Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle

Treball Final de Màster

Màster Universitari en Enginyeria de Telecomunicació

Análisis de los datos recogidos en la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat en el periodo 2011-2016

Alumne

Julia Santo Domingo Gómez

Professor Ponent

Dr. Xavier Vilasis i Cardona

ACTA DE L'EXAMEN
DEL TREBALL FI DE CARRERA

Reunit el Tribunal qualificador en el dia de la data, l'alumne

Dña. Julia Santo Domingo Gómez

va exposar el seu Treball de Fi de Carrera, el qual va tractar sobre el tema següent:

“Análisis de los datos recogidos en la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat en el periodo 2011-2016"

Acabada l'exposició i contestades per part de l'alumne les objeccions formulades pels Srs. membres del tribunal, aquest valorà l'esmentat Treball amb la qualificació de

Barcelona,

VOCAL DEL TRIBUNAL VOCAL DEL TRIBUNAL

PRESIDENT DEL TRIBUNAL

Resumen:

Actualmente, el análisis de datos o lo que se denomina como *business intelligence,* forma cada vez más, parte de nuestra vida cotidiana y día tras día las empresas dedican tanto recursos humanos como económicos para encontrar aquellas claves que les ayuden a aumentar sus ingresos.

Existen numerosas formas y métodos para la minería de datos, algunos de los cuales serán descritos en el presente Trabajo de Fin de Máster. Adicionalmente se analizarán los diferentes softwares de análisis de datos existentes y al igual que hace un analista en su empresa, se decidirá mediante una elección lógica cuál de ellos se usará para ejecutar este proyecto.

Mediante el software elegido se llevará a cabo un análisis y exploración de los datos facilitados por la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat, y se seguirán una serie de objetivos preliminares que convergerán en el siguiente objetivo final: modelizar la variable más característica de un proceso de atención al cliente, el tiempo de espera.

Dado que el análisis de datos se encuentra en pleno auge y es un factor determinante para muchas empresas tanto del sector público como privado, se ha desarrollado este TFM con la intención de generar ciertas conclusiones que sean de gran ayuda para los responsables del Ayuntamiento de Sant Cugat y que con ello puedan optimizar y mejorar la atención del ciudadano en sus oficinas.

Palabras clave: Business Intelligence, análisis de datos, teoría de colas, RStudio, modelos de datos.

Summary:

Currently, the data analysis or what is known as business intelligence, increasingly forms part of our daily life and day after day companies dedicate both human and financial resources to find those keys that help them increase their income.

There are numerous forms and methods for data mining, some of which will be described in this Master's Thesis. Additionally, the different existing data analysis softwares will be analyzed and, as an analyst does in his company, a logical choice will be made in order to determinate which of them is better to execute this project.

Once the software is chosen, an analysis and exploration of the data provided by the citizen service office of the City of Sant Cugat will be carried out, and a series of preliminary objectives will be followed in order to reach the final objective: modeling the most characteristic variable of a customer service process, the waiting time.

Since the data analysis is trending nowdays and is a key factor for many companies in both the public and private sectors, this TFM has the intention of generate solid conclusions that will help the head of the customer service office in the Sant Cugat's town hall to optimize and improve the attention of the citizen in their offices.

Key Words: Business Intelligence, data analysis, queuing theory, RStudio, data models.

Resum:

l'anàlisi de dades o el que es s'anomena com business intelligence, forma part cada cop més, de la nostre vida quotidiana, i dia rere dia, les empreses dediquen recursos humans i econòmics per trobar aquelles claus que les ajudin a augmentar els seus ingressos.

Existeixen nombroses formes i mètodes per la mineria de dades, alguns dels quals seran presentats al Treball de Fi de Màster. Addicionalment s'analitzaran els diferents softwares d'anàlisi de dades existents, i del mateix mode que fa un analista en la empresa, es decidirà mitjançant una elecció lògica quin d'ells s'utilitzarà per dur a terme aquest projecte.

Mitjançant el software escollit es portarà a terme un anàlisi i exploració de les dades facilitades per l'oficina d'atenció al ciutadà de l'Ajuntament de Sant Cugat, i es compliran una sèrie d'objectius preliminars que convergiran al següent objectiu final: modelitzar la variable més característica d'un procés d'atenció al client, el temp d'espera.

Donat que l'anàlisi de dades es troba en ple creixement i es un factor determinant per moltes empreses, tant del sector públic com del privat, s'ha desenvolupat aquest TFM amb la intenció de generar certes conclusions que siguin de gran ajuda pels responsables de l'Ajuntament de San Cugat, i que amb això, puguin optimitzar i millorar l'atenció al ciutadà a les seves oficines.

Palabras clave: Business Intelligence, anàlisi de dades, teoria de cues, RStudio, models de dades.

Agradecimientos

A mi tutor, Xavier por sus buenos consejos, su ayuda y su paciencia, desde el primer día hasta hoy.

A los responsables de la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat, por confiar en LA Salle y en mí para la realización de este trabajo de análisis de datos.

Al resto de profesores de La Salle por todo su conocimiento y por los grandes retos que nos plantearon durante el curso para formarnos como unos excelentes ingenieros de telecomunicaciones.

A mi padre y a mi madre, por estar siempre a mi lado, por apoyarme y quererme siempre, y por demostrarme que con fuerza y voluntad se puede llegar donde uno quiera.

A mi hermana, por ser un ejemplo de valentía.

A todos los compañeros del MET, por su ayuda y colaboración durante todo el curso, y en especial a Luis, por tantas horas de estudio conjunto.

A Kevin, Flor, Rafa, José, Carlos y resto de personas maravillosas que me ha regalado la RESA ya que, gracias a ellos, mi estancia en Barcelona resultó ser una de las mejores experiencias de mi vida.

Gracias en general, a todas aquellas personas que de una manera u otra han colaborado en la satisfactoria realización de este Trabajo de Fin de Máster.

Índice de contenidos

1. INTRODUCCIÓN

Introducción y estado del arte

El proceso de extracción de patrones a partir de datos se llama minería de datos. Es reconocida como una herramienta esencial de los negocios modernos, ya que es capaz de convertir los datos en inteligencia de negocios (*Business Intelligence* o *BI*) dando así una ventaja de información. Actualmente, es ampliamente utilizado en las prácticas de perfil, como vigilancia, comercialización, descubrimientos científicos, y detección de fraudes. (Jimmy W. Maco Elera, 2017)

El *Business Intelligence*, o BI, es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios. (Sinergia e Inteligencia de Negocio S.L., s.f.)

Desde un punto de vista más pragmático, y asociándolo directamente con las tecnologías de la información, podemos definir *Business Intelligence* como el conjunto de metodologías, aplicaciones y herramientas que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (*reporting*, análisis, alertas...) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.

La inteligencia de negocio actúa como un factor estratégico para una empresa u organización, generando una potencial ventaja competitiva, que no es otra que proporcionar información privilegiada para responder a los problemas de negocio: entrada a nuevos mercados, promociones u ofertas de productos, eliminación de islas de información, control financiero, optimización de costes, planificación de la producción, análisis de perfiles de clientes, rentabilidad de un producto concreto, etc...

Los sistemas y componentes del BI se diferencian de los sistemas operacionales en que están optimizados para preguntar y divulgar sobre datos. Esto significa típicamente que, en un *datawarehouse*, los datos están desnormalizados para apoyar consultas de alto rendimiento, mientras que en los sistemas operacionales suelen encontrarse normalizados para apoyar operaciones continuas de inserción, modificación y borrado de datos.

En definitiva, una solución BI completa permite:

- Observar ¿qué está ocurriendo?
- Comprender ¿por qué ocurre?
- Predecir ¿qué ocurriría?
- Colaborar ¿qué debería hacer el equipo?
- Decidir ¿qué camino se debe seguir?

Hay cuatro tipos de tareas que normalmente se involucran en la minería de datos:

- **Clasificación** – la tarea de generalizar una estructura familiar para utilizarla en los nuevos datos.
- **Agrupamiento** – la tarea de encontrar grupos y estructuras en los datos que son de alguna manera u otra lo mismo, sin necesidad de utilizar las estructuras observadas en los datos.
- **Aprendizaje de reglas de asociación** – Busca relaciones entre las variables.
- **Regresión** – Su objetivo es encontrar una función que modele los datos con el menor error.

En capítulos posteriores se detallarán algunas de las mejores herramientas de software para el análisis de datos.

En este capítulo de contextualización de este trabajo de fin de máster es conveniente también hablar de la teoría de colas y de cómo afectan en la calidad de servicio que percibe un cliente que accede al sistema.

A propósito de la teoría de colas, Marcos Singer cree que concentrar los requerimientos en una cola única también puede mejorar la calidad del servicio, porque evita que el bloqueo de una cierta estación perjudique a los ciudadanos que están esperando en su cola respectiva.

Cualquiera que sea el diseño del sistema, los tiempos de espera ocurren por la variabilidad de la llegada y del lapso de servicio, lo cual implica que regularizar ambos procesos redunda en un mejor servicio. Además de la habilitación de recursos adicionales, existen diversas maneras de mejorar la percepción subjetiva del servicio.

Así mismo, es sabido que tomar decisiones correctas en el diseño de los sistemas, permiten entregar un mejor servicio sin necesariamente habilitar recursos adicionales. Por lo tanto, la teoría de colas puede ser una herramienta competitiva que le permite a la organización entender y por ende acceder a su frontera de posibilidades (Marcos Singer, 2008).

Motivación y objetivos

Una vez descrita la contextualización en la que se ha realizado este trabajo fin de máster, puede describirse el fin último con el que se ha realizado:

Análisis temporal y optimización del conjunto de la oficina y de los servicios ofrecidos en la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat.

Cuando un cliente (también llamado ciudadano) termina el proceso regular y abandona la oficina del Ayuntamiento de Sant Cugat debe de haber pasado por tres eventos temporales: llegada, inicio de atención y final de servicio. En el capítulo seis de este TFM nos centraremos con más profundidad en el estudio del tiempo de espera, es decir, el tiempo que transcurre desde que el usuario llega a las oficinas de atención al ciudadano del Ayuntamiento de Sant Cugat hasta que es atendido por alguno de los trabajadores (gestores) del ayuntamiento.

Para llegar a este objetivo final, se han establecido objetivos parciales para guiarnos en el desarrollo de este trabajo fin de máster, que se desarrollan completamente en el capítulo seis de esta memoria. Los objetivos parciales han sido los siguientes:

1. **Caracterización de las variables numeradas en el apartado III, que influyen en los servicios que se realizan en la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat durante los años 2011,2012,2013,2014,2015 y 2016; análisis exploratorio de los datos.**
2. **Optimización del funcionamiento de la oficina de atención al ciudadano. Relación y dependencias existentes entre las variables del modelo estudiado en este TFM: Análisis estadístico.**

Personalmente creo que en este trabajo de fin máster, me centro en realizar un estudio del arte del BI actual y de las diferentes herramientas y software existentes hasta el día de hoy para trabajar con bases de datos. Finalmente, he querido desarrollar un modelo matemático que describe una de las variables más importantes en la gestión de colas del Ayuntamiento de Sant Cugat: **el tiempo de espera**.

Fases y métodos

Para conseguir los objetivos propuestos, se han seguido los siguientes pasos en el desarrollo del proyecto:

- ✓ Estudio del panorama actual del BI y del análisis de datos.
- ✓ Documentación. Realización de un análisis de teoría de colas, redes de colas y modelos de predicción. Estudio de las diferentes herramientas de análisis de datos y elección del software con el que se realizará este TFM.
- ✓ Análisis de diferentes modelos lineales: regresión lineal, anova y ancova.
- ✓ Desarrollo del caso de estudio planteado en este TFM. Ha sido la fase del proyecto que más tiempo ha requerido, debido a la falta de experiencia en la programación en R. Antes de comenzar con el desarrollo del objetivo número dos, hubo una fase de documentación y aprendizaje de *R* y *RStudio*. Esta fase previa al desarrollo se realizó con mucha dedicación e ilusión, pues lo que estaba estudiando tenía un objetivo definido en mi TFM y se realizaba para conseguir desarrollar con éxito el objetivo buscado.
- ✓ Conclusiones y líneas futuras para mejorar y ampliar los objetivos de este trabajo de fin de máster.

Estructura de la memoria

La presente memoria se ha estructurado en ocho capítulos, en los que se pretende abarcar los ámbitos más importantes del desarrollo de un proyecto.

En el presente capítulo se hace una introducción al tema tratado en este trabajo de fin de máster, y se detallan los pasos que se han seguido en el desarrollo de este proyecto.

En el siguiente capítulo, se detallará la problemática que se encontraron en el ayuntamiento de San Cugat para gestionar los asuntos de los usuarios que, acuden diariamente a sus oficinas de atención al ciudadano.

En los tres capítulos posteriores se analiza de manera teórica la teoría de colas, se hace un estudio de las diferentes herramientas de análisis de datos y de los modelos lineales.

A continuación, se definirá la parte técnica del proyecto, y se describirán y desarrollarán los objetivos marcados para conseguir el objetivo final marcado en este trabajo fin de máster.

El capítulo séptimo trata de dar a conocer los gastos que supone la realización de un proyecto como el redactado en esta memoria. Por este motivo y para acercar este TFM lo más posible a un proyecto real, se ha llevado a cabo la realización de una estimación de un presupuesto económico, en el que se ha buscado ajustar los gastos a la realidad, de la manera más precisa posible.

Para concluir esta memoria se describirán en el capítulo octavo los resultados obtenidos en la misma, valorándose si se ha conseguido alcanzar los objetivos marcados en el capítulo uno.

Finalmente expondré mi propia opinión personal, y se explicarán algunas ideas que podrían llevarse a cabo en un futuro para mejorar este proyecto, optimizarlo y poder seguir beneficiando tanto al Ayuntamiento de Sant Cugat como a los usuarios finales.

2. ANÁLISIS Y ESTUDIO DEL PROBLEMA REFERIDO A ESTE TFM:

Situación de partida: Situación actual en las oficinas del Ayuntamiento de San Cugat

Se parte de la siguiente premisa: el ayuntamiento de San Cugat tiene la sensación de que los usuarios de su atención al cliente esperan demasiado tiempo hasta ser atendidos, exceptuando momentos puntuales. Tras realizar una escueta sencilla de satisfacción comprueban que efectivamente la satisfacción del servicio por parte de los usuarios es muy baja.

Llegado este punto, en enero de 2011, los responsables de las oficinas de atención al usuario del ayuntamiento de Sant Cugat inician mecanismos para tratar de estudiar el problema con el tiempo de espera en sus oficinas. Gracias a este motivo comienzan a tomar diferentes medidas que se irán guardando en un Excel: tiempos de llegada y salida de los diferentes usuarios durante un periodo de seis años, junto con el gestor que los atiende.

Su intención es que con esta información sean capaces de encontrar los motivos de la baja satisfacción de los ciudadanos y accionar palancas que mejoren la atención al cliente.

Una vez tienen información suficiente para estudiarla y analizarla, se ponen en contacto con "La Salle, Universitat Ramon Llull", de la cuál soy alumna, para tratar de llegar a un acuerdo: buscar a un alumno interesado en realizar su TFM analizando y estudiando sus datos.

Los profesores del máster, Xavier y Miguel, nos encomiendan dicha tarea repartida a Rachid y a mí, quienes encantados realizaremos el trabajo de fin de máster sobre la atención al ciudadano en las oficinas del ayuntamiento de San Cugat.

Personalmente, una vez finalice la elaboración de esta memoria, y junto con Xavier, tutor de este TFM, haremos llegar este TFM a los interlocutores del ayuntamiento para que puedan hacer uso de la información contenida en él y puedan tomar las medidas que consideren oportunas de cara a mejorar su servicio.

¿Cómo vincular la teoría de colas a la gestión de la oficina de atención al ciudadano del Ayuntamiento de San Cugat?

¿Por qué hablamos de teoría de colas? Porque la teoría de colas apoya la gestión de las empresas y organizaciones que atienden público, cuantificando la manera en que se combinan los indicadores de efectividad (calidad del servicio), de eficiencia (uso de recursos) e internos (de diseño del sistema).

Hoy en día la relación entre teoría de colas y la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat puede basarse en las siguientes cinco premisas que se detallan a continuación:

1. Impacto del tiempo de espera en la calidad del servicio

La opinión que se forman los clientes y usuarios acerca del servicio que se les entrega depende de diversos aspectos subjetivos, tales como la capacidad técnica de quienes atienden, la

amabilidad del trato, la presentación y la limpieza. El mal servicio perjudica la reputación de la firma mucho más que el buen servicio la favorece.

Un aspecto determinante para la calidad del servicio es el tiempo que se debe esperar para obtenerlo. El tiempo se divide en dos componentes: el lapso de servicio y el tiempo de espera. En general se prefiere tiempos de atención breves. (Marcos Singer, 2008)

Es por este motivo que en el capítulo seis de este documento se analiza específicamente la variable denominada como "tiempo de espera", y se analizarán todas aquellas variables que puedan influenciar a ésta.

2. Conjunto de indicadores del desempeño de eficacia y eficiencia con respecto a la teoría de colas

Algunos de los indicadores de gestión más relevantes son los siguientes:

- <u>Parámetros de diseño:</u>
  - λ [u/h] flujo promedio o tasa de recepción de órdenes de atención. (Medido en unidades por hora).

    λ [u/h] = 1/ lapso entre llegadas consecutivas [h/u]
  - μ [u/h]: flujo promedio o tasa de atención de cada servidor cuando opera a máxima capacidad, medido en unidades por hora.

    μ [u/h] = 1/ lapso de servicio [h/u].
  - σ [h/u]: desviación estándar del lapso de servicio. Por definición es la raíz cuadrada de la varianza, que es igual al valor esperado de las diferencias al cuadrado entre cada una de las observaciones y su promedio.
  - k: número de servidores o "recursos" de la estación de trabajo. Por lo tanto, k es una medida de capacidad de atención

- <u>Indicadores de interés del administrador de eficiencia:</u>
  - I: número de servidores ocupados en el sistema. También puede referirse al número de órdenes procesadas en la estación de trabajo. El valor promedio o esperado de I se representa como $\bar{I}$.

    Este inventario promedio se relaciona con el flujo de salida y con el lapso de servicio a través de la *Ley de Little (1961),* posiblemente la fórmula más importante de la teoría de colas:

    Número de ordenes promedio= flujo de salida * lapso de servicio

    Para clarificar la Ley de Little, podemos ver el siguiente ejemplo:

$$\bar{I} = \frac{\lambda}{\mu}$$

- ρ: factor de utilización. Es igual a la razón entre el número promedio de servidores ocupados y el número total de servidores:

$$\rho = \frac{\overline{I}}{k} = \frac{\lambda}{k\,\mu}$$

Los siguientes indicadores se relacionan con la cantidad de clientes en el sistema de espera.

- Q [u]: tamaño de la cola de espera medido en unidades [u]. Su valor promedio es Q.
- N [u]: población de clientes en el sistema medido en unidades [u], es igual a:

$$N = Q + I$$

3. Caracterizar el comportamiento aleatorio de los clientes: Proceso exponencial y distribución de Poisson

Tanto los indicadores del administrador como los del cliente dependen del comportamiento de dos tipos de eventos:

- La llegada de un cliente
- La ejecución de una atención por parte de una persona de la oficina de atención al cliente del Ayuntamiento de Sant Cugat.

La variable aleatoria que mide el intervalo entre la ocurrencia de dos eventos consecutivos tiene una distribución de probabilidad exponencial cuando se dan las siguientes condiciones (Marcos Singer, 2008):

- El número de eventos que ocurren es proporcional al intervalo de tiempo que se considera. Por ejemplo, las ocurrencias durante dos semanas duplican a las ocurrencias durante una semana.
- No pueden ocurrir dos o más eventos de manera simultánea.
- La ocurrencia de un evento no influencia la ocurrencia de un evento posterior.

Estas tres premisas se cumplen en la oficina de atención al ciudadano/cliente del Ayuntamiento de San Cugat, por lo que se puede afirmar que la ocurrencia de dos eventos consecutivos tiene una distribución de probabilidad exponencial.

Dado que esta probabilidad no se ve modificada si inmediatamente antes ocurrió una llegada, los procesos exponenciales se denominan “sin memoria”. Los procesos que muestran estas características son los que cumplen las siguientes condiciones. Primero, los clientes llegan uno por uno, no en grupos. Segundo, no se influencian unos a otros; cada llegada es independiente de la otra. Tercero, el número de llegadas en un periodo de tiempo determinado es, en promedio, proporcional al tamaño de dicho periodo (Marcos Singer, 2008).

Se comprueba que el proceso de llegada tiene una distribución de tipo exponencial ya que el histograma de los intervalos entre las llegadas es similar a la función de densidad de la distribución exponencial: $\lambda \exp(-\lambda t)$.

4. Vincular los indicadores de desempeño en términos del comportamiento de los clientes: Modelos de espera típicos:

   En 1953 David Kendall clasificó los sistemas de espera mediante la nomenclatura A/B/k, donde:

   - A: tipo de distribución de probabilidad de tiempo entre arribos consecutivos;
   - B: tipo de distribución de probabilidad del tiempo de servicio o atención;
   - k: número de servidores de la estación de trabajo.

   Las distribuciones de probabilidad más estudiadas son (Marcos Singer, 2008):

   - M: distribución exponencial o sin memoria (la "M" viene del inglés *memoryless*)

   La nomenclatura de Kendall se extiende a A/B/k/C/N/D, donde:

   - C: capacidad total del sistema, con $C \geq k$. Al igual que en el ejemplo de la estación de servicio, una capacidad limitada puede hacer que se pierdan clientes;
   - N: tamaño de la población desde la que se obtienen los clientes;
   - D: disciplina o política de prioridad con la que se atienden los clientes, la que determina diversos aspectos de calidad del servicio.

   Omitir "/C/N/D" significa que $C = \infty$, $N = \infty$ y D es FIFO. Hacer $C = \infty$ equivale a suponer que los clientes siempre se ponen a la cola, no importa qué tan larga sea ésta. En la práctica, por otro lado, los clientes estiman la espera de la cola y deciden quedarse dependiendo de un cierto valor umbral. Una vez que están en la cola, es raro que la abandonen si no han cambiado las condiciones de servicio.

   La disciplina FIFO cumple con la norma de justicia de atender primero a quien lleva esperando más tiempo, aunque no necesariamente es la disciplina más aconsejable. En el caso de estudio de la oficina de atención al cliente del ayuntamiento de San Cugat se sigue una disciplina FIFO.

   Colas con servidores en paralelo **M/M/k** → Caso en que se encuentra la oficina de atención al cliente del Ayuntamiento de San Cugat para todas aquellas personas que van a realizar **trámites** al ayuntamiento **de la misma naturaleza.**

   En el sistema M/M/k se tiene una llegada de clientes de tipo exponencial con tasa λ y un tiempo de servicio también exponencial de tasa m para cada uno de sus k servidores.

   Un sistema con servidores en paralelo, refiriéndose con *servidores* a los recursos o personas que se dedican a atender a los usuarios de las oficinas de atención al cliente e cuestión, se caracteriza porque hay más de un servidor que ejecuta la misma función con la misma eficiencia. En un sistema con servidores en paralelo no hay varias colas, sino una única cola. (García Sabater, 2016)

5. Conclusiones teoría de colas

   Dependiendo de las características específicas de cada sistema, las fórmulas muestran de manera estilizada las relaciones de transacción entre estos indicadores. Algunas

fórmulas muestran cómo aumenta la probabilidad de que el cliente sea atendido de inmediato cuando se habilitan nuevas estaciones de trabajo. Otras indican cuántos puestos de atención al cliente deben estar ociosas en promedio para ofrecer un determinado tiempo de espera promedio a los clientes.

Todas estas fórmulas son fácilmente definibles en una planilla de cálculo, lo que permite tomar decisiones rápidas con relativamente pocos datos. Si bien la teoría de colas es muy usada para evaluar los sistemas de servicio, existen alternativas tales como la simulación computacional o la prueba y error. (Marcos Singer, 2008)

La teoría de colas hace explícitas las relaciones de causalidad, lo que permite ganar una mayor comprensión de los sistemas estudiados.

Información recibida del Ayuntamiento

El ayuntamiento de Sant Cugat facilitó un Excel como el de la siguiente ilustración donde se recogen todos los datos recopilados en sus sistemas desde el año 2011 hasta el año 2016 incluido.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Cua /Oficina	Data	Año	Di...	Hora_ticke	Numero t	Codi Ges	Hora inici aten	Hora final aten	Tipologia tràmite	F	Da	Tiempo de esp	Tiempo de serv
2	Atenció Ciutadana / Registre	03/01/2011	2011	1	8:00:37	73	1	8:01:01	8:08:36	Certificats digitals	1	0	0:00:24	0:07:35
3	Atenció Ciutadana / Registre	03/01/2011	2011	1	8:03:54	80	2	8:04:38	8:06:50	Altres Registres	1	0	0:00:44	0:02:12
4	Atenció Ciutadana / Registre	03/01/2011	2011	1	8:05:31	81	1	8:08:36	8:16:15	Informacions diverses	1	0	0:03:05	0:07:39
5	Ocupació (SOM)	03/01/2011	2011	1	8:12:55	1	3	8:27:38	8:31:35	Altres	1	0	0:14:43	0:03:57
6	Atenció Ciutadana / Registre	03/01/2011	2011	1	8:18:20	82	1	8:18:38	8:23:52	Padró d'habitants-Volants	1	0	0:00:18	0:05:14
7	Atenció Ciutadana / Registre	03/01/2011	2011	1	8:33:17	83	4	8:33:59	8:56:49	Registre únic	1	0	0:00:42	0:16:50
8	Atenció Ciutadana / Registre	03/01/2011	2011	1	8:50:35	84	1	8:51:30	8:57:54	Registre únic	1	0	0:00:55	0:06:24
9	Atenció Ciutadana / Registre	03/01/2011	2011	1	8:55:18	85	4	8:56:51	9:00:37	Padró d'habitants-Modificacions	1	0	0:01:33	0:03:46
10	Atenció Ciutadana / Registre	03/01/2011	2011	1	8:56:30	86	4	9:00:37	9:03:29	Altres Registres	1	0	0:04:07	0:02:52
11	Atenció Empresa	03/01/2011	2011	1	9:00:44	88	5	9:01:18	9:06:00	Informació	1	0	0:00:34	0:04:42
12	Llicències d'Obres	03/01/2011	2011	1	9:00:57	55	6	9:01:00	9:31:00	Indefinit	1	0	0:00:03	0:30:00
13	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:01:09	87	4	9:03:29	9:10:46	Padró d'habitants- Nous procediments	1	0	0:02:20	0:07:17
14	Atenció Empresa	03/01/2011	2011	1	9:03:27	89	5	9:06:01	9:13:54	Informació	1	0	0:02:34	0:07:53
15	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:06:16	88	1	9:09:47	9:12:28	Padró d'habitants-Volants	1	0	0:03:31	0:02:41
16	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:10:47	87	4	9:10:47	9:18:58	Altres Registres	1	0	0:00:00	0:08:11
17	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:11:08	89	1	9:12:29	9:20:32	Informacions diverses	1	0	0:01:21	0:08:03
18	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:15:00	90	4	9:18:59	9:21:47	Altres Registres	1	0	0:03:59	0:02:48
19	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:15:51	91	1	9:20:33	9:29:16	Padró d'habitants-Modificacions	1	0	0:04:42	0:08:43
20	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:18:00	92	4	9:21:49	9:25:56	Altres Registres	1	0	0:03:49	0:04:07
21	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:18:47	93	4	9:26:14	9:33:00	Altres Registres	1	0	0:07:27	0:10:46
22	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:19:05	94	1	9:29:27	9:36:19	Padró d'habitants-Modificacions	1	0	0:10:22	0:06:52
23	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:25:58	92	4	9:25:58	9:27:45	Altres Registres	1	0	0:00:00	0:01:47
24	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:27:42	95	7	9:30:41	9:36:36	Padró d'habitants-Volants	1	0	0:02:59	0:05:55
25	Atenció Empresa	03/01/2011	2011	1	9:27:54	90	5	9:29:11	9:30:04	Derivació	1	0	0:01:17	0:00:53
26	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:28:06	96	7	9:36:54	9:39:10	Informacions diverses	1	0	0:08:48	0:02:16
27	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:29:06	97	4	9:39:02	9:40:30	Padró d'habitants-Volants	1	0	0:09:56	0:01:28
28	Llicències d'Obres	03/01/2011	2011	1	9:30:09	56	6	9:31:01	9:38:25	Indefinit	1	0	0:00:52	0:07:24
29	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:35:10	98	7	9:39:11	9:46:36	Altres Registres	1	0	0:04:01	0:07:25
30	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:36:08	99	4	9:40:30	9:42:33	Padró d'habitants-Volants	1	0	0:04:22	0:02:03
31	Llicències d'Obres	03/01/2011	2011	1	9:36:32	57	6	9:38:26	9:40:38	Indefinit	1	0	0:01:54	0:02:12
32	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:37:27	1	4	9:42:41	9:43:00	No presentat	1	0	0:05:14	0:00:19
33	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:37:47	2	4	9:43:02	9:52:53	Altres Registres	1	0	0:05:15	0:09:51
34	Llicències d'Obres	03/01/2011	2011	1	9:38:32	58	6	9:40:40	9:41:54	Indefinit	1	0	0:02:08	0:01:14
35	Llicències d'Obres	03/01/2011	2011	1	9:38:52	59	6	9:41:56	9:43:48	Indefinit	1	0	0:03:04	0:01:52
36	Llicències d'Obres	03/01/2011	2011	1	9:42:11	60	6	9:43:49	10:24:48	Indefinit	1	0	0:01:38	0:40:59
37	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:48:49	3	7	9:51:53	9:59:59	Altres Registres	1	0	0:03:04	0:08:06
38	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:49:04	4	2	9:53:01	10:05:25	Altres Registres	1	0	0:03:57	0:12:24
39	Ocupació (SOM)	03/01/2011	2011	1	9:49:17	2	8	9:50:01	10:01:43	Entrevista seguiment	1	0	0:00:44	0:11:42
40	Atenció Ciutadana / Registre	03/01/2011	2011	1	9:50:48	5	4	10:00:42	10:02:44	Altres Registres	1	0	0:09:54	0:02:02
41	Ocupació (SOM)	03/01/2011	2011	1	9:51:09	3	8	10:01:45	10:03:24	No presentat	1	0	0:10:36	0:01:39

Ilustración 1 Captura de pantalla del Excel recibido del Ayto. de Sant Cugat

En este Excel se recoge la información del siguiente modo:

Cada fila del Excel, del número 2 a la numero 415615, hace referencia a cada persona que ha acudido al ayuntamiento desde el día 1 de enero de 2011 hasta el fin de 2016.

Las columnas del Excel recibido del ayuntamiento de Sant Cugat hacen referencia por el siguiente orden a:

- **Columna A**: Oficina
- **Columna B**: Día
- **Columna C**: Año
- **Columna D**: Hora recogida ticket cita previa
- **Columna E**: Número torn

- **Columna F**: Código Gestor
- **Columna G**: Hora inicio de atención personalizada
- **Columna H**: Hora final de atención personalizada
- **Columna I**: Tipología trámites

Una vez vista la información que se facilitó desde el Ayuntamiento de Sant Cugat, empecé a trabajar con ella y a buscar datos o conclusiones relevantes que me ayudaran a entender las situaciones que se viven día a día en dicha oficina de atención al ciudadano.

Mi compañero de máster Rachid Boujaja y yo misma, Julia Santo Domingo Gómez, nos dispusimos a realizar de manera independiente un primer análisis de los datos.

Se estimó conveniente que una vez realizado este primer análisis de datos nos reuniéramos con nuestros interlocutores en el Ayuntamiento de Sant Cugat para informar a los responsables del departamento de atención al cliente todos nuestros avances con el proyecto y todos aquellos aspectos relevantes que consideramos de especial interés para la mejora del servicio de atención al ciudadano en sus oficinas. Esta reunión fue mantenida en febrero de 2016 en el propio Ayuntamiento de Sant Cugat del Vallés y acudimos a ella: dos profesores de "La Salle, Universitat Ramon Llull", dos alumnos, incluida yo misma y tres interlocutores del Ayuntamiento de Sant Cugat, entre ellos el responsable de las oficinas de atención al ciudadano.

En el siguiente apartado, se han descrito la mayoría de conclusiones que se obtuvieron tras el primer análisis de los datos facilitados y que se mostraron a los asistentes en la citada reunión.

Una vez finalizada la reunión se siguió estudiando más en profundidad y con un software más adecuado dichos datos y se obtuvo otra serie de información que se ha adjuntado en la presente memoria en los capítulos posteriores para que pueda ser leída por todas las personas involucradas y con la que se ha realizado todo el análisis sobre el cual se basa este TFM.

Conclusiones presentadas en la reunión realizada en febrero de 2016 a los interlocutores del Ayuntamiento de Sant Cugat

En este capítulo se describe la información relevante obtenida tras el primer análisis de los datos del Ayuntamiento de Sant Cugat que realicé y presenté a los interlocutores del Ayuntamiento de Sant Cugat en la reunión mantenida en febrero de 2016 y que es necesario compartir en esta memoria.

Importante: los siguientes resultados que se muestran son con los datos del primer *excel* recibido del ayuntamiento (400324 líneas) y que solo contenía los datos desde enero de 2011 hasta septiembre de 2016, ya que el resto de datos hasta completar el año no los tenían disponibles para facilitárnoslos.

Nueve meses después nos enviaron una nueva versión del Excel con los datos hasta diciembre de 2016 (415615 líneas), y con los que se realizó el estudio posterior mostrado en capítulos posteriores.

Tiempo medio y tiempo espera

En la primera taba se ha calculado la media de los tiempos de espera y de servicio según el tipo de oficina. En la segunda tabla que se observa a continuación, se indican los tiempos medios de espera y de servicio totales para cada año.

Tiempo medio para cada oficina	Tiempo medio de espera	Tiempo medio de servicio
Altes padró i canvis domicili	0:12:22	0:10:22
Atenció Ciutadana / Registre	0:11:52	0:07:54
Atenció Empresa	0:07:53	0:11:15
Benestar Social i Família (atenció a partir de les 10h)	0:59:36	0:06:51
Llicències d\'Obres	0:06:59	0:15:56
Ocupació (SOM)	0:12:52	0:11:16
Oficina de consum (OMIC)	0:20:51	0:11:35
Oficina Educació	0:14:41	0:07:20
Registre Empresa	0:04:20	0:06:35
Taxes cementiri	0:13:01	0:07:34

Año	Tiempo medio espera	Tiempo medio Servicio
2016	0:22:25	0:09:42
2015	0:12:32	0:08:20
2014	0:10:31	0:08:42
2013	0:10:59	0:09:20
2012	0:12:03	0:08:03
2011	0:08:36	0:06:05

Tabla 1Tiempo medio de espera para cada oficina y por año

- *Desviación estándar y tiempos promedio*

Standar desviation tiempo espera	0:15:27
Standar desviation tiempo servicio	0:10:02
Promedio tiempo espera	0:12:18
Promedio tiempo de servicio	0:08:15

- *Tiempo total de espera y de servicio*

Se ha calculado también el tiempo exacto que se ha empleado en cada servicio. En la segunda columna (azul) el tiempo de espera y de servicio total se ha expresado en días, horas, minutos y segundos.

	TIEMPO TOTAL DE ESPERA	TIEMPO ESPERA DIAS Y HORAS		TIEMPO TOTAL DE SERVICIO	TIEMPO SERVICIO DIAS Y HORAS	
Altes padró i canvis domicili	1,176851852	1	4:14:40	0,986736111	0	23:40:54
Atenció Ciutadana / Registre	2876,729792	2876	17:30:54	1915,583669	1915	14:00:29
Atenció Empresa	41,49326389	41	11:50:18	59,22478009	59	5:23:41
Benestar Social i Família (atenció a partir de les 10h)	138,4552315	138	10:55:32	15,8928588	15	21:25:43
Llicències d\'Obres	21,59268519	21	14:13:28	49,27061343	49	6:29:41
Ocupació (SOM)	215,4381829	215	10:30:59	188,5054514	188	12:07:51
Oficina de consum (OMIC)	13,07758102	13	1:51:43	7,267303241	7	6:24:55
Oficina Educació	99,55340278	99	13:16:54	49,71967593	49	17:16:20
Registre Empresa	0,009016204	0	0:12:59	0,013726852	0	0:19:46
Taxes cementiri	10,50783565	10	12:11:17	6,10380787	6	2:29:29

Tabla 2 Tiempo total de espera y tiempo total de servicio para cada oficna/servicio ofrecido en la oficina de atención al cliente (medido en la primera columna en decimal, y en la segunda columna en dias,horas, minutos y segundos

Número de personas que han acudido a cada oficina

Por oficina se entiende a los diferentes servicios (atención empresas, oficinas de educación, atención ciudadana, …) que tienen atención personalizada en la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat.

Oficina	Suma de PX
Altes padró i canvis domicile (fue un momento puntual solo 1 px)	137
Atenció Ciutadana / Registre	348875
Atenció Empresa	7578
Benestar Social i Família (atenció a partir de les 10h)	3345
Llicències d\'Obres	4451
Ocupació (SOM)	24101
Oficina de consum (OMIC)	903
Oficina Educació	9768
Registre Empresa (fue solo un momento)	3
Taxes cementiri	1162
Total general	**400323**

Tabla 3 Número de personas que han acudido a cada oficina

Oficinas más visitadas cada año

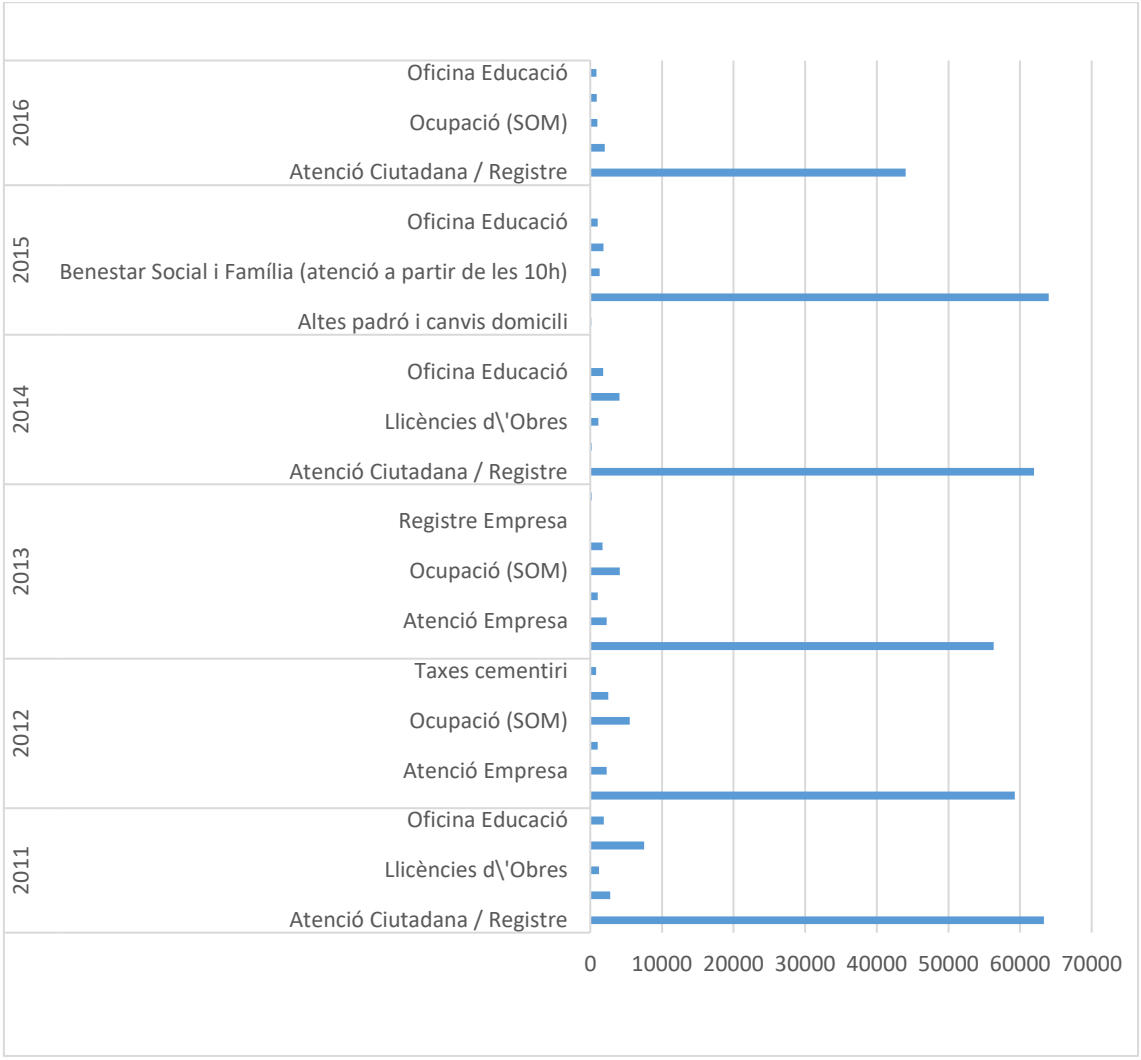


Tabla 4 Oficinas con más afluencia (por año)

Año	Número de Personas
2011	76750
Atenció Ciutadana / Registre	63329
Atenció Empresa	2782
Llicències d\'Obres	1232
Ocupació (SOM)	7509
Oficina Educació	1898
2012	71446
Atenció Ciutadana / Registre	59249
Atenció Empresa	2310

Llicències d\'Obres	1055
Ocupació (SOM)	5519
Oficina Educació	2505
Taxes cementiri	808
2013	65641
Atenció Ciutadana / Registre	56325
Atenció Empresa	2287
Llicències d\'Obres	1026
Ocupació (SOM)	4142
Oficina Educació	1686
Registre Empresa	3
Taxes cementiri	172
2014	69221
Atenció Ciutadana / Registre	61937
Atenció Empresa	199
Llicències d\'Obres	1138
Ocupació (SOM)	4077
Oficina Educació	1784
Taxes cementiri	86
2015	68406
Altes padró i canvis domicili	137
Atenció Ciutadana / Registre	64003
Benestar Social i Família (atenció a partir de les 10h)	1304
Ocupació (SOM)	1845
Oficina Educació	1021
Taxes cementiri	96
2016	48859
Atenció Ciutadana / Registre	44032
Benestar Social i Família (atenció a partir de les 10h)	2041
Ocupació (SOM)	1009

Oficina de consum (OMIC)	903
Oficina Educació	874
Total general	**400323**

Tabla 5 Número de personas que acuden a cada oficina por año

Oficinas más visitadas por horario

Intervalo horario (por hora)	Número de PX
8	21964
Altes padró i canvis domicili	5
Atenció Ciutadana / Registre	20927
Ocupació (SOM)	1026
Oficina Educació	6
9	55366
Altes padró i canvis domicili	14
Atenció Ciutadana / Registre	46859
Atenció Empresa	1128
Benestar Social i Família (atenció a partir de les 10h)	36
Llicències d\'Obres	2071
Ocupació (SOM)	5013
Oficina Educació	14
Registre Empresa	1
Taxes cementiri	230
10	71298
Altes padró i canvis domicili	23
Atenció Ciutadana / Registre	57228
Atenció Empresa	1481
Benestar Social i Família (atenció a partir de les 10h)	1072
Llicències d\'Obres	2363
Ocupació (SOM)	5905
Oficina Educació	2953
Taxes cementiri	273
11	78483

Altes padró i canvis domicili	33
Atenció Ciutadana / Registre	66259
Atenció Empresa	1869
Benestar Social i Família (atenció a partir de les 10h)	1018
Llicències d\'Obres	17
Ocupació (SOM)	6298
Oficina de consum (OMIC)	430
Oficina Educació	2283
Taxes cementiri	276
12	72033
Altes padró i canvis domicili	27
Atenció Ciutadana / Registre	61372
Atenció Empresa	1814
Benestar Social i Família (atenció a partir de les 10h)	729
Ocupació (SOM)	5574
Oficina de consum (OMIC)	261
Oficina Educació	1992
Taxes cementiri	264

13	46937
Altes padró i canvis domicili	19
Atenció Ciutadana / Registre	43228
Atenció Empresa	1276
Benestar Social i Família (atenció a partir de les 10h)	467
Ocupació (SOM)	282
Oficina de consum (OMIC)	189
Oficina Educació	1356
Registre Empresa	2
Taxes cementiri	118
14	9928
Altes padró i canvis domicili	5

Atenció Ciutadana / Registre	9849
Atenció Empresa	10
Benestar Social i Família (atenció a partir de les 10h)	20
Ocupació (SOM)	3
Oficina de consum (OMIC)	20
Oficina Educació	20
Taxes cementiri	1
15	6644
Altes padró i canvis domicili	1
Atenció Ciutadana / Registre	6637
Benestar Social i Família (atenció a partir de les 10h)	3
Oficina de consum (OMIC)	3
16	9353
Altes padró i canvis domicili	3
Atenció Ciutadana / Registre	9031
Oficina Educació	319
17	15875
Altes padró i canvis domicili	3
Atenció Ciutadana / Registre	15365
Oficina Educació	507
18	12004
Altes padró i canvis domicili	4
Atenció Ciutadana / Registre	11685
Oficina Educació	315
19	423
Atenció Ciutadana / Registre	420
Oficina Educació	3
20	15
Atenció Ciutadana / Registre	15
Total general	**400323**

Tabla 6 Oficinas más visitas en cada intervalo horario de una hora

Basándonos en las cifras anteriores y observando la siguiente ilustración (2) se puede concluir sin lugar a dudas que la oficina más demandada por los usuarios es “Atención Ciudadana/Registro”. Esto también puede deberse a que este nombre de oficina es el más genérico (y puede aplicar a todas las solicitudes de atención al cliente por parte de los usuarios) y además es el primero de la lista de oficinas. Es posible que por este motivo muchos usuarios elijan esta oficina para dirigirse a la atención al cliente, aunque exista una oficina más concreta para la atención que desean.

Como se ha apreciado en la tabla e ilustración referidas a esta información, que el número de usuarios que solicitan atención al cliente en la oficina de “Atención ciudadana /registro” es mucho mayor que para el resto de oficinas, el estudio detallado de este TFM que se hará en capítulos posteriores, excluirá la información del resto de oficinas y se centrará en analizar los datos existentes que ocurren exclusivamente en la oficina llamada: “Atención Ciudadana/Registro”.

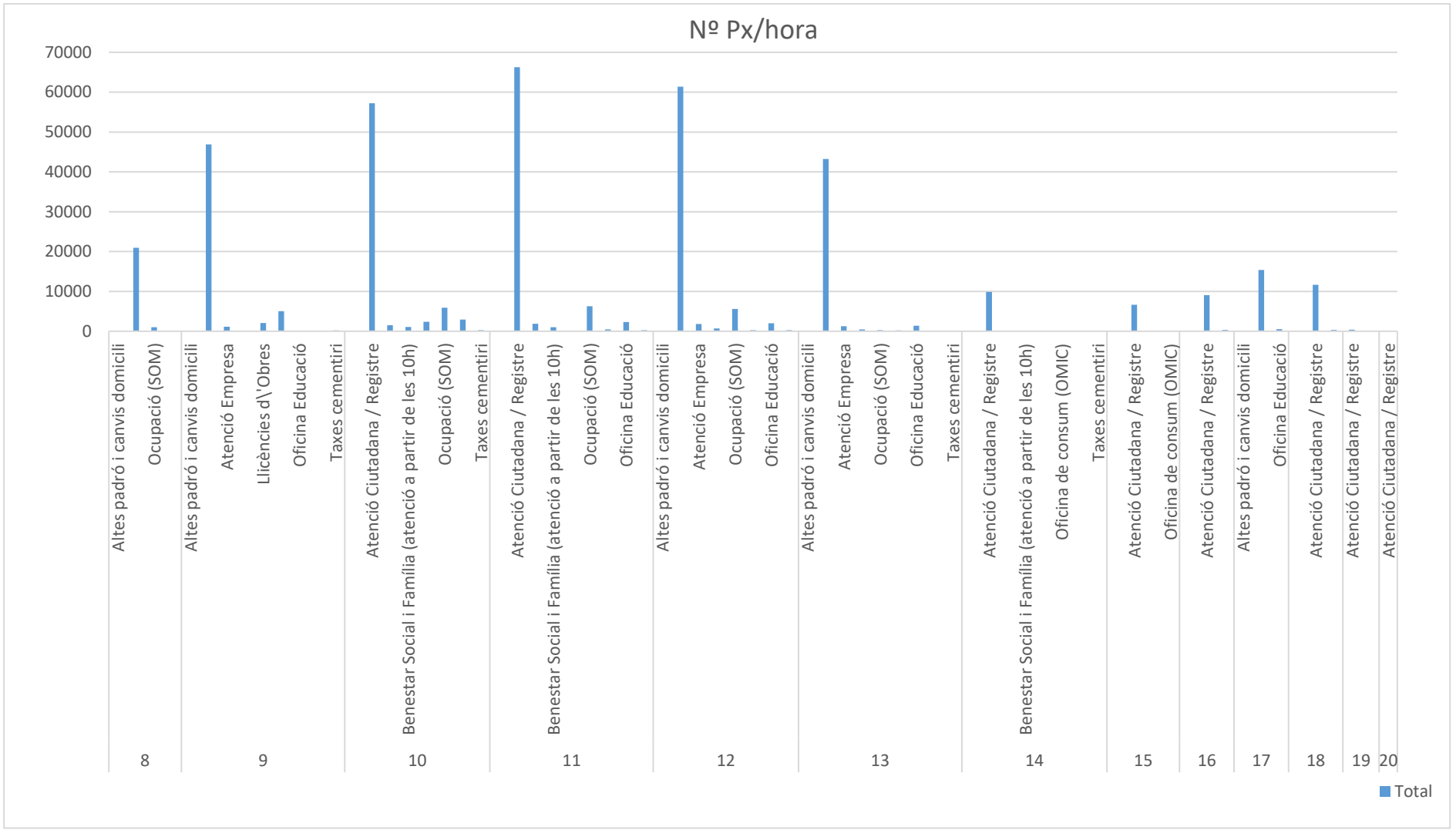


Ilustración 2 Oficinas más visitas en cada intervalo horario de una hora

- *Número de personas que acuden en cada horario*

Año/Hora	8	9	10	11	12	13	14	15	16	17	18	19	20	Total general
2011	4473	10893	13741	15315	14374	8600	1550	1041	1660	2801	2246	55	1	**76750**
2012	3636	9433	12958	14461	13264	8129	1633	1074	1709	2830	2246	72	1	**71446**
2013	3341	8964	11831	13128	11899	7866	1620	1000	1469	2542	1918	63		**65641**
2014	3592	9419	12359	13676	12545	8176	1735	1292	1595	2734	2017	81		**69221**
2015	3737	9620	11810	13081	12111	8210	1969	1249	1678	2784	2094	63		**68406**
2016	3185	7037	8599	8822	7840	5956	1421	988	1242	2184	1483	89	13	**48859**
Total general	**21964**	**55366**	**71298**	**78483**	**72033**	**46937**	**9928**	**6644**	**9353**	**15875**	**12004**	**423**	**15**	***400323***

Tabla 7 Número de personas que acuden por año y por intervalo horario

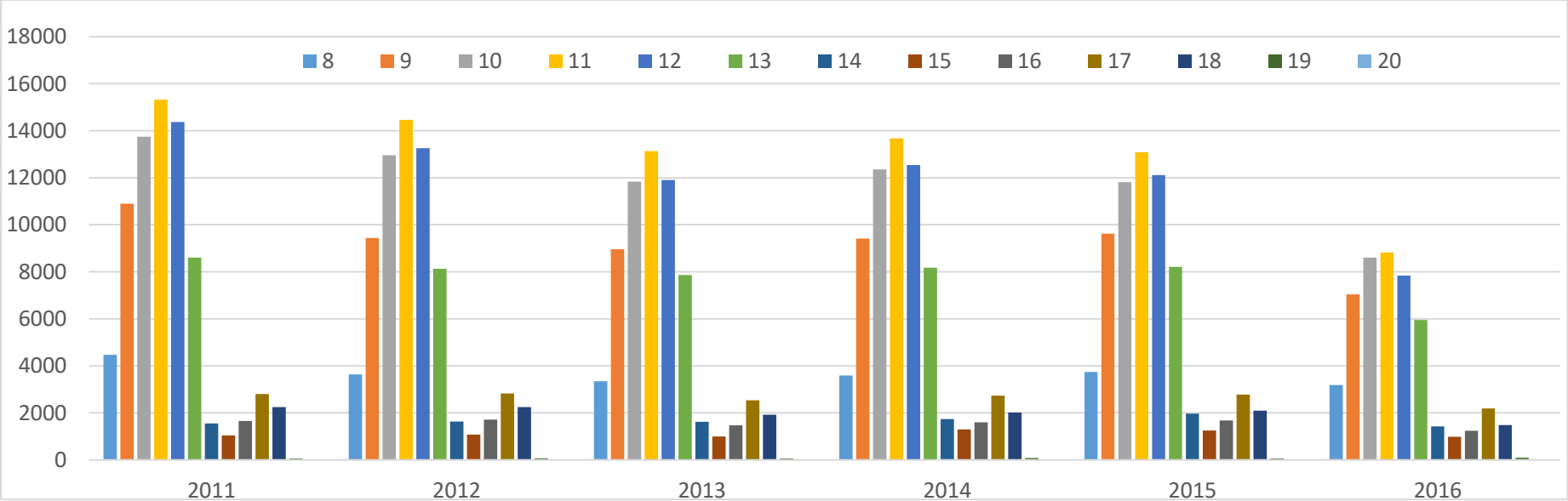


Ilustración 3 Número de personas que acuden por año y por intervalo horario (diferenciación por colores)

- *Número de usuarios que atiende cada gestor durante el periodo 2011-2016*

Código de Gestor	PX Atendidas
1	28615
2	27360
3	2876
4	27404
5	7481
6	271
7	28684
8	1853
9	28693
10	1055
11	24516
12	2477
13	9473
14	9357
15	27837
16	2694
17	25230
18	27118
19	1546
20	30536
21	1852
22	3616
23	1657
24	33947
25	9834
26	13
27	5
28	7883
29	133
30	11838
31	1
32	172
33	37
34	527
35	1

36	3
37	182
38	494
39	400
40	2312
41	6451
42	75
43	97
44	183
45	1036
46	1570
47	11
48	903
49	1
50	13
Total general	**400323**

Tabla 8 Número de usuarios que atiende cada gestor durante el periodo 2011-2016

- *Número de usuarios que acuden por mes y por año*

	2011	2012	2013	2014	2015	2016	
Enero	5925	6404	5499	5305	5446	4385	**32964**
Febrero	6987	5941	5383	6790	5616	5265	**35982**
Marzo	7654	7012	5873	6188	6191	4867	**37785**
Abril	5779	5955	5369	5363	5657	6374	**34497**
Mayo	9014	8671	7078	8220	8689	7897	**49569**
Junio	6650	6971	5915	6041	6018	5719	**37314**
Julio	6094	6168	6197	5382	5496	4996	**34333**
Agosto	3862	2644	3240	3548	3464	3573	**20331**
Septiembre	7770	6104	6262	6305	6126	5783	**38350**
Octubre	6102	6354	6070	6377	5896	0	**30799**
Noviembre	6236	5197	4594	4774	5230	0	**26031**
Diciembre	4677	4025	4161	4928	4577	0	**22368**
	76750	71446	65641	69221	68406	48859	**400323**

Tabla 9 Número de usuarios que acuden cada mes y cada año

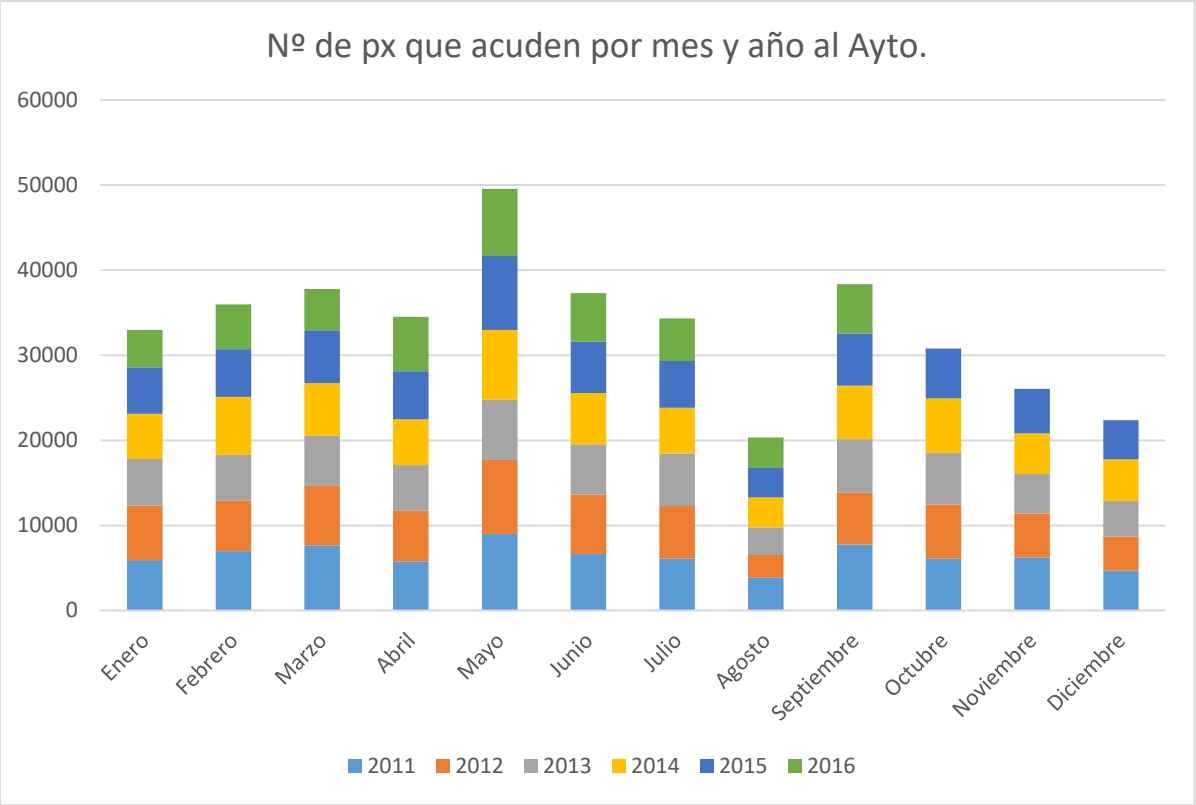


Ilustración 4 Número de usuarios que acuden por mes y año a las oficinas de atención al cliente del Ayto. de San Cugat I

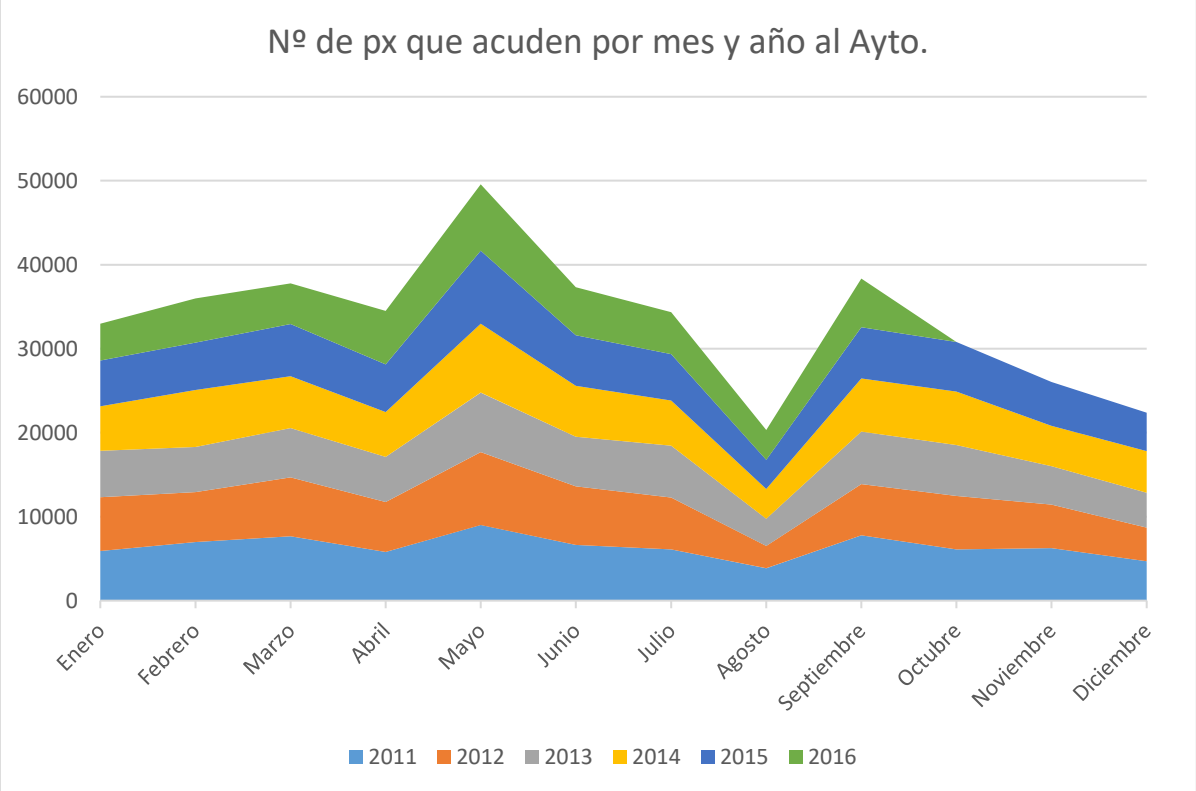


Ilustración 5 Número de usuarios que acuden por mes y año a las oficinas de atención al cliente del Ayto. de San Cugat II

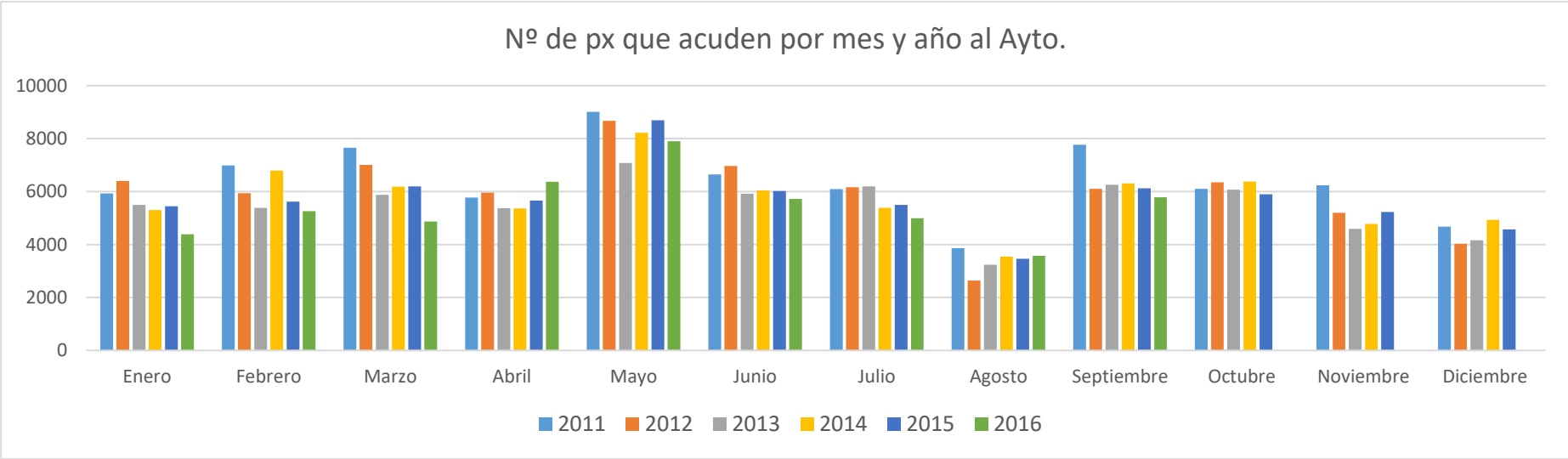


Ilustración 6 Número de usuarios que acuden por mes y año a las oficinas de atención al cliente del Ayto. de San Cugat III

- *Número de usuarios que acuden cada día de la semana y cada año*

	2011	2012	2013	2014	2015	2016	Total
Lunes	17021	16380	14218	14256	14580	9911	86366
Martes	15938	14633	14411	14647	14036	10006	83671
Miércoles	15570	14772	13317	14777	15388	11831	85655
Jueves	14773	13895	12936	13602	13305	9283	77794
Viernes	12952	11300	10563	11636	10806	7601	64858
Sábado	496	466	196	303	291	227	1979
	76750	71446	65641	69221	68406	48859	400323

Tabla 10 Número de usuarios que acuden por día de semana y año a las oficinas de atención al cliente del Ayto. de San Cugat I

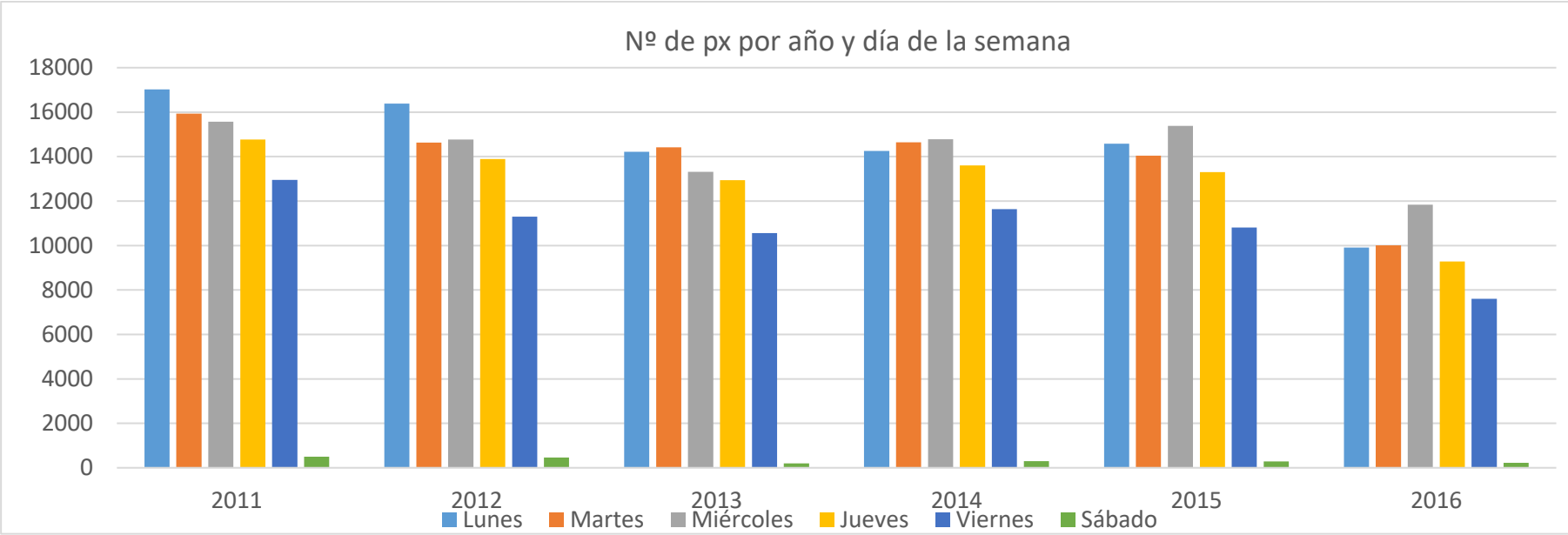


Ilustración 7 Número de usuarios que acuden por día de semana y año a las oficinas de atención al cliente del Ayto. de San Cugat II

- *Número de usuarios que acuden cada día del mes y cada año*

Día del mes/Año	2011	2012	2013	2014	2015	2016	Total general
1	2432	1555	1199	1625	1662	1694	**10167**
2	2477	2656	2044	2145	2437	1527	**13286**
3	2221	2406	1854	2541	1745	1035	**11802**
4	2667	2240	2791	2424	2272	1997	**14391**
5	2445	2611	2255	2382	2288	1767	**13748**
6	1992	2105	1955	2097	1861	1633	**11643**
7	3332	2336	2132	2333	2244	1663	**14040**
8	2475	2082	2433	2001	1987	1938	**12916**

9	2593	2273	2128	2045	2988	1499	**13526**
10	2620	2641	2184	3001	2505	1233	**14184**
11	2777	2187	2502	2146	2040	1731	**13383**
12	2550	2612	2311	2547	2015	1870	**13905**
13	2609	2849	2095	2299	2461	1880	**14193**
14	3416	2020	2148	2248	2498	1385	**13715**
15	2514	2047	2226	2183	2456	1697	**13123**
16	2552	2480	2366	2522	2869	1237	**14026**
17	2537	2705	2317	2944	2387	1165	**14055**
18	2594	2591	2589	2256	2216	1797	**14043**
19	2445	2512	2278	2202	2245	1696	**13378**
20	2419	2682	1697	2366	2517	1698	**13379**
21	3094	2058	2328	1961	2251	1519	**13211**
22	2404	1874	2244	2169	2231	1793	**12715**
23	2255	2370	2127	2026	2644	1561	**12983**
24	2138	2423	1885	2419	2028	1119	**12012**
25	2109	2028	2344	2102	2104	1719	**12406**
26	2012	2292	1992	1964	2182	1949	**12391**
27	2230	2759	2108	2376	2370	1868	**13711**
28	3135	2524	2031	2307	2053	1283	**13333**
29	2065	1681	1808	2319	1844	1489	**11206**
30	2048	2364	2232	2090	2227	1626	**12587**
31	1593	1483	1038	1181	779	791	**6865**
Total general	**76750**	**71446**	**65641**	**69221**	**68406**	**48859**	**400323**

Tabla 11 Número de usuarios que acuden por día del mes y año I

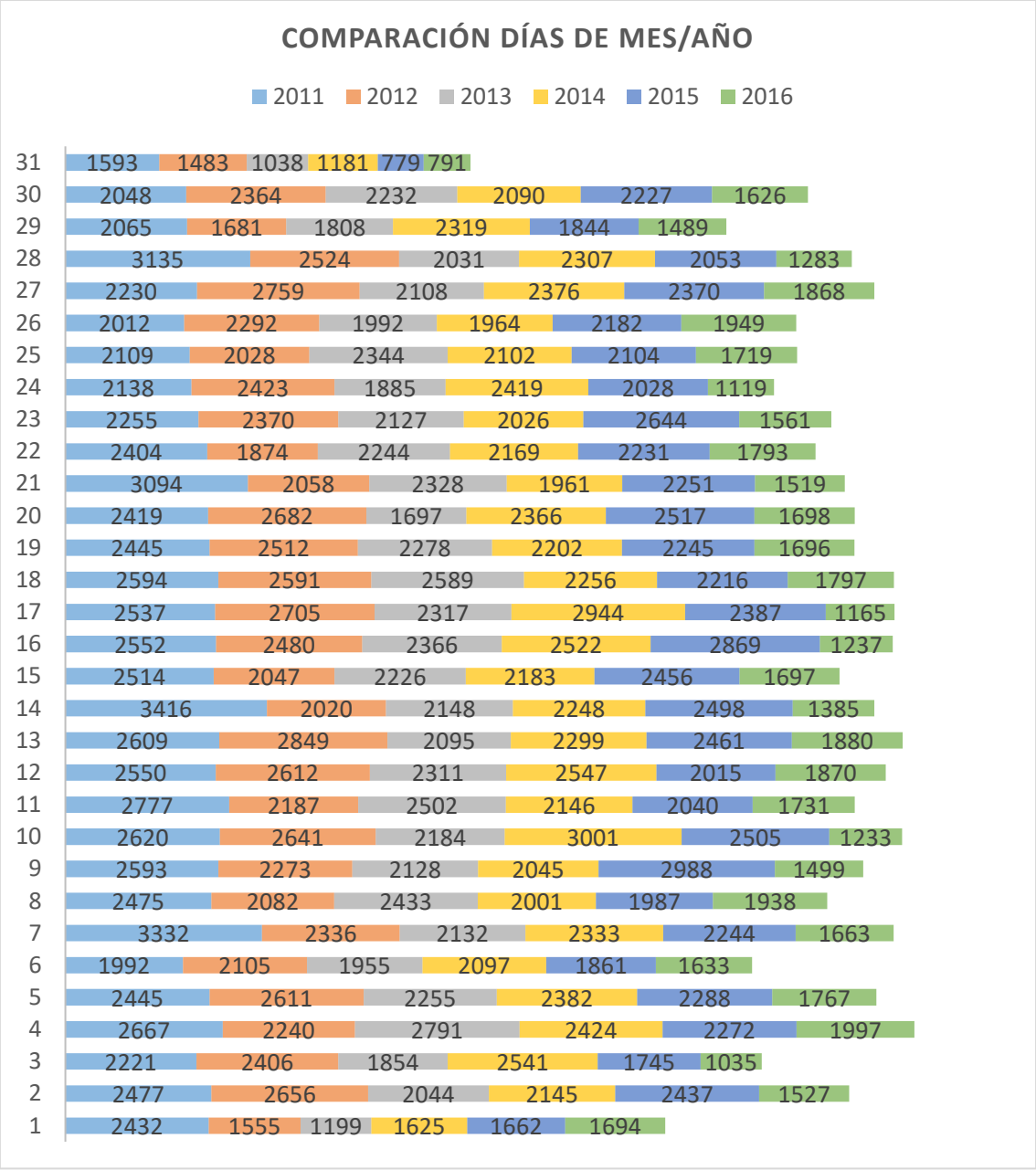


Ilustración 8 Número de usuarios que acuden por día del mes y año (diferenciando por colores) II

Tiempo total empleado en atenciones al cliente por cada gestor

Se ha calculado el tiempo total que dedica cada empleado a atender a las diferentes personas que acuden al Ayuntamiento de Sant Cugat. En la siguiente tabla se muestran de mayor a menor los códigos de gestores del ayuntamiento y el tiempo dedicado a atender clientes en el periodo comprendido entre el 1 de enero de 2011 y diciembre de 2016.

Gestor	Tiempo servicio total (Decimal)
24	178,0996991
4	162,2702546
2	157,3834144

Gestor	Tiempo servicio total (días y horas)	
24	178	2:23:34
4	162	6:29:10
2	157	9:12:07

15	152,4553125		15	152	10:55:39
9	152,0660069		9	152	1:35:03
18	151,9821412		18	151	23:34:17
1	151,6121528		1	151	14:41:30
7	149,0404745		7	149	0:58:17
20	141,4880556		20	141	11:42:48
11	140,4982523		11	140	11:57:29
17	140,1758333		17	140	4:13:12
25	83,35547454		25	83	8:31:53
14	73,52342593		14	73	12:33:44
30	66,93246528		30	66	22:22:45
5	58,49875		5	58	11:58:12
28	44,45775463		28	44	10:59:10
22	42,16071759		22	42	3:51:26
41	40,85739583		41	40	20:34:39
13	34,68678241		13	34	16:28:58
16	23,29746528		16	23	7:08:21
3	21,40394676		3	21	9:41:41
8	19,12258102		8	19	2:56:31
46	14,55601852		46	14	13:20:40
12	13,87386574		12	13	20:58:22
23	13,71893519		23	13	17:15:16
10	11,00546296		10	11	0:07:52
40	10,45768519		40	10	10:59:04
19	7,941226852		19	7	22:35:22
48	7,222673611		48	7	5:20:39
45	5,435497685		45	5	10:27:07
21	5,070231481		21	5	1:41:08
6	3,661168981		6	3	15:52:05
34	3,120393519		34	3	2:53:22
39	2,558229167		39	2	13:23:51
38	2,502673611		38	2	12:03:51
44	1,437222222		44	1	10:29:36
32	1,332858796		32	1	7:59:19
37	1,023703704		37	1	0:34:08
43	0,830717593		43	0	19:56:14
42	0,541921296		42	0	13:00:22
29	0,371539352		29	0	8:55:01
33	0,328333333		33	0	7:52:48
50	0,083113426		50	0	1:59:41
47	0,071516204		47	0	1:42:59
49	0,044965278		49	0	1:04:45
26	0,005648148		26	0	0:08:08
27	0,004664352		27	0	0:06:43
31	0		31	0	0:00:00
35	0		35	0	0:00:00
36	0		36	0	0:00:00

Tabla 12 Tiempo total empleado en atenciones al cliente por cada gestor

Anomalías encontradas en el estudio de los datos:

En algunos casos se encontró que existían tiempos de espera de 0 segundos (outliers) → Estos casos se han omitido en el estudio realizado ya que, aunque una persona sea atendida al llegar a la oficina de atención al ciudadano, siempre va a transcurrir al menos un segundo. Por lo que esas ocurrencias han sido señalizadas y descartadas ya que se ha creído apropiado considerarlo como errores en la medición.

Así mismo se han encontrado tiempos de espera excesivamente altos (outliers), pero estos no han sido descartados ni sacados del estudio ya que no se tenían pruebas o indicios contundentes para demostrar que estos fueran fallos del sistema de medición.

Adicionalmente existen tres códigos de gestor (31, 35, 36) que no han atendido ninguna atención al cliente.

Identificación de los eventos críticos que originan retrasos o anomalías

Inicialmente, haciendo un pequeño análisis de la situación actual y preguntando a distintos usuarios de oficinas públicas de atención al cliente se ha encontrado que los eventos críticos que suelen ocasionar retrasos son:

- Mes en el que acuden
- Tiempo de servicio (que el usuario sea mejor o peor comunicador influye en el tiempo de servicio ya que el gestor puede necesitar más o menos tiempo en atenderlo)
- Número de gestores que atienden a la vez en un momento concreto
- Que existan campañas o eventos ajenos al ayuntamiento que causen que más gente necesite acudir a ellos. Por ejemplo: campañas de vacunación, recepción de nuevos estudiantes, apertura nuevo colegio, nuevas ayudas a familias, etc.

3. MARCO TEÓRICO: ANÁLISIS DE DATOS Y TEORÍA DE COLAS

Aplicaciones de análisis de datos y de teorías de colas

En este capítulo se tratará de explicar algunas aplicaciones de análisis de datos relacionadas con la teoría de colas.

Una vez vista esta relación, se verá como vincular la teoría de colas a la gestión de las organizaciones, en este TFM la organización a la que se refiere es el Ayuntamiento de San Cugat, en primer término, se explicará la relevancia del tiempo de espera en la calidad del servicio. A continuación, se identifica un conjunto de indicadores de desempeño, relacionados con la eficacia, con la eficiencia y con el diseño del sistema.

La teoría de colas es la rama de la investigación de operaciones que estudia el comportamiento de los sistemas de atención, en que los clientes eventualmente esperan por el servicio. (Marcos Singer, 2008)

Hoy en día los costes están principalmente determinados por el personal empleado para atender a los servicios por los cuales los clientes están esperando, y que para algunos servicios requiere un alto grado de especialización técnica y por ende su coste es significativo.

Los modelos de teoría de colas ayudan en la toma de decisiones del propio Ayuntamiento de San Cugat al identificar y relacionar los indicadores de desempeño de interés del administrador (por ejemplo, la capacidad instalada) y los de interés de sus clientes (por ejemplo, el tiempo de espera). Dichos modelos también ayudan a mejorar la calidad del servicio, estimando e informando al cliente cuánto tiempo debe esperar hasta ser atendido.

La aplicabilidad de la teoría de colas es muy amplia en la administración de las organizaciones, pues el dilema entre la eficacia (dar un buen servicio) y la eficiencia (hacerlo con pocos recursos) es universal. (Marcos Singer, 2008)

En el capítulo 4 se abordará como vincular directamente la teoría de colas a la gestión de Ayuntamiento de San Cugat.

Procesos estocásticos:

La teoría de los procesos estocásticos se centra en el estudio y modelización de sistemas que evolucionan a lo largo del tiempo, o del espacio, de acuerdo a unas leyes no determinísticas, esto es, de carácter aleatorio. La forma habitual de describir la evolución del sistema es mediante sucesiones o colecciones de variables aleatorias. De esta manera, se puede estudiar cómo evoluciona una variable aleatoria a lo largo del tiempo. (Marín)

Por ejemplo, el número de personas que espera ante una ventanilla de un banco en un instante t de tiempo; el precio de las acciones de una empresa a lo largo de un año; el número de parados en el sector de Hostelería a lo largo de un año. La primera idea básica es identificar un proceso estocástico con una sucesión de variable aleatoria $\{X_n, n \in N\}$ donde el subíndice indica el instante de tiempo (o espacio) correspondiente. Esta idea inicial se puede generalizar fácilmente, permitiendo que los instantes de tiempo en los que se definen las variables

aleatorias sean continuos. Así, se podrá hablar de una colección o familia de v.a. (variable aleatoria) $\{X_t, t \in R\}$, que da una idea más exacta de lo que es un proceso estocástico.

Se tenía que una v.a. X(s) es una función que va desde un espacio muestral S a la recta real, de manera que a cada punto s ∈ S del espacio muestral se le puede asociar un número de la recta real. De este modo, la probabilidad de cada suceso de S se puede trasladar a la probabilidad de que un valor de X (v.a.) caiga en un cierto intervalo o conjunto de números reales. Si a todo esto se le añade una dimensión temporal, se obtiene un proceso estocástico. La definición formal es la siguiente (Marín):

Un proceso estocástico es una colección o familia de variables aleatorias $\{X_t, t \in R\}$, ordenadas según el subíndice t que en general se suele identificar con el tiempo. (Ruiz)

Es decir, dado el espacio de probabilidad (Ω, a, P) de modo que para todo t ∈ T ⊂ R fijo

X_t: (Ω, a, P) −→ (R, B)

w → $\{X_t(w) \in R\}$

esto es, X_t es una variable aleatoria y ∀w ∈ Ω fijo, X•(w) es una función del tiempo. (Marín)

Los procesos estocásticos pueden clasificarse según:

- La estructura del conjunto paramétrico T y del conjunto de estados E

Los procesos estocásticos se pueden clasificar en cuatro tipos, dependiendo de si T es un conjunto numerable o continuo, y de si E es otro conjunto numerable o continuo. Así:

E/T	**Discreto**	**Continuo**
Discreto	Proceso de estado discreto y tiempo discreto (**Cadena**) (Unidades producidas mensualmente de un producto)	Proceso de estado discreto y tiempo continuo (Proc. Saltos Puros o **Proceso puntual**) (Unidades producidas hasta el instante t)
Continuo	Sucesión De Variables aleatorias: Proceso de estado continuo y tiempo discreto (Toneladas de producción diaria de un producto)	Proceso de estado continuo y tiempo continuo (**Proceso Continuo**) (Velocidad de un vehículo en el instante t)

Tabla 13 Procesos estocásticos: discretos y continuos

Una cadena es un proceso estocástico en el cual el tiempo se mueve en forma discreta y la variable aleatoria sólo toma valores discretos en el espacio de estados. Un proceso de saltos Puros es un proceso estocástico en el cual los cambios de estados ocurren en forma aislada y aleatoria pero la variable aleatoria sólo toma valores discretos en el espacio de estados. En un Proceso Continuo los cambios de estado se producen en cualquier instante y hacia cualquier estado dentro de un espacio continuo de estados.

Como un ejemplo de una Cadena, considere una máquina dentro de una fábrica. Los posibles estados para la máquina son que esté operando o que esté fuera de funcionamiento y la

verificación de esta característica se realizará al principio de cada día de trabajo. Si hacemos corresponder el estado 'fuera de funcionamiento' con el valor 0 y el `estado en operación' con el valor 1, la siguiente figura muestra una posible secuencia de cambios de estado a través del tiempo para esa máquina. (Ruiz)

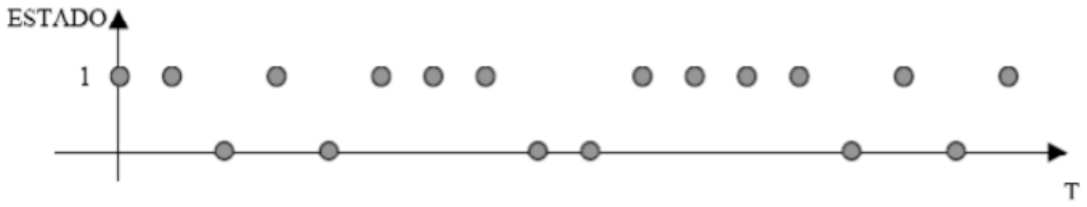


Ilustración 9 Posible secuencia de cambios de estado

Para el caso de los Procesos de Saltos Puros se puede considerar como un ejemplo una señal telegráfica. Sólo hay dos posibles estados (por ejemplo 1 y -1) pero la oportunidad del cambio de estado se da en cualquier instante en el tiempo, es decir, el instante del cambio de estado es aleatorio. La siguiente figura muestra una señal telegráfica. (Ruiz)

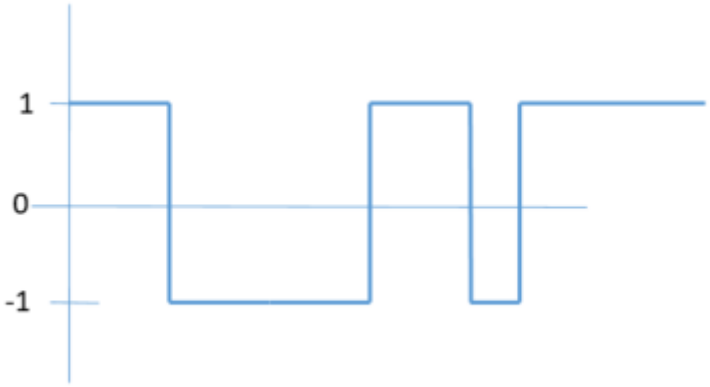


Ilustración 10 Proceso de saltos puros. Ejemplo: Una señal telegráfica aleatoria

Como un ejemplo de un Proceso Continuo, se puede mencionar la señal de voz vista en la pantalla de un osciloscopio. Esta señal acústica es transformada en una señal eléctrica analógica que puede tomar cualquier valor en un intervalo continuo de estados. La figura siguiente muestra una señal de voz la cual está modulada en amplitud. (Ruiz)

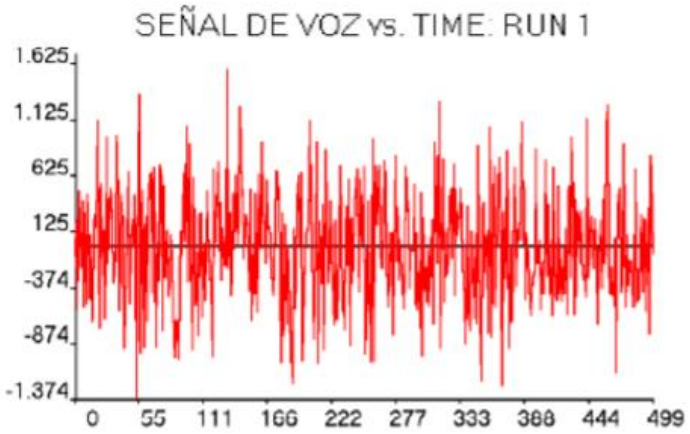


Ilustración 11 Proceso continuo. Ejemplo: Una señal de voz vista en pantalla de un osciloscopio

- Las características probabilísticas de las v.a.

En la vida real se producen distintas relaciones entre las variables aleatorias que constituyen un proceso estocástico. Las propiedades probabilísticas las v.a. son importantes a la hora de identificar y clasificar un proceso estocástico. Se pueden clasificar los procesos en:

- ❖ Procesos estacionarios.
- ❖ Procesos Markovianos.
- ❖ Procesos de incrementos independientes.

El interés de los procesos estocásticos es describir el comportamiento de un sistema en operación durante algunos periodos. Para saber su comportamiento, existen muchas aplicaciones y es usado también como sustento para muchas teorías probabilísticas una de ellas es la teoría de colas. (Ruiz)

Teoría de colas

La teoría de colas estudia el comportamiento de los sistemas de atención sujetos a diferentes condiciones de funcionamiento, en que los clientes a veces deben esperar por el servicio. Su aplicabilidad es muy amplia, pues cuantifica el dilema de muchas empresas e instituciones entre la eficacia (dar un buen servicio) y la eficiencia (mantener bajos los costes).

Podemos describir un sistema de colas como un “conjunto de clientes” que llegan a un sistema buscando un servicio, este servicio puede ser inmediato o no, en cuyo caso deben de esperar, y abandonarán el sistema una vez hayan sido atendidos. En algunos casos se puede admitir que los clientes abandonan el sistema si se cansan de esperar.

Con el término “cliente” se hace referencia a un sentido general y no implica que sea un ser humano, puede significar piezas esperando su turno para ser procesadas o una lista de trabajo esperando para imprimir en una impresora en red.

En nuestro caso de estudio “Oficina de Atención al cliente del Ayuntamiento de San Cugat”, se hará referencia a un cliente, como cada una de las personas que se dirigen al ayuntamiento de San Cugat para realizar algún tipo de gestión o trámite y pasarán por el sistema de colas que aquí se estudia.

En este caso, se entiende que los clientes (usuarios) llegan a un sistema con múltiples colas y varios servidores (gestores que están atendiendo cada uno de los trámites), más adelante se definirán cada uno de los elementos de la teoría de colas.

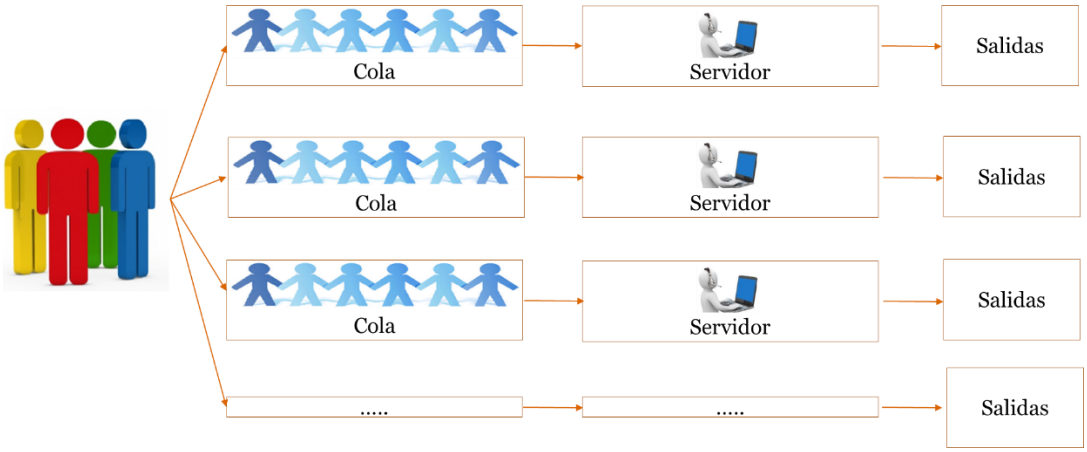


Ilustración 12 Sistema de colas: Varias colas y múltiples servidores. Fuente: Elaboración propia

La teoría de colas fue originariamente un trabajo práctico. La primera aplicación de la que se tiene noticia es del matemático danés Erlang sobre conversaciones telefónicas en 1909, para el cálculo de tamaño de centralitas.

Características de los sistemas de colas:

Seis son las características básicas que se deben utilizar para describir adecuadamente un sistema de colas:

1. Patrón de llegada de los clientes
2. Patrón de servicio de los servidores
3. Disciplina de cola
4. Capacidad del sistema
5. Número de canales de servicio
6. Número de etapas de servicio

Ahora se definirán cada una de las seis características que definen los sistemas de colas (García Sabater, 2016):

1. Patrón de llegadas de los clientes:

En situaciones de cola habituales, la llegada es estocástica, es decir la llegada depende de una cierta variable aleatoria, en este caso es necesario conocer la distribución probabilística entre dos llegadas de cliente sucesivas. Además habría que tener en cuenta si los clientes llegan independiente o simultáneamente.

También es posible que los clientes sean “impacientes”. Es decir, que lleguen a la cola y si es demasiado larga se vayan, o que tras esperar mucho rato en la cola decidan abandonar. Para la realización de este TFM se ha supuesto que ningún cliente abandona nunca la cola.

Por último es posible que el patrón de llegada varíe con el tiempo. Si se mantiene constante le llamamos estacionario, si por ejemplo varía con las horas del día es no-estacionario. (García Sabater, 2016)

2. Patrón de servicio de los servidores

Los servidores pueden tener un tiempo de servicio variable, en cuyo caso hay que asociarle, para definirlo, una función de probabilidad. También pueden atender en lotes o de modo individual.

El tiempo de servicio también puede variar con el número de clientes en la cola, trabajando más rápido o más lento, y en este caso se llama patrones de servicio dependientes. Al igual que el patrón de llegadas el patrón de servicio puede ser no-estacionario, variando con el tiempo transcurrido. (García Sabater, 2016)

3. Disciplina de cola

La disciplina de cola es la manera en que los clientes se ordenan en el momento de ser servidos de entre los de la cola. Cuando se piensa en colas se admite que la disciplina de cola normal es FIFO (atender primero a quien llegó primero).

Sin embargo en muchas colas es habitual el uso de la disciplina LIFO (atender primero al último). También es posible encontrar reglas de secuencia con prioridades, como por ejemplo secuenciar primero las tareas con menor duración o según tipos de clientes.

En este TFM se ha partido de la situación de colas que existe en la atención al cliente del Ayuntamiento de San Cugat: FIFO. Es decir, siempre atienden primero a quién primero llega. (García Sabater, 2016)

4. Capacidad del sistema

En algunos sistemas existe una limitación respecto al número de clientes que pueden esperar en la cola. A estos casos se les denomina situaciones de cola finitas. Esta limitación puede ser considerada como una simplificación en la modelización de la impaciencia de los clientes.

En el ayuntamiento de San Cugat, el número de clientes que pueden esperar en las colas para ser atendidos es infinito, lo cual implica: colas infinitas para la realización de este TFM. (García Sabater, 2016)

5. Número de canales de servicio

Es evidente que es preferible utilizar sistemas multi-servidor con una única línea de espera para todos que con una cola por servidor. Por tanto, cuando se habla de canales de servicio paralelos, se habla generalmente de una cola que alimenta a varios servidores mientras que el caso de colas independientes se asemeja a múltiples sistemas con sólo un servidor. (García Sabater, 2016)

En la figura siguiente el primer sistema tiene una sola cola de espera, mientras que el segundo tiene una sola cola para cada canal.

Ilustración 13 Sistemas de cola multicanal (Varias colas para llegar a ser atenido)

En el caso de estudio que afecta a esta memoria, cuando un cliente llega al ayuntamiento procederá a recoger un ticket según el trámite que quiera realizar. Con ese ticket esperará su turno esperando la cola correspondiente al trámite elegido. Por ese motivo hace referencia a la segunda imagen de la figura anterior, ya que hay diferentes colas que llegan a los distintos servidores.

De cualquiera de las maneras, se asumen que los mecanismos de servicio operan de manera independiente.

6. Etapas de servicio

Un sistema de colas puede ser uni-etapa o multi-etapa. En los sistemas multi-etapa el cliente puede pasar por un número de etapas mayor que uno. Un supermercado es un sistema uni-etapa, salvo que haya diferentes servicios (pago, recogida de productos) y cada uno de estos servicios sea desarrollado por un servidor diferente. (García Sabater, 2016)

En algunos sistemas multi-etapa se puede admitir la vuelta atrás o "reciclado", esto es habitual en sistemas productivos como controles de calidad y reprocesos. Un sistema multi-etapa se ilustra en la figura siguiente:

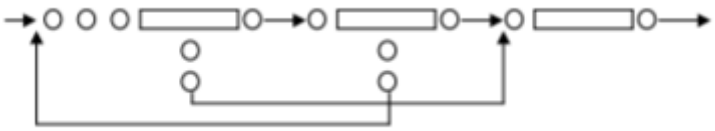


Ilustración 14 Sistema multi-etapa con retroalimentación

El sistema de colas de nuestro caso de estudio, se compone de una sola etapa, que sería el hecho de ser atendido por alguno de los oficinistas del Ayuntamiento de San Cugat.

Es decir, el cliente recoge un ticket con su número identificativo, a partir de este momento el cliente esperará hasta que llegue su turno y sea atendido. Una vez finalice la atención, abandonará el sistema.

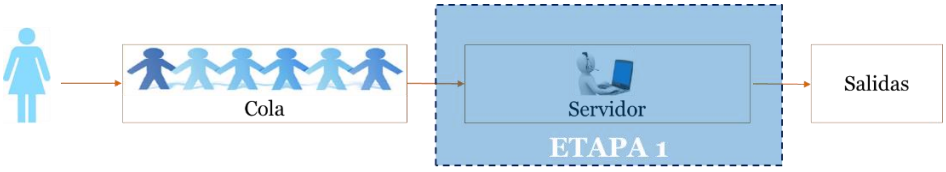


Ilustración 15 Sistema de colas del Ayto. San Cugat. Uni-etapa. Fuente: Elaboración propia

A modo de resumen de las características anteriores, cabe destacar que éstas pueden describir cualquier proceso. Es evidente que se pueden encontrar sistemas de colas muy distintos con sus problemas concretos, es por ello, que antes de comenzar cualquier análisis se debe describir perfectamente el sistema según las características anteriores.

Como conclusión: para poder modelar adecuadamente un sistema de colas, éste debe haber sido definido apropiadamente y tener controladas todas sus variables.

Redes de colas

En este apartado se hará una breve introducción a las redes de colas. Según (García Sabater, 2016), las Redes de Colas se pueden describir como un grupo de nodos (sean k), en el que cada nodo representa una instalación de servicio.

Dicha instalación puede constar de ci servidores (i=,1...k) En el caso más general los clientes pueden entrar en cualquier nodo y, después de moverse por la red, pueden salir en cualquier nodo.

Dentro de las Redes de Colas, se pondrá especial interés en las denominadas "Redes de Jakcson". Estas tienen las siguientes características:

- Las llegadas desde el exterior al nodo i siguen un proceso de Poisson de media γ_i.
- Los tiempos de servicio en cada nodo y son independientes y siguen una distribución negativa exponencial con parámetro μi, que podría ser dependiente del estado.
- La probabilidad de que un cliente que haya completado su servicio en el nodo i vaya al nodo j es rij con i=1, 2, k, j=0, 1, k.
- ri,0 indica la probabilidad de que un cliente abandone desde el nodo i

Si añadimos las características γi y ri, 0 para todo i estamos en el caso de las "Redes de Jackson cerradas" Si no se da el caso anterior el problema se denomina de "Redes de Jackson abiertas". Se consideran tres tipos de redes de Jackson (García Sabater, 2016):

I. Las redes de Jackson "en serie"
II. Las redes de Jackson "en general"
III. Las redes de Jackson "cerradas"

En cualquier caso, las primeras son una variante reducida de las segundas.

Las redes de Jackson "en serie":

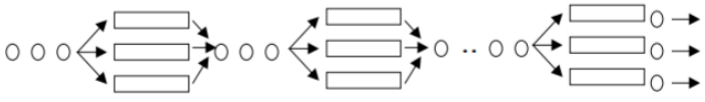


Ilustración 16 Redes de cola. Sistema en serie.

Los clientes entran en el nodo 1 y salen en el nodo k, después de pasar por cada uno de los nodos.

Se puede demostrar que la salida de los clientes de un sistema M/M/c/ ∞ tienen una distribución idéntica a la de la entrada, es decir Poisson con media λ. Por tanto una serie se

compone de k M/M/ci/∞ colas independientes, siempre que la entrada sea Poisson, el servicio sea exponencial y no haya restricciones de capacidad.

Las redes de Jackson "en general":

Se considera que son redes de Jackson abiertas cuando:

- ✓ La llegada externa a cualquier nodo es Poisson γ_i
- ✓ Todos los servidores de cada etapa tienen un servicio exponencial de media μi
- ✓ De cada etapa i un cliente se mueve a otra etapa con probabilidad r_{ij}, y al exterior con probabilidad ri,0

Ilustración 17 Redes de cola. Red de Jackson abierta

Las redes de Jackson "cerradas":

Es un sistema con N clientes continuamente viajando a través de la red.

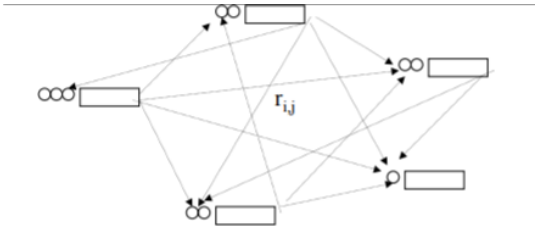


Ilustración 18 Redes de cola. Red de Jackson cerradas

Simulación

Hoy en día existe una problemática con respecto algún caso que no puede resolverse mediante métodos analíticos. Esto puede ser debido por ejemplo a la existencia de patrones no normalizados de entrada y de servicio, una gran complejidad del sistema a modelar o la naturaleza de la disciplina de cola. En otras ocasiones los resultados analíticos son para un estado estacionario que nunca se alcanza, porque el sistema se interrumpe antes de abandonar el estado transitorio.

Para todos estos casos, la mejor solución para encontrar un buen resultado sería analizar las colas mediante simulación.

Resolver un problema mediante simulación se asemeja a encontrarlo por experimentación, por este motivo si existe algún modelo analítico que pueda usarse para resolverlo siempre será más recomendable ya que el resultado será más fiable. Por tanto para la simulación, hay que utilizar todas las herramientas asociadas al diseño y análisis de experimentos: Recogida y Análisis de

Datos, realización de la experimentación, análisis y consistencia de resultados, etc. (García Sabater, 2016)

Para resolver un problema mediante simulación deben seguirse los pasos descritos a continuación:

1. Identificar los elementos de un modelo de simulación
2. Modelizar las entradas
3. Analizar los resultados
4. Validar el modelo (cuando sea posible)

Seguidamente se procede a detallar cada uno de estos pasos:

I) Elementos de un modelo de simulación:

Hay cuatro elementos a tener en cuenta al abordar un modelo de simulación de teoría de colas, supuesto diseñado el modelo "físico" (García Sabater, 2016):

- Selección de los datos de entrada
- Simulación.
- Análisis de los resultados
- Validación del modelo.

II) Modelización de las entradas:

La Modelización de las Entradas es un requerimiento no sólo de la simulación, sino de cualquier tipo de análisis probabilístico y numérico.

Los dos mayores problemas en la modelización de los datos de entrada son la selección de la familia de distribuciones estadísticas y una vez estimada la familia estimar los parámetros que definen la función de las diferentes entradas.

El primero de los dos problemas es evidentemente el más complicado mientras que el segundo sólo es abordable una vez se ha resuelto la selección de la familia de distribuciones estadísticas. (García Sabater, 2016)

III) Análisis de resultados:

Alcanzar conclusiones válidas a partir de los resultados requiere un gran y cuidadoso esfuerzo. Cuando se simulan sistemas estocásticos, no es posible extraer conclusiones a partir de una única simulación que por naturaleza es estadística.

Por tanto para obtener conclusiones es necesario diseñar y ejecutar experimentos de una manera lógica y comprensiva.

IV) Validación del modelo

La validación de los modelos es probablemente el paso más importante, y probablemente también el paso más obviado por aquellos que modelizan. (García Sabater, 2016)

Antes de iniciar el proceso de realizar un modelo de simulación es necesario que el modelizador se familiarice con el sistema que tiene que estudiar. Para ello es necesario involucrar a todos los niveles de personal implicados en el proceso que va a ser simulado. En ese caso uno de los problemas que aparece es el exceso de detalles en el modelo que lo convierten en improductivo.

El primer y fundamental paso en la validación es verificar que el programa hace lo que está previsto que haga. Otro paso es definir el grado de credibilidad, es decir hasta qué punto los que van a usar el modelo consideran que el mismo tiene una utilidad y representa la realidad en la media que nos interesa. Para ello es necesario que los objetivos del estudio, las medidas de rendimiento y el nivel de detalle debe pactarse y mantenerse en el nivel más simple posible.

Cuando sea posible, los resultados de las simulaciones se deben comprobar con la realidad. Si esta no estuviera disponible habría que intentar reproducir modelos teóricos con soluciones conocidas mediante métodos analíticos.

Optimización

Existen varios tipos de optimización usados para diferentes situaciones que por su complejidad se constituyen en un problema. Algunos de esos métodos son (García Sabater, 2016):

- ***Programación lineal***: Procedimiento de resolución de problemas mediante formulación a través de ecuaciones lineales, optimizando la función objetivo teniendo que está restringida por ecuaciones lineales.

- ***Programación entera:*** Al igual que la lineal utiliza un procedimiento o algoritmo para resolver problemas. En estos modelos, algunas o todas sus variables deben ser valores enteros.

- ***Programación dinámica***: Es un enfoque general para la solución de problemas en los que es necesario tomar decisiones en etapas sucesivas. Las decisiones tomadas en una etapa condicionan la evolución futura del sistema y afectan a las situaciones en las que el sistema se encontrará en el futuro (denominadas estados), y a las decisiones que se plantearán en ese momento.

- ***Meta-Heurísticas:*** Es un método para resolver un tipo de problema general, usando los parámetros dados por el usuario sobre unos procedimientos genéricos y abstractos de una manera que se espera eficiente. Las meta-heurísticas generalmente se aplican a problemas que no tienen un algoritmo o heurística específica que dé una solución satisfactoria; o bien cuando no es posible implementar ese método óptimo.

Modelos de predicción

Predicción de eventos futuros a partir de datos históricos

En el análisis predictivo se hace uso de los datos junto con técnicas analíticas, estadísticas y de aprendizaje automático a fin de crear un modelo predictivo para predecir eventos futuros.

El término análisis predictivo no describe ninguna técnica estadística o de aprendizaje automático concreta, sino más bien la aplicación de una técnica para crear una predicción cuantitativa sobre el futuro. Con frecuencia, se utilizan técnicas de aprendizaje automático supervisado para predecir un valor futuro (por ejemplo, la temperatura mañana será de 30º) o para calcular una probabilidad (por ejemplo, la probabilidad de que llueva mañana es del 75 %). (MathWorks)

El análisis predictivo se suele abordar en el contexto de Big Data, ya que las empresas aplican algoritmos a fin de obtener una visión detallada a partir de grandes conjuntos de datos. Las fuentes de datos utilizadas para crear modelos predictivos suelen incluir bases de datos, archivos de registro de equipamiento, imágenes, vídeos, audio y datos de sensores. A continuación, estos modelos predictivos se pueden desplegar para su uso en producción en un entorno de TI o en un sistema embebido.

Las técnicas empleadas en los análisis predictivos a menudo incluyen algunos de los elementos siguientes:

1. Regresión lineal

Descripción de relaciones matemáticas y realización de predicciones a partir de datos experimentales

La regresión lineal es una técnica de modelización estadística que se emplea para describir una variable de respuesta continua a modo de función de una o varias variables predictivas. Puede ayudarle a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales, financieros y biológicos. (MathWorks)

Las técnicas de regresión lineal permiten crear un modelo lineal. Este modelo describe la relación entre una variable dependiente y_y (también conocida como la respuesta) a modo de función de una o varias variables independientes x_i (denominadas predictores). La ecuación general correspondiente a un modelo de regresión lineal es:

$$y = \beta_0 + \sum \beta_i X_i + \in_i \; y = \beta_0 + \sum \beta_i X_i + \in_i$$

Donde β representa las estimaciones de parámetros lineales que se deben calcular y ∈∈ representa los términos de error.

Existen varios tipos de modelos de regresión lineal (MathWorks):

- **Simple:** modelo con un único predictor
- **Múltiple:** modelo con varios predictores
- **Multivariante:** modelo para varias variables de respuesta

La regresión lineal simple se suele realizar en MATLAB, R, Phyton, o cualquier software que permita el análisis de datos y que se detallan en el capítulo 4. Se permite en ellos la regresión múltiple, por pasos, robusta y multivariante:

- Generar predicciones
- Comparar ajustes de modelos lineales
- Representar gráficamente los valores residuales
- Evaluar la bondad del ajuste
- Detectar valores atípicos

Dada la relevancia de los modelos lineales para el caso de estudio analizado en este TFM, se dedicará un capítulo completo (Capitulo 5: Modelos lineales: Regresión lineal, ANOVA y ANCOVA) a la explicación de dichos modelos

Ya que sobre estos modelos se ha construido todo el análisis estadístico e R que se detallará al completo en el capítulo 6.

2. Regresión no lineal:

La regresión no lineal hace referencia a un método para encontrar un modelo no lineal para la relación entre la variable dependiente y un conjunto de variables independientes. A diferencia de la regresión lineal tradicional, que está restringida a la estimación de modelos lineales, la regresión no lineal puede estimar modelos con relaciones arbitrarias entre las variables independientes y las dependientes. Esto se lleva a cabo usando algoritmos de estimación iterativos. (IBM Knowledge Center, 2018)

Consideraciones sobre los datos en una regresión no lineal:

Tanto las variables dependientes e independientes deben ser cuantitativas. Las variables categóricas han de recodificarse como variables binarias (dummy) o como otro de los tipos de variables de contraste.

Los modelos más comunes de regresión no lineales son los siguientes:

- Modelos Potenciales
- Modelos Exponenciales
- Modelos Polinomiales

Para entender mejor este tipo de regresiones se va a realizar una comparación entre regresión no lineal y lineal (Soporte técnico de Minitab, 2017):

Similitudes

- Describen matemáticamente la relación entre una variable de respuesta y una o más variables predictoras.
- Pueden modelar una relación curva.
- Minimizan la suma de los cuadrados del error residual (SSE).
- Tienen los mismos supuestos que usted puede verificar utilizando las gráficas de residuos.

Diferencias:

La diferencia fundamental entre las regresiones lineal y no lineal, y la base para los nombres de los análisis, son las formas funcionales aceptables del modelo. Específicamente, la regresión lineal requiere parámetros lineales mientras que la no lineal no. Se suele utilizar la regresión no lineal en lugar de la regresión lineal cuando no se puede modelar adecuadamente la relación con parámetros lineales.

Una función de regresión lineal debe ser lineal en los parámetros, lo cual restringe la ecuación a una sola forma básica. Los parámetros son lineales cuando cada término del modelo es aditivo y contiene solo un parámetro que multiplica el término.

Sin embargo, una ecuación no lineal puede adoptar muchas formas diferentes. De hecho, el número de posibilidades es infinito. Adicionalmente la regresión no lineal utiliza un procedimiento diferente del que usa la regresión lineal para minimizar la suma de los cuadrados del error residual (SSE). (Soporte técnico de Minitab, 2017)

3. Tipos de análisis predictivos

Para llevar a cabo el análisis predictivo existen 3 modelos que pasaremos a describir (Fugu Software Factory, 2015):

Modelos predictivos:

Los modelos predictivos analizan los resultados anteriores para evaluar qué probabilidad tiene un cliente para mostrar un comportamiento específico en el futuro con el fin de mejorar la eficacia de marketing. Esta categoría también incluye modelos que buscan patrones discriminatorios de datos para responder a las preguntas sobre el comportamiento del cliente, tales como la detección de tipos de fraudes. Los modelos de predicción a menudo realizan cálculos en tiempo real, durante las operaciones, por ejemplo, para evaluar el riesgo o la oportunidad de un determinado cliente o transacción, a fin de orientar una decisión (Fugu Software Factory, 2015).

Modelos descriptivos:

Los modelos descriptivos describen las relaciones en los datos para poder clasificar a los clientes en grupos. A diferencia de modelos de predicción que se centran en predecir el comportamiento de un único cliente (como el riesgo de crédito), los modelos descriptivos identifican diferentes relaciones entre clientes o productos. Pero los modelos descriptivos no clasifican a los clientes según su probabilidad de tomar una acción en particular. Los modelos descriptivos se utilizan a menudo “offline” por ejemplo, clasificar a los clientes por las preferencias de los productos según la etapa de la vida. Las herramientas de modelado descriptivo pueden ser utilizadas para desarrollar modelos basados en agentes simulando una gran cantidad de agentes individuales pudiendo predecir acciones futuras (Fugu Software Factory, 2015).

Modelos de decisión:

Los modelos de decisión describen la relación entre todos los elementos de una decisión – los datos conocidos (incluidos los resultados de los modelos de predicción), la decisión y el plan de variables y valores que determinan la decisión – con el fin de predecir los resultados de las decisiones de muchas variables. Estos modelos pueden ser utilizados en optimización (Fugu Software Factory, 2015).

4. ESTUDIO DE LAS DIFERENTES HERRAMIENTAS DE ANÁLISIS DE DATOS

En este capítulo se ha realizado una comparativa de los distintos software de análisis de datos estadístico y a continuación se ha realizado la elección de la herramienta más adecuada para llevar a cabo el caso de estudio propuesto en este TFM y cuáles han sido los motivos de su elección:

Comparativa de diferentes softwares de análisis de datos:

En un principio se propusieron diferentes software para realizar este TFM:

- *Excel* (Estadística para todos, 2016):
    - En estadística descriptiva representa todos los tipos de gráficos y calcula la media, moda, mediana, recorrido, varianza y desviación típica.
    - En estadística bidimensional representa la nube de puntos y la recta de regresión. Calcula el centro de gravedad, las desviaciones típicas marginales, la covarianza, el coeficiente de correlación, la recta de regresión y buscar objetivos.
    - En la distribución binomial, calcula cualquier probabilidad, la media, varianza y desviación típica.
    - En la distribución normal, calcula cualquier probabilidad en la normal estándar N (0, 1) y en cualquier normal N (m, s) y genera la tabla N (0, 1)
    - En inferencia estadística calcula los intervalos de confianza, el tamaño de la muestra y se puede aplicar al contraste de hipótesis, tanto en el bilateral como en el unilateral.
    - En probabilidad simula todo tipo de lanzamientos.

- *R y RStudio*: **R** es un programa para realizar cálculos estadísticos. Este es un software libre altamente recomendable por lo robusto y potente de las rutinas que tiene implementadas. (Estadística para todos, 2016)

R es un software de análisis estadístico y gráfico, utilizado en áreas tan diversas como: investigación biomédica, telecomunicaciones, bioinformática, finanzas, entre otras características dispone de:

- almacenamiento y manipulación efectiva de datos,
- operadores para cálculo sobre variables indexadas (*Arrays*), en particular matrices,
- una amplia, coherente e integrada colección de herramientas para análisis de datos,
- posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora,
- un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R)
  (W. N. Venables, 2017)

RStudio es un entorno de desarrollo integrado (IDE) para R (lenguaje de programación). Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo.

RStudio está disponible para Windows, Mac y Linux o para navegadores conectados a *RStudio* Server o *RStudio* Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux).

RStudio *tiene las siguientes características:*

IDE construido exclusivo para R

- El resaltado de sintaxis, auto completado de código y sangría inteligente.
- Ejecutar código R directamente desde el editor de código fuente.
- Salto rápido a las funciones definidas.

Colaboración

- Documentación y soporte integrado.
- Administración sencilla de múltiples directorios de trabajo mediante proyectos.
- Navegación en espacios de trabajo y visor de datos.

Potente autoría y depuración.

- Depurador interactivo para diagnosticar y corregir los errores rápidamente.
- Herramientas de desarrollo extensas.
- Autoría con Sweave y R Markdown.

RStudio tiene la misión de proporcionar el entorno informático estadístico R. Permite un análisis y desarrollo para facilitar el análisis de los datos con R.

- *Orange*: Orange es una suite de software para minería de base de datos y aprendizaje automático basado en componentes que cuenta con un fácil y potente, rápido y versátil front-end de programación visual para el análisis exploratorio de datos y visualización, y librerías para Python y secuencias de comando. Contiene un completo juego de componentes para pre-procesamiento de datos, característica de puntuación y filtrado, modelado, evaluación del modelo, y técnicas de exploración. Está escrito en C++ y Python, y su interfaz gráfica de usuario se basa en la plataforma cruzada del framework Qt. (Jimmy W. Maco Elera, 2017)

- *Phyton*: Python es un lenguaje de programación, el cual en los últimos cinco años ha empezado a tener una popularidad que lo convierten en el lenguaje con más crecimiento, particularmente en el pasado año 2017. Hay varias razones para ello: por una parte, parece ser un lenguaje que se está convirtiendo en el favorito para enseñar a programar.

Python es desde hace tiempo un lenguaje de programación muy popular, usado por numerosas empresas, científicos, programadores ocasionales y profesionales (aplicaciones, servicios web y en la nube, sitios web) y generadores de scripts de aplicaciones. Python cuenta con magníficas propiedades (Microsoft: Developer Network, 2018):

- Fiable
- Suele resultar útil para la generación de scripts de programas rápidos, scripting de aplicaciones, aplicaciones de escritorio, servidores web, servicios web, informática científica
- Fácil de aprender a usar y con un diseño adecuado para promover una codificación correcta (numerosas universidades lo usan para cursos de introducción a la programación)
- Admite diversos estilos de programación: imperativa, funcional y orientada a objetos
- Código abierto gratuito que se ejecuta correctamente en los principales sistemas operativos
- Numerosas bibliotecas útiles, gratuitas y bien diseñadas
- Mucha documentación, ejemplos y ayuda disponibles en Internet

- *MiniTab*: **Minitab** es otro de los programas más usados en el mundo para análisis estadístico. Permite calcular la mayoría de metodologías estadísticas habituales, entre las que se cuentan: análisis exploratorio de datos, gráficos estadísticos, control de calidad, estadística no paramétrica, regresión y sus variantes, análisis multivariado de datos, etc. (Estadística para todos, 2016)

- *RapidMiner*: RapidMiner, antes llamado YALE (Sin embargo, otro ambiente de aprendizaje), es un ambiente de experimentos en aprendizaje automático y minería de datos que se utiliza para tareas de minería de datos tanto en investigación como en el mundo real. Permite a los experimentos componerse de un gran número de operadores anidables arbitrariamente, que se detallan en archivos XML y se hacen con la interfaz gráfica de usuario de RapidMiner. RapidMiner ofrece más de 500 operadores para todos los principales procedimientos de máquina de aprendizaje, y también combina esquemas de aprendizaje y evaluadores de atributos del entorno de aprendizaje Weka. Está disponible como una herramienta stand-alone para el análisis de datos y como motor para minería de datos que puede integrarse en tus propios productos. (Jimmy W. Maco Elera, 2017)

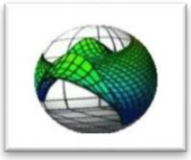
- *Weka*: Escrito en Java, Weka (Entorno Waikato para el Análisis del Conocimiento) es una conocida suite de software para máquinas de aprendizaje que soporta varias tareas típicas de minería de datos, especialmente pre procesamiento de datos, agrupamiento, clasificación, regresión, visualización y características de selección. Sus técnicas se basan en la hipótesis de que los datos están disponibles en un único archivo plano o relación, donde cada punto marcado es etiquetado por un número fijo de atributos.

WEKA proporciona acceso a bases de datos SQL utilizando conectividad de bases de datos Java y puede procesar el resultado devuelto como una consulta de base de datos. Su interfaz de usuario principal es el Explorer, pero la misma funcionalidad puede ser accedida desde la línea de comandos o a través de la interfaz de flujo de conocimientos basada en componentes. (Jimmy W. Maco Elera, 2017)

- *SPSS*: es la herramienta estadística más utilizada a nivel mundial en el entorno académico. Puede trabajar con bases de datos de gran tamaño. . Además, de permitir la recodificación de las variables y registros según las necesidades del usuario. El programa consiste en un módulo base y módulos anexos que se han ido actualizando constantemente con nuevos procedimientos estadísticos. (Estadística para todos, 2016)

- *S-PLUS:* S-Plus es otro de los programas estadísticos más usados a nivel mundial para el análisis de datos. Está disponible al público la versión 8. Incluye dentro de sus principales características: análisis multivariado de datos, análisis de sobrevivencia, escalamiento multidimensional, regresión no paramétrica.
  Entre los estadísticos de cálculo habituales incluye: pruebas de hipótesis y construcción de intervalos de confianza, análisis de varianza, análisis exploratorio de datos, entre otros. (Estadística para todos, 2016)

- *JHepWork*: Diseñado para los científicos, ingenieros y estudiantes, jHepWork es un framework para análisis de datos libre y de código abierto que fue creado como un intento de hacer un entorno de análisis de datos usando paquetes de código abierto con una interfaz de usuario comprensible y para crear una herramienta competitiva a los programas comerciales. Esto se hace especialmente para las ploteos científicos interactivos en 2D y 3D y contiene bibliotecas científicas numéricas implementadas en Java para funciones matemáticas, números aleatorios, y otros algoritmos de minería de datos. jHepWork se basa en Jython un lenguaje de programación de alto nivel, pero codificación en Java también puede ser usada para llamar librerías jHepWork numéricas y gráficas. (Jimmy W. Maco Elera, 2017)



- *KNIME*: KNIME (Konstanz Information Miner) es una plataforma de código abierto de fácil uso y comprensible para integración de datos, procesamiento, análisis, y exploración. Ofrece a los usuarios la capacidad de crear de forma visual flujos o tuberías de datos, ejecutar selectivamente algunos o todos los pasos de análisis, y luego estudiar los resultados, modelos y vistas interactivas. KNIME está escrito en Java y está basado en Eclipse y hace uso de sus métodos de extensión para soportar plugins proporcionando así una funcionalidad adicional. A través de plugins, los usuarios pueden añadir módulos de texto, imagen, procesamiento de series de tiempo y la integración de varios proyectos de código abierto, tales como el lenguaje de

programación R, WEKA, el kit de desarrollo de Química y LIBSVM. (Jimmy W. Maco Elera, 2017)

- *STATGRAPHICS:* es un programa de estadística de fácil manejo y una buena herramienta para la enseñanza de la estadística en secundaria y para la investigación en educación. (Estadística para todos, 2016)

- *STATISTICA:* **Statistica** es otro de los programas más usados a nivel mundial para el análisis estadístico. Entre todas las bondades y pruebas estadísticas que incluye, destaca la variedad de gráficos y la facilidad en el manejo de bases de datos. (Estadística para todos, 2016)
- *PH-STAT 2.5:* **PH-Stat** es un complemento de Excel producido por la Editorial Prentice Hall y acompaña a varios de sus libros de texto sobre estadística. Destaca la posibilidad de poder crear gráficos de control de calidad, diagramas de tallos y hojas, cajas de dispersión, intervalos de confianza en estimación, análisis de varianza, entre otros. El software puede emplearse libremente cuando se adquieren textos de Prentice Hall. (Estadística para todos, 2016)

Elección del software más adecuado para este caso de estudio de mejora de la atención al cliente en el Ayuntamiento de San Cugat:

Inicialmente este proyecto se comenzó haciendo un pre-análisis de los datos con la herramienta Excel, y gracias a ello se obtuvieron las conclusiones presentadas en el capítulo 2, apartado D. Esta herramienta es muy válida en el análisis de datos debido a sus atribuciones:

1. **Entrada de datos.** Excel tiene algunas características interesantes (incluyendo la validación de los mismos) que hacen que la entrada de los datos sea muy ágil. La validación de datos asegura que los datos introducidos se ajustan a los requisitos especificados por el usuario y puede reducir los errores en este tipo de tareas tediosas. Los datos se han ido guardando a lo largo de todo el uso de este software en formato CSV. Este formato es el más ampliamente utilizado ya que puede ser leído por la mayoría de software de datos.

2. **Cálculos básicos.** Excel es muy rápido para cálculos de estadísticas descriptivas.

3. **Ver los datos de un vistazo.** Excel puede ser útil para realizar un rápido resumen visual de los datos. Hay maneras de hacerlo en R, pero la vista de hoja de cálculo no se presenta de forma predeterminada.

4. **Presentación de datos y resumen.** Excel puede ser útil para la presentación estéticamente agradable de hojas de cálculo. (Guido Corradi, 2014)

Excel es un gran aliado en el análisis de dato y con esta herramienta se obtuvieron conclusiones relevantes que se presentaron a los interlocutores del ayuntamiento de San Cugat en nuestra primera reunión con ellos, pero para la segunda parte del proyecto se decidió utilizar el software

R con el objetivo de analizar los datos profesionalmente y obtener resultados estadísticos de interés.

A pesar de que tanto Excel como R son buenos programas a la hora de análisis de datos, surgieron varias razones que hicieron que a medida que se avanzaba el proyecto, empezara a trabajar con *RStudio*.

Los motivos que llevaron a este cambio de herramienta de análisis de datos han sido:

R es un paquete de software y un lenguaje para el análisis estadístico y gráfico. Este hecho conlleva una serie de beneficios de este software con respecto a Excel (Guido Corradi, 2014):

1. **Manipulación de datos.** R te permite manipular (seleccionar, recodifica, recuperar) datos muy rápidamente.

2. **Más fácil automatización.** R utiliza un lenguaje de script en lugar de una interfaz gráfica de usuario, por lo que es mucho más fácil de automatizar código en R que en Excel. Esto hace que se ahorre mucho tiempo, especialmente cuando se ha tenido que volver a ejecutar el mismo análisis varias veces.

3. **Cálculo rápido.** Debido a la automatización proporcionada por R, muchas operaciones son mucho más rápidas para realizar en R de Excel.

4. **Lee cualquier tipo de datos.** R puede leer prácticamente cualquier tipo de datos (.txt, .csv, .dat, etc). También hay paquetes de R específicamente diseñados para leer archivos JSON, SPSS, Excel, SAS, STATA. E incluso se puede usar los datos de cualquier sitio web y ejecutar consultas SQL.

5. **Más fácil Organización de Proyectos.** En Excel, los proyectos se organizan a menudo en diferentes pestañas del mismo libro. Esto puede hacer que los archivos de Excel sean lentos, torpes y difíciles a la hora de trabajar con ellos. Es más fácil mantener un proyecto organizado cuando se trata de R porque las diferentes tareas o sub-proyectos se pueden guardar en archivos separados almacenados en la misma carpeta y unidos entre sí en un mismo proyecto con *RStudio*.

6. **Es compatible con grandes conjuntos de datos.** Excel tiene restricciones para el tamaño pueden tener sus datos. Y aún en el caso de tus datos no superen este tamaño máximo, Excel suele ser lento con grandes conjuntos de datos (sobre todo después de añadir pestañas, fórmulas y referencias.

7. **La replicabilidad.** R tiene características que hacen mucho más fácil replicar los resultados del análisis; algo que es importante para la detección de errores. En primer lugar, es fácil de agregar comentarios a las secuencias de comandos. Por contra, es difícil documentar los pasos que has hecho en Excel. En segundo lugar, los datos y el análisis permanecen separados en R, lo que permite ver la progresión lógica. En Excel, sin embargo, los datos y las fórmulas están juntos, y puede ser difícil de seguir los pasos que dio el analista de los datos. En tercer lugar, se puede utilizar el control de versiones con *git* para realizar un seguimiento (y revertir) los cambios que realicen en el tiempo y para compartir sus scripts con otros y colaborar en proyectos como una comunidad.

8. **Más fáciles de encontrar y corregir los errores.** Debido R utiliza secuencias de comandos en lugar de "hacer clic" y permite comentarios y control de versiones, se puede ver un historial de las acciones realizadas para lograr cada resultado. Esto hace

que sea más fácil encontrar y solucionar errores. En Excel, sin embargo, los errores se pueden ocultar en las fórmulas de cualquier celda y allí pueden ser difíciles de encontrar.

9. **Es código abierto.** A diferencia de Excel y otros paquetes estadísticos que se utilizan en análisis de datos, R no es una caja negra. Puedes examinar el código para cualquier función o cálculo que se realiza. De hecho, puedes incluso modificar y mejorar estas funciones cambiando el código.
10. **Estadística avanzada.** R tiene muchas más (y más avanzadas) capacidades estadísticas que Excel. También es más rápido y flexible. Parte de las capacidades avanzadas de R se deben al punto anterior: R es código abierto.
11. **Los gráficos.** R tiene capacidades avanzadas de gráficos. Se puede crear gráficos chulísimos utilizando tanto el paquete básico de R como *lattice* o *ggplot*. A la gente le gusta ver los datos y R proporciona algunas de las herramientas para la creación de visualizaciones más bonitas del mercado.
12. **Se ejecuta en muchas plataformas.** Puede utilizar R en Windows, Mac, Linux y Unix.

En resumen, hay un número importante de razones que hicieron que a medida que avanzaba el proyecto, se probaran a realizar distintas pruebas con *RStudio* para comprobar que efectivamente gracias a este software podían encontrarse beneficios interesantes para este TFM. Concluyendo se puede decir que debido a los siguientes motivos, se ha trabajado tanto con Excel como con R en la realización de este TFM:

- ✓ R es considerado 'el idioma' de la ciencia de datos,
- ✓ R es mucho más hábil y eficiente en la preparación de datos que Excel.
- ✓ En R puede automatizarse el código.
- ✓ El número de paquetes R está aumentando de forma exponencial (+),
- ✓ R lee cualquier tipo de entrada de datos
- ✓ R es compatible con los datos de mayor tamaño, y puede soportar grandes volúmenes de datos con paquetes. Esto ha sido muy importante a la hora de tomar la decisión, ya que para este trabajo fin de máster, el elevado número de datos (más de 400.000 líneas en la base de datos) estudiados hacía que la herramienta Excel funcionara lentamente.
- ✓ Está demostrado que Excel y otras hojas de cálculo muestran inexactitudes importantes para análisis básicos como la regresión lineal.

R fue diseñado específicamente para hacer análisis estadístico, por lo que es más preciso y exacto para el análisis de datos.

5. MODELOS LINEALES: REGRESIÓN, ANOVA y ANCOVA (técnicas de análisis estadístico)

En el caso de estudio de este TFM se plantea un problema estadístico, el estudio de la variable "tiempo de espera", y para solucionar este problema estadístico se deberá aplicar y usar un modelo lineal. Es por este motivo que este capítulo se centrará en explicar dichos modelos lineales.

El análisis de **regresión** se usa para explicar o modelar la relación entre una variable continua Y, llamada variable respuesta o variable dependiente, y una o más variables continuas X_1,.....,X_p, llamadas variables explicativas o independientes. Cuando p = 1, se denomina regresión simple y cuando p > 1 se denomina regresión múltiple.

Si las variables explicativas son categóricas en vez de continuas entonces nos enfrentamos ante un caso típico de análisis de la varianza o ANOVA. Al igual que antes, si p = 1, el análisis se denomina ANOVA una factorial, mientras que si p > 1 el análisis se denomina ANOVA multifactorial.

En este trabajo se utilizarán técnicas estadísticas que analizan la relación entre varias variables. Con estas técnicas se modelará la dependencia que tienen las distintas variables con la variable respuesta.

Por último, es posible que en el mismo análisis aparezcan tanto variables explicativas continuas como categóricas, y en este caso el análisis pasaría a denominarse análisis de la covarianza o ANCOVA. Aquí ya no haríamos distinción entre único o múltiple ya que este análisis se compone siempre de, al menos, dos variables explicativas (una continua y una categórica). (Cayuela, 2014)

En este TFM, **la variable respuesta será la variable "Tiempo de espera",** y se ha de decir que todos estos modelos que se han nombrado anteriormente están dentro de la categoría de modelos lineales.

Este capítulo se centrará en las técnicas univariadas de la regresión, ANOVA y ANCOVA ya que serán las técnicas que necesitemos para resolver el problema estadístico que se debe resolver en el capítulo 6. Ya que en R todos los análisis univariados de este tipo se ajustan utilizando una única función, la función lm(), ya que la forma de ajustar cualquiera de estos modelos es idéntica, independientemente de que tengamos una o más variables explicativas y de que estas sean continuas o categóricas. (Cayuela, 2014)

Regresión simple

En estadística la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes Xi y un término aleatorio ε.

Cuando hay sólo una variable independiente (p=1) se dice regresión simple y cuando hay más (p>1) regresión múltiple.

El modelo de regresión lineal supone que existen unas constantes β0, β1, …, βp (desconocidas) tales que el valor de Y depende de los valores de X_1, X_2,…,X_p de la forma:

a. Y= β0+ β1∗X1 + β2∗X2 +...+ βp∗Xp + ERROR, donde ERROR es una variable aleatoria NORMAL de media 0 y varianza también desconocida.

Análisis de la varianza (ANOVA)

La técnica de técnicas denominada Análisis de la varianza (ANOVA), del acrónimo *Analysis of variance*, tiene como objetivo básico la comparación de las medias de más de dos poblaciones.

El nombre de Análisis de la varianza, sin embargo, no es muy afortunado. En el ANOVA se comparan siempre las medias de varias poblaciones y se hace a través de un contraste de hipótesis donde se analiza la varianza, es cierto; pero no sólo eso, porque también se analizan las diferencias de medias que hay entre las muestras, y también se analiza el tamaño de muestra.

Un análisis de la varianza permite determinar si diferentes tratamientos muestran diferencias significativas o por el contrario puede suponerse que sus medias poblacionales no difieren. El análisis de la varianza permite superar las limitaciones de hacer contrastes bilaterales por parejas que son un mal método para determinar si un conjunto de variables difiere entre sí. (Cayuela, 2014).

El ANOVA parte de algunos supuestos o hipótesis que han de cumplirse:

- La variable dependiente debe medirse al menos a nivel de intervalo.
- Independencia de las observaciones.
- La distribución de los residuales debe ser normal.
- Homocedasticidad: homogeneidad de las varianzas.

A las variables X_1, X_2, etc. que son cualitativas se les llama FACTORES, y a cada valor que puede tomar se le llama NIVEL.

El modelo ANOVA supone que para cada nivel xi del factor X, se tiene que Y=μ i+ ERROR, donde ERROR es una variable aleatoria NORMAL de media 0 y varianza desconocida, la misma para todos los niveles.

Comparación a posteriori: Test de Fisher:

- Si el contraste resulta aceptar H0 (igualdad de medias) entonces admitimos que el factor X NO influye sobre Y.
- Pero si resulta rechazar H0, entonces admitimos que hay niveles del factor X que dan lugar a mayores medias de Y.
- Las comparaciones a posteriori crean el ránking de niveles que dan medias de Y de mayor a menor, si es que se distinguen.

Análisis de la covarianza (ANCOVA)

El análisis de la covarianza es un modelo lineal, donde la variable respuesta es de tipo continua y cumple con las asunciones del resto de los modelos lineales y donde existen dos tipos de variables explicativas, una de tipo continua y una de tipo categórica o discreta.

Así, este análisis se utiliza para probar diferencias en las relaciones de las dos variables continuas entre los factores de la variable categórica. Es decir, que compara la pendiente de las regresiones (de la variable respuesta y la explicativa continua) entre los diferentes niveles de la variable discreta.

En R, los modelos lineales univariados se ajustan a los datos usando la función base "lm" en la que la fórmula de dependencia en un modelo aditivo es del tipo:

y ~ x + x1, donde y es la variable respuesta (continua), x es la variable respuesta (continua) y x1 es la variable respuesta discreta.

6. CASO DE ESTUDIO: Análisis de los datos de la oficina de atención al cliente del ayuntamiento de San Cugat y optimización de la oficina y de los servicios ofrecidos.

En el presente capítulo de esta memoria describiremos los pasos que se han llevado a cabo para analizar los datos recibidos del ayuntamiento de San Cugat y las conclusiones obtenidas de dicho análisis.

Primeramente, se explicará la estructura de este capítulo: punto de partida, objetivos propuestos, pasos a seguir y finalmente las conclusiones obtenidas.

I) Oficina de atención al ciudadano del ayuntamiento de San Cugat.

II) Datos recolectados de todas las visitas de ciudadanos a la oficina de atención al cliente desde enero del año 2011 hasta diciembre de 2016 y de cada servicio que han solicitado.

III) Para la realización de este TFM se han definido las siguientes variables de interés que afectan durante todo el proceso de atención de un usuario en la oficina de atención al cliente del Ayuntamiento de Sant Cugat y que serán de estudio durante este capítulo:

   a. Número de personas
   b. Número de gestores para cada oficina y servicio
   c. Tiempo de servicio
   d. Tiempo de espera → Es predecible que si el tiempo de espera baja, la satisfacción de los usuarios que acudan a la oficina de atención al cliente aumente. Es por este motivo, que esta variable en particular será, de entre todas, la que más analizaremos y en la cual nos centraremos con el objetivo de poder modelizarla y encontrar con que factores está relacionada.

Por todo lo anterior, concluimos que se ha definido como objetivo general de este TFM el siguiente: Análisis temporal y optimización del conjunto de la oficina y de los servicios ofrecidos en la oficina de atención al cliente del Ayuntamiento de Sant Cugat.

Para la correcta consecución del objetivo general definido anteriormente se hará uso de las herramientas de análisis de datos *R y RStudio*.

Para cumplir el objetivo general que ataña a este TFM, deben cumplirse los dos objetivos particulares siguientes:

3. **Caracterización de las variables numeradas en el apartado III, que influyen en los servicios que se realizan en la oficina de atención al cliente del Ayuntamiento de Sant Cugat durante los años 2011,2012,2013,2014,2015 y 2016; análisis exploratorio de los datos.**
4. **Optimización del funcionamiento de la oficina de atención al cliente. Relación y dependencias existentes entre las variables del modelo estudiado en este TFM: Análisis estadístico.**

Una vez definido el objetivo general, los pasos siguientes que hay que seguir son:

Dentro del objetivo 1 acerca de la caracterización de las variables se han de definir los siguientes objetivos específicos:

- Explorar los datos y estudiar el número de servicios que se ofrecen para cada una de las oficinas de atención al cliente del Ayuntamiento de Sant Cugat.
- Analizar la evolución temporal de las siguientes variables: tiempo de espera, tiempo de servicio, número de gestores y número de usuarios: obtención de gráficas. Se buscarán patrones temporales.

Dentro del objetivo 2 sobre la optimización de las oficinas de atención al cliente, se han definido los siguientes objetivos específicos con el fin de llegar a la construcción del modelo matemático:

- Modelizar la variable tiempo de espera (teoría de colas)
- Identificar las variables explicativas y la variable respuesta. La variable respuesta en este caso será: el tiempo de espera.
- Análisis exploratorio de la relación entre las variables explicativas y la de respuesta a diferentes escalas temporales (patrón diario, semanal, mensual, anual).
- Construcción del modelo matemático que define la variable respuesta: Tiempo de espera (regresión lineal, y=a +bx)

En caso necesario podría repetirse todo este proceso para la variable "número de usuarios" (línea futura).

Una vez vistos los puntos de interés y los objetivos particulares en los que se harán foco en este TFM se procede a la realización mediante el software RStudio de cada uno de los anteriores apartados. Se usará RStudio para la realización de este TFM ya que es un entorno de desarrollo integrado con consola y editor de sintaxis para R.

En este capítulo no se detallará todo el código de R, sino que solo se incluirán las imágenes y gráficas relevantes que se obtengan de la ejecución del código en RStudio. El código en R de este TFM puede consultarse al completo en el último apartado llamado Anexos de esta memoria.

Todas las gráficas e imágenes que se muestren a continuación irán acompañadas de las correspondientes observaciones necesarias para cada una de ellas.

A continuación, se procede a la explicación del código en que se ha usado para la realización de este proyecto:

Pasos comunes tanto para el objetivo 1 como para el objetivo 2:

Todo código en R comienza indicando cual es el directorio en el que se va a trabajar y a continuación se ha cargado en RStudio la base de datos.

Previo a la realización del primer objetivo se ha hecho un estudio de la base de datos facilitada por el Ayuntamiento de Sant Cugat.

La base de datos recibida del ayuntamiento de San Cugat, estaba en formato .csv y Excel, por lo que se ha comprobado que en los datos de los 415.614 registros no había ninguna discordancia que pudiera distorsionar el resultado del análisis de los datos.

Por este motivo se ha comprobado en un primer análisis que no hubiera outliers destacados que indicaran algún dato erróneo en la base de datos. Más adelante, se detallará alguna observación relativa a estos outliers que se encontraron una vez se fue avanzando con el análisis.

Dado que en este trabajo de fin de máster se van a realizar muchos gráficos en R, se ha hecho primero una comprobación de cómo quedará una gráfica dibujada con el comando *plot*.

El ejemplo concreto que se ha probado ha sido el de representar la tipología de tramites existentes para cada Oficina:

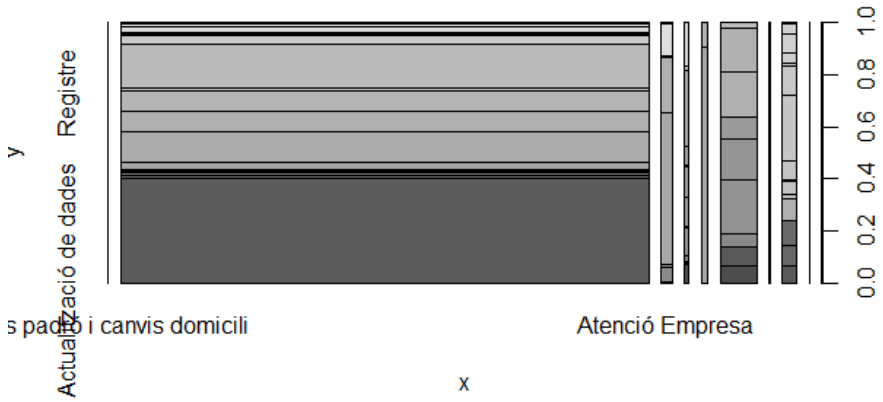


Ilustración 19 Ejemplo gráfica realiza con comando plot de R

Como era de esperar, la gráfica es ilegible, por este motivo se ha instalado el paquete *ggplot2* para poder representar mejores gráficas de los datos analizados (Chang, 2012).

Si representamos el ejemplo anterior con el paquete *ggplot2,* obtenemos la siguiente gráfica en la cual se pueden distinguir perfectamente que tipo de trámites se realizan por cada oficina de atención al cliente del Ayuntamiento de Sant Cugat:

En la primera figura observamos en el eje x las distintas oficinas existentes en el Ayuntamiento de Sant Cugat (atención ciudadana, atención empresas, empadronamientos, bienestar social, licencias de obras, oficina de consumo, educación, …) y en la columna y el número de personas que han acudido a cada una de ellas. Se observa que la **oficina de Atención Ciudadano/Registro** es la oficina a la que más gente ha acudido en los últimos seis años con un porcentaje superior al resto extremadamente alto.

Dentro de la citada oficina, los trámites más realizados han sido temas de guardería (en color naranja) y temas de empadronamientos (en color azul).

Existen 79 tipologías de tramites distintos para todas las oficinas existentes en la atención al cliente del Ayuntamiento de Sant Cugat.

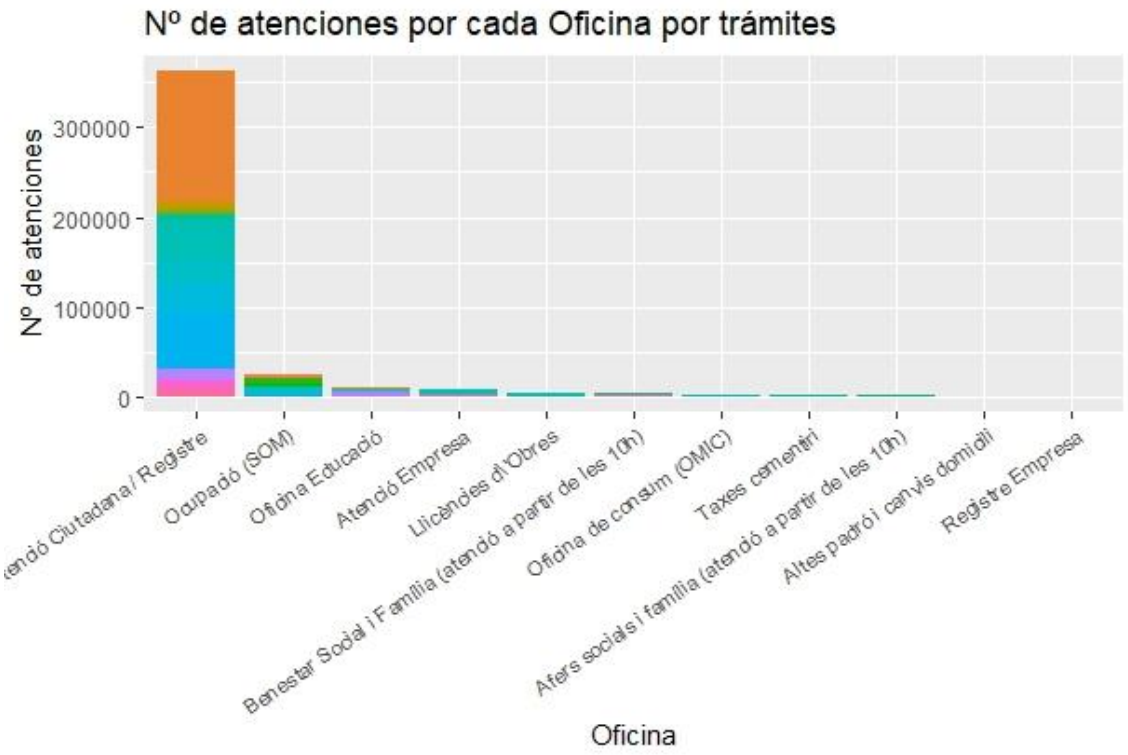


Ilustración 20 Número de atenciones en cada oficina de atención al cliente (diferenciando por tipología de trámite)

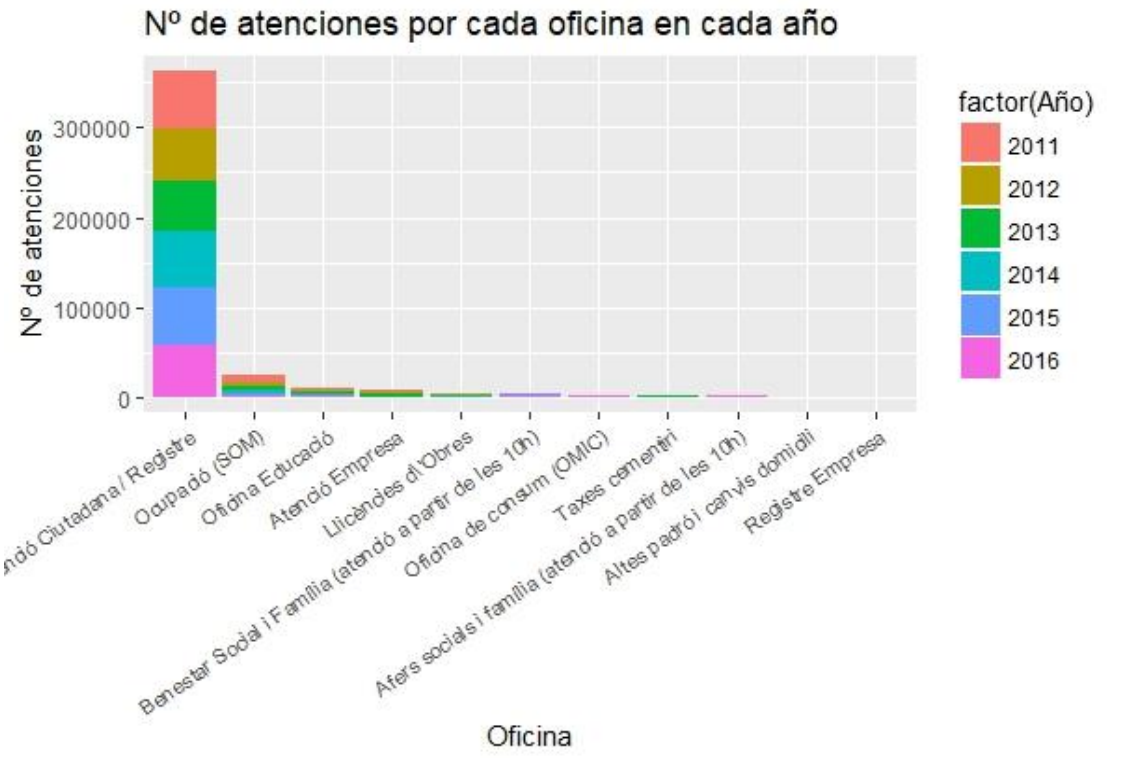


Ilustración 21Número de atenciones en cada oficina de atención al cliente (diferenciado por año)

En la segunda figura mostrada a continuación puede leerse la leyenda que ayuda a interpretar la gráfica anterior.

Ilustración 22 Leyenda de colores sobre la tipología de trámites para cada oficina

Se procede ahora a la redacción de los pasos que se han seguido para cumplir el objetivo marcado en este TFM.

Objetivo 1: Análisis exploratorio de los datos

Como primer paso para cumplir el objetivo general, se necesita la realización del objetivo 1:

1. **Caracterización de las variables numeradas en el apartado III (**Número de personas, Numero de gestores para cada oficina y servicio, Tiempo de servicio y Tiempo de espera)**, que influyen en los servicios que se realizan en la oficina de atención al cliente del Ayuntamiento de Sant Cugat durante los años 2011,2012,2013,2014,2015 y 2016.**

A continuación se procede a la realización de un análisis temporal de las diferentes variables que intervienen en el proceso de atención de un usuario en las oficinas de atención al cliente del ayuntamiento de San Cugat con la intención de caracterizarlas.

Análisis temporal de la variable: Número de atenciones.

Para caracterizar esta variable lo haremos mediante histogramas, ya que se ha observado que es la mejor manera de visualizar todas las variaciones en dicha variable.

Número de atenciones anualmente:

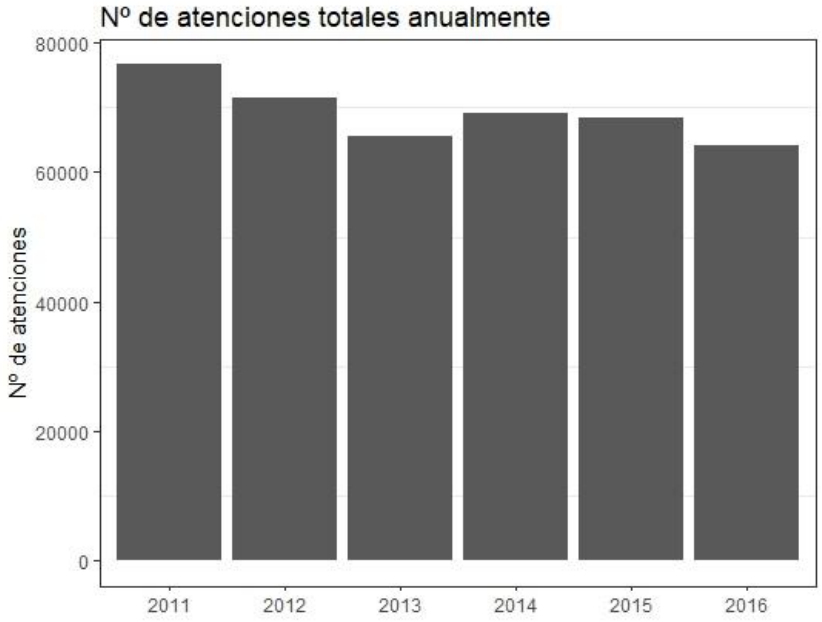


Ilustración 23 Gráfica Análisis temporal del número de atenciones Anualmente

Se observa que el número de atenciones anual tiene una tendencia a disminuir según el paso de los años. Es posible que este fenómeno ocurra ya que haya ciertos trámites que probablemente puedan solucionarse por internet y los usuarios no deban acudir presencialmente al Ayuntamiento de Sant Cugat.

Número de atenciones mensualmente:

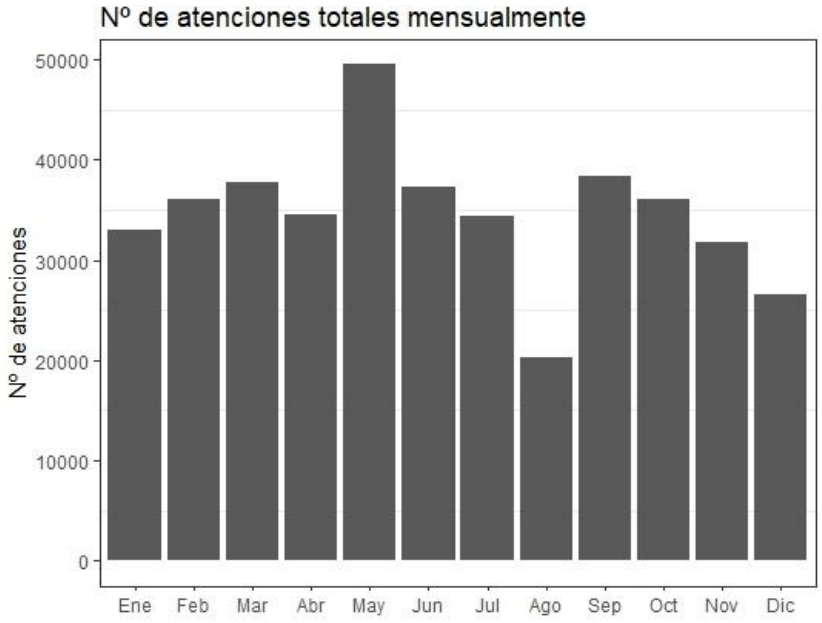


Ilustración 24 Gráfica Análisis temporal del número de atenciones mensualmente

En la gráfica anterior puede observarse que el mayor número de atenciones se hace en mayo, esto puede ser debido a diferentes campañas que ocurran en dicho mes y que obligue a los usuarios a acudir en dicho mes al Ayuntamiento de Sant Cugat. Por el contrario en el mes de

agosto, es cuando menos atenciones se producen, lo cual tiene sentido si se asocia a que la mayoría de la población está de vacaciones.

Número de atenciones mensualmente y diferenciando por año:

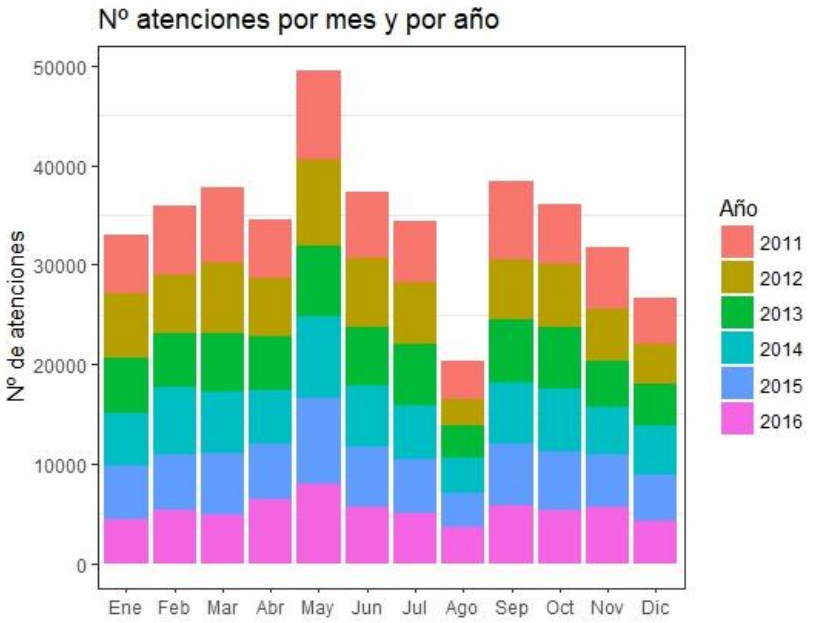


Ilustración 25 Gráfica Análisis temporal del número de atenciones mensualmente (diferenciando por año)

Con la anterior ilustración se observa que, pese a que durante los 6 años estudiados un porcentaje mayor de atenciones suceden en mayo, esta tendencia ha sucedido en los seis años estudiados por igual.

Número de atenciones diarias (1...31):

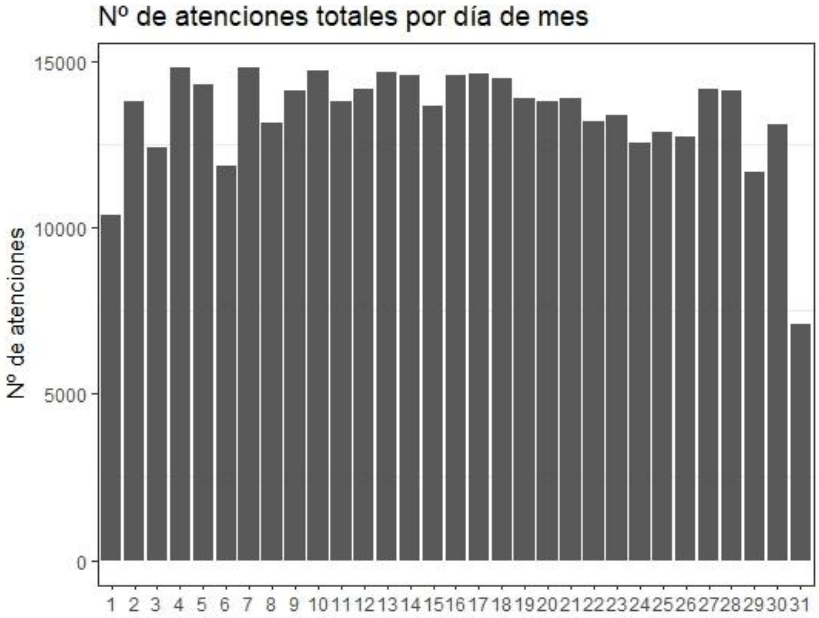


Ilustración 26 Gráfica Análisis temporal del número de atenciones diarias (1...31)

Número de atenciones diarias (1...31) y diferenciando por año:

Con la siguiente imagen se consigue mejorar considerablemente la visualización de la esta variable, ya que puede observarse que días de mes son los de mayor afluencia (7, 4, 10, 13...) y los de menor afluencia (1, 29, 6, 3, ...). Con estos datos se puede dimensionar el número de gestores para estar preparados frente a la afluencia esperada para cada día.

Es obvio que pese a que el día de mes con menor afluencia según el grafico sea el 31, este dato es incorrecto, ya que solo la mitad de los meses tienen dicho día, por lo que no se pueden sacar conclusiones sobre este día, como si pueden hacerse sobre por ejemplo el día 1, que efectivamente es un día con menos afluencia que el resto.

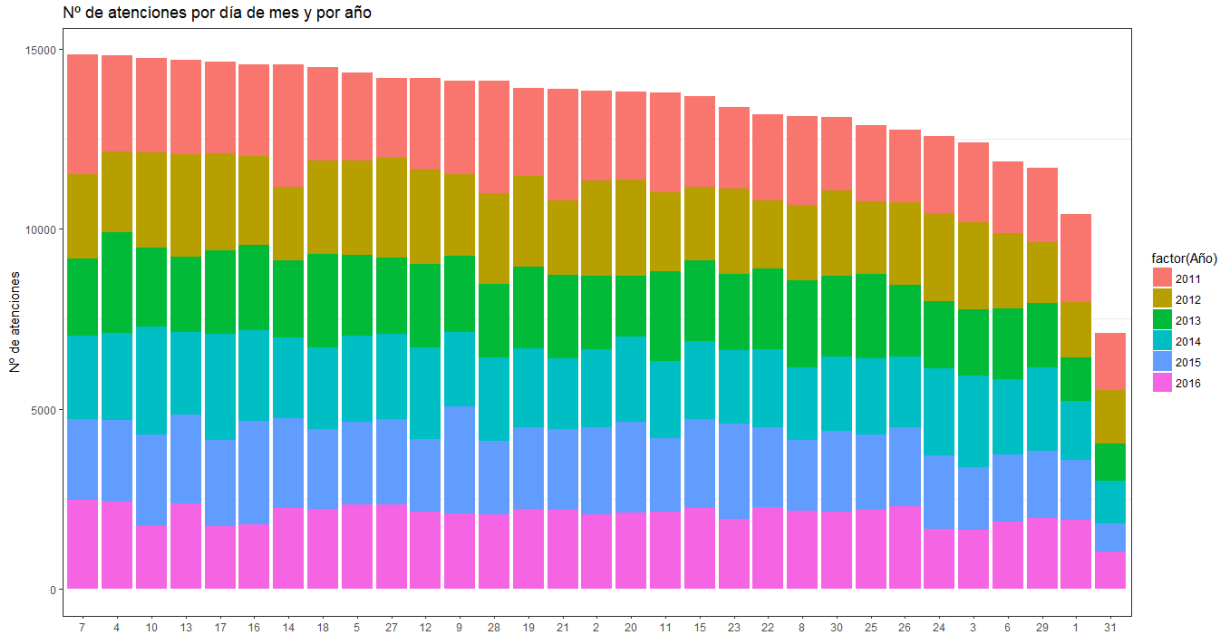


Ilustración 27 Gráfica Análisis temporal del número de atenciones diarias (1...31) (Diferenciando por año)

Número de atenciones cada día de la semana (lunes...domingo):

Se comprueba que el número de atenciones es menor los martes y los viernes.

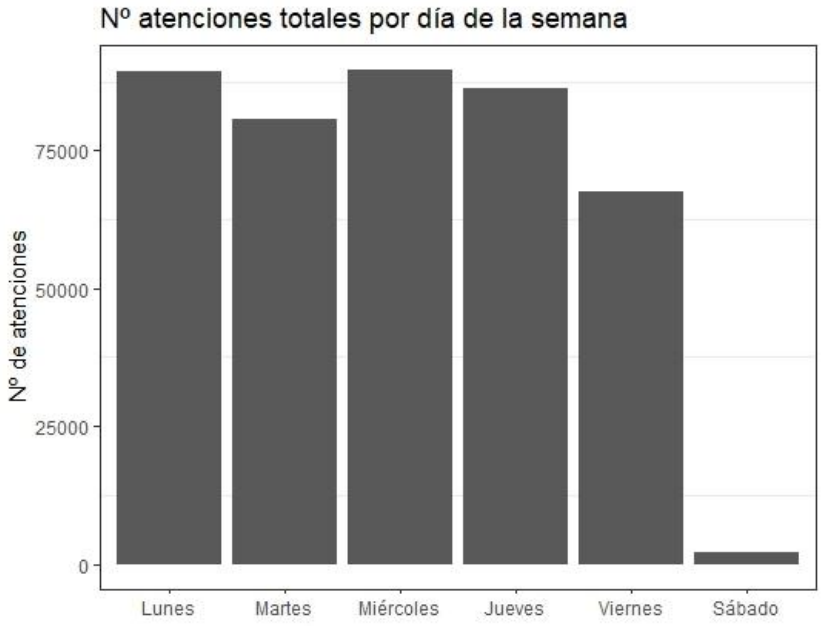


Ilustración 28 Gráfica Análisis temporal de atenciones de personas por día de la semana (lunes...Domingo)

Número de atenciones cada día de la semana (lunes...Domingo) y diferenciando por año:

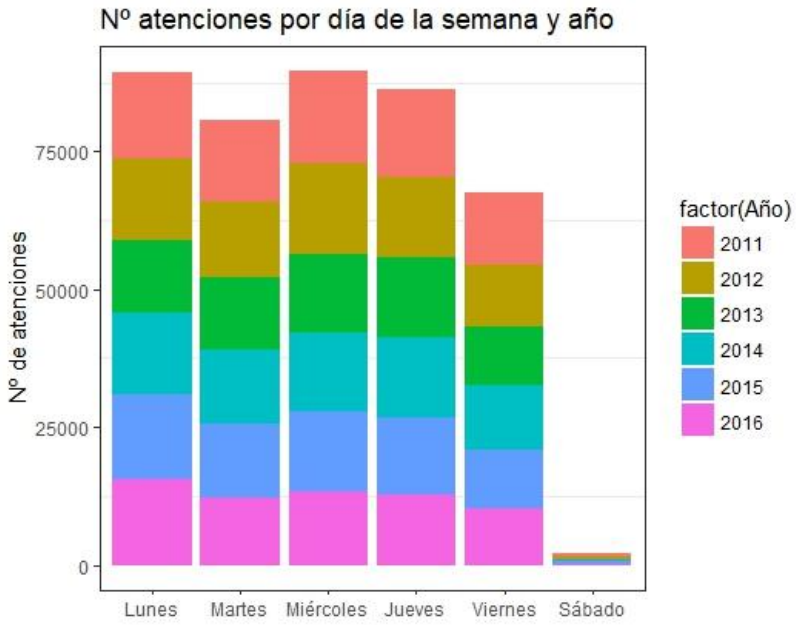


Ilustración 29 Gráfica Análisis temporal del número de atenciones por día de la semana (lunes...Domingo) (Diferenciando por año)

Análisis temporal de la variable: Tiempo de espera

Para poder caracterizar la variable tiempo de espera se ha de tratar adecuadamente la base de datos. Ya que los formatos horarios en Excel no funcionan igual para R.

Para poder trabajar con horas y fechas en R es necesario instalar la librería "*lubridate*" que facilitará todas las operaciones que se deseen realizar.

```
library(lubridate)

Data$Tiempo.espera2 <- hms(Data$Tiempo.espera)
```

En la base de datos se disponía de una columna llamada “Tiempo Espera”, en la que se había calculado el tiempo de espera que esperaba cada usuario antes de ser atendido.

Para trabajar en R con formato horario (HMS), se ha creado una nueva columna en la base de datos con la que se trabaja en R, llamada “Tiempo espera 2” en la cual se han introducido los valores de la columna “Tiempo Espera”, anteriormente mencionada, en formato HMS, formato identificado por R y que permite realizar operaciones y gráficas cómodamente.

Una vez se tiene la variable “Tiempo de espera 2” en formato HMS, transformamos dicho tiempo de espera de segundos a minutos. Primeramente se saca la información como duración (as.duration) para que esté en segundos, luego dejamos solo los valores numéricos de los segundos (as.numeric) y finalmente dividimos entre 60 y obtenemos el tiempo de espera en minutos. Estos minutos se guardarán en una nueva variable que hemos creado y que se llama “Tiempo de espera 3”:

```
Data$Tiempo.espera3<-
as.numeric(as.duration(Data$Tiempo.espera2))/60
```

Hecho esto, se ha comenzado a representar la evolución temporal de la variable Tiempo de espera, al igual que se hizo en el apartado anterior con la variable “Número de atenciones”.

Para caracterizar esta variable lo haremos mediante Boxplot (o gráficos de cajas), con la función:

```
geom_boxplot
```

ya que se ha observado que es la mejor manera de visualizar todas las variaciones en dicha variable y observar también así la presencia de outliers. Estos diagramas de cajas permiten conocer como se distribuyen los datos dentro de una variable, y se puede obtener mucha información relevante, mas que si solo hiciésemos un histograma.

Los gráficos de caja representan la información que se observa en la siguiente figura:

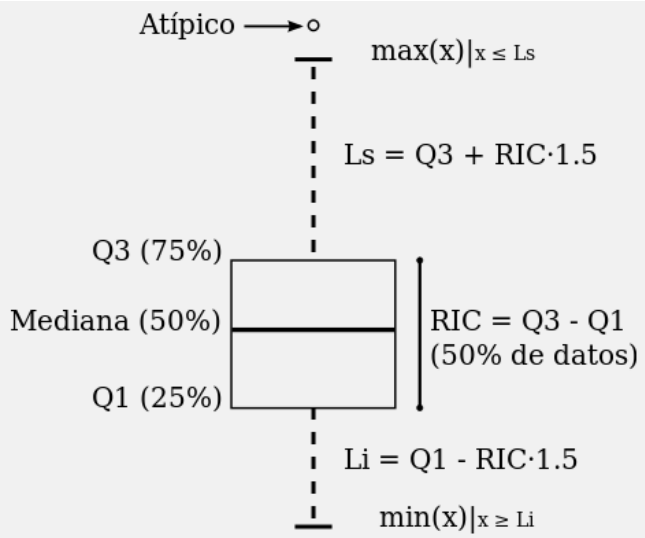


Ilustración 30 Información contenida en un boxplot. (Fuente: Wikipedia)

- **Mediana.** Valor que deja a la mitad de los casos por encima y a la otra mitad por debajo.

- **Primer Cuartil (Q1).** El 25% de los casos se encuentran por debajo de este valor.
- **Tercer Cuartil (Q3).** El 75% de los casos se encuentran por encima de este valor.
- **Rango Intercuartílico (RIC).** Es la diferencia entre el tercer y el primer cuartil.
- **Límites Superior o Inferior (Ls o Li).** Ls contiene los casos por encima de Q3 más 1,5 veces el rango intercuartílico o Li por debajo de Q1 – 1,5xRIC (Tukey).
- Los **valores atípicos** son aquellos que están más a allá de los límites inferior y superior. Cuando los valores atípicos están más allá de 3 veces el RIC en lugar del 1.5 son denominados valores extremos.

Los diagramas de caja son especialmente útiles cuando la distribución de una variable es asimétrica o se aleja de la distribución normal. En este tipo de casos interpretar una variable en función de su media o desviación estándar es un error puesto que estos estimadores no describen fielmente las características de nuestra muestra.

Análisis temporal de la variable tiempo de espera: anualmente

En las primeras pruebas que se hicieron para este gráfico, el eje "y" estaba en escala logarítmica (Ilustración 30) pero finalmente se ha representado el eje "y" con la variable "Tiempo de espera 3", que era en minutos y que facilita la visualización (Ilustración 31). Se ven los outliers de cada año en un color gris más claro.

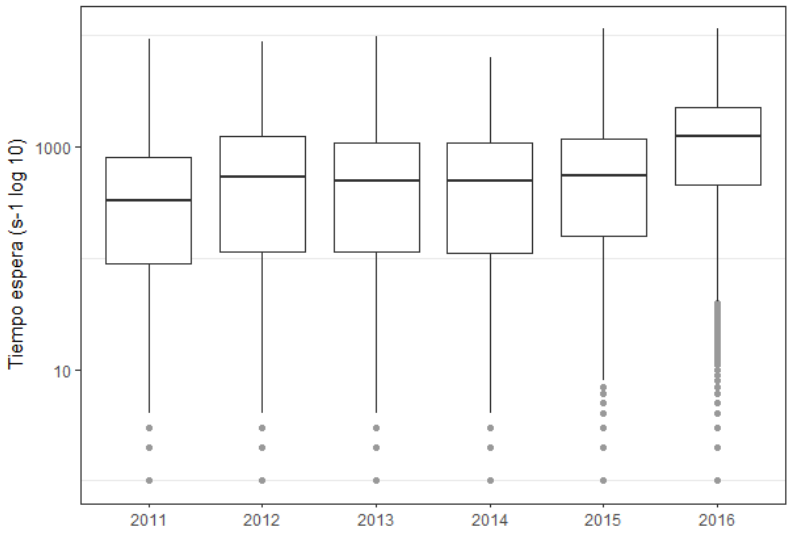


Ilustración 31 Gráfica variación tiempo espera anualmente (Eje Y logarítmico)

Las líneas azul y roja representan el tiempo de espera de 60 minutos y 120 minutos respectivamente para facilitar la comprensión de los gráficos de cajas.

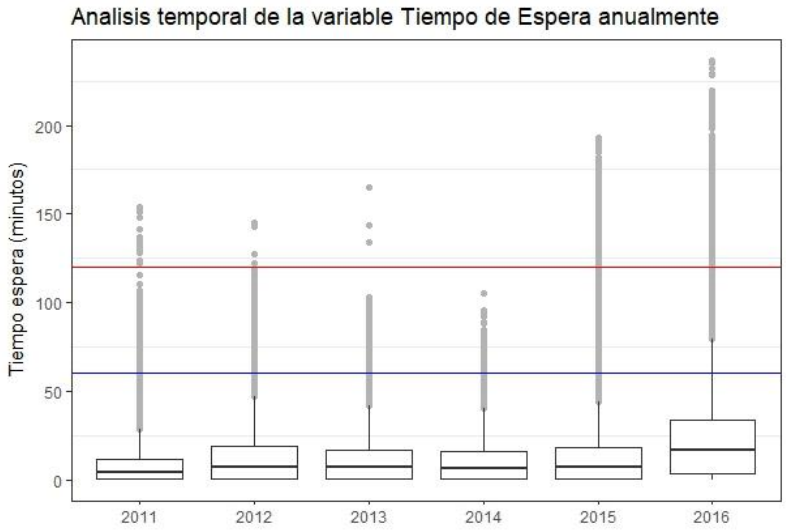


Ilustración 32 Gráfica variación tiempo espera anualmente (Eje Y minutos)

Es notable el aumento considerable del tiempo de espera en el año 2016. Se debe concretar una reunión con el Ayuntamiento de Sant Cugat para tratar este tema y encontrar las posibles casusas que hayan causado este resultado.

Se observa que el tiempo de espera aumenta a partir del año 2014. Adicionalmente para representar la variación de la variable tiempo de espera, se ha creído oportuno añadir un gráfico de violines con la función:

```
ggplot(data=Data, aes(x=Año, y=Tiempo.espera3))+
       geom_violin(trim = FALSE) +
            geom_boxplot(width = 0.2)
```

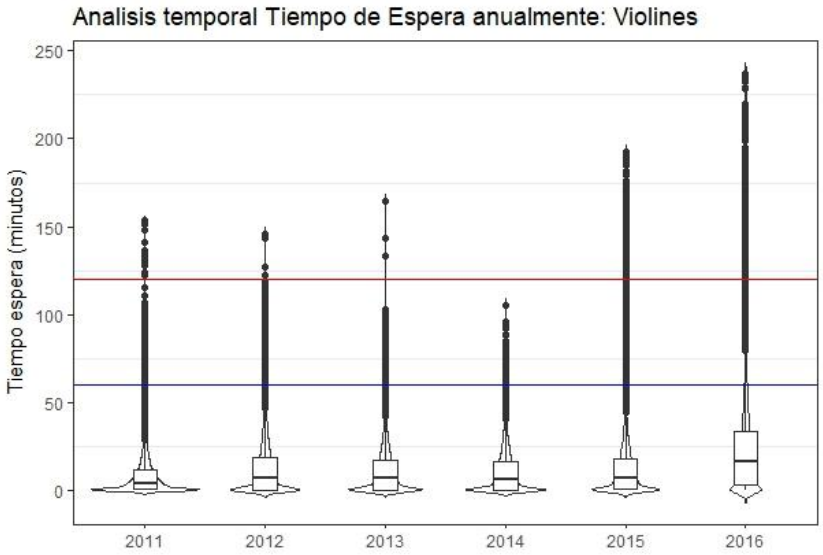


Ilustración 33 Gráfica violines variación tiempo espera anualmente (Eje Y minutos)

Análisis temporal de la variable tiempo de espera: mensualmente

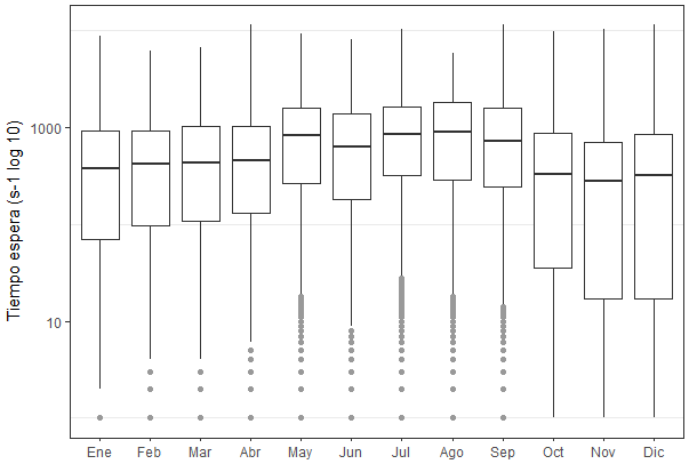


Ilustración 34 Gráfica Análisis temporal del tiempo de espera mensualmente (log)

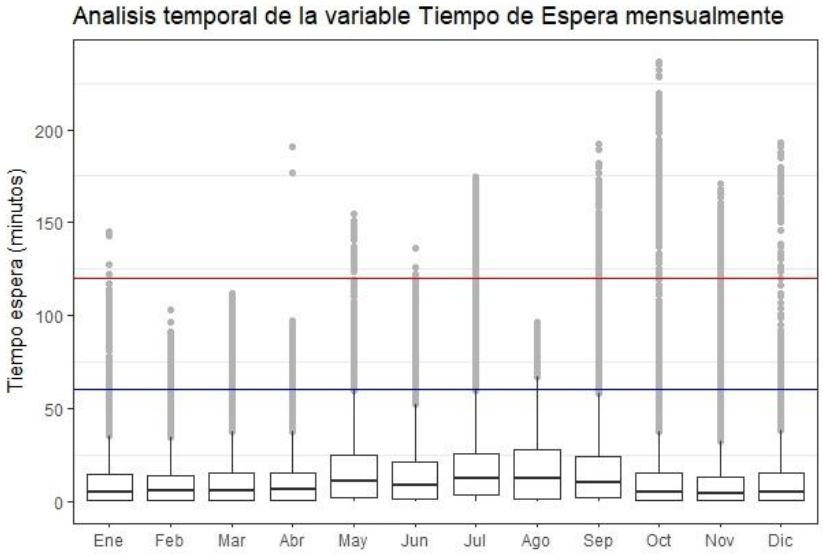


Ilustración 35 Gráfica Análisis temporal del tiempo de espera mensualmente (minutos)

En ambas gráficas se visualiza de igual modo la tendencia que siguen los meses: en mayo, julio y agosto, el tiempo de espera es mayor. El resultado en mayo, puede ser producido por el hecho de ser el mes en que más atenciones se producen y esto radica en un mayor tiempo de espera.

En los meses de julio y agosto probablemente esto sea debido a que hay un número menor de gestores atendiendo a los usuarios, y esto también genera un aumento en el tiempo de espera. Sería conveniente que para estos meses se busquen fórmulas para reforzar los gestores que atienden a usuarios.

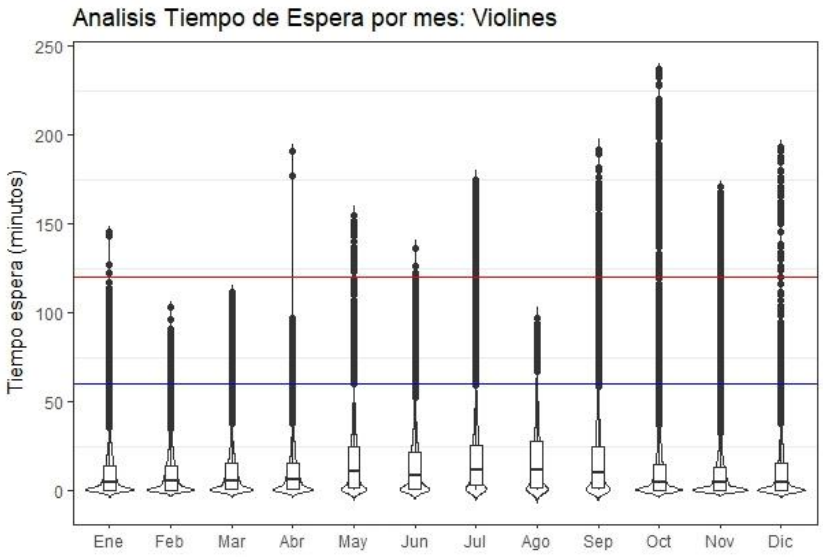


Ilustración 36 Gráfica violines Análisis temporal del tiempo de espera mensualmente (minutos)

Análisis temporal de la variable tiempo de espera: día de mes

En la primera gráfica el eje y es logarítmico, y en la segunda el eje y está en minutos ya que se facilita la visualización. Se ven los outliers de cada año en un color gris más claro.

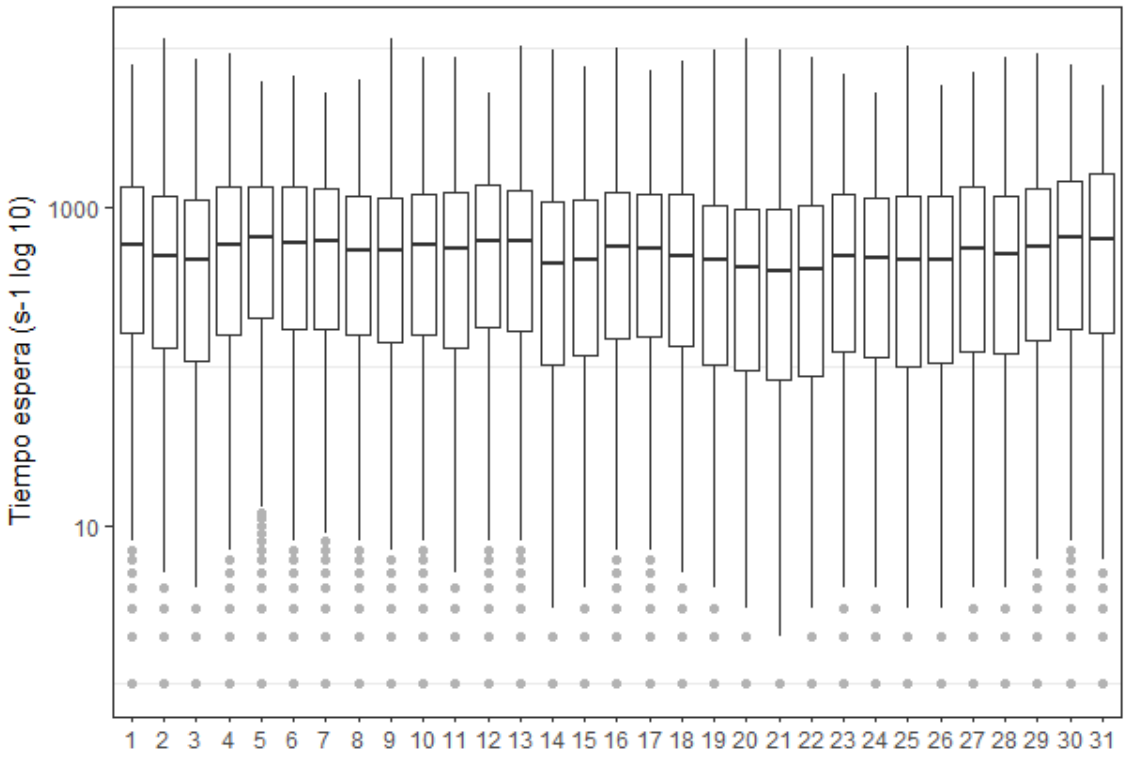


Ilustración 37 Gráfica Análisis temporal del tiempo de espera por día de mes (log)

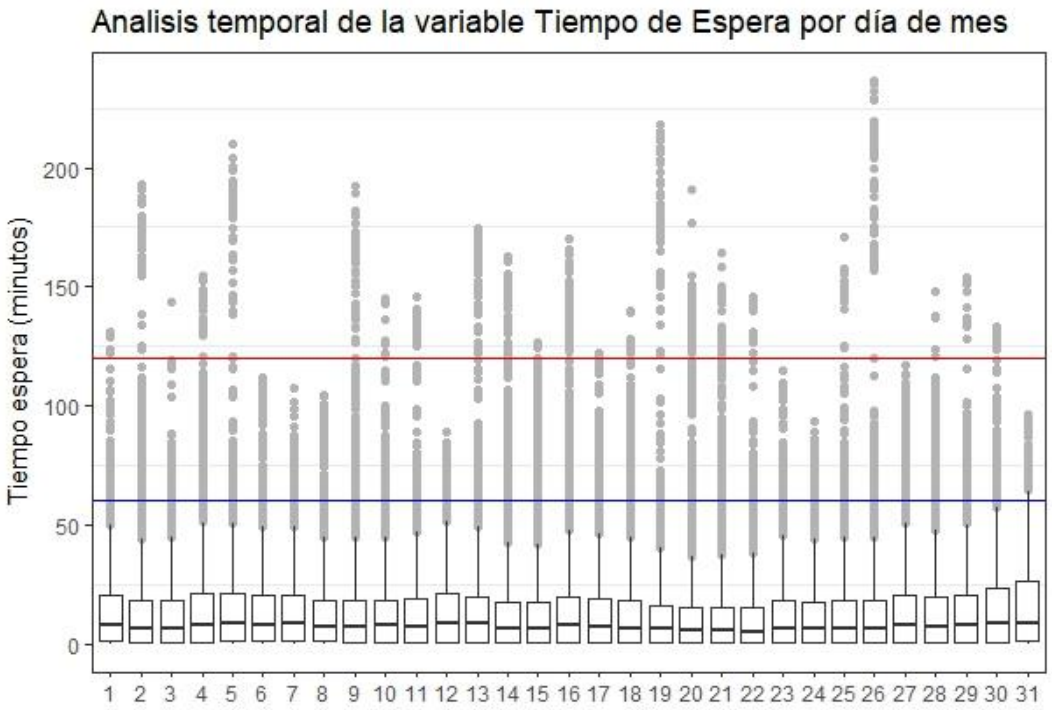


Ilustración 38 Gráfica Análisis temporal del tiempo de espera por día de mes (minutos)

Se observa es que el tiempo de espera no suele ser muy largo, pero lo que si hay son muchos tiempos excesivamente largos (visualizados como outliers).

Análisis temporal de la variable tiempo de espera: día de semana

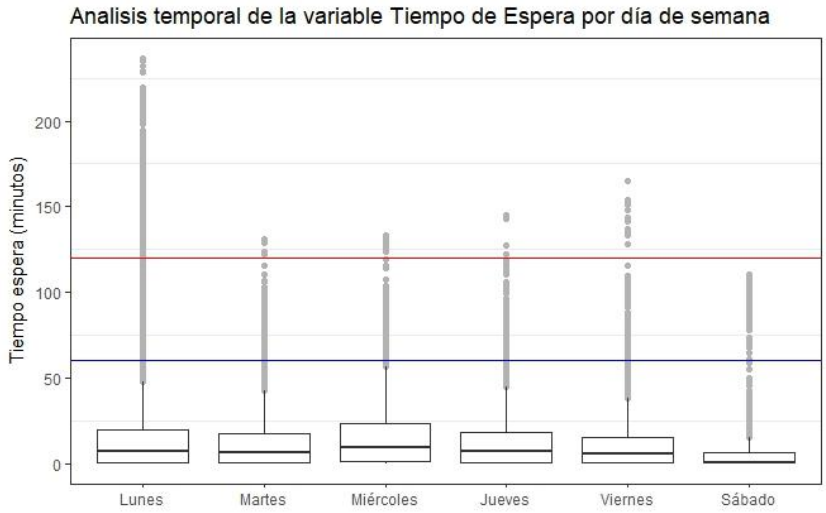


Ilustración 39 Gráfica Análisis temporal del tiempo de espera por día de la semana (minutos)

Tanto en la gráfica 39 como en la 40 se puede observar que la media del tiempo de espera es menor los martes y viernes, lo cual encaja con que dichos días eran los de menor afluencia que vimos en la caracterización de la variable "número de atenciones".

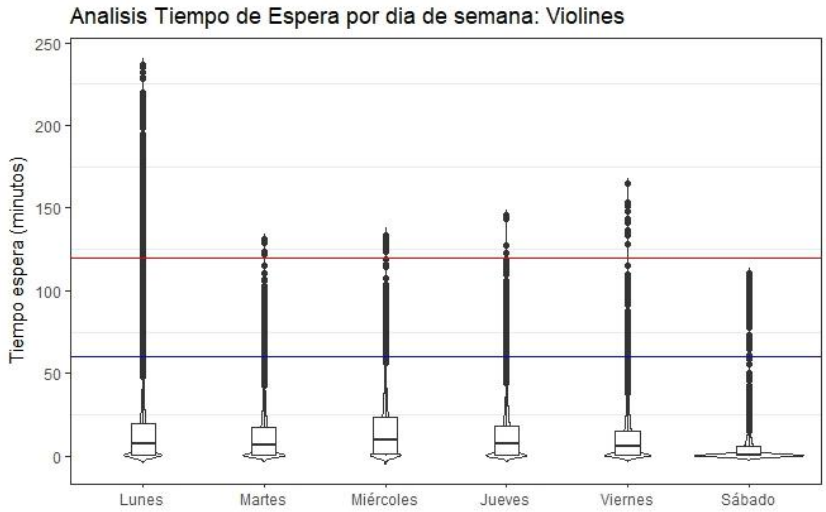


Ilustración 40 Gráfica violines Análisis temporal del tiempo de espera por día de la semana (minutos)

Análisis temporal de la variable: Número de gestor

Número de atenciones que realiza cada gestor a lo largo de los seis años estudiados en este análisis:

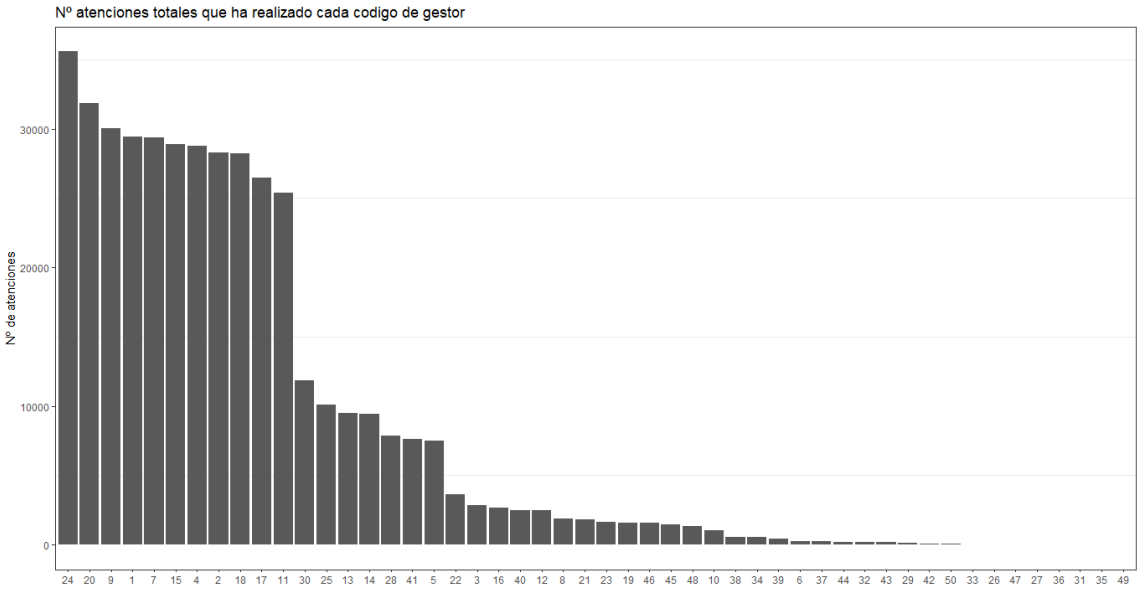


Ilustración 41 Número de atenciones totales que ha realizado cada gestor desde 2011 hasta 2016

Gráfica del número de atenciones que realiza cada gestor. Se ordenan de mayor a menor número de atenciones realizadas. Se observa que los gestores 24, 20, 9, 1 ... son los que más atenciones han realizado. Y los gestores 49, 35, 31, 36, 27,47, … solo han realizado alguna atención puntual. Quizás son gestores que puntualmente han tenido que atender a clientes, pero quizás no sea su trabajo, o han tenido que sustituir a otro gestor.

Para caracterizar esta variable se hará mediante histogramas, ya que se ha observado que es la mejor manera de visualizar todas las variaciones en dicha variable.

En la siguiente ilustración se observa anualmente las atenciones realizadas por cada gestor. En el eje X están numerados los códigos de gestor en el mismo orden que en la anterior ilustración. (Debido al tipo de gráfica, los códigos de gestor del eje X no se han podido representar de manera visible para el lector).

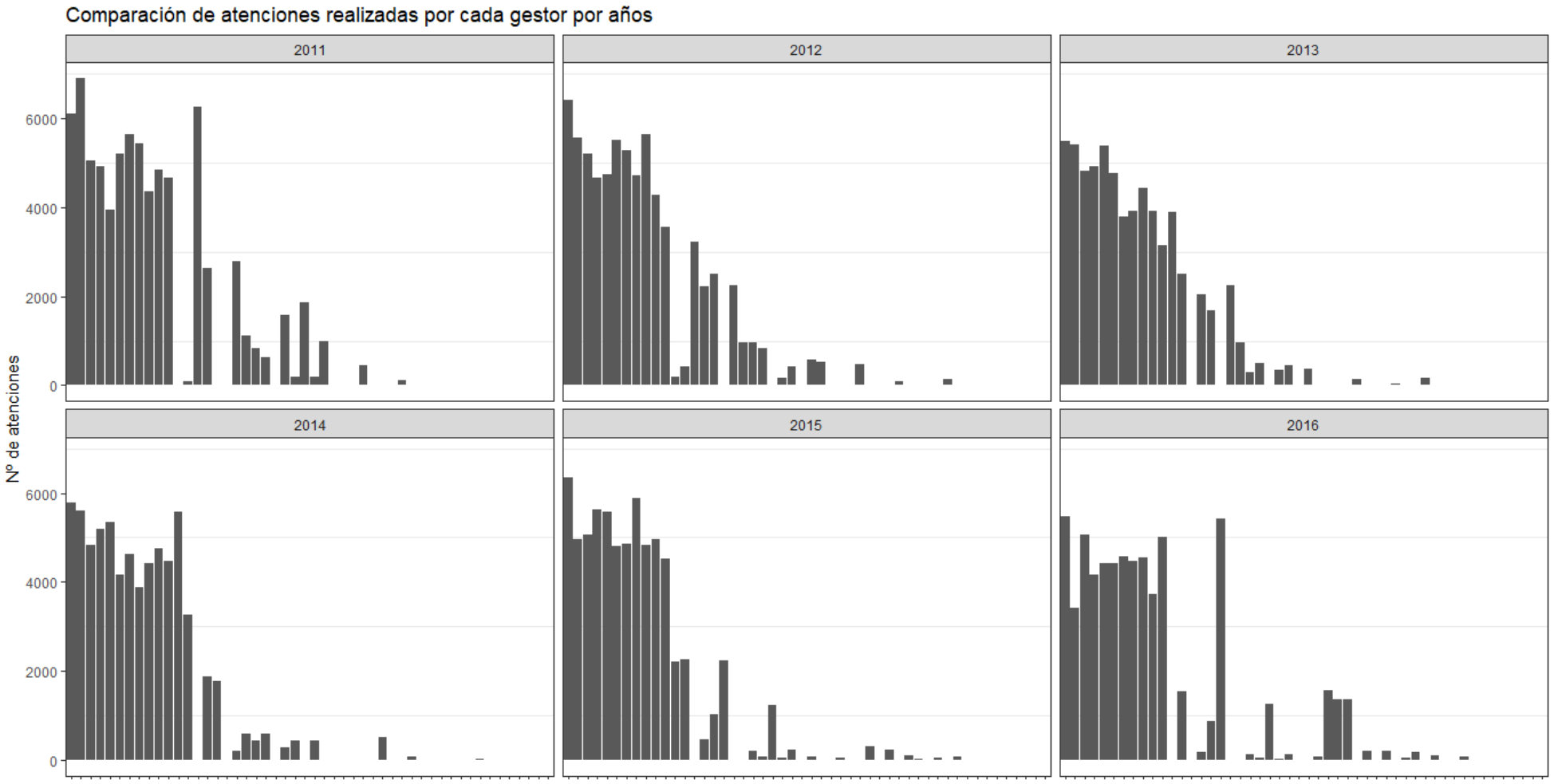


Ilustración 42 Número de atenciones que realiza cada gestor anualmente (orden del eje X igual que gráfica anterior)

Número de personas que atiende cada gestor diferenciando por colores cada año:

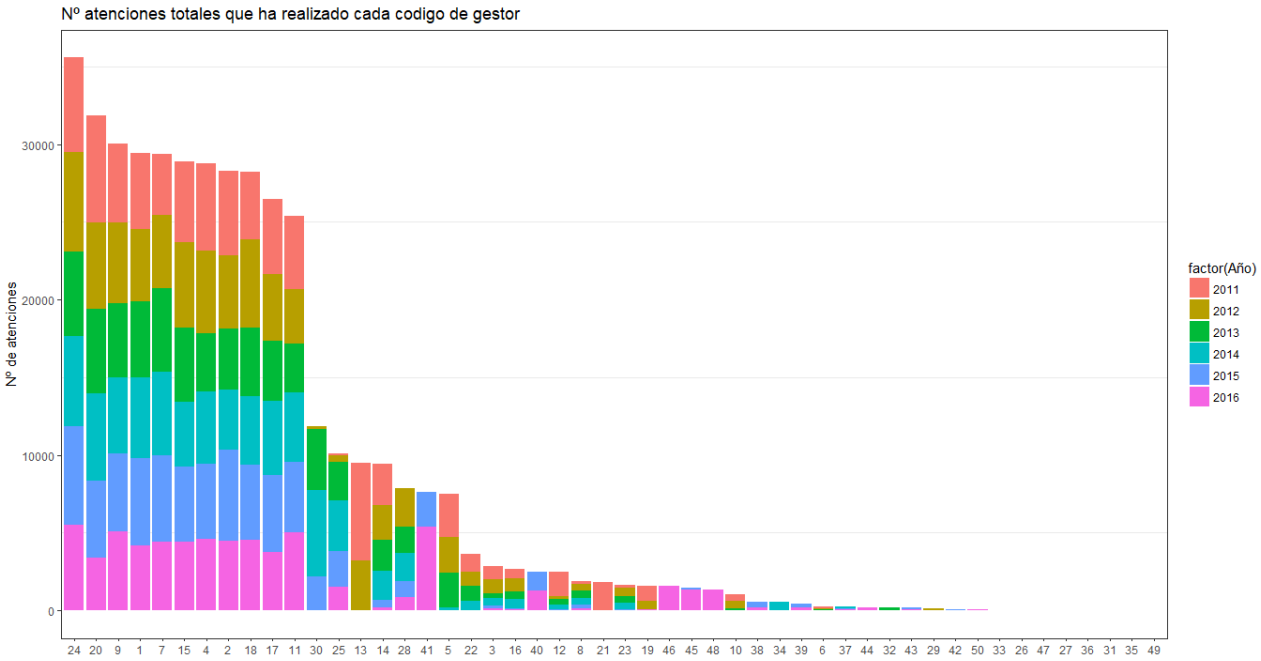


Ilustración 43 Número de atenciones que realiza cada gestor anualmente

Se observa que hay once gestores que han realizado un mayor número de atenciones que el resto de los gestores, esto puede ser debido a que estos once puedan estar destinados a realizar los trámites más rápidos de solucionar y que haya otros gestores dedicados específicamente a trámites más largos como por ejemplo los relacionados con licencias de obras.

Número de personas que atiende cada gestor diferenciando por colores cada mes:

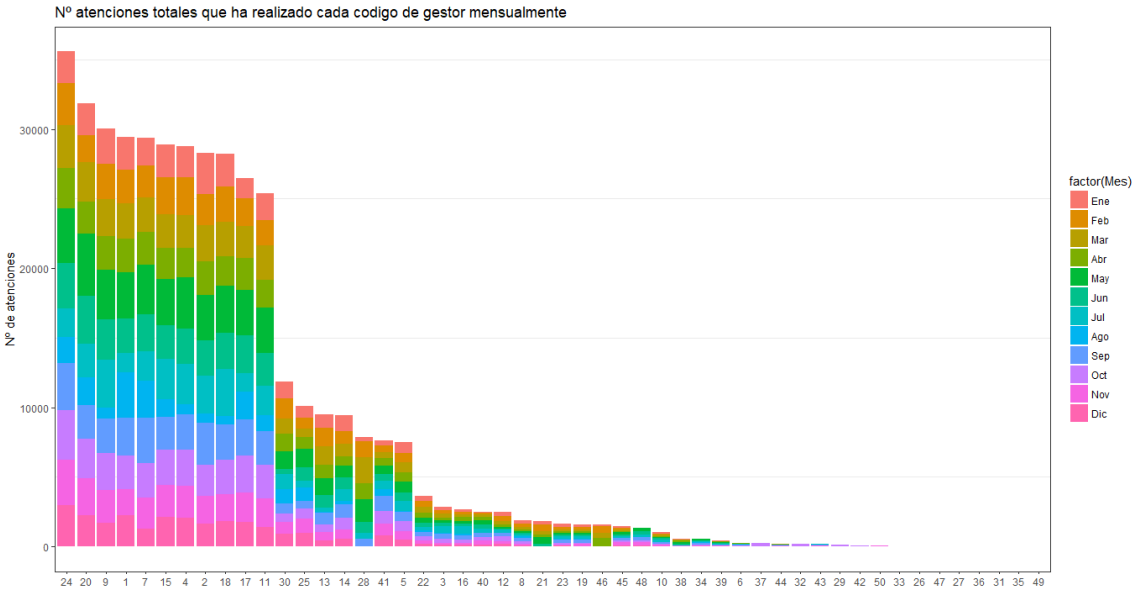


Ilustración 44 Gráfica Análisis temporal del número de atenciones realizadas por cada gestor mensualmente

No haremos el histograma de gestores diferenciado sus atenciones por cada día de mes, ya que, al haber 50 gestores y 31 días por mes, la gráfica no se lee correctamente y se ha pensado que es mejor diferenciar solamente por día de semana.

Número de personas que atiende cada gestor diferenciando por colores cada día de la semana:

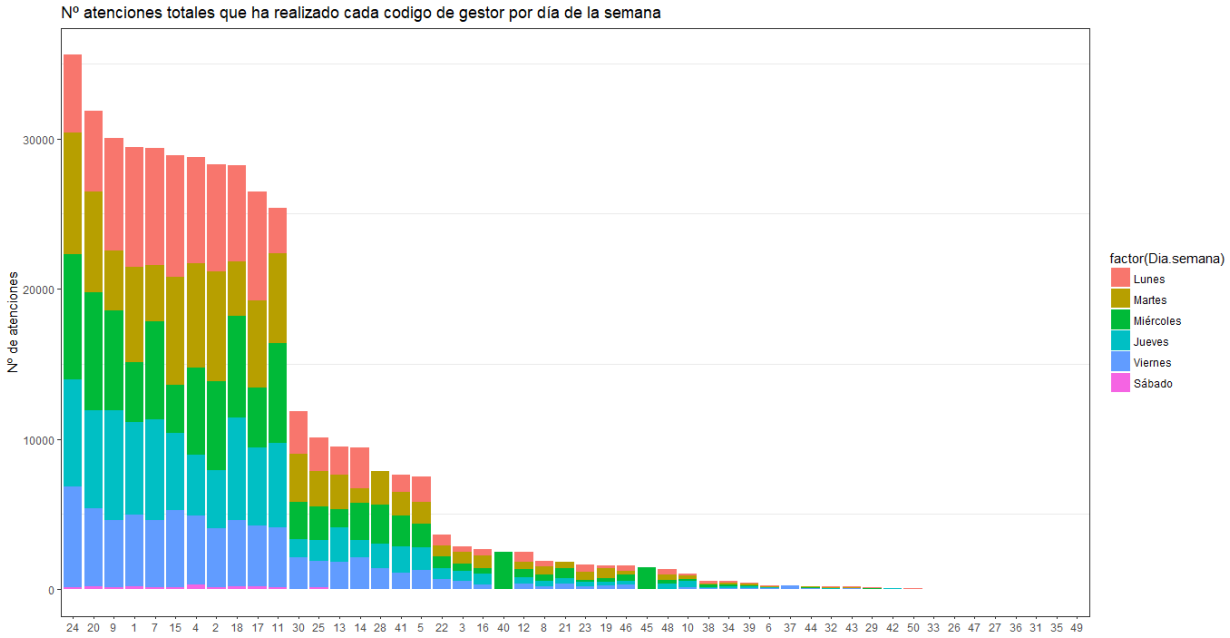


Ilustración 45 Gráfica Análisis temporal del número de atenciones realizadas por cada gestor por día de semana

Análisis temporal de la variable: Tiempo de servicio

En la base de datos se disponía de una columna llamada "Tiempo Servicio", en la que se había calculado el tiempo de servicio que necesitaba cada usuario hasta finalizar su atención.

Para trabajar en R con formato horario (HMS), se ha creado una nueva columna en la base de datos con la que se trabaja en R, llamada "Tiempo servicio 2" en la cual se han introducido los valores de la columna "Tiempo servicio", anteriormente mencionada, en formato HMS, formato identificado por R y que permite realizar operaciones y gráficas cómodamente.

Una vez se tiene la variable "Tiempo de servicio 2" en formato HMS, transformamos dicho tiempo de servicio de segundos a minutos. Primeramente se saca la información como duración (as.duration) para que esté en segundos, luego dejamos solo los valores numéricos de los segundos (as.numeric) y finalmente dividimos entre 60 y obtenemos el tiempo de servicio en minutos. Estos minutos se guardarán en una nueva variable que hemos creado y que se llama "Tiempo de servicio 3":

```
Data$Tiempo.servicio2 <- hms(Data$Tiempo.servicio)
Data$Tiempo.servicio2
Data$Tiempo.servicio3<
as.numeric(as.duration(Data$Tiempo.servicio2))/60
```

Hecho esto, se ha comenzado a representar la evolución temporal de la variable Tiempo de servicio, al igual que se hizo en apartados anteriores con la variable "Número de atenciones".

Para caracterizar esta variable lo haremos mediante Boxplot (o gráficos de cajas), con la función:

```
geom_boxplot
```

ya que se ha observado que es la mejor manera de visualizar todas las variaciones en dicha variable y observar también así la presencia de outliers.

Análisis temporal de la variable tiempo de espera: anualmente

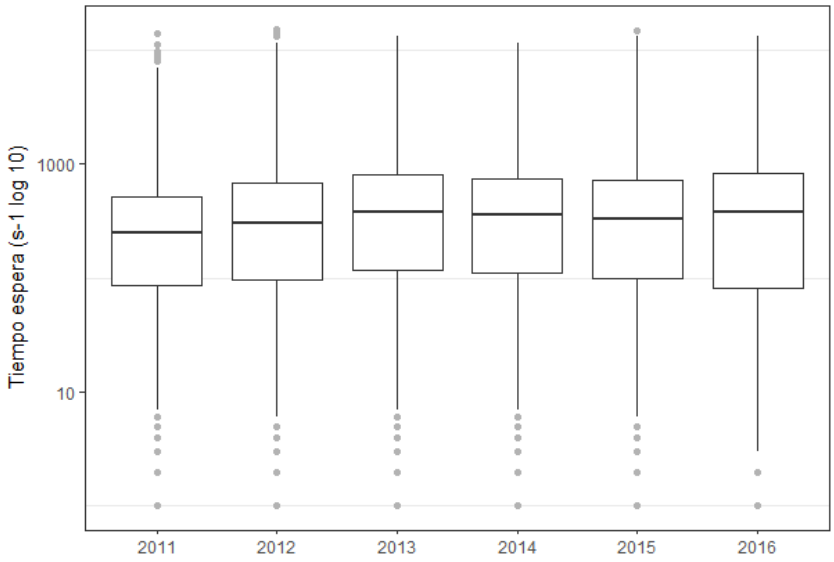


Ilustración 46 Gráfica variación tiempo servicio anualmente (Eje Y logarítmico)

Analizando el tiempo de servicio anualmente se observa que pese a que desde 2013 disminuía hasta 2015, en 2016 ha vuelto a aumentar. Esto tiene sentido relacionarlo directamente con el tiempo de espera: previamente se vio que el tiempo de espera en 2016 había aumentado considerablemente con respecto al resto de años analizados, por lo que esto puede haber sido producido directamente por el aumento en el tiempo de servicio de las atenciones.

Es conveniente por tanto buscar y estudiar de manera paralela a la realización de este TFM las causas que han llevado a que en 2016 se haya producido un aumento de la mediana del tiempo de servicio.

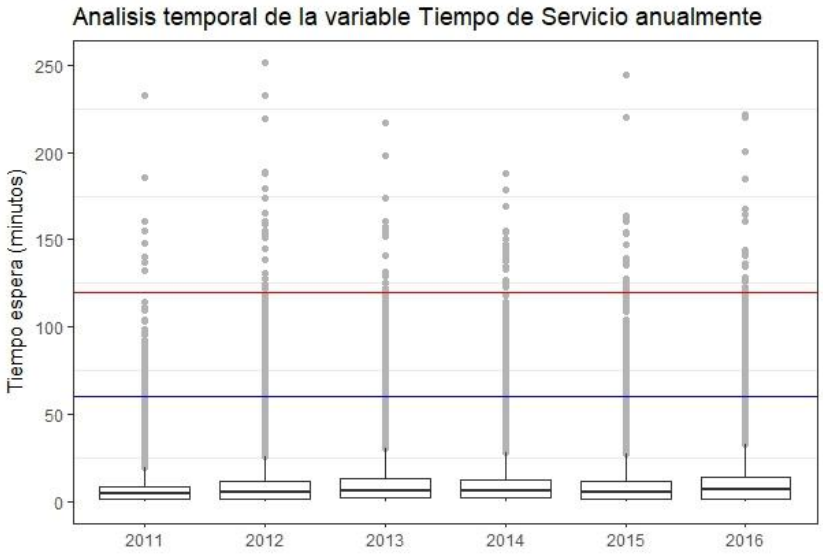


Ilustración 47 Gráfica variación tiempo servicio anualmente (Eje Y minutos)

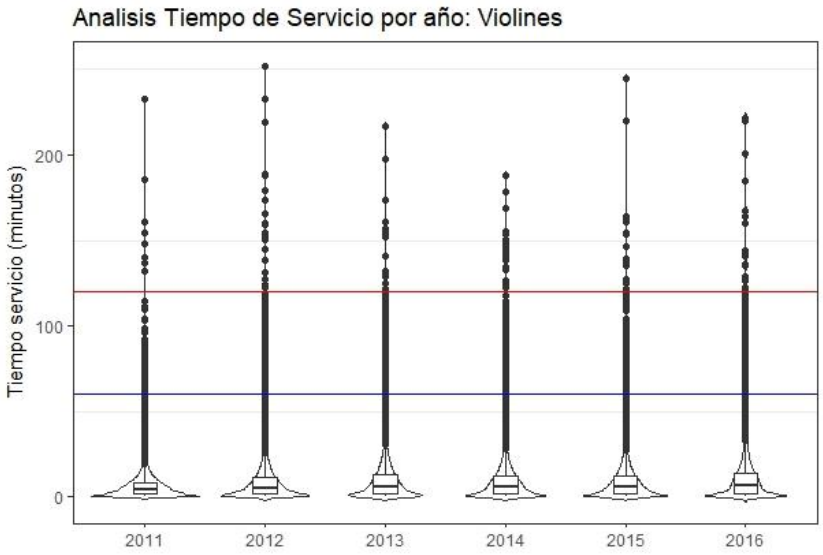


Ilustración 48 Gráfica violines variación tiempo servicio anualmente (Eje Y minutos)

Análisis temporal de la variable tiempo de espera: mensualmente

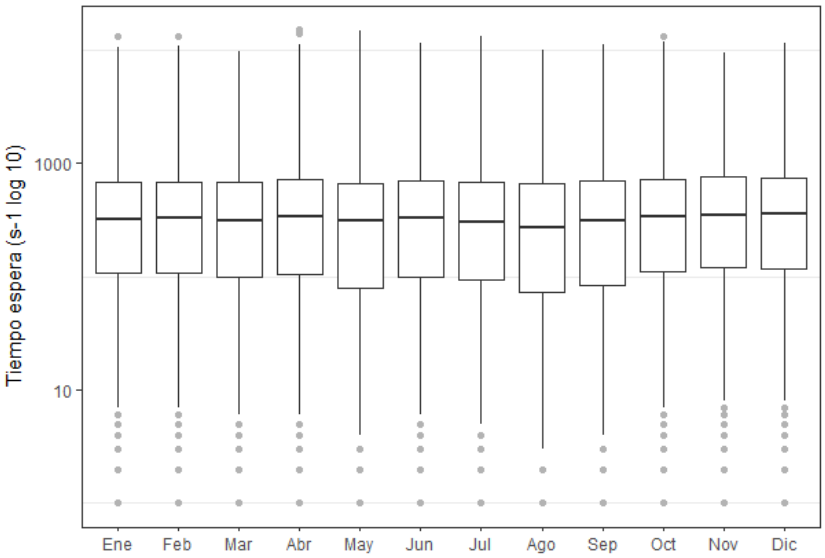


Ilustración 49 Gráfica variación tiempo servicio mensualmente (Eje Y logarítmico)

El tiempo de servicio es ligeramente inferior en agosto que en el resto de los meses analizados.

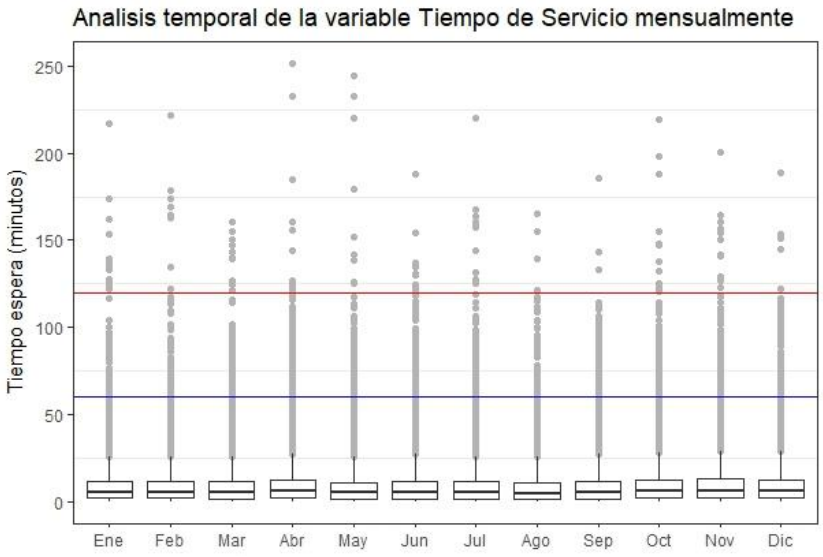


Ilustración 50 Gráfica variación tiempo servicio mensualmente (Eje Y minutos)

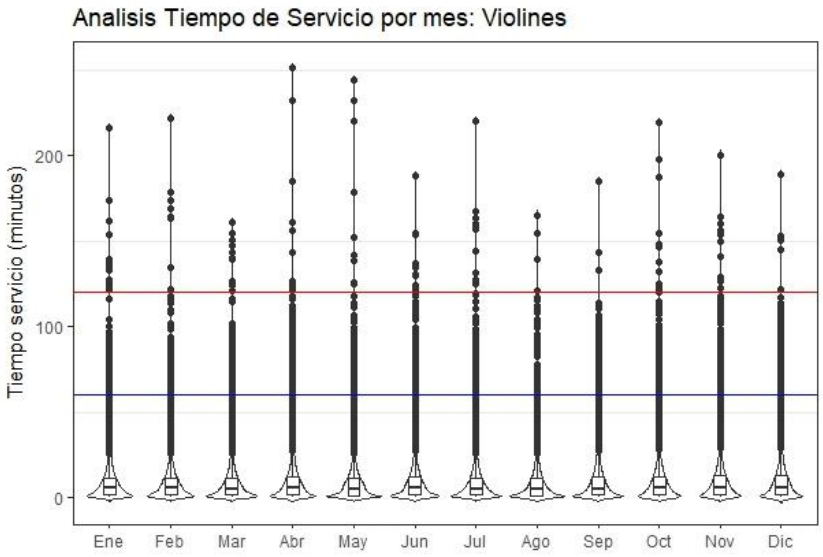


Ilustración 51 Gráfica violines variación tiempo servicio mensualmente (Eje Y minutos)

Análisis temporal de la variable tiempo de espera: día de mes

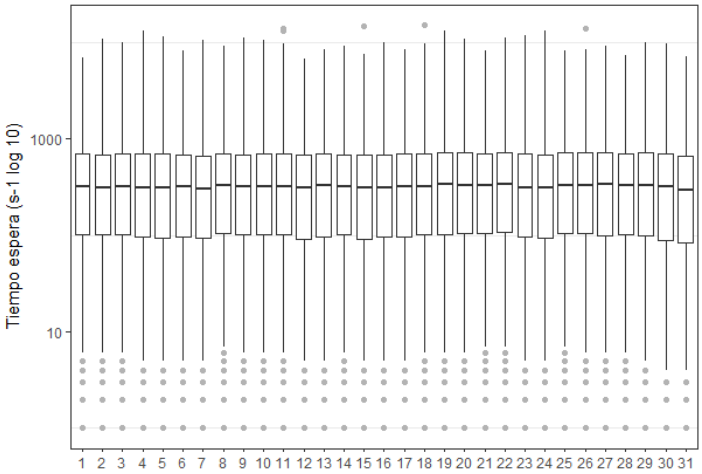


Ilustración 52 Gráfica variación tiempo servicio por día de mes (Eje Y logarítmico)

No se aprecia a simple vista ninguna tendencia del tiempo de servicio con respecto al día de mes.

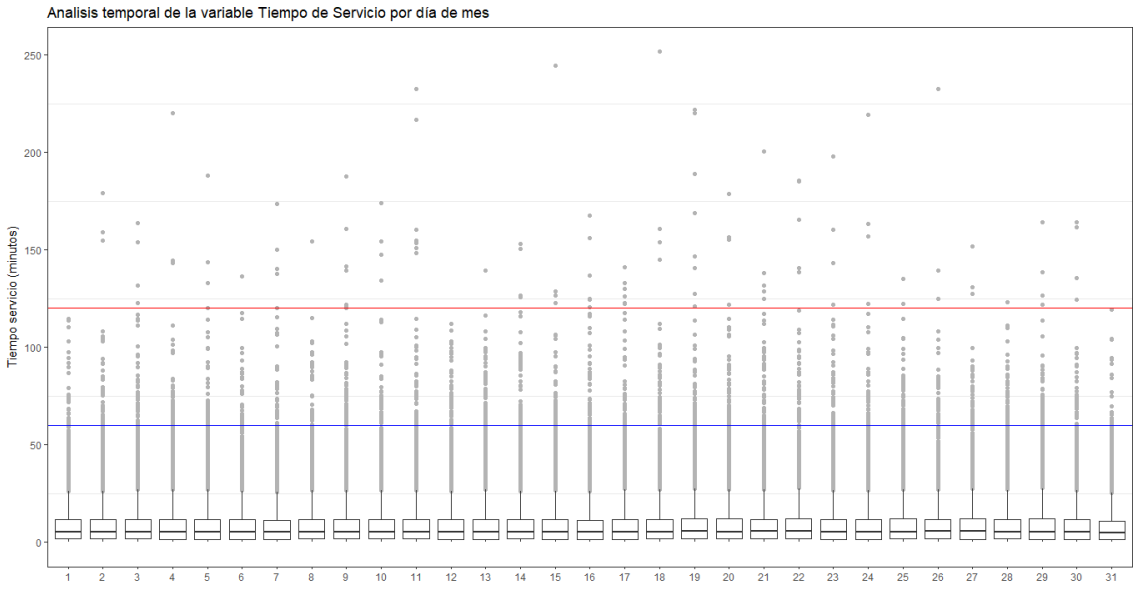


Ilustración 53 Gráfica variación tiempo servicio por día de mes (Eje Y minutos)

Análisis temporal de la variable tiempo de servicio con respecto al gestor que atiende dicha atención:

Se ha creído conveniente realizar este estudio de la relación entre ambas variables ya que es previsible que código de gestor si influya en el tiempo de servicio.

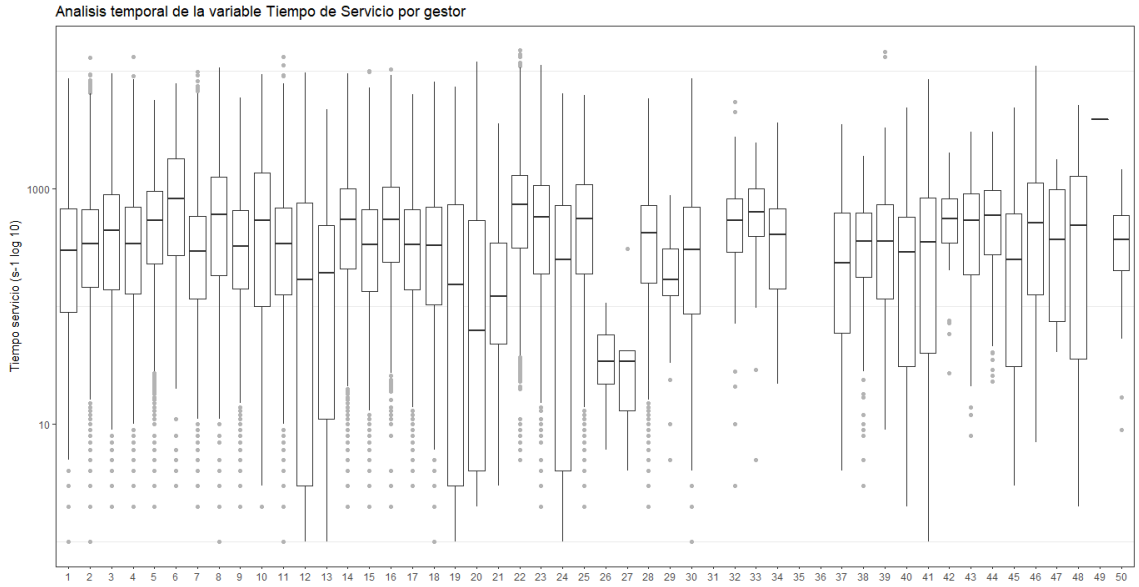


Ilustración 54 Gráfica variación tiempo servicio para cada gestor (Eje Y logarítmico)

Como era previsible hay diferencias muy notables (se observan mejor en la ilustración anterior con la escala del eje y logarítmica) entre los diferentes gestores y el tiempo de servicio.

El gestor 26 y 27 tienes medianas de sus tiempos de servicio muy inferiores al del resto de gestores, quizás sea debido a que ambos gestores estén ocupados de trámites que llevan muy poco tiempo de gestionar y solucionar.

Por otra parte es importante destacar la diferencia tan abrupta que hay entre máximos y mínimos de algunos gestores, por ejemplo, el 12, 13 o 24.

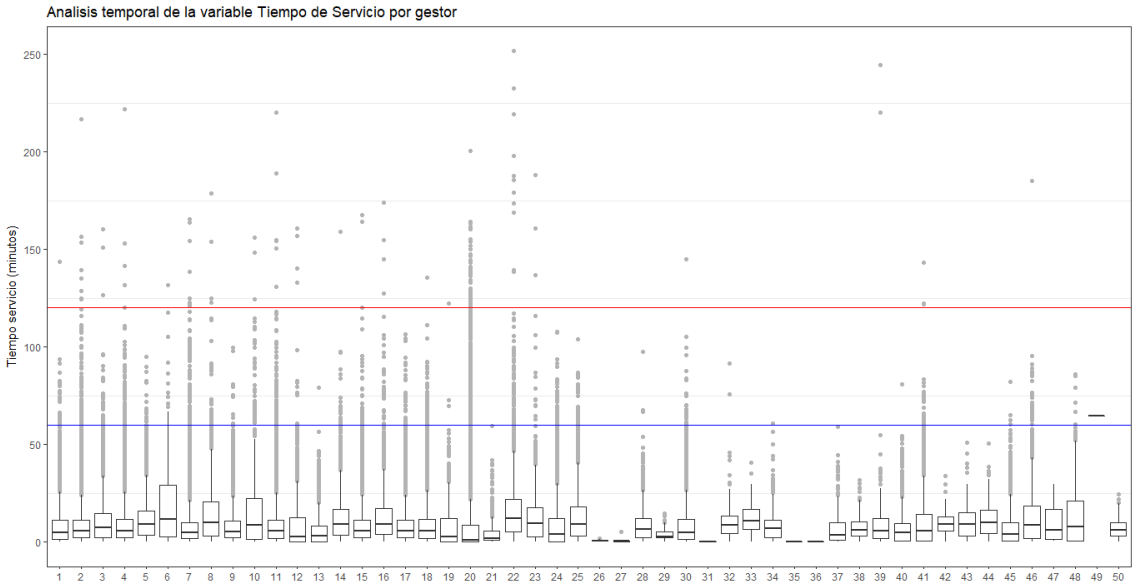


Ilustración 55 Gráfica variación tiempo servicio para cada gestor (Eje Y minutos)

Se observa que el tiempo de servicio suele ser corto para la mayoría de gestores, pero lo que se ve es que hay muchos servicios excesivamente largos (visualizados como outliers).

Análisis temporal de la variable tiempo de espera: día de semana

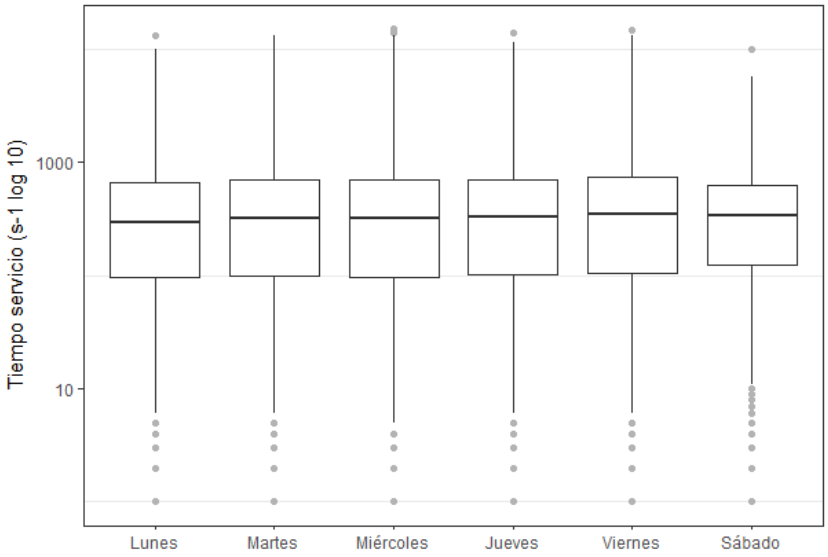


Ilustración 56 Gráfica variación tiempo servicio por día de la semana (Eje Y logarítmico)

El tiempo de servicio no se ve afectado previsiblemente por el día de la semana, pero tiene una variabilidad muy grande.

```
summary(as.numeric(Data$Tiempo.servicio2))

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    #0.0    93.0   315.0   495.2   691.0 15108.0      21


summary(Data$Tiempo.servicio3)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 # 0.000   1.550   5.283   8.320  11.617 251.800      22
```

Si quitásemos al código anterior la escala logarítmica lo que se ve es que el tiempo de servicio suele ser corto, pero lo que se ve es que hay muchos servicios excesivamente largos (visualizados como outliers):

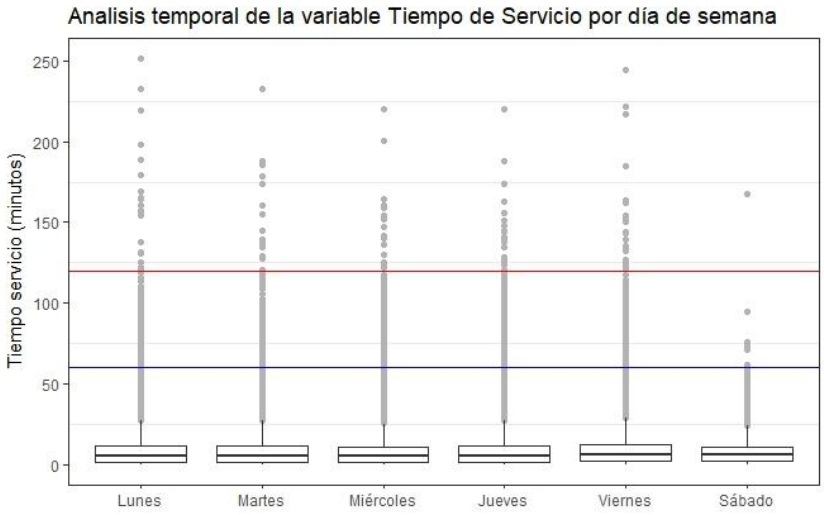


Ilustración 57 Gráfica variación tiempo servicio por día de la semana (Eje Y minutos)

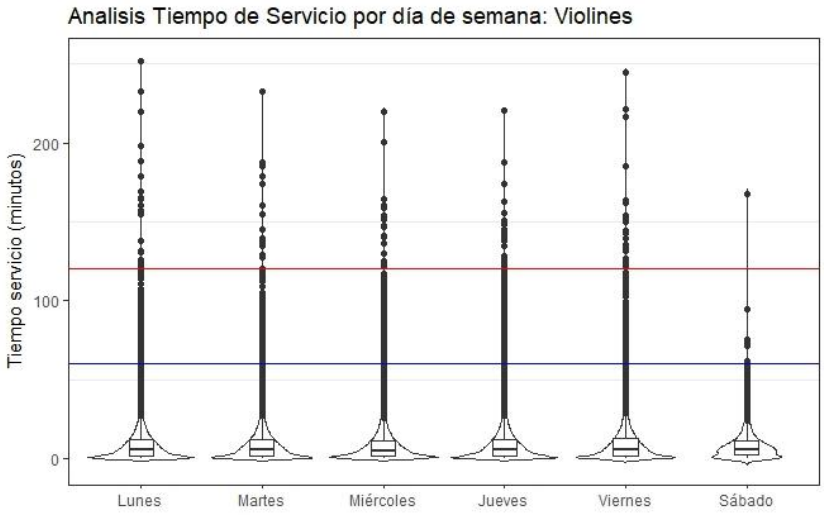


Ilustración 58 Gráfica violines variación tiempo servicio por día de la semana (Eje Y minutos)

Para comparar entre días de semana se necesita una escala que disminuya las observaciones anómalas, es por eso que se usa la escala logarítmica. En la siguiente ilustración se observa en un gráfico de violines y en escala logarítmica la variabilidad del tiempo de espera con respecto al día de la semana.

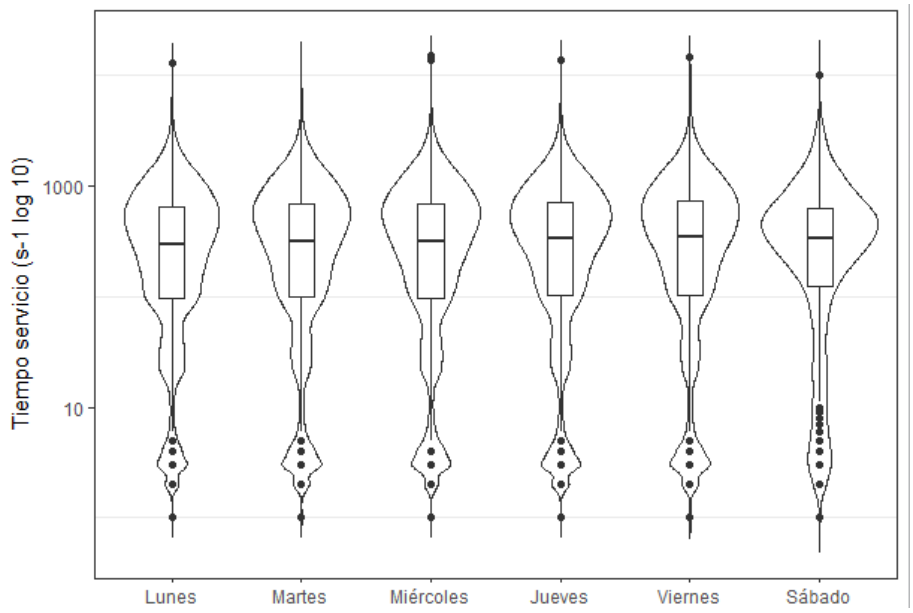


Ilustración 59 Gráfica violines variación tiempo servicio por día de la semana (Eje Y Logarítmico)

¿Qué significa esta gráfica? Hay varia información: la densidad de observaciones las da el grafico de violín (Kernell Density) y muestra que existe un sesgo positivo del tiempo de servicio para todos los días excepto el sábado. Este sesgo implica que existe mayor número de atenciones con más tiempo de servicio que la mediana, excepto el sábado que el sesgo es menor. (Sesgo es desviación).

Se observa también que hay abundantes tiempos de servicio cercanos al cero, que podría estar representando fallas en la atención. (Y que están influenciando la densidad de los violines).

Para finalizar esta sección se modelizarán las variables "tiempo de espera" y "tiempo de servicio"

Modelizar la variable tiempo de espera:

```
summary(Data$Tiempo.espera2)
                  Min.                1st Qu.                 Median
                  "0S"                  "29S"               "7M 3S"
                  Mean                3rd Qu.                   Max.
"12M 35.977286562259S"               "18M 45S"           "3H 56M 52S"
                  NA's
                   "1"
```

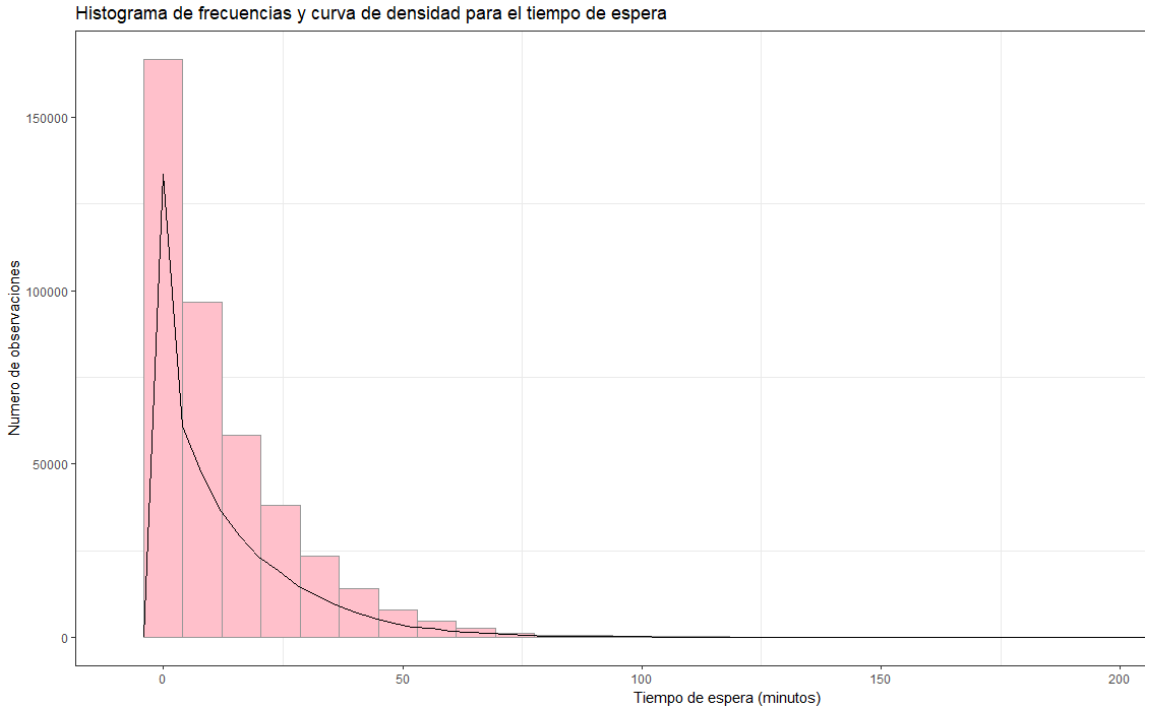


Ilustración 60 Histograma de frecuencias y curva de densidad para el tiempo de espera

Esta gráfica representa el número de casos con diferentes tiempos de espera. La curva representa la misma información que representan las barras de manera continua.

Modelizar la variable tiempo de servicio:

```
summary(Data$Tiempo.servicio2)
                   Min.                 1st Qu.                  Median
                   "0S"              "1M 33S"               "5M 17S"
                   Mean                 3rd Qu.                    Max.
"8M 19.2062936726404S"              "11M 37S"              "4H 11M 48S"
                   NA's
                   "22"
```

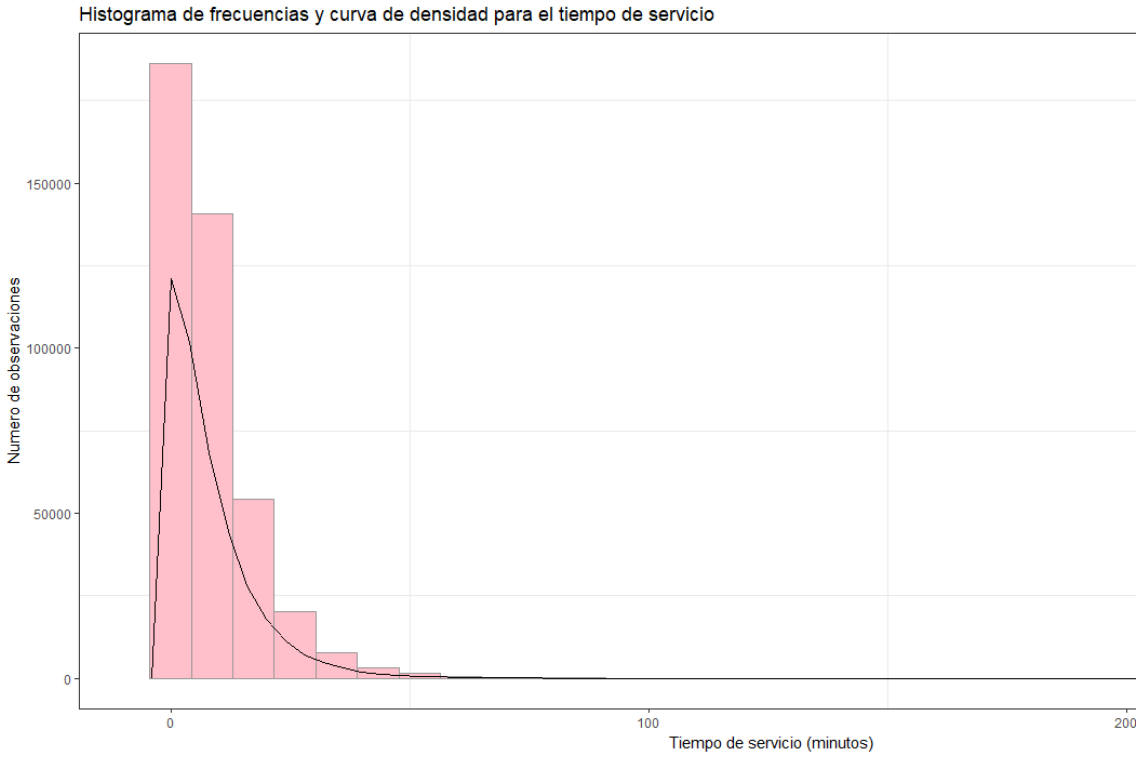


Ilustración 61 Histograma de frecuencias y curva de densidad para el tiempo de servicio

Esta gráfica representa el número de casos con diferentes tiempos de servicio. La curva representa la misma información que representan las barras de manera continua.

```
#-----------------------------------------------------#
#Fin primera parte: Análisis exploratorio de los datos.
#-----------------------------------------------------#
```

Objetivo 2: Análisis estadístico

Como último paso para cumplir el objetivo general, se necesita la realización del objetivo 2:

2. **Optimización del funcionamiento de la oficina de atención al cliente. Relación y dependencias existentes entre las variables del modelo estudiado en este TFM.**

Como se detalló a principio del capítulo seis en el que nos encontramos Dentro del objetivo 2 sobre la optimización de las oficinas de atención al cliente, se han definido los siguientes objetivos específicos con el fin de llegar a la construcción del modelo matemático:

- Modelizar la variable tiempo de espera (teoría de colas)
- Identificar las variables explicativas y la variable respuesta. La variable respuesta en este caso será: el tiempo de espera.
- Análisis exploratorio de la relación entre las variables explicativas y la de respuesta a diferentes escalas temporales (patrón diario, semanal, mensual, anual).
- Construcción del modelo matemático que define la variable respuesta: Tiempo de espera (regresión lineal, y=a +bx)

A continuación se procede a la realización en R los objetivos particulares anteriores.

Lo primero que se debe hacer es abrir un nuevo script en el cuál se programará todo lo relacionado con el análisis estadístico al cual nos referimos en este segundo objetivo.

Previo a la realización de este segundo objetivo se ha hecho un estudio de la base de datos facilitada por el Ayuntamiento de Sant Cugat, al igual que se hizo en el script del objetivo número uno. (Capítulo 6, sección: Pasos comunes tanto para el objetivo uno como para el objetivo dos).

Distribución de la variable respuesta "Tiempo de espera" & Outliers

Una vez finalizados los pasos anteriores, se procede a representar la distribución de los datos de la variable respuesta: Tiempo de espera, mediante un histograma y a buscar la presencia de posibles outliers mediante un boxplot.

```
#Instalamos paquete lubridate
Data$Tiempo.espera2 <- hms(Data$Tiempo.espera)
Data$Tiempo.espera2
Data$Tiempo.espera3 <- as.numeric(as.duration(Data$Tiempo.espera2))/60
#Tiempo de espera 3 : minutos
```

Representamos en el siguiente histograma la distribución de los datos de la variable tiempo de espera. En el eje Y se encuentran el número de atenciones y en el eje X los minutos del tiempo de espera para cada atención.

```
#Representamos el Histograma de tiempo de espera. #Grafica52 (Script
anterior) (Eje X en minutos con Tiempo de espera 3)
ggplot(Data,aes(x=Tiempo.espera3))+
  geom_histogram(fill="pink", colour="grey60", size=0.2, na.rm = TRUE)+
  geom_freqpoly(binwidth=4)+
  labs(x="Tiempo de espera (minutos)", y="Número de observaciones",
title="Histograma de frecuencias y curva de densidad para el tiempo de
espera")+
  theme_bw()+
  theme(
    #axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())
```

El histograma obtenido es el siguiente:

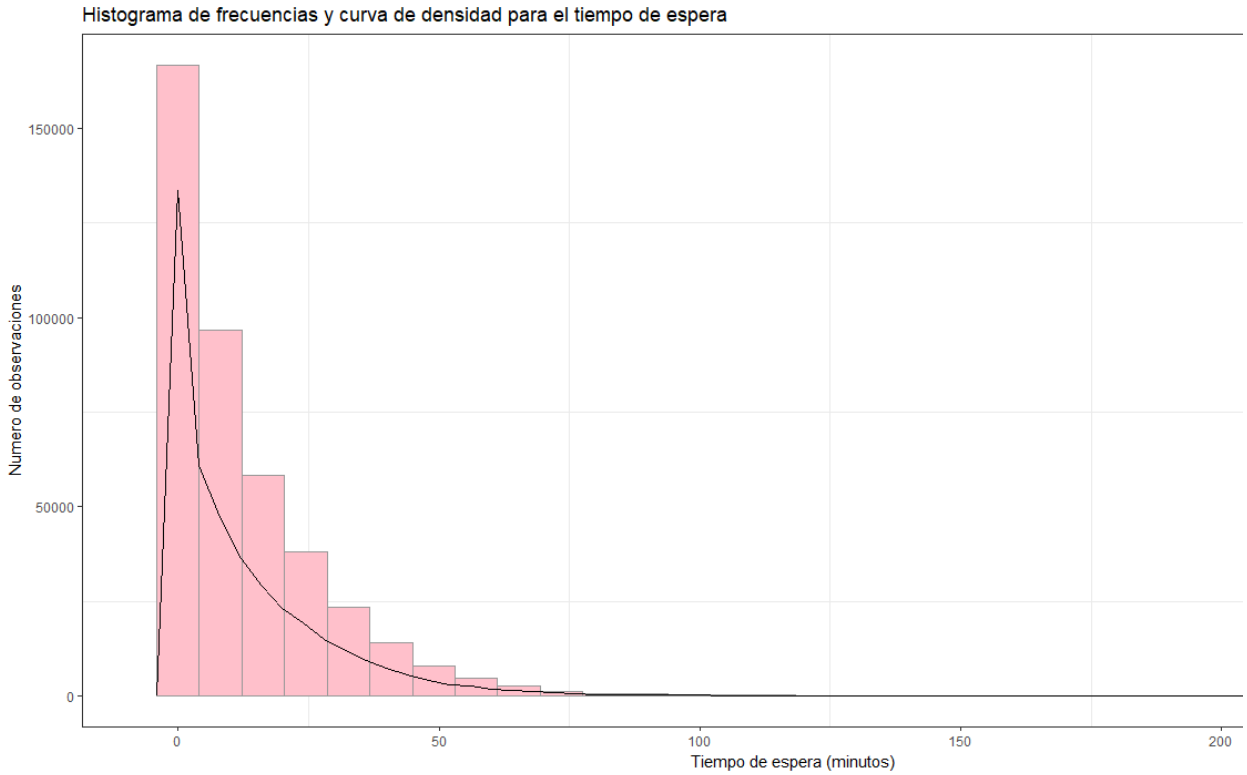


Ilustración 62 Histograma de frecuencias y curva de densidad para el tiempo de espera 3 (minutos)

La respuesta que obtenemos para la distribución de esta variable es de Poisson (binomial negativa). Se observan en este histograma muchas atenciones con tiempos de espera muy cercanos o iguales a cero minutos. Realizamos la siguiente comprobación:

```
> summary(Data$Tiempo.espera3)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
  0.0000   0.4833   7.0500  12.5996  18.7500 236.8667        1
```

Efectivamente se comprueba que el mínimo dato en el vector tiempo de espera 3 (en minutos) es “0 minutos”. Dado este suceso, comprobamos en un boxplot, la presencia de numerosos outliers en torno al 0 (lo que implica cero minutos de espera). Dado que es preferible no tener en cuanta aquellas observaciones con tiempos de espera iguales a cero, vamos a realizar la siguiente operación:

crear una nueva variable llamada “Tiempo de espera4”, en el cual no estén incluidos los tiempos de espera iguales a cero. Para ello se asigna el valor NA a aquellos valores que eran iguales a cero.

```
Data$Tiempo.espera4 <- Data$Tiempo.espera3
Data$Tiempo.espera4[Data$Tiempo.espera4==0] <-NA
```

```
> summary(Data$Tiempo.espera4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.02    2.30    9.02   14.18   20.73  236.87   46217
```

Se comprueba que efectivamente en el nuevo vector “Tiempo de espera 4” no hay tiempos de espera iguales a cero, el mínimo tiempo de espera recogido en este vector es de 0,02 minutos.

Seguidamente representaremos el mismo histograma anterior (ilustración 61) pero en el eje X estará el vector "tiempo de espera 4":

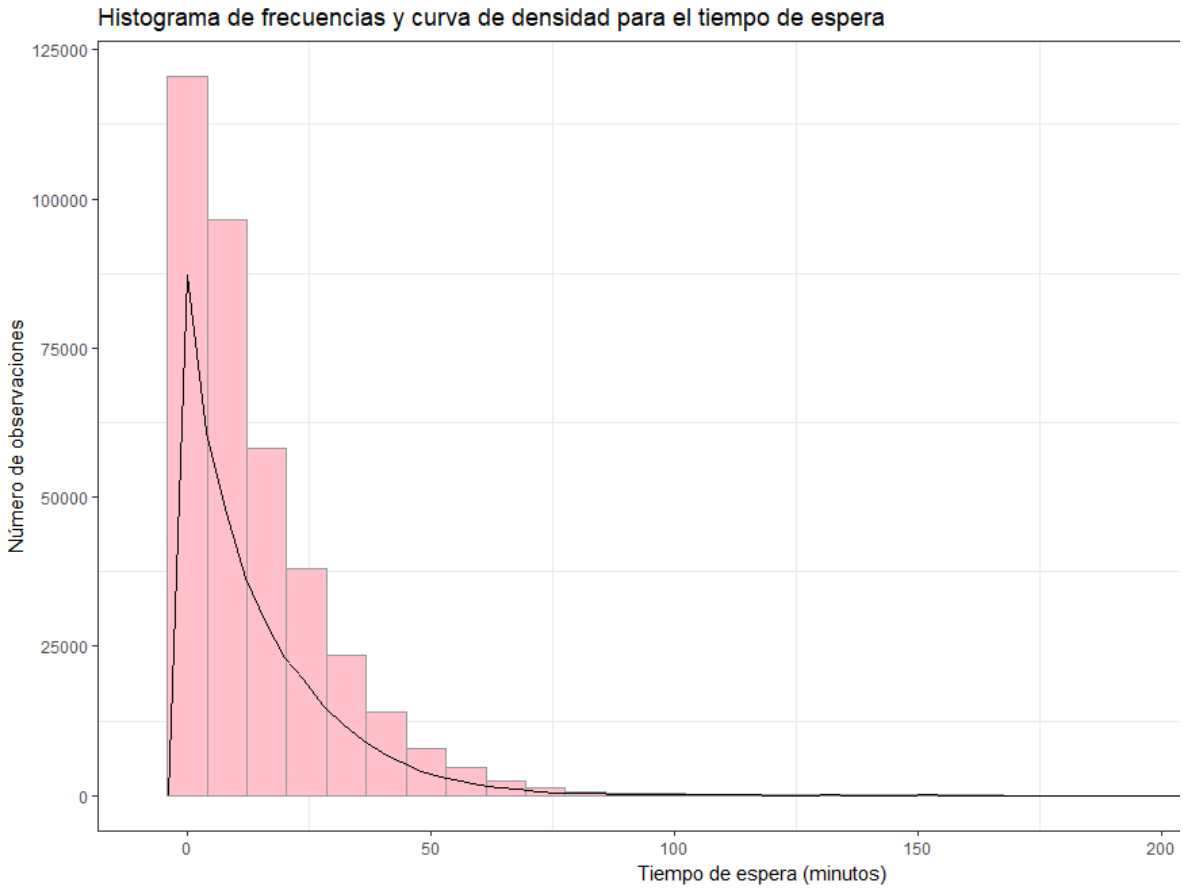


Ilustración 63 Histograma de frecuencias y curva de densidad para el tiempo de espera 4 (minutos, sin 0)

Se observa que en la ilustración anterior, con el eje X sin tiempos de espera iguales a cero, que las atenciones con tiempo de espera cercanas a cero minutos, no llegan a las 125.000 observaciones, en cambio en la ilustración 61 (con eje X tiempo de espera 3) hay más de 150.000 atenciones con tiempos de espera iguales o cercanos a cero.

Outliers: dada la presencia de números outliers (observaciones anómalas) en la variable respuesta, recordemos que la **variable respuesta** en este TFM es la variable "**Tiempo de espera**", se procederá a estudiar y analizar que hacer o como explicar dichos outliers para que no distorsionen el resultado final.

Representamos mediante un boxplot la variable tiempo de espera, primero con el eje Y en minutos (tiempo de espera 4 en minutos y sin ceros); y seguidamente con el eje Y en escala logarítmica, (log (tiempo de espera 4 en minutos y sin ceros)) y vemos como trataremos dichos outliers.

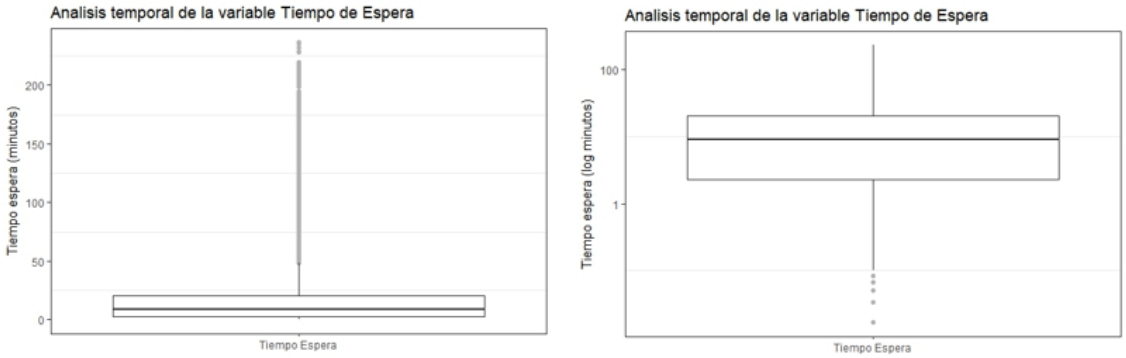


Ilustración 64 ***Izquierda****: Boxplot de la variable Tiempo de espera 4. Eje Y en minutos (sin ceros en el vector).*
Derecha*: Boxplot de la variable tiempo de espera 4. Eje Y logaritmo del tiempo de espera 4 (minutos, sin ceros).*

En ambas gráficas se aprecia que pese a que la mediana del tiempo de espera es inferior a los 25 minutos, existen numerosas observaciones con tiempos de espera excesivamente altos, superando incluso las tres horas de espera.

La mediana es menos sensible a los datos extremos y por ese motivo es más significante para nosotros que la media.

El software R considera que aquellas atenciones con tiempos de espera superiores a cincuenta minutos son consideradas como outliers, (R considera outliers aquellas atenciones que superen dos veces la desviación estándar, es decir aproximadamente 50 minutos). A partir de aquí debemos tomar las consideraciones necesarias para determinar qué tiempos de espera consideraremos como outliers para este TFM.

Debido a la presencia de numerosas observaciones anómalas (outliers) en la ilustración 63, **izquierda**, se procede a transformar la variable respuesta utilizando el logaritmo en base 10 del tiempo de espera, para cambiar así la escala de variación y que los valores extremos (outliers) no tengan tanto peso en los modelos de regresión que se realizarán a continuación.

Si recordamos que la variable respuesta usada hasta ahora la denominábamos “Tiempo de espera 4”, expresada en minutos, sin valores iguales a cero y con un valor mínimo de 0,02 minutos a partir de ahora trabajaremos con una nueva variable denominada: “Tiempo de espera 5”, y será el logaritmo en base 10 de la variable Tiempo de respuesta 4. La nueva variable y columna se crea del siguiente modo:

```
Data$Tiempo.espera5 <- log10(Data$Tiempo.espera4)
```

De esta forma, y a partir de ahora, los modelos de regresión que estudiaremos se realizarán para la variable Tiempo de espera 5 (en escala logarítmica), esta decisión se ha llevado a cabo para asegurarnos que los valores extremos existentes en nuestros datos, no tengan tanto eso en los modelos de regresión

Relación entre variables y construcción modelo matemático para el tiempo de espera:

Identificar la variable respuesta y las variables explicativas de este TFM:

- Variable respuesta: Tiempo de espera (En R será la variable denominada “Tiempo de espera 5”. Esta se considera la variable respuesta ya que se ha visto que a la hora de que un cliente califique la atención recibida en una institución tanto pública como privada, el tiempo esperado es un factor clave a la hora de evaluar dicho servicio.

Dado que el objetivo de este proyecto es la optimización del servicio general de atención al cliente, se cree conveniente que sea en esta variable en la que nos enfoquemos y analicemos con el fin de que se pueda llegar a mejorar dichos tiempos de espera y que con ello que los usuarios del Ayuntamiento de Sant Cugat estén más satisfechos con dichas atenciones.

- La variable respuesta (Tiempo de espera) es predecible que dependa de otras variables que llamaremos explicativas y que concretamos en el siguiente punto. La relación entre ambas se analiza en la siguiente sección una a una.
- Variables explicativas: día de la semana, día de mes, año, mes, gestor que atiende. Todas las variables explicativas son categóricas (ordinales y nominales) por lo que vamos a asumir que no hay colinearidad.

Pairs: Análisis exploratorio de la relación entre las variables explicativas y la variable respuesta a diferente escala temporal. ¿Existen diferencias significativas?

Se van a representar las gráficas de los análisis temporales de la variable respuesta "Tiempo de espera" con cada una de las variables explicativas.

Debido a que el mayor número de observaciones de la base de datos corresponden a la oficina de “Registro/Atención Ciudadana”, los resultados que se obtengan en este capítulo van a representar en gran medida a las atenciones realizadas en esta oficina. Queda para un futuro trabajo, la exclusión de las atenciones de dicha oficina y el análisis por separado de las atenciones en el resto de oficinas existentes en el Ayuntamiento de Sant Cugat.

```
#-----------------------------------------------------#
# Tiempo de espera - Día de la semana
#-----------------------------------------------------#
```

A continuación se muestran dos ilustraciones que corresponden a un análisis temporal de la variable “tiempo de espera” por cada día de la semana mediante la función *boxplot* o diagrama de cajas: en la primera figura se muestra el eje y que corresponde al tiempo de espera en minutos; y en la segunda ilustración se muestra la misma gráfica pero con el eje y en escala logarítmica.

Para esta sección se ha decidido usar diagramas de cajas o boxplot ya que permiten conocer cómo se distribuyen los datos dentro de una variable.

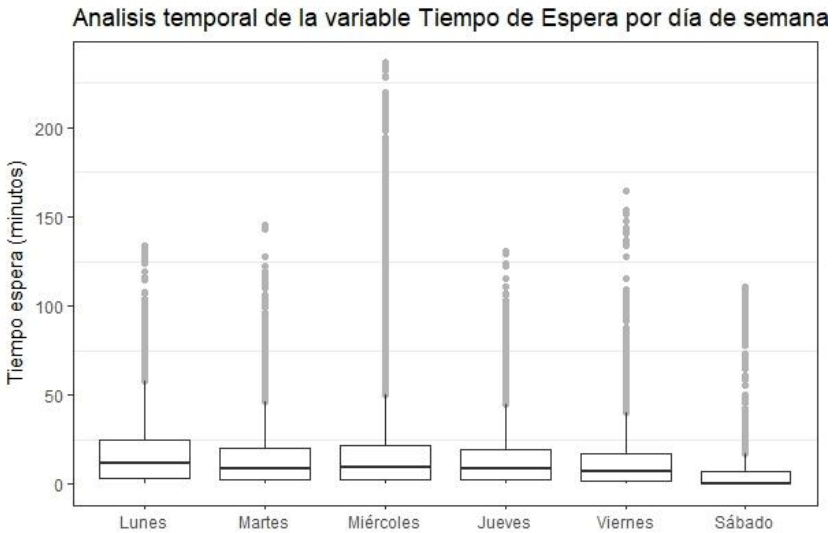


Ilustración 65 Análisis temporal de la variable Tiempo de Espera por día de semana (Eje Y: minutos)

De las ilustraciones que acompañan este análisis se puede deducir que el lunes es el día de la semana que tiene la mediana del tiempo de espera más alta. El sábado es el día de la semana con la mediana de tiempo de espera más baja.

Además la caja del sábado es mucho más alargada que el resto de días de la semana, lo que significa que los sábados hay mucha más dispersión de observaciones.

Puesto que la mediana la no tiende a estar en el centro de la caja, podemos afirmar que la distribución de la variable tiempo de espera es asimétrica positiva durante los cinco primeros días de la semana y asimétrica negativa los sábados.

Con este diagrama no se puede deducir si hay o no diferencias significativas entre los diferentes días de la semana. Por este motivo se procede a realizar a continuación el test estadístico de la variable tiempo de espera en función del día de la semana para conocer entre que días existen diferencias significativas.

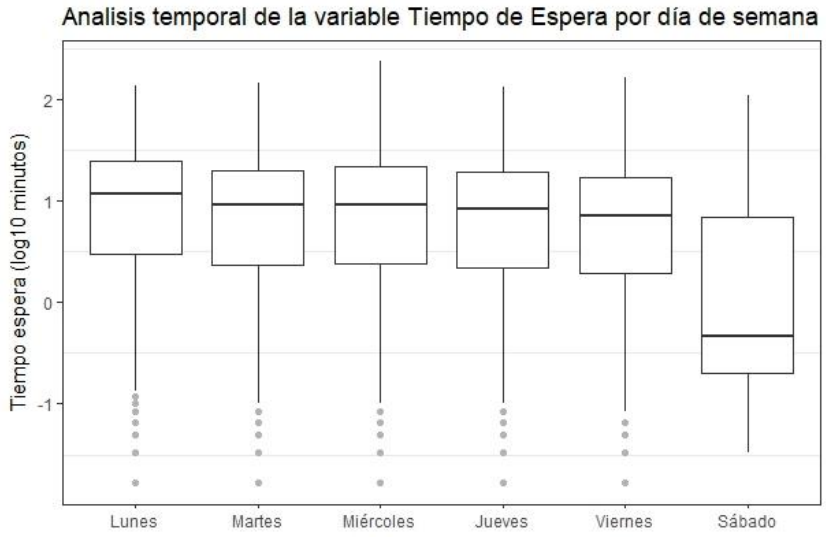


Ilustración 66 Análisis temporal de la variable Tiempo de Espera por día de semana (Eje Y: Log(minutos))

A continuación se muestra parte del código en R que se programó para llevar a cabo esta sección de análisis estadístico y también se han pegado los resultados obtenidos de la propia consola de RStudio al ejecutar dicho código. En caso de que se desee consultar el código de R al completo, está disponible al final de esta memoria en el anexo B.

```
#----------#
```

Todos los test estadísticos que realicemos para cada una de las variables explicativas se realizarán con la variable respuesta que llamaremos en R: ”Tiempo de espera 5” y que hace referencia al logaritmo en base 10 del tiempo de espera medido en minutos.

```
#Modelo1_DiaSemana<- glm(Tiempo.espera5~Dia.semana, family=gaussian,
data=Data)
#Hemos probado a realizar una primera prueba con familia de poisson pero
no salió correctamente porque al usar el logaritmo de la variable tiempo
respuesta: salen negativos, y Possion es sin negativos
#El test anterior es un GLM con familia gaussiana (normal) que es lo
mismo que un LM sin indicar la familia
```

```
Modelo1_DiaSemana<-lm(Tiempo.espera5~Dia.semana, data=Data) #LM hace el
modelo lineal: es decir compara cada día de la semana con un valor por
defecto: este resultado lo entrega el summary del modelo lineal.
(lunes= primer nivel de la variable). Pero si quisiéramos obtener solo
el Análisis de la varianza: se utiliza anova de ese modelo: así
observaremos si el efecto de la variable día de semana es importante o
no sobre el Tiempo de espera.
```

El test anova muestra un Pvalor muy pequeño cercano a cero (se descarta la hipótesis nula) por lo que se demuestra que existe un efecto del día de la semana sobre el tiempo de espera.

```
> anova(Modelo1_DiaSemana) #Mostramos el resultado como una tabla de
anova
Analysis of Variance Table

Response: Tiempo.espera5
               Df Sum Sq Mean Sq F value    Pr(>F)
Dia.semana      5   1812  362.31  616.15 < 2.2e-16 ***
Residuals  369391 217210    0.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> summary(Modelo1_DiaSemana)

Call:
lm(formula = Tiempo.espera5 ~ Dia.semana, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6036 -0.3731  0.2154  0.5716  1.9752

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.825441   0.002685  307.37   <2e-16 ***
Dia.semanaMartes   -0.099247   0.003862  -25.70   <2e-16 ***
Dia.semanaMiércoles -0.072323   0.003829  -18.89   <2e-16 ***
Dia.semanaJueves   -0.110810   0.003930  -28.20   <2e-16 ***
Dia.semanaViernes  -0.156274   0.004139  -37.76   <2e-16 ***
Dia.semanaSábado   -0.756582   0.017907  -42.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7668 on 369391 degrees of freedom
  (46217 observations deleted due to missingness)
Multiple R-squared:  0.008271,      Adjusted R-squared:  0.008258
F-statistic: 616.1 on 5 and 369391 DF,  p-value: < 2.2e-16
```

Este Análisis de anova es suficiente para demostrar que hay diferencia entre los días de la semana y la variable respuesta. Por lo que se concluye que el tiempo de espera no es igual para todos los días de la semana

Dado que el test a posteriori (test de Tukey) solo reconoce el modelo aov (Análisis de la varianza). Haremos dicho modelo y a continuación aplicaremos el test de Tukey a dicho modelo para así obtener un resultado de la comparación de todos los días de la semana entre ellos y conocer si existen diferencias significativas entre alguno de ellos.

```
Modelo1_DiaSemana_2<-aov(Tiempo.espera5~Dia.semana, data=Data)
```

```
> summary(Modelo1_DiaSemana_2)
                Df Sum Sq Mean Sq F value Pr(>F)
Dia.semana       5   1812   362.3   616.1 <2e-16 ***
Residuals   369391 217210     0.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
46217 observations deleted due to missingness
```

En el código anterior, se realiza el resumen del modelo aov, en el cual se puede comprobar al ser un Pvalor cercano a cero que el día de la semana si influye en el tiempo de espera. Ahora se aplica el test de Tukey (con la función *TukeyHSD*) a dicho modelo (*Modelo1_DiaSemana_2*), y seguidamente mostramos el resultado obtenido.

```
posthoc_Diasemana <- TukeyHSD(x=Modelo1_DiaSemana_2, 'Dia.semana',
conf.level=0.95)

posthoc_Diasemana
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Tiempo.espera5 ~ Dia.semana, data = Data)

$Dia.semana
                         diff         lwr           upr     p adj
Martes-Lunes       -0.09924738 -0.11025270 -0.0882420564 0.0000000
Miércoles-Lunes    -0.07232329 -0.08323595 -0.0614106337 0.0000000
Jueves-Lunes       -0.11080981 -0.12200810 -0.0996115222 0.0000000
Viernes-Lunes      -0.15627406 -0.16806830 -0.1444798172 0.0000000
Sábado-Lunes       -0.75658228 -0.80761167 -0.7055528942 0.0000000
Miércoles-Martes    0.02692409  0.01583040  0.0380177780 0.0000000
Jueves-Martes      -0.01156243 -0.02293721 -0.0001876555 0.0437334
Viernes-Martes     -0.05702668 -0.06898862 -0.0450647400 0.0000000
Sábado-Martes      -0.65733490 -0.70840331 -0.6062664950 0.0000000
Jueves-Miércoles   -0.03848652 -0.04977167 -0.0272013719 0.0000000
Viernes-Miércoles  -0.08395077 -0.09582751 -0.0720740245 0.0000000
Sábado-Miércoles   -0.68425899 -0.73530751 -0.6332104706 0.0000000
Viernes-Jueves     -0.04546425 -0.05760396 -0.0333245384 0.0000000
Sábado-Jueves      -0.64577247 -0.69688281 -0.5946621313 0.0000000
Sábado-Viernes     -0.60030822 -0.65155244 -0.5490640106 0.0000000
```

Se comparan todos los días de la semana entre ellos. Si el Pvalor es 0 significa que tienen diferencias significativas. La pareja **jueves - martes** casi no tiene diferencias significativas (porque lo comparamos con la probabilidad de 0,05) entonces al ser el Pvalor=0,044 está en el límite de no ser significativo y se podría decir que entre el martes y el jueves no hay diferencias entre ellos, y el tiempo de espera (si solo dependiera de esta variable) sería igual.

A continuación se representa la gráfica del test de Tuckey, y se observa como efectivamente entre la pareja jueves-martes no hay diferencias significativas ya que la línea que representa ese par en el gráfico está rozando la línea vertical de no diferencias.

El resultado del *posthoc* de una manera más visual sería el siguiente:

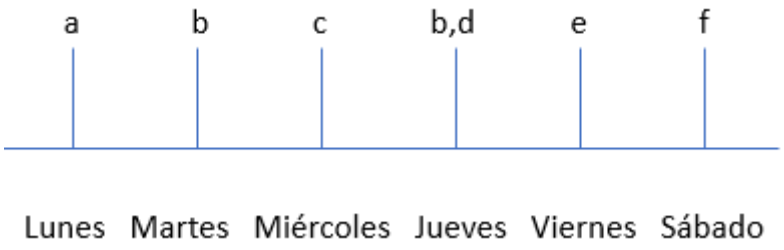


Ilustración 67 Resultado del posthoc para la relación del tiempo de espera y el día de la semana (Fuente: propia)

```
plot(posthoc_Diasemana) #graf104
```

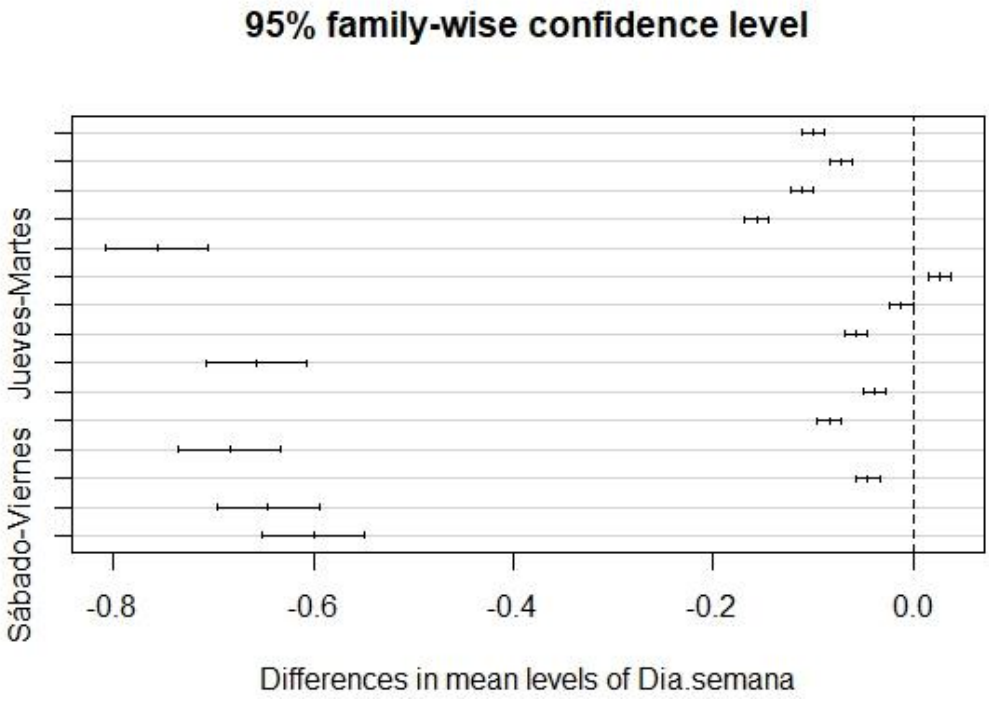


Ilustración 68 Representación gráfica del nivel de no diferencia sobre la variable explicativa día de la semana

```
#-----------------------------------------------------#
# Tiempo de espera - Mes
#-----------------------------------------------------#
```

A continuación se muestran dos ilustraciones que corresponden a un análisis temporal de la variable "tiempo de espera" para cada mes mediante la función *boxplot* o diagrama de cajas: en la primera figura se muestra el eje y que corresponde al tiempo de espera en minutos; y en la segunda ilustración se muestra la misma gráfica pero con el eje y en escala logarítmica.

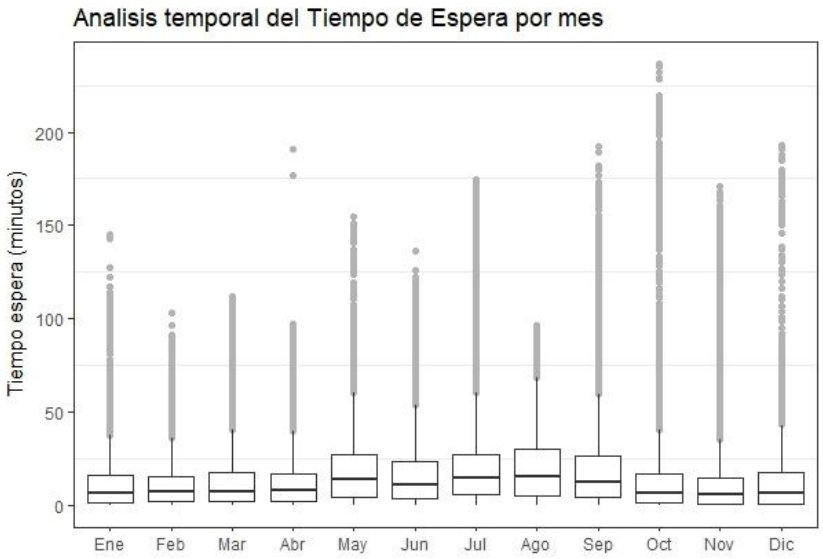


Ilustración 69 Análisis temporal de la variable Tiempo de Espera por mes (Eje Y: minutos)

De las ilustraciones que acompañan este análisis se puede deducir que los meses mayo, julio y agosto son los meses que tienen la mediana del tiempo de espera más alta.

Adicionalmente se observa que dado que en todos los meses, los diagramas de cajas son considerablemente alargadas significa que más dispersa es la distribución de los datos. También se observa más simetría en los meses centrales que en los meses de empiece y fin de año, lo que significa que en estos meses la distribución de datos es sesgada hacia arriba lo que infiere una asimetría positiva.

Con este diagrama no se puede deducir si hay o no diferencias significativas entre los diferentes meses. Por este motivo se procede a realizar a continuación el test estadístico de la variable tiempo de espera en función de los meses para conocer entre cuales existen diferencias significativas.

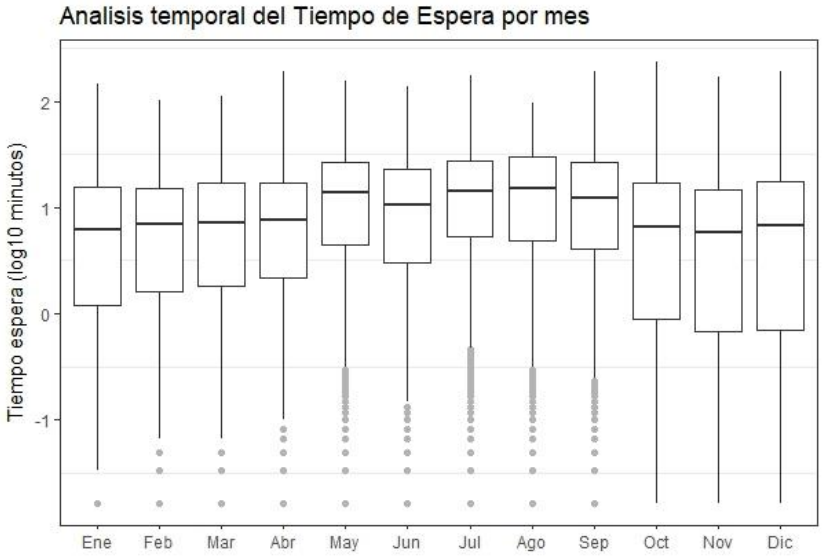


Ilustración 70 Análisis temporal de la variable Tiempo de Espera por mes (Eje Y: Log(minutos))

A continuación se muestra parte del código en R que se programó para llevar a cabo esta sección de análisis estadístico y también se han pegado los resultados obtenidos de la propia consola de RStudio al ejecutar dicho código. En caso de que se desee consultar el código de R al completo, está disponible al final de esta memoria en el anexo B.

Comenzamos aplicando a la variable respuesta el modelo aov y mostrando con la función summary el resultado obtenido.

```
Modelo1_Mes_2<-aov(Tiempo.espera5~Mes, data=Data) # #LM regresión
lineal. Aov variable discreta, como es nuestro caso.

summary(Modelo1_Mes_2)
   Df Sum Sq Mean Sq F value Pr(>F)
Mes            11   9536   866.9    1529 <2e-16 ***
Residuals   369385 209486     0.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
46217 observations deleted due to missingness
```

Efectivamente se demuestra que, dado que el Pvalor es aproximadamente 0, el mes (la variable explicativa en este caso) es significativa para la variable respuesta tiempo de espera estudiada en este proyecto.

Aplicamos al modelo anterior el test a posteriori y confirmamos más adelante en la tabla obtenida entre cuales de los pares de los posibles meses existen diferencias significativas.

```
posthoc_mes <- TukeyHSD(x=Modelo1_Mes_2, 'Mes', conf.level=0.95)
posthoc_mes
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Tiempo.espera5 ~ Mes, data = Data)

$Mes
                diff          lwr          upr     p adj
Feb-Ene  0.0385384737  0.018676355  0.058400593 0.0000000
Mar-Ene  0.0854554907  0.065836542  0.105074440 0.0000000
Abr-Ene  0.1102633065  0.090216315  0.130310298 0.0000000
May-Ene  0.3454807098  0.326950801  0.364010618 0.0000000
Jun-Ene  0.2363113708  0.216614618  0.256008123 0.0000000
Jul-Ene  0.4045293200  0.384510072  0.424548568 0.0000000
Ago-Ene  0.3841343892  0.360765160  0.407503618 0.0000000
Sep-Ene  0.3205513840  0.301006256  0.340096513 0.0000000
Oct-Ene -0.0036320533 -0.023488666  0.016224559 0.9999848
Nov-Ene -0.0500966962 -0.070618037 -0.029575355 0.0000000
Dic-Ene -0.0030410326 -0.024592155  0.018510090 0.9999990
Mar-Feb  0.0469170170  0.027706539  0.066127495 0.0000000
Abr-Feb  0.0717248329  0.052077412  0.091372254 0.0000000
May-Feb  0.3069422362  0.288845363  0.325039109 0.0000000
Jun-Feb  0.1977728971  0.178482969  0.217062826 0.0000000
Jul-Feb  0.3659908464  0.346371734  0.385609959 0.0000000
Ago-Feb  0.3455959155  0.322568537  0.368623294 0.0000000
Sep-Feb  0.2820129104  0.262877829  0.301147992 0.0000000
Oct-Feb -0.0421705270 -0.061623660 -0.022717394 0.0000000
Nov-Feb -0.0886351698 -0.108766358 -0.068503982 0.0000000
Dic-Feb -0.0415795062 -0.062759454 -0.020399559 0.0000000
Abr-Mar  0.0248078158  0.005406256  0.044209376 0.0017482
```

May-Mar	0.2600252191	0.242195576	0.277854862	0.0000000
Jun-Mar	0.1508558801	0.131816428	0.169895332	0.0000000
Jul-Mar	0.3190738293	0.299700937	0.338446721	0.0000000
Ago-Mar	0.2986788985	0.275860933	0.321496864	0.0000000
Sep-Mar	0.2350958933	0.216213343	0.253978444	0.0000000
Oct-Mar	-0.0890875440	-0.108292328	-0.069882760	0.0000000
Nov-Mar	-0.1355521869	-0.155443494	-0.115660880	0.0000000
Dic-Mar	-0.0884965233	-0.109448601	-0.067544446	0.0000000
May-Abr	0.2352174033	0.216917816	0.253516990	0.0000000
Jun-Abr	0.1260480642	0.106567833	0.145528296	0.0000000
Jul-Abr	0.2942660135	0.274459761	0.314072266	0.0000000
Ago-Abr	0.2738710826	0.250684055	0.297058110	0.0000000
Sep-Abr	0.2102880775	0.190961168	0.229614987	0.0000000
Oct-Abr	-0.1138953599	-0.133537214	-0.094253505	0.0000000
Nov-Abr	-0.1603600027	-0.180673614	-0.140046391	0.0000000
Dic-Abr	-0.1133043391	-0.134657752	-0.091950926	0.0000000
Jun-May	-0.1091693390	-0.127084558	-0.091254120	0.0000000
Jul-May	0.0590486102	0.040779420	0.077317800	0.0000000
Ago-May	0.0386536793	0.016765009	0.060542350	0.0000005
Sep-May	-0.0249293258	-0.042677708	-0.007180944	0.0002770
Oct-May	-0.3491127631	-0.367203592	-0.331021934	0.0000000
Nov-May	-0.3955774060	-0.414395441	-0.376759371	0.0000000
Dic-May	-0.3485217424	-0.368457737	-0.328585747	0.0000000
Jul-Jun	0.1682179492	0.148766270	0.187669629	0.0000000
Ago-Jun	0.1478230184	0.124938123	0.170707914	0.0000000
Sep-Jun	0.0842400132	0.065276638	0.103203389	0.0000000
Oct-Jun	-0.2399434241	-0.259227683	-0.220659165	0.0000000
Nov-Jun	-0.2864080670	-0.306376116	-0.266440017	0.0000000
Dic-Jun	-0.2393524033	-0.260377351	-0.218327455	0.0000000
Ago-Jul	-0.0203949309	-0.043557976	0.002768114	0.1488744
Sep-Jul	-0.0839779360	-0.103276066	-0.064679806	0.0000000
Oct-Jul	-0.4081613733	-0.427774911	-0.388547836	0.0000000
Nov-Jul	-0.4546260162	-0.474912248	-0.434339784	0.0000000
Dic-Jul	-0.4075703526	-0.428897722	-0.386242984	0.0000000
Sep-Ago	-0.0635830051	-0.086337531	-0.040828480	0.0000000
Oct-Ago	-0.3877664425	-0.410789072	-0.364743813	0.0000000
Nov-Ago	-0.4342310853	-0.457829429	-0.410632742	0.0000000
Dic-Ago	-0.3871754217	-0.411674549	-0.362676294	0.0000000
Oct-Sep	-0.3241834373	-0.343312803	-0.305054071	0.0000000
Nov-Sep	-0.3706480802	-0.390466582	-0.350829579	0.0000000
Dic-Sep	-0.3235924166	-0.344475387	-0.302709446	0.0000000
Nov-Oct	-0.0464646429	-0.066590398	-0.026338888	0.0000000
Dic-Oct	0.0005910208	-0.020583763	0.021765804	1.0000000
Dic-Nov	0.0470556636	0.025256310	0.068855017	0.0000000

Se observa que en los pares marcados en negrita (por ejemplo, Diciembre- Enero, Agosto-Julio,…), dado que el Pvalor no es cero, no existen diferencias significativas entre dichos pares.

Para el resto de pares de meses existen diferencias significativas que afectan a que dependiendo del mes el tiempo de espera se verá afectado por él.

En la ilustración siguiente (pese a que en la gráfica existente en esta memoria no se aprecie con mucha exactitud) si se ampliara se observaría que los pares marcados en negrita (Pvalor distinto de cero), están tocando la línea vertical discontinua de no diferencias.

El resultado del *posthoc* de una manera más visual sería el siguiente:

Ilustración 71 Resultado del posthoc para la relación del tiempo de espera y el mes (Fuente: propia)

```
plot(posthoc_mes) #Graf106
```

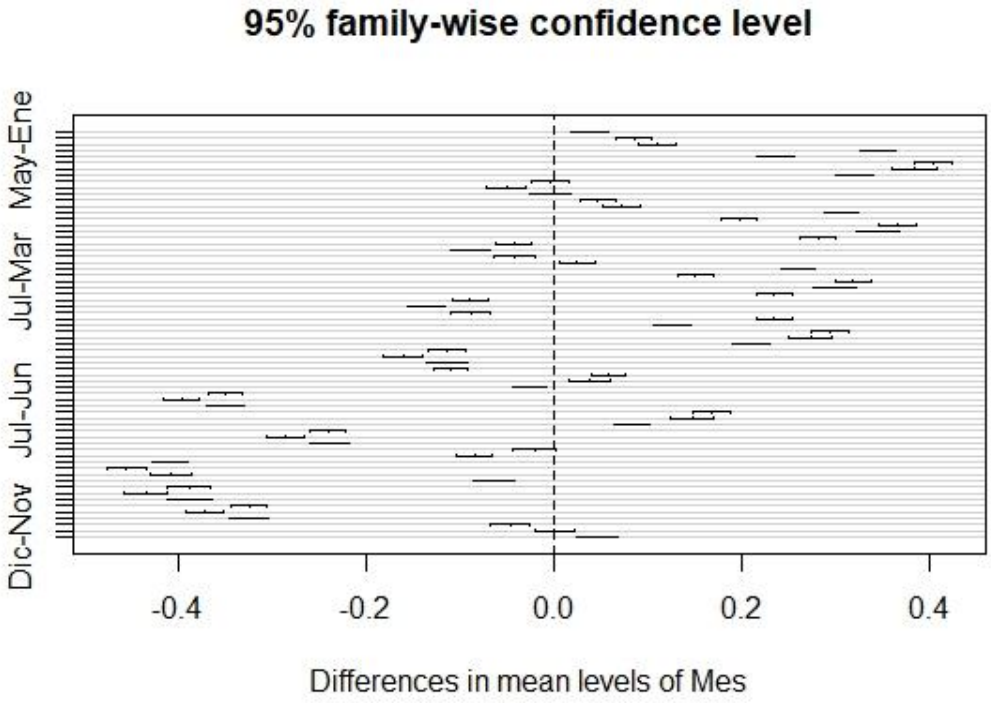


Ilustración 72 Representación gráfica del nivel de no diferencia sobre la variable explicativa mes

```
#-----------------------------------------------------#
# Tiempo de espera - Año
#-----------------------------------------------------#
```

A continuación se muestran dos ilustraciones que corresponden a un análisis temporal de la variable "tiempo de espera" por cada año estudiado mediante la función *boxplot* o diagrama de cajas: en la primera figura se muestra el eje y que corresponde al tiempo de espera en minutos; y en la segunda ilustración se muestra la misma gráfica pero con el eje y en escala logarítmica.

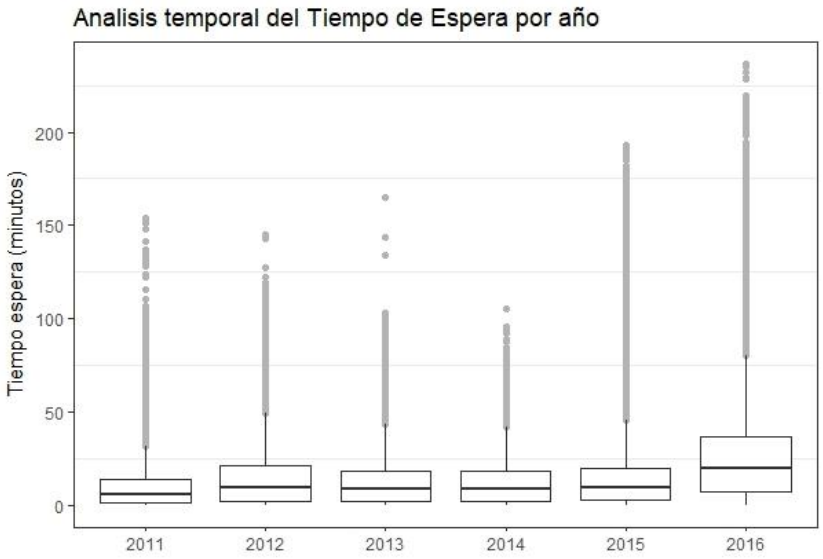


Ilustración 73 Análisis temporal de la variable Tiempo de Espera por año (Eje Y: minutos)

En las ilustraciones que acompañan este análisis se observa que en 2016 ha habido un crecimiento del tiempo de espera. De hecho puede decirse que con respecto al año 2011, año en el que se comenzó a medir los tiempos y la afluencia de gente en el Ayuntamiento de Sant Cugat, ha ido aumentando año a año la mediana del tiempo de espera.

En 2016, parece que la distribución de las observaciones es mucho más simétrica que el resto de años, aunque igualmente tiene un pequeño sesgo positivo.

Con este diagrama no se puede deducir si hay o no diferencias significativas entre los diferentes años, aunque puede predecirse que si será así dada la tendencia a aumentar del tiempo de espera. Por este motivo se procede a realizar a continuación un test estadístico de la variable tiempo de espera en función del año para conocer si entre algunos de ellos no existen diferencias significativas.

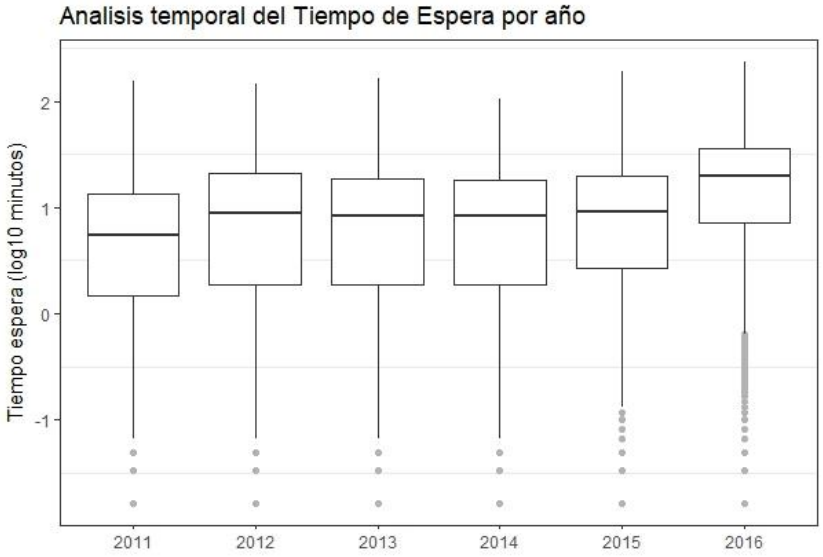


Ilustración 74 Análisis temporal de la variable Tiempo de Espera por año (Eje Y: Log(minutos))

A continuación se muestra parte del código en R que se programó para llevar a cabo esta sección de análisis estadístico y también se han pegado los resultados obtenidos de la propia consola de RStudio al ejecutar dicho código. En caso de que se desee consultar el código de R al completo, está disponible al final de esta memoria en el anexo B.

Aplicamos el modelo aov a la variable tiempo de espera con respecto el año.

```
Modelo1_Año_2<-aov(Tiempo.espera5~Año, data=Data)
summary(Modelo1_Año_2)
                Df Sum Sq Mean Sq F value Pr(>F)
Año              5   8820  1764.0    3100 <2e-16 ***
Residuals   369391 210202     0.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
46217 observations deleted due to missingness
```

Efectivamente se demuestra que, dado que el Pvalor es aproximadamente 0, el año (la variable explicativa en este caso) es significativa para la variable respuesta tiempo de espera estudiada en este proyecto.

Aplicamos al modelo anterior el test a posteriori y confirmamos en la siguiente tabla entre cuales de los pares de los posibles años analizados existen diferencias significativas.

```
posthoc_año <- TukeyHSD(x=Modelo1_Año_2, 'Año', conf.level=0.95)
posthoc_año #2014 y 2013 no tienen diferencia significativa
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Tiempo.espera5 ~ Año, data = Data)

$Año
                  diff         lwr          upr     p adj
2012-2011  0.091857662  0.07998462  0.103730708 0.0000000
2013-2011  0.066126992  0.05402078  0.078233203 0.0000000
2014-2011  0.061696667  0.04969005  0.073703282 0.0000000
2015-2011  0.146139176  0.13409657  0.158181779 0.0000000
2016-2011  0.481125279  0.46890369  0.493346865 0.0000000
2013-2012 -0.025730670 -0.03798992 -0.013471416 0.0000000
2014-2012 -0.030160995 -0.04232191 -0.018000084 0.0000000
2015-2012  0.054281513  0.04208507  0.066477958 0.0000000
2016-2012  0.389267616  0.37689441  0.401640819 0.0000000
2014-2013 -0.004430325 -0.01681898  0.007958335 0.9117522
2015-2013  0.080012183  0.06758864  0.092435725 0.0000000
2016-2013  0.414998287  0.40240117  0.427595400 0.0000000
2015-2014  0.084442508  0.07211600  0.096769018 0.0000000
2016-2014  0.419428611  0.40692718  0.431930040 0.0000000
2016-2015  0.334986103  0.32245011  0.347522100 0.0000000
```

Se observa que en para el par 2014-2013 marcado en negrita dado que el Pvalor no es cero, no existe diferencia significativa entre dicho par.

Para el resto de años existen diferencias significativas que afectan a que dependiendo del año el tiempo de espera se verá afectado por él.

En la ilustración siguiente se observa que el par 2014-2013 marcado en negrita (Pvalor distinto de cero), está tocando la línea vertical discontinua de no diferencias.

El resultado del *posthoc* de una manera más visual sería el siguiente:

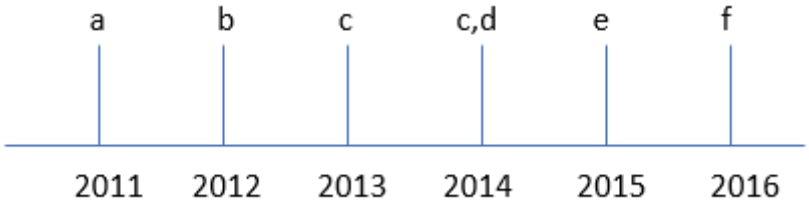


Ilustración 75 Resultado del posthoc para la relación del tiempo de espera y el año (Fuente: propia)

```
plot(posthoc_año) #graf_108
```

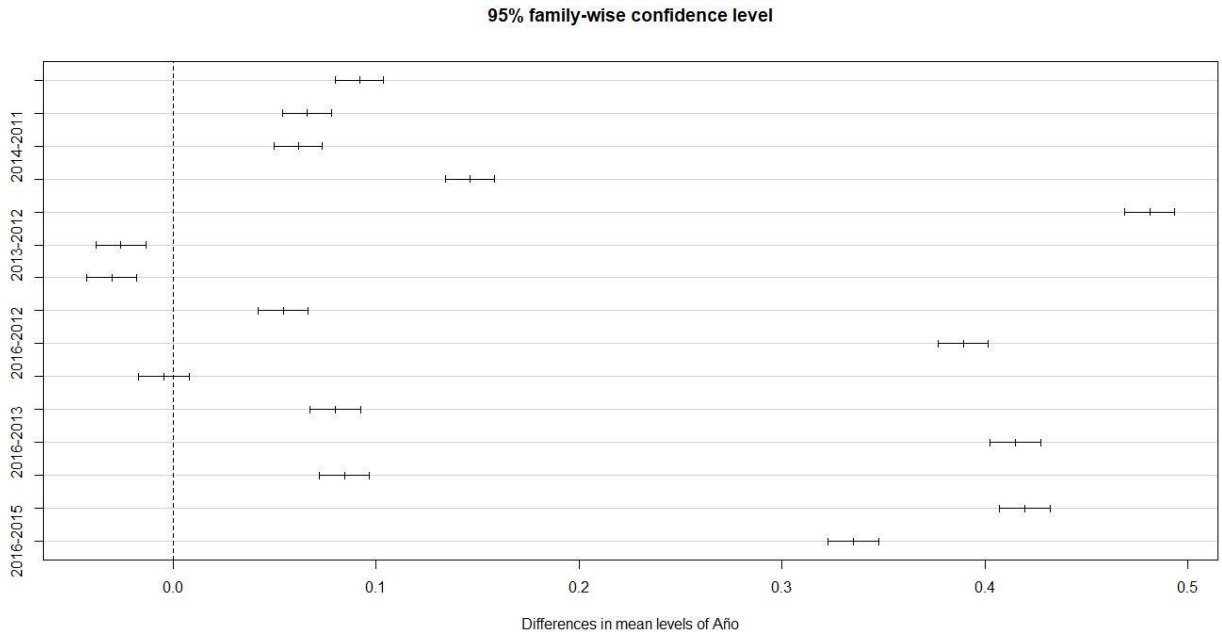


Ilustración 76 Representación gráfica del nivel de no diferencia sobre la variable explicativa año

En este caso, existe una particularidad que debe ser tomada en cuenta para poder predecir lo que sucederá en años posteriores y que no se había dado con las variables explicativas anteriores:

El mes y el día de la semana son variables circulares, por lo que servirá aplicar el modelo anova, pero para el año, dado que no es circular, sería mejor considerar esta variable como una variable numérica y que así facilite la creación del modelo final.

Primeramente crearemos una nueva variable (una nueva columna en la base de datos) para el año que será numérica y no un factor como la variable Año anterior.

```
Data$Año #Es un factor y por eso tiene niveles
#Creamos una nueva variable/columna que sea numérico y no factor
Data$AñoNumerico<- as.numeric(as.character(Data$Año))
summary(Data$AñoNumerico)
```

Representamos de nuevo el gráfico con la variable numérica año, y en este caso lo haremos con la función *geom point* por ser numérico (no sirve el boxplot)

```
ggplot(data=Data, aes(x=AñoNumerico, y=Tiempo.espera5))+
```

```
  geom_point(na.rm=TRUE)+
  geom_smooth(na.rm=TRUE, method = "lm",formula = y~poly(x,2))+
  #geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (log10 minutos)", title="Análisis temporal del
Tiempo de Espera por año")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())
```

Una vez obtenido el gráfico dispondremos del modelo lineal para poder predecir, y este modelo lineal habrá que ejecutarlo de la siguiente forma:

```
Modelo3_Año<-lm(Tiempo.espera5~AñoNumerico, data=Data)
summary(Modelo3_Año)

Call:
lm(formula = Tiempo.espera5 ~ AñoNumerico, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7006 -0.3585  0.2167  0.5611  1.6192

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.424e+02  1.469e+00  -96.92   <2e-16 ***
AñoNumerico  7.108e-02  7.295e-04   97.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7603 on 369395 degrees of freedom
  (46217 observations deleted due to missingness)
Multiple R-squared:  0.02505,  Adjusted R-squared:  0.02505
F-statistic:  9492 on 1 and 369395 DF,  p-value: < 2.2e-16
```

En este caso el summary es distinto que el del anova de los casos anteriores: ahora aparece de la forma y=a+bx --> que el intercepto sea significativo (es casi cero) quiere decir que no pasa por el origen. La variable respuesta que es el año si es significativo, lo que quiere decir que no tienen pendiente igual a cero (la pendiente es 0,07) a cada incremento de año hay un 0,07 más del logaritmo del tiempo de espera: `log10(0.07): -1.154902.`

Pero no sabemos si este modelo es el adecuado, por lo que probamos a crear un nuevo modelo para esta variable que lo explique mejor y comprobaremos si es mejor o peor que el anterior:

```
Modelo4_Año<-lm(Tiempo.espera5~AñoNumerico+I(AñoNumerico^2),
data=Data)

summary(Modelo4_Año)

Call:
lm(formula = Tiempo.espera5 ~ AñoNumerico + I(AñoNumerico^2),
    data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8012 -0.3650  0.2150  0.5606  1.5863
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.164e+05  2.028e+03   57.40   <2e-16 ***
AñoNumerico      -1.157e+02  2.014e+00  -57.44   <2e-16 ***
I(AñoNumerico^2)  2.875e-02  5.003e-04   57.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7569 on 369394 degrees of freedom
  (46217 observations deleted due to missingness)
Multiple R-squared:  0.03369,  Adjusted R-squared:  0.03369
F-statistic:  6440 on 2 and 369394 DF,  p-value: < 2.2e-16
```

Este modelo tiene mayor poder explicativo ya que el AIC es menor cuando se compara con el modelo lineal sencillo, este sería un modelo más óptimo que el anterior y explica el 3,4 % de los datos (Multiple R-squared: 0.03369,).

Pero de nuevo no sabemos si este modelo es el óptimo, por lo que probamos a crear un nuevo modelo para esta variable y comprobaremos si es mejor o peor que el anterior:

```
Modelo5_Año<-
lm(Tiempo.espera5~AñoNumerico+I(AñoNumerico^2)+I(AñoNumerico^3),
data=Data)

summary(Modelo5_Año)

Call:
lm(formula = Tiempo.espera5 ~ AñoNumerico + I(AñoNumerico^2) +
    I(AñoNumerico^3), data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8012 -0.3650  0.2150  0.5606  1.5863

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.164e+05  2.028e+03   57.40   <2e-16 ***
AñoNumerico      -1.157e+02  2.014e+00  -57.44   <2e-16 ***
I(AñoNumerico^2)  2.875e-02  5.003e-04   57.47   <2e-16 ***
I(AñoNumerico^3)         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7569 on 369394 degrees of freedom
  (46217 observations deleted due to missingness)
Multiple R-squared:  0.03369,  Adjusted R-squared:  0.03369
F-statistic:  6440 on 2 and 369394 DF,  p-value: < 2.2e-16
```

Este modelo 5 falla al ajustar el cúbico del año numérico, por eso sabemos que la gráfica tiene pendiente positiva pero que no crece al infinito ya que el cuadrado del tiempo de espera tiene pendiente negativa (log (de una pendiente positiva)).

```
AIC(Modelo3_Año, Modelo4_Año, Modelo5_Año) #df      AIC Modelo3_Año  3
845853.4 Modelo4_Año  4 842567.0 Modelo5_Año  4 842567.0
            df      AIC
Modelo3_Año  3 845853.4
Modelo4_Año  4 842567.0
Modelo5_Año  4 842567.0
```

Comprobamos que el modelo Modelo4_Año es el que más se ajusta a la realidad y este sería el modelo óptimo.

```
#-----------------------------------------------------#
# Tiempo de espera - Día de mes
#-----------------------------------------------------#
```

A continuación se muestran dos ilustraciones que corresponden a un análisis temporal de la variable "tiempo de espera" por cada día de mes mediante la función *boxplot* o diagrama de cajas: en la primera figura se muestra el eje y que corresponde al tiempo de espera en minutos; y en la segunda ilustración se muestra la misma gráfica, pero con el eje y en escala logarítmica.

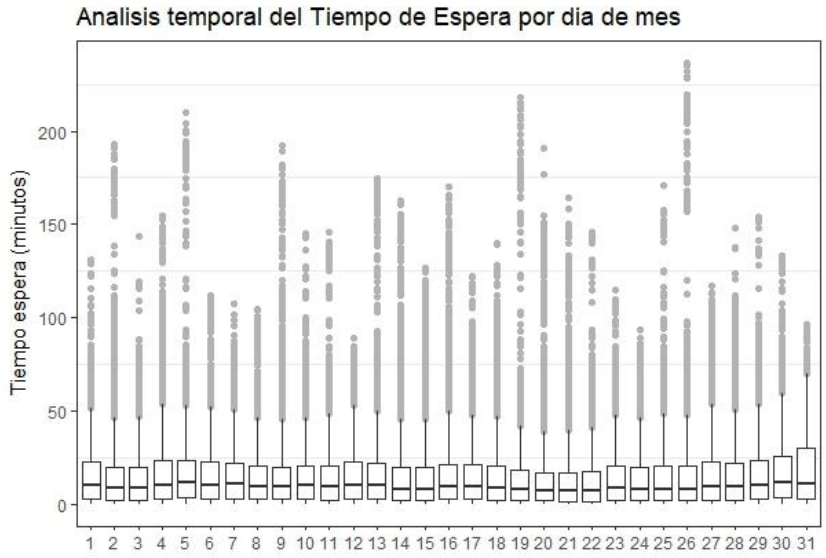


Ilustración 77 Análisis temporal de la variable Tiempo de Espera por día de mes (Eje Y: minutos)

De las ilustraciones que acompañan este análisis se puede deducir que los días 2, 3, 14, 15, 20, 21, 22 tienen la mediana del tiempo de espera más baja.

Se aprecia que en varios días de mes la mediana está situada aproximadamente en el centro de la caja, por lo que se puede deducir que en dichos días la distribución de dichos datos es simétrica.

Con este diagrama no se puede deducir si hay o no diferencias significativas entre los diferentes días de mes. Por este motivo se procede a realizar a continuación el test estadístico de la variable tiempo de espera en función del día de mes para conocer entre que días del mes existen diferencias significativas.

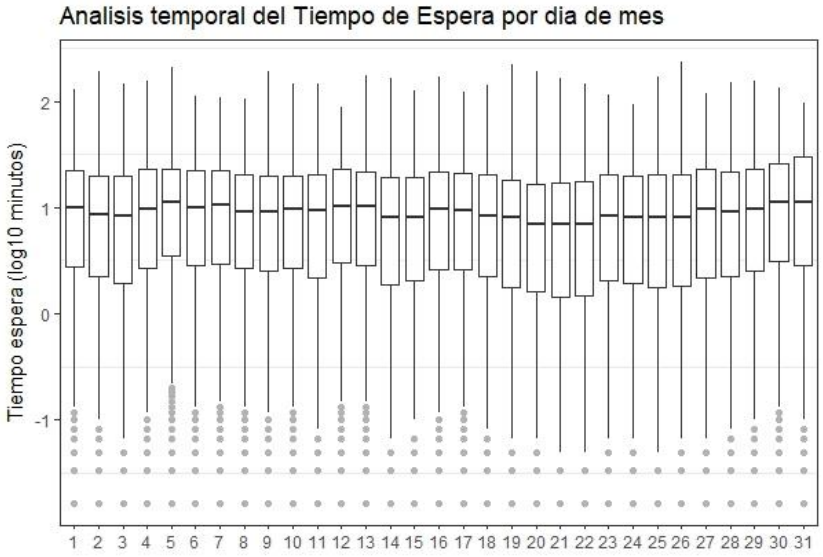


Ilustración 78 Análisis temporal de la variable Tiempo de Espera por día de mes (Eje Y: Log(minutos))

A continuación se muestra parte del código en R que se programó para llevar a cabo esta sección de análisis estadístico y también se han pegado los resultados obtenidos de la propia consola de RStudio al ejecutar dicho código. En caso de que se desee consultar el código de R al completo, está disponible al final de esta memoria en el anexo B.

Aplicamos el modelo aov a la variable tiempo de espera con respecto al día de mes.

```
Modelo1_DiaMes_2<-aov(Tiempo.espera5~Dia.2, data=Data)
summary(Modelo1_DiaMes_2)
               Df Sum Sq Mean Sq F value Pr(>F)
Dia.2          30   1220   40.67   68.97 <2e-16 ***
Residuals   369366 217802    0.59
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
46217 observations deleted due to missingness
```

Efectivamente se demuestra que, dado que el Pvalor es aproximadamente 0, el día de mes (la variable explicativa en este caso) es significativa para la variable respuesta tiempo de espera estudiada en este proyecto.

Aplicamos al modelo anterior el test a posteriori y confirmamos en la siguiente tabla entre cuales de los pares de los posibles días de mes existen diferencias significativas.

```
posthoc_Diames <- TukeyHSD(x=Modelo1_DiaMes_2, 'Dia.2',
conf.level=0.95)
posthoc_Diames
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Tiempo.espera5 ~ Dia.2, data = Data)

$Dia.2
              diff           lwr           upr     p adj
2-1   -0.0544572585 -9.413876e-02 -0.0147757530 0.0001054
3-1   -0.0779551945 -1.185755e-01 -0.0373348822 0.0000000
4-1    0.0042401231 -3.485382e-02  0.0433340629 1.0000000
5-1    0.0550546671  1.559323e-02  0.0945161019 0.0000668
```

```
6-1    0.0062692087 -3.471876e-02  0.0472571805 1.0000000
22-10 -0.1320637881 -1.688747e-01 -0.0952529255 0.0000000
23-10 -0.0486907713 -8.531603e-02 -0.0120655103 0.0002429
24-10 -0.0693480143 -1.066028e-01 -0.0320932055 0.0000000
25-10 -0.0735447319 -1.104413e-01 -0.0366481361 0.0000000
26-10 -0.0641314864 -1.011308e-01 -0.0271321893 0.0000000
 [ reached getOption("max.print") -- omitted … rows ]
```

No colocamos todo lo obtenido e R dado que las comparaciones de todos los días entre ellos es una lista demasiado larga. Es mejor que busquemos tendencias dentro de los días de mes:

Se observa que por ejemplo para los pares: 4-1, 6-1, … marcados en negrita dado que el Pvalor no es cero, no existen diferencias significativas entre dichos pares.

Para el resto de días de mes existen diferencias significativas que afectan a que dependiendo del día el tiempo de espera se verá afectado por él.

En la ilustración siguiente se observa que hay varios pares con Pvalor distinto de cero ya que dichos pares están tocando la línea vertical discontinua de no diferencias.

```
plot(posthoc_Diames) #graf_110
```

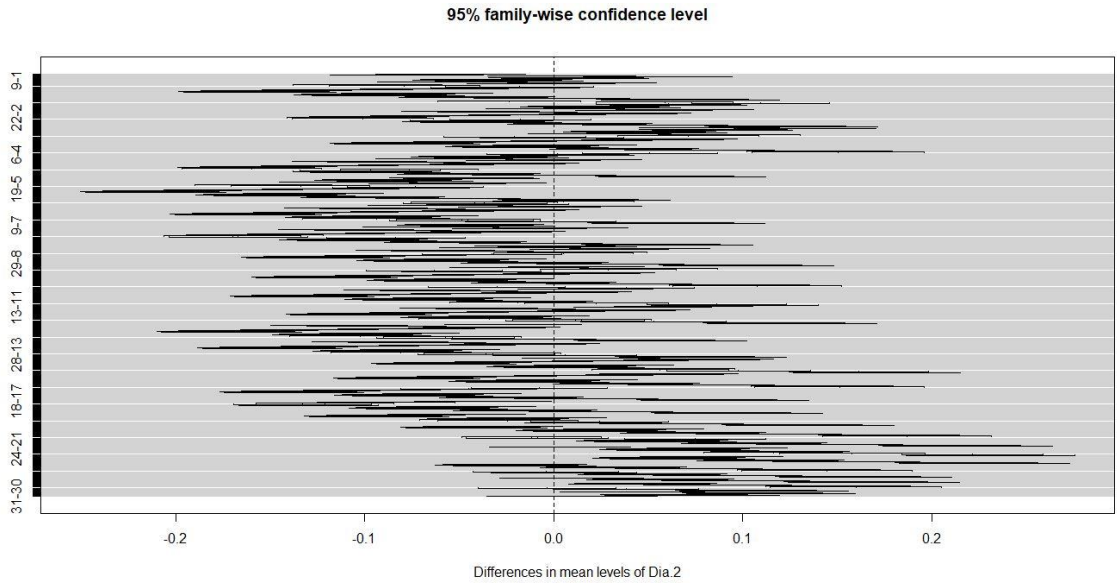


Ilustración 79 Representación gráfica del nivel de no diferencia sobre la variable explicativa día de mes

Si ahora por ejemplo considerásemos la variable día del mes como una variable continua, podríamos ver si hay (o no) tendencias dentro del mes: aplicamos modelo LM.

```
Modelo2_DiaMes_2 <- lm(Tiempo.espera5~Dia.1, data=Data)
summary(Modelo2_DiaMes_2)
```

```
Call:
lm(formula = Tiempo.espera5 ~ Dia.1, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5403 -0.3756  0.2155  0.5771  1.6515
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7637240  0.0026239  291.07   <2e-16 ***
Dia.1       -0.0015660  0.0001463  -10.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7699 on 369395 degrees of freedom
  (46217 observations deleted due to missingness)
Multiple R-squared:  0.0003102,       Adjusted R-squared:  0.0003075
F-statistic: 114.6 on 1 and 369395 DF,  p-value: < 2.2e-16
```

Dado que no sabemos si este modelo es el adecuado, probaremos a crear un nuevo modelo (cuadrado) para esta variable que lo explique mejor y comprobaremos si es mejor o peor que el anterior:

```
Modelo3_DiaMes_2 <- lm(Tiempo.espera5~Dia.1+I(Dia.1^2), data=Data)
summary(Modelo3_DiaMes_2)
Call:
lm(formula = Tiempo.espera5 ~ Dia.1 + I(Dia.1^2), data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5943 -0.3770  0.2159  0.5773  1.6410

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.291e-01  4.177e-03  198.49   <2e-16 ***
Dia.1       -1.334e-02  6.034e-04  -22.11   <2e-16 ***
I(Dia.1^2)   3.715e-04  1.847e-05   20.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7695 on 369394 degrees of freedom
  (46217 observations deleted due to missingness)
Multiple R-squared:  0.001404, Adjusted R-squared:  0.001398
F-statistic: 259.6 on 2 and 369394 DF,  p-value: < 2.2e-16
```

De nuevo creamos el siguiente modelo (cúbico) para comprobar si se ajusta mejor o peor que el anterior:

```
Modelo4_DiaMes_2 <- lm(Tiempo.espera5~Dia.1+I(Dia.1^2)+I(Dia.1^3),
data=Data)
summary(Modelo4_DiaMes_2)
Call:
lm(formula = Tiempo.espera5 ~ Dia.1 + I(Dia.1^2) + I(Dia.1^3),
    data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6458 -0.3764  0.2156  0.5761  1.6657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.050e-01  6.016e-03  117.18   <2e-16 ***
Dia.1        2.894e-02  1.595e-03   18.15   <2e-16 ***
I(Dia.1^2)  -2.891e-03  1.154e-04  -25.05   <2e-16 ***
I(Dia.1^3)   6.862e-05  2.396e-06   28.63   <2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7686 on 369393 degrees of freedom
  (46217 observations deleted due to missingness)
Multiple R-squared:  0.003615, Adjusted R-squared:  0.003607
F-statistic: 446.7 on 3 and 369393 DF,  p-value: < 2.2e-16


AIC(Modelo2_DiaMes_2,Modelo3_DiaMes_2,Modelo4_DiaMes_2)
                 df      AIC
Modelo2_DiaMes_2  3 855110.9
Modelo3_DiaMes_2  4 854708.6
Modelo4_DiaMes_2  5 853891.7
```

Gracias a la función AIC podemos confirmar que el modelo 4 (cúbico) es el que mejor resultado da ya que intenta ajustar una curva polinómica donde se observa que el tiempo de espera disminuye hacia fin de mes, pero esta disminución NO es lineal. A principio y a final de mes (para los días “veinti” poco) el tiempo de espera es menor y esto lo respalda el modelo cúbico

```
#-----------------------------------------------------#
# Tiempo de espera - Código de gestor
#-----------------------------------------------------#
```

A continuación se muestran dos ilustraciones que corresponden a un análisis temporal de la variable “tiempo de espera” para código de gestor mediante la función *boxplot* o diagrama de cajas: en la primera figura se muestra el eje y que corresponde al tiempo de espera en minutos; y en la segunda ilustración se muestra la misma gráfica pero con el eje y en escala logarítmica.

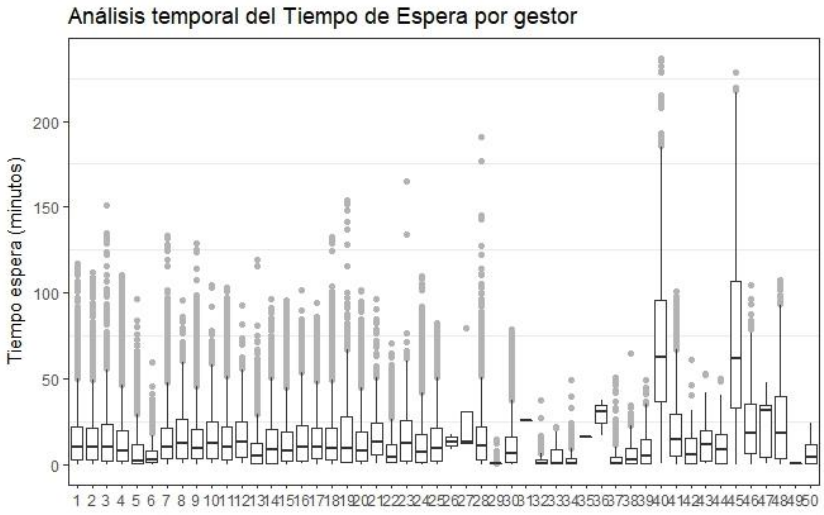


Ilustración 80 Análisis temporal de la variable Tiempo de Espera por código de gestor (Eje Y: minutos)

Las ilustraciones que acompañan este análisis debido a que contienen información de los cincuenta gestores, tienen información más comprimida que en el resto de los casos anteriores pero igualmente pueden observarse algunos datos relevantes como por ejemplo: los gestores 5, 6, 27... tienen la mediana del tiempo de espera más baja que el resto de los gestores. Además en estos casos la mediana no está colocada simétricamente sino se observa un sesgo hacia abajo lo que implica que la distribución de observaciones en este caso es asimétrica negativa.

Con este diagrama no se puede deducir si hay o no diferencias significativas entre los diferentes gestores. Por este motivo se procede a realizar a continuación el test estadístico de la variable tiempo de espera en función de los gestores para conocer entre cuáles de ellos existen diferencias significativas.

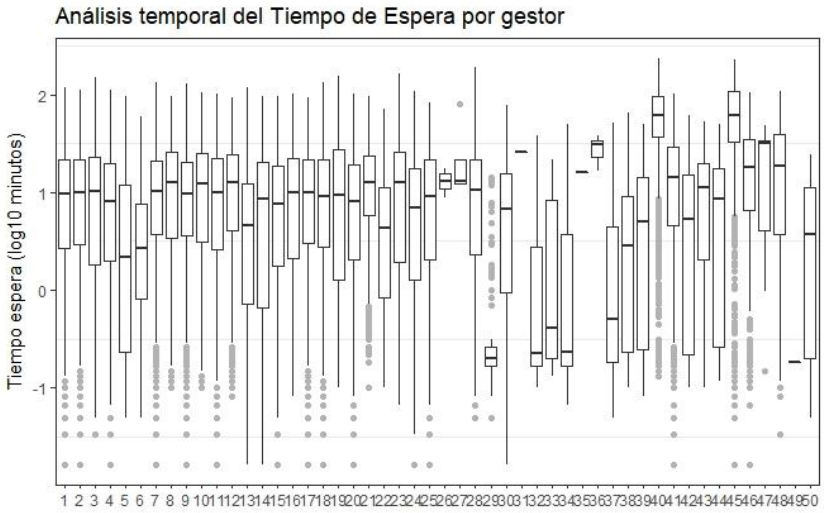


Ilustración 81 Análisis temporal de la variable Tiempo de Espera por código de gestor (Eje Y: Log(minutos))

A continuación se muestra parte del código en R que se programó para llevar a cabo esta sección de análisis estadístico y también se han pegado los resultados obtenidos de la propia consola de RStudio al ejecutar dicho código. En caso de que se desee consultar el código de R al completo, está disponible al final de esta memoria en el anexo B.

```
Data$Codi.Gestor2<-factor(Data$Codi.Gestor)
```

Aplicamos el modelo aov a la variable tiempo de espera con respecto el año.

```
Modelo1_Gestor_2<-aov(Tiempo.espera5~Codi.Gestor2, data=Data)
summary(Modelo1_Gestor_2)
                Df Sum Sq Mean Sq F value Pr(>F)
Codi.Gestor2     49   9395  191.74   337.8 <2e-16 ***
Residuals    369347 209627    0.57
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
46217 observations deleted due to missingness
```

Efectivamente se demuestra que, dado que el Pvalor es aproximadamente cero, el código de gestor (la variable explicativa en este caso) es significativa para la variable respuesta tiempo de espera estudiada en este proyecto.

Aplicamos al modelo anterior el test a posteriori y confirmamos en la siguiente tabla entre cuales de los pares de los posibles gestores analizados existen diferencias significativas.

```
posthoc_gestor <- TukeyHSD(x=Modelo1_Gestor_2, 'Codi.Gestor2',
conf.level=0.95)
posthoc_gestor
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

```
Fit: aov(formula = Tiempo.espera5 ~ Codi.Gestor2, data = Data)

$Codi.Gestor2
               diff           lwr           upr     p adj
2-1    1.468243e-02 -0.0113197552  0.0406846170 0.9897227
3-1   -2.947854e-02 -0.0906872859  0.0317302126 0.9996589
4-1   -6.281541e-02 -0.0888672159 -0.0367636086 0.0000000
5-1   -4.970933e-01 -0.5368069621 -0.4573795467 0.0000000
13-6   1.060222e-01 -0.0821726604  0.2942171493 0.9901466
14-6   2.436184e-01  0.0561372821  0.4310994717 0.0002417
15-6   2.953412e-01  0.1097132584  0.4809691786 0.0000003
16-6   3.689796e-01  0.1752696285  0.5626895270 0.0000000
… [ reached getOption("max.print") -- omitted … rows ]
```

Se observa que por ejemplo para los pares de gestores: 2-1, 3-1, 13-6 … marcados en negrita dado que el Pvalor no es cero, no existen diferencias significativas entre dichos pares.

Para el resto de gestores existen diferencias significativas que afectan a que dependiendo del gestor que atienda una atención el tiempo de espera se verá afectado por él.

En la ilustración siguiente se observa que hay varios pares de gestores con Pvalor distinto de cero ya que dichos pares están tocando la línea vertical discontinua de no diferencias.

```
plot(posthoc_gestor) #graf_112
```

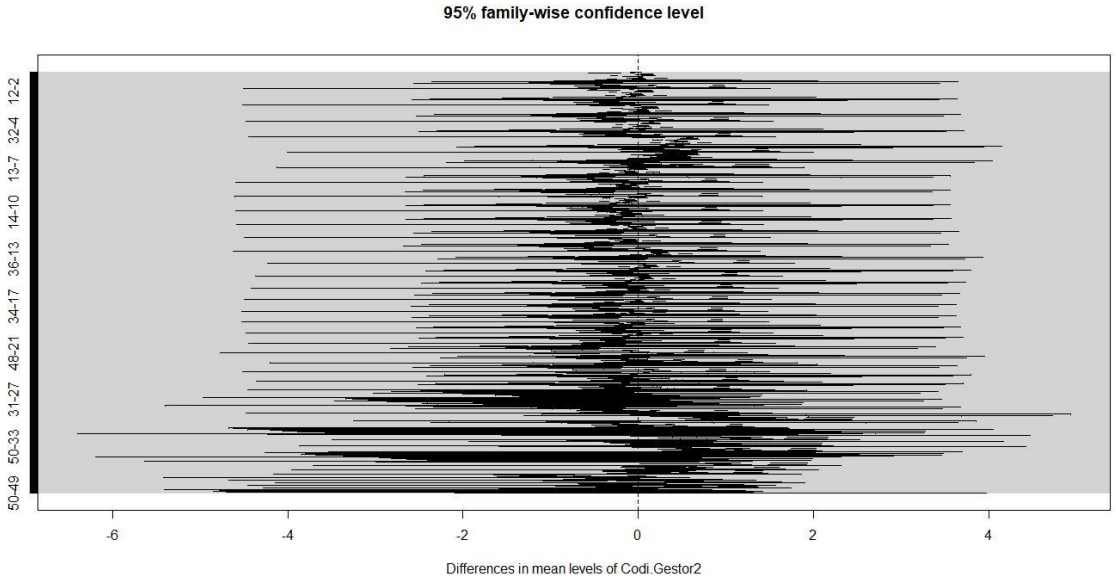


Ilustración 82 Representación gráfica del nivel de no diferencia sobre la variable explicativa código de gestor

Lo significativo de esta gráfica es que hay gestores que son significativamente diferentes. Por ejemplo: el 16 y el 6 (con Pvalor igual a cero) tienen diferencia significativa, pero entre el 13 y el 6 no hay diferencias significativas (tienen similar tiempo espera). Hay que tener precaución a la hora de explicar estos resultados ya que los gestores pueden venir de servicios y oficinas distintos que ya de por si tengan tiempos más o menos altos de espera.

Por lo que quedará como línea futura de este proyecto conseguir información del Ayuntamiento de Sant Cugat para poder diferenciar a los gestores por oficinas y así extraer resultados concretos que puedan aplicarse a cada una de las oficinas de atención al cliente.

Construcción del modelo matemático para la variable "Tiempo de espera":

```
#-----------------------------------------------------------------------#
# Modelo general que explique la variable respuesta: Tiempo de espera
#-----------------------------------------------------------------------#
```

Para construir este modelo debemos tener en cuenta toda la información obtenida en las secciones anteriores.

Se debe unir todo lo que se ha obtenido de cada una de las variables explicativas en un mismo código siguiendo el criterio: ***backward selection criteria***:

ya que primero hemos analizado todas las variables explicativas y luego hemos visto cuáles afectan al modelo. En este caso, todas ellas son significativas para este modelo.

```
MODELOGENERAL <- lm(Data$Tiempo.espera5~ Data$Dia.semana +Data$Mes
+Data$AñoNumerico+I(Data$AñoNumerico^2)+Data$Dia.2+Data$Codi.Gestor2)
```

Seguidamente obtenemos en una tabla los resultados del modelo general para comprobar lo que se ha obtenido por separado para cada variable en el apartado anterior:

```
anova(MODELOGENERAL)
```

```
Analysis of Variance Table

Response: Data$Tiempo.espera5
                          Df Sum Sq Mean Sq   F value    Pr(>F)
Data$Dia.semana            5   1812   362.3   694.447 < 2.2e-16 ***
Data$Mes                  11   9634   875.8  1678.712 < 2.2e-16 ***
Data$AñoNumerico           1   5504  5503.7 10549.033 < 2.2e-16 ***
I(Data$AñoNumerico^2)      1   1907  1906.5  3654.303 < 2.2e-16 ***
Data$Dia.2                30   1142    38.1    72.933 < 2.2e-16 ***
Data$Codi.Gestor2         49   6353   129.7   248.524 < 2.2e-16 ***
Residuals             369299 192671     0.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Una vez llegados a este punto, habríamos conseguido finalizar el último
objetivo de este proyecto.
```

Finalmente se ha incluido la tabla anterior con los resultados obtenidos para disponer del resumen del modelo que representa la variable respuesta tiempo de espera, el cuál era nuestro objetivo final de este trabajo de fin de máster.

¿Qué información nos aporta los datos de la tabla anterior? Explican que todas las variables incluidas en el modelo tienen significancia.

Esto es importante ya que este modelo puede no estar representando la misma información en conjunto que cuando analizamos cada variable por separado.

¿Que se podría hacer ahora para continuar este proyecto? Sería interesante evaluar el efecto de cada variable sobre el tiempo de espera manteniendo el efecto de las otras variables (y que pueden tener valor 0), pueden tener el valor del promedio, ... Esto queda como línea futura para realizarlo con el paquete disponible en R llamado *visreg* (Patrick Breheny, 2017).

Con este paquete y la función *visreg* obtendríamos cinco gráficas (cinco por las variables explicativas que disponemos) del efecto de cada variable sobre el tiempo de espera considerando el efecto del modelo global.

Como comentario global al margen de este modelo, pero en relación a todo el proceso que ejecuta un ciudadano en la oficina de atención al cliente del Ayuntamiento de Sant Cugat, se explica que la afluencia como tal no tiene efecto sobre el tiempo de espera en esta resolución temporal ya que cada visita (afluencia) es un valor del tiempo de espera, ya que cada día o cada visita tiene su propio tiempo de espera. Como otra posible línea futura de análisis podría cambiarse la resolución temporal y hacer un análisis agregado en el tiempo (semanal, mensual, donde si se evaluase el efecto del número de visitas en un periodo determinado sobre el tiempo de espera.

7. PRESUPUESTO ECONÓMICO

En este capítulo se estimará un presupuesto económico del proyecto realizado. Siguiendo las diferentes fases seguidas en la realización de este proyecto, se irán incluyendo los diferentes gastos tanto de software como de hardware que han surgido durante este desarrollo.

Antes de comenzar con dicha estimación, se comentará que existen numerosos criterios para establecer un precio a un proyecto de estas características. Algunas empresas, priman la mano de obra, otras el tiempo dedicado a su desarrollo en R, y otras la calidad del software usado.

Además existen otros factores que incrementarían el precio de este proyecto. Si por ejemplo, además del estudio de la variable "tiempo de espera", se hubiera incluido el estudio de la variable "tiempo de servicio", el precio de la programación aumentaría, no por el hecho de que aumentaran las horas de programación, sino por el hecho de dar solución a otro problema que tienen actualmente en el Ayuntamiento de Sant Cugat.

El presupuesto estimado para la realización de este TFM y su entrega a los responsables del Ayuntamiento de Sant Cugat, suponiendo que esta aplicación se hubiese realizado como una trabajadora autónoma, hubiera sido el siguiente:

Estimación presupuesto económico		
Partidas	**Cálculo**	**Coste**
Servicios de consultoría y asesoramiento + coste mano de obra programador R	20€/h * 240h	4.800,00 €
Cuota autónomo	264.44€/mes	265,00 €
Software libre	R	- €
	Rstudio	- €
	NotePad++	- €
Costes oficina y gastos corrientes	Alquiler + Luz + Conexión Internet + Teléfono + Gas (30 días)	196,00 €
Coste ordenador portátil con Windows 7	Coste total: 800€. Amortización fiscal 25% al año (según Hacienda): 200€. Al mes: 16,7€.	17,00 €
Licencia de Office 365 Empresa	8,80 € por usuario y mes	8,80 €
	Total	5.286,80 €

Tabla 14 Estimación presupuesto económico

El ordenador portátil ha sido necesario durante todo el proceso de desarrollo de aplicación. Primero se usó para la búsqueda de documentación inicial y para el aprendizaje de R y RStudio que se han usado para el desarrollo del proyecto, y que ha sido aprendido gracias a documentación y a ejemplos gráficos encontrados en Internet. Una vez finalizada la fase previa de documentación y aprendizaje, se comenzó a plantear lo que sería el desarrollo en sí. Para ello, se tuvieron reuniones con los responsables de la oficina de atención al cliente del Ayuntamiento de Sant Cugat; dicha reunión fue de vital importancia para el entendimiento del

problema que tenían y así poder mostrar las primeras conclusiones obtenidas; de ese modo pudieron comprender que gracias a nuestro trabajo iban a disponer de un apoyo para la mejora de la atención al cliente en sus oficinas.

A continuación, se empezó el análisis en R de las distintas variables que intervienen en un proceso de atención al cliente en sus oficinas. A la hora de realizar el presupuesto económico, se han tenido en cuenta aproximadamente unas 240 horas, en las que no se tiene en cuenta la fase inicial de aprendizaje, pero si se han tenido cuenta las horas en las que se han llevado a cabo reuniones con los responsables del Ayuntamiento de Sant Cugat, para la resolución de distintas dudas que han surgido durante el desarrollo y programación de esta aplicación.

Es por esto que teniendo en cuenta el salario de un consultor/programador junior, aproximadamente de 20€/hora, si ha necesitado unas 240 horas para la realización de este proyecto, se estima un gasto de mano de obra aproximado de 4.800€. Aproximadamente 240 horas son las horas equivalentes a 30 días de trabajo con una jornada laboral de 8 horas al día. Es por este motivo por el que todos los cálculos de los gastos se estiman aproximadamente para un mes de 30 días.

Aclaraciones a las diferentes partidas de este presupuesto:

- Salario del consultor, que es la única partida de ingresos y que representa el mayor coste del presupuesto. En esta partida se incluyen las horas de consultoría y programación para la realización de este proyecto.
- Cuota mensual de la seguridad social en régimen de autónomos (266.44€/mes). En este presupuesto este gasto se ha prorrateado para un mes.
- Licencias de dos programas usados para la realización de la aplicación:
    - R: Software libre
    - RStudio: Software libre
- Gastos generales correspondientes al lugar de trabajo. Estos son, el coste del alquiler, la luz, el gas, la conexión a Internet y el teléfono. La estimación de estos gastos fijos se ha prorrateado igualmente para 30 días.
- Coste del ordenador portátil. El ordenador tiene una depreciación, y según el Ministerio de Hacienda, se permite que anualmente se incluya como gasto de depreciación en la declaración de impuestos hasta un 25% del valor total. Esto implica que en 4 años el ordenador queda desgravado, por lo que al año se puede poner como gasto de depreciación del ordenador un 25% de su valor (200€). De este importe, se incluirá en este presupuesto la parte proporcional correspondiente a un mes (16.70€), ya que suponemos que cuando se adquiera un nuevo equipo éste tendrá un coste superior al actual (aplicando el criterio contable de coste de renovación).
- Coste de la licencia de Office 365 Empresa: supone un gasto de 8,80€/mes. Esta licencia es necesaria para la entrega del proyecto en formato Word y para la realización de las presentaciones oportunas a los responsables del Ayuntamiento de Sant Cugat en formato PPT.
- Todas las partidas de este presupuesto económico van grabadas con el 21% correspondiente en concepto de IVA.

Finalmente, se ha estimado que el presupuesto económico para el desarrollo de este proyecto y la elaboración del TFM, es de aproximadamente 5.300€.

8. CONCLUSIONES Y LÍNEAS FUTURAS

Conclusiones

En este último apartado de la memoria, me gustaría reflexionar sobre lo que ha supuesto para mí la realización de este Trabajo de Fin de Máster tanto a nivel profesional y académico como a nivel personal.

Este proyecto surgió de la idea de dos profesores del máster de Ingeniería de Telecomunicaciones de "La Salle, Universitat Ramon Llull", Xavier Vilasis i Cardona y Miguel Ramírez, a los cuales me gustaría aprovechar para agradecer su implicación en el mismo.

Mediante la resolución de este trabajo he podido aprender mucho sobre el análisis de datos y la teoría de colas, y gracias a ello he podido completar y ampliar la formación recibida en el máster de manera autodidacta. Así mismo este TFM me ha enseñado a superar dificultades que han ido surgiendo durante la realización y el desarrollo del mismo. Desde aquí me gustaría de nuevo, agradecer a mi tutor Xavier, las facilidades prestadas, su apoyo y ayuda durante todo el proceso de realización de este TFM.

Ha sido fascinante poder vivir el análisis de datos y el estudio de los mismos desde dentro de una organización como es el Ayuntamiento de Sant Cugat. Es muy motivante saber que los datos con los que estoy trabajando son reales, y las conclusiones obtenidas pueden efectivamente, ayudar a mejorar el desempeño de su trabajo diario. Creo sinceramente que he tenido mucha suerte de poder realizar un proyecto que puede ser puesto en marcha en dicho Ayuntamiento, y que tal vez sirva de palanca a esta administración u otras, para comenzar o continuar trabajando en la mejora y optimización de sus servicios gracias al análisis y estudio de los datos recogidos *(Business Intelligence).*

Como bien dice L. Cayuela en su libro sobre modelos lineales, la estadística comienza con un problema, continua con la recogida de datos, y termina con el análisis de los mismos, lo cual conduce a unas conclusiones sobre las hipótesis de partida.

Él cree que es un error muy común enredarse en análisis muy complejos, sin prestar atención a los objetivos que se persiguen. Por este mismo motivo, creo que el reto más grande que me ha supuesto este trabajo de fin de máster, fue hacerme la pregunta correcta que se quería contestar en este trabajo de fin de máster. Una vez concretado el objetivo que se deseaba obtener, era necesario confirmar que los datos recogidos por el Ayuntamiento de Sant Cugat eran apropiados o no para el análisis propuesto.

El camino que seguí para poder dar con las preguntas adecuadas y el objetivo final, fue el siguiente:

1. Comprender el problema de fondo y su contexto: para ello fue necesaria la primera reunión con los responsables de las oficinas de atención al cliente del Ayuntamiento de Sant Cugat. He de decir que me hubiera gustado tener la oportunidad de poder realizar reuniones periódicas con ellos mientras realizaba este proyecto, pues creo que me hubieran facilitado la comprensión de muchas dudas, que surgieron durante el desarrollo del mismo.

2. Comprender bien el objetivo de estudio. Primeramente, se hizo un análisis exploratorio de los datos, pero siendo cuidadosos con los análisis no dirigidos ya que podían desviar la atención del objetivo final.

3. Plantear el problema en términos estadísticos. Éste ha sido uno de los pasos más difíciles, e implica la formulación de hipótesis y modelos. Traducir el problema al lenguaje estadístico fue lo más costoso debido a la falta de experiencia sobre la materia.

En paralelo a la ejecución de los pasos anteriores fue necesario no perder de vista aspectos como: las unidades de medidas, posibles errores en los datos, valores extremos ó outliers que puedan distorsionar el resultado, etc.

De este modo, capítulo a capítulo, he ido progresando y avanzando todos los propósitos necesarios para cumplir de forma satisfactoria los objetivos definidos en el capítulo uno y transformar positivamente los procesos de atención al cliente en esta administración pública.

Finalmente, me gustaría remarcar que con este trabajo de fin de máster pongo punto y seguido a un gran año de máster en “La Salle, Universitat Ramon Llull”, que comencé con mucha ilusión en una nueva ciudad, recién finalizado mi grado de telecomunicaciones. Me siento muy feliz y orgullosa de haber adquirido grandes conocimientos, gracias a profesores excelentes y compañeros extraordinarios con los que he tenido la suerte de compartir mis clases.

Líneas futuras

Durante todo el proceso de ejecución de este trabajo de fin de máster, han ido surgiendo nuevas ideas y mejoras que pueden ser incluidas en un futuro en un nuevo análisis de los datos del Ayuntamiento de Sant Cugat, con el fin de conocer lo sucedido en años posteriores, mejorar la caracterización de las variables involucradas y poder optimizar aún más el proceso de atención al cliente en las oficinas del Ayuntamiento de Sant Cugat.

Al mismo tiempo que surgían nuevas ideas durante este periodo, también emergieron inconvenientes y problemas que aprendí a resolver para continuar correctamente el curso de este proyecto.

En un primer momento, una vez recibidos los datos del Ayuntamiento de Sant Cugat, no teníamos marcado un objetivo concreto para la realización de este TFM, sino que la idea que fijamos con el tutor fue, comenzar a analizar y explorar los datos, observarlos, tratarlos y con eso ir deduciendo puntos o ideas que nos llevaran a deducir cuál sería el objetivo de este trabajo fin de máster. Por este mismo motivo, fue muy relevante la primera reunión que mantuvimos con los responsables de la oficina de atención al ciudadano del Ayuntamiento de Sant Cugat para conocer de primera mano cuales eran los aspectos que más les preocupaban.

Gracias a esta reunión pudimos empezar a trabajar en diferentes ideas que, dieron lugar al desarrollo de este proyecto y de esta memoria.

Una vez finalizado este trabajo, han surgido nuevas ideas que se podrán implementar en una próxima versión de este proyecto o incluso que en un futuro otra persona, pueda continuar con este proyecto con el fin de mejorar los resultados obtenidos en el mismo.

Algunas de las nuevas ideas que pueden ser implementadas en el futuro, son las siguientes:

- En el capítulo número dos se observó en las tablas cuatro y cinco que, durante los seis años analizados, la oficina más visitada fue la denominada “Atención Ciudadana/Registro”. Del total de visitas registradas en el Ayuntamiento de Sant Cugat, aproximadamente un 88% del total fueron consultas en la oficina “Atención Ciudadana/Registro”.
  Por lo que los resultados obtenidos en este proyecto aproximarán mejor los datos que correspondan a esta oficina dado el alto porcentaje de visitas que contiene y queda pendiente para un futuro estudio, omitir las observaciones relativas a dicha oficina y realizar un nuevo estudio concreto para cada una de las otras oficinas de atención al cliente, existentes en el Ayuntamiento de Sant Cugat. De esta forma, se encontrarán resultados más personalizados para estas citadas oficinas.
- En el capítulo seis, en el apartado de caracterización de variables, se observa que la variable “tiempo de servicio”, aumenta considerablemente en el año 2016. Sería interesante poder buscar de dónde procede este aumento en el tiempo de servicio para dicho año, y buscar si tiene alguna relación, que parece predecible, con el aumento tan cuantioso que también aplica a la variable “tiempo de espera” para el mismo año 2016.
- Se sugiere mantener una nueva reunión con los responsables del Ayuntamiento de Sant Cugat para comentar con ellos todos los avances que se han realizado y tratar de buscar las causas que pudieron concluir en un aumento del “tiempo de servicio” en el año 2016.
- En el capítulo seis se explica y caracteriza a los diferentes gestores. Dado que es predecible que ciertos gestores solo atiendan en oficinas concretas del Ayuntamiento de Sant Cugat, es conveniente conocer esta información y así solo tratar la información de los gestores relativa a la oficina que se esté estudiando en el momento.
  Dado que actualmente no pudimos disponer de dicha clasificación de información, todo el proyecto se ha llevado a cabo suponiendo que todos los gestores atienden en cualquiera de las oficinas.
  Sería interesante diferenciar a los gestores por su oficina correspondiente, y así poder realizar un análisis exhaustivo de los datos. De este modo, aseguraremos que ciertas observaciones de una oficina concreta para las cuales el tiempo de servicio es muy alto por la propia naturaleza de dicha oficina, no alteren o influyan en los resultados obtenidos para otras oficinas.
- El modelo general construido para la variable “tiempo de espera”, podría ser probado en la variable “tiempo de servicio” y comprobar si tienen dependencias o no.

Existe todo un abanico de posibilidades dentro del *Business Intelligence,* que podrían ser incluidas en versiones posteriores de este trabajo fin de máster. Esto supondría que el proyecto creciera y si consecuentemente los resultados obtenidos afectasen positivamente al funcionamiento de la oficina del Ayuntamiento de Sant Cugat, podría convertirse en un proyecto de referencia dentro de la mejora y optimización de la atención al ciudadano en las administraciones públicas.

Creo fielmente que con estas nuevas mejoras u otras que algún compañero pueda aportar, este proyecto podría evolucionar de tal forma que se abriría camino en un mercado tan grande como lo es el de análisis de datos y toma de decisiones, en el que con poco tiempo y buena publicidad

podría llegar a un público muy amplio tanto del sector público como del sector privado, que desearán aprovecharse de las ventajas y facilidades que este proyecto les ofrece.

Bibliografía

Balmón, M. A. (2015). *Guía práctica de análisis de datos.* Sevilla: IFAPA.

Cayuela, L. (2014). *Modelos lineales: Regresión, ANOVA y ANCOVA.* Madrid: Area de Biodiversidad y Conservación, Universidad Rey Juan Carlos.

Chang, W. (2012). *Practical Recipes for Visualizing Data: R Graphics Cookbook.* United States of America: O`Reilly.

Chicana, R. C. (2014). Introducción al uso de R y R Commander para el. *R*, 128.

Estadística para todos. (2016). *Sotware Estadístico.* Obtenido de Estadística para todos: http://www.estadisticaparatodos.es/software/software_libre.html

Fugu Software Factory. (2015). *Análisis predictivo.* Obtenido de http://www.fugu.ec/productos-y-servicios/analisis-predictivo/

García Sabater, J. P. (2016). Aplicando Teoría de Colas en Dirección de Operaciones. En G. R. Valencia., *Teroría de colas.* Valencia: Universidad Politécnica de Valencia.

Guido Corradi. (16 de Octubre de 2014). *Por qué R es mejor que Excel (Psicología de datos).* Obtenido de https://psicologiadedatos.wordpress.com/2014/10/16/por-que-r-es-mejor-que-excel/

IBM Knowledge Center. (Enero de 2018). *IBM.* Obtenido de https://www.ibm.com/support/knowledgecenter/es/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/regression/idh_nlre.htm

Jimmy W. Maco Elera. (14 de Junio de 2017). *Cincode los mejores software de minería de datos de Código Libre y Abierto.* Obtenido de http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/

Marcos Singer, P. d.-W. (2008). UNA INTRODUCCIÓN A LA TEORÍA DE COLAS APLICADA A LA GESTIÓN DE SERVICIOS. *Revista ABANTE, VOL: 11, N°2*, 93-120. Obtenido de file:///C:/Users/jsantodomi001/Downloads/Singer.pdf

Marín, J. M. (s.f.). Introducción a los Procesos Estocásticos. En U. C. III. Madrid.

MathWorks. (s.f.). *Predicción de eventos futuros a partir de datos históricos.* Obtenido de Análisis predictivo: https://es.mathworks.com/discovery/analisis-predictivo.html

Microsoft: Developer Network. (Enero de 2018). *https://msdn.microsoft.com/es-es/library/dn705848.aspx?f=255&MSPPError=-2147217396*. Obtenido de https://msdn.microsoft.com/es-es/library/dn705848.aspx?f=255&MSPPError=-2147217396

Patrick Breheny, W. B. (23 de Junio de 2017). *Visualization of Regression Models.* Obtenido de Provides a convenient interface for constructing plots to visualize the fit of regression: https://cran.r-project.org/web/packages/visreg/visreg.pdf

Ruiz, M. C. (s.f.). *PROCESOS ESTOCÁSTICOS (Tema 5).* Obtenido de Universidad Politécnica de Cartagena: http://www.dmae.upct.es/~mcruiz/Telem06/Teoria/apuntes_procesos.pdf

Sinergia e Inteligencia de Negocio S.L. (s.f.). *Sinnexus*. Obtenido de Sinnexus: http://www.sinnexus.com/business_intelligence/

Soporte técnico de Minitab. (Junio de 2017). *Regresión no lineal.* Obtenido de https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/nonlinear-regression/understanding-nonlinear-regression/

W. N. Venables, D. M. (21 de Abril de 2017). *An Introduction to R.* Obtenido de Notes on R: A Programming Environment for Data Analysis and Graphics : https://cran.r-project.org/

ANEXOS

A. Anexo código R del análisis exploratorio de la base de datos (BBDD)

```
#-----------------------------------------------------#
## Julia Santo Domingo Gómez ##
#-----------------------------------------------------#
## TFM: "La Salle, Universitat Ramon Llull", ##
#-----------------------------------------------------#

setwd("C:/Users/User/Documents/MEGA/TFM/Draft de R")
getwd()
#-----------------------------------------------------#

#Cargamos historial
load("Project_1.RData")
#Guardamos historial
save.image("Project_1.RData")

#Limpiamos
rm(list=ls())

#Salimos
q(save="no")
#-----------------------------------------------------#
# LOADING THE DATABASE
Data<-read.csv(file="Datos_bbdd_new.csv", header = TRUE, sep=";")

# Revisamos la base de datos
str(Data)

#-----------------------------------------------------#
Para el aprendizaje de la herramienta R y Rstudio se han usado varios
libros de R y análisis de datos que están incluidos en la bibliografía
mostrada posteriormente a este anexo.


Exploratory Data analysis: Se ha comprobado que los datos estén bien
(que no haya outliers).
#-----------------------------------------------------#

#Vamos a ver el número de columnas que componen la BD y cuáles son
names(Data)

str(Data$Mes) #Para que siempre salgan los labels
Data$Mes<-factor(Data$Mes,      labels      =      c("Ene","Feb","Mar",
"Abr","May","Jun","Jul","Ago","Sep", "Oct", "Nov","Dic"))

str(Data$Año) #Para que siempre salgan los labels
Data$Año<-factor(Data$Año, labels =c("2011", "2012", "2013", "2014",
"2015", "2016"))

str(Data$Dia.semana) #Para que siempre salgan los labels
levels(Data$Dia.semana)
# Primero cambiamos las etiquetas de los dias de la semana
Data$Dia.semana<-factor(Data$Dia.semana, labels =c("Jueves", "Lunes",
"Martes", "Miércoles", "Sábado", "Viernes"))

#Ahora cambiamos el orden de los dias de la semana dentro del vector,
para que no se ordenen alfabeticamente
```

```
Data$Dia.semana<-factor(Data$Dia.semana,levels = c("Lunes", "Martes",
"Miércoles", "Jueves", "Viernes", "Sábado"))

#Levels es para ordenar los item dentro del factor

str(Data$Dia.1) #Para que siempre salgan los labels
#Data$Dia.1<-factor(Data$Dia.1,     labels     =c("1",     "2","3","4",
"5","6","7",   "8","9","10",   "11",   "12","13","14",   "15","16","17",
"18","19",       "20","21",       "22","23","24",       "25","26","27",
"28","29","30","31" ) )

str(Data$Codi.Gestor)#Para que siempre salgan los labels
#Data$Dia.1<-factor(Data$Codi.Gestor, labels =c("1", "2","3","4",
"5","6","7", "8","9","10", "11", "12","13","14", "15","16","17",
"18","19", "20","21", "22","23","24", "25","26","27",
"28","29","30","31","32","33","34", "35","36","37", "38","39",
"40","41", "42","43","44", "45","46","47", "48","49","50" ) )

#Data$X <- NULL #Esto es para borrar la columna X que no lo necesitamos
y que venía de la bbdd antigua del Ayuntamiento de Sant Cugat
Data$Tipologia.tràmites  #vemos todos los tipos de tramites que hay

str(Data$Tipologia.tràmites) #nos dice que es un factor de 79 niveles

length(unique(Data$Tipologia.tràmites)) #hay 79 servicios distintos

plot(x=Data$Cua..Oficina,y=Data$Tipologia.tràmites) #hay que instalar
el paquete ggplot2 para representar mejor los gráficos (ya que solo con
la función plot vemos que quedan muy escuetos)

library(ggplot2)

#Hasta ahora se ha trabajado la bbdd(base de datos) de forma que quede
ordenada y que nos permita realizar todos los análisis necesarios. A
partir de aquí se comienzan a realizar los diferentes cálculos que nos
interesan de cara a caracterizar la base de datos al completo:

#-----------------------------------------------------#
#De ahora en adelante se irán nombrando a las gráficas numéricamente
según se vayan obteniendo (Ejemplo: Gráfica 1, Gráfica 2, Graf3,
...Graf1000). Hay algunas gráficas que las he representado para buscar
información pero que no las he nombrado ya que he decidido no mostrarlas
en memoria, bien porque se ha encontrado otra gráfica que representa
mejor lo que se quería mostrar en primera instancia o bien porque no se
creen relevantes para formar parte de la memoria que se entregará.
#-----------------------------------------------------#

#Calculamos el número de atenciones en cada oficina por año ejemplo
reordenado Gráfica 1#
ggplot(data = Data,
       aes(x=reorder(Cua..Oficina,Cua..Oficina,
                     function(x)-length(x)))) +
  geom_bar(aes(fill=factor(Año)))+
  theme(strip.text=element_text(face='italic', size=12),
        axis.text.x=element_text(size=rel(0.9), angle=35, hjust=1,
vjust=1))+
  labs(x = "Oficina", y="N° de atenciones", title="N° de atenciones
por cada oficina en cada año")

#número de servicios por cada oficina: Se observa una leyenda muy grande
que impide ver el diagrama de barras correctamente
```

```
ggplot(data = Data, aes(x=Cua..Oficina))+
  geom_bar(aes(fill=Tipologia.tràmites))

#número de servicios por cada oficina sin leyenda: labels super
posicionados
ggplot(data = Data, aes(x=Cua..Oficina))+
  geom_bar(aes(fill=Tipologia.tràmites)) +
  theme(legend.position="none")

#número de servicios por cada oficina: sin leyenda y recolocamos los
labels de manera más ordenada Gráfica 2
ggplot(data = Data, aes(x=Cua..Oficina))+
  geom_bar(aes(fill=Tipologia.tràmites)) +
  theme(legend.position="none")+
  theme(strip.text=element_text(face='italic', size=12),
        axis.text.x=element_text(size=rel(0.9), angle=35, hjust=1,
vjust=1))
#Gráfica 2 ordenado de mayor a menor--> Añadir la leyenda
ggplot(data = Data, aes(x=reorder(Cua..Oficina,Cua..Oficina,
                                  function(x)-length(x))))+
  geom_bar(aes(fill=Tipologia.tràmites)) +
  theme(legend.position="none")+
  theme(strip.text=element_text(face='italic', size=12),
        axis.text.x=element_text(size=rel(0.9), angle=35, hjust=1,
vjust=1))+
  labs(x = "Oficina", y="N° de atenciones", title="N° de atenciones
por cada Oficina por trámites")

#A continuación se procede a la realización de un análisis temporal de
las diferentes variables que intervienen en el proceso de atención de
un usuario en las oficinas de atención al cliente del ayuntamiento de
San Cugat

#########################################################################
#Análisis temporal de la variable: número de atenciones
#########################################################################

#Gráfica Análisis temporal del número de atenciones Anualmente
(Gráfica 4)

ggplot(data=Data, aes(x=Año))+
  geom_bar() +
  labs(x= "Año",y="N° de atenciones", title="N° de atenciones totales
anualmente")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Gráfica Análisis temporal del número de atenciones mensualmente (graf
5)
ggplot(data=Data, aes(x=Mes))+
  geom_bar()+
  labs(x="Mes",y="N° de atenciones", title="N° de atenciones totales
mensualmente")+
  theme_bw()+
  theme(
```

```
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Gráfica Análisis temporal del número de atenciones mensualmente
Diferenciamos por año (graf 6)

Graf_Px_Mens23 <-
ggplot(data=Data, aes(x=Mes))+
  geom_bar(aes(fill=factor(Año)))+
  labs(x="",y="N° de atenciones", title="N° atenciones por mes y por
año", fill="Año")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())#+
    #axis.ticks = element_blank(),
    #legend.justification = c(1, 0),
    #legend.position = c(0.4, 0.84),
    #legend.direction = "horizontal")+
 #guides( title.position = "top", title.hjust = 0.5)

#visualizar en una nueva ventana
x11()
Graf_Px_Mens23

#código para guardar imagen (-->)para otras ocasiones, prefiero
exportarla directamente
ggsave(filename = "N_Px_Mes.jpg",plot =Graf_Px_Mens, path =
getwd(),width = 30 ,height = 30,units = "cm",dpi =150,scale = 0.6)

#Gráfica Análisis temporal del número de atenciones diarias (1...31)
#Gráfica Sin diferenciar año graf7
ggplot(data=Data, aes(x=factor(Dia.1)))+
  geom_bar()+
  labs(x="",y="N° de atenciones",title="N° de atenciones totales por
día de mes")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Nueva gráfica diferenciando año graf8

Graf_Px_Diaria <-
  ggplot(data=Data, aes(x=reorder(factor(Dia.1), factor(Dia.1),
function(x)-length(x)))) +
  geom_bar(aes(fill=factor(Año)))+
  labs(x="",y="N° de atenciones", title="N° de atenciones por día de
mes y por año")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
```

```
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

Graf_Px_Diaria


#Gráfica Análisis temporal del número de atenciones por día de la semana
(lunes...Domingo)

#Sin diferenciar año graf9

ggplot(data=Data, aes(x=Dia.semana))+
  geom_bar()+
  labs(x="",y="N° de atenciones", title="N° atenciones totales por día
de la semana")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#diferenciando por día de semana y año graf10
Graf_Px_Diaria <-
  ggplot(data=Data, aes(x=Dia.semana))+
  geom_bar(aes(fill=factor(Año)))+
  labs(x="",y="N° de atenciones", title="N° atenciones por día de la
semana y año")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

Graf_Px_Diaria

#########################################################################
#Análisis temporal de la variable: Tiempo espera
#########################################################################

#usamos la librería lubridate para poder trabajar con fechas y tiempos
library(lubridate)

#una vez que tenemos el tiempo en formato HMS ya podríamos trabajar
con gráficos
Data$Tiempo.espera2 <- hms(Data$Tiempo.espera)
Data$Tiempo.espera2

# Transformamos el tiempo de espera de segundos a minutos
# primero sacamos la información como duración (as.duration) para que
esté en segundos, luego nos dejamos solo los valores numéricos de los
segundos (as.numeric) y finalmente dividimos en 60 y obtenemos minutos
Data$Tiempo.espera3 <- as.numeric(as.duration(Data$Tiempo.espera2))/60

#Gráfica Análisis temporal de la variable tiempo de espera Anualmente
EL eje y estaba en escala logarítmica pero finalmente lo he hecho con
```

```
la variable "tiempo de espera 3" que esta en minutos ya que se
facilita la visualización. Se ven los outliers de cada año en un color
gris mas claro: Graf11

ggplot(data=Data, aes(x=Año, y=Tiempo.espera3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  #scale_y_log10()+
  labs(y="Tiempo espera (minutos)", title="Analisis temporal de la
variable Tiempo de Espera anualmente")+
  geom_hline(yintercept = 120, color="red")+
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Graf11_violines_año
ggplot(data=Data, aes(x=Año, y=Tiempo.espera3))+
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.2)+
  labs(y="Tiempo espera (minutos)", title="Analisis temporal Tiempo de
Espera anualmente: Violines")+
  geom_hline(yintercept = 60, color="blue")+
  geom_hline(yintercept = 120, color="red")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Se ve que el tiempo de espera aumenta a partir del 2014

#Gráfica Análisis temporal de la variable tiempo de espera mensualmente
EL eje y estaba en escala logarítmica pero finalmente lo he hecho con
la variable denominada "tiempo de espera 3" que está en minutos ya que
se facilita la visualización. Se ven los outliers de cada año en un
color gris más claro: graf12

ggplot(data=Data, aes(x=Mes, y=Tiempo.espera3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  #scale_y_log10()+
  labs(y="Tiempo espera (minutos)", title="Analisis temporal de la
variable Tiempo de Espera mensualmente")+
  geom_hline(yintercept = 120, color="red")+
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf12_violines_mes
```

```
ggplot(data=Data, aes(x=Mes, y=Tiempo.espera3))+
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.2)+
  labs(y="Tiempo espera (minutos)", title="Analisis Tiempo de Espera
por mes: Violines")+
  geom_hline(yintercept = 60, color="blue")+
  geom_hline(yintercept = 120, color="red")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())


#Gráfica Análisis temporal de la variable tiempo de espera porcada día
de mes. El eje y en minutos ya que se facilita la visualización. Se ven
los outliers de cada año en un color gris más claro: graf13

ggplot(data=Data, aes(x=factor(Dia.1), y=Tiempo.espera3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  #scale_y_log10()+
  labs(y="Tiempo espera (minutos)", title="Analisis temporal de la
variable Tiempo de Espera por día de mes")+
  geom_hline(yintercept = 120, color="red")+
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Gráfica Análisis temporal de la variable tiempo de espera por cada día
de semana El eje y está en minutos ya que se facilita la visualización.
Se ven los outliers de cada año en un color gris más claro: graf14

ggplot(data=Data, aes(x=Dia.semana, y=Tiempo.espera3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  #scale_y_log10()+
  labs(y="Tiempo espera (minutos)", title="Analisis temporal de la
variable Tiempo de Espera por día de semana")+
  geom_hline(yintercept = 120, color="red")+
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Graf14_viloines_diasemana
ggplot(data=Data, aes(x=Dia.semana, y=Tiempo.espera3))+
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.2)+
```

```
  labs(y="Tiempo espera (minutos)", title="Analisis Tiempo de Espera
por dia de semana: Violines")+
  geom_hline(yintercept = 60, color="blue")+
  geom_hline(yintercept = 120, color="red")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#####################################################################
#Análisis temporal de la variable: Código de gestor
#####################################################################

#Gráfica del número de usuarios que atiende cada gestor. Los ordenamos
de mayor a menor número de personas atendidas (Gráfica 20)

ggplot(data=Data,
aes(x=reorder(factor(Codi.Gestor),factor(Codi.Gestor),
                                  function(x)-length(x))))+
  geom_bar() +
  labs(y="N° de atenciones", title="N° atenciones totales que ha
realizado cada codigo de gestor", x="Codigo de gestor")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Gráfica Análisis temporal del número de personas que atiende cada
gestor anualmente

# graf 21
Graf_Codi_Año <-
  ggplot(data=Data,
aes(x=reorder(factor(Codi.Gestor),factor(Codi.Gestor),
                                    function(x)-length(x))))+
  geom_bar()+
  labs(x="",y="N° de atenciones", title="Comparación de atenciones
realizadas por cada gestor por años")+
  theme_bw()+
  facet_wrap(~Año)+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

Graf_Codi_Año

#graf21 bis

ggplot(data=Data,
aes(x=reorder(factor(Codi.Gestor),factor(Codi.Gestor), function(x) -
length(x)))) +
  geom_bar(aes(fill=factor(Año)))+
```

```
  labs(x="",y="N° de atenciones", title= "N° atenciones totales que ha
realizado cada codigo de gestor")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Gráfica Análisis temporal del número de personas que atiende cada gestor
anualmente
# graf 22
# Graf_Codi_Año1 <-
#   ggplot(data=Data, aes(x=Año))+
#   geom_bar(aes(fill=factor(Codi.Gestor)))+
#   labs(x="",y="N° de Px")+
#   theme_bw()+
#   theme(
#     axis.title.x = element_blank(),
#     #axis.title.y = element_blank(),
#     panel.grid.major = element_blank(),
#     #panel.border = element_blank(),
#     panel.background = element_blank())
#
# Graf_Codi_Año1

#Gráfica Análisis temporal del número de personas que atiende cada gestor
mensualmente graf23
Graf_Codi_Año2 <-
  ggplot(data=Data,
aes(x=reorder(factor(Codi.Gestor),factor(Codi.Gestor), function(x) -
length(x)))) +
  geom_bar(aes(fill=factor(Mes)))+
  labs(x="",y="N° de atenciones", title= "N° atenciones totales que ha
realizado cada codigo de gestor mensualmente")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

Graf_Codi_Año2
###################
#La Gráfica del análisis temporal del código de gestor por cada día de
mes no va a presentarse en el tfm ya que no se llega a distinguir
suficientemente bien las diferencias existentes. Para solventar esto lo
representaremos por día de semana/mes/año

#Graf_Codi_Año4 <-
  ggplot(data=Data, aes(x=factor(Codi.Gestor)))+
  geom_bar(aes(fill=factor(Dia.1)))+
  labs(x="",y="N° de Px")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
```

```
    panel.background = element_blank())

#Graf_Codi_Año4
###################

#Gráfica Análisis temporal del número de personas que atiende cada
gestor por día de la semana graf24
Graf_Codi_Año3 <-
  ggplot(data=Data,
aes(x=reorder(factor(Codi.Gestor),factor(Codi.Gestor), function(x) -
length(x)))) +
  geom_bar(aes(fill=factor(Dia.semana)))+
  labs(x="",y="N° de atenciones", title="N° atenciones totales que ha
realizado cada codigo de gestor por día de la semana")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

Graf_Codi_Año3

#############################################################
#Análisis temporal de la variable: Tiempo de servicio
#############################################################

#una vez que tenemos el tiempo en formato HMS ya podríamos trabajar con
gráficos
Data$Tiempo.servicio2 <- hms(Data$Tiempo.servicio)
Data$Tiempo.servicio2
Data$Tiempo.servicio3                                                <-
as.numeric(as.duration(Data$Tiempo.servicio2))/60

#Gráfica Análisis temporal de la variable tiempo de servicio Anualmente
EL eje y está en escala logarítmica ya que se facilita la visualización.
Se ven los outliers de cada año en un color gris mas claro: graf30


ggplot(data=Data, aes(x=Año, y=as.numeric(Tiempo.servicio2)))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  scale_y_log10()+
  labs(y="Tiempo espera (s-1 log 10)", title="Analisis temporal de la
variable Tiempo de Servicio anualmente")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf30 bis en minutos, sin log
ggplot(data=Data, aes(x=Año, y=Tiempo.servicio3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (minutos)", title="Analisis temporal de la
variable Tiempo de Servicio anualmente")+
  geom_hline(yintercept = 120, color="red")+
```

```
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf30_bis_minutos_violines
ggplot(data=Data, aes(x=Año, y=Tiempo.servicio3))+
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.2)+
  labs(y="Tiempo servicio (minutos)", title="Analisis Tiempo de
Servicio por año: Violines")+
  geom_hline(yintercept = 60, color="blue")+
  geom_hline(yintercept = 120, color="red")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Gráfica Análisis temporal de la variable tiempo de servicio mensualmente
EL eje y está en escala logarítmica ya que se facilita la visualización.
Se ven los outliers de cada año en un color gris más claro: graf31

ggplot(data=Data, aes(x=Mes, y=as.numeric(Tiempo.servicio2)))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  scale_y_log10()+
  labs(y="Tiempo espera (s-1 log 10)", title="Analisis temporal de la
variable Tiempo de Servicio mensualmente")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf31_bis_ sin logaritmos
ggplot(data=Data, aes(x=Mes, y=Tiempo.servicio3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (minutos)", title="Analisis temporal de la
variable Tiempo de Servicio mensualmente")+
  geom_hline(yintercept = 120, color="red")+
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf31_bis_minutos_violines
ggplot(data=Data, aes(x=Mes, y=Tiempo.servicio3))+
```

```
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.2)+
  labs(y="Tiempo servicio (minutos)", title="Analisis Tiempo de
Servicio por mes: Violines")+
  geom_hline(yintercept = 60, color="blue")+
  geom_hline(yintercept = 120, color="red")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Gráfica Análisis temporal de la variable tiempo de servicio según día
de mes El eje y está en escala logarítmica ya que se facilita la
visualización. Se ven los outliers de cada año en un color gris más
claro: graf32

ggplot(data=Data, aes(x=factor(Dia.1),
y=as.numeric(Tiempo.servicio2)))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  scale_y_log10()+
  labs(y="Tiempo servicio (s-1 log 10)", title="Analisis temporal de
la variable Tiempo de Servicio por día de mes")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf32_bis_sin_log
ggplot(data=Data, aes(x=factor(Dia.1), y=Tiempo.servicio3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo servicio (minutos)", title="Analisis temporal de la
variable Tiempo de Servicio por día de mes")+
  geom_hline(yintercept = 120, color="red")+
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#no se aprecia ninguna tendencia

#Gráfica Análisis temporal de la variable tiempo de servicio según el
gestor El eje y está en escala logarítmica ya que se facilita la
visualización. Se ven los outliers de cada año en un color gris más
claro: Graf 300

ggplot(data=Data, aes(x=factor(Codi.Gestor),
y=as.numeric(Tiempo.servicio2)))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
```

```
  scale_y_log10()+
  labs(y="Tiempo servicio (s-1 log 10)", title="Analisis temporal de
la variable Tiempo de Servicio por gestor")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf_300_sin_log
ggplot(data=Data, aes(x=factor(Codi.Gestor), y=Tiempo.servicio3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo servicio (minutos)", title="Analisis temporal de la
variable Tiempo de Servicio por gestor")+
  geom_hline(yintercept = 120, color="red")+
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf300 violines, lo descarto ya que no se aprecia nada

#Gráfica Análisis temporal de la variable tiempo de servicio según día
de semana El eje y está en escala logarítmica ya que se facilita la
visualización. Se ven los outliers de cada año en un color gris más
claro: graf33

ggplot(data=Data, aes(x=Dia.semana, y=as.numeric(Tiempo.servicio2)))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  scale_y_log10()+
  labs(y="Tiempo servicio (s-1 log 10)", title="Analisis temporal de
la variable Tiempo de Servicio por día de semana")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf33 sin logaritmos
ggplot(data=Data, aes(x=Dia.semana, y=Tiempo.servicio3))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo servicio (minutos)", title="Analisis temporal de la
variable Tiempo de Servicio por día de semana")+
  geom_hline(yintercept = 120, color="red")+
  geom_hline(yintercept = 60, color="blue")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
```

```
    #panel.border = element_blank(),
    panel.background = element_blank())


#el tiempo de servicio no se ve afectado por el día de la semana, pero
tiene una variabilidad muy grande

summary(as.numeric(Data$Tiempo.servicio2))

#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    #0.0    93.0   315.0   495.2   691.0 15108.0      21


summary(Data$Tiempo.servicio3)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 # 0.000   1.550   5.283   8.320  11.617 251.800      22

#mismo que anterior, pero grafico de violines graf34
ggplot(data=Data, aes(x=Dia.semana, y=Tiempo.servicio3))+
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.2)+
  labs(y="Tiempo servicio (minutos)", title="Analisis Tiempo de
Servicio por día de semana: Violines")+
  geom_hline(yintercept = 60, color="blue")+
  geom_hline(yintercept = 120, color="red")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#########################################################################
#Por último para finalizar esta sección se caracterizarán las variables
"tiempo de espera" y "tiempo de servicio"
#-------------------------------------------------------#
Modelizar la variable tiempo de espera
#-------------------------------------------------------#

str(Data$Tiempo.espera3)
summary(Data$Tiempo.espera2)
summary(Data$Tiempo.servicio2)
str(Data$Tiempo.espera3)
#graf50
hist(Data$Tiempo.espera3, breaks=100)

#graf 51
# ggplot(Data,aes(x=Tiempo.espera3, y=..density..))+
#   geom_histogram(fill="pink", colour="grey60", size=0.2, na.rm =
TRUE)+
#   geom_density(na.rm = TRUE)
#   #xlim(0,120)

#graf52
ggplot(Data,aes(x=Tiempo.espera3))+
  geom_histogram(fill="pink", colour="grey60", size=0.2, na.rm =
TRUE)+
  geom_freqpoly(binwidth=4)+
```

```
  labs(x="Tiempo de espera (minutos)", y="número de observaciones",
title="Histograma de frecuencias y curva de densidad para el tiempo de
espera")+
  theme_bw()+
  theme(
    #axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#-----------------------------------------------------#
#Modelizar la variable tiempo de servicio
#-----------------------------------------------------#
#graf53 histograma tiempo de servicio
hist(Data$Tiempo.servicio3, breaks=100)

#graf54
ggplot(Data,aes(x=Tiempo.servicio3))+
  geom_histogram(fill="pink", colour="grey60", size=0.2, na.rm =
TRUE)+
  geom_freqpoly(binwidth=4)+
  labs(x="Tiempo de servicio (minutos)", y="número de observaciones",
title="Histograma de frecuencias y curva de densidad para el tiempo de
servicio")+
  theme_bw()+
  theme(
    #axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())
#-----------------------------------------------------#
#Fin primera parte: Análisis exploratorio de los datos.
#-----------------------------------------------------#
```

B. Anexo código R del estudio de la variable tiempo de espera para el caso del Ayuntamiento de Sant Cugat planteado en este proyecto.

```
#-----------------------------------------------------#
## Julia Santo Domingo Gómez ##
#-----------------------------------------------------#
## TFM : "La Salle, Universitat Ramon Llull", ## Análisis estadístico
#-----------------------------------------------------#
setwd("C:/Users/User/Documents/MEGA/TFM/Draft de R")
getwd()
#-----------------------------------------------------#
#Cargamos historial
load("Project_1_Analisis_estadistico.RData")

#Guardamos historial
save.image("Project_1_Analisis_estadistico.RData")

#Limpiamos
rm(list=ls())

#Salimos
q(save="no")
```

```
#-----------------------------------------------------#
# LOADING THE DATABASE
Data<-read.csv(file="Datos_bbdd_new.csv", header = TRUE, sep=";")

# Revisamos la base de dtos
str(Data)

#-----------------------------------------------------#
#Exploratory Data analysis: Se ha comprobado que los datos estén bien
(que no haya outliers)
#-----------------------------------------------------#
#Vamos a ver el número de servicios por mesa
names(Data)

str(Data$Mes) #Para que siempre salgan los labels
Data$Mes<-factor(Data$Mes, labels = c("Ene","Feb","Mar",
"Abr","May","Jun","Jul","Ago","Sep", "Oct", "Nov","Dic"))

str(Data$Año) #Para que siempre salgan los labels
Data$Año<-factor(Data$Año, labels =c("2011", "2012", "2013", "2014",
"2015", "2016"))

str(Data$Dia.semana) #Para que siempre salgan los labels
levels(Data$Dia.semana)
# Primero cambiamos las etiquetas de los días de la semana
Data$Dia.semana<-factor(Data$Dia.semana, labels =c("Jueves", "Lunes",
"Martes", "Miércoles", "Sábado", "Viernes"))
#Ahora cambiamos el orden de los días de la semana dentro del vector,
para que no los ordene alfabeticamente
Data$Dia.semana<-factor(Data$Dia.semana,levels = c("Lunes", "Martes",
"Miércoles", "Jueves", "Viernes", "Sábado"))
#Levels es para ordenar los item dentro del factor

#-----------------------------------------------------#
#Análisis estadístico
#-----------------------------------------------------#

#1. Distribución de los datos

#Instalamos paquete lubridate

Data$Tiempo.espera2 <- hms(Data$Tiempo.espera)
Data$Tiempo.espera2
Data$Tiempo.espera3  <-  as.numeric(as.duration(Data$Tiempo.espera2))/60
#Tiempo de espera 3 : minutos

#Representamos el Histograma de tiempo de espera. #Grafica52 (Script
anterior) (Eje X en minutos con Tiempo de espera 3)

summary(Data$Tiempo.espera3)
#summary(as.numeric(as.duration(Data$Tiempo.espera2)))

#Creamos una nueva columna llamada tiempo de espera 4 en la que no
aparezcan tiempos de espera =0. A aquellos valores = a 0, se les asignará
NA.
summary(Data$Tiempo.espera3[Data$Tiempo.espera3>0])
Data$Tiempo.espera4 <- Data$Tiempo.espera3
Data$Tiempo.espera4[Data$Tiempo.espera4==0] <-NA

summary(Data$Tiempo.espera4)
```

```
#Observamos que el mínimo tiempo de espera en esta nueva columna es de
0,02 minutos. Ya no hay tiempos = a 0.

#Grafica100 es el mismo histograma que el anterior pero con tiempo de
espera 4 (sin ceros)
ggplot(Data,aes(x=Tiempo.espera4))+
  geom_histogram(fill="pink", colour="grey60", size=0.2, na.rm =
TRUE)+
  geom_freqpoly(binwidth=4)+
  labs(x="Tiempo de espera (minutos)", y="Número de observaciones",
title="Histograma de frecuencias y curva de densidad para el tiempo de
espera")+
  theme_bw()+
  theme(
    #axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#1. Buscamos presencia de outliers

#Grafica101 se ven muchos otliers, hay outliers en la variable de
respuesta, tenemos que ver que hacer con estos outliers (para que no
distorsionen el resultado)
ggplot(data=Data, aes(x="Tiempo Espera", y=Tiempo.espera4))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (minutos)", title="Análisis temporal de la
variable Tiempo de Espera")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#graf101_log --> Usaremos logaritmos para evitar que los outliers
distorsionen tanto el resultado
ggplot(data=Data, aes(x="Tiempo Espera", y=Tiempo.espera4))+
  #geom_point(na.rm=TRUE)+
  scale_y_log10()+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (log minutos)", title="Análisis temporal de la
variable Tiempo de Espera")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())
#-----------------------------------------------------#
#Creamos una nueva columna: Tiempo de espera 5:Sera el log en base 10
del tiempo de espera 4 (en minutos y sin 0)
Data$Tiempo.espera5 <- log10(Data$Tiempo.espera4)

#La variable respuesta será el Tiempo de espera 5 (es el log en base 10
del tiempo de espera 4)
#-----------------------------------------------------#
```

```
#2 Análisis estadístico de la variable tiempo de espera:
#2.1 Todas las variables explicativas son categóricas (ordinales y
nominales) por lo que vamos a asumir que no hay colinearidad.
#2.2 Pairs:
#Se van a representar las gráficas de la variable respuesta "Tiempo de
espera" con cada una de las variables explicativas:
#-----------------------------------------------------#
# Tiempo de espera - Día de la semana
#-----------------------------------------------------#
#Gráfica103:  tiempo de espera con dia de la semana (Tiempo espera 4:
minutos y sin ceros)

ggplot(data=Data, aes(x=Dia.semana, y=Tiempo.espera4))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (minutos)", title="Análisis temporal de la
variable Tiempo de Espera por día de semana")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#La gráfica la mostramos con y sin logaritmo

#Graf103_log (Tiempo de espera 5= Log(Tiempo de espera 4))
ggplot(data=Data, aes(x=Dia.semana, y=Tiempo.espera5))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (log10 minutos)", title="Análisis temporal de
la variable Tiempo de Espera por día de semana")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#----------#

#Test estadístico que si lo hacemos con la variable Tiempo de espera5
(Logaritmo)

#Modelo1_DiaSemana<-glm(Tiempo.espera5~Dia.semana,     family=gaussian,
data=Data)
#Hemos probado a realizar el test con familia de poisson pero no salió
correctamente porque al usar el logaritmo de la variable tiempo
respuesta: salen negativos, y Possion es sin negativos
#Lo anterior que es un glm con familia gasussiana (normal) es lo mismo
que un lm sin familia

Modelo1_DiaSemana<-lm(Tiempo.espera5~Dia.semana, data=Data) #LM hace el
modelo lineal (compara cada día con un valor por defecto: este resultado
lo entrega el summary del modelo lineal) (lunes= primer nivel de la
variable). Pero si queremos obtener solo el Análisis de la varianza: se
utiliza Anova de ese modelo: para observar si el efecto de la variable
día de semana es importante o no sobre el Tiempo de espera. Este Análisis
```

```
de anova es suficiente para demostrar que hay diferencia entre los días
de la semana y la variable respuesta. El test anova muestra un Pvalor
muy pequeño cercano a cero (se descarta la hipótesis nula) por lo que
se demuestra que existe un efecto del día de la semana sobre el tiempo
de espera. El tiempo de espera no es igual para todos los días de la
semana

anova(Modelo1_DiaSemana) #esto es para que el resultado se muestre
como una tabla de anova
summary(Modelo1_DiaSemana)
Modelo1_DiaSemana_2<-aov(Tiempo.espera5~Dia.semana, data=Data)
summary(Modelo1_DiaSemana_2)

#Ahora se hacen las comparaciones de todos con todos, para saber que día
es distinto a que día. El test a posteriori solo reconoce AOV (aov modelo
ANOVA)
posthoc_Diasemana   <-   TukeyHSD(x=Modelo1_DiaSemana_2,   'Dia.semana',
conf.level=0.95)

posthoc_Diasemana #Compara todos los días con todos. El pvalor es 0 por
lo que significa que son todos diferentes. La pareja jueves -martes casi
no tiene diferencias significativas (porque lo comparamos con la
probabilidad de 0,05 entonces al ser el Pvalor=0,044 está en el límite
de no ser significativo)
plot(posthoc_Diasemana) #graf104 como la pareja martes-jueves hemos
comprobado que no tiene diferencias significativas vemos en el plot que
está tocando la línea de no diferencias (no importa lo que haya a la
izquierda o derecha, sino que no toque la línea de no diferencias)

#-----------------------------------------------------#
# Tiempo de espera - Mes
#-----------------------------------------------------#

#Gráfica 105 : tiempo de espera con mes (tiempo de espera 4: minutos y
sin 0)

ggplot(data=Data, aes(x=Mes, y=Tiempo.espera4))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (minutos)", title="Análisis temporal del
Tiempo de Espera por mes")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#La gráfica la mostramos con y sin logaritmo

#Graf105_log (tiempo de espera 5: log (tiempo espera 4))
ggplot(data=Data, aes(x=Mes, y=Tiempo.espera5))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (log10 minutos)", title="Análisis temporal del
Tiempo de Espera por mes")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
```

```
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#----------#

#Modelo1_Mes<-lm(Tiempo.espera5~Mes, data=Data)
anova(Modelo1_Mes) #si que es significativo el efecto del mes en la variable respuesta (tiempo de espera) E=existen diferencias significativas entre los niveles de la variable mes sobre la variable tiempo de espera
summary(Modelo1_Mes) # si tiene diferencias significativa
Modelo1_Mes_2<-aov(Tiempo.espera5~Mes, data=Data) # la diferencia entre los dos modelos el primero es para regresión lineal si la variable respuesta era continua, pero nuestra variable respuesta es discreta, por lo que usamos anova y no usamos el primer modelo
summary(Modelo1_Mes_2)
posthoc_mes <- TukeyHSD(x=Modelo1_Mes_2, 'Mes', conf.level=0.95)
posthoc_mes
plot(posthoc_mes) #Graf106

#-----------------------------------------------------#
# Tiempo de espera - Año
#-----------------------------------------------------#

#Gráfica 107 tiempo de espera con año (tiempo de espera 4: minutos y sin 0)
ggplot(data=Data, aes(x=Año, y=Tiempo.espera4))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (minutos)", title="Análisis temporal del Tiempo de Espera por año")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#La gráfica la mostramos con y sin logaritmo

#Graf107_log (tiempo de espera 5: log (tiempo espera 4))
ggplot(data=Data, aes(x=Año, y=Tiempo.espera5))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (log10 minutos)", title="Análisis temporal del Tiempo de Espera por año")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#----------#
#Modelo1_Año<-lm(Tiempo.espera5~Año, data=Data)
#anova(Modelo1_Año)
#summary(Modelo1_Año)
Modelo1_Año_2<-aov(Tiempo.espera5~Año, data=Data)
```

```
summary(Modelo1_Año_2)
posthoc_año <- TukeyHSD(x=Modelo1_Año_2, 'Año', conf.level=0.95)
posthoc_año #2014 y 2013 no tienen diferencia significativa
plot(posthoc_año) #graf_108

# Para poder predecir en años posteriores, dado que el mes y el día de
la semana son variables circulares, sirve con el modelo anova, pero para
el año dado que no es circular será mejor considerar esta variable como
una variable numérica (para así poder predecir)

Data$Año #Es un factor y por eso tiene niveles
#Creamos una nueva variable/columna que sea numérico y no factor
Data$AñoNumerico<- as.numeric(as.character(Data$Año))
summary(Data$AñoNumerico)

#Hacemos de nuevo el gráfico con la variable numérica de año, lo hacemos
con geom point porque es numérico y no sirve el boxplot
####ojo con esta gráfica que peta

ggplot(data=Data, aes(x=AñoNumerico, y=Tiempo.espera5))+
  geom_point(na.rm=TRUE)+
  geom_smooth(na.rm=TRUE, method = "lm",formula = y~poly(x,2))+
  #geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (log10 minutos)", title="Análisis temporal del
Tiempo de Espera por año")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#Una vez obtenido el gráfico dispondremos del modelo lineal para poder
predecir, y este modelo lineal habrá que ejecutarlo

Modelo3_Año<-lm(Tiempo.espera5~AñoNumerico, data=Data)
summary(Modelo3_Año) #el summary es distinto que el del anova: ahora
aparece y=a+bx --> que el intercepto sea significativo (es casi 0 )
quiere decir que no pasa por el origen. La variable respuesta que es el
año si es significativo, lo que quiere decir que no tienen pendiente
igual a cero (la pendiente es0,07) a cada incremento de año hay un 0,07
mas del logaritmo del tiempo de espera : log10(0.07): -1.154902.

Modelo4_Año<-lm(Tiempo.espera5~AñoNumerico+I(AñoNumerico^2), data=Data)
summary(Modelo4_Año) # este modelo tiene mayor poder explicativo ya que
el AIC es menor cuando se compara con el modelo lineal sencillo, este
sería el modelo optimo y explica el 3,4 % de los datos (Multiple R-
squared:  0.03369,)

Modelo5_Año<-
lm(Tiempo.espera5~AñoNumerico+I(AñoNumerico^2)+I(AñoNumerico^3),
data=Data) # este modelo falla al ajustar el cúbico del año numérico,
por eso sabemos que la gráfica tiene pendiente positiva  pero que no
crece al infinito porque el cuadrado del tiempo tiene pendiente negativa
(log(de una pendiente positiva))

summary(Modelo5_Año)

AIC(Modelo3_Año, Modelo4_Año, Modelo5_Año) #df      AIC Modelo3_Año  3
845853.4 Modelo4_Año  4 842567.0 Modelo5_Año  4 842567.0
```

```
#-----------------------------------------------------#
# Tiempo de espera - Día de mes
#-----------------------------------------------------#

#gráfica 109 tiempo de espera con dia mes  (tiempo de espera 4:
minutos y sin 0)

Data$Dia.2 <- factor(Data$Dia.1)

ggplot(data=Data, aes(x=factor(Dia.1), y=Tiempo.espera4))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (minutos)", title="Análisis temporal del
Tiempo de Espera por dia de mes")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#La gráfica la mostramos con y sin logaritmo para compararlo

#Graf109_log (tiempo de espera 5: log (tiempo espera 4))
ggplot(data=Data, aes(x=factor(Dia.1), y=Tiempo.espera5))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (log10 minutos)", title="Análisis temporal del
Tiempo de Espera por dia de mes")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#----------#
#Modelo1_DiaMes<-lm(Tiempo.espera5~Dia.1, data=Data)
#anova(Modelo1_DiaMes)
#summary(Modelo1_DiaMes)
Modelo1_DiaMes_2<-aov(Tiempo.espera5~Dia.2, data=Data)
summary(Modelo1_DiaMes_2)

posthoc_Diames <- TukeyHSD(x=Modelo1_DiaMes_2, 'Dia.2',
conf.level=0.95)
posthoc_Diames

plot(posthoc_Diames) #graf_110

#Hay días que tienen significancia y hay otros que no, no muestra
todos.
##Si considerásemos la variable día del mes como una variable
continua, podríamos ver si hay (o no) tendencias dentro del mes.
Modelo2_DiaMes_2 <- lm(Tiempo.espera5~Dia.1, data=Data)
summary(Modelo2_DiaMes_2)

Modelo3_DiaMes_2 <- lm(Tiempo.espera5~Dia.1+I(Dia.1^2), data=Data)
summary(Modelo3_DiaMes_2)
```

```
Modelo4_DiaMes_2 <- lm(Tiempo.espera5~Dia.1+I(Dia.1^2)+I(Dia.1^3),
data=Data)
summary(Modelo4_DiaMes_2)

AIC(Modelo2_DiaMes_2,Modelo3_DiaMes_2,Modelo4_DiaMes_2)

# El modelo 4 (cúbico) es el que mejor resultado da e intenta ajustar
una curva polinómica donde se observa que el tiempo de espera disminuye
hacia fin de mes, per la disminución NO es lineal. A principio y a final
de mes (los veinti poco) el tiempo de espera es menos y esto lo respalda
el modelo cúbico

#-----------------------------------------------------#
# Tiempo de espera - Código de gestor
#-----------------------------------------------------#

#Gráfica 111 tiempo de espera con código gestor (tiempo de espera 4:
minutos y sin 0)
ggplot(data=Data, aes(x=factor(Codi.Gestor), y=Tiempo.espera4))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (minutos)", title="Análisis temporal del
Tiempo de Espera por gestor")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#La gráfica la mostramos con y sin logaritmo

#Graf111_log (tiempo de espera 5: log (tiempo espera 4))
ggplot(data=Data, aes(x=factor(Codi.Gestor), y=Tiempo.espera5))+
  #geom_point(na.rm=TRUE)+
  geom_boxplot(na.rm = TRUE, outlier.colour = "gray70")+
  labs(y="Tiempo espera (log10 minutos)", title="Análisis temporal del
Tiempo de Espera por gestor")+
  theme_bw()+
  theme(
    axis.title.x = element_blank(),
    #axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    #panel.border = element_blank(),
    panel.background = element_blank())

#----------#

Data$Codi.Gestor2<-factor(Data$Codi.Gestor)

#Modelo1_Gestor<-lm(Tiempo.espera5~Codi.Gestor, data=Data)
#anova(Modelo1_Gestor)
#summary(Modelo1_Gestor)
Modelo1_Gestor_2<-aov(Tiempo.espera5~Codi.Gestor2, data=Data)
summary(Modelo1_Gestor_2)

posthoc_gestor <- TukeyHSD(x=Modelo1_Gestor_2, 'Codi.Gestor2',
conf.level=0.95)
posthoc_gestor
```

```
plot(posthoc_gestor) #graf_112

#Hay gestores que son significativamente diferentes. Por ejemplo: el 16
y el 6 tienen diferencia significativa, pero entre el 13 y el 6 no hay
diferencias significativas (tienen similar tiempo espera). Ojo por que
los gestores pueden venir de servicios distintos que ya de por si tengan
tiempos más o menos altos de espera
#Línea futura: podrían compararse los gestores que vinieran del mismo
servicio

#-----------------------------------------------------------------------#
# Modelo general que explique la variable respuesta: Tiempo de espera
#-------------------------------------------------------#

#MODELO: unir todo lo que se ha obtenido de las distintas variables en
un mismo código: backward selection criteria (porque hemos analizado
todas las variables explicativas y todas ellas son significativas para
el modelo)

MODELOGENERAL<-   lm(Data$Tiempo.espera5~   Data$Dia.semana   +Data$Mes
+Data$AñoNumerico+I(Data$AñoNumerico^2)+Data$Dia.2+Data$Codi.Gestor2)
anova(MODELOGENERAL)

#Pondremos la tabla obtenida en R en la memoria como resumen del modelo
que representa la variable respuesta "tiempo de espera". ¿Qué dicen
estos datos? Dicen que todas las variables incluidas en el modelo tienen
significancia. Este modelo puede no estar representado la misma
información en conjunto, que cuando analizamos cada variable por
separado. Pero claramente este modelo explica la variable "tiempo de
espera".

#Una vez llegados a este punto, habríamos conseguido finalizar el último
objetivo de este proyecto.

#¿Qué se podría hacer ahora? Evaluar el efecto de cada variable sobre
el tiempo de espera manteniendo el efecto de las otras variables
contraste (pueden tener valor 0), pueden tener el valor del promedio,
...con el paquete visreg.


#Línea futura: Podríamos probar el modelo general para el tiempo de
servicio y ver si tienen dependencias o no
#Inspeccionar el modelo por cada una de las variables


#-----------------------------------------------------#
#Fin análisis estadístico.
#-----------------------------------------------------#
```