UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Análisis de los datos de emisiones en la ciudad de Valencia basados en los datos del tráfico de los años 2016,2017,2018 y 2019

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: López Valero, Elena

Tutor/a: Mateo Pla, Miguel Ángel

Cotutor/a: Lemus Zúñiga, Lenin Guillermo

CURSO ACADÉMICO: 2021/2022

Resumen

Investigadores del grupo de investigación Tecnologías de la Información y Comunicación contra en cambio climático (ICTvsCC) han desarrollado una metodología para aproximar los datos de emisiones de gases de efecto invernadero a partir de los datos provenientes del sistema de control de tráfico. Un software prototipo que implementa esta tecnología se ha usado en la ciudad de Valencia [1]. A pesar de la realización de varios estudios preliminares, se desea realizar un estudio más en profundidad de estos datos, seleccionando aquellos que más interés tengan para los ciudadanos y los decisores públicos. Para ellos se dispone de los datos del prototipo para los años de 2016 a 2021, completos y con una resolución horaria y para un conjunto elevado de calles de la ciudad de Valencia. En este TFG se deben definir y documentar los procesos de evaluación de dichos datos y el desarrollo de un informe de la situación de emisiones en la ciudad y su evolución en dichos años.

Tabla de contenidos

Índice de Figuras

Índice de Tablas

1 Introducción

La sociedad está cada vez más concienciada sobre el problema que significa el cambio climático, asumiendo que tiene gran impacto tanto en el desarrollo de las personas como en el de los animales y es una gran amenaza de la sociedad moderna[2]. El concepto de cambio climático es un conjunto de grandes cambios y transformaciones en nuestro planeta producidos como el resultado de las actividades domésticas incluyendo la minería o manufactura [3].

Conforme van pasado los años y los siglos, la atmósfera, los océanos, los suelos y la biodiversidad van cambiando como consecuencia, entre otras actividades humanas, de la intensificación de la pesca, la agricultura y la deforestación. Estas actividades producen gases de efecto invernadero, también producidos de manera natural, que absorben y emiten radiación infrarroja y como resultado se produce el efecto invernadero. La Convención Marco de las Naciones Unidas sobre el Cambio Climático clasifica en este grupo al dióxido de carbono (CO_2), metano (CH_4), óxido nitroso (N_2O), hidrofluorocarbonos (HFC), perfluorocarbonos (PFC) y el hexafluoruro de azufre (SF_6) [4].

Estos gases se encuentran de manera natural en la atmósfera, pero debido a la actividad humana, su cantidad ha aumentado produciendo un empeoramiento que genera la variabilidad del clima mundial, denominado cambio climático. Los efectos que puede producir el aumento de estos gases son: el deshielo, cambios en las estaciones, acidez en los océanos, inundaciones en la costa e islas, huracanes más peligrosos, migraciones, desertificación, daños en la agricultura y la ganadería, hambrunas y escasez de alimentos e incluso enfermedades y pandemias (ver Figura 1) [5].

No hace falta mirar hacia grandes procesos que producen un aumento de los gases de efecto invernadero, basta con mirar dentro de cada familia. Un ejemplo son productos de limpieza o los aerosoles, también el uso del automóvil que produce emisiones de dióxido de carbono ya que requiere de la combustión de fósiles.

Las consecuencias del cambio climático nos afectan a todos, independientemente de nuestro lugar de residencia. Por ello, ya existen numerosas acciones para evitar las peores consecuencias o, al menos, reducir su importancia. La ONU define el cambio climático como "… un cambio de clima atribuido directa o indirectamente a la actividad humana."[6]. En 1992, organizó la Convención Macro de las Naciones Unidas para estabilizar las concentraciones de GEI. Además, el Protocolo de Kioto de 1995 buscaba disminuir las emisiones de CO_2. En septiembre de 2015 se establecieron los 17 Objetivos de Desarrollo Sostenible (ODS) para 2030, con 169 metas relacionadas a cada objetivo. En 2016, en el Acuerdo de París lograron un acuerdo con el objetivo de no superar 2ºC la temperatura. Este acuerdo fue firmado por 175 líderes..."[6].

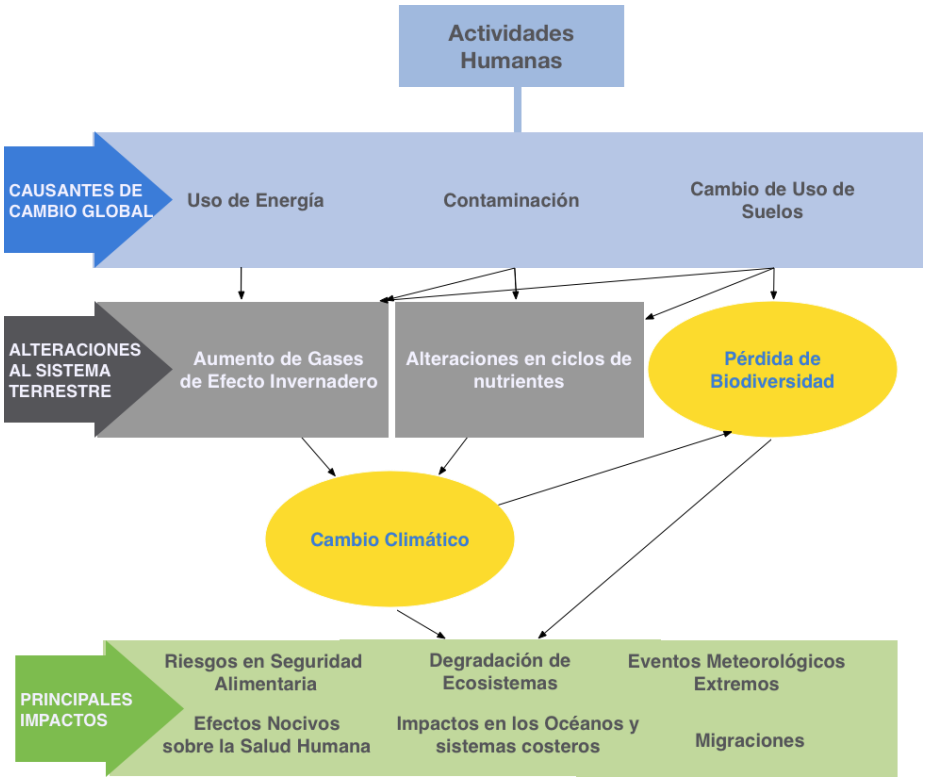


Figura 1. Causas y consecuencias cambio climático (Fuente: https://cambioglobal.uc.cl/comunicacion-y-recursos/que-es-el-cambio-global)

A menor escala, en nuestra ciudad, podríamos ayudar a reducir los gases de efecto invernadero adoptando un transporte público ecológico, construir mejor o reacondicionar los edificios de manera que no emitan ningún tipo de emisiones, más espacios verdes, fomentar los paseos y el uso de la bicicleta, adaptarse a la electricidad y uso inteligente del agua [7].

Por otro lado, emerge hoy día el concepto de ciudad inteligente "Smart City". La Oficina de Ciudad Inteligente de Valencia explica que "Una ciudad Inteligente y Sostenible es una ciudad innovadora que aprovecha las Tecnologías de la Información y la Comunicación (TIC) y otros medios para mejorar la calidad de vida, la competitividad, la eficiencia del funcionamiento y los servicios urbanos, al tiempo que se asegura de que responde a las necesidad de las generaciones presente y futuras en lo que respeta a los aspectos económicos, sociales, medioambientales y culturales" [8].

Se prevé que aproximadamente el 85% de la población mundial viva en ciudades en 2050. Por lo tanto, el modelo ideal de una Smart City sería un uso intensivo de las TIC y otras tecnologías para conseguir:

- Generación distribuida: un abastecimiento de electricidad individualizado y no centralizado.

- Smart Grids: redes inteligentes interconectadas entre los proveedores de servicios y los usuarios.
- Smart Metering: medición inteligente de los datos de gasto energético de cada usuario.
- Smart Buildings: edificios que respeten el medioambiente y que poseen sistemas de producción y gestión de energía integrados.
- Smart Sensors: sensores inteligentes que recopilen datos necesarios para mantener la ciudad conectada e informada, posibilitando que cada subsistema cumpla su función.
- eMobility: vehículos eléctricos y los respectivos puestos de recarga públicos y privados.
- Smart Citizen: los ciudadanos, ya que sin su participación no es posible poder llevar a cabo estas iniciativas.

Este trabajo estará centralizado en la ciudad de Valencia. Los datos que se utilizarán, proporcionados por la Oficina de Ciudad Inteligente (OCI) al grupo de investigación de digitalización, nos permitirán alcanzar nuestro objetivo principal: realizar un estudio riguroso de la movilidad de los automóviles en la ciudad de Valencia. Con esos resultados, se podrá mejorar el proceso de concienciar a los ciudadanos de Valencia para que reduzcamos lo máximo posible el uso del automóvil y con ello, los gases de efecto invernadero, por ejemplo, individualmente comenzando a usar más la bicicleta y globalmente convirtiendo nuestras ciudades en Smart Cities.

En este TFG se realizará un análisis riguroso del tráfico en la ciudad de Valencia por años (desde 2016 hasta 2021) y se realizarán visualizaciones para comprender mejor los resultados. A nivel estadístico, este estudio tendrá un margen de error porque es imposible tener todo el tráfico bien monitorizado en todo momento, pero conseguiremos una aproximación realista a la ciudad de Valencia.

1.1 Motivación

La motivación que me ha llevado a hacer este trabajo ha sido la alta necesidad de reducir los gases de efecto invernadero ya que cada vez las consecuencias se ven más presentes en nuestro día a día.

Vivimos en una sociedad muy avanzada tecnológicamente y en la que siempre se quiere investigar más, pero en muchas ocasiones nos olvidamos de la contaminación, que sigue siendo un tema muy importante. A finales de 2021 y principios de 2022, ya empezamos a vivir los primeros cambios inusuales para esa época del año, un invierno cálido con temperaturas muy impropias en la estación más fría del año [9]. Esto es producido por la cantidad de gases de efecto invernadero que liberamos en las actividades cotidianas de la vida. Simplemente con el uso de aerosoles o de

los automóviles estamos liberando a la atmósfera grandes cantidades. En este trabajo nos centraremos en los automóviles queriendo conseguir una reducción de esos gases concienciando a los valencianos a un mayor uso de la bicicleta. Con los diferentes análisis descubriremos que meses, semanas o incluso horas de la ciudad de Valencia se utiliza más el automóvil y si cada año que pasa se intensifica el uso que conlleve una mayor liberación de gases de efecto invernadero.

Por ello, haciendo este trabajo me gustaría concienciar a la gente y que se dieran cuenta que ya estamos viviendo los efectos de esos gases, como la subida de las temperaturas en los meses más fríos del año y el calor extremo en los meses más cálidos. Se podría concienciar a las personas para convertir las ciudades en Smart Cities, ciudades cada vez más sostenibles e involucradas en el ahorro de energía y de gases de efecto invernadero.

1.2 Objetivos

El objetivo principal de este TFG consiste en realizar un análisis descriptivo del tráfico en Valencia de los coches a lo largo de los años 2016 a 2021. Estos datos son proporcionados por la OCI a nuestro equipo de investigación. Los factores que influyen en este estudio son la intensidad, la ocupación, la velocidad y la fiabilidad de cada uno de ellos.

Una vez sabemos cuál es el objetivo principal, necesitamos alcanzar los siguientes objetivos específicos para lograr nuestro trabajo. Estos son:

- Limpiar los datos: eliminado de caracteres y otros elementos presentes en los archivos de la OCI que no son de utilidad en el estudio.
- Transformar los datos a un formato manejable por algunas de las herramientas más usadas en ciencia de datos: *RStudio* y *Python*.
- Realizar un preanálisis de los datos de los coches para detectar datos anómalos, visualizar correlaciones presentes, etc.
- Comparar las distribuciones de cada año, determinando si existen diferencias o tendencias.
- Realizar gráficos representativos que permitan visualizar mejor el estado y características del tráfico en la Valencia, así como su evolución en los últimos años.
- Aplicar los conocimientos aprendidos en el grado de Ciencia de Datos.

1.3 Impacto Esperado

Una vez se hayan completado los objetivos específicos y podamos cumplir el objetivo general de este trabajo, que consiste en realizar un análisis riguroso de la ciudad de Valencia durante 6 años, se pretende repercutir en la calidad de vida de los valencianos.

Con este análisis lo que se quiere conseguir es tener una vista general del uso del coche mediante el tráfico que captan las espiras repartidas por Valencia. De esta forma, se podrá llegar a conclusiones que podrían ayudar tanto a la implementación del proceso de captación de los datos como para conseguir una ciudad cada vez más sostenible incluyendo transformaciones en la ciudad para un uso eficiente, como podría ser la construcción de más carriles bici para facilitar los desplazamientos por la ciudad.

1.4 Metodología

En este trabajo se ha utilizado una metodología de desarrollo en cascada donde cada proceso va seguido de otro y las tareas se hacen en orden secuencial [10]. Las fases de este desarrollo las mostramos en la Figura 2 y las explicamos a continuación:

- **Requerimientos:** entenderemos los problemas, objetivos y requisitos del trabajo. Hace falta situar el trabajo en un contexto para entenderlo mejor y definir los objetivos tanto específicos como el principal. En este trabajo hay que estudiar y analizar el tráfico en coches entre los años 2016 y 2021 y el primer paso es la comprensión de la procedencia de nuestros datos y la importancia del estudio.
- **Diseño:** esta fase es la más importante ya que consiste en la integración, selección, depuración y transformación de los datos. Para ello, comprenderemos el formato de nuestros datos para poder cargarlos de forma adecuada, fusionaremos todas las bases de datos de los coches que se encuentran separadas por años en un mismo documento para que el tratamiento de los datos sea más cómodo y eficaz. Además, se limpiará la base de datos y se decidirá la calidad de los datos.
- **Implementación:** en esta fase realizaremos un análisis descriptivo de nuestras variables para conocer el tipo que son y explorar posibles valores anómalos en nuestros datos y su correcto tratamiento, posibles correlaciones entre variables y compararemos las distribuciones de distintas muestras de la base de datos.
- **Comprobación:** como fase final, donde realizaremos diferentes visualizaciones para validar el estudio realizado, demostrar que se han alcanzado todos los objetivos y, además, mostrar al cliente de forma más clara los resultados obtenidos.

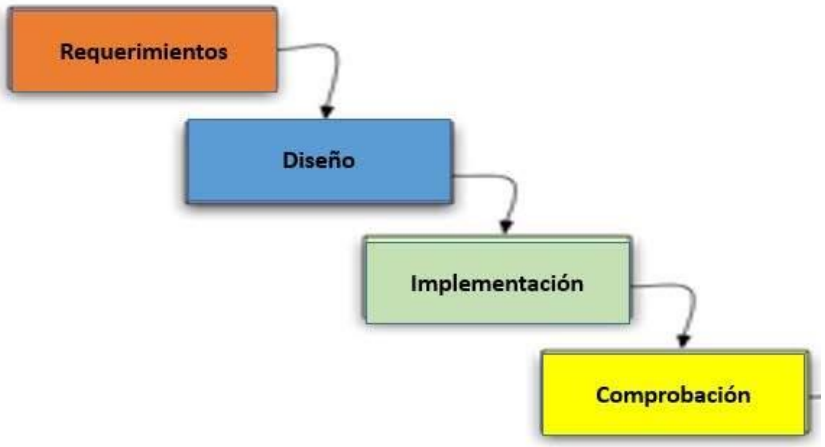


Figura 2: Método de cascada (Fuente: https://www.researchgate.net/figure/Figura-1-Actividades-del-desarrollo-de-software-en-el-modelo-de-cascada_fig1_320935254)

1.5 Estructura

La memoria de este trabajo se ha divido en los siguientes seis capítulos:

En el primer capítulo, podemos encontrar la introducción que incluye la motivación, los objetivos, el impacto esperado, la metodología seguida, la estructura y las colaboraciones. Como podemos observar, esta parte es fundamental ya que explicamos el contexto personal en el que se sitúa este trabajo y cuáles son los objetivos, tanto el general como los específicos, que queremos alcanzar en este Trabajo de Fin de Grado.

En el segundo capítulo, una vez situado el trabajo en el contexto personal hay que situarlo en el contexto profesional y cultural, es decir, en este capítulo se explica a qué nivel se encuentran las Smart Cities en Europa y sobre todo en España y que queremos ofrecer para mejorar, indicando las tecnologías que permitirán el desarrollo de este trabajo.

En el tercer capítulo, explicamos el problema principal que queremos tratar incluyendo un análisis del marco legal y ético, unas posibles soluciones, el plan de trabajo y el presupuesto que hemos necesitado tanto de horas invertidas como de hardware y software, la necesidad de hacer este trabajo, los datos que tenemos y las soluciones que hemos encontrado. Además, en este capítulo también se ha explicado los datos e información inicial que nos fueron facilitados.

En el cuarto capítulo, una vez que sabemos el significado de los datos, pasaremos a la primera herramienta que nos ayudará a tratar nuestros datos (*RStudio*) y que paquetes hemos utilizado. Además, explicaremos como hemos cargado los datos de forma que la codificación fuera correcta, como hemos juntado los datos, porque hemos eliminado unas variables y hemos creado otras.

En el quinto capítulo, se realizará un preanálisis de los datos donde se explicará con más detalle las variables que tenemos, que valores toman y de que formato son. También se realizará una limpieza de valores anómalos y diversas comparaciones de las distribuciones de los distintos años. Además, pasaremos a la herramienta *Power BI* y explicaremos como se han cargado los datos, los procesos realizados y las visualizaciones que hemos creado para observar más claramente los resultados obtenidos.

Por último, en el sexto capítulo, se expondrán las conclusiones tanto del trabajo como las personales e incluso se sugerirán trabajos futuros para la optimización del tiempo ya que ha habido procesos que podrían haber sido más rápidos.

1.6 Colaboraciones

En este trabajo, como hemos comentado anteriormente, se ha colaborado con investigadores del grupo de investigación Tecnologías de la Información y Comunicación contra en cambio climático (ICTvsCC) que se han encargado del desarrollo de una metodología para aproximar los datos de emisiones de gases de efecto invernadero a partir de los datos provenientes del sistema de control de tráfico. Además, también se ha colaborado con Carlos Romero, estudiante de la UPV. Ambos hemos partido de los mismos ficheros de datos proporcionados por la OCI del Ayuntamiento de Valencia: datos a nivel de espira y cada 5 minutos recogidos en tiempo real durante los años 2016, 2017, 2018, 2019, 2020 y 2021. El trabajo realizado por Carlos Romero se ha centrado en implementar una API que facilite la captura y visualización de los datos. En paralelo, yo me he encargado de realizar un análisis riguroso de los datos.

2 Estado del arte

La Ciencia de datos surge en 1962 gracias a John Wilder Tukey gran estadístico del siglo XX que desarrolló algoritmos complejos y escribió varios libros como "Exploratory Data Analysis" [11], pero no es hasta 2001 que empieza a considerarse una disciplina independiente a la estadística [12]. Esta ciencia permite el análisis de grandes volúmenes de datos para encontrar patrones y extraer conocimiento por medio de métodos estadísticos.

En este trabajo se ha necesitado los conocimientos adquiridos en el grado de Ciencia de Datos de la Universidad Politécnica de Valencia para realizar un análisis descriptivo de los datos de tráfico de la ciudad de Valencia. Los datos utilizados se han obtenido gracias a los sensores del Ayuntamiento de Valencia que hay repartidos por toda la ciudad.

2.1 Crítica al estado del arte

Cada vez más, las ciudades apuestan por desarrollarse como un ejemplo del concepto llamado Smart City, es decir, por convertir la ciudad en un sistema más eficiente, que facilite la movilidad de los ciudadanos, mejore su sostenibilidad, etc. De esta forma, se persigue mejorar la calidad de vida de los ciudadanos [13].

En la Unión Europea encontramos que las urbes ocupan el 4% de la superficie con 75% de ciudadanos que emiten el 69% de las emisiones globales de CO_2 [14]. Las principales capitales han optado por las ciudades inteligentes consiguiendo que 5 de ellas destaquen por su innovación. 10 de las ciudades innovadoras con mayor facilidad para desarrollar proyectos tecnológicos son Copenhague, Estocolmo, Múnich, Barcelona, Berlín, Londres, Rotterdam, Costa Azul, Dublín y Tallin [15]. Se prevé que en 2030 en cada país haya al menos 10 Smart Cities, consiguiendo así un total de unas 100 ciudades inteligentes.

Actualmente, las ciudades españolas que tienen como objetivo convertirse en Smart Cities son: Málaga, Zaragoza, Gijón, Donostia, Vitoria, Bilbao, Madrid, Santander y Valencia. Cada ciudad pretende actuar de forma diferente:

- Málaga, se centra en la gestión de la energía, balances eficientes de la demanda y la involucración de todos los agentes del sistema eléctrico.
- Zaragoza, trabaja en una implantación de estructuras de telecomunicaciones o el uso de software libre.
- Gijón, se centra en la administración electrónica, open data y movilidad eléctrica.

- Vitoria, en la reducción de la contaminación acústica, la concienciación ciudadana y el transporte sostenible.
- Bilbao, se encuentra en el top 5 de ciudades inteligentes de España e incrementa la eficiencia energética y de transporte y aumenta su atractivo turístico.
- Madrid, cuenta con sensores que cuentan el número de coches que pasan por las calles
- Santander, utiliza gran cantidad de sensores que recopilan la información del estado de la ciudad.
- Valencia, cuenta con la "Oficina de Ciudad Inteligente" (OCI) [8] y, además, es la ciudad más sensorizada de toda Europa.

2.2 Propuesta

Existen numerosos estudios relacionados con el tráfico en la ciudad de Valencia como por ejemplo "From traffic data to GHG emissions" [1] realizado por los profesores de la Universidad Politécnica de Valencia Miguel A. Mateo Pla y Lenin G. Lemus, entre otros.

Este trabajo consiste en realizar un análisis descriptivo del tráfico de la ciudad de Valencia que no ha sido realizado en otros proyectos, centralizándonos en realizar primeramente una limpieza de los datos para obtener resultados correctos, un preanálisis donde conoceremos más las variables de estudio, se compararán las distribuciones de los diferentes años que tenemos y, finalmente, se visualizarán varios gráficos. Con este análisis se podrá obtener posibles patrones que existen en la ciudad de Valencia y se podrá exponer diferentes mejoras para convertirla en una ciudad cada vez más sostenible y en la Smart City ideal.

3 Análisis del problema

Inicialmente, la Oficina de Ciudad Inteligente (OCI) nos mandó datos procedentes de diferentes tramos repartidos por la ciudad Valencia a nivel de espira y recogidos cada 5 minutos. Los datos con los que disponemos se han creado desde 2016 a 2021 cada mes. Esos datos incluían información sobre el tráfico de bicicletas y de coches de los 6 años y cada año separado en 12 bases de datos, es decir, por meses y con 12 variables donde la información se encuentra explicada en la Tabla 1.

Columnas en .csv	Contenido de las columnas	Nombre del valor almacenado en MongoDB
Columna A	Identificador del tramo (IDTA) – **Puntos aforados**	id1
Columna B	Identificador del punto de medida (IDPM)	id2
Columna C	Nombre del punto de medida	id3
Columna D	Descripción del punto de medida	dirección
Columna E	Fecha y Hora.	fecha1
Columna F	Intensidad de Tráfico por hora	id4
Columna G	Fiabilidad de intensidad de Tráfico	id5
Columna H	Ocupación – Si hay automóviles	id6
Columna I	Fiabilidad de Ocupación	id7
Columna J	Velocidad – Promedio del flujo	id8
Columna K	Fiabilidad de Velocidad	id9
Columna L	Fecha (sin hora)	fecha2

Tabla 1. Variables iniciales

Más tarde, tras visualizar diversos errores en los datos, recibimos nuevas bases de datos, también separadas por años y por meses, pero esta vez no teníamos la información del uso de la bicicleta en la ciudad de Valencia de 4356 tramos, sino que solo teníamos la información del tráfico de coches. La información sobre las nuevas bases de datos se encuentra explicada en la Tabla 2 y, como podemos observar, se eliminó la columna L de la Tabla 1 y se añadió la columna Sentido.

Columnas en .csv	Contenido de las columnas
IdPM	Identificador del punto de medida
Nombre	Nombre del punto de medida
Descripción	Descripción del punto de medida
Sentido	Dirección del tráfico
IdTA	Identificador del tramo
FechaHora	Fecha y hora
Intensidad	Intensidad de Tráfico por hora
FiabilidadIntensidad	Fiabilidad de intensidad de Tráfico
Ocupación	% de tiempo que durante el ciclo semafórico había algún automóvil encima del sensor
FiabilidadOcupacion	Fiabilidad de Ocupación
Velocidad	Velocidad – Promedio del flujo
FiabilidadVelocidad	Fiabilidad de Velocidad

Tabla 2. Variables finales

3.1 Análisis del marco legal y ético

El análisis del marco legal y ético debe ser cumplido por cualquier profesional y se deben de tener en cuenta varios factores:

- La protección de datos. Es importante saber de dónde proceden los datos con los que vamos a trabajar, cómo son almacenados y si se requiere el cumplimiento de algunas normativas para su uso. En este trabajo los datos proporcionados por la empresa Oficina de Ciudad Inteligente de Valencia son datos abiertos que podemos encontrar en la página Open Data del ayuntamiento de Valencia y, además, son datos anónimos con lo cual no se necesitaría de un ningún permiso ni de ningún acuerdo de confidencialidad.
- La propiedad intelectual que se relaciona con la creación del intelecto humano mediante patentes, derechos de autor y marcas. En este trabajo, el software utilizado es libre ya que está preparado para comercialización.
- La ética. En este trabajo contamos con datos que nos ofrece el ayuntamiento de Valencia por lo que no afectan al comportamiento humano ante dilemas morales.
- Otros aspectos legales.

3.2 Identificación y análisis de soluciones posibles

El problema que se quiere llevar a resolver es si en la ciudad de Valencia existen patrones del tráfico según el paso de los años, durante la semana o por horas. Para ello, se realizará el análisis riguroso de los datos de tráfico de diferentes años con diferentes visualizaciones y se podrá llegar a observar si cada vez que pasan los años el uso del coche ha incrementado y si es así, proponer posibles soluciones para la disminución de gases de efecto invernadero como podría ser promover la utilización de la bicicleta y una menor utilización del automóvil.

3.3 Solución propuesta

Para un científico de datos es vital una correcta toma de decisiones que tienen que estar basadas en información fiable. Para ello, es necesario seguir una serie de pasos (ver Figura 3) para lograr el objetivo final.

El primer paso es obtener los datos con los que vamos a trabajar. En este trabajo los datos son obtenidos por la Oficina de Ciudad Inteligente y una vez ordenados y organizados, en la medida de lo posible, nos los envían.

El segundo paso y el más importante, ya que le daremos calidad a nuestros datos, es la preparación. En este paso, limpiaremos los datos que no son fiables, nulos o anómalos, se eliminarán o añadirán variables que no son necesarias para el enfoque de nuestro estudio o realizaremos diversas transformaciones.

El siguiente paso para seguir avanzando para conseguir nuestro objetivo final, consiste en la exploración o análisis de la base de datos. En este paso, se comenzará a conocer de una forma más precisa las variables que tenemos, el tipo que son, si tienen correlación unas variables con otras y haremos las comparaciones oportunas.

A continuación, pasaremos a la visualización donde mostraremos gráficamente los resultados de nuestro estudio de una forma más clara para que el cliente entienda de una primera vista cual es el objetivo y si lo hemos alcanzado.

Finalmente, siguiendo todos estos pasos obtendremos el objetivo final que es realizar un análisis descriptivo del tráfico en Valencia de los coches a lo largo de los años 2016 a 2021 y se podrá compartir con los clientes.

Figura 3. Pasos (Fuente: https://www.r-bloggers.com/2016/08/r-with-power-bi-import-transform-visualize-and-share/)

3.4 Plan de trabajo

Tras identificar el problema y proponer posibles soluciones, será necesario realizar un plan de trabajo para cumplir con todos los objetivos específicos para poder llegar al objetivo general.

En primer lugar, se realiza una documentación de estudios realizados y de diversas aplicaciones para poder ir haciéndonos una idea de cómo enfocar este trabajo. Para ello, se leerán las páginas web: "From traffic data to GHG emissions: A novel bottom-up methodology and its application to Valencia city" [1] y "Valencia traffic" [16]. Además, nos documentaremos del gestor de referencias de Mendeley [17] y ampliaremos los conceptos de *Power BI* como el libro "Fundamentos de Modelado en Estrella" [18]

En segundo lugar, se realiza un análisis previo a los datos obtenidos. Se obtienen nuevas variables y transformaciones, se realizan agrupaciones y la limpieza de los datos para la eliminación de variables no necesarias y/o de casos no fiables. Además, se obtienen los primeros análisis y visualizaciones.

Tras numerosos errores en los primeros datos que teníamos, se obtienen unas nuevas bases de datos a las que se les hace un tratamiento muy parecido a los datos anteriores. Esta vez se cargan los datos teniendo en cuenta la codificación, se limpian muchas de las filas ya que no siguen el mismo formato que las demás, se eliminan variables y se realizan agrupaciones que ayudan a tener bases de datos más pequeñas para reducir el tiempo de trabajo y que ayudará a la posterior unión de todos los meses y de todos los años. Además, se obtiene un preanálisis para observar la información más detallada de nuestras variables finales y posibles correlaciones para ir entendiendo cada vez más nuestra base de datos final.

Una vez ya hemos obtenido nuestra base de datos final y hemos conocido más nuestras variables de estudio, se procederá a realizar las comparaciones oportunas, pero antes nos tendremos que documentar de algunas funciones de *RStudio* ya que nunca había utilizado muchas de ellas. Una vez obtenidas las comparaciones, se procederá a la visualización de datos e interpretación de los resultados que nos ayudarán a observar si las comparaciones están bien hechas y a enseñar al usuario de una forma más clara los resultados obtenidos.

Mientras se han realizado las siguientes fases del trabajo, al mismo tiempo, se ha estado realizando el desarrollo de la memoria para reflejar el trabajo realizado durante todo el proceso.

Todo el tiempo junto a las fechas de inicio y final de cada fase del trabajo, se encuentran mostradas en el diagrama de Gantt de la Figura 4 y, como se puede observar, el trabajo ha sido realizado en 23 semanas.

TAREA	INICIO	FINAL	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18	W19	W20	W21	W22	W23
Contexto proyecto	**27.12.2021**	**09.01.2022**																							
Documentación datos de tráfico a las emisiones GEI	27.12.2021	27.12.2021																							
Documentación TomTom	29.12.2021	29.12.2021																							
Documentación Mendeley	31.12.2021	02.01.2022																							
Documentación PowerBI	03.01.2022	09.01.2022																							
Análisis previo	**10.01.2022**	**06.02.2022**																							
Obtención de nuevas variables y transformaciones	10.01.2022	16.01.2022																							
Agrupaciones y limpieza	16.01.2022	27.01.2022																							
Primeros análisis	28.01.2022	03.02.2022																							
Primeras visualizaciones	3.02.2022	06.02.2022																							
Tratamiento datos	**28.04.2022**	**05.06.2022**																							
Cargar datos	28.04.2022	01.05.2022																							
Limpieza	02.05.2022	16.05.2022																							
Preanálisis	18.05.2022	05.06.2022																							
Análisis descriptivo datos	**06.06.2022**	**17.07.2022**																							
Documentación para realizar comparaciones	06.06.2022	10.06.2022																							
Comparación distribuciones	11.06.2022	27.06.2022																							
Visualizaciones	01.07.2022	10.07.2022																							
Interpretación resultados	11.07.2022	17.07.2022																							
Memoria	**13.01.2022**	**16.08.2022**																							

Figura 4. Diagrama de Gantt

3.5 Presupuesto

Para la realización de este proyecto, como ya he comentado anteriormente, se han utilizado unas 23 semanas con una media de 15 horas semanales, con lo que se obtendría una cantidad de 350 horas dedicadas a este proyecto. Si cada ECTS equivale a 25 horas y este proyecto son 12 ECTS, las horas empleadas son mayores que los créditos asignados.

Un científico o científica de datos recién graduado o graduada tiene un salario alrededor de los 29.400€ al año, con lo que a la hora serían unos 14€. En la Tabla 3 se puede observar las horas dedicadas a cada fase del trabajo y el dinero que habría que invertir en un científico o científica de datos para alcanzar los objetivos.

En el caso del hardware, el ordenador portátil que se ha utilizado en este trabajo tiene un coste de aproximadamente 900€ y una vida útil media de 4 años que equivale a: 4 años x 50 semanas/año = 200 semanas, así pues, la amortización del equipo ha sido:

$$Amortización = \frac{Coste\ del\ equipo\ \times Semanas\ de\ uso}{Vida\ útil\ del\ equipo\ en\ semanas} = \frac{900€ \times 23\ semanas}{200\ semanas} \approx 104€$$

En este caso, el software utilizado no conlleva ningún coste ya que *RStudio*, que es un lenguaje de programación que se centra en la estadística, es un software libre, *NotePad++* que es un editor de texto con código fuente libre, *Power BI* que nos permite analizar nuestros datos con visualizaciones interactivas y aunque existen varias versiones de pago como puede ser *Power BI Pro* o *Power BI Premiun* se ha utilizado la versión gratuita que es *Power BI Desktop* y para realizar la bibliografía se ha utilizado el gestor de referencias *Mendeley* que también ha sido gratuito.

Tarea	Tiempo	Coste unitario	Coste
Contexto del proyecto	30h	14€	420€
Documentación datos de tráfico a las emisiones GEI	3h	14€	42€
Documentación TomTom	2h	14€	28€
Documentación *Mendeley*	5h	14€	70€
Documentación *PowerBI*	20h	14€	280€
Análisis previo	80h	14€	1120€
Obtención de nuevas variables y transformaciones	15h	14€	210€
Agrupaciones y limpieza	25h	14€	350€
Primeros análisis	15h	14€	210€
Primeras visualizaciones	25h	14€	350€
Tratamiento datos	115h	14€	1400€
Cargar datos	5h	14€	70€
Limpieza	65h	14€	700€
Preanálisis	45h	14€	630€
Análisis datos	85h	14€	1400€
Documentación para realizar comparaciones	5h	14€	140€
Comparación distribuciones	40h	14€	560€
Visualizaciones	25h	14€	350€
Interpretación de resultados	15h	14€	350€
Memoria	40h	14€	560€
Total	350h		4900€

Tabla 3. Tareas del trabajo y tiempo y coste invertidos

4 Preparación y Comprensión de Datos

Una vez se han obtenido los datos con los que se va a trabajar, es importante prepararlos y comprenderlos realizándoles una limpieza para realizar las transformaciones necesarias y evitar trabajar con datos no fiables o innecesarios. En primer lugar, se ha decidido trabajar con *RStudio* ya que a lo largo de la carrera se ha estado trabajando con este software estadístico.

Para realizar la limpieza y transformaciones necesarias a nuestros datos se han utilizado las librerías: `readr, dplyr, tidyr y stringr.`

Inicialmente, se realizó un primer análisis con los datos de tráfico del año 2021 de los coches y surge el primer inconveniente. El trabajo con todos los meses del año juntados en un mismo documento *CSV* parecía ser una buena opción para una limpieza más rápida y eficaz. pero en el caso de los coches daba el error: "*No se puede ubicar un vector de tamaño 515.5 Mb*". Esto sucedía porque al ser tan grandes las bases de datos no dejaba juntarlas en una. Así que, se decidió tratar las bases de datos de los meses de los automóviles por separado, mientras que en el caso de las bicicletas se consiguió juntar los meses de cada año en un mismo documento y guardarlo en un documento de *Excel*.

Tanto para las bases de datos de los coches como las bases de datos de las bicicletas, se realizaron las mismas transformaciones, es decir:

- Añadir nombres a las columnas.
- Añadir la semana del año, el día de la semana y mes.
- Separar la columna fecha y, así, obtener la hora.

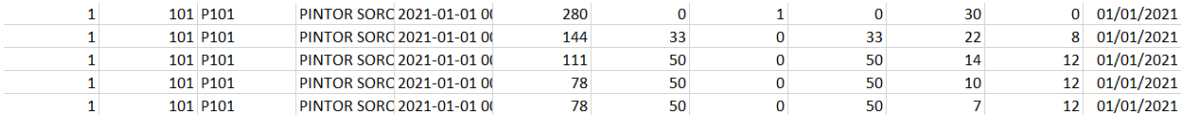
Una vez ya acabadas las trasformaciones, se guardaron todas las bases de datos para no tener que aplicar las trasformaciones cada vez y, de esta forma, tardar más de lo necesario.

Ahora se pasó a analizar si existía algún dato faltante o algún error en los datos. Para ambos grupos de bases de datos, se encontraron los mismos errores en los mismos datos, en 2018 había diversas líneas con datos faltantes o con fallos de codificación, como se puede observar en la Figura 5 y en la Figura 6, y al ser pocas se decidió eliminarlas.

Figura 5. Fallos en las bicis de 2018

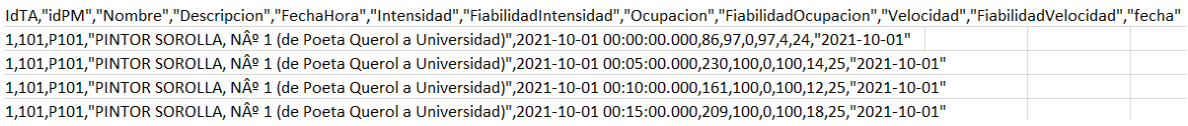
Figura 6. Fallos en los coches 2018

Con 2021, pasaba algo completamente diferente, los datos de los meses de octubre, noviembre y diciembre de los datos de tráfico de coches, no se encontraban en el mismo formato que los demás meses. Así que se cambió a *Excel* para ver en que cambiaba respecto los demás. Como se puede observar en la Figura 7, las variables se encontraban separadas por columnas de las bases de datos desde enero hasta septiembre. En cambio, como se puede observar en la Figura 8, las variables se encontraban separadas por comas para los meses de octubre, noviembre y diciembre. Así que se decidió separar la información de cada variable en columnas.



1	101	P101	PINTOR SORC	2021-01-01 0(	280	0	1	0	30	0	01/01/2021
1	101	P101	PINTOR SORC	2021-01-01 0(	144	33	0	33	22	8	01/01/2021
1	101	P101	PINTOR SORC	2021-01-01 0(	111	50	0	50	14	12	01/01/2021
1	101	P101	PINTOR SORC	2021-01-01 0(	78	50	0	50	10	12	01/01/2021
1	101	P101	PINTOR SORC	2021-01-01 0(	78	50	0	50	7	12	01/01/2021

Figura 7. Base de datos de enero de 2021



```
IdTA,"idPM","Nombre","Descripcion","FechaHora","Intensidad","FiabilidadIntensidad","Ocupacion","FiabilidadOcupacion","Velocidad","FiabilidadVelocidad","fecha"
1,101,P101,"PINTOR SOROLLA, NÂº 1 (de Poeta Querol a Universidad)",2021-10-01 00:00:00.000,86,97,0,97,4,24,"2021-10-01"
1,101,P101,"PINTOR SOROLLA, NÂº 1 (de Poeta Querol a Universidad)",2021-10-01 00:05:00.000,230,100,0,100,14,25,"2021-10-01"
1,101,P101,"PINTOR SOROLLA, NÂº 1 (de Poeta Querol a Universidad)",2021-10-01 00:10:00.000,161,100,0,100,12,25,"2021-10-01"
1,101,P101,"PINTOR SOROLLA, NÂº 1 (de Poeta Querol a Universidad)",2021-10-01 00:15:00.000,209,100,0,100,18,25,"2021-10-01"
```

Figura 8. Base de datos de octubre de 2021

Para poder tratar todas las bases de datos de la misma forma, se procedió a separar las variables de los meses de octubre, noviembre y diciembre en columnas. Para ello, se seleccionará la primera columna de la base de datos, ya que la información se encuentra en las celdas de la primera columna, y en la sección "Datos" se elegirá la opción "Texto en columnas" del apartado "Herramientas de datos" mostrado en la Figura 9. Se nos abrirá una ventana con las opciones de cómo queremos separar las columnas, si con alguna delimitación o con ancho fijo. Como se puede

observar en la Figura 8, las variables se encuentran delimitadas por comas. De esta forma, se elegirá la opción de "Delimitados" que nos llevará a la ventana de opciones de la Figura 10 donde elegiremos el separador de coma. En la vista previa de los datos, que se encuentra en la Figura 10, se puede observar que los datos se encuentran separados correctamente ya que cada variable se encuentra en una columna distinta.

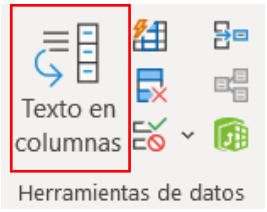


Figura 9. Texto en columnas de Excel

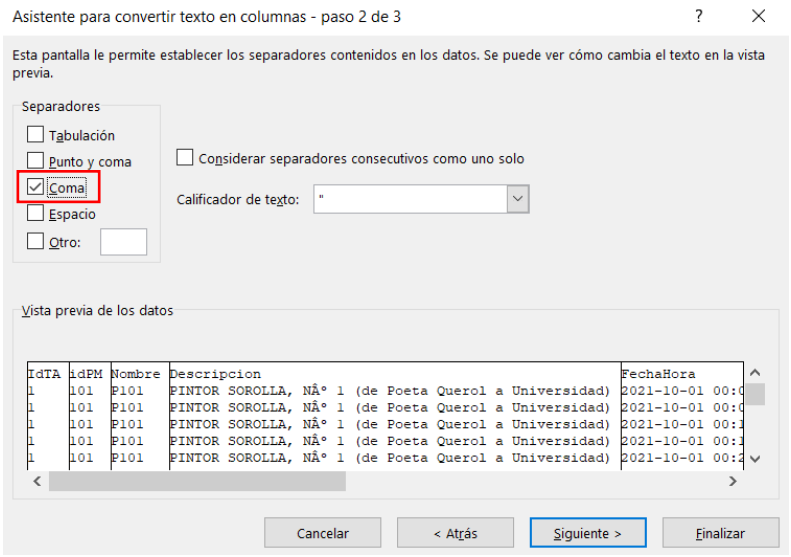


Figura 10. Delimitados de Texto en columnas

Para un mejor tratamiento de los datos, se decidió agrupar por ID, Nombre, Descripción, Fecha y Hora; ya que los datos se encontraban recopilados cada 5 minutos y no era necesario tanto nivel detalle para alcanzar los objetivos propuestos. Además, a esa agrupación se le añadió las sumas por intensidad y ocupación y las medias por fiabilidad de la intensidad, fiabilidad de la ocupación, velocidad y fiabilidad de la velocidad. Una vez se obtuvieron las agrupaciones, las bases de datos de las bicicletas se guardaron en un *Excel* nuevo mientras que, para las bases de datos de los coches, al haber disminuido el tamaño de las bases de datos con las agrupaciones, se pudieron juntar los meses de cada año en documento *Excel*.

Una vez realizada la agrupación de los datos y obtenido las nuevas bases de datos, se realizó el proceso de identificación de valores anómalos en la base de datos de las bicicletas obteniendo

velocidades superiores a 40km/h e intensidades con valores muy elevados como por ejemplo 23.604 coches/hora.

Para ver qué pasaba con la velocidad y con la intensidad, se decidió estudiar las fiabilidades tomando como válidos aquellos valores mayores del 50%, ya que hacer el estudio con datos poco fiables puede llevar a tomar decisiones menos precisas o incorrectas.

Para la base de datos de las bicicletas en cada año y en cada tipo de fiabilidad menor que el 50% encontré la cantidad mostrada en Tabla 4. La Tabla 5 muestra la cantidad de datos que se encuentran con una fiabilidad menor que el 50% en la base de datos de los coches. Como se puede observar en las tablas, los porcentajes de la cantidad de fiabilidad menor que el 50% en la velocidad son del 100%, es decir, que ninguno de los datos de la fiabilidad de la velocidad es fiable. En este estudio ya se puede ir entendiendo el porqué de la aparición de velocidades superiores a 40km/h en bicicletas o de las intensidades con valores muy elevados.

Año	Fiabilidad_intensidad	Fiabilidad_ocupacion	Fiabilidad_velocidad
2016	27 182 (4.22%)	27 182 (4.22%)	643 756 (100%)
2017	29 680 (3.83%)	29 680 (3.83%)	774 465 (100%)
2018	16 416 (2.08%)	16 416 (2.08%)	790 273 (100%)
2019	17 143 (1.47%)	17 143 (1.47%)	960 034 (100%)
2020	26 454 (2.64%)	26 454 (2.64%)	1 002 192 (100%)
2021	23 375 (3.54%)	23 375 (3.54%)	660 144 (100%)

Tabla 4. Fiabilidad menor del 50% en las bicicletas

Año	Fiabilidad_intensidad	Fiabilidad_ocupacion	Fiabilidad_velocidad
2016	552 279 (5.29%)	552 279 (5.29%)	10 436 884 (100%)
2017	443 655 (4.3%)	443 655 (4.30%)	10 329 300 (100%)
2018	237 719 (2.44%)	237 719 (2.44%)	9 720 950 (100%)
2019	233 267 (2.36%)	233 267 (2.36%)	9 902 356 (100%)
2020	435 985 (4.16%)	435 985 (4.16%)	10 458 673 (100%)
2021	288 074 (3.55%)	288 074 (3.55%)	8 105 011 (100%)

Tabla 5. Fiabilidad menor del 50% en los coches

Tras la detección de todos estos errores, se decidió contactar con los responsables de los datos. Gracias a Carlos Hernández, trabajador de la Oficina de Ciudad Inteligente de la ciudad de Valencia y responsable de los datos, se obtuvieron nuevos datos de los que se aseguró que eran los adecuados y que no tendrían tantos fallos.

De obtener dos tipos de bases de datos, una para las bicicletas y otra para el tráfico de los coches, se pasó a analizar solo las bases de datos de los coches ya que solo nos dieron acceso a esas. Se volvió a realizar el proceso de limpieza de los datos y, cargándolos, aparecía un problema de codificación ya que acentos o símbolos no estaban en el formato adecuado. Para saber el tipo de codificación que tenían nuestros datos, se utilizó la aplicación *NotePad++* y se descubrió que la codificación utilizada era 'Windows-1252', así que se cargaron los datos teniéndola en cuenta.

La eliminación de columnas innecesarias para nuestro estudio es uno de los pasos importantes para la limpieza tanto para el ahorro de espacio como para la disminución del tiempo de ejecución de los procesos. Se decidió eliminar las columnas de *Sentido*, ya que en nuestro estudio no iba a ser utilizada porque lo que queríamos obtener era un análisis de la ciudad de Valencia en general y no centrarnos en un estudio de flujo; la *Velocidad* y *FiabilidadVelocidad*, ya que no eran datos reales porque se necesitarían espiras dobles y ninguna espira de balanceo es doble. Además, se añadieron las columnas de *Fecha* y *hora* obtenidas gracias a la partición de la columna de *FechaHora*.

El siguiente paso que se realizó en nuestra limpieza era la detección de valores nulos (Figura 11). Un valor nulo es aquel valor desconocido de nuestra base de datos y está representado como NULL o NA. Si se analiza con *RStudio* con la función `which(is.na(datos))` se observa que sí que existen valores nulos y te identifica la celda a la que corresponde. Una vez identificada las celdas, se descubrió que no había valores nulos, sino que existía un salto de línea ya que había valores que no correspondían a la columna correcta. Como se puede observar en las celdas de la Figura 11, hay veces que se encuentra un valor numérico donde corresponde un valor de tipo texto. Por eso, se decidió volver a utilizar la aplicación *NotePad++* para ver que sucedía con esos casos. En la Figura 12, donde aparece una muestra de nuestra base de datos, se encontraron algunas líneas que incluían un salto de línea o un retorno de carro que provocaba que una observación se partiera en dos líneas. Para la resolución del problema y volver a juntar las observaciones adecuadas, la mejor forma fue reemplazar los valores `s/n\n` y `s/n\r\n` por `s/n` ya que correspondían a un retorno de carro y a una tabulación.

	idPM	Nombre	Descripcion	idTA	FechaHora	Intensidad	FiabilidadIntensidad	Ocupacion	FiabilidadOcupacion	Fecha	hora
1	2210	P2210	INSTITUTO OBRERO DE VALÈNCIA s/n	NA	NA	NA	NA	NA	NA	NA	NA
2	esquina Av. Amado Granell Mesado (de Av. Amado Granell ...	Profesor López Piñero	598	284	NA	1	100	19	25	NA	NA
3	2210	P2210	INSTITUTO OBRERO DE VALÈNCIA s/n	NA	NA	NA	NA	NA	NA	NA	NA
4	esquina Av. Amado Granell Mesado (de Av. Amado Granell ...	Profesor López Piñero	598	284	NA	1	100	19	25	NA	NA
5	2210	P2210	INSTITUTO OBRERO DE VALÈNCIA s/n	NA	NA	NA	NA	NA	NA	NA	NA
6	esquina Av. Amado Granell Mesado (de Av. Amado Granell ...	Profesor López Piñero	598	284	NA	1	100	19	25	NA	NA
7	2210	P2210	INSTITUTO OBRERO DE VALÈNCIA s/n	NA	NA	NA	NA	NA	NA	NA	NA
8	esquina Av. Amado Granell Mesado (de Av. Amado Granell ...	Profesor López Piñero	598	284	NA	1	100	19	25	NA	NA
9	2210	P2210	INSTITUTO OBRERO DE VALÈNCIA s/n	NA	NA	NA	NA	NA	NA	NA	NA
10	esquina Av. Amado Granell Mesado (de Av. Amado Granell ...	Profesor López Piñero	598	284	NA	1	100	19	25	NA	NA
11	2210	P2210	INSTITUTO OBRERO DE VALÈNCIA s/n	NA	NA	NA	NA	NA	NA	NA	NA
12	esquina Av. Amado Granell Mesado (de Av. Amado Granell ...	Profesor López Piñero	598	284	NA	1	100	19	25	NA	NA
13	2210	P2210	INSTITUTO OBRERO DE VALÈNCIA s/n	NA	NA	NA	NA	NA	NA	NA	NA
14	esquina Av. Amado Granell Mesado (de Av. Amado Granell ...	Profesor López Piñero	598	284	NA	1	100	19	25	NA	NA
15	2210	P2210	INSTITUTO OBRERO DE VALÈNCIA s/n	NA	NA	NA	NA	NA	NA	NA	NA
16	esquina Av. Amado Granell Mesado (de Av. Amado Granell ...	Profesor López Piñero	598	284	NA	1	100	19	25	NA	NA

Figura 11. Valores nulos

```
2210;P2210;INSTITUTO OBRERO DE VALÈNCIA s/n
esquina Av. Amado Granell Mesado (de Av. Amado Granell Mesado a Es;Profesor López Piñero;598;2016-01-31 23:45:00.000;151;100;0;100;11;25
2210;P2210;INSTITUTO OBRERO DE VALÈNCIA s/n
esquina Av. Amado Granell Mesado (de Av. Amado Granell Mesado a Es;Profesor López Piñero;598;2016-01-31 23:50:00.000;151;100;0;100;11;25
2210;P2210;INSTITUTO OBRERO DE VALÈNCIA s/n
esquina Av. Amado Granell Mesado (de Av. Amado Granell Mesado a Es;Profesor López Piñero;598;2016-01-31 23:55:00.000;151;100;0;100;11;25
2211;P2211;AV. AMADO GRANELL MESADO (MILITAR) s/n
esquina Alcalde Gisbert Rico (de Av. Hermanos Maristas a Av. ;Peris y Valero;599;2016-01-01 00:00:00.000;198;100;1;100;32;23
2211;P2211;AV. AMADO GRANELL MESADO (MILITAR) s/n
esquina Alcalde Gisbert Rico (de Av. Hermanos Maristas a Av. ;Peris y Valero;599;2016-01-01 00:05:00.000;198;100;1;100;32;23
2211;P2211;AV. AMADO GRANELL MESADO (MILITAR) s/n
esquina Alcalde Gisbert Rico (de Av. Hermanos Maristas a Av. ;Peris y Valero;599;2016-01-01 00:10:00.000;198;100;1;100;32;23
2211;P2211;AV. AMADO GRANELL MESADO (MILITAR) s/n
esquina Alcalde Gisbert Rico (de Av. Hermanos Maristas a Av. ;Peris y Valero;599;2016-01-01 00:15:00.000;198;100;1;100;32;23
```

Figura 12. Valores nulos en Notepad++

Una vez se obtuvieron todas las líneas correctamente, se volvió a utilizar a *RStudio* y se pasó a visualizar si había algún valor más nulo. En ninguna base de datos de ningún año se encontraron, excepto en 2021 en las columnas *Ocupación* y *FiabilidadOcupación*. Así que se decidió analizar qué había pasado con esas filas y columnas. Una vez que se identificó las filas en las que se encontraban los datos nulos, se decidió pasar a *Notepad++* y visualizar que pasaba con los archivos originales. Como se puede observar en la Figura 13, no se encuentra información en las columnas de *Ocupación* y *FiabilidadOcupación* y esto sucede en los meses de mayo hasta diciembre.

```
12211511 921004;P921004;DR. PESET ALEIXANDRE (GENERAL AVILÉS);INVERSO;1587;2021-05-31 23:50:00;232;0;0;0;0;0
12211512 921004;P921004;DR. PESET ALEIXANDRE (GENERAL AVILÉS);INVERSO;1587;2021-05-31 23:55:00;232;0;0;0;0;0
12211513 9200608;P9200608;CARRIL BICI CAMARA 608 GUILLEM DE CASTRO - JESÚS (GUILLEM DE CASTRO);DIRECTO;1601;2021-05-03 00:00:00;0;100;;;0;
12211514 9200608;P9200608;CARRIL BICI CAMARA 608 GUILLEM DE CASTRO - JESÚS (GUILLEM DE CASTRO);DIRECTO;1601;2021-05-03 00:05:00;0;100;;;0;
12211515 9200608;P9200608;CARRIL BICI CAMARA 608 GUILLEM DE CASTRO - JESÚS (GUILLEM DE CASTRO);DIRECTO;1601;2021-05-03 00:10:00;0;100;;;0;
12211516 9200608;P9200608;CARRIL BICI CAMARA 608 GUILLEM DE CASTRO - JESÚS (GUILLEM DE CASTRO);DIRECTO;1601;2021-05-03 00:15:00;0;100;;;0;
12211517 9200608;P9200608;CARRIL BICI CAMARA 608 GUILLEM DE CASTRO - JESÚS (GUILLEM DE CASTRO);DIRECTO;1601;2021-05-03 00:20:00;0;100;;;0;
12211518 9200608;P9200608;CARRIL BICI CAMARA 608 GUILLEM DE CASTRO - JESÚS (GUILLEM DE CASTRO);DIRECTO;1601;2021-05-03 00:25:00;0;100;;;0;
12211519 9200608;P9200608;CARRIL BICI CAMARA 608 GUILLEM DE CASTRO - JESÚS (GUILLEM DE CASTRO);DIRECTO;1601;2021-05-03 00:30:00;0;100;;;0;
12211520 9200608;P9200608;CARRIL BICI CAMARA 608 GUILLEM DE CASTRO - JESÚS (GUILLEM DE CASTRO);DIRECTO;1601;2021-05-03 00:35:00;0;100;;;0;
```

Figura 13. Valores nulos en ocupación y fiabilidad en Notepad++

Para ver la cantidad de valores faltantes en porcentaje por tramos se realizó una visualización en *PowerBI*. Como se puede observar en la Figura 14, solo encontramos 10 tramos con datos faltantes en la variable *Ocupación* o *FiabilidadOcupación*. Esta cantidad corresponde a valores menores que el 0,10% del total de nuestros datos, así que la mejor opción para evitar futuros problemas fue eliminar esos casos.

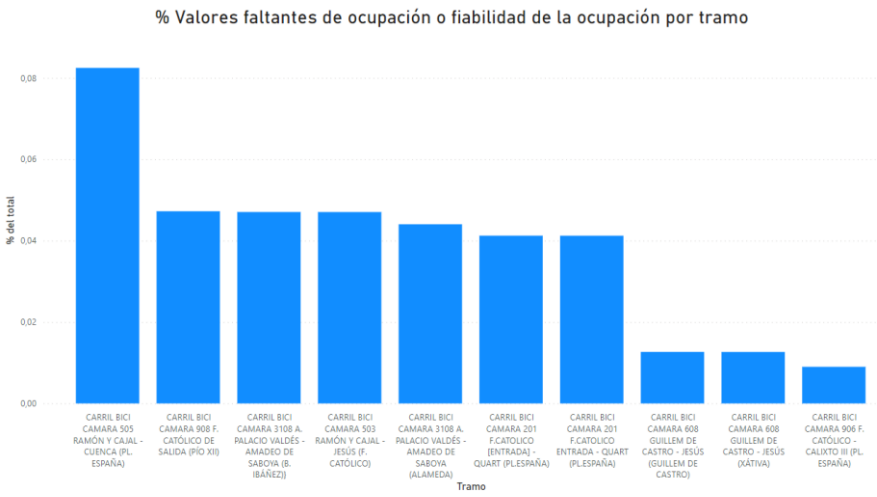


Figura 14. Porcentaje de valores faltantes según tramo

Otro paso en la limpieza de nuestra base de datos es la detección de valores con una fiabilidad menor del 95% tanto en la variable de *Intensidad* como en *Ocupación*. Este valor fue recomendado por el trabajador de la OCI y, además, en un análisis no se puede trabajar con datos no fiables, sino que se tiene que trabajar con los datos más fiables posible porque si no se pueden tomar decisiones erróneas.

Así que, se realizó un análisis de las cantidades con fiabilidad menor del 95%. Como se puede observar en la Tabla 6, hay muy pocos valores que logran alcanzar el 15%, así que teniendo en cuenta que los casos menores del 95% de probabilidad son bajos, se decidió eliminarlos.

Año	Fiabilidad_intensidad	Fiabilidad_ocupacion
2016	18 637 390 (15.32%)	18 637 390 (15.32%)
2017	13 391 301 (10.16%)	13 391 301 (10.16%)
2018	8 360 224 (6.25%)	8 360 224 (6.25%)
2019	8 996 243 (7.86%)	8 996 246 (7.86%)
2020	14 451 025 (10.51%)	14 451 025 (10.51%)
2021	11 516 416 (8.04%)	11 516 416 (8.04%)

Tabla 6. Fiabilidad menor del 95%

Una vez se ha realizado la limpieza, se realiza la agrupación de los datos según *IdPM*, *Nombre*, *Descripción*, *Fecha*, *hora*, y promedio de la *Intensidad*, de la *FiabilidadIntensidad*, de la *Ocupación* y de la *FiabilidadOcupación* ya que se encontraban recopilados cada 5 minutos y no era necesario que tuvieran tanta granularidad. A continuación, se juntaron todos los meses de cada año en un *CSV*, acción que ayudaría a tratar los datos como un bloque y no por separado siendo así más cómodo para tratamientos posteriores y, además, nos ayudaría a ahorrar tiempo.

5 Resultados obtenidos y discusión

Cuando la limpieza de los datos ya está realizada, se realiza el proceso de preanálisis, las comparaciones necesarias y, finalmente, las visualizaciones e interpretaciones para alcanzar nuestro objetivo final.

Primeramente, con la orden *glimpse()* se visualizarán las variables almacenadas, el tipo que son y los primeros valores que almacenan cada una. Como se puede observar en la Figura 15, en nuestra base de datos encontramos 61 579 355 casos y 9 variables que son: *IdPM, Nombre, Descripcion, Fecha, hora, Intensidad, FiabilidadIntensidad, Ocupación,* y *FiabilidadOcupacion.* También se puede visualizar el tipo de cada variable donde se encontrará definido como *<chr>* las variables de tipo texto, *<date>* las de tipo fecha, *<time>* de hora y por último *<dbl>* que significa dobles, o números reales.

```
Rows: 61,579,355
Columns: 9
$ IdPM                <chr> "101", "101", "101", "101", "101", "101", "101", "101", "101", "101", "1…
$ Nombre              <chr> "P101", "P101", "P101", "P101", "P101", "P101", "P101", "P101", "P101", …
$ Descripcion         <chr> "PINTOR SOROLLA, Nº 1 (de Poeta Querol a Universidad)", "PINTOR SOROLLA,…
$ Fecha               <date> 2016-01-01, 2016-01-01, 2016-01-01, 2016-01-01, 2016-01-01, 2016-01-01,…
$ hora                <time> 00:00:00, 01:00:00, 02:00:00, 03:00:00, 04:00:00, 05:00:00, 06:00:00, 0…
$ Intensidad          <dbl> 108, 126, 180, 208, 204, 237, 149, 212, 191, 248, 265, 427, 576, 631, 52…
$ FiabilidadIntensidad <dbl> 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 10…
$ Ocupacion           <dbl> 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 2, 3, 3, 2, 2, 2, 3, 4, 5, 4, 3, 2, 1, …
$ FiabilidadOcupacion <dbl> 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 10…
```

Figura 15. Información base de datos

Una vez se iba sabiendo más información de los datos con los que íbamos a trabajar y que el formato de las variables era el correcto, se pasó a realizar un resumen descriptivo de los datos resultantes con la orden *summary()*.

En la Figura 16 se puede visualizar información según el formato de la variable. Para las columnas de tipo carácter se puede observar la cantidad de datos que hay y la clase y la moda de la variable. Para las columnas de tipo numérico se puede observar el mínimo y el máximo, así como el primer y tercer cuartil y la mediana, la media y si hay algún dato faltante. En la columna de *Fecha*, se observa que los datos se encuentran distribuidos desde el 1 de enero de 2016 hasta el 31 de diciembre de 2021. En la columna de *Intensidad*, el máximo donde se puede apreciar que es un valor muy elevado ya que la mediana es de 164 y la media de 408. En este caso, sería interesante un estudio de los outliers de esta variable. En el caso de las columnas de fiabilidad, se puede deducir que la mayoría de los valores se encuentran en el 100% ya que la mediana es de 100% y la media de 99.96%. Por último, en la columna de *Ocupación* se deduce que la mayoría de los valores son pequeños ya que la mediana es 1 y la media es de 2.181 aunque encontremos uno o varios outliers hasta 100.

```
     IdPM              Nombre          Descripcion           Fecha                hora
 Length:61579355    Length:61579355    Length:61579355    Min.   :2016-01-01   Length:61579355
 Class :character   Class :character   Class :character   1st Qu.:2017-08-23   Class1:hms
 Mode  :character   Mode  :character   Mode  :character   Median :2019-02-05   Class2:difftime
                                                          Mean   :2019-02-12   Mode  :numeric
                                                          3rd Qu.:2020-08-24
                                                          Max.   :2021-12-31

   Intensidad      FiabilidadIntensidad   Ocupacion       FiabilidadOcupacion
 Min.   :      0   Min.   : 95.00       Min.   :  0.000   Min.   : 95.00
 1st Qu.:     45   1st Qu.:100.00       1st Qu.:  0.000   1st Qu.:100.00
 Median :    164   Median :100.00       Median :  1.000   Median :100.00
 Mean   :    408   Mean   : 99.96       Mean   :  2.181   Mean   : 99.96
 3rd Qu.:    505   3rd Qu.:100.00       3rd Qu.:  3.000   3rd Qu.:100.00
 Max.   :4317000   Max.   :100.00       Max.   :100.000   Max.   :100.00
```

Figura 16. Información de cada variable

Para obtener un análisis más minucioso de las variables y de los outliers que puede haber, se realiza un boxplot para cada variable numérica. El diagrama de cajas y bigotes o boxplot es una gráfica que nos permite visualizar más fácilmente medidas estadísticas de las diferentes variables numéricas de una base de datos. Esta visualización (ver Figura 17) consiste en un gráfico que comprende el mínimo y el máximo de la muestra, quitando los outliers; el primer cuartil (Q1) que corresponde al 25% de los casos; la mediana que es el valor que se encuentra en medio de la muestra; el tercer cuartil (Q3) que corresponde al 75% de los datos; el rango intercuartílico (IQR) que corresponde a la diferencia entre Q1 y Q3; y los valores atípicos que son todos aquellos valores que se encuentran fuera del mínimo y del máximo de la muestra.

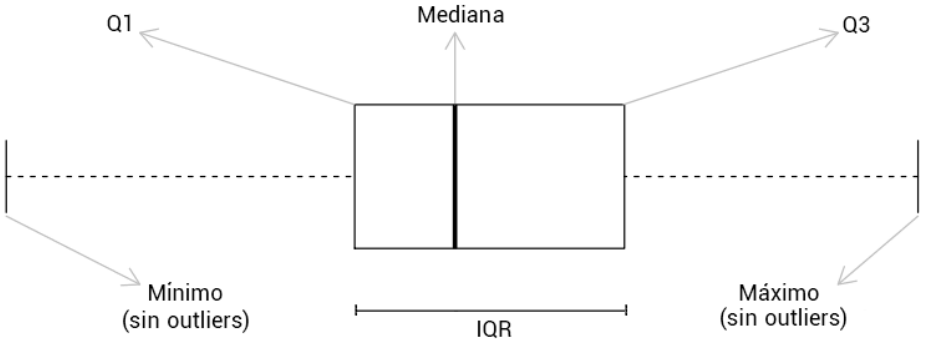


Figura 17. Boxplot (Fuente: https://r-coder.com/boxplot-en-r/)

Primeramente, se realizó un boxplot de la variable *Intensidad*. En la Figura 18 se puede observar que hay cuatro resultados: *$stats* donde se observa el rango intercuartílico que, en este caso, se encuentra de 0 a 1195, el primer cuartil ubicado en 45, la mediana, que como se ha comentado antes, se encuentra en 164 y el tercer cuartil ubicado en 505; la cantidad de datos que tenemos mostrado en *$n* que es igual a 61 579 355; *$conf*, que es el extremo superior e inferior de la muestra; y, por último, se encuentran indicados los outliers (*$out*), es decir, valores anómalos o extremos que se encuentran en la muestra. En la Figura 19 se pueden ver mostrados los valores anómalos ya que son los puntos que se encuentran fuera de los bigotes. En este caso, los outliers constituyen el 9.20% de nuestros datos. Como este valor es bastante pequeño, se eliminarán todos

los datos mayores de 1195 ya que como hemos comentado antes el rango intercuartílico comprendía desde el 0 hasta 1195.

```
$stats
[1]    0   45  164  505 1195

$n
[1] 61579355

$conf
[1] 163.9074 164.0926

$out
   [1] 1372 1498 1237 1236 1366 1482 1315 1343 1293 1302 1430 1555 1772 1884 1633 1360 1508 1538 1466
  [20] 1658 1480 1360 1492 1494 1668 1687 1631 1222 1395 1541 1631 1704 1638 1418 1274 1494 1426 1555
  [39] 1568 1246 1397 1431 1624 1659 1748 1583 1359 1340 1349 1401 1381 1334 1231 1292 1312 1475 1302
  [58] 1420 1428 1467 1513 1583 1268 1390 1444 1376 1445 1486 1524 1461 1428 1344 1205 1327 1378 1364
  [77] 1339 1223 1410 1557 1481 1446 1488 1299 1330 1457 1475 1367 1382 1460 1576 1502 1616 1409 1240
  [96] 1377 1504 1495 1538 1243 1337 1208 1418 1219 1234 1458 1425 1536 1520 1448 1369 1199 1290 1354
 [115] 1400 1303 1329 1502 1469 1495 1464 1510 1335 1300 1366 1359 1362 1203 1358 1267 1450 1349 1548
 [134] 1255 1295 1368 1406 1457 1604 1517 1470 1504 1339 1538 1502 1469 1311 1298 1489 1613 1604 1551
 [153] 1616 1549 1369 1320 1343 1426 1352 1290 1219 1360 1251 1290 1277 1688 1786 1641 1364 1384 1899
```

Figura 18. Valores boxplot Intensidad

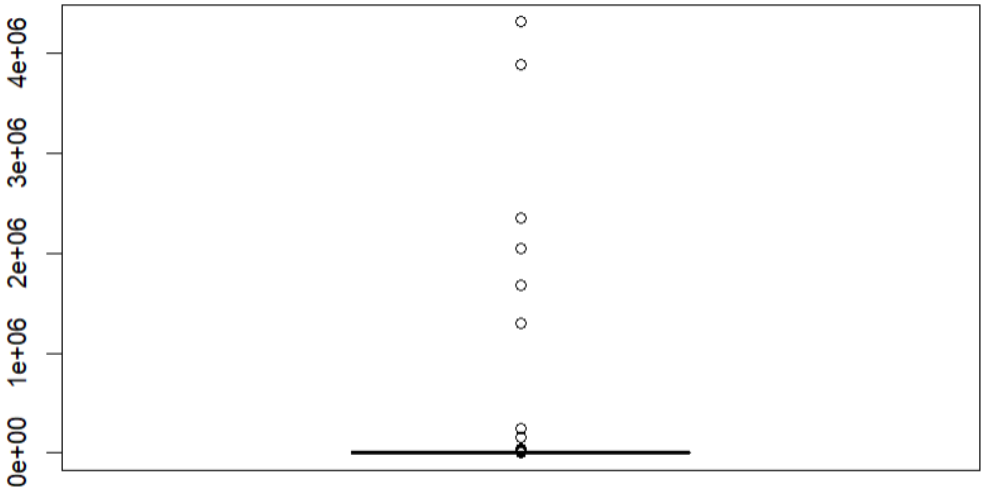


Figura 19. Boxplot Intensidad

El siguiente paso consiste en realizar el boxplot de la variable *FiabilidadIntensidad*. Como se puede observar en la Figura 20 y en la Figura 21, cualquier valor que sea diferente de 100 es anómalo, ya que el rango intercuartílico se encuentra de 100 a 100. Sacando el porcentaje de valores diferentes de 100, en este caso que se han clasificado como anómalos, constituyen el 2.12% de los datos. Por los que se eliminaran todos esos valores.

```
$stats
[1] 100 100 100 100 100

$n
[1] 55912048

$conf
[1] 100 100

$out
   [1] 97 98 98 99 99 97 98 98 98 95 96 97 97 95 95 98 98 97 98 96 96 95 95 97 96 97
95 95 96 96 97 97
  [33] 95 97 97 97 98 96 99 96 97 96 95 95 99 96 96 98 99 96 99 98 95 96 99 99 98 98
98 97 95 99 95 99
  [65] 97 95 96 99 96 98 98 99 98 99 98 99 99 99 95 99 99 99 98 98 99 97 98 99 97 99
95 98 98 99 98 99
```

Figura 20. Valores boxplot FiabilidadIntensidad

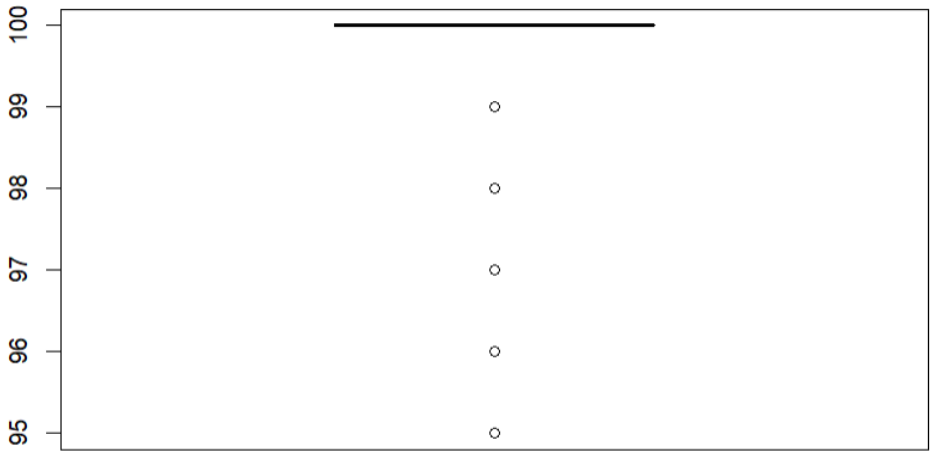


Figura 21. Boxplot FiabilidadIntensidad

Ahora pasa a realizar el boxplot de la siguiente variable numérica, *Ocupación*, y como se puede observar en la Figura 22 el rango intercuartílico engloba del 0 al 5 mientras que en la base de datos se tienen valores hasta 100 como indica en la Figura 23. El porcentaje de datos que se encuentra fuera del rango intercuartílico pertenece al 6.38%, como es un valor bastante pequeño, así que los datos anómalos se eliminarán.

```
$stats
[1] 0 0 1 2 5

$n
[1] 54728272

$conf
[1] 0.9995728 1.0004272

$out
   [1]  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  7  6  6  6  6  6  6  7  8
8  6  6  7  7  8
  [33]  6  7  7  6 11 12 10  7  6  7 10  9 12  7  8 14 15 28  7  7  6  7  9  9  7  6
6  6  7  7  6  6
  [65]  6  7  8 14  6  7  6  7  6  7  6  6  7  6  8  8  6  6  7  7  7  9  9  6  6  7
7  8  7  7  6  6
  [97]  7  7  7  6  7  6  6  6  7  9  8  7  6  6  6  7 11  8  6  6  6  7  9  7  7  7
9  7  6  6  7  7
 [129]  6  6  8  8  6  7  8  6  7  6  7  7  7  8  7  6 12  8  7  6  6  8  8  7  7 11
```

Figura 22. Valores boxplot Ocupación

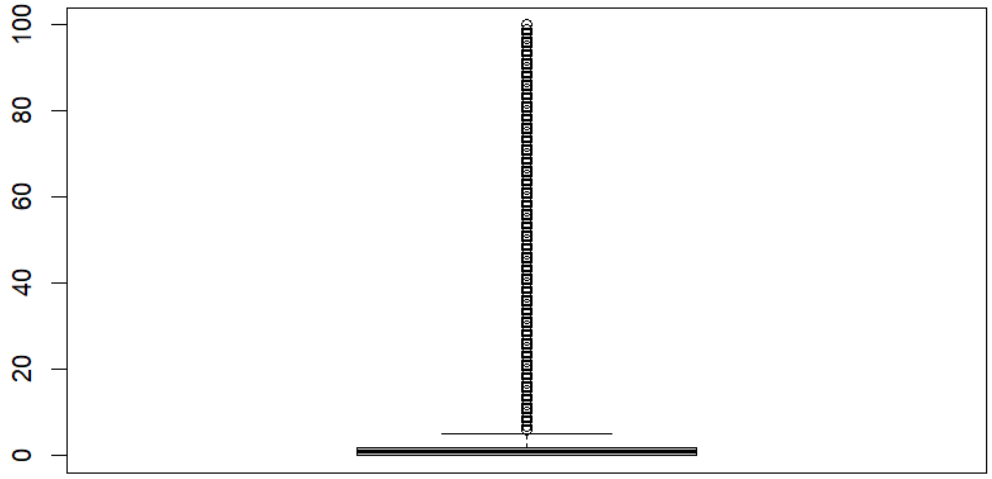


Figura 23. Boxplot Ocupación

Por último, se realiza el boxplot de la última variable numérica, *FiabilidadOcupacion*. Como se puede observar en la Figura 24 y que la Figura 25 confirma, es que solo tenemos valores de 100% y no hay ningún caso que sea un dato anómalo.

```
$stats
[1] 100 100 100 100 100

$n
[1] 51234175

$conf
[1] 100 100

$out
numeric(0)
```

Figura 24. Valores boxplot FiabilidadOcupacion

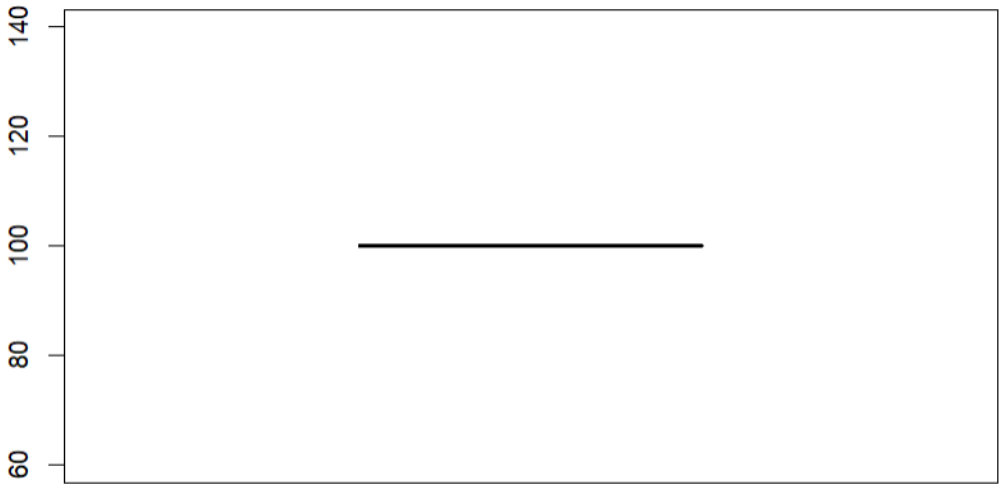


Figura 25. Boxplot FiabilidadOcupacion

Con la limpieza de datos anómalos la base de datos se ha disminuido a 51 234 175 casos. Ahora, se procede a la observación de la correlación de nuestros datos. Como se puede observar en la Figura 26, ninguna de las variables tiene correlación con ninguna otra, es decir, dos de las variables de la base de datos no están relacionadas linealmente ya que mientras una aumenta o disminuye, la otra no tiene ningún efecto. Para la gráfica de *Intensidad* vs *Ocupación*, se observa 6 líneas paralelas que demuestran que para cualquier valor de *Intensidad* se obtiene cualquier valor de *Ocupación*. Para los valores de *FiabilidadIntensidad* y *FiabilidadOcupacion* se muestra que solo se encuentra un punto que significa que solo encontramos el valor 100, tal y como se ha dejado en la limpieza de valores anómalos.

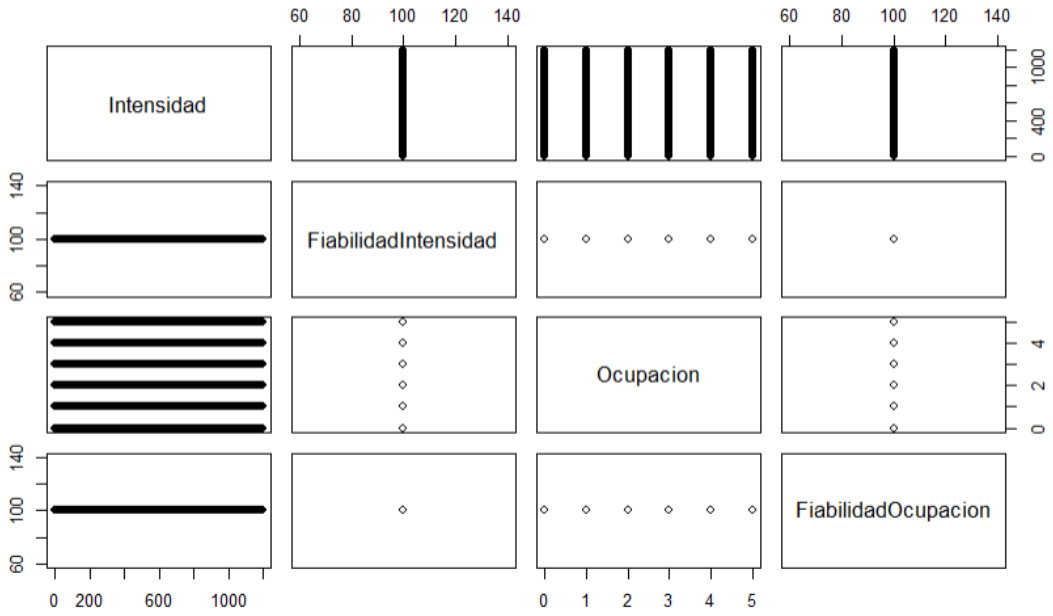


Figura 26. Correlaciones

A continuación, se pasó a realizar una comparativa por años en la ciudad de Valencia. Para ello, se utilizó la prueba de Kolmogórov-Smirnov que compara dos muestras para ver si proceden de la misma distribución. Utiliza tres hipótesis nulas: two-sided, si los dos conjuntos de datos de muestra provienen de la misma distribución; greater, si la distribución *x* es mayor que la distribución *y*; y less, si la distribución *x* es menor que la distribución *y*. La hipótesis nula se aceptará cuando el p-valor sea mayor que 0.05 y se rechazará cuando sea menor que 0.05.

Primeramente, se decidió comparar los años más recientes, 2020 y 2021. Como se puede observar en la Figura 27, la primera hipótesis nos indica si los dos conjuntos de datos de muestra provienen de la misma distribución. Si nos fijamos en el p-valor, es menor que 0.05 por lo que se rechazará la hipótesis y se aceptará la hipótesis alternativa que nos indica que los dos conjuntos de datos provienen de distribuciones distintas. La Figura 28 indica si la distribución *x*, en este caso 2020,

no está por encima de la distribución *y*, en este caso 2021. Por lo que indica el p-valor que es menor de 0.05, se rechaza la hipótesis nula y se acepta que *x* sea una distribución por encima de *y*. Por último, la Figura 29 indica si la distribución de 2020 está por encima de la distribución de 2021 y como el p-valor es mayor que 0.05 se aceptará la hipótesis nula. Por lo que, en conclusión, como se puede observar más claramente en la Figura 30 se obtiene que la distribución de 2020 no es ni menor ni igual que la de 2021, sino que es mayor.

```
D = 0.029253, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figura 27. Test Kolmogórov-Smirnov 2020-2021 two-sided

```
D^+ = 0.029253, p-value < 2.2e-16
alternative hypothesis: the CDF of x lies above that of y
```

Figura 28. Test Kolmogórov-Smirnov 2020-2021 greater

```
D^- = 2.4786e-16, p-value = 1
alternative hypothesis: the CDF of x lies below that of y
```

Figura 29. Test Kolmogórov-Smirnov 2020-2021 less

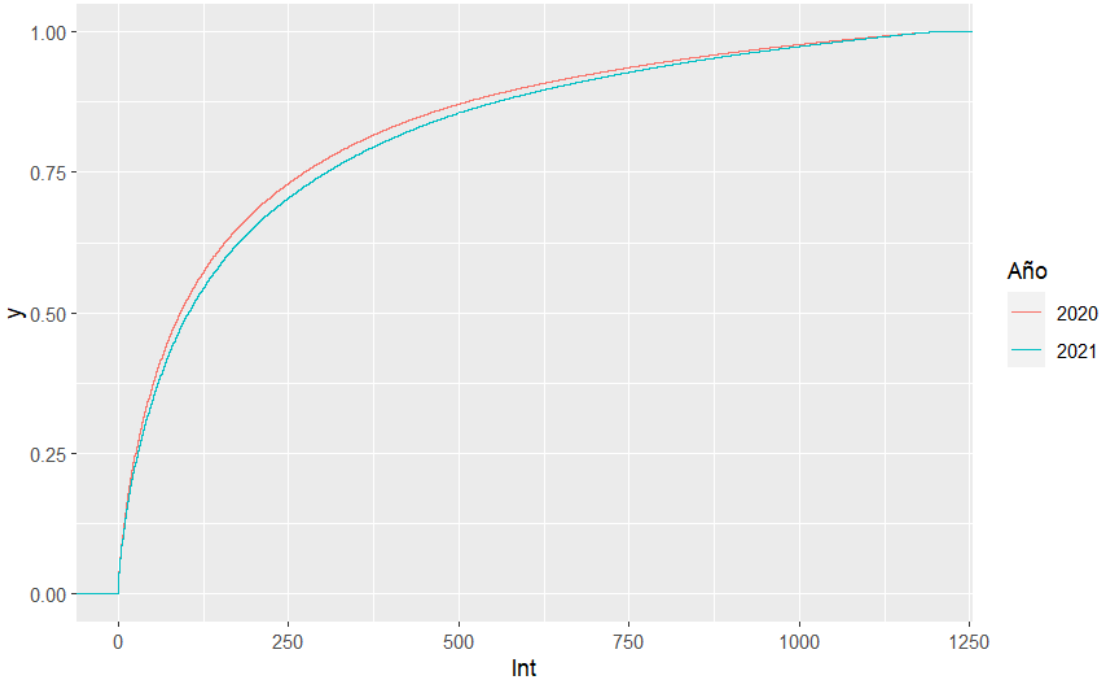


Figura 30. Ecdf 2020-2021

Otro análisis importante es la comparación de los años 2016 y 2021, es decir, el primer y último año con los que se está trabajando para ver si las distribuciones son iguales o hay alguna mayor que la otra. Este análisis es importante ya que podemos llegar a la conclusión de si por lo general el uso del automóvil ha aumentado con los años o en cambio, ha disminuido o sigue igual.

En las pruebas de Kolmogórov-Smirnov que se han realizado, se puede observar que en la Figura 31, donde se compara si las dos distribuciones son iguales, se obtiene un p-valor por debajo de 0.05 por lo que se rechazará la hipótesis nula y se aceptará la hipótesis alternativa que afirma que los dos conjuntos de datos provienen de distribuciones distintas. A continuación, se realizó la comparación *greater*, mostrada en la Figura 32, donde la hipótesis nula afirma que la distribución de 2016 no está por encima de la distribución de 2021 ya que en los resultados que se han obtenido, el p-valor es mayor que 0.05 con lo que se afirma la hipótesis nula. Por último, la Figura 33 muestra la comparación *less*, donde la hipótesis nula indica si la distribución de 2016 está por encima de la distribución de 2021 y como el p-valor es menor que 0.05 se rechazará. En conclusión, se obtendrá que los conjuntos de datos de 2016 y de 2021 no provienen de la misma distribución ni que 2016 está por encima de 2021. Todas estas hipótesis que se han aceptado o rechazado se muestran más claramente en la Figura 34, donde vemos que la distribución de 2021 es mayor que la de 2016. Este estudio nos indica que, en rasgos generales, el uso del automóvil ha crecido desde 2016.

```
D = 0.078746, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figura 31. Test Kolmogórov-Smirnov 2016-2021 two-sided

```
D^+ = -1.3864e-14, p-value = 1
alternative hypothesis: the CDF of x lies above that of y
```

Figura 32. Test Kolmogórov-Smirnov 2016-2021 greater

```
D^- = 0.078746, p-value < 2.2e-16
alternative hypothesis: the CDF of x lies below that of y
```

Figura 33. Test Kolmogórov-Smirnov 2016-2021 less

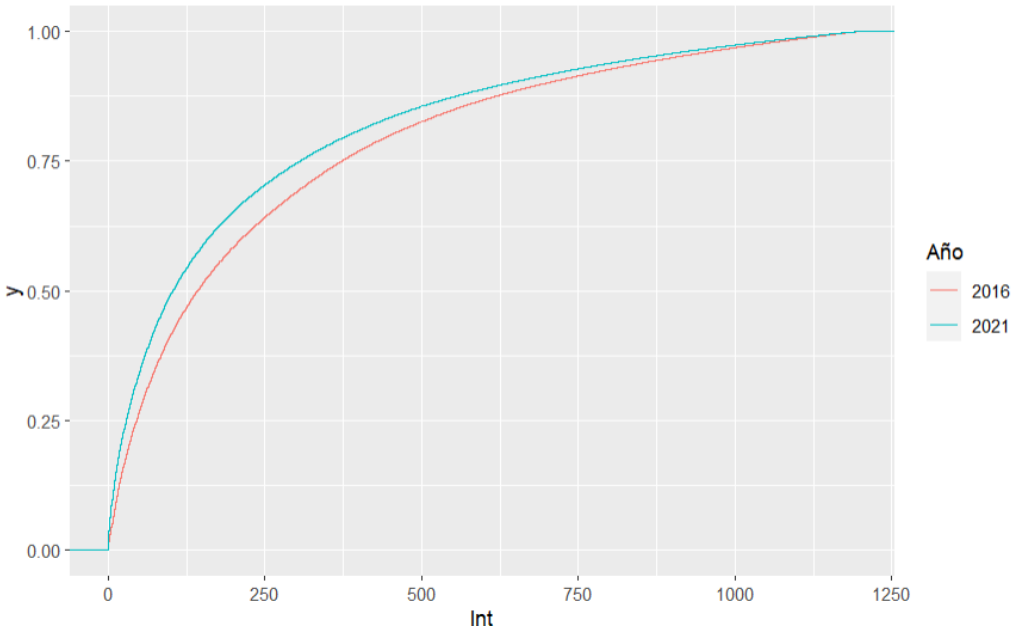


Figura 34. Ecdf 2016-2021

Un gráfico relevante para ver si hay diferencias significativas entre dos muestras, es el gráfico cuantil-cuantil (Q-Q). Este gráfico muestra como distribución normal una línea recta y una línea más gruesa (dx,dy) de las dos muestras que tenemos. Si la tendencia de los puntos (dx,dy) forman una línea recta con pendiente 1 que pasa por la media, se puede afirmar que ambas muestras tienen la misma distribución. En caso contrario, cuanto más se aleje esa línea de la distribución normal, menos parecidas serán las distribuciones.

En la Figura 35 se puede observar el gráfico Q-Q de la variable *Intensidad* de 2020 y de 2021 que se demuestra que las dos muestras vienen de la misma distribución ya que la línea que forman es muy parecida a la formada por una distribución normal. En la Figura 36 se ha obtenido un gráfico Q-Q para cada año que tenemos comparado con todos los demás. Como se puede observar, las comparaciones entre los años 2016, 2017, 2018, 2019 y 2021 siguen una distribución muy parecida a la normal con lo que se podría concluir que todas provienen de la misma distribución. En el caso de las comparaciones con 2020, se puede observar una ligera curvatura en la parte del centro de los datos que podría ser originado por los cambios de los meses de confinamiento. Como ese cambio no es muy significativo, se podría afirmar que provienen de distribuciones muy similares.

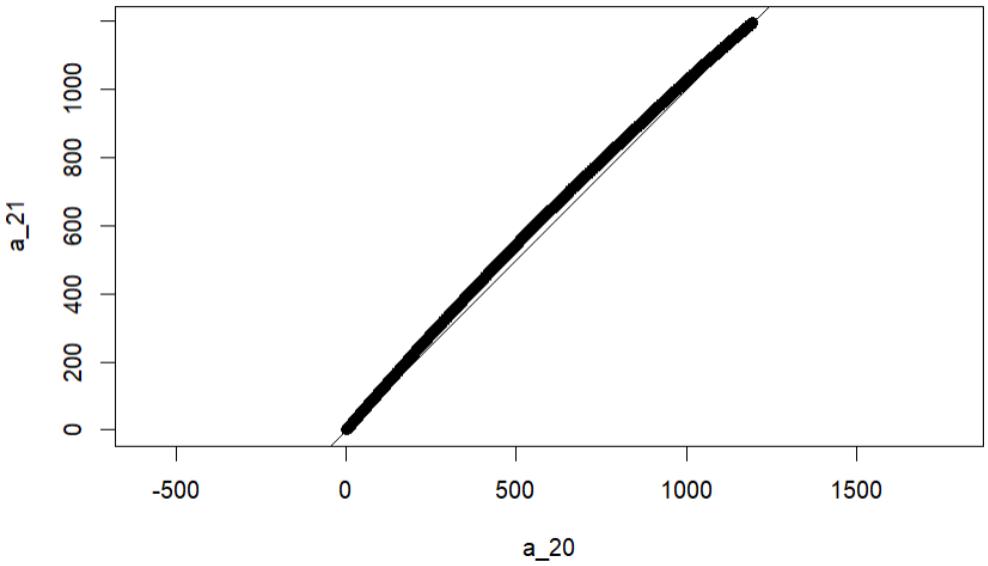


Figura 35. Gráfico Q-Q 2020-2021

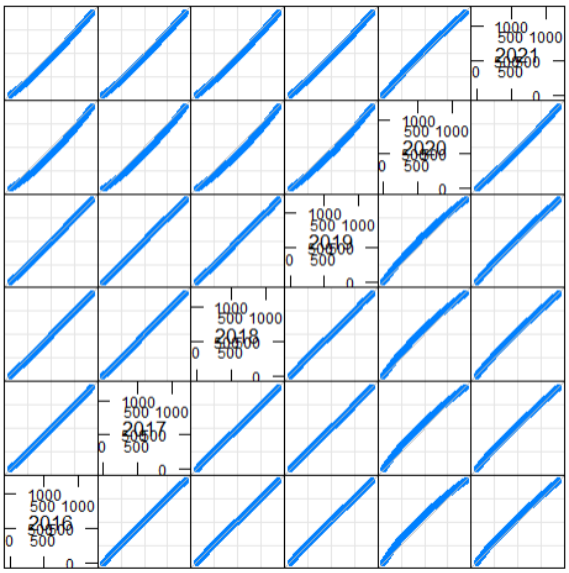


Figura 36. Gráfico Q-Q de todos los años

Una vez se han comparado los años de forma general, se podría separar el estudio en comparar por una parte los días de entre semana (lunes, martes, miércoles, jueves y viernes) para cada año y los días de fin de semana (sábado y domingo). Para ello, se realizará el gráfico Q-Q para comparar las distribuciones de un año respecto a otro. Como se puede observar en la Figura 37, donde se muestran la comparación entre los días de la semana de los diferentes años, existe una línea recta cuando comparamos todos los años excepto 2020 donde podemos observar que la línea es un poco más curva. En la Figura 38, donde se muestran los días de fin de semana, se puede observar los mismos resultados que en la Figura 37 pero la curva de las comparaciones con 2020 es más notable.

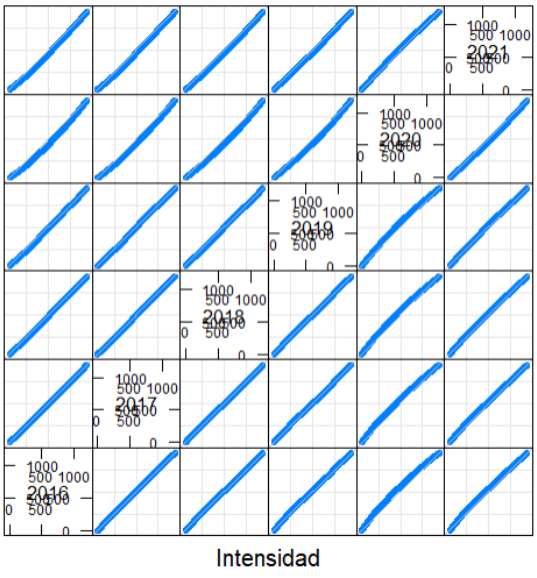


Figura 37. Gráfico Q-Q de los días entre semana

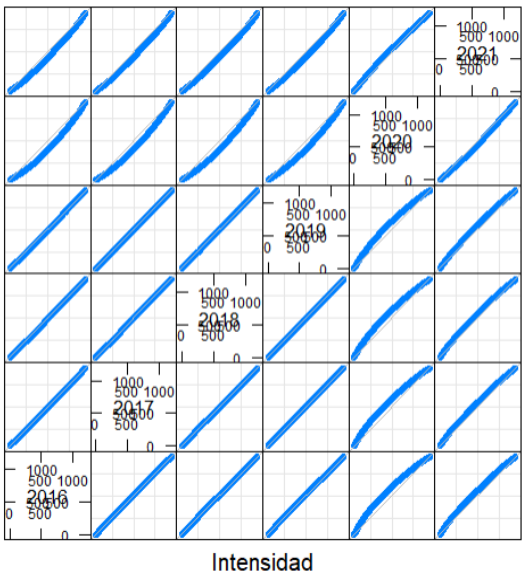


Figura 38. Gráfico Q-Q de los días de fin de semana

Para la visualización de los datos se cambió a *Power BI*, un software que permite la visualización de datos más fácilmente y la automatización de un gran volumen de trabajo de preparación de informes. *Power BI* tiene 3 secciones: Informe donde se encuentran las visualizaciones, Datos donde se encuentra un resumen de los datos que estamos utilizando y Modelo que es donde se encuentran las relaciones de cada tabla.

El primer paso para trabajar con *Power BI* consiste en cargar los datos. Para ello, en la parte de Inicio > Datos se encuentran diversas opciones según el tipo de archivo que vas a cargar. En nuestro caso, el archivo es un documento de *Texto o CSV* y se cargará. La tabla con la información del documento cargado se quedará con el nombre Datos (ya que es el nombre de nuestro documento).

Una vez se han cargado los datos con los que vamos a trabajar, se pasó a analizar de qué forma era mejor tratar las fechas y horas de nuestros datos. *Power BI* automáticamente crea la jerarquía Año-Trimestre-Mes-Fecha usando la inteligencia del tiempo, pero esta opción no es del todo adecuada ya que nos limita en muchas ocasiones. En este trabajo se ha decidido desactivar la opción de inteligencia de tiempo automática ya que aparte de las limitaciones que conlleva, se necesita crear más atributos de fecha. Para ello, se desactiva la opción de inteligencia de tiempo en Opciones > Carga de datos y se desmarcará la opción de: Fecha y hora automáticas para archivos nuevos (ver Figura 39).

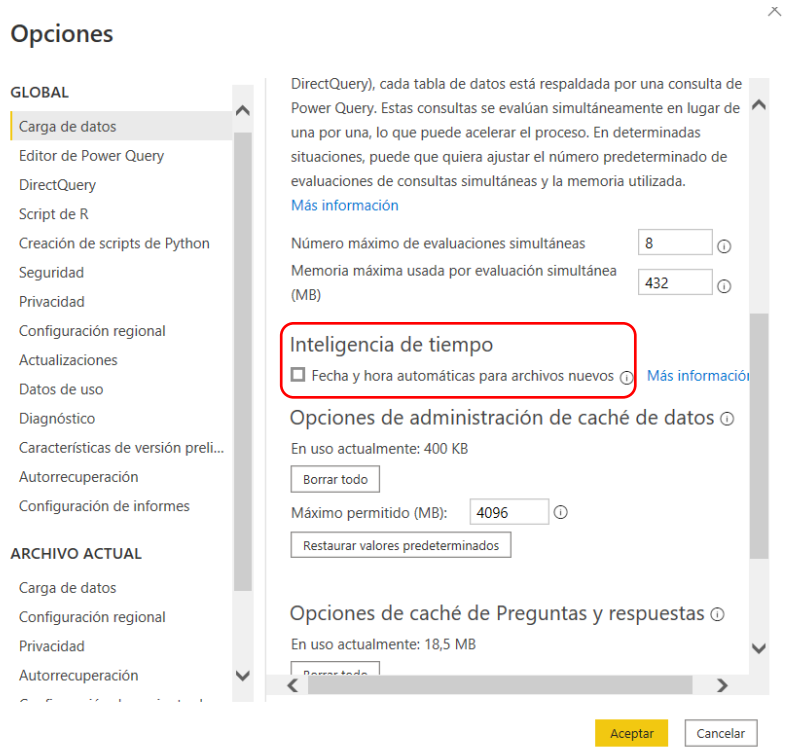


Figura 39. Inteligencia de tiempo

Ahora el siguiente paso consiste en crear una nueva tabla para convertirla en tabla de tiempo. Para ello, se utilizará DAX, un lenguaje de *Power BI* que permite crear objetos y hacer consultas. Para crearla, se irá a la sección de Datos y se creará una nueva tabla donde incluiremos diversas funciones (ver Figura 40). En primer lugar, se escribirá la función *ADDCOLUMNS* donde incluirá todas las demás funciones y permitirá añadir columnas por cada separación de coma. En segundo lugar, se obtendrá la columna Date que será creada con la fecha de inicio y fin de la columna Fecha de la tabla cargada inicialmente. En tercer lugar, se obtendrá el año, el número del mes y el nombre, el número y el nombre del día de la semana y la semana del año que corresponde.

```
1  Calendario =
2  ADDCOLUMNS(
3      CALENDAR(MIN(Datos[Fecha]),MAX(Datos[Fecha])),
4      "Nº Año",YEAR([Date]),
5      "Nº Mes",MONTH([Date]),
6      "Nombre Mes",FORMAT([Date],"mmmm"),
7      "Día de la semana",FORMAT([Date],"dddd"),
8      "Semana del año",WEEKNUM([Date]),
9      "Weekday", WEEKDAY([Date],2)
10 )
```

Figura 40. Calendario utilizando DAX

Una vez se ha creado la tabla Calendario, *Power BI* debe de saber que corresponde a una tabla de fechas. Para ello, se pulsará con el botón derecho en la tabla Calendario y se elegirá la opción Marcar como tabla de fechas > Marcar como tabla de fechas (ver Figura 41). A continuación, aparecerá una ventana (ver Figura 42) donde habrá que indicar que columna deseamos que se convierta en la columna de fechas. En este caso, se elegirá la columna Date y como la tabla es correcta y cumple con los requisitos, será validada correctamente.

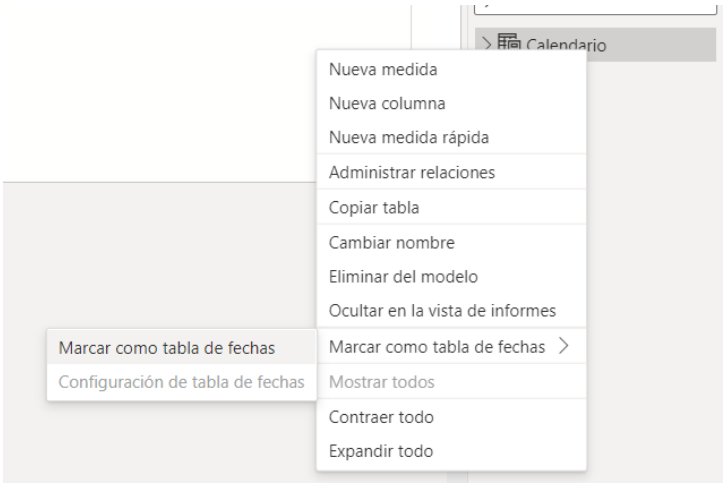


Figura 41. Marcar como tabla de fechas

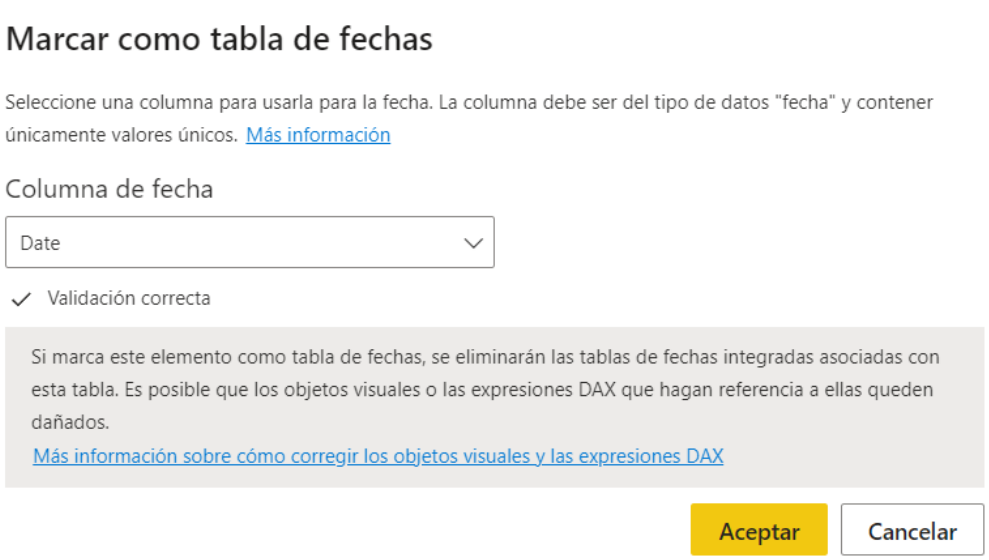


Figura 42. Columna de fecha para marcar como tabla de fechas

Ahora que se ha creado la tabla de Calendario se necesita relacionar con nuestra tabla Datos que es la que se ha cargado inicialmente. Para ello, se pasará al apartado Modelo y se arrastrará la columna Date de la tabla Calendario a la variable Fecha de la tabla Datos (ver Figura 43). Automáticamente, se crea la relación 1 a Varios (*) que indicará que una fecha de la tabla Calendario puede aparecer en múltiples filas de la tabla Datos.

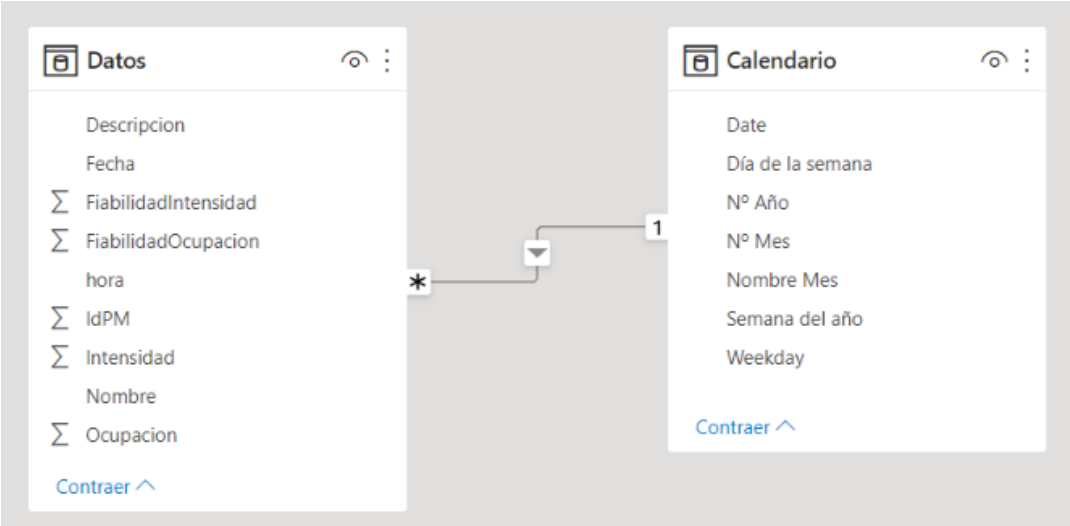


Figura 43. Relaciones Power BI

Cuando ya se han cargado los datos necesarios y se han obtenido las relaciones oportunas, se pasa a la visualización de los datos.

Primeramente, se pasó a visualizar el promedio de la intensidad por mes anualmente. En el eje de abscisas se encuentran los meses del año desde enero hasta diciembre, en el eje de ordenadas se encuentra el promedio de intensidad y cada línea corresponde a un año. Como se puede observar en la Figura 44, los años 2016, 2017, 2018 y 2021, el promedio de la intensidad tiene comportamientos similares mensualmente, aunque diferentes valores de intensidad. En cambio, el año 2020 tiene un cambio muy brusco entre los meses de febrero y junio que podría corresponder a los meses de confinamiento por COVID-19 donde solo se salía de casa para obtener productos de primera necesidad en supermercados. En el caso de 2019, se observan dos picos en el mes de abril y agosto donde habría que hacer un estudio más exhaustivo para comprender el comportamiento.

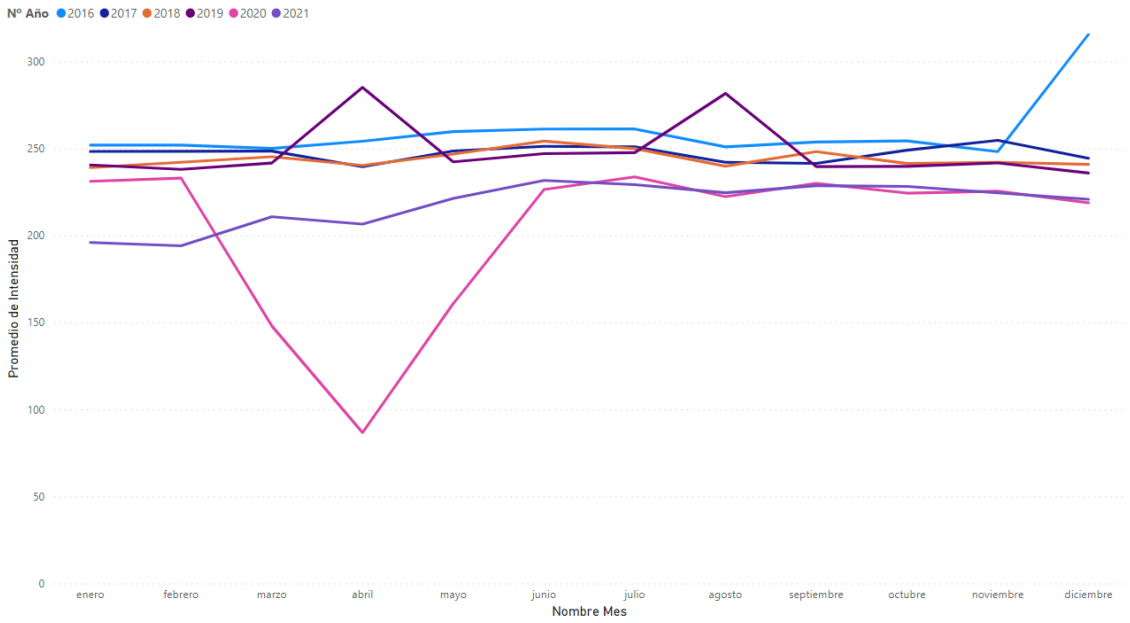


Figura 44. Promedio de la intensidad mensualmente desde 2019 hasta 2021

Una visualización también interesante es el promedio de la intensidad (ver Figura 45) y de la ocupación (ver Figura 46) por día de la semana y por año. Como se puede observar en la Figura 45, todos los años tienen un comportamiento similar. Durante todos los años el valor promedio de la intensidad va aumentando ligeramente hasta que llegan los sábados donde existe una disminución. Esto se puede deber a que los días entre semana la mayoría de las personas en la ciudad de Valencia utiliza el coche para ir a trabajar, ir a la universidad, etc. En cambio, los fines de semana debido a que la gente no trabaja y aprovecha para descansar, el uso del automóvil disminuye. En la Figura 46 se puede observar que durante la semana existe un ligero crecimiento de la ocupación, es decir, conforme va pasando la semana el tiempo que están parados los coches es mayor pero una vez llega el viernes, la ocupación va disminuyendo rápidamente hasta que acaba el domingo.

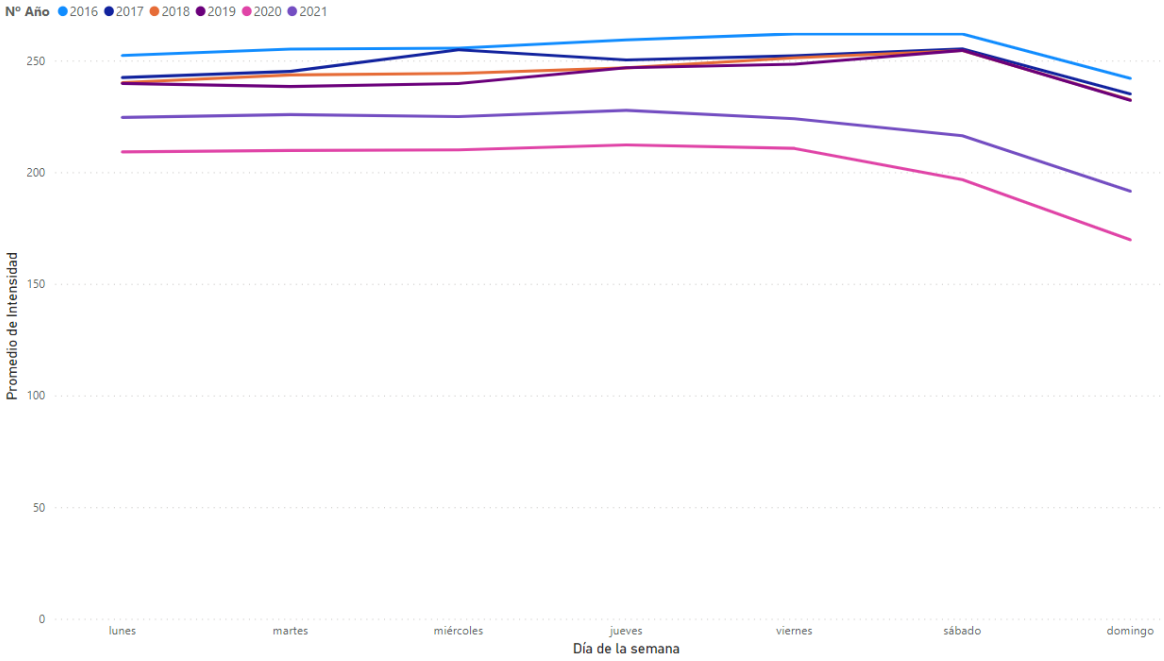


Figura 45. Promedio de la intensidad por día de la semana y año

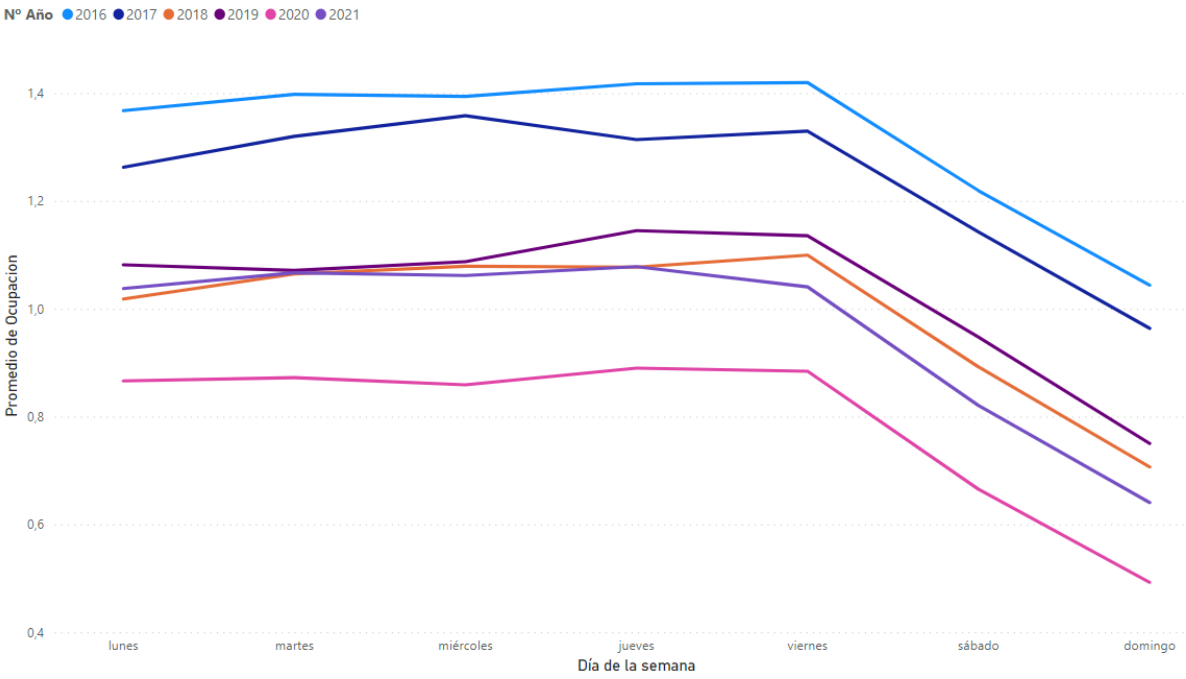


Figura 46. Promedio de la ocupación por día de la semana y año

Saber qué horas está más concurrida la ciudad de Valencia nos puede ayudar a sacar múltiples conclusiones. Para ello, se visualizará en dos gráficos, uno para las horas entre semana (ver Figura 47) y otro para los fines de semana (ver Figura 48), donde el eje de abscisas mostrará las 24h de un día, el eje de ordenadas el promedio de la intensidad y cada línea representará un año.

En el caso de visualizar el promedio de intensidad para cada hora durante los días entre semana, es decir, de lunes a viernes, se puede observar en la Figura 47 que existe una subida de la intensidad sobre las 4 am hasta las 7 am y se mantiene hasta las 7 pm con un promedio de entre 250 y un poco más de 350 vehículos la hora y con una bajada ligera a las 2 pm. Estos resultados podrían corresponder al inicio y final del turno matutino y vespertino y la bajada a la hora de comer. A partir de las 8 pm la intensidad comienza a disminuir hasta las 4 am. En cuanto a los años, se puede observar que todos menos 2020, donde el promedio de la intensidad es menor, tienen un comportamiento similar. Esto podría venir explicado por la enfermedad llamada COVID-19 que hizo que estuviéramos en confinamiento múltiples meses y repercutió de forma negativa en la intensidad del tráfico.

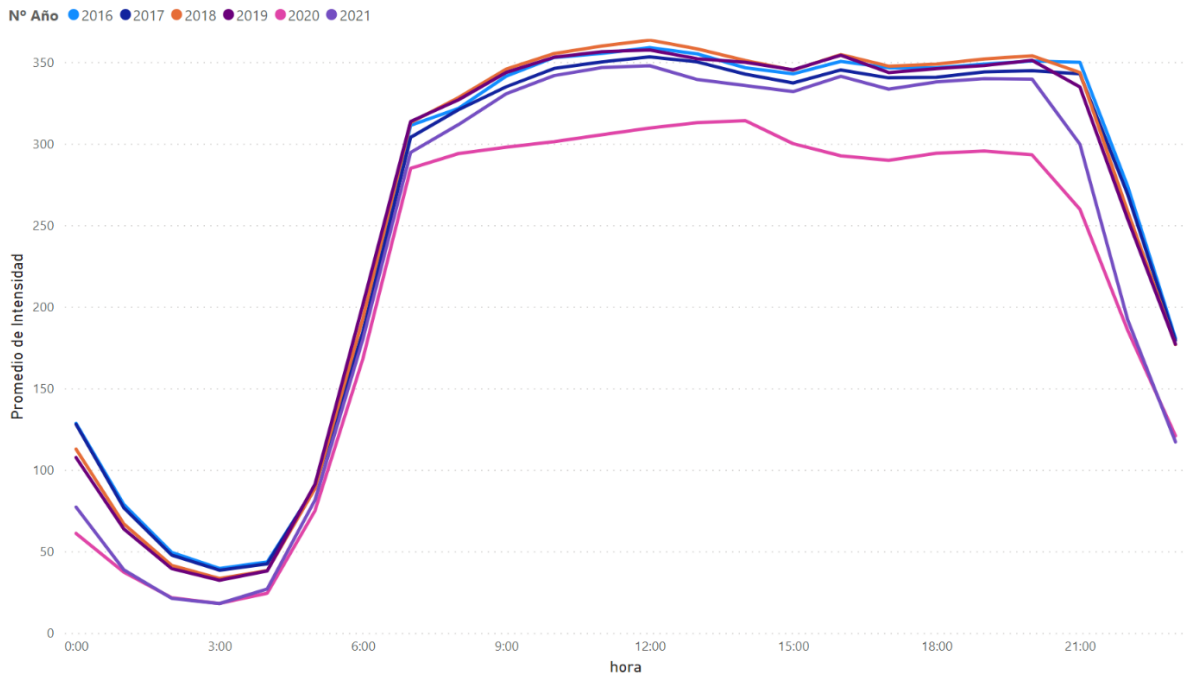


Figura 47. Promedio de la intensidad por horas entre semana para los años estudiados

Si nos fijamos en la Figura 48, que nos indica el promedio de la intensidad por horas los fines de semana, es decir, los sábados y los domingos, se observan subidas y bajadas a lo largo del día. Entre las 4 am y la 1 pm existe una subida más ligera que en el gráfico anterior, entre la 1 y las 3 pm una bajada que puede corresponder a las horas de la comida donde el tráfico disminuye, entre las 3 y 7 pm una subida y finalmente, una bajada entre las 7 pm y las 4 am. Más o menos, todos los años tienen un comportamiento similar, pero 2020 y 2021 tienen menor tráfico los fines de semana. Como se puede observar en ambos gráficos, la actividad comienza sobre las 4 am, pero entre semana va acabando sobre las 8 pm y los fines de semana entre las 7 pm con una disminución del tráfico entre las 1 y 3 pm.

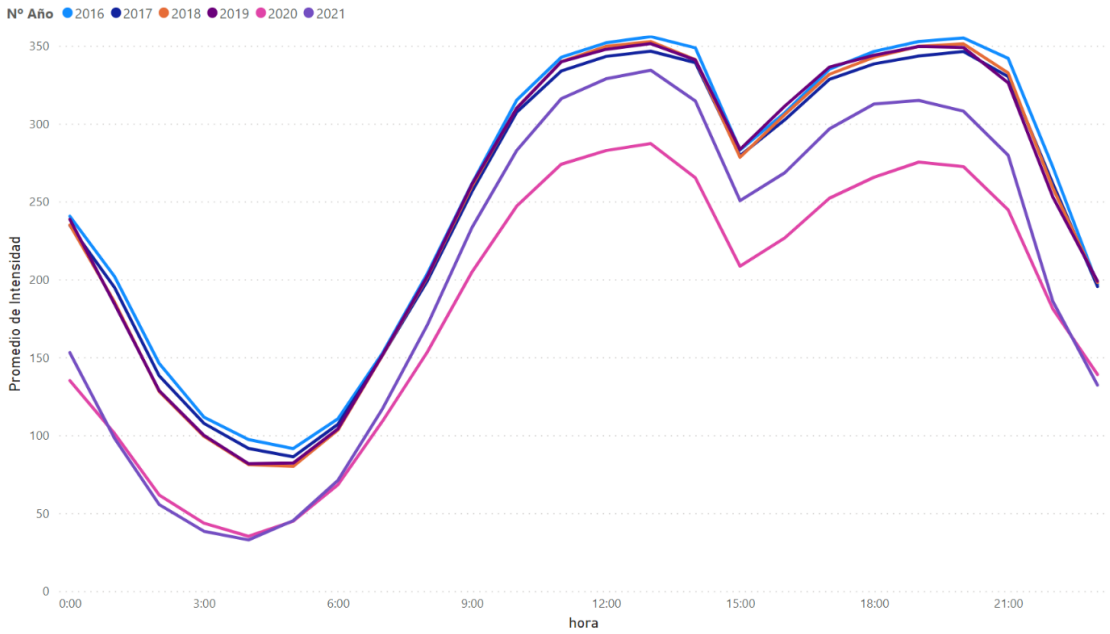


Figura 48. Promedio de la intensidad por horas fines de semana para los años estudiados

6 Conclusiones

El problema que se quería llegar a resolver era si en la ciudad de Valencia existían patrones en el tráfico de coches según el paso de los años realizando un análisis descriptivo de los datos que obtuvimos mediante la OCI. Para ello, se propusieron varios objetivos específicos para lograr alcanzar el objetivo final. Estos objetivos específicos consistían en la limpieza de los datos para ayudar al tratamiento posterior, transformación de los datos a un formato manejable, realizar un preanálisis de los datos de los coches para eliminar posibles outliers, comparar las distribuciones de cada año y realizar visualizaciones representativas. Todo esto, se quería lograr aplicando los conocimientos de análisis de datos que se obtuvieron en el grado de Ciencia de Datos de la Universidad Politécnica de Valencia.

Para lograr todos los objetivos que se propusieron, aparte de saber conocimientos de análisis de datos, también fue necesario documentarse de varias páginas web para aprender de estudios realizados anteriormente o fortalecer conocimientos. Se realizó la limpieza de los datos eliminando variables que en un futuro no nos iban a servir y se crearon otras nuevas que eran importantes para el estudio. Se realizaron agrupaciones para disminuir la granularidad de los datos. Se conocieron más los datos realizando previsualizaciones y para los outliers, se visualizaron gráficos de caja y bigotes. Por último, se compararon las distribuciones de cada año y se visualizaron diferentes gráficos para fortalecer los resultados obtenidos.

Para implementar el trabajo realizado, una visualización interesante que nos podría ayudar a sacar conclusiones más específicas por tramo sería realizar un mapa. Este mapa se podría obtener buscando o realizando un código que obtuviera las coordenadas de cada tramo y visualizarlas con la intensidad de los coches para cada tramo. Esta visualización es interesante ya que obtendríamos una vista general del tráfico de coches en la ciudad de Valencia y podríamos observar las zonas con mayor o menor intensidad para en un futuro ver el porqué del comportamiento de esa zona de la ciudad. También, se podrían utilizar métodos o técnicas que realizarán la recolección de datos de una forma más fiable y sin tantos fallos. Además de obtener datos correctos para tomar buenas decisiones, también se ahorraría tiempo de limpieza de datos ya que como podemos observar en este trabajo ha sido donde más tiempo se ha invertido.

Para realizar este trabajo se han utilizado los conocimientos aprendidos en el grado de Ciencia de Datos de la Universidad Politécnica de Valencia. Este trabajo se encuentra muy vinculado a diversas asignaturas impartidas en el grado. Gracias a la asignatura de Análisis Exploratorio de Datos, Modelos Descriptivos I y II se ha podido alcanzar el objetivo final por medio de los objetivos específicos. Este trabajo consistía en un análisis descriptivo realizado con los conocimientos aprendidos en estas asignaturas ya que se aprendió a tratar los datos desde la

obtención hasta la realización de modelos pasando por la limpieza de los datos. Otras asignaturas que han tenido relevancia en este trabajo son Gestión de Datos, Bases de Datos y Visualización. En estas asignaturas se aprendió a usar el software de *Power BI,* realizar visualizaciones estéticamente atractivas y de fácil comprensión en una primera vista, y los conocimientos necesarios para tratar los datos en una base de datos y realizar las correctas relaciones entre tablas.

En el grado de Ciencia de Datos, hemos tenido varios trabajos que resolver cada año en las asignaturas de Proyecto, pero nunca había realizado uno sola ni me veía capaz de poder alcanzar el objetivo final y, menos, con las horas de trabajo que había que invertir. He aprendido muchas funciones que antes no conocía de *RStudio,* nuevas aplicaciones como *Notepad++* y he reforzado mis conocimientos en *Power BI*. Gracias a este trabajo, he podido comprender muchos de los conocimientos aprendidos en el grado que antes o no tenía claro o no sabía y que el dato está en continuo tratamiento de limpieza, que es un proceso muy largo.

7 Bibliografía

[1] M. A. Mateo Pla *et al.*, "From traffic data to GHG emissions: A novel bottom-up methodology and its application to Valencia city," *Sustain Cities Soc*, vol. 66, p. 102643, Mar. 2021, doi: 10.1016/J.SCS.2020.102643.

[2] "¿Por qué es importante hablar sobre Cambio Climático? - Dirección de Cambio Climático." https://cambioclimatico.go.cr/por-que-es-importante-hablar-sobre-cambio-climatico/ (accessed Mar. 29, 2022).

[3] "Qué es el cambio global." https://cambioglobal.uc.cl/comunicacion-y-recursos/que-es-el-cambio-global (accessed Mar. 07, 2022).

[4] "GASES DE EFECTO INVERNADERO." http://www.ccpy.gob.mx/cambio-climatico/gases-efecto-invernadero.php (accessed Mar. 07, 2022).

[5] "¿Cuáles son las consecuencias del efecto invernadero? | Ingredientes que Suman." https://blog.oxfamintermon.org/cuales-son-las-consecuencias-del-efecto-invernadero/ (accessed Mar. 07, 2022).

[6] "Efecto invernadero, causas y soluciones | Naturgy." https://www.naturgy.es/blog/hogar/efecto_invernadero_soluciones?page=1 (accessed Mar. 07, 2022).

[7] "Siete maneras en que las ciudades pueden actuar contra el cambio climático | CMNUCC." https://unfccc.int/es/blog/siete-maneras-en-que-las-ciudades-pueden-actuar-contra-el-cambio-climatico (accessed Mar. 07, 2022).

[8] "Oficina de Ciudad Inteligente - València Ciudad Inteligente." https://smartcity.valencia.es/oficina-ciudad-inteligente/ (accessed Mar. 07, 2022).

[9] "Bienvenida a 2022 con temperaturas 'inusuales' para esta época del año | Onda Regional de Murcia." https://www.orm.es/informativos/noticias-2021/bienvenida-a-2022-con-temperaturas-inusuales-para-esta-epoca-del-ano/ (accessed Mar. 09, 2022).

[10] "Las 12 metodologías más populares para la gestión de proyectos • Asana." https://asana.com/es/resources/project-management-methodologies (accessed Mar. 29, 2022).

[11] T. John W., Exploratory Data Analysis.

[12] "Una historia muy breve de la ciencia de datos." https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/?sh=6f9602e255cf (accessed Aug. 26, 2022).

[13] "¿Qué es exactamente una ciudad inteligente?" https://blog.bismart.com/que-es-exactamente-una-ciudad-inteligente (accessed May 31, 2022).

[14] "Europa apuesta por 100 Ciudades Inteligentes en 2030 - Andina Link Smart Cities." https://www.andinalinksmartcities.com/europa-apuesta-por-100-ciudades-inteligentes-en-2030/ (accessed May 31, 2022).

[15] "5 Smart Cities europeas que destacan por su innovación - Zemsania Global Group." https://zemsaniaglobalgroup.com/5-smart-cities-europeas-que-destacan-por-su-innovacion/ (accessed May 03, 2022).

[16] "Valencia traffic report | TomTom Traffic Index." https://www.tomtom.com/en_gb/traffic-index/valencia-traffic/ (accessed Apr. 07, 2021).

[17] "Search | Mendeley." https://www.mendeley.com/search/ (accessed Aug. 22, 2022).

[18] A. Jurado, "Fundamentos de Modelado en Estrella".

[19] "Portada - Desarrollo Sostenible." https://www.un.org/sustainabledevelopment/es/ (accessed Aug. 23, 2022).

8 ANEXO: Objetivos de Desarrollo Sostenible

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. Fin de la pobreza.				X
ODS 2. Hambre cero.				X
ODS 3. Salud y bienestar.	X			
ODS 4. Educación de calidad.			X	
ODS 5. Igualdad de género.				X
ODS 6. Agua limpia y saneamiento.				X
ODS 7. Energía asequible y no contaminante.	X			
ODS 8. Trabajo decente y crecimiento económico.			X	
ODS 9. Industria, innovación e infraestructuras.		X		
ODS 10. Reducción de las desigualdades.				X
ODS 11. Ciudades y comunidades sostenibles.	X			
ODS 12. Producción y consumo responsables.	X			
ODS 13. Acción por el clima.	X			
ODS 14. Vida submarina.				X
ODS 15. Vida de ecosistemas terrestres.		X		
ODS 16. Paz, justicia e instituciones sólidas.				X
ODS 17. Alianzas para lograr objetivos.			X	

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

Los 17 Objetivos de Desarrollo Sostenible, aprobados por la ONU en 2015, nacen de la necesidad de mejorar la vida de todos [19]. Incluye numerosos objetivos relacionados con las personas, la prosperidad, la paz, las alianzas y el planeta. En este trabajo encontramos objetivos que se encuentran relacionados alta, media o bajamente o incluso que no se encuentran nada relacionados. A continuación, se explicará él porque del grado de relación que se encuentra.

Los Objetivos de Desarrollo Sostenible que se encuentran relacionados altamente con nuestro trabajo son:

- Salud y bienestar.
- Energía asequible y no contaminante.
- Ciudades y comunidades sostenibles.
- Producción y consumo responsables.
- Acción por el clima.

Como se puede observar, en este trabajo se realiza un análisis descriptivo de los datos de tráfico de los coches de la ciudad de Valencia. Esto se encuentra altamente relacionado con el cambio climático y las llamadas Smart Cities o ciudades sostenibles, ya que con los resultados obtenidos se podría realizar mejoras en la ciudad para disminuir los gases de efecto invernadero. Para ello, se podrían llegar a hacer campañas para el incremento del uso de la bicicleta añadiendo nuevos o reacondicionar los carriles bici que se encuentran repartidos por la ciudad de Valencia. Por todo esto, los ODS presentados anteriormente se encuentran relacionados altamente con nuestro trabajo ya que se busca convertir las ciudades en sostenibles para disminuir los gases de efecto invernadero con energía no contaminante como podría ser el uso de la bicicleta y que para los ciudadanos producirá una mejora de la salud tanto por el ejercicio realizado como por los gases que dejan de liberar.

En cambio, también existen ODS que se encuentran con una relación media y son:

- Industria, innovación e infraestructuras.
- Vida de ecosistemas terrestres

Estos objetivos se encuentran relacionados con un grado medio ya que, aunque la finalidad de estos sea mejorar la vida de los ciudadanos tiene menos que ver con los resultados obtenidos o con los trabajos futuros de este trabajo. Gracias a este trabajo, se podrían mejorar algunas de las infraestructuras que se encuentran en Valencia que ayudaría a convertir la ciudad en una Smart City para obtener o utilizar la energía de maneras más sostenibles.

A continuación, mostraremos todos aquellos ODS que se encuentran relacionados con nuestro trabajo de una forma baja:

- Educación de calidad.
- Trabajo decente y crecimiento económico.
- Alianzas para lograr objetivos.

Podríamos relacionar estos objetivos con nuestro trabajo, ya que, aunque no estén relacionados activamente con nuestro trabajo, podría ayudar a desarrollarlos. Si construimos o mejoramos los carriles bici que se encuentran por la ciudad de Valencia y hacemos campañas, ayudaríamos a la utilización de la bicicleta y a una disminución del uso del automóvil aplicando así una educación de calidad y en un crecimiento económico ofreciendo nuevos empleos con la finalidad de crear alianzas para lograr objetivos.

En cambio, estos objetivos de los ODS aprobados por la ONU no se encuentran relacionados de ninguna forma con nuestro trabajo:

- Fin de la pobreza
- Hambre cero
- Igualdad de género
- Agua limpia y saneamiento
- Reducción de las desigualdades
- Vida submarina
- Paz, justicia e instituciones sólidas.

En conclusión, nuestro trabajo está relacionado con todos los objetivos que consisten en la disminución de la contaminación. Como se puede observar, ninguno de los objetivos presentados en este último apartado promueve las ciudades sostenibles y/o las mejoras del medioambiente. Estos últimos ODS se encuentran relacionados con la calidad de vida del ser humano.