



Universidade de Vigo

Trabajo Fin de Máster

Construcción de aplicaciones avanzadas de data analytics sobre Power BI

Iria Lago Portela

Máster en Técnicas Estadísticas

Curso 2021-2022

Propuesta de Trabajo Fin de Máster

Título en galego: Construcción de aplicacións avanzadas de data analytics sobre Power BI
Título en español: Construcción de aplicaciones avanzadas de data analytics sobre Power BI
English title: Building advanced data analytics applications on Power BI
Modalidad: Modalidad B
Autora: Iria Lago Portela, Universidad de Santiago de Compostela
Directora: Marta Sestelo Pérez, Universidad de Vigo
Tutor: Rafael P. Martínez Álvarez, Gradient
Breve resumen del trabajo: Este trabajo se centrará en la creación de dashboards y visualizaciones interactivas personalizadas para aplicaciones de business intelligence con la herramienta Power BI. También se hará unha comparativa con otras herramientas alternativas como Tableau, Qlik Sense, Metabase, Redash y Superset.
Recomendaciones:
Otras observaciones:

Doña Marta Sestelo Pérez, Profesora Ayudante Doctora en el Departamento de Estadística e Investigación Operativa de la Universidad de Vigo y don Rafael P. Martínez Álvarez, Responsable Técnico de Big Data Analytics de Gradiant, informan que el Trabajo Fin de Máster titulado

Construcción de aplicaciones avanzadas de data analytics sobre Power BI

fue realizado bajo su dirección por doña Iria Lago Portela para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 01 de Junio de 2022.

La directora:
Doña Marta Sestelo Pérez

El tutor:
Don Rafael P. Martínez Álvarez

La autora:
Doña Iria Lago Portela

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, [Disposición 2978 del BOE núm. 48 de 2022](#)), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas, . . .)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración, . . . sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Índice general

Resumen	IX
1. Business Intelligence	1
1.1. Introducción herramientas BI	2
1.1.1. Power BI	2
1.1.2. Tableau	3
1.1.3. Qlik Sense	5
1.1.4. Redash	5
1.1.5. Metabase	6
1.1.6. Superset	7
1.1.7. Resumen	7
2. Comparación de las herramientas BI	9
2.1. Introducción a los datos	9
2.2. Orígenes de datos	11
2.3. Modelado de datos	14
2.4. <i>Data quality</i> y tratamiento de datos	18
2.5. Creación de visualizaciones	22
2.6. Creación de nueva información	29
2.7. <i>Dashboards</i>	32
2.7.1. Resumen	56
3. Aplicaciones avanzadas en Power BI	57
3.1. Integración de R y Python en Power BI	57
3.2. Creación de visualizaciones interactivas con Power BI	59
3.3. Uso de Preguntas y respuestas de Power BI	60
4. Conclusiones y trabajo futuro	63
4.1. Conclusiones	63
4.2. Trabajo futuro	64
Bibliografía	65

Resumen

Resumen en español

Se denomina *Business Intelligence* (BI) al conjunto de estrategias y herramientas que transforman los datos en información para la toma de decisiones de una organización. En el primer capítulo introduciremos distintas herramientas de BI entre las que se encuentran Power BI, Tableau y Qlik Sense, y otras herramientas *open source* como Metabase, Redash y Superset. En el segundo capítulo nos centraremos en los pasos a seguir para la creación de *dashboards* y visualizaciones interactivas personalizadas, utilizando un conjunto de datos público. Además, veremos cuáles son las ventajas y desventajas de cada una de estas herramientas. En el tercer capítulo investigaremos más en profundidad la herramienta Power BI. Nos centraremos en la construcción de aplicaciones avanzadas de *data analytics* sobre Power BI, donde veremos la integración de los lenguajes R y Python, creación de visualizaciones avanzadas y la funcionalidad de preguntas y respuestas de Power BI. En el cuarto y último capítulo haremos la conclusión de los resultados vistos, y propondremos un trabajo a futuro.

English abstract

Business Intelligence (BI) is the set of strategies and tools that transform data into information for decision-making in a company. In the first chapter we will introduce different BI tools including Power BI, Tableau, and Qlik Sense, and other open source tools such as Metabase, Redash and Superset. In the second chapter we will focus on the steps to create dashboards and personalized interactive visualizations, using an open data set. In addition, we will see what are the advantages and disadvantages of each of these tools. In the third chapter we will take a deeper look at the Power BI tool. We will focus on the construction of advanced data analytics applications using Power BI, where we will see the integration of R and Python languages, the creation of advanced visualizations and the Q&A feature of Power BI. In the fourth and final chapter, we will make a conclusion of the results seen, and we will propose a future work.

Capítulo 1

Business Intelligence

Se denomina *Business Intelligence* (BI), inteligencia de negocio o inteligencia empresarial al conjunto de estrategias y tecnologías que transforman los datos en información para una mejor toma de decisiones de una organización.

Debido a la llegada del Big Data, la automatización y el procesamiento de datos en tiempo real, también ha aumentado la complejidad para extraer conocimiento de interés de toda esta información. Por este motivo las herramientas de Business Intelligence se han ido haciendo más sofisticadas y son hoy en día muy potentes, capaces de analizar y procesar grandes cantidades de datos y de ayudar a las empresas a extraer conclusiones para mejorar sus cifras de negocio.

El proceso comienza con la recogida de los datos, para su posterior preparación y análisis. Es importante que los datos sean de buena calidad. Las distintas fuentes de datos se recopilan, transforman, limpian, cargan y almacenan en un depósito de datos o *data warehouse*. El data warehouse [16] es una plataforma utilizada para almacenar y analizar datos provenientes de distintas fuentes, con el objetivo de transformar los datos brutos en información útil para el usuario. Por último, aquellos datos de interés son explotados por un equipo de BI, que tratará de resumir, analizar y visualizar los datos, haciendo así accesible la información contenida en ellos.

Debido a los rápidos avances, los clientes cada vez requieren de soluciones más eficientes para sus negocios. Es por este motivo que las diferencias entre distintas herramientas de BI son más notables.

Cada año la consultora Gartner emite un informe con la valoración de las herramientas de BI con mayor impacto. En la Figura 1.1 se muestra el Cuadrante de Gartner para Analytics 2021.

Cada Cuadrante Mágico incorpora un gráfico con dos ejes. El eje vertical representa el conocimiento de mercado, mientras que el horizontal indica la habilidad de ejecución. Además, este gráfico proporciona un posicionamiento competitivo de cuatro tipos de proveedores de tecnología:

1. Líderes: se desempeñan bien de acuerdo a la visión actual del mercado y están bien posicionados para el futuro.
2. Visionarios: entienden hacia dónde se dirige el mercado o tienen una visión para cambiar sus reglas, pero su capacidad de ejecución aún es limitada.
3. Jugadores de nicho: se centran con éxito en un segmento pequeño, o están dispersos y no innovan por encima de la media ni superan a los demás.

4. Aspirantes: se desempeñan bien hoy e incluso pueden dominar un gran segmento, pero no demuestran poseer una buena comprensión de hacia dónde se dirige el mercado.



Figura 1.1: Cuadrante Mágico de Gartner para Analytics 2021.

En este trabajo se utilizarán las herramientas Power BI (Microsoft), Tableau y Qlik Sense. Como se puede observar en la Figura 1.1, estas tres opciones se posicionan como líderes del mercado, siendo Power BI la más destacada. Además, también se utilizarán otras herramientas *open source*, como Metabase, Superset y Redash. Cuando se habla de herramientas *open source* o de código abierto se refiere a programas informáticos desarrollados y distribuidos con una licencia que permite a cualquier persona ver el código, modificarlo y, dependiendo del tipo de licencia, utilizarlo sin restricciones. Cabe destacar que las herramientas Metabase, Superset y Redash no son tenidas en cuenta en este gráfico.

1.1. Introducción herramientas BI

En esta sección se hará una breve introducción a las distintas herramientas de *Business Intelligence* que se usarán a lo largo del trabajo. Además se verán sus características principales, precios y tipos de licencias.

1.1.1. Power BI

Microsoft Power BI ¹ es una colección de servicios de software, aplicaciones y conectores que funcionan conjuntamente para convertir orígenes de datos sin relación entre sí en información coherente, interactiva y atractiva visualmente. Esta herramienta es utilizada por empresas de distintos ámbitos: energía, servicios financieros, administración pública, educación, sanidad, etc. Por ejemplo, empresas como Nestle, hp, Bayer y Nokia utilizan Power BI para realizar sus análisis.

¹Página oficial: <https://docs.microsoft.com/es-es/power-bi/>.

Power BI presenta tres tipos de licencia: Gratuita, Pro y Premium. El tipo de licencia que necesita un usuario depende de dónde almacene el contenido, de cómo interactuará con él y de si dicho contenido usa las características Premium. Con respecto a los precios (a fecha de 05/2022), se tiene que el precio de la licencia Pro es de 8,40 € por usuario al mes, mientras que la licencia Premium tiene dos opciones: por usuario y por capacidad. El precio por usuario es de 16,90 € al mes, mientras que por capacidad es de 4.212,30 € al mes. La capacidad es un concepto básico de Power BI que representa un conjunto de recursos que se usan para hospedar y facilitar el contenido de Power BI. Las capacidades pueden ser compartidas o dedicadas; una capacidad compartida es aquella que se comparte con otros clientes de Microsoft, mientras que una dedicada está disponible en exclusiva para un solo cliente.

Además del precio existen otras diferencias entre las distintas licencias. Utilizando una licencia gratuita no se puede colaborar con compañeros en el servicio Power BI, ni los compañeros pueden compartir contenido directamente con este usuario. El límite de tamaño del modelo es de 1 GB y la frecuencia de actualización es de 8 veces al día con un espacio mínimo de 30 minutos en cada actualización. El almacenamiento máximo es de 10 GB por usuario. La versión Pro tiene las mismas características que la versión gratuita, a diferencia de que los informes y tableros pueden ser compartidos con otros usuarios, con la única limitación del tamaño. Para la versión Premium, tenemos que el tamaño del modelo es de 100 GB por usuario y 400 GB por capacidad. La frecuencia de actualización es de 48 veces al día respectivamente. Además, a diferencia de las otras versiones posee herramientas de Inteligencia Artificial avanzada (análisis de texto, detección de imágenes o *machine learning* automatizado). Por último, el almacenamiento máximo es de 100 TB.

En el Cuadro 1.1 se muestra un resumen de las características disponibles dependiendo del tipo de licencia de Power BI:

	Gratuito	Pro por usuario	Premium por usuario	Premium por capacidad
Precio	0 €/mes	8,40 € usuario/mes	16,90 € usuario/mes	4212,30 € capacidad/mes
Tamaño modelo	1 GB	1 GB	100 GB	400 GB
Almacenamiento	10 GB	10 GB	100 TB	100 TB
Frecuencia actualización	8 veces/día	8 veces/día	48 veces/día	48 veces/día
Colaborar con usuarios	No	Si	Si	Si

Cuadro 1.1: Resumen comparación de las licencias en Power BI.

1.1.2. Tableau

Tableau ² es una plataforma de análisis empresarial que permite realizar análisis visuales interactivos de forma rápida, fácil y atractiva. Esta herramienta es utilizada hoy en día por grandes empresas como LinkedIn, Lenovo y Cisco.

Tableau consta de distintas herramientas para el análisis y diseño, de las cuales se dará una breve descripción. La primera de ellas es Tableau Server, se trata de una aplicación web que permite a los usuarios acceder a los informes y consultar información en un entorno seguro sin la necesidad de descargar ningún software. Esta herramienta permite consumir

²Página oficial: <https://www.tableau.com/>.

informes en tabletas y móviles iOS o Android, y en ordenadores. La siguiente herramienta es Tableau Online, una plataforma de análisis alojada en la nube. Desde Tableau Online se pueden compartir visualizaciones y cuadros de mando con la comunidad de Tableau. Es accesible tanto desde navegador web en un equipo de escritorio como en smartphone. Tableau Prep Builder es una herramienta diseñada para preparar datos de forma fácil e intuitiva. Puede usarse para combinar datos, darles forma y limpiarlos para su posterior análisis en Tableau. Tableau Desktop es una herramienta de escritorio para la creación de *dashboards* interactivos, que permite realizar análisis de datos visuales ilimitados y en tiempo real. Además, permite conectarse a múltiples orígenes de datos, como bases de datos o aplicaciones en la nube.

Por último, Tableau Public es un servicio web que puede utilizarse para publicar informes en Internet. Además, también ofrece una versión de escritorio gratuita para la creación y publicación de reportes. Esta versión será la que se utilizará para este trabajo, aunque cabe comentar que posee ciertas limitaciones: permite conectarse con pocos orígenes de datos y los informes creados sólo pueden ser guardados públicamente, por lo que estos informes pueden ser vistos y descargados por cualquier usuario. Además, en el caso de tener algún problema utilizando este producto, no se podrá contactar con el departamento de soporte de Tableau. Otra de las desventajas es que el espacio de almacenamiento está limitado a 10GB por cuenta de usuario, por lo que, superada esta capacidad, no se podrán guardar más informes. Por último, el tamaño de las fuentes de datos está limitado a 10 millones de registros.

A continuación se explicarán los distintos tipos de licencias, Creator, Explorer y Viewer, y las características de cada uno. Los usuarios con licencia Creator crean el contenido para el análisis. Esto incluye el diseño, limpieza y selección de las fuentes de datos, así como la creación de visualizaciones y *dashboards*. Aquéllos usuarios con licencia Explorer pueden acceder a los datos publicados por usuarios Creator y analizarlos. Además, también podrán crear y compartir sus propios dashboards. Por último, los usuarios con licencia Viewer pueden acceder a las visualizaciones y los dashboards publicados, e interactuar con ellos. Esto incluye suscribirse al contenido para recibir actualizaciones y alertas. Sin embargo, no podrán crear sus propios informes.

Por último, existen tres opciones de compra dependiendo de las funcionalidades que incluya. En primer lugar se tiene la opción Tableau Creator, con un precio de USD 70 (61,68 €), que incluye las herramientas Tableau Desktop, Tableau Prep Builder y la licencia Creator. En segundo lugar la opción Tableau Explorer, con un precio de USD 42 (37,01 €), que incluye la licencia Explorer. Por último la opción Tableau Viewer con un precio de USD 15 (13,22 €), que incluye la licencia Viewer. Nótese que estos precios se corresponden a fecha de 05/2022 y pueden cambiar en el futuro.

En el Cuadro 1.2 se muestra un resumen de las características incluidas en cada licencia de Tableau:

	Creator	Explorer	Viewer
Precio	61,68 €/mes	37,01 €/mes	13,22 €/mes
Herramientas incluidas	Tableau Desktop, Tableau Prep Builder		

Cuadro 1.2: Resumen comparación de las licencias en Tableau.

1.1.3. Qlik Sense

Qlik Sense ³ es una aplicación avanzada de *Business Intelligence* que permite realizar visualizaciones de datos flexibles e interactivas. Entre las empresas que utilizan Qlik Sense se encuentran Johnson & Johnson y Danone.

Esta herramienta dispone de dos opciones de compra: Qlik Sense Business y Qlik Sense Enterprise SaaS. La primera de ellas tiene, a fecha de 05/2022, un precio de USD 30 (26,43 €), mientras que para la segunda es necesario contactar con ventas para obtener un presupuesto. A diferencia de otras herramientas no tiene versión gratuita, pero puede utilizarse una prueba durante 30 días, que es la que se utilizará para este trabajo.

Además del precio, existen diversas diferencias entre ambas opciones de compra. En primer lugar, el límite de tamaño de aplicación estándar (en memoria) es para Qlik Sense Business de 1,25GB y para Qlik Sense Enterprise SaaS de 5GB ampliables. Además, Qlik Sense Business presenta limitaciones en la carga de datos: por día hasta un máximo de 50 cargas y hasta 3 cargas de datos simultáneas. Por último, Qlik Sense Business puede almacenar apps y datos por un total de 250 GB o menos.

En el Cuadro 1.3 se muestra un resumen de las características que presenta cada una de las licencias de Qlik Sense:

	Qlik Sense Business	Qlik Sense Enterprise SaaS
Precio	26,43 €/mes	A consultar
Memoria app	1,25 GB	5 GB ampliables
Almacenamiento	250 GB	No especificado
Cargas por día	50 cargas	No especificado
Cargas simultáneas	3 cargas	No especificado

Cuadro 1.3: Resumen comparación de las licencias en Qlik Sense.

1.1.4. Redash

La siguiente herramienta de *Business Intelligence* de la que se hablará será Redash ⁴. Se trata de una herramienta *open source* y gratuita con una comunidad de más de 350 contribuidores. A diferencia de las anteriores, esta herramienta está diseñada para usuarios nativos de SQL, por lo que se puede acceder a los datos a través de consultas. Además, también proporciona una amplia variedad de visualizaciones.

Para utilizar la versión de código abierto de Redash es necesario que el usuario haga el propio despliegue de la herramienta, y además deberá cumplir varios requisitos. En primer lugar será necesario utilizar un servidor Linux (o un contenedor VM/Docker [14]). Además, se recomienda también utilizar los navegadores Chrome o Firefox. Para implementaciones básicas se recomienda un mínimo de 4 GB de RAM y una asignación de CPU razonable, aunque a medida que se utilice puede ser necesaria más memoria RAM y potencia de CPU.

³Página oficial: <https://www.qlik.com/es-es/>.

⁴Página oficial: <https://redash.io/>.

Este también será el caso para empresas donde se quiere admitir una gran cantidad de trabajadores. Para crear una instancia se ofrecen diversas opciones: AWS, DigitalOcean, Google Compute Engine Image y Docker. Una vez creada la instancia con la imagen o el *script*, Redash estará disponible usando la IP del servidor o el nombre DNS asignado.

1.1.5. Metabase

Metabase ⁵ es otra herramienta de *Business Intelligence* para la visualización de datos. Se trata de una herramienta *open source*, aunque posee varias opciones de compra cuyos precios y características dependerán de si se quiere utilizar *on-premises* o como servicio en la nube. Dentro del primer grupo existe la opción gratuita, que permite gráficos y *dashboards* ilimitados, conectarse a más de 20 tipos de bases de datos, usar más de 15 visualizaciones y programar actualizaciones. Sin embargo, no tiene ningún tipo de soporte y el despliegue del programa lo debe hacer uno mismo. A continuación está la versión Pro, que a parte de las características anteriores permite inicio de sesión único, más permisos a nivel de fila, personalización de logotipos y colores, y otras opciones avanzadas. Además, en este caso el soporte se realiza por email cada tres días y el precio es de 500\$/mes (475,38 €/mes) o si se incluyen 10 usuarios entonces 10\$ (9,51 €) por usuario al mes. Por último se encuentra la opción Enterprise, que incluye todo lo anterior y además tiene soporte prioritario y facturación anual. El precio variará en función del tipo de usuario. Recordemos que los precios son a fecha de 05/2022 y pueden variar en el futuro. Por otra parte, en la opción Cloud existen también tres opciones: Starter, Pro y Enterprise, cuyas características son similares a las ya vistas, respectivamente. En la opción Starter se mantienen las características de la versión gratuita, salvo que el despliegue se realiza en la nube, el soporte es cada 3 días vía email y el precio es de 85\$/mes (80,82 €/mes), o bien si se incluyen 5 usuarios, 5\$ (4,75 €) por usuario al mes. Además, la nube está completamente administrada, las actualizaciones y copias de seguridad automatizadas y permite la migración desde código abierto. La opción Pro en cloud tiene las mismas opciones que la versión *on-premises*, salvo que el despliegue se realiza en la nube. Lo mismo ocurre con la versión Enterprise. Para este proyecto se utilizará la versión gratuita *on-premises*.

En el Cuadro 1.4 se muestra un resumen de las características incluidas en cada opción de compra de Metabase *on-premises*:

	Gratuita	Pro	Enterprise
Precio	0 €/mes	475,38 €/mes	Variable
Soporte	Foro	Email cada 3 días	Soporte prioritario
Despliegue	Auto-hospedado	Auto-hospedado	Auto-hospedado

Cuadro 1.4: Resumen comparación de las opciones de compra de Metabase *on-premises*.

En el Cuadro 1.5 se muestra un resumen de las características incluidas en cada opción de compra de Metabase en la nube:

⁵Página oficial: <https://www.metabase.com/>.

	Starter	Pro	Enterprise
Precio	85 €/mes	475,38 €/mes	Variable
Soporte	Email cada 3 días	Email cada 3 días	Soporte prioritario
Despliegue	Nube Metabase	Nube Metabase	Nube Metabase

Cuadro 1.5: Resumen comparación de las opciones de compra de Metabase en la nube.

1.1.6. Superset

Por último, se considerará la herramienta de *Business Intelligence* Apache Superset ⁶. Se trata de un programa de código abierto para la exploración y visualización de datos. Es utilizado por una gran variedad de empresas de distintos ámbitos: servicios financieros, tecnología, entretenimiento, educación, salud, etc. Destacamos su utilización en empresas como Airbnb, Zalando, Netflix, Udemy y Twitter.

El despliegue de la herramienta Superset debe hacerla el propio usuario. Para instalarla existen dos opciones, desde cero o mediante Docker Compose. Esta última opción es la recomendada por ser más rápida y sencilla, y es la que se ha utilizado en este trabajo. Esta herramienta está disponible para Mac OS X y Linux, pero no lo está para Windows. Para su instalación en MAC se recomienda reservar una memoria de 6 GB. Para el caso de Windows se podría instalar localmente un escritorio Ubuntu utilizando una máquina virtual como VirtualBox. En este caso se recomienda asignar al menos 8 GB de RAM para la máquina virtual y aprovisionar un disco duro de al menos 40 GB. Además, tanto en el caso de Linux como en el de la máquina virtual de Ubuntu será necesario instalar Docker. Una vez creada la instancia de Superset, el servicio estará disponible en la url <http://localhost:8088>. Nótese que muchos navegadores utilizan por defecto ‘https’, por lo que será necesario un navegador que utilice ‘http’.

1.1.7. Resumen

En las subsecciones anteriores hemos visto que existe una gran variedad de opciones de compra dentro del mundo del *Business Intelligence*. Cada una de las herramientas presenta unas características que la hacen única y por ello difícil de comparar con el resto. En el Cuadro 1.6 se muestra un resumen de las opciones que ofrece cada una de las herramientas, considerando únicamente su versión gratuita o de prueba, es decir, la que se ha utilizado en este trabajo.

En primer lugar, hemos visto que todas las herramientas presentan una opción gratuita, a excepción de Qlik Sense, que ofrece una versión de prueba de 30 días. Consideramos que es importante que se ofrezcan versiones gratuitas, sobre todo para que cualquier usuario pueda introducirse en el mundo del BI. Otra característica de importancia es la disponibilidad en Windows, puesto que es uno de los sistemas operativos más utilizados, y puede resultar complicado el cambio a otro sistema operativo como GNU/Linux. A continuación, la posibilidad de guardar los archivos en local, donde Tableau es la única herramienta que no lo permite. Por último, el espacio de almacenamiento, donde Power BI y Tableau ofrecen 10 GB/usuario, mientras que Qlik Sense ofrece 250 GB. Esta gran diferencia se debe a

⁶Página oficial: <https://superset.apache.org/>.

que en Qlik Sense se está utilizando una versión de prueba. Para las herramientas Redash, Metabase y Superset el almacenamiento depende de lo que cada usuario decida y pueda ofrecer a la hora de instalar el programa.

	Power BI	Tableau	Qlik Sense	Redash	Metabase	Superset
Opción gratuita	Si	Si	No	Si	Si	Si
Disponible Windows	Si	Si	Si	No	Si	No
Informes privados	Si	No	Si	Si	Si	Si
Almacenamiento	10 GB/usuario	10 GB/usuario	250 GB			

Cuadro 1.6: Resumen comparación de las herramientas de BI.

Capítulo 2

Comparación de las herramientas BI

En este capítulo se realizará una comparativa de las herramientas de BI introducidas en el primer capítulo. Esta comparación se realizará sobre un caso práctico con un conjunto de datos públicos. En primer lugar se hará una introducción a este conjunto de datos. A continuación se verá el tipo de conexiones de datos que permite cada herramienta. En la tercera sección se explicará qué es el modelado de datos y cuál es el utilizado en este conjunto de datos. En la cuarta sección se verá qué posibilidades ofrece cada herramienta para el tratamiento de los datos. En la siguiente sección se explicará cómo se crean las visualizaciones en cada programa. A continuación se verá cómo crear nueva información a partir de los datos ya existentes. Por último, se realizará un dashboard con Power BI que se intentará replicar para el resto de herramientas de BI.

2.1. Introducción a los datos

Para la comparación de las herramientas explicadas en el primer capítulo se realizará un ejemplo práctico. Para poder sacar el mayor partido a las herramientas de visualización es necesario un conjunto de datos con un número razonable de tablas y que posea variables tanto continuas como discretas.

El conjunto de datos utilizado proviene de la plataforma Kaggle ¹ y contiene información acerca de los accidentes de tráfico ocurridos en el año 2019 en Francia continental, departamentos de Ultramar (Guadalupe, Guyana, Martinica, Isla de la Reunión y Mayotte) y en los demás territorios de ultramar (Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Martin, Wallis-et-Futuna, French Polynesia y New Caledonia). Este conjunto de datos posee 5 tablas principales:

1. **Fact_Accidentes:** Contiene la identificación del número de accidente y sus características, hora y minuto del accidente, dirección, latitud, longitud y fecha del accidente.

¹Se puede acceder al conjunto de datos en el siguiente enlace: <https://www.kaggle.com/datasets/dorianvoydie/2019-database-of-road-traffic-injuries>.

Id_Accidente	Hrmn	Id_Iluminacion	Id_CondAtm	Id_Colision	Direccion	Latitud	Longitud	Fecha
201900000001	0.0625	4	1	2	AUTOROUTE A3	488962100	2470120	30/11/2019
201900000002	0.118056	3	1	6	AUTOROUTE A1	489307000	23688000	30/11/2019

2. **Dim Lugar:** Contiene el identificador del accidente y las características del lugar donde se produjo, tipo de carretera y condiciones de la superficie.

Id_Accidente	Id_TipoCarretera	Id_CondSuperf
201900000001	1	1
201900000002	1	1

3. **Dim Vehículo:** Contiene el id y las condiciones del choque, tipo de obstáculo, lugar del choque, número de ocupantes.

Id_Accidente	Id_Obs	Id_ObsMovil	Id_Choque	Id_Motor	NumOcupantes
201900000001	0	2	5	1	0
201900000002	1	0	3	1	0

4. **Dim Usuario:** Contiene el id y las características del usuario involucrado, tipo de usuario (conductor, pasajero o peatón), sexo, año de licencia, razón del desplazamiento.

Id_Accidente	Id_CatUsuario	Id_Gravedad	Id_Sexo	Año Licencia	Id_Trayecto	Id_Acompañante
201900000001	2	4	2	2002	0	0
201900000001	1	4	2	1993	5	0

5. **Dim Calendario:** Contiene la fecha del accidente, año, mes, semana y día.

Fecha	Año	Mes	Semana	Dia
01/01/2019	2019	1	1	1
02/01/2019	2019	1	1	2

Además, se usarán otras tablas complementarias para aportar más información.

2.2. Orígenes de datos

El primer paso para trabajar con herramientas de *Business Intelligence* es la carga de los datos. El origen de datos es, como su nombre indica, de dónde provienen los datos.

Power BI destaca por la multitud de orígenes de datos con los que puede conectarse. Estos se dividen en Archivos, Base de datos, Power Platform, Azure, Servicios en línea y Otros. Dentro de la categoría ‘Archivos’ encontramos las opciones de archivos Excel, CSV, XML, JSON, PDF y otros. En la categoría ‘Bases de Datos’ destacamos las bases de datos SQL Server, Oracle Database, MySQL y Snowflake. Power Platform es un conjunto de herramientas de Microsoft integradas dentro de Office 365. En este apartado podemos conectarnos a flujos y conjuntos de datos de Power BI, Flujos de datos, Common Data Service y Dataverse. Azure es un conjunto de servicios en la nube diseñados por Microsoft que está pensado para almacenar información y crear e implementar aplicaciones en la nube. Se trata de un servicio de pago. Power BI permite conectarse a Azure SQL Database, Azure Synapse Analytics, Azure Databricks, etc. Los servicios en línea son los servicios disponibles a través de Internet. En este apartado destacamos Google Analytics, GitHub y el Navegador de ventas de LinkedIn. Por último, en ‘Otras’ podemos encontrar Spark, Hadoop, Scripts de R [19] y Python ², entre otros.

En la Figura 2.1 se muestra la pantalla de carga de los datos en Power BI.

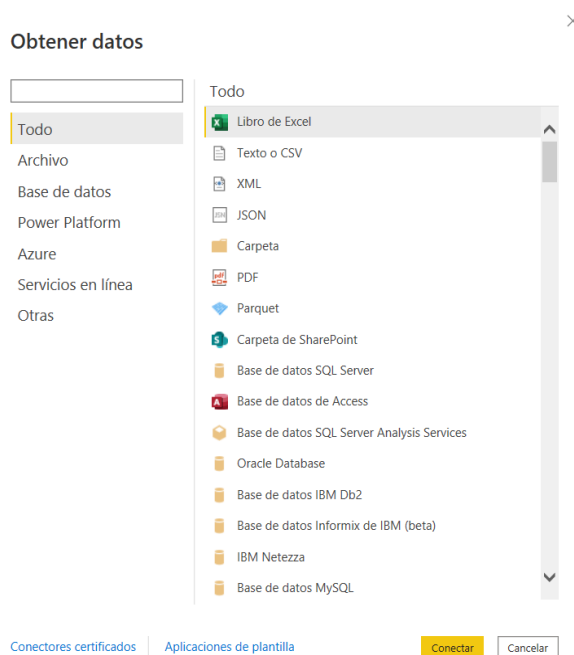


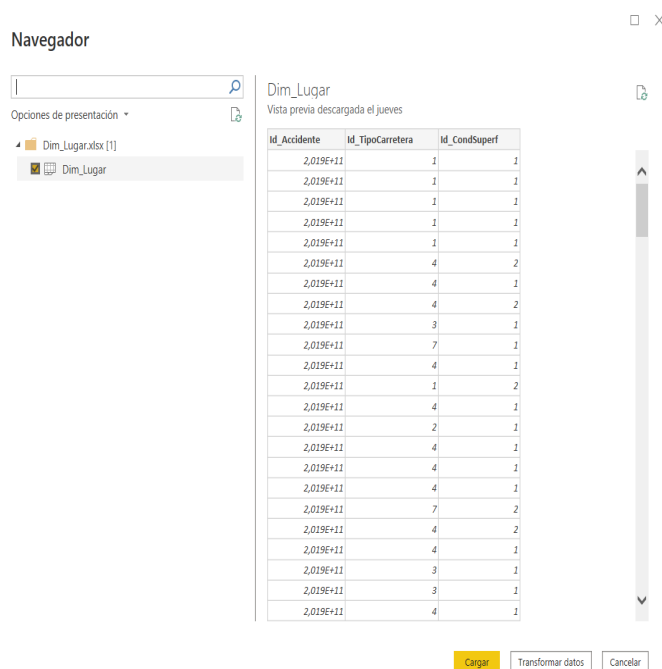
Figura 2.1: Orígenes de datos de Power BI.

Como hemos visto, en la categoría de Power Platform se encuentran las opciones flujos y conjuntos de datos de Power BI. Cada vez que se publica un informe de Power BI éste se guarda directamente en Power BI Services, junto con el conjunto de datos que utiliza. Si se quisiera utilizar este conjunto de datos para un nuevo informe podría cargarse a través del conjunto de datos de Power BI. Por otra parte, un diagrama de flujo de datos traza el flujo de la información para cualquier proceso o sistema. Este diagrama muestra cómo se manejan los datos de una organización, incluyendo quién tiene acceso a los datos, dónde

²Página oficial: <https://www.python.org/>.

se encuentran almacenados y qué se hace con ellos. Luego, cuando se utiliza Power BI en el contexto empresarial, puede resultar más práctico conectarse a un flujo de datos ya creado, o bien crear uno a través de Power BI. Sin embargo estas características sólo están disponibles para los usuarios Pro y Premium.

Para acceder a los datos de Accidentes de tráfico en Francia en el año 2019 se han utilizado dos formas, a través de archivos Excel y a través de la base de datos MySQL. Para cargar un archivo Excel se tendrán que seguir los pasos Obtener Datos > Libro de Excel y seleccionar el archivo. En la Figura 2.2 se observa la previsualización de la tabla Dim_Lugar.



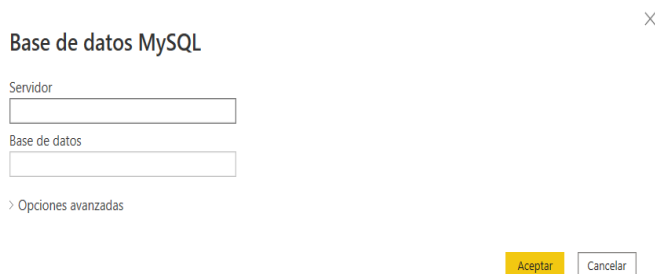
Dim_Lugar
Vista previa descargada el jueves

Id_Accidente	Id_TipoCarretera	Id_CondSuperf
2,019E+11	1	1
2,019E+11	1	1
2,019E+11	1	1
2,019E+11	1	1
2,019E+11	1	1
2,019E+11	4	2
2,019E+11	4	1
2,019E+11	4	2
2,019E+11	3	1
2,019E+11	7	1
2,019E+11	4	1
2,019E+11	1	2
2,019E+11	4	1
2,019E+11	2	1
2,019E+11	4	1
2,019E+11	4	1
2,019E+11	4	1
2,019E+11	7	2
2,019E+11	4	2
2,019E+11	4	1
2,019E+11	3	1
2,019E+11	3	1
2,019E+11	4	1

Cargar Transformar datos Cancelar

Figura 2.2: Carga de datos a partir de un archivo Excel en Power BI.

Por otra parte, para acceder a los datos a través de la base de datos MySQL se deberán seguir los pasos Obtener Datos > Base de Datos MySQL. En este caso se deberá conectar al servidor donde se encuentra la base de datos. En la Figura 2.3 se muestra el acceso a la base de datos.



Base de datos MySQL

Servidor

Base de datos

> Opciones avanzadas

Aceptar Cancelar

Figura 2.3: Carga de datos a partir de una base de datos MySQL en Power BI.

Otra característica de importancia a la hora de cargar los datos es el tipo de conexión que se establece con ellos. En Power BI se diferencian tres tipos de conexión: 'Importación',

Direct Query y *Live Connection*. En el método de importación se coge una copia de los datos con los que se ha conectado y se guarda con el archivo. Las ventajas de este método son que se puede subir gran cantidad de información, combinar con diferentes orígenes de información y se admiten todas las funcionalidades del lenguaje DAX (que explicaremos posteriormente) y transformaciones. Entre sus desventajas está la limitación del tamaño de archivo de Power BI y que para actualizar los datos es necesario volver a importar la información. En el método *Direct Query* se genera una conexión al origen de datos, y es en el momento de abrir el informe o de interactuar con él cuando se conecta a la base de datos y devuelven los datos. La principal ventaja es que los datos siempre están actualizados, sin embargo se trata de un tipo de conexión mucho más lenta, pues tiene que acceder a la base de datos y extraerlos cada vez que se conecta. Además, las funcionalidades DAX y transformaciones están limitadas y no se pueden combinar distintos orígenes de datos. Por último, la conexión *Live Connection* permite conectarse a datos que se están actualizando continuamente. Entre sus ventajas está que permite la conexión con fuentes de datos de gran tamaño, sin embargo, tan sólo permite visualizaciones y DAX muy limitado, no se pueden realizar transformaciones. Por lo tanto, antes de conectarse a algún origen de datos, es necesario decidir qué tipo de conexión es adecuada. Para este ejemplo se ha utilizado el tipo de conexión Importación.

A continuación explicaremos qué orígenes de datos se encuentran disponibles en Qlik Sense. Este programa da la posibilidad de conectarse a Archivos, Bases de datos, Servicios en la nube y Aplicaciones empresariales. Como se muestra en la Figura 2.4 Qlik permite subir un archivo de datos o bien establecer una nueva conexión de datos. En la opción de conectar con los datos se permite la conexión a bases de datos de Amazon, Azure SQL Database, Google BigQuery y MySQL, entre muchos otros, así como la conexión a plataformas como Twitter y YouTube Analytics.

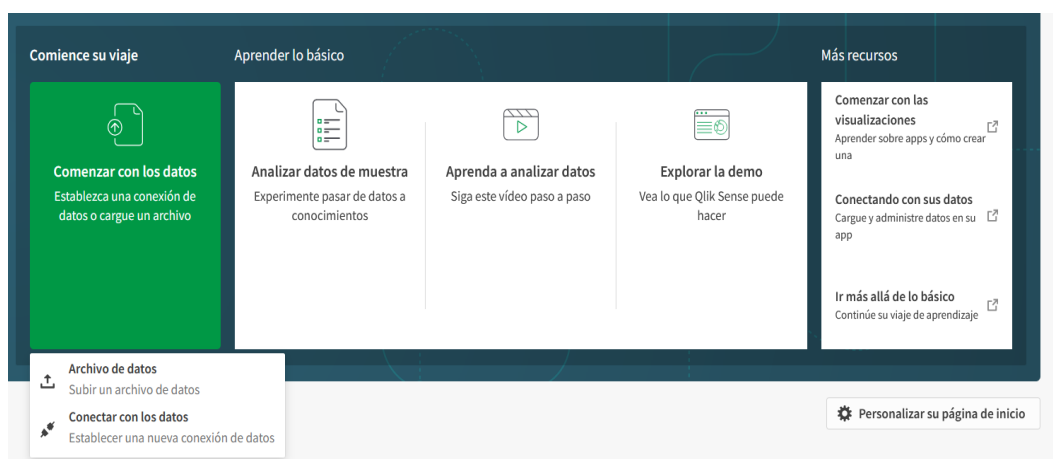


Figura 2.4: Orígenes de datos de Qlik Sense.

Además, se tiene la opción de subir los archivos Excel o bien conectarse a la base de datos creada en MySQL, tal y como se hizo con Power BI.

En Tableau se tienen dos opciones de conexión a orígenes de datos: archivos y servidores. En la sección de archivos se encuentran archivos Excel, de texto, JSON, PDF, archivos espaciales, estadísticos y de Microsoft Access. Mientras que en servidores permite conectarse a datos web, Google Drive, Hojas de cálculo de Google y OData. En la Figura 2.5 se observan las distintas opciones de conexión.

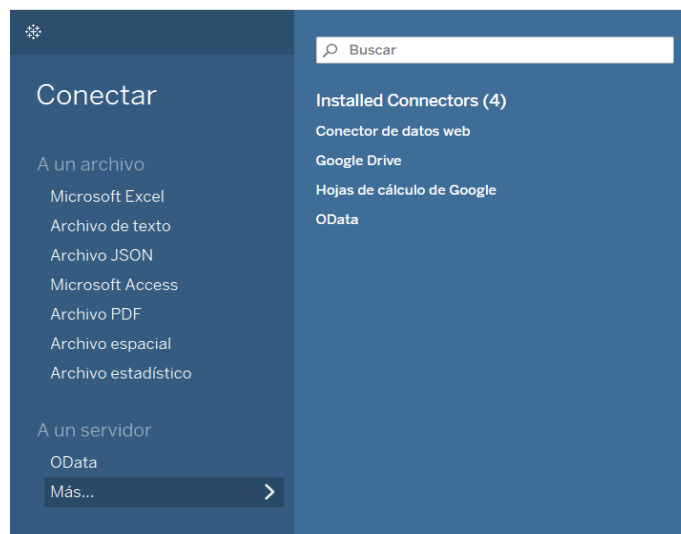


Figura 2.5: Orígenes de datos de Tableau.

Redash ofrece distintas conexiones a bases de datos, entre las que destacan Amazon Redshift, Databricks, Google BigQuery, MySQL, Oracle y Snowflake. Además, también permite la integración de otras herramientas como Google Analytics, Google Spreadsheets, JSON y Python. Sin embargo, no tiene disponible la conexión con archivos, por lo que no se pueden cargar los datos a través de archivos Excel. Por lo que en este caso se utilizará la base de datos creada en MySQL.

Lo mismo ocurre con Superset, que tan sólo permite conexión a bases de datos, entre las que se encuentran Amazon Redshift, Big Query, MySQL, Oracle, Snowflake y SQL Server.

Por último, y al igual que en los dos casos anteriores, Metabase sólo permite conexiones a bases de datos. Destacan la conexión a las bases de datos MySQL, Amazon Redshift, SQL Server, Oracle, SQLite, BigQuery, Snowflake y Google Analytics.

Como resumen, las herramientas Power BI y Qlik Sense permiten la conexión con una gran variedad de orígenes de datos. En el caso de Tableau, con la versión gratuita permite conectarse a pocos orígenes de datos. Por último, las herramientas Metabase, Superset y Redash permiten la conexión con una gran variedad de bases de datos, pero no tienen la opción de conectarse a otro tipo de archivos.

2.3. Modelado de datos

El modelo asociativo de una base de datos es el conjunto de relaciones existentes entre las distintas tablas. Cuando se modelan los datos se crea una estructura lógica simple que facilita la exploración y el análisis de los datos. Además, permite al *software* navegar y obtener resultados de forma rápida y eficiente. Las tablas de dimensión describen entidades empresariales. Estas pueden incluir productos, personas, lugares y conceptos, y responden a las preguntas ‘quién’, ‘qué’, ‘dónde’, ‘cuándo’, ‘cómo’ y ‘por qué’ asociado con el evento. Las tablas de dimensión contienen una columna (o columnas) clave que actúa como identificador único y columnas descriptivas. Las tablas de hechos contienen columnas clave de dimensiones relacionadas con las tablas de dimensiones y columnas de medidas numéricas. Normalmente, las tablas de dimensiones contienen un número relativamente pequeño de filas. Por el contrario, las tablas de hechos pueden contener un gran número de filas y seguir creciendo con el tiempo. En el caso del conjunto de datos utilizado, la

tabla de hechos se corresponde con la tabla `Fact_Accidentes`, pues contiene los registros de todos los accidentes y contiene campos clave que se unen con el resto de tablas. El resto de tablas del conjunto de datos son dimensionales, pues contienen información acerca de los accidentes.

Existen diversas formas de modelar los datos, aquí se explicarán dos de ellas: esquema de estrella y snowflake. El esquema de estrella consta de una tabla central de hechos y varias tablas de dimensión que se disponen alrededor. La característica principal de este modelado es que no existe relación alguna entre las tablas de dimensiones, tan sólo se relacionan con la tabla de hechos. El esquema snowflake es una variación del anterior. En este caso la tabla de hechos deja de ser la única relacionada con las tablas de dimensiones. Existen otras tablas que se relacionan con las dimensiones y que no tienen relación directa con la tabla de hechos.

En la Figura 2.6 se muestra a la izquierda un ejemplo ³ de modelado con esquema de estrella y a la derecha un ejemplo de modelado con esquema de copo de nieve.

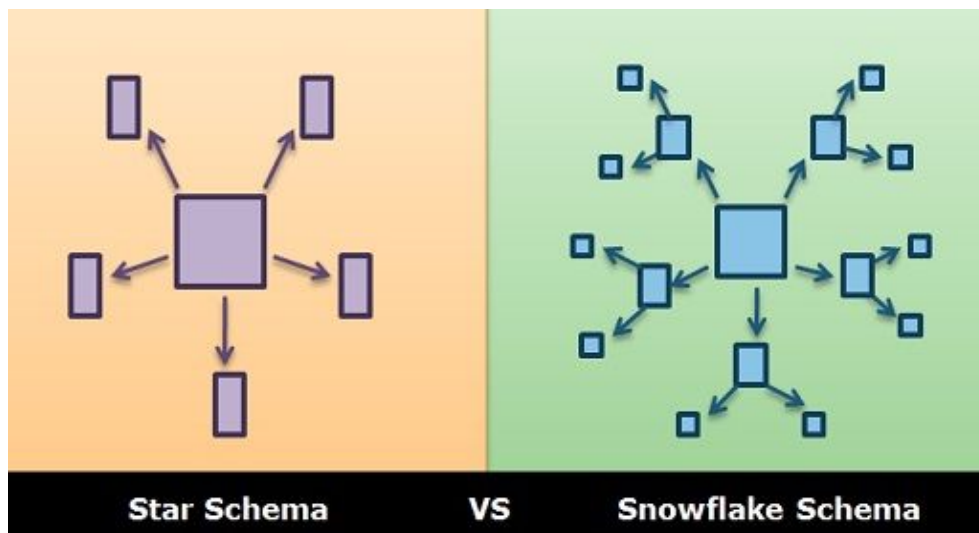


Figura 2.6: Modelados con esquema de estrella y snowflake.

En este caso se ha utilizado el modelado de snowflake, puesto que las tablas adicionales son tablas dimensionales que no se relacionan directamente con la tabla de hechos.

A continuación se mostrará cómo se crea el modelado de datos en las distintas herramientas de Business Intelligence.

En Power BI generalmente las asociaciones en las tablas se crean de forma automática al cargar los datos. Sin embargo, estas asociaciones pueden ser modificadas manualmente, de forma que se puede eliminar una relación o bien crear una nueva. Para establecer una asociación entre ambas tablas basta seleccionar el campo por el cual se unen y arrastrarlo hacia el campo de la otra tabla. Además se puede establecer el tipo de relación (uno a uno, uno a varios, varios a uno y varios a varios) y la dirección del filtro (única o ambas). Estas propiedades conviene tenerlas claras para obtener los resultados deseados.

En la Figura 2.7 se observa el modelo dimensional realizado en Power BI. La tabla inferior representa la tabla de hechos, mientras que el resto de tablas son dimensionales.

³Fuente: <https://techdifferences.com/difference-between-star-and-snowflake-schema.html>.

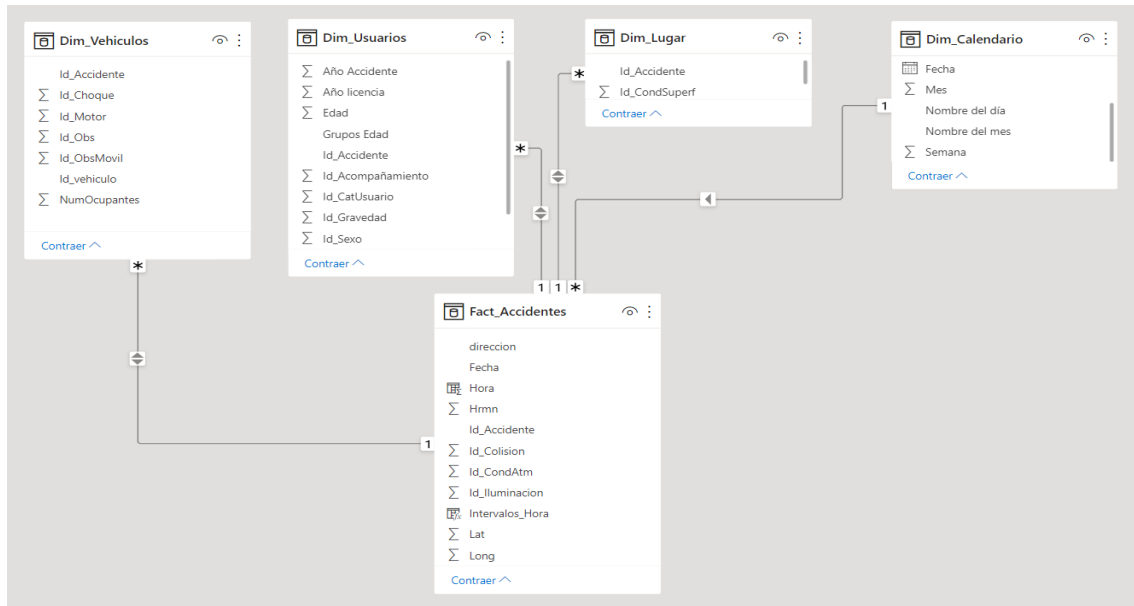


Figura 2.7: Modelo dimensional con Power BI.

En Qlik Sense al cargar los datos se muestran automáticamente las tablas en forma de burbujas en el ‘Data Manager’. A continuación hay que establecer las relaciones entre las tablas. El propio sistema muestra las asociaciones recomendadas, por lo que se podrían aplicar directamente. Otra forma sería seleccionar la tabla y arrastrarla hacia la tabla que se quiere asociar. En este caso, al seleccionar la tabla, el resto de tablas se mostrarán en color, dependiendo de lo segura que sea la relación. Esto es, si la tabla se muestra de color verde será una asociación muy segura, naranja medianamente segura y roja no recomendada.

En la Figura 2.8 se muestra el modelo dimensional en forma de burbujas en Qlik Sense.

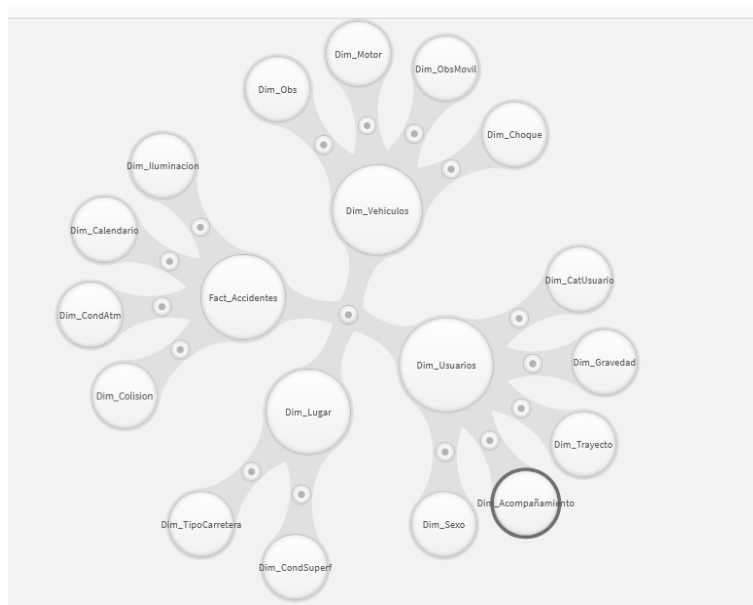


Figura 2.8: Modelo dimensional con Qlik Sense.

Al realizar cada asociación se mostrarán las tablas asociadas y el campo que une a ambas

tablas. Por último, si se quiere romper una asociación entre tablas basta tirar de una tabla para que el enlace se rompa.

Si en vez de ver estas asociaciones en forma de burbujas fuera de interés ver el modelo con las tablas y campos, se puede ir al visor del modelo de datos. La representación del modelo con las tablas es muy similar a la ya vista en la herramienta de Power BI. Visualmente la representación con burbujas es mucho más simple y fácil de interpretar, pero ofrece menos información que el modelado con tablas.

En el caso de Tableau el modelo dimensional no se crea directamente, sino que hay que hacer manualmente las asociaciones entre tablas. A diferencia de otras herramientas, el modelo asociativo de Tableau se divide en dos capas:

1. En primer lugar existe una capa lógica, en la que se establecen las relaciones entre las tablas. Esta capa se corresponde con el modelado de datos en las otras herramientas.
2. En segundo lugar existe una capa física, donde las tablas se combinan utilizando *joins* y uniones. Cada tabla lógica contiene al menos una tabla física. Para acceder a ellas hay que hacer doble click en la tabla lógica.

En la Figura 2.9 se puede observar cómo es el modelo dimensional final con Tableau.

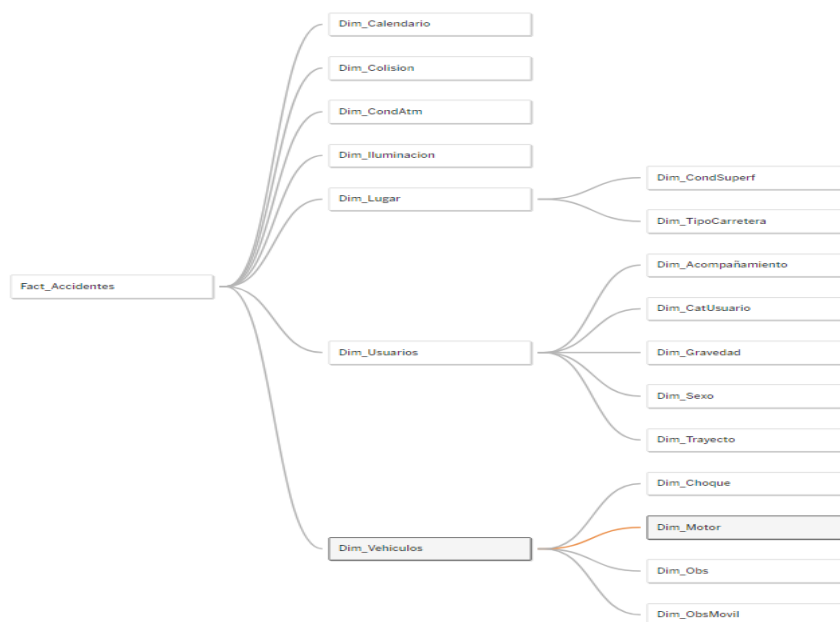


Figura 2.9: Modelo dimensional con Tableau.

Visualmente, los modelos dimensionales con tablas son menos intuitivos pero aportan más información, mientras que el gráfico de burbujas de Qlik Sense y el diagrama de Tableau son más sencillos y fáciles de interpretar.

Por otra parte, las herramientas *open source* Redash, Metabase y Superset no tienen la opción de crear un modelo dimensional. Esto se debe principalmente a que en estos programas los datos se cargan desde sistemas de gestión de bases de datos en las cuales se pueden indicar las relaciones entre las tablas.

Como conclusión, las herramientas Power BI, Qlik Sense y Tableau presentan la opción del modelado de datos, pudiendo establecer el usuario las relaciones existentes en cada

tabla. Además, cabe destacar que tanto Power BI como Tableau permiten el modelado de datos de forma gratuita, mientras que en Qlik Sense es necesario el uso de una licencia. Por otra parte, las herramientas Redash, Metabase y Superset no poseen ninguna opción para el modelado de datos.

2.4. *Data quality* y tratamiento de datos

Se denomina *data quality* al proceso de planificación, implementación y control de actividades que aplican técnicas de gestión de calidad a los datos para garantizar que sean aptos para su uso. La calidad de los datos es de vital importancia dentro del *Business Intelligence*, puesto que unos datos inadecuados o incorrectos pueden dar lugar a una mala toma de decisiones por parte de la empresa, lo que puede suponer un impacto económico.

Para que los datos sean de calidad es necesario que verifiquen ciertas propiedades o ‘dimensiones’, éstas son:

1. **Compleitud:** Mide el grado en el que el dato tiene el valor esperado y cumple con los requerimientos marcados.
2. **Validez:** Mide si un valor se ajusta a un estándar preestablecido con respecto al formato, tipo de datos, valores posibles o rangos especificados.
3. **Pertinencia temporal:** Esta dimensión mide el grado en el que los datos están disponibles cuando se requieren. Esta medida cada vez es más importante debido a la necesidad de datos en tiempo real.
4. **Unicidad:** La unicidad mide la cantidad de datos duplicados dentro de un conjunto de datos, ya sea en una columna en particular o en todos los campos.
5. **Exactitud:** Mide el grado en el que los datos representan correctamente el objeto del mundo real o evento que describen.
6. **Consistencia:** Esta dimensión mide si los datos están libres de contradicción y tienen coherencia lógica, de formato o temporal.

A continuación se explicarán las técnicas de *data quality* disponibles dentro de cada herramienta.

En Power BI existen varias opciones para limpiar, transformar y entender los datos. En primer lugar se explicarán las herramientas para entender los datos. Dentro del Editor de Power Query existen tres herramientas para la visualización y resumen de los datos: ‘Calidad de columnas’, ‘Distribución de columnas’ y ‘Perfil de columna’. La herramienta Calidad de columnas divide los datos en cinco categorías según la calidad del dato: válidos, errores, vacíos, desconocidos y errores inesperados. Cada columna tendrá asignado el porcentaje de valores de cada tipo. La herramienta Distribución de columnas provee a cada columna de un gráfico de barras donde se muestra la frecuencia y distribución de sus valores. Los datos representados se muestran ordenados de forma descendiente desde el valor con mayor frecuencia. Además, también se muestra el porcentaje de valores distintos y únicos. Por último, la herramienta Perfil de columna aporta información más detallada acerca de los datos de cada columna. En primer lugar, se muestran las estadísticas de la columna, esto es, el recuento de valores, los valores mínimo y máximo, el promedio, la desviación estándar y el número de errores, ceros, vacíos, distintos y únicos. Además, se muestra de nuevo el gráfico de barras con la distribución de los valores.

Con respecto a las herramientas para transformar los datos, en Power BI se pueden realizar filtros, agrupaciones y copiar conjuntos de datos. Además, se pueden editar los valores de los datos, poniéndolos en mayúsculas, minúsculas y eliminar espacios.

Por último, para limpiar los datos existen varias opciones. Entre ellas se encuentran eliminar errores, vacíos y duplicados, o bien reemplazarlos por un nuevo valor. Otra opción para la limpieza de los datos sería implementar técnicas de *data quality* a través de un script de R o Python, pero hablaremos posteriormente de esta posibilidad.

Todas estas opciones de preprocesado de los datos, se traducen en un script con lenguaje M. Se trata de un lenguaje funcional que opera tras Power Query y cuya función principal es el preprocesado de los datos antes de cargarlos en el modelo de Power BI. Para acceder al script con lenguaje M basta abrir el Editor de Power Query, ir a inicio y abrir el Editor avanzado. En la Figura 2.10 se muestran los cambios aplicados para la tabla `Fact.Accidentes`.

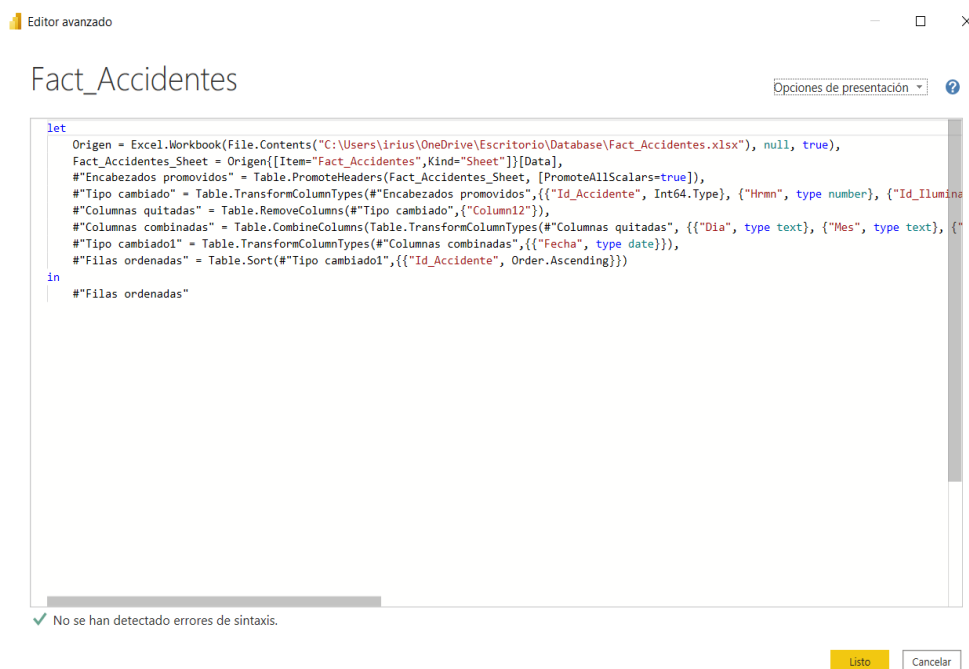


Figura 2.10: Editor avanzado Power Query.

Qlik Sense tiene una amplia funcionalidad en este aspecto. Entre sus funciones se puede filtrar, agrupar y reemplazar valores. Además, cada columna de la tabla presenta un panel de resumen donde se obtiene la siguiente información:

1. El número de valores distintos en el campo.
2. El número total de valores del campo.
3. Una vista previa de la distribución de los valores distintos. En el caso de que todos los valores sean distintos no se mostrará ningún gráfico de barras.
4. Si la columna es una medida o es un campo temporal se mostrará un rango de valores. Para un campo de medida, el rango de valores es un gráfico que muestra los valores mínimo, mediana, media y máximo. En el caso del campo temporal, el rango de valores es el período de tiempo que abarca el campo.

5. El número de valores nulos en los datos. Esta visualización solo se muestra si existen valores nulos.

Por ejemplo, para el caso de la columna `Intervalos Hora`, el resumen se muestra en la Figura 2.11.

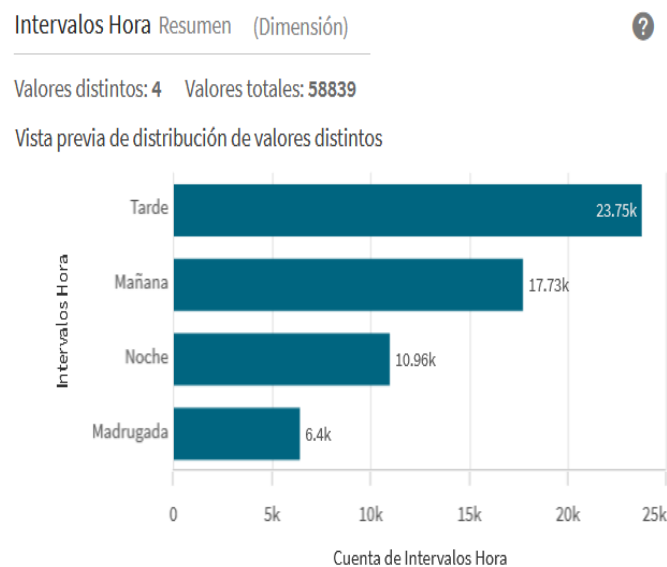


Figura 2.11: Opción resumen de columnas con Qlik Sense.

Qlik Sense también permite reemplazar valores de campo en una tabla, por ejemplo, cuando existen diferentes términos para un mismo objeto. Otra herramienta de utilidad es definir valores de campo como nulos. Además, a diferencia de Power BI, Qlik Sense permite personalizar el orden de los valores de una dimensión. Esto resulta útil cuando no queremos ordenar los valores por orden alfabético o numérico. Un ejemplo sería el caso de la variable `Intervalos Hora`, que toma los valores ‘Madrugada’, ‘Mañana’, ‘Tarde’ y ‘Noche’ y que podemos ver en la Figura 2.12.



Figura 2.12: Personalización del orden de la columna con Qlik Sense.

Qlik también permite dividir un campo existente en nuevos campos. Por ejemplo, si se tiene una dirección y se quiere quedar con la ciudad o con el código postal. Por último, Qlik también permite la integración de scripts de R y Python.

Además, cabe destacar que Qlik Sense también posee un editor de carga de datos donde aparecen predefinidos los formatos de la carga de datos, los decimales, los días de la semana, del mes, etc. Se trata de algo muy parecido al script con lenguaje M de Power BI. En la Figura 2.13 mostramos el editor de carga de datos de Qlik Sense.

```

1 SET ThousandSep='.';
2 SET DecimalSep='.';
3 SET MoneyThousandSep='.';
4 SET MoneyDecimalSep='.';
5 SET MoneyFormat='$ ###0.00;- $ ##0.00';
6 SET TimeFormat='h:mm:ss TT';
7 SET DateFormat='M/D/YYYY';
8 SET TimestampFormat='M/D/YYYY h:mm:ss[.fff] TT';
9 SET FirstWeekDay=0;
10 SET BrokenWeeks=1;
11 SET ReferenceDay=0;
12 SET FirstMonthOfYear=1;
13 SET CollationLocale='en-US';
14 SET CreateSearchIndexOnReload=1;
15 SET MonthNames='Jan;Feb;Mar;Apr;May;Jun;Jul;Aug;Sep;Oct;Nov;Dec';
16 SET LongMonthNames='January;February;March;April;May;June;July;August;September;October;November;December';
17 SET DayNames='Lun;Mar;Mie;Jue;Vie;Sab;Dom';
18 SET LongDayNames='Lunes;Martes;Miércoles;Jueves;Viernes;Sábado;Domingo';
19 SET NumericalAbbreviation='3:k;6:M;9:G;12:T;15:B;18:E;21:Z;24:Y;-3:m;-6:µ;-9:n;-12:p;-15:f;-18:a;-21:z;-24:y';
20

```

Figura 2.13: Editor de carga de datos de Qlik Sense.

Por otra parte, Tableau posee una herramienta propia para la preparación de los datos, conocida como 'Tableau Prep Builder'. Sin embargo, esta herramienta es de pago, en particular, para el uso individual el precio a fecha de 05/2022 es de 61.84€ (70 USD). Dentro de sus funcionalidades se pueden aplicar filtros, añadir columnas, reemplazar valores y agrupar o eliminar campos. También se pueden editar los valores de los datos, convertirlos

a mayúsculas o minúsculas, eliminar letras, números, signos de puntuación, espacios y espacios adicionales. Al igual que en Power BI existe la opción de aplicar tareas de limpieza a través de scripts de R o Python. Por último, una de las novedades es que Tableau puede analizar los datos y recomendar operaciones de limpieza que se pueden aplicar de forma automática para corregir problemas detectados en los datos.

En Superset, Redash y Metabase, dado que tan sólo permiten conexiones a bases de datos, las tareas de limpieza y tratamiento de datos se deben realizar previamente en el gestor de bases de datos.

Como conclusión, las herramientas Power BI, Qlik Sense y Tableau poseen funcionalidades semejantes en cuanto a las opciones de limpieza y tratamiento de los datos. Sin embargo, para utilizar estas herramientas tanto en Qlik Sense como en Tableau es necesario tener una licencia de pago. En el caso de las herramientas *open source*, Superset, Redash y Metabase, no presentan ninguna herramienta para la preparación y limpieza de datos.

2.5. Creación de visualizaciones

La creación de visualizaciones en Power BI es muy sencilla e intuitiva. En primer lugar se debe ir a la pantalla principal de Power BI y seleccionar la página del informe donde se quiera realizar la visualización. A continuación, en el apartado de visualizaciones aparece una lista con múltiples opciones, entre las que se encuentran distintos tipos de gráficos de barras, de líneas, de dispersión, circulares, tablas, KPI, filtros, etc. Seleccionando cualquiera de ellos se añadirá a la página y aparecerán varios campos a cubrir, por ejemplo ejes, leyendas, valores, etc, que variarán dependiendo del tipo de gráfico seleccionado. Aquí se deberán arrastrar las columnas de las tablas que son de interés y ya estará realizada la visualización. Un ejemplo se muestra en la Figura 2.14, donde se ha creado un gráfico de anillos cuya leyenda es la columna **Tipo Gravedad** y los valores son la columna **Total_Accidentes**.

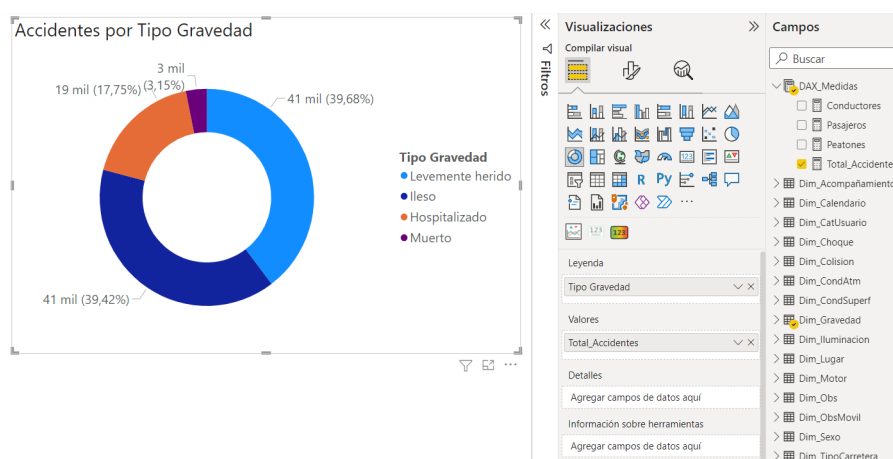


Figura 2.14: Creación de visualización con Power BI.

En el caso de que se quiera una visualización que no se encuentre en la lista se pueden obtener más objetos visuales a través de la opción de los tres puntos. Existen dos formas de hacerlo, con la opción de 'Obtener más objetos visuales' o bien 'Importar un objeto visual de un archivo', que veremos en el Capítulo 3. Si se hace click en la primera de ellas se abrirá una nueva página con múltiples posibilidades, donde se puede buscar por nombre

o filtrar por categoría. En la Figura 2.15 se muestran algunas de estas opciones.

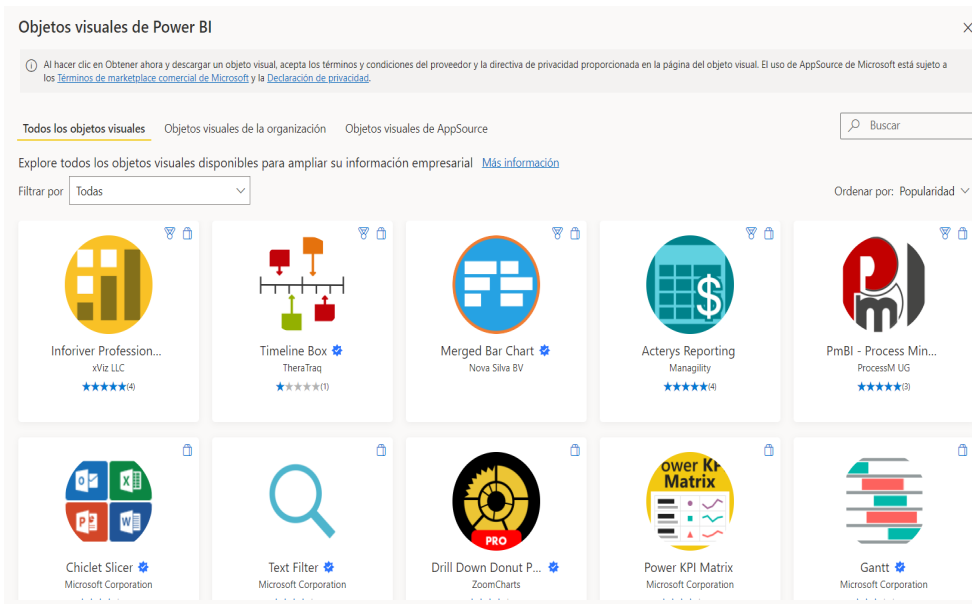


Figura 2.15: Obtención de más objetos visuales con Power BI.

La creación de visualizaciones en Qlik Sense comienza después de la carga de los datos, en la opción de 'Crear nueva hoja'. En esta herramienta existen múltiples opciones de gráficos: de botón, diagramas de cajas, gráficos de barras, de bloques, de dispersión, de líneas, de tarta, de viñetas, en cascada, histogramas, indicadores, mapas, etc. Se debe elegir un gráfico y arrastrarlo a la hoja. Entonces será necesario añadir un título, una dimensión y una medida, como se muestra en la Figura 2.16.

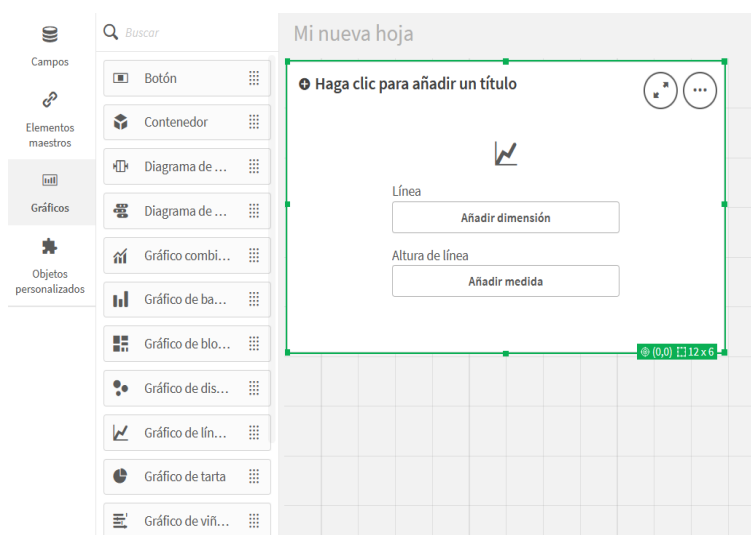


Figura 2.16: Creación de visualización con Qlik Sense.

Al hacer click en 'Añadir dimensión' o 'Añadir medida' se muestra una lista con las opciones que se pueden utilizar. A continuación, en la Figura 2.17, se muestra un gráfico de líneas que muestra el número de accidentes por mes.

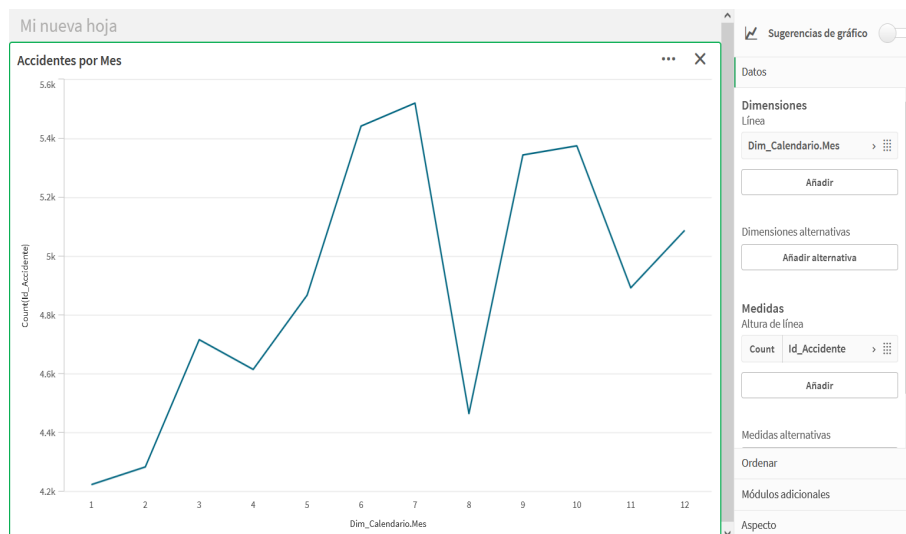


Figura 2.17: Gráfico de líneas con Qlik Sense.

Como se puede observar, en la parte derecha del gráfico aparece un menú con distintas opciones para personalizarlo. En primer lugar existe una opción de ‘Datos’, donde aparecen las dimensiones y medidas seleccionadas. Aquí se puede especificar si queremos incluir o no valores nulos o si se quiere limitar el número de datos que se muestran. A continuación se tiene una opción de ‘Ordenar’, donde se puede utilizar un orden personalizado, ordenar por una expresión, numéricamente o alfabéticamente, ya sea de forma ascendente o descendente.

Otra opción son los ‘Módulos adicionales’, donde se tiene tres opciones: ‘Manejo de datos’, ‘Líneas de referencia’ o ‘Líneas de referencia de dimensión’. En el manejo de datos se tiene la opción de incluir o no valores cero, y especificar si se quiere una condición de cálculo, por ejemplo, representar el gráfico de líneas anterior sólo para ciertos meses. Las líneas de referencia son líneas que cruzan el área del gráfico desde un punto de datos en el eje de la medida. Las líneas de referencia dimensionales son líneas de referencia a lo largo del eje de dimensión. En ambos casos puede ponerse una etiqueta que explique el significado de la línea de referencia, lo cual puede resultar útil para el usuario.

Por último, en la opción de ‘Aspecto’ se tienen distintas opciones para configurar la apariencia de los gráficos. En primer lugar, en la opción de ‘General’, se permite activar o desactivar el título, cambiar su nombre, poner subtítulos o notas a pie de página y mostrar detalles, donde se puede aportar una descripción de las medidas y dimensiones utilizadas. A continuación se tiene la opción de ‘Presentación’, que depende del tipo de gráfico empleado. Por ejemplo, en el gráfico de líneas, se puede especificar que el área bajo la curva esté sombreada, o que el gráfico se disponga de manera vertical en vez de horizontal. Además, se puede elegir cómo desplazarse a través del gráfico, donde se tiene la opción de un mini gráfico, una barra o ninguno, y si se quiere que el ajuste del desplazamiento empiece por el inicio o por el fin. Se puede especificar cómo se muestran los valores perdidos, si se quiere que aparezcan como huecos, como ceros o como conexiones. Se pueden mostrar los puntos de datos y ponerles etiquetas, y si se quiere utilizar o no una rejilla en el gráfico. En la opción de ‘Colores y leyenda’ se puede especificar el color del gráfico, o si se quiere que este color varíe por medida o expresión. Por último, en las opciones de ‘Ejes X e Y’ se puede cambiar la escala utilizada, poner títulos en los ejes, elegir su posición y poner etiquetas.

En la herramienta Tableau Public la creación de visualizaciones es muy diferente al

resto de opciones. En primer lugar no se arrastran los gráficos al *dashboard*, sino que se crean en hojas de trabajo independientes. Además, no existen unos gráficos predefinidos donde elegir, sino que se arrastran campos y medidas a ‘filas’ o ‘columnas’, y dependiendo de la combinación se tiene una visualización u otra. Nótese que esta herramienta facilita unas opciones de cómo combinar las medidas y dimensiones para obtener los gráficos deseados. Estas opciones pueden verse en la Figura 2.18.



Figura 2.18: Opciones de visualización con la herramienta Tableau.

Como opciones de visualización se tienen las tablas de texto, mapas de calor, tablas de resaltado, mapas de símbolo, mapas, gráficos circulares, barras horizontales, barras apiladas, barras paralelas, mapas de árbol, vistas circulares, círculos paralelos, líneas continuas o discretas, líneas dobles, gráficos de áreas continuos o discretos, combinación doble, diagramas de dispersión, histogramas, diagramas de campos y valores, vistas de Gantt, gráficos de balas y burbujas agrupadas.

Una vez creado el gráfico, existen varias opciones de personalización del mismo. En la Figura 2.19 se muestran a la izquierda las distintas opciones disponibles para el gráfico de barras que representa el número de accidentes por grupo de edad.

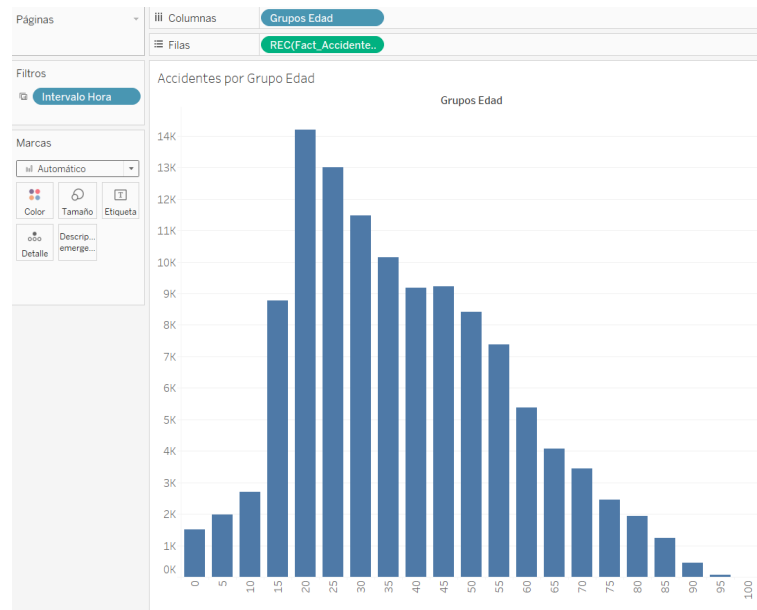


Figura 2.19: Personalización de visualizaciones con Tableau.

En primer lugar se tiene un menú desplegable que permite cambiar la forma del gráfico, en vez de tener un gráfico de barras se puede elegir un gráfico de líneas u otras de las opciones que se veía anteriormente. A continuación se tiene la opción de color, para modificar el color del gráfico. La siguiente opción es el tamaño, que en este caso modifica la anchura de las barras. Se pueden añadir también etiquetas y descripciones emergentes, o bien añadir más detalle al gráfico. Además, si se está utilizando un tipo de gráfico que contenga leyenda, esta podrá ser colocada de forma independiente a la visualización, por lo que puede ser colocada donde interese.

El proceso de creación de visualizaciones en Metabase consta de varias etapas. En primer lugar, y a diferencia de otras herramientas, no se arrastran los campos de las tablas, sino que es necesario hacer una consulta con lenguaje SQL para poder acceder a las columnas de las tablas que serán de interés a la hora de realizar el gráfico. Una vez hecha la consulta se puede elegir qué tipo de visualización se quiere realizar y se muestra como quedaría. Entre las opciones de visualización se tiene el gráfico de línea, de barras, combo, área, fila, cascada, dispersión, pastel, embudo, tendencia, progreso, contador, número, tabla, tabla dinámica y mapa. En la Figura 2.20 se muestra un ejemplo de gráfico de número.

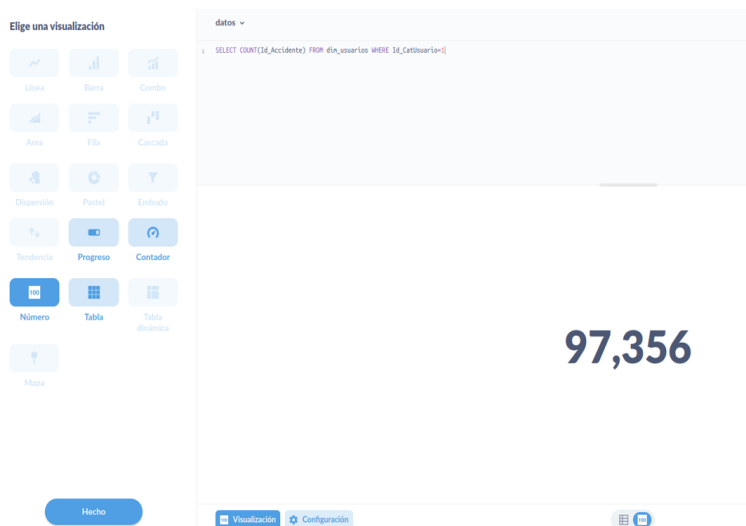


Figura 2.20: Ejemplo de visualización con la herramienta Metabase.

Una vez realizado el gráfico, puede ser guardado. Es entonces cuando se pregunta si se quiere añadir a algún *dashboard*. Otra forma de añadir visualizaciones sería acceder al *dashboard* y desde allí darle al símbolo '+', donde se mostrará una lista de todas las posibles visualizaciones.

En la herramienta Redash el proceso de creación de visualizaciones es muy similar a Metabase. Para crear los gráficos también se debe emplear lenguaje SQL sobre las columnas de las tablas que se quieran seleccionar. Una vez terminada la consulta aparece una opción de nueva visualización, donde se despliega el editor de visualizaciones, que se muestra en la Figura 2.21.

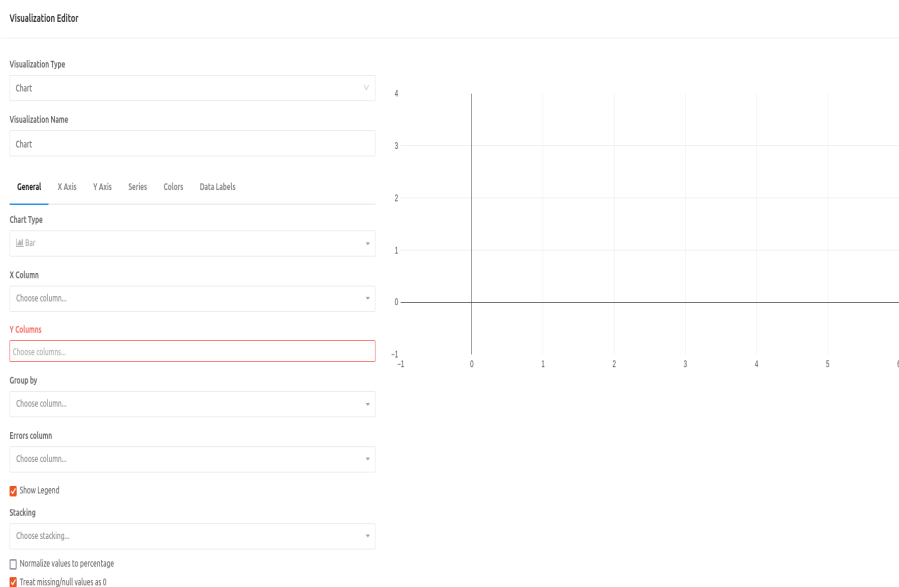


Figura 2.21: Editor de visualizaciones de Redash.

En el editor de visualizaciones se puede especificar el tipo de visualización, título, ejes, colores, etiquetas, leyendas, etc. En la opción tipo de gráfico se despliega una lista con los nombres de los posibles gráficos, junto con una pequeña representación de lo que sería el

gráfico. Una vez analizada la herramienta, esta representación resulta escasa cuando no se tienen demasiados conocimientos de la herramienta.

Otra de las características es que se pueden poner los valores como porcentajes y tratar los datos faltantes o nulos como ceros.

Por último, explicaremos la creación de visualizaciones con la herramienta Superset. Ésta funciona un poco diferente a Metabase y Redash, puesto que si queremos crear una visualización donde se utilicen campos de distintas tablas es necesario crear un conjunto de datos con las columnas que nos interesan. Es decir, en primer lugar realizamos una consulta SQL a nuestro conjunto de datos para conseguir las columnas necesarias. Con esa *query* creamos una vista materializada y finalmente construimos la visualización a partir de este nuevo conjunto de datos. En caso de que los *dashboards* no sean más o menos estáticos puede resultar poco práctico, puesto que cada vez que se necesiten columnas de diferentes tablas para crear una visualización se deberá crear una nueva vista.

Para construir una visualización deberemos ir al apartado de ‘Charts’, donde se nos pedirá el dataset para crear el gráfico y el tipo de visualización que queramos. En este último apartado se muestran dos formas de buscar el gráfico de interés. La primera de ellas es mediante *tags*, como por ejemplo el tag Popular, donde encontramos las visualizaciones más utilizadas. La segunda de ellas es mediante la categoría, donde encontramos las categorías correlación, distribución, evolución, flow, KPI, mapas, ‘Part of a Whole’ ranking, tablas y herramientas. En cada una de estas categorías se muestran varios estilos de gráficos, por ejemplo, para la categoría distribución tenemos histogramas y diagramas de cajas. En la Figura 2.22 mostramos estas opciones.

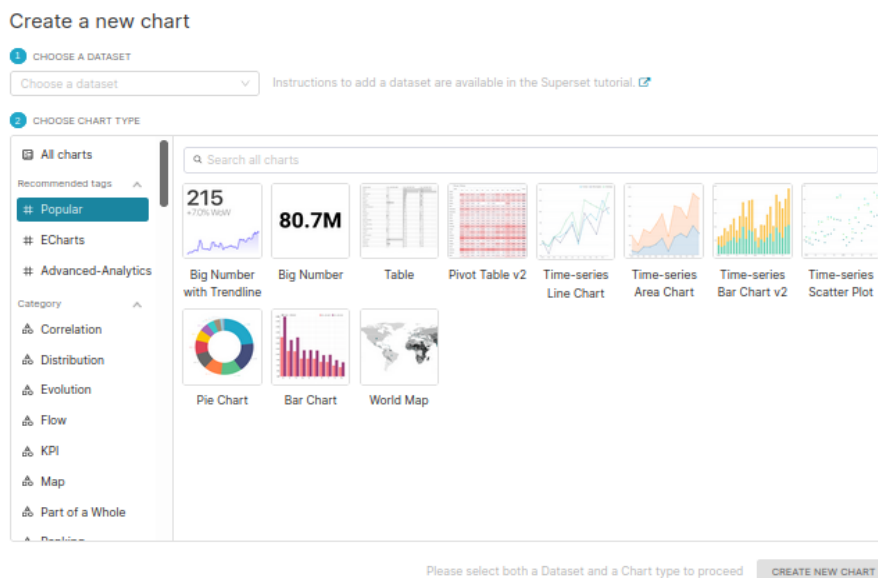


Figura 2.22: Creación de visualizaciones Superset

Una vez seleccionados el conjunto de datos y el tipo de gráfico se tendrán que especificar las características para la visualización. Dependiendo del tipo de gráfico las especificaciones serán distintas. En particular, en la Figura 2.23 se muestra la creación del gráfico de accidentes por mes. En primer lugar se muestra una columna con las métricas y columnas utilizadas, que en este caso son las columnas *Fecha* e *Id.Accidente*, y la función *count*. En la segunda columna hay dos opciones, ‘Data’, donde se deben añadir los campos a utilizar en la visualización, y ‘Customize’, donde se especifican los aspectos visuales. En

Data aparece, en primer lugar, el tipo de gráfico utilizado, que en este caso se corresponde con un gráfico de líneas para series temporales. A continuación se debe seleccionar el Tiempo, que en este caso será la columna Fecha, y la granularidad temporal, que puede ser el valor original, segundos, minutos, horas, días, meses y cuatrimestres. Además, en este apartado se permite aplicar filtros de temporales. A continuación se tiene la opción de Query, donde se debe seleccionar la columna que se quiere representar, en este caso el número de accidentes, que se define como `count(Id.Accidente)`. Aquí se puede especificar si se quiere agrupar por alguna columna, filtros, orden ascendiente o descendiente y número máximo de filas a considerar. Además, incluye también una sección de analítica avanzada, donde se pueden crear ventanas deslizantes, comparaciones temporales o remuestras. La siguiente opción son anotaciones o etiquetas y por último análisis predictivo. Por otra parte, en la sección Customize se pueden especificar los títulos de los ejes, colores, leyendas y otros formatos de gráfico.

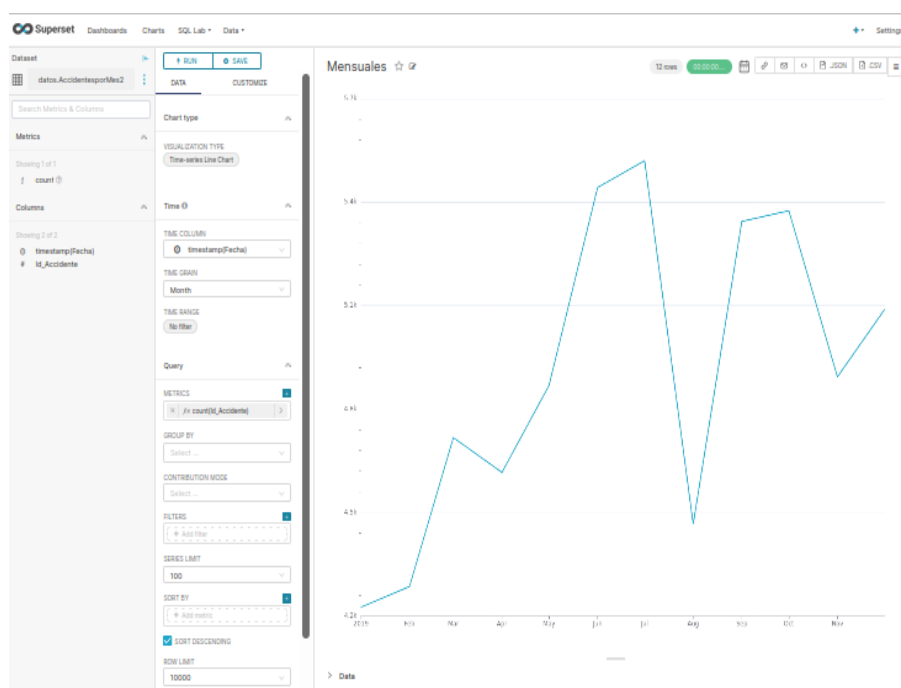


Figura 2.23: Ejemplo visualización Superset

Como resumen, todas las herramientas descritas anteriormente presentan una gran diversidad de gráficos, aunque cabe destacar que Power BI permite la opción de añadir nuevos gráficos desde fuera de la propia plataforma. Además, tanto Power BI como Qlik Sense poseen bastantes opciones para la personalización de las visualizaciones. En el caso de Tableau resulta poco intuitiva la creación de gráficos utilizando las opciones de filas y columnas. En el caso de las herramientas Metabase, Superset y Redash, la creación de visualizaciones se realiza a través del lenguaje SQL, por lo que para los usuarios que estén habituados a este lenguaje lo pueden considerar una gran ventaja.

2.6. Creación de nueva información

En ocasiones, la información almacenada en una base de datos no es suficiente para dar respuestas a ciertas preguntas, por lo que es necesario transformar dicha información, ya sea a través de nuevas columnas o medidas.

Para la creación de nueva información Power BI utiliza el lenguaje DAX. *Data Analysis Expressions* o DAX es un lenguaje creado por Microsoft en el año 2009 para el análisis de datos. Este lenguaje se emplea en programas como Excel, SQL Server Analysis Services y Power BI. DAX posee una colección de funciones, operadores y constantes que se pueden usar en una fórmula o expresión, para calcular y devolver uno o más valores. De este modo se obtiene nueva información a partir de los datos ya incluidos en el modelo.

Dentro del lenguaje DAX existen tres conceptos fundamentales: sintaxis, funciones y contexto. La sintaxis incluye los distintos elementos que componen una fórmula. Si la sintaxis no es correcta Power BI devolverá un error. Las funciones son fórmulas predefinidas que realizan cálculos por medio de valores específicos, denominados argumentos, en un orden o estructura determinados. Las funciones DAX se dividen en las siguientes categorías: Fecha y Hora, Inteligencia de tiempo, Información, Lógicas, Matemáticas, Estadísticas, Texto, Primarias/Secundarias y Otras. Por último se hablará del contexto, donde se distinguirá el contexto de fila y el contexto de filtro. El contexto de fila se aplica siempre que una fórmula tiene una función que use filtros para identificar una fila individual en una tabla. El contexto de filtro se aplica además del contexto de fila. Por ejemplo, para restringir aún más los valores que quiera incluir en un cálculo, puede aplicar un contexto de filtro que especifique un valor determinado en ese contexto de fila.

Un ejemplo de creación de nueva información en Power BI sería la construcción de la columna `Intervalos_Hora`. Ésta se muestra en la Figura 2.24.

Id_Accidente	Hrnm	Id_Iluminacion	Id_CondAtm	Id_Colision	direccion	Lat	Long	Hora	Intervalos_Hora	Fecha
201900000327	0,708333333333333	1	1	3	SAINT-EXUPERY (RUE DE)_SUD	492683000	24596800	16	Tarde	miércoles, 14 de agosto de 2019
201900000556	0,708333333333333	1	1	3	BD PERIPHERIQUE EXTERIEUR	488377373	24129528	16	Tarde	viernes, 29 de noviembre de 2019
201900000997	0,708333333333333	1	1	3	Col du bougnon	433895020	66765960	16	Tarde	domingo, 17 de febrero de 2019
201900001141	0,708333333333333	1	1	3	JEAN JAURES (AV)	457138900	48194600	16	Tarde	martes, 12 de febrero de 2019
201900001501	0,708333333333333	1	1	3	RADOLFZELL AVENUE DE	434991940	49878290	16	Tarde	viernes, 16 de agosto de 2019
201900001549	0,708333333333333	1	1	3	CLUCHY (QUAI DE) DE (104 A 202)	489036700	22900900	16	Tarde	domingo, 16 de junio de 2019
201900001760	0,708333333333333	1	1	3	rd 813	434634580	15684950	16	Tarde	viernes, 25 de octubre de 2019

Figura 2.24: Creación de la columna `Intevalos_Hora` usando DAX

Esta columna se crea utilizando la función lógica IF, de modo que si la hora del accidente es menor o igual que las 6 entonces el Intervalo Hora será ‘Madrugada’, si es menor o igual que 12 será ‘Mañana’, si es menor o igual que 18 será ‘Tarde’ y en otro caso será ‘Noche’.

En Qlik Sense para la creación de nuevas columnas hay que acceder a la tabla y hacer click en la opción ‘Añadir campo’. Se abrirá entonces una pantalla como la que vemos en la Figura 2.25.



Figura 2.25: Creación de la columna Intevalos_Hora en Qlik Sense

Aquí se tendrá que especificar el nombre de la nueva columna y la expresión para crearla. La sintaxis es similar a la utilizada en el lenguaje DAX de Power BI, sin embargo cambia la forma de acceder a las columnas, en el sentido de que si el nombre de la columna es único, es decir, no existe otra tabla con una columna con el mismo nombre, entonces no es necesario especificar de qué tabla proviene. Además, Qlik Sense acepta únicamente las comillas simples y no las dobles.

Como se puede observar en la Figura anterior, una vez creada la columna, se muestra una vista previa de los resultados obtenidos. Para más información de la sintaxis permitida para la creación de campos calculados puede consultarse la página oficial de Qlik.

En la versión Public de Tableau no es posible crear nueva información, es decir, no permite crear nuevas columnas a partir de otras ya existentes. En este caso se han creado las columnas necesarias para la realización de los gráficos mediante Excel y se han cargado los datos completos de la forma habitual.

Las herramientas Superset, Metabase y Redash son muy similares a la hora de crear nueva información. Estas tres opciones emplean el lenguaje SQL para crear nueva información, y lo hacen a la hora de realizar nuevos gráficos. Otra opción sería crear dichas columnas en el gestor de datos, antes de cargar el conjunto de datos en el programa.

Como resumen, tanto Power BI como Qlik Sense poseen un lenguaje propio para la creación de nueva información. En el caso de Power BI, el lenguaje DAX es muy similar al usado en la herramienta Excel, por lo que puede resultar una ventaja para aquellos usuarios con conocimientos de esta herramienta. En el caso de Tableau sería necesaria una licencia de pago para la creación de nuevas columnas. Por último, las herramientas Superset, Metabase y Redash utilizan lenguaje SQL para la creación de nueva información, por lo que se trata de una ventaja para aquellos usuarios con conocimientos de este lenguaje.

2.7. Dashboards

Los *dashboards* o cuadros de mando son herramientas de *Business Intelligence* que representan de manera visual los indicadores clave de desempeño (KPI), métricas y datos fundamentales para hacer un seguimiento del estado de una empresa, departamento o un proceso específico. Los *dashboard* deben ser útiles, es decir, deben contener aquellas métricas que ayuden a responder a las preguntas clave de negocio, y que pueden ir cambiando a lo largo del tiempo. Debe ser visual y comprensible, para poder tomar decisiones y pasar a la acción. Y por último debe ser actual, puesto que los datos evolucionan con gran rapidez y es necesario que la información esté actualizada y se muestre en el *dashboard* en tiempo real.

En Power BI los *dashboard* son un conjunto de hojas, donde en cada una de ellas se arrastran los gráficos de interés y se personalizan sobre la hoja final. A continuación mostramos los dashboard obtenidos con Power BI.

En primer lugar, en la Figura 2.26 se muestra un resumen visual del número de accidentes de Francia en el año 2019. En la parte superior izquierda se ha añadido el logo de la empresa y el título de la hoja, mientras que en la parte izquierda se han añadido varios filtros. En el primero de ellos se puede especificar el tipo de usuario, ya sea conductor, pasajero o peatón; a continuación la gravedad del accidente, que puede ser ileso, levemente hospitalizado, hospitalizado o muerto; el sexo y por último el intervalo horario, que puede ser madrugada, mañana, tarde o noche.

A continuación se muestran cuatro tarjetas, que resumen los datos de accidentes obtenidos. En Francia en el año 2019 hubo un total de 58.840 accidentes, donde se vieron involucrados 97.350 conductores, 23.810 pasajeros y 11.230 peatones. Además, se muestran también tres gráficos de líneas, donde se representa el número de accidentes mensuales, semanales y diarios.

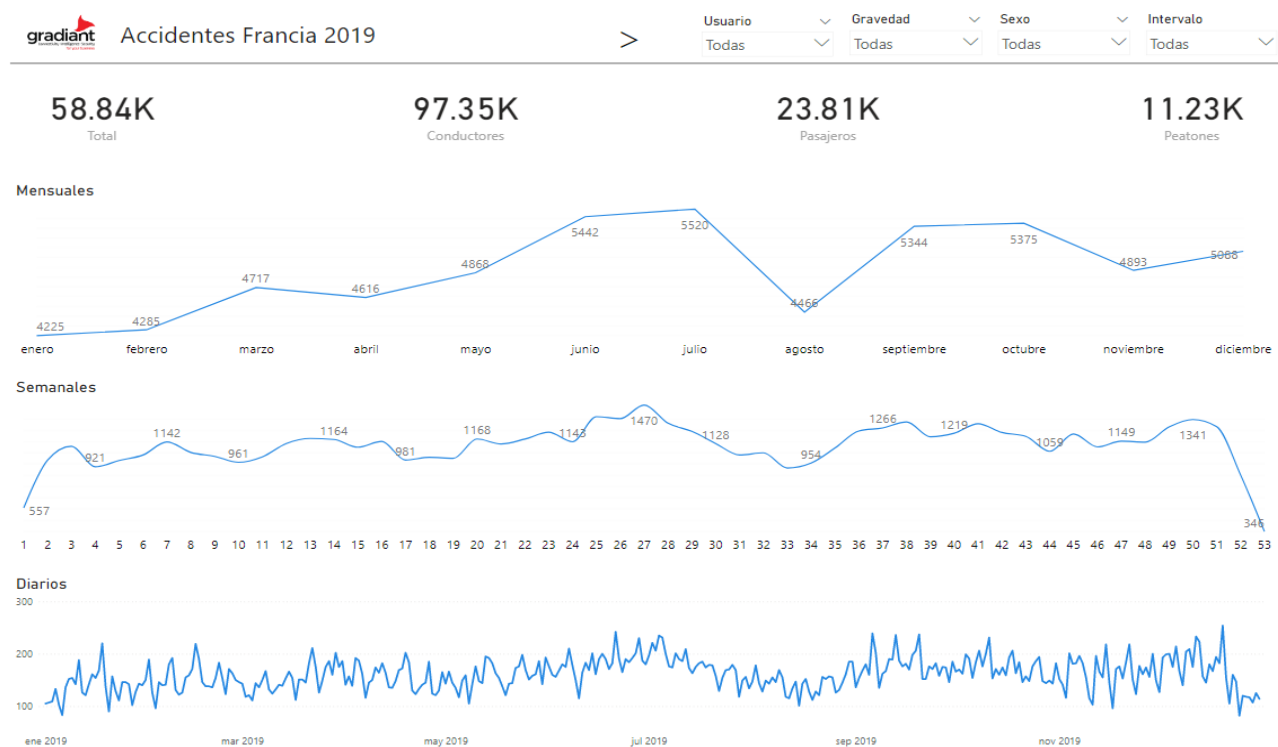


Figura 2.26: Hoja 1 Dashboard Power BI

En la siguiente hoja (Figura 2.27), se muestra la frecuencia con la que ocurren estos accidentes. En este caso los filtros utilizados son el mes, la semana, el día y el intervalo horario. A continuación se muestran dos gráficos de líneas, donde se representa el número de accidentes por día de la semana y por mes. Se tiene que el viernes es el día de la semana con mayor tasa de accidentes, mientras que el mes con más accidentes es julio. Además, se ha representado mediante un gráfico de barras el número de accidentes por intervalo horario, donde se observa que por la tarde ocurren más accidentes, seguido de la mañana, la noche y la madrugada. Por último, se ha representado en un gráfico de anillos el número de accidentes por trayecto. Se observa que el mayor número de accidentes ocurre en los trayectos de ocio.

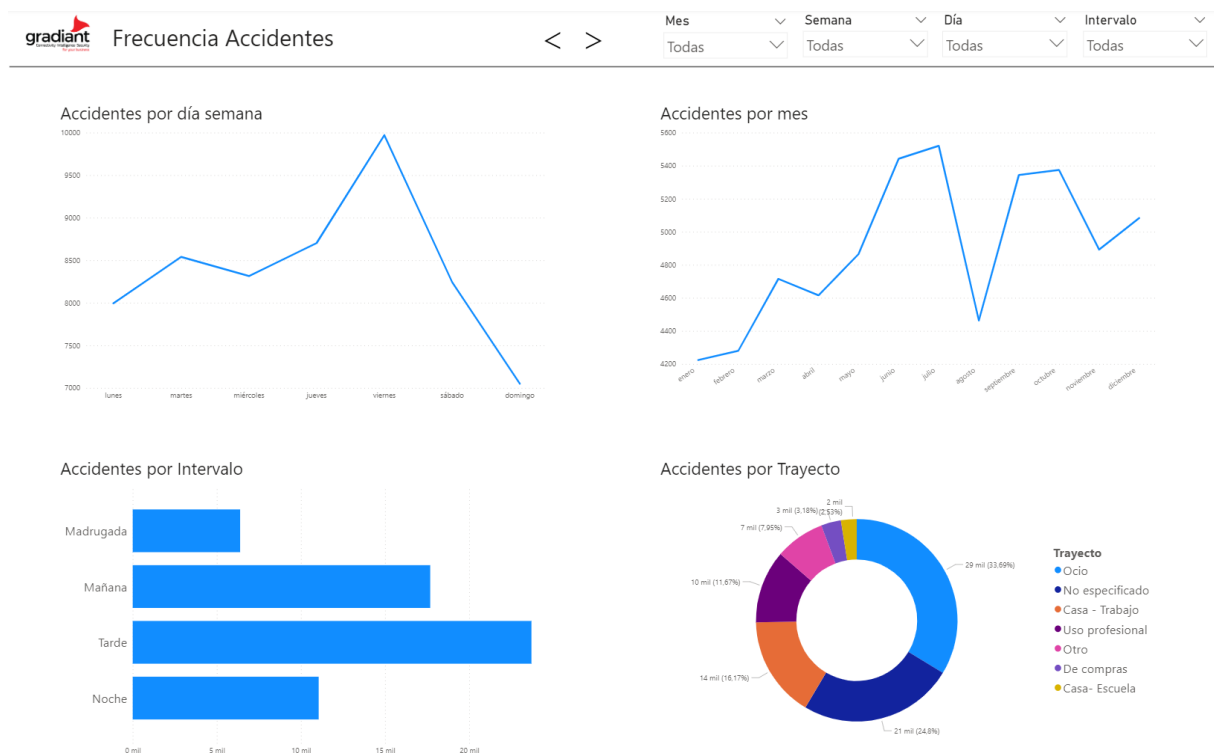


Figura 2.27: Hoja 2 Dashboard Power BI

En la siguiente hoja (Figura 2.28), se muestran las características de los usuarios involucrados en los accidentes. Los filtros utilizados coinciden con los de la segunda hoja del *dashboard*. En primer lugar se ha utilizado un gráfico de anillos para representar el número de accidentes por tipo de gravedad, donde se ha visto que en el 39.68 % de los accidentes el usuario ha salido levemente herido, mientras que el 39.42 % ha sido ileso, el 17.75 % ha sido hospitalizado y un 3.15 % ha fallecido. A su derecha se muestra un diagrama de barras, donde se representa el número de accidentes por tipo de acompañamiento. Se observa que la categoría con más accidentes es no especificado, por lo que no se tiene información acerca del acompañamiento de dichos usuarios. El siguiente gráfico es un gráfico circular, que muestra el número de accidentes según el sexo. Se tiene que en el 63.33 % de los accidentes estuvieron involucrados hombres. Por último se observa el número de accidentes por rango de edad. Como se puede observar el rango de edades va desde los 0 años hasta los 100, puesto que se tienen en cuenta aquellas personas involucradas en un accidente de tráfico, por lo que pueden ser peatones, acompañantes o conductores. Se observa que el rango de edad con mayor tasa de accidentes es de 21 a 25 años.

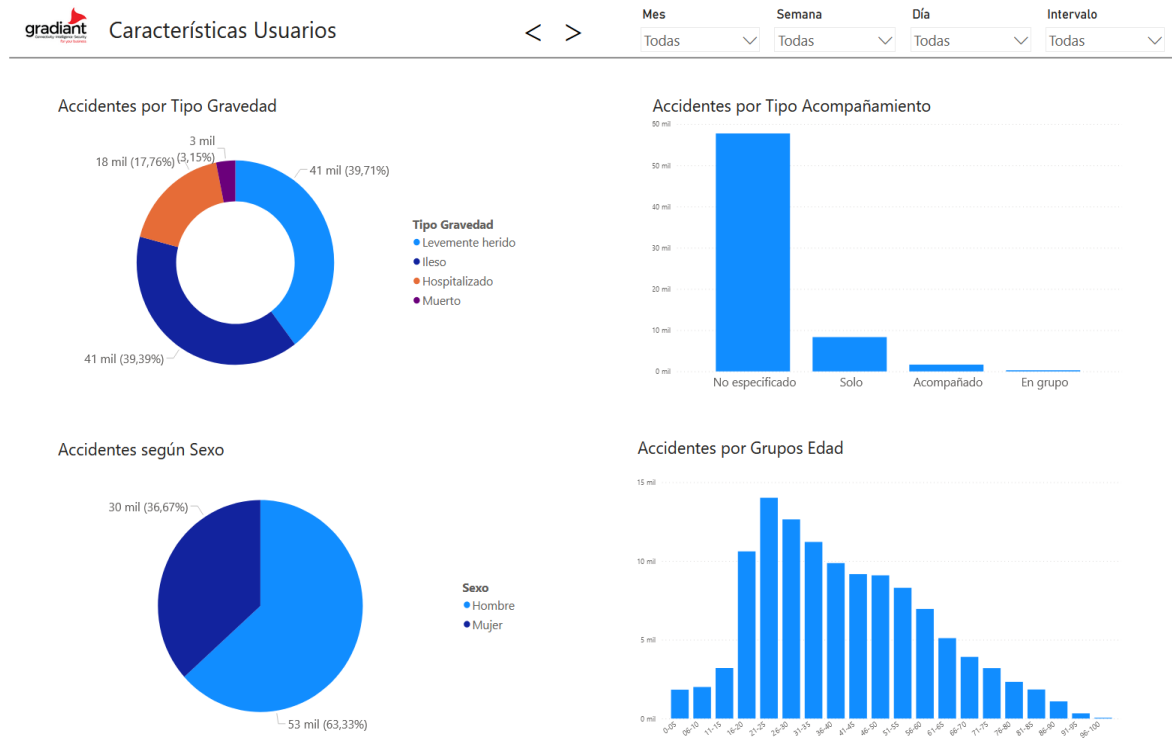


Figura 2.28: Hoja 3 Dashboard Power BI

En la Figura 2.29 se hace un resumen de las condiciones externas al accidente, donde se engloban el tipo de carretera, las condiciones de la superficie y atmosféricas y el tipo de iluminación. En el primer gráfico circular se observa que casi la mitad de los accidentes ocurren en caminos comunales, mientras que otra gran parte de ellos ocurre en vías departamentales. A continuación se muestra un gráfico de barras, donde se observa que la mayor parte de los accidentes ocurren con condiciones atmosféricas normales, en ausencia de lluvia, viento, nieve u otros fenómenos. En el siguiente gráfico se representa el número de accidentes por tipo de iluminación. En este caso el grupo mayoritario es de día, siguiéndole aunque en menor medida de noche con alumbrado. Por último, se puede observar que la mayor parte de los accidentes ocurren en condiciones superficiales normales, seguido de superficies húmedas por la lluvia.

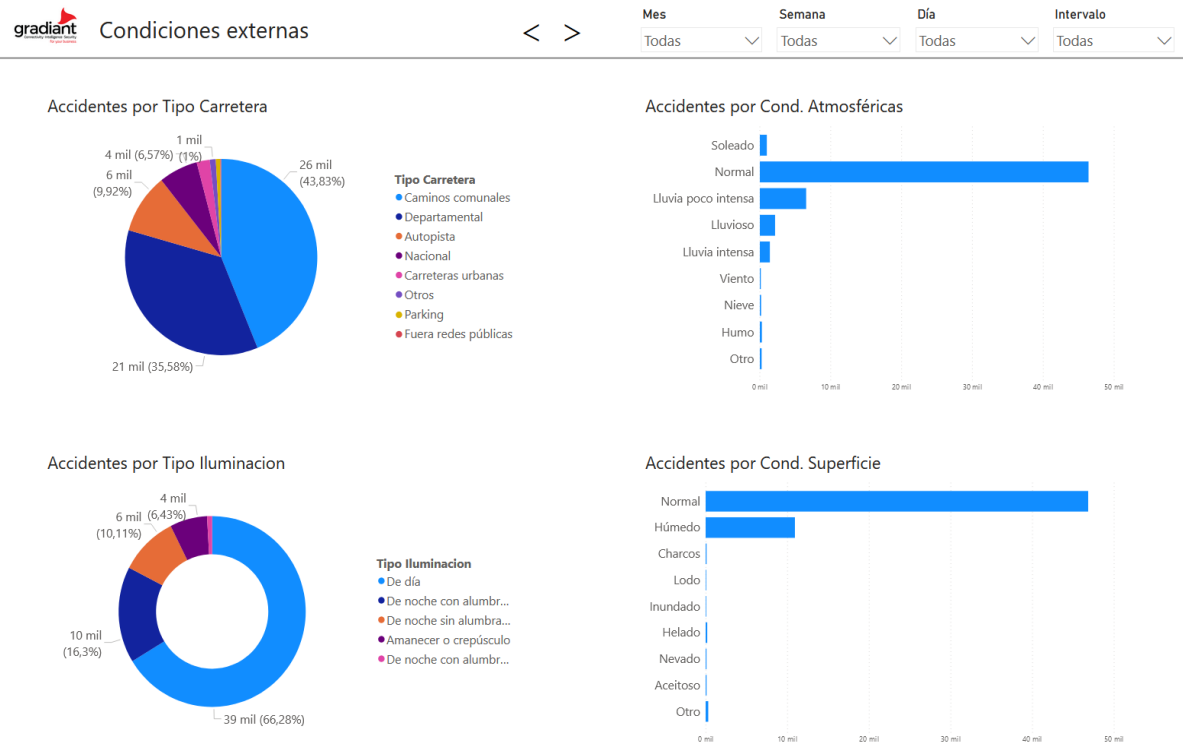


Figura 2.29: Hoja 4 Dashboard Power BI

En la última hoja del *dashboard*, que se muestra en la Figura 2.30, se han representado características relacionadas con el accidente, como puede ser el tipo de obstáculo, el tipo de choque o el tipo de colisión. En primer lugar se observa un gráfico circular que muestra el número de accidentes según el obstáculo fijo. El grupo mayoritario se corresponde con 'No aplicable', por lo que la mayor parte de accidentes se corresponden con obstáculos móviles. En el siguiente gráfico, que representa los accidentes según el obstáculo móvil, se ve que la mayor parte de accidentes tienen como obstáculo móvil un vehículo. En el tercer gráfico se muestra un diagrama de barras que representa el número de accidentes por tipo de choque. En este caso el choque delantero es el más frecuente. Por último, se observa un diagrama de barras que representa el número de accidentes por tipo de colisión, donde el grupo más frecuente es otra colisión, seguido de la colisión lateral.

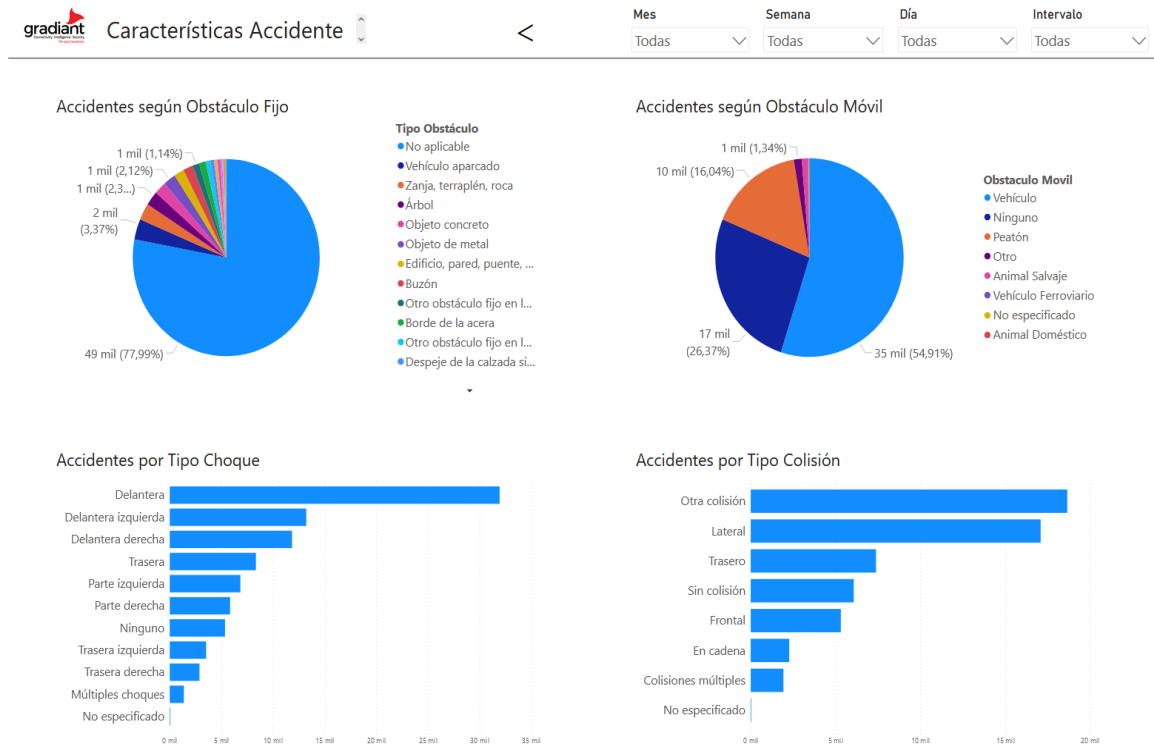


Figura 2.30: Hoja 5 Dashboard Power BI

Indicar que lo que se acaba de mostrar sería un ejemplo de un *dashboard* sencillo que podría ayudar con la toma de decisiones. Además, todos estos gráficos son interactivos, en el sentido de que podemos utilizarlos para filtrar información. Un ejemplo sería el que mostramos en la Figura 2.31, donde si en el primer gráfico se hace click en la opción de Hospitalizados, se filtran el resto de gráficos de esta hoja, aunque se podrían filtrar los gráficos de todas las hojas o bien el que sea de interés. En este caso puede verse que de los accidentes en los que la víctima resultó hospitalizada, 6000 eran mujeres, frente a 14000 hombres. Además, este tipo de accidentes son más frecuentes en los rangos de edad de 16 a 25 años. En base a esto podría tomarse la decisión de intentar concienciar a los jóvenes sobre la importancia de la seguridad vial, tanto como conductores como pasajeros.

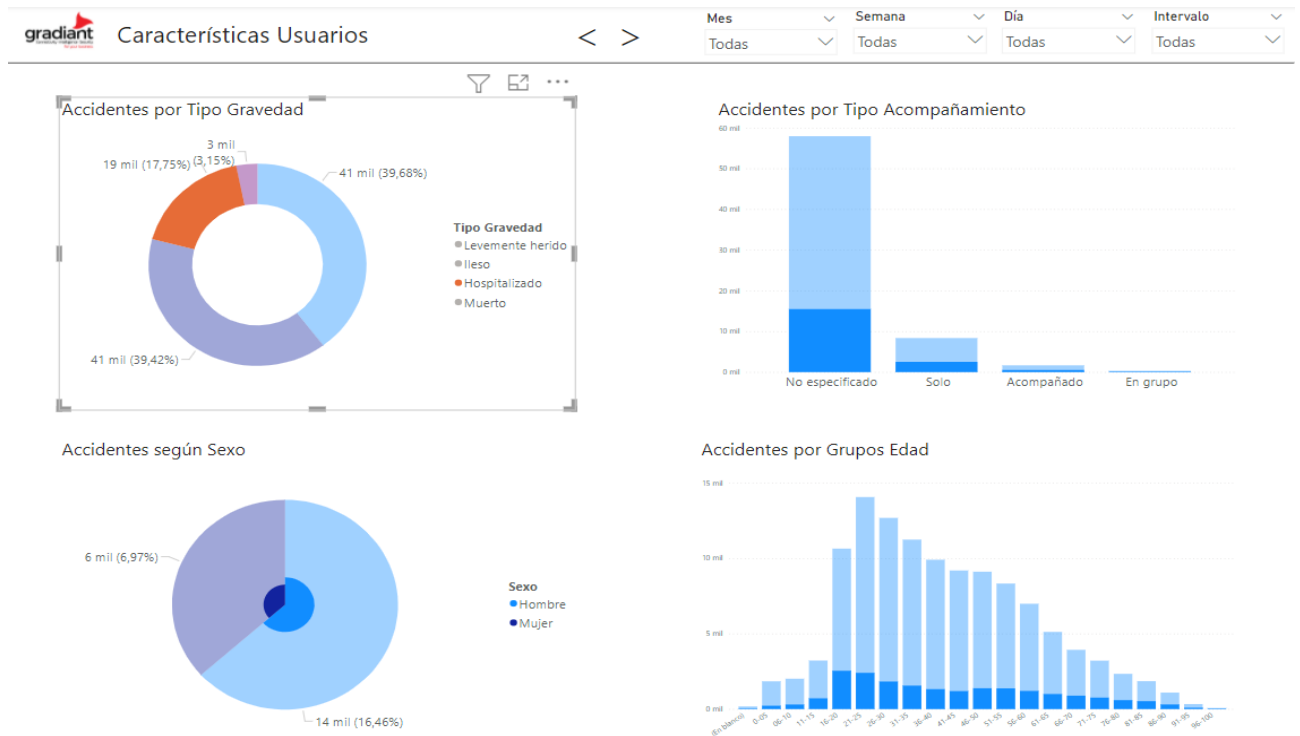


Figura 2.31: Ejemplo dashboard interactivo

En Qlik Sense, al igual que en Power BI, los *dashboard* son un conjunto de hojas, donde en cada una de ellas se arrastran los gráficos de interés y se personalizan sobre la hoja final. A continuación se muestran los *dashboard* obtenidos con esta herramienta, comenzando con la Figura 2.32. El formato de Qlik Sense es distinto del formato de Power BI, el fondo gris resalta los gráficos actualizados y hace que la vista no sea tan uniforme como en el *dashboard* de Power BI. Comparándolo con la Figura 2.26 se han podido realizar cada uno de los gráficos propuestos. Sin embargo, los filtros que se observan en la parte superior derecha son una lista en la cual es necesario navegar y no un menú desplegable como en Power BI, que resulta mucho más atractivo para el usuario. Nótese que en versiones posteriores han solucionado este problema. Además, los valores del número de accidentes por conductores, pasajeros y peatones no coinciden, aunque no se ha encontrado ningún motivo por el cual ocurra esto.

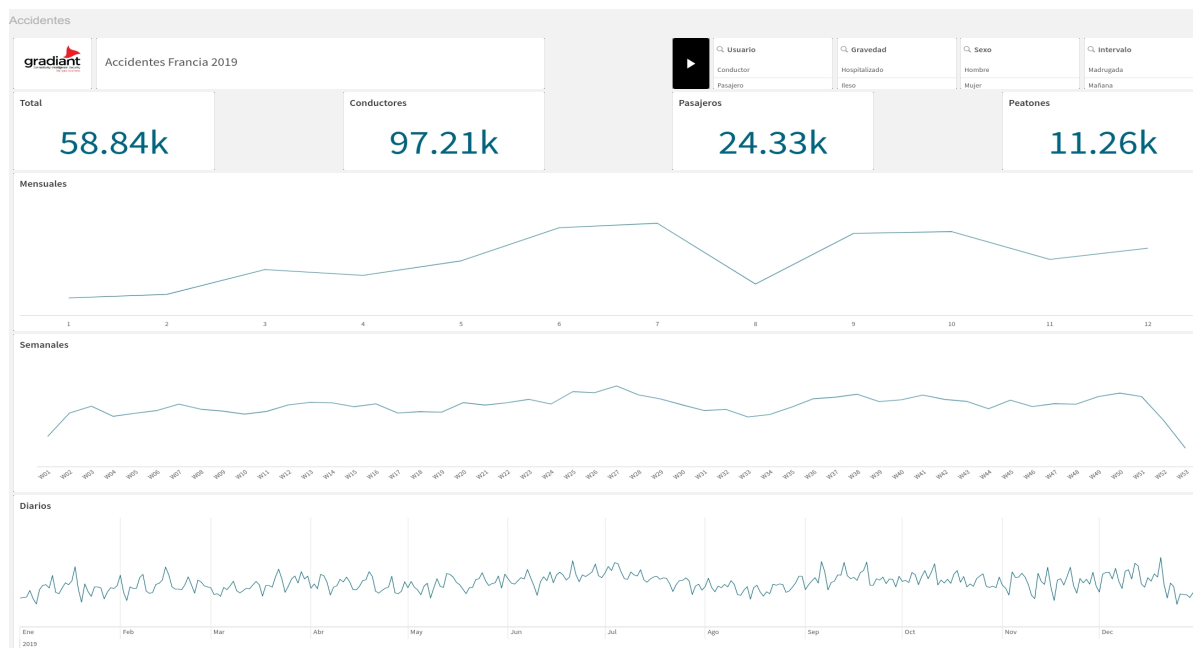


Figura 2.32: Hoja 1 Dashboard Qlik Sense

La segunda hoja del *dashboard* viene dada por la Figura 2.33. En este caso también se han podido obtener todos los gráficos, pero se ha tenido dificultades a la hora de ordenarlos. La primera dificultad ha ocurrido en el gráfico del número de accidentes por día de la semana. En ocasiones, las empresas tienen establecido que el primer día de la semana es Domingo, y lo mismo ocurre con Qlik Sense; si se quisieran ordenar numéricamente los días de la semana, esta herramienta seleccionará el domingo como primer día. En pruebas posteriores de la herramienta se ha visto que en el editor de carga de datos existe la opción ‘Set FirstWeekDay’, que de forma predeterminada es 6, es decir, domingo, por lo que bastaría poner 0, lunes, para que este problema se solucione. La siguiente dificultad surgió en el gráfico de accidentes por Intervalo, que como se puede observar está ordenado por Madrugada, Mañana, Noche y Tarde, cuando la noche debería ser la última etiqueta. En pruebas posteriores de la herramienta se ha podido solucionar este error, basta establecer un orden personalizado, como se ha visto en la Figura 2.12 y a la hora de crear el gráfico no especificar ningún orden, ni automático, ni numérico, ni alfabético ni personalizado.

A continuación, en la Figura 2.34, se muestra la siguiente hoja del *dashboard*. En este caso también han surgido problemas con la ordenación de los gráficos, en particular con el último de ellos, que representa el número de accidentes por grupo de edad. Al ordenar de forma numérica este gráfico, no ha sabido interpretar bien los valores que empiezan por 0, como el ‘06’ por ejemplo, y no los ha colocado al principio de la gráfica.

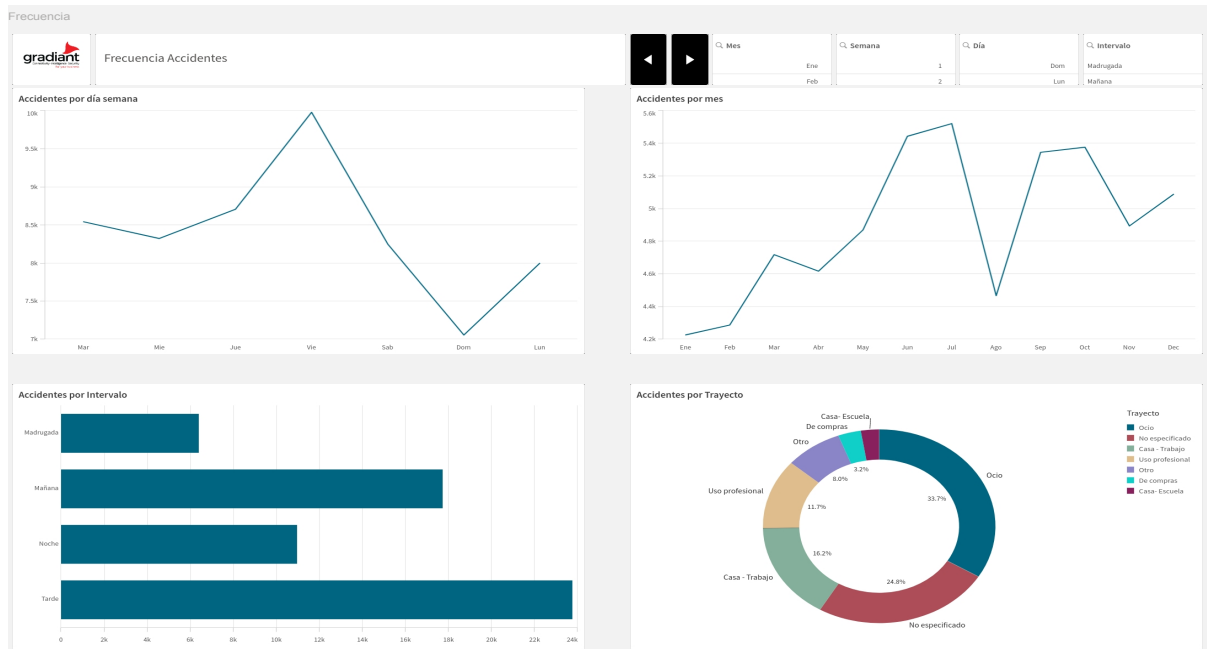


Figura 2.33: Hoja 2 Dashboard Qlik Sense

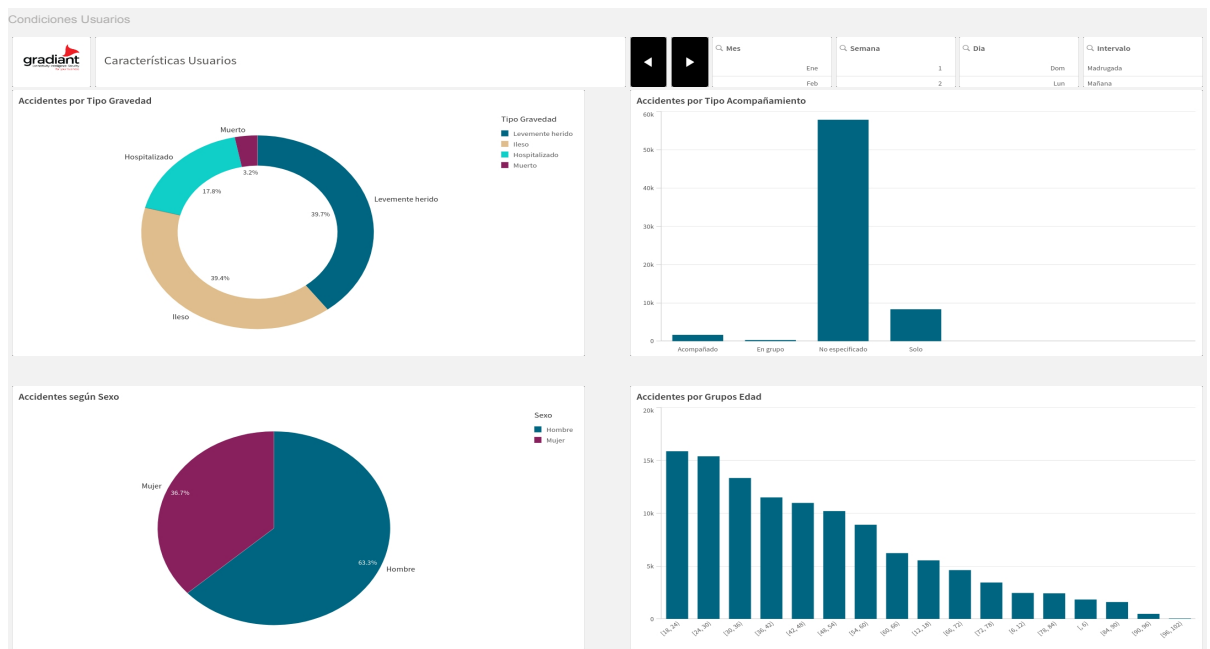


Figura 2.34: Hoja 3 Dashboard Qlik Sense

Finalmente, en las dos últimas hojas del *dashboard*, dadas por las Figuras 2.34 y 2.35,

se han podido representar todos los gráficos de manera correcta.

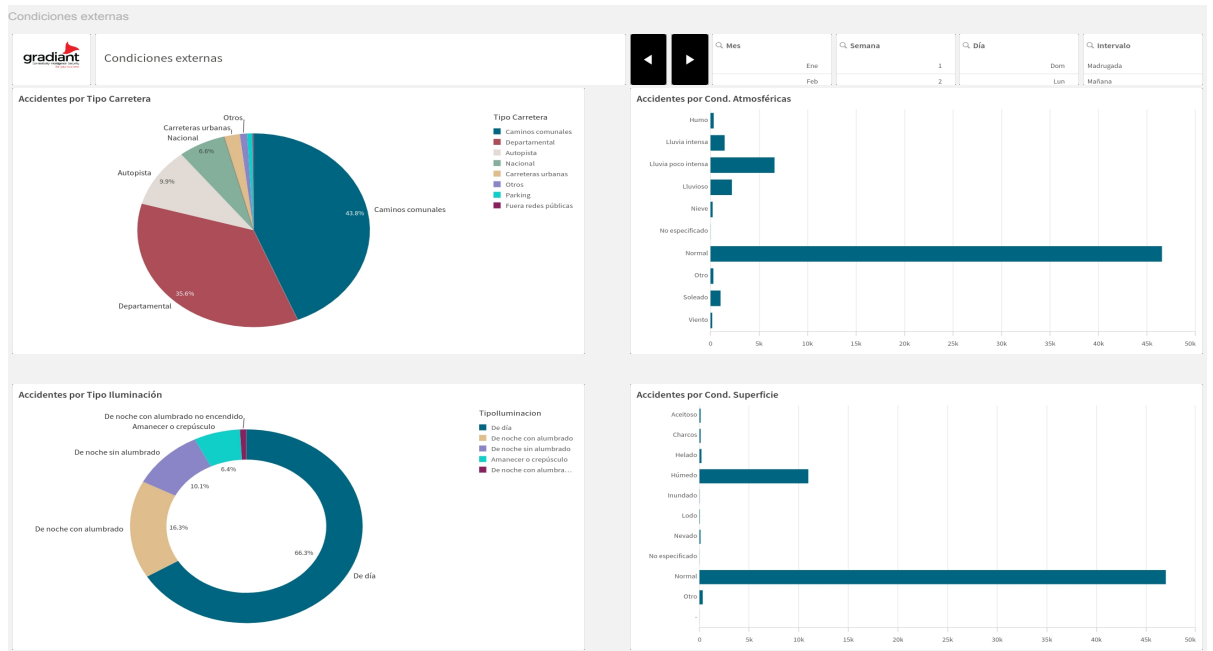


Figura 2.35: Hoja 4 Dashboard Qlik Sense

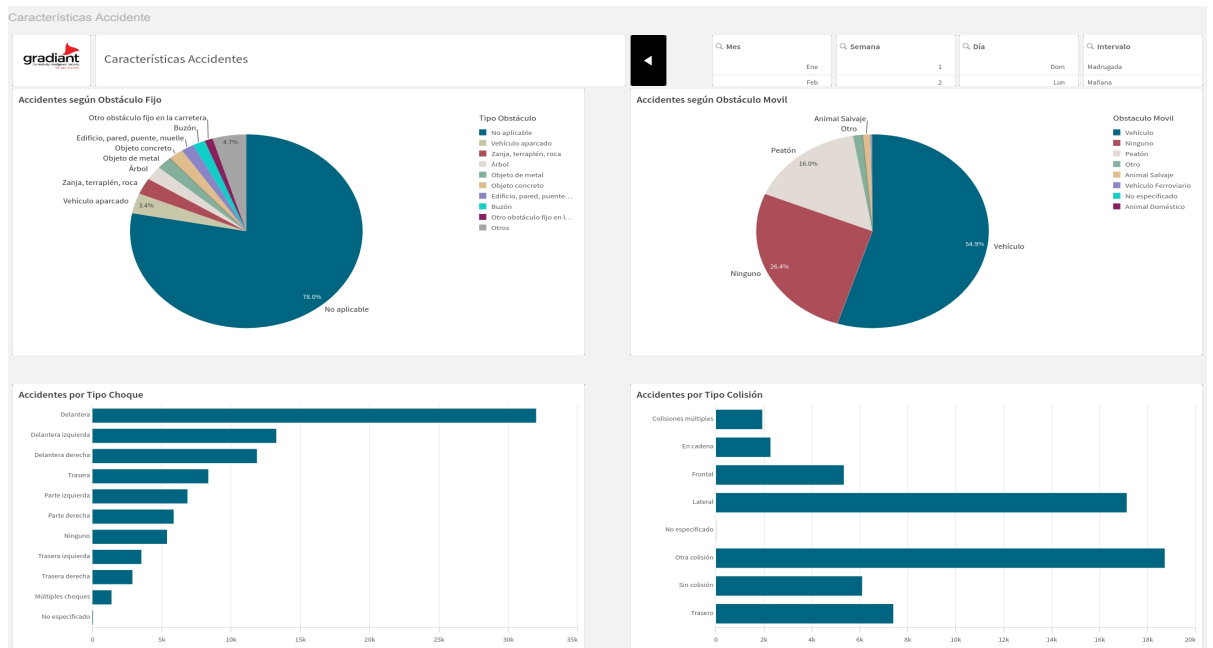


Figura 2.36: Hoja 5 Dashboard Qlik Sense

Como conclusión, visualmente es más atractiva la herramienta Power BI, puesto que los gráficos se integran mejor en la hoja y los filtros son mucho más cómodos. Además, también resulta más cómoda la forma de arrastrar los gráficos y colocarlos en la herramienta de Power BI, puesto que Qlik Sense tiene unos espacios predefinidos donde colocar los gráficos, de forma que no permite superponer un gráfico con otro, lo cual puede resultar útil a la hora de modificar el *dashboard*. En la Figura 2.37 se muestra esta situación.

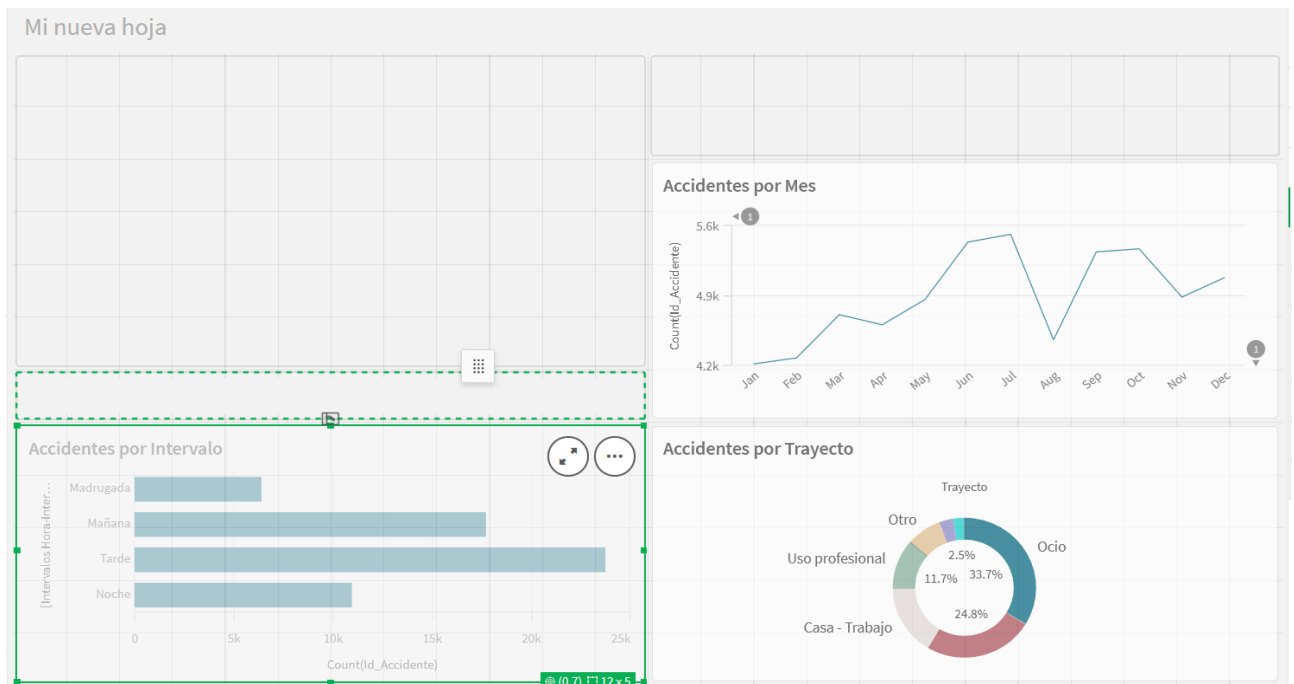


Figura 2.37: Colocación de gráficos Qlik Sense

Al intentar arrastrar el gráfico de Accidentes por Intervalo, nos da varias opciones, que se muestran como recuadros. Si pasamos por encima de alguno de ellos se muestra el borde del recuadro en verde punteado.

Además, resulta complicado y poco intuitivo establecer el orden de los gráficos. Como punto a favor tiene una gran diversidad de objetos visuales.

A continuación se mostrarán los *dashboards* obtenidos en Tableau. En esta herramienta la creación de *dashboards* y gráficos es distinta a las anteriores, puesto que las visualizaciones se crean en hojas de trabajo, y en los *dashboards*, en vez de añadir visualizaciones, se añaden hojas de trabajo. Este método puede ser un inconveniente, puesto que si se tienen tantas hojas como visualizaciones puede resultar difícil encontrar el gráfico que se necesite. Por ejemplo, en este caso se tienen 5 hojas en el informe y se han necesitado 30 hojas de trabajo diferentes.

En la Figura 2.38 se puede ver el primer *dashboard* creado con esta herramienta. Como se puede observar, no se han podido replicar todos los gráficos, en particular no se han podido crear los KPI con el número de accidentes por conductor, peatón y pasajero. Seguramente este fallo se deba a un problema de comprensión del lenguaje, que puede ser solucionado. Otro de los detalles en los que debemos fijarnos es que el nombre de los meses del primer gráfico está en inglés en vez de español, a pesar de que la herramienta fue configurada en español. Además, si nos fijamos en el último gráfico, vemos que el número de accidentes diarios es acumulado, es decir, no representa todos los días del año, sino que hace el recuento de accidentes por el número del día. El resto de gráficos se han realizado

sin problemas y, además, los filtros se muestran como menús desplegables, lo que mejora la experiencia de usuario.

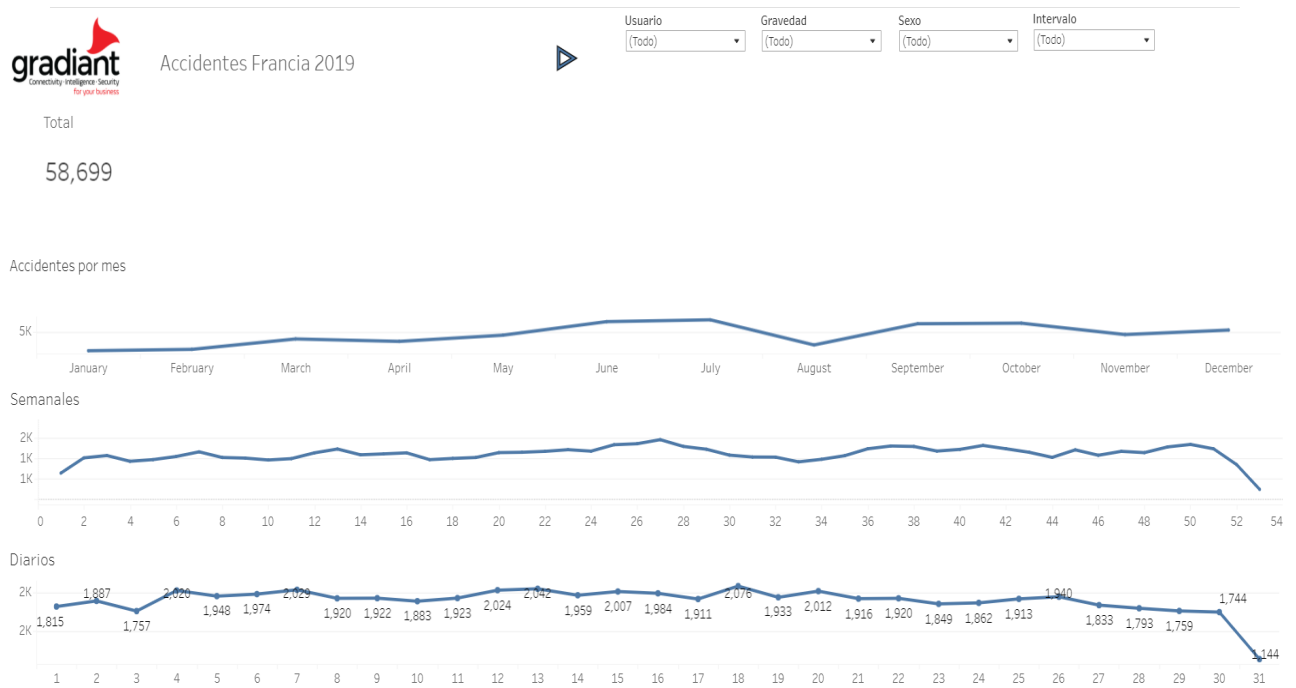


Figura 2.38: Hoja 1 Dashboard Tableau

La siguiente hoja del *dashboard* se muestra en la Figura 2.39, que como podemos ver contiene todos los gráficos esperados. Al igual que en la primera hoja, los nombres de los días de la semana y los meses están en inglés en vez de español. Otro de los problemas que tuvimos fue que no existen los gráficos de anillos de forma predeterminada, tan sólo existen los circulares. Como podemos ver en la siguiente hoja del *dashboard*, finalmente conseguimos de forma costosa replicar este gráfico. Como puntos a favor, cabe destacar que no hemos tenido problemas de ordenación de los gráficos; para las visualizaciones de días de accidentes por día de la semana y mes ha entendido bien la ordenación, y para la visualización de accidentes por intervalo horario basta hacer intercambios entre las barras del diagrama para obtener el resultado deseado.

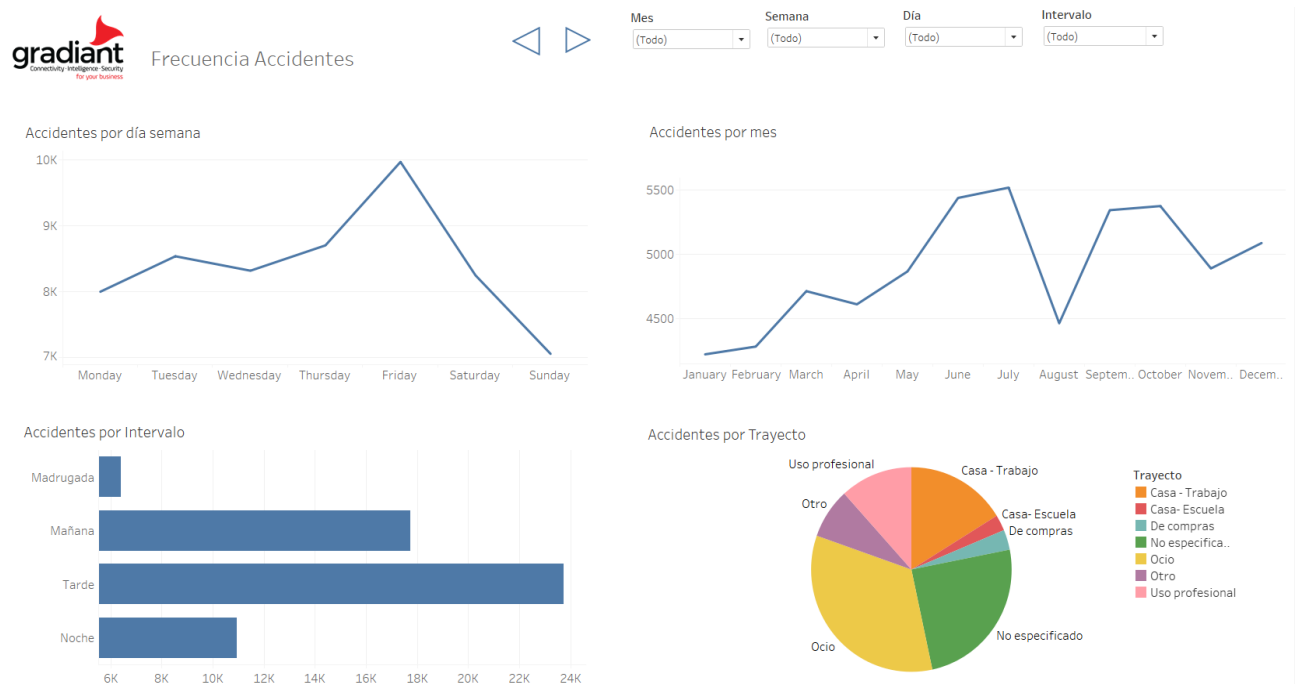


Figura 2.39: Hoja 2 Dashboard Tableau

En la Figura 2.40 mostramos la tercera hoja del *dashboard*. En este caso hemos obtenido todos los gráficos deseados. Tal y como comentamos anteriormente, hemos conseguido realizar un gráfico de anillo, aunque, como se puede observar, se encuentra rodeado por un borde gris que no hemos podido eliminar.

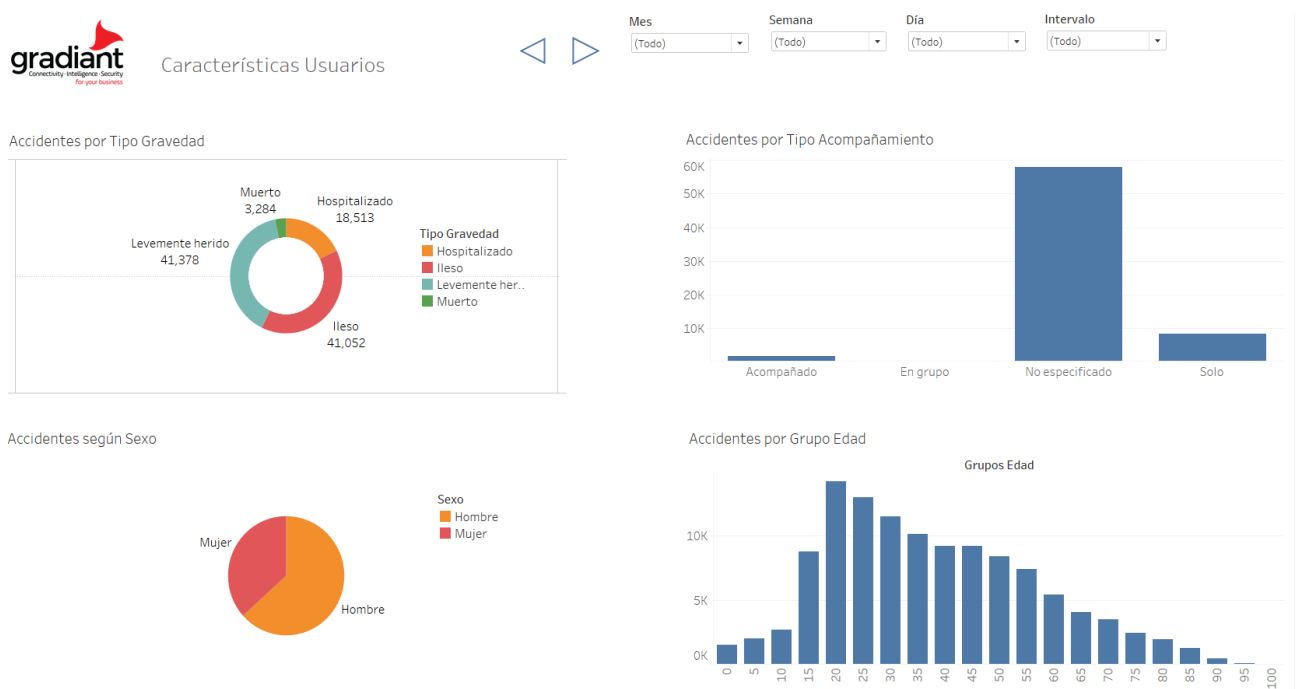


Figura 2.40: Hoja 3 Dashboard Tableau

Las dos últimas hojas del *dashboard* se muestran en la Figura 2.41 y la Figura 2.42,

respectivamente. Como comentarios de ambas hojas podemos ver que, a diferencia de las anteriores herramientas, cuando la visualización es demasiado grande aparece una barra de desplazamiento. Como aspecto negativo podemos decir que, cuando tenemos demasiadas categorías, las leyendas se vuelven demasiado grandes, por lo que sería más adecuado poner en las leyendas las categorías mayoritarias.

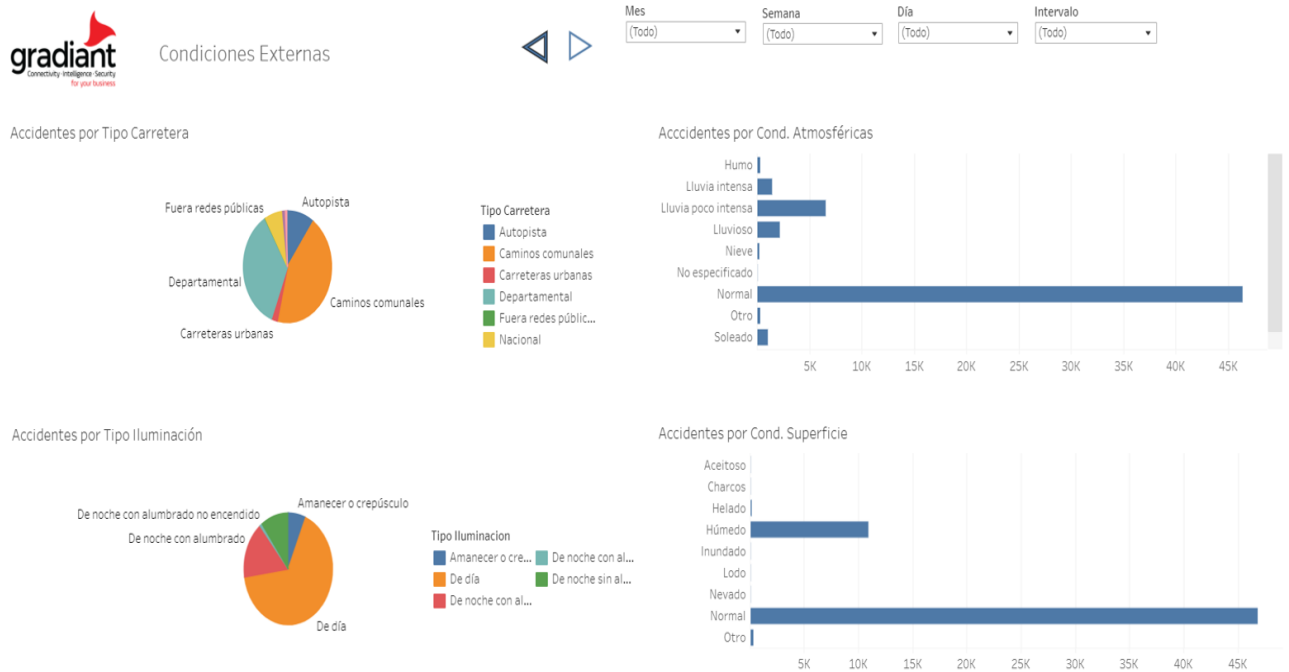


Figura 2.41: Hoja 4 Dashboard Tableau

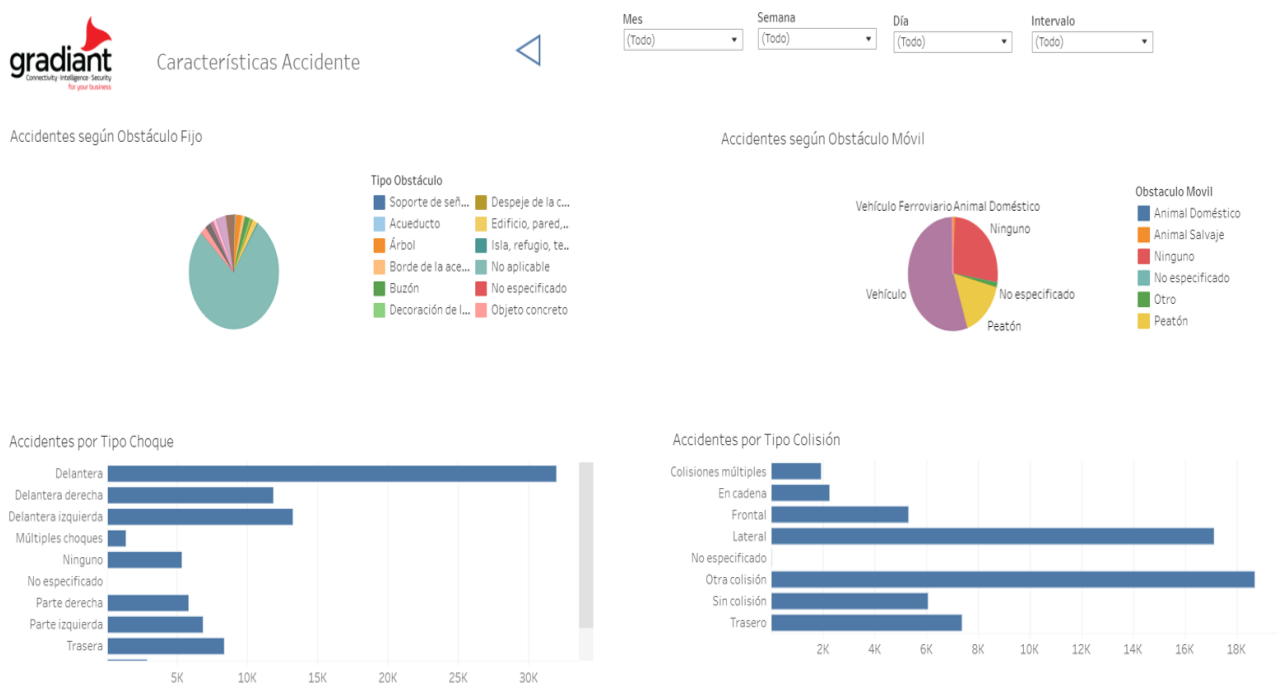


Figura 2.42: Hoja 5 Dashboard Tableau

Además, otra de las desventajas que hemos visto de la herramienta es que para poder filtrar los gráficos es necesario establecer todos los filtros para todos los gráficos, puesto que añadir los filtros en el *dashboard* no implica que las visualizaciones se filtren.

Como conclusión, podemos decir que hemos conseguido la mayor parte de las visualizaciones deseadas, aunque para algunas han surgido complicaciones debido a la dificultad en su usabilidad o una falta de comprensión del lenguaje de Tableau. La ordenación de las visualizaciones no ha resultado un problema, aunque sí el filtrado y el idioma de las fechas (meses y días de la semana).

A continuación, mostraremos el resultado final del informe con la herramienta Metabase. En primer lugar, en la Figura 2.43 se observa la primera hoja del dashboard.

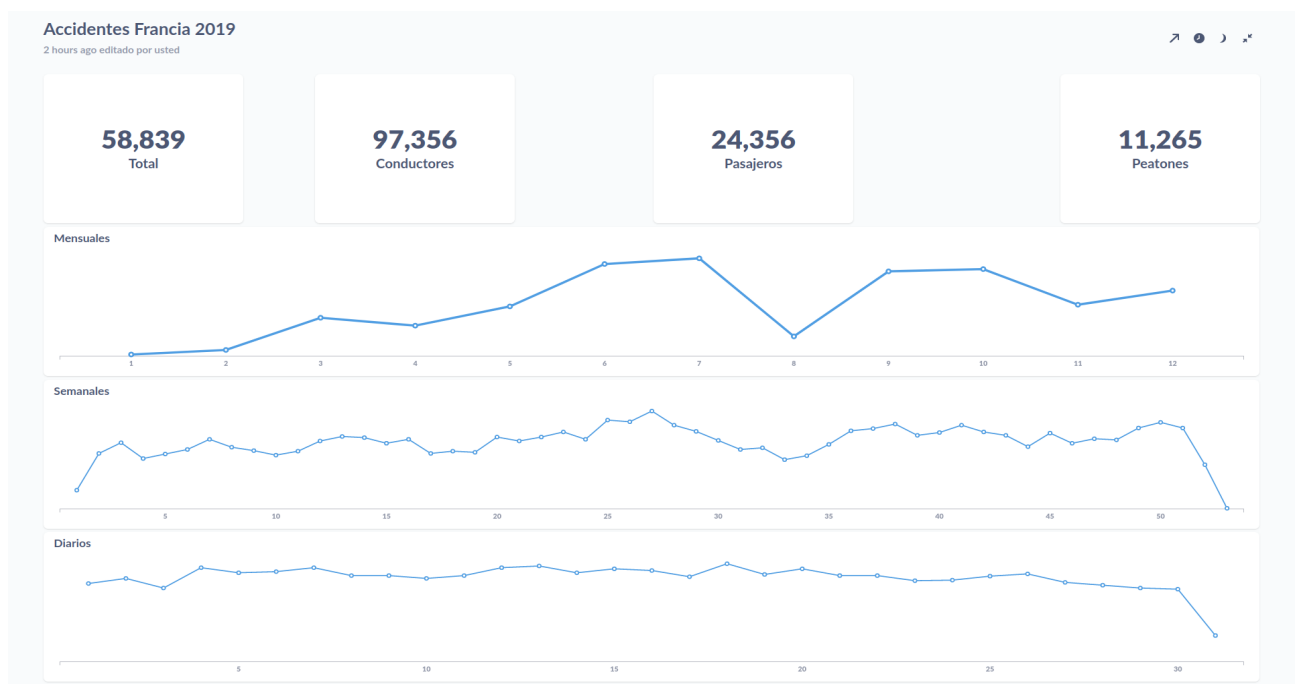


Figura 2.43: Hoja 1 Dashboard Metabase

En esta hoja hemos conseguido replicar todos los gráficos realizados con Power BI, aunque encontramos varias discrepancias. En primer lugar los KPI coinciden para el número de accidentes total, conductores y peatones, pero no para el número de accidentes por pasajeros. No sabemos explicar el motivo de esta diferencia. Para el gráfico de accidentes mensuales no hemos utilizado los nombres de los meses, sino el número del mes. Por último, al igual que en Tableau, el número de accidentes diarios es acumulado, es decir, no representa el número de accidentes por cada día del año. Recordemos que este tipo de gráficos se realiza utilizando lenguaje SQL, por lo que seguramente pueda solucionarse este error con un mayor conocimiento de este lenguaje. Por otra parte, uno de los mayores problemas con esta herramienta es la creación de filtros. En realidad, añadir filtros es una tarea sencilla, en el sentido de que una vez creado el *dashboard* existe una opción de añadir filtros. Aquí se puede elegir entre las opciones de tiempo, localización, ID y otras categorías. Dentro de los filtros de tiempo podemos escoger mes y año, trimestre y año, fecha única, rango de fechas, fecha relativa o todas las opciones. En los filtros de localización podemos escoger ciudad, estado, ZIP o código postal y país. El filtro de ID proporciona un cuadro de entrada donde se puede escribir el ID de un usuario, pedido, etc. Por último, el filtro de otras categorías es un tipo de filtro flexible que permite crear

un menú desplegable o un cuadro de entrada para filtrar cualquier campo de categorías en sus tarjetas. Sin embargo, en la práctica, no conseguimos establecer el filtro que deseamos cuando está en la categoría de otros filtros.

A continuación, en la Figura 2.44 se muestra la segunda hoja del *dashboard*.

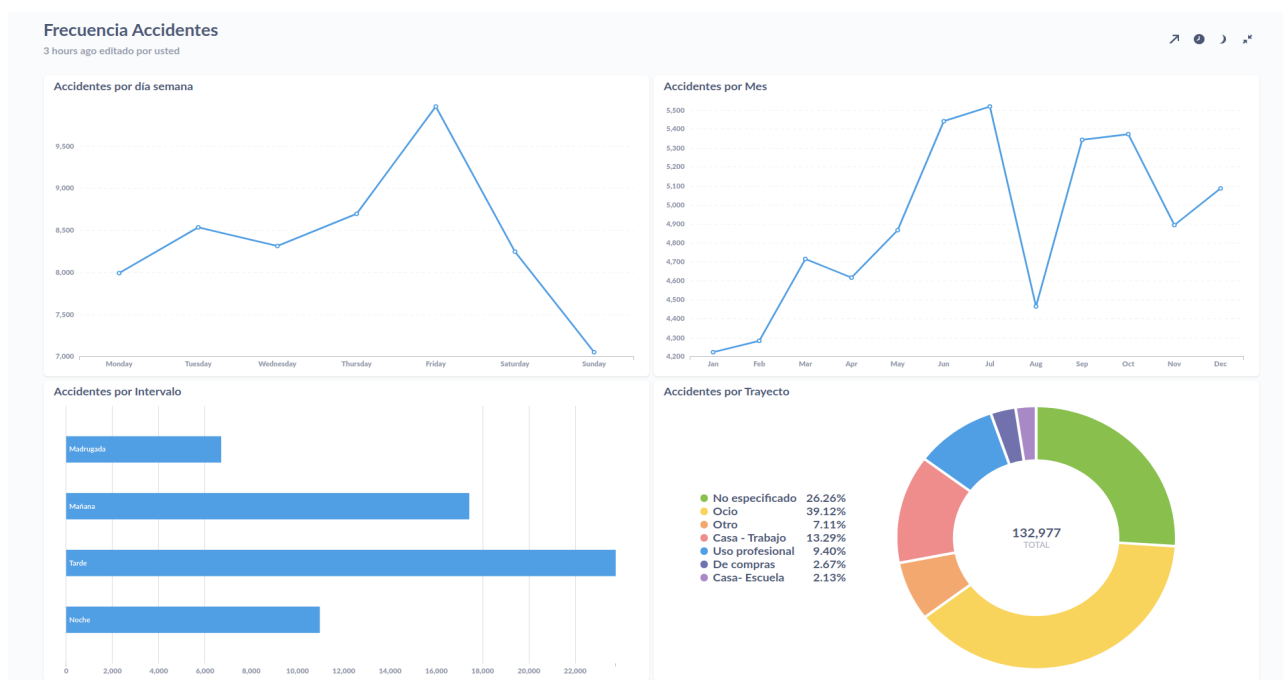


Figura 2.44: Hoja 2 Dashboard Metabase

En este caso tampoco hemos tenido problemas en replicar ninguno de los gráficos. Tanto en el primero como en el segundo gráfico, los nombres de los días de la semana y del mes están en inglés, a pesar de que el idioma de la herramienta estaba configurado en español. No hemos tenido problemas en ordenar los gráficos de manera personalizada. Por último, el gráfico de anillo tiene un formato un poco distinto al resto, en el centro del gráfico se muestra el total de accidentes por trayecto y en la leyenda se desglosa cada categoría junto con un porcentaje del total.

En la Figura 2.45 se muestra la tercera hoja del informe. Al igual que en las hojas anteriores hemos podido replicar todos los gráficos, a excepción de los filtros de página. Nótese también que el gráfico que representa el número de accidentes según el sexo no es un diagrama circular, sino de anillo. Esto se debe a que dentro de las posibles visualizaciones no existe el diagrama circular.

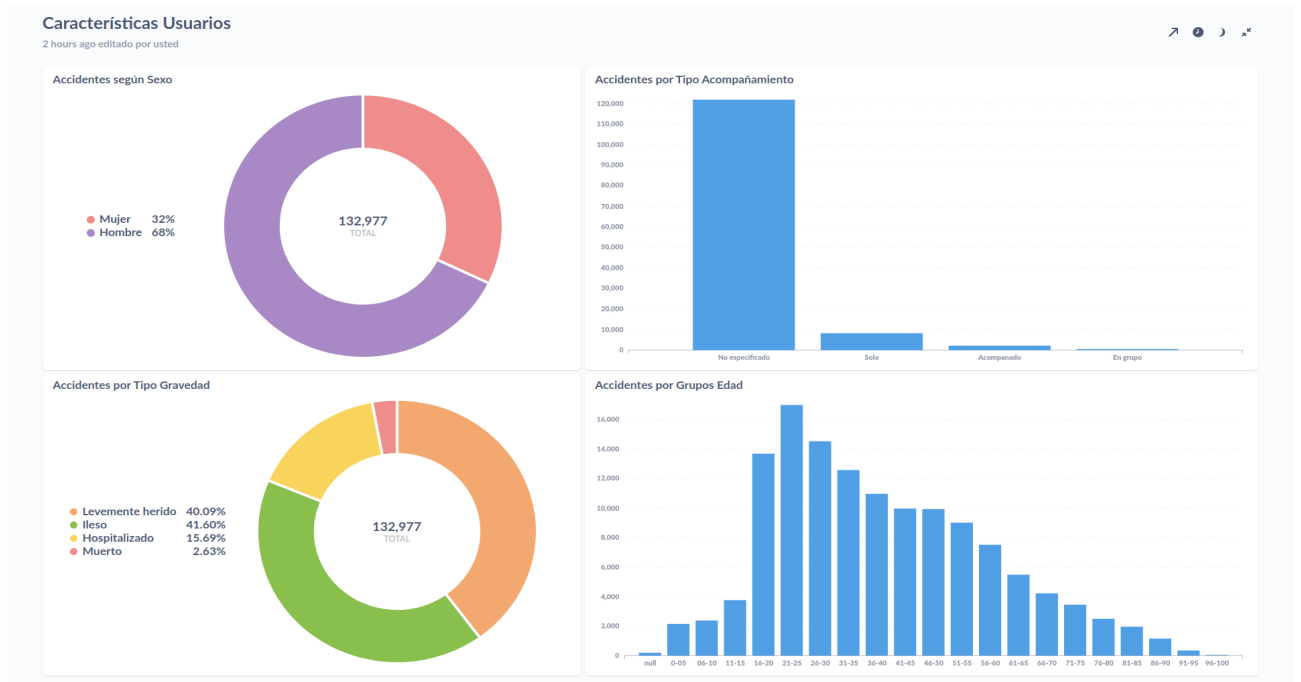


Figura 2.45: Hoja 3 Dashboard Metabase

En las Figuras 2.46 y 2.47 se muestran las dos últimas hojas del informe. Para ambos casos no tenemos ningún comentario a mayores acerca los gráficos utilizados.

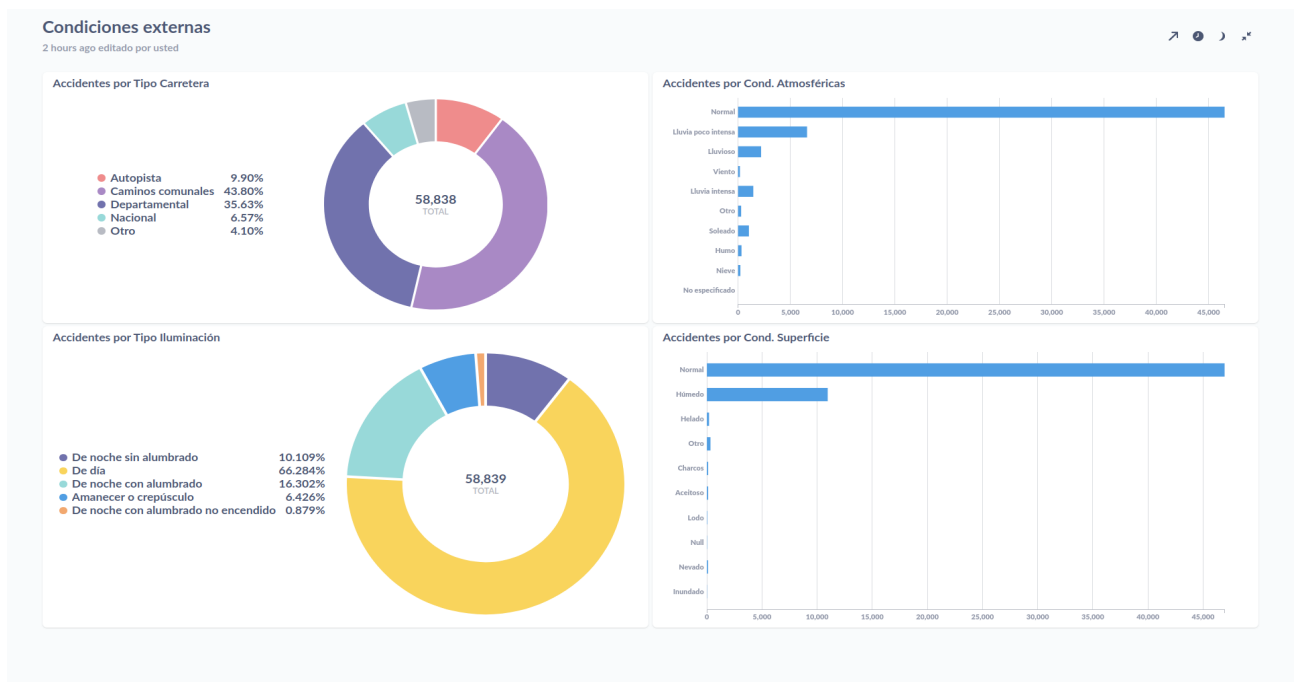


Figura 2.46: Hoja 4 Dashboard Metabase

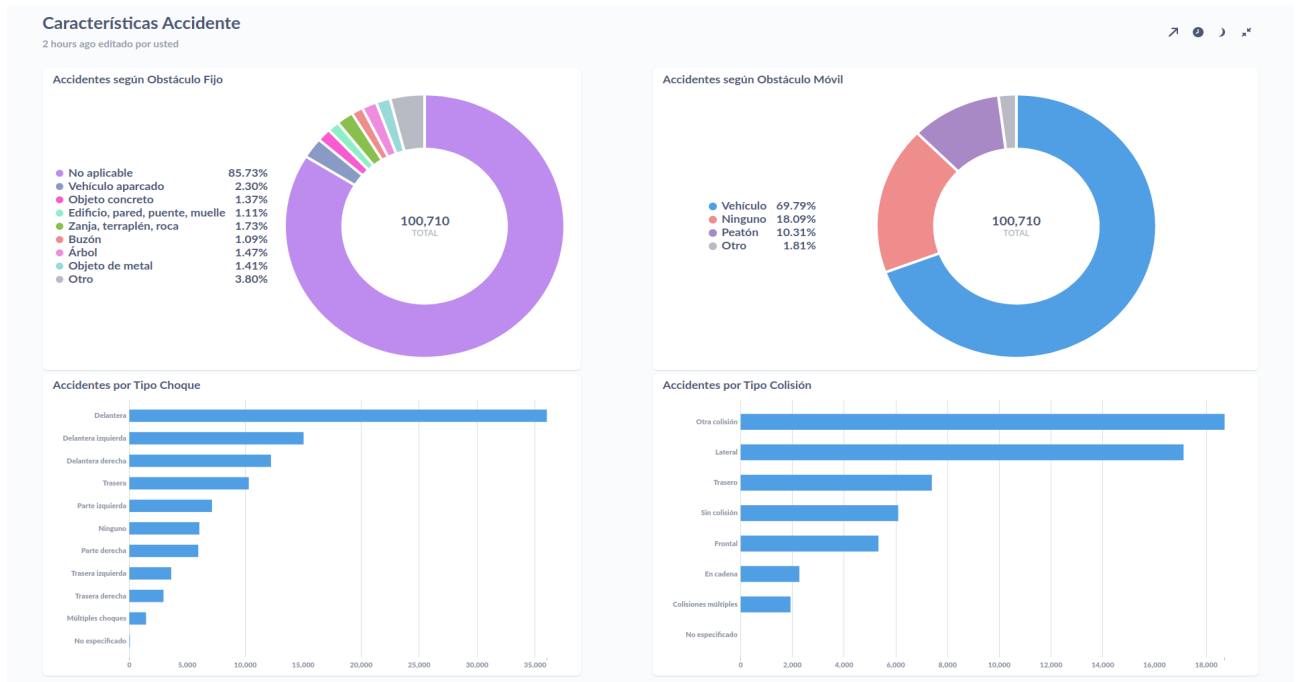


Figura 2.47: Hoja 5 Dashboard Metabase

Metabase incorpora en todos sus dashboard la opción de visión nocturna, es decir, utiliza un fondo oscuro en vez de claro y el color de los caracteres pasa de negro a blanco. Podemos ver un ejemplo en la Figura 2.48.

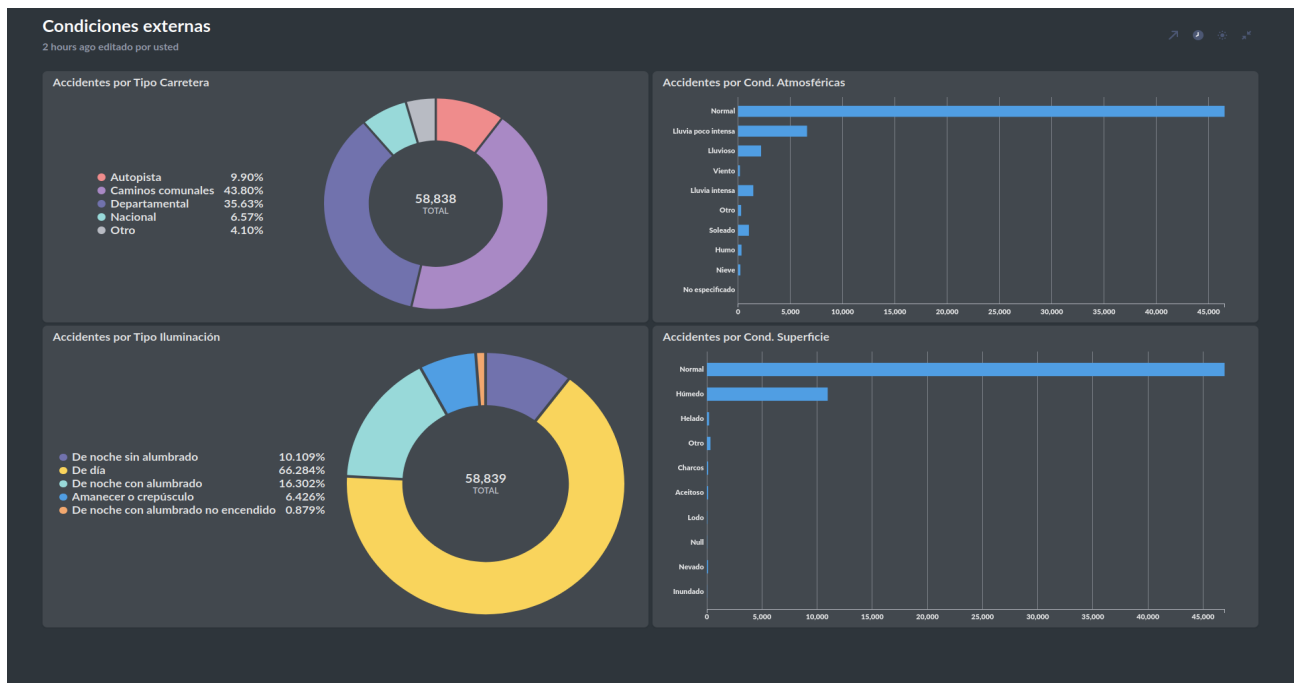


Figura 2.48: Visión nocturna Dashboard Metabase

Como conclusión una vez realizado el análisis, Metabase es una herramienta visualmente atractiva y sencilla de utilizar debido al uso del lenguaje SQL. Por otra parte, las

opciones de visualización son más limitadas que en otras herramientas y los filtros más complicados de utilizar.

A continuación mostraremos los *dashboards* obtenidos con la herramienta Redash. En la Figura 2.49 se muestra la primera hoja del *dashboard*.

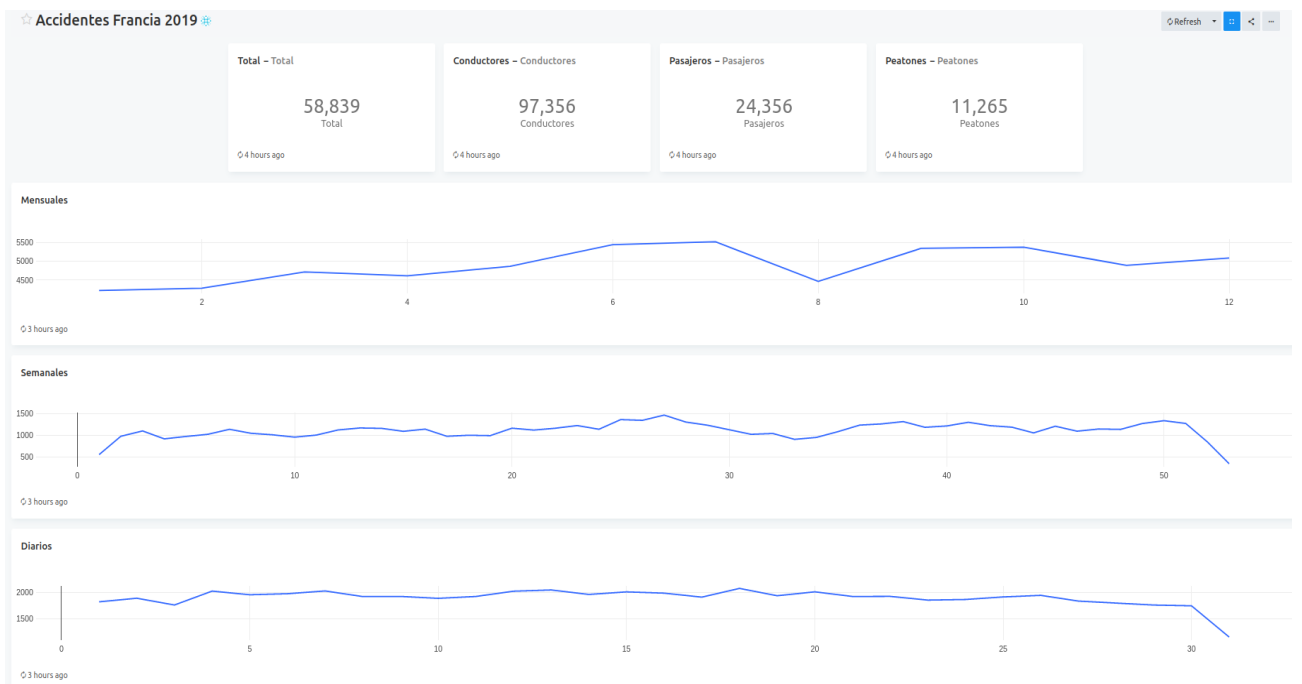


Figura 2.49: Hoja 1 Dashboard Redash

Como podemos observar, hemos podido replicar todas las visualizaciones. En primer lugar cabe destacar que en los KPI, los accidentes totales, por conductor, pasajero y peatón coinciden con las cifras de la herramienta Metabase. En el gráfico de líneas que representa el número de accidentes diarios ocurre lo mismo que con Tableau y Metabase, se representa el número de accidentes diarios acumulado, no el real. Además, en este caso tampoco hemos podido establecer filtros de página. Por último, una novedad con respecto a las demás herramientas es que en cada visualización, en la esquina inferior izquierda aparece el tiempo que pasó desde la última vez que se actualizó el *dashboard*.

A continuación, en la Figura 2.50 se muestra la segunda hoja del *dashboard*.

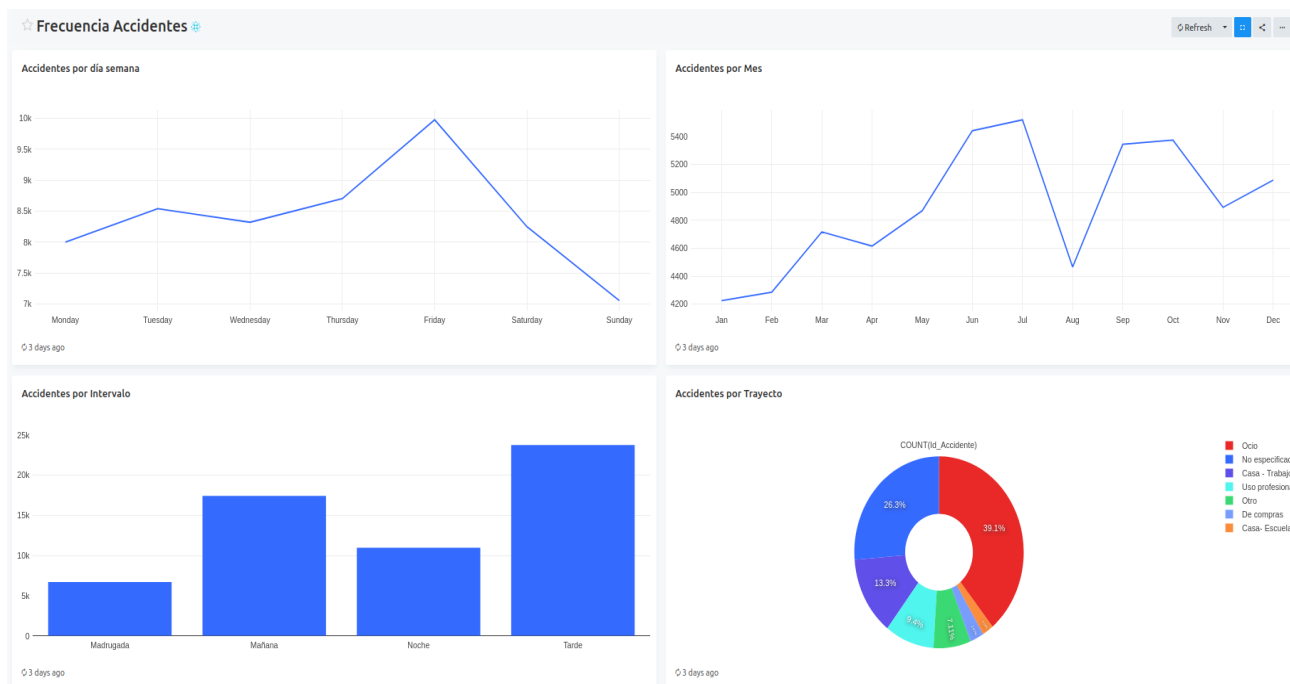


Figura 2.50: Hoja 2 Dashboard Redash

Como podemos observar, en este caso también hemos podido replicar todas las visualizaciones. En los dos primeros gráficos de líneas se observa que los días de la semana y meses del año se encuentran en inglés, sin embargo, esto se debe a que la herramienta por defecto se encontraba en este idioma. Además, en ambos casos no hemos tenido problemas con la ordenación de las fechas. En el tercer gráfico, que muestra los accidentes por intervalo horario, sí hemos tenido problemas de ordenación, puesto que la categoría ‘Noche’ precede a la ‘Tarde’. Por último, no hemos tenido problemas a la hora de realizar el gráfico de anillo, aunque visualmente es menos atractivo y los porcentajes no se encuentran en horizontal, por lo que dificulta la lectura.

En las siguientes hojas del *dashboard*, dadas por las Figuras 2.51, 2.52 y 2.53, las visualizaciones se han realizado correctamente y con el orden deseado. Además, nótese que para aquellos gráficos que ocupen mucho espacio o la leyenda sea muy grande existe una barra de desplazamiento para poder desplazarse y ver todo el gráfico. Por último comentar que en la última hoja, los diagramas de barras se muestran de forma vertical y no horizontal, pues no existe dicha opción.

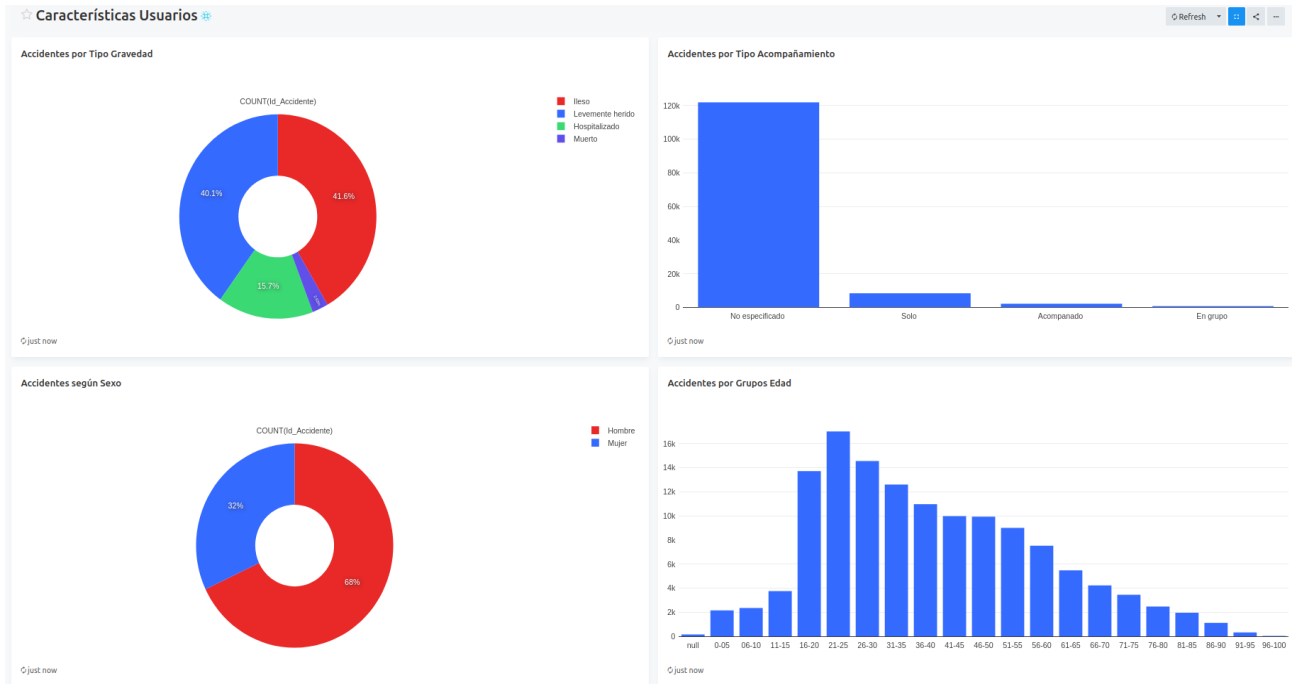


Figura 2.51: Hoja 3 Dashboard Redash

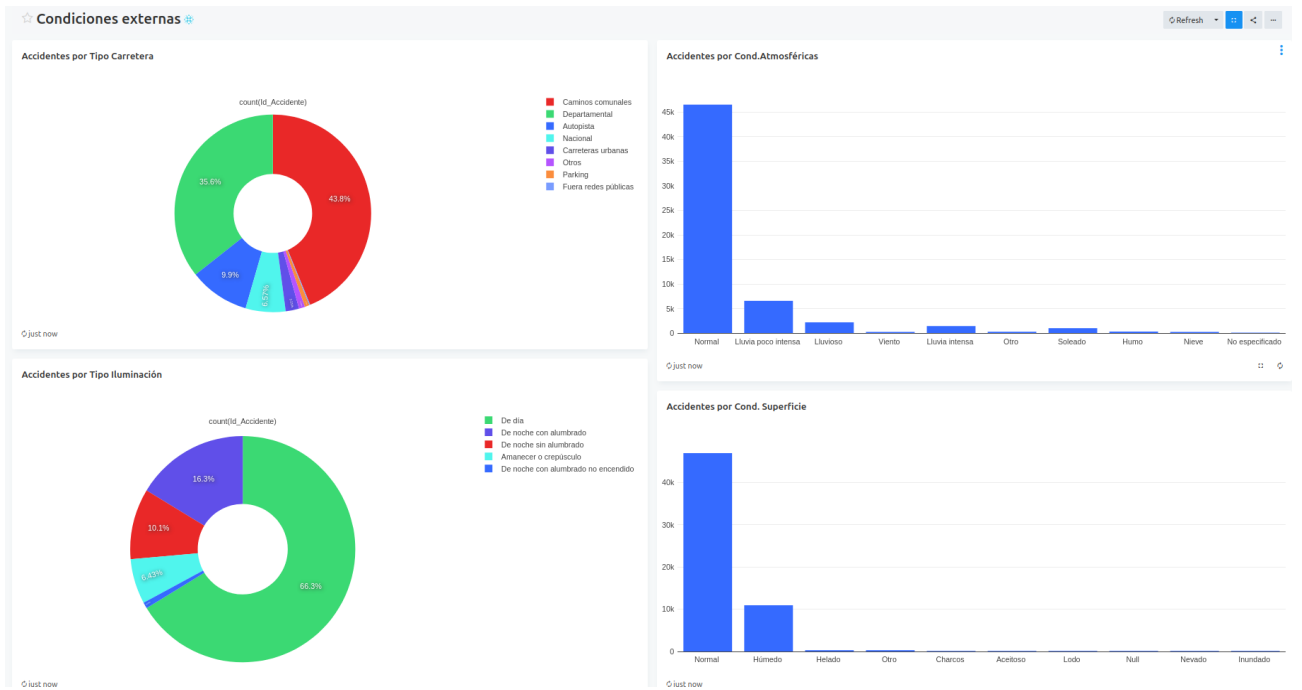


Figura 2.52: Hoja 4 Dashboard Redash

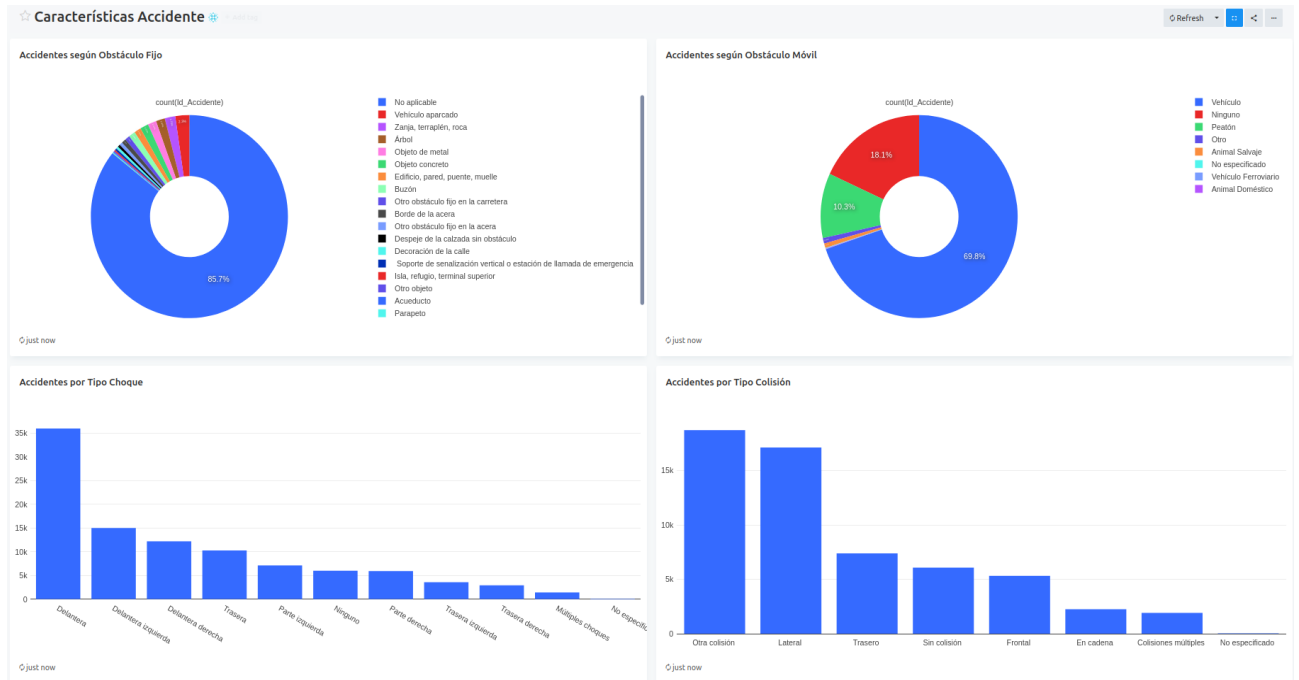


Figura 2.53: Hoja 5 Dashboard Redash

Como conclusión, podemos afirmar que Redash es una herramienta sencilla de utilizar aunque con ciertas limitaciones, como la utilización de filtros y la ordenación personalizada. Visualmente, los gráficos podrían ser mejorables, en el sentido de que los porcentajes de las categorías en los gráficos de anillo podrían mostrarse en horizontal y las leyendas podrían ser menos amplias y contener únicamente las categorías más destacadas.

Por último, mostraremos la creación del informe con la herramienta Superset. En primer lugar, en la Figura 2.54 se muestra la primera hoja del *dashboard*. Con esta herramienta hemos podido replicar todas las visualizaciones. Por un lado, los KPI con el recuento del total de accidentes y del número de accidentes por conductores, pasajeros y peatones coincide con Power BI, Metabase y Redash. Por otra parte las visualizaciones de accidentes mensuales, anuales y diarios tienen las fechas en inglés, puesto que la herramienta por defecto está configurada en este idioma. Cabe destacar que en los tres casos se han realizado los gráficos correctamente, inclusive los accidentes diarios, que no se muestran acumulados sino de forma mensual. Por último, en esta herramienta sí se pueden aplicar filtros, aunque no se muestren dentro del informe. Cuando entramos en el *dashboard* aparece una columna de filtros, donde podemos añadir las columnas por las que nos interese filtrar y ahí es donde debemos seleccionarlos.

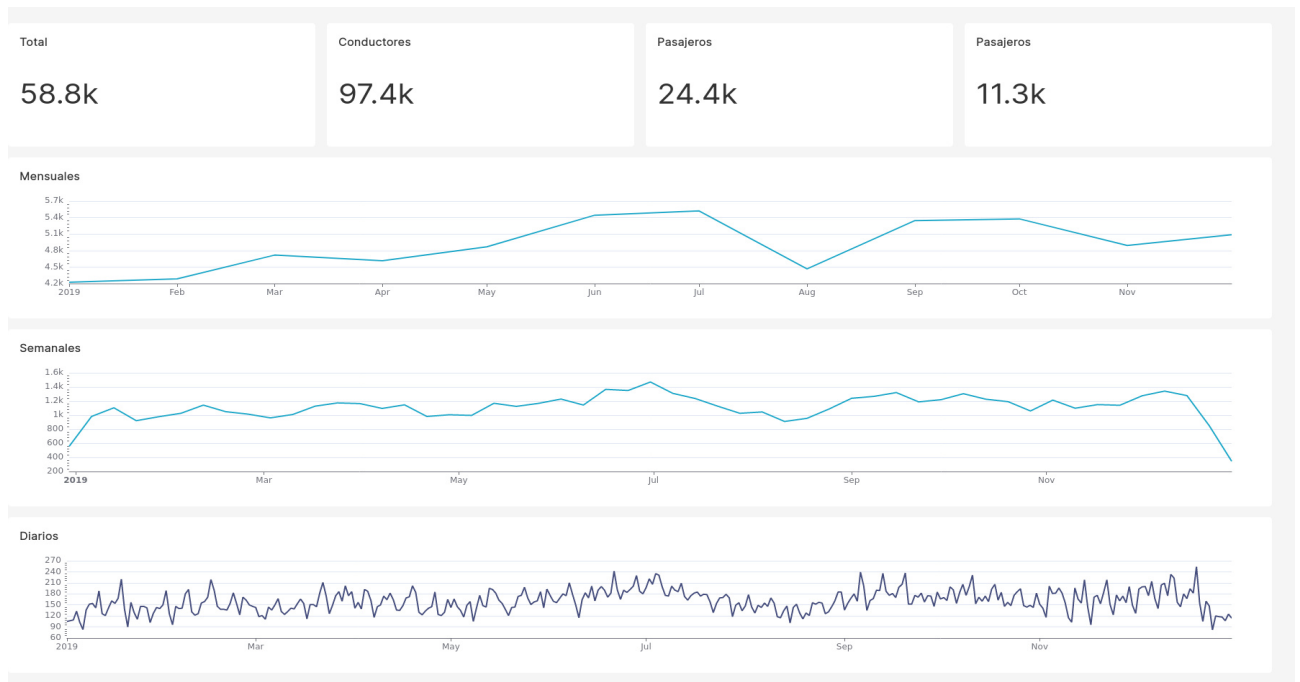


Figura 2.54: Hoja 1 Dashboard Superpet

A continuación, en la Figura 2.55 se muestra la segunda hoja del *dashboard*. Como podemos observar, en este caso no hemos podido replicar todos los casos. En particular no hemos podido representar el número de accidentes por día de la semana. Esto se debe a que en este tipo de gráficos, cuando se selecciona la granularidad temporal deseada no existe la opción de día de la semana, por lo que no podemos representarlos. En la visualización de los accidentes por intervalo horario observamos que las columnas no están ordenadas. Esto se debe a que esta herramienta tan sólo permite un orden ascendente o descendente, y no personalizado. Además, Superset no permite realizar gráficos de barra horizontales, tan sólo verticales. Por último, cuando creamos una visualización, los colores del gráfico no tienen por qué coincidir con los que se utilizan en el *dashboard*, por lo que podemos tener distintos gráficos con distintos colores, como pasa con la última visualización.

En la Figura 2.56 se muestra la tercera hoja del *dashboard*. En este caso tan sólo cabe comentar que en la última visualización, número de accidentes por grupo de edad, no hemos podido ordenar el gráfico de barras por intervalos de edad de forma ascendente. Esto se debe a que esta herramienta tan sólo permite ordenar por número de accidentes de forma ascendente o descendente.

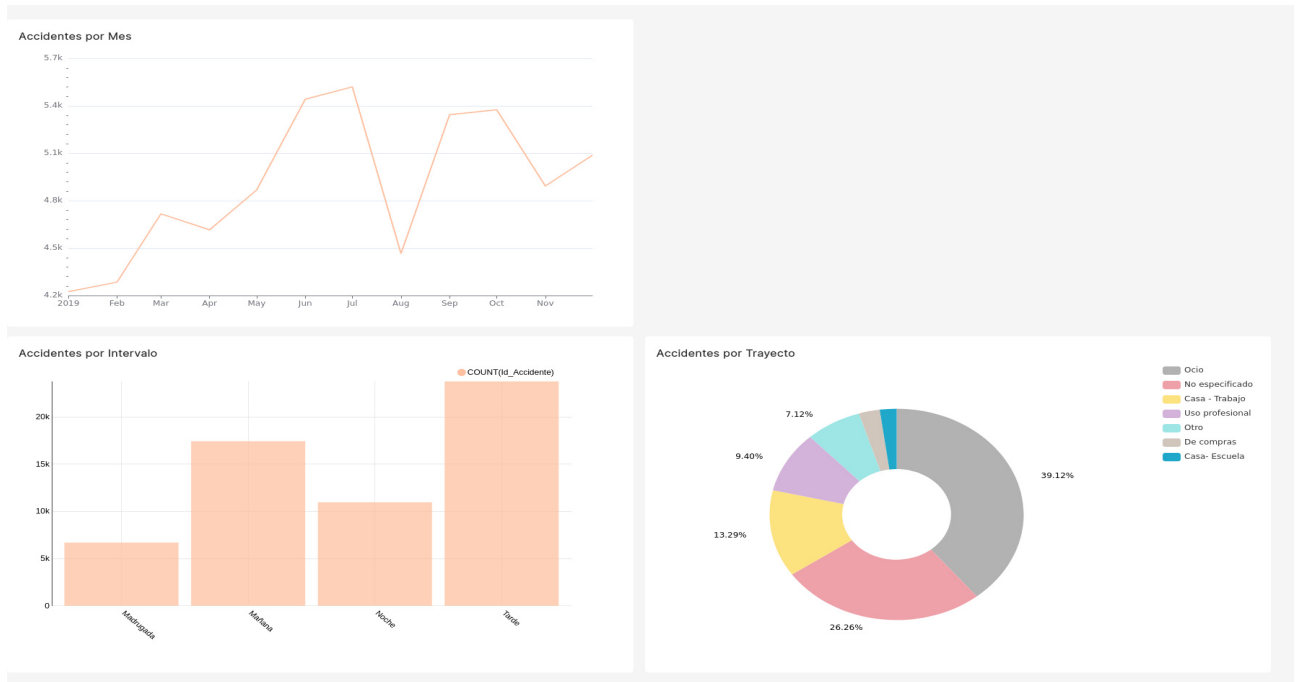


Figura 2.55: Hoja 2 Dashboard Superpet

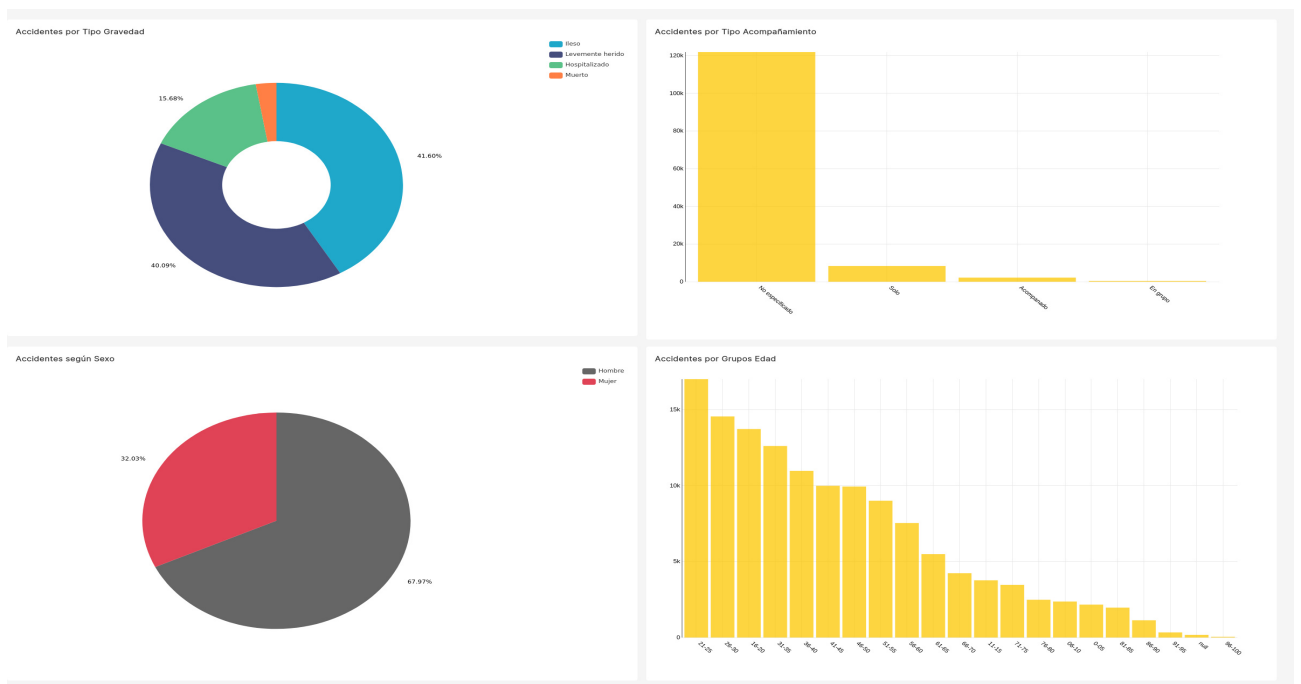


Figura 2.56: Hoja 3 Dashboard Superpet

Por último, en las Figuras 2.57 y 2.58 se muestran las dos últimas hojas del *dashboard*. No hemos tenido problemas para representar ninguna de las visualizaciones. Como defecto podemos decir que en la última hoja, en la visualización del número de accidentes por obstáculo fijo, la leyenda es demasiado grande y se superpone al propio gráfico.

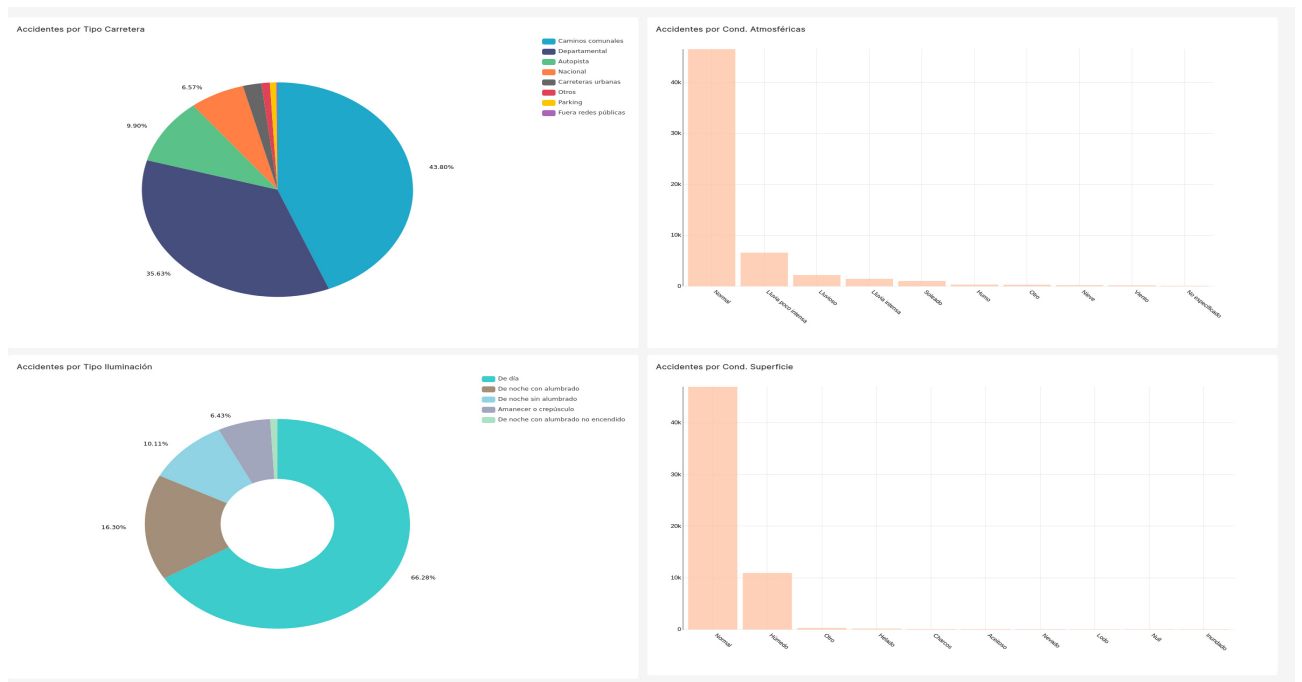


Figura 2.57: Hoja 4 Dashboard Superset

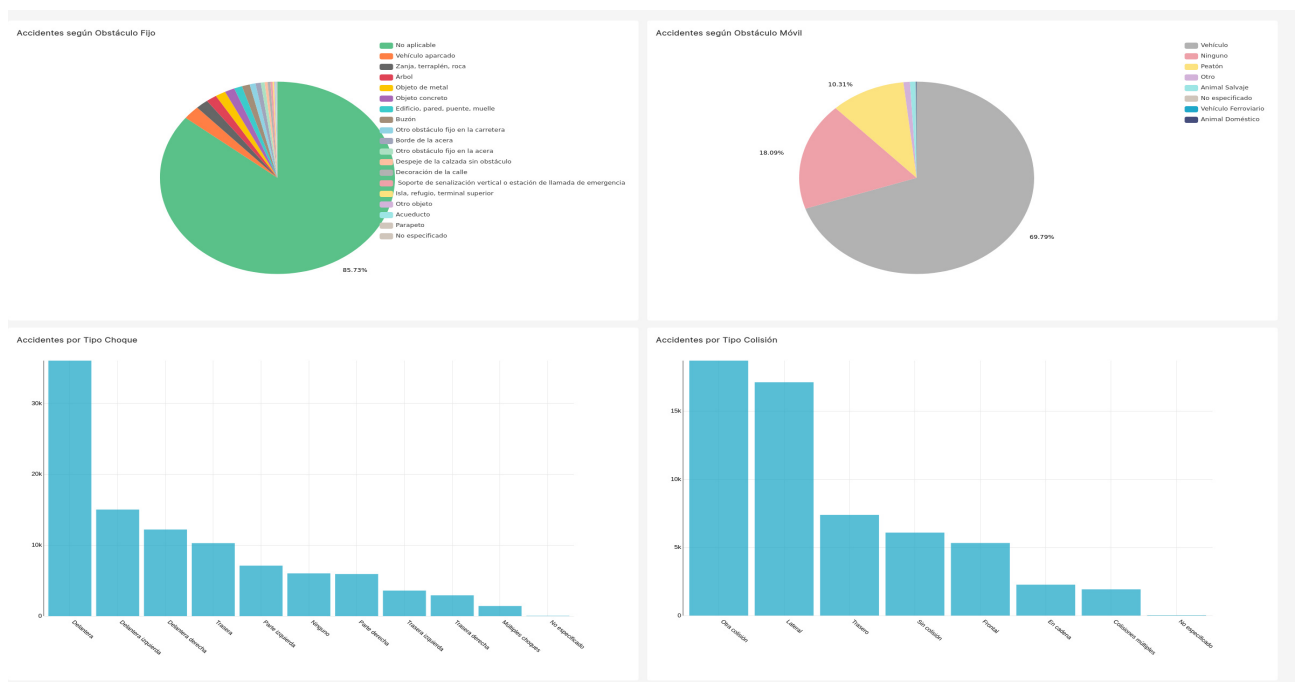


Figura 2.58: Hoja 5 Dashboard Superset

Como conclusión podemos afirmar que Superset es una herramienta menos intuitiva y manejable que otras de las opciones ya vistas. Como desventajas encontramos que es necesario crear un nuevo conjunto de datos cada vez que necesitemos columnas de distintas tablas, no permite un orden personalizado y no tiene la granularidad suficiente para la creación de ciertos tipos de gráficos. Como ventajas, podemos realizar filtros de forma

sencilla y posee una gran diversidad de visualizaciones.

2.7.1. Resumen

Por último haremos un resumen final de las herramientas en relación a los dashboards creados anteriormente. Como ya hemos visto, cada una de estas herramientas presenta sus ventajas y limitaciones. A continuación recogemos los puntos que consideramos más importantes a la hora de realizar los informes:

1. Sencillez: Es fundamental que la herramienta sea sencilla e intuitiva, tanto para principiantes como para usuarios con experiencia.
2. Visualizaciones: Se trata de herramientas de visualización, por lo que es necesario contar con gran variedad de gráficos para poder representar los datos.
3. Filtros: Cuando se trabaja con un dashboard interactivo resulta esencial la posibilidad de filtrar tanto los resultados como los gráficos, para así poder obtener más información acerca de los datos.

Con respecto a la sencillez, consideramos que las herramientas Power BI, Qlik Sense, Metabase y Redash son fáciles y sencillas de utilizar. Destacamos entre ellas las herramientas Metabase y Redash, puesto que utilizan lenguaje SQL para la creación de las visualizaciones, y no un lenguaje propio. Con respecto a las visualizaciones, todas las herramientas poseen una gran variedad de opciones. Aunque quizás la herramienta Metabase posea opciones más limitadas. Por último, con respecto al filtrado de las visualizaciones, es complicado su uso en la herramienta Metabase y en el caso de Redash no hemos conseguido aplicarlos.

Capítulo 3

Aplicaciones avanzadas en Power BI

Este capítulo se centrará en la construcción de aplicaciones avanzadas de *data analytics* sobre Power BI, donde se verá la integración de los lenguajes R y Python, creación de visualizaciones avanzadas y la herramienta de preguntas y respuestas de Power BI.

3.1. Integración de R y Python en Power BI

R es un lenguaje de programación diseñado para hacer análisis estadísticos y representaciones visuales. Este lenguaje es comúnmente utilizado por estadísticos, científicos de datos y analistas de datos. Dentro de Power BI puede ser utilizado para llevar a cabo las siguientes tareas:

1. Realizar tareas de limpieza de datos.
2. Modelado de datos avanzado.
3. Análisis de conjuntos de datos, donde se incluyen predicciones, agrupaciones en clusters, etc.
4. Preparar modelos de datos.
5. Crear informes.

En primer lugar veremos cómo se puede utilizar este lenguaje de programación para realizar la carga de los datos. Para ello, debemos ir a la pantalla principal de Power BI y en el menú hacer click en ‘Obtener datos’, del mismo modo que lo hicimos en el capítulo anterior. A continuación, debemos ir a la opción ‘Otras’, buscar ‘Script de R’ y conectar. Entonces se nos abrirá una ventana con un *script*, tal y como se muestra en la Figura 3.1. Para cargar los datos debemos utilizar el lenguaje R para leer el conjunto de datos, que en este ejemplo será un archivo csv.

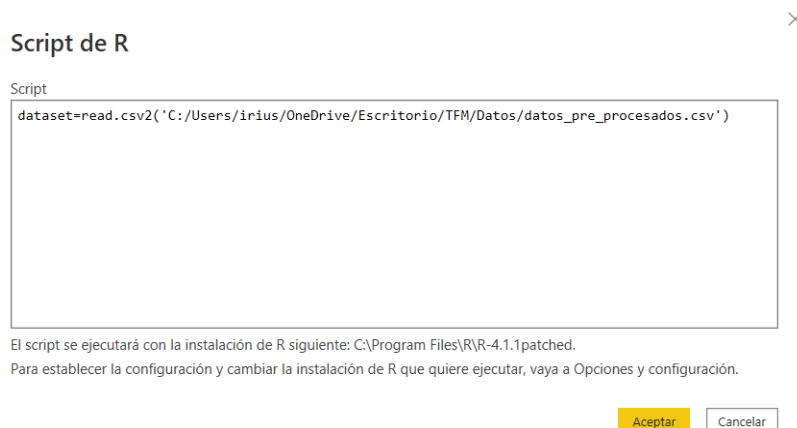


Figura 3.1: Cargar los datos con script de R

Una vez cargados los datos, puede ser de interés crear nuevas columnas. Esto también podemos hacerlo utilizando lenguaje R. En este caso, debemos ir al Editor de Power Query, y en la opción ‘Transformar’ seleccionamos ‘Ejecutar script de R’. Al igual que en el caso anterior, aparecerá una nueva ventana con un script R, como se muestra en la Figura 3.2.

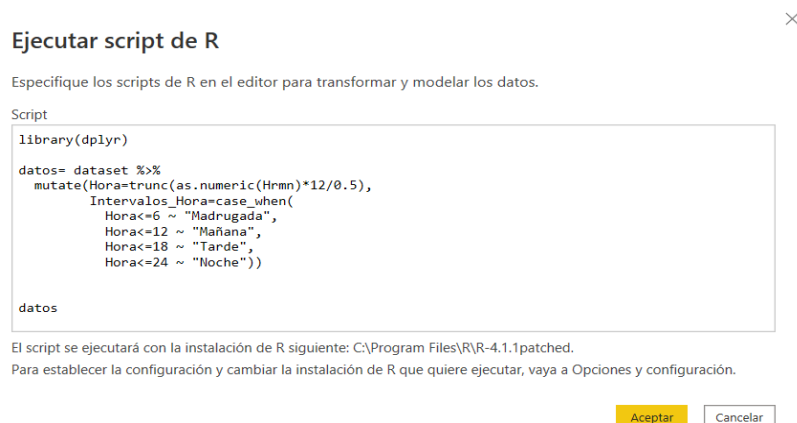


Figura 3.2: Creación de columnas con R

En este ejemplo hemos utilizado la librería dplyr [10] para crear la columna Intervalos_Hora, que en el capítulo anterior ya mostramos cómo construirla en Power BI.

Además, también podemos crear visualizaciones utilizando lenguaje R. En este caso debemos ir a la pantalla principal de Power BI y en la sección de visualizaciones aparece la opción ‘Objeto visual de script de R’. Al igual que con el resto de visualizaciones de Power BI, debemos arrastrar los campos que queramos usar para nuestra visualización, entonces aparecerá un editor de *script* R, como el que se muestra en la Figura 3.3. Aquí debemos introducir un *script* con lenguaje R que realice un gráfico con las columnas seleccionadas. En este ejemplo hemos utilizado la librería rpart [21] para crear un árbol de decisión.

```

Editor de script R
1 set.seed(60)
2 nobs<-nrow(dataset)
3 itrain<-sample(nobs,0.8*nobs)
4 train<-dataset[itrain,]
5 test<-dataset[-itrain,]
6
7
8 library(rpart)
9 tree<-rpart(Disease~.,data=train)
10
11 library(rpart.plot)
12
13 rpart.plot(tree, main="Clasificacion tree heart disease",
14            extra = 104, # show fitted class, probs, percentages
15            box.palette = "GnBu", # color scheme
16            branch.lty = 3, # dotted branch lines
17            shadow.col = "gray", # shadows under the node boxes
18            nn = TRUE)
    
```

Figura 3.3: Editor de script R para creación de visualizaciones

El resultado final de este código se muestra en la Figura 3.4.

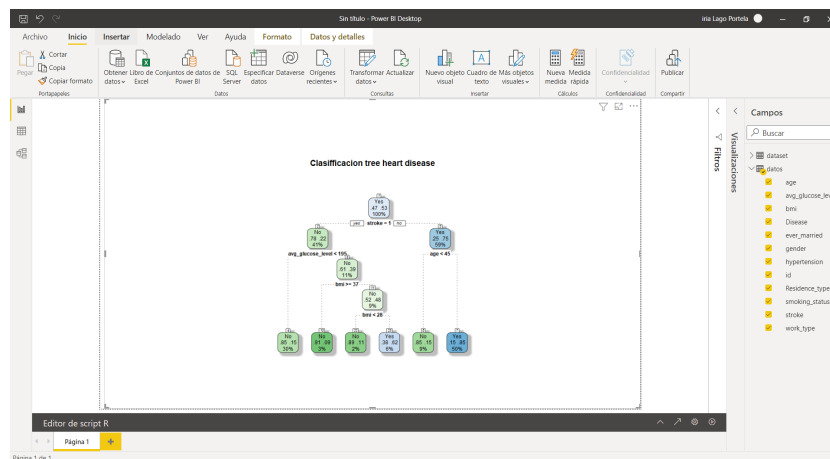


Figura 3.4: Visualización creada con script R

Python es otro lenguaje de programación no tan enfocado al análisis estadístico, sino en la programación en general, pudiendo ser utilizado para desarrollo web, análisis de datos, automatización de operaciones, etc. Además, Python utiliza una sintaxis lógica y sencilla que facilita su uso para los más principiantes. Su utilización en Power BI es prácticamente igual a la ya vista en R. En particular, para hacer la carga de datos debemos seguir las mismas instrucciones, solo que en vez de elegir ‘Script de R’ será ‘Script de Python’. Por otra parte, para crear nueva información deberemos seguir los pasos descritos anteriormente y seleccionar ‘Ejecutar script de Python’. Por último, para crear visualizaciones utilizando este lenguaje debemos ir a la sección de visualizaciones y elegir ‘Objeto visual de Python’.

3.2. Creación de visualizaciones interactivas con Power BI

Otra de las herramientas avanzadas es la utilización de visualizaciones interactivas y personalizadas con la librería Plotly [3]. Este tipo de gráficos amplían las capacidades de Power BI mediante la introducción de visualizaciones y características de visualización

que actualmente no se encuentran disponibles en Power BI. Sin embargo, este tipo de visualizaciones sólo se encuentran disponibles utilizando lenguaje R, a pesar de que Python también dispone de la librería Plotly.

Para poder utilizar este tipo de herramienta es necesario cumplir varios prerequisites:

1. Tener una cuenta de Power BI Pro o Premium por usuario.
2. Tener descargado Visual Studio Code u otro tipo de editor de código.
3. Tener la versión 4 o posterior de Windows PowerShell en el caso de Windows, o el Terminal para macOS.
4. Disponer de un entorno preparado para desarrollar visualizaciones en Power BI. ¹

En resumen, deberemos instalar Node.js ², un entorno en tiempo de ejecución de JavaScript que permite a desarrolladores ejecutar cualquier aplicación creada en JavaScript, y la herramienta pbiviz, que nos permite compilar el código fuente visual del paquete pbiviz. Además, para que un cliente (un ordenador) y un servicio (Power BI) puedan interactuar de manera segura se requiere de un certificado SSL, por lo que deberemos crearlo e instalarlo. A continuación debemos configurar Power BI para desarrollar visualizaciones y, por último, instalar las siguientes librerías: D3[6], TypeScript definitions [22], core-js [4] y powerbi-visual-api [15].

Llegados a este punto, ya podemos comenzar a crear gráficos interactivos. En primer lugar, debemos diseñar una visualización en R utilizando las librerías ggplot2 [11] y plotly. A continuación debemos ir a la Terminal de Visual Studio Code e introducir el comando ‘pbiviz new Nombre-t rhtml’. Esta acción creará en una carpeta llamada ‘Nombre’, una plantilla visual personalizada de R y los componentes necesarios para crear la visualización. Desde VSCode abriremos dicha carpeta y seleccionaremos el documento ‘script.r’. Aquí deberemos introducir nuestro código R de la visualización, y además reemplazar ‘dataset’ por ‘Values’, añadir puntos y coma al final de cada línea de código y reemplazar ‘->’ de R por el signo ‘=’. Después deberemos abrir el archivo ‘visual.ts’ y reemplazar ‘NodeListOf’ por ‘HTMLCollectionOf’, en caso de que esto no sea así. Por último, ir al archivo ‘pbiviz.json’ y asegurarse de haber introducido nombre, email, url de soporte y una descripción de la visualización. A continuación, debemos volver a la terminal e introducir el comando ‘pbiviz package’ y ver que no hay ningún error. En ese caso ya tendremos nuestra visualización creada y deberemos importarla a Power BI. En la pantalla de inicio de Power BI, en el apartado de visualizaciones, debemos ir a los tres puntos y a ‘Importar un objeto visual de un archivo’. Luego ya tendremos nuestra visualización interactiva de R en Power BI.

3.3. Uso de Preguntas y respuestas de Power BI

En ocasiones, la forma más rápida de obtener información acerca de nuestros datos es formular una pregunta. La característica Preguntas y respuestas de Power BI permite explorar los datos utilizando un lenguaje natural.

Esta herramienta puede utilizarse en informes en el servicio de Power BI o bien en Power BI Desktop. Para la opción de escritorio de Power BI deberemos ir a la página

¹Para más información consultar la página web <https://docs.microsoft.com/en-us/power-bi/developer/visuals/environment-setup?tabs=windows>

²Página para descarga: <https://nodejs.org/es/>.

principal y en la opción de insertar seleccionar ‘Preguntas y respuestas’. Otra opción más sencilla es hacer doble click sobre una página del informe. Entonces aparecerá una nueva visualización, que mostramos en la Figura 3.5.

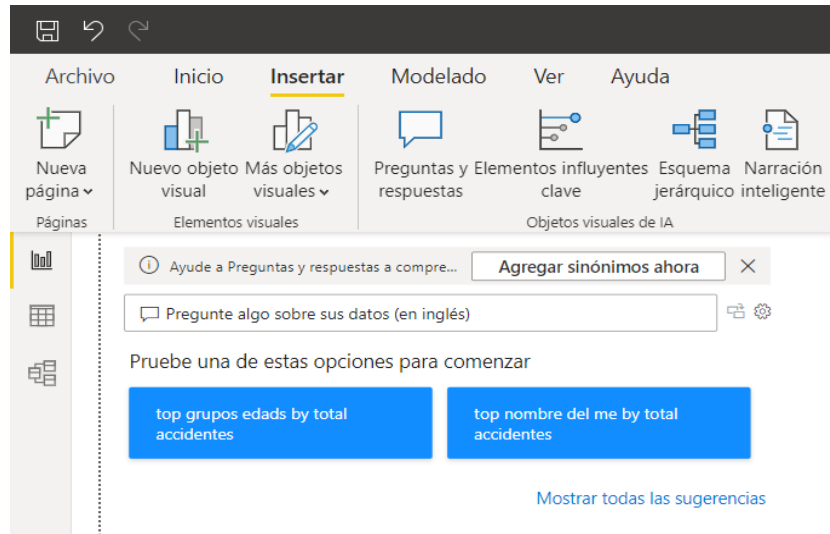


Figura 3.5: Herramienta Preguntas y respuestas de Power BI

A continuación, debemos ir al cuadro de preguntas y escribir la pregunta que queremos formularle a Power BI. A medida que vayamos escribiendo se mostrarán sugerencias para ayudar a construir la pregunta, así como la mejor visualización para mostrar la respuesta. Además, en la opción de ‘Agregar sinónimos ahora’ podemos crear sinónimos sencillos de campos o palabras que vayamos a utilizar de forma repetida. En la Figura 3.6 se muestran varios ejemplos de pregunta y respuesta.

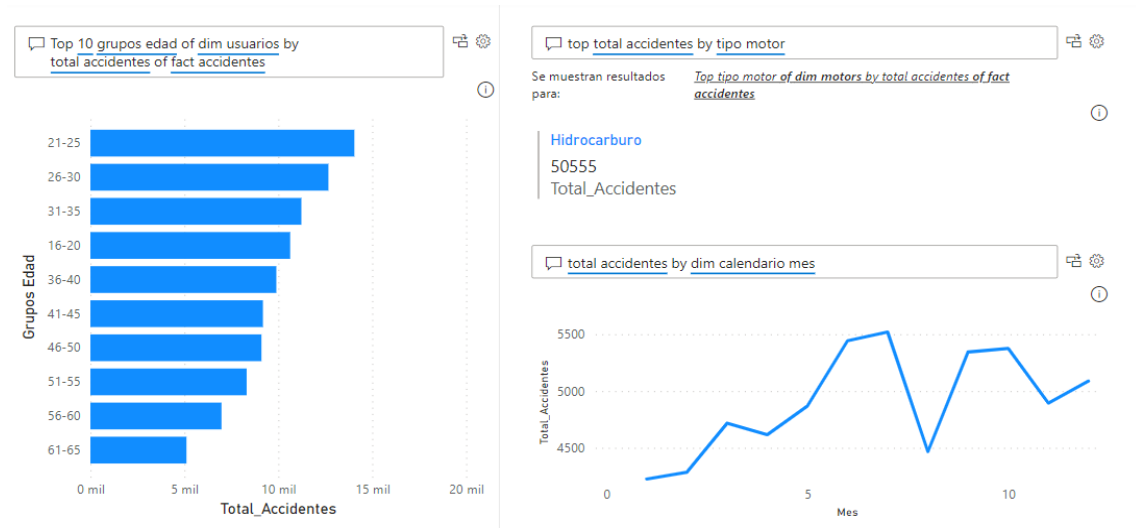


Figura 3.6: Ejemplo Preguntas y respuestas de Power BI

Capítulo 4

Conclusiones y trabajo futuro

Este último capítulo estará dividido en dos secciones: Conclusiones y Trabajo futuro. En la sección de Conclusiones veremos qué ventajas y desventajas posee cada herramienta dependiendo de si es un usuario principiante o bien una empresa. Por último, en la sección de Trabajo futuro, propondremos dos líneas de investigación que quedan abiertas una vez terminado este trabajo.

4.1. Conclusiones

A lo largo de este trabajo hemos analizado las características y funcionalidades de las herramientas de *Business Intelligence* Power BI, Qlik Sense, Tableau, Redash, Metabase y Superset.

Para hacer una comparación entre las herramientas de BI no sólo hay que fijarse en las opciones que cada una ofrece, sino en el motivo por el cual se utiliza. En particular, un usuario que quiere iniciarse en el *Business Intelligence* no tiene las mismas necesidades que una organización con cientos de empleados.

Supongamos que estamos en el primer caso, es decir, un usuario que se está iniciando. Los puntos a tener en cuenta serían los siguientes:

1. Precio: Se valoraría que la herramienta tenga una opción gratuita o de prueba, o quizás alguna opción con un precio asequible. Hemos visto que todas poseen opción gratuita, a excepción de Qlik Sense, que posee una versión de prueba de 30 días.
2. Descarga y sistema operativo: Se valoraría la facilidad para descargarlo y que exista alguna opción para el sistema operativo que utiliza. Con respecto al sistema operativo utilizado, las herramientas Redash y Superset no disponen de versión en windows, por lo que sería necesario utilizar Linux o bien una máquina virtual. Además, la descarga de las herramientas *open source* es más complicada, puesto que el despliegue lo hace el propio usuario.
3. Conexiones a datos: Lo habitual será comenzar con archivos de dominio público, que suelen ser archivos de texto, Excel, etc, y no bases de datos. Hemos visto que Metabase, Redash y Superset tan sólo permiten conexiones a bases de datos.
4. Sencillez: Se valoraría la sencillez de uso de la herramienta, que sea intuitiva.

Teniendo en cuenta estas características consideramos que las herramientas Power BI, Tableau y Qlik Sense son las más adecuadas para un usuario principiante. Por último,

un aspecto importante a tener en cuenta es la exigencia del mercado actual, que como ya vimos en el Cuadrante Mágico de Gartner para Analytics 2021 (Figura 1.1), Power BI es líder en el sector.

Consideremos ahora la opción de una empresa que quiere tomar una decisión acerca de qué herramienta utilizar. En este caso debemos tener en cuenta otras necesidades, que vendrán también marcadas por el tamaño de la empresa. A continuación mostramos los puntos que consideramos más importantes:

1. Precio: Empresas grandes podrán permitirse presupuestos mayores que empresas pequeñas. Dado que en este trabajo tan sólo estamos considerando la opción gratuita de las herramientas, no será un factor de importancia, aunque quedaría descartada la herramienta Qlik Sense, por sólo ofrecer una versión de prueba de 30 días.
2. Privacidad: En cualquier empresa tanto los datos que se utilicen como los informes creados deberán ser privados. Hemos visto que en la versión gratuita de Tableau los informes son de dominio público, por lo que pueden ser vistos y descargados por cualquier usuario, luego esta opción es inviable para cualquier tipo de empresa.
3. Conexiones a datos: Esto dependerá de lo que la empresa utilice habitualmente. Power BI destaca en este aspecto por la gran variedad de conexiones que permite; por otra parte, Metabase, Redash y Superset permiten únicamente conexiones a bases de datos.
4. Facilidad a la hora del desarrollo: Aquéllos usuarios que tengan conocimientos en Excel les resultará mucho más sencillo utilizar el lenguaje DAX de Power BI, mientras que aquellos que hayan trabajado con lenguaje SQL les resultará más cómodo trabajar con Metabase, Redash y Superset.
5. Colaboración: No todas las herramientas permiten la colaboración con otros usuarios. Con la versión gratuita de Power BI esto no es posible. Sin embargo, sí lo permiten las herramientas *open source*.

Por tanto, si se tratase de una pequeña empresa, en la que no es necesario compartir dashboards con otros usuarios, podrían utilizarse las herramientas Power BI, Metabase, Superset o Redash. En el caso de empresas pequeñas, medianas y grandes que necesiten colaborar con otros usuarios, podrían utilizarse Metabase, Superset o Redash.

4.2. Trabajo futuro

Como continuación de este trabajo, existen diversas líneas de investigación que quedan abiertas y en las que es posible continuar trabajando. A lo largo de este trabajo hemos realizado una investigación de las herramientas de *Business Intelligence* considerando en mayor medida la versión gratuita o de prueba. Esta versión puede no ser suficiente para pequeñas o grandes empresas que manejan grandes cantidades de datos o que quieren realizar análisis más profundos. Por este motivo podría resultar de interés realizar una comparación de las herramientas ya vistas teniendo en cuenta su versión de pago. Otra de las líneas de investigación posibles sería utilizar otras herramientas de BI, donde podría considerarse alguna de las otras herramientas que aparecen en el Cuadrante Mágico de Gartner para Analytics 2021 (Figura 1.1), como por ejemplo Looker (Google) ¹ o Domo ².

¹Página oficial: <https://www.looker.com/>.

²Página oficial: <https://www.domo.com/>.

Bibliografía

- [1] Abellán, E. (2020). Qué es un dashboard de negocios y cuáles sus beneficios. We are marketing. <https://www.wearemarketing.com/es/blog/que-es-un-dashboard-de-negocios-y-cuales-sus-beneficios.html>. Accedido 6 de junio de 2022.
- [2] Ataccama (2022). What is Data Quality and Why Is It Important?. <https://www.ataccama.com/blog/what-is-data-quality-why-is-it-important>. Accedido 6 de junio de 2022.
- [3] Carson Sievert, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec, Pedro Despouy and Salim Brüggemann (2021). plotly: Create Interactive Web Graphics via 'plotly.js'. R package version 4.10.0. <https://cran.r-project.org/web/packages/plotly/index.html>. Accedido 9 de junio de 2022.
- [4] core-js. Package version 3.22.8. (2022) <https://www.npmjs.com/package/core-js>. Accedido 9 de junio de 2022.
- [5] Data Flair (2022). Qlik Sense Data Model. Associations in Qlik Sense. <https://data-flair.training/blogs/qlik-sense-data-model/>. Accedido 6 de junio de 2022.
- [6] D3. Package version 7.4.4. (2022). <https://www.npmjs.com/package/d3>. Accedido 6 de junio de 2022.
- [7] Escobar, M. (2019). Data Profiling, Quality & Distribution in Power BI / Power Query. The Power User. <https://www.thepoweruser.com/2019/08/13/data-profiling-quality-distribution-in-power-bi-power-query/>. Accedido 6 de junio de 2022.
- [8] Gartner (2021). Magic Quadrant de Gartner. <https://www.gartner.es/es/metodologias/magic-quadrant>. Accedido 6 de junio de 2022.
- [9] Gartner (2021). Magic Quadrant for Analytics and Business Intelligence Platforms.
- [10] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.9. <https://cran.r-project.org/web/packages/dplyr/index.html>. Accedido 9 de junio de 2022.
- [11] Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani and Dewey Dunnington (2022). ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.3.6. <https://cran.r-project.org/web/packages/ggplot2/index.html>. Accedido 6 de junio de 2022.
- [12] Levy, R. (2021). Data Quality Dimensions: How Do You Measure Up?. Precisely. <https://www.precisely.com/blog/data-quality/data-quality-dimensions-measure>. Accedido 6 de junio de 2022

- [13] Ioannou, K. (2019). How to create R custom visual (html) in Power BI. Medium. <https://medium.com/@Konstantinos.Ioannou/how-to-create-an-r-custom-visual-html-for-powerbi-7f2d2e44e453>. Accedido 6 de junio de 2022.
- [14] Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239), 2.
- [15] Microsoft (2022). Microsoft PowerBI Custom Visuals API. Package version 4.2.0. <https://www.npmjs.com/package/powerbi-visuals-api>. Accedido 6 de junio de 2022.
- [16] Mike A. (2022). Data Warehouse: ¿qué es y cómo utilizarlo?. Data Scientist. <https://datascientest.com/es/data-warehouse-que-es-y-como-utilizarlo>. Accedido 6 de junio de 2022.
- [17] Ortiz, D. (2021). ¿Qué es un dashboard y para qué se usa?. Cyberclick. <https://www.cyberclick.es/numerical-blog/que-es-un-dashboard>. Accedido 6 de junio de 2022.
- [18] Ranjan, J. (2009) Business Intelligence: Concepts, components, techniques and benefits. *Journal of Theoretical and Applied Information Technology* Vol 9. No 1: 60-70.
- [19] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>. Accedido 6 de junio de 2022.
- [20] Samos, J. (2021) Sistemas Multidimensionales. <https://lsi2.ugr.es/jsamos/sm2019/>. Accedido 6 de junio de 2022.
- [21] Terry Therneau, Beth Atkinson and Brian Ripley (2022). rpart: Recursive Partitioning and Regression Trees. R package version 4.1.16. <https://cran.r-project.org/web/packages/rpart/index.html>. Accedido 9 de junio de 2022.
- [22] TypeScript. Package version 4.7.3. (2022) <https://www.npmjs.com/package/typescript>. Accedido 9 de junio de 2022.