Universitat Oberta de Catalunya

M2.859 Visualización de datos

A9 - Creación de la visualización y entrega del proyecto (Práctica)

Marc Moreno Galimany
mmorenogali

Junio, 2021

Índice

1. Presentación de la visualización

El título de la visualización *Accidentes en las carreteras de UK en función del género, edad y fecha* se ha escogido con el objetivo de resumir el dashboard presentado.

En resumen, la visualización presenta los datos de accidentes en el Reino Unido (UK) a excepción de Irlanda del Norte, según el género del conductor y la víctima, la edad del conductor y la fecha de ocurrencia.

La visualización de los datos se puede encontrar en `https://mmorenogali.github.io/UKaccidents/`,

El presente informe se ha estructurado en cuatro bloques principales. El primero, es la presentación de la visualización, que hace una presentación a modo de *abstract* de la visualización.

A continuación, en el punto 2 se encuentra una descripción de las preguntas y cuestiones que pretende responder la visualización, así como una explicación de los motivos por los que se ha utilizado cada tipo de gráfico y las facilidades que le proporciona al usuario.

En el punto 5 Se encuentra la descripción técnica del proyecto: Software utilizado, código, lenguajes, librerías, datos...

Por último, en el punto 4 se justifica la elección del color, el modo como se puede responder a las preguntas planteadas, las interacciones, títulos, leyendas...

Los enlaces a los repositorios y a la visualización se encuentran al final del documento.

2. Explicación de la visualización

2.1 Preguntas

La visualización se ha realizado con unas preguntas muy claras en mente. Tal y como se vio en los requisitos de la práctica anterior, la visualización debía tener en cuenta la perspectiva del género.

Como en muchos otros ámbitos, el género es un generador de tópicos que en muchos casos se desmienten cuando se contrastan con los datos. En el caso de la conducción, es típico oír expresiones relacionadas con la mala destreza de las mujeres al volante.

Por este motivo, y tal y como se definió en la anterior práctica, las cuestiones entorno a las que se ha realizado esta visualización son:

- La severidad del accidente en función del género.
- Si hay relación entre los días/horas de la semana, el género y los accidentes.
- La relación entre la edad y el número de víctimas.
- Si en función de la zona hay más accidentes de un género o de otro.
- Si hay alguna correlación entre el género del conductor/a y el de la víctima.

Sin embargo, gracias a la potencia de la visualización y los filtros, y a que se ha seguido el mantra de Schneiderman (Filtrado, Zoom-in y muestra de datos bajo demanda) se pueden se pueden contestar a otras preguntas además de las originales.

Por ejemplo, la visualización permite el filtrado por zona, límite de velocidad, edad del conductor, fecha... Por lo que se puede responder a preguntas como: *¿Cuantos accidentes hay en una zona determinada?* o bien *¿Dónde se concentran los accidentes más severos?*

2.2 Tipos de gráficos y filtros. Caso de uso

Los gráficos y filtros se han escogido teniendo en cuenta lo que se ha estudiado en la teoría y el uso que le podría dar un usuario medio que esté interesado en responder a alguna de las preguntas planteadas.

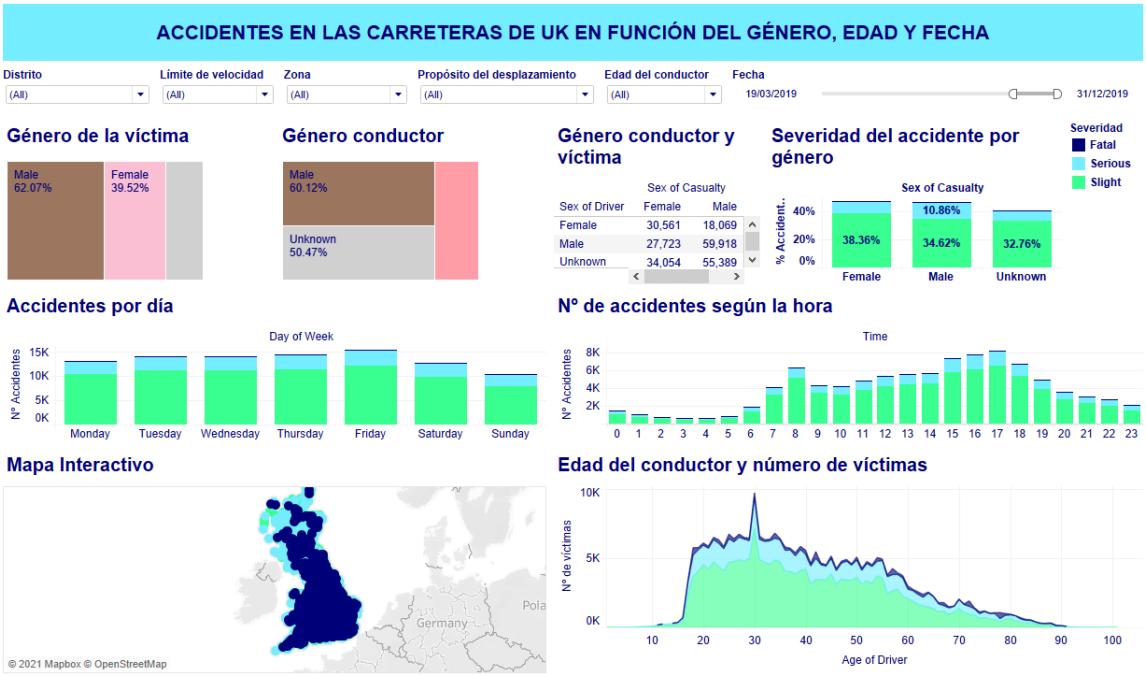

Figure 2.1: *Dashboard. Elaboración propia*

En este sentido los **filtros** se han colocado en la parte superior del dashboard, una posición visible y fácilmente localizable. Los títulos son cortos pero explicativos, y en general son del tipo dropdown de múltiples opciones. Esto permite que no ocupe espacio innecesario en la visualización pero que a su vez se desplieguen cuando el usuario quiera centrar su atención en los filtros.

El **filtro temporal** se ha definido como un rango temporal en sobre un eje lineal, ya que culturalmente comprendemos el tiempo como una línea en 1D. Esto permite al usuario filtrar rápidamente y con relativa precisión el intervalo de fechas que quiere estudiar, además de ver rápidamente la proporción de datos que se muestran. En el caso de que necesite un ajuste más preciso, el filtro permite establecer la fecha sobre el calendario al hacer clic.

Finalmente el usuario puede encontrar una **leyenda** sobre la severidad del accidente que funciona a su vez como filtro.

Cabe comentar que todos los gráficos actúan a su vez de filtro, por lo que si el usuario realiza la selección de un grupo de interés, este se aplicará a la visualización completa.

En la visualización de los datos se pueden observar varios tipos de gráficos. Por un lado hay **dos mapas de árbol** `https://datavizcatalogue.com/ES/metodos/mapa_de_arbol.html` con una única jerarquía, que se utiliza como elemento visual para comparar la proporción entre el género de la víctima en un gráfico y el género del conductor/a en el otro.

Para facilitar la comprensión de los datos se han añadido las etiquetas y el porcentaje que ocupa cada área.

Por otro lado encontramos una **tabla** textual que permite ver y comparar el número de accidentes según el género del conductor y la víctima. Se ha optado por una tabla que permite al usuario comparar los datos directamente. De este modo no sólo comprende el porcentaje, sino el valor total.

El usuario encuentra también tres **gráficos de barras apiladas** `https://datavizcatalogue.com/ES/metodos/grafico_de_barras_apiladas.html`. Uno de ellos, el de severidad del accidente por género, es un **gráfico de barras apiladas al 100%**. Esto es así ya que como se ha podido observar en los dos gráficos anteriores, la proporción de accidentes no es similar entre hombres y mujeres. De este modo se pueden comparar la proporción de la severidad de accidentes según el género fácilmente.

Los otros dos **gráficos de barras** se utilizan para que el usuario pueda comprender el total de accidentes según el día de la semana y la hora del día, y a su vez, la severidad del accidente.

El usuario encuentra también **un mapa** donde se pueden ver todas las localizaciones de los accidentes registrados. Estos están diferenciados por color según la severidad del accidente.

Por último encuentra un **gráfico de área apilada** `https://datavizcatalogue.com/ES/metodos/grafico_de_area_apilada.html` donde se puede ver la evolución del número de víctimas en función de la edad del conductor.

Todos los gráficos tienen la función de hoovering que permiten al usuario conocer los detalles de los datos en cada gráfico.

3. Descripción técnica

3.1 Software

Para la realización de la práctica se ha utilizado el siguiente software:

- **Tableau 2021.1** Para la realización de la visualización.
- **Overleaf** Para la generación del informe
- **Github** Para alojar el código y los documentos.
- **Google Drive** Para alojar las tablas que no se han podido alojar en Github debido a la limitación de tamaño de archivo (25 MB).
- **Python 3.8.6** Para realizar la exploración, unión y procesado de los datos
- **OpenOffice y Excel** Para la visualización general de las tablas y ficheros csv.

En **Tableau** se ha realizado el dashboard a partir de las distintas visualizaciones realizadas. Además, con la expresión *COUNTD([Accident_Index]* se ha creado un nuevo campo para realizar el recuento del número de accidentes.

Overleaf `https://es.overleaf.com/` es un editor de LaTeXonline, se ha utilizado para la realización del informe, ya que permite organizar de forma fácil la estructura del documento.

El código de **Python** utilizado se encuentra alojado en el repositorio de **github**

Los pasos realizados para obtener los datos finales para realizar la visualización se detallaron en la práctica anterior. Sin embargo, para una mayor facilidad del flujo de trabajo se ha realizado el filtrado para incluir únicamente los años 2014-2019, ya que incorporar datos anteriores hacía que el peso del fichero csv fuese demasiado pesado para el ordenador que se dispone para realizar la práctica.

Tal y como se puede ver en el notebook https://github.com/mmorenogali/UKRoadSafety, se han unido todas las tablas proporcionadas por el gobierno de Reino Unido de 2005 a 2019 en una sola tabla, por lo que se han unificado los datos en un solo conjunto. Para cada año el gobierno libera tres tablas: una correspondiente a las circunstancias del accidente, otra que hace referencia al vehículo y otra que hace referencia a los daños personales.

Las tres tablas están unidas por el índice de accidente. Mediante esta unión se han juntado un total de 18 tablas. También se ha realizado el mapeado de las variables categóricas con las categorías del documento "variable lookup.xls". De este modo no necesitamos consultar distintos ficheros para entender qué se esta viendo en el dataset.

Es decir, se han utilizado un total de 18 tablas que se han procesado hasta llegar a la tabla final. Durante este procesado se han modificado algunos nombres de campos para poder realizar el mapeado de forma automática, ya que se presentaban algunas abreviaciones en los nombres de las páginas de la hoja de cálculo anteriormente mencionada.

Mediante las herramientas Python, un editor de hojas de cálculo y Tableau se ha realizado la exploración del conjunto de datos. Se han encontrado un total de 69 variables y 2459050 registros.

3.2 Justificación

El software utilizado se ha seleccionado por varios motivos, que se pueden clasificar según el software.

Por un lado está el motivo de utilizar **Tableau** como herramienta principal de visualización. Se ha utilizado este software ya que se dispone de una licencia gratuita por ser estudiante de la UOC, es una herramienta muy potente y intuitiva de utilizar, con un amplio abanico de funciones y además, permite la publicación en abierto para cualquiera, aunque no tenga Tableau.

Por otro lado está la utilización de **Python**. En este caso se ha utilizado por ser el lenguaje de programación que utilizo laboralmente, por lo que estoy familiarizado. Sin embargo, también se podría haber realizado con otro lenguaje como por ejemplo R.

Las librerías utilizadas son:

- **pandas**. Librería para trabajar con datos en formato tabla. Gracias a esta librería se ha podido realizar la carga, unión, procesado y guardado de las distintas tablas.

- **os** Útil para trabajar con archivos y directorios
- **numpy** Para operaciones vectoriales y selección de valores
- **gc** Garbage Collector. Debido a que Python no dispone de control de memoria como por ejemplo, C, con esta función podemos eliminar de la memoria variables que ya no se utilizan y liberar al sistema.
- **time** Para calcular los tiempos de ejecución.
- **xlrd** Ha sido útil para cargar archivos .xls y tomar los nombres de las pestañas, así como los valores de cada una de ellas.

La utilización de **Overleaf** como editor de texto y no otro es, principalmente, la facilidad de organización, la maquetación automática, el auto guardado y la disponibilidad online, que permite reprender el trabajo desde cualquier ordenador.

Alojar el trabajo en **Github** es debido a la facilidad de compartir el código, controlar el historial de versiones, gestionar el repositorio, la sincronización con los dispositivos y la previsualización online sin necesidad de software de terceros ni la necesidad de disponer de una cuenta para acceder.

Sin embargo el tamaño de archivo está limitado a 25MB, lo que hace que se haya necesitado de una plataforma Cloud para poder compartir libremente los archivos. En este caso se ha escogido **Google Drive** ya que la UOC proporciona almacenaje ilimitado con la cuenta de estudiante.

Por último el hecho de utilizar un editor de hojas de texto como **OpenOffice** o **Excel** ha facilitado la previsualización de los campos de las tablas.

3.3 Licencia

La licencia de los datos es *Open Government Licence for public sector information*, y se puede encontrar en detalle en la página `http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/`.

Esta licencia permite la copia, publicación, distribución y transmisión libre de la información, la adaptación y explotar la información de forma comercial y no comercial, incluyendo otra información de forma libre.

Lo único que se pide es referenciar los datos en el proyecto. El enlace a los datos se puede encontrar al final de este documento.

Las excepciones incluidas en la licencia no se aplican en nuestro caso, ya que son datos agregados y no proporcionan información como datos personales, patentes, logotipos, documentos de identidad...

4. Visualización de datos

4.1 Respuesta a las preguntas

La visualización responde correctamente a las preguntas descritas con anterioridad, tal y como se muestra a continuación.

La primera pregunta, *Severidad del accidente en función del género* se responde mediante el gráfico de barras apiladas que lleva el mismo título. Con este gráfico se pueden observar la proporción al y ver que los hombres sufren accidentes más severos que las mujeres.

Para responder a la pregunta de si hay relación entre los días/horas de la semana podemos ver claramente en los gráficos centrales que sí. Durante la noche ocurren pocos accidentes, mientras que los picos se producen a las 8 de la mañana y a media tarde. Al parecer coincide con el horario laboral y escolar.

También se concentran más accidentes el viernes que el resto de días de la semana. Para poder diferenciar estos valores según el género, podemos escoger entre si queremos visualizar según el género de la víctima o del conductor (o ambos). Haciendo clic en cualquier gráfico que contenga el género, se actualizan los valores para que el usuario pueda responder a la pregunta.

Por otro lado vemos que la respuesta entre la edad y el número de víctimas es inmediata con el gráfico de áreas apiladas. Observamos un importante pico en la edad de 30 años, y una concentración más general entre los 20 y los 55-60. La subida cercana a la edad de 20 años es debida a que legalmente, a los 18 años se puede empezar a conducir.

Para responder a la última pregunta se puede consultar la tabla numérica "Género conductor y víctima", que proporciona los valores numéricos para responder rápidamente a esta pregunta.

Pero como hemos visto en el punto 2, gracias a la ayuda de los filtros y el mapa se puede responder a otras preguntas y refinar las respuestas por zonas, límites de velocidad, fecha...

4.2 Interactividad

Como se ha comentado anteriormente, se ha realizado la visualización de la manera más interactiva dentro de las posibilidades de Tableau. De este modo encontramos, por ejemplo, la posibilidad de hacer zoom en el mapa, que permite visualizar los accidentes directamente en el lugar donde se han producido y ver así las zonas con registros de accidentes.

También se ha aprovechado la interactividad que proporciona del hoovering para proporcionar al usuario los datos agregados bajo demanda.

La utilización de filtros también permite un grado de interactividad para obtener las respuestas a las preguntas de un modo menos general. El hecho de utilizar los gráficos como filtro aumenta el grado de participación del usuario.

4.3 Colores

La selección de los colores se ha realizado de acuerdo al libro de estilo de la UOC `https://www.uoc.edu/portal/es/llibre-estil/index.html`. Esto es así ya que si estuviéramos realizando una tarea profesional, es conveniente utilizar los colores y estilos de la marca.

Además, se han utilizado los colores para describir el mismo hecho. Por ejemplo los colores del grado de severidad de los accidentes se comparten para toda la visualización, mientras que en el caso del género se han utilizado dos colores distintos. Para el género se han escogido colores que, aunque sean objeto de estereotipos, permiten al usuario clasificar inconscientemente el género. El rosa para el género femenino y el marrón para el masculino. En el caso de los datos desconocidos se ha tomado el color gris.

Para los colores de la severidad del accidente se ha escogido el verde para los accidentes menos severos, el azul celeste para los graves y el azul oscuro para los fatales. De este modo se asocia inconscientemente el verde como bueno, y la intensidad de los azules como incremento de la gravedad.

5. Enlaces externos

- **Enlace a Github** `https://github.com/mmorenogali/AccidentsUK`
- **Enlace a Google Drive**[1] Click to Google Drive
- **Enlace a la fuente de datos** https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data
- **Enlace a Tableau Public** https://public.tableau.com/app/profile/marc3064/viz/PRAC_mmorenogali/Dashboard1
- **Enlace a Github Pages** `https://mmorenogali.github.io/UKaccidents/`

[1] Únicamente se ha podido dar acceso a los usuarios con un correo ...@uoc.edu