

VISUALIZACIÓN AVANZADA

PRÁCTICA 5.8: ANÁLISIS ORIGINAL DE DATOS SOBRE FILMES EN TABLEAU

USC, CURSO 2022-2023

DIANA MASCAREÑAS SANDE

ÍNDICE

INTRODUCCIÓN.....	3
ELECCIÓN DE UN DATASET	3
TEMA A TRATAR	3
PREPROCESADO DE LOS DATOS	3
ESTUDIO DE LOS DATOS Y ANÁLISIS DE RESULTADOS.....	4
ANÁLISIS DE LOS DASHBOARDS	14
CONCLUSIONES.....	17
BIBLIOGRAFÍA	17

INTRODUCCIÓN

En esta práctica se llevará a cabo un estudio, empleando la herramienta Tableau, sobre los datos de un dataset específico. Se explicará el contenido del mismo, modificaciones que se tengan que hacer para poder llevar a cabo el análisis... entre otros factores. Se detallará qué información concreta se busca en el estudio y cuando se construyan las visualizaciones, se concretarán los resultados en este informe.

Finalmente, se expondrán unas conclusiones tras el trabajo realizado.

ELECCIÓN DE UN DATASET

Para la elaboración de esta práctica, se ha elegido la base de *Filmaffinity dataset (Spanish)*, que se puede encontrar en la fuente *kaggle.com*. La url exacta está disponible en el apartado de bibliografía.

Este dataset contiene información de 119003 filmes registrados entre los años 1900 y 2020 en la página web Filmaffinity (esta url también está disponible en la bibliografía).

El conjunto de datos contiene un único csv con un información sobre estos filmes: *filmaffinity_dataset.csv*. Este tiene diferentes campos: título del filme, año del filme, país del filme, nombre/nombres de las personas que forman la dirección del filme, reparto del mismo, nota proporcionada por los usuarios de *Filmaffinity*, tipo de filme (película, serie...) y género.

TEMA A TRATAR

Se realizará un estudio a partir del dataset elegido. En un primer momento, el análisis se basará en los datos en función de los géneros registrados de los filmes, para a continuación centrarse en información relativa a los propios filmes. Finalmente, se analizarán los datos referidos al reparto y su dirección.

En definitiva, se llevará a cabo un pequeño estudio sobre aspectos que resultan de interés sobre la tabla que conforma el dataset.

PREPROCESADO DE LOS DATOS

En este caso, no se ha necesitado realizar una limpieza de datos previa, más bien, durante el análisis, se descartarán valores nulos según convenga, ya que eliminarlos desde un

principio puede alterar enormemente el contenido del dataset, al eliminar filas completas del mismo.

ESTUDIO DE LOS DATOS Y ANÁLISIS DE RESULTADOS

En la sección de fuente de datos, se especifican las propiedades del fichero de texto del que se obtiene el dataset:

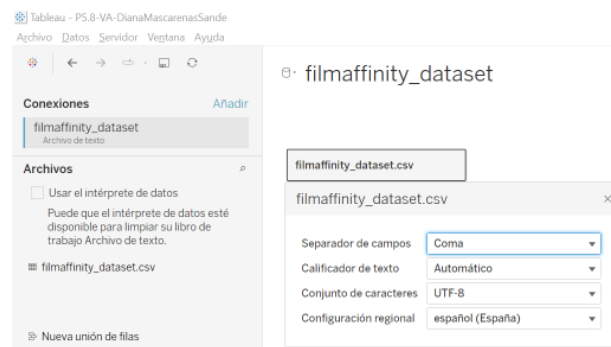


Ilustración 1: propiedades del csv

Primer dashboard: Información relativa a los géneros.

La primera página o dashboard estará centrado en el estudio de los géneros registrados. En este caso, en el dataset, la mayoría de filmes tiene más de un género asociado, por lo que se ha decidido elegir solo aquellos filmes que tengan un único género asociado. Dichos géneros son 14: acción, animación, aventuras, ciencia ficción, comedia, documental, drama, fantástico, infantil, intriga, romance, terror, thriller y western.

Este filtro que selecciona los géneros indicados anteriormente se mantendrá en todos los gráficos de este primer dashboard.

Número de filmes por género

Se ha querido averiguar el número de películas que se han registrado en filmaffinity por género entre 1900 y 2020. Para ello, se ha creado un campo calculado: *Total filmaciones*, que cuenta el número de filas de la tabla, es decir, el número de filmes.

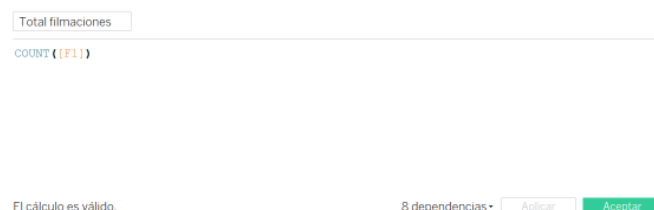


Ilustración 2: campo calculado Total filmaciones

En la imagen, “F1” es el nombre asignado a la primera columna de la tabla, que es un identificador de la fila/filme.

A continuación, se arrastra este campo calculado a las filas, y el campo *Género* a columnas. Este último también se coloca en filtros para elegir los 14 géneros de interés. El campo calculado *Total filmaciones* se arrastra, en el menú de Marcas, al campo *Etiquetas*, para que de cada género se vea con claridad el número total de películas asociadas.

Finalmente, en el panel *Mostrarme*, se selecciona el tipo de gráfico. En este caso se considera adecuado un gráfico de barras verticales, para apreciar qué género tiene asociadas más y menos películas, en función de la altura de las barras. Se ordenan los resultados de forma descendente para apreciar aún mejor esta información.

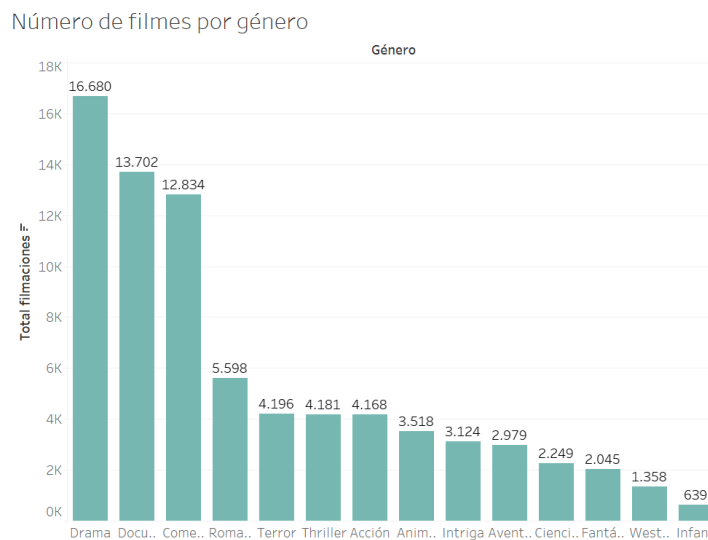


Ilustración 3: número de filmes por género

En la segunda ilustración, se puede apreciar que, entre los años 1900 y 2020, del género del que se ha registrado un mayor número de filmes es *Drama*, con una cantidad asociada de 16680. Por otra parte, el género con menos filmes asociados en estos años es *Infantil*. Además, se puede ver un gran salto entre las filmaciones de *Comedia*, con un total de 12834, respecto a las de *Romance*, que es la siguiente en la clasificación, pero con un total de 5598 filmes asociados. Así, los géneros con un mayor número de filmes asociados, y con una gran diferencia respecto al resto, son *Drama*, *Documental* y *Comedia*.

Nota media por género

Lo siguiente que se ha querido estudiar es la nota media que los usuarios han dado a cada género. Para ello, se ha construido un gráfico de barras, en esta ocasión horizontales, ya que no se tiene por qué apreciar tanta diferencia como en el caso anterior (ahora se va a trabajar con puntuaciones del 1 al 10, no en unidades de miles), y esta disposición permite ver ordenadamente, de arriba abajo, la puntuación de cada género. Se ordenarán de forma descendente para que, en la primera fila, se vea el género con mayor puntuación. En las columnas se ha colocado el campo calculado *Media nota*, que realiza la media sobre el campo numérico *Nota*:

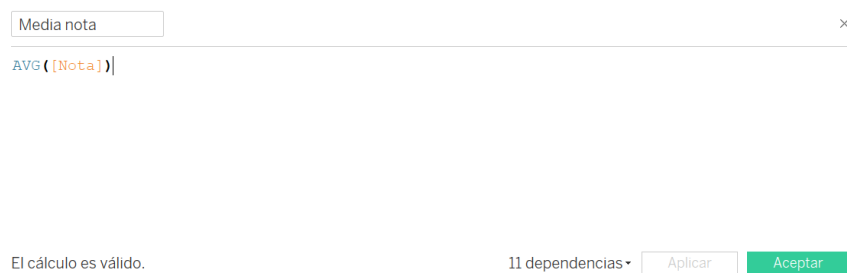


Ilustración 4: campo calculado media nota

Este campo calculado se ha arrastrado a *Etiqueta*, para que se vea la nota media exacta de cada género situada al lado de la barra correspondiente. Por otra parte, se ha arrastrado el campo *Género* a las filas y a los filtros, para seleccionar los 14 géneros de interés.

Desde el panel *Mostrarme* se ha elegido el gráfico de barras horizontales.

A continuación se muestra la gráfica resultante:

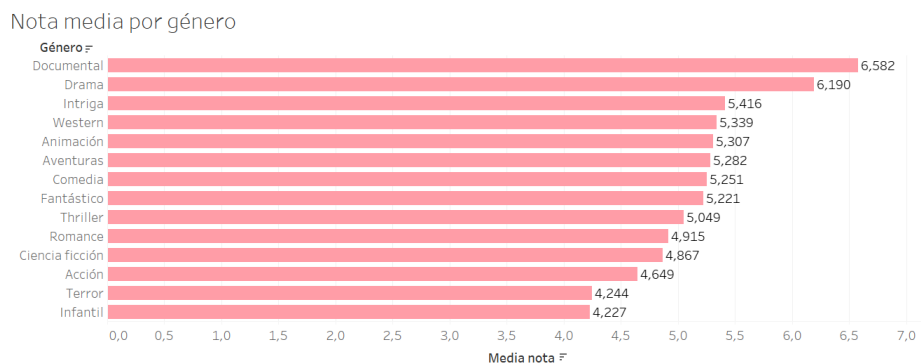


Ilustración 5: nota media por género

Como se puede apreciar, los géneros mejor puntuados son *Documental* y *Drama*, ambos en la cabeza de la clasificación del número de filmes asociados a cada género. Por otra

parte, los géneros con peor puntuación son *Terror* e *Infantil*. Se puede ver cómo las puntuaciones dadas entre 1900 y 2020 oscilan entre el 4 y 7 sobre 10. Sería interesante conocer el número de opiniones proporcionadas por filme, ya que no es lo mismo que puntúe un único usuario a que lo hagan miles. Este sería un aspecto a mejorar del dataset.

Filmes por año

Finalmente, se quiere conocer el número de filmes que se han registrado por año, dentro de esos 14 géneros. Para ello, se empleará una tabla, en la que se colocará en las filas los años (se ha cambiado el tipo de dato de este campo para que sea una fecha), y se arrastrará a las etiquetas el campo de *Total filmaciones*. Por otra parte, en los filtros, se incluirán los 14 géneros de interés.

Desde el panel *Mostrarme*, se selecciona el mapa en árbol, para poder visualizar mejor en qué años se grabaron más filmes a partir del tamaño de la cuadrícula de este tipo de gráfico.

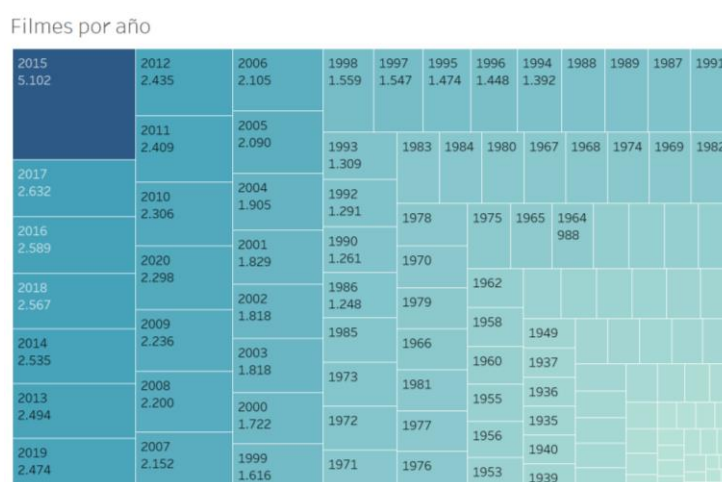


Ilustración 6: filmes por año

Así, se aprecia cómo es en 2015 el año en que más filmes se grabaron sobre los géneros seleccionados.

Segundo dashboard: Información relativa a los filmes.

Nota media por país

Cada filme tiene asociado su país. Se ha querido conocer la nota media que obtiene cada país desde 1900 a 2020. Para ello se ha colocado en las filas los países (se ha cambiado el tipo de dato de este campo para que se reconozca como país/región), en las etiquetas del panel *Marcas*, el campo calculado *Media nota* y, desde el panel *Mostrarme*, se ha

seleccionado *Mapas*. Es necesario destacar que Tableau no reconocía una serie de países porque el nombre con el que se les registró en el dataset no es exactamente igual al nombre que tiene Tableau registrado para ellos. Se ha arreglado manualmente introduciendo el nombre de los países correcto. Una vez hecho esto todavía quedan 8 países desconocidos, que se ha decidido filtrar ya que en la mayoría ha habido errores de formato que impiden conocer el nombre del país al que quieren hacer referencia.

El gráfico resultante es el siguiente:

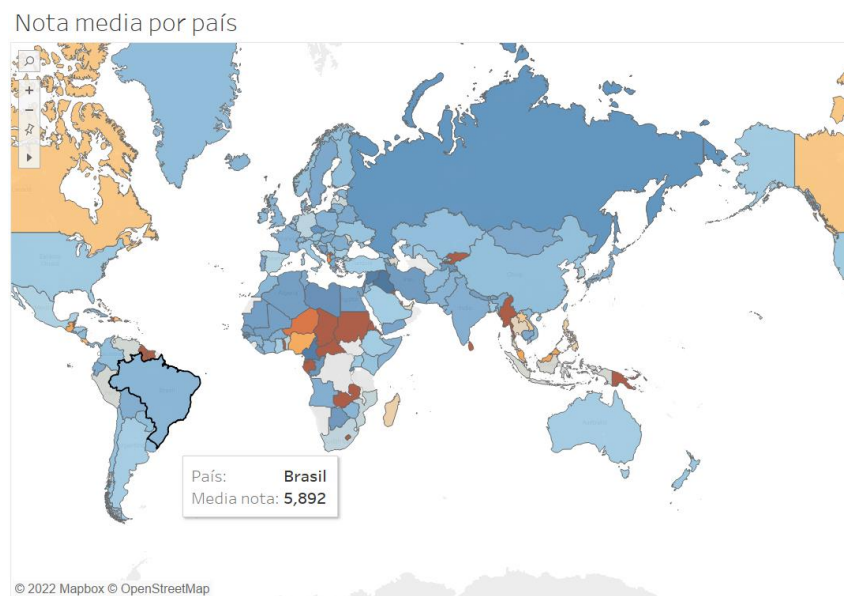


Ilustración 7: nota media por país

En la ilustración 6 no se ha incluido la leyenda para poder apreciar mejor el contenido del mapa, pero esta está incluida en la hoja *nota media por país* de Tableau. Los tonos azules indican mayor nota, cuanto más oscuros, mayor nota media obtienen. Los tonos naranjas señalan los valores más bajos de la nota media. En los más oscuros, como Sudán, no aparece dicha nota ya que hay países de los que se desconoce esta cualificación en el dataset.

Ahora, pasando el cursor por encima de los países que interesen, se podría conocer la nota media de los filmes grabados en cada país.

Evolución de tipo de filmes

Se quiere conocer la evolución en cuanto al número de filmes por cada tipo de los mismos a lo largo del tiempo, entre 1900 y 2020. Así, se podrá apreciar qué tipo de filme es el más común, además de la tendencia que sigue cada uno.

Para ello, se ha colocado el tiempo en las columnas, mientras que en las filas se ha arrastrado el campo de *Total filmaciones*. Para representar la evolución en el tiempo de cada tipo de filme en función del total de filmes, se arrastra el campo *Tipo filme* al campo *Color* del panel de marcas. El gráfico será de líneas.

El resultado obtenido es el siguiente:

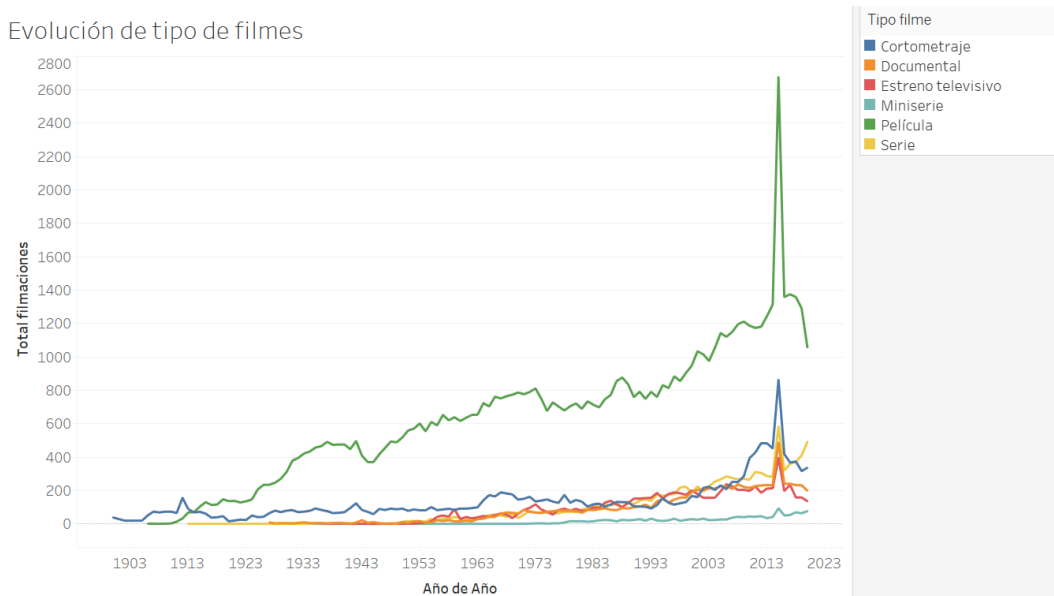


Ilustración 8: evolución de tipo de filmes

Como se puede apreciar, salvo en la primera década del 1900, el tipo de filme que ha ido en cabeza respecto a los demás, registradas en filmaffinity, son las películas. El resto, han tenido un menor crecimiento, sobre todo las miniseries. Sin embargo, en 2015 hubo un pico de registros de filmes de todos los tipos.

Si se selecciona la opción de ver las líneas de tendencia, se vería lo siguiente:

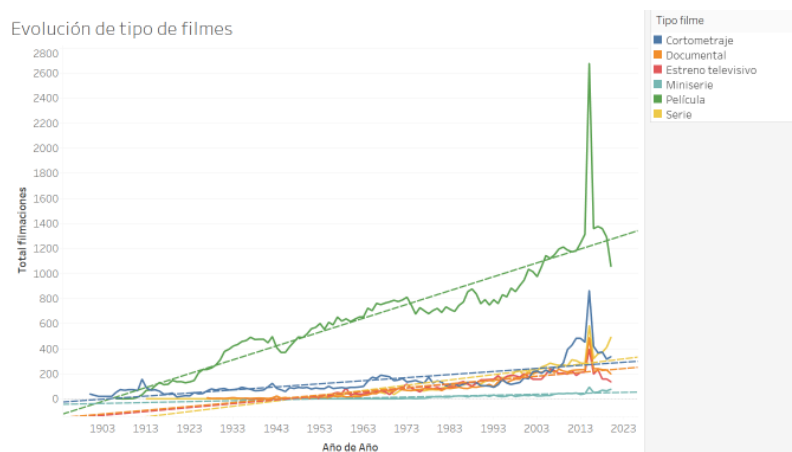


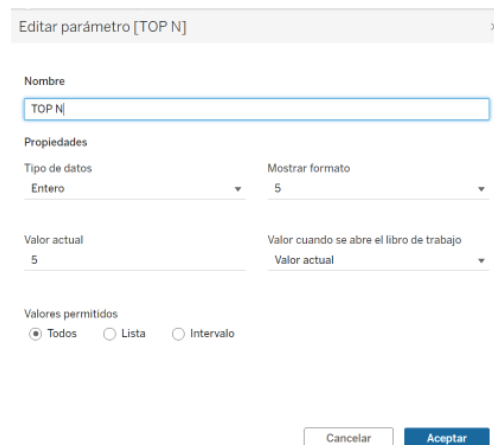
Ilustración 9: evolución tipo de filmes (líneas de tendencia)

A pesar de que el tipo de filme **cortometraje** tuviera por lo general, mayor número de filmes que las **series** a lo largo del tiempo, en los últimos años parece que se ha registrado un mayor número de estas últimas, viendo que su línea de tendencia tiene mayor pendiente en su tramo final y, superando además la de los cortometrajes. Esto indica que en los próximos años, las series podrían convertirse en el segundo tipo de filme con mayor número de filmes de filmaffinity.

5 mejores y peores valoraciones

Se ha querido reflejar un ranking con las 5 mejores y peores puntuaciones dadas a los filmes, con sus títulos correspondientes.

En el caso de los 10 mejor valorados se ha creado un parámetro llamado TOP N, que servirá para seleccionar 5 valores para el posterior ranking.



Editar parámetro [TOP N]

Nombre
TOP N

Propiedades

Tipo de datos
Entero

Mostrar formato
5

Valor actual
5

Valor cuando se abre el libro de trabajo
Valor actual

Valores permitidos
☒ Todos ☐ Lista ☐ Intervalo

Cancelar Aceptar

Ilustración 10: configuración del parámetro TOP N

Una vez se tiene este parámetro, se crean dos campos calculados: *RANKING* y *ÚLTIMOS*. En ambos se empleará la opción *rank* sobre el valor de *Media nota*, y se indicará que su rango será menor o igual al parámetro *TOP N*. La única diferencia entre *RANKING* y *ÚLTIMOS* es que, en el primero, se indicará que se siga un orden descendiente, para seleccionar las notas más altas y, en el segundo, su orden será ascendiente para elegir las más bajas.

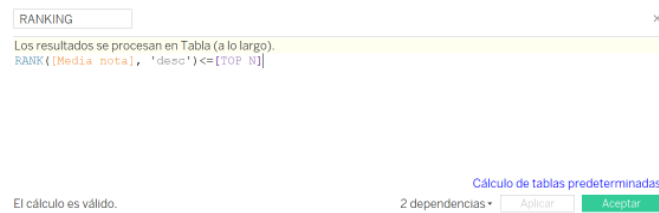


Ilustración 11: campo calculado *RANKING*



Ilustración 12: campo calculado *ÚLTIMOS*

El siguiente paso, en el caso de las 5 mejores valoraciones, es colocar el campo calculado *Media nota* en las columnas, el título de los filmes en las filas, el género en los filtros (se trabajará con los del primer dashboard, para mantener mayor coherencia) y el campo calculado *RANKING* con sus valores a *true*. En el caso de las 5 peores, se sigue el mismo procedimiento cambiando el campo calculado *RANKING* por *ÚLTIMOS*. Se colocan como etiquetas la nota media para apreciar mejor las puntuaciones de los rankings.

Los resultados son los siguientes:

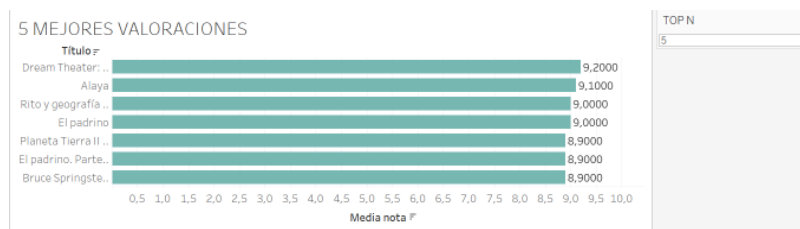


Ilustración 13: 5 mejores valoraciones

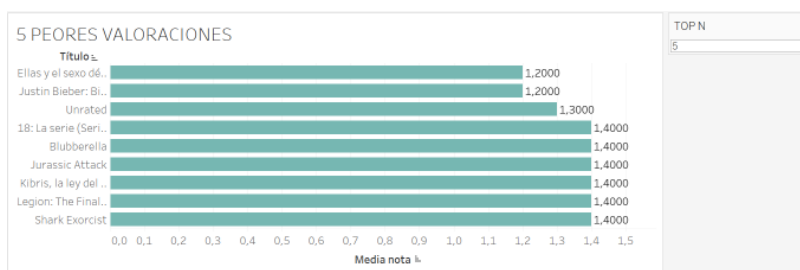


Ilustración 14: 5 peores valoraciones

En este caso se puede ver cómo las mejores valoraciones oscilan alrededor del 9 sobre 10, mientras que las peores no llegan al punto y medio sobre 10. En el menú de la derecha

(donde está el parámetro *TOP N*) se podría cambiar el 5 por cualquier otro número, en caso de querer comprobar otro ranking (top 3, top 10... entre otros).

Tercer dashboard: Información relativa al reparto y a la dirección.

Se ha querido obtener información relativa al número de filmes registrados por reparto, así como la puntuación obtenida para dichos filmes. Esto mismo se ha querido conocer sobre los directores de los filmes.

En el caso del reparto estos datos son más complejos, ya que hay filmes de las que se conoce un único actor o actriz y en otros donde se registran más actores del elenco. De este modo, realmente será más útil analizar la nota obtenida que el número de filmes, para saber qué reparto ha participado en filmes con mejor puntuación. Sin embargo se analizarán ambas cosas para ver qué resultados se obtienen.

Para obtener estos datos, se han colocado dos valores en las filas: *Total filmaciones* y *Media nota*. Por otra parte, se arrastra a filtros *Media nota*, seleccionando los valores no NULL. Se coloca en filtros el campo reparto y dirección según el caso, descartando también los NULL en ambos casos. Se selecciona en el panel *Mostrarme* los diagramas de campos y valores.

Se obtienen los siguientes resultados:

Reparto

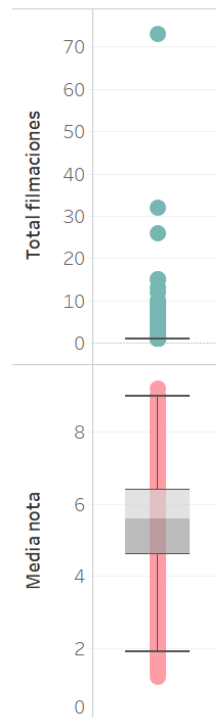


Ilustración 15: reparto

Dirección

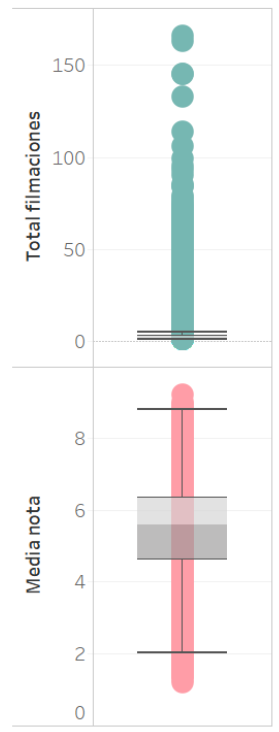


Ilustración 16: dirección

En el caso del reparto, se aprecia como la tendencia es que cada combinación de reparto haga únicamente una película. Este resultado es lógico. Por otra parte, se ven algunos

valores atípicos de actores que han participado en más de un filme registrado en filmaffinity, con hasta un máximo de 73 películas. Por otra parte, la nota media de los filmes coincide en ambas gráficas (lo que es lógico), pero se puede conocer qué reparto participó en el filme mejor valorado, en el peor valorado...

En el caso de la dirección, se ve cómo la tendencia es que un director dirija entre 1 y 5 películas (mayoritariamente 1 frente a 5). Sin embargo, hay muchos valores atípicos, llegando a un máximo de 166 filmes dirigidos por un director concreto. Por otra parte, se puede conocer qué director ha dirigido el filme mejor valorado, peor valorado...

ANÁLISIS DE LOS DASHBOARDS

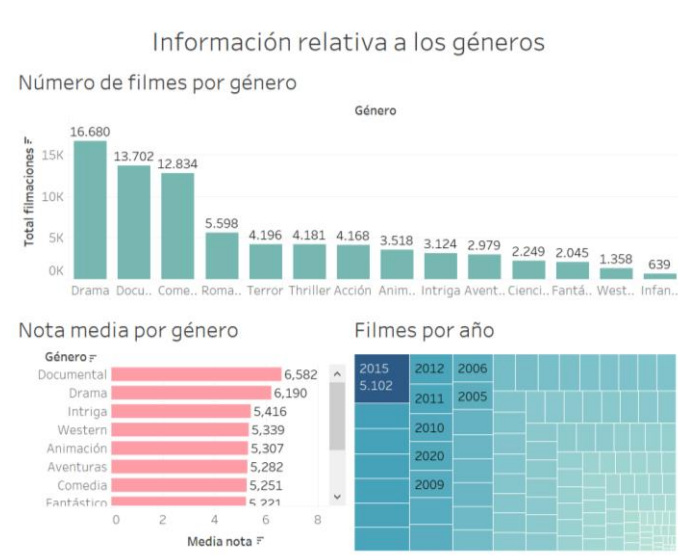


Ilustración 17: dashboard información relativa a los géneros

En apartados anteriores, ya se habían analizado los resultados de forma individual sobre los gráficos: *número de filmes por género*, *nota media por género* y *filmes por año*. Sin embargo, este dashboard se ha construido de modo que el gráfico *filmes por año*, actúa como filtro para las demás gráficas. Así, se podrá analizar el número de filmes por género o su nota media en un año o varios años concretos.

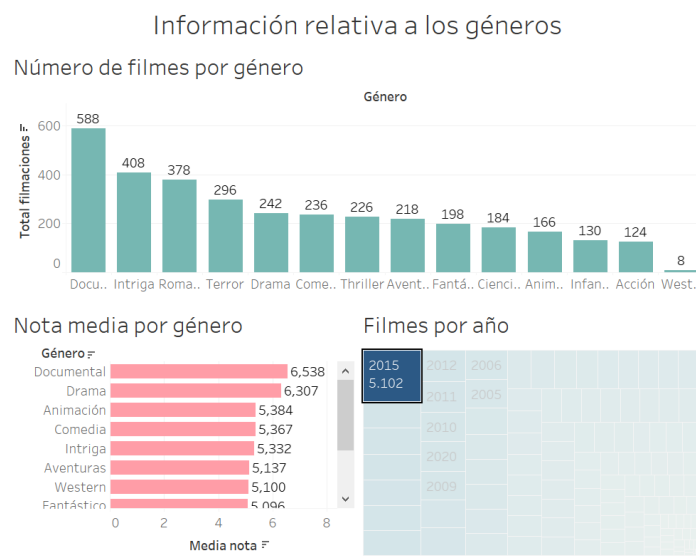


Ilustración 18: ejemplo uso de filtro en el dashboard 1

En la ilustración 15 se muestra un ejemplo de uso de los filtros. Se selecciona el año 2015, en el que el género en el que se han registrado más filmes es el *Documental*, así como el mejor valorado. La puntuación media máxima dada por los usuarios ese año ha sido de 6.538.

El segundo dashboard recogerá las gráficas: *nota media por país*, *evolución tipo de filmes*, *5 mejores valoraciones* y *5 peores valoraciones*.

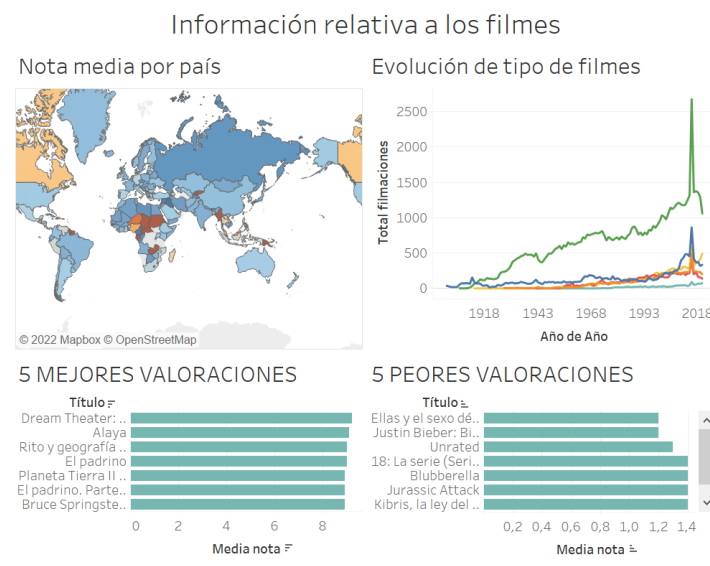


Ilustración 19: dashboard información relativa a los filmes

Cada gráfica ha sido analizada individualmente. Sin embargo, para analizar los resultados de otra forma, se ha seleccionado la gráfica *Evolución de tipo de filmes* como filtro para

las gráficas de este dashboard, para poder analizar la nota obtenida en distintos países o las 5 peores y mejores valoraciones, pero para un tipo de filme concreto.

Un ejemplo del uso de este filtro se muestra en la siguiente imagen, en la que se filtran las gráficas para solo tratar con películas. Así, por una parte se obtiene que, de entre los países con mejores puntuaciones de películas está Suecia, que la máxima puntuación dada a una película a lo largo de los años es un 8,9 y, que la más baja es de un 1,8.

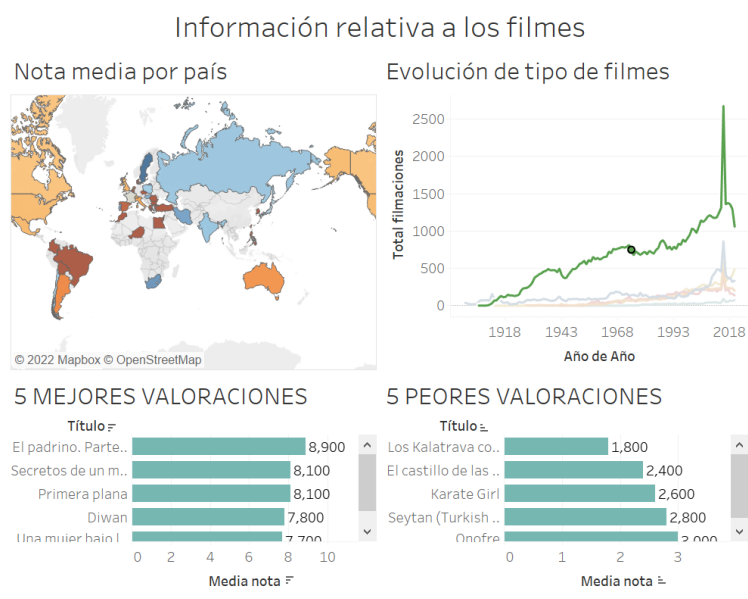


Ilustración 20: ejemplo de uso de filtro en el dashboard 2

En el último dashboard, que recoge información sobre el reparto y dirección de los filmes, simplemente recoge los dos últimos gráficos creados: *reparto* y *dirección*. No se ha añadido ningún filtro a mayores.

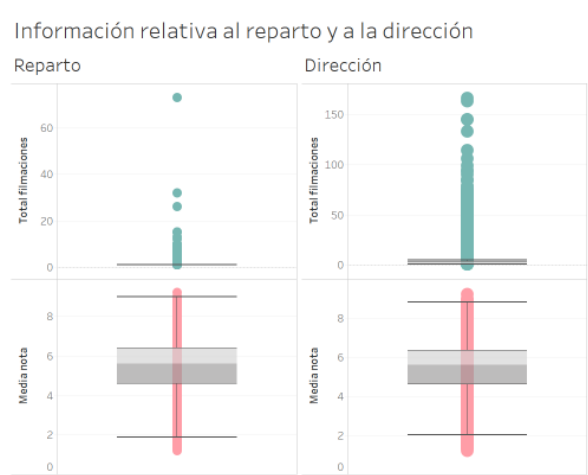


Ilustración 21: dashboard de información relativa al reparto y a la dirección

CONCLUSIONES

Tras el trabajo realizado, se puede afirmar que Tableau resulta una herramienta muy útil para visualizar datos. Resulta muy intuitivo realizar distintos tipos de visualizaciones: el mecanismo de arrastrar los campos a los filtros, columnas y filas, crear campos calculados con funciones que resultan conocidas, seleccionar el tipo de gráfico desde la pestaña *Mostrarme...* Además, resulta más sencilla que las herramientas con las que se ha trabajado anteriormente. Sin embargo, la personalización de los dashboards es más limitada que en Power Bi (no se puede elegir las interacciones entre dos gráficos concretos sin afectar a otros repetidos de otros dashboards, no se puede colocar los elementos en cualquier lugar del lienzo..., entre otros).

Se concluye, por tanto, que Tableau es una herramienta muy adecuada para crear visualizaciones apropiadas de los datos.

BIBLIOGRAFÍA

- [1] Fuente de datos. *Filmaffinity Dataset (Spanish)* (por Iván González, actualizado en noviembre de 2022). <https://www.kaggle.com/datasets/gan2gan/filmaffinity-dataset-spanish> (última visita realizada el 5/12/2022).
- [2] Página de la que [1] obtiene los datos. <https://www.filmaffinity.com/es/main.html> (última visita realizada el 5/12/2022).