Coderhouse – Data Analytics
Tercera entrega del Proyecto Final

IMDB: *Top 130 según la calificación de los votantes regulares de IMDB.*

Fecha de entrega: 21/02/2023
Comisión: 32650
Alumno: Santiago Ribon

ÍNDICE

Descripción de la temática de los datos

IMDB, Internet Movie Database (Base de Datos de películas en Internet) es la base de datos de películas más grande y completa de la web que almacena información relacionada con películas, personal del equipo de producción, directores, actores, bandas sonoras, series de televisión, programas de televisión, videojuegos, actores de doblaje, reseñas, avances de películas, críticas cinematográficas y, más recientemente, personajes ficticios que aparecen en los medios de entretenimiento visual.

Considerado como uno de los sitios más importantes de cine y televisión, IMDb, originalmente, nació oficialmente en 1990, y en 1998 fue adquirida por Amazon.

El dataset que se utilizará en este proyecto contiene el top de las 130 películas con mejor calificación hasta junio del 2022, detallando: Año de estreno, cantidad de votos, clasificación, certificación, duración, genero, sinopsis, actores, directores y escritores de estas películas.

Hipótesis

En el top de las películas mejor calificadas predominan producciones y elencos de Hollywood, aunque también se pueden observar producciones del resto del mundo. Puede observarse que también la mayoría de las películas pertenecen a estrenos del siglo XX, aunque también se pueden ver películas estrenadas el año pasado, habiendo actores, directores y escritores que repitieron su éxito tanto en este siglo como en el siglo pasado.

Alcance, usuario final y nivel de aplicación

La idea es analizar en qué año se estrenaron más películas del top, quienes son los directores que más películas dirigieron, quienes son los actores que más películas protagonizaron y cuáles fueron las películas más votadas, es decir, las más populares.

El resultado de este análisis nos permitirá observar quienes fueron los actores y directores que más se destacan a lo largo de estos años y cuál fue el año “dorado” de la industria del cine.

Este análisis podría ser utilizado por diferentes productoras cinematográficas para guiarse y tomar como ejemplo el trabajo que hicieron los equipos más exitosos del top y, de tratarse de una productora que cuenta con un presupuesto multimillonario, realizar contratación de integrantes de dichos equipos.

A su vez, Amazon también podría utilizar este análisis para estudiar la posibilidad de conseguir los derechos de algunas películas del top y así poder nutrir su plataforma de streaming Amazon Prime Video.

Dataset

Los dataset utilizados para este proyecto son:

- movies.csv
- actors.csv
- directors.csv
- writers.csv

Estos dataset estarán adjuntados en el siguiente link del drive:

https://drive.google.com/drive/folders/1vlzvt9xL1DdUhEm_aR7E4l_TCBa7dvVk?usp=sharing

Diagrama entidad-relación inicial

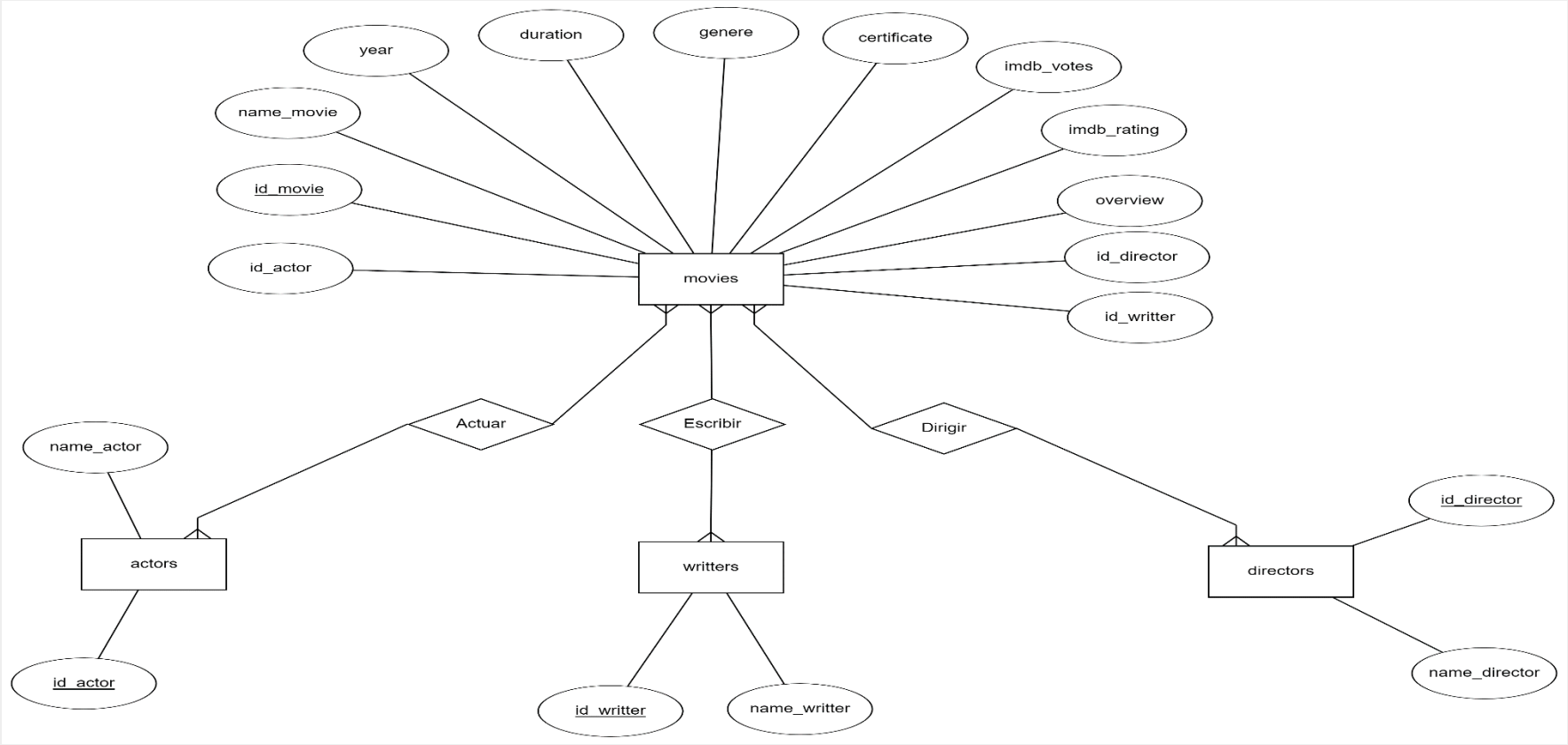
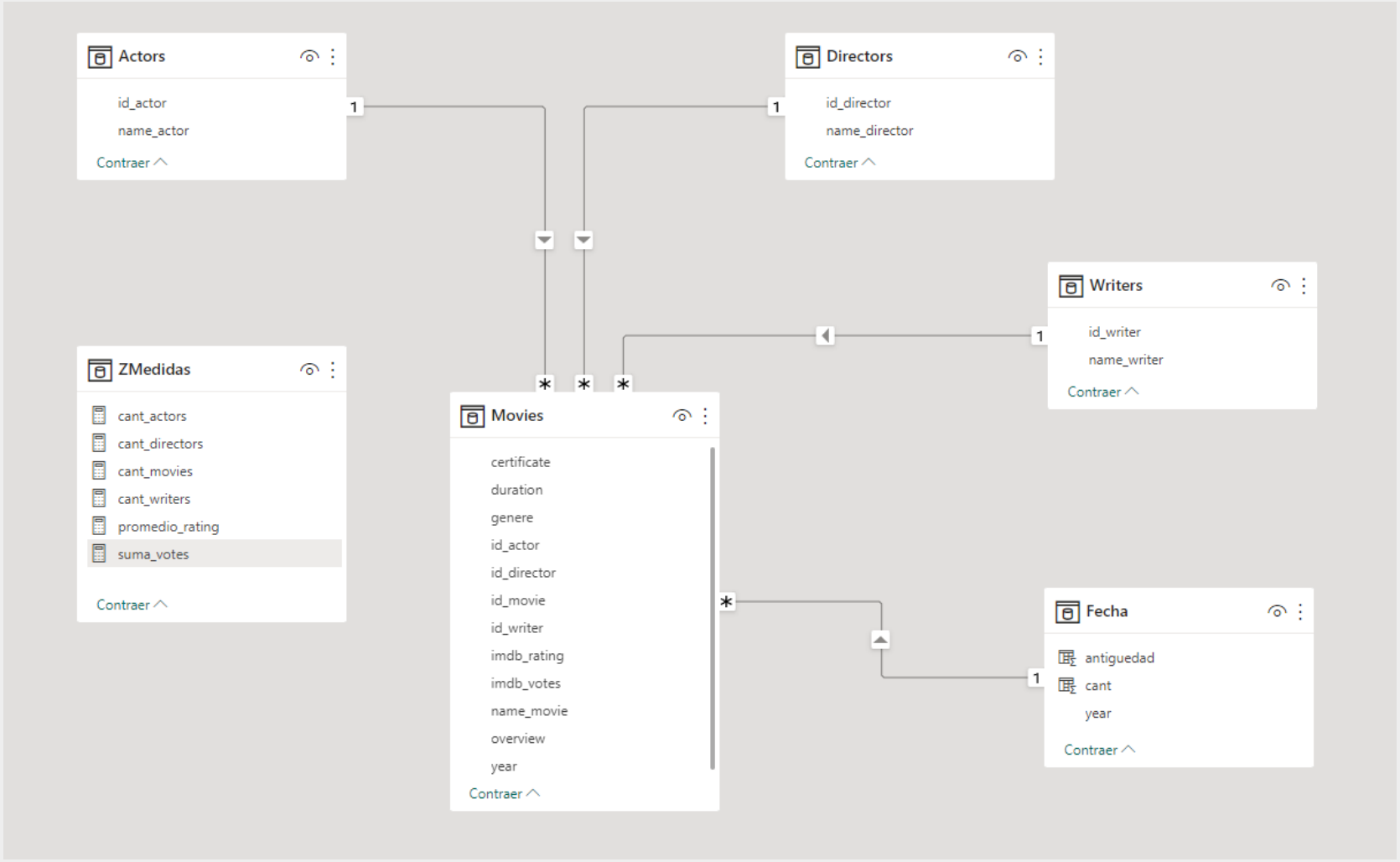


Diagrama entidad-relación final



Listado de tablas final

1. Movies: Esta tabla contiene los datos principales de las películas. Estos son: el código de identificación de la película, el nombre de la película, el año de estreno, cantidad de votos, calificación, certificación, duración, genero, sinopsis, código de identificación del actor, código de identificación del director y código de identificación del escritor.

Tipo de Clave	Campo	Tipo de dato	Detalle
PK	id_movie	VARCHAR(50)	Cada película tiene su código único de identificación.
-	name_movie	VARCHAR(100)	Nombre de la película.
FK	year	INT	Año de estreno de la película.
-	imdb_votes	INT	Cantidad de votos en IMDB.
-	imdb_rating	FLOAT	Clasificación en IMDB.
-	certificate	VARCHAR(50)	Calificación de audiencia de la película.
-	duration	INT	Duración calculada en minutos de la película.
-	genere	VARCHAR(50)	Genero/s de la película.
-	overview	VARCHAR(250)	Resumen muy breve y general de la película.
FK	id_actor	VARCHAR(50)	Cada actor tiene su código único de identificación. Este campo permitirá relacionar esta tabla con la tabla 'actors'.
FK	id_director	VARCHAR(50)	Cada director tiene su código único de identificación. Este campo permitirá relacionar esta tabla con la tabla 'directors'.
FK	id_writers	VARCHAR(50)	Cada escritor tiene su código único de identificación. Este campo permitirá relacionar esta tabla con la tabla 'writers'.

Listado de tablas final

2. Actors: En esta tabla se puede observar los datos de los actores. Estos son: el código de identificación del actor y el nombre completo.

Tipo de clave	Campo	Tipo de dato	Detalle
PK	id_actor	VARCHAR(50)	Cada actor tiene su código único de identificación.
-	name_actor	VARCHAR(50)	Nombre y apellido del actor.

3. Directors: Esta tabla contiene los datos de los directores. Estos son: el código de identificación del director y el nombre completo.

Tipo de clave	Campo	Tipo de dato	Detalle
PK	id_director	VARCHAR(50)	Cada director tiene su código único de identificación.
-	name_director	VARCHAR(50)	Nombre y apellido del director.

Listado de tablas final

4. Writers: Esta tabla presenta los datos de los escritores. Estos son: el código de identificación del escritor y el nombre completo.

Tipo de clave	Campo	Tipo de dato	Detalle
PK	id_writer	VARCHAR(50)	Cada escritor tiene su código único de identificación.
-	name_writer	VARCHAR(50)	Nombre y apellido del escritor.

5. Year: Esta tabla fue creada para obtener datos relacionados con el tiempo. Estos datos son: el año de estreno, los años transcurridos entre el estreno y el año actual, y la cantidad de películas estrenadas por año.

Tipo de clave	Campo	Tipo de dato	Detalle
PK	year	INT	Año de estreno de la pelicula.
-	antiguedad	INT	Años transcurridos entre el año de estreno y el año actual.
-	cant	INT	Cantidad de películas estrenadas por año.

Transformaciones realizadas

- Se les puso mayúscula a los nombres de las tablas.
- Se cambió el tipo de datos de las columnas year, imdb_votes, imdb_rating, duration. Todas ellas fueron designadas automaticamente por el programa con el tipo de dato Text y se las cambió a Numero entero.
- A las tablas Actors, Directors y Writers se les modificó el encabezado, ya que los nombres de las columnas formaban parte de la primera fila designado automáticamente por el programa.

Actors:

```
#"Encabezados promovidos" = Table.PromoteHeaders(#"Tipo cambiado", [PromoteAllScalars=true]),
#"Tipo cambiado1" = Table.TransformColumnTypes(#"Encabezados promovidos",{{"id_actor", type text}, {"name_actor", type text}})
```

Directors:

```
#"Encabezados promovidos" = Table.PromoteHeaders(#"Tipo cambiado", [PromoteAllScalars=true]),
#"Tipo cambiado1" = Table.TransformColumnTypes(#"Encabezados promovidos",{{"id_director", type text}, {"name_director", type text}})
```

Writers:

```
#"Encabezados promovidos" = Table.PromoteHeaders(#"Tipo cambiado", [PromoteAllScalars=true]),
#"Tipo cambiado1" = Table.TransformColumnTypes(#"Encabezados promovidos",{{"id_writer", type text}, {"name_writer", type text}})
```

Transformaciones realizadas

- Se duplicó la tabla Movies y se la renombró como Fecha, dejando solamente el campo year, y se eliminaron los duplicados de dicho campo.

```
#"Tipo cambiado" = Table.TransformColumnTypes(#"Encabezados promovidos",{{"year", Int64.Type}, {"imdb_votes", Int64.Type}, {"imdb_rating", type text}}),
#"Valor reemplazado" = Table.ReplaceValue(#"Tipo cambiado",".",",",Replacer.ReplaceText,{"imdb_rating"}),
#"Tipo cambiado1" = Table.TransformColumnTypes(#"Valor reemplazado",{{"imdb_rating", type number}, {"duration", Int64.Type}}),
#"Otras columnas quitadas" = Table.SelectColumns(#"Tipo cambiado1",{"year"}),
#"Duplicados quitados" = Table.Distinct(#"Otras columnas quitadas"),
#"Filas filtradas" = Table.SelectRows(#"Duplicados quitados", each true)
```

- En la tabla Fecha se crea una nueva columna llamada ‘antiguedad’ que calcula la diferencia entre el año de la fecha actual y el año de estreno.

```
antiguedad = YEAR(TODAY()) - Fecha[year]
```

- En la tabla Fecha se crea una nueva columna llamada ‘cant’ que cuenta la cantidad de películas estrenadas por año.

```
cant = CALCULATE(DISTINCTCOUNT(Movies[id_movie]))
```

Mediciones realizadas

- Se crea una medición llamada '*cant_actors*', que cuenta la cantidad de actores que hay en la tabla Actors.

  cant_actors = COUNT(Actors[id_actor])
- Se crea una medición llamada '*cant_directors*', que cuenta la cantidad de directores que hay en la tabla Directors.

  cant_directors = COUNT(Directors[id_director])
- Se crea una medición llamada '*cant_movies*', que cuenta la cantidad de películas que hay en la tabla Movies.

  cant_movies = COUNT(Movies[id_movie])
- Se crea una medición llamada '*cant_writers*', que cuenta la cantidad de escritores que hay en la tabla Writers.

  cant_writers = COUNT(Writers[id_writer])
- Se crea una medición llamada '*promedio_rating*', que promedia las calificaciones de IMDB.

  promedio_rating = AVERAGE(Movies[imdb_rating])
- Se crea una medición llamada '*suma_votes*', que suma la cantidad de votos de IMDB.

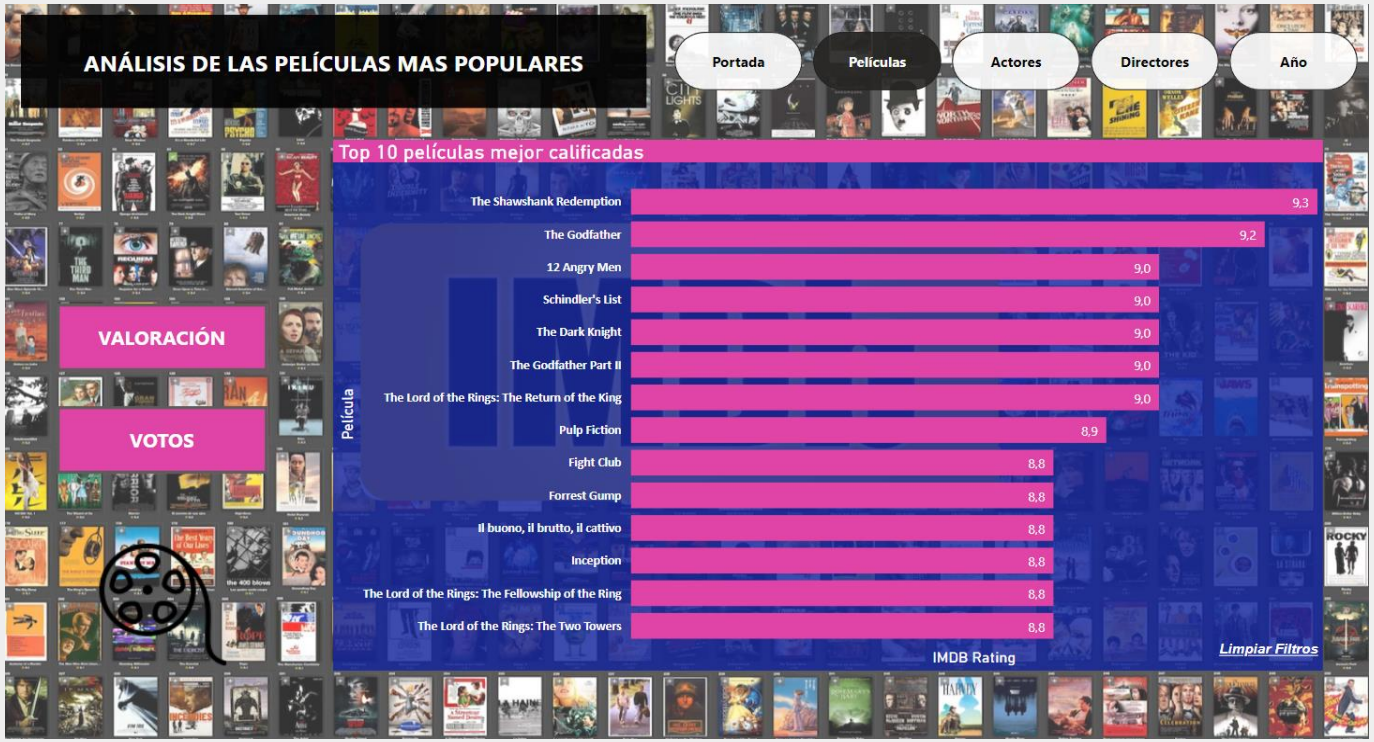
  suma_votes = SUM(Movies[imdb_votes])
- Finalmente se crea una tabla llamada *ZMedidas,* que contiene todas las medidas calculadas.

Dashboard y descripción

En la primer solapa se presenta la portada del tablero, con botones de navegación para cada una de las pestañas.

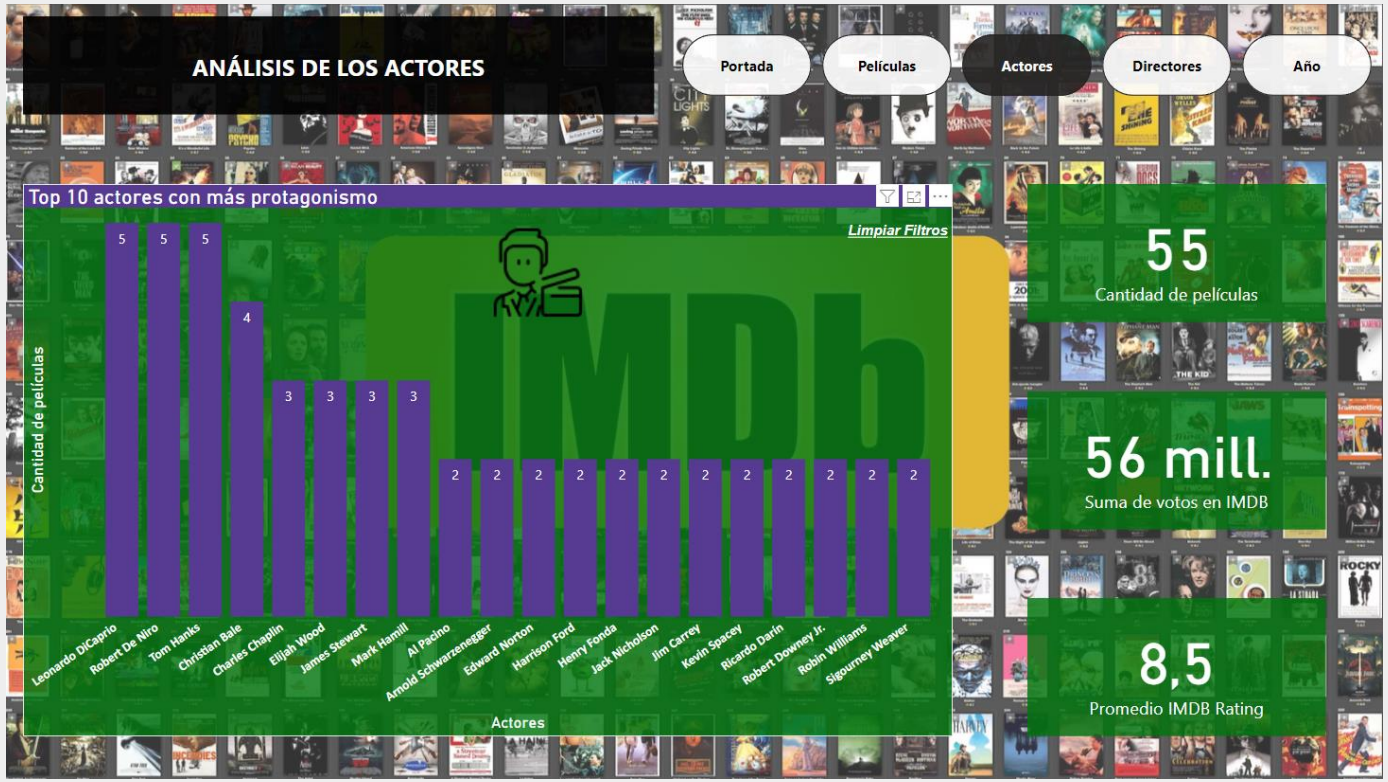
Dashboard y descripción

Esta solapa contiene el análisis de las películas más populares. Es decir, las más votadas y las mejores valoradas. Se podrá filtrar por cualquiera de estas dos opciones y mediante un tooltip se podrá acceder información detallada sobre las películas.



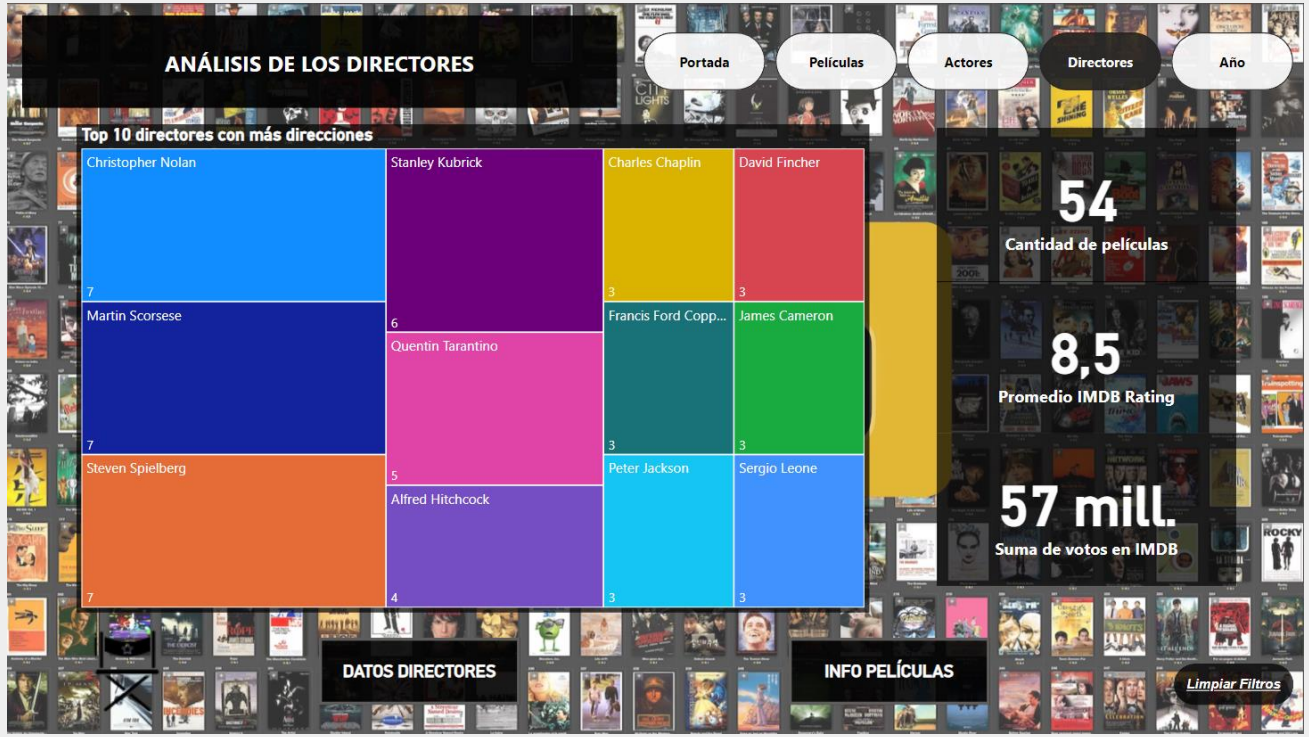
Dashboard y descripción

La tercera solapa contiene el análisis de los actores, donde se podrá observar los actores que mas películas del top protagonizaron. Mediante un tooltip se podra acceder a datos de las películas. Al filtrar por actor podremos ver detalladamente en las tarjetas las estadísticas de los actores.



Dashboard y descripción

En la cuarta solapa se puede observar el análisis de los directores que más películas dirigieron. Al filtrar por director en ‘DATOS DIRECTORES’ se representará en las tarjetas las estadísticas de estos. En ‘INFO PELÍCULAS’ se podrá acceder a una tabla con detalles de las películas que estos directores dirigieron.



Dashboard y descripción

En esta última solapa se encuentra el análisis por año de estreno. En el gráfico se puede observar la cantidad de películas estrenadas por año. Al filtrar por uno de estos se observará en las tarjetas las estadísticas de las películas y podremos ver detalles de estas en la tabla que se encuentra en la parte inferior.