

PRUEBA BI

Objetivo: Análisis, extracción, ingestión, transformación y visualización de la información de películas y shows de Netflix en Power BI.

Herramientas utilizadas:

- Anaconda (Jupyter)
- Google colab
- Python (Pandas, Numpy, Seaborn)
- MySQL
- Power Query
- Power BI
- Excel

ANACONDA (JUPYTER)

The screenshot shows a Jupyter Notebook interface with the following content:

```

# Análisis de la información de Netflix
## Importación de librerías
In [9]: import numpy as np
import pandas as pd
import seaborn as sns

## Conector a MySQL
In [3]: !pip install mysql-connector-python
Requirement already satisfied: mysql-connector-python in c:\users\karen\anaconda3\lib\site-packages (8.0.32)
Requirement already satisfied: protobuf<=3.20.3,>=3.11.0 in c:\users\karen\anaconda3\lib\site-packages (from mysql-connector-python) (3.20.3)

In [81]: from sqlalchemy import create_engine
In [82]: from pandas.io import sql
In [83]: engine = create_engine("mysql+pymysql://{}:{}@{}:{}/{}".format(host='localhost', db='netflix', user='root', pw='*****'))

## Creación y lectura de dataframes
In [5]: df_actores = pd.read_csv("Downloads/Power_BI/Actores.csv")
In [6]: df_data_netflix = pd.read_csv("Downloads/Power_BI/data_netflix.csv")
In [7]: df_mejores_peliculas_netflix = pd.read_csv("Downloads/Power_BI/mejores_peliculas_netflix.csv")
In [8]: df_mejores_shows_netflix = pd.read_csv("Downloads/Power_BI/mejores_shows_netflix.csv")

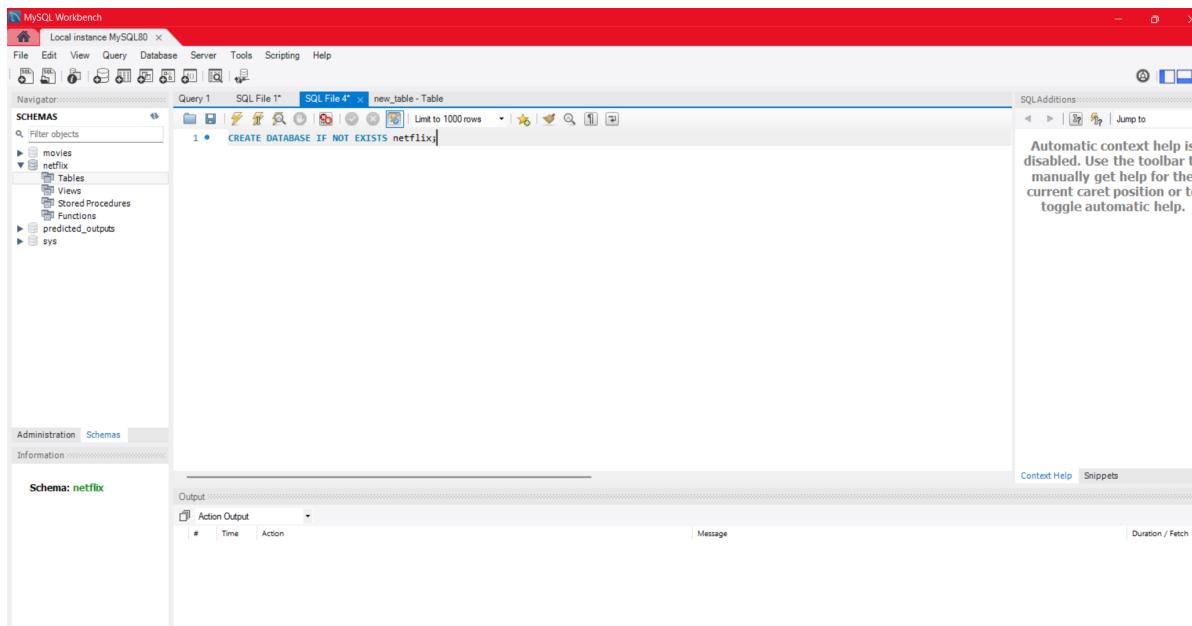
```

Se comparte la liga de Colab donde se puede encontrar el archivo:

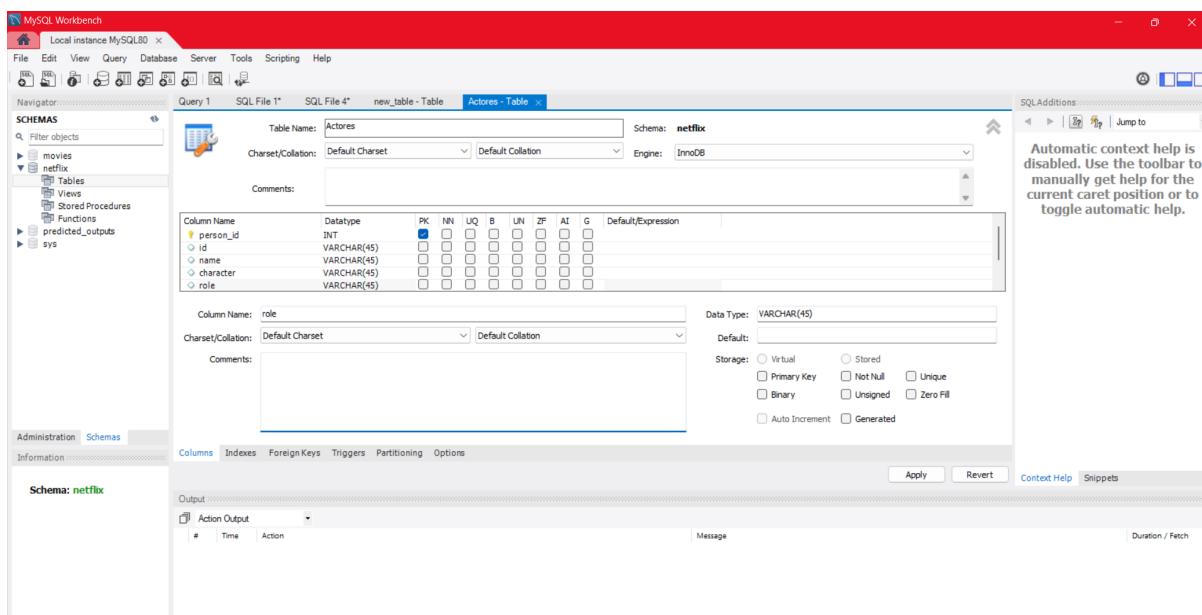
<https://colab.research.google.com/drive/1s8CKOuZAmhPDHSWH7ZO15puQiMGCmmTH?usp=sharing>

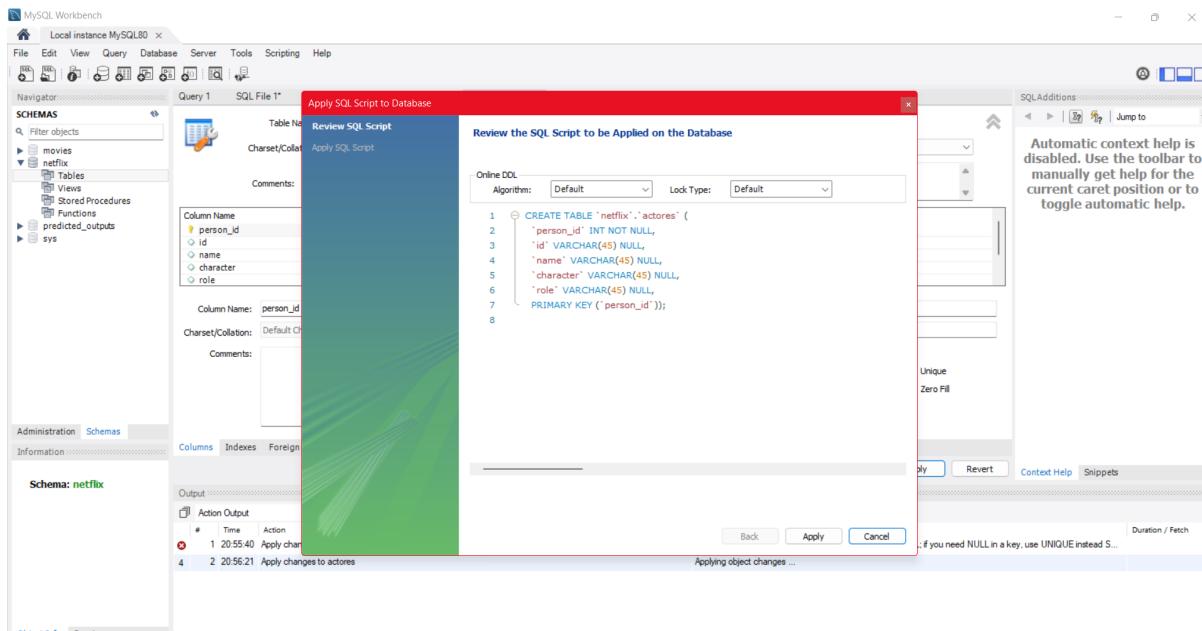
MYSQL

- Creación de la base de datos netflix en MySQL.



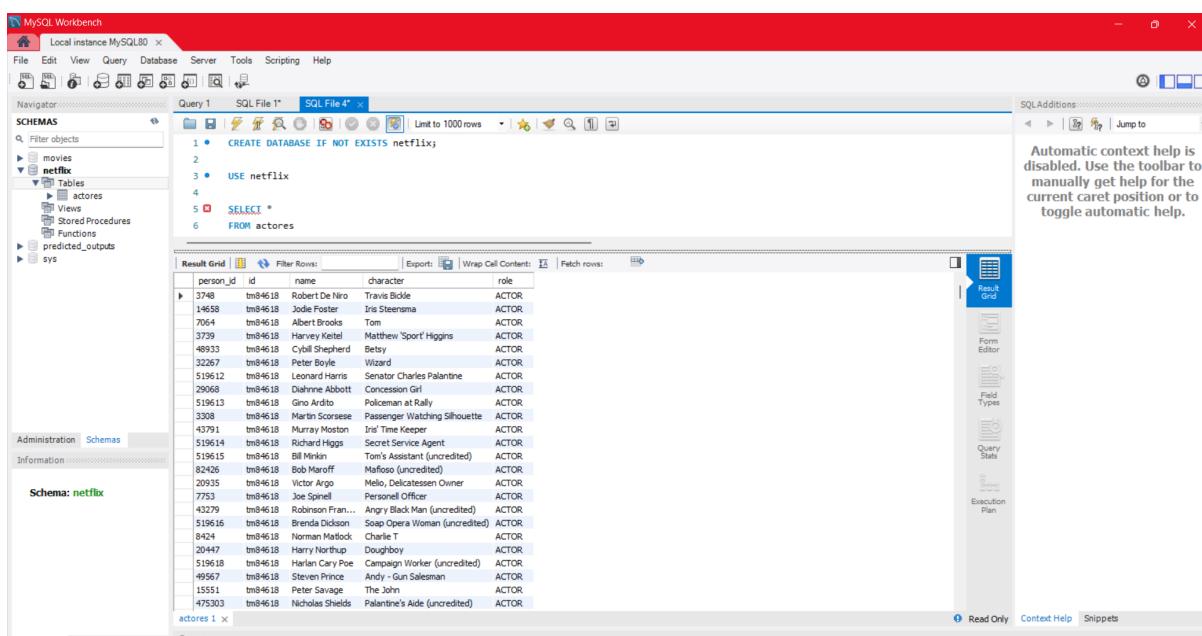
- Creación de la tabla Actores en MySQL dentro de la base de datos netflix



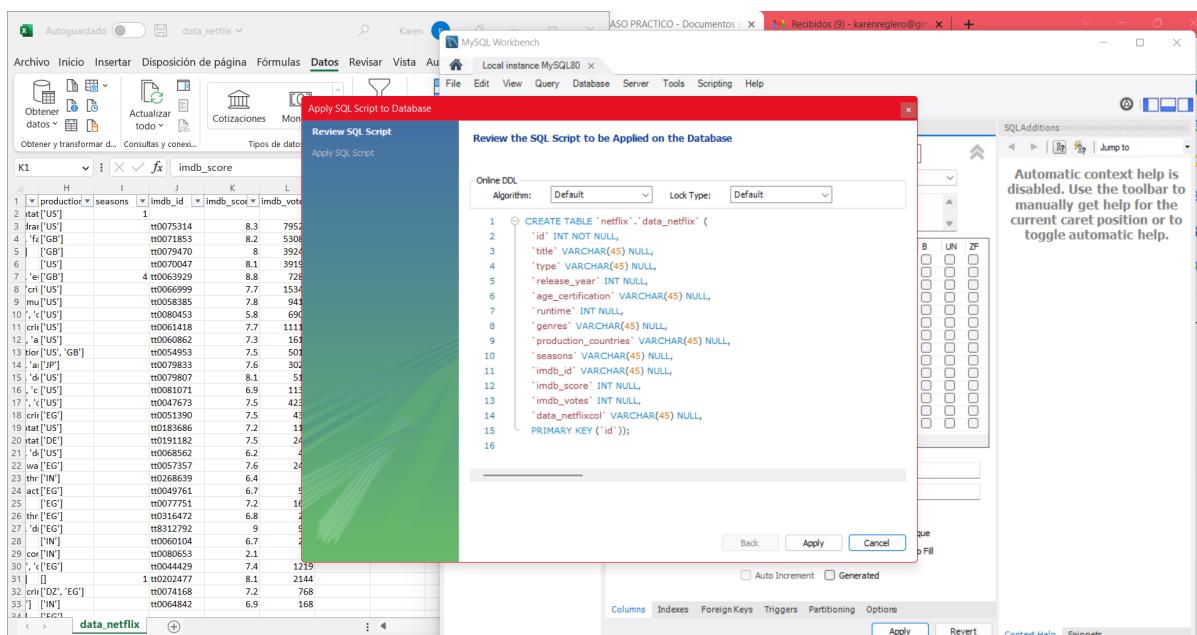
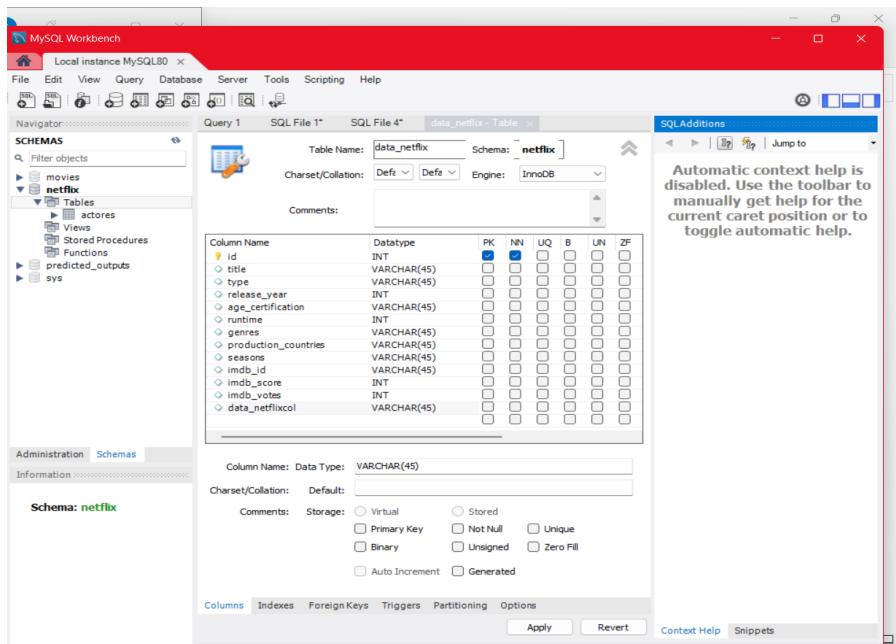


```
CREATE TABLE `netflix`.`actores` (
`person_id` INT NOT NULL,
`id` VARCHAR(45) NULL,
`name` VARCHAR(45) NULL,
`character` VARCHAR(45) NULL,
`role` VARCHAR(45) NULL,
PRIMARY KEY (`person_id`));
```

- SELECT de la tabla actores para confirmar que la información se ingestó de forma correcta:



- Creación de la tabla data_netflix en MySQL dentro de la base de datos netflix



```
CREATE TABLE `netflix`.`data_netflix` (
  `id` INT NOT NULL,
  `title` VARCHAR(45) NULL,
  `type` VARCHAR(45) NULL,
  `release_year` INT NULL,
  `age_certification` VARCHAR(45) NULL,
  `runtime` INT NULL,
  `genres` VARCHAR(45) NULL,
  `production_countries` VARCHAR(45) NULL,
  `seasons` VARCHAR(45) NULL,
  `imdb_id` VARCHAR(45) NULL,
```

```
'imdb_score` INT NULL,
`imdb_votes` INT NULL,
`data.netflixcol` VARCHAR(45) NULL,
PRIMARY KEY (`id`);
```

- SELECT de la tabla data.netflix para confirmar que la información se ingestó de forma correcta:

The screenshot shows the MySQL Workbench interface with the following details:

- Navigator:** Shows the schema structure, including the movies and netflix databases, and their respective tables (actores, data.netflix).
- Query Editor:** Contains the following SQL code:


```
4
5  SELECT *
6   FROM actores
7
8  SELECT *
9   FROM data.netflix
```
- Result Grid:** Displays the results of the query, listing numerous movies from the data.netflix table. The columns include id, title, type, release_year, age_certification, runtime, genres, production_countries, seasons, and imdb_id.
- SQLAdditions:** A panel on the right provides context help for the current caret position.

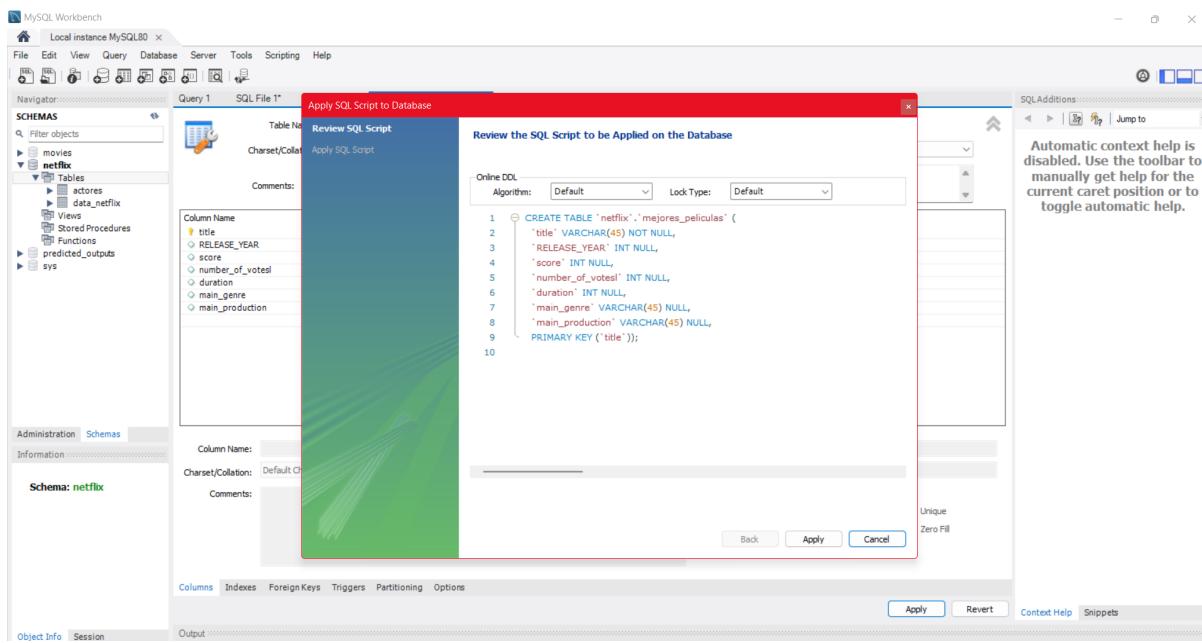
- Creación de la tabla mejores_peliculas en MySQL dentro de la base de datos netflix

The screenshot shows the MySQL Workbench interface with the following details:

- Navigator:** Shows the schema structure, including the movies and netflix databases, and their respective tables (actores, data.netflix).
- Query Editor:** Contains the following SQL code to create the table:


```
Table Name: mejores_peliculas
Schema: netflix
CharSet/Collation: Default Charset
Engine: InnoDB
```
- Table Definition:** The table structure is defined with the following columns:

Column Name	Datatype	PK	NN	UQ	B	UN	ZF	AI	G	Default/Expression
title	VARCHAR(45)	<input checked="" type="checkbox"/>	<input type="checkbox"/>							
RELEASE_YEAR	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
score	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
number_of_votes	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
duration	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
main_genre	VARCHAR(45)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
main_production	VARCHAR(45)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
- SQLAdditions:** A panel on the right provides context help for the current caret position.



```

CREATE TABLE `netflix`.`mejores_peliculas` (
  `title` VARCHAR(45) NOT NULL,
  `RELEASE_YEAR` INT NULL,
  `score` INT NULL,
  `number_of_votes` INT NULL,
  `duration` INT NULL,
  `main_genre` VARCHAR(45) NULL,
  `main_production` VARCHAR(45) NULL,
  PRIMARY KEY (`title`));
  
```

- **SELECT** de la tabla `mejores_peliculas` para confirmar que la información se ingestó de forma correcta:

The screenshot shows the MySQL Workbench interface with the 'Schemas' tree on the left showing the 'netflix' schema. The 'Query 1' tab contains the following SQL code:

```

SELECT *
FROM data.netflix
  
```

The results grid displays a list of movies from the 'data.netflix' table, including columns like TITLE, RELEASE_YEAR, SCORE, NUMBER_OF_VOTES, DURATION, MAIN_GENRE, and MAIN_PRODUCTION. The results are as follows:

TITLE	RELEASE_YEAR	SCORE	NUMBER_OF_VOTES	DURATION	MAIN_GENRE	MAIN_PRODUCTION
David Attenborough: A Life on Our Planet	2020	9	31180	83	documentary	GB
Inception	2010	8.8	2265288	148	sci-fi	GB
Forrest Gump	1994	8.8	1999999	142	drama	US
American Beauty	2003	8.7	20395	160	comedy	IN
Be Burnham: Inside	2021	8.7	14974	87	comedy	US
Saving Private Ryan	1998	8.6	1346020	169	drama	US
Django Unchained	2012	8.4	1472668	165	western	US
Dangal	2016	8.4	180247	161	action	IN
Be Burnham: Make Happy	2016	8.4	14356	60	comedy	US
Louis C.K.: Hilarious	2010	8.4	11973	84	comedy	US
Dave Chappelle: Sticks & Stones	2019	8.4	25687	65	comedy	US
3 Idiots	2009	8.4	385782	170	comedy	IN
Black Friday	2004	8.4	20611	143	crime	IN
Super Deluxe	2019	8.4	1368	176	thriller	IN
Winter on Fire: Ukraine's Fight for Freedom	2015	8.3	17710	98	documentary	UA
Once Upon a Time in America	1984	8.3	342335	229	drama	US
Taxi Driver	1976	8.3	79522	113	crime	US
Like Stars on Earth	2007	8.3	188234	165	drama	IN
Be Burnham: What.	2013	8.3	11488	60	comedy	US
Full Metal Jacket	1987	8.3	723306	116	drama	GB
Warrior	2011	8.2	463276	140	drama	US
Drishyam	2015	8.2	79075	163	thriller	IN
Queen	2014	8.2	64805	146	drama	IN
Paan Singh Tomar	2012	8.2	35888	135	drama	IN

- Creación de la tabla mejores_shows en MySQL dentro de la base de datos netflix

The screenshot shows two instances of MySQL Workbench. The top instance displays the 'mejores_shows' table creation dialog. The table has the following structure:

Column Name	Datatype	PK	NN	UQ	B	UN	ZF	AI	G	Default/Expression
title	VARCHAR(45)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>						
release_year	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
score	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
number_of_votes	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
duration	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
number_of_seasons	INT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
main_genre	VARCHAR(45)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
main_production	VARCHAR(45)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

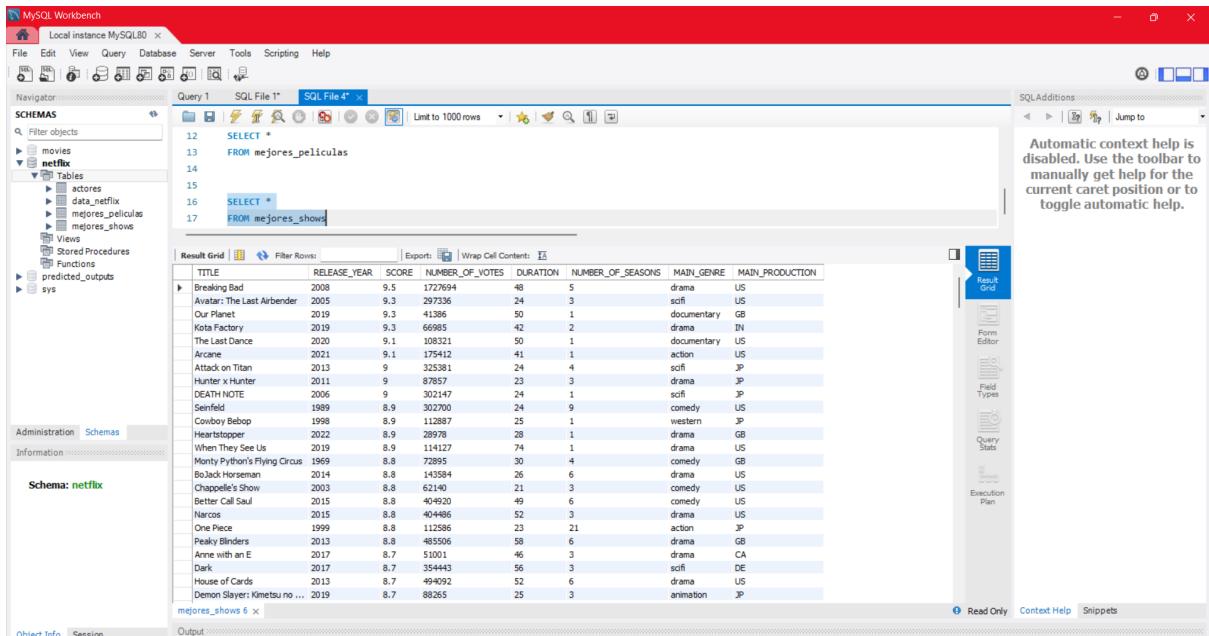
The bottom instance shows the 'Apply SQL Script to Database' dialog, displaying the generated SQL code:

```

CREATE TABLE `netflix`.`mejores_shows` (
  `title` VARCHAR(45) NOT NULL,
  `release_year` INT NULL,
  `score` INT NULL,
  `number_of_votes` INT NULL,
  `duration` INT NULL,
  `number_of_seasons` INT NULL,
  `main_genre` VARCHAR(45) NULL,
  `main_production` VARCHAR(45) NULL,
  PRIMARY KEY (`title`)
);

```

- SELECT de la tabla mejores_shows para confirmar que la información se ingestó de forma correcta:



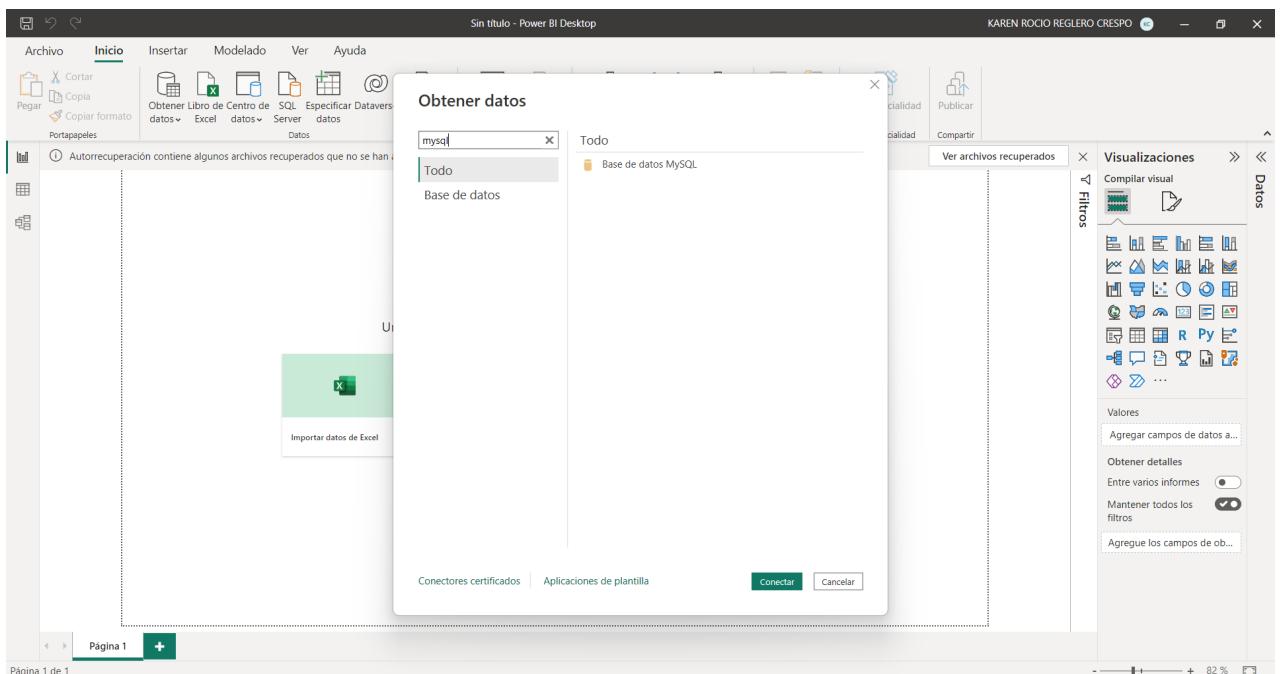
The screenshot shows the MySQL Workbench interface with the following details:

- File Bar:** File, Edit, View, Query, Database, Server, Tools, Scripting, Help.
- Schemas:** Local instance MySQLBD, Schemas (movies, netflix, sys), Tables (actores, data_netflix, mejores_peliculas, mejores_shows).
- Query Editor:** SQL File 1*, SQL File 4* (contains the following code):

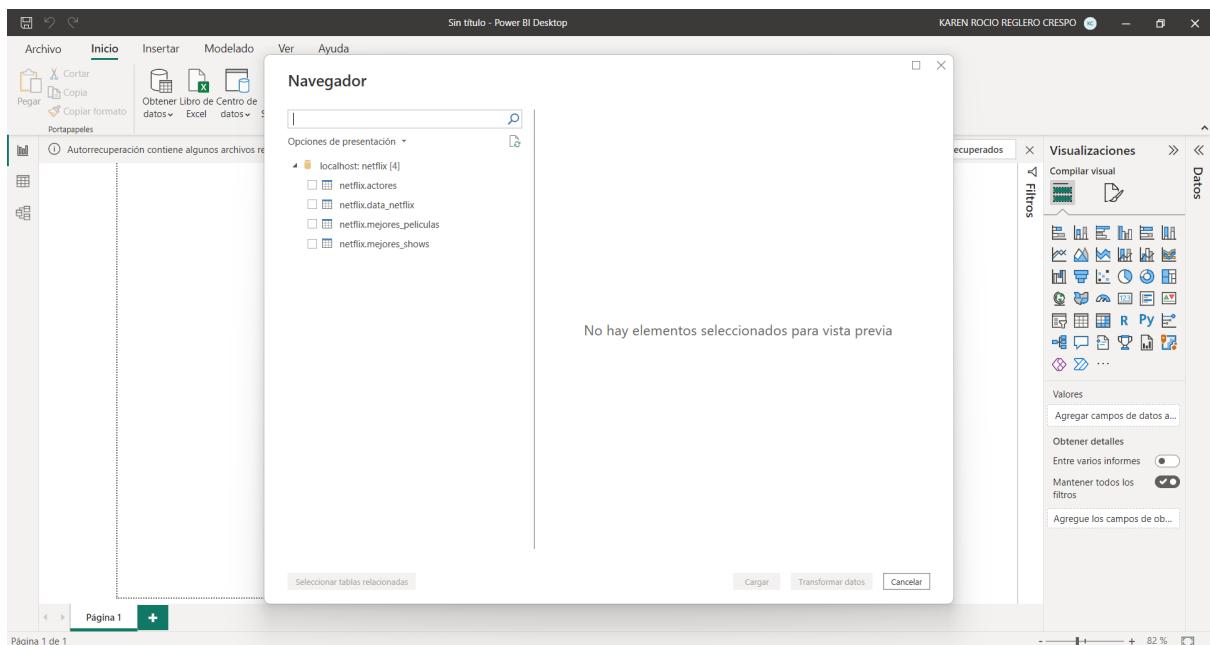
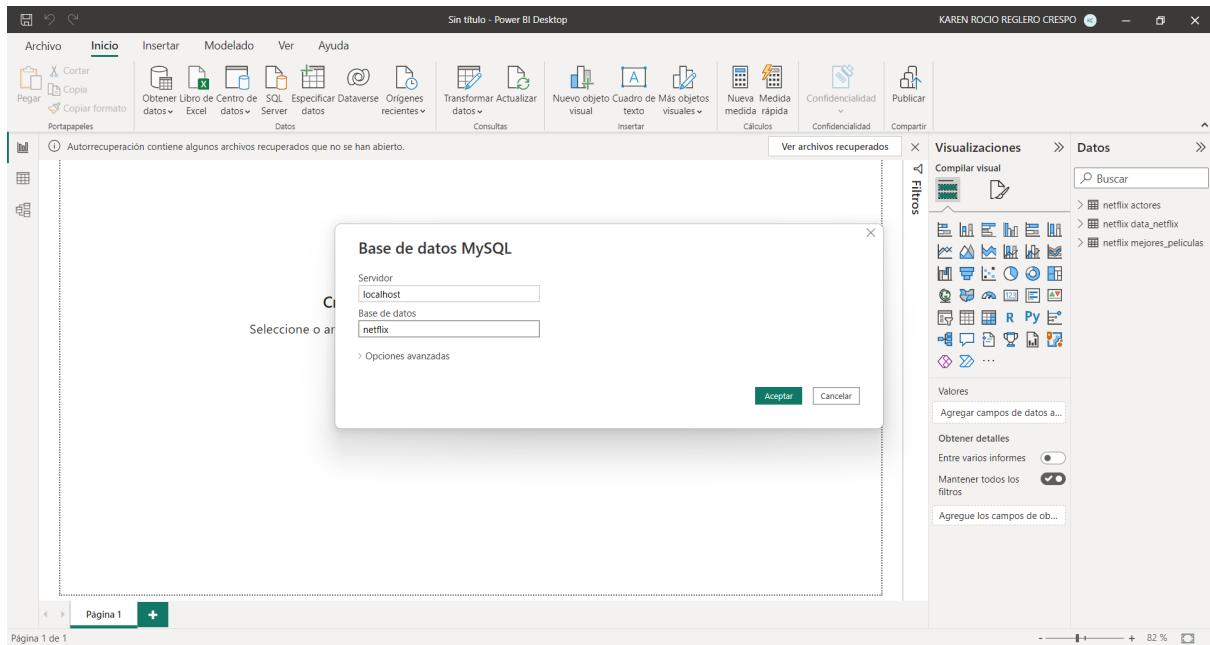

```

12  SELECT *
13  FROM mejores_peliculas
14
15  SELECT *
16  FROM mejores_shows
17
      
```
- Result Grid:** Displays the results of the SELECT query on the 'mejores_shows' table. The columns are: TITLE, RELEASE_YEAR, SCORE, NUMBER_OF_VOTES, DURATION, NUMBER_OF_SEASONS, MAIN_GENRE, and MAIN_PRODUCTION. The data includes rows for various TV shows like 'Breaking Bad', 'Avatar: The Last Airbender', 'Our Planet', etc.
- SQL Additions:** A panel on the right with the message: "Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help."
- Toolbar:** Includes icons for Run, Stop, Refresh, Save, and others.
- Status Bar:** Read Only, Context Help, Snippets.

Power BI (Conexión a MySQL)



- Conexión a la base de datos de netflix de MySQL



Se decidió ingestar toda la información de los 4 archivos provenientes de archivos csv a MySQL y posteriormente la conexión a Power BI, para su posterior limpieza y transformación de los datos.

POWER QUERY

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table named 'Table.RemoveColumns(netflix_actores, {"person_id"})'. The table has four columns: 'tm_id', 'name', 'character', and 'role'. The 'character' column is currently selected. The configuration pane on the right is titled 'Configuración de la consulta' (Query configuration). It shows the 'PROPIEDADES' (Properties) section with 'Nombre' (Name) set to 'netflix actores' and the 'PASOS APLICADOS' (Applied steps) section, which includes 'Origen' (Source) and 'Navegación' (Navigation), with 'Columnas quitadas' (Removed columns) listed.

LIMPIEZA DE DATOS

- Lo primero que se realizó fue cambiar el nombre de las columnas para cada conjunto de datos para identificar mejor el contenido de cada columna.

EJEMPLO: release_year → year
RELEASE_YEAR → year

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table named 'Table.RenameColumns(netflix_mejores_peliculas, {"TITLE", "title"}, {"RELEASE_YEAR", "year"}, {"SCORE", "score"}, {"NUMBER_OF_VOTES", "number_of_votes"}, {"DURATION", "duration"}, {"MAIN_GENRE", "main_genre"})'. The table has six columns: 'title', 'year', 'score', 'number_of_votes', 'duration', and 'main_genre'. The 'year' column is currently selected. The configuration pane on the right is titled 'Configuración de la consulta' (Query configuration). It shows the 'PROPIEDADES' (Properties) section with 'Nombre' (Name) set to 'netflix mejores_peliculas' and the 'PASOS APLICADOS' (Applied steps) section, which includes 'Origen' (Source) and 'Navegación' (Navigation), with 'Columnas con nombre cambi...' (Changed column names) listed.

- Se reemplazaron valores nulos por 0 para una mejor visualización.
 - Se removieron columnas innecesarias para este análisis:
 - person_id
 - character
 - age_certification
 - imdb_id

The screenshot shows the Microsoft Power Query Editor interface. The ribbon at the top has tabs for Archivo, Inicio, Transformar, Agregar columna, Vista, Herramientas, and Ayuda. The 'Transformar' tab is selected. The main area displays a table titled 'Table.RemoveColumns(#"Filas filtradas", "person_id")'. The table contains 28 rows of data about actors, with columns: Id, name, character, and role. The configuration pane on the right shows 'PROPIEDADES' with 'Nombre' set to 'netflix actores' and 'Todas las propiedades'. Under 'PASOS APLICADOS', 'Filas filtradas' is listed as a step, and 'Columnas quitadas' is highlighted with its value set to 'Valor reemplazado'.

TRANSFORMACIÓN DE DATOS

En esta fase del proceso, se realiza la transformación de 2 columnas principalmente:

- ❖ genres
 - ❖ production_countries

Se opta por realizar una anulación de dinamización de columnas para visualizar los resultados de ['drama', 'comedy'] → drama y comedy por separado.

Para lo cual, se agrega una columna personalizada como se muestra:

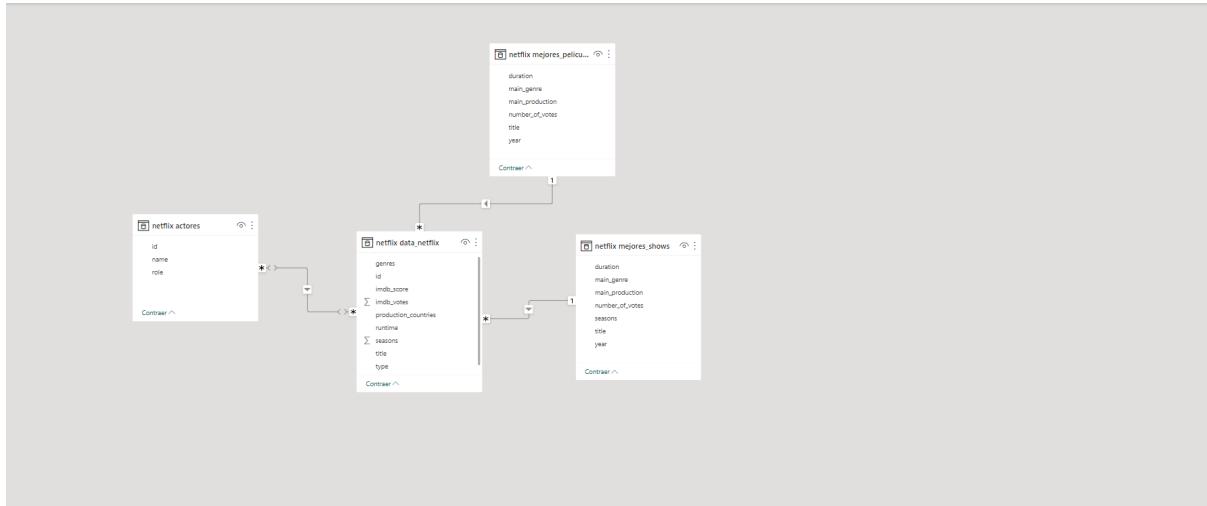
The screenshot shows the Microsoft Power Query Editor interface. A context menu is open over the 'genres' column, specifically at row 28. The menu path 'Transformar' > 'Añadir columna' > 'Columna personalizada...' leads to the 'Columna personalizada' dialog box. In this dialog, the new column name is set to 'genre', and the formula is defined as '=Text.Select([genres], {"A..","Z", " ", ","})'. The formula bar also shows the full formula: '=Text.Select([genres], {"A..","Z", " ", ","})'. The right side of the screen displays the 'Configuración de la consulta' pane, which includes sections for 'PROPIEDADES' (Nombre: netflix data.netflix) and 'PASOS APLICADOS' (Origen: Navegación, Valor reemplazado, Valor reemplazado1, Personalizada agregada).

- Posteriormente, se divide la columna por delimitador (,) y se realiza el despivoteo de la tabla nuevamente para contener valores únicos.
- Se realiza el mismo proceso con la columna production_countries:

The screenshot shows the Microsoft Power Query Editor interface. A context menu is open over the 'production_countries' column, specifically at row 28. The menu path 'Transformar' > 'Añadir columna' > 'Columna personalizada...' leads to the 'Columna personalizada' dialog box. In this dialog, the new column name is set to 'production_countries', and the formula is defined as '=Text.Select([production_countries], {"A..","Z", " ", ","})'. The formula bar also shows the full formula: '=Text.Select([production_countries], {"A..","Z", " ", ","})'. The right side of the screen displays the 'Configuración de la consulta' pane, which includes sections for 'PROPIEDADES' (Nombre: netflix data.netflix) and 'PASOS APLICADOS' (Origen: Navegación, Valor reemplazado, Valor reemplazado1, Personalizada agregada).

MODELADO EN POWER BI

- El modelado en Power BI de las 4 tablas queda de la siguiente forma:
 - Las tablas de actores y data_netflix se unen mediante el id
 - Las tablas data_netflix, mejores_películas y mejores_shows se unen mediante el title



EXCEL

- La herramienta Excel se utilizó como apoyo para comprobar algunos de los resultados generados en Power BI.

Actores						data_netflix					
A1	B1	C1	D1	E1	F1	A2	B2	C2	D2	E2	F2
person_id	name	character	role	type	release_year	title	type	release_year	age_certificate	runtime	
1	38632	tm67635	Shah Rukh Khan	ACTOR	1995	Five Came Back: The Reference Films	SHOW	1945	TV-MA	4	
2	38632	tm55100	Shah Rukh Khan	ACTOR	1976	Taxi Driver	MOVIE	1976	R	11	
3	38632	tm88593	Shah Rukh Khan	ACTOR	1975	Monty Python and the Holy Grail	MOVIE	1975	PG	9	
4	38632	tm12482	Shah Rukh Khan	ACTOR	1979	Life of Brian	MOVIE	1979	R	9	
5	38632	tm34502	Shah Rukh Khan	ACTOR	1980	The Exorcist	MOVIE	1980	R	13	
6	38632	tm34503	Shah Rukh Khan	ACTOR	1989	Monty Python's Flying Circus	SHOW	1989	TV-14	3	
7	38632	tm34504	Shah Rukh Khan	ACTOR	1991	Dirty Harry	MOVIE	1971	R	10	
8	38632	tm37601	Shah Rukh Khan	ACTOR	1994	My Fair Lady	MOVIE	1964	G	17	
9	38632	tm105716	Shah Rukh Khan	ACTOR	1997	The Blue Lagoon	MOVIE	1980	R	10	
10	38632	tm189773	Shah Rukh Khan	ACTOR	1999	Bonnie and Clyde	MOVIE	1967	R	11	
11	38632	tm58893	Shah Rukh Khan	ACTOR	2000	The Professionals	MOVIE	1966	PG-13	11	
12	38632	tm110225	Shah Rukh Khan	ACTOR	2002	The Guns of Navarone	MOVIE	1961		15	
13	38632	tm118438	Shah Rukh Khan	ACTOR	2003	Urin the Third: The Castle of Cagliostro	MOVIE	1970	PG	10	
14	38632	tm125000	Shah Rukh Khan	ACTOR	2004	Rio Bravo Live in Concert	MOVIE	1970	R	7	
15	38632	tm125058	Shah Rukh Khan	ACTOR	2005	The Lost Riders	MOVIE	1980	R	9	
16	38632	tm154578	Shah Rukh Khan	ACTOR	2006	White Christmas	MOVIE	1954	R	11	
17	38632	tm30170	Shah Rukh Khan	ACTOR	2007	Cairo Station	MOVIE	1958		7	
18	38632	tm82765	Shah Rukh Khan	ACTOR	2008	The Queen	MOVIE	1968		6	
19	38632	tm144345	Shah Rukh Khan	ACTOR	2009	Hilter: A Career	MOVIE	1977	PG	15	
20	38632	tm23441	Shah Rukh Khan	ACTOR	2010	FTA	MOVIE	1972	R	9	
21	38632	tm88366	Shah Rukh Khan	ACTOR	2011	Saladin the Victorious	MOVIE	1963		18	
22	38632	tm125070	Shah Rukh Khan	ACTOR	2012	Singapore	MOVIE	1960		15	
23	38632	tm125071	Shah Rukh Khan	ACTOR	2013	Dark Waters	MOVIE	1956		12	
24	38632	tm141117	Shah Rukh Khan	ACTOR	2014	Alexandria! Why?	MOVIE	1979		13	
25	38632	tm183037	Shah Rukh Khan	ACTOR	2015	Raya and Sasha	MOVIE	1953		10	
26	38632	tm176481	Shah Rukh Khan	ACTOR	2016	No Longer Kids	MOVIE	1979		23	
27	38632	tm365824	Shah Rukh Khan	ACTOR	2017	Amarapali	MOVIE	1966		12	
28	38632	tm232464	Shah Rukh Khan	ACTOR	2018						