Karen Rocío Reglero Crespo

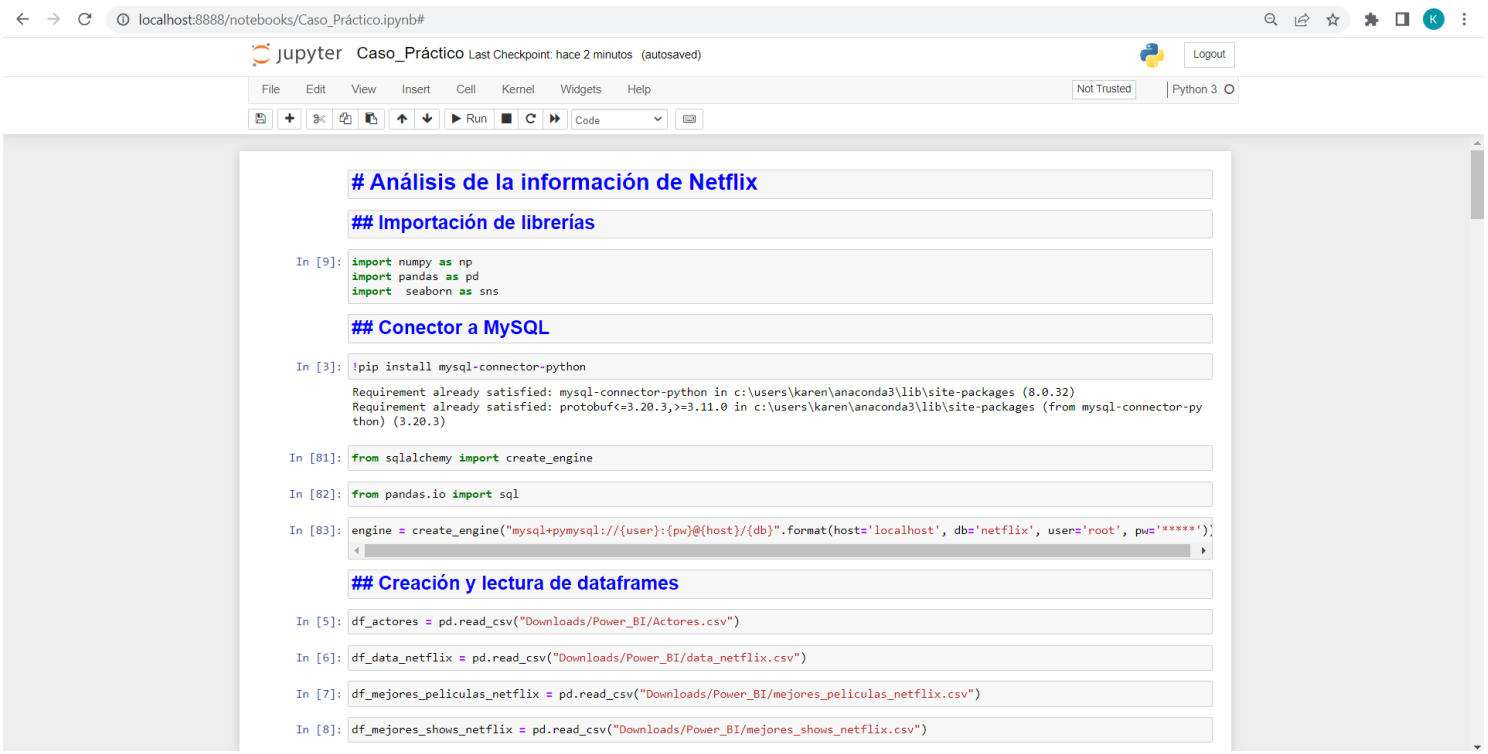
PRUEBA BI

Objetivo: Análisis, extracción, ingesta, 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

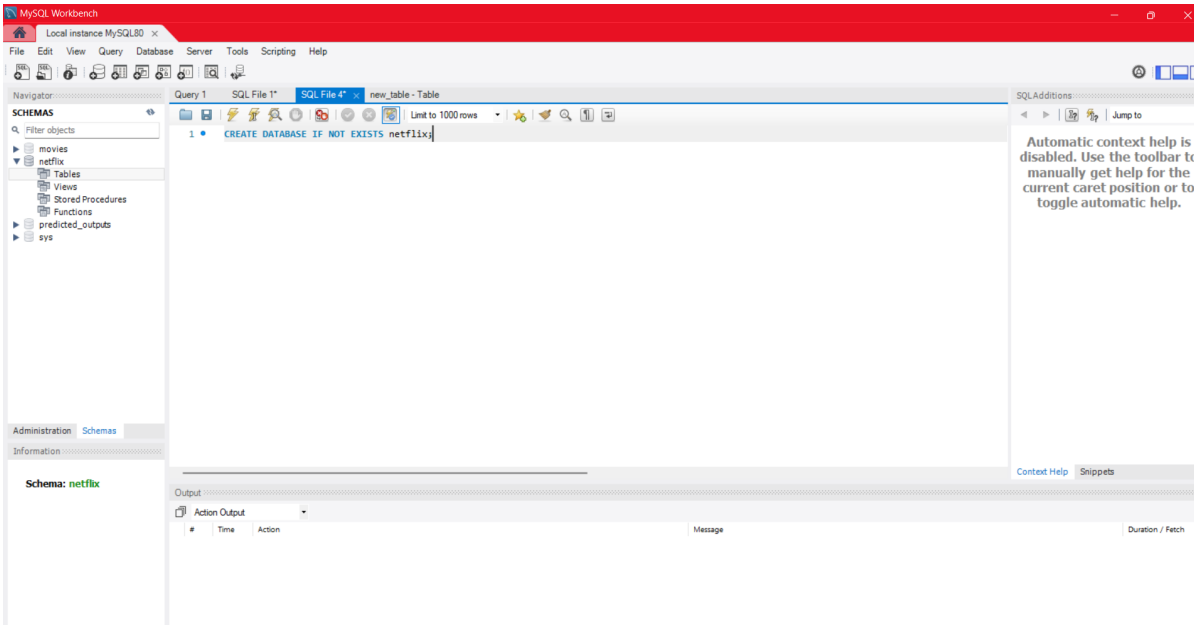
ANACONDA (JUPYTER)



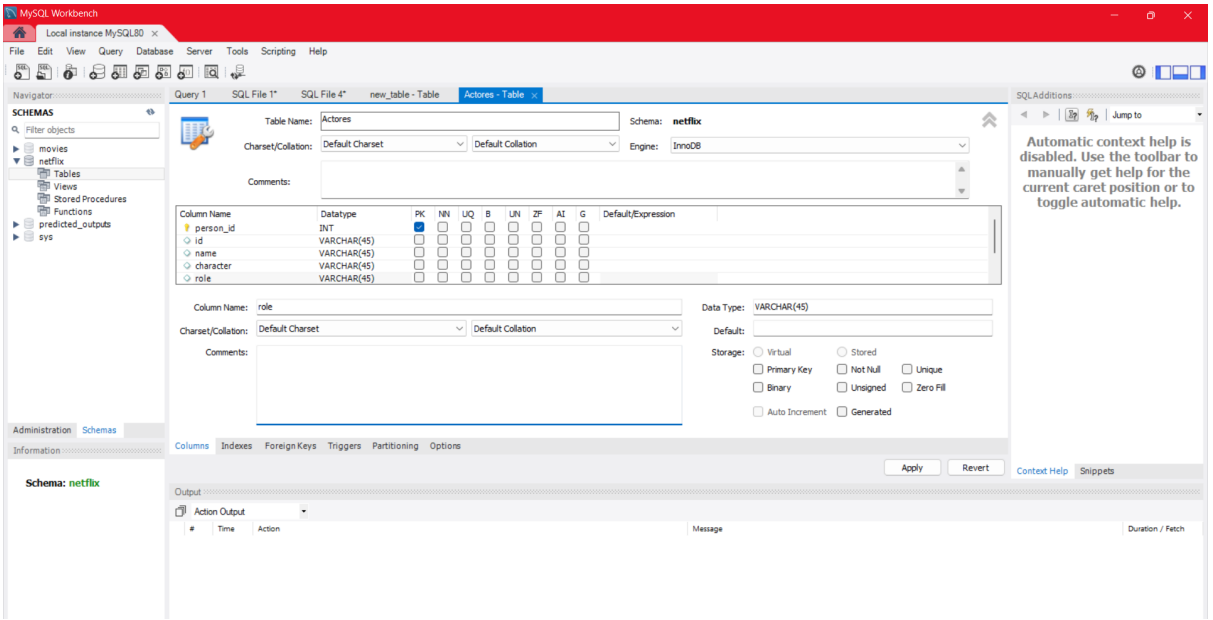
Se comparte la liga de Colab donde se puede encontrar el archivo: https://colab.research.google.com/drive/1s8CKOuZAmhPDHSWH7ZO15puQiMGCmmTH?usp=sharing

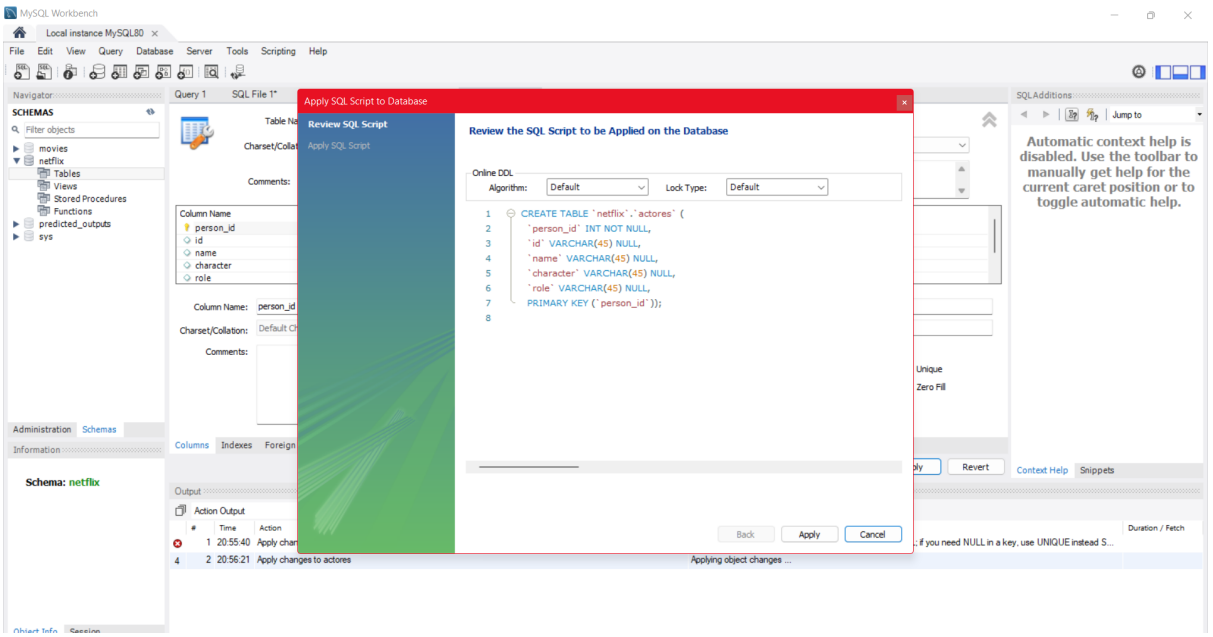
MYSQL

- Creación de la base de datos netflix en MySQL.



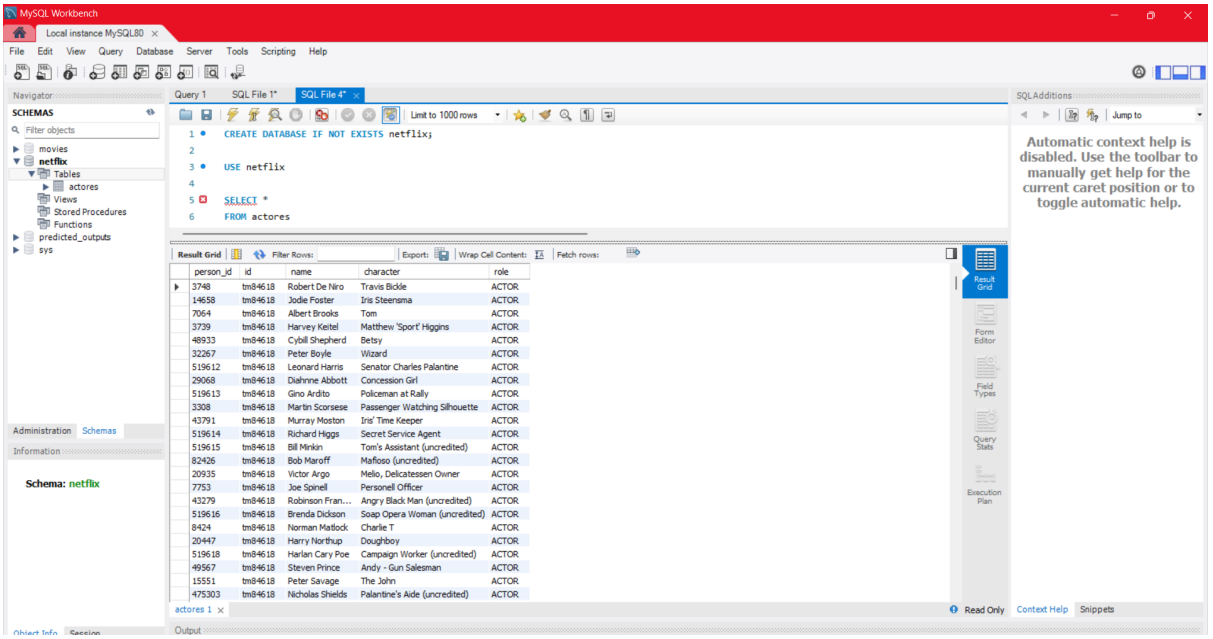
- Creación de la tabla Actores en MySQL dentro de la base de datos netflix



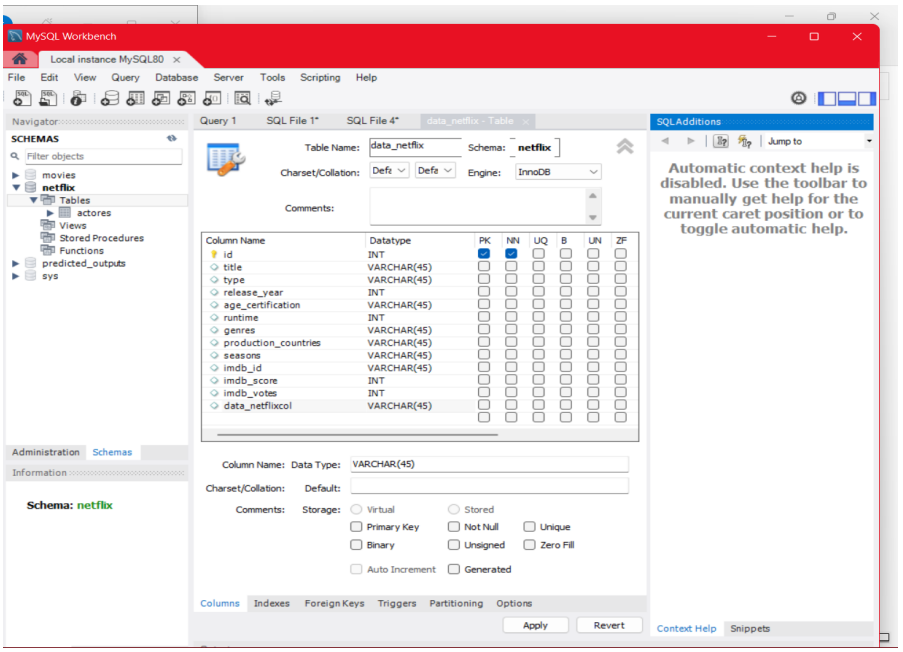


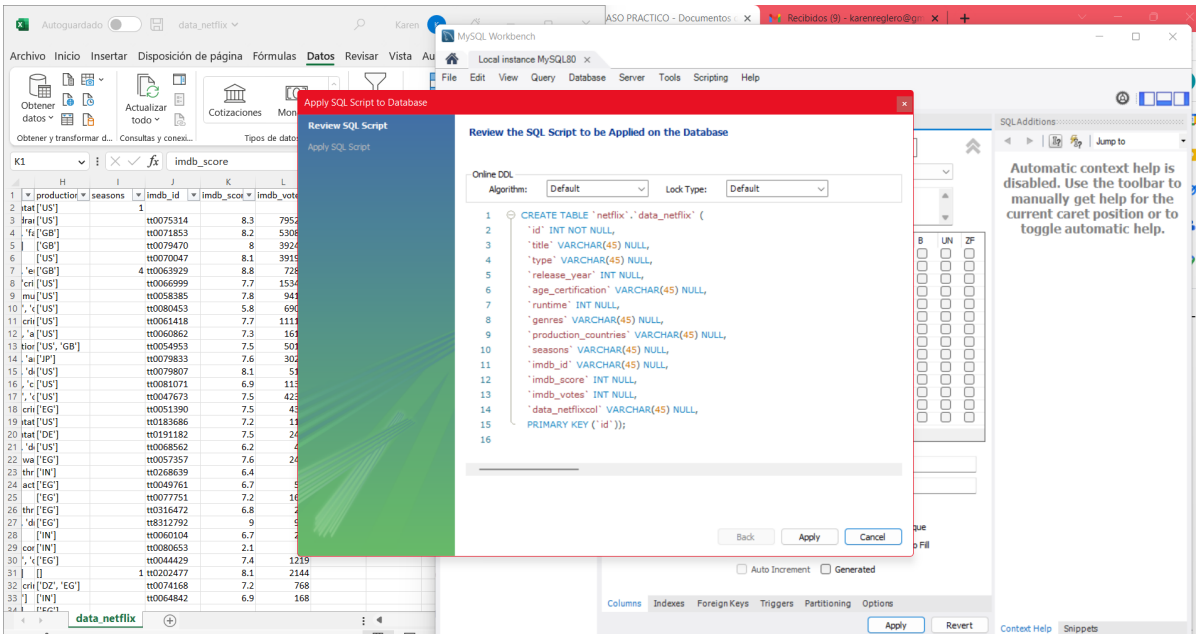

```
CREATE TABLE `netflix`.`actores` (
  `person_id` INT NOT NULL,
  `id` VARCHAR(45) NULL,
  `name` VARCHAR(45) NULL,
  `character` VARCHAR(45) NULL,
  `role` VARCHAR(45) NULL,
  PRIMARY KEY (`person_id`));
```

- 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

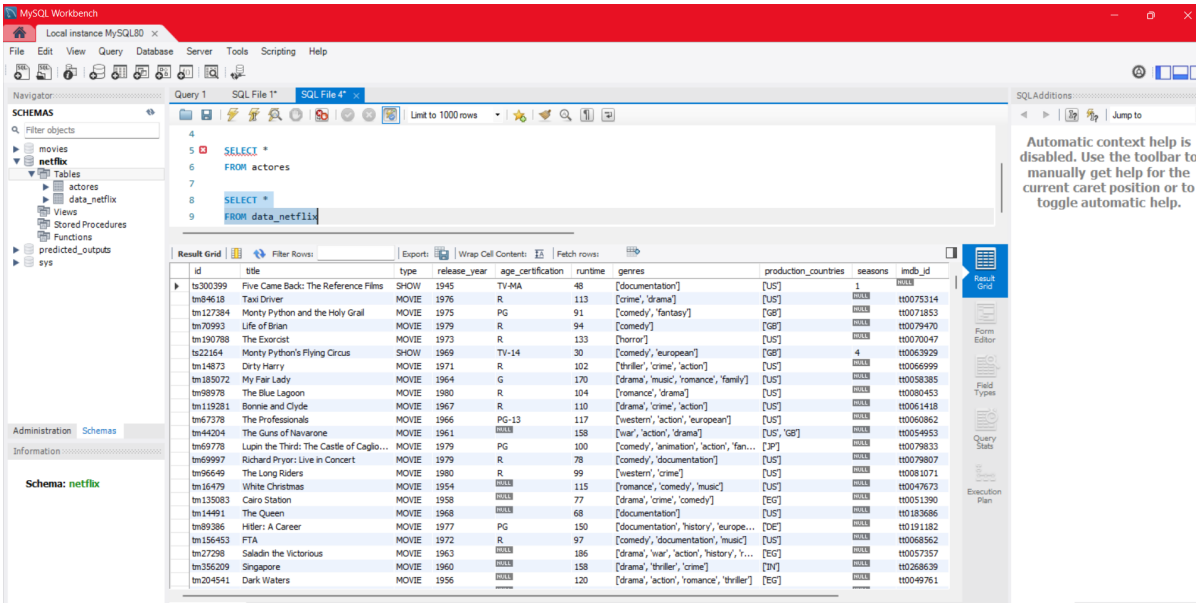




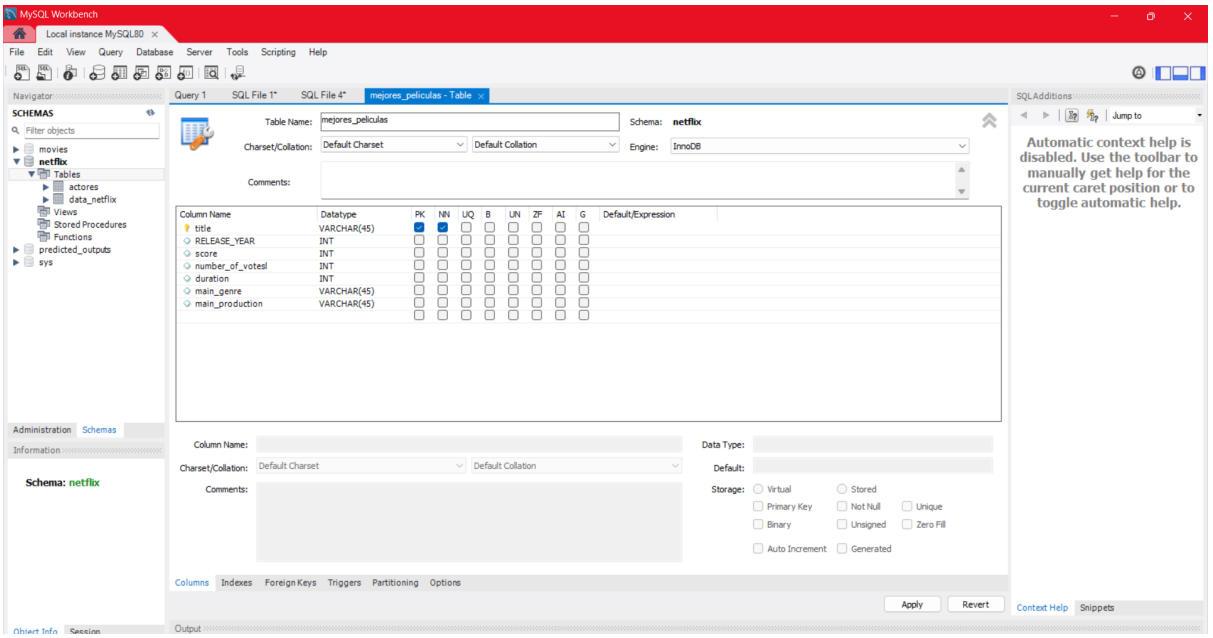

```
CREATE TABLE `netflix`.`data_netflix` (
  `id` INT NOT NULL,
  `title` VARCHAR(45) NULL,
  `type` VARCHAR(45) NULL,
  `release_year` INT NULL,
  `age_certification` VARCHAR(45) NULL,
  `runtime` INT NULL,
  `genres` VARCHAR(45) NULL,
  `production_countries` VARCHAR(45) NULL,
  `seasons` VARCHAR(45)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

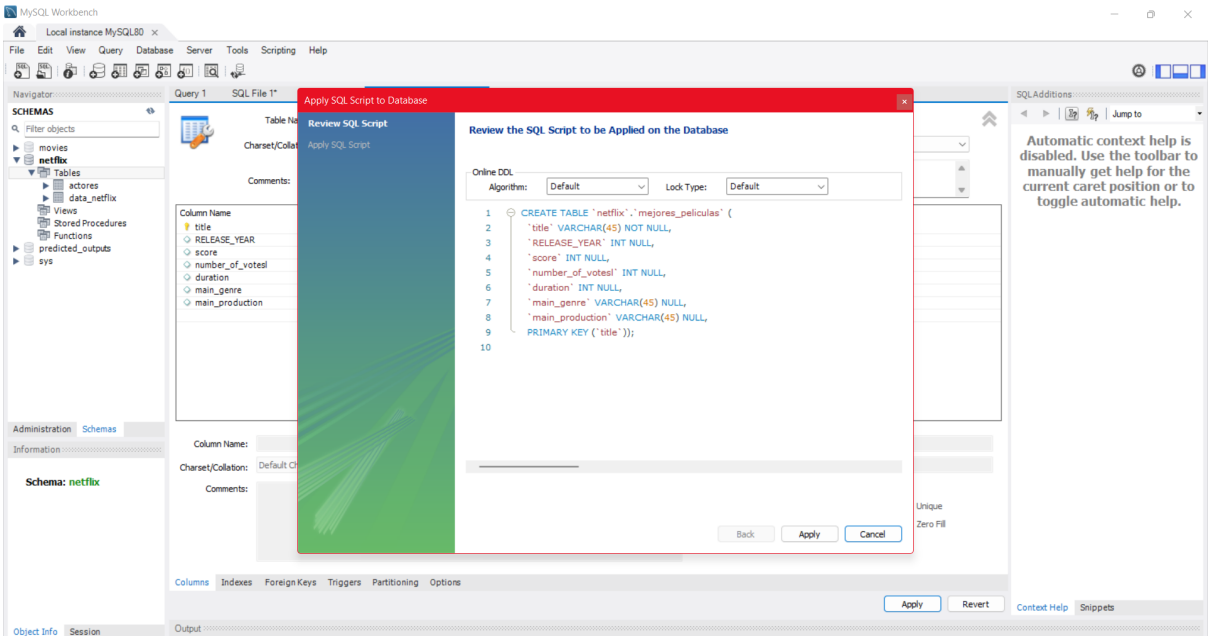
```
`imdb_score` INT NULL,
`imdb_votes` INT NULL,
`data_netflixcol` VARCHAR(45) NULL,
PRIMARY KEY (`id`));
```

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



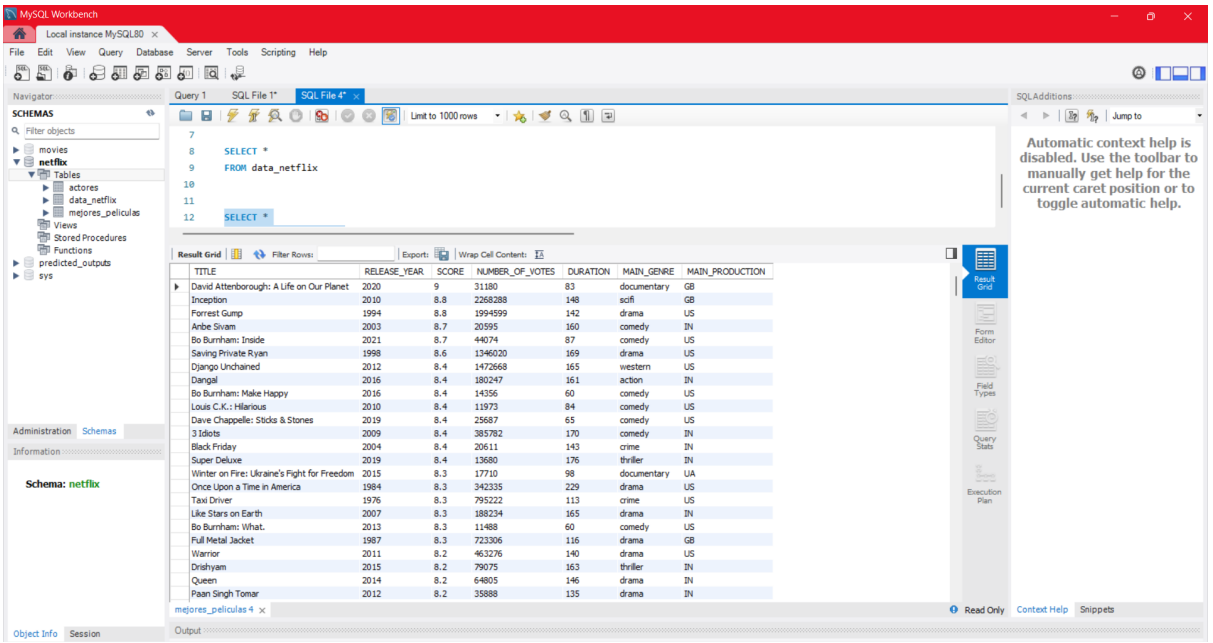
- Creación de la tabla mejores_peliculas en MySQL dentro de la base de datos netflix



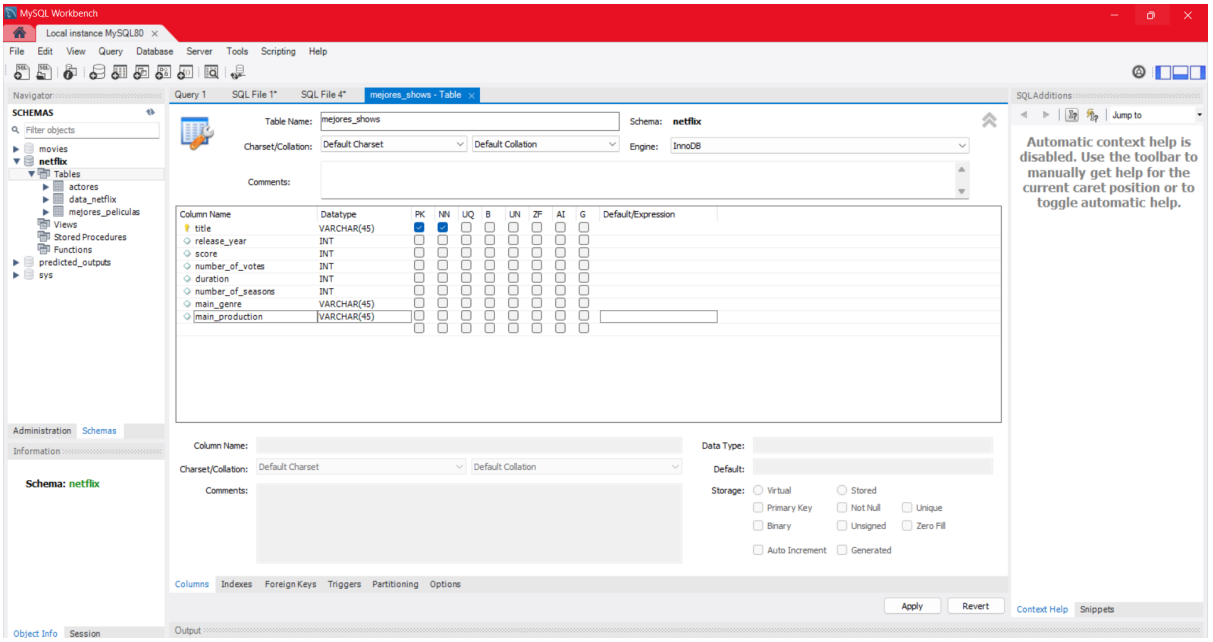


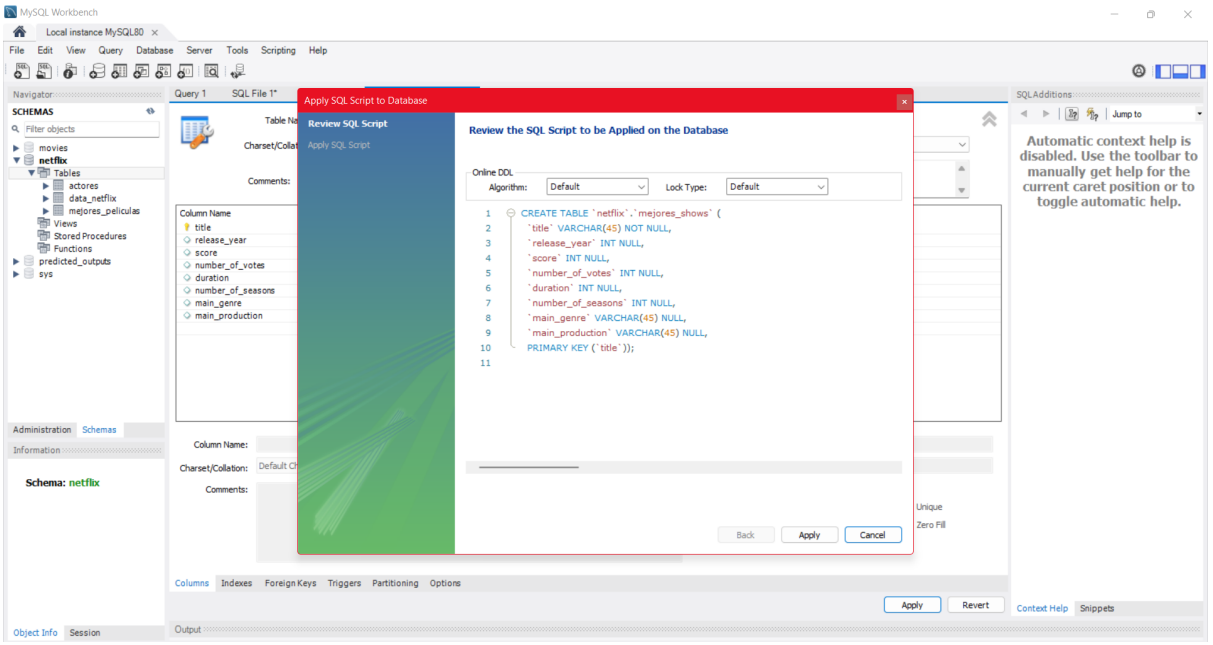

```
CREATE TABLE `netflix`.`mejores_peliculas` (
  `title` VARCHAR(45) NOT NULL,
  `RELEASE_YEAR` INT NULL,
  `score` INT NULL,
  `number_of_votesl` INT NULL,
  `duration` INT NULL,
  `main_genre` VARCHAR(45) NULL,
  `main_production` VARCHAR(45) NULL,
  PRIMARY KEY (`title`));
```

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

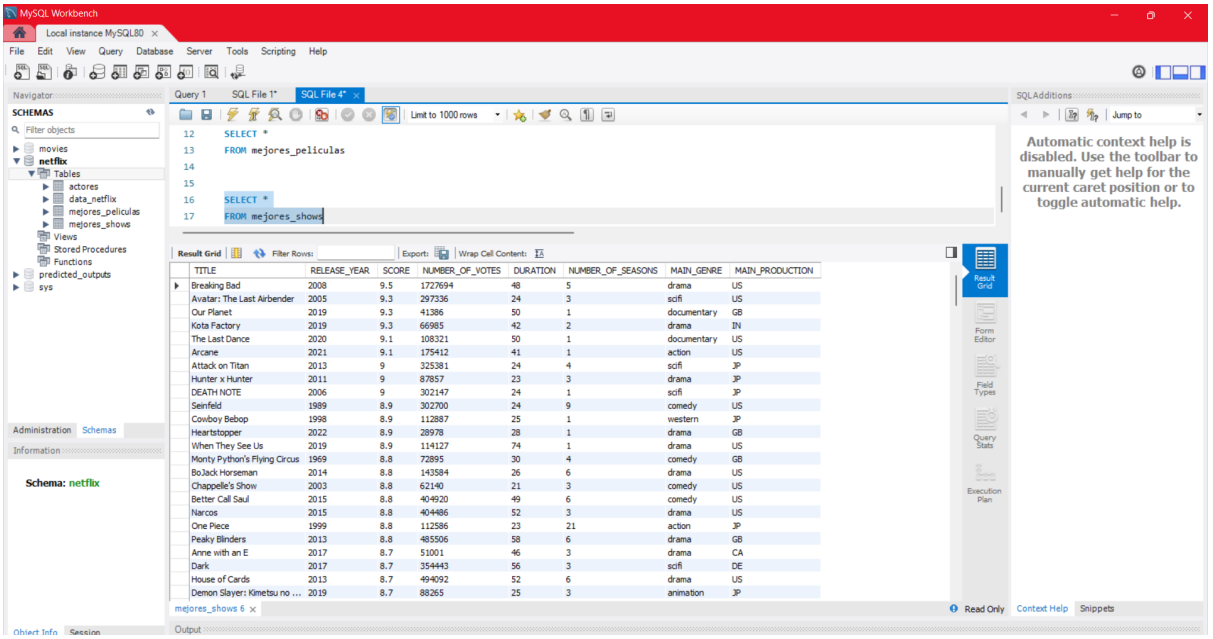


- Creación de la tabla mejores_shows en MySQL dentro de la base de datos netflix

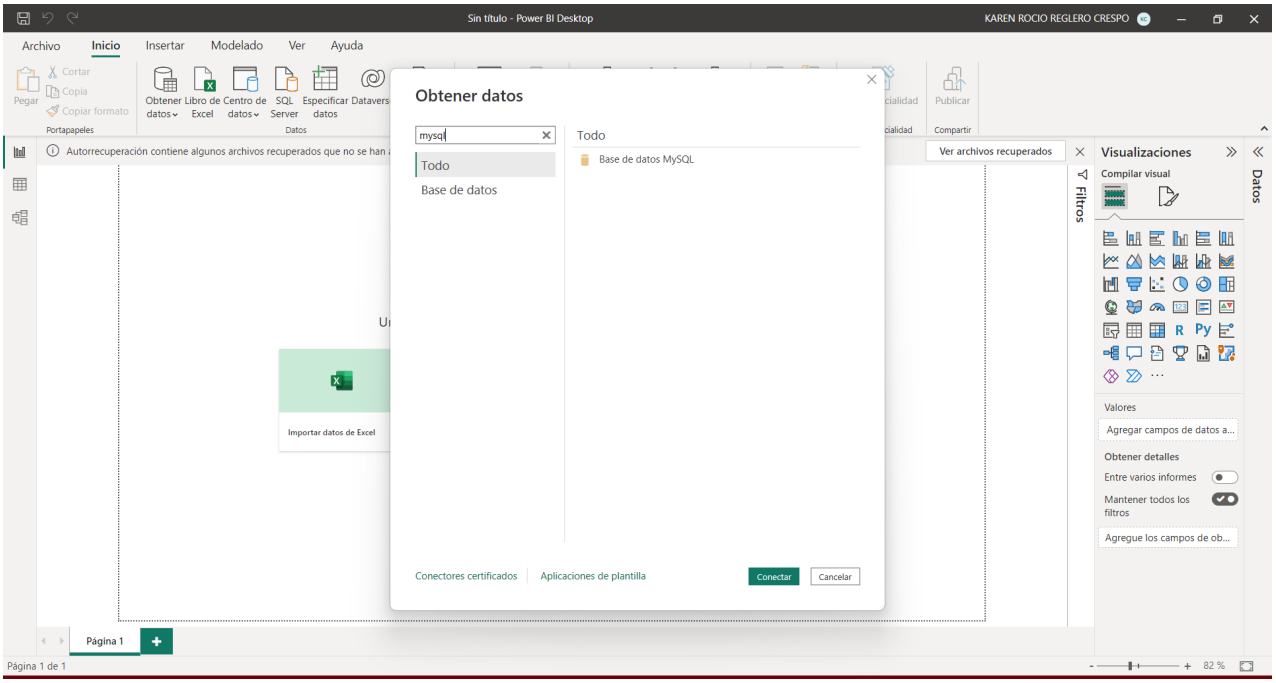




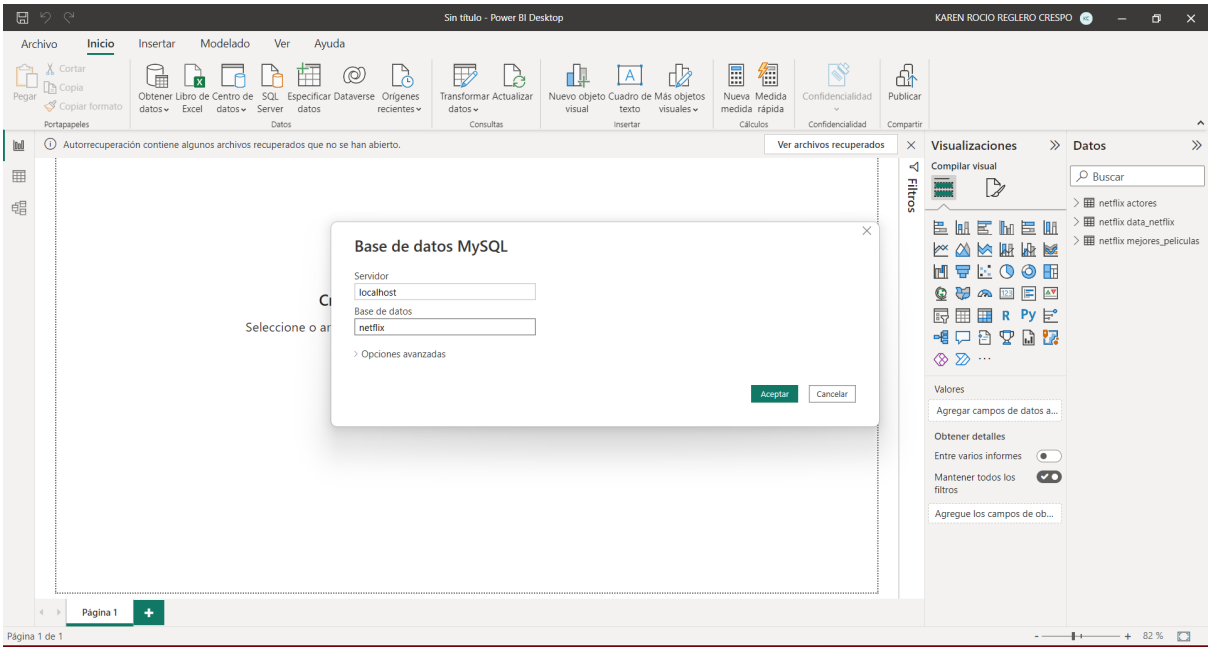
- SELECT de la tabla mejores_shows para confirmar que la información se ingestó de forma correcta:

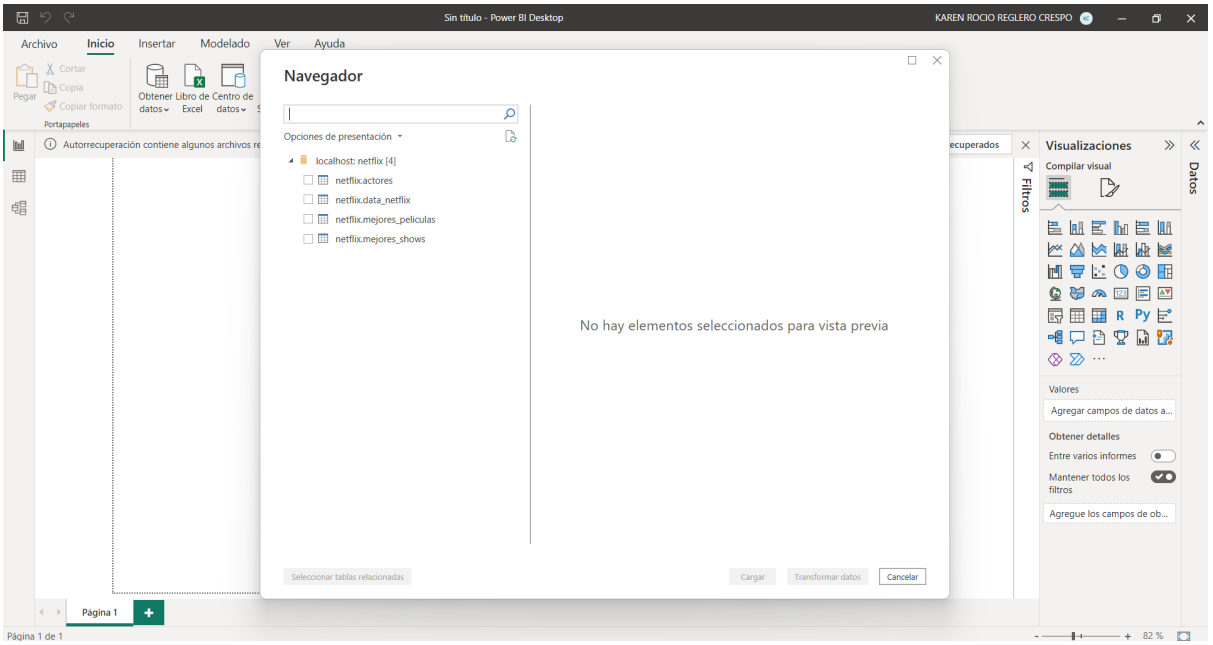


Power BI (Conexión a MySQL)



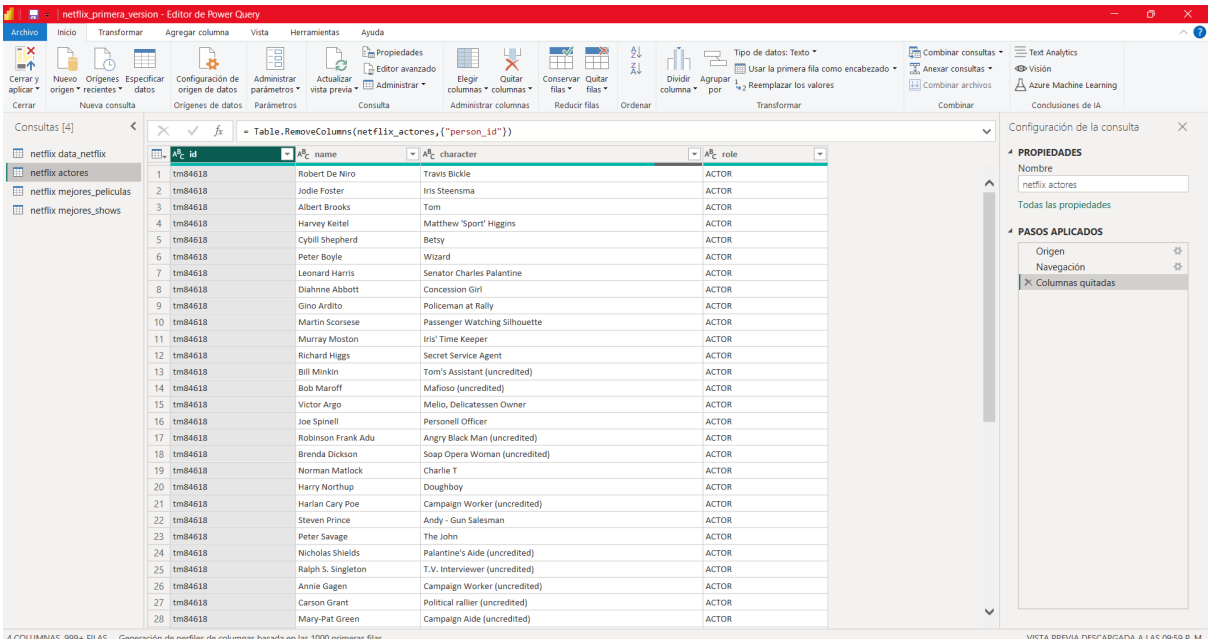
- Conexión a la base de datos de netflix de MySQL





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

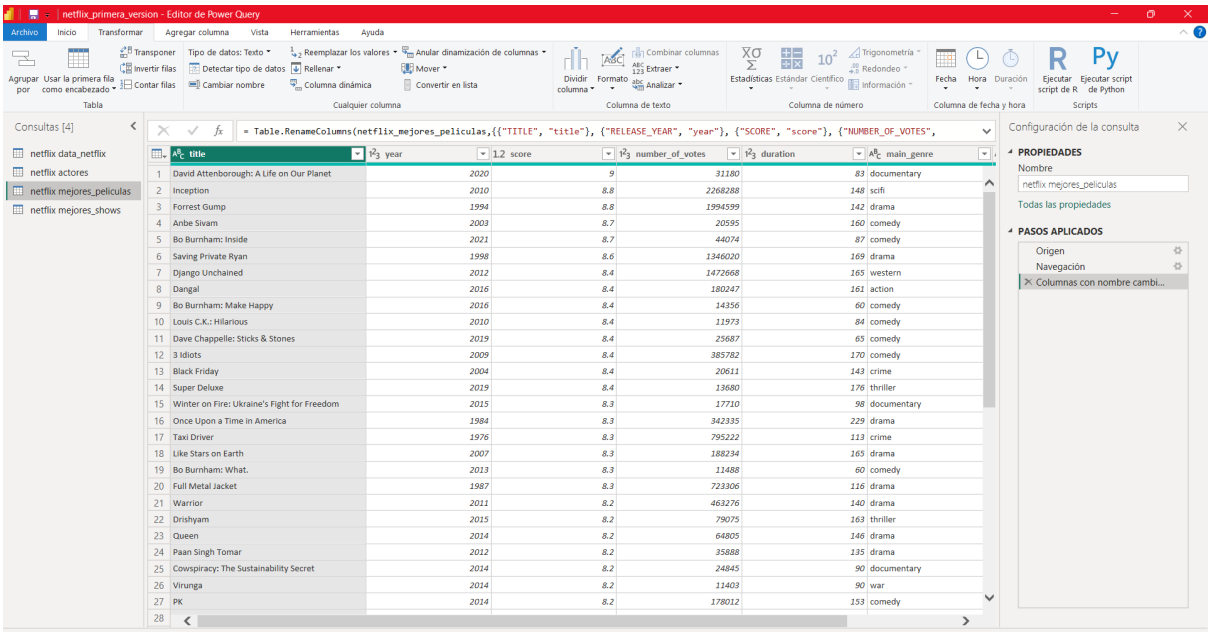


	id	name	character	role
1	tm84618	Robert De Niro	Travis Bickle	ACTOR
2	tm84618	Jodie Foster	Iris Steensma	ACTOR
3	tm84618	Albert Brooks	Tom	ACTOR
4	tm84618	Harvey Keitel	Matthew 'Sport' Higgins	ACTOR
5	tm84618	Cybill Shepherd	Betsy	ACTOR
6	tm84618	Peter Boyle	Wizard	ACTOR
7	tm84618	Leonard Harris	Senator Charles Palantine	ACTOR
8	tm84618	Diahnne Abbott	Concession Girl	ACTOR
9	tm84618	Gino Ardito	Policeman at Rally	ACTOR
10	tm84618	Martin Scorsese	Passenger Watching Silhouette	ACTOR
11	tm84618	Murray Moston	Iris' Time Keeper	ACTOR
12	tm84618	Richard Higgs	Secret Service Agent	ACTOR
13	tm84618	Bill Minkin	Tom's Assistant (uncredited)	ACTOR
14	tm84618	Bob Maroff	Mafioso (uncredited)	ACTOR
15	tm84618	Victor Argo	Melio, Delicatessen Owner	ACTOR
16	tm84618	Joe Spinell	Personell Officer	ACTOR
17	tm84618	Robinson Frank Adu	Angry Black Man (uncredited)	ACTOR
18	tm84618	Brenda Dickson	Soap Opera Woman (uncredited)	ACTOR
19	tm84618	Norman Matlock	Charlie T	ACTOR
20	tm84618	Harry Northup	Doughboy	ACTOR
21	tm84618	Harlan Cary Poe	Campaign Worker (uncredited)	ACTOR
22	tm84618	Steven Prince	Andy - Gun Salesman	ACTOR
23	tm84618	Peter Savage	The John	ACTOR
24	tm84618	Nicholas Shields	Palantine's Aide (uncredited)	ACTOR
25	tm84618	Ralph S. Singleton	T.V. Interviewer (uncredited)	ACTOR
26	tm84618	Annie Gagen	Campaign Worker (uncredited)	ACTOR
27	tm84618	Carson Grant	Political rallier (uncredited)	ACTOR
28	tm84618	Mary-Pat Green	Campaign Aide (uncredited)	ACTOR

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.

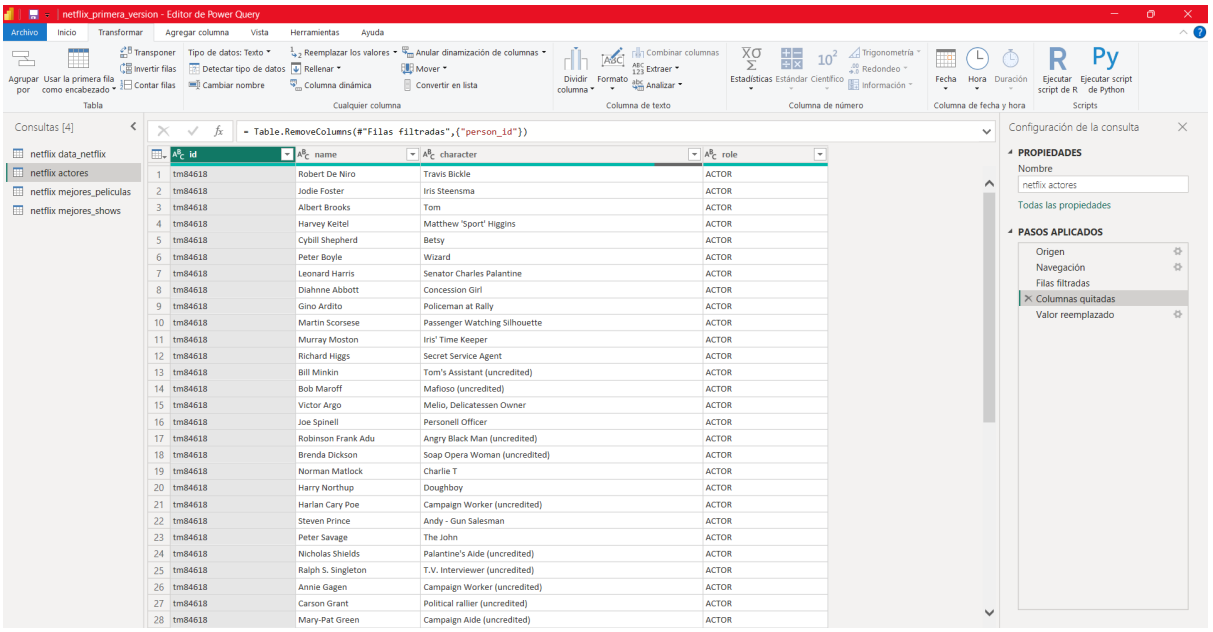
EJEMPLO: release_year → year
RELEASE_YEAR → year



	title	year	score	number_of_votes	duration	main_genre
1	David Attenborough: A Life on Our Planet	2020	9	31180	83	documentary
2	Inception	2010	8.8	2268288	148	scifi
3	Forrest Gump	1994	8.8	1994599	142	drama
4	Anbe Sivam	2003	8.7	20595	160	comedy
5	Bo Burnham: Inside	2021	8.7	44074	87	comedy
6	Saving Private Ryan	1998	8.6	1346020	169	drama
7	Django Unchained	2012	8.4	1472668	165	western
8	Dangal	2016	8.4	180247	161	action
9	Bo Burnham: Make Happy	2016	8.4	14356	60	comedy
10	Louis C.K.: Hilarious	2010	8.4	11973	84	comedy
11	Dave Chappelle: Sticks & Stones	2019	8.4	25687	65	comedy
12	3 Idiots	2009	8.4	385782	170	comedy
13	Black Friday	2004	8.4	20611	143	crime
14	Super Deluxe	2019	8.4	13680	176	thriller
15	Winter on Fire: Ukraine's Fight for Freedom	2015	8.3	17710	98	documentary
16	Once Upon a Time in America	1984	8.3	342335	229	drama
17	Taxi Driver	1976	8.3	795222	113	crime
18	Like Stars on Earth	2007	8.3	188234	165	drama
19	Bo Burnham: What.	2013	8.3	11488	60	comedy
20	Full Metal Jacket	1987	8.3	723306	116	drama
21	Warrior	2011	8.2	463276	140	drama
22	Drishyam	2015	8.2	79075	163	thriller
23	Queen	2014	8.2	64805	146	drama
24	Paan Singh Tomar	2012	8.2	35888	135	drama
25	Cowspiracy: The Sustainability Secret	2014	8.2	24845	90	documentary
26	Virunga	2014	8.2	11403	90	war
27	PK	2014	8.2	178012	153	comedy

- Se reemplazaron valores nulos por 0 para una mejor visualización.

- Se removieron columnas innecesarias para este análisis:
  - person_id
  - character
  - age_certification
  - imdb_id



	id	name	character	role
1	tm84618	Robert De Niro	Travis Bickle	ACTOR
2	tm84618	Jodie Foster	Iris Steensma	ACTOR
3	tm84618	Albert Brooks	Tom	ACTOR
4	tm84618	Harvey Keitel	Matthew 'Sport' Higgins	ACTOR
5	tm84618	Cybill Shepherd	Betsy	ACTOR
6	tm84618	Peter Boyle	Wizard	ACTOR
7	tm84618	Leonard Harris	Senator Charles Palantine	ACTOR
8	tm84618	Diahnne Abbott	Concession Girl	ACTOR
9	tm84618	Gino Ardito	Policeman at Rally	ACTOR
10	tm84618	Martin Scorsese	Passenger Watching Silhouette	ACTOR
11	tm84618	Murray Moston	Iris' Time Keeper	ACTOR
12	tm84618	Richard Higgs	Secret Service Agent	ACTOR
13	tm84618	Bill Minkin	Tom's Assistant (uncredited)	ACTOR
14	tm84618	Bob Maroff	Mafioso (uncredited)	ACTOR
15	tm84618	Victor Argo	Melio, Delicatessen Owner	ACTOR
16	tm84618	Joe Spinell	Personell Officer	ACTOR
17	tm84618	Robinson Frank Adu	Angry Black Man (uncredited)	ACTOR
18	tm84618	Brenda Dickson	Soap Opera Woman (uncredited)	ACTOR
19	tm84618	Norman Matlock	Charlie T	ACTOR
20	tm84618	Harry Northup	Doughboy	ACTOR
21	tm84618	Harlan Cary Poe	Campaign Worker (uncredited)	ACTOR
22	tm84618	Steven Prince	Andy - Gun Salesman	ACTOR
23	tm84618	Peter Savage	The John	ACTOR
24	tm84618	Nicholas Shields	Palantine's Aide (uncredited)	ACTOR
25	tm84618	Ralph S. Singleton	T.V. Interviewer (uncredited)	ACTOR
26	tm84618	Annie Gagen	Campaign Worker (uncredited)	ACTOR
27	tm84618	Carson Grant	Political rallier (uncredited)	ACTOR
28	tm84618	Mary-Pat Green	Campaign Aide (uncredited)	ACTOR

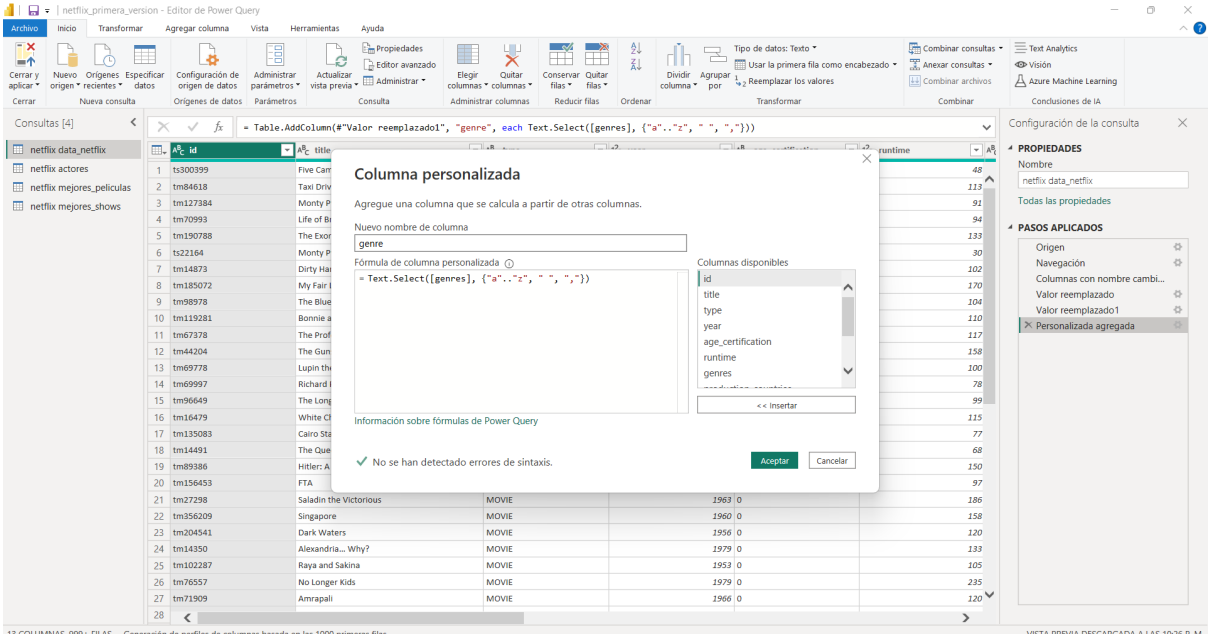
TRANSFORMACIÓN DE DATOS

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

- ❖ genres
- ❖ production_countries

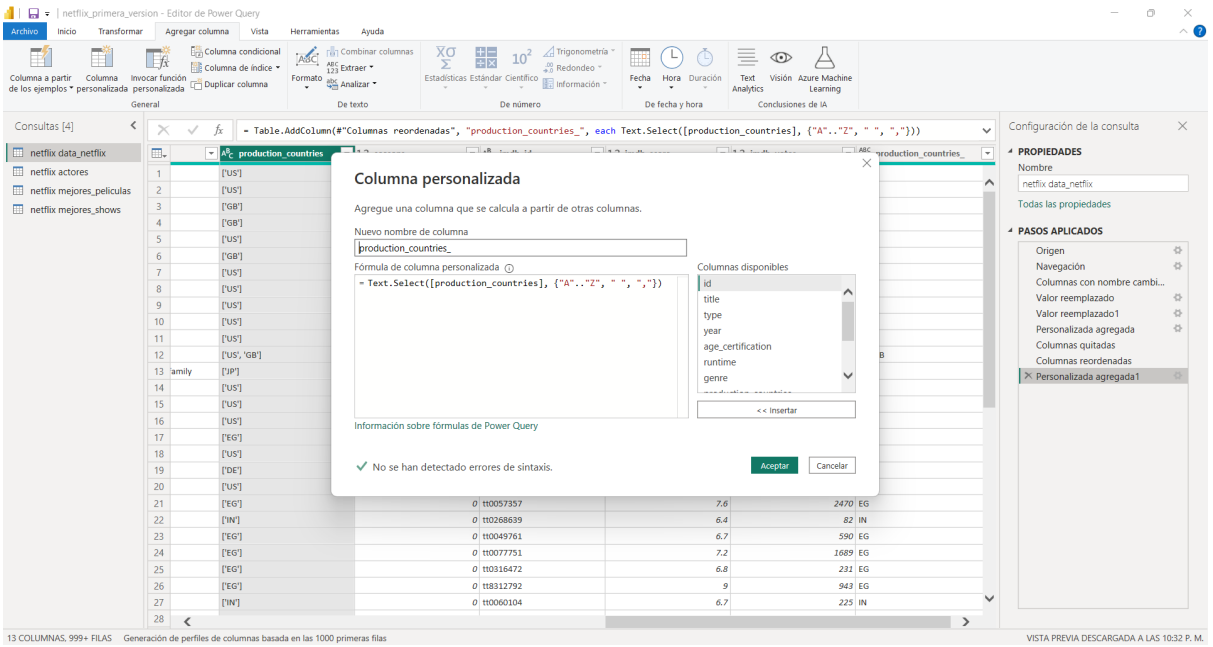
Se opta por realizar una anulación de dinamización de columnas para visualizar los resultados de [‘drama’, ‘comedy’] → drama y comedy por separado.

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



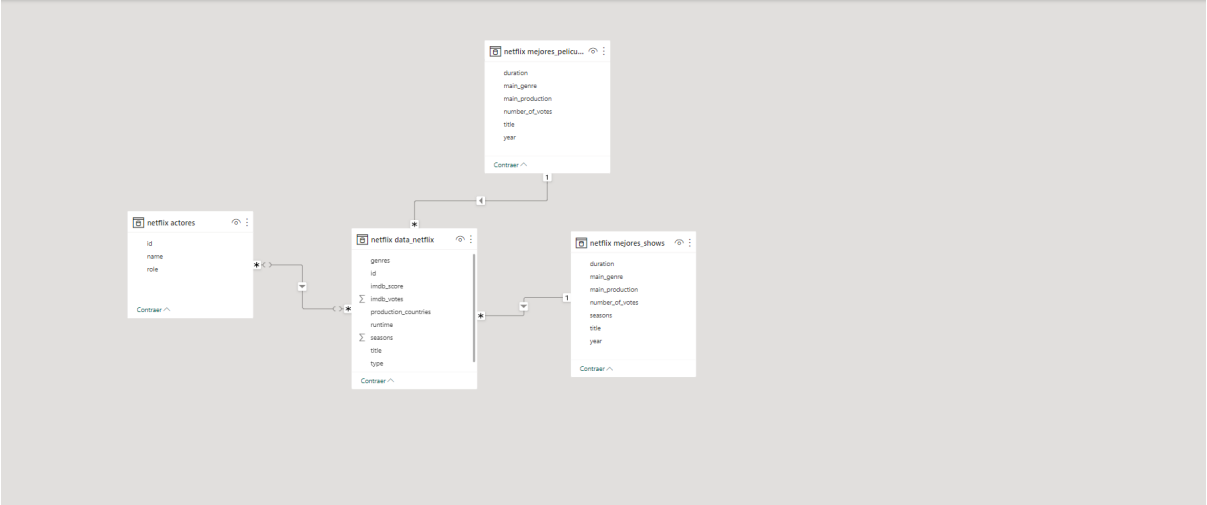
- Posteriormente, se divide la columna por delimitador (,) y se realiza el despivoteo de la tabla nuevamente para contener valores únicos.

- Se realiza el mismo proceso con la columna production_countries:



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_peliculas 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.