Project Title: Exploratory Data Analysis on Film Data Using PostgreSQL

Prepared by Thomas, Nelson Chinenyeze

Date: 5th August 2024

Table of Contents

1. INTRODUCTION

Project Description: In this project, I applied the SQL skills I have learned throughout the course to perform exploratory data analysis on the "film" database, which contains information about films, people, reviews, and roles. The project focused on retrieving, filtering, and analyzing data to answer real-life business questions related to the film industry.

Project Objectives:

- Apply SQL queries to retrieve and analyze data from the "film" database.
- Gain practical experience in using SQL for data analysis.
- Answer real-life business questions using SQL.
- Present findings in a clear and organized manner.

2. METHODOLOGY

Tools and Databases Used:

- PostgreSQL and PgAdmin4 for SQL queries and data retrieval.
- Excel for descriptive analysis.
- Tableau for data visualization and presentation.

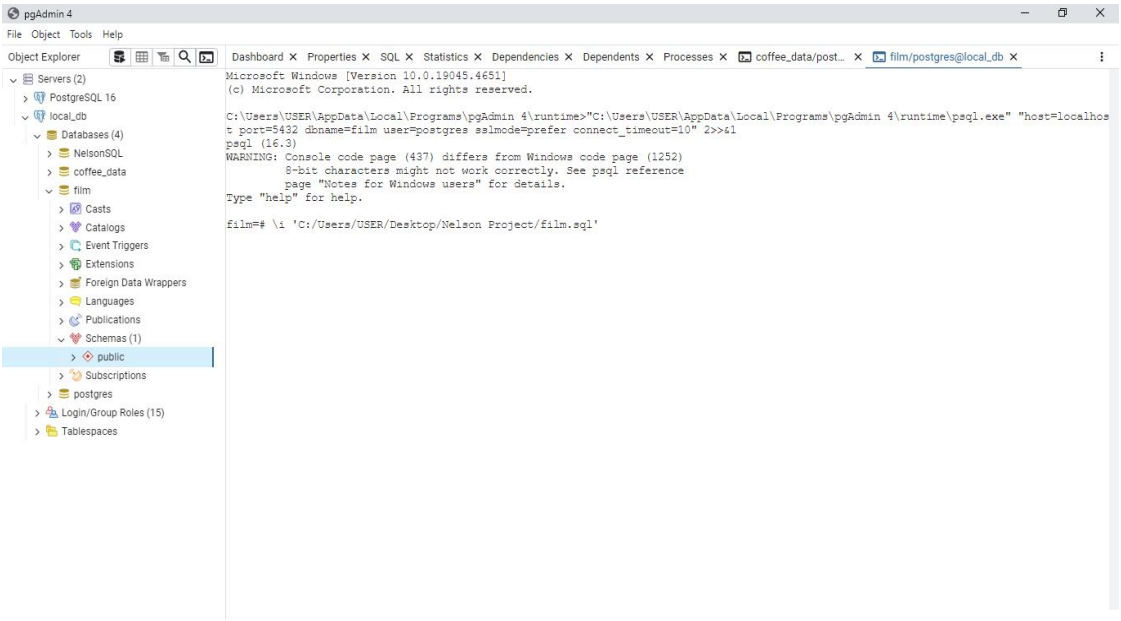
SQL Queries and Analysis Process: A series of SQL queries were used to extract data from the "film" database, followed by analysis to derive insights related to film performance, distribution, and audience engagement.

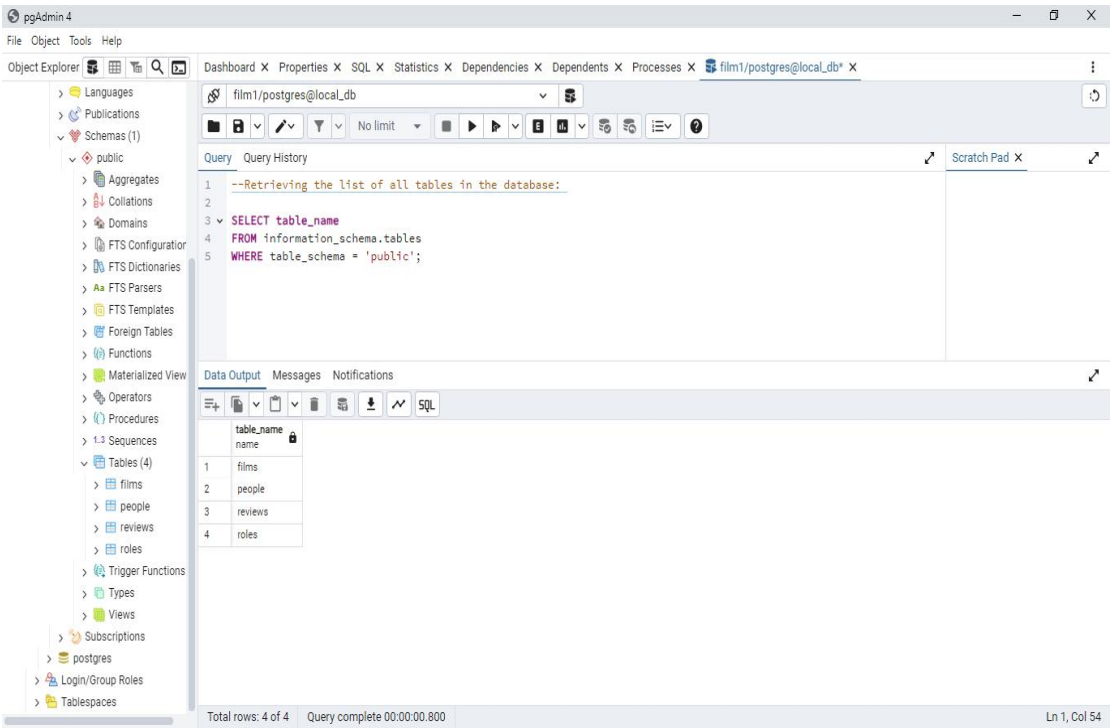
3. Data Retrieval and Analysis

Task 1: Data Retrieval

Here I connected to the “film” database using PostgreSQL and PgAdmin4 and I ensured the tables in the film database are related. Below images show the SQL code queries.

SQL Code:



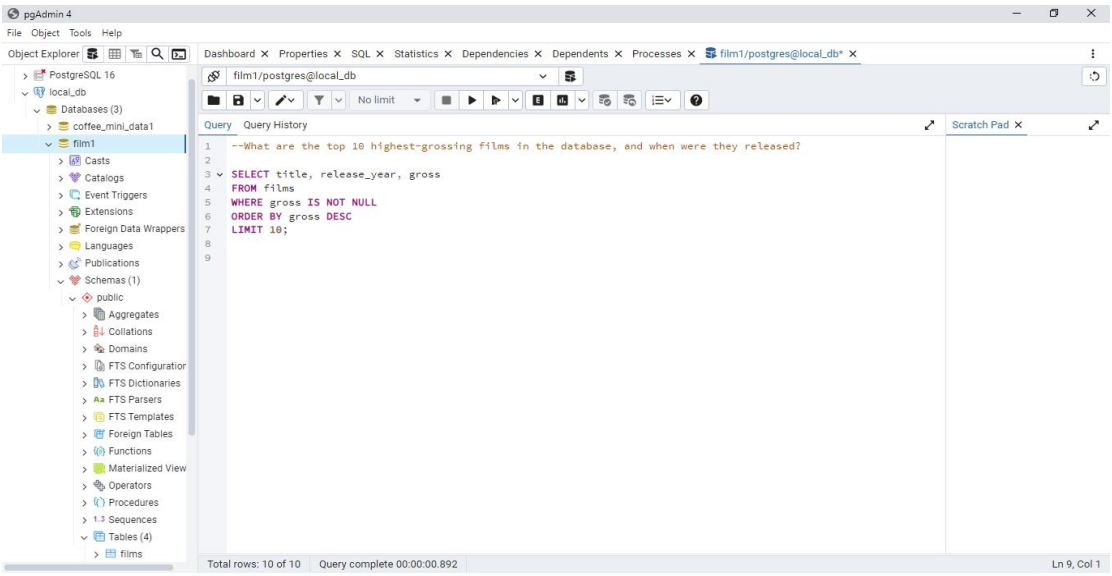


Task 2: Film Analysis

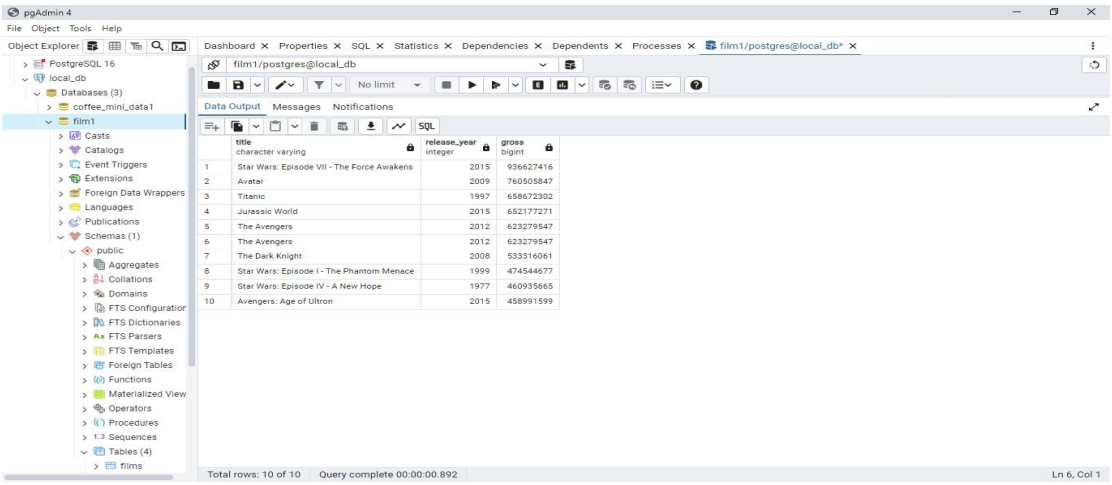
In this section, I used SQL queries to retrieve, filter, and analyze data to answer real-life business questions related to the film industry. Below are the questions;

Question 2.1 : What are the top 10 highest-grossing films in the database, and when were they released?

SQL Code:



Results:

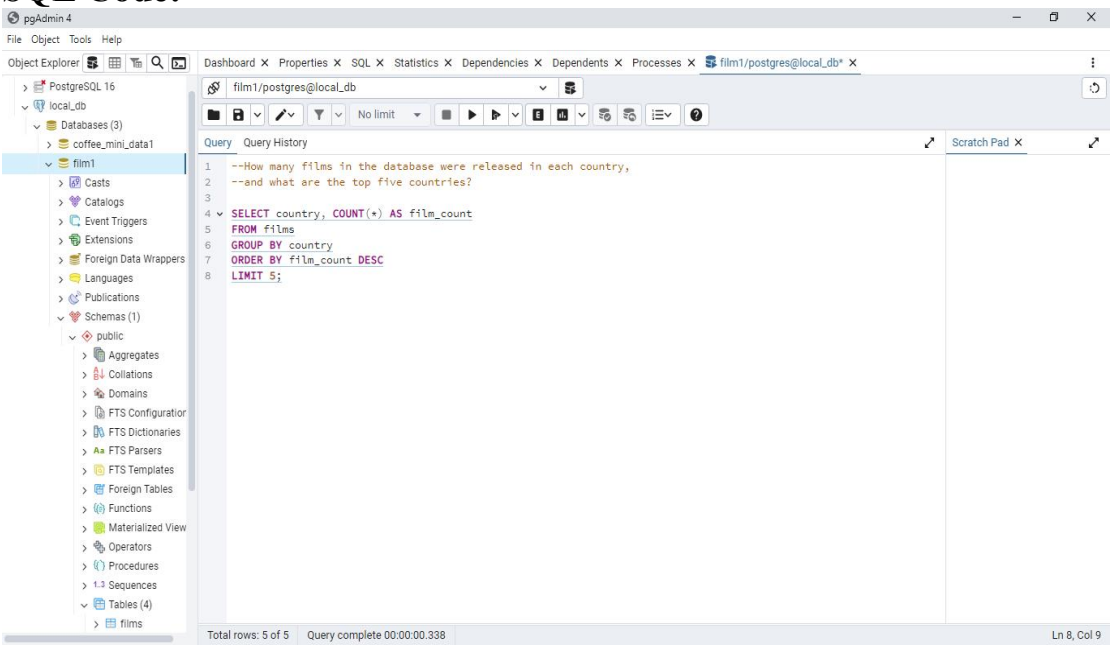


	title character varying	release_year integer	gross bigint
1	Star Wars: Episode VII - The Force Awakens	2015	936627416
2	Avatar	2009	760505847
3	Titanic	1997	658672302
4	Jurassic World	2015	652177271
5	The Avengers	2012	623279547
6	The Avengers	2012	623279547
7	The Dark Knight	2008	533316061
8	Star Wars: Episode I - The Phantom Menace	1999	474544677
9	Star Wars: Episode IV - A New Hope	1977	460935665
10	Avengers: Age of Ultron	2015	458991599

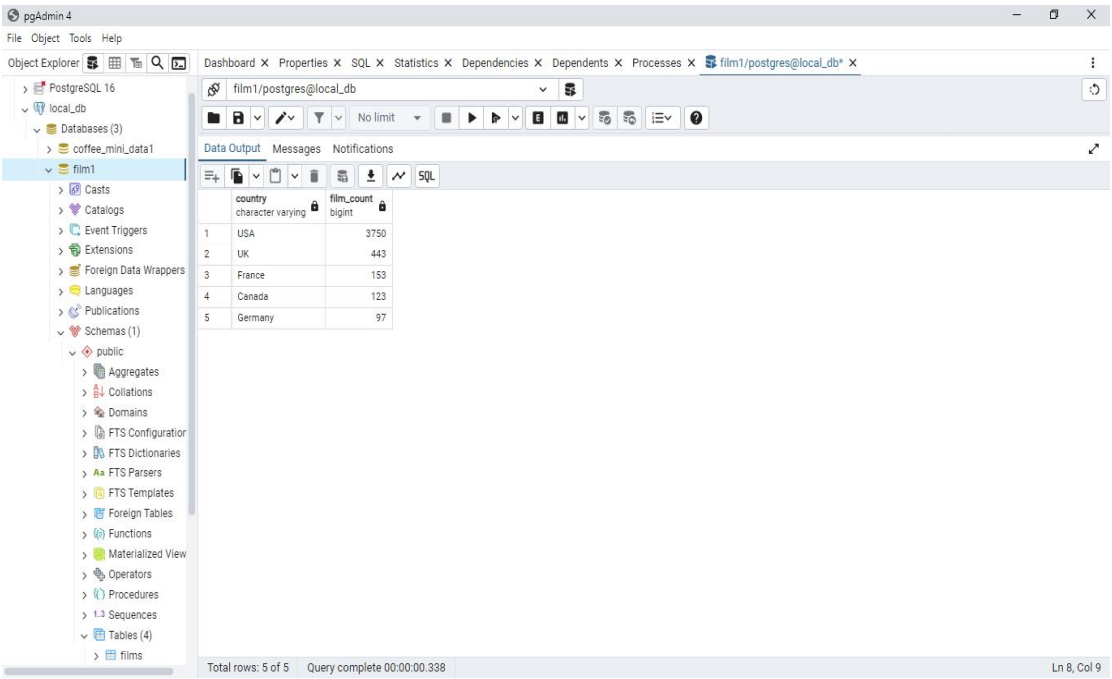
Analysis: The top 10 highest-grossing films are listed, with their release dates. This provides insights into the most financially successful films in the database.

Question 2.2: How many films in the database were released in each country, and what are the top five countries?

SQL Code:



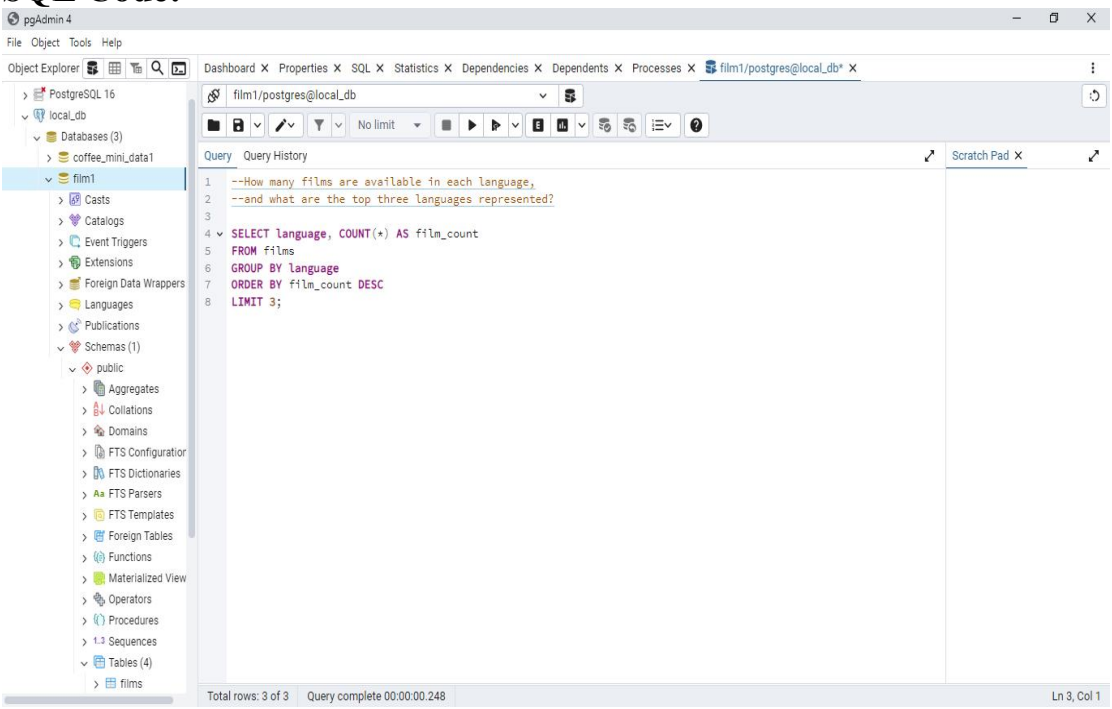
Results:



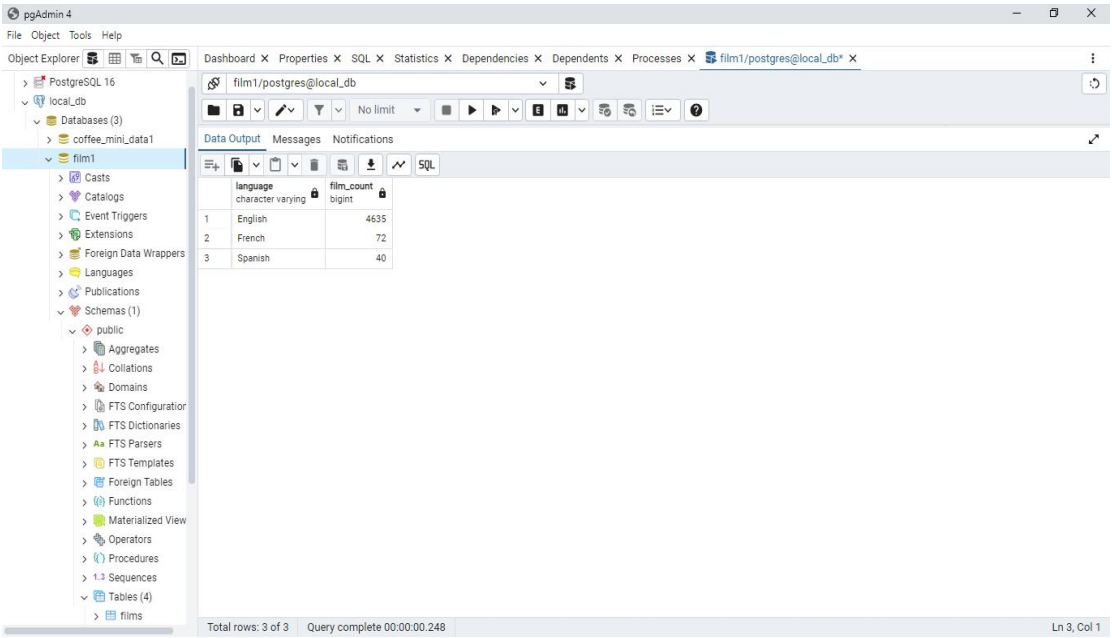
Analysis: The data shows the number of films released in each country, highlighting the top five countries with the most film release.

Question 2.3: How many films are available in each language, and what are the top three languages represented?

SQL Code:



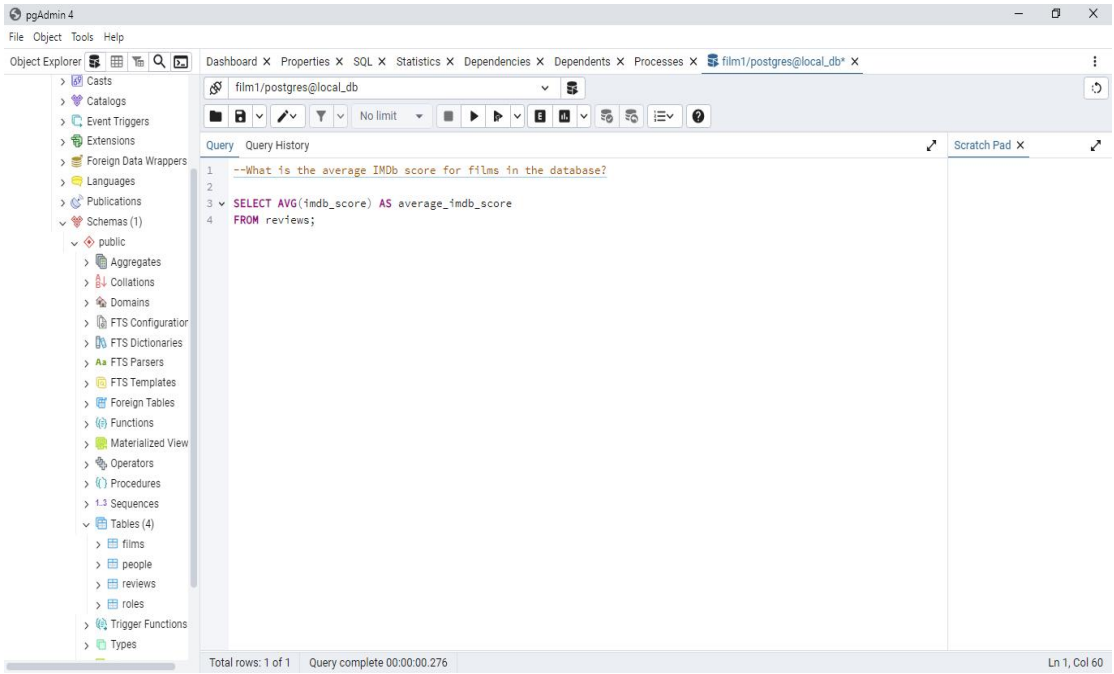
Results:



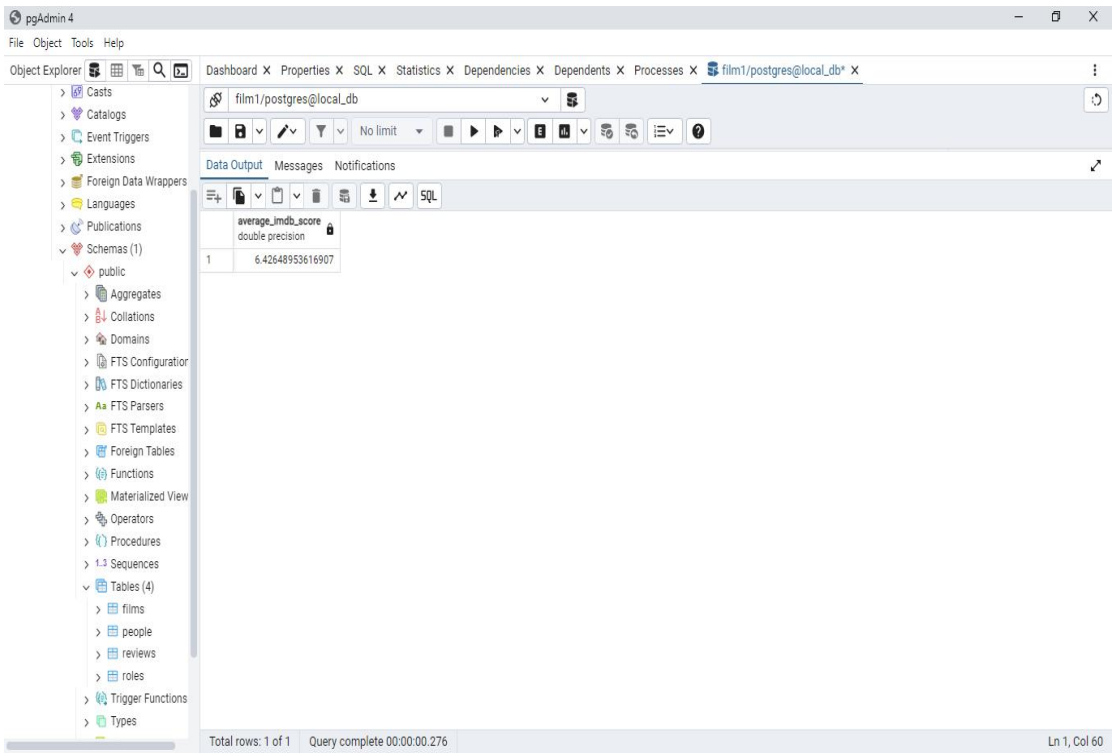
Analysis: The data shows the top three languages with the highest number of films, along with the count of films available in each of the languages.

Question 2.4: What is the average IMDb score for films in the database?

SQL Code:



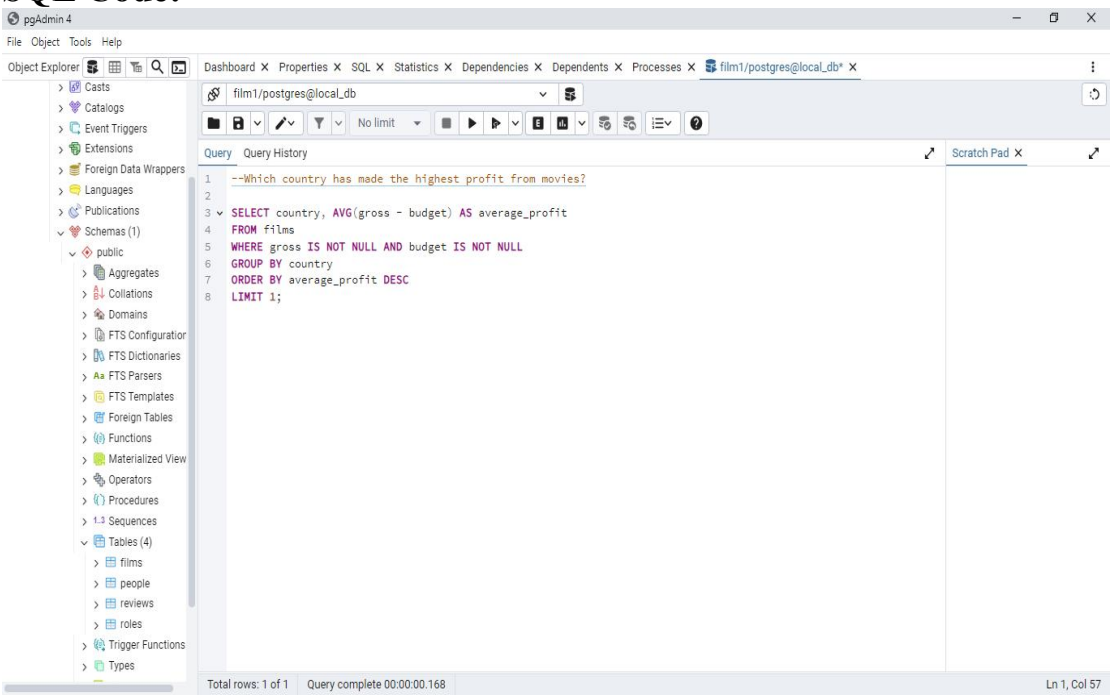
Results:



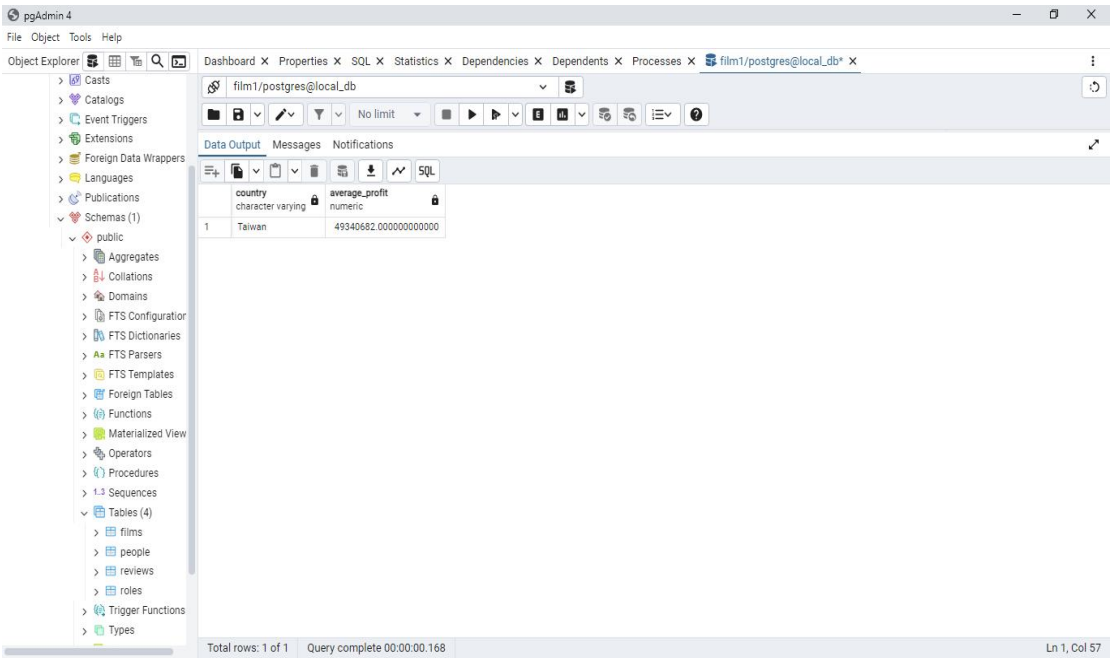
Analysis: The query result shows the average IMDb score for all the films in the database.

Question 2.5: Which country has made the highest profit from movies?

SQL Code:



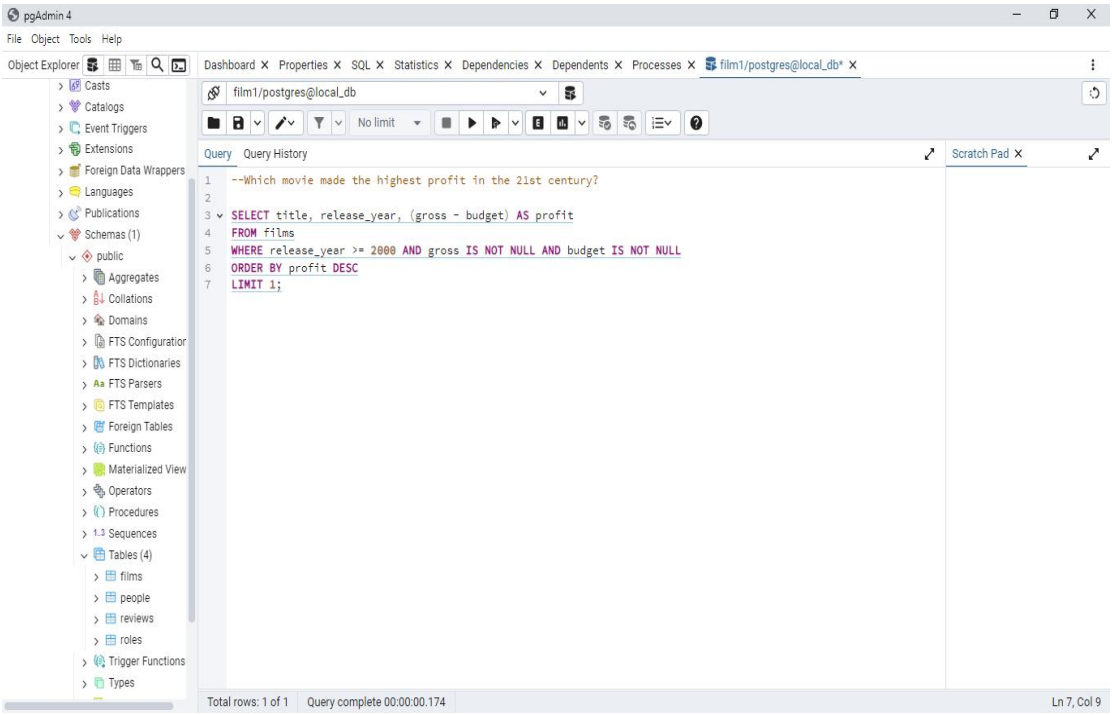
Results:



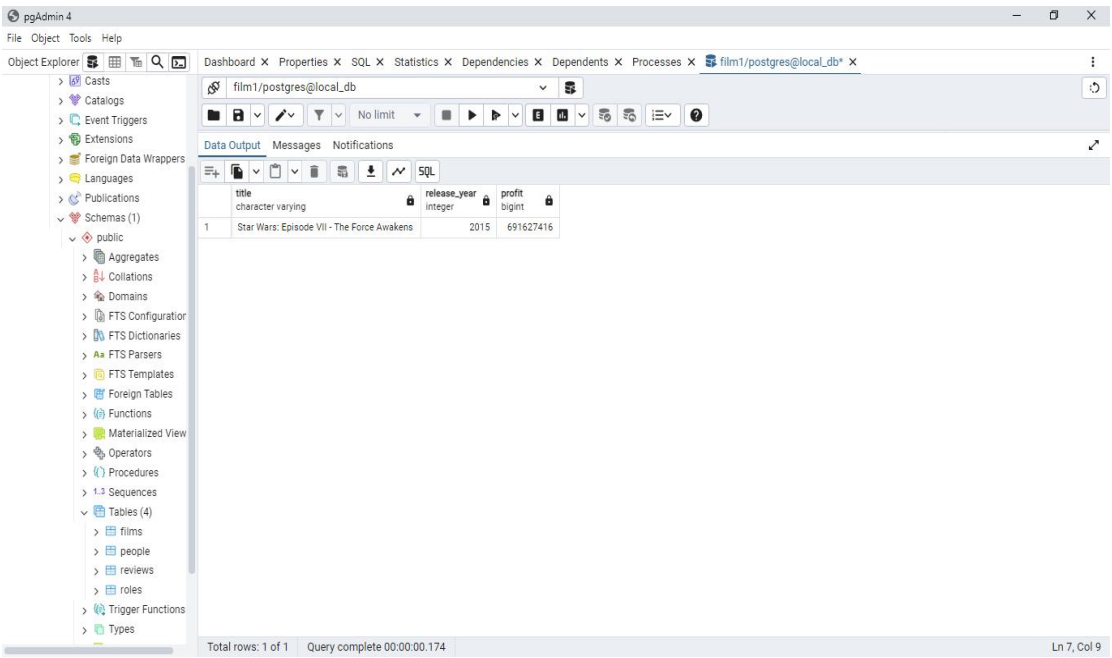
Analysis: This query filters out films where the gross or budget values are null, before computing the average profit, ensuring that only valid data is used in the calculation.

Question 2.6: Which movie made the highest profit in the 21st century?

SQL Code:



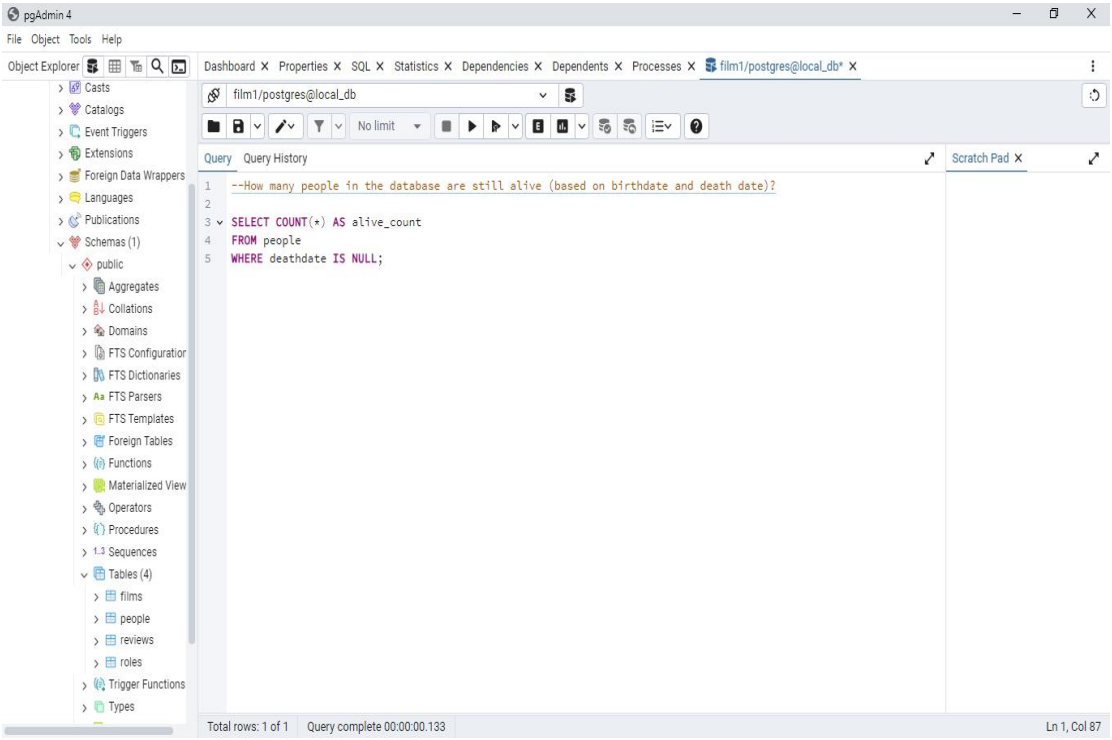
Results:



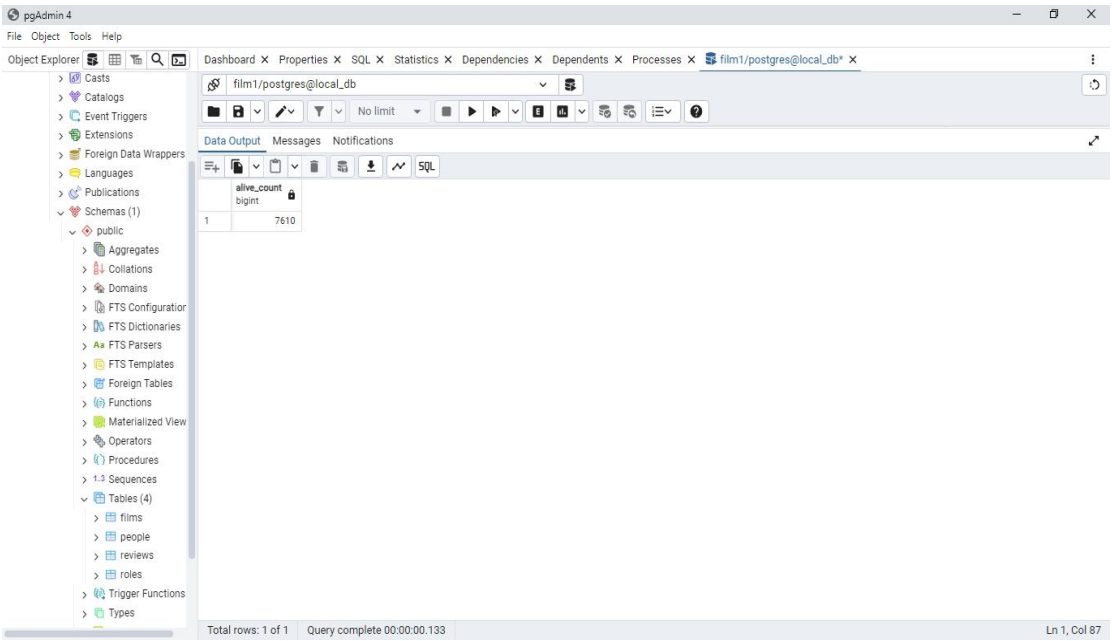
Analysis: The data shows the movie with the highest profit released in the 21st century.

Question 2.7: How many people in the database are still alive (based on birth date and death date)?

SQL Code:



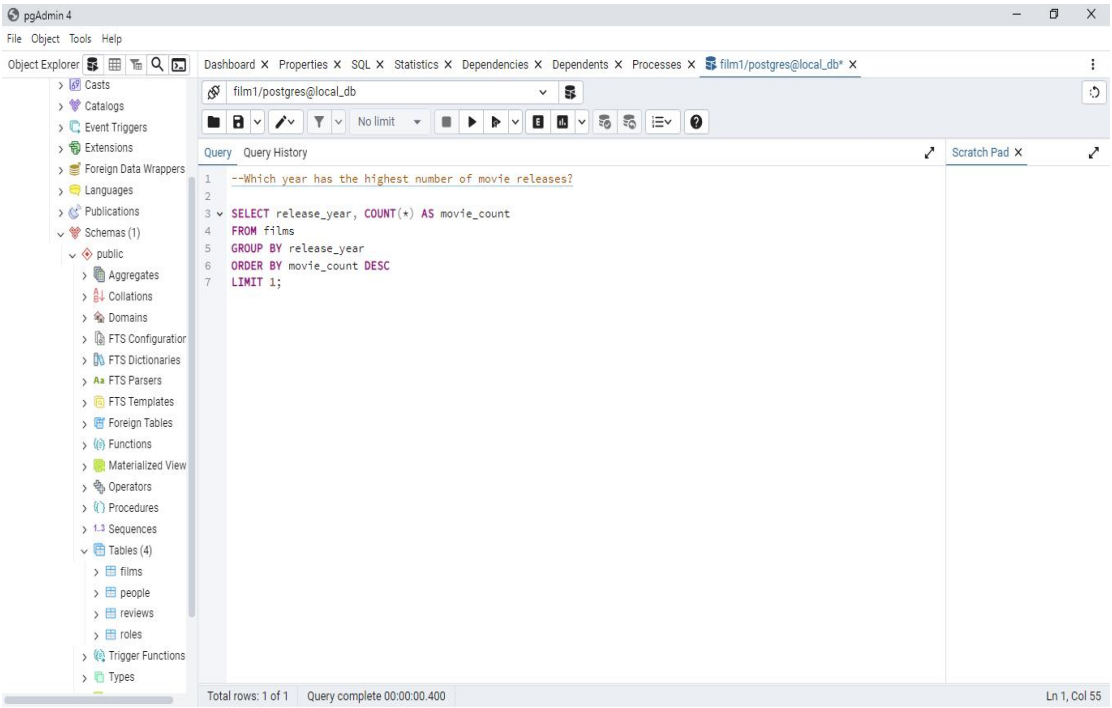
Results:



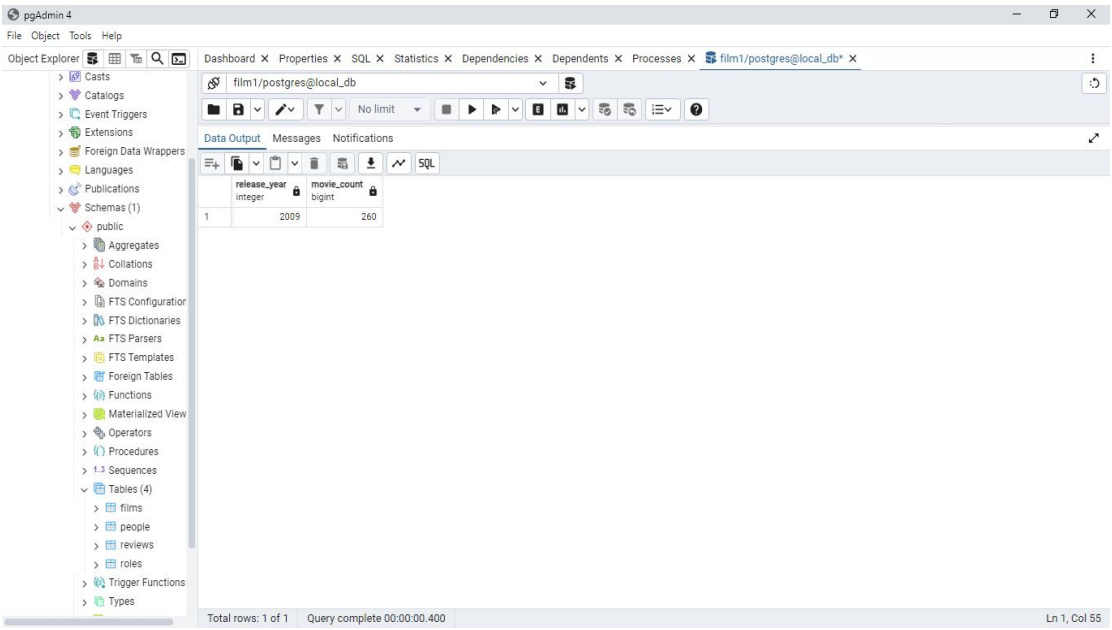
Analysis: The query shows the count of people who are still alive in the database.

Question 2.8: Which year has the highest number of movie releases?

SQL Code:



Results:



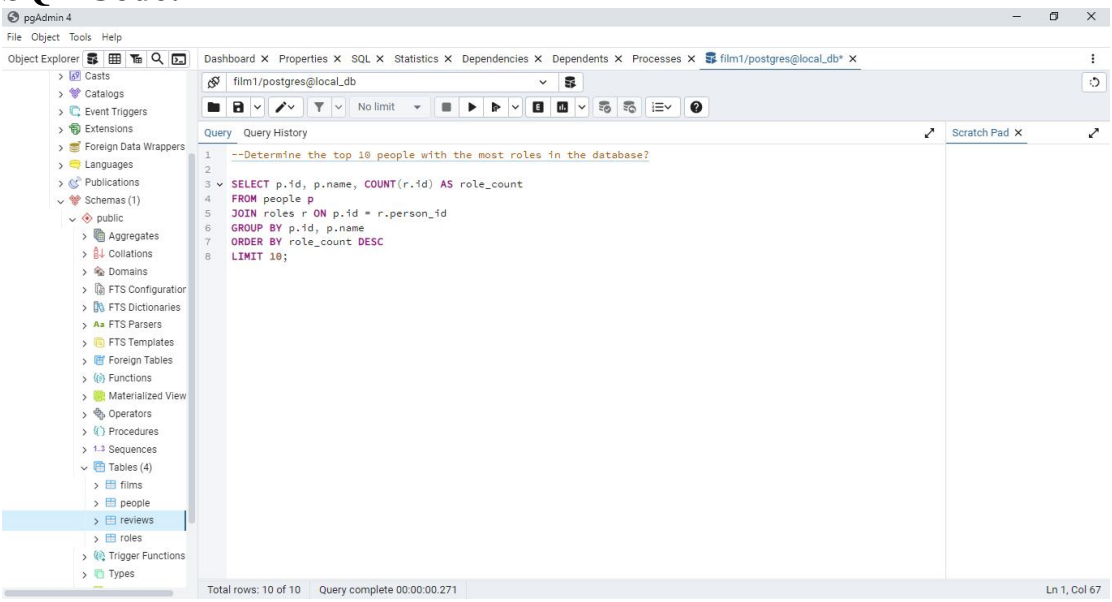
Analysis: This query result shows, the movies grouped by their release year, counts the number of movies for each year, orders the results by the count in descending order, and then limits the result to just the top year.

Task 3: People and Roles Analysis

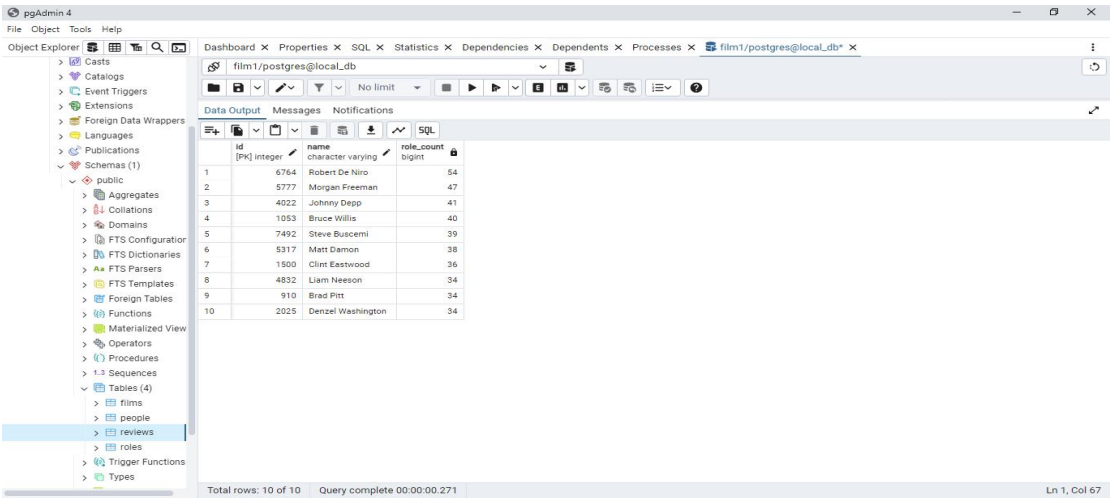
For a better understanding of role distributions, identification of key personnel, and optimization of resource allocation, I leveraged SQL to answer the real-life business questions related to the film industry. Below are the questions;

Question 3.1: Determine the top 10 people with the most roles in the database.

SQL Code:



Results:

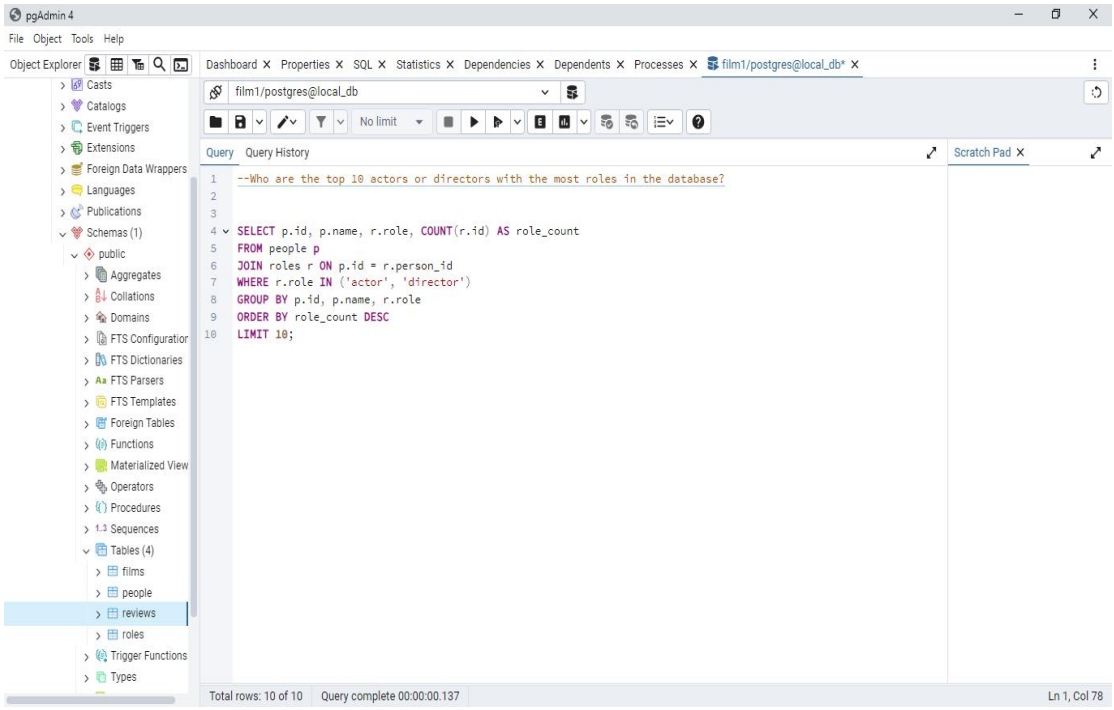


	id [PK] integer	name character varying	role_count bigint
1	6764	Robert De Niro	54
2	5777	Morgan Freeman	47
3	4022	Johnny Depp	41
4	1053	Bruce Willis	40
5	7492	Steve Buscemi	39
6	5317	Matt Damon	38
7	1500	Clint Eastwood	36
8	4832	Liam Neeson	34
9	910	Brad Pitt	34
10	2025	Denzel Washington	34

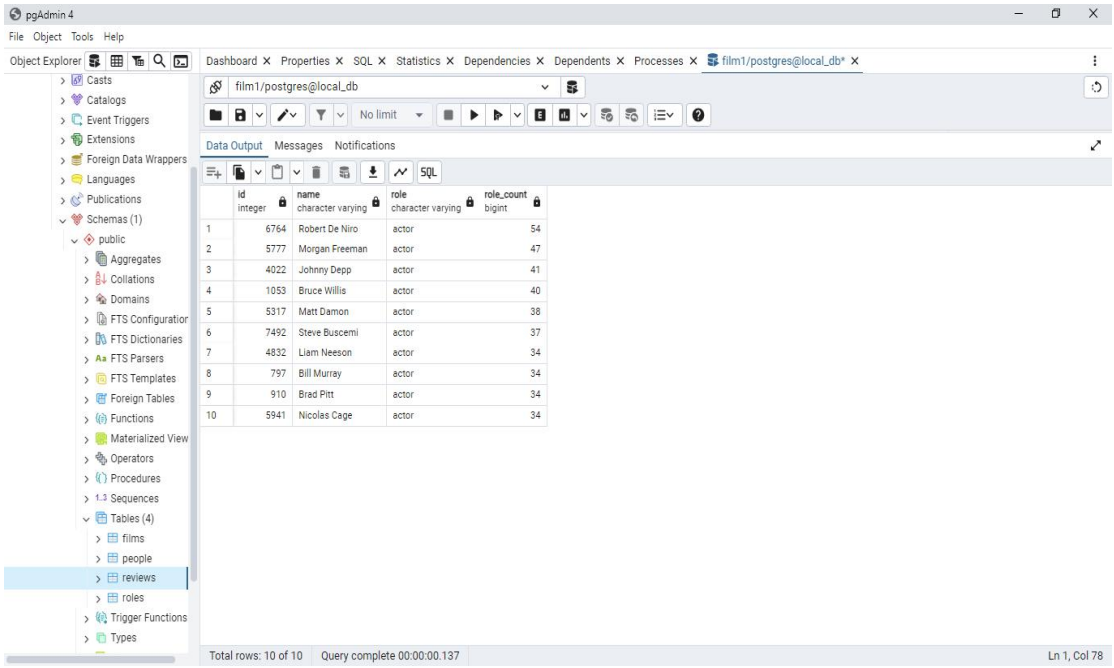
Analysis: p.id and p.name are selected from the people table. COUNT (r.id) counts the number of roles each person has. The JOIN clause connects the people table with the roles table using person_id. GROUP BY groups the results by person. ORDER BY role_count DESC sorts the people by their role count in descending order. LIMIT 10 shows only the top ten people.

Question 3.2: Who are the top 10 actors or directors with the most roles in the database?

SQL Code:



Results:

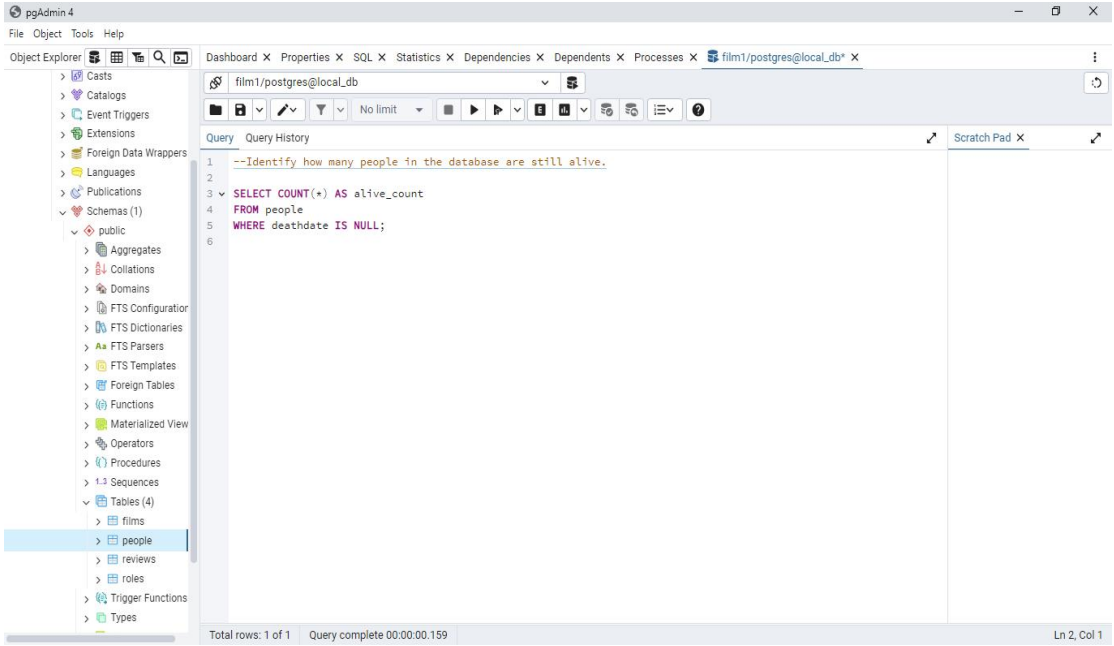


	id integer	name character varying	role character varying	role_count bigint
1	6764	Robert De Niro	actor	54
2	5777	Morgan Freeman	actor	47
3	4022	Johnny Depp	actor	41
4	1053	Bruce Willis	actor	40
5	5317	Matt Damon	actor	38
6	7492	Steve Buscemi	actor	37
7	4832	Liam Neeson	actor	34
8	797	Bill Murray	actor	34
9	910	Brad Pitt	actor	34
10	5941	Nicolas Cage	actor	34

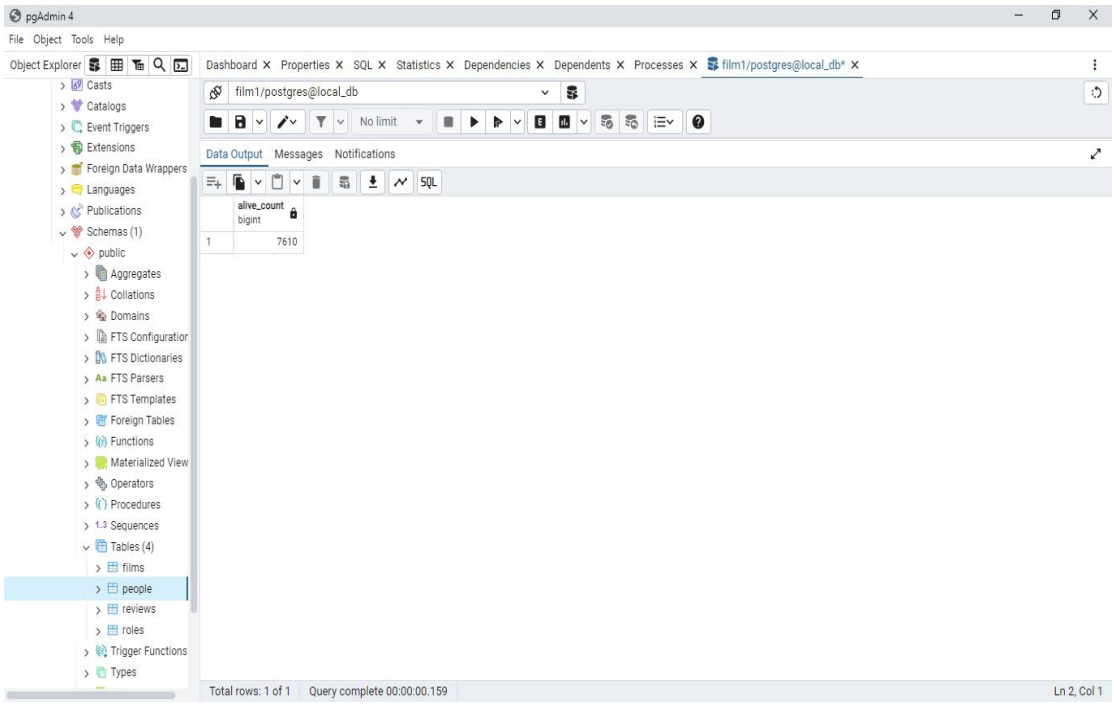
Analysis: This data shows the list of the top 10 people with the most roles.

Question 3.3: Identify how many people in the database are still alive.

SQL Code:



Results:



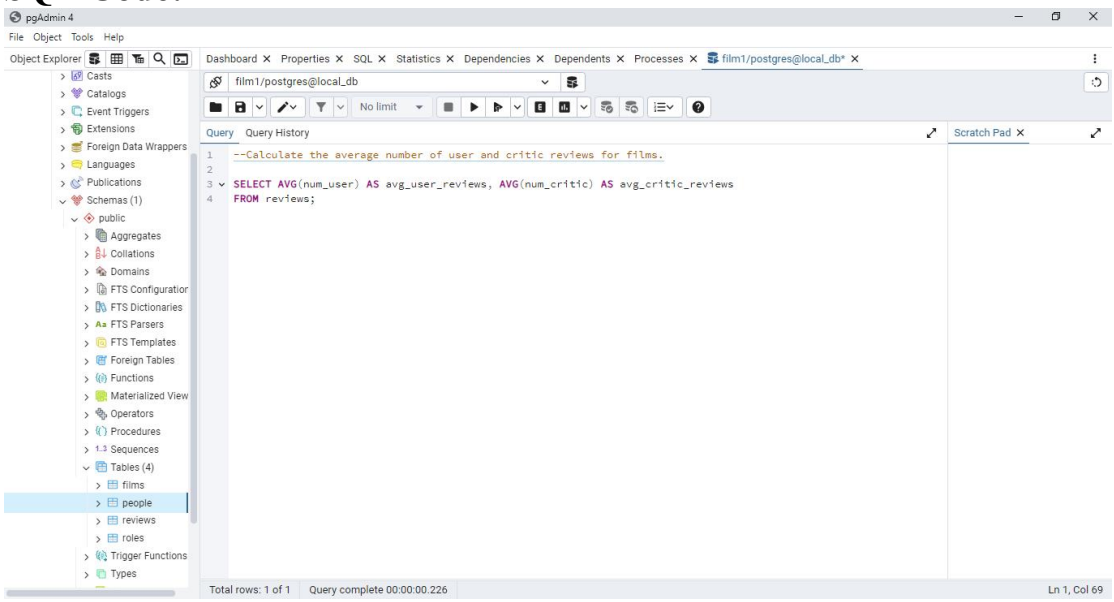
Analysis: This query counts the number of people alive in the database.

Task 4: User Reviews and Votes Analysis

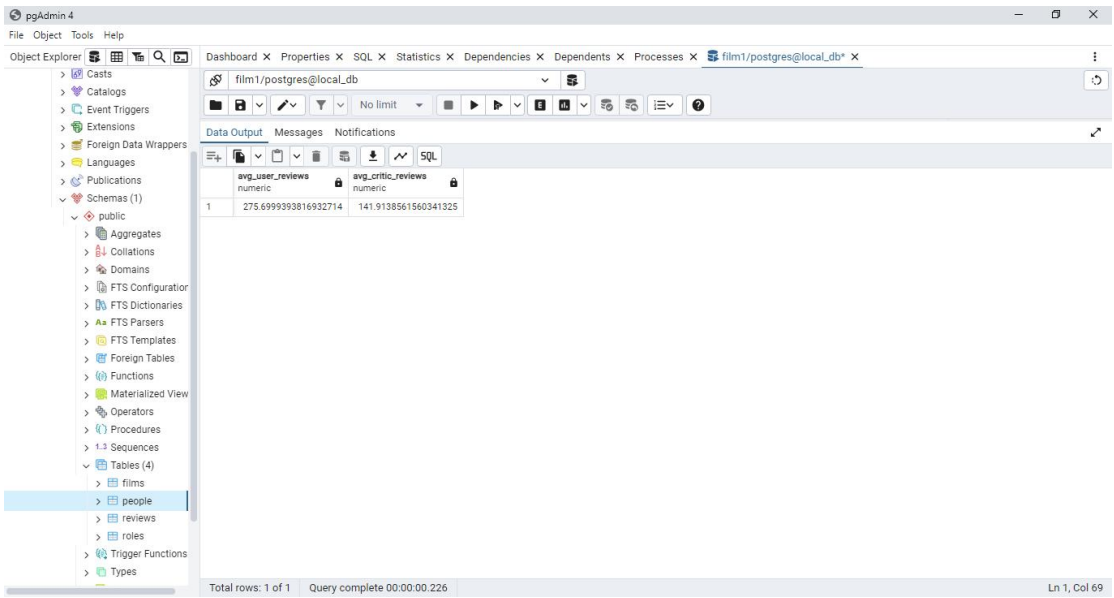
In order to gain insights into customer satisfaction, preferences, and sentiments, I utilized SQL to analyze user reviews and voting patterns. This analysis helps in understanding customer behaviour, identifying key trends, and optimizing product or service offerings. Below are the questions addressed:

Question 4.1: Calculate the average number of user and critic reviews for films?

SQL Code:



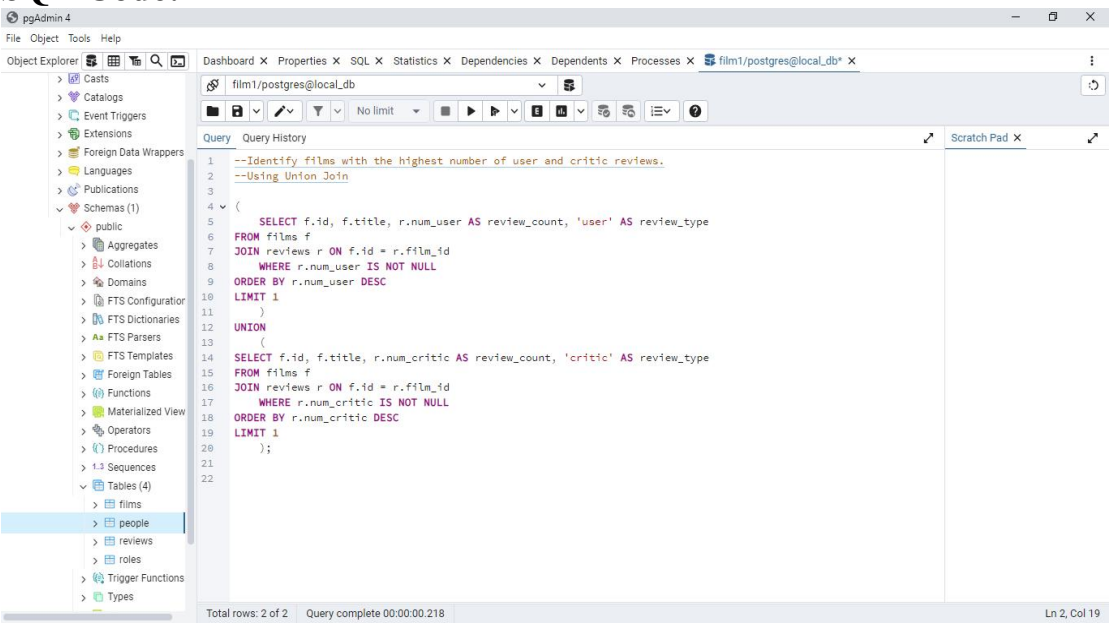
Results:



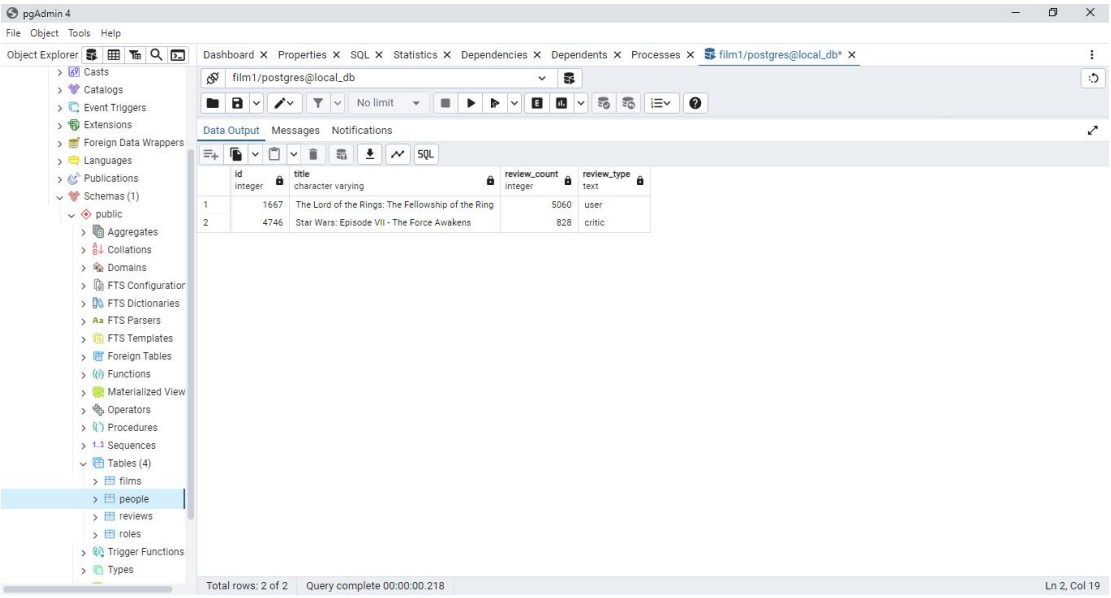
Analysis: The query result shows the average number of user and critic reviews for films in the database.

Question 4.2: Identify films with the highest number of user and critic reviews?

SQL Code:



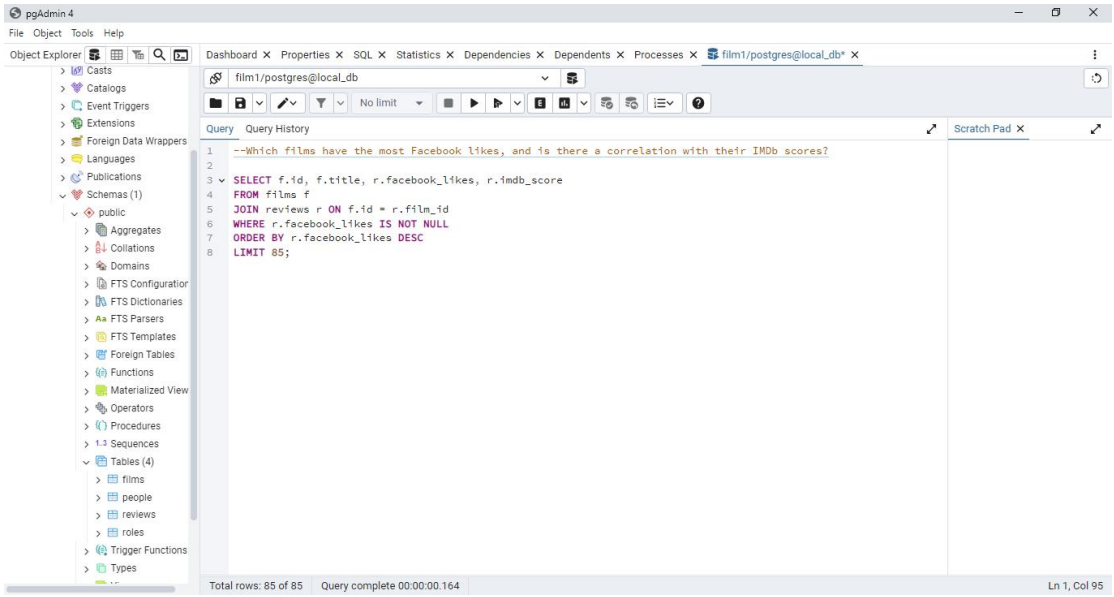
Results:



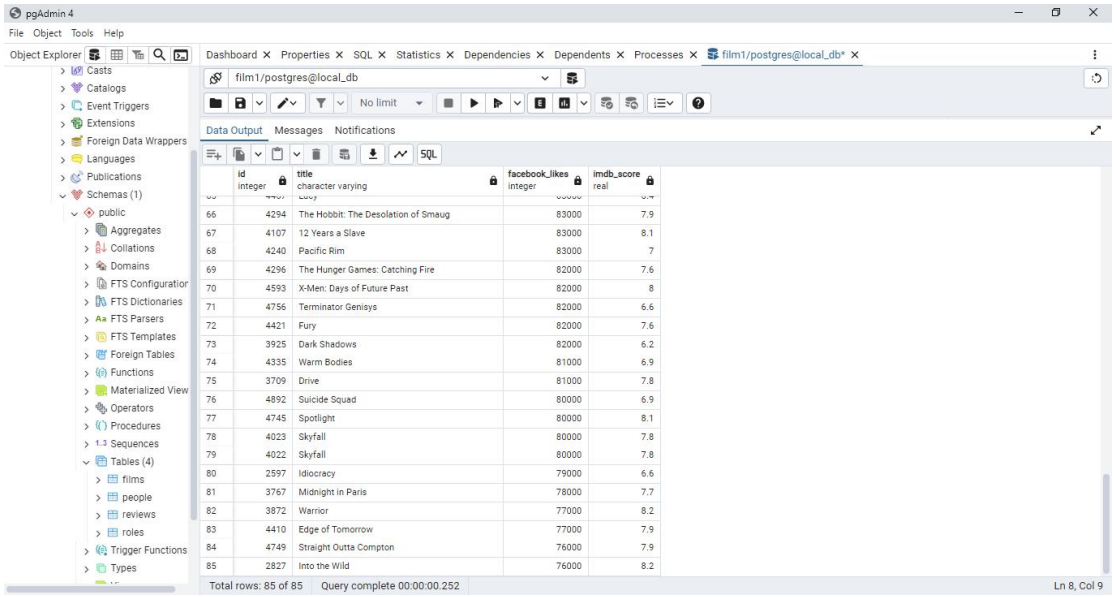
Analysis: The WHERE clauses r.num_user IS NOT NULL and r.num_critic IS NOT NULL filter out rows where the number of user or critic reviews is null. The data identifies the films with the highest number of user and critic reviews.

Question 4.3: Which films have the most Facebook likes, and is there a correlation with their IMDb scores?

SQL Code:



Results:



id integer	title character varying	facebook_likes integer	imdb_score real
66	4294	The Hobbit: The Desolation of Smaug	83000	7.9
67	4107	12 Years a Slave	83000	8.1
68	4240	Pacific Rim	83000	7
69	4296	The Hunger Games: Catching Fire	82000	7.6
70	4593	X-Men: Days of Future Past	82000	8
71	4756	Terminator Genisys	82000	6.6
72	4421	Fury	82000	7.6
73	3925	Dark Shadows	82000	6.2
74	4335	Warm Bodies	81000	6.9
75	3709	Drive	81000	7.8
76	4892	Suicide Squad	80000	6.9
77	4745	Spotlight	80000	8.1
78	4023	Skyfall	80000	7.8
79	4022	Skyfall	80000	7.8
80	2597	Idiocracy	79000	6.6
81	3767	Midnight in Paris	78000	7.7
82	3872	Warrior	77000	8.2
83	4410	Edge of Tomorrow	77000	7.9
84	4749	Straight Outta Compton	76000	7.9
85	2827	Into the Wild	76000	8.2

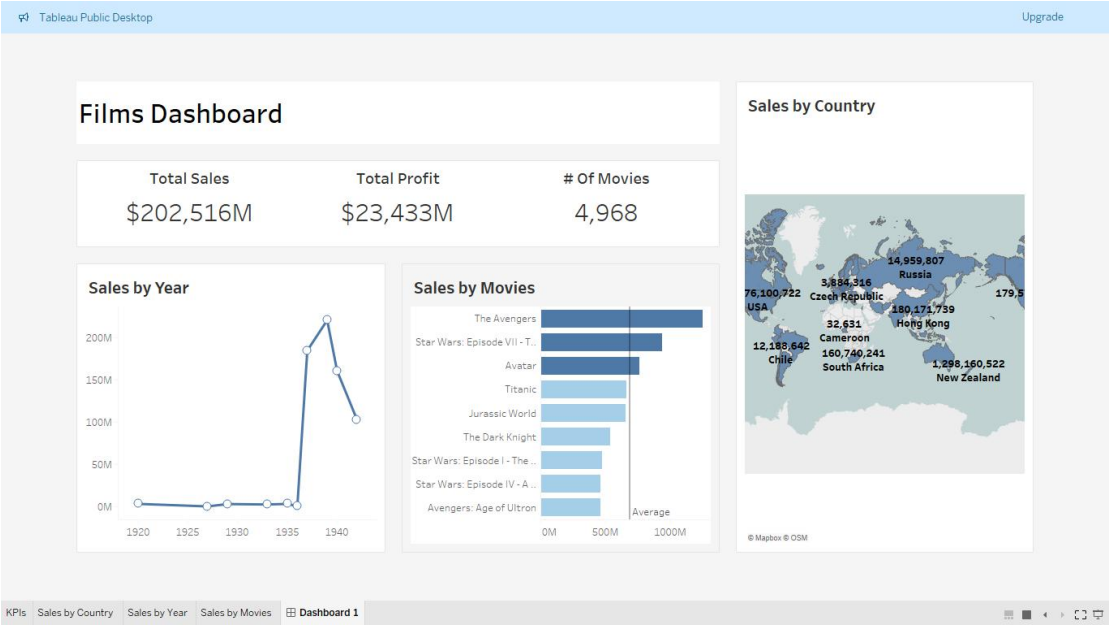
Analysis:

Here the data result shows the films which have the most Facebook likes and IMDb scores. In order to get the correlation between Facebook likes and IMDb scores, I took a limit of 85 records as a sample size out of 4968 records of the both variables, to get a medium effect size.

The datasets was imported to Excel, where I further used the correlation function to get the correlation between Facebook likes and IMDb scores, and my result was 0.16220513.

The value 0.16220513 is relatively close to 0, which indicates a weak correlation. This Suggests that the relationship between Facebook likes and IMDb scores is not very strong. While there is a positive trend, it is not pronounced, and many other factors likely influence IMDB scores independently of Facebook likes.

4. Visualizations



In this section, I created a film dashboard using Tableau to present key performance indicators (KPIs) and visualizations based on the film database given. The dashboard includes:

1) KPIs:

- Total sales (Gross) for movies
- Total profit
- Number of Movies

2) Visualizations:

- Sales by Year (Trend Line)
- Sales by Movie (Bar Chart)
- Sales by Country (Map)

5. Conclusion

Using Tableau, I developed this dashboard to provide clear insights into film performance. Key observations include:

- **Sales Trends Over Time:** The trend line reveals fluctuations in yearly sales, indicating varying periods of revenue.
- Top-Grossing Films: The bar chart highlights the most financially successful movies.
- **Geographical Revenue Distribution:** The map shows revenue by country, aiding in strategic planning.

Overall, this dashboard helps productive decisions towards optimizing production, marketing, and distribution strategies in the film industry.

6. References

TechWay Consult. (2024). Film database. TechWay Consult.