IMDB Movie Analysis

Name: Prince Kumar | Email: prince22495@gmail.com

Project Description:

Problem Statement:

The IMDB movie dataset is a large and complex dataset. It contains a variety of information about movies, including the director, title, year, actor, language, number of votes, IMDb score and various other columns.

This dataset can be used to answer a variety of questions about movies, such as:

- Which movies have the highest profit?
- What are the top 250 movies?
- What are the top movies in foreign language?
- Who are the best directors?
- What are the most popular genres?

Approach:

The project will be divided into three phases:

1. Data cleaning

The first phase of the project will involve cleaning the dataset. This will include removing errors and missing values, and formatting the data so that it is consistent.

Tools: Excel formulas, SQL queries

Tasks: Identify and remove errors, identify and fill in missing values, format data

2. Data analysis

The second phase of the project will involve analyzing the dataset. This will include answering questions about movies, such as which movies have the highest profit, what are the top 250 movies, and who are the best directors.

Tools: Excel formulas, SQL queries, data visualization tools

Tasks: Ask and answer questions about movies, create visualizations of the data

3. Data visualization

The third phase of the project will involve creating visualizations of the data. This will help to communicate the results of the analysis and make it easier to understand.

Tools: MS Excel for charts.

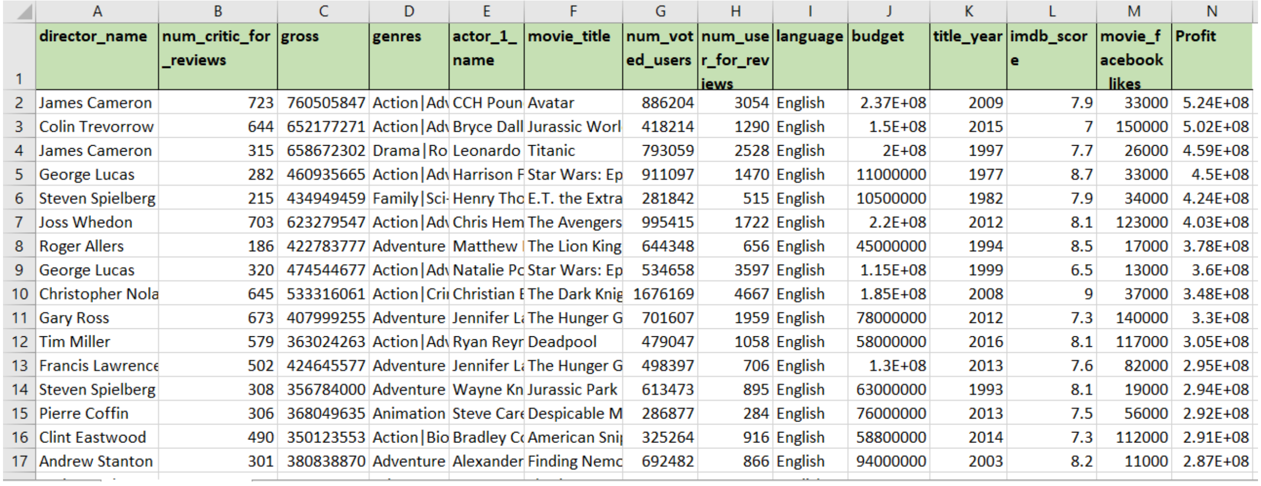
Tasks: Create visualizations of the data, communicate the results of the analysis

Tech-Stack Used:

- MS Excel
- MYSQL

Analysis & Insights:

A. Cleaning the data:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook likes	Profit
2	James Cameron	723	760505847	Action|Adv	CCH Poun	Avatar	886204	3054	English	2.37E+08	2009	7.9	33000	5.24E+08
3	Colin Trevorrow	644	652177271	Action|Adv	Bryce Dall	Jurassic Worl	418214	1290	English	1.5E+08	2015	7	150000	5.02E+08
4	James Cameron	315	658672302	Drama|Ro	Leonardo	Titanic	793059	2528	English	2E+08	1997	7.7	26000	4.59E+08
5	George Lucas	282	460935665	Action|Adv	Harrison F	Star Wars: Ep	911097	1470	English	11000000	1977	8.7	33000	4.5E+08
6	Steven Spielberg	215	434949459	Family|Sci-	Henry Thc	E.T. the Extra	281842	515	English	10500000	1982	7.9	34000	4.24E+08
7	Joss Whedon	703	623279547	Action|Adv	Chris Hem	The Avengers	995415	1722	English	2.2E+08	2012	8.1	123000	4.03E+08
8	Roger Allers	186	422783777	Adventure	Matthew	The Lion King	644348	656	English	45000000	1994	8.5	17000	3.78E+08
9	George Lucas	320	474544677	Action|Adv	Natalie Pc	Star Wars: Ep	534658	3597	English	1.15E+08	1999	6.5	13000	3.6E+08
10	Christopher Nola	645	533316061	Action|Cri	Christian E	The Dark Knig	1676169	4667	English	1.85E+08	2008	9	37000	3.48E+08
11	Gary Ross	673	407999255	Adventure	Jennifer La	The Hunger G	701607	1959	English	78000000	2012	7.3	140000	3.3E+08
12	Tim Miller	579	363024263	Action|Adv	Ryan Reyr	Deadpool	479047	1058	English	58000000	2016	8.1	117000	3.05E+08
13	Francis Lawrence	502	424645577	Adventure	Jennifer La	The Hunger G	498397	706	English	1.3E+08	2013	7.6	82000	2.95E+08
14	Steven Spielberg	308	356784000	Adventure	Wayne Kn	Jurassic Park	613473	895	English	63000000	1993	8.1	19000	2.94E+08
15	Pierre Coffin	306	368049635	Animation	Steve Care	Despicable M	286877	284	English	76000000	2013	7.5	56000	2.92E+08
16	Clint Eastwood	490	350123553	Action|Bio	Bradley Co	American Snip	325264	916	English	58800000	2014	7.3	112000	2.91E+08
17	Andrew Stanton	301	380838870	Adventure	Alexander	Finding Nemc	692482	866	English	94000000	2003	8.2	11000	2.87E+08

I cleaned the dataset using the following steps:

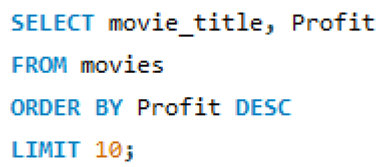
1. I used the filter function in Excel to filter out all the blank cells in each column.
2. I deleted all the blank cells from each column.
3. I deleted a few columns that were not required for my analysis.
4. I removed duplicates from the dataset.

After completing these steps, I was left with 14 columns and 3724 rows. The dataset was now clean and ready to be imported into MySQL for further analysis.

B. Movies with highest profit:

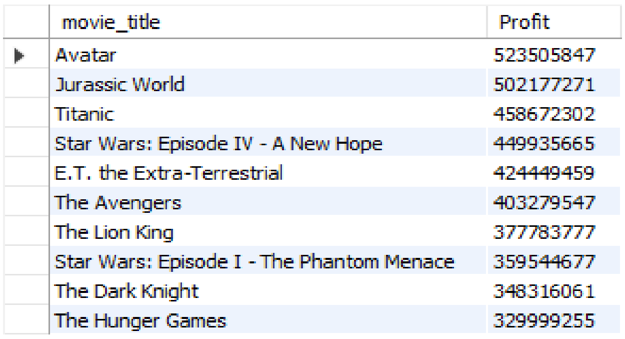
Task: Find the movies with the highest profit?

SQL query:



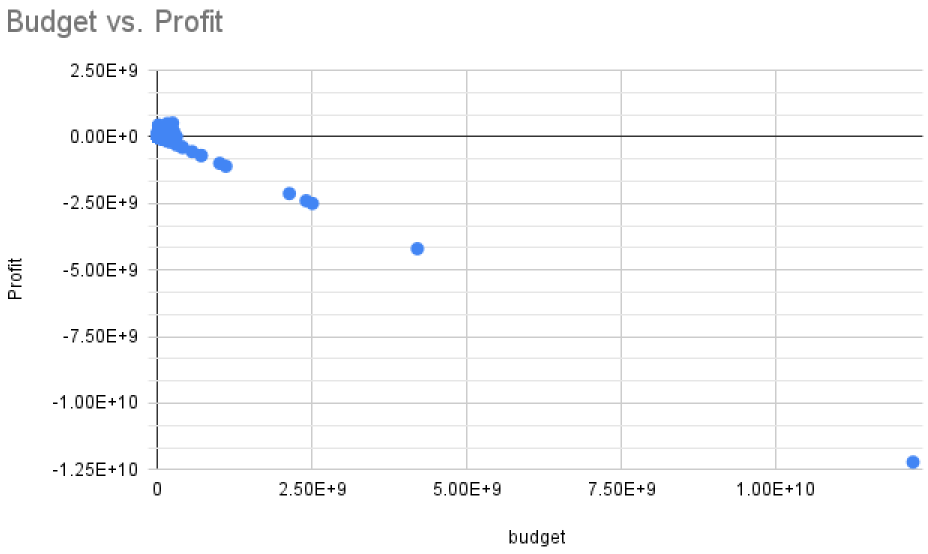
```
SELECT movie_title, Profit
FROM movies
ORDER BY Profit DESC
LIMIT 10;
```

Output table:



movie_title	Profit
Avatar	523505847
Jurassic World	502177271
Titanic	458672302
Star Wars: Episode IV - A New Hope	449935665
E.T. the Extra-Terrestrial	424449459
The Avengers	403279547
The Lion King	377783777
Star Wars: Episode I - The Phantom Menace	359544677
The Dark Knight	348316061
The Hunger Games	329999255

Scatter Plot:



Insights:

Movie Avatar has made the highest profit.

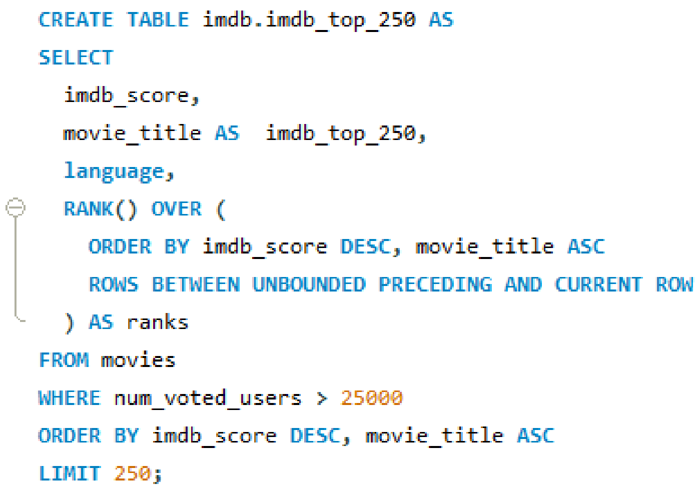
In this scatter plot, we can see an outlier at around -1.25E+10 point.

C. What are the top 250 IMDB movies?

Task: Find IMDB Top 250

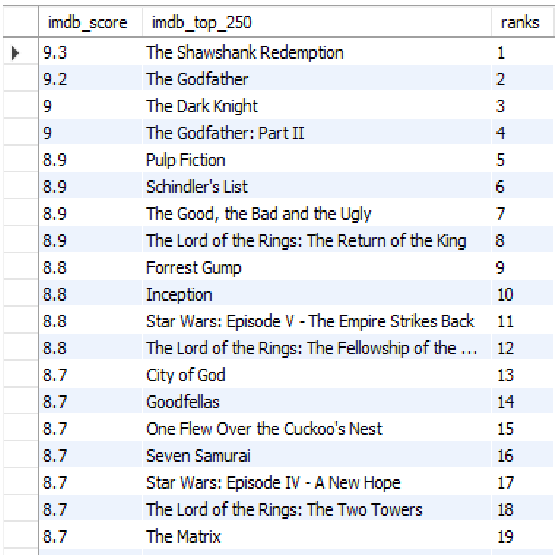
Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

SQL query:



```
CREATE TABLE imdb.imdb_top_250 AS
SELECT
  imdb_score,
  movie_title AS  imdb_top_250,
  language,
  RANK() OVER (
    ORDER BY imdb_score DESC, movie_title ASC
    ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW
  ) AS ranks
FROM movies
WHERE num_voted_users > 25000
ORDER BY imdb_score DESC, movie_title ASC
LIMIT 250;
```

Output Table:



imdb_score	imdb_top_250	ranks
9.3	The Shawshank Redemption	1
9.2	The Godfather	2
9	The Dark Knight	3
9	The Godfather: Part II	4
8.9	Pulp Fiction	5
8.9	Schindler's List	6
8.9	The Good, the Bad and the Ugly	7
8.9	The Lord of the Rings: The Return of the King	8
8.8	Forrest Gump	9
8.8	Inception	10
8.8	Star Wars: Episode V - The Empire Strikes Back	11
8.8	The Lord of the Rings: The Fellowship of the ...	12
8.7	City of God	13
8.7	Goodfellas	14
8.7	One Flew Over the Cuckoo's Nest	15
8.7	Seven Samurai	16
8.7	Star Wars: Episode IV - A New Hope	17
8.7	The Lord of the Rings: The Two Towers	18
8.7	The Matrix	19

Insights:

First movie is "The Shawshank Redemption" with a 9.3 IMDB score, among the top 250 IMDB movies list.

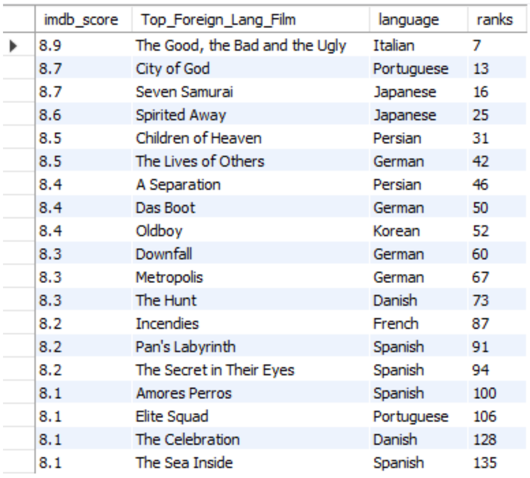
Top IMDB Foreign Language Movies:

Task: Extract all the movies in the IMDb_Top_250 column, which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

SQL query:

```
SELECT imdb_score, imdb_top_250 as Top_Foreign_Lang_Film, language, ranks
FROM imdb_top_250
WHERE language <> "English";
```

Output Table:



imdb_score	Top_Foreign_Lang_Film	language	ranks
8.9	The Good, the Bad and the Ugly	Italian	7
8.7	City of God	Portuguese	13
8.7	Seven Samurai	Japanese	16
8.6	Spirited Away	Japanese	25
8.5	Children of Heaven	Persian	31
8.5	The Lives of Others	German	42
8.4	A Separation	Persian	46
8.4	Das Boot	German	50
8.4	Oldboy	Korean	52
8.3	Downfall	German	60
8.3	Metropolis	German	67
8.3	The Hunt	Danish	73
8.2	Incendies	French	87
8.2	Pan's Labyrinth	Spanish	91
8.2	The Secret in Their Eyes	Spanish	94
8.1	Amores Perros	Spanish	100
8.1	Elite Squad	Portuguese	106
8.1	The Celebration	Danish	128
8.1	The Sea Inside	Spanish	135

Insights:

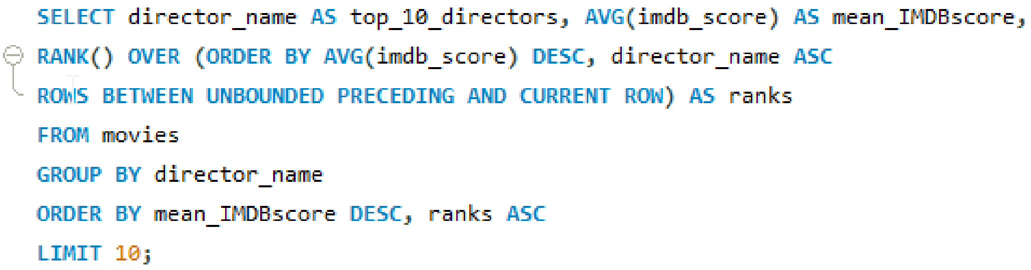
Top foreign Language Movie came out to be “The Good, the Bad and the Ugly”.

D. Who are the best directors?

Task: Find the best directors

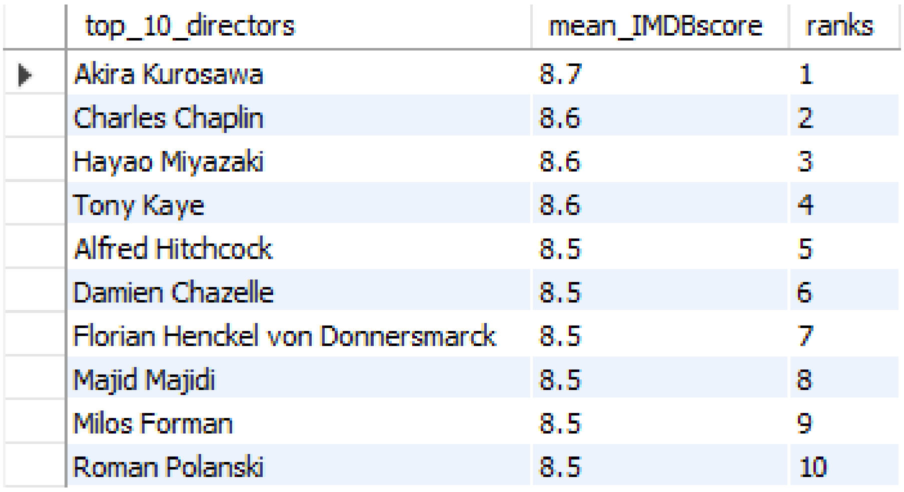
Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.

SQL query:



```
SELECT director_name AS top_10_directors, AVG(imdb_score) AS mean_IMDBscore,
RANK() OVER (ORDER BY AVG(imdb_score) DESC, director_name ASC
ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW) AS ranks
FROM movies
GROUP BY director_name
ORDER BY mean_IMDBscore DESC, ranks ASC
LIMIT 10;
```

Output Table:



top_10_directors	mean_IMDBscore	ranks
Akira Kurosawa	8.7	1
Charles Chaplin	8.6	2
Hayao Miyazaki	8.6	3
Tony Kaye	8.6	4
Alfred Hitchcock	8.5	5
Damien Chazelle	8.5	6
Florian Henckel von Donnersmarck	8.5	7
Majid Majidi	8.5	8
Milos Forman	8.5	9
Roman Polanski	8.5	10

Insights:

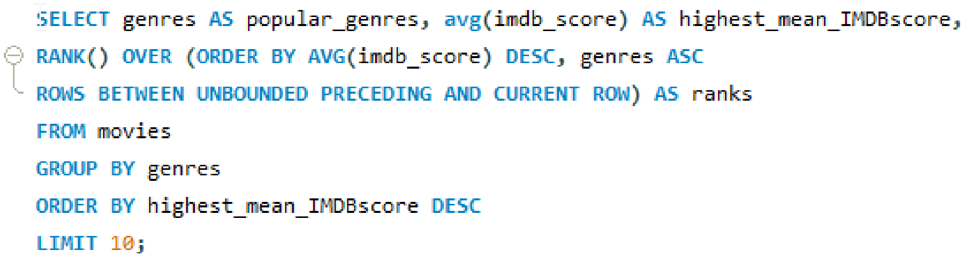
The best director is Akira Kurosawa with the mean IMDB score of 8.7.

E. What are the most popular genres?

Task: Find popular genres

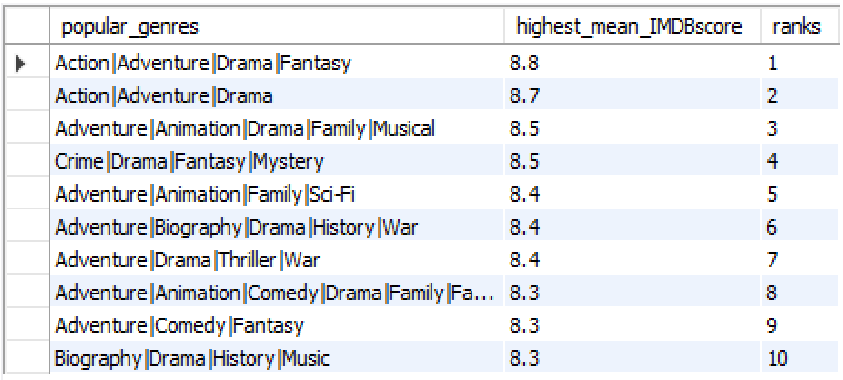
Perform this step using the knowledge gained while performing previous steps.

SQL query:



```
SELECT genres AS popular_genres, avg(imdb_score) AS highest_mean_IMDBscore,
RANK() OVER (ORDER BY AVG(imdb_score) DESC, genres ASC
ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW) AS ranks
FROM movies
GROUP BY genres
ORDER BY highest_mean_IMDBscore DESC
LIMIT 10;
```

Output Table:



popular_genres	highest_mean_IMDBscore	ranks
Action|Adventure|Drama|Fantasy	8.8	1
Action|Adventure|Drama	8.7	2
Adventure|Animation|Drama|Family|Musical	8.5	3
Crime|Drama|Fantasy|Mystery	8.5	4
Adventure|Animation|Family|Sci-Fi	8.4	5
Adventure|Biography|Drama|History|War	8.4	6
Adventure|Drama|Thriller|War	8.4	7
Adventure|Animation|Comedy|Drama|Family|Fa...	8.3	8
Adventure|Comedy|Fantasy	8.3	9
Biography|Drama|History|Music	8.3	10

Insights:

Action|Adventure|Drama|Fantasy came out to be the most popular genre with the highest mean IMDB score of 8.8.

F. Charts

Task: Find the critic-favorite and audience-favorite actors

Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

Append the rows of all these columns and store them in a new column named Combined.

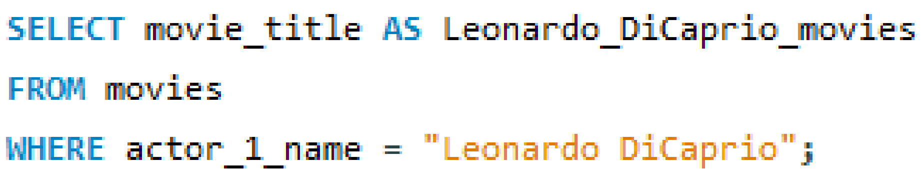
Group the combined column using the actor_1_name column.

Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

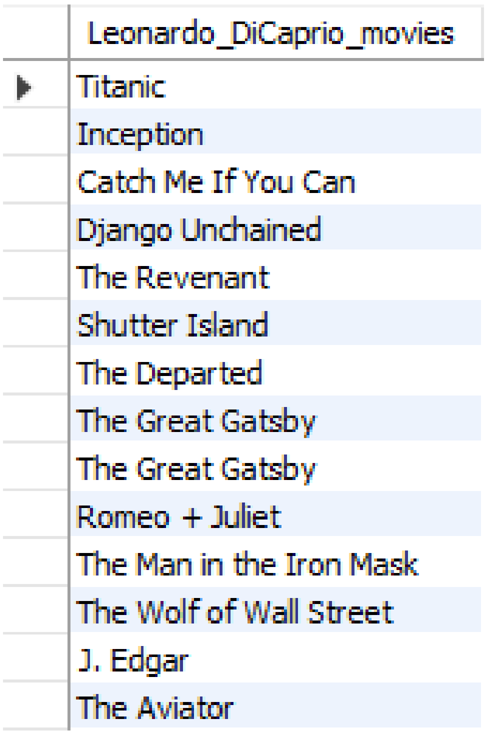
Leonardo DiCaprio Movies List

SQL query:



```
SELECT movie_title AS Leonardo_DiCaprio_movies
FROM movies
WHERE actor_1_name = "Leonardo DiCaprio";
```

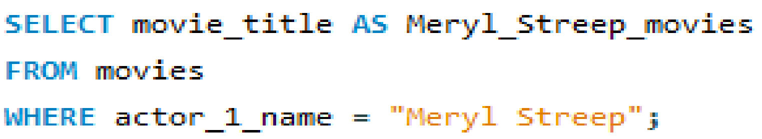
Output Table:



Leonardo_DiCaprio_movies
Titanic
Inception
Catch Me If You Can
Django Unchained
The Revenant
Shutter Island
The Departed
The Great Gatsby
The Great Gatsby
Romeo + Juliet
The Man in the Iron Mask
The Wolf of Wall Street
J. Edgar
The Aviator

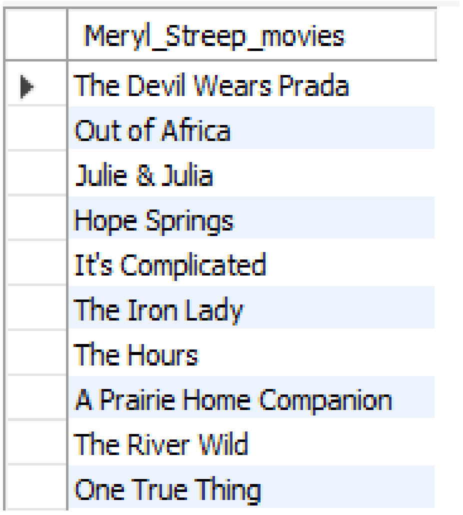
Meryl Streep Movies List

SQL query:



```
SELECT movie_title AS Meryl_Streep_movies
FROM movies
WHERE actor_1_name = "Meryl Streep";
```

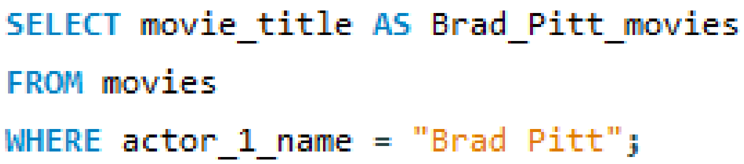
Output Table:



	Meryl_Streep_movies
▶	The Devil Wears Prada
	Out of Africa
	Julie & Julia
	Hope Springs
	It's Complicated
	The Iron Lady
	The Hours
	A Prairie Home Companion
	The River Wild
	One True Thing

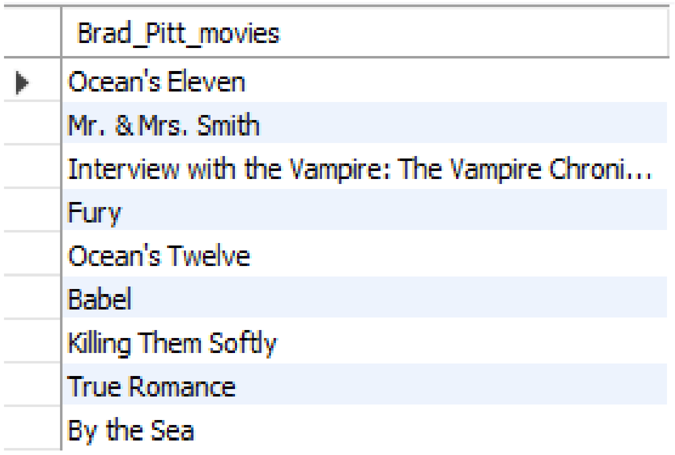
Brad Pitt Movies List

SQL query:



```
SELECT movie_title AS Brad_Pitt_movies
FROM movies
WHERE actor_1_name = "Brad Pitt";
```

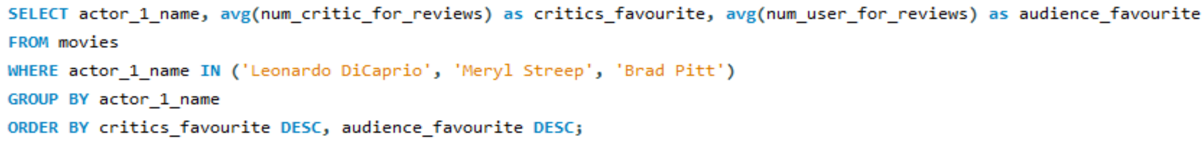
Output Table:



	Brad_Pitt_movies
▶	Ocean's Eleven
	Mr. & Mrs. Smith
	Interview with the Vampire: The Vampire Chroni...
	Fury
	Ocean's Twelve
	Babel
	Killing Them Softly
	True Romance
	By the Sea

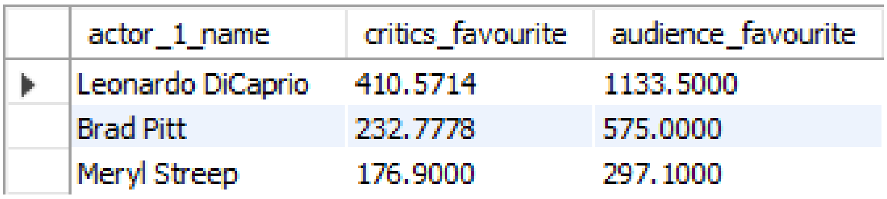
Critic-favourite and audience-favourite actors

SQL query:



```
SELECT actor_1_name, avg(num_critic_for_reviews) as critics_favourite, avg(num_user_for_reviews) as audience_favourite
FROM movies
WHERE actor_1_name IN ('Leonardo DiCaprio', 'Meryl Streep', 'Brad Pitt')
GROUP BY actor_1_name
ORDER BY critics_favourite DESC, audience_favourite DESC;
```

Output Table:



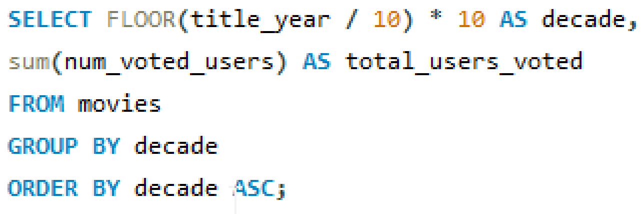
actor_1_name	critics_favourite	audience_favourite
Leonardo DiCaprio	410.5714	1133.5000
Brad Pitt	232.7778	575.0000
Meryl Streep	176.9000	297.1000

Insights:

Leonardo DiCaprio is the critics favourite as well as audience favourite actor, having the most number of reviews in from both critics and audience.

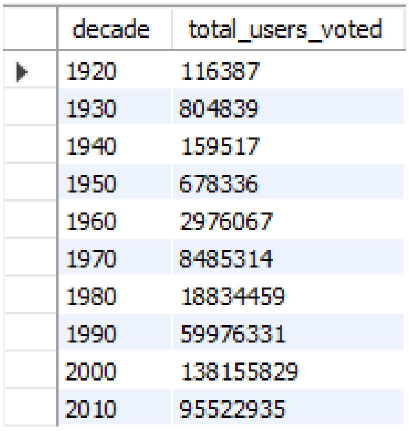
Change in number of voted users over decades

SQL query:



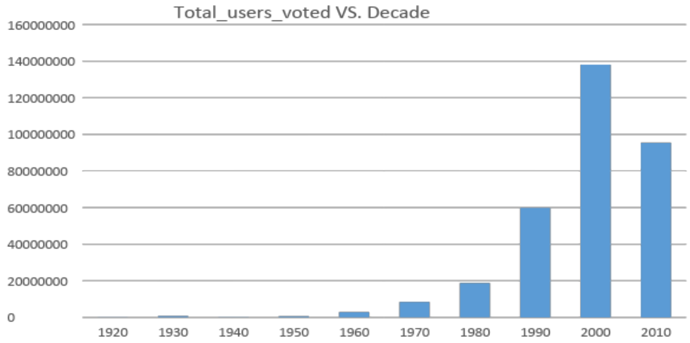
```
SELECT FLOOR(title_year / 10) * 10 AS decade,
sum(num_voted_users) AS total_users_voted
FROM movies
GROUP BY decade
ORDER BY decade ASC;
```

Output Table:



decade	total_users_voted
1920	116387
1930	804839
1940	159517
1950	678336
1960	2976067
1970	8485314
1980	18834459
1990	59976331
2000	138155829
2010	95522935

Bar Graph:



Insights:
Number of voted users are the most in 2000s.

Conclusion

- After cleaning the dataset, I was left with 14 columns and 3724 rows. The dataset was now clean and ready to be imported into MySQL for further analysis.
- Movie Avatar has made the highest profit.
- First movie is “The Shawshank Redemption” with a 9.3 IMDB score, among the top 250 IMDB movies list.
- Top foreign Language Movie came out to be “The Good, the Bad and the Ugly”.
- The best director is Akira Kurosawa with the mean IMDB score of 8.7.
- Action|Adventure|Drama|Fantasy came out to be the most popular genre with the highest mean IMDB score of 8.8.
- Leonardo DiCaprio is the critics favourite as well as audience favourite actor, having the most number of reviews in from both critics and audience.
- Number of voted users are the most in 2000s.

File Links:

Project Link: Click Here!

Working File Link: csv file SQL file