IMDB Movie Analysis

(Data Analyst Project)

PROJECT DESCRIPTION

The goal of this project is to analyze a dataset of IMDB movies and draw insights from the data. The dataset includes various columns such as movie names, budgets, gross revenue, and IMDB ratings. To complete the project, you will need to use a combination of Excel formulas and SQL commands to clean and manipulate the data. You will be asked to complete specific tasks, such as identifying the movie with the highest profit or the top IMDB movies, as well as share your own insights by identifying any problems or trends in the data. You may also be asked to use charts and visualizations to present your findings. The overall objective of the project is to gain a better understanding of the movie industry by analyzing the data and drawing meaningful conclusions.

Approach:

1.Understand the data: Before beginning the analysis, I took some time to familiarize with the data. Look at the structure of the data and get a sense of the overall content. This help me identify any potential issues or challenges that I may need to address as I proceed with my analysis.

2.Check for missing or incomplete data: Make sure to check for any blank values or missing data in your dataset.

3.Identify and handle outliers: Outliers are data points that are significantly different from the rest of the data. They can have a significant impact on summary statistics and can distort the results of your analysis. It's important to identify any outliers and decide how to handle them, such as by excluding them from the analysis or by treating them as separate cases.

4.Communicate your findings: Once completed with analysis, present your findings to your audience in a clear and concise way. Use visualizations, such as charts and graphs, to help communicate your results. Be sure to clearly explain your methodology and the implications of your results.

Tech Stack Use:

MS-Excel, MySQL, PowerBI

A.Cleaning the data::

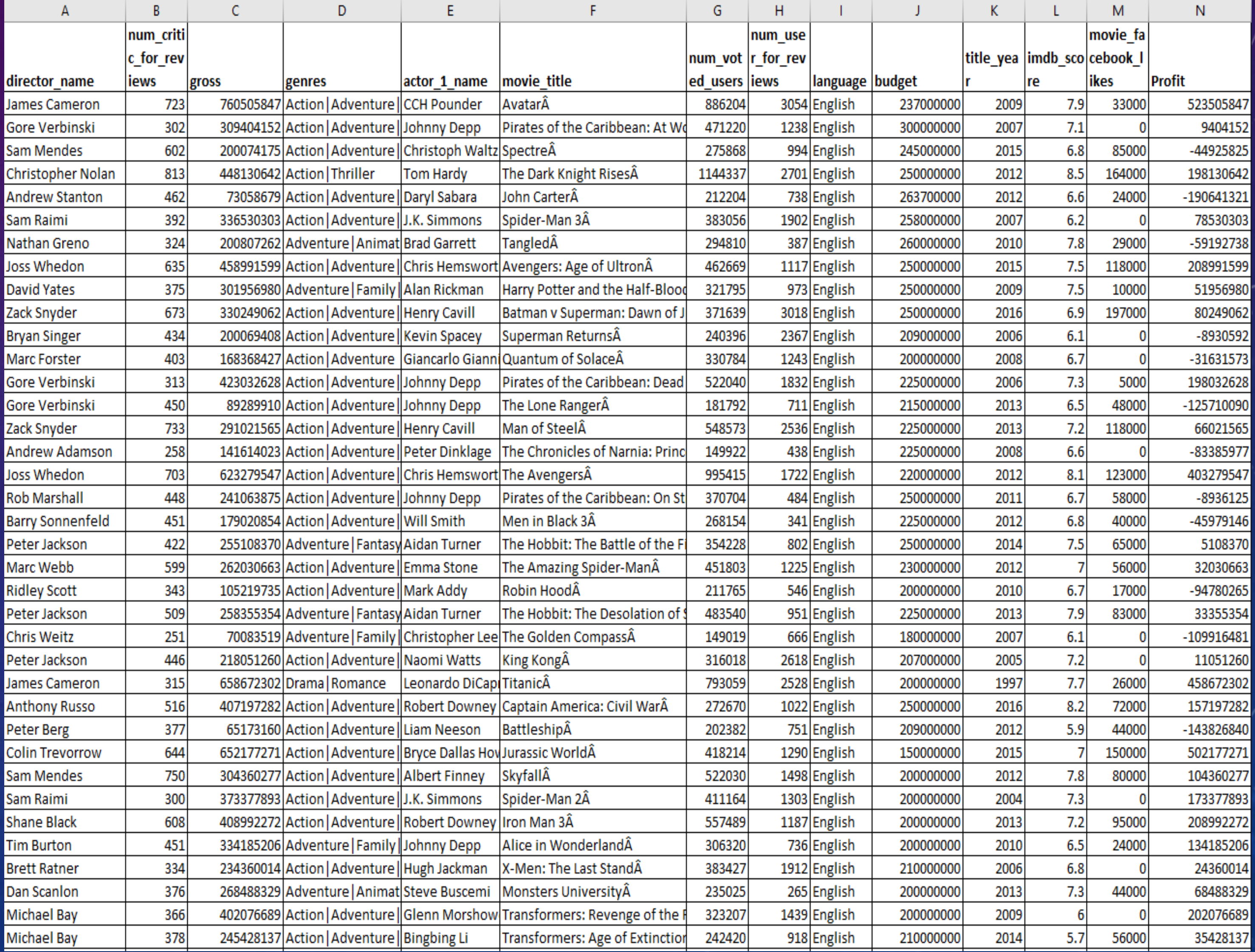
This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Your task: Clean the data.

Ans: To clean the data, I arranged the columns in the correct format and increased the column width to improve readability. I also removed null values and duplicates by selecting them using the "Find & Select" option and deleting the entire row. This helped to ensure that the data was accurate and ready for analysis.

Before Cleaning: 5044 rows and 28 columns

After Cleaning: 3850 rows and 13 columns



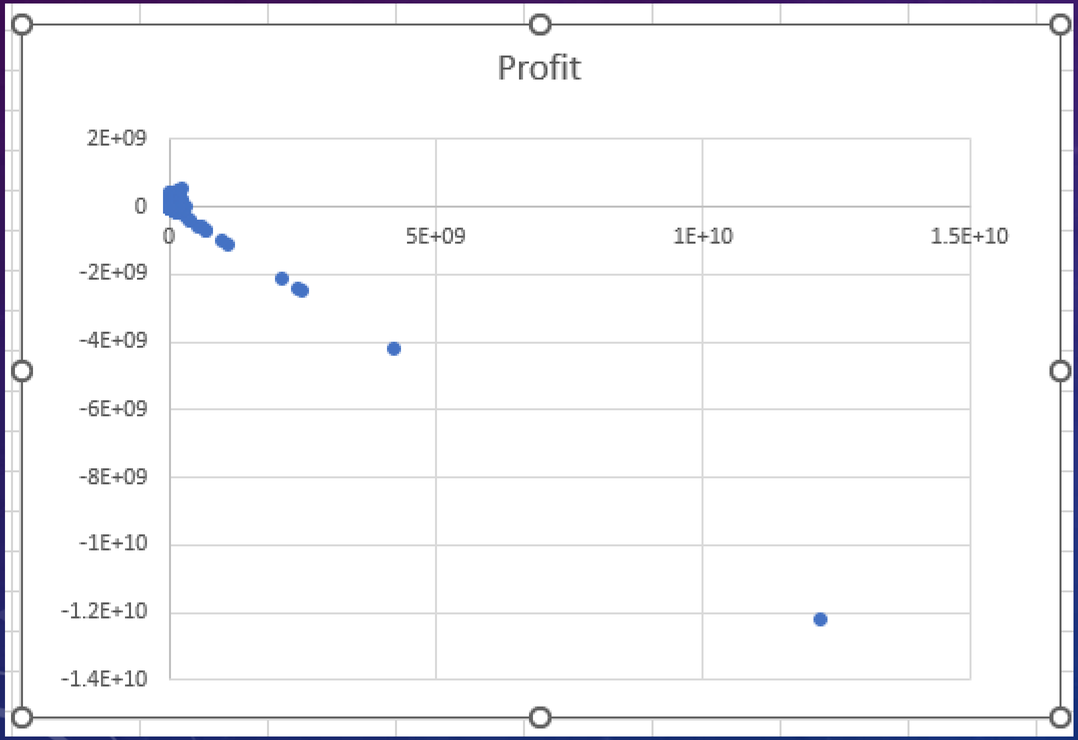
A	B	C	D	E	F	G	H	I	J	K	L	M	N
director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	budget	title_year	imdb_score	movie_facebook_likes	Profit
James Cameron	723	760505847	Action|Adventure|	CCH Pounder	AvatarÂ	886204	3054	English	237000000	2009	7.9	33000	523505847
Gore Verbinski	302	309404152	Action|Adventure|	Johnny Depp	Pirates of the Caribbean: At Wo	471220	1238	English	300000000	2007	7.1	0	9404152
Sam Mendes	602	200074175	Action|Adventure|	Christoph Waltz	SpectreÂ	275868	994	English	245000000	2015	6.8	85000	-44925825
Christopher Nolan	813	448130642	Action|Thriller	Tom Hardy	The Dark Knight RisesÂ	1144337	2701	English	250000000	2012	8.5	164000	198130642
Andrew Stanton	462	73058679	Action|Adventure|	Daryl Sabara	John CarterÂ	212204	738	English	263700000	2012	6.6	24000	-190641321
Sam Raimi	392	336530303	Action|Adventure|	J.K. Simmons	Spider-Man 3Â	383056	1902	English	258000000	2007	6.2	0	78530303
Nathan Greno	324	200807262	Adventure|Animat	Brad Garrett	TangledÂ	294810	387	English	260000000	2010	7.8	29000	-59192738
Joss Whedon	635	458991599	Action|Adventure|	Chris Hemswort	Avengers: Age of UltronÂ	462669	1117	English	250000000	2015	7.5	118000	208991599
David Yates	375	301956980	Adventure|Family|	Alan Rickman	Harry Potter and the Half-Blood	321795	973	English	250000000	2009	7.5	10000	51956980
Zack Snyder	673	330249062	Action|Adventure|	Henry Cavill	Batman v Superman: Dawn of J	371639	3018	English	250000000	2016	6.9	197000	80249062
Bryan Singer	434	200069408	Action|Adventure|	Kevin Spacey	Superman ReturnsÂ	240396	2367	English	209000000	2006	6.1	0	-8930592
Marc Forster	403	168368427	Action|Adventure	Giancarlo Gianni	Quantum of SolaceÂ	330784	1243	English	200000000	2008	6.7	0	-31631573
Gore Verbinski	313	423032628	Action|Adventure|	Johnny Depp	Pirates of the Caribbean: Dead	522040	1832	English	225000000	2006	7.3	5000	198032628
Gore Verbinski	450	89289910	Action|Adventure|	Johnny Depp	The Lone RangerÂ	181792	711	English	215000000	2013	6.5	48000	-125710090
Zack Snyder	733	291021565	Action|Adventure|	Henry Cavill	Man of SteelÂ	548573	2536	English	225000000	2013	7.2	118000	66021565
Andrew Adamson	258	141614023	Action|Adventure|	Peter Dinklage	The Chronicles of Narnia: Princ	149922	438	English	225000000	2008	6.6	0	-83385977
Joss Whedon	703	623279547	Action|Adventure|	Chris Hemswort	The AvengersÂ	995415	1722	English	220000000	2012	8.1	123000	403279547
Rob Marshall	448	241063875	Action|Adventure|	Johnny Depp	Pirates of the Caribbean: On St	370704	484	English	250000000	2011	6.7	58000	-8936125
Barry Sonnenfeld	451	179020854	Action|Adventure|	Will Smith	Men in Black 3Â	268154	341	English	225000000	2012	6.8	40000	-45979146
Peter Jackson	422	255108370	Adventure|Fantasy	Aidan Turner	The Hobbit: The Battle of the Fi	354228	802	English	250000000	2014	7.5	65000	5108370
Marc Webb	599	262030663	Action|Adventure|	Emma Stone	The Amazing Spider-ManÂ	451803	1225	English	230000000	2012	7	56000	32030663
Ridley Scott	343	105219735	Action|Adventure|	Mark Addy	Robin HoodÂ	211765	546	English	200000000	2010	6.7	17000	-94780265
Peter Jackson	509	258355354	Adventure|Fantasy	Aidan Turner	The Hobbit: The Desolation of S	483540	951	English	225000000	2013	7.9	83000	33355354
Chris Weitz	251	70083519	Adventure|Family|	Christopher Lee	The Golden CompassÂ	149019	666	English	180000000	2007	6.1	0	-109916481
Peter Jackson	446	218051260	Action|Adventure|	Naomi Watts	King KongÂ	316018	2618	English	207000000	2005	7.2	0	11051260
James Cameron	315	658672302	Drama|Romance	Leonardo DiCapr	TitanicÂ	793059	2528	English	200000000	1997	7.7	26000	458672302
Anthony Russo	516	407197282	Action|Adventure|	Robert Downey	Captain America: Civil WarÂ	272670	1022	English	250000000	2016	8.2	72000	157197282
Peter Berg	377	65173160	Action|Adventure|	Liam Neeson	BattleshipÂ	202382	751	English	209000000	2012	5.9	44000	-143826840
Colin Trevorrow	644	652177271	Action|Adventure|	Bryce Dallas Hov	Jurassic WorldÂ	418214	1290	English	150000000	2015	7	150000	502177271
Sam Mendes	750	304360277	Action|Adventure|	Albert Finney	SkyfallÂ	522030	1498	English	200000000	2012	7.8	80000	104360277
Sam Raimi	300	373377893	Action|Adventure|	J.K. Simmons	Spider-Man 2Â	411164	1303	English	200000000	2004	7.3	0	173377893
Shane Black	608	408992272	Action|Adventure|	Robert Downey	Iron Man 3Â	557489	1187	English	200000000	2013	7.2	95000	208992272
Tim Burton	451	334185206	Adventure|Family|	Johnny Depp	Alice in WonderlandÂ	306320	736	English	200000000	2010	6.5	24000	134185206
Brett Ratner	334	234360014	Action|Adventure|	Hugh Jackman	X-Men: The Last StandÂ	383427	1912	English	210000000	2006	6.8	0	24360014
Dan Scanlon	376	268488329	Adventure|Animat	Steve Buscemi	Monsters UniversityÂ	235025	265	English	200000000	2013	7.3	44000	68488329
Michael Bay	366	402076689	Action|Adventure|	Glenn Morshow	Transformers: Revenge of the F	323207	1439	English	200000000	2009	6	0	202076689
Michael Bay	378	245428137	Action|Adventure|	Bingbing Li	Transformers: Age of Extinction	242420	918	English	210000000	2014	5.7	56000	35428137

B. Movies with highest profit:

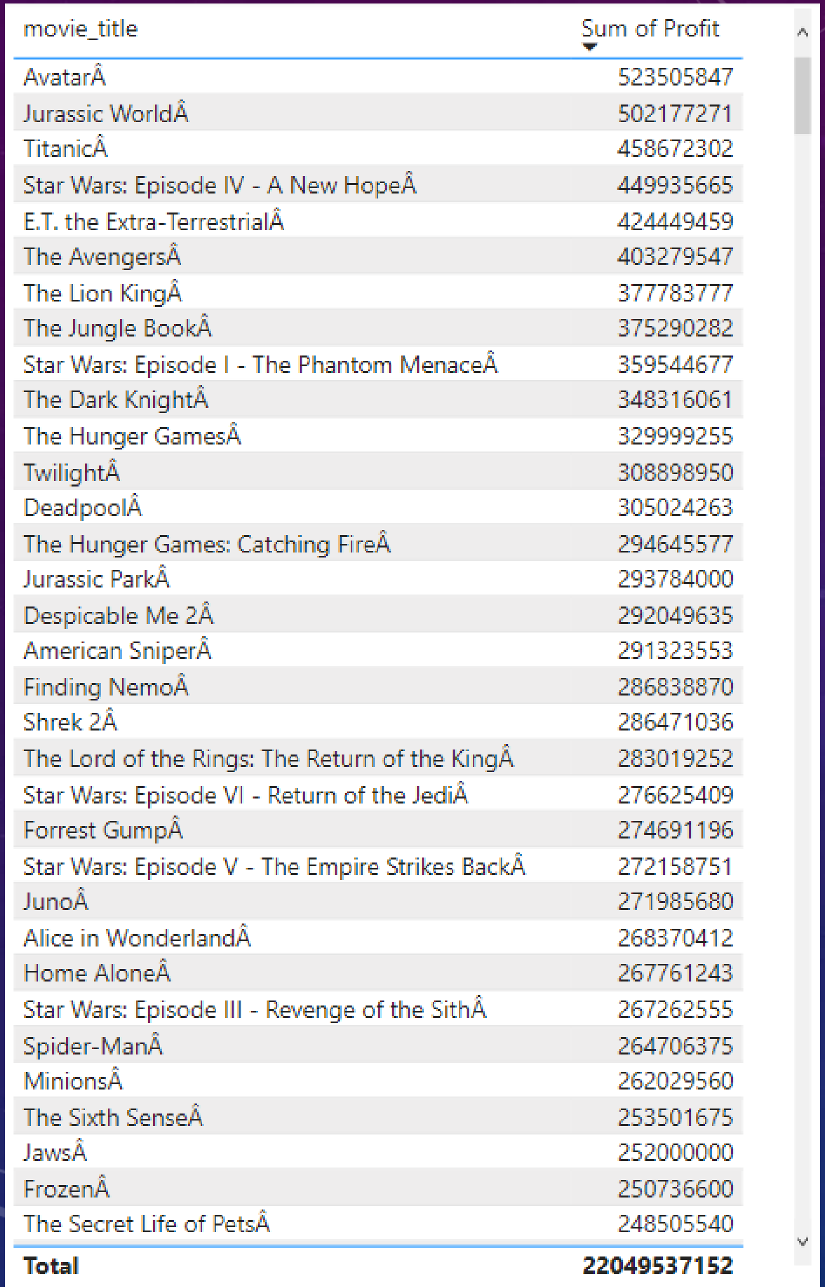
Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

Your task: Find the movies with the highest profit?

Ans: In this task we have to first create a new column to store the profit of the movies by taking the difference of the gross and budget



:To identify outliers in the data, I plotted a chart and looked for any unusually high or low values. One example of an outlier that I observed was a value of -1.2E+10.



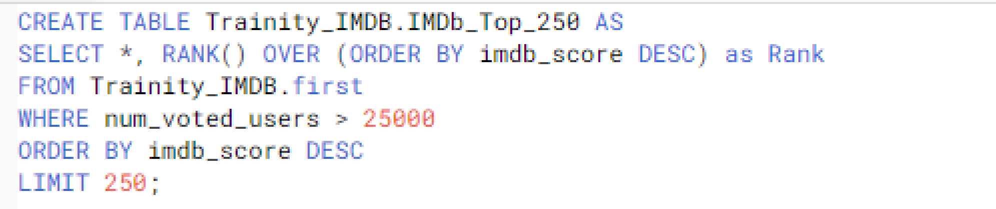
movie_title	Sum of Profit
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The AvengersÂ	403279547
The Lion KingÂ	377783777
The Jungle BookÂ	375290282
Star Wars: Episode I - The Phantom MenaceÂ	359544677
The Dark KnightÂ	348316061
The Hunger GamesÂ	329999255
TwilightÂ	308898950
DeadpoolÂ	305024263
The Hunger Games: Catching FireÂ	294645577
Jurassic ParkÂ	293784000
Despicable Me 2Â	292049635
American SniperÂ	291323553
Finding NemoÂ	286838870
Shrek 2Â	286471036
The Lord of the Rings: The Return of the KingÂ	283019252
Star Wars: Episode VI - Return of the JediÂ	276625409
Forrest GumpÂ	274691196
Star Wars: Episode V - The Empire Strikes BackÂ	272158751
JunoÂ	271985680
Alice in WonderlandÂ	268370412
Home AloneÂ	267761243
Star Wars: Episode III - Revenge of the SithÂ	267262555
Spider-ManÂ	264706375
MinionsÂ	262029560
The Sixth SenseÂ	253501675
JawsÂ	252000000
FrozenÂ	250736600
The Secret Life of PetsÂ	248505540
Total	**22049537152**

:I used a tool called Power BI to create this visualization showing that the movie with the highest profit was "Avatar".

C. 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.

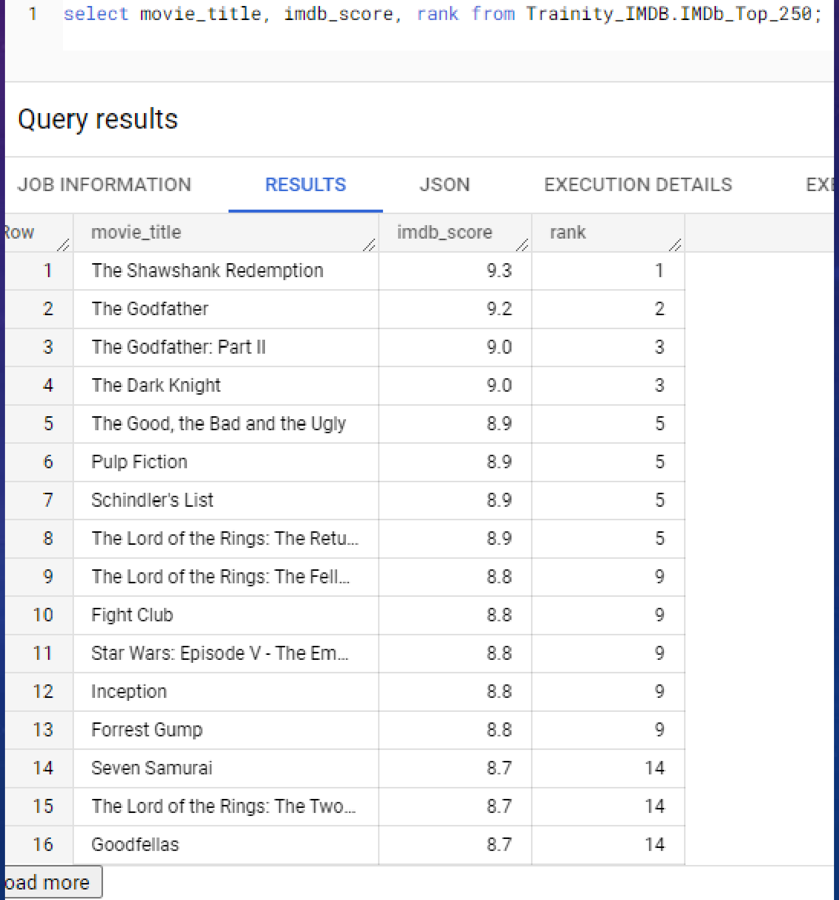
Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

Your task: Find IMDB Top 250



```sql
CREATE TABLE Trainity_IMDB.IMDb_Top_250 AS
SELECT *, RANK() OVER (ORDER BY imdb_score DESC) as Rank
FROM Trainity_IMDB.first
WHERE num_voted_users > 25000
ORDER BY imdb_score DESC
LIMIT 250;
```

:I used an SQL query to identify the top 250 movies with the highest IMDB scores and a minimum of 25,000 voted users. Here is the list:

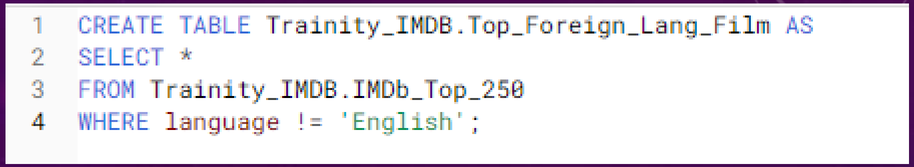


```sql
select movie_title, imdb_score, rank from Trainity_IMDB.IMDb_Top_250;
```

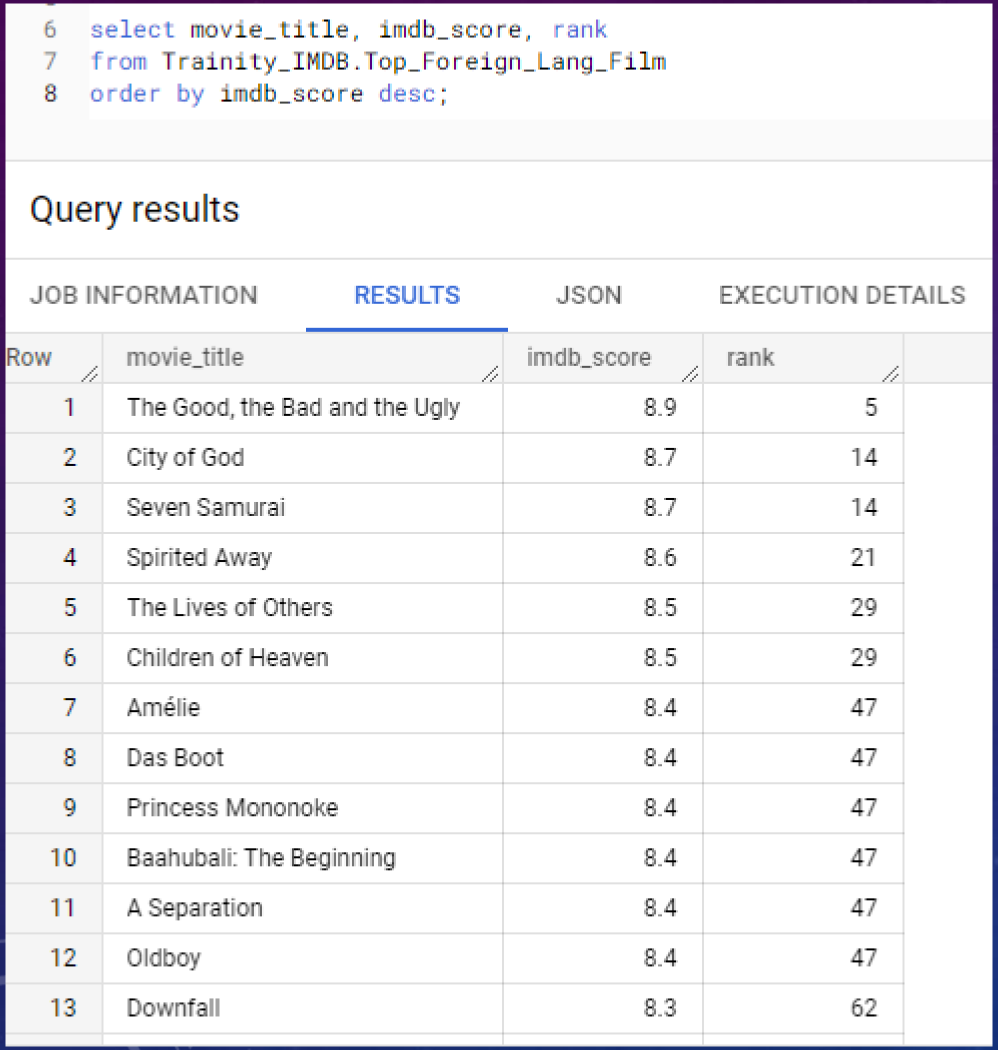
Query results

Row	movie_title	imdb_score	rank
1	The Shawshank Redemption	9.3	1
2	The Godfather	9.2	2
3	The Godfather: Part II	9.0	3
4	The Dark Knight	9.0	3
5	The Good, the Bad and the Ugly	8.9	5
6	Pulp Fiction	8.9	5
7	Schindler's List	8.9	5
8	The Lord of the Rings: The Retu…	8.9	5
9	The Lord of the Rings: The Fell…	8.8	9
10	Fight Club	8.8	9
11	Star Wars: Episode V - The Em…	8.8	9
12	Inception	8.8	9
13	Forrest Gump	8.8	9
14	Seven Samurai	8.7	14
15	The Lord of the Rings: The Two…	8.7	14
16	Goodfellas	8.7	14

..250 rows



```
CREATE TABLE Trainity_IMDB.Top_Foreign_Lang_Film AS
SELECT *
FROM Trainity_IMDB.IMDb_Top_250
WHERE language != 'English';
```



```
select movie_title, imdb_score, rank
from Trainity_IMDB.Top_Foreign_Lang_Film
order by imdb_score desc;
```

Query results

Row	movie_title	imdb_score	rank
1	The Good, the Bad and the Ugly	8.9	5
2	City of God	8.7	14
3	Seven Samurai	8.7	14
4	Spirited Away	8.6	21
5	The Lives of Others	8.5	29
6	Children of Heaven	8.5	29
7	Amélie	8.4	47
8	Das Boot	8.4	47
9	Princess Mononoke	8.4	47
10	Baahubali: The Beginning	8.4	47
11	A Separation	8.4	47
12	Oldboy	8.4	47
13	Downfall	8.3	62

..37 rows

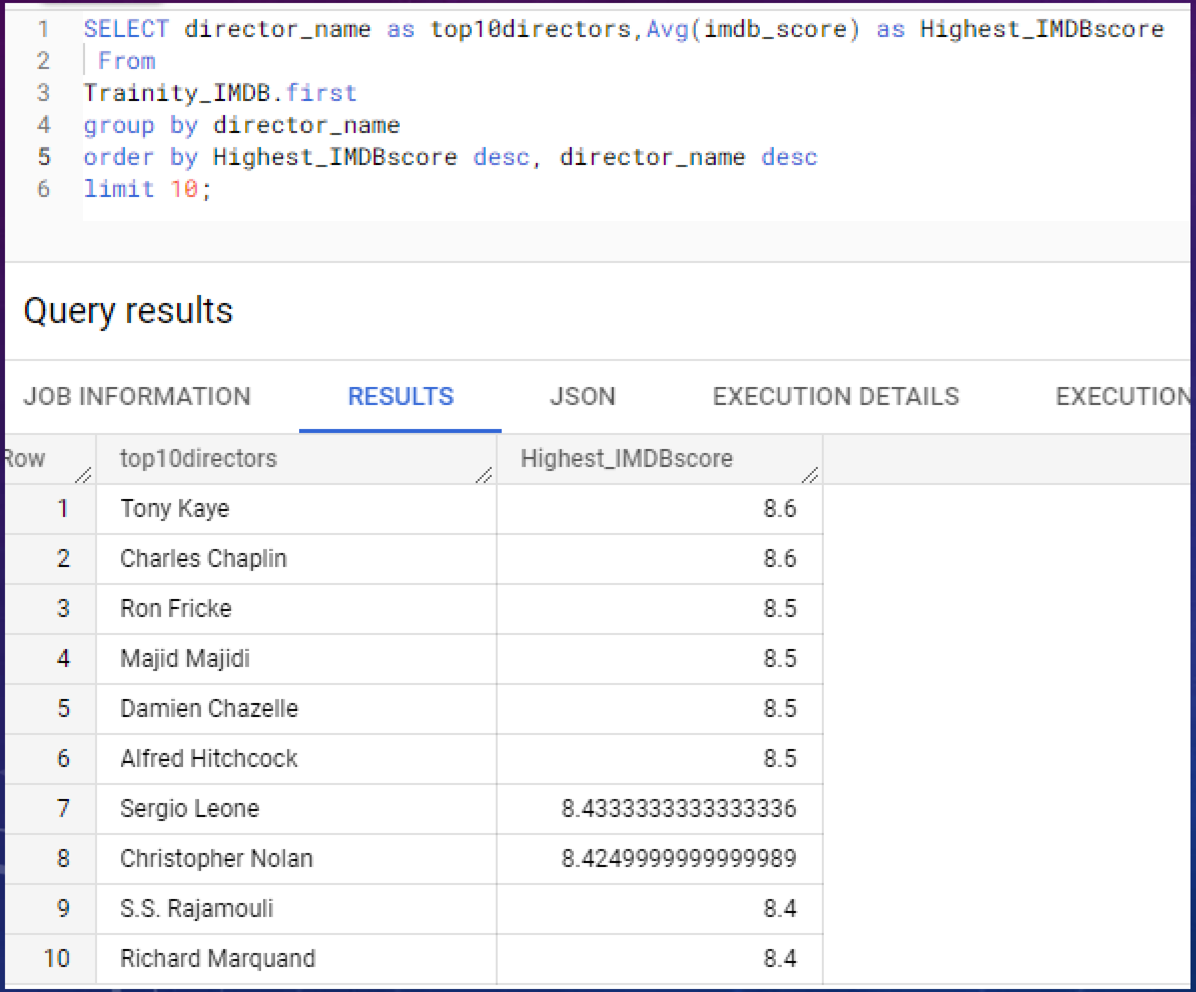
:I used an SQL query to create a table of top foreign language films from the top 250 IMDB movies. The films in this table are those whose language is not English.

:From the top 250 IMDB movies, we can conclude that only 37 of them are not in the English language. This suggests that English is a more preferable language for these films.

D. Best Directors: Group the column using the director_name column.

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.

Your task: Find the best directors

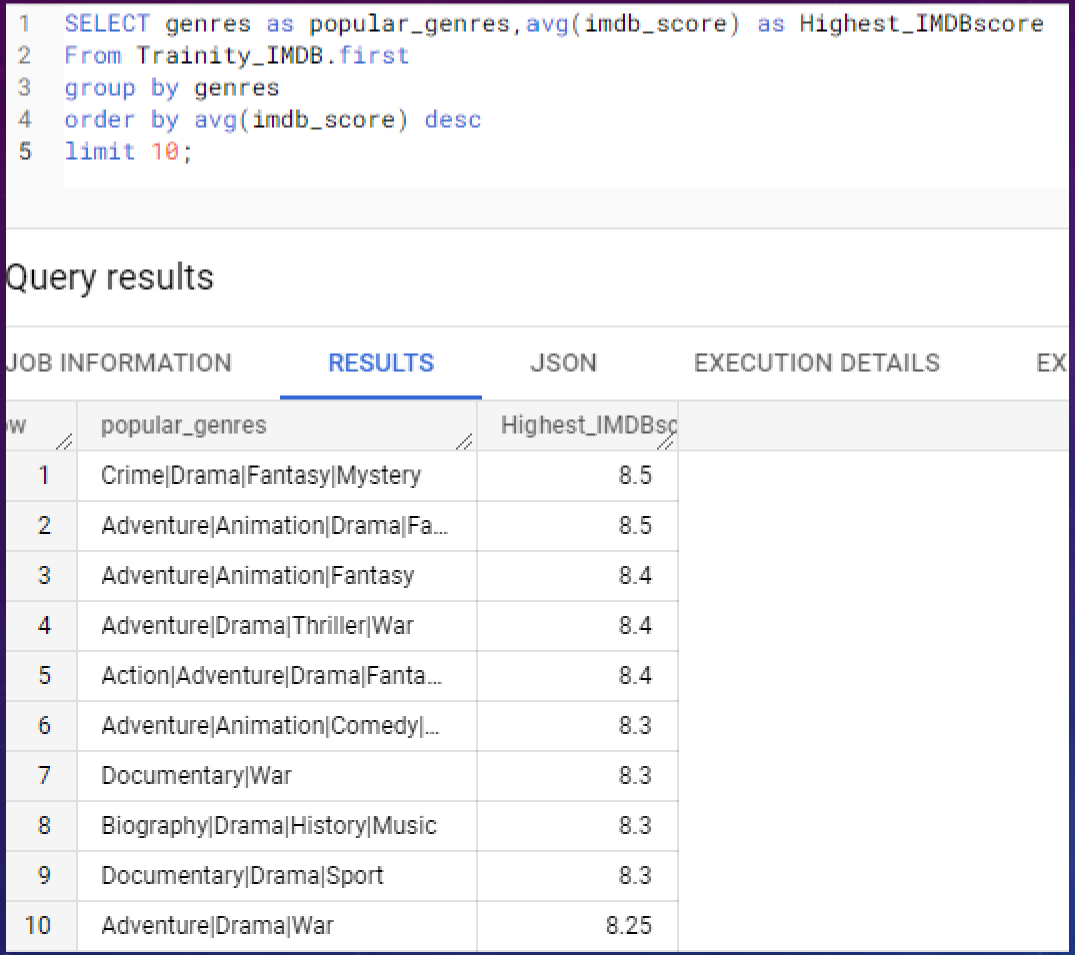


```
SELECT director_name as top10directors,Avg(imdb_score) as Highest_IMDBscore
 From
Trainity_IMDB.first
group by director_name
order by Highest_IMDBscore desc, director_name desc
limit 10;
```

Row	top10directors	Highest_IMDBscore
1	Tony Kaye	8.6
2	Charles Chaplin	8.6
3	Ron Fricke	8.5
4	Majid Majidi	8.5
5	Damien Chazelle	8.5
6	Alfred Hitchcock	8.5
7	Sergio Leone	8.4333333333333336
8	Christopher Nolan	8.4249999999999989
9	S.S. Rajamouli	8.4
10	Richard Marquand	8.4

Based on the data provided, it appears that Tony Kaye and Charles Chaplin are the best director, with an average IMDB score of 8.6 for his movies.

E. Popular Genres: Perform this step using the knowledge gained while performing previous steps. Your task: Find popular genres



```
SELECT genres as popular_genres,avg(imdb_score) as Highest_IMDBscore
From Trainity_IMDB.first
group by genres
order by avg(imdb_score) desc
limit 10;
```

Query results

w	popular_genres	Highest_IMDBsc
1	Crime|Drama|Fantasy|Mystery	8.5
2	Adventure|Animation|Drama|Fa...	8.5
3	Adventure|Animation|Fantasy	8.4
4	Adventure|Drama|Thriller|War	8.4
5	Action|Adventure|Drama|Fanta...	8.4
6	Adventure|Animation|Comedy|...	8.3
7	Documentary|War	8.3
8	Biography|Drama|History|Music	8.3
9	Documentary|Drama|Sport	8.3
10	Adventure|Drama|War	8.25

Based on the data provided, it appears that the Crime|Drama|Fantasy|Mystery genre has the highest average IMDB score, indicating that it is a more preferable genre.

F. Charts: 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.

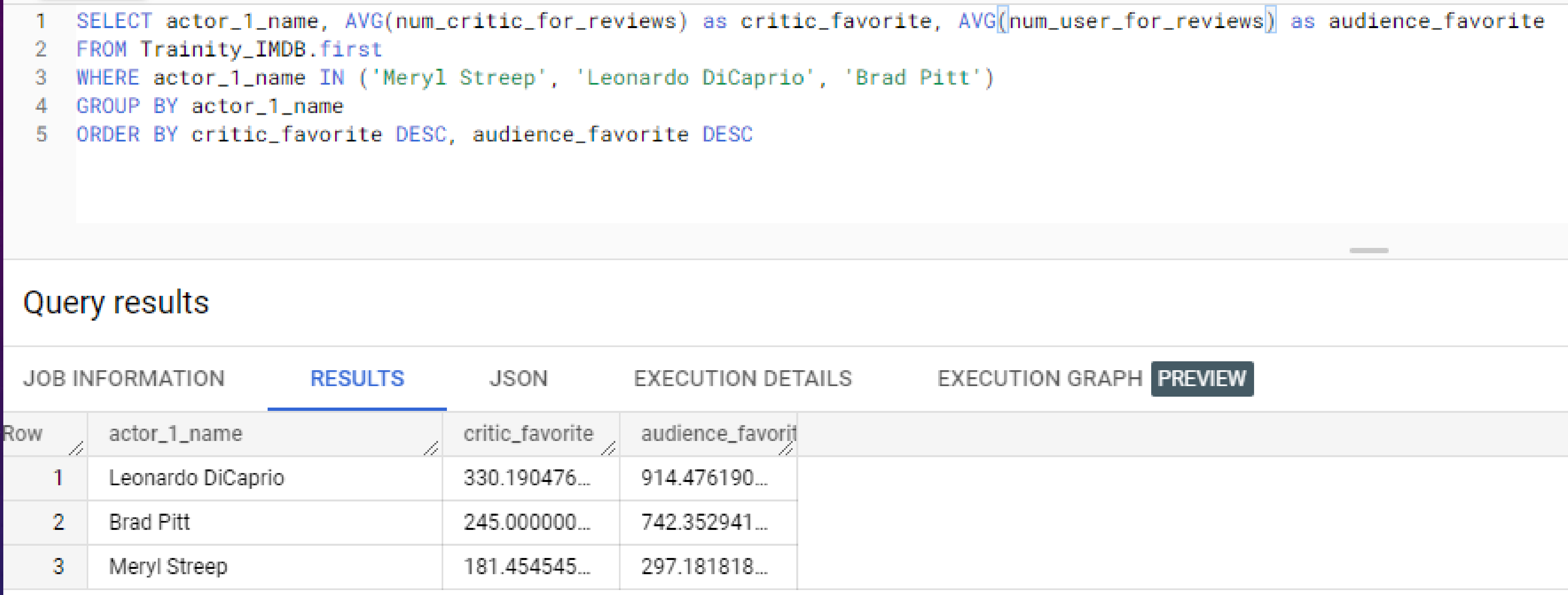
Append the rows of all these columns and store them in a new column named Combined.

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.

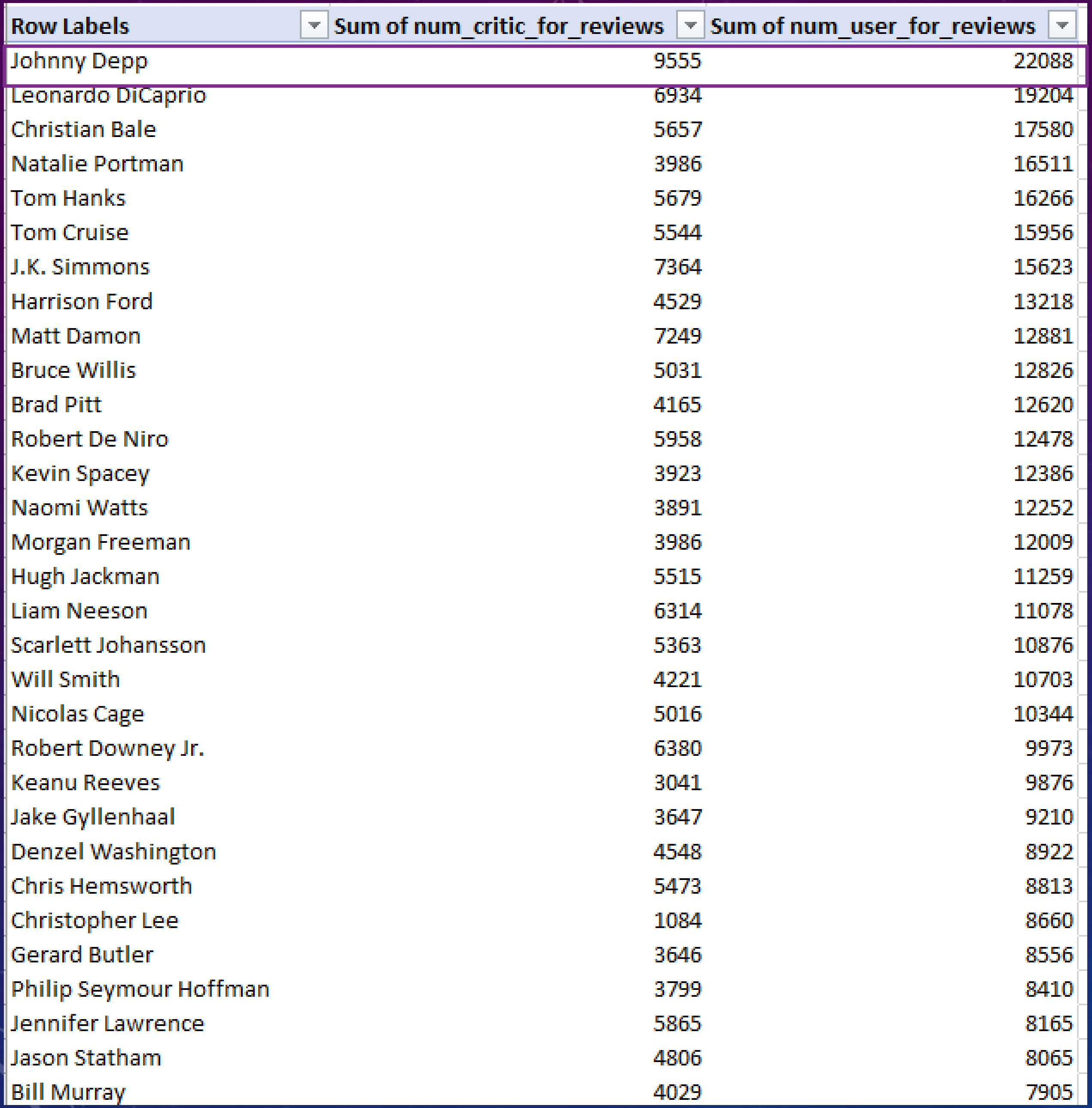
Your task: Find the critic-favorite and audience-favorite actors



```sql
SELECT actor_1_name, AVG(num_critic_for_reviews) as critic_favorite, AVG(num_user_for_reviews) as audience_favorite
FROM Trainity_IMDB.first
WHERE actor_1_name IN ('Meryl Streep', 'Leonardo DiCaprio', 'Brad Pitt')
GROUP BY actor_1_name
ORDER BY critic_favorite DESC, audience_favorite DESC
```

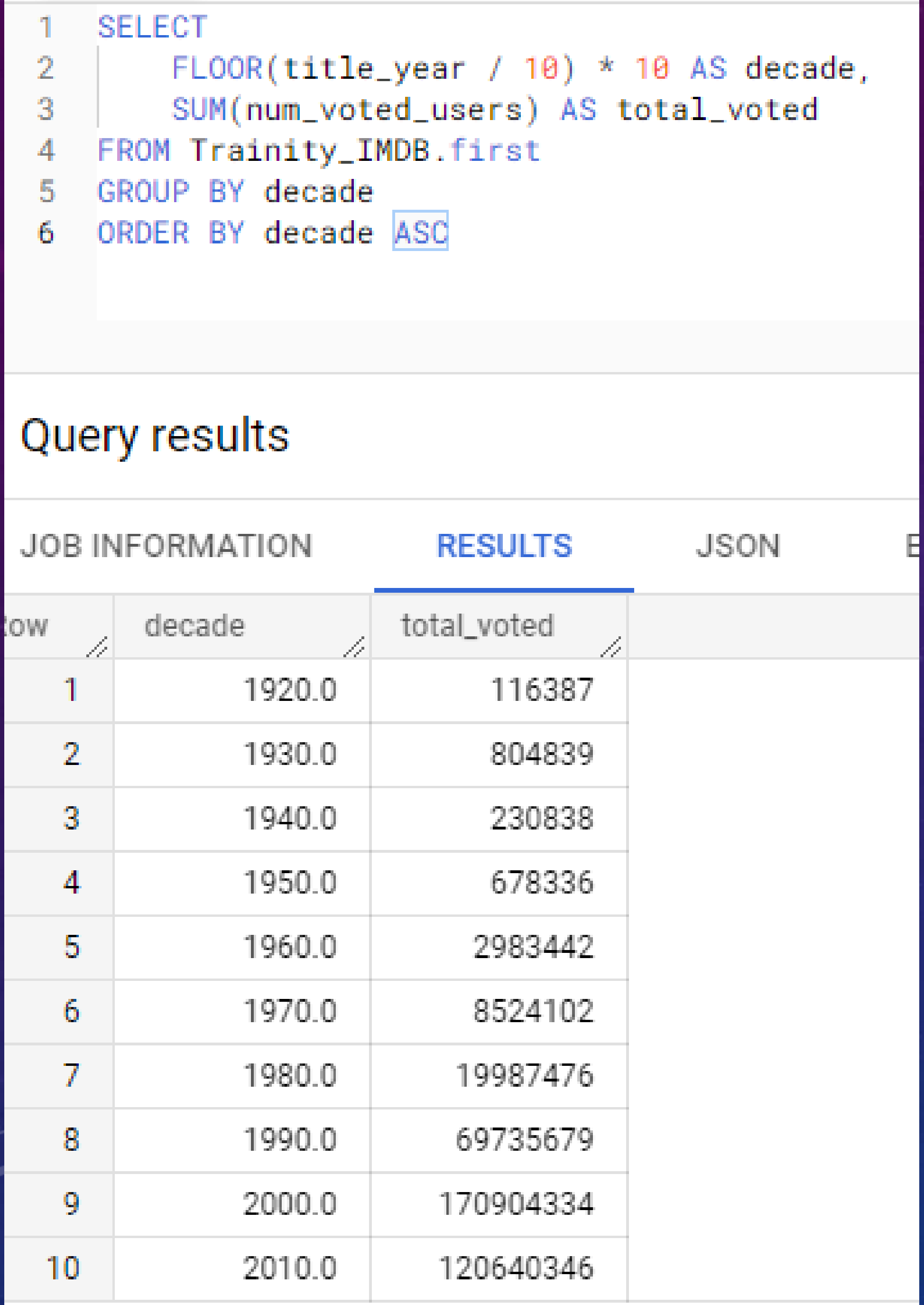
Row	actor_1_name	critic_favorite	audience_favorit
1	Leonardo DiCaprio	330.190476...	914.476190...
2	Brad Pitt	245.000000...	742.352941...
3	Meryl Streep	181.454545...	297.181818...

Based on the data provided, it appears that Leonardo DiCaprio is the audience favorite and critic favorite actor.



Row Labels	Sum of num_critic_for_reviews	Sum of num_user_for_reviews
Johnny Depp	9555	22088
Leonardo DiCaprio	6934	19204
Christian Bale	5657	17580
Natalie Portman	3986	16511
Tom Hanks	5679	16266
Tom Cruise	5544	15956
J.K. Simmons	7364	15623
Harrison Ford	4529	13218
Matt Damon	7249	12881
Bruce Willis	5031	12826
Brad Pitt	4165	12620
Robert De Niro	5958	12478
Kevin Spacey	3923	12386
Naomi Watts	3891	12252
Morgan Freeman	3986	12009
Hugh Jackman	5515	11259
Liam Neeson	6314	11078
Scarlett Johansson	5363	10876
Will Smith	4221	10703
Nicolas Cage	5016	10344
Robert Downey Jr.	6380	9973
Keanu Reeves	3041	9876
Jake Gyllenhaal	3647	9210
Denzel Washington	4548	8922
Chris Hemsworth	5473	8813
Christopher Lee	1084	8660
Gerard Butler	3646	8556
Philip Seymour Hoffman	3799	8410
Jennifer Lawrence	5865	8165
Jason Statham	4806	8065
Bill Murray	4029	7905

Johnny Depp is a highly popular actor among both critics and audiences.



```
SELECT
    FLOOR(title_year / 10) * 10 AS decade,
    SUM(num_voted_users) AS total_voted
FROM Trainity_IMDB.first
GROUP BY decade
ORDER BY decade ASC
```

Query results

ow	decade	total_voted
1	1920.0	116387
2	1930.0	804839
3	1940.0	230838
4	1950.0	678336
5	1960.0	2983442
6	1970.0	8524102
7	1980.0	19987476
8	1990.0	69735679
9	2000.0	170904334
10	2010.0	120640346

During the 2000s, there was a high number of users who voted for movies.

KEY INSIGHTS

- It appears that movies like "Avatar" and "Jurassic Park" have the potential to earn high profits. If the goal is to maximize profit, it may be advisable to consider making movies with similar themes or characteristics.
- It appears that the movie "The Shawshank Redemption" has the highest IMDB score among those with a minimum of 25,000 voted users.
- From the top 250 IMDB movies, we can conclude that only 37 of them are not in the English language. This suggests that English is a more preferable language for these films.
- Consider working with Tony Kaye or Charles Chaplin as a director on future projects, as their past work has received high ratings from audiences and critics.
- It appears that the Crime|Drama|Fantasy|Mystery genre has the highest average IMDB score, indicating that it is a more preferable genre.
- It appears that Johnny Depp is the audience favorite and critic favorite actor.

Result:

During this project, I discovered that a range of factors contribute to the success of a movie. I also learned how to utilize various tools, such as SQL, Excel, and Power BI, to analyze and understand data. By using these tools together, I gained a more comprehensive understanding of what makes a movie successful. This project helped me to see the importance of considering multiple variables and viewpoints when analyzing data.

Completing this project allowed me to improve my skills in crafting and executing queries, as well as giving me a glimpse into the tasks and responsibilities of a data analyst in a professional setting.

THANK YOU