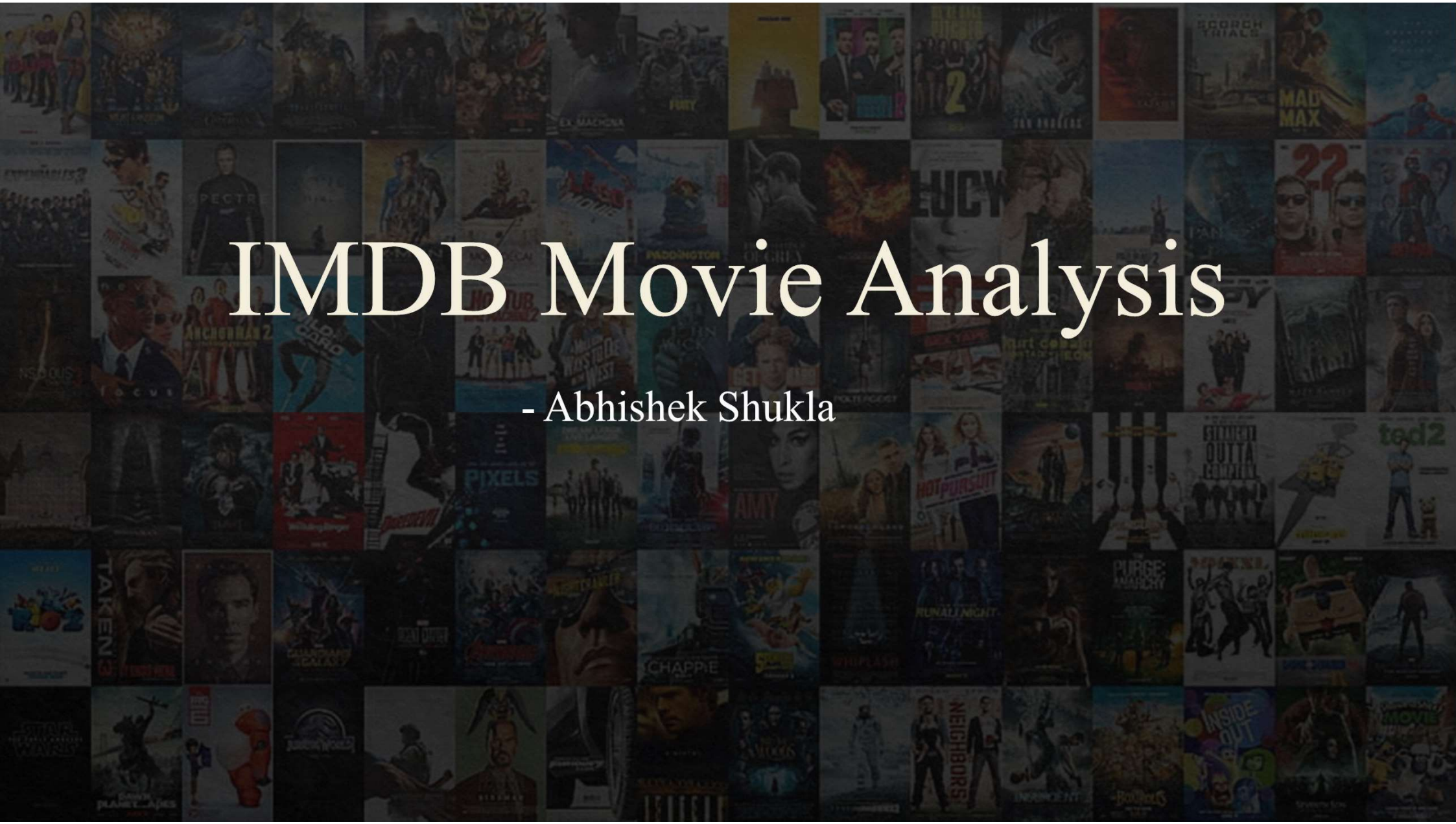

IMDB Movie Analysis
- Abhishek Shukla

Project Description

The IMDB Movie Analysis Project is based on the world wide realised movie records, it is known also as an analytical website on this website we collect information about movies from real world and Analyze them on the basis of users reviews, likes and other a lot of parameter so the project begins with some csv dataset which is totally row in this project we are going to clean the data set and Analyze the records given task by stakeholders to provide some meaningful insights.

Approach of the Analysis

- ✓ First Clean the dataset to start the process of Analysis.
- ✓ Review the dataset and understand the different columns and their meanings.
- ✓ Use the provided guiding questions to frame the problem you want to shed light on.
- ✓ Compile the findings from the analysis into a detailed report.
- ✓ Create Visuals to better understanding of the stakeholders.
- ✓ Present the data story in a clear and concise manner, making it easy for stakeholders to understand.

Tech-Stack Used In The Analysis

- ✓ Excel 365 for complete analysis and visualization.
- ✓ Power Point Presentation 365 is used to create reports with figures.

Task A : Cleaning the data

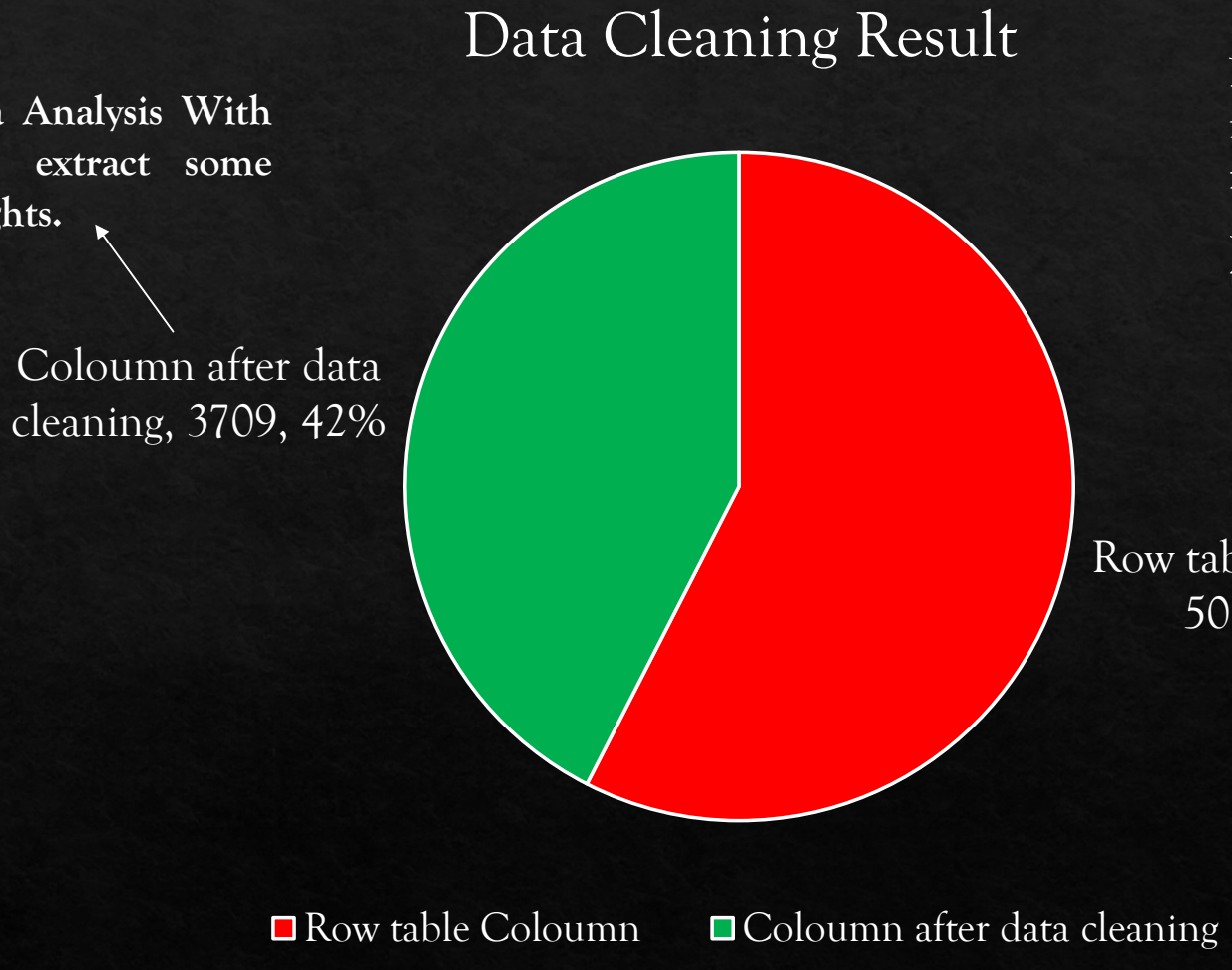
Cleaning the data :: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Your task: Clean the data

Task A - Data Cleaning

Now Starts data Analysis With Clean Data to extract some meaningful insights.

In row dataset I found lots of null values and duplicates after removing this I get 3709 out of 5025 records, which will serve as the starting point for analysis.



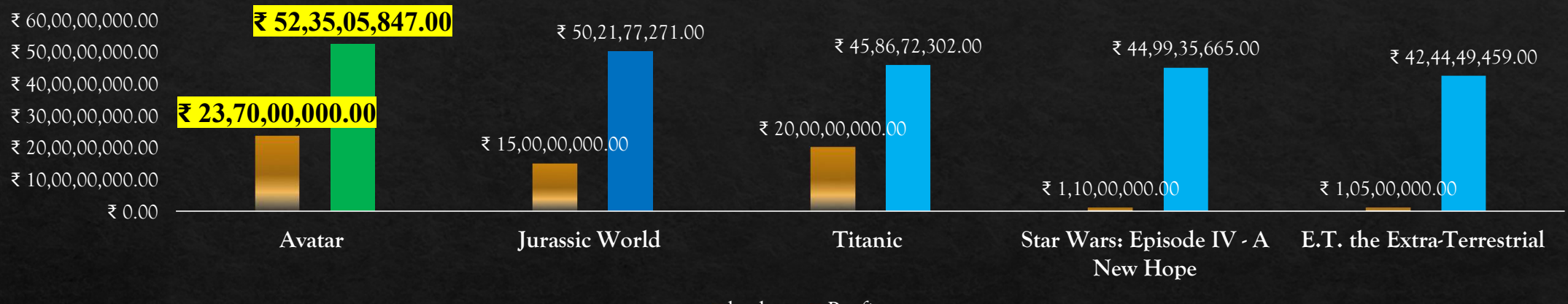
Task B : Movies with highest profit

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?

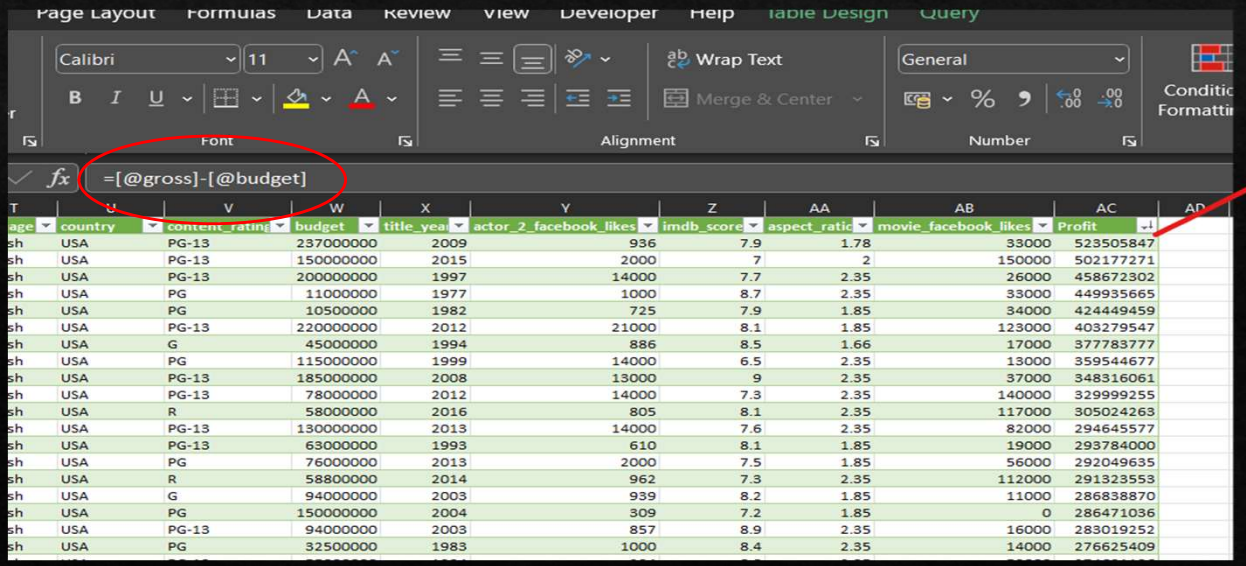
Task B - Find the movies with the highest profit?

TOP 5 MOST PROFITABLE MOVIES



Insights :

These are the top 5 profitable movies in given database and the most profitable movie is “Avatar”.



country	content_rating	budget	title_year	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes	Profit
USA	PG-13	237000000	2009	936	7.9	1.78	33000	523505847
USA	PG-13	150000000	2015	2000	7	2	150000	502177271
USA	PG-13	200000000	1997	14000	7.7	2.35	26000	458672302
USA	PG	11000000	1977	1000	8.7	2.35	33000	449935665
USA	PG	10500000	1982	725	7.9	1.85	34000	424449459
USA	PG-13	220000000	2012	21000	8.1	1.85	123000	403279547
USA	G	45000000	1994	886	8.5	1.66	17000	377783777
USA	PG	115000000	1999	14000	6.5	2.35	13000	359544677
USA	PG-13	185000000	2008	13000	9	2.35	37000	348316061
USA	PG-13	78000000	2012	14000	7.3	2.35	140000	329999255
USA	R	58000000	2016	805	8.1	2.35	117000	305024263
USA	PG-13	130000000	2013	14000	7.6	2.35	82000	294645577
USA	PG-13	63000000	1993	610	8.1	1.85	19000	293784000
USA	PG	76000000	2013	2000	7.5	1.85	56000	292049635
USA	R	58800000	2014	962	7.3	2.35	112000	291323553
USA	G	94000000	2003	939	8.2	1.85	11000	286838870
USA	PG	150000000	2004	309	7.2	1.85	0	286471036
USA	PG-13	94000000	2003	857	8.9	2.35	16000	283019252
USA	PG	32500000	1983	1000	8.4	2.35	14000	276625409

There is Formula How To Calculate Profit.

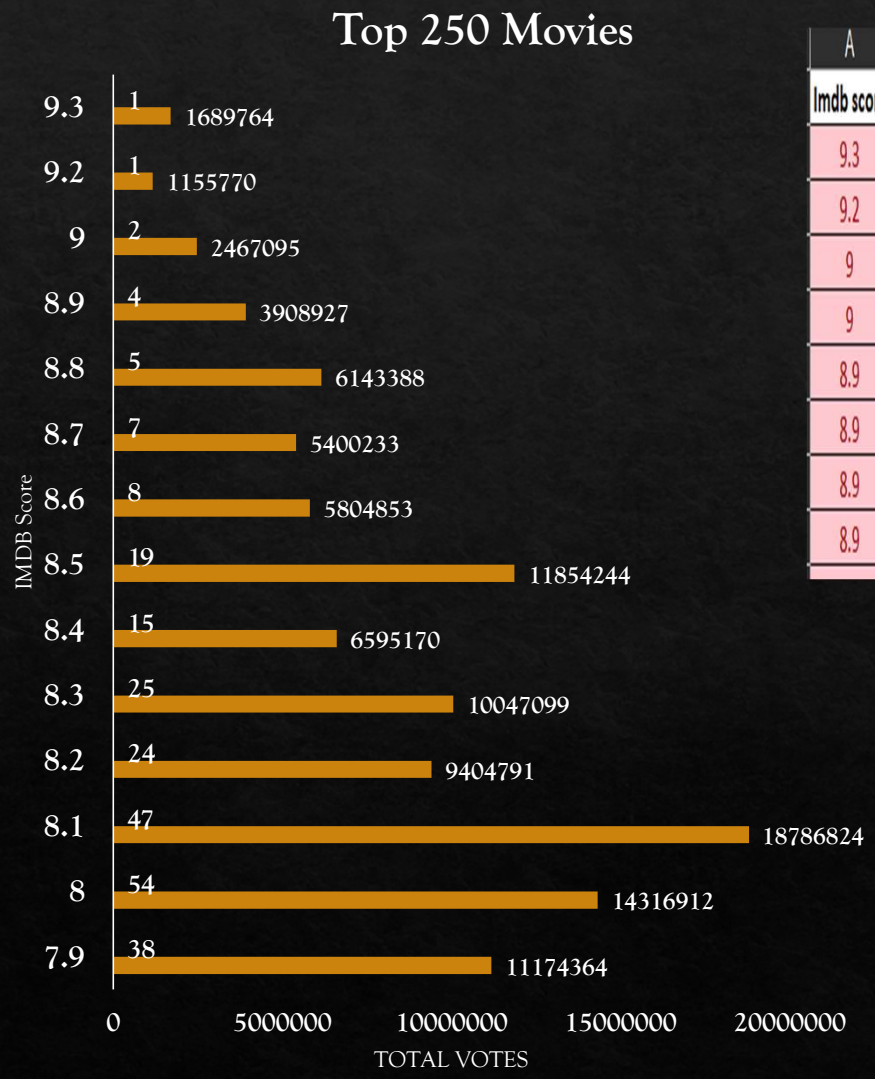
Task C : Top 250

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!

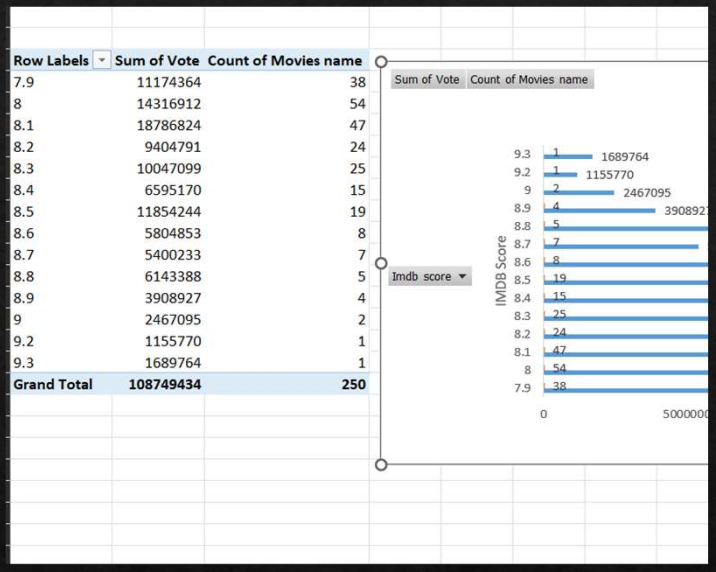
Your task: Find IMDB Top 250

Task C - Find IMDB Top 250



A	B	C	D	E
Imdb score	Top 250	Vote	Rank	
9.3	The Shawshank Redemption	1689764	1	
9.2	The Godfather	1155770	2	
9	The Dark Knight	1676169	3	
9	The Godfather: Part II	790926	4	
8.9	Pulp Fiction	1324680	5	
8.9	The Lord of the Rings: The Return of the King	1215718	6	
8.9	Schindler's List	865020	7	
8.9	The Good, the Bad and the Ugly	503509	8	

Created top 250 Movies Column in



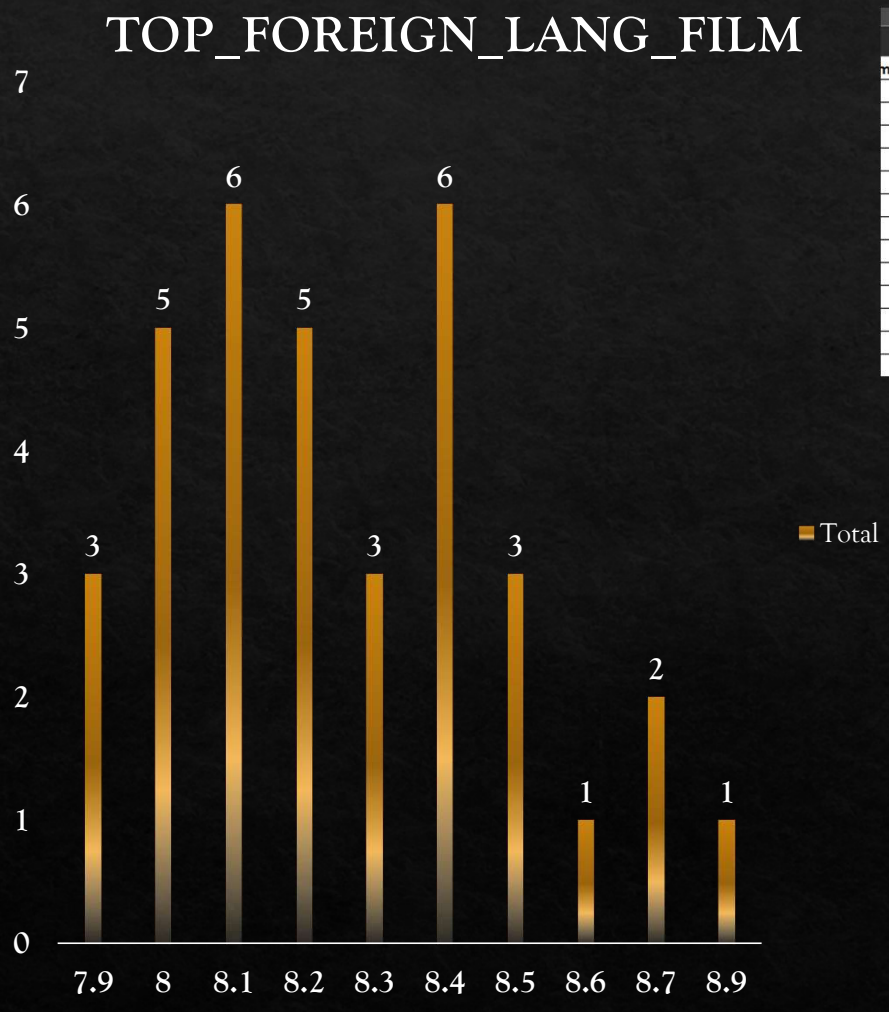
Row Labels	Sum of Vote	Count of Movies name
7.9	11174364	38
8	14316912	54
8.1	18786824	47
8.2	9404791	24
8.3	10047099	25
8.4	6595170	15
8.5	11854244	19
8.6	5804853	8
8.7	5400233	7
8.8	6143388	5
8.9	3908927	4
9	2467095	2
9.2	1155770	1
9.3	1689764	1
Grand Total	108749434	250

Use these column to create pivot chart and then create bar chart to visualize the data.

Insights :

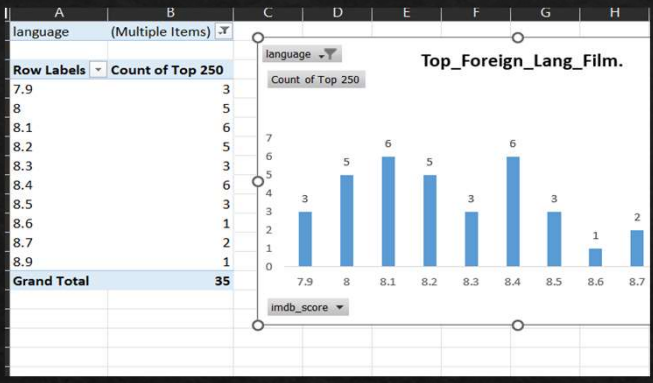
There is top 250 movies according to obtained IMDB Score and sum of user vote I found in this analysis that maximum number (above 20 movies) of movies are in bucket of 7.9 IMDB Score to 8.3.

Task C - Top_Foreign_Lang_Film Find into IMDB Top 250



Z	AA	AB	AC	
mdb_score	aspect_ratio	movie_facebook_likes	Profit	Top 250
9.3	1.85	108000	3341469	The Shawshank Redemp
9.2	1.85	43000	128821952	The Godfather
9	2.35	37000	348316061	The Dark Knight
9	1.85	14000	44300000	The Godfather: Part II
8.9	2.35	45000	99930000	Pulp Fiction
8.9	2.35	16000	283019252	The Lord of the Rings: T
8.9	1.85	41000	74067179	Schindler's List
8.9	2.35	20000	4900000	The Good, the Bad and t
8.8	2.35	175000	132568851	Inception
8.8	2.35	48000	-25976605	Fight Club
8.8	2.35	59000	274691196	Forrest Gump
8.8	2.35	21000	220837577	The Lord of the Rings: T
8.8	2.35	17000	272158751	Star Wars: Episode V - T

Created top 250 Movies Column in



Use these column to create pivot chart, in the pivot chart I put filter on language column and then create Stacked chart to visualize the data.

Insights :

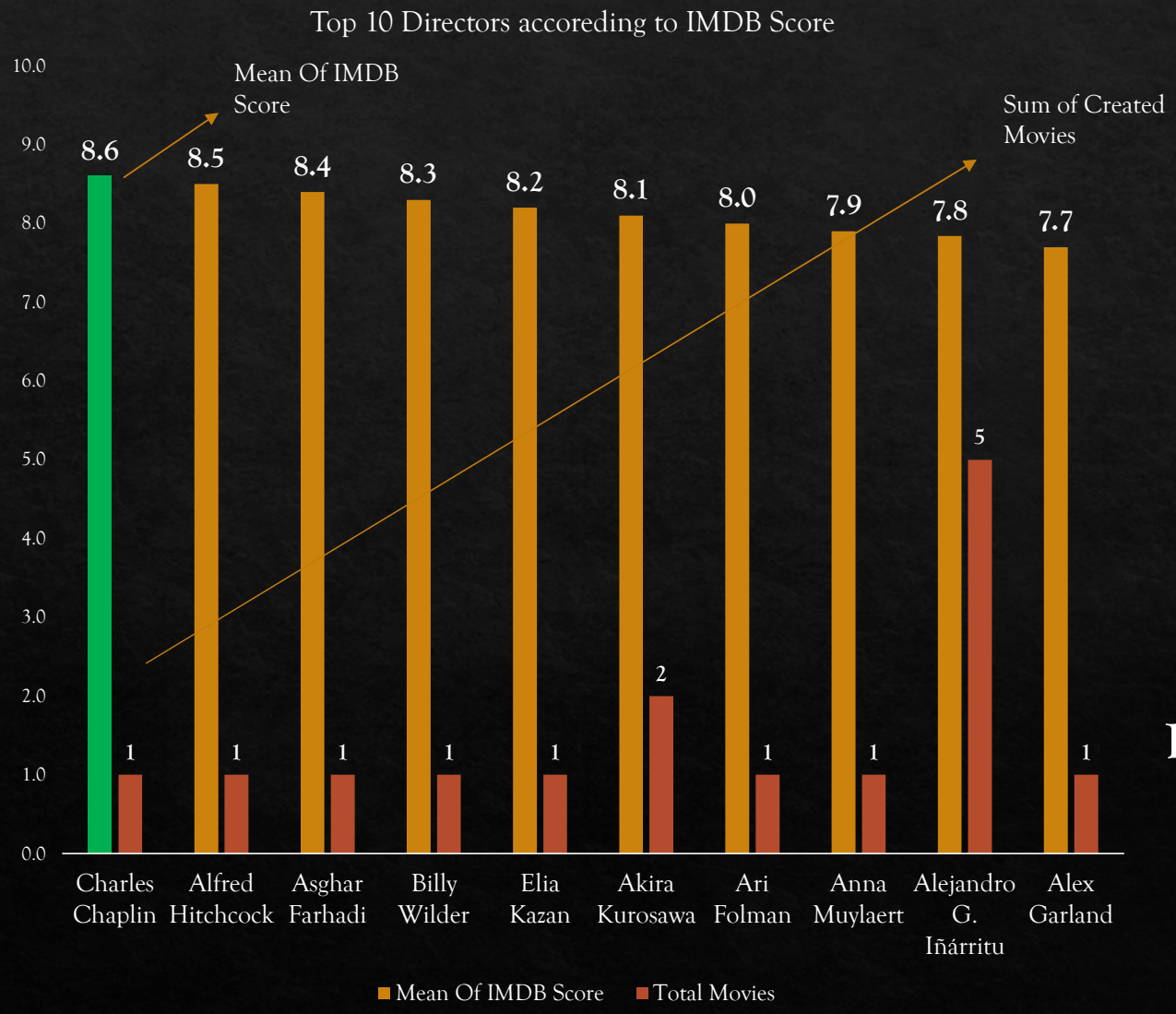
There is top 250 movies according to obtained IMDB Score and sum of user vote I found in this analysis that the total **35** foreign language movies in this list and maximum number (above 5 movies) of movies belongs to 8.1 & 8.4 IMDB Score.

Task D : Best Directors

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

Task D - TOP 10 DIRECTORS



Director Name	Mean Of IMDB Score	Total Movies
Charles Chaplin	8.6	1
Alfred Hitchcock	8.5	1
Asghar Farhadi	8.4	1
Billy Wilder	8.3	1
Elia Kazan	8.2	1
Akira Kurosawa	8.1	2
Ari Folman	8.0	1
Anna Muylaert	7.9	1
Alejandro G. Iñárritu	7.8	5
Alex Garland	7.7	1
Tony Kaye	8.6	1
Damien Chazelle	8.5	1
Majid Majidi	8.5	1
Ron Fricke	8.5	1
Sergio Leone	8.4	3
Christopher Nolan	8.4	8
Richard Marquand	8.4	1
S.S. Rajamouli	8.4	1

Apply pivot chart on given dataset, in the pivot chart I put directors name on rows and IMDB score on values as mean of IMDB Score and also put total movies on values the short IMDB Score on the basis of largest to smallest and then I did some conditional formatting for figure out top 10 directors and then create a clustered chart to visualize the data.

Insights :

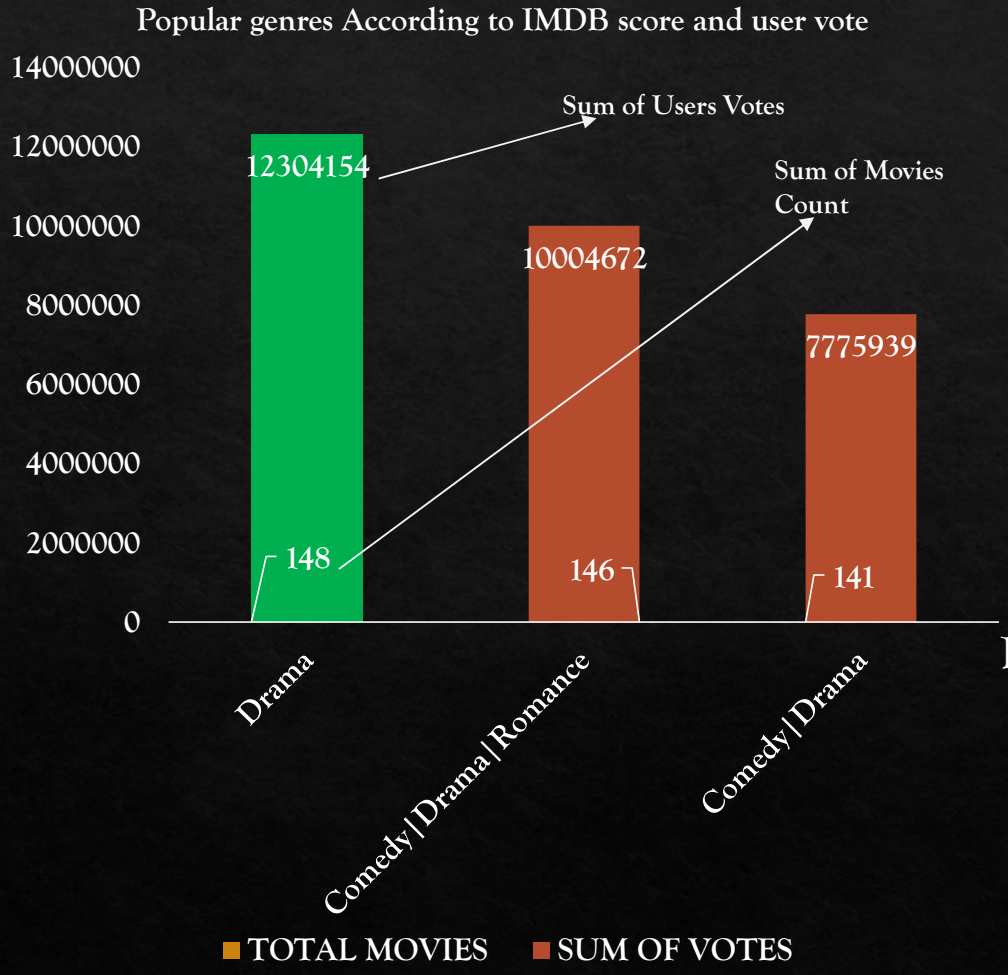
There is Top 10 Directors name along with the mean of IMDB Score and Total Created Movies Mr. Charles Chaplin obtained the maximum IMDB Score

Task E : Popular Genres

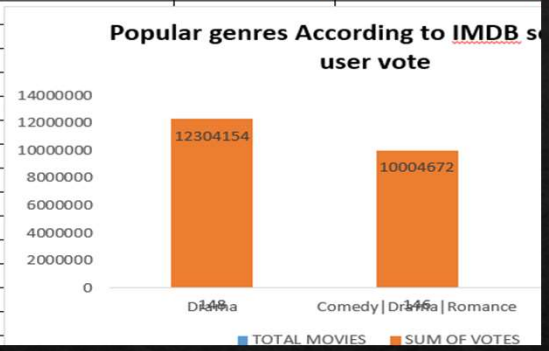
Popular Genres : Perform this step using the knowledge gained while performing previous steps.

Your task : Find popular genres

Task E - Find the Popular Genres



GENRES	TOTAL MOVIE	SUM OF VOTES
Drama	148	1230
Comedy|Drama|Romance	146	1000
Comedy|Drama	141	777
Comedy	137	1111
Comedy|Romance	130	991
Drama|Romance		
Crime|Drama|Thriller		
Action|Crime|Thriller		
Action|Crime|Drama|Thriller		
Action|Adventure|Sci-Fi		
Comedy|Crime		
Action|Adventure|Thriller		
Crime|Drama		
Horror		
Crime|Drama|Mystery|Thriller		
Drama|Thriller		
Action|Adventure|Sci-Fi|Thriller		
Horror|Mystery|Thriller		
Horror|Thriller		



Apply pivot chart on given dataset, in the pivot chart I put filter large to small on genres count as well as apply same on num_voted_users column and then create Stacked chart to visualize the data.

Insights :

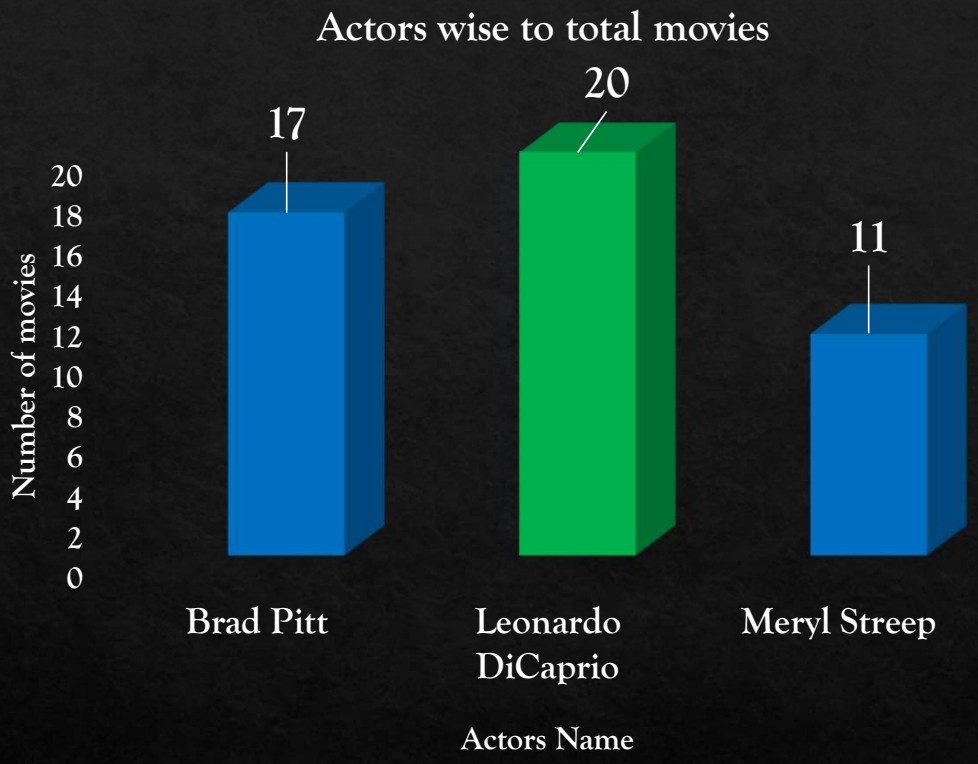
There is Top 3 Genres as it shown in chart the most popular Genres is Drama which is use 148 times as well as the most voted Genres by users is also Drama "The Sum of Users Vote is 12304154 which is the highest one."

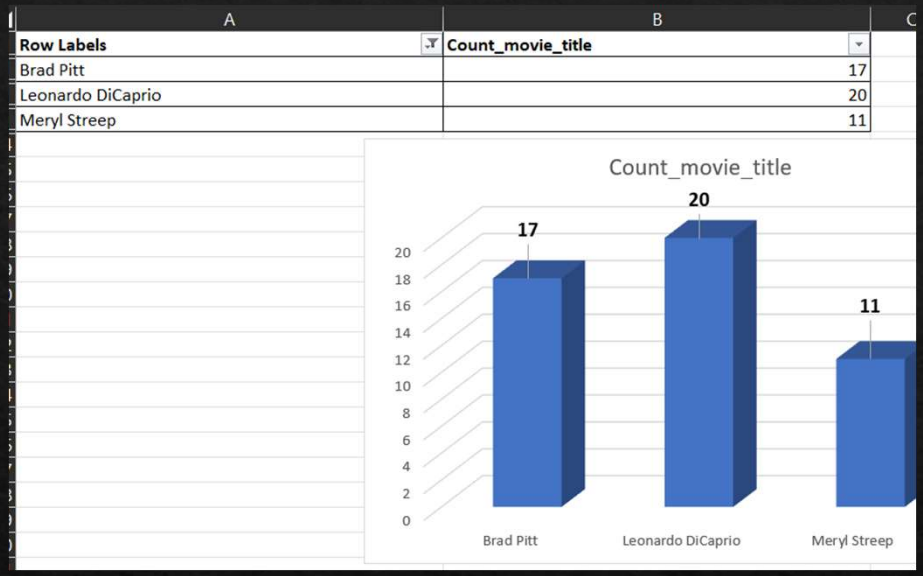
Task F : Charts

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

Task F - Actors wise to total movies





Row Labels	Count_movie_title
Brad Pitt	17
Leonardo DiCaprio	20
Meryl Streep	11

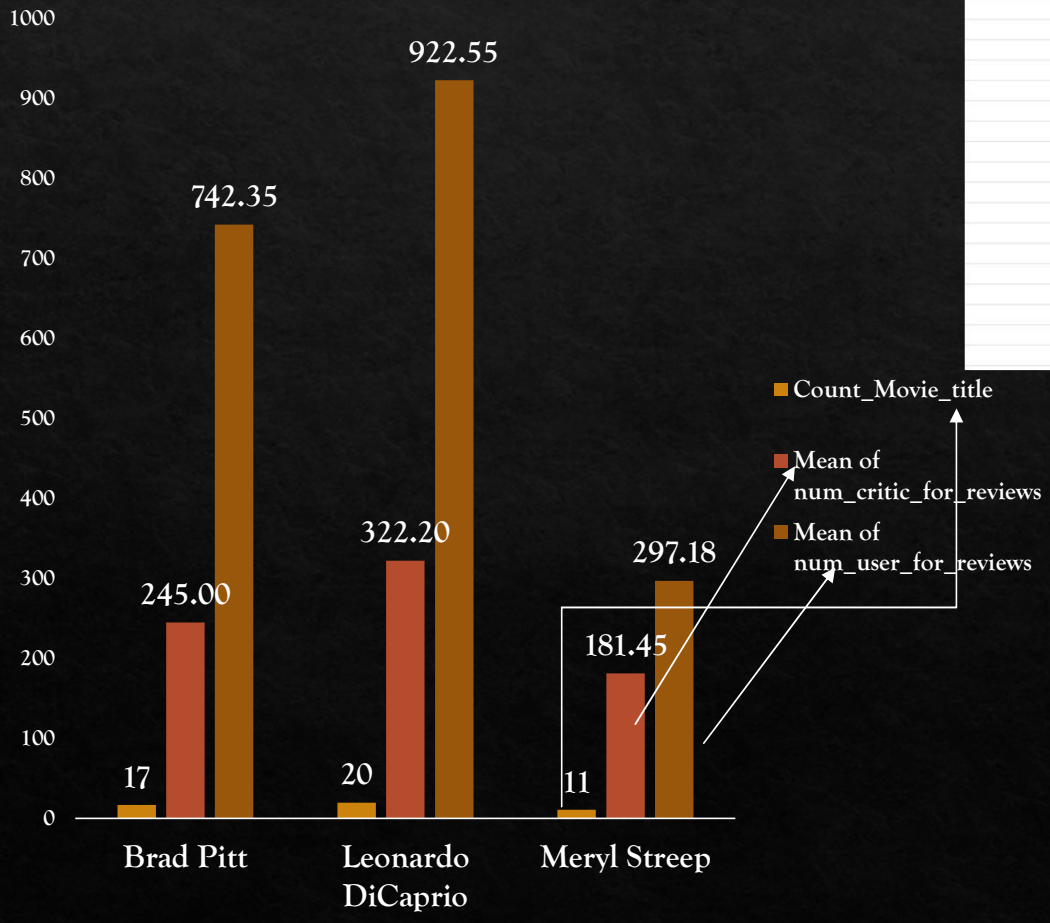
Apply pivot chart on given dataset, in the pivot chart I put filter large to small on count_movie_title as well as apply filter by given actors name on actor_1_name column and then create 3D Stacked chart to visualize the data.

Insights :

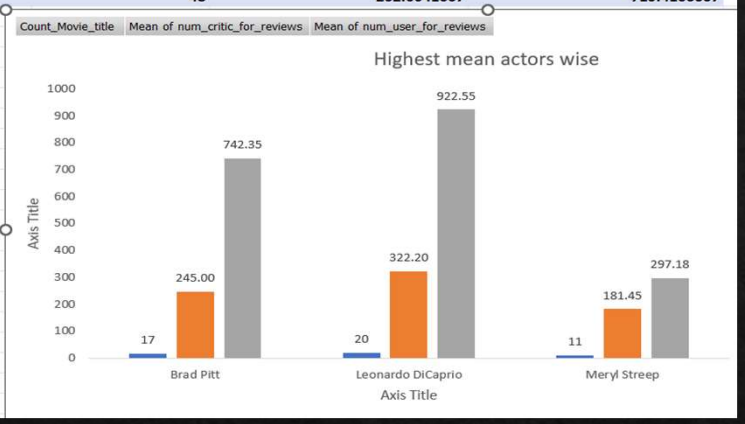
There is 03 Actors with his total movies as per the analysis the chart I found Leonardo DiCaprio worked as a lead actor in 20 movies as well as Bred Pitt 17 and Meryl Streep 11 Hence the max movies done tag goes to Leonardo DiCaprio.

Task F - Actors wise highest mean

Actors wise mean of the num_critic_for_reviews and num_users_for_review



Row Labels	Count_Movie_title	Mean of num_critic_for_reviews	Mean of num_user_for_reviews
Brad Pitt	17	245	742.3529412
Leonardo DiCaprio	20	322.2	922.55
Meryl Streep	11	181.4545455	297.1818182
Grand Total	**48**	**262.6041667**	**715.4166667**

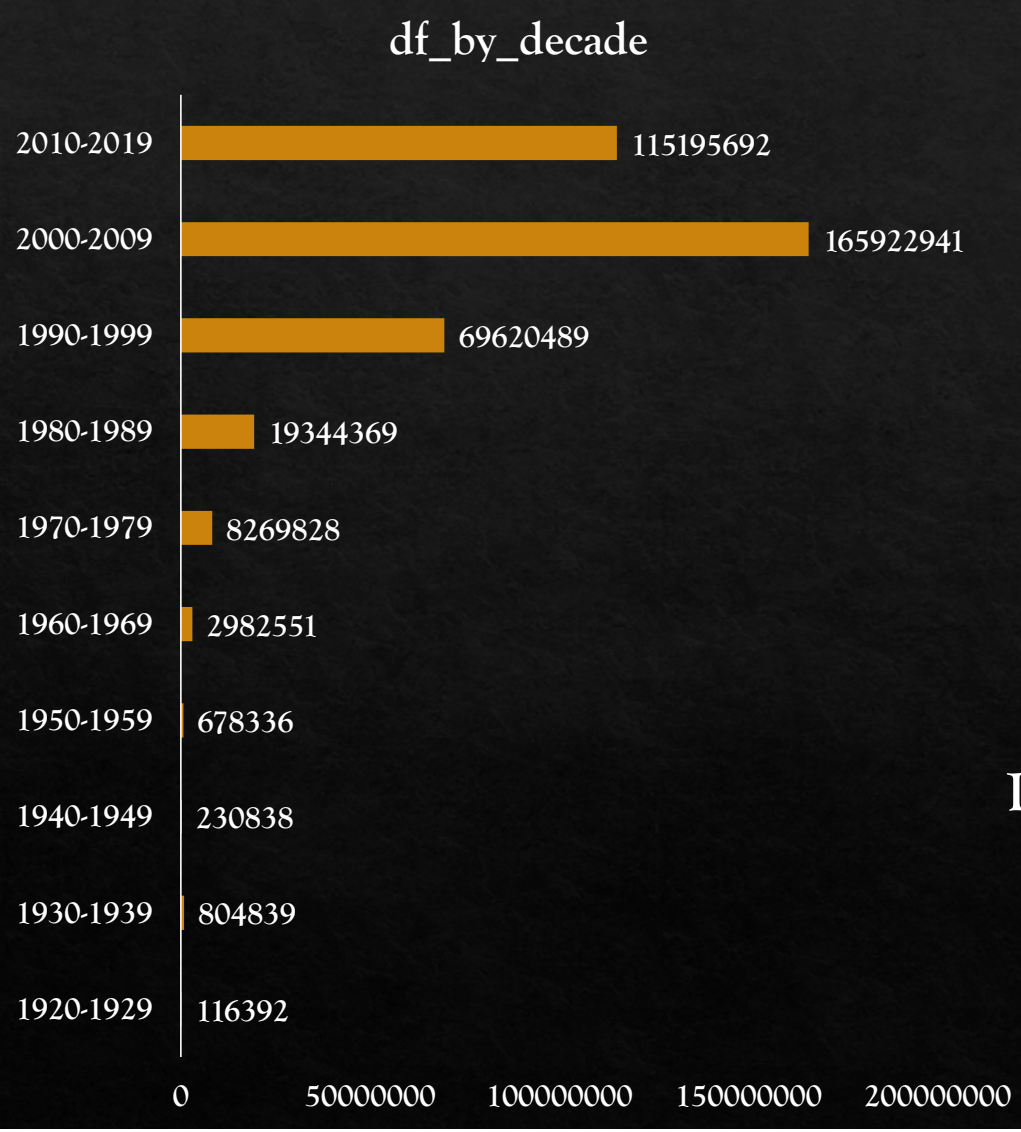


Apply pivot chart on extracted data, in the pivot chart I take mean of critic review and mean of user review in respective of actor name and then create clustered chart to visualize the data.

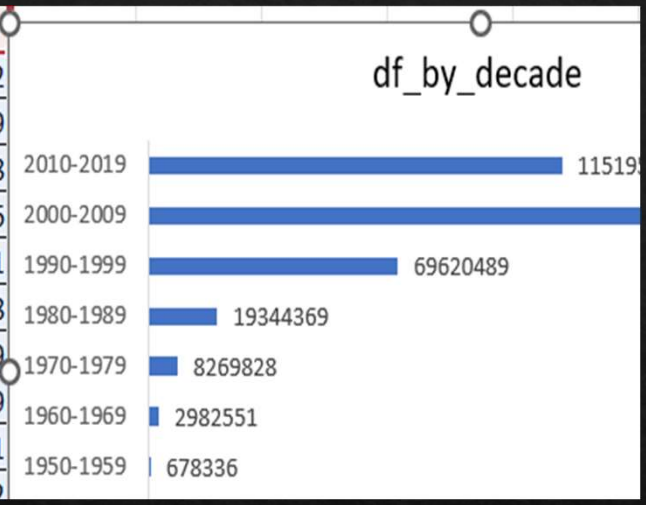
Insights :

There is 03 Actors with his total movies, the Mean of critic review and mean of user review as per the analysis the chart I found Leonardo DiCaprio worked as a lead actor in 20 movies and also he get roundoff 922 users review and roundoff 322 critic review as well as Bred Pitt 17 movies and also he get roundoff 742 users review and roundoff 245 critic review and Meryl Streep 11 movies and also Meryl Streep get roundoff 297 users review and roundoff 181 critic review Hence the most favourite actor tag goes to Leonardo DiCaprio.

Task F - number of voted users over decades using a bar chart



Decade	Sum of num_voted_users
1920-1929	116392
1930-1939	804839
1940-1949	230838
1950-1959	678336
1960-1969	2982551
1970-1979	8269828
1980-1989	19344369
1990-1999	69620489
2000-2009	165922941
2010-2019	115195692



Apply pivot chart on given dataset, in the pivot chart I put title years in rows and Sum of num_voted_users in values and group years by 10 Ex- 1920-1929 then I found 10 decades and also it's voted user value to use this value I created a Bar chart to visualize the data.

Insights :

There is 10 Decades in given bar chart to analysis the chart I found the maximum number of users are voted between year 2000 to 2009 in this decades number of votes is 165922941 and also the minimum number of votes obtained year between 1920 to 1929 the figure of votes is 116392

Thank You !