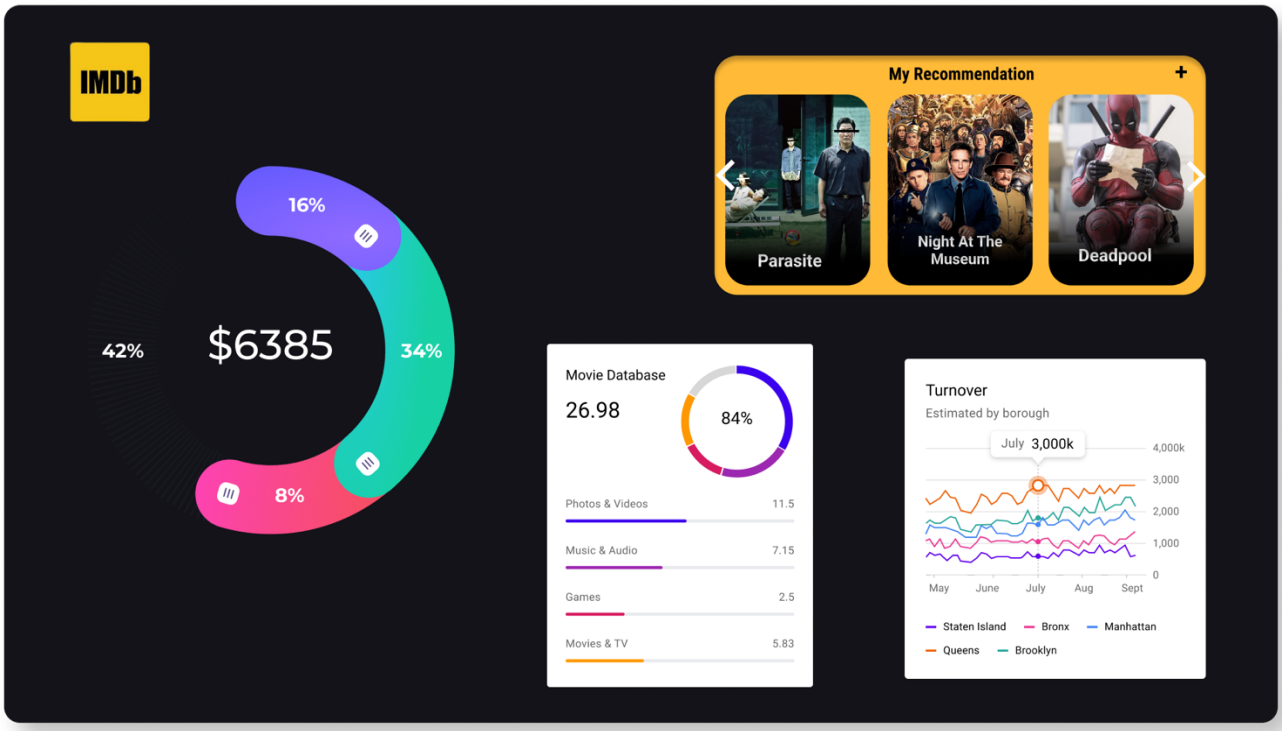
IMDB Movie Analysis

Final Project-1

By Mishree Bagdai



Project Description-

The objective of this project is to analyze a dataset containing information about various movies from IMDb. The dataset consists of multiple columns with details such as movie title, director name, IMDb rating, budget, gross, and other relevant information. The project involves framing a problem, cleaning the data, exploring the dataset, and deriving insights from it.

The analysis tasks include identifying movies with the highest profit, determining the IMDb Top 250 movies and top foreign language films, finding the best directors based on IMDb scores, identifying popular genres, and exploring the impact of lead actors such as Meryl Streep, Leonardo DiCaprio, and Brad Pitt. Additionally, trends in user voting over decades will be visualized. The project aims to provide a detailed report that presents the analysis findings in a cohesive and engaging data story.

Approach-

The approach for this project involves several key steps. First, the dataset will be thoroughly understood, examining the columns and their meanings. Any missing values, outliers, or inconsistencies will be identified and addressed

through data cleaning operations, including dropping irrelevant columns and handling missing data.
Next, a problem statement will be framed based on the dataset and initial observations. This problem statement will guide the analysis and exploration of the data, aiming to shed light on specific aspects of the movie dataset. The 5 Whys technique will be applied to delve deeper into the root causes of the identified problem.
After problem framing, various analysis tasks will be performed. This includes identifying movies with the highest profit by creating a new column and sorting based on the profit values. Additionally, the IMDb Top 250 movies will be determined by considering the IMDb rating and minimum number of voted users. Foreign language films within the IMDb Top 250 will also be identified. Furthermore, the dataset will be grouped by director name to find the top 10 directors with the highest mean IMDb scores, with tiebreakers sorted alphabetically. The popular genres will be explored based on previous analysis results.
Moreover, the dataset will be enhanced by creating columns specific to lead actors such as Meryl Streep, Leonardo DiCaprio, and Brad Pitt. The rows of these columns will be combined, and the mean of critical and audience reviews will be calculated for each actor to determine the audience and critic-favorite actors.
Lastly, the change in the number of voted users over decades will be analyzed by creating a decade column and visualizing the results using a bar chart. This will provide insights into voting trends throughout the years.
Throughout the project, data analysis techniques and visualization tools will be utilized to derive meaningful insights from the dataset. The findings will be compiled into a detailed report that tells a compelling data story, presenting the analysis results and addressing the problem statement in a clear and informative manner.

Tech-Stack used-

The tech stack used for this project consists of Microsoft Excel and Microsoft Word. Excel is employed for data cleaning, analysis, and visualization tasks, utilizing its functionalities for data transformation, calculations, and charting. Word is utilized for creating a detailed report that presents the analysis findings, incorporating text, tables, charts, and images in a visually appealing and structured format. The combination of Excel and Word enables efficient data processing, analysis, and reporting for the project.

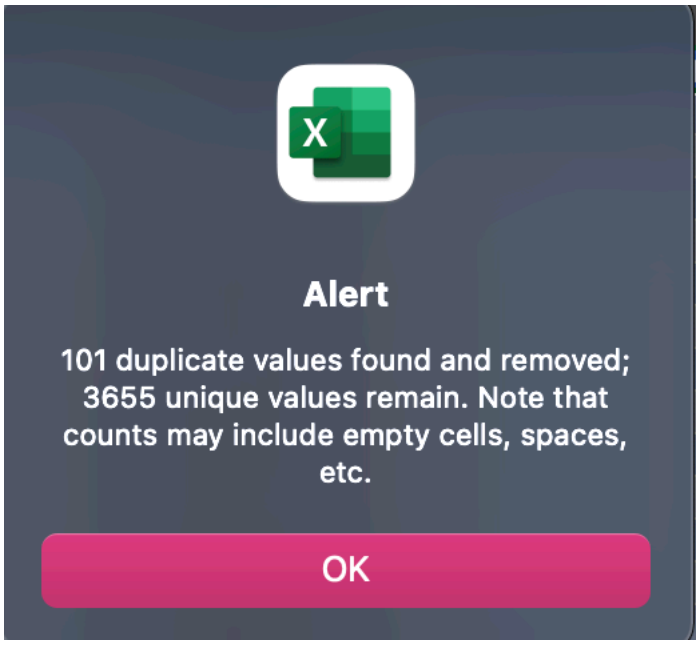
Results & Insights-

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

Before proceeding with the analysis, it is crucial to clean the dataset. This involves tasks such as dropping unnecessary columns, handling null values, and performing any other necessary data cleaning operations.

To clean the dataset I first formatted the dataset into a table form and added filters for all the columns, this helped me uncheck the (blank) values in each column, and that's how I was able to remove all the NULL values in the dataset.

Once all the NULL values were removed there were multiple duplicates in the dataset; in order to remove those values I used the built-in 'remove duplicates' function for the entire dataset and I realised that there were 101 duplicates.

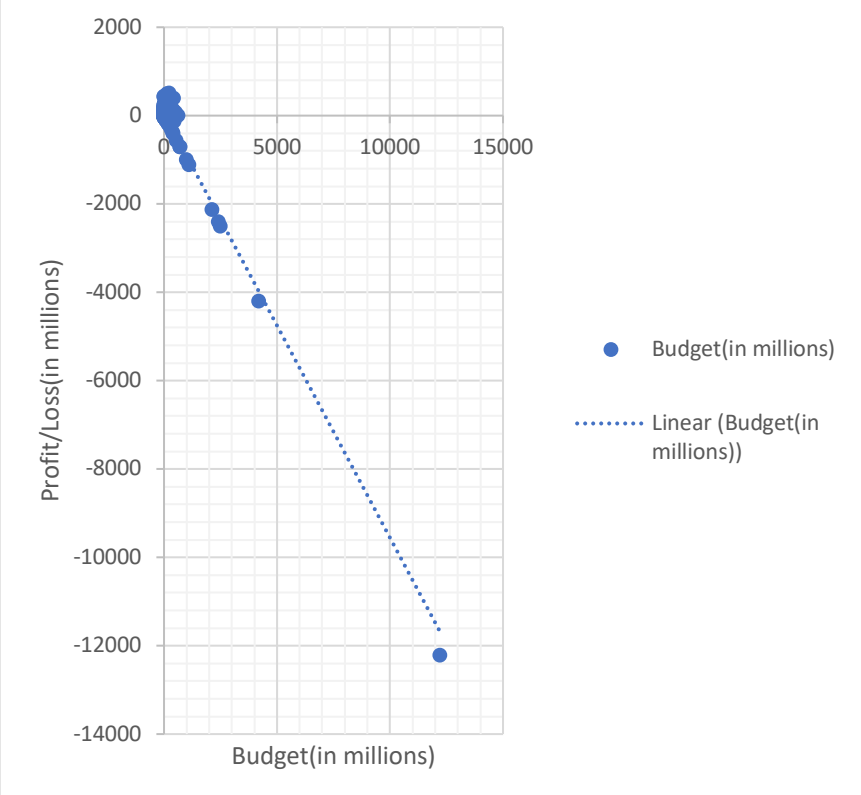


Later, I created a new sheet and renamed it as 'CLEAN_data' which consisted of the cleaned data that can be further used for analysis.

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?

AB	AC	AD
ebook_likes	profit	
33000	523505847	
0	9404152	
85000	-44925825	
164000	198130642	
24000	-190641321	
0	78530303	
29000	-59192738	
118000	208991599	
10000	51956980	
197000	80249062	
0	-8930592	
0	-31631573	
5000	198032628	
48000	-125710090	

	AC	AD
_likes	profit	
7000	-12213298588	
4000	-4199788333	
607	-2499804112	
11000	-2397701809	
973	-2127109510	
0	-1099560838	
339	-989962610	
539	-698312689	
659	-696724557	
0	-553005191	
124	-399545745	
0	-375868702	
0	-299897945	
24000	-190641321	
0	-188094481	
10000	-164334574	
44000	-143826840	



Result- we first created a column with the title 'profit' and the sorted the same in ascending order.

We find the following points to be our outliers,

profit	budget
-12213298588	12215500000
-4199788333	4200000000
-2499804112	2500000000
-2397701809	2400000000
-2127109510	2127519898

Insight –
The top 10 profitable movies were:

Avatar¬†	523505847
Jurassic World¬†	502177271
Titanic¬†	502177271
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

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

Rank	director_name	genres	IMDb_Top_250	num_voted_users	language	country	imdb_score
1	Frank Darabont	Crime|Drama	The Shawshank Redemption¬†	1689764	English	USA	9.3
2	Francis Ford Coppola	Crime|Drama	The Godfather¬†	1155770	English	USA	9.2
3	Francis Ford Coppola	Crime|Drama	The Godfather: Part II¬†	790926	English	USA	9
4	Christopher Nolan	Action|Crime|Drama|Thriller	The Dark Knight¬†	1676169	English	USA	9
5	Sergio Leone	Western	The Good, the Bad and the Ugly¬†	503509	Italian	Italy	8.9
6	Steven Spielberg	Biography|Drama|History	Schindler's List¬†	865020	English	USA	8.9
7	Quentin Tarantino	Crime|Drama	Pulp Fiction¬†	1324680	English	USA	8.9
8	Peter Jackson	Action|Adventure|Drama|Fantasy	The Lord of the Rings: The Return of the King¬†	1215718	English	USA	8.9
9	David Fincher	Drama	Fight Club¬†	1347461	English	USA	8.8
10	Christopher Nolan	Action|Adventure|Sci-Fi|Thriller	Inception¬†	1468200	English	USA	8.8
11	Peter Jackson	Action|Adventure|Drama|Fantasy	The Lord of the Rings: The Fellowship of the Ring¬†	1238746	English	New Zealand	8.8
12	Irvin Kershner	Action|Adventure|Fantasy|Sci-Fi	Star Wars: Episode V - The Empire Strikes Back¬†	837759	English	USA	8.8
13	Robert Zemeckis	Comedy|Drama	Forrest Gump¬†	1251222	English	USA	8.8
14	Akira Kurosawa	Action|Adventure|Drama	Seven Samurai¬†	229012	Japanese	Japan	8.7

Top 250 movies with num_voted_users greater than 25,000.

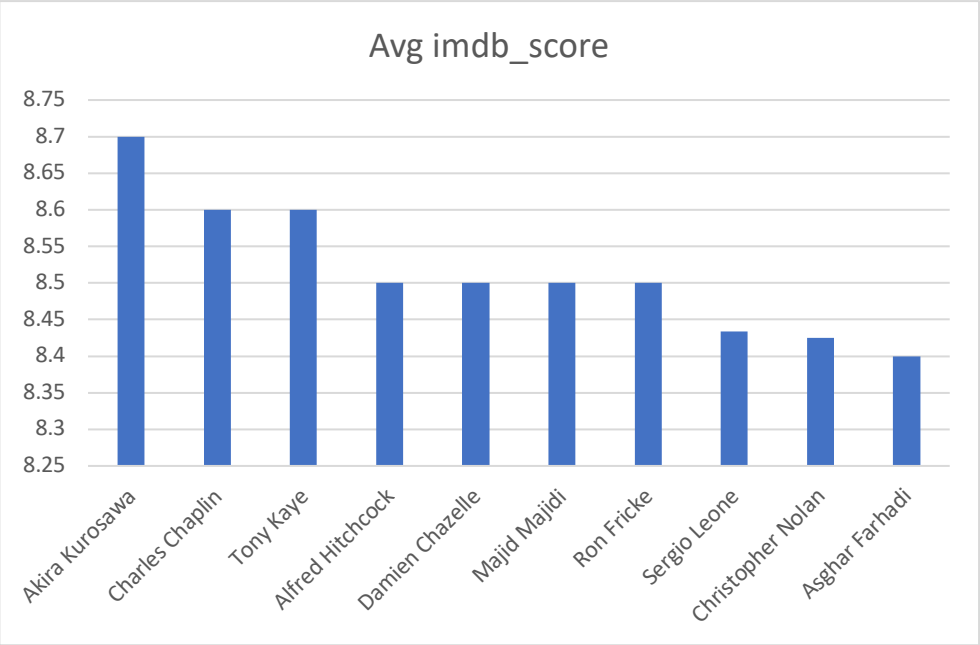
Rank	director_name	genres	IMDb_Top_250	num_voted_users	language	country	imdb_score
5	Sergio Leone	Western	The Good, the Bad and the Ugly¬†	503509	Italian	Italy	8.9
14	Akira Kurosawa	Action|Adventure|Drama	Seven Samurai¬†	229012	Japanese	Japan	8.7
15	Fernando Meirelles	Crime|Drama	City of God¬†	533200	Portuguese	Brazil	8.7
21	Hayao Miyazaki	Adventure|Animation|Family|Fantasy	Spirited Away¬†	417971	Japanese	Japan	8.6
30	Majid Majidi	Drama|Family	Children of Heaven¬†	27882	Persian	Iran	8.5
31	Florian Henckel von Donnersmarck	Drama|Thriller	The Lives of Others¬†	259379	German	Germany	8.5
47	Hayao Miyazaki	Adventure|Animation|Fantasy	Princess Mononoke¬†	221552	Japanese	Japan	8.4
48	Jean-Pierre Jeunet	Comedy|Romance	AmV©lie¬†	534262	French	France	8.4
51	Wolfgang Petersen	Adventure|Drama|Thriller|War	Das Boot¬†	168203	German	West Germa	8.4
53	Chan-wook Park	Drama|Mystery|Thriller	Oldboy¬†	356181	Korean	South Korea	8.4
56	Asghar Farhadi	Drama|Mystery	A Separation¬†	151812	Persian	Iran	8.4
62	Oliver Hirschbiegel	Biography|Drama|History|War	Downfall¬†	248354	German	Germany	8.3
63	Fritz Lang	Drama|Sci-Fi	Metropolis¬†	111841	German	Germany	8.3
64	Thomas Vinterberg	Drama	The Hunt¬†	170155	Danish	Denmark	8.3
85	Hayao Miyazaki	Adventure|Animation|Family|Fantasy	Howl's Moving Castle¬†	214091	Japanese	Japan	8.2
90	Denis Villeneuve	Drama|Mystery|War	Incendies¬†	80429	French	Canada	8.2
98	Juan JosV© Campanella	Drama|Mystery|Thriller	The Secret in Their Eyes¬†	131831	Spanish	Argentina	8.2
99	Guillermo del Toro	Drama|Fantasy|War	Pan's Labyrinth¬†	467234	Spanish	Spain	8.2
108	Katsuhiro vîtomo	Action|Animation|Sci-Fi	Akira¬†	106160	Japanese	Japan	8.1
109	Je-kyu Kang	Action|Drama|War	Tae Guk Gi: The Brotherhood of War¬†	31943	Korean	South Korea	8.1
111	Alejandro AmenV"bar	Biography|Drama|Romance	The Sea Inside¬†	64556	Spanish	Spain	8.1
112	JosV© Padilha	Action|Crime|Drama|Thriller	Elite Squad¬†	81644	Portuguese	Brazil	8.1
114	Thomas Vinterberg	Drama	The Celebration¬†	65951	Danish	Denmark	8.1
116	Alejandro G. IV±V"rritu	Drama|Thriller	Amores Perros¬†	173551	Spanish	Mexico	8.1
160	Karan Johar	Adventure|Drama|Thriller	My Name Is Khan¬†	69759	Hindi	India	8
166	Vincent Paronnaud	Animation|Biography|Drama|War	Persepolis¬†	70194	French	France	8
168	Ari Folman	Animation|Biography|Documentary|Drama|History|War	Waltz with Bashir¬†	46107	Hebrew	Israel	8
170	Walter Salles	Drama	Central Station¬†	28951	Portuguese	Brazil	8
172	Sergio Leone	Action|Drama|Western	A Fistful of Dollars¬†	147566	Italian	Italy	8
210	Yimou Zhang	Action|Adventure|History	Hero¬†	149414	Mandarin	China	7.9
211	Yimou Zhang	Action|Adventure|History	Hero¬†	149414	Mandarin	China	7.9
216	Michael Haneke	Drama|Romance	Amour¬†	70382	French	France	7.9
218	Clint Eastwood	Drama|History|War	Letters from Iwo Jima¬†	132149	Japanese	USA	7.9
219	Christophe Barratier	Drama|Music	The Chorus¬†	44151	French	France	7.9
220	FabiV"n Bielinsky	Crime|Drama|Thriller	Nine Queens¬†	38215	Spanish	Argentina	7.9
222	Cristian Mungiu	Drama	4 Months, 3 Weeks and 2 Days¬†	44763	Romanian	Romania	7.9

D. Best Directors: TGroup 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

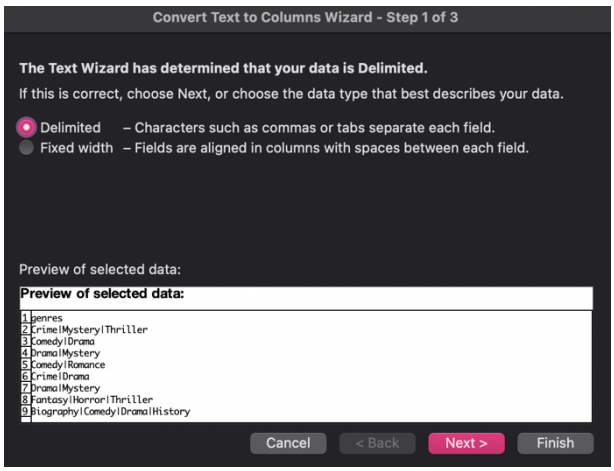
Result- I used pivot tables to find the average imdb score for each directors and the top 10 directors are as follows.

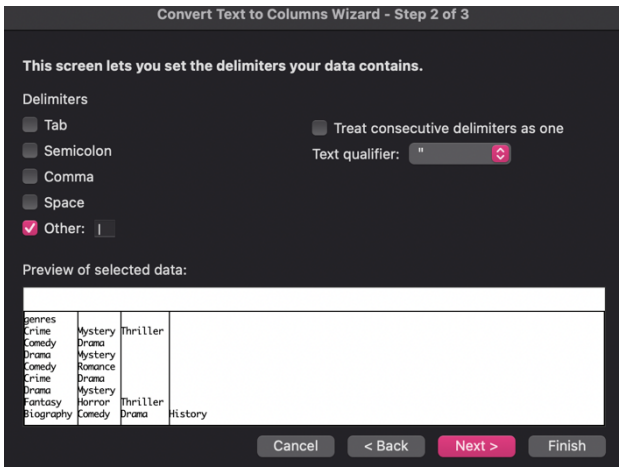
Top 10 Directors	Avg imdb_score
Akira Kurosawa	8.7
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4



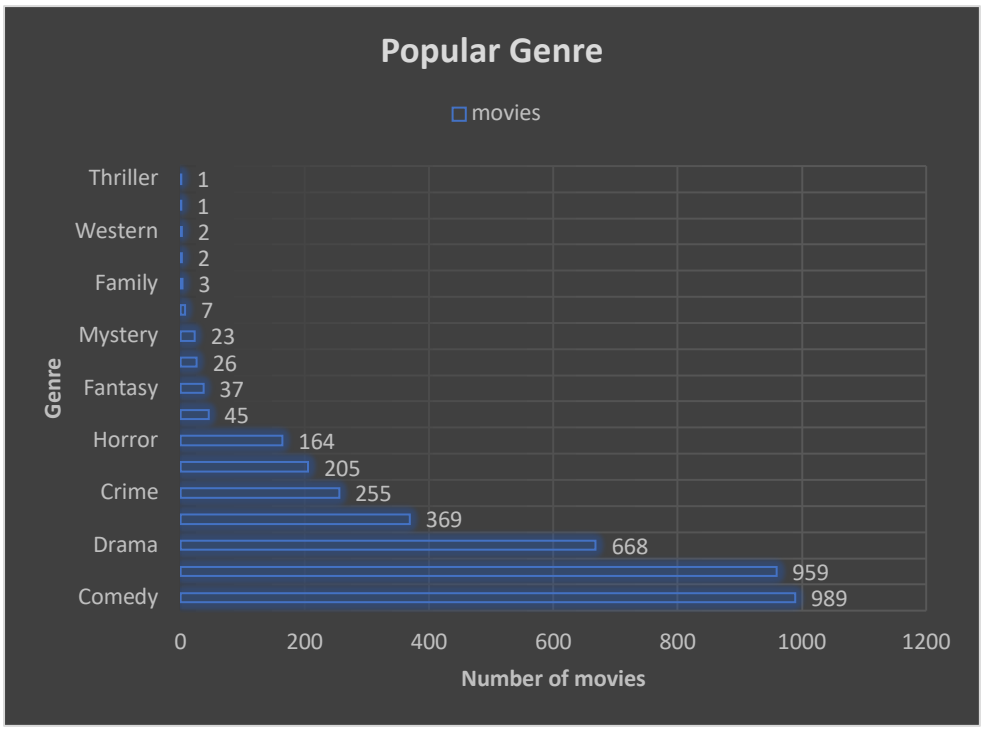
E. Popular Genres: Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres

Result – since there were multiple genres for each movie, I used text-to column feature and separated the genres and used the first genre as the primary genre for each movie, and then I used the pivot table feature to find the count of movies in each genre.





primary genre	movies
Comedy	989
Action	959
Drama	668
Adventure	369
Crime	255
Biography	205
Horror	164
Animation	45
Fantasy	37
Documentary	26
Mystery	23
Sci-Fi	7
Family	3
Musical	2
Western	2
Romance	1
Thriller	1



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.

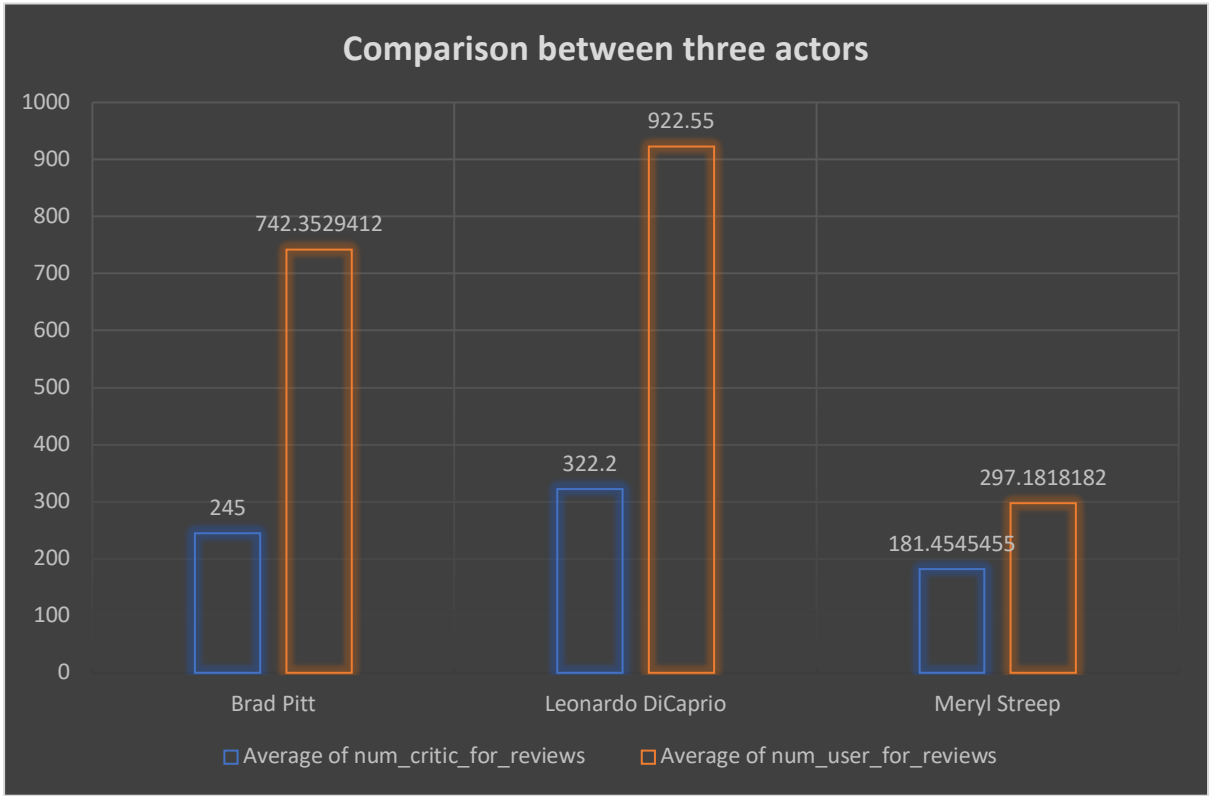
Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

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

Brad Pitt	Leonardo DiCaprio	Meryl Streep
Ocean's Eleven¬†	Titanic¬†	The Devil Wears Prada¬†
Mr. & Mrs. Smith¬†	Inception¬†	Out of Africa¬†
Interview with the Vampire: The Vampire Chronicles¬†	Catch Me If You Can¬†	Julie & Julia¬†
Fury¬†	Django Unchained¬†	Hope Springs¬†
Ocean's Twelve¬†	The Revenant¬†	It's Complicated¬†
Babel¬†	Shutter Island¬†	The Iron Lady¬†
Killing Them Softly¬†	The Departed¬†	The Hours¬†
True Romance¬†	The Great Gatsby¬†	A Prairie Home Companion¬†
By the Sea¬†	Romeo + Juliet¬†	The River Wild¬†
The Tree of Life¬†	The Man in the Iron Mask¬†	One True Thing¬†
The Curious Case of Benjamin Button¬†	The Wolf of Wall Street¬†	Lions for Lambs¬†
Fight Club¬†	J. Edgar¬†	
The Assassination of Jesse James by the Coward Robert Ford¬†	The Aviator¬†	
Seven Years in Tibet¬†	Marvin's Room¬†	
Sinbad: Legend of the Seven Seas¬†	The Beach¬†	
Troy¬†	Revolutionary Road¬†	
Spy Game¬†	The Quick and the Dead¬†	
	Gangs of New York¬†	
	Body of Lies¬†	
	Blood Diamond¬†	

Row Labels	Average of num_critic_for_reviews	Average of num_user_for_reviews
Brad Pitt	245	742.3529412
Leonardo DiCaprio	322.2	922.55
Meryl Streep	181.4545455	297.1818182
Grand Total	**262.6041667**	**715.4166667**

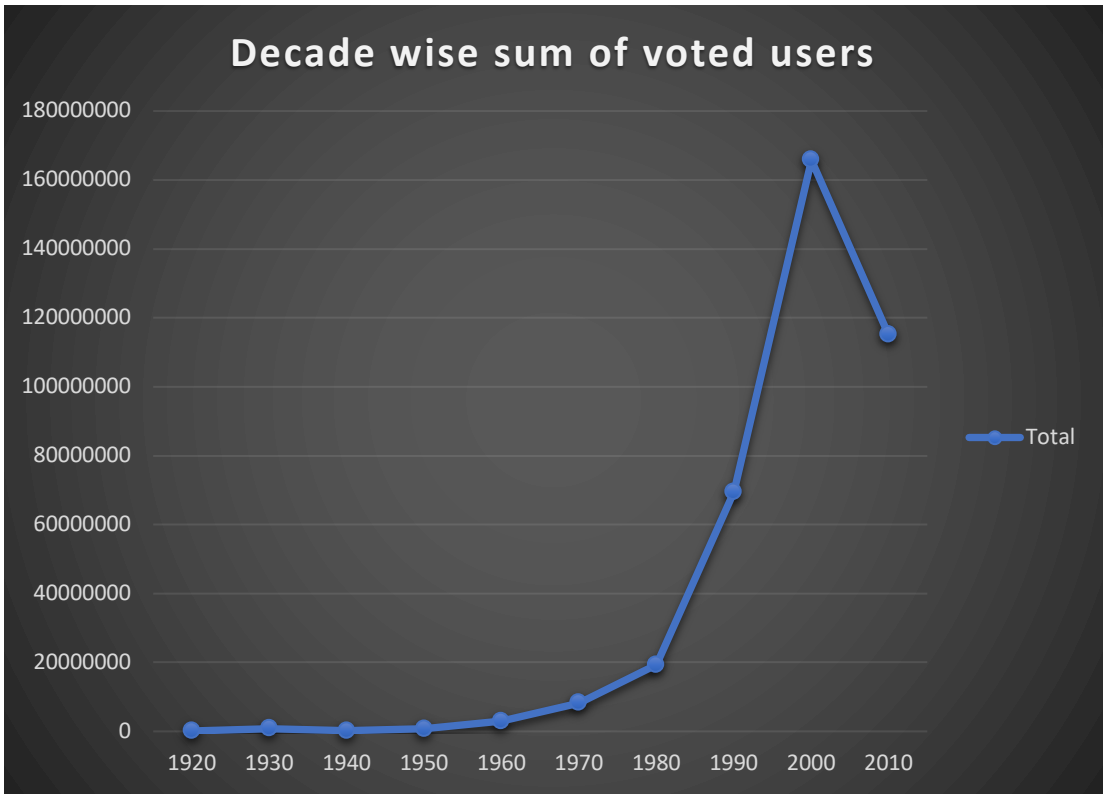


Result - I used pivot tables to find out the required mean for all the three actors.

Insight – we can see that Leonardo DiCaprio is the user-favourite the critic-favourite actor.

To calculate the decade I used the following formula in excel, and created a new column called decade.
=LEFT([@[title_year]],3)&"0"

decade	Sum of num_voted_users
1920s	116387
1930s	804839
1940s	159517
1950s	678336
1960s	2982551
1970s	8269025
1980s	19344369
1990s	69482050
2000s	165749275
2010s	115032219



Insight – Therefore we can see the trend in num_voted_users for each decade using the line chart.

Conclusion –

In conclusion, this project utilized a combination of data analysis techniques, data cleaning operations, and visualization tools to derive meaningful insights from a dataset of IMDb movies. By leveraging the functionalities of Microsoft Excel and Microsoft Word, the dataset was cleaned, analyzed, and visualized to address specific problem statements and uncover key trends and patterns.

The project highlighted movies with the highest profit, identified the IMDb Top 250 movies and top foreign language films, and determined the best directors based on IMDb scores. Additionally, popular genres were explored, and the impact of lead actors such as Meryl Streep, Leonardo DiCaprio, and Brad Pitt was examined. Moreover, the change in the number of voted users over decades was analyzed to understand voting trends.

Link to excel file -
https://1drv.ms/x/s!AgXHafHWEd65as0Ktkwq_0_c5C4?e=puo7Yt