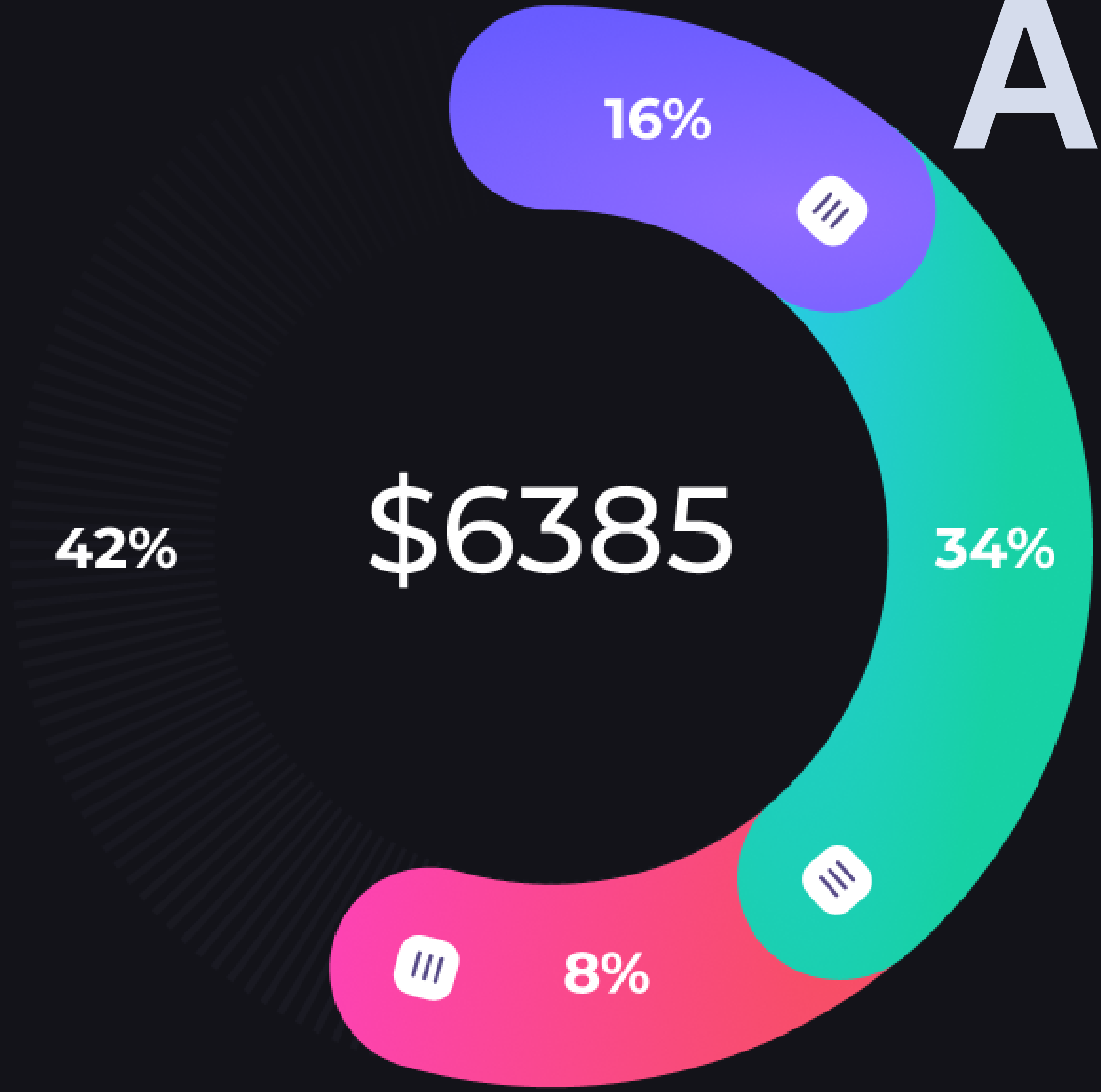
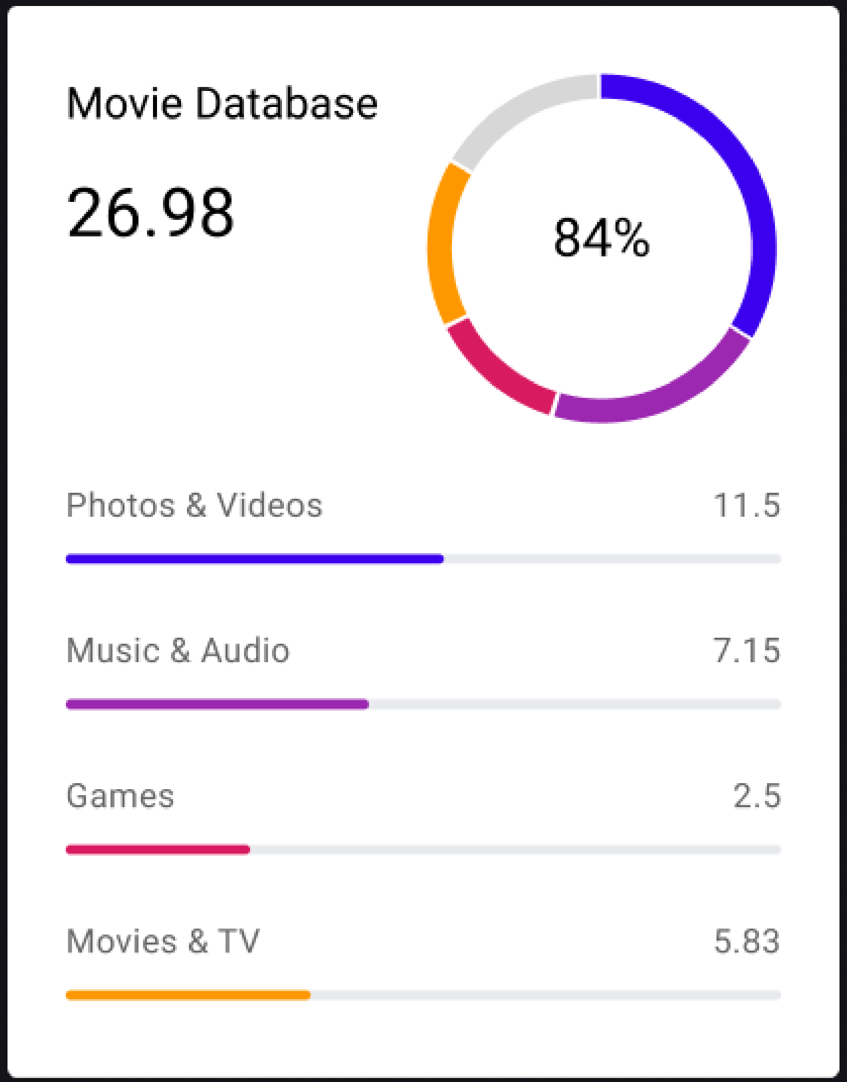
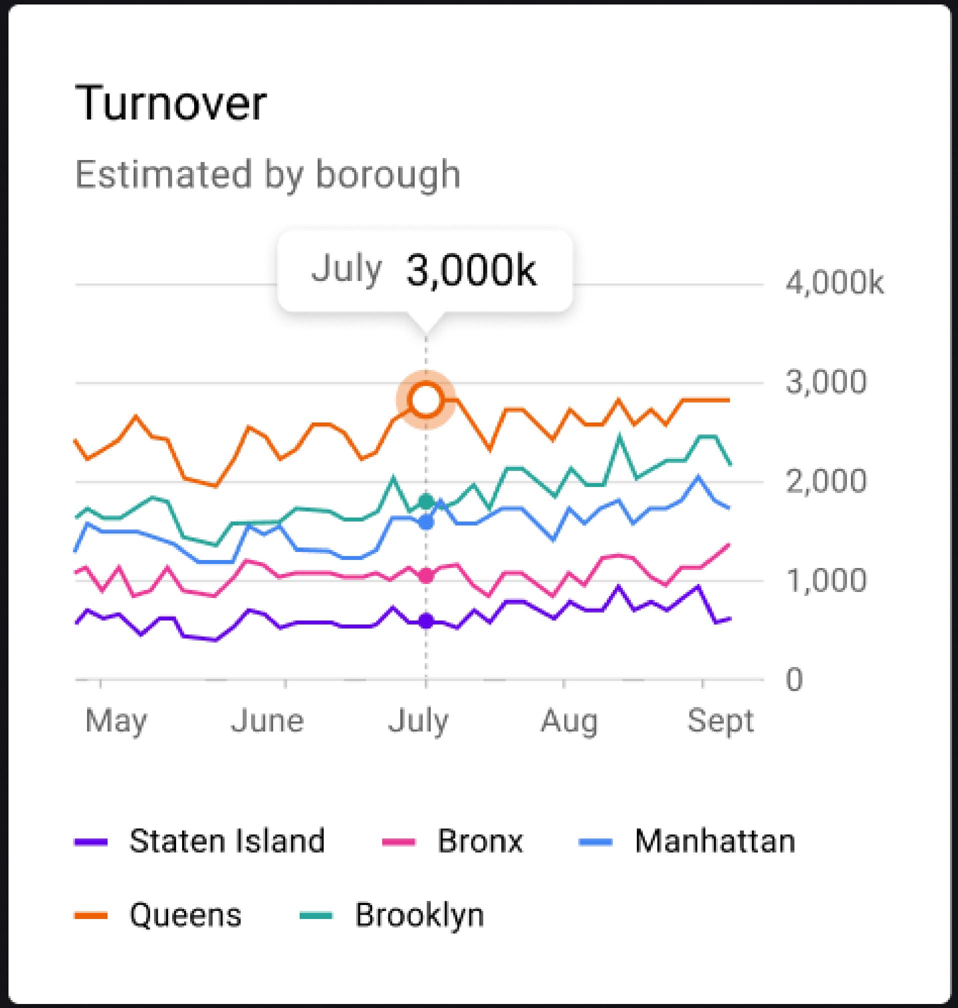
IMDB Movie Analysis







by Suraj Beloshe

Powered by

Project Description

1. The project involves analyzing a movie dataset using various data analysis techniques.
2. There are some tasks to perform like cleaning the data, finding movies with the highest profit, identifying the top 250 movies, finding the best directors, and identifying popular genres.
3. Additionally, the project involves creating new columns for some specific actors and combining them to identify the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.
4. The project also involves observing the change in number of voted users over decades using a bar chart
5. The main objective is to derive insights and trends from the dataset using data analysis techniques.

Approach

1. Understand Dataset
2. Evaluate Insights
3. Use specific columns
4. Perform Analysis
5. Note Down insights

Tech-Stack Used

Microsoft 365 Excel

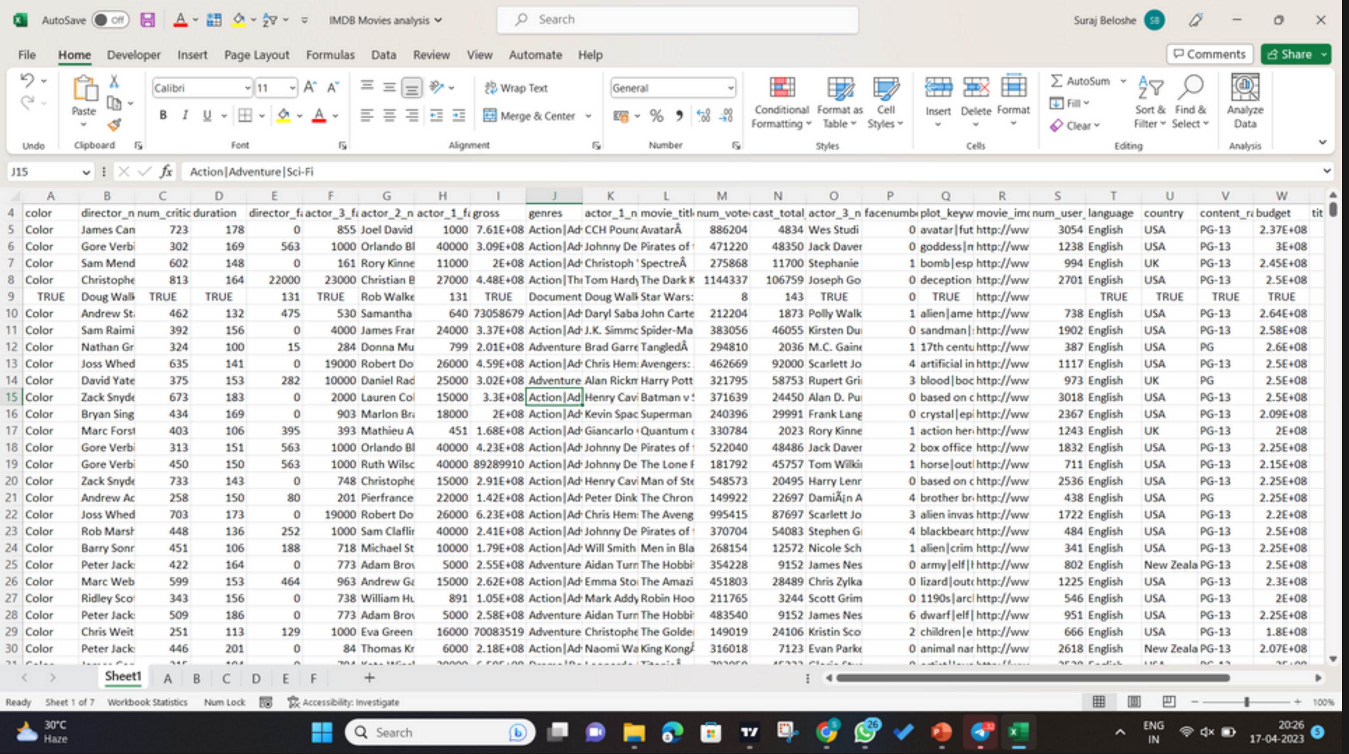
- Familiarity: MS Excel is a widely used software for data analysis and management. You might already be familiar with its interface and functions, which can make the project easier to complete.

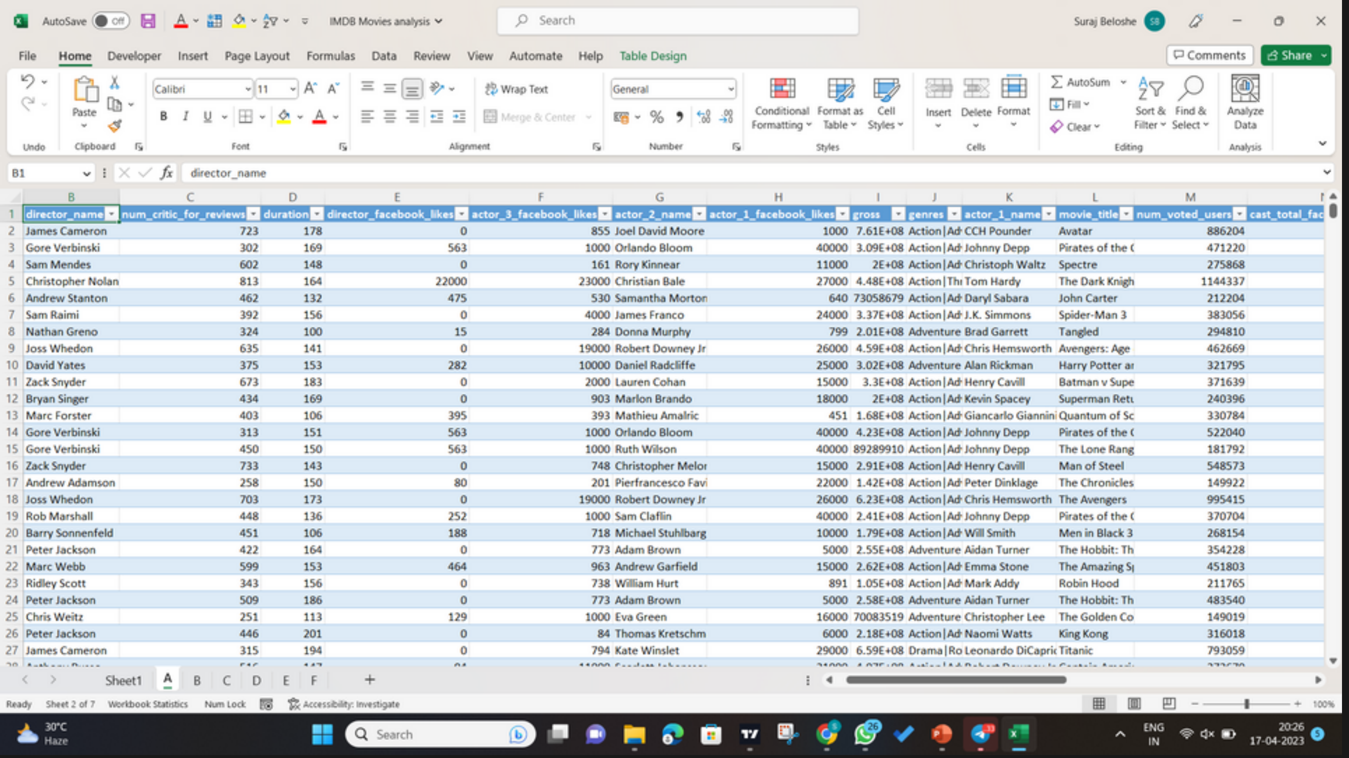
- Accessibility: MS Excel 365 is a cloud-based software that can be accessed from anywhere with an internet connection. This means that you can work on your project from different devices and collaborate with others easily.

- Data analysis tools: MS Excel 365 provides a variety of data analysis tools, such as sorting, filtering, pivot tables, and charts, which can help you derive insights and trends from your data.

- Integration with other Microsoft tools: MS Excel 365 can be integrated with other Microsoft tools such as Power BI, SharePoint, and Teams, which can enhance your data analysis and presentation capabilities.

Insights





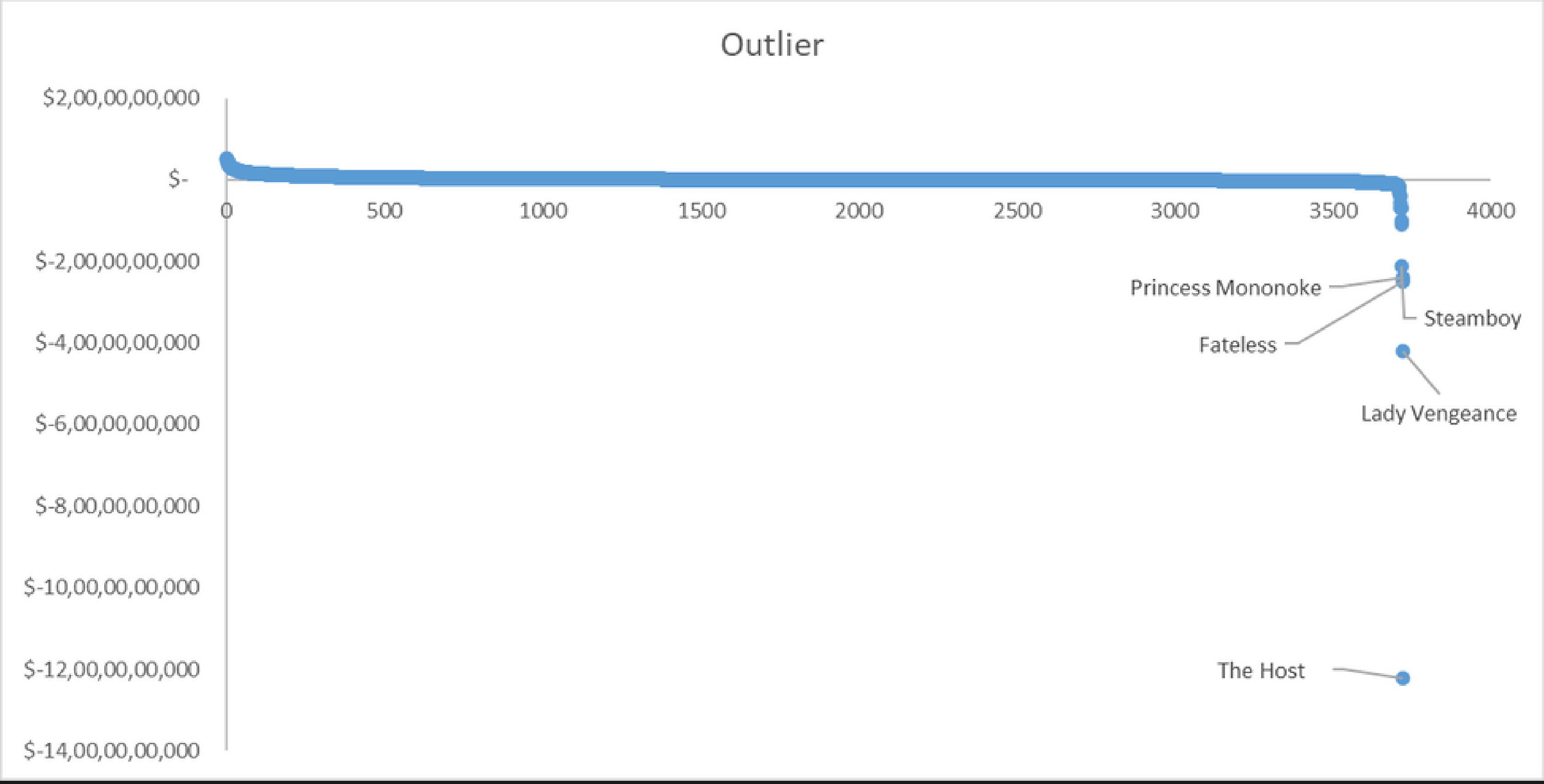
A. Cleaning the data:

- Data cleaning is an **essential step** in ensuring data consistency and accuracy. Removing any duplicate records in the dataset is crucial for this purpose.
- To ensure the integrity of complete data analysis, any missing values or null data should be replaced with appropriate values such as mean, median, or mode.
- The given dataset had **28 columns and 5044 rows. total 1,41,232 cells**
- There was a **special character** in the Movie Title column that was removed to maintain accuracy & readability.
- The dataset had **2697** blank cells, which could have had a significant impact on the findings. Therefore, it was crucial to remove this blank data.
- Using the "Go to Special" function, we identified the blank rows and replaced them with "TRUE".
- I used an **advanced filter with "TRUE"** to subtract the required data and moved it to another sheet, which we then converted into a table.
- Finally, I identified and removed duplicate values. In total, **33 duplicate** values were found, leaving **3723** unique rows.

Insights

B. Movies with highest profit

movie_title	budget	gross	Profit
Avatar	$ 23,70,00,000	$ 76,05,05,847	$ 52,35,05,847
Jurassic World	$ 15,00,00,000	$ 65,21,77,271	$ 50,21,77,271
Titanic	$ 20,00,00,000	$ 65,86,72,302	$ 45,86,72,302
Star Wars: Episode IV - A New Hope	$ 1,10,00,000	$ 46,09,35,665	$ 44,99,35,665
E.T. the Extra-Terrestrial	$ 1,05,00,000	$ 43,49,49,459	$ 42,44,49,459
The Avengers	$ 22,00,00,000	$ 62,32,79,547	$ 40,32,79,547
The Lion King	$ 4,50,00,000	$ 42,27,83,777	$ 37,77,83,777
Star Wars: Episode I - The Phantom Menace	$ 11,50,00,000	$ 47,45,44,677	$ 35,95,44,677
The Dark Knight	$ 18,50,00,000	$ 53,33,16,061	$ 34,83,16,061
The Hunger Games	$ 7,80,00,000	$ 40,79,99,255	$ 32,99,99,255
Deadpool	$ 5,80,00,000	$ 36,30,24,263	$ 30,50,24,263



- In order to determine the most profitable movies, I first needed to calculate the profits. Using a simple arithmetic formula, I calculated the profits for each movie.
- Next, I sorted the table with a "largest to smallest" filter and selected the **top 10 movies with the highest profits.**
- However, I discovered an outlier in our data. An outlier is an observation that significantly differs from other observations in a dataset.
- To identify the outliers, **we used a scatter chart** in Excel to plot the data points and look for points that are far away from the general trend of the data.
- Outliers can be easily visualized as points that are outside the main cluster of data points on the scatter chart.
- In our data, we found that the movie **'The Host' had a profit of $-12,21,32,98,588.00, which was significantly lower than the other movies and thus stood out as an outlier.**
- There were four other movies that also stood out as outliers, as shown in the scatter chart.

Insights

C. Top 250

Rank	movie_title	num_voted_users	language	imdb_score
1	The Shawshank Redemption	1689764	English	9.3
2	The Godfather	1155770	English	9.2
3	The Dark Knight	1676169	English	9
248	The Last of the Mohicans	113068	English	7.8
249	Apocalypto	236000	Maya	7.8
250	Fantastic Mr. Fox	139114	English	7.8

Rank	movie_title	num_voted_users	language	imdb_score
8	The Good, the Bad and the Ugly	503509	Italian	8.9
19	City of God	533200	Portugues	8.7
20	Seven Samurai	229012	Japanese	8.7
219	Letters from Iwo Jima	132149	Japanese	7.9
223	Amour	70382	French	7.9
249	Apocalypto	236000	Maya	7.8

- I were instructed to find the top 250 movies based on their IMDB score, with a minimum of 25,000 votes.
- Using an advanced filter, I separated the movies with 25,000 or more voted users from the rest.
- I then used a **LARGE & SEQUENCE formula** to get the top 250 movies based on their IMDB score.
- However, I faced the challenge of duplicate values.
- To solve this, I created a helper column to get unique values.
- Finally, I used an "XLOOKUP" function to obtain all the top 250 movies as required.
- I also wanted to find foreign language movies from the top 250 list.
- I used an **advanced filter with the criterion "not equal to English"** to extract all the foreign language movies from the top 250.

Insights

D. Best Directors

Sr. N	director_name	imdb_score
1	Akira Kurosawa	8.70
2	Charles Chaplin	8.60
3	Tony Kaye	8.60
4	Alfred Hitchcock	8.50
5	Damien Chazelle	8.50
6	Majid Majidi	8.50
7	Ron Fricke	8.50
8	Christopher Nolan	8.43
9	Sergio Leone	8.43
10	Asghar Farhadi	8.40
11	Richard Marquand	8.40
12	Billy Wilder	8.30
13	Fritz Lang	8.30
14	Lee Unkrich	8.30
15	Lenny Abrahamson	8.30
16	Hayao Miyazaki	8.23
17	Pete Docter	8.23
18	Elia Kazan	8.20
19	George Roy Hill	8.20
20	Joshua Oppenheimer	8.20

- To determine the best directors based on IMDb scores, I separated the two columns and grouped them as per the instructions.
- I used a pivot table to group the data and find the mean IMDb score for each director.
- By summarizing the values by average, we obtained a list of all directors with their respective mean IMDb scores.
- I then decided to find the top 20 directors from this list.
- To accomplish this, we used the LARGE formula and the SEQUENCE formula.
- With the help of a supporting column, I fetched the names of the directors and their IMDb scores.
- Further Sorted Data as instructed Alphabetically
- Finally, I obtained a list of the top 20 directors based on their mean IMDb scores.

Insights

E. Popular Genres

Popular Genres on the basis of Profit		
Sr. No.	genres	Sum of Profit
1	Comedy	3,00,43,60,840
2	Comedy|Romance	2,69,28,38,991
3	Action|Adventure|Fantasy|Sci-Fi	2,39,47,17,100

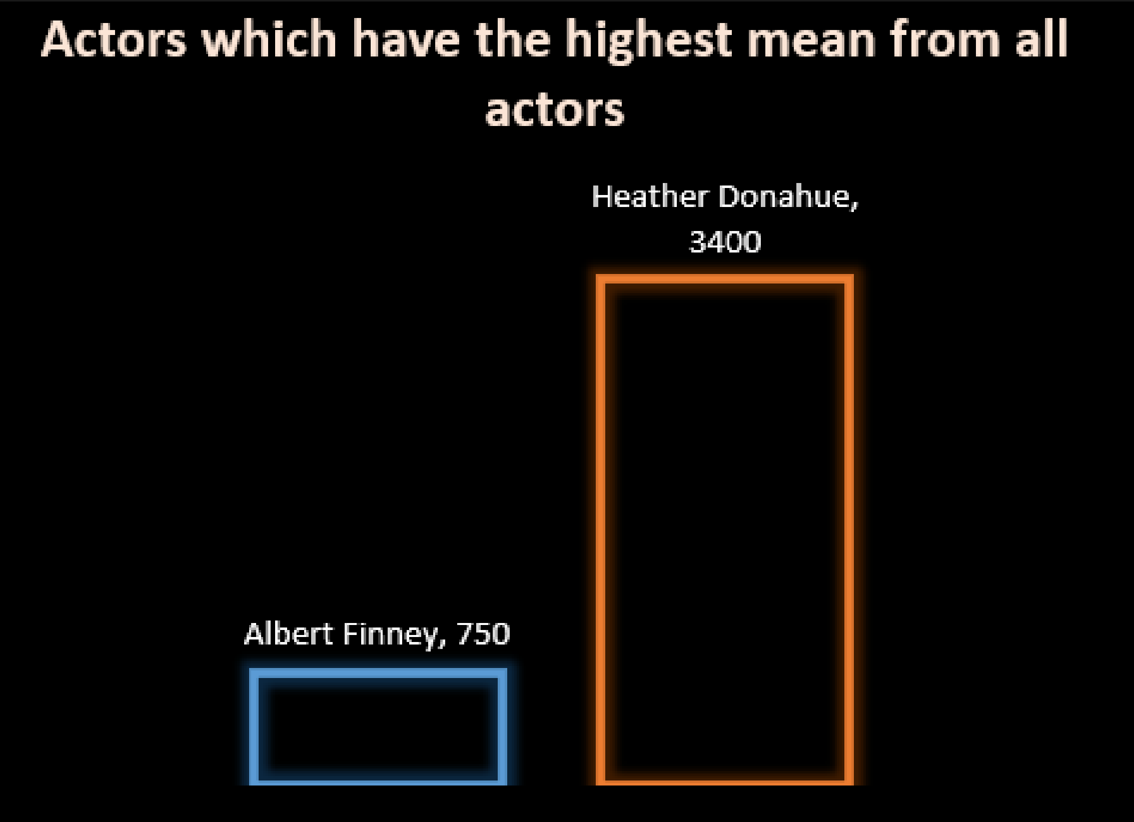
Popular Genres on the basis of number of voted users		
Sr. No.	genres	Sum of Num_Voted Users
1	Action|Adventure|Sci-Fi	1,46,23,810
2	Drama	1,23,19,080
3	Comedy	1,11,12,881

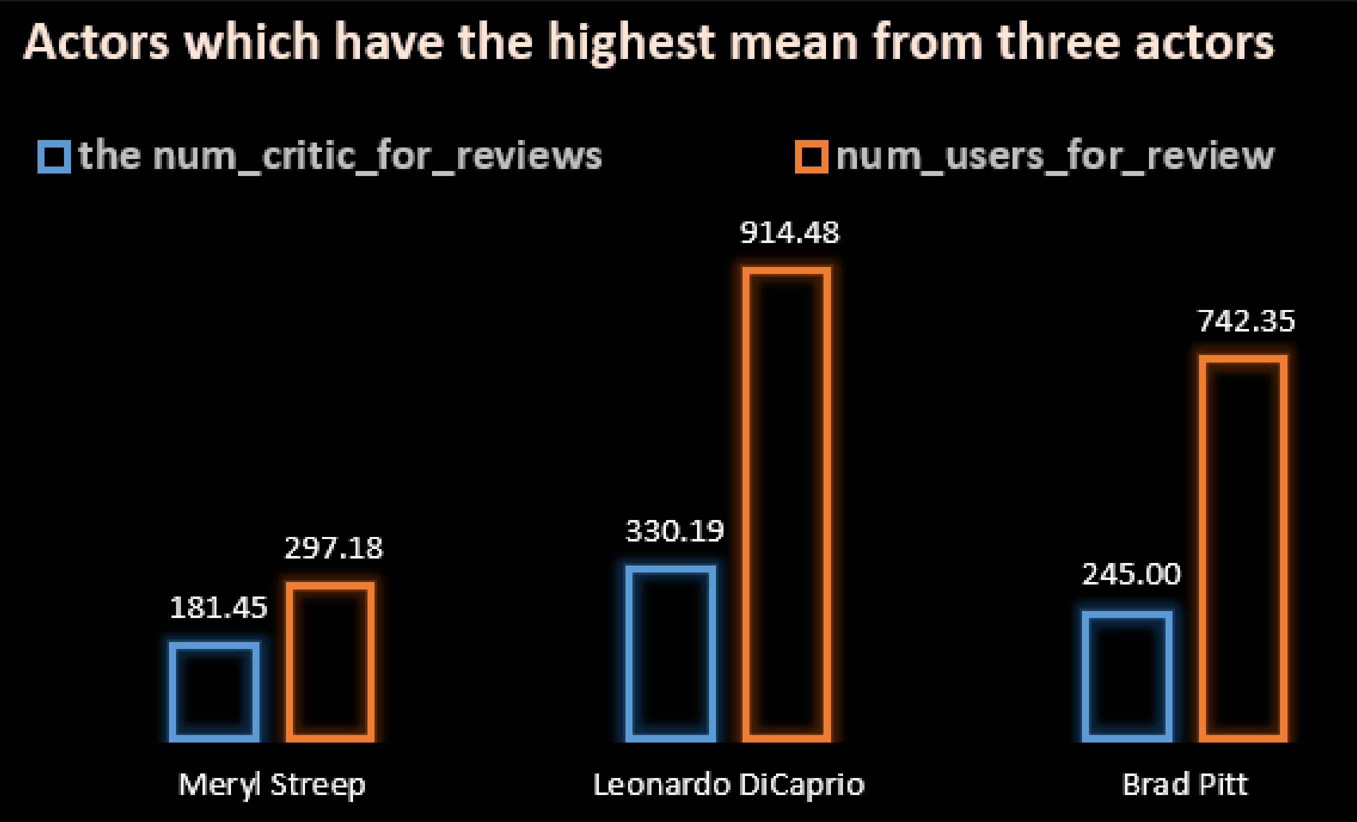
Popular Genres on the basis of Average of imdb score		
Sr. No.	genres	Average of imdb_score
1	Adventure|Animation|Drama|Family|Musical	8.50
2	Crime|Drama|Fantasy|Mystery	8.50
3	Adventure|Animation|Fantasy	8.40

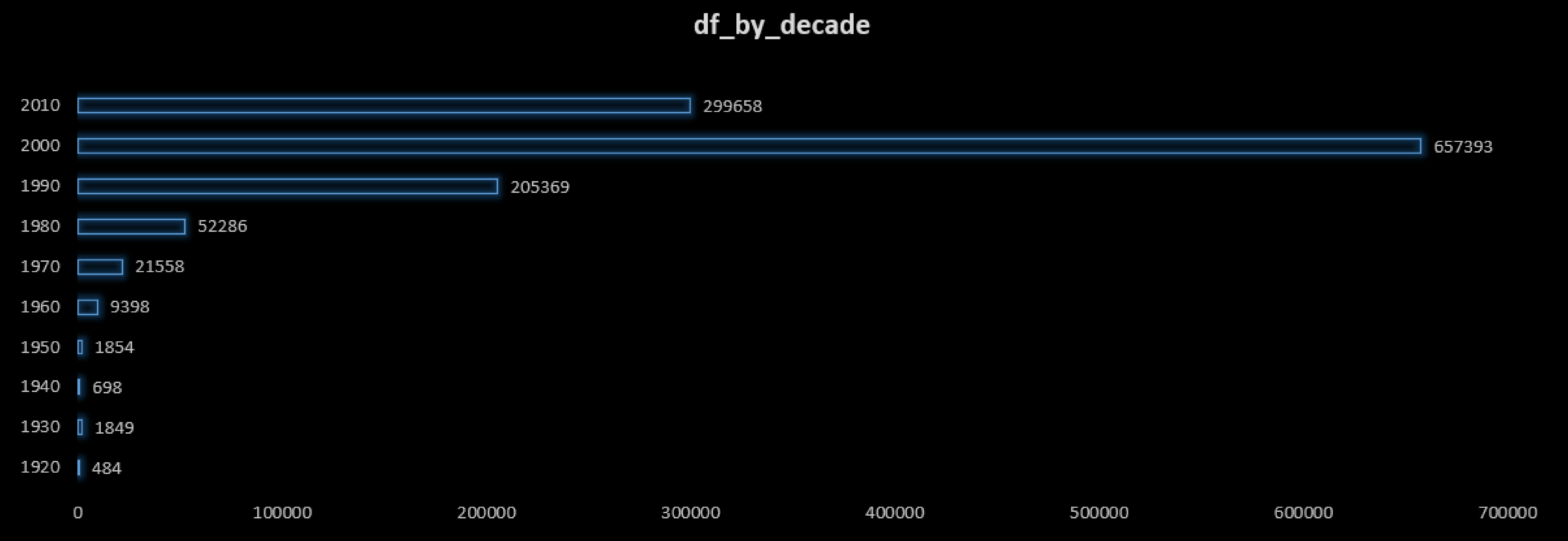
- For the first criterion, I calculated **movie profits** and used a pivot table with large sequence and XLOOKUP formulas to find the top three most popular genres based on earnings.
- For the second criterion, I used a pivot table to group data and find the **most voted genres.**
- For the third criterion, I used a unique value column and the sequence large and XLOOKUP formulas to calculate the **average IMDb** score for each genre.
- Using this method, I identified the top three most popular genres based on IMDb ratings.
- By using different criteria and data-driven methods, I were able to obtain a comprehensive list of the most popular genres in the movie industry.
- This information can be useful for movie producers and distributors in making decisions about which genres to invest in.

Insights

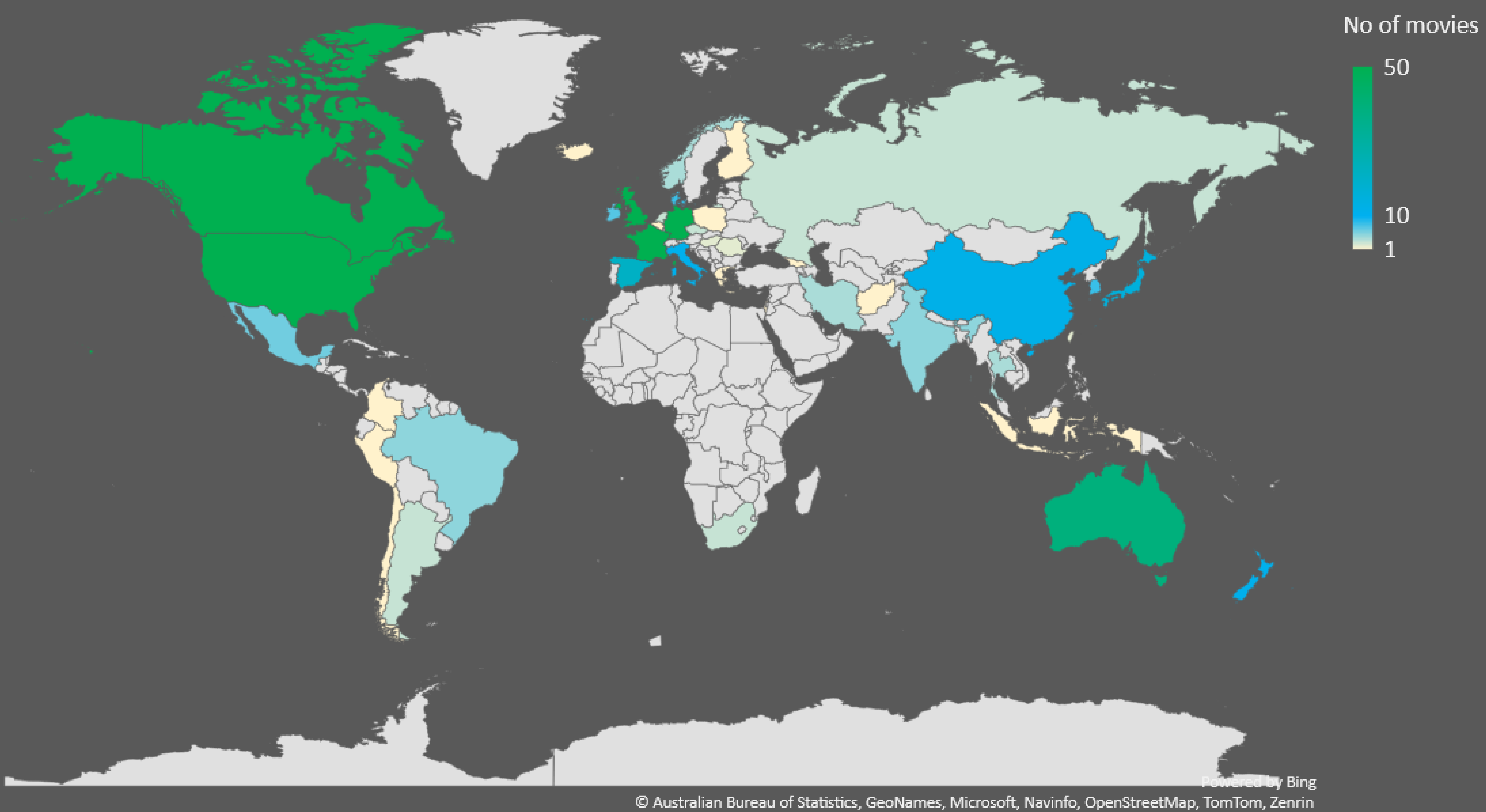
F. Charts







Country wise listing of movies



Result

1. Data cleaning is a crucial step in the data analysis process as it helps to ensure the accuracy and consistency of the data.
2. Data manipulation techniques like creating new columns, grouping the data by certain variables, and sorting the data can help in identifying patterns and insights.
3. Visualization techniques like scatter plots, bar charts and map charts can help in identifying trends, outliers and geographical representation in the data.
4. Grouping the data by certain variables can help in identifying trends and patterns in the data.
5. Using MS 365 to completed this project helps me to understand more about it and realize when some complex operation are taking more time than other language. Still it helps for data analysis with great accuracy and reliability.

Attachment:- Link for project file

Thank You for your time....