IMDB MOVIE ANALYSIS

Project Description:

This project aims to analyse a given dataset containing information about various IMDB movies. The dataset includes details such as movie name, year of release, genres, director and actor names, overall budget, gross collection, and IMDB ratings, etc. The analysis will focus on answering specific questions, including:

- **Movies with the Highest Profit:** Identify and analyse the movies that have generated the highest profits, considering the difference between the overall budget and gross collection.
- **IMDB Top 250 Movies:** Determine the top 250 movies based on IMDB ratings, showcasing the highest-rated films according to the audience.
- **Best Directors:** Explore and identify the most accomplished directors based on their filmography, considering factors such as critical acclaim, commercial success, and overall impact on the industry.
- **Most Popular Genres:** Analyse the dataset to identify the most popular movie genres among the listed movies, helping understand the audience's preferences.
- **Critic and Audience-Favourite Actors:** Identify the actors who receive consistent acclaim from both critics and audiences based on their involvement in highly rated movies and commercial successes.

By conducting a comprehensive analysis of this dataset, we aim to gain insights into the profitability of movies, audience preferences, and the contributions of directors and actors to the film industry.

Tech Stack Used:

I chose to work on Microsoft Excel for this project because it not only allows us to clean and manipulate the data based on our requirements using built-in functions and pivot tables but also provides a platform to visualize our findings to draw quick and meaningful insights at a glance.

Approach:

1. **Clean the dataset:**
   The first step in analysing the dataset and gaining valuable insights was to clean the data to ensure accurate and error-free analysis. The following steps were taken to clean the dataset:

- ***Download Dataset:*** The dataset was downloaded in CSV format and formatted into a table in MS Excel for easy manipulation and analysis.
- ***Dropping Unnecessary Columns:*** Columns that were not required to answer the analysis questions were dropped. The following columns were removed:
    - Colour
    - Duration
    - director_facebook_likes
    - actor_3_facebook_likes
    - actor_2_name
    - actor_1_facebook_likes
    - cast_total_facebook_likes
    - actor_3_name
    - facenumber_in_poster
    - plot_keywords
    - movie_imdb_link
    - country
    - content_rating
    - actor_2_facebook_likes
    - aspect_ratio
    - movie_facebook_likes

After removing these unnecessary columns, the dataset was left with 13 relevant columns for analysis.

- ***Removing Duplicates:*** To ensure data integrity, a check was performed for duplicate movie names along with the language. While duplicate entries for columns like director and actor names, gross, and genre were allowed (as multiple movies can have the same director or actor), movie titles should not have duplicates unless they are in different languages.
  Using a pivot table in Excel, it was identified that certain movie titles, such as 'King Kong', had three entries in the same language. To remove these duplicates, the 'Remove Duplicate' function in Excel was applied based on the 'movie_title' and 'language' columns.
- ***Removing null values:*** The next step was to identify and handle null values in the dataset. Firstly, the number of null values in each column was calculated using the 'COUNTBLANK()' function.
  The number of blank entries was divided by the total number of entries for each column and converted into a percentage.
  The 'gross' and 'budget' columns showed the highest percentage of null values. To address this the blank values in both the 'gross' and 'budget' columns were removed. This step removed almost all the null values in the dataset.
- ***Removing special characters:*** The movie titles ended with a special character ‘Â’ which was removed using ‘Find & Replace’ to clean the movie names.

By performing these data cleaning steps, the dataset was prepared for further analysis, ensuring accurate and reliable insights.

2. Movies with Highest Profit:

The first task was to find out the most profitable movies based on the difference in the gross revenue and the budget. The steps taken are as follows:

- A new column named ‘profit’ was introduced in the table which contained the difference between the ‘gross’ and ‘budget’ columns.
- Once the ‘profit’ was calculated, the table was sorted based on the profit in largest to smallest order, using ‘Custom Sort’.
- To gain a better understanding of the distribution of profits and identify any potential outliers, an X Y-Scatter Plot was created. The 'profit' column was plotted on the y-axis, while the 'budget' column was plotted on the x-axis. This visualization allowed us to observe any extreme values or anomalies in the data.
- During the analysis, several outliers were noticed in the data, which deviated significantly from the majority of movies in terms of profit. These outliers were labelled on the scatter plot for easy identification and further investigation.

movie_title	budget	gross	profit
Avatar	237000000	760505847	523505847
Jurassic World	150000000	652177271	502177271
Titanic	200000000	658672302	458672302
Star Wars: Episode IV - A New Hope	11000000	460935665	449935665
E.T. the Extra-Terrestrial	10500000	434949459	424449459
The Avengers	220000000	623279547	403279547
The Lion King	45000000	422783777	377783777
Star Wars: Episode I - The Phantom Menace	115000000	474544677	359544677
The Dark Knight	185000000	533316061	348316061
The Hunger Games	78000000	407999255	329999255
Deadpool	58000000	363024263	305024263
The Hunger Games: Catching Fire	130000000	424645577	294645577
Jurassic Park	63000000	356784000	293784000
Despicable Me 2	76000000	368049635	292049635
American Sniper	58800000	350123553	291323553
Finding Nemo	94000000	380838870	286838870
Shrek 2	150000000	436471036	286471036
The Lord of the Rings: The Return of the King	94000000	377019252	283019252
Star Wars: Episode VI - Return of the Jedi	32500000	309125409	276625409
Forrest Gump	55000000	329691196	274691196
Star Wars: Episode V - The Empire Strikes Back	18000000	290158751	272158751
Home Alone	18000000	285761243	267761243
Star Wars: Episode III - Revenge of the Sith	113000000	380262555	267262555
Spider-Man	139000000	403706375	264706375

3. Top 250 movies:

Part 1

To find the top 250 movies based on the IMDb score and number of user votes, following steps were taken:

- A column named ‘IMDb_top_250’ was created and the following function was inserted in it:
  =IF (num_voted_users > 25000, movie_title, “”)
  The function checked if the 'num_voted_users' value was greater than 25,000 and returned the movie title if it met this criterion; otherwise, it left the cell empty.
- To eliminate any empty cells in the 'IMDb_top_250' column, a filter was applied to identify and select the blank values. These rows were then deleted, ensuring that only the movies qualifying for the top 250 list remained in the dataset.
- The 'IMDb_top_250' column was sorted first based on the 'imdb_score' in descending order, from the highest to the lowest. Afterward, an additional level of sorting was applied using the 'Custom Sort' feature to arrange the movies in descending order of 'num_voted_users.
- To assign a rank to each movie in the top 250 list, a new column named 'rank' was introduced. The rank was determined by subtracting 1 from the row number of each entry, using the formula:
  = ROW([@IMDb_top_250]) – 1
- To finalize the top 250 movie list, all movies with a rank greater than 250 were deleted from the dataset. This step ensured that only the top 250 movies, based on IMDb score and user votes, remained in the final dataset.

imdb_top_250	num_voted_users	language	imdb_score	Rank
The Shawshank Redemption	1689764	English	9.3	1
The Godfather	1155770	English	9.2	2
The Dark Knight	1676169	English	9	3
The Godfather: Part II	790926	English	9	4
Pulp Fiction	1324680	English	8.9	5
The Lord of the Rings: The Return of the King	1215718	English	8.9	6
Schindler's List	865020	English	8.9	7
The Good, the Bad and the Ugly	503509	Italian	8.9	8
Inception	1468200	English	8.8	9
Fight Club	1347461	English	8.8	10
Forrest Gump	1251222	English	8.8	11
The Lord of the Rings: The Fellowship of the Ring	1238746	English	8.8	12
Star Wars: Episode V - The Empire Strikes Back	837759	English	8.8	13
The Matrix	1217752	English	8.7	14
The Lord of the Rings: The Two Towers	1100446	English	8.7	15
Star Wars: Episode IV - A New Hope	911097	English	8.7	16
Goodfellas	728685	English	8.7	17
One Flew Over the Cuckoo's Nest	680041	English	8.7	18
City of God	533200	Portuguese	8.7	19
Seven Samurai	229012	Japanese	8.7	20
Se7en	1023511	English	8.6	21
Interstellar	928227	English	8.6	22
The Silence of the Lambs	887467	English	8.6	23
Saving Private Ryan	881236	English	8.6	24
American History X	782437	English	8.6	25
The Usual Suspects	740918	English	8.6	26
Spirited Away	417971	Japanese	8.6	27
Modern Times	143086	English	8.6	28

Part 2

Building upon the resulting table from Part 1, the next task was to identify the top IMDb movies in foreign languages. The following steps were taken:

- Using the resulting table that we got from part 1 of this task, the filter was applied in ‘language’ column, unselecting everything, only ‘English’ was selected to show all the movies in English.
- Then those rows were deleted, and we were left with all the top IMDb movies in foreign language.

imdb_top_250	num_voted_users	languag	imdb_s
The Good, the Bad and the Ugly	503509	Italian	8.9
City of God	533200	Portugues	8.7
Seven Samurai	229012	Japanese	8.7
Spirited Away	417971	Japanese	8.6
The Lives of Others	259379	German	8.5
Children of Heaven	27882	Persian	8.5
AmÃ©lie	534262	French	8.4
Oldboy	356181	Korean	8.4
Princess Mononoke	221552	Japanese	8.4
Das Boot	168203	German	8.4
A Separation	151812	Persian	8.4
Baahubali: The Beginning	62756	Telugu	8.4
Downfall	248354	German	8.3
The Hunt	170155	Danish	8.3
Metropolis	111841	German	8.3
Pan's Labyrinth	467234	Spanish	8.2
Howl's Moving Castle	214091	Japanese	8.2
The Secret in Their Eyes	131831	Spanish	8.2
Incendies	80429	French	8.2

4. Best Directors:

To find the best directors based on average imdb_score, following steps were taken:

- A pivot table was inserted with column ‘director_name’ in Rows and ‘imdb_score’ in Values field in the PivotTable Fields.
- By default, the Values field displayed the ‘sum’ of IMDb scores which was changed to show ‘Average’ in ‘Value Field Settings’.
- With the directors, grouped based on their mean IMDb score, the pivot table was sorted in descending order of the average IMDb score.
- To organize the sorted table in ascending alphabetical order, the top 10 directors and their corresponding average IMDb scores were copied and pasted outside the pivot table. Using the 'Custom Sort' feature, an additional level of sorting was applied to arrange the directors' names in alphabetical order.

- Finally, a horizontal bar chart was plotted to visually represent the top 10 directors and their respective mean IMDb score.

Top_10_Directors	Avg _IMDb_sco
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
Marius A. Markevicius	8.4

5. Popular Genres:

The method used to find the popular genres was similar to the one used to find the top directors that is, based on the average IMDb score of each genre and steps taken were:

- Using a pivot table, the ‘genres’ column was inserted in Rows and ‘Average’ of imdb_score in Values.
- Then the result was sorted in largest to smallest order of average imdb_score.
- The top 10 genres were selected, and a horizontal bar chart was plotted. This chart type showcased the relative proportions of the movie counts for each genre, allowing for a quick and intuitive comparison.

Genres	Average of imdb_score
Crime|Drama|Fantasy|Mystery	8.5
Adventure|Animation|Drama|Family|Musical	8.5
Adventure|Animation|Fantasy	8.4
Action|Adventure|Drama|Fantasy|War	8.4
Adventure|Drama|Thriller|War	8.4
Documentary|Drama|Sport	8.3
Documentary|War	8.3
Biography|Drama|History|Music	8.3
Adventure|Animation|Comedy|Drama|Family|Fantasy	8.3
Adventure|Drama|War	8.25

6. **Critic-Favourite and Audience-Favourite Actors:**

Part 1

The first part of this task was to find out the critic-favourite and audience-favourite actors based on the average number of critic and audience reviews for actors – 'Meryl Streep', 'Leonardo DiCaprio', 'Brad Pitt'. The steps are mentioned below:

- Columns 'Meryl_Streep', 'Leo_Caprio', 'Brad_Pitt' were introduced and the following formula was used to find out the movies where they were the lead actors:
  =IF([@[actor_1_name]]="Meryl Streep", [@[movie_title]], "")
  =IF([@[actor_1_name]]="Leonardo Dicaprio", [@[movie_title]], "")
  =IF([@[actor_1_name]]="Brad Pitt", [@[movie_title]], "")
  The formula returned movies with the actors mentioned otherwise null.
- Then the three columns were merged into a new column named 'Combined' using the following formula:
  =CONCAT([@[Meryl_Streep]]&[@[Brad_Pitt]]&[@[Leo_Caprio]])
- Next the 'Combined' column was grouped using 'Custom Sort' on basis of 'actor_1_name' column in ascending order.
- A new column named 'Actor' was introduced to display the names of respective actors of the movies in the 'Combined' column.
- The movies with their respective lead actors i.e., 'Combined' and 'Actor' were highlighted using 'Conditional Formatting'.
- The resulting table was then used to create a pivot table with 'Actor' column in Rows and 'num_user_for_reviews' in Values and the setting was changed from 'Sum' to 'Average' to find out the 'Average user reviews' of the actors.
- Similarly, another pivot table was created, this time with 'num_critic_for_reviews' in the Value field and the setting was changed to 'Average' to find the 'Average Critic Reviews'.
- Horizontal Bar charts were plotted for both the pivot tables for visualisation of results.

Combined	Actor
The Assassination of Jesse James by the Coward Robert Ford	Brad Pitt
The Curious Case of Benjamin Button	Brad Pitt
The Tree of Life	Brad Pitt
Troy	Brad Pitt
True Romance	Brad Pitt
Blood Diamond	Leonardo DiCaprio
Body of Lies	Leonardo DiCaprio
Catch Me If You Can	Leonardo DiCaprio
Django Unchained	Leonardo DiCaprio
Gangs of New York	Leonardo DiCaprio
Inception	Leonardo DiCaprio
J. Edgar	Leonardo DiCaprio
Marvin's Room	Leonardo DiCaprio
Revolutionary Road	Leonardo DiCaprio
Romeo + Juliet	Leonardo DiCaprio
Shutter Island	Leonardo DiCaprio
The Aviator	Leonardo DiCaprio
The Beach	Leonardo DiCaprio
The Departed	Leonardo DiCaprio
The Great Gatsby	Leonardo DiCaprio
The Man in the Iron Mask	Leonardo DiCaprio
The Quick and the Dead	Leonardo DiCaprio
The Revenant	Leonardo DiCaprio
The Wolf of Wall Street	Leonardo DiCaprio
Titanic	Leonardo DiCaprio
A Prairie Home Companion	Meryl Streep
Hope Springs	Meryl Streep
It's Complicated	Meryl Streep
Julie & Julia	Meryl Streep

Actor	Average of num_critic_for_reviews
Leonardo DiCaprio	322.20
Brad Pitt	245
Meryl Streep	181.45

Actor	Average of num_user_for_reviews
Leonardo DiCaprio	922.55
Brad Pitt	742.35
Meryl Streep	297.18

Part 2

The next task was to observe the change in number of voted users over decades and was done as follows:

- A pivot table was created with ‘title_year’ as Rows and ‘sum’ of ‘num_voted_users’ in the Value field.
- Then the ‘title_year’ was grouped into decades starting from 1921 to 2020 and with that the Value field automatically updated.
- The table was copied, and the values were pasted outside the table. A new column ‘Decade’ was added to the new table and the respective decade was mentioned for each row.
- Finally, a horizontal bar chart was plotted to show the result and visualise the pattern and trend from the data.

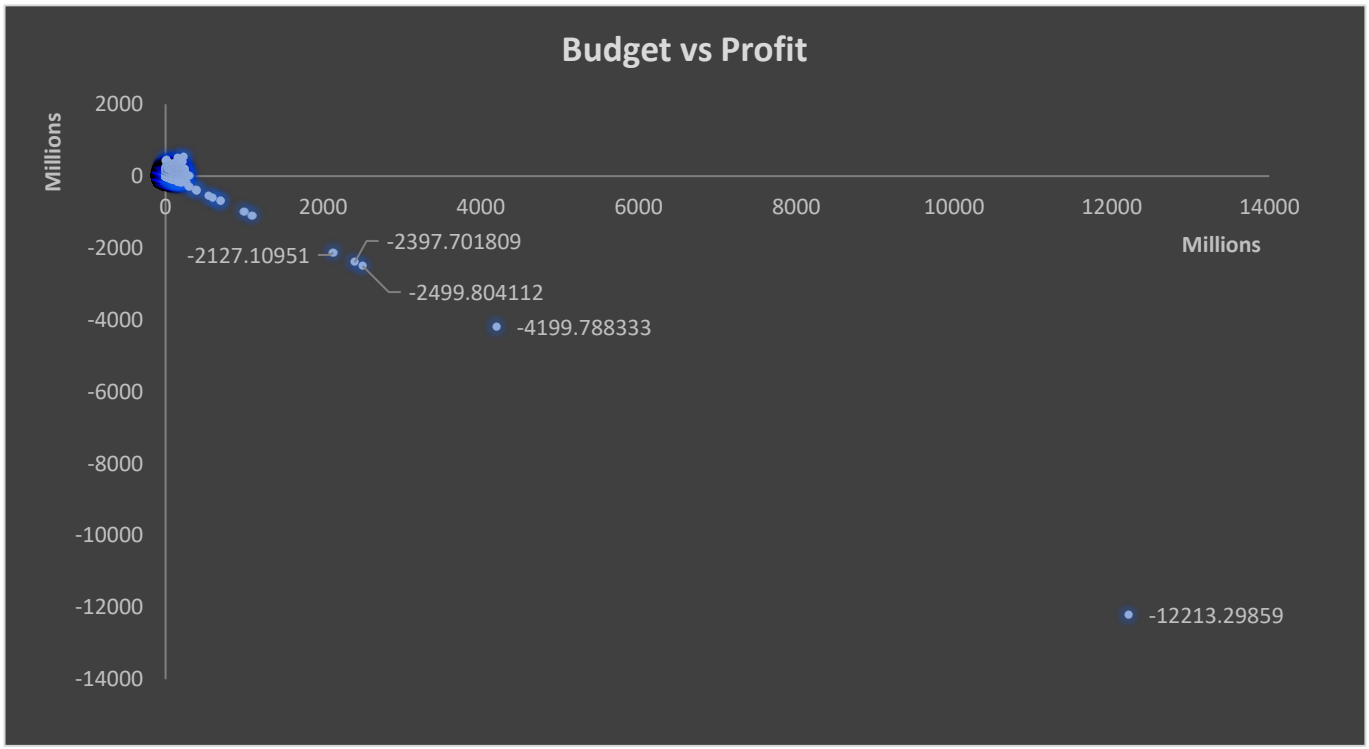
Decade	df_by_decade	Sum of num_voted_users
1920s	1920-1929	116392
1930s	1930-1939	804839
1940s	1940-1949	230838
1950s	1950-1959	678336
1960s	1960-1969	2983442
1970s	1970-1979	8269828
1980s	1980-1989	19344369
1990s	1990-1999	69635863
2000s	2000-2009	166045491
2010s	2010-2019	116240252

Insights:

The insights gathered from the project are mentioned below:

1. **Movies with Highest Profit:**

- The table showed that ‘Avatar’ was the most profiting movie with 523 million dollars of profit, followed by ‘Jurassic World’ with 502 million dollars of profit.



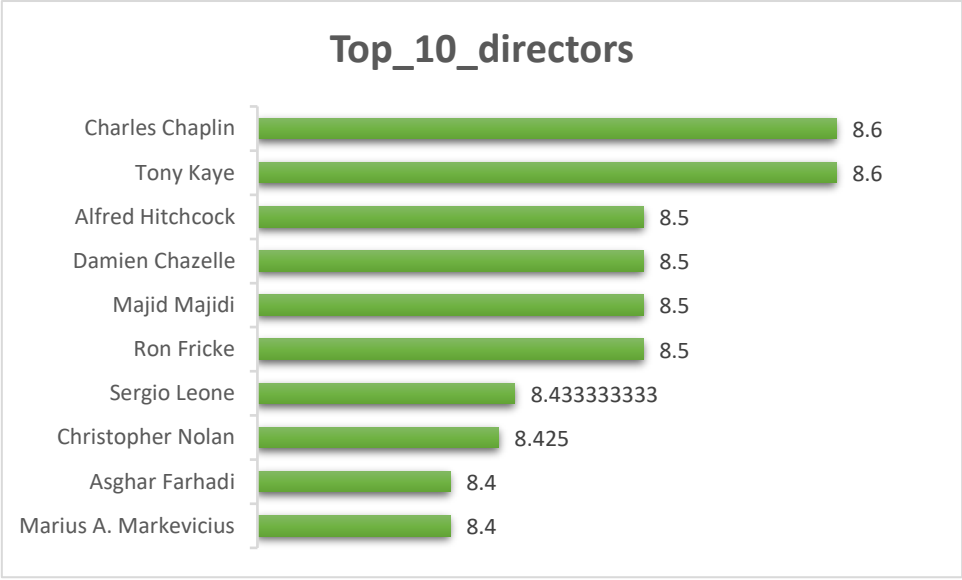
- The X Y Scatterplot of ‘Budget vs Profit’ showed that the data had various outliers, some of them being-

- -2127109510
- -2397701809
- -2499804112
- -4199788333
- -12213298588

2. Top 250 movies:

- The resulting table showed that 'Shawshank Redemption' was the top movie with an IMDb rating of 9.3 and 1689764 votes by users, followed by 'The Godfather' with IMDb score of 9.2 and 1155770 votes.
- Among the top 250 movies, 35 were in languages other than English.
- The top movie in non-English language was the Italian movie 'The Good, the Bad and the Ugly' with an IMDb score of 8.9 and 503509 votes, followed by the Portuguese movie 'City of God' with IMDb rating of 8.7 and 533200 votes.

3. Best Directors:

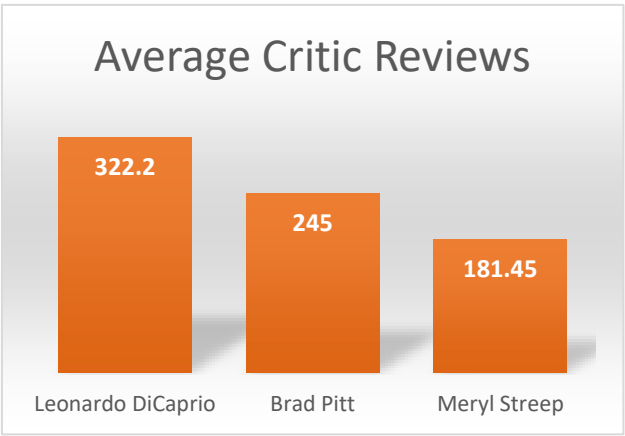


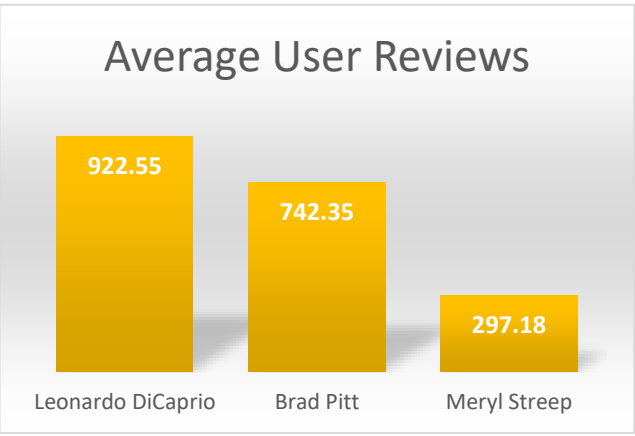
- The top 10 directors based on the average IMDb score of their movies revealed **'Charles Chaplin'** to be the top director with average IMDb rating of 8.6 along with **'Tony Kaye'**.

4. Popular Genres:

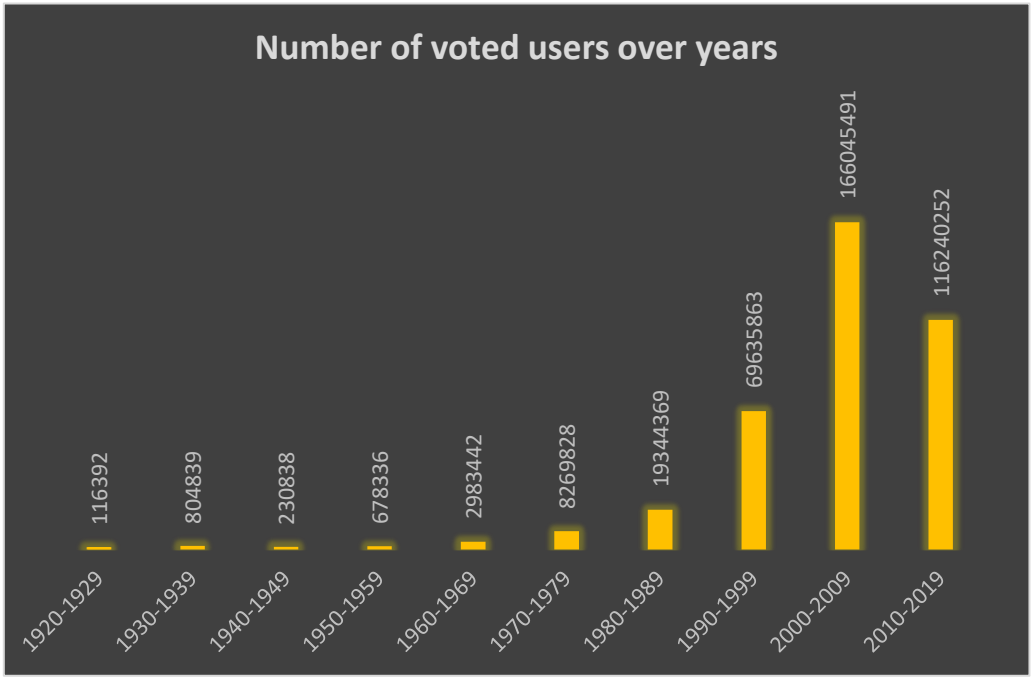
- The genre **Crime| Drama| Fantasy| Mystery** along with **Adventure| Animation| Drama| Family| Musical** had the highest average imdb_score of 8.5 among all the genres.

5. Critic-Favourite and Audience-Favourite Actors:





- The graphs for Average Critic Reviews and Average User Reviews showed that **Leonardo DiCaprio** was both the critic and audience favourite actor.



- The bar chart for the number of voted users over years showed that voted users number increased over the decades from 1920 till 2009 but showed decline after that.
- The highest number of voted users was reported during the 2000s decade while the lowest was in the 1920s decade.

Result:

- With this project I realised the importance of preparing the data for analysis before working with the dataset, such as removing the unnecessary columns to make the dataset compact and data processing efficient, removing duplicate and null values that could result in errors in the analysis and formatting the data with spell checks and removing special characters to make the data make sense.
- One of the most important learnings was to understand the demand of the task and figuring out the most efficient method to go about it.
- Using pivot tables to easily and efficiently do the calculations for us like grouping the data and perform various aggregation functions like sum, average, count etc.
- I learnt how powerful the filter and sort methods can be for data analysis.
- I utilised a lot of basic to advance Excel functions like nested functions to find the results.
- This real-time dataset allowed me to understand how scattered a dataset can get and to manipulate it to derive meaningful insights.

- By making the charts and graphs it was easy to understand what was happening with the data which could eventually lead to finding out the cause of the problem and a possible solution.
- Another insight that I gathered while doing this project was that there are multiple ways to go about finding the answer to a problem statement and they can all be meaningful and correct as long as they are justified.

Excel Sheet Link: ***IMDB_Movies_Analysis.xlsx***

Loom video explanation Link:
https://www.loom.com/share/6805921ed58b49eeb75cc1d95bddec44