IMDB Movie Analysis

Project Description: Using Data Analytics to calculate the movies with highest IMDB ratings and foreign language films with highest IMDB ratings, profit calculations, best directors, audience favourite and critic favourite actors and popular genres increase in number of users voting from the dataset provided on IMDB movies which contains the information about names of actors, director and their facebook likes count and also the information on films like their language, year of release, budget, IMDB and aspect ratios, country and content rating.

Tech-Stack Used: Microsoft Excel.

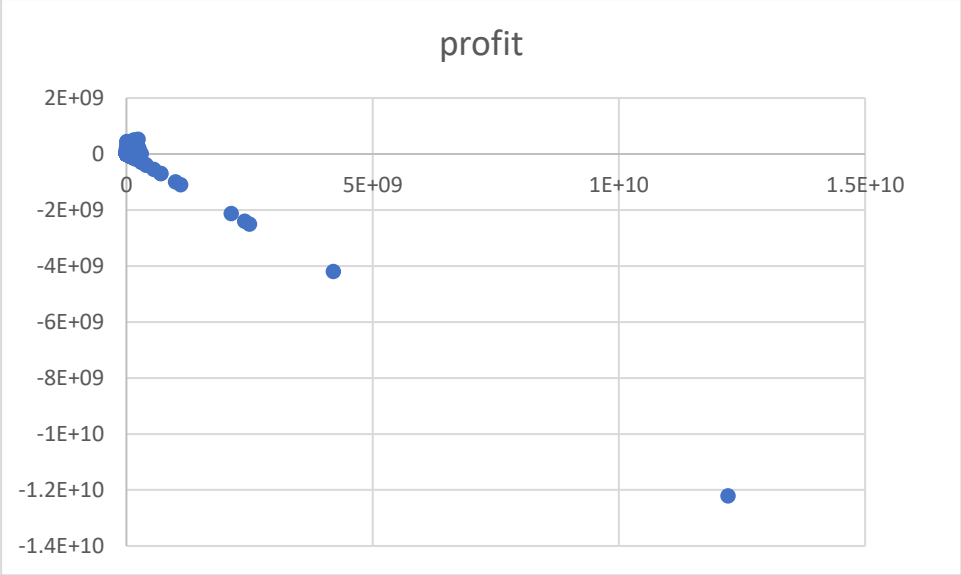
A. **Cleaning the data:** This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)
Your task: Clean the data

Answer:

- Removing the blank values from the columns and re arranging them
- Link: https://docs.google.com/spreadsheets/d/13ybJ2Bk1E8p7QtPz1p6PiKUfpA88eCDY/edit?usp=sharing&ouid=117379562832253850032&rtpof=true&sd=true

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?

Answer:



budget	profit	movie_title
237000000	523505847	AvatarÂ
150000000	502177271	Jurassic WorldÂ

- Movies with highest profit.
- Link: https://docs.google.com/spreadsheets/d/1vYGyPV0U_Vl_omMwVkIwsaHKNG1AV-bY/edit?usp=sharing&ouid=117379562832253850032&rtpof=true&sd=true

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

Answer:

- List of IMDB_Top_250 movies with num_voted_users >25000 and Top_Foreign_Lang_Film
- Link: https://docs.google.com/spreadsheets/d/1xgBgEV3aSvV6ujlEn9P0hUlP-6qR9evM/edit?usp=sharing&ouid=117379562832253850032&rtpof=true&sd=true

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

Row Labels	Average of imdb_score
Akira Kurosawa	8.7
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Majid Majidi	8.5
Damien Chazelle	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Richard Marquand	8.4
Grand Total	**8.47**

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

- Create new columns from the genre column thus separating the second and third genres for a film and then combine all the columns into a single column for the analysis.

Answer:

Row Labels	Count of genre
Drama	1896
Comedy	1462
Thriller	1116
Action	961
Romance	859
Adventure	785
Crime	708
Fantasy	509
Sci-Fi	497
Family	443

Horror	392
Mystery	385
Biography	240
Animation	197
Music	152
War	150
Sport	148
History	148
Musical	96
Western	59
Documentary	45
Film-Noir	1
Grand Total	**11249**

- Drama, comedy and thriller are the most popular genres.

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

Answer:

- Link: https://docs.google.com/spreadsheets/d/1T6VizbW982ONFVdnhysM7FB2KC9_VTzP/edit?usp=sharing&ouid=117379562832253850032&rtpof=true&sd=true
- Sheet1 contains the table

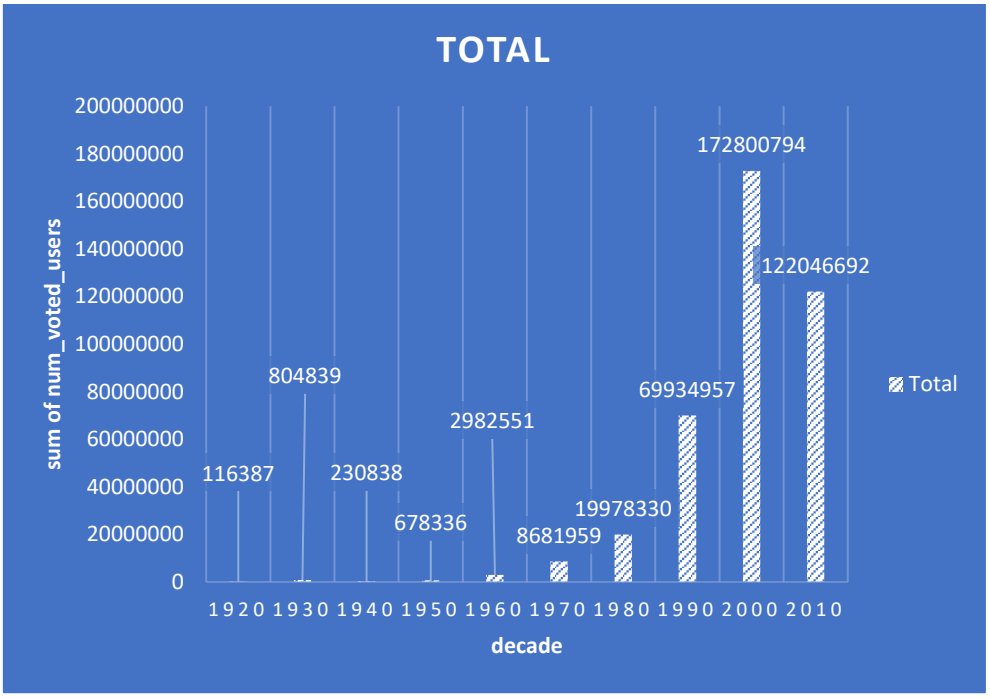
- Sheet2 contains the combined table of actors 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt'
- Sheet3 contains the movies grouped by actor names (if it's not visible in google sheets please open it in excel)
- Sheet4 contains the actor name, average of num_user_for_reviews and average of num_critic_for_reviews
- Sheet5 contains the decade, sum of num_voted_users and the bar graph indicating them.

- Critic favourite:

Row Labels	Average of num_user_for_reviews	Average of num_critic_for_reviews
Albert Finney	1498	750
Phaldut Sharma	1885	738

- Audience favourite:

Row Labels	Average of num_user_for_reviews	Average of num_critic_for_reviews
Heather Donahue	3400	360
Christo Jivkov	2814	406
Steve Bastoni	2789	275



Result:

The project helped me in getting a clarity of the concepts that I learned and in learning new concepts that were required for the project and it helped me understand how to do the IMDB analysis of large datasets.