Himanshu Jain
himjain567@gmail.com

IMDB Movie Analysis

Project Description:

The project is to analyse the movies trend, which movie is performing better, reason behind that, which movie is more popular, which actor is popular and its root cause. The data set given for this project consist of 28 columns and 5044 data points. Which contains actor names, IMDB rating, movie title, year, genres, director, reviews, language, budget, gross, etc. This also helps us to understand root cause analysis 'Five Whys' approach.

Approach:

First download the data and analyse it by checking for the null values in each row as well as in each column, duplicate rows and in which column data can be interpolated. Identified for each of the question which approach to be best. Which column is most useful to answer the questions and the null values removed from it. For each of the result the chart is the best approach to present it. So, I created chart for each of the result.

Tech-Stack Used:

To perform tasks for the project I have used Microsoft Excel (2016), Microsoft Word (2016).

Insights:

This assignment helps me to understand the functions in MS Excel and the pivot table. This gave me a complete idea of using the excel and the power of excel itself. Also, how a movie performance affect with the time, actor selection, director, reviews and IMDB rating.

Results:

A. **Cleaning the data:** This is one of the most important steps 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

The total number of datapoint before cleaning was 5043.

Step No.	Step	Rows Removed	Rows Left
1	Remove Duplicates	45	4998
2	Removes those rows which have more than 7 columns Null	19	4979
3	Remove rows where gross = Null	859	4120
4	Remove rows where budget = Null	263	3857
5	Replace 'Null' values in language column with 'English'	3857	3857
6	Removes those rows which have more than 3 columns Null	6	3851

Rows left after the cleaning is 3851 which is 76.3% of the original data.

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?

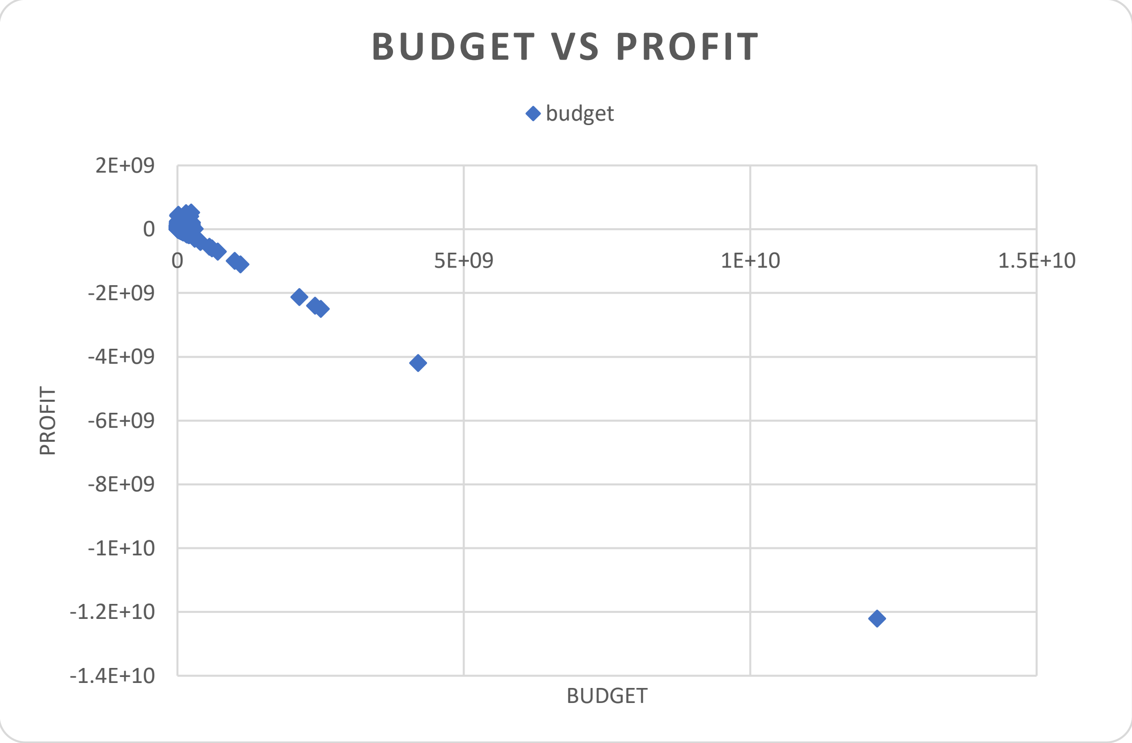

Fig. 1: Budget vs Profit with outliers.

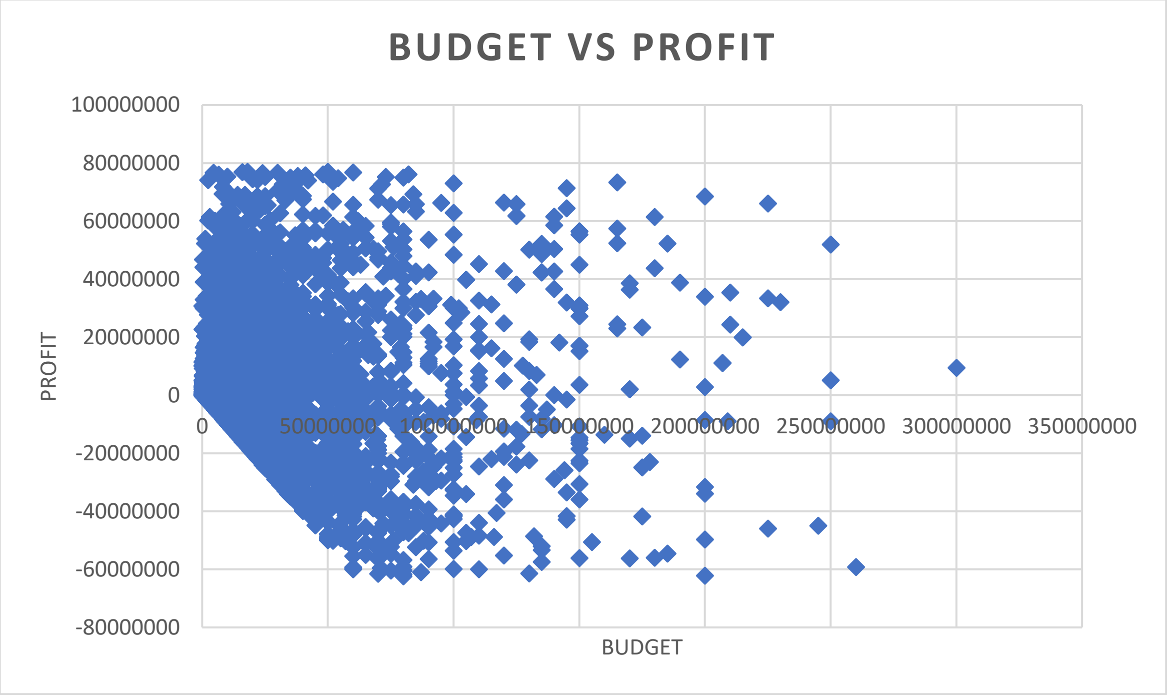

Fig. 2: Budget vs Profit without outliers.

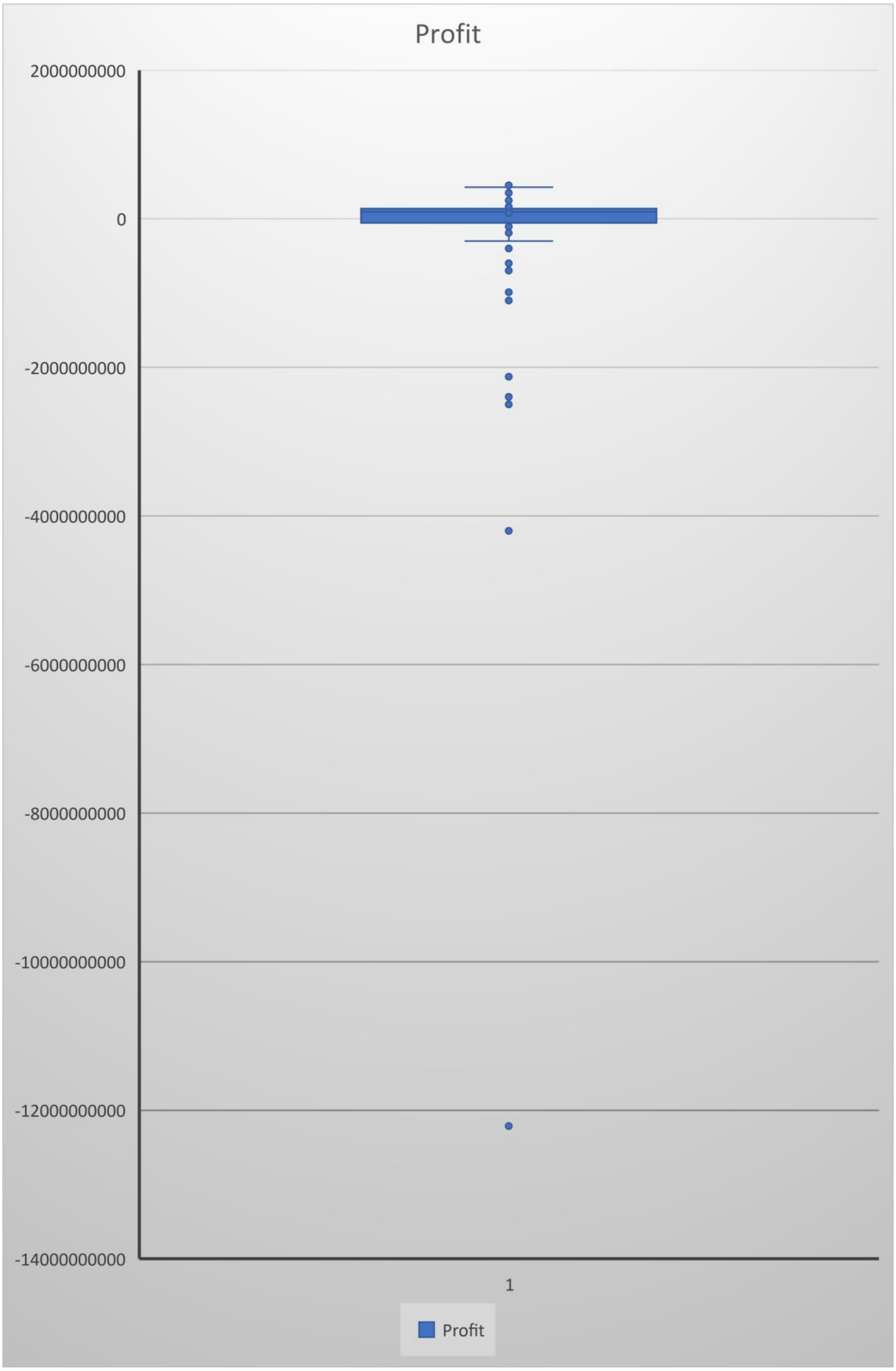

Fig. 3: Box & Whisker Plot with Outliers for profit.

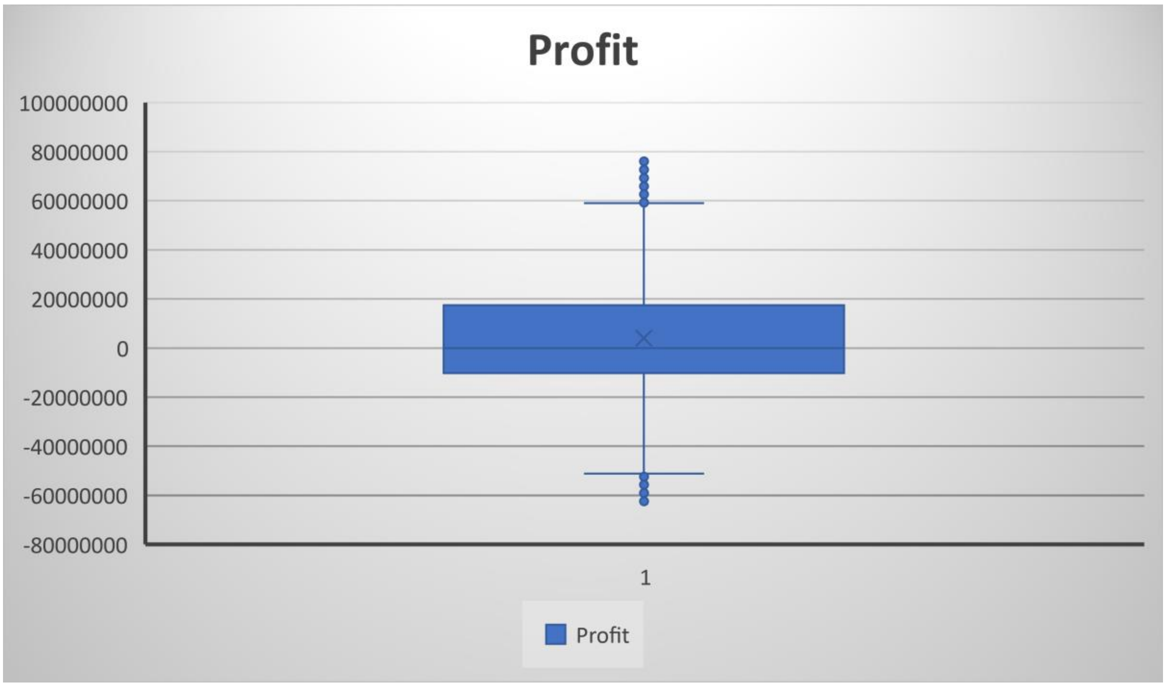

Fig. 4: Box & Whisker plot without Outliers for profit.

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 are 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 25

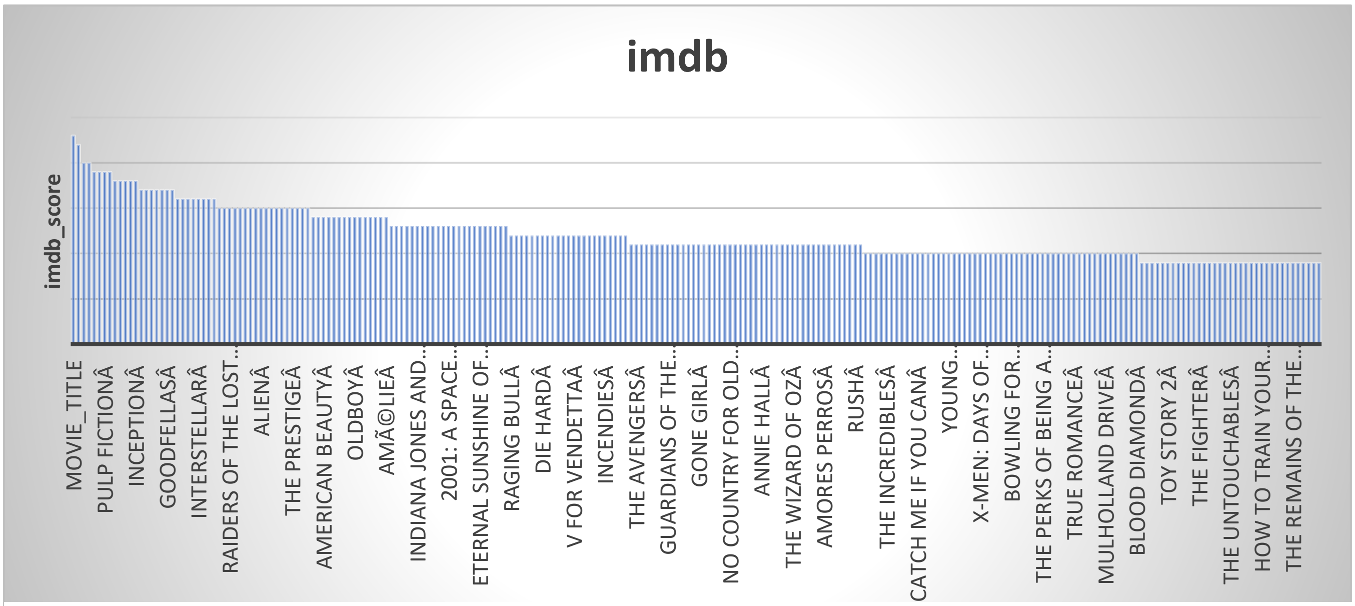

Fig. 5: Top 250 Movies

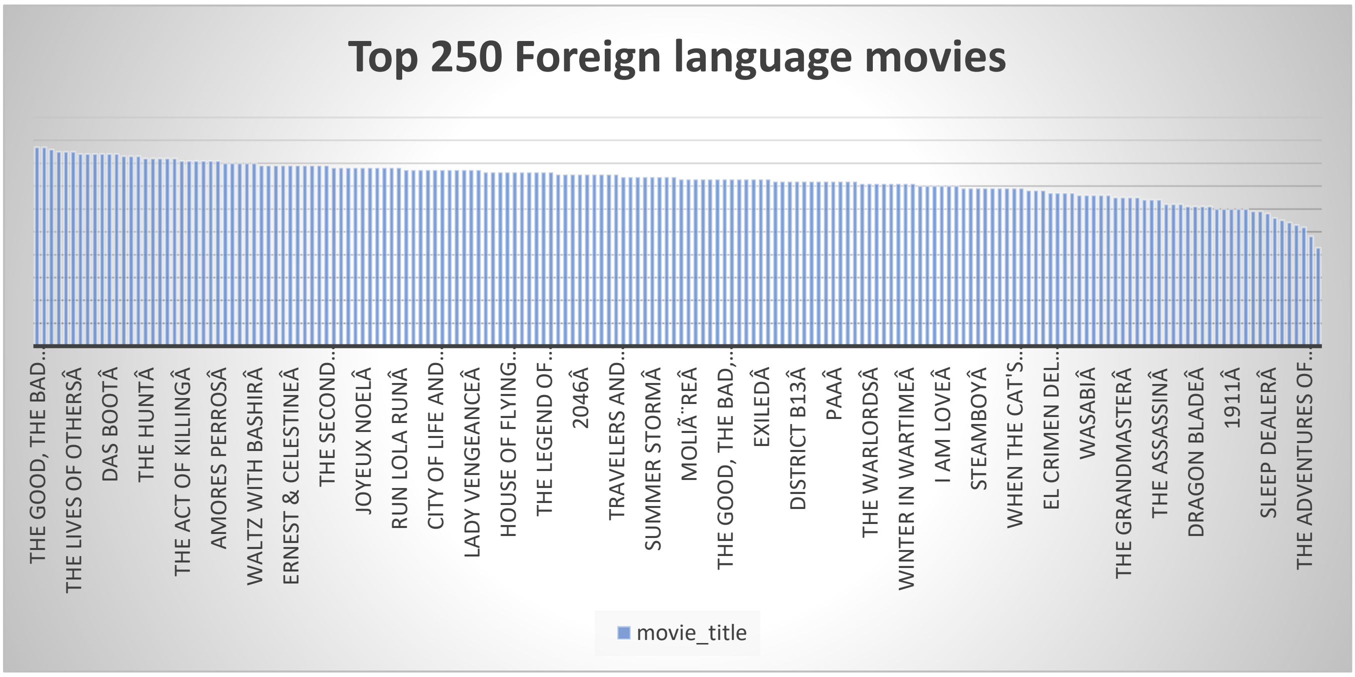

Fig. 6: Top 250 Foreign language movies.

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

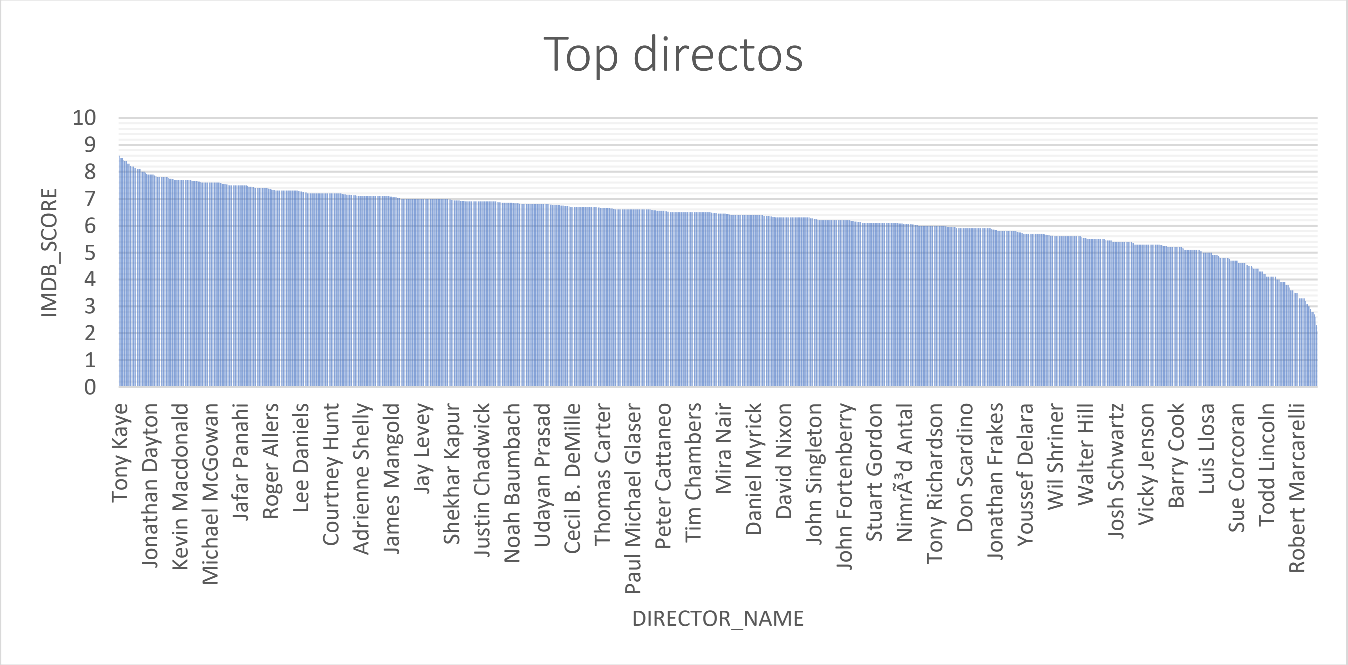

Fig. 7: Top director

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

Your task: Find popular genres

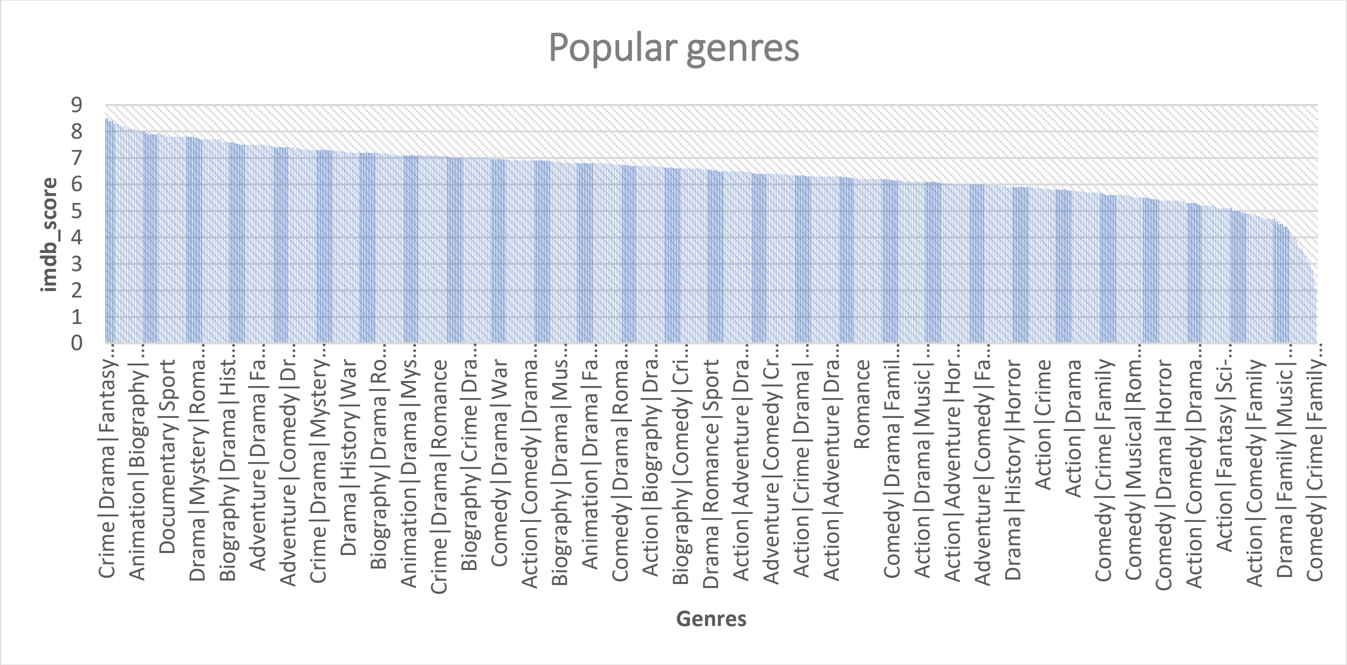

Fig. 8: Popular Genres.

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

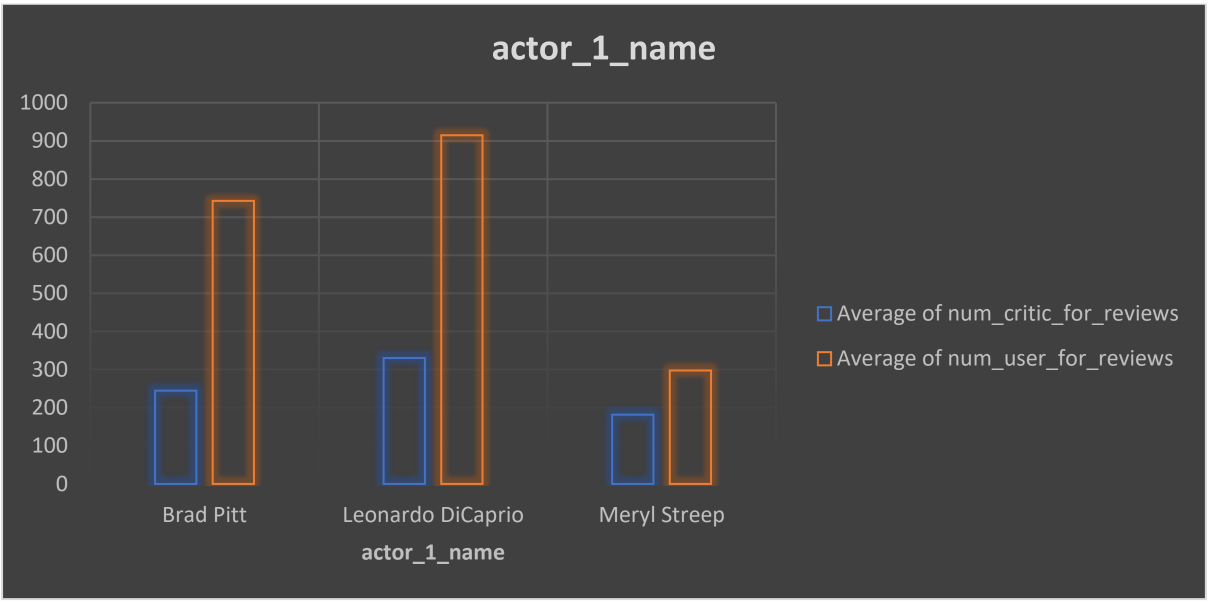

Fig. 9: actor_1_name vs num_critic_for_reviews and num_user_for_reviews.

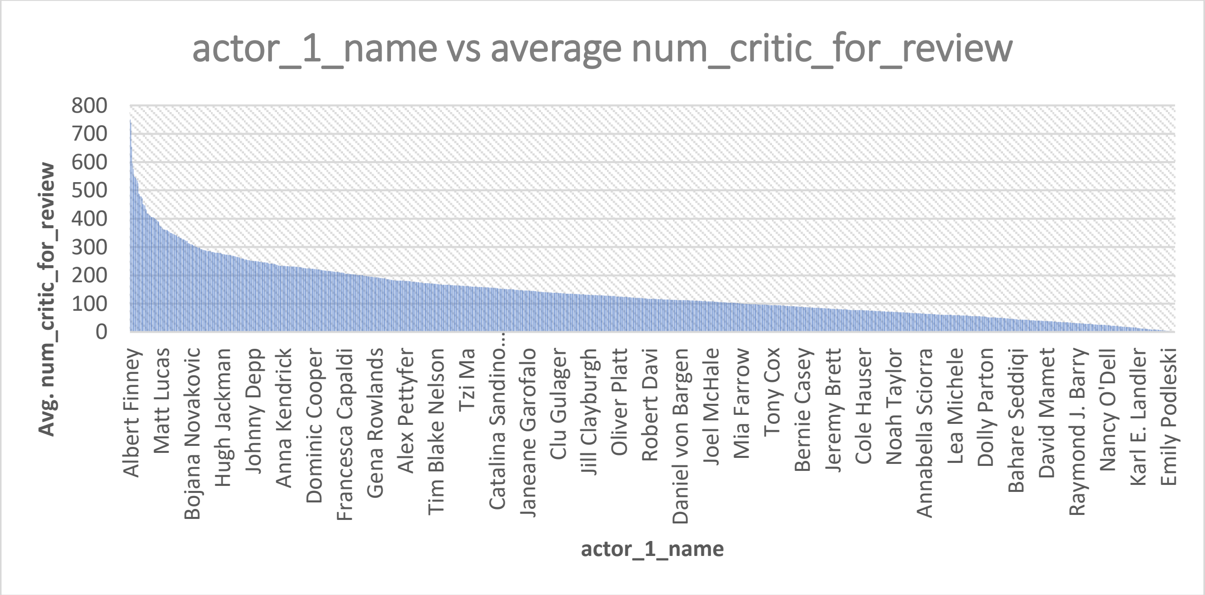

Fig. 10: actor_1_name vs avg, num_critic_for_review.

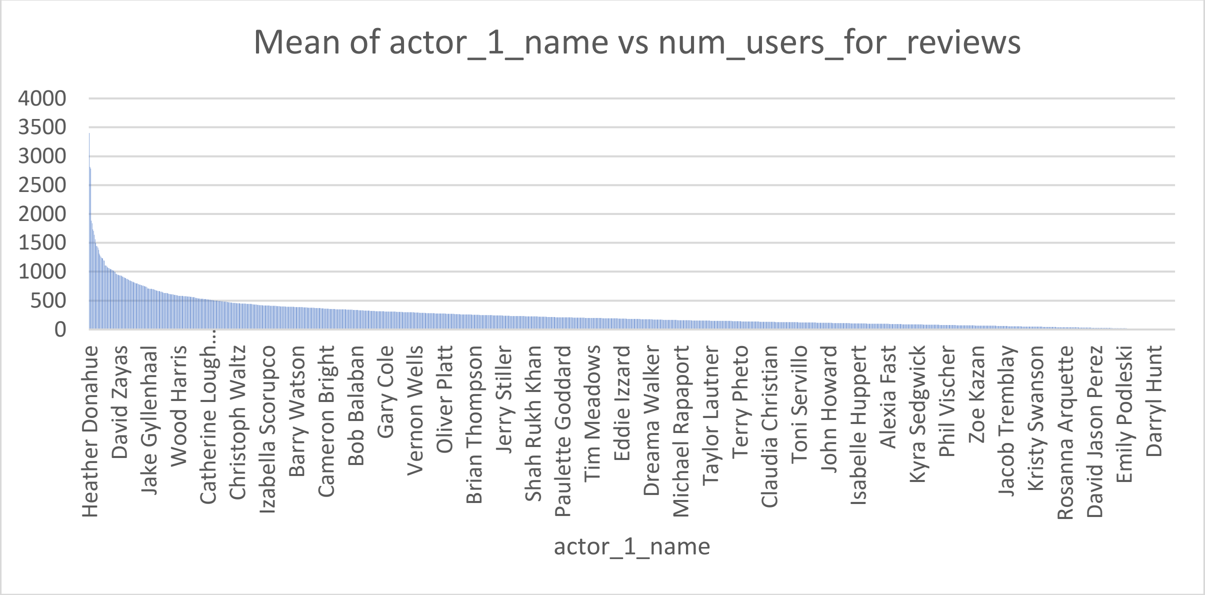

Fig. 11: actor_1_name vs num_critic_for_reviews

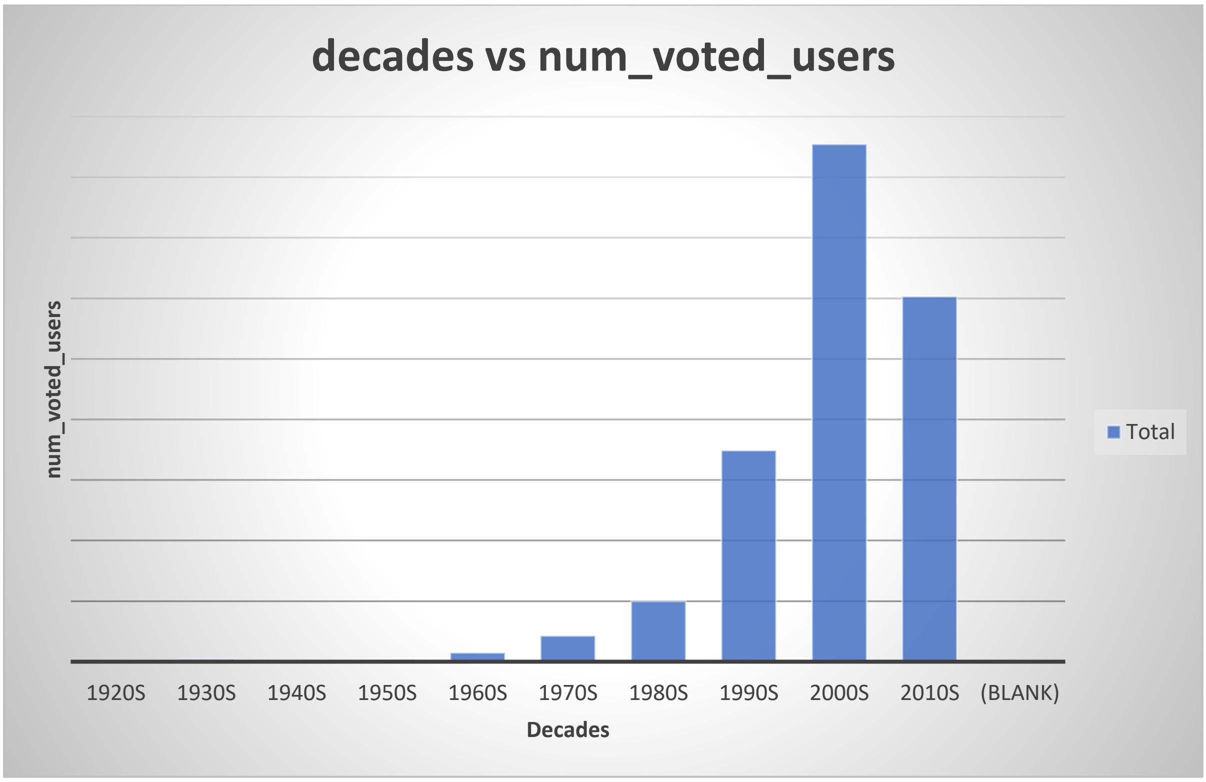

Fig. 12: Decades vs num_voted_users