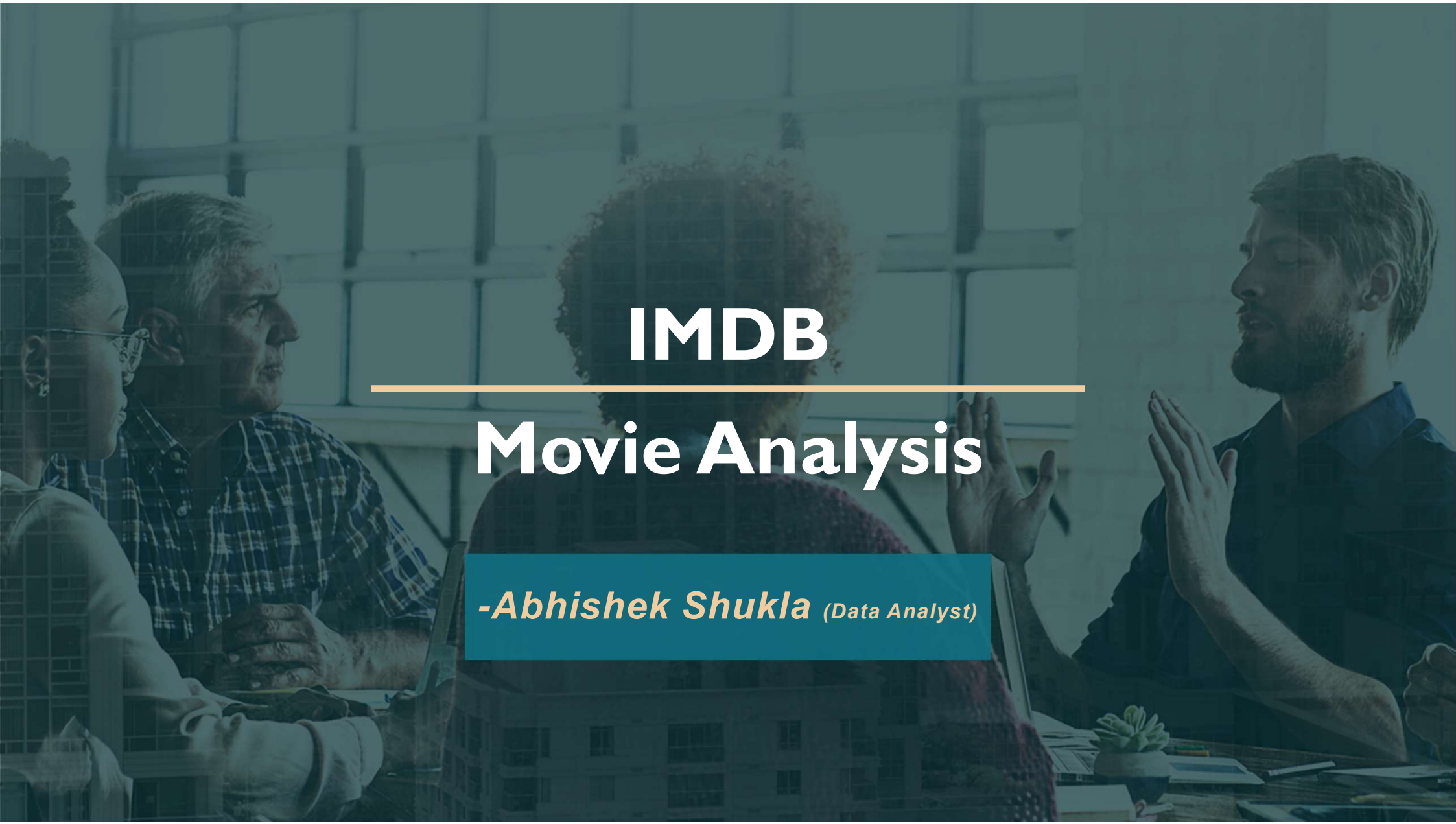

IMDB
Movie Analysis
-Abhishek Shukla (Data Analyst)

Project Description

The final project involves working with a dataset containing various columns of different IMDb movies. The project aims to frame a problem, clean the data, and derive insights from the dataset through data analysis. The following tasks need to be completed:

1. Cleaning the data: This step involves removing unnecessary columns, handling null values, and other data cleaning operations.

2. Movies with highest profit: Create a new column called "profit" by calculating the difference between the "gross" and "budget" columns. Sort the dataset based on the "profit" column and visualize outliers using an appropriate chart type.

3. Finding movies with the highest profit: Identify and report the movies with the highest profit based on the "profit" column.

4. Top 250: Create a new column called "IMDb_Top_250" and store the top 250 movies with the highest IMDb rating. Ensure that these movies have a "num_voted_users" value greater than 25,000. Add a "Rank" column to indicate the ranks of these movies.

5. Extracting non-English movies from the IMDb Top 250: Identify non-English movies from the "IMDb_Top_250" column and store them in a new column called "Top_Foreign_Lang_Film."

6. Finding the best directors: Group the dataset by the "director_name" column. Determine the top 10 directors with the highest mean IMDb score. In case of a tie, sort the directors alphabetically.

7. Finding popular genres: Analyze the dataset to identify popular genres based on the movies' characteristics.

8. Creating actor-specific columns: Create three new columns named "Meryl_Streep," "Leo_Caprio," and "Brad_Pitt," which contain movies where the actors "Meryl Streep," "Leonardo DiCaprio," and "Brad Pitt" are the lead actors, respectively. Group the combined column by the "actor_1_name" column.

9. Identifying critic-favorite and audience-favorite actors: Calculate the mean of the "num_critic_for_reviews" and "num_users_for_review" columns and identify the actors with the highest mean values.

10. Analyzing the change in number of voted users over decades: Create a column called "decade" to represent the decade to which each movie belongs. Sort the dataset based on the "decade" column, group it by decade, and find the sum of users voted in each decade. Store the results in a new data frame called "df_by_decade."

The project requires thorough data cleaning, analysis, and visualization techniques. By addressing the defined problem and answering the provided questions, a comprehensive report should be created to convey the data story discovered during the analysis.

IMDb

Approach

- ✓ Review the dataset and understand the different columns and their meanings.
- ✓ Use the provided guiding questions to frame the problem you want to shed light on.
- ✓ Consider what you see happening, hypothesis for the cause of the problem, and the impact of the problem on stakeholders.
- ✓ Define the problem you are trying to solve and the data required to address it.
- ✓ Compile the findings from the analysis into a detailed report.
- ✓ Include the problem statement, data cleaning process, insights derived from each analysis task, and visualizations.
- ✓ Use the 5 Whys technique to dig deeper into the problem and find the root cause.
- ✓ Present the data story in a clear and concise manner, making it easy for stakeholders to understand.

Tech-Stack Used

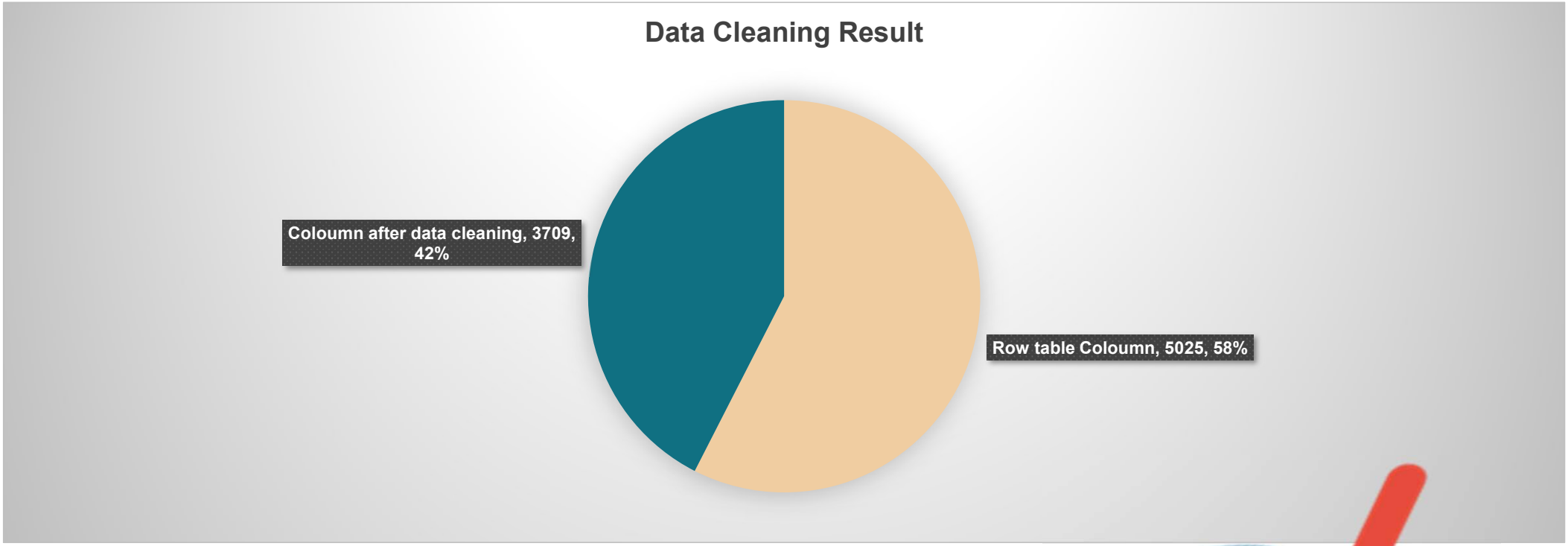
- 1. Microsoft Excel 365: Excel is the primary software used for data manipulation, analysis, and visualization in this project. It provides a familiar interface and a wide range of functionalities for working with tabular data.
- 2. Data Cleaning in Excel: Excel provides various features and functions to clean and preprocess the dataset. These include sorting, filtering, removing duplicates, filling missing values, and correcting data formats.
- 3. Formulas and Functions: Excel's built-in formulas and functions are utilized for data transformations, calculations, and aggregations. Functions like IF, SUM, AVERAGE, VLOOKUP, and CONCATENATE are commonly used to perform complex operations on the data.
- 4. PivotTables: PivotTables in Excel are powerful tools for summarizing and analyzing data. They allow for easy grouping, filtering, and aggregating data based on different criteria. PivotTables are useful for tasks like finding highest profits, top directors, and popular genres.
- 5. Data Visualization in Excel: Excel offers various chart types and customization options to create visually appealing graphs and charts. These charts, such as column charts, bar charts, and line charts, can be used to represent insights derived from the dataset.
- 6. Data Analysis Tools: Excel includes built-in data analysis tools such as Descriptive Statistics, Regression Analysis, and ANOVA. These tools enable advanced statistical analysis and help derive meaningful insights from the data.
- 7. Conditional Formatting: Excel's conditional formatting feature allows for visual highlighting of data based on specified conditions. It can be utilized to identify outliers, trends, or specific patterns in the dataset.
- 8. Data Reporting: Excel provides the capability to create well-structured reports by combining data, charts, and text. Worksheets, cells, and formatting features can be used to present the analysis process and findings in a clear and organized manner.
- Excel, with its extensive functionalities and user-friendly interface, serves as an effective tool for data cleaning, analysis, visualization, and reporting in this project.

Task A : Cleaning the data

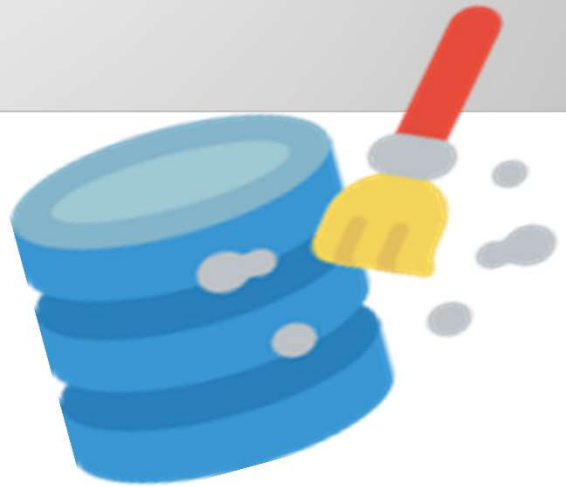
Cleaning the data :: This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Your task: Clean the data

Task A : Cleaning the data



“In Excel, various features and functions are available to clean and preprocess datasets. These include sorting, filtering, removing duplicates, filling missing values, and correcting data formats. After applying these steps, I was able to obtain 3709 out of 5025 records, which will serve as the starting point for analysis.”

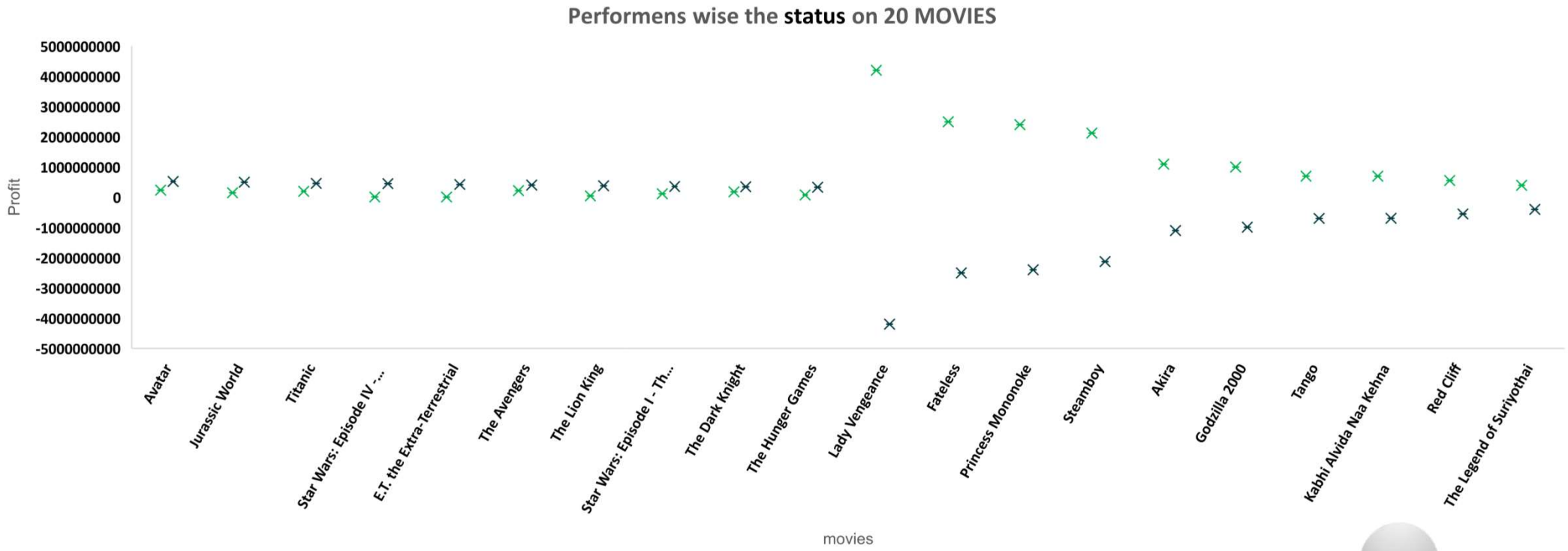


Task B : Movies with highest profit

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.

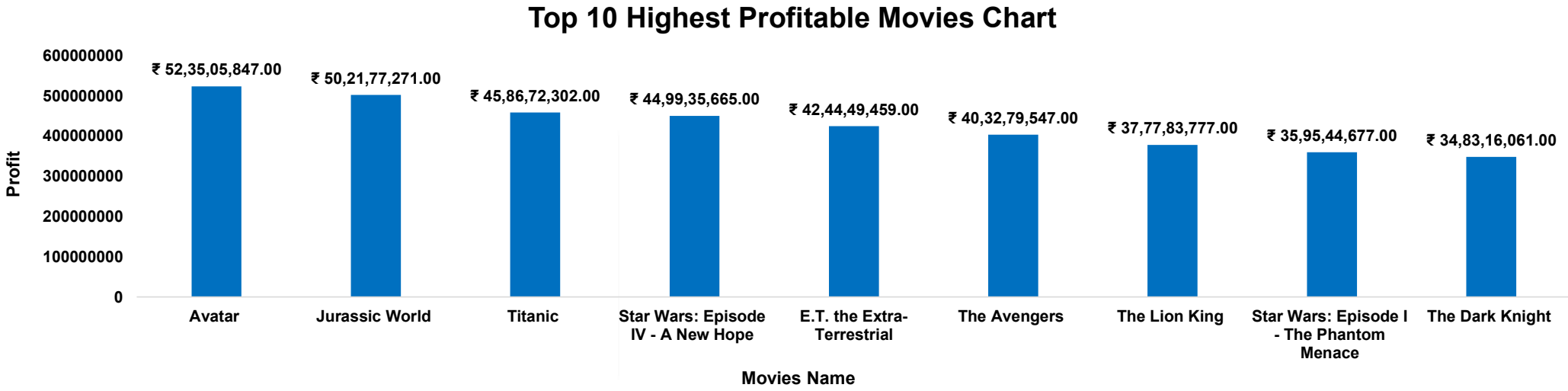
Your task: Find the movies with the highest profit?

Task B : Profit – Budget Analysis and Observing Outliers



“(Plot profit (y-axis) vs budget (x- axis) and observe the outliers
using the appropriate chart type)”

Task B :Top 10 Highest Profitable Movies Chart



"These are the top 10 most profitable movie names and their corresponding profits, as shown in the above chart.

Note : Avatar is the movie with highest profit Rs.523505847, Its budget is Rs.237000000 and its gross income is Rs.760505847."

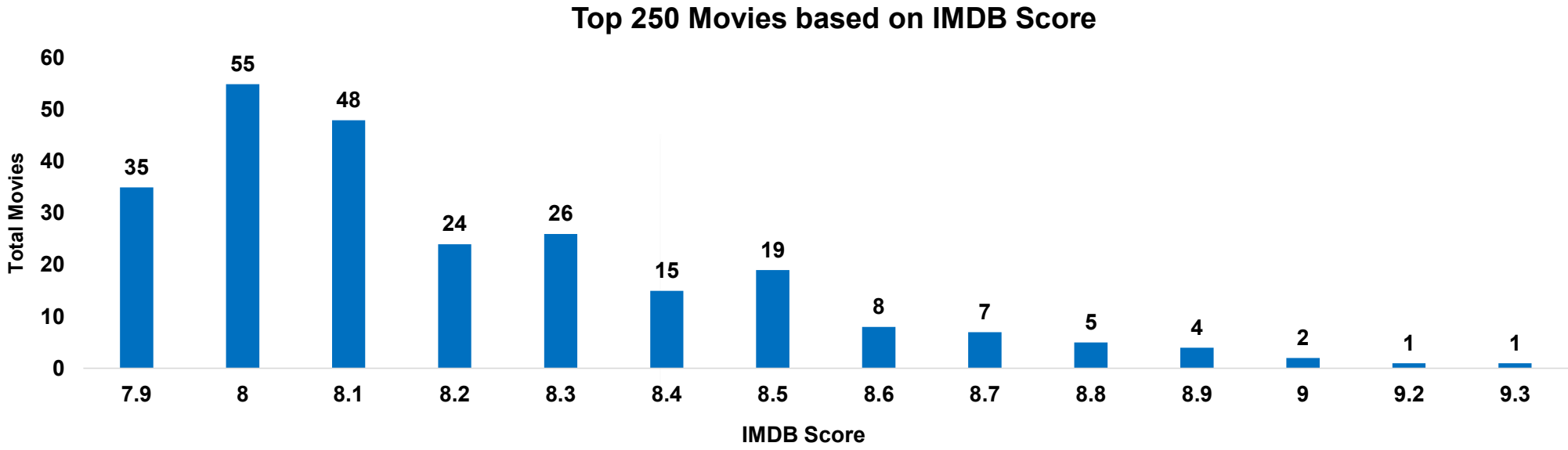
Task C :Top 250

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!

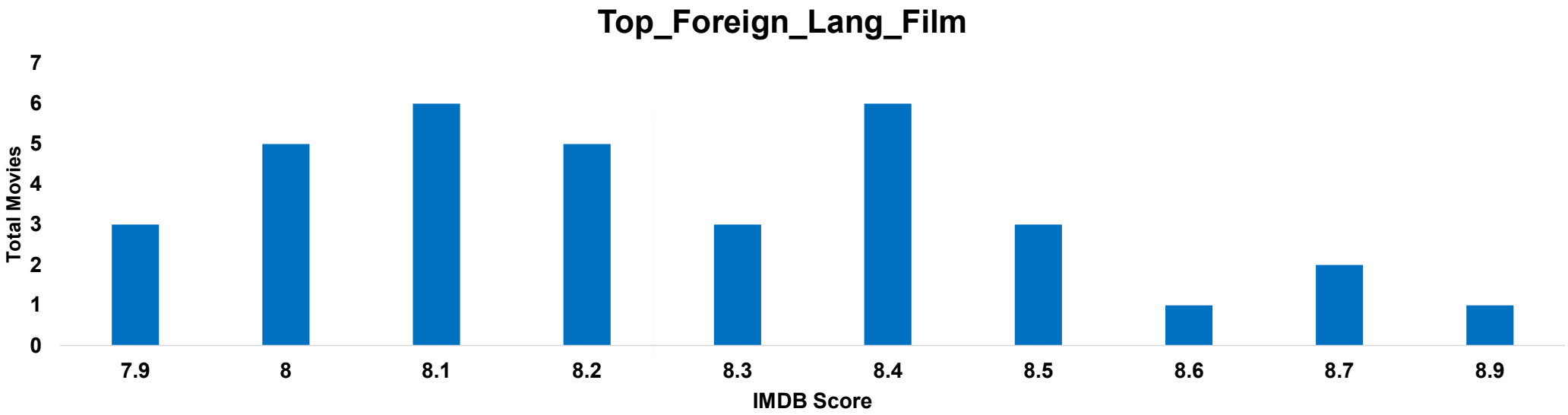
Your task: Find IMDB Top 250

Task C : Top 250



"These are the top 250 movies with more than 25,000 votes, along with the IMDB score."

Task C : Top_Foreign_Lang_Film



“After extracting all the movies in the IMDb_Top_250 column that are not in the English language, I obtained a total of 35 movies. Along with the IMDb score.”

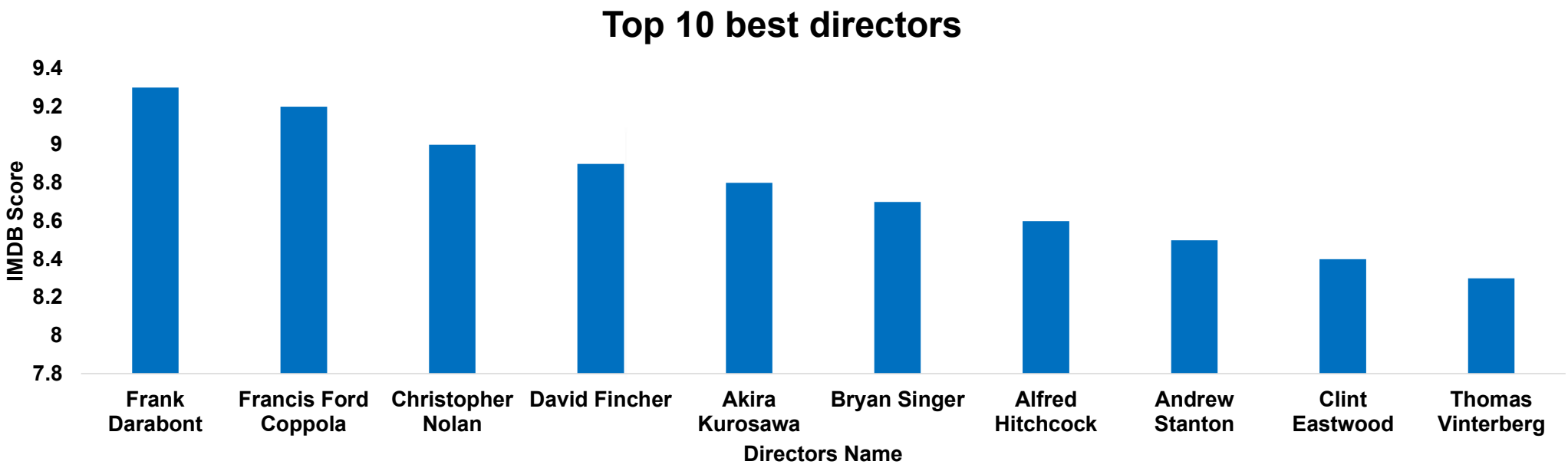
Task D : Best Directors

Best Directors : Group 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

Task D : Best Directors



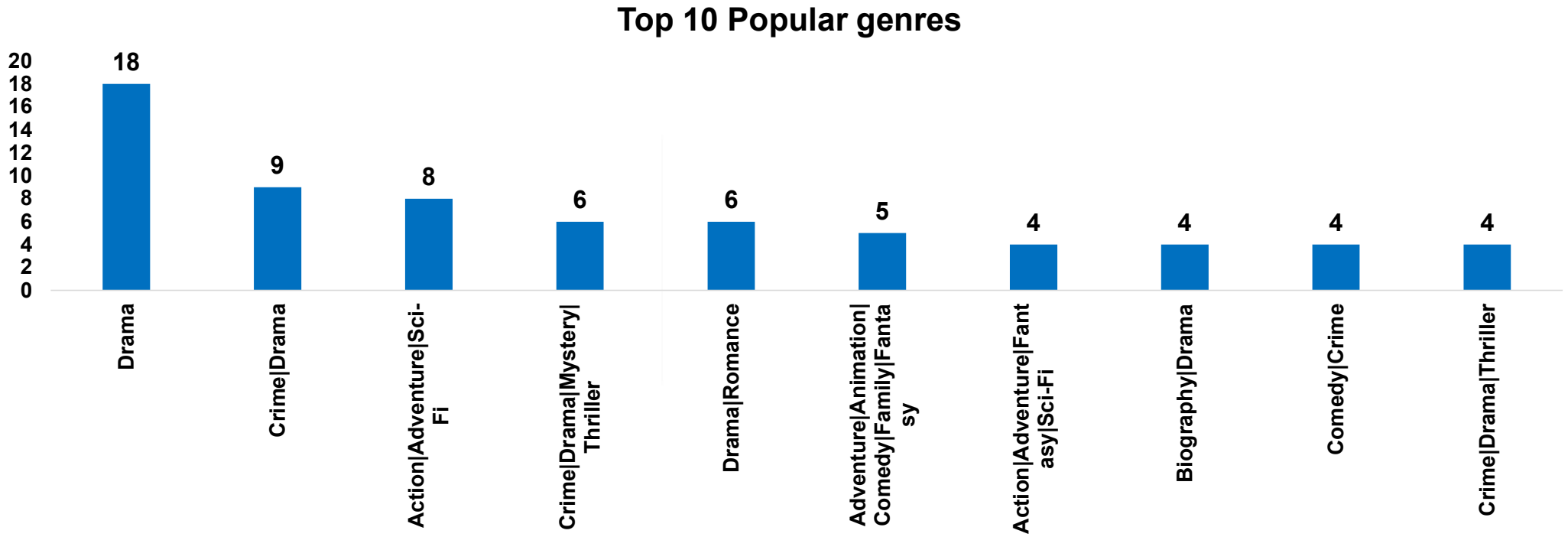
"These are the top 10 directors shown in the above chart."

Task E : Popular Genres

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

Your task : Find popular genres

Task D : Popular Genres



"These are the top 10 popular genres shown in the above chart, with Drama being the most popular genre according to the number of movies and the max number of votes 7330380."

Task F : Charts

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.

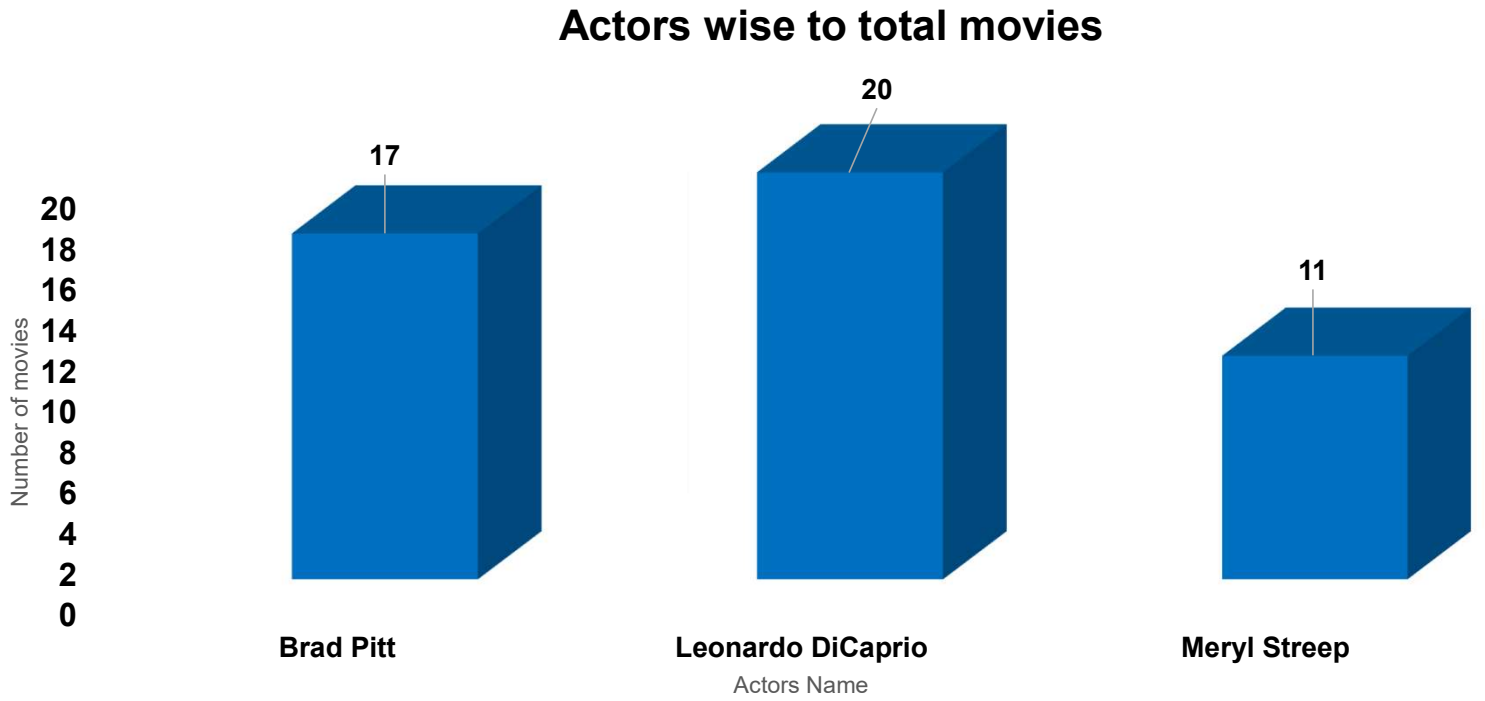
Group the combined column using the actor_1_name column.

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.

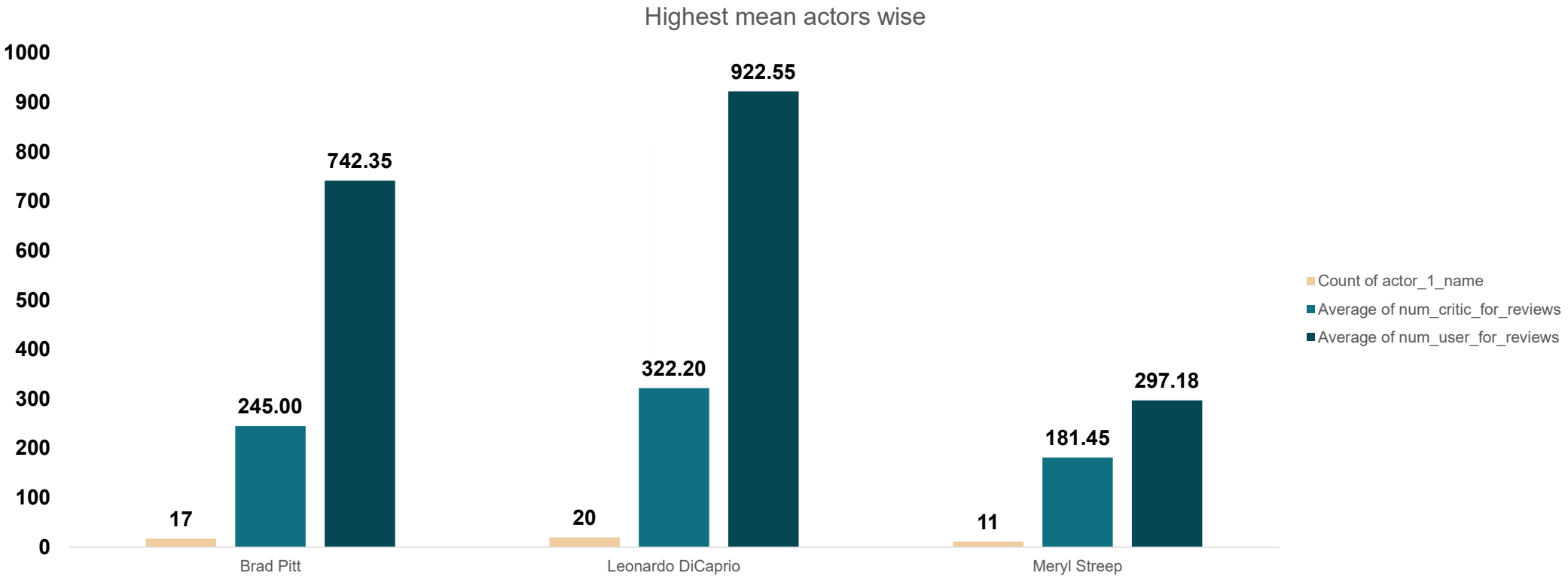
Your task: Find the critic-favorite and audience-favorite actors

Task F : Actors wise to total movies



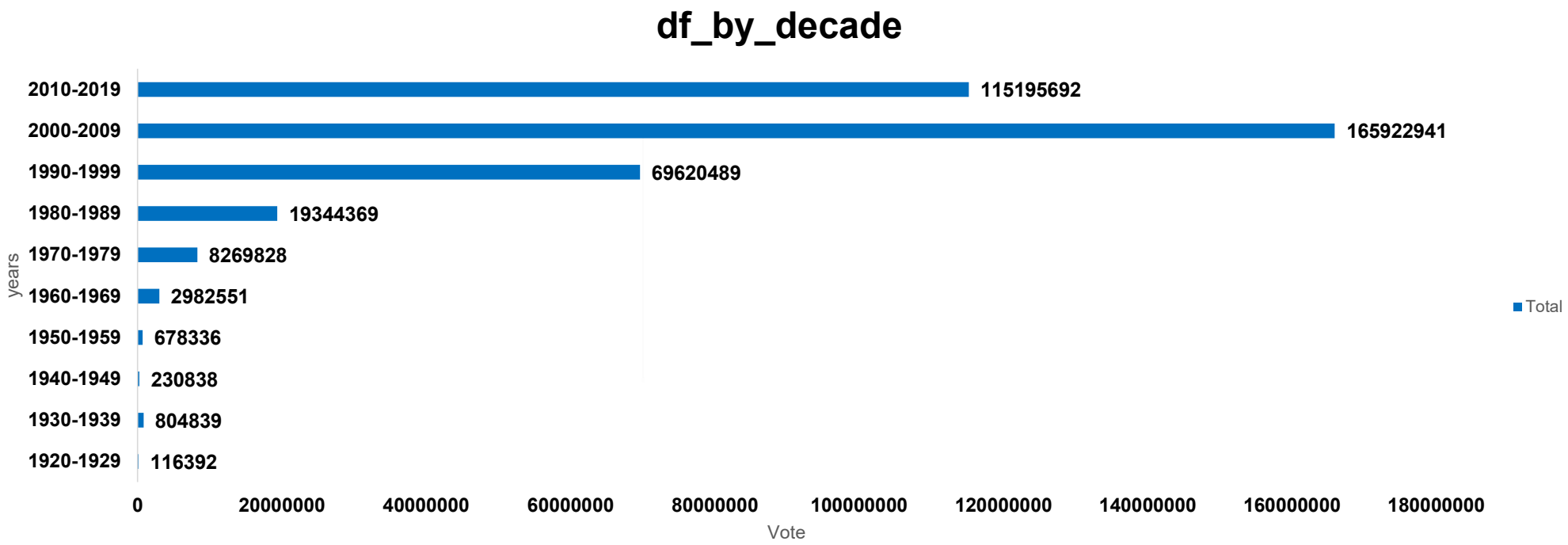
“These are the actors with total movies shown in the above chart.”

Task F : Actors wise highest mean



“Based on the analysis, it was found that Leonardo DiCaprio is the winner as per the critic reviews and users reviews.”

Task F : Decade wise Votes



“Based on the analysis, it was found that the years 2000 to 2009 had the maximum number of voters obtained, while the years 1920 to 1929 had the minimum number of voters obtained.”

IMDb

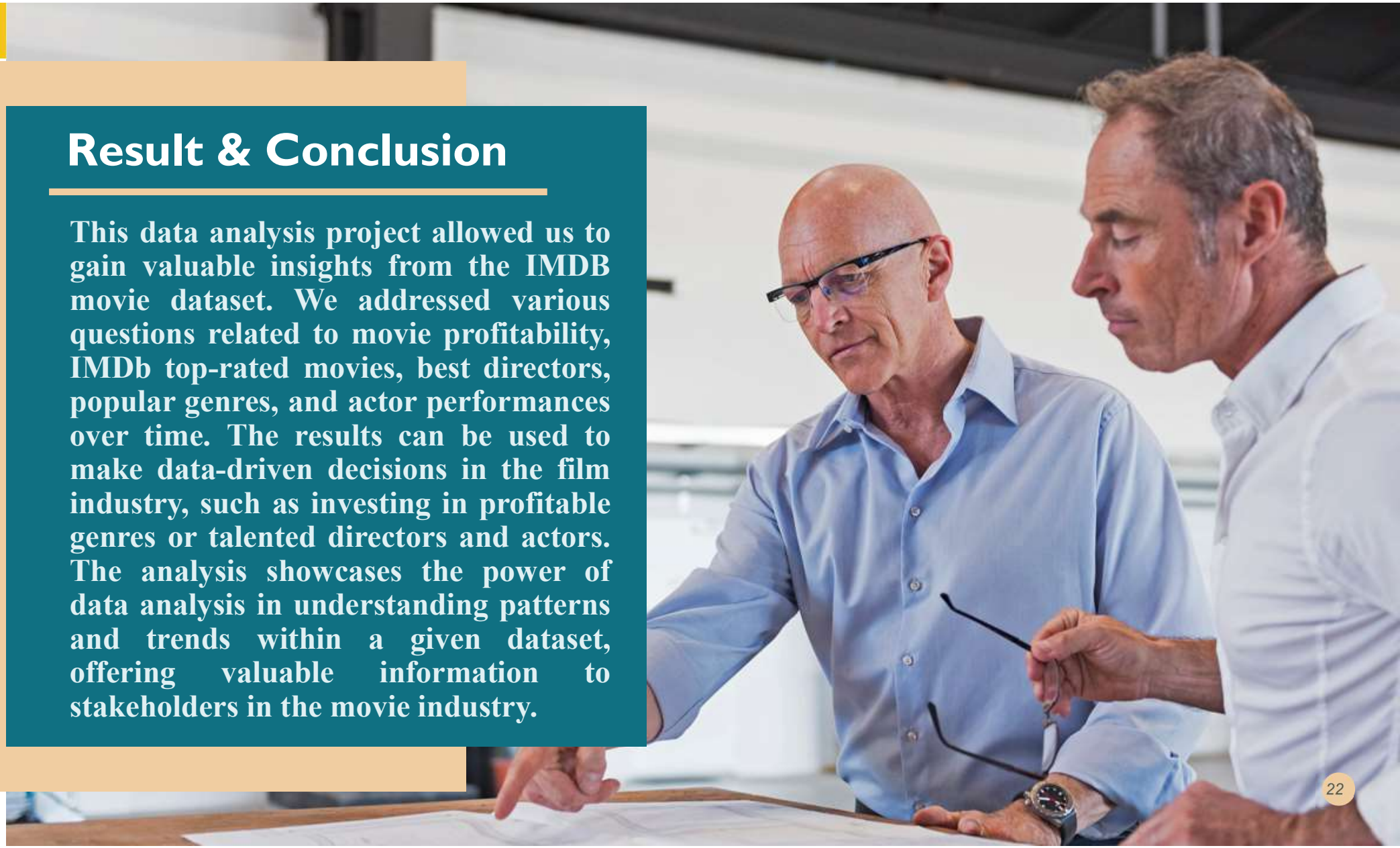
Task F : most favourite actor critic and audience wise

“Johnny depp is the most favourite actor critic and audience wise.”

IMDb

Result & Conclusion

This data analysis project allowed us to gain valuable insights from the IMDB movie dataset. We addressed various questions related to movie profitability, IMDb top-rated movies, best directors, popular genres, and actor performances over time. The results can be used to make data-driven decisions in the film industry, such as investing in profitable genres or talented directors and actors. The analysis showcases the power of data analysis in understanding patterns and trends within a given dataset, offering valuable information to stakeholders in the movie industry.



Thank You !