IMDB Movie Analysis

Final Project-1

Project Description:

The project aims to analyze a dataset containing information about various movies from the IMDB database. The goal is to derive insights and answer specific questions related to the data. The analysis will involve cleaning the data, identifying movies with the highest profit, determining the top 250 movies based on IMDb ratings, finding the best directors, identifying popular genres, exploring the movies featuring specific actors, and analyzing the change in the number of voted users over decades.

Approach:

The project will involve cleaning the dataset by removing irrelevant columns and handling missing values. Exploratory data analysis techniques will be employed to gain insights from the data. This will include visualizations, statistical analysis, and grouping data based on relevant criteria. The analysis will be performed using Excel or Google Sheets.

Tech-Stack Used:

The project will be conducted using Excel or Google Sheets for data analysis and visualization. These tools provide the necessary functionalities for data manipulation, calculations, and creating visualizations.

Insights:

Throughout the analysis, several insights can be derived. These may include identifying movies with the highest profit, understanding the characteristics of the top-rated movies, recognizing the best directors based on IMDb scores, discovering popular genres among the movies, and examining the impact of actors on the critics and audience.

1. **Cleaning the data::** PThis 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.)
   task: Clean the data —
   1. drop unnecessary columns
   2. Remove Blank Cell / Null Value
   3. Removing Duplicates.

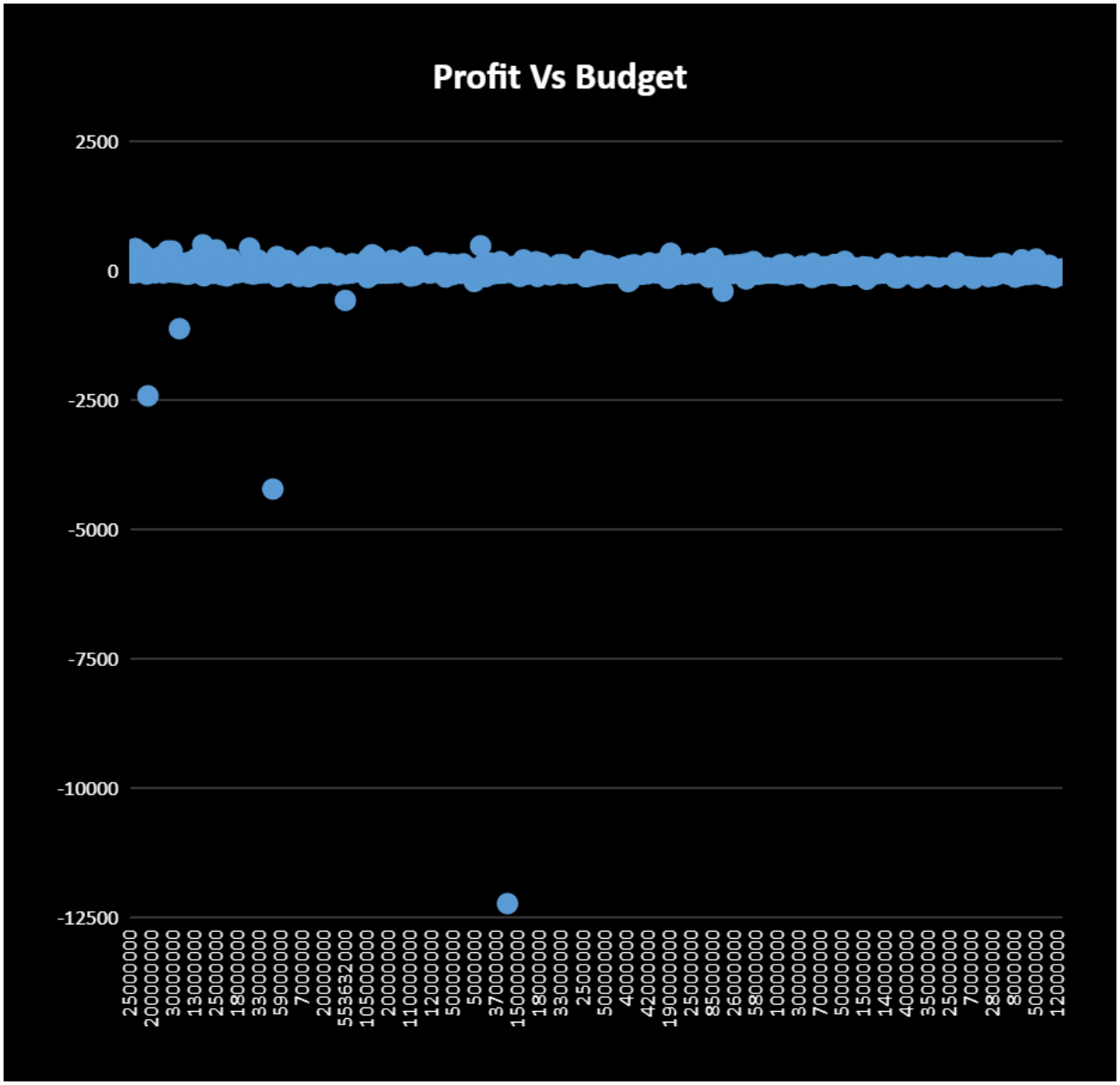
      Hence,columns to be removed for above 3 conditions are as follows :

      (Color, director_facebook_likes, actor_3_facebook_likes,actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_posts, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes)

2. **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?

For this, we will make a scatterPlot

Profit vs budget



The Outliers are : -12213298588 -4199788333 -2499804112 -2397701809 -2127109510

Top 5 Profitable Movies :

director_name	actor_1_name	movie_title	title_year	imdb_score	Profit
James Cameron	CCH Pounder	Avatar	2009	7.9	523505847
Colin Trevorrow	Bryce Dallas Howard	Jurassic World	2015	7	502177271
James Cameron	Leonardo DiCaprio	Titanic	1997	7.7	458672302
George Lucas	Harrison Ford	Star Wars: Episode IV – A New Hope	1977	8.7	449935665
Steven Spielberg	Henry Thomas	E.T. the Extra-Terrestrial	1982	7.9	424449459

1. **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

   Use the filter to filter data with num_voted_users > 25,000.

   Sort data using IMDB_score column
   Use first 250 entries for analysis section

Section **= SEQUENTIAL(COUNTA(G2:G251),1,1,1) Section**
Do not select English, filter language to get only foreign language

This puts our foreign films in the top 250 list.

5749

Top 250 Movies:

https://docs.google.com/spreadsheets/d/1yxLqx_7ZnV85ceV4N5Z_vxRCVISGn-3U/edit?usp=sharing&ouid=110358868122257470939&rtpof=true&sd=true

Top 250 Foreign Language Movies:

https://docs.google.com/spreadsheets/d/1yxLqx_7ZnV85ceV4N5Z_vxRCVISGn-3U/edit?usp=sharing&ouid=110358868122257470939&rtpof=true&sd=true

D. 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

Group the column using the director_name column.

Sort alphabetically

Find top 10 directors where the mean of imdb_score is the highest.

store them in a new column with name top10director.

To perform this task, we first make the pivot table,do filtering as above and sort

Top 10 Directors	Average of imdb_score
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Asghar Farhadi	8.40
Marius A. Markevicius	8.40

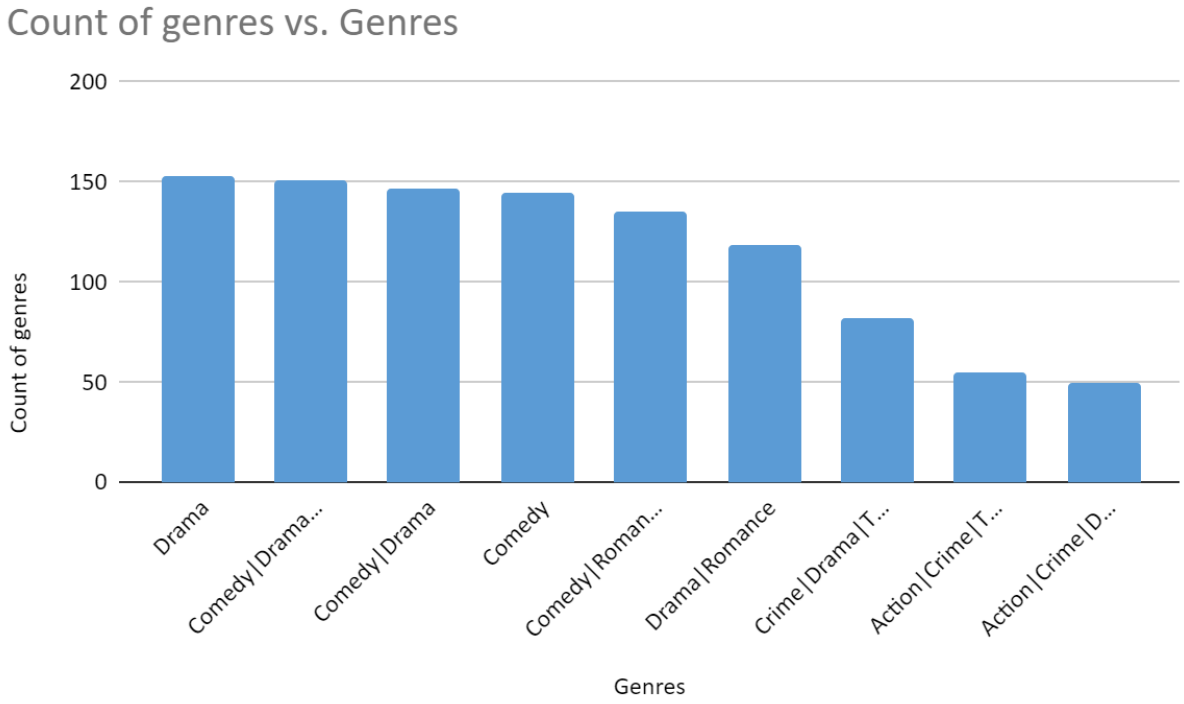
Top 10 directors:

https://docs.google.com/spreadsheets/d/1yxLqx_7ZnV85ceV4N5Z_vxRCVISGn-3U/edit?usp=sharing&ouid=110358868122257470939&rtpof=true&sd=true

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

   To perform this task, we first make the pivot table,do filtering as above and sort

Genres	**Count of genres**
Drama	153
Comedy|Drama|Romance	151
Comedy|Drama	147
Comedy	145
Comedy|Romance	135
Drama|Romance	119
Crime|Drama|Thriller	82
Action|Crime|Thriller	55
Action|Crime|Drama|Thriller	50



As we can conclude from above chart that Drama is the most watched popular genre

https://docs.google.com/spreadsheets/d/1yxLqx_7ZnV85ceV4N5Z_vxRCVISGn-3U/edit?usp=sharing&ouid=110358868122257470939&rtpof=true&sd=true

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

Charts: Create three new columns with name Meryl_Streep, Leo_Caprio and Brad_Pitt,select movies where these 3 are the lead actors

For this ,use only the actor_1_name column to do 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.

We will use bar chart so as to Observe the changes in number of voted users over decades.

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.

To find critic-favorite and audience-favorite actors : PIVOT TABLE

actor_1_name	Mean of num_user_for_reviews	Mean of num_critic_for_reviews
Brad Pit	742.35	245.00
Leonardo DiCaprio	914.48	330.19
Meryl Streep	297.18	181.45

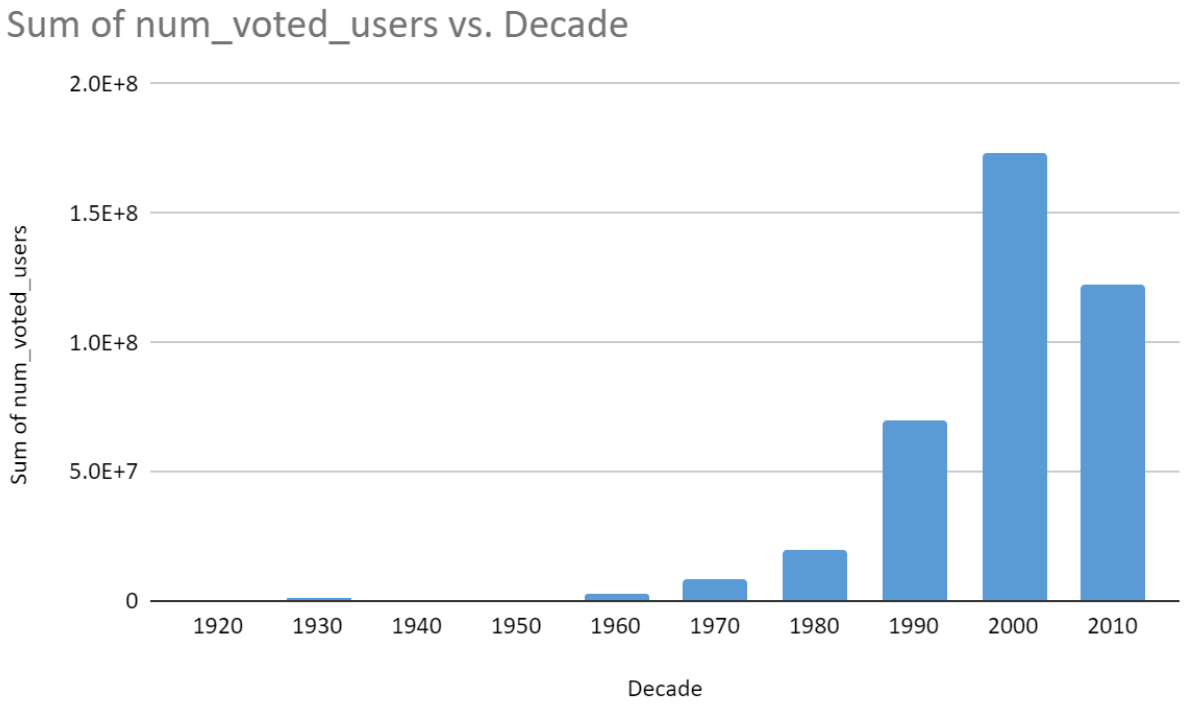
Here We can conclude that Leonardo DiCaprio is the audience's and Critic's favorite actor.

User voting by decade:PIVOT TABLE

Decade	Sum of num_voted_users
1920	116392
1930	804839
1940	230838
1950	678336

1960	2985581
1970	8704723
1980	20101705
1990	70090204
2000	173033966
2010	122492496

This shows that number of voters have increaded over years



https://docs.google.com/spreadsheets/d/1yxLqx_7ZnV85ceV4N5Z_vxRCVISGn-3U/edit?usp=sharing&ouid=110358868122257470939&rtpof=true&sd=true

Result:

By completing the project, we will have gained a better understanding of the dataset and obtained valuable insights related to movies, including financial success, ratings, directors, genres, and actor popularity. The analysis will provide a comprehensive overview of the dataset, enabling us to make informed observations and conclusions.