IMDB Movie Analysis

Project Description

The project involves analysis of IMDB movies data wherein we analyse various aspects of the dataset such as movies with highest profit, top 250 movies, best directors, popular genres and get insights on them using relevant charts and graphs using Excel.

Approach

First, we clean the dataset and clear out any duplicate or null values, outliers if found. Develop an understanding of the dataset and then process it and solve the problems laid out. We also use charts/graphs to get better understanding of the insights we get from our analysis.

Tech-Stack Used

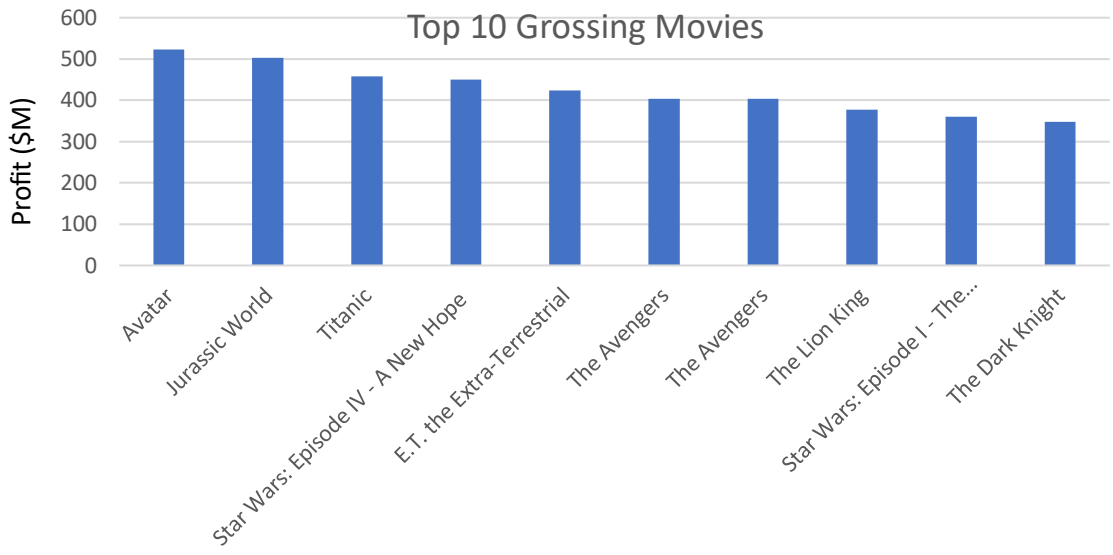
For this project, we used Microsoft Excel 2021 as it is the flagship software used across the world for spreadsheets and is highly reliable and easy to work on.

Insights

The analysis of the dataset gave us various insights which are as follows –

1. Cleaning the data.
    - For this, we used the 'Go To Special' command from the 'Find & Select' option in the 'Editing' section of Home bar. We then selected 'Blank' option in the box that opened which highlighted all the cells with null values.
    - Then, we right clicked on such a cell and deleted the entire rows having any null values.
    - This way before cleaning we had 5044 rows (including the Title row) and after cleaning we had 3757 rows left.

2. Movies with the highest profit.
    - For this, we take two columns namely 'gross' and 'budget' and divide all the cells in these columns with 1,000,000 to get the figures in million dollars.
    - Then, we make a new column named 'profit' and calculate the profit earned by each movie by subtracting 'budget' from 'gross'.
    - We then arrange the columns in descending order of the profit of all the movies and then we select the top 10 movies and its profit (in $M) to make a bar graph which comes out as follows-



3. Find Top 250 IMDB movies.
    - To get the Top 250 movies of IMDB we first set filters on the imdb_score and arrange it in descending order using the sort option from 'Editing' bar.
    - As we must select only those movies that have a vote count of more than 25,000 votes, we put a filter on the 'num_voted_users' column and set it to greater than or equal to 25,000 votes using the 'Number Filters' in the sorting option and also set it to descending order as well.
    - These two filters set would give us the list of all the movies having a vote count of more than 25,000 and a high imdb score. We now simply extract the top 250 movies in the list.
    - The final list comes out as shown in the file mentioned below- https://docs.google.com/spreadsheets/d/1nkTeFF2lHDS-_Red-n0noItLMvmGU5DF/edit?usp=sharing&ouid=106547893611024433703&rtpof=true&sd=true
    - To further get a list of Top non-English or foreign language movies we use the 'language' column and using the filters remove 'Englsih' language from the options and keep all the other languages in it. This way we get a list of movies only in foreign languages.
    - The final list is as shown in the following file- https://docs.google.com/spreadsheets/d/13M4CU_-GMlucPUPaHej-dTphleLvxy8j/edit?usp=sharing&ouid=106547893611024433703&rtpof=true&sd=true

4. Find the top 10 directors.
    - To get the top 10 directors with the highest mean of their IMDb rating we first used a pivot table to arrange the directors and get their average imdb score. We placed 'director_name' in the Rows section and the 'imdb-score' in the Values section and set it to average of. This way we got the average of every director's imdb scores.
    - Then extracted the list and put a filter on it to sort it in descending order from largest to smallest imdb score.

- Putting the top 10 directors in a separate list gave us the desired result which came out as follows-

Director	imdb_score
Akira Kurosawa	8.7
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	8.4333333
Christopher Nolan	8.425
Asghar Farhadi	8.4

5. Find the popular genres.
    - For this, as the genres are separated by '|' so we first took out each genre in the rows and put them into a separate column each using the 'Text to Column' function in Data bar.
    - As the films are primarily categorised based on the first two genre mentioned so we used the first two genres we put in separate columns named 'genre1' and 'genre2'.
    - We selected the data and used pivot tables to get the count of each genre and thereby took the mean of counts of both the genres.
    - Through this we were able to arrive at the desired result-
      The most popular genres according to IMDb are Drama, Comedy and Action.

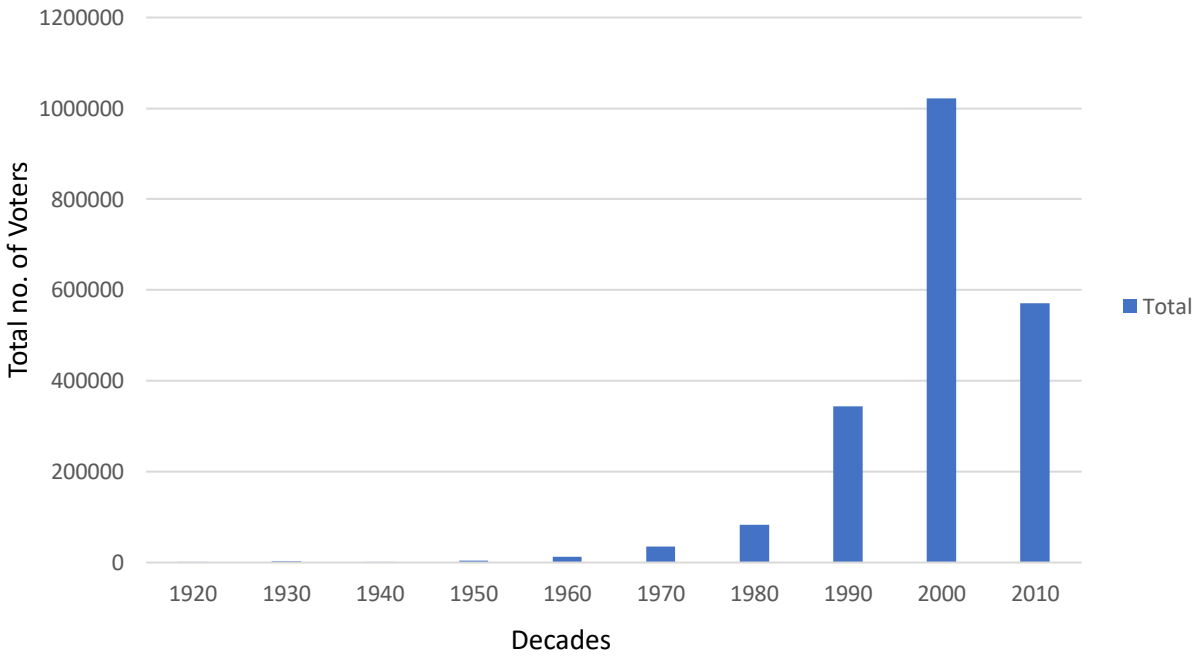
Genre	Mean
Drama	850.5
Comedy	710
Action	482.5
Adventure	391.5
Crime	297

6. Find the critic favourite and audience favourite actors.
    - First, we extracted the actor_1_name, movie_title, num_critic_for_reviews and num_user_for_reviews columns to a new sheet and selected all the data in the new sheet and used a pivo table on it.
    - We then applied filters in the pivot table and selected the desired actors whose mean critic and user reviews we wanted to calculate and set the values column to 'Average' from 'Sum' to get the average of the counts of critic and user reviews.
    - Then we identified the actors with the highest mean which came out to be Leonardo DiCaprio.

actor	mean_critic	mean_user
leo_caprio	330.19	914.47
brad_pitt	245	742.35
meryl_streep	181.45	297.18

- Thus, Leonardo DiCaprio is the critic as well as audience favourite.
- Furthermore, to observe the changes in the number of voted users over the decades we summed up all the voters in a decade and then plotted a bar graph on the data which resulted in the following chart.

decade	num_voted_users
1920	780
1930	2867
1940	1174
1950	3948
1960	13132
1970	35416
1980	82560
1990	343423
2000	1021534
2010	570265



From this graph we can see that the decade 2000s had the maximum number of voters in a span of a century with 1021534 voters in total. The numbers kept rising and saw a sudden and exponential rise starting from 1980s and saw a decline in the 2010s but let's also consider the fact that the data is available till 2016 only.

Result

From this project, we were able to get an insight into the IMDb movies dataset and analyse various elements of the dataset like movies with the highest profits, top 250 movies in English and foreign languages as well, best directors, popular genres, and an insight into the number of voters for the critic and audience reviews over the decades. These results gave us a better understanding of the movie world all across the globe and increased our knowledge as well.