Project Report: IMDB Data Analysis using Tableau

By: Malay Gupta 20BCE10069

Neelesh Shukla 20BCE10236

Sagar Dubey 20BCE10242

Vatsalya Singh 20BCE10044

1. Introduction

The purpose of this project is to perform a comprehensive analysis of the IMDB (Internet Movie Database) dataset using Tableau, a powerful data visualization tool. The dataset contains information about movies, including their titles, genres, ratings, and other relevant details. Through this analysis, we aim to gain insights into various aspects of the movie industry, such as popular genres, highest-rated movies, and the relationship between movie attributes and their success.

2. Data Source

The IMDB dataset used for this analysis was obtained from [source]. It consists of a collection of movies released over the years, along with their associated attributes and ratings. The dataset includes information like movie title, genre, release year, runtime, IMDB rating, number of votes, and more.

3. Data Cleaning and Preparation

Before conducting the analysis, the dataset underwent a cleaning and preparation process to ensure data quality and consistency. This involved handling missing values, removing duplicates, and formatting data types appropriately. Outliers, if present, were also treated to prevent their influence on the analysis.

4. Analysis and Visualizations

4.1. Movie Genre Analysis

- Bar chart: The distribution of movies across different genres was analyzed using a bar chart. This visualization provided insights into the most popular genres within the dataset.

- Pie chart: A pie chart was created to represent the proportion of movies belonging to each genre. This visualization helped identify the relative popularity of different genres.

4.2. Movie Ratings Analysis

- Histogram: A histogram was used to visualize the distribution of movie ratings. This allowed us to understand the overall rating distribution and identify any patterns or trends.

- Line chart: A line chart was created to display the average movie ratings over the years. This visualization provided insights into the changing trends in movie ratings over time.

4.3. Movie Attributes Analysis

- Scatter plot: A scatter plot was used to analyze the relationship between movie runtime and ratings. This visualization helped determine if there was any correlation between these two attributes.

- Box plot: A box plot was created to compare the ratings of different genres. This visualization allowed us to understand the variation in ratings across various movie genres.

4.4. Top Rated Movies Analysis

- Table: A table was created to display the top-rated movies along with their attributes, such as title, genre, rating, and number of votes. This allowed us to identify the highest-rated movies within the dataset.

5. Key Findings

Based on the analysis performed on the IMDB dataset, the following key findings were observed:

- The most popular movie genres within the dataset were Action, Drama, and Comedy.

- The majority of movies fell into the average rating range, with a few exceptional outliers.

- The average movie ratings remained relatively stable over the years, with a slight upward trend.

- There was a weak positive correlation between movie runtime and ratings, indicating that longer movies tended to have slightly higher ratings.

- The genre of a movie had a significant impact on its rating, with certain genres consistently receiving higher ratings compared to others.

6. Conclusion

In conclusion, this project successfully utilized Tableau for analyzing the IMDB dataset and extracting meaningful insights. The visualizations and analysis provided a comprehensive understanding of movie genres, ratings, attributes, and top-rated movies. These findings can be valuable for filmmakers, production companies, and movie enthusiasts to make informed decisions and gain a deeper understanding of the movie industry.

7. Limitations and Future Work

This project has a few limitations that should be considered:

- The analysis is based on a single dataset, and the findings might not be representative of the entire movie industry.

- The dataset's time range might affect

the analysis of trends, as it only covers a specific period.

- The accuracy of the ratings relies on user votes and might be subjective.

For future work, additional datasets could be incorporated to conduct more extensive analysis, such as analyzing the relationship between movie budgets and revenues or exploring the impact of actors/directors on movie success. Furthermore, sentiment analysis on user reviews could be performed to gain insights into the audience's perceptions and opinions of movies.

8. Bibliography

https://datasets.imdbws.com/

https://www.boxofficemojo.com/franchise/?ref_=bo_nb_fr_secondarytab

https://www.boxofficemojo.com/brand/?ref_=bo_nb_frs_secondarytab

https://grouplens.org/datasets/movielens/25m/

https://www.the-numbers.com/movies/franchises

https://www.the-numbers.com/movies/franchise/Marvel-Cinematic-Universe#tab=summary

https://www.the-numbers.com/movie/Avengers-The-(2012)#tab=box-office

www.kaggle.com

www.imdb.com

www.tableau.com