Nolan Arendt & Robert Ransom

Syracuse univeristy ischool | june 2023

Analyzing Viewer Voting Trends in Classic and Modern Cinema

IST 652 – Dr. d. Landowski

Nolan Arendt and Robert Ransom

Final Project

**INTRODUCTION**

The goal of our final project is to create a program that demonstrates our ability to write Python scripts to access and amass data from a dataset of movie reviews. Next, we will carefully construct our panda’s data frames for our processed data that will be utilized to assist in answering our business-related questions. Lastly, we will create several visualizations to help represent our findings using pandas, matplotlib, seaborn, etc. The purpose of this final project is to gain experience with the collection process of data, data transformations, and deriving key insights from the data we process.

**GROUP MEMBER WORK**

Robert Ransom

* Data cleaning
* Feature engineering
* Data visualization

Nolan Arendt

* Data selection
* Exploratory data analysis
* Data visualization

**METHODOLOGY**

Following topic selection, data sources were identified using Kaggle (<https://www.kaggle.com/datasets/simhyunsu/imdbextensivedataset?select=IMDb+movies.csv>) that provided semi-structured and structured data describing voting trends on global movie titles produced between 1894 and 2020. The decision to remove multi-genre films was made early on for ease of categorization and cleaning to avoid differing input patterns for similar fields (i.e. Action, Adventure, Sci-Fi > Adventure, Action, Sci-Fi). Exploratory Data Analysis proved that paring down the dataset for single-genre analysis still provided valuable insight for business questions while affording topics for future investigation and research in the film industry.

Business Analysis focus was placed on identifying trends by genre to inform production executives on historical successes, and failures, in the movie industry. By identifying the most successful genres, countries of origin, target audience, and duration, recommendations can be made to tailor movie-going experiences for increased box-office success.

* Identify trends by genre that indicate higher levels of viewer interaction (vote counts), responses (vote score) and areas where single-genre films fall short of expected successes.
  + Use case and Output: Specifically targeting number of responses and average score, analysis will highlight historical successes and overall viewer interaction; script nomination and production likelihood are particularly emphasized based on these results.
* Conduct a sentiment analysis of plot description and provide comparison across genres for trends on positive, negative, and neutral descriptions.
  + Use case and Output: Examination of plot description provides insight into the impact of marketing and socialization of themes, topics, and descriptive nature of language selection.
* Identify voting trends between listed genders; are males or females more likely to score a particular genre higher or lower?
  + Use case and Output: Addressed for marketing evaluation, are different genres expected to perform at specific levels based on target audience; marketing and socialization efforts to mitigate potential risk based on expected audience response.
* Examine voting trends by film duration.
  + Use case and Output: Scatter plot with trend lines comparing average votes against runtime. Editing for runtime could impact average score by genre.
* Compare successes and failures by country of origin.
  + Use case and Output: Provide insight into the impact of geopolitical situations that could determine the success or failure of a particular movie industry by nation. Likely a target for future research, additional information required for examination of viewer scores, viewer nationality (potentially politically motivated), density of national film industry in the global market (does a nation have vastly more films produced and therefore more votes for/against).
* Identify viewer interaction trends in title production by year; identify a particular era where viewer interaction peaked.
  + Use case and Output: Identify popularity of different eras of film; script selection or marketing potentially focused on period pieces or production styles or increase viewer interaction and attendance.

**PROGRAM DESCRIPTION**

The program is designed to ingest locally stored data, merge movie descriptions and ratings datasets, clean based on selected parameters (single-genre films) and drop incomplete data fields. Following cleaning, Plot Description Sentiment Analysis and feature engineering occur. Business Analysis focus areas are addressed, visualization creation and formatting conducted using bar, histogram, stacked bar, line, and scatterplot representations of metric and categorical information.

**ABOUT THE DATA**

The dataset comprises approximately 19,000 rows and 24 columns, with each row representing an individual movie. The focus of the analysis is on movie voting trends, exploring various factors such as genre, country of origin, plot sentiment derived from the description, average vote, and gender-related statistics. The columns include information such as the movie's title, release year, genre, duration, country, description, sentiment score and category for the description, average vote, number of votes, user, and critic reviews, weighted average vote, mean and median votes, total votes, average vote and number of votes by males and females of all ages, ratings and votes from US voters and non-US voters. This dataset provides a comprehensive range of variables to investigate and understand patterns and preferences in movie voting trends across different genres, countries, and demographics.

**DATA CLEANING AND EXPLORATION**

The dataset was prepared for analysis by merging the IMDb Movies and IMDb Ratings datasets. To simplify the analysis of average ratings by genre, movies with multiple genres were excluded, allowing for easier comparison and interpretation of ratings across different genres.

To ensure an adequate sample size for analysis, genres with fewer than 100 instances and countries with fewer than 10 movies were excluded. Outliers in movie duration, exceeding 300 minutes, were also removed. These data transformations improved the dataset's integrity and enabled comprehensive exploratory analysis of movies and their ratings by genre.

**DATA DICTIONARY**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Type** | **Nulls** | **Definition** |
| Title | Object | 0 | Title of the film. |
| Year | Int32 | 0 | Year the film was released. |
| Genre | Object | 0 | Genre of the film. |
| Duration | Int64 | 0 | How long the film is in minutes. |
| Country | Object | 0 | The country where the film was released. |
| Description | Object | 0 | Short text description of the film. |
| Description\_Sentiment\_Score | Float64 | 0 | Scale from -1 to 1 measuring the description sentiment. -1 being the most negative, and 1 being the most positive sentiment. |
| Description\_Sentiment | Object | 0 | If the description\_sentiment\_score is above 0, the description sentiment is positive. If the description\_sentiment\_score is below 0, the description\_sentiment is negative. If the description\_ sentiment is 0, the description\_sentiment is neutral. |
| Avg\_Vote | Float64 | 0 | The films average vote by all voters. |
| Votes | Int64 | 0 | The number of votes for each film. |
| Reviews\_From\_Users | Float64 | 0 | The number of reviews from non-critics. |
| Reviews\_From\_Critics | Float64 | 0 | The number of reviews from critics. |
| Weighted\_Average\_Vote | Float64 | 0 | The weighted average vote from all the film reviews. |
| Mean\_Vote | Float64 | 0 | The mean review score for each film from all the votes. |
| Median\_Vote | Float64 | 0 | The median review score for each film from all the votes. |
| Total\_Votes | Int64 | 0 | The number of votes / reviews for each film. |
| Males\_AllAges\_Votes | Float64 | 0 | The total number of votes for each film by men. |
| Males\_AllAges\_Avg\_Vote | Float64 | 0 | The average vote for each film by men. |
| Females\_AllAges\_Votes | Float64 | 0 | The total number of votes for each film by females. |
| Females\_AllAges\_Avg\_Vote | Float64 | 0 | The average vote for each film by females. |
| Us\_Voters\_Rating | Float64 | 0 | The average vote for each film by individuals in the United States. |
| Us\_Voters\_Votes | Float64 | 0 | The total number of votes for each film by individuals in the United States. |
| Non\_Us\_Voters\_Rating | Float64 | 0 | The average vote for each film by individuals not in the United States. |
| Non\_Us\_Voters\_Votes | Float64 | 0 | The total number of votes for each film by individuals not in the United States. |

**DATA PREPROCESSING**

**DATA TRANSFORMATION – NEW VARIABLES**

*Description\_Sentiment\_Score*

Vader Sentiment analysis was conducted on movie descriptions, resulting in the creation of the 'description\_sentiment\_score' column. This column assigned sentiment scores ranging from -1 to 1, providing a quantitative measure of positivity or negativity. Like a correlation matrix analysis, values farther away from 0 indicate a stronger relationship toward positive or negative sentiment.

*Description\_Sentiment*

Categorical labels were then applied to the sentiment scores, creating the 'description\_sentiment' column, which identified movies as 'Positive' or 'Negative' based on the sentiment score. This sentiment analysis facilitated exploring the relationship between sentiment and ratings.

**BUSINESS ANALYSIS**

**GENRE –** Voting trends and plot description analysis.

In our dataset of movies, it is evident that Drama and Comedy movies have captured the audience's attention the most. With approximately 33 million votes, Drama emerged as the genre with the highest number of votes, closely followed by Comedy with 24 million votes. In stark contrast, the other genres in the dataset received considerably fewer votes, all below 7 million. This analysis highlights the widespread popularity and appeal of Drama and Comedy films among the viewers, indicating their enduring relevance in the film industry.

Moreover, when considering the average vote ratings, Drama once again takes the lead with an impressive score of 6.4. This finding signifies the appreciation and positive reception that Drama movies have garnered from the audience. Conversely, Sci-Fi experiences the lowest performance in terms of average vote, scoring around 4.33. This observation suggests that the Sci-Fi genre might not resonate as strongly with the viewers in this dataset, indicating a potential area for improvement or exploration within the genre. An important caveat for these insights is that they examine films with only a single genre listed; additional statistical analysis could reveal that combining genres (e.g.: Action/Adventure/Sci-Fi) may result in higher viewer interaction and different voting scores and trends.

These insightful bar graphs portraying the number of votes and average vote by genre offer valuable insights into the movie preferences and opinions of the dataset's audience. By understanding which genres receive the most votes and have the highest average ratings, filmmakers and producers can make informed decisions about their movie offerings. This analysis helps guide the industry in identifying trends, meeting audience expectations, and potentially catering to the demand for Drama and Comedy movies, while also highlighting areas where other genres may require further attention and improvement.

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated

By analyzing the sentiment of movie descriptions, we established a relationship between plot sentiment and average vote. Preliminary results indicate that movies with more positive or emotionally resonant descriptions tend to receive higher average votes. This suggests the importance of captivating storytelling and engaging narratives in driving audience satisfaction and subsequent ratings.

A picture containing text, screenshot, line, diagram

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

**AVERAGE VOTE –** Compare voting trends by gender, duration, country of origin.

An intriguing finding within the dataset is that, across all genres of movies, women consistently provided higher average votes than men. This observation suggests a gender-based divergence in movie preferences and opinions, emphasizing the importance of considering the diverse perspectives of both male and female viewers. Particularly noteworthy is the genre of Family movies, which exhibits the most significant difference in average vote between genders. This valuable information holds the potential to shape marketing strategies, production decisions, and content creation in the film industry. By recognizing and addressing these gender-based variations in movie ratings, filmmakers can cater to the preferences of their target audience more effectively, ensuring that their movies resonate with viewers of all genders and ultimately enhancing audience satisfaction.

A picture containing screenshot, text, plot, line

Description automatically generatedA picture containing diagram, plot, screenshot, line

Description automatically generated

Our analysis reveals a consistent positive correlation between movie runtime and average vote, regardless of the genre. Notably, as movies become longer in duration, their average vote tends to increase. This observation emphasizes the significance of storytelling and narrative depth in captivating audiences and garnering higher ratings.

A picture containing screenshot, map, diagram, line

Description automatically generated

A screen shot of a graph

Description automatically generated with medium confidence

Examining the countries with the highest and lowest average vote offers intriguing insights. Czechoslovakia, the Soviet Union, and Pakistan emerge as the top-ranking countries in terms of average vote, while the United States, Thailand, and Canada exhibit lower average vote. These findings raise questions about the influence of cultural backgrounds, storytelling styles, and regional audience preferences on movie ratings.

A picture containing text, screenshot, diagram, plot

Description automatically generated A picture containing text, screenshot, diagram, plot

Description automatically generated

Our analysis examines average vote trends across genders and history. Interestingly, women consistently provide higher average votes than men in each year of the dataset dating back to the 1800s. However, when considering the total number of votes, men surpass women. These findings raise intriguing questions about potential variations in rating criteria and voting behavior based on gender. Analyzing average voting trends over time reveals a general decline in ratings across all genres, particularly starting in the 1960s, when we see a sudden increase in voting. This observation prompts further exploration into the societal and cultural shifts that may have influenced audience expectations and preferences over the years.

A picture containing text, screenshot, plot, line

Description automatically generated

**FINAL CONCLUSION**

Our analysis highlights the significance of genre in shaping movie ratings and voting patterns. Certain genres, such as Drama and Comedy, attract a larger number of votes and receive higher average ratings, indicating their popularity and appeal among audiences. The lower average votes for Sci-Fi and Horror genres present opportunities for improvement in catering to audience preferences. The gender-based voting disparities emphasize the importance of considering the diverse perspectives and preferences of viewers when creating and marketing movies.

This study provides valuable insights into the factors impacting the average vote in the domain of movies. The analysis reveals a consistent positive correlation between movie runtime and average vote, highlighting the significance of narrative depth and storytelling in capturing audience appreciation. Moreover, the examination of countries of origin demonstrates intriguing variations in average vote, suggesting the influence of cultural backgrounds and regional audience preferences on movie ratings. Additionally, the analysis of plot sentiment indicates the importance of captivating storytelling and emotionally resonant narratives in driving audience satisfaction and subsequent ratings.

By examining the significance of genre, runtime, country of origin, and plot sentiment, we offer actionable information for the film industry to better understand audience preferences and create more successful and engaging movies. These insights can lead to improved marketing strategies, content development, and overall audience satisfaction, ultimately driving the success of movie productions.