Homework 1
Problem 1
Ayad Masud
733009045

I certify that I have <u>*personally*</u> *done the coding, generated the figures and written the report without aid from anybody else, and that I have not plagiarized, self-plagiarized, or used AI-generated text. I certify that I have acknowledged any sources I used to complete this assignment*. ARM.

1 Part 1: Exploring Votes vs Gross Earnings

The purpose of question 1 is to dive deeper into the dataset and understand all the features and information that it provides. To lay some groundwork for upcoming analytical questions, different data visualization techniques were used to find correlations and relationships between features. These techniques included correlograms, box plots, scatter plots, and density plots. For this first part the relationship between the number of votes and the total gross income was analyzed using a scatterplot. Results are shown in **Figure 1**.

The scatterplot shows the relationship between votes and gross earnings. We can see that the data is focused mostly in one spot and slowly spreads up and to the right in a fan shape. This indicates that most movies' gross earnings are somewhat affected by the votes they get. There are some outliers where the more votes the movie got the more it earned. This may show that a movie can get many votes meaning that it was appreciated by the audience and critics alike, but did not attract a large enough theater going audience. Many factors such as limited marketing, niche genre, or competition from other movie releases at the time may have caused this. There are also popular streaming platforms where the movie may have been popular when released, but did not perform as expected in box offices, revealing why a movie got so many votes, but the gross earnings does not reflect that.

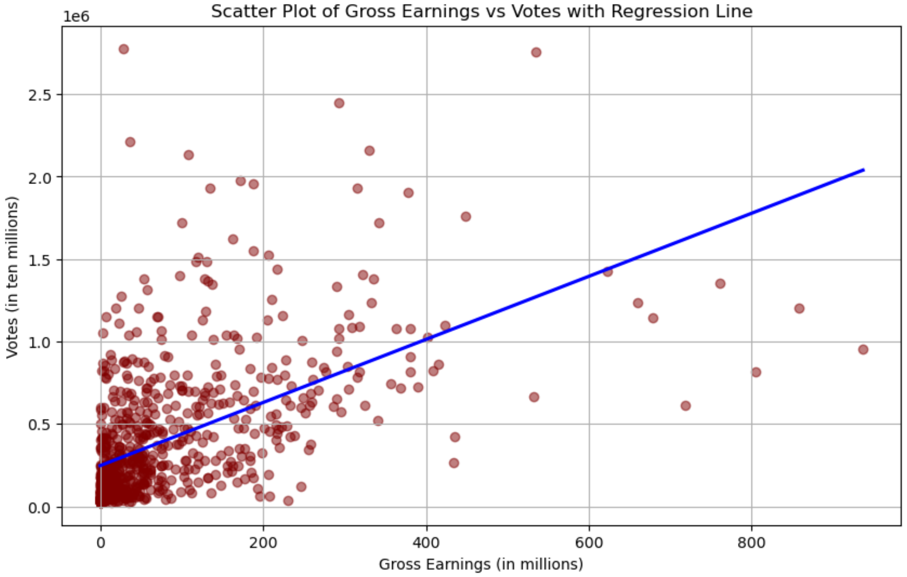

Figure 1. A scatterplot with a regression line showcasing the relationship between Gross Earnings (in millions) and Votes (in ten million).

2 Part 2: Exploring Number of Movies Released in 5 Year Intervals.

For this second part we analyze the number of movies that are released in 5-year intervals. This was done by using a bar plot to visualize the number of movies and extract certain relationships. Results are shown in **Figure 2.**

The box plot shows a steady upward trend from 1925 - 2020. Something interesting to notice is the steep decline in the year 2020. The steady increase in movies being released can be explained by the ease of use of cameras, microphones, CGI, and other technology that make is much easier and faster to film and produce movies. The influx of viewers and ways to view movies such as the development of the internet and streaming platforms have also provided the funding and attention needed to produce more movies. The sharp decline in movies released during the year 2020 can be explained by the worldwide pandemic. The lockdown enforced during this time limited actors, directors, crew, and entire companies from producing movies, which explains the much less number of movies produced compared to previous years.

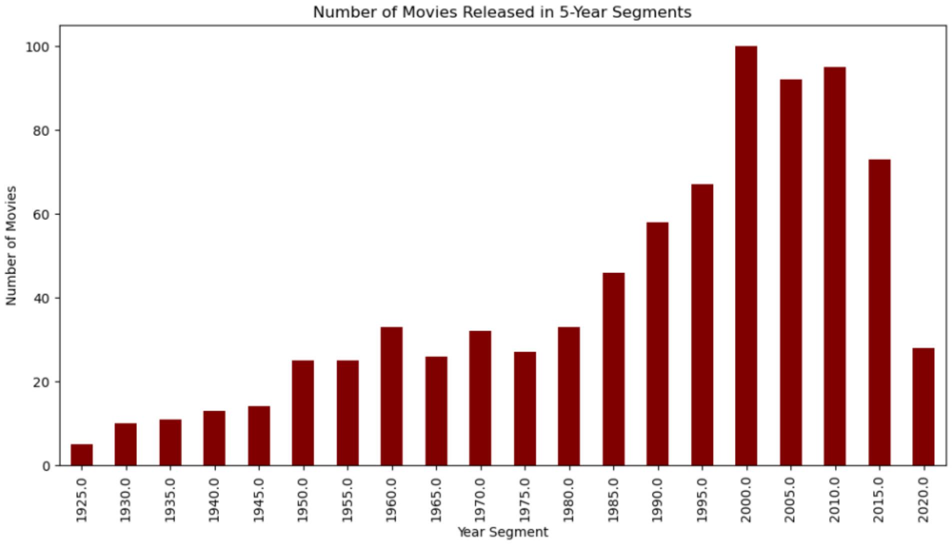

Figure 2. A bar plot showing the number of movies released in each 5-year segment.

3 Part 3: Exploring Density Plots of Watch Time.

For the third part we will analyze the density plot of watch time. Through this graph we will be able to extract useful information about the distribution of this features and resulting effects. Results show in **Figure 3.**

We can see that the density plot for watch time is a right skewed distribution. This means that the median and mode of the watch time is less than the overall mean of the watch time. There a several things that this right skewed distribution can imply about the watch time. We can see that the most common length for a movie is around 110 minutes. There is a specific reason for this in that less than that amount of time audience members say it is too short and there is no substance to the movie, however, if its longer than 110 minutes audience members tend to get bored and lose attention. Another reason why this distribution could be right skewed is the screenplay structure that has been used for several years. For example, many movies have a very distinct introduction, story, and conclusion. Well that conveniently fits into around 30-minute sections, hence why the median and mode of watch time is to the left of the mean.

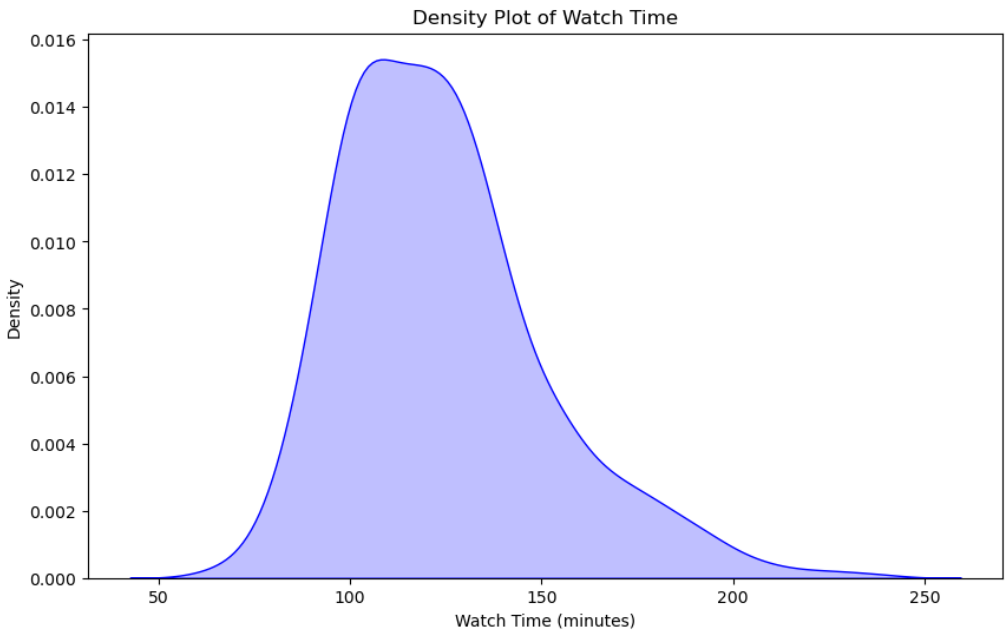

Figure 3. A density plot showing the distribution of watch time (minutes).

4 Resources used to achieve this goal

Python Libraries: Pandas, NumPy, matplotlib, and seaborn. These were used to clean and parse through data as well as display statistical visualizations of the results.

Canvas Home Page: Resources such as lecture slides and homework template.

5 References

- inductive_anks. (2023). *Top 1000 IMDb Movies Dataset*. Kaggle.com. https://www.kaggle.com/datasets/inductiveanks/top-1000-imdb-movies-dataset?resource=download
- Khushi Pitroda. (2023, August 16). *EDA + Sentiment Analysis Top_1000_IMDb_movies*. Kaggle.com; Kaggle. https://www.kaggle.com/code/khushipitroda/eda-sentiment-analysis-top-1000-imdb-movies