Final Report

Movie Data Mining

Andrew Turangan
DSCI 510

Introduction

This project is titled "Movie Data Mining" and was done as a solo project by Andrew Turangan. The aim of this project is to exercise methods of web scraping and data analysis to extract valuable information concerning popular movie titles. Such insights can be used to create more efficient recommendation systems or to build upon large-scale databases.

Data

The raw data was extracted via an HTML parser from a web page by the popular movie critic site Rotten Tomatoes. This particular web page listed all 96 Oscar-winning Best Picture movie titles from 1928 to 2022 along with various information useful for analysis, making it a copious source of movie data excellent to exercise web scraping and data analysis techniques.

The webpage can be reached through the Internet with the following link:
https://editorial.rottentomatoes.com/guide/oscars-best-and-worst-best-pictures/

The data was parsed with BeautifulSoup. 96 samples were collected. Each sample contained the following attributes:

- Movie Title
- Year of Release/Oscar Win
- Critic Consensus
- Synopsis
- Starring Cast
- Director
- Rotten Tomatoes Score (RT Score, primary metric for movie performance)

The approach was to web scrape from the source's HTML file, extract and clean the above attributes, and gather potentially useful information from data analyses.

The greatest challenges mostly revolved around questions of where to gather the raw data necessary for the project in the first place. Some sites on the Internet had hundreds of movie titles but very few attributes to be extracted from while others had very few movie titles but had many attributes to be extracted from. With more time and more team members, it would have been more feasible to gather movie titles from multiple sites and fill in necessary missing attributes. The data source used in this project had more than a few samples to be extracted with the advantage of having all the above attributes easily found.

Analysis & Visualization

The following analyses were performed:

- Calculation of average RT Score for each director
- Calculation of average RT Score for each cast member (actor)
- Calculation of top 10 directors and top 25 actors based on average RT Score
- Distribution Visualization of director average RT Score
- Distribution Visualization of actor average RT Score
- Regression Analysis of the following 2 attributes: Year of Release/Oscar Win, RT Score

Figure 1

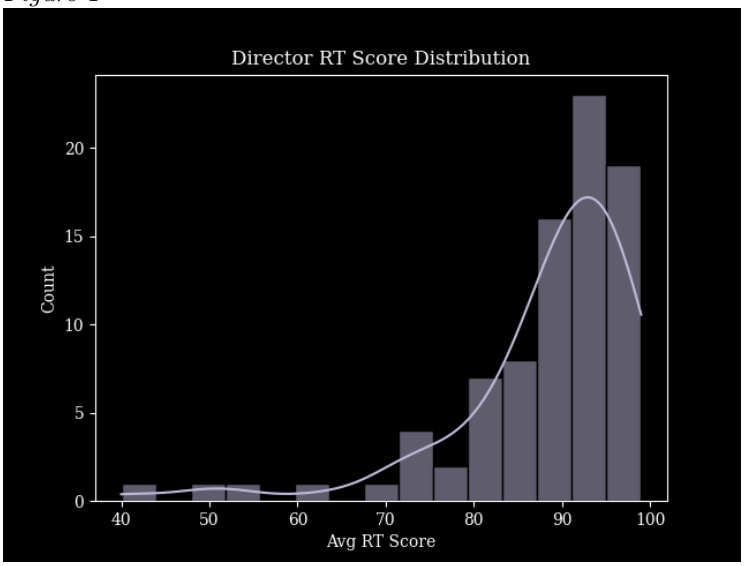

Figure 1 visualizes the distribution of average RT Score for directors included in the samples. A histogram and a probability density function are included. Observe a left-skewed distribution, suggesting a tendency for the directors in the sample to have RT scores in the higher ranges of possibility.

Figure 2

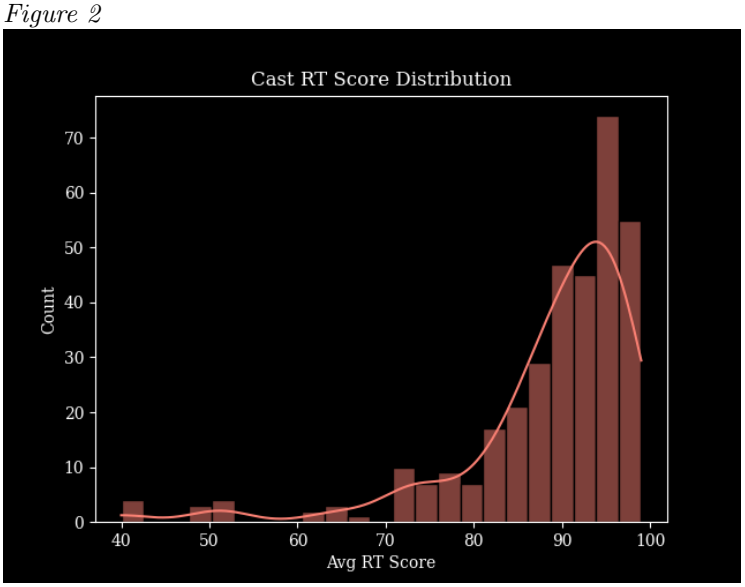

Figure 2 visualizes the distribution of average RT Score for cast members included in the samples. A histogram and a probability density function are included. Observe a left-skewed distribution, suggesting a tendency for the actors in the sample to have RT scores in the higher ranges of possibility.

Figure 3

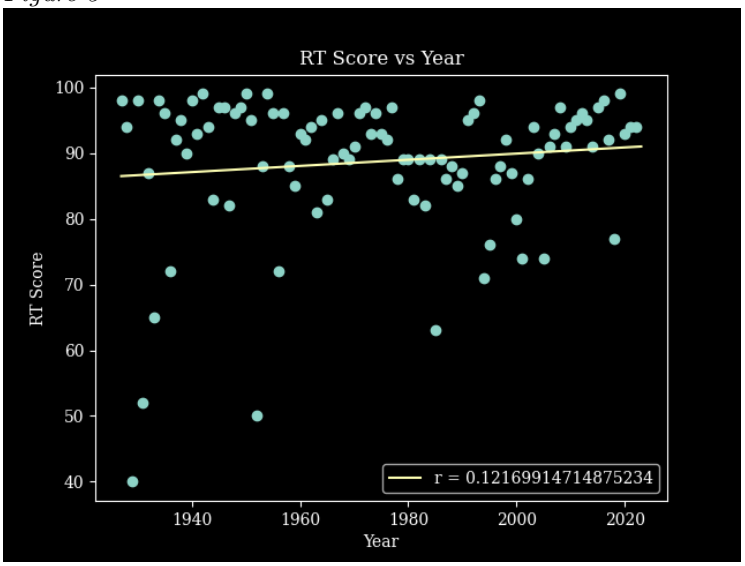

Figure 3 visualizes regression analysis between Year of Release/Oscar Win and RT Score. Weak correlation between the attributes can be observed both visually and numerically with a Pearson correlation coefficient value r of approximately .12.

These findings strongly reflect the nature of the data source in that all the samples include Oscar Best Picture winners. This fact is reflected by the visualizations of average RT Scores being in the upper ranges, irrespective of being tied to directors or actors.

The question of whether there was a tendency with time for RT Score to trend within a certain range was

met with challenge as there was very little correlation found between Year of Release/Oscar Win and RT Score. This is probably because of the "timeless" nature of these movies, as they are all samples of Oscar Best Picture winning movies. Different samples of non-award-winning movies might suggest otherwise.

Future Work

With more project members and time, this project could evolve to build larger scale databases that can be used as valuable samples for future analyses and development of machine learning models and recommendation systems. More data sources could be mined and thus more samples with greater versaility could be captured in future development.