Predicting IMDb Movie Ratings

By: Owen Doyle and Justin Self

1: Introduction

At their best, movies are an art form that make people feel some sort of emotion. Some movies entertain us, some make us sad, and others make us question the world around us. Some of the best movies just have something special about them that is difficult to describe, let alone quantify. Yet many people still attempt to describe what makes a good or bad movie, and there are a wide range of tastes and opinions about movies. These are often expressed in movie reviews. We can likely learn a lot about common characteristics of good movies by examining these reviews and data about the making of the movies these reviews are based on. Understanding this could help those who are involved in the movie-making process use their resources more effectively in order to create movies that people will like. Making movies people like also increases profits for these companies.

IMBd is one of the most popular movie websites that allows ordinary people to review movies they have watched. Aside from being a place to review the movies, IMDb contains vast information about the more than 6 million movies it hosts including plot descriptions, release dates, and where to watch the movie. Public websites like these are great resources for data analysis projects as they contain large amounts of data. One such use of this large dataset is to predict average IMDb rating using characteristics of the movie. The goal is to conduct an exploratory data analysis of the IMDb movie data set to identify what model and predictors are best suited for our research.

2: Methodology

The data set was published on Kaggle by Stefano Leone. It was created via web scraping, containing movies with more than 100 votes on the IMDb website. The repository contains four files, each describing various characteristics of movies, actors, and reviewers. For the purposes of this project, only the dataset consisting of variables that pertain to the movie was used.

After obtaining the data set from Kaggle, additional data cleaning was needed. All data cleaning was done in RStudio. To begin with, the data set was trimmed to only contain movies made in the USA, had English as its main language, and more than 500 votes on IMDb. Then the data was clipped to only contain observations, or movies, that did not have any null values. Finally, any movies produced before 1990 were clipped.

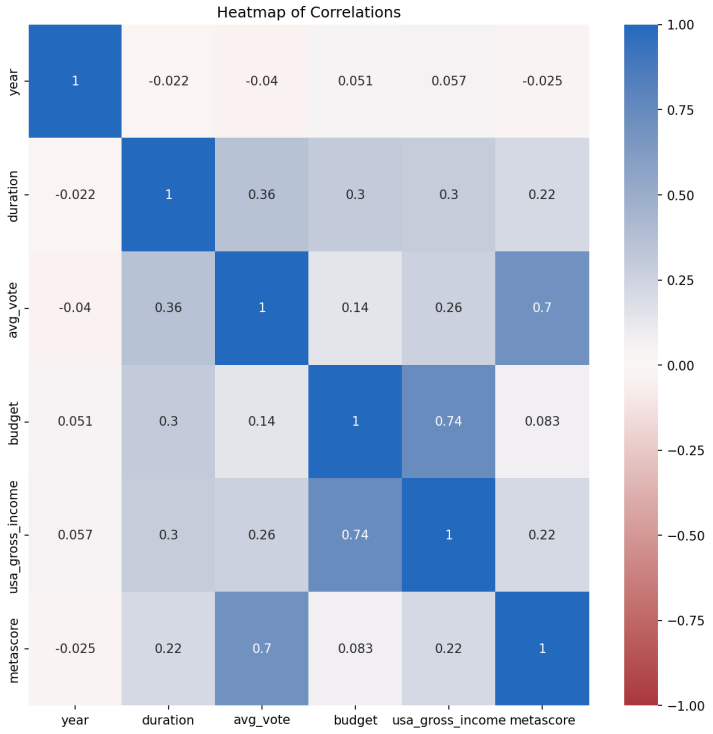
Aside from the necessary clipping of the data, some data manipulation was needed. Specifically, a few of the variables, namely usa gross income, world wide gross income, budget, and year, were read as character variables rather than integers. To correct this, the variables were coerced into integer form.

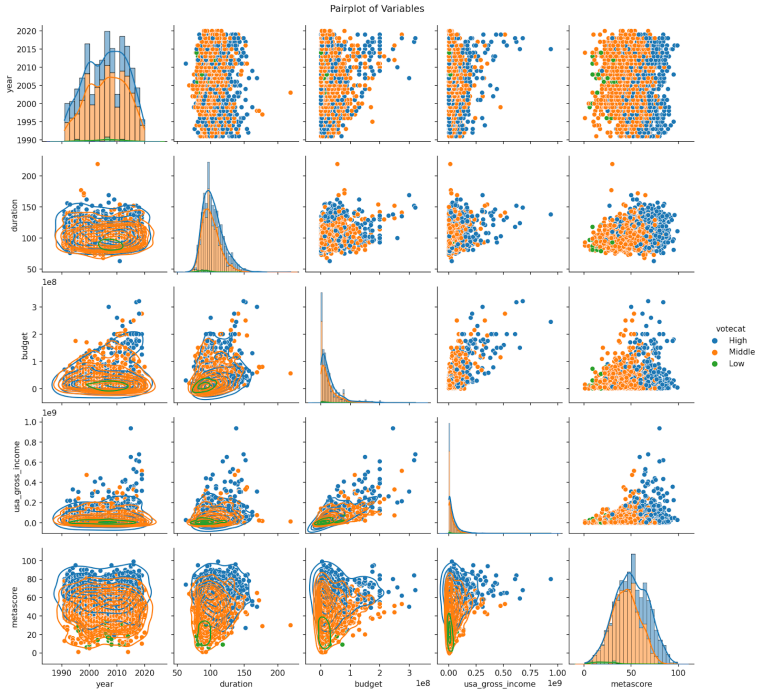
The last piece of the data manipulation was creating a new variable titled “votecat”, for vote category. This variable creates labels for each movie, namely low, middle and high, based on the IMDb average vote. The ten point scale was broken into thirds for this categorical variable.

The main purpose that this variable serves is to help break up the data for visualization purposes.

3: Analysis

The goal of the exploratory data analysis was to determine which variables will be most important in determining the average score that reviewers give a movie. The correlation heatmap shows the variables that are likely most highly correlated with the average vote. These variables include duration, budget, USA gross income, and metascore (a weighted average of reviews from reputable critics). Duration, USA gross income, and metascore all have moderate correlation with average vote, while the correlation between budget and average vote is surprisingly low.





The pairplot shows relationships between these variables of interest. The class used to separate the colors is based on a range of average votes. High votes are considered any average vote over 6.6. Medium votes are average votes between 3.3 and 6.6. Low votes are average votes below 3.3. From the pairplot, it appears that the variable that provides the most distinction of ranges of scores is the metascore. Duration and USA gross income also provide some distinction. These three variables will likely be most important to a successful modeling effort.

4: Conclusions

Based on the results of the exploratory data analysis, metascore, duration, and USA gross income will be the factors that will be used in the model to predict the average ratings of movies. Metascore and USA gross income do present some shortcomings, however, considering the goal of the project. Metascore might be considered too similar to the average votes, as it is also based on reviews from people. It is not useful to predict something based off of another variable that already presents what the predictand deals with. USA gross income also presents a limitation, as it is unknown until after a movie has been released to the general public. It would not have much practical significance in trying to create a movie that people enjoy.

Despite these limitations, these variables will still be used in the modeling, but other variables related to the cast and director might be necessary to create a more successful model. The model will be a regression model, as the goal is to estimate the relationship between independent variables and a dependent variable. The goal of the model is to predict the average rating of a movie based on the variables identified above.

5: Updated Schedule

A more specific tentative schedule for the remainder of the project is included below.

Part 3:

Modeling Efforts: Justin and Owen

Research Paper Draft

- Introduction: Justin
- Data and methods: Owen
- Results: Justin
- Discussion: Justin
- Conclusion: Owen
- References: Justin and Owen

Demonstration: Justin and Owen

Part 4:

Paper Editing: Justin

Presentation Preparation: Owen

Presentation/Demo: Justin and Owen