SI 618 Project Proposal 2

Xinye Jiang

1. Summary and Motivation:

The movies' revenues show the success of movies in a way. In this project, I want to find out the trend of movie revenues, what are its influencing factors and how these factors influence revenues. I will accomplish this goal by performing exploratory analysis such as extracting the relationship between movie revenue and release time, director and so on respectively by plots and applying some machine learning techniques.

The datasets that I will use are the same datasets that I used in project part 1. The datasets provide movie basic information including revenue, release date, genre, budget and the like, and cast and crew information including director and actors. After some manipulations of the datasets, I could use data.table and ggplot to explore the relationship between revenue and other factors of interest like release year, release month, genre, director, budget and etc.

2. Analyses:

1) Exploratory analysis: Investigate the importance of temporal factors by looking at the trend of average/total movie revenues over years. I can pursue this goal by computing the mean and standard deviation of revenues (y-axis) for each year (x-axis) using data.table and plot the relationship using ggplot (box plots or lines with 95% confidence Intervals). Or investigate the distribution of movie revenues for each year using histograms.
2) Exploratory analysis: Similar to the analysis in 1) above but for month-of-the-year effects. (box plots)
3) Exploratory analysis: Investigate the trend of the relationship between revenue and genre by plotting the mean/total revenue of each genre over years. The x-axis will be genres, y-axis is the corresponding mean/total revenues and year decides the facet (bar plots). Or the x-axis is year, y-axis is revenue and color shows genres (line charts).
4) Exploratory analysis: Investigate the trend of the relationship between revenue and budget over years. The x-axis is budget, y-axis is revenue, and year decides facet (scatterplots).
5) Exploratory analysis: Investigate the importance of movie directors to revenues over years. The x-axis will be the top 25 movie directors, y-axis shows revenues and year decides the facet (bar plots).
6) Exploratory analysis: Similar to the analysis in 5) above but for production companies (bar plots).
7) Exploratory analysis: Discuss the relationship between ratings, i.e., vote average and revenues, and its trend over years. The x-axis is rating, y-axis is revenue, and year decides facet or color (scatterplots).
8) Exploratory analysis: Similar to the analysis in 7) above but for vote count (scatterplots).
9) Exploratory analysis: Similar to the analysis in 7) above but for popularity (scatterplots).
10) Machine Learning: Investigate the relationship between revenues and the factors discussed above including year, month, genre, budget, and so on by machine learning methods such as random forest classifier. The response variable is revenue, and explanatory variables are year, month, genre, budget, rating, vote count, popularity. I will use variable importance plot to see variables' significance, where the x-axis shows mean decrease Gini and y-axis is variable name.