Individual Project Proposal

Li Zhaozhi (李兆智)

November 9, 2020

My preferred topic for this research / analytics project would be the movie industry. IMDB is a reliable source for movie data that include movie titles, genres, ratings, box office, country of origin, and much more. The IMDB website offers up-to-date movie data at [www.imdb.com/interfaces](http://www.imdb.com/interfaces) for non-commercial use.

Kaggle provides IMDB data at <https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>. These datasets include the required number of categorical and numerical variables, over 50 observations, and data definition tables that will fulfill the requirements for the data source of this project.

Some questions I would like to explore with these data include 1) the average ratings and box office of each genre, 2) which genre tends to get higher ratings and box office, 3) how are ratings and box office correlated, 4) what are the highest rated adult and non-adult movies, and so on. I plan to use Excel, Power BI, and Tableau, etc. and any additional tools deemed helpful to analyze the data.

The biggest challenge would be data transformation. These datasets may contain missing values and multiple spreadsheets that will require some data cleansing work before performing any analysis. Certain texts in the datasets may have character encoding issues. My ability to transform these data will have an impact on the quality of my analysis.

I will use the internet, the textbook, the recorded lectures, as well as Prof. Borchardt for assistance when necessary.