Project Proposal PPOL670

Xun Zhang (xz461)

The analysis you aim to generate:

Going to cinema and watching a film is an important entertainment activity in people's daily life. In recent years, the global box office has grown rapidly. How to make more blockbuster movies is the primary concern of investors. Therefore, box office analysis and forecasting have emerged.
The focus of this case is to use the R to read box office data and perform descriptive and exploratory data analysis. Specifically includes: the relationship between the box office of the film and its type, the relationship between the box office of the film and the starring / director, to what extent does the early publicity investment matter, how the release date is selected, and the competition between the films showing on the same period

Data source you intend to use:

https://www.boxofficemojo.com/
https://www.imdb.com/

Your plan to obtain that data:

Download data I need from the website and store it in the file in CVS format, then let R read the file .

The methods (learned in class) that you aim to employ:

Data wrangling:
Do data cleaning and deal with missing data; convert the structure to facilitate some analysis
Visualizations:
Employ ggplot to visualize, making data more readable and intuitive
Machine Learning:
Do analysis and make prediction using multiple linear regression; observe and try to interpret coefficients

What "success" means:

1. Make use of three methods above efficiently, codes run smoothly.
2. After visualizing, charts clearly communicate information and fully mobilize the reader's visual senses and instincts
3. The relationship has been explored comprehensively and regression model is interpreted thoroughly