CAPSTONE PROJECT I EXPLORATORY DATA ANALYSIS

Film Industry in Asia

Patrick Shan

Introduction

The previous sections involved cleaning and visualizing the datasets on the box office for China and Japan. For this exploratory data analysis portion, inferential statistics tests were used in determining which variables significantly affect the box office grosses. The statistical tests were performed in Python files in Jupyter Notebook.

Questions

1. Are there variables that are particularly significant in terms of explaining the answer to your project question?
2. Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?
3. What are the most appropriate tests to use to analyse these relationships?

Methods

In analyzing the box office of both countries, the Pearson correlation test and the independent T-test were used for comparing the different variables for each country. The main variables are tickets sold, number of screens, ticket price, and number of movies released. These variables are treated as independent variables whereas the total gross in each country is treated as a dependent variable. These variables were chosen because they were increasing alongside total gross in both countries, indicating that there may be a correlation between these variables and total gross. The Pearson correlation test was used to examine if the variables correlate with the total gross of each country. Each correlation test would produce both a correlation coefficient (which determines how closely the values correlate with each other) and a p-value (if it is less than significance level of 0.05, then it indicates that the test is significant). Furthermore, independent t-tests were used to determine if variables significantly differ by country. This test was chosen because of its compatibility with small sample sizes, which is rather fitting given how the datasets has values only for the years 2012 to 2016.

Outcomes

The correlation tests and independent t-tests offer some interesting insights into the overall effects of the different variables affecting box office grosses for both China and Japan. The t-tests conclusively show that variables like tickets sold, number of screens, ticket price, and number of movies are significantly different between each country. In contrast, the Pearson correlation tests for each individual market offers more conflicting answers. In China, the number of tickets sold, number of screens, and number of movies share strong positive correlation with the China's total box office. In contrast, correlation tests involving the same variable categories and Japan's box office yield more muddled results. None of these variables appear to show strong positive correlation with Japan's total grosses and the test also have p-values greater than 0.05. As such, the results cannot conclusively determine if the variables affect box office gross in Japan. In summation, data analysis show that parameters differ by country and identified the variables that affect China's box office, yet the testing also is unable to draw definitive conclusions regarding the variables that affect Japan's box office.