CAPSTONE PROJECT I DATA WRANGLING

Film Industry in Asia

Patrick Shan

Introduction

For my project, I wish to study the box office growth in China and Japan. As such, the primary data source I gathered was from Kaggle, which details the box office statistics of China and Japan. Additional datasets on currency conversion, wages, and population growth were extracted from the Economist, Organisation for Economic Co-operation and Development (OECD), The Global Social Change Research Project, National Bureau of Statistics of China.

Compiling Data in Excel

As most of the datasets are available in excel files that I can edit, the majority of the data cleaning was carried out in excel. In two separate excel files for China and Japan, I uploaded the Kaggle dataset, which details the tickets sold, total gross, number of screens, and ticket prices for both countries from 2012 to 2016. I also added average wage information from OECD and Bureau of Statistics datasets along with population figures from the Global Social Change data. Given how the financial values were listed in Chinese Yuan (CNY) and Japanese Yen (JPY), I needed to convert those to US Dollars (USD) for direct comparisons. However, I couldn't directly convert these values to USD base on currency exchange rates because this process wouldn't scale to living standards and consumer purchasing power. As such, I used the purchasing power parity (PPP) from the Economist dataset to calculate more appropriately scaled values in USD. The values in USD were computed by dividing the values for each country by their appropriate PPP to get the adjusted values. Once these excel files were completed, they were uploaded to Jupyter Notebook in csv format.

Jupyter Notebook Compilation

In the Jupyter Notebook, I further cleaned and modified the dataset in a Python file and using the pandas library. The separate csv files for China and Japan were uploaded and converted into dataframes. I removed all commas from the value in the dataframes while also converting the values from string to numeric types for calculations. Once each dataframe has been polished, I concatenated the China and Japan dataframes into a single dataframe. Using this concatenated dataframe, I was able to carry out plotting and correlation analysis between the different columns of data.

Plotting and Correlation

Using the information gathered, I decided to plot the different columns against each other to identify potential trends. Plotting the total gross of Japan and China over time showed that that China has been experiencing greater growth compared to Japan. Furthermore, in China, there are more screens available and more tickets sold compared to Japan. However, while ticket prices and average wages are higher in Japan than in China, prices and wages in both countries have not experienced any increases over time. To further explore the effects of ticket prices and sold ticket quantity, I conducted several correlation tests pitting total gross versus either tickets sold or ticket price. Interestingly, for both countries, there was a near one-to-one positive correlation between number of tickets sold and total gross. However, there was no consistent or strong correlation between ticket price and total gross. As such, it may be possible that the main driving factor of box office revenue growth is increased ticket sales and access to movie screenings.