Film Industry in Asia Final Report

Patrick Shan
patricklshan@gmail.com
(650) 862-7528
June 8th, 2019

I. Introduction

Problem

The cinematic box office in the US has been rather inconsistent since 2008 owing to growing audience disinterest and competition from other forms of entertainment. However, the overseas box office is consistently growing in China and Japan. Studying the underlying factors behind this growth trend would be of interest for the film industry and its partners. Figuring out which factors would significantly affect box office numbers would help guide business strategies to promote and capitalize on the growing industry.

Approach

This project would focus primarily on comparing the parameters that affect box office in Japan and China. The initial dataset for each country contains the parameters of the year, number of ticket sold, number of screens, average ticket price, and number of new movies from 2012 to 2016. Furthermore, additional variables were added such as gross per capita, tickets sold and ticket price per year and etc. Through correlation tests and linear regression, these variables would be plotted against either each other or the net gross via linear regression models to determine if there is any correlation.

Furthermore, additional this dataset will be supplemented with additional datasets involving purchasing power parity, average income, and population sizes in both China and Japan. The purchasing power parity will determine costs of tickets in these countries compared to the US. The average income can determine whether citizens have enough disposable income to spend on movies. Finally, the population size can determine whether there is a growing population that can become consumers. Comparing these additional datasets can determine whether the increase in box office revenue is driven by currency changes, ticket prices, or disposable income levels.

1.3 Clients and Applications

The primary clients who would benefit from this research are the film industry and its partners like theater chains and marketing firms. Deciphering trends and understanding the driving forces could guide the film industry in their approaches to making and releasing film. For instance, if the growing box office gross is driven by an increase in the number of tickets sold and number of screens, then it would indicate that the growth is driven by greater outreach to audiences; subsequently the right approach to promote this growth is to focus investment on more theaters and appealing to a wide audience. Alternatively, if growth is driven by ticket price, then the industry should focus on catering towards a more niche audience, providing premium services, or raising ticket prices. Most notably, advertising firms could adjust between either a mass marketing or a niche marketing approach depending on the type of audience driving box office growth.

Understanding trends would be vital for determining the right business decisions to capitalize on a growing cinematic trend.

II. Data Collection and Wrangling

Data Sources

Although main datasets are primarily focused on film industry figures, they were supplemented with addition data sets focused on price adjustment, wages, and population. The box office dataset for China and Japan was sourced from the website Kaggle, and is a time-series that contained the variables of number tickets sold, ticket price, number of screens, and number of new movies from 2012 to 2016. A Purchasing Power Parity (PPP) index was provided using information from the Economist's Big Mac Index, which dictates the appropriate pricing and financial value relative to the US dollar. The annual wages for each country was provided by Organization for Economic Co-operation and Development (OECD) and National Bureau of Statistics of China. The Global Social Change Research Project provided information on population size.

Data Cleaning Compilation

As most of the datasets are available in CSV format, the majority of the data cleaning was carried out in Microsoft Excel. The Kaggle dataset was downloaded as separate dataset files for China and Japan. Average wage information from OECD and Bureau of Statistics were added to the datasets along with population figures from the Global Social Change data. Given how the financial values like gross, ticket prices, and wages were listed in the local currencies of Chinese Yuan (CNY) and Japanese Yen (JPY), they needed to be converted to US Dollars (USD) for direct comparisons. However, a direct conversion based on currency exchange rates wouldn't work because this process wouldn't scale according to living standards and consumer purchasing power. As such, the values were scaled using the Economist’s PPP index for more approachable financial figures in USD. For each country's dataset, the values in USD were computed by dividing the values by their appropriate PPP to get the adjusted values. After the adjustment, these datasets can be uploaded to Jupyter notebooks for analysis.

Jupyter Notebook Compilation

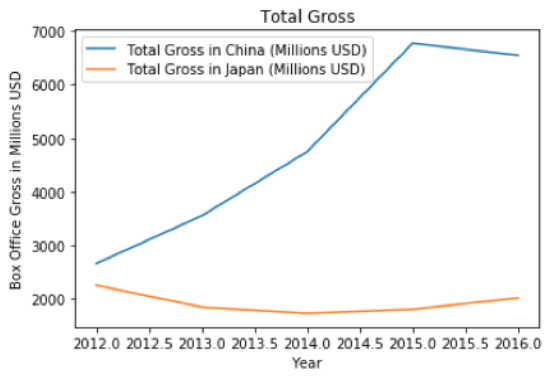
The Excel files were further and cleaned in Python format files of Jupyter Notebook. Jupyter Notebook was used as the main platform for Python analysis since it supports the usage of different modules. The separate files for China and Japan were uploaded into Jupyter notebook as CSV and converted into data frames with pandas. The commas were removed from the values in the data frame, and the values were converted from string to numeric types for. Once each data frame has been polished, they were concatenated into a single data frame for simplicity.

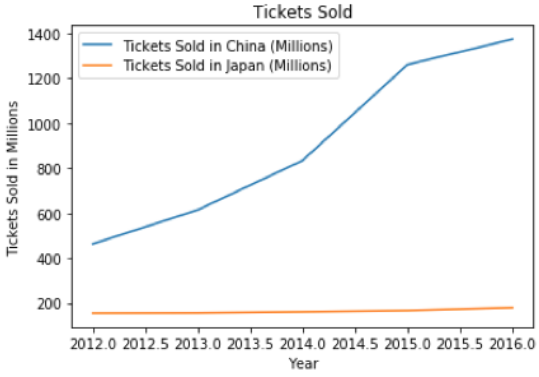
III. Exploratory Data Analysis

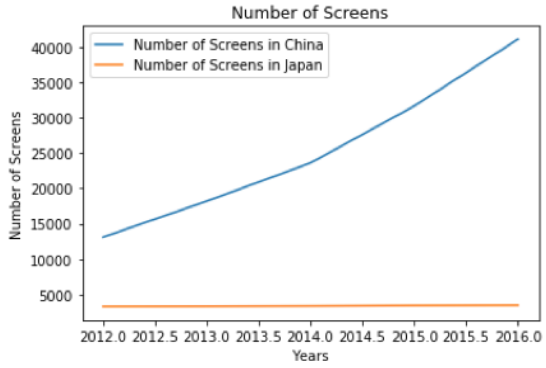
The main hypothesis for the exploratory data analysis is that box office growth is driven primarily by increased attendance. From 2012 to 2016, the ticket prices of both China and Japan do not seem to be increasing during this time period. At the same time, the number of the number of screens and number of tickets sold appear to be rising, which is backed by societal trends as both China and Japan are among the most populous and wealth nations that are capable of supporting a sizable audience who attend movies *en masse*. As

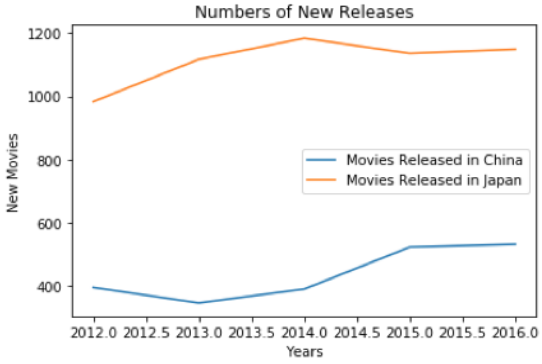
such, it could be possible that the box office revenue is driven by audience attendance and not by increasing prices.

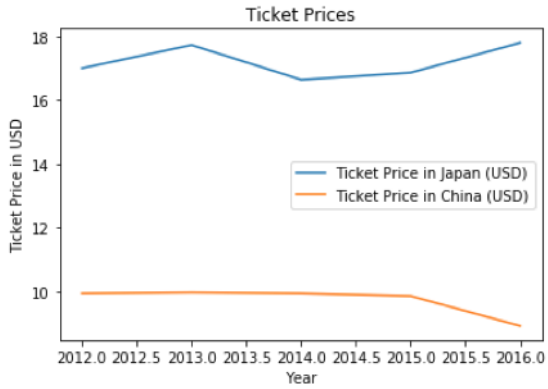
Plotting Variables Over Time





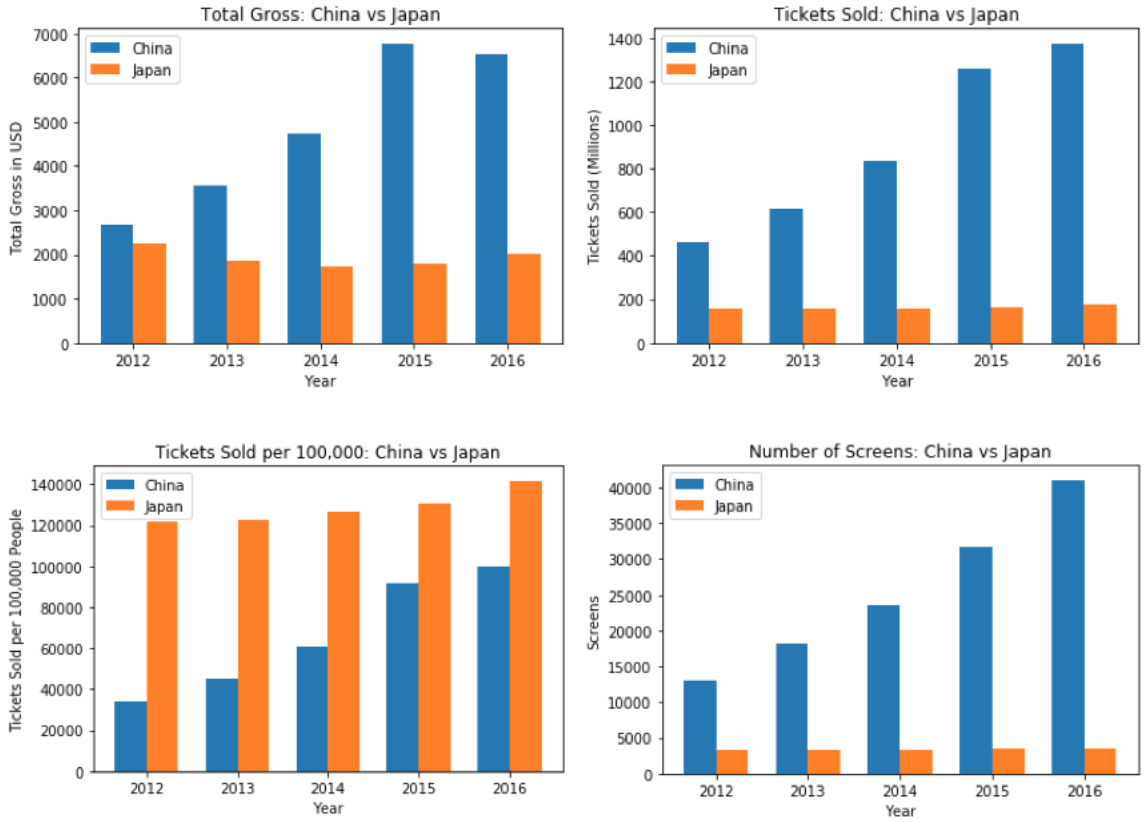






Line graphs showed the countries' total grosses, ticket sales, ticket prices, number of screens, and the number of new releases over time. Not only were total gross, the number of tickets sold, and the number of screens higher in China than Japan, but also that China was experiencing growth in these areas whereas Japan was stagnating in these areas. Ticket prices in Japan are higher although neither country experienced consistent ticket price increases. While Japan releases more new movies than China though both countries are showing more new installments each year.

Comparing Variables of Countries



In addition to the line graphs, bar charts were used for directly visualizing the disparities between the countries. The bar charts show the differences between the countries in gross, tickets sold, tickets sold per hundred thousand people, and number of screens. China and Japan's grosses are closely matched in 2012 yet began diverging from each other over time. China has been continuously selling more tickets than Japan yet Japan proportionally sells more tickets relative to population, although China is closing this gap. The number of screens has been increasing in China unlike in Japan. These charts show that not only does China surpass Japan in most film industry metrics, but also the disparities between the countries are increasing over time.

Comparing Variables of Countries

Correlation heat maps aided in identifying the variables that correlate with box office gross for each country. The heat map for China shows that China's box office gross shows strong positive correlation with the number of screens, number of tickets sold, and number of movies released while showing negative correlation with ticket prices. In contrast, the heat map for Japan shows that only ticket price seem to correlate with box office gross in that country. All the other variables seem to negatively correlate with Japan's box office.

Visualization Summary

The visualization of the data offers some insights in potential trends that could be further explored with statistical analysis. The line graphs show that the different parameters of the China data have been increasing over time while these same parameters have been largely stagnant in Japan data. The bar graphs show that not only is China outpacing Japan in many metrics, but also that the gap between the countries has been gradually increasing. The color-mapping graphs showed that increased box office returns has a strong positive correlation with ticket sales and the number of screens in China while showing a weak correlation in regards to ticket prices; conversely, the only variable that

positively correlates with Japan's gross is ticket price. As such, these results show that not only is China's movie industry growing faster than Japan's, but also it would suggest that the increasing box office returns in these countries are driven primarily by the increased turnout. These trends could then function as starting points for inferential statistics.

Inferential Statistics

Although the visualization seems to suggest possible trends, inferential statistics are needed to mathematically confirm any concrete relationships between the variables. Independent T-test determines whether there are significant differences between countries regarding their respective parameters of gross, number of tickets sold, number of tickets sold per hundred thousand people, number of screens, ticket price, and number of new movie releases. The T-test was chosen over the Z-test because of its compatibility with small sample sizes, which is rather fitting given how each parameter contains data contains only values from the years 2012 to 2016. If the T-test yields a p-value less than 0.05, then the values are significantly different and we can reject the null hypothesis that there is no difference between the countries.

The Pearson correlation test examines 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). A strong correlation value would indicate that the variable might play a key role in affecting the box office gross. The dependent variable is the respective gross in each country and the independent variables and number of tickets sold, number of tickets sold per hundred thousand people, number of screens, ticket price, and number of new movies released.

Table 1. China vs. Japan T-Test. This table shows the results of t-tests of the different parameters between China and Japan. If p-value is less than 0.05, then the parameter for each country is significantly different from each other.

Parameters	T-Statistic	P-Value
Grosses	3.594	0.007
Tickets Sold	4.196	0.003
Tickets Sold per Hundred Thousand People	-4.678	0.002
Screens	4.474	0.002
Ticket Price	-24.120	0.000
New Movies Released	-13.231	0.000

Table 2. Parameters vs. Total Gross. This table shows the results of conducting Pearson correlation test between the variables. If the coefficient is close to 1.0, then the correlation is strong and positive. If the coefficient is close to -1.0, then the correlation is strong and negative. If the p-value is less than 0.05, then it indicates that the correlation is statistically significant.

Variable 1	Variable 2	Coefficient	P-Value
Tickets Sold in China	Total Gross in China	0.986	0.002
Tickets Sold per Hundred Thousand People in China	Total Gross in China	0.987	0.002
Screens in China	Total Gross in China	0.937	0.019
Ticket Price in China	Total Gross in China	-0.587	0.298
New Movies Released in China	Total Gross in China	0.885	0.046
Tickets Sold in Japan	Total Gross in Japan	-0.077	0.902
Tickets Sold per Hundred Thousand People in Japan	Total Gross in Japan	-0.076	0.903
Screens in Japan	Total Gross in Japan	-0.310	0.612
Ticket Price in Japan	Total Gross in Japan	0.219	0.723
New Movies Released in Japan	Total Gross in Japan	0.877	0.051

The inferential statistic testing offers insights into the data parameters. When comparing the parameters between the different countries, they are statistically significant with the independent t-tests yielding p-values of less than 0.05. Furthermore, the Pearson correlation showed that for China, the number of tickets sold, number of screens, and number of movies share strong positive correlation with the China's total box office; their coefficients range from 0.885 to 0.986, indicating strong correlation, while their p-values of less than 0.05 indicate that such correlation is statistically significant. Interestingly, ticket price is the only variable that negatively correlates with China's box office though this relationship is of dubious value given its p-value is greater than 0.05. As for Japan, only ticket prices and number of new releases positively correlate with gross yet even this correlation is rather questionable given how the p-values are greater than 0.05.

EDA Summary

The visualization offers some interesting insights into the box office trends in these countries. The line graphs show that many of the parameters and variables in the China data set appear to be increasing over time. In contrast, the variables in the Japan box office are static. The bar graphs show that the two countries have noticeable disparities in the value of the parameters; not only has China been outpacing Japan in many parameters like box office gross and number of tickets, but the gap in parameter values has been increasing over time. A heat map showed that China's box office gross positively correlates with number of tickets sold, number of tickets sold per hundred thousand people, number of screens, and number of new movie releases. In contrast, Japans' box office gross only positively correlates with ticket prices. These graphs show that there are trends present in the data and that the two countries have distinct patterns.

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 having 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.

IV. Machine Learning

Hypothesis

The machine learning processes of linear regression and R^2 analysis will support the trends found in the EDA portions. The EDA show that total gross in China positively correlates with number of tickets sold, number of screens, and number of movies released. In contrast, the total gross in Japan only positively correlates with ticket prices. As such the linear regression models should yield a positive fitted linear regression with the parameters of China's data while yielding an negative fitted linear regression with Japan's data.

Methods

The machine learning process of the data is comprised of linear regression and R^2 analysis. Linear regression offers many advantages like the flexibility in working with smaller datasets, the ability to provide a fitted regression line showing relationship between independent and dependent variables, and the expediency of providing results. This linear regression is used to test whether the independent variables, which are the box office parameters, show correlation with the dependent variable of box office gross. The linear regression was conducted on the individual parameters separately and then with all parameters together in a single aggregate regression. The aggregate linear regression could then be used to generate a linear equation with the box office parameters as variables.

After performing linear regression, R^2 analysis was used in determining how closely the predicted data points adhere to the fitted regression. The independent and dependent variables are each split into training and testing portions with the training portions becoming fitted to a regressor; the test size used for the split is 0.3. The regressor in turn is used for predicting the output dependent variable based test portion of the independent variable. These values are then used for computing R^2 score and Residual Mean Square Error (RMSE); an R^2 value closer to 1 would indicate close adherence to the fitted regression while a low RMSE score indicates that there is little deviation. R^2 means error

(RSME) of relatively low value would suggest that the predicted data reliably adheres to the fitted regression pattern.

Results

For the machine-learning portion, linear regression was performed on individual and aggregate parameters for both countries. Regression for individual parameters seems to confirm that the number of tickets sold, number of screens, and number of movies released correlate with a positive increase in box office in China. Conversely, ticket price seems to induce a decrease in box office. Aggregate regression for China's parameters show that the positive parameters increase with box office while negative parameter (ticket price) decreases with box office. Conversely, for Japan, both the individual and aggregate regression show that the number of tickets sold, number of screens, and number of movies released decreases with box office while ticket prices increases with box office. The linear equations for each country's box office, which are calculated from the coefficients of the aggregate regression, show the hypothetical box office yields.

China's box office: *Total Gross* = 1.5468 x *Tickets Sold (Millions)* + 0.1453 x *Number of Screens* - 0.0009 x *Ticket Price (USD)* + 1.4862 *Movies Released* - 549.2727

Japan's box office: *Total Gross* = -0.0420 x *Tickets Sold (Millions)* - 0.5222 x *Number of Screens* + 0.0049 x *Ticket Price (USD)* + -2.4207 *Movies Released in Japan* - 6356.8774

To determine the reliability of the linear regression, R^2 and RMSE scores were calculated based off of the training and testing data. For China's aggregate box office, the R^2 is 0.296 and the RMSE is 1258.38. Conversely, for Japan's aggregate box office, the R^2 is -4.064 and the RMSE is 4626.96. The positive R^2 value and smaller RMSE for China indicates that the regression for that country is more reliable than Japan's regression since it has a negative R^2 value and larger RMSE. As the R^2 is a relative measure of fit while RMSE is an absolute measure of fit, the RMSE is the main determining factor for the reliability of regression fit. It should be noted that the calculated RMSE values are relatively large compared to the ranges of box office gross. While a smaller RMSE would've been more preferable, these are chosen as smaller values would've lead to over fitting and skewed coefficients. Since the Japan's RMSE is larger than China's, it would suggest that the machine learning for the China data is more reliable.

Machine Learning Summary

The machine learning results seems to support the hypothesis. Based on linear regression and RMSE values, China's box office seems to be primarily driven by tickets sold, number of screens, and number of movies released. Conversely, Japan's box office seems to be driven by ticket prices. Furthermore, the RMSE value of China is lower than Japan's, suggesting that the China's machine learning model is more reliable. At face value, this would support the hypothesis and confirm the results of the EDA.

However, there are some limitations in the machine learning, much of it stemming from the limited data set size. Since the box office for China only covers between 2012 and 2016, any box office comparisons between China and Japan could only use data from those years. This limited data size, along with the time series nature, means that it is less

compatible with other supervised machine learning algorithms like naive Bayes, decision trees, and k-nearest neighbors. Future areas of improvement could focus on obtaining larger data sets, which would offer greater flexibility and reliability for machine learning.

V. Conclusion

Summary

The goal of this project is to the driving factors behind the growth of the film industry in China and Japan. Data visualization shows that China's film industry has been generally growing while Japan's is largely stagnant, leading to a wide discrepancy between the various parameters of the countries. Inferential statistics confirm that the box office parameters for each country are statistically significant, inferring that the two country's industries operate differently. The inferential statistics also identified clear variables responsible for the growth of China's box office while yielding inconclusive results for Japan. The machine learning further supports many of the patterns shown in the inferential statistics while also showing a projection for the future predictions.

The project identified several patterns that are behind the box office in China and Japan. In China, the box office growth correlates with increases in number of tickets sold, number of screens, and number of new movies. At the same time, prices appear to static and negatively correlate with box office in this region. These trends suggest that the box office in China is driven primarily by an increasing audience attendance as more people are going to the theaters. Conversely, the research suggests that rising ticket prices are the main driving force behind the box office in Japan since this is the only variable that positively correlates with Japan's grosses. However, the Japan data can be questionable given the resulting statistical analysis and machine learning.

Recommendations

1. There is a noticeable discrepancy in box office parameters between the countries. The gross and number of tickets sold have been increasing in China but are static in Japan. This suggests that Japan's film industry is actually stagnating while China's represents more promising growth. China would therefore represent a more financially lucrative market for the film industry in the foreseeable future.
2. China's box office grosses increases along with increased number of tickets sold, number of screens, and number of new movies. As such this seems to suggest that greater film attendance would lead to increased revenue as opposed to increased ticket prices. To capitalize on this trend, the industry should invest in additional theaters to promote accessibility. Furthermore, film advertising should go for a more broad appeal and reach out to as wide of an audience as possible.
3. Japan's box office grosses increases only in conjunction with ticket price. This would imply that the growth is driven more by changes in Japan's consumer good prices rather than increasing turnout. The lack of growing audience attendence .

Sources and Links:

Big Mac Index: https://www.economist.com/node/21569171
Film Industry in China and Japan: https://www.kaggle.com/clouds0715/thefilmindustry
China Average Income: http://data.stats.gov.cn/english/easyquery.htm?cn=C01

Japan Average Income:
https://stats.oecd.org/Index.aspx?DataSetCode=AV_AN_WAGE#
World Population: http://gsociology.icaap.org/dataupload.html
Github Portfolio: https://github.com/dinohunterpat/Capstone-Project-I