Carol Weng, Rong Shi, Alex Gao

Professor Chakraborty

CMSC 205 Data-Scientific Programming

Nov. 18th, 2021

Discussion:

In this project, we were examining the relationship between the response variables (i.e., boxoffice, average ticket price) and the explanatory variables (i.e., year, month) within the top 297 highest box office national and international films with release time ranging from 1998 to 2021. We obtained our data through web scraping from a Chinese professional and neutral box office statistics website (https://piaofang.maoyan.com/). Later we joined them in a tidy datatable in R studio and plotted a scatter plot with a trend line to help us observe the trend. Below is our interpretation based on the scatter plot and the trend line.

Box office + release year

- National films: Most films share similar box office 1e+05 ten thousand, as time goes (getting closer to 2021), there are more outliers with extremely high box office which drags the trend line upward
- International films: There are more variabilities in box office in international films as time gets closer to 2021. We are excluding the point representing the international film released in 1998 because there is a huge time gap between this film and the other films. We also observe an outlier with extremely high box office in 2019 (Avenger: Endgame).

- National & International: The variabilities in both national and international box office is increasing when the year gets close to 2021.

Average ticket price + release year

- National films: We observe a downward quadratic trend in relation between average ticket price and release year. We see the lowest average ticket price is around 2016 and 2017, and the average ticket price higher when time gets closer to 2021 or 2010. The highest average ticket price is in 2021.
- International films: We observe a similar trend in average ticket price with international films that the lowest average ticket price is shown in 2016. We excluded the outlier of international films released in 1998 because there is a huge time gap between this film and the other films.
- National & International: The trend in average ticket price is almost the same when we consider national and international films at the same time.

Box office + release month

- National films: In February, July, August, and September, we have multiple outliers in national films with extremely high box office which drags the average box office in those months higher than the rest. The average box office excluding these specific months is about the same across the year.
- International films: In international films, the box office is about the same. We can see that there are three outliers in April with high box office, and one of them is extremely high (Avenger: Endgame).

- National & International: When we consider both national and international films, the average box office is relatively stable across the year when compared to the individual average box office of national or international films.

Average ticket price + release month

- National films: In national films, the average ticket price throughout the year is about the same, however, we can observe that the average ticket price is the highest in February in all years.
- International films: In international films, although there are some outliers, the average ticket price is roughly stable.
- National & International: In both national and international films, the trend in average ticket price is pretty much the same as the one in national films.

In our interpretation, we found some interesting trends. When looking into the scatter plot with box office as the response variable and release month as explanatory, we found that, on average, the box office reached the peak in February and the whole summer (July, August, & September) every year. Based on our knowledge, we speculate that the reason for the difference in box office is because the Chinese Spring Festival is always in February and people will have a long break without any work during this time. During this long and relaxing festival break, watching a movie with the family is always a popular choice and also a convention. The total box office of films released during the Chinese Lunar New Year holiday reached 7.8 billion yuan ($1.2 billion), breaking the previous record of 5.9 billion yuan set in 2019. According to the website boxoffice Mojo, the cumulative U.S. box office haul has been just $86 million since the start of the year. And for the summer time, as we know that the 2-month long summer break for all students started around July and during this time movie theater is actually a cozy and relaxing place for a rest. Hence, there are many movies released during the period to make sure that the audience will always have new choices and oftentimes some of them will get an extremely high box office.

Report:

In this project, we have used programming skills such as website scraping, shiny app building, data joining, and graph plotting. For website scraping, we chose the "Cat eye professional version box office summary website" as our resource. We first scrap the release time as a whole and then use "str_split" to separate the year, month, and date information by "-". For shiny app building, we utilized the "checkboxGroupInput" and "varSelectInput" to create a sidebar that can provide the users choices to modify their desired x and y variables; we then created a scatterplot with a trend line to provide visualized information. After we built the shiny app, we then used statistical skills such as combining the data and interpreting the trend in graphs. When interpreting the data, we compared the different box office and average ticket price among different release years and release months between national, international, and combination of national and international films. Specifically, we looked into whether there are any outliers or any trends.

In the process of our data programming process, we met some problems.

First, with a 403 error, we must know the importance of permissions. Because without permission, we cannot scrape the data from the website. For example, when we were trying to scrape the grading and price for the products on the website “Sephora”, we found it is impossible for R to turn them into valid data because the website did not provide permission.

Moreover, many websites do not provide enough information that meets our needs. For example, we couldn’t find the categorical information from the second-hand trading website: “Vestiaire Collective”. We need to find a website that can provide all the desired explanatory variables that we need to interpret the change in the response variable.

Furthermore, even though we found the website that provides the desired explanatory variables, some of the variables cannot be turned into valid data after web scraping (the numbers were shown in boxes form). The way to solve this problem is manual input by ourselves.