Xi'an Jiaotong-Liverpool University
西交利物浦大學

School of Advanced Technology

Project 1 Report

Project Title: Web Scraping & Data Analysis

Student Name: Hangming Ye

Student ID: 1822766

Project field: Big Data Analytics

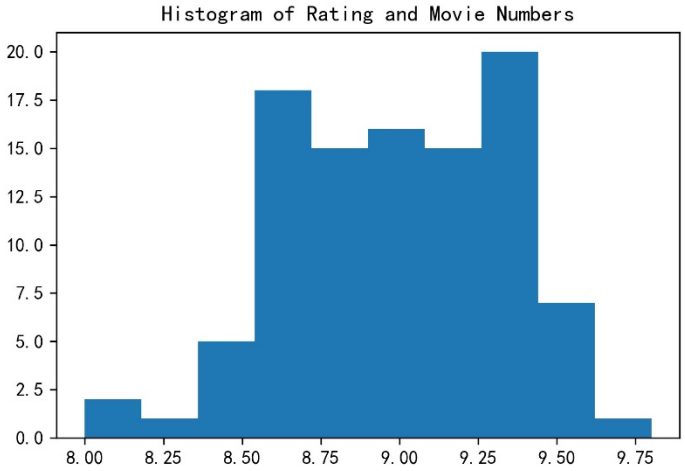
Supervisor: Jia Wang

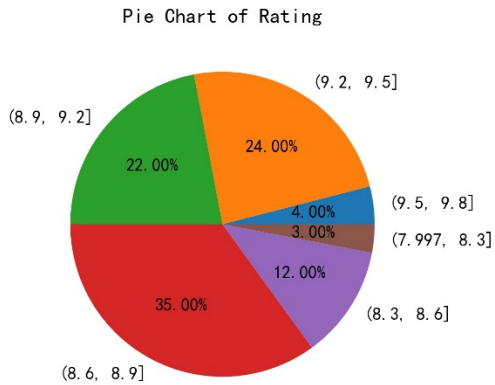
Co-supervisor (if applicable):

The anlysis is trying to find out the common features of a good movie (maoyan Top 100), and to figure out the element affect the rating of the movie. So that, it could be meaningful for the producer to make a good movie.

The attributes of the movie contain region, time, director, actor, 1-week cumulative income, all cumulative income, duration, type, rating. And each analysis will focus on one certain attribute.

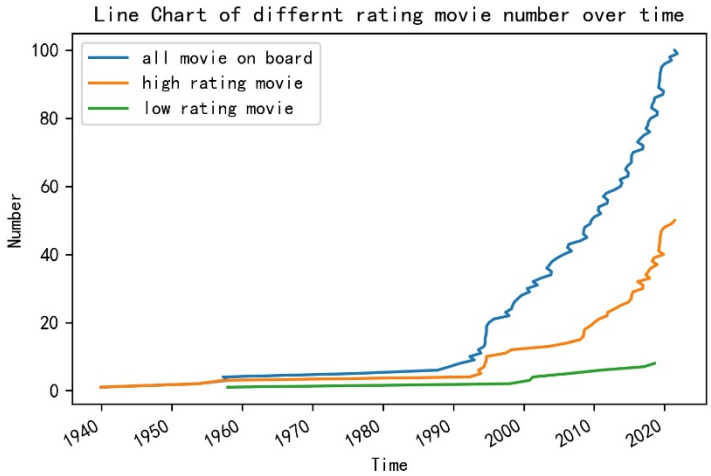
1. Distribution of the Rating

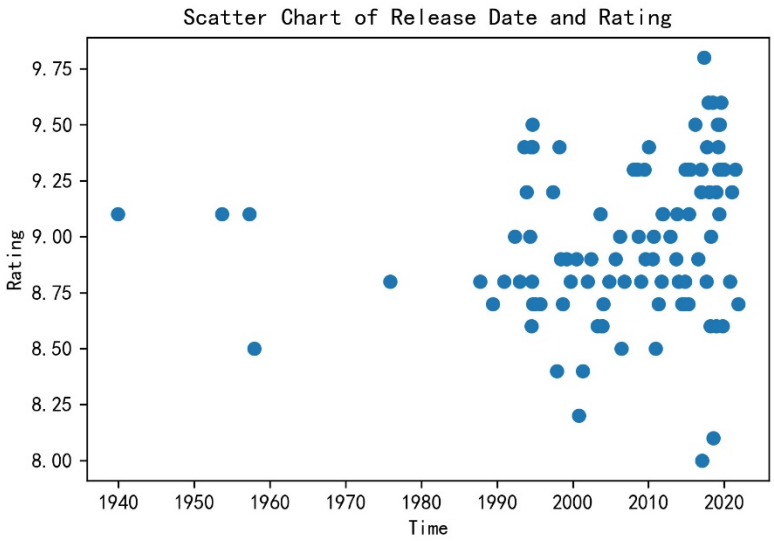




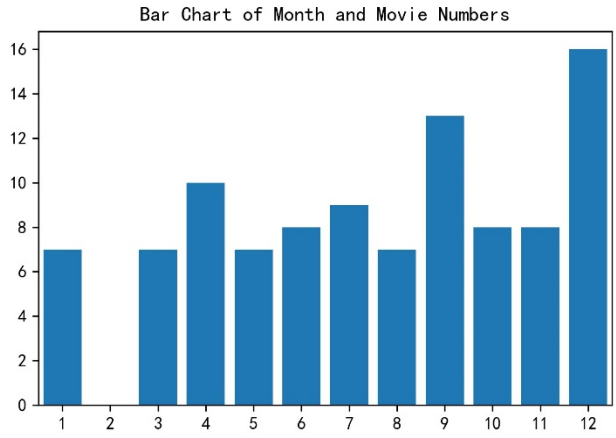
From the hist diagram and pie diagram we could see most popular movies have the rating between 8.5-9.5, and the number of extreme high mark and extreme low mark is low, if the score is too low, it is hard to go to the Top 100 list. Therefore, the movies meet with public opinions are more likely to succeed.

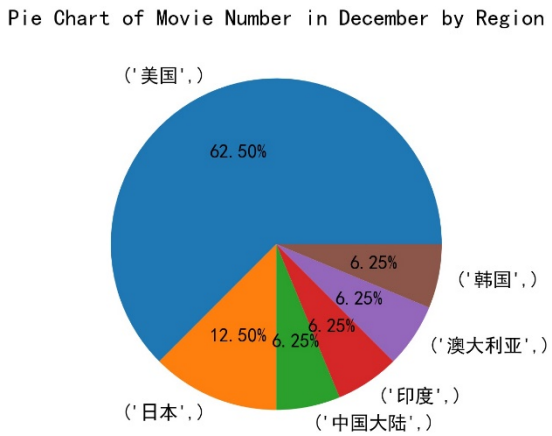
2. Time





From the scatter chart, it is easy to find that most movies in the Top 100 board are released after 1980. It reflects that people now prefers movie shoot by modern techniques. From the line chart, we could see that in the 2010-2020, the number of high rating (>9.0) movie have a significant increase However, the low rating (<8.6) movies also increases. Therefore, movies with modern technology, ideas or elements are more likely to succeed.

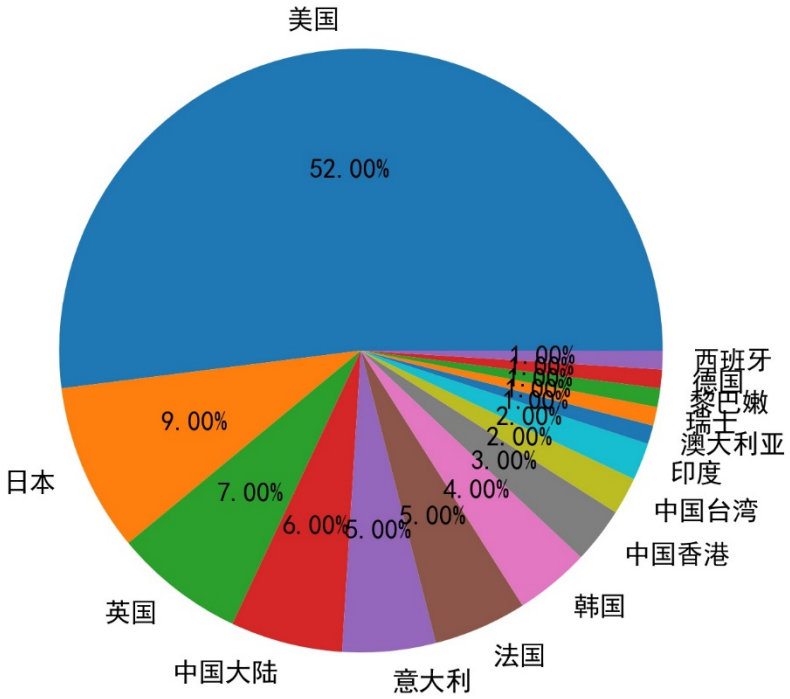


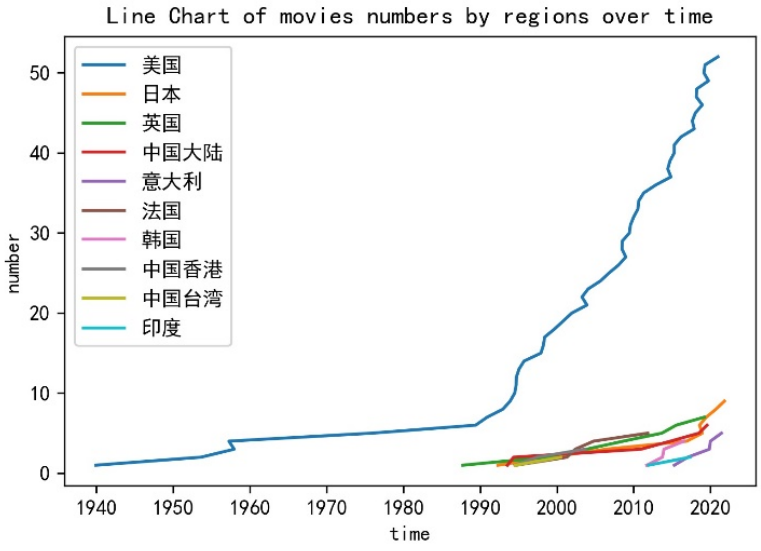


From the bar chart, we could see most of the Top movies are released in December, which is almost two times than other months. It may because the Christmas and other festivals are in December and the movie company want more incomes. The pie chart of movie number in December by region could prove this hypothesis. Therefore, movies released in festival season are more likely to succeed.

3. Region

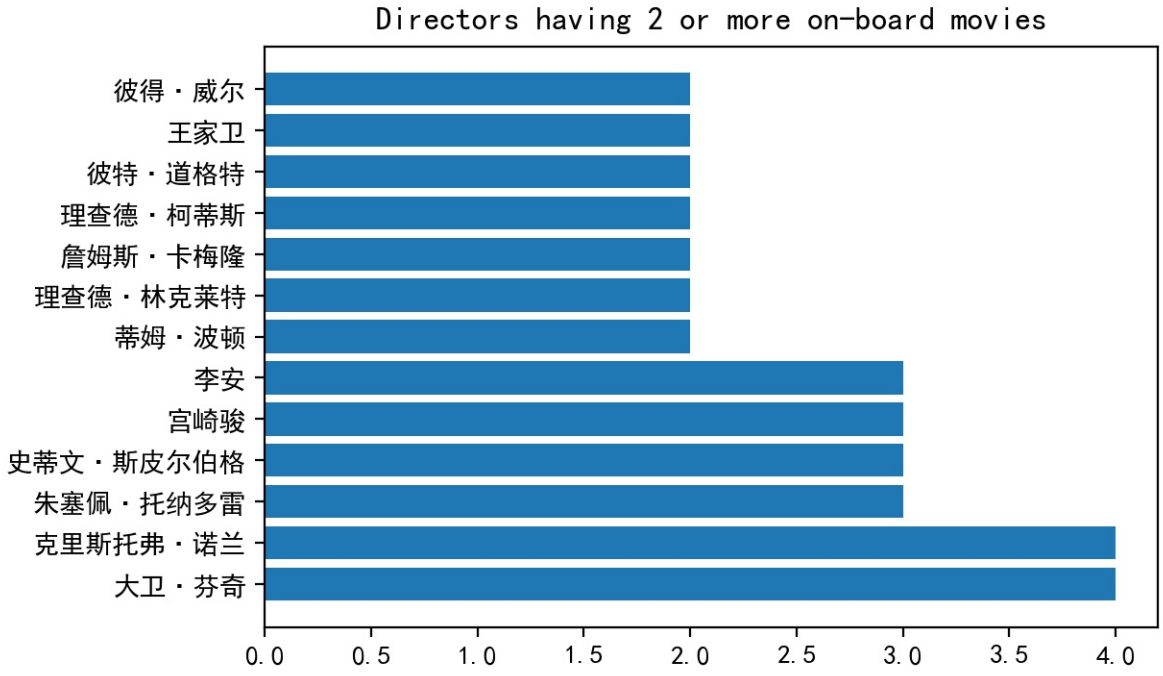
The region proportion of the Top 100 movies is shown by the Pie chart. The US produce most of the movies on board. And from the line chart, we could find that, in the early years, only US had the ability to produce high quality films. Since 1990, other country started producing high quality movies, but the US film industry are still the most powerful, from both amount and growth rate.



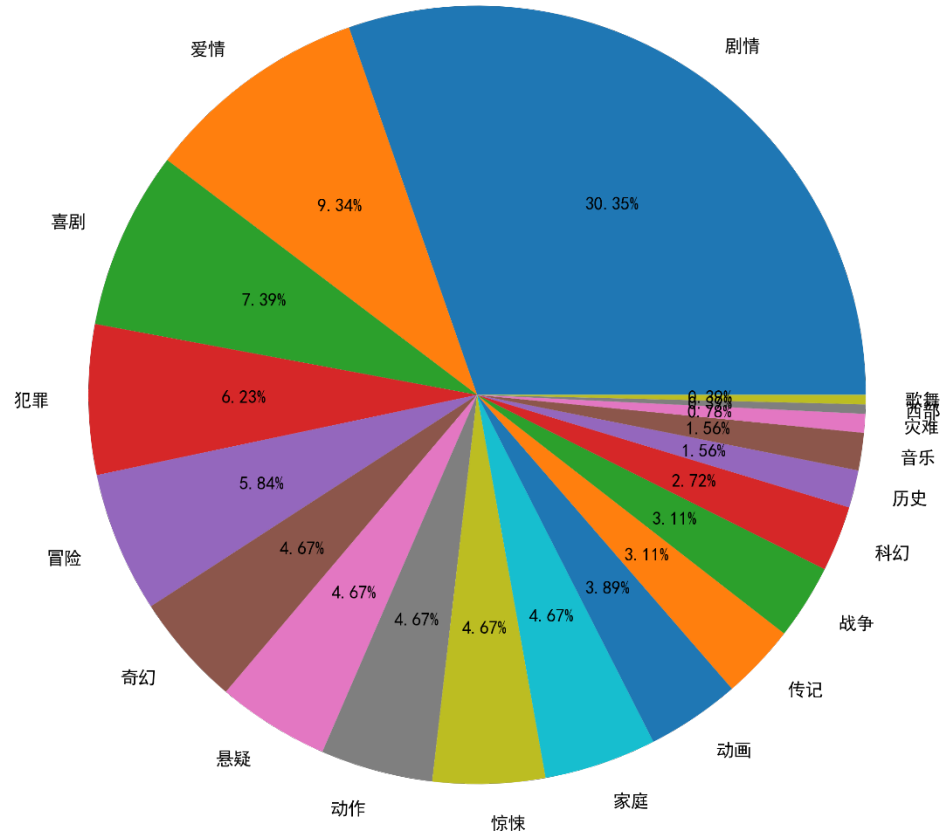


4. Director

Some directors have more than 1 on-board movies. These directors are experienced and could produce high-quality movies in the future.

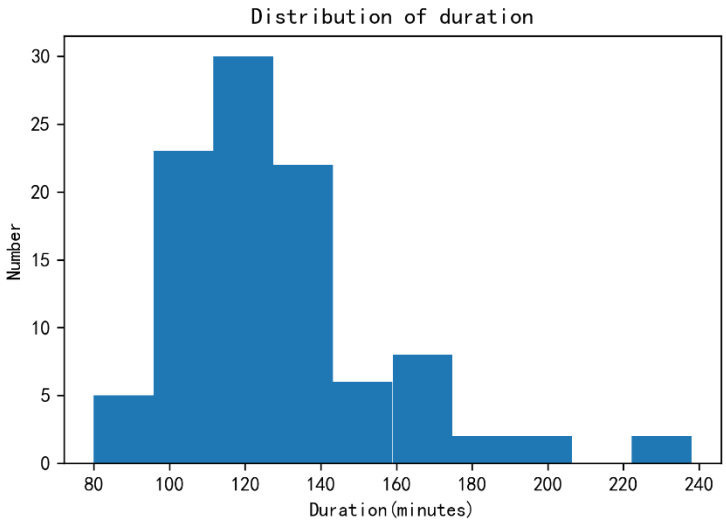


5. Type



Firstly, most of the on-board movies are drama. And romance, comedy, crime elements are also popular. The musical, disaster, dancing, historical elements have rare audience.

6. Duration



From the diagram, we could see most of the movies have a duration of moderation, 100 to 140 minutes is most on-board movies' duration.

Conclusion

To sum up, a movie contains following elements is more likely to get high marks and become popular:

1. Meet with public opinions
2. Using modern technology, ideas or elements
3. Released in festival season
4. Produced by the region with good conditions
5. Using some popular movie type
6. Directed by experienced director
7. Having a reasonable duration

This is the conclusion using the above figure and the "Mao Yan Top 100" data. Limited to the amount of data, it is not definitely correct. Some anlysis and figures are not in the report since they are not decisive. I will put them in the notebook and hope bring some inspiration.