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

School of Advanced Technology

Project 1 Report

Project Title: Assignment 1: Web Scraping & Data Analysis

Student Name: Tianlei Shi

Student ID: 1824152

Project field: Big Data Analytics

Supervisor: Jia Wang

Co-supervisor (if applicable):

1. Introduction

To collect data on a specified topic and analyze it to produce reliable, revealing results, which is always been an important ability, and particularly in today's data-rich world. Therefore, in this project, we scraped data from the TOP 100 movies of Maoyan Movies (https://maoyan.com/board/4) by using python, and then analyzed the data in detail, and obtained enlightening results.

2. Web Scraping and Data Analysis

2.1 Web Scraping

TOP 100 movies of Maoyan Movies means that the 100 most popular (or classic) movies in China. In the beginning, we scraped 10 features of movie from Maoyan Movies TOP 100, which includes ranking, title, name of actors, release date, rating, type, duration, region, name of director, and cumulative income. The example data is shown as Fig.1.

	Rank	Title	Name of actors	Release date	Rating	Type	Duration	Region	Name of director	Cumulative income
0	1	我不是药神	徐峥,周一围,王传君	2018/7/5	9.6	剧情，喜剧	117分钟	中国大陆	文牧野	310002万
1	2	肖申克的救赎	蒂姆·罗宾斯,摩根·弗里曼,鲍勃·冈顿	1994/9/10	9.5	剧情，犯罪	142分钟	美国	弗兰克·德拉邦特	暂无
2	3	绿皮书	维果·莫腾森,马赫沙拉·阿里,琳达·卡德里尼	2019/3/1	9.5	剧情，喜剧，传记	130分钟	美国	彼得·法雷里	47872万
3	4	海上钢琴师	蒂姆·罗斯,比尔·努恩 ,克兰伦斯·威廉姆斯三世	2019/11/15	9.3	剧情，爱情，音乐	126分钟	意大利	朱塞佩·托纳多雷	14376万
4	5	哪吒之魔童降世	吕艳婷,囧森瑟夫,瀚墨	2019/7/26	9.6	动画，喜剧，奇幻	110分钟	中国大陆	饺子	66507万
5	6	霸王别姬	张国荣,张丰毅,巩俐	1993/7/26	9.4	剧情，爱情	171分钟	中国香港	陈凯歌	5万
6	7	小偷家族	中川雅也,安藤樱,松冈茉优	2018/8/3	8.1	剧情，犯罪	121分钟	日本	是枝裕和	9675万
7	8	美丽人生	罗伯托·贝尼尼,朱斯蒂诺·杜拉诺,赛尔乔·比尼·布斯特里克	2020/1/3	9.3	战争，剧情，爱情	116分钟	意大利	罗伯托·贝尼尼	5979万
8	9	这个杀手不太冷	让·雷诺,加里·奥德曼,娜塔莉·波特曼	1994/9/14	9.4	剧情，动作，犯罪	110分钟	法国	吕克·贝松	暂无
9	10	盗梦空间	莱昂纳多·迪卡普里奥,渡边谦,约瑟夫·高登-莱维特	2010/9/1	9.0	动作，悬疑，惊悚，科幻	148分钟	英国	克里斯托弗·诺兰	49620万

Fig.1. Example data of scraped data

The questions I am interested in about these data are: what the base situations of these 100 movies are, whether audiences would prefer new movies or old ones, whether audiences like the movies of certain directors more than others, and whether there is some implicit relationship between the type, rating, ranking, cumulative income of movie.

2.2 Data Analysis

2.2.1 Base Situations

Firstly, we drew a word cloud to show the frequency of occurrence for movie types (shown as Fig.2.), as well as a pie chart to represent the distribution of TOP100 movie-shooting countries (shown as Fig.3.).

Fig.2. Word cloud of TOP100 movie types

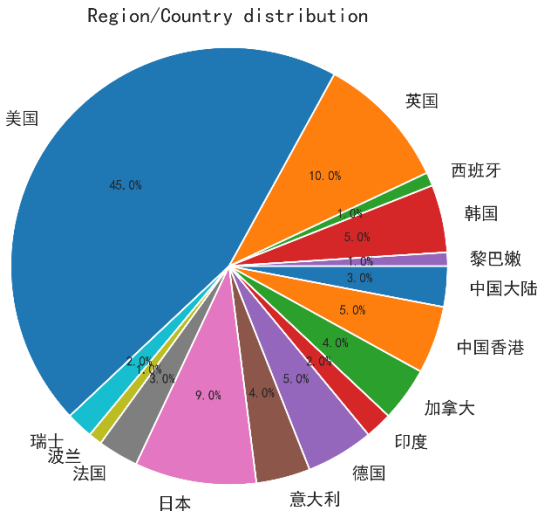

Fig.3. Distribution of TOP100 movie-shooting countries

As word cloud shows, there are 19 types of TOP100 movies, of which *drama* movies dominate the list with an overwhelming number, and which means drama movies are the most numerous and most appreciated by audiences. What's more, type of *romance*, *comedy*, and *adventure* movies are also popular. However, movies type of *musical*, *westerns*, and *disasters* are the least numerous, probably because they attracted fewer audiences and were less likely to become classics. Additionally, as pie chart shows, the *United States* produced the most TOP100 movies, accounting for 45%, and *Spain* and *Lebanon* produced the least, accounting for 1%. It is worth mentioning that *China* accounts for 8% (including *mainland China* and *Hong Kong*), this means that China's film industry still has much room for development.

2.2.2 Era Preferences

To analyze the audience's preference for the movie age, we drew a line chart to show the number of selected movies for each era, shown as Fig.4. The line chart clearly shows that most of the TOP100 movies were released after 2000, suggesting that audiences are more receptive to modern movies, and it is not true that older movies are more classic.

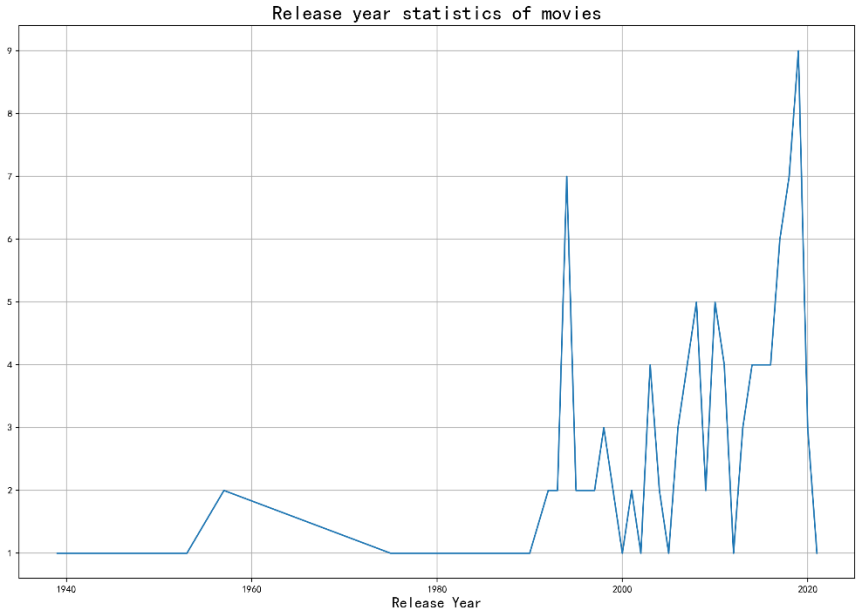

Fig.4. Selected movies in each era

2.2.3 Director and Cumulative Income

Cumulative income is an important factor in evaluating movies and directors. It seems that director will affect the box office and that people prefer to watch movies directed by certain

directors. Thus, to verify this viewpoint, we drew a bar chart showing the director's cumulative box office of all movies in Maoyan TOP100 (shown as Fig.5.).

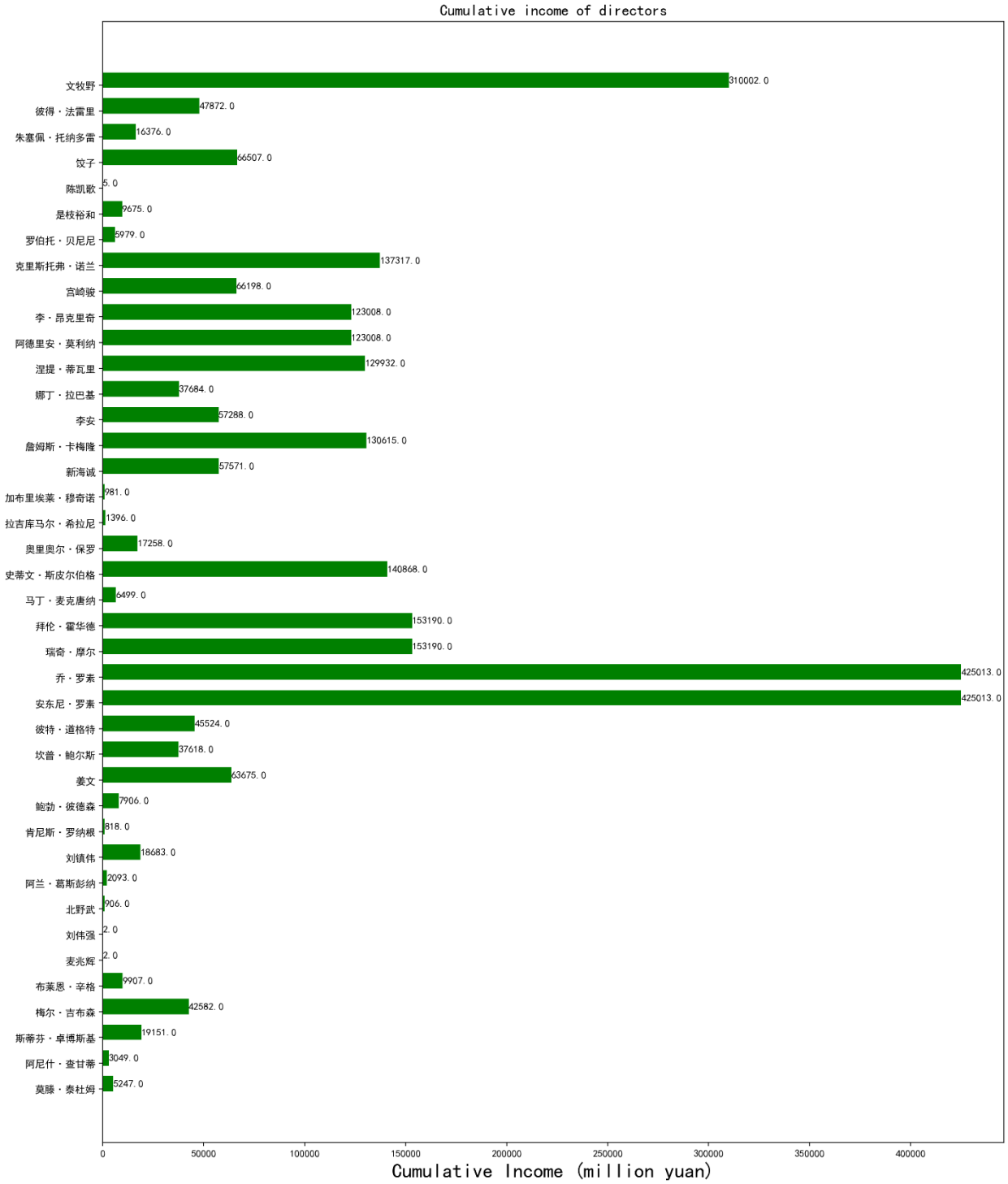

Fig.5. Cumulative income of directors

The three directors with the highest cumulative incomes are Joe Russo, Anthony Russo, and Muye Wen. Joe Russo and Anthony Russo tied for first place among directors with their film *Avengers: Infinity War*, which cumulative income is 425,013 million yuan. Muye Wen came in third with his film *Dying to Survive*, which earnings is 310,002 million yuan. Above directors earned a high box office with only one film. In contrast, the two films of famous director Christopher Nolan *Inception* and *Interstellar* earned 137,317 million yuan combined, it's only a third of Joe Russell's. Therefore, the director's influence on the cumulative income is not decisive, and audiences pay more attention to the quality.

2.2.4 Relationship between Rating, Type, and Cumulative Income

To analyze the relationship between rating, type, and cumulative income, we drew a three-dimensional scatter chart shown as Fig.6. As the scatter chart shows, the Maoyan TOP100 movies are almost all concentrated in the middle and lower part of the 3D space. Most of the movies are rated between 8.75 and 9.50, and the box office of most movies is less than 100,000 million yuan.

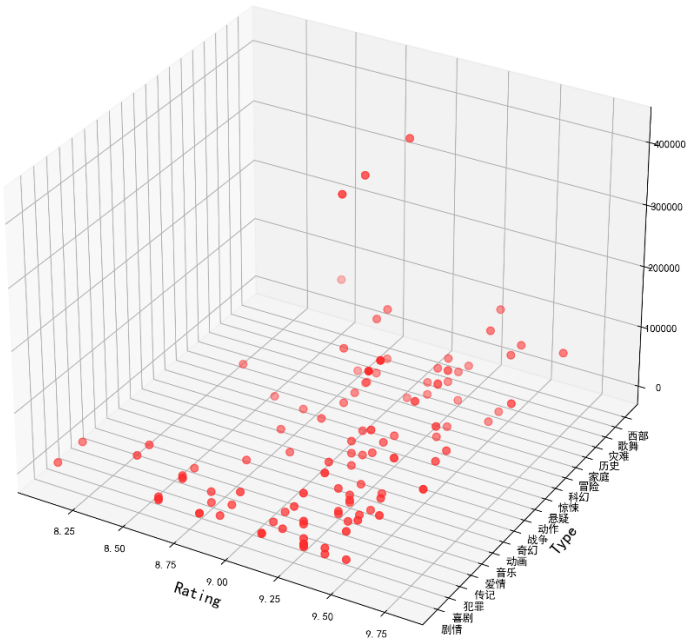

Fig.6. Relationship between rating, type, and cumulative income

2.2.5 Relationship between Rating, Rank, and Cumulative Income

Finally, we drew a three-dimensional line chart to represent the relationship between rating, rank, and cumulative income, and it shown as Fig.7. By analyzing the projection of the original image, we can understand the relationship between the three more clearly. According to the projection on the *rank-rating* plane, we can find that top-ranked films do not necessarily have higher rating, and there is no special relationship between rating and rank. According to the projection on the *rank-income* plane, we can see that films ranked in the top 60 have higher box office than others, so the rating and income of films are positively correlated to some extent. Moreover, according to the projection on the *rating-income* plane, we can know that films with higher rating also have higher box office, so there is a positive correlation between the two factors.

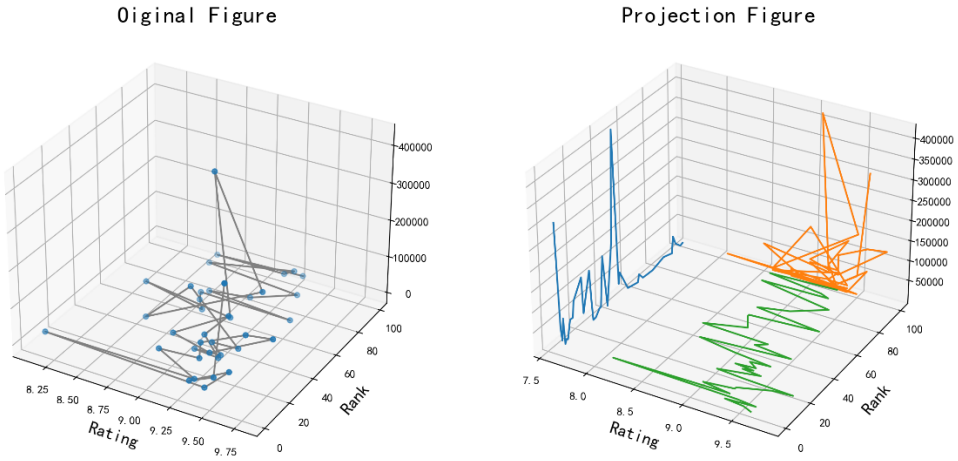

Fig.7. Relationship between rating, rank, and cumulative income

3. Conclusion

In conclusion, in this project, we found the basic situation of the current film industry: the quantity and quality of American films are the first of world, and the audience prefers drama type films and modern films. Nevertheless, we also explored the implicit relationships between some features of film: audiences pay more attention to the quality of films than anything else, and there is a positive correlation between the film's rank and income, as well as between rating and income. Eventually, films may also have other features and other potential relationships, which we are expected to be researched in the future.