**Data analysis of TOP100 movies**

The report focuses on two stories: 1) what features do high-quality movies have; 2) what factors affect the movie’s rating and income. With data analysis, the report can provide advice to audience to select high-quality movies and provide practical value to film makers to produce more attractive films.

**1. The features which high-quality movies have**

1) Description of quantitative data of TOP100 movies

As shown in table ([appendix 1](#appendix1_)), the diversity in ratings is not too large because the standard deviation is only 0.3486. If a movie is scored above 8.0 by Maoyan’s scoring system, it may become one of the TOP100 movies.

The interesting thing is that the most acceptable duration for movie fans is about 2 hours. According to statistics ([appendix 2](#appendix2_)), most viewers can accept movies within 139 minutes, and the duration of movies that viewers favorite is quite different.

The cumulative income is an important metric to evaluate whether a movie is successful. The description of income is exhibited in the table ([appendix 3](#appendix3_)). The diversity between the box office is large, which may be related to the year. Some movies were made in an earlier age when the film's penetration rate is low, so the audience's contribution to the box office is small. However, with the development of the film industry and the rising consumption power of residents, the box office has grown rapidly. The deviation between the box office may be related to the different purchasing power of currencies in different periods.

2) Analysis of category data of TOP100 movies

The pie chart (shown in [appendix 4](#appendix4_)) has calculated the percentage of movies’ type. Plot, romance, and crime are the three most popular movie types by audiences, and the top 100 movies belong to these three themes, accounting for 50.6%. The remaining themes of movies account for about 5%, except the music, which only has 1.7%.

It can be seen from the statistics (shown in [appendix 5](#appendix5_)) that American movies are very popular among Chinese. Among the 15 countries, the United States participated in producing most of films, with the figure of 58. China and the United Kingdom tied for second place and produced 15 films.

The films produced by 13 directors (seen in [appendix 6](#appendix6_)) accounted for 36% of the TOP100 movies, which can show that the director really affects the audiences’ choice of movies. Among these 13 directors, there are only 3 Chinese directors, namely Li Ang, Wang Jiawei and Jiang Wen, and the total number of their works is 7. This is enough to reflect that there is still a certain gap between the films produced by Chinese directors and the international ones.

3) Histograms of Year ([appendix 7](#appendix7_)), Month ([appendix 8](#appendix8_)), and Rating ([appendix 9](#appendix9_))

Among the movies on the list, there were few movies before 1990, especially in the 1970s, almost no movies entered the top 100. From 1990 to 2010, the number of movies that were loved by Chinese gradually increased. The number of Chinese favorite movies during 2010 to 2020 is almost two times more than the number of movies during former decade. Most of the movies were released in December of that year. There may be more holidays in December, so many directors decide to release the movie in December to earn more box office. The ratings of TOP100 movies are well in line with the normal distribution. The mean of this normal distribution is 8.98 and the std is 0.36.

**2. The interesting relationships between variables of movies**

The correlation matrix can more intuitively reflect the correlation between variables, and therefore, before conducting specific pairwise analysis, it is important to take a generally look at the information reflected by the matrix. The first matrix has a larger sample size, being more universally representative. The dark blue defines correlation as 1, the white defines correlation as 0, and the dark red defines correlation as -1.

1) Correlations between rating and other variables

In correlation matrix 1 ([appendix 10](#appendix10_)), there are 16 correlations between four variables, but the report focuses on exploring which factors affect rating. Though correlation matrix 1 shows the negative correlation between rating and month, the scatter does not dramatically show such relationship, so I just analyze how duration and year affect rating.

Movies with ratings above 9.0 are mainly concentrated in 80 to 100 minutes and movies with ratings below 9.0 are mainly concentrated in the range of 120 to 140, which verifies the slight negative correlation between movie duration and ratings (in [appendix 12](#appendix12_)). Movies with ratings above 9.0 are mainly concentrated in 2010 to 2020 and movies with ratings below 9.0 are mainly concentrated in around 2000 (in [appendix 13](#appendix13_)). Movies in the past 20 years have gradually met the public's demand.

2) Correlations between cumulative income and other variables

In correlation matrix 2 ([appendix 11](#appendix11_)), 25 correlations are exhibited, but the report just focuses on the factors that affect cumulative income. Although the matrix shows non-correlation between income and rating as well as the negative correlation between income and month, scatter of each sample gives us the different explanation.

Scatter (in [appendix 14](#appendix14_)) shows the movies with high score usually have high cumulative income, especially for movies whose ratings are over 9.0. Scatter (in [appendix 15](#appendix15_)) does not give us evidence to prove the negative correlation between income and month.

Appendix:

1. Description of Rating

|  |  |
| --- | --- |
|  | **Rating** |
| **count** | 100 |
| **mean** | 8.98 |
| **std** | 0.34641 |
| **min** | 8 |
| **25%** | 8.775 |
| **50%** | 8.95 |
| **75%** | 9.3 |
| **max** | 9.8 |

There are 100 effective ratings out of 100 movies. As shown above, the mean of the rating is 8.98 with the standard deviation 0.34641, the maximum score is 9.8, the minimum score is 8.0, and the median is 8.95.

2. Description of Duration

|  |  |
| --- | --- |
|  | **Duration(min)** |
| **count** | 100 |
| **mean** | 128.55 |
| **std** | 28.38173 |
| **min** | 80 |
| **25%** | 109.75 |
| **50%** | 124.5 |
| **75%** | 140 |
| **max** | 238 |

As shown above, the mean of TOP100 movies’ duration is 128.55 minutes with the standard deviation 28.382, the shortest movie is 80 minutes, the longest movie is 238 minutes, and 75% of the movies are less than 139 minutes.

3. Description of Income

|  |  |
| --- | --- |
|  | **Income** |
| **count** | 41 |
| **mean** | 562,287,561 |
| **std** | 871,131,798 |
| **min** | 20,000 |
| **25%** | 52,470,000 |
| **50%** | 173,700,000 |
| **75%** | 636,750,000 |
| **max** | 4,250,130,000 |

There are only 41 valid data in 100 movies, and 59 movies lack box office information. The mean of income from these 41 movies is 562.287 million yuan, and the standard deviation is 871 million. The lowest box office is 20,000 yuan, which belongs to *Infernal Affairs*, and the highest box office is 4.25 billion yuan, which belongs to *Avengers: Endgame.*

4. Pie chart of types

图表, 饼图

描述已自动生成

5. Bar chart of countries

图表, 直方图

描述已自动生成

6. Table of directors

|  |  |  |
| --- | --- | --- |
|  | **Director** | **Times** |
| **0** | 大卫·芬奇 | 4 |
| **1** | 克里斯托弗·诺兰 | 4 |
| **2** | 史蒂文·斯皮尔伯格 | 3 |
| **3** | 李安 | 3 |
| **4** | 朱塞佩·托纳多雷 | 3 |
| **5** | 宫崎骏 | 3 |
| **6** | 理查德·柯蒂斯 | 2 |
| **7** | 彼得·威尔 | 2 |
| **8** | 理查德·林克莱特 | 2 |
| **9** | 王家卫 | 2 |
| **10** | 詹姆斯·卡梅隆 | 2 |
| **11** | 姜文 | 2 |
| **12** | 蒂姆·波顿 | 2 |
| **13** | 彼特·道格特 | 2 |

7. Histogram of Year

图表, 直方图

描述已自动生成

Among the movies on the list, there were few movies before 1990, especially in the 1970s, almost no movies entered the top 100. From 1990 to 2010, the number of movies that were loved by Chinese gradually increased. The number of Chinese favorite movies during 2010 to 2020 is almost two times more than the number of movies during former decade.

8. Histogram of Month

图表, 直方图

描述已自动生成

Most of the movies were released in December of that year. There may be more holidays in December, so many directors decide to release the movie in December to earn more box office.

9. Histogram of Rating

图表, 直方图

描述已自动生成

The ratings of TOP100 movies are well in line with the normal distribution. The mean of this normal distribution is 8.98 and the std is 0.36.

10. Correlation Matrix 1

图片包含 游戏机, 华美, 人

描述已自动生成

The correlation matrix does not contain income, and therefore, all of 100 data are valid.

11. Correlation Matrix 2

形状, 正方形

描述已自动生成

The matrix contains income, and the calculation is only based on 41 valid information because 59 movies lack box office information.

12. Scatter between Duration and Rating

图表, 散点图

描述已自动生成

13. Scatter between Year and Rating

图表, 散点图

描述已自动生成

14. Scatter between Rating and Income

图表, 散点图

描述已自动生成

15. Scatter between Month and Income

图表, 散点图

描述已自动生成