Data_Analysis

Group 23

2023-03-19

Exploratory data

Correaltion

Figure 3 shows that there are strong correlations between cast total facebook likes and the actors' facebook likes. Because cast total Facebook likes, it contains Facebook likes of actors 1, 2, and 3. When the Facebook likes of actors 1, 2, and 3 increase, so do the cast total Facebook likes. Principal component analysis can be used to eliminate the multicollinearity of the data, or select one of the two data for analysis.

Plot keyword & Film categorization

We are interested in the distribution of the top 20 plot keywords. Figure 4 shows that The most frequently seen plot keyword is wedding, which is repeated 12 times. This was followed by sex, terrorist, and writer, which are repeated 8 times each.

The figure 5 shows that the most repeated movie genre is drama, which is repeated 236 times. This was followed by Thriller, Romance and Comedy, which were repeated 131, 108 and 102 times, respectively.

Imdb score

As we can see from the figure 6, the average imdb score of successful movies is higher than that of unsuccessful movies. The average score of successful movies is about 6.8 to 6.9. Among the unsuccessful movies, there are some outliers, which have an imdb score below 4. We learn from the figure that successful movies have higher imdb scores overall than unsuccessful ones, but this difference is not very large. Some of the successful movies also have lower scores

Content rating

From figure 7, for movies rated PG, PG_13, and R, the percentage of unsuccessful movies was over 50 percent, while for the movies are not be rated, the percentage of successful films was greater.

Title year

The figure 8 shows that the proportion of successful vs unsuccessful movies in each title year. We can tell from the figure that the

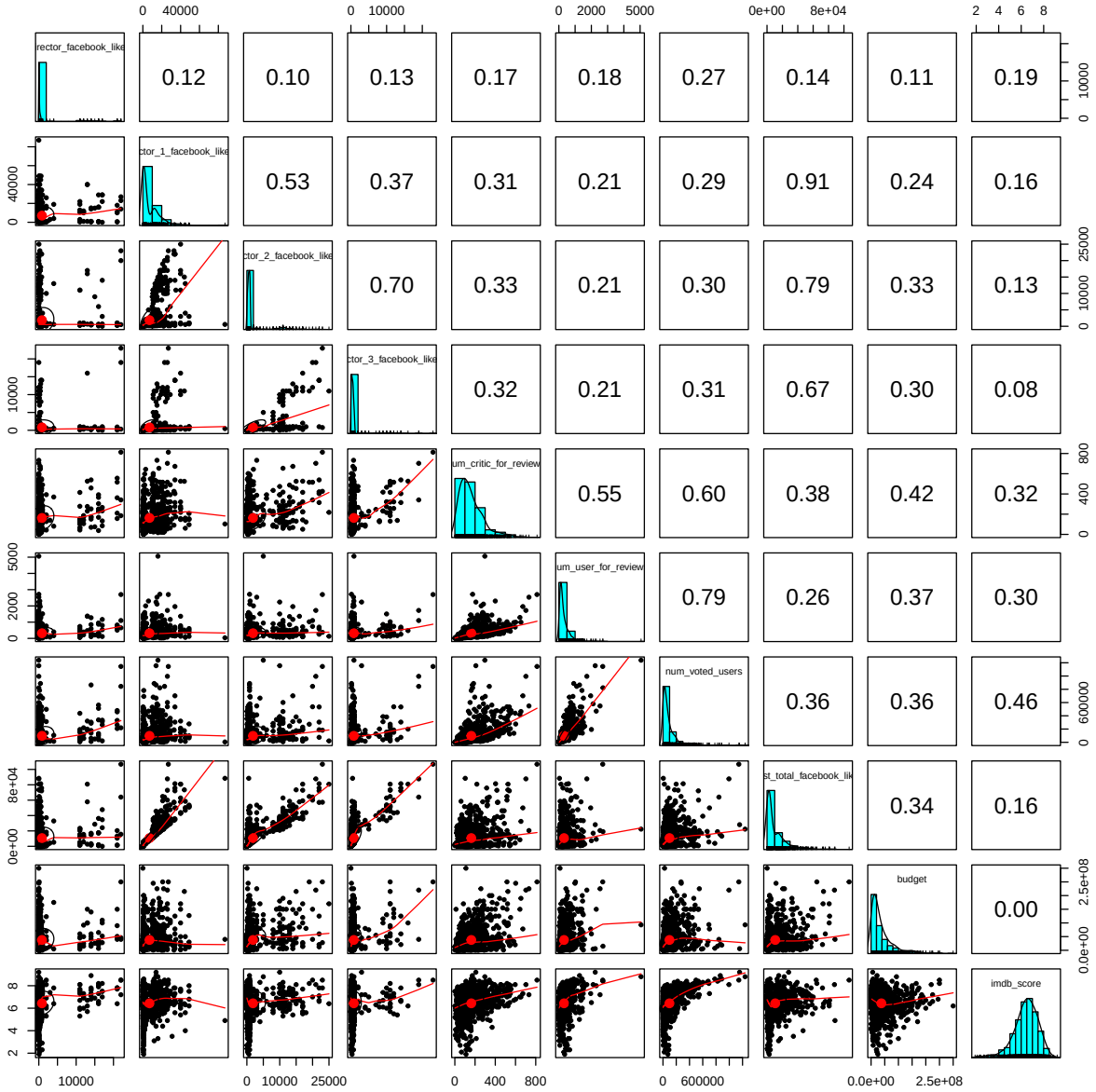

Figure 1: Correlation of the facebook likes

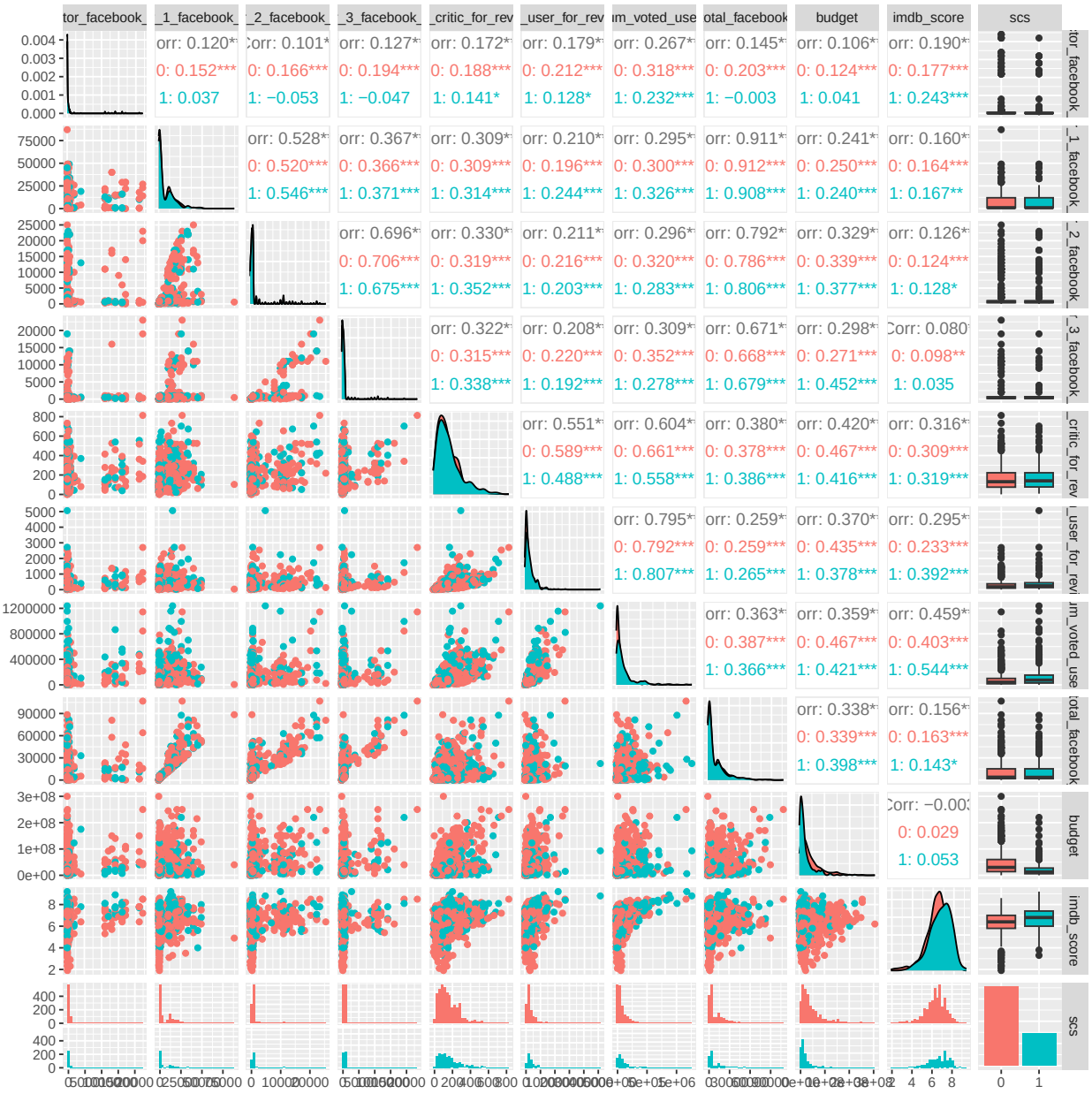

Figure 2: Correlation of the facebook likes

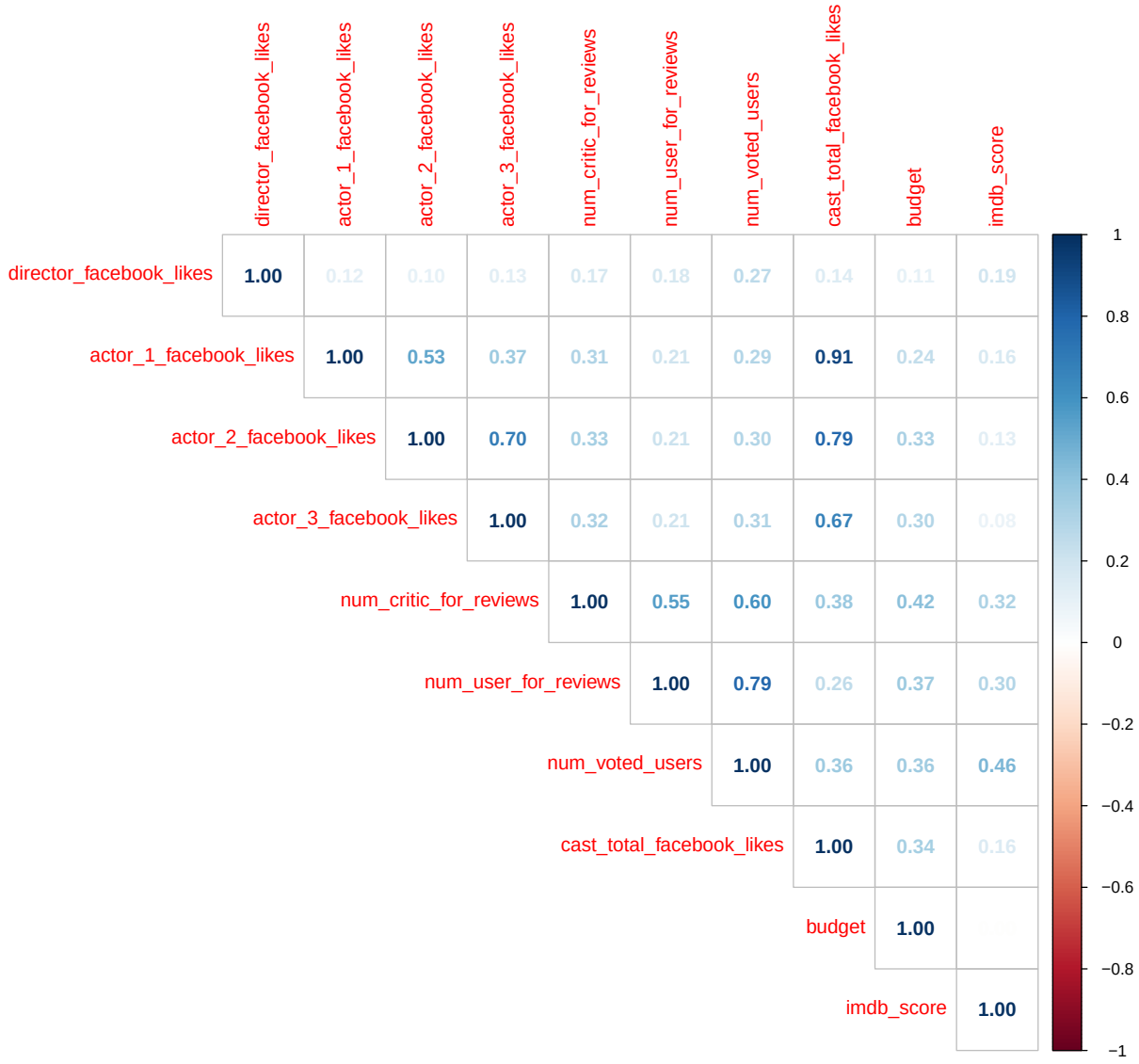

Figure 3: Correlation of the facebook likes

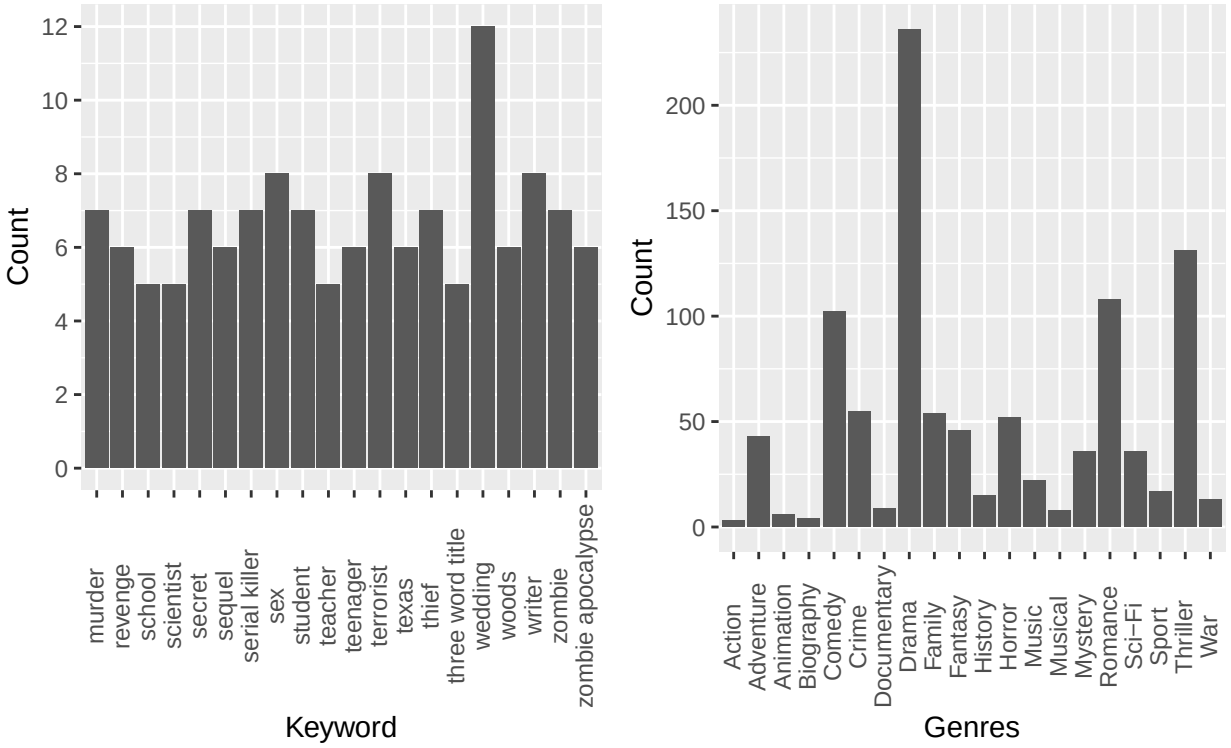

Figure 4: The distribution of the top 20 plot keywords

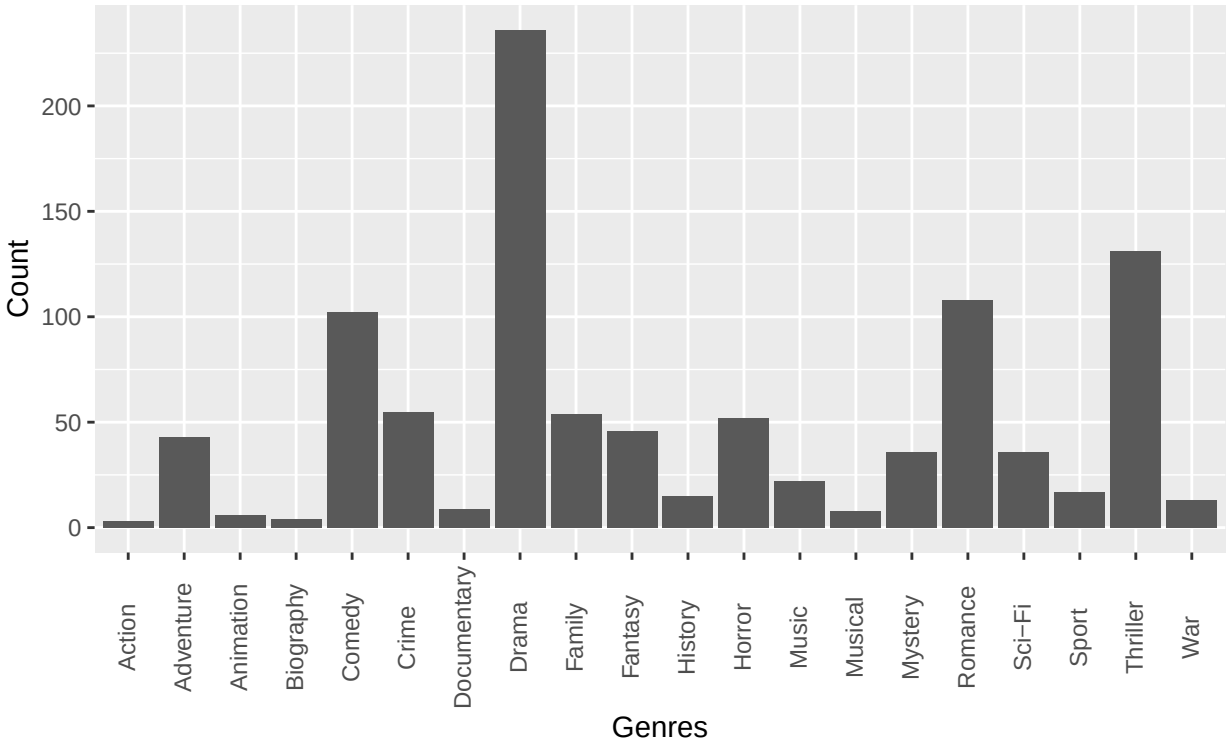

Figure 5: The distribution of the top 20 movie categories

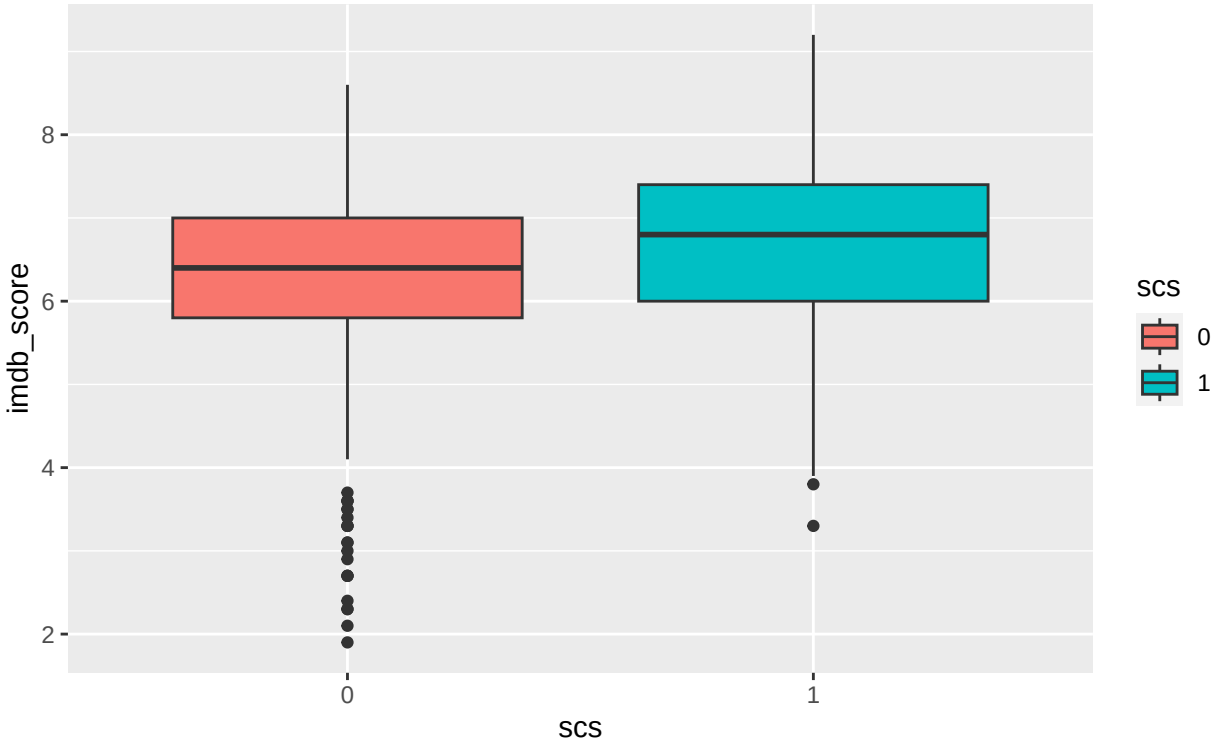

Figure 6: Boxplot of imdb scores of successful and unsuccessful movies

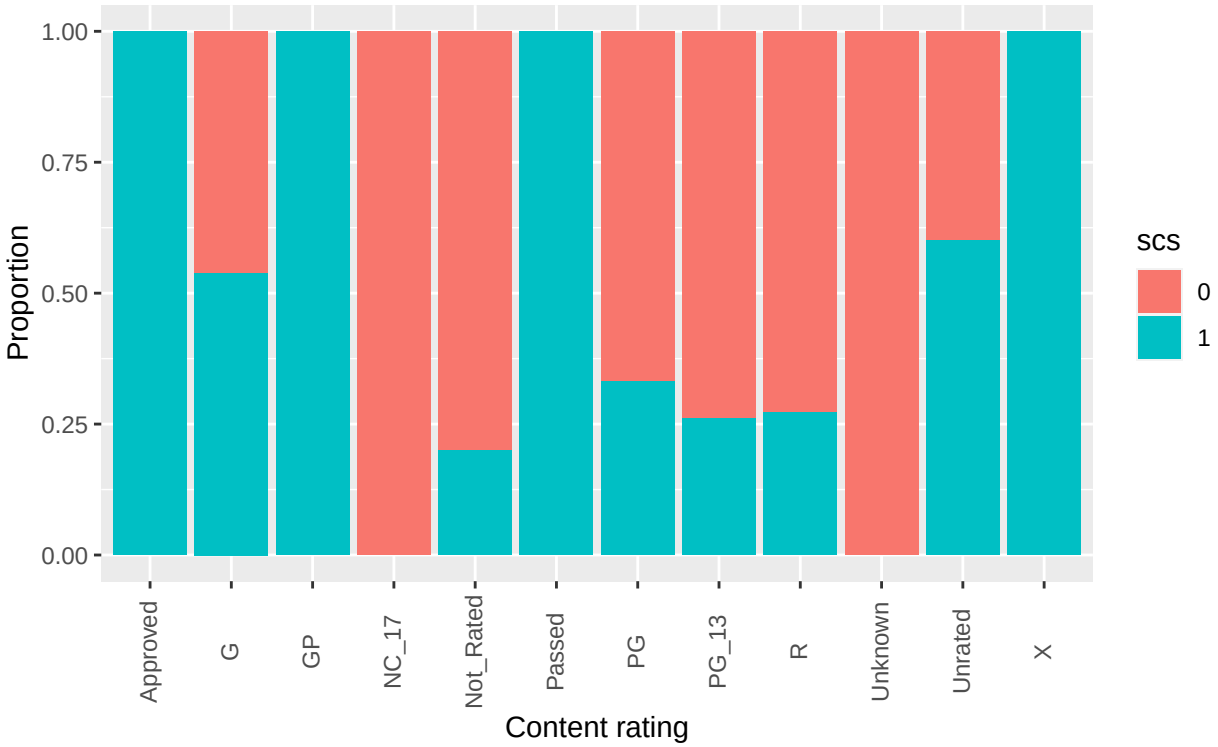

Figure 7: Proportion of successful vs. successful movies in content rating

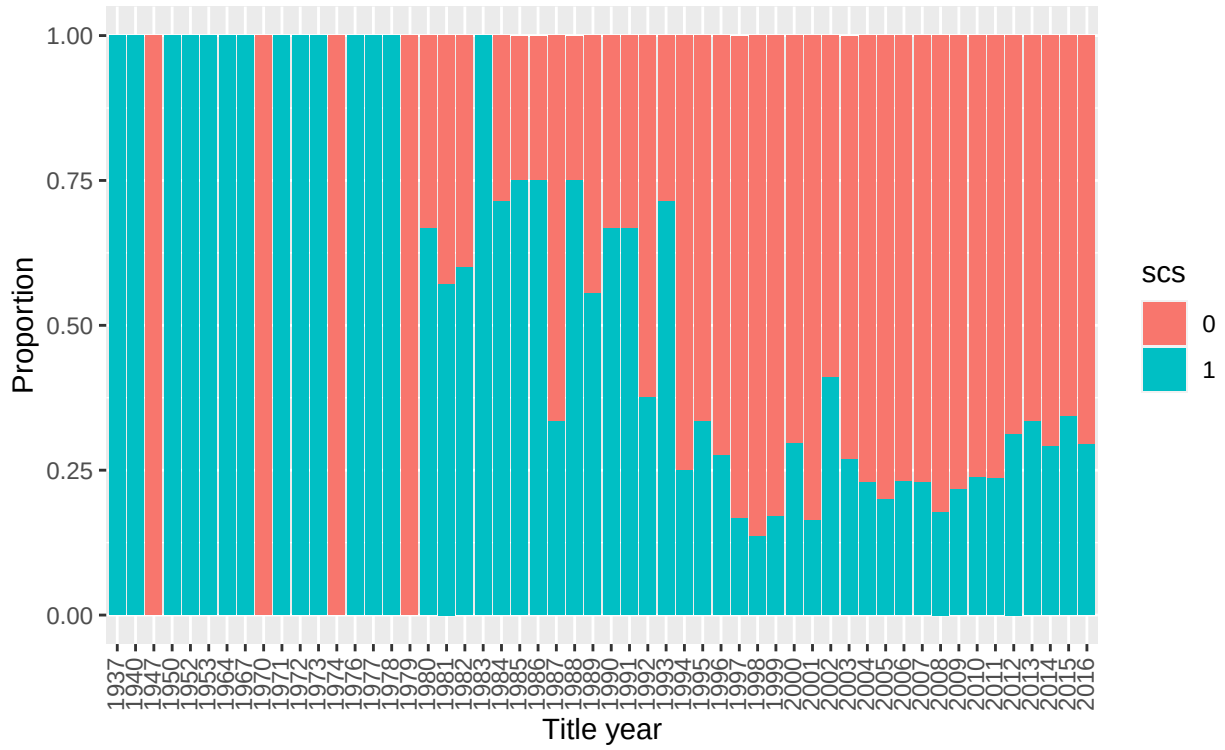

Figure 8: Proportion of successful vs unsuccessful movies in title year

Language

The figure 9 shows that the proportion of successful vs unsuccessful movies in each type of language. We can tell from the figure that movies whose languages are Cantonese and Indonesian are all successful movies in our dataset. The rest of the non-English movies are not a high percentage of successful movies, and the main languages of successful movies are English, French, Hindi, Spanish.

Aspect ratio

We can tell from the figure 10 that more than 60% of the movies with an aspect ratio of 1.3 are successful movies, and about 50% of the movies with aspect ratios of 1.37 and 1.66 are successful movies. Aspect ratios of 1.77, 2.4, and 2.55 are all unsuccessful movies, while movies with aspect ratios of 2 are all successful movies. It can be tentatively stated that the aspect ratio of a movie may be a factor that affects whether a movie is successful or not, because the aspect ratio may affect the audience's enjoyment.

Color

As we can see from the figure 11, most of the movies are in color. The number of successful movies in color is about 270. However, the small number of black and white movies may lead to the conclusion that the analysis of this factor is not convincing. Therefore it may be possible to round off this factor depending on the model requirements.

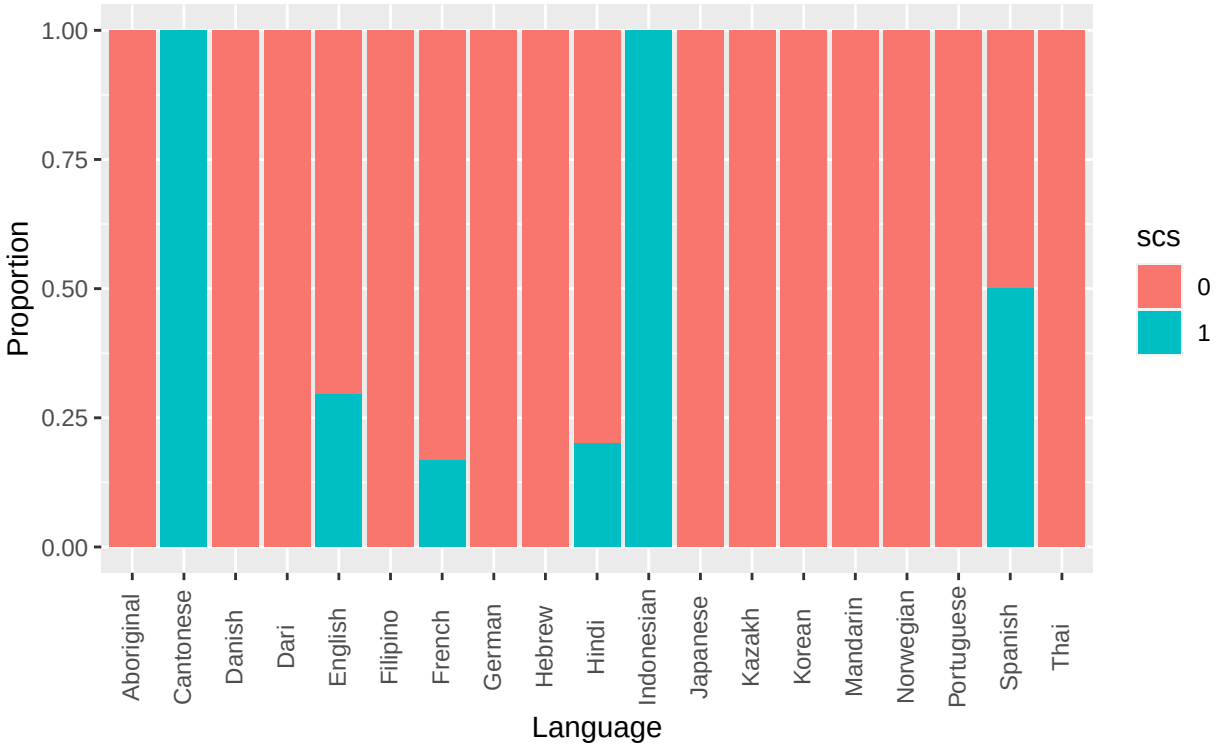

Figure 9: Proportion of successful vs unsuccessful movies in language

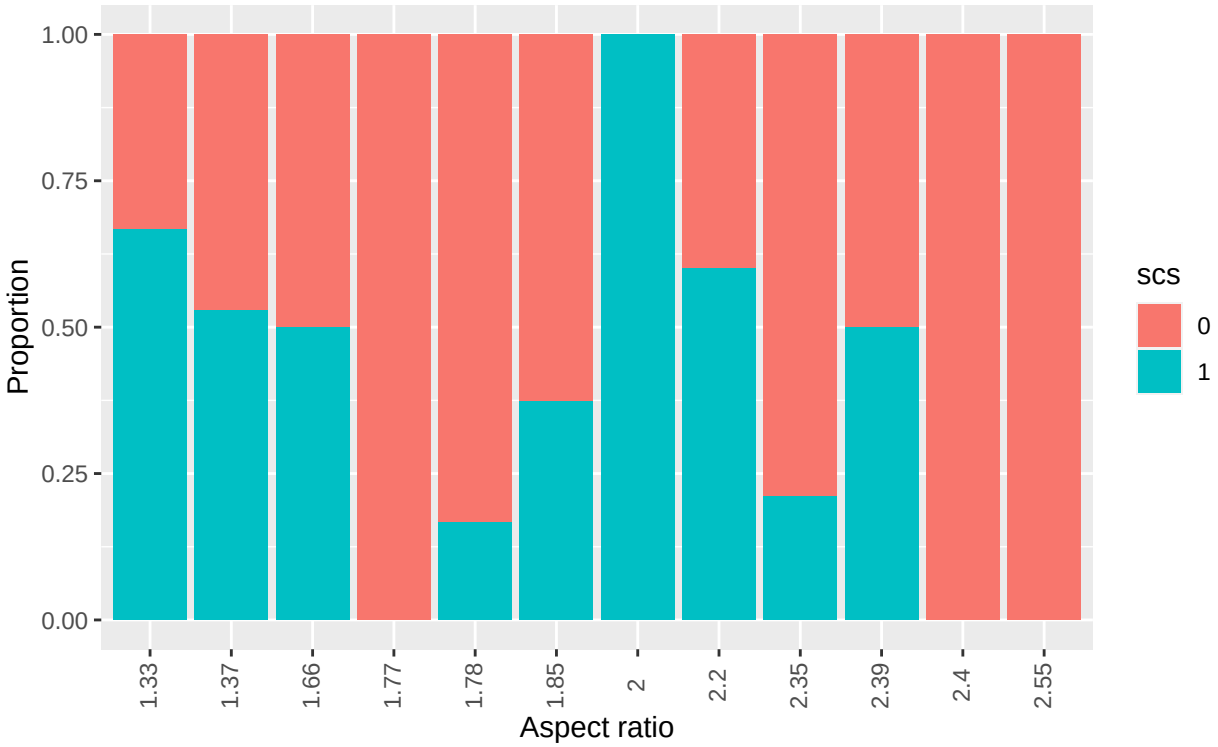

Figure 10: Proportion of successful vs unsuccessful movies in aspect ratio

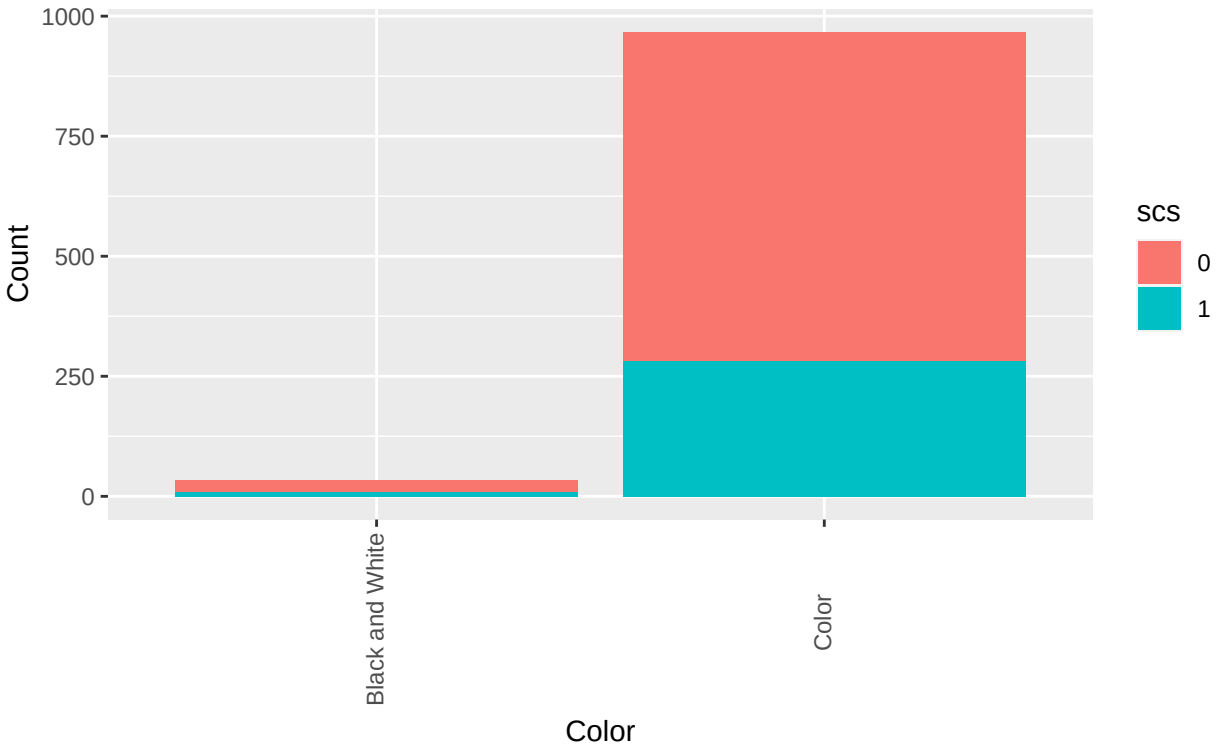

Figure 11: Proportion of successful vs unsuccessful movies in different color