University of St Andrews | FOUNDED 1413

# MT5758 Individual Report

Student ID: 180009169

## 1. Executive Summary

This report mainly has two aims. The first is to uncover any hidden relationship between social media and movie performance. Another is to find any characteristics that might have impact on the monetary aspects of the movie, such as gross and budget. Methods applied in this report is Principal Components Analysis (PCA). The report finds that there are clusters when colouring observations according to some variables. Furthermore, some plots show similar patterns. The analysis conducted in this report also have some imitations. For instance, we cannot determine which variables cause high values in the first principal component, which might need further analysis by other methods.

## 2. Introduction

Due to the increasing popularity of social platforms such as Facebook, the social media campaign has now become an important marketing tactic in the movie industry. According to a previous research by Pardo (2013), Hollywood, one of the biggest players in the movie industry, has changed its reluctant attitude and progressively embraced new technology. This report is supposed to investigate relationships between different aspects of the movie. To achieve this goal, we will first clean our dataset, then conduct Principal Components Analysis, finally plot observations that coloured with different variables to find whether there is any cluster.

## 3. Data Cleaning

This report is based on *IMDB 5000 Movie* dataset (Sun, 2016) and analysed with *RStudio* (RStudio Team, 2015). The original Internet Movie Database (IMDB) 5000 Movie dataset has 28 variables for 5043 movies which spanning across 100 years and 66 countries. According to Figure1, around 1.9% values in the original dataset are missing. Especially in gross and budget variable, 17.53% and 9.76% values are missing respectively.

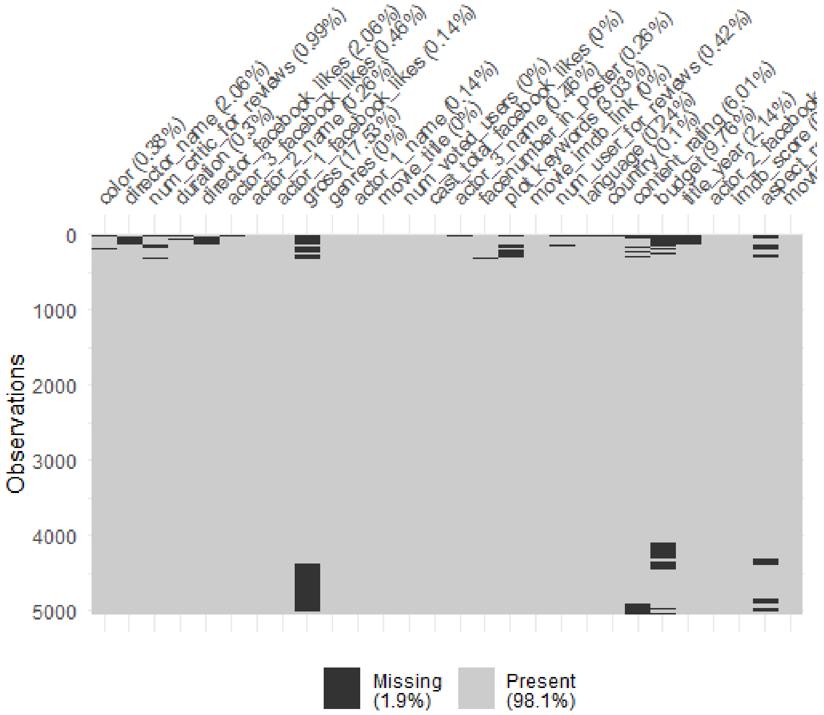


**Figure1. Bird's eye view of missing values in the original dataset** The columns are variables and rows are observations. Black blocks are missing values. The number after variable name is the percentage of missing values for each variable.

To clean the data, we first removed four variables which are "movie_title", "plot_keywords", "movie_imdb_link" and "aspect_ratio" because these variables are either so unique (i.e. movie_title) or useless (i.e. aspect_ratio). Second, we found all missing values in the colour variable appear to be colour movies, all missing values in the country variable appear to be American movies, while all missing values in the content rating variable are grouped as “Not Rated”. After that, observations with missing values in “actor_name” or “actor_facebook_likes” are removed. Then observations with more than two missing values are also removed, which loses about 2.57% of the current number of observations. Finally, the rest variables with missing values are all numeric variables. Therefore, we imputed these missing values with the mean for each variable. After cleaning data, now we have 24 variables for 4891 observations with no missing values.

## 4. Principal Components Analysis

Principal Components Analysis (PCA) is a technique for multivariate data analysis. It is not a model but a method to explore our dataset. The main purpose of PCA is to reduce dimensions of the data cloud, while retaining as much information as possible. To reduce the number of variables, PCA builds new variables which are linear combination of original variables with different weights. These new variables are called Principal Components (PCs).

In the report, some variables such as gross and budget have much bigger value than others, they would get most weights and dominate in principal components if we do PCA on variance matrix. Therefore, we decided to perform PCA on correlation matrix instead of covariance matrix. Correlation matrix can be thought as a standardised covariance matrix with all values are between negative one and positive one. After the scaling problem, the next task of PCA is to choose the number of principal components. According to the Scree plot (Figure2), The curve decreases slowly after component 4, which is an indication that we should keep the first four principal components. In addition, the Kaiser’s criterion also suggests us to retain the first four. Therefore, we made the cut off points as four.

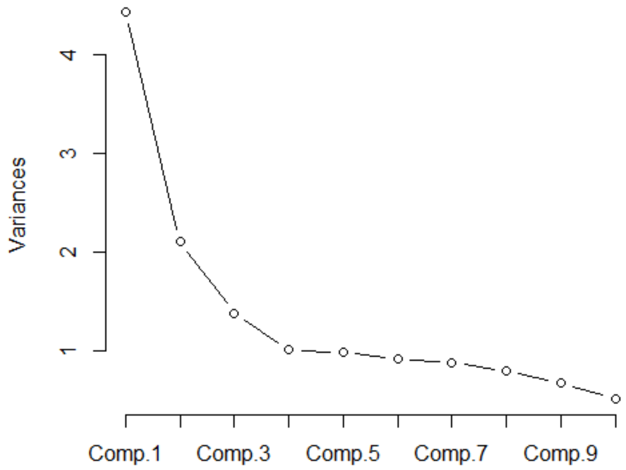


**Figure2. Scree Graph for PCA** This graph plots eigenvalues against each principal component. Eigenvalues reflect the importance of principal components. The slope get mild after the fourth component.

Table1 displays the loadings, which can be considered as the weights of original variables for each principal component. The positive/negative signs show the direction that original variables contribute to principal components. Blanks in the table are not missing values but values that are too small to present.

**Table1. Loadings for the first four Principal Components**

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| num_critic_for_reviews | 0.364 | 0.153 | 0.294 | |
| duration | 0.207 | 0.185 | -0.320 | -0.235 |
| director_facebook_likes | 0.157 | | -0.226 | |
| actor_3_facebook_likes | 0.258 | -0.263 | | -0.202 |
| actor_1_facebook_likes | 0.226 | -0.468 | -0.207 | 0.232 |
| gross | 0.312 | 0.133 | | |
| num_voted_users | 0.392 | 0.234 | | |
| cast_total_facebook_likes | 0.287 | -0.503 | -0.173 | 0.149 |
| facenumber_in_poster | | -0.173 | | -0.849 |
| num_user_for_reviews | 0.353 | 0.249 | | |
| budget | | | 0.103 | 0.297 |
| title_year | | -0.160 | 0.662 | |
| actor_2_facebook_likes | 0.270 | -0.351 | | |
| imdb_score | 0.207 | 0.249 | -0.360 | |
| movie_facebook_likes | 0.316 | 0.125 | 0.311 | |

The biplot in Figure3 is a visualisation of loadings for the first two components. All variables have positive impact on the first component. However, variables about social media (i.e. facebook likes) have negative influence on the PC2. Therefore, the biplot shows a “fan-like” pattern. In addition, the black points in the biplot are scores, which are observations mapped on new axes.

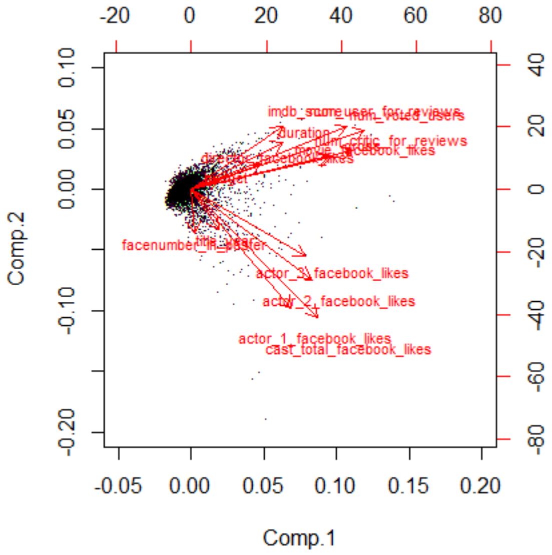


**Figure3. Biplot for PC1 and PC2** The black dots are projected observations. The red arrows show the weights and directions that variables contribute to principal components.

## 5. Discussion

As Figure4 shows, the three plots have clear and similar clusters. The left plot is coloured by the number of critical reviews on imdb. The middle plot is coloured by the number of users who gave a review. The right plot is coloured by the number of people who voted for the movie. Different groups are divided by the tertile and coloured differently. Green group contains the first third of lowest data, the blue group is the middle part of data, while the red points are the highest values. There are clear clusters along PC1. It means the movie with high value in the number of critical reviews, or the number of users who gave a review, or the number of people who voted is more likely to have high value in PC1. However, it is difficult to tell which variables contribute to this result because all loadings for PC1 are positive and no variable dominates in PC1. They have similar patterns might because these three variables are related. The movie with a lot of reviews means it is likely to be popular and well-known. Therefore, it usually has more critical reviews and voted users. This can also be proved by calculating correlation coefficients. The correlation coefficient between any two of them is great than 0.6. However, not all plots show some clusters. For instance, there is no clear pattern when colouring by director's facebook likes (Appendix A) and imdb scores (Appendix B).

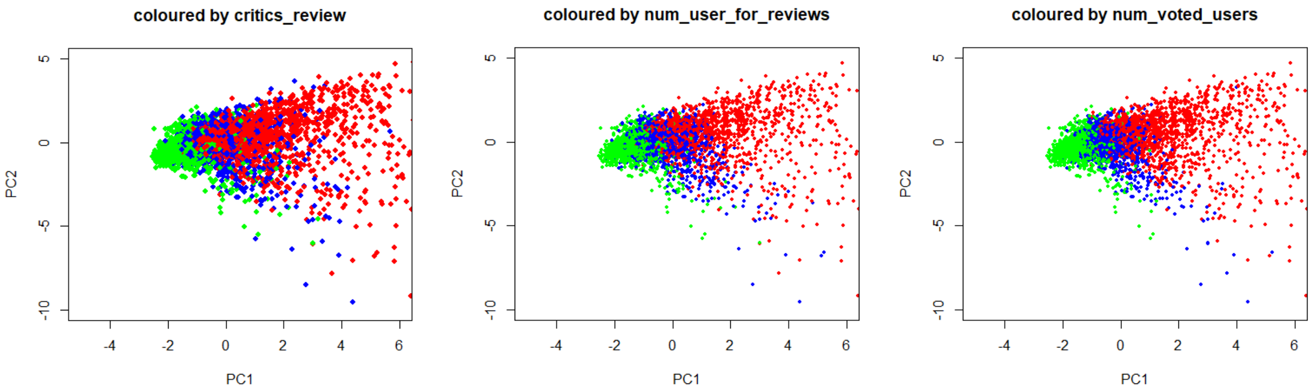


**Figure4. Scores plotted on PC1 and PC2, coloured by different levels of number of critic reviews, user for reviews and voted users** Green points are "low" group. Blue points are "medium" group. Red points are "high" group.

As for the second aim of the report, we found no clear pattern along any principal component when plotting scores with different groups of gross (Figure5). But when plotting scores with different budget, there are clusters along PC1. It means that large budget might be related to large value in PC1. Like the problem we met before, it is hard to determine which variables cause the high values in PC1.

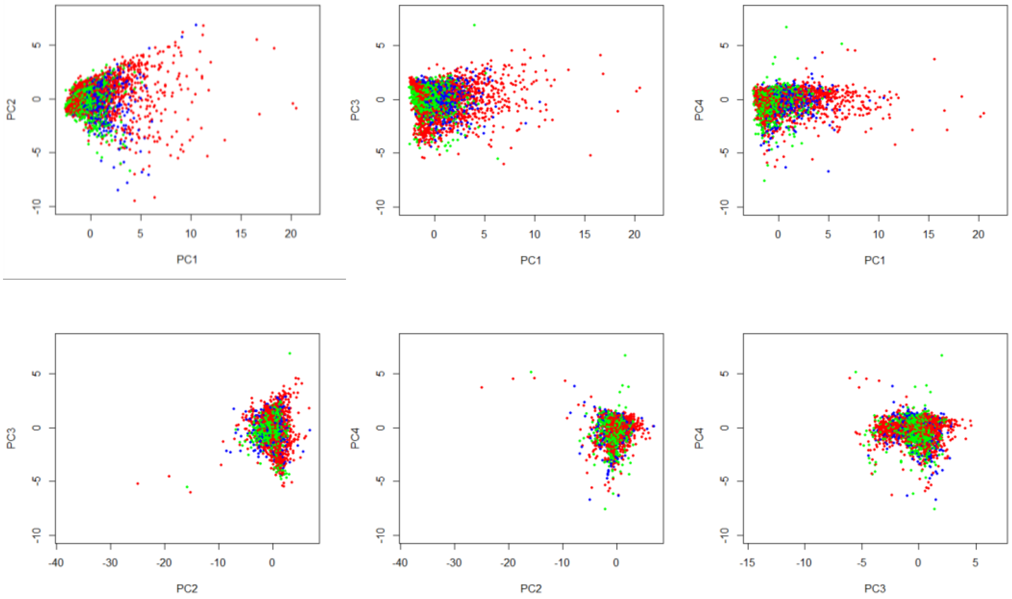


**Figure4. Scores plotted on pairs of PCs, coloured by different levels of gross** Green points are “low” group. Blue points are “medium” group. Red points are “high” group.

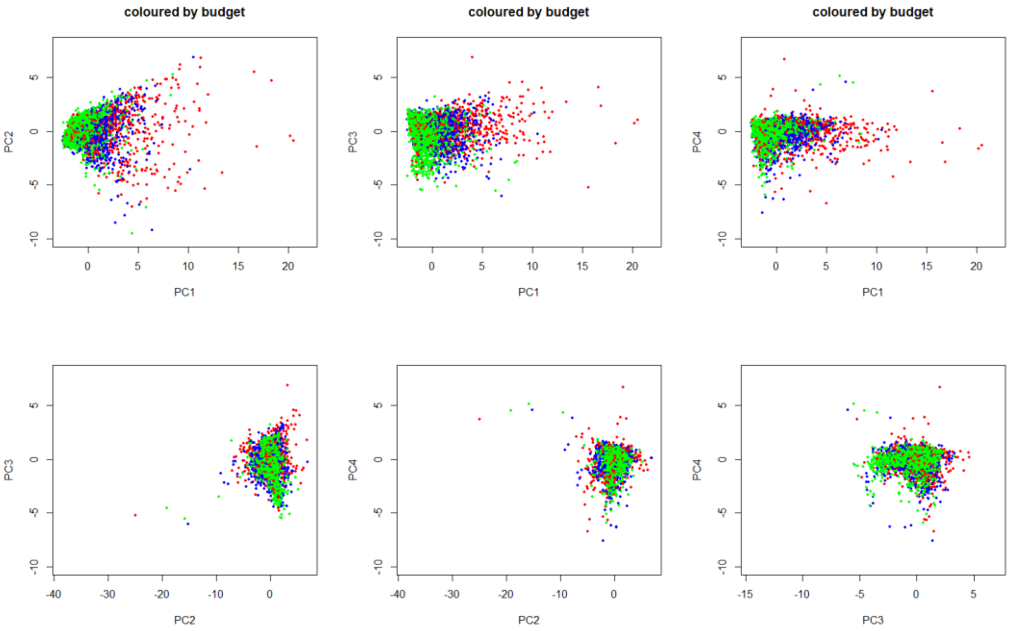


**Figure5. Scores plotted on pairs of PCs, coloured by different levels of budget** Green points are “low” group. Blue points are “medium” group. Red points are “high” group.

## 6. Conclusion

To sum up, there are four  main conclusions. First, all original variables have positive weights in PC1 and only variables which related to facebook have negative weights in PC2. Second, there are clusters when plotting scores and coloured with some variables such as the number of critical reviews and the number of voted users. In addition, patterns are similar when colouring by the number of critical reviews, the number of users who gave review and the number of voted users. Finally, for monetary variables, we could not find any clear pattern in gross. However, budget is likely to have positive relationship with PC1.

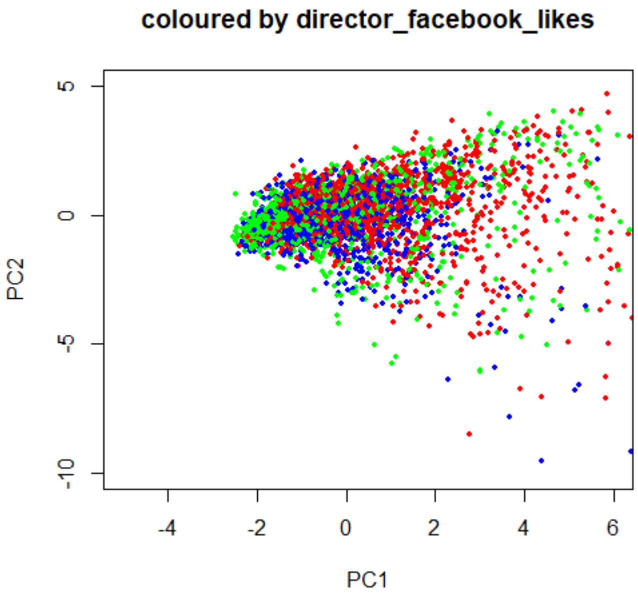
Reference:

Pardo, A., (2013). Digital Hollywood: How Internet and Social media are changing the movie business. In Handbook of Social Media Management (pp. 327-347). Springer, Berlin, Heidelberg.

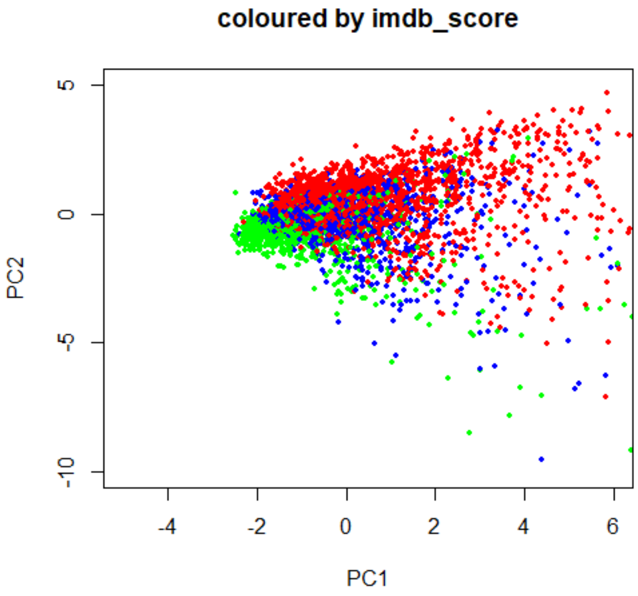
RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA, URL: http://www.rstudio.com/.

Sun, C., (2016). IMDB 5000 Movie Dataset. data.world. e675d8a8. Available at: https://data.world/popculture/imdb-5000-movie-dataset. (Accessed 22nd April 2019)

Appendix A. Scores plotted on PC1 and PC2, coloured by director's facebook likes



Appendix B. Scores plotted on PC1 and PC2, coloured by imdb scores


coloured by imdb_score
PC2
PC1