CASS
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON
EST 1894

Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Arora, Akshay 2. Chawla, Sarthak 3. Gaur, Siddharth	4. Li, Jiaying 5. Wu, Ziyan **GROUP NUMBER:** 2
MSc in: Business Analytics	
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer: Zhu, Rui	**Submission Date:** 05/04/2023

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: %

Group coursework 2

1.Motivation

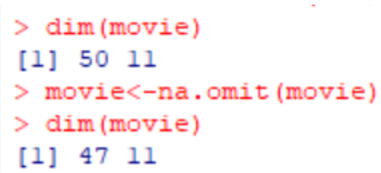
Based on IMDb's movie rating data and movie feature data, this study conducts PCA analysis and cluster analysis to explore the impact of different movie features on movie ratings. This helps us better understand the film market and audience needs, and provides filmmakers with useful information to improve film production and promotion strategies (Yadav et al., 2021; Tripathi et al., 2023). IMDb's movie ratings and other feature data can be used to objectively characterize movie quality and audience satisfaction. However, movie ratings are not solely determined by the quality of the movie itself, but are also influenced by many other factors (Sun et al., 2021). The motivation of this study is to deeply analyze the impact of these factors on movie ratings to provide recommendations for movie promotion and audience demands.

2.Experiment setting

2.1 Data processing and descriptive statistics

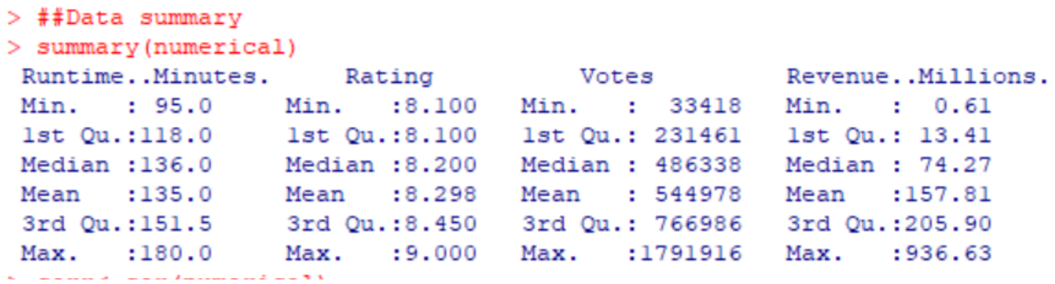
In this study, a sample of 50 top-rated movies from IMDb was collected for PCA and cluster analysis. For the preliminary data collection, a total of 11 variables and 50 samples were included. Data processing was carried out in this study. After deleting the missing values, there were 47 effective samples left in the cleaned samples, as shown in Figure 1. Since it is difficult to effectively extract information from text variables, only numerical variables are retained in this study. Then, this study summarizes the data of the movie database, as shown in Figure 2. Among them, the average movie score is 8.298, and the average movie duration is 135 minutes.

Furthermore, this study calculated the correlation coefficient of movie duration, rating, number of votes and revenue, and drew the correlation coefficient matrix, as shown in Figure 3. Among them, the movie income and the number of votes have the greatest correlation, which is 0.55. However, the correlation between movie earnings and ratings is very weak, only -0.01, which means that movie quality and earnings are not significantly related. Similarly, there is no obvious connection between movie revenue and duration.



```
> dim(movie)
[1] 50 11
> movie<-na.omit(movie)
> dim(movie)
[1] 47 11
```

Figure 1 Data cleaning



```
> ##Data summary
> summary(numerical)
 Runtime..Minutes.      Rating          Votes         Revenue..Millions.
 Min.   : 95.0     Min.   :8.100   Min.   :  33418   Min.   :  0.61    
 1st Qu.:118.0     1st Qu.:8.100   1st Qu.: 231461   1st Qu.: 13.41    
 Median :136.0     Median :8.200   Median : 486338   Median : 74.27    
 Mean   :135.0     Mean   :8.298   Mean   : 544978   Mean   :157.81    
 3rd Qu.:151.5     3rd Qu.:8.450   3rd Qu.: 766986   3rd Qu.:205.90    
 Max.   :180.0     Max.   :9.000   Max.   :1791916   Max.   :936.63    
```

Figure 2 Data summary

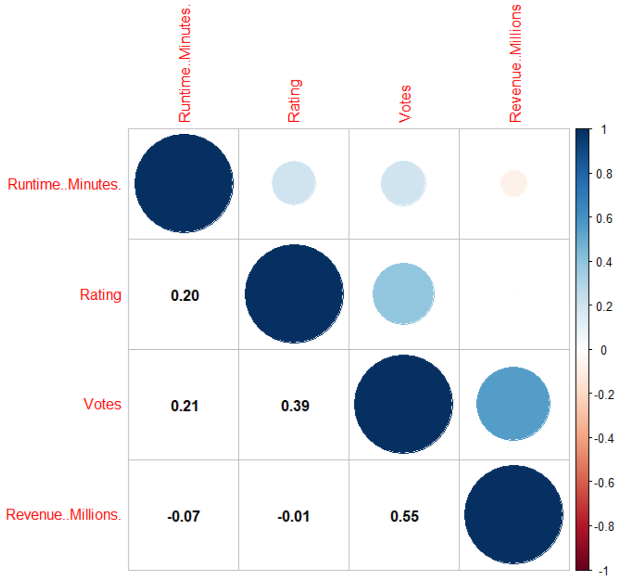


Figure 3 Correlation

2.2 PCA

Through PCA analysis, this study extracts principal components from movie feature data to explain the variability among movies. The reason for choosing PCA in the experimental design is that it can reduce the number of data set features and data redundancy, and improve analysis efficiency by streamlining data information. Colleagues, this study chose to retain the number and explanatory power of principal components to facilitate subsequent cluster analysis.

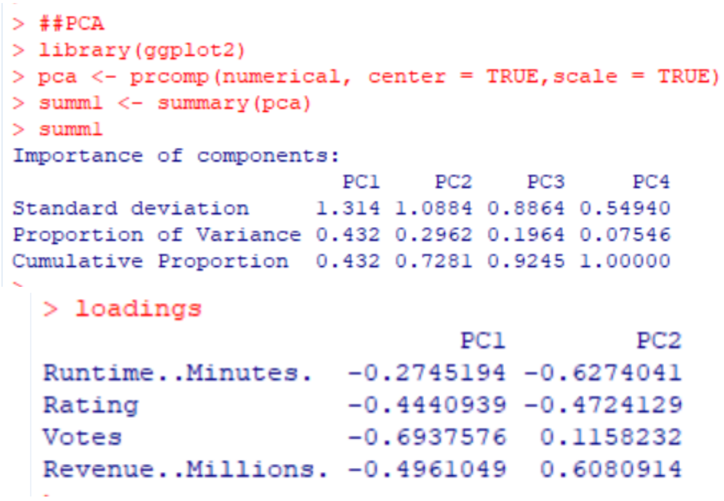
2.3 Cluster

Based on the cluster analysis method, this study was able to divide the movies into different groups, which was used to compare the movie features and ratings between different groups. These analyzes can reveal similarities between movies and the impact of different movie features on movie ratings. The purpose of cluster analysis in the experimental design is to classify movies according to their characteristics and dig out effective information and conclusions. In order to use the data set obtained by principal component analysis for cluster analysis, this research needs to select an appropriate clustering algorithm and distance measurement method, such as using the Kmean method and determining the number of clusters.

3.Analysis of the result

3.1 PCA result

First, PCA analysis was performed using R language, as shown in Figure 4, 4 principal components were obtained, and the explained variance ratios were 43.2%, 29.6%, 19.6 and 7.5, respectively. Among them, PC1 and PC2 can explain 72.8% of the total variance change ratio, indicating that these two principal components have extracted most of the information in the data. The correlation coefficient between voting and PC1 is the largest, indicating that this variable occupies the largest weight in PC1, followed by box office revenue. For PC2, duration and box office receipts hold most of the weight. Therefore, PC1 is mainly used to measure votes, while PC2 is mainly used to measure movie duration and box office revenue.Then, this study took PC1 and PC2 as an example to draw a scatter diagram, as shown in Figure 5. In order to explore the differences in PCA analysis results in different years, this study also divides categories according to years. The results show that PC1 and PC2 have a certain positive correlation, but there is no significant difference in the results of PCA analysis in different years.



```
> ##PCA
> library(ggplot2)
> pca <- prcomp(numerical, center = TRUE,scale = TRUE)
> summ1 <- summary(pca)
> summ1
Importance of components:
                         PC1    PC2    PC3     PC4
Standard deviation     1.314 1.0884 0.8864 0.54940
Proportion of Variance 0.432 0.2962 0.1964 0.07546
Cumulative Proportion  0.432 0.7281 0.9245 1.00000
```

```
> loadings
                           PC1        PC2
Runtime..Minutes.   -0.2745194 -0.6274041
Rating              -0.4440939 -0.4724129
Votes               -0.6937576  0.1158232
Revenue..Millions.  -0.4961049  0.6080914
```

Figure 4 PCA results

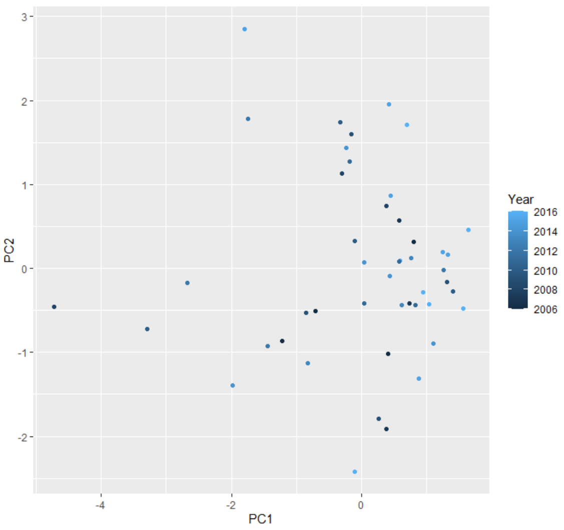
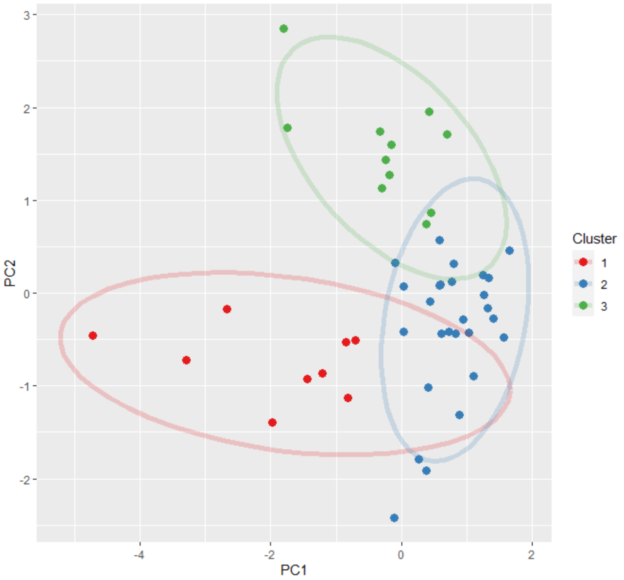


Figure 5 PC1 and PC2

3.2 Cluster the movies

Since the first two principal components can explain more than 70% of the total variance, this study uses PC1 and PC2 as the main input variables of cluster analysis, and uses the K-Mean method for cluster analysis, as shown in Figure 6. In order to further distinguish the clustering results, this study marked the categories with three circles of different colors. Among them, the values of PC1 and PC2 in the green circle range are relatively large, which means that the votes, box office revenue and movie duration of this type of movie are relatively high, but the IMDb score is relatively low. The PC1 value in the blue circle is larger, and the PC2 value is smaller, indicating that the votes of such movies are relatively high, while the box office revenue and movie duration are low. The PC1 and PC2 values in the red circle range are at a small level, which means that the IMDb score of this type of movie is relatively high, and the second vote, box office revenue and movie duration are relatively low. Clustering results can provide filmmakers and investors with important information about movie genres. For example, if a producer's goal is to make a movie that is popular with audiences and earns a lot of money at the box office, then the movie can be made with reference to the "high box office, high voter turnout, high rating, short running time" genre. If the producer's goal is to make a more artistic film, then you can refer to the "high box office, high voter turnout, low rating, long duration" genre.



4. Conclusions

This study uses PCA and K-Mean cluster analysis to carry out data mining on the IMDb movie database. The results show that the 4 numeric variables in the data can be measured by two principal components and can explain more than 70% of the total variance. According to the cluster analysis results, this study was able to classify three different categories of movies. In conclusion, PCA analysis and cluster analysis can provide important information of film types, which has certain reference value for the development of film industry and investment decision-making.

Reference

Sun, Wei & Jiang, Jianhua & Huang, Yubo & Li, Jinlai & Zhang, Mengmeng. (2021). An Integrated PCA DAEGCN Model for Movie Recommendation in the Social Internet of Things. Internet of Things Journal, 2(1), 1-12.

Tripathi, Jyoti & Tiwari, Sunita & Saini, Anu & Kumari, Sunita. (2023). Prediction of movie success based on machine learning and twitter sentiment analysis using internet movie database data. Indonesian Journal of Electrical Engineering and Computer Science, 29(3), 1750.1759.

Yadav, Vikash & Shukla, Rati & Maurya, Anamika. (2021). A New Approach for Movie Recommender System using K-means Clustering and PCA. Journal of Scientific and Industrial Research. 80(2), 159-165.