Comprehensive Analysis of the Last Four Decades of Movie Industry: Implication for Film Investment

Junpeng Yang*

19107412d@connect.polyu.hk

Hong Kong Polytechnic University, Hong Kong, 999077, China

Abstract—On the way to maturity of the film industry, the box office has been accentuated as one of the representative indicators to measure a film's success and intuitively signify profitability, and there are other critical factors for film success that have been intensely debated and divided within the research community. This article aims to combine statistical and machine learning methods, applying SPSS and Python in turn, to comprehensively analyze the IMDb dataset of the film industry. From analysis, budget and votes are selected as the most predictive variables for movie revenue in the multivariate linear regression section. Time series decomposition demonstrates a fluctuating upward trend and evident seasonality toward movie revenue. Two principal components retained after varimax rotation are summarized as income and satisfaction factors in the scheme of factor analysis, and the cross-distribution of primary genres and ratings was found to be consistent in the cross-tabulations with the R-rated comedy being the most significant pair. The U.S. leadership in the film industry is confirmed after the rate of return is introduced in ANOVA. Moreover, the entire dataset's movies are roughly divided into three major categories and latent partnerships between representative directors and their followers are detonated in the K-means clustering analysis part. The major finding of each analysis outlined is expected to enhance the performance of prediction, and clustering models associated with related research, meanwhile, suggesting a more comprehensive industry status for further decision-making of movie investment.

Keywords—Film Industry, Linear Regression, Time Series Analysis, Factor Analysis, ANOVA, K-means Clustering.

1 INTRODUCTION

The impact of the movie industry on people’s lives has involved various fields, including the publishing industry, tourism, education industry, stock market, e-commerce platform, etc. It is a typical microcosm of the national economy (Li & Wang & Wu et al., 2021). As the seventh performing art, the film is a significant vehicle for cultural dissemination, a primary means of mass entertainment and an essential measure of consumer economic growth [1], and a powerful medium for educating citizens or inculcating ideas. As such, the film industry has naturally been a popular investment option in the entertainment sector. On the way to maturity, the box office has been emphasized as one of the representative indicators to measure a film's success and intuitively signify profitability. Early film industry studies mainly explore and analyze the potential factors resulting in the movie's success due to the massive investment involved. Some recent studies focus on discussing the negative influence on the movie industry under the circumstance of the outbreak of covid-19. Li et al. have noted that the continuous decline of the

ITEI 2022, December 23-25, Harbin, People's Republic of China

DOI 10.4108/eai.23-12-2022.2329074

global movie industry [1], the crisis of industry and the phenomenon of movie revenue being diminished may be originated from the long-lasting pandemic. Taking the Korean film industry as the research object, Kim et al. studied the short-run effect of social distancing on movie demand and box-office revenue [2], discovering that Covid-19 has dramatically impacted the overall quality of movies and delayed the release of some major movies, resulting in a 34% decrease in sales and 52 million dollars of revenue loss nationwide. However, before the pandemic, the film industry has gone through a hundred-year journey from origin to maturity; further study on the global dataset of the movie industry assists filmmakers adjust direction or style in a better way or provides implications for industry analysts towards long-term investment in movies, especially on a decade's timescale.

Linear regression and clustering methods are widely adopted in research assessing the success of movies or the basics of movie recommendation systems. To assess whether a movie generates high profit, Walanaraya et al. studied the relationship between movie factors and its revenue and constructed a model via linear regression, polynomial regression, and support vector regression (SVR)", regard R-square and the root-mean-square error (RMSE) as a performance indicator, attempt to improve the accuracy of regression model through K-means clustering[3]. Moreover, Ahmad et al. have put forward a mathematical model and claimed that "the successful prediction of a movie plays a vital role in the movie industry" and defined budget, actors, director, producer, story writer, release day, and so on as critical criteria in calculating movie success [4]. In their paper, the lower weight was given for lower-budget movies, while films released on the weekend were assigned a higher weight. This research has driven us to speculate whether the budget is the most predictive attribute, and time series analysis might consider incorporating models to study the periodicity of movie success over a week or even a month. In addition to optimizing the performance of the linear regression model, K-means clustering can integrate with another advanced machine learning approach to develop a new movie-related application; Nayyar et al. built a movie recommender system that combined the K-Means Clustering with K-Nearest Neighbor algorithms [5].

Except above methodology, factor analysis is an alternative in related scientific experiments or statistical research. Another study conducted by Bhave, Kulkarni, Biramane and Kosamkar et al. has different views and suggested that these conventional factors such as cast, producer, director and so on can be classified as 'classical factor' and highlighted 'social factors' as new factor type which is in form of response of the society on various online platforms [6]. Additionally, as Sand argued, geographical location is a key factor easily overlooked by researchers when discussing filmmaking because of film and television production at different places contribute to cultural diversity [7]. By the same logic, it is of high research value to examine if geographical factors such as country can be considered as an essential attribute in movie success. Cross-tabulation is generally applied to gender-related issues in film analysis; Lindner et al. perform cross-tabulations to examine the relationship between the Bechdel Test and box office performance on the bivariate level on the topic of analyzing the effect of female presence in movies on box office returns [8]. Other than cross-tabulations, Analysis of variance have a vital position in content analysis. Schultz et al. applied it to conduct comparisons across the three types of films and found out that popular or award films contained significantly more sensational death actions [9]. With the growth of new social platforms and popularity of various social media, recent studies pay more attention to the social media analytics of movie success prediction. A shift in research towards the sentiment analysis method is a trend. As Timani et al. insisted that "few studies have

tried to demonstrate by analyzing and scrutinizing various features of tweets sent during the movie release" [10], therefore, they miss a more accurate and specific way than the market-based prediction, that is, twitter-based prediction. Sharma et al. share the same research direction of sentiment analysis towards Twitter and YouTube but import more machine learning concepts during the movie's success prediction process [11]. Mitchell et al. fit his collected movie data into time series model and finds a positive and statistically significant relationship between successful films and tourism [12]. In the meantime, Markey et al. conducted a time series analysis to examine whether substantial increasing violence in movies was related to trends in brutal acts of violence [13]. However, few scholars in the same research area attempt to explore the trend or seasonality of movie revenue over a fixed period. Furthermore, the success of movies is also inseparable from stable and large investment; hence it is of extraordinary value to reduce the risk of investment companies and establish a credible business decision model for film investment. Sinha et al. have built their own model that applied random forest classification for predictive analysis [14], contributing to profitable investment decision-making of film production houses.

In this paper, the academic purpose of research is to excavate valuable industry information from thousands of movie samples over 4 decades crawled from the well-known Internet Movie Database (IMDb) and brings enlightenment for potential film investors or researchers. Different from previous papers that emphasize on a few effective methods, for deeper data mining, we attempt to combine machine learning and statistical approach to conduct a comprehensive analysis of the data set and give investment suggestions or industry speculation based on the experimental results.

For the methodology involved in this paper, initially, multiple linear regression methods are performed to predict movie revenue (gross) in a supervised way; meanwhile, on the basis of experiment results, find out, discard, and record variables that fail to significantly increase R square. Such trail may help the researcher in the same field select appropriate variables and build up a more effective linear regression model. Subject to data dependency and lacking research on trend analysis towards movie revenue, time series analysis is an innovative approach exercised to examine underlying trend or seasonality of gross over decades; the period information extracted from decomposition result will be valuable findings in the area of movie investment. Then further observe cross-distribution between prominent genres and ratings via comparative analysis, which might guide potential filmmakers in choosing which genre under the specific rating of work is preferable. Additionally, the subsequent factor analysis continues previous studies on the key factor of movie revenue and redefined budget, gross, and year as the "income factors" of film, while score, runtime, and votes are summarized as the movie's satisfaction factors. Through one-way analysis of variance, the study of exploring whether there is a significant difference in ratings, genres, or countries on movie returns is conducted as a test to indicate the potential influence of the geographic factor (country) on movie success. Furthermore, some of the research has over-excavated box office as the key factor in movie investment and lacked consideration of returns ratio. To fill in the research gap, the second analysis of variance experiments uses mean value to replace missing values of budget and gross and recalculated rate of return to explore the influence of genre, rating, and country on it to provide better investment decisions than the first trial for potential film investors. Ultimate K-means clustering makes use of all scales and nominal factors mentioned before for movie classification and finally classifies the overall samples into three primary clusters; post clustering analysis is performed to discover the latent partnership between representative directors and their followers. The outcome of the

outlined topic addressed in this paper can raise more attention on researchers in the same field and might offer comprehensive insight into the movie industry for potential investors, analysts, or investment model builders.

2 MOVIE DATASET

2.1 Source

The dataset applied in this essay titled Movie Industry is retrieved from Kaggle, covering 7,512 unique films, which are originally crawled from IMDb and contain 15 attributes(variables) including name, company, director, writer, star, genre, rating, budget, gross, votes, score, runtime, country, year and released. More details are summarized in the Table 1.

2.2 Briefing of Variables

Table 1. Movies Dataset Summary

Variables	Measure	Content	Unique	Missing
budget	Scale	Initial budget of a movie		28%
year	Scale	Year of the movie release		0%
runtime	Scale	Duration of a movie		0%
score	Scale	IMDb user rating		0%
votes	Scale	Number of user votes		0%
gross	Scale	Revenue of a movie		2%
name	Nominal	Name of each movie	7512	0%
company	Nominal	Production company	2385	0%
country	Nominal	Country of origin	59	0%
director	Nominal	Director of each movie	2949	0%
genre	Nominal	Main genre of a movie	19	0%
released	Nominal	Release date (YYYY-MM-DD)	3414	0%
rating	Nominal	Rating of the movie (R, PG, etc.)	12	1%
star	Nominal	Main actor/actress	2814	0%
writer	Nominal	Writer of a movie	4535	0%
Measure: The measurement scale of the variable				
Content: Meaning of variables				
Unique: Represents the number of types of variables				
Missing: The proportion of missing data in the total data for the specific variable				

As shown in Table 1, except for “budget”, there are only a few missing values of variables in each column. Under the circumstance that the total number of samples reaches 7,668, the missing rate does not exceed 1%. There are very few null values in rating, genre, and country, which can be ignored in the later analysis since they won’t affect the outcome of data mining. However, the missing value of budget data accounts for 28% totally, which may have a certain impact on subsequent data analysis. Therefore, data preprocessing will be necessarily performed on the analysis involving budget in the following chapters.

3 PREDICTION OF MOVIE REVENUE

3.1 Analysis Method and Tool

In the field of statistics, linear regression is a regression analysis that models the relationship between one or more independent variables and the number of dependent variables using a least squares function called the linear regression equation. Such an approach is simple and commonly adopted by scholars in predicting movie revenues. Since the dataset include several scales variables ("budget", "year", "runtime", "score", "votes") that can feed into the model, while other string type variables are not applicable in regression, this section mainly focuses on exploring the influence of "year", "budget", "votes", "runtime" and "score" as independent variables on the dependent variable ("gross") and attempt to find out whether they are linearly related to each other. Therefore, multivariate linear regression (MLR) is better than univariate linear regression for this issue. Since several scale variables may affect the final revenue (gross) of movies, it is necessary to apply MLR to predict the gross value and verify the fit performance. Such analysis is used via Statistical Product and Service Solutions (SPSS) tool to observe the key independent variables that affect the dependent variable. Furthermore, for the existing concern of the missing value problem, the SPSS analysis tool automatically skips missing values of each variable and only focuses on the intact data section without missing values. Thus, such dataset issues will seldomly affect the actual performance of the MLR.

3.2 Procedure and Key Steps

For the procedure of this MLR, it's feature important to adopt stepwise to select the most critical variables; the variables with the most predictive power are screened first, then the variables with the second predictive power are screened, and so on. Using the probability of F value based on the following stepping method criteria: Entry (if significance level ≤ 0.05) or removal (if significance level≥ 0.10). Moreover, predicted values can be unstandardized, and prediction intervals are averaged. Durbin-Watson test is applied to test whether the observations were independent of each other. Additionally, make use of collinearity diagnosis to determine whether there is multicollinearity in the independent variables by observing the variance inflation factors (VIFs) and tolerances (Tols).

3.3 Goodness of Fit

Table 2. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.740[a]	0.548	0.548	125785485.2	
2	.807[b]	0.652	0.652	110429683.4	
3	.808[c]	0.653	0.653	110190372.0	
4	.809[d]	0.654	0.654	110081264.5	1.889
a. Predictors: (Constant), budget					
b. Predictors: (Constant), budget, votes					

c. Predictors: (Constant), budget, votes, runtime
d. Predictors: (Constant), budget, votes, runtime, score
e. Dependent Variable: gross

Table 2 is generated from SPSS analysis tools and comprehensively summarizes the key indicators from these four multivariate models constructed based on the above methods and criteria. The model summary demonstrates the goodness of fit (GOF) of these four regressions models and proves that the dependent variables can be explained by these five independent variables to some extent. Since R-square is a significant index to measure how well the predicted value fits the actual value (observations), we can assess the GOF of models by comparing their R-square. From observation of model summary, "budget" is the most predictive variable, followed by "votes", "runtime", and "score". It's noticeable that model 1 with a single variable ("budget") can explain nearly 54.8% of gross data; hence the first variable is significant in explaining the dependent variable. After adding the new variable "votes" in model 2, the increase in R-square is approximately 10%, 65.2% of "gross" can be explained by these two variables, and the second variable is of high predictive value in the regression model as well. However, the R-square does not increase significantly after introducing "score" in model 3 and adding "runtime" in model 4 (0.652→0.653→0.654); the new predictors are worth to be kept in the model if the prediction effect is significantly improved. Maintaining the model simpler with fewer predictors is a critical standard; hence "scores" and "runtime" can be discarded from the final linear regression model, and "budget" and "votes" are retained due to the fact that it necessarily enhances the prediction model.

3.4 Analysis of Variance and F-test

Table 3. Analysis Of Varianc[a]

Model		Sum of Square	df	Mean Square	F	Sig.
1	Regression	1.043E+20	1	1.043E+20	6592.417	.000[b]
	Residual	8.596E+19	5433	1.582E+16		
	Total	1.903E+20	5434			
2	Regression	1.240E+20	2	6.201E+19	5085.167	.000[c]
	Residual	6.624E+19	5432	1.219E+16		
	Total	1.903E+20	5434			
3	Regression	1.243E+20	3	4.144E+19	3413.059	.000[d]
	Residual	6.594E+19	5431	1.214E+16		
	Total	1.903E+20	5434			
4	Regression	1.245E+20	4	3.112E+19	2567.814	.000[e]
	Residual	6.580E+19	5430	1.212E+16		
	Total	1.903E+20	5434			

a. Dependent Variable: gross
b. Predictors: (Constant), budget
c. Predictors: (Constant), budget, votes
d. Predictors: (Constant), budget, votes, runtime
e. Predictors: (Constant), budget, votes, runtime, score

Table 3 is the presentation of the result of the Analysis of Variance (ANOVA) and F-test. Degree of freedom (df) and F-value, as well as significance value (P-value), can also be obtained from the above table. The P values of significance in the F-test are all less than 0.05, which denoted more than 95% confidence that there is a significant linear relationship between the dependent variable ("gross") and independent variables("budget", "votes", "runtime", and "score").

3.5 T-test and Multiple Linear Regression Equation

Additionally, the independent variable year is excluded from the regression model due to its significance value is more significant than 0.05. Therefore, only those four independent variables ("budget", "votes", "runtime", and "score") are selected at the beginning of the stepping method. Through the t-test, it can be seen that the significant values are all less than 0.05, and the coefficients in front of each independent variable are not zero, which indicates that the selected variables are valuable in prediction. There is no collinearity among these independent variables since the VIFs of each regression model are less than 2 with Tols greater than 0.5. The ultimate multiple linear regression equation (1):

$$Y(gross) = -33125994.6 + 2.624 * budget + 364.111 * votes \quad (1)$$

4 TREND AND SEASONALITY OF MOVIE REVENUE

4.1 Data Dependency and Time Series

As discussed in the background introduction, time series analysis is not as popular as regression analysis or clustering analysis when it comes to the studies of major trends or seasonality of movie revenue. Such a scientific approach is a specific means of analyzing a sequence of data points collected over a period of time. Data analysts will typically log data points at consistent intervals over a fixed period, but movie releases are not a daily update cycle, so the datasets contain monthly or yearly published releases. The number of movies is not the same. However, the number of samples acquired from this dataset is adequately large, and time series analysis typically requires a large number of data points to ensure consistency and reliability; hence this extensive dataset is suitable for such type of analysis, ensuring a representative sample size. In the previous chapter, the premise of linear regression is that the variables of the data should be independent of each other. We found that "released" in the dataset records the specific release date of each year's movies, and a time series might exit. In other words, in addition to the previously mentioned variables contained in the dataset, time serves as a critical variable in analyzing movie gross, demonstrating how the gross is adjusted throughout the data points and the final result, providing extra information sources as well and a set of dependencies between data in order. It is reasonable to speculate that there is a relationship between movie revenue

(gross) and time; that is, the gross might be dependent on time. Therefore, following the above assumption, it is necessary to adopt time series analysis through Python to decompose time series data and explore whether the gross of the movie manifests a clear trend over time or significant seasonality.

4.2 Data Preprocessing

Due to concern of missing value of gross and the problem that the released date is string type instead of DateTime format, data preprocessing is necessary before time series analysis. Initially, manually remove the rows that contain null values in the "gross" columns of the dataset; 131 invalid samples are removed afterward. The final valid and the unique number of samples is 6,598 when data cleaning is completed. The next essential step is to change the "release" variable from the original string type to date format, then convert it into datetime64[ns] so that Python can recognize it as a date. To enhance the informativity of the index, the current date read by Python is a better candidate and can be set as a new index. Afterward, a new time series data is established on the basis of the initial dataset. Statsmodels is a powerful Python package specializing in descriptive statistics and is generally applied for time series analysis in data mining; some tools and functions from this package is exercised for subsequent test and decomposition.

4.3 White Noise Test, Autocorrelation, and Stationarity

Table 4.White Noise Test

	lb_stat	lb_pvalue
6	3266.8495	0.00
12	5499.319624	0.00

Before conducting decomposing on time series data constructed on previous steps, it is necessary to perform a white noise test, that is, a pure randomness test. If the time series of the dataset is confirmed to be a white noise series, the covariance/correlation coefficient of any two variables is zero, indicating that there is no correlation between any two different variables; thus, it is meaningless for follow-up analysis since it is difficult to find valuable patterns from a purely random sequence. Apply "acorr_ljungbox" (pure randomness test function) from "statsmodels.stats.diagnostic" (module) to conduct a white noise test; the two "lb_pvalue" has indicated the p-value based on chi-square distribution, both of them are 0.0, which are less than 0.05 (see Table 4), it proves that such time series is not a white noise series.

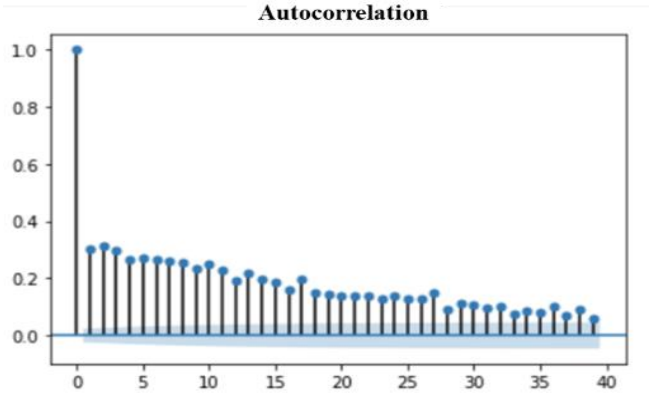

Figure 1. Autocorrelation Plot of Time Series Analysis

Then, to verify the assumption put forward previously, autocorrelation is critical indicator, use "plot_acf" (function) from "statsmodels.graphics.tsaplots" (module) to draw an autocorrelation plot; the diagram shows that the autocorrelation coefficient is greater than 0 for a long time (see Figure 1), revealing that there is a strong long-term correlation between time and movie revenue (gross), the speculation has proved. ADF test is one of the common approaches to test the stationarity of time series. It aims to judge whether there is a unit root in the time series, which denotes that there is no unit root if the series is stationary, and vice versa. As supplementary, implement an ADF test for exploring whether the series is stationary. Deploy "unitroot_adf" (function) from "statsmodels.stats.diagnostic" (module); the result contains test statistic, p-value, lag order, degrees of freedom, and so on. We can find that the test statistic is nearly -6.769, which is less than the 1% critical value (-2.566); it denotes that the actual p-value is much less than 0.01, so it is reasonable to reject the null hypothesis and consider the time series to be stationary. (The null hypothesis here is that there is a unit root, that is, the time series is non-stationary).

4.4 Trend and Seasonality

In order to find out the potential trend and seasonality in the time series, with the help of the "seasonal_decompose" function under module: "statsmodels.tsa.seasonal", the classical decomposition method is implemented to decompose the time series data into "Trend", "Seasonal" and "Residual" respectively. Since this time series is multiplicative, for parameter setting, it is appropriate to select model "additive" as type of seasonal component. The "freq" set as 60 (nearly 2 months) to better observe the seasonality from the chart generated afterwards.

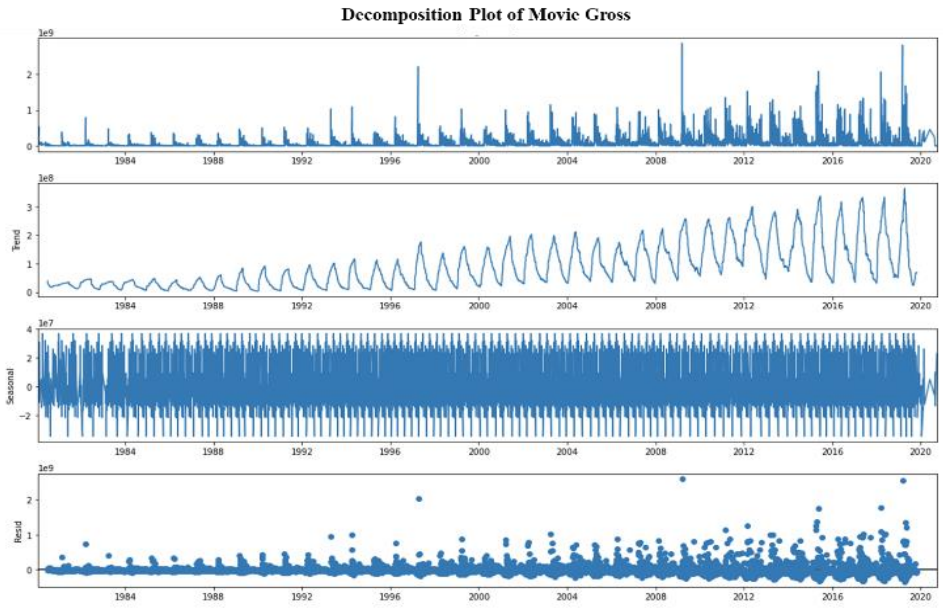

Figure 2. Time Series Decomposition Plot of Movie Gross

By analyzing the Figure 2, in the subgraph representing its trend, it is evident that movie revenue (gross) has a clear upward trend accompanied by fluctuation. There are periodic peaks and troughs, which appear alternately throughout the time series. Such a phenomenon may manifest the existence of periodicity or seasonality. Observing the subgraph showing seasonality will have a more intuitive judgment to characterize it accurately. For the seasonal chart, it can be observed that there are 12 periods between each 4 years, and each period is approximately 4 months, hence the observation reveals that the movie revenue (gross) throughout the entire time series has significant and regular seasonality.

4.5 Discussion

The overall upward trend of film gross in the past four decades may be due to the rapid economic development of various regions, the improvement of people's living standards in the world, and the gradual increase of material desires. Consumers are more inclined to choose movies as the primary way of entertainment consumption, especially in the movie industry, which generally presents a trend of prosperity and positive development, which is commercially valuable and worthy of continued investment and attention in area of entertainment. The seasonality of movie gross is expected to inspire film investors and creators to make better decisions. Gross in each period has experienced a boom and recession, the up and down may be because the film in the industry is concentrated on the selection of actual shooting time and the final release, and there are time gaps exits between two activities. In another speculation, the seasonality of gross might also be due to the periodicity of movie consumption, and it may indirectly reveal the cyclical changes in the economic aggregate. Meanwhile, it can be acquired from the trend of the decomposition diagram that the fluctuation becomes more and more significant with time, and the amplitude becomes larger and larger, which significantly reflects the gradual increase of

volatility with time, which appear as an additional feature of this time series. Such a situation reflects the instability of the film market in the decades of development, which may trigger by the refinement of commercial films, as well as the increase in the number of films entering the market, but the overall quality remains uneven, leading to the drastic changes in the mean and variance of the overall gross.

5 INCOME AND SATISFACTION FACTORS

5.1 Analysis Method

As another statistical approach, factor analysis is widely applied in scientific hypotheses toward movie revenue or seeking success factors of a film. Unlike multivariate linear regression analysis, which aims to effectively predict certain dependent variables, the main purpose of factor analysis is to describe some latent variables hidden in a set of measured variables, which cannot directly measurable but are more basic and statistically important. Such an approach specializes in grouping similar variables together for dimensionality reduction to find out potential latent variables as well. In this section, variables “budget”, “runtime”, “score”, “votes” and “gross” is selected as candidates for factor analysis and observe the explanatory power of the main factor to other variables (how many percentages of variables can be explained by the main factor). Moreover, factors are rotated via the varimax method after applying Principal Component Analysis (PCA) in extraction, guaranteeing that the factors are uncorrelated or orthogonal. This operation avoids the concern of multicollinearity in subsequent regression analysis.

5.2 KMO and Bartlett Sphericity Test

Table 5. Kmo and Bartlett’S Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.672
Bartlett’s Test of Sphericity	Approx. Chi-Square	10766.844
	df	15
	Sig	0.000

The premise of implementing factor analysis is that the data are correlated; thus, it is necessary to apply the Kaiser-Mayer-Olkin Measure of Sampling Adequacy (KMO) as well as Bartlett’s Test of Sphericity (see Table 5). When the KMO statistic is closer to 1, the correlation between variables is more vital, while partial correlation is weaker; hence the effect of factor analysis is better. From the table 5 generated from SPSS, the KMO value obtained is 0.672, signifying that the data selected is suitable for factor analysis. Meanwhile, the significance value of Bartlett’s Test (0.000) is less than 0.05, implying that the null hypothesis is rejected, the correlation matrix of the selected variables is significantly different from the identity matrix, the standard is satisfied, data is spherically distributed. Both tests confirm the existence of a correlation between data, and these factors are fit for factor analysis.

5.3 Retained Principal Components

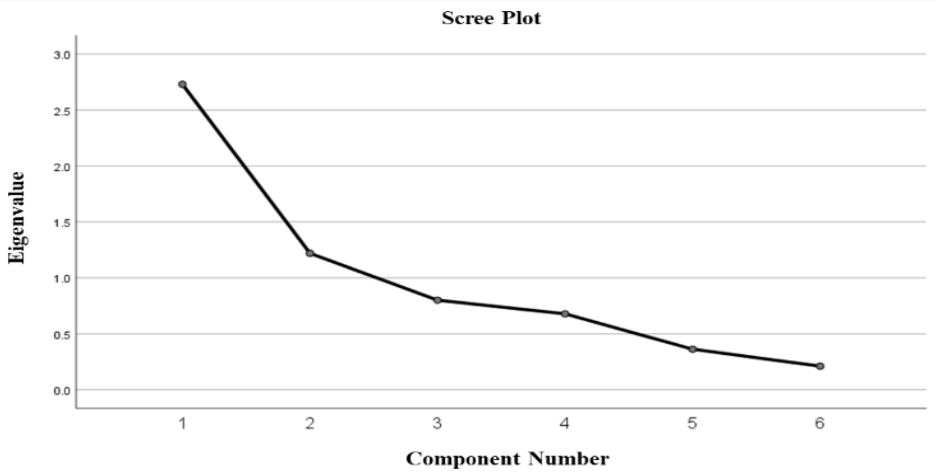

Figure 3. Scree Plot in Factor Analysis

Table 6. Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.729	45.489	45.489
2	1.219	20.322	65.811
3	0.800	13.339	79.150
4	0.678	11.306	90.456
5	0.362	6.036	96.492
6	0.210	3.508	100.000

According to the scree plot collected from the SPSS dimension reduction tool, PCA as an extraction method aims at retaining the most valuable factors. Extract principal components based on a standard that eigenvalue of which is larger than 1, only two components (the first two) are retained, as shown in the Figure 3, and approximately 65.811% of total variance could be explained (see Table 6). Additionally, the other four components fail to fulfill the extraction criteria since they have relatively more minor eigenvalues that are far away from 1.0.

5.4 Communalities and Rotated Component Matrix

Table 7. Communalities

	Initial	Extraction
year	1.000	0.411
score	1.000	0.749

votes	1.000	0.676
budget	1.000	0.776
gross	1.000	0.785
runtime	1.000	0.552
Extraction Method: Principal Component Analysis.		

Inspecting from the result of the communalities (see Table 7), not each factor can be explained by the common factor to a degree of more than 0.7. Apparently, less than 70% of the information of "year", "votes" and "runtime" are explained, which implies that the explanatory power of these two common factors retained is limited for the above factors. While for the other three original variables ("score"," budget", "gross"), common factors still own considerable explanatory power in explaining more than 70% information of them.

Table 8. Rotated Component Matrix[a]

	Component	
	1	2
budget	0.862	
gross	0.819	
year	0.634	
score		0.863
runtime		0.729
votes	0.501	0.652
Extraction Method: Principal Component Analysis.		
Rotation Method: Varimax with Kaiser Normalization.		
a. Rotation converged in 4 iterations.		

After the two components are extracted, adopting the varimax method to rotate, the rationale behinds are to maximize the variances among variables. With the above operation, SPSS generates the following rotated component matrix (refers to Table 8), that is, a more discrete representation of how each factor correlated with these two extracted principal components. This matrix has demonstrated the two extracted principal components, including mentioned original variables; it is noticeable that "votes" is more inclined to the second principal component (component 2); since 0.652 is greater than 0.501, component 2 consists of a greater proportion of "votes" compared with the first component (component 1).

5.5 Summarization

According to the exploration outcome of the previous linear regression chapter, budget and gross are directly related to movie revenue; Therefore, they can be classified into the category of income of films. The score and votes represent the movie's popularity and word-of-mouth among the target audiences, hence regarded as reasonable indicators of customer satisfaction. Inspired from the rotated component matrix, component 1 contained "budget", "gross", and "year" is

summarized as income factors of the movie industry; along the same lines, component 2 involved “score”, “runtime”, and “votes” can be named as satisfaction factors.

6 PROMINENT GENRES AND RATINGS

6.1 Comparative Analysis

In descriptive statistics, the comparative analysis is typically applied to discrete variables to compare the cross-distribution relationship between two or more items. Genre and rating are sometimes overlooked in movie analysis, and few scholars have studied the correlation between genre and rating, especially in terms of their respective frequency of cross-distribution. There are 12 ratings and 19 genres in this dataset, from which we can reasonably surmise different categories of genres probably presents different rating distributions. Verifying such an assumption has considerable research value, and it is worth conducting comparative analysis for deeper data mining towards genres and ratings. Other than previous factor analysis involving multivariate, this chapter emphasizes exploring the cross relationship between these two variables: “genre” and “rating”. Crosstab analysis serves as a commonly used approach in the later analysis of the cross-frequency distribution relationship between them. Chi-Square test, as a nonparametric test and hypothesis testing method for enumeration data, is applied as well to conduct correlation analysis of “rating” and “genre” of this dataset.

6.2 Labeling of Genre and Rating

Data preprocessing is necessary for subsequent analysis since the original data type of “genre” and “rating” is “string”. Among the twelve categories of “rating”, only three categories (R, PG-13, PG) significantly have a quantity of samples more than 1,000, and the rest of the categories (G, NC-17, X, etc.) are far smaller than it. Similarly, among the 19 categories of genre, the number of movie samples in the first three categories (Comedy, Action, Drama) is relatively prominent (quantity of samples of each category is more significant than 1000); by contrast, the sample quantities of remaining categories (Fantasy, Thriller, Sci-Fi, etc.) is insignificant. Due to the distribution of the respective categories of rating and genre is unbalanced, a special labeling strategy is adopted: For rating, adopt labelling in order like R: 1, PG-13: 2, PG: 3, others including missing values are all regarded as ‘Others’ and label as 0 (Others: 0). For genre, specific labeling in order is as follow, Comedy: 1, Action: 2, Drama: 3, others including missing values are treated as ‘Others’ and marked as 0 (Others: 0). Focusing on prominent category is conducive to reduce potential noises in analysis and simplify the labeling process, make it easier to observe following cross-tabulation.

6.3 Crosstabulation and Chi-Square Tests

Table 9. Genre * Rating Crosstabulation

Count		rating				
		0	1	2	3	Total
genre	0	248	1103	361	488	2200
	1	95	984	738	428	2245

	2	58	843	623	181	1705
	3	206	767	390	155	1518
Total		607	3697	2112	1252	7668

Table 10. CHI-SQUARE TESTS

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	487.365[a]	9	0.000
Likelihood Ratio	508.958	9	0.000
Linear-by-Linear Association	37.635	1	0.000
N of Valid Cases	7668		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 120.17

From observation on the crosstab (see Table 9), it is impressive that the number of R-rated movies is the largest among all three prominent genres (Comedy, Action, Drama), significantly larger than the number of PG-13-rated or PG-rated movies, while the quantity of PG-rated movies is the least. The similarity is that comedy has the most significant number of movies among the three main categories of rating (R, PG-13, PG), and drama has the least number of films. In short, the cross-distribution of major genres and ratings is consistent. After the chi-square test(see Table 10), all of the P values of asymptotic significance are less than 0.05, which further proves that there is a significant difference between genre and rating; that is, a strong correlation between them exists, null hypothesis suggesting few differences between them is rejected.

6.4 Discussion

With the ever-increasing pressure on citizens' lives, comedy films, as the most effective film theme for reducing the burden and eliminating fatigue, are very popular and have extensive market value. R-rating, as a relatively loose rating, enables directors and screenwriters to have more space for artistic development without concern for the negative impact of sensitive adult content function on the psychology of underage audiences. It serves as the first choice for commercialization since adult audiences are the larger target groups. Thus, it could be a plausible explanation for why there are the most R-rated comedies in cross-tabulation. Compared with easy-to-understand and witty comedies, dramas may not quickly provide entertainment value to the audience in a short period of time. It embodies more artistic accomplishments and profound themes and requires audiences to possess a high degree of artistic understanding, mindset, and patience. PG-rated movies are relatively limited in terms of material selection and creative space. It is normal to shorten or even delete sensitive clips involving pornography, violence, and blood factors, which violate the artistry and integrity of the movie and weaken the viewing value. It is more suitable for young children to enjoy. Therefore, the number of audiences is the least under such a situation (the least number of PG-rated drama movies in the crosstab).

7 RATE OF RETURN

7.1 One-Way Analysis of Variance

Analysis of variance (ANOVA) is a commonly used statistical model in data analysis, which is applied to test the significance of the difference between two or more samples. It can be seen in plenty of experimental scenarios as well to determine if there are any differences in means between several groups. Continuing the discussion in the last chapter, we wonder whether there exists a difference in the mean value of the gross of each group of movie ratings or genres, which may indirectly lead to different cross-distribution between these two variables (rating & genre). Since each set of experiments involves only two levels of independent variables, one-way ANOVA will be frequently adopted in this section, and the existence of a difference in means can be mainly judged by observing the P value of significance in the ANOVA table. Furthermore, we noticed that quite a few high-grossing films came from the United States (U.S.). Therefore, this chapter will additionally consider "country" as an independent variable in subsequent analysis and use a similar labeling method to make it from a string variable to a nominal one, then continue to apply one-way ANOVA to explore whether prominent countries have a significant difference in the mean of the gross of movies.

7.2 Labeling of Country

From the distribution of the number of movies included in each country, it is evident to find out that the U.S. occupies the largest number of films (5,475) and has a dominant position. And none of the remained countries have a movie quantity of more than 1,000. Therefore, the following labeling strategy is adopted: the movie of U.S. is recorded as "1" (U.S.: 1), while other countries are labeled as "0", and treated as "Others" (Others: 0).

7.3 Analysis Result

Before stepping into ANOVA, the test of normality is necessary to conduct beforehand to examine whether variables involved in ANOVA (except "gross") fulfilled normality. The normality test here adopts two methods: Shapiro–Wilk test (S-W test) and the Kolmogorov-Smirnov test (K-S test). It is noticeable that whether it is the S-W test or K-S test, the P-values of significance (0.00) are all less than 0.05, indicating that the variables including "country", "genre" and "rating" do not satisfy the normal distribution. Inspecting on the result of one-way ANOVA, regardless of the "rating", "genre" or "country", the P value of their respective significance in the homogeneity of variance test table is still less than 0.05, implying that their variances are not homogenous. There are significant differences between groups for "genre", "grade", and "country", as the P value of significance in the ANOVA table is all less than 0.05.

Table 11. Descriptives (Genre & Gross)

					95% Confidence Interval for Mean			
Genre	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
0	2146	88230928.95	175174895.50	3781438.211	80815261.83	95646596.07	682	1670727580
1	2192	44331874.30	71029066.81	1517105.679	41356758.29	47306990.30	309	611257819

2	1673	145508580.80	247515833.10	6051388.207	133639485.90	157377675.70	2970	2847246203
3	1468	38930959.49	95928404.59	2503710.980	34019724.13	43842194.85	596	2201647264
Total	7479	78500541.02	165725124.30	1916313.622	74744027.32	82257054.72	309	2847246203

Table 12. Descriptives (Rating & Gross)

					95% Confidence Interval for Mean			
Rating	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
0	533	52816324.20	138925323.50	5907705.686	41211990.09	64420658.30	596	1083720877
1	3613	42668819.72	79183293.38	1317345.160	40086005.17	45251634.27	1400	1074427370
2	2091	130877145.50	235271042.10	5145072.036	120787146.30	140967144.60	309	2847246203
3	1222	106441540.20	191336623.00	5473466.971	95703097.37	117179983.10	5073	1670727580
Total	7479	78500541.02	165725124.30	1916313.622	74744027.32	82257054.72	309	2847246203

Table 13. Descriptives (Country & Gross)

					95% Confidence Interval for Mean			
Country	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
0	2102	49308650.52	128520782.5	2803218.853	43811275.58	54806025.46	596	1342321665
1	5377	89912360.59	176875022.0	2412106.553	85183653.99	94641067.19	309	2847246203
Total	7479	78500541.02	165725124.3	1916313.622	74744027.32	82257054.72	309	2847246203

More details can be retrieved from the descriptive table (as shown in Table 11, 12, 13); 95% confidence interval for the mean and mean difference of the three variables for the "gross" has shown below. Among the three main genres, "action" has the largest mean (145,508,580.8) as well as the largest span of 95% confidence intervals (133639485.9 ~ 157377675.7) of mean with the most significant standard deviation (247,515,833.1), and "drama" has the most minor mean (38,930,959.49). As for the three prominent movie ratings, the PG-13 rating movies have the highest mean of gross (130,877,145.5); both mean of R-rated (42,668,819.72) and PG-rated movies (106,441,540.2) are significantly smaller than the PG-13-rated one, and PG-13-rated movies have the largest span of 95 % confidence intervals of the mean (120787146.3 ~ 140967144.6) and standard deviation (235,271,042.1) as well. In terms of countries, it is evident that the average gross of films in the U.S. (89,912,360.59) is almost double that of other countries (49,308,650.52). To sum up, from the perspective of movie revenue, action and PG-13-rated movies have the highest movie revenue, while drama and R-rated movies have the most negligible revenue. Compared with other countries, American movies have higher gross (revenue). Moreover, the larger confidence interval of mean and greater standard deviation directly indicate that the action and PG-13-rated movies are highly polarized, with a certain

number of actual values deviating from the mean, and the mean values of these sets of data have a high degree of dispersion and instability.

7.4 Rate of Return

However, emphasizing the final gross (revenue) without considering the budget (cost) is not cater to the logic of solid investment decisions. When inspecting the operation of an enterprise, data analysts typically use the cost-benefit ratio to indicate the profit obtained per unit cost to reflect whether the operation of the enterprise is in satisfactory condition. A higher cost-benefit ratio also suggests that the enterprise is worth investing in. Along the same lines, a new metric is expected to measure whether the economic return of investing in a film is worth its cost to help investors make more prudent decisions in film investment. Hence, the "rate_of_return" is defined as a new indicator to be introduced here, which is calculated by the following formula:

$$Rate\_of\_return = (gross - budget) / budget * 100\% \qquad (2)$$

Data preprocessing is particularly critical in the process of introducing "rate_of_return". Since there are a certain number of missing values in the "budget" and "gross" variables in the dataset, the respective means of budget and gross are applied to fill these missing values. Afterward, for each pair of "gross" and "budget" of a fixed movie, conduct a calculation based on the above formula to acquire the rate_of_return of all samples and record it in a new column (named rate_of_retuen) in the dataset. Ultimately, apply normality and one-way ANOVA test to observe the different outcomes towards mean value.

7.5 Different Outcomes of One-Way ANOVA

Similar to the previously conducted test of normality towards gross and these three nominal variables ("rating", "genre" and "country"), none of the variable groups passed the normality test. However, the one-way ANOVA test results are entirely different from the previous ones focusing on "gross". For the P value of significance in both table of ANOVA and the table of test of homogeneity of variance, only the groups of "rating" are all less than 0.05, which denoted that even if the dependent variable is changed to rate_of_return, the variance is still uneven. The mean value difference between groups is significant as well. However, the P values of significance for other variables groups ("genre" and "country") of these two tables are distinctly greater than 0.05, suggesting that their variances are homogeneous, and the between-group differences are not significant. The results of such one-way ANOVA are utterly divergent from the previous one focusing on "gross".

Table 14. Multiple Comparisions (Rating & Rate_of_Return)

Dependent Variable: rate_of_return						
LSD					95% Confidence Interval	
(I) rating	(J) rating	Mean Difference (I-J)	Std.Error	Sig.	Lower Bound	Upper Bound
0	1	74.547207	18.693143	0.000	37.90353	111.19088
	2	87.481962	19.657513	0.000	48.94786	126.01606
	3	87.380055	21.110983	0.000	45.99675	128.76336
1	0	-74.547207	18.693143	0.000	-111.19088	-37.90353

	2	12.934754	11.642447	0.267	-9.88763	35.75714
	3	12.832847	13.957162	0.358	-14.52701	40.1927
2	0	-87.481962	19.657513	0.000	-126.01606	-48.94786
	1	-12.934754	11.642447	0.267	-35.75714	9.88763
	3	-0.101907	15.224539	0.995	-29.94617	29.74235
3	0	-87.380055	21.110983	0.000	-128.76336	-45.99675
	1	-12.832847	13.957162	0.358	-40.1927	14.52701
	2	0.101907	15.224539	0.995	-29.74235	29.94617

*.The mean difference is significant at the 0.05 level

In ANOVA, Post-Hoc Multiple Comparison (PHMC) applies the least significance difference (LSD) method, the significant level (α) is set to 0.05 as default, and the pairwise comparison between groups is completed with a T-test. Because of the high sensitivity of this test, even small differences in means between levels can be detected for further analysis. Since there are significant differences in "rating" between groups, we apply PHMC to further explore which specific groups had larger differences. From the results of PHMC (see Table 14), the significant difference between group "0" and group "1", group "2", group "3" are worth paying more attention to. Due to the P value of significance is all zero between them, indicating that the other types of ratings do have evident mean differences towards the three mainstream ratings (comedy, action, drama), which directly verifies the correctness of the labeling strategy in data preprocessing procedure. Moreover, a significant value of 0.995 also represents that group "2" (PG-13) is highly similar to group "3" (PG) with few differences on the mean.

Table 15. Descriptives (Genre & Rate_of_Return)

					95% Confidence Interval for Mean			
Genre	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
0	2200	12.81113	293.795107	6.263733	0.52768	25.09458	0.000	12889.387
1	2245	17.63298	576.954435	12.176805	-6.24600	41.51196	0.000	26165.847
2	1705	6.25598	190.208979	4.606476	-2.77896	15.29093	0.000	7849.054
3	1518	24.49018	514.477464	13.204759	-1.41134	50.39170	0.000	15699.108
Total	7668	15.07734	427.357136	4.880338	5.51055	24.64414	0.000	26165.847

Table 16. Descriptives (Rating & Rate_of_Return)

					95% Confidence Interval for Mean			
Rating	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
0	607	89.38130	1216.914267	49.393026	-7.62099	186.38359	-1.000	26165.847
1	3697	14.83410	367.349724	6.041642	2.98882	26.67938	-1.000	15699.108
2	2112	1.89934	5.298989	0.115304	1.67322	2.12546	-1.000	88.176
3	1252	2.00125	8.526610	0.240976	1.52849	2.47401	-1.000	179.461

Total	7668	15.07734	427.357136	4.880338	5.51055	24.64414	-1.000	26165.847

Table 17. Descriptives (Country & Rate_of_Return)

					95% Confidence Interval for Mean			
Country	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
0	2193	1.74897	19.958438	0.426194	0.91319	2.58476	-1.000	784.005
1	5475	20.41600	505.511645	6.831862	7.02283	33.80916	-1.000	26165.847
Total	7668	15.07734	427.357136	4.880338	5.51055	24.64414	-1.000	26165.847

Additional divergence of experimental results of the two one-way ANOVA can be more intuitively captured from the new series of the descriptive table (see Table 15, 16, 17). Among the three main genres, drama movies have the largest mean of the rate of return (24.49018), the largest standard deviation (514.477464), and a 95% confidence interval of mean with the most extensive range (-1.41134 ~ 50.3917), action movies possess the most minor mean of the rate of return (6.25598). For the prominent movie ratings, the R-rated movies own the highest mean value (14.83410), the widest 95% confidence interval of the mean (2.98882 ~ 26.67938), and the largest standard deviation (367.349724). In contrast, both means of rate of return of PG-13-rated (1.89934) and PG-rated movies (2.00125) are much smaller than the R-rated one. The category of countries signifies that the average return rate of U.S. movies (20.416) is approximately 12 times that of other countries (1.74897).

7.6 Further Discussion

In summary, when taking the rate of return into account, rather than purely focusing on gross, drama films are more worth investing in than action movies, the latter simply has high box office but not considerable returns, yet there is a certain risk of unstable returns (high standards deviation). What's more, R-rated movies can also be a good investment direction in all categories of movie rating. Surprisingly, R-rated and drama films with the lowest gross values have the best return rate, implying they have meager budgets. These categories of movies are suitable for film investors eager for low cost and high return on investment. The gap between U.S. films and films from other countries is more significant in the mean value of the rate of returns than the average gross. This phenomenon manifests the prosperity of the American film industry. Under the premise of high gross, the production cost is relatively lower in the U.S.; thus, the rate of return can reach a considerable level. Consequently, investing in films made in the U.S. is more robust regarding expected returns. The above discussion results might bring profound enlightenment to potential film investors.

The reason for the boom of the American film industry can be traced back through history; Hollywood in Southern California has become the world center of the film industry after the Hollywood period. It is one of the most densely populated film regions worldwide and is home to an industrial settlement with a high concentration of film-related industries and manufacturers. Compared with other countries and regions, the production there is more efficient, and the division of labor is precise. The blockbuster production, contract system, and different operating

modes are more mature; hence the great films produced possess a more avant-garde aesthetic style and higher profits.

In the previous section, the number of frequency distributions in the cross-tabulation between genre and rating indicated that R rating and comedy are the most popular ratings and genres. However, from the analysis outcome and discussion of this chapter, it is evident that the genre that chooses drama as the genre of a movie will have a more lucrative return. We surmise that it may be due to the production, shooting, and directing thresholds of dramas being relatively high, which require filmmakers to have profound artistic skills and experience; thus, it is trickier to get started than simple and easy-to-understand comedy films, especially for those young directors, screenwriters and leading actors without adequate artistic precipitation and experience. From another perspective, it is also possible that there are lacking filmmakers in the talent market who are skilled in drama production. Training more filmmakers specializing in drama movies for profits could be a potential business opportunity in the near future.

8 MOVIE CLASSIFICATION AND REPRESENTATIVE DIRECTORS AND COMPANIES

8.1 Purpose of K-means Clustering

As mentioned in the introduction, K-means clustering, a fast and simple clustering analysis approach for large and complex datasets, frequently appears in the field of movie industry analysis and data mining, especially applied in movie rating and popular movie recommendation systems. Given that this dataset possesses several string variables (“director”, “writer”, “company” and “star”) that are not used in the previous analysis, they can be regarded as informative labels in the investigation after clustering to supplement the interpretation of the clustering results. Therefore, in this chapter, all of the non-string variables are treated as the features of movies to perform K-means clustering and classify thousands of movie samples into several specific clusters so that each data point (movie sample) belongs to the cluster corresponding to the nearest centroid and more insight of industry information from the clustering outcomes can be gained afterward.

8.2 Optimal Number of Clusters

In order to avoid the negative effect of missing values in “votes”, “score” and “runtime” on the results of K-means analysis, a small number of missing values are filled with the mean of the respective variables. Given the context, it is crucial to determine the appropriate number of clusters to optimize the performance and results of K-means clustering. When determining the number of clusters, the sum of squared errors within groups (SSE) is a crucial indicator in the elbow method, and the silhouette coefficient serves as another vital evaluation indicator of the density and dispersion of the cluster in the average silhouette method. As a powerful tool, Python provides more stable and superior performance in evaluating the SSE and silhouette score. The figure 4 demonstrates the visualization of the elbow method applied in the movie dataset that removed string variables.

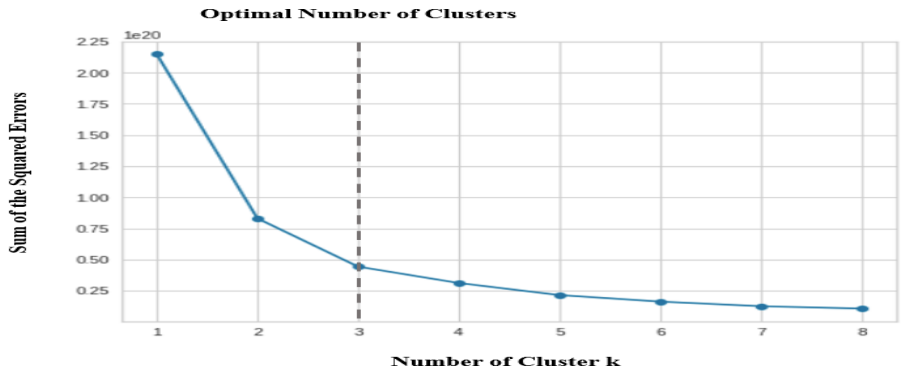

Figure 4. SSE Plot (Elbow Method)

As the number of clusters increases, the number and distance of samples within each cluster become smaller and closer, and the SEE value decreases; thus, it is essential to pay attention to the slope change. When observing SEE decrease slowly and change of slope is inapparent, it is considered that the effect of further increasing the number of clusters cannot be enhanced, and the elbow point indicates the optimal number of clusters. Based on this principle, the optimal number of clusters is selected as "3" from above SSE plot since the decline after "3" is not significant as before.

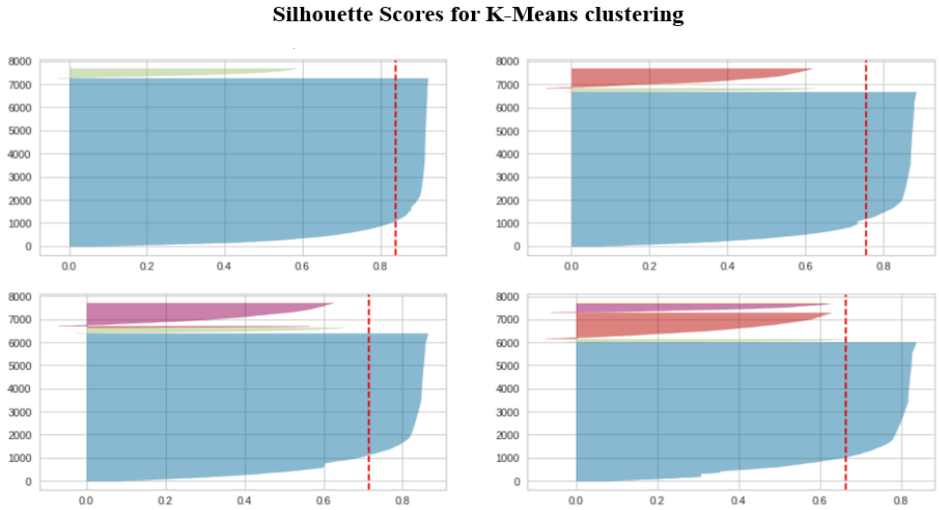

Figure 5. Silhouette Scores for 2, 3, 4, 5 Clusters

The Silhouette score for 2, 3, 4, 5 clusters are 0.839, 0.755, 0.715 and 0.665 separately, as shown in the Figure 5, the value of number of clusters as "4" and "5" looks to be suboptimal due to lower silhouette scores accompanied by clusters with below-average silhouette scores and wide fluctuations in the size of each silhouette subgraph. Thus, one can select the optimal number of

clusters as “3” because its higher silhouette score and the thickness (size) of each cluster in the second subgraph (number of clusters as “3”) is more uniform than the subgraph located at the top left (number of clusters as “2”) with one cluster thickness much more than the other. Ultimately, combining both results of the elbow method and the test of silhouette score, 3 clusters (K=3) can perform the best clustering outcome.

8.3 Results of Clustering

Table 18. Final Cluster Centers

	Cluster		
	1	2	3
rating	1	2	2
genre	1	1	1
year	1999	2012	2007
score	6.3	7.2	6.7
votes	54410	583606	276641
country	1	1	1
budget	27200270	161593473	82075797
gross	32491158	966851204	299788837
runtime	106	130	115
rate_of_return	13.967	6.650	25.507

Conduct K-means clustering with a maximum number of iterations as “100”, the table recording final cluster centers has generated as shown above. Convergence is achieved at iteration “25” due to no or minimal change in cluster centers. After counting the number of cases in each cluster, we found that cluster 1 has the most significant number of cases (6,700), cluster 3 has only 827 cases, and cluster 2 has the least number of cases, only 141. More implicit information can be extracted from the above table.

Take the three final cluster centers as representatives of each cluster. It is noticeable that 1999, 2007, and 2012 are relatively representative years in four decades from 1980 to 2020 (see Table 18). America and comedy are the leading countries and genres in these three clusters. The R rating in cluster 1 is dominant, unlike clusters 2 and 3 (PG-13 is the primary rating). Moreover, cluster 1 has the smallest runtime and score, while cluster 2 has the largest one for both of them. Taking cluster 2 as a reference, the movies with low budgets and with small gross are concentrated in cluster 1, and the film within cluster 3 are all high budget with large gross. But the cases in cluster 2 have the lowest returns, and the returns for cases in cluster 3 are significantly higher than the other clusters. In terms of votes, the polarization is more evident between clusters. Cluster 2 has the highest votes, much higher than the other two clusters, and is nearly 11 times higher than that of cluster 1, which has the lowest votes.

8.4 Discussion on Final Cluster Centers

According to the information of the ultimate cluster centers, it is reasonable to speculate that the runtime length seems to be positively correlated with both the score and votes values. But votes and score as satisfaction factors fail to reflect the film's rate of return credibly. Notably, the movies in cluster 2 possess the highest scores and votes, but the lowest rate of return among the three clusters. Popular movies with audiences may be of admirable artistry and considerable entertainment value but not necessarily worthy of commercial investment. In conclusion, K-means clustering conducted before roughly divides all movies into three distinctive categories. The first category (cluster 3) is collection of commercially successful great films that are large-scale production with considerable returns, while the second category (cluster 2) movies are generally in a low rate of return but remain artistic or entertaining with a good audience reputation and high satisfaction. The third category (cluster 3) movies are somewhere in between, primarily normal movies with moderate income of films and audience satisfaction but occupy the largest number.

8.5 Representative Director, Company, and partnership within Clusters

After a concise discussion of the final cluster centers, we continue to count the most significant directors and companies under each cluster as representatives and the main objects of post-cluster analysis. To start with, Steven Spielberg appears most frequently in cluster 3; thus, he is selected as the representative director of this cluster. In his directed films, Harrison Ford has emerged as a prominent star multiple times, especially in the series of Indiana Jones, and has a long-term binding partnership with him. But the screenwriter of each of Spielberg's films is unique and different; hence, they are only a short-term cooperation relationship with him. Since he is an American director, all the films he directs belong to the U.S., and the film genres span a wide range. The Indiana Jones series, as his masterpiece, has the highest votes, score, and rate_of_return and are commercially successful film. But his other independent films possess normal or even low audience satisfaction and returns.

Both Peter Jackson and David Yates are representative directors in Cluster 2. The former has long-term collaborated on The Lord of the Rings series with Elijah Wood (star) and The Hobbit series with Fran Walsh (writer) and Ian McKellen (star). In comparison, David Yates built long-term cooperation with Steve Kloves (writer), and Daniel Radcliffe (star) Harry Potter series (from the fourth to the seventh) and established long-term cooperation on the Fantastic Beasts series with J.K. Rowling (original author) and Eddie Redmayne (star). Compared to Steven Spielberg's masterpieces in the late 20th century, these movies are classic series films of the early 21st century. However, the rate of return, votes, and scores of the Lord of the Rings series are generally better than the Harry Potter series, and the former is more commercially successful. Moreover, each fantasy movie directed by Peter Jackson has a remarkable long runtime (larger than 2 hours). And the other difference is that neither David Yates nor Peter Jackson is American director. Therefore, none of these famous film series were produced in the U.S. Unlike Steven Spielberg's genre-rich style, David Yates and Peter Jackson have a fixed art style, and the genre mainly focuses on action and fantasy.

As the dominant director in cluster 1, Woody Allen is one of the most respected and famous directors in the United States; and also served as a screenwriter in most of the films he directed and even acted as the main actor (star) in the early films (1980~1985). Mia Farrow, and Scarlett

Johansson are stars who have collaborated with him continually. But the movie he directed and released between 1980 and 2004 typically earned a few incomes, sometimes with negative returns, which can be treated as his growth stage and accumulated experience in these two decades. After 2004, the situation improved greatly, and the commercialization of his directed films matured with considerable returns. Unlike the representative directors of the first two clusters, the genres of movies he participated in and directed focus on comedy. In short, similar directors are clustered together. In contrast, directors with different styles belong to disparate clusters, which strongly proves that the outcomes of this K-means clustering are satisfactory and trustworthy.

Additionally, in terms of movie companies, Universal Pictures is the representative in cluster 3, and Warner Bros. released the most movies in cluster 2. The quantity of released movies by Universal Pictures, Columbia Pictures, Paramount Pictures, and Warner Bros. in cluster 1 is significant. Judging from the characteristics of each cluster summarized earlier, Universal Pictures may be the most investment-minded company because cluster 3 concentrates the largest number of movies with high return rates. Warner Bros. is probably the most popular movie company with the highest audience favorability, as movies in cluster 2 generally own high audience satisfaction. Columbia Pictures and Paramount Pictures and these two companies are well-known American film production and distribution companies. A large number of films released by themselves indicates their large-scale manufacture and excellent operating status as well, denoting the undeniable leading position of the U.S. in the movie industry.

9 CONCLUSION

Movie revenue prediction in the conversation has been gaining popularity among industry analysts and researchers. In this work, we have discovered that the budget and votes of a movie are of high predictive value in the multivariate regression model, such a finding can draw the attention of scholars in the same field and assist optimize their corresponding prediction model since they mainly concern on budget, but few researchers have previously noticed the role of votes in predicting movie revenue. Time series analysis serves as an innovative approach in such research area, revealing the data dependency between released time and movie revenue, evident seasonality (four months as a fixed period), and upward trend of movie revenue with high volatility. The limitation of this section is the lack of analysis of heteroscedasticity to deeply dissect the reasons for the gradual increase in volatility. Due to the different preprocessing of missing values, this section does not incorporate the rate of return introduced in later chapters into the time series analysis. Still, there is potential research value in continuing to delve into the trends and periodicities of budgets or rate of return. Additionally, the two retained principal components, including “budget”, “gross”, “year” and “votes”, “score”, “runtime” separately, are summarized as the income factor and satisfaction factor of a movie in factor analysis. It is worth an extra experiment to examine whether these two retained principal components help to improve the performance of movie clustering.

Moreover, the consistency of the cross-distribution of prominent genres and ratings and their significant correlation are pointed out via comparative analysis. Furthermore, introducing the rate of return in ANOVA demonstrates a more informative and practical investment scheme than purely gross: R-rated and drama films with the lowest gross values have the best return rate. Continuing the conclusion of ANOVA, the distribution of film companies in the cluster analysis

double verifies the United States' leading position in the movie industry. Movies in the entire dataset can be divided into three main categories, one with high returns, one embodying considerable artistic value and audience satisfaction, and the last one in between and more common. In addition, the distribution of directors within a cluster is consistent with the characteristics of the cluster to which they belong and latent partnership between representative directors and their followers are detonated. We believe that the finding of each analysis outlined in this paper will not only enhance the prediction model or performance of clustering associated with the research area of the film industry, but also provide comprehensive insight into the movie industry for potential investors, filmmaker, or investment models builders.

REFERENCES

[1] Z. Li, D. Wang & Y. Wu. (2021) Sentiment Analysis on Chinese Movie Comment with LDA Model. *2021 2nd International Conference on Big Data Economy and Information Management (BDEIM)*. pp. 424-428.

[2] Kim, I. K. (2020) The impact of social distancing on box-office revenue: Evidence from the COVID-19 pandemic. *Quantitative Marketing and Economics*, *19*(1): 93–125.

[3] P. Walanaraya, W. Puengpipattrakul & D. Sutivong. (2021) Movie Revenue Prediction Using Regression and Clustering. *2018 2nd International Conference on Engineering Innovation (ICEI)*. pp. 63-68.

[4] J. Ahmad, P. Duraisamy, A. Yousef & B. Buckles. (2017) Movie success prediction using data mining. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. pp. 1-4.

[5] R. Ahuja, A. Solanki & A. Nayyar. (2019) Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor. *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. pp. 263-268.

[6] A. Bhave, H. Kulkarni, V. Biramane & P. Kosamkar. (2015) Role of different factors in predicting movie success. *2015 International Conference on Pervasive Computing (ICPC)*. pp. 1-4.

[7] Sand, S. A. (2019) Small places, universal stories. Diversity, film policy and the geographical dimension of filmmaking. *Nordisk Kulturpolitisk Tidskrift*, *22*(1): 8–25.

[8] Lindner, A. M., Lindquist, M., & Arnold, J. (2015) Million dollar maybe? The effect of female presence in movies on box office returns. *Sociological Inquiry*, *85*(3): 407-428.

[9] Schultz, N. W., & Huet, L. M. (2001) Sensational! violent! popular! death in American movies. *OMEGA-Journal of Death and Dying*, *42*(2): 137-149.

[10] H. Timani, P. Shah & M. Joshi. (2019) Predicting Success of a Movie from Youtube Trailer Comments using Sentiment Analysis. *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. pp. 584-586.

[11] T. Sharma, R. Dichwalkar, S. Milkhe & K. Gawande. (2020) Movie Buzz - Movie Success Prediction System Using Machine Learning Model. *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*. pp. 111-118.

[12] Mitchell, H., & Stewart, M. F. (2012) Movies and holidays: the empirical relationship between movies and tourism. *Applied Economics Letters*, *19*(15): 1437-1440.

[13] Markey, P. M., French, J. E., & Markey, C. N. (2015) Violent movies and severe acts of violence: Sensationalism versus science. *Human communication research*, *41*(2): 155-173.

[14] A. A. Sinha, S. V. V. Krishna, R. Shedge & A. Sinha. (2017) Movie production investment decision system. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. pp. 494-498.