COSC 3000

Data Visualization Project Report

Student: Yinhuang Huang

43767092

Content List

Introduction……3

Data……4

Data source……..4

Data preparation……4

Method & Results……5

Film industry development trend……..5

Correlation with rating……..9

Correlation with revenues…… ……..14

Conclusion……17

Findings ……17

Future investigation……17

Introduction

Internet Movie Database(IMDb) is an online database of information related to different types of resources of entertainment, including world films, television programs, music and so on. IMDb is one of the biggest movie data base online, it allows the registered users to rate a movie according to their personal like. The rating of a movie can be calculated with IMDb's algorithms based on each user's rates on the movie, and an overall rating would be represented as evaluation of registered users' opinion.

The rating score from IMDb is usually considered as an important criterion for judging the quality of a movie. Because the significant number of registered users who are volunteer to rate, the rating can represent the opinion of the most of audiences more precisely than other rating system like Metascore which is based on experts' and critics' opinions. It is also the reason that IMDb rating was chosen in this project over the others. The purpose of this project is to analyse the factors that can influence the rating and revenue of movies, and the general development tendency of film industry. The potential relation between some interesting factors like budget, revenue etc also would be investigated. This can be useful for the following reasons: it can help the potential audiences to pick a movie without knowing much detail of the movie. It can also help the film maker to make well-founded prediction, based on the development track in past few decades, about the possible response of market and audiences and future development.

To achieve the goal of the project, the following aspects would be covered: the total number of movies, total revenues and the development trend; the correlation between rating and released date, genres, production countries and budget; the correlation between revenues and rating and budgets.

The objective of this report is to investigate the correlation between rating of movie and factors such as genres, released year, production country and so on. A variety of data visualisation methodologies would be used to present the data intuitively and efficiently and analyse how

the factors affect the rating of movies in this report.

Data

■ Data source

As mentioned previously, the data used for this project was sourced from the IMDb online datasets which is available to download from https://datasets.imdbws.com/. The datasets downloaded originally was a multi-dimension table with the mixture of numbers and strings. The table contains the following information including movie IDs, movie titles, genres, production companies, production countries, released dates, budgets, runtime, revenue, ratings and vote counts. Matlab would be used as the main tool for the data visualisation, in this case, Matlab would read through the table and store the rows of the table as cells which contains information of each movie in the workspace.

■ Data Preparation

The IMDb datasets contains information about 4,734,693 movies produced since 1873. The large size of this dataset makes data processing take a long time, to accelerate the process, information of movies produced between 1935 and 2016 was selected as the source for visualisation. It is also worth to notice that there are some missing information in the raw data and some are incorrect, for example few films have some runtime information, and few films have rating over 10 which is impossible since the scale of rating was set between 1 and 10. Before performing any statistical preparation, the entries with incorrect information would be filtered out and the entries with missing information should be completed by filling with correct information found.

After the filtering and patching the raw data, some statistical preparations were carried out based on that. A few cells were created to store the ratings data corresponding to different countries, the number of element in each cell represents the number of movie production of the country and the average rating of each country can also be determined easily from that. Arranging and storing the data with respect to their genres, production country, released data

etc. is required for further operation. The Matlab built-in function ‘median’ would be used to calculate the average rating. In the raw dataset, data of runtime, budget, revenue, rating and vote counts are recognised as string type data, converting them to number in Matlab was necessary to compare and calculate the average.

Methods & Results

The first aim of the project was to investigate the development track of film industry, which requires to show the change of the number of movies with respect to different properties, such as number of production, rating, revenues and so on. To show that, univariate data and bivariate data were used to generate the graph. The univariate data can indicate the quantitative relations of different categories of same variable quickly and clearly. The aim of bivariate visualization was to highlight certain trends in relationships between two fields. For correlation of multiple variables, multivariate visualisation could help with the greater insight into relationship with higher degree of complexity.

■ Film industry development trend

This section aims to highlight the information about the quantitative relations between different genres and production of countries of movies and seek a certain trend of change of movie production. To show the quantitative relations between different genres and production of countries of movies, the pie chart was chosen over bar chart because it indicates the difference in terms of the share of categories more efficiently. The method used to see the development trend of movie production was plotting the dots against the time and then applying the linear fit to seek to general trends.

The data in the figure below indicates the shares of different genres among of all movie produced between 1935 to 2016. Different colour represents different genres in the pie chart below. The drama and the comedy take up the most area of the pie, which means these two are the most common genres of films between 1935 to 2016. The drama movie takes up about 20 percent of all movie. Second to that, comedy movies occupy about

16 percent of all. Action movies and thriller both take up about 11 percent.

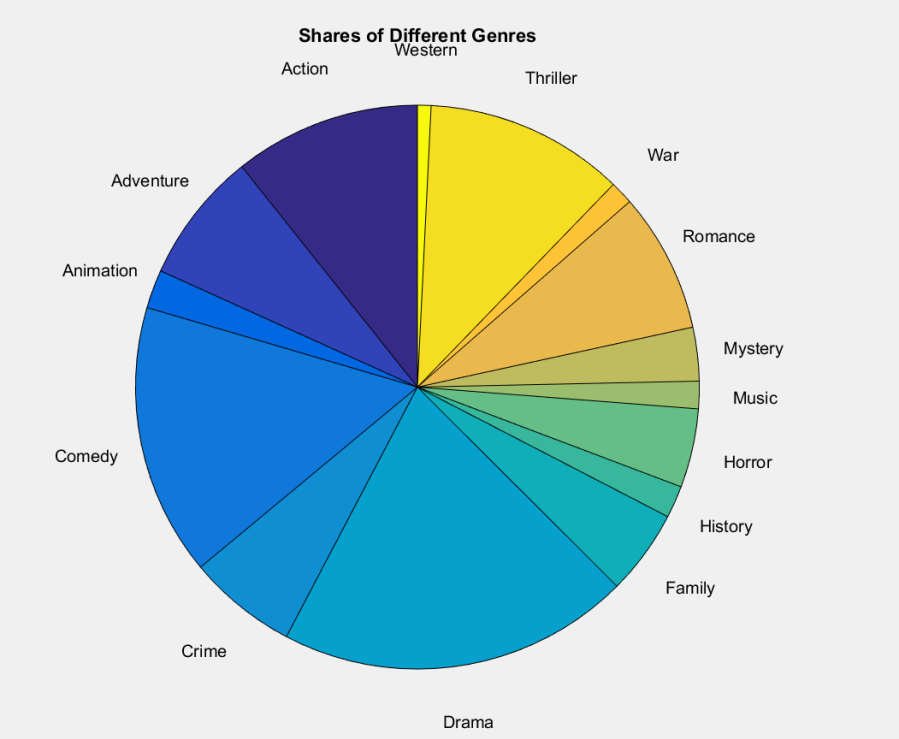

Figure 1 Pie chart of shares of different genres

The graph shown above says that that the film industry tends to produce drama movies and comedy movies. It might be because some certain genres over outperform the others in terms of public preferences, and it also might be the result of adjustment of industry itself. To show that the drama and comedy stand out from all genres, another pie chart about the comparison of high quality movie of different genres would be generated, along with a line chart shows that change of each genre of movie over time.

The follow figure highlights the shares of high quality movies of different genres. The movies with rating greater than 8 were considered as high quality in this case. As shown below, as that drama movies have the greatest part of shares of all movies, the drama movie takes up about 30% of all high-quality movies, which is more than the share it occupies in all movies. However, the comedy movie appears to weight less in the quality movie, and the crime movie and animation movie increase their shares in high quality movie. That differences between the

shares of different genres in all movies and the shares of different genres in high quality movies indicates that the rating does vary between genres. Movies of some genres like crime, animation etc. outperform the others. The more specific correlation between genres and rating would be discussed in the next section.

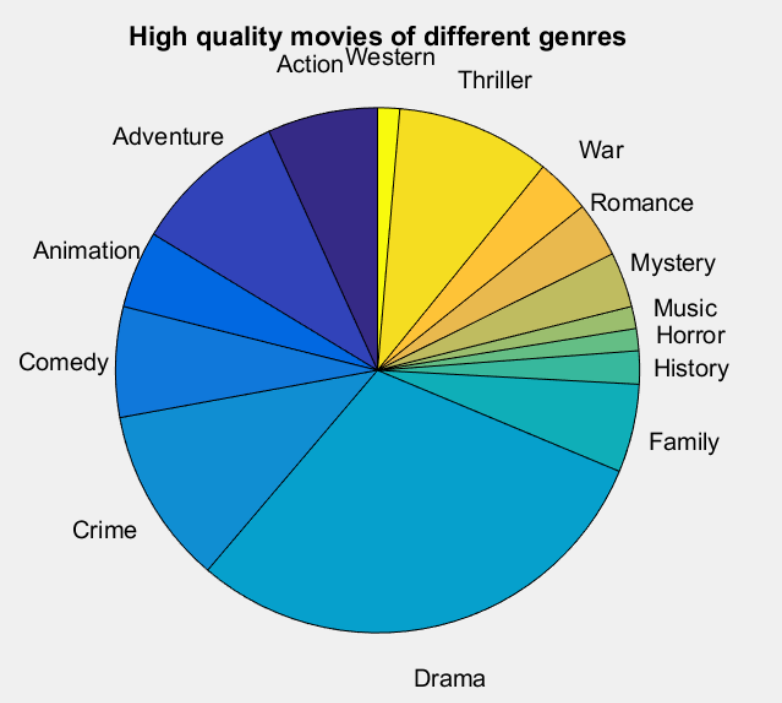

Figure 2 Shares of high quality movies of different genres

To show the development tendency of the film industry, data should be plotted against time. The graph below highlights the change of the number of movie production and total revenues of all movies with the change of time from 1935 to 2016. Two group of data were plotted with respect to different vertical axes on the same graph, the black line represents the number of number of output of movie every year which corresponds to the left y-axis. And the red line shows the change of total revenues of movie produced every year, which is corresponding to the right y-axis. The two lines were generated by connecting data points with straight line to show the general trend of growth. Because two data sets were in different coordinates, the value cannot be compared directly.

From the figure, it can be found that both number of movie produced every year and total

revenues every year have increased significantly since 1935. Specially between 1990 and 2000, they both grew much more quickly than any other period on the graph. After 2000, there was fluctuations occurring to the output of film, but the total revenues maintain its growth till 2010. It makes sense that the total revenues increase as the output of films increasing, but the data visualizations show that when the number of films produced dropped, the total revenues does not necessarily decrease.

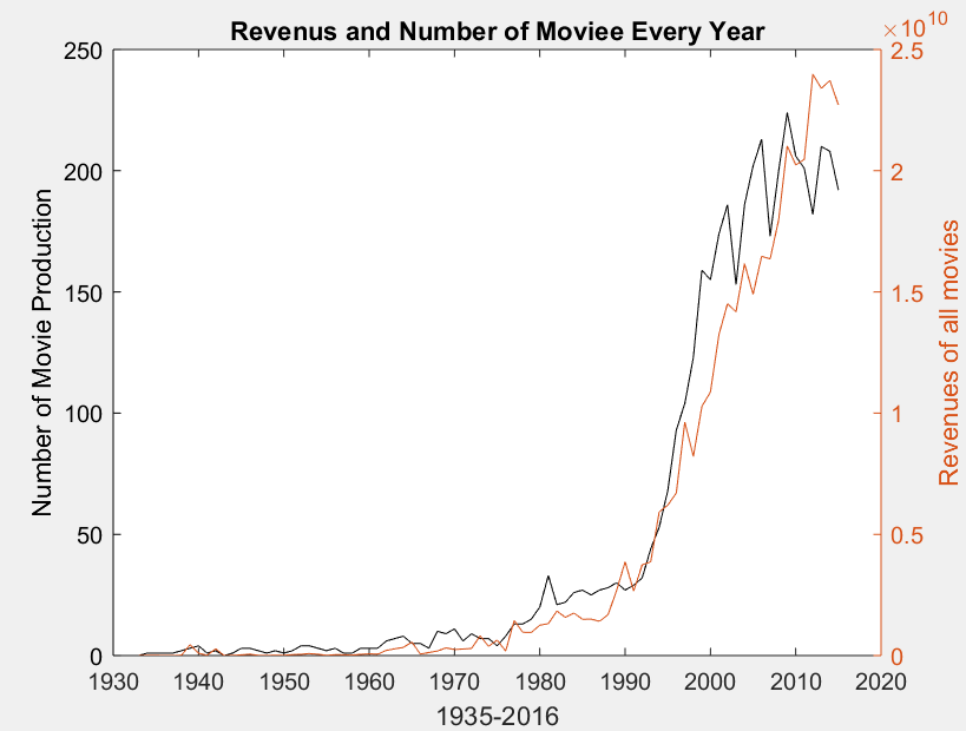

Figure 3 Revenues and Number of Movie Every year

The total revenues of movies also related to the income of each movie, the average revenue of single movie might have varied with different time. The average revenue of single movie can be determined by dividing the total revenues by the total number of movie produced every year, which would generate a group of data that contains the average revenue of single movie against time. To show the change of average revenue, data points of the group would be plotted with a linear fit for indicating the general trend over time. The figure generated is as following:

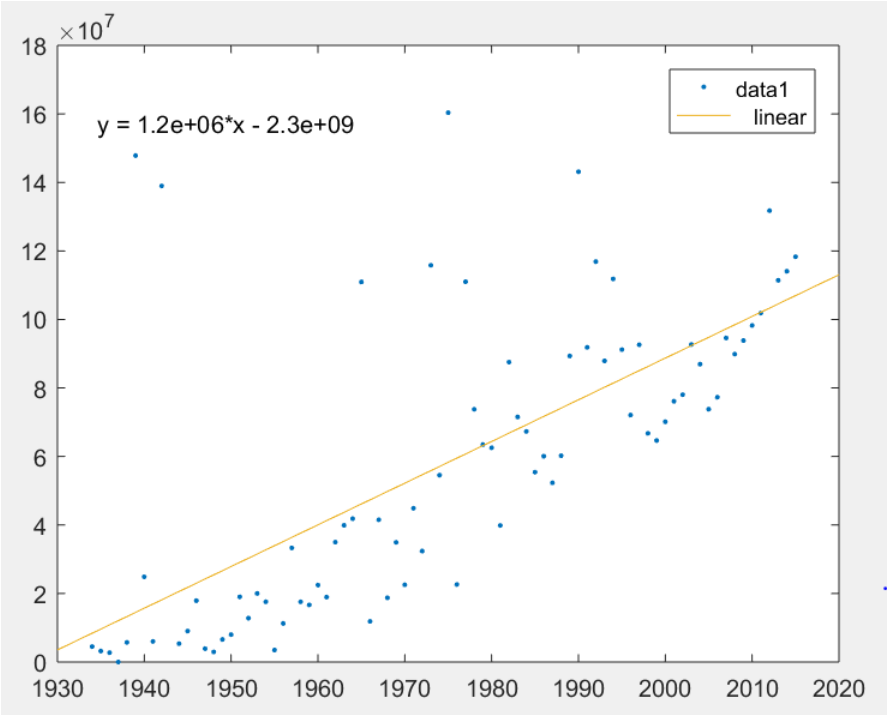

Figure 4 Average revenue of single movie every year

In the graph above, the blue data points of data set 'data 1' represent the average revenue of single movie every year. The yellow straight line is the linear fit of the data set, which highlights that the average revenue of single movie increases as time increasing. There were some years in which the average revenues were relatively high or low, and they did not fit the model. But, overall, the general tendency is growth. This can explain why the total revenues maintain increasing while the number of production of movie decreasing, the growth of average revenues causes the increasement total revenues.

■ Correlation with rating

The section would investigate the correlation between rating and production countries, release year, genres. The bivariate date and multivariate data were used for the visualization. For genres and countries, because they are not quantified variable, to show their correlation with ratings, diagrams of probability distribution function were used. First, a bar chart of production counties against rating would be generated to show ratings vary for different production. And then apply the same method for ratings and genres.

To generate the desired bar chart, data was firstly sorted by the production countries and genres, and then put the sorted data into two different data group along with the corresponding ratings. Each data group contains the information about the ratings of movies of different production counties or genres. The average movie ratings of each country or genre was obtained and used to plotted against the production countries and genres in the bar charts below.

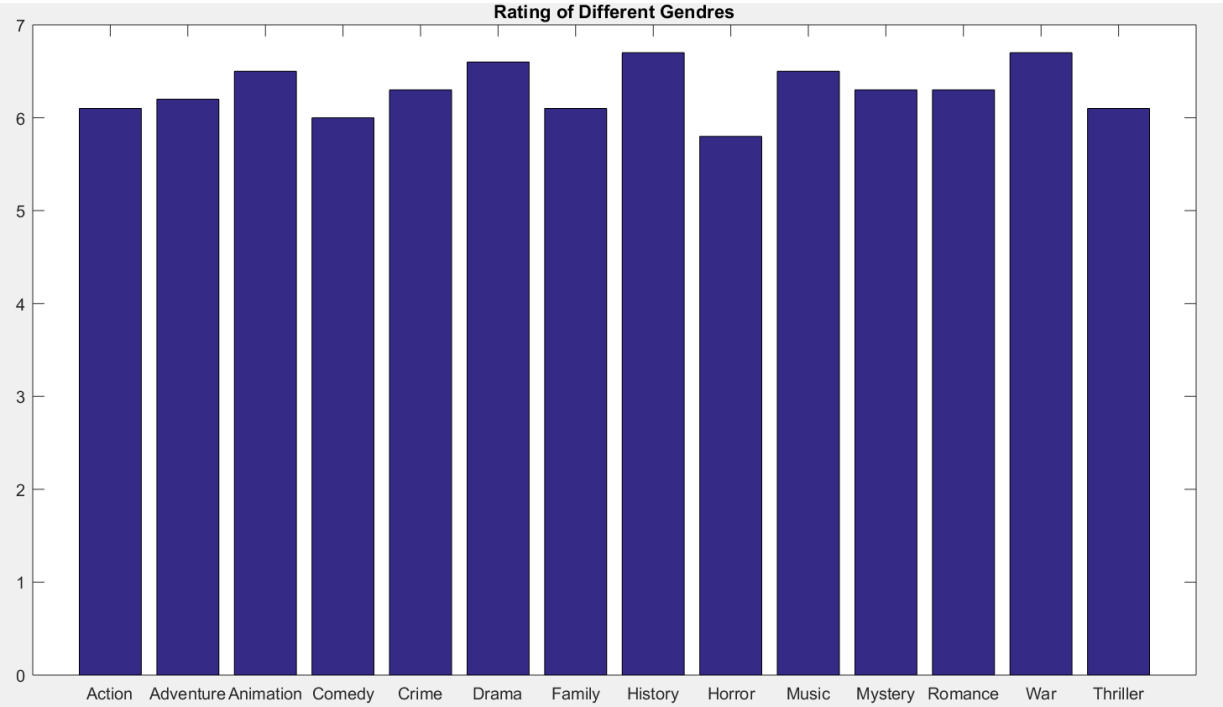

Figure 5 Ratings of different genres

As shown above, the genres of movie affect the rating of movie. War movies and history movies are generally rated higher than the others, other the other hand, the honor movies are not as favored as other genres. To see more specific correlation between ratings and production, the same data set was used to generate the diagrams of probability distribution function(pdf). The plot of pdf can help to determine how significantly one genre is better than the average or another genre.

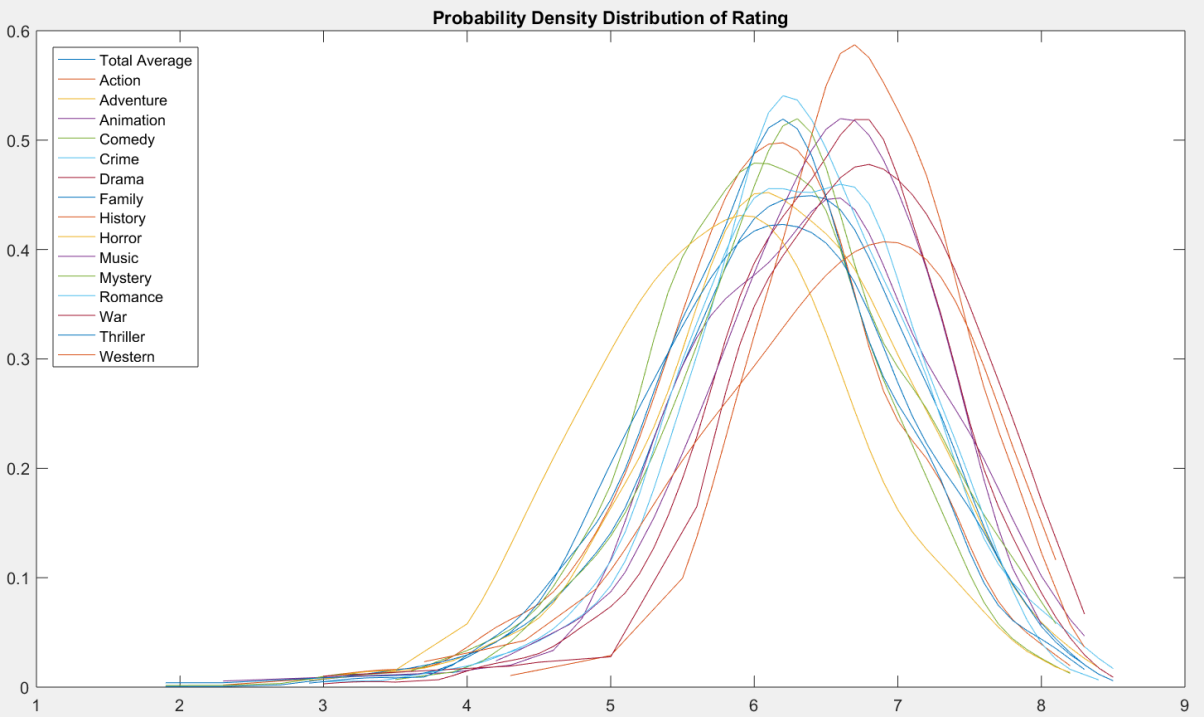

Figure 6 Probability distribution function of rating vs genres

As shown in the graph, the pdf plot does not just show the average of each genre but also the medium and the distribution. The graph says that the history movies has the better rating than the average and the distribution is more concentrated, which means most of history movies are good movies in terms of ratings. Some other genres also stand out like war movies and drama movies. They both have better average rating than the total average, but the mediums do not locate close to the average (low than the average), which means they might still a large part of movies score lower than average of the genre. Based on the figure, it is well-grounded to say that some movies of certain genres like history and drama have better ratings than the others such as horror over all.

The same process can apply for data set of production countries, the generated figures are as following. Similar results can be drawn from the graph, the ratings of movies vary amount different countries. Generally, movies produced in Italy are more likely to have relative high rating than some other countries.

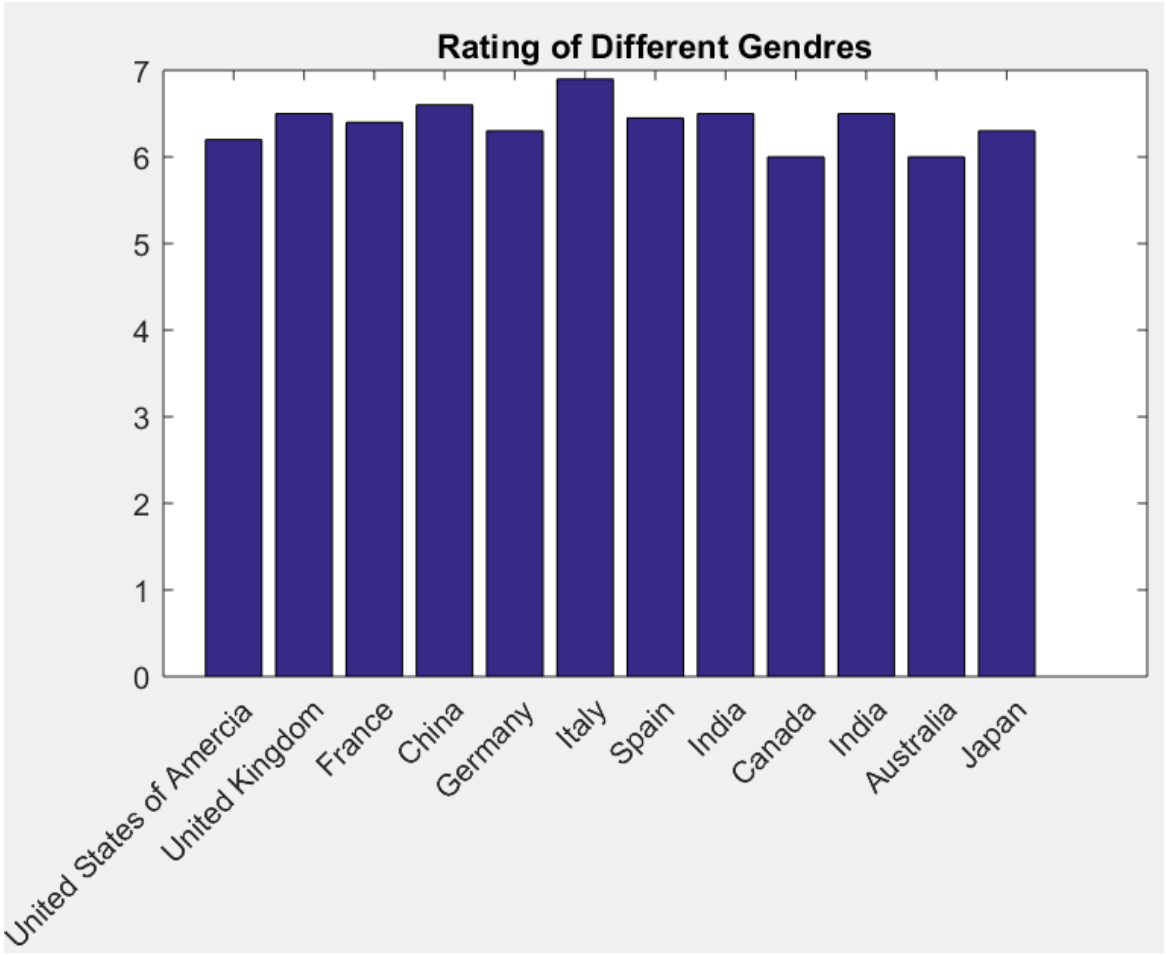

Figure 7 Ratings of different production countries

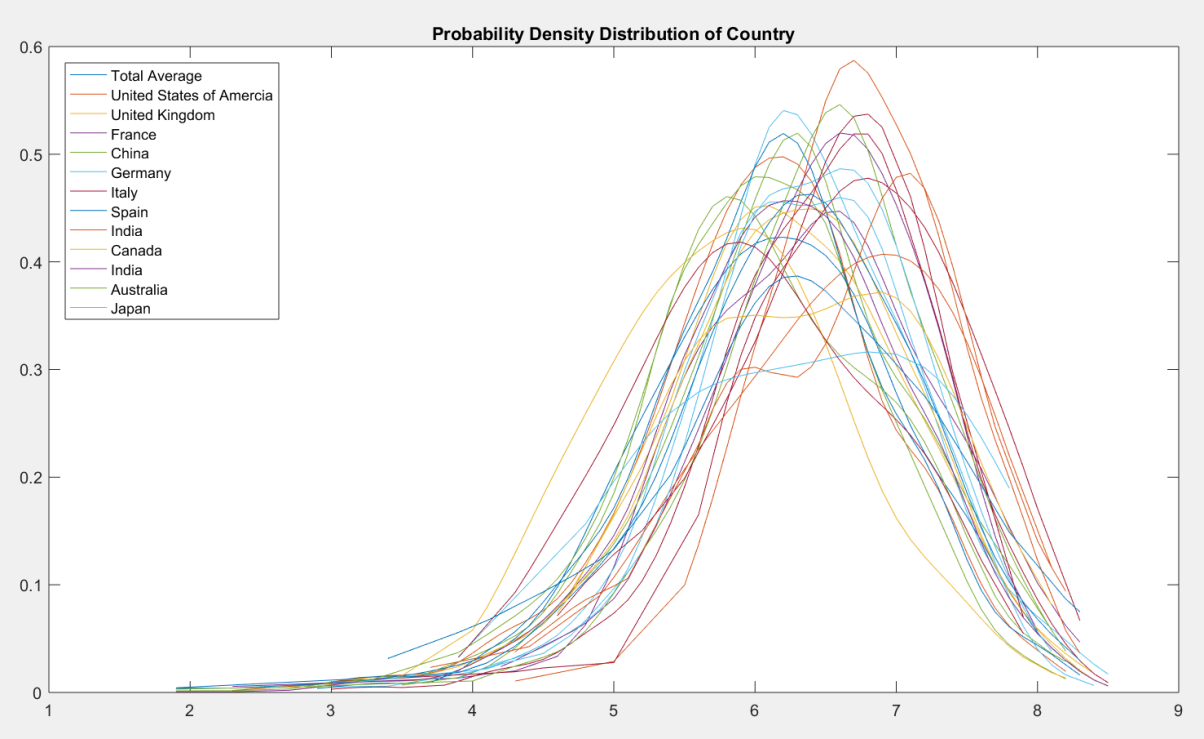

Figure 8 Probability Density Distribution of Country

The rating was also predicted to change with the changes of time, to show that, a data set of the average ratings of every year was created and plot as a bar chart, see the figure below.

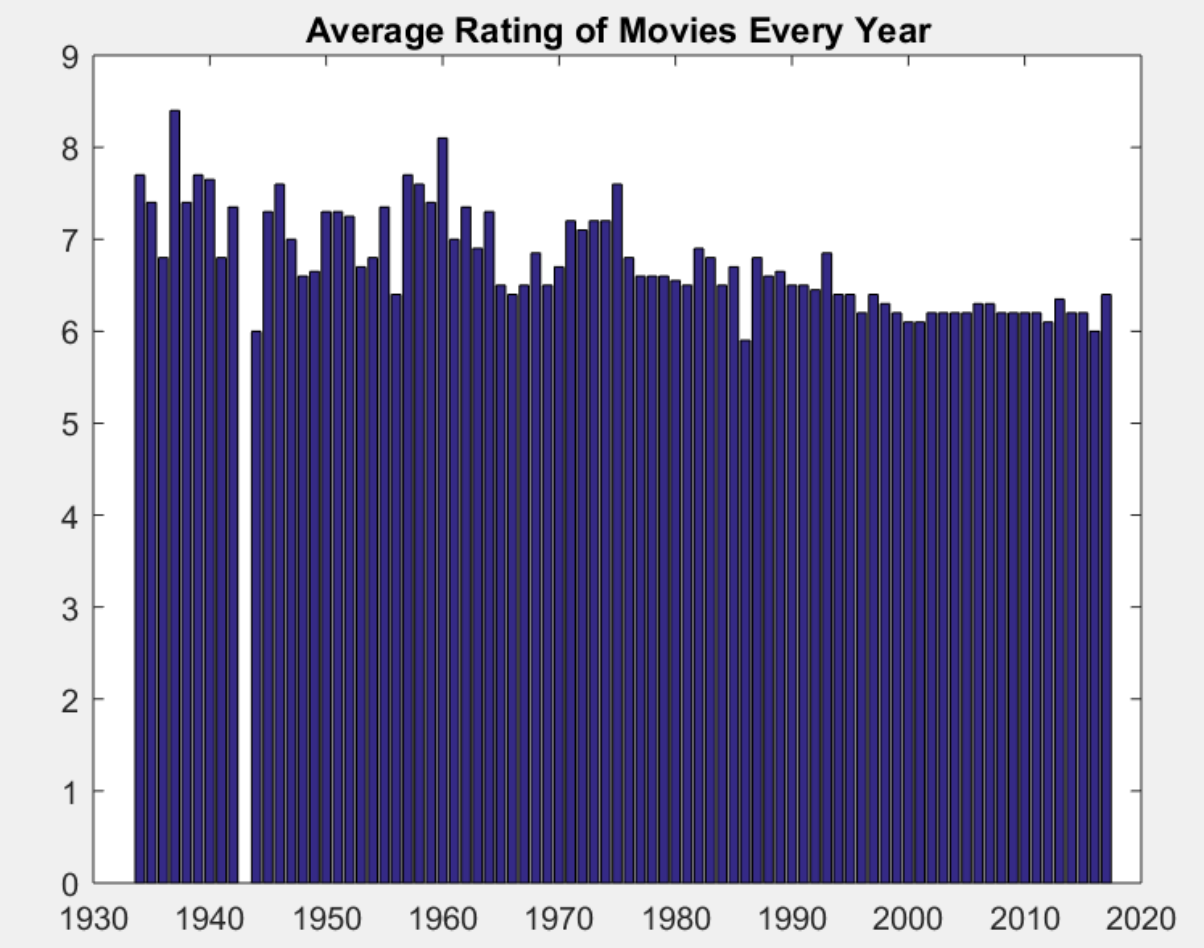

Figure 9 Average rating of movie every year

The bar chart shows the average ratings of movie every year are different for different year. It appears to be higher in the early years and after fluctuation in the middle, the it dropped to a relative level and remain stable. A data points plot would be used to show the trend over time, the data points are the still the average ratings of movie every year. And a linear fit was applied to determine the correlation. The plot is as following.

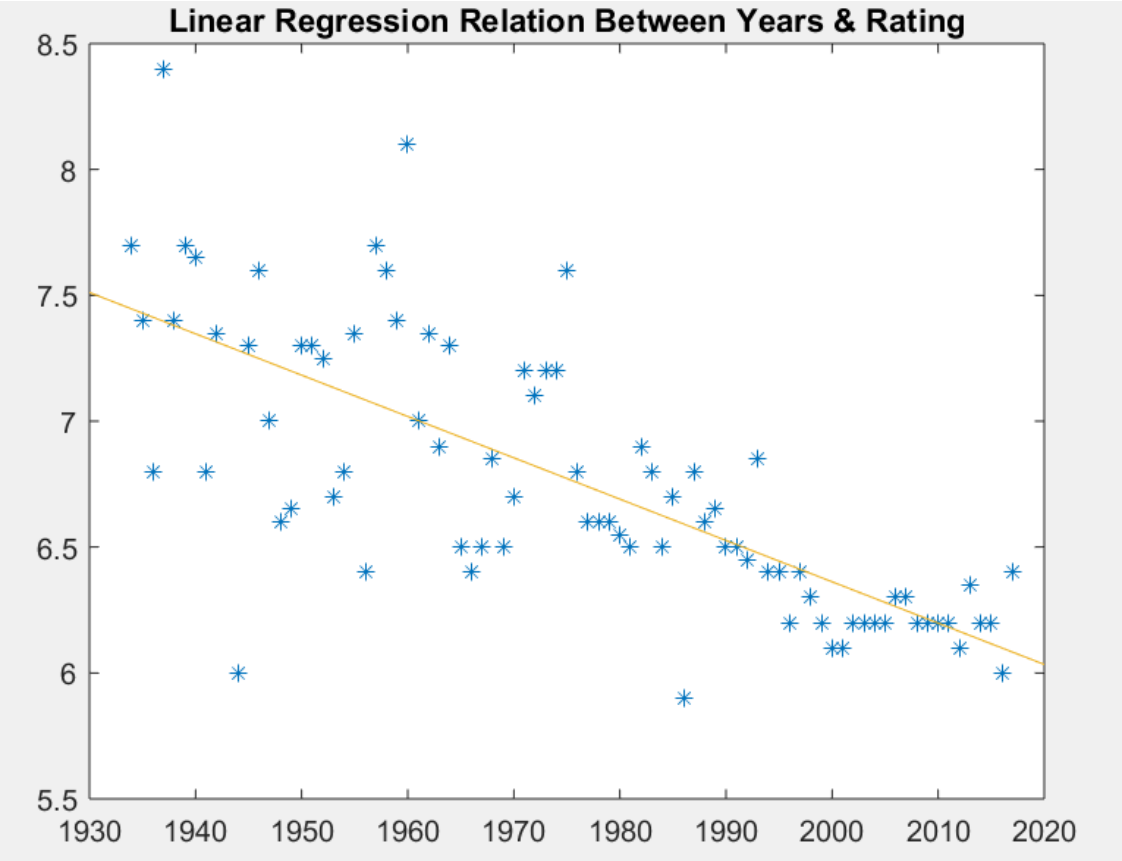

Figure 10 Average rating changes over time

The graph above indicates that average rating of movies decreases as the time increasing. It should be noticed that with the increasing of number of movie production, the quality of movie has stepped backward.

■ Correlation with revenues

The revenues are the most important motivation of film industry development, revenues might have related other factors. The revenues were normally considered to have a strong connection with budgets. The section was to investigate the impact of other factors, such as rating, on the movie of revenues. A data points plot of budgets against revenues would firstly be plotted along with the linear expression. The figure is as shown below,

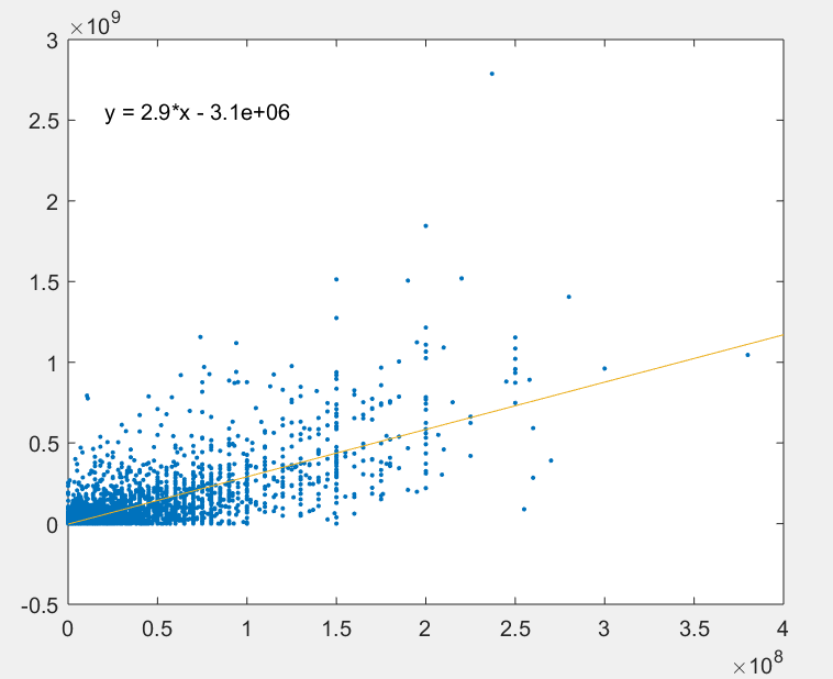

Figure 11 Relation between budgets and revenues

As expectation, the revenues have positive correlation with budgets, the larger budgets, the larger revenues. To compare the correlation between ratings and revenues with budget, similar plot was generated with data set of rating. The figure is as below, the linear fit shows that the rating also has positive correlation with revenues. But the correlation between budgets and revenues is strong due to the larger coefficient of linear expression.

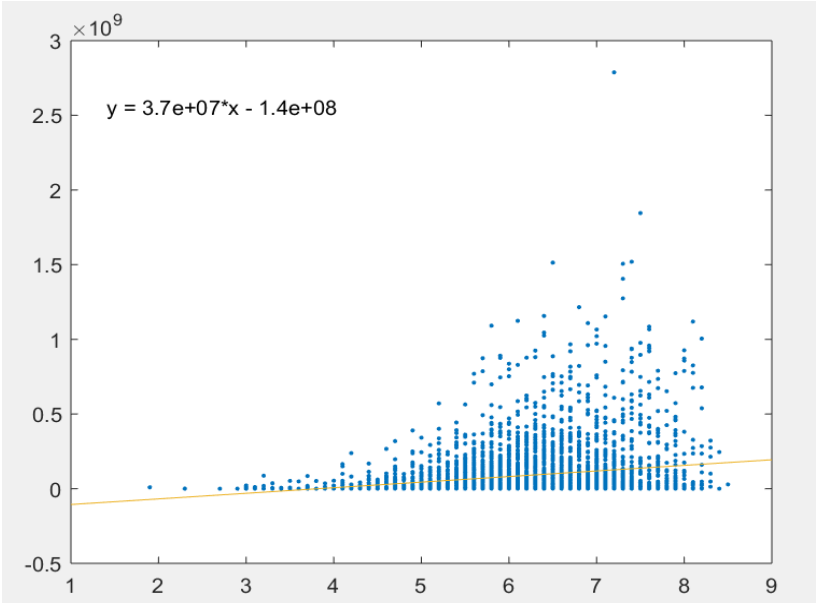

Figure 12 Relation between rating and revenues

Since both factors have correlation with revenues, the impact of these two factors together on the revenues should be investigated. From the previous steps, the multivariate data can be obtained by combining the bivariate data. A 3D plot of rating, budget and revenue can be generated from the multivariate data, as following. To see a clear relation between these three variable, a surface fit generated by using 'Curve fitting tool' was used. The surface fit is a polynomial fit with both degree of x and y as 2.

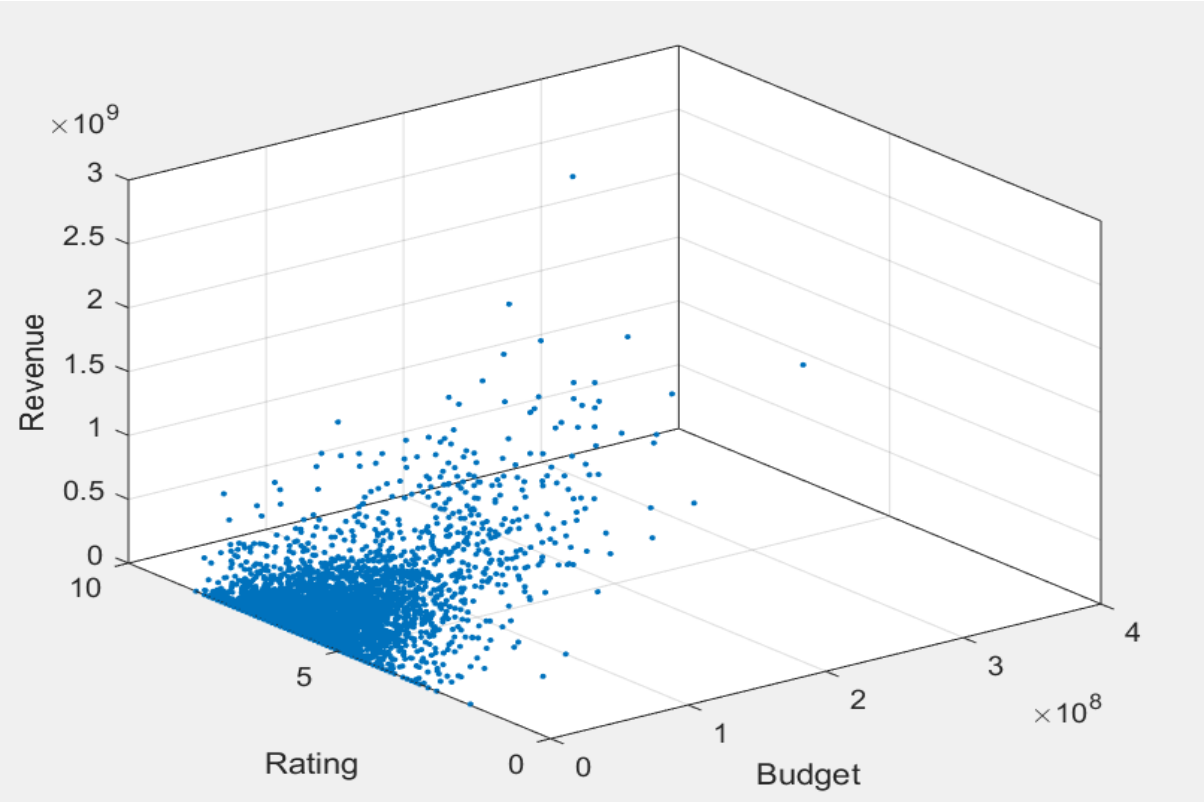

Figure 13 Revenues vs Rating &Budget

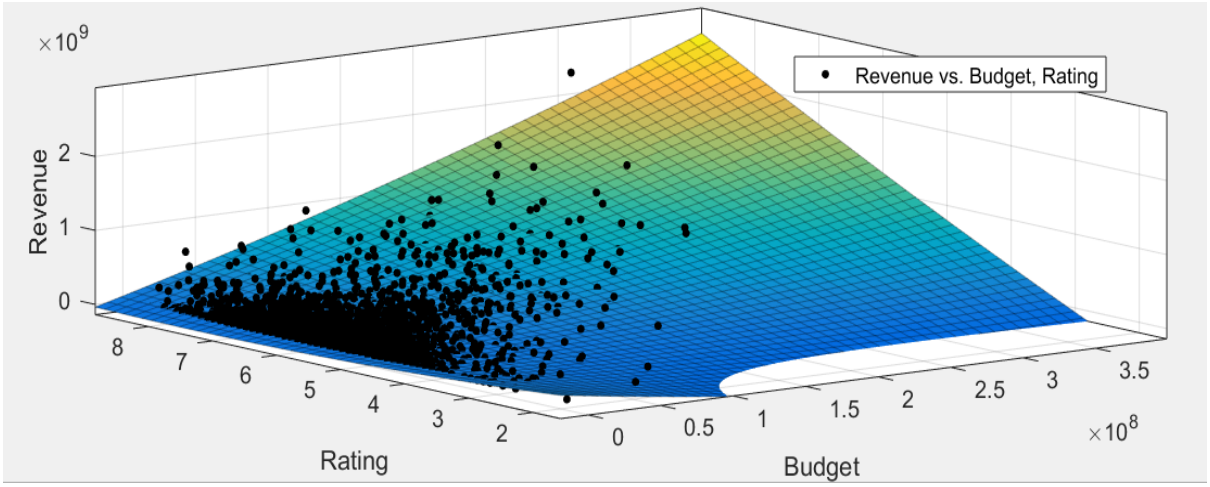

Figure 13 Revenues vs Rating &Budget with fit

From the figure above, it is clear that revenues increase when ratings or budgets increase. Revenues might reach its potential maximum with the maximum rating and maximum budget. The model might be varied for different coefficients used for fit, but the general relationship between factors should be similar.

Conclusion

■ Findings

The project has succeed in some ways. It has found a certain pattern of development of film industry in the past decades. US has appear to be the country with most powerful film industry that produce the most number of movies every year. The drama movie is the most common movie genre among all genres, it is also take up the majority of high-quality movies. The number of movie produced and the total revenues have increased significantly since 1935, but the momentum of increasement has been slowed in recent years. It should be noticed that while the size of movie market expanding, the average quality of movie has stepped backward.

By the data visualization, it is also shown that the production country and genre of a movie effect the rating. The movie from Italy tends to get better rating, and the history movie and the drama movie tend to have better rating than movies of other genres. The correlation between revenues and ratings and budgets is also investigated. The results highlight that both budgets and rating have positive correlation with revenues, the budget appears to have stronger correlation.

■ Future investigation

Although the part of project succeeded, there are still some area can be improved, some mistake made in the process can be avoided. In the data preparation, a number of data were discarded because of missing or incorrect information, those discarded data might effect the final result. For an more accurate data visualization, the missing information should be completed and incorrect data should be replaced with correct information. Some aspects that were not cover in the project are worth investigating in future study. For example, the change of shares of

different genres of movies and the change of budget and revenues of different genres overtime. This might be important because it was the response of the film industry to the film market. By investigating that, the prediction of future could be made.