Project 1 - DATS 6101: Movie Industry

Randomly-Generated: Tran Hieu Le, Fahim Ishrak, Zhilin Wang, Totyana Hill

10/23/2019

### Chapter 1: Introduction

The movie industry is “the first form of industrialized mass entertainment” (Bakker, 2003, p.579). In recent years, it has become an indispensible source of entertainment. The number of people to the cinema is increasing annually. Famous movies such as Avenger, John Wick and Harry Potter are the topics in daily conversations.

Besides being a source of entertainment, the film industry is an important factor to an economy. IBISWorld reports that movie industry generates $103 billion in revenue, globally, in 2019.

According to Kiprop (2018), in 2016, the United States and Canada earned a profit of $11.4 billion from film industry. China - the second biggest economy in the world was benefit from an income of $6.6 billion, and one of the Europe giants - United Kingdom earned 6.5 billion from their movies.

The evidences and data above explained the role of movie industry in our today’s world. Therefore, our group decided to do research on this topic.

This project will concentrate on different movie genres. Before analysis, we will interpret our dataset and bring an overview of movie industry. The rest contains three main contents. Chapter 3 focuses on analyzing the classification of movies and its impacts. In chapter 4, we study the IMDb score system and observe its relationships with other attributes. Finally, we will conduct a research on movie commerce. Chapter 6 is the summary.

### Chapter 2: Description of Data

#### 2.1 Data Source

The movie dataset for our project is taken from Kaggle (link: <https://www.kaggle.com/danielgrijalvas/movies>).

This data was scraped from The Internet Movie Database (IMDb) website (<https://www.imdb.com/>) using Python scripts.

Limitation of this dataset:

* Because the source’s owner scraped the dataset from a website, there may be a percentage of data that were not successfully retrieved.
* IMDb does not have enough information in countries where the movie industry is less developed. Therefore, we do not actually study the global movie industry. Instead, we are doing research on most popular movies.
* Since the reviews of IMDb comes solely from audiences, the score of movies is neither objective nor well-evaluated. The fanboys of the movie or its actors/ actress will likely grade the movie 10 score, whereas the haters often give a rating of one.

#### 2.2 Data Explanation

The CSV contains records of 6820 movies in about three decades (1986-2016). There are lots of information included in the CSV file.

The following is the full list of movie attributes in the columns with their explanation:

* budget: the amount of money was spent to produce a movie
* company: the movie’s producer
* country: the country where the movie is produced
* director: the director of the movie
* genre: the main genre of the movie
* gross: the revenue of the movie
* name: the name of the movie
* rating: a film’s suitability for certain audiences based on its content
* released: the release date (YYYY-MM-DD)
* runtime: the duration of the movie
* score: IMDB user rating
* votes: the number of user votes
* star: the main actor/actress
* writer: the writer of the movie
* year: the year of release

Motion Picture Association of America (MPAA) film rating system:

* G – General Audiences: All ages admitted.
* PG – Parental Guidance Suggested: Some material may not be suitable for children. Parents urged to give “parental guidance.
* PG-13 – Parents Strongly Cautioned: Some material may be inappropriate for children under 13 and pre-teenagers.
* R – Restricted: Under 17 requires accompanying parent or adult guardian. Contains some adult material.
* NC-17 – Adults Only: No One 17 and Under Admitted. Any movies which are not assigned the final rating by MPAA are uncut/extended versions with different contents from their theatrical releases, or commercial products used to advertise for the main television or DVDs series.

Any movies which are not assigned the final rating by MPAA are extended versions or commercial products to advertise for the main TV or DVDs series. Some movies with local labels are published to serve audiences in a certain country or region and are not submitted for a rating.

Since our goal is to research the theatrical exhibition worldwide, we will remove all movies without MPAA rating. Furthermore, any movie attributes that do not support our analysis will be excluded to ensure our dataset is soft. This process will be conducted in the next part.

#### 2.3 Data Cleaning

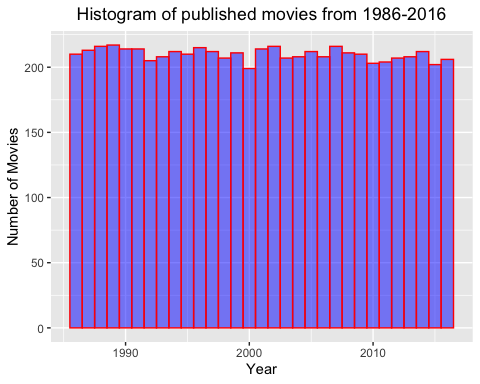
Step 1: Load dataset from csv file and save it as movies.

Step 2: Clean the dataframe. We will save our new dataframe as movies\_new

* At first, we remove duplicates.
* Secondly, we remove columns that will not be used in our analysis.
* As stated in the data explanation, movies without MPAA rating are excluded.
* There are some unrecorded budget data. Our treatment is to change them to NAs.
* Finally, we create two new columns to support our analysis on movie commerce: Profit and Return on Investment.

Step 3: Show our new dataframe.

#### 2.3 Data Visualization

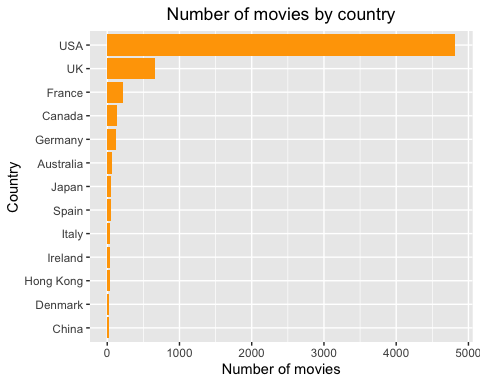


* Our dataset has over 200 movies per year and it is pretty good for our analysis. We do not have to treat any years with significant number of movies.

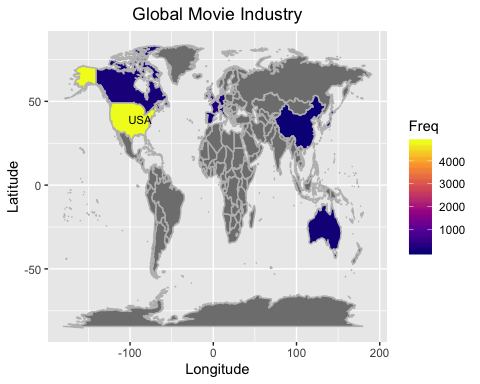
#### 2.4 Movie industry

In this part, we will have an overall visualization about the global movie industry and figure out which country is the largest movie industry.

* Let’s see the which countries have the highest number of published movies.



* Geographic visualization of the global movie industry:



* The country with the highest number of published movies over the three decades is USA.
* USA accounts for approximately 74 percentage of total number of movies.

It is easy to realize that the biggest movie industries are also the most developed countries.

* USA is the largest movie industry, as expected from the biggest economy in the world.
* In Europe, the biggest movie industries are Western giants such as United Kingdom, France and Germany.
* In Asia, China, Japan and Australia are the most growing movie industries.

Now, it is interesting to investigate if the largest movie industry can reflect the global movie industry. To do this, we will conduct a Chi-squared test Goodness of Fit:

* Null Hypothesis: The USA data distribution matches the global dataset.
* The p-value of the test is 0.0000051, much smaller than significance level.
* We reject the null and conclude that the USA data distribution does not match the whole dataset at 0.05 level.
* Unfortunately, we do not have enough evidence to use USA as a sample to study the global movie industry.

### Chapter 3: Movie categories

In this section, we will analyze the movie categories and figure out which genres are the most popular. We will also examine if movie genre and votes have any relationship and discover the most favorite genre based on the votes per movie.

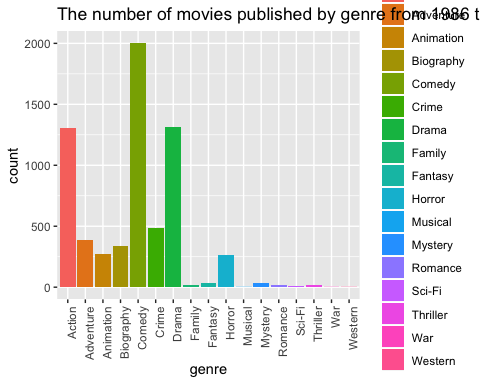
Since the genre of a movie is determined by its content (adult scenes, level of violence, romance elements, criminal activities), it is argued that there is a relationship between movie genre and movie rating, i.e. movie genre is a factor to define the rating. We will examine this critic by a hypothesis test.

#### 3.1 SMART Question

##### Does the movie genre have any influences on the movie rating, and which is the most favorite genre?

#### 3.2 Number of published movies by genres

* We will calculate the number of published movies by genre and show the results in the table format:
* Below is the visualization of the number of movies by genre.



* As we see from the table and the bar charts, Action, Adventure, Animation, Biography, Comedy, Crime, Drama, and Horror are the most popular genres. Action, Comedy and Drama have most products in the three decades.
* In contrast, Family, Fantasy, Musical, Mystery, Romance, Sci-Fi, Thriller, War and Western movies seem not to be preferred by film companies.
* Indeed, the elements of Family, Fantasy, Musical, Mystery, Romance, Sci-Fi, Thriller also exist in other genres and improve their contents. However, when a single of these elements is attributed to the main content of a movie, it often makes audiences feel monotonous and boring.
* War and Western movies seem not to interest international viewers. The reason may be that they focus a lot on historical events and stories which appear frequently in papers, books and documentaries. Their main contents on the history bring little entertainment to audiences.
* In the next parts, we will focus on most popular genres: Action, Adventure, Animation, Biography, Comedy, Crime, Drama, and Horror.

#### 3.3 Genre and rating relationship

* To test the dependence of genre and rating, we will conduct a Chi-squared Test Independence to examine the null hypothesis: Movie genre and movie rating are independent.
* The p-value in the chi-squared test is 0, much smaller than the significance level. We reject the null.
* Conclusion: The movie genre and rating are NOT independent.
* Therefore, there is an evidence that movie genre affects a movie’s rating.

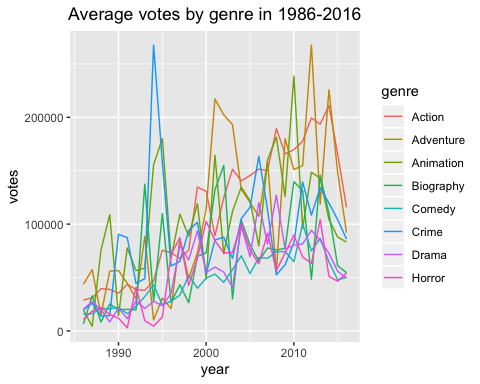
#### 3.4 Vote and Genre relationship

* It is obvious that some genres are of great interest to audiences. This may affect users’ votes for a movie since they will always suggest their favorite genre. We will identify whether the genre of a movie influences users’ voting.
* To do this we will conduct an anova test to see if different genres perform similarly in users’ votes.
* From our anova test, the p-value is much smaller than significance level. We reject the null that the voting averages of different genres are all the same.
* Since the voting averages by genre are not all the same, there is an evidence that movie genre has some relationship with users’ votes. There are some certain genres that greatly interest audiences, hence, they receive more votes than others.
* From the Tukey test, we can divide 8 genres into two groups. Each pair of genres within a group are not significant but two genres from two different groups will have significant means.
* Group 1: Action, Adventure, Animation. Group 2: Biography, Comedy, Horror, Drama. Crime is an exception since Crime-Adventure and Crime-Biography are not significant. We do not put this genre into any groups.
* Prediction: One of two groups contains the most favorite genres, while the other is least favorited. Crime seems to stay in the middle.

#### 3.4 Most favorite genres

* We will confirm our prediction above and figure out the most favorite group.
* We divide the timeline into three decades: 1986-1995, 1996-2005, 2006-2016 to see the votes in different periods.

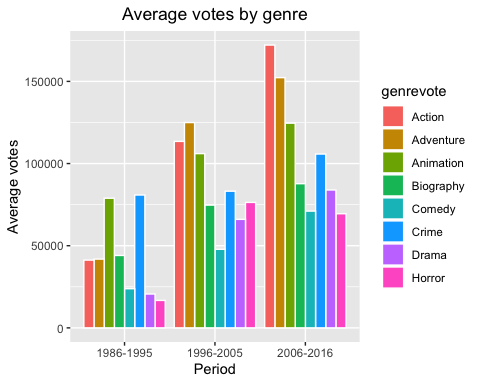
At first, we will visualize the change of average votes.



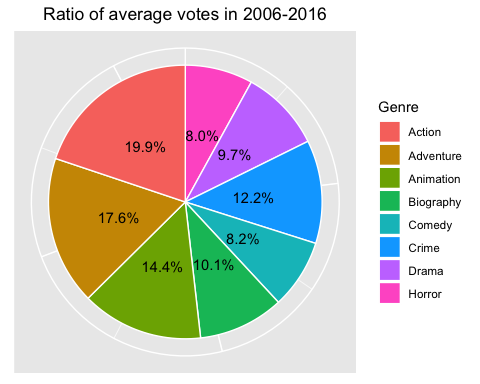
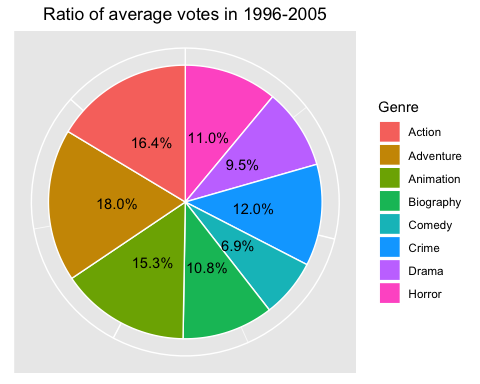
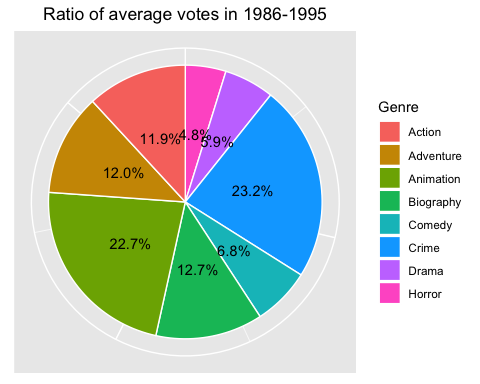
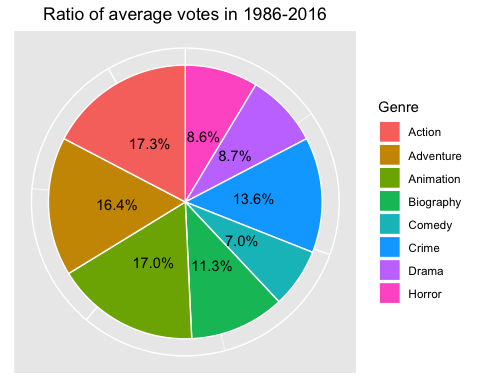
* The graph is not clear but it is easy to realize a significant drop in the average votes of Crime. Comedy, Horror, Drama and Biography have small increases between 1986 and 2016, while Action, Adventure and Animation show a strong growth in general. It seems that in recent years Action, Adventure and Animation have become the interests of audiences.

To figure out the most favorite genre, we will observe the average votes by genre in different decades.

* Below is the barplot showing the average votes by genre in three periods of time:



* And we have pie charts showing ratio of average votes by genre:



* As we see from the plots, Action, Animation and Adventure which belong to group 1 always receive more votes than Biography, Drama, Comedy and Horror which belong to group 2. Crime is always in the middle of the two groups except the first decade. This observation explained our prediction from previous part.
* During the first period, Crime were the most favorite genre and Animation placed in the second. In the next periods, Action and Adventure grew as the most favorite genres and still maintained their posistions until 2016. In contrast, Animation and Crime dropped to third and fourth places after the first 10 years, respectively.
* Starting as the least favorite genre in the top 4, Action became the most favorite type of movie. The first pie chart also shows that Action is the most favorite genre in three decades. The growth of Action movies may be the result of the development of technology. The appearance of modern technologies in recent years has enhanced the visual effects, sounds and camera views … And Action movies seem to take huge advatages from these improvements compare to other genres. Additionally, the groundbreaking ideas and diverse contents in Action movies seem to be of great interest to people.

##### Summary

We will conclude this section by answering our SMART questions at the beginning. There is an evidence that the genre of a movie is a factor to define its rating. Furthermore, from our analysis we found out that Action is generally the most favorite genre.

### Chapter 4: Movie Scores

As we stated in the limitation of our dataset, the fanboys will likely grade the movie 10 score, while the haters often give a rating of one. This will significantly skew the movie score with either huge fanboys or huge haters. If we assume that the number of votes depend on fanboys and haters. We expect to see that movies with a lot of votes will score well while movies with few votes will receive bad grades.

In this section, we research the relationship between user votes and the score to examine our prediction. Furthermore, we also build a hypothesis to test whether movie genre is a factor to the score.

#### 4.1 SMART Question

##### Is there any relationships between the votes and the scores, and are the score means the same across all genres?

#### 4.2 Scores and Votes Relationship

Votes’ and scores’ categories by our own criteria:

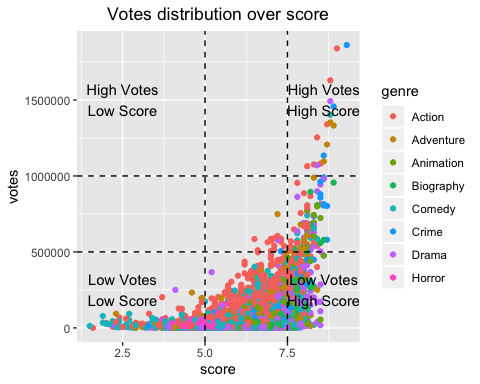
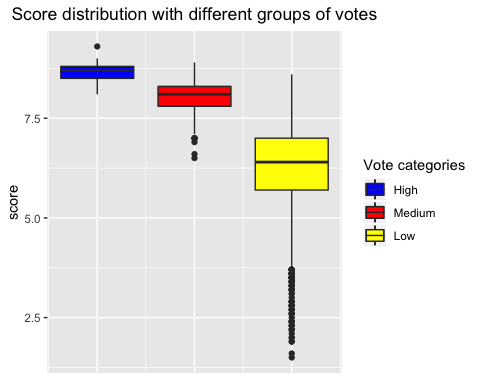
Votes:

* High votes: from 1,000,000
* Medium votes: from 500,000 to below 1,000,000
* Low votes: below 500,000

Scores:

* Good score: from 7.5
* Medium score: from 5.0 to below 7.5
* Bad score: below 5.0

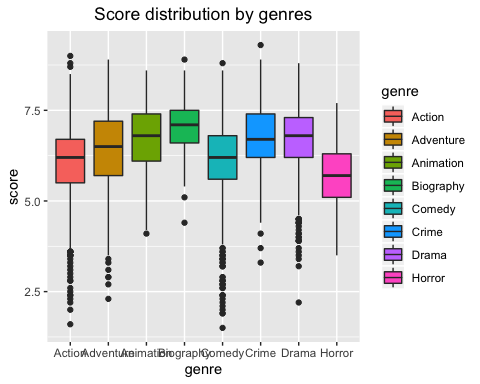
Now, we plot the distribution of votings across the scores using boxplots and mapping plots.



* As shown in boxplots, the movies with high votes always score well, while the movies with bad scores always have very low votes.
* However, high score movies do not certainly have high votes and low votes movies do not mean that they will earn bad scores.
* The mapping plot also gives us the same observation, movies with votes over 1 million always get high scores. And at around 500000 votes and below, the score varies extensively. Movies with scores below 5.0 have low votes.
* Our prediction is not totally correct. The visualization matches our expectation for most favorite movies (with high votes) but in the cases of least voted movies the scores are not certainly bad. The reason is that there are many good movies which belong to least favorite group (Biography, Crime, Horror and Drama). Therefore, these movies have a small number of viewers which makes them earn few votes despite their high grades.

#### 4.3 Score by genres

* We will examine if the score means the same across all genres with Anova test:
* Null Hypothesis: The score means of all genres are the same.
* The p-value of the test is 2.63e-156, much smaller than significance level.
* We reject the null and conclude that the average scores of different genres are NOT all the same.
* It seems that movie genre has some influences on the score. It is understandable since people tend to give good scores for their favorite genres.



* As shown in the boxplots, the following pairs have pretty the same tall boxplots and overlapped interquartile ranges: Action-Comedy, Animation-Crime, Animation-Drama and Crime-Drama. These pairs seem to be NOT significant in their means.
* We will conduct a Tukey test to examine our visualization.
* As we see from our Tukey test, there are no significant differences in means of the following pairs: Action-Comedy, Animation-Crime, Animation-Drama, Crime-Drama, while the other pairs are significant.
* The results match our observation from the boxplots.

##### Summary

In this section, we disovered the relationship between the score and the number of votes and the results did not show any specific correlations. The clearest characteristic is that movies with high votes always get good scores while bad-graded movies do not receive many votes from audiences. Additionally, the mean values of the score are not the same across all genres. It shows that movie genres perform differently in scoring, and the reason may be that people tend to give good scores for their favorite types of movies.

### Chapter 5: Movie Commerce

Besides bringing entertainment to people, another main purpose of producing movies is the business. All movie organizations expect to earn high profit with their products.

In this last section, we will analyze the movie commerce in terms of profit and return on investment (ROI). We will figure out some possible factors to the profit and ROI and visualize their patterns. Our goal is to give suggestions for companies and directors who want to be succesful in the movie industry.

#### 5.1 SMART Question

In the previous sections, we found out that certain kinds of movie are of great interest to audiences. We have a curiousity whether some genres perform better than others in movie business.

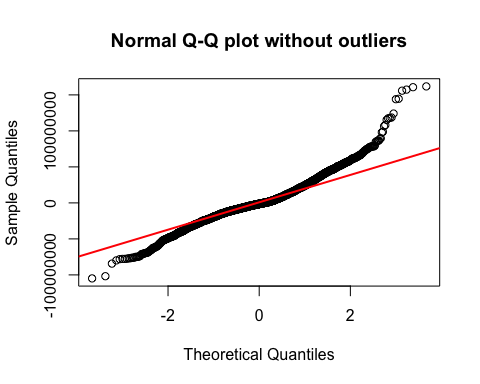
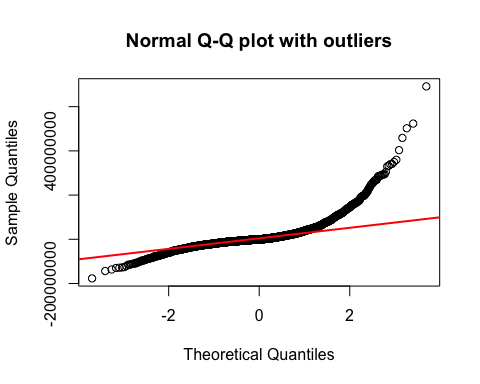
Therefore, we come up with our SMART question:

##### Which genres are good selections for commercial success?

To answer this question, we will analyze the profit and the return on investment (ROI) of different genres in 1986 - 2016. After that, we will conclude which suggestions we should give to the movie companies.

#### 5.2 Profit distribution

Below is an overview of profit distribution.



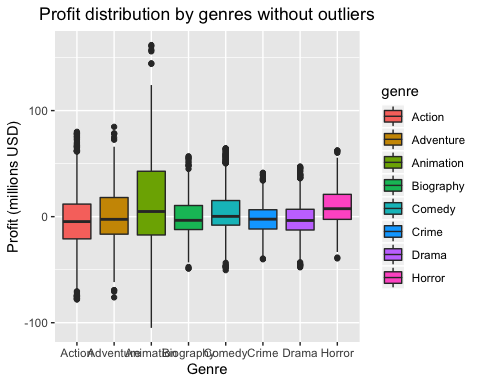
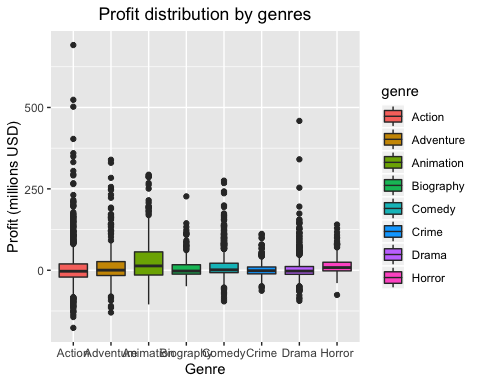
* The Q-Q plot with outliers shows a distribution with over-dispersed data due to outliers. To have a better visualization, we will remove outliers and plot the Q-Q plot again.
* The profit distribution looks different from a normal distribution. It looks like a leptokurtic distribution. The extreme right skew shows that the majority of movies focus on the left. In other words, a random movie will likely earn a low profit. Furthermore, the leptokurtic distribution means that the profit distribution has fatter tails than a normal distribution. The probability of having a net-loss is higher in our movie data.
* These patterns will not statisfy movie organizations. They expect a profit distribution with more concentration on the right or the positive side of profit).

Now, we have a look at the 5-number summary of profit (including the mean) in the table format:

* The median is small than zero, which means that more than 50% of published movies have no profit. This shows that there is a high chance for a movie company to get a net-less with their products.
* To support movie companies, we will conduct a deeper analysis on movie profit.

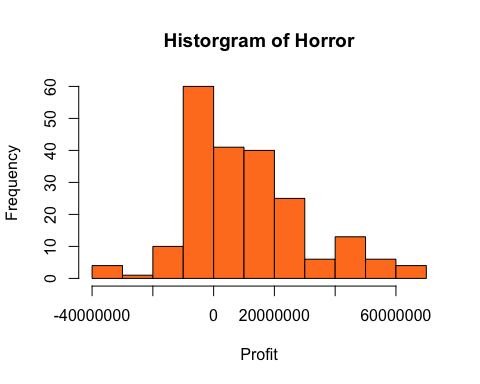
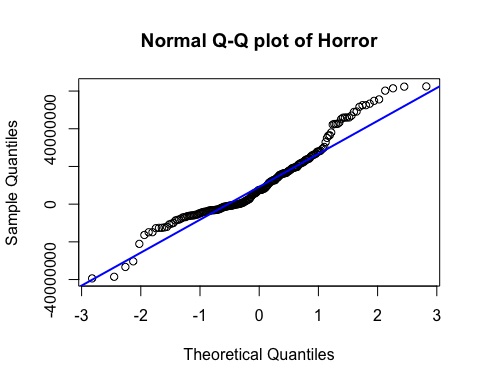
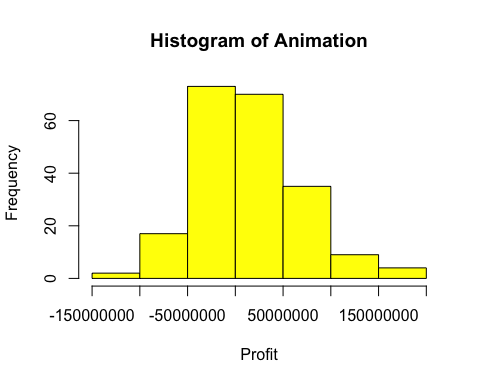
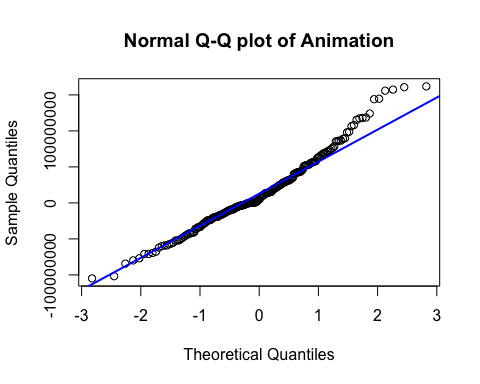
#### 5.3 Profit by genres

This part will identify the genre with the highest ratio of profitable movies and the most profitable genre in absolute.



* From the first figure (with outliers), both the most net-income movies and the most net-loss movies belong to Action genre. Therefore, it is risky to spend on Action despite its potential high profit.
* The Animation and Horror have their median lines at the highest positions and lie over zero, showing that these two genres have good ratios of profitable movies (over 50%). Since the majority of the interquartile range lies above zero, Horror seems to have the highest ratio of profitable movies.

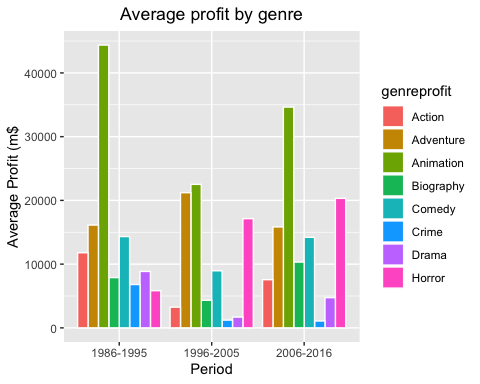
Let’s have a look at the normality of Animation and Horror distributions.



* As we see from the Q-Q plots, profit distributions of Animation and Horror look roughly normally distributed. Both distributions have a slight right skew. Although movie companies expect a left skew distribution but a slight right skew distribution is still acceptable.

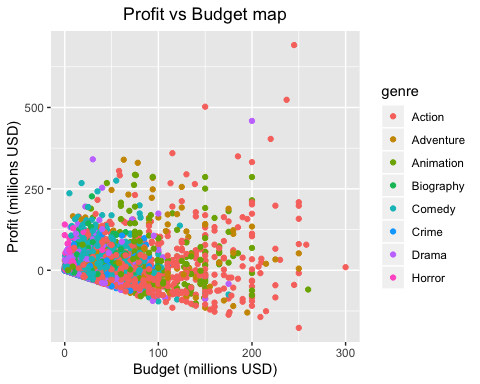
Now, we will calculate the ratio of profitable movies in each genre to examine our visulizations.

* As shown in the table, Horor and Animation have highest ratio of profitable movies (nearly 67% and 60% respectively), matching our observation from the boxplots.
* To contine, we will figure out which genre has the most profit in absolute. Below is the graph of the profit averages (profit per movie) by genre.



* As we see from our barplots, Animation is the most profitable genre. The reason may be that Animation does not require budget to hire actors and invest on visual effects and movie set.
* Through the periods, Horror has grown as one of the most profitable genres. It was ranked second during 2006-2016.
* Action does not earn good profit despite being favorited by people. The reason is that Action movies require a considerable investment on expert directors, famous actors, visual effects and movie set.

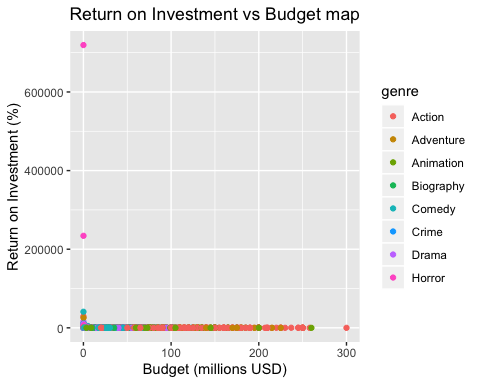
Profit and Budget map:



* Low budget movies tend to make profit, while higher budget movies are at risk of having losses: As budget value increases, the range of gross-loss value also increases.
* Horror movies require low budgets but their incomes are impressive. This genre seems to be the most successful. We will confirm this with analysis on Return on Investment (ROI).

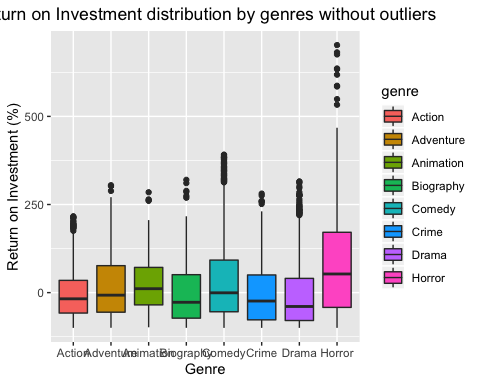
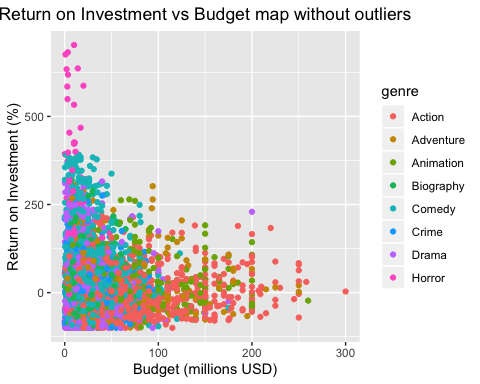
#### 5.4 Return on Investment

* Return on Investment (ROI) = (Profit) / (Budget).
* Return on Investment determines whether a movie is successful.



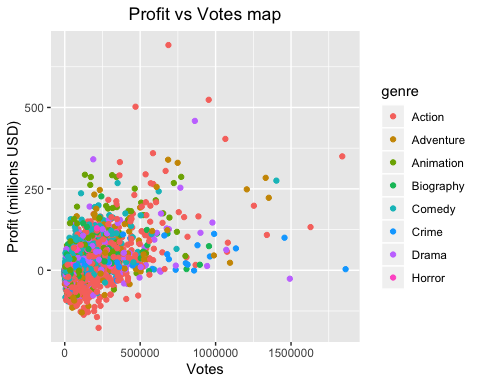
* As we claimed, Horror movies have low budget and high profit, resulting in a very high return on investment (ROI=Profit/Budget).
* We can conclude that Horror is the most successful genre in movie business.

However, it is hard for us to visualize other genre. We need to remove outliers for a better observation. Below is the new graph.

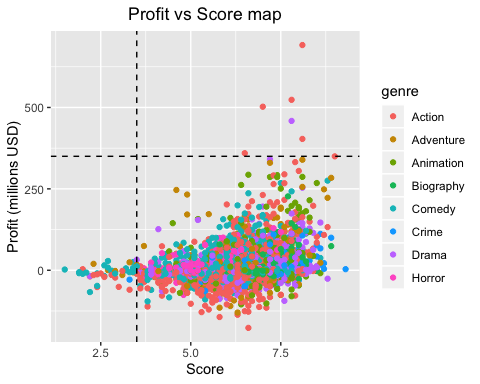


* As expected, Horror movies still possess high return on investment.
* Action movies seem not to do well in Return on Investment. Due to its characteristics, Action requires high expenditures which lead to its low return on investment.
* The Return on Investment tends to reduce as the budget increases. Therefore, low budget movies tend to be more successful than high budget movie in movie commerce.

#### 5.5 Profit vs Votes and Scores



#### 5.6 Profit and Scores



* It seems that there is no relationship between profit and score or votes.

##### Summary

After a research on profit and return on investment, we figured out that Horror is the most successful genre. Horror movies often earn incredibly high return on investment and they also have a decent profit per movie. A Horror product has small chancce to get no profit. Therefore, we would suggest movie managers to invest on Horror if their purpose is the business.

### Chapter 6: Conclusion

##### In summary, there are many insights from our analysis on movie industry:

1. USA is the biggest movie industry since around 74% of published movies worldwide come from this country. However, USA is not an appropriate sample to study the global movie industry because there is no evidence that USA data match the wolrd data.
2. From our testing, we observe a relationship between movie genre and ratings. In other words, genre is a factor to decide a movie rating. It is understandable since different categories are determined by their target audiences, contents and level of violence which are also the basis to restrict the range of viewers.
3. Although there is no clear relationship between movie scores and its votes, we can conclude that a movie with low score will always receive few votes and a movie with lots of votes will score well.
4. We do not have enough evidence to determine if movie genre has any effect on customers’ score and votes.
5. The most profitable movie belongs to Action but the most successful genre is Horror. Horror has the highest ratio of profitable movies and its products often earn high return on investment and maitain a decent profit per movie.

##### Finally, we have some recommendations for movie companies:

1. If companies have little budget or have inexperienced directors actors or the goal is to achieve business success, they should invest on Horror movies which will grant great return on investment. Horror also possess the second highest profit per movie in few recent years.
2. If a company prioritizes the profit, Animation is a safe and stable choice.
3. Although we do not suggest this, but if a company has strong economic potential to invite expert directors and famous actors/actress and to invest on modern cinematic technologies, they can consider Action movies. This genre is risky but it can reward an incredible amount of income if the movie is successful.