Text

Description automatically generated

**ISSS616 – Applied Statistical Analytics with R**

**Group Project: Analysis of Movie Profitability using IMDb dataset**

**Done By:**

**Filbert**

**Koh Jit Woon**

**Pang Hui Qi Michelle**

**Shi JunJie**

**Uthara Venkatachari**

Contents

[**Executive Summary** 3](#_Toc86695701)

[**Data Preparation** 3](#_Toc86695702)

[**Descriptive Statistics** 4](#_Toc86695703)

[**Shiny Application - Descriptive Statistics** 11](#_Toc86695704)

[**Shiny Application - Inferential Statistics** 13](#_Toc86695705)

[**Recommendations** 15](#_Toc86695706)

[**Appendix** 16](#_Toc86695707)

# **Executive Summary**

In this report, we aim to investigate the IMDB dataset to draw meaningful insights on movies for the past two decades, and identify attributes of movies that may have an impact on their profitability. As movie production requires a significant amount of time and resources, it is important for movie production companies to make informed decisions on the types of movies to produce in order to maximize their profits.

# **Data Preparation**

The datasets were taken from Kaggle. There were 20 attributes for movies (title, genre, duration, country, budget, worldwide gross etc.) and aggregated data for ratings of each movie, broken down into each gender age group. After feature exclusion and extraction, the main variables are summarized below.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Title | Title of the Movie |
| Genre | The first genre listed |
| Duration | Duration in minutes |
| Country | Country where movie was produced |
| Continent | Continent where movie was produced |
| Director | Movie director |
| Avg vote | Average rating out of 10 |
| Budget | Budget of the movie |
| Worldwide gross | Worldwide box office proceeds |
| Profit | (Worldwide gross – budget) |
| Metascore | Average score out of 100, given by critics |
| Allgenders\_age\_avg vote | Average rating from each age group |
| Males\_age\_avg vote | Average rating from each male age group |
| Females\_age\_avg vote | Average rating from each female age group |

The following data cleaning steps were performed:

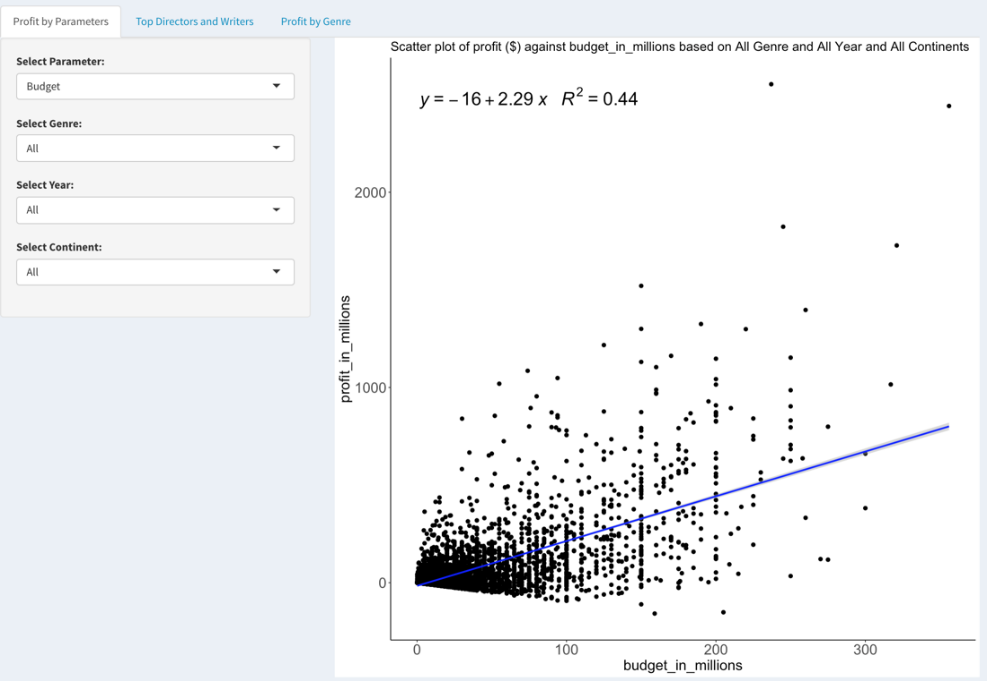
* Filtered only movies produced after year 2000
* Filtered only movies with 100 votes or more
* Filtered out movies with no available information on budget/worldwide gross income
* Took only the first listed genre and ignored the other 2 genres (if any) for our analysis
* Filtered only movies with budget > 100k and worldwide gross income > 1000

There were some limitations from the dataset, due to missing data on budget and worldwide gross income. Movie genres were concatenated as the genre field contained up to 3 genres per movie, separated by commas. We took a simplified approach of taking only the first genre in the list for our analysis.

# **Descriptive Statistics**

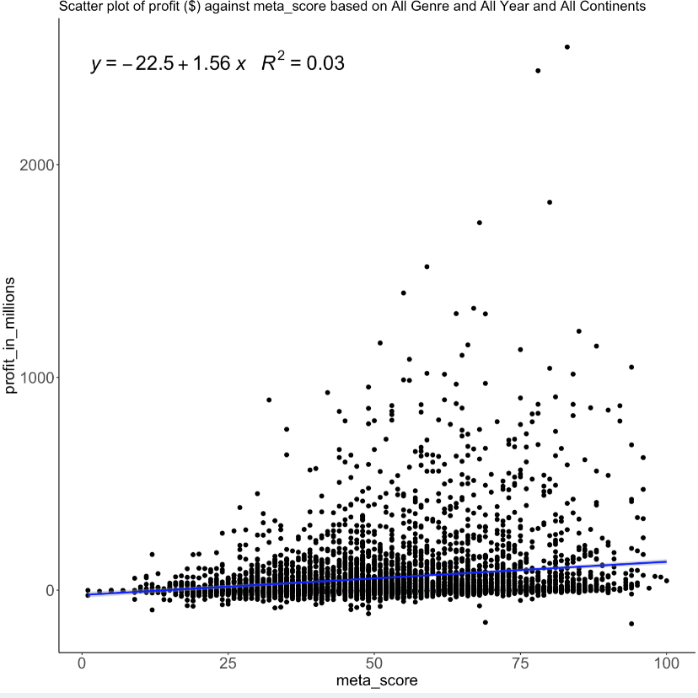
**Correlation Analysis**

Interactive UI was developed to analyze the correlation between the movie parameters such as budget, duration, average vote, meta score, genre and continent with movie profitability. The data was visualized by using scatterplot and linear regression graph. The formula and R squared were also generated to describe the strength of the linear relationship between the two variables.



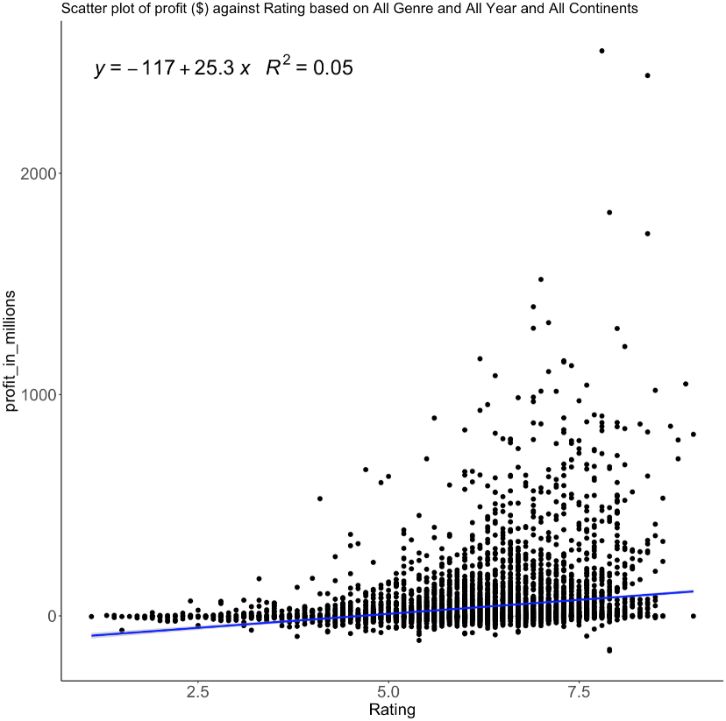
Correlation Analysis depicted on the Shiny Application

The upward sloping line of best fit shows that budget and profit have a strong positive linear relationship. The Pearson correlation indicates an r value of **0.663**. As such, when more budget is invested into making a movie, the profits returned is likely to increase. However, closer observation of the scatterplot shows that the amount of losses incurred (negative profit) also rises as the budget increases, albeit to a lesser extent. This means that there is risk inherent in allocating more budget in a movie as although there is a higher possibility of profit, there is also a small chance that the revenue from producing the movie does not cover the cost expended into it.



Scatter plot of Metascore against profit

There was no significant linear relationship based on the scatterplot comparing metascore and profit. This shows that the correlation between metascore and profit is close to 0, and the relationship between both variables is not strong enough to derive at a conclusion.

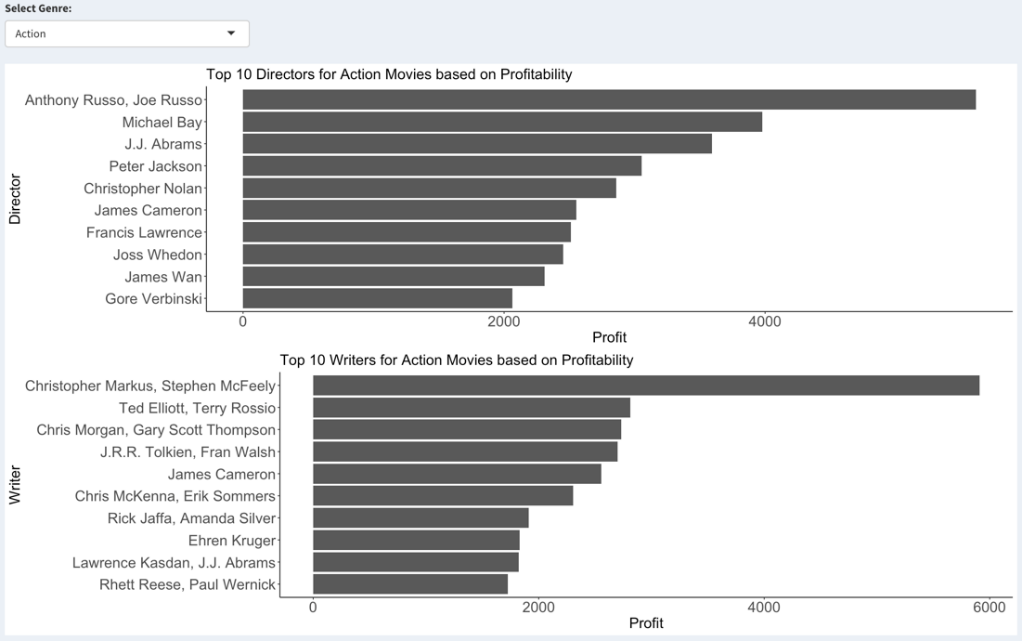


Scatter plot of mean vote vs profit

A very weak positive relationship can be seen in the scatterplot comparing mean vote and profit. The Pearson correlation score at **0.202** shows a slight correlation between mean vote and profit. This means that to a small extent, a higher average vote in the movie ratings might indicate that the movie made a profit, although there is also a high likelihood that a movie with high ratings made a loss from production. Even though a movie may not be profitable, it still has a chance to be received positively by the public.

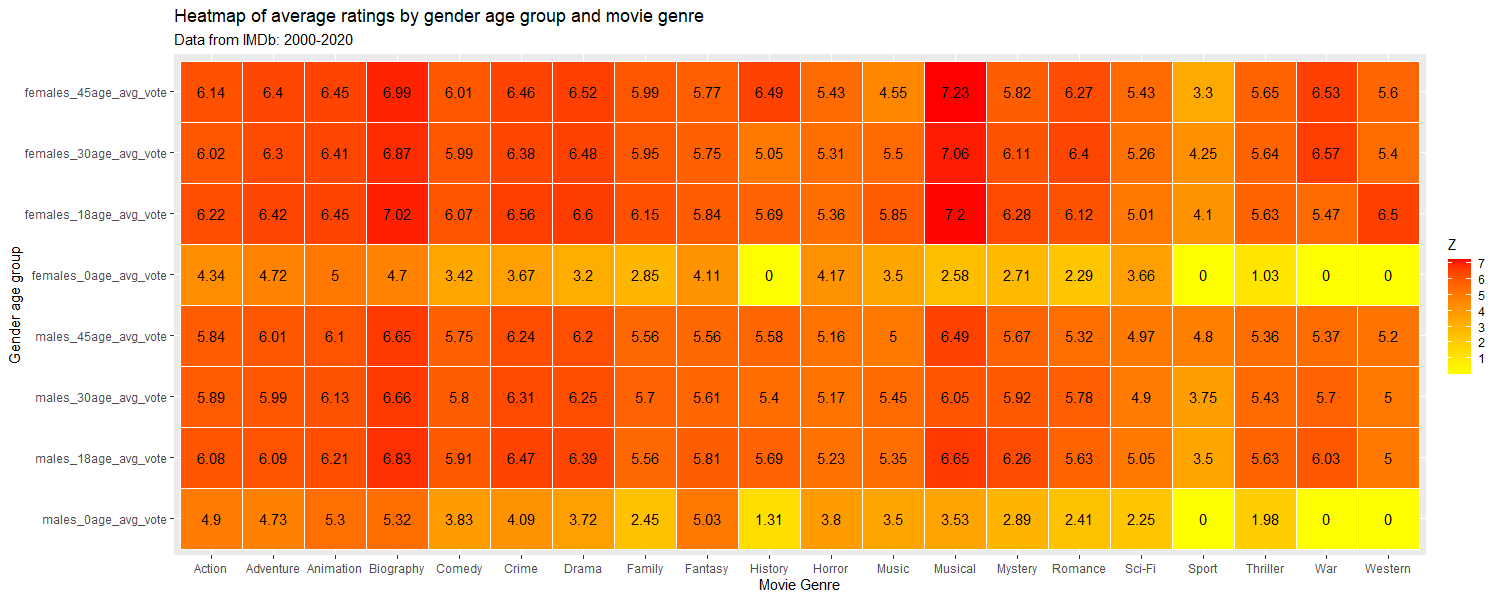
**Top 10 Directors and 10 Writers**

The app could generate the top 10 Directors and Writers that produced the most profitable movies from 2000 to 2020 based on Genre. Knowing the top Directors and Writers would assist the movie production companies in choosing the right Directors and Writers for the movies.



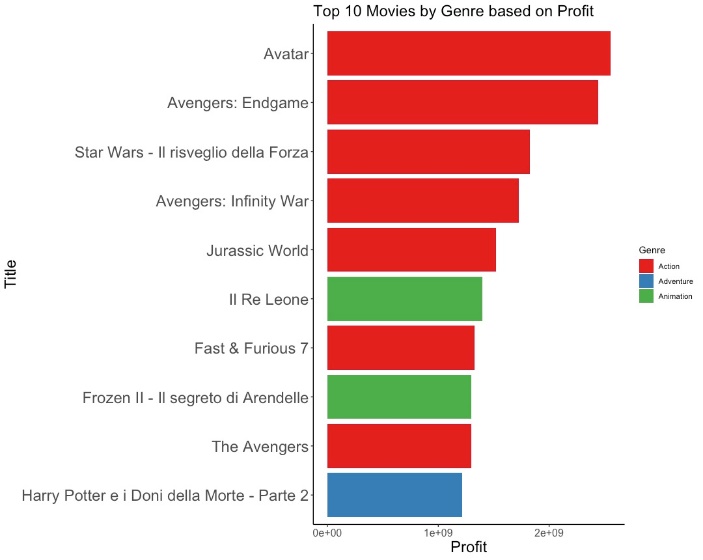
Top 10 directors and writers of the most profitable movie genre

For action genre movies, directors that produced the most profitable movies are Anthony and Joe Russo. They are also the directors who directed the top 10 most profitable movies, namely Avengers: Endgame and Avengers: Infinity War. On the writer side, writers that writes the most profitable films are Christopher Markus and Stephen McFeely. They are also the same writers for Avengers: Endgame and Avengers: Infinity War.



Heatmap of average ratings by gender age group and movie genre

Among the movie genres, Biographies and Musicals ranked the highest in terms of average ratings, while sport has ranked the lowest. There is also a clear divergence between average ratings for males and females for war and romance genres, with both genres ranking higher among females compared to males.



Bar plot of the Top 10 movies by Genre Based on Profit

Based on the top 10 most profitable movie genres, 7 are Action movies, 2 are Animation movies and 1 is Adventure movie. Therefore, action movie appears to be the most lucrative genre.

**Bar plot of the genres of top 100 profitable movies vs Bar plot of the genres of top 100 most highly rated movies**

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Genres of the top 100 most profitable movies Genres of the top 100 rated movies

From the two plots shown above, the genres of Action, Adventure, Animation, Biography, Crime, Drama and Mystery can be seen to be present in the top 100 most profitable and most highly rated movies. This shows that there is already the possibility for these genres to be both highly profitable and highly rated. The action genre has the best statistics as it is in the top 3 for both the number of movies in the top 100 most profitable and most highly rated movies. Aside from the action genre, movies in the genre of adventure and animation are the next best choice for receiving the highest profits and ratings. On the other hand, while Dramas tend to bring about high ratings, the number of dramas in the top 100 profitable movies are in the lower range.

**Bar Plot of all profitable & non-profitable movies by genre**

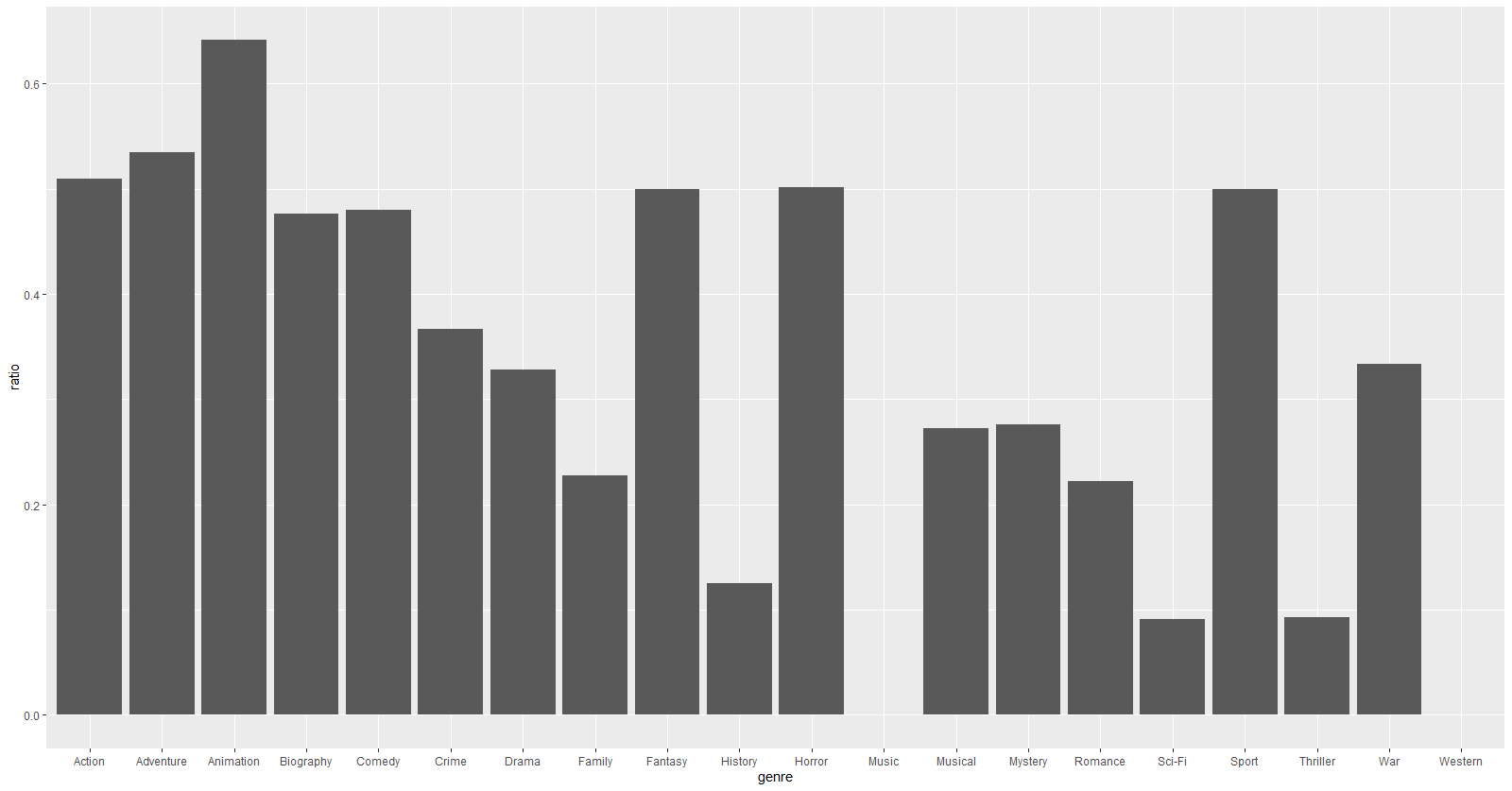
Calendar

Description automatically generated

Bar plot of all profitable vs non-profitable movies for each Genre

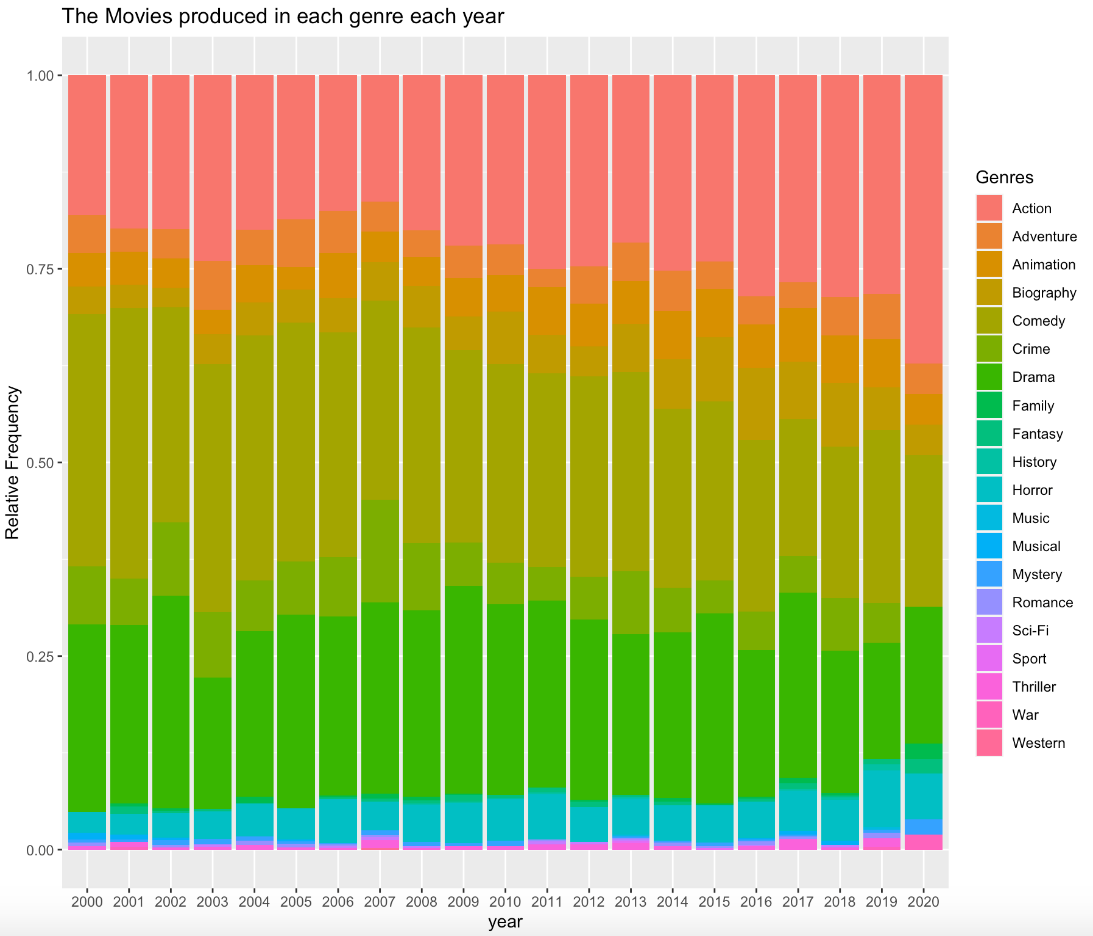
Based on the graph above which shows the number of profitable movies versus the number of nonprofitable movies made in the past years. The number of movies made for the genres of Family, Fantasy, History, Music, Musical, Mystery, Romance, Sci-Fi, Sport, War and Western, are significantly lower compared to the other genres. It is harder to accurately tell whether it is a profitable genre to invest in, as some genres only have two movies in the dataset.

The graph below shows the ratio of the number of profitable movies over the total number of movies made, per genre. The genres that are above the 0.5 mark are Animation and Adventure. These two genres have a higher chance of being profitable movies than non-profitable movies. Genres like Action, Fantasy, Horror, and Sport hit the 0.5 mark exactly. This shows that there is an equal chance for the movie to be profitable or not profitable.



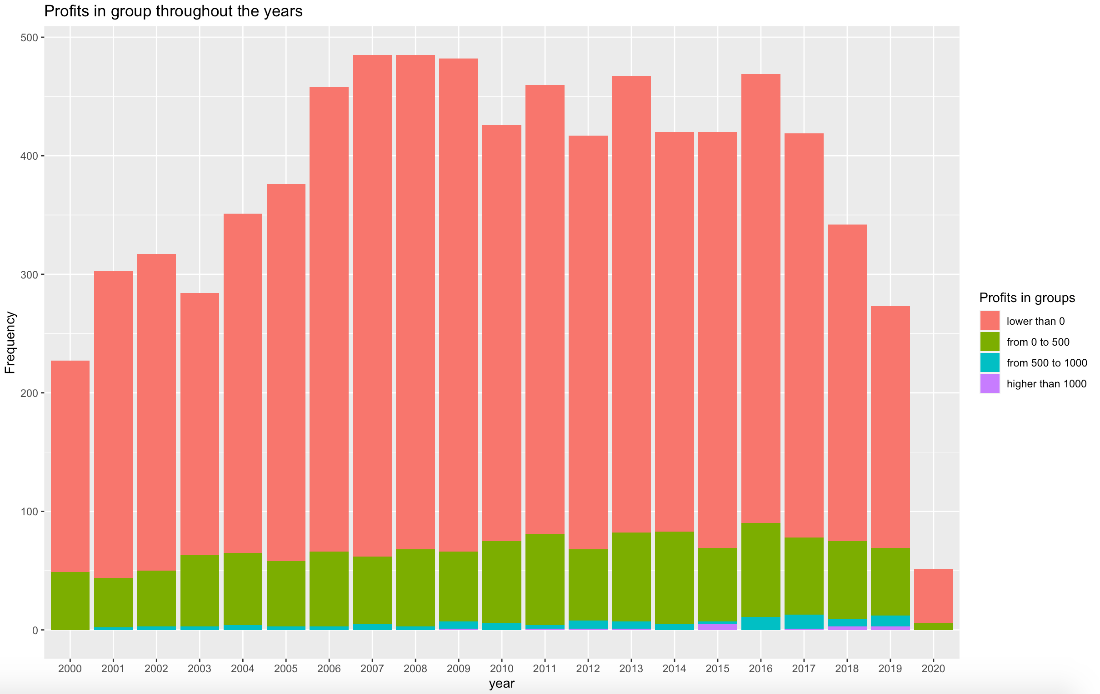
Bar plot of the ratio of the number of profitable movies over the total number of movies made for each genre

**Bar plots of profit and other variables with time**



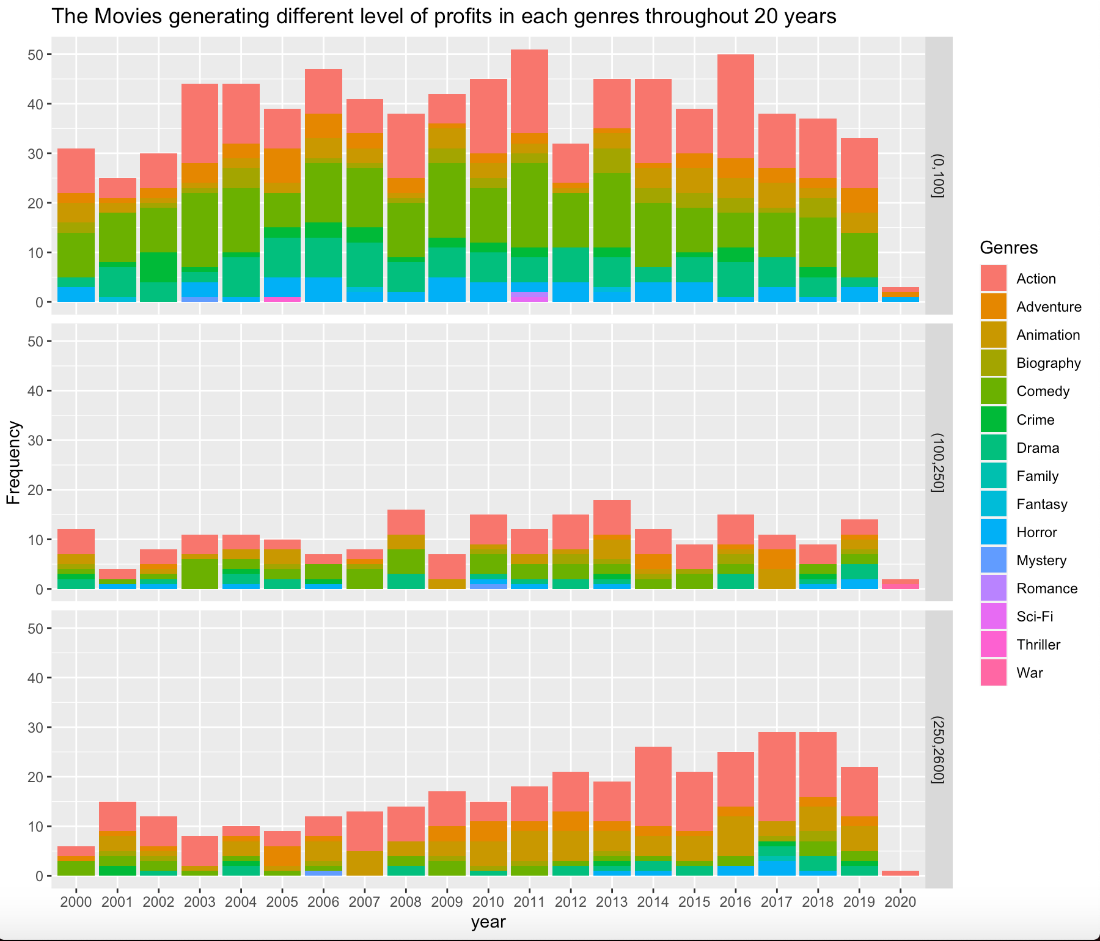
Proportion of movies produced of each genre for each year

Over the past 20 years, the movie genres that was produced the most were Action movies, Comedies and Drama. On the contrary, the number of Romance, Sci-Fi and Western movies had been stably low during this period. The graph also shows that the Action movies had been taking up the market since the percentage of the Action movies produced these years were in an increasing trend.



proportion of profits of each genre for each year

From 2000 to 2020, not all the movies created were profitable. By dividing the range of the profits generated into 4 groups according to the quantiles of the profit variable, as is shown in the chart, a vast number of films created was not profitable, having the profits below 0. The graph also shows that most of the movies which had positive profits only had profits generated below 500 million dollars, in which the green bar represents. And it is exceedingly rare to see movies having net incomes higher than 1,000 million dollars.



Timeline of the proportion of profits for each genre

From the graph shown above, profitable movies, as being divided into 3 groups based on the first and third quantile of the variable, mostly had profits either below 100 million dollars or above 250 million dollars. The number of movies landing in the profit group of 100 to 250 million dollars is very low, indicating the distribution of the profits from movies might have a shape of “M” which means the greatest number of samples in the distribution have either high or low profit. It also can be observed that Action movies take up the most percentage of the movies having profits from 250 to 2600 million dollars, meaning that action films are possible to make the most amount of money.

# **Shiny Application - Descriptive Statistics**

**Trends in profitability**

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Firstly, we only select the movies that made profit, since they are the observations that we care the most. Secondly, we set up three groups (budget, profit, and reviews from critics) for further analysis by using quantiles to divide the profitable movies into three or four subsets. To make the grouping work much clearer, we write three functions for each of the three variables, setting the threshold in distinct functions.

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedInstances of the Bar plots of available based on user inputs in the shiny application

In Panel 1, we would like to see whether the profits of movies depend on factors like genre, budget (group), review from critics (group) and year. By combining the observations having same factors, and comparing the data combined, we have four bar chats which represent the mean of profits generated against each different factor class.

For example, it is shown in the first chart that family movies have the highest mean profits among all movie genres, this is probably because the amount of family movies over the years is not a lot, but all the family movies that had generated money all made a great of net income out of its low-cost product. We also discovered that more reviews from critics, more profit the movie will generate. One probable reason is that comments or critics make the movie to be more controversial and people like to watch the movies out of curiosity, leading to higher box-office records.

**Metascore-Profit Matrix**

Text

Description automatically generated with medium confidence

Graphical user interface, chart, scatter chart

Description automatically generated

Instance of scatter plot result based on year range selected

In Panel 2, we want to see what the relationship between profit and metascore is over the years. By adding lines standing for the mean of profit and metascore, we are able to have a clear understanding about how many movies earning profit had high metascore and high net income during different time period. We can also see the changes in the number of profitable movies over the years. As time goes, the mean of the metascore is also moving toward right, showing that profitable movies are more likely to have higher metascore from 2000 to 2020.

**Word clouds of money-spinning crews**

Graphical user interface, text, application

Description automatically generated

Text, letter

Description automatically generated

Using for loop, we can extract the names of writers splitting by comma and store it in the columns of data frame. What we see want to explore here is that the crews or casts participating in the movies that have the greatest number of profitable movies, giving us ideas that we should invite more casts among them to make much higher profit. The output plot here is word cloud, showing the writers who participated in the movies that earn profit with their names in bigger font.

Text

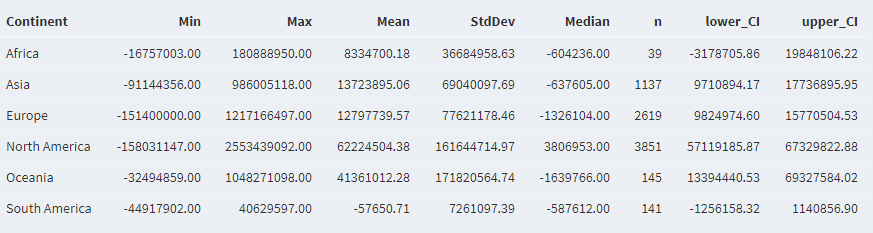
Description automatically generated

Word cloud of writers based on profits earned

# **Shiny Application - Inferential Statistics**

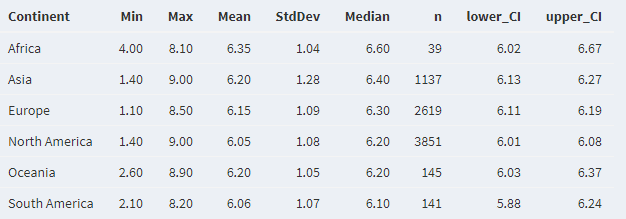
**Confidence Interval**

The confidence interval for profit and rating was calculated in our Shiny App.



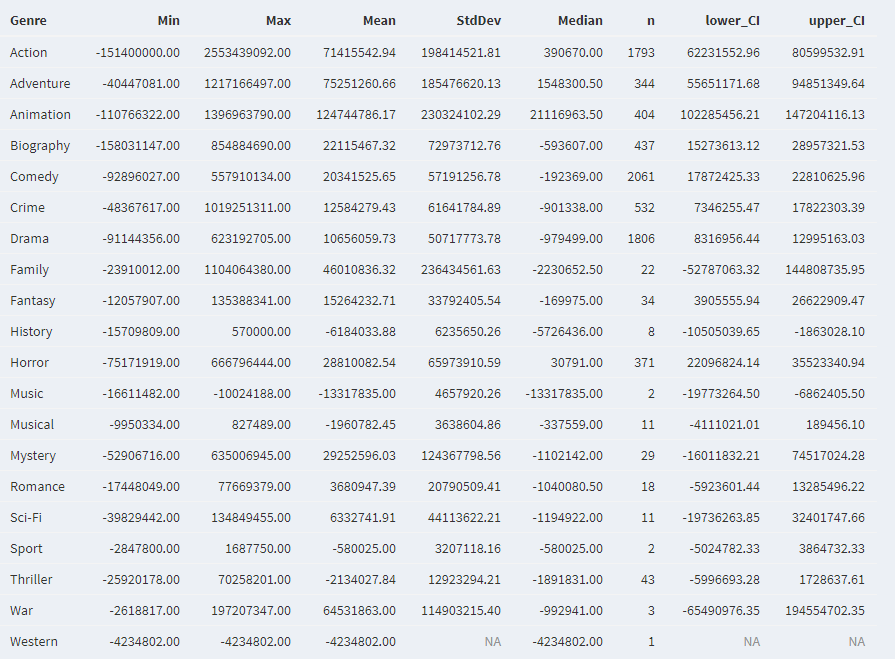
Summary table for Continents based on mean profitability

Among the 6 continents, North America has the highest mean profit, followed by Oceania. North America has a 95% confidence interval for profit falling between US$57.1m and US$67.3m. However, the data is skewed as most movies are produced in North America and Europe, and very few movies are produced in Africa and South America.



Summary Table for Continents based on ratings

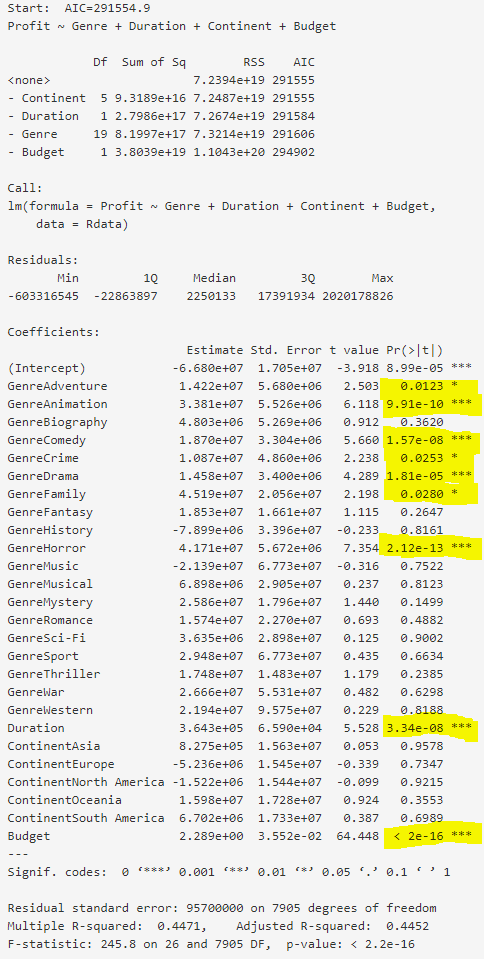
Interestingly, the movie ratings did not vary significantly across the 6 continents. Movies produced in Africa had the highest mean and median ratings.



Summary Table for Genre based on mean profitability

Out of the 20 different movie genres, Action and Animation films have the highest mean profitability, and their confidence intervals (Action: US$62.2m to US$80.6m, Animation: US$102.3m to US$147.2m) indicate that these genres are commercially successful. On the other hand, Family, History and Music movies have performed poorly at the box office, with the latter two having negative profitability for both lower and upper limits of their 95% confidence intervals.

**Multiple Linear Regression**



We tried to apply multiple linear regression on 4 independent variables (Genre, Duration, Continent and Budget). However, the results were inconclusive as two of the independent variables were categorical data and the adjusted R squared value was 0.4452 (only 44.5% of total variation in profit is explained by variation in the independent variables). The statistically significant variables are highlighted in the diagram on the left.

Results of Multiple Linear Regression based on 4 independent variables

# **Recommendations**

The R Shiny application can be further improved as additional independent variables such as directors, writers and actors can be considered to refine the correlation analysis and Profit prediction model. As for the top 10 directors and writers that were shown in the descriptive analysis, it was based on the profitability of the movies that they produced. The frequency of the directors and writers producing profitable movies could also be considered to assess the consistency of the directors and writers' performance.

For the dataset, the recommendations listed below could be used to enhance the insights obtained from the Shiny App:

1. To include the other two listed genres in one movie to refine the genre analysis. There are three genres listed in one movie. Assumption was made to use the first listed genre and ignore the other two listed genres,
2. To include other datasets to refine the unbalanced data. In terms of continent, the data is skewed as most movies are produced in North America and Europe, and very few movies are produced in Africa and South America. As for genre, the data is skewed towards Action, Comedy and Drama movies,
3. The number of movies for each genre is not evenly distributed, making the result unreliable to some extent. For example, the number of family movies is way less than that of action movies in our dataset, the observations for family movies are not large enough to determine whether it is significant to the profit even though it passes the statistical test.

# **Appendix**

R packages

* shiny
* shinydashboard
* lreadr
* dplyr
* matrixStats
* ggplot2
* ggpubr
* ggpmisc
* tidyverse
* MASS
* data.table
* wordcloud2