# REFINEMENT AND REPORTING

## 1. Introduction

### 1.1. Overview

Against the backdrop of a booming loneliness economy, we are interested in whether the movie industry, often considered the antidote to loneliness, facing decline? Our project aim to develop a shiny app, incorporating R code and d3 objects to help clients to conduct in-depth market analysis to better strategize their future development. The dataset is from Kaggle, which can be accessed from this link:

https://www.kaggle.com/datasets/danielgrijalvas/movies/data

### 1.2 Literature Review

Our visualization project draws insights from various literature resources. Devashree et al.'s Netflix analysis inspires our data reduction and visualization methods. Nathan Yau's diverse visualizations guide our presentation approach, while Lev Manovich's "direct visualization" concept informs our logical representation. Nan Cao's work on social media influences how we communicate evolving audience preferences. Nannan Zhang's web-crawling approach informs our data retrieval strategies. Ahmed's IMDb network visualization inspires scalable and dynamic analysis integration. These sources collectively shape our project, aligning our objectives with established and innovative data analysis practices.

## 2. Static Visualizations

Before generating an interactive visualization, we have integrated some static plots explaining the following issues. All code for static visualization is posted to the GitHub repository which can be accessed from this link: https://github.com/feiyunyan2333/stat679_proj

### 2.1 Budget Analysis

For this part, we aimed to help movie companies make budgeting decisions by figuring out the relationship between budget and other movie features at first.

Considering some features in the original dataset are character factors that should have impacts on the budget, we transform them into numeric features to calculate the correlation. We used heatmap to visualize correlation among features.

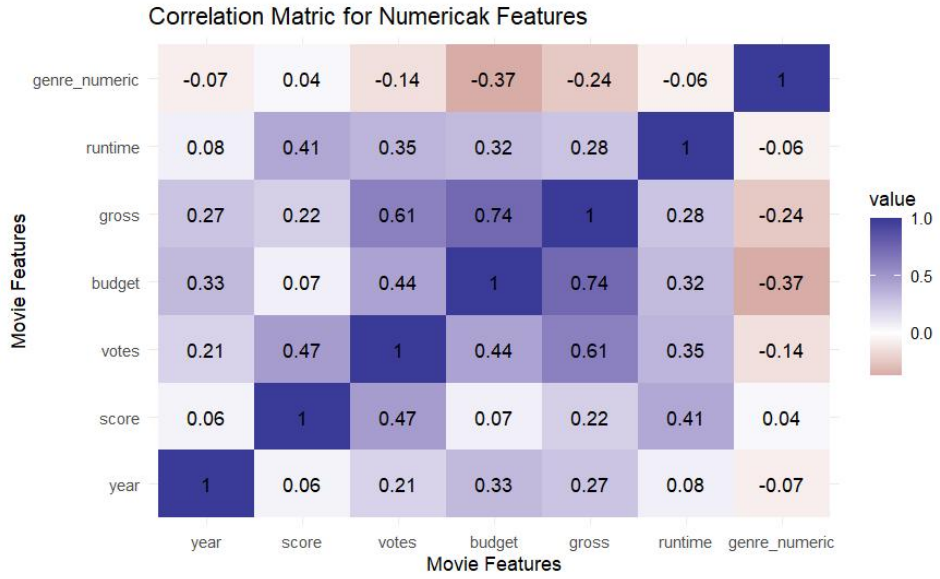


*Figure 1. Correlation in heatmap*

The heatmap shows a strong positive linear relationship between gross and budget. Then we intend to explore more. We used scatter plots with regression lines since they're clear and understandable. Each point on the scatter plot represents a data point, and the blue line is the visualization of linear regression on the two specific movie features.

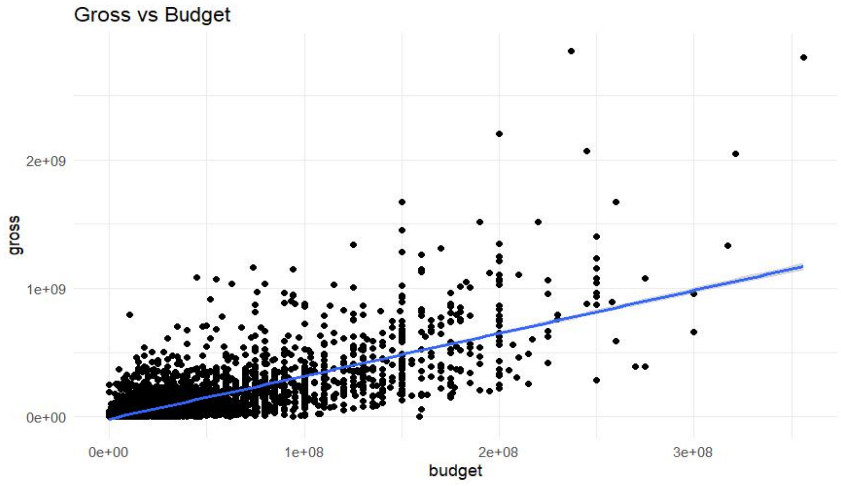

*Figure 2. Scatter plot and linear regression on gross and budget*

The slope of regression lines placed in *Figure 2* consistently below 1, which means most movie companies probably don't have a rational budget. The phenomena that low-budget movies generated high revenues also exist.

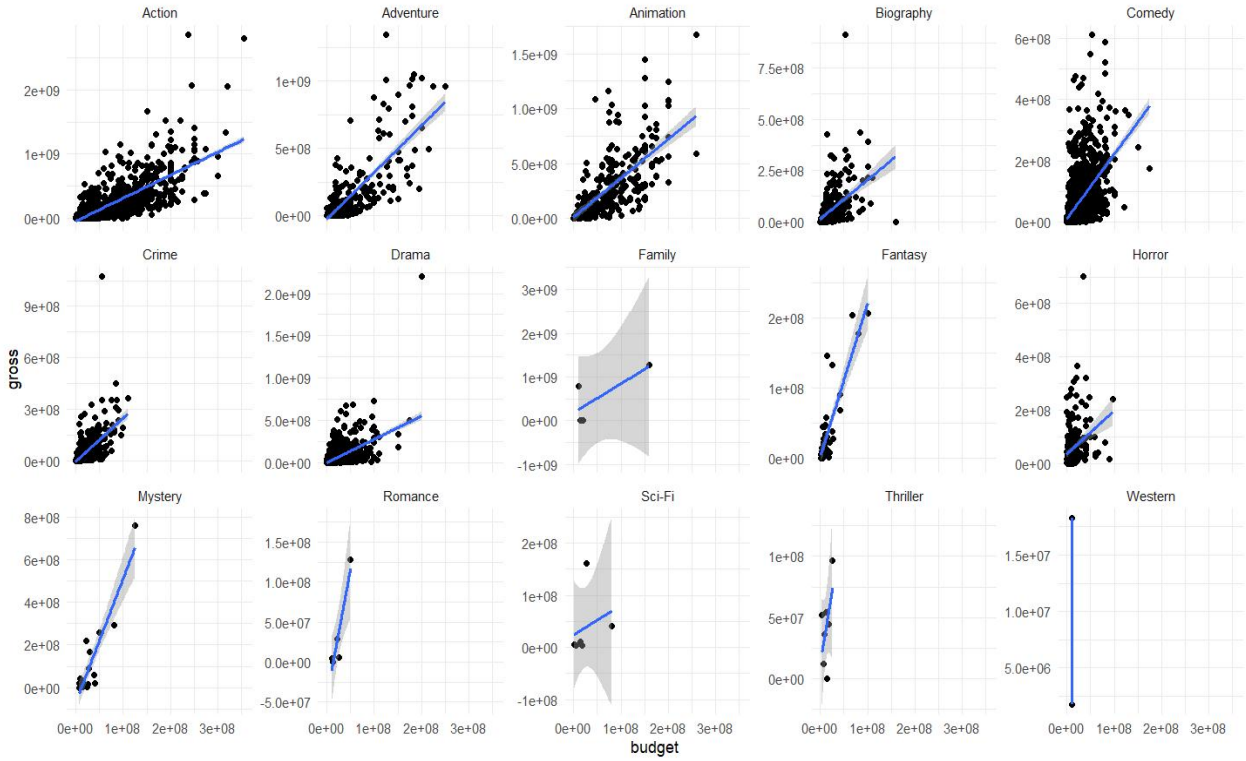


*Figure 3. Gross-Budget analysis on different movie genre*

*Figure 3.* indicates that action, adventure, animation, and comedy movies dominated the market most from 1908 to 2020. However, only comedies exhibited a consistent alignment between budgets and gross with a slope greater than 1. Thus comedy might be a kind of high-return genre.

## 2.2 Movie Releases Analysis

### 2.2.1 Release Date

We designed bar plots to visually convey the time-based distribution of movie releases.

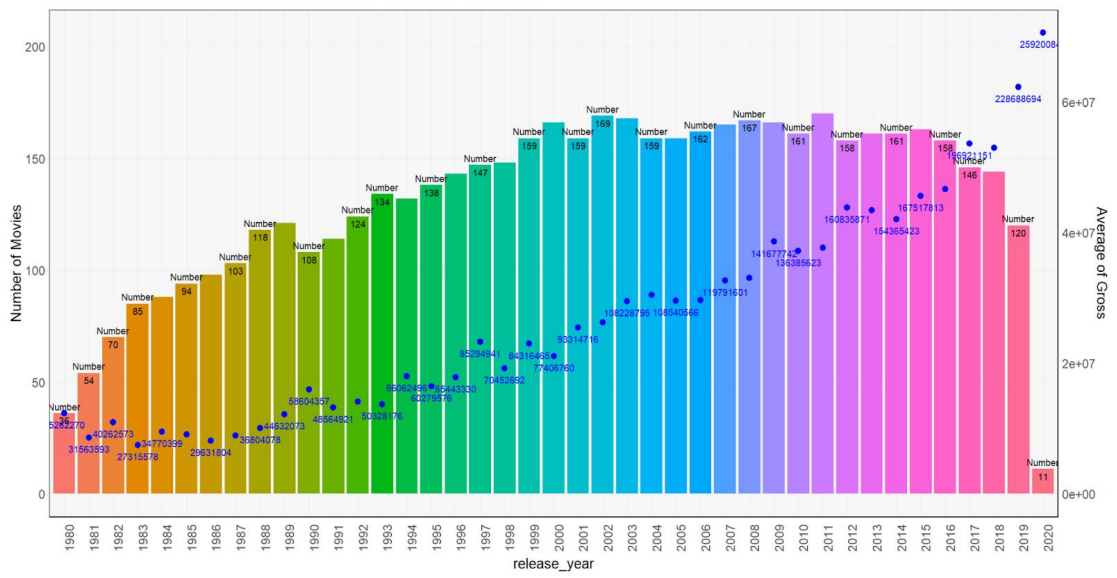


*Figure 4. Gross and release number of movies by year*

According to *Figure 4*, the annual number of movies released increased gradually from 1980 to 1999 and stayed steadily until 2015. We noticed there was a sharp decrease in 2019 and 2020, which is probably caused by COVID-19 pandemic. Reversely, the average gross experienced a steady increase.

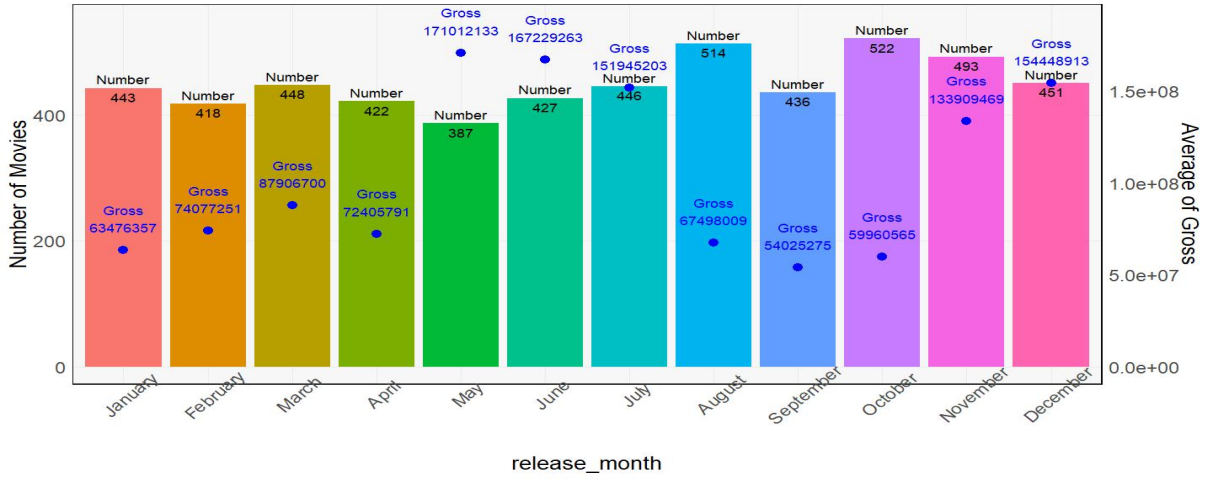


*Figure 5. Gross and release number of movies by month*

The most popular months for movie release are October and August, followed by November, which might relate to holidays. Movies released in May and June have largest average gross, probably related to film festivals. Therefore, releasing movies in these months may earn larger gross.

### 2.2.2 Release Region Analysis

We compared various features across countries, using gross as an example here since we'll show more in our Shiny App. According to *Figure 6*, we found that Chinese movies generated the most average grosses from 1980 to 2020, followed by the United States. Focusing on marketing and promotion in these two countries is likely to contribute to a movie's success.

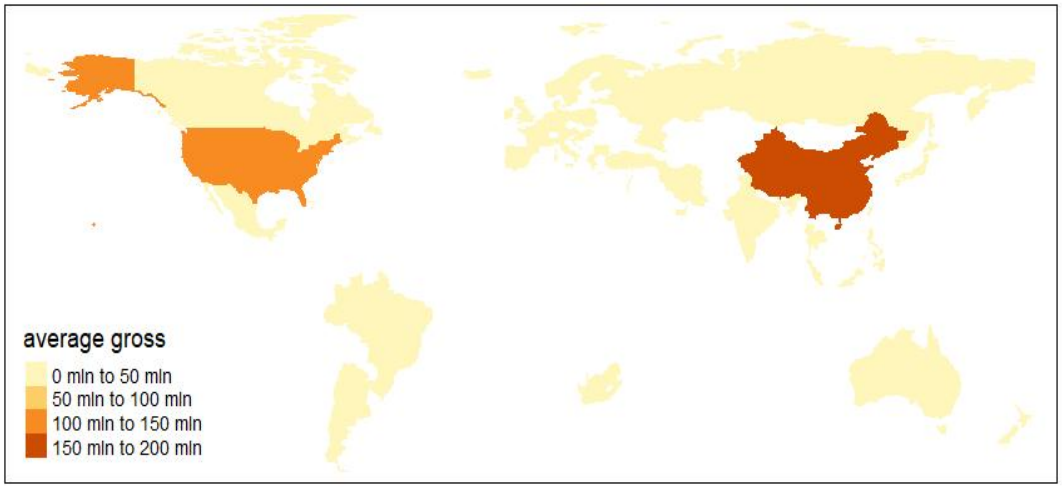


*Figure 6*

We found movies released by the United States are far more ahead of other countries. So, we then compare detailed movie features only in the US..

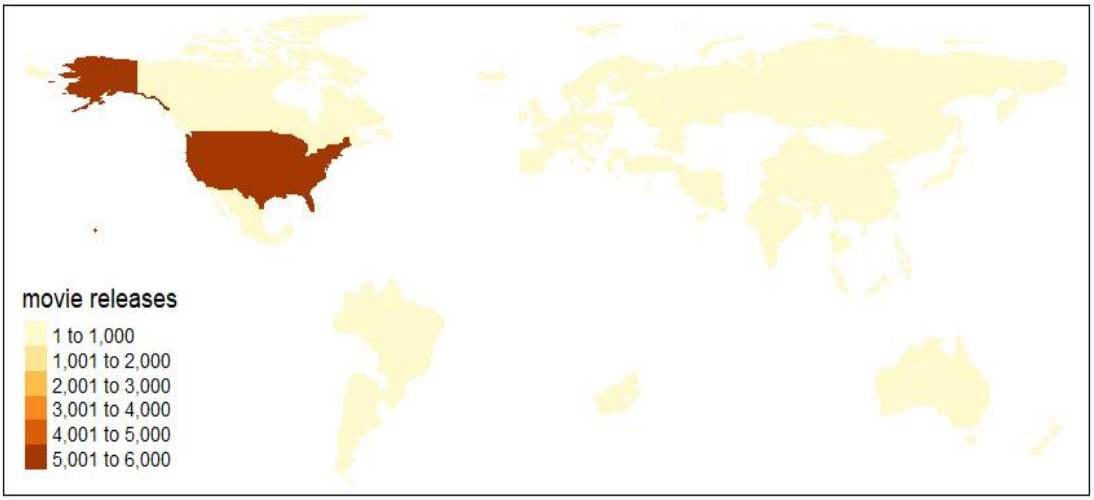


*Figure 7. Number of movie releases worldwide*


*Figure 8. Comedy and Adventure movie released in the US by year*

We made comparisons among different genres. The United States produced less Comedy but more Adventure movies from 2014 to 2019. Considering there was a steady increase in grosses earned by the US movie industry, we conclude adventure movies might be more popular.


*Figure 9. American movies gross by year*

## 2.3 Directors and Actors

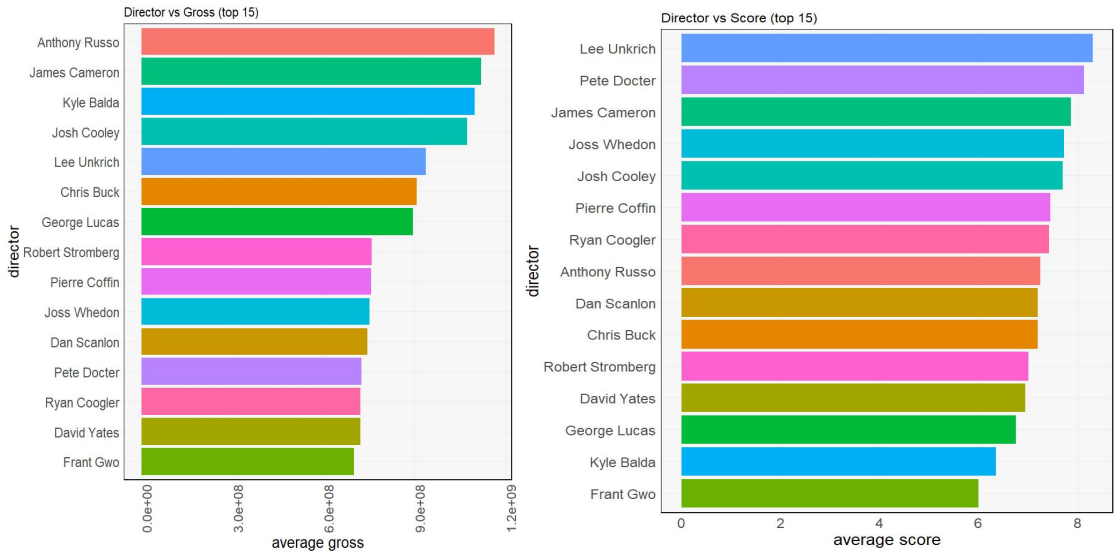


*Figure 10.*

Proportion of overlapping directors between these two plots is 100%, which means directors whose movies could earn large grosses can also be high-score.

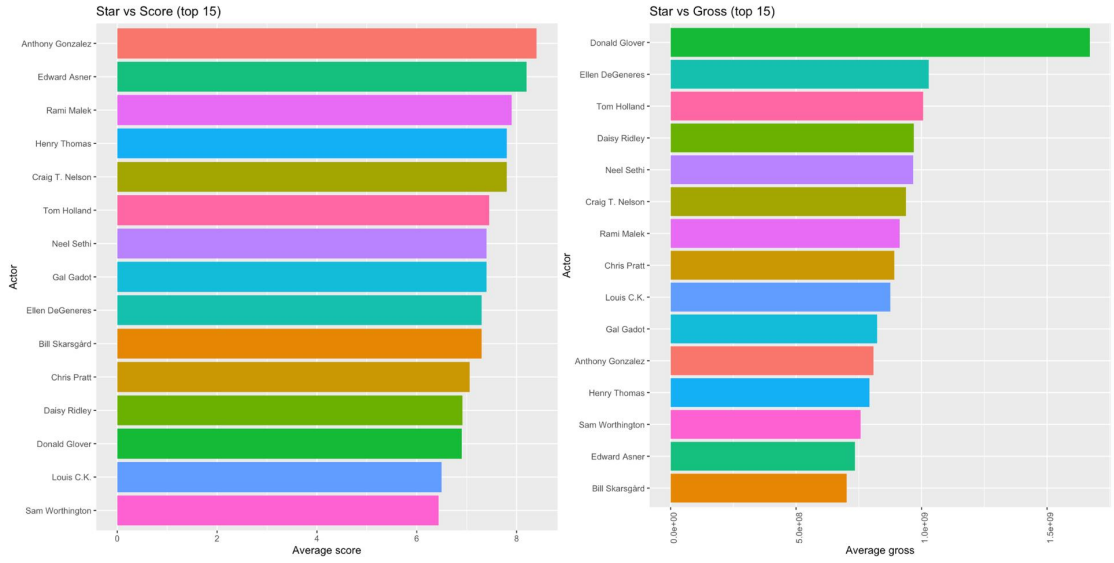


*Figure 11.*

*Figure 11* shows the relationships between stars and average grosses or average scores. Proportion of overlapping is also 100%, which provides the same inspiration as *Figure 10.*

Then, we use Bipartite Graph to find relationships between directors and stars.

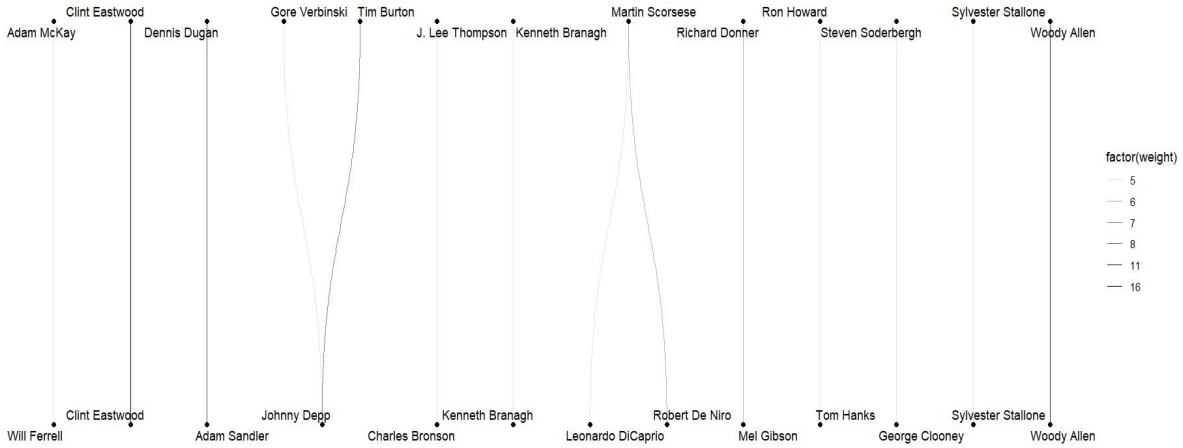


*Figure 12. Bipartite Graph*

Transparency of edges represents cooperation times. Here we only focus on cooperation times greater than 4. We find endpoints of brightest edges are the same, which means a director most likely participates in his movie.

## 2.4 Runtime Analysis

In this module, we analyzed the relationship between runtime and cross in different rating groups, approximately classified in R, PG-13 and other. The scatter plots effectively showcase outliers in the dataset, which represent the "legendary" movies that have achieved unprecedented success in gross revenue compared to their counterparts. Within the genre of comedy, we can see from the graph that a higher gross often aligns with shorter runtime.

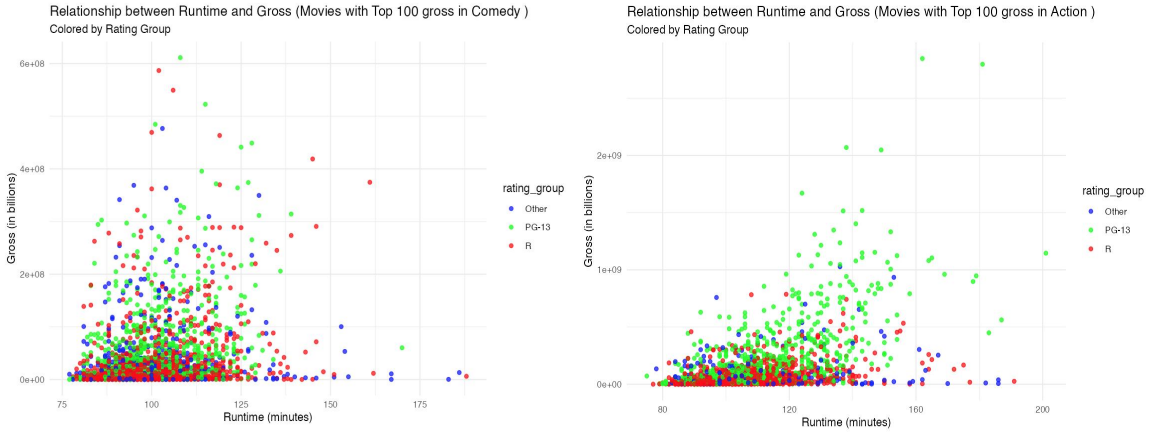


*Figure 13. Runtime-Gross by different rating group in comedy and adventure*

## 2.5 Prediction

To practicalize, we built a decision tree model to predict budget. We used features analyzed before: genre, release month, release year, release region and runtime. For features except runtime, we transformed them to factor vectors. Before transforming, we assume feature's impact is positively related to its corresponding gross. Taking release month as an example, we found May, June, December, July and November had five largest grosses, then we changed "May" to 5, .., "November" to 1 and others to 0. We also transformed runtime and output budget to their logarithm value to make predictions more accurate. Given the required data for a new movie, we can predict its corresponding budget.

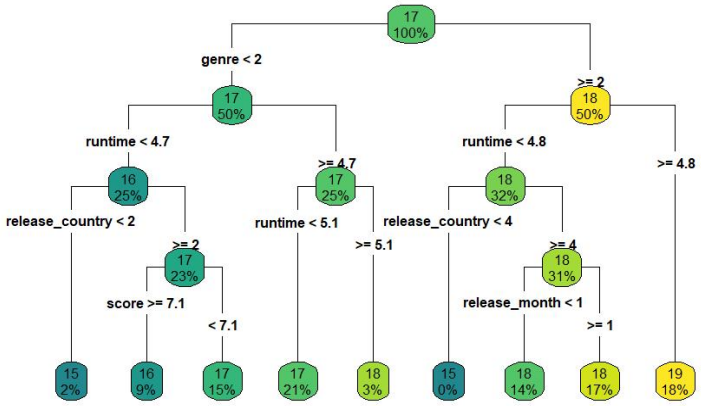


*Figure 16. Display of Decision Tree Model (RMSE: 1.14)*

# 3. Interactivity Design

In this section, we aim at enhancing user experience and providing a dynamic exploration of the dataset. We created a shiny app displaying the analysis we have done above, delving into the interactive features implemented using d3, where users can see the corresponding visual results by selecting different questions of interest. The link to shiny app is here: https://siyanwang.shinyapps.io/679groupproject

## 3.1 Data Overview

When users select “Dataset Overview” as an interesting question, the shiny app provides 3 options for analysis type: Correlation Matrix for Numerical Features, Data table sorted by score, and prediction analysis.

The first option displays Figure 1. for users' reference, and the second option is linked to a special page showing the organized movie data set on which our analysis is based. The third selection provides a prediction system which is mentioned in 2.5. Users can choose Genre, Release region, Release Month, and Release Year in a slider, and stipulate the movie's runtime. The shiny app will provide a prediction result on the movie's budget.

## 3.2 Region and Release Time Analysis

Both these analyses are based on temporal data. In the first section, users can choose their interest analysis type between Release Movie Number and Overall Country Distribution. When selecting the former one, users can explore detailed movie release conditions of specific regions and genres, visualized by representing movie numbers using gradient colors across 5 intervals on the selected region's map plot. For the latter one, we provided an additional link where users are able to explore the overall distribution of movies across countries by d3 visualizations including pie charts with mouse move, covering tags, etc.

The link to d3 visualization: https://vivianbeagle.github.io/STAT679/pie.html

In the second section, users can easily choose to explore the relationship between the number of released movies, year or month. We also displayed the average gross in the specific month or year on the plot.

## 3.3 Runtime Analysis

Users have the option to explore the relationship between runtime and two features: gross and score. When selecting gross, they can easily choose the year range and movie genre to observe the scatterplot and line plot which is similar to the plot in 2.4.

Specifically, we added the linear regression result on scatterplot, along with the insight into the overall trend and possible mistake area. However, if their choose score, we provided an additional link to an interactive html page using d31 where users can click to select a bar chart

---

1 Reference code source: https://github.com/krisrs1128/stat679_code/blob/main/examples/week6/week6-3/imdb-linked.js

depicting the count of genres. This allows users to observe a scatter plot of runtime versus score for all movies in the selected genre. When hovering over any point, detailed information about the corresponding movie will be displayed. Link to the page is here:

https://vivianbeagle.github.io/STAT679/runtime.html

### 3.4 Budget Analysis

In this part, we exhibited our results on the analysis between movie budget and gross, as well as the analysis on movie budget and movie score. Excepting the exhibition of static plot, we created a d3 page showing the trend of budget in different movie genres. Each point on the scatter plot represents a movie, while each line in the line chart represents the trend of budget by year on one kind of movie. When users' mouse hovering over one point in the scatter plot, the line chart will highlight one line whose category is exactly the same as the chosen movie..

The link to d3 page: https://siyanwang123.github.io/stat679_budget_score/imdb-table.html

### 3.5 Directors and Actors Analysis

In this section, users can choose analysis subjects from director or actor versus grosses or movie scores, which will display the bar plots above in *figure 10* and *figure 11*, which only shows the relationship between the top 15 directors, actors and the average grosses and movie scores respectively.

## 4. Conclusion

Inspired by a thorough literature review, we explored and conducted comparative analyses of the film industry dataset, delved into multiple variables and their inter-relationship, and integrated diverse multi-level visualizations into a Shiny app for dynamic exploration. Our analyses cover budget relationships, movie releases, country-based comparisons, and insights into directors, actors, and runtimes. The interactive features implemented using d3 enhance user experience, providing a nuanced understanding of the dataset. Our project not just aims at data analysis, which is also a strategic asset for decision-makers in the film industry and other sectors,

but also showcases our commitment to providing clients with data-driven decision-making standards-setting and practical, user-friendly solutions for navigating industry complexities.