SI 618 Project Part 2 Report

Exploratory Data Analysis of the Movie Data

Xinye Jiang

1 Motivation

People nowadays love watching movies. Whether a movie is a success can always be shown by its revenue in a way, as it is not likely for an unsuccessful movie to make money. So the general goal of this project is focused on movie revenue. To be more specific, I want to find out the trend of movie revenue, what are the influencing or related factors of revenue and how these factors influence or relate to revenue. I will accomplish this goal by performing exploratory data analysis such as extracting and visualizing the relationships between movie revenue and release time, genres and so on by data.table operations and ggplot2 functions.

The three specific questions that I decide to explore are as follows.

Question 1: How do temporal factors influence movie revenue?

Question 2: What is the relationship between movie revenue and movie budget for each genre?

Question 3: How do vote count and vote average relate to movie revenue?

2 Data Source

Movie Basic Information Dataset

Source: https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv

This dataset regarding the basic information of movies is available on kaggle and can be downloaded as a CSV file. It contains 24 variables and 45,466 records which cover movies released from 1874 to 2017. Considering the completeness of the movie information, I decided to use the 29,373 records of the movies that were released from 1985 to 2015. Each record mainly shows the genres, budget, revenue, original language, popularity, production companies, production countries, release date, runtime, release status, vote average and vote count of a movie.

The variables of interest in this project are *id*, *revenue*, *release_date*, *runtime*, *genres*, *budget*, *vote_average* and *vote_count*. *release_date* has a data type of date. *genres* shows a list of dictionaries with a disparate genre in each dictionary. Each movie may have several genres, such as adventure, animation, comedy and etc. Other important variables are all numeric. The *vote_count* variable displays how many ratings a film's score is based on and the *vote_average* variable shows the weighted average ratings of movies. *vote_average* is a reliable measure of the movie ratings, as it applies filters to eliminate and reduce attempts at vote stuffing.

3 Methods

The dataset was not in a clean and nice format and thus some data manipulations were still needed to prepare the dataset for the three proposed questions.

The source code can be found in the corresponding part of **si618-project-part-2-xinyej.Rmd**.

3.1 Question 1: How do temporal factors influence movie revenue?

3.1.1 Question 1: Manipulation

As for data preprocessing, I firstly changed all these variables into the right data types, because some original records mixed a few features and the dataset thus had unreasonable data types for some variables like *id*. I then picked out the movies that have *status* "Released" and were released from 1985 to 2015 based on *release_date*. Finally, I kept only the necessary variables that I am interested in for convenience, i.e., *id*, *revenue*, *release_date* and *runtime* in this case. These manipulations were accomplished by the basic functions regarding the data type transformation as.integer(), as.numeric(), as.Date() and basic data.table operations in R.

To prepare for the analysis, I computed the movie counts and total movie revenues for each year by data.table operations and length(), sum(), year() functions. Besides, I calculated the average revenues respectively for each year and each month, each year and each day of week, and each month and each day of week by data.table operations and mean(), year(), month(), weekdays() functions.

3.1.2 Question 1: Missing/Incomplete/Noise

I simply dropped the incomplete records, as they only account for a small part of the whole data. Most movies have the abnormal value 0 for *revenue* which might probably represent missing values out of the difficulty in information acquisition. I chose to keep these movie records because of their large amount. When it comes to calculate the mean revenue of movies, all the retained records would be under consideration though those abnormal records did slightly affect the results.

3.1.3 Question 1: Challenges

I encountered a challenge that some variables such as *id* had unreasonable data types when they were read in. It is because some original records mixed a few features. In order to avoid weird bugs in the following parts, I checked every variable of interest extremely carefully, changed them into the right data types and got rid of all the wrong records.

3.2 Question 2: What is the relationship between movie revenue and movie budget for each genre?

3.2.1 Question 2: Manipulation

In order to prepare the data for analysis, after changing the variables into the right data types, I selected the movies that were released from 1985 to 2015 based on *status* and *release_date*. I removed those records that had "[]" as their *genres* from the dataset and retained only the variables *id*, *revenue*, *budget* and *genres* for convenience.

To carry out this analysis regarding *genres*, the dataset needed to be reorganized to have one line of movie information for each genre. I used gsub() function and regular expression to replace all the unnecessary characters with the empty string and then applied strsplit(), unlist() and sapply() functions to grasp the

genre(s) for each movie into a vector. Finally I utilized the pipe operator %>% and tidyr::unnest() function to make each genre its own row, generating the desired dataset for question 2.

3.2.2 Question 2: Missing/Incomplete/Noise

The incomplete records were dropped due to their small amount. Most movies have the abnormal value 0 for *revenue* or *budget* which might probably represent missing values. I decided to keep these records because they account for a large part of the data.

3.2.3 Question 2: Challenges

The biggest challenge was to reorganize the dataset to make each genre its own row for all the movies. It is easy to do this manipulation by flatMap in PySpark. R also provides packages to pursue this goal similarly. I used another way that processed the data mainly by string manipulation functions. I used regular expression and gsub(), strsplit(), unlist() and sapply() functions to get the genre(s) as a vector for each movie and applied pipe operator %>% and tidyr::unnest() function to obtain the desired dataset.

Another challenge was that the variable *budget* had wrong data type "factor" when it was read in and I got inconsistent numbers when I changed it into type "numeric". In order to fix this problem, I added an argument "stringsAsFactors = FALSE" when I read the original dataset using read.csv() function. In this way, I could get consistent numbers when correcting the type of *budget*.

3.3 Question 3: How do vote count and vote average relate to movie revenue?

3.3.1 Question 3: Manipulation

Similar to the manipulations in the previous questions, I changed the variables into the correct data types and retained only the movies that were released from 1985 to 2015 based on *status* and *release_date*. I reserved the variables *id*, *revenue*, *vote_count* and *vote_average* to facilitate the following computation. These data manipulations in question 3 were accomplished by the basic data.table operations and the data type transformation functions as.integer() and as.numeric() in R.

To prepare for analysis in question 3, I rounded the vote average to the nearest 0.5 and calculated the mean profit for each vote average group by round() and data.table operations. Similar manipulation was also done to the vote count that I rounded the vote count to the nearest 1000 and computed the mean revenue for each vote count group.

3.3.2 Question 3: Missing/Incomplete/Noise

I also removed the incomplete records as a result of their small amount. As mentioned before, most movies have the abnormal value 0 for *revenue* which might represent missing values. I chose to keep them as they account for most of the data.

3.3.3 Question 3: Challenges

The challenge that I had was also the same challenge that I had in question 1. Some records mixed a few features and caused some variables to have wrong data types. I solved this problem by carefully checking the variables one by one and changing them into the right types.

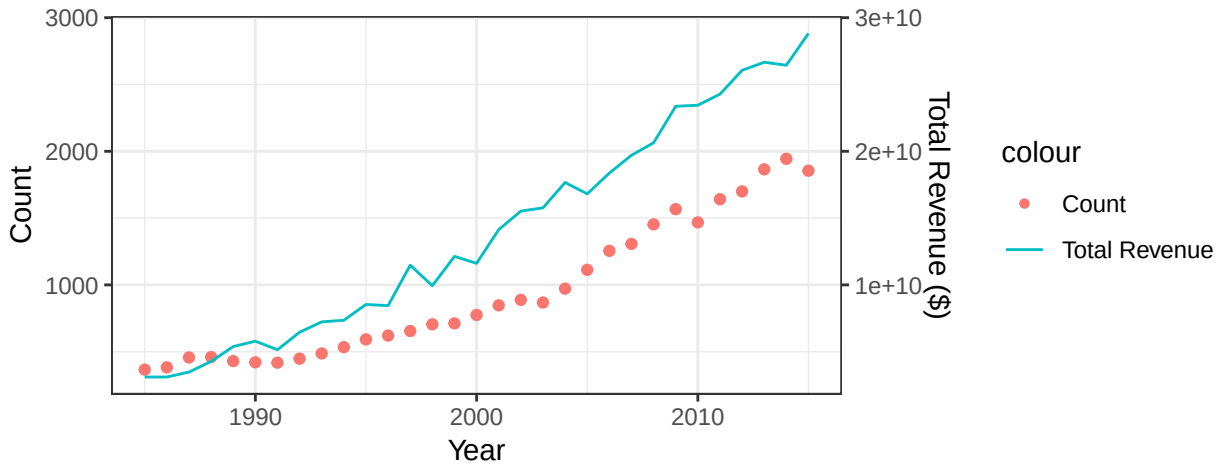

Figure 1: Overall Trend of Movie Revenues and Counts

4 Analysis and Results

The analysis and visualization work was done in R using data.table methods and ggplot2 functions. The source code can be found in the corresponding part of **si618-project-part-2-xinyej.Rmd**.

4.1 Question 1: How do temporal factors influence movie revenue?

The first data analysis was to investigate the importance of temporal factors to movie revenue and how do temporal factors influence movie revenue.

4.1.1 Question 1: Workflow of the source code

To have an overall idea of the trend of movie revenues, I first computed the total movie revenues and the number of movies for each year by data.table operations and visualized the result using lines and points in Figure 1 by ggplot2 functions. Note that I adjusted the y scales and added two different y axes on the same plot to make the results shown in one plot by setting the sec.axis argument in the scale_y_continuous() function.

Secondly, in order to find out the temporal patterns of movie revenues, I calculated the average revenues respectively for each year and each month, each year and each day of week, and each month and each day of week by data.table methods and mean(), year(), month(), weekdays() functions. I then displayed the results in Figure 2 using ggplot2 functions and grid.arrange() by three heat maps with darker color representing higher mean profits.

I finally examined the effect of a movie's runtime to its profit by a scatterplot in Figure 3 with runtime as x and revenue as y, using mainly ggplot() and geom_point() functions.

4.1.2 Question 1: Result and Visualization

Figure 1 shows that both the movie counts and total revenues have increased steadily over years, and the total revenues grow faster than movie counts. So we can conclude that more movies are produced and they overally tend to make more money as time goes on.

The first two heat maps in Figure 2 also indicate that the average profitability of movies has increased in the recent years, as in general the color is darker on the right side of the figures.

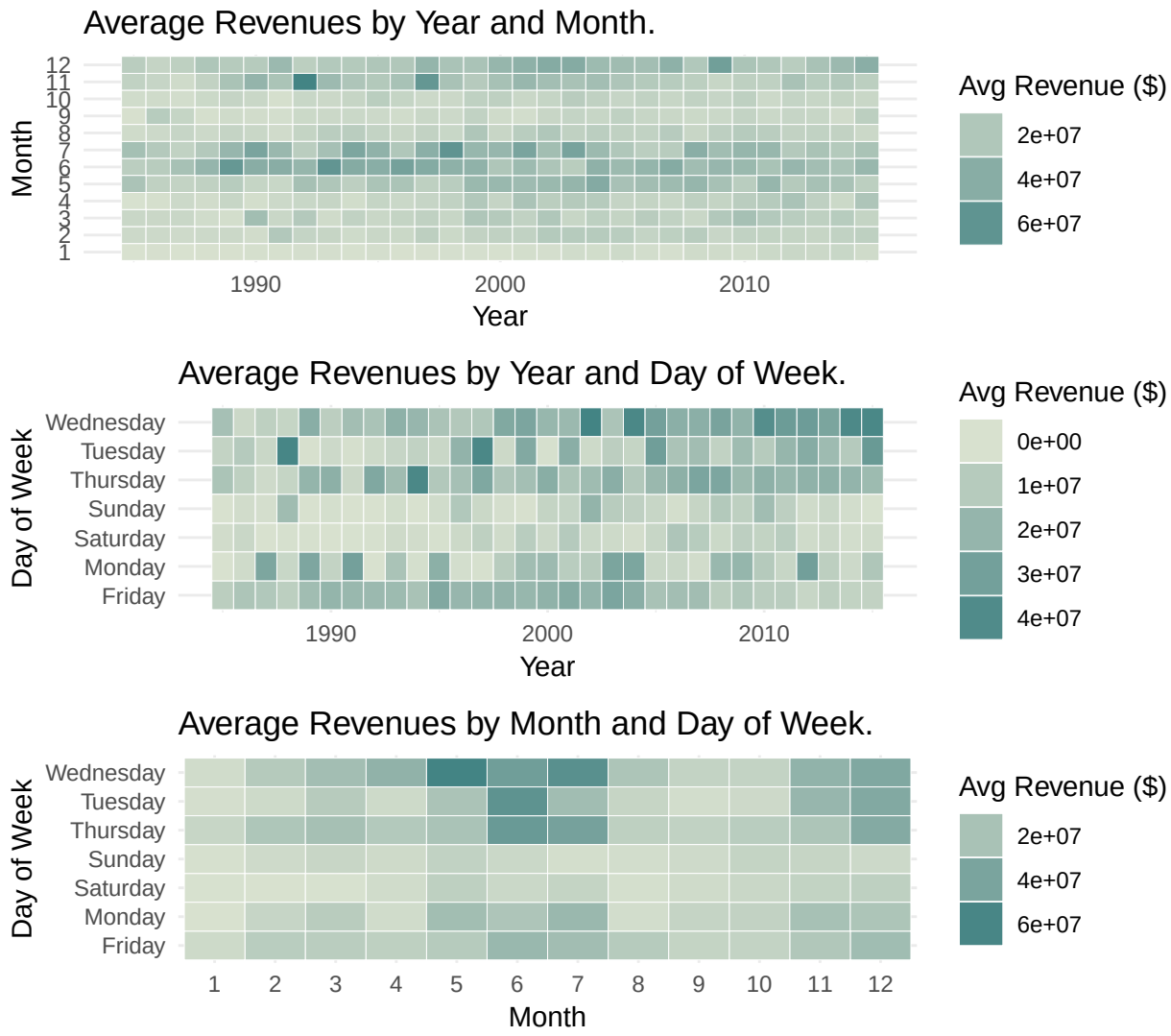

Figure 2: Temporal patterns of average revenues

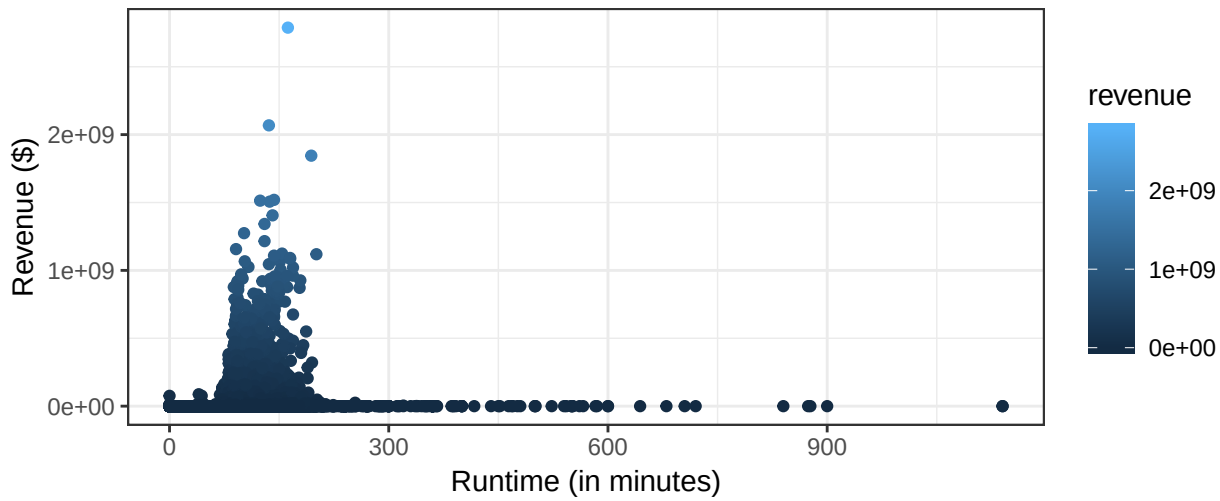

Figure 3: Revenue vs Runtime

The first and third heat maps show that months like June and July, followed by May, November and December, are particularly important for movie releases, as the average revenue of movies released in these months can earn 2 times more than the average revenue in other months. It might result from the fact that people, especially some groups such as students, may have more free time to watch movies during summer and winter. From the first heat map, the average revenue differences between months decrease over years, probably due to the convenience to watch films nowadays.

The last two heat maps indicate that movies released on Tuesday through Thursday are most likely to earn better profits, while those released during the weekends have a tendency to make less money. This might be because people can make plans in advance to see the movies that are released and promoted for a few days on Friday night or at the weekend.

Figure 3 has no specific pattern and the range of revenues seem random for movies of different runtimes. Movies that have extremely long or short runtimes tend to have no record for *revenue*, i.e., revenue value 0 in this case. Runtime is not likely to have much effect on movie revenue.

In conclusion, movies have increasing numbers and average profitability over years. Movies released on Tuesday through Thursday in June and July overally earn best revenues. The runtime of a movie does not have much effect on its revenue.

4.2 Question 2: What is the relationship between movie revenue and movie budget for each genre?

In the second analysis, we looked into the relationship between revenue and budget for each genre.

4.2.1 Question 2: Workflow of the source code

To check how budget influences revenue for various genres, I utilized scatterplots to show the relationships. Considering the intelligibility of the plots, I showed the plots of 20 genres in four rows by using %in% (to filter data), data.table operations and grid.arrange(..., nrow=4). I also added grey points and a black "y=x" line for better comparison of budget and revenue for each genre by geom_point(color="grey"), geom_abline(intercept=0, slope=1) and facet_grid(.~genres).

4.2.2 Question 2: Result and Visualization

Figure 4 shows the relationship between revenue and budget for each genre. We see that various genres produce different numbers of movies. The documentary, foreign and tv movie genres have the least number of movies. Movies of the adventure, fantasy, science fiction and action genres, however, account for a large part of all the movies.

The scatterplots also show the positive correlation between profit and budget. In general, movies that have higher budgets seem to actually see the benefits of these budgets through their higher profits. It is understandable that higher budget movies normally could attract more people, as the high budget might be used to hire well-known actors or actresses, build elaborate sets, or acquire rights regarding some famous topics.

Most genres except history, foreign and tv movie produce movies that overally have revenues higher than their budgets, so generally movies are profitable. Adventure, fantasy, science fiction and action movies normally have larger budgets along with extremely high profits. Popular movies, such as Avatar, Avengers and Pirates of the Caribbean, mostly belong to these genres and could make quite a fortune despite their large budget. Movies of the genres animation, comedy, drama, family, romance and so on have moderate budgets and high revenues. And documentary and music movies have small budgets and moderate revenues.

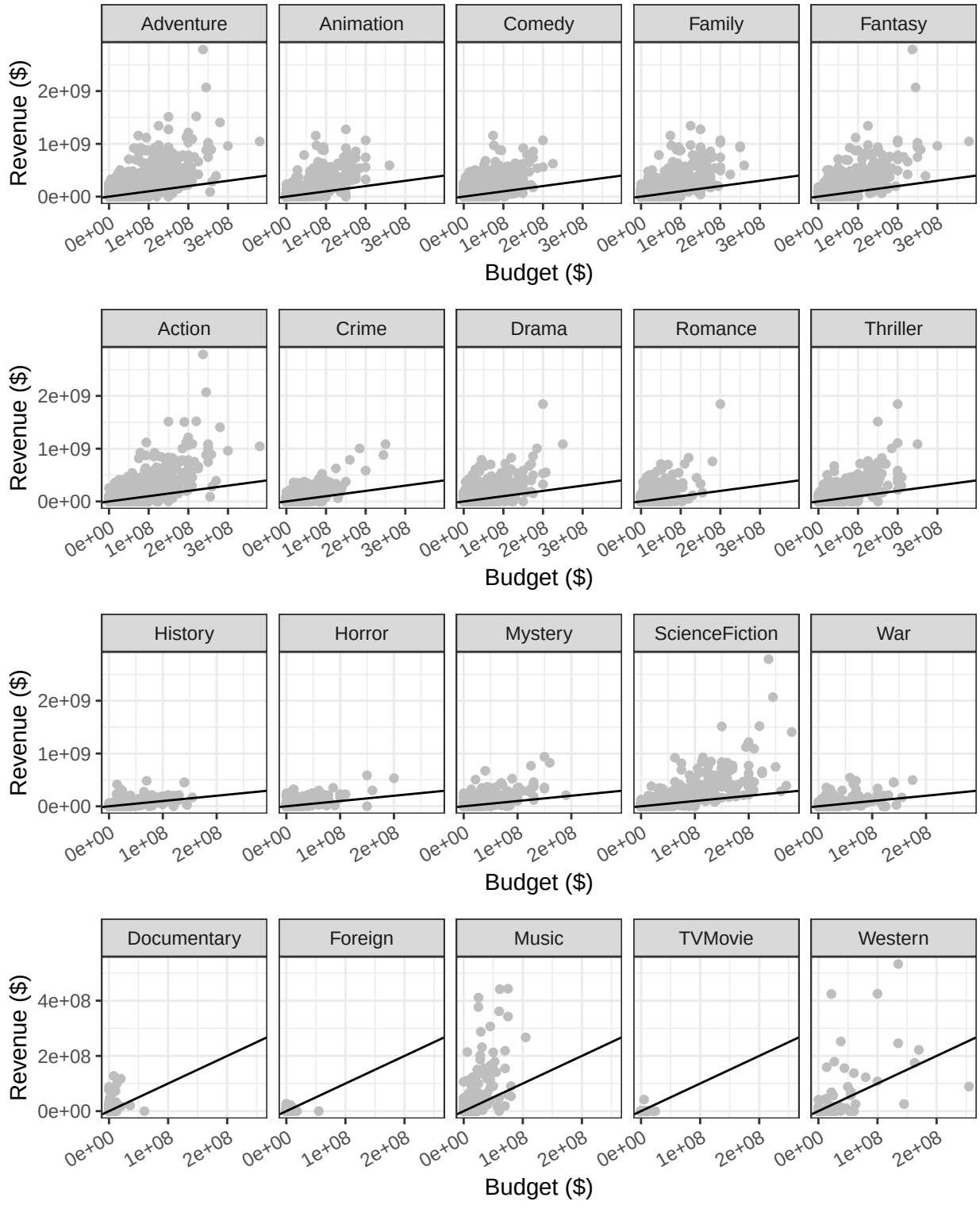

Figure 4: The relationship between revenue and budget for each genre

4.3 Question 3: How do vote count and vote average relate to movie revenue?

The goal of the third question was to examine the relationships between movie revenue and vote count, vote average. *vote_count* here displays the number of ratings and *vote_average* shows the weighted average ratings of movies.

4.3.1 Question 3: Workflow of the source code

I applied the scatterplot to show the desired result by setting vote average as x, movie revenue as y and vote count as point size and point color using ggplot(..., aes(x=vote_average, y=revenue)) and geom_point(aes(size=vote_count, color=vote_count), alpha=0.8). I also added scale_color_gradient(low="blue", high="red") to create a two color gradient showing the values of vote count from low to high.

In order to see the relationships more clearly, I showed the results for vote average and vote count separately by bar plots in Figure 6. I rounded the vote average to the nearest 0.5 and calculated the mean profit for each vote average group by round() and data.table operations and visualized the result by geom_bar(stat="identity"). Similar manipulation was also done to vote count that I rounded the vote count to the nearest 1000 and visualized the mean profit for each vote count group.

4.3.2 Question 3: Result and Visualization

The relationships between movie revenue and vote count, vote average are shown in Figure 5 and Figure 6.

Figure 5 indicates that despite that some high-rated movies have very small vote count value, there is positive correlation between vote count and vote average. In other words, high rating always comes along with moderate or high vote count. We could see that movies with higher vote count tend to make more profits, as the bigger and redder points representing the movies with higher vote count seem to correspond to larger revenues in Figure 5. Vote average also seems to be positively related to movie revenue.

Figure 6 shows the results for vote average and vote count separately. We see a pattern that the average revenue first increases when vote average increases until vote average reaches around 7.5 and then it begins to decrease when the vote average increases. After checking the data, I found out several factors that lead to this pattern. One is that the extremely high average rating like 10.0 for a movie is always made by only one person, i.e., vote count of this movie equals 1. So in this case, this rating is doubtable, as it does not reflect an honest overall evaluation of the movie. Another reason is probably that movies with high average rating over 8.5 are not necessarily the movies that are popular and are loved by everyone, and thus they have merely mediocre performance on revenue.

The average revenue also first increases and then decreases when vote count increases in Figure 6. I checked the data and found that this change might result from the fact that there is only one movie with the high vote count over 14,000 in this dataset. So the observed pattern is unreliable. We suppose that movie revenue normally first increases and later flattens when vote count increases.

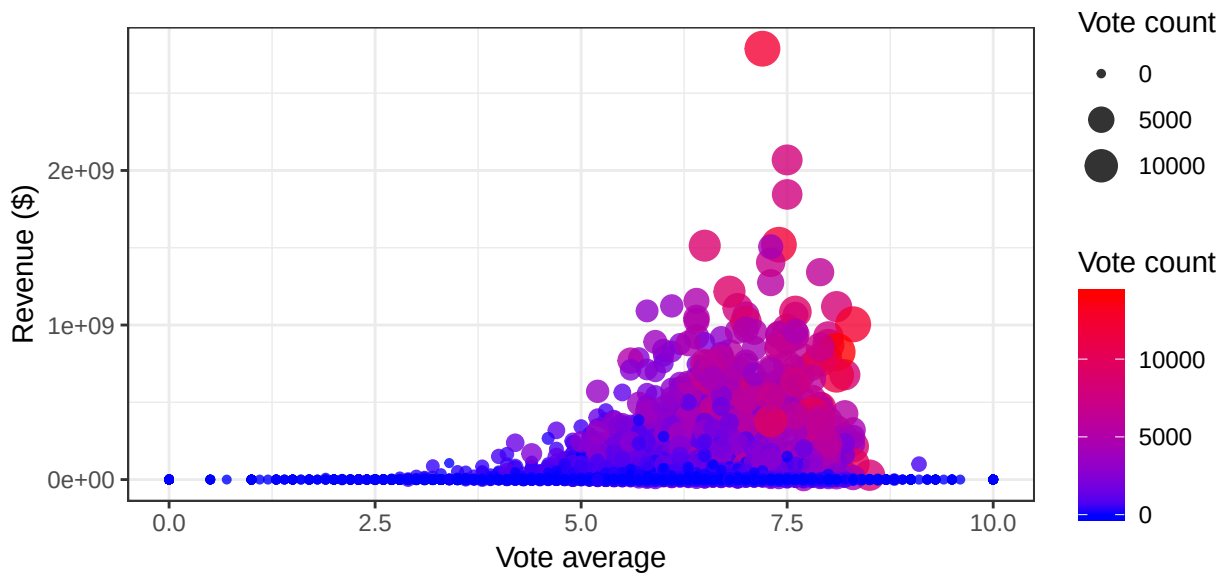

Figure 5: Revenue vs Vote average and Vote count

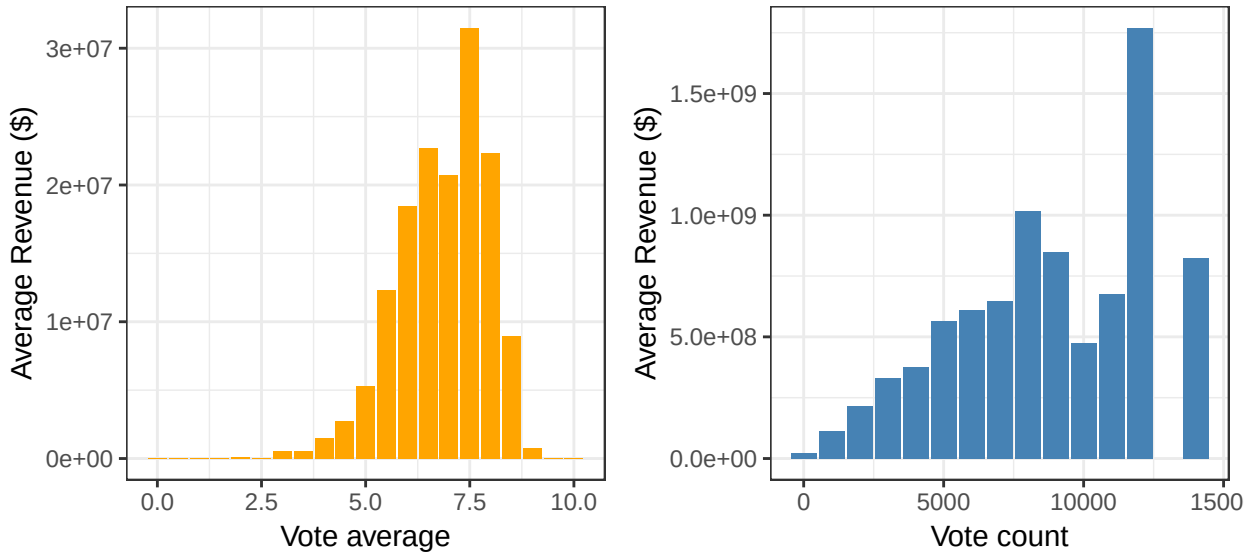

Figure 6: The relationships between revenue and vote average, vote count separately