Final Project - Step 2 (15 Points)

PSTAT100: Data Science Concepts and Analysis

STUDENT NAMES

- Luke Drushell (drushell)
- David Pacheco (dpacheco)
- Rafael Romero (romeroflores)
- Kai Barker (kaibarker)
- Diana Lee (dianadlee)
- Davian Valencia (davianvalencia)

Due Date

The deadline for this step is **Friday, May 9, 2025**.

Instructions

In this step, you will develop clear research questions and hypotheses based on your selected dataset, and conduct a thorough Exploratory Data Analysis (EDA). This foundational work is crucial for guiding your analysis in the following steps.

1 Step 2: Research Questions, Hypotheses, and Exploratory Data Analysis (EDA)

1.1 Research Questions

Question 1 How are genres linked to the vote_average of movies? And does release year play a measurable role?

Question 2 Are movies getting longer as time goes on? Does this affect any aspect of the movie? (ex: ratings and revenue)

Question 3 Does the budget of the movie affect how good it is. In other words, is the average rating affected by the movies budget.

1.2 Hypotheses

Hypothesis 1 Movies are becoming more popular but less liked: As year goes up popularity goes up proportionally, but vote average drops. Maybe this is only in action/adventure genres, whereas comedies/romances/dramas the opposite might be seen.

Hypothesis 2 Despite the shorter attention span of our generation, movies have been slowly increasing runtime, which has been seen to increase the ratings and revenue of a movie.

Hypothesis 3 While users rating or a movies quality is not always a reflection of how much money was spent, a higher budget should afford better actors, better production, etc and should lead to an overall better movie in some aspects.

1.3 Exploratory Data Analysis (EDA)

1.4 Data Cleaning

Data came relatively clean, some columns such as indices, or long descriptions are often missing so those columns will be directly removed. Finally some rows have a few zero'd values, which will be delt with on the case by case basis described in the comments.

```
library(dplyr)
dataset <- read.csv("top_1000_popular_movies_tmdb.csv")
dataset <- dataset %>% select(-tagline)
dataset <- dataset %>% select(-X)
dataset <- dataset %>% select(-id)
dataset <- dataset %>% select(-overview)

#since half the data is missing this info we'll include it
revenue_zeros <- sum(dataset$revenue == 0, na.rm = TRUE)
budget_zeros  <- sum(dataset$budget == 0, na.rm = TRUE)
cat("Rows with revenue == 0:", revenue_zeros, "\n")
```

```
Rows with revenue == 0: 4395
```

```
cat("Rows with budget  == 0:", budget_zeros,  "\n")
```

```
Rows with budget  == 0: 4649
```

```
#------

#only 175 are missing runtimes though, so they can be purged
runtime_zeros <- sum(dataset$runtime == 0, na.rm = TRUE)
cat("Rows with runtime == 0:", runtime_zeros, "\n")
```

```
Rows with runtime == 0: 175
```

```
#------

#the dataset to be manipulated from here on out!
dataset <- dataset[!(dataset$runtime == 0), ]
```

1.5 Descriptive Statistics

Some things that are interesting here: The mean average vote is 6.395, with the 1st quartile being 6. It would be possible to filter true vote_averages of 0 from movies with no votes by simply taking taking the complement of when vote_count is 0, which may be something someone mapping votes should consider. Another interesting thing here is how low the popularity mean is – many of the top 10,000 movies are not actually that popular compared to true blockbusters as shown with the max (Fast X). Finally, it is interesting to see how runtime cany vary: why is there a movie with a runtime of 2 minutes for instance? It appears there are a variety of short films in the top 10,000.

```
dataset %>%
  summary()
```

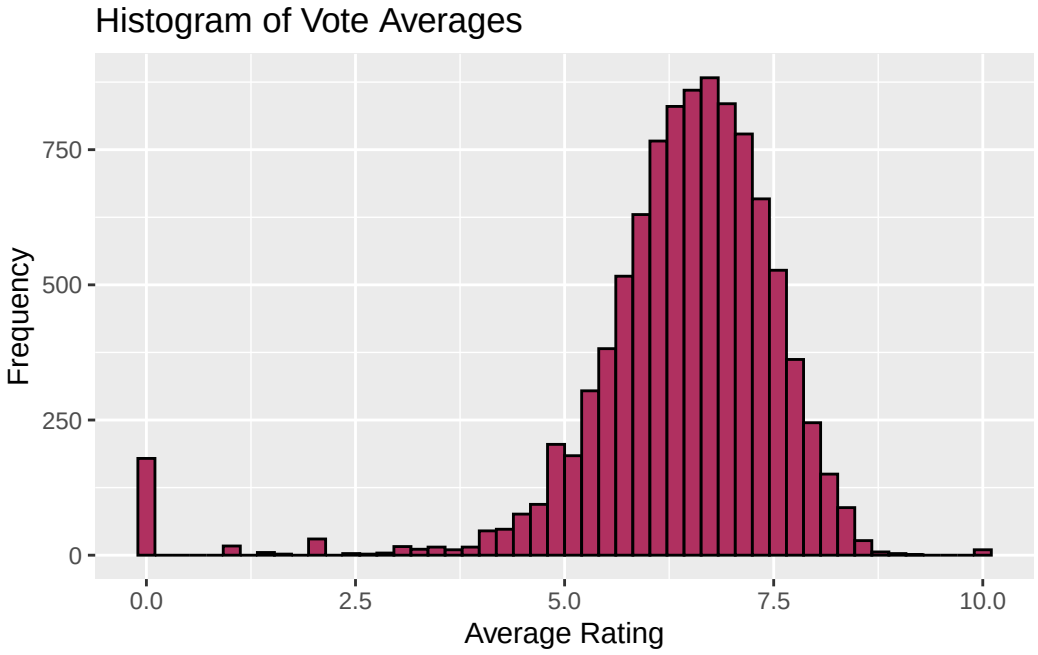
```
   title           release_date         genres          original_language
Length:9826        Length:9826        Length:9826        Length:9826
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
 vote_average      vote_count      popularity          budget
Min.   : 0.000   Min.   :    0   Min.   :  12.30   Min.   :        0
1st Qu.: 6.000   1st Qu.:  155   1st Qu.:  14.60   1st Qu.:        0
Median : 6.600   Median :  540   Median :  18.57   Median :  1800000
Mean   : 6.395   Mean   : 1586   Mean   :  33.65   Mean   : 20175183
3rd Qu.: 7.200   3rd Qu.: 1616   3rd Qu.:  27.99   3rd Qu.: 24000000
Max.   :10.000   Max.   :33822   Max.   :8363.47   Max.   :460000000
NA's   :2        NA's   :2       NA's   :2         NA's   :2
production_companies    revenue             runtime
Length:9826          Min.   :0.000e+00   Min.   :  2.0
Class :character     1st Qu.:0.000e+00   1st Qu.: 90.0
Mode  :character     Median :2.715e+06   Median :101.0
                     Mean   :6.193e+07   Mean   :102.6
                     3rd Qu.:5.343e+07   3rd Qu.:115.0
                     Max.   :2.924e+09   Max.   :366.0
                     NA's   :2           NA's   :2
```

1.6 Data Visualization

Since the vote_average variable provides some of the most meaningful data, providing valuable insight into audiences feelings on the movie, we will make several visual plots so we can explore its distribution and relationship with other variables to better understand viewer preferences and trends.

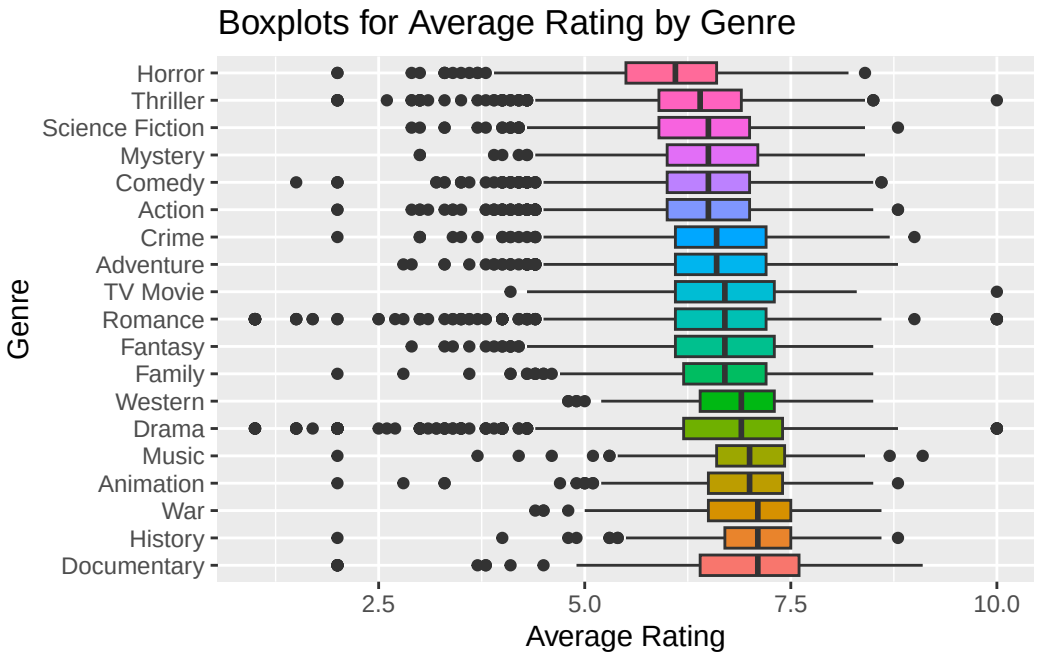
First, we can explore the distribution of the vote average over all the movies in the dataset. As a reminder, the variable is the average rating of each movie, not the average number of votes.



We can observe that the distribution of the vote averages looks approximately normal, which is further evidenced by the large sample size (~10,000) and the central limit theorem. Interestingly, there is a large cluster of movies with an

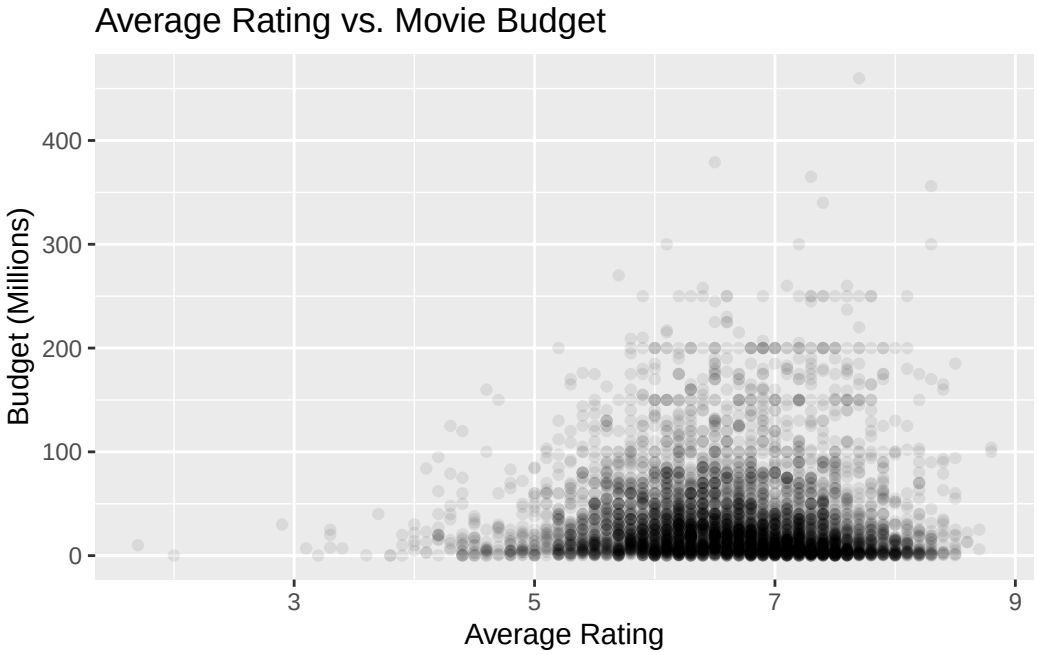
average rating of 0.0. This could be due to relevancy of the movie, where few people have cared to rate it, lack of data due to the release date of the movie, genuine dislike of the movie, or some other error. Most likely it is due to an error or lack of data for the movies rating, since any actual value of 0.0 would be nearly impossible unless it lacked a single rating. As discussed previously, these could be omitted when exploring the relationship between vote average and other variables.

Now that we have an idea of how the vote average variable looks and how it is distributed, we can explore some of its relationships with other variables in the dataset. One relationship that could be interesting to explore is the average ratings by genre. We can model this relationship with a set of boxplots, each representing a genre, ordered by median average rating (as boxplots are centered about the median), with the boxes themselves modeling the middle 50% of the data, and the whiskers indicating the spread of the non-boxed values.



From our boxplots, we can observe that the smallest median vote average comes from Horror movies, and the largest from Documentaries. While this may make more or less sense depending on personal opinions, Horror movies and Documentaries might tend to have more polarizing opinions, where documentaries might appeal to larger audiences, or audiences that were already interested in the movie, depending on the subject covered, and Horror movies might receive less praise if people are turned off by their gory nature.

Another relationship that could be important to explore is that between the vote average and budget of the movies. While budget should not inherently increase the quality of the movie, it should eliminate some aspects that could lower ratings, such as production, camera quality, visual effects, and the level of the cast afforded. We can model this relationship with a scatter plot



Values in the scatterplot have had their opacity turned down to better highlight large clusters of data. The scatterplot might not be the best method of analyzing this relationship as the vast number of datapoints leads to more clutter. However, we can observe that a majority of movies, regardless of budget seem to fall between an average rating of 5-7.5. Further, most of the movies appear to have a budget of about 50 million or less. We can also observe that the movie with the largest budget, around 450 million, does not have the highest average rating, and there is a large amount of movies with a low budget that are mixed in with the cluster of the 5-7.5 vote average. It would appear that the movies with the highest average rating have a budget of around 100 million dollars. To make a better conclusion about the relationship between budget and ratings, further statistical analysis would be warranted.