Final Project: Proposal/Dataset Selection

Morgyn Joubert, Sean McLean

College of Professional Studies, Northeastern University

ALY6015: Intermediate Analytics

Prof. Thomas Goulding

February 9, 2024

INTRODUCTION

The entertainment industry, particularly the movie sector, is characterized by the considerable financial investments and uncertainties over return on those investments. This preliminary analysis looks first at the dataset of movies to understand the relationship between the budgets and revenue. The other portion of the analysis focuses on potential correlations between movie features and its user interactions. Utilizing statistical techniques such as descriptive statistics, correlation analysis and regression modeling, we hope to derive the fundamental aspects of what makes a successful movie.

DATASET OVERVIEW

Initial exploration of the dataset revealed a dataset with a variety of movie budgets, popularity, revenue, and other characteristics. The summary statistics provided insight that the budget and revenues of movies range widely and as a result there will be that much variability within our dataset. For instance, the median budget was at $15 million, and the median revenue was approximately at $19.17 million. This sort of variability underscores the need for further analysis to understand the key drivers of movie success.

```
> # Summary statistics
> summary(movie_dataset)
     index          budget               genres            homepage              id        
 Min.   :   0   Min.   :        0   Length:4803        Length:4803        Min.   :     5  
 1st Qu.:1200   1st Qu.:   790000   Class :character   Class :character   1st Qu.:  9014  
 Median :2401   Median : 15000000   Mode  :character   Mode  :character   Median : 14629  
 Mean   :2401   Mean   : 29045040                                         Mean   : 57166  
 3rd Qu.:3602   3rd Qu.: 40000000                                         3rd Qu.: 58610  
 Max.   :4802   Max.   :380000000                                         Max.   :459488  

   keywords         original_language  original_title       overview            popularity     
 Length:4803        Length:4803        Length:4803        Length:4803        Min.   :  0.000  
 Class :character   Class :character   Class :character   Class :character   1st Qu.:  4.668  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 12.922  
                                                                             Mean   : 21.492  
                                                                             3rd Qu.: 28.314  
                                                                             Max.   :875.581  

 production_companies production_countries  release_date           revenue         
 Length:4803          Length:4803          Min.   :1916-09-04   Min.   :         0  
 Class :character     Class :character     1st Qu.:1999-07-14   1st Qu.:         0  
 Mode  :character     Mode  :character     Median :2005-10-03   Median :  19170001  
                                           Mean   :2002-12-27   Mean   :  82260639  
                                           3rd Qu.:2011-02-16   3rd Qu.:  92917187  
                                           Max.   :2017-02-03   Max.   :2787965087  
                                           NA's   :1                                
    runtime       spoken_languages      status            tagline             title          
 Min.   :  0.0   Length:4803        Length:4803        Length:4803        Length:4803       
 1st Qu.: 94.0   Class :character   Class :character   Class :character   Class :character  
 Median :103.0   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :106.9                                                                              
 3rd Qu.:118.0                                                                              
 Max.   :338.0                                                                              
 NA's   :2                                                                                  
  vote_average      vote_count          cast               crew             director        
 Min.   : 0.000   Min.   :    0.0   Length:4803        Length:4803        Length:4803       
 1st Qu.: 5.600   1st Qu.:   54.0   Class :character   Class :character   Class :character  
 Median : 6.200   Median :  235.0   Mode  :character   Mode  :character   Mode  :character  
 Mean   : 6.092   Mean   :  690.2                                                           
 3rd Qu.: 6.800   3rd Qu.:  737.0                                                           
 Max.   :10.000   Max.   :13752.0                                                           
```

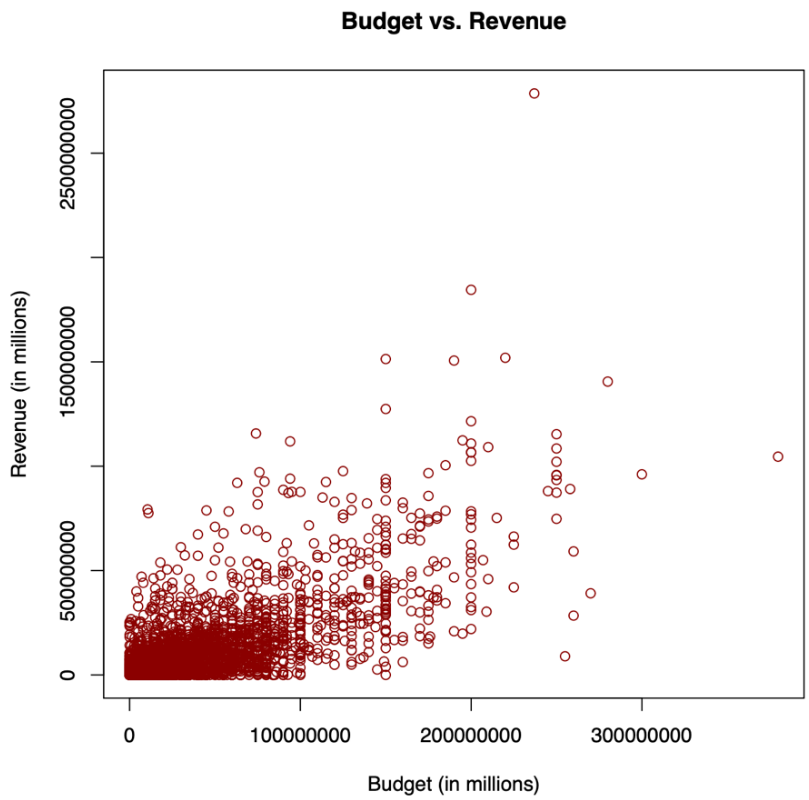
The summary statistics for the other question being studied had one small modification to allow for proper modeling. The variable ‘runtime’ had two observations with missing values, so these were removed for the subsequent heatmap to be executed.

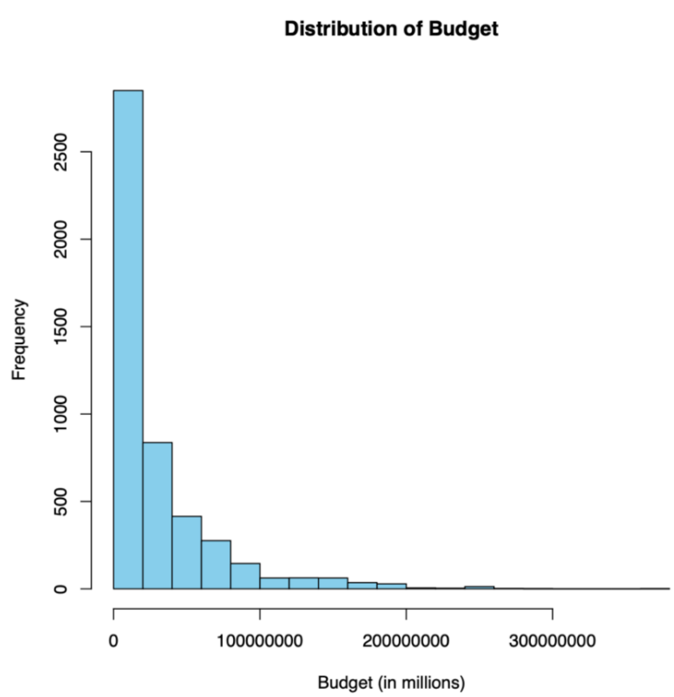
```
R 4.3.2 · C:/Users/seanm/OneDrive/Desktop/McLean_FinalProject_ALY6015/
> #Remove all observations with 'NA' in the 'runtime' variable
> movies <- movie_dataset[complete.cases(movie_dataset$runtime), ]
> #Display the modified dataset
> summary(movies)
```

```
R 4.3.2 · C:/Users/seanm/OneDrive/Desktop/McLean_FinalProject_ALY6015/
production_countries release_date        revenue               runtime
Length:4801          Length:4801        Min.   :0.000e+00   Min.   :  0.0
Class :character     Class :character   1st Qu.:0.000e+00   1st Qu.: 94.0
Mode  :character     Mode  :character   Median :1.918e+07   Median :103.0
                                        Mean   :8.229e+07   Mean   :106.9
                                        3rd Qu.:9.292e+07   3rd Qu.:118.0
                                        Max.   :2.788e+09   Max.   :338.0
```

PRELIMINARY ANALYSIS

We started our preliminary analysis of the relationship between movie budget and revenue by examining the Pearson correlation coefficient. Visualization, through histograms and scatter plots, to understand the distribution and relationship between budgets and revenues, helped us get a feel for the data.


Budget vs. Revenue
Revenue (in millions)
Budget (in millions)
0
500000000
1500000000
2500000000
0
100000000
200000000
300000000


Distribution of Budget
Frequency
Budget (in millions)
0
500
1000
1500
2000
2500
0
100000000
200000000
300000000

A linear regression model was fitted to assess the relationship between budget and revenue. The regression analysis indicated a statistically significant positive association between the two variables. The positive correlation between budget and revenue indicates that higher investments in production typically result in higher returns. For each additional unit of currency spent on the budget, revenue increased by approximately 2.9227 units of currency.

```
> # Summary of the regression model
> summary(lm_model)

Call:
lm(formula = revenue ~ budget, data = movie_dataset)

Residuals:
       Min         1Q     Median         3Q        Max 
-653371282  -35365659    2250851    8486969 2097912654 

Coefficients:
                 Estimate    Std. Error t value            Pr(>|t|)    
(Intercept) -2629555.3399  1970427.0736  -1.335               0.182    
budget             2.9227        0.0394  74.188 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 111200000 on 4801 degrees of freedom
Multiple R-squared:  0.5341,	Adjusted R-squared:  0.534 
F-statistic:  5504 on 1 and 4801 DF,  p-value: < 0.00000000000000022
```

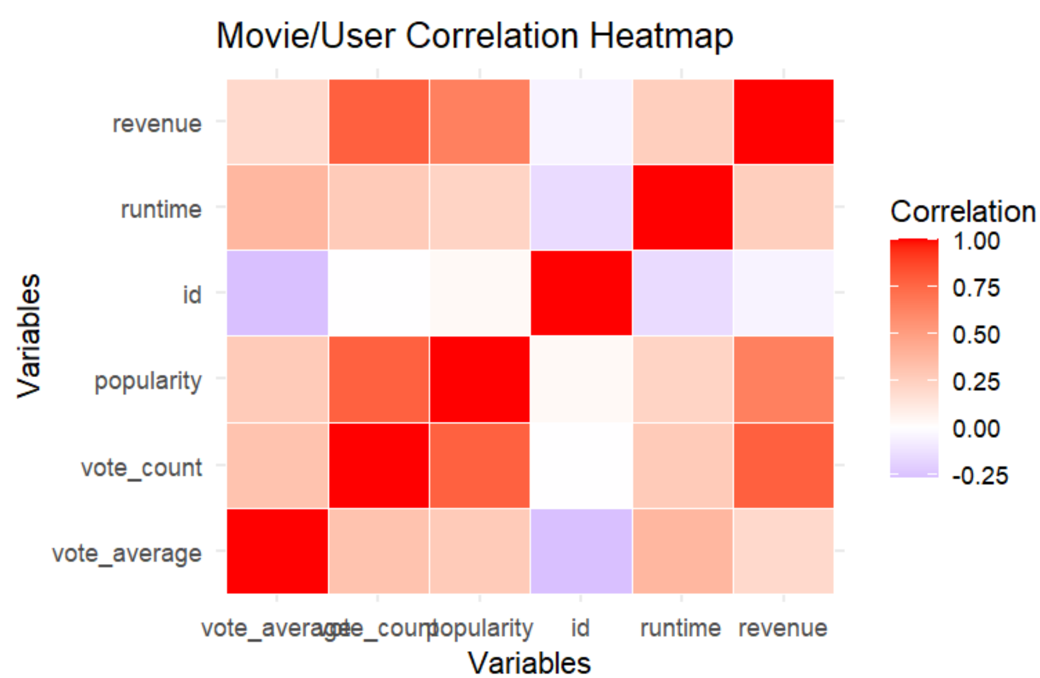
It is important to note, however, that this study naturally has its limitations, and one of them might be the inclusion of factors such as genre, release date, and marketing strategies which could have a significant influence on a movie's success.

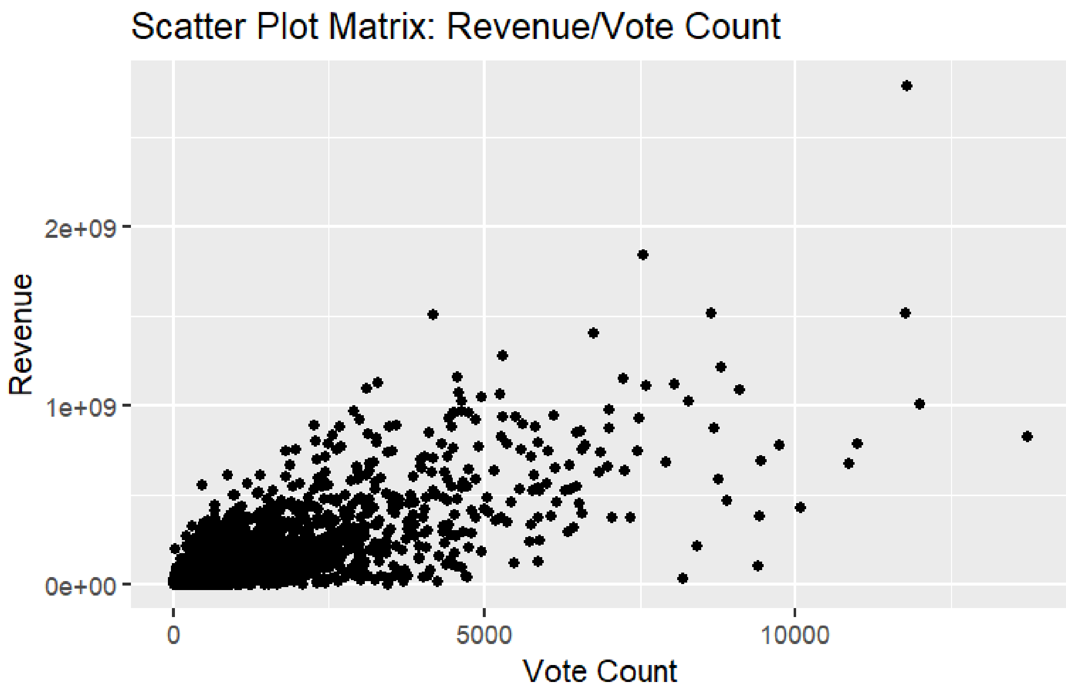
The other question in the final project looked at the correlation between user interactions and movie features in the data set. Six numerical variables were selected for a correlation analysis that pertained to movie characteristics and user actions like voting ratings. A correlation matrix was constructed to identify any positive linear relationships among the chosen variables.

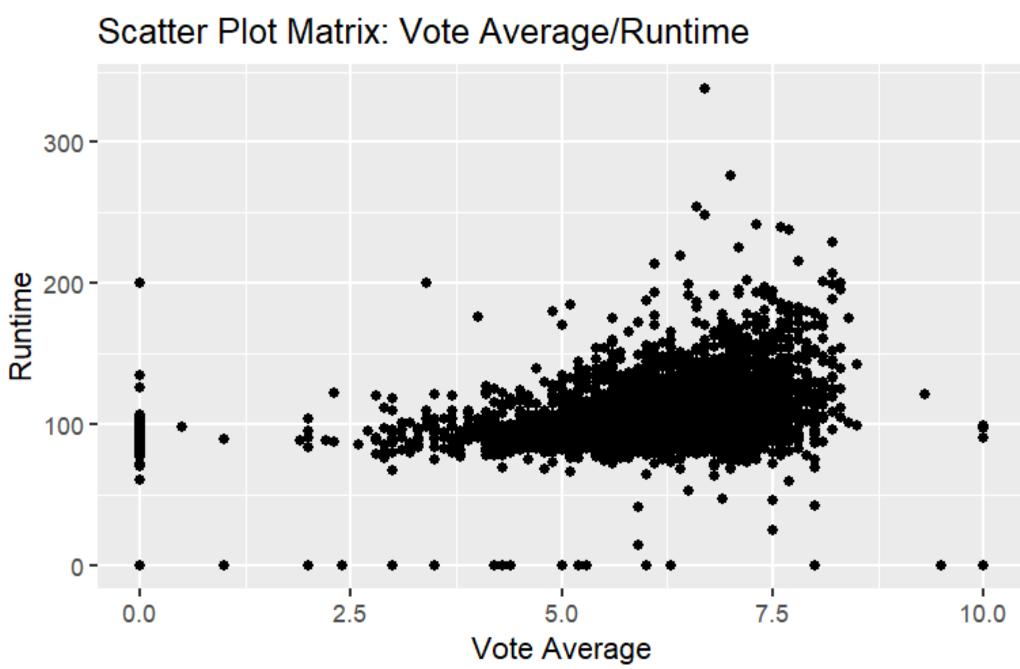
```
untime", "revenue")])
> correlation_matrix
             vote_average   vote_count popularity          id   runtime     revenue
vote_average    1.0000000  0.31342259  0.2741711 -0.26820034 0.3750457  0.19728596
vote_count      0.3134226  1.00000000  0.7780978 -0.00320646 0.2719442  0.78146223
popularity      0.2741711  0.77809783  1.0000000  0.03243460 0.2255021  0.64467729
id             -0.2682003 -0.00320646  0.0324346  1.00000000 -0.1535360 -0.04975024
runtime         0.3750457  0.27194420  0.2255021 -0.15353603 1.0000000  0.25109314
revenue         0.1972860  0.78146223  0.6446773 -0.04975024 0.2510931  1.00000000
> 
```

Interpreting the matrix there are several strong relationships with revenue, popularity and vote count standing out among the variables. The revenue variable has a high association with the vote count and popularity variables, suggesting that it affects user responses in some capacity. The id variable could be eliminated from the mix due to having negative or indifferent correlations with all the other variables.

A heatmap is used to visualize correlation matrices and identify trends and patterns in the data. To understand the relationships between two variables with strong positive linear correlations, a pair of scatterplots are built for visual representation of the interactions.







For the final submission, we will expand our investigation by including additional variables, such as genre, release date, and marketing expenditure. These will enable us to build out our regression model and account more fully for multiple factors that influence a movie's revenue. Furthermore, an in-depth analysis would include the chi-square test, ANOVA, and other advanced statistical methodologies, which will allow us to gain insights into relationships present within the dataset.

CONCLUSION

Our preliminary analysis from the first question indicates that budget plays a significant role in determining movie revenue. While it's true that higher investments are generally aimed at higher returns, there are several other success factors for filmmakers and industry stakeholders to consider. The correlations involving movie and user relationships show mainly positive associations, with certain aspects of movies like revenue and vote counts having an impact on its popularity and vote average. By conducting further analysis and utilizing more advanced statistical techniques, we can identify actionable insights that can help decision-making in the dynamic and hyper-competitive movie business.

REFERENCES

Bluman, A. (2018). Elementary statistics: A step by step approach (10th ed.). McGraw Hill.

Kabacoff, R. I. (2022). R in action: Data analysis and graphics with R and tidyverse (3rd ed.). Manning Publications.

Northeastern University – Canvas – (Panopto) videos by Prof. Thomas Goulding

APPENDIX

```
#Question 1: Correlation between Movie Budget and Popularity
# Load necessary libraries
library(readr)
```

```
library(ggplot2)

# Load the dataset
movie_dataset <- read_csv("/Users/m.joubert/Documents/Final Project: Initial Analysis Report - Joubert/movie_dataset.csv")

# Explore the structure of the dataset
str(movie_dataset)

# Summary statistics
summary(movie_dataset)

# Histogram of budget
hist(movie_dataset$budget, breaks = 20, col = "skyblue", main = "Distribution of Budget", xlab = "Budget (in millions)")

# Scatter plot of budget vs. revenue
plot(movie_dataset$budget, movie_dataset$revenue,
    main = "Budget vs. Revenue",
    xlab = "Budget (in millions)",
    ylab = "Revenue (in millions)",
    col = "darkred")

# Fit linear regression model
```

```
lm_model <- lm(revenue ~ budget, data = movie_dataset)


# Summary of the regression model
summary(lm_model)


#Question: Correlation between user interactions and movie features
#Loading necessary libraries
library(readr)
library(dplyr)
library(ggplot2)


#Importing the dataset
movie_dataset <- read.csv("movie_dataset.csv", header = TRUE)
movie_dataset


#Summary statistics of the dataset
summary(movie_dataset)
View(movie_dataset)


#Remove all observations with 'NA' in the 'runtime' variable
movies <- movie_dataset[complete.cases(movie_dataset$runtime), ]


#Display the modified dataset
summary(movies)
```

```
#Perform correlation analysis
correlation_matrix <- cor(movies[c("vote_average", "vote_count", "popularity", "id",
"runtime", "revenue")])
correlation_matrix

# Reshape the correlation matrix to long format for heatmap
library(tidyr)
cor_long <- as.data.frame(as.table(correlation_matrix))
names(cor_long) <- c("Var1", "Var2", "Correlation")

# Create a heatmap using ggplot2
heatmap_plot <- ggplot(data = cor_long, aes(x = Var1, y = Var2)) +
  geom_tile(aes(fill = Correlation), color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) + # Adjust
color scale
  labs(title = "Movie/User Correlation Heatmap",
       x = "Variables",
       y = "Variables") +
  theme_minimal()
print(heatmap_plot)

#Individual Scatter Plots Showing Relationship Between Vote Count and Revenue.
scatter_plot_matrix <- ggplot(movies, aes(x = vote_count, y = revenue)) +
```

```
  geom_point() +
  labs(title = "Scatter Plot Matrix",
       x = "Vote Count",
       y = "Revenue")
scatter_plot_matrix


#Individual Scatter Plots Showing Relationship Between Vote Average and Runtime.
scatter_plot_matrix_2 <- ggplot(movies, aes(x = vote_average, y = runtime)) +
  geom_point() +
  labs(title = "Scatter Plot Matrix",
       x = "Vote Average",
       y = "Runtime")
scatter_plot_matrix_2
```