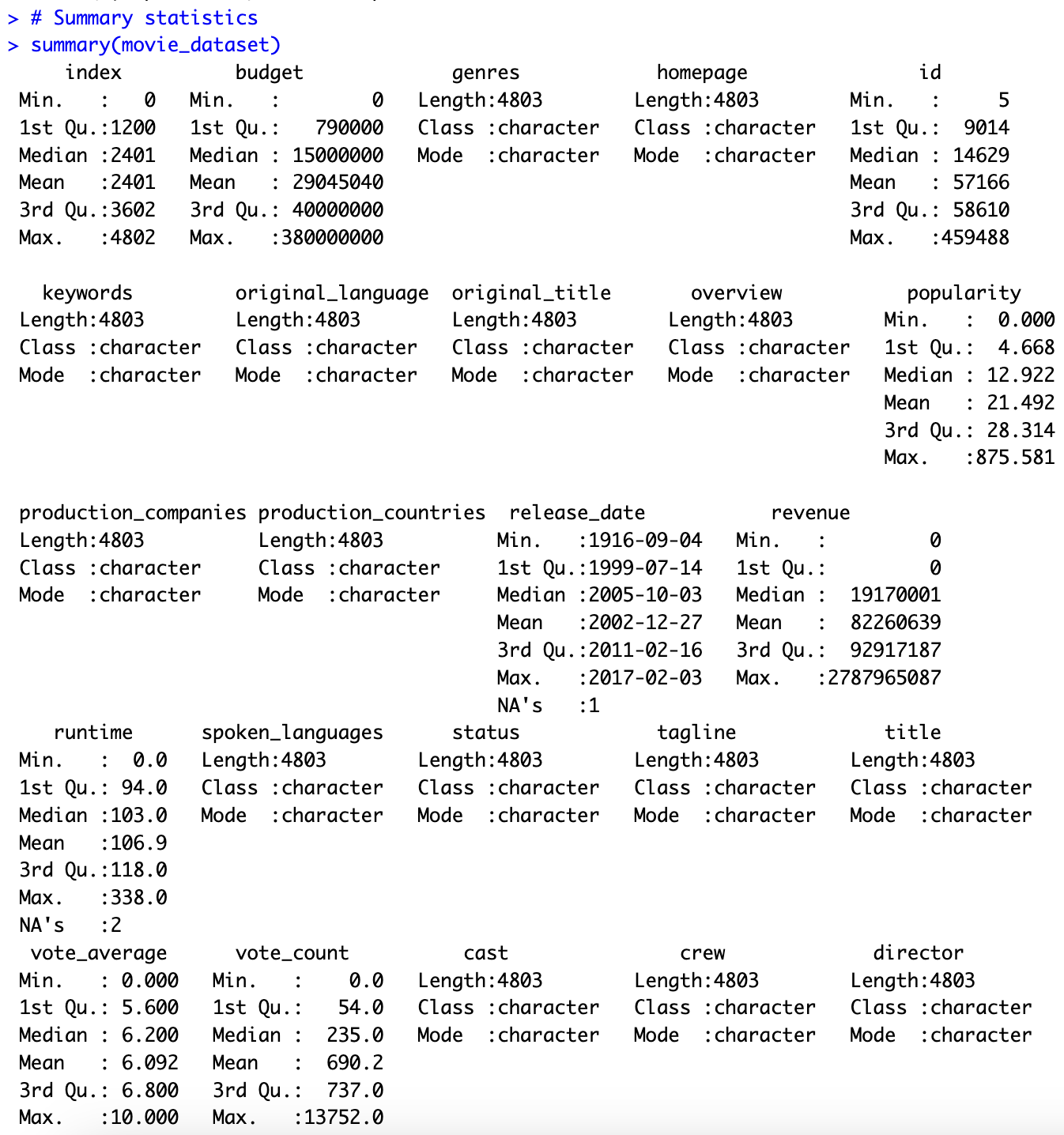
Final Project: Proposal/Dataset Selection   
Morgyn Joubert, Sean McLean  
College of Professional Studies, Northeastern University   
ALY6015: Intermediate Analytics  
Prof. Thomas Goulding  
February 9, 2024

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.



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.

A screenshot of a computer

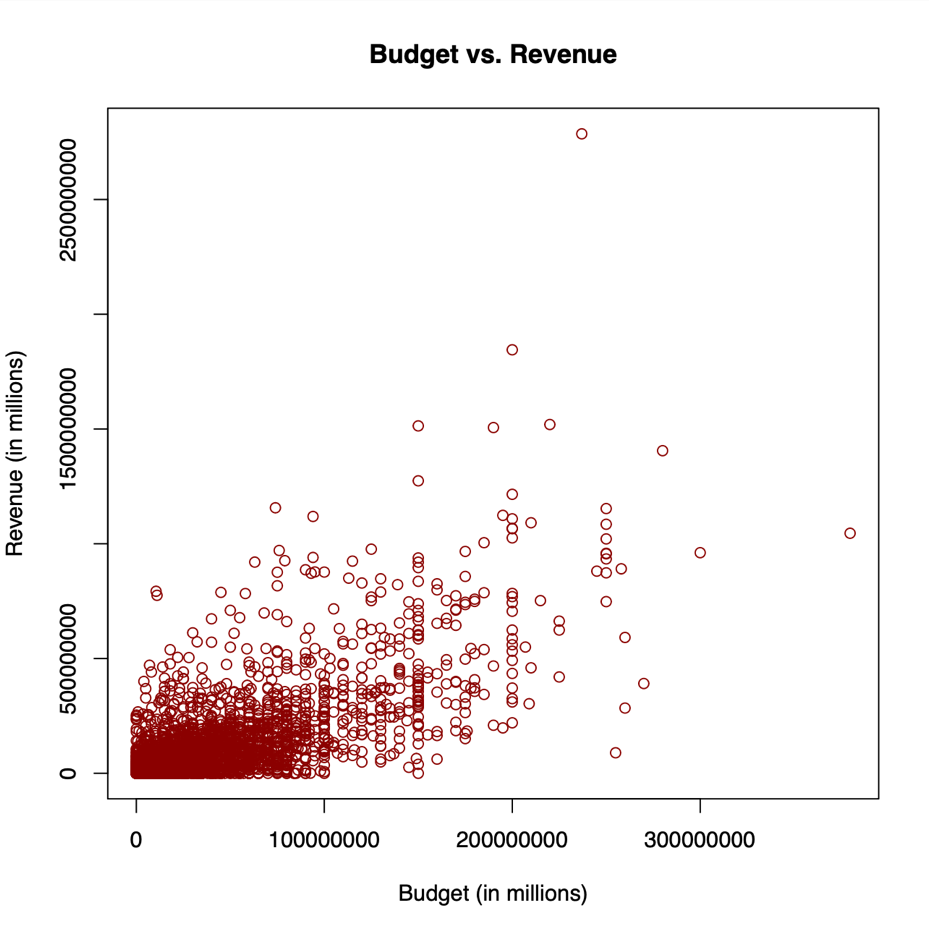
Description automatically generated

A screenshot of a computer

Description automatically generated

PRELIMINARY ANALYSIS

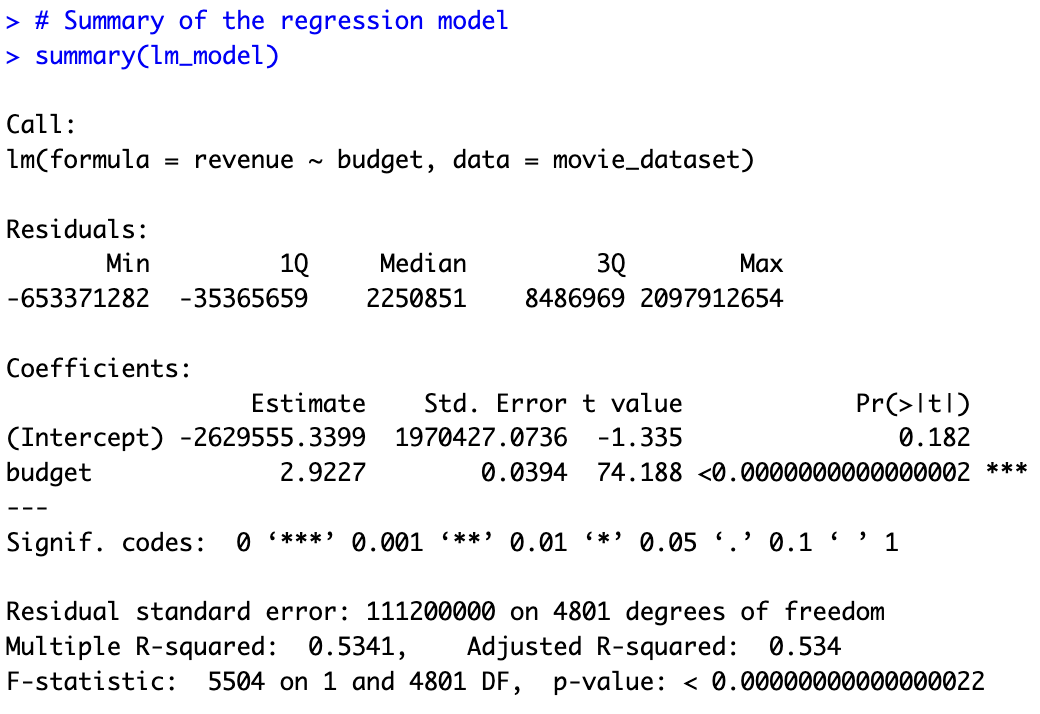
We started our preliminary analysis of the relationship between movie budget and revenue by examining the Pearson correlation coefficient. Visualization, through histograms and scatter plots, to understand the distribution and relationship between budgets and revenues, helped us get a feel for the data.



A graph of a number of bars

Description automatically generated

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.



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.

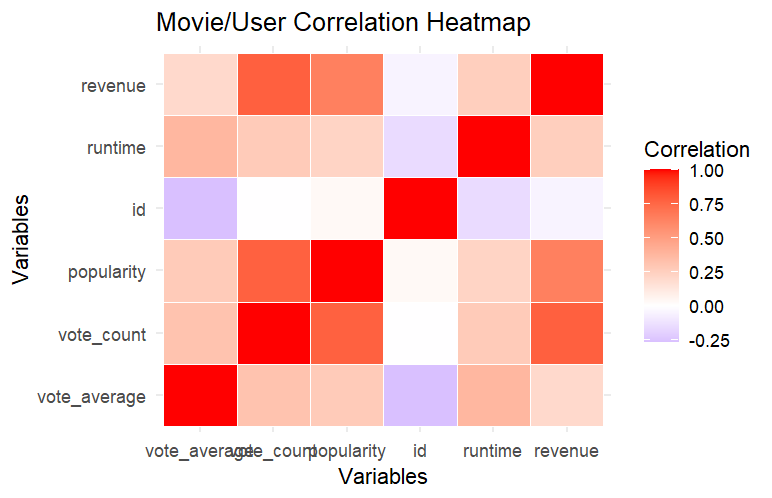
The other question in the final project looked at the correlation between user interactions and movie features in the data set. Six numerical variables were selected for a correlation analysis that pertained to movie characteristics and user actions like voting ratings. A correlation matrix was constructed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

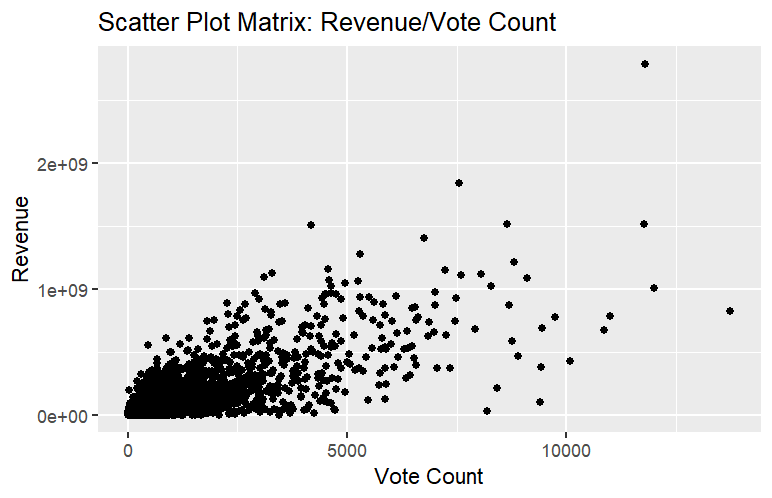
A number of numbers on a white background

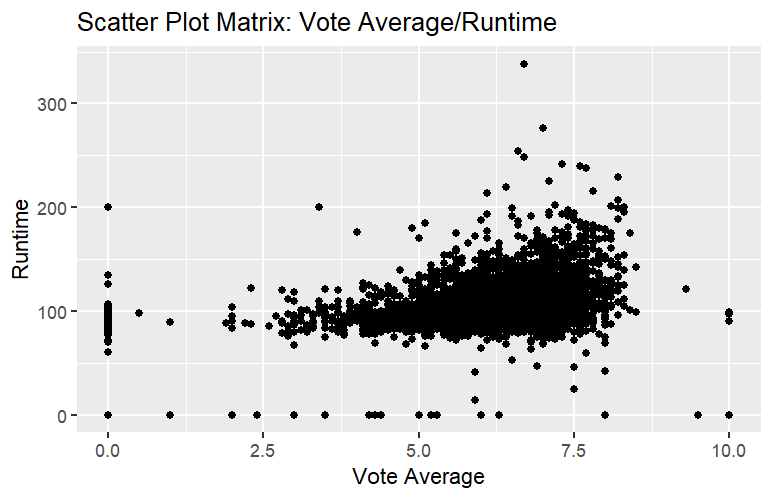
Description automatically generated

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 shows 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