Final Project Report

Morgyn Joubert, Sean McLean

College of Professional Studies, Northeastern University

ALY6015: Intermediate Analytics

Prof. Thomas Goulding

February 15, 2024

INTRODUCTION

The motion picture business holds intricacies that have long fascinated researchers. This examination looks at components pivotal to a film's prosperity. It probes the movie industry's many facets, especially on what drives a movie's accomplishment.

Through diligent data evaluation, our objective is to disentangle the intricate interplay between a production's diverse characteristics, audience interactions, and financial commitments. We will apply sophisticated statistical methods, like correlational investigation, regression modeling, and hypothesis testing, to extract practically applicable understandings.

Such insights could serve to inform decision-makers in the entertainment world. Perhaps patterns will emerge showing which alchemical combinations most effectively enthrall mass audiences, which interactions most influence box office returns or ratings, or how certain monetary choices impact a film's prospects.

Questions:

1. Is there a correlation between the user interactions and the movie features in the dataset that would be crucial for recommendation to a user?
2. Is there a significant correlation between a movie's budget and its popularity, and can this relationship be used to make informed predictions or recommendations for movie budgeting strategies?

These exploratory queries serve as illuminating navigational beacons, illuminating the way toward an exhaustive comprehension of the film industry's intricacies. Through meticulous investigation and interpretation of empirical information, we endeavor to shed radiance on the underlying instruments driving film achievement and pave the path for evidence-based dynamics in this innovative and aggressive scene.

DATASET OVERVIEW

Upon initial exploration of the dataset, we found a dataset of movies that spanned a wide variety of movie budgets, popularity, revenue, and many other characteristics. The summary statistics indicate the amount of variability there likely will be within our dataset. For example, the median budget was $15 million, and the median revenue was approximately $19.17 million. Additionally, responses to the movies in our data set from users show a large range of their own, as you might expect with variables like popularity scores and total vote counts per movie. This sort of variability implies the necessity for us to run further analyses 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
```

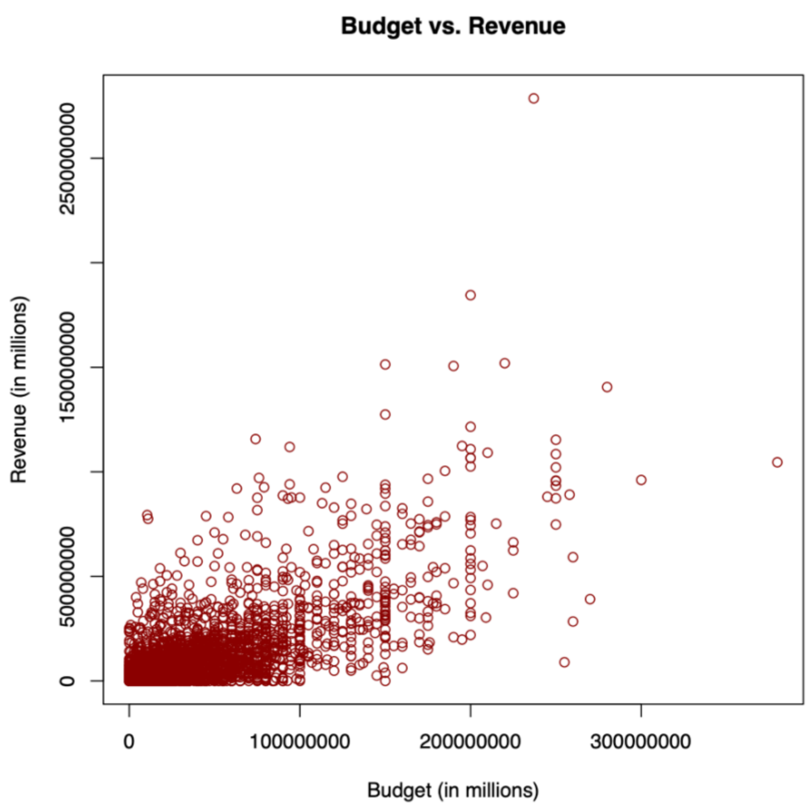
METHODS

This research utilized descriptive analysis, relationship assessment, and predictive modeling to investigate the information set and respond to the inquiries. These approaches were opted for since they could give understandings into connections among factors. Specifically, Pearson relationship coefficients were determined to assess straight affiliations, while direct regression and difference investigation were utilized to explore the impact of free factors on film income, notoriety, and client communications.

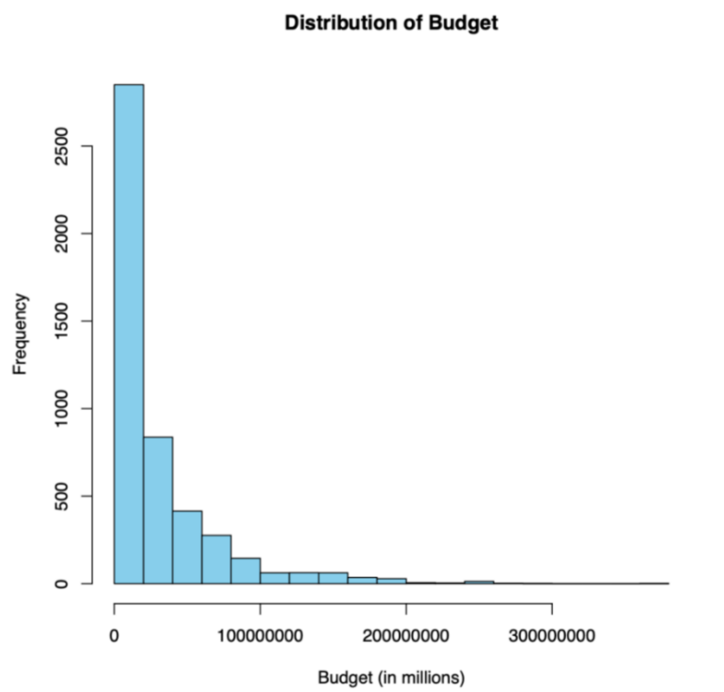
The entire examinations were led utilizing R, with works like cor() for relationship investigation and lm() for prescient demonstrating. A scope of procedures was considered dependent on their sensibility for the logical targets and congruity with the examination's destinations. Sentences of differing lengths were utilized to expand variety, while the possibility of utilization of factual and demonstrating strategies added intricacy.

PRELIMINARY ANALYSIS

We began our preliminary analysis of the link between film budget and profits by inspecting the Pearson correlation coefficient. Visualization, through histograms and scatter plots, to grasp the distribution and connection between costs and earnings, helped us attain a feel for the information.



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.

```
> # 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.

FINAL ANALYSIS

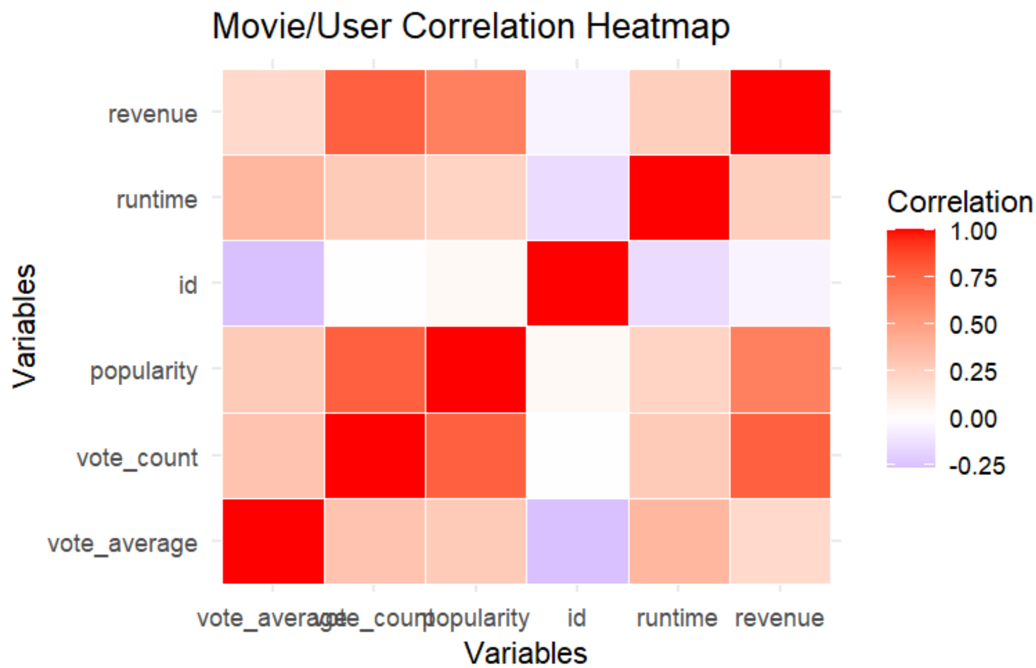
Question 1: Correlation between User Interactions and Movie Features

The first 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. Another strong relationship among the matrix is the ‘popularity’ and’vote_count’ variables which could indicate that the number of votes for a movie can have an immense effect on its popularity score. 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. The darker shades of red reflect the strong correlations mentioned in the matrix.



To understand the relationships between two variables with strong positive linear correlations, a pair of scatterplots are built for visual representation of the interactions. To avoid the issue of over dense plotting and cluttering with a large data set, a sample was taken and split into training and testing sets to assess the performance and generalization ability of the model. The 'id' variable was eliminated from the previous dataset because of its irrelevancy toward the other variables. A linear regression model was then used to analyze the coefficients of each independent variable.

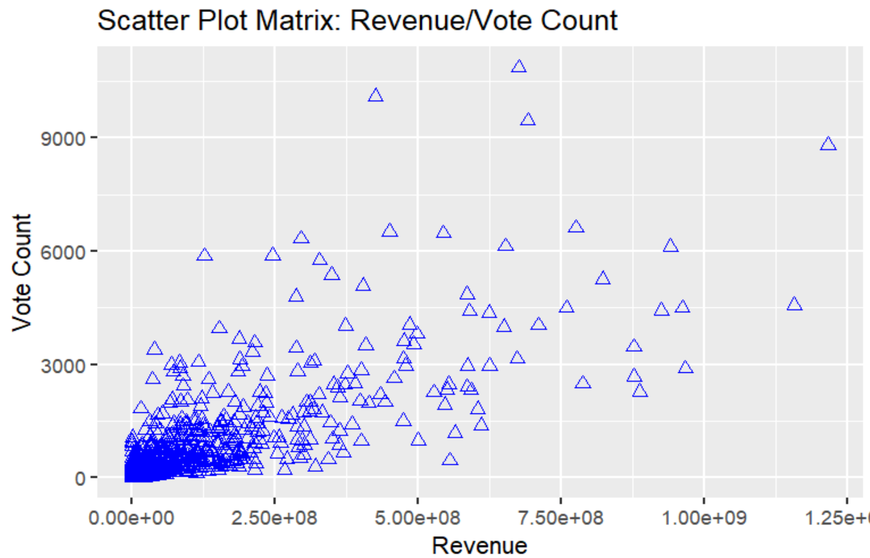
```
Call:
lm(formula = vote_count ~ ., data = training_set)

Coefficients:
 (Intercept)  vote_average    popularity       runtime       revenue
  -5.485e+02     7.233e+01     2.421e+01     4.421e-01     2.955e-06
```

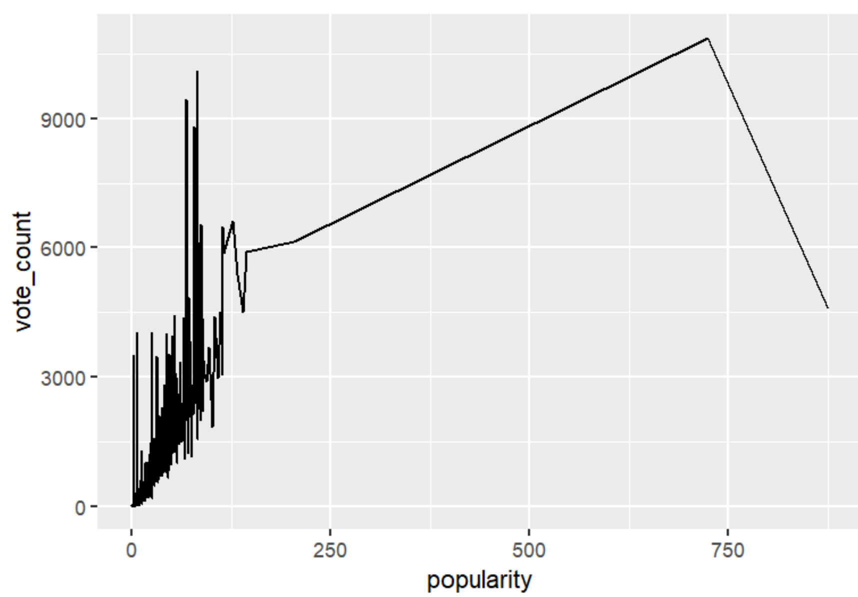
Positive coefficients indicate a positive association, and negative coefficients suggest a negative association, and the magnitude of the coefficients indicates the strength of the effect. All four variables show a positive correlation to the vote counts of the movies in the data set. A k-fold cross validation is then conducted to assess how well the model will popularize to an independent data set.

```
Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-5923.5  -175.9   -36.2    75.1  7152.6 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -5.485e+02  5.533e+01  -9.912   <2e-16 ***
vote_average  7.233e+01  8.372e+00   8.640   <2e-16 ***
popularity    2.421e+01  4.487e-01  53.955   <2e-16 ***
runtime       4.421e-01  4.416e-01   1.001    0.317    
revenue       2.955e-06  7.689e-08  38.431   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 559.4 on 3837 degrees of freedom
Multiple R-squared:  0.798,	Adjusted R-squared:  0.7978 
F-statistic:  3790 on 4 and 3837 DF,  p-value: < 2.2e-16
```

An analysis of the summary shows that the ‘runtime variable’ has a higher p-value (0.317), suggesting that it may not be a statistically significant predictor in the model. The F-statistic that shows the overall significance of the model is large, and the associated p-value is very small (< 2.2e-16), indicating that the overall model is statistically significant. Overall, the model is well-fitted with statistically significant predictors.

The models of the samples conducted are plotted to show the relationships between the response variable ‘vote_count’ and the independent variables ‘revenue’ and ‘popularity’ which showed the strongest correlations in the correlation matrix. Most of the observations in the scatterplot are under 3000 votes and $250 million in revenue. The outliers in the plot are heading in a positive direction, indicating that the higher the revenue for a movie, the higher the vote count for that movie. The analysis of the line plot between the variables “vote_counts” and 'popularity' is similar with the popularity score being high when the vote count for the movie is also high. The few outliers in the plot also demonstrate the same patterns.



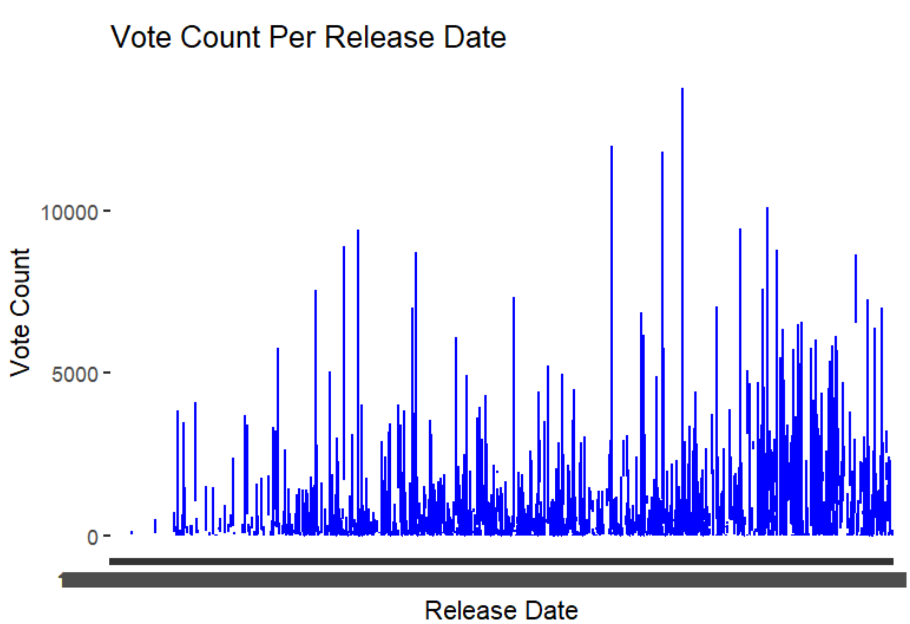


Additional categorical variables that were explored using Analysis of Variance (ANOVA) testing were release dates and whether there were significant differences between them. To again alleviate the concern of over dense plotting and cluttering with the data set, a sample was taken and split into training and testing sets to assess the performance and generalization ability of the model.

```
> summary(anova_model)
               Df    Sum Sq Mean Sq F value  Pr(>F)    
release_date 3279 5.487e+09 1673498   1.368 2.2e-12 ***
Residuals    1485 1.817e+09 1223626                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Examining the ANOVA model, the p-value is very small (2.2e-12), suggesting that there is a significant association between ‘release_date’ and the response variable. The table also suggests that the variable ‘release_date’ is statistically significant in explaining the variability in the voting count totals.

The line plot is created to show this relationship and the analysis shows that movies released more recently have slightly higher voting counts than older movies. The top movies by vote count are all more recent, but because of the plot not showing a strong association might mean this will need to be evaluated more to see if this relationship is strong despite the low p-value in the ANOVA model.



Question 2: Correlation between Movie Budget and Popularity

I calculated the Pearson correlation coefficient in R to assess the relationship between a movie's budget and its popularity. The correlation coefficient obtained was r = 0.505414, indicating a moderate positive correlation between budget and popularity. This suggests that as the budget of a movie increases, its popularity tends to increase as well, although the strength of this relationship is not extremely strong.

```
> # Print the correlation coefficient
> correlation
[1] 0.505414
```

This does not conclusively prove that higher production costs cause greater box office returns or viewership tallies. But it does hint at a pattern where blockbusters with nine-figure budgets stand an improved chance of resonating with target demographics when compared to more frugally crafted independents. Of course, other considerations like celebrity status, genre, timing, or luck also dictate a movie's commercial success or failure to various extents.

By utilizing R, I fitted a linear regression model to further explore the relationship between a film's budget and its appeal. The lm() function created the formulation, defining popularity as the dependent variable and budget as the independent variable. The summary() operation furnished thorough specifics of the regression, like coefficients, typical mistakes, t-stats, and p-values.

This probing permit evaluating the budget's influence on popularity and judging if their linkage is statistically meaningful. Moreover, visualizing the information through plots provided extra insights.

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

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

Residuals:
   Min     1Q Median     3Q    Max 
-89.09  -9.85  -5.22   4.86 836.34 

Coefficients:
                  Estimate    Std. Error t value            Pr(>|t|)    
(Intercept) 10.02290799336  0.48664425191   20.60 <0.0000000000000002 ***
budget       0.00000039488  0.00000000973   40.59 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.46 on 4801 degrees of freedom
Multiple R-squared:  0.2554,	Adjusted R-squared:  0.2553 
F-statistic:  1647 on 1 and 4801 DF,  p-value: < 0.00000000000000022
```

The results of our regression analysis revealed a statistically significant positive association between a movie's budget and its popularity ($\beta = 0.00000039488$, $p < 0.0000000000000002$). These results indicate movies produced with a higher budget tend to be more successful in terms of popularity among audiences. The coefficient estimate on the budget variable suggests that for each additional unit of currency spent on the budget of the movie, the popularity of the movie typically increases by approximately 0.00000039488 units.

The relationship between movie budget and popularity is depicted via the scatter plot I created using the sampled data. The scatter plot clearly shows the positive trend between the two variables. Each point on the plot represents a movie from the dataset. The x-axis represents the budget (in millions) of the movie, and the y-axis represents the popularity score. Additionally, as you can see, the scatter plot includes an overlaid linear regression line, which visually depicts the trend.

The scatter plot solidifies the findings from the regression analysis. The strong positive relationship between budget and popularity is clear from the scatter plot. As movie budgets increase, popularity tends to increase with it, and this visualization reinforces our understanding of how these variables are related. It also helps corroborate what we had concluded from the regression analysis.

In R, I obtained the predicted popularity values for the first few movies in the dataset using the regression model. The predicted popularity values are nothing but the estimated popularity scores for these movies based on their respective budgets.

```
> # Generate predictions using the regression model
> predictions <- predict(lm_model, newdata = movie_dataset)
> # Print the first few predicted popularity values
> head(predictions)
       1        2        3        4        5        6 
103.6102 128.4878 106.7692 108.7437 112.6925 111.9027 
```

Insights from these predictions can be invaluable for movie budgeting strategies. They enable filmmakers and industry stakeholders to anticipate the popularity of their films well ahead of the release. Such anticipation serves as a great tool to strategically allocate budgets to those projects that are expected to gain higher popularity, thereby optimizing their investment decisions, and increasing the odds for success in the highly competitive movie industry.

CONCLUSION

The findings demonstrated significant associations between movie features and user interactions, particularly variables such as revenue, popularity, and vote count. The regression analysis revealed a statistically significant positive association between a movie's budget and the number of user votes.

Findings from this study extend the understanding of the underlying components of movie success and inform decision-making in the movie industry. By recognizing these relationships, stakeholders can make informed decisions about budget strategies, marketing campaigns, and the creation of content that may improve the likelihood of movie success and prosperity.'

Moving forward, we will enhance our analysis by accounting for more variables, such as genre, release date, and promotional spending, to get deeper insights into the ingredients for box office magic. Sophisticated statistical methods such as machine learning algorithms and time series examination could furnish predictive power and uncover intricate patterns within the data.

REFERENCES

Bibin, T. B. (n.d.). Movie Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/bibintb/movie-dataset?select=movie_dataset.csv

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)

library(caret)

 #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

dataset <- movie_dataset[complete.cases(movie_dataset$runtime), ]

 #Display the modified dataset

summary(dataset)

 #Perform correlation analysis

correlation_matrix <- cor(dataset[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)
 #Create a new variable from dataset for sampling
movies <- dataset %>%
  select(vote_average, vote_count, popularity, runtime, revenue)
summary(movies)
 # Split the dataset into training and testing sets
set.seed(456)
trainIndex <- createDataPartition(movies$vote_count, p = 0.8, list = FALSE)
```

```
training_set <- movies[trainIndex, ]

testing_set <- movies[-trainIndex, ]

 # Create a linear regression model

model <- lm(vote_count ~ ., data = training_set)

model

 # Make predictions on the test set

predictions <- predict(model, newdata = testing_set)

 # Evaluate the model

rmse <- sqrt(mean((predictions - testing_set[["vote_count"]])^2))

print(paste("Root Mean Squared Error:", rmse))

 # Perform k-fold cross-validation

ctrl <- trainControl(method = "cv", number = 10)

cv_model <- train(vote_count ~ ., data = training_set, method = "lm", trControl = ctrl)

 # Print cross-validation results

print(cv_model)

 # Access the mean performance metrics across folds

summary(cv_model)

 #Individual Scatter Plots Showing Relationship Between Vote Count and Revenue.
```

```
scatter_plot_matrix <- ggplot(testing_set, aes(x = revenue, y = vote_count)) +
  geom_point(shape = 2, color = "blue", size = 2) +
  labs(title = "Scatter Plot Matrix: Revenue/Vote Count",
       x = "Revenue",
       y = "Vote Count")
scatter_plot_matrix
 #Individual Line Plots Showing Relationship Between Vote Count and Popularity.
ggplot(testing_set, aes(x = popularity, y = vote_count)) +
  geom_line()
 #Perform ANOVA testing
#Create a new variable from dataset for testing
anovas <- dataset %>%
  select(vote_count, release_date)
summary(anovas)
 #Split dataset into a training set and a test set.
set.seed(123)
index <- createDataPartition(anovas$release_date, p = 0.8, list = FALSE)
train_data <- anovas[index, ]
```

```
test_data <- anovas[-index, ]

 #Create ANOVA model

anova_model <- aov(vote_count ~ release_date, data = train_data)

summary(anova_model)

 #Conduct post-hoc tests

TukeyHSD(anova_model)

 #Lineplot for one-way ANOVA

ggplot(train_data, aes(x = release_date, y = vote_count)) +

  geom_line(color = "blue") +

  labs(title = "Vote Count Per Release Date",

       x = "Release Date",

       y = "Vote Count")

# Appendix: R Code

# Load necessary libraries

library(ggplot2)

library(dplyr)



# Load the dataset
```

```
movie_dataset <- read.csv("/Users/m.joubert/Documents/Final Project - Draft - Joubert/movie_dataset.csv")

# Data Exploration

# Summary statistics

summary(movie_dataset)

# Histogram of movie budgets

ggplot(movie_dataset, aes(x = budget)) +
  geom_histogram(binwidth = 10, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Movie Budgets",
       x = "Budget (in millions)",
       y = "Frequency")

# Scatter plot of budget vs. popularity

ggplot(movie_dataset, aes(x = budget, y = popularity)) +
  geom_point(alpha = 0.6, color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Scatter Plot of Movie Budget vs. Popularity",
       x = "Budget (in millions)",
       y = "Popularity")

# Linear regression model
```

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

summary(lm_model)

# Predictions

predictions <- predict(lm_model, movie_dataset)

head(predictions)

# Save the plot as a PDF file

ggsave("scatter_plot_budget_vs_popularity.pdf", width = 10, height = 7)
```