Introduction to Data Science

Project #1: Impact of Variables on Profits for Dramatic Theatrical Films

Statistical Analysis in R

Tammy Walker, 2020

**Contents**

[Introduction 3](#_Toc52847976)

[Methods 3](#_Toc52847977)

[Variables 3](#_Toc52847978)

[EDA and Simple Statistical Analysis Techniques 3](#_Toc52847979)

[Exploratory and Expository Data Visualizations 5](#_Toc52847980)

[Plots of One Variable 5](#_Toc52847981)

[Plots of Two Variables 11](#_Toc52847982)

[Predictions and Linear Regression 13](#_Toc52847983)

[Logistic Regression Model 15](#_Toc52847984)

[Conclusion 17](#_Toc52847985)

[Appendix 18](#_Toc52847986)

# Introduction

The main goal of the project is to understand the variables that impact the gross box office revenues for dramatic films released from 2010 to 2019 and to predict the profits of a film given a new value for one of the predictors.

## Methods

This project will use data sets from IMDB and box office data sets on Kaggle and will use the R language for statistical analysis. The project will involve acquiring data, cleaning data, using various exploratory data analysis techniques, plotting, and machine learning.

## Variables

The IMDB data set includes 81,273 observations that represent movie titles and its associated data, such as the title, box office, budget, average review, and number of critics reviews. There were a total of five data sets: movies, ratings, principals, names, and boxoffice. The final set of variables after joining tables and subsetting include the following with 17 variables and 153 observations:

> names(dramaticFilms)

[1] "title" "genre"

[3] "name" "director"

[5] "year" "profitMargin"

[7] "lifetime\_gross" "budget"

[9] "reviews\_from\_critics" "votes"

[11] "avg\_vote" "females\_45age\_avg\_vote"

[13] "males\_45age\_avg\_vote" "females\_30age\_avg\_vote"

[15] "males\_30age\_avg\_vote" "females\_18age\_avg\_vote"

[17] "males\_18age\_avg\_vote"

After joining the data sets, the data was further subsetted to evaluate films that had the following criteria:

* an average vote of 7 or more on a scale of 1 to 10
* lifetime gross projects of more than $1M
* genre was drama or a combination of drama, comedy, biography, and history
* domestic films in the USA
* released between 2010 and 2019

## EDA and Simple Statistical Analysis Techniques

For the main goal of the project in regards to box office profits, the scatterplot and matrix below shows that there is a moderately strong correlation between the lifetime gross profits and the following: 1) the budget (66%) and 2) average vote from females ages 45 and above (62%). There is a moderate correlation between the lifetime gross profits and the following: 1) average vote from females ages 30 to 44 (58%) and 2) average vote from females ages 18 to 29 (55%).

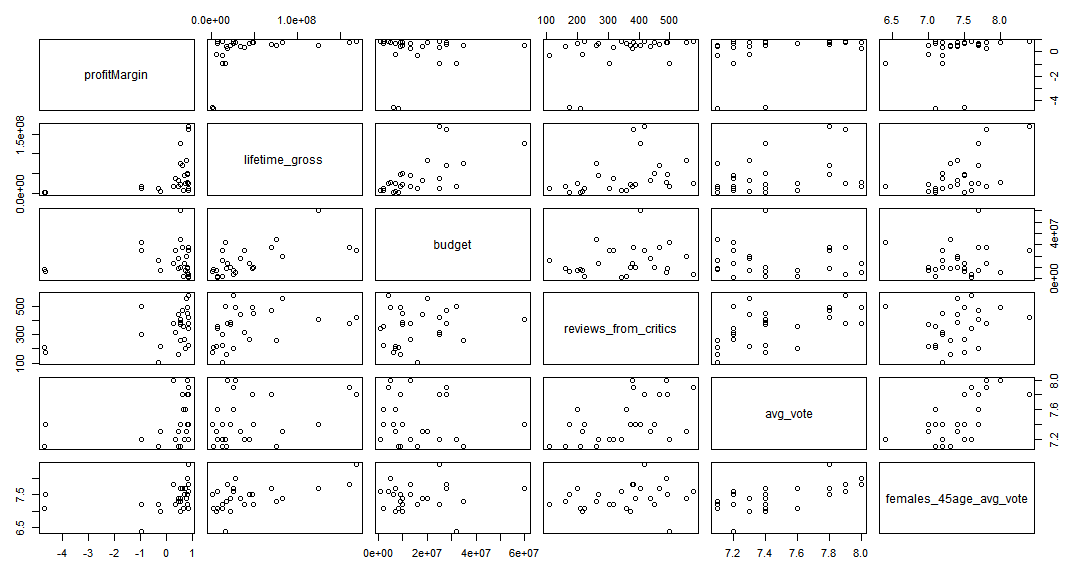
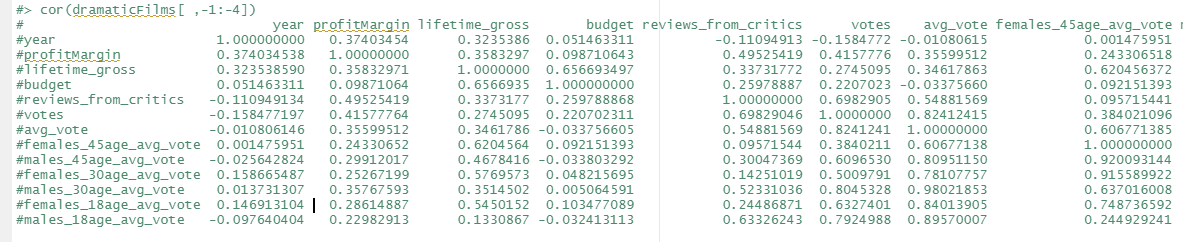


Figure : Scatterplot for the quantitative variables in the dramatic films data set

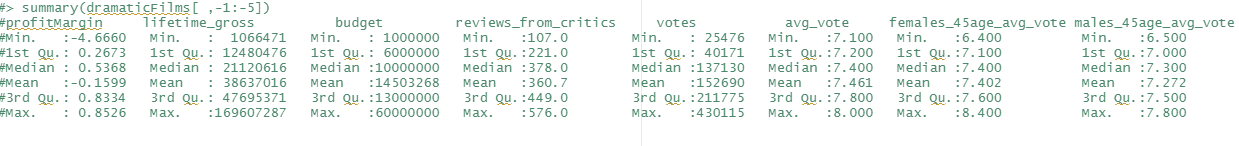
In addition, there is a moderate correlation between the profit margin and number of reviews from critics (50%). The weakest correlation for profit margin are males ages 18 to 29 (22%).

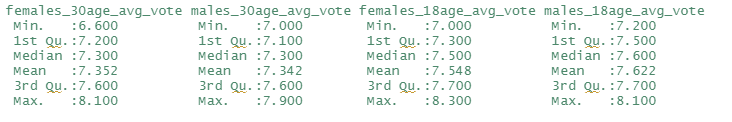
The weakest correlation for the life time gross profits are average votes from males ages 18 to 29 (13%). It is interesting to note that the number of reviews from critics has a strong correlation to the number of votes (70%). The correction between reviews from critics and males is moderately strong with the average vote of males ages 18 to 29 (63%), moderate for #males 30 to 44 (52%), and weak for males 45 and above (30%).

There is a strong correlation between the number of votes and the average vote (82%). The other correlations are not a factor of lifetime gross revenues and will not be evaluated for this matrix.



The observations below show that the mean profit margin of dramatic films with an average vote of 7 or above and that were released between 2010 and 2019 had a negative 15% profit margin; however, the median profit margin was 53%. The mean budget was $14.5M and the median budget was $10M for dramatic films. Females had a higher average vote than males, and younger males had a higher average vote than older males. Per the correlation of 63% for the average vote of males 18 to 29 and number of reviews from critics, there is a possibility that younger males had a higher average vote because they were influenced by reading reviews from critics.





## Exploratory and Expository Data Visualizations

### Plots of One Variable

The budget had a 66% correlation to the lifetime gross profits. The barplot below shows that 50% of the films in the data set (70 films of the 153 films) had a budget between $5M and $10M. The range was $1M to $60M+.

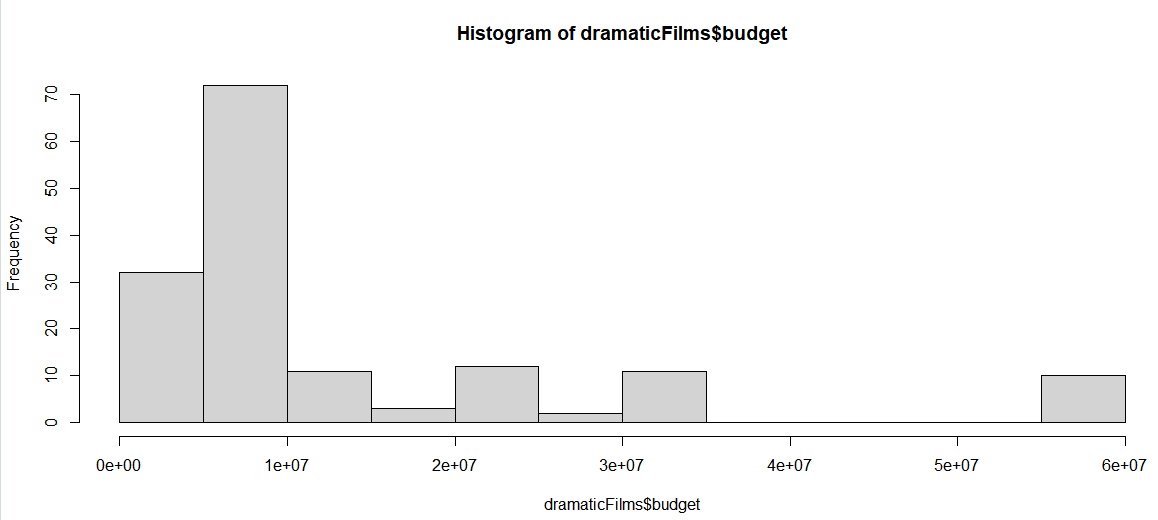


Figure : Histogram shows the frequency of values for the budget variable

The boxplot below shows that the median budget is $10M and the median budget is $14.5M with the 1st quartile at $6M and 3rd quartiles at $13M.

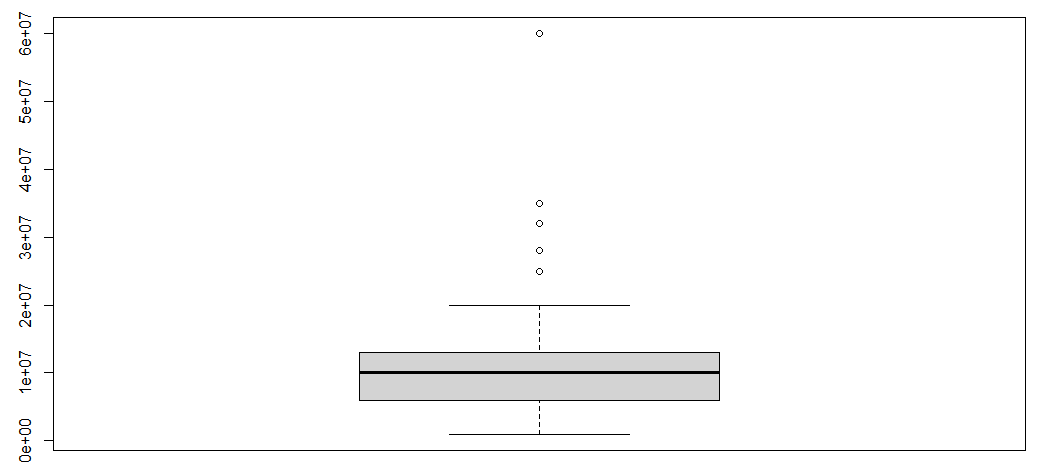


Figure : Boxplot showing the values of the budget variable

The barplot below shows the largest distribution of the budget in the range of $8M to $16M.

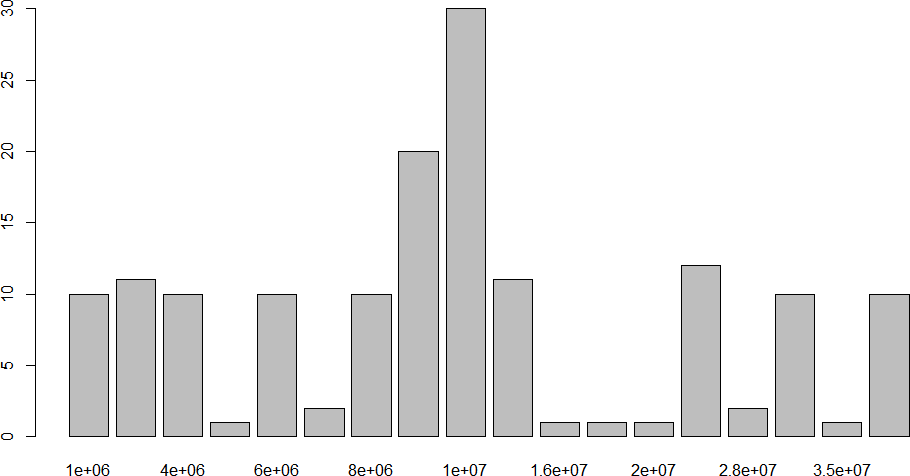
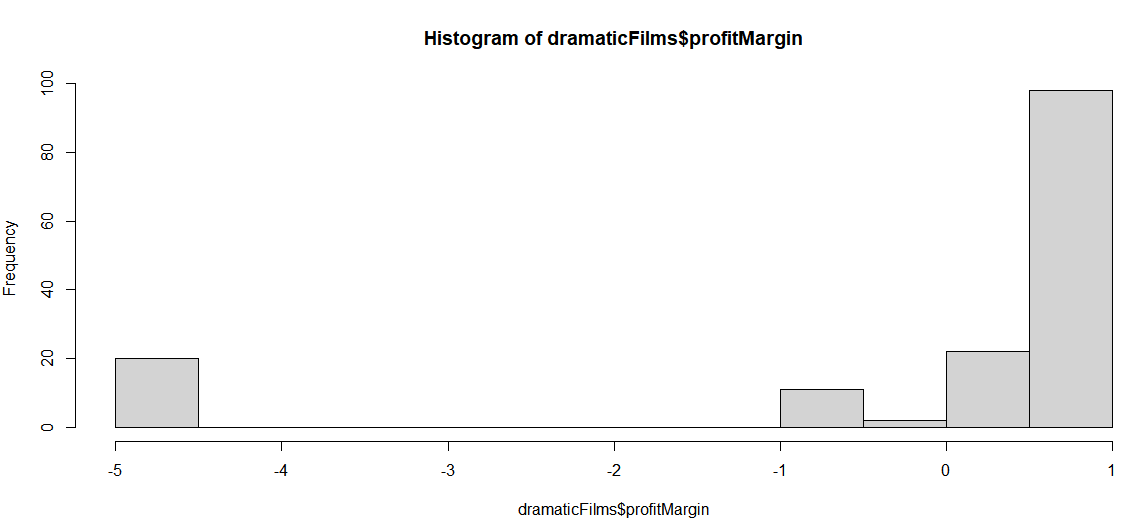


Figure : Barplot showing the distribution of values by number of observations for the budget variable

The barplot below shows that of the 153 films in the data set, about 30 films had a negative profit margin, 20 films had a profit below 50%, and 100 films had a profit margin above 50%.



The smoothing line in the barchart below shows the greatest density of the profit margin at approximately 80%.

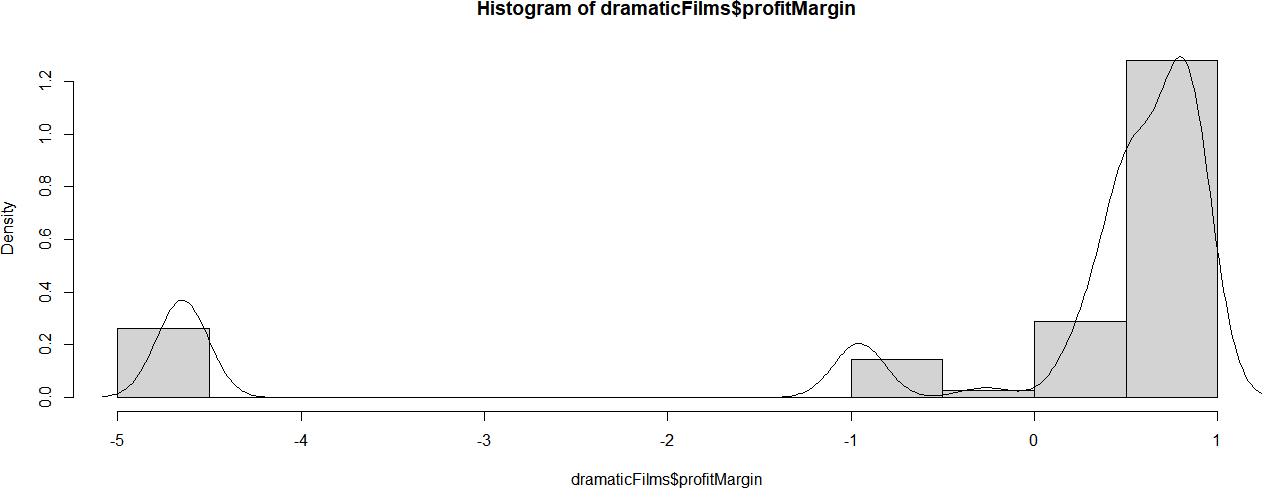


Figure : Histogram with smoothing line shows the density of values for the profitMargin variable

The boxplot below shows that most of the films had a profit margin above 50% with a few outliers far below the minimum boundary.

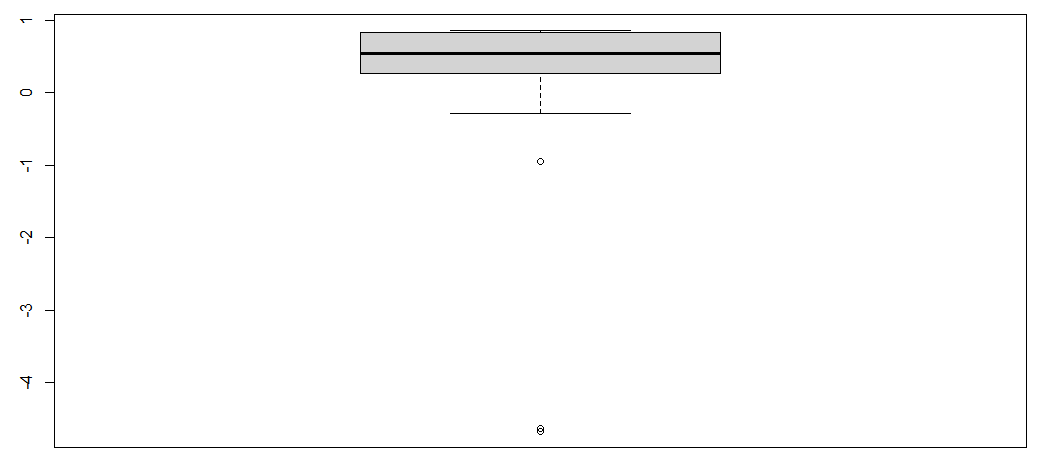


Figure : Boxplot of the profitMargin variable

The barchart below shows the frequency of the profit margins with more than 40 of 153 (30%) films in the range of approximately 52% . Also, there are approximately 30% with a profit margin above 80% and about 20% of the films that had a negative profit margin.

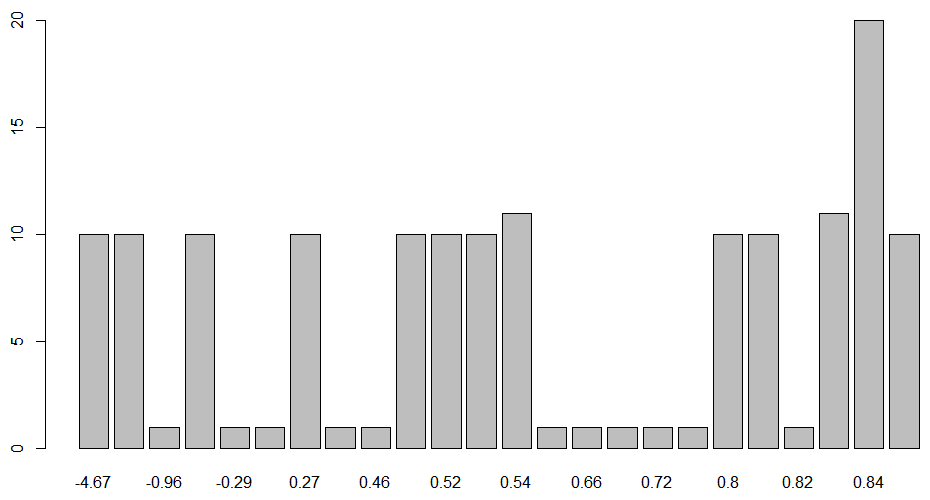


Figure : Barplot showing the distribution of values by number of observations for the profitMargin variable

The barplot below shows that about 120 of the 153 (80%) of films had a life time gross of less than $50M with 70 of the 153 (50%) grossing $12M or less. There are about 50 films (30%) that grossed between $12.5 and $50M.

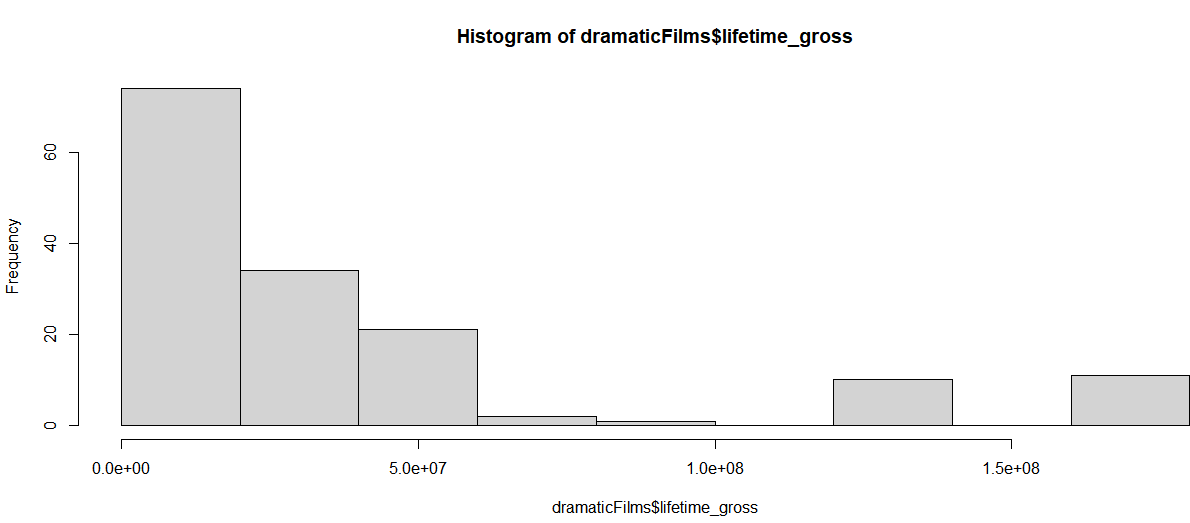


Figure : Histogram shows the frequency of values for the lifetime\_gross variable

The barchart below breaks down the midrange showing that approximately 45% of films grossed between $16M and $25M.

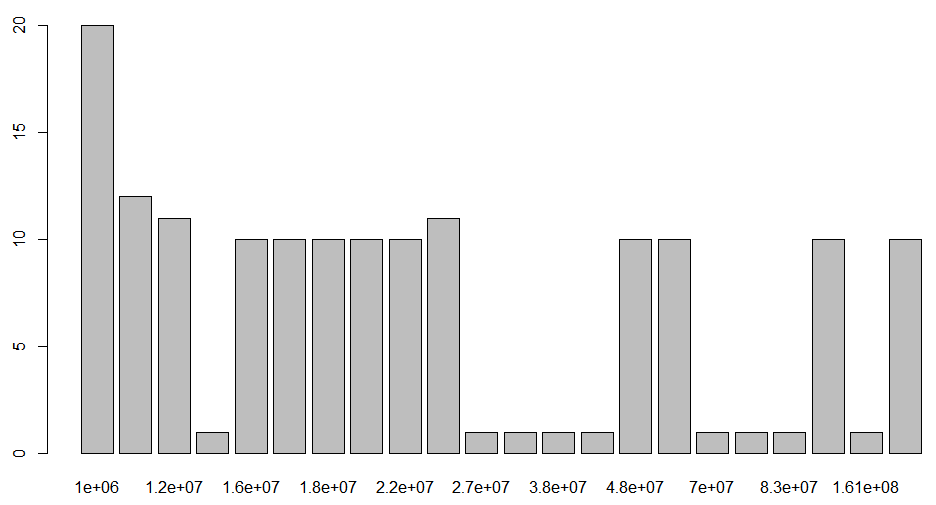


Figure : Barplot showing the distribution of values by number of observations for the lifetime\_gross variable

The boxplot below represents the mean gross in the summary of this variable, which has a mean of $39M lifetime gross and median of$21M lifetime gross.

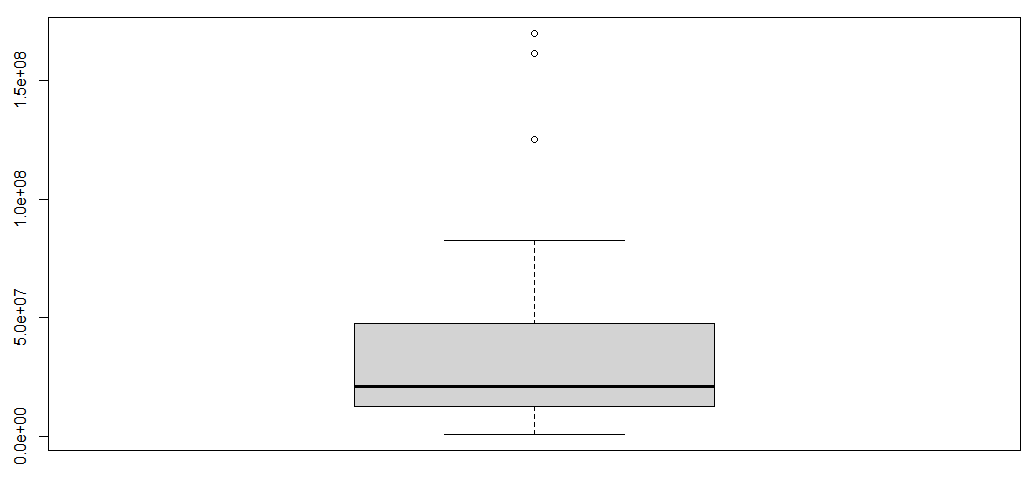


Figure : Boxplot of lifetime\_ gross variable

Reviews from critics had a moderate correlation of 50% to the profit margin, but only a weak 33% correlation to the lifetime gross. Approximately 70% of the films were reviewed by critics 300 to 500 times.

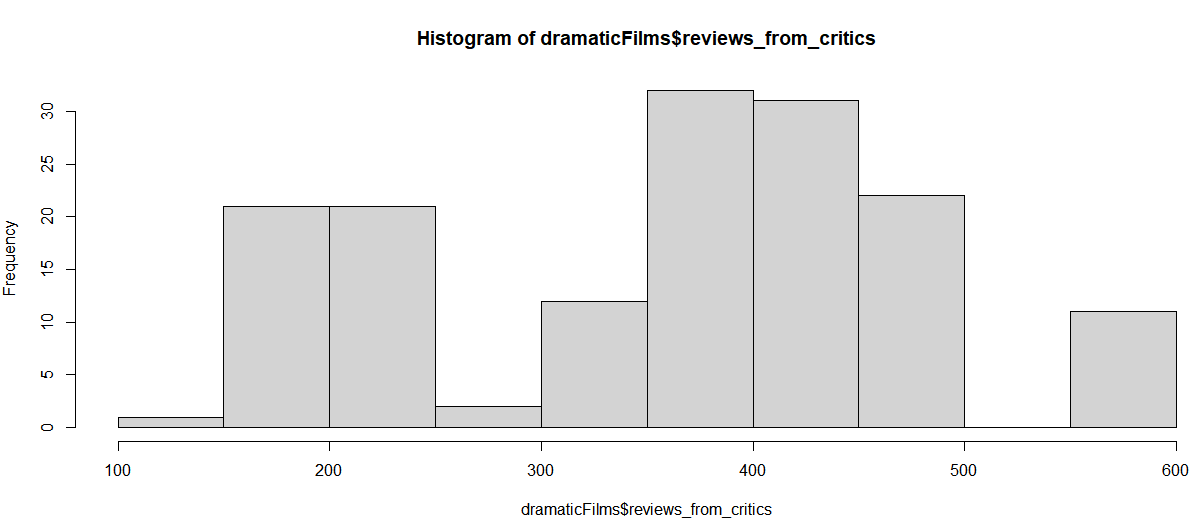


Figure : Histogram shows the frequency of values for the number of reviews\_from\_critics variable

The number of votes had a weak correlation of 27% to the lifetime gross and moderately weak correlation of 42% to the profit margin. The barplot below illustrates that 70% of films with a mean average rating more the 7 had more than 100K votes. Approximately 50% of the films had 100K to 250K votes.

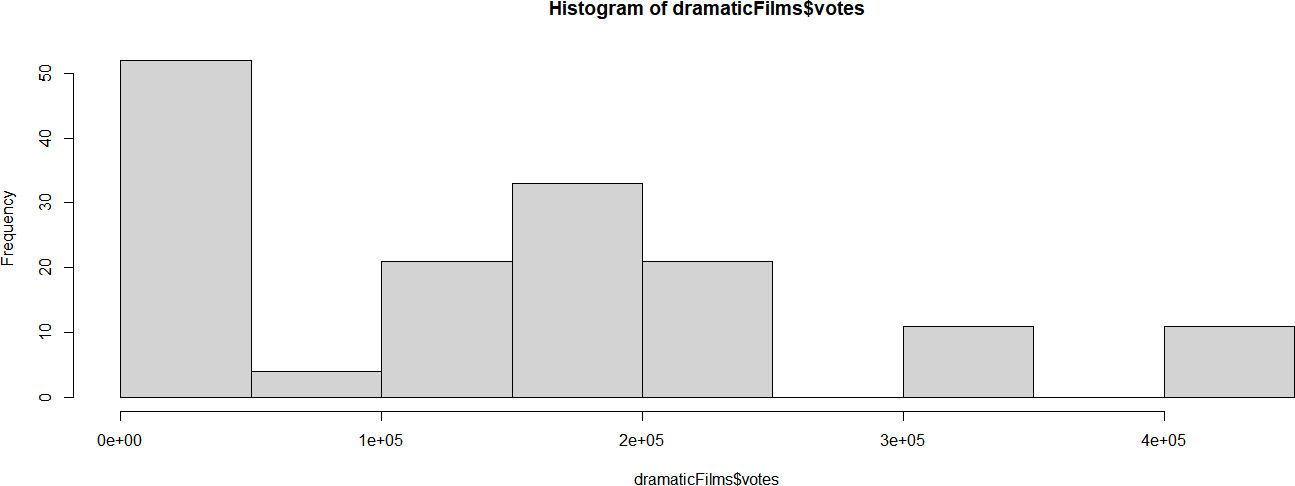


Figure : Histogram shows the frequency of values for the votes variable

### Plots of Two Variables

The scatterplot below shows that the film's budget did not show much of an increase or decrease in a films profit margin.

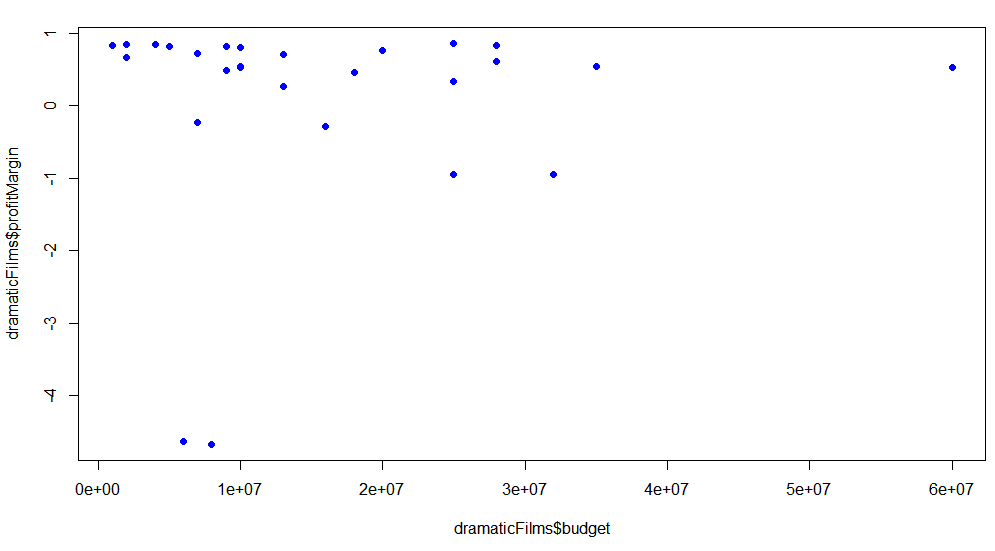


Figure : Scatterplot comparing the profitMargin and budget variables

The scatterplot below shows there is a moderate correlation between the budget and lifetime gross profits. As the budget increases, the lifetime gross profits moderately increase.

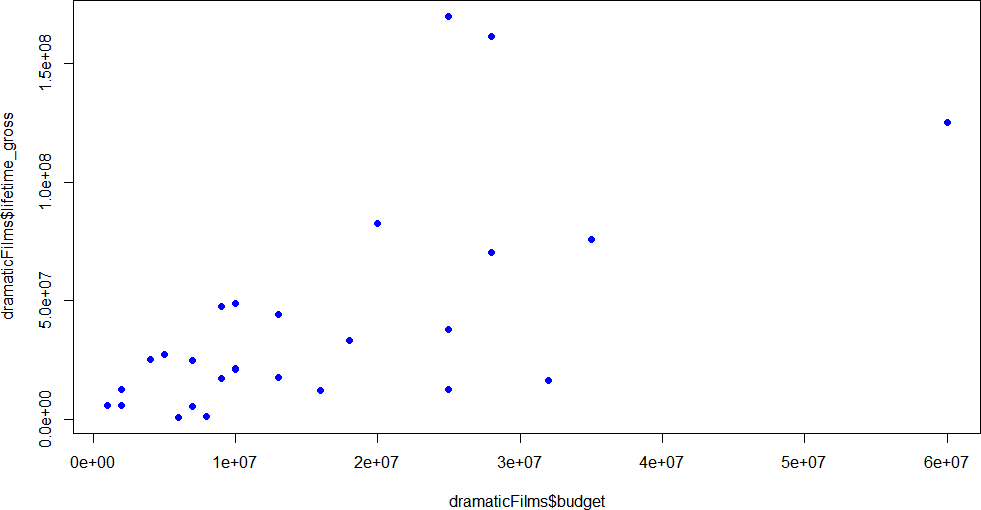


Figure : Scatterplot comparing the lifetime\_gross and budget variables

The scatterplots below show that as the budget increases, the film's gross profits increase and as the average vote of females 45 and above increases, the lifetime gross profits increased. Less so for females between 30 and 44, and even less so for females between 18 and 29.

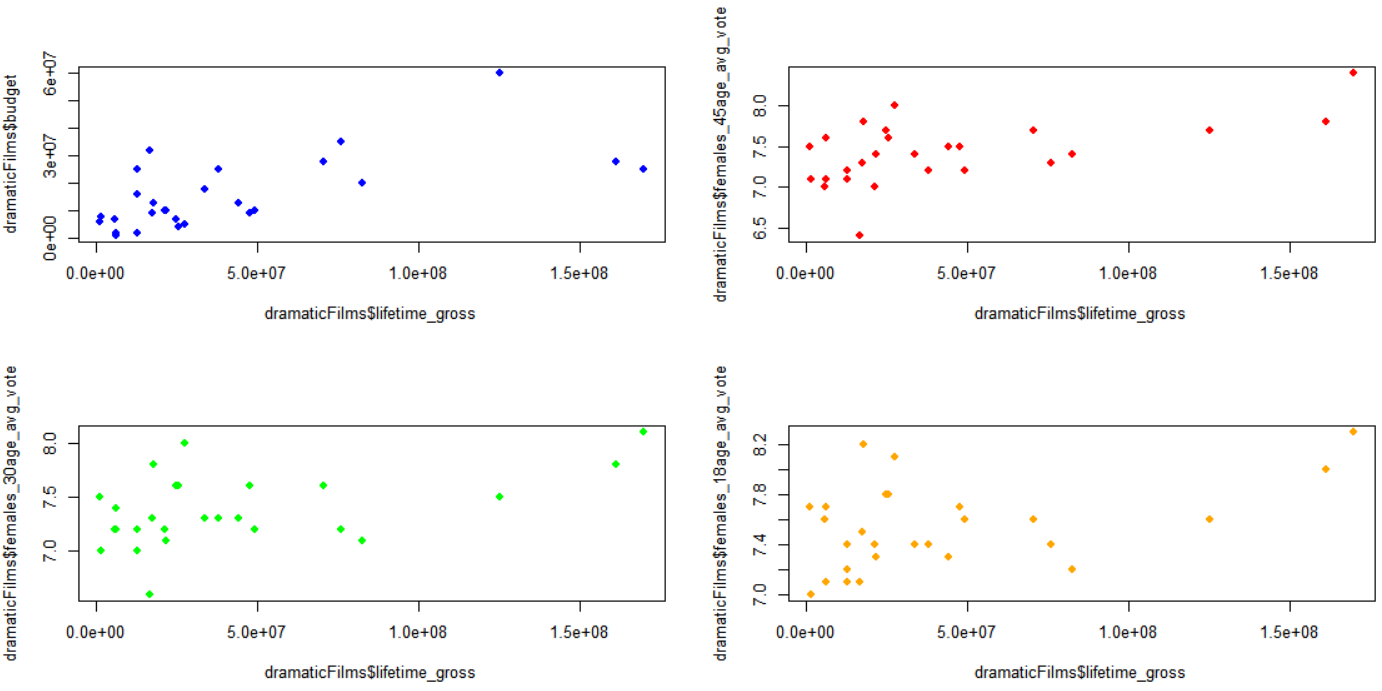
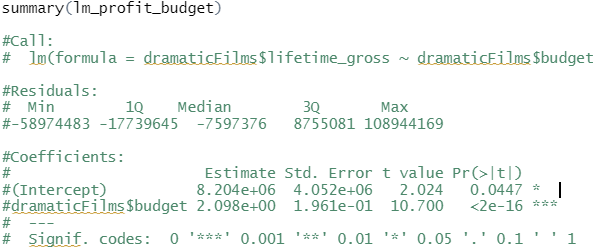


Figure : Comparison of variables that impact lifetime gross profits

The focus of the linear regression will be to compare the impact of the budget and other variables above on the lifetime gross profits.

## Predictions and Linear Regression

In using the lm() function to perform a simple linear regression with lifetime gross profits as the response variable and budget as the predictor, the following summary is generated and shows the budget has a strong impact on the lifetime gross.



* Residual standard error is 35700000 on 151 degrees of freedom
* Multiple R-squared is 0.4312,Adjusted R-squared is 0.4275
* F-statistic is 114.5 on 1 and 151 DF, p-value: < 2.2e-16
* Estimate: The value of lifetime\_gross at budget = 0 is 8.204e+06, this is a positive relationship
* Std. Error: The standard error for the budget variable is 1.961e-01, which is 10 times less (in absolute value), than the the estimate of the coefficient 2.098e+00. At least one order of magnitude less is better.
* Pr(>|t|): The p-value for the t-test and is the probability the variable is not relevant, so it should be small and is confirmed by its value in scientific notation of 2e-16 (2 x 10^16)
* Adjusted R-squared: The model explained 43 percent of the cause of lifetime gross is due to the budget. As indicated by the three asterisks in the table, budget has a strong impact on lifetime gross.
* As indicated by the three asterisks in the table, budget has a strong impact on lifetime gross.

The result is the simple linear regression model predicts that the lifetime gross profit for a dramatic film budgeted at $15M will be $11.3M.

Below is a plot of the response variable and predictor, using the abline() function to display the least square regression line. The plot shows data points loosely around the line, which is good. A pattern may indicate the fit is not good.

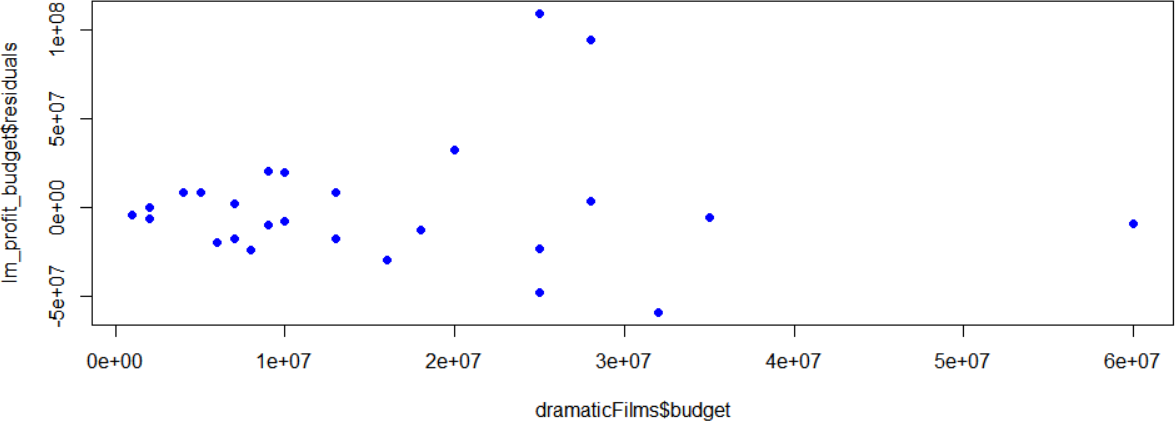
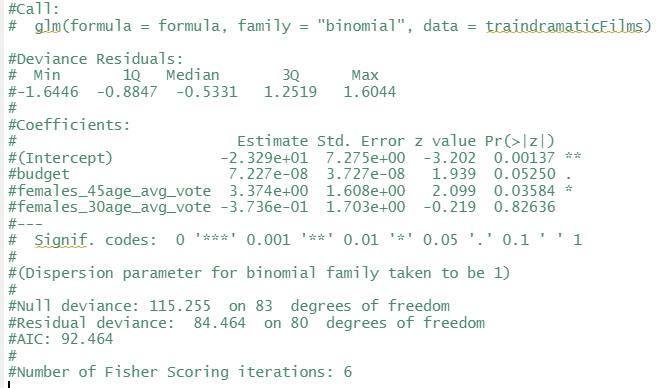
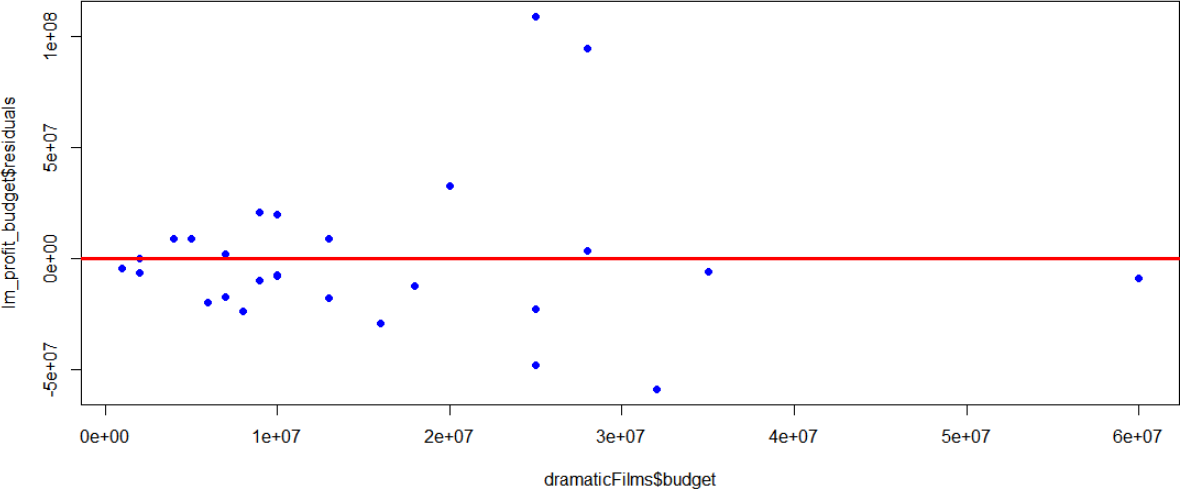


Figure : Plot of the response lm\_profit\_budget variable and predictor budget

In the table below, the smallest p-value 0.03584 is associated with females\_45age\_avg\_vote, meaning there is an association between females\_45age\_avg\_vote and lifetime\_gross01. The standard error for the females\_45age\_avg\_vote variable is 1.608e+00, which is 2.1 times less than the estimate of the coefficient 3.374, a good sign the model is fit.



Below is a plot of the response variable and predictor, using the abline() function to display the least square regression line.



The plot shows a diagnostic plot of the fit, displaying the Residuals vs. Fitted plot with equally distributed values.

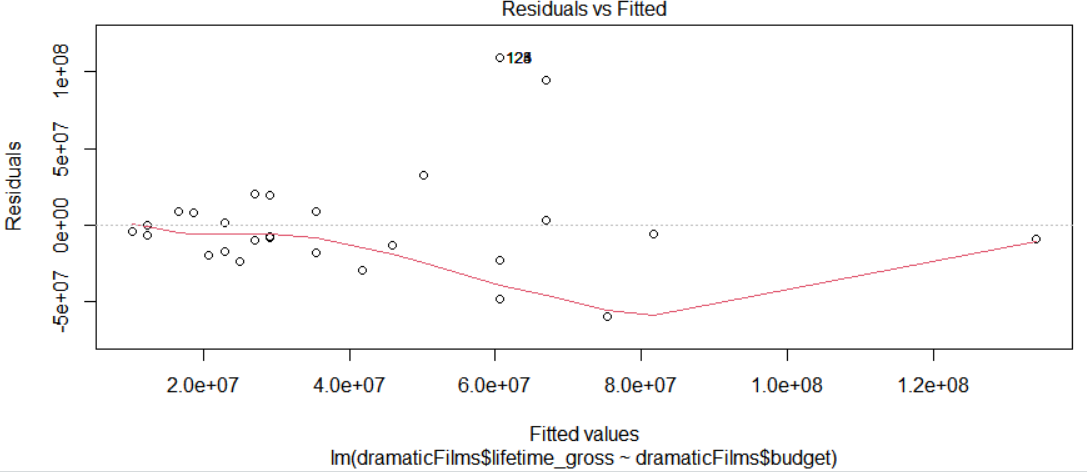


Figure : Residuals vs Fitted plot with equally distributed value

### Logistic Regression Model

The purpose of the scatterplot below is to confirm a good distribution. There appears to be a few clusters.

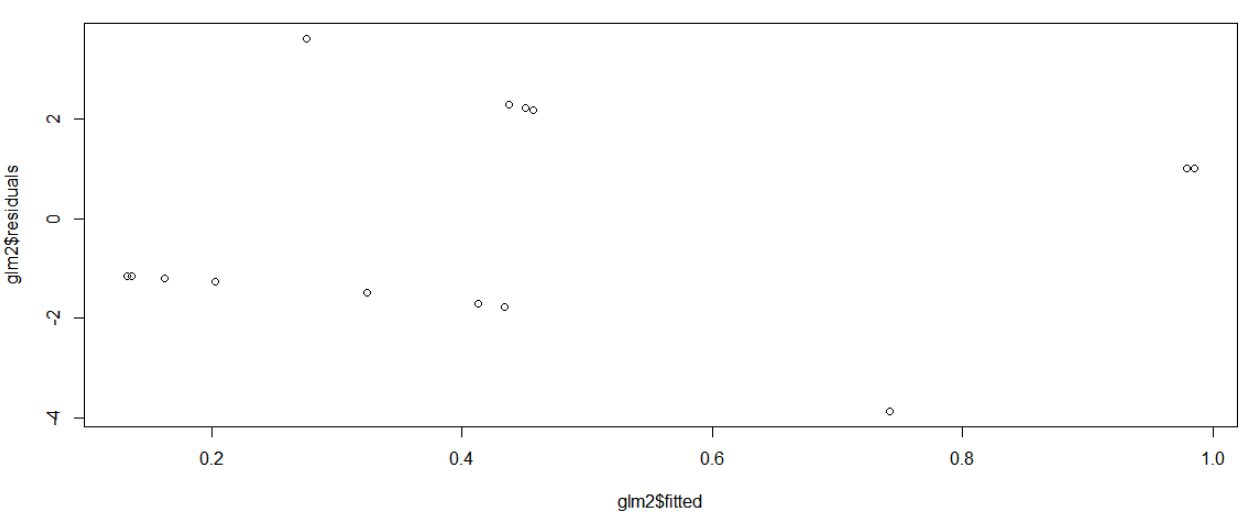


Figure : Plot purpose is to confirm a good distribution

The purpose of the plot below is to show no trend. There does not appear to be a trend.

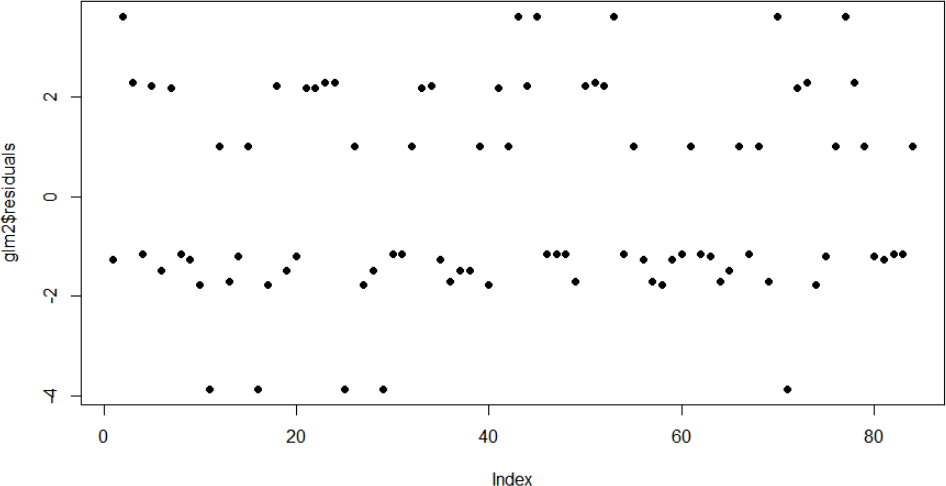


Figure : Plot purpose is to show no trend

The test error metric for the logistic regression model is 0.375 or approximately 37.5%.

# Conclusion

This project explored various statistical analysis and EDA techniques to understand the data, plot the most significant variables, and apply machine learning algorithms to predict the lifetime gross profits of a dramatic film.

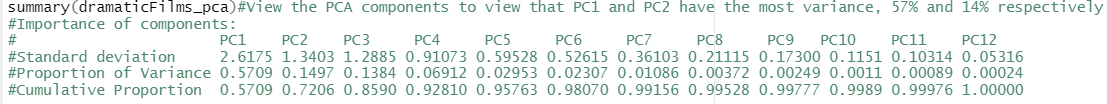
It was evident that the budget had the strongest impact on the gross profits. A qualitative study outside of the scope of this project may examine further how the budget impacts the gross profits, which may be marketing funds, high-budget lead actors, and overall higher production values among other variables. If this project were for the purpose of marketing analysis, the key target segment would be females for dramatic films, particularly for females ages 45 or above. It would be important to have good reviews from critics to obtain a higher profit margin and there should be social media as a part of a marketing campaign to allow moviegoers to vote and rate a film to attract more viewers.

In addition, a production would need to plan a minimum budget of $10 to $15M to be competitive with other dramatic films that generate a higher lifetime gross profit. Finally, the 37% error rate is somewhat high for the model; however, films unlike other consumer goods are somewhat unpredictable in that a lower budget movie may go viral and make millions while a high budget film may not perform well at the box office.

A future iteration of this project may expand on demographics and other data sets to form a more comprehensive view of the film market and models that may predict how well a film performs at the box office.

# Appendix

The purpose of the Principal Component Analysis (PCA) to examine whether to reduce the dimensions and remove variable from the data set. The variables were reduced from 12 to 3 in this exercise.



Step 5.2.3 PCA

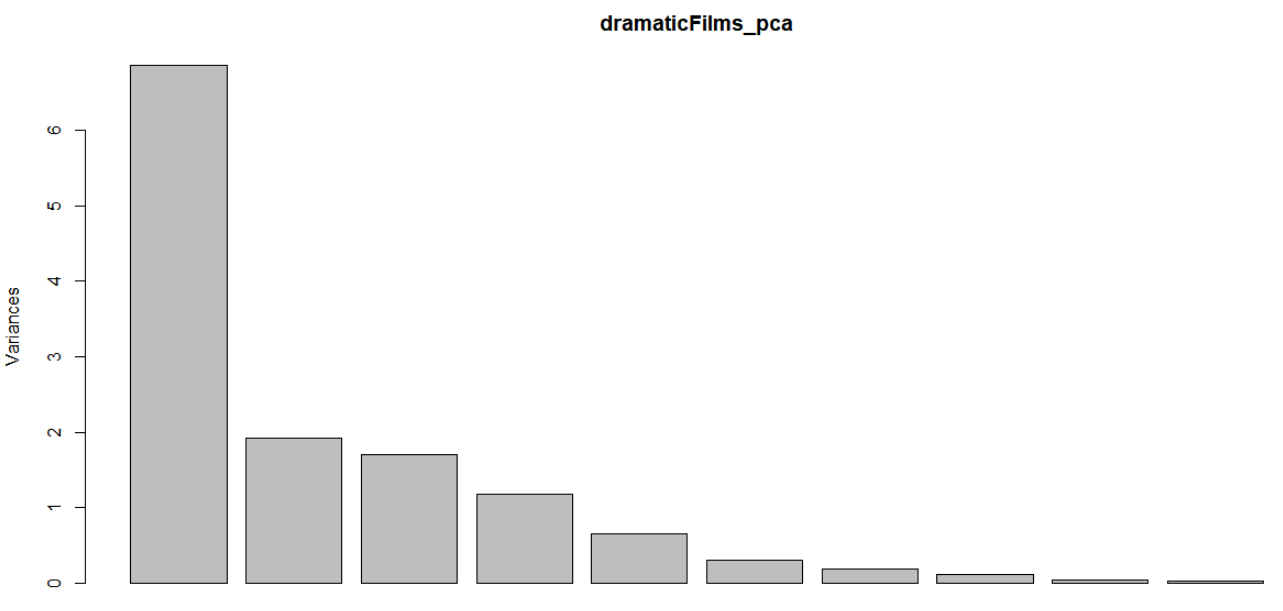


Figure : PCA plot showing relative variance for each PCA component

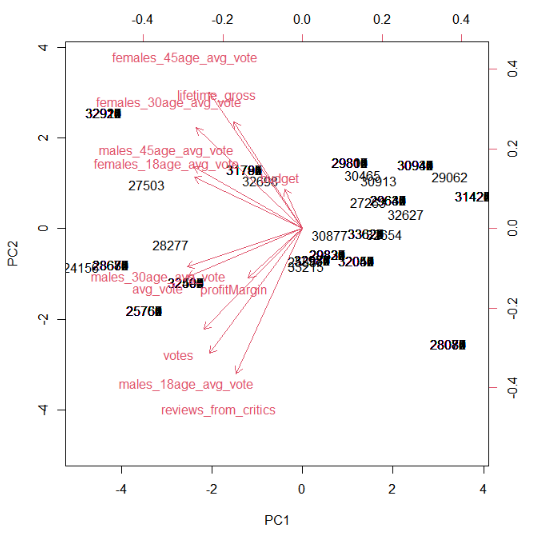


Figure : Plot of the first two principal components

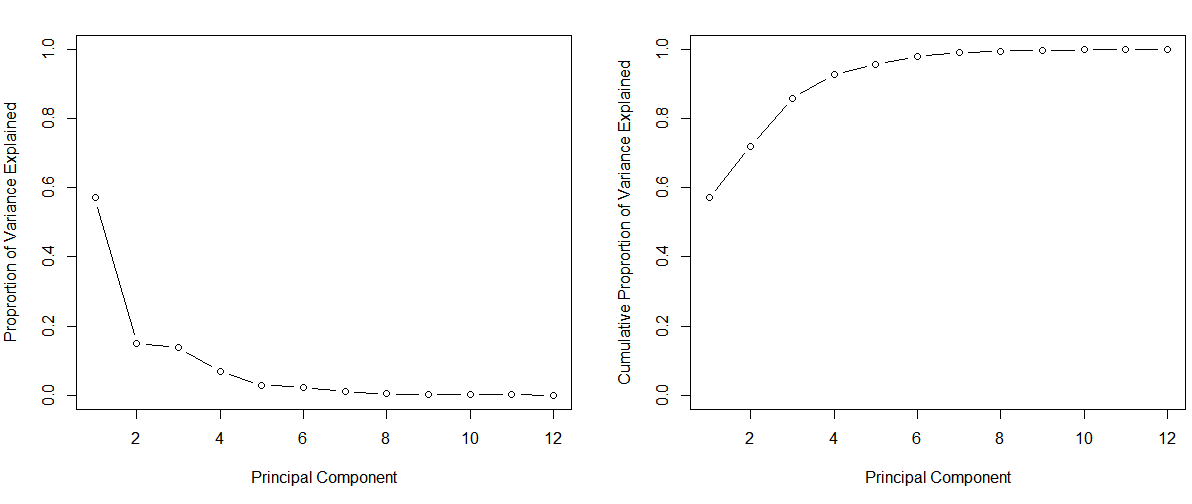


Figure : Scree plot showing proportion of variance and cumulative proportion explained

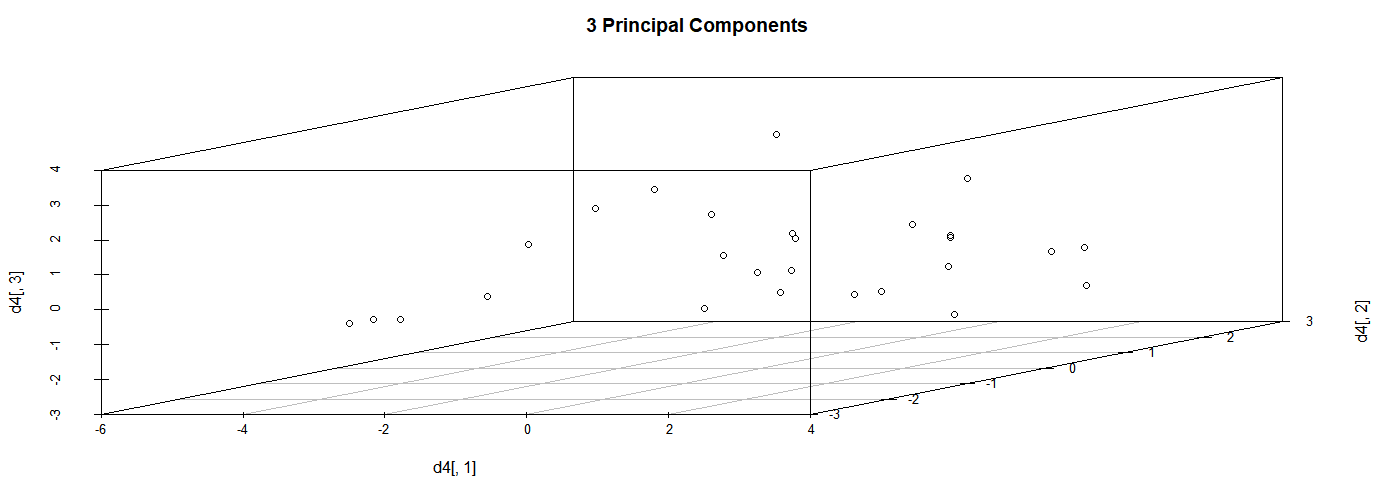


Figure : 3D Scatterplot using three principal components