Statistical Learning in Movies

Max Sellers, Claire Martino, Ari Augustine, and Rim Nassiri

Abstract

To be a producer of film, understanding the factors influencing movie success is extremely important. This project uses statistical learning techniques to predict how successful a film will be critically and popularity-wise. Using a dataset compiled from the Internet Movie Database (IMDB), we delve into the relationships between these variables and movie ratings. Our goal in this project is to use variables like the movie's budget, revenue, runtime, rating, and popularity we can determine what the film's critical and commercial success could be by using models like simple and multiple linear regression.

I. Introduction

In the dynamic film industry, understanding the factors that contribute to a movie's success can be unpredictable. Our data project examines a movie dataset, aiming to uncover relationships between key variables and a movie's performance. This analysis is motivated by an interest in understanding if there are underlying patterns that govern movie success. By using statistical models, particularly linear regression, we aim to find out how significant each variable is in affecting a movie's success.

The movie dataset utilized in this project has a wide range of films spanning different genres, release years, and production scales. Each movie entry in the dataset includes information on key variables providing insights into various aspects such as financial investment and audience engagement. The variables we've selected for further analysis include revenue, budget, popularity, rating, and runtime. By analyzing the data, we aim to identify patterns and trends driving success in the film industry.

In our initial exploration of the dataset, we have formulated hypotheses regarding the potential impact of these variables—budget, popularity, rating, and runtime—on a movie's revenue. These hypotheses are based on common assumptions and consumer behavior. We anticipate that a larger budget will be positively correlated with higher revenue. Greater investment typically enables higher production values, marketing campaigns, and wider distribution, all of which contribute to higher returns. Additionally, higher levels of popularity and rating are predicted to correlate positively with revenue since an increased level of interest usually results in more sales. There may be a non-linear relationship between movie runtime and revenue since it's difficult to anticipate and may not have substantial significance. Through exploratory data analysis and model building, we hope to gain deeper insights into the factors influencing a movie's revenue.

II. Methods

a. Cleaning Data

```
movies_data = movies_data %>%
  group_by(title) %>%
  filter(budget >= 1000000 & revenue >= 1000000 & vote_count > 100) %>%
  arrange(desc(vote_average))
```

b. Choosing Important Values on Dataset

```
#number of categories and variables in the dataset
categories = ncol(movies_data)
var = nrow(movies_data)

#means of variables that will be using in analysis
mean_budget = format(round(mean(movies_data$budget),2),scientific=F)
mean_popularity = format(round(mean(movies_data$popularity),2), scientific = F)
mean_revenue = format(round(mean(movies_data$revenue),2), scientific = F)
mean_vote_average = format(round(mean(movies_data$vote_average),2), scientific = F)
mean_runtime = format(round(mean(movies_data$runtime),2), scientific = F)
```

There are 12 categories in this movies dataset that we will be using. There are 3566 rows in this dataset that we will be using.

The mean budget of the dataset is 42149142. The mean popularity of the movies is 12.34. The mean revenue of the movies is 130185255. The mean voter average of the movies 6.39. The mean runtime of the movies is 110.84.

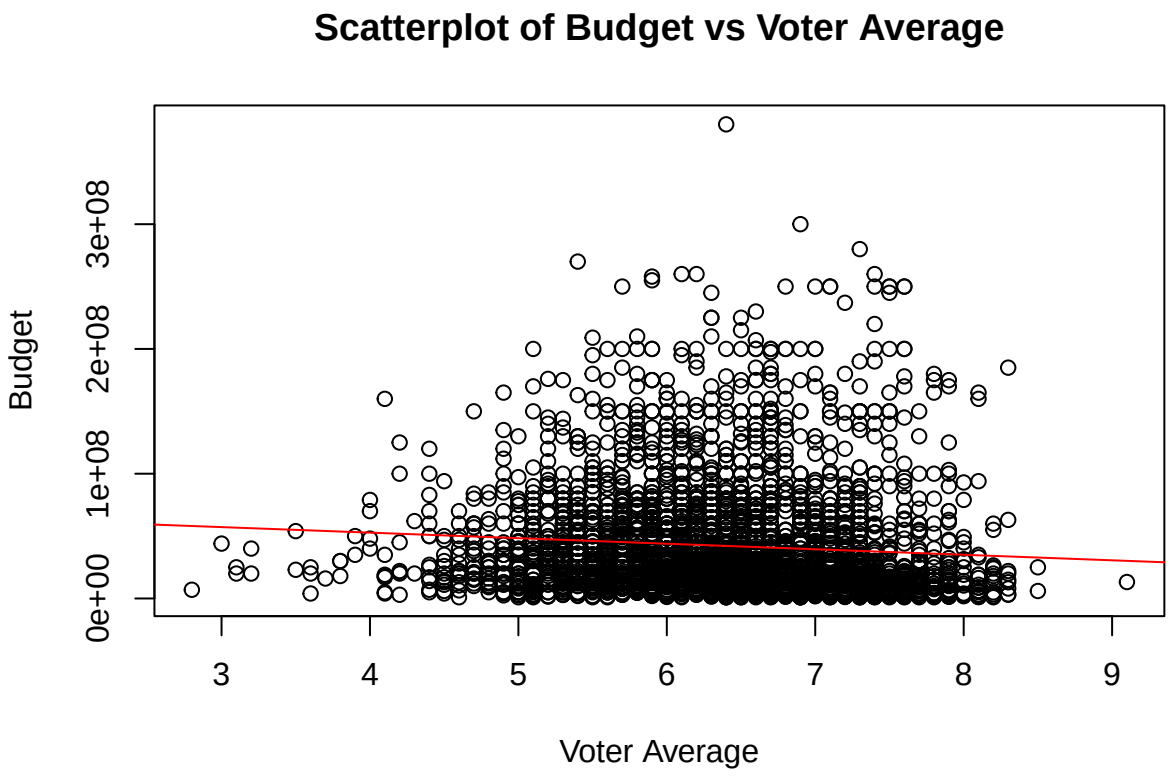
c. Training/Testing Data

```
num_obs = nrow(movies_data)  #extracting the total rows from the data set

movies_indx = sample(num_obs, size = trunc(0.5*num_obs))

#training data set:
movies_train = movies_data[movies_indx,]

#test data set:
movies_test = movies_data[-movies_indx,]
```

III. Results

a. Linear Regression Summary and Line

```
#linear regression model of budget vs vote_average
linear_1 = lm(budget ~ vote_average, data = movies_data)
summary_linear1 = summary(linear_1)

r_2_lin_1 = summary_linear1$r.squared


#graph of above data
plot(budget~ vote_average, data= movies_data, xlab = "Voter Average", ylab = "Budget", main = "Scatterpl
abline(linear_1, col = "red")
```

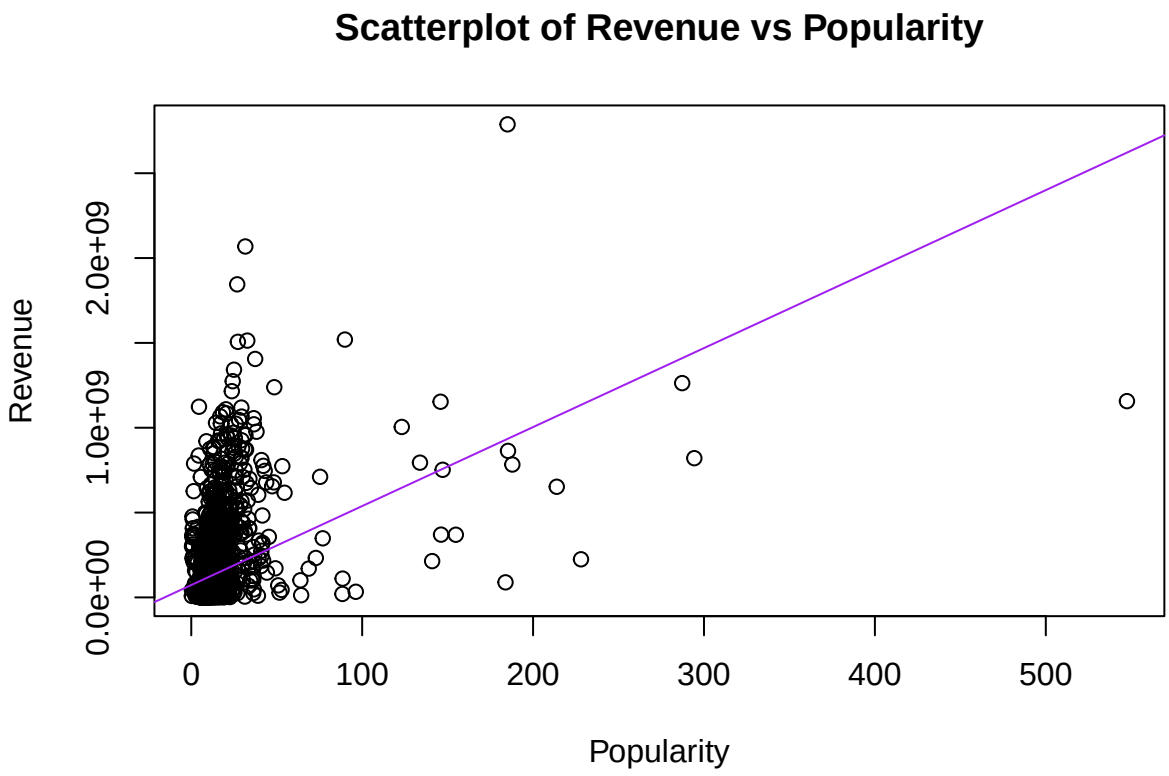


Since the Pr(>|t|) value of voter average is 8.77e_07, and this value is less than the standard level of significance of 0.05, this shows that there is a statistically significant relationship between voter average and budget.

In order to assess the relationship between the predictor and the response variable, you must look at the R^2 value. In this case, $R^2 = 0.006763$. Since this value is closer to 0 than it is 1, this indicates a weak relationship between voter average and budget.

This scatterplot with the linear regression line shows a weak and negative relationship between between budget and voter average.

```
#linear regression model of revenue vs popularity
linear_2 = lm(revenue ~ popularity, data = movies_data)
summary_linear2 = summary(linear_2)

r_2_lin_2 = summary_linear2$r.squared


plot(revenue ~ popularity, data= movies_data, xlab = "Popularity", ylab = "Revenue", main = "Scatterplo
abline(linear_2, col = "purple")
```

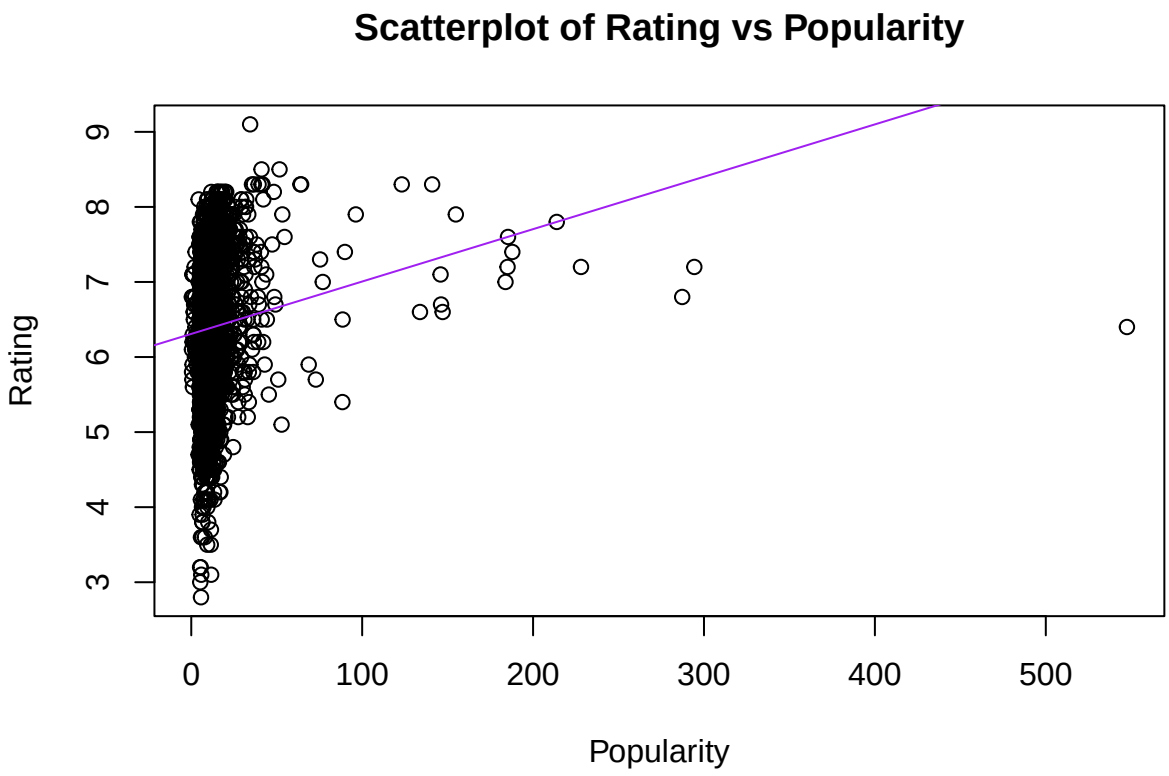


Since the Pr(>|t|) value of popularity is <2e-16, and this value is less than the standard level of significance of 0.05, this shows that there is a statistically significant relationship between popularity and revenue.

In order to assess the relationship between the predictor and the response variable, you must look at the R^2 value. In this case, $R^2 = 0.1547861$. Since this value is closer to 0 than it is 1, this indicates a mildly weak relationship between revenue and popularity.

The scatter plot with the linear regression line shows this mildy weak positive relationship between revenue and popularity.

```
linear_3 = lm(vote_average ~ popularity, data = movies_data)
summary_linear3 = summary(linear_3)

r_2_lin_3 = summary_linear3$r.squared

plot(vote_average ~ popularity, data= movies_data, xlab = "Popularity", ylab = "Rating", main = "Scatte
abline(linear_3, col = "purple")
```



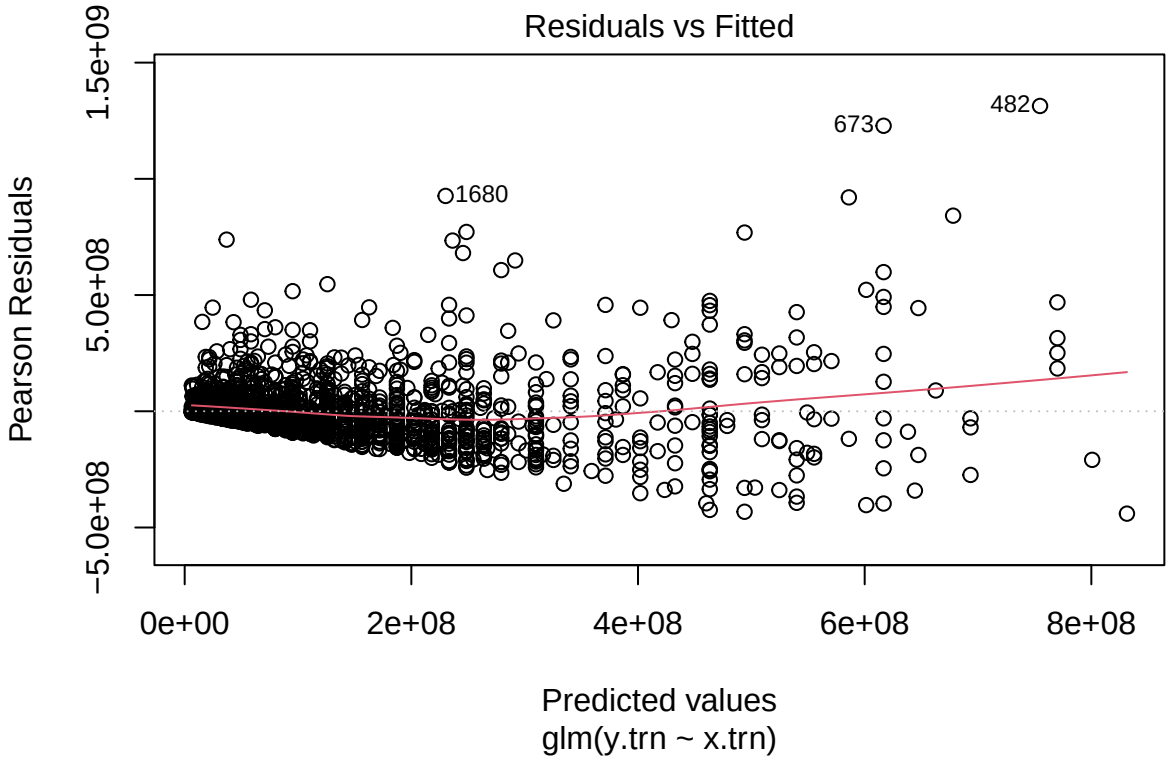
Since the Pr(>|t|) value of popularity is 8.77e-07, and this value is less than the standard level of significance of 0.05, this shows that there is a statistically significant relationship between popularity and rating.

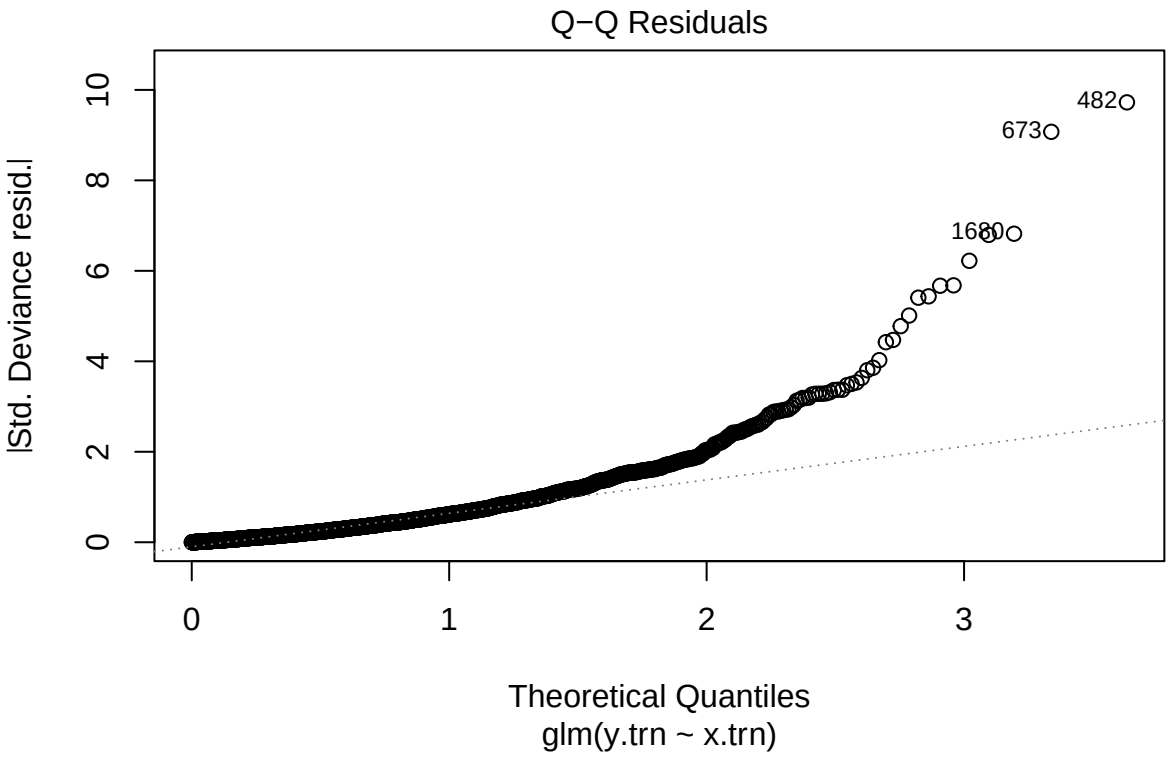
In order to assess the relationship between the predictor and the response variable, you must look at the R^2 value. In this case, $R^2 = 0.0185173$. Since this value is closer to 0 than it is 1, this indicates a weak relationship between rating and popularity.

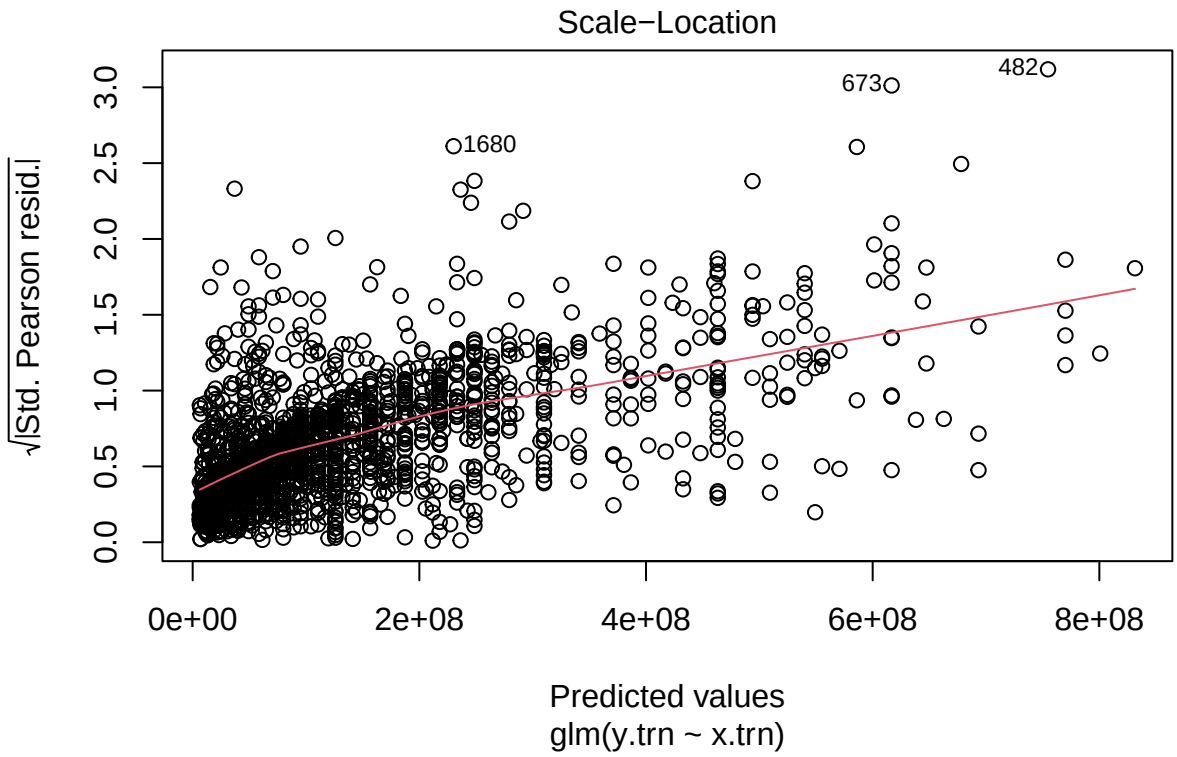
The scatter plot with the linear regression line shows this weak positive relationship between rating and popularity.

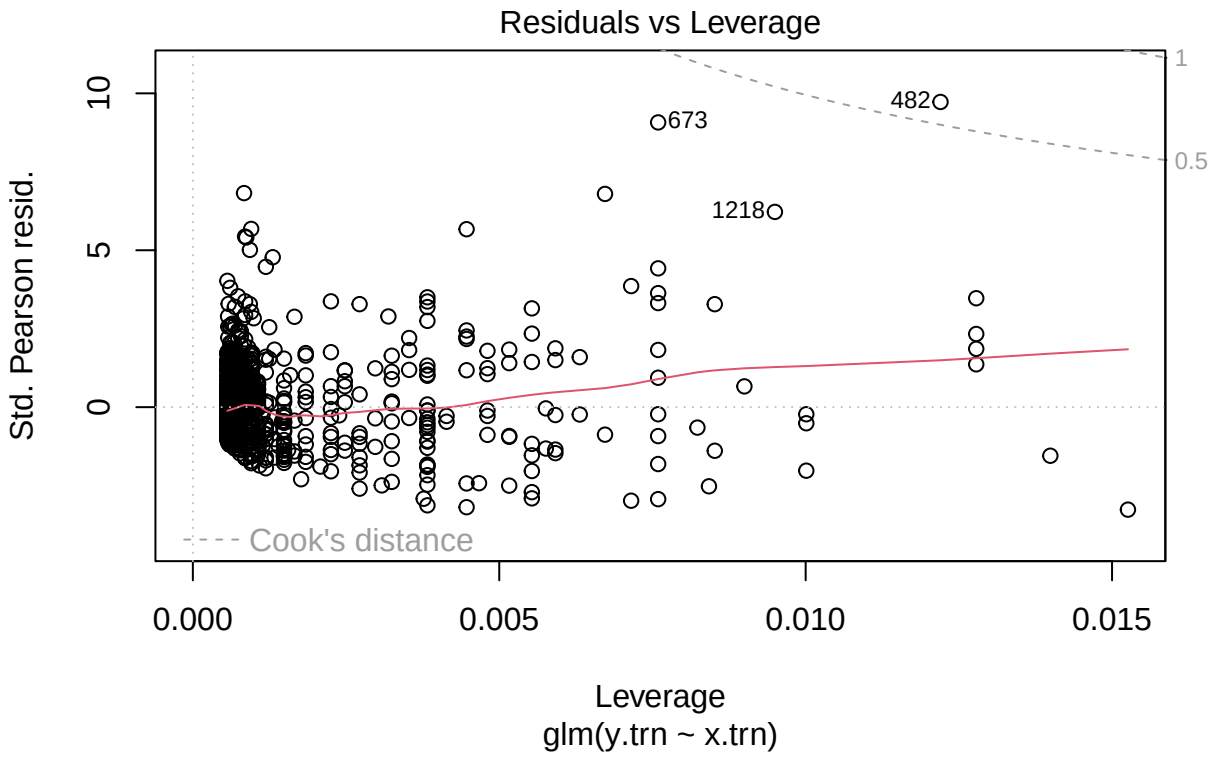
b. Trained Simple Linear Regression Summary

```
y.trn = movies_train$revenue
x.trn = movies_train$budget

train.data_frame = data.frame(y.trn, x.trn)

model_train = glm(formula = y.trn ~ x.trn, data = train.data_frame)

plot(model_train)
```


Residuals vs Fitted
Pearson Residuals
1.5e+09
5.0e+08
−5.0e+08
1680
673
482
0e+00
2e+08
4e+08
6e+08
8e+08
Predicted values
glm(y.trn ~ x.trn)


Q−Q Residuals
|Std. Deviance resid.|
0
2
4
6
8
10
482
673
1680
0
1
2
3
Theoretical Quantiles
glm(y.trn ~ x.trn)


Scale−Location
√|Std. Pearson resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1680
673
482
0e+00
2e+08
4e+08
6e+08
8e+08
Predicted values
glm(y.trn ~ x.trn)




```
actual = ifelse(movies_test$revenue >= 100000000, "Box Office Success", "Box Office Flop")
predicted = ifelse(predict(model_train) >= 100000000, "Box Office Success", "Box Office Flop")

error_function = function(actual, predicted){
  mean(actual != predicted)
}

error_function(actual = actual, predicted = predicted)
```

```
## [1] 0.4901851
```

```
accuracy = round(100*(1 - error_function(actual = actual, predicted = predicted)),2)
library(caret)
```

```
## Loading required package: lattice
```

```
train_table = table(predicted = predicted, Actual = actual)

train_confusion_matrix = confusionMatrix(train_table, positive = "Box Office Success")

sensitivity = round(100*(specificity(train_table)),2)

train_confusion_matrix
```

```
## Confusion Matrix and Statistics
##
##                     Actual
## predicted            Box Office Flop Box Office Success
##   Box Office Flop                629                353
##   Box Office Success             521                280
##
##                Accuracy : 0.5098
##                  95% CI : (0.4863, 0.5333)
##     No Information Rate : 0.645
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0101
##
##  Mcnemar's Test P-Value : 1.615e-08
##
##             Sensitivity : 0.4423
##             Specificity : 0.5470
##          Pos Pred Value : 0.3496
##          Neg Pred Value : 0.6405
##              Prevalence : 0.3550
##          Detection Rate : 0.1570
##    Detection Prevalence : 0.4492
##       Balanced Accuracy : 0.4946
##
##        'Positive' Class : Box Office Success
##
```

IV. Discussion

With the results shown in this project, the data shows how unpredictable a film's success can be. Given the provided variables of budget, voter average, popularity, and revenue it is hard to determine how critically acclaimed and how profitable a movie could be. Throughout all of our actual simple linear regression models, the models' had a very low p-value meaning the model was statistically significant, but had an R-squared value that was much closer to 0 than 1 meaning the relationship was insignificant.

This means having a high budget would not correlate to a high rating, which shows that producers do not need to always need to make the big budget movie to get critical aclaim. For example, the film Whiplash had a high rating of 8.3, but only had a budget of $3.3 million. Having a popular movie does not mean it made a lot of money in its box office run, rather that it had a cult following over time or had better replay value. John Wick is one the most popular movies in the past decade, but had a relatively low box office run of $88 million. Also, a popular movie does not mean it would be critically aclaimed. The most popular movie in this dataset was Minions and it had a tame rating of 6.4 which is widely considered as an average rating on the website.

For our training data set we compared the budget to revenue, which is the most valuable comparison for movie producers. In the confusion matrix, we used the metric of if the revenue is greater than $100 million it is a success and if not it is a flop. After training the model, the training data was accurate 50.98% which is below standard for training datasets. The sensitivity of 44.23% is bad because that means when a movie is a box office success it is only predicted to be a success 44.23% of the time.

Overall, it is incredibly hard to determine how successful a movie would be given these parameters. This project gave us more questions than answers. Do the writers or directors determine more of the success of a movie, than the budget of the movie? Is it wise to use Machine Learning on creative arts like film, art, and music?

Appendix

```
library(readxl)
Data_Dictionary = read_excel("Data_Dictionary.xlsx")
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
```

```
Data_Dictionary
```

```
## # A tibble: 10 x 5
##    `File Data Dictionary` ...2            ...3                       ...4  ...5
##    <chr>                  <chr>           <chr>                      <chr> <chr>
##  1 File Name:             data_dictionary <NA>                       <NA>  <NA>
##  2 File Format:           XLSX            <NA>                       <NA>  <NA>
##  3 <NA>                   <NA>            <NA>                       <NA>  <NA>
##  4 <NA>                   <NA>            <NA>                       <NA>  <NA>
##  5 Field Name             Field Type      Description                Null~ Cons~
##  6 Budget                 int             Total amount of money all~ N     Cont~
##  7 Popularity             float           Metric indicating the lev~ N     Cont~
##  8 Revenue                int             Total income generated by~ N     Cont~
##  9 Runtime                int             Duration of a movie, meas~ N     n/a
## 10 Rating                 float           Average rating given to a~ N     n/a
```

Data from https://www.kaggle.com/datasets/jacopoferretti/idmb-movies-user-friendly