Statistical Learning in Movies

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

Abstract

I. Introduction

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 [USE SOMEWHERE]

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

movies_indx = sample(num_obs, size = trunc(0.6*num_obs))
```

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

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

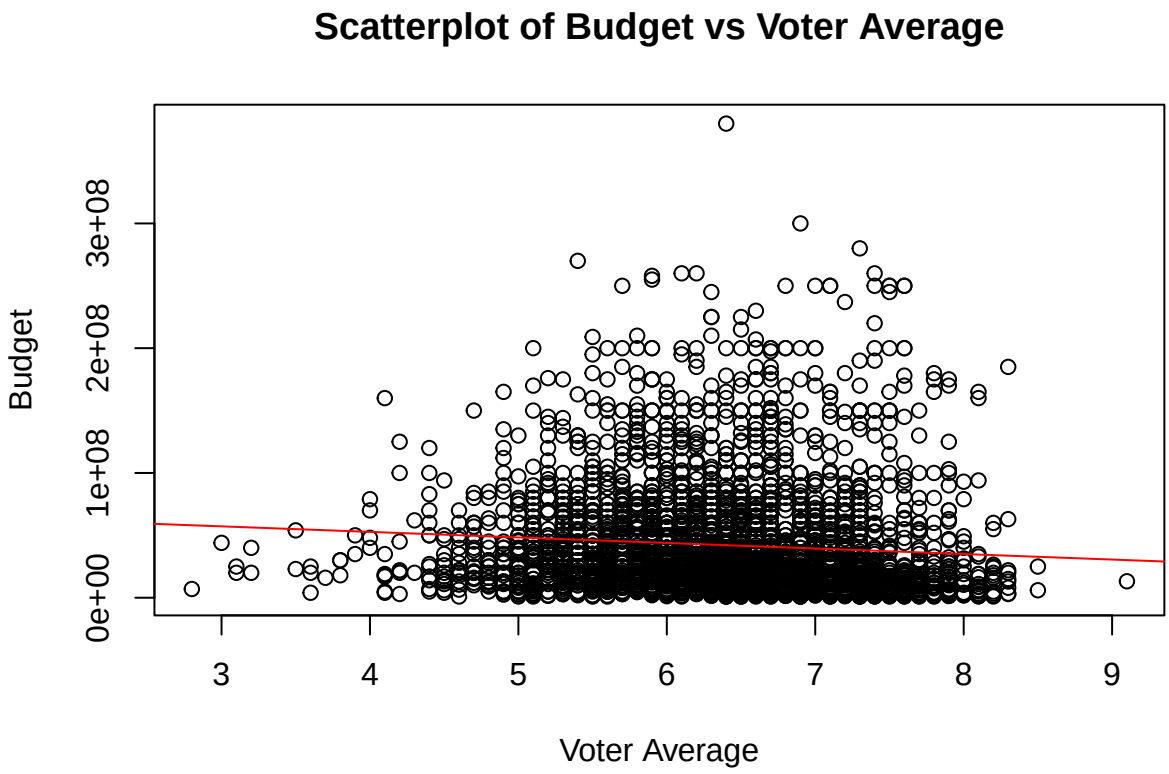
III. Results

a. Linear Regression Summary and Line

```r
#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.

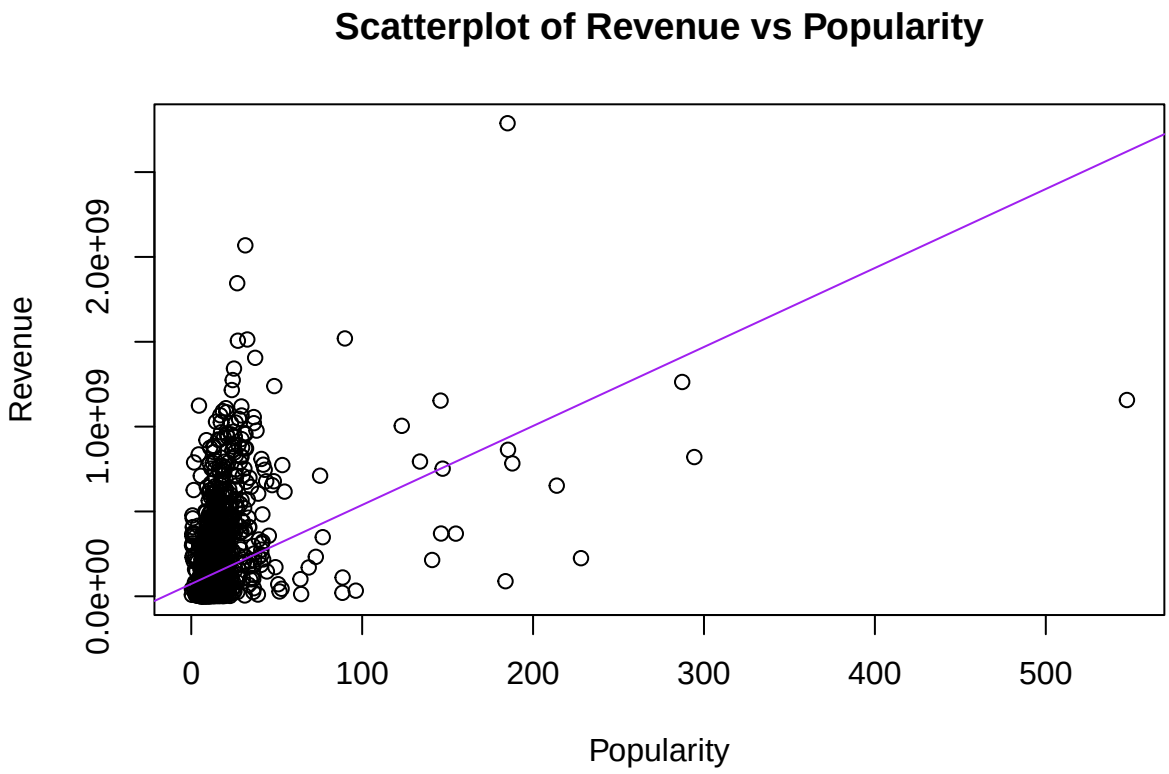
In order to assess the relationship between the predictor and the response variable, you must look at the R^2 value. In this case, $R^2 = 0.006763$. Since this value is closer to 0 than it is 1, this indicates a weak relationship between voter average and budget.

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.

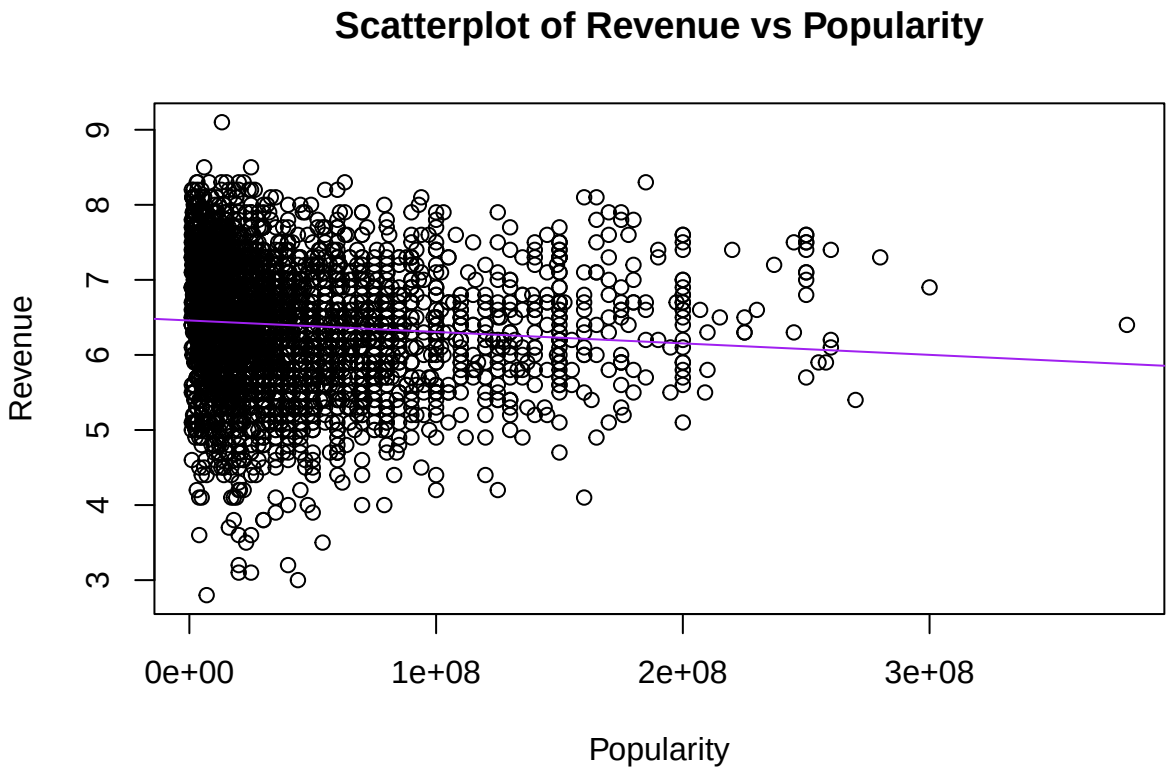
The scatter plot with the linear regression line shows this mildy weak positive relationship between revenue and popularity.

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

r_2_lin_3 = summary_linear3$r.squared
r_2_lin_3
```

```
## [1] 0.006763019
```

```r
plot(vote_average ~ budget, data= movies_data, xlab = "Popularity", ylab = "Revenue", main = "Scatterplo
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 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, $R^2 = 0.006763$. Since this value is closer to 0 than it is 1, this indicates a weak relationship between revenue and popularity.

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

b. Multiple Linear Regression Summary

```r
multi_linear = lm(revenue ~ budget + vote_average, data = movies_data)
summary(multi_linear)
```

```
##
## Call:
## lm(formula = revenue ~ budget + vote_average, data = movies_data)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -672617496  -60126512  -17169262   37828453 2017931375
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.125e+08  1.728e+07  -18.09   <2e-16 ***
## budget        3.082e+00  4.887e-02   63.06   <2e-16 ***
## vote_average  4.892e+07  2.635e+06   18.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.3e+08 on 3563 degrees of freedom
## Multiple R-squared:  0.5385, Adjusted R-squared:  0.5382
## F-statistic:  2078 on 2 and 3563 DF,  p-value: < 2.2e-16
```

```r
multi_linear2 = lm(vote_average ~ budget + revenue + popularity + runtime, data = movies_data)
summary(multi_linear2)
```

```
##
## Call:
## lm(formula = vote_average ~ budget + revenue + popularity + runtime,
##     data = movies_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3287 -0.4499  0.0311  0.4755  2.5441
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.558e+00  6.797e-02  67.058  < 2e-16 ***
## budget      -7.877e-09  3.770e-10 -20.894  < 2e-16 ***
## revenue      1.428e-09  9.308e-11  15.345  < 2e-16 ***
## popularity   3.022e-03  8.032e-04   3.763 0.000171 ***
## runtime      1.755e-02  6.180e-04  28.397  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.711 on 3561 degrees of freedom
## Multiple R-squared:  0.2651, Adjusted R-squared:  0.2643
## F-statistic: 321.2 on 4 and 3561 DF,  p-value: < 2.2e-16
```

```r
#confidence interval insert here
```

IV. Discussion