# Statistical Learning in Movies

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

### Abstract

In

## I. Introduction

## II. Methods

### a. Cleaning Data

```
movies_data = movies_data %>%
  group_by(title) %>%
  filter(budget >= 1000 & revenue >= 1000 & vote_count > 100) %>%
  arrange(desc(vote_average))
movies_data
```

```
## # A tibble: 3,778 x 12
## # Groups:   title [3,711]
##      budget popularity revenue runtime title vote_average vote_count day_of_week
##       <int>      <dbl>   <dbl>   <int> <chr>        <dbl>      <int> <chr>
##   1 13200000       34.5  1   e8     190 Dilw~          9.1        661 Friday
##   2 25000000       51.6  2.83e7     142 The ~          8.5       8358 Friday
##   3  6000000       41.1  2.45e8     175 The ~          8.5       6024 Tuesday
##   4  8000000      141.   2.14e8     154 Pulp~          8.3       8670 Saturday
##   5 22000000       41.7  3.21e8     195 Schi~          8.3       4436 Monday
##   6  3000000       35.5  1.09e8     133 One ~          8.3       3001 Tuesday
##   7   806948       36.8  3.20e7     109 Psyc~          8.3       2405 Thursday
##   8 13000000       36.6  4.75e7     200 The ~          8.3       3418 Friday
##   9 20000000       39.4  2.29e8     116 Life~          8.3       3643 Saturday
##  10 63000000       63.9  1.01e8     139 Figh~          8.3       9678 Friday
## # i 3,768 more rows
## # i 4 more variables: month <chr>, season <chr>, year <int>, genre <chr>
```

### b. Choosing Important Values on Dataset

```
#number of cateogories and varaibles in the dataset
catagories = ncol(movies_data)
var = nrow(movies_data)

#means of variables that will be using in analysis
mean_budget = round(mean(movies_data$budget),2)
mean_popularity = round(mean(movies_data$popularity),2)
mean_revenue = round(mean(movies_data$revenue),2)
```

### c. Training Data

### d. Testing Data

## III. Results

### a. Graphs of Specific Data

### b. Linear Regression Summary and Line

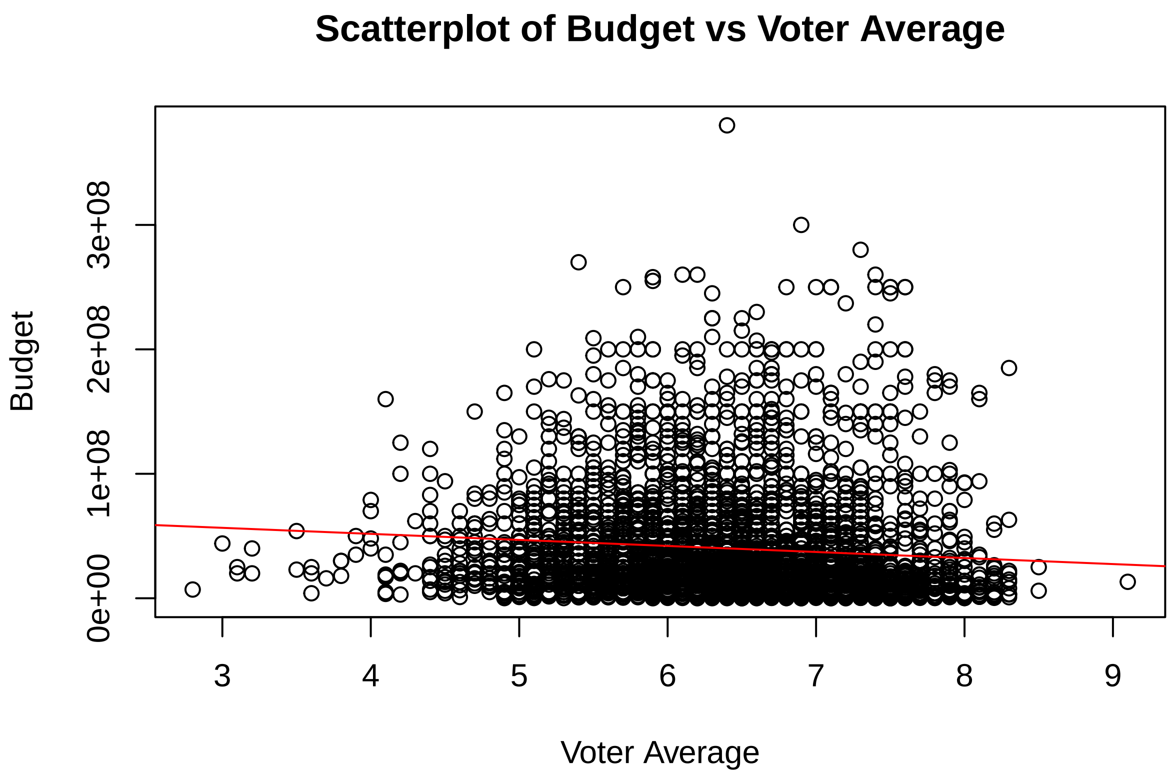
```
#linear regression model of budget vs vote_average
linear_1 = lm(budget ~ vote_average, data = movies_data)
summary_linear1 = summary(linear_1)
summary_linear1
```

```
##
## Call:
## lm(formula = budget ~ vote_average, data = movies_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -50551064 -29115284 -15085588  12518347 339914412
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71135322    5577958  12.753  < 2e-16 ***
## vote_average -4851521     863493  -5.618 2.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44140000 on 3776 degrees of freedom
## Multiple R-squared:  0.008291,   Adjusted R-squared:  0.008028
## F-statistic: 31.57 on 1 and 3776 DF,  p-value: 2.065e-08
```

```
r_2_lin_1 = summary_linear1$r.squared
r_2_lin_1
```

```
## [1] 0.008290695
```

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



```
#plot(linear_1)
```

Since the Pr(>|t|) value of voter average 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 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, $R^2 = 0.0082907$. Since this value is closer to 0 than it is 1, this indicates a weak relationship between voter average and budget.

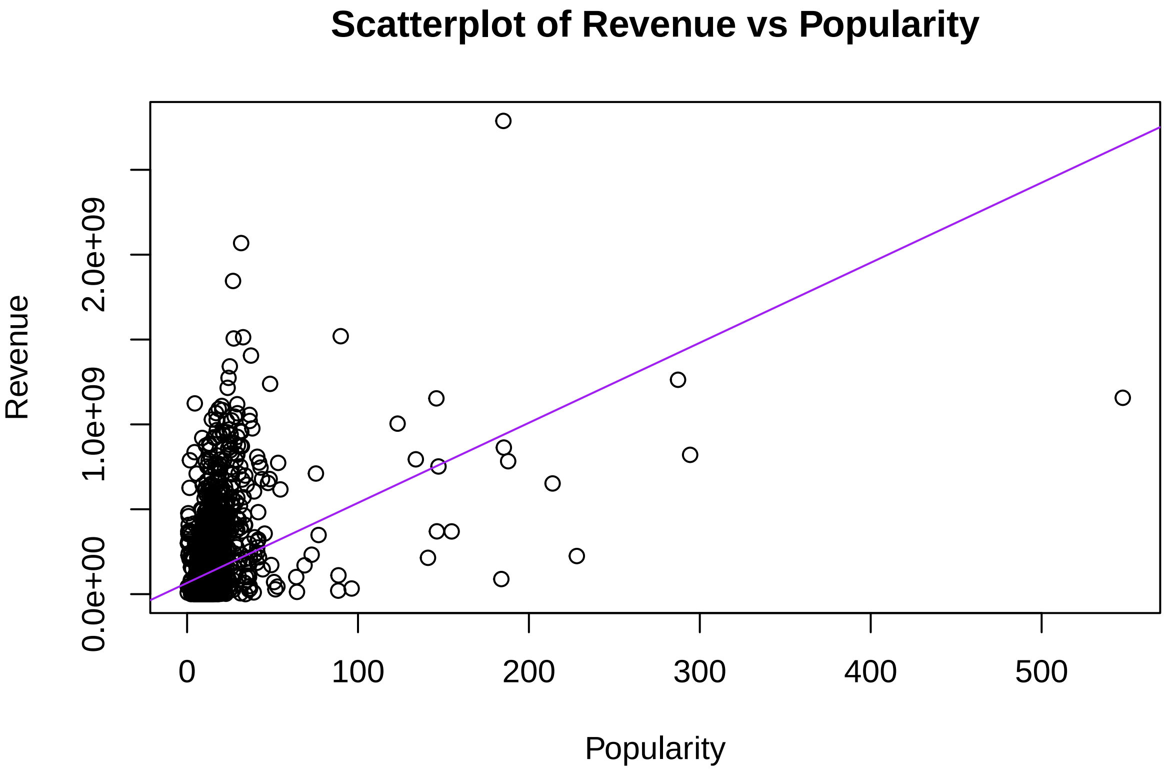
```
#linear regression model of revenue vs popularity
linear_2 = lm(revenue ~ popularity, data = movies_data)
summary_linear2 = summary(linear_2)

r_2_lin_2 = summary_linear2$r.squared
r_2_lin_2
```

```
## [1] 0.1561118
```

```
#plot(linear_2)

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



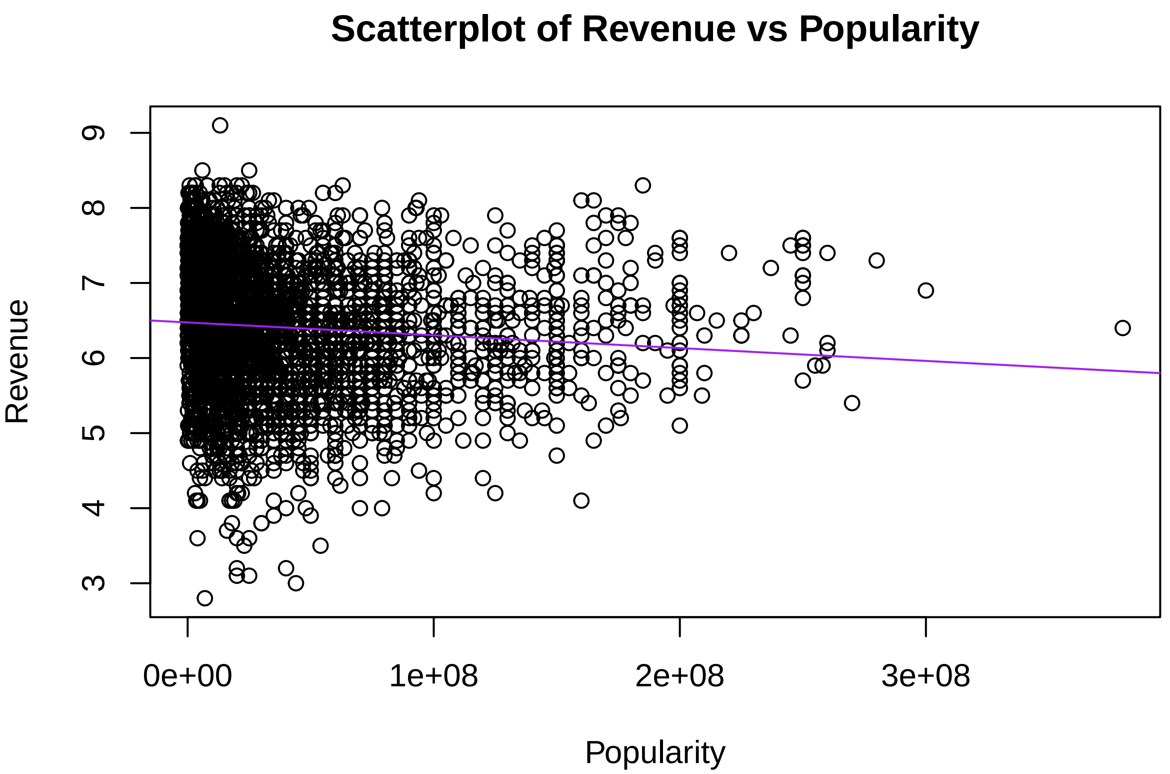
```
linear_3 = lm(vote_average ~ budget, data = movies_data)
summary_linear3 = summary(linear_2)

r_2_lin_3 = summary_linear3$r.squared
r_2_lin_3
```

```
## [1] 0.1561118
```

```
#plot(linear_2)

plot(vote_average ~ budget, data= movies_data, xlab = "Popularity", ylab = "Revenue", main = "Scatterpl
abline(linear_3, 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, $R^2 = 0.1561118$. Since this value is closer to 0 than it is 1, this indicates a mildly weak relationship between revenue and popularity.

### c. Multiple Linear Regression Summary

```
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
## -674088075  -57034859  -16385006   35784504 2020061109
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.962e+08  1.640e+07  -18.06   <2e-16 ***
## budget        3.086e+00  4.684e-02   65.89   <2e-16 ***
## vote_average  4.621e+07  2.496e+06   18.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.27e+08 on 3775 degrees of freedom
## Multiple R-squared:  0.5437, Adjusted R-squared:  0.5435
## F-statistic:  2249 on 2 and 3775 DF,  p-value: < 2.2e-16
```

```
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 
## -3.3575 -0.4577  0.0266  0.4800  2.4580 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.661e+00  6.653e-02  70.070  < 2e-16 ***
## budget      -8.174e-09  3.788e-10 -21.579  < 2e-16 ***
## revenue      1.441e-09  9.413e-11  15.305  < 2e-16 ***
## popularity   3.153e-03  8.138e-04   3.875 0.000109 ***
## runtime      1.684e-02  6.094e-04  27.628  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7219 on 3773 degrees of freedom
## Multiple R-squared:  0.2475, Adjusted R-squared:  0.2467 
## F-statistic: 310.2 on 4 and 3773 DF,  p-value: < 2.2e-16
```

```
#confidence interval insert here
```

## IV. Discussion