Statistical Learning in Movies

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

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

H_T : There is not difference in Average Wi-Fi speed between Hanson Hall of Science and Old Main.

H_A : There is a difference between in average Wi-Fi speeds between Hanson Hall of Science and Old main.

$\alpha = 0.05$

Abstract

I. Introduction

II. Methods

a. Cleaning Data

```
movies_data = movies_data %>%
  group_by(title) %>%
  filter(budget != 0 & revenue != 0 & vote_count > 100) %>%
  arrange(desc(vote_average))
```

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

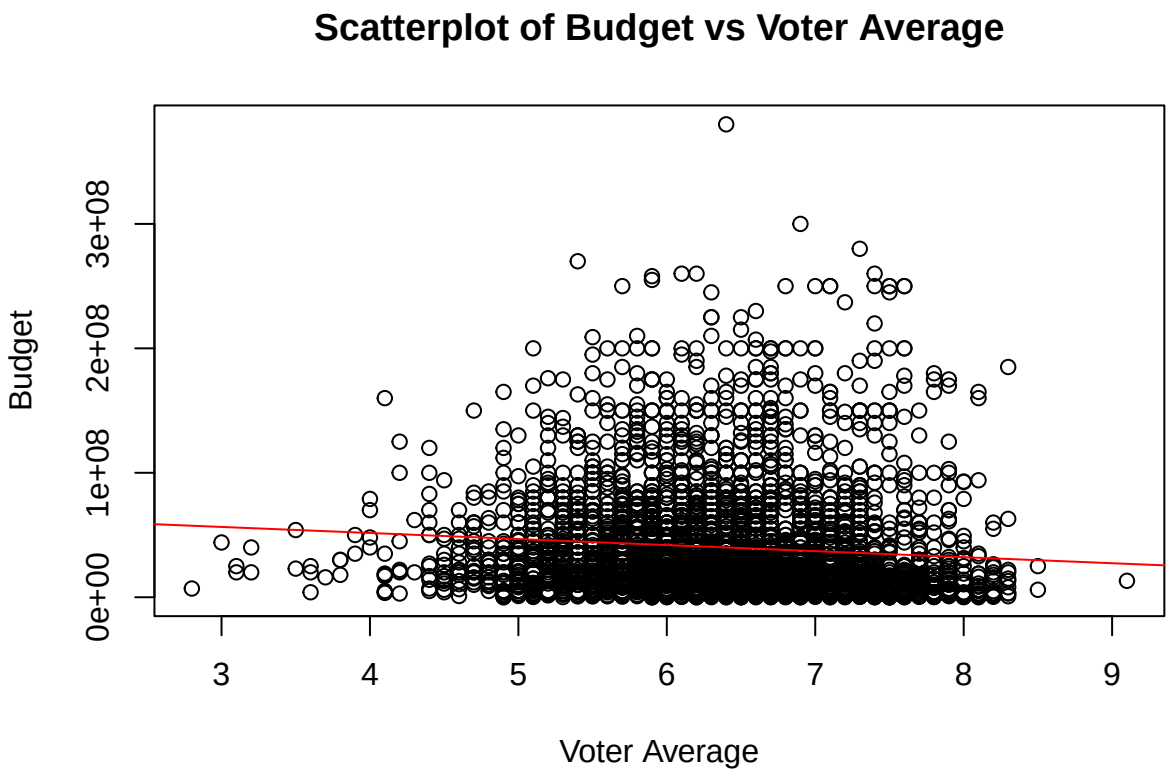
```
#linear regression model of budget vs vote_average
linear_1 = lm(budget ~ vote_average, data = movies_data)
summary_linear1 = summary(linear_1)
summary_linear1
```

```
##
## Call:
## lm(formula = budget ~ vote_average, data = movies_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -50458935 -29103923 -15096835  12601623 340030745
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71062020    5565826  12.768  < 2e-16 ***
## vote_average -4858244     861558  -5.639 1.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44120000 on 3788 degrees of freedom
## Multiple R-squared:  0.008324,   Adjusted R-squared:  0.008063
## F-statistic:  31.8 on 1 and 3788 DF,  p-value: 1.836e-08
```

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

```
## [1] 0.008324328
```

```
#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, $R^2 = 0.0083243$. Since this value is closer to 0 than it is 1, this indicates a weak relationship between voter average and budget.

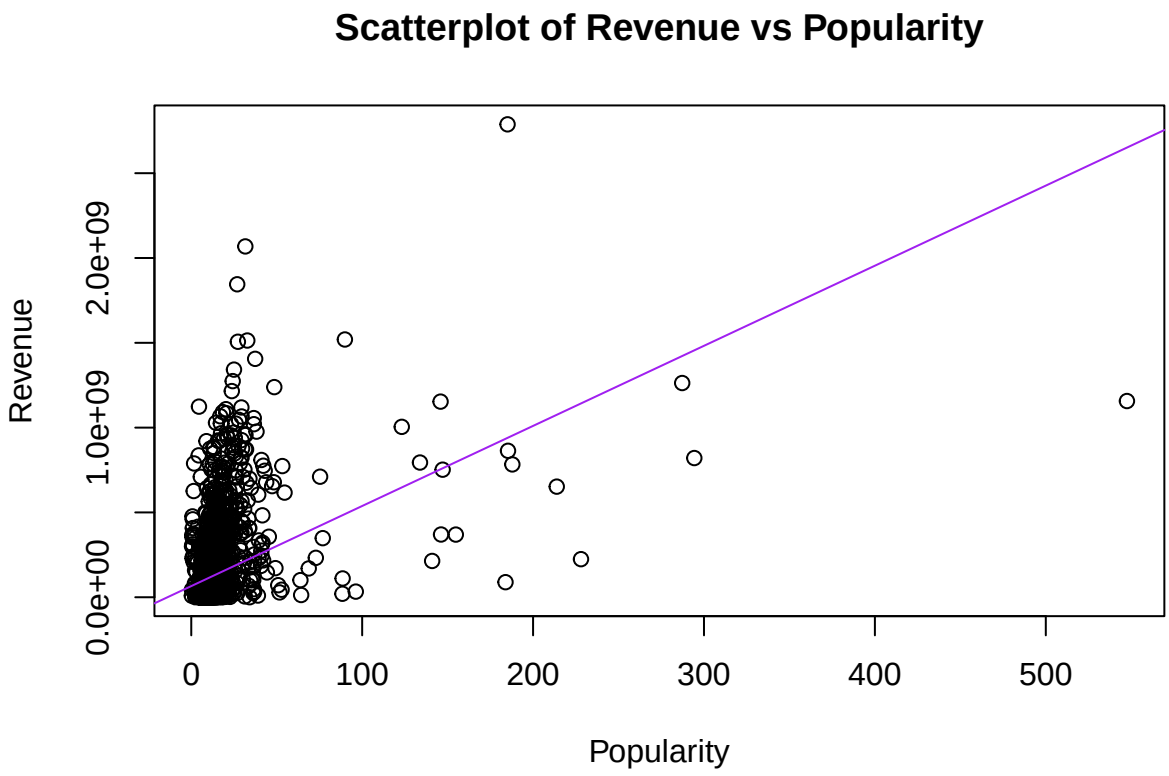
```
#linear regression model of revenue vs popularity
linear_2 = lm(revenue ~ popularity, data = movies_data)
summary_linear2 = summary(linear_2)

r_2_lin_2 = summary_linear2$r.squared
r_2_lin_2
```

```
## [1] 0.1563509
```

```
#plot(linear_2)

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, $R^2 = 0.1563509$. 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
## -674093484  -56948534  -16254609   35600930 2020292697
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.950e+08  1.635e+07  -18.04   <2e-16 ***
## budget        3.086e+00  4.673e-02   66.03   <2e-16 ***
## vote_average  4.602e+07  2.488e+06   18.50   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 126900000 on 3787 degrees of freedom
## Multiple R-squared:  0.544,  Adjusted R-squared:  0.5437
## F-statistic:  2259 on 2 and 3787 DF,  p-value: < 2.2e-16
```

```
#confidence interval insert here
```

```
#logit.model.test = glm(revenue ~ budget + vote_average,
                        #data = movies_data,family = binomial(link = "logit"))
```

IV. Discussion