Examining the Influence of Film Directors on Movie Opening Gross

2023-10-05

Created variables

In this analysis, we will look at the effect of the director on the opening gross of the movie. As stated before, we merged several datasets from Imdb with the boxofficemojo data on the opening gross of movies. For measuring the effect of the director, we constructed the variables "director_count" and "directed_above_median". The variable director_count measures the amount of times a specific director has directed a movie in the movie_director_data dataset. The variable directed_above_median is a dummy variable that equals 1 if the director has directed more movies than the median.

```
library(readr)
# Import the dataset
csv_file <- "filtered_merged_dataset.csv"
folder <- "../../gen/analysis"

folder_path <- file.path(folder, csv_file)

movie_director_data <- read.csv2(folder_path)
```

Utilized models

T-test

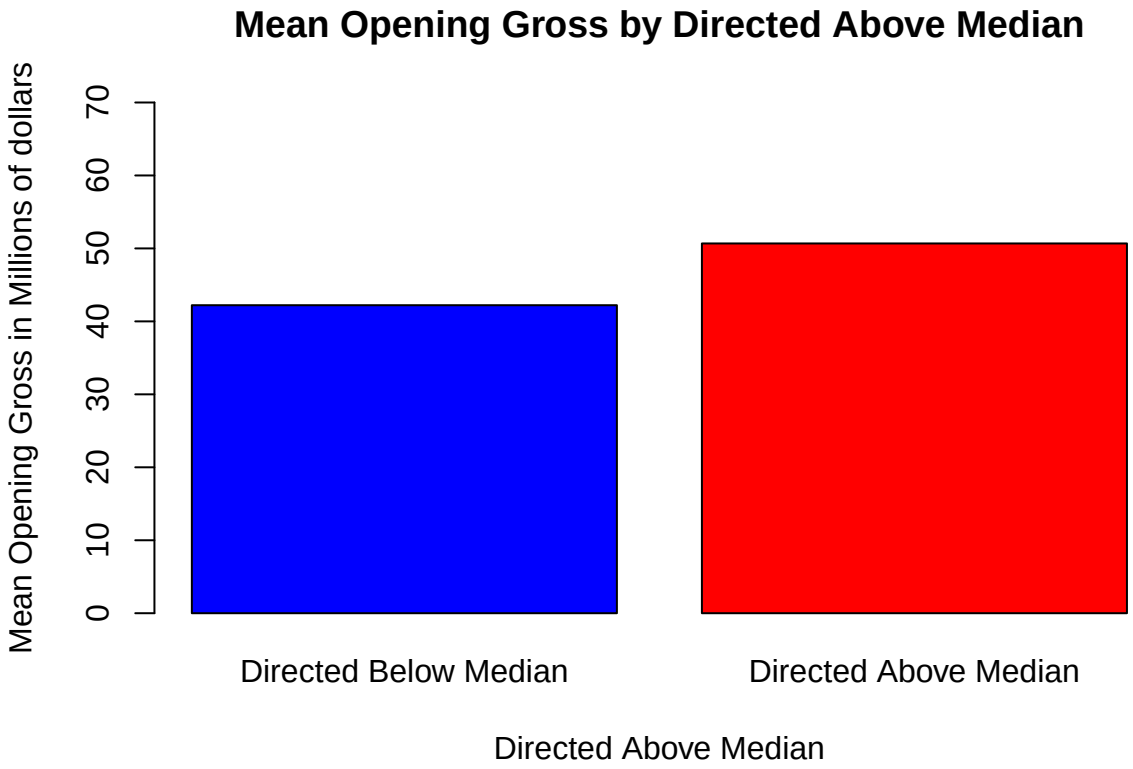
Firstly, we conduct a t-test to examine whether there is a significant difference in the means of the "directed_above_median" groups concerning opening gross.

```
# T-test + Bar plot to make the t-test visual
  # T-test with opening gross DV and director_above_median IV
t_test_result <- t.test(openinggross ~ directed_above_median, data = movie_director_data)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  openinggross by directed_above_median
## t = -3.6277, df = 710.85, p-value = 0.0003064
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -13049390  -3884769
## sample estimates:
## mean in group 0 mean in group 1
##        42213705        50680785
```

```r
# Calculate the average opening gross per group
means <- tapply(movie_director_data$openinggross / 1e6, movie_director_data$directed_above_median, mean)

# Make a bar chart of the two groups directed above the median
barplot(means,
        names.arg = c("Directed Below Median", "Directed Above Median"),
        ylab = "Mean Opening Gross in Millions of dollars",
        xlab = "Directed Above Median",
        col = c("blue", "red"),
        main = "Mean Opening Gross by Directed Above Median",
        ylim = c(0, 70))  # Limit the y-axis to 0-70
```



Linear regression

We will do a regression analysis to find out if the significance in the mean of the groups is caused by the missing covariables. Thus, we will conduct a regression analysis by including the covariables theaters and runtime_dummy. In this regression, openinggross is the independent variable. The numbers of theaters and the runtimedummy serve as covariables in the regression.

```r
# Regression analysis
openinggross_lm1 <- lm(openinggross ~ theathers + directed_above_median + runtime_dummy, movie_director
summary(openinggross_lm1)
```

```
##
```

```
## Call:
## lm(formula = openinggross ~ theathers + directed_above_median +
##     runtime_dummy, data = movie_director_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -35406416 -17060130  -8765505   7999597 203411744
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             36076988    1697200  21.257  < 2e-16 ***
## theathers                 -14357      23204  -0.619   0.5363
## directed_above_median    4279071    2297908   1.862   0.0629 .
## runtime_dummy           16433018    2269404   7.241 1.01e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31930000 on 840 degrees of freedom
## Multiple R-squared:  0.0742, Adjusted R-squared:  0.07089
## F-statistic: 22.44 on 3 and 840 DF,  p-value: 5.54e-14
```

Conclusion

Looking at the results, we clearly observe a difference in the two means between the opening gross of films directed by directors with higher recognition (directed_above_median = 1) against directors with a lower recognition (directed_above_median = 0), and this difference is statistically significant. However, upon reviewing the results of the linear regression, it becomes clear that this outcome is predominantly driven by other variables. By including the variables "theaters" and "runtime_dummy," we observe that the p-value for the variable "directed_above_median" becomes insignificant. The remaining variables in this model maintain their significance. Therefore, based on this dataset, it cannot be concluded that the director of a film has a significant impact on the opening gross of a film.