# STATS 101A - Final Project Report: Predicting Gross Revenue from Film Metrics

*Authors: Moulik Chatterjee, Isaac Yu, Jason Chung, Peyton Garrett, Uriel Santa Cruz*

## INTRODUCTION

This research project aims to predict gross box office revenue based on several film metrics. We utilized a dataset from Kaggle that includes 7,545 observations and 11 variables: index, *MovieID, Title, MPAA Rating, Budget, Gross, Release Date, Genre, Runtime, Rating, and Rating Count.* Here is a breakdown of the variables and their interpretation:

1. *MovieID*: An index of the observations — movies; unitless
2. *Title*: Name of the movie; unitless
3. *MPAA Rating*: The ratings of individual motion pictures as determined by the Motion Pictures Association (MPA) based on criteria like mature themes, language, violence, etc[1]
4. *Budget:* The total cost of filming, producing, and releasing the movie; dollars
5. *Gross:* Total revenue generated by the movie; dollars
6. *Release Date*: When the movie was released; yy-mm-dd
7. *Runtime:* The duration of the movie; minutes
8. *Rating:* An average score voted on by viewers that operates on a scale of 1-10; unitless
9. *Rating.Count:* The number of people who have rated the movie; unitless

Predicting a film's box office performance is crucial for studies, investors, and distributors to optimize marketing strategies, manage budgets, and anticipate investment returns. Global box office returns are still attempting to return to pre-COVID levels[2], thus driving the increasing implementation of data-driven decision-making in the entertainment industry as it pertains to movie production.

This dataset includes several key predictive factors like *Budget, MPAA Rating, Runtime, Genre, and Release Date*, which can significantly influence the likelihood of a movie being a flop or blockbuster. Before beginning our research, we had several assumptions we wanted to explore, like whether genre and rating are statistically significant predictors of movie revenue. For example, the combined worldwide revenue of the top five highest-grossing PG-13 movies is more than 2.5 times the combined worldwide revenue of the top five highest-grossing R-rated films, and this is a trend we wanted to verify in our research. Another key assumption we explored through this project is the impact of production budget - whether its impact on revenue can be modeled as an increasing linear relationship.

While this project attempts to provide a surface-level prediction basis for box office revenue, we acknowledge a few shortcomings and challenges in prediction: unpredictable cultural trends, competition, and biases that should be considered when designing models that improve upon our approach in the future.

## DATA DESCRIPTION

For our research, we wanted to predict a movie's Gross Revenue (*Gross*) based on the following numerical variables: *Budget(Dollars), Runtime(Minutes), Rating, and Rating count* as categorical variables like Genre and Release Date were difficult to convert into binary predictors.

To pare down the extent of our research and isolate its objective, we started by manipulating the dataset. First, from Observation 511 onward, the data was either empty or irrelevant to our analysis. Thus, we only utilized 510 observations for our analysis. Additionally, we converted the data types of each variable to create a proper predictive model. Furthermore, we removed the *Summary* variable as a qualitative description of movies is irrelevant to our analysis. Lastly, variables such as *Rating* and *Rating.Count* contained NULL values, so we removed them accordingly (Refer to Appendix A Figures 1-3).

Now, with a dataset of 510 observations and 5 different variables for our research, we first observed the characteristics of our dataset before conducting any further analysis.

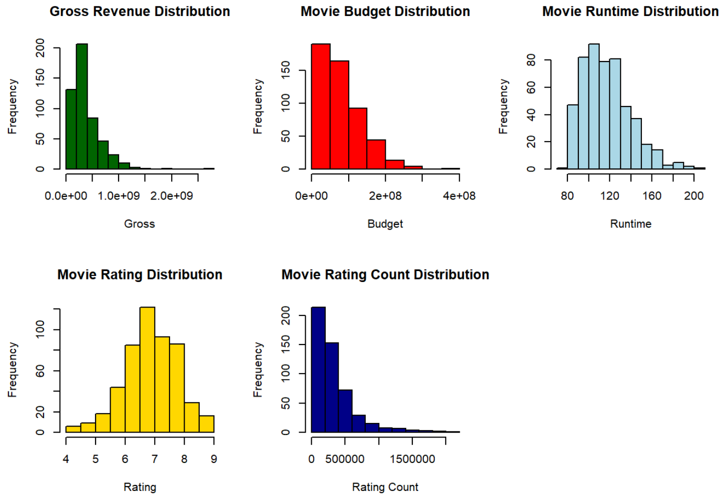


*Figure 1 - Histograms of Analyzed Variables*

The histograms show that all variables are right skewed except for *Rating*. We may be expected to manipulate the response and predictor variables to make them jointly normal for the model.

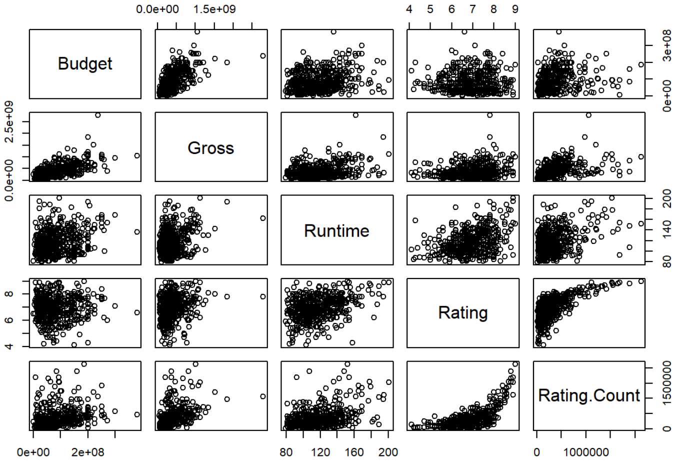
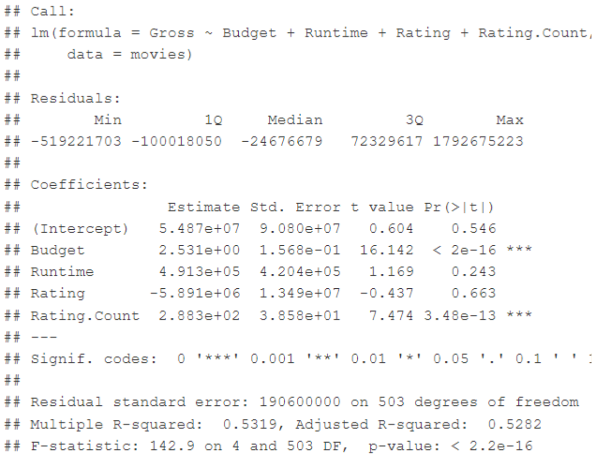


*Figure 2 - Scatterplot Matrix of Analyzed Variables*

Additionally, our correlation scatterplot matrix shows non-linear relationships between variables, especially *Rating* and *Rating.Count*. We suspect that *Rating* and *Rating.Count* may contribute to multicollinearity in our model. However, considering other variables do not display a strong correlation, we must use VIF to check if *Rating* and *Rating.Count* does contribute towards multicollinearity.

## RESULTS AND INTERPRETATION

**Model 1**:
Findings from our first model are shown below.



```
## Call:
## lm(formula = Gross ~ Budget + Runtime + Rating + Rating.Count,
##     data = movies)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -519221703 -100018050  -24676679   72329617 1792675223
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.487e+07  9.080e+07   0.604    0.546
## Budget        2.531e+00  1.568e-01  16.142  < 2e-16 ***
## Runtime       4.913e+05  4.204e+05   1.169    0.243
## Rating       -5.891e+06  1.349e+07  -0.437    0.663
## Rating.Count  2.883e+02  3.858e+01   7.474 3.48e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 190600000 on 503 degrees of freedom
## Multiple R-squared:  0.5319, Adjusted R-squared:  0.5282
## F-statistic: 142.9 on 4 and 503 DF,  p-value: < 2.2e-16
```

*Figure 3 - Summary Statistics for Model 1*

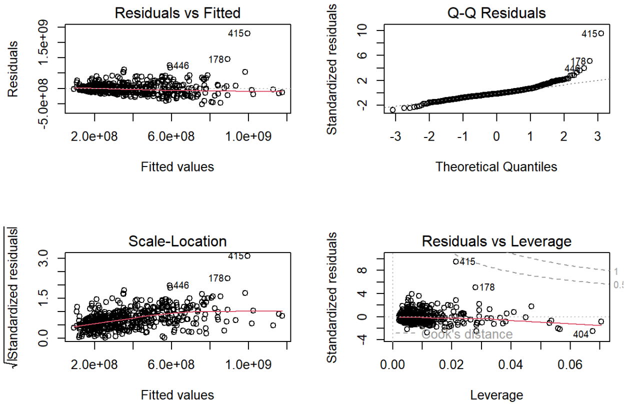


*Figure 4 - Diagnostic Plots for Model 3*

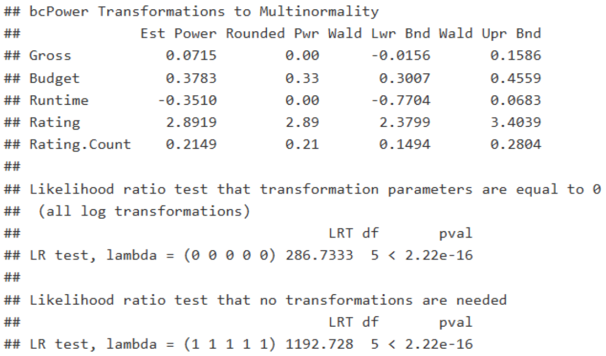
From the figure, we achieve the following equation:

$$\widehat{Gross} = 54870000 + 2.531(Budget) + 493100(Runtime) - 5891000(Rating) + 288.3(Rating.Count)$$

We can inspect that only two variables are significant predictors of *Gross*: *Budget* and *Rating.Count*. Since the other two predictor variables, *Runtime* and *Rating*, are not significant, it indicates that either transformations or variable selection are required. In addition, the residual standard error is very high, and the $R^2$ value is at a moderate 0.5319, which indicates that 53.19% of the variation in *Gross* is explained by all four predictors. Finally, the F-test indicates that at least one of the predictor coefficients is significant, signified by a p-value far less than the significance level of 0.05.

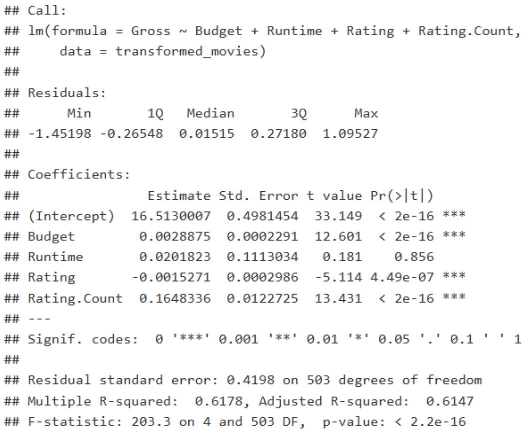
The diagnostic plots above show that our initial model had some flaws and violated some of the model assumptions. To begin, our Residuals vs. Fitted Plot appears to be centered around 0 and relatively straight, upholding the linearity assumption between the response and predictor variables. However, our Normal Q-Q Plot shows that the distribution of the error term is not normal due to the heavy tails and the deviance from the reference line, indicating that the normality assumption is violated. Our Scale-Location Plot is, more or less, randomly scattered which would uphold the constant variance assumption. However, there does seem to be some funneling near larger fitted values. Finally, the Residuals vs. Leverage Plot does show a number of outliers, leverage points, and potential influential points that would need to be addressed if we continued using this model. Since these assumptions are violated, we must transform the data to develop a more reliable and efficient model.

**Model 2**:
To develop the transformed model, we conducted a box-cox test simultaneously on both the response and predictor variables. The results are shown below.



```
## bcPower Transformations to Multinormality
##              Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Gross           0.0715        0.00      -0.0156       0.1586
## Budget          0.3783        0.33       0.3007       0.4559
## Runtime        -0.3510        0.00      -0.7704       0.0683
## Rating          2.8919        2.89       2.3799       3.4039
## Rating.Count    0.2149        0.21       0.1494       0.2804
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                    LRT df       pval
## LR test, lambda = (0 0 0 0 0) 286.7333  5 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                    LRT df       pval
## LR test, lambda = (1 1 1 1 1) 1192.728  5 < 2.22e-16
```

*Figure 5 - Box-Cox Transformation on all Variables*



```
## Call:
## lm(formula = Gross ~ Budget + Runtime + Rating + Rating.Count,
##     data = transformed_movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45198 -0.26548  0.01515  0.27180  1.09527
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.5130007  0.4981454  33.149  < 2e-16 ***
## Budget        0.0028875  0.0002291  12.601  < 2e-16 ***
## Runtime       0.0201823  0.1113034   0.181    0.856
## Rating       -0.0015271  0.0002986  -5.114 4.49e-07 ***
## Rating.Count  0.1648336  0.0122725  13.431  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4198 on 503 degrees of freedom
## Multiple R-squared:  0.6178, Adjusted R-squared:  0.6147
## F-statistic: 203.3 on 4 and 503 DF,  p-value: < 2.2e-16
```

*Figure 6 - Summary Statistics for Model 2*

From the box-cox transformation output, we can see that the p-value for the likelihood ratio test for only log transformation and no transformations was less than the significance level of 0.05, which indicates that only log transformations aren't recommended and that a transformation is necessary.
From the summary output, one can examine how these three predictors are significant. All variables except *Runtime* are consequential predictors as shown by a p-value less than 0.05 for their respective T-tests. In addition, the residual standard error is significantly lower than the previous model, and the R$^2$ value increased to 0.6178, which means that the predictor variables explain 61.78% of the variation in *Gross*. Finally, the F-test indicates that at least one of the predictor coefficients is significant given its p-value is much smaller than a significance level of 0.05.

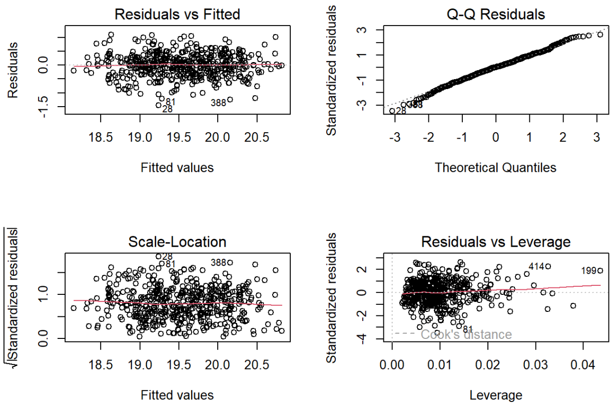
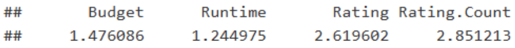


*Figure 7 - Diagnostic Plots for Model 2*

*VIF table*

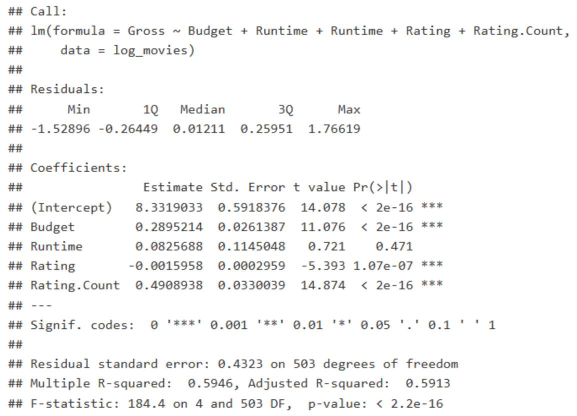


```
##       Budget      Runtime       Rating Rating.Count
##     1.476086     1.244975     2.619602     2.851213
```

*Figure 8 - VIF Output for Model 2*

As we can see, the model is valid. All the assumptions (linearity, normality of error terms, heteroscedasticity) are upheld as shown in the diagnostic plots. The VIF table also tells us that none of the variables are victims of multicollinearity. However, power transformations pose a challenge regarding interpretation. As such, we will not analyze this model's findings in-depth because we wanted to develop a model that would make our findings more interpretable and practical for the real world.

**Model 3**:
As noted earlier, we want to make our model more interpretable, so we will conduct log transformations to *Gross*, *Budget*, *Runtime*, and *Rating.Count* since the box-cox lambda values are all near 0. Below is the result:



```
## Call:
## lm(formula = Gross ~ Budget + Runtime + Runtime + Rating + Rating.Count,
##     data = log_movies)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52896 -0.26449  0.01211  0.25951  1.76619
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.3319033  0.5918376  14.078  < 2e-16 ***
## Budget       0.2895214  0.0261387  11.076  < 2e-16 ***
## Runtime      0.0825688  0.1145048   0.721    0.471
## Rating      -0.0015958  0.0002959  -5.393 1.07e-07 ***
## Rating.Count 0.4908938  0.0330039  14.874  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4323 on 503 degrees of freedom
## Multiple R-squared:  0.5946, Adjusted R-squared:  0.5913
## F-statistic: 184.4 on 4 and 503 DF,  p-value: < 2.2e-16
```

*Figure 9 - Summary Statistics for Model 3*

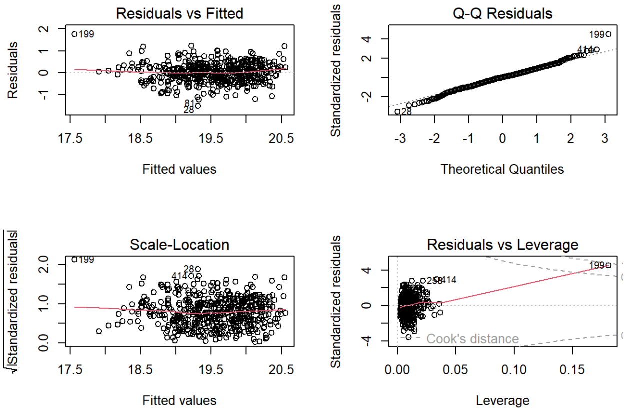


*Figure 10 - Diagnostic Plots for Model 3*

From the summary output, we get the equation:

$$log(\widehat{Gross}) = 8.33 + 0.290log(Budget) + 0.083log(Runtime) - 0.002(Rating)^{2.89} + 0.491log(Rating.Count)$$

The summary output, again, shows that *Runtime* is not a significant predictor in our new model. In addition, the residual standard error is significantly lower than **Model 1**, and the $R^2$ value increased to 0.5946, which means that the predictor variables explain 59.46% of the variation in *Gross*. Finally, the F-test indicates that at least one of the predictor coefficients is significant, signified by the p-value which is far less than the significance level of 0.05.

From the diagnostic plots, we can see that most of the assumptions are upheld (linearity and constant variance). However, one point that needs to be explored is Case 199, which is shown to deviate heavily in the Normal Q-Q Plot and the Residuals vs. Leverage plot. As a result, plotting Standardized Deviance Residuals vs. Leverage gives us the following graph:

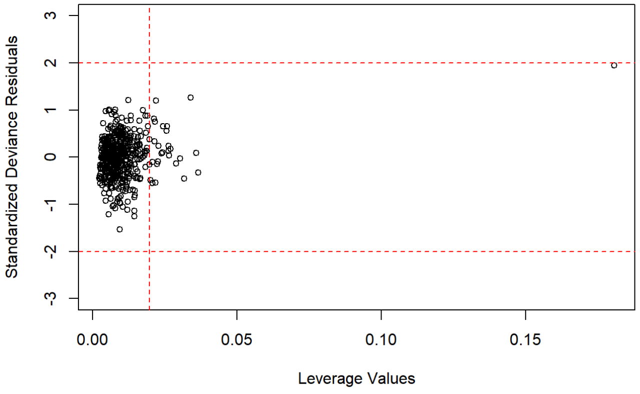


*Figure 11 - Standardized Deviance Residuals vs. Leverage Plot for Model 3*

As one can see, Case 199 is not contained within the top right nor the top left of the graph, which would indicate that it is a leverage point that negatively affects the regression line in our model. As a result, we can assert Case 199 is a good leverage point and follows the regression trend. We see that *The Blair Witch Project* is Case 199, a movie with a $60k budget and made a $250M profit, which would explain why it is a leverage point.

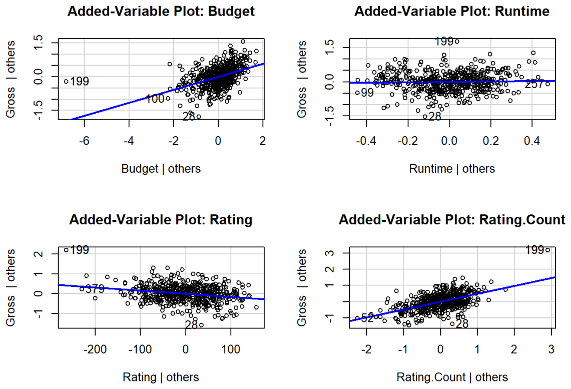
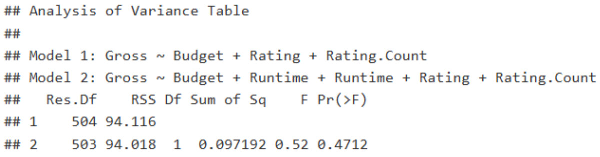


*Figure 12 - Added-Variable Plots for Model 3*



```
## Analysis of Variance Table
##
## Model 1: Gross ~ Budget + Rating + Rating.Count
## Model 2: Gross ~ Budget + Runtime + Runtime + Rating + Rating.Count
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1    504 94.116
## 2    503 94.018  1  0.097192 0.52 0.4712
```

*Figure 13 - ANOVA Test for Reduced Model 3 vs. Full Model 3*

From the Added-Variable plots and the ANOVA table's results—testing whether the reduced model (without Runtime) is better—we can see that Runtime is not a significant predictor and should be omitted in the final model. Therefore, our final model is the following:

$$log(\widehat{Gross}) = 8.629 + 0.294log(Budget) - 0.002(Rating)^{2.89} + 0.491log(Rating.Count)$$

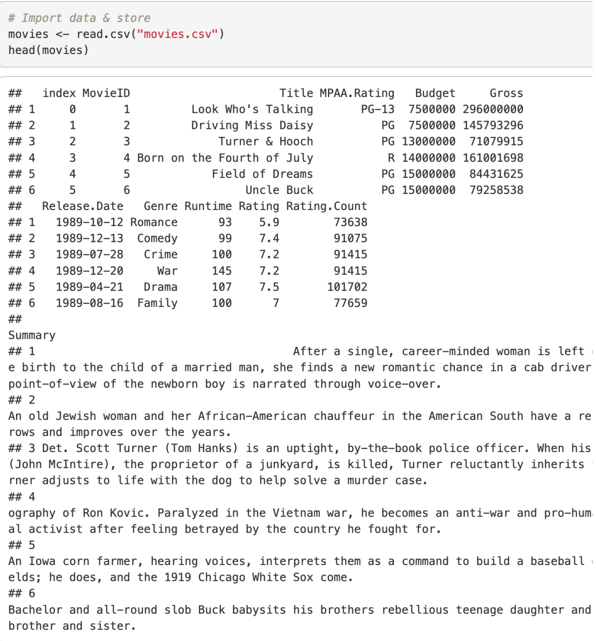
## DISCUSSION

Our final model indicates that a 1% increase in *Budget* results in a 0.29% increase in *Gross Revenue*, and a 1% increase in *Rating.Count* results in a 0.49% increase in *Gross Revenue*. However, due to the power transformation, the effects of *Rating* are difficult to interpret. Still, we see that *Rating* has a small negative effect on *Gross Revenue*. Therefore our final model shows that increasing *Budget* and *Rating.Count* is the most effective way to increase *Gross Revenue*.

These findings are reasonable as films with a larger budget are able to make larger productions and can promote more. Yet, as seen by *Rating* hurting *Gross Revenue* films can be overrated which is why *Rating.Count* is a better predictor. We can speculate that the higher the *Rating.Count*, the more people are willing to rate the movie, thus speaking to how the movie was extremely satisfactory or disappointing.

One of the main challenges we faced was that no single variable has a strong statistical relationship with revenue. This meant that our predictors individually had limited explanatory power, making it harder to build a highly predictive model. To improve model performance, we had to apply log and power transformations to our predictors. While this helped meet regression assumptions, it also introduced interpretation challenges because the coefficients are no longer in their original units.

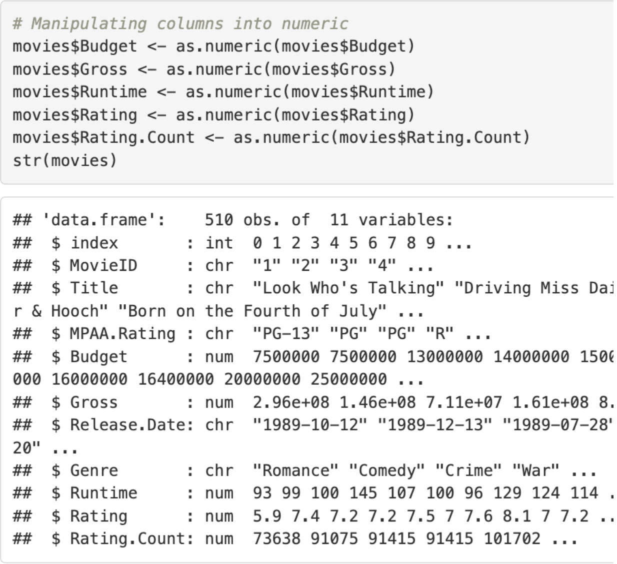
# APPENDIX A



```
# Import data & store
movies <- read.csv("movies.csv")
head(movies)
```

```
##   index MovieID                      Title MPAA.Rating   Budget     Gross
## 1     0       1         Look Who's Talking       PG-13  7500000 296000000
## 2     1       2         Driving Miss Daisy          PG  7500000 145793296
## 3     2       3             Turner & Hooch          PG 13000000  71079915
## 4     3       4 Born on the Fourth of July           R 14000000 161001698
## 5     4       5            Field of Dreams          PG 15000000  84431625
## 6     5       6                 Uncle Buck          PG 15000000  79258538
##   Release.Date   Genre Runtime Rating Rating.Count
## 1   1989-10-12 Romance      93    5.9        73638
## 2   1989-12-13  Comedy      99    7.4        91075
## 3   1989-07-28   Crime     100    7.2        91415
## 4   1989-12-20     War     145    7.2        91415
## 5   1989-04-21   Drama     107    7.5       101702
## 6   1989-08-16  Family     100      7        77659
##
Summary
## 1                                          After a single, career-minded woman is left
e birth to the child of a married man, she finds a new romantic chance in a cab driver
point-of-view of the newborn boy is narrated through voice-over.
## 2
An old Jewish woman and her African-American chauffeur in the American South have a re
rows and improves over the years.
## 3 Det. Scott Turner (Tom Hanks) is an uptight, by-the-book police officer. When his
(John McIntire), the proprietor of a junkyard, is killed, Turner reluctantly inherits
rner adjusts to life with the dog to help solve a murder case.
## 4
ography of Ron Kovic. Paralyzed in the Vietnam war, he becomes an anti-war and pro-hum
al activist after feeling betrayed by the country he fought for.
## 5
An Iowa corn farmer, hearing voices, interprets them as a command to build a baseball
elds; he does, and the 1919 Chicago White Sox come.
## 6
Bachelor and all-round slob Buck babysits his brothers rebellious teenage daughter and
brother and sister.
```

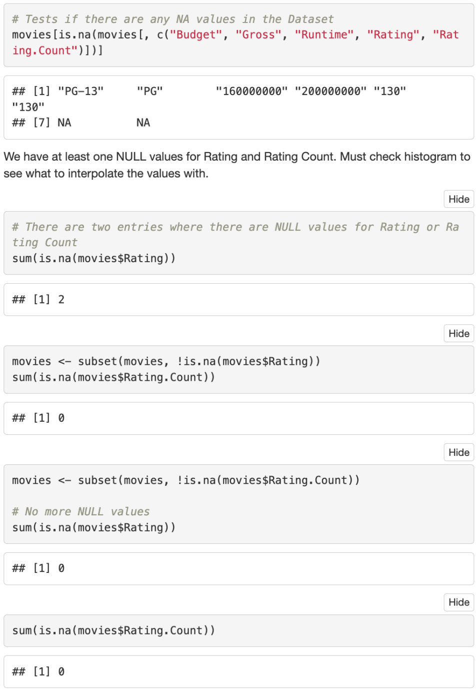
*Figure A1 - head() output of Full Dataset*



```
# Manipulating columns into numeric
movies$Budget <- as.numeric(movies$Budget)
movies$Gross <- as.numeric(movies$Gross)
movies$Runtime <- as.numeric(movies$Runtime)
movies$Rating <- as.numeric(movies$Rating)
movies$Rating.Count <- as.numeric(movies$Rating.Count)
str(movies)
```

```
## 'data.frame':    510 obs. of  11 variables:
##  $ index       : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ MovieID     : chr  "1" "2" "3" "4" ...
##  $ Title       : chr  "Look Who's Talking" "Driving Miss Dai
r & Hooch" "Born on the Fourth of July" ...
##  $ MPAA.Rating : chr  "PG-13" "PG" "PG" "R" ...
##  $ Budget      : num  7500000 7500000 13000000 14000000 150
000 16000000 16400000 20000000 25000000 ...
##  $ Gross       : num  2.96e+08 1.46e+08 7.11e+07 1.61e+08 8.
##  $ Release.Date: chr  "1989-10-12" "1989-12-13" "1989-07-28"
20" ...
##  $ Genre       : chr  "Romance" "Comedy" "Crime" "War" ...
##  $ Runtime     : num  93 99 100 145 107 100 96 129 124 114 .
##  $ Rating      : num  5.9 7.4 7.2 7.2 7.5 7 7.6 8.1 7 7.2 ..
##  $ Rating.Count: num  73638 91075 91415 91415 101702 ...
```

*Figure A2 - Converting Data types of variables*



```
# Tests if there are any NA values in the Dataset
movies[is.na(movies[, c("Budget", "Gross", "Runtime", "Rating", "Rat
ing.Count")])]
```

```
## [1] "PG-13"     "PG"        "160000000" "200000000" "130"
"130"
## [7] NA          NA
```

We have at least one NULL values for Rating and Rating Count. Must check histogram to see what to interpolate the values with.

```
# There are two entries where there are NULL values for Rating or Ra
ting Count
sum(is.na(movies$Rating))
```

```
## [1] 2
```

```
movies <- subset(movies, !is.na(movies$Rating))
sum(is.na(movies$Rating.Count))
```

```
## [1] 0
```

```
movies <- subset(movies, !is.na(movies$Rating.Count))

# No more NULL values
sum(is.na(movies$Rating))
```

```
## [1] 0
```

```
sum(is.na(movies$Rating.Count))
```

```
## [1] 0
```

*Figure A3 - Identifying and Removing Null Values*

## APPENDIX B

1. “Top Lifetime Grosses by MPAA Rating.” *Box Office Mojo*, IMDb.com, https://www.boxofficemojo.com/chart/mpaa_title_lifetime_gross/?area=XWW&by_mpaa=R, Accessed 9 Mar. 2025.
2. “Box Office 2025 Predictions: Will ‘Superman,’ ‘Jurassic World 4’ and ‘Wicked 2’ Rule Theaters?”, Rubin, Rebecca, 2025, https://variety.com/2025/film/box-office/box-office-2025-predictions-superman-jurassic-world-4-wicked-2-1236261990/. Accessed 10 Mar. 2025.