Paige Johnston and Jenn Hwang

Final Project

Dr.Gabrosek

STA 321

December 16, 2022

Paige Johnston
Jenn Jung-Hye Hwang

Table of Contents

Paige Johnston
Jenn Jung-Hye Hwang

Introduction

Research Question: What variable(s) influence the revenue for a movie?
Response Variable: Movie revenue

Variables we believe could affect the revenue for a movie: Before any reading or research we assumed that revenue could be influenced by genre, runtime, and voting average. The more popular the movie genre is, such as Marvel superhero blockbusters and action movies, the more money they make. Also, it is not uncommon to look at ratings from sites such as rotten tomato and letterboxd before going into a movie, which is why it makes sense to study voting averages.

After further research we've decided to also look at language and release date. We've decided to look at language because not all languages are widely spoken and may have a limited appeal and generate lower revenue. Release date is also an interesting variable to look into, as holidays, school breaks, and other seasonal factors may affect the movie's revenue.

https://www.researchgate.net/publication/281730174_The_determinants_of_box_office_performance_in_the_film_industry_revisited
https://www.inquirer.com/philly/entertainment/20100509_Americans_are_seeing_fewer_and_fewer_foreign_films.html
https://www.jstor.org/stable/25046296

Variable Name	Variable Type	Possible Categories
Genre	Categorial	Music, Romance, Family, War, Foreign, TV Movie, Adventure, Fantasy, Animation, Drama, Horror, Comedy, History, Western, Thriller, Crime, and Science Fiction
Runtime	Quantitative	-
Language	Categorical	Africans, Chinese, Danish, German, English, Spanish, French, Hindi, Magyar, Italian, Japanese, Korean, Portuguese, Romanian, Russian, Swedish, Turkish, Mandarin, Other
Vote Average	Quantitative	-
Release Date	Categorical	Year and month the movie was released
Popularity	Quantitative	-
Vote Count	Quantitative	-
Status	Categorical	Released, Post Production. Rumored

Table 1: List of all potential predictor variables

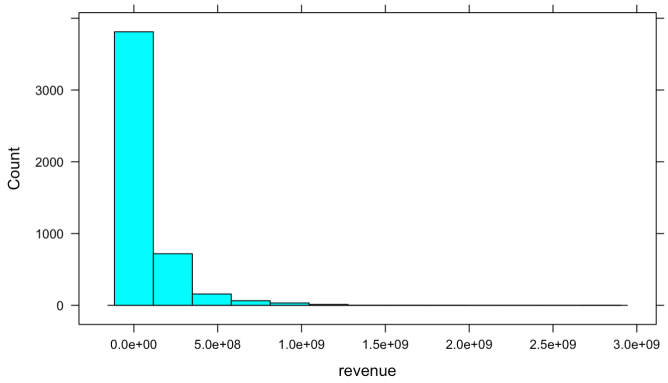
The variable genre is the type of movie produced. Run time is the length of time the movie plays for in minutes. Language represents the language that the movie was produced in. Table 1 shows the amount of languages present. Vote average is a measure of quality with no given units, but the higher the number the higher the quality of the movie. Release date is what time of year and what year the movie was released on.

The data does abide by the rules of having 50 or more observations per predictor, the response variable is quantitative

Modeling- Exploratory Data Analysis

Univariate EDA:
We are going to begin the exploratory data analysis by observing each categorical/quantitative predictor variable, as well as the response variable, to see its distribution. By doing this, we can gauge whether or not it is necessary to transform our data from looking at how normal the distributions look.

Y variable: Revenue:
The distribution of Revenue is uni-model and very right-skewed. This means that there are few outliers with a few movies who have a much higher revenue than the majority of the movies.



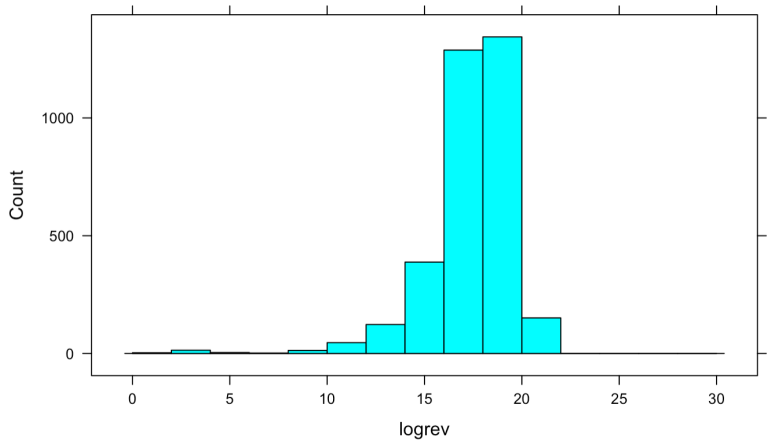
min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0	0	19170001	92917187	2787965087	82260639	162857101	4803	0

1 row

Log of Y variable: Revenue

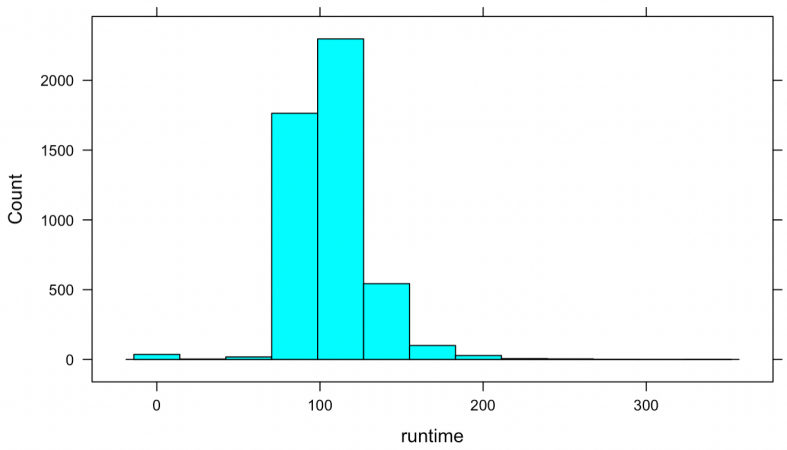
Paige Johnston
Jenn Jung-Hye Hwang

By taking the log of the Y variable Revenue, we are essentially pulling back its larger values towards the smaller ones. By transforming those small values, we are able to analyze the distribution better.



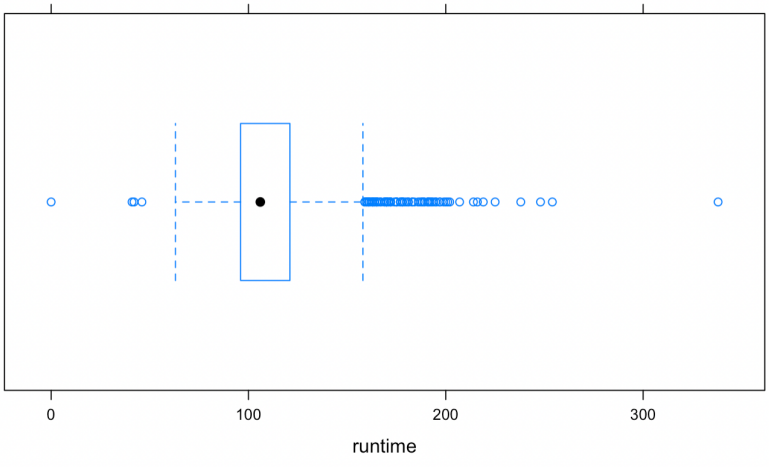
Runtime:

The distribution of Runtime is unimodal and has a right skew. This means that there are few outliers with longer runtimes who have a lower runtime than the majority of the films.



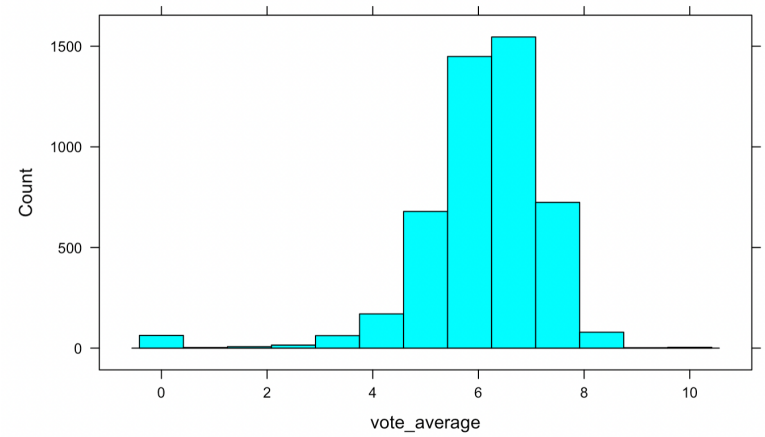
From the boxplot below, we are able to see a clearer representation of the outliers mentioned above. Outstanding outliers appear for runtimes of longer than 200 minutes.

Paige Johnston
Jenn Jung-Hye Hwang



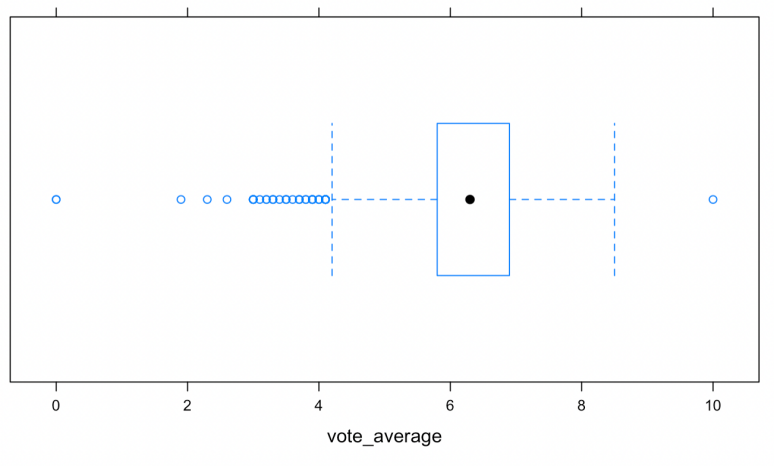
Vote Average:

The distribution of Vote Average is uni-model with a left skew. There are a few outliers with voting averages being less than 4, causing the data to be skewed left.

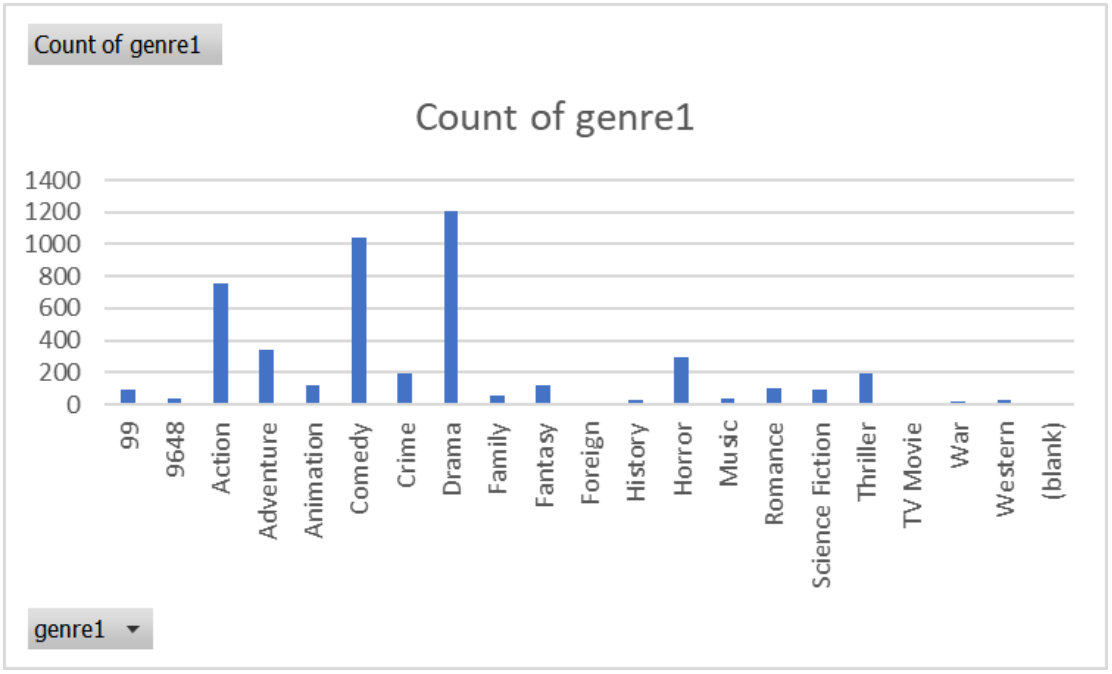


From the boxplot below, we are able to see a clearer representation of the outliers mentioned above. Outstanding outliers appear for points where vote_average is less than 4.

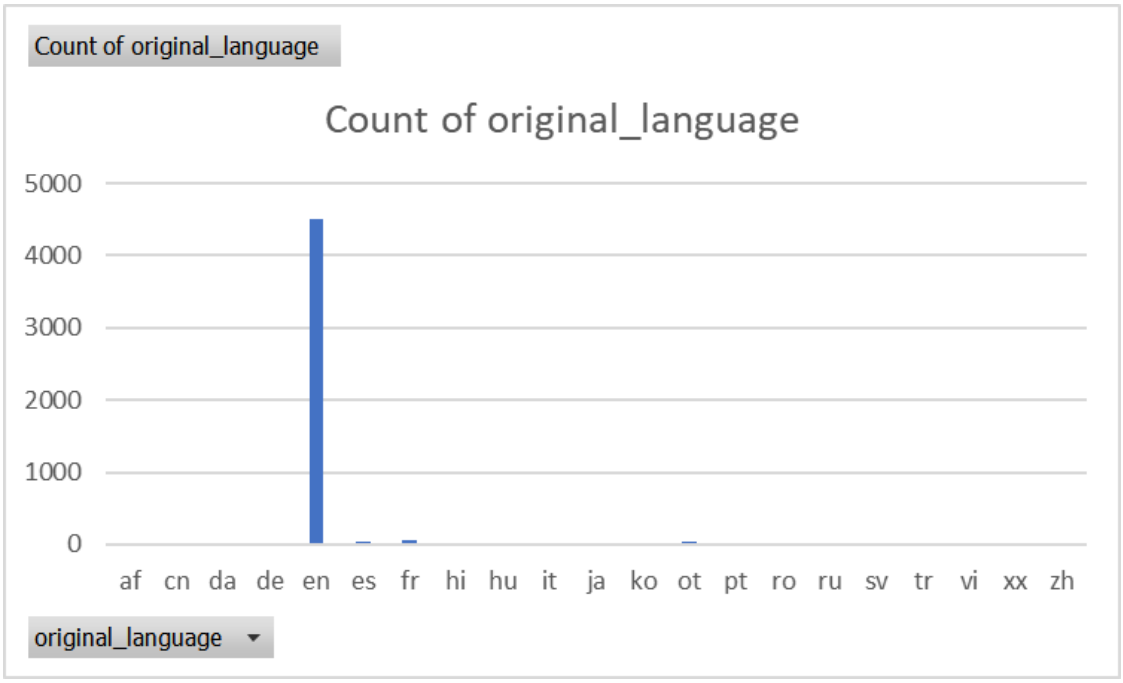
Paige Johnston
Jenn Jung-Hye Hwang



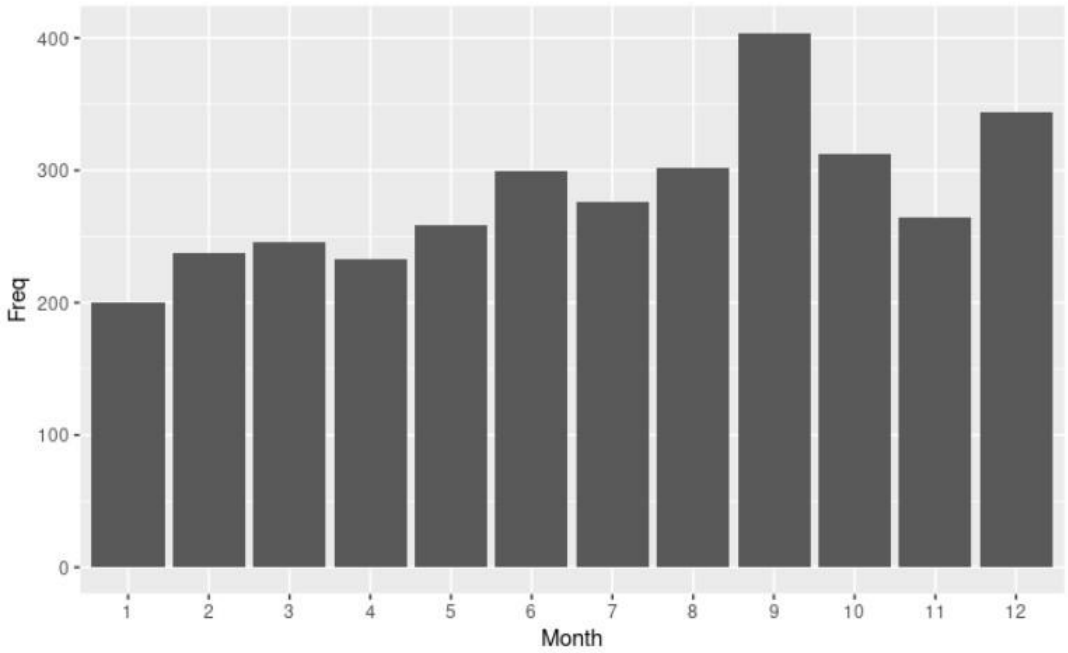
Genre



Language:



Release Month:



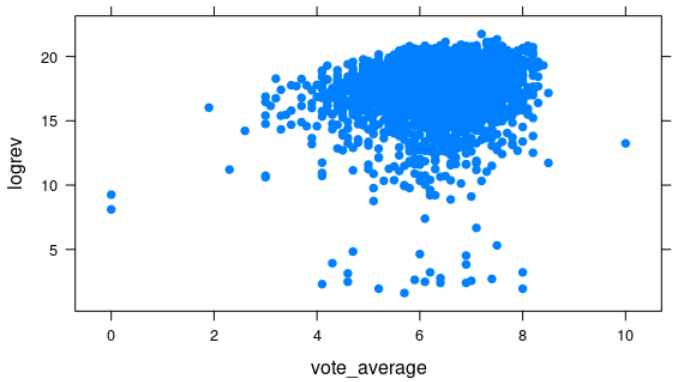
Bivariate EDA:

Paige Johnston
Jenn Jung-Hye Hwang

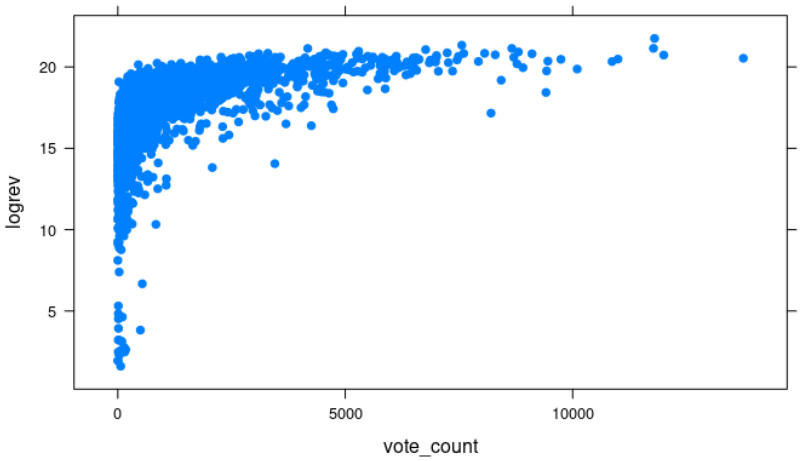
Based off of the univariate EDA we decided to transform revenue by taking it's log. The next step is to explore the relationship between the logrev and the predictors. Below are the two quantitative variables.

There does appear to be a linear relationship between run time and logrev. There are some outliers in the data that are shaping the scatterplot by a lot.

There does seem to be a linear relationship between logrev and vote average.

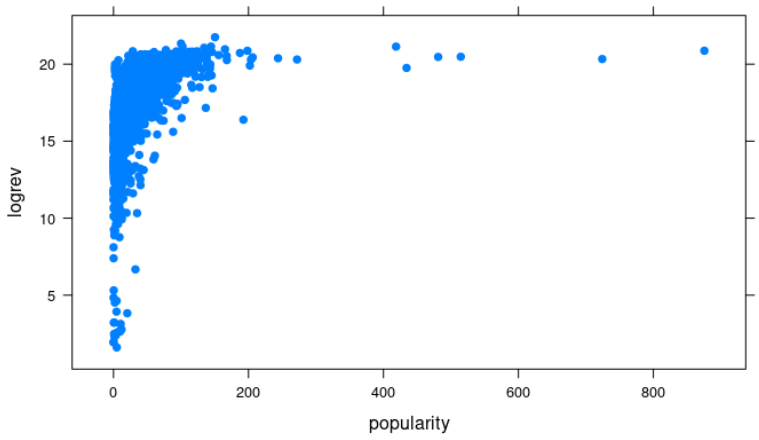


The relationship between log rev and vot count does not seem to be linear but this could be due to outliers.



The scatterplot for log rev and popularity is very similar to the scatter plot above.

Paige Johnston
Jenn Jung-Hye Hwang



Modeling- Full Model Specification

The variables are renamed with simple notation for this section.

Notation	Variable	Notation	Variable	Notation	Variable
Y	Revenue	X1	Runtime	X_2	Budget
X_3	Vote Average	X_4	Popularity	X5	Vote Count
X6	Year	X7	Month	X8	Day
$L1.e$	English	$L1.f$	French	$G1.d$	Drama
$G1.c$	Comedy	$G1.a$	Action	$G1.h$	Horror

Table 2: Table of predictors

Before making the full model, we've created a dummy variable for the categorical variables genre and language. Genre involves five corresponding dummy variables, containing drama, comedy, action, horror and if all 0 it is other. Language involves two corresponding dummy variables, containing English, French, and other which is every other language besides English and French which is 0 for both variables.

Interactions

It would be interesting to see if runtime and voting average would have an impact on the amount of revenue a movie receives. The interaction variable is created, and is denoted as I.1.

I.1 = vote_average_*runtime

Full Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 L_{1.e} + \beta_{10} L_{1.f} + \beta_{11} G_{1.d} + \beta_{12} G_{1.c} + \beta_{13} G_{1.a} + \beta_{14} G_{1.h} + \beta_{15} I.1 + \varepsilon$$

Variable selection

The full model currently has a lot of predictor variables for predicting the revenue for a movie, and assuming that not every variable will be significant the model needs to be reduced. It is important that we start by checking the assumptions for the full model. For example checking for normality, homogeneity of variances, and linearity. We want to use stepwise selection, along with forward and backward selection to fit the reduced model with the variables to determine revenue.

We will also utilize overall and multiple partial F-tests to see the significance of the reduced model as we narrow down the variables.

Paige Johnston
Jenn Jung-Hye Hwang

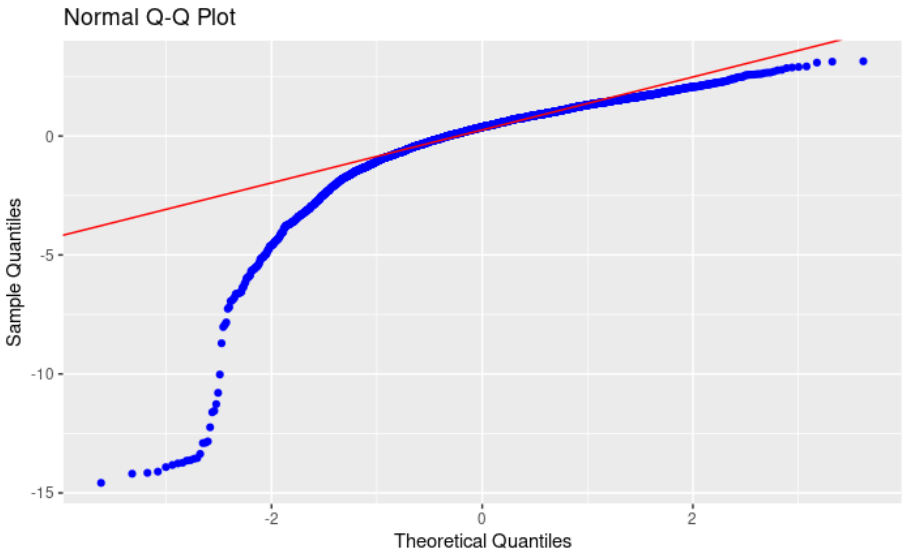
Modeling fitting

To start off we want to fit the full model into a regression equation. We accomplished this by using the anova function results are below:

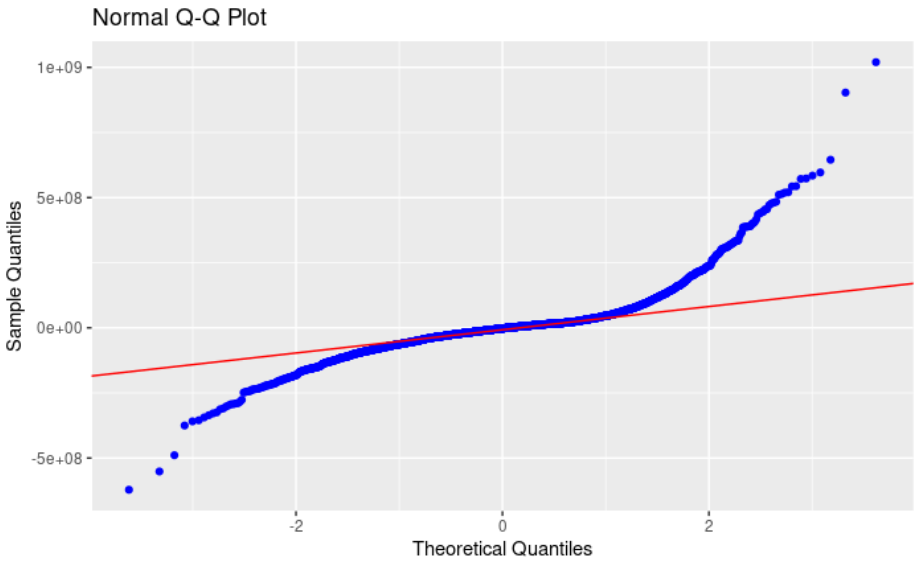
```
Residuals:
     Min       1Q   Median       3Q      Max
-14.7277  -0.5289   0.3606   1.0224   3.0832

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.758e+01  5.261e+00   3.342 0.000840 ***
runtime       5.178e-03  1.732e-03   2.990 0.002807 **
budget        1.933e-08  9.482e-10  20.390  < 2e-16 ***
vote_average  2.158e-01  4.453e-02   4.845 1.32e-06 ***
popularity    6.748e-03  1.318e-03   5.119 3.24e-07 ***
vote_count    2.019e-04  3.880e-05   5.204 2.07e-07 ***
L1.e          8.114e-01  1.743e-01   4.654 3.38e-06 ***
L1.f          9.865e-02  3.773e-01   0.261 0.793747
G1.d         -3.224e-01  8.910e-02  -3.618 0.000301 ***
G1.c          3.109e-01  9.106e-02   3.414 0.000647 ***
G1.a          3.409e-02  9.191e-02   0.371 0.710761
G1.h          4.758e-01  1.405e-01   3.387 0.000716 ***
year         -2.088e-03  2.575e-03  -0.811 0.417462
month         2.727e-03  9.356e-03   0.291 0.770715
day           6.275e-03  3.629e-03   1.729 0.083841 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.788 on 3360 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3428,    Adjusted R-squared:   0.34
F-statistic: 125.2 on 14 and 3360 DF,  p-value: < 2.2e-16
```

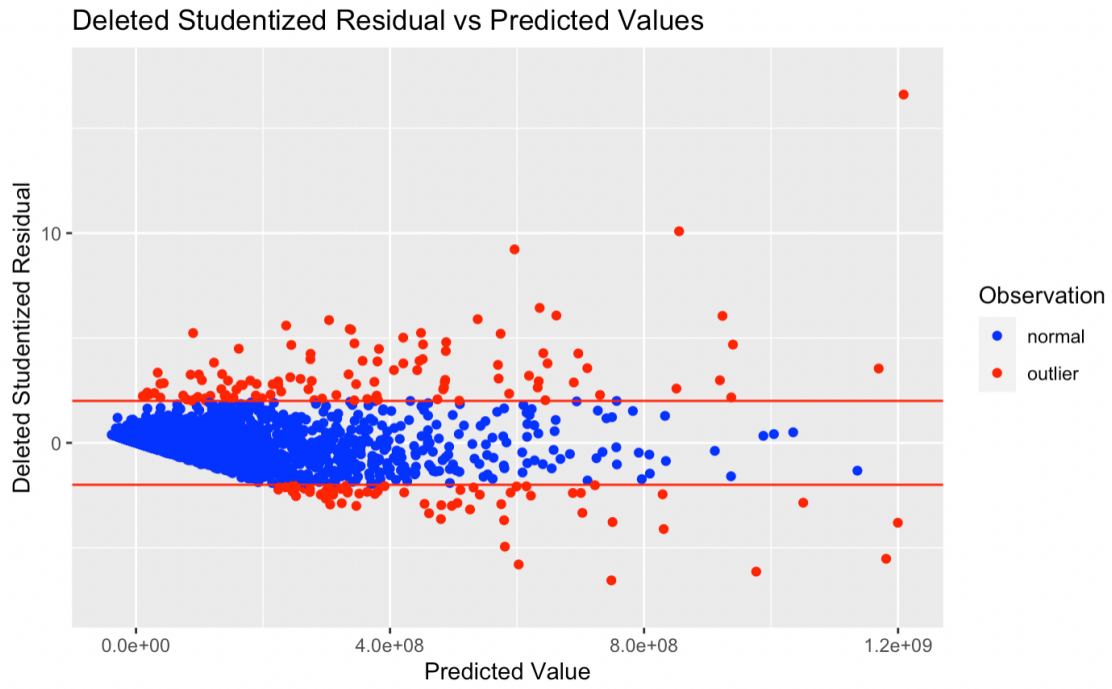
Full Model Residuals:

The full model's normality is not met, there is a high level of skewness on the left tail of the Q-Q plot. This leads us to believe that the data needs to be transformed



The Q-Q plot below shows the turning back logrevenue to revenue, where we were able to see less of a skew on the left tail end of the plot. From this, we came to the conclusion that our reduced model we should also reverse logrevenue back to revenue.





Forward Selection:

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	budget	0.6989	0.6987	72.7601	134011.7022	100722312.3790
2	popularity	0.7021	0.7018	38.3320	133977.7693	100202569.0390
3	vote_count	0.7042	0.7038	17.1379	133917.0161	99875503.8300
4	year	NA	NA	NA	NA	NA
5	G1.a	NA	NA	NA	NA	NA

Backwards Selection:

Paige Johnston
Jenn Jung-Hye Hwang

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	G1.h	0.706	0.7048	13.0061	133912.8547	99696073.4980
2	L1.f	0.706	0.7049	11.0766	133910.9255	99682291.0508
3	runtime	0.7059	0.705	9.2980	133909.1479	99670752.8543
4	day	0.7058	0.705	8.2770	133908.1310	99670453.5903
5	L1.e	0.7057	0.7049	7.5550	133907.4140	99674585.8237
6	month	0.7055	0.7048	8.0075	133907.8747	99696116.5778
7	vote_average	0.7053	0.7047	8.4240	133908.2975	99717094.8636
8	G1.c	0.705	0.7045	9.6176	133909.4968	99749557.7058
9	G1.d	0.7042	0.7038	17.1379	133917.0161	99875503.8300

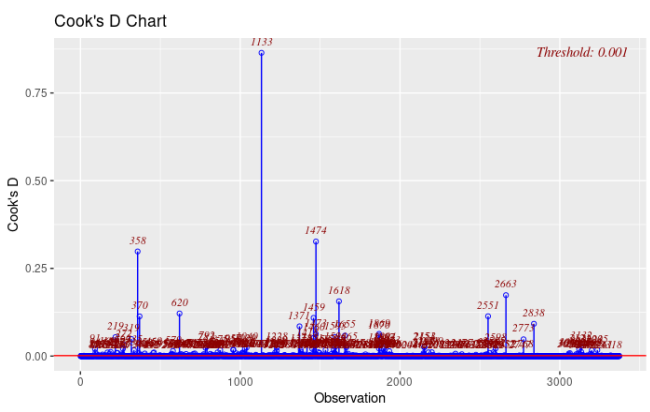
Step-wise Selection:

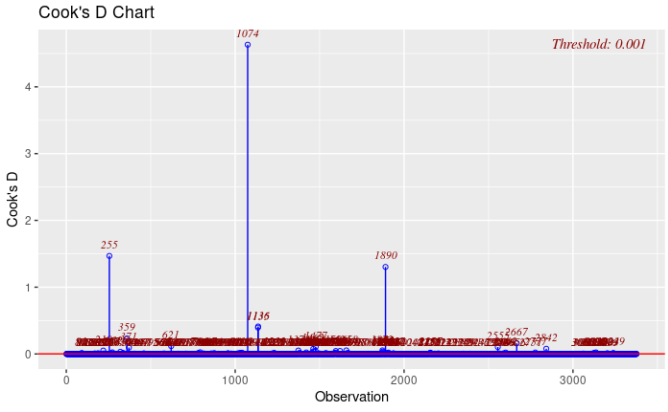
Stepwise Selection Summary

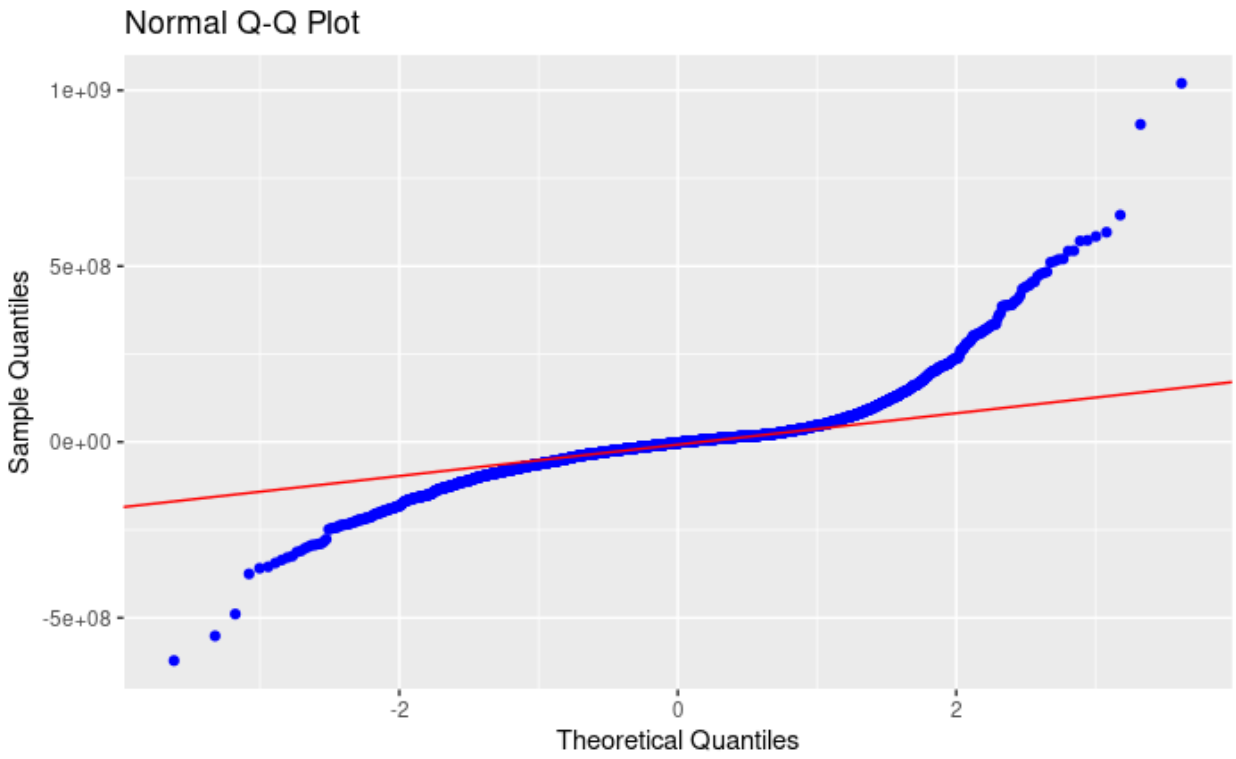
Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	budget	addition	0.699	0.699	72.7600	134011.7022	100722312.3790
2	popularity	addition	0.702	0.702	38.3320	133977.7693	100202569.0390
3	vote_count	addition	0.704	0.704	17.1380	133917.0161	99875503.8300

Findings:

Budget, popularity, and vote_count remain in all three of the processes which is evidence that they will be placed into the reduced model. In forward selection year and G1.a are included but they aren't in the other two processes therefor we will not consider them in the model. We decided to go with the step-wise selection model because it gave the highest R-squared value.







The assumptions are still violated because of the skewness on both tails of the Q-Q plot. So some data modification is needed in order to meet these assumptions. Based on this we don't have approval to make the reduced and final model. The cooksd plot shows us the influential points in this case we had to removed 4 data points:1074,255,1136,1890.

Final Model:

$$Y = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Fitted Regression Equation:

$$\hat{Y} = -(1.906e + 07) + (1.712)X_2 + (4.200e + 5)X_4 + (6.079e + 04)X_5$$

Model Usage

The final model for predicting the revenue of a movie consists of budget, popularity, and vote count. The R-squared value is on the higher end with a .7069. This means that 70.69% of the variation in movie revenue is explained by the final model. Another way to interpret the model is by predicting the regression coefficients. For example by looking at the predictor budget, which is the amount of money that is allocated for the production of the movie. The predicted movie revenue will increase by 1.712 dollars.

As an example I am going to predict movie revenue of observation 103: A Cinderella Story (ACS). ACS had a budget of 19,000,000, a popularity score of 25.281320 and a vote count of 713. The regression equation would look like this:

$$\hat{Y} = -(1.906e + 07) + (1.712)(19,000,000) + (4.200e + 5)(25.281320) + (6.079e + 04)(713)$$

$\hat{Y}$ = 67,429,424.4

From this equation it tells us that the predicted revenue is 67,429,424.4 dollars. The residual is 70,067,909 - 67,429,424.4 = 2,638,484.6

There are a couple reasons why the predicted revenue undershoots the actual revenue. One being that there could be bias on who the actors are or who directed it. Another might be unforeseen circumstances. This is when an unexpected event occurs that might affect the revenue for a movie. For example, if a movie were to be released at the same time as a natural disaster, it might not perform as well as expected. Lastly, major movie studios may hold pre-screenings for film-critics. If the movie isn't well-received it may not perform as well as expected.