Final Project Step 3

PSTAT126: Regression Analysis

Ali Abuzaid

STUDENT NAMES AND NETID

- AARTI GARAYE (aartigaraye)
- MOIRA KEATING (mskeating)
- YIWEN XIAO (yiwenxiao)
- SYLVIA LI (sylvia_li)
- SIDDHARTH SINGH (siddharthsingh)

1.1 Regression Model Application

1.1.0 Our Objective and Hypotheses

Research Objective

How can artists use our model to strategically structure their music and plan the release dates to maximize their streams?

Research Questions

Question 1

How does the number of playlists a song is in influence the number of streams?

Question 2

Out of acousticness %, number of playlists, bpm, spotify charts, which variables are statistically significant for our model to predict the number of streams?

Question 3

How does release month impact the number of streams?

Hypotheses

Hypothesis 1

Null Hypothesis: H_0: $\beta_1 = 0$ The number of playlists a song is in (predictor variable) and the number of streams (response variable) have no linear relationship.

Alternate Hypothesis: H_A: $\beta_1 \neq 0$ The number of playlists a song is in (predictor variable) and the number of streams (response variable) have some linear relationship.

Hypothesis 2

Null Hypothesis: H_0: $\beta_1 = 0, \beta_2 = 0, ..., \beta_p = 0$ None of the variables listed have a statistical impact on the number of streams.

Alternate Hypothesis: H_A: $\beta_1 \neq 0, \beta_2 \neq 0, ..., \beta_p \neq 0$ At least one of the variables listed above have a statistical impact on the number of streams.

Hypothesis 3

Null Hypothesis: H_0: $\beta_1 = 0, \beta_2 = 0, ..., \beta_p = 0$ None of the months are statistically significant to impact the number of streams.

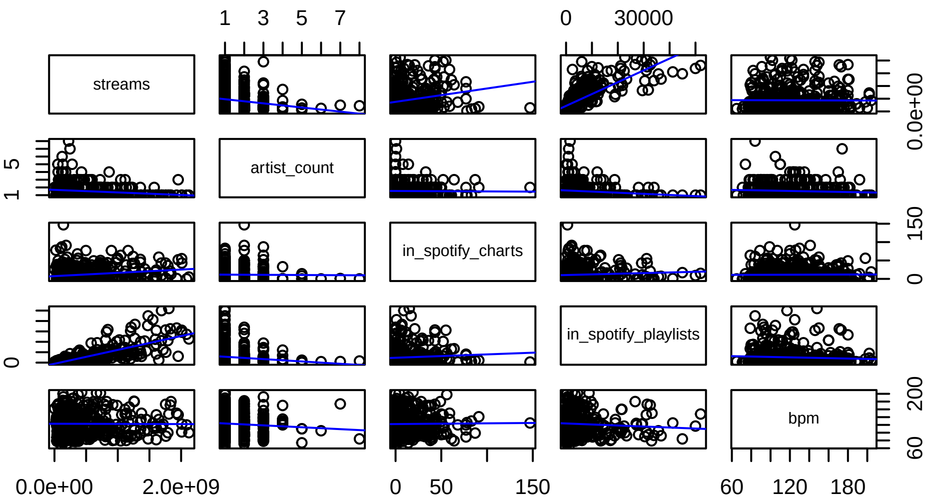
Alternate Hypothesis: H_A: $\beta_1 \neq 0, \beta_2 \neq 0, ..., \beta_p \neq 0$ At least one of the months are statistically significant to impact the number of streams.

Introduction of Dataset

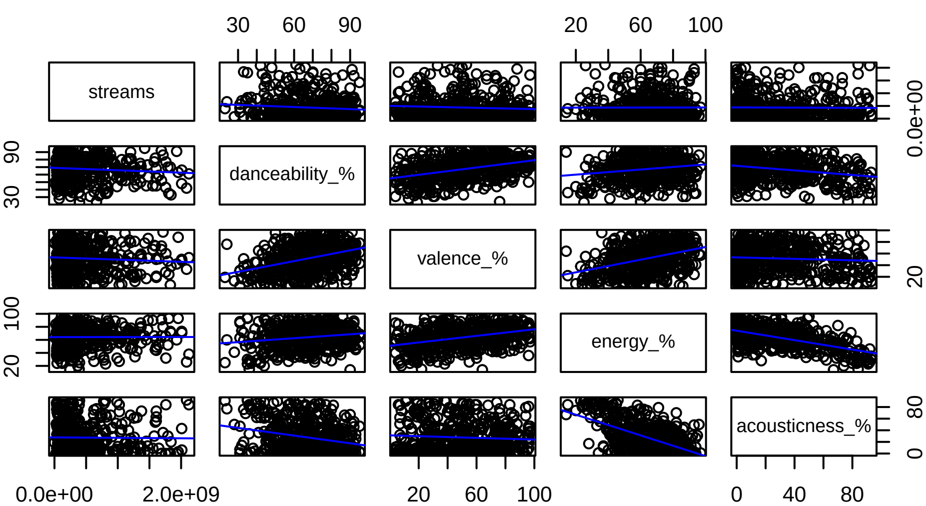
Variables

- Artist_count tells us the number of artists that have contributed to the song.
- Released_year, released_month, and released_day tell us the year, month, and day the song is released.
- in_spotify_playlists and in_spotify_charts tell us the number of Spotify playlists the song is in and its rank in charts.
- Streams is the total number of streams the song has in spotify. BPM tell us the tempo of the song by measuring the number of beats per minute.
- danceability_% represents the suitability of the song for dancing
- valence_% is the positivity of the song's musical content
- energy_%: is the perceived energy level of the song
- acousticness_% measures acoustic sound presence in the song
- instrumentalness_% measures the proportion of instrumental content in the track
- liveness_% tell us the presence of live performance elements
- speechiness_% measures the number of words spoken in a song

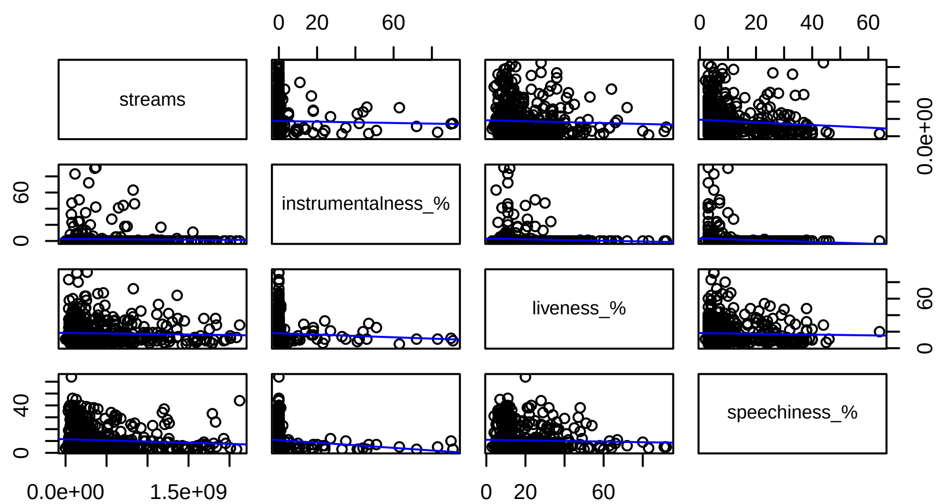
e 1.1: Scatterplot Matrix Comparing Streams and Predictor Variables with Regressio



e 1.2: Scatterplot Matrix Comparing Streams and Predictor Variables with Regressio



e 1.3: Scatterplot Matrix Comparing Streams and Predictor Variables with Regressio



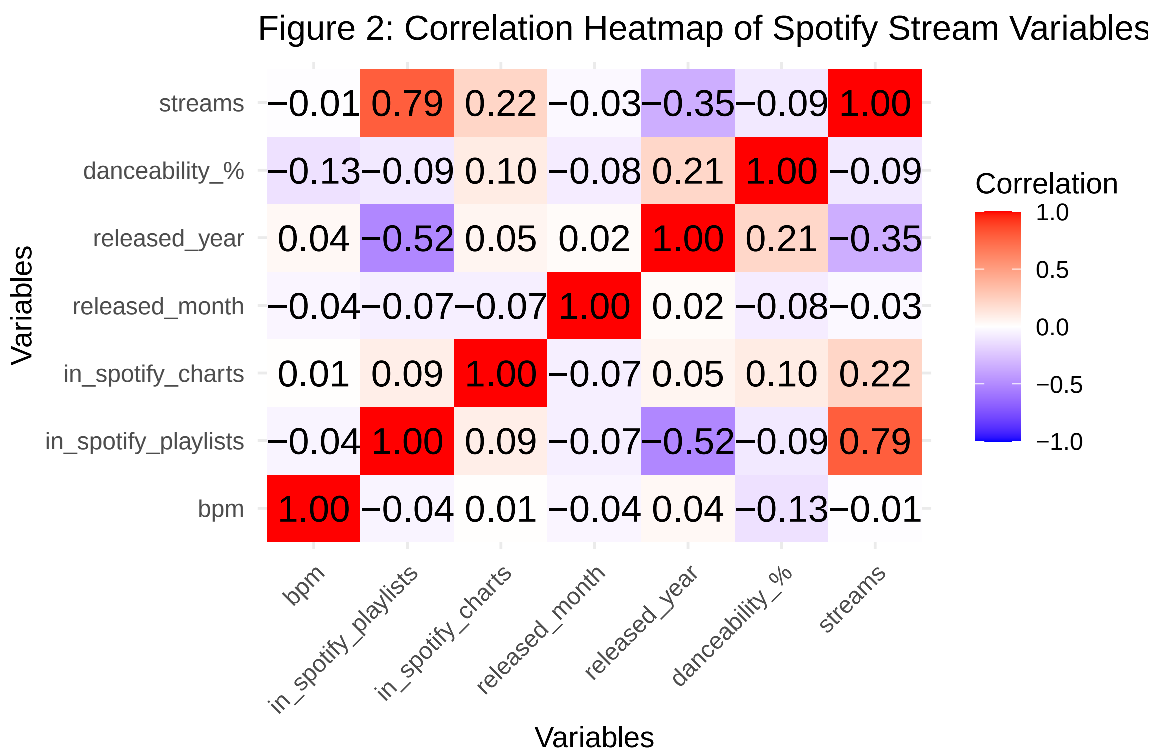


Figure 1: The positive relationship between **streams** and **in_spotify_playlists** suggests that being in more playlists and high streams are closely related, with the most notable correlation. **Acousticness_%** and **bpm** show very weak correlations to streams and **in_spotify_charts** shows a positive correlation but it is still quite weak. There appears to be a somewhat negative correlation between **released_year** and number of streams but it's still relatively weak.

For our research question 1, "How does the number of playlists a song is in influence the number of streams?," we analyzed the correlation heatmap and chose to interpret the relationship between streams and in_spotify_playlists. This was based on the fact that the relationship showed the most correlation among the included variables, suggesting that there could be some significance explained in a simple regression model.

Model 1: Simple Linear Regression

Streams = 217,699,474 + 46,503(in_spotify_playlists)

```r
model1 <- lm(streams ~ in_spotify_playlists, data = spotify_data)
summary(model1)
```

```

Call:
lm(formula = streams ~ in_spotify_playlists, data = spotify_data)

Residuals:
       Min         1Q     Median         3Q        Max 
-852220670 -154179262  -68303547   95697840 1453377238 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          217699474   14520088   14.99   <2e-16 ***
in_spotify_playlists     46503       1650   28.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 271100000 on 487 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.6198,	Adjusted R-squared:  0.619 
F-statistic: 793.8 on 1 and 487 DF,  p-value: < 2.2e-16
```

Interpretation: The simple linear regression model for the number of playlists a song is in as a predictor for the number of streams as the response has an intercept of 217,699,474, reflecting that a song in no Spotify playlists is predicted to have a base 217,699,474 streams. The coefficient is 46,503, representing that for every additional playlist a song is added to, there is an expected increase of 46,503 streams.

Model Fit: This model has an R-squared value of 0.6198, indicating that the model explains 61.98% of the variability in the number of streams. This is more than half of the variability, although some is still unexplained and may be caused due to other variables.

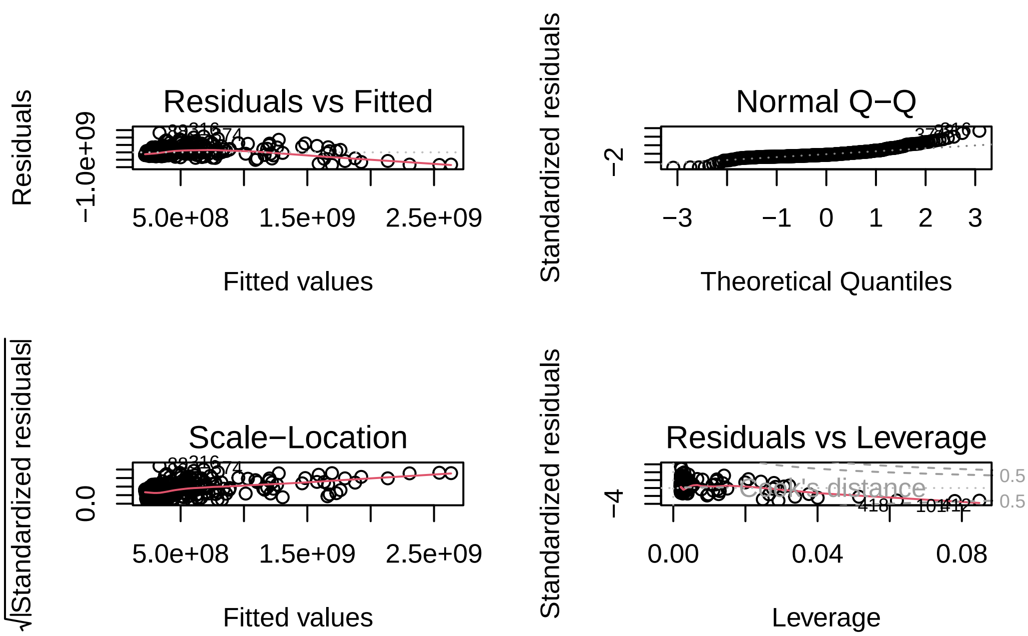
Statistical Significance: The coefficient has a p-value of <2e-16 which is statistically significant at a 95% significance level, since it is much less than $\alpha = 0.05$. The model has a p-value of 2.2e-16, so it is statistically significant overall because it is much lower than $\alpha = 0.05$.

Just by looking at the simple linear regression model, we can tell that the overall p-value of the linear model is less than the alpha significance level of 0.05. This means that the model is statistically significant, however, we haven't checked the assumptions about the error and that's why we can't fully comment on the significance of the model and the predictor variables.

Variable Selection/Interaction Terms/Complex Regressors:

We skipped this since we are only analyzing one variable in relation to streams.

Diagnostic Checking:

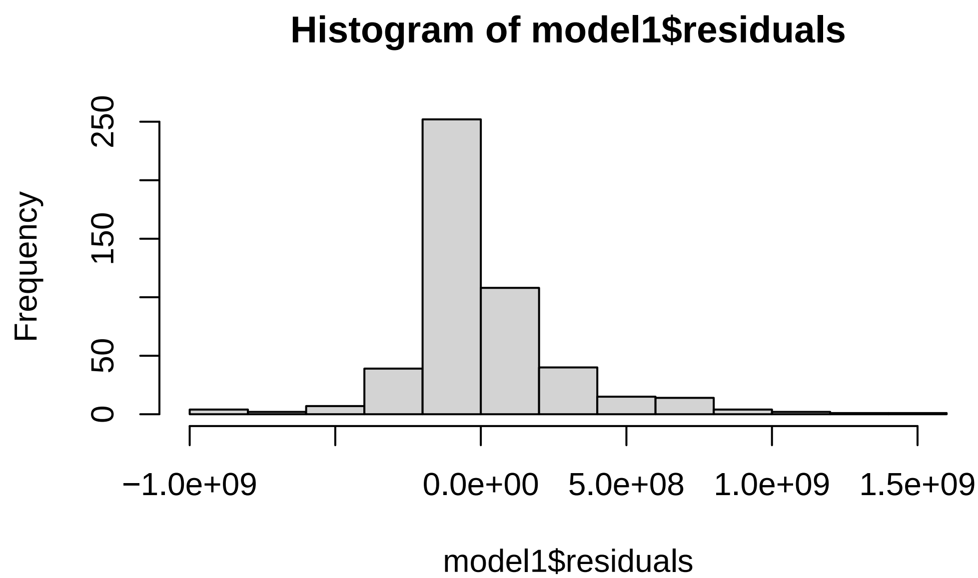



```
    Shapiro-Wilk normality test

data:  model1$residuals
W = 0.87068, p-value < 2.2e-16
```

```
    studentized Breusch-Pagan test

data:  model1
BP = 75.267, df = 1, p-value < 2.2e-16
```

```
    Durbin-Watson test
```

```
data:  model1
DW = 1.9908, p-value = 0.4595
alternative hypothesis: true autocorrelation is greater than 0
```



From the evidence above we can conclude that this model is not meeting the basic assumptions we have about the error term. The residuals seem to be not randomly scattered, there seems to be heteroscedasticity and it doesn't follow the normal distribution. Looking at the qqnorm, the histogram, as well as the numerical tests we need some transformation on this.

Model 1 Scatterplot with Fitted Line

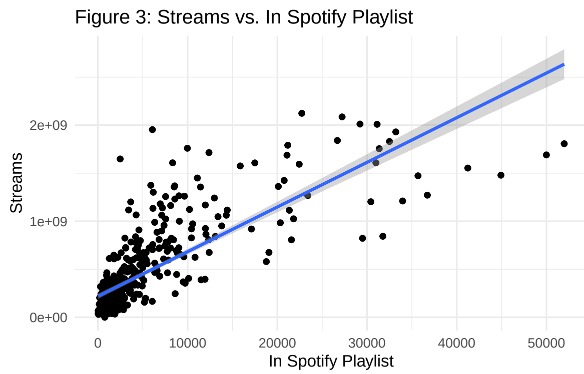


Figure 2: The scatter plot shows a moderately strong positive relationship, as In Spotify Playlist increases, Streams increases. The variance increases as In Spotify Playlist increases, and there are a few identifiable outliers.

From looking at the scatter plot of the linear regression, we chose to further clean our data and analyze songs that are in less than or equal to 2500 playlists on Spotify to focus on how they affect the number of streams.

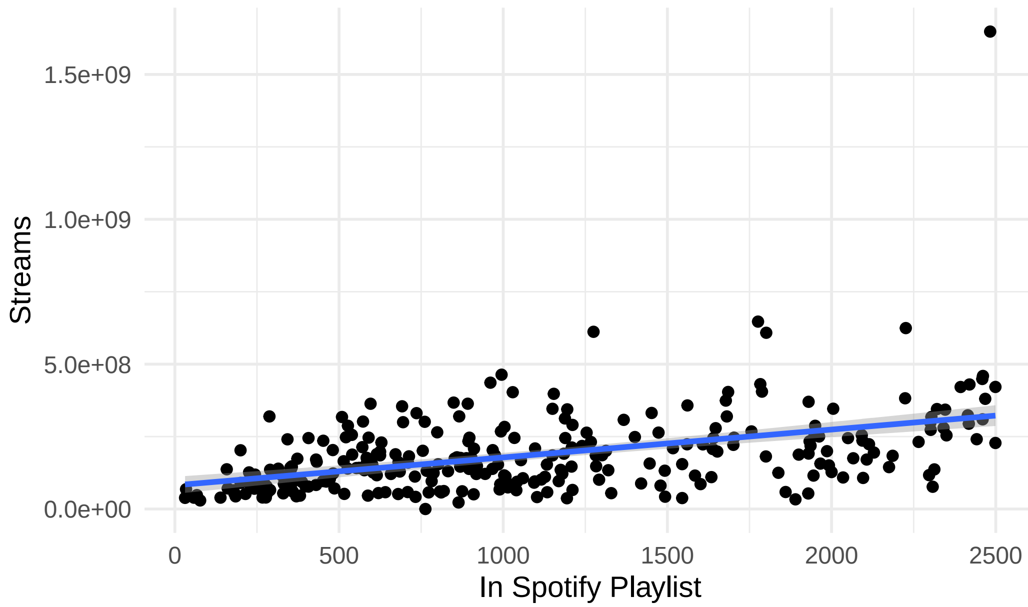
Model 1 with Limit on Playlist Count

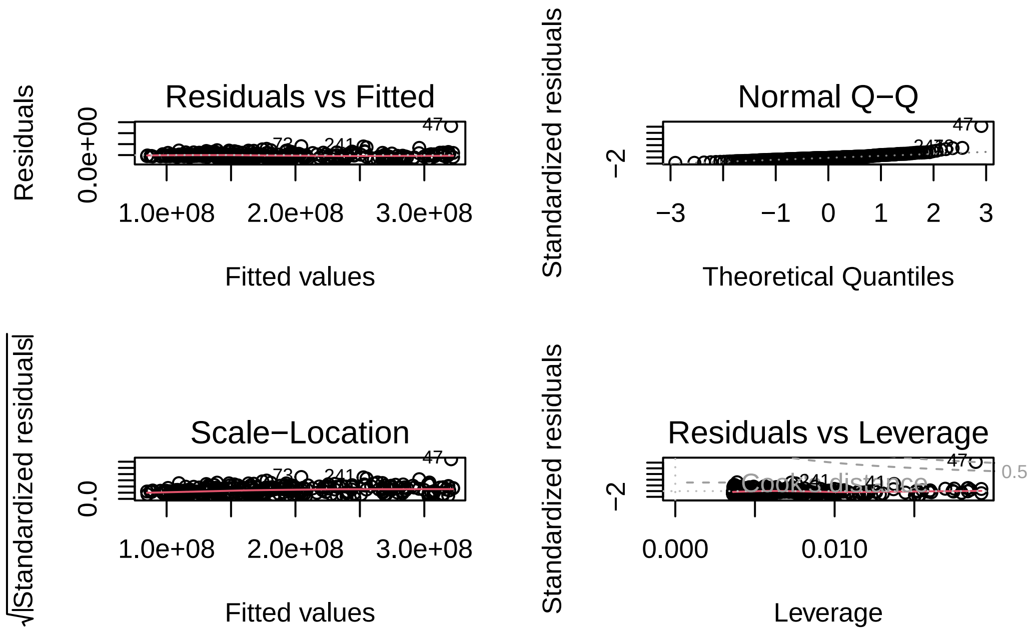
```r
library(dplyr)

new_data <- spotify_data %>%
  filter(in_spotify_playlists<=2500)

new_model1 <- lm(streams~in_spotify_playlists, data = new_data)
summary(new_model1)
```

```
Call:
lm(formula = streams ~ in_spotify_playlists, data = new_data)

Residuals:
       Min         1Q     Median         3Q        Max
-230347985  -72669314  -20083313   40112131 1327188477
```

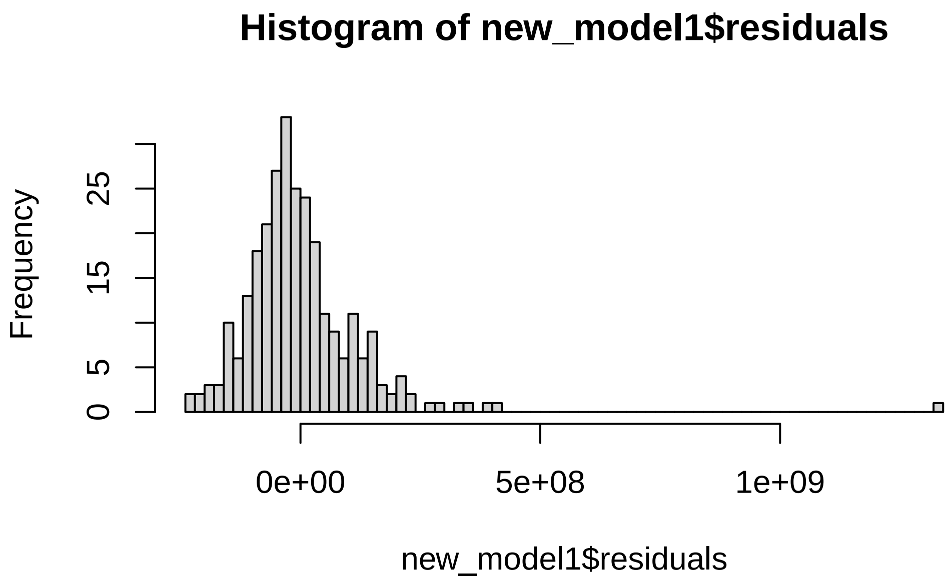
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           81828940   15021062   5.448 1.14e-07 ***
in_spotify_playlists     96244      11662   8.253 6.59e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 131100000 on 274 degrees of freedom
Multiple R-squared:  0.1991,    Adjusted R-squared:  0.1962 
F-statistic:  68.1 on 1 and 274 DF,  p-value: 6.594e-15
```

Figure 4: Streams vs. In Spotify Playlist






```
    Shapiro-Wilk normality test

data:  new_model1$residuals
W = 0.75361, p-value < 2.2e-16


    studentized Breusch-Pagan test

data:  new_model1
BP = 7.0444, df = 1, p-value = 0.007951


    Durbin-Watson test

data:  new_model1
DW = 2.2017, p-value = 0.9544
alternative hypothesis: true autocorrelation is greater than 0
```



After looking at the residual plots for this model, our plots are still not what we are wanting for the model. In order to meet our assumptions, it would be ideal to do the log transformation on our dependent variable i.e. streams. Although our R^2 decreased, we decided to perform our assumption analysis on this model to see if it improved interpretability.

Model 1 Logarithmic Transformation

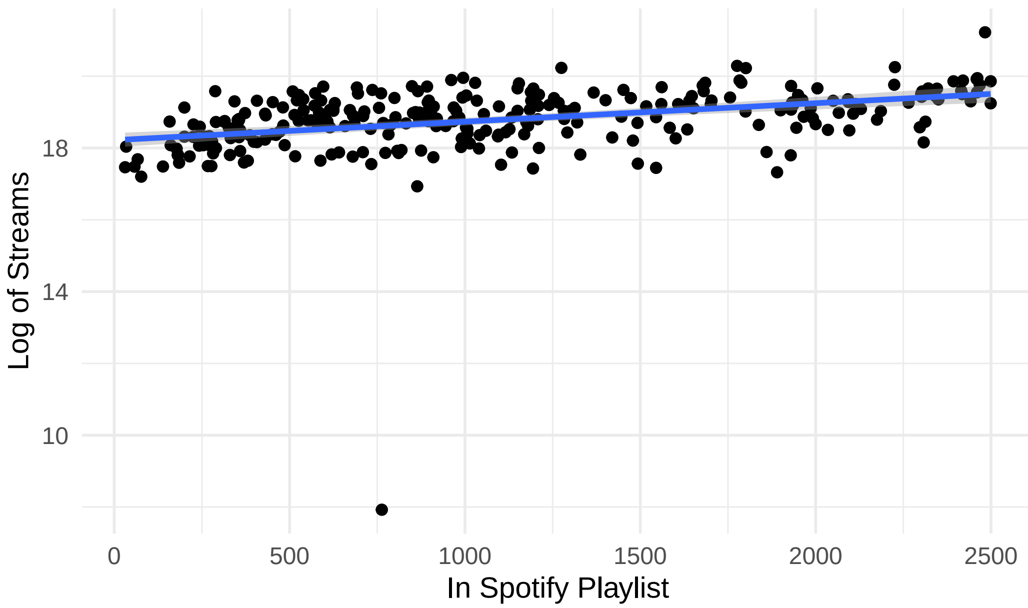
log(Streams) = 18.22 + .0005148(in_spotify_playlists)

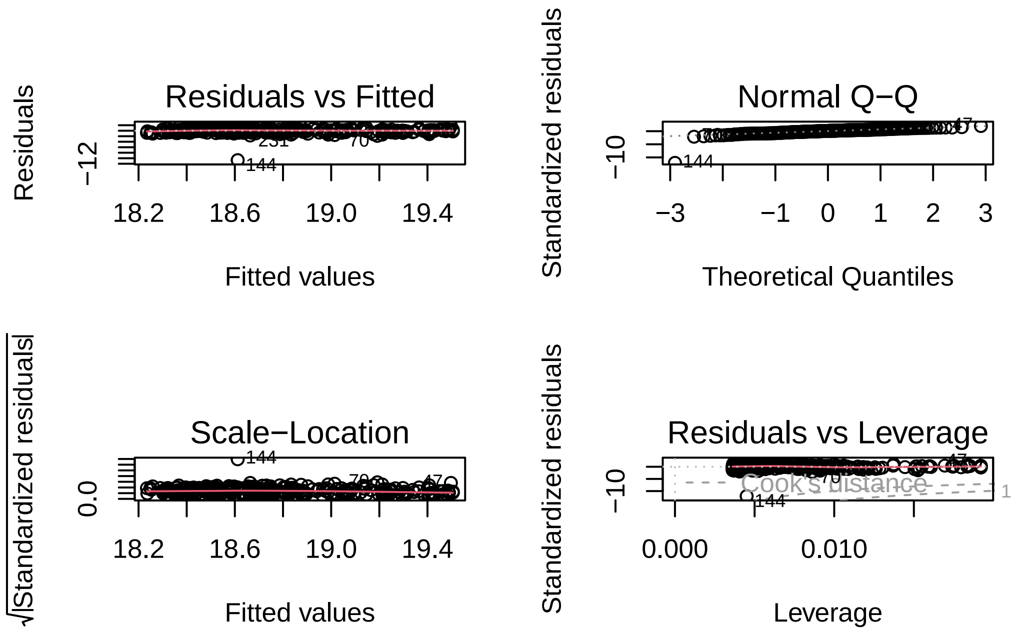
```
new_model1log <- lm(log(streams)~in_spotify_playlists, data = new_data)
summary(new_model1log)
```

```
Call:
lm(formula = log(streams) ~ in_spotify_playlists, data = new_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.6881  -0.3499   0.0843   0.4135   1.7255 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
```

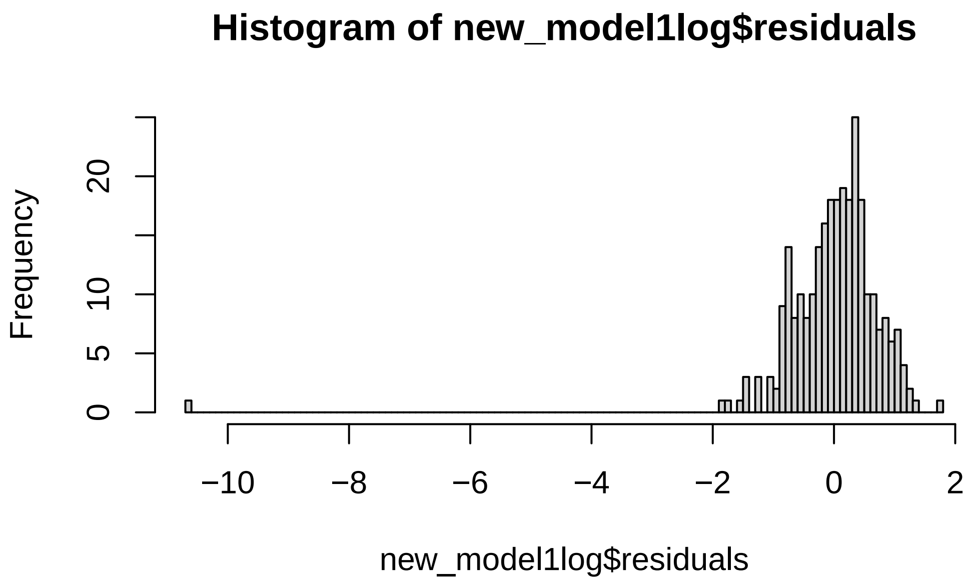
```
(Intercept)          1.822e+01  1.016e-01 179.285  < 2e-16 ***
in_spotify_playlists 5.148e-04  7.890e-05   6.526 3.26e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8866 on 274 degrees of freedom
Multiple R-squared:  0.1345,    Adjusted R-squared:  0.1313
F-statistic: 42.58 on 1 and 274 DF,  p-value: 3.264e-10
```

Figure 5: Streams vs. In Spotify Playlist






```
    studentized Breusch-Pagan test

data:  new_model1log
BP = 0.22299, df = 1, p-value = 0.6368


    Durbin-Watson test

data:  new_model1log
DW = 2.0019, p-value = 0.5097
alternative hypothesis: true autocorrelation is greater than 0
```



After performing the log transformation, we can see that the residuals vs. the fitted value actually look like the random scatter, the qqnorm seems way better than before and the histogram looks like a normal distribution, if we overlook the outlier in the extreme left. The numerical tests for independence and constant variance also pass and our assumptions are met. Finally, we can say that this is the best model we have so far with all the transformations. We still note that our R^2 decreased, but since our assumptions are now met, we chose to stick with this.

• Linearity: Our residual plot after the transformation remains showing a linear relationship since there is a horizontal line without distinct patterns.

• Independence: In our dwtest, our p-value was not significant so we fail to reject the null. Thus, our model shows independence. This suggests that there is no significant autocorrelation in the residuals of our model.

• Homoscedasticity: Our bptest has a non-significant p-value as well, in which we do not reject the null of homoscedasticity. Our Scale-Location test also shows a horizontal line with equally spread points which supports homoscedasticity.

• Normality of Errors: Based on our Q-Q plot, our residuals are located along the line, demonstrating that the residuals are normally distributed. We have some marked points but the majority are on the the dotted line.

Residual Analysis/ Outliers and Influential Points

```r
outlierTest(new_model1log)
```

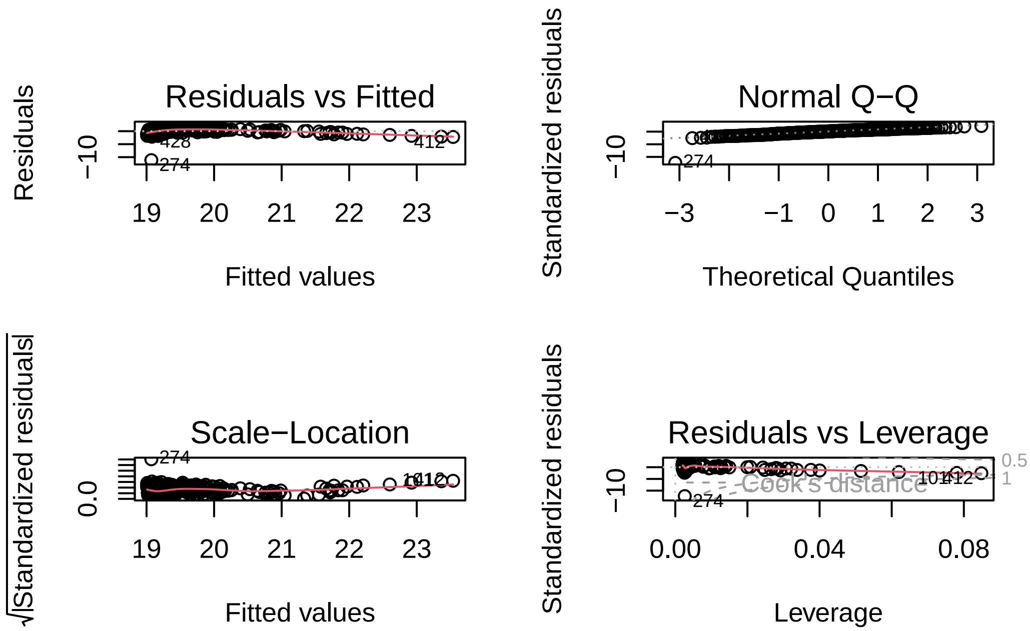
```
     rstudent unadjusted p-value Bonferroni p
144 -17.64366          5.121e-47   1.4134e-44
```

The Cook's distance didn't show any outliers or leverage points but from our qq plot after the transformation, we see that the 47th and the 144th observation are a little different from the trend. From the outlierTest, we confirmed that the 144th observation was an outlier. We could just explore them a little further and see if there is any valid justification to omit them from the data or not. For now, we opted to keep them in.

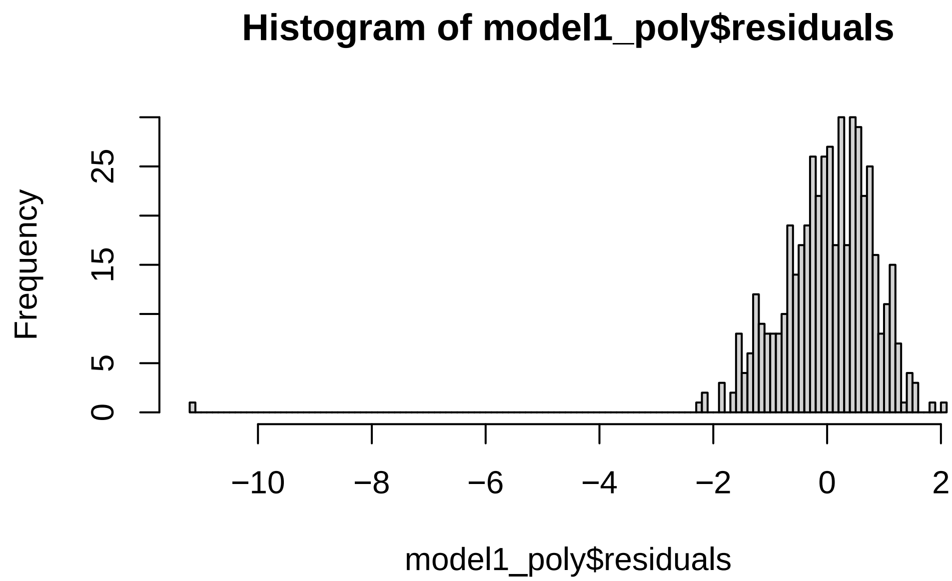
In an attempt to increase our R^2, we tried to fit a polynomial transformation after the log transformation.

```r
model1_poly <- lm(log(spotify_data$streams)~spotify_data$in_spotify_playlists + spotify_data$
summary(model1_poly)
```

```

Call:
lm(formula = log(spotify_data$streams) ~ spotify_data$in_spotify_playlists +
    spotify_data$in_spotify_playlists^2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1474  -0.4536   0.0781   0.5629   2.0017

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.900e+01  4.842e-02  392.49   <2e-16 ***
spotify_data$in_spotify_playlists 8.721e-05  5.504e-06   15.85   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.904 on 487 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.3402,    Adjusted R-squared:  0.3388
F-statistic: 251.1 on 1 and 487 DF,  p-value: < 2.2e-16
```




```
    Shapiro-Wilk normality test

data:  model1_poly$residuals
W = 0.80384, p-value < 2.2e-16


    studentized Breusch-Pagan test

data:  model1_poly
BP = 0.00091705, df = 1, p-value = 0.9758


    Durbin-Watson test

data:  model1_poly
DW = 1.9608, p-value = 0.3321
alternative hypothesis: true autocorrelation is greater than 0
```



This did increase our R^2 slightly, but this caused our assumptions to not be met, thus we chose to interpret our previous log transformed model.

Transformed Model 1 Interpretation

The intercept of 18.22 represents the predicted log-transformed number of streams for a song that is not included in any Spotify playlists. The coefficient 0.0005148 indicates that for each additional playlist a song is added to, there is an expected increase of 0.0005148 in the log-transformed number of streams.

The R^2 value for the model is 0.1345 suggesting that only 13.45% of the variability in the log-transformed number of streams is explained by the number of playlists. This means we need to explore adding more variables to the model to see if an altered model can help explain the variability in streams in a more meaningful way.

The F-value of 42.58 is large, indicating that the predictors contribute significantly to the model. The overall p-value of 3.264 e^{-10} suggests strong evidence against the null hypothesis, confirming that the number of playlists significantly contributes to predicting the log-transformed number of streams.

Transformed Model 1 Conclusion

An initial residual plot revealed a funnel-shaped pattern, indicating heteroscedasticity, which violates the assumption of constant variances. Both the residual plot and the bp test confirmed the presence of heteroscedasticity. Furthermore, the residuals did not meet the assumption of normality based on the diagnostic plots.

The data also showed a deviation from linearity when the number of playlists exceeded 2500. To address these issues, we restricted the data to include only observations where the number of playlists was less than 2500. Additionally, a log transformation on the response variable was applied to address heteroscedasticity and improve normality.

After making these adjustments, the updated simple linear regression model yielded the following results: the intercept is 18.22 with p-value of 2eˆ-16$ and the coefficient for number of playlists is 0.0005148 with p-value of 3.26e-10. Both the intercept and the coefficient were statistically significant, indicating they are meaningful contributors to the prediction of the log-transformed number of streams.

Final Insights

Null Hypothesis: H_0: $\beta_1 = 0$ The number of playlists a song is in (predictor variable) and the number of streams (response variable) have no linear relationship.

Alternate Hypothesis: H_A: $\beta_1 \neq 0$ The number of playlists a song is in (predictor variable) and the number of streams (response variable) have some linear relationship.

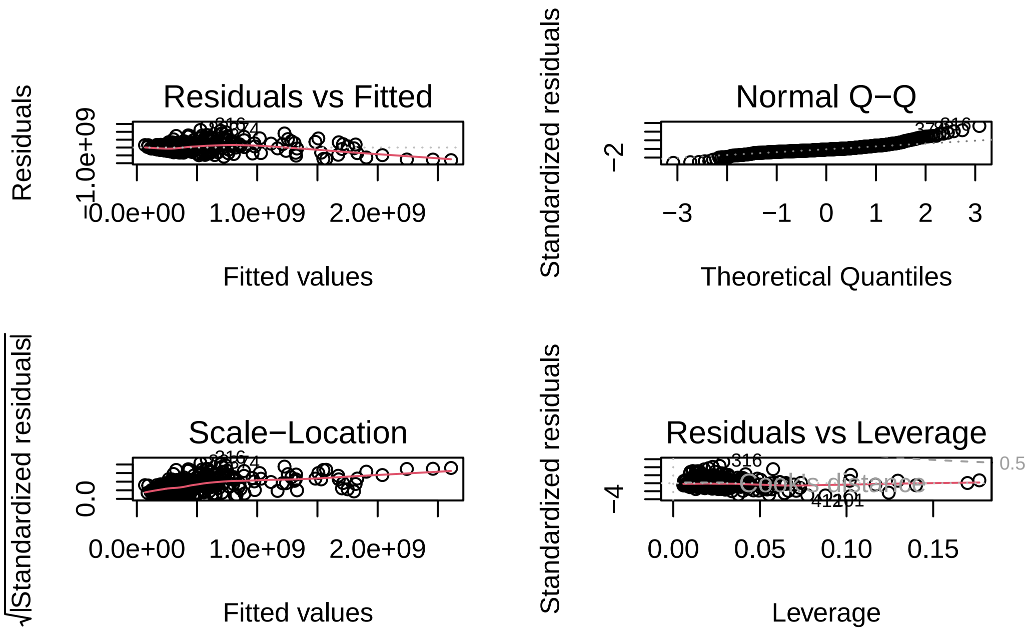
Looking at the p value of 3.264e-10 which is much smaller than our alpha significance level of 0.05 for our transformed model, we can see that we have sufficient evidence to reject the null hypothesis. This suggests that number of playlists a song is in has a statistically significant impact on the number of streams. This demonstrates that a strong indicator for more streams of a song is the number of playlists it is in. We would suggest to artists looking at our model to think about this in their marketing strategy. They could do this by encouraging their listeners to add their songs to playlists. Overall, our model does indicate a relationship between number of playlists a song is in and the number of streams a song receives.

Model 2: Multiple Linear Regression

From the full model, which variables are statistically significant for our model to predict the number of streams?

```r
model2 <- lm(spotify_data$streams ~ spotify_data$artist_count + spotify_data$in_spotify_char
summary(model2)
```

```
Call:
lm(formula = spotify_data$streams ~ spotify_data$artist_count +
    spotify_data$in_spotify_charts + spotify_data$in_spotify_playlists +
    spotify_data$bpm + spotify_data$`danceability_%` + spotify_data$`valence_%` +
    spotify_data$`energy_%` + spotify_data$`acousticness_%` +
    spotify_data$`instrumentalness_%` + spotify_data$`liveness_%` +
    spotify_data$`speechiness_%`)

Residuals:
       Min         1Q     Median         3Q        Max
-803792084 -137914190  -39266298   94766609 1372523954

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       214701869  114957144   1.868   0.0624 .
spotify_data$artist_count         -22474660   14282070  -1.574   0.1162
spotify_data$in_spotify_charts      3674074     672160   5.466 7.43e-08 ***
spotify_data$in_spotify_playlists     45451       1642  27.674  < 2e-16 ***
spotify_data$bpm                     428721     428408   1.001   0.3175
spotify_data$`danceability_%`        165160     949678   0.174   0.8620
spotify_data$`valence_%`            -618510     602670  -1.026   0.3053
spotify_data$`energy_%`             -428495     991311  -0.432   0.6658
spotify_data$`acousticness_%`        624828     600401   1.041   0.2985
spotify_data$`instrumentalness_%`   -705633    1208773  -0.584   0.5597
spotify_data$`liveness_%`           -117194     897494  -0.131   0.8962
spotify_data$`speechiness_%`       -1601743    1243062  -1.289   0.1982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 262600000 on 477 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.6507,    Adjusted R-squared:  0.6426
F-statistic: 80.77 on 11 and 477 DF,  p-value: < 2.2e-16
```



By looking at our full model, we see that only some of the coefficients are significant. This means that we need to further break it down and do the process of variable selection since there could be better models out there. We chose to rewrite our full model with our dataset limited to songs in less than or equal to 2500 Spotify playlists.

We can see that the Residual vs Fitted plot is cone shaped, which does not align with what we want for our model. Our other assumptions are not being met which supports our choice to look at the model using our limited data set adjusted for spotify playlists.

```r
new_data <- spotify_data %>%
  filter(in_spotify_playlists<=2500)
model2_new <- lm(new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts + new_
summary(model2_new)
```

```
Call:
lm(formula = new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$bpm + new_data$`danceability_%` +
    new_data$`valence_%` + new_data$`energy_%` + new_data$`acousticness_%` +
    new_data$`instrumentalness_%` + new_data$`liveness_%` + new_data$`speechiness_%`)

Residuals:
       Min         1Q     Median         3Q        Max
```

```
-232312881  -66104166  -13649318   50785618 1230994587

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   148661416   75599930   1.966   0.0503 .
new_data$artist_count         -16967337    8122715  -2.089   0.0377 *
new_data$in_spotify_charts      1099152     461482   2.382   0.0179 *
new_data$in_spotify_playlists     98253      11818   8.314 4.94e-15 ***
new_data$bpm                      89142     282507   0.316   0.7526
new_data$`danceability_%`       -410643     616841  -0.666   0.5062
new_data$`valence_%`             762265     396781   1.921   0.0558 .
new_data$`energy_%`             -468686     631983  -0.742   0.4590
new_data$`acousticness_%`       -647550     395308  -1.638   0.1026
new_data$`instrumentalness_%`    549505     748334   0.734   0.4634
new_data$`liveness_%`           -797563     547934  -1.456   0.1467
new_data$`speechiness_%`       -1134181     733191  -1.547   0.1231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 127700000 on 264 degrees of freedom
Multiple R-squared:  0.2674,	Adjusted R-squared:  0.2369
F-statistic: 8.759 on 11 and 264 DF,  p-value: 3.096e-13
```

Model 2 Variable Selection:

To select the variables that would be statistically significant we would perform variable selections. Using stepwise backwards selection, we will interpret and discuss what variables need to be omitted.

```r
step(model2_new, direction = "backward", scope = formula(model2_new))
```

```
Start:  AIC=10314.89
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$bpm + new_data$`danceability_%` +
    new_data$`valence_%` + new_data$`energy_%` + new_data$`acousticness_%` +
    new_data$`instrumentalness_%` + new_data$`liveness_%` + new_data$`speechiness_%`

                                Df  Sum of Sq        RSS   AIC
- new_data$bpm                   1 1.6235e+15 4.3063e+18 10313
- new_data$`danceability_%`      1 7.2263e+15 4.3119e+18 10313
- new_data$`instrumentalness_%`  1 8.7920e+15 4.3135e+18 10314
```

```
- new_data$`energy_%`            1 8.9679e+15 4.3136e+18 10314
<none>                                        4.3047e+18 10315
- new_data$`liveness_%`          1 3.4547e+16 4.3392e+18 10315
- new_data$`speechiness_%`       1 3.9018e+16 4.3437e+18 10315
- new_data$`acousticness_%`      1 4.3753e+16 4.3484e+18 10316
- new_data$`valence_%`           1 6.0179e+16 4.3648e+18 10317
- new_data$artist_count          1 7.1148e+16 4.3758e+18 10317
- new_data$in_spotify_charts     1 9.2500e+16 4.3972e+18 10319
- new_data$in_spotify_playlists  1 1.1271e+18 5.4318e+18 10377

Step:  AIC=10312.99
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`danceability_%` +
    new_data$`valence_%` + new_data$`energy_%` + new_data$`acousticness_%` +
    new_data$`instrumentalness_%` + new_data$`liveness_%` + new_data$`speechiness_%`

                                Df  Sum of Sq        RSS   AIC
- new_data$`danceability_%`      1 8.2696e+15 4.3146e+18 10312
- new_data$`instrumentalness_%`  1 9.4318e+15 4.3157e+18 10312
- new_data$`energy_%`            1 9.5236e+15 4.3158e+18 10312
<none>                                        4.3063e+18 10313
- new_data$`liveness_%`          1 3.4541e+16 4.3408e+18 10313
- new_data$`speechiness_%`       1 3.9217e+16 4.3455e+18 10314
- new_data$`acousticness_%`      1 4.7518e+16 4.3538e+18 10314
- new_data$`valence_%`           1 6.4048e+16 4.3703e+18 10315
- new_data$artist_count          1 7.2706e+16 4.3790e+18 10316
- new_data$in_spotify_charts     1 9.4440e+16 4.4007e+18 10317
- new_data$in_spotify_playlists  1 1.1351e+18 5.4414e+18 10376

Step:  AIC=10311.52
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`energy_%` +
    new_data$`acousticness_%` + new_data$`instrumentalness_%` +
    new_data$`liveness_%` + new_data$`speechiness_%`

                                Df  Sum of Sq        RSS   AIC
- new_data$`energy_%`            1 8.5034e+15 4.3231e+18 10310
- new_data$`instrumentalness_%`  1 9.2530e+15 4.3238e+18 10310
- new_data$`liveness_%`          1 3.0420e+16 4.3450e+18 10312
<none>                                        4.3146e+18 10312
- new_data$`acousticness_%`      1 4.1494e+16 4.3561e+18 10312
- new_data$`speechiness_%`       1 4.5712e+16 4.3603e+18 10312
- new_data$`valence_%`           1 5.6035e+16 4.3706e+18 10313
```

```
- new_data$artist_count          1 7.7388e+16 4.3919e+18 10314
- new_data$in_spotify_charts     1 8.9599e+16 4.4042e+18 10315
- new_data$in_spotify_playlists  1 1.1393e+18 5.4538e+18 10374

Step:  AIC=10310.07
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`acousticness_%` +
    new_data$`instrumentalness_%` + new_data$`liveness_%` + new_data$`speechiness_%`

                                Df  Sum of Sq        RSS   AIC
- new_data$`instrumentalness_%`  1 8.7011e+15 4.3318e+18 10309
<none>                                        4.3231e+18 10310
- new_data$`acousticness_%`      1 3.3508e+16 4.3566e+18 10310
- new_data$`liveness_%`          1 3.6000e+16 4.3591e+18 10310
- new_data$`speechiness_%`       1 4.2859e+16 4.3659e+18 10311
- new_data$`valence_%`           1 4.7533e+16 4.3706e+18 10311
- new_data$artist_count          1 8.2867e+16 4.4059e+18 10313
- new_data$in_spotify_charts     1 8.8181e+16 4.4112e+18 10314
- new_data$in_spotify_playlists  1 1.1559e+18 5.4789e+18 10374

Step:  AIC=10308.62
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`acousticness_%` +
    new_data$`liveness_%` + new_data$`speechiness_%`

                                Df  Sum of Sq        RSS   AIC
<none>                                        4.3318e+18 10309
- new_data$`acousticness_%`      1 3.3727e+16 4.3655e+18 10309
- new_data$`liveness_%`          1 3.8345e+16 4.3701e+18 10309
- new_data$`valence_%`           1 4.3055e+16 4.3748e+18 10309
- new_data$`speechiness_%`       1 4.7728e+16 4.3795e+18 10310
- new_data$artist_count          1 8.5768e+16 4.4175e+18 10312
- new_data$in_spotify_charts     1 8.6924e+16 4.4187e+18 10312
- new_data$in_spotify_playlists  1 1.1789e+18 5.5107e+18 10373


Call:
lm(formula = new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`acousticness_%` +
    new_data$`liveness_%` + new_data$`speechiness_%`)

Coefficients:
```

```
                 (Intercept)          new_data$artist_count
                   110723144                      -18429999
   new_data$in_spotify_charts  new_data$in_spotify_playlists
                     1056645                          99819
         new_data$`valence_%`      new_data$`acousticness_%`
                      557628                        -460586
        new_data$`liveness_%`       new_data$`speechiness_%`
                     -820920                       -1229161
```

```
reduced_model2 <- lm(streams~artist_count + in_spotify_charts + in_spotify_playlists, data=n

BIC(model2_new)
```

```
[1] 11147.21
```

```
BIC(reduced_model2)
```

```
[1] 11113.21
```

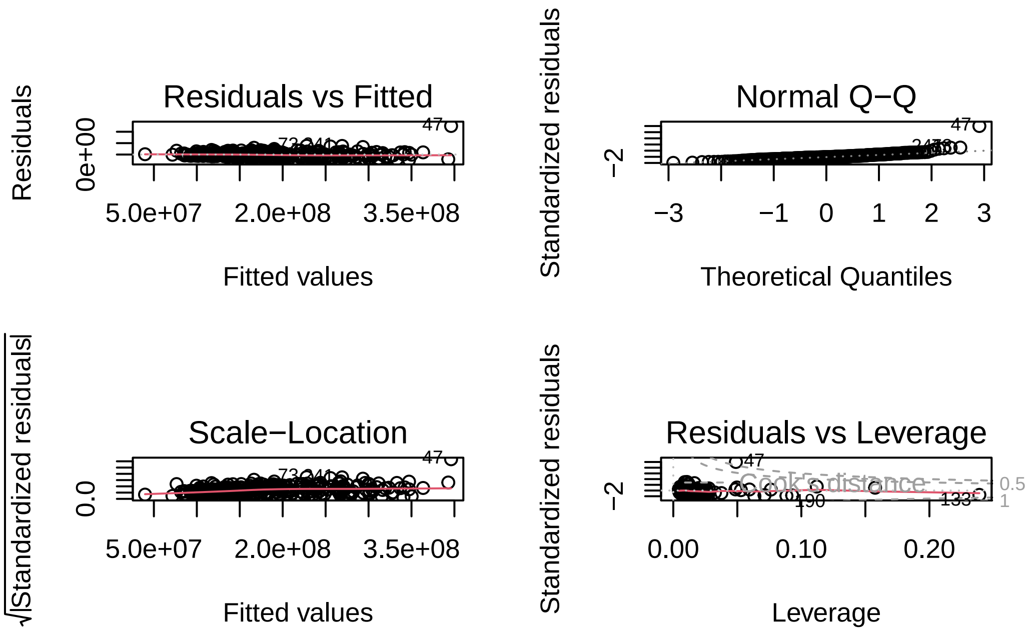
We used stepwise selection in forward, back, and both and were able to reduce the model. The model with the smallest AIC was from the backwards stepwise.

Based on our stepwise selection results using the backward direction, it is shown that several variables have little power in explaining the model. When we remove BPM, danceability_%, energy_%, and instrumentalness_%, our AIC decreased, demonstrating a better model fit. Given that our p-values for these variables were all over 0.4, which is above the threshold for statistical significance, we felt safe omitting them from our model. This resulted in a simpler multiple regression model, with a lower and better AIC and variables that contribute more to explaining the number of streams.

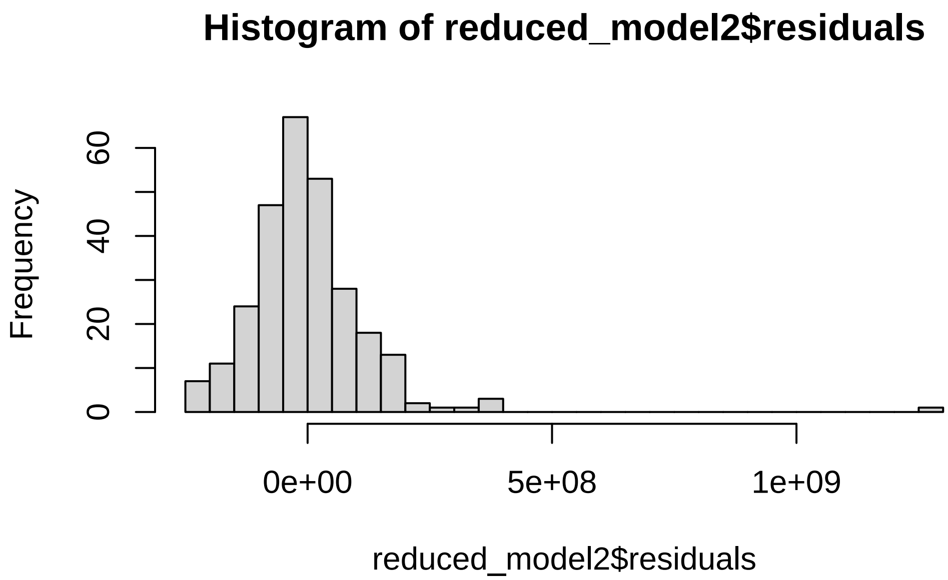
Thus we had a model with predictors as acousticness%, liveness%, valence%, speechiness%, artist_count, in_spotify_charts, and in_spotify_playlists. Furthermore, after looking at the scatterplot matrix from the very beginning of this report we know that there is no linear relationship between acousticness%, liveness%, valence%, and speechiness%. After this, we found the BIC for our original model and compared this to the model omitting these variables, which indicated that our reduced model had a lower BIC, demonstrating increased goodness of fit and supporting our decision to omit them. We also noted that the BIC for our model before stepwise selection and after was the same, while our current reduced model without these 4 variables had a much smaller BIC. This is enough evidence to omit these from the full model. Thus, our reduced model looks something like this:

```r
reduced_model2 <- lm(streams~artist_count + in_spotify_charts + in_spotify_playlists, data=ne
summary(reduced_model2)
```

```

Call:
lm(formula = streams ~ artist_count + in_spotify_charts + in_spotify_playlists,
    data = new_data)

Residuals:
       Min         1Q     Median         3Q        Max
-244803796  -71142608  -13500358   46823282 1251849386

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           99210185   19747819   5.024 9.18e-07 ***
artist_count         -17491846    7826942  -2.235   0.0262 *
in_spotify_charts      1326921     447565   2.965   0.0033 **
in_spotify_playlists     95100      11427   8.322 4.23e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 128300000 on 272 degrees of freedom
Multiple R-squared:  0.2377,    Adjusted R-squared:  0.2293
F-statistic: 28.27 on 3 and 272 DF,  p-value: 6.067e-16
```




```
    Shapiro-Wilk normality test

data:  reduced_model2$residuals
W = 0.78434, p-value < 2.2e-16


    Durbin-Watson test

data:  reduced_model2
DW = 2.1849, p-value = 0.9396
alternative hypothesis: true autocorrelation is greater than 0


    studentized Breusch-Pagan test

data:  reduced_model2
BP = 20.38, df = 3, p-value = 0.0001416
```



- Linearity is met from looking at our Residuals plot.
- The Q-Q plot shows some deviation from the dotted line, indicating that the residuals are not normally distributed.
- The Scale-Location plot shows a more horizontal line than before but is still suggesting that we do not have heteroscedasticity.
- There are some identified points we have to look further at from the Scale-Location plot.

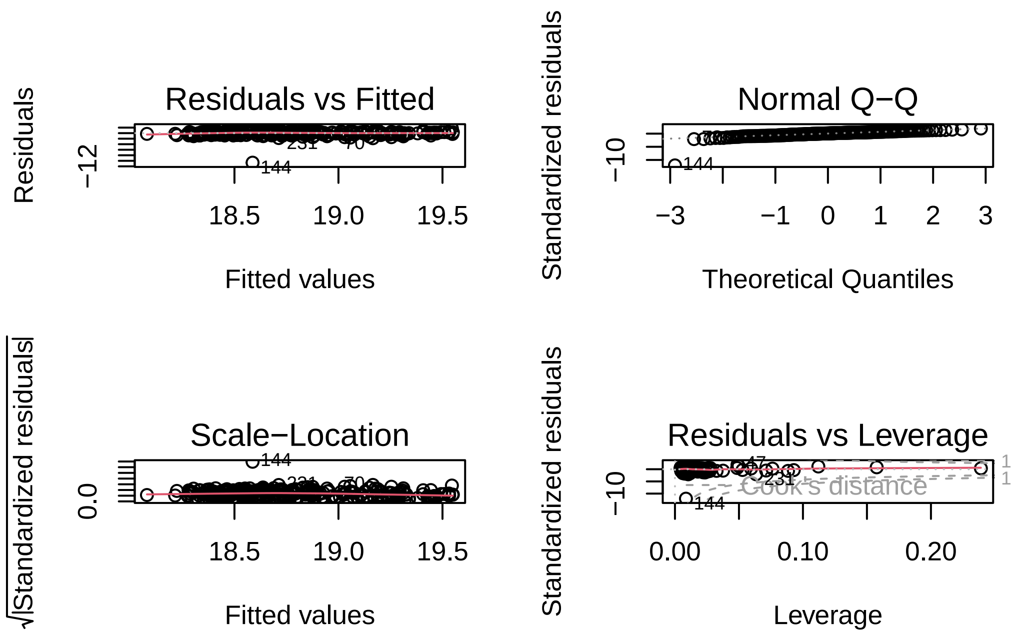
Looking at the evidence above, we aren't meeting the assumptions about our error term. Our residuals need to follow a normal distribution, have a random scatter in residual vs. fitted plot, have independence, and constant variance. To fix these, a log transformation would be ideal.

Model 2 Logarithmic Transformation

```
reduced_model2_log <- lm(log(streams)~artist_count + in_spotify_charts + in_spotify_playlists
summary(reduced_model2_log)
```

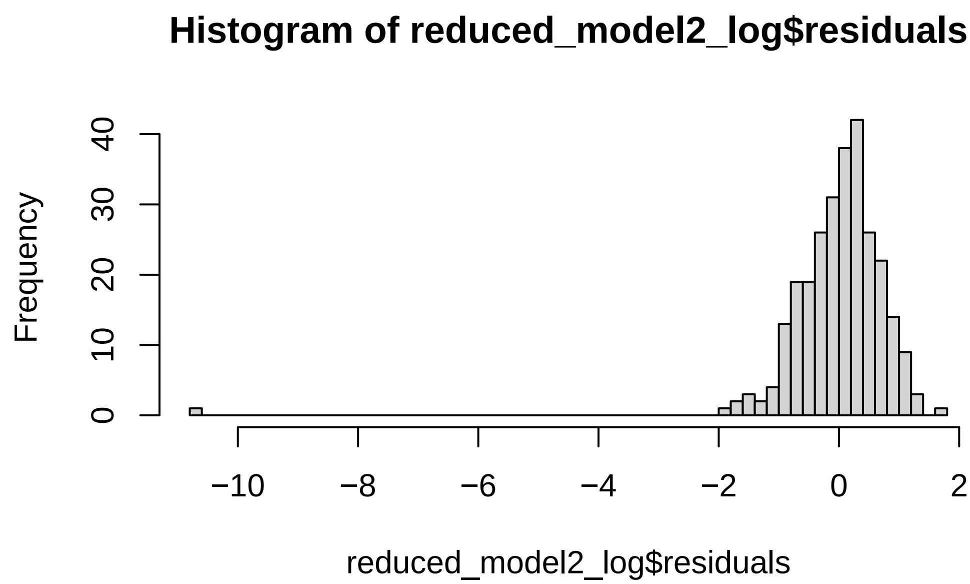
```
Call:
lm(formula = log(streams) ~ artist_count + in_spotify_charts +
```

```
    in_spotify_playlists, data = new_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6635  -0.3492   0.0842   0.4194   1.6775

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.833e+01  1.365e-01 134.251  < 2e-16 ***
artist_count         -6.893e-02  5.411e-02  -1.274    0.204
in_spotify_charts     9.288e-05  3.094e-03   0.030    0.976
in_spotify_playlists  5.152e-04  7.901e-05   6.521  3.4e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8872 on 272 degrees of freedom
Multiple R-squared:  0.1396,    Adjusted R-squared:  0.1302
F-statistic: 14.72 on 3 and 272 DF,  p-value: 6.587e-09
```




```
    Durbin-Watson test
```

```
data:  reduced_model2_log
DW = 1.9989, p-value = 0.5002
alternative hypothesis: true autocorrelation is greater than 0


    studentized Breusch-Pagan test

data:  reduced_model2_log
BP = 1.7161, df = 3, p-value = 0.6334
```



After performing the log transformation, we can see that the residuals vs. the fitted value actually look like the random scatter, the qqnorm seems way better than before and the histogram looks like a normal distribution, if we overlook the outlier in the extreme left. The numerical tests for independence and constant variance also pass and our assumptions are met. Finally, we can say that this is the best model we have so far with all the transformations.

- Linearity: Our scatter plot after the transformation remains showing a linear relationship.

- Independence: In our dwtest, our p-value was not significant so we fail to reject the null. Thus, our model shows independence. This suggests that there is no significant autocorrelation in the residuals of our model.

- Homoscedasticity: Our bptest has a non-significant p-value as well, in which we do not reject the null of homoscedasticity. The variance of the residuals is constant.

- Normality of Errors: Based on our Q-Q plot, our residuals are located along the striaght line, demonstrating that the residuals are normally distributed.

Thus, we are meeting our assumptions.

VIF

```r
vif(reduced_model2_log)
```

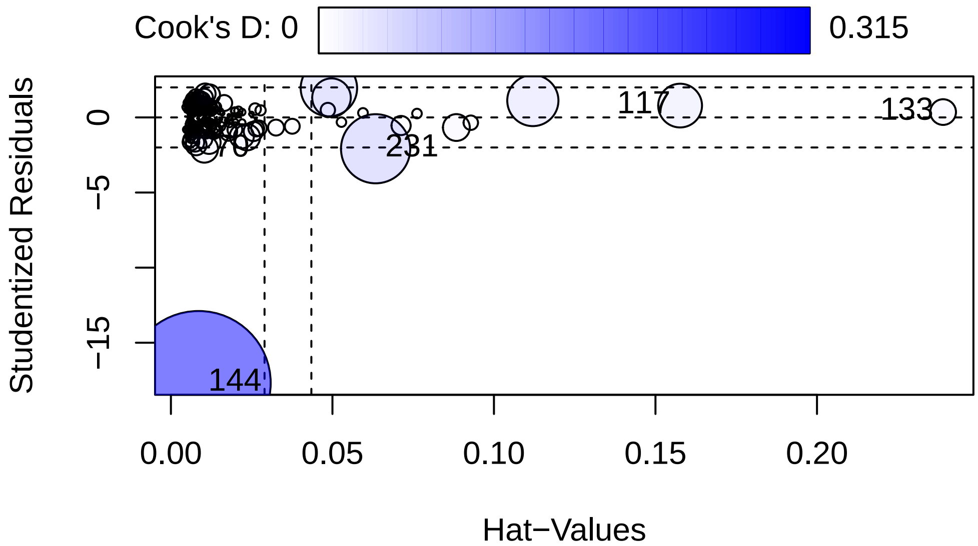
```
        artist_count    in_spotify_charts in_spotify_playlists
            1.000016             1.001350             1.001365
```

Based on the vif analysis for our model, since every value for the coefficients was greater than 1 and less than our cutoff value of 5, we can deduce that multicollinearity between our variables is not a problem for our model. Thus we did not omit any of these variables on multicollinearity issues.

Residual Analysis/ Outliers and Influential Points

From the graph and looking at the Cook's distance, it seems we are safe since none of the points seem to cross the Cook's lines. There might be some outliers like the 144th observation. We can conduct the outlier test to confirm this. However, we do not have sufficient background information whether we can omit these outliers.

```r
influencePlot(reduced_model2_log)
```




```
        StudRes         Hat      CookD
70   -2.0999100 0.010343702 0.01137949
117   0.7881993 0.157623514 0.02910262
133   0.3568990 0.239040278 0.01003542
144 -17.6822652 0.008578712 0.31520042
231  -2.0862487 0.063366330 0.07271784
```

```r
outlierTest(reduced_model2_log)
```

```
     rstudent unadjusted p-value Bonferroni p
144 -17.68227         4.6951e-47   1.2958e-44
```

Our influence test identified 5 influential points. These were observations 70, 117, 133, 144, and 231. We then used the outlier test to define possible outliers.

The Cook's distance didn't show any outliers but from our qq plot after the transformation, we see that the 47th and the 144th observation are a little different from the trend. From the outlierTest, we confirmed that the 144th observation was an outlier. We could just explore them a little further and see if there is any valid justification to omit them from the data or not. For now, we opted to keep them in.

Transformed Model 2 Interpretation

An initial residual plot revealed a funnel-shaped pattern, indicating heteroscedasticity, which violates the assumption of constant variances. Both the residual plot and the bp test confirmed the presence of heteroscedasticity. Furthermore, the residuals did not meet the assumption of normality based on the diagnostic plots.

The data also showed a deviation from linearity when the number of playlists exceeded 2500. To address these issues, we restricted the data to include only observations where the number of playlists was less than 2500. Additionally, a log transformation on the response variable was applied to address heteroscedasticity and improve normality.

Once we made these adjustments, our multiple linear regression gave us an intercept of 18.33, which is the log-transformed number of streams for a song with no artists, not included in any spotify playlists, and is unranked; the intercept has a p-value of 2e-16. The artists count coefficient is -6.893e-02, which tells us that for each unit increase in the number of artists, the number of streams for that song decreases by exp(6.893e-02); this coefficient has a p-value of 0.204, which means that it is not statistically significant. The in spotify charts coefficient is 9.288e-05, which means that the number of streams increases by exp(9.288e-05) for when a song has a one unit increase in ranking; this coefficient's p-value is 0.976, which means that it is not statistically significant . The in spotify playlists is 5.152e-04, which means that when the number of playlists a song is in increases, the number of streams goes up by exp(5.152e-04); the p-value for this coefficient is 3.4e-10, meaning that it is statistically significant.

The R^2 value for the model is 0.1396 suggesting that only 13.96% of the variability in the log-transformed number of streams is explained by the number of playlists. The attempts to increase this did not preserve our assumptions thus we opted to interpret the model with the lower R^2.

Transformed Model 2 Conclusion and Final Insights

Null Hypothesis: $H_0 : \beta_1 = 0, \beta_2 = 0, ..., \beta_p = 0$ None of the variables listed have a statistical impact on the number of streams.

Alternate Hypothesis: $H_A : \beta_1 \neq 0, \beta_2 \neq 0, ..., \beta_p \neq 0$. At least one of the variables listed above have a statistical impact on the number of streams.

The F-value of 14.72 indicates that the overall regression model is statistically significant, with a strong ability to explain variability in the log-transformed number of streams. This large F-value suggests that the predictors, artist count, the number of charts a song appears in, and the number of Spotify playlists a song is included in, collectively contribute meaningfully to the model. The corresponding p-value of $6.587e^{-9}$ provides strong evidence against the null hypothesis, which assumes that all regression coefficients are zero, meaning the predictors have no effect. Rejecting the null hypothesis confirms that at least one of the predictors is significantly associated with the log-transformed number of streams. This result supports the inclusion of these variables in the model to capture meaningful relationships with streaming performance.

We would suggest to artists looking at our model to focus on the results from the variables **in_spotify_playlists**, **in_spotify_charts**, and **artist_count** when creating a song and marketing it to listeners.

Model 3: Multiple Linear Regression with Categorical Variables

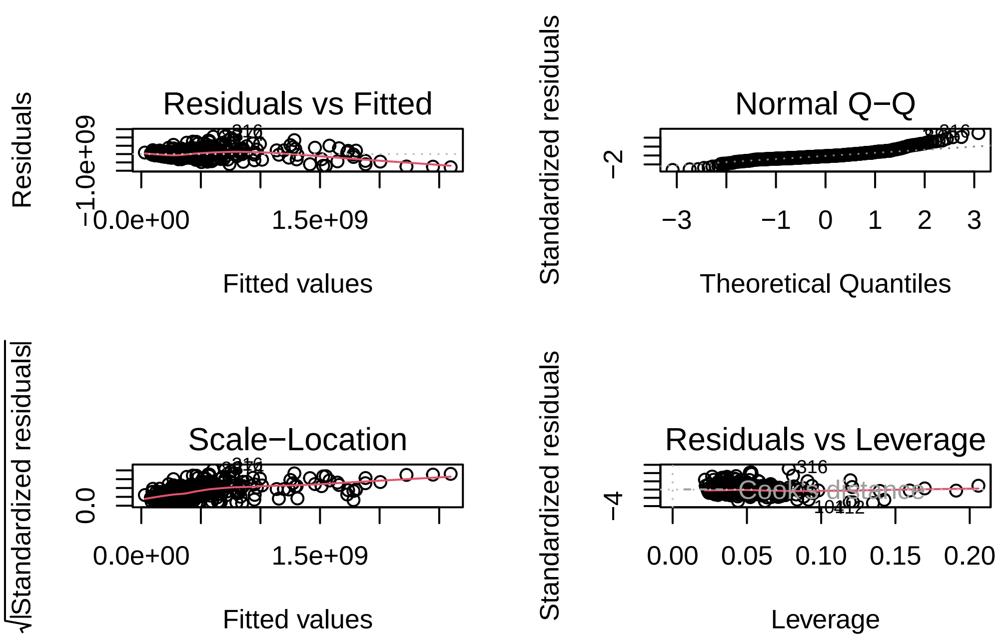
How does release month impact the number of streams?

To start with our third model, we should look at the full model with all the numerical variables and release month as our categorical. Even though, this model gives us an adjusted r^2 of 0.6448 none of our assumptions about the error or the residuals are met. We need to perform transformation on this to fix these.

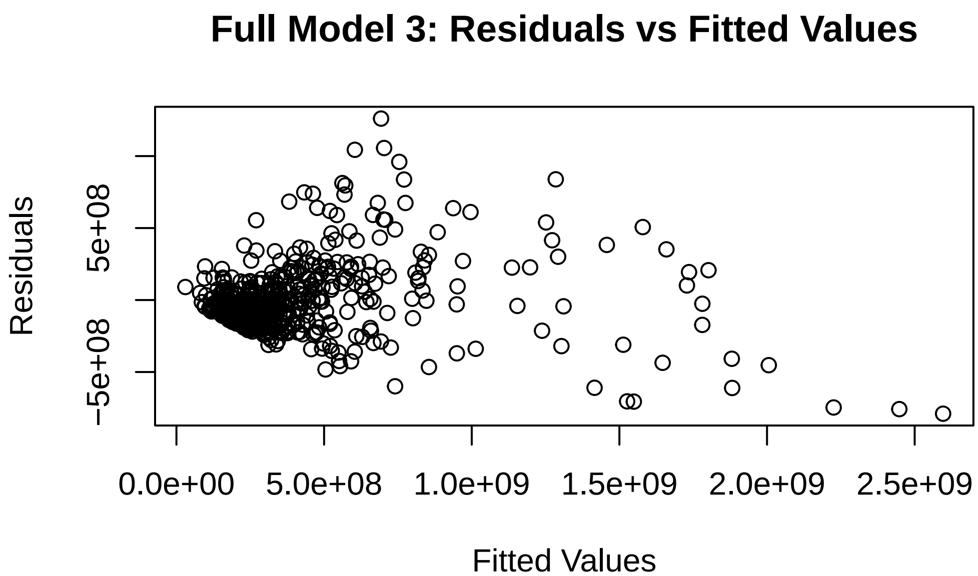
```r
full_model3 <- lm(spotify_data$streams ~ spotify_data$artist_count + spotify_data$in_spotify_
summary(full_model3)
```

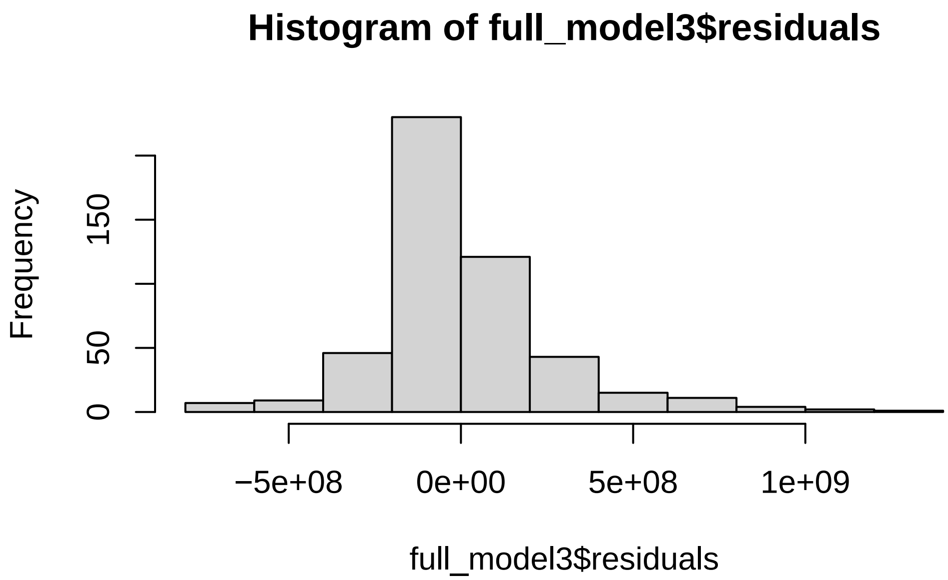
```

Call:
lm(formula = spotify_data$streams ~ spotify_data$artist_count + 
    spotify_data$in_spotify_charts + spotify_data$in_spotify_playlists + 
    spotify_data$bpm + spotify_data$`danceability_%` + spotify_data$`valence_%` + 
    spotify_data$`energy_%` + spotify_data$`acousticness_%` + 
    spotify_data$`instrumentalness_%` + spotify_data$`liveness_%` + 
    spotify_data$`speechiness_%` + spotify_data$released_month)

Residuals:
       Min         1Q     Median         3Q        Max 
-789669920 -143660616  -41621521   95238706 1260559090 

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)                       159000010  121502236   1.309    0.191    
spotify_data$artist_count         -22495022   14447964  -1.557    0.120    
spotify_data$in_spotify_charts      3960910     694623   5.702  2.1e-08 ***
spotify_data$in_spotify_playlists     44925       1763  25.477  < 2e-16 ***
spotify_data$bpm                     510642     431273   1.184    0.237    
spotify_data$`danceability_%`        337609     959310   0.352    0.725    
spotify_data$`valence_%`            -548004     609029  -0.900    0.369    
```

```
spotify_data$`energy_%`             -210792     995939  -0.212     0.832
spotify_data$`acousticness_%`        746522     605180   1.234     0.218
spotify_data$`instrumentalness_%`   -720047    1215916  -0.592     0.554
spotify_data$`liveness_%`           -166972     899470  -0.186     0.853
spotify_data$`speechiness_%`       -1468133    1249083  -1.175     0.240
spotify_data$released_month10      14700448   56458738   0.260     0.795
spotify_data$released_month11      79087161   53044587   1.491     0.137
spotify_data$released_month12      11019188   54430165   0.202     0.840
spotify_data$released_month2      -50327973   57153646  -0.881     0.379
spotify_data$released_month3       51118604   51119292   1.000     0.318
spotify_data$released_month4       66404381   54537686   1.218     0.224
spotify_data$released_month5      -21404062   47673562  -0.449     0.654
spotify_data$released_month6      -56379296   53307938  -1.058     0.291
spotify_data$released_month7      -21150341   60347992  -0.350     0.726
spotify_data$released_month8      105796436   68632360   1.541     0.124
spotify_data$released_month9       23977706   59406660   0.404     0.687
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261800000 on 466 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.6608,    Adjusted R-squared:  0.6448
F-statistic: 41.27 on 22 and 466 DF,  p-value: < 2.2e-16
```



```
    Shapiro-Wilk normality test

data:  full_model3$residuals
W = 0.90977, p-value < 2.2e-16


    studentized Breusch-Pagan test

data:  full_model3
BP = 108.17, df = 22, p-value = 2.336e-13


    Durbin-Watson test

data:  full_model3
DW = 2.0749, p-value = 0.7941
alternative hypothesis: true autocorrelation is greater than 0
```





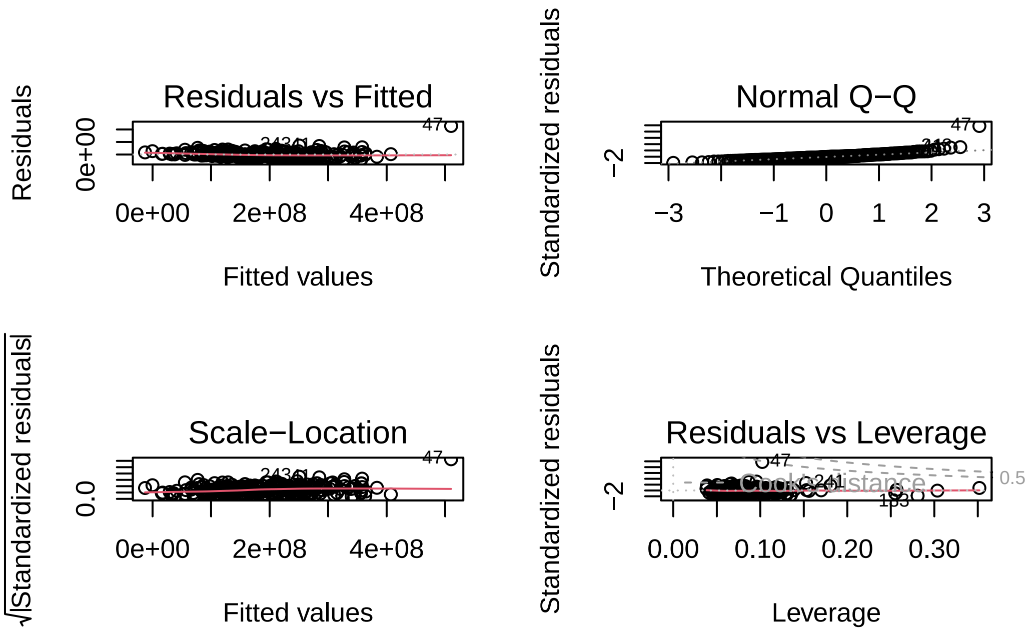
- Once again, we notice that our residual plots do not show constant variance. Our very small p-value of 2.336e-13 for our bp tests supports our findings of heteroscedasticity.The Scale-Location plot shows non-equally spaced points also suggesting that we do not have homoscedasticity.
- Our residuals vs fitted values shows a cone pattern which is not in accordance with our assumptions, suggesting a transformation of the variables.
- Our Shapiro Wilk test is suggesting that we do not have normality. This is supported by our Q-Q plot in which we have strong deviation from the line.
- The dw test has a large p-value, suggesting that we do have independence in our model.

From previous models, the transformed data of songs that are in less than or equal to 2500 playlists have given better results. We should make this model with the same data as well and test out the assumptions.

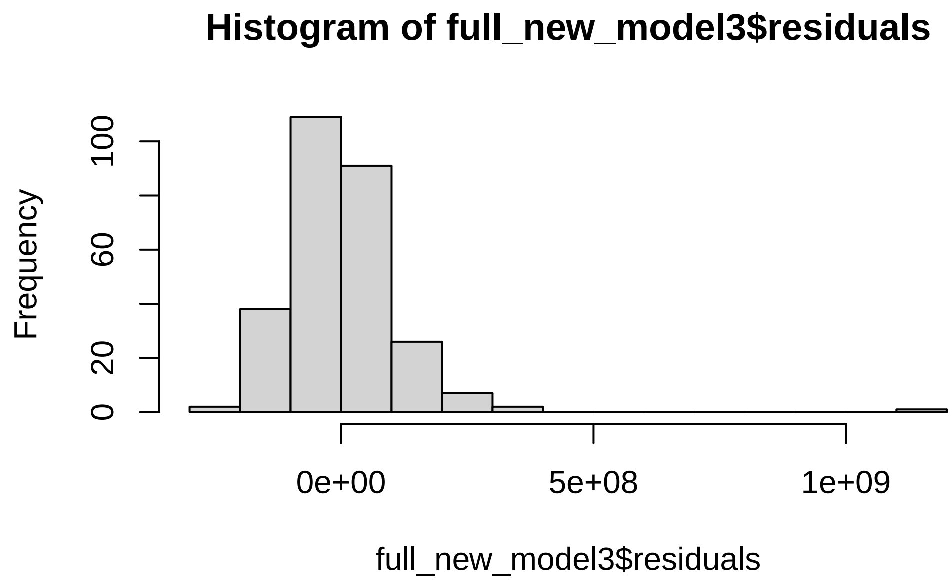
Model 3 Using New Data with Adjustment for Large Spotify Playlist Value

```
full_new_model3 <- lm(new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts
summary(full_new_model3)
```

```
Call:
lm(formula = new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$bpm + new_data$`danceability_%` +
    new_data$`valence_%` + new_data$`energy_%` + new_data$`acousticness_%` +
    new_data$`instrumentalness_%` + new_data$`liveness_%` + new_data$`speechiness_%` +
    new_data$released_month)

Residuals:
       Min         1Q     Median         3Q        Max
-224486853  -65551820  -11086931   49644408 1137846723

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    83915389   77968814   1.076  0.28283
new_data$artist_count         -18269578    7981847  -2.289  0.02291 *
new_data$in_spotify_charts      1310281     466113   2.811  0.00532 **
new_data$in_spotify_playlists     94754      11591   8.174 1.44e-14 ***
new_data$bpm                     149916     276475   0.542  0.58813
new_data$`danceability_%`       -284609     613910  -0.464  0.64333
new_data$`valence_%`             844393     394096   2.143  0.03310 *
new_data$`energy_%`             -218302     616313  -0.354  0.72348
new_data$`acousticness_%`       -339435     391148  -0.868  0.38633
new_data$`instrumentalness_%`    292224     734681   0.398  0.69115
new_data$`liveness_%`           -684405     534639  -1.280  0.20167
new_data$`speechiness_%`       -1112521     712775  -1.561  0.11981
new_data$released_month10      63194208   40489693   1.561  0.11983
new_data$released_month11     118017456   36010470   3.277  0.00119 **
new_data$released_month12      15027847   34921465   0.430  0.66732
new_data$released_month2       39565122   34803456   1.137  0.25669
new_data$released_month3       12909709   34134973   0.378  0.70560
new_data$released_month4       21214058   36553518   0.580  0.56219
new_data$released_month5      -25694379   30712864  -0.837  0.40361
new_data$released_month6      -16348925   34374679  -0.476  0.63476
new_data$released_month7        2511956   41606287   0.060  0.95191
new_data$released_month8      110263823   44674360   2.468  0.01424 *
new_data$released_month9       42352117   43733114   0.968  0.33376
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.23e+08 on 253 degrees of freedom
Multiple R-squared:  0.3489,    Adjusted R-squared:  0.2923
F-statistic: 6.162 on 22 and 253 DF,  p-value: 3.693e-14
```




```
    Shapiro-Wilk normality test

data:  full_new_model3$residuals
W = 0.79639, p-value < 2.2e-16


    studentized Breusch-Pagan test

data:  full_new_model3
BP = 35.031, df = 22, p-value = 0.03845


    Durbin-Watson test

data:  full_new_model3
DW = 2.1195, p-value = 0.8399
alternative hypothesis: true autocorrelation is greater than 0
```



Even this model the doesn't meet our assumptions, we should look at cutting down the variables to reduce the model.

Model 3 Variable Selection:

```
step(full_new_model3, direction = "backward", scope = formula(full_new_model3))
```

```
Start:  AIC=10304.34
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$bpm + new_data$`danceability_%` +
    new_data$`valence_%` + new_data$`energy_%` + new_data$`acousticness_%` +
    new_data$`instrumentalness_%` + new_data$`liveness_%` + new_data$`speechiness_%` +
    new_data$released_month

                                Df  Sum of Sq        RSS   AIC
- new_data$`energy_%`            1 1.8972e+15 3.8277e+18 10302
- new_data$`instrumentalness_%`  1 2.3924e+15 3.8282e+18 10302
- new_data$`danceability_%`      1 3.2501e+15 3.8291e+18 10303
- new_data$bpm                   1 4.4462e+15 3.8303e+18 10303
- new_data$`acousticness_%`      1 1.1388e+16 3.8372e+18 10303
- new_data$`liveness_%`          1 2.4781e+16 3.8506e+18 10304
```

```
<none>                                        3.8258e+18 10304
- new_data$`speechiness_%`       1 3.6840e+16 3.8627e+18 10305
- new_data$`valence_%`           1 6.9421e+16 3.8953e+18 10307
- new_data$artist_count          1 7.9224e+16 3.9051e+18 10308
- new_data$in_spotify_charts     1 1.1950e+17 3.9453e+18 10311
- new_data$released_month       11 4.7882e+17 4.3047e+18 10315
- new_data$in_spotify_playlists  1 1.0105e+18 4.8363e+18 10367

Step:  AIC=10302.48
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$bpm + new_data$`danceability_%` +
    new_data$`valence_%` + new_data$`acousticness_%` + new_data$`instrumentalness_%` +
    new_data$`liveness_%` + new_data$`speechiness_%` + new_data$released_month

                                Df  Sum of Sq        RSS   AIC
- new_data$`instrumentalness_%`  1 2.2316e+15 3.8300e+18 10301
- new_data$`danceability_%`      1 3.0369e+15 3.8308e+18 10301
- new_data$bpm                   1 4.8199e+15 3.8326e+18 10301
- new_data$`acousticness_%`      1 9.7593e+15 3.8375e+18 10301
- new_data$`liveness_%`          1 2.7042e+16 3.8548e+18 10302
<none>                                        3.8277e+18 10302
- new_data$`speechiness_%`       1 3.5935e+16 3.8637e+18 10303
- new_data$`valence_%`           1 7.0452e+16 3.8982e+18 10306
- new_data$artist_count          1 8.1897e+16 3.9096e+18 10306
- new_data$in_spotify_charts     1 1.1906e+17 3.9468e+18 10309
- new_data$released_month       11 4.8589e+17 4.3136e+18 10314
- new_data$in_spotify_playlists  1 1.0167e+18 4.8444e+18 10366

Step:  AIC=10300.64
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$bpm + new_data$`danceability_%` +
    new_data$`valence_%` + new_data$`acousticness_%` + new_data$`liveness_%` +
    new_data$`speechiness_%` + new_data$released_month

                                Df  Sum of Sq        RSS   AIC
- new_data$`danceability_%`      1 2.9645e+15 3.8329e+18 10299
- new_data$bpm                   1 5.4828e+15 3.8355e+18 10299
- new_data$`acousticness_%`      1 9.4595e+15 3.8394e+18 10299
<none>                                        3.8300e+18 10301
- new_data$`liveness_%`          1 2.7992e+16 3.8580e+18 10301
- new_data$`speechiness_%`       1 3.8234e+16 3.8682e+18 10301
- new_data$`valence_%`           1 6.8463e+16 3.8984e+18 10304
- new_data$artist_count          1 8.3908e+16 3.9139e+18 10305
```

```
- new_data$in_spotify_charts     1 1.1737e+17 3.9473e+18 10307
- new_data$released_month       11 4.9181e+17 4.3218e+18 10312
- new_data$in_spotify_playlists  1 1.0304e+18 4.8604e+18 10364

Step:  AIC=10298.85
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$bpm + new_data$`valence_%` +
    new_data$`acousticness_%` + new_data$`liveness_%` + new_data$`speechiness_%` +
    new_data$released_month

                                Df  Sum of Sq        RSS   AIC
- new_data$bpm                   1 6.4880e+15 3.8394e+18 10297
- new_data$`acousticness_%`      1 7.7375e+15 3.8407e+18 10297
- new_data$`liveness_%`          1 2.5789e+16 3.8587e+18 10299
<none>                                        3.8329e+18 10299
- new_data$`speechiness_%`       1 4.1864e+16 3.8748e+18 10300
- new_data$`valence_%`           1 6.6921e+16 3.8999e+18 10302
- new_data$artist_count          1 8.6835e+16 3.9198e+18 10303
- new_data$in_spotify_charts     1 1.1473e+17 3.9477e+18 10305
- new_data$released_month       11 4.9479e+17 4.3277e+18 10310
- new_data$in_spotify_playlists  1 1.0296e+18 4.8626e+18 10362

Step:  AIC=10297.32
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`acousticness_%` +
    new_data$`liveness_%` + new_data$`speechiness_%` + new_data$released_month

                                Df  Sum of Sq        RSS   AIC
- new_data$`acousticness_%`      1 9.8135e+15 3.8492e+18 10296
- new_data$`liveness_%`          1 2.5445e+16 3.8649e+18 10297
<none>                                        3.8394e+18 10297
- new_data$`speechiness_%`       1 4.3256e+16 3.8827e+18 10298
- new_data$`valence_%`           1 6.9253e+16 3.9087e+18 10300
- new_data$artist_count          1 9.0735e+16 3.9302e+18 10302
- new_data$in_spotify_charts     1 1.1893e+17 3.9584e+18 10304
- new_data$released_month       11 4.9234e+17 4.3318e+18 10309
- new_data$in_spotify_playlists  1 1.0406e+18 4.8801e+18 10362

Step:  AIC=10296.03
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`liveness_%` +
    new_data$`speechiness_%` + new_data$released_month
```

```
                                Df  Sum of Sq        RSS   AIC
- new_data$`liveness_%`          1 2.1465e+16 3.8707e+18 10296
<none>                                        3.8492e+18 10296
- new_data$`speechiness_%`       1 4.1746e+16 3.8910e+18 10297
- new_data$`valence_%`           1 7.0030e+16 3.9193e+18 10299
- new_data$artist_count          1 8.6496e+16 3.9357e+18 10300
- new_data$in_spotify_charts     1 1.2486e+17 3.9741e+18 10303
- new_data$released_month       11 5.1625e+17 4.3655e+18 10309
- new_data$in_spotify_playlists  1 1.0544e+18 4.9036e+18 10361

Step:  AIC=10295.56
new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`speechiness_%` +
    new_data$released_month

                                Df  Sum of Sq        RSS   AIC
<none>                                        3.8707e+18 10296
- new_data$`speechiness_%`       1 4.0406e+16 3.9111e+18 10296
- new_data$`valence_%`           1 7.0646e+16 3.9413e+18 10299
- new_data$artist_count          1 8.4033e+16 3.9547e+18 10300
- new_data$in_spotify_charts     1 1.2394e+17 3.9946e+18 10302
- new_data$released_month       11 5.2435e+17 4.3950e+18 10309
- new_data$in_spotify_playlists  1 1.0639e+18 4.9347e+18 10361


Call:
lm(formula = new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`speechiness_%` +
    new_data$released_month)

Coefficients:
                  (Intercept)          new_data$artist_count
                     52537080                      -18539347
   new_data$in_spotify_charts  new_data$in_spotify_playlists
                      1316723                          96523
         new_data$`valence_%`       new_data$`speechiness_%`
                       740796                       -1144141
    new_data$released_month10        new_data$released_month11
                     63583765                      120243279
    new_data$released_month12         new_data$released_month2
                     11589495                       41893321
     new_data$released_month3         new_data$released_month4
```

```
                 7713771                         21137439
     new_data$released_month5       new_data$released_month6
                -28279221                       -18625771
     new_data$released_month7       new_data$released_month8
                  7827337                       112420377
     new_data$released_month9
                 38088583
```

```
stepmodel3 <- lm(formula = new_data$streams ~ new_data$artist_count + new_data$in_spotify_cha
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`speechiness_%` +
    new_data$released_month)
summary(stepmodel3)
```

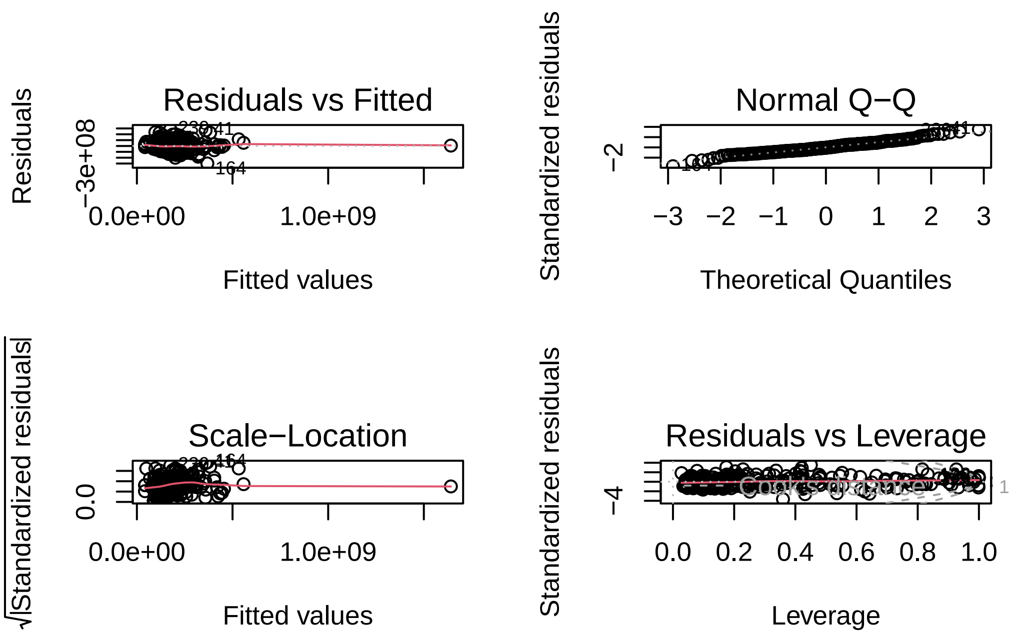
```
Call:
lm(formula = new_data$streams ~ new_data$artist_count + new_data$in_spotify_charts +
    new_data$in_spotify_playlists + new_data$`valence_%` + new_data$`speechiness_%` +
    new_data$released_month)

Residuals:
       Min         1Q     Median         3Q        Max
-224407067  -66060987  -12607225   44929912 1146624909

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    52537080   36445351   1.442  0.15064
new_data$artist_count         -18539347    7818340  -2.371  0.01846 *
new_data$in_spotify_charts      1316723     457236   2.880  0.00431 **
new_data$in_spotify_playlists     96523      11440   8.438 2.31e-15 ***
new_data$`valence_%`             740796     340721   2.174  0.03060 *
new_data$`speechiness_%`       -1144141     695828  -1.644  0.10133
new_data$released_month10      63583765   39458473   1.611  0.10831
new_data$released_month11     120243280   35287566   3.408  0.00076 ***
new_data$released_month12      11589495   34484761   0.336  0.73709
new_data$released_month2       41893321   34551685   1.212  0.22643
new_data$released_month3        7713771   33418773   0.231  0.81764
new_data$released_month4       21137439   35960023   0.588  0.55718
new_data$released_month5      -28279221   30103527  -0.939  0.34840
new_data$released_month6      -18625771   33646006  -0.554  0.58034
new_data$released_month7        7827337   40543491   0.193  0.84706
new_data$released_month8      112420377   44068210   2.551  0.01132 *
new_data$released_month9       38088583   42752120   0.891  0.37380
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 122200000 on 259 degrees of freedom
Multiple R-squared:  0.3412,    Adjusted R-squared:  0.3005
F-statistic: 8.385 on 16 and 259 DF,  p-value: 3.01e-16
```

By looking at the model and numerical tests, we chose to reduce the model. Our Shapiro Wilk test showed non-normality in our residuals thus we chose to further analyze. Using stepwise selection in the backward direction, our model with the lowest AIC is below. This model has the variables artist count, in spotify charts, in spotify playlists, valence, and speechiness. Through further analysis, we looked at the significance levels and scatterplot with streams, and opted to remove speechiness.

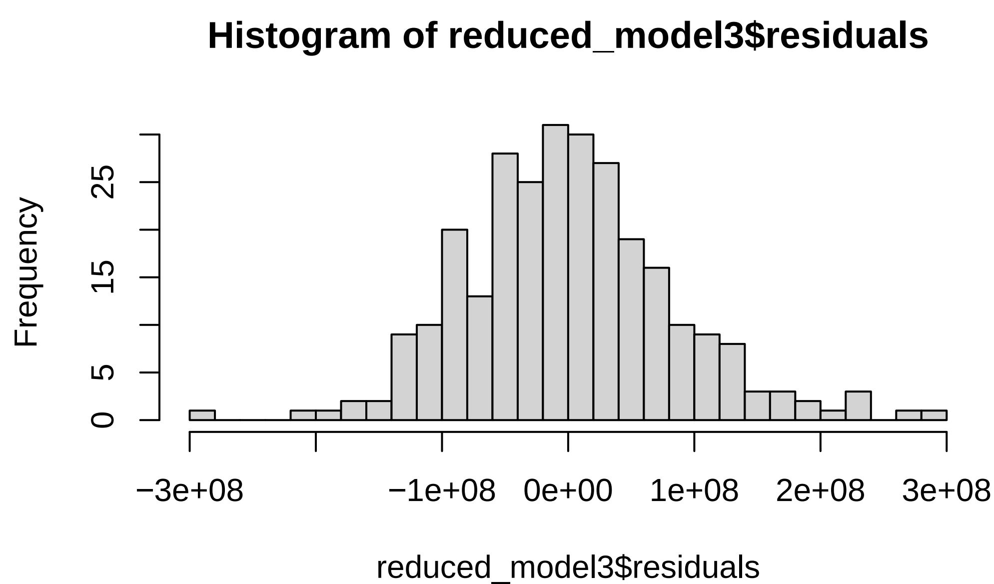
In an effort to increase R^2, we chose to see if interactions between the variables would help explain more variability.

```
reduced_model3 <- lm(streams~artist_count * new_data$`valence_%` * in_spotify_charts *  relea
```




```
    Shapiro-Wilk normality test
```

```
data:  reduced_model3$residuals
W = 0.98553, p-value = 0.006965


    Durbin-Watson test

data:  reduced_model3
DW = 2.1568, p-value = 0.7786
alternative hypothesis: true autocorrelation is greater than 0


    studentized Breusch-Pagan test

data:  reduced_model3
BP = 119.23, df = 91, p-value = 0.02517
```

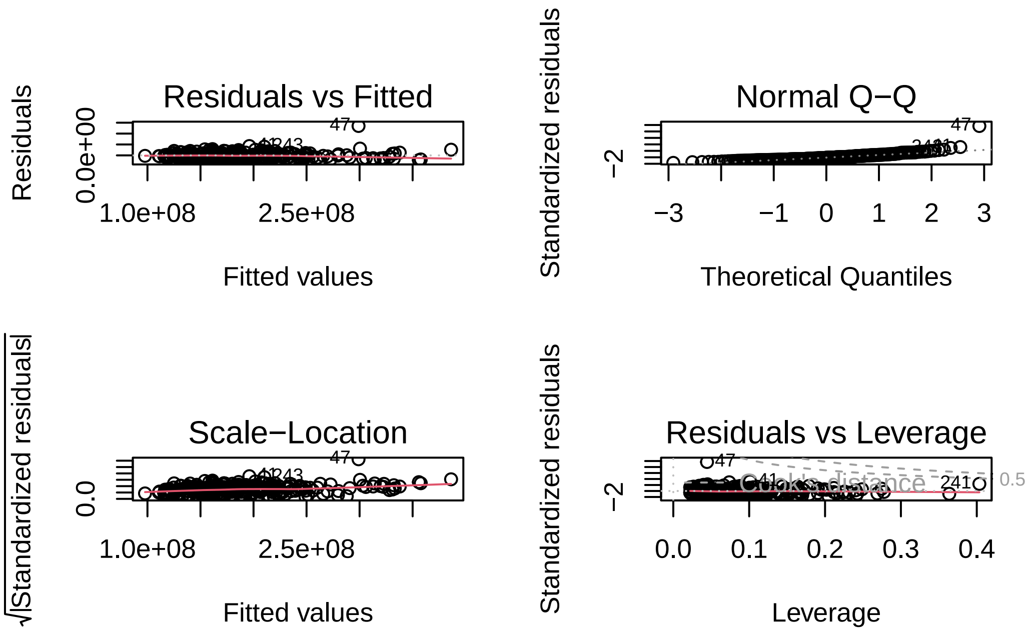


After the analyzing the summary table for our reduced model, as well as our plots and numerical tests, we decided to only keep valence and released month with interactions as the other interactions did not have significant p-values. We wanted to explore the idea that more positive songs and the month with which they were released had an affect on the number of streams so we chose to interpret this model more in depth.

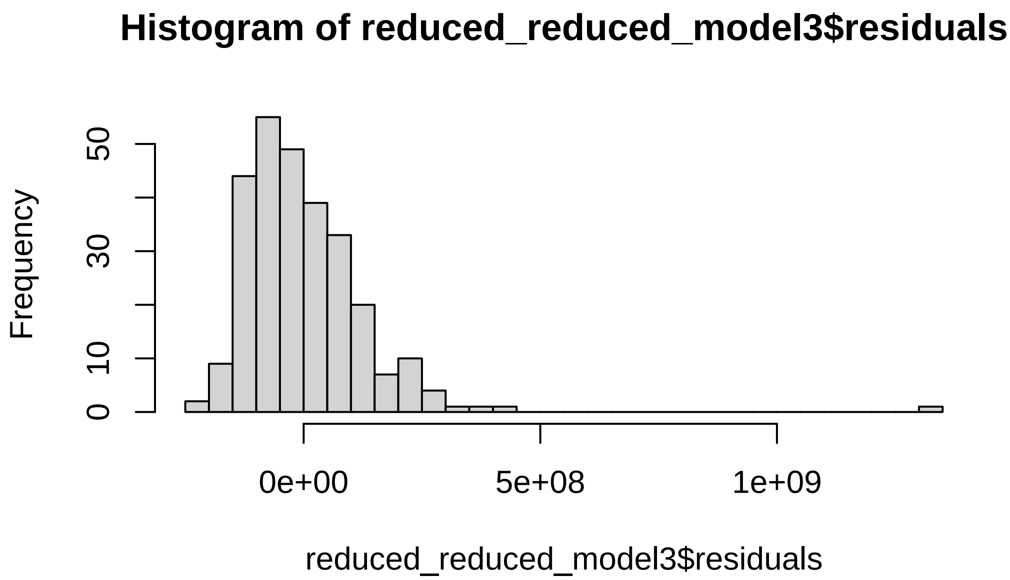
```r
reduced_reduced_model3 <- lm(new_data$streams~new_data$`valence_%`*new_data$released_month)
summary(reduced_reduced_model3)
```

```

Call:
lm(formula = new_data$streams ~ new_data$`valence_%` * new_data$released_month)

Residuals:
       Min         1Q     Median         3Q        Max
-219194617  -83724087  -20634770   62877334 1348972971

Coefficients:
                                               Estimate Std. Error t value
(Intercept)                                    78922365   82673755   0.955
new_data$`valence_%`                            1445604    1237192   1.168
new_data$released_month10                     135683428  102497089   1.324
new_data$released_month11                     149358394  107098644   1.395
new_data$released_month12                     141574820  114977839   1.231
new_data$released_month2                      157474913  121550941   1.296
new_data$released_month3                       86171759  111960260   0.770
new_data$released_month4                       72579134  104902941   0.692
new_data$released_month5                       33337974   98148356   0.340
new_data$released_month6                       84739768  104775500   0.809
new_data$released_month7                       37974021  148346070   0.256
new_data$released_month8                      315530387  127260526   2.479
new_data$released_month9                      230274615  148416704   1.552
new_data$`valence_%`:new_data$released_month10  -1024579    1756112  -0.583
new_data$`valence_%`:new_data$released_month11    -58610    1785733  -0.033
new_data$`valence_%`:new_data$released_month12  -2071876    1841833  -1.125
new_data$`valence_%`:new_data$released_month2   -2045501    1892040  -1.081
new_data$`valence_%`:new_data$released_month3   -1566420    1802816  -0.869
new_data$`valence_%`:new_data$released_month4    -760061    1864003  -0.408
new_data$`valence_%`:new_data$released_month5    -823210    1509407  -0.545
new_data$`valence_%`:new_data$released_month6   -1996603    1674913  -1.192
new_data$`valence_%`:new_data$released_month7      -9642    2345175  -0.004
new_data$`valence_%`:new_data$released_month8   -3487974    2140084  -1.630
new_data$`valence_%`:new_data$released_month9   -3773396    2442584  -1.545
                                              Pr(>|t|)
(Intercept)                                     0.3407
new_data$`valence_%`                            0.2437
new_data$released_month10                       0.1868
new_data$released_month11                       0.1644
```

```
new_data$released_month12                           0.2194
new_data$released_month2                            0.1963
new_data$released_month3                            0.4422
new_data$released_month4                            0.4897
new_data$released_month5                            0.7344
new_data$released_month6                            0.4194
new_data$released_month7                            0.7982
new_data$released_month8                            0.0138 *
new_data$released_month9                            0.1220
new_data$`valence_%`:new_data$released_month10   0.5601
new_data$`valence_%`:new_data$released_month11   0.9738
new_data$`valence_%`:new_data$released_month12   0.2617
new_data$`valence_%`:new_data$released_month2    0.2807
new_data$`valence_%`:new_data$released_month3    0.3857
new_data$`valence_%`:new_data$released_month4    0.6838
new_data$`valence_%`:new_data$released_month5    0.5860
new_data$`valence_%`:new_data$released_month6    0.2344
new_data$`valence_%`:new_data$released_month7    0.9967
new_data$`valence_%`:new_data$released_month8    0.1044
new_data$`valence_%`:new_data$released_month9    0.1236
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 142600000 on 252 degrees of freedom
Multiple R-squared:  0.1284,    Adjusted R-squared:  0.04887
F-statistic: 1.614 on 23 and 252 DF,  p-value: 0.04076
```




```
	Durbin-Watson test

data:  reduced_reduced_model3
DW = 2.0816, p-value = 0.7565
alternative hypothesis: true autocorrelation is greater than 0


	studentized Breusch-Pagan test

data:  reduced_reduced_model3
BP = 14.524, df = 23, p-value = 0.9109
```



After performing the interactions transformation, we can see that the residuals vs. the fitted plot is closer to a random scatter, the qqnorm seems way better than before and the histogram looks closer to a normal distribution, if we overlook the outlier in the extreme left. The numerical tests for independence and constant variance also pass and our assumptions are met. Finally, we can say that this is the best model we have so far with all the transformations.

- Linearity: Our scatter plot after the transformation remains showing a linear relationship.

- Independence: In our dwtest, our p-value was not significant so we fail to reject the null. Thus, our model shows independence. This suggests that there is no significant autocorrelation in the residuals of our model.

- Homoscedasticity: Our bptest has a non-significant p-value as well, in which we do not reject the null of homoscedasticity. The variance of the residuals is constant.

- Normality of Errors: Based on our Q-Q plot, our residuals are located along the line, demonstrating that the residuals are normally distributed.

VIF

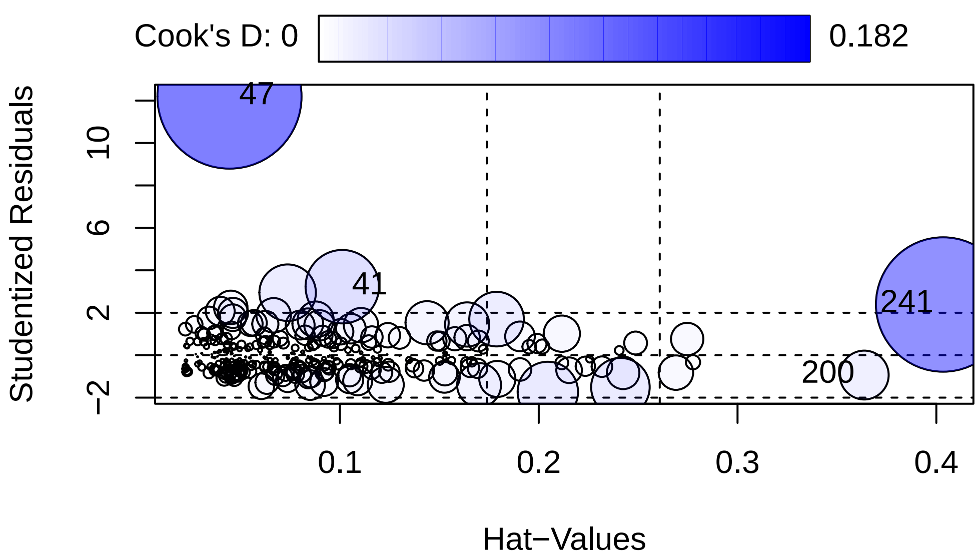
```r
vif(reduced_reduced_model3)
```

```
                                                      GVIF Df GVIF^(1/(2*Df))
new_data$`valence_%`                          1.151596e+01  1        3.393517
new_data$released_month                       1.027772e+09 11        2.568217
new_data$`valence_%`:new_data$released_month 1.859744e+09 11        2.638389
```

Based on the vif analysis for our model, since every value for the coefficients was greater than 1 and less than our cutoff value of 5, we can deduce that multicollinearity between our variables is not a problem for our model. Thus we did not omit any of these variables on multicollinearity issues.

Outliers and Influential Points

From the graph and looking at the Cook's distance, it seems we are safe since none of the points seem to cross the Cook's lines. There might be some outliers like the 47th and 241st observation. We can conduct the outlier test to confirm this. However, we do not have sufficient background information whether we can omit these outliers.




```
       StudRes       Hat      CookD
41    3.2288986 0.1011778 0.04713692
47   12.1897598 0.0444409 0.18158918
200  -0.9353511 0.3637370 0.02084993
241   2.3887678 0.4034102 0.15782355
```

```
   rstudent unadjusted p-value Bonferroni p
47 12.18976          3.7155e-27   1.0255e-24
```

Our influence test identified 4 influential points. These were observations 41, 47, 200, and 241. We then used the outlier test to define possible outliers.

The Cook's distance didn't show any outliers or leverage points but from our qq plot after the transformation, we see that the 47th and the 241st observation are seemingly different than the trend. From the outlierTest, we confirmed that the 47th observation was an outlier. However, we do not have the data on the background of data collection and there is no reasonable and justifiable reason to exclude this outlier. Thus, we decide to keep it in our analysis but also to overlook it at our convenience. We could just explore them a little further and see if there is any valid justification to omit them from the data or not. For now, we opted to keep them in.

Transformed Model 3 Interpretation

The intercept represents the baseline estimate for the reference month, indicating that the expected number of streams is 78,922,365 in January when the valence percentage is 0%. However, the p-value is relatively small, suggesting that this value is not statistically significant and is not a reliable estimate. This is reasonable because a valence of 0% is unlikely.

The valence coefficient suggests that for every 1% increase in valence, the number of streams increases by approximately 1,445,604 in January.

The coefficients for each released month estimate the additional expected streams for songs released in that particular month compared to January when valence is 0%. The only statistically significant release month is August. For songs released in August, we expect to have 315,530,387 more streams than January when valence_% is 0%. Our p-value 0.0138 is small, which means that= it is statistically significant, suggesting that August releases may perform better regardless of valence percentage. For the rest of the months, the lack of statistical significance shows no clear evidence of a consistent difference from January in terms of number of streams.

The interaction terms show how the effect of valence on streams changes across different months relative to January. None of these interaction coefficients are statistically significant, meaning that there is no strong evidence that the relationship between valence percentage and streams differs meaningfully across months.

Transformed Model 3 Conclusion and Final Insights

Null Hypothesis: H_0: $\beta_1 = 0, \beta_2 = 0, ..., \beta_p = 0$ None of the months are statistically significant to impact the number of streams.

Alternate Hypothesis: H_A: $\beta_1 \neq 0, \beta_2 \neq 0, ..., \beta_p \neq 0$ At least one of the months are statistically significant to impact the number of streams.

The R^2 value for the model is 0.1284, suggesting that only 12.84% of the variability in the number of streams is explained by release month and valence percentage. The R^2 adjusted is 0.04887. Although this value is not high, we choose to analyze this model as it is the best model that satisfies all the assumptions.

The F-statistics is 1.614, with a p-value 0.04076. This means that we reject the null hypothesis, suggesting that we have sufficient evidence to conclude that valence percentage or at least one of the released months is significant in predicting the number of streams.