**STAT 512 Final Report – Group 14**

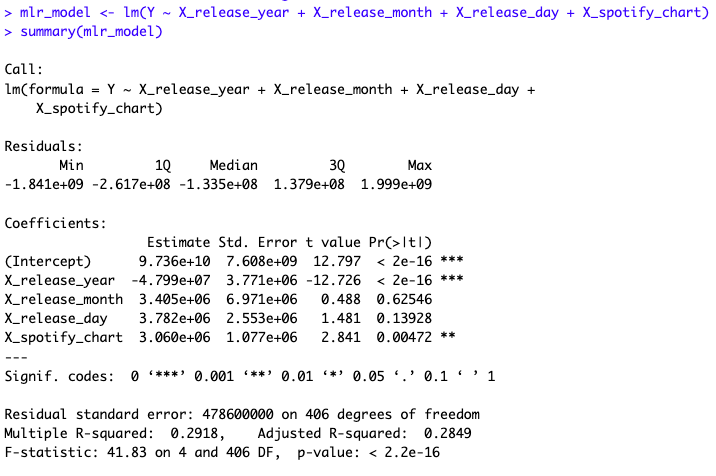
**Research Question 1: Jiya**

The purpose of this analysis was to investigate how the release timing of a song, including the year, month, and day, impacts its streaming statistics and presence on various music platforms, with the research question: How does the release timing of a song (year, month, and day) impact its streaming statistics and presence on various music platforms? We hypothesized that songs released earlier in the year (considering month and day) are likely to have more streams on Spotify. Understanding how the timing of a song's release influences its popularity and performance on streaming platforms is crucial for the music industry. This information can guide artists, record labels, and music platforms in optimizing release strategies to maximize reach and engagement.

The dataset includes key features such as Release year, Release Month, Release Day, Spotify Chart Position, and the total number of streams of the song on Spotify.

Prior to model building, the dataset underwent a thorough screening process to identify and address any data anomalies, missing values, or outliers.

A full multiple linear regression (MLR) model was constructed, including the release year, month, day, and Spotify chart position.

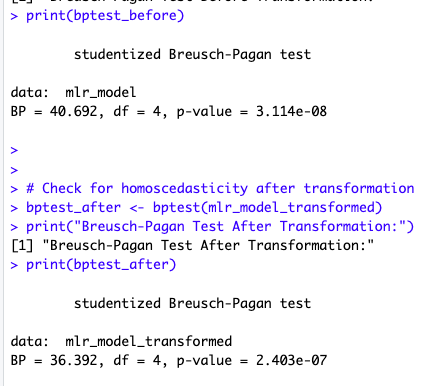
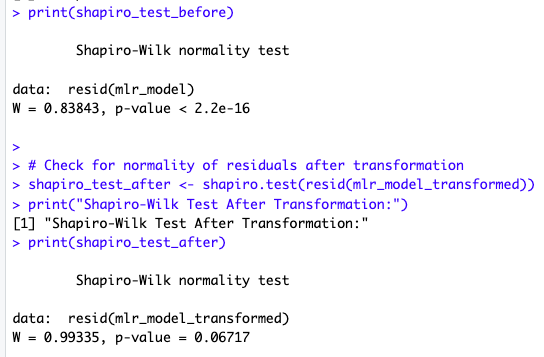


To assess the individual impact of each predictor, separate single linear regression (SLR) models were fitted for each predictor against the dependent variable (streams). Assumptions of linearity, normality of residuals, and homoscedasticity were evaluated through residual plots, Shapiro-Wilk tests, and Breusch-Pagan tests, respectively.

The Shapiro-Wilks normality test assesses whether the residuals of a regression model follow a normal distribution. In each case, the p-values obtained from the tests are remarkably small (all less than 2.2e-16), indicating strong evidence against the null hypothesis of normality. The test statistics (W values) for the residuals of models based on release year, release month, release day, and Spotify chart position are 0.82001, 0.81883, 0.81724, and 0.83301, respectively. These values are considerably lower than 1, suggesting deviations from normality. Therefore, the assumption of normality for the residuals in these models is rejected.

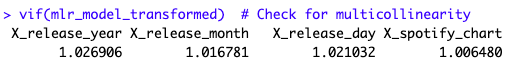
The results from the studentized Breusch-Pagan tests assess the presence of heteroscedasticity in the residuals of the regression models. For the model based on release year, the test yields a significant p-value of 2.225e-08, indicating evidence of heteroscedasticity. In contrast, the models based on release month, release day, and Spotify chart position show non-significant p-values of 0.8291, 0.6687, and 0.1789, respectively. This suggests no strong evidence of heteroscedasticity in the residuals for these models. Heteroscedasticity implies that the variability of the residuals is not constant across all levels of the predictor variables, which can impact the efficiency of parameter estimates.

Given non-normality in the residuals, a Box-Cox transformation was applied to the response variable (streams) to achieve normality.

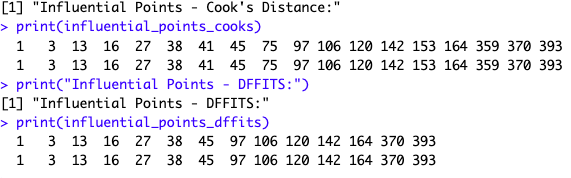


The results of the Shapiro-Wilk normality tests before and after the Box-Cox transformation provide insights into the distribution of residuals in the linear regression model. Prior to the transformation, the test indicated a departure from normality (W = 0.83843, p-value < 2.2e-16), indicating a non-normal distribution of residuals. Following the Box-Cox transformation, there was a substantial improvement in normality, as evidenced by the higher W statistic (W = 0.99335) and a p-value of 0.06717, suggesting that the transformed residuals are reasonably consistent with a normal distribution. Regarding homoscedasticity, the Breusch-Pagan tests both before and after the transformation assessed the variance of residuals. The test before transformation yielded a significant result (BP = 40.692, p-value = 3.114e-08), indicating heteroscedasticity. However, after the Box-Cox transformation, the Breusch-Pagan test still indicated some heteroscedasticity, though the p-value (p-value = 2.403e-07) suggests a substantial improvement. Overall, the Box-Cox transformation has successfully addressed non-normality in residuals.

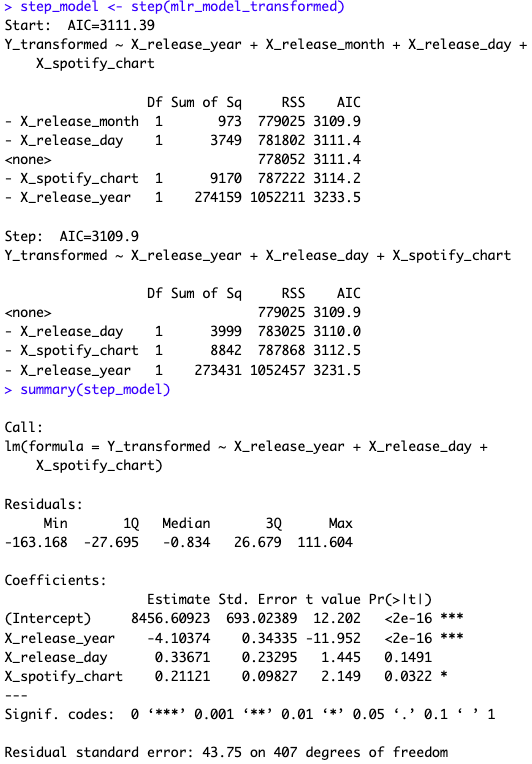
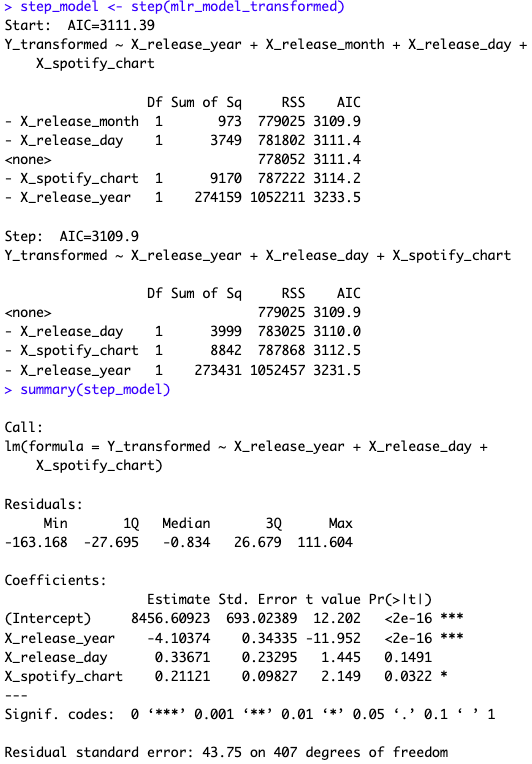
Variance inflation factors (VIF) were computed to evaluate multicollinearity among predictors. In this case, the VIF values for the predictors are all close to 1. Generally, VIF values close to 1 indicate low levels of multicollinearity, suggesting that these predictors are not highly correlated with each other. Hence, there is no substantial risk of multicollinearity influencing the stability and reliability of the regression coefficients.



Influential points were identified using DFFITS, Cook's distance, and DFBETAS. The output reveals influential points detected using DFBETAS, Cook's Distance, and DFFITS measures. Specifically, rows with indices 1, 3, 13, 16, 27, 38, 41, 45, 75, 97, 106, 120, 142, 153, 164, 359, 370, and 393 are identified as potentially influential based on their impact on estimated coefficients, the overall model, and predicted values.

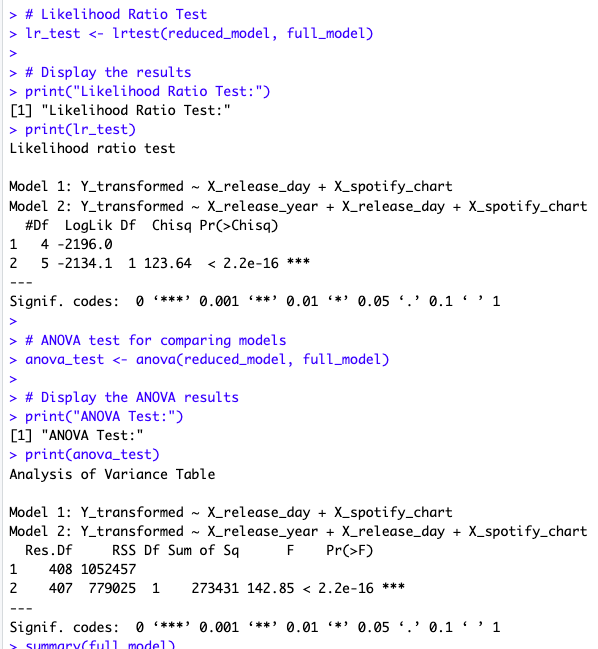


The stepwise elimination process involved iteratively testing the removal of each predictor to minimize the BIC, with the final model achieving an AIC of 3109.9. The selected model exhibits a residual standard error of 43.75, indicating the average deviation of observed data points from the model's predictions. The coefficients provide insights into the impact of each predictor on the transformed response variable. Notably, the negative coefficient for release year implies that, on average, newer songs have fewer streams. Release day, with a positive coefficient, suggests that songs released later in the month tend to have more streams, while a higher Spotify chart position is associated with increased streams. The adjusted R-squared value of 0.2578 indicates that approximately 25.78% of the variability in the transformed streaming statistics is explained by the predictors in the model. The F-statistic with a p-value less than 2.2e-16 further supports the overall significance of the model. In summary, the selected model, derived through stepwise regression, provides a statistically significant representation of the relationship between release timing variables and song streaming statistics.

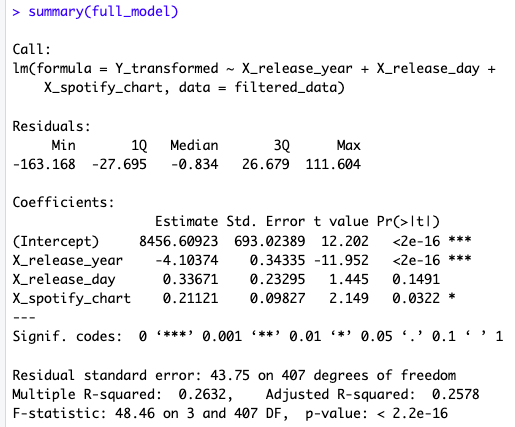


To test the model, the null hypothesis [:There is no impact of release year on the transformed streams ()] was compared to the alternative hypothesis[: There is an impact of release year on the transformed streams ()]. The equation for the reduced model was hence as follows: , with the hypothesis testing for . The reduced model considers only the impact of release day and Spotify chart position on transformed streams. The hypothesis test will determine whether excluding release year significantly impacts the model. The full model’s equation is thus as follows:

The full model includes the impact of release year, release day, and Spotify chart position on transformed streams.



The likelihood ratio test and ANOVA results shown above provide compelling evidence of the significant impact of release timing variables on the streaming statistics of songs. The likelihood ratio test compares two models, one including only release day and Spotify chart position and the other incorporating release year. The resulting chi-squared statistic of 123.64, with a p-value well below the significance threshold, strongly supports the full model with release year. This implies that the inclusion of release year significantly improves the model's ability to explain variations in streaming statistics. The ANOVA results reinforce this, indicating that the full model, including release year, release day, and Spotify chart position, is superior in explaining the variability in the transformed streaming data compared to the reduced model without release year.



In conclusion, contrary to the hypothesis that "Songs released earlier in the year (considering month and day) are likely to have more streams on Spotify," the analysis reveals a nuanced picture. The negative coefficient for release year (-4,103,740) suggests that, on average, newer songs experience fewer streams. In contrast, the positive coefficients for release day (0.33671) and Spotify chart position (0.21121) indicate that songs released later in the day and those with higher positions on the Spotify chart are associated with more streams. The release month did not exhibit a statistically significant impact on streaming statistics.

To quantify, the results indicate that, holding other factors constant, a one-unit increase in the release day is associated with a 0.34 increase in streams, and a one-unit increase in Spotify chart position is associated with a 0.21 increase in streams. The multiple R-squared value of 0.2632 suggests that the model explains approximately 26.32% of the variance in the streaming statistics.

While the initial hypothesis is not entirely supported, the nuanced findings underscore the importance of considering various temporal factors in predicting a song's popularity on Spotify.

**Research Question 2: Ethan Carter**

How do the audio attributes of a song (beats per minute, mode, energy, and danceability) influence its ranking on Spotify charts? I will be testing the following hypothesis: Songs with higher BPM, major mode, and higher danceability percentage are likely to rank higher on Spotify charts

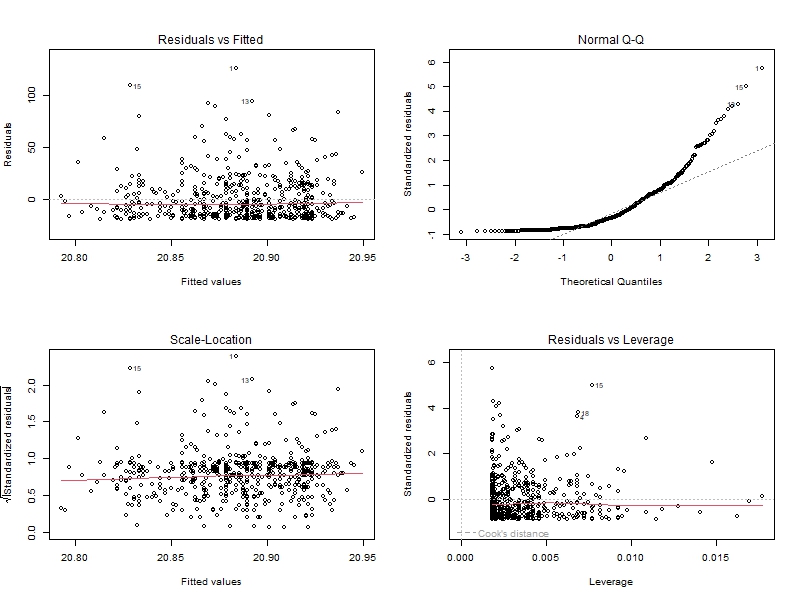
Investigating the intricate relationship between a song's inherent audio attributes—such as tempo, key, mode, and danceability—and its prominence on streaming platforms like Spotify and Apple Music is the focal point of this analysis. Harnessing a comprehensive dataset encompassing diverse song attributes, streaming rankings, and listener engagement metrics across platforms, this study seeks to unveil the nuanced impact of these sonic elements on a song's chart performance. Understanding how these fundamental audio traits interplay with audience preferences and platform algorithms holds immense potential for reshaping music strategies. Insights gained here could empower musicians, labels, and streaming services to refine their content curation, playlist algorithms, and promotional tactics. Unveiling the significance of beats per minute, mode, energy, and danceability in song success can redefine the narrative of what captivates audiences and ascends to the top tiers of the ever-evolving music charts.

in\_spotify\_charts, bpm, mode, danceability\_%, and energy\_% are specifically address the audio attributes (tempo, key, mode, and danceability) of songs and their influence on the rankings of songs on Spotify and Apple Music charts.

Prior to model building, the dataset underwent a thorough screening process to identify and address any data anomalies, missing values, or outliers. Entries that are not present in the Spotify charts were filtered out to focus on songs with meaningful platform presence.

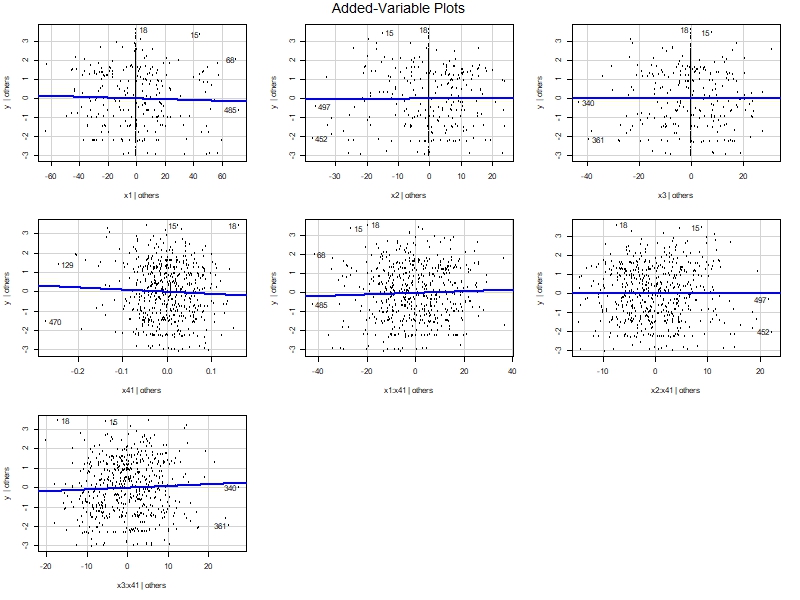
A full multiple linear regression (MLR) model was constructed, including beats per minute, mode, energy, and danceability. Quadratic terms for relevant predictors were introduced to capture potential nonlinear relationships.

To assess the individual impact of each predictor, separate single linear regression (SLR) models were fitted for each predictor against the dependent variable (chart rating). Residuals were checked for normality and homoscedasticity. Assumptions of linearity, normality of residuals, and homoscedasticity were evaluated through residual plots, Shapiro-Wilk tests, and Breusch-Pagan tests, respectively.



From this initial analysis we can see that there is a constant variance, but the data is not distributed normally. In the following section we will handle this issue of non-normality to improve the p-value from the Shapiro test.

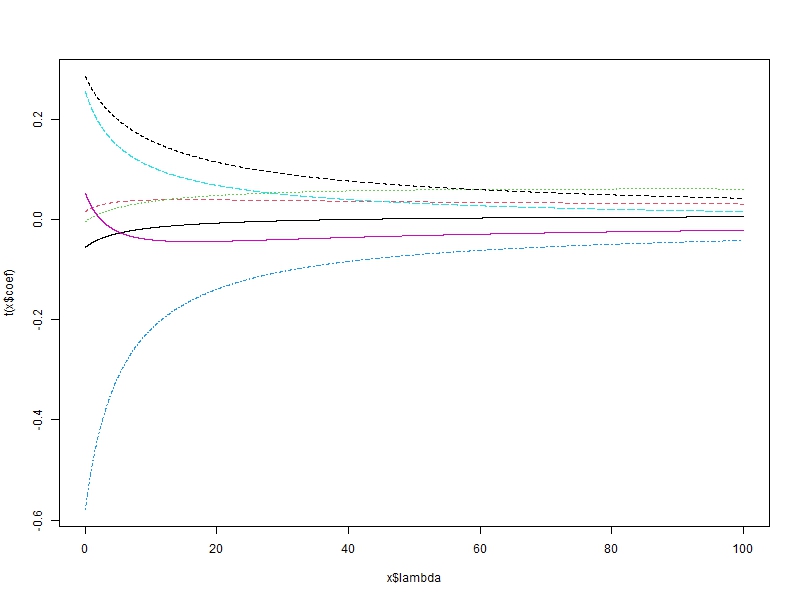
To assess the marginal impacts of each predictor in the multiple linear regression (SLR) model was fitted for each predictor against the dependent variable (chart rating). Residuals were checked for normality and homoscedasticity. Assumptions of multicollinearity, normality of residuals, and homoscedasticity were evaluated through added variable plots. variance inflation factors, Shapiro-Wilk tests, and Breusch-Pagan tests.

Added variable (or partial regression) plot allows us to visualize the relationship between each of the predictor variables and the response this is done while holding all other predictors constant if the coefficient changes drastically from the SLR model this can suggest multicollinearity.



The categorical variable (Mode) presents high levels of multicollinearity suggesting that it presents little value as a predictor variable.

Box cox transformation and robust regression were run on the SLR models and the MLR model to address the non-normality. Improving the p-value from 2.2e-16 to 0.0001285.



The GCV that is related to lambda is the max value of the sequence. Since Lambda manages the shrinkage of the distribution, we would want to balance the size of lambda with the amount of variance that is described by the predictor variable, (i.e. the max might not be as desirable as the sequence grows to infinity). This suggests that there is some amount of multicollinearity that cannot be removed without trivializing the model.

Validation was run comparing the original full model to the full model after being transformed to see if these measures improved the predictive capabilities of the model.

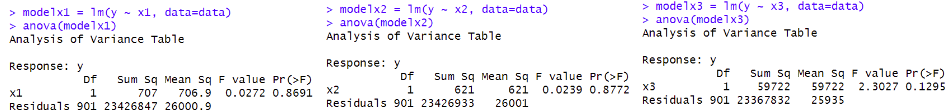
The results suggest that the transformations have helped the model.

The SLR analysis, the added variable plots and the large sum squared error present in the model suggests that these predictor variables do not play statistically significant role in predicting how a song is going to perform on the Spotify charts. This can suggest that there are different predictors that can create more effective models, or the diverse genres can also play a role in the challenge of predicting a song's performance on the charts.

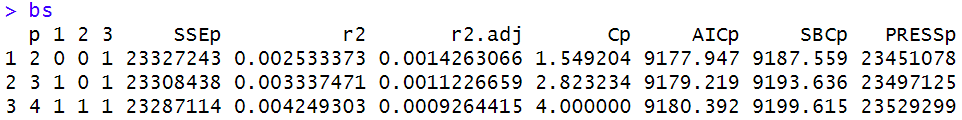
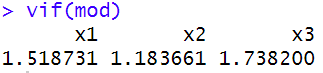
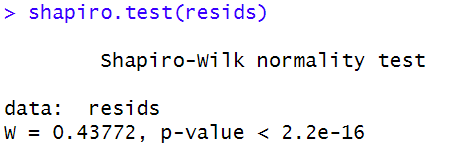
**Research Question 3: Saksham Gupta**

How does the energy of a song affect its positioning in Shazam charts? I will test the hypothesis that energy will have an impact on a song's ranking on Shazam Charts given valence and acousticness are already in the model. The variables used are “Shazam Charts”, “Valence”, “Acousticness”, and “Energy”. “Valence”, “Acousticness”, and “Energy” are the predictors.

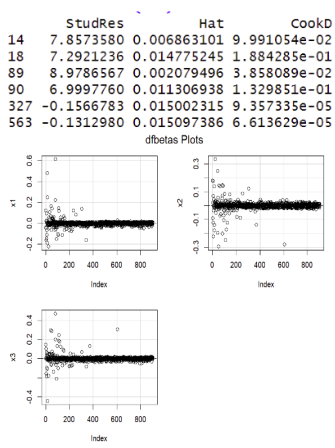
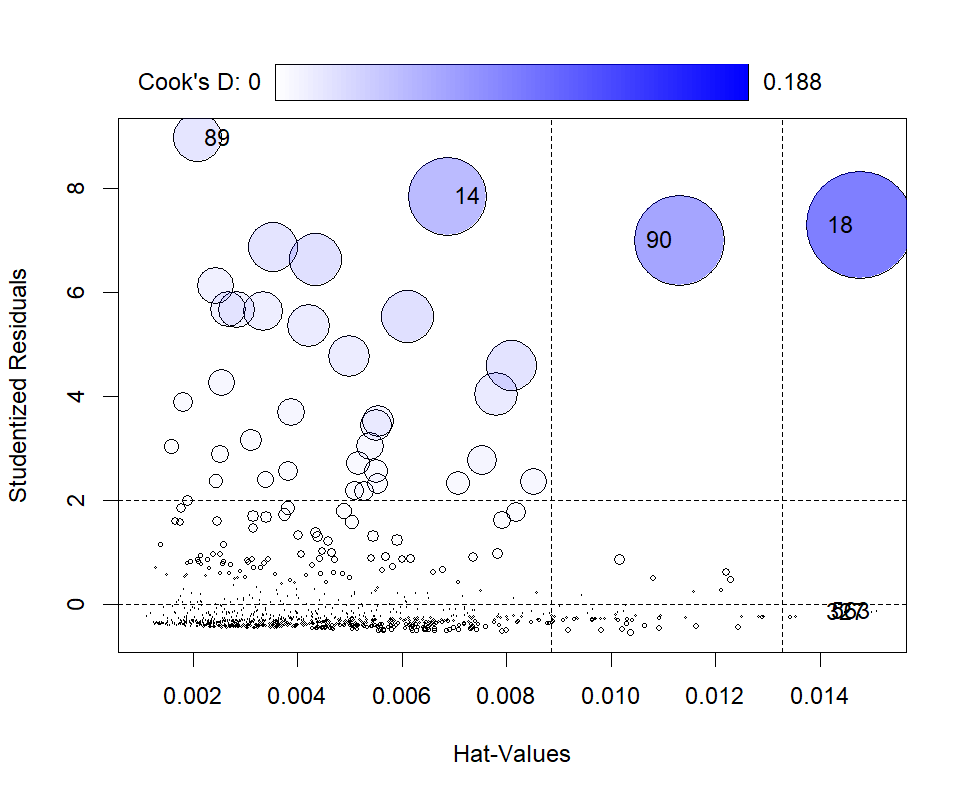
Only the data in columns in\_shazam\_charts, acousticness\_., valence\_., energy\_. are selected from the data. Some of the values are NA. Some values have a comma ( , ) in them. Some values are 0. All 0 values are dropped. All NA values are dropped. All commas are removed such that values like 1,032 become 1032.

SLRs are created with each individual predictor. None of these models are statistically significant, as shown by the following Anova tables:  


Then, BestSub method is used to check for the best subset. As shown by the following output, the BestSub method also agrees that the model with all 3 variables is the best one.

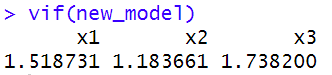
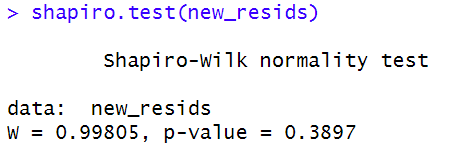
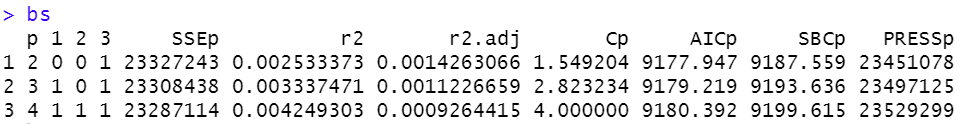
  


Shapiro Test shows that the residuals are not normal. The VIF values show that there is no multicollinearity present.

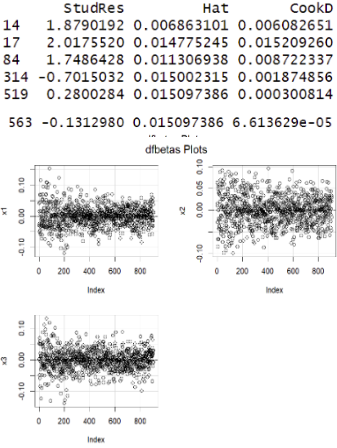
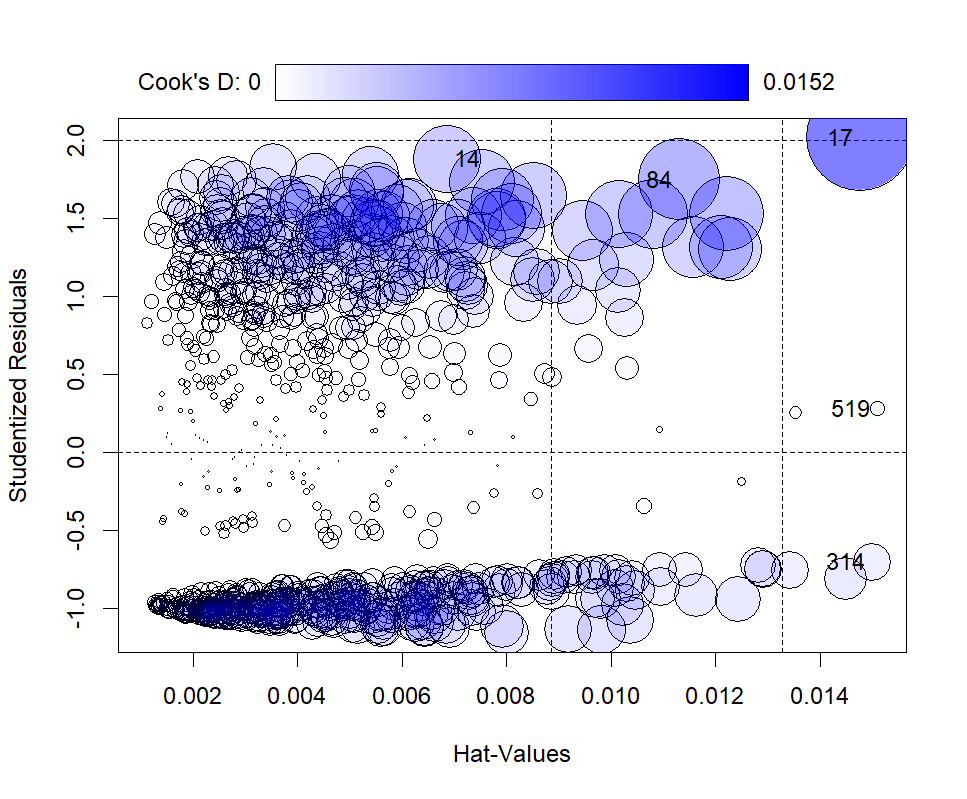


As shown here, there are very few influential points. Even those few influence points do not have a very large Cook’s Distance. Thus, we can conclude that these points do not mess up our model significantly.

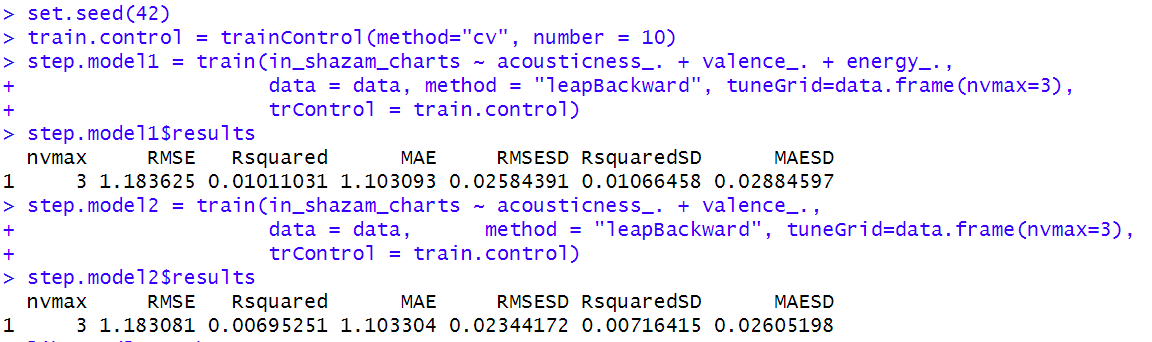
Once again, the BestSub method indicates that the model with all 3 variables is the best one.



Shapiro Test shows that the residuals are normal. These VIF values show that there is no multicollinearity present.



As shown above, there are very few influence points. Even those few influence points do not have a very large Cook’s Distance. Thus, we can conclude that these points do not interfere with our model significantly.



Validation is performed with model1 as full model and model2 as reduced model. The full model has all 3 variables, and the reduced model has all variables except energy. The results strongly indicate that the full model is better than the reduced model.

Hypothesis: Energy of a song has an impact on the song's ranking in Shazam's charts when Acousticness and Valence are already in the model.  
H0: β3 = 0 : Reduced model: ShazamRanking = β0 + β1{Acousticness} + β2{Valence} + β3{Energy}  
Ha: Not H0:Full model: ShazamRanking = β0 + β1{Acousticness} + β2{Valence}

A screenshot of a computer code

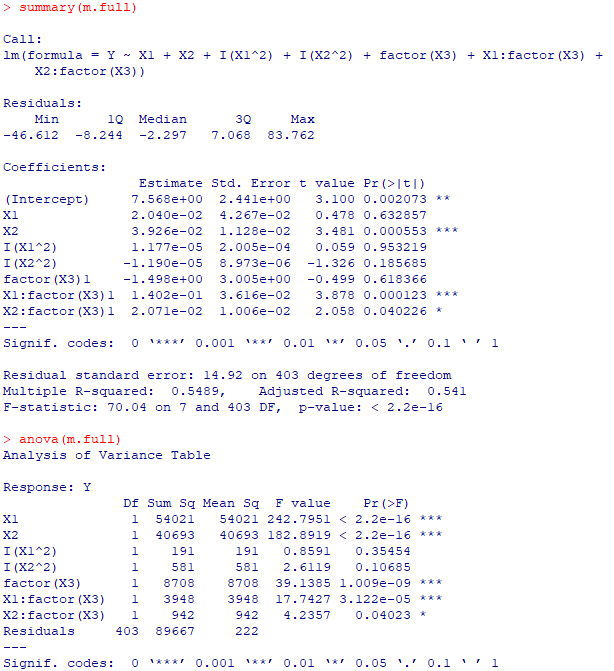
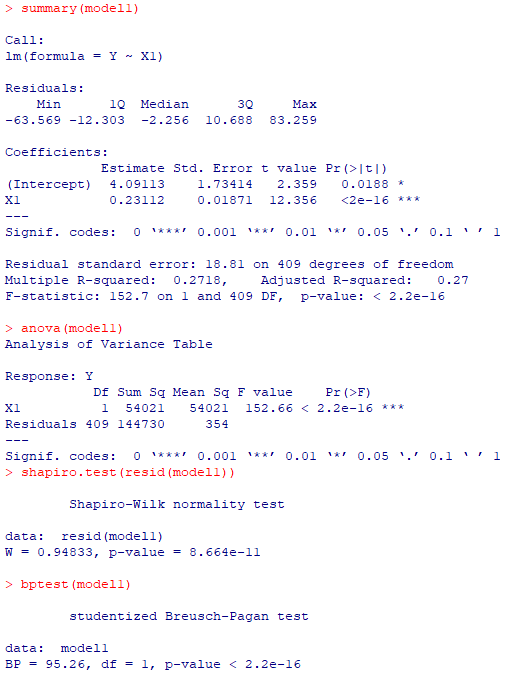
Description automatically generated

Both tests reject the null hypothesis at a 94.8% Confidence interval. We can conclude that energy significantly affects the position of a song in Shazam charts.

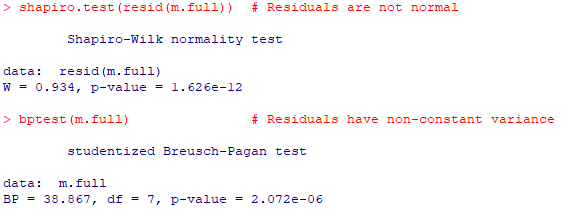
**Research Question 4: Zach Davis**

Does the presence of a song on Deezer’s charts have any predicting power on its position in Spotify’s Charts? This research question is of some interest, as it is reasonable to think that people who listen to music on Deezer have different tastes than people who listen to music on Spotify. However, it could also be that popular songs on Deezer are also popular on Spotify. The testable hypothesis for this question is as follows: Songs that are not on Deezer’s charts and songs that are on Deezer’s charts have the same average Spotify Chart Ranking for any ranking in Shazam’s and Apple Music’s charts. In this question, the response variable Y is Spotify Chart Ranking. Apple Chart ranking is X1, Shazam Chart ranking is X2, and the presence of a song on Deezer’s charts is categorical variable X3. X3=0 is a song that is not in Deezer’s Charts, and X3=1 otherwise. In this research question, we filter out songs that are not present in Spotify, Apple, or Shazam charts to deal with strong non-constant residual variance and non-normal residual issues.

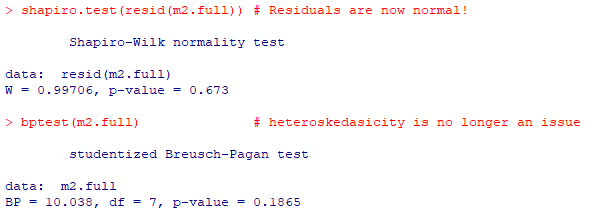
When running analysis on SLR models Y~X1, Y~X2, and Y~factor(X3), the summary and Anova tables show that they X1, X2, and factor(X3) have significant linear impact on Y when they are by themselves. However, Shapiro tests and bptests show that there is severe non-normality and non-constant variance issues with the residuals. For the sake of space, only the results of Y~X1 (model1) are shown below:



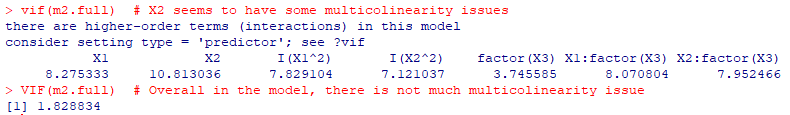
For our full model (m.full), we include the 2nd order factors for X1 and X2, as well as X1 and X2’s interaction effect with X3. When checking the summary and type II ANOVA tables for this model, we can see that the Intercept, X2, X1:factor(X3), and X2:factor(X3) has significant marginal linear impact, and that X1, X2, X1:factor(X3) and X2:factor(X3) have significant marginal effect on the residuals.



When performing assumption checks on this model, we can see that there is still major non-normality and non-constant variance issues with the residuals. To deal with this, we perform a Box-Cox transformation on Y with a lambda value of .2626263, with the transformed Y value becoming Y2. After performing the Box-Cox transformation, we can see that the non-normality and non-constant variance issues are fully handled.

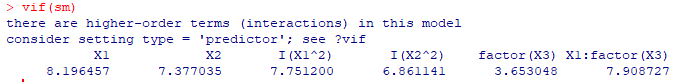


On this transformed model, we see that X2 has some multicollinearity issues, while the model altogether has no multicollinearity issues.

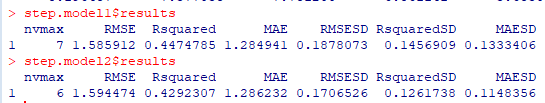


Additionally, we see find that the influential points in the model are 29, 44, 74, 124, 187, 202, 254, 330, 370, 393, 394, and 396. As there are only minor multicollinearity issues and not many influential points compared to n (n=411), we will move on to our model selection.

Using stepwise selection, only the interaction factor X2:factor(X3) is removed. This is to be expected, as it has shown little marginal linear impact and little marginal effect on the residuals compared to other variables in the model. With this selected model, there are no non-normality or non-constant variance issues with the residuals. Additionally, the multicollinearity issue with X2 is ameliorated.

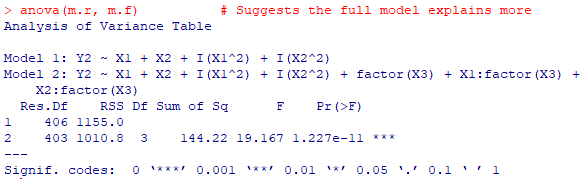


When comparing the full model and the selected model, we get the results shown below:



Step.model1 in the figure above is the full model, while step.model2 is the model chosen by the stepwise selection. In the selected model, we see that RMSE is slightly higher and Rsquared is slightly lower. This may suggest that the selected model is wors. However, the fact that RMESD, RsquaredSD, and MAESD are all much lower in the selected model than the full model suggests that on average, the selected model is more consistent when predicting than the full model.

Now that model selection is done, we can test our hypothesis. The hypotheses are H0: β5= β6= β7=0 and Ha: not H0.



The results of the test between these two indicate that the full model (Ha) explains more of the variation in Y2 than the reduced model (H0) does. In context, that means that the presence of a song on Deezer’s charts contributes meaningfully to the prediction of a song’s rank on Spotify Charts.

**Results and Conclusions - Jiya**

The project's major findings reveal crucial insights into the dynamics of music streaming and chart performance, . Firstly, the analysis underscores the impact of release timing on streaming statistics, showcasing that newer songs tend to garner fewer streams. Songs released later in the year and month exhibit a notable increase in streaming numbers. The inclusion of the release year significantly improves the model's predictive capabilities, offering valuable guidance for stakeholders aiming to understand and optimize promotional strategies.

The project addressed the relationship between BPM, mode, energy, danceability, and Spotify chart rankings. Despite constructing a full MLR model, challenges emerged, indicating that these audio traits might not singularly determine chart positions. Single linear regression analysis, added variable plots, and sum squared errors suggested a potential influence of other predictors or the diverse genre landscape, posing challenges in predicting Spotify chart performance solely based on these audio features.

Furthermore, the project delves into the nuanced relationship between song energy and positioning on Shazam charts. Rigorous statistical analyses, including single linear regression, Best Subset method, and validation processes, reveal that energy has a significant impact on a song's Shazam chart ranking when considering valence and acousticness. This finding adds a layer of sophistication to our understanding of the factors influencing chart performance.

Finally, the project’s investigation of the predictive power of Deezer's charts on Spotify's charts found that the presence of a song on Deezer's charts is a meaningful predictor of its position on Spotify. Its higher consistency in predictions suggests that Deezer's chart data plays a valuable role in forecasting a song's performance on Spotify. This revelation provides industry professionals with nuanced insights into the interconnectedness of music platforms and user preferences.