Assessing Key Factors of Song Popularity on Spotify

Jeffrey Foster, Anika Achari, Abdullah Ahmed, Jinglong Yu and Peter Yee

1. INTRODUCTION

1.1. MOTIVATION

Digital streaming platforms have rapidly emerged as the dominant force in the music and audio industries in recent years. The ubiquity and accessibility of such streaming platforms have transformed the way in which audiences consume musical content. With millions of users utilizing these platforms and streaming billions of songs daily, these platforms have amassed extensive datasets that provide ample opportunity for extensive analysis regarding the elements that influence song popularity and other trends pertaining to music consumption (1-2). Spotify, which is currently the most popular streaming platform, leverages advanced algorithms to analyze various audio features of songs including tempo, danceability, and energy, among others (2). We intend to explore the relationship between these structural elements of a song and its popularity, with the aim of deepening our understanding of the intricate factors that contribute to the widespread appeal of certain songs. Ultimately, this analysis serves to enhance our comprehension of the factors that drive listener engagement and underscores the complexity of musical appeal in the modern era.

1.1.1. Context

This project is rooted in the domain of digital music streaming, a rapidly evolving industry that has redefined the global music landscape (1). Streaming platforms such as Spotify and Apple Music, among others, have transformed traditional music distribution by offering unprecedented access to music for audiences while also enabling artists to reach listeners worldwide. Spotify, as the leading platform, connects millions of users to their favorite songs and generates extensive data on song attributes and user engagement (2).

Within this aforementioned context, our project focuses on analyzing Spotify song data to understand the dynamics of song popularity. By analyzing the attributes of music, we aim to better understand the factors influencing song popularity and provide practical insights for artists, producers, and marketers. Our work is situated at the intersections of data science, music consumption patterns and to a larger extent, the creative industries.

1.1.2. Problem

The problem this project aims to address is developing a better understanding of the factors that drive the popularity of songs on Spotify. Although the platform hosts millions of songs and millions of songs are uploaded and streamed daily, only a small percentage manage to achieve significant popularity. Thus, it is imperative to understand the factors that make certain songs resonate more profoundly with listeners than others. Despite the availability of data for song attributes such as tempo, energy, and danceability, there is minimal clarity on how these factors specifically contribute to popularity. This lack of understanding poses challenges for artists and marketers who aim to create music that performs well on streaming platforms. To tackle this pressing problem, our project proposes a data-driven solution by building a regression model that

analyzes the relationship between song attributes and popularity. This approach will provide insights into the most influential features, offering valuable guidance for interested parties who wish to create music that aligns with listener preferences and thrives in the competitive streaming landscape.

1.1.3. Challenges

The challenges faced in this analysis stem from the complexities of ensuring the regression model meets all theoretical assumptions required for valid statistical inference. Despite efforts to refine the model, several significant issues arose. First, the final model exhibited signs of non-linearity, indicating that the relationship between the predictors and the dependent variable "track popularity" could not fully be captured by a linear equation (3). This presented difficulties in maintaining the simplicity and interpretability inherent to linear regression models, often requiring transformations or the addition of interaction terms to better fit the data.

Additionally, the model contained multiple outliers, which could have influenced the regression results (3). Identifying and addressing these outliers was challenging because their removal or transformation had to be carefully balanced to avoid compromising the dataset's integrity. These influential data points complicated efforts to draw reliable conclusions about the predictors' effects on track popularity.

The residuals of the final model also deviated from normality and exhibited unequal variance, further compounding the difficulties of this analysis. The lack of normality affected the validity of hypothesis tests and confidence intervals, while heteroscedasticity compromised the reliability of parameter estimates. Remedying these issues requires substantial adjustments, such as exploring alternative modeling approaches and introducing weighted regression techniques, which added complexity to the process.

In addition, the dataset possessed certain limitations as it failed to account for factors such as who produced the song, how much a song was marketed and the use of the song in other media; all of which are known to significantly contribute to a song's popularity. These aforementioned issues underscore the challenges of working with real-world datasets, where theoretical assumptions are often violated (4). Addressing these issues required iterative testing, exploration of non-linear models, and careful consideration of trade-offs to improve model robustness while maintaining its interpretability.

1.2. OBJECTIVES

1.2.1. Overview

The primary objective of this work is to analyze and identify the key factors that influence the popularity of songs on Spotify. By investigating the relationship between various auditory

attributes and track popularity, we aim to determine which characteristics of a song contribute most significantly to popularity. This analysis will enhance our understanding of the elements that enhance musical appeal and additionally, provide insight regarding the nuances of musical appeal. Furthermore, the insights gained from this work will be valuable for creators, artists, producers and marketers, and potentially allow them to leverage these insights to optimize their content and facilitate stronger connections with their audiences.

1.2.2. Goals & Research Questions

The primary research question we aim to answer is "*What factors contribute to a Spotify song's popularity?"* To successfully explore this question, we intend to accomplish the three following goals:

- Develop a regression model to examine the relationship between attributes and a song's popularity.
- Identify the variables that most significantly influence track popularity.
- Uncover insights into the factors driving a song's success on the platform.

The insights generated from this work will be invaluable for musical artists and marketers, enabling them to create music that resonates with listeners and achieves success on the Spotify platform.

2. METHODOLOGY

2.1. DATA

The dataset we intend to utilize in the execution of this project is a publicly available dataset on Kaggle entitled "Audio Features and Lyrics of Spotify Songs." (5) This dataset has been aggregated by a third party on Kaggle from Spotify's open data using their web API (6). The dataset was created in 2020, with songs selected from a collection of playlists developed by the authors of the dataset. It is unclear how these playlists were developed and if they used true random sampling means to obtain this data, or if bias exists in the sampled songs. Given the breadth of variance across the independent variables, it is assumed that the dataset authors used random sampling to identify these songs.

The dataset (5) is composed of information pertaining to over 18,000 songs and contains a comprehensive blend of both qualitative and quantitative attributes. Each row of the dataset represents a unique song with 24 associated variables. The 25 total elements contained in our dataset are outlined in the table below.

Table 1: Variables and associated data type, description and inclusion/exclusion in regression model.

Variable	Data Type	Description	Included in Regression (Y/N)?
track_id	String (Qualitative)	Song unique ID. The primary key of the dataset.	N
track_name	String (Qualitative)	Song Name	N
track_artist	String (Qualitative)	Song Artist	N
lyrics	String (Qualitative)	Lyrics for the song	N
track_popularity	Integer (Quantitative)	Song Popularity (0-100) where a higher value indicates more popular	Y
track_album_id	String (Qualitative)	Album unique ID	N
track_album_name	String (Qualitative)	Song album name	N
track_album_release_date	String (Qualitative)	Date when album released	N
playlist_name	String (Qualitative)	Name of playlist song is ong	N
playlist_id	String (Qualitative)	Playlist ID	N
playlist_genre	String (Qualitative)	Primary genre of the playlist	Y
playlist_subgenre	String (Qualitative)	Secondary genre of the playlist	Y
danceability	Float (Quantitative)	Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.	Y

energy	Float (Quantitative)	A measure from 0.0 - 1.0 representing a perceptual measure of intensity and activity. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.	Y
key	Integer (Quantitative)	The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.	Y
loudness	Float (Quantitative)	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks.	Y
mode	Integer (Quantitative)	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.	Y
speechiness	Float (Quantitative)	Speechiness detects the presence of spoken words in a track. The more exclusively speech-	Y

		like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.	
acousticness	Float (Quantitative)	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.	Y
instrumentalness	Float (Quantitative)	Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values	Y

		above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.	
liveness	Float (Quantitative)	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.	Y
valence	Float (Quantitative)	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).	Y
tempo	Float (Quantitative)	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.	Y
duration_ms	Float (Quantitative)	Duration of song in milliseconds	Y
language	String (Qualitative)	Language of the	Y

		lyrics	

A collection of dataset attributes were excluded from our methodology. This is due to these variables containing a high number of near-unique strings, which we have not learned how to process using a multiple regression approach. These variables may lead to interesting insights that could aid us in answering our guiding questions, and it is encouraged that future iterations of this project utilize these variables, if possible.

2.2. APPROACH

To answer our primary research question, our solution will be to leverage our dataset to identify the best possible multiple linear regression model that will predict the "track_popularity" variable. We will utilize the R software environment for statistical computing, which, in conjunction with several built-in packages, facilitates the use of several multiple linear regression modelling techniques and statistical tests.

2.3. WORKFLOW

To conduct our analysis, we implemented a series of multiple linear regression modeling tasks and statistical tests. These included Multiple Linear Regression, Interaction Model Selection, Stepwise Model Selection, and All-Possible-Regression Selection. The specific steps for each method are detailed below:

2.3.1. Multiple Linear Regression Model Selection

One of the regression models we will complete is the Multiple Linear Regression Model. The goal of this step is to develop a multiple linear regression equation where all of the independent variables included in the final model are statistically significant. An initial linear regression model will be tested that includes all of our valid quantitative and qualitative independent variables with our dependent variable, *"track_popularity"*. With this initial model complete, we will then sequentially reduce our model to its optimal final model with only statistically significant independent variables. An Individual Coefficients Test (t-test) will be completed to test if each individual variable should be included or excluded in our reduced model. The hypotheses for the t-test are:

$$H_0 : \beta_k = 0$$
$$H_a : \beta_k \neq 0 \ (k = 1, 2, 3, ..., p)$$

We will test all of our independent variables against an α of 0.05. Completing the t-test for each variable will result in a reduced multiple linear regression equation that we will leverage in our Interaction Model Selection.

2.3.2. Interaction Model Selection

After identifying our reduced multiple linear regression equation, we noticed that certain variables might have a greater impact when interacting with others. For example, combining the attributes of “loudness” and “genre” reveals that some soft folk songs with high loudness scores might decrease in popularity, while loud sounds in rock music could enhance audience engagement. This observation led us to incorporate interaction terms into our model.

To do this, we first added all possible interaction terms to the model. We then conducted individual t-tests for each variable based on the model output, using a significance level of $\alpha = 0.05$. Variables that did not pass the test were dropped.The hypotheses for the t-test are:

$$H_0 : \beta_k = 0$$
$$H_a : \beta_k \neq 0 \ (k = 1, 2, 3, \ldots, p)$$

An important consideration during this process was the handling of categorical factors with numerous interaction terms. For instance, while most subgenres interacting with energy passed the t-test, the interaction between hip-hop and energy had a high p-value. Since this result aligns with reasonable expectations, we chose to retain it in the model.

2.3.3. Stepwise Model Selection

Another key regression model we implemented was Stepwise Model Selection. This method systematically includes or excludes independent variables based on their statistical significance and contribution to the overall performance of the model. The primary goal of stepwise regression is to develop an optimized model that balances complexity and predictive accuracy by retaining only the most relevant variables. It ensures that unnecessary predictors are excluded, thereby, minimizing the risk of overfitting while enhancing the interpretability of the model. Stepwise regression is particularly useful in identifying the most influential factors in scenarios where numerous predictors are available, as is the case in our dataset.

2.3.4. All-Possible-Regressions Selection Procedure

We also intend to utilize another algorithm that looks at all possible regressions using all subsets of predictors from our full regression model. This algorithm will select the best subset of predictors for regression of all the subsets of the same size and report metrics for the best regressions for each size of predictor subset. From this, we can select the best model out of the best models of each size from the full regression model by directly comparing the metrics of the regressions like adjusted R-squared or Mallow’s CP criterion.

2.3.5. Final Model Comparison and Selection

To evaluate the four unique multiple regression model development techniques completed above and identify the best overall model, we will isolate two primary metrics that will inform us which of the four models (see above) is the best at predicting the dependent variable.

Adjusted R-squared

The Adjusted R-squared of a model measures the proportion of variance in the dependent variable ("track_popularity") that is explained by the independent variables and has been adjusted for the number of predictor variables in the models. The larger the Adjusted R-squared for a valid model, the better that model is, at explaining the dependent variable.

Residual Standard Error (RSE)

The RSE of a model measures the standard deviation of the unexplained variance, or the average amount that the actual observed values deviate from the regression line. The smaller the RSE, the closer the model prediction values are to the actual observed values.

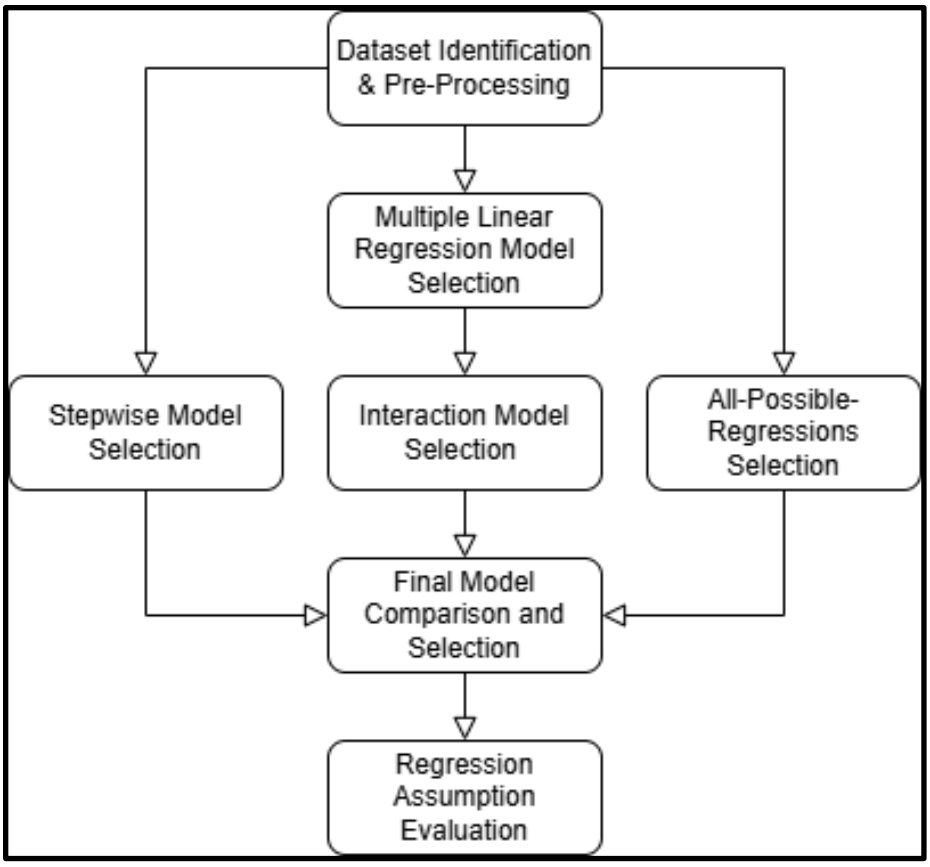
Each model created will be judged based on these two metrics to identify our final selected model.

2.3.6. Regression Assumption Evaluations on Final Model

To ensure the validity and reliability of the final regression model, we rigorously tested it against standard assumptions of multiple linear regression. After identifying the best-fit model using techniques such as the All-Possible-Regressions Selection Procedure and Stepwise Regression, we conducted a series of assumption evaluations.

First, the linearity assumption was tested by examining residual plots for any systematic patterns. The random scatter observed in these plots is used to indicate that the relationship between the independent variables and the dependent variable ("track popularity"), was appropriately modeled as linear.The equal variance assumption, or homoscedasticity, was evaluated by reviewing residuals versus fitted values plots and using the Breusch-Pagan test. To address the normality assumption, we examined the residuals using Q-Q plots. Multicollinearity could be safely ignored due to the presence of higher power variables and interaction terms. Hence, we did not conduct the test. For independence since the data was not related to time, space, or group, we can assume that their measurements are independent. Finally, we analyzed potential outliers using Cook's Distance and leverage statistics. These measures helped identify data points that could disproportionately influence the model. Identified outliers were carefully evaluated to determine whether adjustments, such as removal or transformation, were necessary. By systematically addressing these assumptions, we can see how the final regression model holds up in terms of the required multiple linear regression assumptions.

2.3.7. Dependent Variable Logit Transformation Model Identification



2.3.8: Workflow for Dependent Variable Logit Transformation Model Identification

We have identified a possible issue with our dependent variable ("track_popularity"). It appears to be a bounded percentile of how popular a song is among all songs. However, this type of variable is not the best suited for linear regression, as there is a possibility of our model outputting a predicted value outside of the domain of the dependent variable. To address this issue, we made the decision to perform a logit transformation of the popularity variable in order to obtain a continuous variable without a restrictive domain that would be better suited for linear regression. Therefore, we intend to take our best regression model procedure for the untransformed popularity variable and apply it on the logit transformation of popularity, to compare that model to our original analysis.

3. MAIN RESULTS OF THE ANALYSIS

3.1. RESULTS

Below, we have outlined results for each regression method that we implemented:

3.1.1. Multiple Linear Regression Model Selection

An initial multiple linear regression has been completed on all relevant variables previously identified. The code and output is included in the accompanying appendix document under the *"Multiple Linear Regression Model Selection"* section**.**

To identify the final model, an initial model has been developed that will be reduced using an Individual Coefficients Test (t-test) to evaluate each variable for statistical significance, as outlined in the methodology section. This process has been demonstrated in the accompanying appendix document under the "*Multiple Linear Regression Model Selection*" section**.** In total, we failed to reject five independent variables that had p-values greater than 0.05. The five variables that were removed were"'key", "mode", "speechiness", "acousticness", and "language".

Additionally, a sixth variable was removed due to the presence of multicollinearity between variables. The "playlist_genre" and "playlist_subgenre" variables were found to be multicollinear with one another, resulting in singularities in regression of these variables. This is demonstrated in the accompanying appendix document under the "Variance Inflation Factor (VIF) Test" section**,** which shows the R output for two linear regression equations, one reduced and including the "playlist_genre" variable (with the "playlist_subgenre" removed), and one reduced and including the "playlist_subgenre" variable (with the "playlist_genre" removed). This is to demonstrate the difference in model accuracy between these variables, where the model including the "playlist_subgenre" was found to be more accurate with an improved Adjusted R-Squared value and RSE over the model that preferred the "playlist_genre" variable.

The reduced initial regression equation identified is:

$$
\begin{aligned}
Y_{\text{trackpopularity}} = 63.54 &- 12.72 \cdot \text{big room} - 1.87 \cdot \text{classic rock} + 10.77 \cdot \text{dance pop} \\
&- 4.79 \cdot \text{electro house} - 2.22 \cdot \text{electropop} - 12.47 \cdot \text{gangster rap} \\
&- 6.03 \cdot \text{hard rock} + 9.34 \cdot \text{hip hop} + 3.28 \cdot \text{hip pop} \\
&- 4.48 \cdot \text{indie poptimism} - 4.04 \cdot \text{latin hip hop} + 3.55 \cdot \text{latin pop} \\
&- 12.14 \cdot \text{neo soul} - 16.46 \cdot \text{new jack swing} + 12.15 \cdot \text{permanent wave} \\
&- 7.22 \cdot \text{pop edm} + 10.45 \cdot \text{post-teen pop} - 16.61 \cdot \text{progressive electro house} \\
&- 0.45 \cdot \text{reggaeton} - 8.50 \cdot \text{southern hip hop} + 3.56 \cdot \text{trap} \\
&- 2.22 \cdot \text{tropical} + 2.66 \cdot \text{urban contemporary} \\
&+ 11.54 \cdot \text{danceability} - 18.80 \cdot \text{energy} + 0.856 \cdot \text{loudness} \\
&- 6.785 \cdot \text{instrumentalness} - 3.392 \cdot \text{liveness} - 2.891 \cdot \text{valence} \\
&+ 0.01693 \cdot \text{tempo} - 0.00002987 \cdot \text{duration\_ms}
\end{aligned}
$$

The model identified above has been leveraged in the following Interaction Model Selection.

3.1.2. Interaction Model Selection

After constructing the multiple linear regression model, we compared the adjusted R-squared and residual standard error (RSE) with the previous model. The RSE decreased from 22.67 to 22.5, and the adjusted R-squared increased from 15.2% to 16.86%. This improvement indicates that the updated model explains a greater proportion of the variance in the response variable, “track_popularity”.

The preferred Multiple Linear Regression formula with interaction terms is:

$$
\begin{aligned}
Y_{\text{trackpopularity}} = {} & 56.25 + 25.56 \cdot \text{big room} + 7.30 \cdot \text{classic rock} + 21.06 \cdot \text{dance pop} \\
& + 30.50 \cdot \text{electro house} + 36.50 \cdot \text{electropop} + 12.90 \cdot \text{gangster rap} \\
& - 3.56 \cdot \text{hard rock} + 22.26 \cdot \text{hip hop} + 28.34 \cdot \text{hip pop} \\
& + 12.21 \cdot \text{indie poptimism} + 10.33 \cdot \text{latin hip hop} + 36.12 \cdot \text{latin pop} \\
& + 5.82 \cdot \text{neo soul} - 13.18 \cdot \text{new jack swing} + 8.39 \cdot \text{permanent wave} \\
& + 24.19 \cdot \text{pop edm} + 28.44 \cdot \text{post-teen pop} + 4.24 \cdot \text{progressive electro house} \\
& + 22.44 \cdot \text{reggaeton} - 1.95 \cdot \text{southern hip hop} + 1.73 \cdot \text{trap} \\
& - 3.86 \cdot \text{tropical} + 19.36 \cdot \text{urban contemporary} \\
& + 7.21 \cdot \text{danceability} - 5.09 \cdot \text{energy} + 1.559 \cdot \text{loudness} \\
& - 33.43 \cdot \text{instrumentalness} - 17.22 \cdot \text{liveness} - 2.545 \cdot \text{valence} \\
& + 0.01595 \cdot \text{tempo} - 0.00002592 \cdot \text{duration\_ms} \\
& - 28.28 \cdot \text{big room:energy} - 13.95 \cdot \text{classic rock:energy} - 17.46 \cdot \text{dance pop:energy} \\
& - 29.41 \cdot \text{electro house:energy} - 27.55 \cdot \text{electropop:energy} - 21.47 \cdot \text{gangster rap:energy} \\
& + 4.49 \cdot \text{hard rock:energy} - 18.61 \cdot \text{hip hop:energy} - 12.23 \cdot \text{hip pop:energy} \\
& - 7.61 \cdot \text{indie poptimism:energy} - 6.08 \cdot \text{latin hip hop:energy} - 17.22 \cdot \text{latin pop:energy} \\
& - 11.10 \cdot \text{neo soul:energy} - 20.45 \cdot \text{new jack swing:energy} - 0.45 \cdot \text{permanent wave:energy} \\
& - 22.00 \cdot \text{pop edm:energy} - 30.34 \cdot \text{post-teen pop:energy} - 8.42 \cdot \text{progressive electro house:energy} \\
& - 27.97 \cdot \text{reggaeton:energy} - 6.25 \cdot \text{southern hip hop:energy} - 0.28 \cdot \text{trap:energy} \\
& - 14.80 \cdot \text{tropical:energy} - 20.12 \cdot \text{urban contemporary:energy} \\
& + 22.78 \cdot \text{big room:instrumentalness} + 10.01 \cdot \text{classic rock:instrumentalness} + 23.19 \cdot \text{dance pop:instrumentalness} \\
& + 19.26 \cdot \text{electro house:instrumentalness} + 11.54 \cdot \text{electropop:instrumentalness} + 28.47 \cdot \text{gangster rap:instrumentalness} \\
& + 1.001 \cdot \text{classic rock:instrumentalness} + 2.319 \cdot \text{dance pop:instrumentalness} \\
& + 1.926 \cdot \text{electro house:instrumentalness} + 1.154 \cdot \text{electropop:instrumentalness} \\
& + 2.847 \cdot \text{gangster rap:instrumentalness} + 2.344 \cdot \text{hard rock:instrumentalness} \\
& - 1.762 \cdot \text{hip hop:instrumentalness} + 1.346 \cdot \text{hip pop:instrumentalness} \\
& + 3.572 \cdot \text{indie poptimism:instrumentalness} - 1.062 \cdot \text{latin hip hop:instrumentalness} \\
& + 4.213 \cdot \text{latin pop:instrumentalness} + 3.547 \cdot \text{neo soul:instrumentalness} \\
& - 2.001 \cdot \text{new jack swing:instrumentalness} + 3.657 \cdot \text{permanent wave:instrumentalness} \\
& + 1.341 \cdot \text{pop edm:instrumentalness} + 2.447 \cdot \text{post-teen pop:instrumentalness} \\
& + 1.883 \cdot \text{progressive electro house:instrumentalness} + 2.567 \cdot \text{reggaeton:instrumentalness} \\
& - 1.762 \cdot \text{southern hip hop:instrumentalness} + 1.478 \cdot \text{trap:instrumentalness} \\
& + 2.011 \cdot \text{urban contemporary:instrumentalness} + 3.897 \cdot \text{danceability} \\
& - 4.005 \cdot \text{energy} + 1.241 \cdot \text{loudness} \\
& - 3.113 \cdot \text{liveness} - 2.765 \cdot \text{valence} + 0.01672 \cdot \text{tempo} \\
& - 0.00002712 \cdot \text{duration\_ms}
\end{aligned}
$$

3.1.3. Stepwise Model Selection

The stepwise regression model produced an adjusted R-squared value of 0.04654, indicating that approximately 4.7% of the variability in track popularity is explained by the predictors included in the model. Although this value suggests a relatively low explanatory power, the overall model is statistically significant, as evidenced by an F-statistic of 89.8 and a p-value of less than 2.2e-16. This indicates that, despite the low R-squared, the model as a whole provides meaningful insights into the relationships between the predictors and the response variable. During the stepwise selection process, the variables 'key' and 'valence' were excluded from the final model because their removal reduced the Akaike Information Criterion (AIC), thereby improving the model's efficiency. The lowest AIC value achieved by stepwise regression was 115779, reflecting the optimization between model complexity and fit. The final model retained key variables such as

danceability, energy, loudness, and tempo, among others, resulting in the following first-order regression equation:

$$\begin{aligned} Y_{\text{trackpopularity}} = 69.45 + 6.129 \cdot \text{danceability} - 21.40 \cdot \text{energy} + 1.054 \cdot \text{loudness} \\ + 0.724 \cdot \text{mode} - 6.994 \cdot \text{speechiness} + 3.640 \cdot \text{acousticness} \\ - 7.960 \cdot \text{instrumentalness} - 5.254 \cdot \text{liveness} + 0.03110 \cdot \text{tempo} \\ - 0.00005071 \cdot \text{duration\_ms} \end{aligned}$$

This equation (above) highlights the contribution of each predictor to track popularity, with attributes such as “danceability”, “loudness”, and “acousticness” positively influencing popularity, while the attributes of “energy”, “instrumentalness”, and “liveness” exhibit negative relationships.

3.1.4. All-Possible-Regressions Selection Procedure

Using the all possible regressions algorithm on the 12 quantitative predictors and the subgenre categorical predictor, the model with the highest adjusted-R^2 of 0.1521 was the best model with 32 predictors and removes the “key”, “mode”, and “reggaeton” predictors from the full regression model. However, if we select the model based on a different regression statistic, such as Mallow’s CP criterion, the best model is the model with 31 predictors and removes “speechiness” in addition to the predictors removed by the best model by adjusted-R^2. The results regarding these qualitative predictors concur with the results of the multiple linear regression and only “acousticness” was removed by our previous model that was not excluded in this all possible regression model. This model also removes the “reggaeton” indicator as a predictor, however, in the multiple linear regression, that specific predictor has a high p-value which may be the reason why the all possible regressions algorithm removed it in the best models. Logically it would change the interpretation of songs in this genre, as removing the “reggaeton” indicator variable would combine “reggaeton” songs with the “album rock” songs that are indicated by being 0 for all the subgenre indicator variables. Since this idea of combining subgenres would be a much more extensive process, combining “reggaeton” and “album rock” because “album rock” happens to be the category indicated by being 0 for all the subgenre indicator variables, it would be an arbitrary decision to combine these specific subgenres and therefore the “reggaeton” indicator variable will not be excluded from our final model.

$$
\begin{aligned}
Y_{trackpopularity} = 62.46 &+ 11.77 \cdot \text{danceability} - 17.94 \cdot \text{energy} + 0.862 \cdot \text{loudness} \\
&+ 1.488 \cdot \text{acousticness} - 6.750 \cdot \text{instrumentalness} - 3.491 \cdot \text{liveness} \\
&- 3.022 \cdot \text{valence} + 0.01693 \cdot \text{tempo} - 0.00002937 \cdot \text{duration\_ms} \\
&+ 14.39 \cdot \text{latin} + 27.06 \cdot \text{pop} + 19.27 \cdot \text{r\&b} + 20.17 \cdot \text{rap} \\
&+ 16.61 \cdot \text{rock} - 12.78 \cdot \text{big room} - 1.88 \cdot \text{classic rock} \\
&- 4.901 \cdot \text{electro house} - 2.250 \cdot \text{electropop} - 12.87 \cdot \text{gangster rap} \\
&- 6.02 \cdot \text{hard rock} + 9.009 \cdot \text{hip hop} + 3.041 \cdot \text{hip pop} \\
&- 4.572 \cdot \text{indie poptimism} + 3.343 \cdot \text{latin pop} - 12.36 \cdot \text{neo soul} \\
&- 16.47 \cdot \text{new jack swing} + 12.19 \cdot \text{permanent wave} - 7.286 \cdot \text{pop edm} \\
&+ 10.41 \cdot \text{post-teen pop} - 16.64 \cdot \text{progressive electro house} \\
&- .6866 \cdot \text{reggaeton} - 8.769 \cdot \text{southern hip hop} + 3.278 \cdot \text{trap} \\
&- 2.352 \cdot \text{tropical} + 2.392 \cdot \text{urban contemporary}
\end{aligned}
$$

For the results of the all possible regressions algorithm, we will select the model based on adjusted-R^2, since it is a common metric among the other procedures, which is the best 32 predictor regression model. This model removes fewer predictors from the model than the multiple linear regression process we conducted, however, two predictors in the all possible regression model, “speechiness” and “acousticness” are not significant and the 32 predictor model is not significantly stronger than the model produced by the multiple linear regression process based on adjusted R-squared value or residual standard error.

3.1.5. Final Model Comparison and Selection

Each of the four unique multiple regression model development techniques completed above will be judged by their Adjusted R-squared and Residual Standard Error (RSE) to identify our final selected model. An interaction model is more reliable in explaining the response variable because it accounts for the possibility that different genres may respond differently to certain predictors, such as “loudness” and “energy”. For example, in some music genres, high loudness might be strongly associated with popularity, while in others, the relationship could be weaker or even negative. Similarly, energy levels might influence the popularity of a track differently depending on the genre.

Table 2: Comparison of Regression Models Based on Adjusted R-squared and Residual Standard Error (RSE)

Model	**Adjusted R-squared**	**Residual Standard Error (RSE)**
Multiple Linear Regression Model	0.152 (15.2%)	22.67
Interaction Model	0.1639 (16.39%)	22.51
Stepwise Model	0.04654 (4.65%)	24.03
All-Possible-Regressions Model	0.1521 (15.21%)	22.67

It can be observed that the largest Adjusted R-squared value is 0.1639 (Table 2) and the smallest Residual Standard Error value is 22.51 (Table 2) among all of the models. Based on this, the Interaction Model has been identified as the optimal model developed.

The Adjusted R-Squared value of 0.1639 (Table 2) indicates that 16.39% of the variance in the "track_popularity" variable is explained by the model, the largest value of the developed models.

The Residual Standard Error value of 22.51 (Table 2) indicates that the actual observed values of the "track_popularity" variable deviate from the predicted values by an average of approximately 22.51 (Table 2), the smallest value of the developed models.

Table 3: Interpretation of Coefficients of the Final Regression Model

Coefficient	Interpretation
playlist_subgenre	Most of the significant linear subgenre predictors have positive coefficients, meaning that most of the subgenres have on average higher popularity than the "album rock" subgenre, the subgenre that is indicated when all other subgenre indicator variables are 0.
danceability	Danceability has a positive coefficient and therefore, positively affects the popularity of songs.
loudness	Loudness has a positive coefficient and therefore, positively affects the popularity of songs.
instrumentalness	Instrumentalness has a slightly negative coefficient and therefore, negatively affects the popularity of songs.
liveness	Liveness has a negative coefficient and therefore, negatively affects the popularity of songs.
valence	Valence has a negative coefficient and therefore, negatively affects the popularity of songs.
tempo	Tempo has a slightly positive coefficient and therefore, positively affects the popularity of songs.
duration_ms	Duration has a negative coefficient and therefore, negatively affects the popularity of songs.
playlist_subgenre : energy	Most of the significant interactive terms for energy and subgenre are negative, therefore, energy negatively impacts the popularity of songs of those subgenres.
playlist_subgenre :	Most of the significant interactive terms for instrumentalness and

instrumentalness	subgenre are positive, therefore, instrumentalness positively impacts the popularity of songs of those subgenres.
playlist_subgenre : duration_ms	Most of the significant interactive terms for duration and subgenre are negative, therefore, duration negatively impacts the popularity of songs of those subgenres.
loudness : instrumentalness	The coefficient for the interaction term of loudness and instrumentalness is negative and therefore, these two predictors interact to negatively affect the popularity of songs.
loudness : duration_ms	The coefficient for the interaction term of loudness and duration is slightly negative and therefore, these two predictors interact to negatively affect the popularity of songs.
danceability : liveness	The coefficient for the interaction term of danceability and liveness is positive and therefore, these two predictors interact to positively affect the popularity of songs.

3.1.6. Regression Assumption Evaluation on Final Model

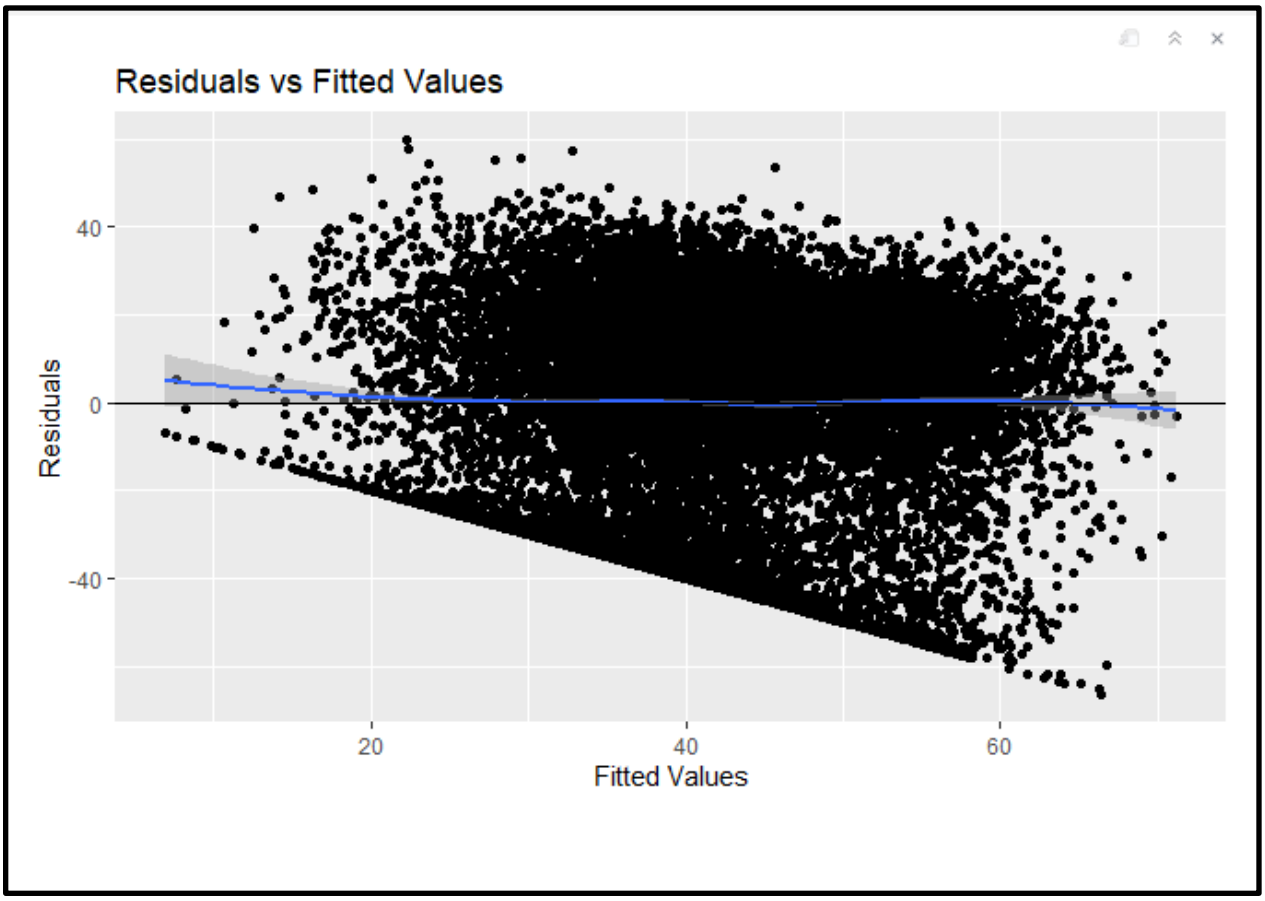


Figure 1: Residuals vs Fitted values plot for Linearity testing.

The Residuals vs. Fitted Values plot (Figure 1) is a diagnostic tool for assessing the linearity assumption in regression analysis, which requires that the relationship between predictors and the response variable be linear. In this plot, the residuals are primarily centered around zero, and the smooth line is nearly flat, indicating that the model captures the linear trend reasonably well. This suggests that the relationship between the fitted values and the residuals can be somewhat considered linear. However, there are slight deviations from randomness, as the residuals form subtle patterns and exhibit some curvature, particularly at the lower and higher ends of the fitted values. These patterns imply that the model may not fully capture non-linear aspects of the data. While the overall flatness of the smooth line supports the linearity assumption to a certain extent, the observed deviations suggest potential model refinement, such as incorporating non-linear terms or interactions, could improve the fit.

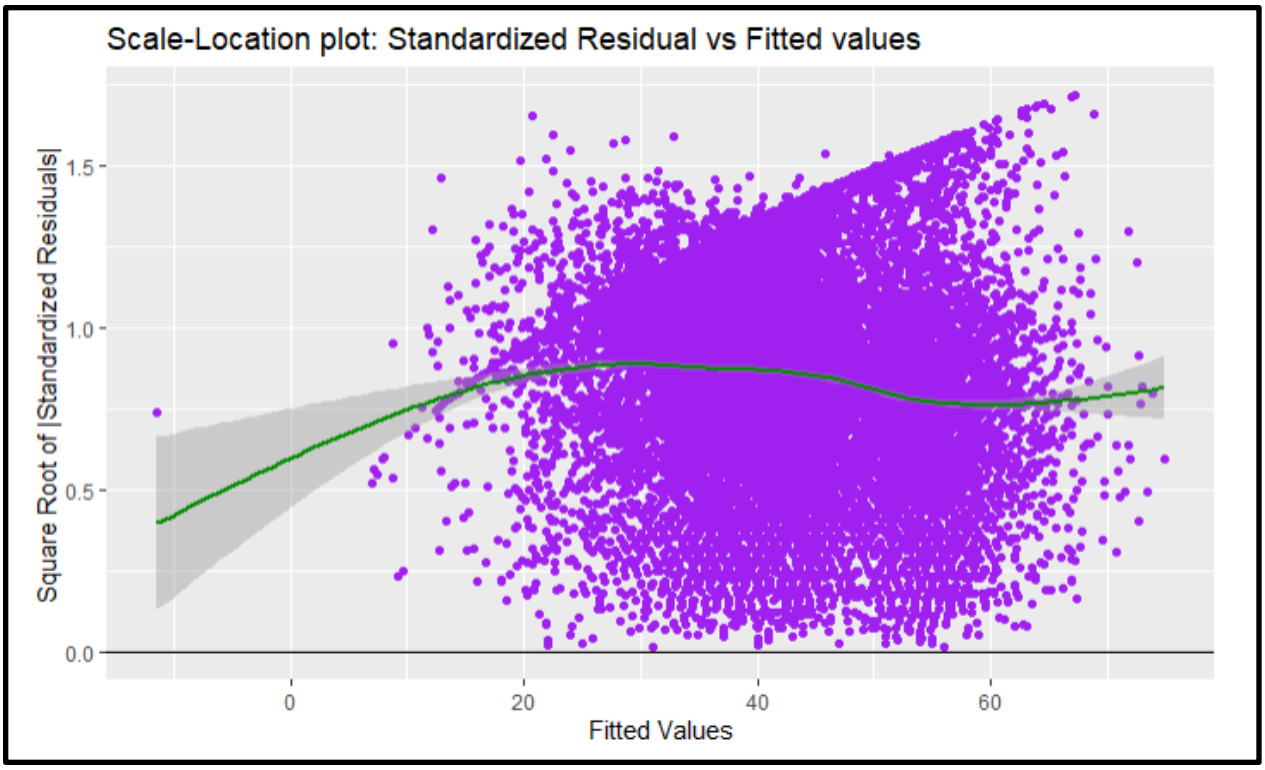


Figure 2: Scale-Location plot for equal variance testing.

The Scale-Location plot, which displays the square root of the standardized residuals against the fitted values, is used to assess the equal variance assumption in regression analysis (Figure 2). In this plot, the residuals should ideally be randomly scattered with a flat smooth line, indicating consistent variance across the range of fitted values. However, the plot shows a slight increase in the spread of residuals as fitted values grow, suggesting potential heteroscedasticity, where the variance of residuals is not constant. This violation of the equal variance assumption can affect the reliability of confidence intervals and predictions generated by the regression model. Additionally, the smooth line is not perfectly flat and shows subtle variations, reinforcing concerns about non-constant variance. There is also some clustering of residuals, particularly at lower fitted values, which warrants further investigation. To address these issues, transformations of the response variable, such as a logarithmic, could be considered.

The Breusch-Pagan test was also conducted to assess the presence of heteroscedasticity. The null hypothesis (H_0) assumes homoscedasticity (equal variance), while the alternative hypothesis (H_a) suggests heteroscedasticity.

H_0 : heteroscedasticity is not present (homoscedasticity).

H_a : heteroscedasticity is present.

The test produced a p-value < 2.2e-16, indicating strong evidence against the null hypothesis. These results confirm that heteroscedasticity is present in the dataset, reinforcing the findings from the Scale-Location plot and suggesting that transformations or weighted regression methods may be required to address this issue.

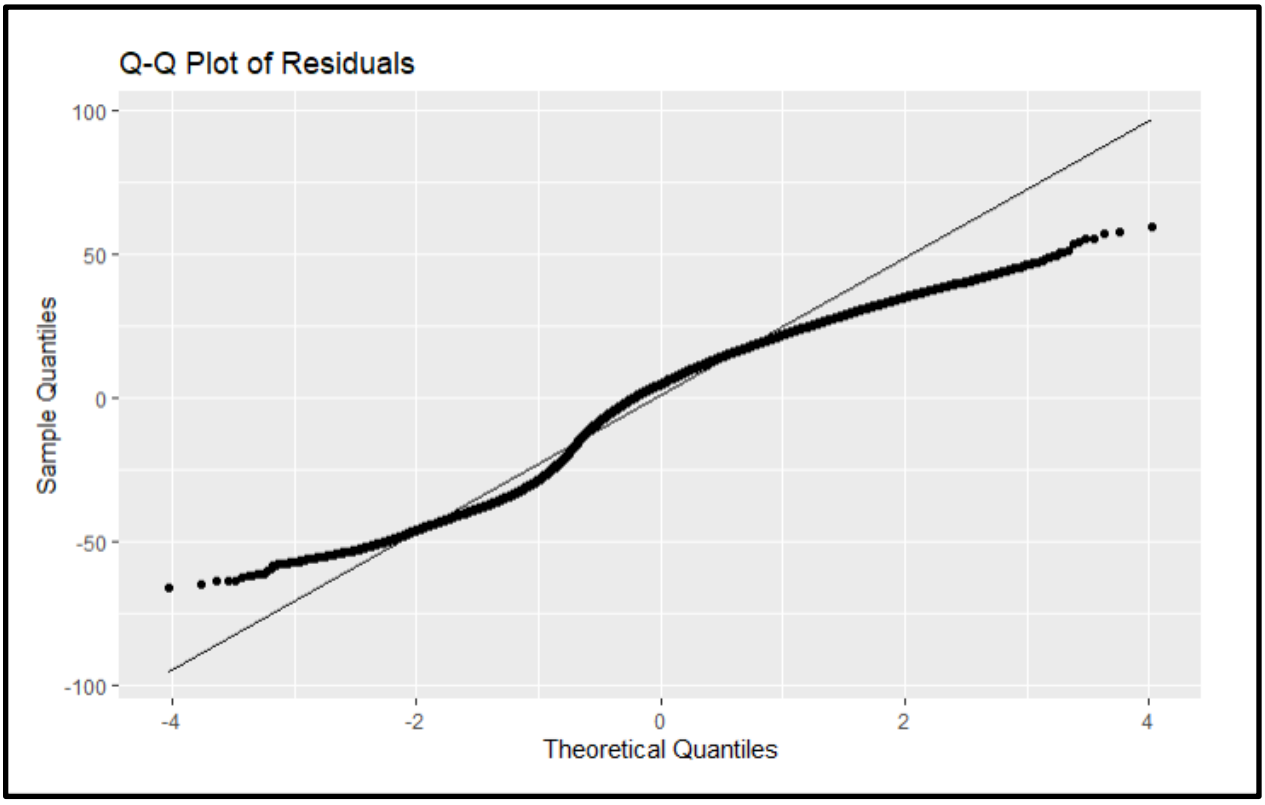


Figure 3: Q-Q plot for linearity testing.

The Q-Q plot assesses the normality of residuals, which should ideally align with the diagonal line if they follow a normal distribution (Figure 3). The residuals deviate systematically from the diagonal line, forming an S-shaped pattern. This indicates excessive kurtosis, suggesting that the data may have heavier or lighter tails than a normal distribution. Specifically, the deviation in the lower tail and the upper tail (rising above the line) reflects heavy tails and potentially the presence of outliers or extreme residuals in both directions. Kurtosis measures whether the data has heavy tails (high kurtosis) or light tails (low kurtosis) compared to a normal distribution. The observed high kurtosis in this dataset implies the presence of more extreme residuals than expected under normality, which can influence hypothesis testing and confidence intervals.

Despite this deviation, the central portion of the residuals aligns moderately well with the diagonal line, indicating that for a large portion of the data, the residuals approximate normality. However, the tails highlight areas where the model does not fully adhere to the normality assumption, which could impact the robustness of statistical inferences drawn from the model. Transformation

techniques might be required to address these deviations for better model performance and validity.

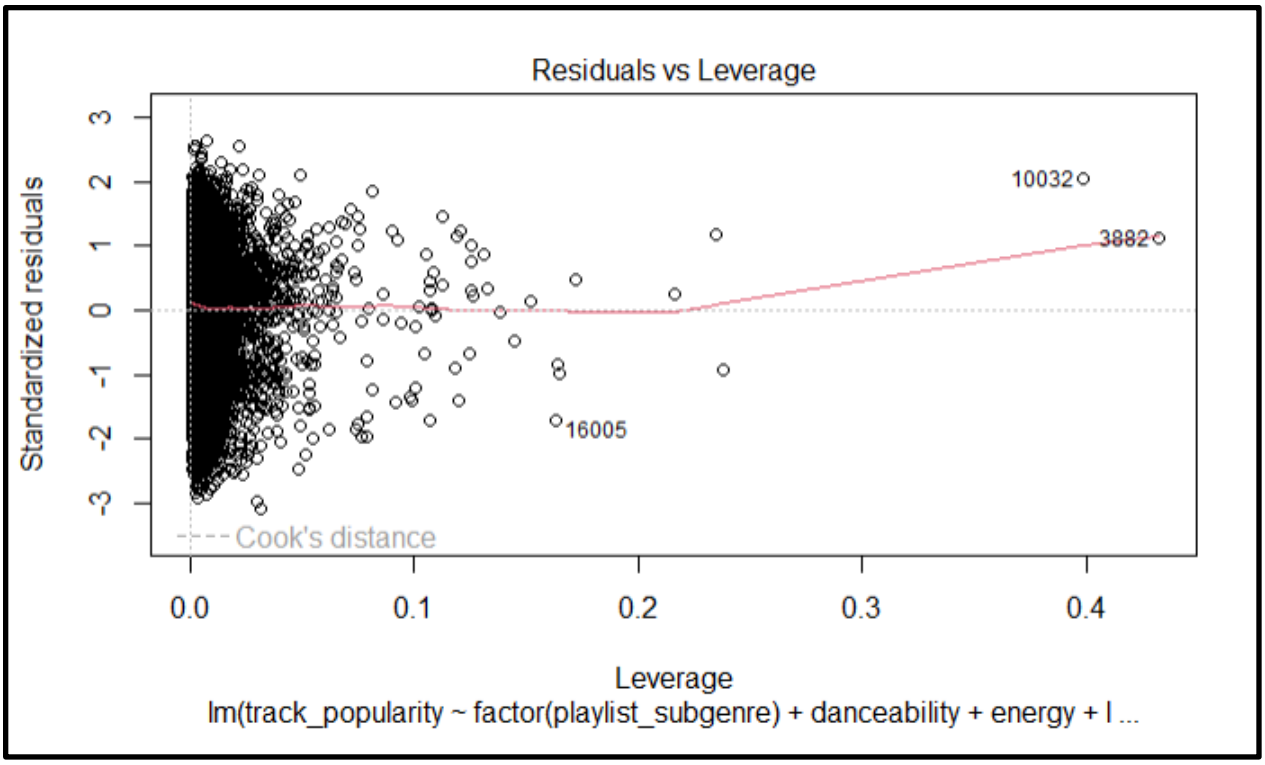


Figure 4: Residuals vs. Leverage plot for outlier leverage testing.

The Residuals vs. Leverage plot, used to identify influential data points, becomes particularly relevant when analyzing a dataset with over 18,000 data points, as in this case. The majority of points are clustered near the center, indicating low leverage and minimal influence on the regression mode (Figure 4). However, a small number of points, such as those labeled 3882 and 10032, show high leverage and fall near or beyond the Cook's distance threshold, suggesting they may have a disproportionate impact on the model. Given the size of the dataset, these few high-leverage points represent a very small fraction of the data and are unlikely to significantly skew the overall results. With few major leverage points in comparison to the dataset size, this can safely be considered a reliable dataset to use, with minimal effects from the present outliers.

Outlier analysis was conducted, and high-leverage points identified through the outlier analysis were removed to assess their impact on the regression model. However, this did not result in any noticeable changes to the diagnostic plots, including the Residuals vs. Fitted Values, Scale-Location, and Q-Q plots. This suggests that the underlying issues within the assumptions, such as heteroscedasticity and deviations from normality, are not solely attributable to outliers and require further model refinements. Detailed outlier analysis can be found in the appendix under the *"Rerun of tests on outlier adjusted data"* section.

3.1.7. Dependent Variable Logit Transformation Model Identification

The logarithmic transformation of popularity did not improve the model's fit, with the R square decreasing to **8.2%**, indicating a poor overall variance explanation. However, the predictors' statistical significance improved, with most variables having p-values less than $\alpha = 0.05$, following variable selection from the previous model. This outcome might indicate a mismatch between the predictors and the transformed scale or potential overfitting. Therefore, we chose to retain our interaction multiple linear regression model as the final model, as it provides a better explanation of the variance in the response variable, popularity (R square = 16.4).

4. CONCLUSION AND DISCUSSION

4.1. APPROACH

Our approach effectively addresses the primary research question by identifying key variables that strongly influence the dependent variable and quantifying their impact. We developed a model that selects all significant predictors and demonstrates their effects through coefficients. However, the model has limitations. It exhibits a weak fit, with an R square value of only **16.4%**, and fails to meet several assumptions.

The diagnostic plots provide a concise assessment of the regression model's adherence to key assumptions, revealing areas of strength and potential refinement. The linearity assumption is generally satisfied, as the Residuals vs. Fitted Values plot shows a near-flat smooth line and residuals mostly centered around zero. However, slight curvature and clustering at the extremes indicate minor non-linear patterns, suggesting potential improvement by adding non-linear terms or interactions. The influence diagnostics (Residuals vs. Leverage plot) confirm that the dataset is reliable, with few high-leverage points relative to its size of over 18,000 observations. These influential points are unlikely to significantly skew results, confirming the dataset's suitability for analysis. Overall, this approach is promising as it effectively identifies key areas for refinement without compromising the dataset's integrity.

Despite these challenges, our current approach provides valuable insights into how musical attributes influence a song's popularity, particularly across different subgenres. For example, the effect of song duration varies by subgenre: in hard rock, longer songs tend to lose popularity (β_hard rock:duration_ms=-3.76×10^5) whereas New Jack Swing fans prefer longer songs (β_New Jack Swing:duration_ms=3.06×10^5). Additionally, energy levels demonstrate distinct patterns. Hard rock is the only subgenre positively associated with energy, with higher energy slightly increasing popularity (β_hard rock:energy=4.489). In

contrast, all other subgenres exhibit a negative interaction between energy and popularity, indicating that rock fans prefer higher energy in their music.

We can transform our response variable from a numeric scale to a binary (dummy) variable by setting a threshold as a different approach. For example, we can classify a track as "popular" if it falls within the top 25% and as "not popular" otherwise. This transformation allows us to apply logistic regression and compare its results with the multiple linear regression model to assess whether the findings differ between the two approaches.

Understanding these nuanced relationships—whether positive or negative—offers musicians actionable advice for tailoring their creations to specific audiences, ultimately helping them reach a broader market and expand their fan base.

4.2. FUTURE WORK

In terms of the validity of the final model, the equal variance assumption was found to be partially violated, with the Scale-Location plot showing increasing spread of residuals. The Breusch-Pagan test confirmed heteroscedasticity with a p-value < 2.2e-16. Additionally, the normality assumption was moderately supported, though the Q-Q plot revealed deviations in the tails, indicating excessive kurtosis.While we have gained important preliminary insights into the factors contributing to track popularity, it is clear several refinements to our methodology could enhance future work in this area. For instance, incorporating polynomial features or splines may better capture the nonlinear relationships observed in the dataset. Additionally, addressing outliers through either removal or transformation, could lead to more robust results. Considering alternative regression models, such as logistic regression, might also be worthwhile, since popularity measured as a percentile aligns well with the assumptions of logistic regression.

If we interpret the findings of this study as an initial exploration of song popularity, we can see that there are several opportunities to expand the scope of our analysis. One significant avenue involves extending the dataset to include external factors beyond the intrinsic characteristics of a song. For example, how the song sounds, data on the marketing budget, the popularity of the artist's other songs, or social media engagement metrics could also offer valuable insights, as these external factors undoubtedly influence a song's popularity.

A notable finding from our regression models was that a song’s subgenre has a significant impact on increasing the accuracy of our models and additionally, had significant interaction within our model. An interpretation of this finding is that the factors driving popularity may vary significantly across subgenres. In future iterations of this work, we could consider separating our data into different genres or subgenres as this approach would allow for tailored predictions that align with the unique expectations of listeners within each category.

5. References

1. Hocking College. (n.d.). *The pros and cons of streaming music*. Retrieved December 1, 2024, from https://blog.hocking.edu/the-pros-and-cons-of-streaming-music
2. Temple University. (2024, September 6). How do streaming services impact song popularity? *PUSHAT*. https://sites.temple.edu/pushat/2024/09/06/how-do-streaming-services-impact-song-popularity/
3. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-Hill Education.
4. Aguiar, L., & Waldfogel, J. (2018). Platforms, promotion, and product discovery: Evidence from Spotify playlists. *Journal of Industrial Economics, 66*(4), 735–771.
5. Audio features and lyrics of Spotify songs. (2020, June 14). Kaggle. https://www.kaggle.com/datasets/imuhammad/audio-features-and-lyrics-of-spotify-songs
6. Home | Spotify for Developers. (n.d.). https://developer.spotify.com/
7. Comprehensive R Archive Network (CRAN). (n.d.). CRAN: Package latexpdf. https://cran.r-project.org/web/packages/latexpdf/index.html
8. R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/
9. Xie, Y. (2021). knitr: A general-purpose package for dynamic report generation in R. https://yihui.org/knitr/