STA302 Final Report

Group 20

Contributions

Our group consisted of four members Danish Ahmed, Shamayla Durrin Islam, Rufaida Khan, and Ari Aynedjian (Group 20). In part 1 of the project, Danish covered the literature review and summary statistics, Shamayla contributed by cleaning the dataset while building and analysing the preliminary model results. Rufaida covered the research question and data description, and Ari covered the residual analysis and assumptions tests. In part 2 of the project, all group members contributed to the flowchart analysis, by reviewing the modules to develop a rough outline of the final project. Everyone was part of the three meetings to make sure that the flowchart is in the right direction and order. Lastly, in part 3, Danish and Shamayla were primarily responsible for generating diagnostic plots to check for model assumptions, implementing transformations to correct for those assumptions, and developing a reduced model through partial F-tests including the R coding. Rufaida and Ari then covered the analysis required for the results and discussion sections of the report, they were primarily responsible for the writing the report. Everyone worked well and equivalently depending on their respective skill sets.

Introduction

In the era of music streaming, the factors that contribute to the popularity of certain tracks have become a topic of interest in recent times. The central question guiding this analysis is 'Do variations in audio attributes such as loudness, valence etc. have substantial influence on the popularity of songs on Spotify?'. In this report we uncover the relationships between the popularity of songs on Spotify and audio attributes that a particular track has. Spotify, being one of the most widely used platforms in the world today with over 551 million monthly active users, led us to choose the dataset related to the platform. The predictors of interest in our analysis include both continuous and categorical variables. Continuous predictors, such as danceability, energy, tempo, and loudness, offer insights into the quantitative aspects of a track's composition. On the other hand, categorical predictors, including time signature and mode, contribute to our understanding of the qualitative musical attributes that might impact popularity. It should also be noted that 'musical key,' though treated as a categorical variable, is transformed into a dummy variable so that we can accurately grasp the nuances that are associated with different musical keys. In our study we used a linear regression model to perform the data analysis because it allows us to express the relationship between the chosen predictor variables and popularity. The coefficients of the respective predictor variables will provide us with an intuition of their impact on the popularity of the track. In the quest of understanding the factors that influence track popularity on Spotify, it is crucial to first examine and understand existing knowledge pertaining to the research question. The article "Music Popularity: Metrics, Characteristics, and Audio-Based Prediction" aims to measure the popularity of a track through sales and the number of streams. However, there is a gap in knowledge on how certain aspects of the track itself determine popularity. The article "What Makes a Music Track Popular in Online Social Networks?" correlates popularity with music content, the artist's reputation, and the social

context of the track. Like the first article, the authors used predictor variables that measure external factors rather than the musical aspect of the tracks. The article "A Model for Predicting Pop Music Popularity and Its Different Characteristics Based on Multiple Linear Regression" measured the popularity of pop music on YouTube through streams and musical-related predictors that we plan to use. However, the gap in knowledge is that the article does not cover many genres of music. Also, the article uses YouTube as the music platform instead of Spotify. This can impact results because the demographics of users differ between platforms, and some songs might be exclusive to a certain platform. These peer-reviewed articles provide us existing knowledge and helps us identify gaps, to provide a more intricate understanding of the research question. Nevertheless, most articles aimed to predict the popularity metrics of the study. We aim to restrict ourselves into finding the most appropriate and best model to find the associations between popularity and other covariates.

Methods

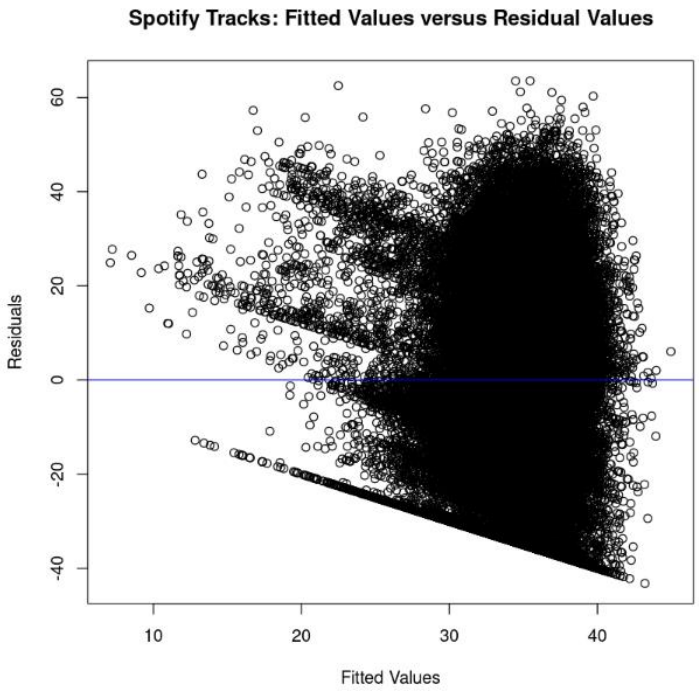
To pursue our analysis of the research question, we are applying multivariable linear regression concepts. Firstly, we ran diagnostic tests on our full model, which included all the predictor variables, such as danceability, energy, tempo, time signature, valence, 'acousticness', loudness, and mode (major or minor). The primary step was to check if the regression assumption holds or not. With the aid of the residual plots, such as the residuals vs fitted values plot, normal Q-Q (Quantile-Quantile) plot and response vs fitted values plot we were able to identify the short comings our data had. Alongside additional histograms, and scatterplot matrix, we determined whether the model assumptions such as linearity and normality, homoscedasticity, and independence of the error terms, hold. After careful observation of the plots, appropriate transformations were implemented to rectify them, specifically the Yeo-Johnson transformation (an extension of box cox function in R). We utilized the Yeo-Johnson since it allowed us to implement box-cox method with the negative values in our data. Following the transformation, the diagnostic tests were once again run on the new model to check the validity and effect of the transformations. Furthermore, outliers were removed from the data to improve cohesion with the model assumptions and skewness issues achieving normality to an extent, which were then verified by diagnostic plots again. Next, backward model selection was conducted using the MASS package in R, in a bid to find a refined model. Moreover, individual t-tests and analysis of p-values were conducted to determine which predictors to exclude, providing us with potential reduced model for partial F-tests, which were then conducted. The null hypothesis of the partial F-test states that the reduced model is sufficient, while the alternative hypothesis states that the reduced model is not sufficient. We reached our final model once the partial F-test failed to reject the null hypothesis at the 1% significance level.

Results

Since, we aimed to answer our research question based on a linear regression model, we needed to make sure that the regression assumptions including linearity, normality, and homoscedasticity hold. We performed a regression with our preliminary model and observed that the linearity assumption holds largely since in the fitted values vs residual values the horizontal line approximately pass through zero as seen Plot 1. Moreover, with the same observations we

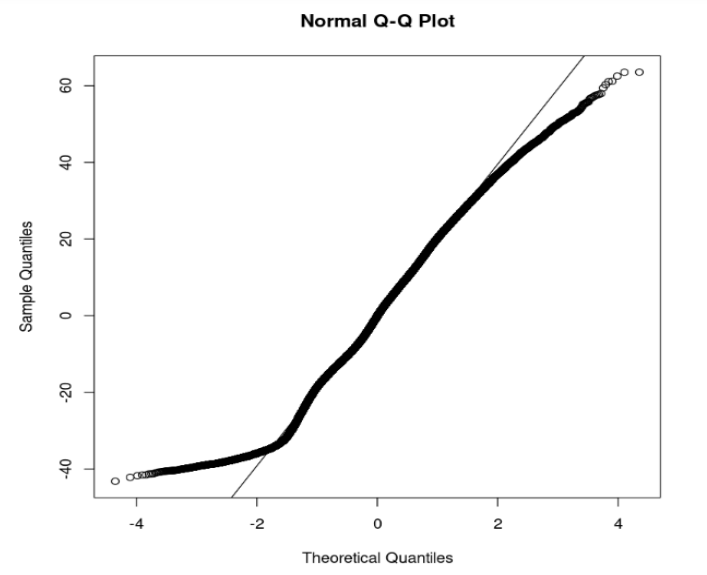
perform the check if the homoscedasticity assumption holds. The lower part of the graph in Plot 1 shows that there must be some outliers which led to the random data points pointing in a downward direction. Nevertheless, the homoscedasticity assumption did not hold.

Plot 1



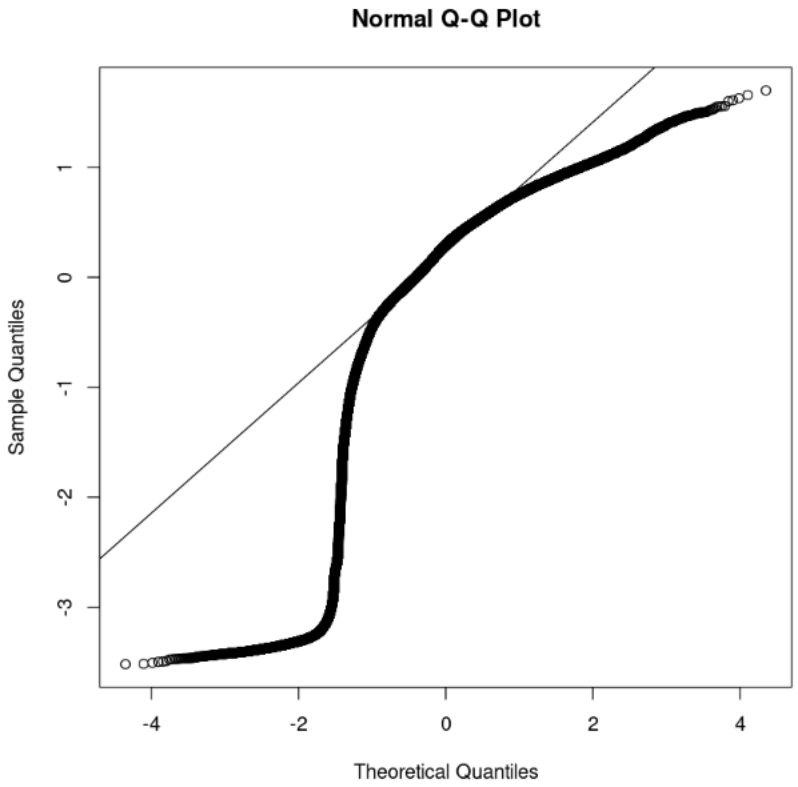
For the final normality assumption we plot a Quantile-Quantile plot (see plot 2). With the lower tail heavily above the line and the upper tail heavily below the line our normality assumption also fails.

Plot 2



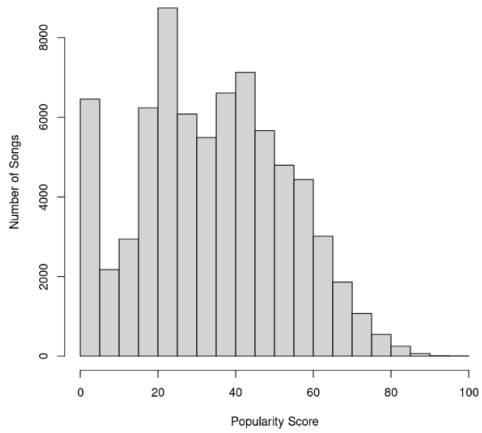
To correct for the regression assumptions, we decided to perform a suitable transformation to ensure that accurate and reliable relationships among the variables can be established. Our first resort was to apply a logarithmic transformation. Unfortunately, it was not possible since our *loudness* variable contained negative values. We then moved towards trying a box-cox transformation but that was also not possible for the same reason. After exploring different types of transformations, we came up with an extension of the box-cox transformation called the Yeo-Johnson transformation which corrected transformation for negative data points in the data. Once the transformation was applied, we ran diagnostic plots to check if the transformation corrected for our regression assumptions. Regrettably, the transformation method failed and made our plots much worse (refer to Plot 3).

Plot 3

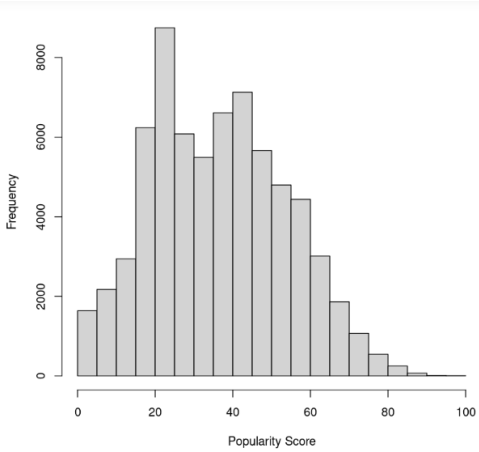


At this stage we had to make the decision of removing outliers for our popularity score variable. The idea of resorting to this step was that since our research question aimed to find the associative relationship between popularity score and other audio attributes. Our data contained a lot of observations with songs with a popularity score of zero. This would only hinder our research since songs which such a popularity score would not really give any useful information. Below is a difference between the distribution of popularity with and without the outlier observations.

Popularity Pre-Outlier Removal



Popularity Post-Outlier Removal



Removing outliers from our cleaned data set allowed us to perform the regression since now the regression assumption largely holds. The linear regression results are shown below.

Regression Table 1

```
lm(formula = popularity ~ danceability + energy + tempo + time_signature +
    valence + acousticness + loudness + major, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-43.792 -13.271  -0.539  12.154  61.768

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     50.455874   0.828168  60.925  < 2e-16 ***
danceability     5.390440   0.447736  12.039  < 2e-16 ***
energy         -16.332157   0.480598 -33.983  < 2e-16 ***
tempo           -0.009551   0.002272  -4.204 2.62e-05 ***
time_signature   0.701279   0.149912   4.678 2.90e-06 ***
valence         -1.855842   0.300196  -6.182 6.36e-10 ***
acousticness    -1.205731   0.288800  -4.175 2.98e-05 ***
loudness         0.779412   0.019943  39.081  < 2e-16 ***
major            0.362150   0.137287   2.638  0.00834 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.19 on 68781 degrees of freedom
Multiple R-squared:  0.03286,   Adjusted R-squared:  0.03275
F-statistic: 292.1 on 8 and 68781 DF,  p-value: < 2.2e-16
```

Our regression equation is.

$$y = \beta_0 + \beta_1 * danceability + \beta_2 * energy + \beta_3 * tempo + \beta_4 * time\_signature + \beta_5 * valence + \beta_6 * acousticness + \beta_7 * loudness + \beta_8 * major + \varepsilon$$

Note that the p-values are less than the significance level of 5% are statistically significant for all the variables and that all variables have a significant relationship with the popularity variable based off the T-tests values.

Based on the regression results, we run F-Tests and partial F-tests. After some trial and error we decided to choose the variables with the highest p-values from our regression results and remove them from our reduced model namely, tempo, acousticness, major and valence. We run an overall F-test and observed an F-statistic in our ANOVA table of 3.43 and a p-value of less than the significance value which means that reject the null-hypothesis and conclude that our model provides a better fit than the intercept-only model.

When conducting the partial F-test, after several trial-and-error models, we found a reduced model which did better than our primary full model. This reduced model includes an interaction term as well between *loudness* and *tempo*. At the 1% significance level the the p-value was 0.03 indicating reduced model is sufficient when considering the residual sum of squares.

Finally, we conducted an AIC test for model selection where we used the backward selection method in the *stepAIC* function in R which got us a very large AIC score 39,1430. The called model was very similar to our model but without the *major* variable.

Discussions

It is integral to verify the model assumptions to ensure validity of the proceeding linear regression model since interpretation of the regression coefficients rely on the linearity assumption. Violated assumptions can also lead to bias in coefficient estimates and reduced efficiency of the estimates resulting in wider confidence intervals. Moreover, hypothesis tests such as t-tests and F-tests assume those model assumptions to hold, such tests become invalid when they are violated. Although the variable *key* was a potential candidate as a categorical variable, we unfortunately had to drop the variable due to computing power issues. To cater to that, we decided to move with the *mode* as the second-best categorical predictor.

When looking over the regression results, the R-squared value of 0.03286 of our models shows that our model is slightly weak, and the goodness of fit could be potentially improved. This maybe possible if we tended to remove outlier observations of our independent variable. We avoided this since, it could potentially lead to losing additional information about the dataset overall. Moreover, the p-value of the model is also quite low and less than the significance level which allows us to reject the null hypothesis that there is evidence to believe that the corresponding predictor variable has a statistically significant effect on the dependent variable. AIC tests also gave us a better understanding of more possible models which one can look over.

(word count: 1723)

Bibliography

Lee, J., & Lee, J.-S. (2018). Music Popularity: Metrics, Characteristics, and Audio-Based Prediction. IEEE. https://ieeexplore.ieee.org/document/8327835

Ren , J., Shen, J., & Kauffman, R. J. (2016, April 1). What Makes a Music Track Popular in Online Social Networks?. ACM Digital Library. https://dl.acm.org/doi/abs/10.1145/2872518.2889402

Guo, B. (2021). A Model for Predicting Pop Music Popularity and Its Different Characteristics Based on Multiple Linear Regression. Academic Journal of Computing & Information Science. https://francis-press.com/uploads/papers/Spr4ld4vt2GMM6wzwLG04ePKKYuUImZJHC5m8Tjo.pdf

Appendix

Correlation Matrix of Spotify Dataset