Explaining the Popularity of Songs by Audio Features

Introduction

It is of interest to unravel the deeper logic of music, as this could hypothetically explain the great sense of resonance one experiences when listening to some music. Indeed, there have already been many related pieces of literature on audio features, but most of them do not incorporate either a sufficient amount of or more plausible features to be studied (Boyle et al., 1981; North & Hargreaves, 1996; Delsing et al., 2008). This, in turn, motivates a more tenable explanation to address this matter. Therefore, in this study, I attempt to employ statistical analysis to provide a more plausible account of what audio attributes contribute to the popularity of music.

Methods

Variable Selection

Among different pieces of literature, certain audio features are suggested to the preference for music. One study found danceability, acousticness, and tempo of the music are significant in different genders. Another study revealed that audio valence is important in different situations. Also, a study suggests the significance of energy and loudness in different personality types of people. Given the reliability of these studies, the statistical model will contain those features as variables to predict the popularity of the music.

To get the most informative and accurate dataset that facilitates the application of the statistical tools, at the very beginning, I cleaned the dataset by removing some insensible data according to EDA and the related knowledge. Then, after splitting the dataset into training and testing datasets for model validation, I constructed a rough model with song_popularity as the response and the rest as the predictors, and had this model checked against the conditions and assumptions to verify the suggestiveness of the model for later statistical tools applied (more detail later).

Since the first model failed to satisfy the assumptions, I transformed several variables of the model that violate the assumptions through the Box-Cox transformation in the hope to mitigate the violations. With a transformed model at hand, I went through the conditions and assumptions checking another time to verify its informativeness.

As this model mostly satisfied the assumptions, I later wanted to see the effect of some observations on the model and checked the leverage points, influential points, and outliers of the transformed model by quantifying the amount of influence each observation has. This step helped later model comparison. Since these observations in my model are all legit, I will discuss them more in the discussion section.

Later, I examined the predictors of the model. To start, I wanted to see a general picture of all the predictors in relation to the response. Firstly, I wanted to see if my predictors consider the conditional nature of regression and the possibility that multiple predictors are correlated to one another by detecting possible multicollinearity in my model by checking the variance inflation factor. In my case, my models, either a full or later reduced one, satisfy. Secondly, I

applied an F-test on the model by employing the ANOVA table and summary table to detect if there is an overall significant linear association between the predictors and the response by checking the p-value. Upon this, I also wanted to detect the significance of the individual predictor to the response, so I employed a t-test to see if there is a significant association by checking the p-value. Here, I created a reduced model which later went through all the steps mentioned from checking the conditions to verify if it satisfies the assumptions.

With these two models at hand, I finally wanted to select the most appropriate one. To achieve this, I checked the partial F-test against the ANOVA table and compared them by different indicators, Rsq_adj, AIC, AIC_c, BIC. If I have the one with higher Rsq_adj, smaller AIC, AIC_c, and BIC, it might be the most preferred one among the choices.

Model Validation

After I had a cleaned dataset, I created two independent datasets, a training and a testing dataset by 50/50 split from the randomly sampling 18,554 observations. The training dataset is used for constructing the model, whereas the testing one is used for validating the model.

After having my final model, I made a comparison of the characteristics between the model that fitted the training dataset and that fitted the testing dataset to check whether the model might be overfitted to the training dataset or incapable enough to capture the actual result. This comparison would be made by comparing the predictors, VIF information, number of influential points by Cooks and DFFITS, summaries of the coefficients, and whether assumptions hold. The success of validating the model will be determined by the difference and the failure of validating would be discussed in the discussion section.

Model Violations and Diagnostics

After dataset cleaning, I wanted to check the model against the two additional conditions with a pairwise scatterplot and a y vs. fitted y plot first, which suggested if the first or the second assumption from the later constructed residual plot respectively is being violated. If the violation has been suggested, I would not be able to have reasonable and convincing interpretations from the residual plots and I would make a note of it in the discussion section.

Later, I constructed a residual plot to check the model against the three assumptions and constructed a Q-Q plot to check the last assumption. The first assumption is violated when there seems to have no linear trend on the plot between the response and the predictors; the second assumption is violated when there seems to be an obvious cluster of data on the plot; the third assumption is violated when the residuals do not have constant variance (i.e. there is a funnel-like pattern detected); the fourth assumption is violated when the datapoints do not fall along the straight line in the Q-Q plot. In my case, since there did not seem any violation against the second assumption, I applied only the Box-Cox transformation to adjust the models.

After the assumption check, I proceeded to identify the leverage points, influential points, and outliers of the model by quantifying the amount of influence each observation has. To clarify, p is the number of predictors minus one and n is the number of observations. To figure out the leverage point, I checked if the observations' h_ii is greater than 2*(p+1)/n. To figure out the outliers, I checked if the absolute value of the observations' r_i is greater than 4, since our

sample size was big. To figure out the influential points, I applied three methods. I checked if the observations' D_i is greater than F_(0.5)(p+1)(n-p-1), if the absolute value of the observations' DFFITS_i is greater than 2*((p+1)/n)^(1/2), and if the absolute value of the observations' DFBETAS_i is greater than 2/((n)^(1/2)).

The observations identified were all legit in my case, and I made a note in the discussion section. Otherwise, I would remove the illegit observations and refit the model.

After identifying the influential points, I detected possible multicollinearity in my model by checking the variance inflation factor of the predictors. If all the variance inflation factors are insignificant, less than 5.0, which is in my case, then there is nothing to worry about. Otherwise, I might consider removing the associated variables and refitting a model.

Result

Description of Data

The dataset contains the response song_popularity as the response in the model and the rest of the variables, danceability, acousticness, tempo, audio_valence, loudness, and energy as the predictors in the model. The histogram of each variable is shown (Figure 1).

While most of the data distributed ranged from 0.00 to 1.00, the distribution of tempo has a range from 50 to 200, the distribution of song_popularity has a range from 0 to 100, and the distribution of loudness has a range from -40 to 0.

Specifically, for the graph of danceability, the data is slightly left-skewed with its mode placed between 0.65 to 0.75, which resembles the shape of the graph of song_popularity which is also slightly left-skewed yet has a mode between 65 to 70. This left-skewed distribution also applies to the graph of energy, which has a longer tail than the graph of danceability and with the mode placed between 0.65 to 0.90. The graph of loudness shows a much heavy left-skewed distribution with the majority of data distributed between -20 to 0. The opposite skewed distribution, left-skewed, applies to the distribution of acousticness, where it has the majority of data distributed between 0.00 to 0.25. The distribution of the data on tempo seems to have two modes, one between 120 and 130 and another between 90 and 100 and it seems to be right-skewed. Lastly, the distribution of audio_valence has a slight left skew and is evenly distributed from 0.80 to 1.00.

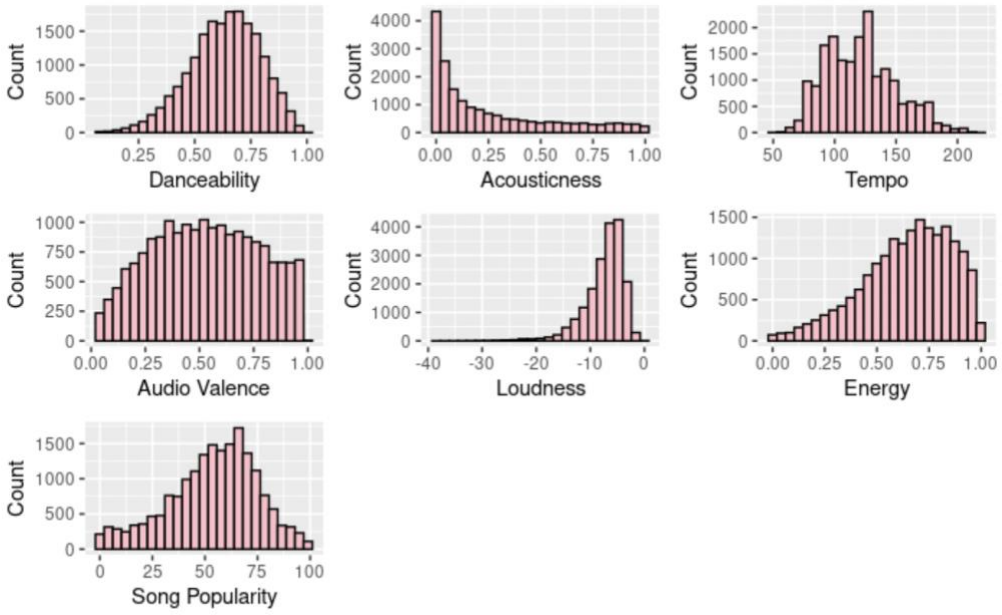


Figure 1: Histograms of the individual variable.

Presenting the Analysis Process and the Results

After cleaning the dataset according to EDA and related knowledge, I split the dataset into two groups, one for training and another for testing in a 50/50 manner. I first employed the training dataset to construct a rough model, with song_popularity as a response and the rest as the predictors.

By checking against the y and fitted y plot (Figure 2), I was comfortable that condition one is satisfied given that the points are randomly scattered around the line, whereas, by checking the pairwise scatter plot (Figure 2), I detected that the variable loudness formed a curved relationship with the predictor acousticness and energy. Thus, the second condition failed, and I would make a note in the discussion section.

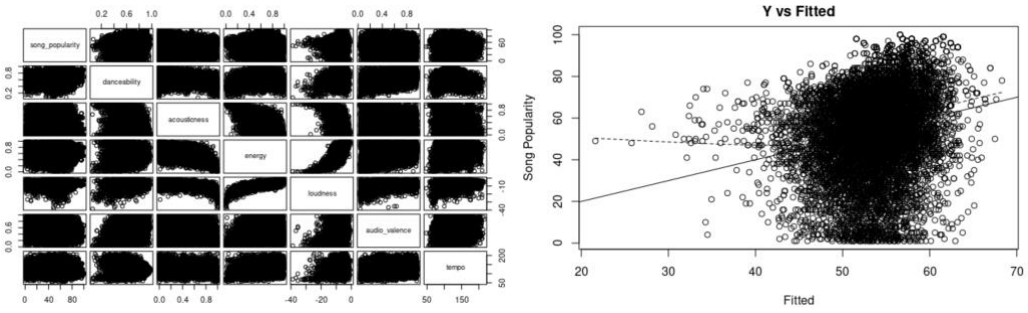


Figure 2: The pairwise scatter plot (left) and the
Y and fitted Y plot (right) for the initial model

By checking against the residual plot (Figure 3), I noticed that assumption one was decent, where there was a negative linear relationship between the residuals and the fitted values, that assumption two was valid, given there was no discernible pattern shown in the plot, that assumption three seemed invalid, as there was a funnel-like pattern detected, that assumption four seemed decent, as the datapoints were put closely along the straight line in the Q-Q plot with a slight negative skew. In addition, the residual plots of the predictor energy, loudness, and danceability showed discernable inconstant variance which violated assumption three. In this case, I performed a Box-Cox transformation on the response and the predictors, energy, loudness, and danceability, and I had a transformed model.

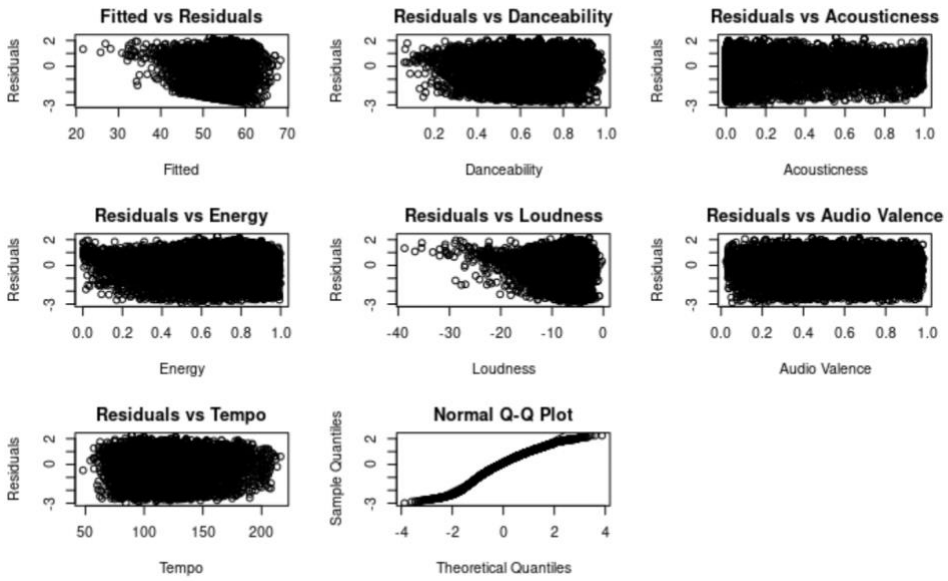


Figure 3: The residual plots for each variable and the Q-Q plot at the end of the initial model.

The transformed model was then checked against the additional conditions through the pairwise scatter plot and y vs fitted y plot (Appendix A). It was found that the curved relationship still occurred between the predictors tenergy and tloudness, but condition one seemed to be more comfortably acceptable. In addition, the transformed model was checked against the residual plot (Appendix A). It turned out that assumption three still got slightly violated, while other assumptions seem more comfortably acceptable. In this case, I adopted the transformed model since it was preferable to the assumptions.

After that, any outliers, leverage points and influential points were to be identified in the transformed model. Here, by applying the method as proposed in the method section, I obtained 389 leverage points, 414 influential points according to DFFIS and 4040 influential points by DFBETAS.

Afterwards, I turned to detect the possible multicollinearity of this model. Since the values of VIF for all of the predictors were less than the cut-off, 5.0, multicollinearity did not seem to appear in our predictors.

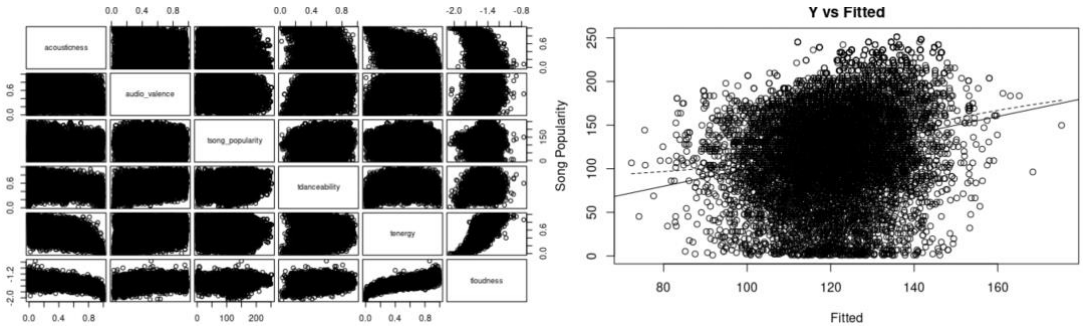
Proceeding forward, I applied F-test to the dataset to check the overall significance of the predictors to the response. Given the p-value was 2.2e-16, I failed to reject that all predictors have no relationship to the response. Next, I turned to check the t-test for the individual predictor if it forms a significant relationship to the response. It turned out that the predictor tempo was the only predictor which had a moderate relationship to the response, which was 0.134, unlike other predictors with which the p-values were close to 0. Therefore, I decided to remove the predictor, tempo, to refit a new model (the discussion of the checking new model will be extended in the next section).

Therefore, I conducted a partial F-test with the third and the second model to check if the variable tempo is statistically significant for the model. Since the p-value was 0.1336, we had little evidence that the additional predictor tempo is statistically significant to the response. In this case, I suspected that the reduced model, the third model, should be a more reasonable model.

In addition, the indicators were checked against the models. Given the smaller AIC, AIC_c, BIC, and the bigger Rsg_adj, the better the model would be, the full model seemed to be a better model. However, the difference is too small, and the partial F-test suggested the reduced model. Therefore, I would believe the reduced model was a better model to account for the relationship.

Goodness of the Final Model

As suggested in the previous section, I had a final or reduced model. After checking against the additional conditions and the assumptions (Figure 4), where they were quite similar to the full model, this reduced model was then transformed as what has been done for the first model, with its danceability, energy, loudness, song_popularity transformed. Then, I turned to identify the influential points of the model. It showed that there were 455 leverage points, 458 influential points by DFFIS and 3430 influential points by DFBETAS. After this, there was no multicollinearity occurred in this model by checking against VIF.



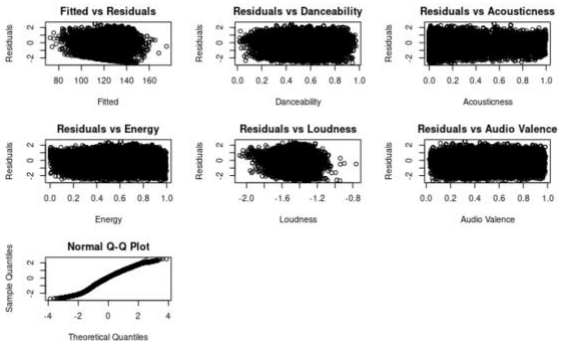


Figure 4: The pairwise scatter plot (top left) and the
Y and fitted Y plot (top right) for the final model. The residual plots
for each variable (bottom) and the Q-Q plot at the end.

After checking against the conditions and diagnosing the model, I attempted to validate the model by fitting it with the testing dataset set at the beginning. Here, I went through the same process of building the model as the final model starting all the way from checking the additional conditions. After fitting the model with the testing dataset, I compared the significant values between the model constructed by the training dataset and by the testing dataset shown in the table (Table 1).

Characteristics	Model (Train)	Model (Test)
Largest VIF value	3.464	3.459
Rsq_adj	0.052	0.045
# Cook's D	0	0
# DIFFITS	414	458
Violations	None	None
Intercept	272.119 ± 11.304 (*)	268.196 ± 11.229 (*)
Danceability	34.079 ± 3.483 (*)	29.424 ± 3.506 (*)
Loudness	87.934 ± 6.173 (*)	84.474 ± 6.113 (*)
Energy	-46.787 ± 4.401(*)	-41.946 ± 4.370 (*)
Audio Valence	-16.323 ± 2.583 (*)	-19.042 ± 2.593 (*)
Acousticness	-17.137 ± 2.629 (*)	-12.671 ± 2.609 (*)

Table 1: Summary of characteristics of the final model in the training and test datasets.
Coefficients are presented as estimate ± SE (* = significant t-test at $\alpha = 0.05$).

As can be seen, despite the model with the training dataset and the model with the testing dataset shared similarity in the largest VIF value, the number of Cook's D, number of DIFFITS, no violation against the assumptions, significance in the coefficients, coefficients except for the intercept and the variable, loudness, fitted by the model from the testing dataset failed to be captured by the interval of the corresponding coefficient in the model with the training dataset. In

addition, there is a noticeable difference in Rsq_adj. Therefore, the proposed final model failed to be validated by the testing dataset.

Discussion

Final Model Interpretation and Importance

The coefficients of the final model suggested the relationship to the response.

Take the coefficient for danceability as an example. It represents the average change in song popularity when the input of danceability increases by one unit when all other predictors are held fixed.

Take a general picture of the model. While the predictors, energy, audio valence, and acousticness form a negative relationship with the song population, the predictors, danceability and loudness form a positive relationship with the song population. The relationship of the coefficients to the response is analogous to that, for example, when I created a piece of music that I adjusted those audio features associated with the positive relationship with the song's popularity to the greatest and adjusted the opposite ones to the lowest, then I could make the music hypothetically popular according to this model.

Though the final model failed to be validated, the final model still suggested a potential relationship between the audio features and the popularity of a song and suggested later research to be done in this area.

Limitations of the Analysis

The crucial part of constructing a model was the condition and assumption checks. In my case, as the additional conditions might suggest, the second condition was violated for the final model. This violation suggested that the assumptions coming from the residual plot might not be informative, especially the second assumption of the model. Therefore, the final model did not always account for the relationship accurately. It was possible to overcome this by changing the functional form of the suggested predictor that violated the condition with a transformation. Yet, it is unnecessary since the condition check simply served to suggest later assumptions.

Also, due to the huge sample size to construct the model, it was unclear to eyeball the clustered data points on the plots to check against the conditions and the assumptions. In this case, the checks of the conditions and the assumptions were all built on the general trend or general pattern, and it was hard indeed to explicitly point out the violation. Therefore, there might be a potential violation of the assumption that was not pointed out. To make a more accurate decision on the violation, one might consider rescaling the plots to check the distribution in a much clearer way.

In addition, the final model was selected with not all conditions it performed better than another candidate model, the transformed full model. The effect might be that the final reduced model would still not able to account for certain cases, while the transformed full model could. It was hard to overcome this since there was always a trade-off among different indicator checks, even though one might perform better at some aspects. In this case, since it performs better at the

partial F test and other indicators suggested slightly worse performance than the full mode, I would take the reduced one as the final model.

Lastly, the final model did not seem to be validated. The effect might be that the final model might not produce an accurate result. It was hard to overcome this, given that there are too many influential observations in both the training and the testing datasets, which are also legit observations.

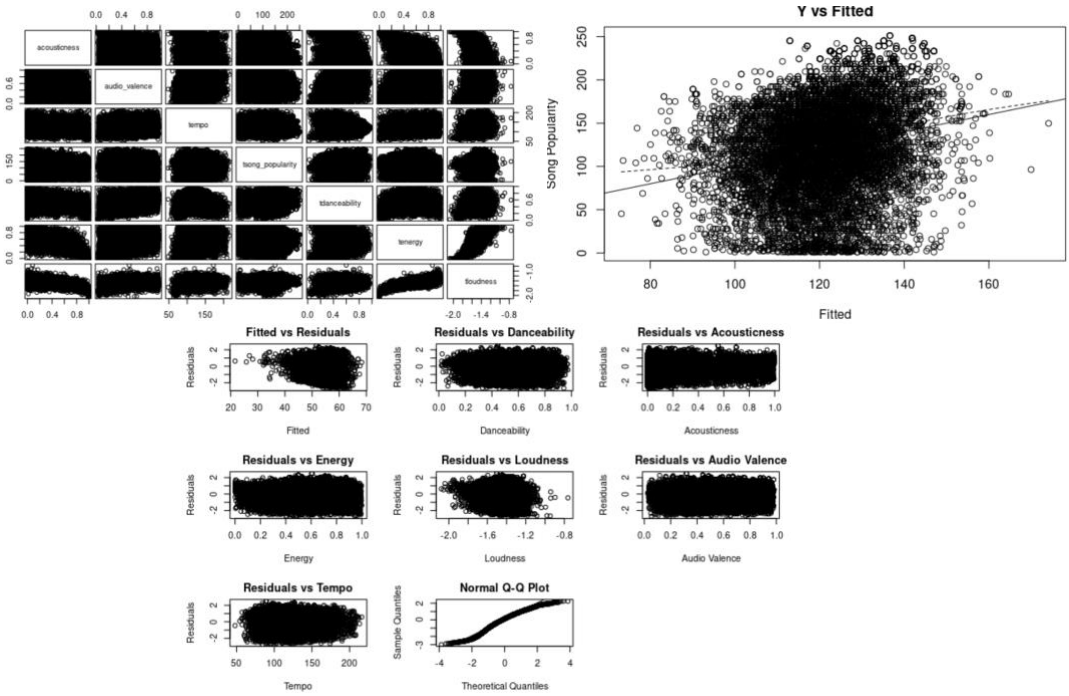
Bibliography

Boyle, J. D., Hosterman, G. L., & Ramsey, D. S. (1981). Factors influencing pop music preferences of young people. *Journal of Research in Music Educa ion*, *29*(1), 47–55.

Delsing, M. J., Ter Bogt, T. F., Engels, R. C., & Meeus, W. H. (2008). Adolescents' music preferences and personality characteristics. *European Journal of Personality: Published for the European Association of Personality Psychology*, *22*(2), 109–130.

North, A. C., & Hargreaves, D. J. (1996). Situational influences on reported musical preference. *Psychomusicology: A Journal of Research in Music Cognition*, *15*(1–2), 30.

Appendix A



Appendix A: The pairwise scatter plot (top left) and the Y and fitted Y plot (top right) for the transformed model. The residual plots (bottom) for each variable and the Q-Q plot at the end.