SPOTIFY DATA ANALYSIS

A LINEAR REGRESSION MODEL

SYBMITTED TO :
Mr. Santanoo Pattnaik

SUBMITTED BY:
SHREYA SISODIA
SIDDHI SHRIVASTWA
SNEHA GOYAL

CONTENT

PROBLEM STATEMENT

We aim to predict the energy level of songs based on various audio features using linear regression model.

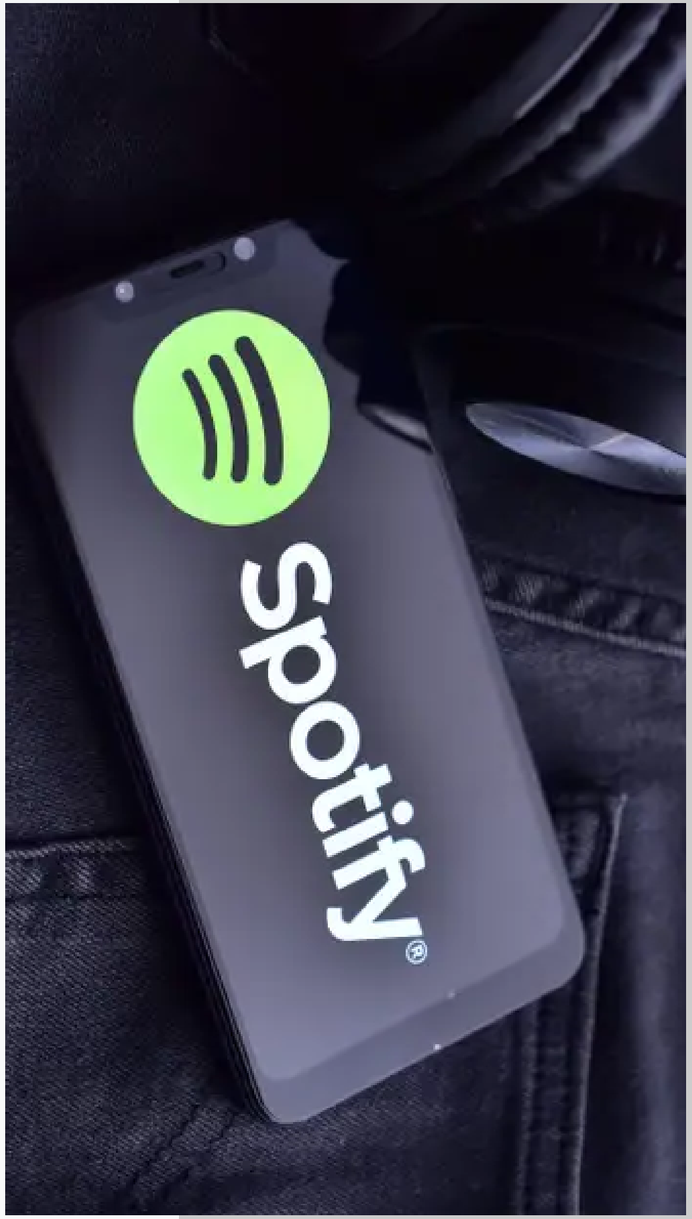
The purpose of analyzing this data is to gain insights into user behavior, preferences, and trends, which can be used to improve the platform's features, personalize recommendations, and make data-driven business decisions.

INTRODUCTION

The emergence of digital music streaming platforms has revolutionized the way we consume music, providing access to an extensive library of songs at our fingertips.

Spotify data analysis refers to the process of exploring, analyzing, and deriving meaningful insights from the data generated by users on the platform.

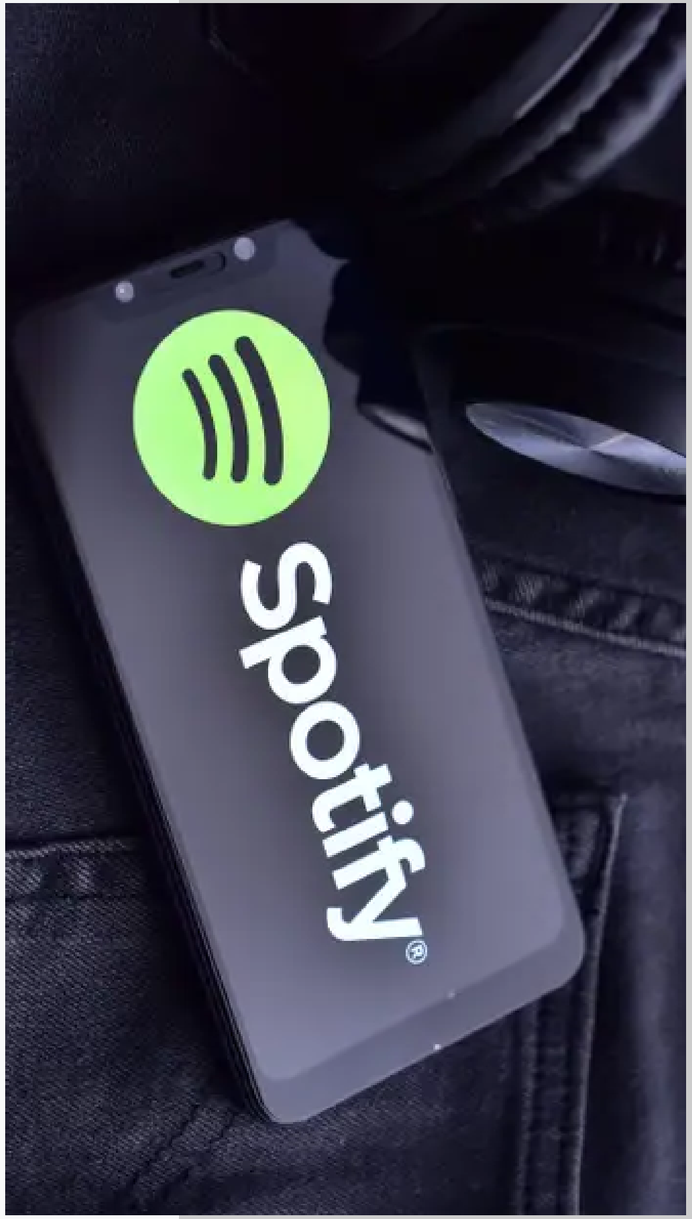
Spotify has emerged as a prominent platform with millions of users worldwide.



INTRODUCTION

With its vast user base and rich dataset, Spotify presents a valuable opportunity for data analysis to uncover insights into music preferences, user behavior, and industry trends.

In addition to improving the user experience, Spotify data analysis also provides valuable insights for artists, record labels, and music industry professionals

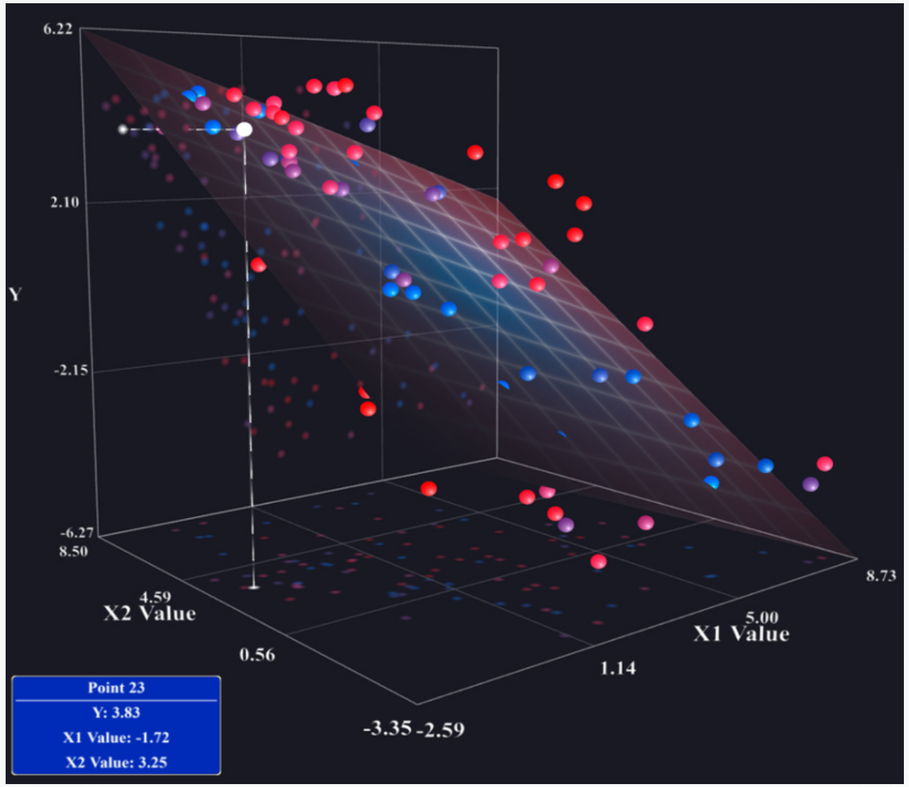


MULTIPLE LINEAR REGRESSION

Multiple linear regression is a statistical modeling technique used to predict the relationship between multiple independent variables (various type of audio feature) and a dependent variable.

The equation can be expressed as:

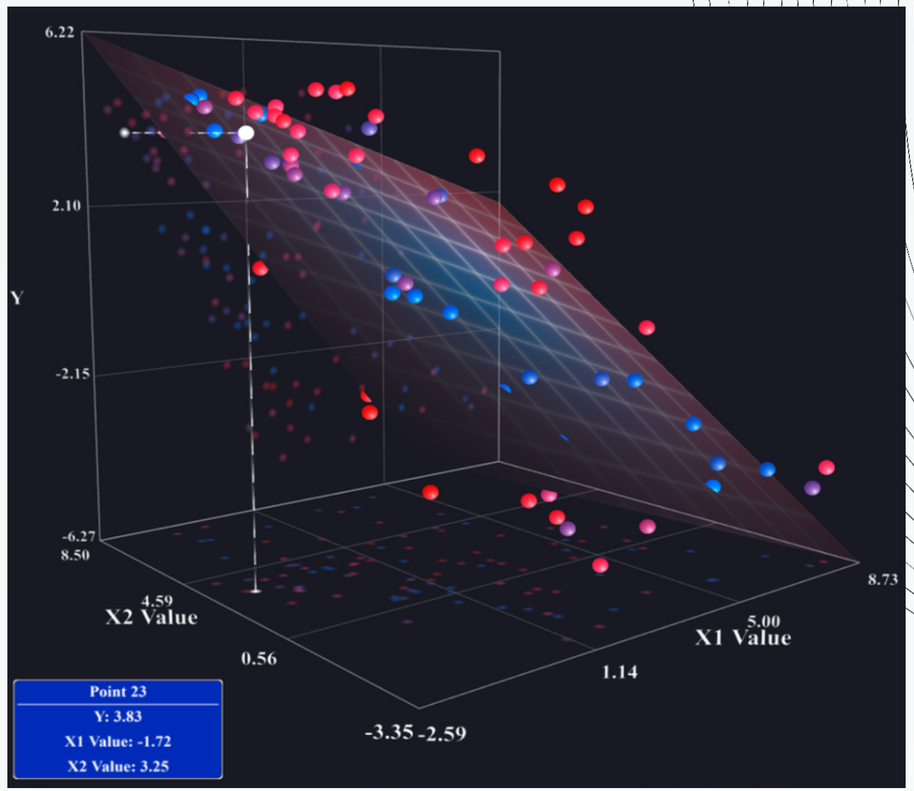
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$



MULTIPLE LINEAR REGRESSION

CONDITIONS OF APPLICATION

1. Linearity: The relationship between the two variables should be linear (at least roughly).
2. Independence: Observations must be independent.
3. Normality of the residuals: Normality testing is used to determine whether sample data has been drawn from a normally distributed population (within some tolerance).
4. Homoscedasticity of the residuals: The variance of the errors should be constant.
5. No influential points: If the data contain outliers, it is essential to identify them so that they do not, on their own, influence the results of the regression.
6. No Multicollinearity: It arises when there is a strong linear correlation between the independent variables, conditional on the other variables in the model and it may lead to an imprecision or an instability of the estimated parameters when a variable changes.

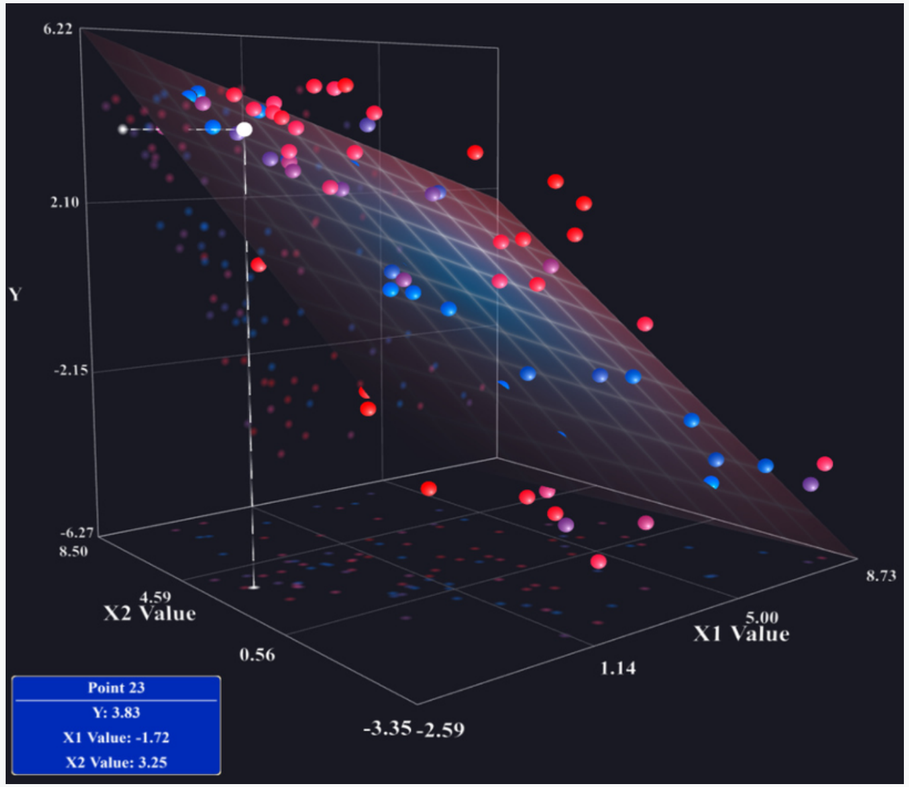


MULTIPLE LINEAR REGRESSION

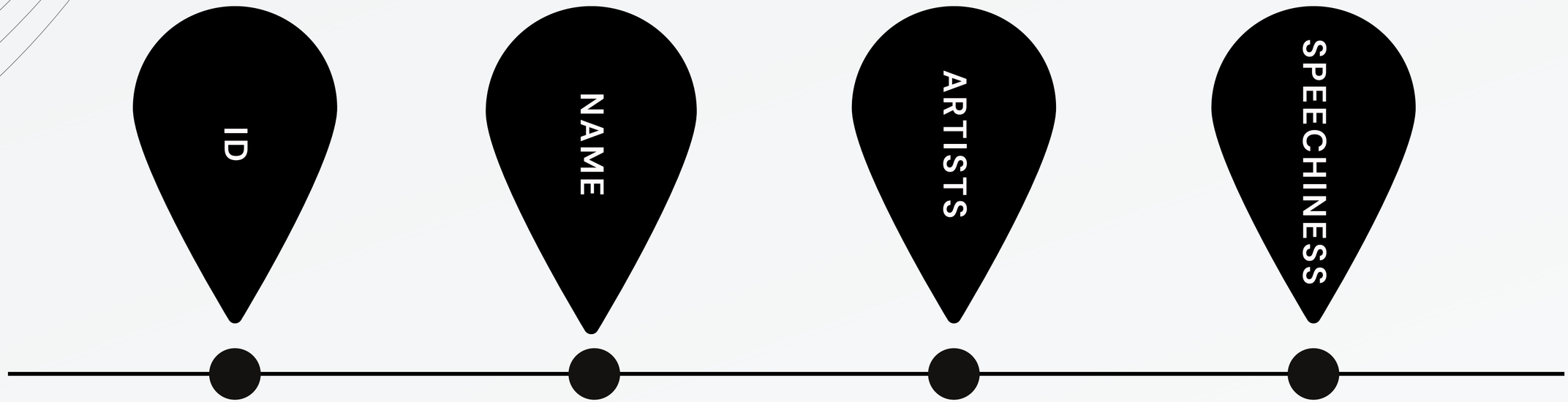
CHOOSING A GOOD MODEL

Three most common tools to select a good linear model are:

- The p-value associated to the model,
- The coefficient of determination R2
- The Akaike Information Criterion (AIC)



ABOUT THE DATASET



ID

Spotify URI of the song

NAME

Name of the song

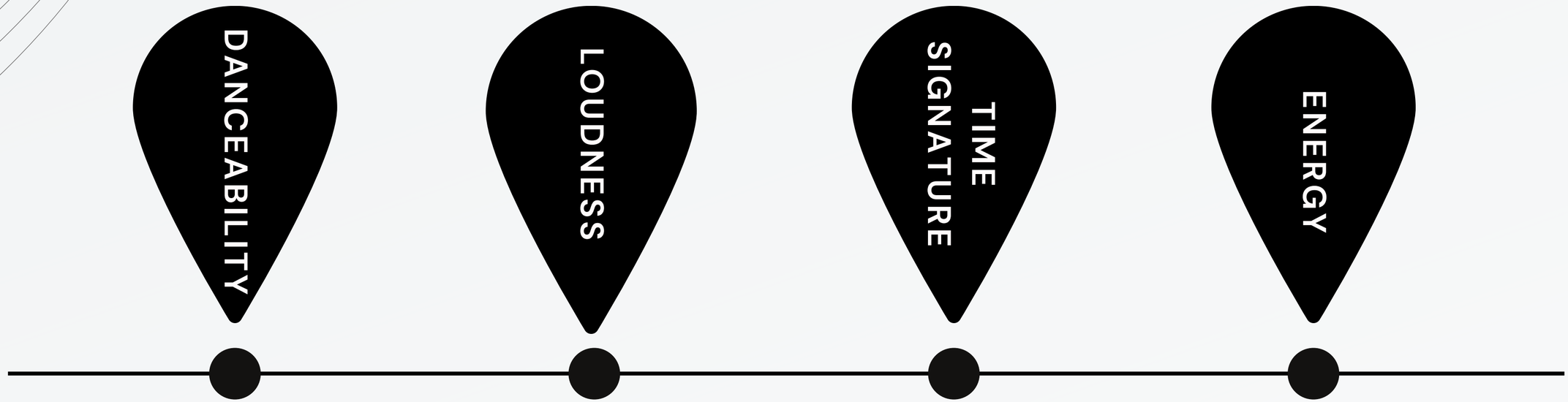
ARTISTS

Artist(s) of the song

SPEECHINESS

It detects the presence of spoken words in a track

ABOUT THE DATASET



DANCEABILITY

How suitable a track is for dancing based on a musical elements including tempo, rhythm stability, beat strength, and overall regularity.

LOUDNESS

It is the quality of sound, correlates to amplitude.

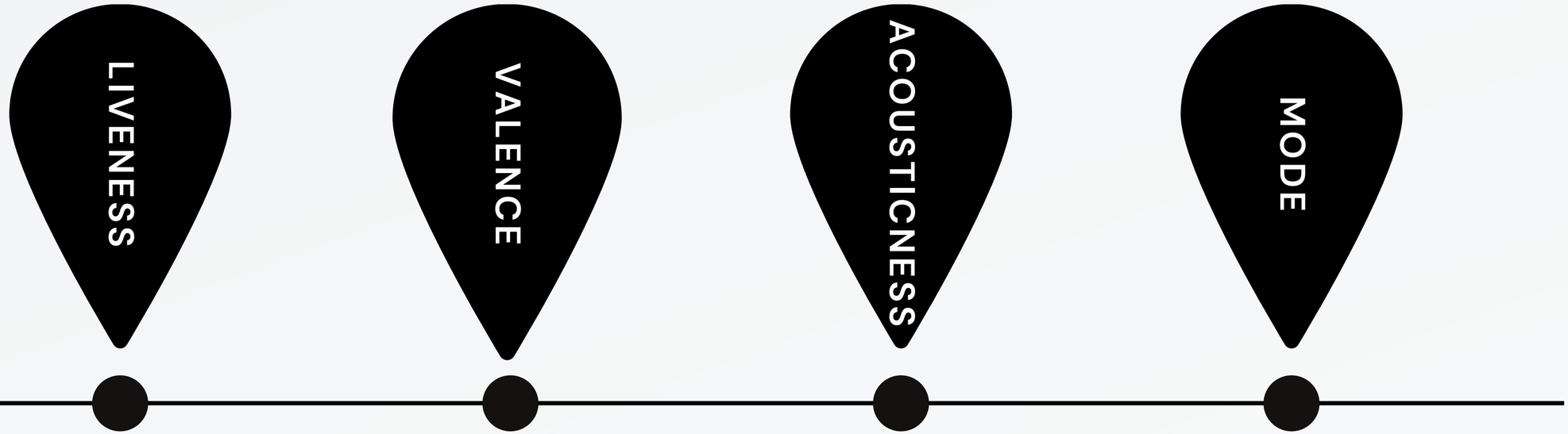
TIME SIGNATURE

A notational convention to specify how many beats are in each bar .

ENERGY

Energy represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.

ABOUT THE DATASET



LIVENESS

Detects the presence of an audience.High values represent an increased probability that the track was performed live.

VALENCE

0.0 to 1.0 describing the musical positiveness conveyed by a track.

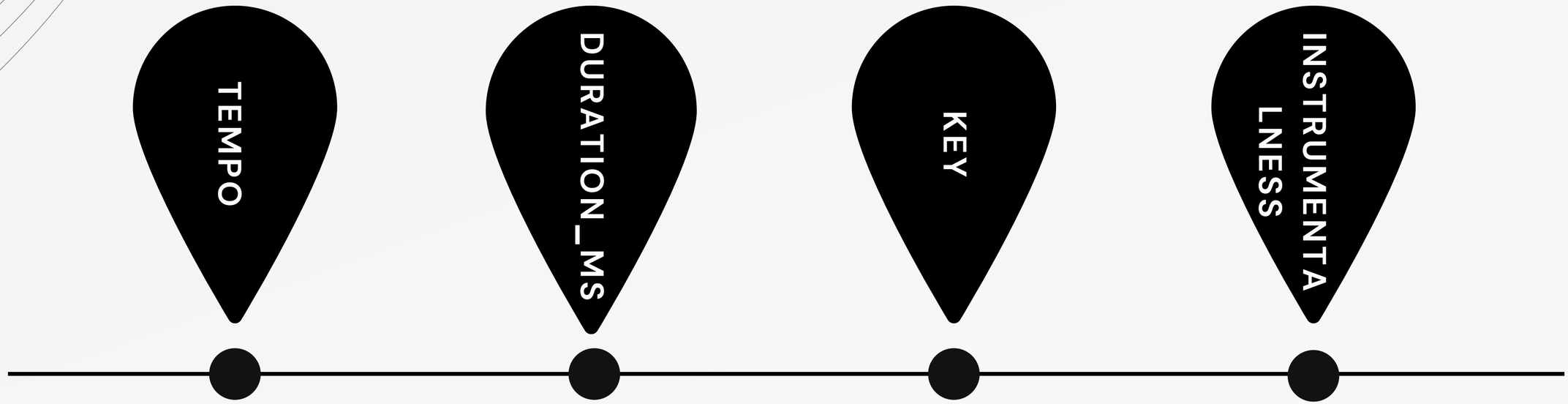
ACOUSTICNESS

A confidence measure From 0.0 to 1.0.

MODE

It indicates the modality (major (1) or minor(0)) of a track, the type of scale from which its melodic content is derived.

ABOUT THE DATASET



TEMPO

It is the pace of a given piece and derived directly from the average beat duration.

DURATION_MS

The duration of the track in milliseconds

KEY

The key the track is in, integers map to pitches.

INSTRUMENTALNESS

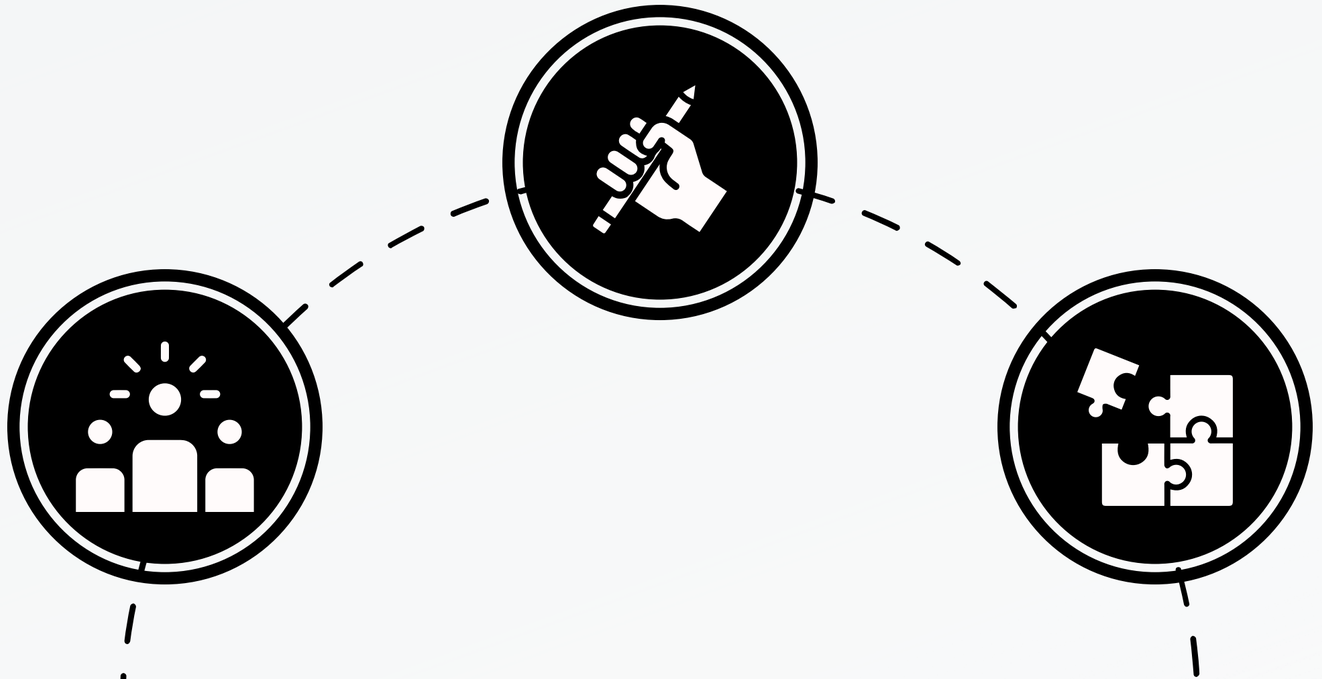
Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal".

METHODOLOGY

DATA ACQUISITION

DATA PREPROCESSING

DATA FITTING



DATA ACQUISITION

The dataset of Top 100 Songs of 2018 from Spotify was obtained and used as the primary source of data for this project.

The dataset consists of 100 songs and their corresponding audio features, providing a comprehensive pool of user preferences and song information. The audio features were extracted from spotify web API and spotify python library

DATA PREPROCESSING

DATA CLEANING

The dataset was checked and corrected, for errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Our Data has cleaned.

DATA TRANSFORMATION

- Audio features like key, valence, mode and time signature were converted to type factor.
- Numerical keys were converted to the actual musical keys.
- The duration milliseconds was converted to mins .
- The values of valence were categorized as happy,sad,Euphoric.

DATA PREPROCESSING

DATA REDUCTION

No Requirement of any deduction.
Each attribute of the data for our model is important.

SPLITTING THE DATASET

The data is split into 2 parts in a ratio of 8:2.
The data set is randomly distributed and the distribution is done with replacement.

DATA FITTING

CHOOSING AN ALGORITHM

Models train based on the data that has been processed and annotated.

Regression helps to look for this correlation and predict an output. Thus, here, energy will be evaluated based on other audio features.

EVALUATION

Cross-validation technique is used to evaluate the performance of a machine learning model.

It helps in obtaining performance estimates for the model across different subsets of the data.

FEATURE SELECTION

Model uses backward elimination and both elimination to achieve best set of features.

Combining backward elimination and both elimination provides a more thorough exploration of the predictor variables,

RECREATE MODEL

Final model was made after performing feature selection techniques and using the best suggested set of audio features to predict energy of a song.

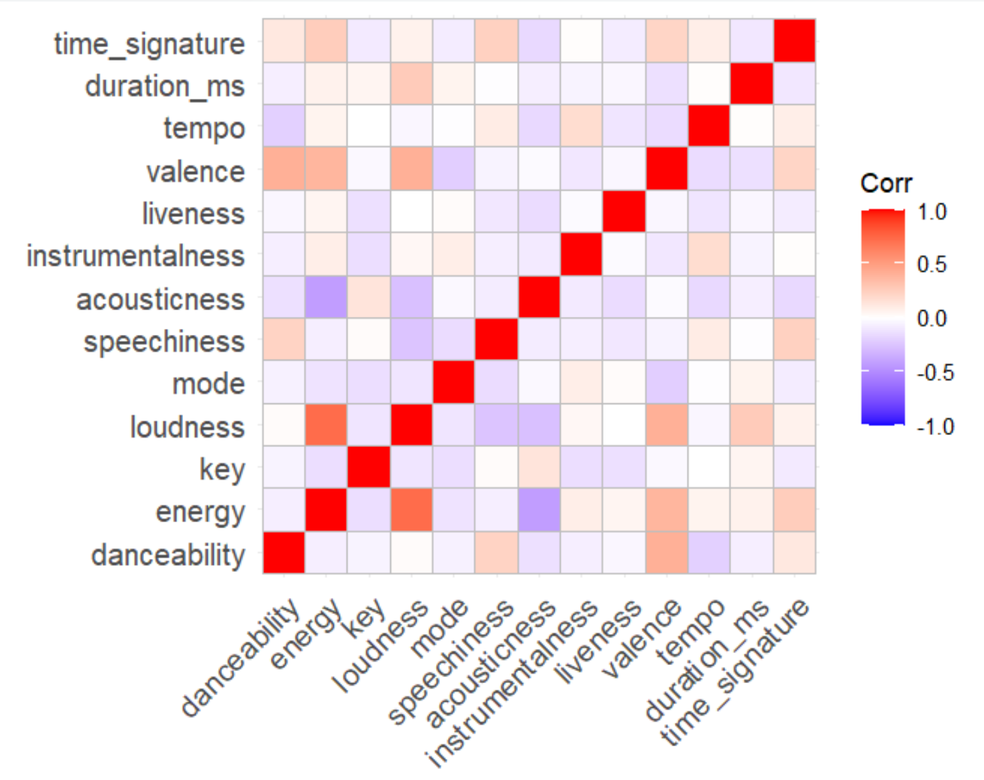
RESULT

The Root Mean Squared Error (RMSE) provides an estimation of how well the model is able to predict the target value (accuracy).
For our model, a RMSE value of 0.08134404 is obtained indicating a very small average difference between values predicted by a model and the actual values.

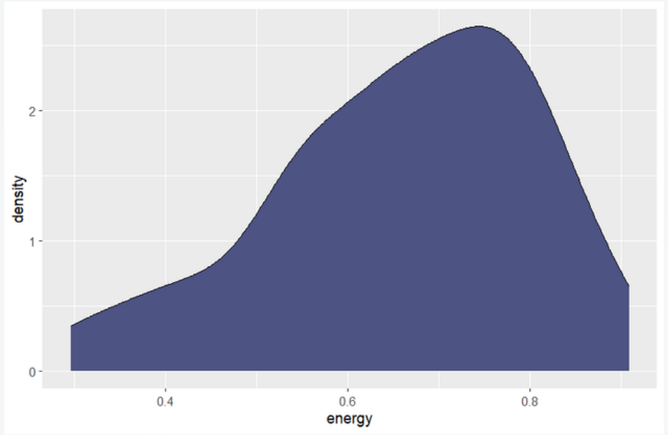
R-squared is a goodness-of-fit measure for linear regression models and indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.
For our model, a R2 value of 0.7877681 is obtained indicating that the independent variables greatly explain all the variance in the dependent variable.

RESULT

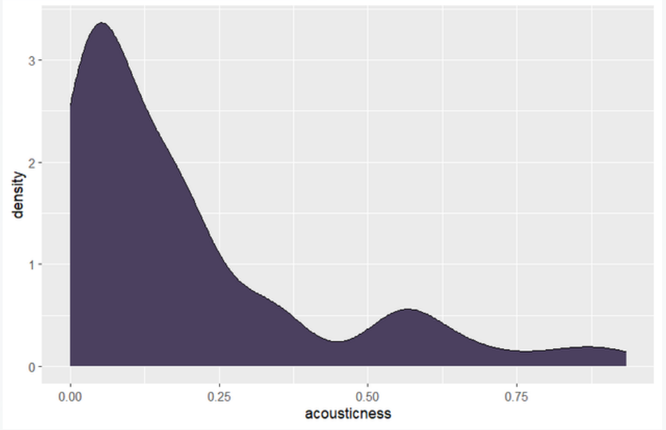
CORRELATION GRAPH



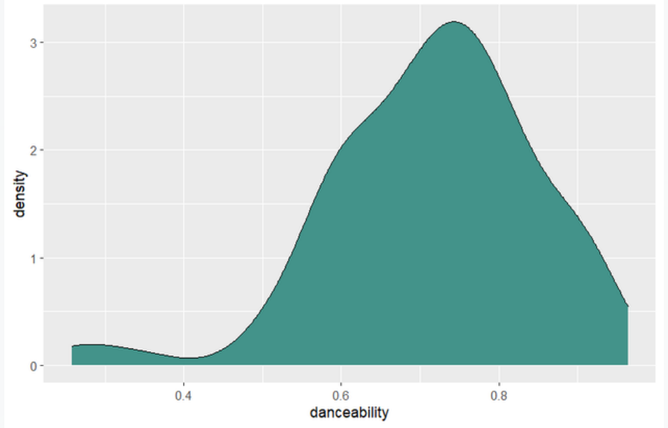
ENERGY ANALYSIS



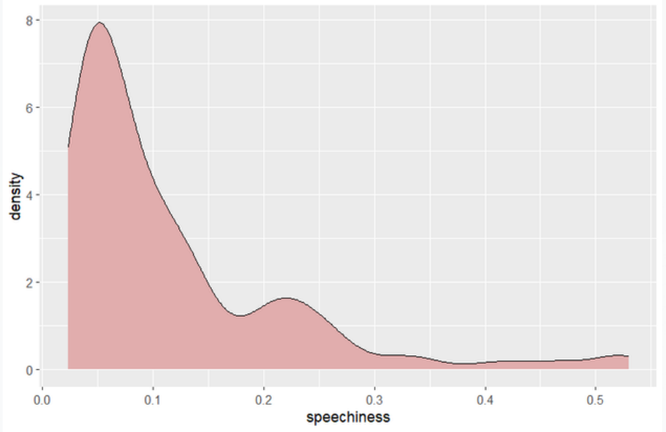
ACOUSTICNESS ANALYSIS



DANCEABILITY ANALYSIS



SPEECHINESS ANALYSIS



EVALUATION

p-value indicates if the model is better than a model with only the intercept. **We have a p value of 5.17 e-12.** so we conclude that our model is better because at least one coefficient β is significantly different from 0.

P - VALUE

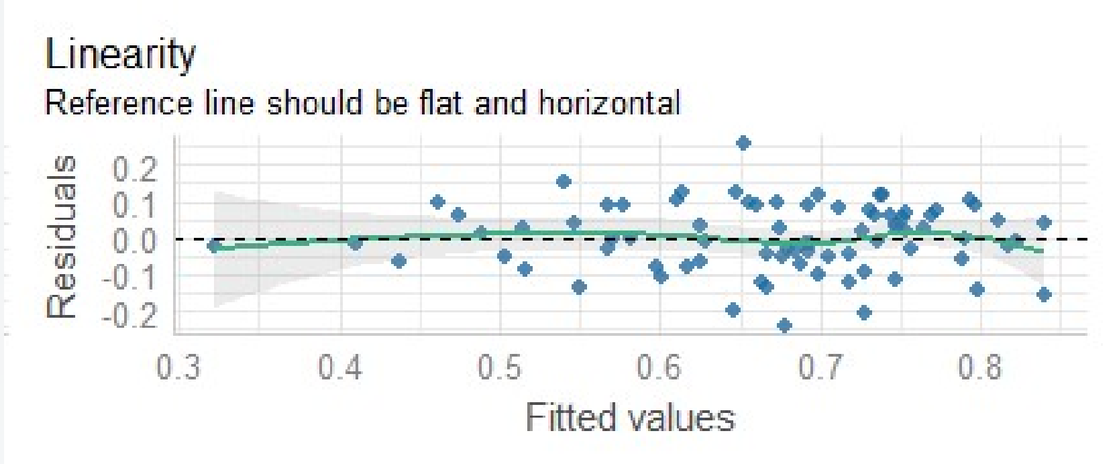
The coefficient of determination, R^2, is a measure of the goodness of fit of the model. As a rule of thumb, a R2>0.7 indicates a good fit of the model.
We have a R^2 value of 0.7877681.

R^2

AIC expresses a desire to fit the model with the smallest number of coefficients possible and allows to compare models. The best model is the one with the lowest AIC. After comparing different variable sets, **the set of variables with AIC = -385.58 (lowest)was chosen.**

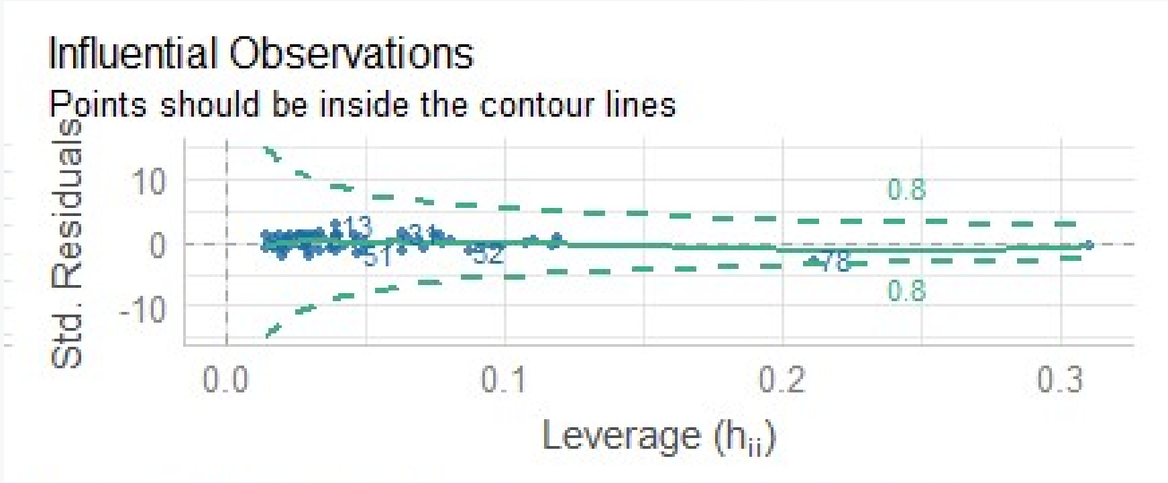
AIC

VALIDATION



- Linearity of a model gives the proportionality between dependent and independent variables. The relationship btw the 2 variables should be linear (at least roughly).

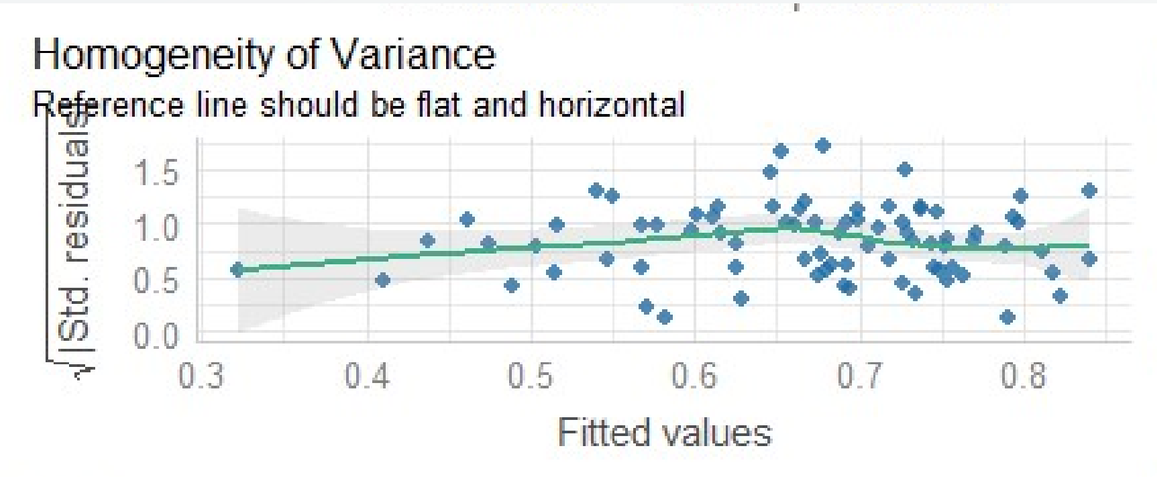
Linearity is almost a straight line.



- If the data contain outliers, it is essential to identify them so that they do not, on their own, influence the results of the regression.

There is no influential points

VALIDATION

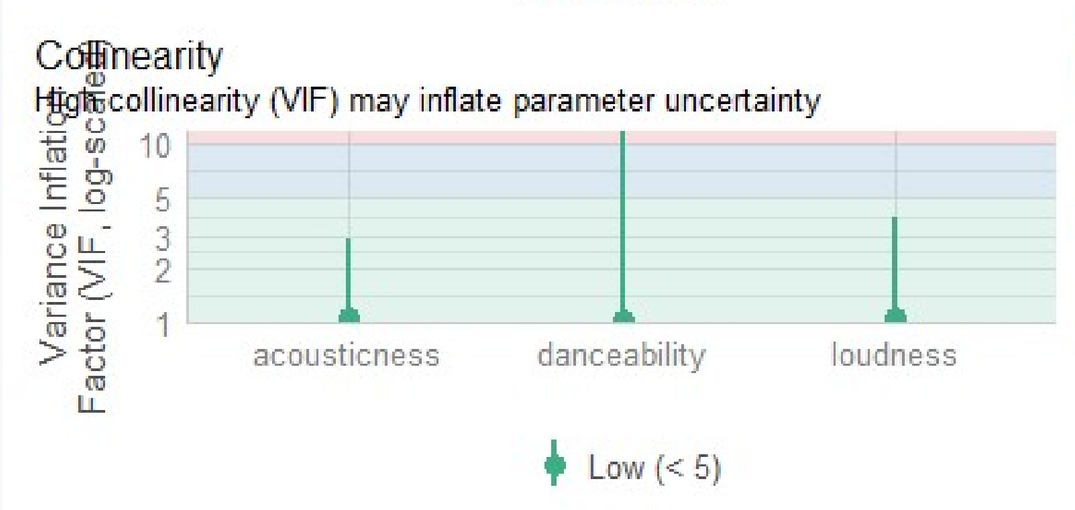


- The variance of the errors should be constant.

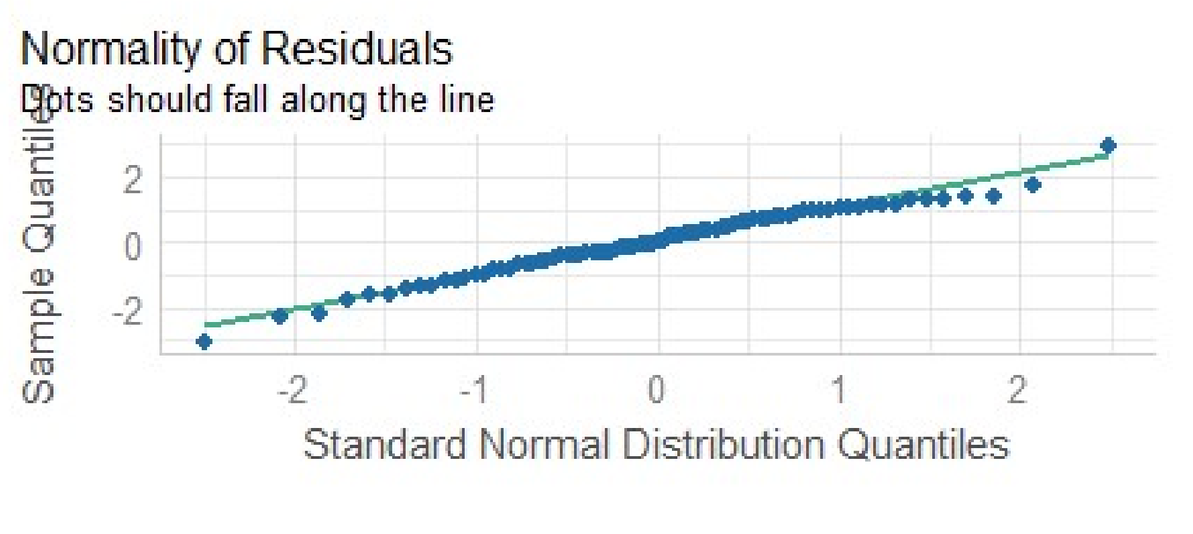
Homogeneity of variance (middle left plot) is respected.

- Multicollinearity arises when there is a strong linear correlation between the independent variables, conditional on the other variables in the model, it may lead to an imprecision or an instability of the estimated parameters when a variable changes.

Multicollinearity is not an issue (I tend to use the threshold of 10 for VIF, and all of them are below 10).



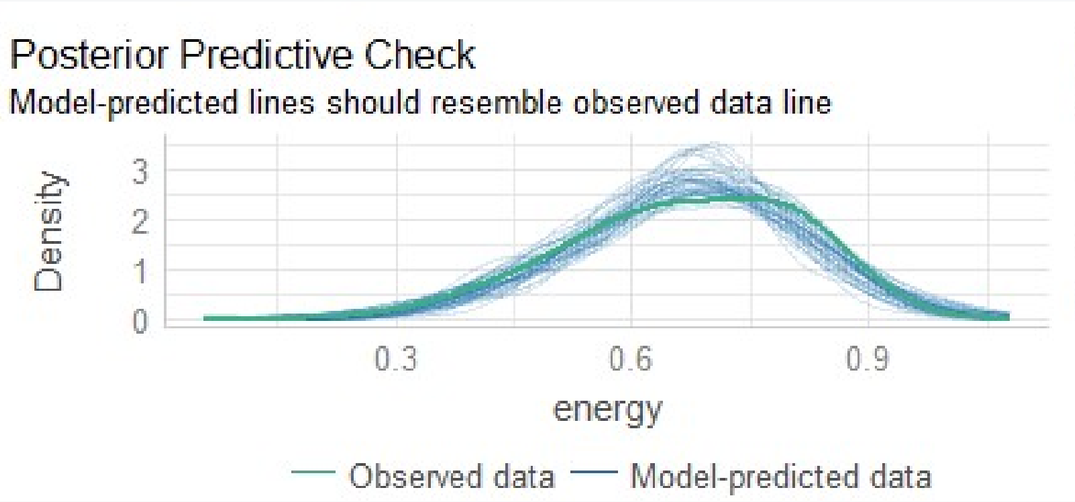
VALIDATION



Normality of the residuals is not perfect due to very few points deviating from the reference line but it still seems acceptable. In any case, the number of observations is large enough given the number of parameters and given the small deviation from normality so tests on the coefficients are (approximately) valid.

Posterior Predictive Check: Model- predicted lines (blue) are somewhat resembling the observed data line(green).

This shows that our fitted model is adequate to describe the data.



CONCLUSION

People like songs with the following audio features:

1. Energetic songs
2. Time duration of the songs between 3-4 mins.
3. High dancebility.
4. Low Speechiness and low instrumentalness (very less speech in the songs)
5. Low accousticness(more inclusion of electric sounds).

Audio features Energy and Loudness correlate the most.

We predicted one audio feature based on the others. With the above model we obtained a considerable r2 value while predicting the energy of a song using the variables loudness, danceability, acousticness

THANKYOU