**Final Project | Statistical**

**Computing**

BANA 6043 | Bhawesh Joshi

04 December 2021

# Project Contents

1. Project Overview
   1. Introduction 1.2.
   2. Basic EDA
2. Visualizations

## 2.1. Correlation Map

## 2.2. PCA Analysis

## 2.3. Box Plot

1. Linear Regression
2. Analyzing Residuals

4.1. Jarque-Bera test

4.2. QQ PLOT

4.3. Shapiro Test

1. Model Performance Measures

## 1.Project Overview

**1.1 Introduction:**

After exploring possible datasets to use for the final project, I decided to analyze Spotify data available in Kaggle. Spotify is a digital music, podcast, and video service that gives you access to millions of songs and other content from creators all over the world.

This project will deal with 50 songs (Observations )having different Track names and artist's name. There are also other variables like Beats per minute, Energy, danceability, Loudness, Liveness, Valence, Length, Acousticness, Speechiness, Popularity. We here to analyze what variable affects the Popularity the most based on the other variables.

This data is in the form of a csv and we will be exploring it in R. we will try fitting a linear regression model to predict the Popularity (independent variable) and evaluate the model's performance.

**1.2 Basic EDA:**

We first conducted basic Exploratory Data Analysis on the dataset in R. After realizing that the first column is only numbering each row, I decided to drop it from our analysis. I then used functions such as head (), summary (), dim(), names(), str(), and is.na,colnames(). The head function showed us the first five observations, verifying that the first variable had been removed. Summary shows each variable’s minimum, 1st quartile, median, 3rd quartile, and maximum. This is only useful for numerical data. Dimension shows us that there are 14 columns and 50 rows, proving that we are working with a small dataset. Names showed us each column name, which is a good reference when exploring individual variables. String shows us the data type; there is numerical data (11 variables), character data (it’s called factor in R – artist name, song name) data in this dataset. Lastly, there are no N/A values in the data set. I used the is.na () function to test that.

## 2.Visualizations

**2.1 Correlation :**

Background pattern

Description automatically generated

From the correlation map(corrgram()) we can see none of the variables are very highly correlated with the Popularity. The most that are correlated are Speechiness and beats per minute.Actually, many of the variables we could use predictors are correlated with each other, raising the chance of multicollinearity. This will be kept in mind as the linear regression model is built.

**2.2 PCA Analysis:**

**Chart, scatter chart

Description automatically generated**

Since there are many numerical variables, the idea of PCA is Used— reduce the number of variables of a data set, while preserving as much information as possible. The relationship between variance and information here, is that the larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has. We again found Popularity, Speechiness are most correlated.

**2.3 Box Plot:**

A picture containing text, screenshot, white, tiled

Description automatically generated

A picture containing chart

Description automatically generated

From the box plot, We observe an outlier which may dissuade our result in our end.as I observed later, the outlier did create an issue in our model and hence it was wise to remove the outlier data.

## 3.Linear Regression

This first model we ran was a multiple linear regression model that tested Speechiness and beats per minute variables to see if they correlated with the Popularity variable.

This produced the following results:

Text

Description automatically generated

This model produced an extremely low R-Squared value of 0.03455, a single digit percentage, indicating this model does not explain the variation in the results very well. The p-value was also high, but not positive significance as it is suggesting that there it is very unlikely we will observe a relationship between Popularity and these selected variables. We then made a second model and tested the Beats per minute variable with respect to Popularity. 46 beats per minute variables and output a slightly higher R-Squared of 0.06 as shown in the figure below.

Text, letter

Description automatically generated

**summarizing model**

**residuals**

Residuals are essentially the difference between the actual observed response values

(Popularity to Beats\_per\_min in our case) and the response values that the model predicted Smaller residuals are better.

When assessing how well the model fit the data, you should look for a symmetrical distribution across these points on the mean value zero (0). In our example, we can see that the distribution of the residuals do not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points.

**Coefficients**

Intercept:

Theoretically, in simple linear regression, the coefficients are two unknown constants that

represent the intercept and slope terms in the linear model. The intercept is the left over when you average the independent and dependent variable. In the simple regression we see that the intercept is much larger meaning there’s a fair amount left over.

The coefficient Estimate contains two rows; the first one is the intercept. The intercept, in our example, is essentially the expected value of the popularity required for a track to be popular when we consider the average beats per min for each track in the dataset. In other words, it takes an average track in our dataset 84.07766 Popularity level.

The second row in the Coefficients is the slope, or in our example, the effect Beats\_per\_min has in popularity level. The slope term in our model is saying that for every 1 bpm increase, the required popularity goes up by 0.02851.

Text

Description automatically generated

## 4.Analyzing Residuals

**4.1 Jarque-Bera test:**

The null hypothesis of the Jarque-Bera test is a joint hypothesis of the skewness being zero and the excess kurtosis being zero. With a p-value >0.05, one would usually say that the data are consistent with having skewness and excess kurtosis zero

**4.2 Q-Q Plot**

Chart, scatter chart

Description automatically generated

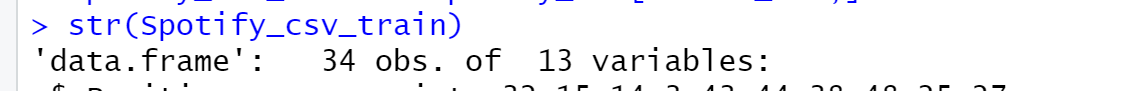
**4.3 Shapiro Test:**

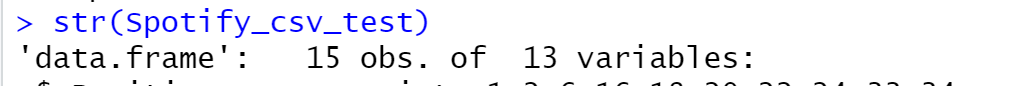
Graphical user interface, application

Description automatically generated

## 5.Model Performance Measures

The data was divided into 2 parts,the train and the test.





Then the model summary was run,to test it.

Text, letter

Description automatically generated

Text, letter

Description automatically generated

Lastly, the conclusion was reached that to maintain the average popularity,the beats per minute should be in the below range:

Table

Description automatically generated

To test the accuracy of the model, RMSE, MAE,MPSE was found out and it revealed there was a very less impacts of errors.

Text

Description automatically generated with medium confidence