# Machine Learning Model

Description of preliminary data preprocessing:

1. Extract, transform and load:
    - Two datasets were extract from Kaggle.com .
    - The datasets were merged by using sql inner join query.
    - The dataset was loaded to notebook by engine via PostGresSql server.

2. Exploratory Data Analysis:
    - Display Data Types.
    - Determine if there is any Null or Missing Values • Display Summary of Statistics • Creating Visualization:
        i. Popularity Density Plot with Seaborn.
        ii. Range and Distribution Plots for numerical features. iii. Multi-Variate Analysis.
        iv. Histogram Plots for each feature
        v. Correlation Heatmap.
    - Creating categorical column for 'popularity' by qcut function. Below the mean popularity will be 0 and above the mean popularity will 1.
    - Convert binary column 'explicit' to numerical datatype.
    - Export the processed dataset to sql table with PostGres server.

Description of preliminary feature engineering and preliminary feature selection, including their decision-making process:

1. Feature engineering and Preliminary Feature Selection
    - Mode column is indicator of modality of each song such Minor and Major and since encoding removes redundancies from data, this column was encoded as "Minor" and "Major".
    - Using bucketing on column 'artist, song and genre 'to improve query performance and sampling then encode mentioned column for removing redundancies.
    - **Update July 3, 2022**: Removed object features such as 'Artist' and 'Song' columns as well as discrete feature 'year' and liveness appears to be no impact on the model it was also removed.
    - Upon completion of bucketing and encoding, initial columns dropped from the dataset and ready for splitting as training and testing data.
    - Since we algorithm is still on early stage we are still discovering on selection of features. However, currently we are keeping every feature from the datasets as of now.
    - **Update July 3, 2022:** Upon completion of PCA analysis we discovered that there are many features come with variance as close as zero and therefore those features were removed prior to ML training and testing.

Description of how data was split into training and testing sets:

- Target: 'Popularity'
- The data was split on 20% test to 80% training.

Explanation of model choice, including limitations and benefits:

- Neural Network: 2 Hidden Layers with 200 and 100 nodes respectively and activation function as 'Relu' – accuracy: 0.63
- **Update July 3, 2022:** Accuracy **:0.67**
- Random Forest Classifier: accuracy: 0.63
- Support Vector Model: accuracy: 0.518
- Logistic Regression: accuracy: 0.50

**Benefits:**

At this stage, although Neural Network has the highest accuracy, it comes with a high loss as 1.18. Neural Network provides ways to improve accuracy, but it tends to overfit the data. We are currently working on the model to improve accuracy and decrease loss.

**Update July 3, 2022:** ANNs shows the ability to learn and model non-linear and complex relationships, which is important for our model, many of the relationships between inputs and outputs are non-linear as well as complex. After learning from the initial inputs and their relationships, it infer unseen relationships on unseen data as well, thus making the model generalize and predict on unseen data.

**Limitations:**

Since neural network architecture has a fixed number of input layers, it can only take a fixed sized input and output for any task. This is a limiting factor for many patterns recognition tasks.

**Update July 3, 2022:** Upon completion of PCA analysis we discovered there are many features to removed and increased the accuracy dramatically.

However**,** gathering more data and creating stronger correlation among features would give us better accuracy and confident on making prediction.