**Grade: 96%**

Group - 3

ischool | Syracuse University

**IST718 – PROJECT PROPOSAL**

GROUP 3: ANANTH ADISESH, JUSTIN LIM & YASHILA UPENDRA KUMAR

# **OBJECTIVE:**

To determine and predict the trends in popularity of a song from the Spotify dataset by analyzing appropriate features/variables such as acousticness, danceability, duration, energy level, liveness, speechiness, tempo and instrumental and classifying the data in terms of playlists, genres, artists etc. for a time span of 6 years from 2015 to 2020 by applying advanced machine learning concepts to derive insightful information from data models.

# **DATASET DESCRIPTION:**

* Overview/description: The dataset consists of Spotify songs from the year 2015 to 2020 with variables:

#### **Primary:**

**- id** (Id of track generated by Spotify)

#### **Numerical:**

**- acousticness** (Ranges from 0 to 1)

**- danceability** (Ranges from 0 to 1)

**- energy** (Ranges from 0 to 1)

**- duration\_ms** (Integer typically ranging from 200k to 300k)

**- instrumentalness** (Ranges from 0 to 1)

**- valence** (Ranges from 0 to 1)

**- popularity** (Ranges from 0 to 100)

**- tempo** (Float typically ranging from 50 to 150)

**- liveness** (Ranges from 0 to 1)

**- loudness** (Float typically ranging from -60 to 0)

**- speechiness** (Ranges from 0 to 1)

**- year** (Ranges from 1921 to 2020 reduced to 2015 to 2020)

#### **Dummy:**

**- mode** (0 = Minor, 1 = Major)

**- explicit** (0 = No explicit content, 1 = Explicit content)

#### **Categorical:**

**- key** (All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on…)

**- artists** (List of artists mentioned)

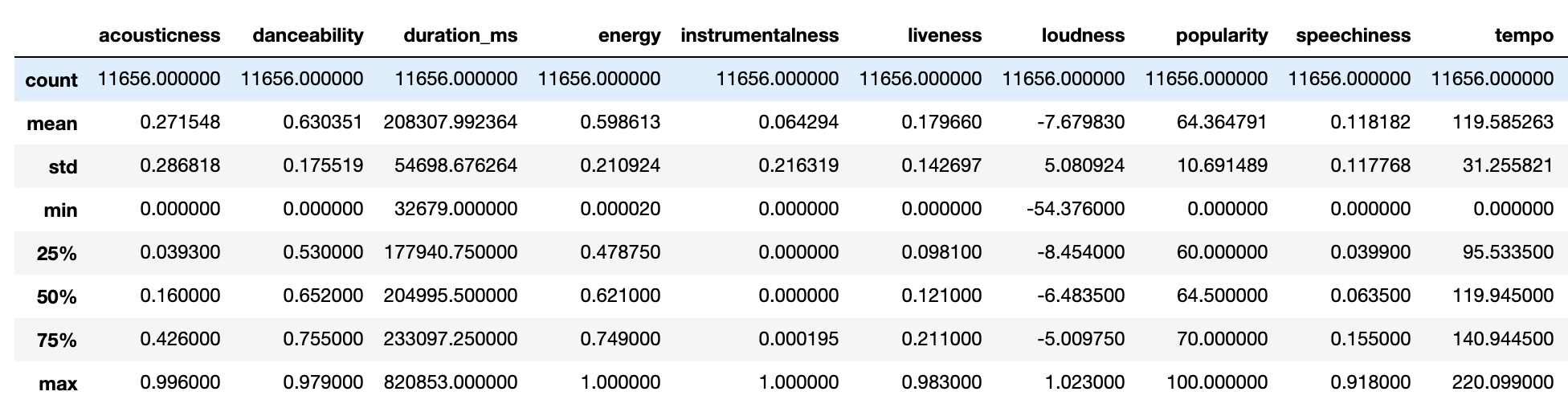
**- release\_date** (Date of release mostly in yyyy-mm-dd format, however precision of date may vary)

**- name** (Name of the song)

* Rows and columns: The dataset contains 19 columns and 11656 rows in total.
* Sample predictors: acousticness, danceability, duration, energy level, liveness, speechiness, tempo, instrumental, id, mode, release date, valence, song name and artists.
* Link to the dataset: <https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks>
* Surprising facts about the data: The popularity column indicates a score ranging from 0 – 100 but most of the songs do not have a rating more than 2 indicating there are very few popular songs in the dataset.

# **PRELIMINARY DATA EXPLORATION:**

The original dataset contains Spotify music data from 1921-2020 with 160k+ rows and 19 columns. The dataset is reduced to more recent years (2015 to 2020) to analyze the current trend in the music industry. After the reduction, the dataset contains 11,656 rows and 19 columns. Figure 1 shows the summary statistics of numeric variables in the data.



<Figure 1: Summary Statistics of numeric variables>

Chart, histogram

Description automatically generatedAs mentioned in the previous section, the popularity column from the original dataset contains significantly greater number of zeros with the mean value of 31.55. However, with the recent data, the mean of the popularity column increased to 64.36. Figure 2 compares two histograms.

Chart, histogram

Description automatically generated<Figure 2: Popularity histogram comparison>

The lack of measuring popularity in the past compared to recent years could be a reason why data from 1921 to 2020 have significantly a greater number of lower values in the popularity column. Additionally, changes in the music characteristics from 2015 to 2020 is explored. The characteristics included in figure 3 are acousticness, danceability, energy, instrumentalness, liveness, speechiness, and valence. These variables are in same scale, which make the figure comprehensible. Danceability and speechiness shows upward trend, while energy and instrumentalness shows downward trend.

Chart, line chart

Description automatically generated

<Figure 3: Changes in music characteristics from 2015 to 2020>

# **PREDICTIONS:**

* Predict a popularity rate based on other features of the song such as **danceability, acousticness, tempo, duration, energy etc.**
* Analyzing the variation in the song trends over the years, i.e. by performing a time series analysis of the data
* Predicting the popularity of a Spotify artist based on popularity of the song, i.e. identifying the most popular artists in the dataset
* Analyzing the popularity in terms of song genres, i.e. determining which genres of songs in the dataset has the highest level of popularity.

# **INFERENCE:**

* Determining most popular songs and artists based on their popularity.
* Interpreting what factors drives popularity and what insights can be determined from the data analysis.

# **NON-SPARK PACKAGES**

Tentative list of Non-spark packages:

* Sklearn, warnings.

Sklearn: Since Sklearn supports python numerical and scientific libraries such as SciPy and NumPy we plan to use it to support our analysis

Warnings: We will be using this package to ignore insignificant warnings