# **Spotify Dataset Analysis (group16)**

# **Authors**: Rahul Dilip Wable | Sanjana Sanjay Bhot | Jaishree Palaniswamy | Aditya Vinod Kini

# **Objective:**

# Spotify is the largest music streaming service available. Spotify has around 286 million users including 131 million paying subscribers. It is the ideal platform for artists to reach their audiences across the world. In this project, we have intended to understand the evolution of music over time as well as to understand the characteristics of various genres of music popular among the users. The main goal of this project is to ultimately provide recommendations of songs to users based on their preference, based on different features of the user’s playlist.

# **Data Set Description:**

# **Overview / Description:**

# Spotify dataset consists of around 1,60,000 songs released between 1921-2020, including the details about the artist, the year it was published, duration, etc. This dataset was collected from Spotify Web API and it comprises explanatory variables — audio features such as valence, loudness, and danceability as well as more general characteristics such as genre, title, artist, and year released.

* **Number of rows and cols:**

This dataset consists of **19** columns including the target variable and **169910** rows.

Also, certain data are grouped by artist, year, or genre.

Grouped datasets:

Data by the artist – 15 columns and 27622 rows

Data by genres – 14 columns and 2665 rows

Data by year – 14 columns and 101 rows

* **Sample predictors (does not need to be an exhaustive list):**

From our initial analysis, we have identified some attributes that significantly influence the popularity, they are genre, artists follower count, danceability.

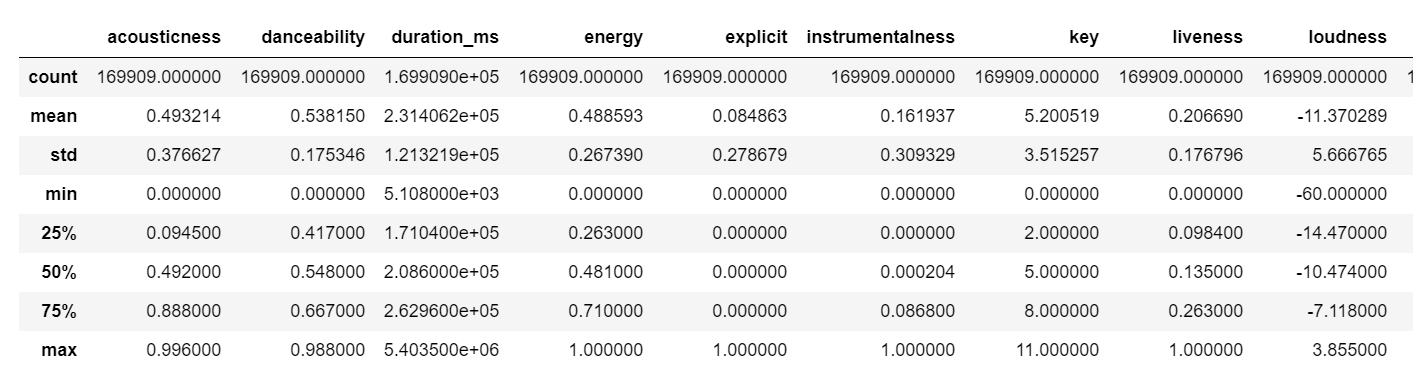
* **Anything interesting or surprising about the data:**

The most interesting aspect of the data is that we have several factors that make recommending songs difficult because there could be no similarity in features between two similar songs which completely has different genres and audiences. Thus, it would be a challenging task to analyze the data to come up with an appropriate model to recommend the songs.

**Dataset Link:** [**https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks**](https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks)

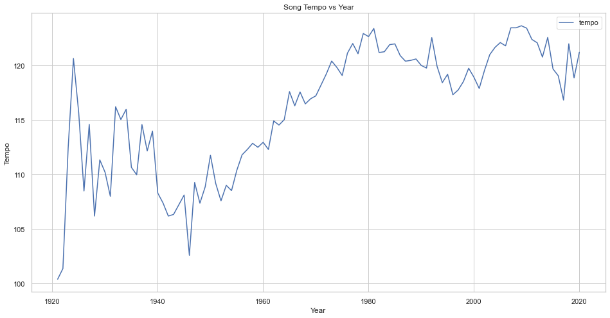
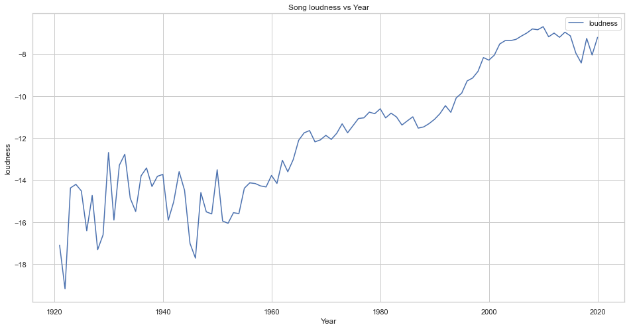
**Preliminary Data Exploration:**

**Summary Statistic:**

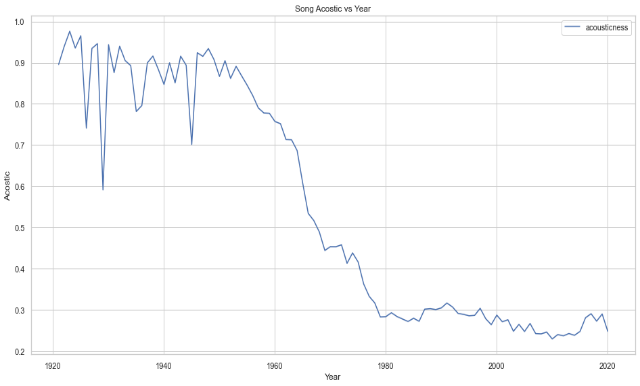
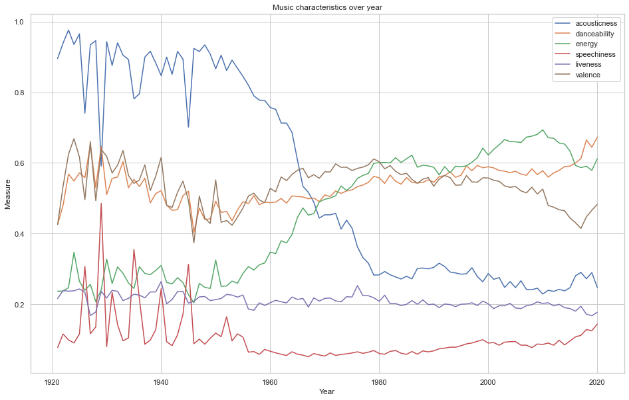
****.

**Interesting interpretations from the data:**

* Over the years the loudness and tempo of the songs have increased.



* The tracks have become less "Acoustic" over the years.
* A song has various characteristics that may be responsible for its popularity. It is difficult to make a conclusion if the song will be popular or not just based on these attributes. After plotting the line graph for some of the characteristics, we can see that these characteristics follow a trend that is highly influenced by when the song was released (Year). People tend to like certain types of music with similar characteristics for certain periods of time. These characteristics may vary with change in people’s preferences over the years.

****

**Prediction:**

Business problem- The central issue for our project is to provide recommendations of songs to a user. Our system will help the users to improve their listening experience which is unique and more personalized to them. It will also help them explore similar artists and genres from a wide range of Spotify playlists of songs. We aim to gain the following insights:

* **Analysis of artists-** Our aim will be to create an artist-based recommendation system that recommends users with similar artists.
* **Performing time-series analysis-** We will also be performing a time series analysis that visualizes the evolution of songs over time to understand the trends in music.
* **Recommending similar songs-** We will be recommending songs by analyzing, differentiating, and clustering the genres.

**Inference:**

Our project revolves around analyzing and recommending artists, songs from the Spotify dataset.

* **Artist based recommender**- This will be carried out by first choosing user top liked genres. Then, filtering out artists liked by the user. Lastly, sorting our artists who are popular in this user favorite genre. This will offer users with more set of similar artists to create a personalized experience.
* **Time series analysis of songs**- To build this we will first find the mean of the audio features contained in the dataset (acoustic ness, danceability, energy, etc.) and group them by year to which they are associated. We will then be plotting them using line graphs to observe the trends in music over the years. This will help us understand the trends over the years which will eventually improve our recommendation system.
* **Music recommendations**- We will be clustering the genres by making use of the K-means clustering algorithm. The K- means the algorithm will be used to find the nearest distance between two songs. Songs that have less distance will be clustered to form one genre. Songs in one genre will be considered similar songs and will thus be recommended to the user.

**Non-Spark packages:**

We will be making use of various packages during our project

* **Sklearn**: We will be using it for the purpose of data analysis and data mining. Additionally, we will make use of various categories here such as clustering, pre-processing, creating alterative split function.
* **Plotly**: We will make use of this for creating informational and interactive data visualizations to improve recommendations and time-series analysis for trends in music over the years for the users.
* **SciPy**: We will be making use of this package for Data statistics.

As we move ahead into the project, we also might make use of additional packages if required.