**Future of Music – Spotify**

---

title: "Spotify Data Analysis and Descriptive Statistics"

output: html\_notebook

author: CS Duals

date: November 18, 2020

---

1. Project Value and Impact

The overarching objective for this project is to predict the next hit songs and understand what features drives popularity over time. This will give music professionals (music producer, song writer, singer, etc) a data-driven approach to be successful in the industry.

In order to do this, we will be analyzing different variable relations, music trends over different decades, and feature significance with popularity. In the next few sections, we have shown in details different steps to clean the dataset, draw different insights/visualization based on initial analysis of the data, and create two predictive models to meet our objective.

2. Data Set Description

For this project, our team decided to take a data-driven approach to evaluate music. We are using the Spotify Dataset from 1921 to 2020 that consists of over 160,000 different tracks (from Kaggle). In order to stay consistent in the evaluation, all data was sourced from the Spotify Web API. The following set of data is combination of Primary, Numerical, Dummy(binary), and Categorical data. This allows us to explore different types of models and draw different insights. Here are all the variables that is included in the data set:

Primary

- id: this an unique key comprised of numbers and characters that is assigned to each track generated by Spotify

Numerical

- acousticness: range from 0(LOW) to 1(HIGH)

- danceability: range from 0(LOW) to 1(HIGH)

- energy: range from 0(LOW) to 1(HIGH)

- duration\_ms: majority range from 200,000 to 300,000

- instrumentalness: range from 0(LOW) to 1(HIGH)

- valence: range from 0(LOW) to 1(HIGH)

- popularity: range from 0(LOW) to 100(HIGH) – this is the main dependent variable

- tempo: majority range from 50(LOW) to 150(high)

- liveness: range from 0(LOW) to 1(HIGH)

- loudness majority range from -60 to 0

- speechiness: range from 0(LOW) to 1(HIGH)

- year: range from 1921 to 2020

Dummy/Binary

- mode: 0 represents minor and 1 represents major

- explicit: 0 represents no explicit content and 1 represents explicit content

Categorical

- key: this consists of all different music keys on octave encoded from 0 to 11 (i.e. C = 0, C# = 1, etc...)

- artists: the artist of the track

- release\_date: the date of release in yyyy-mm-dd format

- name: the name of the track