# Group Project

CS 488

Spring 2022

Stage 2

Group: Kerry Forsythe, Roderick Hutterer, David Torres

## Motivation

Data is fundamental to the operation of modern cloud-based services. In addition to organizing their catalogs of products or materials, companies have the opportunity to collect data from their users’ interactions with the app. Understanding who, what, when, where, and why can help a company better serve their customers through analysis such as classifying products, identifying purchase trends, identifying patterns in users’ past behavior, predict their future behavior, identify commonalities between users, and recommending products and services.

## Problem

In this project, we propose to data mine a sub-set of the Spotify music catalog and user data to provide descriptive and predictive analytics. Spotify songs include attributes such as the following:

* Beats Per Minute (BPM) — The tempo of the song.
* Energy — The energy of a song, the higher the value, the more energetic.
* Danceability — The higher the value, the easier it is to dance to this song.
* Loudness — The higher the value, the louder the song (in dB).
* Valence — The higher the value, the more positive mood for the song.
* Length — The duration of the song.
* Acousticness — The higher the value the more acoustic the song is.
* Release Year — The year each song was released.
* Popularity — The higher the value the more popular the song is.

Some questions might include:

Do certain characteristics of a song predict popularity? Using a **classification** technique, we can construct a training and testing data set to model and predict song popularity.

What other songs might a user like? Using **clustering**, we can a user’s play history to try to identify songs with similar characteristics to recommend to a user.

Other possible tasks might include:

Which artists are often played together? Using association to identify commonalities (similar to the beer, milk, diaper, shopping basket example from class)

How long will a song be popular on the top charts? Using time-series analysis to identify trends in song popularity.

**Software & Datasets**

For our dataset, we have a few different selections of data to use as training data and testing data. Our first possible option would be to download a dataset from Kaggle.com. There are numerous datasets that have been collected from Spotify, consisting of statistics from many different songs, where each song is an instance in their respective dataset. These options, however, may be limiting to how we can analyze the data. Many of these sets tend to be specialized, such as consisting of strictly unpopular songs or representing only one or two genres of music. Because of these limitations, it would be optimal to use a software package that can collect data for us.

The Spotipy Python package is a useful Python plugin that collects information on Spotify songs by using a URL as input. This could be useful for our project since it allows us to create our own datasets however we wish. We could easily collect a wide variety of songs from all different genres, regardless of popularity. We could choose specific attributes of songs to record as well, such as mode, key, runtime, genre, number of listens, etc. This tool provides us with a large range of dataset options for our project.

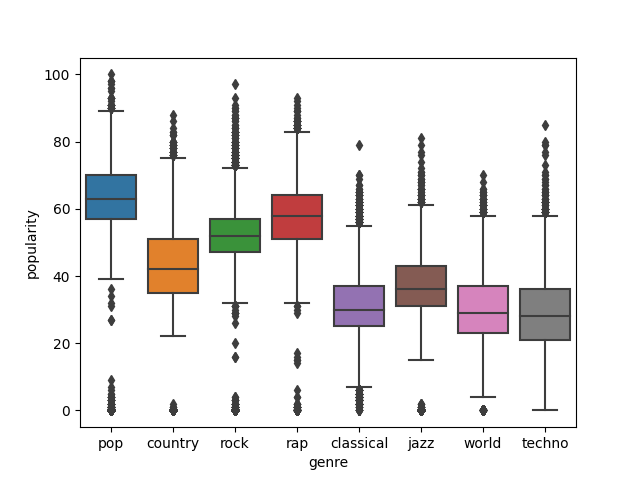
Techniques

After collecting our dataset(s), we would most likely pre-process the data to eliminate noise using means and variances to single out instances that would greatly differ from the rest of the set. Additionally, we would use general statistics of songs and histograms of multiple attributes in order to identify grouping and similarities that we could explore when training a training model. Seeing as a Spotify dataset will have a lot of continuous data, we might even use linear/logistic regression techniques to help with our model. A clustering analysis of different attributes would help us find close correlations that we could exploit when training the data. While developing a decision tree would be ideal, it may not be feasible with such an expansive dataset that has so much continuous data. During the early parts of this project, we will attempt to see how and if we can model a decision tree.

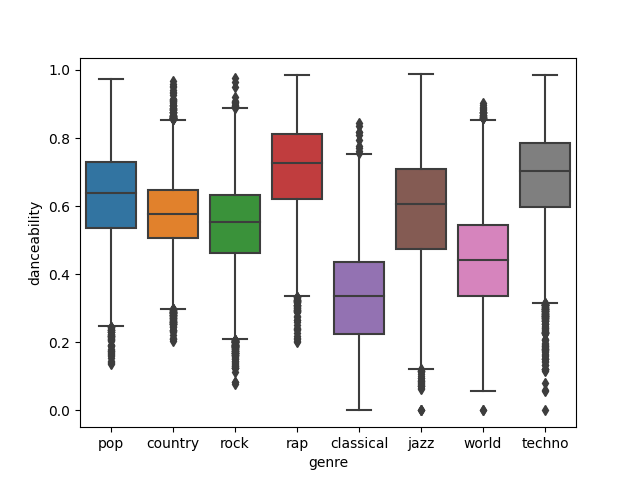
**Data Collection**

**Data Preprocessing**

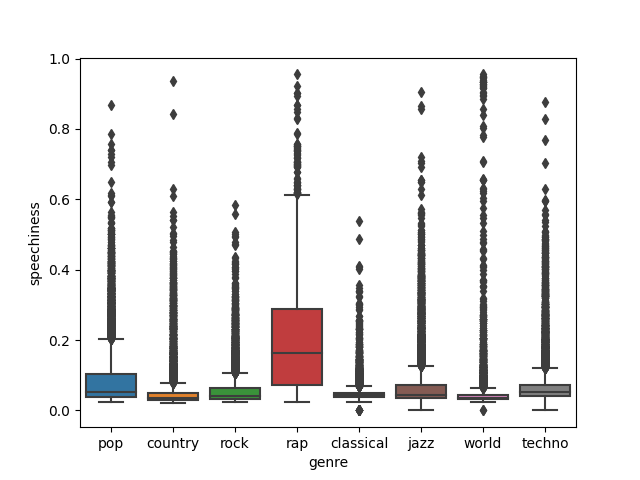
The analysis started with the eight different datasets with about 10,000 songs each. Then, they were all compiled into one large dataset while removing the redundant information, such as disc number, id, track number, album type, album name, is playable, and linked from. Next, various histograms and boxplots were created to try to find the most interesting information about the data. For instance, we found that pop music tends to be the most popular out of the eight chosen genres as shown below:



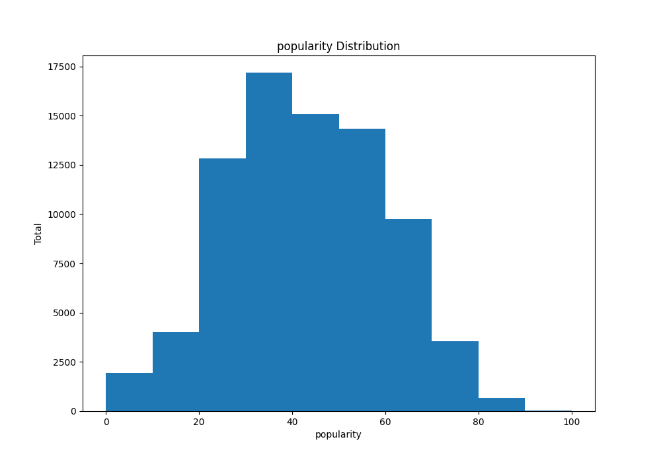
As expected, the least popular would be classical and techno music. So far, the data that was collected seems to be indicative of current music. Additionally we found that the danceable music is rap, pop, jazz, and techno:



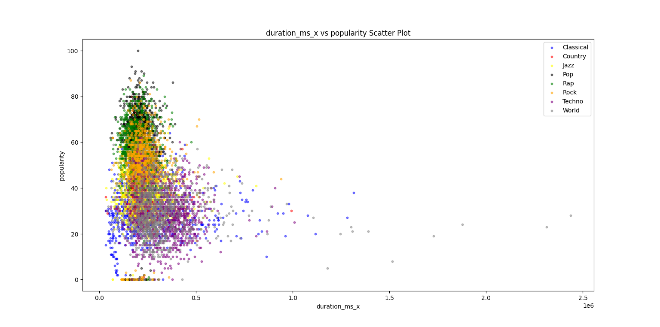
Another important correlation we found (albeit a rather obvious one) was that rap music tends to be the most vocal type, while classical music is the least as shown below:



These boxplots show a distinctness between these genres of music, indicating that it may be possible to predict the genres of Spotify tracks. Another important piece of information is that for all eight genres, popularity is most dense in the 30-60 range (on a scale of 0-100) while still having a good amount of variability:

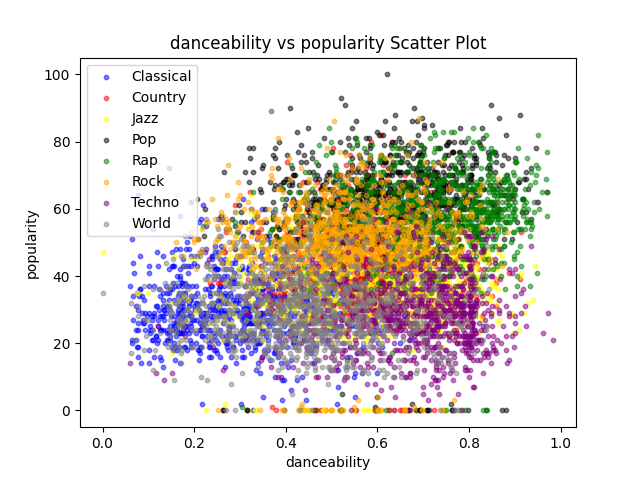
This indicates that our sample data may be useful enough to predict popularity based on all other attributes.

After plotting these initial graphs, a Pearson correlation matrix was created using the attributes of the songs. Some of the highest correlations were between: (popularity, explicit), (popularity, instrumentalness), (valence, danceability), (acousticness, loudness), (genre, duration), (genre, energy), (genre, loudness), and (genre, acousticness). After this analysis, a stratified sample of about 8,000 songs was taken from the dataset (grouped by popularity). This was used to generate scatter plots in order to find any correlation between popularity and genre. One of the correlations found was techno tends to have longer duration and lower popularity than rock music:

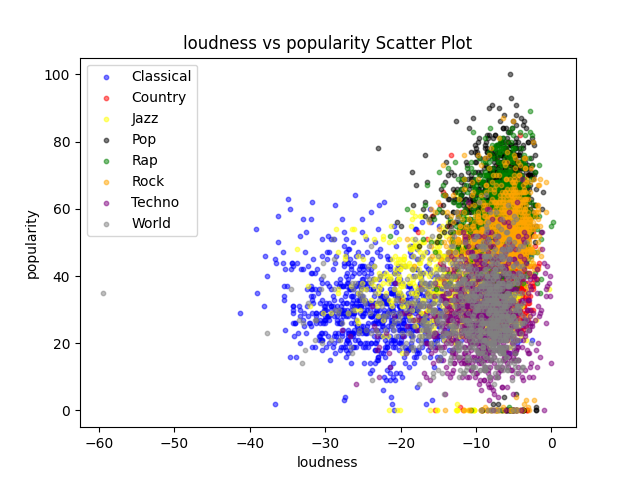


(Make sure to zoom in for clarity)

It can also be noticed that classical has much lower popularity and danceability than rap music:



Another good point to notice is that loudness is more closely related to genre than popularity. In the plot below, it is shown that even though pop, rock, rap, techno and world music generally have the same loudness, but techno and world music have much lower popularity:

With many more scatter plots that visualize similarity and dissimilarity between popularity, genres, and their attributes, we will be able to train and run models using all of this important information. These scatter plots show once more that our data is concrete and that the genres are distinct enough to develop training models with good accuracies that may predict genre or popularity.