CMSC 491 Data Science Project Report

Rohan Ahuja, Daniyal Khan

UMBC

Abstract

The goal of this project was to analyze Spotify data and use various techniques to predict the popularity score of a song, and then popularity growth over time of a song. A dataset of Spotify Audio Features from April 2019, which consisted of data from the official Spotify Web API for 130,000 songs, was used to predict the popularity of a song. Then two versions of the Spotify Audio Features dataset, one from April 2019 and the other from November 2018, were integrated together to predict growth or decline of a song popularity over time.

CMSC 491 Data Science Project Report

# Datasets

The dataset used for the project was Spotify Audio Features, which contains audio features for 130,000 tracks collected from the official Spotify Web API. The data within the dataset was obtained through Spotify for Developers, which gives users access to the extensive catalog of Spotify data.

Two versions of this dataset were used. The first version consists of Spotify song data from April 2019, and the second version consists of Spotify song data from November 2018. The dataset from April 2019 was used to predict song popularity, and the dataset from November 2018 was integrated with the April 2019 dataset to predict growth or decline of song popularity over time.

The features in the dataset consist of the artist name, track id, track name, and numerous audio features of the song. The following is a description taken directly from Spotify of each audio feature.

**Acousticness:**

A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

**Danceability:**

Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

**Duration\_ms:**

It is the duration of the track in milliseconds. We might want to convert this to seconds later.

**Energy:**

Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. Loudness, timbre, onset rate and general entropy are some of the features contributing to this.

**Instrumentalness**:

Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.

**Key:**

The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. If there was no key detected, a value of -1 was assigned.

**Liveness:**

Detects whether the song was played live by looking for audience in the background. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

**Loudness:**

The overall loudness of a track in decibels (dB), averaged across the entire track. Loudness directly relates to amplitude. Values typical range between -60 and 0 db.

**Mode:**

Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

**Speechiness:**

Speechiness detects the presence of spoken words in a track. Values > 0.66 indicate that the song pretty much contains spoken words everywhere. Values < 0.33 indicate that the track is mostly music and represent non-speech like sounds.

**Tempo:**

The overall estimated tempo of a track in beats per minute (BPM). Tempo is the speed or pace of a given piece and derives directly from the average beat duration.

**Time\_Signature:**

The time signature is a notational convention used in Western musical notation to specify how many beats are contained in each measure, and which note value is equivalent to a beat.

**Valence:**

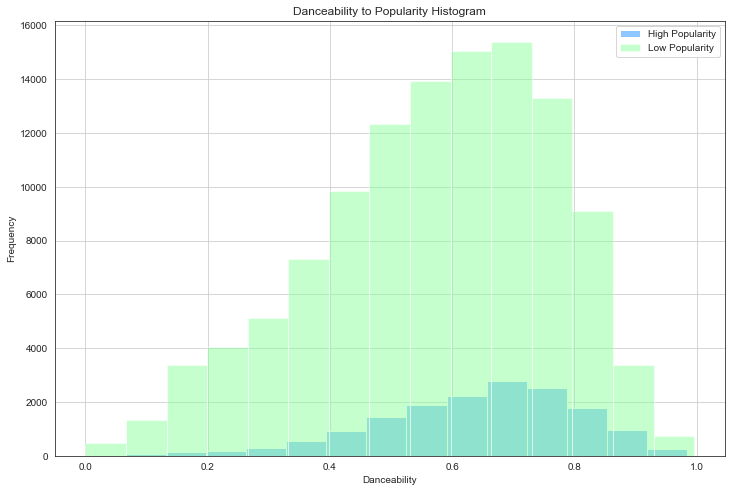
A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy), while tracks with low valence sound more negative (eg. sad)

**Popularity:**

A popularity score between 0 to 100, describing how popular a track is.

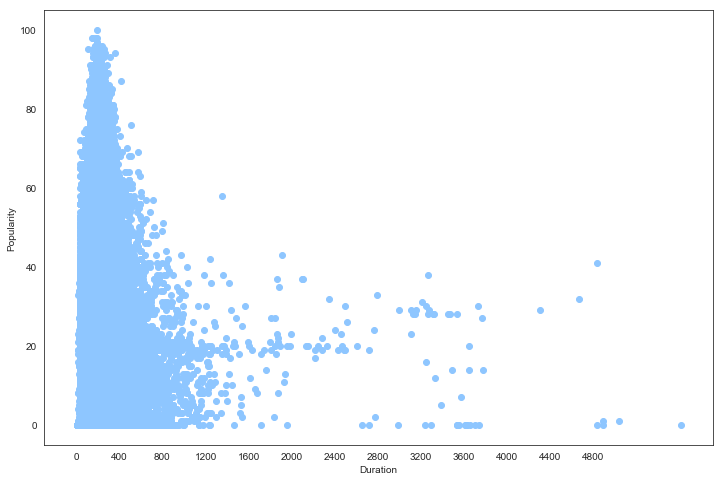
# Analysis of Datasets

The first dataset analyzed was the Spotify Audio Features from April 2019. Popularity was examined to see how it related to the other audio features. First, popularity was related to danceability in the form of a histogram.



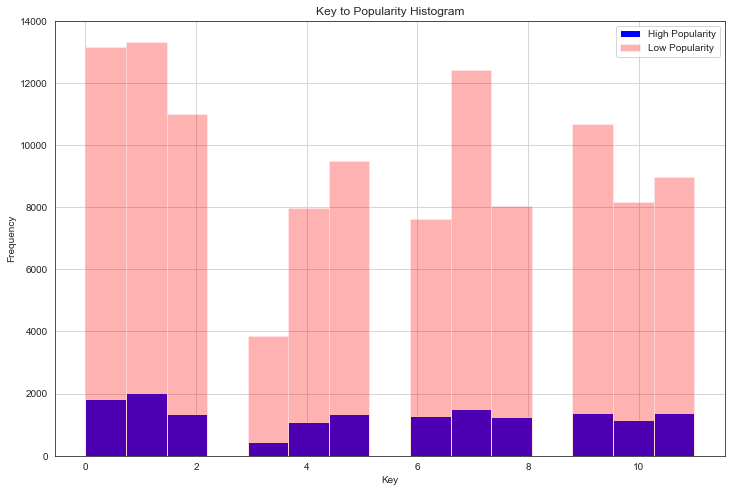
Songs were categorized to have high popularity if their popularity score was greater than 50, and low popularity if their popularity score was less than or equal to 50. There was not a clear correlation between popularity and danceability, but the number of popular songs was greatest when they had a danceability of around 0.7.

Next, popularity vs duration of a song was observed in the form of a scatter graph.



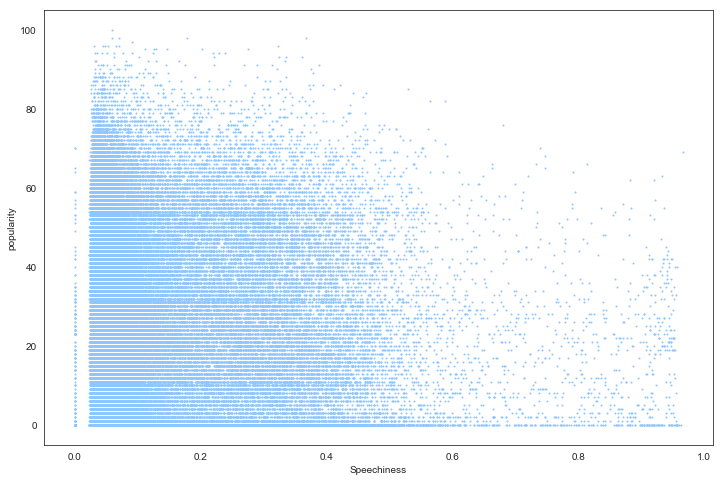
Songs which lasted between 200 to 300 seconds (3.5 to 5 minutes) were the most popular. Songs that were much less or much greater than that range were much lower in popularity. However, there was a sudden surge around the 3000 to 4000 second mark, which could be podcasts.

The key of a song could influence the song’s popularity, so a histogram was created for it. Songs were categorized to have high popularity if their popularity score was greater than 50, and low popularity if their popularity score was less than or equal to 50.



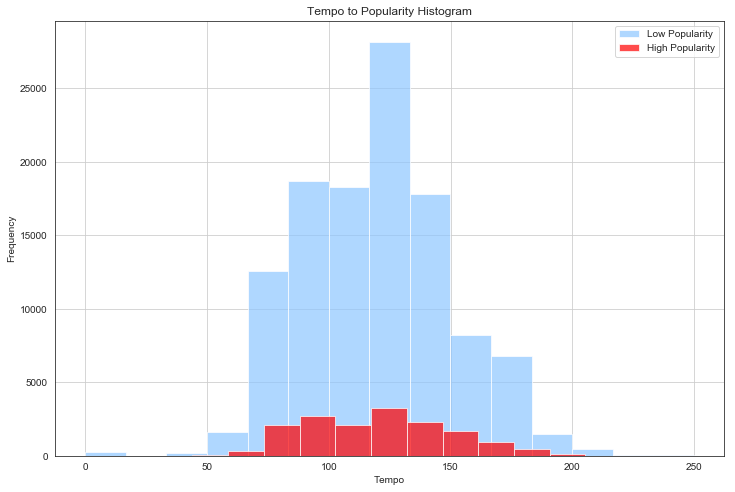
Unfortunately, nothing was obtained from the histogram since no clear correlation could be observed.

Popularity was then compared with the Speechiness of a song through a scatter plot.



Extremely high speechiness was not seen to be very popular. For the most popular songs, a speechiness around 0.10 appeared to be a good amount.

Next, tempo and popularity were analyzed through a histogram. Songs were categorized to have high popularity if their popularity score was greater than 50, and low popularity if their popularity score was less than or equal to 50.



From the histogram, a clear correlation can be deduced, that a tempo range of around 70-160 results in more popularity.

Modality was then compared to popularity