Project Proposal: What Makes a Banger a Banger?

Ben Carew & Simone Garnero

November 2023

1 Data Sources

For this project, we will utilise a data-set simply named "30000 Spotify Songs", available on the data science website Kaggle.com at the following url: `https://www.kaggle.com/data-sets/joebeachcapital/30000-spotify-songs/`. The data-set contains information on more than 30000 of the most popular songs currently available on the well-known music streaming platform, Spotify. Specifically, the data file contains roughly 5000 songs from each of the 6 most popular genres: EDM, Latin, Pop, R&B, Rap, and Rock.

The information provided on each song includes both standard meta-data such as title, artist, album, release date, genre, sub-genre and popularity, as well as information provided by Spotify that attempts to quantify the somewhat more subjective and "fuzzy" musical characteristics of the song. These include the song's "danceability", "acousticness", "instrumentalness", "liveness" and "speechiness", to name a few.

The numerical quantification of these musical characteristics allows us to statistically analyse them and seek interesting correlations, such as which characteristics make music more "danceable", or what musical trends have occurred in popular songs across the decades, to name a few examples.

The data-set has already been pre-processed such that any further pre-processing required is minimal. Spotify is a data-driven company that relies on information like this for it's various machine learning algorithms that recommend songs to users, generate playlists, or predict what song a user wants to hear next. As such, the data is clean and well-maintained, simplifying the pre-processing procedure.

2 Project goals

The aim of this project is to systematically investigate the dynamics of musical evolution through the analysis of quantitative musical features provided by the Spotify data-set. Our objectives include observing temporal variations in musical characteristics and assessing the changing relevance of specific attributes for a song's success over the years and decades. To achieve these goals, our analysis will consist of two main modes: firstly, the use of standard statistical techniques to find correlations in the data; and secondly, the use of machine-learning algorithms such as clustering in order to produce productive algorithms of song meta-data.

In the first category, we aim to study the musical characteristics which share the highest correlation, such as the impact of song volume and tempo on overall energy, or the "valence" (how happy or sad a song sounds) on it's quantified danceability. Another important goal of this first part of the analysis is to find which characteristics are the strongest predictors of song popularity, in order to answer the titular question, ”what makes a banger a banger?”, and discuss trends in this result over time. Finally, our statistical analysis will focus on determining which songs are the most similar to one another in terms of the given quantitative musical characteristics, in order to determine how well genre labels do at grouping together sonically similar music.

The second phase of our analysis will involve the training of unsupervised learning algorithms in order to build models that can predict a song's popularity, it's genre, and the year/decade it was released. Utilising clustering techniques, we aim to train the computer to identify a song's genre (and sub-genre) based on parameters such as volume, danceability, and acousticness, examining whether the results reflect traditional musical genres or give rise to new categories. We also want to construct a predictive model for a song's popularity based on its musical features in order to explore key elements of mainstream success.

This data-set provides class labels for genre, sub-genre and decade, so we will be able to quantify the efficacy of the machine learning algorithms with both internal and external indexes. This will allow us to critically analyse the results and optimise our models for maximum predictive power and confidence. In summary, we aim to conduct an in-depth study into the complex relationships that define the evolution and success of music. We hope to find statistically significant results that allow us to quantitatively understand music quality, which is often considered to be a purely qualitative and subjective metric.

3 Project timeline

Since the time available for the project is 4 weeks, here is a timeline.

- **Week 1:** Exploratory Data Analysis: explore the data-set, identify patterns, correlations, and outliers. Gain insights into the distribution and characteristics of musical features and study the evolution of music over time.
- **Week 2:** Statistical Analysis: investigate and analyse correlations between musical features like loudness, energy, valence, and danceability, and study the similarities between songs in terms of the given quantitative musical characteristics.
- **Week 3:** Model Development: implement clustering algorithms to categorise songs into genre, sub-genre, and decade, and experiment with model parameters to optimise the model's predicitive power, using external and internal indices to quantify the model's success.
- **Week 4:** Report: Finalise the preceding tasks, extract meaningful knowledge from the results of our analysis, and present the information in a report.