Final Project: Spotify Analysis

Luke Syverson

May 17, 2023

Introduction

This research aims to holistically and categorically understand how my music taste (as embodied by my Spotify usage data) can be measured, qualified, and compared within its distinct sets of preferences. I plan to implement the findings into my descriptive musical vocabulary to better relate to other listeners, find new music within my taste, and expand my tastes into new genres.
Principles of Data Science will give me a new opportunity to objectively address my listening habits with tools for quantification and qualification. The ability to measure a song's behavior and my subjective appreciation of it accurately and with a large sample of a year's data should generate novel insights for my taste and the types of listeners I may represent.

Research Questions

Below is a list of queries that this analysis may resolve :
* What song attributes distinguish my favorite songs from the rest?
* What qualities generally comprise my favorite songs?
+ How do these qualities vary across genres?
+ How might I weigh and measure my enjoyment of a song saved in my library?
* What determines the 'replayability' of a song for me?
* How do my tastes compare to my peers?
+ Do I perceive certain aspects of music differently from others?
+ What deviations does my taste have against the population?
+ Can I quantify the degree of 'indie' my taste in music trends?
* What new genres might I enjoy based on my current taste?

Data

While much of Spotify's algorithmic data is proprietary, a feature is provided where a user may download all of their demographic and transactional data. For this analysis, I selected the last year of streaming history, including all songs played, my custom playlists, and saved songs (my library). The scope of this analysis renders most of the demographic information irrelevant, meaning I'll just be analyzing the playlists and streaming history. The data are included in a few .json files sent in a package from Spotify via email, which I will need to transform and then join together.
The Spotify API provides specific information about songs not included in the original data export. I'll need to retrieve the data for the subsets of songs I'm analyzing the characteristics of before conducting a qualifying analysis; however, a simple analysis of demographics like artist preferences and song playback frequency allows for some rudimentary exploration without configuring API access.

Most of the research questions relate to my music taste, but the remainder addresses the general population compared to my 'sample.' While it's certainly a valid topic for investigation, preliminary research has revealed a deep complexity regarding the Spotify API and methodology to scrape popular songs from ever-changing standards. For the sake of quality, I've elected to forgo the additional labor required to attain popular song data and restrict the analysis to my data; this reduces the pool of research to trends and patterns within my behavior.

Approach

The initial priority is to compile the draft versions of the musical-qualities dataset for the subset of songs I would like to begin analysis on. I'll begin by using my 'Favorites' playlist, which has 342 songs and a duration of just under 24 hours, providing a range of genres and a large data set. I started using Spotify around 2018, and this playlist was created in January 2019. The 'Favorites' playlist within the raw 'Playlist' .json is clean enough to filter before identifying keys to join the Spotify API dataset. Each track has a name, artist name, album name, and 'URI,' which functions as a key to the Spotify API.

The API dataset will be compiled on-demand per the analysis of a given subset, per the complexity and potential size of song-specific data. I'll have to decide on a methodology within R and utilize the 'Spotifyr' package to create the new dataset.

The portion of API data within the analysis scope is called 'audio-features,' referred to in this paper as song qualities or behavior, and is detailed in the Spotify for Developers documentation. All the non-metadata variables are valid candidates for analysis.

Later, I may perform an outer join with some portion of my streaming history to my library (saved, but not favorite songs) to compare more generally what I listen to vs. what I save; this depends on the suitability of the data within the 'Favorites' playlist to address the relevant research questions.

After the data transformation is complete, I'll need to continue researching the API parameters and their meanings, specifically, the significance of their ranges and actual musical parallels. Understanding these parameters is critical to interpreting the song-feature analysis of the data. I expect the parameter-oriented analysis to be the most intriguing and pattern-describing within my listening activity.

Dataset Adequacy

The datasets utilized, while cleanly formatted .json files pose some experimental challenges to the research. The preliminary disclaimer for the entirety of the analysis is the inherently subjective nature of quantification and appreciation of musical qualities. As a practiced and active musician, my exposure to and participation in music influences the perception of all kinds of qualities of songs I hear. My perception, and therefore my interpretation of an objective rating, will innately carry with it a bias that defies simple quantification. Subsequently, the results of this analysis may be skewed by my lack of awareness of a foundational bias I may harbor regarding musical perception.

Ideally, statistical methods will prove an apt foil to this subjectivity, and the results may be generalized to some extent to characterize the type of listener that I am. I have confidence in the presumably-unbiased algorithmic generation of song parameters from the Spotify API, which I anticipate will serve as a grounding influence to analyze my preferences.

On the practical side, a significant consideration is the change of my taste from Jan 2019 (the creation date of my 'Favorites' playlist) to the last year of data (May 2022: May 2023), which could induce noise into the final predictions given my natural change in taste over time. The data may require additional subsetting to address time-series discrepancies or an additional weighting method (akin to forecasting) to retain value from older, possibly less-relevant taste data.

Requisite Packages

In short, the currently required packages would be:
* Spotifyr (API calls)
* dplyr (data munging)
* jsonlite (.json parsing)
* ggplot (exploratory analysis)

Plot & Table Requirements

Understanding the ordinal qualities of song data related to their position within various subsets would require distribution and density analysis; I'm partial to histograms and violin plots for these purposes. Identifying possible relationships and correlations requires using multi-axis charts, which may be serviced by scatter and violin plots. Creating an array of song feature charts should guide further analysis with primitive insights while warning of possible regression assumption considerations (non-parametric distributions, heteroskedasticity, multicollinearity, etc.)

Future Questions

I'll need to assess the effectiveness of the 'Favorites' playlist in addressing research questions to determine if additional data are needed for the same or other, more suitable hypotheses. I'm not sure how the Spotify song features will relate to each other and, consequently, how the trends will describe the qualities of each song within different genres. I may need to split up datasets according to genres, especially considering the differences in attributes between some, i.e., 'Extreme Death Metal' and 'Smooth Jazz'; it's also possible that the Spotify measurements are neutral enough to function well in generalized data sets. I'll need to assess the performance of future models with these data limitations in mind.
Another unknown is the time-series aspect of taste: assuming that taste is an evolving concept for the individual, analyzing older data might introduce irreparable noise into predictive models targeting current taste; alternatively, it could be possible that older data is crucial to quality predictions of present data. I'll have to test multiple sets to attain the greatest accuracy, but I plan to err on the side of caution and exclude old data if noise can't be mitigated.