Weather-Activity-Spotify Trends

STAT 5010 Final Project Proposal

Spencer Zeigler

18 Mar 2022

I will be using my personal Spotify data, personal FitBit data, and weather data for Boulder, CO from NOAA to explore relationships and/or differences between my personal music habits, audio features, weather, and daily activity amounts. These data are observational and cover a mostly continuous time period from 2016-current. Depending on the continuity of the data, I might just look at the past 1-3 years, but I hope to examine all 5 years. There are so many questions I could answer, and once I have the main data set in hand (by April 2nd), it will be easier to narrow down the possibilities. But for now, I will introduce each data set and the variables of interest, and then discuss methods and possible questions I could tackle.

My Spotify data set comes in a JavaScript Object Notation (json) format which can be read into R using `jsonlite`, which automatically transforms it into a dataframe. Data cleaning and summarizing using the `tidyverse` and more specifically, the `lubridate` package will take place in order to get the data into a usable format. For example, I might sum up the minutes listened over a day, or take the mean score of the certain audio features for all songs listened to in a day. Using the `spotifyr` package, I will be able to get the audio features for each song I have streamed: acousticness, danceability, duration, energy, instrumentalness, key, liveliness, loudness, mode, speechiness, tempo, time signature, valence, and popularity. These scores might not all be on the same scale, so I will have to be careful to normalize them. I expect to do a fair amount of exploratory data analysis to see which questions are the most interesting and to see if the data meets the assumptions of linear regression and/or time series analysis.

The second data set I intend to work with is my FitBit data, specifically my activity and step counts for each day. This data is already in a rectangular format, but I will have to deal with `NA` rows, since there are days I forgot to wear my watch or it was broken. The third data set is a simple day-by-day temperature (high and low) and precipitation (amount) for Boulder from NOAA. Again, this data set has `NA` values, so I will have to clean the data a little bit. My final goal is to have these three data sets combined so that relationships between music habits (time spent listening or audio features) can be examined against trends like activity or weather. To this end, I expect a significant potion of my project to be focused on exploratory analysis and data wrangling.

Without my Spotify data in hand, I have brainstormed some potential ideas for things to explore during my first phase of exploratory analysis:

1. Is there a correlation between the weather (temperature) and the amount of music I listen to?
2. Does the valence (mood) score correlate with precipitation or temperature?
3. Does the loudness/energy score correlate with the amount of activity?
4. Does any audio feature correlate with activity better than others?
5. Do deviations from my average music taste* increase with deviations in sleep, exercise, or temperature?
   *music taste defined by the mean, normalized score of all audio features.
6. Does the popularity score correlate with other audio features?
7. Is there a seasonal pattern to the amount of music I listen to? (basic time series analysis)

Although no literature or scientific theory that I know of will dictate how my personal music habits might shift with the weather or activity, I might expect more 'upbeat' music on warm/high activity days and more

'subdue' music on cold/wet/low activity days. Additionally, I expect to have a problem with multicollinearity with Spotify's audio features. For example, I might expect energy, loudness, and danceability to be highly correlated.

I intend to use data wrangling/cleaning techniques that I have learned in class and on my own (eg. `dplyr`, `tidyr`, `purrr`, `forcats`, `VIM` etc.). For exploratory data analysis, I will use the techniques discussed in class to assess normality, deal with missing values, and explore basic patterns and trends of interest through visualizations and summary tables. I hope to explore some basic time-series analysis techniques on my own since all of this data is collected over time (e.g. question 7). If that does not work out, I will rely on linear (and hopefully!) non-linear regression techniques to answer questions like 3, 4, 5. For example, a regression to answer #3 might look like: `lm(daily_steps ~ average_daily_energy_score)`

It might also be interesting to use ANOVA/ANCOVA to examine if there are differences in the amount of music I listen to on sunny vs. rainy days by encoding a dummy variable for the weather. There are many possible directions to go in once I receive my Spotify data set and do some exploratory data analysis, but I intend for this project to answer questions about potential relationships or differences between the types/amount of music I listen to and other variables, like weather and activity.