Eric Lin

In my data exploration, I chose a dataset containing information on over 20,000 songs that can be found on YouTube and the Spotify music streaming app. This topic initially piqued my interest as most everyone likes music, and overall, looking into song analytics may help to prove or disprove my current beliefs on mainstream music.

The dataset contained information in the form of 26 unique variables for each song. Some of the variables included were: song name, artists, views, likes, comments, Spotify streams, album type, and duration. The more interesting variables that I used were: speechiness and valence. All three of these variables are measurements of musicality and represented by numbers ranging from 0.0 to 1.0.

My first test was to check the relationship between valence and Spotify streams. According to the data card, valence is defined as the musical positiveness. Happier tracks have higher valence while more depressing tracks have lower valence. Therefore, my initial hypothesis was the higher the valence, the higher the streams. People use music as a safe space and time to relax, so I believed that they would more likely listen to things that make them feel uplifted. Going into all of my tests, I first subsetted the dataset so I could focus on what variables I needed the most. In this case, I subsetted the dataset to only include the music title, valence measurement, and view count. I noticed the stream counts for a few songs were labeled as “NA”, so I made sure to remove those from the testing as well. I separated the valence into high, medium, and low categories and then used tapply() to compare the average streams. To my surprise, the lower valence songs appeared to have more streams than the higher level valence songs. This result however does not necessarily mean that people like to listen to depressing songs. The data card described valence similarly to major vs minor key music. As a musician, I know minor songs are not necessarily sad songs. It’s possible that many people just prefer a minor key song due to other factors that are not related to feeling happy or sad. Either way, the results were unexpected.

For my second test, I wanted to see the relationship between speechiness and streams. Speechiness is predefined as the presence of words in a recording. The higher the speechiness value, the more exclusively speech-like the recording is. The data card defined speechiness values above 0.66 as tracks almost exclusively words. Values between 0.33 and 0.66 are defined as songs with both music and lyrics, while values below 0.33 are likely more instrumental. Mainstream music tends to focus on pop or rap songs which typically emphasize lyrics; therefore, my hypothesis was that a mid-level of speechiness between 0.33 and 0.66 would likely garner more streams. I first cleaned and subsetted my data as previously mentioned. I noticed some high outliers from previous tests, so I removed them. I then plotted speechiness vs streams on a scatter plot to observe any clustering. The results were opposite to my hypothesis as the stream count for speechiness levels below 0.33 were higher than anything above 0.33. This goes to show that fewer lyric to music ratio songs tend to be more popular on Spotify. I believe this is explainable as most popular music tends to have long sections of just instrumental before any lyrics. In addition, people listen to acoustic study music on Spotify, so that could contribute to the higher stream counts.

For my third test, I wanted to see the relationship between duration and album type. My initial hypothesis was that different album types should have different song lengths. For example, songs in an album should be shorter because albums are a collection of songs made for one another. On the other hand, songs released as singles should typically be longer than their album counterparts. Therefore, album songs should be shorter while singles should be longer. I subsetted my code before separating duration into categorical data. From there, I used table() to compare. Unexpectedly, my data showed that while it is true that albums produced the shortest songs, they also produced the longest songs compared to singles. Therefore, it is not conclusive to say that singles are typically longer than album songs.