Bryan Emer

DSC 500

Final paper

Spotify Exploratory Data Analysis

In my final project I used Spotify’s data set to try and determine what musical features, if any, have a relationship to a track’s overall popularity. I was able to identify the correlations between these variables and popularity and create a model using these variables. However, there was not a strong enough relationship to create a model that could confidently predict a track’s popularity given just these variables. One issue that makes this analysis difficult is that a song’s popularity is determined by much more than just the music itself, and this is harder to measure in data.

I believe that the key to creating a more accurate model involves how the tracks being analyzed are selected. It is possible that the model would be a better predictor of a tracks popularity if we begin with only artists that are above a certain threshold of popularity. We could select, for example, only the tracks from the top 1000 most popular artists on a platform. While this would be a significantly smaller data set than the one used in our analysis, it would still be tens of thousands of songs and perhaps would provide a more relevant analysis of just the actual musical features of a track because there is already a threshold of popularity for the artists.

I think one assumption I made that was incorrect was that using a larger dataset would necessarily be better for analysis than a smaller one. While the dataset used contained lots of tracks, as mentioned previously it may have been possible to create a better model using a smaller dataset of only popular artists. Perhaps a good approach would be to start with an even smaller dataset, for example just the top 100 artists, and then work your way up with more artists to see how it impacts the model.

One challenge I faced was when running the hypothesis tests, as the results for many of them produced odd results with a p-value of zero which was not helpful. From looking at the graphs I did not expect to find significant correlations with these variables, but only using the chi-squared test was I able to see any correlation at all. Overall I thought that this was an interesting process and would enjoy further exploring this data and trying to find a better model to predict song popularity.