Implementation Summary

For this section of the assignment, the work was again divided fairly evenly. After the presentation last week we wanted to answer some of the questions that we were unsure of. The most important of these questions was exactly what question we wanted to answer and whether the data we had already collected was sufficient to answer this question. Together we decided that it would be best to try predict whether a song would be “popular” or “unpopular” based on the audio characteristics of each song. Trystan investigated our current data as there were a few strange occurrences in the data that we were unsure of. He found that the spike towards the end of the data that could be seen in a few of the graphs was indeed caused by Christmas songs taking many of the top places on the charts for a few days and falling off quickly after the 25th. However, we continued to run into issues with our data set from the Spotify top 200 for a few reasons. Firstly, there was a lot of data in each data point that we did not need to use such as number of plays, making the data more difficult and time consuming to process. Additionally, there was data from every day. Having data from every day made processing data over longer periods very difficult. Daily data also meant that relatively short events, such as Christmas, also were able to make a noticeable visual impact on the data.

In order to help remedy these problems, Ethan collected some additional but similar data from Billboard top 100 instead of from the Spotify top 200. This new data is published only weekly and has less unneeded information like number of plays. The data was then cleaned and stored in pandas dataframes which took a few hours of processing but little additional effort. This data for this section is stored in a set of csvs in the Deliverables/Deliv\_2 directory of our repository. Ethan and David processed the new data and used the Spotify API wrapper spotipy to find audio metrics about each song from EchoNest. The data was then used to create graphs showing the change in these metrics overtime. We all then met as a group to discuss this new data and to run some statistical tests to attempt to gain new insights into the data.