Capstone 1 Data Wrangling Report

This is the data wrangling report for my first Springboard capstone project for the data science career track. This project is creating a model that predicts what decade a song is from using lyrics of songs from that decade. The data for this project is relatively simple in terms of composition. The dataset contains the 100 songs from the Billboard Top 100 for each year from 1965-2015, with fields containing the rank in the Billboard Top 100, name of the song, name of the artist, the year the song was from for the Billboard Top 100, and the lyrics of the song. This data itself is not particularly messy, as there are not many ways for this data to be messy. The most obvious place this data needs to be cleaned up is addressing songs that are missing lyrics. Another obvious error in the data is that songs seem to start with the artists' name instead of the proper lyrics. Some of the artist names for songs are a bit messy, because of the way many pop songs work, there are many artist titles that have and when two artists work together, as well as featuring. The last error that seems to be in the data enough to mention is that there are songs that include structure words, such as "Verse 1" or "Chorus", which is an interesting problem since some songs will include those words purposefully.

It was cumbersome to even begin finding the songs which were missing lyrics. The songs missing lyrics were not filled with an empty string, they were filled with an empty space or two empty spaces, so replacing the blanks with NaN values took a bit of trial and error. In total there are 234 songs missing lyrics in the dataset. 1971 and 1972 are missing the most data, missing 13 and 12 songs respectively. Most other years are missing an average of five songs. Songs that were missing lyrics are going to be dropped from the dataset, primarily because of the scope of this project and it may be too difficult to scrape for the missing lyrics considering they were already missing from the dataset. Songs that are instrumental are also going to be dropped from the dataset, considering since they do not have lyrics, they cannot contribute to the model.

When it came to the obvious lyrical errors that start with the artists name, it took a bit of time to fix those errors. After removing the erroneous artists names from the beginning of the lyrics, many of the songs had ,more obvious nonsense words like 'miscellaneous' and then the artist name again. These songs were cleaned as best they could be until it seemed the songs that fell into this category seemed to start with only the beginning of the songs lyrics. My Mentor and I also decided to remove the words verse and chorus from the data, since it was obviously erroneous in many songs as just structure words as artifacts from scraping. This only addressed songs that had obvious errors that were easily identifiable, so there may be songs that are still erroneous, however finding many of these other errors seems like it would be more trouble than it is worth.

Finally, the last thing to be done to the data is to remove the source column since it will have no relevance to the model. The source column in the data regarded where the scraped lyrics

came from, and since that will not be relevant at any point in this analysis, so that column will simply be removed from the data.