Marielle Billig
Data Science Midterm
March 9, 2017

LINK TO CODE REPO: https://github.com/mbillig

The premise of this project revolves around combining data about the characteristics of songs, with data about their popularity. The first and primary dataset being analyzed has distinguishing data about the songs, such as title and artist, date data, such as year, decade and quarter, as well as many quantitative musical measurements regarding harmonic and timbral aspects of the songs. The harmonic features have to do with which chords are used during the song, whereas the timbral features have to do with the sound of the song. For example, a song might use dominant 7^{th} chords, which is a harmonic variable, and be energetic, percussive and have male vocals, which would all be timbral. This data was originally used in a paper that analyzed how popular music in America has changed between the 1960s and the 2010s [3]. One of the research questions being pursued in this project is to see if these musical characteristics can be used to predict which decade a song came from. Humans can fairly easily determine if a pop song is from 1964 or 2004, so it seems reasonable that the decade of a song's origin could be predicted.

The second aspect of this project involves incorporating data from another source. Although all the songs that I have characteristic data on are from the top 100 billboard charts, they were not all equally successful on the charts. Therefore, I scraped the data from the Ultimate Music Database site, which had data about every song that has been on the top 100 billboard chart such as what it's position was during that week, the date it entered the list, and how many weeks it had been on the list [1]. The scraper I used was lightly modified from a web scraper I found on github [2]. After converting the file to a csv, I initially viewed the output in excel to determine the state

of the data. A lot of wrangling was required to fix up the output of the scrapper. It seems that whenever there was a supporting artist, the scrapper interpreted that as the next field, which should have been an integer, thus offsetting the rest feature by one or more cells. Additionally, text that was bolded on the Ultimate Music Database website came through with extra b characters, apostrophes, and quotation marks. Some of these errors I fixed directly in excel, but others I fixed after I read the file into a python notebook.

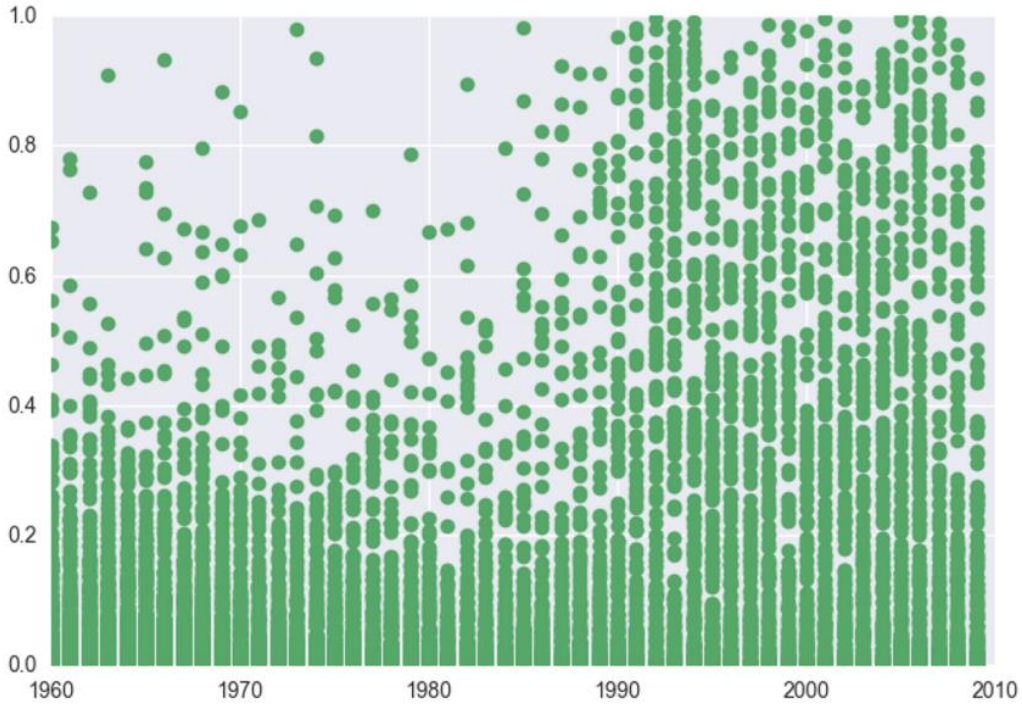
I did not realize this when I was first tidying up the data but the biggest issue with the data from Ultimate Music Database was that each song had as many entries in the dataset as weeks that it was on the chart. For example, the song *(Don't Fear) The Reaper* was on the top 100 billboard charts for 20 weeks, and thus has 20 entries in the table, each with features that sometimes are the same (ie entry date) and others the differ from week to week (ie previous week rank). This was an issue because it was my goal to merge the billboard data with the characteristic data, and the characteristic data had only one entry per song. Thus I decided to combine the data from multiple rows for the same song into one data entry. I decided to compute that for each song, I would like the maximum rank, the minimum rank, the average rank, and the total number of weeks that the song was on the chart.

Another issue that came up regarding merging was the fact that I was using the title of the song as the key to join the tables, and sometimes the title was represented differently in the two data sets. The most obvious case of this was that some songs like *(Don't Fear) The Reaper* have a title, but also have a parenthetical aspect of the title that isn't always included. I resolved this by excising the parts of titles within parentheses after I read in the datasets. Although this greatly reduced the number of data entries that could not be matched, I am considering trying to

incorporate the artist of the song into the key since songs with the same name are currently being treated as the same song.

After the data was cleaned enough to be workable, and merged into a single dataframe, analysis was finally able to be considered. I was hoping there might be some obvious clustering, but after running several k-means cluster analyses with different numbers of cluster, it seemed that this was not the case.

I also attempted to fit linear regression models to the data as well, with limited success. While I was able to implement the model fitting and prediction, the scores were very low. Because of this, I began to wonder if perhaps the data did not actually show the changes that people are easily able to perceive between different decades. To determine if the harmonic and timbral characteristics had any sort of relationship with the year, maximum rank, or weeks on the chart, several scatter plots were generated. Below is an example of the harmonic topic of 'no chords.'



From the above plot, it seems intuitively likely that if a song is high for ‘no chords’ it is more likely from post 1990. Some of the characteristics shown little variation with time but several showed change as time progressed.

Next steps for this project include improving the accuracy of the keys for both tables and potentially smoothing the characteristic data. Furthermore, at this point only 16 characteristics are being considered, but there are actually over 200 more specific features included per song in the characteristic dataset. Therefore, some sort of feature analysis such as PCA will likely be needed soon.

It is also necessary that additional models be considered for prediction. Although little success was seen with the linear regression and the cluster analysis, these were both very rough attempts and could likely be more fine-tuned. It would also be interesting to consider other models such as neural networks or svms.

References

1. Jurceka, M. Ultimate Music Database. Last Update 19 February 2017. DOI: http://www.umdmusic.com/
2. Kling M. umdmusic-downloader. Last Published 3 April 2016. DOI: https://github.com/mwkling/umdmusic-downloader/commits/master
3. Mauch, M., MacCallum R.M., Levy, M., Leroi A.M., The Evolution of Popular Music: USA 1960-2010. in R. Soc. open sci. 2015 2 150081; DOI 10.1098/rsos. Published 6 May 2015