Link to data story notebook:
https://github.com/mbuck86/Capstone1/blob/master/Data%20Story/Data%20Story%20Notebook.ipynb

Capstone 1 Data Story

The data in this project has been incredibly interesting to work with, considering it is a large set of text data scraped from different websites. There was an interesting set of problems to deal with, such as nonsense 'typo' words as well as strange formatting inconsistencies that are mostly unique to representing lyrics; such as using 'x2' to represent a line being repeated twice. The following will be a description of the process I used to generate the initial insights into my data. Code and visualizations that are described in the text can be seen in the Jupyter notebook linked at the top of this document.

The first process that I went through when trying to process my data story was to look at aggregates of all of the data, primarily to confirm the data that was being worked with. I have around 1000 songs to work with from the 70's, 80's, 90's, and 00's to work with, but due to the way the data is structured, cleaning processes, and time, I only have around 450 to work with from the 60's, and around 600 to work with form the 2010's. Most of the top-ranked songs from the dataset have not been removed from the cleaning process, and the three artists that occurred most frequently in the Billboard Top 100 from 1965-2015 are Madonna, Elton John, and Mariah Carey.

After visualizing the big picture of the data, I moved to performing more granular looks at what was in the data. The first steps I took in processing my data were to take the lyrics from the songs and break them up into individual tokens(seperated the lyrics into word by word pieces as opposed to continuous strings). While performing this process, I removed words that were just numbers from the data and removed stopwords(words that tend not to carry information such as 'of' or 'is') using the stopwords dictionary from the nltk library. I then calculated word counts for the data in order to begin to identify trends in word usage across decades. The most common word in the dataset is love easily, with several thousand occurrences over 'know' and 'like' the next most common words. 'Love' is also the most common word for every decade in the dataset excluding the 2000's and 2010's, where it is overtaken by 'like'. I then generated word clouds for the data by decade which can be observed in the notebook. While these are nice informative visualizations, they did not provide much more valuable insight than wordcounts, besides showing a few minor changes by decade, which is valuable, however I feel due to the nature of the similarity of the decades, it is fairly difficult to gleam those differences using word clouds. Next I calculated two-word and three-word n-grams for the data. N-grams are words in sequential order in the data of size n, which is a value you choose when creating the n-grams. Creating the n-grams did bring some interesting information out about the data, such as the fact that 'la la' and 'la la la' are both extremely common n-grams across the data, as well as 'yeah

yeah yeah' . Given additional time and resources these n-grams would be extracted as features and used in the classification algorithm, but on the advice of my mentor, this step is being avoided for this project, primarily due to the number of features this would generate.

After viewing word counts and n-grams from the data, I looked into generating term frequency-inverse document frequency(tf-idf) values for the data. These values are a way to quantify the importance of words in a song across the entire lyrical library that we have. These tf-idf values are going to be the primary source of features for the classification algorithm. For this data story part of my project, I am going to observe 'big picture' values and try to summarize important aspects of the data, while going into more depth into the exploratory data analysis part of this project. On that end, I grouped the tf-idf data by decades and then calculated the mean tf-idf value of each word in the data. The plots for the top ten tf-idf means by decade can be seen in the notebook. These visualizations show potentially promising differences across decades of the words and their usage, as these are the most significantly different plots we have seen in terms of lyrical content by decade. Moving forward, I am going to study the differences between these values more in depth in my exploratory data analysis notebook, to see what quantitative insights may be drawn from these values, as it will be important for these values to be different across decades for them to be useful features for predicting what decade a song is from.