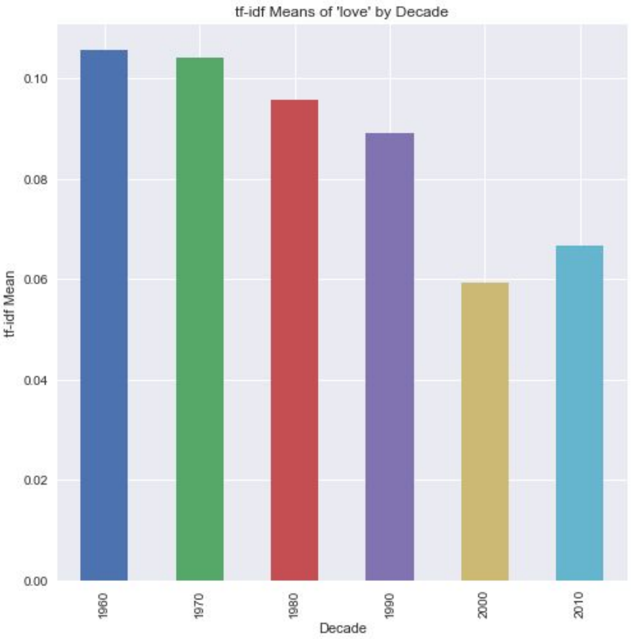
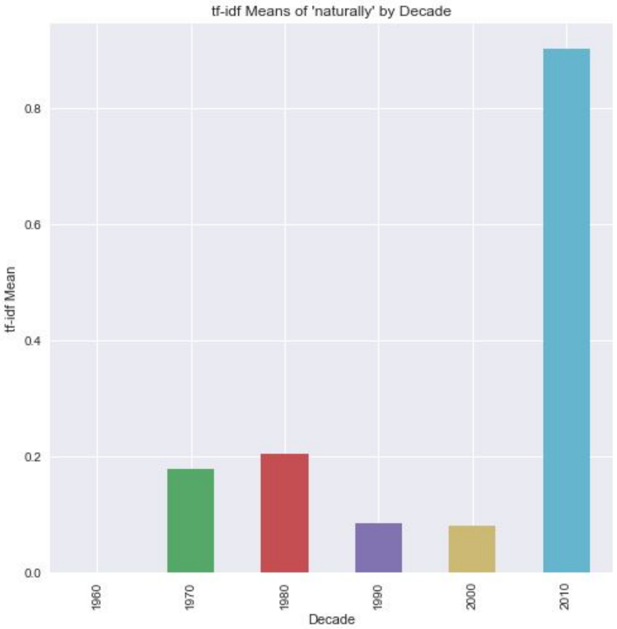
Capstone 1 Exploratory Data Analysis

This exploratory data analysis will primarily involve looking at the tf-idf data that was generated for this project previously. The primary point of this analysis is to hopefully find significant differences in word usage across decades that will provide some insight into the creation of a classification algorithm for this data. I am hoping to discover interesting insights into the variations of this word usage through the analysis I am about to perform. This analysis will be short based on the advice of my mentor, who mentioned that due to the nature of the data that I am using, many of the more traditional statistical analyses were not necessarily useful in terms of this project.



I started with observing the tf-idf means of love across the decades. I was interested in the fact that love seemed to consistently become less popular and important to songs over time, and wanted to see if the difference between the tf-idf mean for the 1960's and the tf-idf mean from the 2010's was significant. I wanted to perform a t-test for the difference of means, so I plotted histograms of the values from their respective decades and found that both of the values were exponentially distributed. Since you cannot perform a t-test with exponentially distributed data, I took the log of each set of data, and observed that the log values of the data did seem to be normally distributed, so I performed a t-test with the log values and found that there was a statistically significant difference between the two means with a p-value of 1.15×10^{-10}.



After performing that experiment, I moved to making another observation because I noticed something strange while creating my Data Story document. 'Naturally' was the word with the highest tf-idf mean in the 2010's, and had not really been present or significant in any other visualizations I had made of the data, so I was interested in seeing what might have caused this to be the case. It turns out that only one song in the data even contains the word naturally in the 2010’s, and it is a song titled ‘Naturally” by Selena Gomez. This was an enlightening quirk in the data to discover, as if this is the case for many of the 3199 words in the data, any classification algorithm that I write may be inundated with noise similar to this for each word.

This analysis brought me insight into the fact that the words might have clear identifiable trends like ‘love’ seems to have, or strange quirks caused by just a single song, like ‘naturally’. I will use this insight as I move forward to doing more in depth work on this data, such as principal component analysis and finally the construction of my classifier.