Methods

I will look into future research later.

As previously stated, we relied on natural language processing methods to perform our analysis. The methods that were most useful for this research were keyword extraction and sentiment analysis. Both methods of analysis provided us with crucial insights, so the details of each seem pertinent to share.

Keyword extraction is largely aimed at identifying the most used words in different texts and utilizing those words to understand common theme or popular topics. [] We used the *tm* library in R to perform keyword extraction on the sonnets. This process involves first prepping the text by removing unwanted punctuation or numbers, eliminating stop words, changing everything to lower case. These transformations are necessary because when working with strings, you not only have to be precise, but you also must be exact. The cleaning process sometimes includes stemming the words, but we opted not to do that. Once the text is clean the idea is to create a matrix containing each word used in the text and the frequency with which it is used. To identify the top ten key words we sorted the words in the matrix by their frequency. (Good place to insert word cloud or key word table)

Once Shakespeare’s keywords were identified, the next step was to identify the keywords for each of the 45 artists in our dataset. This process was similar to that of extracting keywords from the sonnets except we utilized a loop to automate the process for each artist. Once keywords were identified we converted the frequencies for Shakespeare and all other artists into proportions, in order to standardize for comparison. To find the artists who were most similar to Shakespeare based on keywords, we checked for artist who had the highest number of matching keywords. Finally, then we ranked those by who used keywords in similar proportions to Shakespeare by calculating the Euclidian distance between the proportions of each of the matching keywords.