The goal with this project was to explore a **Yelp dataset** from Kaggle that serves as a collection of reviews that the website cultivated during the year of 2013. More information regarding the dataset can be found at the link above.

To my surprise, and benefit, the dataset itself had no missing values that could potentially make the analysis process much harder. Being a review-based dataset, the “text” and “stars” columns caught my eye and gave me an idea to analyze potential relationships. I wanted to approach it from a numerical and text-based point of view, so I created a “WordCount” column for each review. Using this column to plot a set of plots to visualize any potential correlation. The main goal, overall, to understand what constitutes a “good” review and if we can predict for one using a Naïve Bayes model. One major point of note is that there seemed to be more 4- & 5-star ratings in the dataset that could’ve potentially affected the result of the model training. The code itself can be adjusted to train on any combination of the star ratings but I chose to use a randomly sampled mixture of 1- and 5-star ratings since the results were so interesting.