Progress report

To approach exploratory data analysis (EDA), I generated summary statistics and visualized the data using histograms and boxplots. Using summary statistics, measures of center were analyzed as well as the spread of the data. Columns were assessed to ensure data completeness. When looking at the data distribution for the “funny” and “cool” columns, it was elected to drop the columns from analysis due to sparsity of data. Visual and descriptive analysis was undertaken to identify and eliminate outliers. In addition to analyzing the yelp data, I also vectorized the text (review) column and performed EDA on those generated columns.

Initial results of the EDA show that the vectorized data is filled with outliers, which may make modeling difficult. The size of the corpus may also make modeling challenging. It should be noted that there may be other limitations of the data that may make it hard to model high scores. Some models may be more suited for this kind of binary classification than others.

A small set back was noticed when generating summary statistics for the text (review) column. It had had incorrect regex run on it in the data cleaning stage, and needed to be rectified. I fixed the mistake in the data cleaning notebook and regenerated the required data for EDA.

For proposed next steps, three steps will be undertaken. Firstly, I will vectorize the data, using TF-IDF or bag of words. From there, I will model the data using at least three models. Once I have initial model results, I will then select and tune the model with the best outcome.