The first step when trying to create a predictive model is always examining, analyzing, and visualizing each element of the dataset that you are working with. In our case, we are trying to create a model that best automatically grades essays based on important essay-writing components, which we will discuss later. We will be looking at 8 different datasets (taken from Kaggle) that all contain essays with given scores, which we are going to parse through and visualize based on a series of scatter plots.

Because ultimately, we want to figure out a score for the essays, all of the factors that we are examining are plotted against the scores for the essays. For this part, we plotted the word counts against the scores and the vocabulary sizes against the scores for each of the datasets to if there was a pattern between the quantities. In order to calculate the word count for each essay, we used the python split function to split the text into words based on alpha characters. We then counted the length of that. For the vocabulary sizes, we used a count vectorizer that contained an index for each unique word, and counted the length of that.

Typically, one might assume that longer essays equate to higher essays scores because writing more means more effort, more ideas, and possibly a deeper dive into the subject matter. However, teachers do stress the importance of being succinct, and refraining from adding “fluff” and unnecessary details to papers. As a result, from personal experiences we predicted that longer essays would correlate to better scores, but we didn’t know if that was the case after seeing both sides of the argument.

Our logic behind vocabulary sizes affecting the essay score is that individuals who typically read and write more use a wide-variety of words. They have more exposure and understanding of how words are used, and in what context they make sense. An essay that frequently uses exciting, new words keeps the reader engaged, while also giving the writer the opportunity to best express his/her ideas from using words that fit best. Therefore, we would assume that a larger vocabulary list would signify a higher essay score.

Examining the scatterplots, the first thing that sticks out is that the word counts vs. score scatterplots for each of the sets were virtually the same as the vocabulary size vs. score scatterplots for each of the sets. That shows that there might be a relationship between the length of the essay and the different number of words that the writer uses. To rationalize this, one might think that someone who reads and writes more is keener on his/her ideas. That may cause them to write more detailed essays. This also might mean they enjoy writing more than someone who does not read or write as much. As a result, for later explanations of the graph, it would make sense to talk about the trends of both graphs for each set as one, since they are very similar.

From the word counts vs. score scatterplots, we see that in general, there seems to be an upward, positive trend between the essay words counts and the score. For sets 1, 2, 4, 5, 6, and 7, the groups of essays for each score shift over to the right for the most part. In set 4, we see how there are a couple of essays with the score of 1 that have a smaller word count and vocabulary list than the essays with a score of 0, but that is probably because essays with a score of 0 are either incomplete or are not about the topic. As a result, that is an outlier in this case, and does not speak to the general, positive relationship.

For set 3, we see that the essays for each score do not **shift** over to the right, but instead the groups of essays **extend** further to the right as the scores increase. That shows that there are many essays, which have the same word and vocabulary length, but have different scores. However, the essays that do tend to have a distinctly greater word count and vocabulary size do tend to receive higher scores. Similarly, for sets 1, 2, 4, 5, 6, and 7, we see that the shifts are very slight as the scores increase, and that there still is a large overlap between the counts and scores, as they increase. This points to the same conclusion that word length might not correlate greatly to score because there are many essays with the same counts that earn different scores, but the essays that do have noticeably larger counts, do warrant higher scores.

Finally in set 8, we see an opposite trend in that essays with large word counts and vocabulary sizes range greatly in their scores. However, it is obvious that the essays that have low counts have very low scores. So unlike the other datasets, where noticeably higher counts equate to higher scores and the rest of the essays seem to be unaffected by the counts, we see that here, noticeably lower counts equate to lower scores and the rest of the essays seem to unaffected by the counts. This may be because this set of essays had a maximum word length, while the others didn’t.

In the future, we are going to examine more than just word counts and vocabulary size to determine essay grades. Some of the other factors we will be observing include number of sentences, number of paragraphs, paragraph lengths, number of grammatical errors, and number of punctuations used. Many of these numbers relate to the depth of the writing, and the number of ideas presented. With this information, hopefully, we will be able to find trends in the sets that we have, and see which predictors are the most significant when determining the score.