**3 Experiments:**

1. My first experiment will measure the frequencies of the most common words and I will show a frequency table as well as a bar graph of the 25 most common words among the large, combined dataset. Additionally, I will use pandas to create a word cloud (similar to the first experiment mentioned in the project description).
2. For my second experiment, I will look at the differences between the individual datasets (instead of looking at the merged dataset), and I will focus more on the length of the most common words. For each of these datasets, I will create a histogram of the length of some of the more commonly used words (don’t know yet how many letters I’ll go up to – will have to see once I start). After this, I will calculate, display, and compare the average word length for the top 25 words within each dataset.
3. My last experiment will shift the focus away from very common words such as “the”, “and”, “like”, and other similar words. I will use the merged dataset and eliminate any words that are not at least 6 letters long, hopefully leaving words that have a little more context when left by themselves. I will display how many entries this got rid of and then I will create another bar graph showing the frequency of what are now the most common words in the dataset.

**Updates on data collection:**

After going through the feedback you left for the first progress report, I decided it would be better to use option 1 for my data. This was the option with the four different word frequency datasets I found from the following sites:

* <https://www.wordfrequency.info/samples.asp>
* <https://www.kaggle.com/datasets/rtatman/english-word-frequency>
* <http://norvig.com/google-books-common-words.txt>
* <https://github.com/filiph/english_words/blob/master/data/word-freq-top5000.csv>

I figured I would be able to focus a little bit more on completing my actual experiments rather than gathering, processing, and cleaning the data with this option instead of option 2. Additionally, depending on how everything looks and how much data I have, there’s a strong chance that I only end up using three of these four sources since they all have a lot of data, and they seem to be mostly similar.

**Current status:**

Currently, I am in the process of downloading all of the data from these websites, creating separate csv files, and will soon be in the process of merging all of these datasets into one large, combined dataset. From there, I will begin my three experiments and document the results. As of now, I do not see any problems with completing the project on time. However, there is a chance that I run into an issue that could cause a setback when using pandas during my experiments due to a lack of much experience using it. At the moment, I do not think I will need to limit my original project deliverables.