Overall I think I am in a better position this check in period with this project than I was last time. I have not only finished cleaning and getting some basic visualizations, but have already drafted many of the helper functions that I want to help simplify what is happening in the notebook, and created some rudimentary models. I don't think I am all the way where I want to be, but I think I am just on the cusp of finishing the basic notebook and draft of all the code and results.

Being this far along is going to make the draft of the project paper will be much easier and allow me to better write the abstract to incorporate the results in my early stages of the paper.

Any surprises from your domain from these data?
What I am surprised about this data is how dirty the text is, but I guess I shouldn't be too surprised. I was questioning how much I was going to need to clean the data, but had little idea of what would be best for creating deep learning models. For example, I knew some degree of NLP was going to be used, here it was primarily tokenization that was required in order to get it into a numeric matrix format the models would be able to interpret.

The dataset is what you thought it was?
The dataset was exactly what I had expected when it came to just being text messages, lots of text messages with many of them being spam. The only thing I could ask for is more test data, especially if I could get more spam messages incorporated as well.

Have you had to adjust your approach or research questions?
Yes I have adjusted my research questions and approach just slightly. I am no longer trying to improve and build non-deep learning models but instead looking at the cleaning and NLP operations that are required for deep learning models. I am curious what will likely yield the best possible models because this type of NLP set for deep learning models with text is a crucial step to build a viable model.

Is your method working?
Yes it is working, concentrating on NLP actions on the data to simplify the tokenization process and make the dictionary as simple and accurate as possible. My original thought was to do as little cleaning and NLP actions as possible, but it seems like that would not work so well because the tokenization with NLP libraries gets more and more precise and can possibly improve the same models. One of the things that I was hoping would help is having certain representations of words like FREE or CASH with the capitalization would help, but it really just kinda adds junk somewhat in the tokenization vocabulary. I would really like to dig into that issue further, thinking about possibly having certain specific NLP tokenization for these types of message might help.

What challenges are you having?
Some challenge I am having is finding an NLP library that is able to translate shorthand used in text messages to full english messages, and that this process using the GingerIT library is rather slow because it uses an external API that is likely throttled. The problem is text shorthand can have a many to many relationships. I also need to just slightly refine some of my helper

functions so I don’t have so much repetitive code which is sometimes harder to read and ugly, and of course writing the project paper draft is always a challenge to get started.