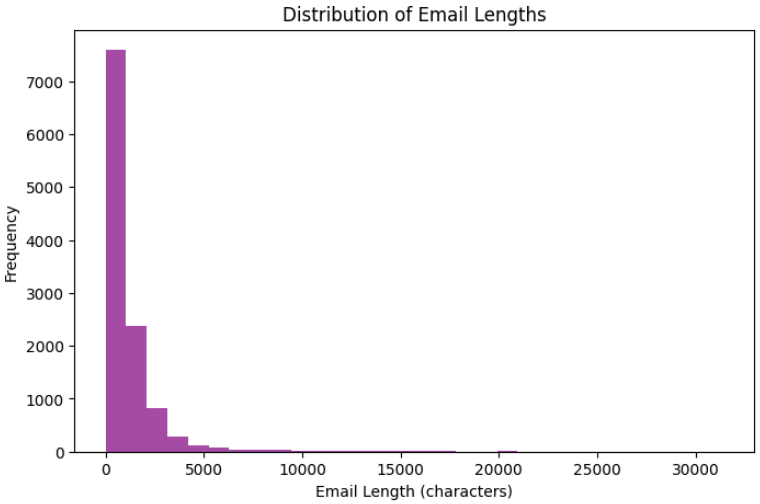
1. What is in your data?

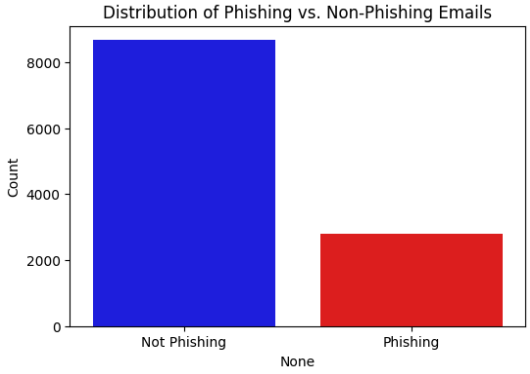
We decided to look at data from Kaggle, and discovered a dataset that included information from phishing emails, and around 15,000 messages. We decided to focus our energy on TREC_06, a recently updated but smaller size file of information. The variables included are “sender”, “receiver”, “subject”, “body”, “date”, “label”, and “urls”. Sender and receiver refer to metadata used to keep track of information sent, subject and body are the main email content, date refers to when the email was sent, label highlights whether the email was flagged as spam, and urls contains information if the email contained any.

```
Emails Containing URLs:
 0    7246
1    4225
```



2. How will these data be useful for studying the phenomenon you're interested in?

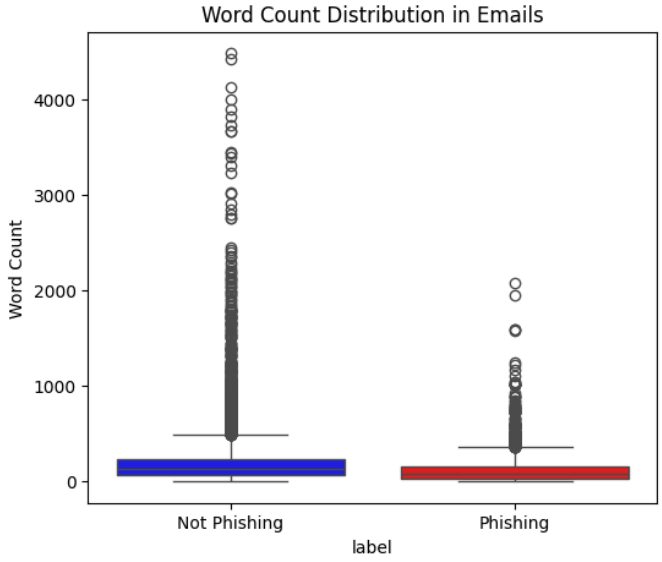
Our main goal with this data is to eventually create a predictive model that can successfully categorize phishing emails. These Text Retrieval Conference (TREC) Spam Track datasets were actually developed for this exact purpose; to simulate incoming emails that required filters to determine if they were spam or not spam. Since we will have the label variable available for us, we will know whether an email was flagged or not, so the model can successfully train and be tested without this category. Based on initial statistical analyses, we know that there are 8,668 non-spam and 2,803 spam emails present.

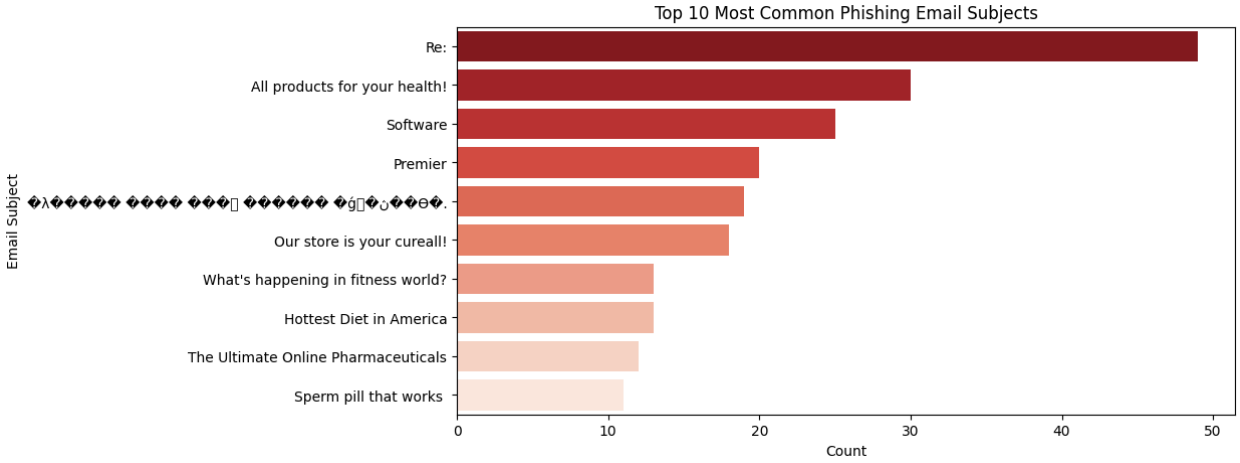



```
Phishing Email Counts:
 0    8668
1    2803
```

There are a few different aspects of the data that we might predict the model will learn to use in its distinctions. Phishing emails seem to include less words, often include “A0” and demanding

language such as “will”, “need”, “now”, “want”, and are likely to include “money” as well. The subject line might also be incredibly telling, with distinct lines including “Re:”, long strings of special characters, exclamation points and question marks, and other eye-catching titles.





3. What are the challenges you've resolved or expect to face in using them?

The main issue we expect to deal with is the sheer size of the dataset. At 15.5 MB, processing with python will be tedious, especially if we attempt to filter it for specific entries. Unfortunately, there is not much that can be done about the processing delays we will likely experience, writing the model code and testing it will take a lot of time, so we will need to carefully plan ahead. Furthermore, as of right now we are planning on splitting the data into training and non-training for model accuracy testing, however we are unsure of exactly how we will want to split it percentage wise to avoid overfitting or underfitting our model.

1. A. I. Champa, M. F. Rabbi, and M. F. Zibran, “Why phishing emails escape detection: A closer look at the failure points,” in *12th Interna- tional Symposium on Digital Forensics and Security (ISDFS)*, 2024, pp. 1–6.

2. A. I. Champa, M. F. Rabbi, and M. F. Zibran, “Curated datasets and feature analysis for phishing email detection with machine learning,” in 3rd IEEE International Conference on Computing and Machine Intelligence (ICMI), 2024, pp. 1–7.