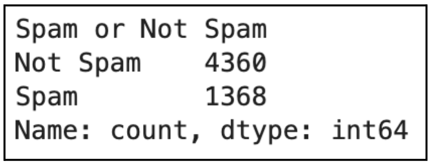
Data Appendix

Our final dataset (preprocessed_data.csv) was derived from our raw dataset (emails.csv) and it includes six columns. The two columns ‘Email Text’ and ‘Spam or Not Spam’ were cleaned/renamed from the original data. The final dataset also includes numerical columns tracking the character count, word count, and sentence count of the emails. The final column in the dataset is the preprocessed email text. Because each row in this dataset represents a singular email, the unit of observation is “email.”

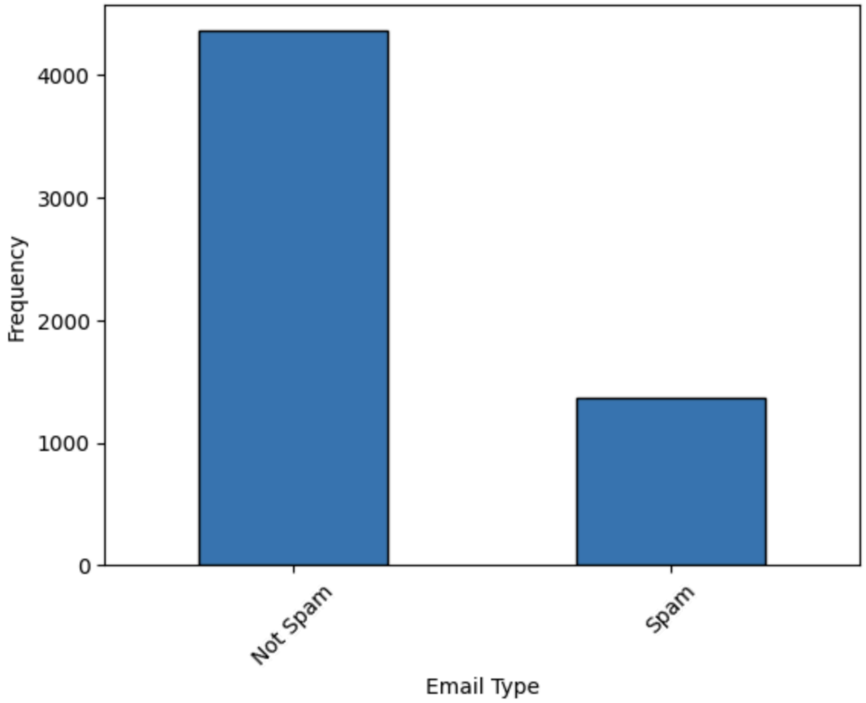
Variables in preprocessed_data.csv:

- **Email Text**
  - *Definition:* Contains the textual content of the email.
  - *Missing values: 5728(0)*
- **Spam or Not Spam**
  - *Definition:* Contains a value of “Spam” or “Not Spam” depending on the type of email.
  - *Missing values: 5728(0)*
  - *Frequency table:*

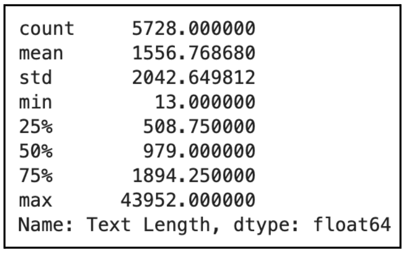


```
Spam or Not Spam
Not Spam    4360
Spam        1368
Name: count, dtype: int64
```

  - *Frequency distribution:*

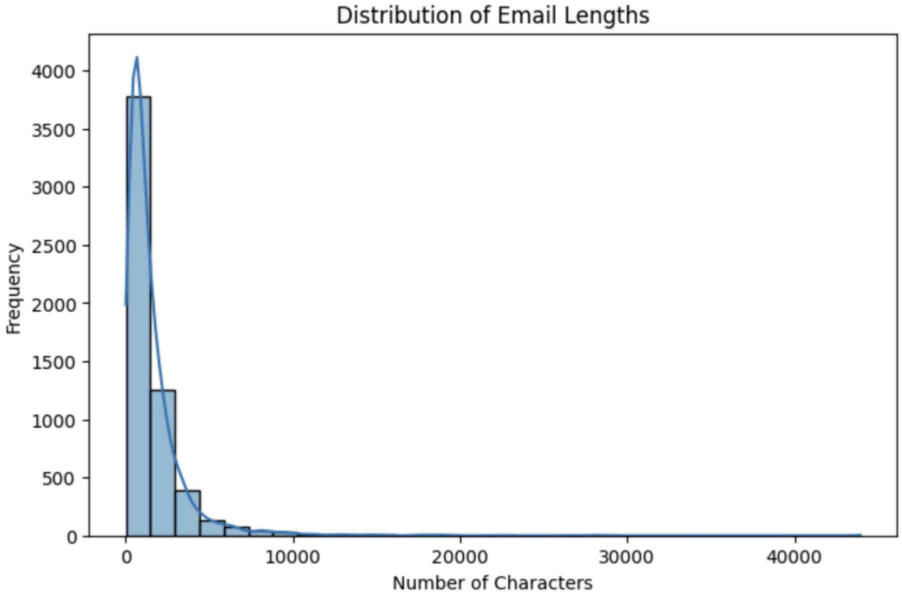


- **Text Length**
    - *Definition:* Contains the number of characters in the email.
    - *Missing values: 5728(0)*
    - *Variable production:* This variable was produced by finding the length of each row in the ‘Email Text’ column.
    - *Summary statistics:*

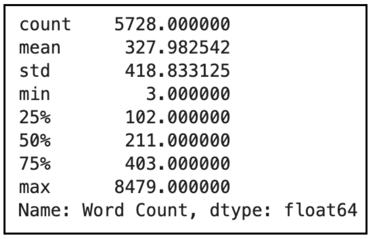


```
count     5728.000000
mean      1556.768680
std       2042.649812
min         13.000000
25%        508.750000
50%        979.000000
75%       1894.250000
max      43952.000000
Name: Text Length, dtype: float64
```

    - *Histogram:*

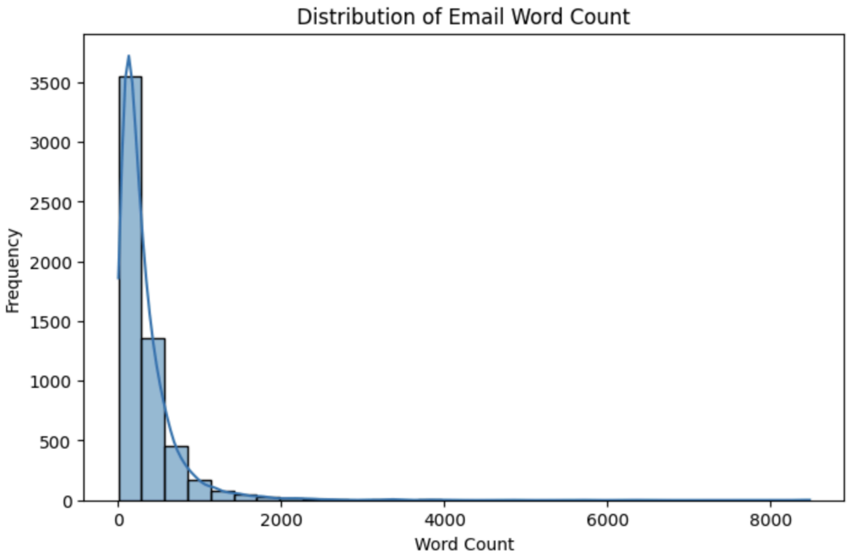


- **Word Count**
    - *Definition:* Contains the word count for each email.
    - *Missing values: 5728(0)*
    - *Variable Production:* This variable was produced by applying NLTK’s word_tokenize() function on the ‘Email Text’ column, and then finding the corresponding length.
    - *Summary statistics:*

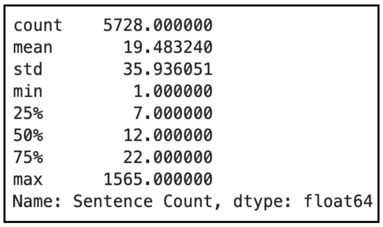


```
count    5728.000000
mean      327.982542
std       418.833125
min         3.000000
25%       102.000000
50%       211.000000
75%       403.000000
max      8479.000000
Name: Word Count, dtype: float64
```

  - *Histogram:*

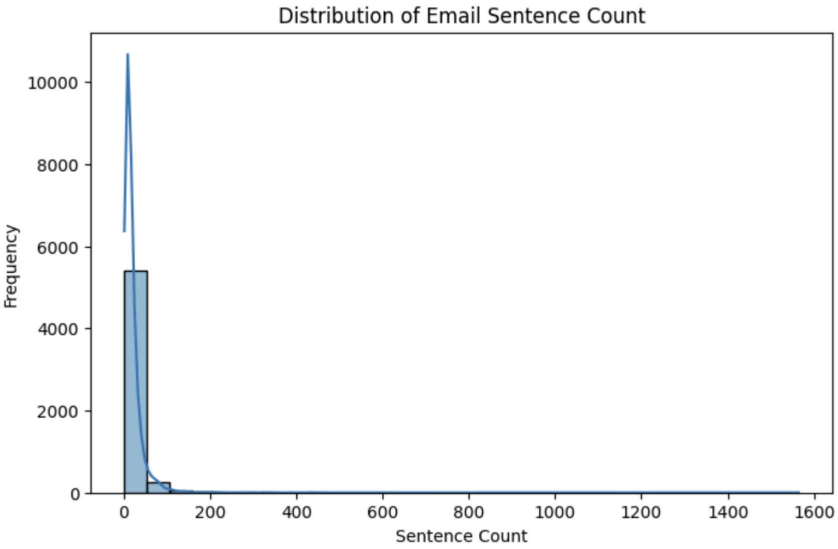


- **Sentence Count**
  - *Definition:* Contains the sentence count for each email.
  - *Missing values: 5728(0)*
  - *Variable production:* This variable was produced by applying NLTK's sent_tokenize() function on the 'Email Text' column, and then finding the corresponding length.
  - *Summary statistics:*



```
count    5728.000000
mean       19.483240
std        35.936051
min         1.000000
25%         7.000000
50%        12.000000
75%        22.000000
max      1565.000000
Name: Sentence Count, dtype: float64
```

  - *Histogram:*



- **Preprocessed Text**
    - *Definition:* Contains the textual content of the email after textual preprocessing (lowercase, tokenization, stopword removal, and stemming).
    - *Missing values: 5728(0)*
    - *Variable production:* This variable was produced by applying the necessary preprocessing steps onto the ‘Email Text’ column. ‘Email Text’ was first converted to lowercase and saved into the new ‘Preprocessed Text’ column. Then, tokenization was applied to ‘Preprocessed Text.’ Stopwords and punctuation were also removed from all entries in the column before it was stemmed and converted back to string format.