Project 1 DS 4002 Data Appendix

1. Dataset 1: [Raw.csv]

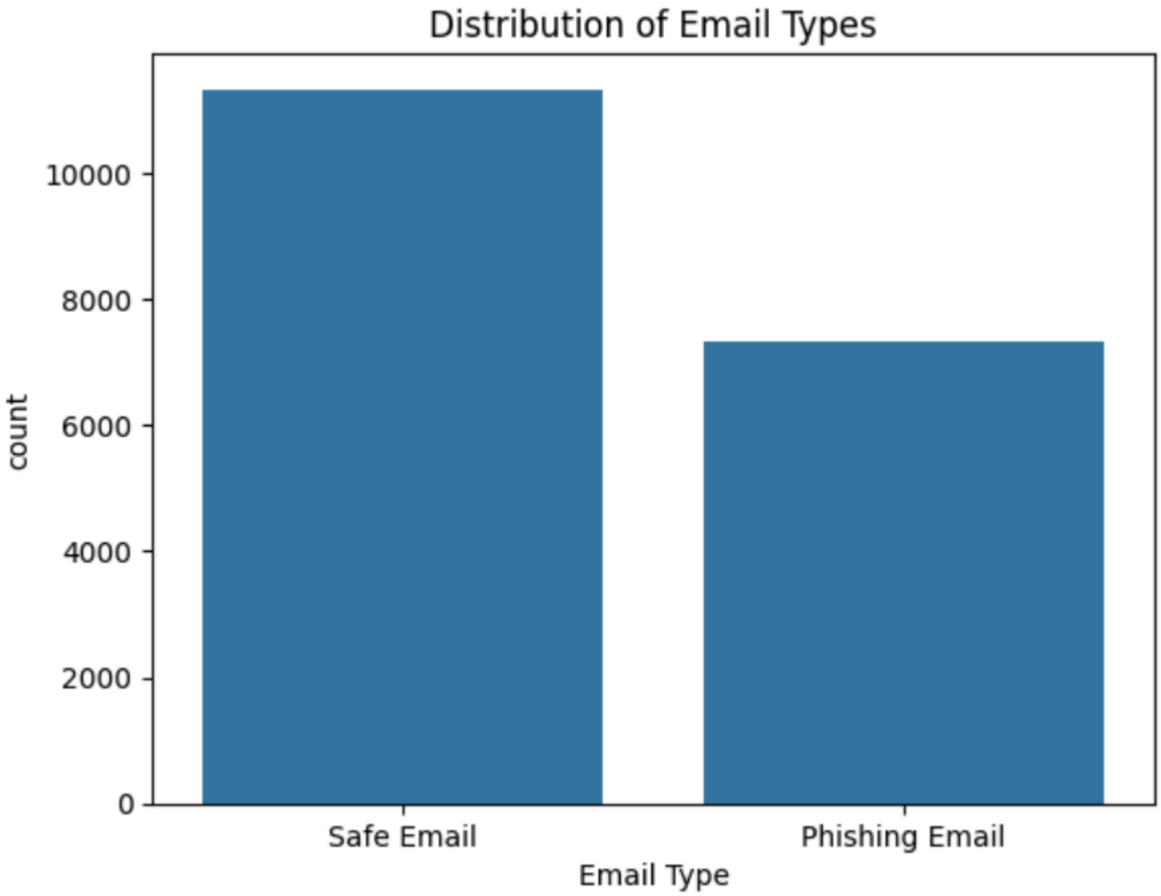
- **1.1 Unit of Observation:** Each row contains information about a single email
- **1.2 Variables Overview:** Each observation contains email_text (str) and email_type (str)

  1.3 Variable Subsections :
  - **[Variable Name]:** email_text, contains the text from the body of the email, all emails contain text leading to no missing value.
  - **Data Type:** Text
  - **Notes/Observations:** Some of the emails contain their subject line aswell as their body text.

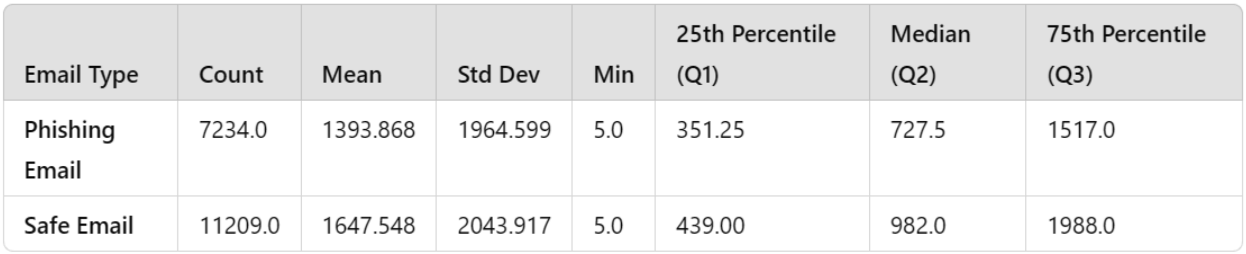
  1.4 Variable Subsections :

  - **[Variable Name]:** email_type, contains one of two values, either Safe Email or Phishing Email.
  - **Data Type:** Text
  - **Notes/Observations:** All emails contain a value for this variable, no emails have missing values.
- **1.4 Overall Dataset Descriptive Statistics:**
  - The Dataset contains a total of 18634 rows, each denoting their own email, containing both the text and the type.

- **2.5 Figures and Visualizations:**



- Distribution of Email Text Length by Email Type



Email Type	Count	Mean	Std Dev	Min	25th Percentile (Q1)	Median (Q2)	75th Percentile (Q3)
Phishing Email	7234.0	1393.868	1964.599	5.0	351.25	727.5	1517.0
Safe Email	11209.0	1647.548	2043.917	5.0	439.00	982.0	1988.0

---

2. Dataset 2: [Sentiment.csv]

- **2.1 Unit of Observation:** Each row represents the sentiment score created by VADER for a single email.
- **2.2 Variables Overview:** Each observation contains email_text (str), email_type (str), and four values representing the sentiment score of the text: positive, neutral,

negative, and compound.

2.3 Variable Subsections :

- **[Variable Name]:** email_text, contains the text from the body of the email, all emails contain text leading to no missing value.
  - **Data Type:** Text
  - **Notes/Observations:** Some of the emails contain their subject line aswell as their body text.
- **[Variable Name]:** email_type, contains one of two values, either Safe Email or Phishing Email.
  - **Data Type:** Text
  - **Notes/Observations:** All emails contain a value for this variable, no emails have missing values.
- **[Variable Name]:**Pos (Positive), contains a float between 0 and 1, 0 indicating no positive sentiment and 1 indicating all positive
  - **Data Type:** Float
- **[Variable Name]:**Neu (Neutral), contains a float between 0 and 1, 0 indicating no neutral sentiment and 1 indicating all neutral sentiment
  - **Data Type:** Float
- **[Variable Name]:**Neg (Negative), contains a float between 0 and 1, 0 indicating no negative sentiment and 1 indicating all negative sentiment.
  - **Data Type:** Float
- **[Variable Name]:** Compound, contains the compound score of all pos, neg, and neu variables. This variable serves as the overall sentiment score.
  - **Data Type:** Float

- **2.4 Overall Dataset Descriptive Statistics:**
  - The Dataset contains a total of 18634 rows and 5 columns

3. Dataset 3: email.parquet

- **Unit of Observation:** Each row represents a single email transformed using TF-IDF vectorization.
- **Variables Overview:** Each observation contains email_type and TF-IDF features.
  Variable Subsections:
  - Email_type: Same classification as in previous datasets: Safe Email or Phishing Email.
    - *Data Type:* Text
  - **TF-IDF Features:** Numeric values representing the frequency of unigrams and bigrams.
    - ***Data Type:*** Numeric (sparse matrix format)
    - ***Notes/Observations:*** The TF-IDF matrix is large and sparse due to the vectorization of terms.
- **Overall Dataset Descriptive Statistics:**

- The dataset contains 18,634 rows and over 5,000 columns representing TF-IDF features.