Data Science Methodology Exercise

Which topic did you choose to apply the data science methodology to?
E-mail

Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer.
You are required to:
Describe the problem, related to the topic you selected.
Phrase the problem as a question to be answered using data.

Problem Description: Email communication is widely used today, but it also brings various security risks. Threats such as spam messages, phishing attempts, and malicious software targeting email accounts can jeopardize users' data and privacy. Therefore, email security is of great importance.

Problem Formulation: Regarding email security and spam filters, is it possible to develop a system that automatically identifies incoming emails as spam and filters them before reaching users' primary inbox?

Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with.
Analytic Approach
Data Requirements
Data Collection
Data Understanding and Preparation
Modeling and Evaluation

Analytical Approach: In this stage, we would analyze the problem of email security and spam filters. We would review the current state of email security technologies, existing spam filtering algorithms, and their effectiveness. Additionally, we would examine relevant literature and previous studies in this field to gain insights.

Data Requirements: We would identify the data requirements for email security and spam filters. This involves obtaining a dataset containing examples of spam and non-spam emails. The dataset should be labeled, indicating whether each email is spam or not.

Data Collection: During the data collection stage, we would gather a dataset containing spam and non-spam emails from various sources. This may involve searching for open datasets, collecting data from email providers, or creating a dataset in our own laboratory.

Data Understanding and Preparation: We would use data exploration and preprocessing techniques to understand and prepare the collected data. This includes visualizing the dataset, handling missing data, removing unnecessary features, and preprocessing the data to make it suitable for modeling.

Modeling and Evaluation: Finally, we would build and train a spam filtering model using machine learning or deep learning techniques. After training the model, we would evaluate its performance using metrics such as accuracy, precision, recall, and others. We would then fine-tune the model as necessary to improve its performance and report the result