1) Which topic did you choose to apply the data science methodology to? (2 marks)

I've opted to explore the realm of emails, given their integral role in our professional routines and their daily prevalence in our lives.

2) Next, you will play the role of the client and the data scientist.

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. (3 marks)

You are required to:

Describe the problem, related to the topic you selected.

Phrase the problem as a question to be answered using data.

For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

Problem Description: In a professional environment, emails are an indispensable means of communication. However, the influx of emails can also bring about challenges such as identifying and managing spam emails, as well as potentially malicious threat emails. These issues can disrupt workflow, compromise security, and waste valuable time. Therefore, it is crucial to have an effective system in place that can accurately classify and handle various types of emails, ensuring a smooth and secure email communication process.

Problem Question: "How can we develop a robust email classification system that accurately distinguishes between legitimate work-related emails, spam emails, and potential threat emails? Furthermore, how can we efficiently import and manage emails from the organization's email servers to facilitate streamlined communication and security?"

This question aims to leverage data science techniques to create a solution that addresses the challenges posed by spam and threat emails, while also enhancing the organization's overall email management process.

3)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. (5 marks):

Analytic Approach

Data Requirements

Data Collection

Data Understanding and Preparation

Modeling and Evaluation

You can always refer to the labs as a reference with describing how you would complete each stage for your problem.

1. Analytic Approach:

The analytic approach for this problem would involve implementing a supervised machine learning classification model. The model will be trained on a labeled dataset of emails, with features extracted from the email content, sender information, and other relevant metadata. Classification algorithms such as Naive Bayes, Support Vector Machines, or deep learning approaches like recurrent neural networks (RNNs) could be explored.

2. Data Requirements:

The required data for this problem includes a diverse and representative dataset of emails. This dataset should cover a wide range of scenarios, including legitimate work-related emails, spam emails, and potential threat emails. Alongside the labeled email data, additional metadata such as sender details, timestamps, and email subjects will also be needed for feature extraction.

3. Data Collection:

Data collection involves sourcing emails from the organization's email servers. Collaborating with IT teams and administrators, access to both incoming and historical emails can be secured. Additionally, labeled examples of spam and threat emails might be gathered from past incidents and external sources to enhance the training dataset.

4. Data Understanding and Preparation:

In this stage, the collected data will be cleaned, preprocessed, and transformed into suitable formats for analysis. Text processing techniques like tokenization, stemming, and stop-word removal will be applied to extract meaningful features from the email content. Feature engineering will also include extracting sender characteristics, attachments, and other relevant attributes.

5. Modeling and Evaluation:

Multiple classification models will be developed using the preprocessed data. The dataset will be split into training and testing sets to evaluate model performance. Various evaluation metrics such as accuracy, precision, recall, and F1-score will be used to assess the model's effectiveness in distinguishing between different types of emails. The chosen model will then be fine-tuned to achieve optimal performance.

Through these stages, a data-driven solution will be developed to effectively classify and manage emails, addressing the challenges of spam and threat emails while streamlining communication processes and ensuring security within the organization.