In this Assignment, you will demonstrate your understanding of the data science methodology by applying it to a given problem. Pick one of the following topics to apply the data science methodology to:

1. Emails
2. Hospitals
3. Credit Cards

You will have to play the role of the client as well as the data scientist to come up with a problem that is more specific but related to these topics.

Please note that this assignment is worth 10% of your final grade.

Project Title: **Email Organizer**

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

A: **Emails**

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 points)**

You are required to:

1. Describe the problem, related to the topic you selected.

A: **The XYZ company receives a lot of emails from our clients and suppliers. Unfortunately, some ads and spam also arrive in the box. Employees always need to spend time selecting useful messages from useless messages. This happens because our email server cannot handle effectively separating the spam messages, and sometimes we lose an important message in the junk folder.**

**Now we need some help from a Data Scientist to create a useful model to separate and classify our messages using the historical data from our suppliers and clients. Notwithstanding, organize these messages according to priority.**

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

A: **Can we effectively organize email messages from clients and suppliers’ historical data into a visually organized app?**

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?".

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 points)**:

1. Analytic Approach

A: **The problem was divided into two parts, the first was to separate the useful from useless messages, and the second part was to rank the messages according to priority.**

**The first part of the problem can use the email addresses and the historical message context to generate a reasonable database for separating spam/ads from business-related messages.**

**The second part of the problem needs more data and is dependent on the first model that should provide the business-related messages. With the useful messages, the second model could use historical supplies from suppliers and the sales for the customers.**

1. Data Requirements

A: **The first model should use email-related data such as the addresses, the subject, and the content of the messages. The second model needs to use historical data from the customers and suppliers. These data include sales, products, quantities, delivery methods, delivery time, prices, payment methods, discounts, and fees.**

1. Data Collection

A: **Data collection would involve gathering the email's raw data from the email servers. The historical data from customers and suppliers should be provided from the sales database and the supplier's purchases database.**

1. Data Understanding and Preparation

A: **In the data understanding and preparation stage, the data for the first model would provide all the information for the useful messages in the inbox. Additionally, data mining should be implemented to extract useful data from the content of these useful emails. The extracted data will be converted and stored in a structured format. With the separation of the useless messages, the second model can be made by the useful emails. All information from customers and suppliers should be retrieved from historical databases and will be provided and converted to a structured format.**

1. Modeling and Evaluation

A: **At the modeling stage, different techniques would be applied to achieve the results. For the first problem, natural language processing (NLP) should be used to extract insights from the email contents. For spam detection, a machine learning classifier algorithm could be used for the useful messages. The second model should be constructed using forecast models and a machine-learning approach.**

**The models should be evaluated using the accuracy, precision, recall, cross-validation, and ROC curves to assess the model's effectiveness on the email ranking process. Feedback methods can be used to improve the models by discussing results with stakeholders and reviewing the model parameters with new data along the model functionalities.**

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