Assignment

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

The topic that I have chosen to apply data science methodology to is Emails.

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:

1. **Describe the problem, related to the topic you selected.**
2. **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?".

Every day, we receive hundreds of emails, and it is unlikely that we will be able to read them all. We may sort emails into categories like Promotions, Updates, Social, Order Receipts, Important/Not Important, Spam, and so on to see which ones are worth looking at again.

Question: Is it possible to automatically classify emails by its content?

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):**

1. **Analytic Approach**
2. **Data Requirements**
3. **Data Collection**
4. **Data Understanding and Preparation**
5. **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 As long as we're looking for businesses that might be related in some way, we'll need to employ a descriptive model. It would also be beneficial to define clusters within a specific area.

2.Data Requirements To build the model, we'll need information on the sender, such as their email address, domain, subject, language, whether the email contains an attachment, and the email body to determine if it contains a list (presence of a list could help classify the email as an order).

3.Data Collection All of this information can be gathered from numerous email accounts and inboxes (Gmail, Hotmail, yahoo, outlook etc.). To produce a useful dataset, we can combine the emails from the multiple inboxes. Descriptive statistics & visualizations can be applied to the data set to assess the content quality and if we have the required information.

4.Data Understanding and Preparation The redundant data in our dataset should be removed. It's possible that two identical emails were delivered to two distinct inboxes. We must perform text analysis because we are working with text. We should ensure proper groupings to help classify the emails properly. These groupings should be done based on certain keywords present in the subject or content of the email.

5.Modeling and Evaluation The classification model is created by us. We examine the model's output to see how much of it is labelled properly or inaccurately. We may tune the model by adding parameters and making appropriate modifications based on this feedback to verify that we're obtaining the desired outcomes.