**Project Title: Give your project a descriptive name**

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

I have chosen ‘emails’ as the topic.

*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:*

*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?".*

There have been lots of surveys done over the years that show that the majority of the workforce feel overwhelmed by the number of emails they receive, and can find it difficult to prioritise. The problem/ question I wish to solve is this: Is it possible to automate the process of scanning these emails and putting them into a to-do list? This entails:

1. Categorising emails by subject matter
2. Prioritising the emails in order of urgency/ deadline
3. Transcribe the requests within these emails to a to-do list in an app like OneNote

**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):**

Analytic Approach

For this project we are categorising the emails in terms of subject matter and urgency. For that reason the classification model is most appropriate.

Data Requirements

We would need to know:

* The subject categories that are relevant to the job role
* Employees and their job titles to help classify by subject. Also a list of regular outside contacts
* Deadlines relating to each subject. For finance related roles this could be e.g. when journals need to be posted, when budget holder meetings are due to take place

Data Collection

In this stage we need to collect the required data from a number of sources:

* A list of internal and external contacts, with their job titles. This can probably be obtained from the address list of Outlook
* A list of job tasks relevant to the role, can be obtained from the job description and from the employee doing the job
* A list of deadlines relevant to the role, which can also be taken from the job description and those doing the job

Data Understanding and Preparation

This step is for cleaning up the collected data, identifying missing information, checking data for accuracy. This would involve:

* Are the job descriptions up to date?
* Are the address lists in Outlook well maintained/ up to date?
* Are there any regular contacts that are not included in the address lists?
* Is the list of subjects we are using as classifications clearly defined, or do they overlap?

Modeling and Evaluation

We would need a prescriptive model to solve this question as we are trying to predict what categories the emails fall into, what tasks are contained within the emails (if any) and the relative urgency of the task(s).

We would need to make use of machine learning to train on all of the information available to correctly categorise emails and rank by urgency.

We would need to create a training set of data where we know the desired outcomes in order to calibrate the model to make it more accureate.

After the model has been prepared it will be deployed to stakeholders. Any feedback we get from those who use the model will be used to help improve the model.

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