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

I have chosen to apply the data science methodology to emails as my parents are constantly inundated with spam emails phishing for their credit card details.

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.

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?".

Almost everyday many people get sent emails from multiple sources, including their; friends, work, delivery services, streaming services and many other sources. Occasionally amongst this mix of truthful emails, there are some that are promising special discounts or trying to obtain personal info from the receiver. In the worst cases they are trying to scam the receiver. How can we automate the detection of these spam emails and redirect them to a spam folder?

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.

1. **Analytic Approach**
2. **Data Requirements**
3. **Data Collection**
4. **Data Understanding and Preparation**
5. **Modeling and Evaluation**

1. Analytic Approach Need to define the analytic approach to solve the problem, in this case this will be classifying the emails that are received as either spam or regular emails. Therefore a classification ML algorithm would need to be used.

2. Data Requirements The classification algorithm would require data in the form of all the emails the receiver has been sent

3. Data Collection Need to start collecting data from all our email inbox, to increase the amount of data emails could be obtained from other family members or from the internet. Getting data from different sources and of different types such as; structured, semi-structured, unstructured.

4. Data Understanding and Preparation To understand the data's content I would need to perform descriptive statistics on the words that appear in the emails and visualise the occurrences of different words, their distributions and occurrences of any phrases. Top prepare the data, I would need to clean the dataset by removing any duplicates - for example if there are multiple emails posing as delivery services or offering special discounts etc. Once all the operations have been performed on the data the data would need to be merged into one data frame.

5. Modelling and Evaluation Modelling would involve using various algorithms such as; KNN, Naive Bayes, SVM and Neural networks to implement classification and even using various libraries. Then comparing the respective accuracies of each algorithm at classifying an email as spam. Evaluation involves checking the model's quality, this can be done by splitting the dataset into testing and training and then training the model using the training set and testing it on the testing set. The model can then be evaluated by how accurately it predicts the actual spam emails.