**Section 1: Theory Questions**

* 1. **In your own words, what does the role of a data scientist involve?** [2 points]

The role of a Data scientist involves the key responsibilities below:

1. **Data Cleaning & Data exploration & Data Preparation:** Data scientists gather large datasets, clean it up, and prepare it for analysis. They familiarise themselves with the data to find interesting insights. During the data cleaning process, they handle the missing values and outliers whilst maintaining data quality.
2. **Data Analysis & Trend prediction:** Data scientists uses various methods to extract meaningful patterns and insights from the data. They utilise statistical methods and Machine Learning models to develop predictive models, identify trends and establish the relationships within the data.
3. **Data Visualisation:** Data scientists create dashboards and reports using data visualisation tools like Tableau and Power BI to share their results with stakeholders and aid in their understanding of complex data patterns.
4. **Continuous Learning:** Data scientists have to stay curious and creative in order to solve data-related challenges.

**1.2 What is an outlier? Here we expect to see the following:** [4 points]

1. **An outlier** is a data point that deviates drastically from the rest of the data in a dataset.
2. **Examples:** Below is an example of a scatterplot which was used to analyse the titanic data set. In the Titanic dataset, the majority of fares fall within the range of 0 to just under 300. However, two fares at 500 are considered outliers.

A graph of blue dots

Description automatically generated

1. The decision to remove outliers depends on the individual and the project's goals. In some cases, eliminating outliers may enhance data visualization, while in other situations, retaining outliers is crucial as they can convey valuable information. Outliers might signify uncommon events or errors, yet they can also reveal significant data trends or anomalies.
2. **What are other possible issues that you can find in a dataset?**

* Inaccuracies in data collection
* Missing data values
* Errors or anomalies
* Data Imbalance

**1.3** **Describe the concepts of data cleaning and data quality. Here we expect to see the following:** [4 points]

1. **Data cleaning** involves the removal of inaccurate data, including improperly formatted data and duplicated entries. This process involves handling missing values, normalising the data, handling any outliers whilst maintaining data integrity.
2. Failure to complete the **data cleaning** process will impact the accuracy of the insights derived from the data. The reliability of data outcomes and algorithms will be compromised by inaccurate data, even if they appear to be correct.
3. **What type of mistakes do we expect to commonly see in datasets?**

* Incomplete or missing Data values
* Outliers
* Duplicate values
* Unreliable data or datasets result from inaccuracies.

**1.4 Discuss what is Unsupervised Learning - Clustering in Machine Learning using an example. Here we expect to see the following:** [7.5 points]

1. **Unsupervised learning** analyses and clusters unlabelled datasets using machine learning methods. The algorithm is not provided with labelled output data to learn from. And these algorithms find hidden patterns in data without human interaction.
2. **Unsupervised learning** is used to perform three significant tasks such as **Clustering, Association and Dimensionality reduction. It is used to** identifies hidden patterns, structures, or relationships within the input data on its own.
3. A real-world application of unsupervised learning is **customer segmentation.**

* **Problem:** identifying a separate group of customers based on their purchase habits and without any predefined categories.
* **Application:** The goal of unsupervised learning is to discover patterns and similarities among customers based on several features, such as identifying segments with similar tastes, demographics, and purchasing behaviour, which can help with personalised marketing techniques

1. While Unsupervised learning is extremely powerful, it also comes with it own set of limitations such as:
   * **Struggling with Complex Patterns:** especially when the data is noisy or contains outliers.
   * **Lack of Clear Accuracy Measure**: There are no clear accuracy measures because there are no labelled outputs.
   * **Difficulty in Performance Assessment:** It can be challenging to critically assess the algorithm's performance.

**1.5 Discuss what is Supervised Learning - Classification in Machine Learning using an example. Here we expect to see the following:** [7.5 points]

1. **Supervised learning** is a machine learning approach defined by the usage of labelled datasets. These datasets are crafted to guide and instruct computers in recognising patterns in data and predicting outcomes.
2. When engaging in data mining, **supervised learning** is applicable to two main types of problems: classification and regression. Labelled inputs and outputs are used to train the algorithm in these instances, allowing it to quantify its accuracy over time. Based on the patterns acquired from the labelled data, the model learns to classify data into specified categories (classification) or predict numerical outcomes (regression).
3. A real-world application of supervised learning is the task of  **Email Spam Classification.**

* **Problem:** Classifying emails as either spam or non-spam based on the labelled historical data.
* **Application:** Supervised learning models can learn patterns from labelled examples to classify incoming emails.

1. **What data do we need for it?**

* For supervised learning, we need labelled datasets including input features and their associated output labels.
* Using the real-world application example, the dataset would need to include certain characteristics extracted from emails such as the frequency of particular phrases and labels which indicates whether each email was a spam or non-spam.

1. **Is there any processing that needs to be done?**

* Yes, there are pre-processing steps that will need to be done such as cleaning and organising the data, handling missing values, and most importantly transforming categorical variables into a format suited for the selected algorithm.
* The dataset is commonly divided into training and testing sets. Training sets typically consist of 70%, 80%, or 90% of the data, while the remaining 30%, 20%, or 10% is reserved for testing.