THEORY QUESTIONS ASSIGNMENT

The answers need to be explained clearly and illustrated with relevant examples where necessary.  
Your examples can include code snippets, diagrams or any other evidence-based representation of your answer.

**Theory questions 10 point each  
1. What does “Data Cleansing” mean? What are the best ways to practice this?**

“Data Cleansing” is the process of make the identifying and correcting or removing data that is incorrect, corrupted, incomplete, duplicated or incorrectly formatted. This is usually the first step of any data driven process after obtaining the data as it allowed the following steps to be done without errors. Data cleansing is important in achieving data integrity and ensures that the conclusion made from the data are accurate. The data cleansing stages start with importing data, merging data sets, rebuilding missing data, standardisation, normalisation, deduplication, verification & enrichment and exporting data.

**2. What is the difference between data profiling and data mining?**

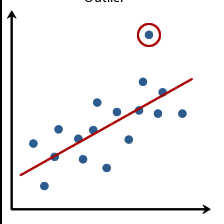
Data mining is the action of identifying patterns in a database. It evaluates large raw data and turns into useful information by finding patterns and trends in the existing data. It is done to discover knowledge in the database. The data mining techniques are regression, clustering, classification, association.

Data profiling analyses raw data to gather statistics and summarise the data. It is used to check the quality of the data and assess it for consistency, uniqueness and logic. This process helps identify incorrect values and missing values. The types of data mining are structure discovery, content discovery and relationship discovery.

Data mining is used to identify patterns in the existing data whereas data profiling is used to asses the quality of the data. The aim of data mining is to build machine learning techniques and the aim or data profiling is to achieve accurate and consistent data .

**3. Define Outlier with an example.**

An outlier is a value that lies outside the interquartile or a “normal” range of that set of values. The normal range needs to be defined to be able to identify outliers. Outliers are easier to observe from graphs than from raw data. If there is a graph showing the average height for ages of females between 0-18, you can plot a line of best fit. Any data that considerably far from that line can be identified as an outlier. These values are sometimes left out of average calculations and line of best as they can skew the data and provide inaccurate results. The outleir is corcles below



**4. What is “Collaborative Filtering”?**

Collaborate filtering, filters out information based on interaction and data from other users and is based on the concept that people who have similar opinions on certain items will have similar opinions on other items. For example, Netflix can use algorithms based on collaborative filtering to recommend you movies based on data from other users who liked similar movies as you. It focuses on the relationship between users and items and links items based on users who have rated both items to work out a similarity between them. It uses this similarity system to recommend items to other users. Collaborate filtering can measure the similarity between users or between items.

**5. What is “Time Series Analysis”?**

Time series analysis is a way of analysing data in respect to time. Data is recorded at consistent intervals to see the effect of time on the variable. It shows how data changes from on point to the other rather than just showing the start and end result. This can allow the users to see seasonal trends and then perform analysis to see why these trends occur. For example a Card shop could use time series analysis to plot the sales at a daily interval for a year to identify when there is an increase in sales in the year and use it to market there product more around those times. So time series analysis can be a very useful data analysis tool. It is also be used to forecast future values.

**6. Explain the core steps of a Data Analysis project?**

The core steps of a data analysis project are:

1. Define the goal and understand the activity: It is important to understand the aim and setting an objective and the question you are aiming to answer. Once you know what you are trying to achieve, it allows you to break the steps down and start your project.
2. Get the data: After identifying your objective and the question you are trying to answer, you need to find suitable data to help answer that question. Sometimes one data sources I enough and other time you need to merge data from multiple sources. You can get data from databases and data sets and use APIs.
3. Clean the data: After you have obtained the data you want to analyse, you have to clean and prepare the data. This is done by correcting or removing data that is incorrect, corrupted, incomplete, duplicated or incorrectly formatted. This step ensures data integrity.
4. Data modelling: You can build models to test your data and fins answers for your chosen question. You can use regression to predict future values, and classification to identify values.
5. Validation: After you have some results, the models should be assessed to ensure they is working as desired.
6. Visualise: Data visualisation is sued to represent your data so that it can be communicated easily. This can be done through a various number of graphs like bar charts, scatter graphs and charts and maps. This provides an effective way to show your conclusions.

**7. What are the characteristics of a good data model?**

The characteristics of a good data model is it provides data that can be “easily consumed”, large data changes are scalable, it provides predictable performance.

**8. Explain and provide examples of univariate, bivariate, and multivariate analysis?**

Univariate analysis is when you are analysing one variable. The most common analysis is checking the mean, median and mode, the range, the maximum and minimum values and standard deviation of the variable. You could do a box plot to see the spread of the variables and show the outliers.

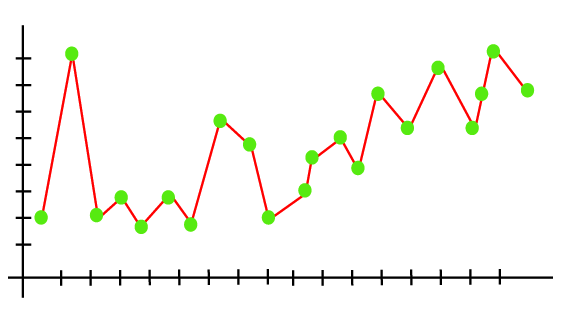
Bivariate analysis is when you are comparing two variables to study their relationship between each other. A common example is a scatter plot where the relationship between x and y can be seen.

Multivariate Analysis is when you are comparing more than two variables to study their relationship between each other. This can be done by a 3D model to study their relationship.

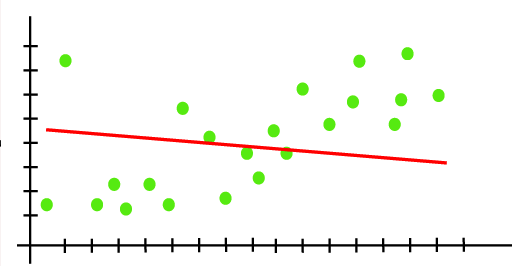
**9. What is a Linear Regression?**

Linear regression is a supervised machine learning model which finds the linear relationship between two variables, an independent and dependent variable. It predicts a value (y) based on the given independent variable (x). When the model is being built and trained it is given training data and then it fits the line if best fit to predict the value based on x. It finds the y intercept and the coefficient of x to find the line of best fit so that there is the smallest error. The error is the vertical distance between the data points and the regression line.

**10. In terms of modelling data, what do we mean by Over-fitting and Under-fitting?**

Over-fitting is when the model is tries to cover all data points. This is likely to occur when the model is trained on too much data. This model includes outliers and inaccurate values. This model low bias and high variance. Overfitting can be reduced by stopping the training of the data earlier.

This graph tries to cover all of the data points and the line is over-fitted.

Under-fitting occurs when not enough data is fed to the model may not learn enough to find a trend in the data. You can avoid underfitting, by increasing the training time of the model and increasing the number of features. This model has high bias and low variance.

The graph shows a line which does not capture the trend of the data as the model is underfitted.