General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

→ Linear regression is a supervised machine learning model that uses the strength of the straight line equation for prediction of a target variable. The straight line equation y = mx + c, y = target/dependent variable, m = slope of the line, x = independent variable, c = y intercept.

In a data set we try to predict the value of y for a value of x by getting the best m and c values or the best fit line. The best fit line is the line that has the minimum deviation for the predicted and actual values, and the target is to obtain a line that fits all the data points with the minimum sum of squares.

2. Explain the Anscombe's quartet in detail. (3 marks)

→ Anscombe's quartet is a combination of four datasets with similar statistical properties (mean, standard deviation and correlation) yet different graphical representations. This gives us an idea on the importance of analysing any dataset graphically.

3. What is Pearson's R? (3 marks)

→ Pearson's R is used to determine the strength of relationship between two variables. If the relationship between two variables are:

- Close to +1, there is a strong positive relationship and a positive slope in the regression line.
- Close to -1, there is a strong Negative relationship and a Negative slope in the regression line.
- If 0, then there is no relationship between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

→ Scaling is a method applied those data in a dataset that have a very high variation in it. It is normalization of data to restrict its variation in a specific range.

→ Consider a dataset with a column 'length', the values in the said column are 10cm, 10m, 10mm…, etc. We might find ways to clean the data in the column by removing the string and retaining the integer values. But, that does not make sense and the magnitude of the data is the same, hence the unit also needs to be considered. The solution to this is scaling, we bring all the values in the column to a standardized unit.

→ Normalization scaling brings all the data in the range [0, 1], while standardization scaling replaces the values by their Z scores.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

→ The value of VIF is sometimes infinite because there exists a perfect correlation between the two variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

→ Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.