General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   ANS: Linear regression algorithm is a a machine learning algorithm based on supervised learning. It is used to find relationship between input and target variables

2. Explain the Anscombe's quartet in detail. (3 marks)
   ANS: Anscombe's quartet comprises 4 data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when plotted a graph. Each data set has 11 x,y points

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

   ANS: It measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviation. The result always lies between -1 and 1 as it is normalized measurement of covariance

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

   ANS: Scaling is a process done to normalize data to a particular range and speed up calculations in an algorithm. It is done to prevent incorrect modelling as algorithm takes only magnitude into account and not the units from the given data. Normalization/min-max scaling brings data to the range 0 to 1 whereas in standardized scaling it brings data into standard normal distribution with mean zero and standard deviation. Normalization loses outliers data

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

ANS: The VIF is infinite when two independent variables have a perfect correlation and r2=1. We need to drop one of variable which is not required from business prospective.

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

ANS: Q-Q plot is probability plot or graphical method to determine whether two samples of data came from same population or not. It is a plot of quantiles of one dataset with other to compare two probability distributions. where quantile is percentage of values below a given point