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

- Linear Regression is one of the most fundamental algorithms in the Machine Learning world which comes under supervised learning. Basically it performs a regression task. Regression models predict a dependent (target) value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used.

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

- **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.
- The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

- It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
- However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

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

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- Normalized scaling brings all of the data in the range of 0 and 1. where as Standard scaling Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables
- To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

- The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not
- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.