General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Introduction to Linear regression-

Linear regression may be defined as the statistical model that analyse the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

$$Y=mX+bY=mX+b$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the *Y*Y-intercept. If X = 0,Y would be equal to *b*b.

Furthermore, the linear relationship can be positive or negative in nature as explained below:

1)Positive Linear Relationship

A linear relationship will be called positive if both independent and dependent variable increases.

2)Negative Linear Relationship

A linear relationship will be called positive if independent increases and dependent variable decreases.

Types of Linear Regression

1)Simple Linear Regression

2)Multiple Linear Regression

2. Explain the Anscombe's quartet in detail

Ans: **Anscombe's Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots. It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analysing and model building, and the effect of other **observations on statistical properties**.

Application:

1)The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

2)This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Conclusion:

We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

Ans: **Correlation coefficients** are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson. **Pearson's correlation** (also called Pearson's *R*) is a **correlation coefficient** commonly used in linear regression. If you're starting out in statistics, you'll probably learn about Pearson's *R* first. In fact, when anyone refers to **the** correlation coefficient, they are usually talking about Pearson's.

For the Pearson r correlation, both variables should be **normally distributed** that is the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'. A simple way to do this is to determine the normality of each variable separately using the Shapiro-Wilk Test.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.
The most common techniques of feature scaling are Normalization and Standardization.
-Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data **unitless**.
Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

Few ways of feature Scaling:

1)Min Max Scaler
2) Standard Scaler
3) Max Abs Scaler
4) Robust Scaler
5) Quantile Transformer Scaler
6) Power Transformer Scaler
7) Unit Vector Scaler

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

Ans: If there is perfect correlation, then **VIF** = **infinity**. A large **value of VIF** indicates that there is a correlation between the variables. If the **VIF** is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

Ans: 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. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Importance: a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behaviour