1. Explain the linear regression algorithm in detail.

   Linear Regression is a machine learning algorithm based on **supervised learning**. It is used to predict the relationship between two variables by applying a linear equation to observed data. It is commonly used for predictive analysis. There are two types of variables – independent variable and dependent variable. The linear model is built on the set of independent variables and is used to predict the dependent variable.

   Example of linear regression: predicting the salary of an employee using variables like years experience, department, etc.

   The measure of the relationship between two variables is shown by the correlation coefficient. The range of the coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data between two variables.

The equation of Linear Regression is given as:

$Y=a+bX$

where X is the independent variable and it is plotted along the x-axis

Y is the dependent variable and it is plotted along the y-axis

b is the slope of the line, and a is the intercept (the value of y when x = 0).

A scatter plot can be used to visualize the relationship between variables.

In simple linear regression, we just have one independent variable, whereas in multiple linear regression, we have multiple independent variables and the equation is given as:

$Y=a+bX1+cX2+dX3…$

Properties of Linear Regression

For the best fit line where the parameters a and b are defined, the following properties are applicable:

- The best fit line reduces the sum of squared differences between observed values and predicted values.
- The best fit line passes through the mean of X and Y variable values.
- The regression constant a is equal to the y-intercept of the linear regression.
- The regression coefficient b is the slope of the regression line. Its value is equal to the change in the dependent variable (Y) for a unit change in the independent variable (X)

Explain the Anscombe's quartet in detail.

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

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 analyzing and model building, and the effect of other **observations on statistical properties**.

There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.

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.

What is Pearson's R?

Correlation is used in simple linear regression analysis to determine if two numeric variables are significantly related. A correlation analysis gives the **strength** and **direction** of the linear relationship between two variables

The Pearson correlation coefficient, r, measures the strength of correlation *and* can take on values between -1 and 1. The higher is the r, the stronger the linear relationship between the two variables. The sign of *r* corresponds to the direction of the relationship. If *r* is positive, then as one variable increases, the other tends to increase. If *r* is negative, then as one variable increases, the other tends to decrease. A perfect linear relationship (r=-1 or r=1) means that one of the variables can be perfectly explained by a linear function of the other.

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

Feature **Scaling** or Standardization: It is a step of Data Pre Processing which is applied to independent variables or features of data. It is a process in which different variables that have values in different ranges are bought to similar range of values. For e.g. if sales in in millions and units sold in hundreds then the values of sales and units are scaled to bring them into comparable units.

It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm by optimizing the gradient descent process.

Normalized scaling **is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.**

Standardized scaling **is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.**

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

VIF is infinite if there is a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.