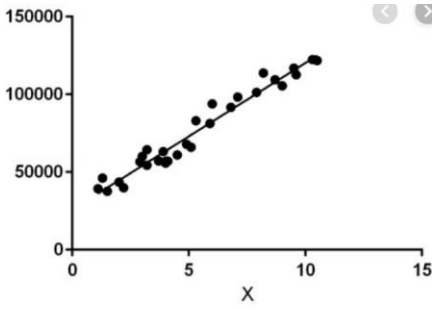
1. Explain the linear regression algorithm in detail.

   Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction 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 dependent and independent variables and the number of independent variables being used.



   STEPS:

   a. DATA PREPARATION
      i. In case of prediction, divide the data into train and test data set
      ii. In case of forecast, consider all the data as training dataset and forecast is been made on future data
   b. IDENTIFYING ESTIMATORS
      i. Identifying estimators based on analysis for using RFE
   c. BUILDING LINEAR REGRESSION MODEL
      i. Build a model based on identified estimators using OLS (Ordinary Least Square) after introducing constant if statsmodels.api is used
      ii. Check the summary and VIF of the model
      iii. Eliminate the estimators having high p-value (greater than 0.05)
      iv. If all the p-values are well within range, check VIF and eliminate estimators having VIF more than 5
      v. One thing to note is, we should remove/add estimators one by one
      vi. Repeat the above steps until we get a model with good R-square, Adjusted R-square, p-value and VIF
   d. PREDICTING DEPEND VARIABLE IN TRAIN DATA
      i. Predict the train data
   e. RESIDUAL ANALYSIS
      i. Calculate y_test - y_pred and plot the graph → the graph should be normally distributed with mean at zero
   f. EVALUATING MODEL
      i. Using the above generated linear model, predict the y_test values
      ii. Calculate the R-square for test data → the R-square for training and test data set should be close enough (difference should not be more than 5%)

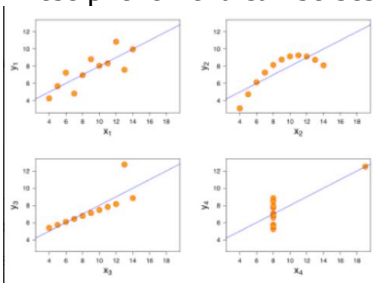
2. What are the assumptions of linear regression regarding residuals?
    - Normality assumption: Error Terms are normally distributed
    - Zero mean assumption: The mean of the residuals are assumed to be zero
    - Constant variance assumption: The variance of the residuals is same
    - Independent error assumption: The residual terms are independent of each other i.e. pairwise covariance is zero

3. What is the coefficient of correlation and the coefficient of determination?

    Coefficient of Correlation (R) is the degree of relationship between two variables. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way.
    Coefficient of Determination is the square of Coefficient of Correlation. It is generally used in multilinear regression to show percentage of variation in dependent variable which is explained by all independent variables together. Its value is from 0 to 1 and higher the value better it is.

4. Explain the Anscombe's quartet in detail.
    There are few shortcomings to linear regression
    - It is sensitive to outliers
    - It models the linear relationships only
    - A few assumptions are required to make the inference

    These phenomena can be best explained by the Anscombe's Quartet, shown below:

    From above image, it is evident that the model is same in all the four cases. In the first case it does pretty good job in predicting while in second case it is incapable to handling any other kind of data. In third and fourth case, there are outliers present and hence it fails to address them.

5. What is Pearson's R?
    Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. It has a value between +1 and −1,

where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

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

   Scaling is a technique in Machine Learning used for converting the values of independent variables to a certain scale.
   The range of raw values varies widely, to bring them to acceptable scale is called scaling.

   Since values can be wide spread for different independent variables, the reason behind performing scaling is to ensure one feature does not dominate over other and while plotting the graph, we can observe clear pattern.

   Normalization: Normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

   Standardization: It rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

   In case where the independent variables are perfectly correlated, we observe infinite VIF. Suppose there are two variables A and B having 1 as the correlation value, then the VIF of A and B would be infinite.

8. What is the Gauss-Markov theorem?

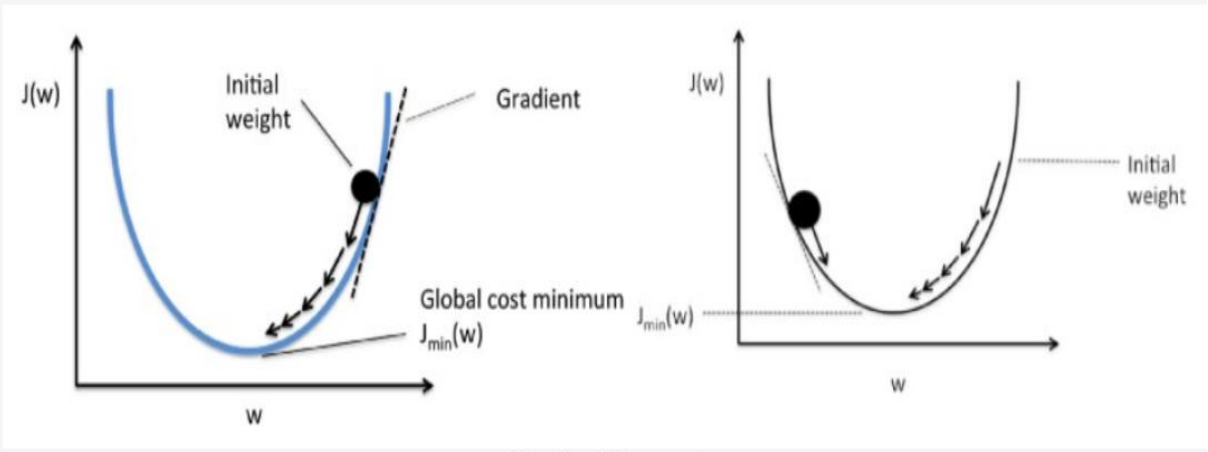
   The Gauss-Markov theorem states that if your linear regression model satisfies the first six classical assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators.

   The Gauss-Markov theorem famously states that OLS is BLUE. BLUE is an acronym for Best Linear Unbiased Estimator.

   In this context, the definition of “best” refers to the minimum variance or the narrowest sampling distribution. More specifically, when your model satisfies the assumptions, OLS coefficient estimates follow the tightest possible sampling distribution of unbiased estimates compared to other linear estimation methods.

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model.
In linear regression, it is used to optimise the cost function and find the values of the βs (estimators) corresponding to the optimised value of the cost function.
Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).
Gradient Descent starts with a random solution, and then based on the direction of the gradient, the solution is updated to the new value where the cost function has a lower value.



Gradient Descent

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight