1. Explain the linear regression algorithm in detail.

   Linear regression is the most well-known and well understood algorithm. It is an attractive model because the representation is very simple. Linear regression is a predictive modeling by which it can predict the relationship b/w dependent variable(target) and independent variable(predictors). There are 2 types of linear regression 1) single linear regression 2) multiple linear regression

   Single linear regression:

   this explains the relationship b/w a single independent variable and target variable with the equation of y = b0 + b1*x. The best model should have best fit line by minimizing the expression (RSS – Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residual for any point is found by subtracting predicted value of independent variable from its actual value. There are 2 metrics. 1) R2 (Coefficient of Determination) and 2) RSE (Residual Standard Error).

   There are below steps involved.

   a. Importing the dataset.
   b. Splitting dataset into training set and testing set (2 dimensions of X and y per each set). Normally, the testing set should be 5% to 30% of dataset.
   c. Visualize the training set and testing set to double check
   d. Initializing the regression model and fitting it using training set (both X and y).
   e. calculate the coefficient
   f. predict the values in test data
   g. check r^2 value and mean square error

   Multiple linear regression:

   This has multiple independent variable and one target variable. We need to make sure that we do data cleaning before starting to model and creating some dummy variable is a important part of it. After splitting into test and train data, following are the steps involved,

   a. normalization/standardization might be needed
   b. build the model using RFE.
   c. check the VIF values and p-values and keep repeating the process until VIF values <5 and p value is < 0.05 which implies that they are statistically significant.

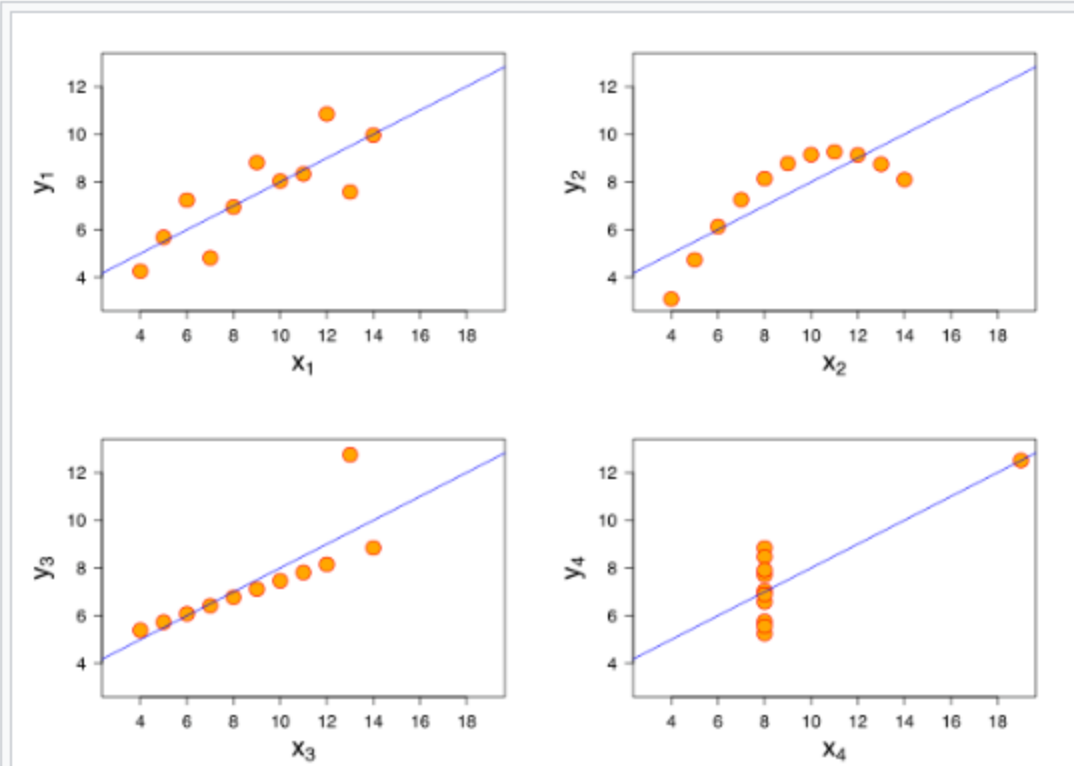
2. What are the assumptions of linear regression regarding residuals?

   There are four assumptions associated with a linear regression model:

a. Linearity: The relationship between X and the mean of Y is linear.
b. Homoscedasticity: The variance of residual is the same for any value of X.
c. Independence: Observations are independent of each other.
d. Normality: For any fixed value of X, Y is normally distributed.

3. What is the coefficient of correlation and the coefficient of determination?
   coefficient of correlation (r): it measures the strength and direction of linear relationship b/w 2 variables. The value of r lies b/w -1 and 1 including both. – and + signs show the negative and positive correlation. If there is no relation, then r would be 0. A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak.
   coefficient of determination(r2): this is useful in order to give proportion of variance of one variable that is predictable from another variable. It lies b/w $0 < r\ 2 < 1$. It represents the percent of the data that is the closest to the line of best fit. It also measures of how well the regression line represents the data.

4. Explain the Anscombe's quartet in detail.
   It comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. It uses to mimic the importance of looking dataset graphically before analyzing, seeing the data virtually to find a relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

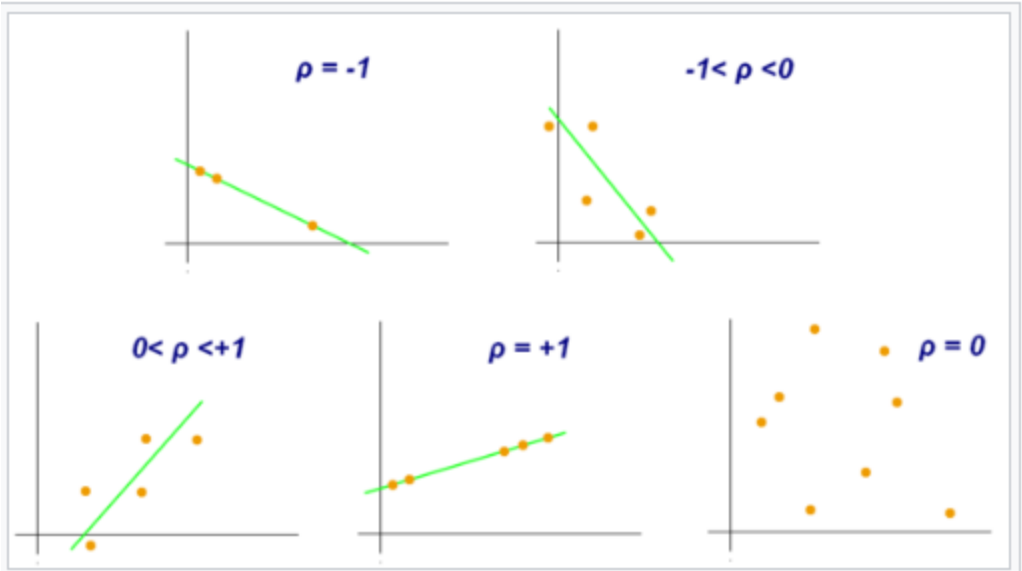


All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

- The top left is a simple linear relationship, where there is a gaussian relationship b/w 2 variables
- The top right is a nonlinear relationship
- The bottom left graph, the distribution is linear, the calculated regression is offset by outlier and the correlation coefficient from 1 to 0.816
- The bottom right shoes the example when one high leverage point is enough to produce a high correlation coefficient.

5. What is Pearson's R?

   The Pearson's R is also known as bivariate correlation. It is a statistic that measures linear correlation b/w 2 variables X and y. it has values b/w +1 and -1. A positive 1 indicates that there is a positive relationship b/w X and Y and a negative 1 indicates that there is a negative relationship b/w X and Y. 0 is no linear correlation.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Feature scaling is an important technique in ML and this can make a difference b/w weak ML model and a strong ML model. This is the most important steps during data preprocessing.
   Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. They are 2 most important scaling techniques: Standardization and Normalization.
   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: typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
   Normalization is good when you know that the distribution of the data does not follow Gaussian distribution whereas Standardization can be helpful in dataset which follow Gaussian distribution. But that can't be true in all the cases. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   An infinite VIF value shows that the corresponding variable may be defined exactly by a linear relationship with other variables. VIF of X1 = 1/(1-R2 of X1 on all other Xs). If R2 values is 1 then VIF can be infinite. And VIF can be infinite when there is a prefect collinearity of a variable with all other variable and when we have chosen completely redundant variables.

8. What is the Gauss-Markov theorem?
   Gauss Markov theorem states that the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible if the following assumptions are met.

   Linearity: the parameters should be linear.
   Random: data must have been randomly sampled from the population.
   Non-Collinearity: the regressors being calculated shouldn't be perfectly correlated with each other.
   Exogeneity: the regressors aren't correlated with the error term.
   Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

   The above assumptions make sure that the validity of the ordinary leat squares for the estimating regression coefficients. In practice, the Gauss Markov's assumptions are very hard to meet, but they are still useful as a benchmark.

9. Explain the gradient descent algorithm in detail.
   Gradient Descent is an optimization algorithm used to minimize some function by riteratively moving in the steepest direction as defined by negative gradient. In linear regression, we use gradient descent as parameter referred as coefficient.
   a. The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.
   b. The coefficients' cost is evaluated by plugging them into the function and calculating the cost.
   c. The cost's derivative is calculated. It is refer as the slope of the function at a given point. The slope indicates the direction in which we can move the coefficient values in order to get a lower cost on the next iteration
   d. The derivate is downhill, we can update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.
   e. We can repeat the process until the cost of the coefficients is 0.0 or close to 0.

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

    A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. 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. It also helps in determining the data sets come from the same common distribution from population.
    This helps in linear regression when we have training and test data set which are received separately and then we can confirm that both come from same distribution of the populations.
    It can be used with sample sizes also. Many outliers can be detected from this plot. It can check that the 2 data sets have common location and scale, similar distributional shapes and similar tail behavior.
    Below are the possible interpretations for two data sets.

a. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b. Y-values < X-values: If y-quantiles are lower than the x-quantiles.
c. X-values < Y-values: If x-quantiles are lower than the y-quantiles.
d. Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis