1. Explain the linear regression algorithm in detail.
   Linear Regression algorith is an algorithm based on supervised learning that is widely used in Machine Learning process.
   Linear Regression process models a target variable for prediction based on one or more independent variables.
   Based on the number of independent variables it can classified into two categories:
   1) Simple Linear Regression Model
   2) Multiple Linear Regression Model
   It is mostly used for finding out the relationship between target variables and independent variables.

   In any linear regression process we perform the task to predict a dependent variable value (y) based on a given independent variable (x).
   This tries to fit the model using the line/hyperplane eqution

   $y = mx + c$ (for simple Linear Regression Technique)

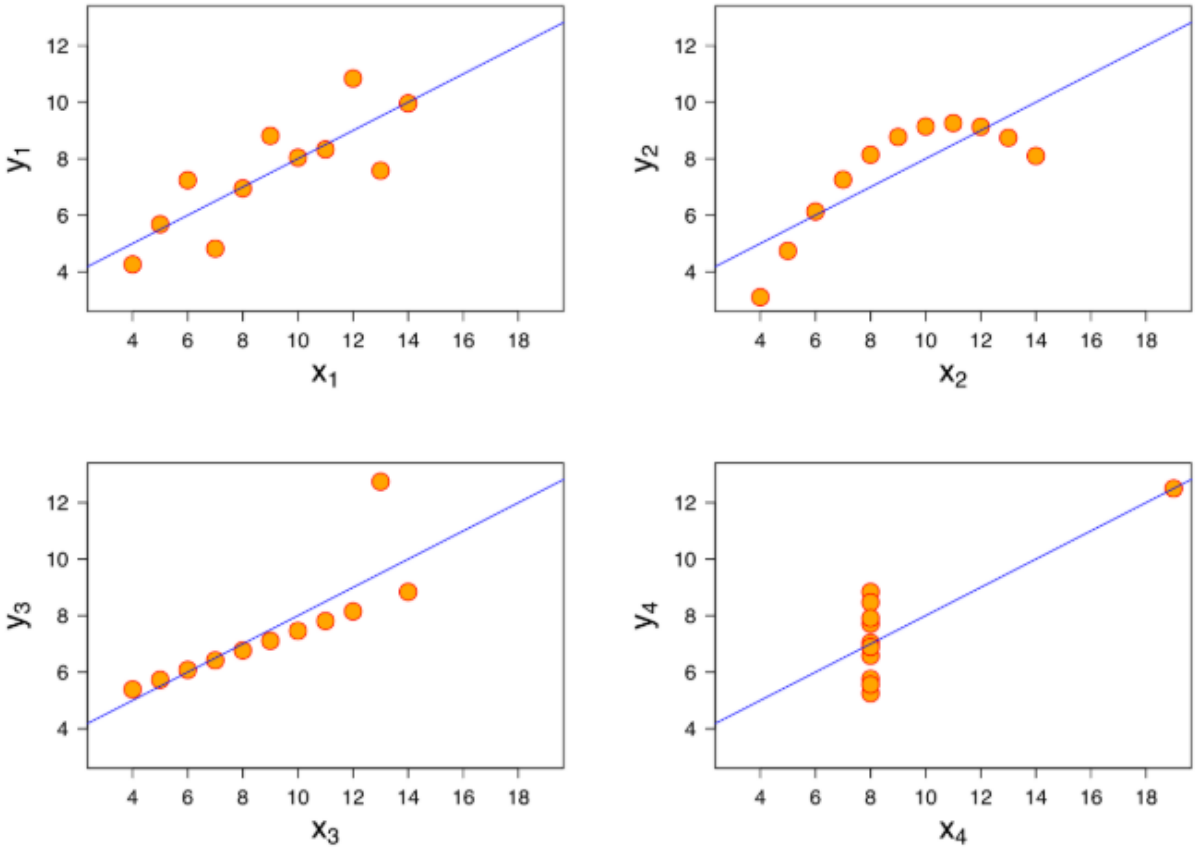
   $y = m1x1 + m2x2 + m3x3 + .. + mnXn + c$ (for multiple Linear Regression Technique)

2. What are the assumptions of linear regression regarding residuals?
   Assumptions of simple linear regression residuals are:
   1. Residuals or error terms are normally distributed
   2. Residuals or error terms are independent of each other
   3. Residuals or error terms have constant variance (homoscedasticity)

3. What is the coefficient of correlation and the coefficient of determination?
   Coefficient of correlation:
   A correlation coefficient is a numerical measure of correlation or a statistical relationship between two variables.
   There are several types of correlation coefficient exist. Each with their own definition and own range of usability and characteristics.
   All correlation coefficient values assume values in the range from −1 to +1, where ±1 indicates the strongest possible agreement and 0 the strongest possible disagreement.

   Coefficient of determination
   The coefficient of determination, denoted R2, is a number which explains what portion of the given data variation is explained by the developed model.
   It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information.
   $R^2 = 1 - (RSS / TSS)$
   Where RSS = Residual sum of squares
   TSS = Total sum of squares

4. Explain the Anscombe's quartet in detail.
   Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises of four datasets, each containing eleven (x,y) pairs.
   The significant thing to note about these datasets is that they share the same descriptive statistics. But things change when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



   This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

5. What is Pearson's R?
   A Pearson correlation® is a number between -1 and 1 that indicates the extent to which two variables are linearly related. 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.
   Formula for Pearson R

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

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

   Scaling is a process that is used to build features that have similar ranges to each other.
   It helps analysts to do the analysis of two correlated variables which are of completely different magnitude and scale.
   Normalizing a feature to a [0,1] range, through (x−min(x))/(max(X)−min(x))
   Standardizing the feature (also referred to as z-score), through (x−μ)/σ, where μ is the mean and σ is the standard deviation.

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

   VIF represents the measure the multi-collinearity among the predictor variables.
   Higher the value of VIF, higher the co-linearity of the predictor variables.
   So one of the interpretation of the VIF being infinite is that the predictor variables highly co-linear or they can be represented as an exact linear combination of other variables like below:
   X1 = X2 or X1 = X2 + X3

   Another interpretation of VIF being infinite value can be following:

   Formula for VIF

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

   VIF will become infinite if R2 is equal to 1 i.e. the fitted line is covering 100% percent of the training points. Which essentially means that the model has over-fitted the data points by remembering all of them.

8. What is the Gauss-Markov theorem?

   Gauss-Markov theorem states that Ordinary Least Square (OLS) regression produces Best Linear Unbiased Estimator (BLUE).
   Here the term “best” refers that the sampling distribution will have the minimum variance among all the unbiased liner estimators.
   It can be summarized using the following equations:
   Let the linear regression model be represented as $y_i = x_i \beta + \varepsilon_i$ if it is generated by OLS then it wil satisfy the following conditions:
   1. $E\{\varepsilon_i\} = 0$, i = 1, . . . , N i.e. expected value of the errors is zero
   2. $\{\varepsilon_1 \ldots\ldots \varepsilon_n\}$ and $\{x_1 \ldots.., x_N\}$ are independent, i.e. all the error term and x values are linearly independent
   3. $\mathrm{cov}\{\varepsilon_i, \varepsilon_j\} = 0$, i, j = 1,…., N I ≠ j. i.e. all the error terms are un-correlated
   4. $V\{\varepsilon_1 = \sigma^2$, i= 1, ….N i.e. the error terms has equal variances

   And it will be a the Best Linear Unbiased Estimator(BLUE)

9. Explain the gradient descent algorithm in detail.

   Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function.
   It is used to minimizing the cost function in machine learning algorithm. Usually the cost function will be the sum of least square.
   After many iteration, if the cost doesn't improve then we will reach at the state of convergence. The value of the parameters at that very last step is known as the optimum set of parameters (in the case of the linear regression model, we have the optimized value for both Betas).
   Types of gradient Descent:

   Batch Gradient Descent: This is a type of gradient descent which processes all the training examples for each iteration of gradient descent.

   Stochastic Gradient Descent: This is a type of gradient descent which processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is quite faster than batch gradient descent.

   Mini Batch gradient descent: This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Here b examples where b<m are processed per iteration. So even if the number of training examples is large, it is processed in batches of b training examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    A Q-Q plot or qualtile-quantile 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 will be nearly straight.

    The advantages of the q-q plot are:

    The sample sizes do not need to be equal.
    Many distributional aspects can be simultaneously tested. For example, shifts in scale, changes in symmetry, the presence of outliers can all be detected from this plot.
    Importance in Linear Regression Model: If the plot don't lie in a line then the residuals aren't Gaussian and thus the errors aren't either. This implies that for small sample sizes, we can't assume that the estimator is Gaussian either, so the standard confidence intervals and significance tests are invalid.