## 1. Explain the linear regression algorithm in detail?

**Linear Regression** is a machine learning algorithm based on **supervised learning.** This method is mostly used for forecasting and finding out cause and effect relationship between variables. To predict a target value based on independent predictors.

Linear regression algorithm finds the best linear-fit relationship between independent and dependent variables from the given data. Based on the data, it learns a hypothesis function which fits the data well. It models a relationship between the dependent variable and explanatory variables.

A linear regression model with two variables can be expressed with the following equation:
y = B0 + B1*X1 + B2*X2 + e.

The variables in the model are:
y is the target variable.
X1 is the first input variable.
X2 is the second input variable and
e is the residual error, which is an unmeasured variable.

The parameters in the model are:
B0 is the Y-intercept
B1 is the first regression coefficient and
B2 is the second regression coefficient.

It uses Sum of Squared Residuals Method to calculate the error terms and find the best fit. An optimization algorithm Gradient descent is used to find the value of optimal parameters (Beta coefficients) by minimizing the cost function.

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

It is assumed that the residuals have a mean of zero. Residuals form a normal distribution centered around zero when plotted. In general, there are no identifiable patterns observed on the distribution of residuals. residual terms are Homoscedasticity having the same variance. Also, residuals are independent of each other.

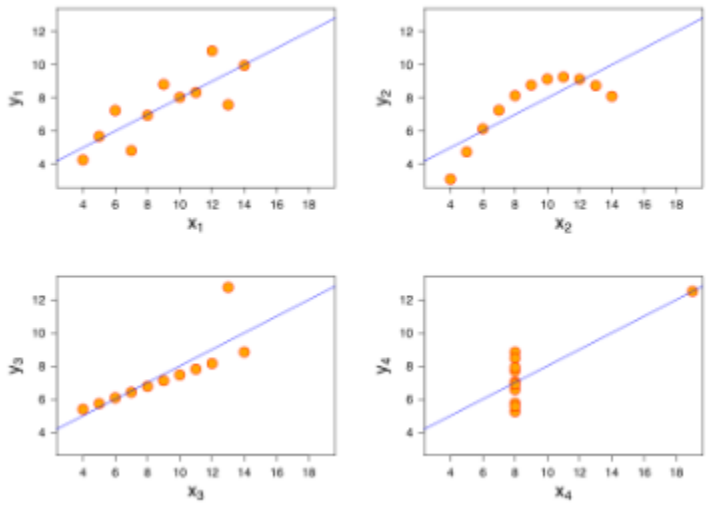
## 3. What is the coefficient of correlation and the coefficient of determination?

Coefficient of Correlation is also called as R tells us the relationship between two variables say x and y. It can go between -1 and 1. Whereas -1 and 1 shows a strong relationship between variables. Value 0 tells they are not correlated.

Coefficient of determination also called R-square; it is a squared value of R Coefficient of correlation). Since it is a squared value it is always between 0 and 1. It can never be negative. It shows the strength of a linear regression, how good is the fit. Because in general we use multiple variables in linear regression we need to see the overall fit irrespective of the positive and negative correlation

## 4. Explain the Anscombe's quartet in detail?

Anscombe's Quartet explains us "we should never just run a regression without having a good look at your data". This was explained by Anscombe's Quartet four datasets plotted on X,Y coordinate plane with same mean and variance. Each plot tell us a different story. 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.
Plot 1 appears to have clean and well-fitting linear models.
Plot 2 is not distributed normally.
Plot 3 the distribution is linear, but the calculated regression is thrown off by an outlier.
Plot 4 shows that one outlier is enough to produce a high correlation coefficient.



## 5. What is Pearson's R?

Pearson's R is the normal correlation coefficient R we use in the linear regression two find the linear relationship between two variables. But there are certain limitations when using this. This Pearson's R is designed for linear relationships and it might not be a good measure if the relationship between the variables is non-linear.

Though there are different techniques like Spearman's R to determine the correlation if the relationship between the variables is not linear. Though Pearson's R still give a correlation coefficient for non-linear relationships, it is not be reliable.

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

Scaling is used to standardize the independent features present in the data in a fixed range.

Scaling helps with ease of interpretation. If we have variables with different scales and ranges. Scaling makes the comparing coefficients easier. It will also fasten the convergence for gradient descent method.

Normalized scaling transforms the data into a range between 0 and 1.

Standardization scaling transforms the data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

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

Variance influence Factor (VIF) calculates how well one independent variable is explained by all the other independent variables combined.

If there is perfect correlation, then VIF = infinity.

In VIF, each feature is regression against all other features. If R2 of feature is more which means this feature is correlated with other features. While calculating the VIF = 1 / (1 – R2). IF R2 reaches 1, VIF reaches infinity which signifies a perfect correlation.

The common heuristic is removing features for which VIF > 5.

## 8. What is the Gauss-Markov theorem?

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

Some Classical assumptions

a. Linearity: the parameters we are estimating using the OLS method must be themselves linear.
b. Random: our data must have been randomly sampled from the population.
c. Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
d. Exogeneity: the regressors aren't correlated with the error term.
e. Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

## 9. Explain the gradient descent algorithm in detail?

Gradient descent is an optimization algorithm used to find optimal parameters (coefficients) of a function that minimizes a cost function. Gradient descent is used when the parameters cannot be calculated analytically and must be searched for by an optimization algorithm.

Gradient descent works like a ball rolling down a graph. The ball moves along the direction of the greatest gradient and comes to rest at the flat surface.

The goal is to continue to try different values for the coefficients, evaluate their cost and select new coefficients that have a slightly better (lower) cost. Repeating this process enough times will lead to the bottom of the surface and we will know the values of the coefficients that result in the minimum cost.

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.

coefficient = 0.0

The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient)

or

cost = evaluate(f(coefficient))

The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.

delta = derivative(cost)

Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

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.

## 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. 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.