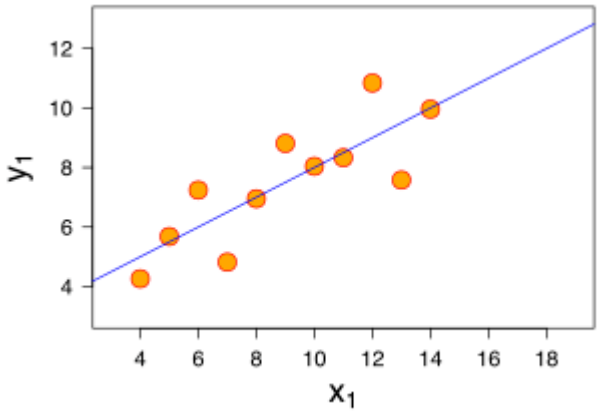
1. **Explain the linear regression algorithm in detail.**

Ans:

- Linear regression is a linear model, where a model that assumes linear relationship between the input variables (x) and single output variable(y).
- When there is a single input variable then it is called a simple linear regression. When there are multiple input variables then it is called multiple linear regression
  Ex: y = B0 + B1*x - > simple linear regression with single input variable X
  $f(x,y,z)=w1x+w2y+w3z$ - > multiple linear regression with multiple input variables (X,Y,Z)
- Different techniques are used to train the linear model, the most common one used is Ordinary least squares.
- Best fit Line – the line that minimizes the sum of squares of distances of points from the regression line

  An example of Linear regression model with best fit line.



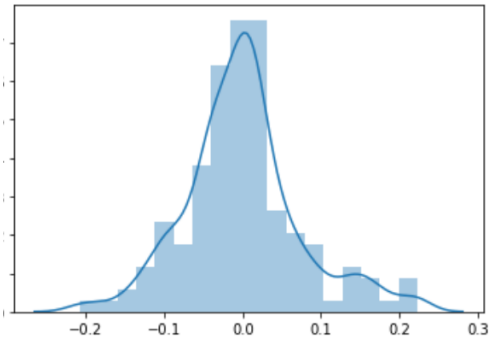
- *R2*, called the coefficient of determination, is used to evaluate how good the fit of the regression model is, it is calculated by ESS/TSS. i.e, the ratio of explained variance by total variance
- ESS – Is known as explained variance, it is the portion of total variation that measures how well the regression equation explains the relationship between X and Y

  $$ESS = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$
- Residual sum of squares (RSS): This expression is also known as unexplained variation and is the portion of total variation that measures discrepancies (errors) between the actual values of Y and those estimated by the regression equation

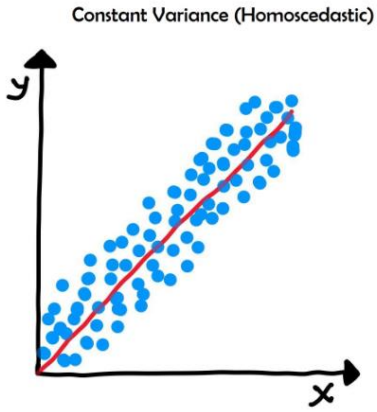
  $$RSS = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$
- Total sum of squares TSS = RSS+ESS
  R2 is the ratio of explained sum of squares (ESS) to total sum of squares (TSS):
  R2 = ESS/TSS

2. **Asumptions of Linear regression**

- Residuals follow the normal distribution with mean at 0. In the below histogram of residuals, we can see that they are normally distributed with mean at 0.



- Residuals have a constant variance like below. The variance should not have any visible pattern



- Regression function is linear.
- A pattern does not exist when residuals are plotted in a time or run-order sequence.
- There are no outliers.

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

Coefficient of determination :

R square or coefficient of determination shows percentage variance in y which is explained by all the x variables together. Higher the value, better the model. It is always between 0 and 1. It can never be 0

Coefficient of Correlation:

Is the degree of relationship between two variables x and y. 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.

- From regression point of view, correlation can be rightfully explained for simple linear regression, because it has only one input and one output variable.
- For multiple linear regression, R square is a better term.
- We can explain R square for both simple linear regressions and also for multiple linear regressions.

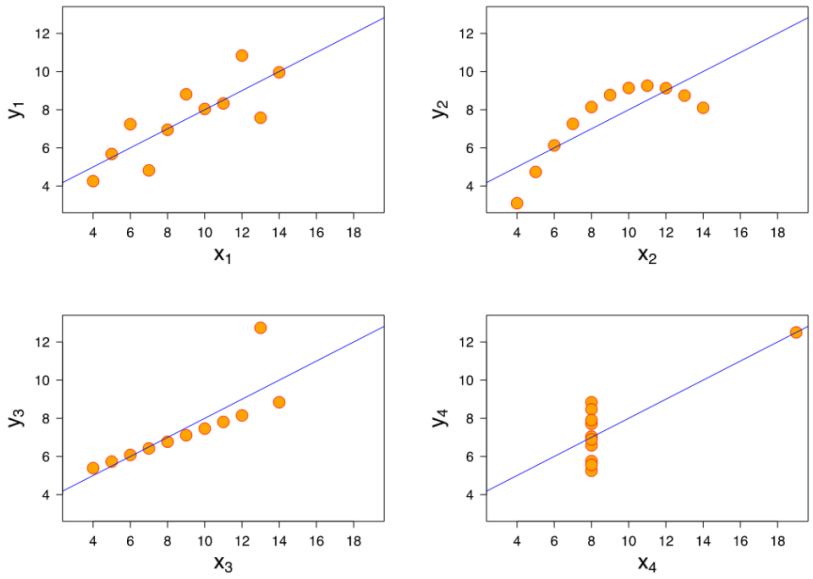
4. Explain the Anscombe's quartet in detail?

Anscombe's quartet is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven *(x,y)* pairs

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

These four datasets appear to be very similar but they tell a different story when plotted in a graph. They are shown as below.



Below are our observations based on the above graph

- Dataset I appears to be clean and well fitting linear model
- Dataset II is not normally distributed
- Dataset III is distributed linear, but an outlier throws off the calculated regression
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient

5. **What is Pearson's R**?

   Pearson correlation coefficient , also referred to as Peasron's R is a measure of the linear correlation between two variables X and Y

   - It is the test statistic that measures the association or relationship between two continuous variables
   - It is defined by covariance of two variables divided by the product of their standard deviations

     Below is the formula

     $$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

   - The values range between -1 and 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables

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

   - Feature scaling is a technique to standardize the independent features present in the data to fixed range.
   - It is performed at pre-processing stage to handle highly varying magnitude or values or units.

   Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

   Normalized Scaling: This techniques re-scales a variable values with distribution value between 0 and 1.

   $$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

   Standardized scaling: It is a technique which re-scales a feature value so that it has distribution with 0 mean and variance as 1.

   $$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

   In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated to other features

   VIF= 1/(1-R2)

   The value of VIF becomes infinity if R2 becomes 1, that means the variables are highly correlated.

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

   The **Gauss Markov theorem** tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate *(BLUE)* possible.

   Below are the Gauss Markov assumptions:

   - **Linearity** – The parameters we are estimating using OLS must be linear in nature
   - **Random** – Data must be randomly sampled
   - **Non -correlation** – The variables are not perfectly correlated with each other
   - **Exogeneity** – The variables are not correlated with error term
   - **Homoscedasticity** – Irrespective of variable values, the variance of errors is constant

9. **Explain the gradient descent algorithm in detail**

   Gradient descent is an optimization technique to find the values of coefficients of a function that minimizes a const function

   Example:
   you are at the top of a mountain, and you have to reach a lake which is at the lowest point of the mountain. A twist is that you are blindfolded and you have zero visibility to see where you are headed. So, what approach will you take to reach the lake?

   The best way is to check the ground near you and observe where the land tends to descend. This will give an idea in what direction you should take your first step. If you follow the descending path, it is very likely you would reach the lake.

   Procedure:
   - The procedure starts with initial values of coefficients it could be 0.0
     Coefficient= 0
   - The cost of coefficients are calculated by plugging them into a function and calculating cost
     Cost=f(coefficient)

- The derivative of the cost is calculated. Derivative refers to slope of the function at the given point. We need to know the slope so that the direction of the to move the coefficient values in order to get a lower cost the next iteration

  Delta=derivative(cost)
- Since we know the delta, we can now update the coefficient values. A learning rate parameter(alpha) must be specified that controls how much the coefficient can change in each step
  Coefficient =coefficient – (alpha*delta)
- This process is continued until the cost of coefficient is 0 or close to zero

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 help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

Importance of a Q-Q plot in linear regression:

The plot will plot on x-axis the quantiles of one variables, on y axis the quantiles of another variable. This method is better used to confirm our below assumptions for linear regression

- Are the error terms (residuals of the regression) normally distributed?

To prove this, we can plot Q-Q plot and if the quantiles of the error terms are near enough to the quantiles of the corresponding values computed from the normal distribution, then we might begin to accept the idea that the error terms are normally distributed

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. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.