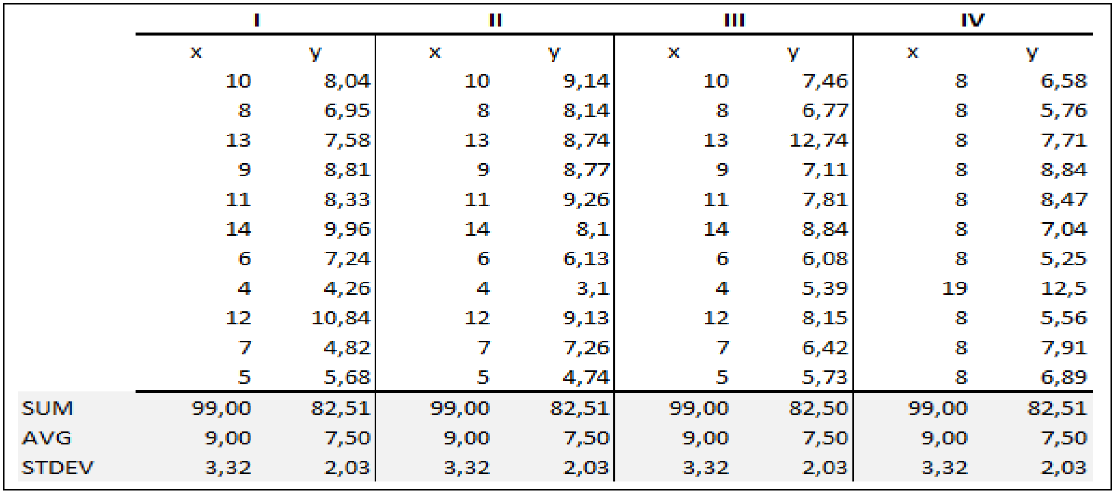
1. **Explain the linear regression algorithm in detail.**

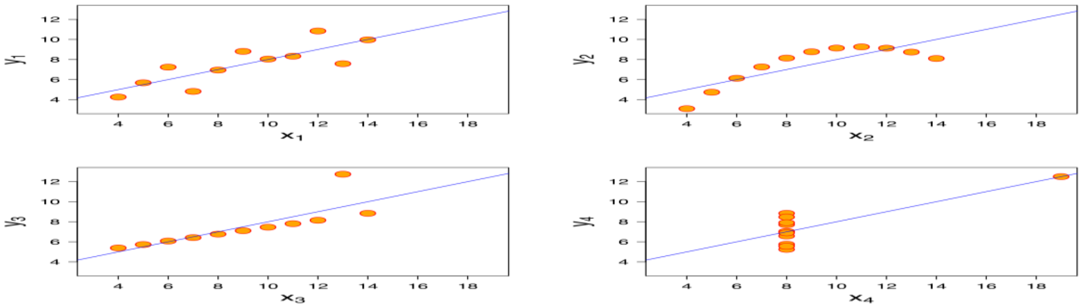
   Linear regression is an approach to design a model that describe linear relationship between the input variables (x) and the single output variable (y). we can also say that it is defining the linear relationship between dependent variable and one or more independent variable. If it defines relation between one input and one output, it is known as linear regression. If defined relation is between one output and two or more input combinations . Then this kind of relationship is known as multiple linear regression. The problem of linear regression is finding the best fit line. It can be found by using OLS(ordinary least square).It minimises the average error .The line with least error is known as the best fit. The parameters are being defined by using OLS. In computer it can be found by using Gradient Descent algorithm. Once the slope and intercept is found, it can be used for prediction purpose. eg :Y=mX+c if m and c is known then we can find out Y.
2. **What are the assumptions of linear regression regarding residuals?**
   1. **Normality Assumption:**
      Error Terms are normally distributed.
   2. **Zero Mean Assumption:**
      Error terms are normally distributed around zero.
   3. **Constant variance assumption:**
      Residual terms has constant variation. It is also known as assumption of homogeneity or homoscedasticity.
   4. **Independent error assumption:**
      Residual terms are independent of each other ,i.e. their pairwise covariance is zero.
3. **What is the coefficient of correlation and the coefficient of determination?**

   Coefficient of correlation explain about the correlation between two variables. It means that it tells about the impact of change in one variable if another variable changes. Coefficient of determination is the square of the coefficient of correlation. In case of simple linear regression it is R-squared value.
4. **Explain the Anscombe's quartet in detail.**

   Anscombe's Quartet was developed by statistician Francis Anscombe. It is collection of four dataset as (X,Y) pair. All of them share the same descriptive statistics. Overall it shows the importance of data viasulization.



	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



- From first dataset and first figure it seems to have clean and well-fitting linear models
- From second dataset and image it can be found that dataset is normally distributed.
- From third dataset and image it can be found that it seems to have linear regression but comes out with outliers.
- In last one it has an outlier that can contribute large correlation coefficient.

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

Pearson product-moment correlation coefficient (often abbreviated as Pearson's r), which measures the linear dependence between pairs of features. The correlation coefficients are in the range –1 to 1. Two features have a perfect positive correlation if r = 1, no correlation if r = 0, and a perfect negative correlation if r = –1. As mentioned previously, Pearson's correlation coefficient can simply be calculated as the covariance between two features, x and y (numerator), divided by the product of their standard deviations (denominator).

There are two types of correlation coefficient one is Pearson's R correlation coefficient which is mostly used in linear regression model. For non linear relationship it is not good choice in that case the coefficient correlation are used known as Spearman's R

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

Scaling is used to transform the data at the time of pre –processing of data so that data variation for different columns comes under a single scale. Also known as Normalization of data. It is being done at the time of data preparation stage for model building.

There are mainly two methods for scaling data.

- Normalization
  It scales numerical variable in the range of (0,1)

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Standardization:
  Transform all the data points in normal distribution with zero mean and standard deviation one.

$$x_{new} = \frac{x - \mu}{\sigma}$$

Drawbacks: If dataset has outliers then applying Normalization bring all data points under very small range. On the other applying Standardization new data doesn’t comes under any range or without boundation..

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

   The formula of VIF is

   $$VIF = 1/(1-R^2)$$

In VIF finding calculation of VIF for one independent variable is done with the help of other independent variable if the R^2 is equal to 1 then in this case VIF becomes infinite. Since VIF shows multi-collinearity if any how one variable is highly correlated with all other variables. Then in that case it becomes infinite.

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.

   BLUE – Best Linear Unbiased Estimator

   Gauss Markov Assumptions There are five Gauss Markov assumptions (also called conditions):

   1. Linearity: the parameters we are estimating using the OLS method must be themselves linear.

   2. Random: our data must have been randomly sampled from the population.

   3. Non-Collinearity: the regressors being calculated aren’t perfectly correlated with each other.

   4. Exogeneity: the regressors aren’t correlated with the error term.

   5. Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

   Purpose of the Assumptions:
   The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.
   Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.
   In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what ‘ideal’ conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

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

   Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

   Types of gradient Descent:

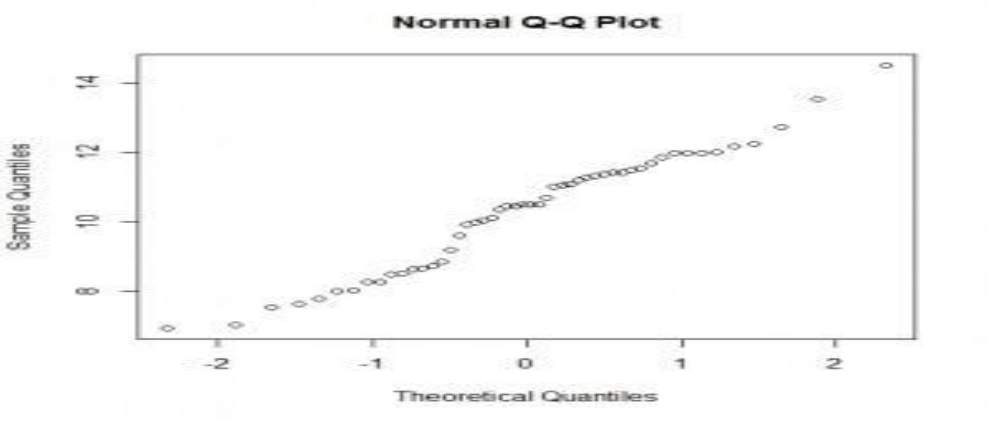
   - **Batch Gradient Descent:** This is a type of gradient descent which processes all the training examples for each iteration of gradient descent. But if the number of training examples is large, then batch gradient descent is computationally very expensive. Hence if the number of training examples is large, then batch gradient descent is not preferred. Instead, we prefer to use stochastic gradient descent or mini-batch 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. But again, when the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.
- **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?**

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



d