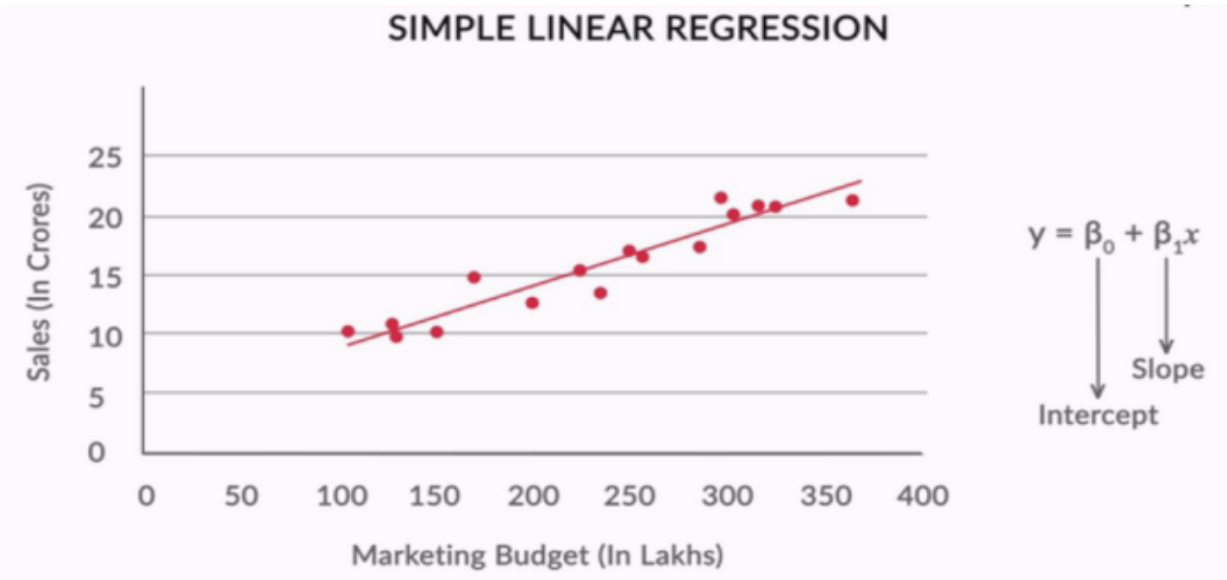
1. **Explain the linear regression algorithm in detail.**

Regression is a method of modelling a target value based on independent predictors. This method is used to forecast and find out the relationship between these independent variables to predict the target value.
Linear regression algorithm is one such method which is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple linear regression.



Simple Linear Regression

Simple linear regression model explains the relationship between a dependent variable and one independent variable using a straight line.

$Y = \beta_0 + \beta_1.X$

where,
Y - Dependent variable
X - Independent variable
βo - Intercept (βo being a constant that is to be determined. It is referred to as the intercept because, when X is 0 then Y = βo)
β1 – Slope (βo being a value that is to be determined. It is referred to as the coefficient, and is sort of like magnitude of change that Y goes through when X changes.)

Multiple Linear Regression

The multiple linear regression explains the relationship between **one continuous dependent variable**(*Y*) and **two or more independent variables**(*x*1, *x*2, *x*3... etc.)**.**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

y - Dependent variable
x1, x2..... Xn - Independent variable
βo - Intercept (βo being a constant that is to be determined. It is referred to as the intercept because, when X is 0 then Y = βo)
β1 – The coefficient for X1(the first feature)
β2 – The coefficient for X2(the second feature)
βn – The coefficient for Xn(the nth feature)

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

The important assumptions in regression analysis are

- Relationship between your independent and dependent variables should always be linear i.e. you can depict a relationship between two variables with help of a straight line.
- Mean of residuals should be zero or close to 0 as much as possible. It is done to check whether our line is actually the line of “best fit”.

  Residual = Observed – Predicted

  We want the arithmetic sum of these residuals to be as much equal to zero as possible.

- There should be homoscedasticity or equal variance in our regression model. This assumption means that the variance around the regression line is the same for all values of the predictor variable (X).
- All the dependent variables and residuals should be uncorrelated.
- The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
- There should be no perfect multicollinearity in your model. Multicollinearity generally occurs when there are high correlations between two or more independent variables. In other words, one independent variable can be used to predict the other. This creates redundant information, skewing the results in a regression model. We can check multicollinearity using VIF(variance inflation factor). Higher the VIF

for an independent variable, more is the chance that variable is already explained by other independent variables.
- Residuals should be normally distributed

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

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient: **Pearson's correlation** (also called **Pearson's R**) is a correlation coefficient commonly used in linear regression.

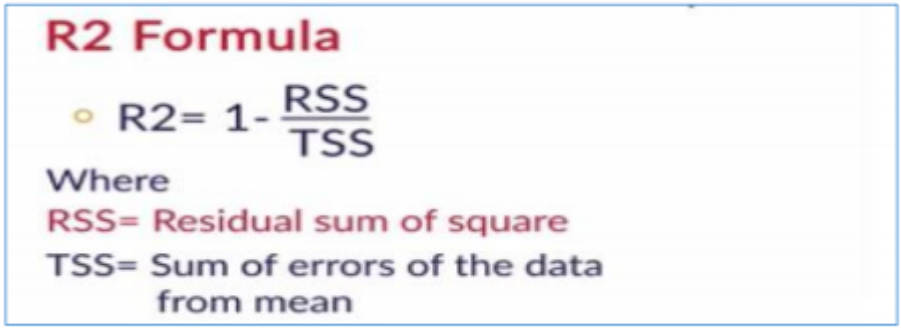
Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,]\,[\, n\sum y^2 - (\sum y)^2 \,]}}$$

For example, if Correlation coefficient of Gold and crude is 1 which means they are highly correlated, hence a rise in the price of oil increases gold rate

The coefficient of determination, R2, is used to analyze how differences in one variable can be explained by a difference in a second variable. More specifically, R-squared gives you the percentage variation in y explained by x-variables. The range is 0 to 1 (i.e. 0% to 100% of the variation in y can be explained by the x-variables)
In general term, it provides a measure of how well actual outcomes are replicated by the model. Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as: R 2 = 1 - (RSS / TSS)



Where,

RSS (Residual Sum of Squares): The total sum of error across the whole sample.

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable.

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

For example, if r-squared = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.

4. **Explain the Anscombe's quartet in detail.**

Anscombe's Quartet was developed by statistician **Francis Anscombe**. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics.

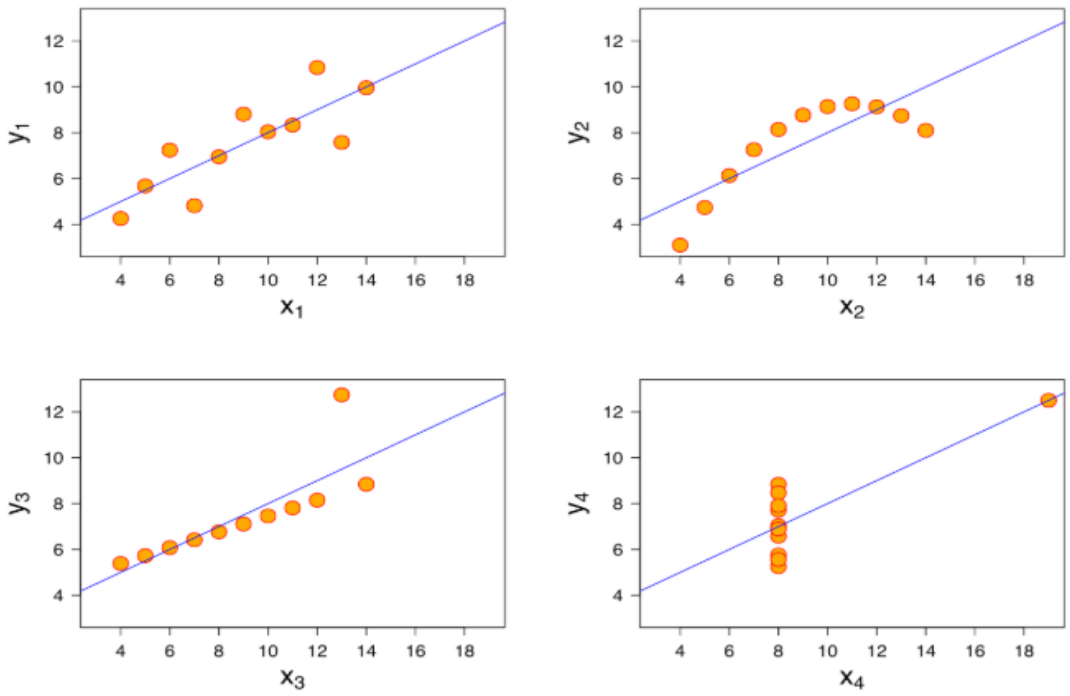
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

But while plotting these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

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.

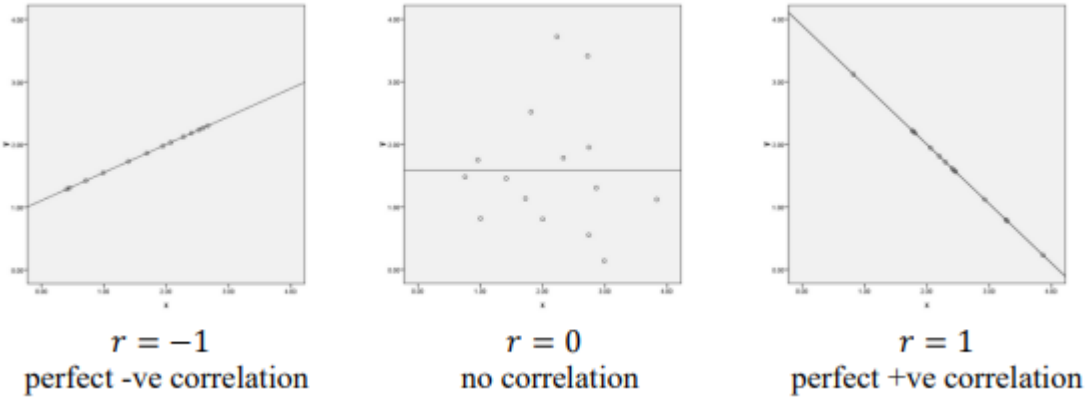
5. What is Pearson's R?

Pearson's correlation coefficient is a statistical measure of the strength of a linear relationship between paired data. In a sample it is denoted by r and is by design constrained as follows

Furthermore:

- Positive values denote positive linear correlation
- Negative values denote negative linear correlation;
- A value of 0 denotes no linear correlation;
- The closer the value is to 1 or –1, the stronger the linear correlation.

In the figures various samples and their corresponding sample correlation coefficient values are presented. The first three represent the "extreme" correlation values of -1, 0 and 1:



$r = -1$
perfect -ve correlation

$r = 0$
no correlation

$r = 1$
perfect +ve correlation

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

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Eucledian distance between two data points in their computations, this is a problem.

To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

The two most discussed scaling methods are Normalization and Standardization.

Normalization typically means rescales the values into a range of [0,1].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

$$x' = \frac{x - \bar{x}}{\sigma}$$

Scaling is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

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

The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. The Variance Inflation Factor (VIF) is 1/Tolerance

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where,

R2i is the coefficient of determination of a regression model where the ith factor is treated as a response variable in the model with all of the other factors.

Tolerance **= 1-R2i**

If VIF is infinite, it means,

Tolerance = 0

i.e. 1-R2i = 0 which implies R2 value is 1 which means variables are highly correlated , hence multicollinearity issues exists and model is not stable.

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.

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.

Cost Function is a way to determine how well the machine learning model has performed given the different values of each parameters.

For example, the linear regression model, the parameters will be the two coefficients, Beta 1 and Beta 2.

The cost function will be the sum of least square methods.

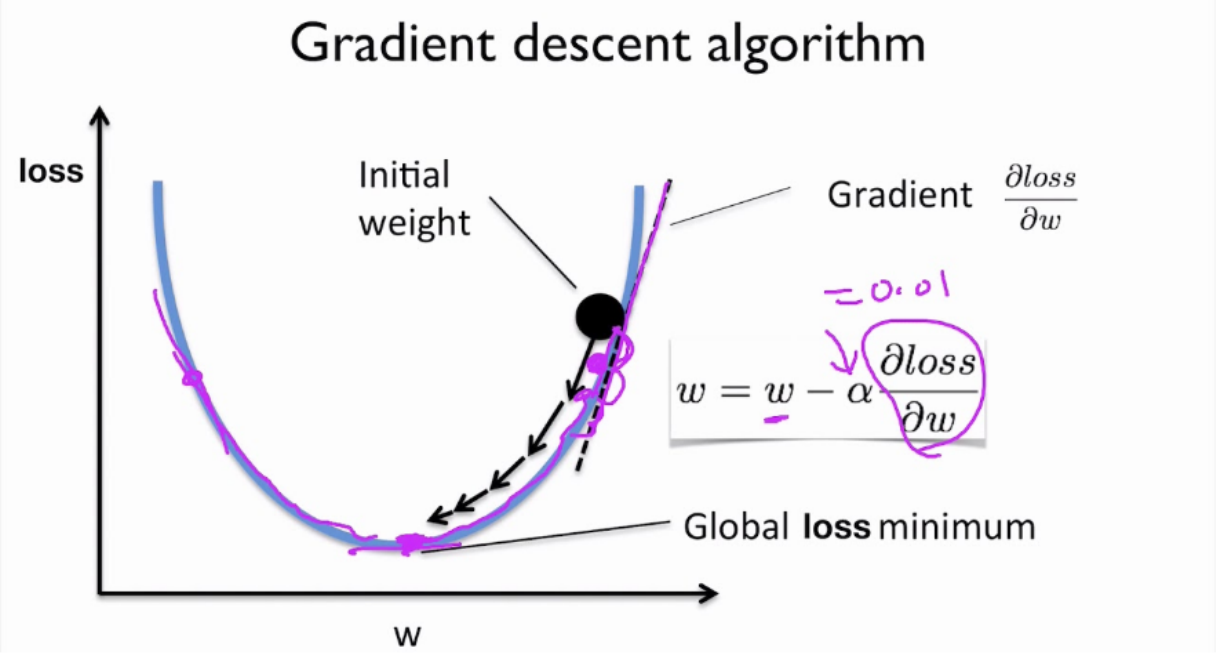
$$y = \beta_0 + \beta_1 x_1$$

Linear Regression Model

Since the cost function is a function of the parameters Beta 1 and Beta 2, we can plot out the cost function with each value of Beta. (I.e. Given the value of each coefficient, we can refer to the cost function to know how well the machine learning model has performed.)

When we are training the model, we are trying to the find the values of the coefficients (the Betas, for the case of linear regression) that will give us the lowest cost. In other words, for the case of linear regression, we are finding the value of the coefficients that will reduce the cost to the minimum

Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost).



<u>Gradient Descent Procedure</u>
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.

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

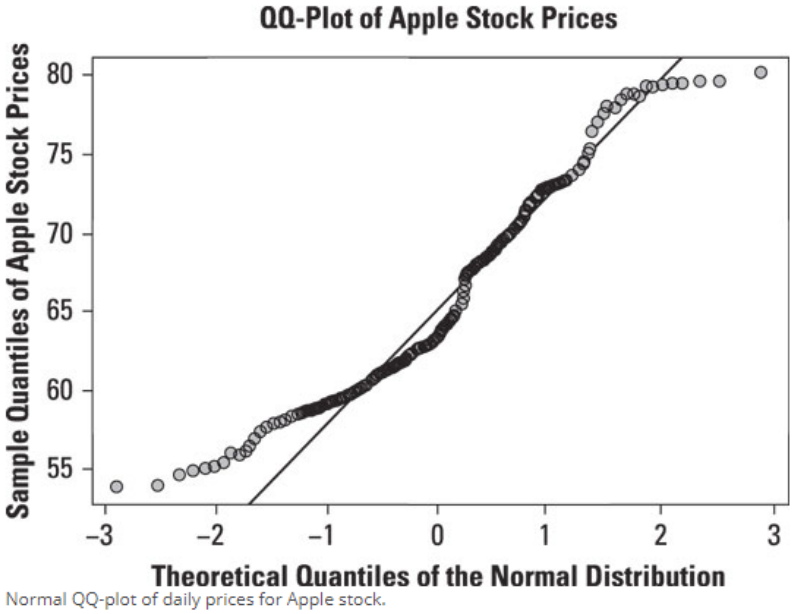
A quantile-quantile plot (also known as a QQ-plot) is another way you can determine whether a dataset matches a specified probability distribution. **QQ-plots are often used to determine whether a dataset is normally distributed.** Graphically, as the name suggests, the horizontal and vertical axes of a QQ-plot are used to show quantiles.

Quartiles divide a dataset into four equal groups, each consisting of 25 percent of the data. But there is nothing particularly special about the number four. You can choose any number of groups you please.
Another popular type of quantile is the percentile, which divides a dataset into 100 equal groups. For example, the 30th percentile is the boundary between the smallest 30 percent of the data and the largest 70 percent of the data. The median of a dataset is the 50th percentile of the dataset. The 25th percentile is the first quartile, and the 75th percentile the third quartile.

With a QQ-plot, the quantiles of the sample data are on the vertical axis, and the quantiles of a specified probability distribution are on the horizontal axis. The plot consists of a series of points that show the **relationship between the actual data and the specified probability distribution**. If the elements of a dataset

perfectly match the specified probability distribution, the points on the graph will form a 45 degree line.

For example, this figure shows a normal QQ-plot for the price of Apple stock from January 1, 2013 to December 31, 2013.



Normal QQ-plot of daily prices for Apple stock.

The QQ-plot shows that the prices of Apple stock do not conform very well to the normal distribution. In particular, the deviation between Apple stock prices and the normal distribution seems to be greatest in the lower left-hand corner of the graph, which corresponds to the left tail of the normal distribution. The discrepancy is also noticeable in the upper right-hand corner of the graph, which corresponds to the right tail of the normal distribution.