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

Ans: Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

Simple regression

Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

$$y=mx+b$$

Multivariable regression

A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$f(x,y,z)=w1x+w2y+w3z$$

The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

$$Sales=w1Radio+w2TV+w3News$$

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

Ans: We make a few assumptions when we use linear regression to model the relationship between a response and a predictor. These assumptions are essentially conditions that should be met before we draw inferences regarding the model estimates or before we use a model to make prediction.

- **Linear Relationship:** The relationship between the independent and dependent variables should be linear. This can be tested using scatter plots.
- **Multivariate Normal:** All the variables together should be multivariate normal. For all the variables to be multivariate normal each variable separately has to be univariate normal means a bell shaped curve. And any subset of variables should also be multivariate normal. This can be tested by plotting a histogram.

- **No Multicollinearity:** There is little or no multicollinearity in the data. Multicollinearity happens when the independent variables are highly correlated with each other. Multicollinearity can be tested with correlation matrix.

- **No Autocorrelation:** There is little or no autocorrelation in the data. Autocorrelation means a single column data values are related to each other. In other words f(x+1)is dependent on value of f(x). Autocorrelation can be tested with scatter plots.

- **Homoscedasticity:** Homoscedasticity is there. This means "same variance" .In other words residuals are equal across regression line. Homoscedasticity can also be tested using scatter plot.

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

Ans: Coefficient of correlation:

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movement of the two variables.

Coefficient of Determination

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
- With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R^2 of 0.20 means that 20 percent is predictable; and so on.

4. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

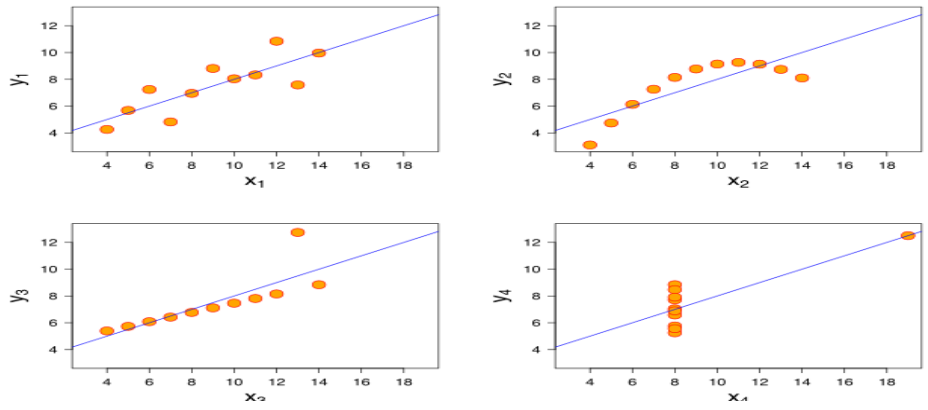
Anscombe's Dataset

Each of the datasets in the quartet consists of 11 (x,y) points:

Anscombe 1		Anscombe 2		Anscombe 3		Anscombe 4	
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

Each Dataset in the quartet consists of the following statistical analysis:

	X	Y
Mean	9	7.50
Variance	11	4.127
Correlation between X and Y	0.816	
Linear Regression Line	y=0.5x+3.00	



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

5. What is Pearson's R?

Ans: Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. 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.

One of the most commonly used formulas in stats is Pearson's correlation coefficient formula

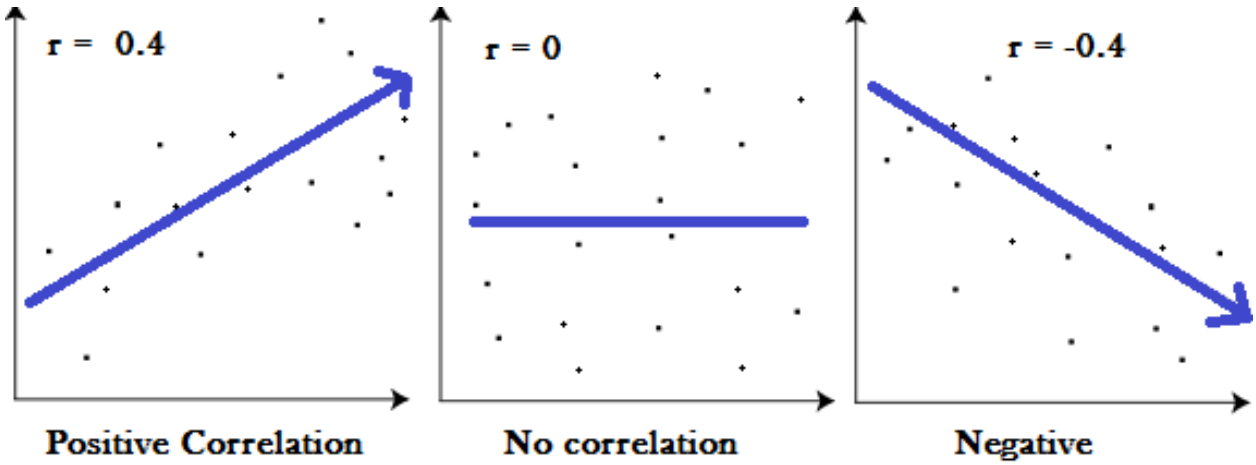
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

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: A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

-1 indicates a strong negative relationship: A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

A result of zero indicates no relationship at all: Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.



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

Ans: Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

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.

Normalization:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1

- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

Ans: An **infinite VIF value** indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an **infinite VIF** as well)

Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the squared multiple correlation of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes infinite.

8. What is the Gauss-Markov theorem?

Ans: the **Gauss–Markov theorem** states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero.[1] The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance. See, for example, the James–Stein estimator (which also drops linearity) or ridge regression.

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.

9. Explain the gradient descent algorithm in detail.

Ans: Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

Gradient Descent Procedure
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

Ans: 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.

Q-Q plots let you check that the data meet the assumption of normality. They compare the distribution of your data to a normal distribution by plotting the quartiles of your data against the quartiles of a normal distribution. If your data are normally distributed then they should form an approximately straight line.