Q1. Explain the linear regression algorithm in detail.

A1. A linear regression algorithm tries to model the relationship between independent variables and dependent variable. It assumes that the relationship between them is linear.

The following are the steps that are a part of linear regression –

A. It tries to fit a best fitting line which can describe the linear relationship between the target variable and the feature variables. For that it assumes the equation of the best fit line as –

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Here, Xs are the feature variables, Epsilon is the error term and Y is the dependent variable. The error term accounts for the variation in thedependent variablethat theindependent variablesdo not explain.

B. In order to get the best fitting line, the algorithm tries to minimize the error between the predicted target variables and the actual target variable value. This it can do using 2 ways -
    a. Ordinary Least Squares Method where it calculates RSS(Residual Sum of Squares) and tries to minimize the sum of errors between actual target values and predicted target values. The line which corresponds to the minimum error is chosen.
    b. Using Gradient Descent. It takes partial derivative of the best fit line wrt the coefficients and constant term and then step by step moves closer to the point where the error is the least.
C. Once the best line is reached. The algorithm has done its job. Now its on us to see if the model would be able to predict values on the test set too. For that we make sure that all the assumptions regarding the residuals are satisfied. We also make sure that the model had no variables which were collinear. For that we may use various feature selection techniques and once the features are filtered out, we then again model using the Linear Regression algorithm.

Q2. What are the assumptions of linear regression regarding residuals?

A2. The following are the Assumptions which need to be true in case we want the OLS model to correctly give inferences and predictions –

First Assumption - The regression **model is linear in the coefficients and the error term** :-
The model that we make follows the following equation. This equation assumes that Y (dependent variable) is a linear combination of independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

In the equation, the betas (βs) are the parameters that OLS estimates. Epsilon (ε) is the random error.

Second Assumption - The **error term has a population mean of zero :-**

The error term accounts for the variation in the dependent variable that the independent variables do not explain. Random chance should determine the values of the error term. For your model to be unbiased, the average value of the error term must equal zero.

Suppose the average error is +7. This non-zero average error indicates that our model systematically underpredicts the observed values. Statisticians refer to **systematic error like this as bias**, and it signifies that our model is inadequate because it is not correct on average.

Stated another way, we want the expected value of the error to equal zero. If the expected value is +7 rather than zero, **part of the error term is predictable**, and we **should add that information to the regression model itself**. We want only random error left for the error term.

Third Assumption - **All independent variables are uncorrelated with the error term** :-

If an independent variable is correlated with the error term, we can use the independent variable to predict the error term, which violates the notion that the error term represents unpredictable random error. We need to find a way to incorporate that information into the regression model itself.

This assumption is also referred to as **exogeneity**. **When this type of correlation exists, there is endogeneity**.

Violating this assumption biases the coefficient estimate. To understand why this bias occurs, keep in mind that the error term always explains some of the variability in the dependent variable. However, when an independent variable correlates with the error term, OLS incorrectly attributes some of the variance that the error term explains to the independent variable instead.

Fourth Assumption - Observations of the **error term are uncorrelated with each other :-**

One observation of the error term should not predict the next observation. For instance, if the error for one observation is positive and that systematically increases the probability that the following error is positive, that is a positive correlation. If the subsequent error is more likely to have the opposite sign, that is a negative correlation. **This problem is known both as serial correlation and autocorrelation**.

Assess this assumption by **graphing the residuals in the order that the data were collected**. **You want to see a randomness in the plot**. In the graph for a sales model, there appears to be a cyclical pattern with a positive correlation

Fifth Assumption - The error term has a constant variance (**no heteroscedasticity**)

The variance of the errors should be consistent for all observations. In other words, the variance does not change for each observation or for a range of observations. **This preferred condition is known as homoscedasticity (same scatter)**. If the variance changes, we refer to that as heteroscedasticity (different scatter).

The easiest way to check this assumption is to create a residuals versus fitted value plot. On this type of graph, heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction

Sixth Assumption - No independent variable is a linear function of other explanatory variables(**Non Collinearity**) :-

Perfect correlation suggests that two variables are different forms of the same variable. For example, games won and games lost have a perfect negative correlation (-1). The temperature in Fahrenheit and Celsius have a perfect positive correlation (+1).

Ordinary least squares cannot distinguish one variable from the other when they are perfectly correlated. If you specify a model that contains independent variables with perfect correlation, your statistical software can't fit the model, and it will display an error message. You must remove one of the variables from the model to proceed.

However, the statistical software can fit OLS regression models with imperfect but strong relationships between the independent variables. **If these correlations are high enough, they can cause problems. We refer to this condition as multicollinearity, and it reduces the precision of the estimates in OLS linear regression.**

Seventh Assumption: **The error term is normally distributed**

OLS does not require that the error term follows a normal distribution to produce unbiased estimates with the minimum variance. However, satisfying this assumption allows you to perform statistical hypothesis testing and generate reliable confidence intervals and prediction intervals.

Q3. What is the coefficient of correlation and the coefficient of determination?

A3. Coefficient of Correlation R –

The quantity *r*, called the *linear correlation coefficient*, **measures the strength and the direction of a linear relationship between two variables**. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honor of its developer Karl Pearson.
The mathematical formula for computing ***r*** is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

where *n* is the number of pairs of data.

The value of *r* is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

Note that *r* is a dimensionless quantity

The ***coefficient of determination*, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable**.
The *coefficient of determination* is the ratio of the explained variation to the total variation.
The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between *x* and *y*. The *coefficient of determination* represents the percent of the data that is the closest to the line of best fit

Q4. Explain the Anscombe's quartet in detail.

A4. Anscombe quartet is a **set of 4 plots which bring home the importance of data visualization** and how 4 datasets **having the same summary statistics – mean, sum, standard deviation and correlation - can be totally different from each other when visualized**.

It shows us that over reliance on summary statistics is not a good way to draw inferences and that data distribution should also be visualized. The same is explained using an example below -

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

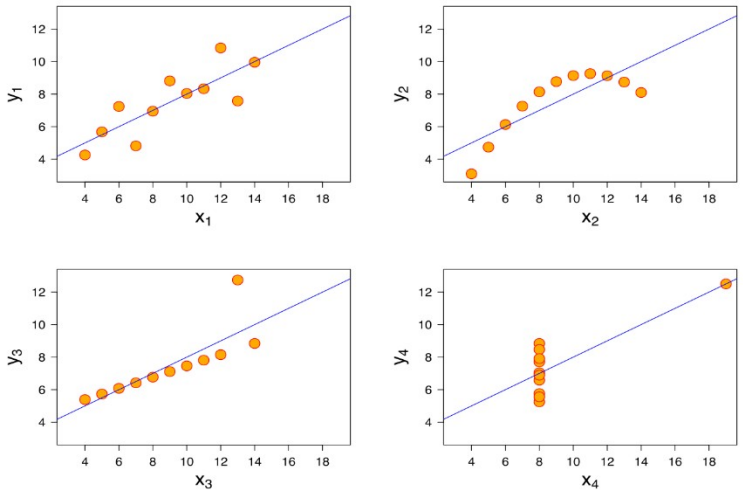
The summary statistics show that the means and the variances were identical for x and y across the groups :

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

When we plot 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 model.

Dataset II is not linearly distributed.

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

Q5. What is Pearson's R?

A5. **Correlation** is a bivariate analysis that **measures the strength of association between two variables** and the direction of the relationship. Pearson *r* correlation is the **most widely used correlation statistic to measure the degree of the relationship between linearly related variables**. The following formula is used to calculate the Pearson *r* correlation:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y

n = number of observations

x_i = value of x (for ith observation)

y_i = value of y (for ith observation)

Assumptions

For the Pearson *r* correlation, **both variables should be normally distributed** (normally distributed variables have a bell-shaped curve). Other assumptions include **linearity and homoscedasticity**. Linearity assumes a straight line relationship between each of the two variables and homoscedasticity assumes that data is equally distributed about the regression line.

Q6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A6. **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 range**. Sometimes, it also helps in **speeding up the calculations in an algorithm.**

The result of **standardization** (or **Z-score normalization**) is that the features will be rescaled so that they'll have **the properties of a standard normal distribution with mean zero and standard deviation 1**

Standardizing the features so that they are centred around 0 with a standard deviation of 1 is not only important **if we are comparing measurements that have different units,** but it is also a **general requirement for many machine learning algorithms**. Intuitively, we can think of **gradient descent** as a prominent example **with features being on different scales, certain weights may update faster than others since the feature values play a role in the weight updates**

An alternative approach to Z-score normalization (or standardization) is the so-called **Min-Max scaling**(Normalisation). In this approach, the data is **scaled to a fixed range - usually 0 to 1**. The cost of having this bounded range - in contrast to standardization - is that we will **end up with smaller standard deviations, which can suppress the effect of outliers.** A popular **application is image processing**, where pixel intensities must be normalized to fit within a certain range (i.e., 0 to 255 for the RGB color range). Also, **typical neural network algorithm require data that on a 0-1 scale**.

Q7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A7. VIF is infinite when **denominator is not defined which is when R square is 1.** That is if R is 1. Hence it is only possible if the **variable whose VIF is being measured is a linear combination of all the other variables in the model**.

Q8. What is the Gauss-Markov theorem?

A8. 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. i.e. **BLUE (Best linear unbiased estimator).**

Q9. Explain the gradient descent algorithm in detail

A9. Gradient descent is an **iterative algorithm that is used to minimize a function/loss function/cost function**. In linear regression where we are modelling the relationship between a dependent variable and one or more independent variables - Let **X** be the independent variable and **Y** be the dependent variable. We will define a linear relationship between these two variables as follows:

$$Y = mX + c$$

Our challenge is to **determine the value of m and c, such that the line corresponding to those values is the best fitting line or gives the minimum error**. The loss is the error in our predicted value of m and c. Our goal is to minimize this error to obtain the most accurate value of m and c. We use the Mean Squared Error function to calculate the loss.

There are **three steps in this function**:

1. Find the difference between the actual y and predicted y value (y = mx + c), for a given x.
2. Square this difference.
3. Find the mean of the squares for every value in X. **This will be our loss function**

$$E = \frac{1}{n}\sum_{i=0}^{n}(y_i - \bar{y}_i)^2$$

Minimizing the loss function and finding **m** and **c –**

Imagine a valley and a person with no sense of direction who wants to get to the bottom of the valley. He goes down the slope **and takes large steps when the slope is steep and small steps when the slope is less steep**. He **decides his next position based on his current position and stops when he gets to the bottom** of the valley which was his goal. This is the same approach that the gradient descent algorithm follows –

Applying gradient descent to **m** and **c** and approach it step by step:

1. Initially let m = 0 and c = 0. Let **L be our learning rate**. This controls how much the value of **m** changes with each step. L could be a small value like 0.0001 for good accuracy.
2. Calculate the **partial derivative of the loss function with respect to m**, and plug in the current values of x, y, m and c in it to obtain the derivative value **D**.

$$D_m = \frac{1}{n}\sum_{i=0}^{n} 2(y_i - (mx_i + c))(-x_i)$$

$$D_m = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - \bar{y}_i)$$

D_m is the value of the partial derivative with respect to **m**.

Similarly lets find the partial derivative with respect to **c**, Dc :

$$D_c = \frac{-2}{n} \sum_{i=0}^{n} (y_i - \bar{y}_i)$$

Now we update the current value of **m** and **c** using the following equation:

$$m = m - L \times D_m$$

$$c = c - L \times D_c$$

We **repeat this process until our loss function is a very small value or ideally 0** (which means 0 error or 100% accuracy). The value of **m** and **c** that we are left with now will be the optimum values.

Now going back to the analogy –

m can be considered the **current position of the person**.

D is equivalent to the **steepness of the slope**

L can be the **speed with which he moves**.

Now the new value of **m** that we calculate using the above equation will be his next position, **and L×D will be the size of the steps** he will take.

When the **slope is more steep (D is more) he takes longer steps and when it is less steep (D is less), he takes smaller steps.**

Finally he arrives at the bottom of the valley which corresponds to our loss = 0.

Q10 .What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. -this

A10 The quantile-quantile (q-q) plot is a graphical technique for determining **if two data sets come from populations with a common distribution**.

A q-q plot is a **plot of the quantiles of the first data set against the quantiles of the second data set**. By a quantile, we mean the fraction (or percent) of points below the given value. **That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value**.

In Linear Regression Q-Q plot is **used to see if the residuals are gaussian in nature**. It is used to **compare the quantiles of the error terms against the quantiles of the normal distribution**. If the

quantiles of the error terms are near enough to the quantiles of the corresponding values computed from the normal distribution i.e **if the lie along the 45 degree straight line**, then we might begin to accept the idea that the **error terms are normally distributed and if they don't, the residuals aren't Gaussian and thus the errors aren't either. This implies that for small sample sizes, one can't assume the estimator $\hat{\beta}$ is Gaussian either, so the standard confidence intervals and significance tests are invalid**.