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

   Linear regression is ML algorithm used to build predication model. Basic assumption for linear regression is that there is some sort of linear relation (negative or positive) between the predictor variables and the dependent variable which we are going to predict. A simple linear regression can be given by equation

   $y = B0 + B1*x$

   where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable(predictor variable), and Y is the value of the dependent variable.

   A linear regression model starts by taking in

   x: input training data (univariate – one input variable(parameter))
   y: labels to data (supervised learning)

   Then the model predicts value of y from given value of x and tries to for best fit line from given datapoints. It also calculates B0(intercept of constant) and B1 which is coefficient of x i.e. what is the factor by which x effects output y.

   This best fit line is calculated using gradient descent method in which the Root Mean Square Error should be minimum.

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

   A linear regression assumes the residual should be normally distributed i.e. it should have its center around zero. Also, there should be no clear pattern between residuals versus predicted value. This 2nd principle is called Homoscedasticity which states variance of error term should be similar across distribution.

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

   R is called as coefficient of correlation and R*R is called coefficient of determination. In other words, Coefficient of Determination is the square of Coefficients of Correlation.

   Coefficient of determination explain percentage variation in y with respect to x. Higher the R square better is our model(assuming your model is not overfitted). Coefficient of determination ranges between 0 and 1.

   Coefficient of correlation is degree of relation between x and y . It ranges between -1 and 1. 1 means the variables move in unison and -1 means exact opposite ie one rises other falls

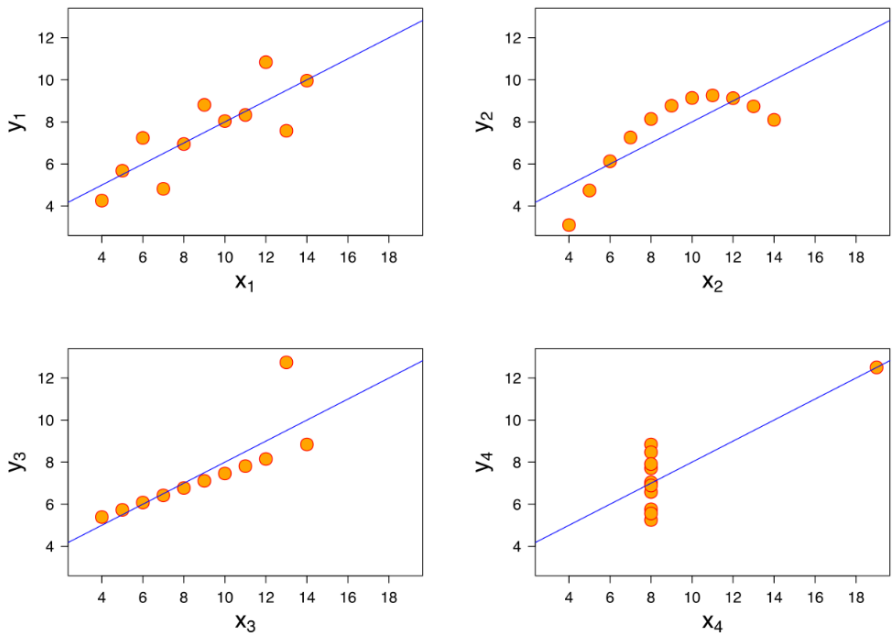
4. Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. All 4 dataset shows similar statistics when calculated numerically using summary statistics but show different properties when examined by graphs.

	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

All four dataset has same properties like sum avg and standard deviation. Even when plotted graphically they have same regression line but are totally different as seen below



This shows effect of outliers which could only be seen when data is plotted and visualized and would have been missed using statistical calculation only.

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

Pearson's R is another name for coefficient of correlation and as discussed above it tells relation between two variables. It values lies between 1 and -1. 1 indicating strong positive linear relation and -1 indicate strong negative relation. R=0 indicates that there is no relation between 2 variables

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

In Linear regression a best fit line is plotted which aim to describe all points with least RMSE. This is done using gradient descent method. Scaling is a method where we bring all independent variables in one standard range say between 0 and 1 in case of minmax scaling or where mean is 0 (Standardization). This process improves the overall speed of algorithm because now we have every independent variable in same range or almost near range to each other.

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.

Below is formula for normalization

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

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.

The major difference is that if Standardization is applied on dummy variables it would change it values as mean of variables will be at zero hence some values will be negative on other hand normalization will not change their values as they are already at zero and one. MinMax scaling also takes care of outlier as all outlier will be mapped to maximum value which is 1 whereas in standardization there still will be outliers.

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

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). This happens if your independent variable is perfectly described by combination of other independent variables.

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.

Let's see each assumption one by one

- Linearity: the parameters we are estimating using the ordinary least square method must be themselves linear.
- Random: our data must have been randomly sampled from the population.
- Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
- Exogeneity: the regressors aren't correlated with the error term.
- 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.

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.

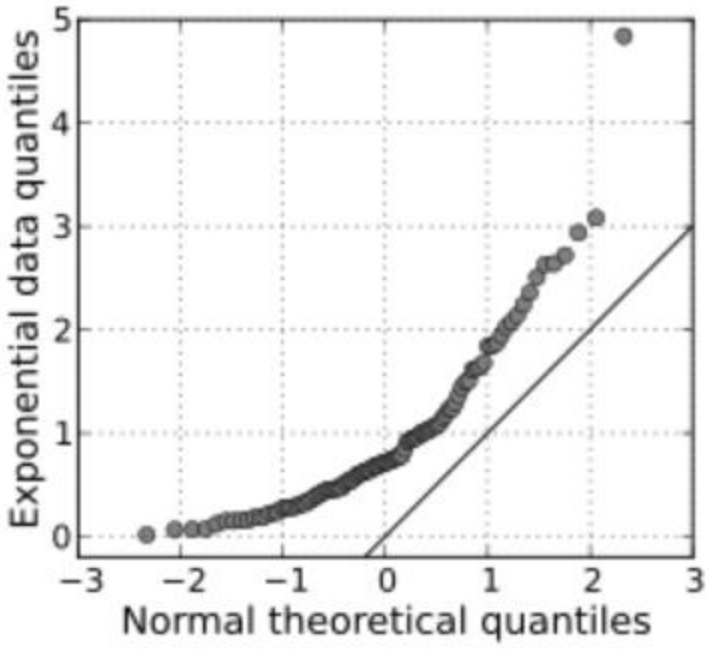
The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value. The cost of the coefficients is evaluated by plugging them into the function and calculating the cost and the derivative of the cost is calculated. The slope or derivative tells us which direction we need to move the coefficients in order to reduce the cost

Now that we know from the derivative which direction is downhill, we can now update the coefficient values. 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.

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



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.