1. Explain the linear regression algorithm in detail.

   Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

   Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
   In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

   Hypothesis function for Linear Regression :

   While training the model we are given :
   x: input training data (univariate – one input variable(parameter))
   y: labels to data (supervised learning)
   When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
   θ_1: intercept
   θ_2: coefficient of x
   Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.
   How to update θ_1 and θ_2 values to get the best fit line ?
   Cost Function (J):
   By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

   Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).
   Gradient Descent:
   To update θ_1 and θ_2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ_1 and θ_2 values and then iteratively updating the values, reaching minimum cost.

2. What are the assumptions of linear regression regarding residual

- X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.
- Error terms are ***normally distributed*** with mean zero(not X, Y)
- Error terms are ***independent*** of each other
- Error terms have ***constant variance*** (homoscedasticity):

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

Correlation Coefficient, *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.

(Aren't you glad you have a graphing calculator that computes this formula?

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

Positive correlation: If *x* and *y* have a strong positive linear correlation, *r* is close to +1. An *r* value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between *x* and *y* variables such that as values for *x* increases, values for *y* also increase.

Negative correlation: If *x* and *y* have a strong negative linear correlation, *r* is close to -1. An *r* value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between *x* and *y* such that as values for *x* increase, values for *y* decrease.

No correlation: If there is no linear correlation or a weak linear correlation, *r* is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables

Note that *r* is a dimensionless quantity; that is, it does not depend on the units employed.

A *perfect* correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If *r* = +1, the slope of this line is positive. If *r* = -1, the slope of this line is negative.

A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

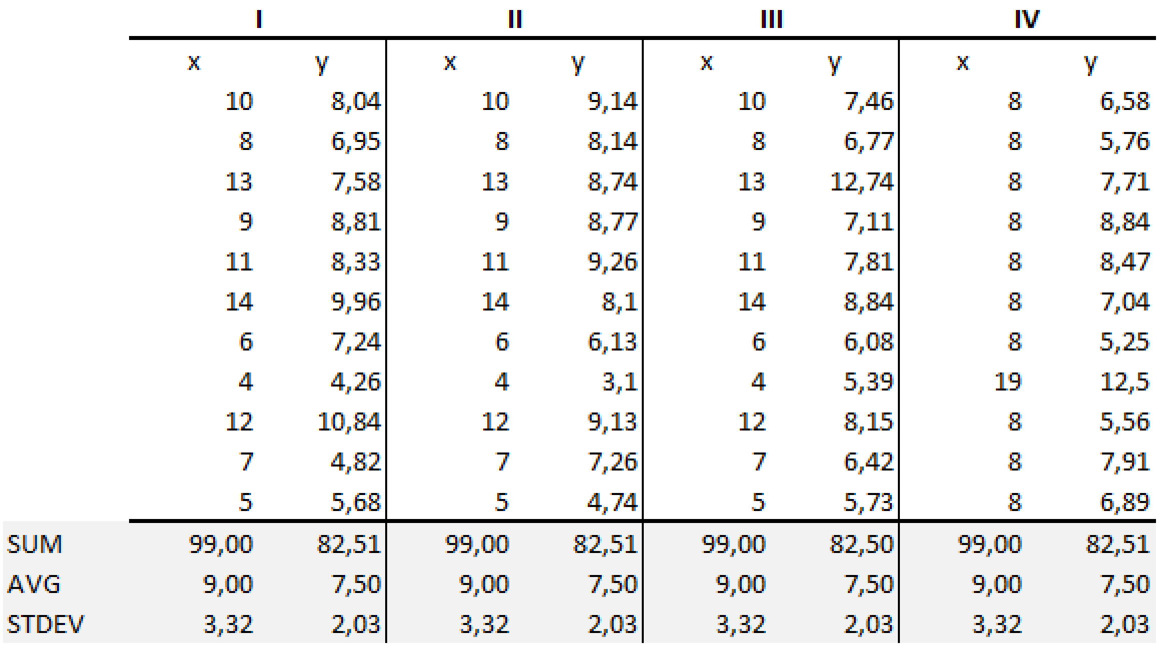
Coefficient of Determination, r^2 or R^2 :

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. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

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. For example, if $r = 0.922$, then $r^2 = 0.850$, 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.
The *coefficient of determination* is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

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. But things change completely, and I must emphasize **COMPLETELY,** when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



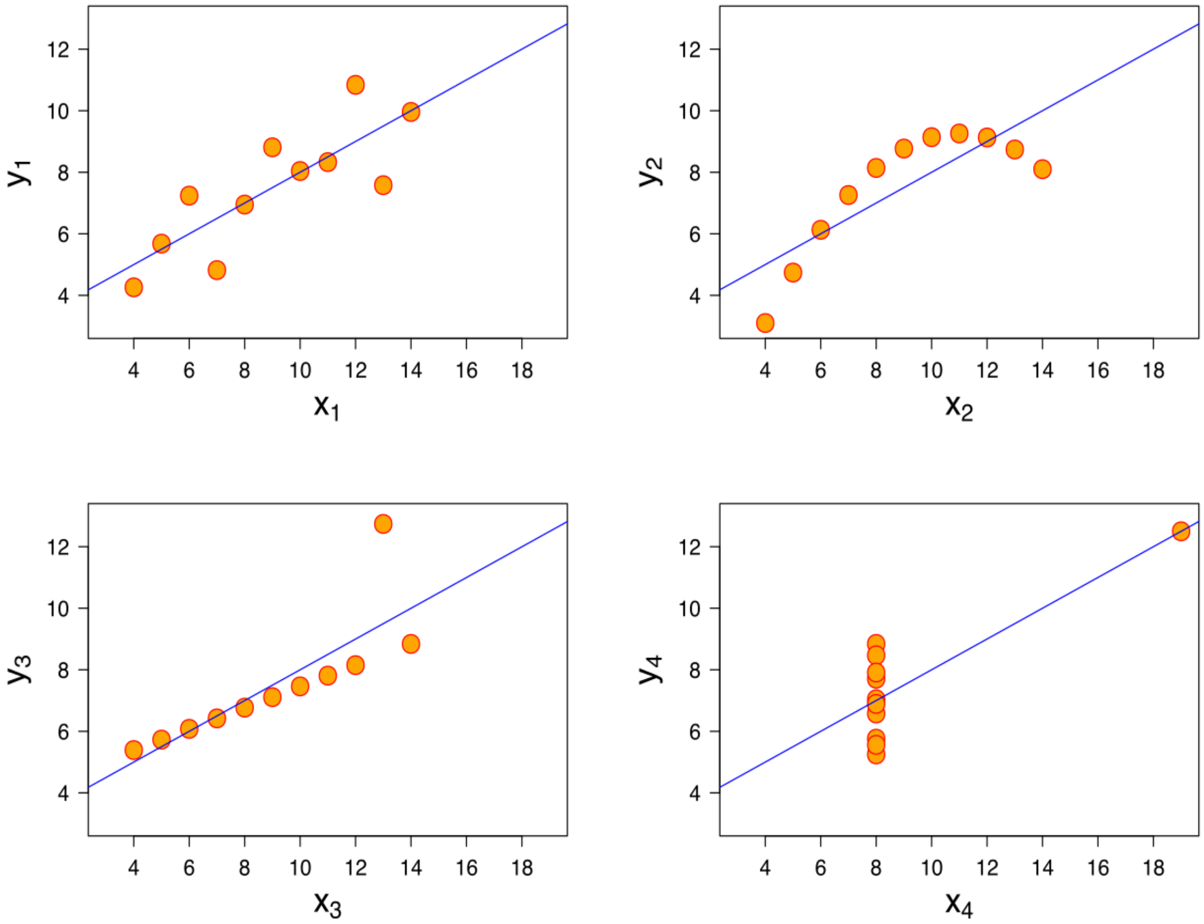
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15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

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 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 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)

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

Scaling

In scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1].

It is performed because 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

Two methods are usually well known for rescaling data. ***Normalization***, which scales all numeric variables in the range [0,1]. One possible formula is given below:

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

On the other hand, you can use **standardization** on your data set. It will then transform it to have zero mean and unit variance, for example using the equation below:

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

Both of these techniques have their drawbacks. If you have outliers in your data set, normalizing your data will certainly scale the “normal” data to a very small interval. And generally, most of data sets have outliers. When using standardization, your new data aren’t bounded (unlike normalization).

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).

8. What is the Gauss-Markov theorem?

   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.

Gradient Descent algorithm

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:

1. **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.
2. **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.
3. **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.

Purpose: Check If Two Data Sets Can Be **Fit** With the Same Distribution. 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.