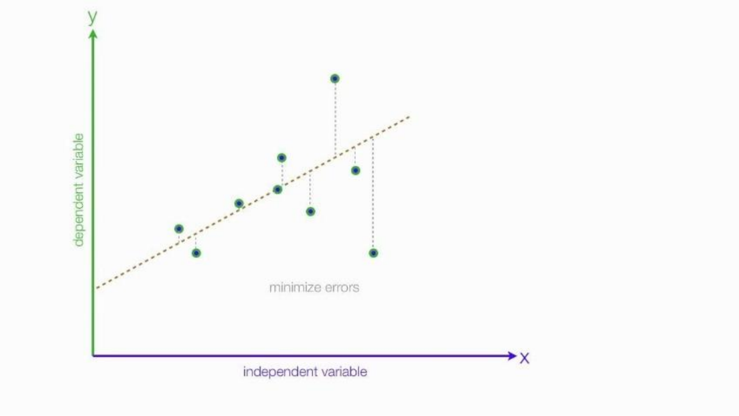
1. Explain the linear regression algorithm in detail.
   Solution: Linear Regression is a method of finding the best straight-line fitting to the given data points, it is the way of finding the best Linear Relationship between the independent and dependent variables. In technical terms the Linear Regression is a ML algorithm that finds the best linear-fit relationship between input variables (x) and output variables (y), i.e. y can be calculated from a linear combination of the input variables x.



   For example, if we want to predict the price of house based on Size of the House Carpet Area. Let's say we have data for three houses: Sizes 200, 700, 1400 sqft and prices are respectively 500$, 200$ and 1000 $ and if someone wants to buy a house of 1000sqft what's the expected prices is.

   The Linear Equation assigns one scaler factor (coefficient) to each feature or input variables and is it usually represented as β and one additional coefficient is added i.e. intercept or bias, and ε is the error term.
   Equation of Simple Linear Regression is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

   Now in Linear Regression algorithm the main objective is to find the right coefficient values β's, so that the line perfectly fits all the data keeping the residuals as less as possible (or the cost function as less as possible). To minimize the error function either **Closed Form** approach is taken or **Iterative Solution**, for the first order (Gradient Descent) is followed.

2. What are the assumptions of linear regression regarding residuals?
   Solution: The assumptions of Linear Regressions are
   I. The error terms/residual **ε** must be normally distributed. (Normality Assumption)
   II. It is assumed that the errors/residuals must have a zero mean i.e. the errors are normally distributed around zero (zero mean assumption)
   III. Residuals are independent of each other i.e. the pairwise covariance is zero. Absence of this phenomenon is also known as Autocorrelation (Independent Error Assumption)
   IV. The error terms/residuals must have a constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity. (Constant variance assumption)

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

   Solution: To answer to How well is the regression equation truly represent the set of data. One way to answer that question is *Correlation Coefficient* and the *Coefficient of determination*.

   Correlation of Coefficient is *r* or *R* and **Coefficient of determination** r^2 or R^2

   a. **Correlation of Coefficient**

      i. The quantity r or R is called the *Linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables.

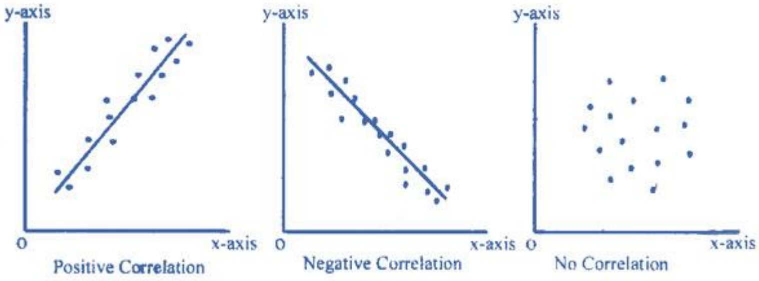
      ii. Mathematical Formula is $r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\ \sqrt{n(\sum y^2) - (\sum y)^2}}$

      iii. Value of r is -1 ≤ r ≤ +1, The + and – signs are used for positive linear correlations and negative linear correlations

      iv. Positive Correlation: r is close to +1, i.e. if x increase y also increases

      v. Negative Correlation: r is close is -1, i.e. if x increase y decreases

      vi. No Correlation: r is close to 0, i.e. value of one variable change and the other variable remains constant



   b. **Coefficient of determination**

      i. This 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. It gives how well the regression line represents the data.

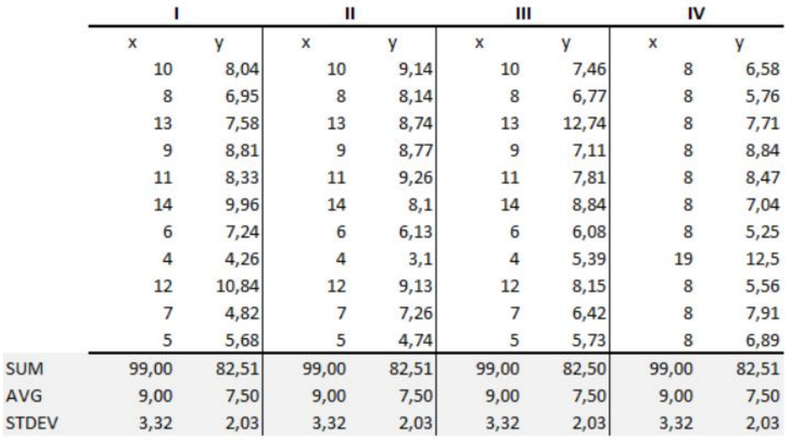
      ii. This represents the percent of the data that is the closest to the line of best fit. If R^2 = 0.92, i.e. 92% of total variance in *y* can be explained by the linear Relationship between *x* and *y*. Other 8% remains unexplained.

4. Explain the Anscombe's quartet in detail.

   Solution: Anscombe's Quartet was developed by statistician Francis Anscombe.

   To explain the phenomenon, we take help of some data, there are **4 Data set** and **11 number of Data points**. So, we can't just simply put the regression algorithm.

   So, these four data sets have nearly identical simple **descriptive statistics**.

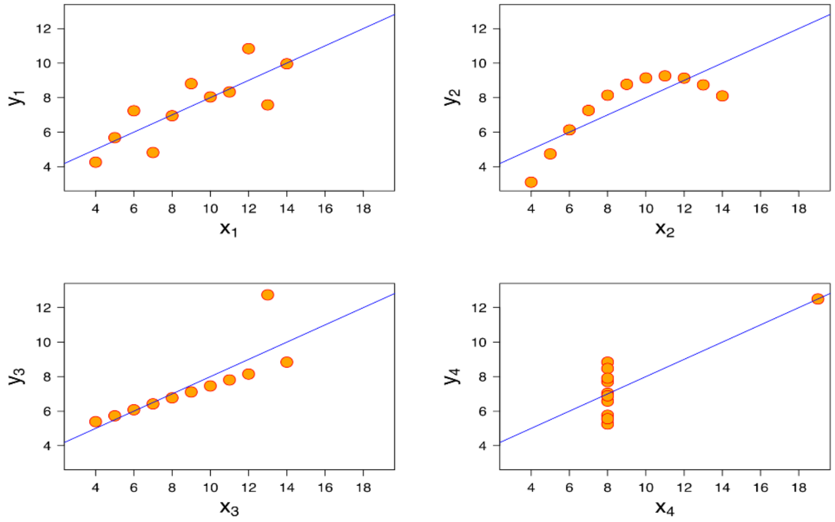


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15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

Let's see the summary statistics for each of the dataset and we will calculate Mean, Standard Deviation, so we can see all the stats are quite identical.

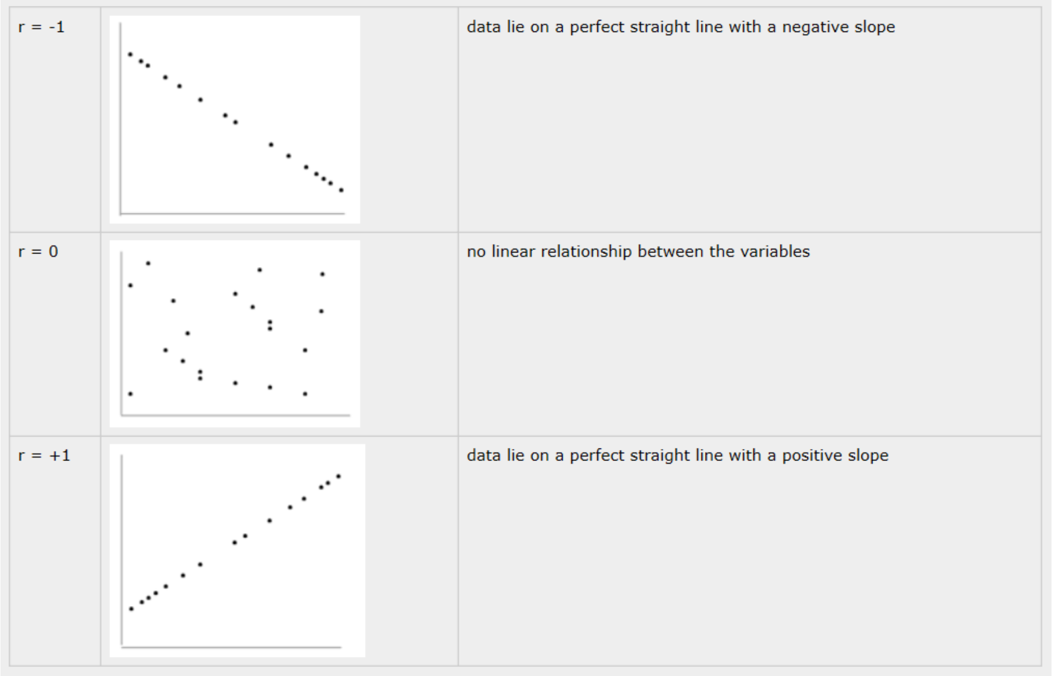
So, when we plot it, they have different distribution and appear very different.



i. First scatter plot **Simple Linear Relationship,** and it fulfills assumptions of normality
ii. Second plot is not **Distributed Normally**
iii. For the third plot, the distribution is Linear, but regression line should be different, because the calculated regression is completed thrown off by an outlier.
iv. And for the last plot One outlier is more than enough to produce a high Correlation Coefficient.

So, Quartet is often used to illustrate the importance of visualization of the data points prior to analyzing and build the model.

5. What is Pearson’s R?
   Solution: This is the test statistics that measures the statistical relationship or association between quantitative, continuous variables. **Pearson's correlation coefficient** (r) is a measure of the strength of the association between the two variables. It is based on method of covariance. This correlation coefficient is designed for linear relationships, and it might not be a good measure for a non-linear relationship between the variables.
   Pearson's correlation coefficient (r) lies between **-1 ≤ r ≤ 1**
   -1 means negative correlation, **+1** Total Positive correlation, **0** is no linear correlation



r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables
r = +1		data lie on a perfect straight line with a positive slope

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa, and for the Zero correlation, value of one variable change and the other variable remains constant.

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

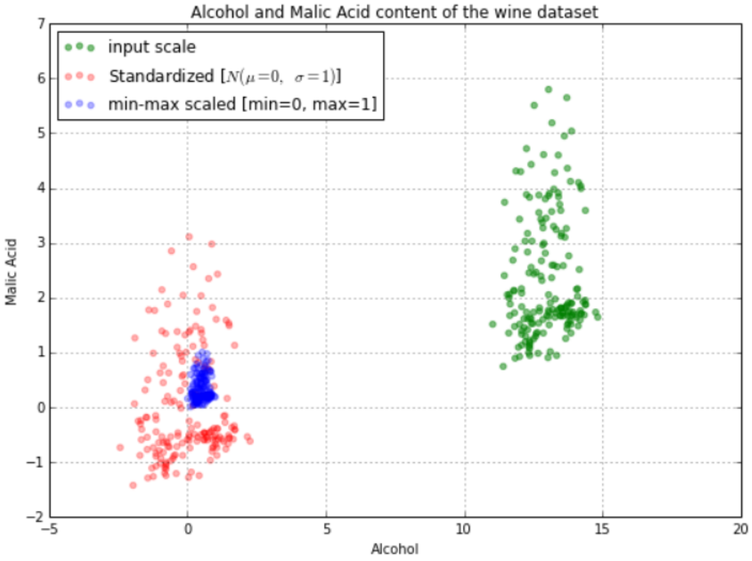
   Solution: There are often happens the dataset contains features and they have different scales, units, range. Now in most ML algorithm, **Euclidian Distance** is used between the two data points. If one of the features has a broad range of values, the distance will be governed by this feature. If left alone, these algorithms only take in the magnitude of features neglecting the units. This result will vary a lot, the features with high magnitude weigh more compared low magnitude’s one.

   To suppress the effect, all the features must be brought to the same level of magnitude.
   Different Scaling:

i. **Standardization (Z Score Normalization)**: $x' = \frac{x - \bar{x}}{\sigma}$ where x is original value and $\bar{x} = \text{average}(x)$ mean of the feature vector and σ is the standard deviation. This results into a distribution with mean **μ = 0** and standard deviation **σ = 1**.

ii. **Min Max Scaling:** $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$ This scales the data point between 0 and 1.

Normalization typically means scaling the data between [0,1] and Standardization typically means rescale the data mean of 0 and standard deviation of 1 (unit variance). One major disadvantage of **normalization** over **standardization**, it loses some information in the data, especially about **outliers**.



Scaled data are very close together, which might not be the desired result. It might cause algorithms such as gradient descent to take longer to converge.

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Solution: In modelling, multicollinearity in the set of predictors is a problem, VIF (Variance Inflation Analysis). When other variables gives a nearly perfectly fit (**R^2 = 1**) against the other variables, which results in undefined VIF. If all the independent variables are orthogonal to each other, then VIF = 1.0.
VIF Formula is **VIF**$_i$ = $\mathbf{\frac{1}{1 - R^2{}_i}}$, so when R^2 = 1 then the VIF becomes infinite.

8. What is the Gauss-Markov theorem?
Solution: It tells if a **certain set of assumption** are met, the **Ordinary Least Square** (OLS), estimates the Regression coefficient, which results in **Best Linear Unbiased Estimator** (BLUE) possible.
There are Six Assumptions:
   1. **Linearity:** The parameters that are being estimated by OLS, must be linear
   2. **Random:** Data must have been randomly sampled from Population
   3. **Non-Collinearity:** The calculated regressors are not perfectly correlated with each other
   4. **Exogeneity:** The calculated regressors are not correlated with the error term
   5. **Homoscedasticity:** the error/residual of the variance must be constant.
   6. The **expected value** of the **error term** is **zero**

   Checking how well the data matches these assumptions is an important part of estimating regression coefficients.

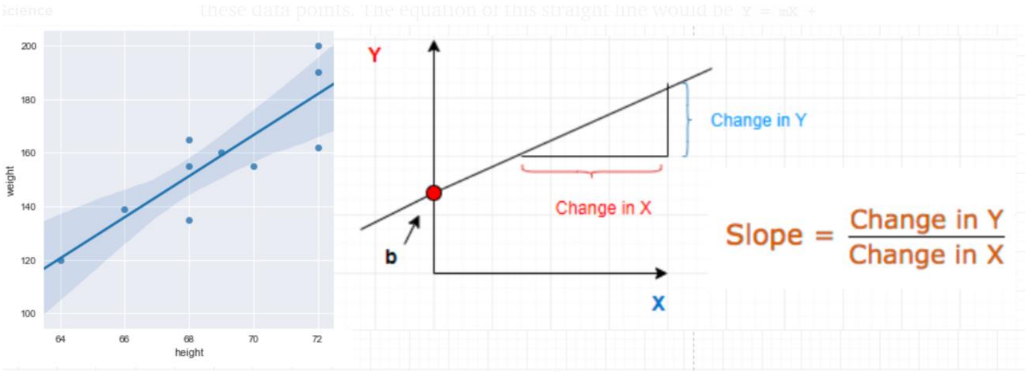
9. Explain the gradient descent algorithm in detail.
Solution: Gradient Descent Algorithm is an iterative process that takes to the minimum of a function. The final Formula is

$$\boldsymbol{\theta^1} = \boldsymbol{\theta^0} - \boldsymbol{\alpha \nabla J(\theta)}$$

$\boldsymbol{\theta^1} = \text{Next Position,}$ $\boldsymbol{\theta^1} = \text{Current Position,}$ $\boldsymbol{\alpha} = \text{Learning rate (small step),}$

$\boldsymbol{\nabla J(\theta)} = \text{Direction of Fastest Decrease}$

Now the Question is how to derive at this formula.
The equation of this straight-line would-be $Y = mX + b$ where m is the slope and b is its intercept on the Y-axis.


Science
these data points. The equation of this straight line would be Y = mX +
200
180
160
140
120
100
weight
64
66
68
70
72
height
Y
Change in Y
Change in X
b
X
Slope = Change in Y / Change in X

Now for a ML algorithm it predicts outputs based on inputs in this case it predicts the coefficient that itself will result in determination of Y. Now the residual or error is

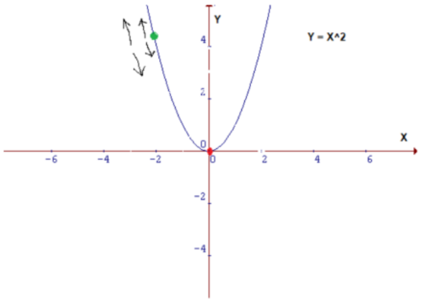
$$Error = Y(Pred) - Y\ (Actual)$$

Now comes the idea of **Cost Function/Loss Function.** This helps in evaluating the performances of the ML Algorithm. **Loss Function** is error for a Single Training and **Cost Function** is average of the loss functions for the whole epoch.
For N number of datapoint, so as to minimize the error.

$$\text{Cost}\,\boldsymbol{J(\theta)} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} (\mathrm{Y\_pred - Y\_actual})^2$$

The main goal of all ML algorithm is to minimize the Cost Function, because lower error between the actual and the predicted values signifies that the algorithm has done a significant job in learning. This is lead to smallest possible error.

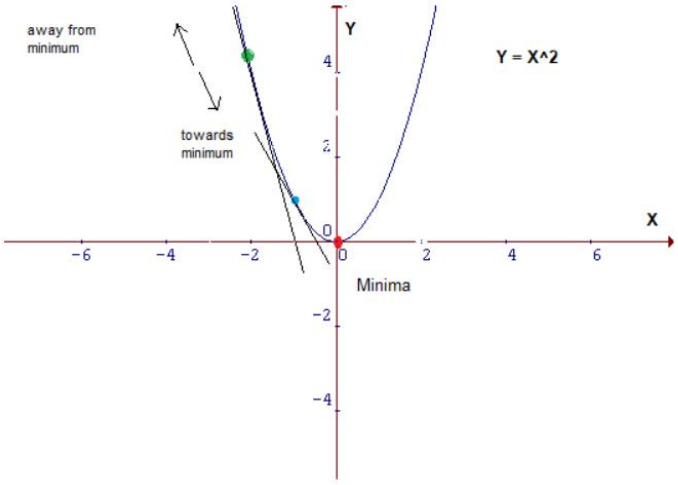
To find the local minima, the algorithm that's being used is **Gradient Descent.** Main idea is walking along the graph below, and currently at the 'green' dot. Aim is to reach the minimum i.e. the 'red' dot, but from the current position.



Now the possible actions are

i. Go upward or Downward
ii. If decided which way to go now remains is the step size, bigger or smaller.

Gradient Descent here takes the help of **Derivatives.** It gives the slope of the graph, at a point. So, after computing the Derivatives, directions are determined to determine the **Local Minima (Convergence Point)**.

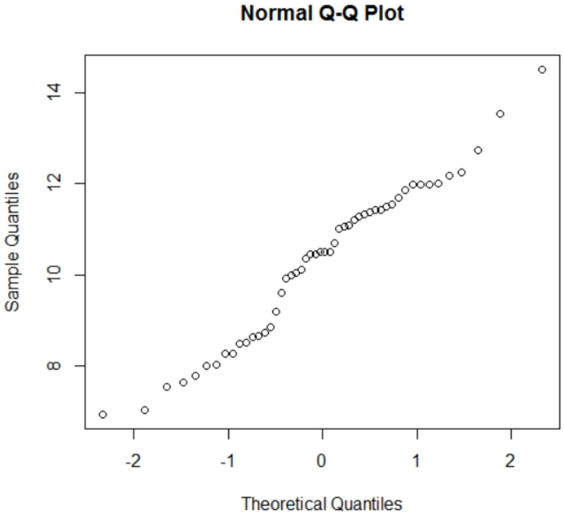


The slope at the blue point is less steep than that at the green point which means it will take much smaller steps to reach the minimum from the blue point than from the green point. This size of steps taken to reach the minimum or bottom is called **Learning Rate**.

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

Solution: Q-Q plot is Quantile-Quantile Plot is a Probability plot, is a graphical method for comparing two probability distributions by plotting their quantiles against each other. It is basically required to determine if two data sets come from populations with a common distribution. 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.

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, it can be seen the points forming a line that's roughly straight.



In Linear regression we have an assumption, that the residuals/errors are Normally distributed. To check the same, we use Q-Q Plot. Objective is to fit a linear regression model, check if the points lie approximately on the line, and if they don't, residuals aren't Gaussian and thus your errors aren't either.