Correlation

Correlation is used to measure the mutual relationship between two or more variables. Correlation coefficient is the numerical measure of the relationship between two variables. Correlation shows the presence of a relationship between variables, strength of the relationship and its direction. It can be illustrated in graphs which simplify the interpretation. Correlation demonstrates the relationship between the measurable variables, but it doesn't identify the cause of the relationship. Therefore there can be underlying variables which affects the relationship.

Scatter diagrams, Karl Pearson's correlation coefficient and Spearsman's Rank correlation are few methods which used measure the correlation. Correlation coefficient of sample is represented by "r", Correlation coefficient of population is represented by "ρ". Correlation Coefficient ranges from +1 to -1. If the correlation coefficient is +1, there is a perfect positive correlation between two variables. When correlation coefficient is -1 there is a perfect negative correlation between two variables. When there is no correlation at all the correlation coefficient is 0. When it is below ± 0.5 relationship considered to be not strong. On the other hand when it is above ± 0.75 relationship considered to be strong. Between ± 0.5 to ± 0.75 it is considered to have a moderate correlation between variables.

1.1 Scatter Diagram

Scatter diagram is the method of drawing a graph with (x,y) coordinates. According to the spread of the coordinates, one can easily guess the direction and the strength of the relationship. X is used for independent variables and Y is used for dependent variables. Dependent variable is the variable affected by one or more independent variables. Independent variables are the cause of the dependent variable. When there are multiple independent variables they are represented by x_1 ,x_2,x_3 , ….,x_n.

Followings are few examples for scatter plot diagrams.

Figure 1.1 : Perfect Positive Correlation

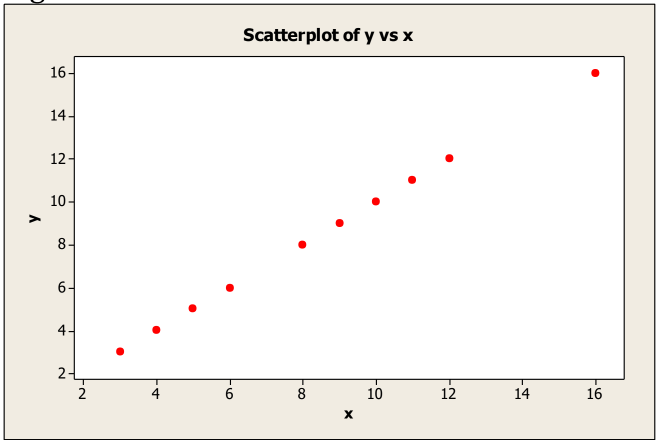


Figure 1.2: Perfect Negative Correlation

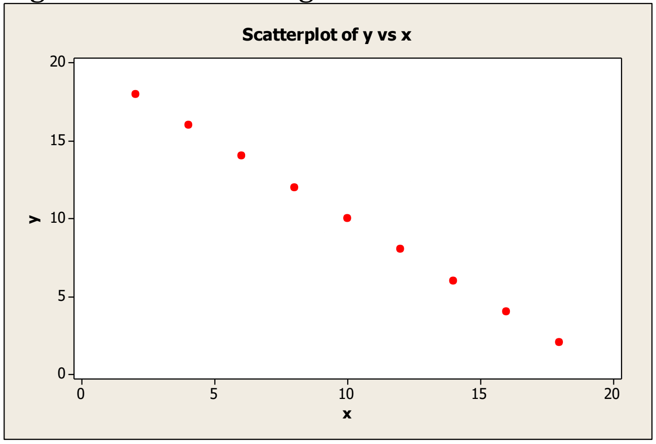


Figure 1.3 : Strong Positive Correlation

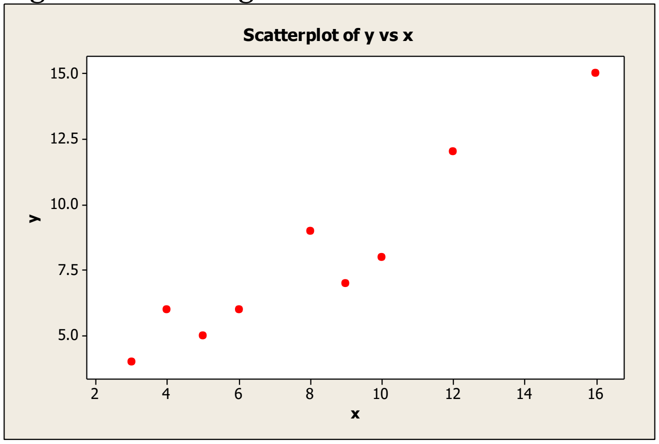


Figure 1.4: Strong Negative Correlation

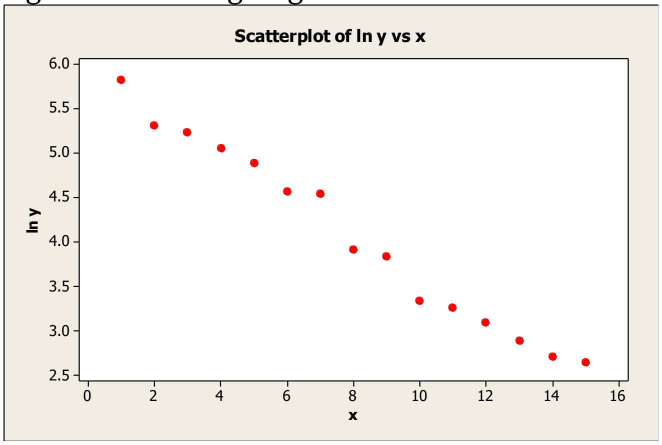


Figure 1.5: Moderate Positive Correlation

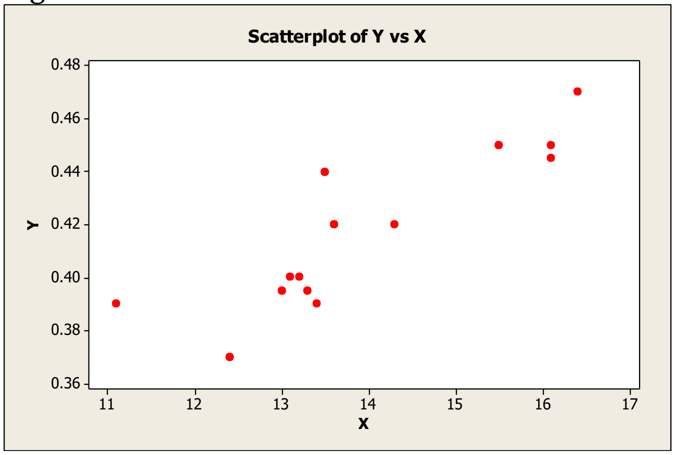


Figure 1.6: Moderate Negative Correlation

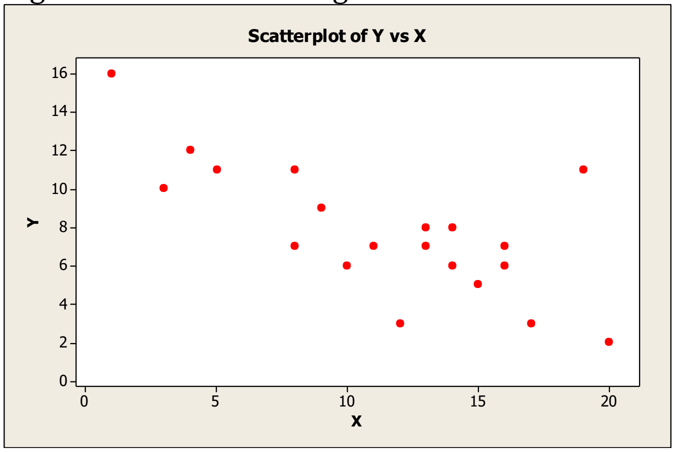


Figure 1.7: Non Linear Correlation

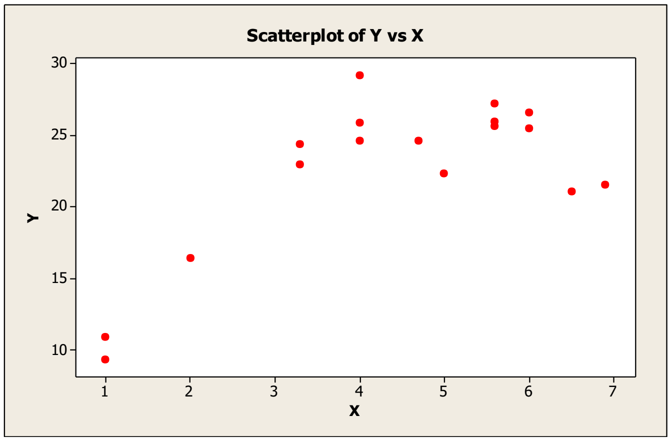
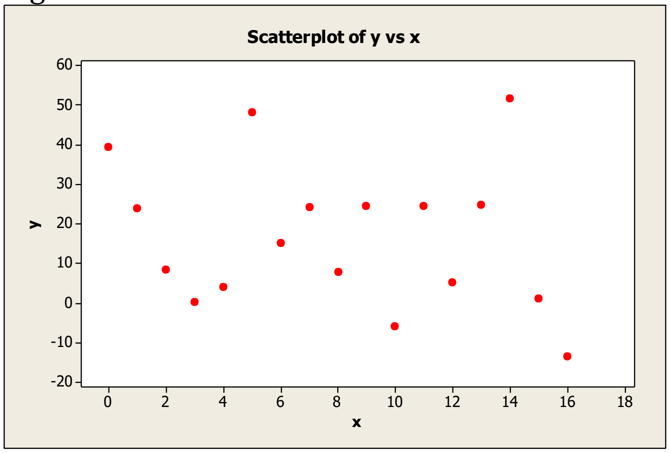


Figure 1.8: No Correlation



When there is more than one independent variable used in a regression, relationship among independent variables; relationship between independent variables and dependent variable should be analyzed using a correlation matrix.

Karl Pearson's Correlation Coefficient

Karl Pearson's correlation coefficient measures both strength and direction of the relationship. The coefficient gives a measurable value to the strength of relationship. Karl Pearson's coefficient is calculated assuming there are no other factors influencing the dependent variable other than one dependent variable.

Covariance of (x,y) or Cov(x,y) is equal to $E[(x- \bar{x})- (y-\bar{y})]$, on assumption of variances are positive, correlation of (x,y) = $\frac{cov(x,y)}{sd(x)sd(y)}$, where sd is standard deviation

Below equations which can be used to calculate the correlation are based on the covariance of relationship.

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

$$r = \frac{\sum[(x - \bar{x})(y - \bar{y})]}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

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

Example 1.1:

Given below is the data set of marks scored for the year end mathematics examination and hours spent for mathematic homework per week. Find out if there is a relationship between these variables and describe the nature of the relationship.

x	1	2	3	4	5
y	35	60	75	85	95

x- hours spent for mathematic homework per day

y- marks scored for the year end mathematics examination

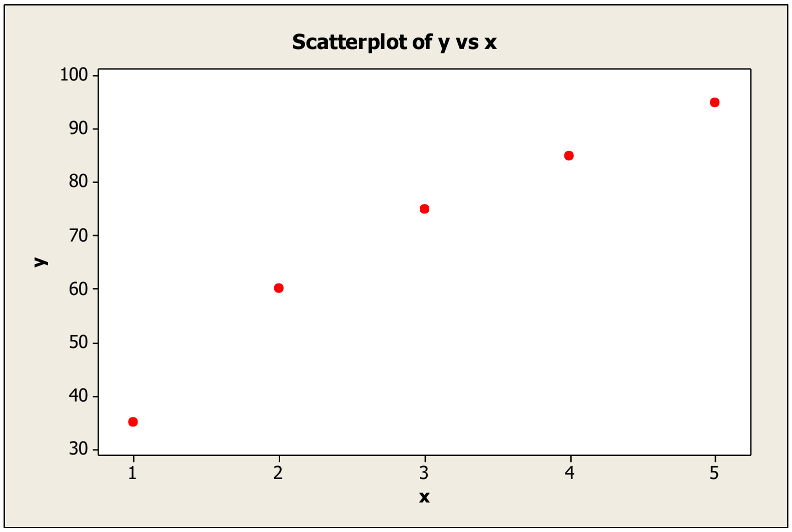
Answer

Calculate $\sum x$, $\sum y$, $\sum xy$, $\sum x^2$, $\sum y^2$ as explained in table 1.1.

Table 1.1: Descriptive Statistics

	x	**y**	**xy**	**x^2**	**y^2**
	1	35	35	1	1225
	2	60	120	4	3600
	3	75	225	9	5625
	4	85	340	16	7225
	5	95	475	25	9025
Total	**15**	**350**	**1195**	**55**	**26700**

Figure 1.9: Scatter plot



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

$$r = \frac{5 * 1195 - 15 * 350}{\sqrt{(5 * 55) - (15)^2}\sqrt{(5 * 26700) - (350)^2}}$$

$$r = 0.978$$

In above example correlation coefficient is +0.978 which is close to +1. It can be concluded that there is a strong positive relationship between hours a student spend for home work per week and year end mathematics exam results.