# Distance Measures available in Cluster Analysis - StatTools

## Euclidean Distance

Euclidean distance between two vectors and , is defined as:



In many analysis tasks, the variables under consideration are measured on different scales or levels and this can affect the Euclidian distance. We can solve this problem by standardizing the variables.

## Squared Euclidean Distance

The standard Euclidean distance is usually used in order to place progressively greater weight on objects that are farther apart.

## Manhattan Distance

It is similar to the walking distance between two points in a city like New York’s Manhattan district. For two vectors  and , this distance equals the sum of the horizontal and vertical components.

$$d\left(x,y\right)=\sum\_{i=1}^{p}\left|y\_{i}-x\_{i}\right|$$

## Mahalanobis Distance

Also called statistical distance because it takes into account the correlation between the random variables. The Mahalanobis distance between two random vectors **x** and **y** is defined as:



where **S** is the sample covariance matrix.

## Correlation and Absolute Correlation

It is a measure of linear [relation](https://en.wikipedia.org/wiki/Correlation) between two variables X and Y. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. The mathematical expression for computing it is:



where σXY is the covariance, σX is the standard deviation of the variable X, and σY is the standard deviation of the variable Y.

Since most clustering methods use dissimilarities (such as distances), we need to convert the correlation matrix to a dissimilarity matrix. This can conveniently be done by replacing each rij by 1-|rij| if **Absolute Correlation** is used or 1-rij if just **Correlation** is used. Notice that if Absolute Correlation is not used, only positive correlation is considered as similarity measure.