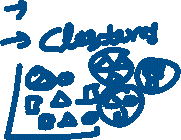
**Cluster Analysis**

We will first learn about the fundamentals of clustering, then proceed to explore its applications, various methodologies ad implement the Rmap package and our own K-Means clustering algorithm in R.

Shape

Description automatically generated with low confidence



***“Clustering is a technique of data segmentation that partitions the data into several groups based on their similarity”.***



Basically, we group the data through a statistical operation. These smaller groups that are formed from the bigger data are known as clusters. These cluster exhibit the following properties:

* They are discovered while carrying out the operation and the knowledge of their number is not known in advance.
* Clusters are the aggregation of similar objects that share common characteristics.

Clustering is the most widespread and popular method of Data Analysis and Data Mining. It used in cases where the underlying input data has a colossal volume and we are tasked with finding similar subsets that can be analysed in several ways.

**For example –** A marketing company can categorise their customers based on their economic background, age and several other factors to sell their products, in a better way.

**Applications of Cluster Analysis:**

Applications of R clustering are as follows:

* **Marketing –** In the area of marketing, we use clustering to explore and select customers that are potential buyers of the product. This differentiates the most likeable customers from the ones who possess the least tendency to purchase the product. After the clusters have been developed, *businesses can keep a track of their customers and make necessary decisions to retain them in that cluster.*
* **Retail –**Retail industries make use of clustering to group customers based on their preferences, style, choice of wear as well as store preferences. This allows them to manage their stores in a much more efficient manner.
* **Medical Science –** Medicine and health industries make use of clustering algorithms to *facilitate efficient diagnosis and treatment of their patients as well as the discovery of new medicines.* Based on the age, group, genetic coding of the patients, these organisations are better capable to understand diagnosis through robust clustering.
* **Sociology –** Clustering is used in Data Mining operations to *divide people based on their demographics, lifestyle, socioeconomic status,* etc. This can help the law enforcement agencies to group potential criminals and even identify them with an efficient implementation of the clustering algorithm.

In different fields, R clustering has different names, such as:

* **Marketing –** In marketing, ‘*segmentation*’ or ‘*typological analyses*’ term is available for clustering.
* **Medicine –** Clustering in medicine is known as *nosology*.
* **Biology –** It is referred to as *numerical taxonomy* in the field of Biology.

**Methods for Measuring Distance between Objects:**

For calculating the distance between the objects in K-means, we make use of the following types of methods:

* **Euclidean Distance –** It is the most widely used method for measuring the distance between the objects that are present in a multidimensional space.



In general, for an n-dimensional space, the distance is

* **Squared Euclidean Distance –** This is obtained by squaring the Euclidean Distance. The objects that are present at further distances are assigned greater weights.



* **City-Block (Manhattan) Distance –** The difference between two points in all dimensions is calculated using this method. It is similar to Euclidean Distance in many cases but it has an added functioning in the reduction of the effect in the extreme objects which do not possess squared coordinates.



**K-Means Clustering in R:**

One of the most popular partitioning algorithms in clustering is the K-means cluster analysis in [R](https://en.wikipedia.org/wiki/R_(programming_language)). It is an unsupervised learning algorithm. It tries to cluster data based on their similarity. Also, we have specified the number of clusters and we want that the data must be grouped into the same clusters. The algorithm assigns each observation to a cluster and also finds the centroid of each cluster.

**The K-means Algorithm:**



* Selects K centroids (K rows chosen at random).



* Then, we have to assign each data point to its closest centroid.

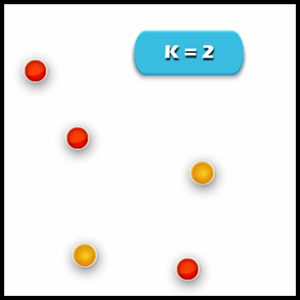
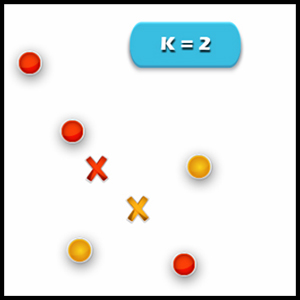


* Moreover, it recalculates the centroids as the average of all data points in a cluster.

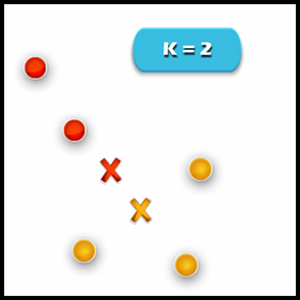
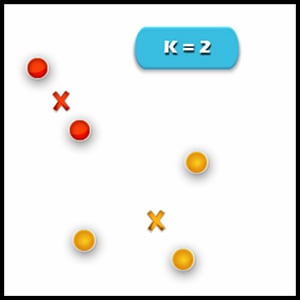


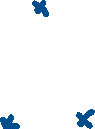
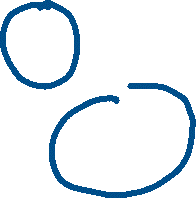
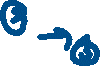
* Assigns data points to their closest centroids.
* Moreover, we have to continue steps 3 and 4 until the observations are not reassigned.



[](https://data-flair.training/blogs/wp-content/uploads/sites/2/2017/07/k-means-clustering-1.jpg) [](https://data-flair.training/blogs/wp-content/uploads/sites/2/2017/07/k-means-clustering-2.jpg)



[](https://data-flair.training/blogs/wp-content/uploads/sites/2/2017/07/k-means-clustering-3.jpg) [](https://data-flair.training/blogs/wp-content/uploads/sites/2/2017/07/k-means-clustering-4.jpg)

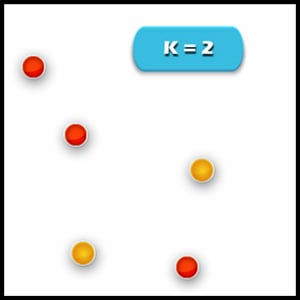


This algorithm works in these steps:

**1. Specify the desired number of clusters K**: Let us choose k=2 for these 5 data points in 2D space.



**2. Assign each data point to a cluster**: Let’s assign three points in cluster 1 using red colour and two points in cluster 2 using yellow colour (as shown in the image).

[](https://data-flair.training/blogs/wp-content/uploads/sites/2/2017/07/k-means-clustering-5.jpg)

**3. Compute cluster centroids**: The centroid of data points in the red cluster is shown using the red cross and those in a yellow cluster using a yellow cross.

**4. Re-assignment of points to their closest cluster in centroid:**Red clusters contain data points that are assigned to the bottom even though it’s closer to the centroid of the yellow cluster. Thus, we assign that data point into a yellow cluster.

**5. Re-compute cluster centroids**: Now, re-computing the centroids for both the clusters. We perform the repetition of step 4 and 5 and until that time no more improvements can be performed. We’ll repeat the 4th and 5th steps until we’ll reach global optima. This continues until no more switching is possible. Then it will mark the termination of the algorithm if not mentioned.

We will now understand the k-means algorithm with the following example:

## Cluster Analysis in R

**Clusters**are a group of the same items or elements like a cluster of stars or a cluster of grapes or a cluster of networks and so on.

Clustering in R refers to the assimilation of the same kind of data in groups or clusters to distinguish one group from the others(gathering of the same type of data). This can be represented in graphical format through R. We use the KMeans model in this process.

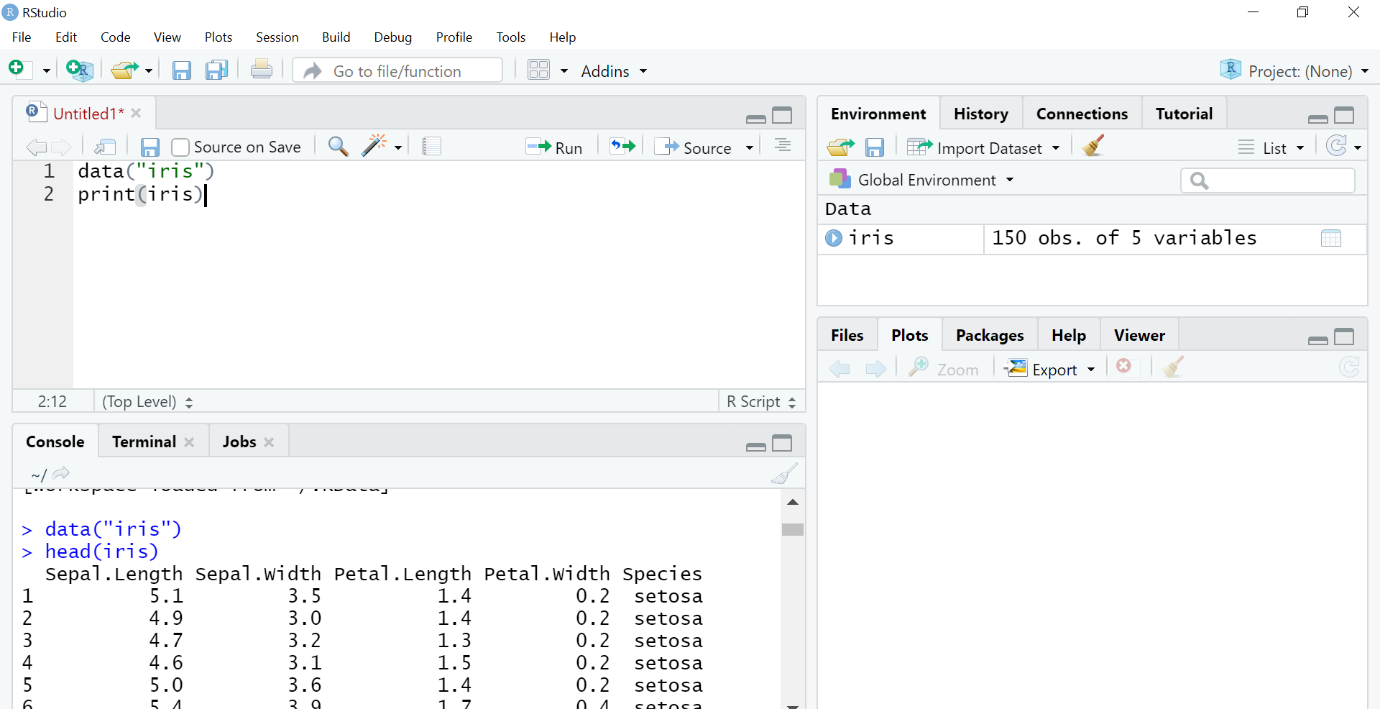
All this is theory but in practice, R has a clustering package that calculates the above steps.

## Step 1

I will work on the Iris dataset which is an inbuilt dataset in R using the Cluster package. It has 5 columns namely – Sepal length, Sepal width, Petal Length, Petal Width, and Species. Iris is a flower and here in this dataset 3 of its species Setosa, Versicolor, Verginica are mentioned. We will cluster the flowers according to their species. The code to load the dataset:

**data("iris")**

**head(iris)**

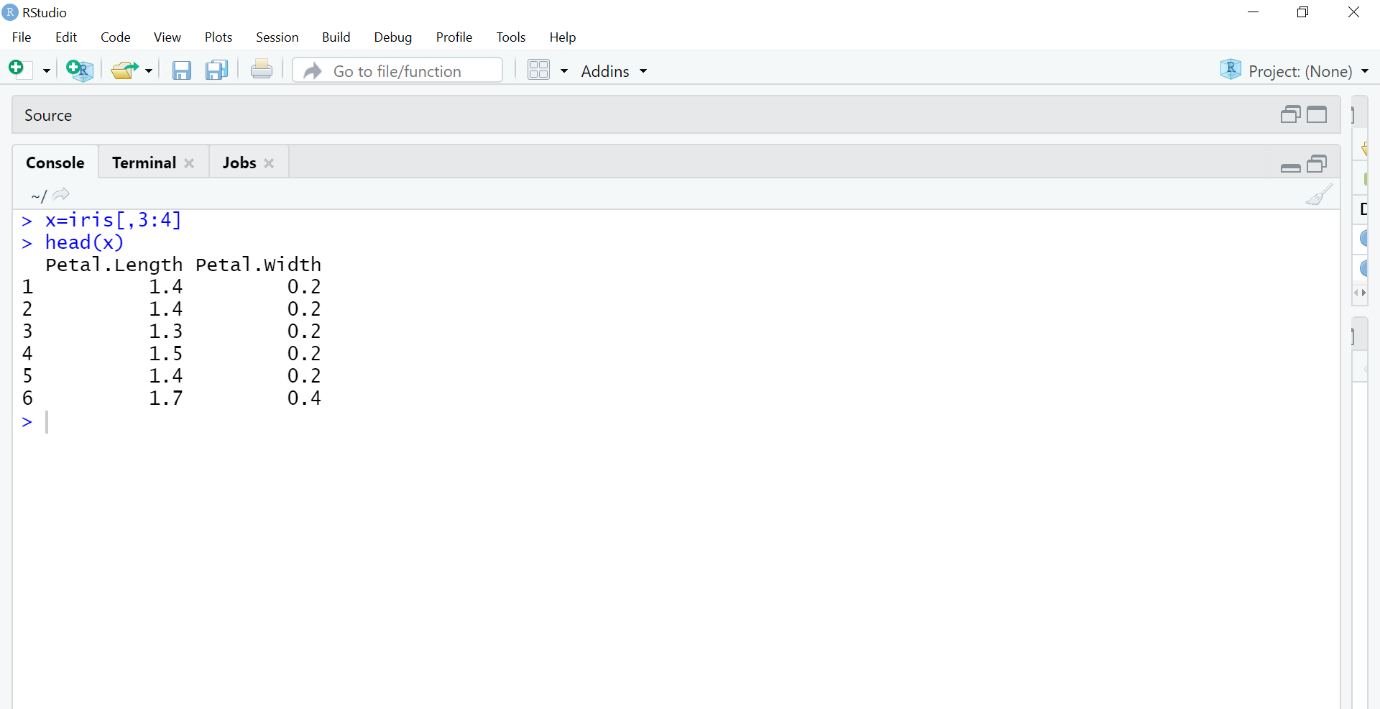


## Step 2

The next step is to separate the 3rd and 4th columns into separate object x as we are using the unsupervised learning method. We are removing labels so that the huge input of petal length and petal width columns will be used by the machine to perform clustering unsupervised.

x=iris[,3:4] #using only petal length and width columns

head(x)



## Step 3

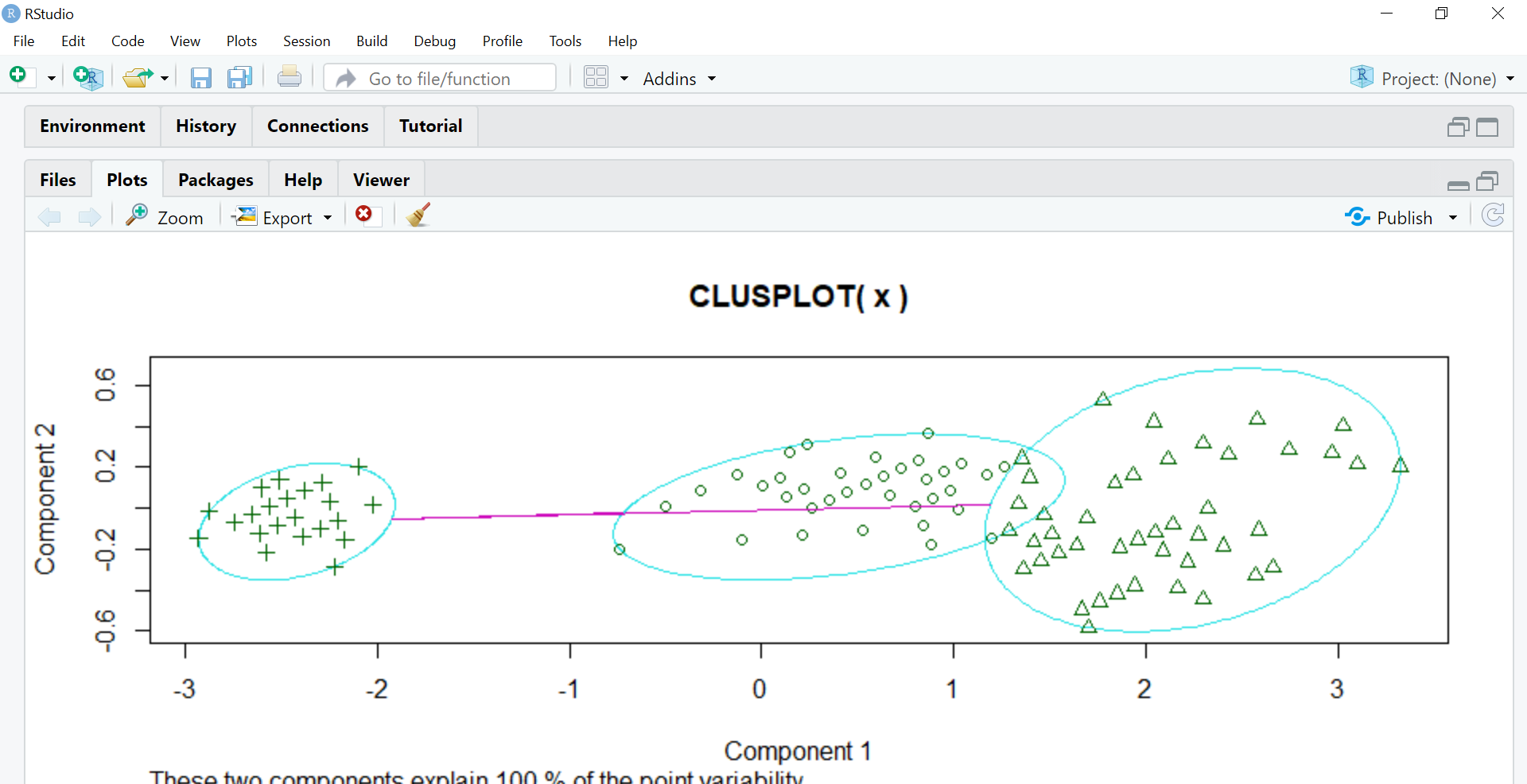
The next step is to use the K Means algorithm. K Means is the method we use which has parameters (data, no. of clusters or groups). Here our data is the x object and we will have k=3 clusters as there are 3 species in the dataset.

Then the ‘cluster’ packageis called. Clustering in R is done using this inbuilt package which will perform all the mathematics. Clusplot function creates a 2D graph of the clusters.

**model=kmeans(x,3)**

**library(cluster)**

**clusplot(x,model$cluster)**



Component 1 and Component 2 seen in the graph are the two components in PCA (Principal Component Analysis) which is basically a feature extraction method that uses the important components and removes the rest. It reduces the dimensionality of the data for easier KMeans application. All of this is done by the cluster package itself in R.

These two components explain 100% variability in the output which means the data object x fed to PCA was precise enough to form clear clusters using KMeans and there is minimum (negligible) overlapping amongst them.

## Step 4

The next step is to assign different colors to the clusters and shading them hence we use the color and shade parameters setting them to T which means true.

**clusplot(x,model$cluster,color=T,shade=T)**

