K-means is efficient, and perhaps, the most popular clustering method. It is a way for finding natural groups in otherwise unlabeled data. You specify the number of clusters you want defined and the algorithm minimizes the total within-cluster variance.

Here, we will play around with the base R inbuilt k-means function on some *labeled* data.

```r
data("iris")

x <- 1:100
km1 <- kmeans(x,5)
km1$cluster
```

```
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
##  [36] 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 2 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
```

## Exercise 1

**Feed the columns with sepal length measurements in the inbuilt iris data-set to the k-means; save the cluster vector of each observation. Use 3 centers and set the random seed to 1 before.**

## Exercise 2

**Check the proportions of each species by cluster (use table()).**

## Exercise 3

**Make a plot with sepal length on the horizontal axis and sepal width on the vertical axis. Find a way to visualize both the actual species and the cluster the algorithm is categorized into.**

## Exercise 4

**Repeat the clustering from step one, but include petal measurements also. Does the clustering reflect the actual species better now?**

## Exercise 5

**Create a new data-set identical to iris, but multiply the "Petal.Width" by 2. Are the results different now?**

## Exercise 6

**Standardize your new data-set so that each variable has a mean of 0 and a variance of 1; check once more if multiplying by two makes a difference.**