k means cluster with R

In this section we will learn to use R library in Machine learning for prediction using Unsupervised ML. Here, I consider Iris dataset which is already available in package. and we will predict the optimum number of clusters and visualize it.

k means cluster

k-means method is an algorithm that assign each object to the cluster having nearest centroid (mean).

Algorithm
1. Partition the objects into k-initial clusters.
2. Re-assign objects to the cluster whose centeroid is nearest.
3. Re-calculate the centroid for the cluster receiving the new objects .
4. Repeat step 2 until no more assignment is possible.

Data Prepration

Import require package for this problem.

```
library(datasets)
library(tidyverse)
library(cluster)
library(factoextra)
```

Import the data

```
data<-iris
data<-na.omit(data)
```

Successfully imported data

```
 summary(data)
```

```
Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
```

```
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
Species
setosa    :50
versicolor:50
virginica :50
```

Then, we will read the nature and structure of imported data.

```
str(data)
```

```
’data.frame’: 150 obs. of  5 variables:
$ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
$ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
$ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
$ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
$ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Here, we observe that first four column have numeric structure and column Species are factor. so we eliminate last column.

```
Data<-select(data,c(1,2,3,4))
head(Data)
```

```
Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
4          4.6         3.1          1.5         0.2
5          5.0         3.6          1.4         0.2
6          5.4         3.9          1.7         0.4
```

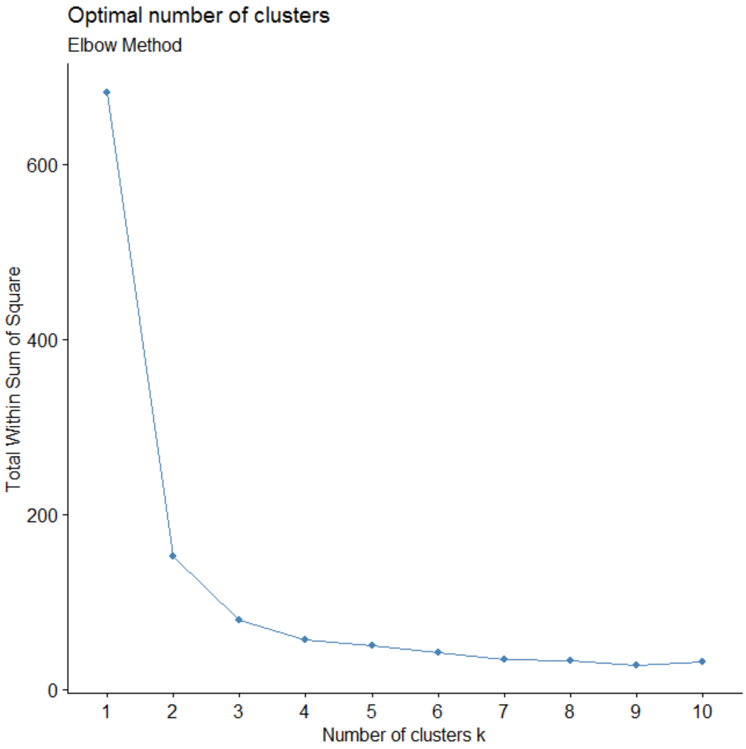
Interpretation

Now we will predict k-means cluster using several methods for Iris dataset and visualize it.

1. **Elbow Method** The basic idea of k-mean cluster is to define a cluster such that intra-cluster variation is minimized. The Elbow method looks to the

total WSS as a function of the number of cluster. One should choose number of cluster so that adding another cluster doesn't improve much better total WSS. while plotting it, location of bend (knee) in the plot is generally considered as an indicator of optomal number of cluster.

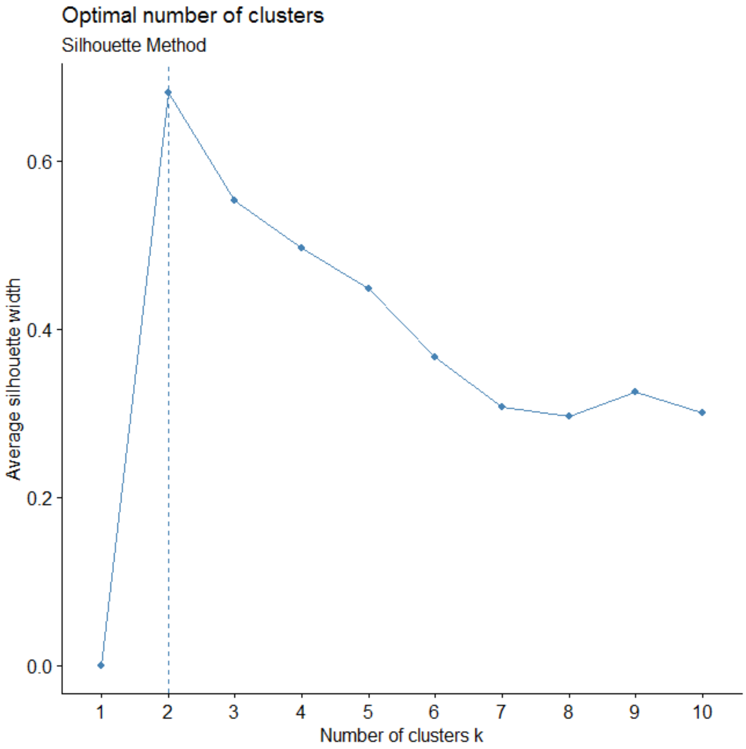
```
fviz_nbclust(Data,kmeans,method = "wss")+labs(subtitle = "Elbow Method")
```



Here, Elbow method suggest 4 cluster as optimal number of cluster.

2. **Silhouette Method** measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average Silhouette width indicates a good clustering. The optimal number of cluster k is the one that maximize the average Silhouette over a range of possible values for k.

```
fviz_nbclust(Data,kmeans,method = "silhouette")+labs(subtitle = "Silhouette Method")
```

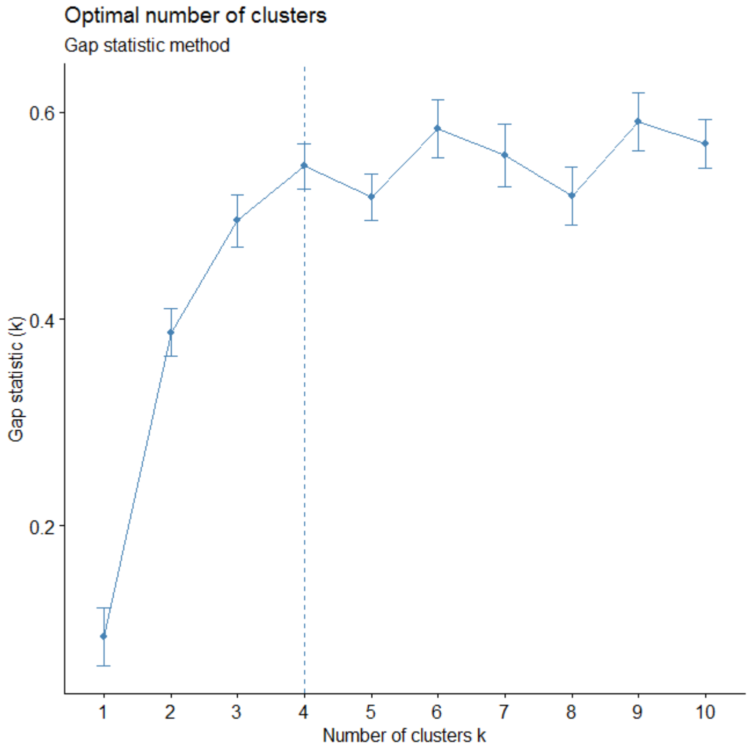


Here, Silhouette method suggest 2 cluster as optimal number of cluster.

3. **Gap statistic Method** compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic.

```
fviz_nbclust(Data,kmeans,method = "gap_stat",nboot = 50)+labs(subtitle =
 "Gap statistic method")
```

```
Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 50)  [one "." per sample]:
.................................................. 50
```



Gap statistic method suggest 4 cluster as optimal number of cluster.

Using PCA plot the majority of variance.

```
set.seed(120)
k<-kmeans(Data,3,nstart = 20);k
fviz_cluster(k,Data)
```

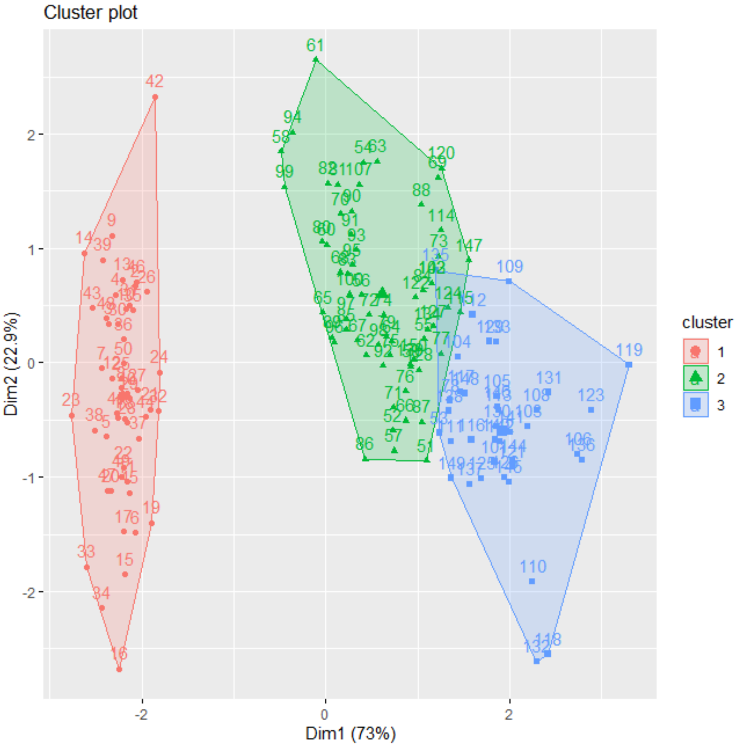
```
K-means clustering with 3 clusters of sizes 50, 62, 38

Cluster means:
Sepal.Length Sepal.Width Petal.Length Petal.Width
1      5.006000    3.428000     1.462000    0.246000
2      5.901613    2.748387     4.393548    1.433871
3      6.850000    3.073684     5.742105    2.071053

Clustering vector:
1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
```

```
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   2   2   3   2
 55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
 73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
  2   2   2   2   2   3   2   2   2   2   2   2   2   2   2   2   2   2
 91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
  2   2   2   2   2   2   2   2   2   2   3   2   3   3   3   3   2   3
109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
  3   3   3   3   3   2   2   3   3   3   3   2   3   2   3   2   3   3
127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
  2   2   3   3   3   3   3   2   3   3   3   3   2   3   3   3   2   3
145 146 147 148 149 150
  3   3   2   3   3   2

Within cluster sum of squares by cluster:
[1] 15.15100 39.82097 23.87947
 (between_SS / total_SS =  88.4 %)
```


Cluster plot
Dim2 (22.9%)
Dim1 (73%)
cluster