Rituraj Anand **19BEC1089**

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Programme	:	**B Tech – ECE and ECM**	Semester	:	**Win 2022**
Course	:	**Essentials of Data Analytics Lab**	Code	:	**CSE3506**
Faculty	:	**Gobinath N**	Slot	:	**L51 + L52**

Ex_06_K-Means Clustering

Importing packages:

```
rm(list = ls())
install.packages('cluster')
install.packages('ClusterR')
```

Setting working directories and reading csv file:

```
setwd("C:\\Users\\Rituraj Anand\\Desktop\\Sem6\\CSE3506\\LAB\\Lab 6")
dt=read.csv("seeds_K Means.csv")
```

summary(dt)

```
> summary(dt)
       ID              area          perimeter      compactness     lengthOfKernel  widthOfKernel   asymmetryCoefficient
 Min.   :  1.00   Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899   Min.   :2.630   Min.   :0.7651
 1st Qu.: 53.25   1st Qu.:12.27   1st Qu.:13.45   1st Qu.:0.8569   1st Qu.:5.262   1st Qu.:2.944   1st Qu.:2.5615
 Median :105.50   Median :14.36   Median :14.32   Median :0.8734   Median :5.524   Median :3.237   Median :3.5990
 Mean   :105.50   Mean   :14.85   Mean   :14.56   Mean   :0.8710   Mean   :5.629   Mean   :3.259   Mean   :3.7002
 3rd Qu.:157.75   3rd Qu.:17.30   3rd Qu.:15.71   3rd Qu.:0.8878   3rd Qu.:5.980   3rd Qu.:3.562   3rd Qu.:4.7687
 Max.   :210.00   Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.675   Max.   :4.033   Max.   :8.4560
 lengthOfKernelGroove    seedType
 Min.   :4.519        Min.   :1
 1st Qu.:5.045        1st Qu.:1
 Median :5.223        Median :2
 Mean   :5.408        Mean   :2
 3rd Qu.:5.877        3rd Qu.:3
 Max.   :6.550        Max.   :3
```

seed distribution based on compactness:

```
> dt$system=as.factor(dt$compactness)
> str(dt)
'data.frame':	210 obs. of  10 variables:
 $ ID                  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ area                : num  15.3 14.9 14.3 13.8 16.1 ...
 $ perimeter           : num  14.8 14.6 14.1 13.9 15 ...
 $ compactness         : num  0.871 0.881 0.905 0.895 0.903 ...
 $ lengthOfKernel      : num  5.76 5.55 5.29 5.32 5.66 ...
 $ widthOfKernel       : num  3.31 3.33 3.34 3.38 3.56 ...
 $ asymmetryCoefficient: num  2.22 1.02 2.7 2.26 1.35 ...
 $ lengthOfKernelGroove: num  5.22 4.96 4.83 4.8 5.17 ...
 $ seedType            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ system              : Factor w/ 186 levels "0.8081","0.8082",..: 88 118 176 155 174 154 115 148 99 141 ...
> |
```

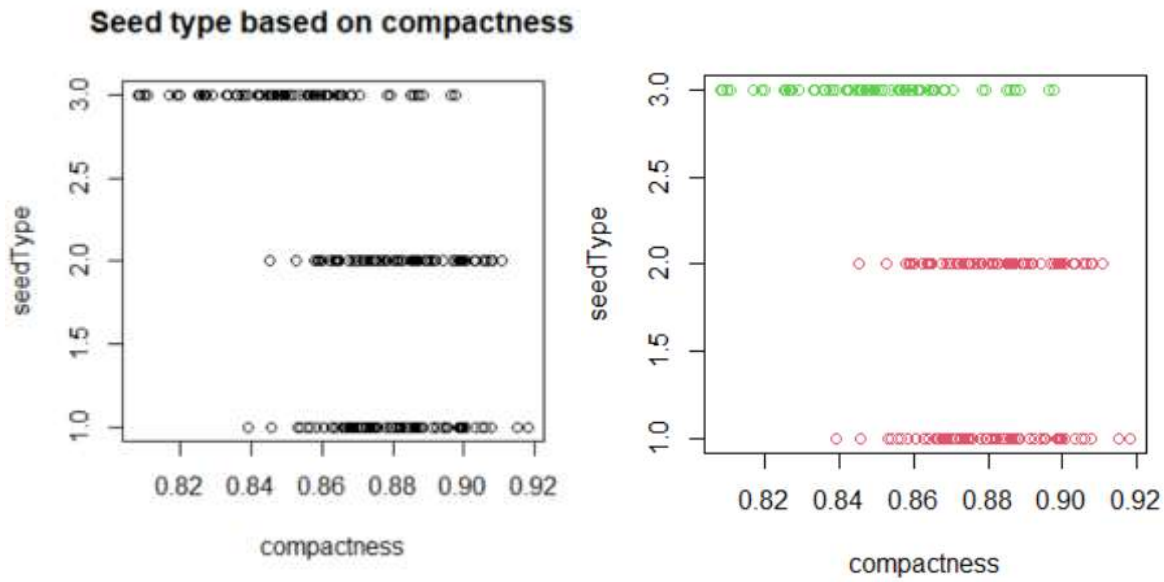
K-mean clustering with these two

```
> #seed distr based on compactness
> pdt=dt[,c(4,9)]
> View(pdt)
> plot(pdt,main="Price based on compactness")
> km=kmeans(pdt,2)    #
> View(km)
> plot(pdt,col=(km$cluster+1)) # when k(  ,3)
> km
K-means clustering with 2 clusters of sizes 140, 70

Cluster means:
  compactness seedType
1   0.8817936      1.5
2   0.8494086      3.0

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [64] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[127] 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 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[190] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 35.03508132  0.03267023
 (between_SS / total_SS =  75.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

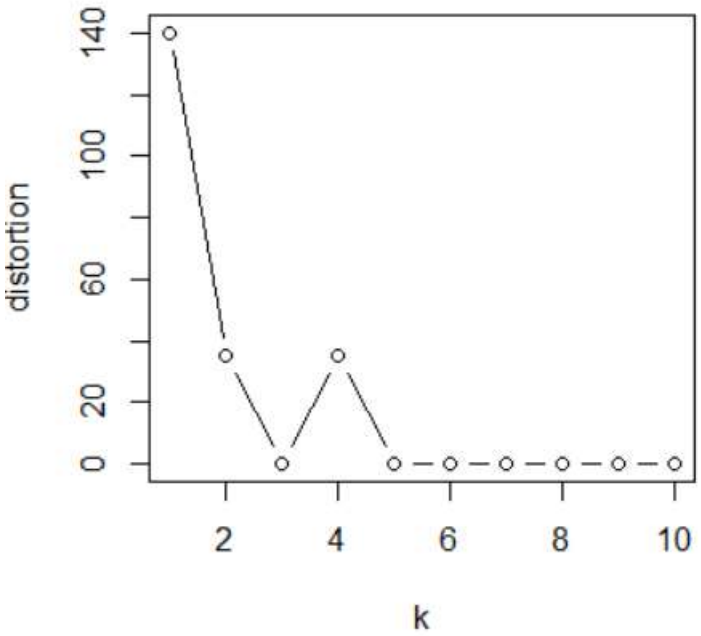
	compactness	seedType
1	0.8710	1
2	0.8811	1
3	0.9050	1
4	0.8955	1
5	0.9034	1
6	0.8951	1
7	0.8799	1
8	0.8911	1
9	0.8747	1
10	0.8880	1
11	0.8696	1
12	0.8796	1
13	0.8880	1
14	0.8759	1
15	0.8744	1
16	0.8993	1

Showing 1 to 17 of 210 entries, 2 total columns

Since, there are three types of seeds,



We can see the distortion:



Elbow point~=3

Appendix:

Code:

```
rm(list = ls())
#install.packages('cluster')
#install.packages('ClusterR')
setwd("C:\\Users\\Rituraj Anand\\Desktop\\Sem6\\CSE3506\\LAB\\Lab 6")
dt=read.csv("seeds_K Means.csv")
View(dt)
summary(dt)
dt$system=as.factor(dt$compactness)
str(dt)

#seed distr based on compactness
pdt=dt[,c(4,9)]
plot(pdt,main="Seed type based on compactness")
km=kmeans(pdt,2)    #
plot(pdt,col=(km$cluster+1)) # when k(  ,3)
km

#checking for optimal 'k'
dt2=pdt
ss=(nrow(dt2)-1)*sum(apply(dt2,2,var))
for(i in 2:10) ss[i]=sum(kmeans(dt2,centers = i)$withinss)

plot(1:10,ss,type = 'b',xlab='k',ylab='distortion')
```

Result and Inference:

As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid.

Hence, we saw the seed types and its distribution with proper inference through the distortion curve.