Homework 5

Vaishnavi Mule **A20516627**

1 Recitation Exercises

Chapter 12

Exercises: 1

1.a)

a] Eq^n $= \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$

where, $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$, is the mean of feature j in cluster C_k. Expanding LHS, we get $\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$

$$= \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{ij}^2 + \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{i'j}^2 - \frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{ij}x_{i'j}$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - \frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{ij}x_{i'j}$$

Expanding RHS further & substituting the value of $\bar{x}_{kj}$, we get

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 4\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}\bar{x}_{kj} + 2\sum_{i\in C_k}\sum_{j=1}^{p}\bar{x}_{kj}^2$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 4|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 + 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - \frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{ij}x_{i'j}$$

Hence, LHS = RHS.

1.b)

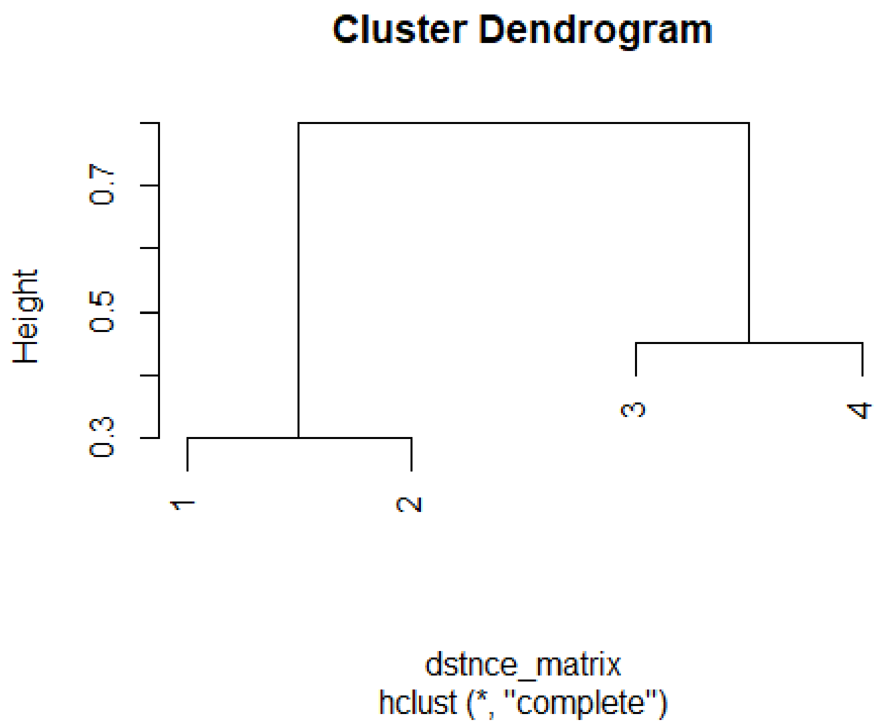
1) The equation shown above is used to minimize the sum of the squared Euclidean distance for each cluster. This is equivalent to minimizing the within-cluster variance for each cluster, which is what happens in K-means clustering.

2) During the initial step of each iteration, wherein we shift the centroid of each cluster towards the vector of the feature means, our objective is to minimize the sum of deviations to the center within each cluster. Subsequently, as we reassign observations to the nearest center, this process inherently leads to a reduction in the overall sum of deviations.

Exercises: 2

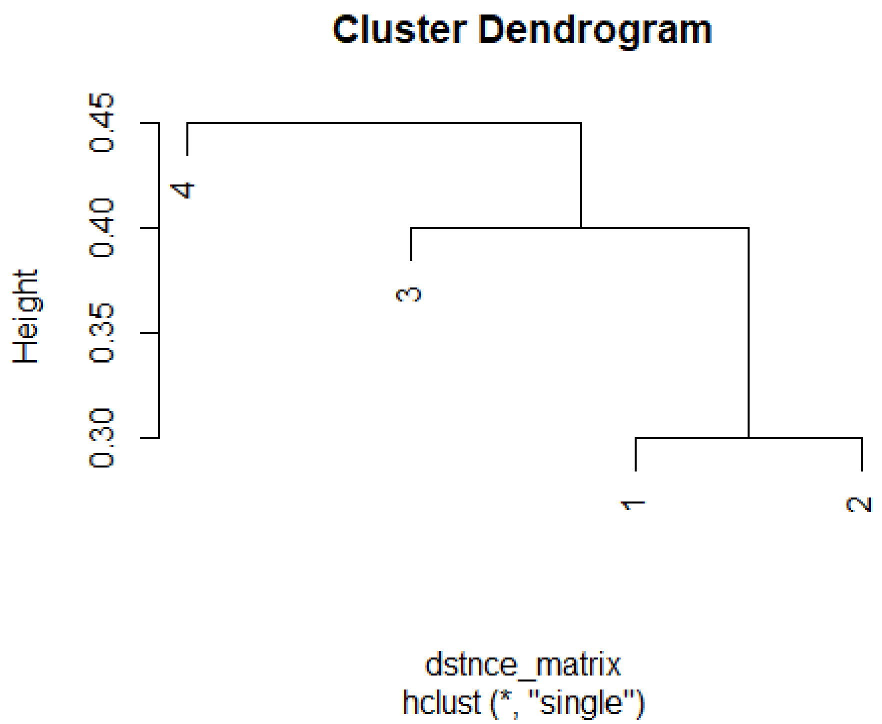
2.a)

```
dstnce_matrix = as.dist(matrix(c(0, 0.3, 0.4, 0.7,
                     0.3, 0, 0.5, 0.8,
                     0.4, 0.5, 0.0, 0.45,
                     0.7, 0.8, 0.45, 0.0), nrow=4))
hc_complt=hclust(dstnce_matrix, method="complete")

#Heights where the fusion occurs
print(hc_complt$height)

## [1] 0.30 0.45 0.80

plot(hc_complt)
```



2.b)

```
hclust_1 = hclust(dstnce_matrix, method="single")
#Heights where the fusion occurs
print(hclust_1$height)

## [1] 0.30 0.40 0.45

plot(hclust_1)
```



2.c)

```
hclust_complete_cut=cutree(hc_complt,k=2)
hclust_complete_cut

## [1] 1 1 2 2
```

The clusters we see from (a) are : (1,2), (3,4)

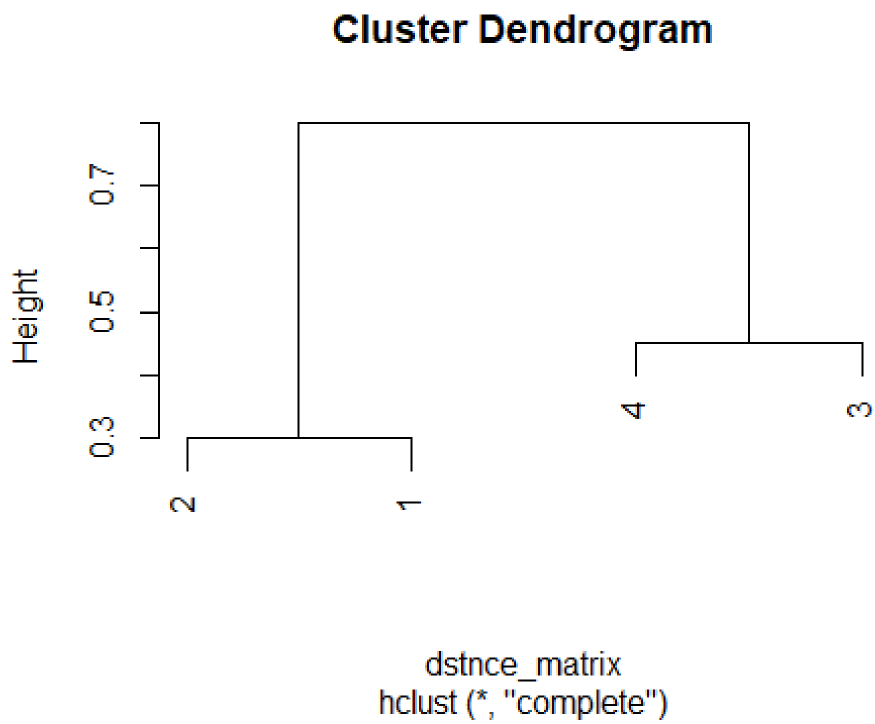
2.d)

```
hclust_single_cut=cutree(hclust_1,k=2)
hclust_single_cut

## [1] 1 1 1 2
```

The clusters we see from (b) are : ((1,2),3), (4)

2.e)

```
plot(hclust(dstnce_matrix, method="complete"), labels=c(2,1,4,3))
```



The above dendrogram is equivalent to the dendrogram of (a)

Exercises: 3

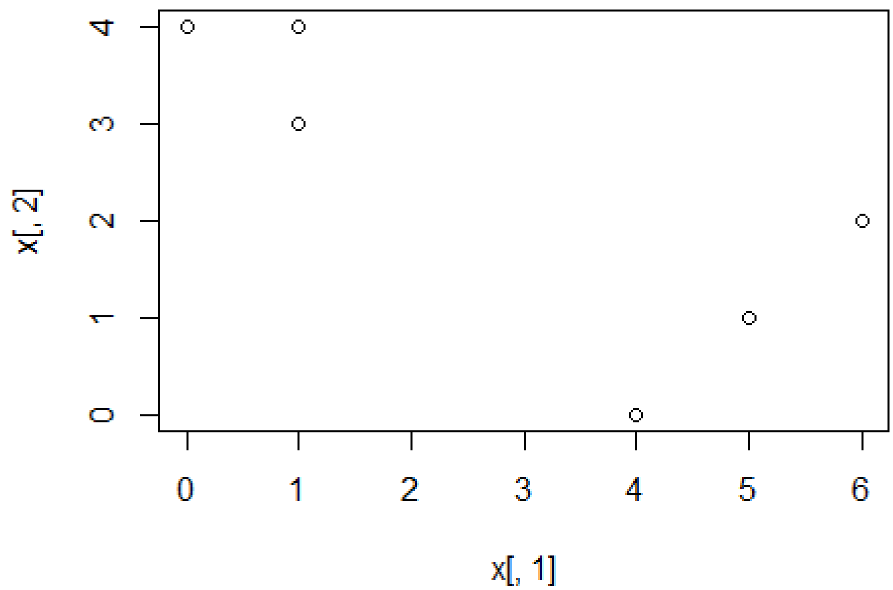
Given observations :

```
set.seed(1)
x = cbind(c(1, 1, 0, 5, 6, 4), c(4, 3, 4, 1, 2, 0))
x

##      [,1] [,2]
## [1,]    1    4
## [2,]    1    3
## [3,]    0    4
## [4,]    5    1
## [5,]    6    2
## [6,]    4    0
```

3.a)

plotting :

```
plot(x[,1], x[,2])
```



3.b)

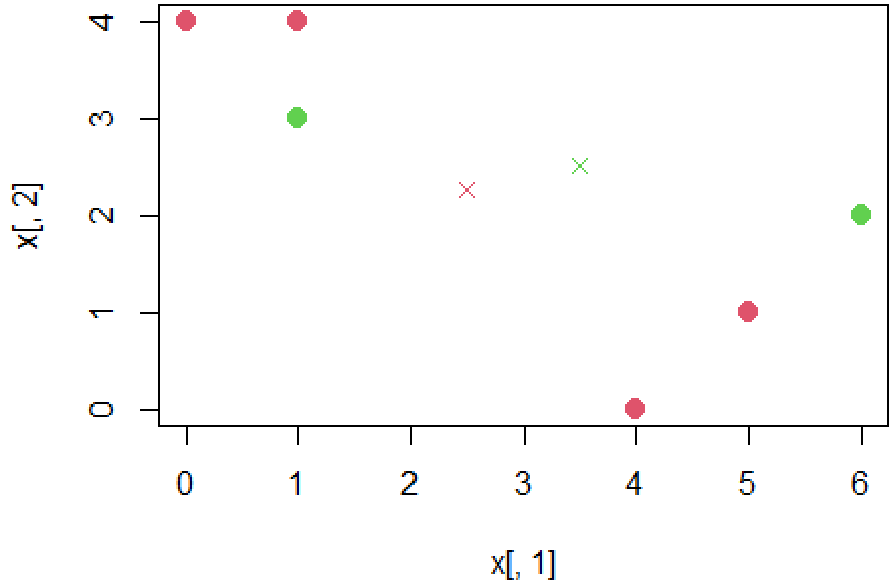
Assigning labels randomly :

```
labels = sample(2, nrow(x), replace=T)
labels

## [1] 1 2 1 1 2 1
```

3.c)

compute centroid for each cluster :

```
centroid1 = c(mean(x[labels==1, 1]), mean(x[labels==1, 2]))
centroid2 = c(mean(x[labels==2, 1]), mean(x[labels==2, 2]))
print(centroid1)

## [1] 2.50 2.25

print(centroid2)

## [1] 3.5 2.5

plot(x[,1], x[,2], col=(labels+1), pch=20, cex=2)
points(centroid1[1], centroid1[2], col=2, pch=4)
points(centroid2[1], centroid2[2], col=3, pch=4)
```



3.d)

assign oservations to centroids clostest with the help of euclidean distances calculated

```
euc_dis = function(a, b) {
  return(sqrt((a[1] - b[1])^2 + (a[2]-b[2])^2))
}

assn_labels = function(x, centroid1, centroid2) {
  labels = rep(NA, nrow(x))
  for (i in 1:nrow(x)) {
    if (euc_dis(x[i,], centroid1) < euc_dis(x[i,], centroid2)) {
      labels[i] = 1
    } else {
      labels[i] = 2
    }
  }
  return(labels)
}
#Function call
lbls = assn_labels(x, centroid1, centroid2)
print(lbls)

## [1] 1 1 1 2 2 2
```

3.e)

loop till values doen't change :

```
last_labels = rep(-1, 6)
while (!all(last_labels == lbls)) {
  last_labels = lbls
  centroid1 = c(mean(x[lbls==1, 1]), mean(x[lbls==1, 2]))
  centroid2 = c(mean(x[lbls==2, 1]), mean(x[lbls==2, 2]))
  print(centroid1)
  print(centroid2)
  lbls = assn_labels(x, centroid1, centroid2)
}
```

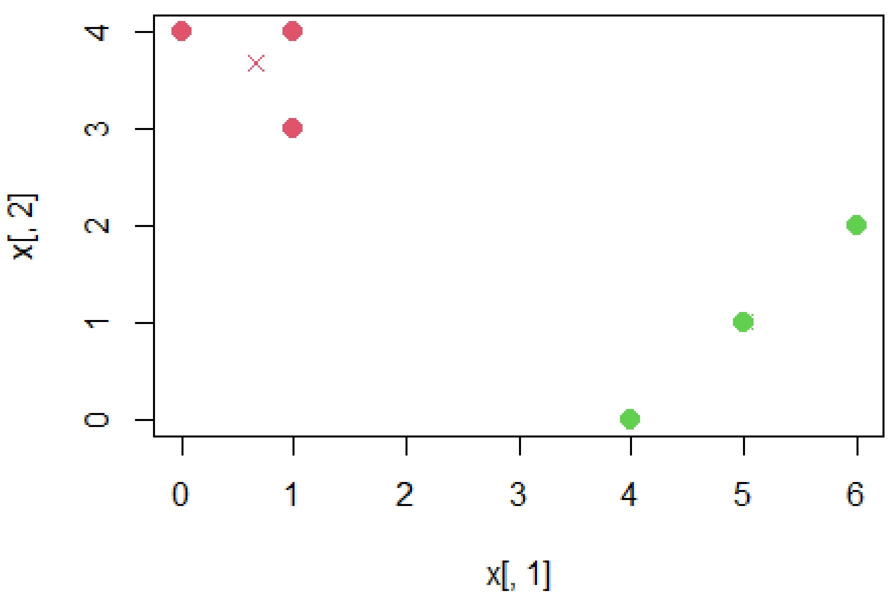
```
## [1] 0.6666667 3.6666667
## [1] 5 1
```

```
print(lbls)
```

```
## [1] 1 1 1 2 2 2
```

3.f)

```
plot(x[,1], x[,2], col=(lbls+1), pch=20, cex=2)
points(centroid1[1], centroid1[2], col=2, pch=4)
points(centroid2[1], centroid2[2], col=3, pch=4)
```



Exercises: 4

4.a)

It is difficult to determine which fusion will occur at a higher position on the tree due to a lack of information. In problem 2 presented earlier, three clusters fused at varying heights, which is dependent on the dissimilarity matrix. If the dissimilarities are identical, they would fuse at the same height, but if not, the single linkage dendrogram would typically fuse at a lower height.

4.b)

The linkage method only impacts how clusters are merged and does not affect how the leaf nodes of the tree merge. The leaf nodes would merge at the same height in both single and complete linkage dendrograms.

2 Practicum Problems

2.1 Problem 1

```
# making scipen=999 to prevent scientific notation when calulating variances
and means (to remove e^-2 etc)
options(scipen = 999)
#Loading Wine.data
wine_df=read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databas
es/wine/wine.data"),sep=",",header=F)
column_s=c('Alcohol','Malic acid','Ash','Alcalinity of ash','Magnesium','Tota
l phenols','Flavanoids','Nonflavanoid','phenols','Proanthocyanins','Color int
ensity','Hue','OD280/OD315 of diluted wines','Proline')

#looking at summary
summary(wine_df)

##        V1              V2              V3              V4
##  Min.   :1.000   Min.   :11.03   Min.   :0.740   Min.   :1.360
##  1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210
##  Median :2.000   Median :13.05   Median :1.865   Median :2.360
##  Mean   :1.938   Mean   :13.00   Mean   :2.336   Mean   :2.367
##  3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558
##  Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230
##        V5              V6               V7              V8
##  Min.   :10.60   Min.   : 70.00   Min.   :0.980   Min.   :0.340
##  1st Qu.:17.20   1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205
```

```
##  Median :19.50   Median : 98.00   Median :2.355   Median :2.135  
##  Mean   :19.49   Mean   : 99.74   Mean   :2.295   Mean   :2.029  
##  3rd Qu.:21.50   3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875  
##  Max.   :30.00   Max.   :162.00   Max.   :3.880   Max.   :5.080  
##       V9              V10              V11              V12        
##  Min.   :0.1300   Min.   :0.410   Min.   : 1.280   Min.   :0.4800  
##  1st Qu.:0.2700   1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825  
##  Median :0.3400   Median :1.555   Median : 4.690   Median :0.9650  
##  Mean   :0.3619   Mean   :1.591   Mean   : 5.058   Mean   :0.9574  
##  3rd Qu.:0.4375   3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200  
##  Max.   :0.6600   Max.   :3.580   Max.   :13.000   Max.   :1.7100  
##      V13             V14        
##  Min.   :1.270   Min.   : 278.0  
##  1st Qu.:1.938   1st Qu.: 500.5  
##  Median :2.780   Median : 673.5  
##  Mean   :2.612   Mean   : 746.9  
##  3rd Qu.:3.170   3rd Qu.: 985.0  
##  Max.   :4.000   Max.   :1680.0
```

```
colnames(wine_df)=column_s

#checking column names
print(names(wine_df))
```

```
##  [1] "Alcohol"                      "Malic acid"                  
##  [3] "Ash"                          "Alcalinity of ash"           
##  [5] "Magnesium"                    "Total phenols"               
##  [7] "Flavanoids"                   "Nonflavanoid"                
##  [9] "phenols"                      "Proanthocyanins"             
## [11] "Color intensity"              "Hue"                         
## [13] "OD280/OD315 of diluted wines" "Proline"
```

```
#checking the mean of the columns in data
print(apply(wine_df, 2, mean))
```

```
##                      Alcohol                   Malic acid 
##                    1.9382022                   13.0006180 
##                          Ash            Alcalinity of ash 
##                    2.3363483                    2.3665169 
##                    Magnesium                Total phenols 
##                   19.4949438                   99.7415730 
##                   Flavanoids                 Nonflavanoid 
##                    2.2951124                    2.0292697 
##                      phenols              Proanthocyanins 
##                    0.3618539                    1.5908989 
##              Color intensity                          Hue 
##                    5.0580899                    0.9574494 
## OD280/OD315 of diluted wines                      Proline 
##                    2.6116854                  746.8932584
```

```r
#checking Variance of the columns in data
print(apply(wine_df, 2, var))
```

```
##                      Alcohol                   Malic acid
##                   0.60067924                   0.65906233
##                          Ash            Alcalinity of ash
##                   1.24801540                   0.07526464
##                    Magnesium                 Total phenols
##                  11.15268616                 203.98933536
##                   Flavanoids                  Nonflavanoid
##                   0.39168954                   0.99771867
##                      phenols              Proanthocyanins
##                   0.01548863                   0.32759467
##              Color intensity                          Hue
##                   5.37444938                   0.05224496
## OD280/OD315 of diluted wines                      Proline
##                   0.50408641               99166.71735542
```

```r
# Performing PCA using prcomp with scaling of columns
pr.out=prcomp(wine_df , scale=TRUE)

# Checking values of center, scale and rotation matrix
print(pr.out$center)
```

```
##                      Alcohol                   Malic acid
##                    1.9382022                   13.0006180
##                          Ash            Alcalinity of ash
##                    2.3363483                    2.3665169
##                    Magnesium                 Total phenols
##                   19.4949438                   99.7415730
##                   Flavanoids                  Nonflavanoid
##                    2.2951124                    2.0292697
##                      phenols              Proanthocyanins
##                    0.3618539                    1.5908989
##              Color intensity                          Hue
##                    5.0580899                    0.9574494
## OD280/OD315 of diluted wines                      Proline
##                    2.6116854                  746.8932584
```

```r
print(pr.out$scale)
```

```
##                      Alcohol                   Malic acid
##                    0.7750350                    0.8118265
##                          Ash            Alcalinity of ash
##                    1.1171461                    0.2743440
##                    Magnesium                 Total phenols
##                    3.3395638                   14.2824835
##                   Flavanoids                  Nonflavanoid
##                    0.6258510                    0.9988587
##                      phenols              Proanthocyanins
##                    0.1244533                    0.5723589
```

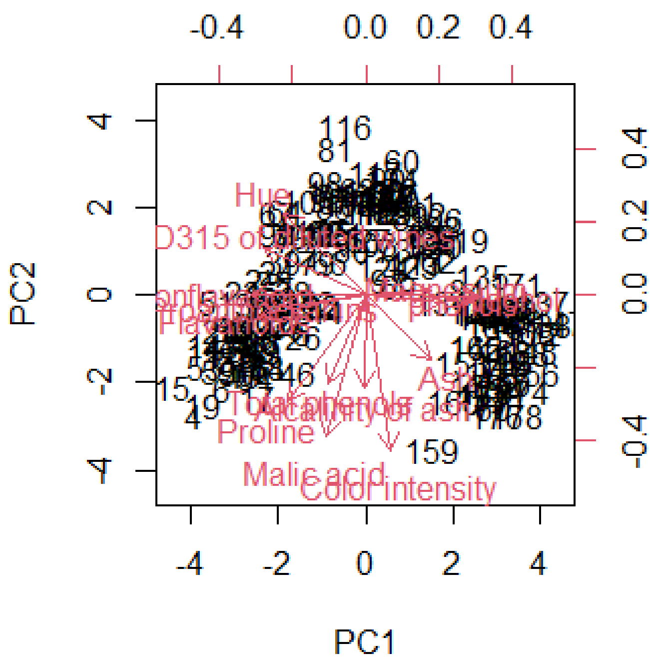
```
##                Color intensity                          Hue
##                      2.3182859                    0.2285716
## OD280/OD315 of diluted wines                      Proline
##                      0.7099904                  314.9074743
```

```
print(pr.out$rotation)
```

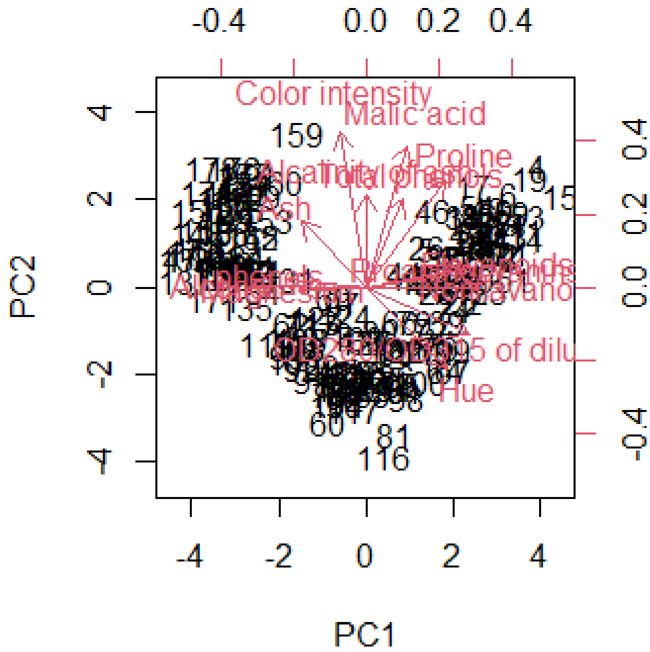
```
##                                        PC1          PC2          PC3         
PC4
## Alcohol                       0.393669533 -0.005690412  0.001217953 -0.122
46373
## Malic acid                   -0.136325011 -0.484160868 -0.207400812  0.081
91848
## Ash                           0.222676383 -0.223590947  0.088796064 -0.469
88824
## Alcalinity of ash            -0.002257932 -0.315855884  0.626102363  0.249
84122
## Magnesium                     0.224298489  0.011615737  0.611989600 -0.071
99322
## Total phenols                -0.124630159 -0.300551432  0.130984580  0.163
21412
## Flavanoids                   -0.359264042 -0.067119829  0.146507749 -0.190
98521
## Nonflavanoid                 -0.390711715  0.001313454  0.150962746 -0.144
61667
## phenols                       0.267001203 -0.026988703  0.169975512  0.328
01272
## Proanthocyanins              -0.279062504 -0.041222563  0.149879586 -0.462
75771
## Color intensity               0.089318293 -0.529782740 -0.137266298 -0.072
11248
## Hue                          -0.276822650  0.277907354  0.085328539  0.434
66618
## OD280/OD315 of diluted wines -0.350526181  0.162776250  0.166204360 -0.156
72341
## Proline                      -0.269515252 -0.366058862 -0.126686846  0.255
79490
##                                      PC5         PC6         PC7          P
C8
## Alcohol                       0.15758395 -0.20033864  0.05938234 -0.071795
53
## Malic acid                   -0.25089415  0.13517139  0.09269887 -0.421544
35
## Ash                          -0.18860015  0.59841948 -0.37436980 -0.087575
56
## Alcalinity of ash            -0.09352360  0.10799983  0.16708856  0.172080
34
## Magnesium                     0.04656750 -0.08811224  0.26872469 -0.413248
57
## Total phenols                 0.77833048  0.14483831 -0.32957951  0.148811
```

```
89
## Flavanoids                   -0.14466563 -0.14809748  0.03789829  0.363438
84
## Nonflavanoid                 -0.11200553 -0.06247252  0.06773223  0.175405
00
## phenols                      -0.43257916 -0.25868639 -0.61111195  0.230751
35
## Proanthocyanins               0.09158820 -0.46627764 -0.42292282 -0.343739
20
## Color intensity              -0.04626960 -0.42525454  0.18613617  0.040696
17
## Hue                          -0.02986657  0.01565089 -0.19204101 -0.483625
64
## OD280/OD315 of diluted wines -0.14419358  0.21770365  0.07850980  0.068651
16
## Proline                      -0.08440794  0.06656550 -0.05420370 -0.111466
71
##                                       PC9        PC10        PC11         P
C12
## Alcohol                      -0.162368819  0.19899373 -0.01444169  0.01575
769
## Malic acid                   -0.450190708 -0.31127983  0.22154641 -0.26411
262
## Ash                          -0.006025687  0.32592413 -0.06839251  0.11921
210
## Alcalinity of ash             0.262494455  0.12452347  0.49452428 -0.04502
305
## Magnesium                    -0.118633417 -0.15716811 -0.47461722 -0.06131
271
## Total phenols                -0.252536278 -0.12773363 -0.07119731  0.06116
074
## Flavanoids                   -0.406373544  0.30772263 -0.29740957 -0.30087
591
## Nonflavanoid                 -0.090919334  0.14044000  0.03219187 -0.05001
396
## phenols                      -0.159122818 -0.24054263 -0.12200984  0.04266
558
## Proanthocyanins               0.265786794 -0.10869629  0.23292405 -0.09334
264
## Color intensity              -0.075264592  0.21704255 -0.01972448  0.59795
428
## Hue                          -0.212416815  0.50966073  0.06140493  0.25774
292
## OD280/OD315 of diluted wines -0.084264837 -0.45570504 -0.06646166  0.61109
218
## Proline                       0.544905394  0.04620802 -0.55130818 -0.07268
036
##                                     PC13         PC14
## Alcohol                      -0.49224318 -0.669045280
## Malic acid                   -0.05610645 -0.090626055
```

```
## Ash                           0.06675544  0.025225306
## Alcalinity of ash            -0.19201787  0.001635816
## Magnesium                     0.20007784  0.095361066
## Total phenols                 0.05829909 -0.022300745
## Flavanoids                   -0.35952714  0.253037788
## Nonflavanoid                  0.59834288 -0.601909165
## phenols                       0.06403952 -0.082230935
## Proanthocyanins              -0.11013538  0.058641979
## Color intensity               0.15917751  0.178821145
## Hue                          -0.04923091  0.022582562
## OD280/OD315 of diluted wines -0.32941979 -0.135092159
## Proline                      -0.17322892 -0.216043617

# Plotting first two principal components
biplot(pr.out,scale=0)
```

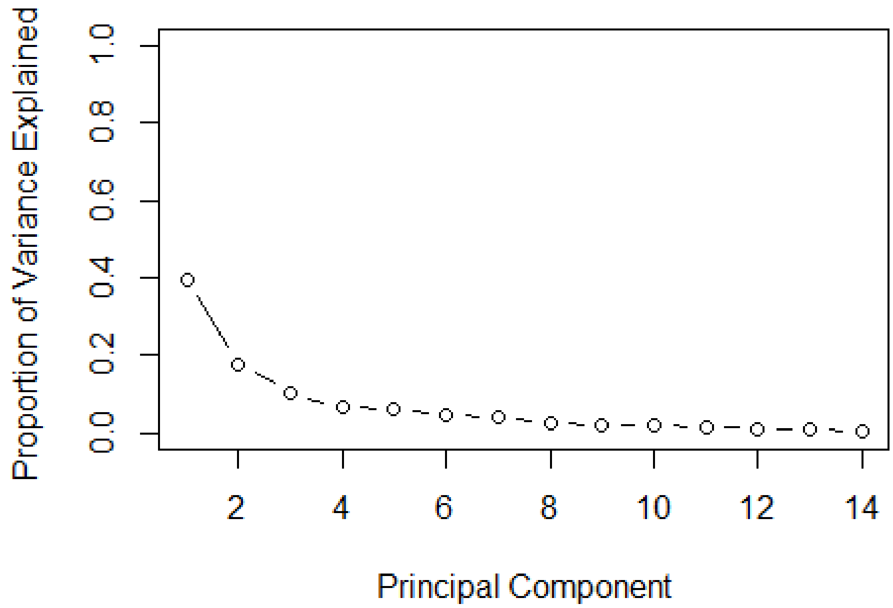



```
# making sign change to produce a mirror image
pr.out$rotation=-pr.out$rotation
pr.out$x=-pr.out$x

biplot (pr.out , scale =0)
```

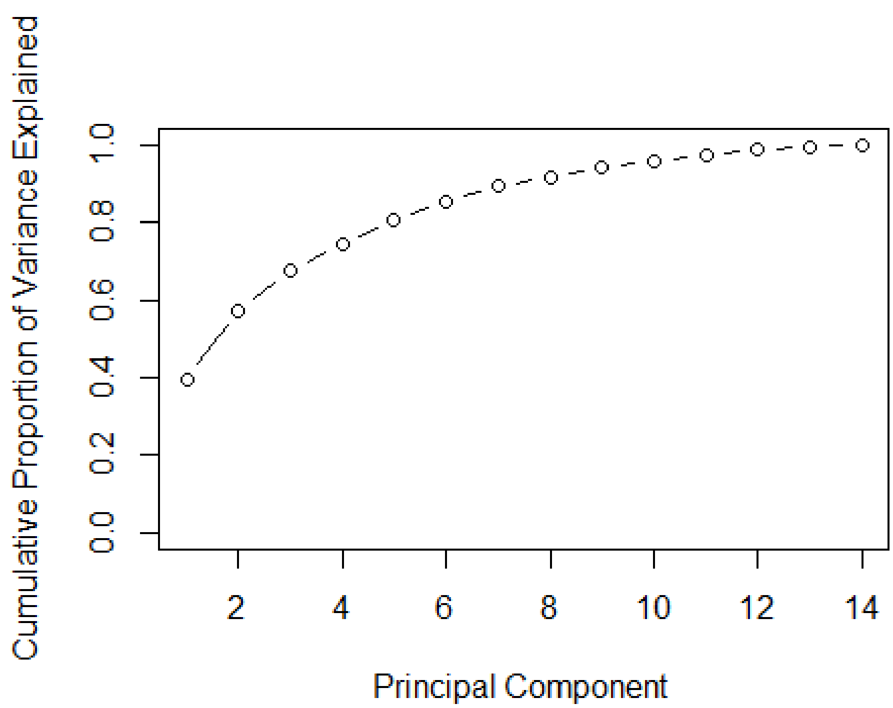


As the feature ash is pointed on the opposite direction of hue it is seen that they are inversely correlated.

```
# Calculate variance explained by each principal component
pr_var=pr.out$sdev ^2
print(pr_var)

##  [1] 5.53594804 2.49707625 1.44607422 0.92791783 0.87750252 0.67277834
##  [7] 0.55379896 0.35003417 0.29454194 0.26230610 0.22584842 0.16879672
## [13] 0.12956418 0.05781232

# proportion of variance explained by each principal component
p=pr_var/sum(pr_var)
print(p)

##  [1] 0.395424860 0.178362589 0.103291016 0.066279845 0.062678751 0.0480555
96
##  [7] 0.039557068 0.025002441 0.021038710 0.018736150 0.016132030 0.0120569
08
## [13] 0.009254584 0.004129451

#Plot Proportion explained by each PC
plot(p , xlab=" Principal Component ", ylab="Proportion of Variance Explained
", ylim=c(0,1),type="b")
```




```
plot(cumsum(p), xlab="Principal Component ", ylab="Cumulative Proportion of V
ariance Explained",ylim=c(0,1),type="b")
```



The prcomp function is preferred over princomp for numerical computation because it uses n-1 instead of n to perform calculations. Scaling of variables is considered a best practice, even when the deviation of each feature is not large. Scaling helps when a few features have a large variance, as it brings all features onto a unit scale. This ensures that the principal components are not dominated by features with large variance.

2.2 Problem 2

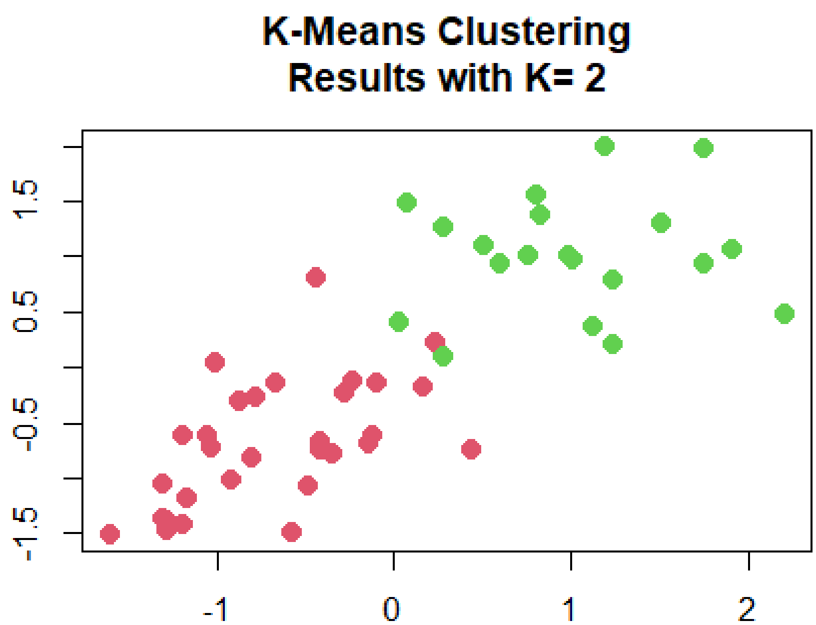
```
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.2.3

## Welcome! Want to learn more? See two factoextra-related books at https://g
oo.gl/ve3WBa

library(collections)

## Warning: package 'collections' was built under R version 4.2.3

##
## Attaching package: 'collections'

## The following object is masked from 'package:utils':
##
##     stack

set.seed(20)
#Loading USArrests data
data("USArrests")

summary(USArrests)

##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00

# Checking means and variances of columns
print(apply(USArrests,2,mean))

##   Murder  Assault UrbanPop     Rape
##    7.788  170.760   65.540   21.232

print(apply(USArrests,2,var))

##     Murder    Assault   UrbanPop       Rape
##   18.97047 6945.16571  209.51878   87.72916
```

Applying scaling to the columns is advisable due to the noticeable variance differences between the features. Scaling will standardize all features to a unit normal space. This is important because the K-means algorithm is isotropic, and scaling prevents any one feature with larger magnitudes from exerting undue influence over the algorithm compared to other features.

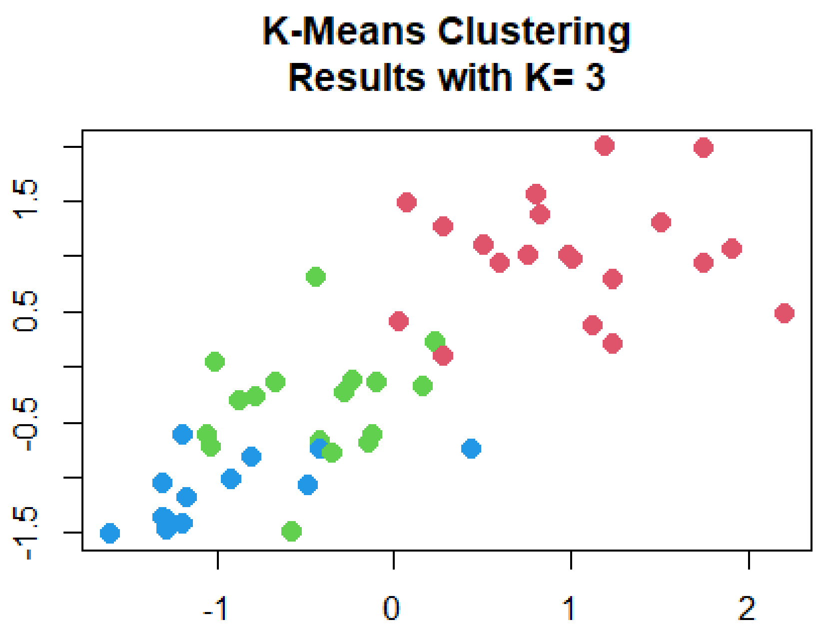
```
#Scaling the dataset
scaled_data = scale(USArrests,center = TRUE,scale=TRUE)

dict_ss=dict()

iter_in=c(2:10)

# Perform K means clustering for k=2 to k=10 and plot the clusters
for (val in iter_in){
  km.out = kmeans(scaled_data,val,nstart = 30)
  print(km.out$cluster)

  #Basic plotting for cluster
  plot(scaled_data, col=(km.out$cluster +1), main=paste("K-Means Clustering
Results with K=",val,sep=" "), xlab="", ylab="", pch=20, cex=2)

  #fviz_cluster to visualize clusters
  print(fviz_cluster(km.out, data = scaled_data,geom = "point",ellipse.type =
"convex",ggtheme = theme_bw()))

  #Store within cluster sum of squares values and the respective K value in a
dict
  w = km.out$tot.withinss
  dict_ss$set(val,w)
}
```

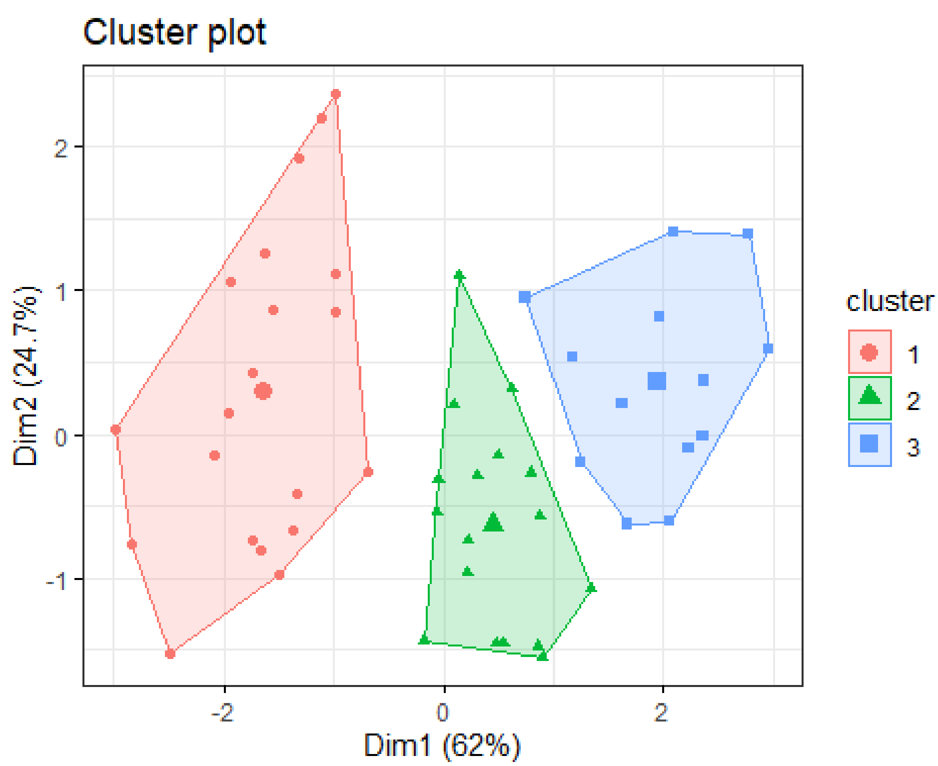
```
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              2              2              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              2              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              2              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              2              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              2              1              2              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              2              1              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              2              1              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              1              1              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
```

```
##              1              2              2              1              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              1              1              1
```



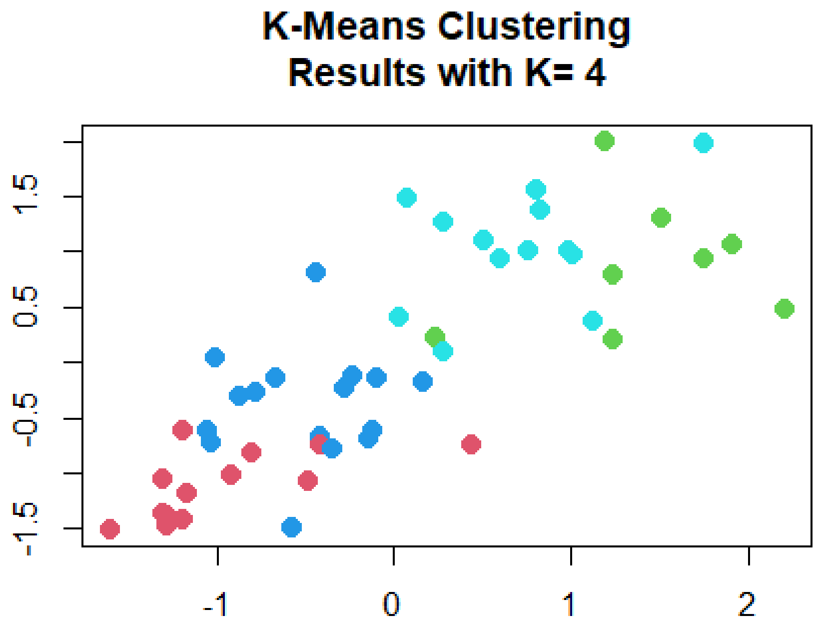
```
##        Alabama         Alaska        Arizona       Arkansas     California 
##              1              1              1              2              1 
##       Colorado    Connecticut       Delaware        Florida        Georgia 
##              1              2              2              1              1 
##         Hawaii          Idaho       Illinois        Indiana           Iowa 
##              2              3              1              2              3 
##         Kansas       Kentucky      Louisiana          Maine       Maryland 
##              2              3              1              3              1 
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##              2              1              3              1              1 
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##              3              3              1              3              2 
##     New Mexico       New York North Carolina   North Dakota           Ohio 
##              1              1              1              3              2 
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##              2              2              2              2              1 
##   South Dakota      Tennessee          Texas           Utah        Vermont 
##              3              1              1              2              3 
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##              2              2              3              3              2
```

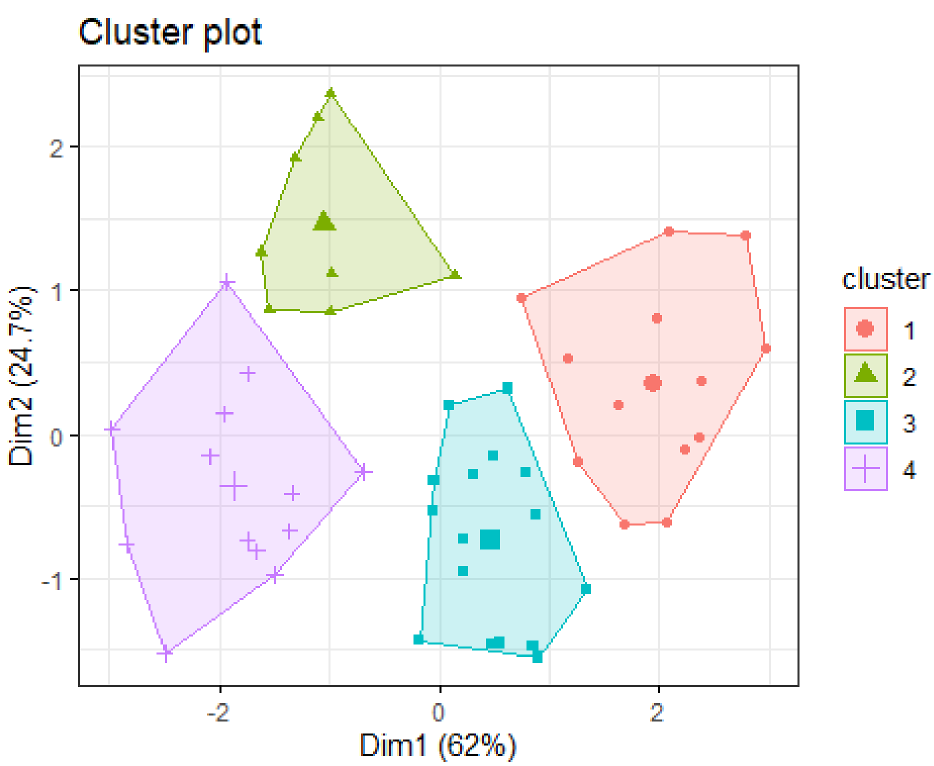





```
##          Alabama        Alaska       Arizona      Arkansas    California
##                2             4             4             2             4
##         Colorado   Connecticut      Delaware       Florida       Georgia
##                4             3             3             4             2
```

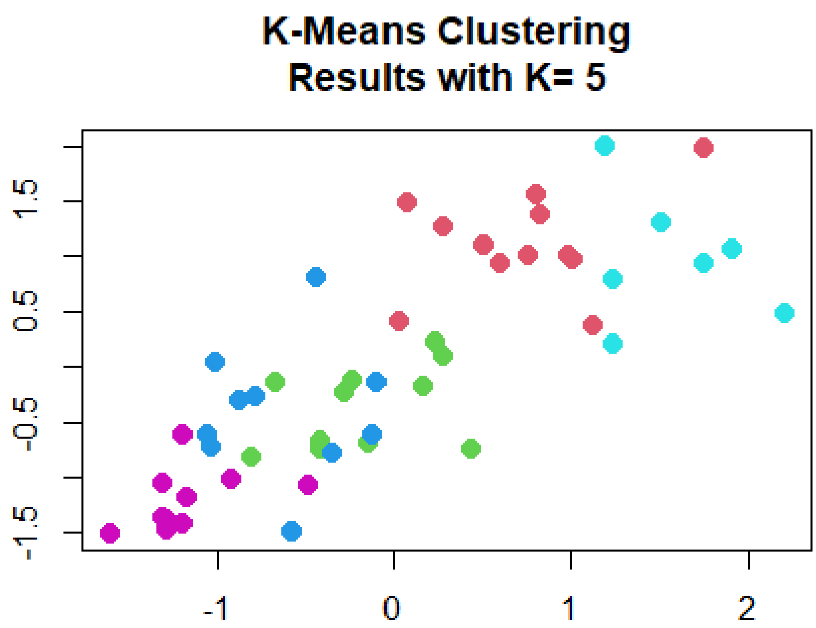
```
##         Hawaii          Idaho       Illinois        Indiana           Iowa 
##              3              1              4              3              1 
##         Kansas       Kentucky      Louisiana          Maine       Maryland 
##              3              1              2              1              4 
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##              3              4              1              2              4 
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##              1              1              4              1              3 
##     New Mexico       New York North Carolina   North Dakota           Ohio 
##              4              4              2              1              3 
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##              3              3              3              3              2 
##   South Dakota      Tennessee          Texas           Utah        Vermont 
##              1              2              4              3              1 
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##              3              3              1              1              3 
```

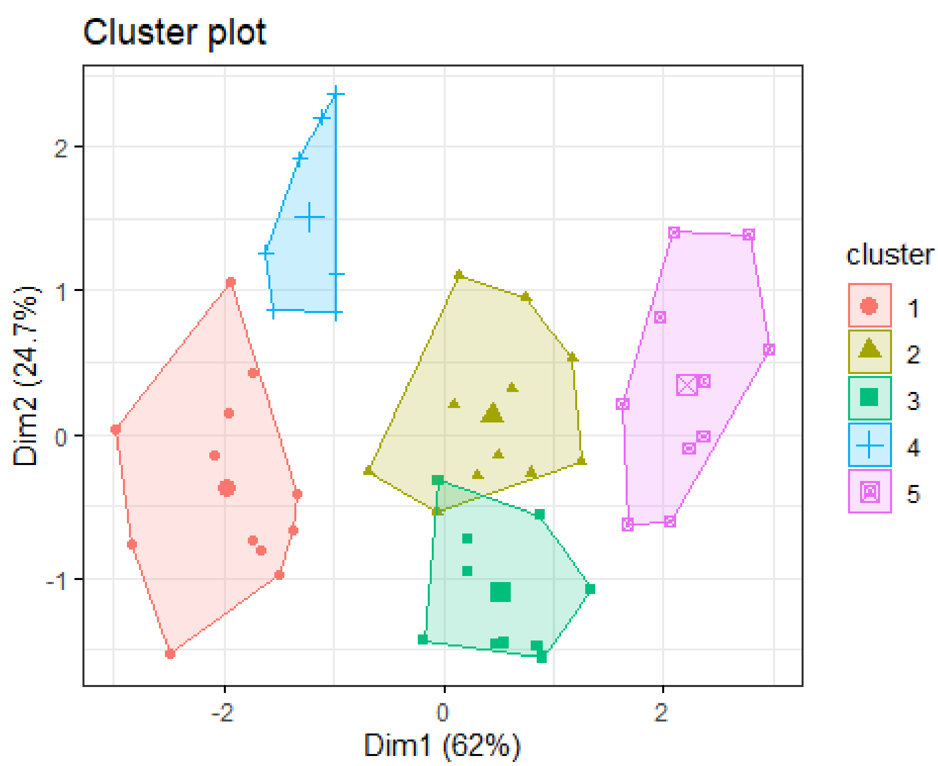





```
##          Alabama          Alaska         Arizona        Arkansas      California
##                4               1               1               2               1
##         Colorado     Connecticut        Delaware         Florida         Georgia
##                1               3               3               1               4
```

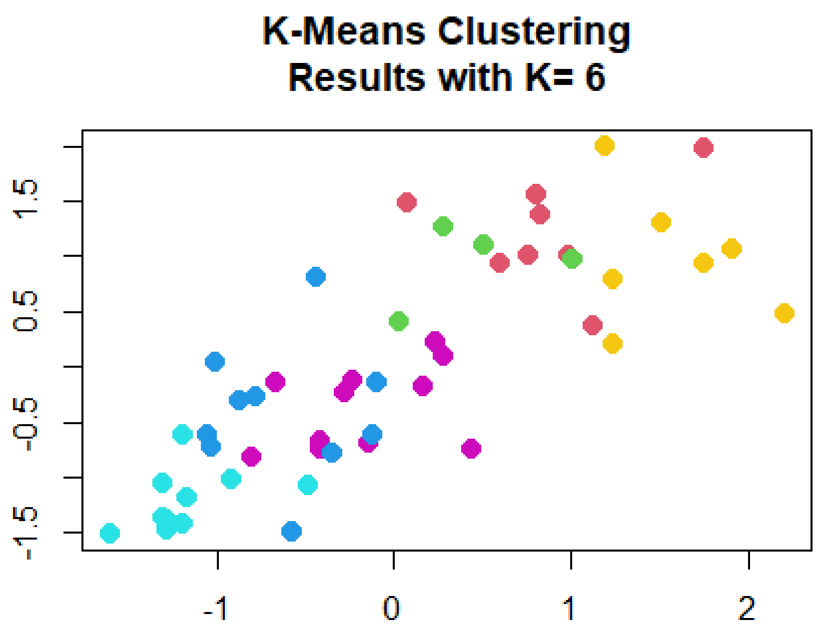
```
##          Hawaii          Idaho       Illinois        Indiana           Iowa 
##               3              5              1              2              5 
##          Kansas       Kentucky      Louisiana          Maine       Maryland 
##               2              2              4              5              1 
##   Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##               3              1              5              4              2 
##         Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##               2              2              1              5              3 
##      New Mexico       New York North Carolina   North Dakota           Ohio 
##               1              1              4              5              3 
##        Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##               2              2              3              3              4 
##    South Dakota      Tennessee          Texas           Utah        Vermont 
##               5              4              1              3              5 
##        Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##               2              3              5              5              2
```

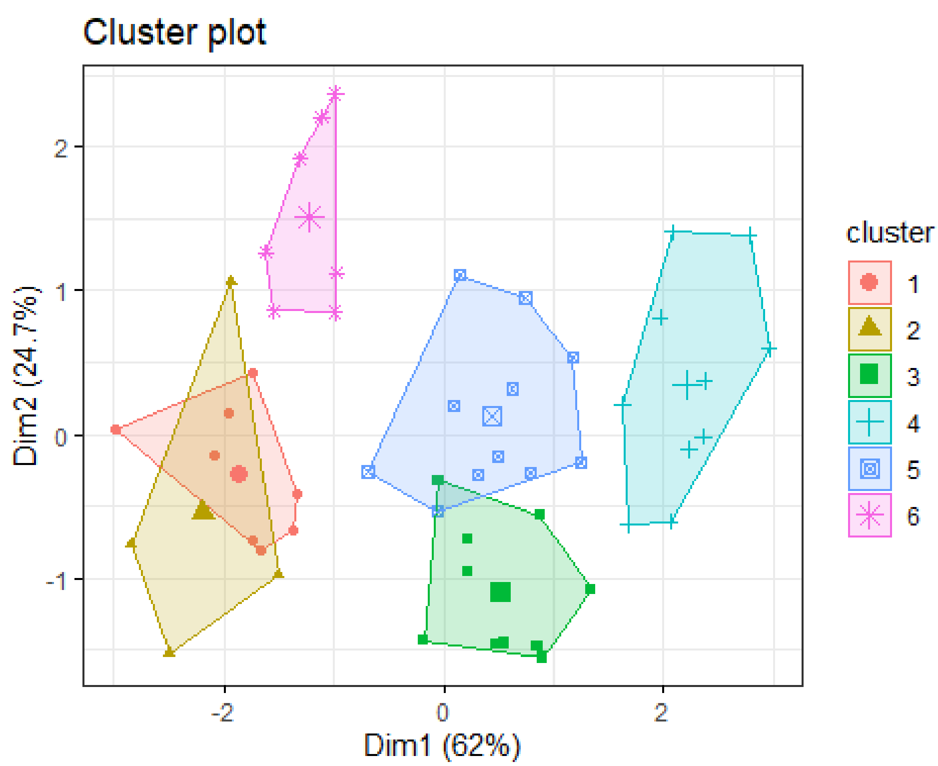





```
##          Alabama        Alaska       Arizona      Arkansas    California
##                6             2             1             5             2
##         Colorado   Connecticut      Delaware       Florida       Georgia
##                2             3             3             1             6
```

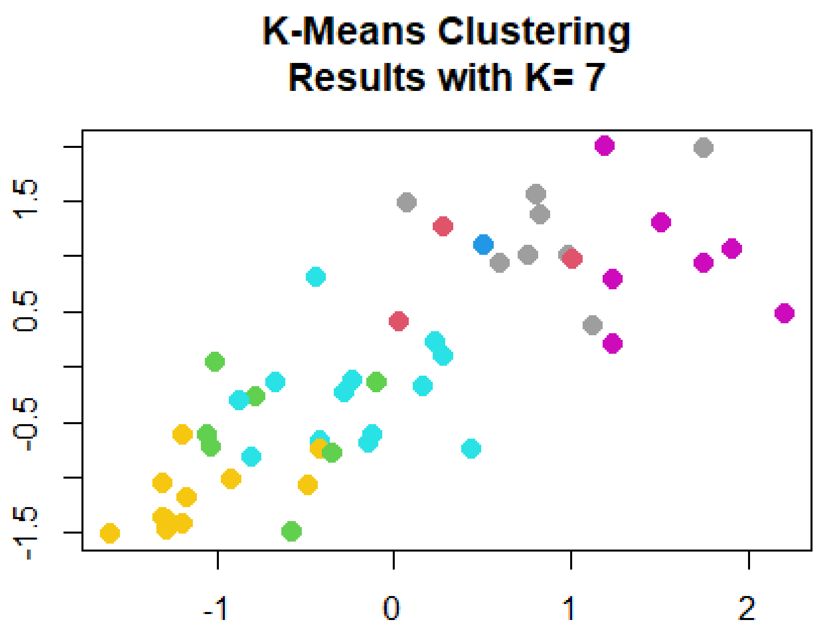
```
##          Hawaii          Idaho       Illinois        Indiana           Iowa 
##               3              4              1              5              4 
##          Kansas       Kentucky      Louisiana          Maine       Maryland 
##               5              5              6              4              1 
##   Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##               3              1              4              6              5 
##         Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##               5              5              2              4              3 
##      New Mexico       New York North Carolina   North Dakota           Ohio 
##               1              1              6              4              3 
##        Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##               5              5              3              3              6 
##    South Dakota      Tennessee          Texas           Utah        Vermont 
##               4              6              1              3              4 
##        Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##               5              3              4              4              5 
```

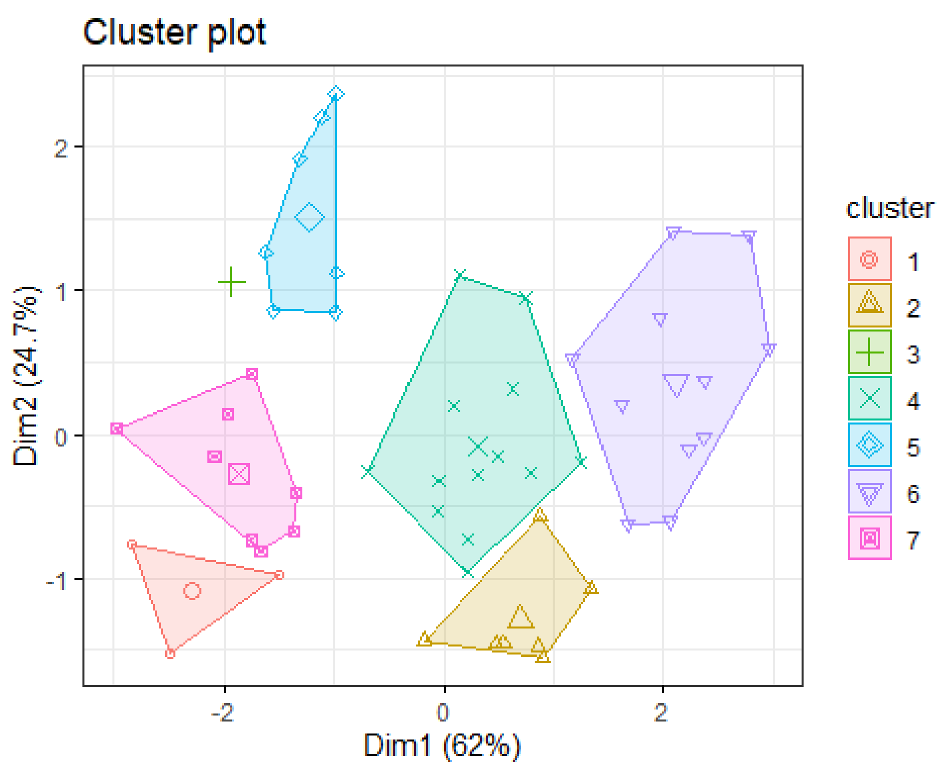





```
##         Alabama        Alaska       Arizona      Arkansas    California
##               5             3             7             4             1
##        Colorado   Connecticut      Delaware       Florida       Georgia
##               1             2             4             7             5
```

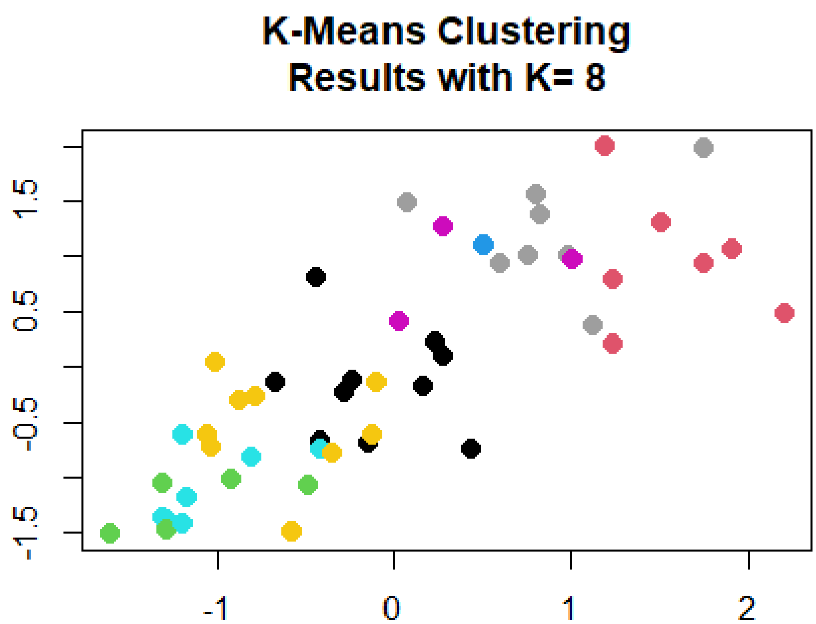
```
##         Hawaii          Idaho       Illinois        Indiana           Iowa 
##              2              6              7              4              6 
##         Kansas       Kentucky      Louisiana          Maine       Maryland 
##              4              4              5              6              7 
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##              2              7              6              5              4 
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##              6              4              1              6              2 
##     New Mexico       New York North Carolina   North Dakota           Ohio 
##              7              7              5              6              4 
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##              4              4              2              2              5 
##   South Dakota      Tennessee          Texas           Utah        Vermont 
##              6              5              7              2              6 
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##              4              4              6              6              4
```

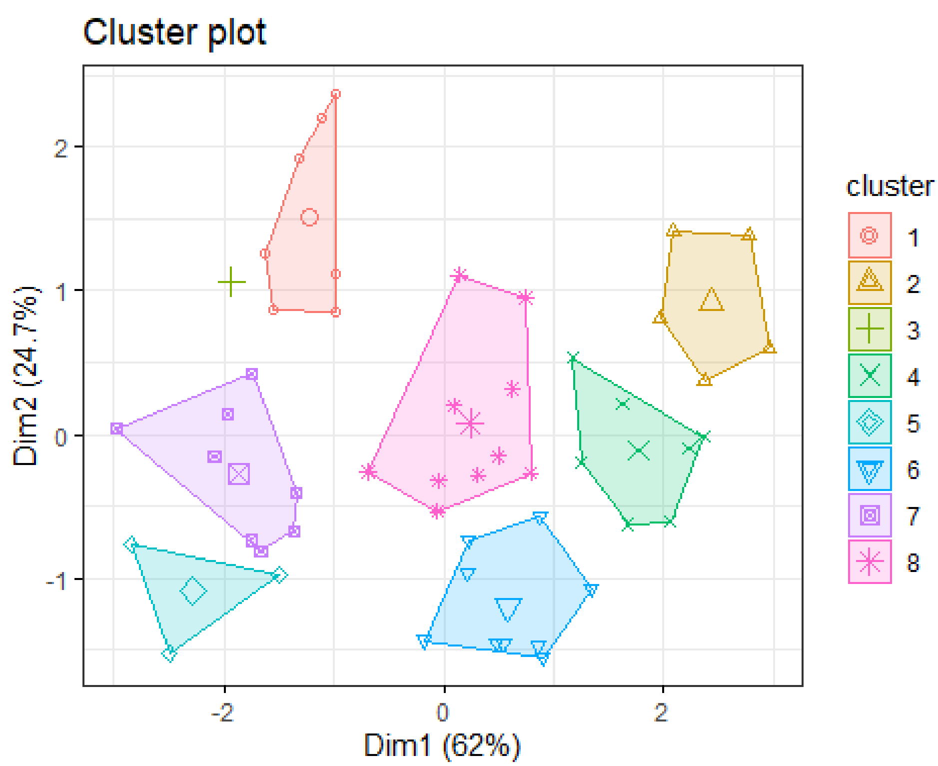





```
##       Alabama      Alaska     Arizona    Arkansas  California
##             1           3           7           8           5
##      Colorado Connecticut    Delaware     Florida     Georgia
##             5           6           8           7           1
```

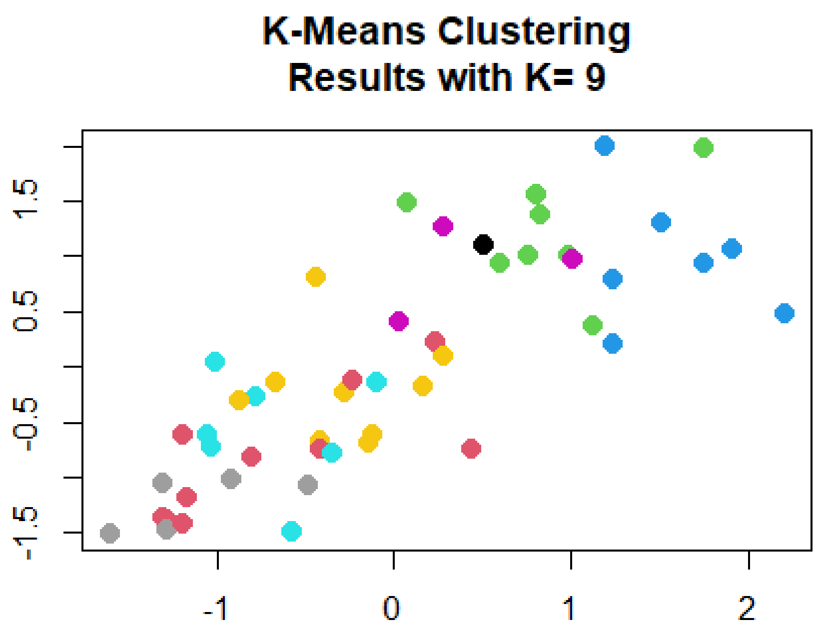
```
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              6              4              7              8              4
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              8              8              1              2              7
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              6              7              4              1              8
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              4              4              5              4              6
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              7              7              1              2              6
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              8              8              6              6              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              7              6              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              8              6              2              4              8
```

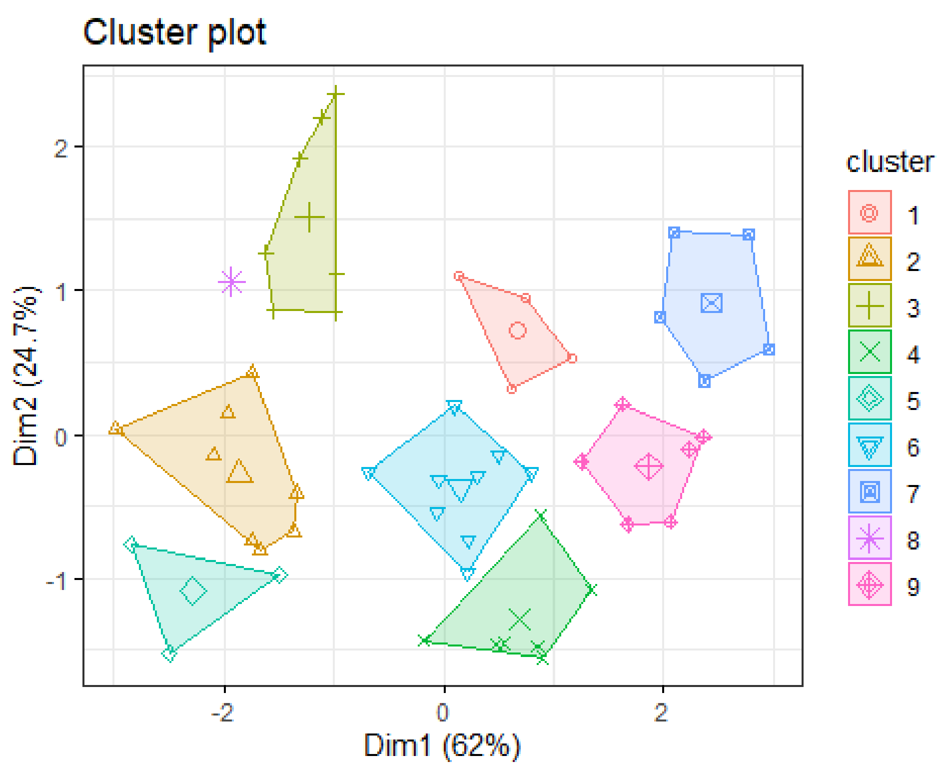





```
##          Alabama          Alaska         Arizona        Arkansas      California
##                3               8               2               1               5
##         Colorado     Connecticut        Delaware         Florida         Georgia
##                5               4               6               2               3
```

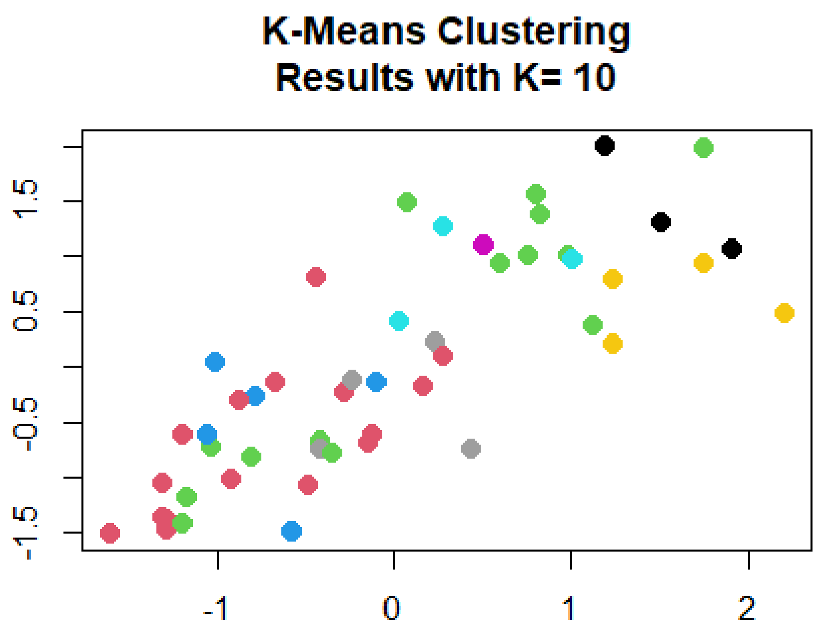
```
##         Hawaii          Idaho       Illinois        Indiana           Iowa 
##              4              9              2              6              9 
##         Kansas       Kentucky      Louisiana          Maine       Maryland 
##              6              1              3              7              2 
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##              4              2              9              3              6 
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##              1              9              5              9              4 
##     New Mexico       New York North Carolina   North Dakota           Ohio 
##              2              2              3              7              6 
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##              6              6              4              4              3 
##   South Dakota      Tennessee          Texas           Utah        Vermont 
##              7              3              2              4              7 
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##              6              6              7              9              1
```

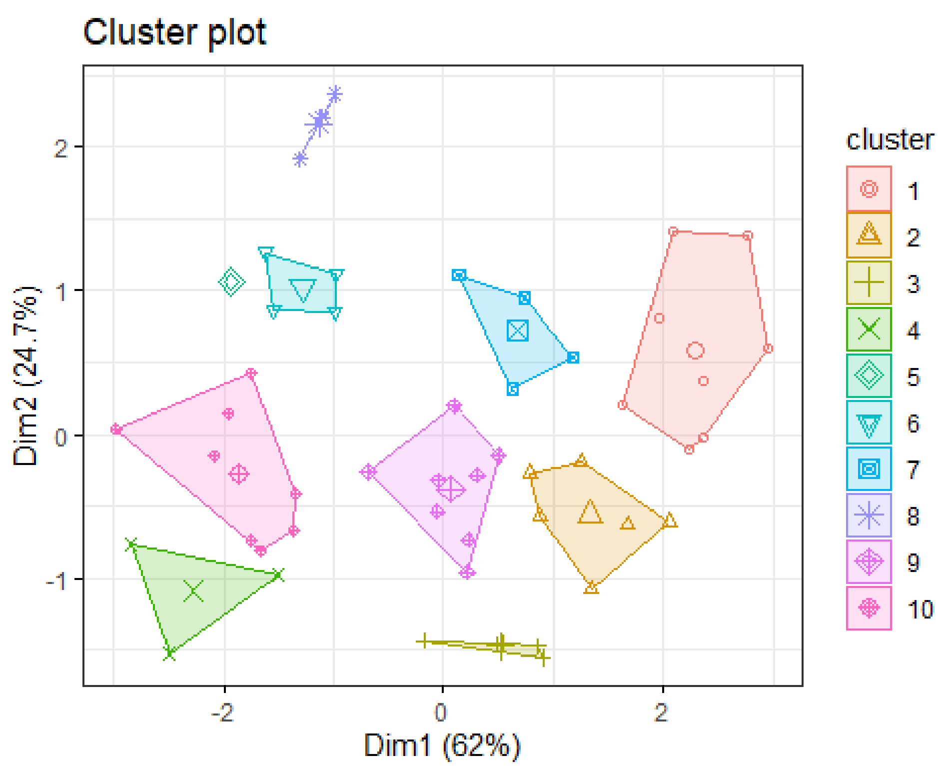





```
##        Alabama       Alaska     Arizona    Arkansas   California
##              6            5          10           7            4
##       Colorado  Connecticut    Delaware     Florida      Georgia
##              4            2           9          10            6
```

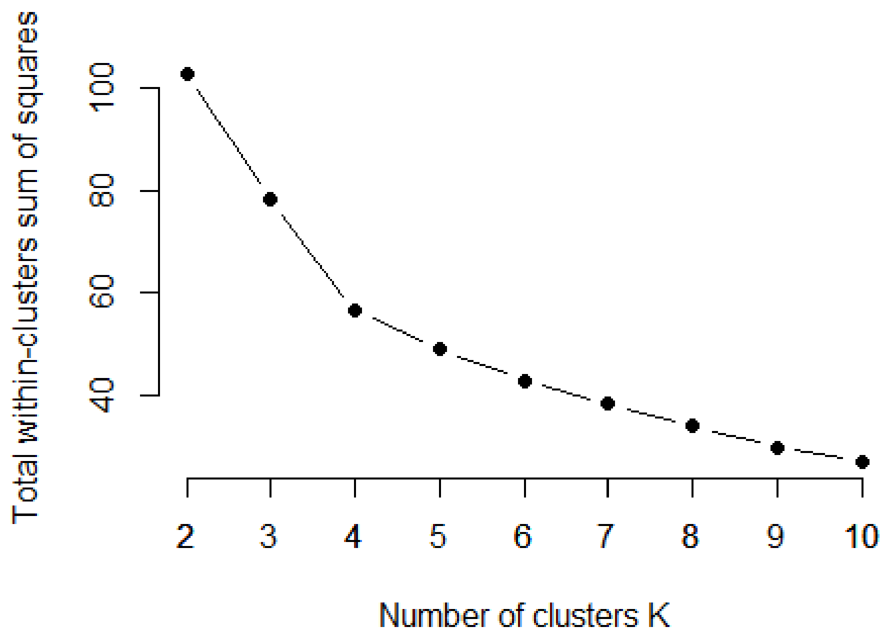
```
##         Hawaii          Idaho       Illinois        Indiana           Iowa 
##              3              1             10              9              1 
##         Kansas       Kentucky      Louisiana          Maine       Maryland 
##              2              7              6              1             10 
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##              3             10              2              8              9 
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##              7              2              4              1              3 
##     New Mexico       New York North Carolina   North Dakota           Ohio 
##             10             10              8              1              9 
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##              9              9              2              3              8 
##   South Dakota      Tennessee          Texas           Utah        Vermont 
##              1              6             10              3              1 
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##              9              9              1              2              7 
```






```
# Create an elbow plot to find the optimal K value
plot(dict_ss$keys(), dict_ss$values(),
       type="b", pch = 19, frame = FALSE,xlim = c(2,10),
       xlab="Number of clusters K",
```

```
       ylab="Total within-clusters sum of squares")
axis(side = 1, at = c(2:10))
```



From the above elbow plot we can see that k=4 is the optimal value of k.

2.3 Problem 3

```
library(factoextra)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

#Loading dataset

white_w=read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databas
es/wine-quality/winequality-white.csv"),header = T,sep=";")
```

```r
summary(white_w)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar  
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600  
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700  
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200  
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391  
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900  
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density      
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871  
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917  
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937  
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940  
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961  
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390  
##        pH          sulphates         alcohol         quality     
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000  
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000  
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000  
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878  
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000  
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000  
```

```r
#Checking the mean and variance among the available features
print(apply(white_w,2,mean))
```

```
##       fixed.acidity     volatile.acidity          citric.acid 
##          6.85478767           0.27824112           0.33419151 
##       residual.sugar            chlorides  free.sulfur.dioxide 
##          6.39141486           0.04577236          35.30808493 
## total.sulfur.dioxide              density                   pH 
##        138.36065741           0.99402738           3.18826664 
##            sulphates              alcohol              quality 
##          0.48984688          10.51426705           5.87790935 
```

```r
print(apply(white_w,2,var))
```

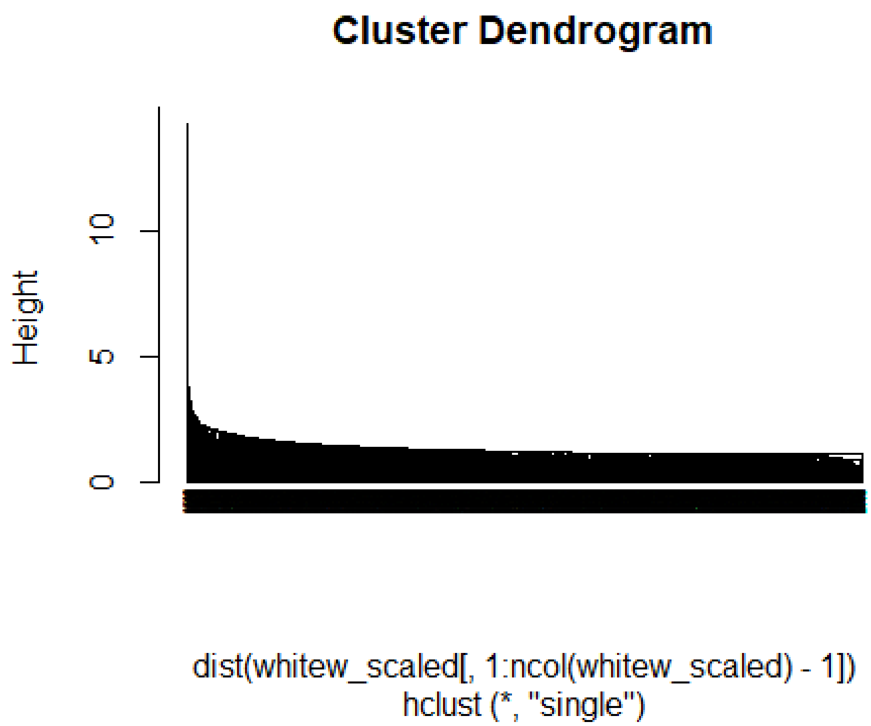
```
##       fixed.acidity     volatile.acidity          citric.acid 
##      0.712113585700       0.010159540992       0.014645793009 
##       residual.sugar            chlorides  free.sulfur.dioxide 
##     25.725770164386       0.000477333710     289.242719999320 
## total.sulfur.dioxide              density                   pH 
##   1806.085490848098       0.000008945524       0.022801181084 
##            sulphates              alcohol              quality 
##      0.013024705975       1.514426981787       0.784355685471 
```

The free sulfur dioxide and total sulfur dioxide features exhibit the greatest differences in means. Moreover, the variances of these features, as well as those of residual sugar and other features, are significantly larger than those of the remaining features. Therefore, it would be beneficial to scale these features to unit normal space, in order to prevent their large magnitudes from disproportionately affecting the algorithm.

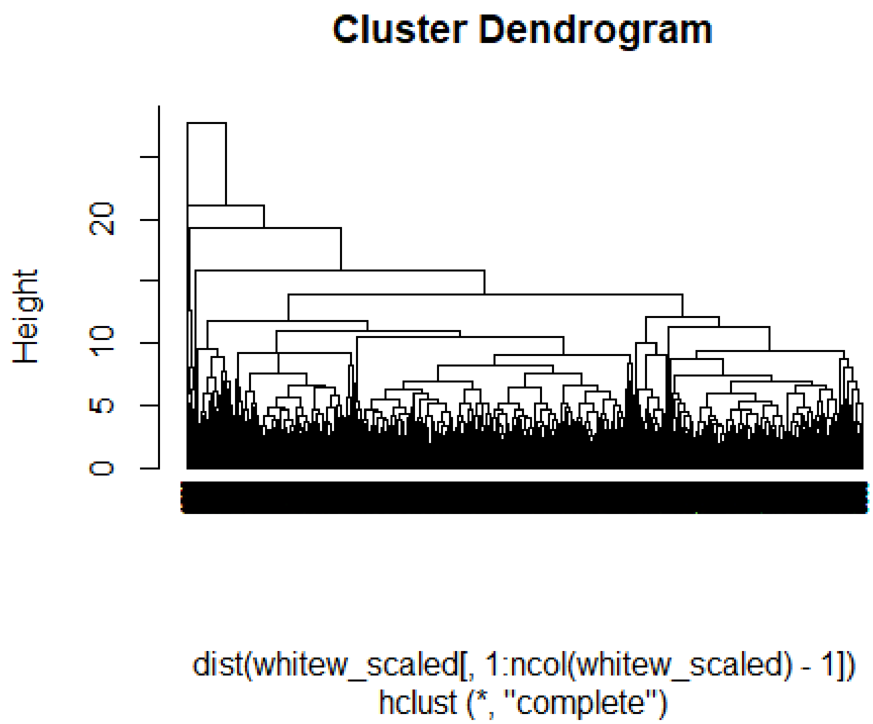
```
library(dplyr)
#Scaling the dataset
whitew_scaled=scale(white_w,center = TRUE,scale = TRUE)
#white_wine_scaled=white_wine
colnames(whitew_scaled)
```

```
##  [1] "fixed.acidity"        "volatile.acidity"     "citric.acid"
##  [4] "residual.sugar"       "chlorides"            "free.sulfur.dioxide"
##  [7] "total.sulfur.dioxide" "density"              "pH"
## [10] "sulphates"            "alcohol"              "quality"
```

```
hclust_1=hclust(dist(whitew_scaled[ , 1:ncol(whitew_scaled)-1]),method = "sin
gle")
plot(hclust_1,cex = 0.3, hang = -1)
```




```
hclust_complete=hclust(dist(whitew_scaled[ , 1:ncol(whitew_scaled)-1]),method
= "complete")
plot(hclust_complete, cex = 0.5, hang = -1)
```




```
sub_1=cutree(hclust_1,k=2)
sub_complete=cutree(hclust_complete,k=2)

sub_2=cutree(hclust_1,k=3)
sub_complete2=cutree(hclust_complete,k=3)

print(table(sub_1))

## sub_1
##    1    2
## 4897    1

print(table(sub_1,white_w$quality))

##
## sub_1    3    4    5    6    7    8    9
##     1   20  163 1457 2197  880  175    5
##     2    0    0    0    1    0    0    0

print(table(sub_complete))

## sub_complete
##    1    2
## 4897    1

print(table(sub_complete,white_w$quality))
```

```
##
## sub_complete    3    4    5    6    7    8    9
##            1   20  163 1457 2197  880  175    5
##            2    0    0    0    1    0    0    0
```

```
print(table(sub_2))
```

```
## sub_2
##    1    2    3
## 4896    1    1
```

```
print(table(sub_2,white_w$quality))
```

```
##
## sub_2    3    4    5    6    7    8    9
##     1   19  163 1457 2197  880  175    5
##     2    0    0    0    1    0    0    0
##     3    1    0    0    0    0    0    0
```

```
print(table(sub_complete2))
```

```
## sub_complete2
##    1    2    3
## 4896    1    1
```

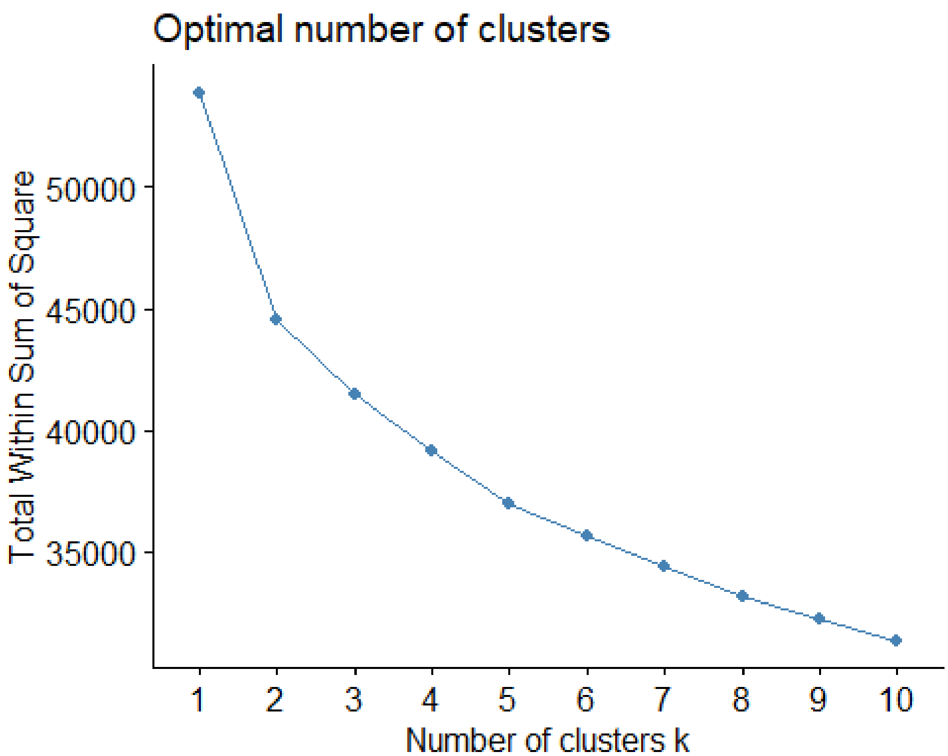
```
print(table(sub_complete2,white_w$quality))
```

```
##
## sub_complete2    3    4    5    6    7    8    9
##             1   19  163 1457 2197  880  175    5
##             2    0    0    0    1    0    0    0
##             3    1    0    0    0    0    0    0
```

```
sub_single_height = cutree(hclust_1,k=3)
sub_complete_height = cutree(hclust_complete,k=3)
```

Elbow plot :

```
fviz_nbclust(whitew_scaled[,1:ncol(whitew_scaled)-1], FUN = hcut, method = "w
ss")
```



The complete linkage method produces more balanced clustering from the dendrograms above