KENTUCKY STATE UNIVERSITY, COMPUTER SCIENCE

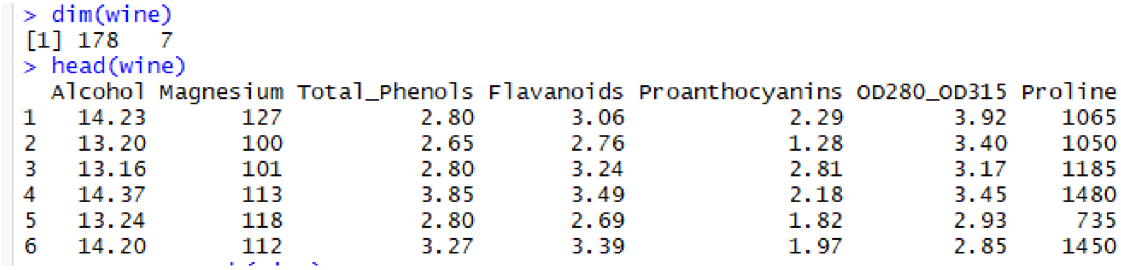
Data Mining
HW8

Anil Varma B

November 29, 2019

1. Use attached Wine_sub data to find 4 clusters by using R. You need to report SSE within cluster and between clusters. You also need to report the ratio of betweenSSE over total SSE. Discus each cluster's behavior in terms of its density.

```
library(cluster)
library(fpc)
wine <- read.csv(file.choose(),head=TRUE)
dim(wine)
head(wine)
```

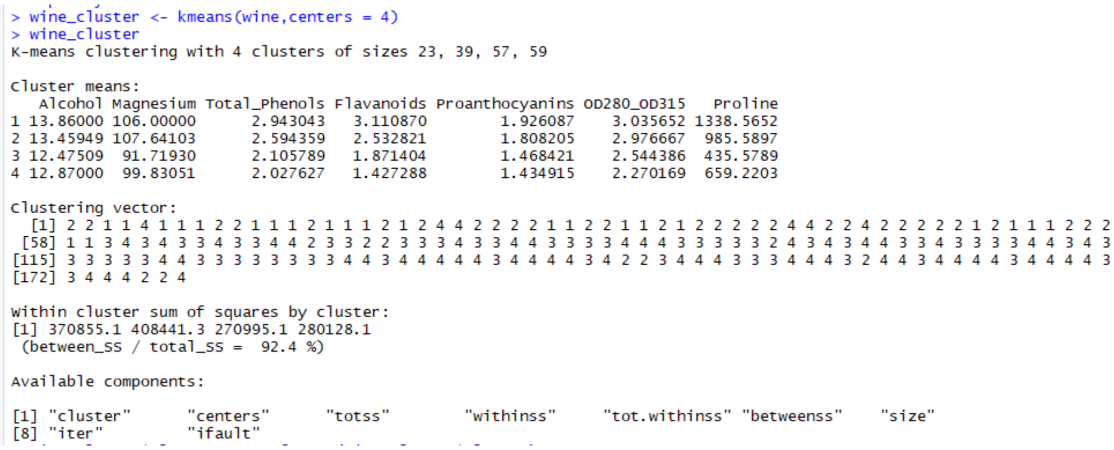


```
> dim(wine)
[1] 178   7
> head(wine)
  Alcohol Magnesium Total_Phenols Flavanoids Proanthocyanins OD280_OD315 Proline
1   14.23       127          2.80       3.06            2.29        3.92    1065
2   13.20       100          2.65       2.76            1.28        3.40    1050
3   13.16       101          2.80       3.24            2.81        3.17    1185
4   14.37       113          3.85       3.49            2.18        3.45    1480
5   13.24       118          2.80       2.69            1.82        2.93     735
6   14.20       112          3.27       3.39            1.97        2.85    1450
```

Figure 0.1: Dimensions of the data set

For finding 4 clusters, we give the option for 4 centers in the kmeans method and plot the four clusters.

```
wine_cluster <- kmeans(wine,centers = 4)
wine_cluster
wine_cluster$cluster <- as.factor(wine_cluster$cluster)
#Plot the data using clusters
plotcluster(wine, wine_cluster$cluster)
```



```
> wine_cluster <- kmeans(wine,centers = 4)
> wine_cluster
K-means clustering with 4 clusters of sizes 23, 39, 57, 59

Cluster means:
   Alcohol Magnesium Total_Phenols Flavanoids Proanthocyanins OD280_OD315   Proline
1 13.86000 106.00000      2.943043   3.110870        1.926087    3.035652 1338.5652
2 13.45949 107.64103      2.594359   2.532821        1.808205    2.976667  985.5897
3 12.47509  91.71930      2.105789   1.871404        1.468421    2.544386  435.5789
4 12.87000  99.83051      2.027627   1.427288        1.434915    2.270169  659.2203

Clustering vector:
  [1] 2 2 1 1 4 1 1 1 2 2 1 1 1 2 1 1 1 2 1 2 4 4 2 2 2 2 1 1 2 2 1 1 2 1 2 2 2 2 2 4 4 2 2 4 2 2 2 2 2 1 2 1 1 1 2 2 2
 [58] 1 1 3 4 3 4 3 3 4 3 3 4 4 2 3 3 2 2 3 3 3 4 3 3 4 4 3 3 3 3 4 4 4 3 3 3 3 3 2 4 3 4 3 4 4 3 3 4 3 3 3 3 4 4 3 4 3
[115] 3 3 3 3 3 4 4 3 3 3 3 3 3 3 3 4 4 3 4 4 4 4 4 3 4 4 4 4 3 4 2 2 3 4 4 4 3 3 3 4 4 4 3 2 4 4 3 4 4 4 4 3 4 4 4 4 3
[172] 3 4 4 4 2 2 4

Within cluster sum of squares by cluster:
[1] 370855.1 408441.3 270995.1 280128.1
 (between_SS / total_SS =  92.4 %)

Available components:

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

Figure 0.2: Kmeans summary

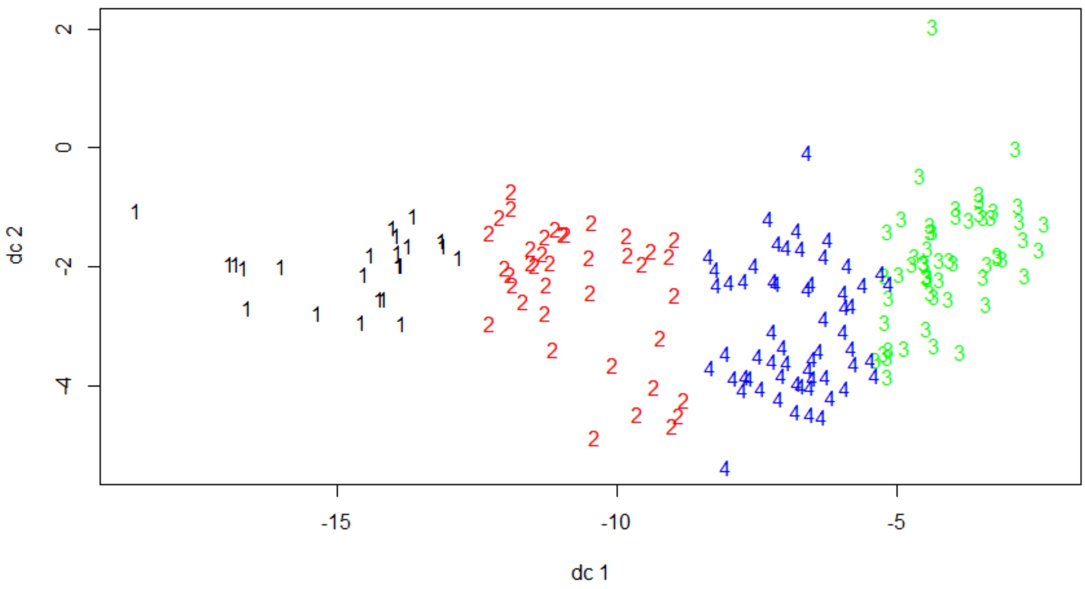
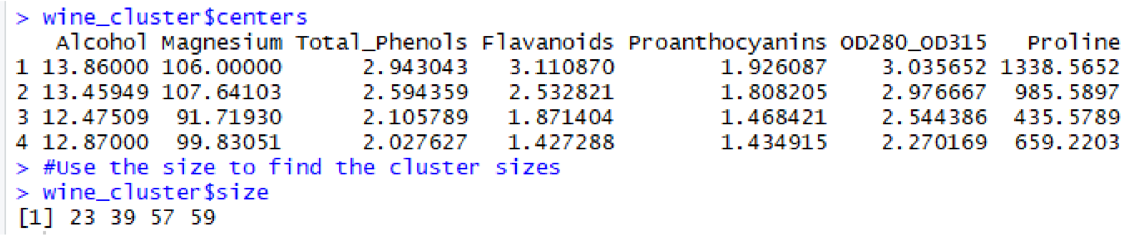


Figure 0.3: Plot with 4 clusters

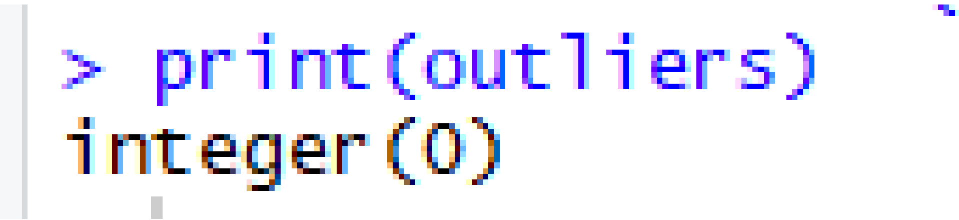
```
#Use the centers to find the cluster centers
wine_cluster$centers
#Use the size to find the cluster sizes, The number of points in each cluster
wine_cluster$size
```



```
> wine_cluster$centers
   Alcohol Magnesium Total_Phenols Flavanoids Proanthocyanins OD280_OD315   Proline
1 13.86000 106.00000      2.943043   3.110870        1.926087    3.035652 1338.5652
2 13.45949 107.64103      2.594359   2.532821        1.808205    2.976667  985.5897
3 12.47509  91.71930      2.105789   1.871404        1.468421    2.544386  435.5789
4 12.87000  99.83051      2.027627   1.427288        1.434915    2.270169  659.2203
> #Use the size to find the cluster sizes
> wine_cluster$size
[1] 23 39 57 59
```

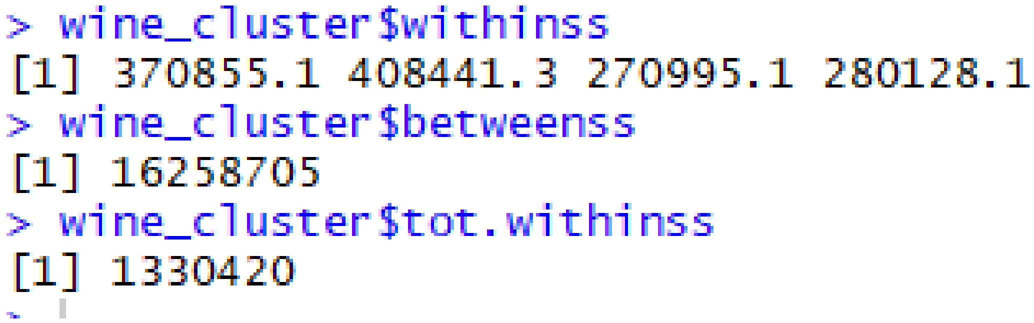
Figure 0.4: Centers and cluster size

```
#Outlier detection
distances <- sqrt(rowSums((wine - wine$centers)^2))
outliers <- order(distances, decreasing=T)
print(outliers)
```



```
> print(outliers)
integer(0)
```

Figure 0.5: Outlier detection



```
> wine_cluster$withinss
[1] 370855.1 408441.3 270995.1 280128.1
> wine_cluster$betweenss
[1] 16258705
> wine_cluster$tot.withinss
[1] 1330420
```

Figure 0.6: Other parameters of cluster

Below are the observations and reports from the clusters. SSE within cluster for each cluster are

```
Cluster 1 = 370855.1
Cluster 2 = 408441.3
Cluster 3 = 270995.1
cluster 4  = 280128.1
Total within for all clusters = 1330420
and between clusters = 16258705.
```

```
The ratio of betweenSSE over total SSE =  (between_SS / total_SS) =  92.4 %
```

From the cluster size, we can see that the fourth cluster is the most dense with the maximum size followed by cluster three which has the second highest size. The first cluster is of the least size and least dense.