# DSA 5013-001 Homework 8

Group 19: Zachary White, Azadeh Gilanpour, Precious Jatau

December 09, 2018

These are the libraries used:

```
library(cluster) # silhouette model
library(rgl) # silhouette model
library(useful)
library(dplyr)
library(dbscan) #dencity based clustering
library(factoextra) #fviz_cluster
library(fpc) #Compute DBSCAN
```

## Problem 1. Cluster Analysis

### Problem 1a

*Identify a data set for cluster analysis. The data set must have a minimum of 8 numeric variables and 500 observations.*

We are using a red wine classification set from kaggle. (https://www.kaggle.com/piyushgoyal443/red-wine-dataset/) It contains 1599 observations and 11 numeric variables. We show that there is no missing data, remove X and the variable quality, then scale the data.

```
wine=  read.csv("C:\\Users\\zackw\\Documents\\Classes\\Intelligent Data
Analytics\\HW\\HW8\\wineQualityReds.csv")

#finding missigness
propMissing = function (x) mean(is.na(x))
missing = apply(wine,2, propMissing)
missing

##                    X        fixed.acidity     volatile.acidity 
##                    0                    0                    0 
##          citric.acid       residual.sugar            chlorides 
##                    0                    0                    0 
##  free.sulfur.dioxide total.sulfur.dioxide              density 
##                    0                    0                    0 
##                   pH            sulphates              alcohol 
##                    0                    0                    0 
##              quality 
##                    0

# delete variable x "it is row number"
wine$X = NULL
z= wine[,-c(12,12)]

#Scale the data
wineScaled = data.frame(scale(z))
```

## Problem 1b

*Briefly describe the data set.*

The data contains information on red wine. It contains 1599 observations and 11 numeric variables. There is a 12th variable in the dataset named *quality* that ranges from 0-10. Ideally, the cluster data will be grouped together similar quality. The information about the 11 variables is given from the kaggle website.

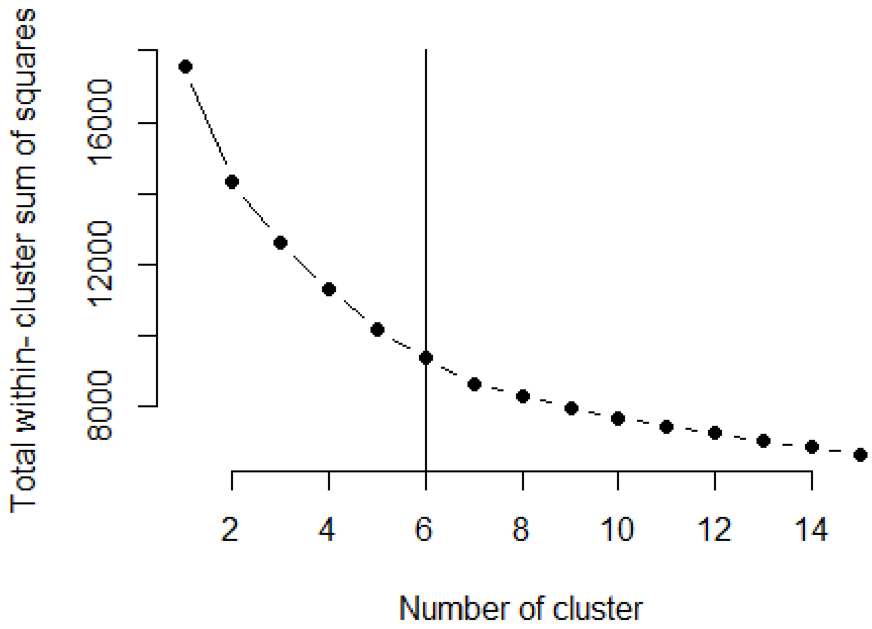
```
1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not
evaporate readily)
2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels
can lead to an unpleasant, vinegar taste
3 - citric acid: found in small quantities, citric acid can add 'freshness' and
flavor to wines
4 - residual sugar: the amount of sugar remaining after fermentation stops, it's rare
to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter
are considered sweet
5 - chlorides: the amount of salt in the wine
6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular
SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the
oxidation of wine
7 - total sulfur dioxide: amount of free and bound forms of S02; in low
concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations
over 50 ppm, SO2 becomes evident in the nose and taste of wine
8 - density: the density of water is close to that of water depending on the percent
alcohol and sugar content
9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14
(very basic); most wines are between 3-4 on the pH scale
10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02)
levels, wich acts as an antimicrobial and antioxidant
11 - alcohol: the percent alcohol content of the wine
```

## Problem 1c

*Perform a clustering analysis of this data set using:*

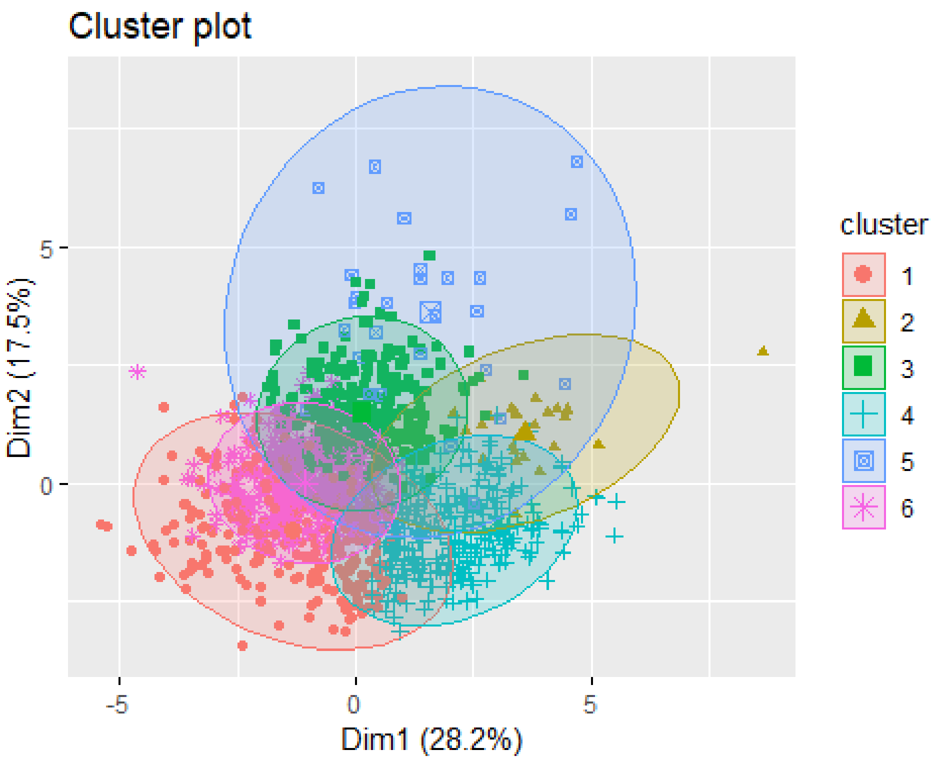
- *Partitional Clustering (use statistical and/or rational reasoning to determine k)*

```
#kmeans
# elbow method
set.seed(7)
k.max = 15
data = wineScaled
wass = sapply
(1:k.max,function(k){kmeans(data,k,nstart=50,iter.max=15)$tot.withinss})
plot(1:k.max,wass,
     type= "b",pch =19,frame= F,
     xlab= "Number of cluster",
     ylab= "Total within- cluster sum of squares")
abline(v=6)
```



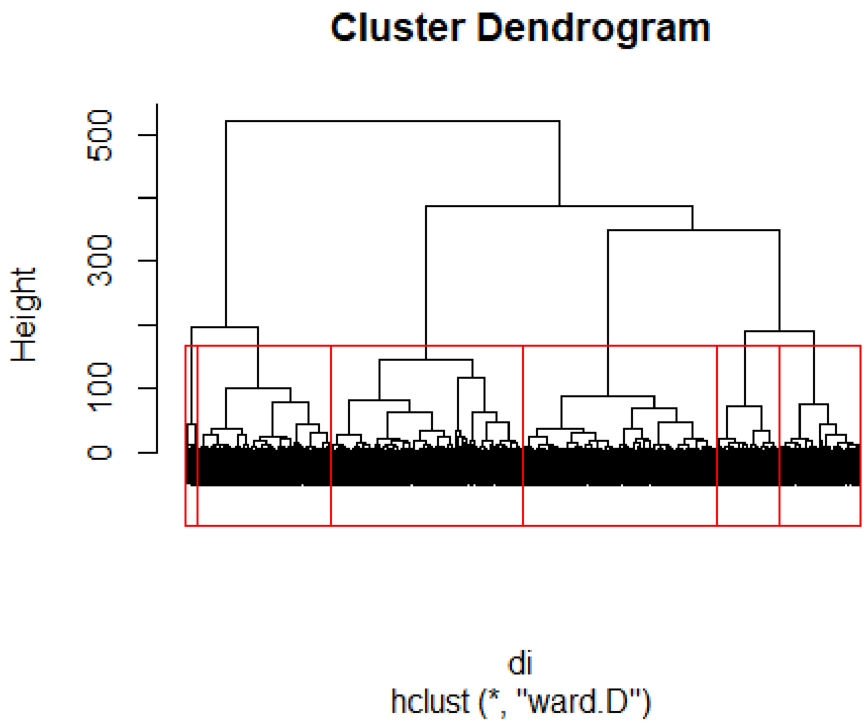
By using elbow method, we determine through the chart that k = 6 is the best number clusters

```
# chossing k=6 as clustring
winekm1= kmeans(wineScaled,6,nstart = 20)

# visualization of k-,means clustering
fviz_cluster(winekm1, data = wineScaled, geom = "point",
             stand = FALSE, frame.type = "norm")
```


Cluster plot
Dim2 (17.5%)
Dim1 (28.2%)
5
0
-5
0
5
cluster
1
2
3
4
5
6

- *Hierarchical Clustering (use statistical and/or rational reasoning to determine cut height)*

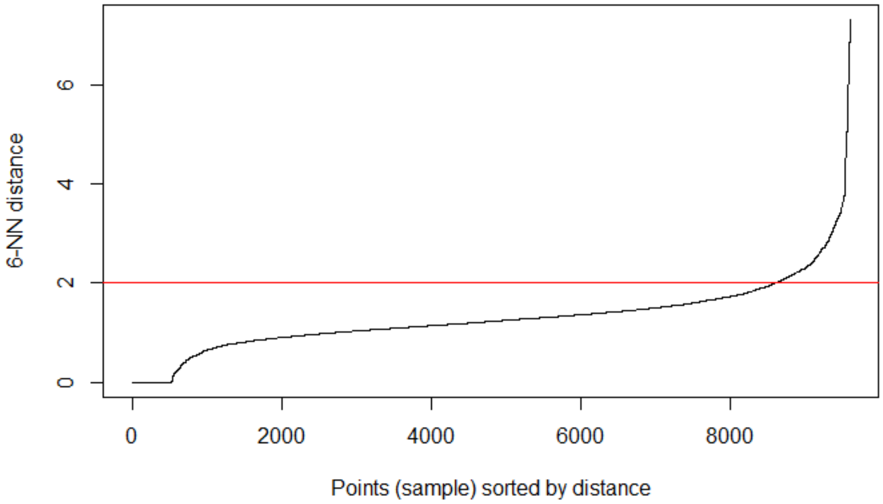
```
# with hiearchical clustering
di <- dist(wineScaled, method="euclidean")

hc <- hclust(di, method="ward.D")
plot(hc, labels=FALSE)

rect.hclust(hc, k=6, border="red")
```



By observing the dendrogram, we estimate that 6 is the optimal cut height. So hierarchical base Clustering says that there are 6 clusters.

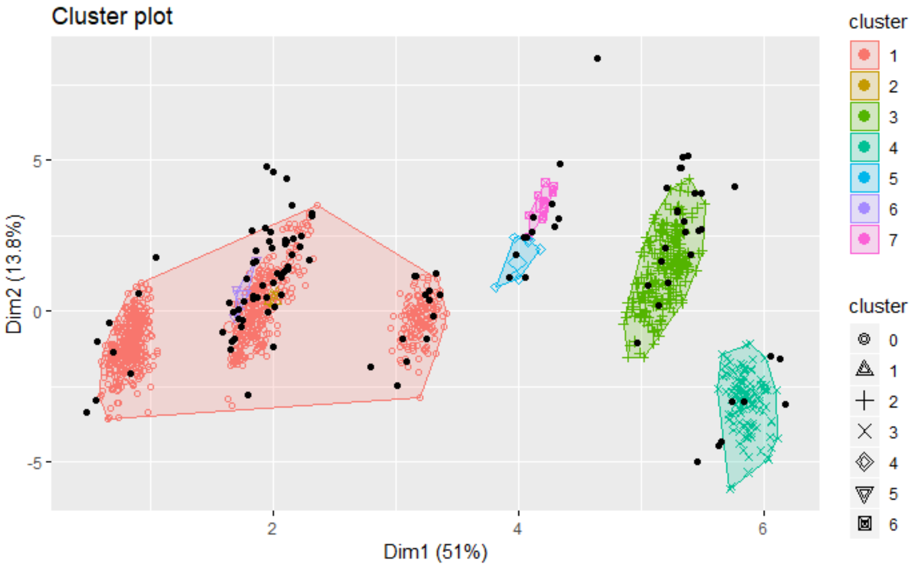
- *Density-Based Clustering (use statistical and/or rational reasoning to determine $\epsilon$ and min points)*

```
#3-dencity based clustring
winewmatrix= data.matrix(wineScaled)
kNNdistplot(winewmatrix, k=6)
abline(h=2, col="red")
```



By looking at the knee chart, we estimate $\epsilon = 2$. Through experiments, we estimate that minPt = 5.

```
res <- dbscan(winewmatrix, eps=2, minPts = 5)

# Compute DBSCAN using fpc package
set.seed(123)
db <- fpc::dbscan(winewmatrix, eps = 2, MinPts = 5)

fviz_cluster(db, winewmatrix, stand = FALSE, frame = FALSE, geom = "point")
```



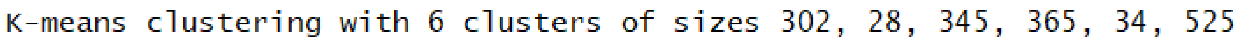
Density based clustering says there are 7 clusters.

## Problem 1d

*Provide a concise, statistical description of the final cluster results for each method.*

Kmeans:

6 clusters



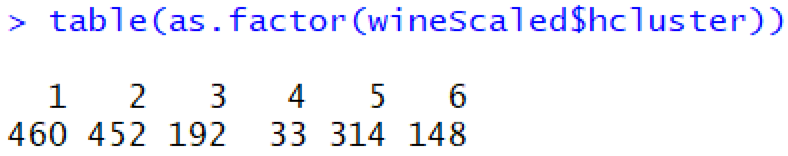
```
K-means clustering with 6 clusters of sizes 302, 28, 345, 365, 34, 525
```

```
winekm1$centers

##   fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides
## 1   -0.69270212     -0.437750230 -0.14745270   -0.257678973 -0.416913732
## 2    0.09538646      0.002199115  1.18118314   -0.389750233  5.782950580
## 3   -0.06888024      0.056768744  0.07177018   -0.009945823 -0.030112762
## 4    1.33443302     -0.671842755  1.12280156    0.073230184 -0.007875253
## 5   -0.08560643     -0.034641327  0.41472600    4.960215800  0.296295201
## 6   -0.48355911      0.683722264 -0.83281214   -0.196596170 -0.062524449
##   free.sulfur.dioxide total.sulfur.dioxide     density         pH
## 1          0.12449435           -0.2294327 -1.24608499  0.6334369
## 2         -0.04950011            0.5101700  0.18001552 -1.7352487
## 3          0.98015875            1.2141818  0.24707145 -0.1281191
## 4         -0.56740982           -0.5433378  0.76180825 -0.8453491
## 5          1.74964380            1.6953018  1.22461740 -0.3253578
## 6         -0.43190355           -0.4251634 -0.06411318  0.4211518
##     sulphates    alcohol
## 1  0.13496581  1.2951673
## 2  3.66226647 -0.8694593
## 3 -0.17786196 -0.5665520
## 4  0.34646284  0.1742501
## 5 -0.02378189 -0.3637992
## 6 -0.39541164 -0.4239378
```

Hierarchical:

6 clusters

Cluster size:

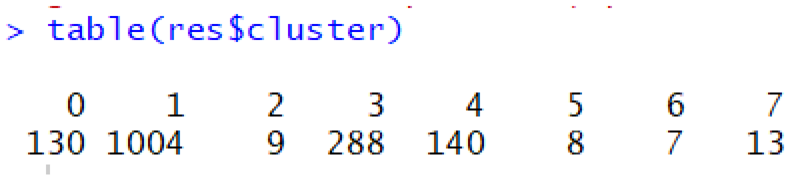


```
> table(as.factor(wineScaled$hcluster))

  1   2   3   4   5   6 
460 452 192  33 314 148 
```

Density:

7 clusters

Cluster Size:



```
> table(res$cluster)

   0    1    2    3    4    5    6    7 
 130 1004    9  288  140    8    7   13 
```

## Problem 1e

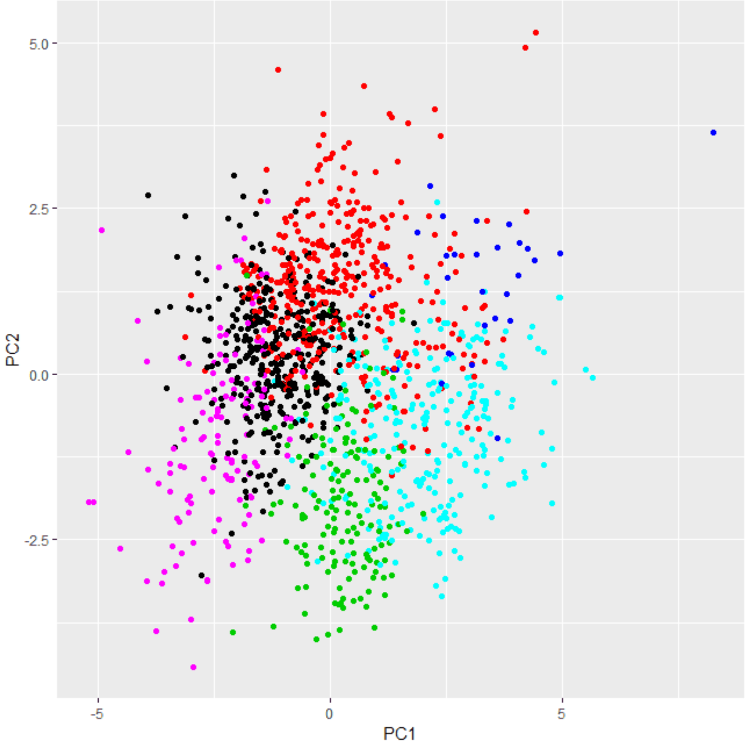
*Identify the cluster solution you think is best and provide a rationale for your choice.*

Hierarchical Clustering is the best solution. When looking at the three charts of the clusters: k-means clustering has significant overlap and density based clustering shows 6 clusters but reports 7 clusters and has two clusters within another cluster. Hierarchical clustering's dendrogram does not have any of these errors. Therefore we believe that hierarchical clustering to be the best method.

## Problem 1f

*Provide an interpretation of the clusters belonging to your preferred solution. This interpretation will require a subjective analysis of the clusters.*

```
pcWine = prcomp(wine, scale. = T)
biplot(pcWine)
biplot(pcWine)
newData = pcWine$x
newData = newData[,1:2]
ggplot(data = newData, aes(x = PC1, y = PC2))
newData = as.data.frame(newData)
ggplot(data = newData, aes(x = PC1, y = PC2)) + geom_point(col = wineScaled$hcluster)
```



We plotted the 1st two principle components and it shows 6 clusters.