**Assignment 3: Orthopedic Materials Sales (100 points)**

**Student Name: William Joseph Wadford**

**Purpose:** To perform cluster analysis to identify potential business for orthopedic material sales

**Description:** The objective of this study is to find ways to increase sales of orthopedic material from our company to hospitals in the United States. The data include information about over 4000 hospitals. Below is the data dictionary:

ZIP: US POSTAL CODE

HID: HOSPITAL ID

CITY: CITY NAME

STATE: STATE NAME

BEDS: NUMBER OF HOSPITAL BEDS

RBEDS: NUMBER OF REHAB BEDS

OUT-V: NUMBER OF OUTPATIENT VISITS

ADM: ADMINISTRATIVE COST (In $1000's per year)

SIR: REVENUE FROM INPATIENT

SALESY: SALES OF REHABILITATION EQUIPMENT SINCE JAN 1

SALES12: SALES OF REHAB. EQUIP. FOR THE LAST 12 MO

HIP: NUMBER OF HIP OPERATIONS FOR TWO YEARS AGO

KNEE: NUMBER OF KNEE OPERATIONS FOR TWO YEARS AGO

TH: TEACHING HOSPITAL? 0, 1

TRAUMA: DO THEY HAVE A TRAUMA UNIT? 0, 1

REHAB: DO THEY HAVE A REHAB UNIT? 0, 1

HIP12: NUMBER HIP OPERATIONS FOR THE LAST 12 MO

KNEE12: NUMBER KNEE OPERATIONS FOR THE LAST 12 MO

FEMUR12: NUMBER FEMUR OPERATIONS FOR THE LAST 12 MO

**Instructions:** Please follow these steps:

1. In Canvas, navigate to Assignments and then Assignment4
2. Download and save the data set hospital\_ortho.csv
3. Read the file:

data <- fread("hospital\_ortho.csv", sep=",", header=T, strip.white = T, na.strings = c("NA","NaN","","?"))

1. The original data includes hospitals across the US. However, we can only sell our products in NC and the nearby states of SC, VA, GA, and TN. Use the following code to narrow down the data to hospitals in these states.

nc\_data <- data[(data$state == "NC") | (data$state == "SC") | (data$state == "VA") | (data$state == "GA") | (data$state == "TN")]

* 1. **(3 points)** Look at each individual variable and decide if it should be included in cluster analysis. For those variables that you decide not to include, give your reasons for exclusion.

**I would exclude HID (Hospital ID), State, City Name, and Zip Code because I would want to find common bonds that do not have to do with location or the hospital ID. We would want to understand characteristics about operations and spending. I do not want to cluster just based on similar location.**

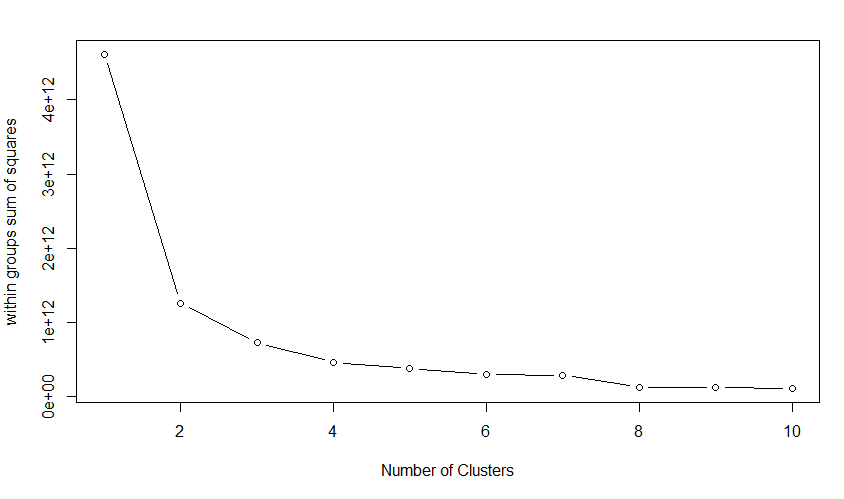
* 1. **(3 points)** Do you need to scale this data? Why?

**Yes, because there are differences in scales. Rbeds is as low as the tens and out-v is in the hundreds of thousands.**

1. Perform k-means clustering:
   1. **(3 points)** Use “Within Groups SSE” to determine the number of clusters for k-means. How many clusters you would like to create?

**After reviewing the plot, I would like to create 2 clusters. It appears as though the SSE would not significantly be lowered if we go from 2 to 3.**

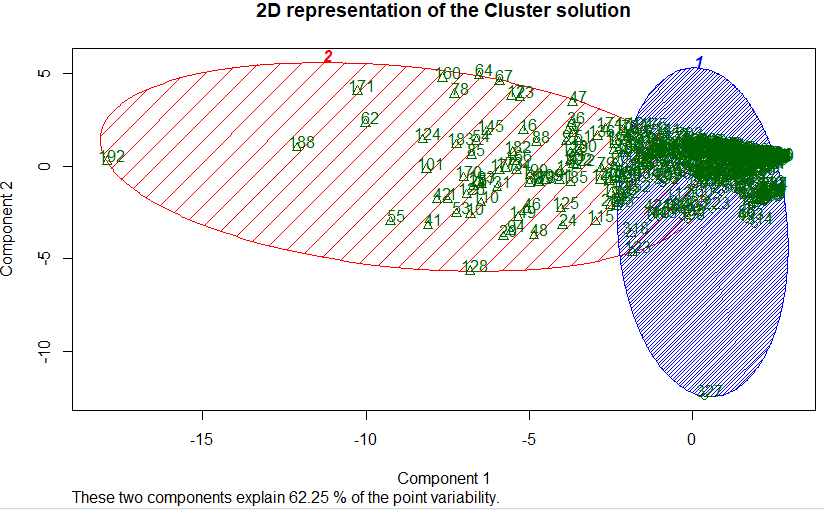
* 1. **(3 points)** Paste the “Within Groups SSE” plot in the space below:



* 1. **(3 points)** Perform k-means clustering using the number of clusters you recommended in 5.1. How many hospitals fall in each cluster?

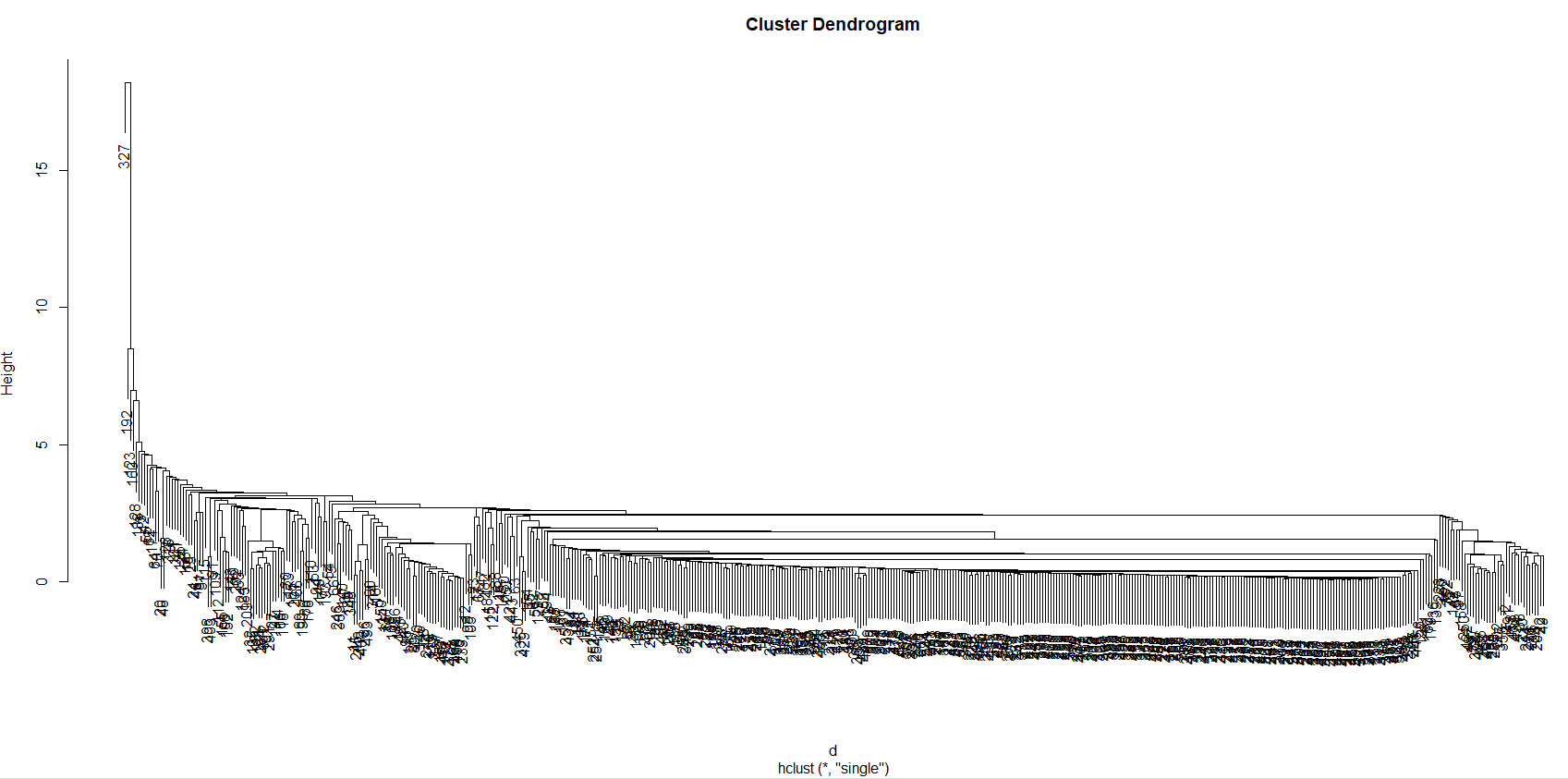
**418; 90**

* 1. **(3 points)** Create a two-dimensional representation of the clusters and paste it below:

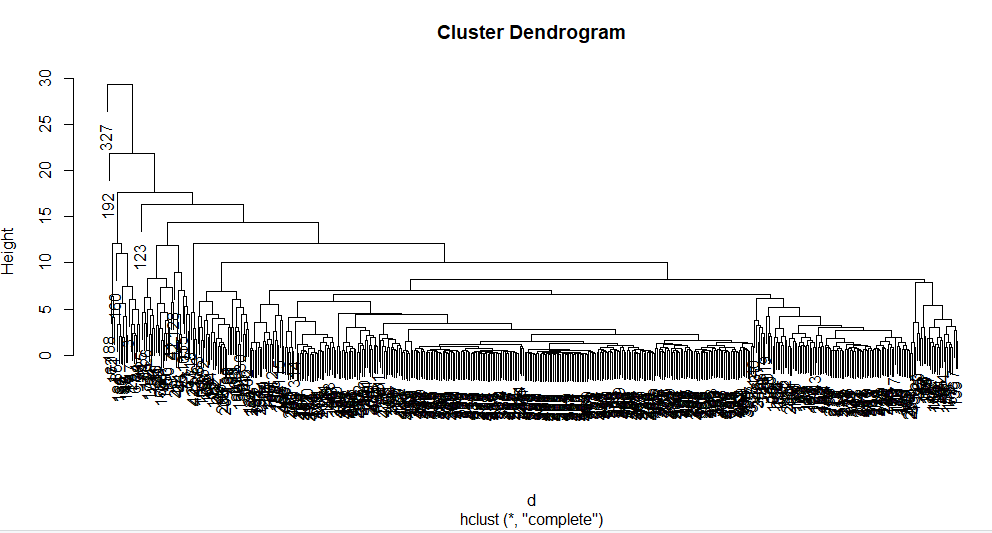


1. Perform Hierarchical clustering.
   1. **(4 points)** Try different hierarchical clustering and paste the dendrograms in the space below:

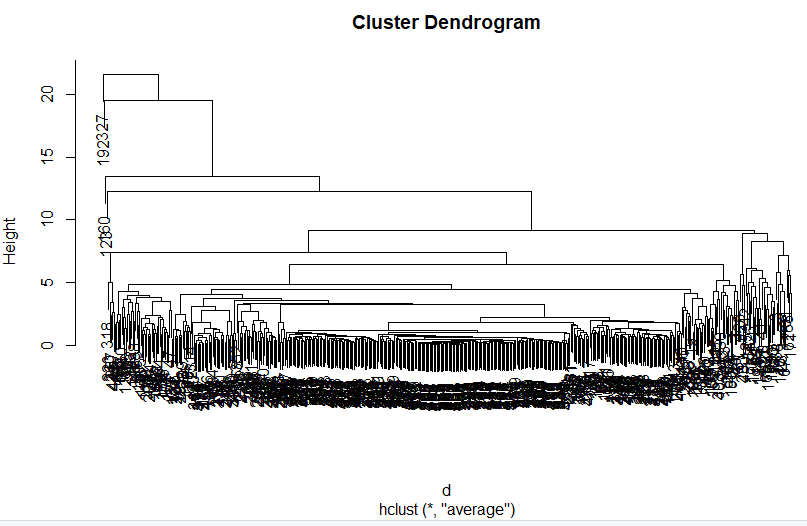
**H.single**



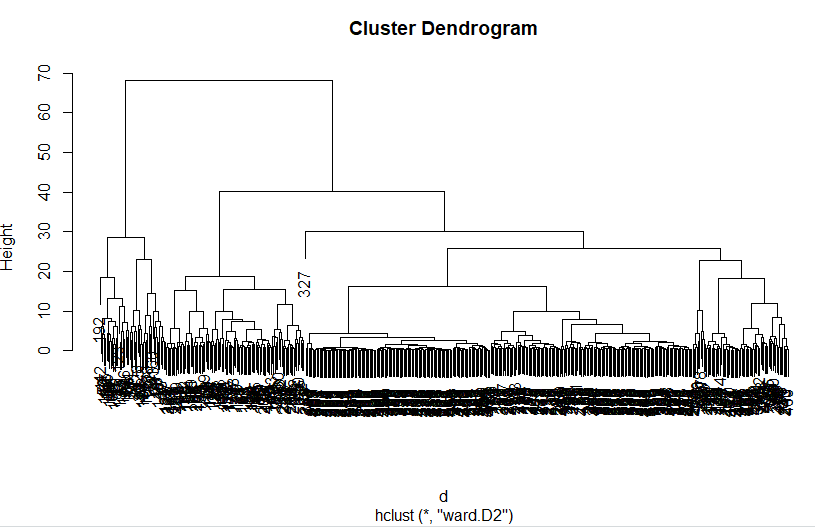
**H.complete**



**H.average**



**H.ward**

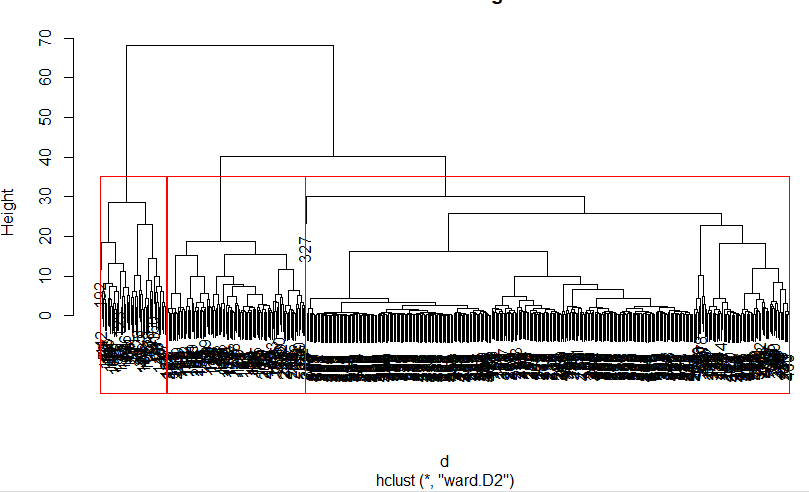


* 1. **(3 points)** Determine which hierarchical clustering method would be more appropriate for this data. Why?

**Based on appearances, it appears the ward method for hierarchical clustering is more appropriate for this data. It appears to be the best clustering method with distinct clusters.**

* 1. **(3 points)** Based on hierarchical clustering results, how many clusters do you find in this data?

**There are three clusters.**

* 1. **(3 points)** Paste the dendrogram that you chose with the red borders around the clusters in the space below:

1. Perform DBSCAN cluster analysis:
   1. **(7 points)** First, you need to determine minPts. The rule of thumb for minPts is the number of dimensions of the data + 1. Suggest a method to determine the number of dimensions of this data? Implement your method and suggest a good minPts.

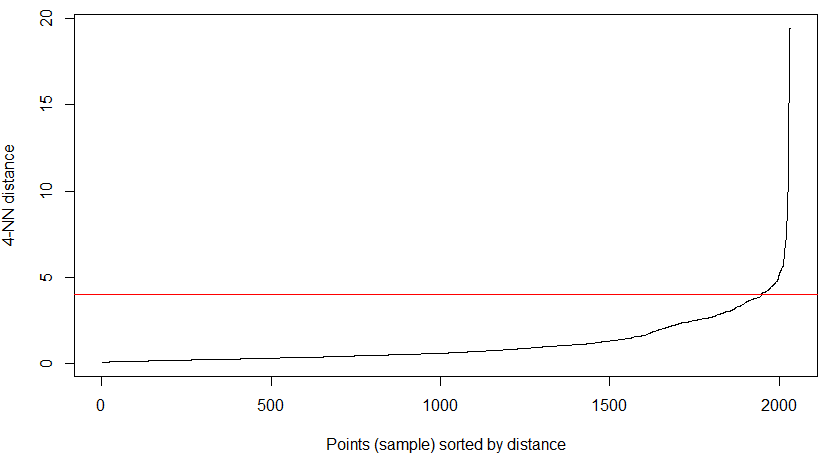
**Reviewing the independent (predictor) variables, I would suggest 4 dimensions as follows:**

* **Patient info**
  + **BEDS: Number of hospital beds**
  + **RBEDS: Number of rehab beds**
  + **OUT-V: Number of outpatient visits**
  + **HIP: number of hip operations from two years ago**
  + **KNEE: number of knee operations from two years ago**
  + **HIP12: number of hip operations in the past 12 months**
  + **KNEE12: number of knee operations in the past 12 months**
  + **FEMUR12: number of femur operations in the past 12 months**
* **Hospital Characteristics**
  + **TH: Is it a teaching hospital? Yes – 1; no - 0**
  + **TRAUMA: Do they have a trauma unit? 1, 0**
  + **REHAB: Do they have a rehab unit? 1,0**
* **Sales data**
  + **SIR: Revenue from inpatient**
  + **ADM: Administrative costs**

**The measure would be sales data, such as SALESY and SALES12. This would allow us to slice and dice patient and sales data by hospital characteristics.**

* 1. **(3 points)** Based on your suggested minPts, determine the eps. Explain your recommendation for eps.

**Based on my suggested minPts, I suggest 4 as the eps. 4 is the knee of the kNN plot.**



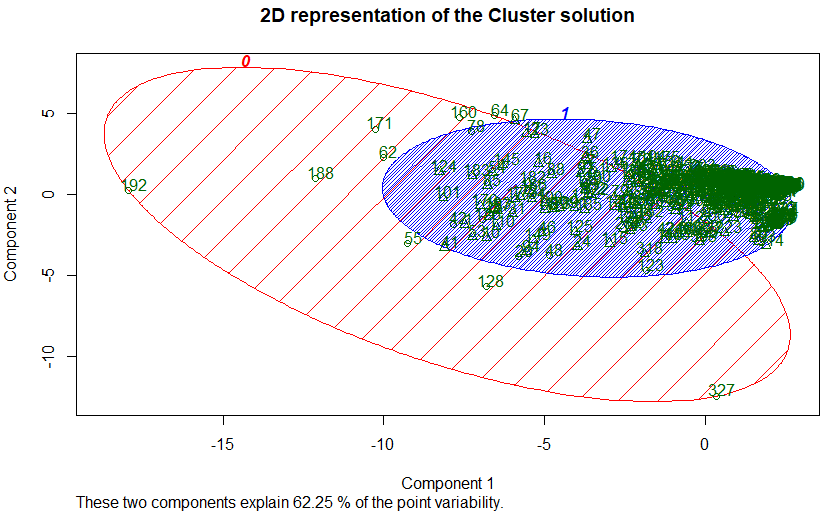
* 1. **(3 points)** Perform DBSCAN clustering using the minPts and eps that you recommended. How many clusters DBSCAN returns?

**1 cluster**

* 1. **(3 points)** How many noise points it returns?

**16 noise points**

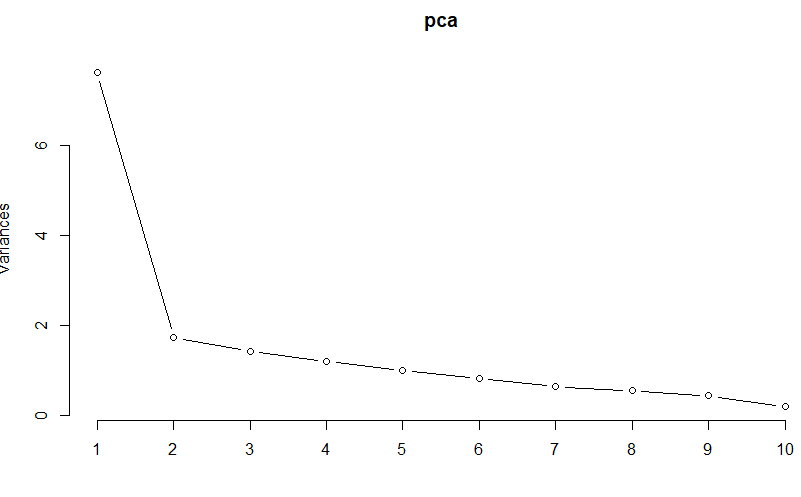
* 1. **(3 points)** Create a two-dimensional representation of DBSCAN cluster(s) and paste it in the space below:



1. Perform principal component analysis on the original data (nc\_data). Then select the number of principal components based on PCs variance plot. Let’s call the number of PCs n\_pc. Then we can use the best PCs instead of the data to perform cluster analysis. To do this, run:

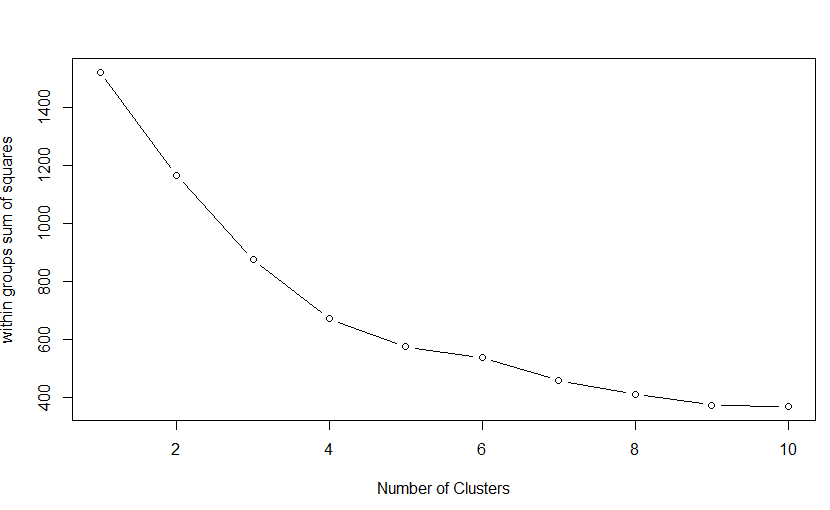
pca\_data <- predict(pca, newdata = nc\_data)

pc\_df <- as.data.frame(scale(pca\_data[,c(1:n\_pc)])) # replace n\_pc with the number of PCs you recommend.

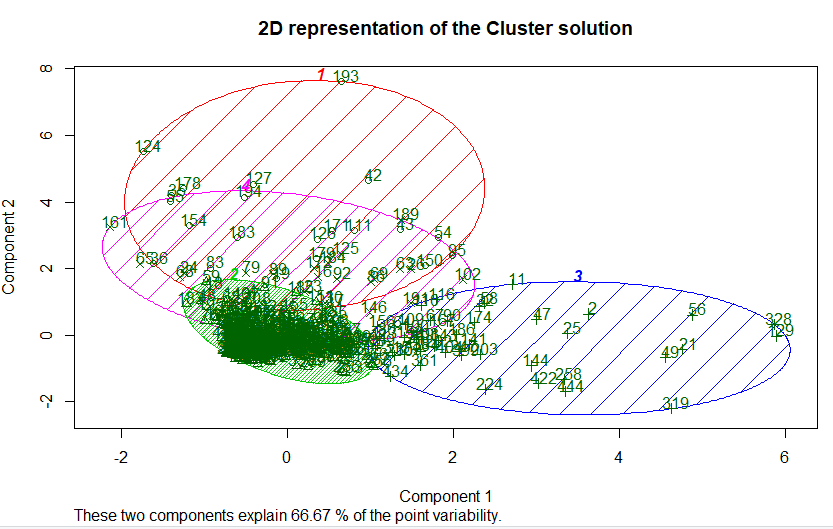


* 1. **(10 points)** Repeat your analysis in question 5 using the new pc\_df. What is the best k? Paste the two-dimensional representation in the space below:

**NOTE: used 3 for variance even though the increase was only 8.936%**

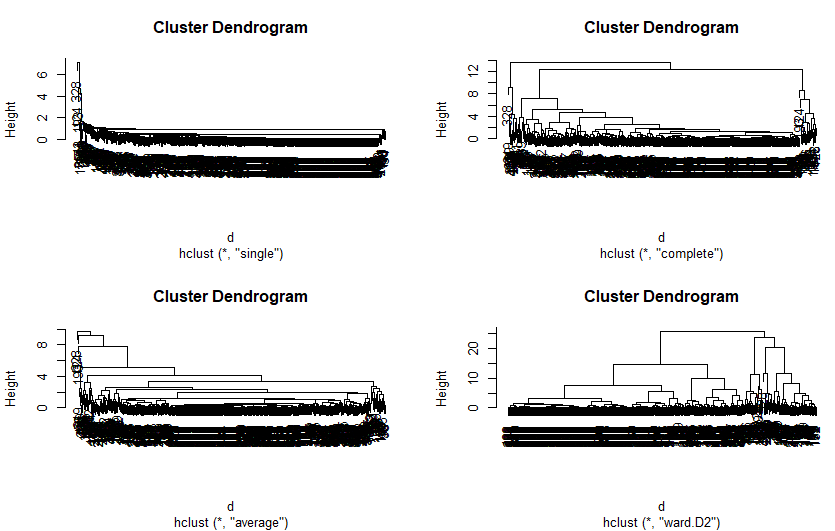


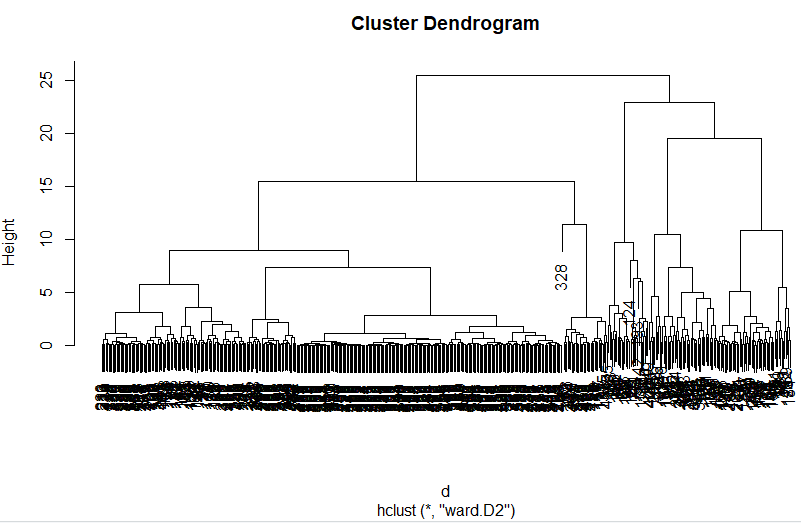
**The best k, according to the above chart, is 4**

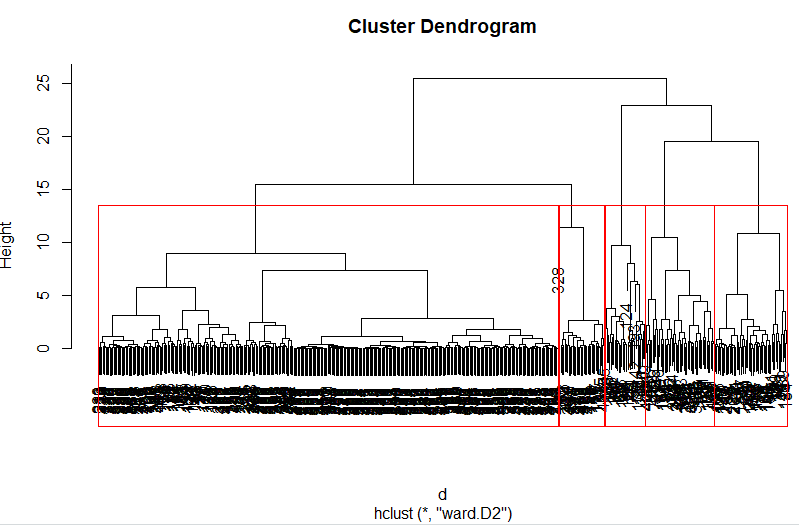


* 1. **(10 points)** Repeat your analysis in question 6 using the new pc\_df. What is the best method? What is the best k? Paste the dendrogram in the space below:

**The best method is ward and the best k is 5. See the screenshots below:**

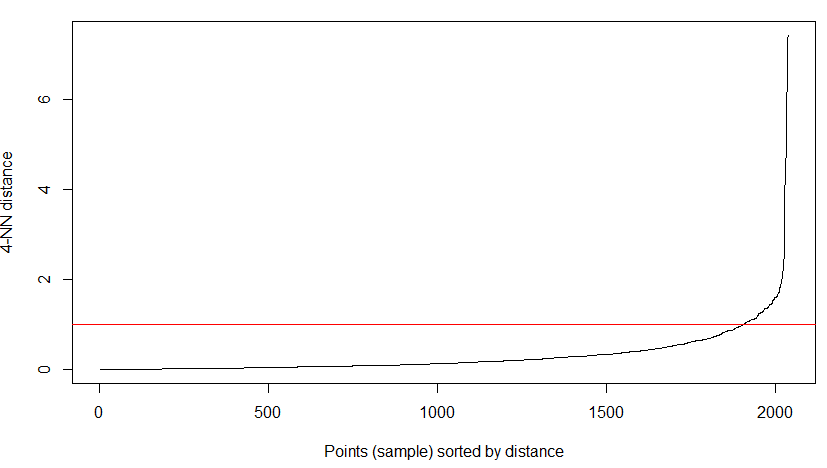


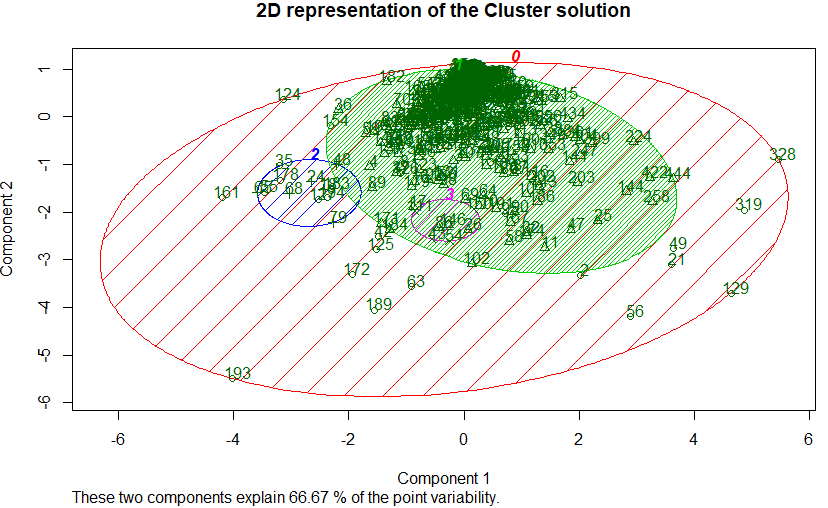




* 1. **(10 points)** Repeat your analysis in question 7 using the new pc\_df. What is the best minPts? What is the best eps? How many clusters DBSCAN returns? Perform the DBSCAN clustering and paste the two-dimensional representation in the space below:

**We use the same minPts as above. After plotting the kNN distribution, we see the eps at 1 dissects the knee. The DBSCAN returns 3 clusters and 24 noise points.**





1. For each hospital, determine the cluster (based on pc\_df) to which they belong. Then determine the value of "sales12","rbeds","hip12","knee12", and "femur12" for each cluster for each clustering method (e.g. k-means, hierarchical, DBSCAN). To do this, you need to run the following lines:

pc\_df$kmeans <- k.means.fit$cluster

pc\_df $hclust <- groups # these groups are created in hierarchical clustering

pc\_df $db <- db$cluster

pc\_df $hid <- nc\_data$hid # Add hospital id to pc\_df data

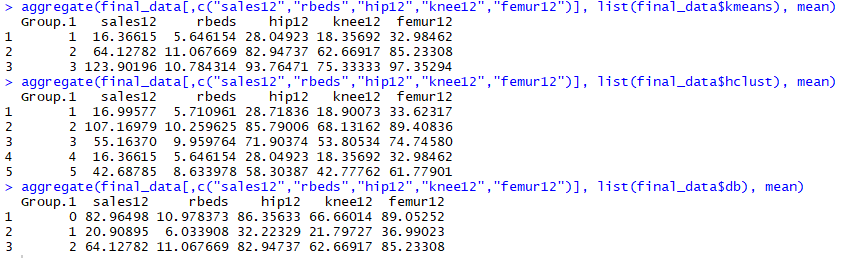
final\_data <- merge(x=pc\_df, y=nc\_data, key="hid")

aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$kmeans), mean)

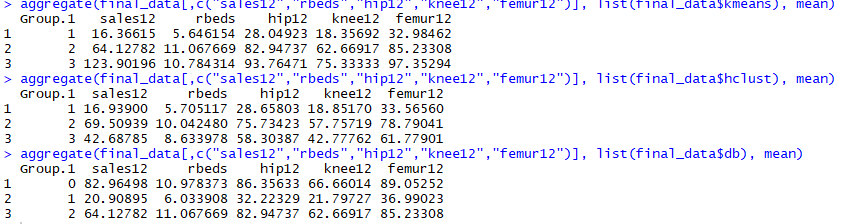
aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$hclust), mean)

aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$db), mean)

* 1. **(20 points)** Based on these results for each clustering method (e.g. k-means, hierartchical, and DBSCAN), recommend which cluster we should immediately reach out to. Give your reasons.



**Per our analysis, the ward method of clustering yielded five groups. In order for proper continuity, we edit the ward group to 3 clusters.**



**Aggregating the data from the screenshot above, we see the following totals per group**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **kmeans** | **sales** | **rbeds** | **hip12** | **knee12** | **femur12** | **sales/rbeds** | **sales/operation** |
| **group 1** | 16.36615 | 5.65 | 28 | 18.36 | 32.98 | $ 2.90 | $ 0.21 |
| **group 2** | 64.12782 | 11.07 | 82.95 | 62.67 | 85.23 | $ 5.79 | $ 0.28 |
| **group 3\*** | 123.9 | 10.78 | 93.76 | 75.33 | 97.35 | $ 11.49 | $ 0.47 |
|  |  |  |  |  |  |  |  |
| **hierarchical** | **sales** | **rbeds** | **hip12** | **knee12** | **femur12** | **sales/rbeds** | **sales/operation** |
| **group 1** | 16.94 | 5.71 | 28.66 | 18.85 | 33.57 | $ 2.97 | $ 0.21 |
| **group 2\*** | 69.51 | 10.04 | 75.73 | 57.76 | 78.79 | $ 6.92 | $ 0.33 |
| **group 3** | 42.69 | 8.63 | 58.3 | 42.78 | 61.78 | $ 4.95 | $ 0.26 |
|  |  |  |  |  |  |  |  |
| **DBSCAN** | **sales** | **rbeds** | **hip12** | **knee12** | **femur12** | **sales/rbeds** | **sales/operation** |
| **group 1\*** | 82.96 | 10.98 | 86.36 | 66.66 | 89 | $ 7.56 | $ 0.34 |
| **group 2** | 20.91 | 6.03 | 32.22 | 21.8 | 36.99 | $ 3.47 | $ 0.23 |
| **group 3** | 64.13 | 11.07 | 82.95 | 62.67 | 85.23 | $ 5.79 | $ 0.28 |

**There was no uniformity among the clustering. The kmeans saw group 3 as the leader while hierarchical had group 2 leading sales and group 1 was the sales leader of clustering method DBSCAN. The clustering method that yielded the highest sales and ROI cluster is kmeans’ group 3. If I had to choose one group, it would be group 3 from the kmeans method.**

**Appendix 1 – R Code**

setwd("C:/Users/joewa/OneDrive/Documents/MBA/MBAD 6211/2- assignments/assignment 3")

library("data.table")

data <- fread("hospital\_ortho.csv", sep=",", header=TRUE, strip.white=T, na.strings=c("NA","NaN","","?"))

data\_NC <- data[(data$state=="NC")|(data$state=="SC")|(data$state=="TN")|(data$state=="VA")|(data$state=="GA")]

library("cluster")

library("clustertend")

library("dbscan")

data\_NC$zip <- NULL

data\_NC$hid <- NULL

data\_NC$state <- NULL

data\_NC$city <- NULL

if (TRUE){

df <- scale(data\_NC[-1]) #standardize the data

} else{

df <- data\_NC[-1]

}

k.means.fit <- kmeans(df, 2)

attributes(k.means.fit)

k.means.fit$centers

k.means.fit$cluster

k.means.fit$size

hopkins(df, n = nrow(df)-1)

withinsplot <- function(data\_NC, nc=15, seed=1234){

wss <- (nrow(data\_NC)-1)\*sum(apply(data\_NC,2,var))

for (i in 2:nc){

set.seed(seed)

wss[i] <- sum(kmeans(data\_NC, centers = i)$withinss)}

plot(1:nc, wss, type="b", xlab="Number of Clusters",

ylab="within groups sum of squares")}

withinsplot(df, nc=10)

############################################################################

k.means.fit1 <- kmeans(df, 3)

attributes(k.means.fit1)

k.means.fit1$centers

k.means.fit1$cluster

k.means.fit1$size

hopkins(df, n = nrow(df)-1)

withinsplot <- function(data, nc=15, seed=1234){

wss <- (nrow(data\_NC)-1)\*sum(apply(data\_NC,2,var))

for (i in 2:nc){

set.seed(seed)

wss[i] <- sum(kmeans(data\_NC, centers = i)$withinss)}

plot(1:nc, wss, type="b", xlab="Number of Clusters",

ylab="within groups sum of squares")}

withinsplot(df, nc=10)

##################################################################################

clusplot(df, k.means.fit$cluster, main='2D representation of the Cluster solution', color=TRUE, shade=TRUE,

labels=2, lines=0)

data\_NC$kmeans <- k.means.fit1$cluster

d <- dist(df, method = "euclidean")

H.single <- hclust(d, method = "single")

plot(H.single)

H.complete <- hclust(d, method = "complete")

plot(H.complete)

H.average <- hclust(d, method = "average")

plot(H.average)

H.ward <- hclust(d, method = "ward.D2")

plot(H.ward)

groups <- cutree(H.ward, k=3)

plot(H.ward)

rect.hclust(H.ward, k = 3, border="red")

clusplot(df, groups, main="2D representation of the Cluster Solution",

color = TRUE, shade = TRUE,

labels = 2, lines = 0)

##############################################################################################

kNNdistplot(df, k = 4)

abline(h=4, col = "red")

db <- dbscan(df, eps = 4, minPts = 4)

db

clusplot(df, db$cluster, main='2D representation of the Cluster solution',

color=TRUE, shade=TRUE,

labels=2, lines=0)

pca <- prcomp(data\_NC[-1], center = TRUE, scale = TRUE)

summary(pca)

plot(pca, type = "lines")

library("devtools")

pca\_data <- predict(pca, newdata = data\_NC)

pc\_df <- as.data.frame(scale(pca\_data[,c(1:3)]))

k.means.fit.pc <- kmeans(pc\_df, 3)

attributes(k.means.fit.pc)

k.means.fit.pc$centers

k.means.fit.pc$cluster

k.means.fit.pc$size

hopkins(pc\_df, n = nrow(pc\_df)-1)

withinsplot <- function(data\_NC, nc=15, seed=1234){

wss <- (nrow(data\_NC)-1)\*sum(apply(data\_NC,2,var))

for (i in 2:nc){

set.seed(seed)

wss[i] <- sum(kmeans(data\_NC, centers = i)$withinss)}

plot(1:nc, wss, type="b", xlab="Number of Clusters",

ylab="within groups sum of squares")}

withinsplot(pc\_df, nc=10)

####################################################################################

k.means.fit.pc1 <- kmeans(pc\_df, 4)

k.means.fit.pc1$cluster

clusplot(pc\_df, k.means.fit.pc1$cluster, main='2D representation of the Cluster solution', color=TRUE, shade=TRUE,

labels=2, lines=0)

d <- dist(pc\_df, method = "euclidean")

H.single <- hclust(d, method = "single")

plot(H.single)

H.complete <- hclust(d, method = "complete")

plot(H.complete)

H.average <- hclust(d, method = "average")

plot(H.average)

H.ward <- hclust(d, method = "ward.D2")

plot(H.ward)

par(mfrow=c(1,1))

plot(H.single)

plot(H.complete)

plot(H.average)

plot(H.ward)

groups <- cutree(H.ward, k=3)

plot(H.ward)

rect.hclust(H.ward, k = 5, border="red")

############################################################################################

kNNdistplot(pc\_df, k = 4)

abline(h=1, col = "red")

db <- dbscan(pc\_df, eps = 1, minPts = 4)

db

clusplot(pc\_df, db$cluster, main='2D representation of the Cluster solution',

color=TRUE, shade=TRUE,

labels=2, lines=0)

pc\_df$kmeans <- k.means.fit.pc1$cluster

pc\_df$hclust <- groups

pc\_df$db <- db$cluster

pc\_df$hid <- data\_NC$hid

final\_data <- merge(x=pc\_df, y=data\_NC, key="hid")

aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$kmeans), mean)

aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$hclust), mean)

aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$db), mean)