Cluster Analysis

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: Steps I followed.

1. Read the file:

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

2. 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.

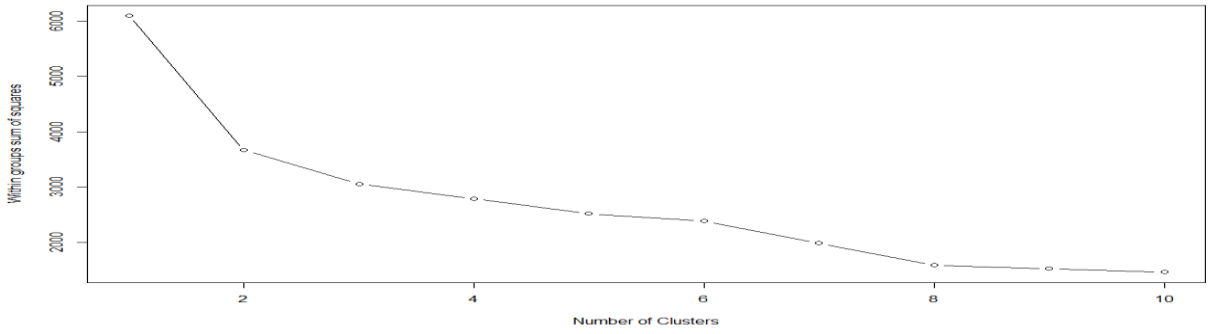
```
nc_data <- data[(data$state == "NC") | (data$state == "SC") | (data$state == "VA") | (data$state == "GA") | (data$state == "TN")]
```

2.1. **(3 points)** Looking at each individual variable I decide if it should be included in cluster analysis. For those variables that I decide not to include,
Below mentioned variables were removed and reasons for removing them

1. ZIP: US POSTAL CODE - Does not hold specific relevance to the analysis as we already have sorted the hospitals according to state.
2. HID : Unique reference number for each hospital does not hold much significance in the analysis.
3. CITY: CITY NAME – Character cannot be included for a clustering analysis
4. STATE: STATE NAME - Character cannot be included for a clustering analysis
5. TH: TEACHING HOSPITAL? 0, 1 – Binary variable removed and does not makes the correct judgement if it drives the sales of orthopedic items. The mean value is not meaningful on this kind of data.
6. TRAUMA: DO THEY HAVE A TRAUMA UNIT? 0, 1 - Binary variable removed and does not makes the correct judgement if it drives the sales of orthopedic items. The mean value is not meaningful on this kind of data.
7. REHAB: DO THEY HAVE A REHAB UNIT? 0, 1 - Binary variable removed and does not makes the correct judgement if it drives the sales of orthopedic items. The mean value is not meaningful on this kind of data.

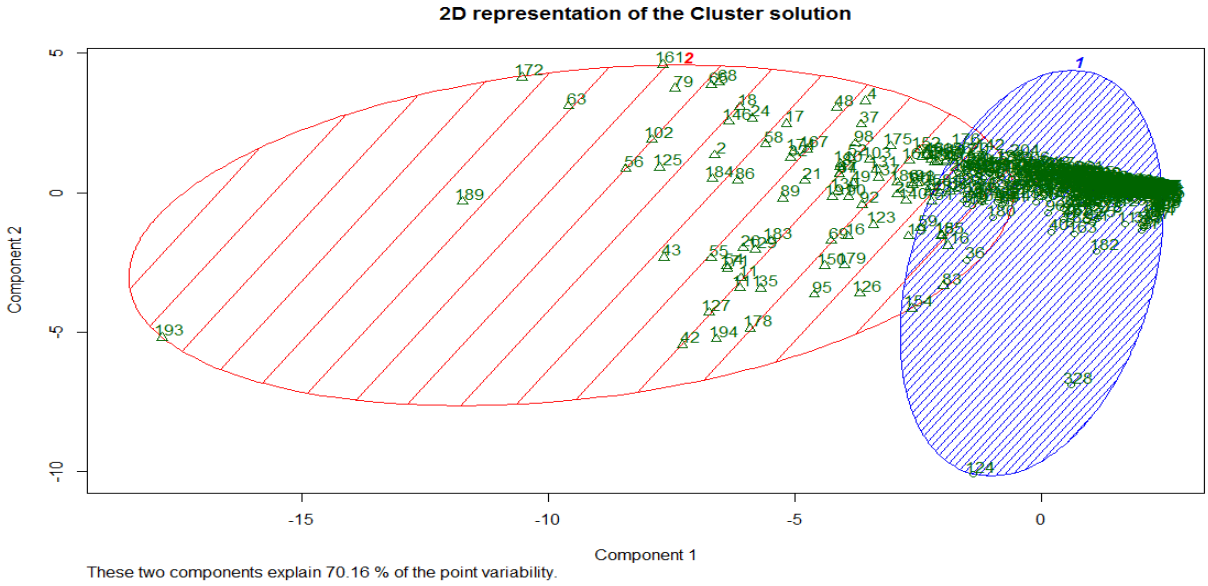
2.2. We need to scale this data as we have mixed numerical data, where each attribute is something entirely different, has different units attached and these values aren't really comparable hence scale to give equal weight to them.

3. Perform k-means clustering:
   3.1. Using "Within Groups SSE" to determine the number of clusters for k-means. How many clusters you would like to create and looking at the graph below I would create 2 clusters
   3.2. "Within Groups SSE" plot in the space below:



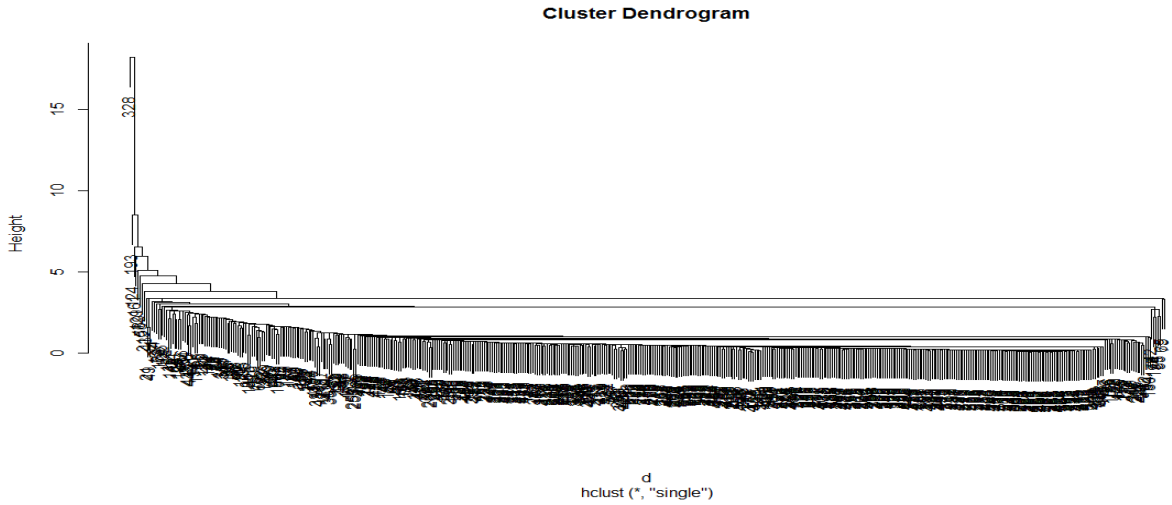
3.3. k-means clustering using the number of clusters recommended in 3.1. Cluster distribution Cluster 1 – 417 and Cluster 2 – 92

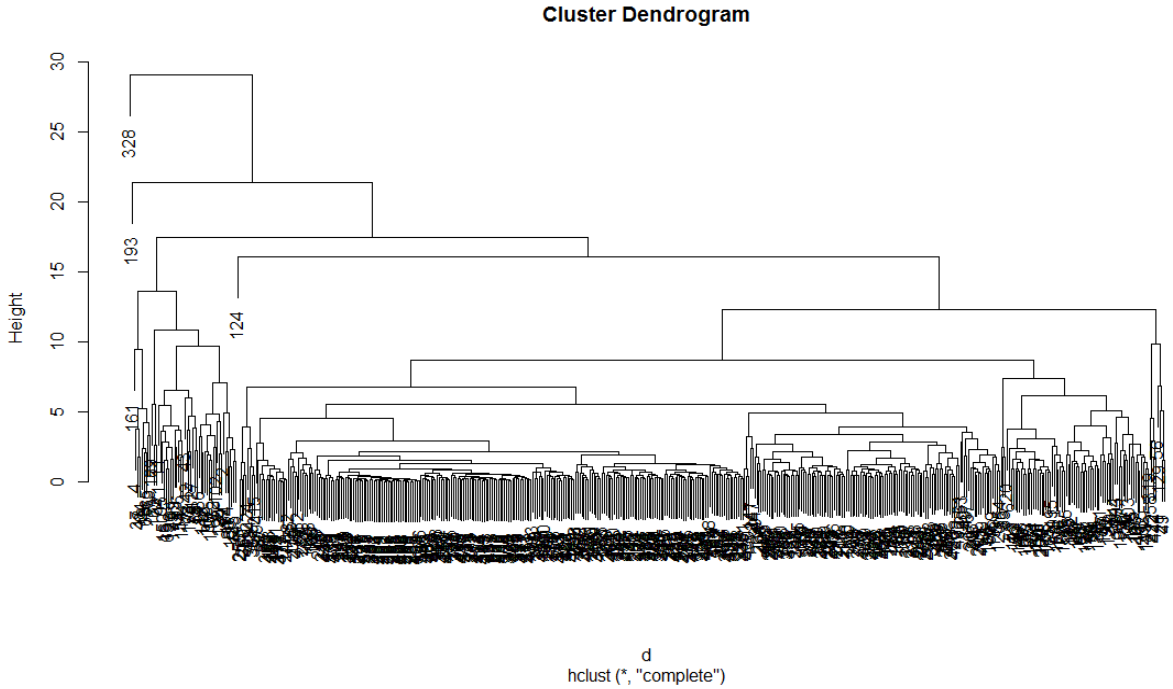
3.4. Two-dimensional representation of the clusters

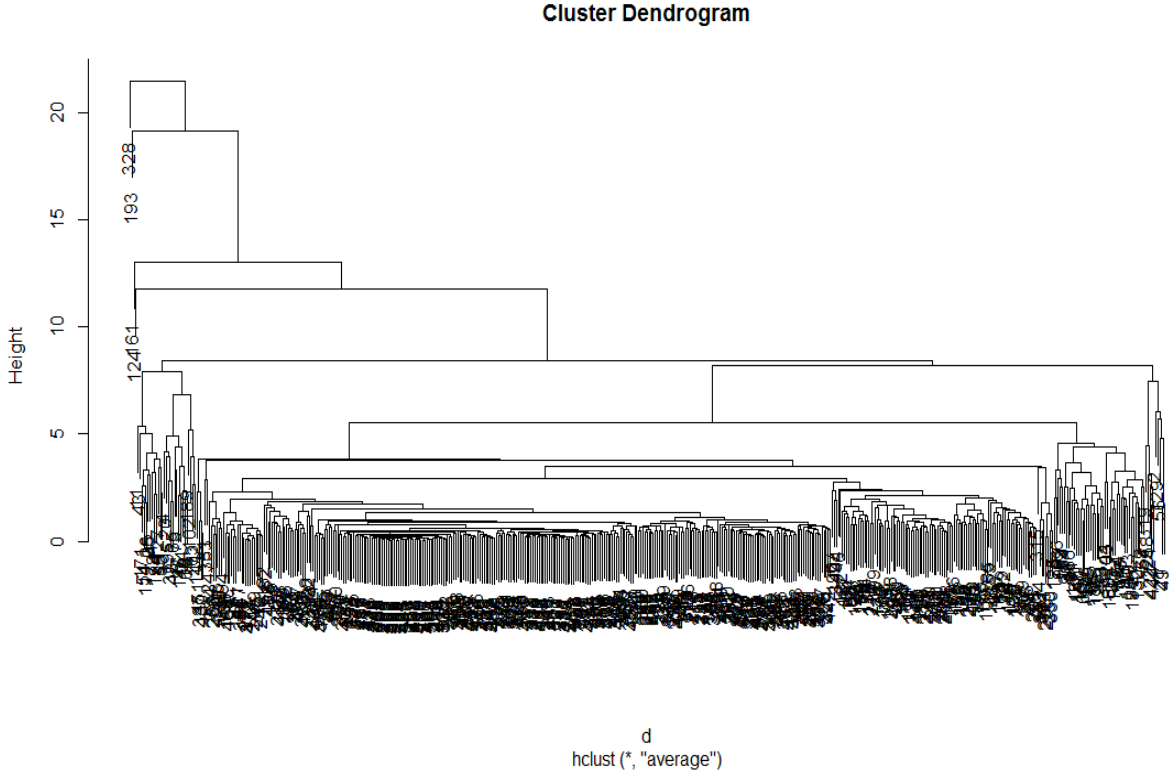


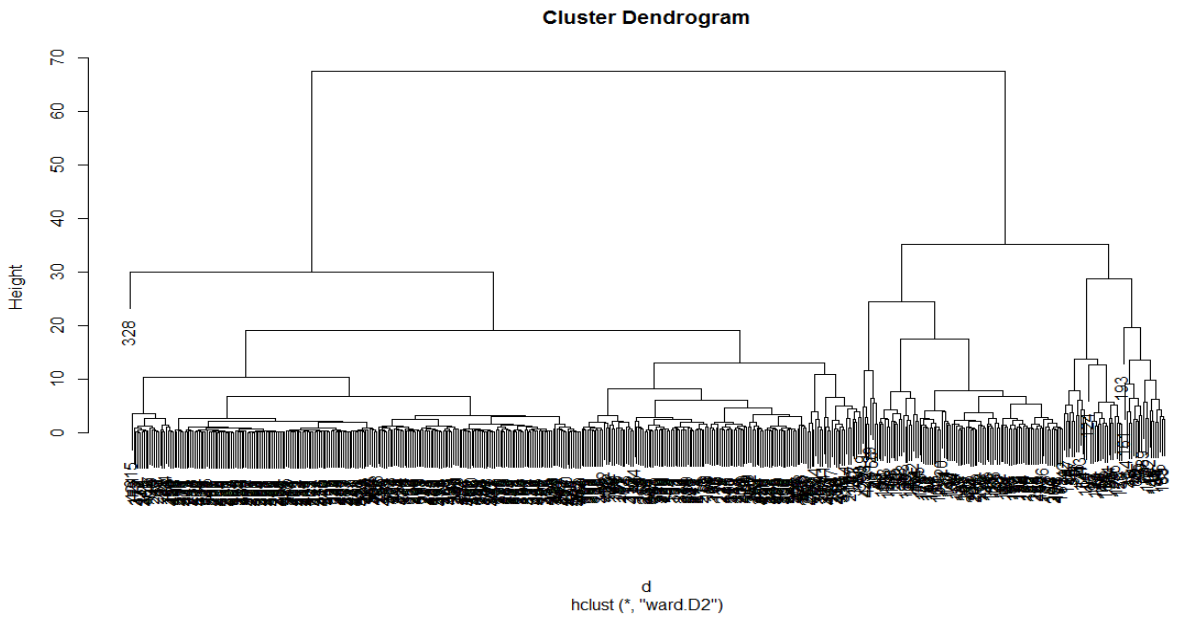
4. Hierarchical clustering.

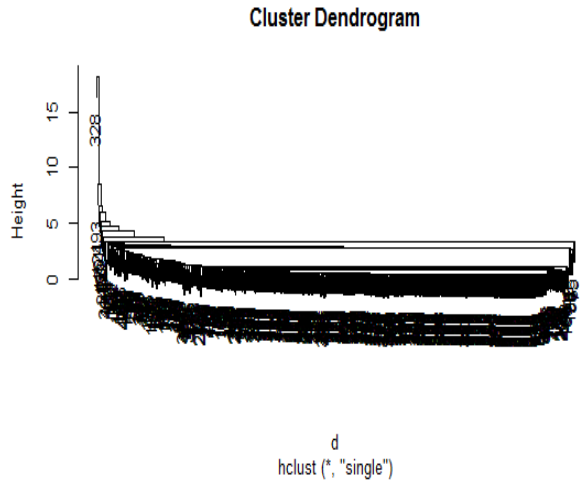
4.1. Different hierarchical clustering and the dendrograms in the space below:

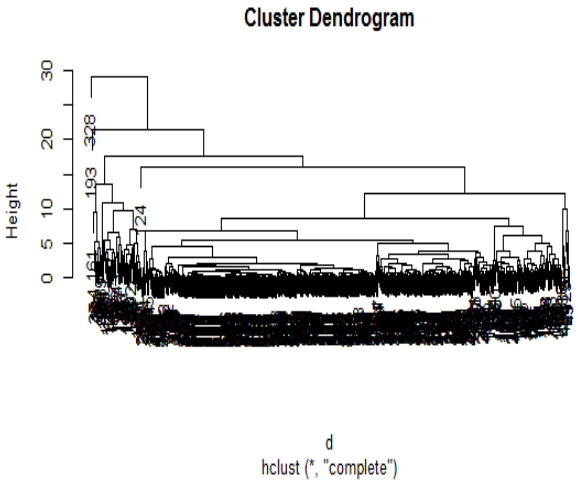


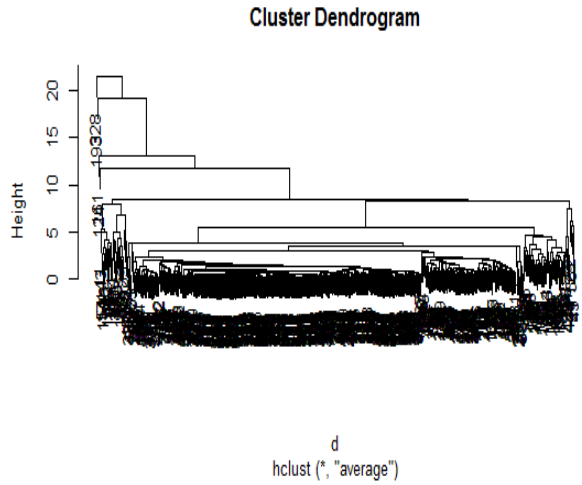

Cluster Dendrogram
Height
0
5
10
15
20
25
30
328
193
124
161
d
hclust (*, "complete")

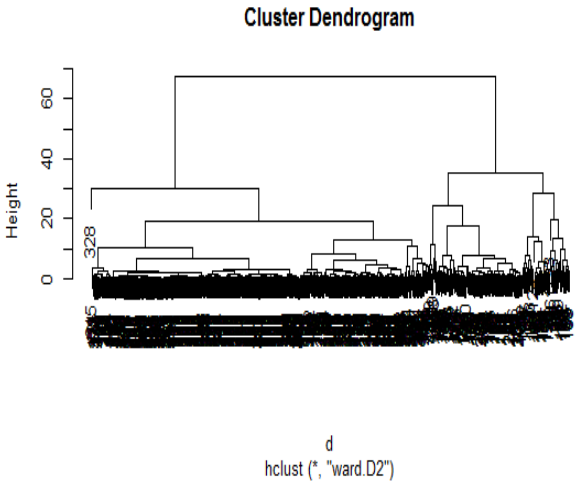

Cluster Dendrogram
Height
0
5
10
15
20
328
193
124
161
d
hclust (*, "average")


Cluster Dendrogram
Height
0
10
20
30
40
50
60
70
328
d
hclust (*, "ward.D2")


Cluster Dendrogram
Height
0
5
10
15
328
d
hclust (*, "single")


Cluster Dendrogram
Height
0
5
10
20
30
328
193
24
161
d
hclust (*, "complete")

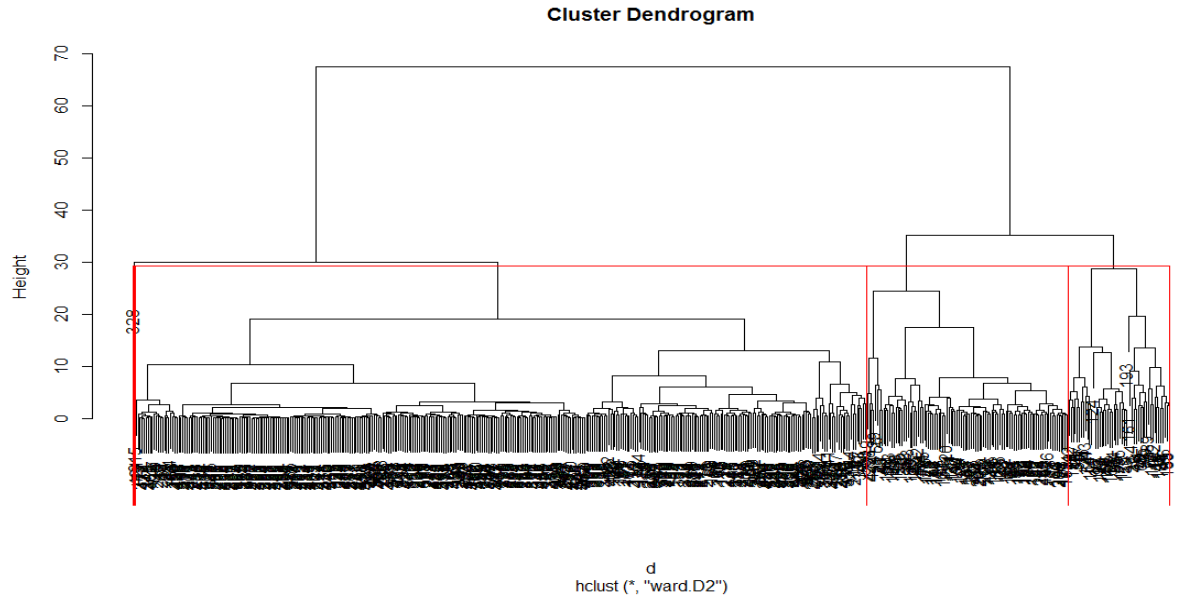

Cluster Dendrogram
Height
0
5
10
15
20
328
193
d
hclust (*, "average")


Cluster Dendrogram
Height
0
20
40
60
328
d
hclust (*, "ward.D2")

4.2. Ward Visually the dendogram seems well split for Ward then for rest of the others In the Kmeans we found few noise and outliers for the clusters and ward algorithm is less susceptible to them.

4.3. Based on hierarchical clustering results, I find 4 clusters in the data.

4.4. The dendrogram that with the red borders around the clusters in the space below:

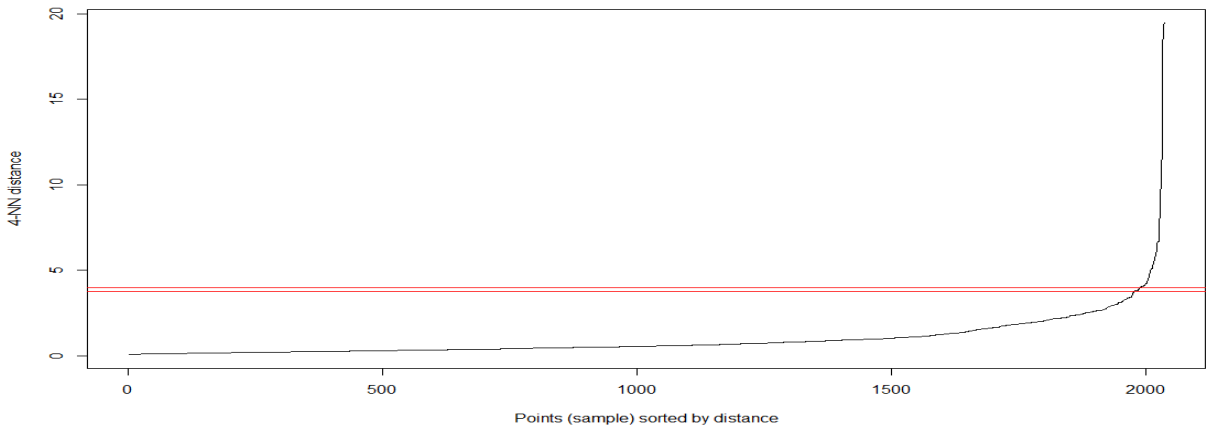


5. DBSCAN cluster analysis:

5.1. First, you need to determine minPts. The rule of thumb for minPts is the number of dimensions of the data + 1. I have used PCA to determine the minimum number of dimensions. A good number of minPts would be 3+1 = 4.

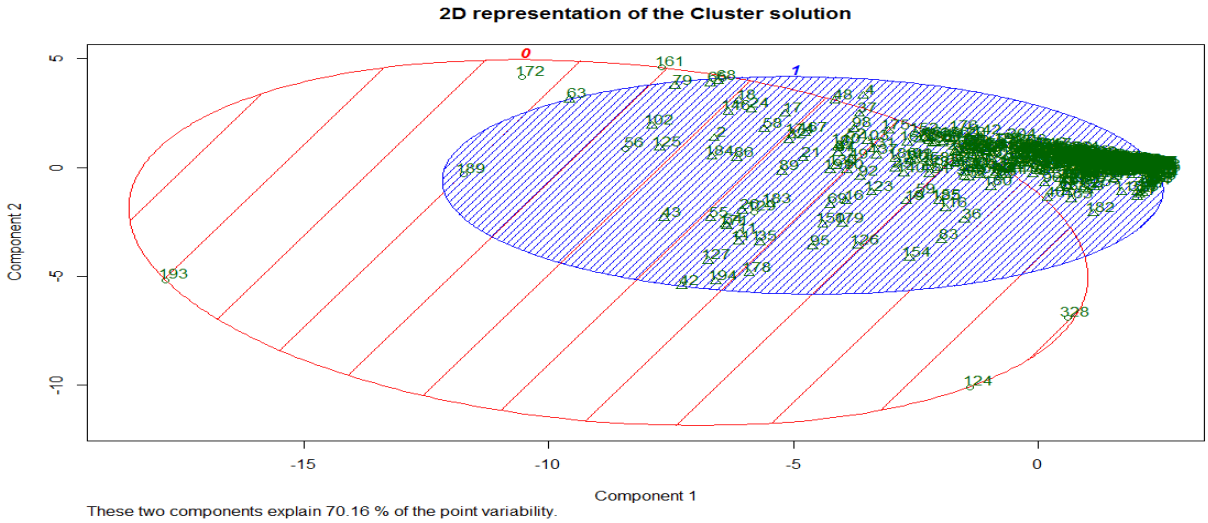
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
PC8
Standard deviation     2.6476 1.1872 1.1571 0.98627 0.8154 0.45574 0.37531 0.
29731
Proportion of Variance 0.5842 0.1175 0.1116 0.08106 0.0554 0.01731 0.01174 0.
00737
Cumulative Proportion  0.5842 0.7016 0.8132 0.89426 0.9497 0.96697 0.97871 0.
98608
                           PC9    PC10    PC11    PC12
Standard deviation     0.25274 0.22008 0.18243 0.14663
Proportion of Variance 0.00532 0.00404 0.00277 0.00179
Cumulative Proportion  0.99140 0.99543 0.99821 1.00000
```

5.2. Based on the minPts the eps was found to be 4.0. The knee in the graph was found to be at the point 4.0 there on the distance between the nearest neighbors increases.



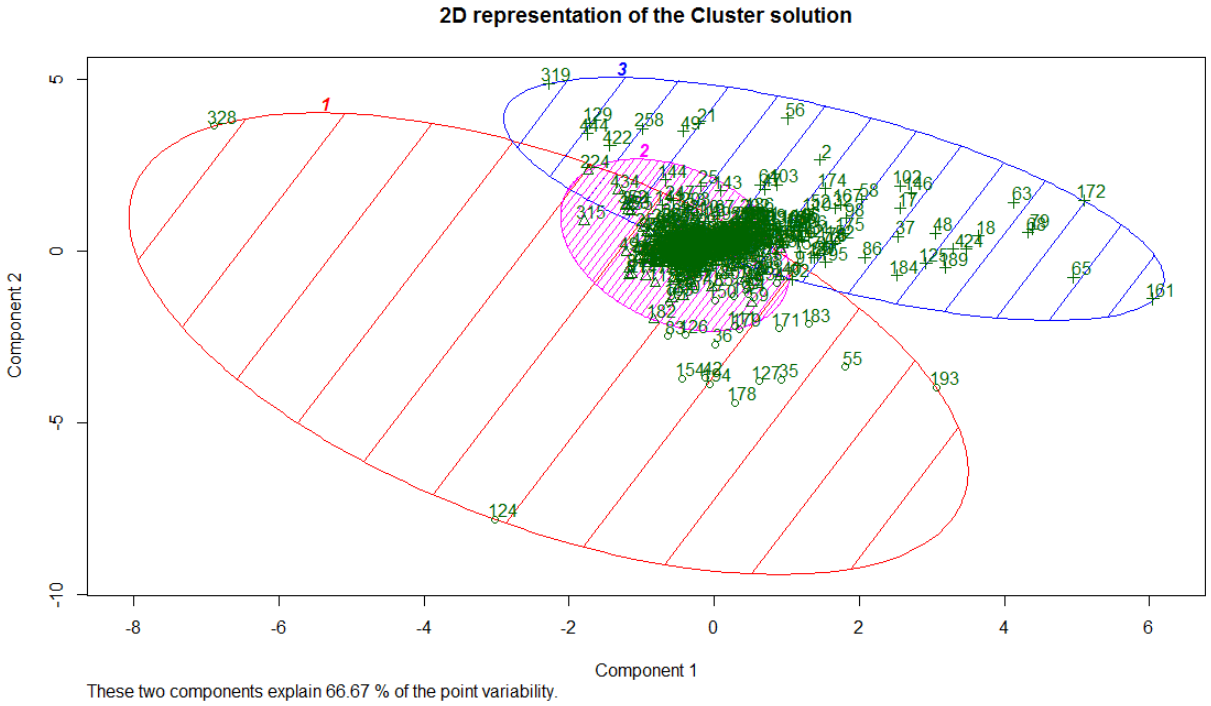
5.3. DBSCAN clustering using the minPts and eps was performed DBSCAN returns 1 cluster And 8 Noise points.

5.4. Two-dimensional representation of DBSCAN cluster(s) in the space below:



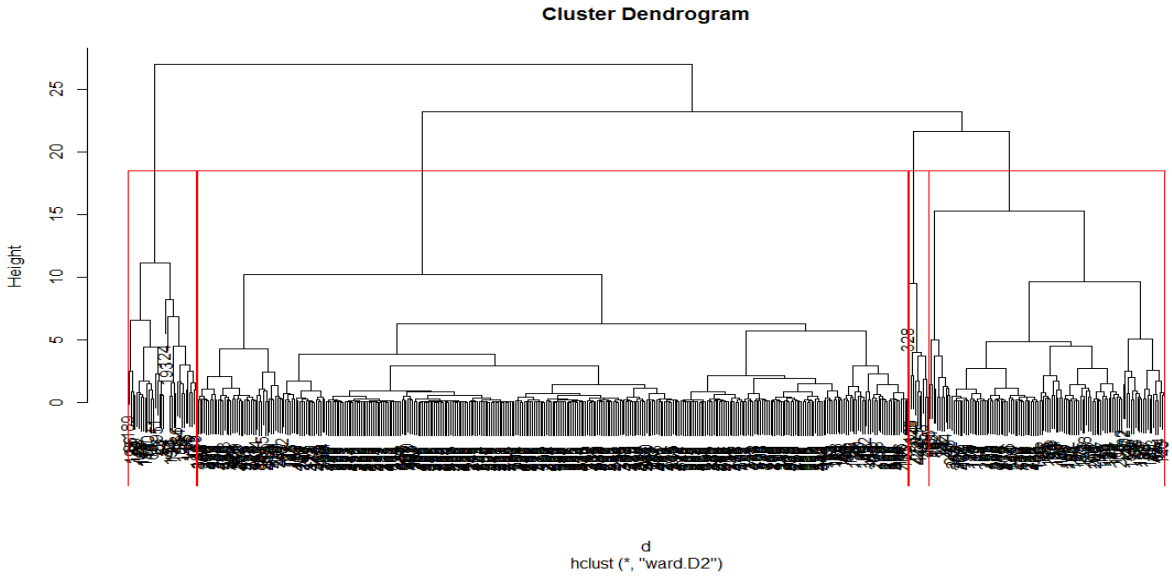
6. Principal component analysis on the original data (nc_data) was performed to select the number of principal components based on PCs variance plot. Let's call the number of PCs n_pc. Then we can use the best PCs instead of the data to perform cluster analysis. To do this, run:

6.1. Repeating the analysis using the new pc_df. Best k was found to be 3

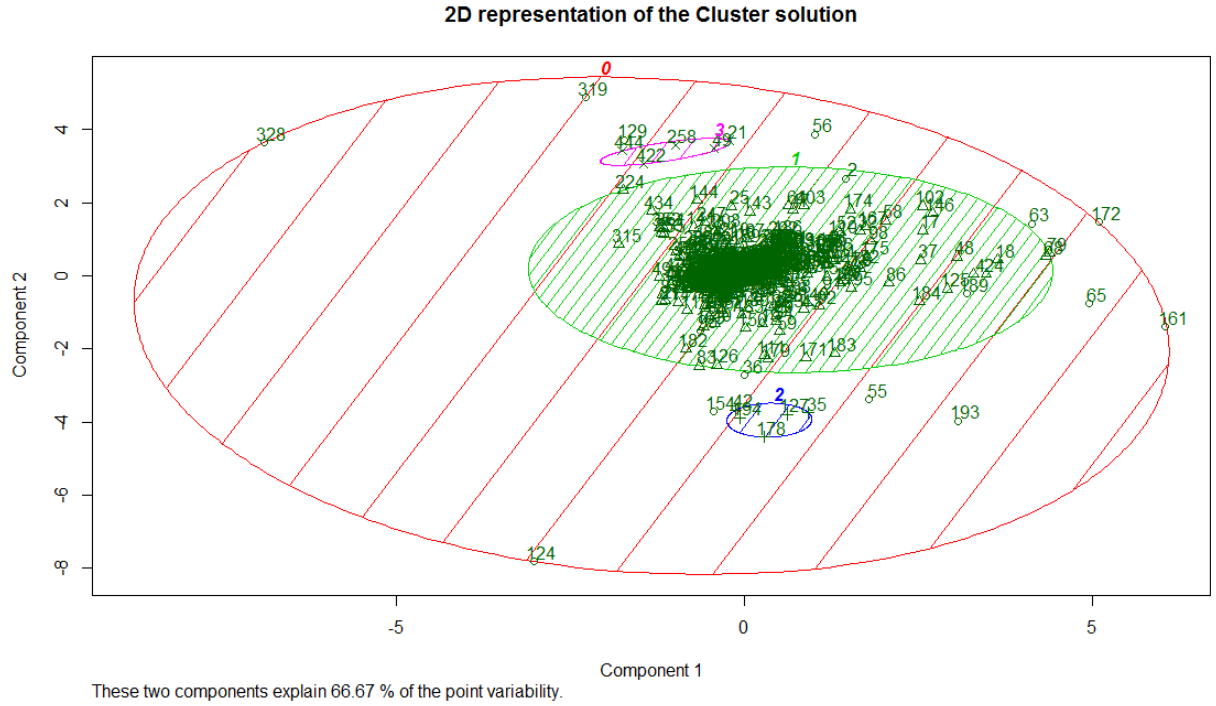


6.2. Repeating the hierarchical analysis in using the new pc_df to select the best method .The dendrogram in the space below:

The best method is Ward. The best k is 4



6.3. **(10 points)** Repeating the DBSCAN analysis in using the new pc_df. What is the best minPts? What is the best eps? How many clusters DBSCAN returns? the two-dimensional representation in the space below:
Best minPts – 4 as the PC selected is 3, Best eps 1 . Clusters returned – 3 and 17 noise points.



7. For each hospital, determining the cluster (based on pc_df) to which they belong. Then determining the value of "sales12","rbeds","hip12","knee12", and "femur12" for each cluster for each clustering method (e.g. k-means, hierarchical, DBSCAN).

Based on these results for each clustering method (e.g. k-means, hierartchical, and DBSCAN), recommend which cluster we should immediately reach out to is the question

KMEANS

Group	Sales12	Rbeds	Hip12	Knee12	Femur12
1	392.96	50.26	131.84	90.00	138.00
2	15.65	4.99	24.95	15.65	29.87
3	40.77	6.45	124.42	102.38	124.65

HCLUST:

Group	Sales12	Rbeds	Hip12	Knee12	Femur12

1	12.93	5.07	19.34	11.43	24.72
2	28.49	5.4	109.36	91.11	108.51
3	357.67	14.02	141.44	96.64	142.94
4	18.50	97.50	68.40	29.60	94.30

DBSCAN

Group	Sales12	Rbeds	Hip12	Knee12	Femur12
0	254.29	64.23	169.52	140.35	165.58
1	27.07	5.36	43.21	31.26	47.91
2	558.40	24.20	168.60	111.20	141.40
3	0.600	11.40	75.20	26.00	81.20

The objective of this study was to find ways to increase sales of orthopedic material from the company to hospitals in the United States. Orthopedic materials are used in operations and procedures related to bones. Hence considering that the numbers of knee, hip and femur operations from previous year are compared against the sales of the company for the year

Looking at the results we should immediately reach out to cluster number 3 in DBSCAN as mean of number of hip knee and femur operation is quite high as compared to the mean of the sales. The number of rehab beds seems decent enough to be targeted. This indicates there is an window of opportunity of increasing sales if the hospitals in the mentioned clusters are approached and targeted. In order to increases the current sales of the company the company should target hospitals where there sales are low but the number of operation and procedures related to orthopedics are reasonable.

Rest of the clusters indicates sales complementing the number of operations.