***#1. Run K-means clustering on the above heart disease dataset and answer the***

***#following questions***

***#1) Why should the attribute “class” in heart-c.csv (“num”) not be included for clustering?***

***# 2) Run K-means algorithm by choosing different numbers of clusters, numCluster = 2, 3, 4,5,***

***#then observe the differences of clusters generated:***

***#a. How are the Within Cluster Sum of Squared Errors1***

***#changed for different numbers of***

***# clusters?***

***#b. What can you conclude?***

***#c. How can you explain this conclusion from clustering analysis point of view?***

***#Library used for assignment***

***# caret Library for confusion matrix***

***library(caret)***

***library(e1071)***

***library(cluster)***

***library(factoextra)***

***# pulling over heart dataset.***

***heart\_class <- read\_csv("heart-c.csv")------------1***

***head(heart\_class)***

***# Appyling summary function to determine its statistical data.---2***

***summary(heart\_class)***

***heart\_class <- heart\_class %>%***

***mutate(num=ifelse(num == "<50", 0,***

***ifelse(num == ">50\_1", 1, NA)))***

***# removing X1 identifier in order to blind the subject---3***

***heart\_identifier<- heart\_class[ , !(names(heart\_class) %in% c("X1","num"))]***

***# determining the data type of the variables.-----4***

***str(heart\_identifier)***

***#Assiging the values as per the specification provided.---5***

***heart\_identifier <- heart\_identifier %>%***

***mutate(sex=ifelse(sex == "male", 1,***

***ifelse(sex == "female", 0, NA))) %>%***

***mutate(cp=ifelse(cp == "typ\_angina", 1,***

***ifelse(cp == "atyp\_angina", 2,ifelse(cp == "non\_anginal", 3,ifelse(cp == "asympt", 4, NA))))) %>%***

***mutate(fbs=ifelse(fbs == "TRUE", 1,***

***ifelse(fbs == "FALSE", 0, NA))) %>%***

***mutate(restecg=ifelse(restecg== "normal", 0,***

***ifelse(restecg == "st\_t\_wave\_abnormality", 1,***

***ifelse(restecg=="left\_vent\_hyper",2,NA)))) %>%***

***mutate(exang=ifelse(exang=="yes", 1, ifelse(exang == "no", 0, NA)))%>%***

***mutate(slope=ifelse(slope=="up",1,ifelse(slope=="flat",2,ifelse(slope=="down",3,NA))))%>%***

***mutate(thal=ifelse(thal=="normal",3,ifelse(thal=="fixed\_defect",6,ifelse(thal=="reversable\_defect",7,NA))))***

***# sacling of data.---6***

***heart\_sacle <- scale(heart\_identifier)***

***# Determining the statistical values post scaling up the data values.---7***

***summary(heart\_sacle)***

***#Removing the missing value from the dataset----8***

***library(mice)***

***repaired\_heart<-mice(heart\_sacle)***

***heart\_missing <- complete(repaired\_heart,1)***

***summary(heart\_missing)***

***str(heart\_missing)***

***heart\_hierarchial <- heart\_missing***

***# seed function is used in order to keep the reproducibility.----9***

***seed\_val <- 302***

***set.seed(seed\_val)***

***#run K-means with clusters=2---10***

***run1= kmeans(heart\_missing, centers = 2, nstart = 1)***

***run2= kmeans(heart\_missing, centers = 2, nstart = 1)***

***run3= kmeans(heart\_missing, centers = 2, nstart = 1)***

***#In all 3 runs above the cluster size is consistent, results are reproducible.--11***

***heart\_res <- heart\_missing***

***#Create a heart dataset copy and assign predicted cluster label of each row--12***

***heart\_res[, "first\_clust"] <- run1$cluster***

***heart\_res <- heart\_res %>%***

***mutate(first\_clust = ifelse(first\_clust == "2", 0,***

***ifelse(first\_clust == "1", 1, NA)))***

***#Forming a matrix matching actual label & predicted label ---13***

***Cmatrix <- confusionMatrix(as.factor(heart\_res$first\_clust),as.factor(heart\_class$num))***

***ConfusionMatrix <- Cmatrix$table***

***ConfusionMatrix***

***#Accuracy is no.of valid prediction/total samples\*100---14***

***Accuracy\_2cluster=(Cmatrix$table[1,1]+Cmatrix$table[2,2])/(nrow(heart\_res))***

***\*100Accuracy\_2cluster***

***#run K-means with clusters=3---15***

***run31= kmeans(heart\_missing, centers = 3, nstart = 1)***

***run32= kmeans(heart\_missing, centers = 3)***

***run33= kmeans(heart\_missing, centers = 3)***

***heart\_res[, "second\_clust"] <- run31$cluster***

***#run K-means with clusters=4---16***

***run41= kmeans(heart\_missing, centers = 4, nstart = 1)***

***run42= kmeans(heart\_missing, centers = 4)***

***run43= kmeans(heart\_missing, centers = 4)***

***heart\_res[, "third\_clust"] <- run41$cluster***

***#run K-means with clusters=5--17***

***run51= kmeans(heart\_missing, centers = 5, nstart = 1)***

***run52= kmeans(heart\_missing, centers = 5)***

***run53= kmeans(heart\_missing, centers = 5)***

***#running mean function in order to get the dimension---18***

***heart\_K <- kmeans(heart\_missing , centers = 5)***

***heart\_K***

***#dimension of clusters ---19***

***fviz\_cluster( heart\_K, geo="points", data=heart\_missing)***

***heart\_K $tot.withinss***

***# calculating Sum square error, which will be ---20***

***#further used for plotting to determine the elbow region***

***error<-vector()***

***for (i in 1:12){***

***error[i]<-kmeans(heart\_missing,i)$tot.withinss***

***}***

***plot(error)***

***## hierarchical clustering---21***

***#calculating the distance using dist function.***

***dist\_mat <- dist(heart\_hierarchial)***

***#Running hierarchical clustering algorithm and plotting the same ,using complete method--22***

***hclust\_avg <- hclust(dist\_mat, method = 'complete')***

***plot(hclust\_avg)***

***#Running hierarchical clustering algorithm and plotting the same ,using single method--23***

***hier\_clust\_2 <- hclust(dist\_mat, method = "single")***

***plot(hier\_clust\_2)***

***#Running hierarchical clustering algorithm and plotting the same ,using average method--24***

***hier\_clust\_3 <- hclust(dist\_mat, method = "average")***

***plot(hier\_clust\_3)***

***# Considering completed method for hierarchical clustering, below code brings***

***#in cosmetic changes ---25***

***library(dendextend)***

***avg\_dend\_obj <- as.dendrogram(hclust\_avg)***

***avg\_col\_dend <- color\_branches(avg\_dend\_obj, h = 5)***

***plot(avg\_col\_dend)***

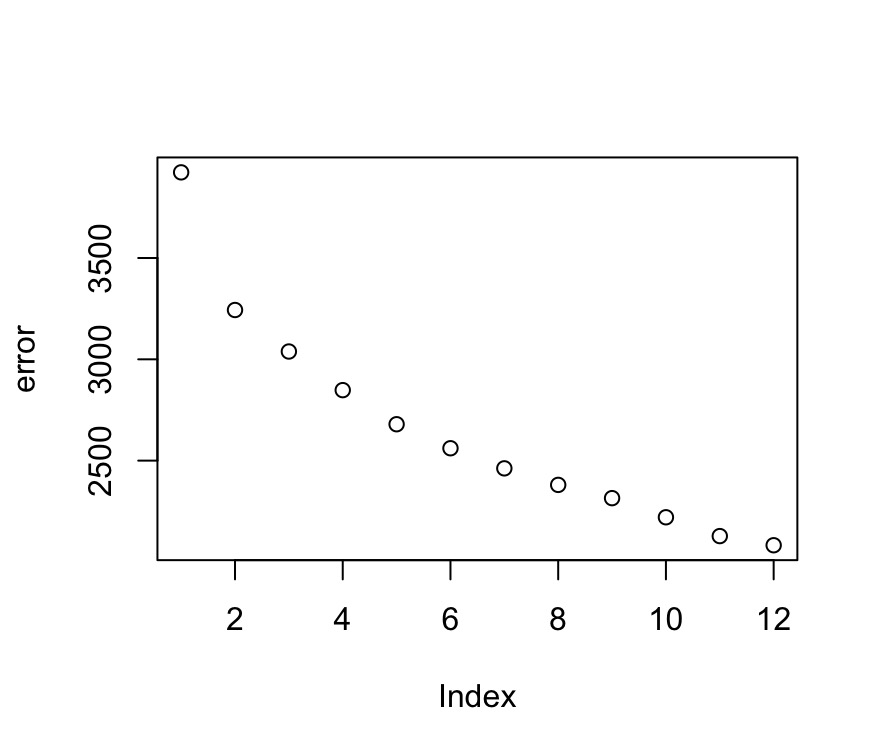
***1) Run K-means clustering on the above heart disease dataset and answer the following questions:***

1) Why should the attribute “*class”* in ***heart-c.csv*** (“***num***”) **not** be included for clustering?

Ans 1) Since K- means is an unsupervised learning, the expected output cluster labels are not passed as inputs.These labels are passed to the algorithm in case of supervised learning in order to train the model, as per the labels.

If we observe “num” column in the heart\_c.csv file which holds a ground truth labels as < 50 and >50\_1, so it cannot be passed as input.

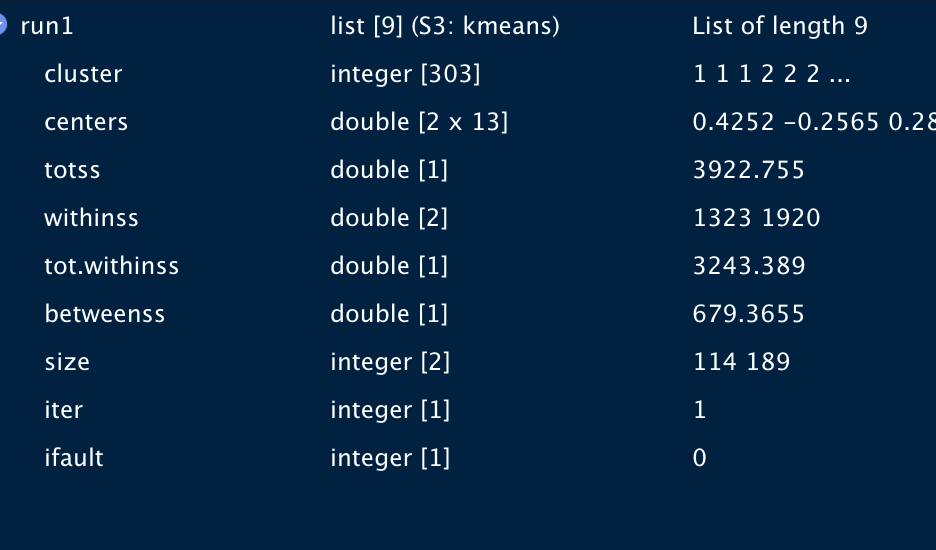
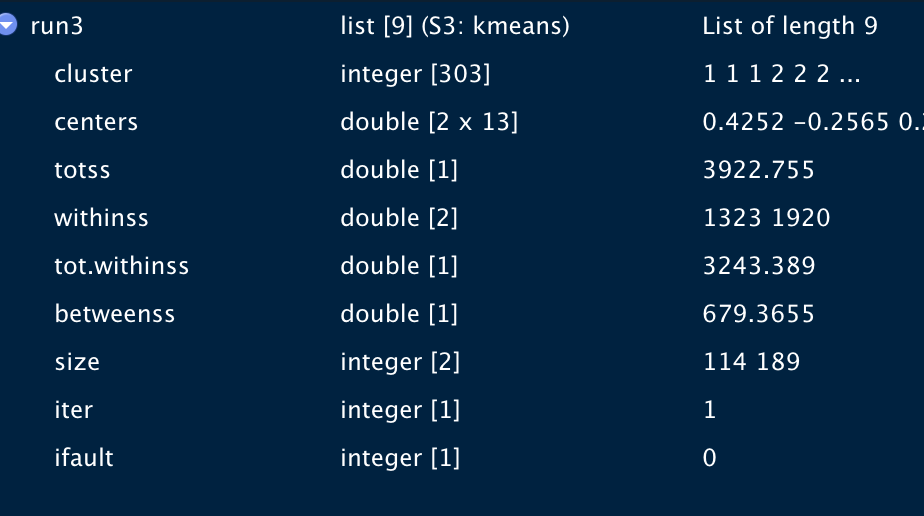
1. ***Run K-means algorithm by choosing different numbers of clusters, num Cluster = 2, 3, 4,5, then observe the differences of clusters generated:( refer code (14-18,9)***
2. ***How are the Within Cluster Sum of Squared Errors changed for different numbers of clusters? ( refer code=18,19,20)***

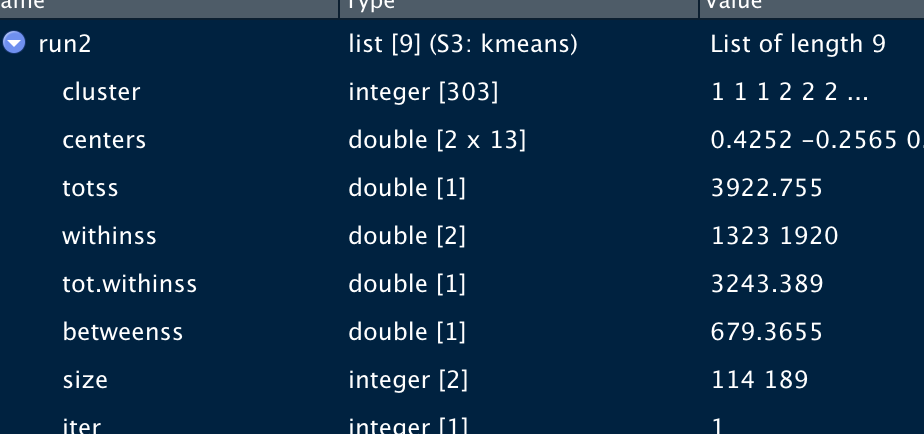


Ans a) From the above graph where X axis is the number of clusters and the Y axis which is sum of squared error, it is been observed that for cluster one the error is quite high but for cluster 2 there is a significant drop in the error the elbow region of the graph and gradually the error rate decreases as the number of clusters increase , ideally the error rate will reach zero if the number of clusters is equal to the number of data points we have , but as cluster 2 shows a significant drop in the error we will consider it as optimal solution.

Ans b) Post running K mean for num Cluster = 2, 3, 4,5, we can infer that num cluster = 2 is an optimal solution. For the below reasons.(14 to18, 9, 13)

1. K mean execution results are reproducible for cluster= 2 i.e the cluster size for each execution is consistent.

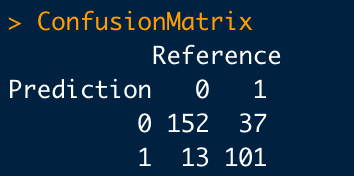
******

******

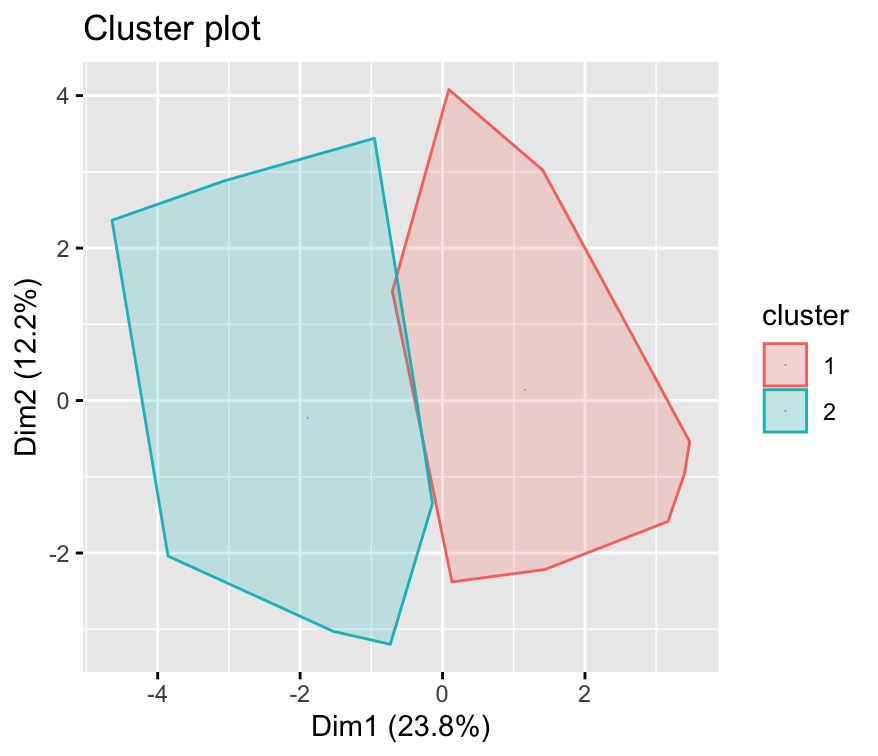
2) As per the the sum squared error graph it can be concluded that cluster num =2 , is the elbow region of the graph and is considered as suitable solution.

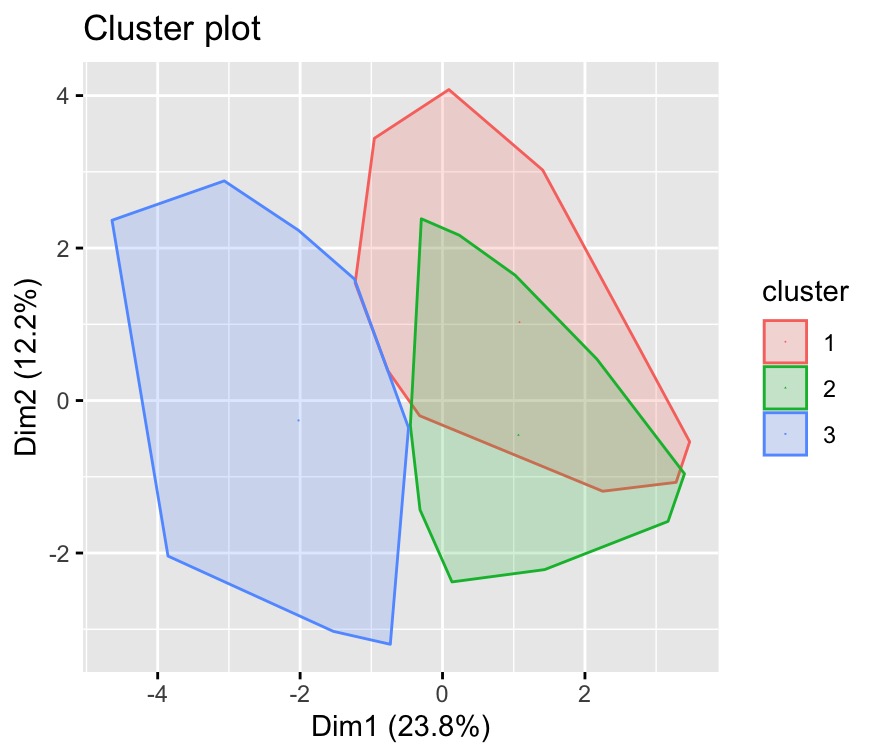
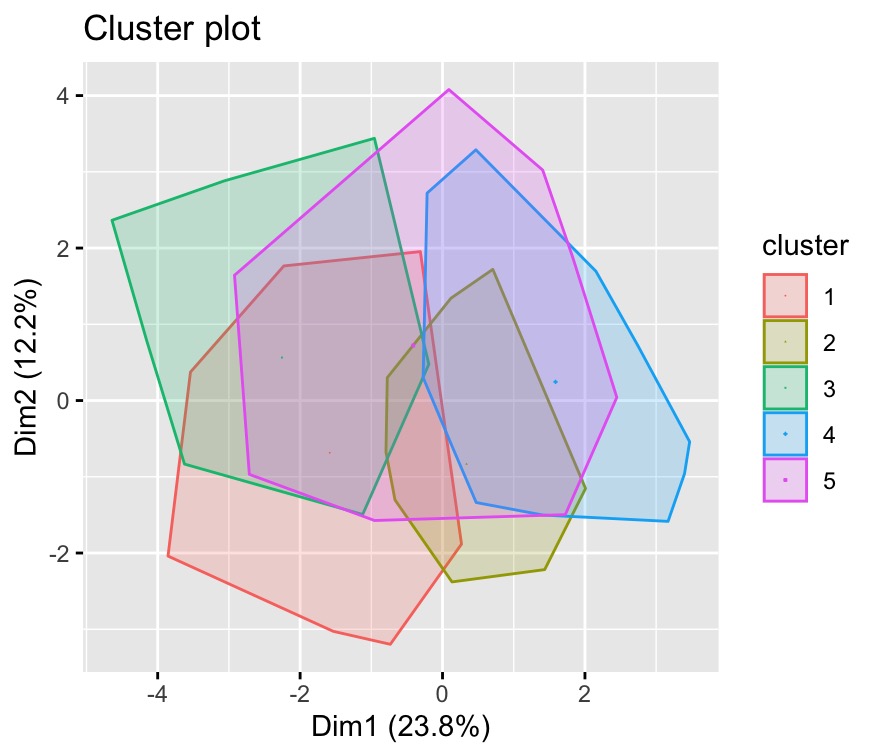
***3) How can you explain this conclusion from clustering analysis point of view? ( refer code:18,19,13)***

*Ans3) From the results of the clustering analysis i.e post running confusion matrix the value of the true positive and true negative seems to be* very prominent.



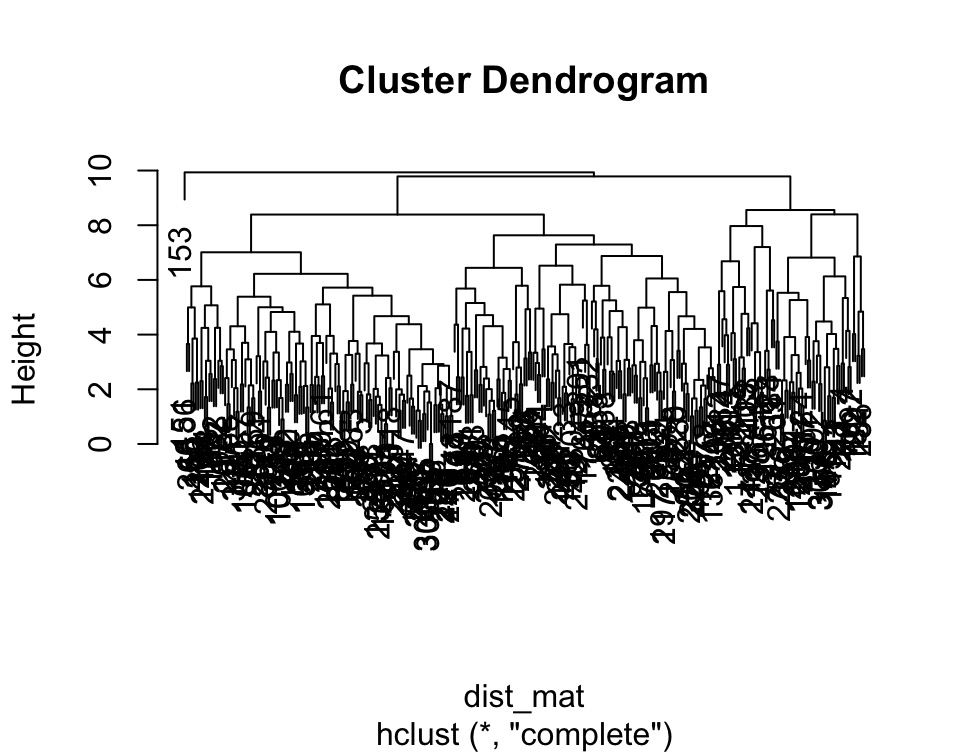
Therefore the accuracy of the K mean clustering with num cluster= 2 is 83.82838, which quite a significant value for a model to be considered as suitable for this data set.Furthermore as per below graph ,when cluster visualization was performed using “fviz\_cluster” function it is been observed that the 2 cluster are quite distinct,and for any given data point in this dimension plane, it is accurately associating with its corresponding cluster, also it doesn't overlap which infer that our model has categorized data point effectively, which is not observed in figure2 ,cluster num=3 and Figure3 that is cluster num=5.

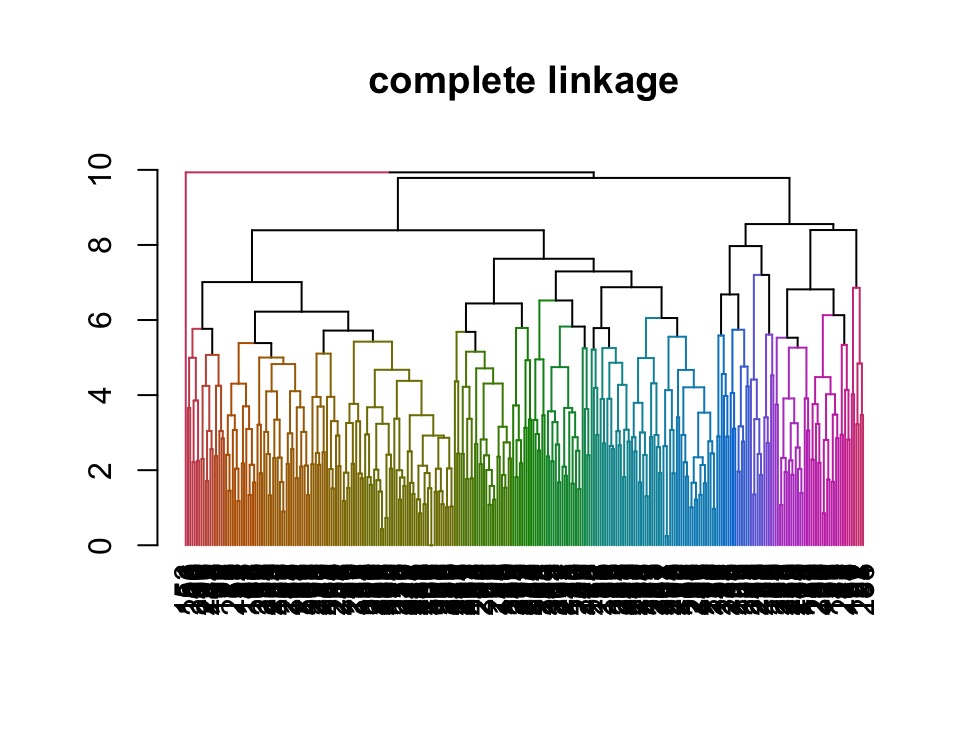




**2. Run the hierarchical clustering on above heart disease dataset, and answer the following questions**

1) Show the clustering results in tree structure;





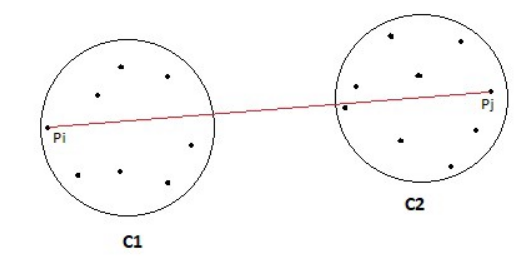
1. Describe the link method you used;

Hierarchical Clustering is an unsupervised clustering algorithm that involves constructing clusters which have a predominant ordering from top-to-bottom. The endpoint is a set of clusters, each of which is distinct from the others while the artefacts within each cluster are similar.

Linkage method: Complete linkage

One of the agglomerative hierarchical clustering approaches is complete-linkage clustering. Each element is in its own cluster at the start of the process. The clusters are then grouped into larger clusters one by one until all of the components are in the same cluster.  
In complete linkage the distance between two clusters is defined as the longest distance between two points in each cluster.

As shown in below picture, we pick the two farthest points such that one point lies in cluster one and the other point lies in cluster 2 and takes their similarity and declares it as the similarity between two clusters.



In Complete linkage two clusters will be considered close only if all points in their union are relatively similar. This method creates compact clusters with small diameters.

3) What are the strengths and limitations of this link method in hierarchical clustering?

Strengths and Limitations on Complete linkage:

Strengths:

* Complete linkage approach does well in separating clusters if there is noise between clusters.
* Complete linkage clustering eliminates a downside of single linkage process, which is "chaining". Chaining is phenomena where clusters created by single linkage clustering may be pushed together due to single elements being near to each other, despite the fact that several of the elements in each cluster may be quite far apart.

Limitations:

* Complete linkage is sensitivity to outliers. A single document located far from the centre might significantly increase the diameters of candidate merge clusters, completely altering the final clustering.