**CS 5900/STAT 46700 Topics in Data Science Spring 2025**

**Homework 4- Solution**

**Q.N. 1)** The data on the total spending by customers and their ages is provided below

ID: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T

Age: 18, 21, 22, 24, 26, 26, 27, 30, 31, 35, 39, 40, 41, 42, 44, 46, 47, 48, 49, 54

Spend: 10, 11, 22, 15, 12, 13, 14, 33, 39, 37, 44, 27, 29, 20, 28, 21, 30, 31, 23, 24

Perform the cluster analysis using the K-means clustering and Cluster Dendrogram and identify the members in three clusters.

*Solution: First we create the data frame df using R code below. Note that we don’t want to have column names for ID so that names appear in the clusters.*

> ID=LETTERS[1:20]

> Age=c(18, 21, 22, 24, 26, 26, 27, 30, 31, 35, 39, 40, 41, 42, 44, 46, 47, 48, 49, 54)

> Spend=c(10, 11, 22, 15, 12, 13, 14, 33, 39, 37, 44, 27, 29, 20, 28, 21, 30, 31, 23, 24)

> df=data.frame(row.names=ID,Age,Spend)

> head(df)

Age Spend

A 18 10

B 21 11

C 22 22

D 24 15

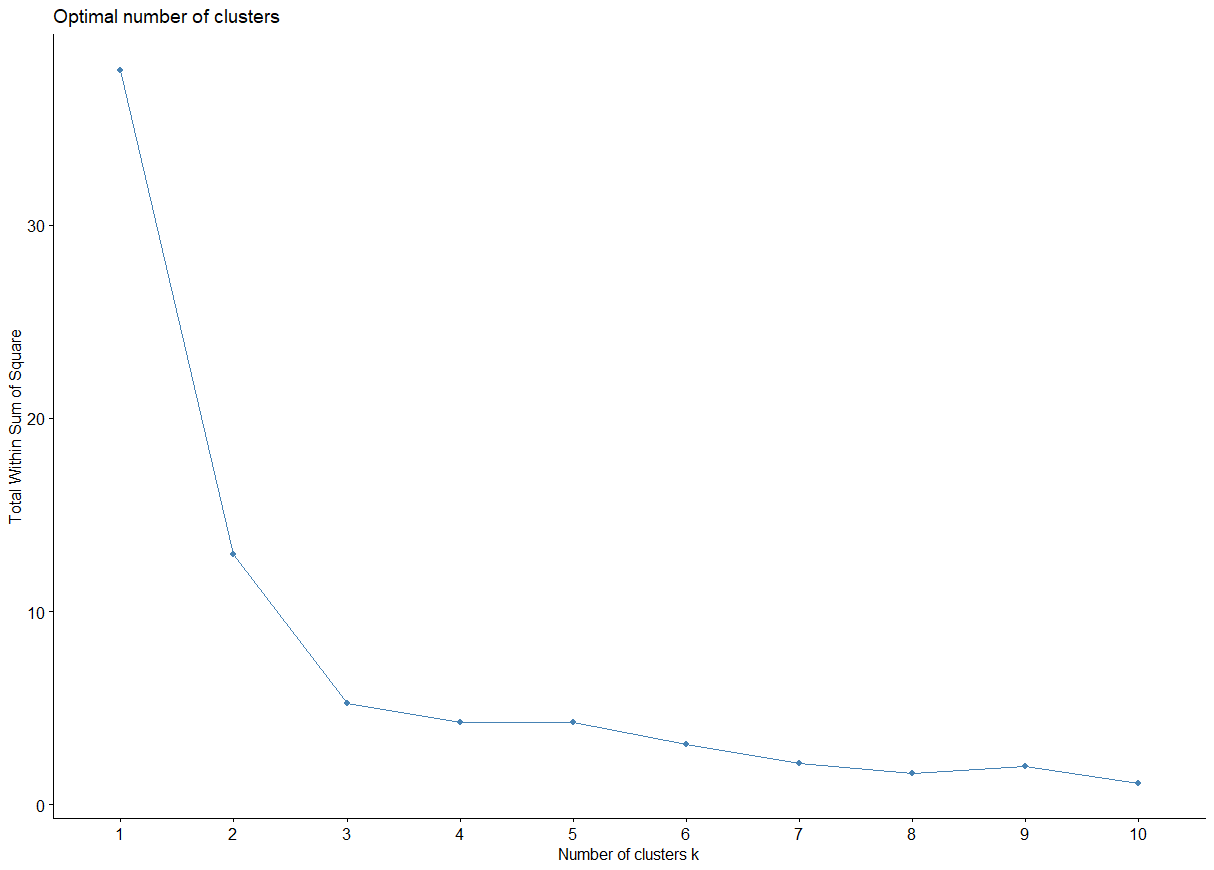
E 26 12

F 26 13

> dn=scale(df)

> library(factoextra)

> fviz\_nbclust(dn,kmeans, method="wss")

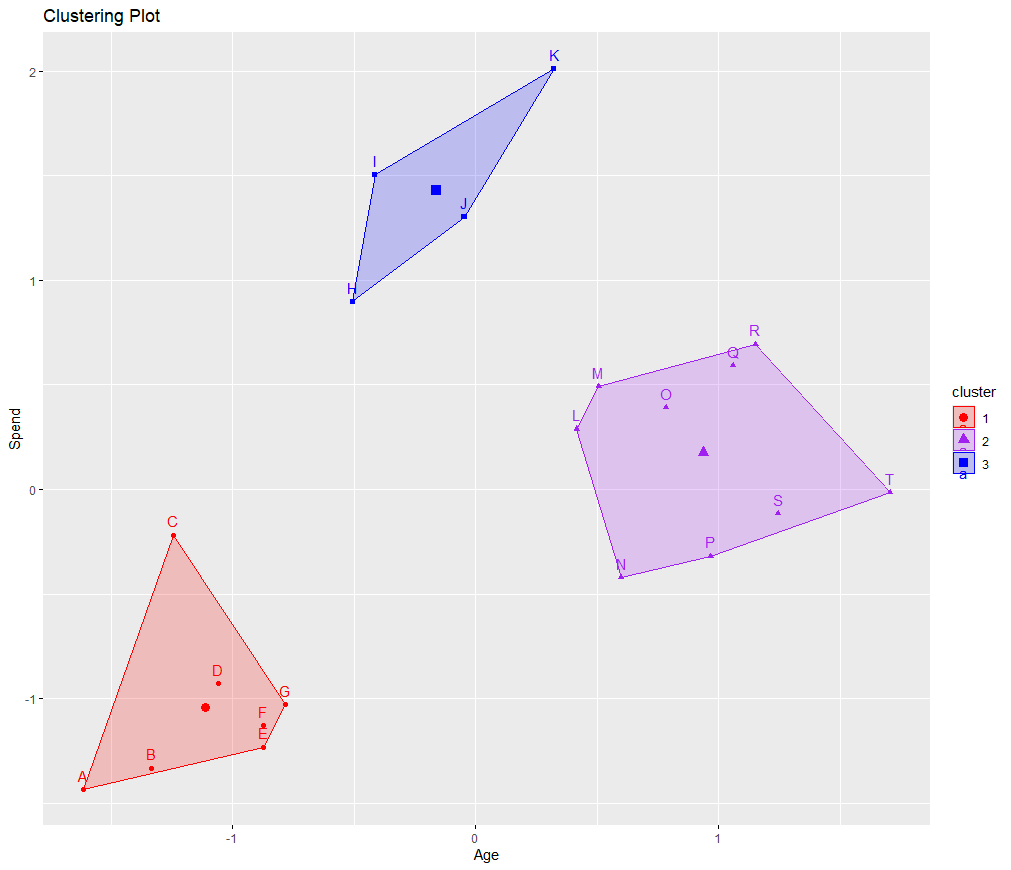


*We will create three clusters*

|  |
| --- |
| > result<- kmeans(dn,3)  > result$centers  Age Spend  1 -1.1130040 -1.0429472  2 0.9373820 0.1761969  3 -0.1613526 1.4287147  > result$size  [1] 7 9 4  > result$cluster  A B C D E F G H I J K L M N O P Q R S T  1 1 1 1 1 1 1 3 3 3 3 2 2 2 2 2 2 2 2 2 |
| *There are 7 individuals in cluster 1, 9 in cluster 2 and 4 in cluster 3.* |
| |  | | --- | |  | |

> library(ggplot2)

> fviz\_cluster(result, data = df, palette =c("red","purple", "blue"), main = "Clustering Plot")

****

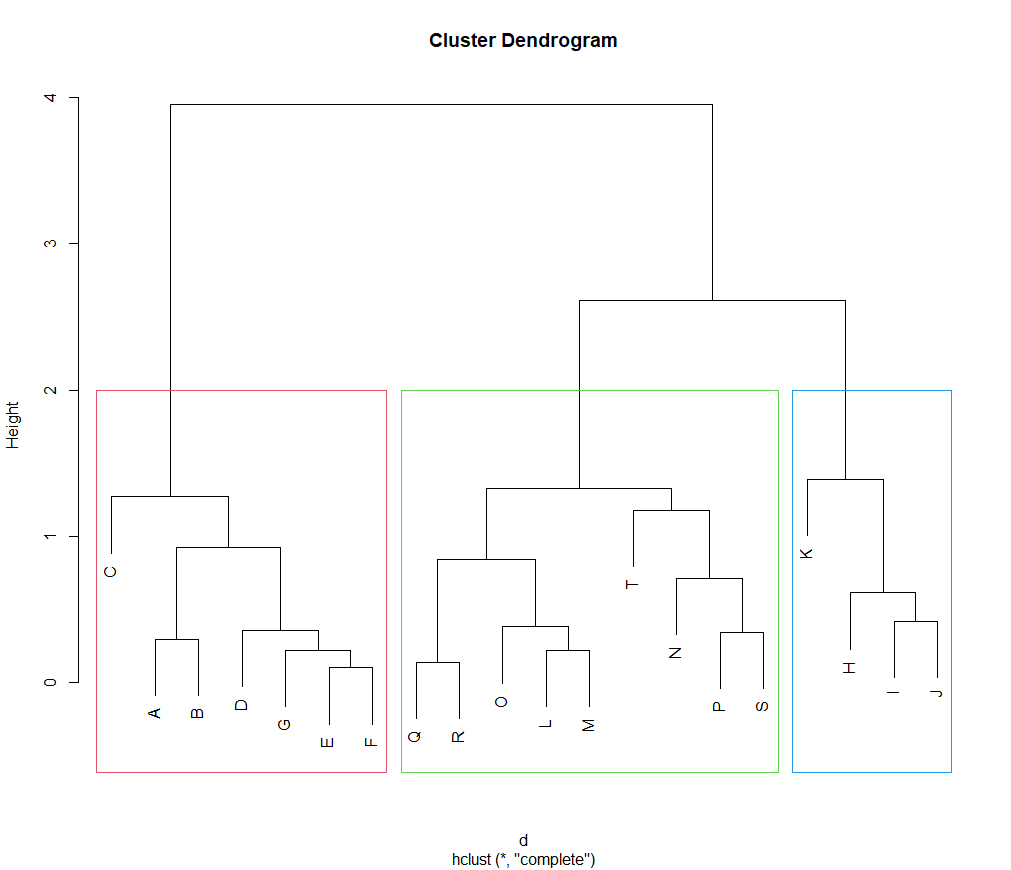
*We used R code below to perform the hierarchical clustering*

> d=dist(dn)

> hc=hclust(d)

> plot(hc)

> rect.hclust(hc,k=3,border = 2:6)

****

*Further we can draw a colored dendrogram using R code below*

> library(dendextend)

> dend=as.dendrogram(hc)

> dend=color\_branches(dend, k=3)

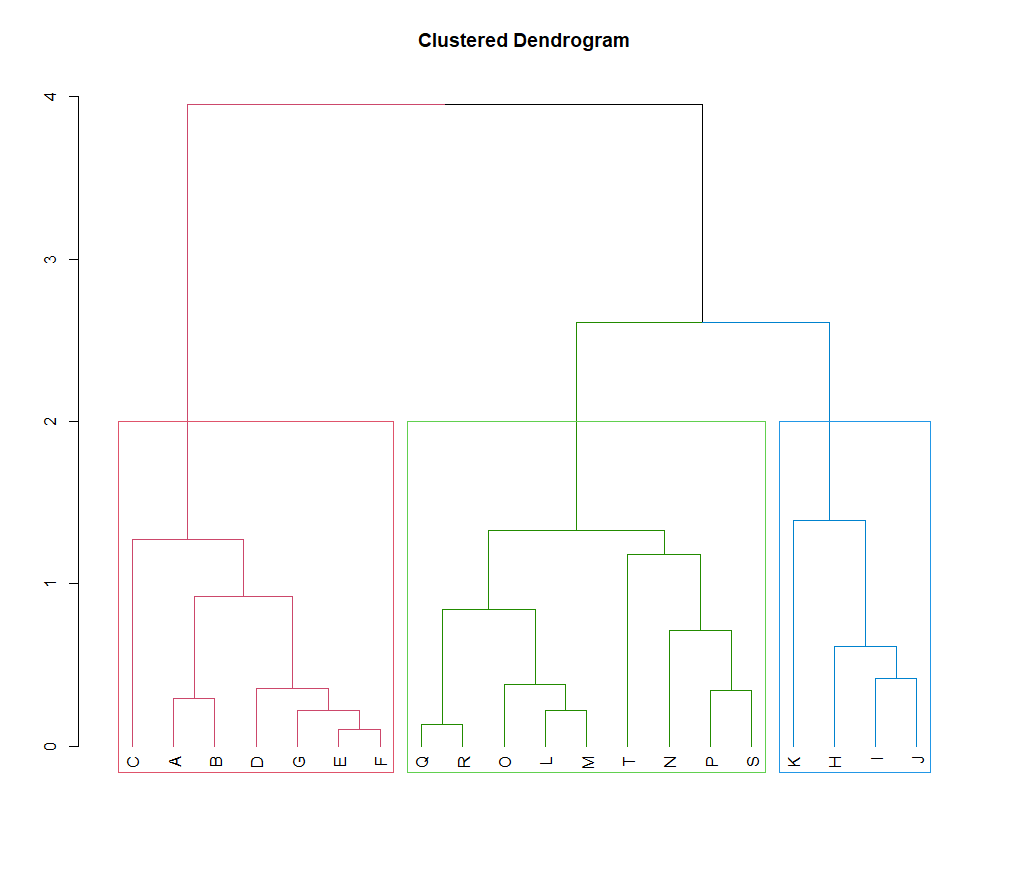
> plot(dend, main="Clustered Dendrogram")

> rect.hclust(hc,k=3,border = 2:6)

> cutree(hc, k=3)

A B C D E F G H I J K L M N O P Q R S T

1 1 1 1 1 1 1 2 2 2 2 3 3 3 3 3 3 3 3 3

****

**Q.N. 2)** The dataset USArrests in the base package contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

**a)** Calculate the numerical summary of the variables provide in the dataset.

**b)** Use K-means clustering methods to determine the clusters of the states with similar rate of Murder, Assault, Urban Population, and Rape. Justify the number of clusters that you have chosen.

**c)** Use hierarchical clustering to find the clusters and display the findings using a dendrogram.

Solution:

1. We used R code below to calculate the summary statistics of each variables.

> data(USArrests)

> dim(USArrests)

[1] 50 4

> summary(USArrests)

Murder Assault UrbanPop Rape

Min. : 0.800 Min. : 45.0 Min. :32.00 Min. : 7.30

1st Qu.: 4.075 1st Qu.:109.0 1st Qu.:54.50 1st Qu.:15.07

Median : 7.250 Median :159.0 Median :66.00 Median :20.10

Mean : 7.788 Mean :170.8 Mean :65.54 Mean :21.23

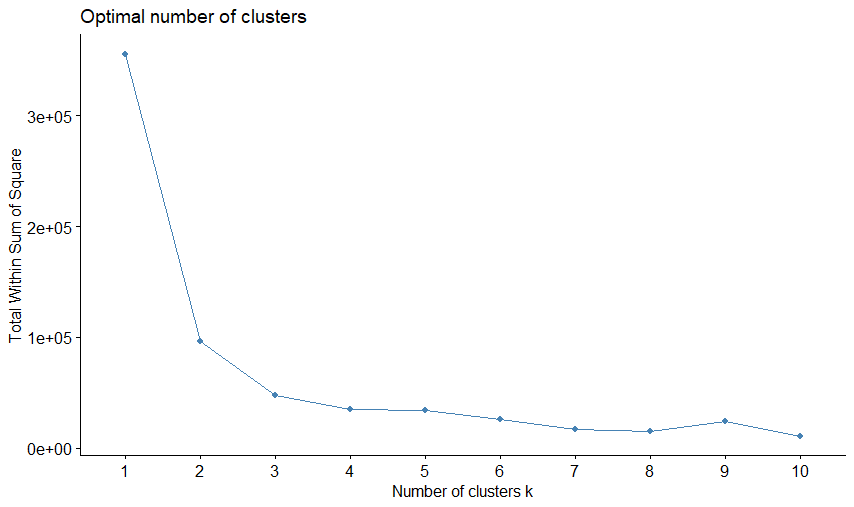
3rd Qu.:11.250 3rd Qu.:249.0 3rd Qu.:77.75 3rd Qu.:26.18

Max. :17.400 Max. :337.0 Max. :91.00 Max. :46.00

1. *Based R code below, we can create 4 clusters.*

*>library(factoextra)*

*> fviz\_nbclust(USArrests,kmeans,method = "wss")*

**

> result = kmeans(USArrests,4)

> result$size

[1] 19 7 14 10



> Cluster1<-subset(result$cluster,result$cluster==1)

> Cluster2<-subset(result$cluster,result$cluster==2)

> Cluster3<-subset(result$cluster,result$cluster==3)

> Cluster4<-subset(result$cluster,result$cluster==4)

> Cluster1

Connecticut Idaho Indiana Kansas Kentucky Massachusetts Missouri Montana Nebraska New Jersey Ohio Oklahoma Oregon Pennsylvania Rhode Island

Utah Virginia Washington Wyoming

> Cluster2

Arizona California Florida Maryland New Mexico

North Carolina South Carolina

> Cluster3

Alabama Alaska Arkansas Colorado Delaware Georgia Illinois Louisiana Michigan Mississippi Nevada New York

Tennessee Texas

> Cluster4

Hawaii Iowa Maine Minnesota New Hampshire

North Dakota South Dakota Vermont West Virginia Wisconsin

> library(factoextra)

> library(ggplot2)

> fviz\_cluster(result, data = USArrests, palette =c("#00AFBB","#2E9FDF", "#E7B800", "#FC4E07"), main = "Clustering Plot", xlab="", ylab="")

Chart, radar chart

Description automatically generated

If we normalize the data we will get a different classification as below

> data(USArrests)

> attach(USArrests)

> result = kmeans(scale(USArrests),4) # see the scale function

> result$size

[1] 13 8 13 16

> library(factoextra)

> library(ggplot2)

> fviz\_cluster(result, data = USArrests, palette =c("#00AFBB","#2E9FDF", "#E7B800", "#FC4E07"), main = "Clustering Plot", xlab="", ylab="")

Chart, radar chart

Description automatically generated

*b)We use R code below to perform the hierarchical clustering*

arrestdata <- dist(USArrests)

hc <- hclust(arrestdata)

plot(hc)

library(dendextend)

dendrogram=as.dendrogram(hc)

dendrogram=color\_branches(dendrogram, k=4)

plot(dendrogram, main="Clustered states”)



If we used the normalized data with scale function than the dendrogram will be as below

arrestdata <- dist(scale(USArrests))

hc <- hclust(arrestdata)

plot(hc)

library(dendextend)

dendrogram=as.dendrogram(hc)

dendrogram=color\_branches(dendrogram, k=4)

plot(dendrogram, main="Clustered states")

Chart

Description automatically generated

**Q.N. 3)** The penguins dataset included in the palmerpenguins package provides the size measurements for adult foraging penguins near Palmer Station, Antarctica.

a) Access the data and determine its dimension.

b) Omit the missing values from the dataset

c) Choose 50 observations at random from the clean dataset. Please make sure that your results are reproduceable so choose a set.seed () value.

d) Extract the numerical variables and standardize them.

e) Perform the cluster analysis using the K-means clustering and Cluster Dendrogram and identify the members in the clusters.

*Solution: a) Based on R output below there are 344 observations with 8 variables*

|  |
| --- |
| > library(palmerpenguins)  > dim(penguins)  [1] 344 8 |

*b)We have used R code below to delete the missing values and it has been determined that there are 333 observations with complete information.*

> Q3=na.omit(penguins)

> dim(Q3)

[1] 333 8

1. *We can use R code below to choose 50 observations at random*

> set.seed(100)

> index=sample(1:nrow(Q3),50)

> samplepen=Q3[index,]

> head(samplepen)

# A tibble: 6 × 8

species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g sex year

*<fct>* *<fct>* *<dbl>* *<dbl>* *<int>* *<int>* *<fct>* *<int>*

1 Gentoo Biscoe 43.8 13.9 208 4300 fema… 2008

2 Adelie Torge… 37.3 20.5 199 3775 male 2009

3 Gentoo Biscoe 45.3 13.8 208 4200 fema… 2008

4 Adelie Torge… 36.7 19.3 193 3450 fema… 2007

5 Chinst… Dream 50.8 18.5 201 4450 male 2009

6 Chinst… Dream 51.9 19.5 206 3950 male 2009

> table(samplepen$species)

Adelie Chinstrap Gentoo

20 12 18

*It appears that the sample data has 20 Adelie, 12 Chinstrap and 18 Gentoo species***.**

1. *We used R code to select the numerical variables and standardized them*

> samplepen\_num=samplepen[,c(3,4,5,6)]

> scaledata=scale(samplepen\_num)

> head(scaledata)

bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g

[1,] -0.0485991 -1.4776644 0.47555162 0.10833483

[2,] -1.1531241 1.4352941 -0.19952166 -0.57691559

[3,] 0.2062913 -1.5218002 0.47555162 -0.02218906

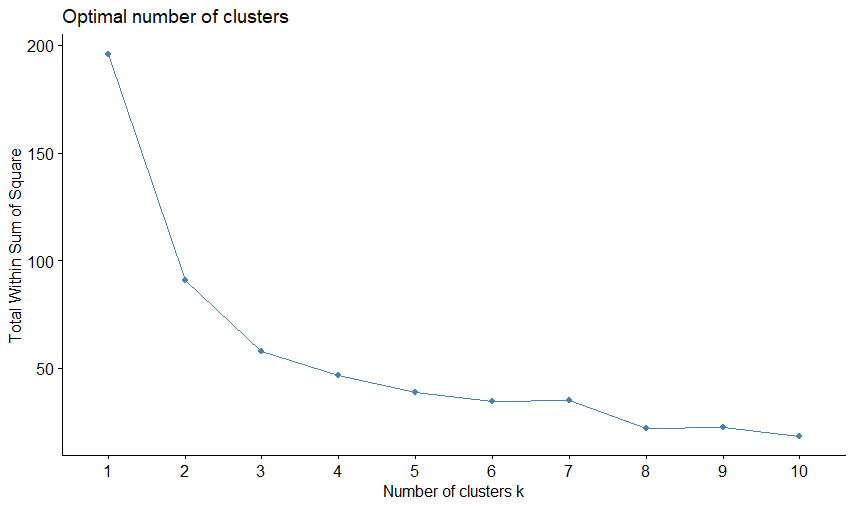
[4,] -1.2550803 0.9056653 -0.64957051 -1.00111824

[5,] 1.1408894 0.5525794 -0.04950537 0.30412066

[6,] 1.3278090 0.9939368 0.32553534 -0.34849879

1. In order to perform the cluster analysis, first we determine the value of k. Based on the R output we will choose k=3

|  |
| --- |
| > library(factoextra)  > fviz\_nbclust(scaledata,kmeans,method = "wss") |



> result=kmeans(scaledata, 3)

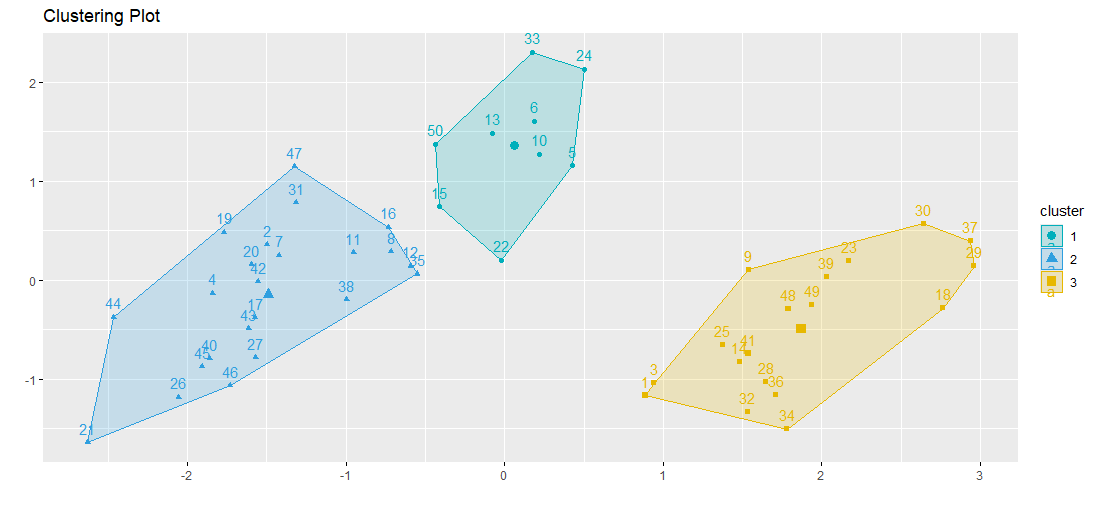
> result$size

[1] 9 23 18

> result$cluster

[1] 3 2 3 2 1 1 2 2 3 1 2 2 1 3 1 2 2 3 2 2 2 1 3 1 3 2 2 3 3 3

[31] 2 3 1 3 2 3 3 2 3 2 3 2 2 2 2 2 2 3 3 1



|  |
| --- |
| > dist=dist(scaledata)  > hc <- hclust(dist)  > plot(hc)  > library(dendextend)  > dendrogram=as.dendrogram(hc)  > dendrogram=color\_branches(dendrogram, k=3)  > plot(dendrogram, main="Clustered penguins") |
|  |
| |  | | --- | |  | |

Remark: In order to display the names we could use the R code below:

> scaledata=scale(samplepen\_num)

> row.names(scaledata)=samplepen$species

> head(scaledata)

> dist=dist(scaledata)

> hc <- hclust(dist)

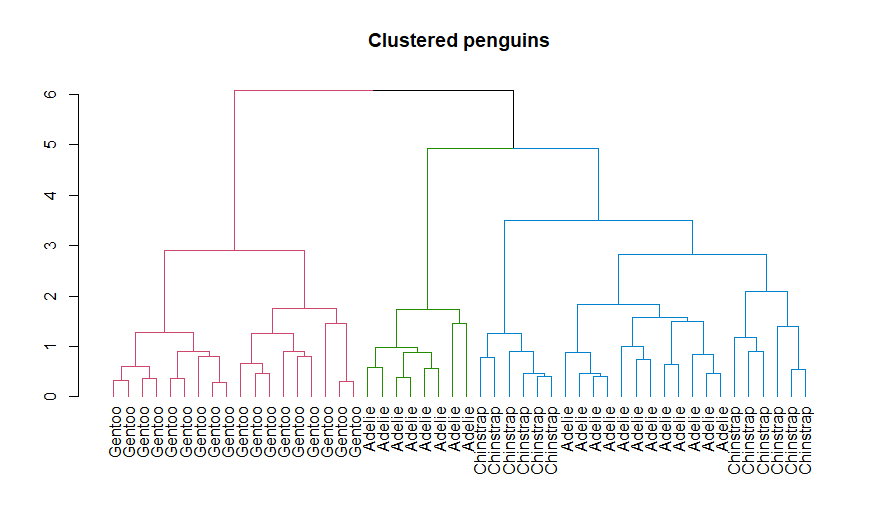
> plot(hc)

> library(dendextend)

> dendrogram=as.dendrogram(hc)

> dendrogram=color\_branches(dendrogram, k=3)

> plot(dendrogram, main="Clustered penguins")



**Q.N. 4)** The diabetes data set provided with this assignment consists of 520 observations on 17 features. The target variable is the Class variable which can take on two values, Positive and Negative. All but one of the features are binary. The nonbinary feature is the Age feature.

a) Import the dataset in R and print the variable names.

b) Split the data set in training and test set with 70% in training and 30% in test set.

c) Creating a Decision Tree Model using the training data.

d) Predict the response on the test data and produce a confusion matrix comparing the test labels to the predicted labels. What is the Accuracy rate?

*Solution:*

1. *Based on the R output below, there are 520 observations with 17 variable. The variable names are provided below*

|  |
| --- |
| > Q4=read.csv("C:\\Users\\aryalgSTAT 46700\\Homework\\diabetes.csv")  > dim(Q4)  [1] 520 17  > names(Q4)  [1] "Age" "Gender" "Polyuria" "Polydipsia"  [5] "weightloss" "weakness" "Polyphagia" "Genital\_thrush"  [9] "visual\_blurring" "Itching" "Irritability" "delayed\_healing"  [13] "partial\_paresis" "muscle\_stiffness" "Alopecia" "Obesity"  [17] "class" |
|  |
| |  | | --- | |  | |

1. *We used R code below to create the training and test sets. It appears that there are 364 observations in training set and 156 observations in test set.*

> library(caret)

> set.seed(200)

> split = createDataPartition(Q4$class, p = 0.70, list=FALSE)

> train.data <- Q4[split,]

> test.data <- Q4[-split,]

> dim(train.data)

[1] 364 17

> dim(test.data)

[1] 156 17

1. We have created a decision tree model using R code below

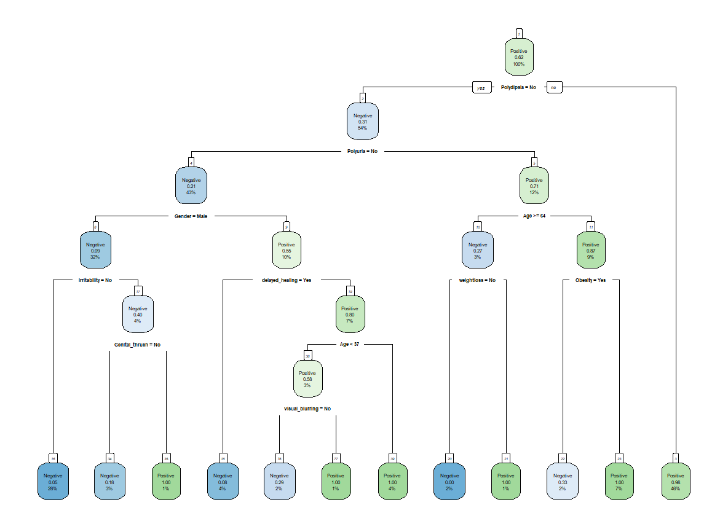
> library(rpart)

> library(rpart.plot)

> model <- rpart(class ~ ., data = train.data, method = "class",

+ control = rpart.control(minsplit = 10, cp = 0.01))

> rpart.plot(model, box.palette = "auto", nn = TRUE)



1. Based on the R output below, the accuracy rate of the model is 89.74%.

> pred <- predict(model, test.data, type = "class")

> test.class = as.factor(test.data$class)

> confusionMatrix(pred,test.class,positive="Positive")

Confusion Matrix and Statistics

Reference

Prediction Negative Positive

Negative 56 12

Positive 4 84

Accuracy : 0.8974

95% CI : (0.8388, 0.9402)

No Information Rate : 0.6154

P-Value [Acc > NIR] : 2.24e-15

Kappa : 0.7886

Mcnemar's Test P-Value : 0.08012

Sensitivity : 0.8750

Specificity : 0.9333

Pos Pred Value : 0.9545

Neg Pred Value : 0.8235

Prevalence : 0.6154

Detection Rate : 0.5385

Detection Prevalence : 0.5641

Balanced Accuracy : 0.9042

'Positive' Class : Positive