ASSIGNMENT 02

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

Date: Nov 22, 2019 Group Mates: Hamid, Mateen & Saadoon

Note: *R Commands are provided at the end of each respective answer*

Question No 01:

The data in table 3.6 show the nutritional content of different foodstuffs (the quantity involved is always three ounces). Use R to create a scatterplot matrix of the data labelling the foodstuffs appropriately in each panel. On the basis of this diagram undertake what you think is an appropriate principal component analysis and try to interpret your results.

Solution:

Scatterplot matrix for the dataset is as below (on next page), we can observe that:

- Energy and Fat are strongly correlated and may turn up as a single component
- Calcium seems negatively corelated with others and may turn up a separate component

We further apply Principal Component Analysis in order to:

- Reduce the number of variables
- To have uncorrelated variables

Note: As the provided data is already in a standardized scale and available in units of ounces, we didn't apply any normalization technique.

Correlation Matrix for the data is follows:

	Energy	Protein	Fat	Calcium	Iron
Energy	1	0.17	0.99	-0.32	-0.10
Protein	0.17	1	0.02	-0.09	-0.17
Fat	0.99	0.02	1	-0.31	-0.06
Calcium	-0.32	-0.09	-0.31	1	0.04
Iron	-0.10	-0.17	-0.06	0.04	1

We can see that Energy and Fat have high covariance (0.99). From which we can expect that they may compose a component in PCA.

Note: In our PCA, we have not fixed any attribute and identified components to predict that. We execute PCA for all attributes to identify uncorrelated variables and reduce our variables.

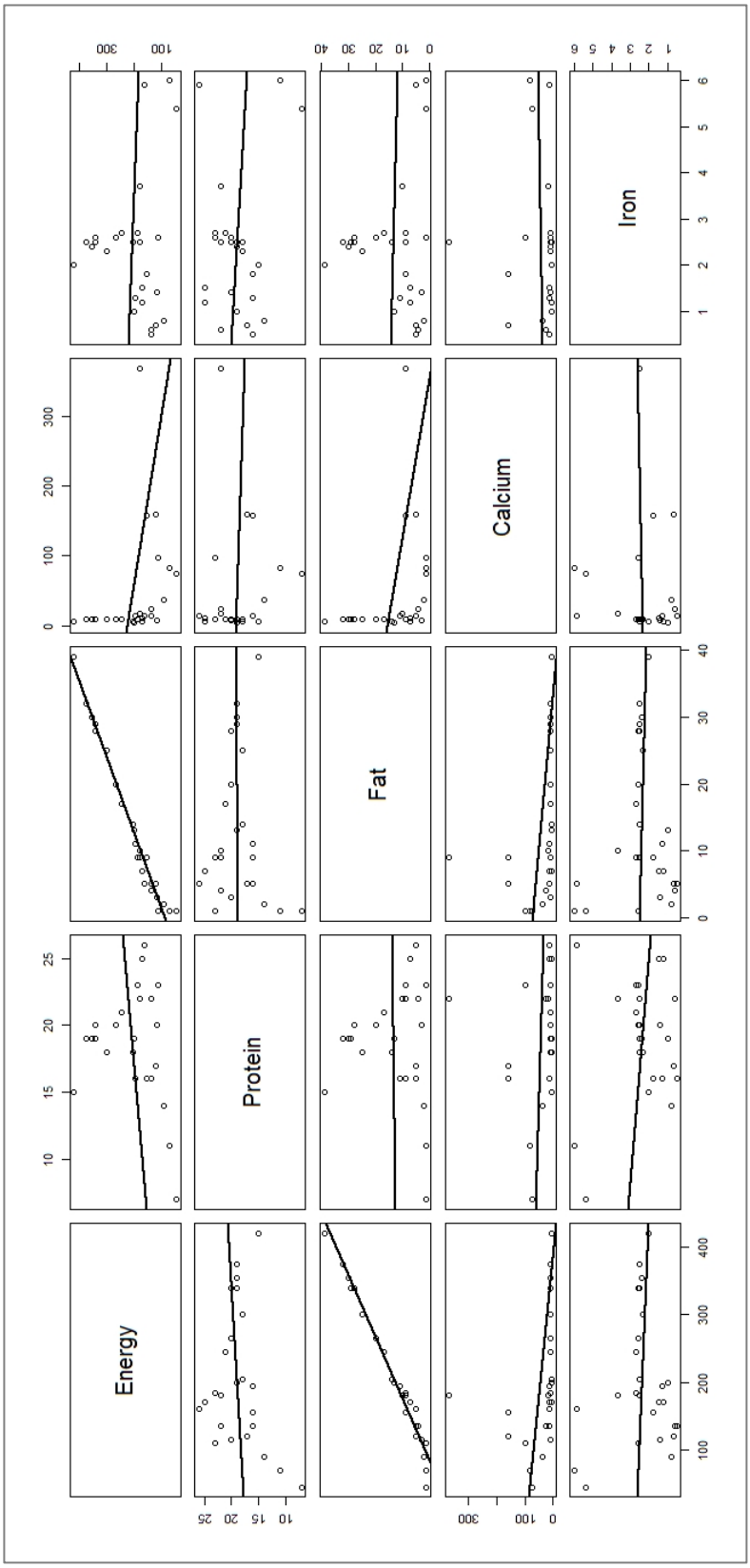


Figure 1 - Scatterplot Matrix of Data

Principal Component Analysis:

	Comp1	Comp2	Comp3	Comp4	Comp5
Standard Deviation	1.48	1.07	0.92	0.90	0.04
Proportion of Variance	0.44	0.23	0.17	0.16	0.00
Cumulative Proportion	0.44	0.67	0.84	1	1

General practice is to consider components which captures around 80% of the cumulative proportion of all variables. In our case first three components are covering 84% hence we are considering Comp1, Comp2 and Comp3 as our Principal Components.

LOADINGS	Comp1	Comp2	Comp3	Comp4	Comp5
Energy	0.65		0.15	0.20	0.71
Protein	0.15	-0.69	-0.46	0.53	-0.10
Fat	0.64	0.20	0.22	0.13	-0.70
Calcium	-0.36		0.65	0.67	
Iron	-0.12	0.69	-0.54	0.47	

From Loadings, we can observe that Comp1 is capturing Energy and Fat which supports our identification made from Scatterplot. Comp2 captures Protein and Iron whereas Comp3 captures Calcium. Null values show that sum of square was nearly zero.

Labeling: We would like to give our Components appropriate labels, and will say as:

Component 1 -> E-Fat & **Component 2**-> P-Iron & **Component 3**-> Calcium

From Components, we can get values for all observations. Here we have shown first five:

PC Score	Comp1	Comp2	Comp3
BB Beef, braised	1.89	0.33	-0.01
HR Hamburger	0.66	-0.08	-0.52
BR Beef, roast	2.94	1.14	1.09
BS Beef, steak	2.33	0.55	0.27
BC Beef, canned	-0.26	0.05	-1.17

Interpretation of Results:

Principal Component Analysis have deduced for us three components which captures 84% of the overall variance of the dataset and will allow us to continue further analysis on these three components. Loadings show us that data is appropriately bifurcated into components capturing variance of all uncorrelated elements. Some additional outputs are provided below:

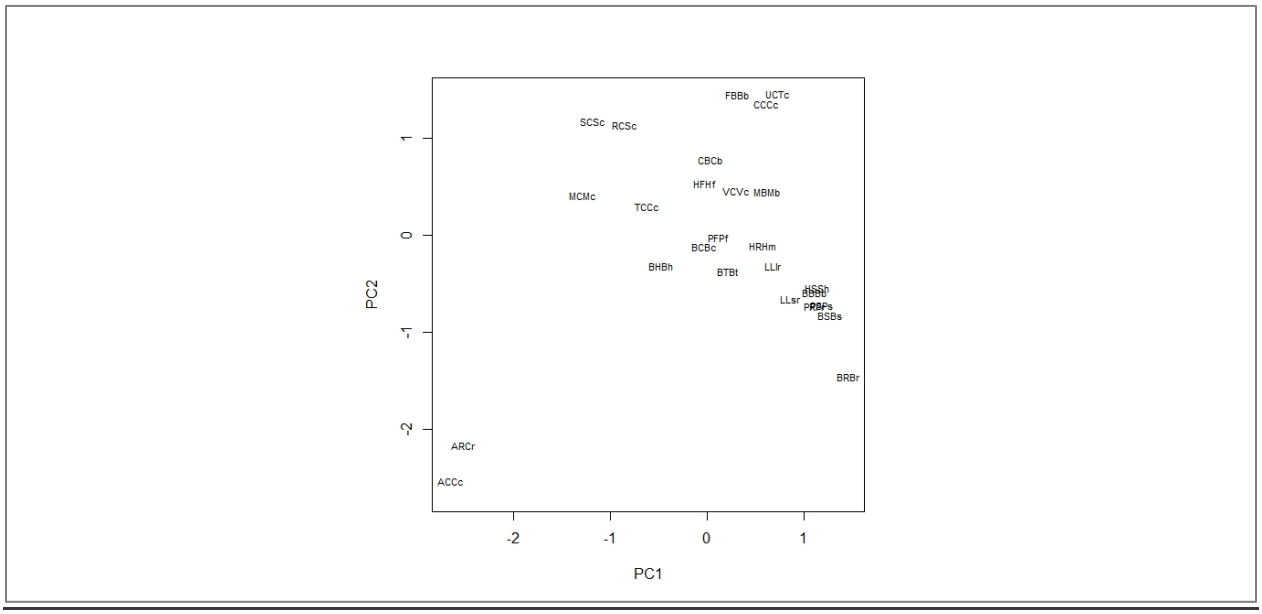


Figure 2 - PC1 plot vs PC2

Remarks: We can identify possible outliers. We can see components are not correlated in any manner.

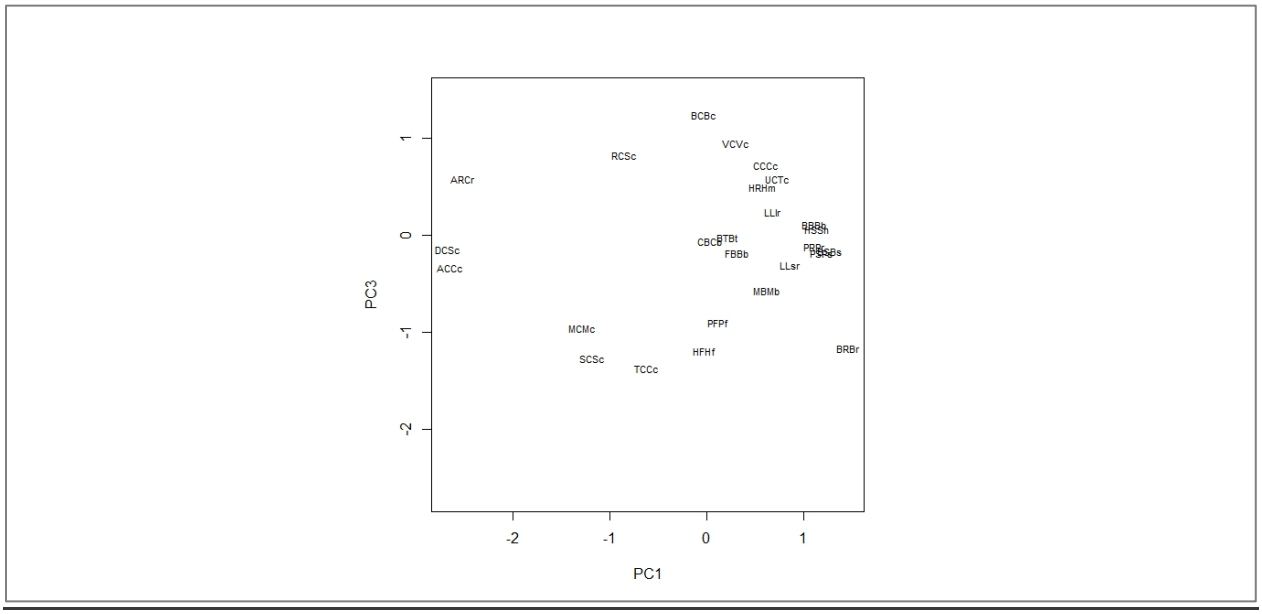


Figure 3 - PC1 plot vs PC3

Remarks: Uncorrelation is visible from plot as data is scattered with no linear pattern.

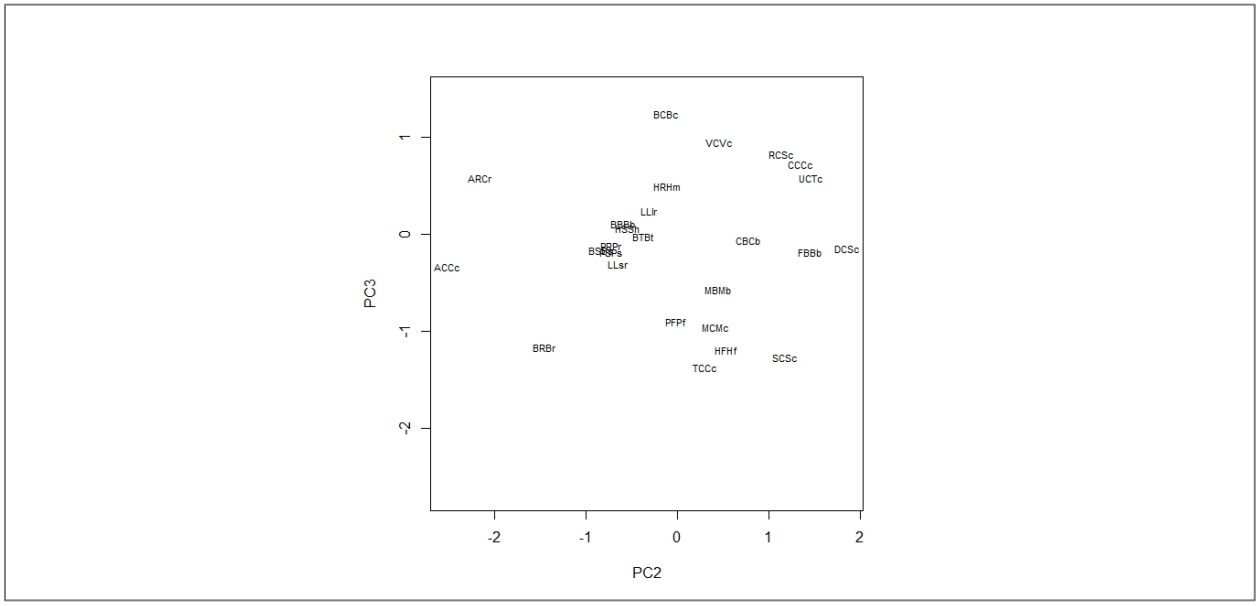


Figure 4 - PC2 plot vs PC3

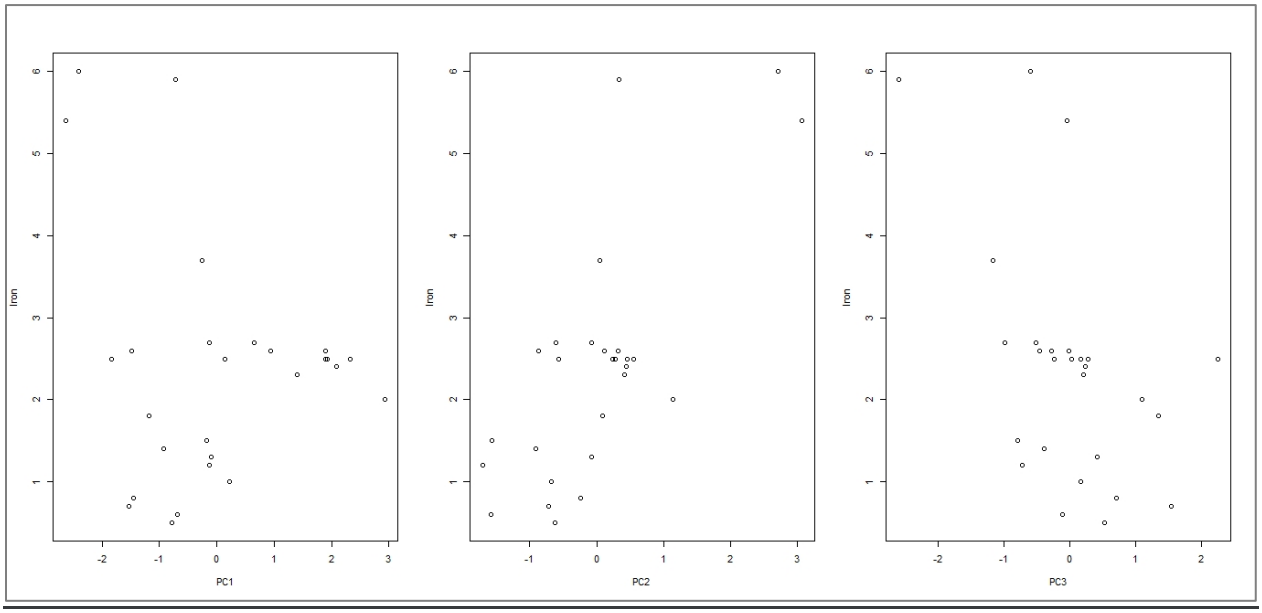


Figure 5 - Iron's plot against all Principal Components

Remarks: Below provided loadings are visible in above plot as Comp2 is corelated with Iron in capturing its variance the most.

LOADINGS	Comp1	Comp2	Comp3
Iron	-0.12	0.69	-0.54

R Commands (Qs 01)

```
foodstuff<-source("F:\\UJI\\Courses\\Applied Mathematics\\Assignment 2\\foodstuffs.dat")$value

//calculates scatterplot

pairs(foodstuff)

pairs(foodstuff,panel=function(x,y) {abline(lsfit(x,y)$coef,lwd=2)

                        points(x,y)})

cor(foodstuffs.dat[,])

//calculates the Principal components

foodstuff.pc<-princomp(foodstuffs,cor=TRUE)

summary(foodstuff.pc,loadings=TRUE)

// outputs the component score

foodstuff.pc$scores[,1:3]


par(pty="s")

plot(foodstuff.pc$scores[,1],foodstuff.pc$scores[,2],

ylim=range(foodstuff.pc$scores[,1]),

xlab="PC1",ylab="PC2",type="n",lwd=2)

text(foodstuff.pc$scores[,1],foodstuff.pc$scores[,2],

labels=abbreviate(row.names(foodstuff)),cex=0.7,lwd=2)



par(pty="s")
```

```
plot(foodstuff.pc$scores[,1],foodstuff.pc$scores[,3],
ylim=range(foodstuff.pc$scores[,1]),
xlab="PC1",ylab="PC3",type="n",lwd=2)
text(foodstuff.pc$scores[,1],foodstuff.pc$scores[,3],
labels=abbreviate(row.names(foodstuff)),cex=0.7,lwd=2)


par(pty="s")
plot(foodstuff.pc$scores[,2],foodstuff.pc$scores[,3],
ylim=range(foodstuff.pc$scores[,1]),
xlab="PC2",ylab="PC3",type="n",lwd=2)
text(foodstuff.pc$scores[,2],foodstuff.pc$scores[,3],
labels=abbreviate(row.names(foodstuff)),cex=0.7,lwd=2)


par(mfrow=c(1,3))
plot(foodstuff.pc$scores[,1],Iron,xlab="PC1")
plot(foodstuff.pc$scores[,2],Iron,xlab="PC2")
plot(foodstuff.pc$scores[,3],Iron,xlab="PC3")
```

Question No 02:

Reanalyze the life expectancy data by clustering the countries on the basis of differences between the life expectancies of men and women at corresponding ages.

Considering the groups/clusters previously found, classify into these groups the following two new observations: Country1:(65,50,33,15,69,57,37,16); Country2:(59,46,31,15,64,56,33,16)

Solution:

First of all, we developed Dendrograms to observe the data observations. As we look them (on next page), we cannot clearly identify the cutoff value to produce right number of clusters on the data.

Note: Hierarchal clustering do not offer any appropriate value for cut off. We use K-means clustering to find appropriate value for cutoff and then go back to Hierarchal clustering to obtain specified number of clusters, K.

Just in order to view clusters based on intuition, we considered value of 'h' as 4 and constructed below 08 clusters.

Cluster #	Observations (Total 31)
1	Algeria, Cameroon, Mauritius, South Africa(C) El Salvador, Greenland, Grenada, Jamaica, Trinidad (67)
2	Madagascar
3	Reunion
4	Seychelles
5	South Africa(W), Canada, United States (66) United States (NW66), United States (W66), United States (67), Argentina, Chile
6	Tunisia
7	Costa Rica, Dominican Rep, Guatemala, Honduras, Mexico, Nicaragua, Panama, Columbia, Ecuador
8	Trinidad(62)

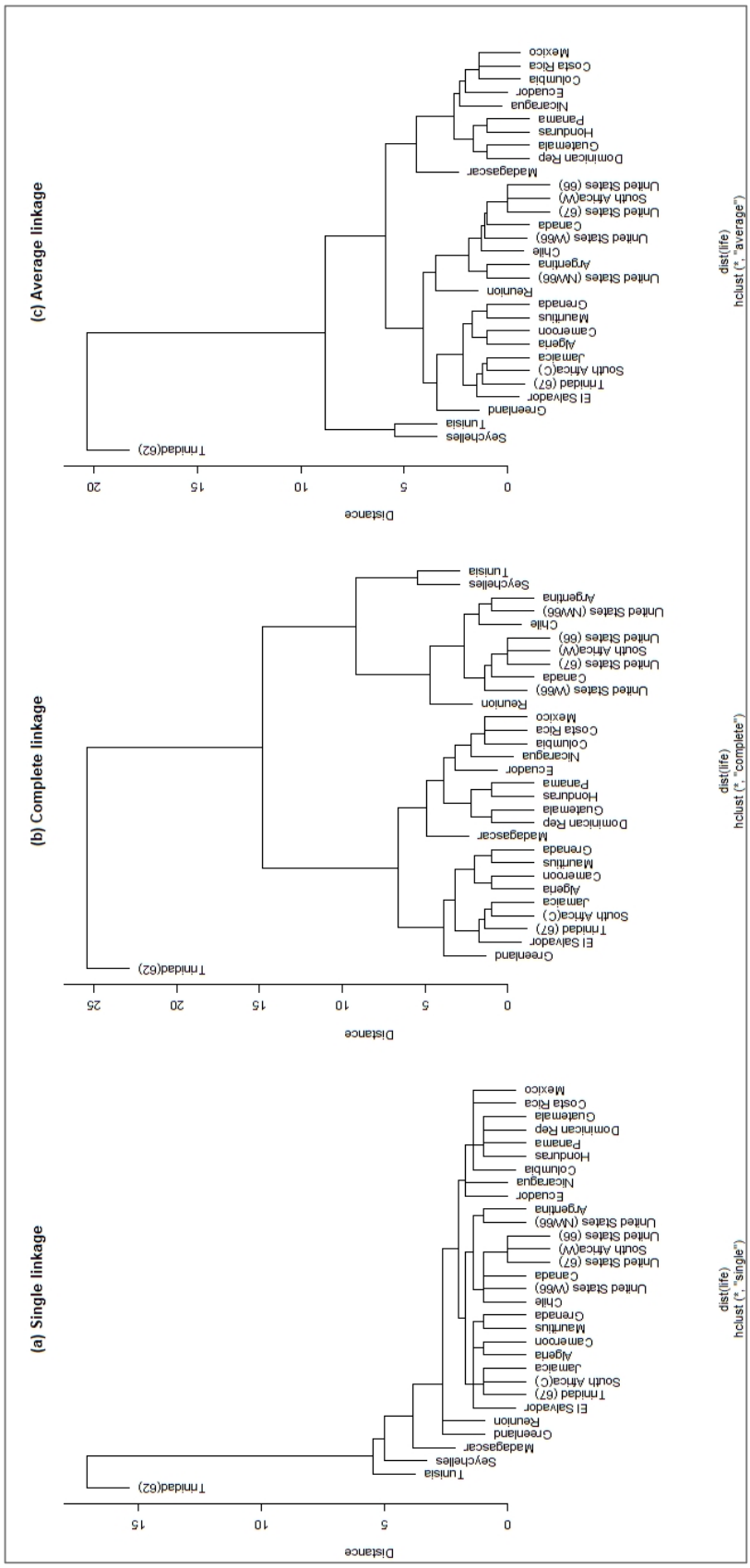


Figure 6 - Dendrogram

K-MEANS:

After applying the K-Means function, we obtained the following graph from where we can find the appropriate value for cut off:

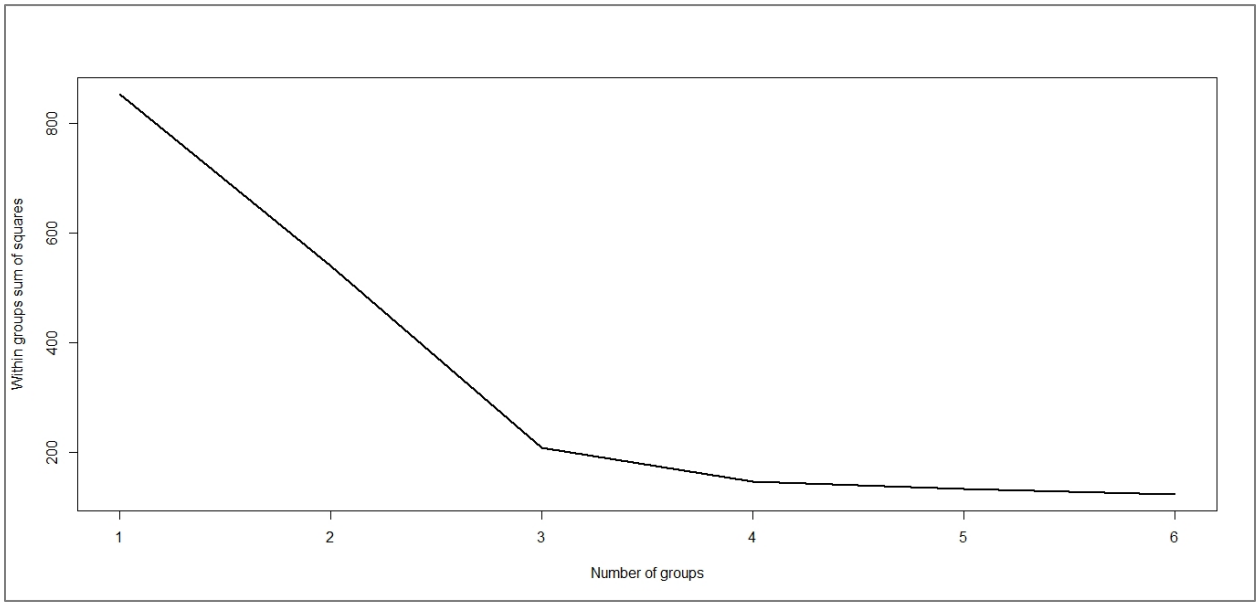


Figure 7 - K-Means

We can observe from the plot that the slope is breaking stepper at 3. So, we considered 3 number of clusters for data.

Sized of Cluster:

Cluster #	No. of Observations
1	13
2	1
3	17

Cluster #	Observations (Total 31)
1	Mauritius, Reunion, Seychelles, South Africa(W), Tunisia, Grenada, United States (66), United States (NW66), United States (W66), United States (67), Argentina, Chile
2	Trinidad(62)
3	Algeria, Cameroon, Madagascar, South Africa(C), Canada, Costa Rica, Dominican Rep, El Salvador, Greenland, Guatemala, Honduras, Jamaica, Mexico, Nicaragua, Panama, Trinidad (67), Columbia, Ecuador

CLUSTER MEANS	mo	m25	m50	m75
1	-6.31	-5.92	-5.23	-2.77
2	-4.00	-16.00	-4.00	-2.00
3	-3.29	-2.59	-1.88	-0.94

New Samples:

Now we have the following observations on two new Countries,

New Observation	mo	m25	m50	m75	f0	f25	f50	f75
Country-1	65	50	33	15	69	57	37	16
Country-2	59	46	31	15	64	56	33	16

We have taken the difference at corresponding ages between male and female gender for Country 1 and Country-2 and obtained following result:

DIFFERENCE VALUES	Mo-f0	m25-f25	m50-f50	m75-f75
Country-1	-4	-7	-4	-1
Country-2	-5	-10	-2	-1

We have applied Fisher's linear discriminant analysis by using the "MASS" library to assign these new observations into existing clusters. We get the following classification probabilities for each cluster:

DISCRIMINANT ANALYSIS	Type-I	Type-II	Type-III
Country-1	0.95	0.00	0.05
Country-2	1.00	0.00	0.00

Interpretation of Results:

We can see that both samples have high probability to be of Type1, hence we will consider the same.

ASSIGNMENT 02

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

R Commands (Qs 02)

```
life<-source("F:\\UJI\\Courses\\Applied Mathematics\\Topic 1\\RSPCMA\\Data\\chap4lifeexp.dat")$value

attach(life)

//these commands calculate the difference of ages between genders
life[,1]= life[,1]-life[,5]
life[,2]= life[,2]-life[,6]
life[,3]= life[,3]-life[,7]
life[,4]= life[,4]-life[,8]

//this command will remove the remaining columns in data frame which are  not required for statistical analysis
life=life[,-(5:8)]

par(mfrow=c(1,3))
plclust(hclust(dist(life),method="single"),labels=row.names(life),ylab="Distance")
title("(a) Single linkage")
plclust(hclust(dist(life),method="complete"),labels=row.names(life),ylab="Distance")
title("(b) Complete linkage")
plclust(hclust(dist(life),method="average"),labels=row.names(life),ylab="Distance")
title("(c) Average linkage")

pairs(life,panel=function(x,y) {abline(lsfit(x,y)$coef,lwd=2)
                        points(x,y)})


country.clus<-lapply(1:8,function(nc) row.names(life)[cutoff==nc])
country.mean<-lapply(1:8,function(nc) apply(life[cutoff==nc,],2,mean))
country.mean
country.clus

kmean<-kmeans(life,3)
kmean
lapply(1:3,function(nc) apply(life[kmean$cluster==nc,],2,mean))
```

```
kmean$cluster

Type<-cbind(kmean$cluster)
life[,5]<-Type
colnames(life)<- c ("m0-w0","m25-w25","m50-w50","m75-w75", "Type")

library(MASS)
dis<-lda(Type~m0-w0+m0-w0+m0-w0+m0-w0,data=life,prior=c(0.33,0.33,0.34))


newdata<-rbind(c(-4,-7,-4,-1),c(-5,-10,-2,-1))
colnames(newdata)<-colnames(life[,-5])

newdata<-data.frame(newdata)
predict(dis,newdata=newdata)
```

Question No 03:

From Table 7.1 (Chapter 7 Everitt): (a) Perform a PCA and interpret the results; (b) Find homogeneous clusters amongst the skulls. Compare the results with existing groups given by variable "Type".

Solution:

Firstly, we developed a Scatterplot Matrix for the data which is provided below (on next page):
We can observe that:

- **Fbreadth** seems to be the attribute most correlated with others attribute
- We can observe that **Length, Fheight** and **Fbreadth** are strongly correlated so they may be considered as a single component to illustrate behavior of all

We further applied Principal Component Analysis in order to:

- Reduce the number of variables
- To have uncorrelated variables

Note: As the provided data is already in a standardized scale and available in units of millimeters, we didn't apply any normalization technique.

Correlation Matrix for the data is follows:

	Length	**Breadth**	**Height**	**Fheight**	**Fbreadth**
Length	1	0.11	0.43	0.75	0.57
Breadth	0.11	1	0.01	0.09	0.55
Height	0.43	0.01	1	0.29	0.20
Fheight	0.75	0.09	0.29	1	0.62
Fbreadth	0.57	0.55	0.20	0.62	1

We can see that Length and Fheight have high covariance (0.75). Similarly, Fheight and Fbreadth have high covariance (0.62). Breadth and Height have very low value of covariance. From which we can expect that they will make different components in PCA.

Note: In our PCA, we have not fixed any attribute and identified components to predict that. We execute PCA for all attributes to identify uncorrelated variables and reduce our variables.

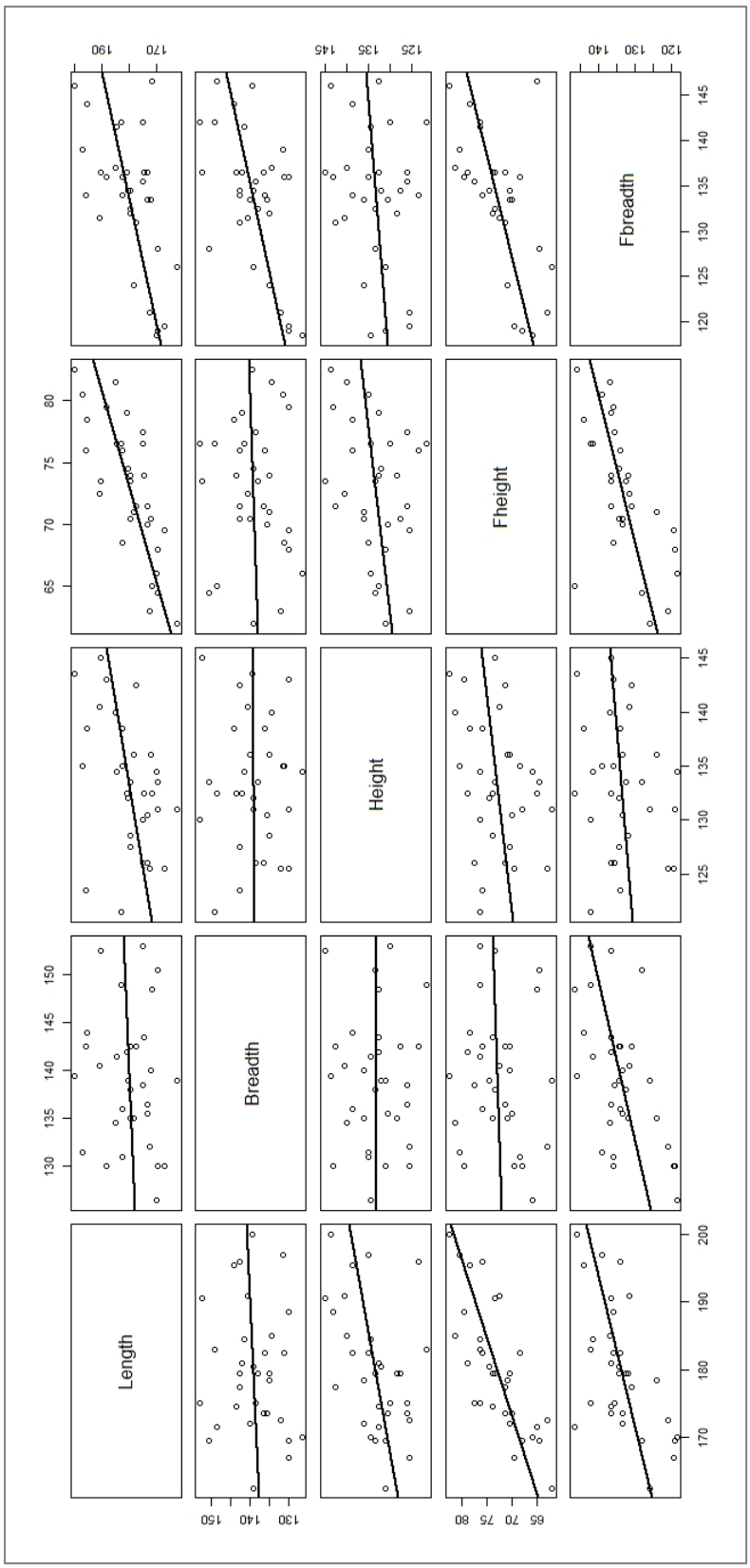


Figure 8 - Scatter plot Matrix

ASSIGNMENT 02

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

Principal Component Analysis:

	Comp1	Comp2	Comp3	Comp4	Comp5
Standard Deviation	1.61	1.09	0.87	0.51	0.45
Proportion of Variance	0.52	0.24	0.15	0.05	0.04
Cumulative Proportion	0.52	0.76	0.91	0.96	1

General practice is to consider components which captures around 80% of the cumulative proportion of all variables. In our case first two components are covering 75% and third component will capture 91% of proportion which we believe is no necessary in an analysis of five attributes (unless specifically specified), hence we are considering Comp1 and Comp2 as our Principal Components.

LOADINGS	Comp1	Comp2	Comp3	Comp4	Comp5
Length	0.54	0.26	0.16	0.70	0.37
Breadth	0.25	-0.76	-0.36	0.35	-0.33
Height	0.31	0.44	-0.82	-0.18	
Fheight	0.53	0.19	0.41	-0.21	-0.69
Fbreadth	0.52	-0.35		-0.57	0.52

From Loadings, we can observe that Comp1 is capturing Length, Fheight and Fbreadth which supports our identification made from Scatterplot. Comp2 captures Breadth and Height. Null values show that sum of square was nearly zero.

Labeling: **Component 1** -> Height & **Component 2**-> Width

From Comp1 and Comp2, we can get values for all 31 observations. We have shown first five:

PC Score	Comp1	Comp2
1	1.97	-0.48
2	-2.99	0.28
3	-2.85	0.66
4	-1.42	-1.60
5	-0.10	-0.53

Interpretation of Results:
Principal Component Analysis have deduced for us two components which captures 75% of the overall variance of the dataset (five attributes) and will allow us to continue further analysis on these two components.

ASSIGNMENT 02

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

CLUSTER ANALYSIS

First of all, we developed Dendrograms to observe our data. Dendrogram is provided below (on next page). As we look them, we cannot clearly identify value of cutoff to produce our clusters.

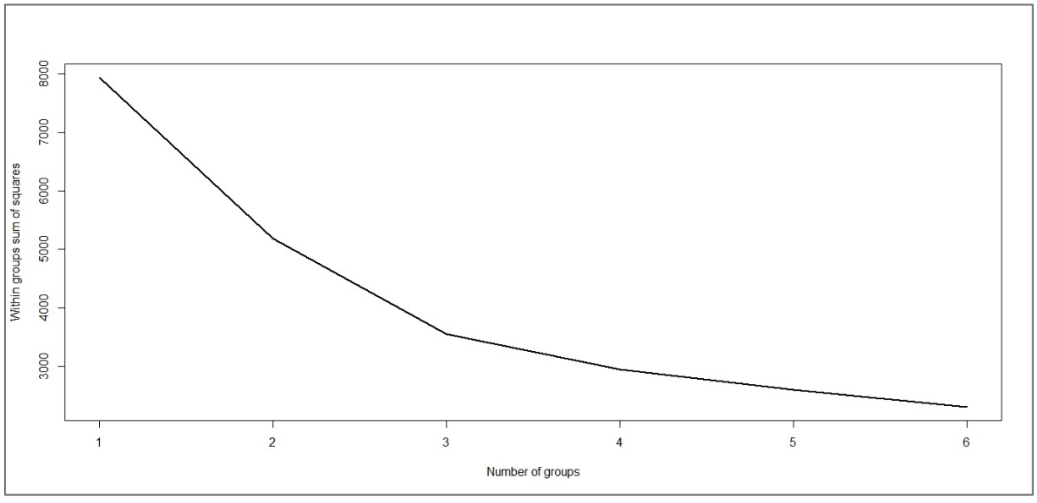
Note: Hierarchal clustering do not offer any formal cut off value. We use K-means to find formal value of cutoff and then go back to Hierarchal clustering.

Just in order to view clusters based on intuition, we considered value of 'h' as 22 and constructed below 07 clusters.

Cluster #	Observations (Total 32)
1	1, 20
2	2, 3, 15, 17
3	4, 10
4	5, 6, 7, 8, 9, 11, 13, 16, 18, 19, 23, 29, 32
5	12, 14, 21, 22, 25
6	24, 27, 28, 30, 31
7	26

K-MEANS

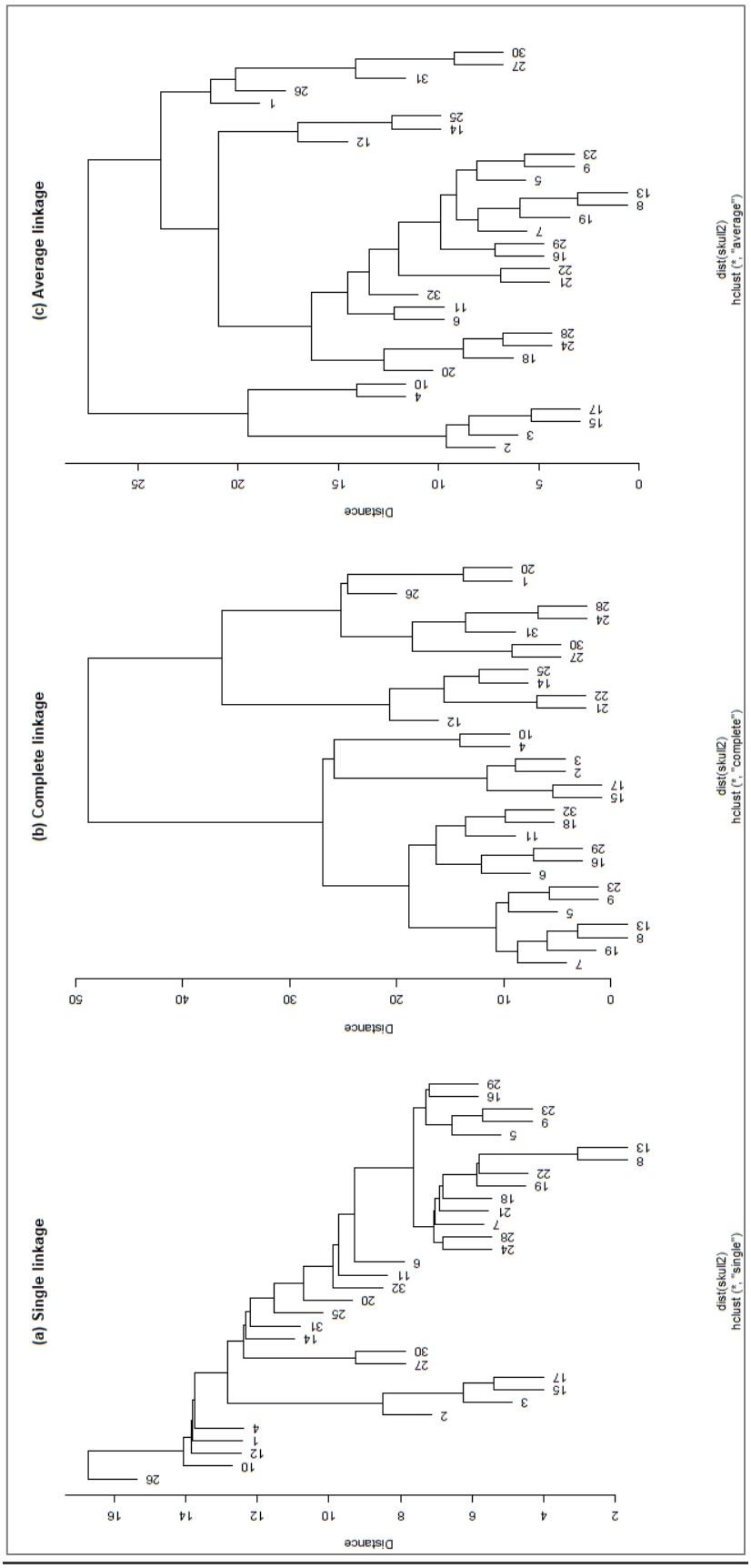
As we applied K-means, we obtained below graph:



We can see slope is breaking at two spots, at 2 and 3. We considered it at 3 as its steeper there.

ASSIGNMENT 02

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics



ASSIGNMENT 02

M.Sc. Geospatial Technologies
SIW004 – Applied Mathematics, Logic & Statistics

Sized of Cluster:

Cluster #	No. of Observations
1	6
2	16
3	10

Cluster #	Observations (Total 32)
1	2, 3, 10, 11, 15, 17
2	4, 5, 6, 7, 8, 9, 12, 13, 14, 16, 19, 22, 23, 25, 29, 32
3	1, 18, 20, 21, 24, 26, 27, 28, 30, 31

CLUSTER MEANS	Length	Breadth	Height	Fheight	Fbreadth
1	170	132.08	130.58	66.58	121.33
2	176.72	141.56	131.25	72.34	135.69
3	191.05	139.25	138.20	77.70	137.95

New Clusters assignment is shown for only first observations:

Observation	Old Cluster (Type)	New Cluster (Type2)
1	1	3
2	1	1
3	1	1
4	1	2
5	1	2

COMPARISON WITH EXISITNG 'TYPES':

Old Clusters were two in numbers, and our results suggested three. We could have selected two as well as slope break was also present at point 2 however, we believe that considering the break at three (steeper break) and adding the third cluster will increase similarity among a single group and dissimilarity between groups. This way boundary conditions of both old groups have been assigned with a new group to allow for more variance.

R Commands (Qs 03)

```
skulls<-source("F:\\UJI\\Courses\\Applied Mathematics\\Topic 1\\RSPCMA\\Data\\chap7tibetskull.dat")$value

attach(skulls)

skull2<-skulls[,-6]

attach(skull2)

pairs(skull2)

pairs(skull2,panel=function(x,y) {abline(lsfit(x,y)$coef,lwd=2)
                        points(x,y)})


cor(skull2)

skull2.pc<-princomp(skull2,cor=TRUE)
summary(skull2.pc,loadings=TRUE)

skull2.pc$scores[,1:2]

par(pty="s")
plot(skull2.pc$scores[,1],skull2.pc$scores[,2],
ylim=range(skull2.pc$scores[,1]),
xlab="PC1",ylab="PC2",type="n",lwd=2)
text(skull2.pc$scores[,1],skull2.pc$scores[,2],
labels=abbreviate(row.names(skull2)),cex=0.7,lwd=2)


par(mfrow=c(1,3))
plclust(hclust(dist(skull2),method="single"),labels=row.names(skull2),ylab="Distance")
title("(a) Single linkage")
plclust(hclust(dist(skull2),method="complete"),labels=row.names(skull2),ylab="Distance")
title("(b) Complete linkage")
plclust(hclust(dist(skull2),method="average"),labels=row.names(skull2),ylab="Distance")
```

```
title("(c) Average linkage")

pairs(skull2,panel=function(x,y) {abline(lsfit(x,y)$coef,lwd=2)
                        points(x,y)})

cutoff<-cutree(hclust(dist(skull2),method="complete"),h=22)



skull2.clus<-lapply(1:7,function(nc) row.names(skull2)[cutoff==nc])
skull2.mean<-lapply(1:7,function(nc) apply(skull2[cutoff==nc,],2,mean))
skull2.mean
skull2.clus


n<-length(skull2[,1])

wss1<-(n-1)*sum(apply(skull2,2,var))
wss<-numeric(0)
for(i in 2:6) {
        W<-sum(kmeans(skull2,i)$withinss)
        wss<-c(wss,W)
}

wss<-c(wss1,wss)
plot(1:6,wss,type="l",xlab="Number of groups",ylab="Within groups sum of squares",lwd=2)

kmean<-kmeans(skull2,3)
kmean
lapply(1:3,function(nc) apply(skull2[kmean$cluster==nc,],2,mean))


kmean$cluster
```