Assignment3

Aniket Guha Roy-19200164

5/23/2020

UCD School of Mathematics and Statistics Exam Honour Code.

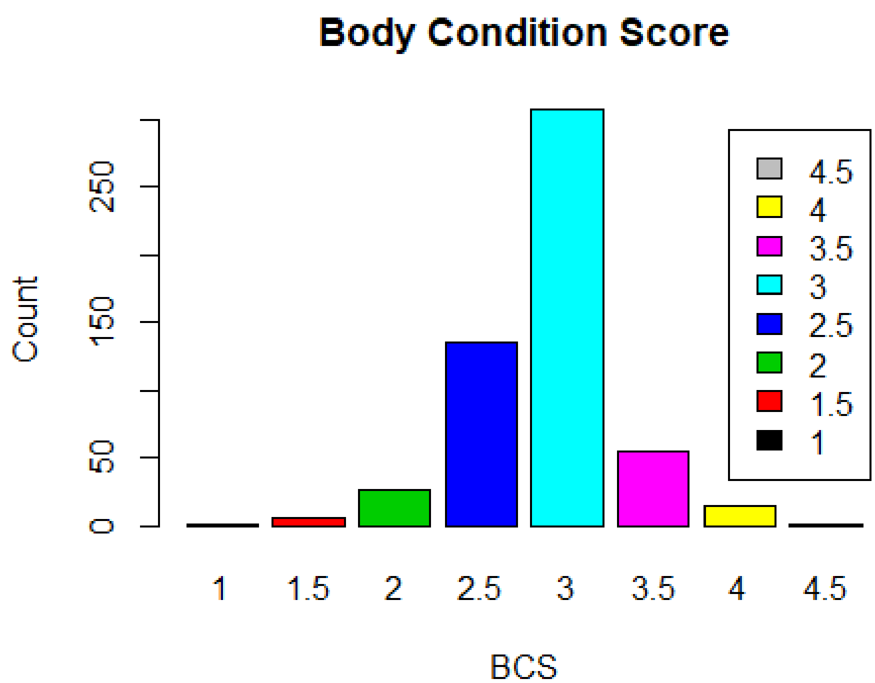
I confirm that I have not given aid, or sought and/or received aid for this assignment.
Name: Aniket Guha Roy Student Id: 19200164

Question 1a)

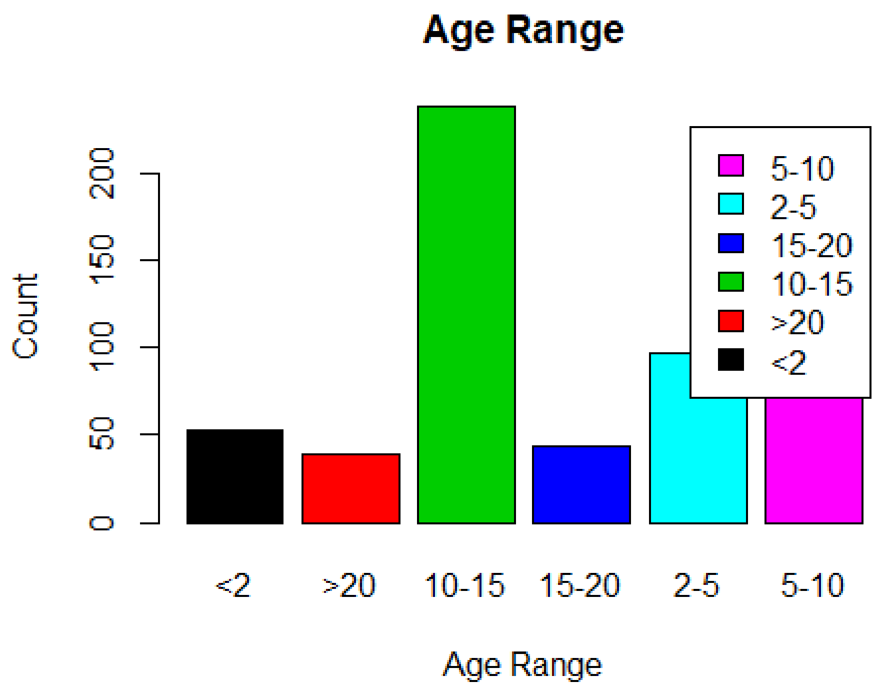
```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

#loading the dataset

donkey = read.csv(file.choose())
head(donkey)

##   BCS Age      Sex Length Girth Height Weight
## 1 3.0  <2 stallion     78    90     90     77
## 2 2.5  <2 stallion     91    97     94    100
## 3 1.5  <2 stallion     74    93     95     74
## 4 3.0  <2   female     87   109     96    116
## 5 2.5  <2   female     79    98     91     91
## 6 1.5  <2   female     86   102     98    105

str(donkey)

## 'data.frame':    544 obs. of  7 variables:
##  $ BCS   : num  3 2.5 1.5 3 2.5 1.5 2.5 2 3 3 ...
##  $ Age   : Factor w/ 6 levels "<2",">20","10-15",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ Sex   : Factor w/ 3 levels "female","gelding",..: 3 3 3 1 1 1 3 3 3 3
...
##  $ Length: int  78 91 74 87 79 86 83 77 46 92 ...
##  $ Girth : int  90 97 93 109 98 102 106 95 66 110 ...
##  $ Height: int  90 94 95 96 91 98 96 89 71 99 ...
##  $ Weight: int  77 100 74 116 91 105 108 86 27 141 ...

#distribution of the variables
summary(donkey)

##       BCS            Age              Sex           Length
##  Min.   :1.00   <2    : 53   female  :251   Min.   : 46.00
##  1st Qu.:2.50   >20   : 39   gelding : 79   1st Qu.: 92.00
```

```
##  Median :3.00   10-15:238   stallion:214   Median : 97.00
##  Mean   :2.89   15-20: 43                  Mean   : 95.67
##  3rd Qu.:3.00   2-5   : 97                 3rd Qu.:101.00
##  Max.   :4.50   5-10 : 74                  Max.   :112.00
##      Girth            Height           Weight
##  Min.   : 66.0   Min.   : 71.0   Min.   : 27.0
##  1st Qu.:112.8   1st Qu.: 99.0   1st Qu.:139.0
##  Median :117.0   Median :102.0   Median :155.0
##  Mean   :115.9   Mean   :101.3   Mean   :152.1
##  3rd Qu.:121.0   3rd Qu.:104.0   3rd Qu.:170.0
##  Max.   :134.0   Max.   :116.0   Max.   :230.0
```

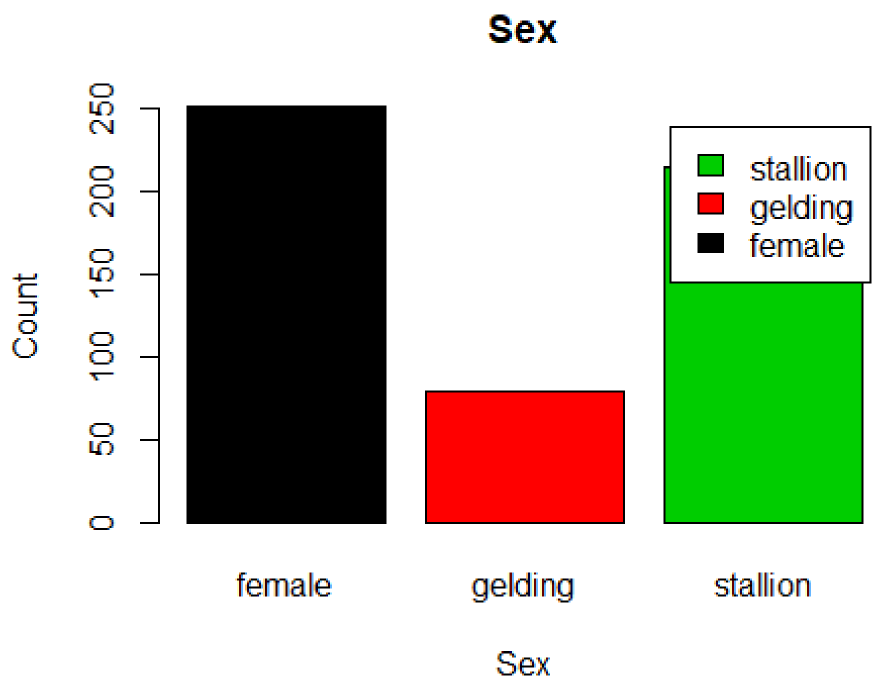
```
tab<-table(donkey[,1],donkey[,1])
colors = 1: length(unique(donkey$BCS))
barplot(tab,col = colors,legend=rownames(tab),
        xlab = "BCS",ylab = "Count",
        main="Body Condition Score" )
```



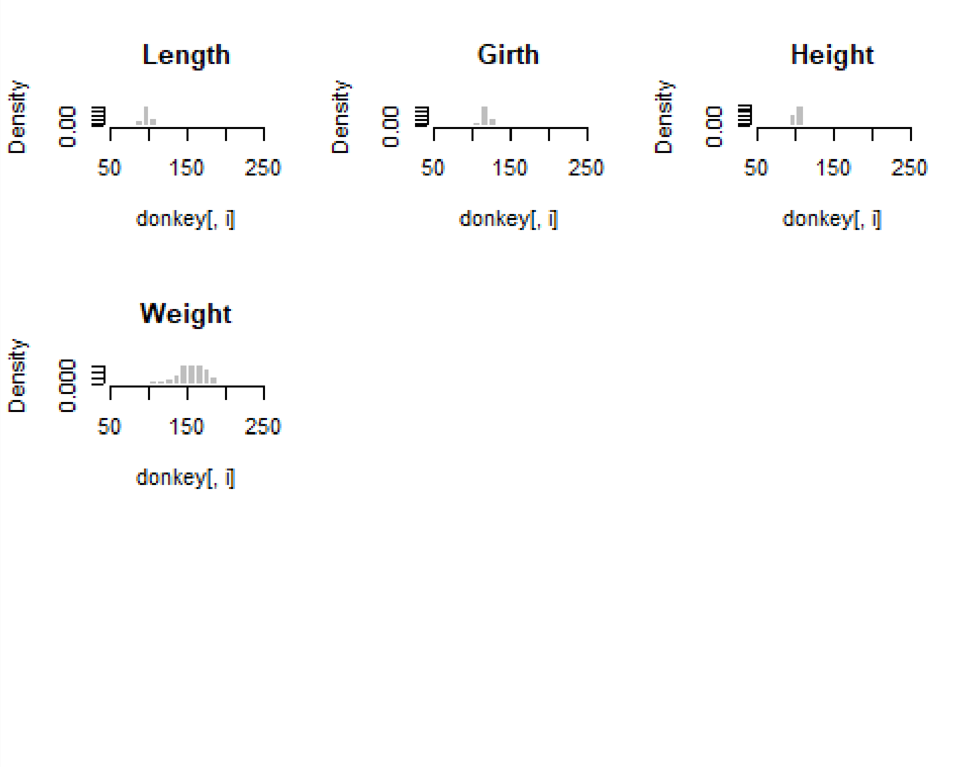

```
colorsA = 1:length(unique(donkey$Age))
tab2<-table(donkey[,2],donkey[,2])
barplot(tab2,col = colorsA,legend=rownames(tab2),
        xlab = "Age Range",ylab = "Count",
        main="Age Range")
```



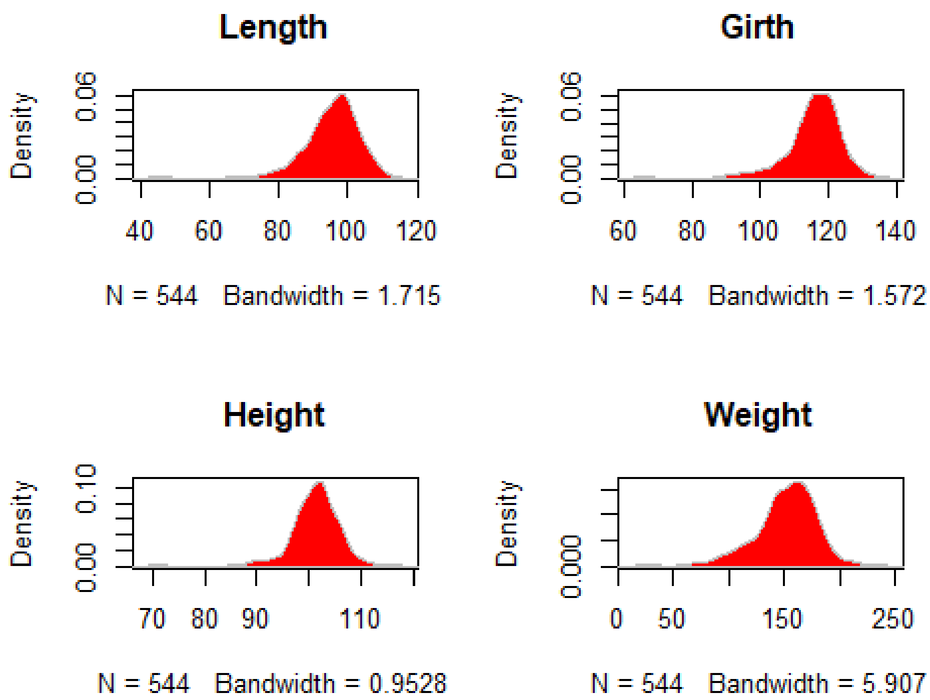

```
colorsS = 1:length(unique(donkey$Sex))
tab2<-table(donkey[,3],donkey[,3])
barplot(tab2,col = colorsA,legend=rownames(tab2),
        xlab = "Sex",ylab = "Count",
        main="Sex")
```




```
#histogram
par(mfrow=c(3, 3))
cols <- colnames(donkey)
for (i in 4:7) {
    hist(donkey[,i], xlim=c(50, 250), breaks=seq(0, 250, 10), main=cols[i],
probability=TRUE, col="gray", border="white")
}
```




```r
## Density plot
par(mfrow=c(2, 2))

for (i in 4:7) {
    d <- density(donkey[,i])
    plot(d, type="n", main=cols[i])
    polygon(d, col="red", border="gray")
}
```




```
#removing the outlying dokey from the datatset
donk<-donkey[-which.min(donkey$Length),]
```

From the summary, histograms, boxplots, and density plots, we get an idea on the distributions of each variable. Few are listed below: 1. Majority of the body conidtion score is of value 3 while 1 and 4 have the least frequency. 2. donkeys in the age group of 10-15 have the highest frequency and donkeys above 20 have the least frequency. 3. Females majorly dominate the dataset followed by stallion and gelding. 4: The lengths of the donkeys ranges mainly within 80-110. There are few outliers as is observed from the boxplots. 5: The girth also has a similar distribution where its values ranges mailny within 100-130. We have outliers here as well. 6. The Heights are strictly concentrated within 90-110 with few values outside this range. 7. The distribution of the Weight is spread but is almost normally distributed with a slight tendency of being positively skewed. #outlier removal We remove the donkey with the minimum length as we assume an observation point which is an outlier would be common for other variables as well.Hence we remove one observation as outlier.

Question 1b)

The main objective of PCA is dimensionality reduction method while retaining most of the variation in the data set.Principal components analysis involves breaking down the variance structure of a group of variables.Its difficult to apply PCA on Categorical variables as they not numerical,and thus have no variance structure or mean. Though we can convert categorical variables to a series of binary (0 or 1) variables and then perform principal components analysis on the result but its quite cumbersome. Hence we will consider the last 4 variables: Length, Girth, Height and WEight of the datasets. Though BCS, Body condition score is numerical we won't include that in our observation since it is not in sync with the dimensions of the variables and it can be treated as categorical as most of the observations can be classified into certain levels.

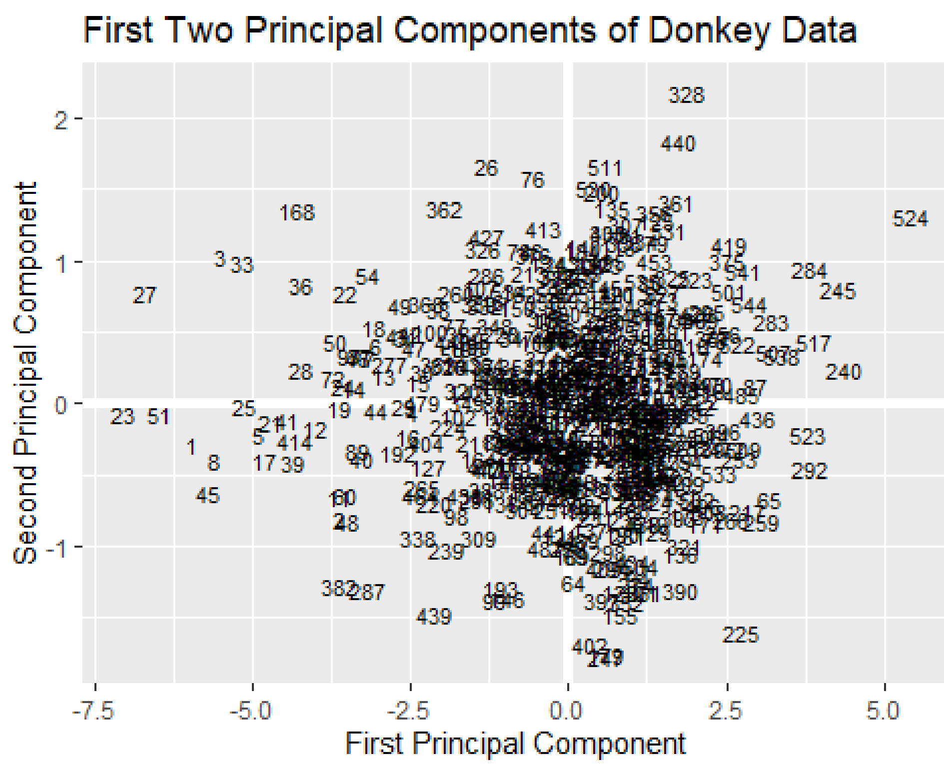
Question 1c)

We generally use correlation matrix when the scales of the variables are different covariance when the variables are not scaled. Correlation is actually a function of the covariance and are standardizd sets.In our case, since we would deal with only the variables related to the dimensions of the donkeys, its preferred to use covariance matrix rather correlation.

Question 1d)

```
#removing other variables
mat = donk[,4:7]
scaled_data = apply(mat,2,scale)
pca = function(data){
  d.cov = cov(scaled_data)
  d.eigen = eigen(d.cov)
  w =  d.eigen$vectors[,1:2]
  w = -w
  eigen.values = d.eigen$values
  row.names(w) = colnames(data)
  colnames(w) = c("PC1","PC2")
  PVE <- d.eigen$values / sum(d.eigen$values)
  return(list(w,PVE=PVE, eigenval =eigen.values ))
}
p = pca(scaled_data)
p

## [[1]]
##              PC1         PC2
## Length 0.4776173 -0.64618629
## Girth  0.5166499  0.05170898
## Height 0.4680907  0.75031229
## Weight 0.5346454 -0.12961841
##
## $PVE
## [1] 0.80123407 0.10284592 0.07595044 0.01996957
##
## $eigenval
## [1] 3.20493628 0.41138369 0.30380175 0.07987829

pca_values = p[[1]]
pve_values = p[[2]]
PC1 = as.matrix(scaled_data) %*% p[[1]][,1]
PC2 = as.matrix(scaled_data) %*% p[[1]][,2]
PC <- data.frame(dimensions = row.names(mat),PC1, PC2)
```

```r
ggplot(PC, aes(PC1, PC2)) +
  modelr::geom_ref_line(h = 0) +
  modelr::geom_ref_line(v = 0) +
  geom_text(aes(label = dimensions), size = 3) +
  xlab("First Principal Component") +
  ylab("Second Principal Component") +
  ggtitle("First Two Principal Components of Donkey Data")
```

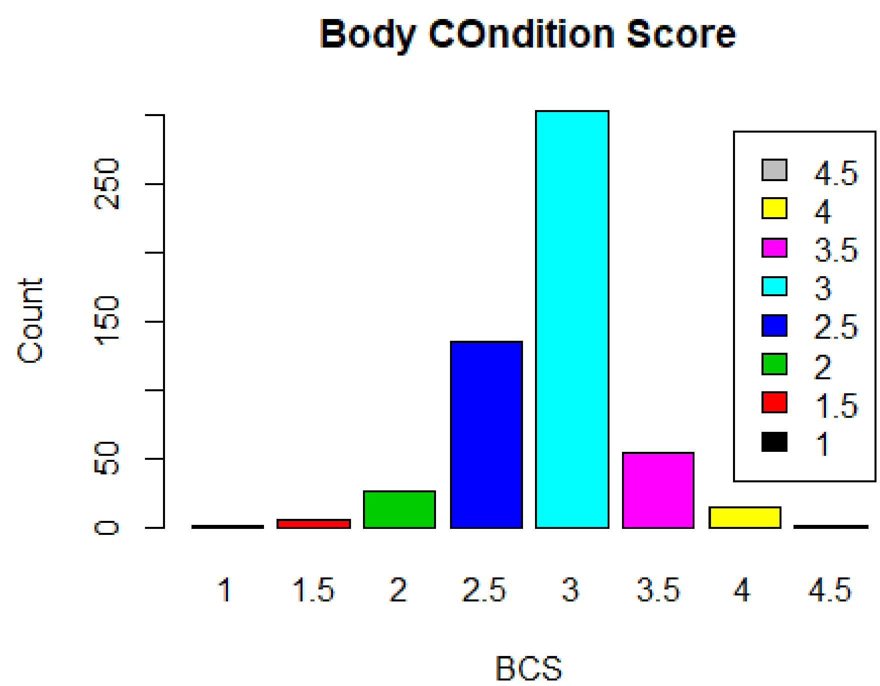


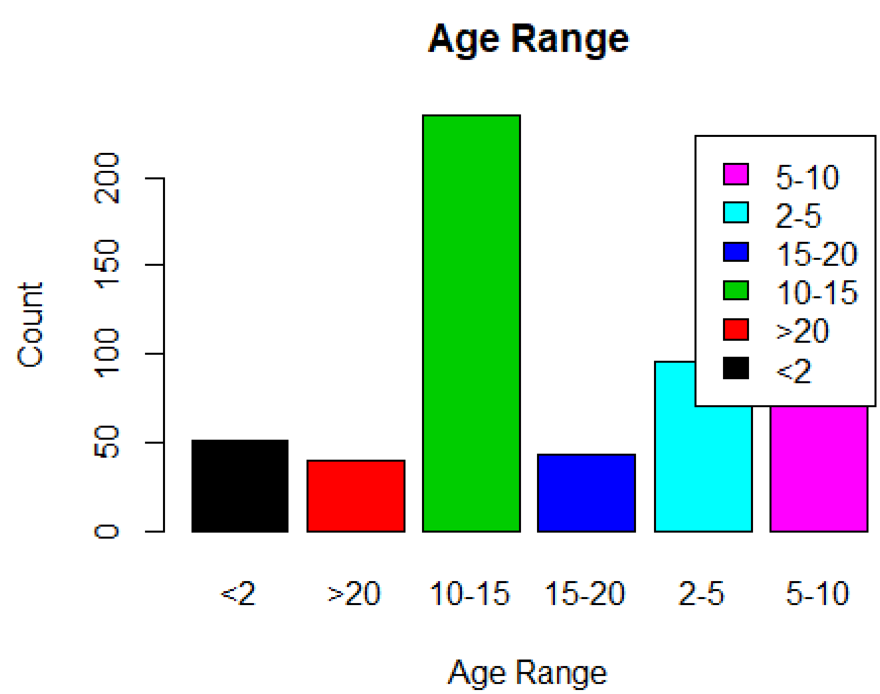
By default, eigenvectors in R point in the negative direction. During the calculation of PCA function, we have multiplied the default loadings by -1 as we'd prefer the eigenvectors point in the positive direction since it leads to more logical interpretation of graphical results. From the PCA function we get the loading factors and the proportion of variation explained by the respective principal components. The principal component scores are also calculated by projecting the n data points onto the first eigen vector and stored in PC. We also plot the graph of the two principal components that gives us an idea about the how the variabes are influenced by the principal component factors.

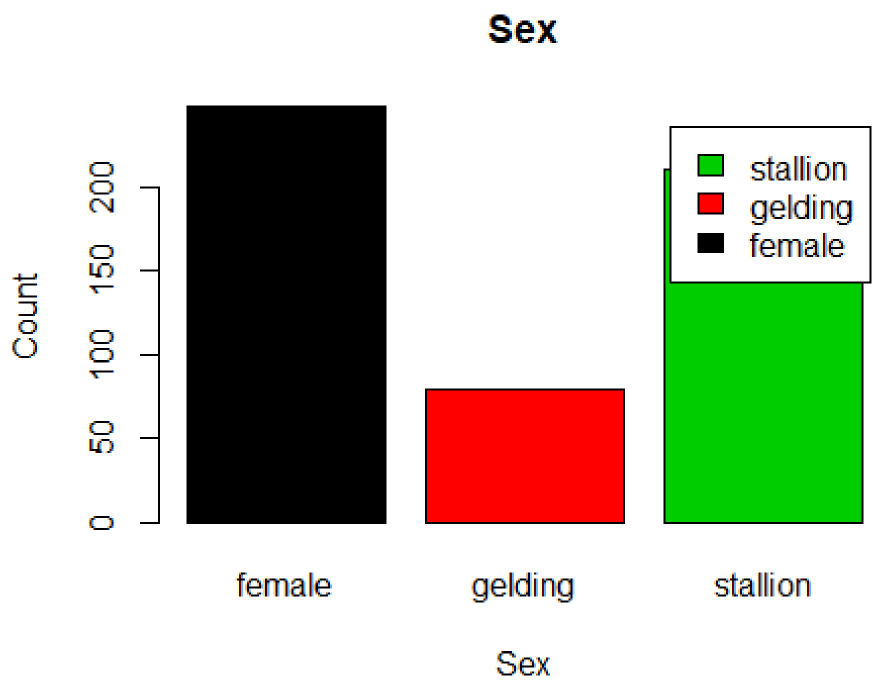
Question 1e)

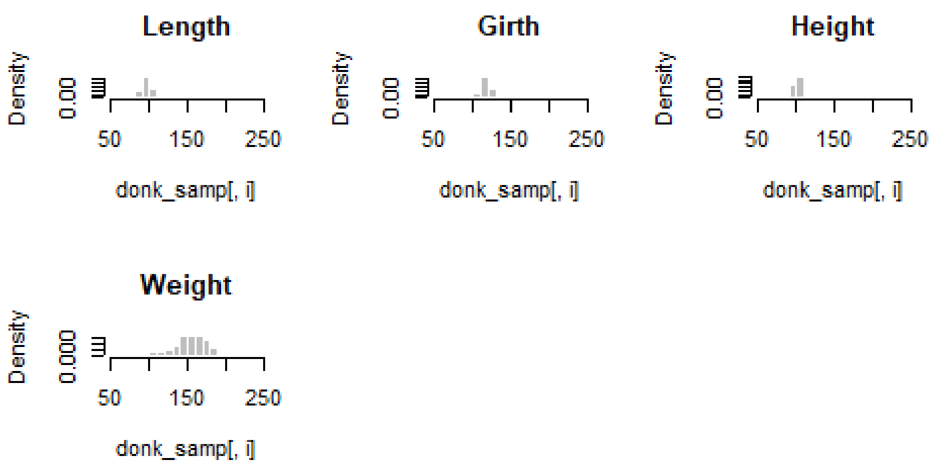
```
##        BCS              Age            Sex          Length
##  Min.   :1.000   <2    : 51   female  :248   Min.   : 68.00
##  1st Qu.:2.500   >20   : 39   gelding : 79   1st Qu.: 92.00
##  Median :3.000   10-15:235    stallion:211   Median : 97.00
##  Mean   :2.888   15-20: 43                   Mean   : 95.79
##  3rd Qu.:3.000   2-5   : 96                  3rd Qu.:101.00
##  Max.   :4.500   5-10  : 74                  Max.   :112.00
```

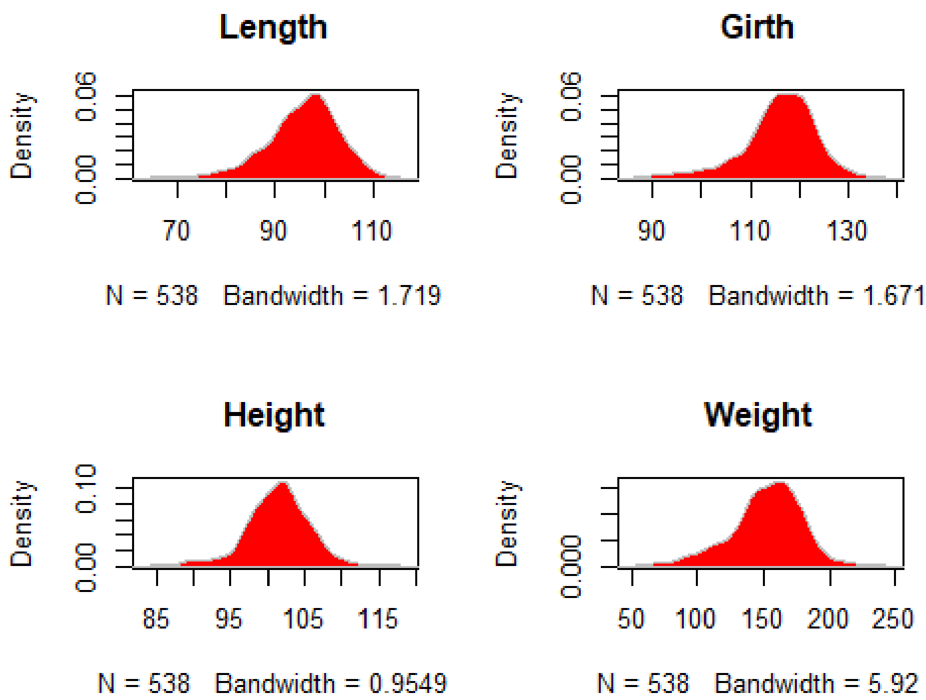
```
##      Girth           Height          Weight
##  Min.   : 90.0   Min.   : 86.0   Min.   : 65.0
##  1st Qu.:112.2   1st Qu.: 99.0   1st Qu.:139.0
##  Median :117.0   Median :102.0   Median :155.0
##  Mean   :116.0   Mean   :101.4   Mean   :152.3
##  3rd Qu.:121.0   3rd Qu.:104.0   3rd Qu.:170.0
##  Max.   :134.0   Max.   :116.0   Max.   :230.0
```


Body COndition Score
Count
0
50
150
250
1
1.5
2
2.5
3
3.5
4
4.5
BCS
4.5
4
3.5
3
2.5
2
1.5
1


Age Range
Count
0
50
100
150
200
<2
>20
10-15
15-20
2-5
5-10
Age Range
5-10
2-5
15-20
10-15
>20
<2

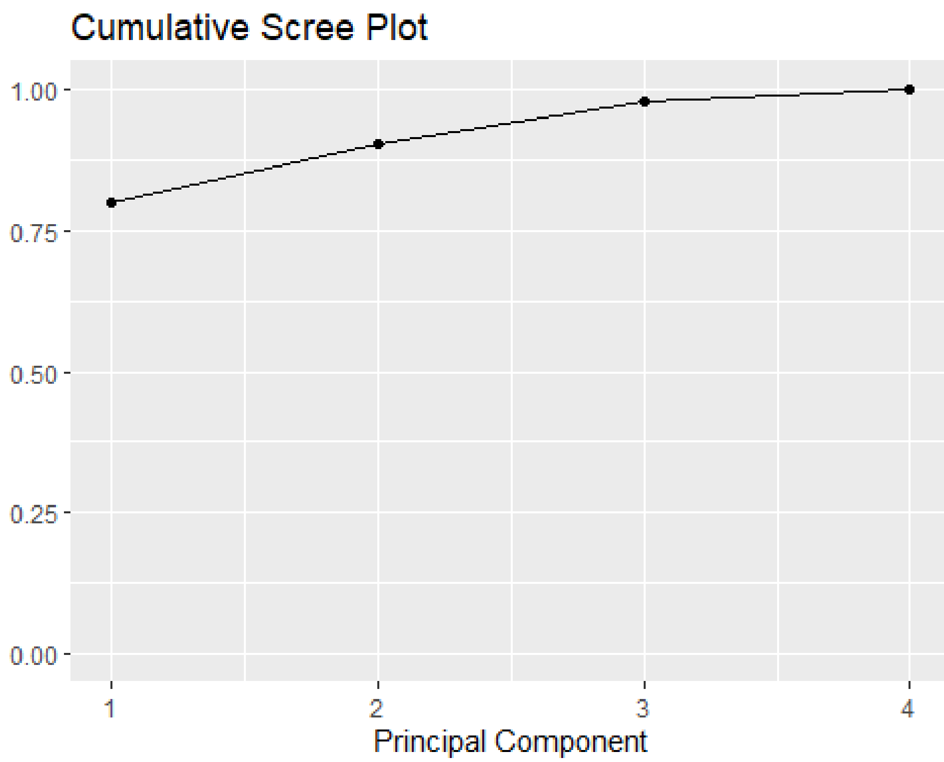

Sex
stallion
gelding
female
Count
0
50
100
150
200
female
gelding
stallion
Sex


Length
Girth
Height
Weight
Density
0.00
0.000
50
150
250
donk_samp[, i]




```
## $eigenvectors
##              PC1         PC2        PC3         PC4
## Length 0.4776173 -0.64618629 -0.5543660  0.21680254
## Girth  0.5166499  0.05170898  0.6163752  0.59201409
## Height 0.4680907  0.75031229 -0.4666717  0.01183748
## Weight 0.5346454 -0.12961841  0.3081842 -0.77612876
##
## $PVE
## [1] 0.80123407 0.10284592 0.07595044 0.01996957
##
## $eigenvalues
## [1] 3.20493628 0.41138369 0.30380175 0.07987829
```

```
##   dimensions       PC1         PC2
## 1          1 -5.941791 -0.28833822
## 2          2 -3.622530 -0.80340221
## 3          3 -5.480797  1.03865595
## 4          4 -2.466832 -0.06030295
## 5          5 -4.893142 -0.20690852
## 6          6 -3.035719  0.40500573
```



The most common technique that we use for determining how many principal components to keep is 'scree plot'. To determine the number of components, we look for the "elbow point", where the PVE significantly drops off. In our case, because we only have 4 variables to begin with, reduction to 2 variables while still explaining close to 90% of the variability is a good improvement, which is inferred from the cumulative screeplot.

Question 1f)

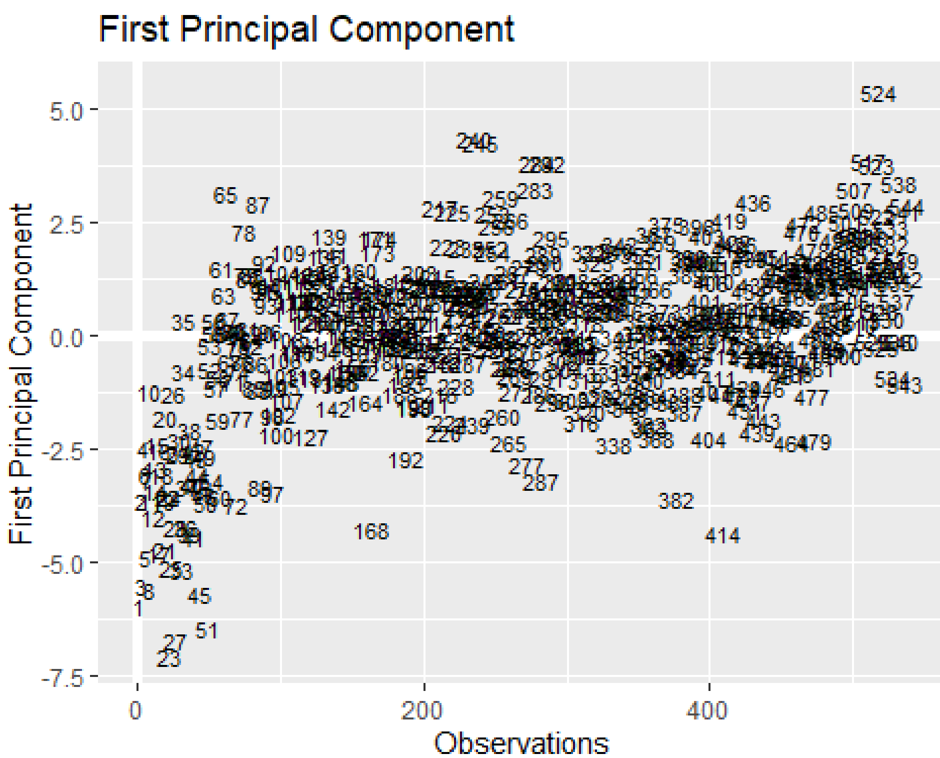
```
p$eigenvectors[,1]

##    Length     Girth    Height    Weight
## 0.4776173 0.5166499 0.4680907 0.5346454
```

The first principal component of a data set with columns X1, X2...Xn is the linear combination of the features: Z1 = (phi11*X1) + (phi21*X2) +... (phin1*Xn) that has the largest variance (i.e 80% in our case) and where phi11, phi21,...phin1 are the loadings of the first principal component. Hence the phi vector that maximizes the variance will be the first column of the loading matrix. Combining the loadings matrix along with the values of the dataset help to compute the principal component scores of each observation.We can see that for PC1, almost all the variables have high values with girth and weight being the highest. This explains the fact that PC1 explains around 80% of the variance of the whole dataset.

Question 1g)

```
ggplot(PC, aes(seq_along(PC1),PC1)) +
  modelr::geom_ref_line(h = 0) +
```

```
  modelr::geom_ref_line(v = 0) +
  geom_text(aes(label = dimensions), size = 3) +
   xlab("Observations") +
  ylab("First Principal Component") +
  ggtitle("First Principal Component")
```



The first principal component roughly corresponds to all the dimensions(lenght, girth, height and weight of the donkeys) in the context as it explains most of the variation (around 80%) in the dataset. So we could say the the observations points like 524,547,538 have quite high values of length, girth, height and weight. Similarly, we can say the observation points like 1,10,23,27 have less values of the variables. Most of the other observation points are normally distributed around 0 indicating average values of the dimensions. The distribution has a slight negative tendency since few observation points lie at the extreme corners signifying the lower and high values of the variables.

Question 1h)

```
n =  nrow(scaled_data1)
jack_score<-matrix(NA,nrow=n,ncol = 4)
#scaled.df<-scale(donkey.df[,c(4,7)],center = TRUE,scale = TRUE)
 for(i in 1:n)
 {
  jack = scaled_data1[-i,]
  pseudo = scaled_data1[i,]
  pca_pseudo = pca(jack)
  s<-as.matrix(pseudo)
  jack_score[i,]<-t(s)%*%as.matrix(p[[1]])
```

```
}
cat("PCA evaluation\n")

## PCA evaluation

cat("variance of the jackknife :")

## variance of the jackknife :

cat(diag(var(jack_score)))

## 3.212499 0.4061454 0.3029389 0.07841632

cat("\nvariance of the PCA function :")

##
## variance of the PCA function :

cat(p$eigenvalues)

## 3.204936 0.4113837 0.3038017 0.07987829
```

Jackknife is a resampling method which can be used to evalulate the quality of the PCA model. We apply the PCA function on the dataset as many times as the number of observations in the dataset while omitting one row each time. In this way, we conduct a PCA analysis for all all possible subsets of size (n-1). Consequently , we compare the variance of the jackknife and variance of the first prinicipal component and observe the values are alsmot equal.This indicates the goodness of the model and the accuracy.

Question 2

Aniket Guha Roy
19200164

STAT - 40150

2a) $x_i = \mu + \Lambda f_i + \epsilon_i$

where $x_i \in (i = 1, \dots N)$

$f_i \sim MVN_q(0, I)$

$\epsilon_i \sim MVN_p(0, \psi)$

$\Rightarrow$ f_i + ϵ_i are assumed independent;

$\Rightarrow$ $q \ll p$;

Let, $x_i - \mu = \Lambda f_i + \epsilon_i$

Λ : $(p \times q)$ matrix of factor loadings.

Under the orthogonal factor model,

$E[f_i] = 0$ $\quad$ $Cov(f_i) = I$.

$E[\epsilon_i] = 0$ $\quad$ $Cov(\epsilon_i) = \psi$.

f & ϵ are independent and so $Cov(f_i, \epsilon) = E[f_i \epsilon^T] = 0$

Expected Value of $(X - \mu)$

$E[(X-\mu)] = E[\Lambda f_i + \epsilon_i]$

$= \Lambda E[f_i] + E[\epsilon_i]$

$E[X_i] - \mu = 0$ $\quad$ as $E[f_i] = E[\epsilon_i] = 0$.

or $E[X_i] = \mu$

Expected value of covariance:

$E[(x_i - \mu)(x_i - \mu)^T] = E[(\Lambda f_i + \epsilon_i)(\Lambda f_i + \epsilon_i)^T]$

$= E[(\Lambda f_i)(\Lambda f_i)^T + (\Lambda f_i)^T \epsilon_i + (\Lambda f_i)\epsilon_i^T + \epsilon_i \epsilon_i^T]$

$= E[\Lambda f_i f_i^T \Lambda^T + f_i^T \Lambda^T \epsilon_i + \epsilon_i^T \Lambda f_i + \epsilon_i \epsilon_i^T]$

$= \Lambda E[f_i f_i^T] \Lambda^T + E[f_i^T \epsilon_i] \Lambda^T + \Lambda E[\epsilon_i^T f_i] + E[\epsilon_i \epsilon_i^T]$

$= \Lambda \Lambda^T + 0 + 0 + \psi$

as $E[f_i f_i^T] = Cov(f_i) = I$

and $E[\epsilon_i \epsilon_i^T] = Cov(\epsilon_i) = \psi$

$\therefore Cov(X_i - \mu) = \Lambda \Lambda^T + \psi$.

Hence, X_i has expected value μ and covariance $\Lambda\Lambda^T+\psi$.

μ has $p\times 1$ dimension.

$\Lambda\Lambda^T$ has $p\times p$ dimensions.

ψ has $p\times p$ dimensions

$\therefore \quad X_i \sim MVN_p(\mu, \Lambda\Lambda^T+\psi)$

2.6) From a mathematical viewpoint, a factor rotation is immaterial. But from interpretation point, the rotation of the axes of the factors sometimes results in making it more easily interpretable. Rotation does not change the position of variables relative to each other in the space of the factors ie: correlations between the variables are being preserved. The coordinates of the variable vector end points are changed onto the factor axes. After an orthogonal rotation of the loading matrix, factor variances get changed, but factors remain uncorrelated and variable communalities are preserved. In oblique rotation, factors are allowed to lose their uncorrelatedness if that will provide clearer 'simpler structure'. But interpretation of correlated factors is more difficult since we have to derive meaning from one factor so that it does not contaminate the meaning of another one that it correlates with. Hence, we can say that rotation is done in the pursuit of some structure, which may be called simpler structure. A simpler structure is where clusters of correlated variables show up.

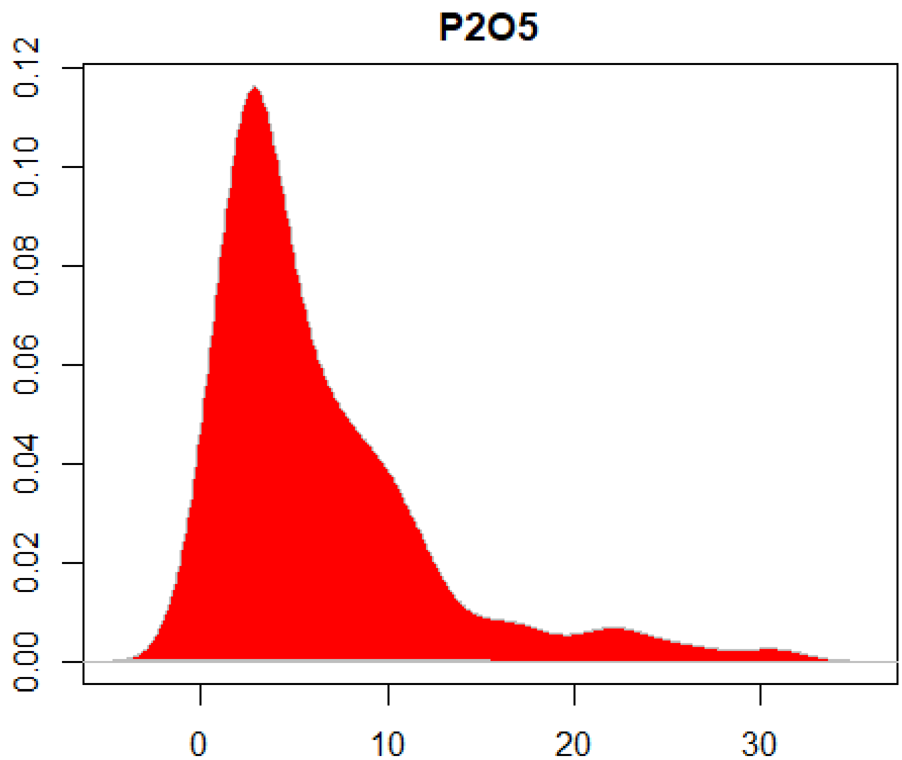
Question 2

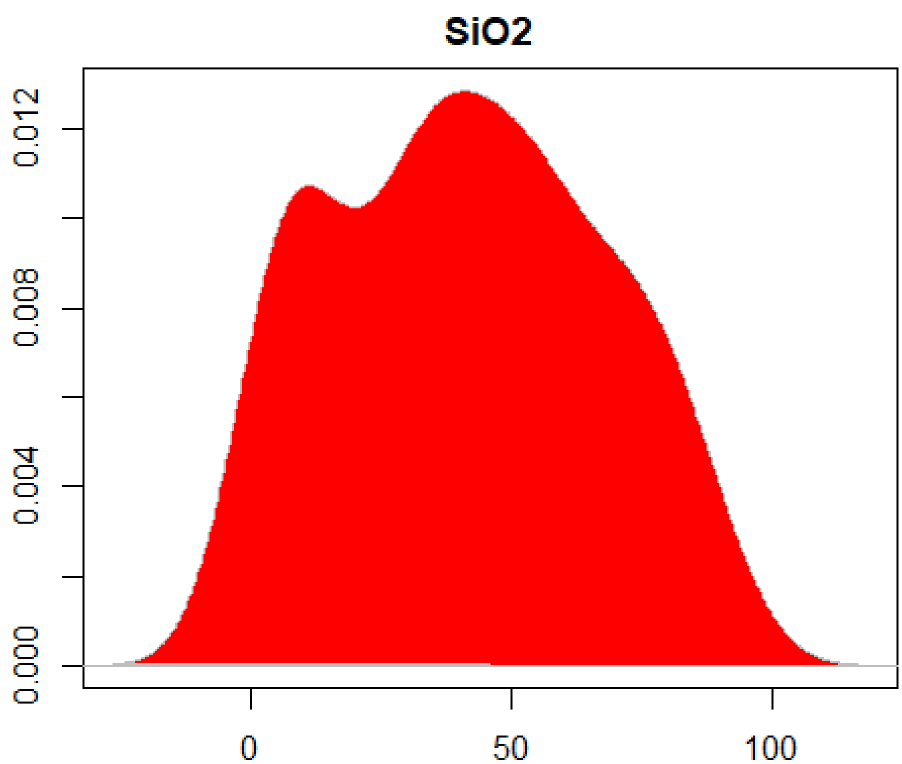
```
##loading the ash data
ash = read.csv(file.choose(),header = TRUE)
set.seed(19200164)
ss = sample(1:99, 5)
ash_samp = ash[-ss,]
head(ash_samp)
```

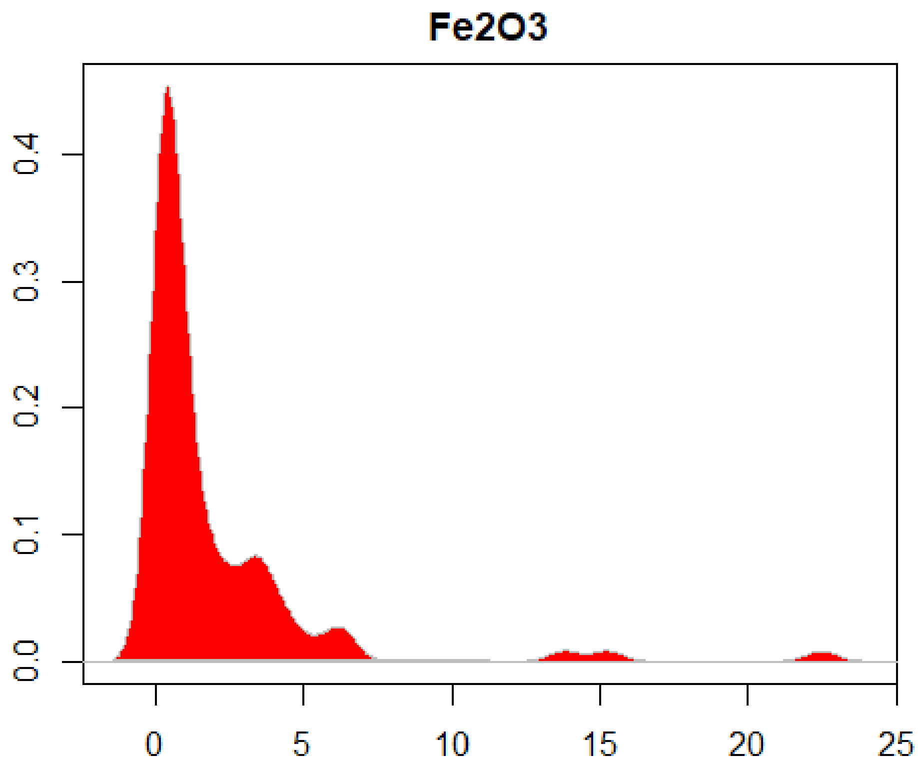
```
##    SOT   P2O5   SiO2 Fe2O3 Al2O3    CaO   MgO  Na2O    K2O
## 1  680  7.509  3.454 0.285 0.360 23.427 2.943 0.300 61.721
## 2  680  9.595  6.159 0.328 0.438 30.933 3.627 0.192 48.727
## 3  680  7.868  3.221 0.261 0.322 27.914 3.834 0.291 56.288
## 4 1070 11.956  5.593 0.333 0.378 23.881 5.532 0.181 52.146
## 5 1350 10.796  3.085 0.411 0.499 27.321 5.802 0.235 51.851
## 6  730  9.465 10.532 0.323 0.379 28.788 5.617 0.379 44.516
```

Question 2c-i)

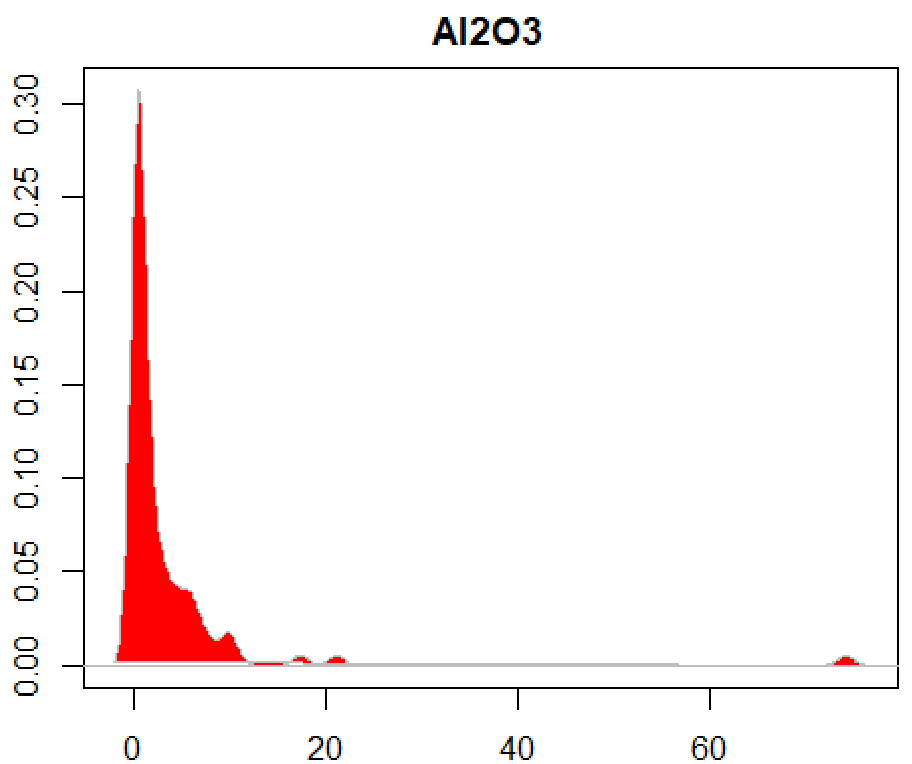
```
## Density plot
cols_ash <- colnames(ash_samp)
par(mar = rep(2, 4))
for (i in 2:9) {
    d <- density(ash_samp[,i])
    plot(d, type="n", main=cols_ash[i])
    polygon(d, col="red", border="gray")
}
```


P2O5
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
10
20
30

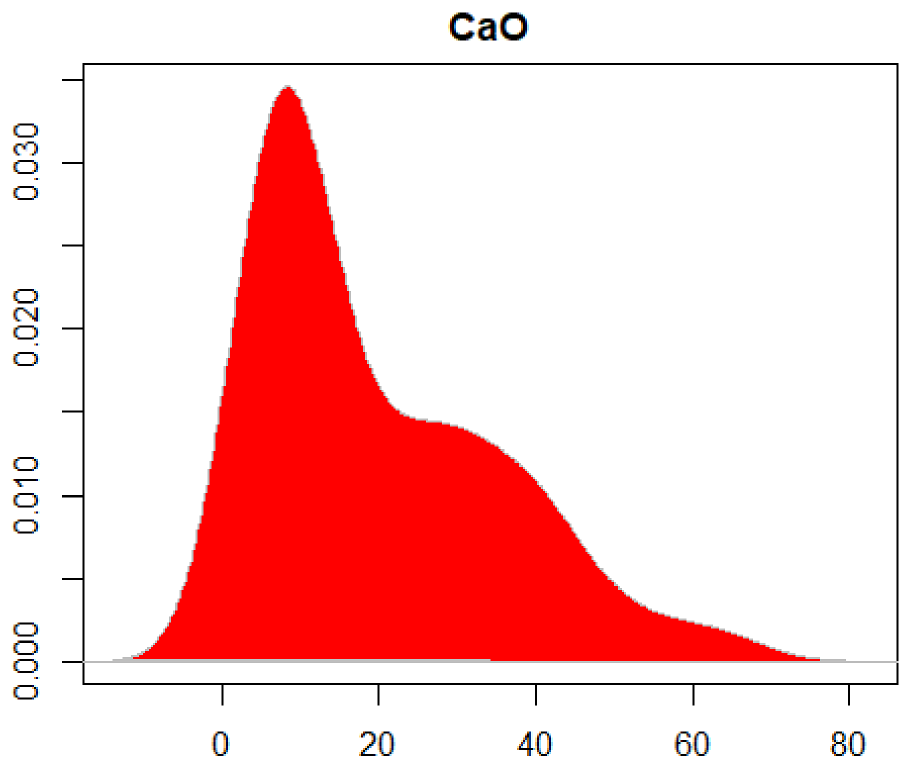

SiO2
0.000
0.004
0.008
0.012
0
50
100

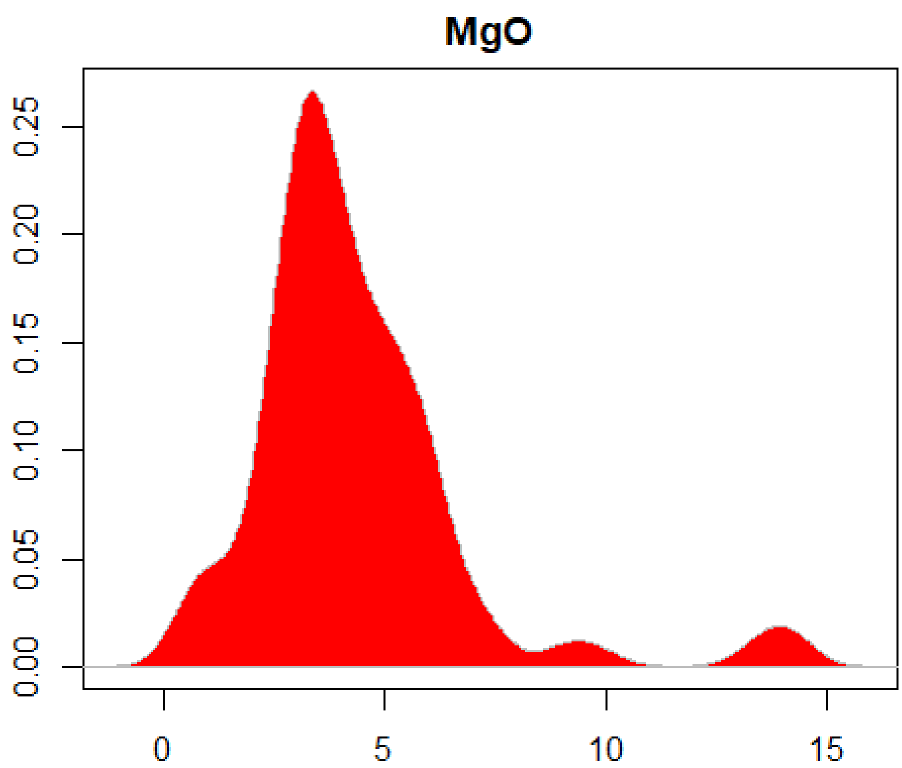


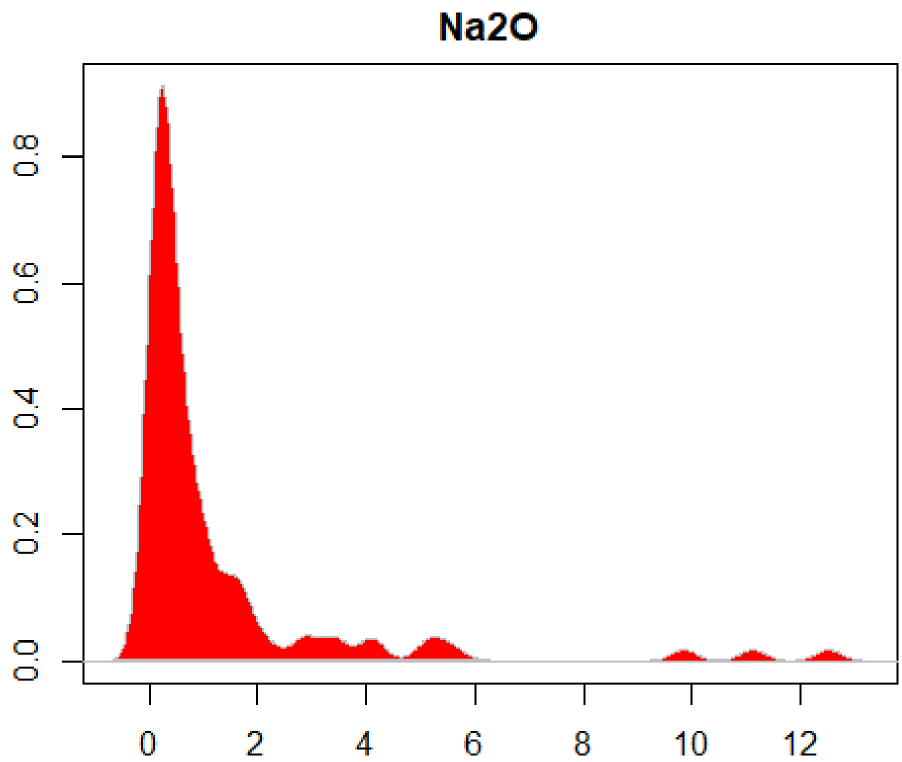
Fe2O3
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25

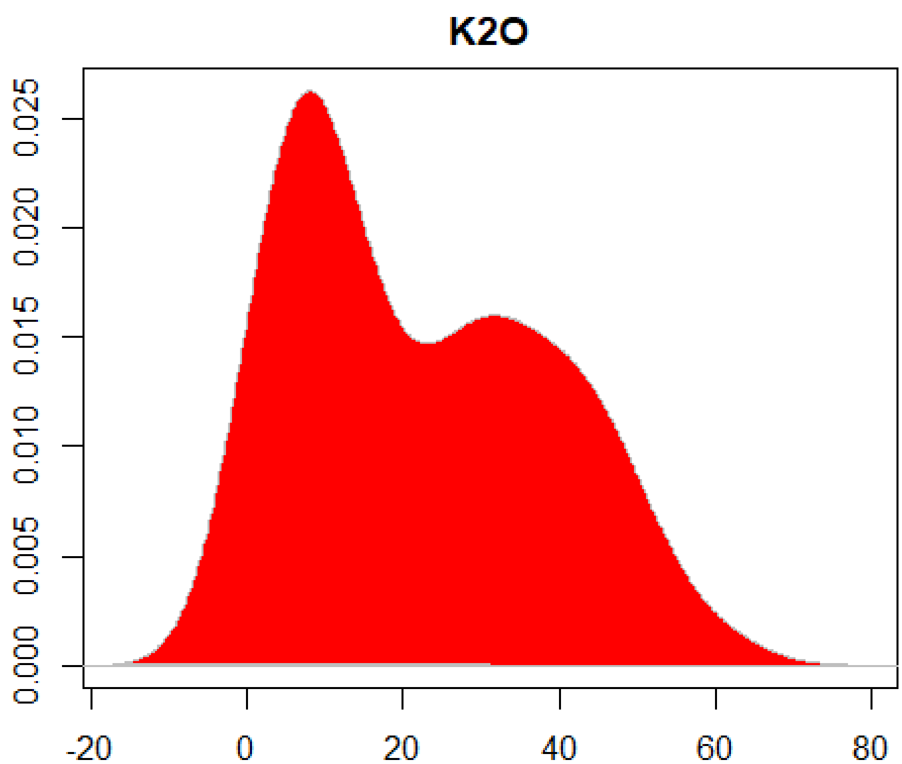


Al2O3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
20
40
60


CaO
0.000
0.010
0.020
0.030
0
20
40
60
80

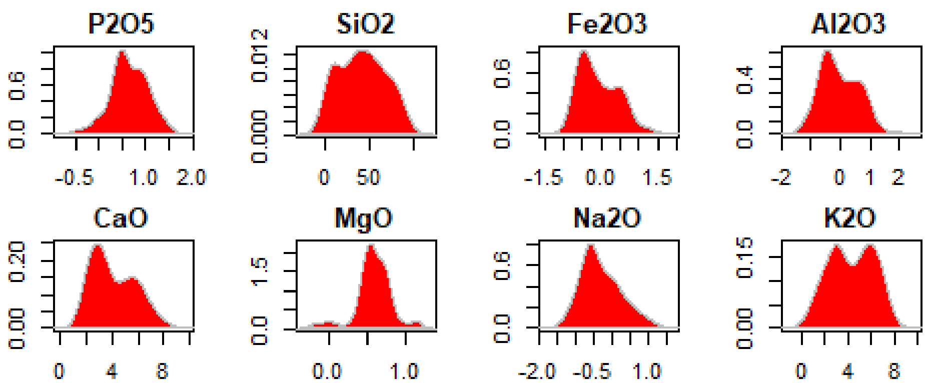

MgO
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15






```
## applying transformation
##ash_tfm = apply(ash_samp,2,log10)
ash_tfm = ash_samp
ash_tfm[,2] = log10(ash_samp[,2])
```

```
ash_tfm[,4] = log10(ash_samp[,4])
ash_tfm[,5] = log10(ash_samp[,5])
ash_tfm[,6] = sqrt(ash_samp[,6])
ash_tfm[,7] = log10(ash_samp[,7])
ash_tfm[,8] = log10(ash_samp[,8])
ash_tfm[,9] = sqrt(ash_samp[,9])

## Density plot
par(mfrow=c(4, 4))
for (i in 2:9) {
    d <- density(ash_tfm[,i])
    plot(d, type="n", main=cols_ash[i])
    polygon(d, col="red", border="gray")
}
```



The density plots of the mass concentrations of the ash samples indicate that most of the distributions of the variales are positively skewed. Among which, the variables P2O5, Fe2O3, Al2O3, MgO and Na2O are highly skewed and elements Ca0 and K2O are moderately skewed. Hence we need to apply transformation to the skewed data. Normal distribution of the variables are a requirement while calculating factor loadings in factor analysis.

After applying transformation we find that the highly skewed data are transformed using log10 transformation and moderately skewed data are transformed using sqrt transformation

Question 2c-ii)

```
fa2 = factanal(ash_tfm[,-1],factors = 2,rotation = "varimax")
fa3 = factanal(ash_tfm[,-1],factors = 3,rotation = "varimax")
fa4 = factanal(ash_tfm[,-1],factors = 4,rotation = "varimax")
fa2

##
## Call:
## factanal(x = ash_tfm[, -1], factors = 2, rotation = "varimax")
##
## Uniquenesses:
##  P2O5  SiO2 Fe2O3 Al2O3   CaO   MgO  Na2O   K2O
## 0.586 0.005 0.184 0.274 0.305 0.655 0.751 0.215
##
## Loadings:
##       Factor1 Factor2
## P2O5           0.640
## SiO2  -0.474  -0.877
## Fe2O3  0.890  -0.153
## Al2O3  0.838  -0.155
## CaO    0.733   0.398
## MgO    0.341   0.479
## Na2O   0.477   0.144
## K2O   -0.436   0.771
##
##                Factor1 Factor2
## SS loadings      2.796   2.230
## Proportion Var   0.350   0.279
## Cumulative Var   0.350   0.628
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 75.41 on 13 degrees of freedom.
## The p-value is 7.99e-11

fa3

##
## Call:
## factanal(x = ash_tfm[, -1], factors = 3, rotation = "varimax")
##
## Uniquenesses:
##  P2O5  SiO2 Fe2O3 Al2O3   CaO   MgO  Na2O   K2O
## 0.549 0.086 0.177 0.205 0.005 0.553 0.638 0.113
##
## Loadings:
##       Factor1 Factor2 Factor3
## P2O5           0.646   0.181
## SiO2  -0.330  -0.622  -0.647
## Fe2O3  0.789  -0.292   0.338
## Al2O3  0.827  -0.256   0.213
```

```
## CaO    0.316           0.946
## MgO    0.140   0.302   0.580
## Na2O   0.572   0.150   0.111
## K2O   -0.292   0.895
##
##                Factor1 Factor2 Factor3
## SS loadings      1.949   1.871   1.854
## Proportion Var   0.244   0.234   0.232
## Cumulative Var   0.244   0.478   0.709
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 29.87 on 7 degrees of freedom.
## The p-value is 1e-04
```

```
fa4
```

```
##
## Call:
## factanal(x = ash_tfm[, -1], factors = 4, rotation = "varimax")
##
## Uniquenesses:
##  P2O5  SiO2 Fe2O3 Al2O3   CaO   MgO  Na2O   K2O
## 0.370 0.022 0.005 0.214 0.005 0.458 0.658 0.185
##
## Loadings:
##       Factor1 Factor2 Factor3 Factor4
## P2O5           0.755   0.188   0.152
## SiO2  -0.377  -0.618  -0.610   0.286
## Fe2O3  0.838  -0.204   0.335   0.372
## Al2O3  0.809  -0.293   0.212
## CaO    0.327           0.941
## MgO    0.117   0.372   0.595   0.189
## Na2O   0.563   0.104   0.103
## K2O   -0.277   0.822          -0.250
##
##                Factor1 Factor2 Factor3 Factor4
## SS loadings      2.016   1.904   1.816   0.348
## Proportion Var   0.252   0.238   0.227   0.044
## Cumulative Var   0.252   0.490   0.717   0.760
##
## Test of the hypothesis that 4 factors are sufficient.
## The chi square statistic is 6.53 on 2 degrees of freedom.
## The p-value is 0.0381
```

```
##two columns of the loadings matrix
fa4$loadings[,1:2]
```

```
##           Factor1     Factor2
## P2O5  -0.03707183  0.75492170
## SiO2  -0.37689602 -0.61790751
## Fe2O3  0.83825398 -0.20435977
```

```
## Al2O3  0.80927542 -0.29292755
## CaO    0.32715530 -0.01874288
## MgO    0.11745675  0.37175006
## Na2O   0.56328502  0.10428894
## K2O   -0.27668866  0.82188766
```

Factor analysis with 2, 3 and 4 factors reveals below highlights: 1.factors=2: chi-square statistic is 75.41 and p value 7.99e-11 2.factors=3: chi-square statistic is 29.87 and p value 1e-04 3.factors=4: chi-square statistic is 6.53 and p value 0.0381

Since a low chi-squared statistic suggests a good model, we would select the model with 4 factors. The p values are quite less for models with factors 2 and 3, so it suggests that we can reject the null hyposthesis and can say that 2 and 3 factors are suufiecient as is also stated by 'hypothesis test'. For 4 factors the p-value is slighlt less. Overall, the model with 4 factors seem to be the best to capture the correlation structure in the variables.

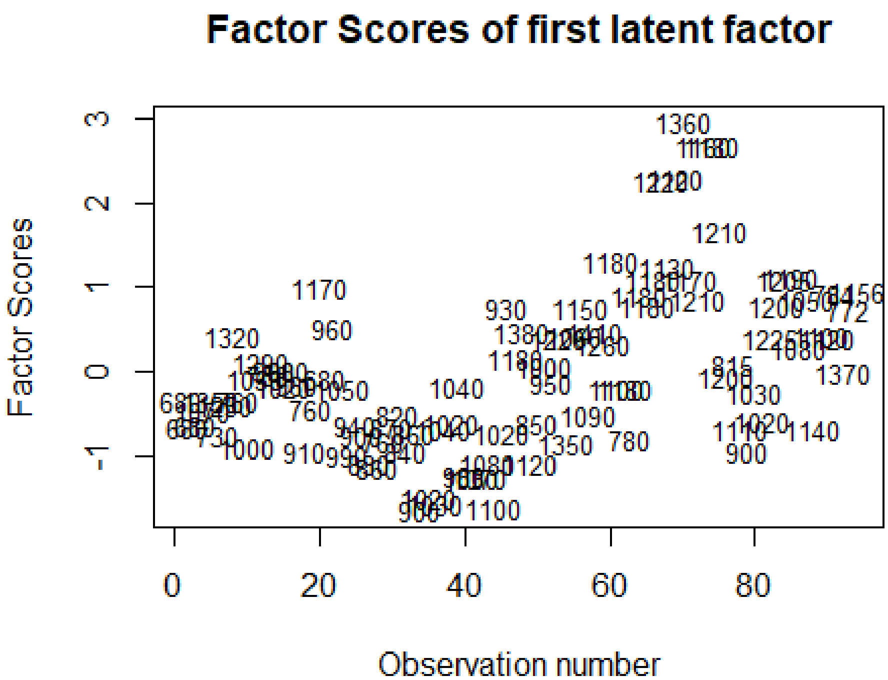
##Interpretation of first two columns of loadings matrix The range of loadings is between [-1,1]. A value close to 0 suggests the factor loading does not have significant impact on the variable and value cloase to -1,1 suggests significant impact of factor loading

We see that Factor 1 strongly influence Fe2O3, Al2O3 with 0.83 and 0.80 loading factors respectively. We also have significant impact on NA2o with 0.56. It has the least impact on P2o5 with a loading of .03. Factor 1 also have quite less influence on Mg0 and K2o with 0.11 and 0.27 loading.

On the other hand, Factor 2 have strong influence on P2o5, K2o and Si02 with 0.75, 0.82 and 0.61 laoding fators respectively. It has the least impact on Cao with only 0.01 loading. Na2o,FE2o3, AL2o3 and MGo also seem to have quite less impacts by factor 2 as they have loadng factors of 0.10,0.20,0.29 and 0.37. Conbbining both the factors, we see that Cao and Mg0 have quite less impacts compared to other variables.

Question 2c-iii)

```
fa_lf = factanal(ash_tfm[,-1],scores = "regression",factors = 4,rotation = "v
arimax")
{plot(fa_lf$scores[,1],type="n",xlab="Observation number",ylab="Factor Scores
",main="Factor Scores of first latent factor")
text(x=seq_along(fa_lf$scores[,1]),y=fa_lf$scores[,1], labels=ash_tfm$SOT,cex
=0.8)
}
```



To calculate the factor analysis scores for 4 factors, we apply regression method. The scores will help define the variation of the data by the 4 factors. The factor scores for the first latent factor define the variation of data explained by the first factor for each observation. Most the of the data lies below 1, hence we can say that these observations have not been explained by factor 1. Few observation points lie at the top, which indicates that their high variance with scores greater than 2.