business1.R

2023-03-15

```
# Import used libraries

library(readxl)
library(rgl)

# Remember to set the correct working directory!
# Import dataset

data <- read_excel("BankClients.xlsx")
data = data.frame(data)
head(data)
```

```
##   ID Age Gender Job Area CitySize FamilySize    Income    Wealth      Debt
## 1  1  24      1   1    2        2          4 0.6680457 0.7027862 0.2620704
## 2  2  47      1   2    2        3          1 0.8584531 0.9150432 0.7304303
## 3  3  38      0   2    1        2          2 0.9268178 0.8983156 0.4412723
## 4  4  67      0   2    1        2          3 0.5387970 0.4231798 0.6004011
## 5  5  33      0   2    1        3          1 0.8066585 0.7314038 0.8314495
## 6  6  81      0   2    1        3          4 0.4616797 0.3802429 0.4907270
##      FinEdu       ESG   Digital BankFriend LifeStyle    Luxury    Saving
## 1 0.7418525 0.4836840 0.6986246  0.6182594 0.6078768 0.8973694 0.2832216
## 2 0.8594230 0.5371667 0.9590247  0.7859364 0.8622712 0.9137287 0.8215896
## 3 0.4859534 0.6494336 0.7502654  0.6997249 0.7554039 0.7651989 0.5037898
## 4 0.4931437 0.5338292 0.5901652  0.6753534 0.3344322 0.5172088 0.6912398
## 5 0.8562864 0.7849399 0.7100256  0.7587931 0.9088782 0.6116103 0.6159157
## 6 0.2121204 0.4227478 0.3467378  0.3648083 0.3073098 0.5008489 0.4730362
##   Investments
## 1           1
## 2           3
## 3           3
## 4           2
## 5           2
## 6           3
```

```
dim(data)
```

```
## [1] 5000   18
```

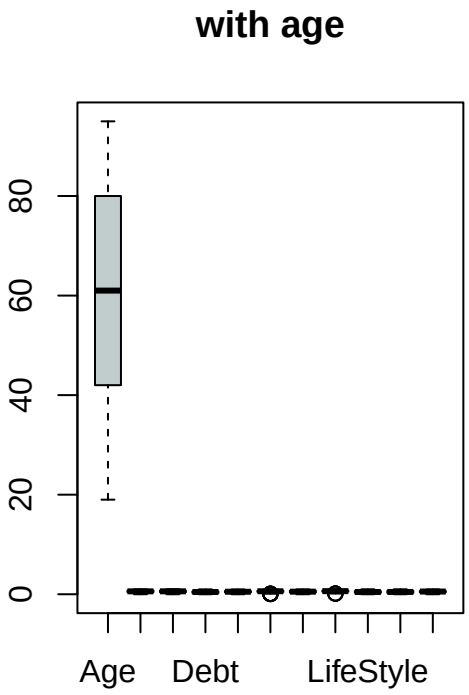
```
data = data[, -1]
colnames(data)
```

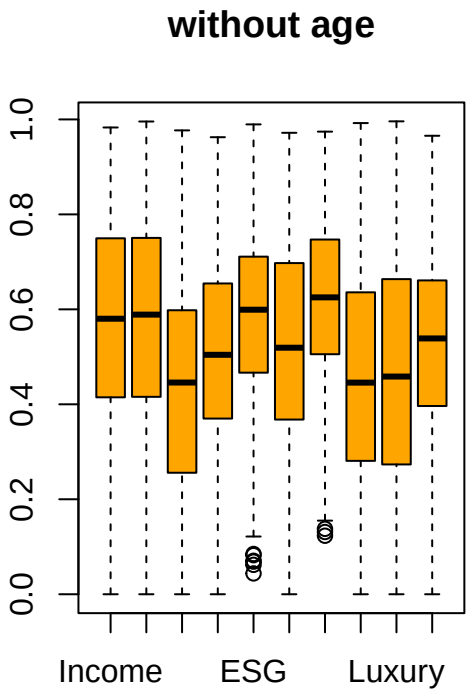
```
##  [1] "Age"         "Gender"      "Job"         "Area"        "CitySize"
##  [6] "FamilySize"  "Income"      "Wealth"      "Debt"        "FinEdu"
## [11] "ESG"         "Digital"     "BankFriend"  "LifeStyle"   "Luxury"
## [16] "Saving"      "Investments"
```

```r
# Divide cathegorical and numerical variables

ctg = c("Gender", "Job", "Area", "CitySize", "FamilySize", "Investments")
ctg_id = c(which(colnames(data)==ctg[1]),
           which(colnames(data)==ctg[2]),
           which(colnames(data)==ctg[3]),
           which(colnames(data)==ctg[4]),
           which(colnames(data)==ctg[5]),
           which(colnames(data)==ctg[6])
)
num = colnames(data)[-ctg_id]
num_id = numeric(length(num))
for (i in 1:length(num))
{
  num_id[i] = which(colnames(data)==num[i])
}
d_num = data[, num_id]
head(d_num)
```

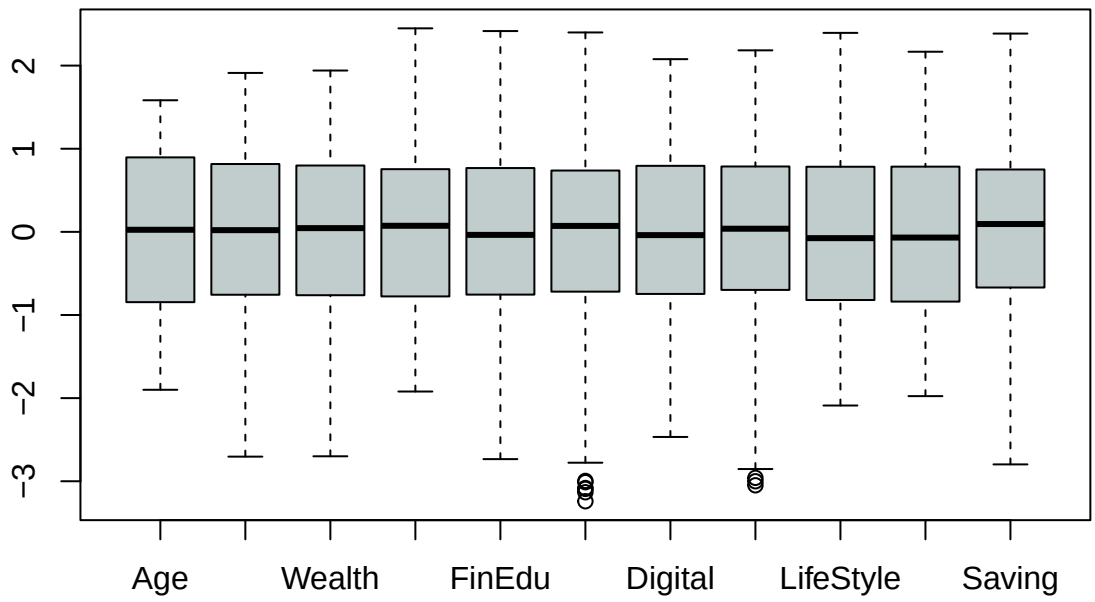
```
##   Age    Income    Wealth      Debt    FinEdu       ESG   Digital BankFriend
## 1  24 0.6680457 0.7027862 0.2620704 0.7418525 0.4836840 0.6986246  0.6182594
## 2  47 0.8584531 0.9150432 0.7304303 0.8594230 0.5371667 0.9590247  0.7859364
## 3  38 0.9268178 0.8983156 0.4412723 0.4859534 0.6494336 0.7502654  0.6997249
## 4  67 0.5387970 0.4231798 0.6004011 0.4931437 0.5338292 0.5901652  0.6753534
## 5  33 0.8066585 0.7314038 0.8314495 0.8562864 0.7849399 0.7100256  0.7587931
## 6  81 0.4616797 0.3802429 0.4907270 0.2121204 0.4227478 0.3467378  0.3648083
##   LifeStyle    Luxury    Saving
## 1 0.6078768 0.8973694 0.2832216
## 2 0.8622712 0.9137287 0.8215896
## 3 0.7554039 0.7651989 0.5037898
## 4 0.3344322 0.5172088 0.6912398
## 5 0.9088782 0.6116103 0.6159157
## 6 0.3073098 0.5008489 0.4730362
```

```r
# Some data visualization

par(mfrow = c(1,2))
boxplot(d_num, col = "azure3", main="with age")

# Age variable is out of scale

boxplot(d_num[,-1], col = "orange", main = "without age")
```





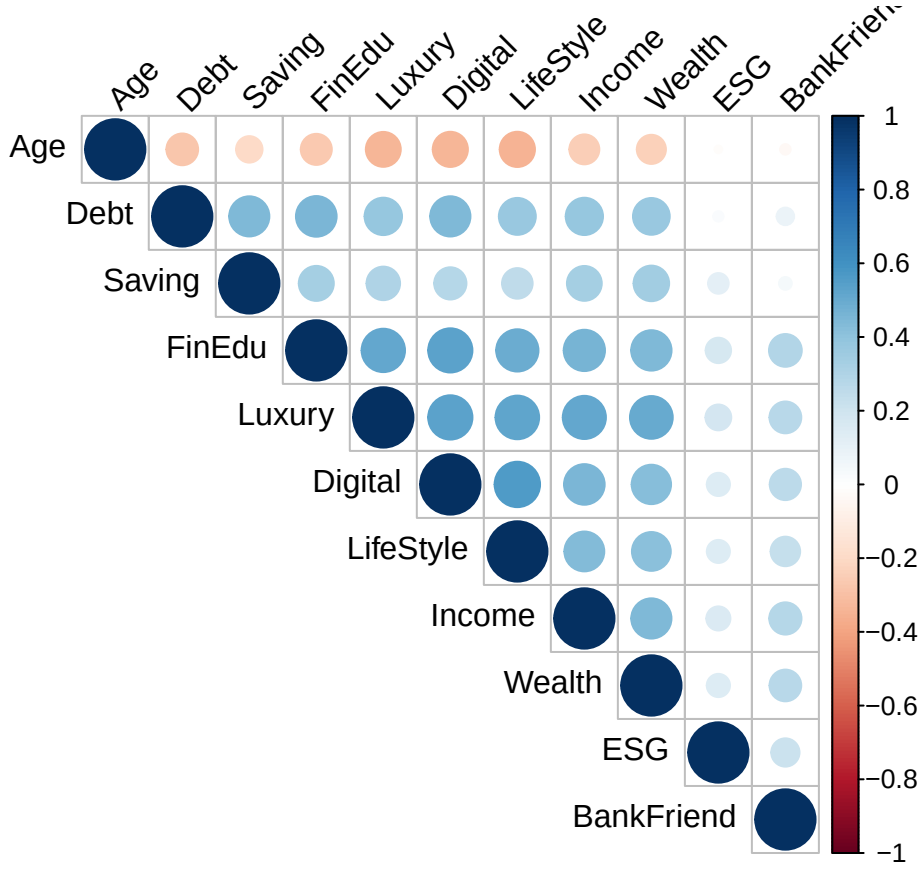

```
# Rescale Age variable

par(mfrow = c(1,1))
boxplot(scale(d_num), col = "azure3", main ="rescaled dataset")
```




```
# Correlation plot to analyze the general behaviour
x11()
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
corrplot(cor(d_num), type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



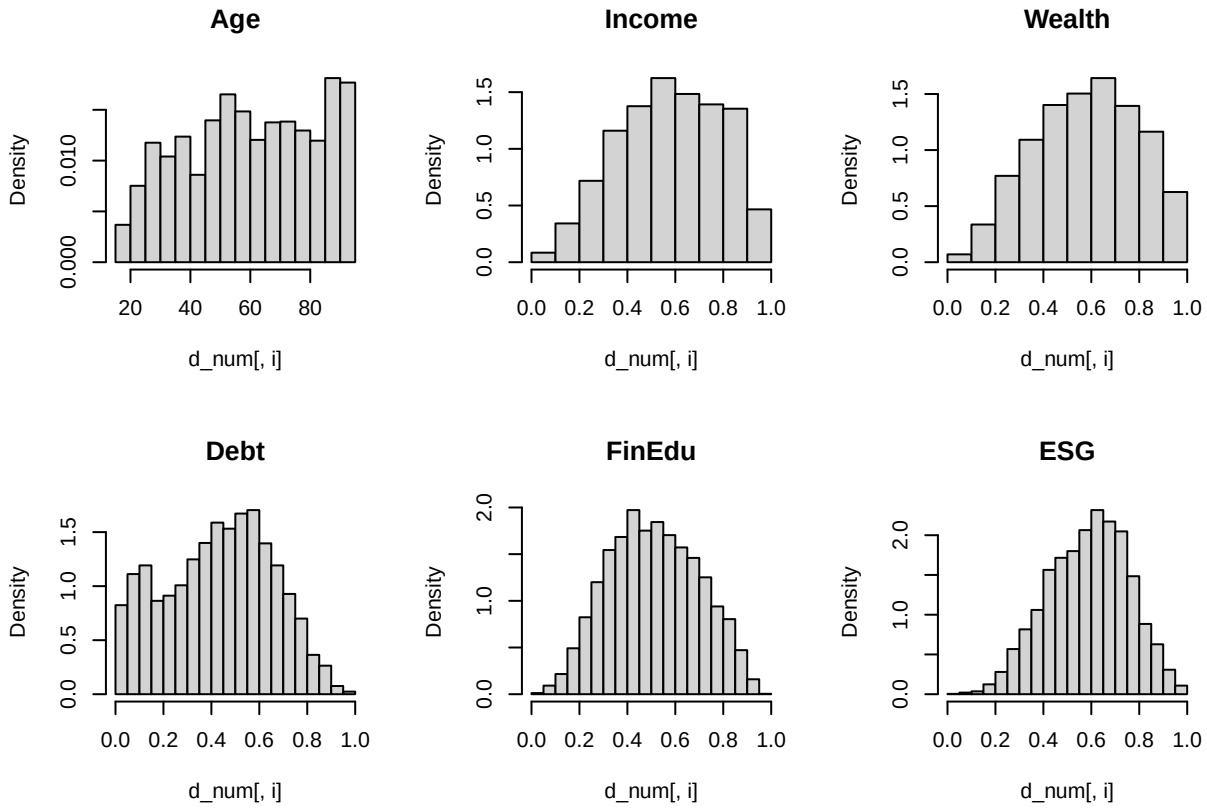

```
cor(d_num)
```

```
##                    Age      Income      Wealth        Debt      FinEdu          ESG
## Age         1.00000000 -0.2461308 -0.2326888 -0.27670543 -0.2646004 -0.01551783
## Income     -0.24613084  1.0000000  0.4445542  0.38206250  0.4682080  0.15940551
## Wealth     -0.23268881  0.4445542  1.0000000  0.37088272  0.4477582  0.14812417
## Debt       -0.27670543  0.3820625  0.3708827  1.00000000  0.4545482  0.02942753
## FinEdu     -0.26460035  0.4682080  0.4477582  0.45454817  1.0000000  0.17724825
## ESG        -0.01551783  0.1594055  0.1481242  0.02942753  0.1772482  1.00000000
## Digital    -0.33944312  0.4532769  0.4265600  0.44454869  0.5362112  0.14759467
## BankFriend -0.03035117  0.2852859  0.2763207  0.08459625  0.2948406  0.21946571
## LifeStyle  -0.34099190  0.4366634  0.4137667  0.37457476  0.4906263  0.14615618
## Luxury     -0.33457080  0.5116091  0.5022372  0.38462120  0.5172210  0.18295184
## Saving     -0.19237991  0.3333938  0.3477872  0.44240385  0.3368712  0.11692119
##                Digital  BankFriend  LifeStyle     Luxury      Saving
## Age        -0.3394431 -0.03035117 -0.3409919 -0.3345708 -0.19237991
## Income      0.4532769  0.28528588  0.4366634  0.5116091  0.33339381
## Wealth      0.4265600  0.27632068  0.4137667  0.5022372  0.34778721
## Debt        0.4445487  0.08459625  0.3745748  0.3846212  0.44240385
## FinEdu      0.5362112  0.29484062  0.4906263  0.5172210  0.33687122
## ESG         0.1475947  0.21946571  0.1461562  0.1829518  0.11692119
## Digital     1.0000000  0.26933884  0.5666284  0.5385245  0.28543880
## BankFriend  0.2693388  1.00000000  0.2394929  0.2782174  0.04474126
## LifeStyle   0.5666284  0.23949288  1.0000000  0.5229818  0.25695308
## Luxury      0.5385245  0.27821740  0.5229818  1.0000000  0.30892118
## Saving      0.2854388  0.04474126  0.2569531  0.3089212  1.00000000
```

```
dim(d_num)
```

```
## [1] 5000   11
```

```
x11()
par(mfrow = c(2,3))
for (i in 1:6)
{
  hist(d_num[,i], main = colnames(d_num)[i], freq = FALSE)
}
```




```
x11()
par(mfrow = c(2,3))
for (i in 7:length(num))
{
  hist(d_num[,i], main = colnames(d_num)[i], freq = FALSE)
}
```

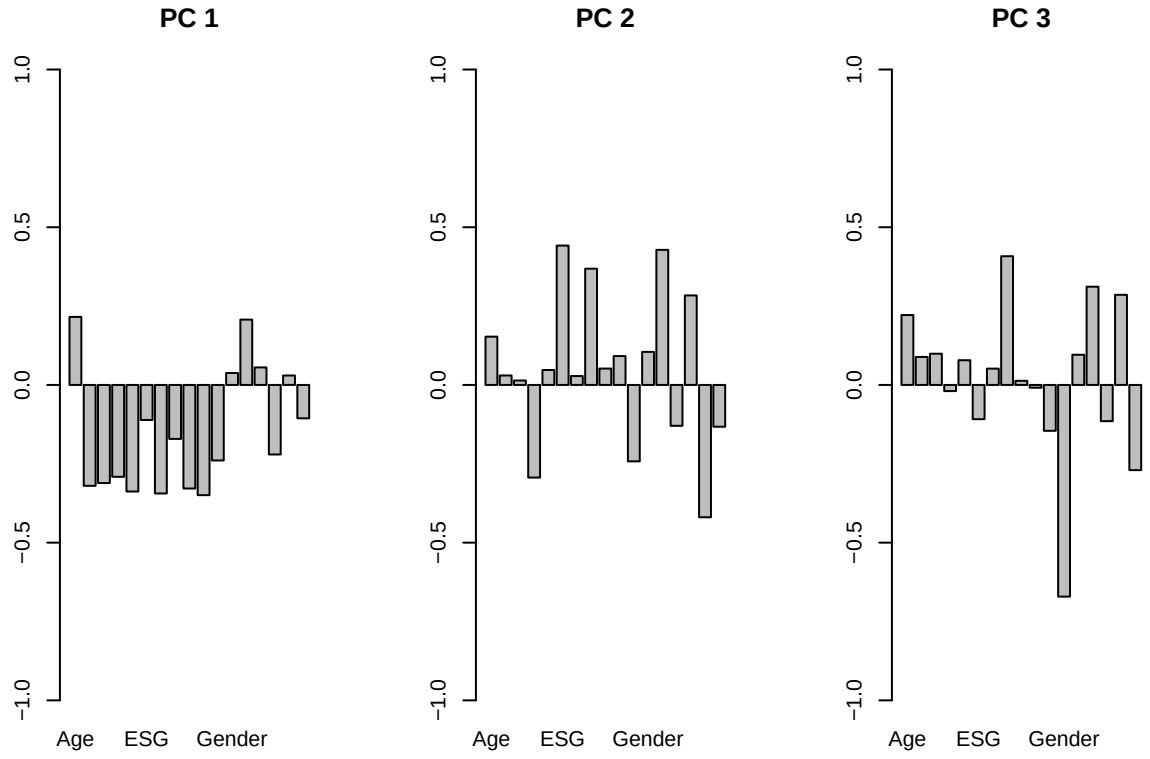
```
d_sc=data.frame(scale(data[,num_id]))
d_ctg=data[,ctg_id]
dnew=cbind(d_sc,d_ctg)
```

```
# PCA
```

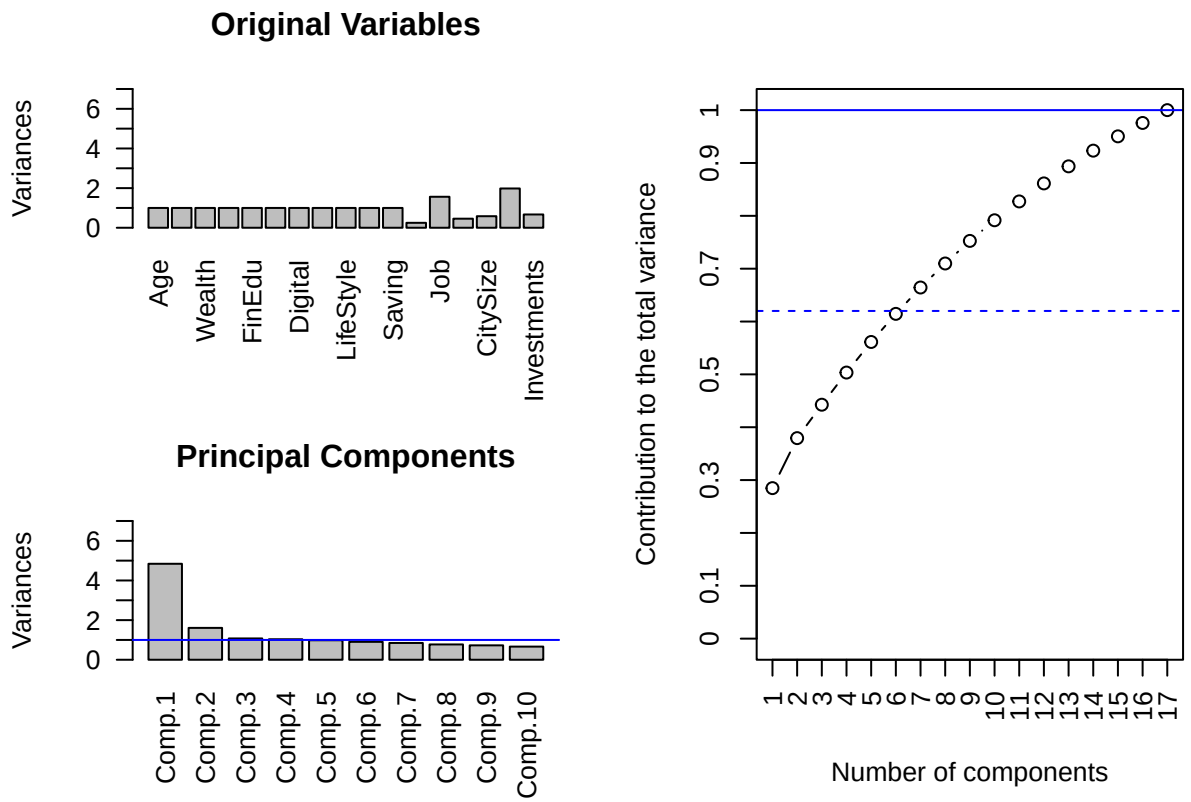
```
pc.data <- princomp(scale(dnew), scores=T)
summary(pc.data)
```

```
## Importance of components:
##                           Comp.1     Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.2000729 1.26805576 1.0365512 1.01705864 0.99157911
## Proportion of Variance 0.2847817 0.09460512 0.0632149 0.06085972 0.05784858
## Cumulative Proportion  0.2847817 0.37938682 0.4426017 0.50346144 0.56131002
##                           Comp.6     Comp.7     Comp.8     Comp.9   Comp.10
## Standard deviation     0.95100818 0.92028437 0.87902883 0.85231115 0.8146081
## Proportion of Variance 0.05321162 0.04982899 0.04546154 0.04273998 0.0390423
## Cumulative Proportion  0.61452164 0.66435062 0.70981217 0.75255214 0.7915944
##                           Comp.11    Comp.12    Comp.13    Comp.14    Comp.15
## Standard deviation     0.77855270 0.76103714 0.74081421 0.71057806 0.67847198
## Proportion of Variance 0.03566268 0.03407608 0.03228915 0.02970719 0.02708331
## Cumulative Proportion  0.82725712 0.86133321 0.89362235 0.92332954 0.95041285
##                           Comp.16    Comp.17
## Standard deviation     0.65589630 0.64234958
## Proportion of Variance 0.02531094 0.02427621
## Cumulative Proportion  0.97572379 1.00000000
```

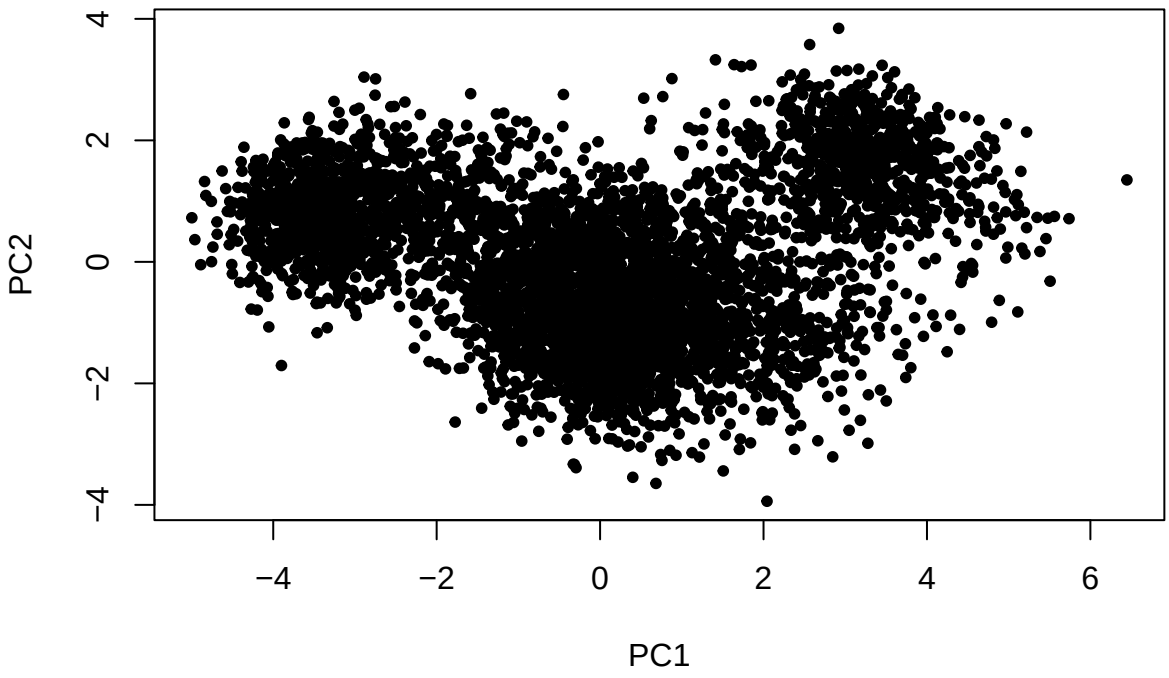
```
load.data <- pc.data$loadings
x11()
par(mfcol = c(1,3))
for(i in 1:3) barplot(load.data[,i], ylim = c(-1, 1), main=paste("PC",i))
```



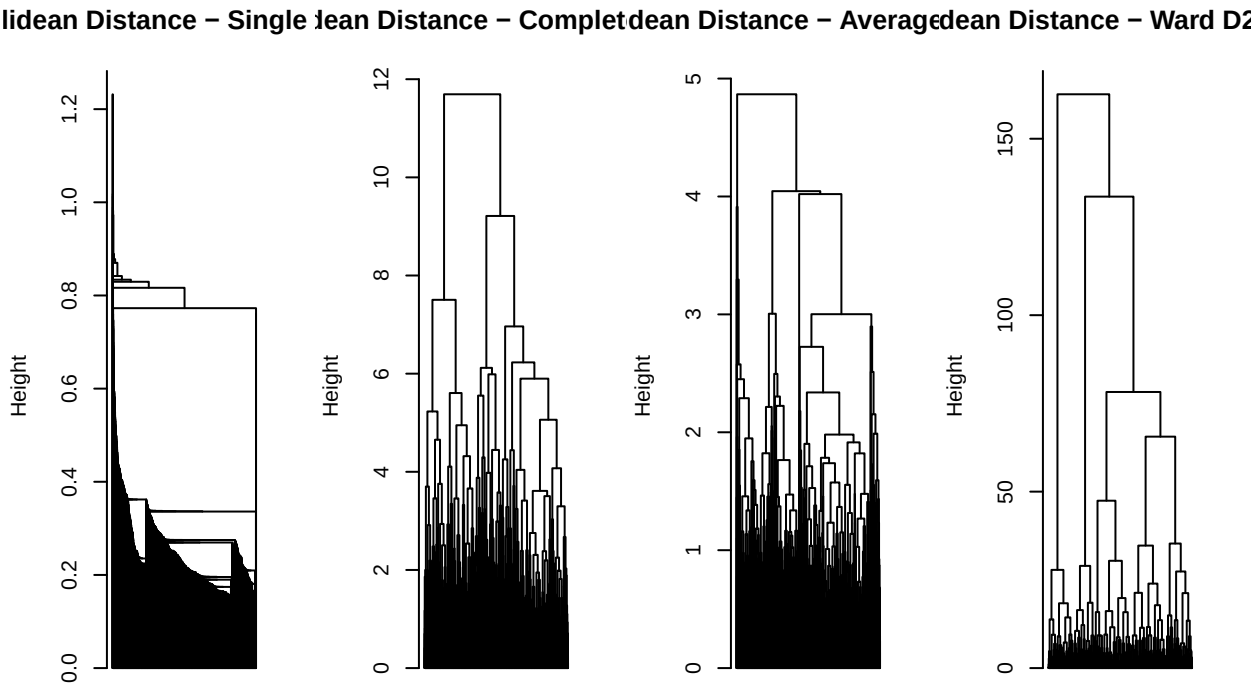
```
data_reduced_scaled = dnew
layout(matrix(c(2,3,1,3),2,byrow=T))
plot(pc.data, las=2, main='Principal Components', ylim=c(0,7))
abline(h=1, col='blue')
barplot(sapply(as.data.frame(data_reduced_scaled),sd)^2, las=2, main='Original Variables', ylim=c(0,7),
plot(cumsum(pc.data$sde^2)/sum(pc.data$sde^2), type='b', axes=F, xlab='Number of components', ylab='Con
abline(h=1, col='blue')
abline(h=0.62, lty=2, col='blue')
box()
axis(2,at=0:10/10,labels=0:10/10)
axis(1,at=1:ncol(data_reduced_scaled),labels=1:ncol(data_reduced_scaled),las=2)
```



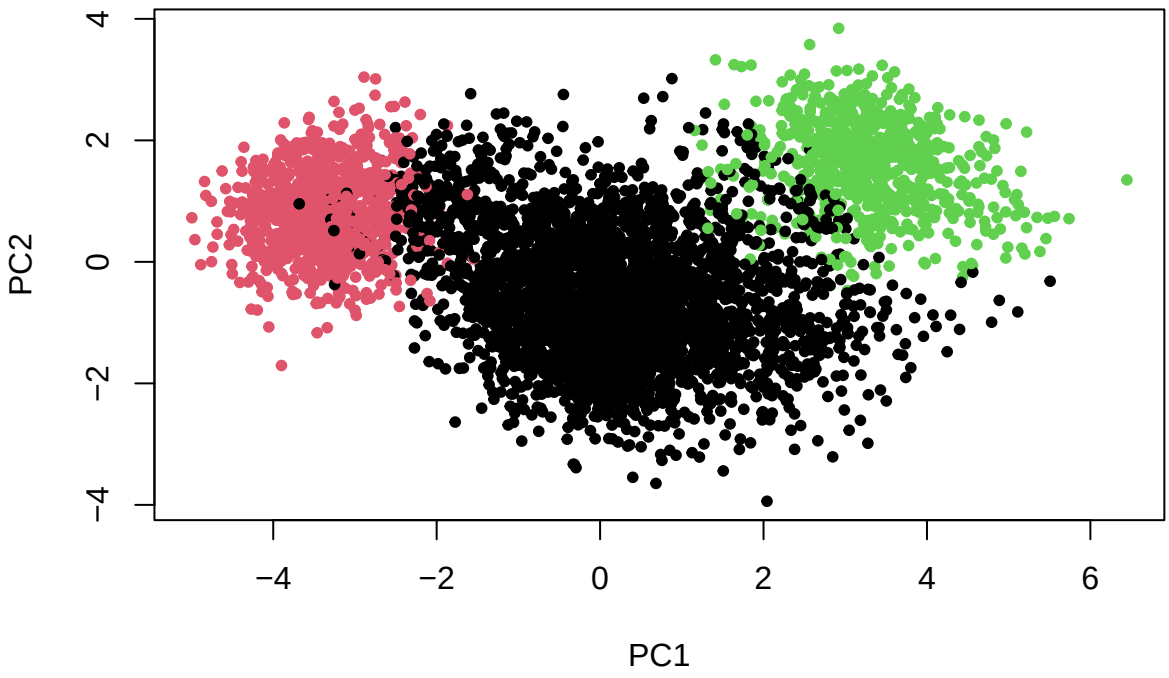

```
data_reduced_scaled = as.data.frame(data_reduced_scaled)

scores = pc.data$scores
par(mfrow=c(1,1))
plot(scores[,1], scores[,2], pch = 20, xlab="PC1", ylab="PC2")
```




```
plot3d(scores[,1], scores[,2], scores[,3])



# Hierarchical Clustering

dfNum.e <- dist(scores[,1:3], method='euclidean')
dfNum.es <- hclust(dfNum.e, method='single')
dfNum.ea <- hclust(dfNum.e, method='average')
dfNum.ec <- hclust(dfNum.e, method='complete')
dfNum.ew <- hclust(dfNum.e, method='ward.D2')

par(mfrow=c(1,4))
plot(dfNum.es, main='Euclidean Distance - Single Linkage', hang=-0.1, xlab='', labels=F, cex=0.6, sub='
plot(dfNum.ec, main='Euclidean Distance - Complete Linkage', hang=-0.1, xlab='', labels=F, cex=0.6, sub
plot(dfNum.ea, main='Euclidean Distance - Average Linkage', hang=-0.1, xlab='', labels=F, cex=0.6, sub=
plot(dfNum.ew, main='Euclidean Distance - Ward D2 Linkage', hang=-0.1, xlab='', labels=F, cex=0.6, sub=
```



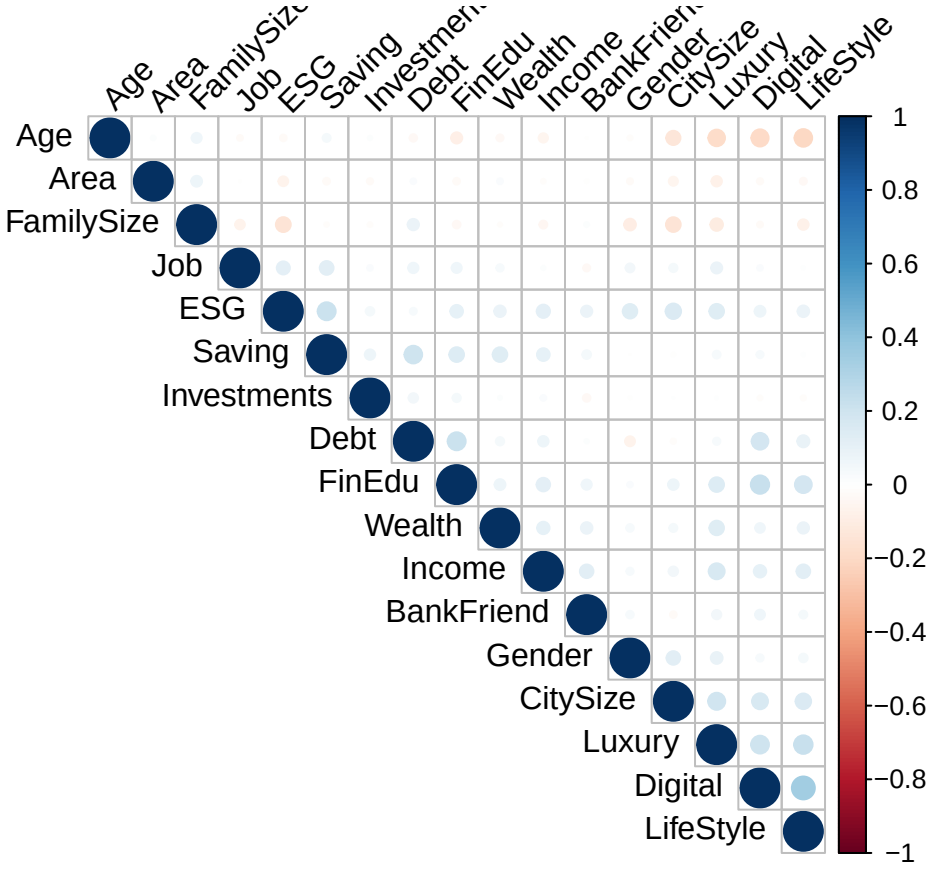

```r
cluster.ew.3 <- cutree(dfNum.ew, k=3)

par(mfrow=c(1,1))
plot(scores[,1], scores[,2], pch = 20, xlab="PC1", ylab="PC2", col=cluster.ew.3)
```



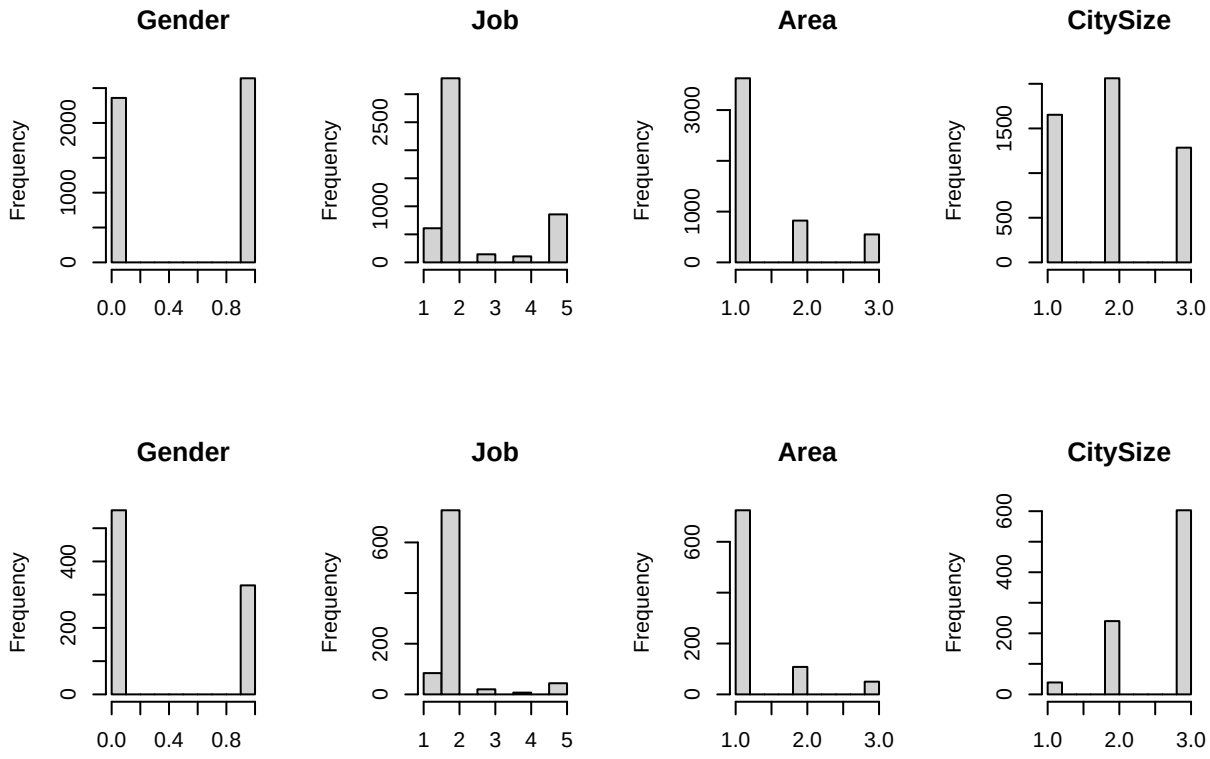

```
plot3d(scores[,1], scores[,2], scores[,3], col=cluster.ew.3)

feature = cluster.ew.3
num_clusters = 3

i1 = which(cluster.ew.3 == 1)
i2 = which(cluster.ew.3 == 2)
i3 = which(cluster.ew.3 == 3)
names = colnames(dnew)

table(cluster.ew.3)
```

```
## cluster.ew.3
##    1    2    3
## 3392  882  726
```

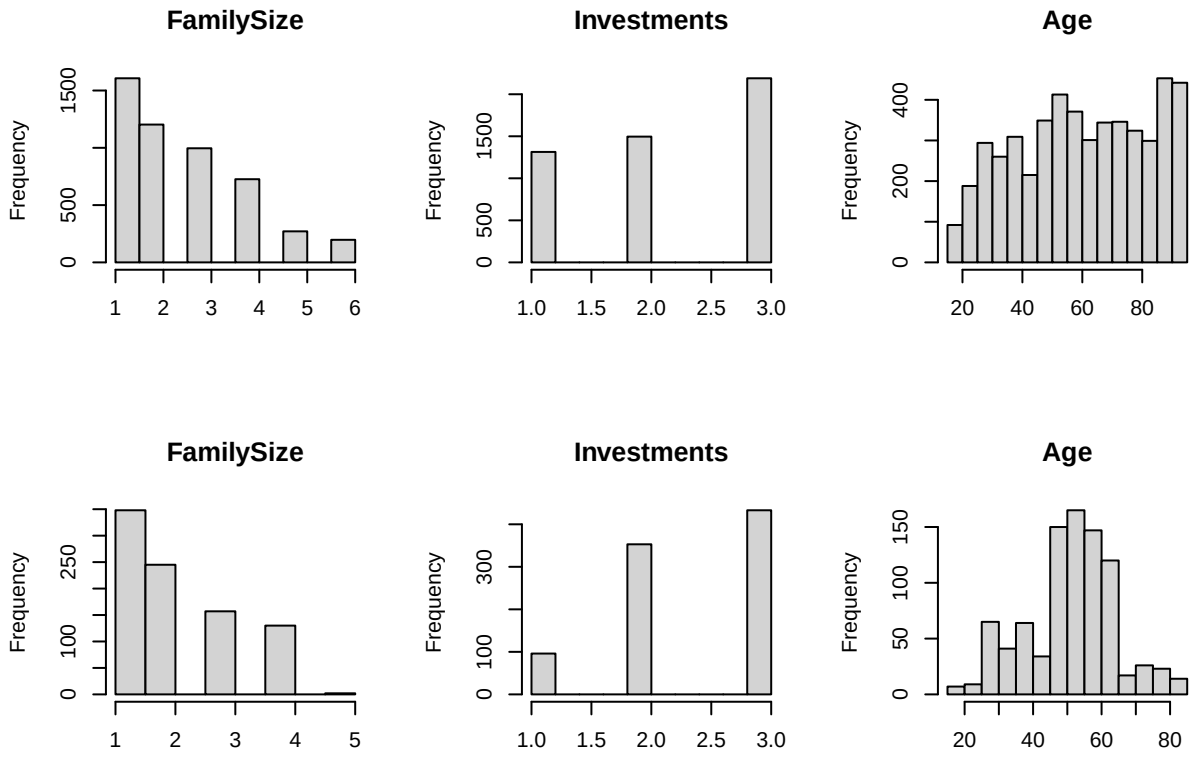
```
library(corrplot)
corrplot(cor(dnew[i1,]), type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



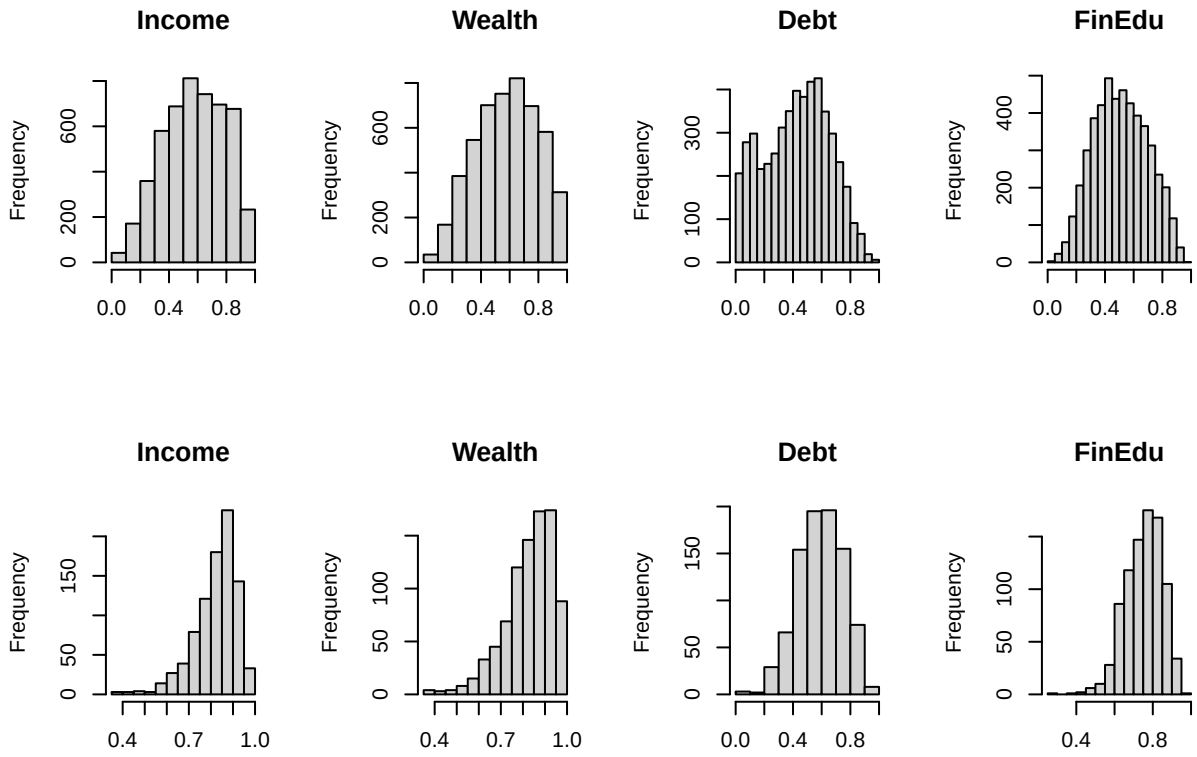

```
x11()
par(mfrow=c(2,4))
hist(data$Gender, freq=TRUE, main = "Gender", xlab = "")
hist(data$Job, freq=TRUE, main = "Job", xlab = "")
hist(data$Area, freq=TRUE, main = "Area", xlab = "")
hist(data$CitySize, freq=TRUE, main = "CitySize", xlab = "")
hist(data[i2,]$Gender, freq=TRUE, main = "Gender", xlab = "")
hist(data[i2,]$Job, freq=TRUE, main = "Job", xlab = "")
hist(data[i2,]$Area, freq=TRUE, main = "Area", xlab = "")
hist(data[i2,]$CitySize, freq=TRUE, main = "CitySize", xlab = "")
```



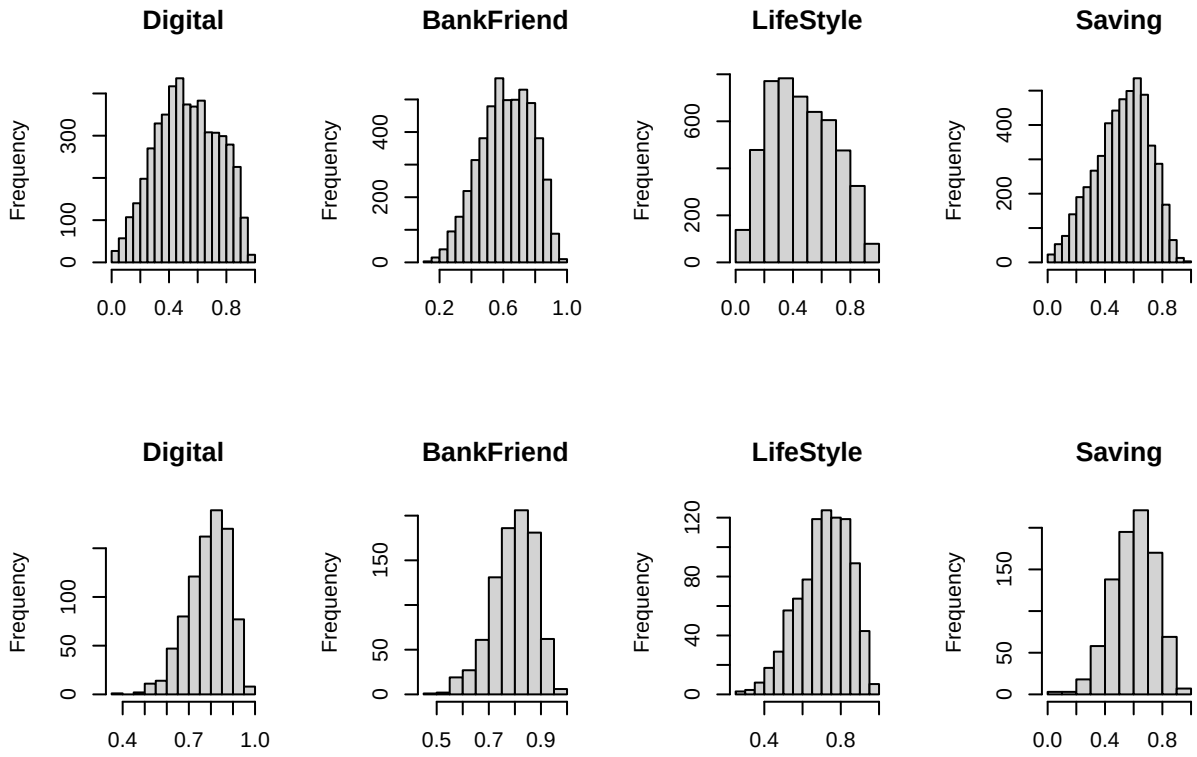

```
par(mfrow=c(2,3))
hist(data$FamilySize, freq=TRUE, main = "FamilySize", xlab = "")
hist(data$Investments, freq=TRUE, main = "Investments", xlab = "")
hist(data$Age, main = "Age", xlab = "")
hist(data[i2,]$FamilySize, freq=TRUE, main = "FamilySize", xlab = "")
hist(data[i2,]$Investments, freq=TRUE, main = "Investments", xlab = "")
hist(data[i2,]$Age, main = "Age", xlab = "")
```



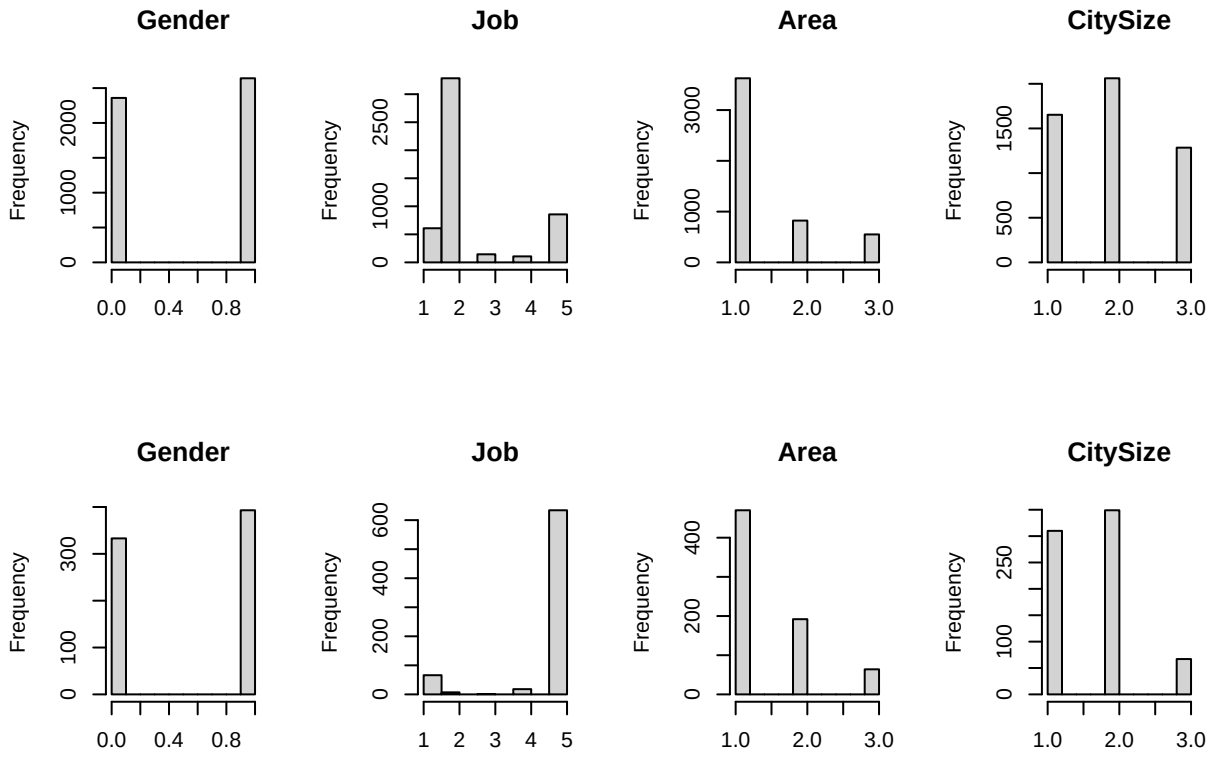

```
x11()
par(mfrow=c(2,4))
hist(data$Income, main = "Income", xlab = "")
hist(data$Wealth, main = "Wealth", xlab = "")
hist(data$Debt, main = "Debt", xlab = "")
hist(data$FinEdu, main = "FinEdu", xlab = "")
hist(data[i2,]$Income, main = "Income", xlab = "")
hist(data[i2,]$Wealth, main = "Wealth", xlab = "")
hist(data[i2,]$Debt, main = "Debt", xlab = "")
hist(data[i2,]$FinEdu, main = "FinEdu", xlab = "")
```



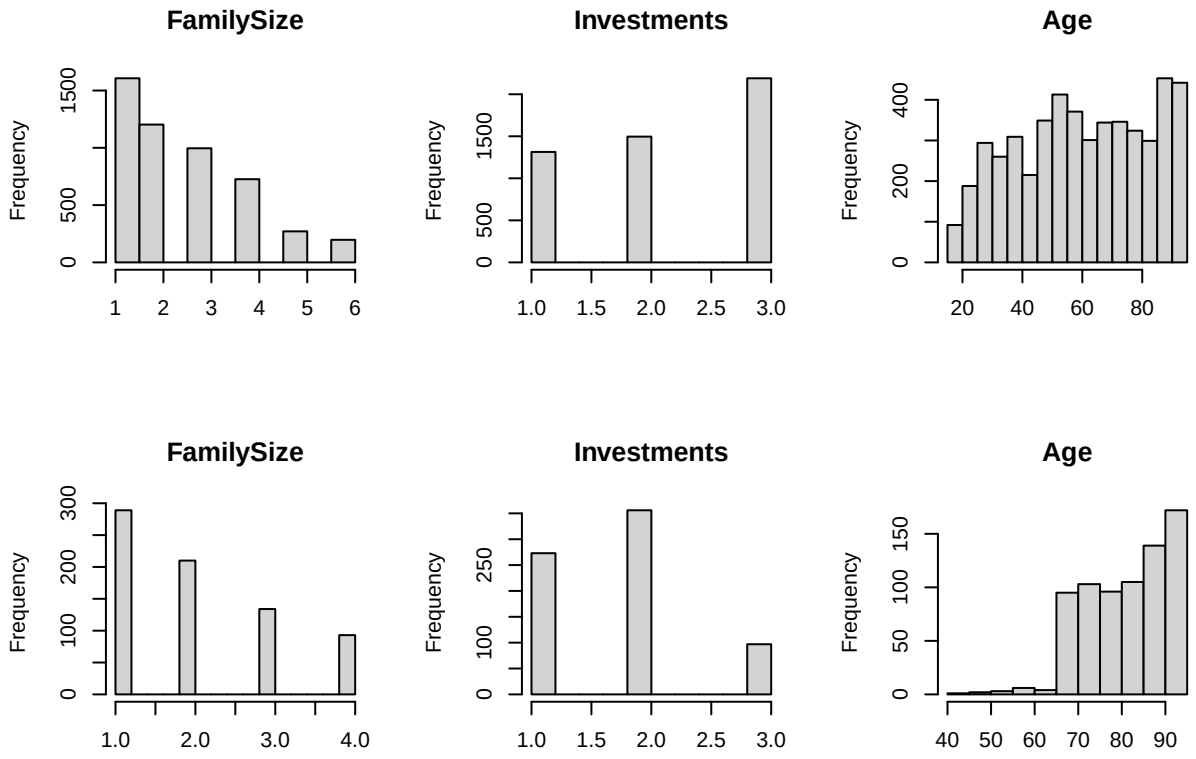

```
par(mfrow=c(2,4))
hist(data$Digital, main = "Digital", xlab = "")
hist(data$BankFriend, main = "BankFriend", xlab = "")
hist(data$LifeStyle, main = "LifeStyle", xlab = "")
hist(data$Saving, main = "Saving", xlab = "")
hist(data[i2,]$Digital, main = "Digital", xlab = "")
hist(data[i2,]$BankFriend, main = "BankFriend", xlab = "")
hist(data[i2,]$LifeStyle, main = "LifeStyle", xlab = "")
hist(data[i2,]$Saving, main = "Saving", xlab = "")
```



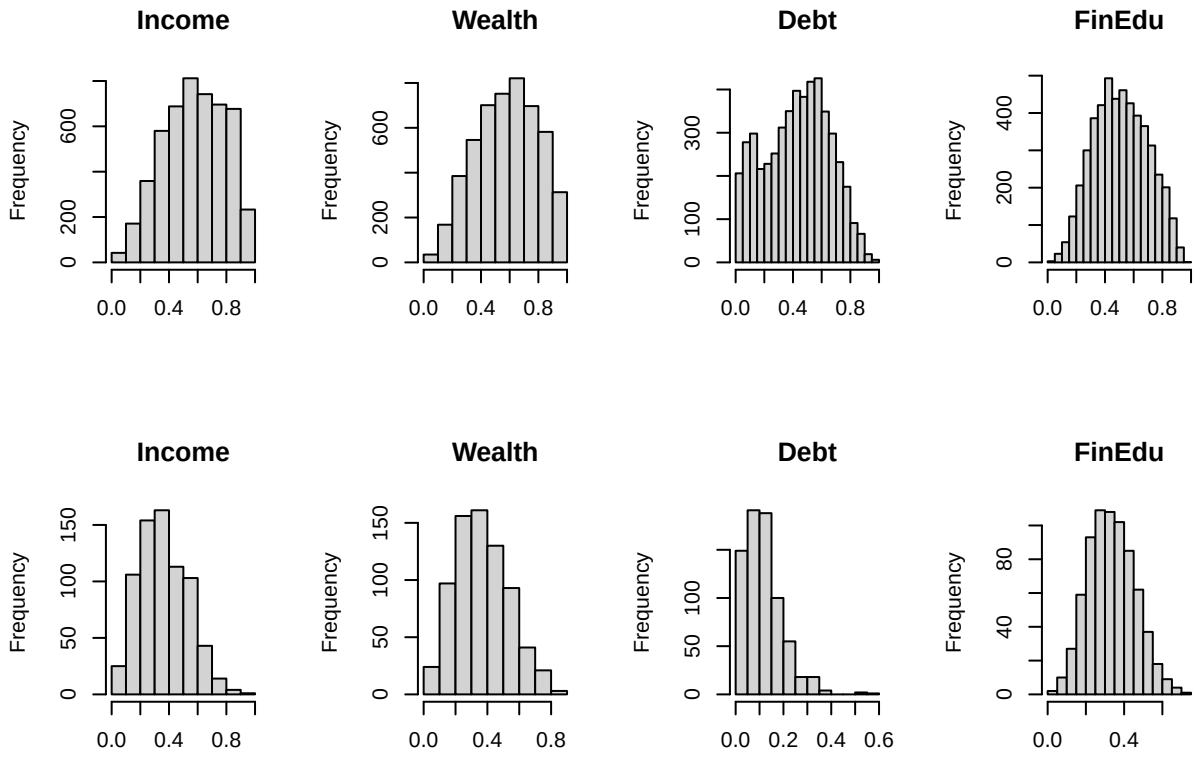

```
x11()
par(mfrow=c(2,4))
hist(data$Gender, freq=TRUE, main = "Gender", xlab = "")
hist(data$Job, freq=TRUE, main = "Job", xlab = "")
hist(data$Area, freq=TRUE, main = "Area", xlab = "")
hist(data$CitySize, freq=TRUE, main = "CitySize", xlab = "")
hist(data[i3,]$Gender, freq=TRUE, main = "Gender", xlab = "")
hist(data[i3,]$Job, freq=TRUE, main = "Job", xlab = "")
hist(data[i3,]$Area, freq=TRUE, main = "Area", xlab = "")
hist(data[i3,]$CitySize, freq=TRUE, main = "CitySize", xlab = "")
```



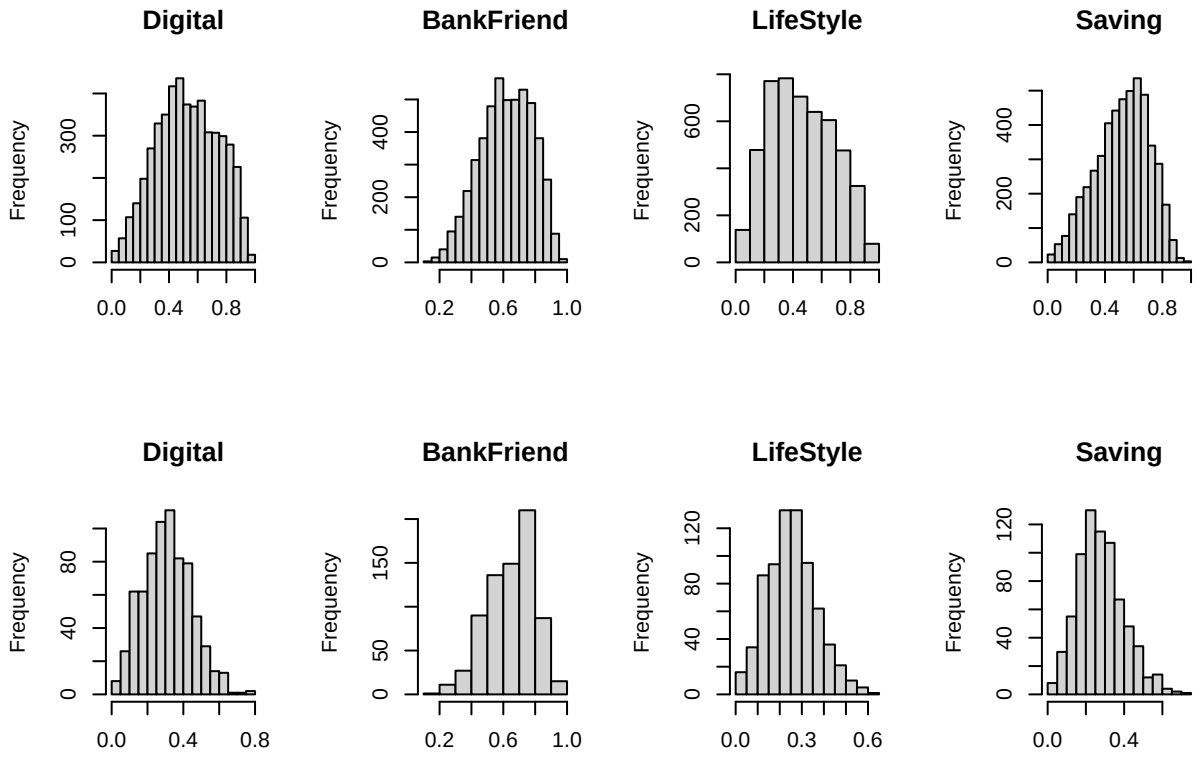

```
par(mfrow=c(2,3))
hist(data$FamilySize, freq=TRUE, main = "FamilySize", xlab = "")
hist(data$Investments, freq=TRUE, main = "Investments", xlab = "")
hist(data$Age, main = "Age", xlab = "")
hist(data[i3,]$FamilySize, freq=TRUE, main = "FamilySize", xlab = "")
hist(data[i3,]$Investments, freq=TRUE, main = "Investments", xlab = "")
hist(data[i3,]$Age, main = "Age", xlab = "")
```



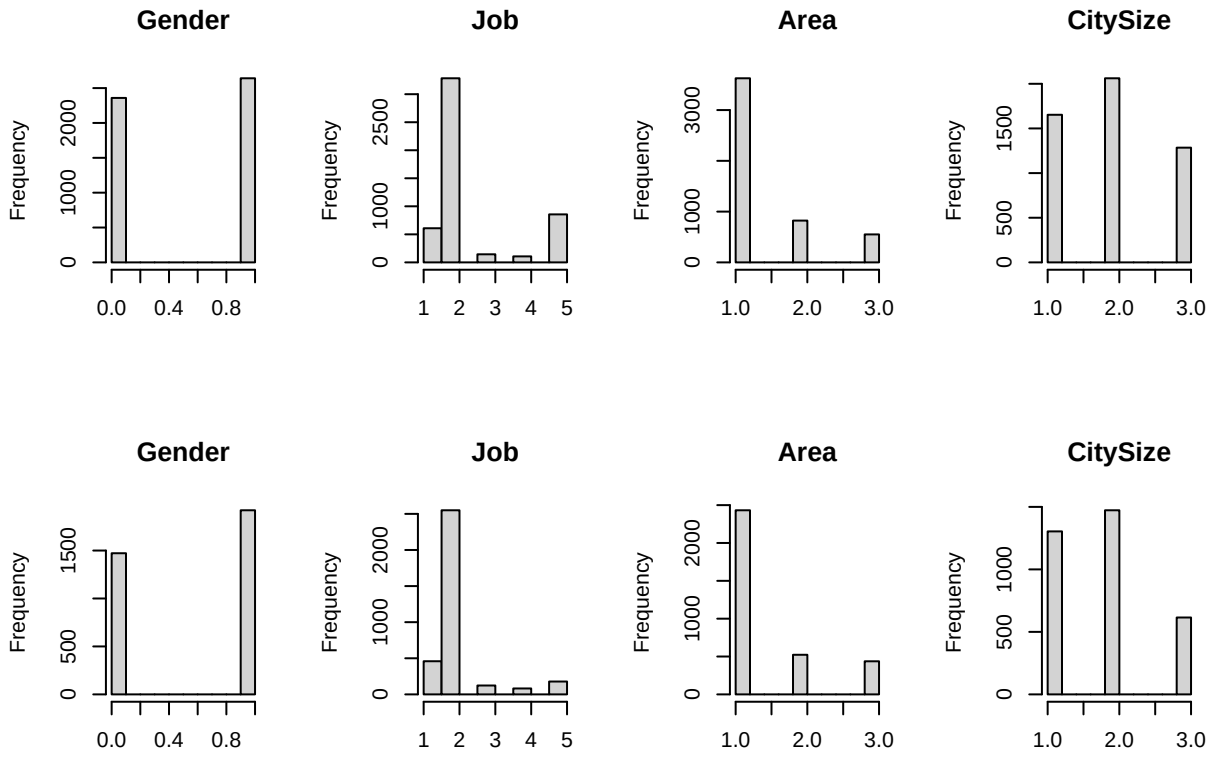

```
x11()
par(mfrow=c(2,4))
hist(data$Income, main = "Income", xlab = "")
hist(data$Wealth, main = "Wealth", xlab = "")
hist(data$Debt, main = "Debt", xlab = "")
hist(data$FinEdu, main = "FinEdu", xlab = "")
hist(data[i3,]$Income, main = "Income", xlab = "")
hist(data[i3,]$Wealth, main = "Wealth", xlab = "")
hist(data[i3,]$Debt, main = "Debt", xlab = "")
hist(data[i3,]$FinEdu, main = "FinEdu", xlab = "")
```



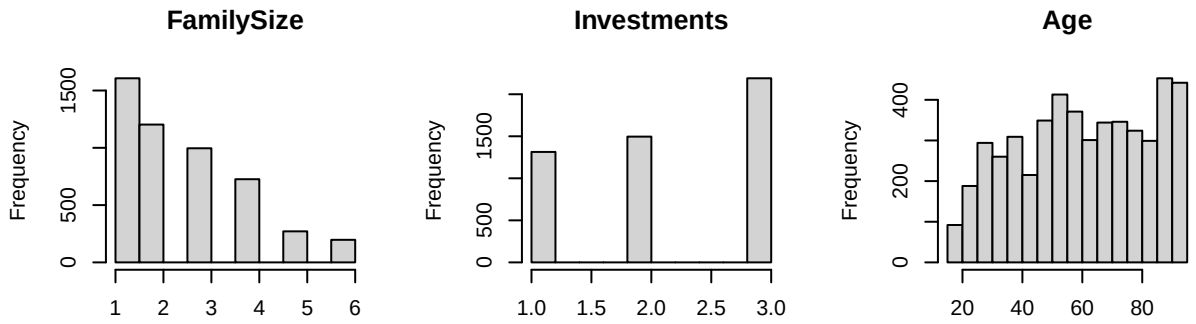

```
par(mfrow=c(2,4))
hist(data$Digital, main = "Digital", xlab = "")
hist(data$BankFriend, main = "BankFriend", xlab = "")
hist(data$LifeStyle, main = "LifeStyle", xlab = "")
hist(data$Saving, main = "Saving", xlab = "")
hist(data[i3,]$Digital, main = "Digital", xlab = "")
hist(data[i3,]$BankFriend, main = "BankFriend", xlab = "")
hist(data[i3,]$LifeStyle, main = "LifeStyle", xlab = "")
hist(data[i3,]$Saving, main = "Saving", xlab = "")
```

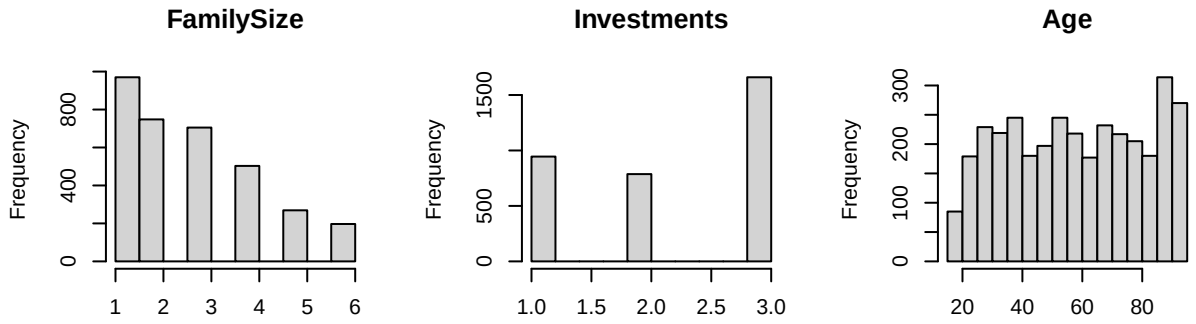



```
x11()
par(mfrow=c(2,4))
hist(data$Gender, freq=TRUE, main = "Gender", xlab = "")
hist(data$Job, freq=TRUE, main = "Job", xlab = "")
hist(data$Area, freq=TRUE, main = "Area", xlab = "")
hist(data$CitySize, freq=TRUE, main = "CitySize", xlab = "")
hist(data[i1,]$Gender, freq=TRUE, main = "Gender", xlab = "")
hist(data[i1,]$Job, freq=TRUE, main = "Job", xlab = "")
hist(data[i1,]$Area, freq=TRUE, main = "Area", xlab = "")
hist(data[i1,]$CitySize, freq=TRUE, main = "CitySize", xlab = "")
```

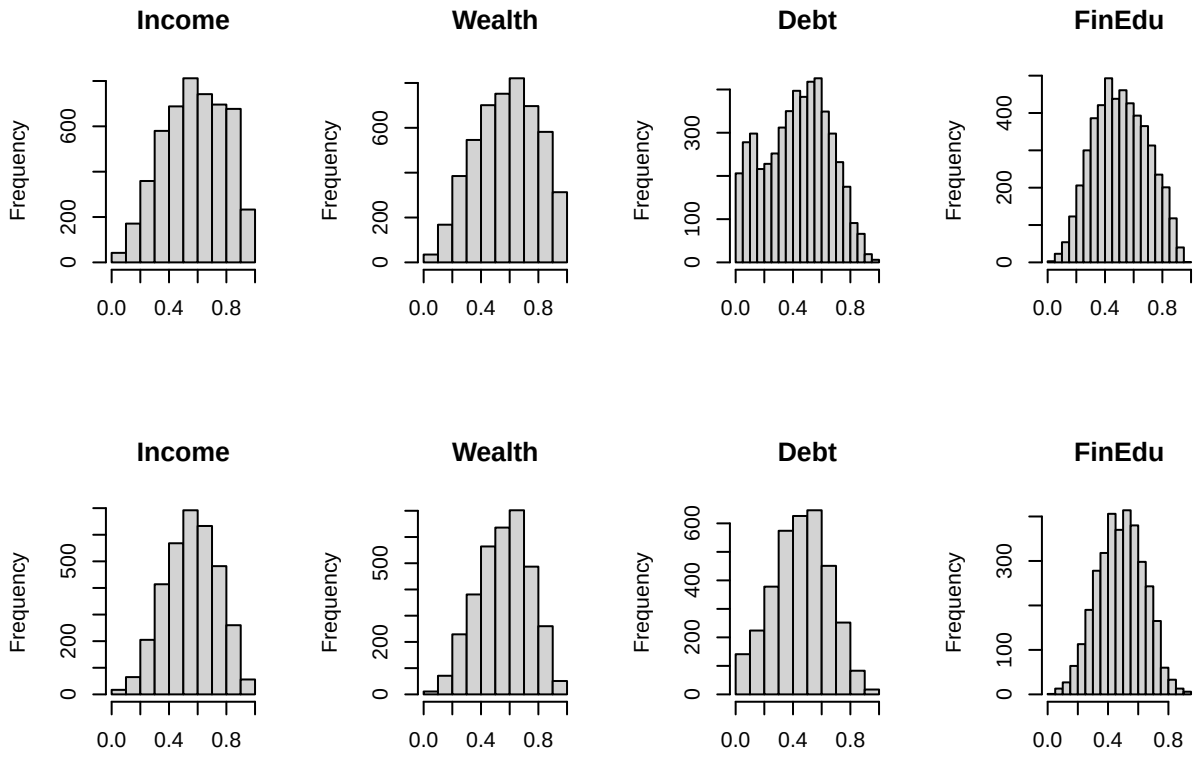



```
par(mfrow=c(2,3))
hist(data$FamilySize, freq=TRUE, main = "FamilySize", xlab = "")
hist(data$Investments, freq=TRUE, main = "Investments", xlab = "")
hist(data$Age, main = "Age", xlab = "")
hist(data[i1,]$FamilySize, freq=TRUE, main = "FamilySize", xlab = "")
hist(data[i1,]$Investments, freq=TRUE, main = "Investments", xlab = "")
hist(data[i1,]$Age, main = "Age", xlab = "")
```

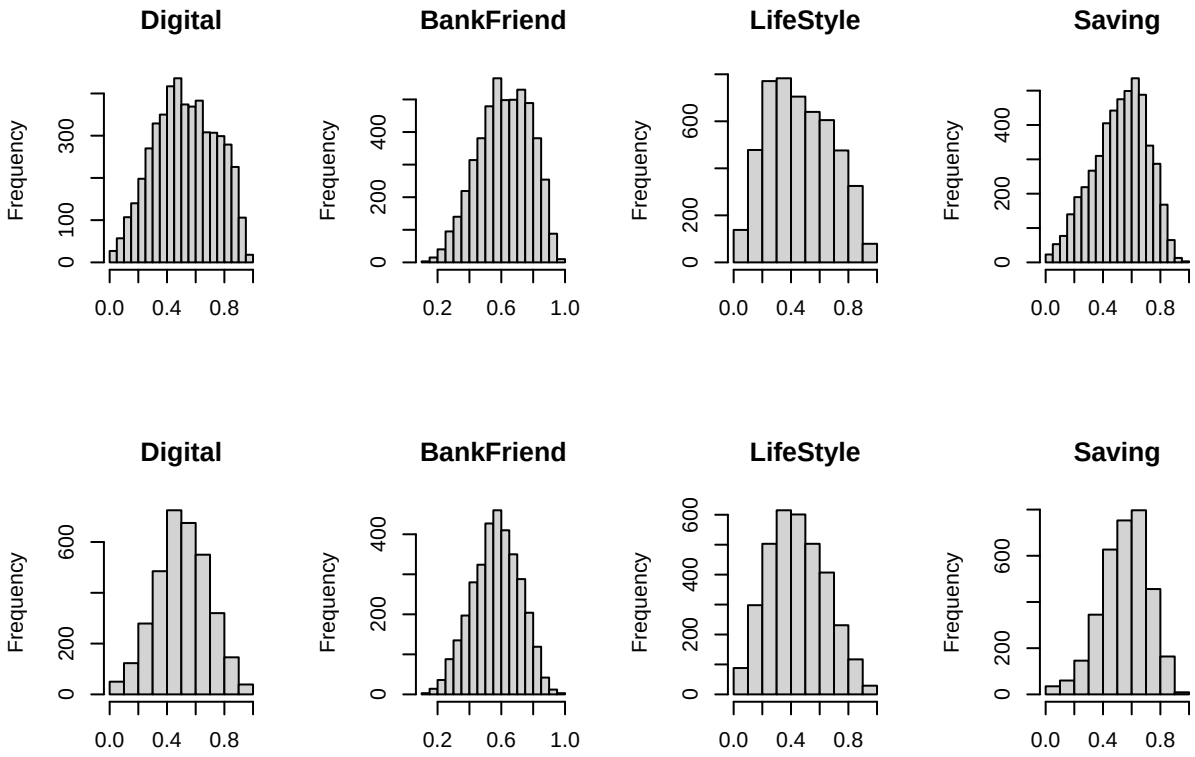





```
x11()
par(mfrow=c(2,4))
hist(data$Income, main = "Income", xlab = "")
hist(data$Wealth, main = "Wealth", xlab = "")
hist(data$Debt, main = "Debt", xlab = "")
hist(data$FinEdu, main = "FinEdu", xlab = "")


hist(data[i1,]$Income, main = "Income", xlab = "")
hist(data[i1,]$Wealth, main = "Wealth", xlab = "")
hist(data[i1,]$Debt, main = "Debt", xlab = "")
hist(data[i1,]$FinEdu, main = "FinEdu", xlab = "")
```




```
par(mfrow=c(2,4))
hist(data$Digital, main = "Digital", xlab = "")
hist(data$BankFriend, main = "BankFriend", xlab = "")
hist(data$LifeStyle, main = "LifeStyle", xlab = "")
hist(data$Saving, main = "Saving", xlab = "")
hist(data[i1,]$Digital, main = "Digital", xlab = "")
hist(data[i1,]$BankFriend, main = "BankFriend", xlab = "")
hist(data[i1,]$LifeStyle, main = "LifeStyle", xlab = "")
hist(data[i1,]$Saving, main = "Saving", xlab = "")
```




```
# pc_ <- princomp(scale(dnew[i1,]), scores=T)
# summary(pc_)
# load.data_ <- pc_$loadings
# scores_g1 = pc_$scores
#
# library(fpc)
#
# d = as.matrix(dist(scores_g1[,1:2], method='euclidean'))
# k = 200 # 100/150 works
# knee_plot = numeric(table(cluster.ew.3)[1])
# for (i in 1:table(cluster.ew.3)[1])
# {
#   d_i = as.numeric(sort(d[i,]))
#   knee_plot[i] = mean(d_i[1:k])
# }
# plot(sort(knee_plot), ylab = "")
#
# set.seed(220)  # Setting seed
# Dbscan_cl <- dbscan(scores_g1[,1:4], eps = 1, MinPts = k)
# # Dbscan_cl
# # Dbscan_cl$cluster
# table(Dbscan_cl$cluster)
#
# plot(scores[i1,1:2], col = "white", pch = 20)
# points(scores[i1,1:2])
# points(scores[i1,1:2], col = Dbscan_cl$cluster, pch = 20)
```

```
#
#
# d = as.matrix(dist(dnew[i1,-1], method='euclidean'))
# k = 150 # <= 90 100/150 works
# knee_plot = numeric(table(cluster.ew.3)[1])
# for (i in 1:table(cluster.ew.3)[1])
# {
#   d_i = as.numeric(sort(d[i,]))
#   knee_plot[i] = mean(d_i[1:k])
# }
# plot(sort(knee_plot), ylab="")
#
# set.seed(29061999)
# Dbscan_cl <- dbscan(dnew[i1,-1], eps = 2.7, MinPts = k)
# table(Dbscan_cl$cluster)
#
# plot3d(scores[i1,1:3], col = Dbscan_cl$cluster+1, pch = 20)
#
# group1 = dnew[i1,]
# group1_glob = data[i1,]
# i11 = which(Dbscan_cl$cluster == 1)
# i12 = which(Dbscan_cl$cluster == 2)
# i13 = which(Dbscan_cl$cluster == 3)
# i14 = which(Dbscan_cl$cluster == 4)
# i15 = which(Dbscan_cl$cluster == 5)
# i16 = which(Dbscan_cl$cluster == 6)
#i10 = which(Dbscan_cl$cluster == 0)



#pc.data <- princomp(scale(group1), scores=T)
#summary(pc.data)
#load.data <- pc.data$loadings
#x11()
#par(mfcol = c(4,3))
#for(i in 1:11) barplot(load.data[,i], ylim = c(-1, 1), main=paste("PC",i))
#x11()
#par(mfcol = c(1,3))
#for(i in 1:3) barplot(load.data[,i], ylim = c(-1, 1), main=paste("PC",i))
```