Final Project - Group 1: Melissa Theodore, Filipe Soares, Gabriel Fearon, Sina Aligholizadeh, Jagos Radovic

Set Directory, Importing data and libraries

```
setwd('C:/Users/filip/Desktop/Back To School/Data1010/Final Project')
adult<- read.csv('adult.csv', header = TRUE)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scales)
library(ggplot2)
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(ROCR)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(caTools)
library(rpart.plot)
```

```
## Loading required package: rpart
```

```
library(Metrics)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:Metrics':
##
##     precision, recall
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v tibble  3.0.3     v purrr   0.3.4
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------ tidyv
## x plyr::arrange()     masks dplyr::arrange()
## x readr::col_factor() masks scales::col_factor()
## x purrr::compact()    masks plyr::compact()
## x plyr::count()       masks dplyr::count()
## x purrr::discard()    masks scales::discard()
## x plyr::failwith()    masks dplyr::failwith()
## x dplyr::filter()     masks stats::filter()
## x plyr::id()          masks dplyr::id()
## x dplyr::lag()        masks stats::lag()
## x purrr::lift()       masks caret::lift()
## x plyr::mutate()      masks dplyr::mutate()
## x plyr::rename()      masks dplyr::rename()
## x plyr::summarise()   masks dplyr::summarise()
## x plyr::summarize()   masks dplyr::summarize()
```

```
library(cluster)
library(readr)
library(Rtsne)
library(rpart.plot)
library(ISLR)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(NbClust)
library(knitr)


#Increase memory limit

memory.limit(size = 100000)
```

```
## [1] 1e+05
```

DATA UNDERSTANDING

```
str(adult)
```

```
## 'data.frame':    32561 obs. of  15 variables:
```

```
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : chr  "?" "Private" "?" "Private" ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...
##  $ education     : chr  "HS-grad" "HS-grad" "Some-college" "7th-8th" ...
##  $ education.num : int  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: chr  "Widowed" "Widowed" "Widowed" "Divorced" ...
##  $ occupation    : chr  "?" "Exec-managerial" "?" "Machine-op-inspct" ...
##  $ relationship  : chr  "Not-in-family" "Not-in-family" "Unmarried" "Unmarried" ...
##  $ race          : chr  "White" "White" "Black" "White" ...
##  $ sex           : chr  "Female" "Female" "Female" "Female" ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 4356 4356 3900 3900 3770 3770 3683 3683 3004 ...
##  $ hours.per.week: int  40 18 40 40 40 45 40 20 40 60 ...
##  $ native.country: chr  "United-States" "United-States" "United-States" "United-States" ...
##  $ income        : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
```

```
summary(adult)
```

```
##       age          workclass              fnlwgt         education
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.58                      Mean   : 189778
##  3rd Qu.:48.00                      3rd Qu.: 237051
##  Max.   :90.00                      Max.   :1484705
##  education.num   marital.status      occupation        relationship
##  Min.   : 1.00   Length:32561       Length:32561       Length:32561
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.08
##  3rd Qu.:12.00
##  Max.   :16.00
##      race               sex             capital.gain    capital.loss
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0
##                                        Mean   : 1078   Mean   :  87.3
##                                        3rd Qu.:    0   3rd Qu.:   0.0
##                                        Max.   :99999   Max.   :4356.0
##  hours.per.week  native.country        income
##  Min.   : 1.00   Length:32561       Length:32561
##  1st Qu.:40.00   Class :character   Class :character
##  Median :40.00   Mode  :character   Mode  :character
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
```

```
colSums(is.na(adult))
```

```
##            age      workclass         fnlwgt      education  education.num
##              0              0              0              0              0
## marital.status     occupation   relationship           race            sex
##              0              0              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0              0              0
```

```
colSums(adult=="")
```

```
##            age      workclass         fnlwgt      education  education.num
##              0              0              0              0              0
## marital.status     occupation   relationship           race            sex
##              0              0              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0              0              0
```

```
colSums(adult=="?")
```

```
##            age      workclass         fnlwgt      education  education.num
##              0           1836              0              0              0
## marital.status     occupation   relationship           race            sex
##              0           1843              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0            583              0
```

```
# Convert "?" to NA, remove NA. Create data frames which will have numeric values and factor values

adult_na<- replace(adult, adult == "?", NA)
adult_num<-  na.omit(adult_na)
adult_fact<- adult_num

# Set Non-numeric to factors

fact <- c(2,4,6,7,8,9,10,14,15)
adult_num[,fact] <- lapply(adult_num[,fact] , factor)
str(adult_num)
```

```
## 'data.frame':    30162 obs. of  15 variables:
##  $ age           : int  82 54 41 34 38 74 68 45 38 52 ...
##  $ workclass     : Factor w/ 7 levels "Federal-gov",..: 3 3 3 3 3 6 1 3 5 3 ...
##  $ fnlwgt        : int  132870 140359 264663 216864 150601 88638 422013 172274 164526 129177 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 6 16 12 1 11 12 11 15 10 ...
##  $ education.num : int  9 4 10 9 6 16 9 16 15 13 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 1 6 1 6 5 1 1 5 7 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical",..: 4 7 10 8 1 10 10 10 10 8 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 5 4 5 5 3 2 5 2 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 5 5 5 5 3 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 1 1 2 1 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3900 3770 3770 3683 3683 3004 2824 2824 ...
##  $ hours.per.week: int  18 40 40 45 40 20 40 35 45 20 ...
##  $ native.country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 39 39 39 39 39 39 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 2 1 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:2399] 1 3 10 15 19 25 45 49 50 66 ...
##   ..- attr(*, "names")= chr [1:2399] "1" "3" "10" "15" ...
```

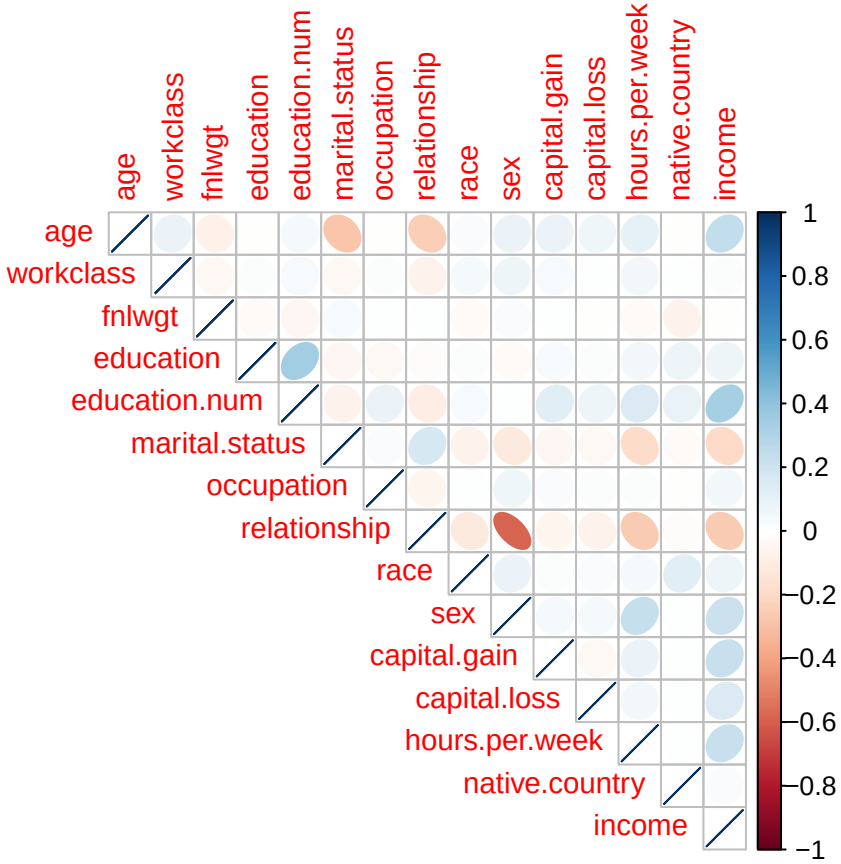
```
# convert factors into numeric for hierarchical clustering and corrplot
adult_num[, c(2,4,6:10,14,15)] <- sapply(adult_num[, c(2,4,6:10,14,15)], as.numeric)

str(adult_num)
```

```
## 'data.frame':    30162 obs. of  15 variables:
##  $ age           : int  82 54 41 34 38 74 68 45 38 52 ...
```

```
##  $ workclass     : num  3 3 3 3 3 6 1 3 5 3 ...
##  $ fnlwgt        : int  132870 140359 264663 216864 150601 88638 422013 172274 164526 129177 ...
##  $ education     : num  12 6 16 12 1 11 12 11 15 10 ...
##  $ education.num : int  9 4 10 9 6 16 9 16 15 13 ...
##  $ marital.status: num  7 1 6 1 6 5 1 1 5 7 ...
##  $ occupation    : num  4 7 10 8 1 10 10 10 10 8 ...
##  $ relationship  : num  2 5 4 5 5 3 2 5 2 2 ...
##  $ race          : num  5 5 5 5 5 5 5 3 5 5 ...
##  $ sex           : num  1 1 1 1 2 1 1 1 2 1 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3900 3770 3770 3683 3683 3004 2824 2824 ...
##  $ hours.per.week: int  18 40 40 45 40 20 40 35 45 20 ...
##  $ native.country: num  39 39 39 39 39 39 39 39 39 39 ...
##  $ income        : num  1 1 1 1 1 2 1 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:2399] 1 3 10 15 19 25 45 49 50 66 ...
##   ..- attr(*, "names")= chr [1:2399] "1" "3" "10" "15" ...
```

```
# Correlation Matrix
corrplot(cor(adult_num), type = "upper", method = "ellipse", tl.cex = 0.9)
```




```
# Convert character values to factors for Gower Dist/Matrix
adult_fact[,fact] <- lapply(adult_fact[,fact] , factor)
str(adult_fact)
```

```
## 'data.frame':    30162 obs. of  15 variables:
##  $ age           : int  82 54 41 34 38 74 68 45 38 52 ...
##  $ workclass     : Factor w/ 7 levels "Federal-gov",..: 3 3 3 3 3 6 1 3 5 3 ...
```

```
##  $ fnlwgt        : int  132870 140359 264663 216864 150601 88638 422013 172274 164526 129177 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 6 16 12 1 11 12 11 15 10 ...
##  $ education.num : int  9 4 10 9 6 16 9 16 15 13 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 1 6 1 6 5 1 1 5 7 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical",..: 4 7 10 8 1 10 10 10 10 8 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 5 4 5 5 3 2 5 2 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 5 5 5 5 3 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 1 1 2 1 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3900 3770 3770 3683 3683 3004 2824 2824 ...
##  $ hours.per.week: int  18 40 40 45 40 20 40 35 45 20 ...
##  $ native.country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 39 39 39 39 39 39 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 2 1 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:2399] 1 3 10 15 19 25 45 49 50 66 ...
##   ..- attr(*, "names")= chr [1:2399] "1" "3" "10" "15" ...
```

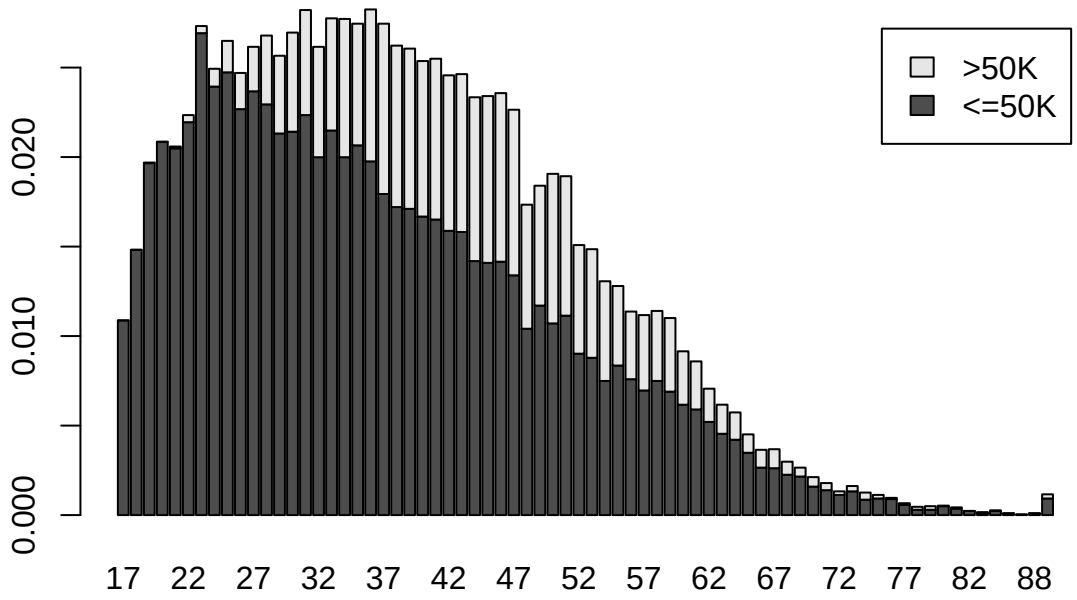
```
# Remove Unimportant Variables with regards to INCOME
adult_fact<- select(adult_fact, -c(fnlwgt, education, native.country, workclass, occupation, race))
adult_num<- select(adult_num, -c(fnlwgt, education, native.country, workclass, occupation, race))
```

###Visualizations *We removed the rows containing "?"*

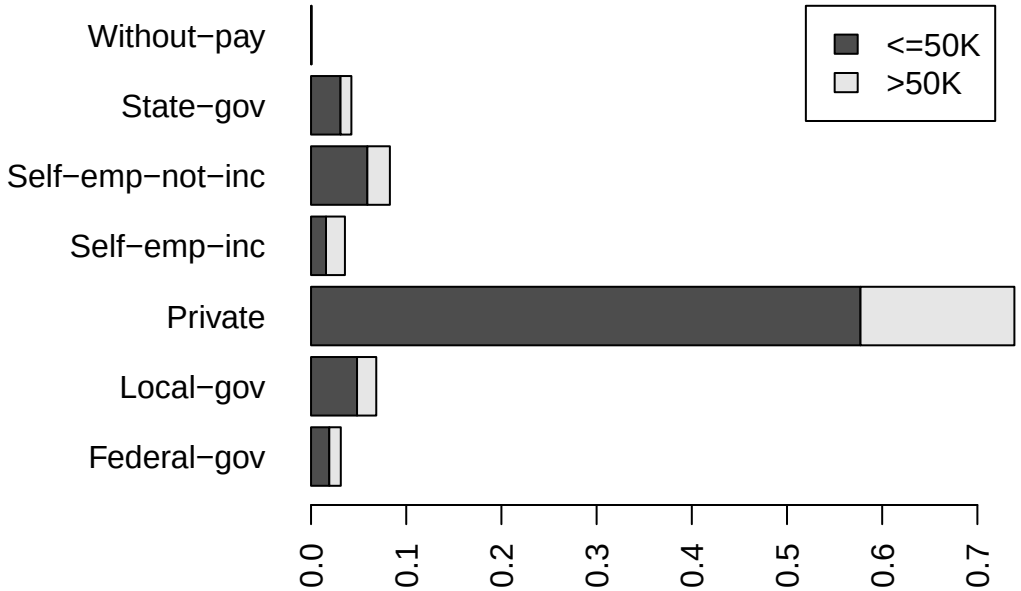
```
adult2=subset(adult, workclass!="?" & occupation!="?" & native.country!="?")
```

We used a combination of proportion barplots and mosaic plots with standardized residuals to derive visual insights from the data, focusing on the relationship of "income" feature with other features

```
freq_age=xtabs(~adult2$income+adult2$age)
barplot(prop.table(freq_age),legend=rownames(freq_age))
```

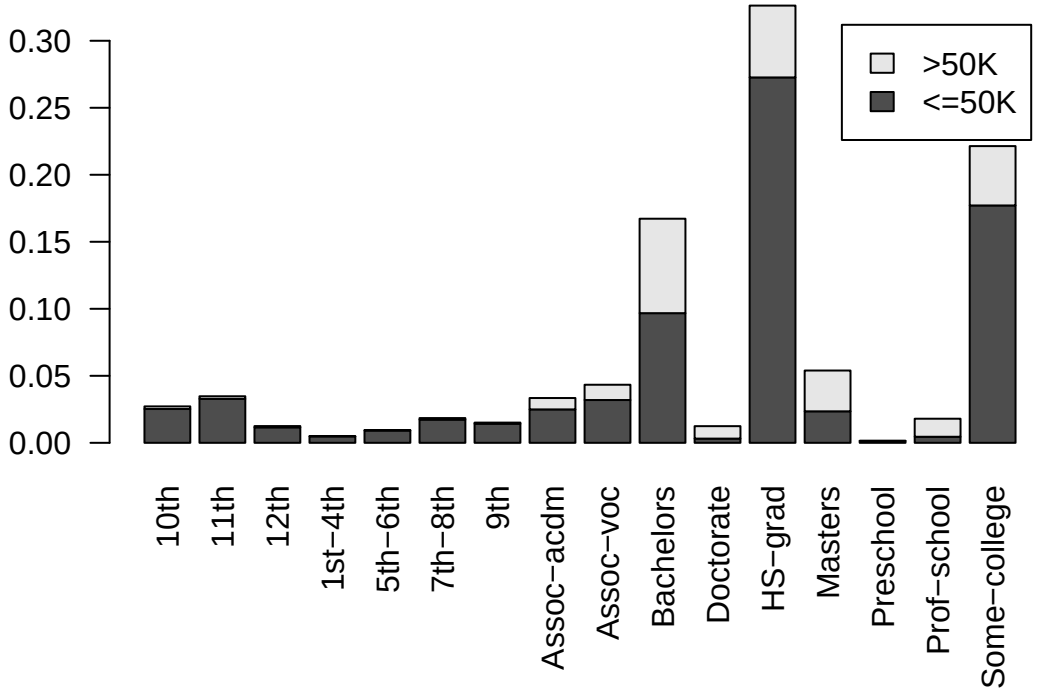



```
freq_workclass=xtabs(~adult2$income+adult2$workclass)
par(mar=c(5,10,4,4))
barplot(prop.table(freq_workclass),legend=rownames(freq_workclass),horiz=TRUE, las=2)
```

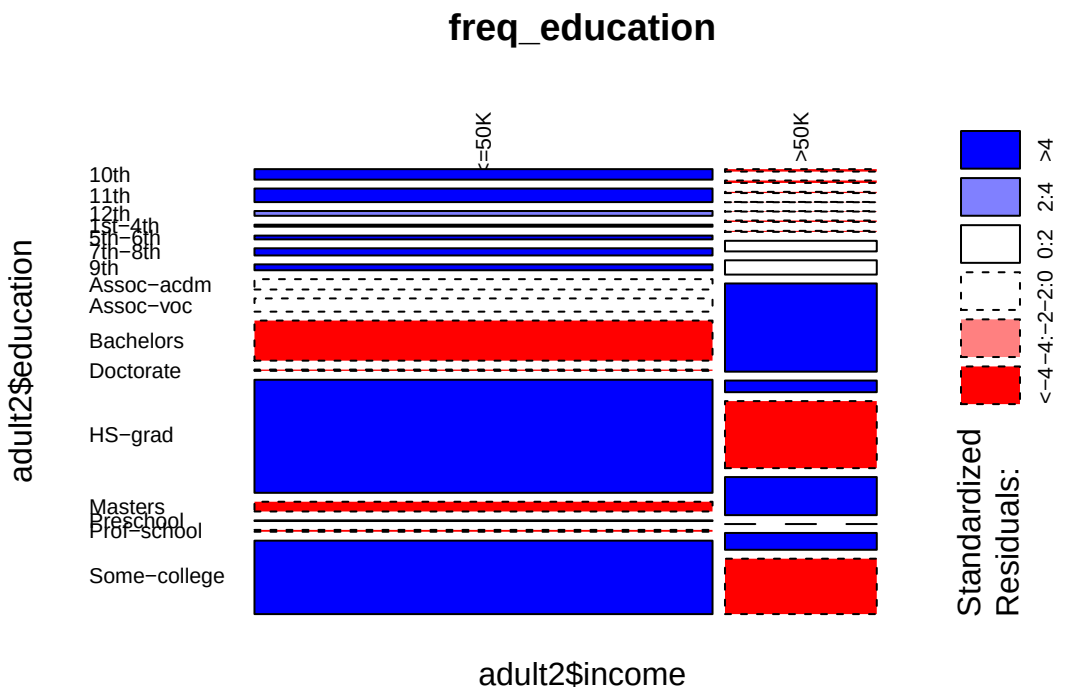



```
mosaicplot(freq_workclass, border = "black",
           shade = TRUE, las=2)
```

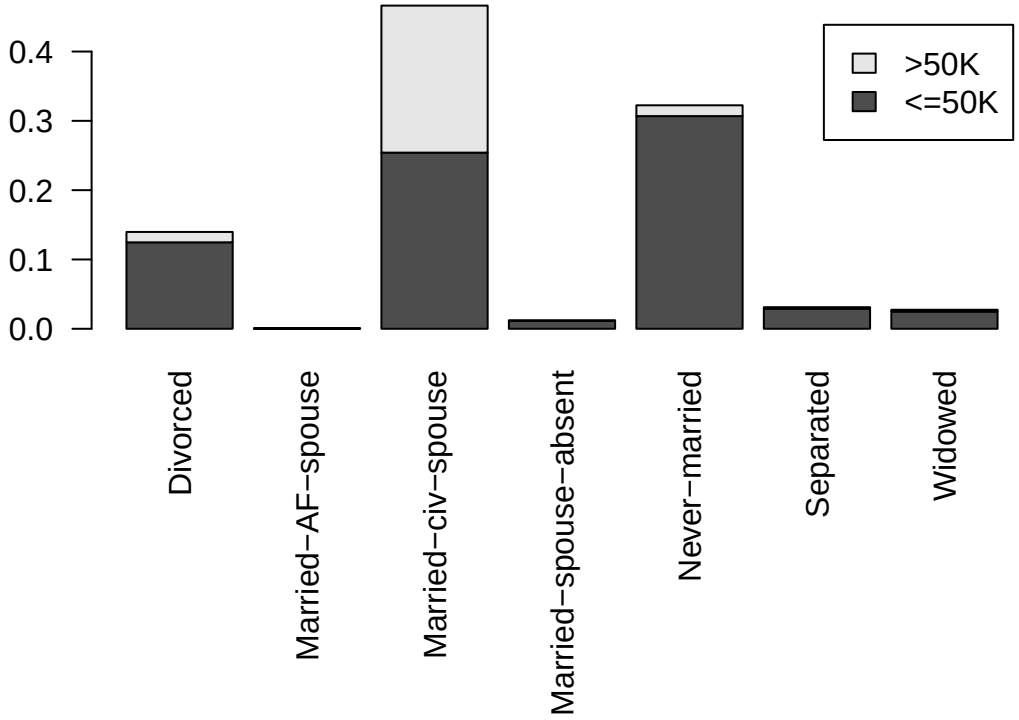

```
freq_education=xtabs(~adult2$income+adult2$education)
par(mar=c(7,4,4,4))
barplot(prop.table(freq_education),legend=rownames(freq_education),las=2)
```



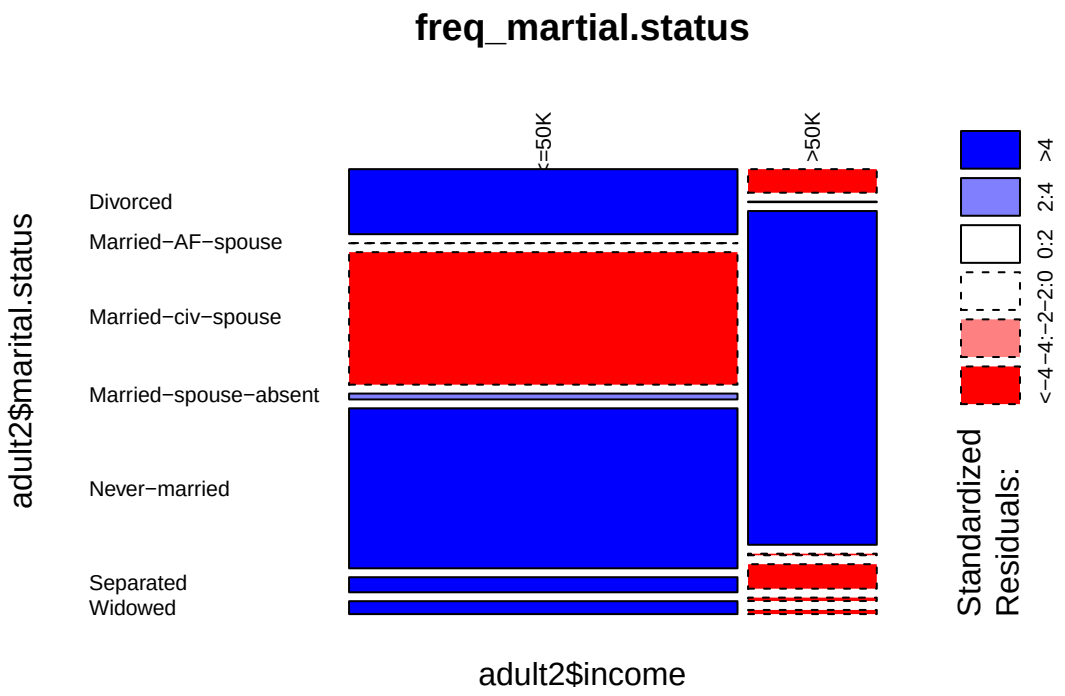

```
mosaicplot(freq_education, border = "black",
           shade = TRUE, las=2)
```



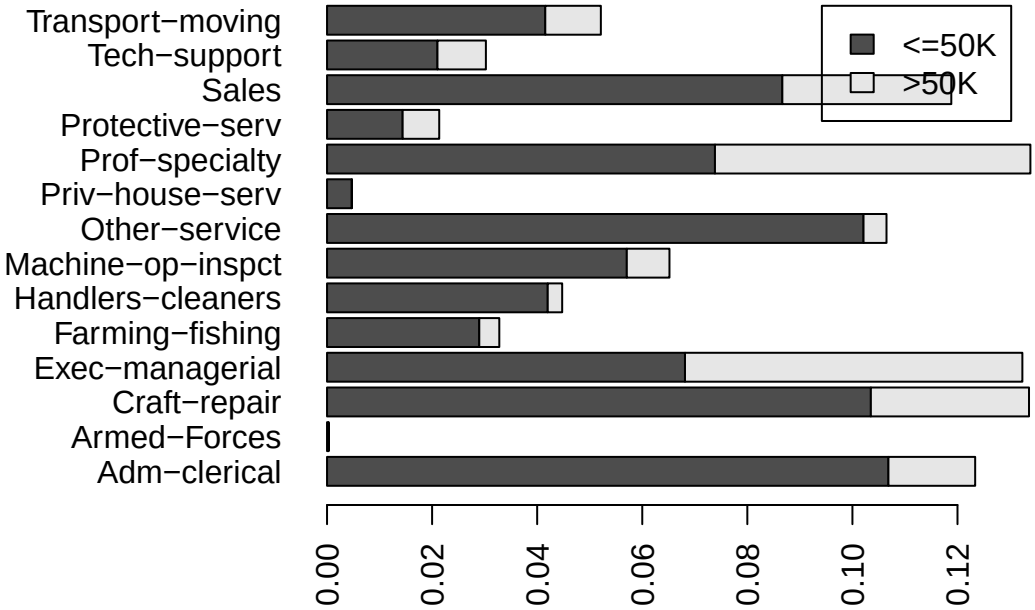

```
freq_martial.status=xtabs(~adult2$income+adult2$marital.status)
par(mar=c(10,4,4,4))
barplot(prop.table(freq_martial.status),legend=rownames(freq_martial.status),las=2)
```



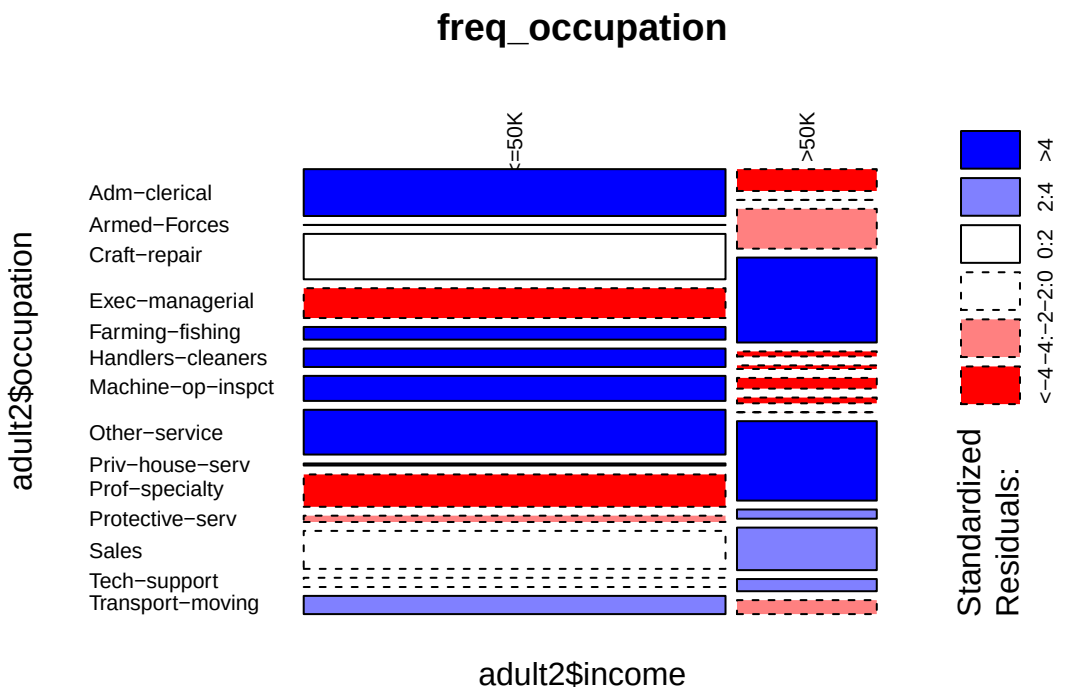

```
mosaicplot(freq_martial.status, border = "black",
           shade = TRUE, las=2)
```



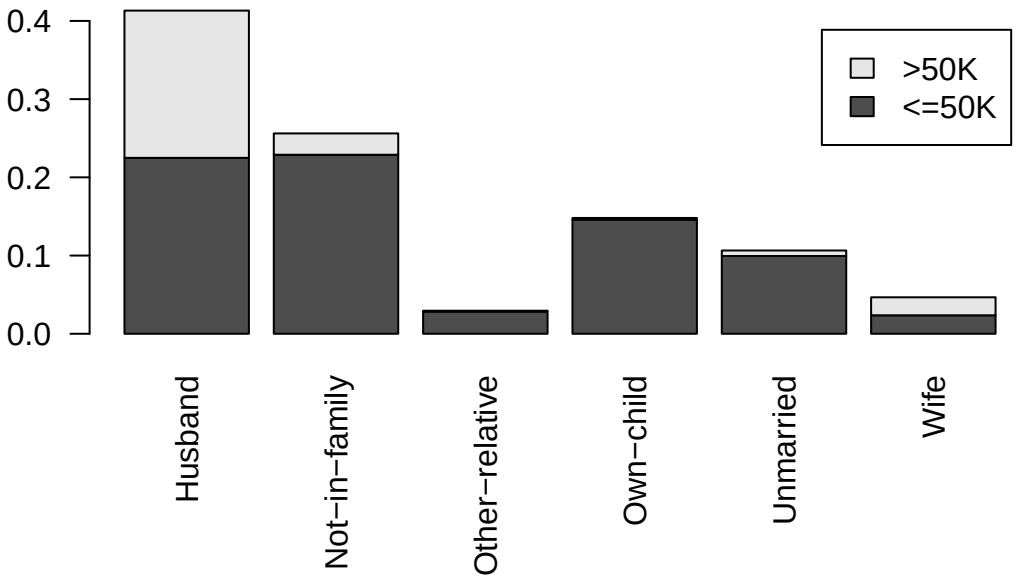

```
freq_occupation=xtabs(~adult2$income+adult2$occupation)
par(mar=c(5,10,4,4))
barplot(prop.table(freq_occupation),legend=rownames(freq_occupation),horiz=TRUE, las=2)
```



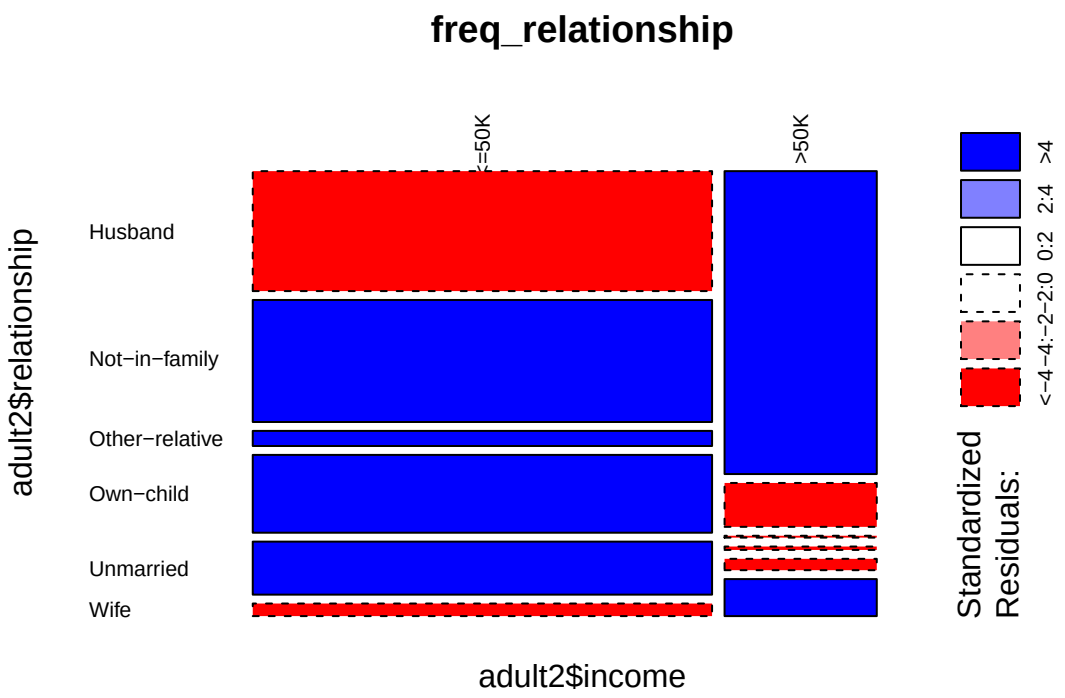

```
mosaicplot(freq_occupation, border = "black",
           shade = TRUE, las=2)
```




```
freq_relationship=xtabs(~adult2$income+adult2$relationship)
par(mar=c(10,4,4,4))
barplot(prop.table(freq_relationship),legend=rownames(freq_relationship),las=2)
```




```
mosaicplot(freq_relationship, border = "black",
           shade = TRUE, las=2)
```



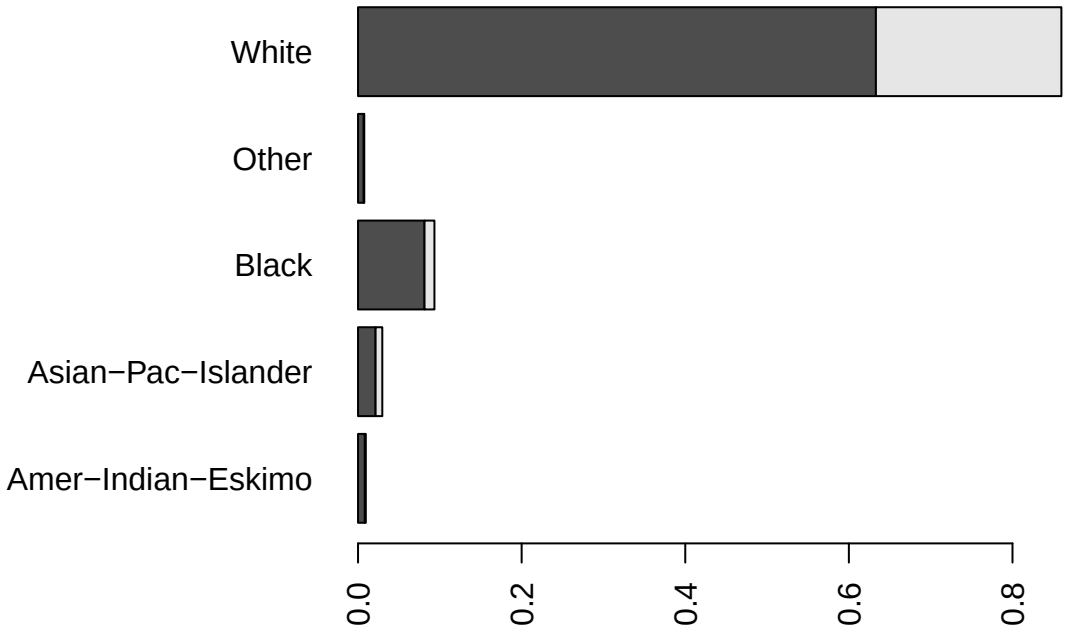

```
freq_race=xtabs(~adult2$income+adult2$race)
par(mar=c(4,10,4,4))
barplot(prop.table(freq_race),legend=rownames(freq_race),horiz=TRUE, las=2, legend.text=FALSE)
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "legend" is not a graphical
## parameter
```

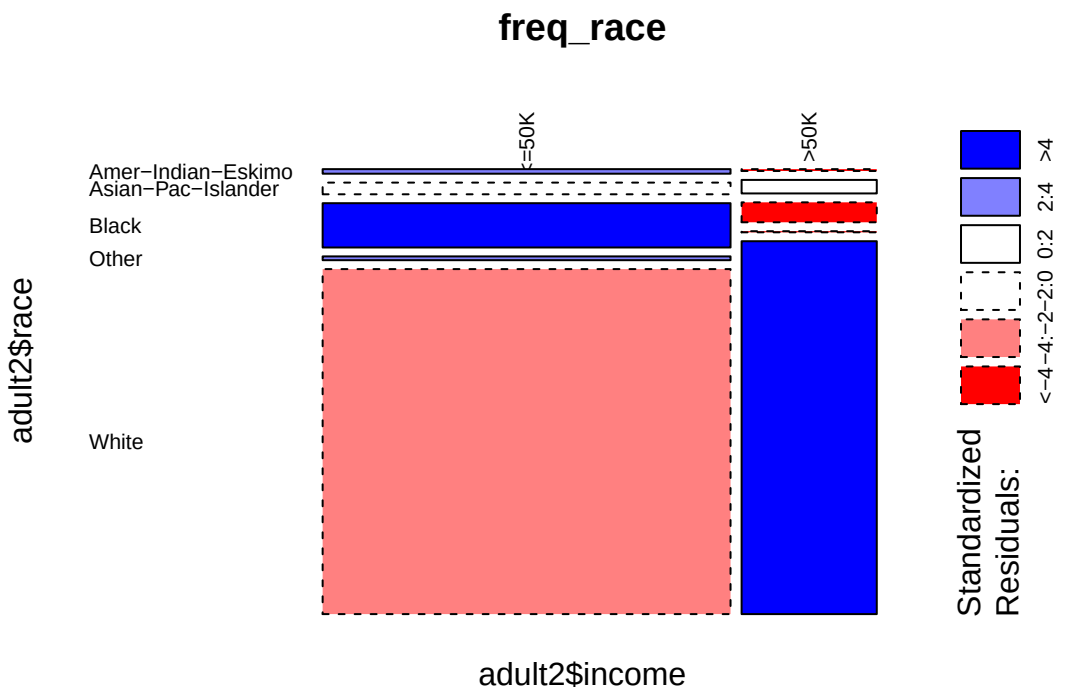
```
## Warning in axis(if (horiz) 2 else 1, at = at.l, labels = names.arg, lty =
## axis.lty, : "legend" is not a graphical parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "legend" is not a graphical parameter
```

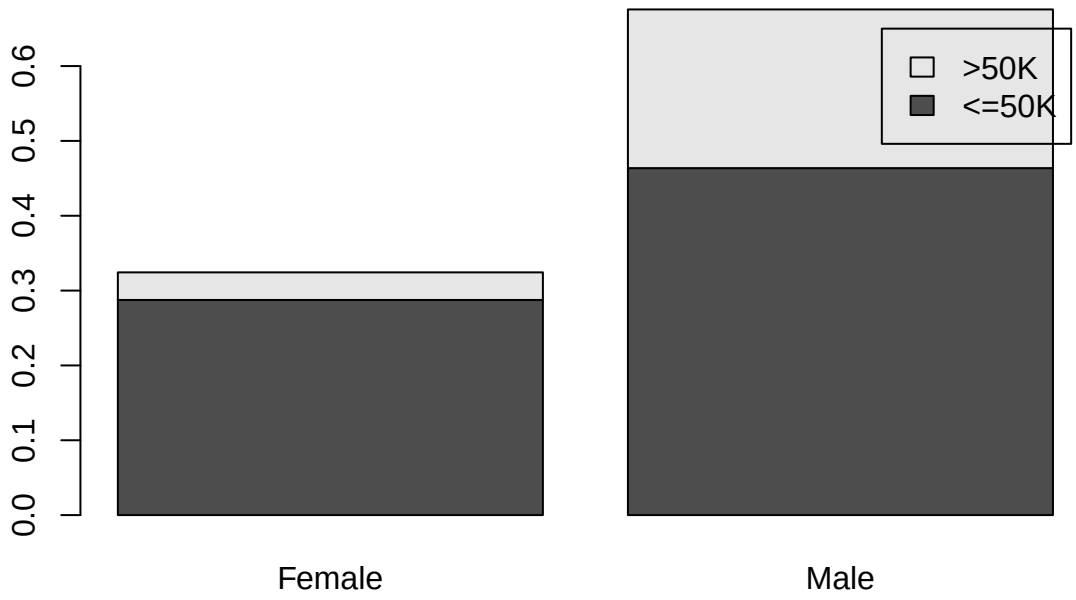
```
## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "legend" is not
## a graphical parameter
```



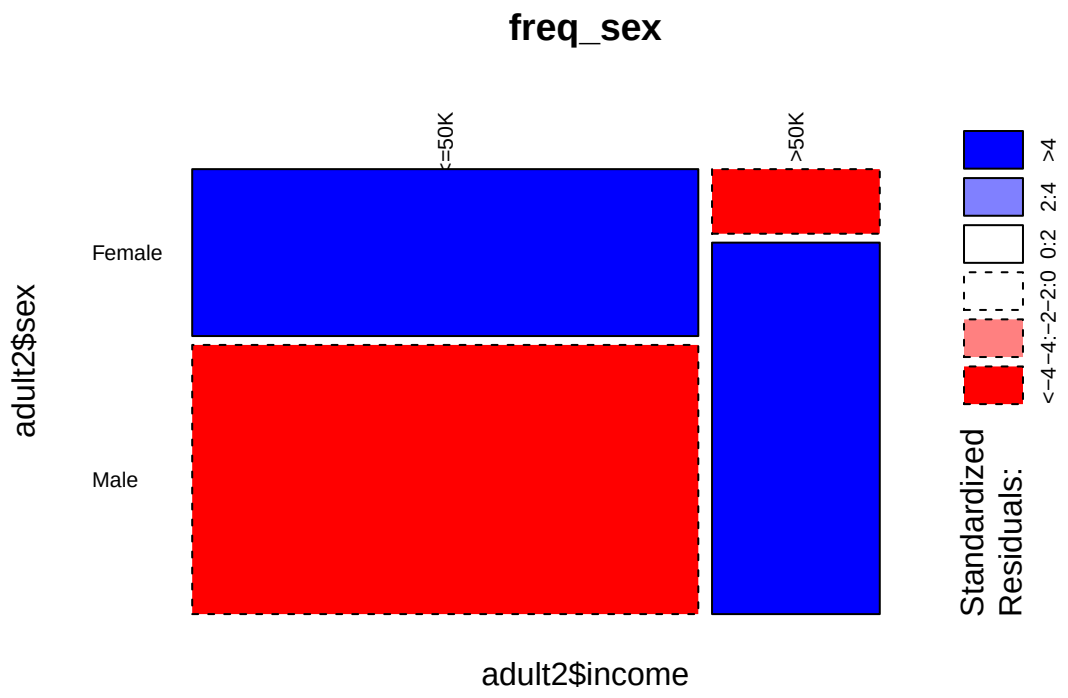

```
mosaicplot(freq_race, border = "black",
           shade = TRUE, las=2)
```



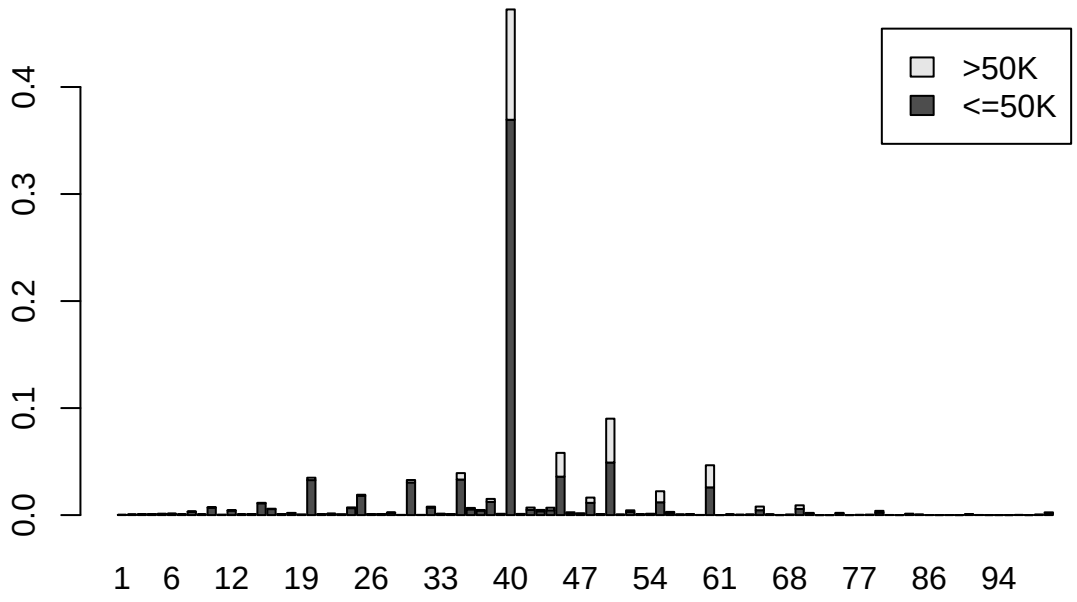

```
freq_sex=xtabs(~adult2$income+adult2$sex)
barplot(prop.table(freq_sex),legend=rownames(freq_sex))
```



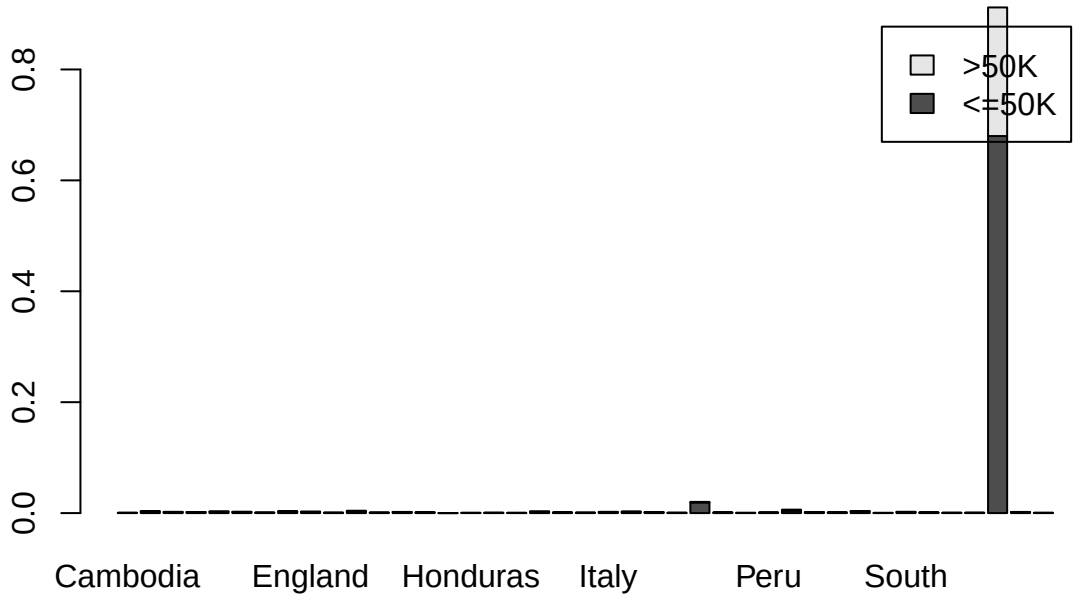

```
mosaicplot(freq_sex, border = "black",
           shade = TRUE, las=2)
```




```
freq_hours.per.week=xtabs(~adult2$income+adult2$hours.per.week)
barplot(prop.table(freq_hours.per.week),legend=rownames(freq_hours.per.week))
```




```
freq_native.country=xtabs(~adult2$income+adult2$native.country)
barplot(prop.table(freq_native.country),legend=rownames(freq_native.country))
```



FEATURE ENGINEERING

Creating vectors with the variables in each categorical variable

```
wc <-c("Private", "Self-emp-not-inc", "Self-emp-inc", "Federal-gov", "Local-gov", "State-gov", "Without
ed <-c("Bachelors", "Some-college", "11th", "HS-grad", "Prof-school", "Assoc-acdm", "Assoc-voc", "9th",
ms <-c("Married-civ-spouse", "Divorced", "Never-married", "Separated", "Widowed", "Married-spouse-absen
oc <-c("Tech-support", "Craft-repair", "Other-service", "Sales", "Exec-managerial", "Prof-specialty", "
rs <-c("Wife", "Own-child", "Husband", "Not-in-family", "Other-relative", "Unmarried")
ra <-c("White", "Asian-Pac-Islander", "Amer-Indian-Eskimo", "Other", "Black")
nc <-c("United-States", "Cambodia", "England", "Puerto-Rico", "Canada", "Germany", "Outlying-US(Guam-US
```

Transforming income and sex variables in to binary variables

```
adult$income <- as.factor(adult$income)
adult$income <- as.numeric(adult$income)
adult$sex <- as.factor(adult$sex)
adult$sex<- as.numeric(adult$sex)
```

Matching variables in the data set with the category variables which will transform the categorical

```
adult$workclass <-match(adult$workclass, wc)
adult$education <-match(adult$education, ed)
adult$marital.status <-match(adult$marital.status, ms)
adult$occupation <-match(adult$occupation, oc)
adult$relationship <-match(adult$relationship,rs)
adult$race <-match(adult$race, ra)
adult$native.country <-match(adult$native.country, nc)
```

Creating a new dataframe with out any missing observations

```
dataframe <-na.omit(adult)
str(dataframe)
```

```
## 'data.frame':    30162 obs. of  15 variables:
##  $ age           : int  82 54 41 34 38 74 68 45 38 52 ...
##  $ workclass     : int  1 1 1 1 1 6 4 1 2 1 ...
##  $ fnlwgt        : int  132870 140359 264663 216864 150601 88638 422013 172274 164526 129177 ...
##  $ education     : int  4 9 2 4 13 14 4 14 5 1 ...
##  $ education.num : int  9 4 10 9 6 16 9 16 15 13 ...
##  $ marital.status: int  5 2 4 2 4 3 2 2 3 5 ...
##  $ occupation    : int  5 8 6 3 9 6 6 6 6 3 ...
##  $ relationship  : int  4 6 2 6 6 5 4 6 4 4 ...
##  $ race          : int  1 1 1 1 1 1 1 5 1 1 ...
##  $ sex           : num  1 1 1 1 2 1 1 1 2 1 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3900 3770 3770 3683 3683 3004 2824 2824 ...
##  $ hours.per.week: int  18 40 40 45 40 20 40 35 45 20 ...
##  $ native.country: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ income        : num  1 1 1 1 1 2 1 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:2399] 1 3 10 15 19 25 45 49 50 66 ...
##   ..- attr(*, "names")= chr [1:2399] "1" "3" "10" "15" ...
```

```
summary(dataframe)
```

```
##       age          workclass         fnlwgt          education
##  Min.   :17.00   Min.   :1.000   Min.   :  13769   Min.   : 1.000
##  1st Qu.:28.00   1st Qu.:1.000   1st Qu.: 117627   1st Qu.: 2.000
##  Median :37.00   Median :1.000   Median : 178425   Median : 4.000
##  Mean   :38.44   Mean   :1.737   Mean   : 189794   Mean   : 4.373
##  3rd Qu.:47.00   3rd Qu.:2.000   3rd Qu.: 237629   3rd Qu.: 5.000
##  Max.   :90.00   Max.   :7.000   Max.   :1484705   Max.   :16.000
##  education.num   marital.status    occupation      relationship
##  Min.   : 1.00   Min.   :1.000   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:1.000   1st Qu.: 3.000   1st Qu.:3.000
##  Median :10.00   Median :2.000   Median : 5.000   Median :3.000
##  Mean   :10.12   Mean   :2.053   Mean   : 5.742   Mean   :3.393
##  3rd Qu.:13.00   3rd Qu.:3.000   3rd Qu.: 8.000   3rd Qu.:4.000
##  Max.   :16.00   Max.   :7.000   Max.   :14.000   Max.   :6.000
##       race            sex         capital.gain    capital.loss
##  Min.   :1.000   Min.   :1.000   Min.   :    0   Min.   :   0.00
```

```
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:    0   1st Qu.:   0.00
##  Median :1.000   Median :2.000   Median :    0   Median :   0.00
##  Mean   :1.445   Mean   :1.676   Mean   : 1092   Mean   :  88.37
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:    0   3rd Qu.:   0.00
##  Max.   :5.000   Max.   :2.000   Max.   :99999   Max.   :4356.00
##  hours.per.week  native.country       income
##  Min.   : 1.00   Min.   : 1.000   Min.   :1.000
##  1st Qu.:40.00   1st Qu.: 1.000   1st Qu.:1.000
##  Median :40.00   Median : 1.000   Median :1.000
##  Mean   :40.93   Mean   : 2.516   Mean   :1.249
##  3rd Qu.:45.00   3rd Qu.: 1.000   3rd Qu.:1.000
##  Max.   :99.00   Max.   :41.000   Max.   :2.000
```

MODELING - REGRESSION

```
reg1 <-lm(hours.per.week ~ age + factor(workclass) + factor(education) + factor(marital.status) + facto
summary(reg1)
```

```
##
## Call:
## lm(formula = hours.per.week ~ age + factor(workclass) + factor(education) +
##     factor(marital.status) + factor(occupation) + sex + capital.gain +
##     capital.loss + income, data = dataframe, weights = fnlwgt)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -33647  -1883    -35   1945  40625
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.458e+01  6.033e-01  57.327  < 2e-16 ***
## age                      -5.328e-02  6.029e-03  -8.838  < 2e-16 ***
## factor(workclass)2        8.922e-01  2.472e-01   3.609 0.000307 ***
## factor(workclass)3        4.466e+00  3.613e-01  12.360  < 2e-16 ***
## factor(workclass)4       -3.269e-02  3.712e-01  -0.088 0.929825
## factor(workclass)5       -3.546e-01  2.652e-01  -1.337 0.181200
## factor(workclass)6       -1.641e+00  3.236e-01  -5.070 4.01e-07 ***
## factor(workclass)7       -7.842e+00  3.011e+00  -2.604 0.009215 **
## factor(education)2       -2.217e+00  2.154e-01 -10.294  < 2e-16 ***
## factor(education)3       -6.056e+00  3.829e-01 -15.815  < 2e-16 ***
## factor(education)4       -9.505e-01  2.130e-01  -4.462 8.13e-06 ***
## factor(education)5        3.168e+00  5.174e-01   6.123 9.32e-10 ***
## factor(education)6       -6.672e-01  3.746e-01  -1.781 0.074911 .
## factor(education)7       -6.736e-01  3.485e-01  -1.933 0.053258 .
## factor(education)8       -2.125e+00  5.277e-01  -4.027 5.67e-05 ***
## factor(education)9       -1.901e+00  5.014e-01  -3.792 0.000150 ***
## factor(education)10      -5.062e+00  5.731e-01  -8.833  < 2e-16 ***
## factor(education)11       9.390e-01  3.201e-01   2.933 0.003356 **
## factor(education)12      -9.811e-01  8.038e-01  -1.221 0.222266
## factor(education)13      -3.781e+00  4.184e-01  -9.036  < 2e-16 ***
## factor(education)14       4.268e+00  5.961e-01   7.160 8.28e-13 ***
## factor(education)15      -1.899e+00  6.099e-01  -3.114 0.001845 **
## factor(education)16      -3.006e+00  1.454e+00  -2.068 0.038661 *
```

```
## factor(marital.status)2  1.590e+00  2.133e-01   7.453 9.36e-14 ***
## factor(marital.status)3 -3.082e+00  1.817e-01 -16.962  < 2e-16 ***
## factor(marital.status)4  4.024e-01  3.625e-01   1.110 0.266988
## factor(marital.status)5 -3.410e+00  4.322e-01  -7.889 3.15e-15 ***
## factor(marital.status)6 -3.669e-01  5.689e-01  -0.645 0.519028
## factor(marital.status)7 -1.635e-01  2.380e+00  -0.069 0.945248
## factor(occupation)2      1.962e+00  4.042e-01   4.854 1.21e-06 ***
## factor(occupation)3     -2.317e+00  4.126e-01  -5.616 1.98e-08 ***
## factor(occupation)4      1.107e+00  4.029e-01   2.748 0.006007 **
## factor(occupation)5      3.397e+00  4.004e-01   8.483  < 2e-16 ***
## factor(occupation)6      8.865e-01  4.103e-01   2.161 0.030721 *
## factor(occupation)7     -5.222e-01  4.676e-01  -1.117 0.264050
## factor(occupation)8      1.765e+00  4.398e-01   4.014 5.98e-05 ***
## factor(occupation)9      1.748e-02  4.013e-01   0.044 0.965252
## factor(occupation)10     5.949e+00  5.253e-01  11.325  < 2e-16 ***
## factor(occupation)11     4.449e+00  4.591e-01   9.690  < 2e-16 ***
## factor(occupation)12    -2.751e+00  9.602e-01  -2.865 0.004176 **
## factor(occupation)13     2.611e+00  5.593e-01   4.668 3.05e-06 ***
## factor(occupation)14     2.381e+00  3.409e+00   0.698 0.484880
## sex                      3.116e+00  1.631e-01  19.112  < 2e-16 ***
## capital.gain             3.478e-05  8.705e-06   3.996 6.46e-05 ***
## capital.loss             3.477e-04  1.584e-04   2.195 0.028171 *
## income                   2.884e+00  1.809e-01  15.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4694 on 30116 degrees of freedom
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.1598
## F-statistic: 128.4 on 45 and 30116 DF,  p-value: < 2.2e-16
```

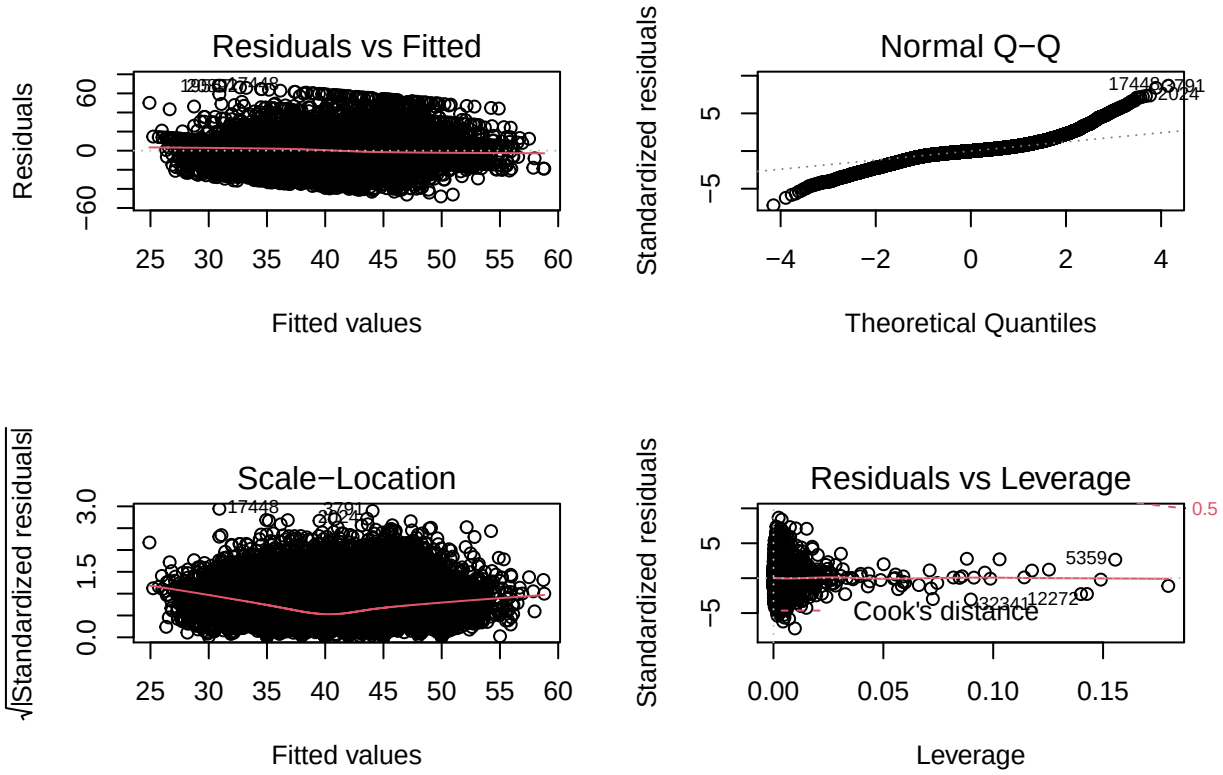
Testing the OLS model assumptions

```
cor(dataframe)
```

```
##                        age   workclass        fnlwgt    education education.num
## age             1.00000000  0.13548584 -0.0765108361  0.115388572   0.043526087
## workclass       0.13548584  1.00000000 -0.0268182086  0.033766347   0.179178398
## fnlwgt         -0.07651084 -0.02681821  1.0000000000  0.021348999  -0.044991742
## education       0.11538857  0.03376635  0.0213489990  1.000000000  -0.232985821
## education.num   0.04352609  0.17917840 -0.0449917421 -0.232985821   1.000000000
## marital.status -0.22917937 -0.05471039  0.0288431980 -0.014320869  -0.103013927
## occupation      0.02271652  0.12877924  0.0037586754  0.012513423  -0.039234711
## relationship    0.12299719  0.01346658  0.0152713820  0.028990304  -0.033293279
## race           -0.02382669  0.03498514  0.0992509173  0.007865332  -0.079068980
## sex             0.08199259 -0.00531661  0.0253622713  0.032918804   0.006156852
## capital.gain    0.08015423  0.01370186  0.0004215674  0.022471922   0.124415995
## capital.loss    0.06016548  0.02254063 -0.0097495278  0.020413255   0.079646410
## hours.per.week  0.10159876  0.02751856 -0.0228857516  0.010587438   0.152522071
## native.country -0.03585423 -0.05575043  0.1017145838  0.160303517  -0.164778860
## income          0.24199814  0.08566249 -0.0089574234  0.009566776   0.335286197
##                marital.status    occupation relationship         race
## age               -0.22917937  0.0227165206   0.12299719 -0.023826688
## workclass         -0.05471039  0.1287792427   0.01346658  0.034985135
```

```
## fnlwgt            0.02884320  0.0037586754   0.01527138  0.099250917
## education        -0.01432087  0.0125134226   0.02899030  0.007865332
## education.num    -0.10301393 -0.0392347107  -0.03329328 -0.079068980
## marital.status    1.00000000  0.0176419349   0.36794411  0.132561975
## occupation        0.01764193  1.0000000000   0.01134532  0.042564655
## relationship      0.36794411  0.0113453172   1.00000000  0.121135759
## race              0.13256198  0.0425646551   0.12113576  1.000000000
## sex              -0.38310737 -0.0446579359  -0.17905119 -0.118940193
## capital.gain     -0.07352647 -0.0121902969  -0.02687211 -0.020829315
## capital.loss     -0.06938392 -0.0144896182  -0.03175774 -0.028396801
## hours.per.week   -0.22675463  0.0398336125   0.05289354 -0.061667023
## native.country    0.05246859  0.0008498021   0.04398917  0.054022615
## income           -0.37868408 -0.0451286242  -0.17911623 -0.096061474
##                         sex  capital.gain capital.loss hours.per.week
## age             0.0819925947  0.0801542263  0.060165480     0.10159876
## workclass      -0.0053166101  0.0137018565  0.022540630     0.02751856
## fnlwgt          0.0253622713  0.0004215674 -0.009749528    -0.02288575
## education       0.0329188037  0.0224719225  0.020413255     0.01058744
## education.num   0.0061568521  0.1244159953  0.079646410     0.15252207
## marital.status -0.3831073699 -0.0735264673 -0.069383917    -0.22675463
## occupation     -0.0446579359 -0.0121902969 -0.014489618     0.03983361
## relationship   -0.1790511890 -0.0268721096 -0.031757742     0.05289354
## race           -0.1189401928 -0.0208293154 -0.028396801    -0.06166702
## sex             1.0000000000  0.0488140537  0.047011240     0.23126813
## capital.gain    0.0488140537  1.0000000000 -0.032229327     0.08043180
## capital.loss    0.0470112398 -0.0322293265  1.000000000     0.05241705
## hours.per.week  0.2312681272  0.0804318007  0.052417049     1.00000000
## native.country  0.0005954787 -0.0171080007 -0.020918092    -0.01883877
## income          0.2166986811  0.2211962145  0.150053308     0.22948013
##                native.country       income
## age             -0.0358542313  0.241998136
## workclass       -0.0557504271  0.085662491
## fnlwgt           0.1017145838 -0.008957423
## education        0.1603035173  0.009566776
## education.num   -0.1647788602  0.335286197
## marital.status   0.0524685888 -0.378684084
## occupation       0.0008498021 -0.045128624
## relationship     0.0439891710 -0.179116226
## race             0.0540226149 -0.096061474
## sex              0.0005954787  0.216698681
## capital.gain    -0.0171080007  0.221196215
## capital.loss    -0.0209180923  0.150053308
## hours.per.week  -0.0188387657  0.229480130
## native.country   1.0000000000 -0.059455067
## income          -0.0594550668  1.000000000
```

```
par(mfrow=c(2,2))
plot (reg1)
```



creating the test and train data set

```
spl = sample.split(dataframe$hours.per.week, SplitRatio = 0.7)
train = subset(dataframe, spl==TRUE)
test = subset(dataframe, spl==FALSE)
```

creating a train model

```
trainreg <-lm(hours.per.week ~ age + factor(workclass) + factor(education) + factor(marital.status) + fa
testreg <- predict(trainreg, test)
regrmse <- rmse(actual = test$hours.per.week, predicted = testreg)
print(regrmse)
```

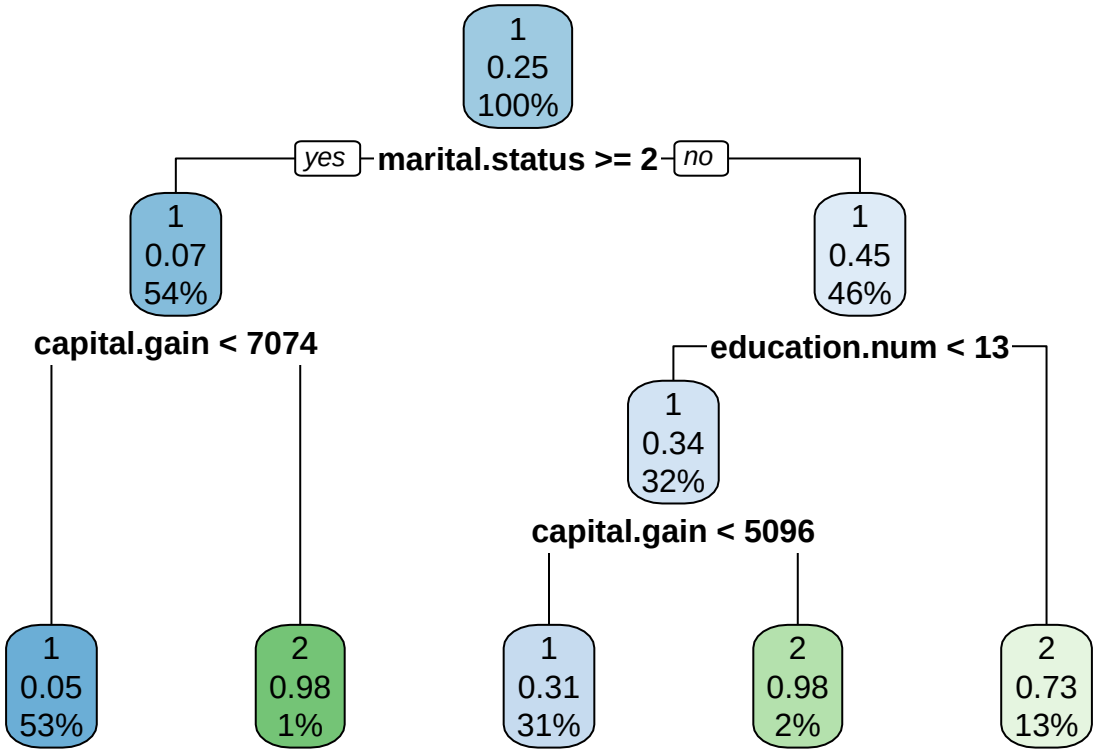
```
## [1] 11.01644
```

DECISION TREE

```
dataTree = dataframe

spl = sample.split(dataTree$income, SplitRatio = 0.8)
train_tree = subset(dataTree, spl==TRUE)
test_tree = subset(dataTree, spl==FALSE)
```

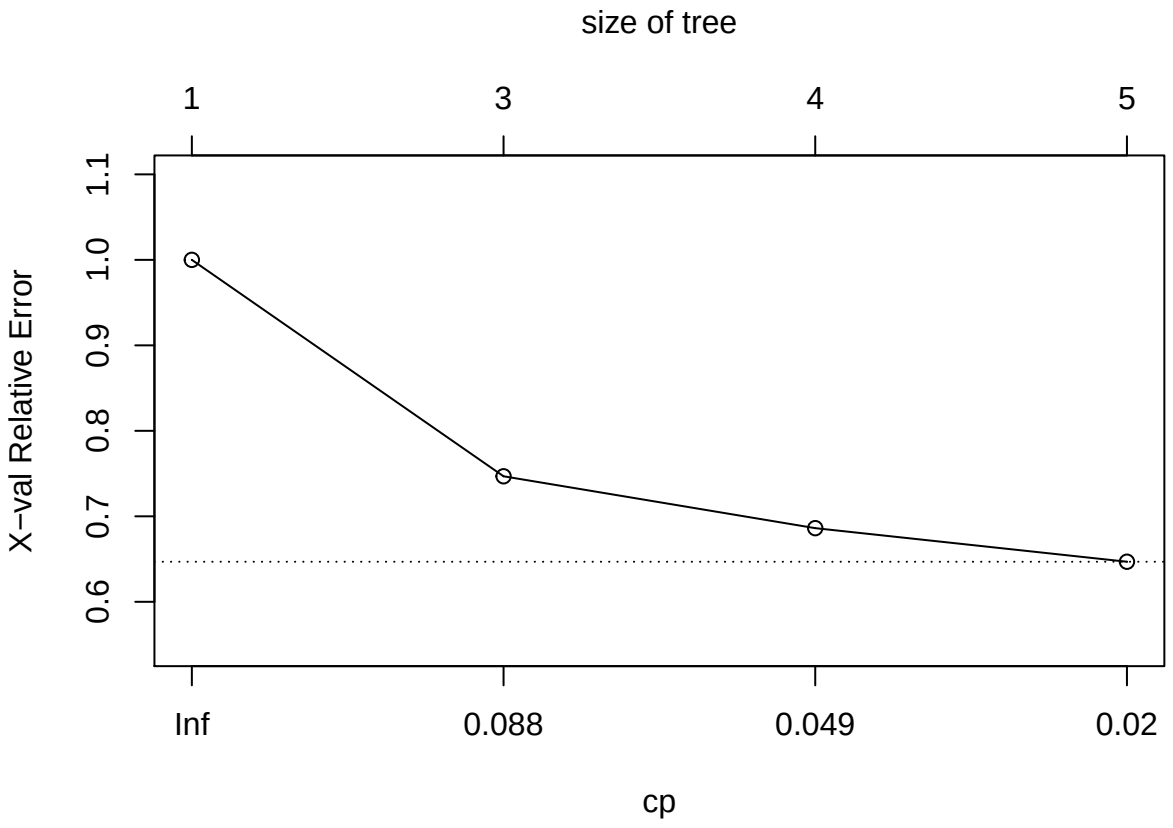
```
dtree <- rpart(income~., data = train_tree, weights=c(fnlwgt), method = 'class')
rpart.plot(dtree, extra = 'auto')
```



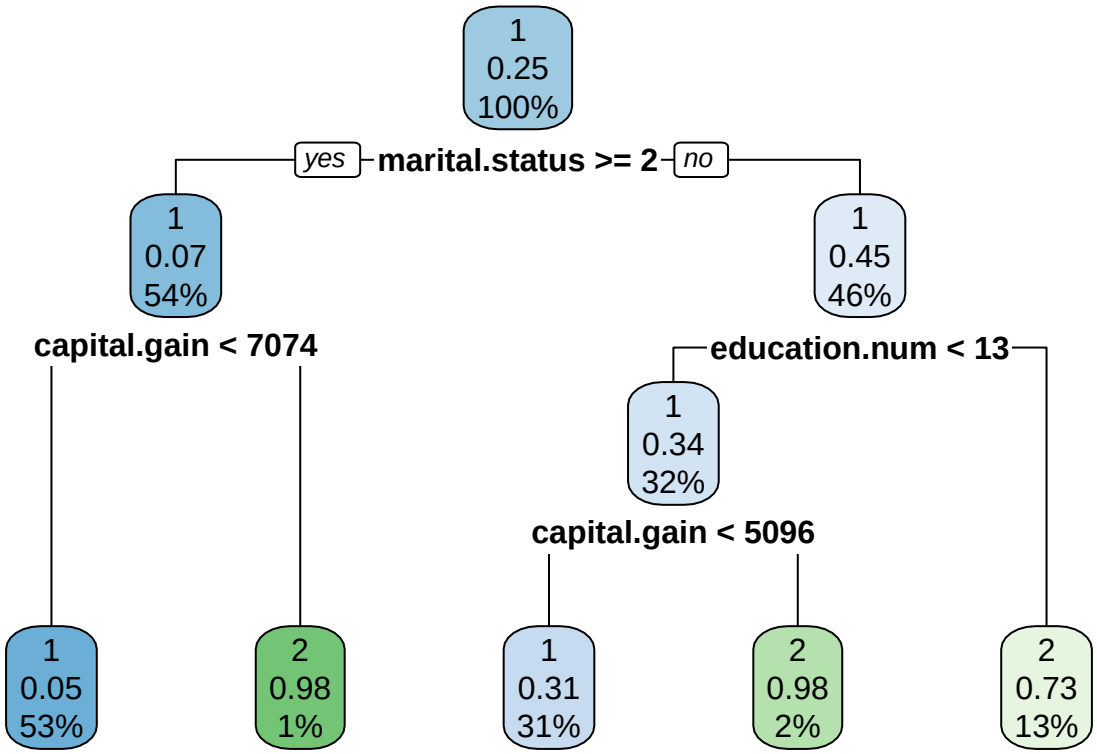

```
printcp(dtree)
```

```
##
## Classification tree:
## rpart(formula = income ~ ., data = train_tree, weights = c(fnlwgt),
##     method = "class")
##
## Variables actually used in tree construction:
## [1] capital.gain   education.num  marital.status
##
## Root node error: 1126324778/24129 = 46679
##
## n= 24129
##
##         CP nsplit rel error  xerror       xstd
## 1 0.126583      0   1.00000 1.00000 2.5883e-05
## 2 0.060671      2   0.74683 0.74683 2.3271e-05
## 3 0.039289      3   0.68616 0.68616 2.2508e-05
## 4 0.010000      4   0.64687 0.64687 2.1980e-05
```

```
plotcp(dtree)
```




```
ptree<- prune(dtree, cp= dtree$cptable[which.min(dtree$cptable[,'xerror']),'CP'])
rpart.plot(ptree)
```




```
predTree <-predict(ptree, test_tree, type = 'class')
```

Confusion Matrix

```
confmat_Tree <- table(test_tree$income, predTree)
confmat_Tree
```

```
##    predTree
##       1    2
##   1 4311  220
##   2  713  789
```

Precision

```
precision_Tree <- confmat_Tree[2,2]/(confmat_Tree[2,2] + confmat_Tree[1,2])
precision_Tree
```

```
## [1] 0.7819623
```

Recall

```
recall_Tree <- confmat_Tree[2,2]/ (confmat_Tree[2,2] + confmat_Tree[2,1])
recall_Tree
```

```
## [1] 0.5252996
```

Accuracy

```
accuracy_Tree <- (confmat_Tree[2,2]+confmat_Tree[1,1])/ nrow(test_tree)
accuracy_Tree
```

```
## [1] 0.8453506
```

Balanced Accuracy

```
balancedacc_Tree <- (confmat_Tree[1,1]/(confmat_Tree[1,1] + confmat_Tree[1,2]) + confmat_Tree[2,2]/ (co
balancedacc_Tree
```

```
## [1] 0.7383726
```

Clustering mixed data types, gower distance

```
# Normalize Data

preproc1 <- preProcess(adult_fact[,c(1:9)], method=c("center", "scale"))
norm1 <- predict(preproc1, adult_fact[,c(1:9)])
summary(norm1)
```

```
##       age          education.num                    marital.status
##  Min.   :-1.6322   Min.   :-3.57699   Divorced              : 4214
##  1st Qu.:-0.7947   1st Qu.:-0.43973   Married-AF-spouse     :   21
##  Median :-0.1095   Median :-0.04757   Married-civ-spouse    :14065
##  Mean   : 0.0000   Mean   : 0.00000   Married-spouse-absent:  370
##  3rd Qu.: 0.6519   3rd Qu.: 1.12890   Never-married         : 9726
##  Max.   : 3.9257   Max.   : 2.30537   Separated             :  939
##                                       Widowed               :  827
##          relationship       sex         capital.gain      capital.loss
##  Husband       :12463   Female: 9782   Min.   :-0.1474   Min.   :-0.2186
##  Not-in-family : 7726   Male  :20380   1st Qu.:-0.1474   1st Qu.:-0.2186
##  Other-relative:  889                  Median :-0.1474   Median :-0.2186
##  Own-child     : 4466                  Mean   : 0.0000   Mean   : 0.0000
##  Unmarried     : 3212                  3rd Qu.:-0.1474   3rd Qu.:-0.2186
##  Wife          : 1406                  Max.   :13.3544   Max.   :10.5556
##
##  hours.per.week       income
##  Min.   :-3.33316   <=50K:22654
##  1st Qu.:-0.07773   >50K : 7508
##  Median :-0.07773
##  Mean   : 0.00000
##  3rd Qu.: 0.33963
##  Max.   : 4.84715
##
```

```
preproc2 <- preProcess(adult_num[,c(1:9)], method=c("center", "scale"))
norm2 <- predict(preproc2, adult_num[,c(1:9)])
summary(norm2)
```

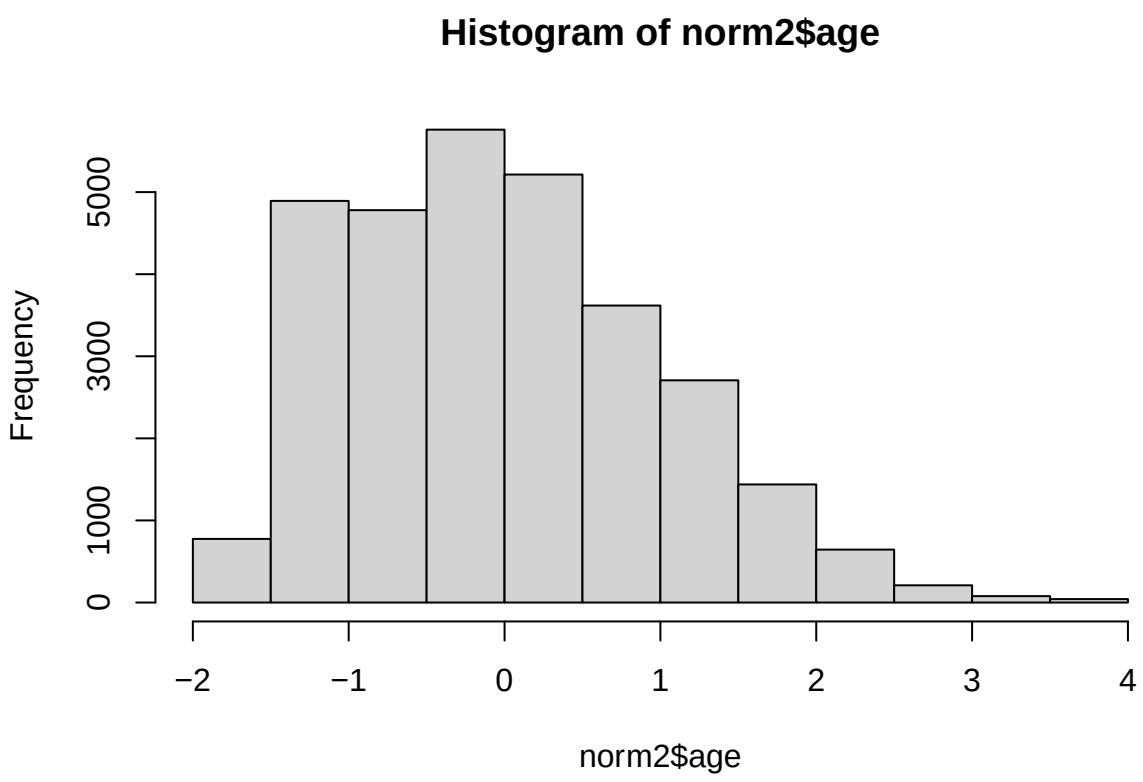
```
##       age          education.num      marital.status     relationship
##  Min.   :-1.6322   Min.   :-3.57699   Min.   :-1.7224   Min.   :-0.8857
##  1st Qu.:-0.7947   1st Qu.:-0.43973   1st Qu.:-0.3873   1st Qu.:-0.8857
##  Median :-0.1095   Median :-0.04757   Median :-0.3873   Median :-0.2612
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.6519   3rd Qu.: 1.12890   3rd Qu.: 0.9478   3rd Qu.: 0.9877
##  Max.   : 3.9257   Max.   : 2.30537   Max.   : 2.2829   Max.   : 2.2367
##       sex           capital.gain      capital.loss     hours.per.week
##  Min.   :-1.4434   Min.   :-0.1474   Min.   :-0.2186   Min.   :-3.33316
```

```
##  1st Qu.:-1.4434   1st Qu.:-0.1474   1st Qu.:-0.2186   1st Qu.:-0.07773
##  Median : 0.6928   Median :-0.1474   Median :-0.2186   Median :-0.07773
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.6928   3rd Qu.:-0.1474   3rd Qu.:-0.2186   3rd Qu.: 0.33963
##  Max.   : 0.6928   Max.   :13.3544   Max.   :10.5556   Max.   : 4.84715
##      income
##  Min.   :-0.5757
##  1st Qu.:-0.5757
##  Median :-0.5757
##  Mean   : 0.0000
##  3rd Qu.:-0.5757
##  Max.   : 1.7370
```

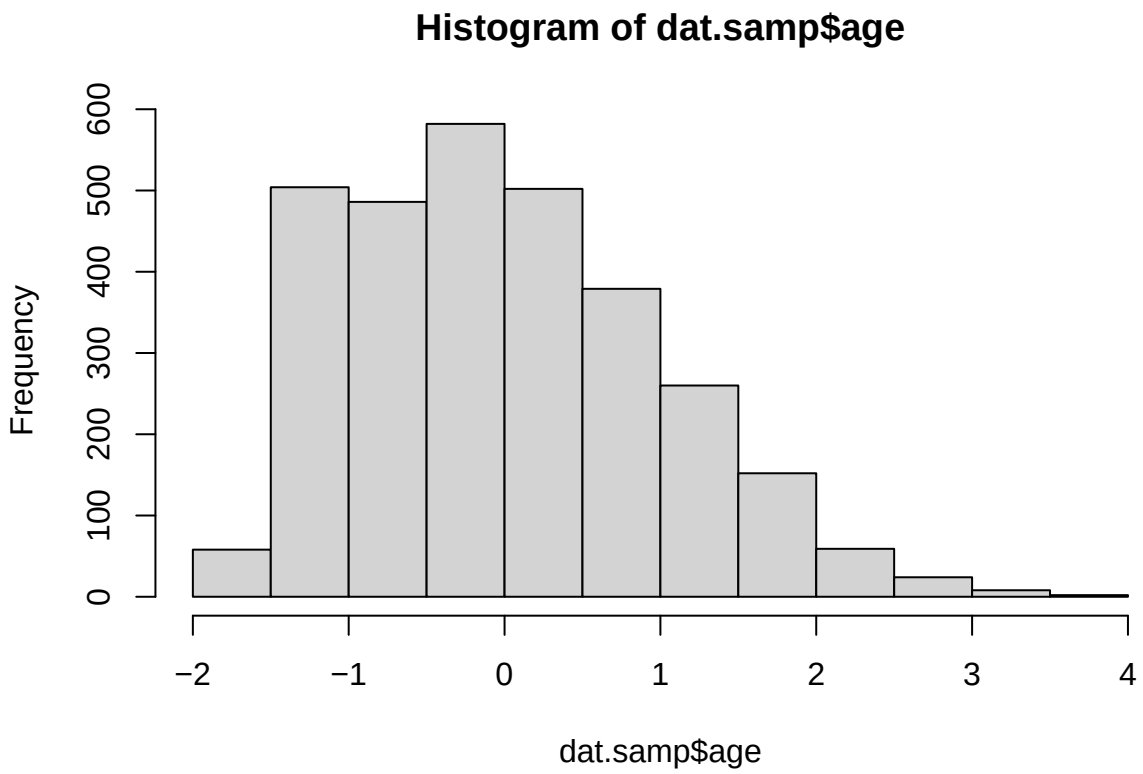
```
# Split  Data for assessing sample size, we will compare histograms using the numeric dataset
set.seed(100)
dat.split = sample.split(norm2$income, SplitRatio = 0.9)
dat.discard = subset(norm2, dat.split==TRUE)
dat.samp = subset(norm2, dat.split==FALSE)
```

Histograms Comparing distrubution in full dataset and sampled data

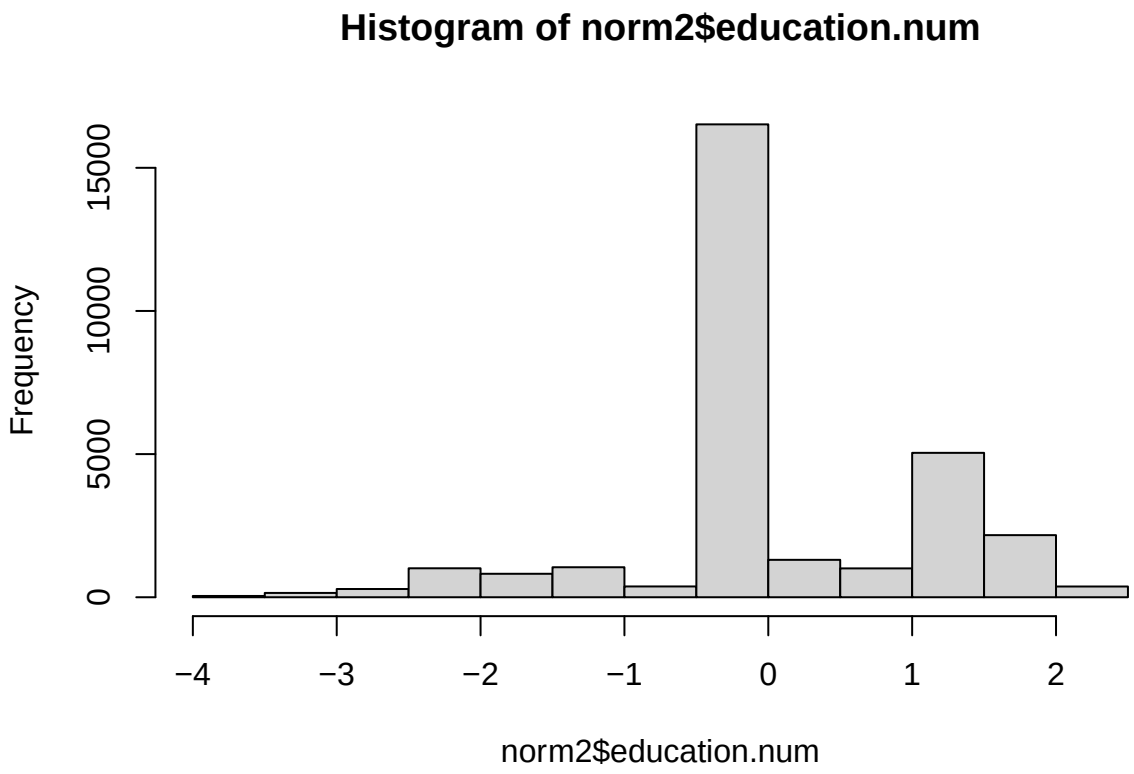
```
hist(norm2$age)
```



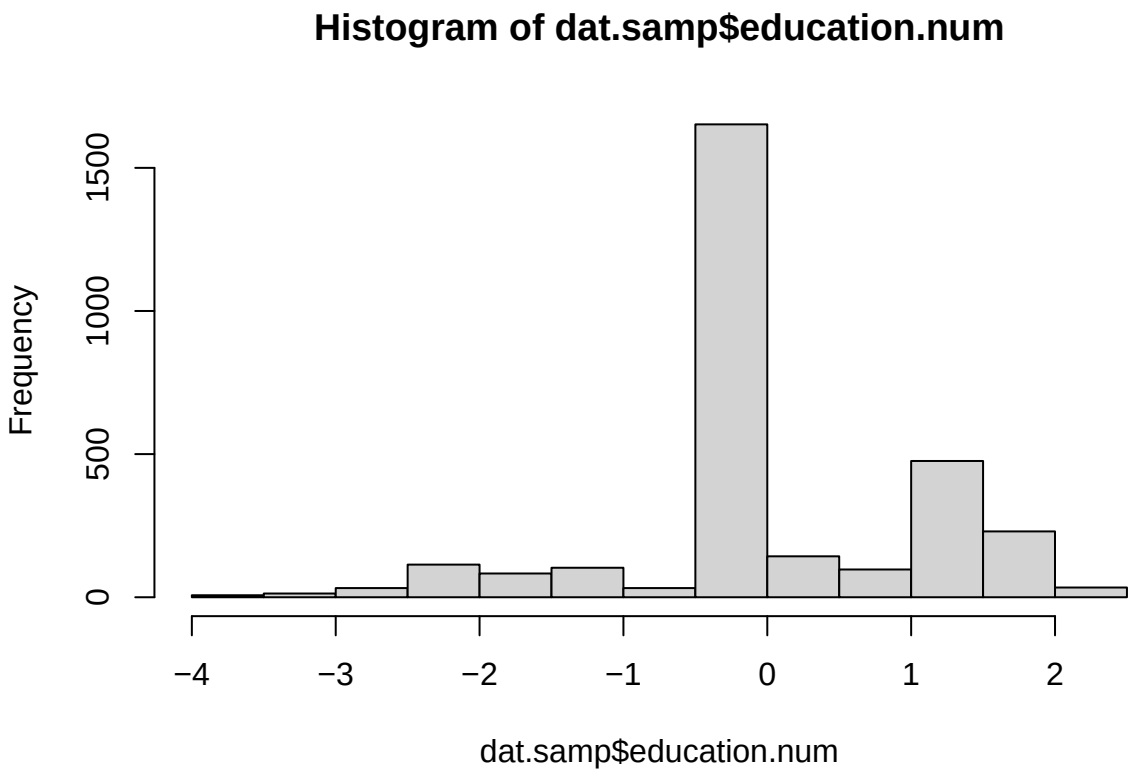
```r
hist(dat.samp$age)
```



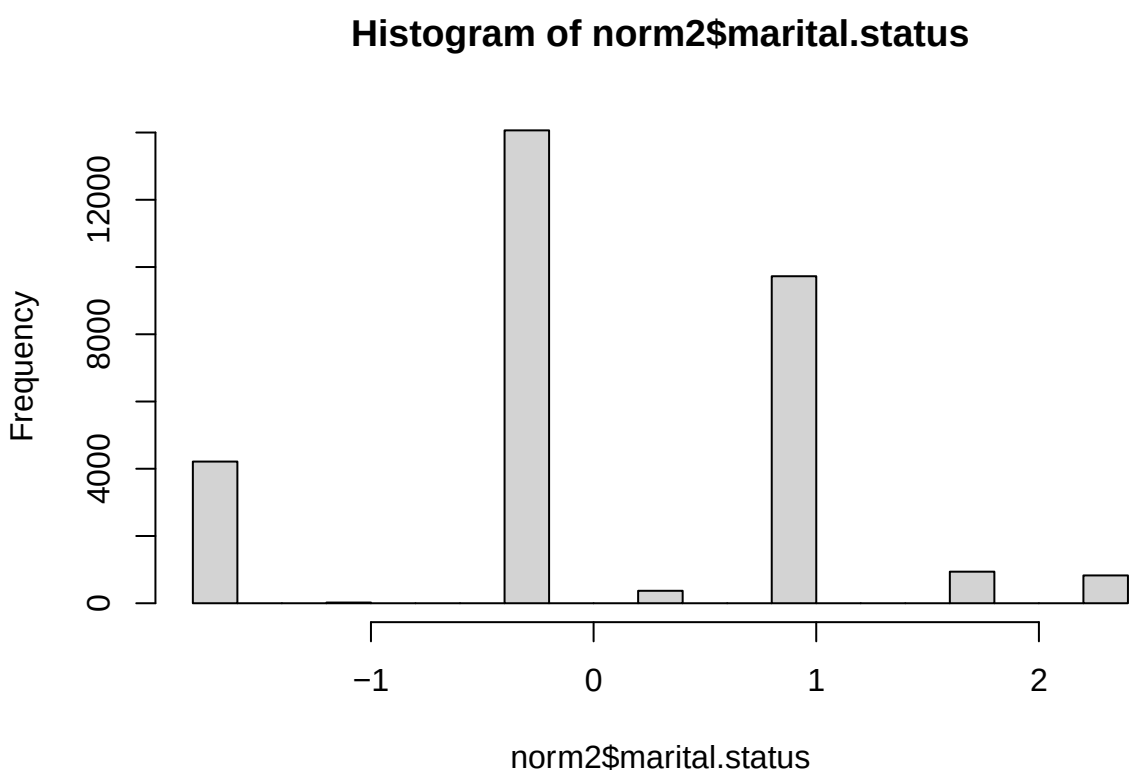

```r
hist(norm2$education.num)
```



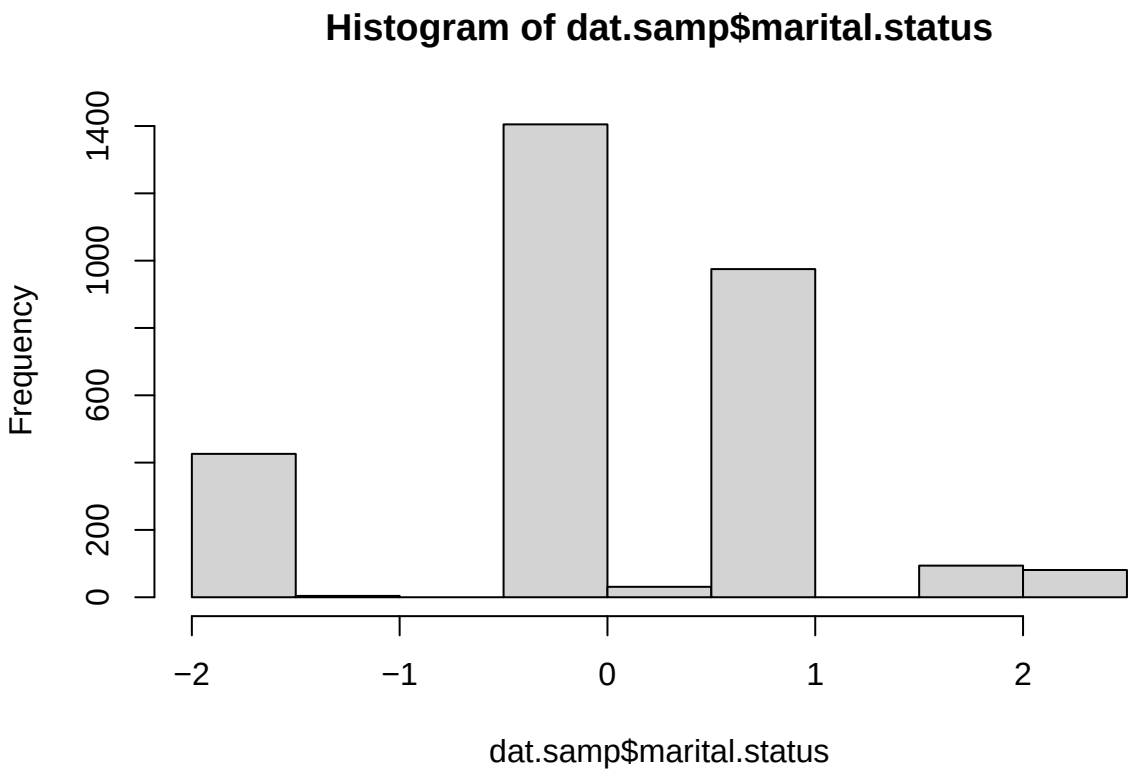

```
hist(dat.samp$education.num)
```



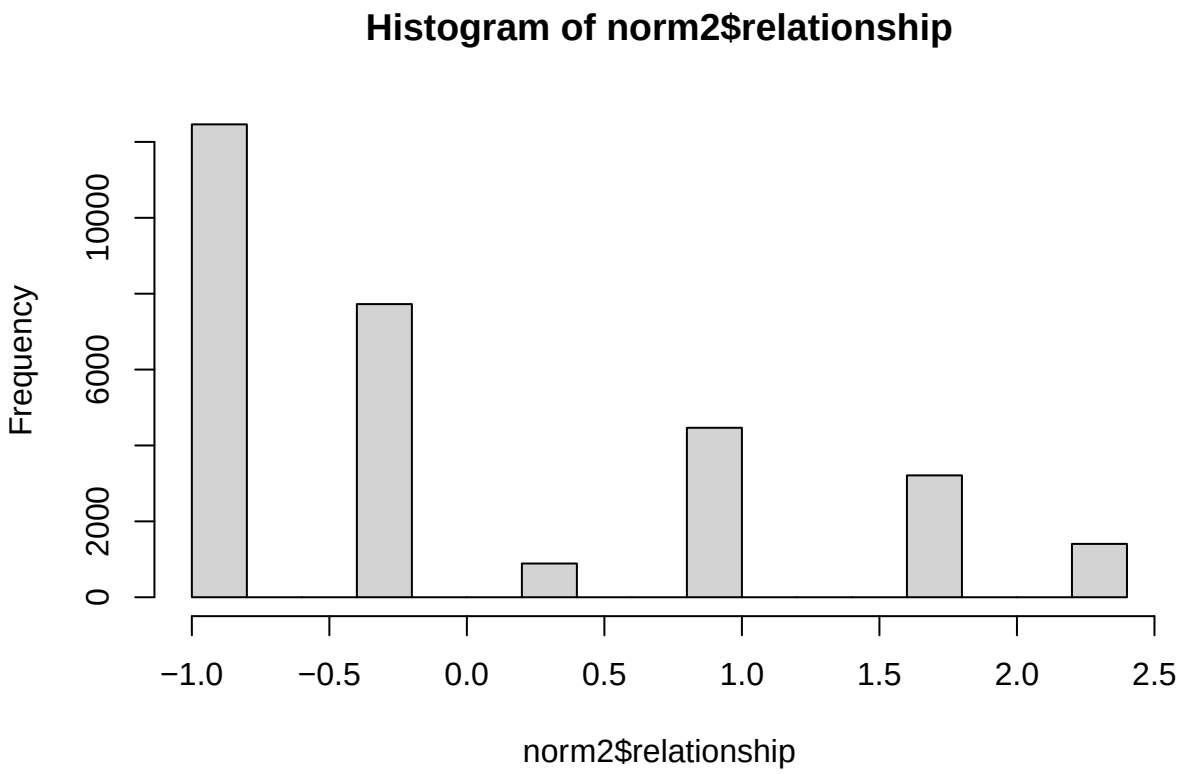

```
hist(norm2$marital.status)
```



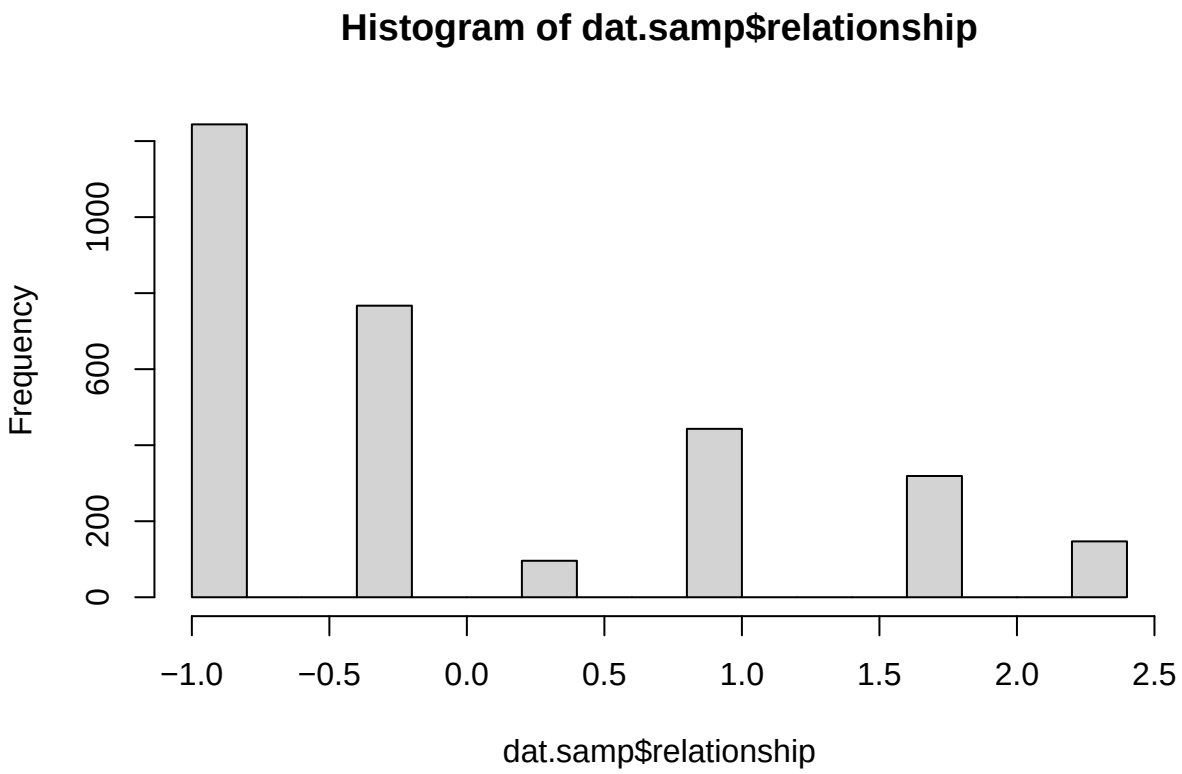

```
hist(dat.samp$marital.status)
```



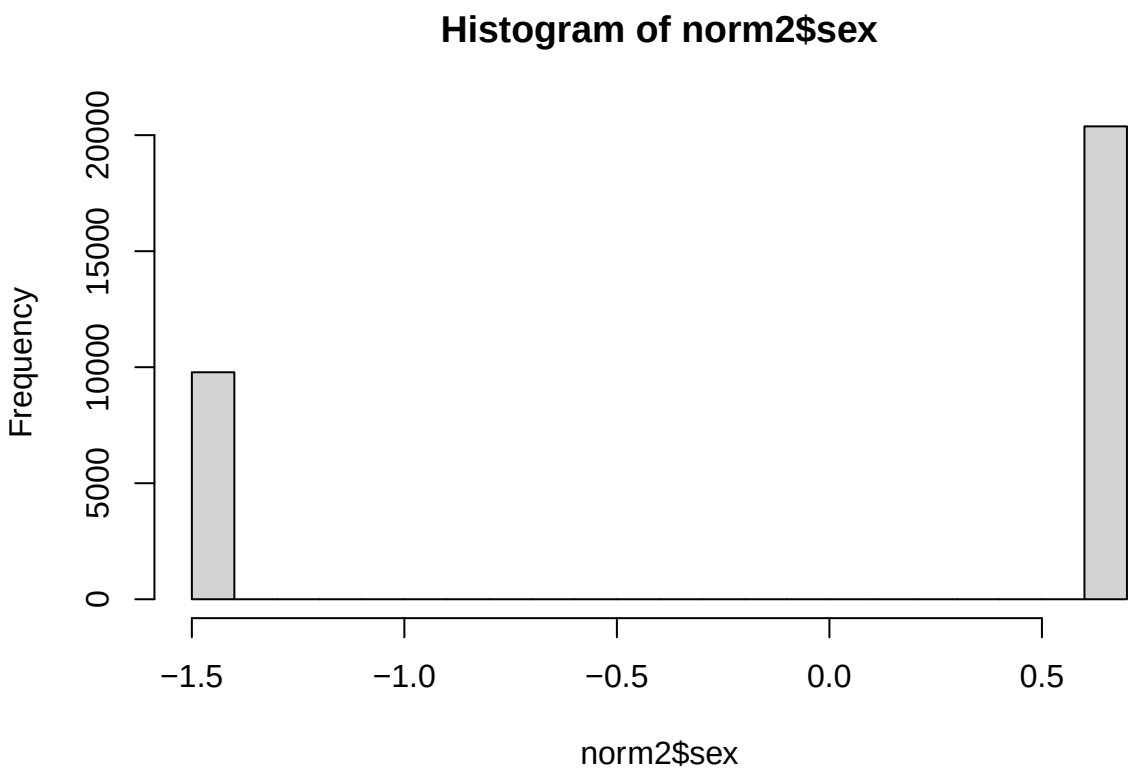

```
hist(norm2$relationship)
```



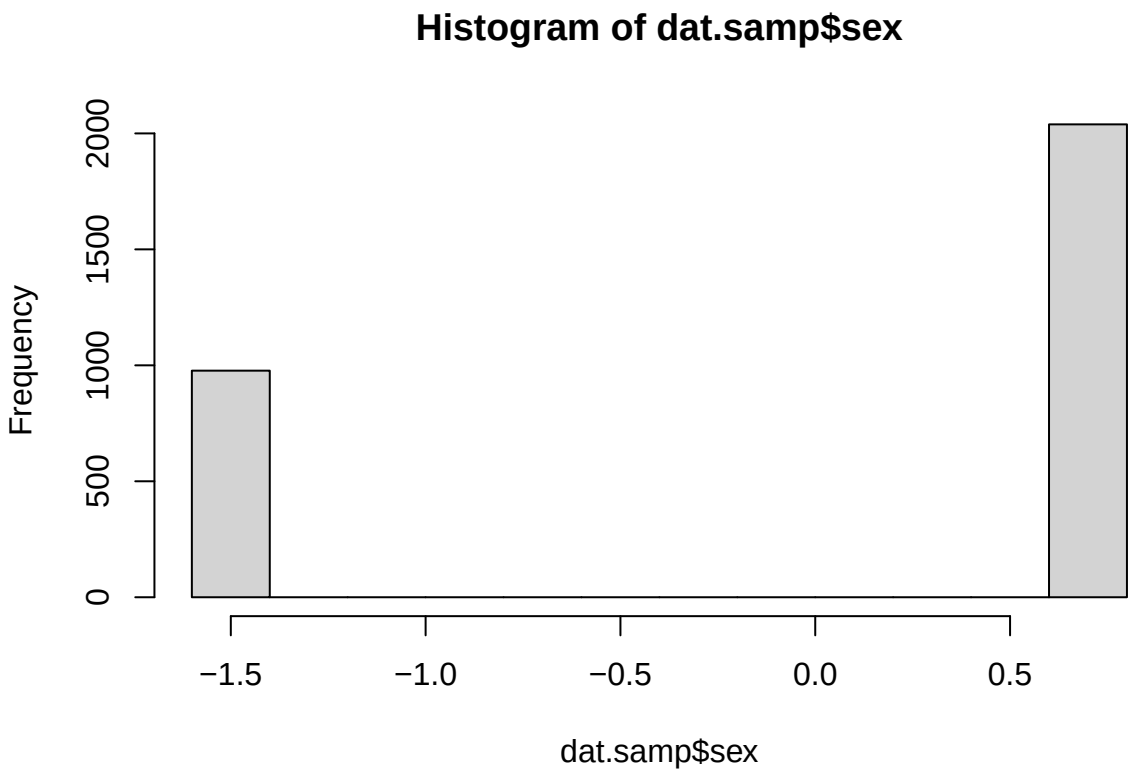

```
hist(dat.samp$relationship)
```



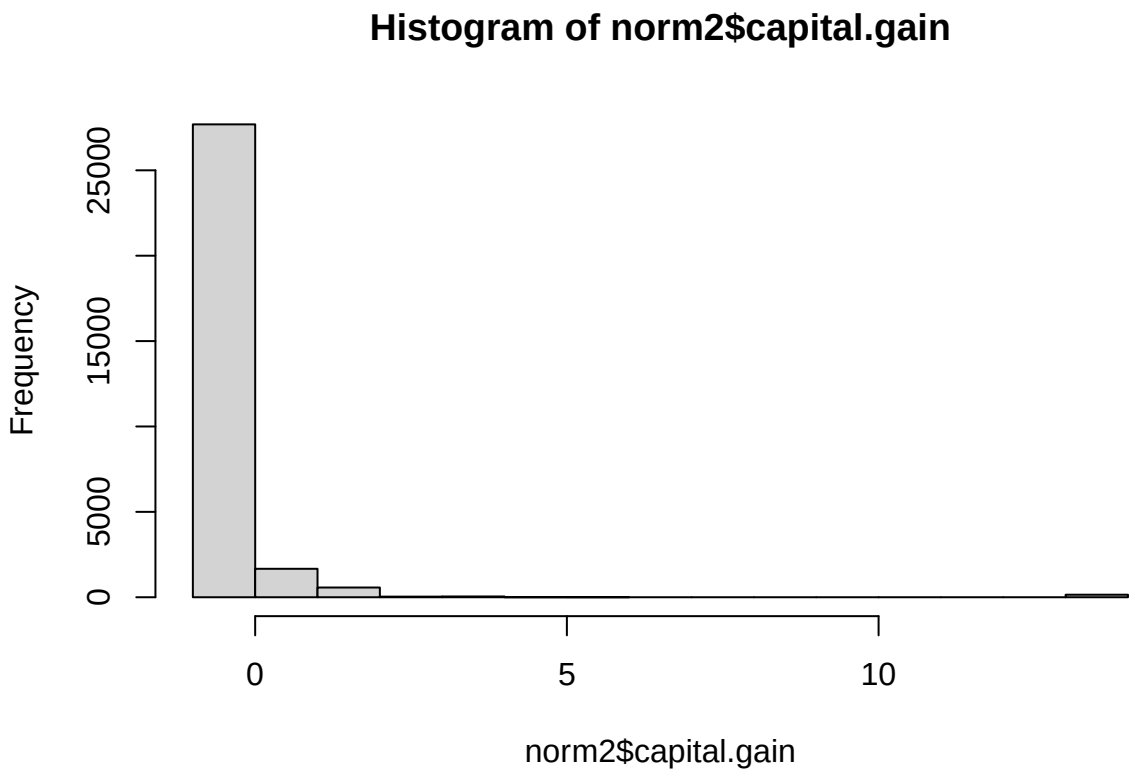

```
hist(norm2$sex)
```



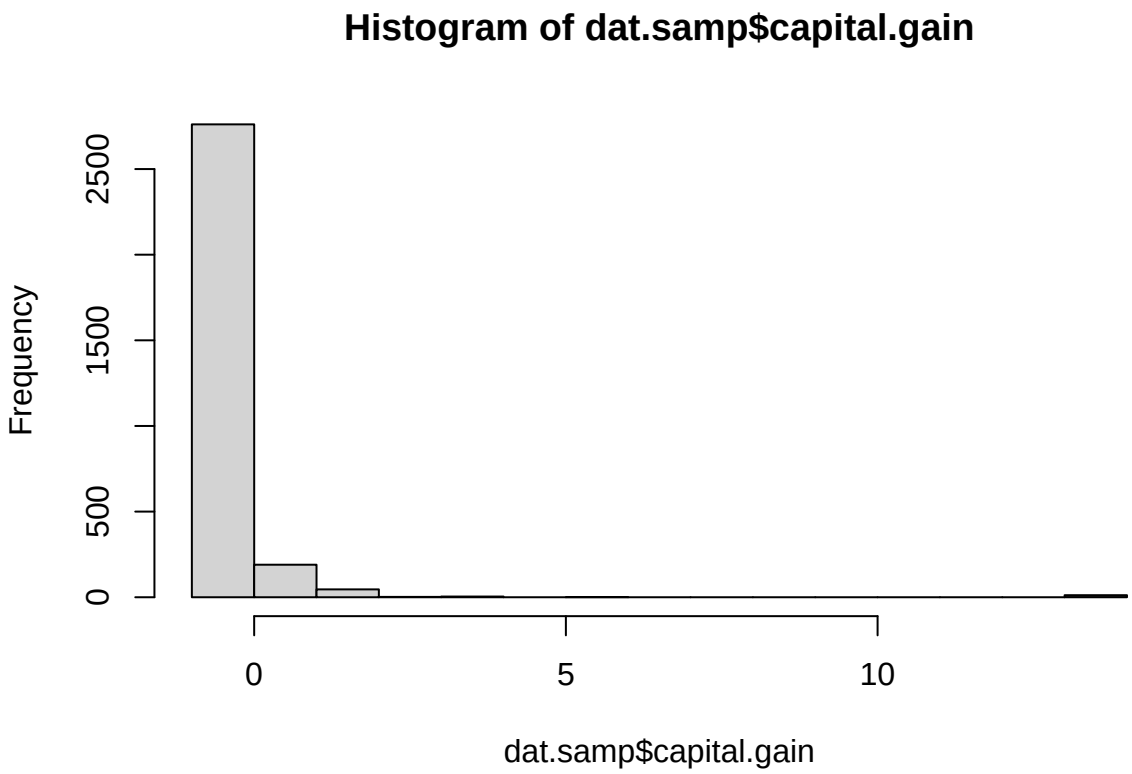

```
hist(dat.samp$sex)
```



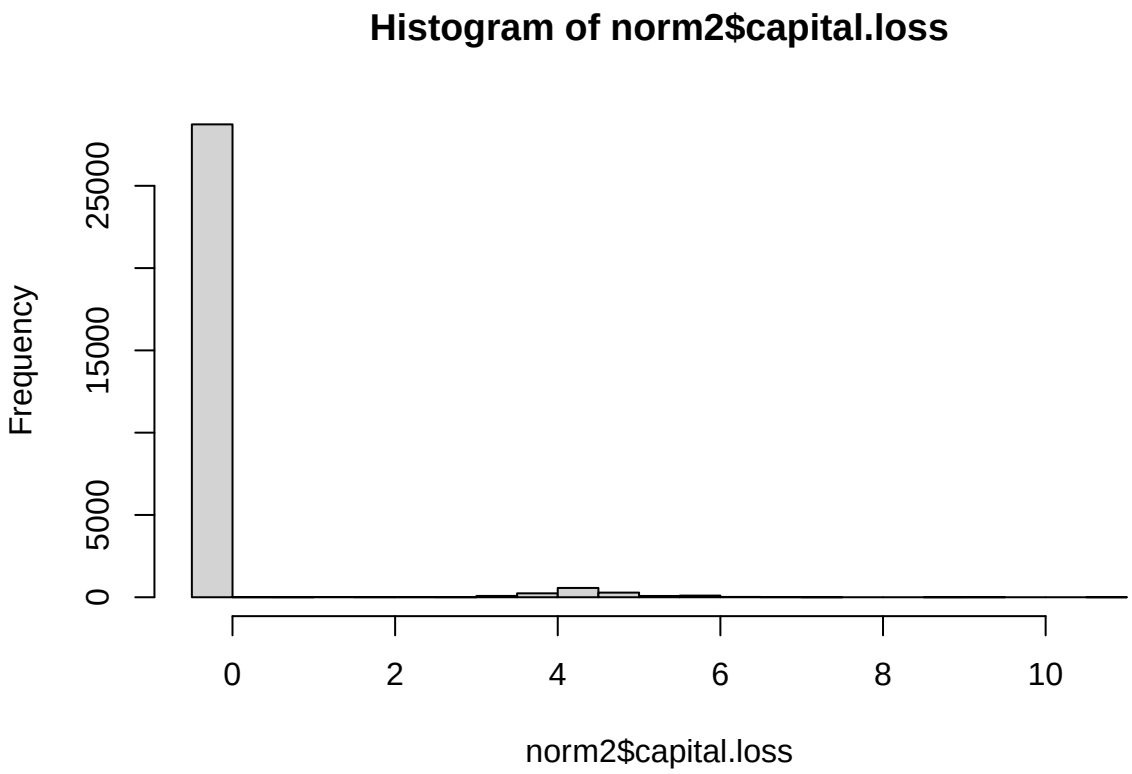

```
hist(norm2$capital.gain)
```



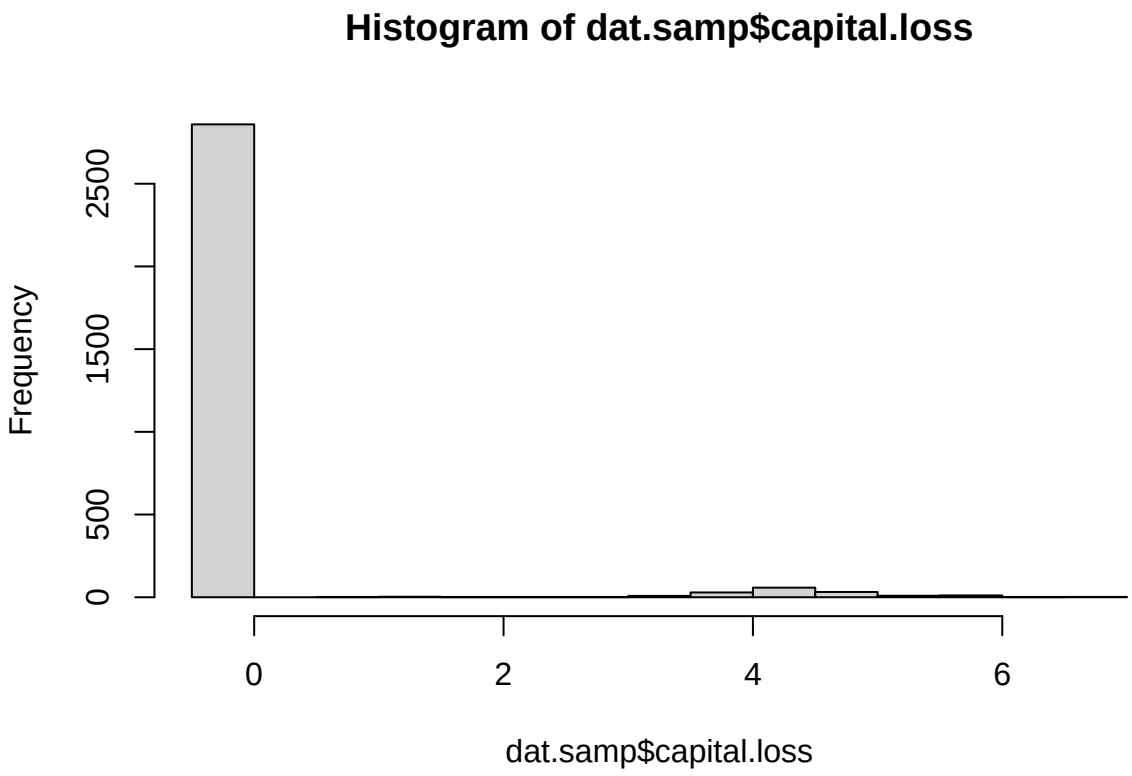

```
hist(dat.samp$capital.gain)
```



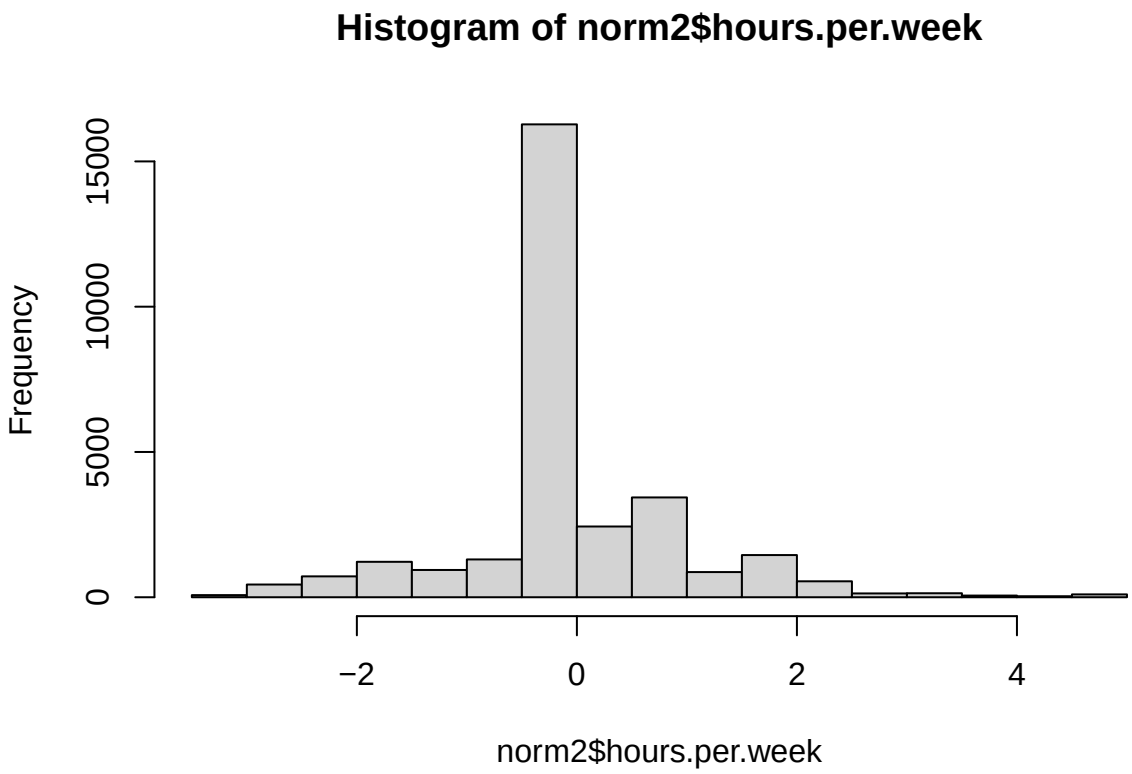

```
hist(norm2$capital.loss)
```



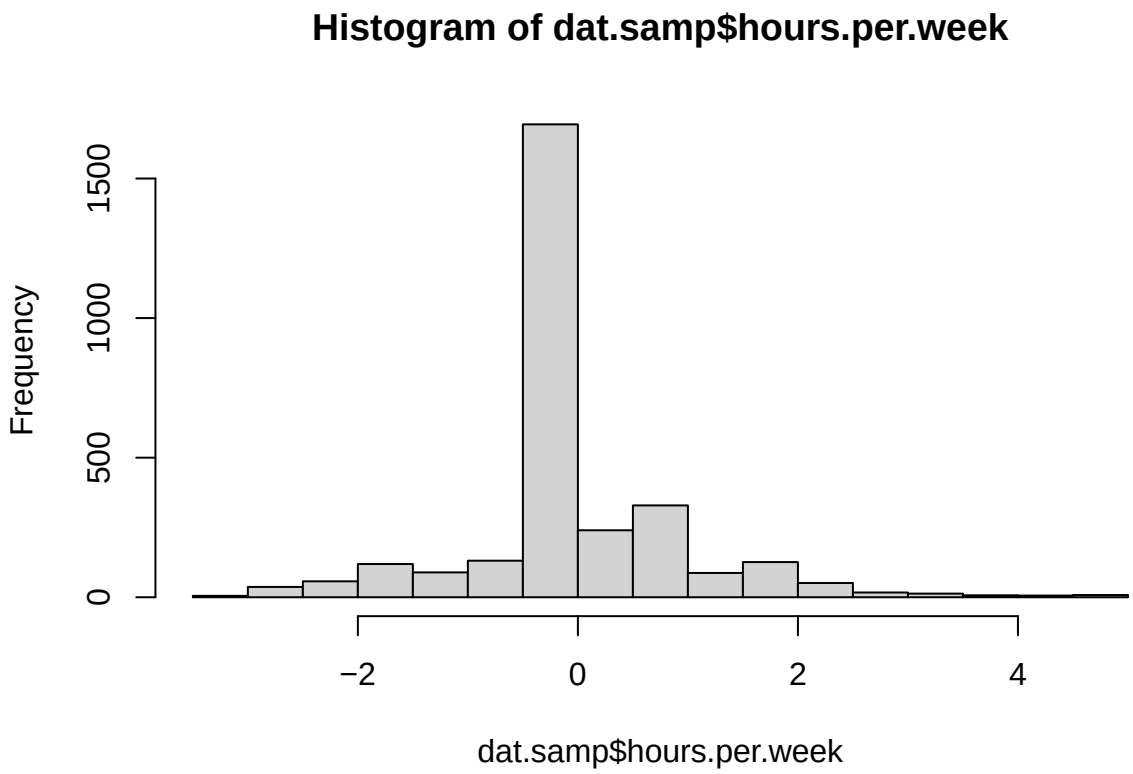

```
hist(dat.samp$capital.loss)
```



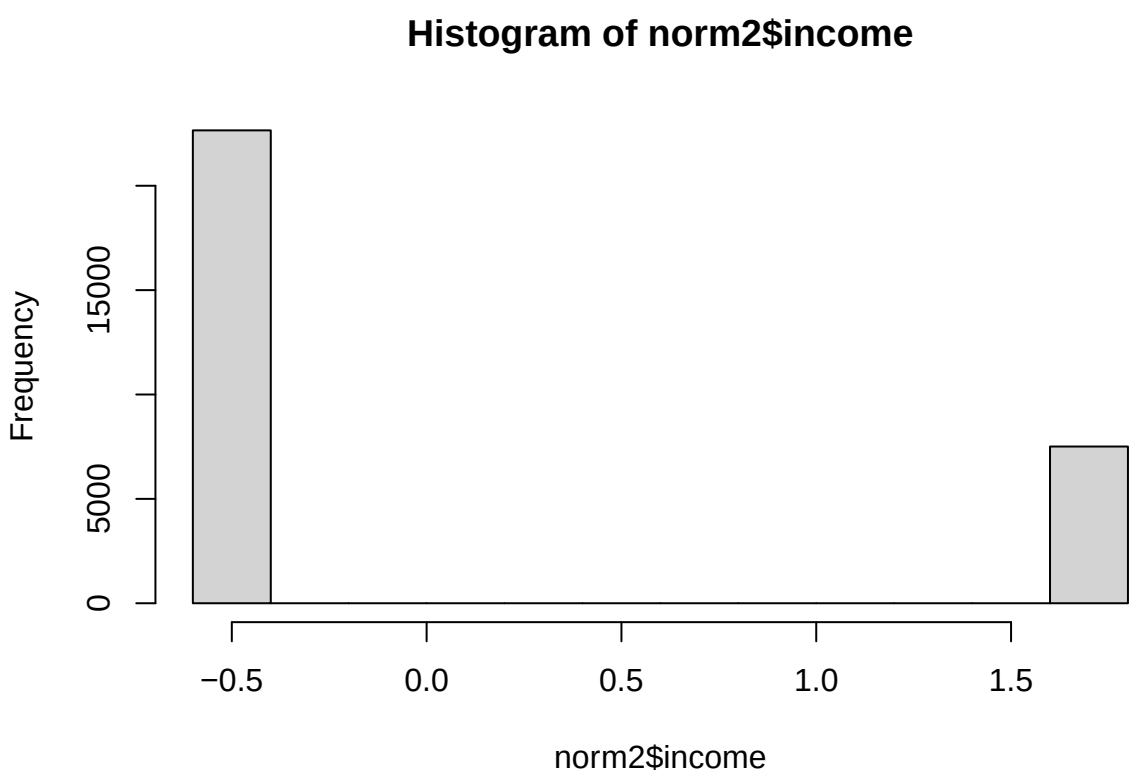

```
hist(norm2$hours.per.week)
```



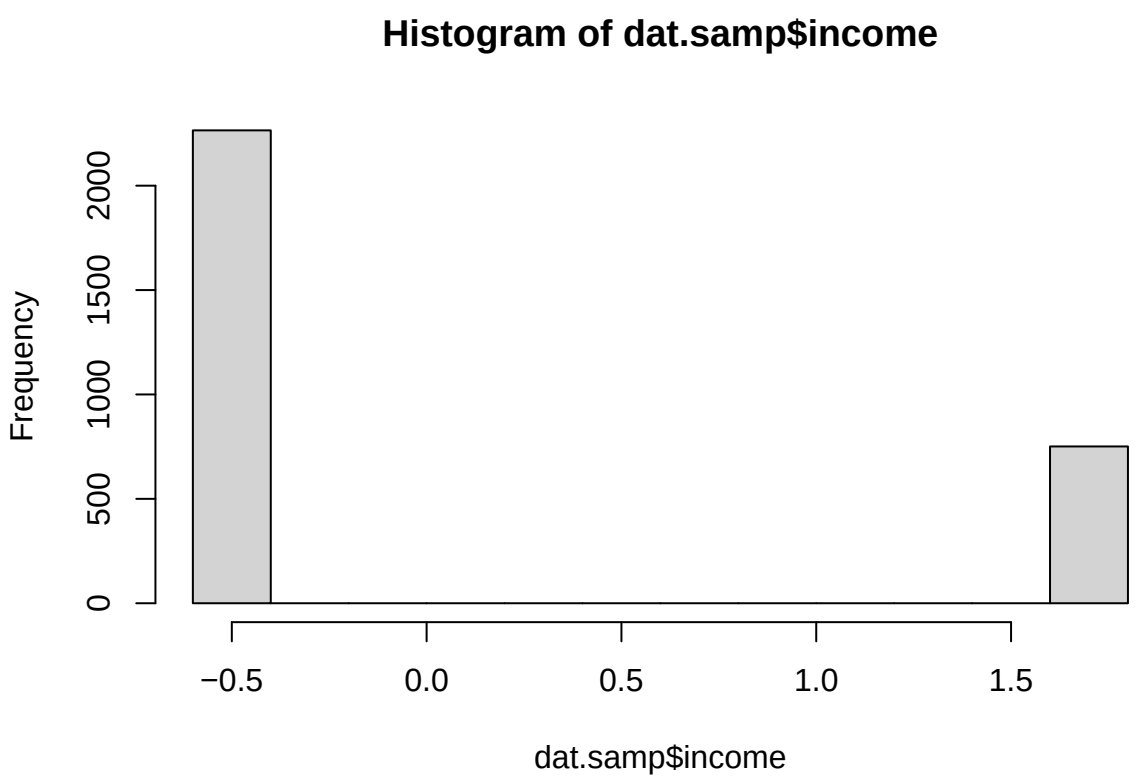

```
hist(dat.samp$hours.per.week)
```




```
hist(norm2$income)
```




```
hist(dat.samp$income)
```




```
# Split  Data For Gower/K-Med into Discard and Test
set.seed(100)
spl_gow = sample.split(norm1$income, SplitRatio = 0.9)
discard_gow = subset(norm1, spl_gow==TRUE)
test_gow = subset(norm1, spl_gow==FALSE)

# Distance matrix

gower_dist <- daisy(test_gow, metric = "gower")
gower_mat <- as.matrix(gower_dist)


# Print most similar
test_gow[which(gower_mat == min(gower_mat[gower_mat != min(gower_mat)]), arr.ind = TRUE)[1, ], ]
```

```
##             age education.num     marital.status relationship  sex capital.gain
## 3566 -0.4140115    -0.4397309 Married-civ-spouse      Husband Male    0.3131088
## 3481 -0.4140115    -0.4397309 Married-civ-spouse      Husband Male    0.3630659
##      capital.loss hours.per.week income
## 3566   -0.2185824    -0.07773282  <=50K
## 3481   -0.2185824    -0.07773282  <=50K
```

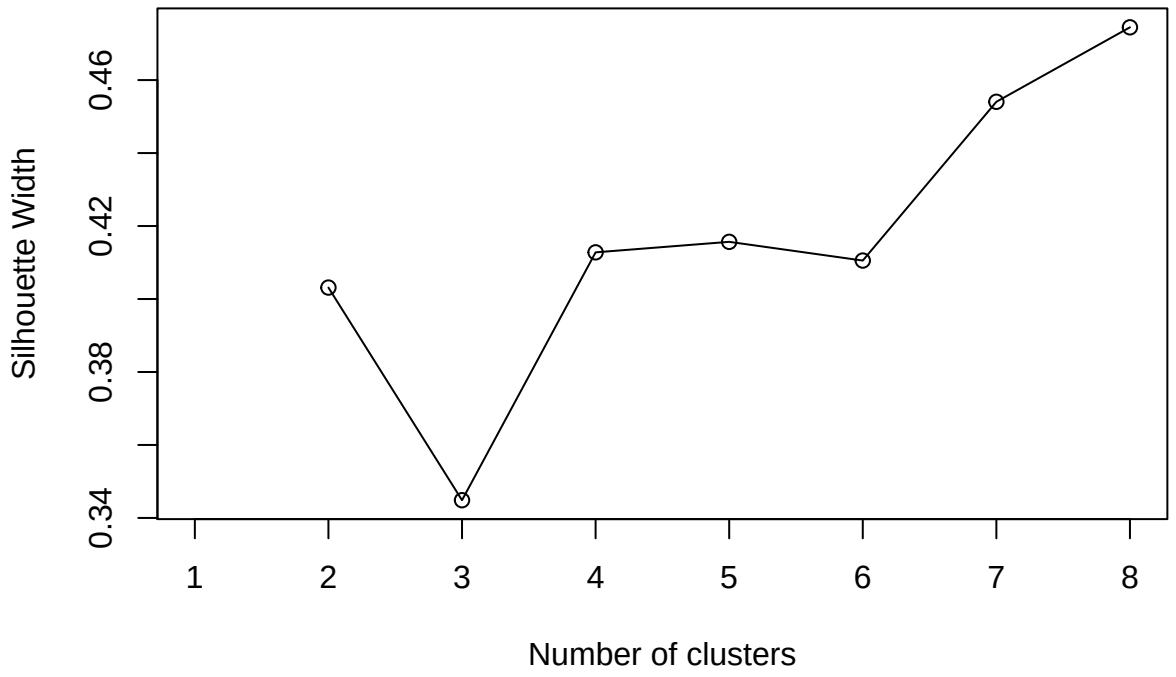
```
# Print most dissimilar
test_gow[which(gower_mat == max(gower_mat[gower_mat != max(gower_mat)]), arr.ind = TRUE)[1, ], ]
```

```
##            age education.num     marital.status relationship    sex
## 1609  2.707499      1.913215 Married-civ-spouse      Husband   Male
```

```
## 1242 -1.632162     -1.616204      Never-married    Own-child Female
##      capital.gain capital.loss hours.per.week income
## 1609   13.3543566   -0.2185824      0.7569928   >50K
## 1242   -0.1474422    3.7438378     -1.9141292  <=50K
```

```r
# Search for meaningful and easy to remember clusters (2-8), plot

sil_width <- c(NA)
for(i in 2:8){
  pam_fit <- pam(gower_dist, diss = TRUE, k = i)
  sil_width[i] <- pam_fit$silinfo$avg.width
}
plot(1:8, sil_width,
     xlab = "Number of clusters",
     ylab = "Silhouette Width",lines(1:8, sil_width))
```




```r
# Pick k=7 due to simplicity despite it being less accurate, summary for each cluster

k <- 7
pam_fit <- pam(gower_dist, diss = TRUE, k)
pam_results <- test_gow %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))
pam_results$the_summary
```

```
## [[1]]
```

```
##       age          education.num                   marital.status
##  Min.   :-1.2515   Min.   :-3.1848   Divorced              : 18
##  1st Qu.:-0.1856   1st Qu.:-0.4397   Married-AF-spouse     :  1
##  Median : 0.3473   Median : 0.7367   Married-civ-spouse    :565
##  Mean   : 0.4481   Mean   : 0.5446   Married-spouse-absent:  4
##  3rd Qu.: 1.0325   3rd Qu.: 1.1289   Never-married         :  1
##  Max.   : 3.9257   Max.   : 2.3054   Separated             :  3
##                                      Widowed               :  4
##          relationship     sex       capital.gain      capital.loss
##  Husband       :561   Female:  0   Min.   :-0.1474   Min.   :-0.2186
##  Not-in-family : 20   Male  :596   1st Qu.:-0.1474   1st Qu.:-0.2186
##  Other-relative:  3                Median :-0.1474   Median :-0.2186
##  Own-child     :  4                Mean   : 0.3195   Mean   : 0.3170
##  Unmarried     :  8                3rd Qu.:-0.1474   3rd Qu.:-0.2186
##  Wife          :  0                Max.   :13.3544   Max.   : 6.7664
##
##  hours.per.week       income       cluster
##  Min.   :-2.99927   <=50K:  0   Min.   :1
##  1st Qu.:-0.07773   >50K :596   1st Qu.:1
##  Median : 0.33963               Median :1
##  Mean   : 0.47983               Mean   :1
##  3rd Qu.: 0.75699               3rd Qu.:1
##  Max.   : 4.84715               Max.   :1
##
##
## [[2]]
##       age           education.num                     marital.status
##  Min.   :-1.47989   Min.   :-3.57699   Divorced              :240
##  1st Qu.:-0.03334   1st Qu.:-0.43973   Married-AF-spouse     :  0
##  Median : 0.49960   Median :-0.43973   Married-civ-spouse    :  0
##  Mean   : 0.51935   Mean   :-0.15833   Married-spouse-absent:  6
##  3rd Qu.: 0.95641   3rd Qu.:-0.04757   Never-married         : 20
##  Max.   : 2.63137   Max.   : 2.30537   Separated             : 41
##                                        Widowed               : 40
##          relationship     sex       capital.gain       capital.loss
##  Husband       :  0   Female:309   Min.   :-0.14744   Min.   :-0.2186
##  Not-in-family : 71   Male  : 38   1st Qu.:-0.14744   1st Qu.:-0.2186
##  Other-relative: 20                Median :-0.14744   Median :-0.2186
##  Own-child     : 13                Mean   :-0.07278   Mean   :-0.1211
##  Unmarried     :243                3rd Qu.:-0.14744   3rd Qu.:-0.2186
##  Wife          :  0                Max.   :13.35436   Max.   : 6.1109
##
##  hours.per.week       income       cluster
##  Min.   :-2.99927   <=50K:330   Min.   :2
##  1st Qu.:-0.16120   >50K : 17   1st Qu.:2
##  Median :-0.07773               Median :2
##  Mean   :-0.10516               Mean   :2
##  3rd Qu.:-0.07773               3rd Qu.:2
##  Max.   : 4.84715               Max.   :2
##
##
## [[3]]
##       age          education.num                     marital.status
##  Min.   :-1.6322   Min.   :-3.57699   Divorced              : 39
```

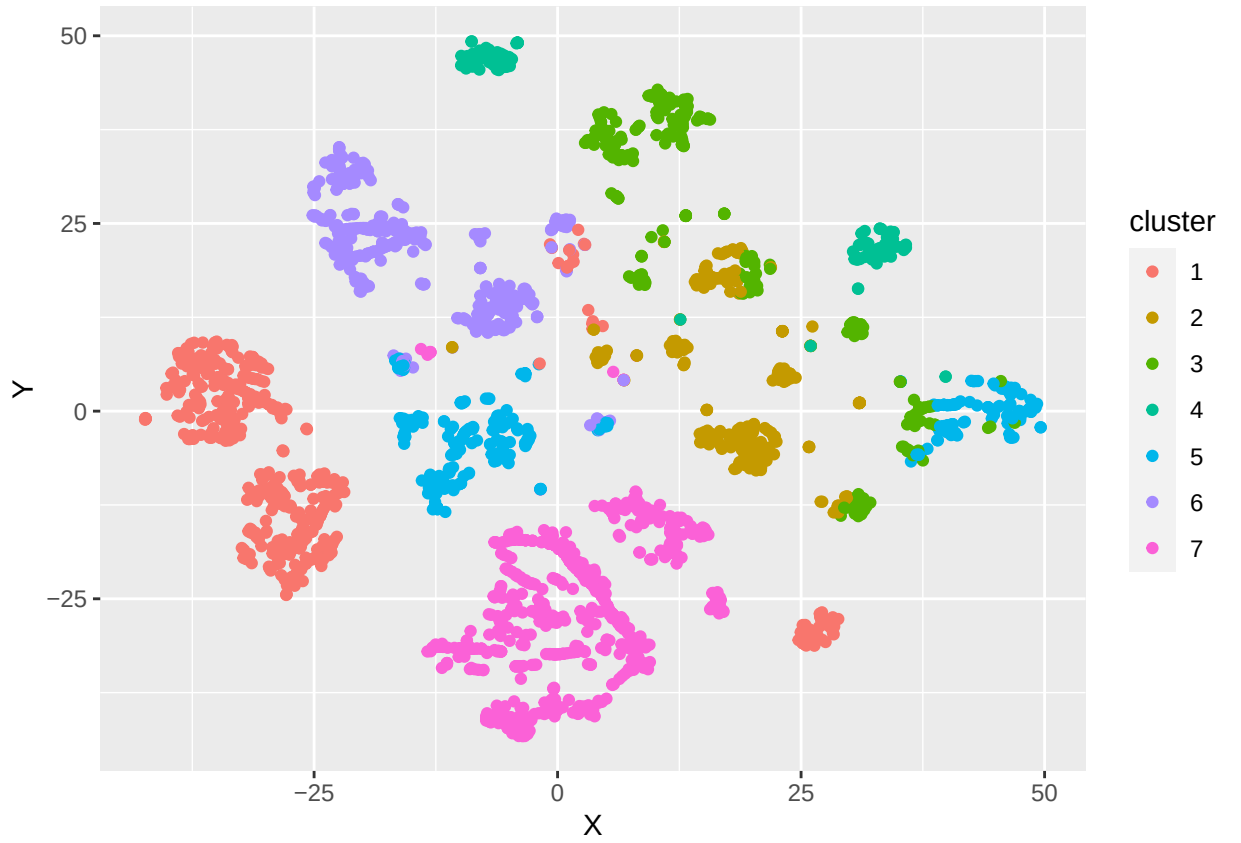
```
##  1st Qu.:-0.9470   1st Qu.:-0.43973   Married-AF-spouse    :  1
##  Median :-0.5663   Median :-0.04757   Married-civ-spouse   :  0
##  Mean   :-0.2764   Mean   : 0.13116   Married-spouse-absent:  6
##  3rd Qu.: 0.1189   3rd Qu.: 1.12890   Never-married        :283
##  Max.   : 3.9257   Max.   : 2.30537   Separated            : 16
##                                       Widowed              : 28
##          relationship     sex       capital.gain       capital.loss
##  Husband       :  0   Female:373   Min.   :-0.14744   Min.   :-0.21858
##  Not-in-family :260   Male  :  0   1st Qu.:-0.14744   1st Qu.:-0.21858
##  Other-relative: 31                Median :-0.14744   Median :-0.21858
##  Own-child     : 50                Mean   :-0.08981   Mean   :-0.07228
##  Unmarried     : 32                3rd Qu.:-0.14744   3rd Qu.:-0.21858
##  Wife          :  0                Max.   : 3.60988   Max.   : 5.82646
##
##  hours.per.week       income       cluster
##  Min.   :-2.74885   <=50K:352   Min.   :3
##  1st Qu.:-0.49510   >50K : 21   1st Qu.:3
##  Median :-0.07773               Median :3
##  Mean   :-0.20082               Mean   :3
##  3rd Qu.:-0.07773               3rd Qu.:3
##  Max.   : 2.84381               Max.   :3
##
##
## [[4]]
##       age           education.num                    marital.status
##  Min.   :-1.25149   Min.   :-3.18483   Divorced             :  0
##  1st Qu.:-0.41401   1st Qu.:-0.43973   Married-AF-spouse    :  2
##  Median : 0.04279   Median :-0.04757   Married-civ-spouse   :150
##  Mean   : 0.13072   Mean   : 0.06628   Married-spouse-absent:  0
##  3rd Qu.: 0.65187   3rd Qu.: 0.93282   Never-married        :  0
##  Max.   : 2.63137   Max.   : 2.30537   Separated            :  2
##                                        Widowed              :  1
##          relationship     sex       capital.gain       capital.loss
##  Husband       :  0   Female:155   Min.   :-0.14744   Min.   :-0.2186
##  Not-in-family :  0   Male  :  0   1st Qu.:-0.14744   1st Qu.:-0.2186
##  Other-relative:  4                Median :-0.14744   Median :-0.2186
##  Own-child     :  4                Mean   :-0.05116   Mean   : 0.1163
##  Unmarried     :  0                3rd Qu.:-0.14744   3rd Qu.:-0.2186
##  Wife          :147                Max.   : 2.55983   Max.   : 5.7547
##
##  hours.per.week       income      cluster
##  Min.   :-2.99927   <=50K:81   Min.   :4
##  1st Qu.:-0.66204   >50K :74   1st Qu.:4
##  Median :-0.07773              Median :4
##  Mean   :-0.28668              Mean   :4
##  3rd Qu.:-0.07773              3rd Qu.:4
##  Max.   : 4.84715              Max.   :4
##
##
## [[5]]
##       age          education.num                    marital.status
##  Min.   :-1.632   Min.   :-3.18483   Divorced             : 13
##  1st Qu.:-1.404   1st Qu.:-0.43973   Married-AF-spouse    :  0
##  Median :-1.251   Median :-0.04757   Married-civ-spouse   :  4
```

```
##  Mean   :-1.123   Mean   :-0.25576   Married-spouse-absent:  3
##  3rd Qu.:-1.023   3rd Qu.:-0.04757   Never-married        :378
##  Max.   : 1.565   Max.   : 2.30537   Separated            :  7
##                                      Widowed              :  0
##         relationship     sex       capital.gain      capital.loss
##  Husband       :  0   Female:140   Min.   :-0.1474   Min.   :-0.2186
##  Not-in-family :  0   Male  :265   1st Qu.:-0.1474   1st Qu.:-0.2186
##  Other-relative: 18                Median :-0.1474   Median :-0.2186
##  Own-child     :371                Mean   :-0.1377   Mean   :-0.0686
##  Unmarried     : 16                3rd Qu.:-0.1474   3rd Qu.:-0.2186
##  Wife          :  0                Max.   : 1.7542   Max.   : 5.5667
##
##  hours.per.week       income       cluster
##  Min.   :-3.33316   <=50K:402   Min.   :5
##  1st Qu.:-1.32982   >50K :  3   1st Qu.:5
##  Median :-0.07773               Median :5
##  Mean   :-0.62309               Mean   :5
##  3rd Qu.:-0.07773               3rd Qu.:5
##  Max.   : 2.42644               Max.   :5
##
##
## [[6]]
##       age          education.num                    marital.status
##  Min.   :-1.5560   Min.   :-3.57699   Divorced             :116
##  1st Qu.:-0.8708   1st Qu.:-0.43973   Married-AF-spouse    :  0
##  Median :-0.4140   Median :-0.04757   Married-civ-spouse   :  0
##  Mean   :-0.2074   Mean   :-0.08741   Married-spouse-absent: 10
##  3rd Qu.: 0.2712   3rd Qu.: 0.73674   Never-married        :293
##  Max.   : 3.3927   Max.   : 2.30537   Separated            : 18
##                                       Widowed              :  6
##         relationship     sex       capital.gain       capital.loss
##  Husband       :  0   Female:  0   Min.   :-0.14744   Min.   :-0.2186
##  Not-in-family :416   Male  :443   1st Qu.:-0.14744   1st Qu.:-0.2186
##  Other-relative: 13                Median :-0.14744   Median :-0.2186
##  Own-child     :  0                Mean   :-0.08976   Mean   :-0.0405
##  Unmarried     : 14                3rd Qu.:-0.14744   3rd Qu.:-0.2186
##  Wife          :  0                Max.   : 3.60988   Max.   : 5.3664
##
##  hours.per.week       income       cluster
##  Min.   :-2.99927   <=50K:403   Min.   :6
##  1st Qu.:-0.07773   >50K : 40   1st Qu.:6
##  Median :-0.07773               Median :6
##  Mean   : 0.05209               Mean   :6
##  3rd Qu.: 0.33963               3rd Qu.:6
##  Max.   : 4.09589               Max.   :6
##
##
## [[7]]
##       age          education.num                    marital.status
##  Min.   :-1.4038   Min.   :-3.57699   Divorced             :  0
##  1st Qu.:-0.4901   1st Qu.:-0.43973   Married-AF-spouse    :  0
##  Median : 0.1189   Median :-0.43973   Married-civ-spouse   :686
##  Mean   : 0.2608   Mean   :-0.31370   Married-spouse-absent:  2
##  3rd Qu.: 0.8803   3rd Qu.:-0.04757   Never-married        :  0
```

```
##  Max.   : 3.0882   Max.   : 2.30537   Separated           :  7
##                                       Widowed             :  2
##          relationship     sex       capital.gain      capital.loss
##  Husband       :683   Female:  0   Min.   :-0.1474   Min.   :-0.21858
##  Not-in-family :  0   Male  :697   1st Qu.:-0.1474   1st Qu.:-0.21858
##  Other-relative:  7                Median :-0.1474   Median :-0.21858
##  Own-child     :  1                Mean   :-0.1076   Mean   :-0.05231
##  Unmarried     :  6                3rd Qu.:-0.1474   3rd Qu.:-0.21858
##  Wife          :  0                Max.   : 5.4302   Max.   : 5.17100
##
##  hours.per.week       income       cluster
##  Min.   :-3.24969   <=50K:697   Min.   :7
##  1st Qu.:-0.07773   >50K :  0   1st Qu.:7
##  Median :-0.07773               Median :7
##  Mean   : 0.17819               Mean   :7
##  3rd Qu.: 0.42310               3rd Qu.:7
##  Max.   : 4.84715               Max.   :7
##
```

Cluster Plot

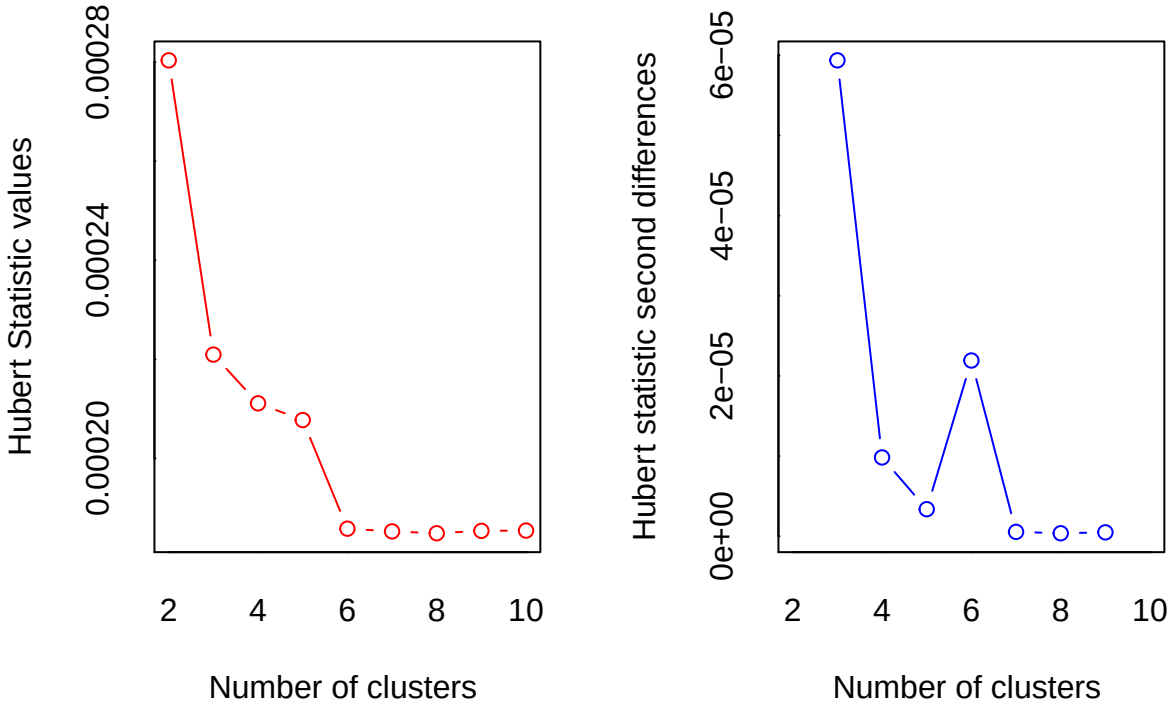
```
tsne_obj <- Rtsne(gower_dist, is_distance = TRUE)
tsne_data <- tsne_obj$Y %>%
  data.frame() %>%
  setNames(c("X", "Y")) %>%
  mutate(cluster = factor(pam_fit$clustering))
ggplot(aes(x = X, y = Y), data = tsne_data) +
  geom_point(aes(color = cluster))
```



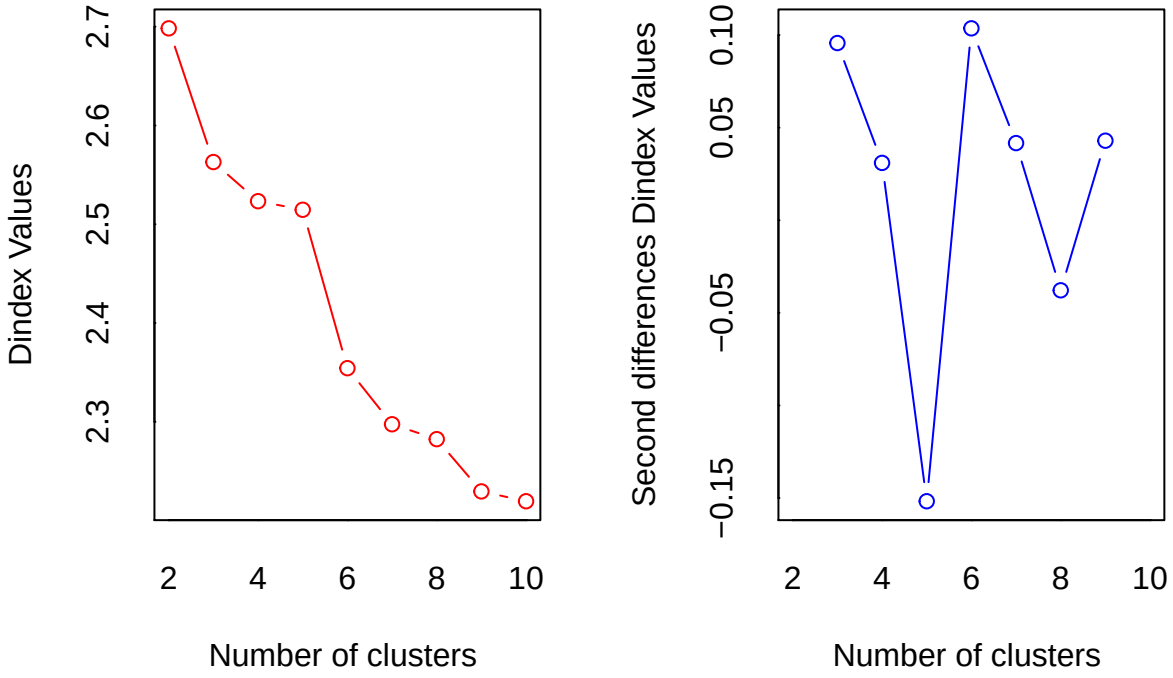
Hierarchical Clustering

```
#Split data
set.seed(100)
spl_hc = sample.split(norm2$income, SplitRatio = 0.9)
discard_hc = subset(norm2, spl_hc==TRUE)
test_hc = subset(norm2, spl_hc==FALSE)

# Determine optimal number of clusters
opt_clust <- test_hc %>%
     scale() %>%
     NbClust(distance = "euclidean",
             min.nc = 2, max.nc = 10,
             method = "complete", index ="all")
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```




```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 7 proposed 2 as the best number of clusters
## * 12 proposed 3 as the best number of clusters
## * 2 proposed 6 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 2 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```

```
fviz_nbclust(opt_clust, ggtheme = theme_minimal())
```

```
## Warning in if (class(best_nc) == "numeric") print(best_nc) else if
## (class(best_nc) == : the condition has length > 1 and only the first element
## will be used
```

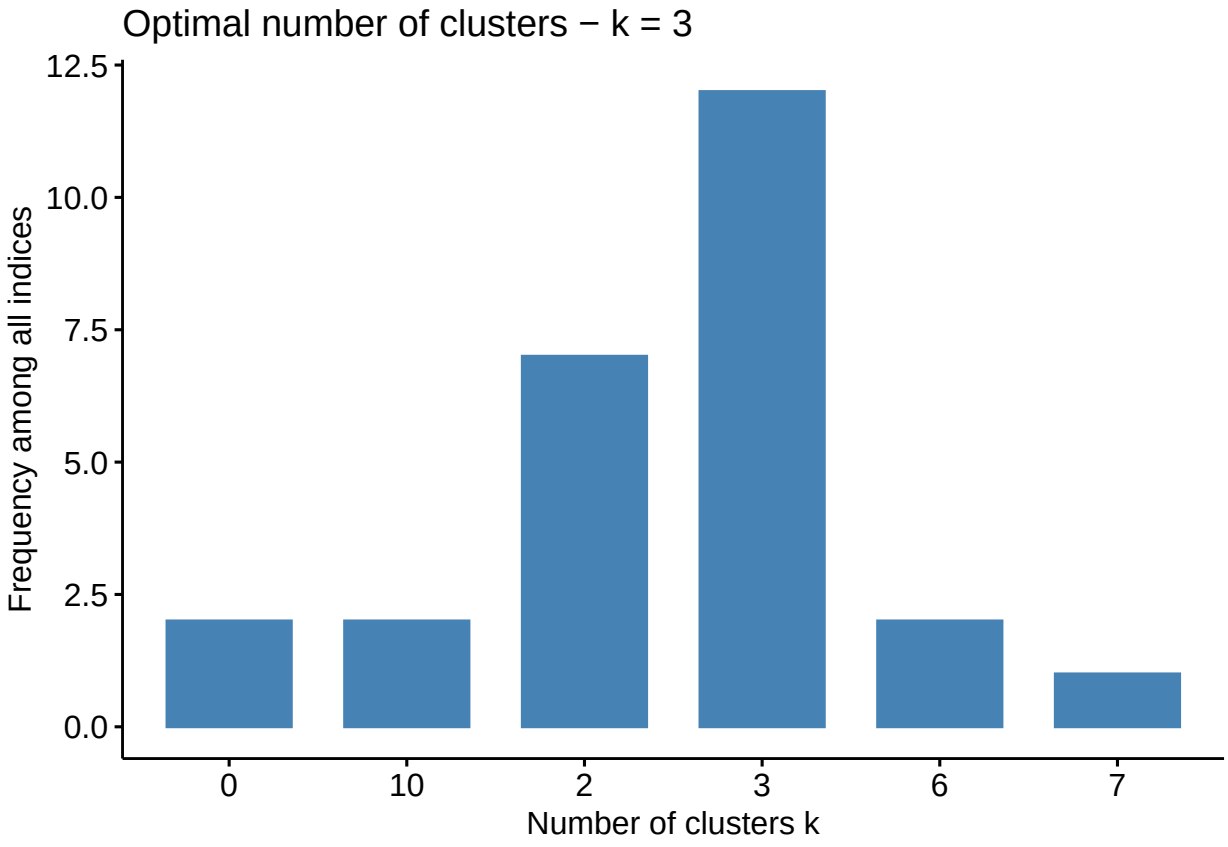
```
## Warning in if (class(best_nc) == "matrix") .viz_NbClust(x, print.summary, : the
```

```
## condition has length > 1 and only the first element will be used
```

```
## Warning in if (class(best_nc) == "numeric") print(best_nc) else if
## (class(best_nc) == : the condition has length > 1 and only the first element
## will be used
```

```
## Warning in if (class(best_nc) == "matrix") {: the condition has length > 1 and
## only the first element will be used
```

```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 7 proposed  2 as the best number of clusters
## * 12 proposed  3 as the best number of clusters
## * 2 proposed  6 as the best number of clusters
## * 1 proposed  7 as the best number of clusters
## * 2 proposed  10 as the best number of clusters
##
## Conclusion
## =========================
## * According to the majority rule, the best number of clusters is  3 .
```

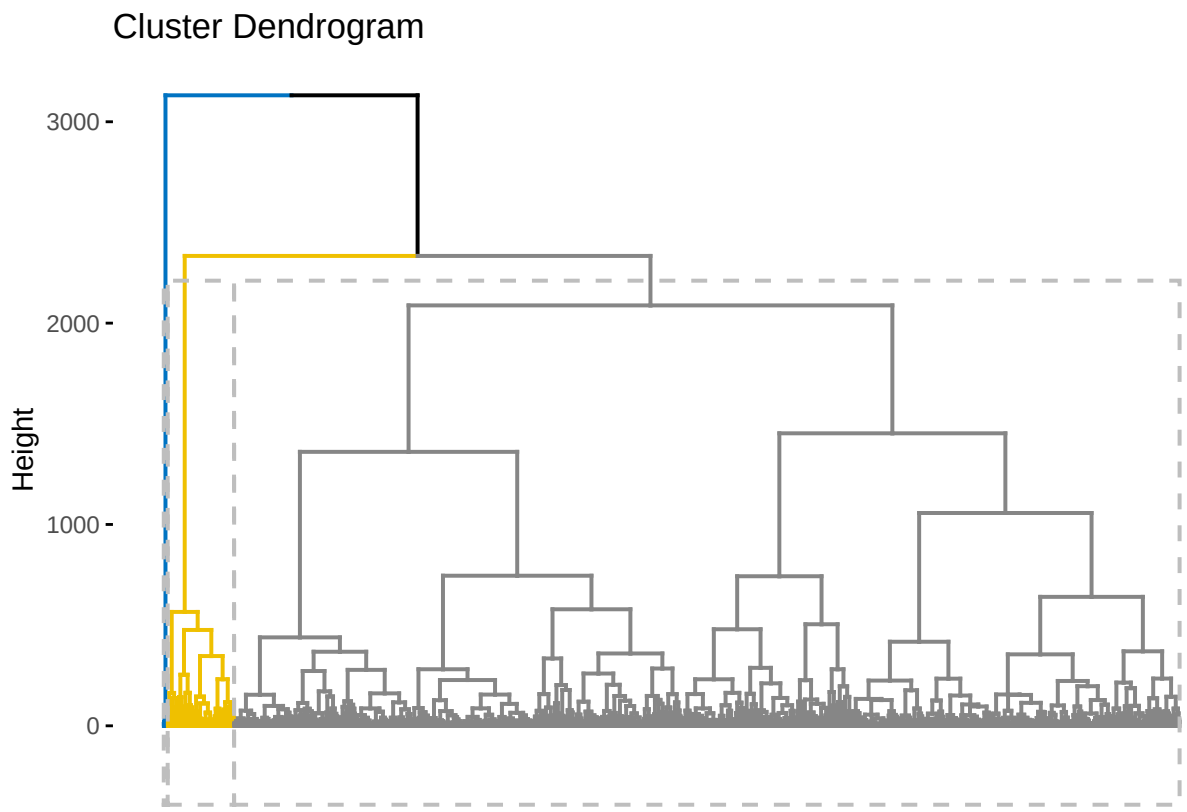



```r
# Hierarchical Clustering

adult_hc <- test_hc %>%
  scale() %>%
  dist(method = "euclidean") %>%
  eclust("hclust", k = 3, graph = FALSE)
```

Dendrogram

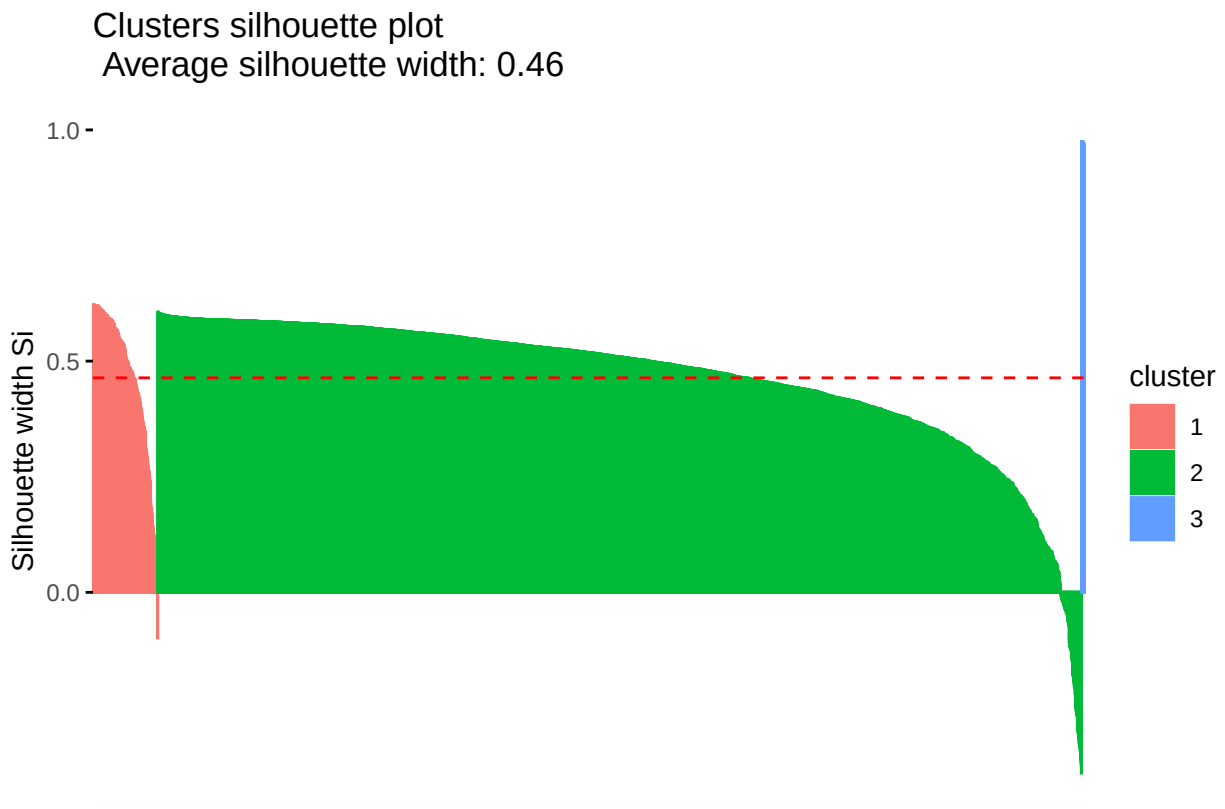
```
fviz_dend(adult_hc, palette = "jco",
          rect = TRUE, show_labels = FALSE)
```



Silhouette Validation

```
# Silhouette plot
fviz_silhouette(adult_hc)
```

```
##   cluster size ave.sil.width
## 1       1  197          0.46
## 2       2 2807          0.46
## 3       3   12          0.97
```




```
# Silhouette width of observations
sil <- adult_hc$silinfo$widths[, 1:3]

# Objects with negative silhouette
neg_sil_index <- which(sil[, 'sil_width'] < 0)
sil[neg_sil_index, , drop = FALSE]
```

```
##       cluster neighbor    sil_width
## 15193       1        2 -0.098925104
## 4098        2        1 -0.000745809
## 27369       2        1 -0.007323274
## 31428       2        1 -0.007507364
## 32486       2        1 -0.012639648
## 13346       2        1 -0.014709365
## 25312       2        1 -0.017993069
## 13210       2        1 -0.019081455
## 19366       2        1 -0.019930125
## 17324       2        1 -0.020369596
## 4128        2        1 -0.024059790
## 17063       2        1 -0.028021202
## 28156       2        1 -0.029440285
## 9387        2        1 -0.034019851
## 2207        2        1 -0.034304714
## 24356       2        1 -0.035884165
## 17524       2        1 -0.038477593
## 2958        2        1 -0.039302248
```

```
## 7985        2        1 -0.044049437
## 2229        2        1 -0.044333015
## 6951        2        1 -0.044514864
## 10580       2        1 -0.047017952
## 29918       2        1 -0.058839646
## 15099       2        1 -0.061337459
## 10417       2        1 -0.063364033
## 3194        2        1 -0.069565257
## 32171       2        1 -0.071135974
## 1494        2        1 -0.106853644
## 27792       2        1 -0.115471849
## 2369        2        1 -0.116344634
## 25919       2        1 -0.119659341
## 27375       2        1 -0.121316740
## 28826       2        1 -0.123385328
## 16978       2        1 -0.124871886
## 13173       2        1 -0.135585739
## 20853       2        1 -0.144819327
## 2064        2        1 -0.146782756
## 20031       2        1 -0.158915745
## 28951       2        1 -0.179049341
## 4086        2        1 -0.180477944
## 17879       2        1 -0.181326016
## 19011       2        1 -0.193393505
## 23542       2        1 -0.200885286
## 5911        2        1 -0.205065008
## 6422        2        1 -0.206165882
## 7873        2        1 -0.218247084
## 2373        2        1 -0.226879419
## 16312       2        1 -0.249994625
## 10735       2        1 -0.255499357
## 9445        2        1 -0.261592932
## 16489       2        1 -0.264201842
## 29804       2        1 -0.267550077
## 5503        2        1 -0.298343613
## 9956        2        1 -0.299352966
## 31905       2        1 -0.303055552
## 3214        2        1 -0.307257445
## 27862       2        1 -0.315975542
## 1772        2        1 -0.324264198
## 20622       2        1 -0.327830736
## 32203       2        1 -0.333010439
## 9605        2        1 -0.339157728
## 27499       2        1 -0.340872136
## 25614       2        1 -0.359911585
## 2300        2        1 -0.369899258
## 6090        2        1 -0.390914406
```