week1_R_solution.R

Nate George

January 9, 2018

Load libraries we will use.

```
library(data.table)
library(DMwR)
library(corrplot)
library(raster)
library(ggplot2)
```

Load the data; replace the placeholder -9 with NA. Throw out the dm column (history of diabetes) because almost all values are missing.

```
fn <- '/home/nate/Dropbox/MSDS/MSDS680_ncg_S8W1_18/week1/heart.disease.data'
dt <- fread(fn)
summary(dt)
```

```
##       age             sex               cp           trestbps    
##  Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0  
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :3.000   Median :130.0  
##  Mean   :54.41   Mean   :0.6773   Mean   :3.163   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0  
##       chol            cigs           years            fbs        
##  Min.   :126.0   Min.   :-9.00   Min.   :-9.00   Min.   :0.0000  
##  1st Qu.:213.0   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:0.0000  
##  Median :244.0   Median :10.00   Median :15.00   Median :0.0000  
##  Mean   :249.1   Mean   :16.46   Mean   :14.83   Mean   :0.1489  
##  3rd Qu.:277.0   3rd Qu.:30.00   3rd Qu.:30.00   3rd Qu.:0.0000  
##  Max.   :564.0   Max.   :99.00   Max.   :54.00   Max.   :1.0000  
##        dm            famhist          restecg         thalach     
##  Min.   :-9.000   Min.   :0.0000   Min.   :0.000   Min.   : 71.0  
##  1st Qu.:-9.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:133.2  
##  Median :-9.000   Median :1.0000   Median :2.000   Median :153.5  
##  Mean   :-8.184   Mean   :0.6206   Mean   :1.014   Mean   :149.8  
##  3rd Qu.:-9.000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:165.8  
##  Max.   : 1.000   Max.   :1.0000   Max.   :2.000   Max.   :202.0  
##      exang             thal             num        
##  Min.   :0.0000   Min.   :-9.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.: 3.000   1st Qu.:0.0000  
##  Median :0.0000   Median : 3.000   Median :0.0000  
##  Mean   :0.3262   Mean   : 4.582   Mean   :0.9078  
##  3rd Qu.:1.0000   3rd Qu.: 7.000   3rd Qu.:2.0000  
##  Max.   :1.0000   Max.   : 7.000   Max.   :4.0000  
```

```
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   282 obs. of  15 variables:
##  $ age     : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : int  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp      : int  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: int  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ cigs    : int  50 40 20 0 0 20 0 0 0 20 ...
##  $ years   : int  20 40 35 0 0 20 0 0 0 25 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ dm      : int  -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 ...
##  $ famhist : int  1 1 1 1 1 1 1 1 0 1 ...
##  $ restecg : int  2 2 2 0 2 0 2 0 2 2 ...
##  $ thalach : int  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : int  0 1 1 0 0 0 0 1 0 1 ...
##  $ thal    : int  6 3 7 3 3 3 3 3 7 7 ...
##  $ num     : int  0 2 1 0 0 0 3 0 2 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
dt <- dt[, lapply(.SD, as.numeric)]
# replace all -9 with NA
dt[dt == -9] <- NA
summary(dt)
```

```
##       age             sex                cp           trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :3.000   Median :130.0
##  Mean   :54.41   Mean   :0.6773   Mean   :3.163   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0
##
##       chol            cigs           years            fbs
##  Min.   :126.0   Min.   : 0.00   Min.   : 0.00   Min.   :0.0000
##  1st Qu.:213.0   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:0.0000
##  Median :244.0   Median :10.00   Median :15.00   Median :0.0000
##  Mean   :249.1   Mean   :16.92   Mean   :15.26   Mean   :0.1489
##  3rd Qu.:277.0   3rd Qu.:30.00   3rd Qu.:30.00   3rd Qu.:0.0000
##  Max.   :564.0   Max.   :99.00   Max.   :54.00   Max.   :1.0000
##                  NA's   :5       NA's   :5
##        dm          famhist           restecg         thalach
##  Min.   :1     Min.   :0.0000   Min.   :0.000   Min.   : 71.0
##  1st Qu.:1     1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:133.2
##  Median :1     Median :1.0000   Median :2.000   Median :153.5
##  Mean   :1     Mean   :0.6206   Mean   :1.014   Mean   :149.8
##  3rd Qu.:1     3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:165.8
##  Max.   :1     Max.   :1.0000   Max.   :2.000   Max.   :202.0
##  NA's   :259
##      exang             thal            num
##  Min.   :0.0000   Min.   :3.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:0.0000
##  Median :0.0000   Median :3.000   Median :0.0000
##  Mean   :0.3262   Mean   :4.679   Mean   :0.9078
##  3rd Qu.:1.0000   3rd Qu.:7.000   3rd Qu.:2.0000
##  Max.   :1.0000   Max.   :7.000   Max.   :4.0000
##                   NA's   :2
```

```r
dim(dt)
```

```
## [1] 282  15
```

```r
# almost all 'dm' values are missing, so throw out that column
dt[, dm:=NULL]
```

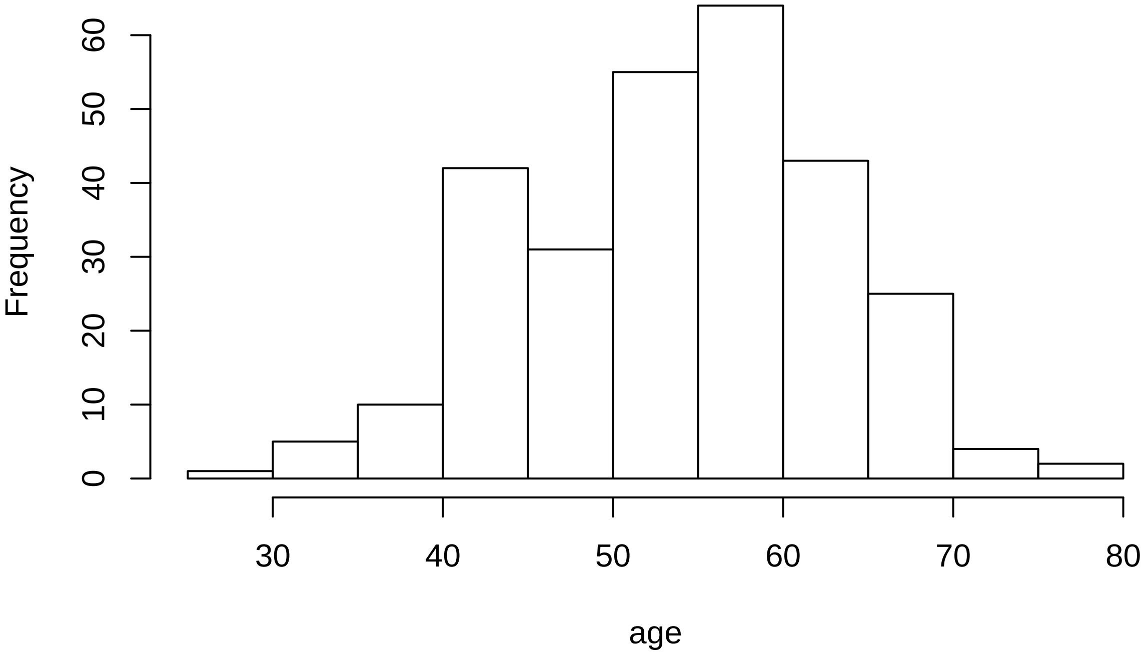
Impute NA values with KNN.

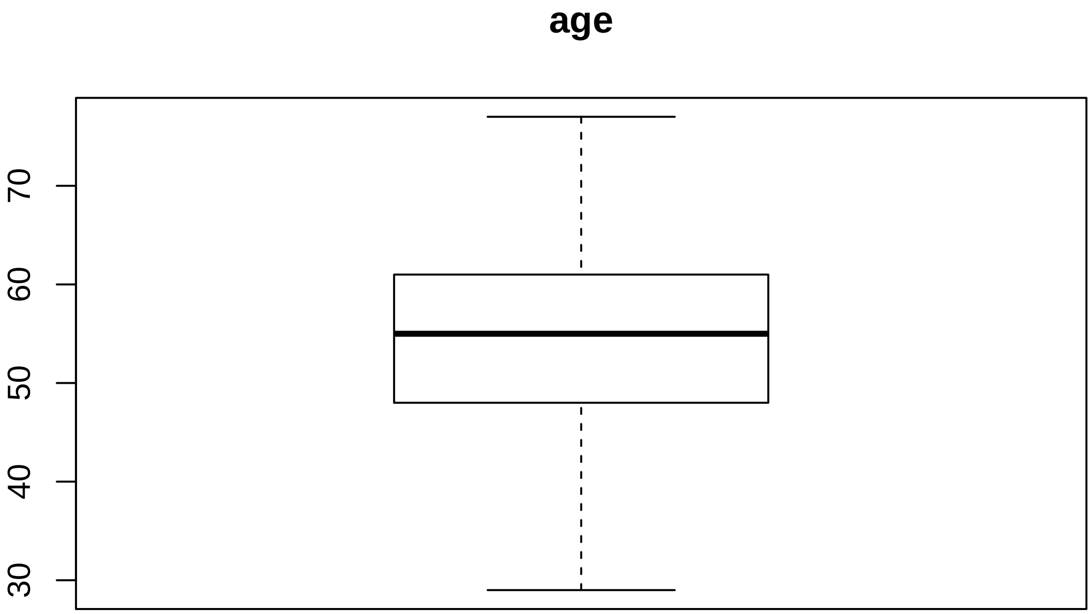
```r
# impute missing values
dt.nona <- knnImputation(dt)
summary(dt.nona)
```

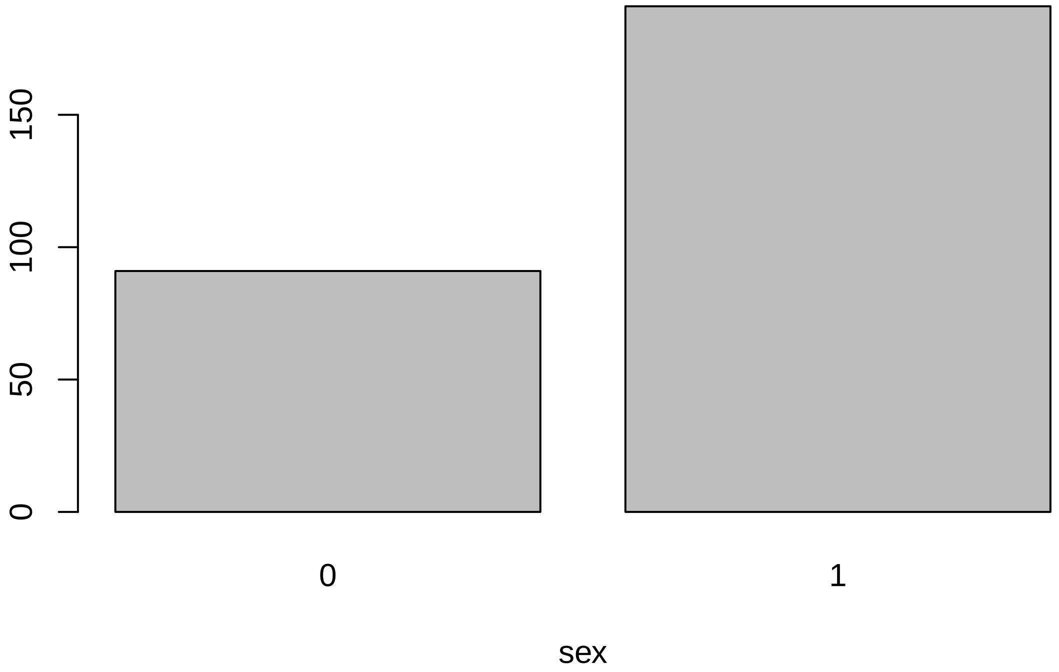
```
##       age             sex               cp           trestbps    
##  Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0  
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :3.000   Median :130.0  
##  Mean   :54.41   Mean   :0.6773   Mean   :3.163   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0  
##       chol            cigs           years            fbs        
##  Min.   :126.0   Min.   : 0.00   Min.   : 0.00   Min.   :0.0000  
##  1st Qu.:213.0   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:0.0000  
##  Median :244.0   Median :11.98   Median :15.00   Median :0.0000  
##  Mean   :249.1   Mean   :16.96   Mean   :15.35   Mean   :0.1489  
##  3rd Qu.:277.0   3rd Qu.:30.00   3rd Qu.:30.00   3rd Qu.:0.0000  
##  Max.   :564.0   Max.   :99.00   Max.   :54.00   Max.   :1.0000  
##     famhist          restecg         thalach          exang       
##  Min.   :0.0000   Min.   :0.000   Min.   : 71.0   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:133.2   1st Qu.:0.0000  
##  Median :1.0000   Median :2.000   Median :153.5   Median :0.0000  
##  Mean   :0.6206   Mean   :1.014   Mean   :149.8   Mean   :0.3262  
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:165.8   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :2.000   Max.   :202.0   Max.   :1.0000  
##       thal            num        
##  Min.   :3.000   Min.   :0.0000  
##  1st Qu.:3.000   1st Qu.:0.0000  
##  Median :3.000   Median :0.0000  
##  Mean   :4.677   Mean   :0.9078  
##  3rd Qu.:7.000   3rd Qu.:2.0000  
##  Max.   :7.000   Max.   :4.0000  
```

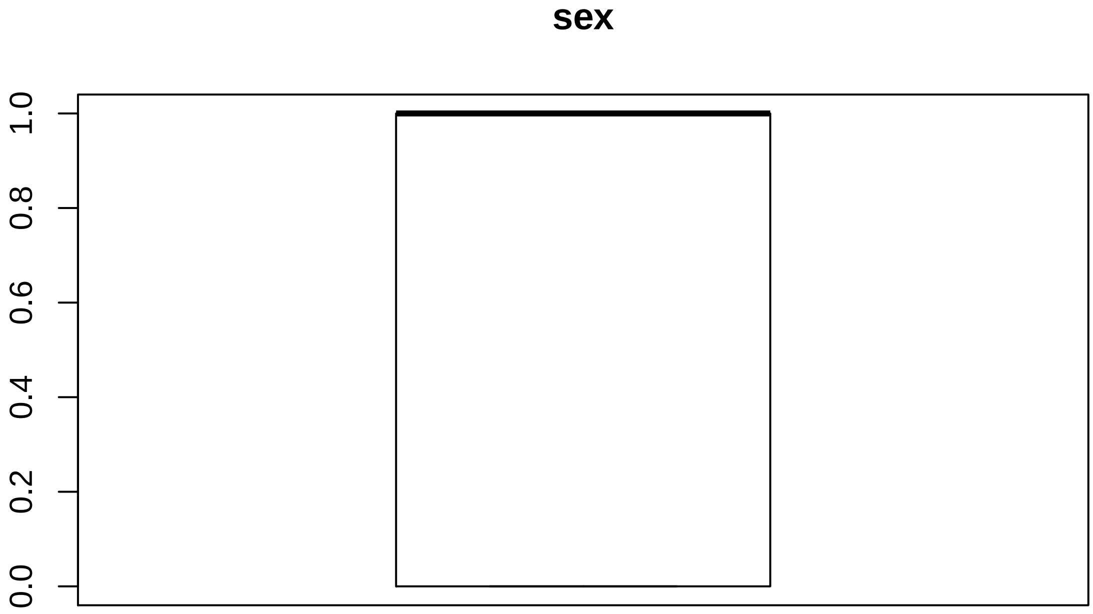
Examine the data with histograms.

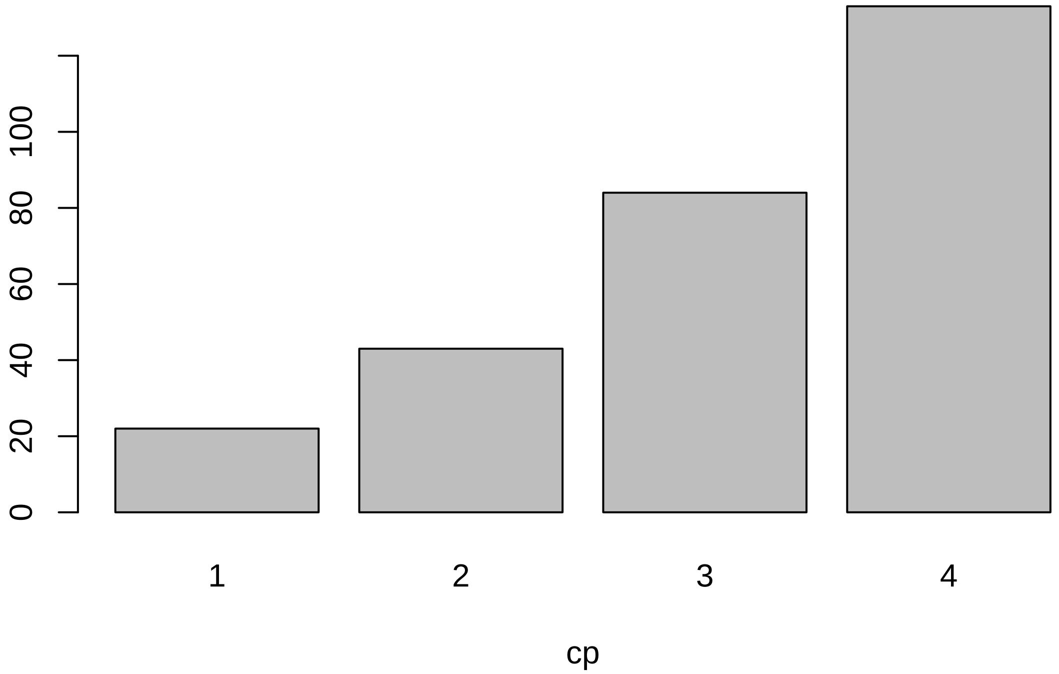
```r
# cholesterol and cigs appear to have large outliers
labels <- colnames(dt.nona)
for (i in seq(dim(dt)[2])) {
  col.data <- dt.nona[, get(labels[i])]
  nlevs <- nlevels(as.factor(col.data))
  if (nlevs <= 10) {
    barplot(table(col.data), main = NULL, xlab = labels[i])
    # axis(1, at=seq(nlevs), labels=levels(as.factor(col.data)))
  } else {
    hist(as.numeric(col.data), main = NULL, xlab = labels[i])
  }
  cat('\n\n')
  boxplot(col.data, main = labels[i])
  cat('\n\n')
}
```

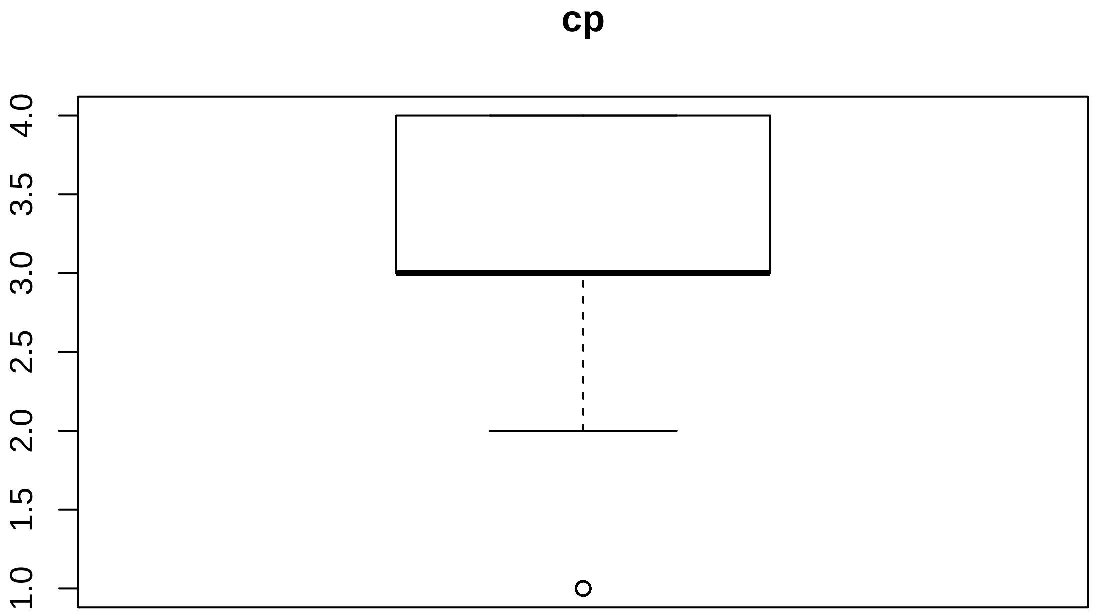


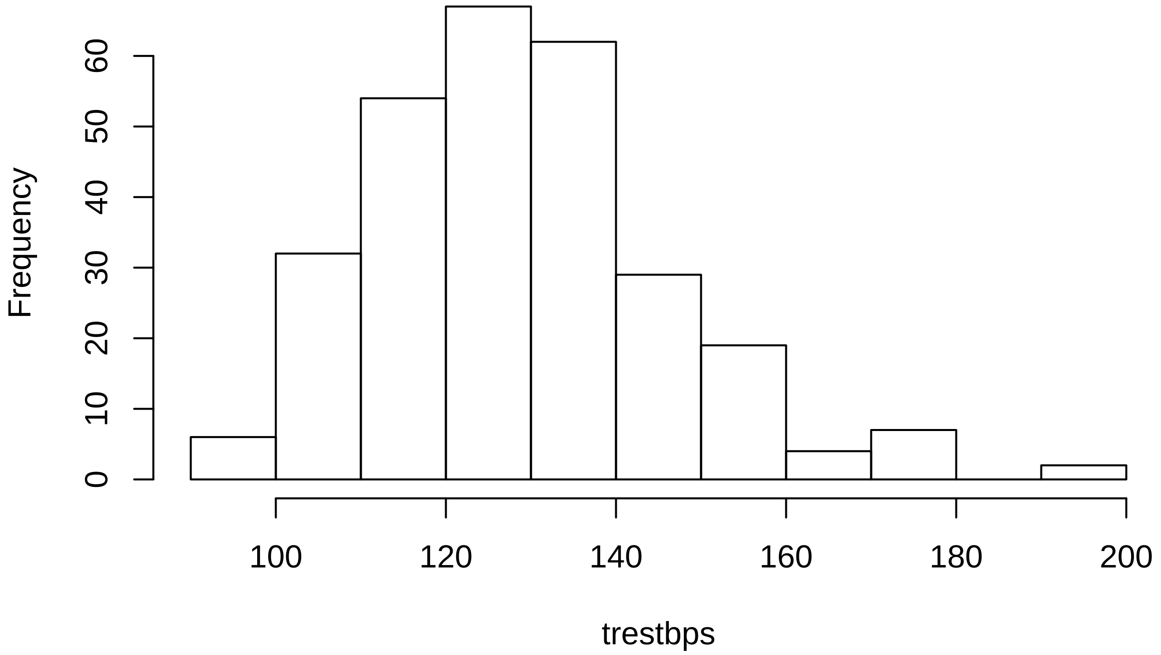


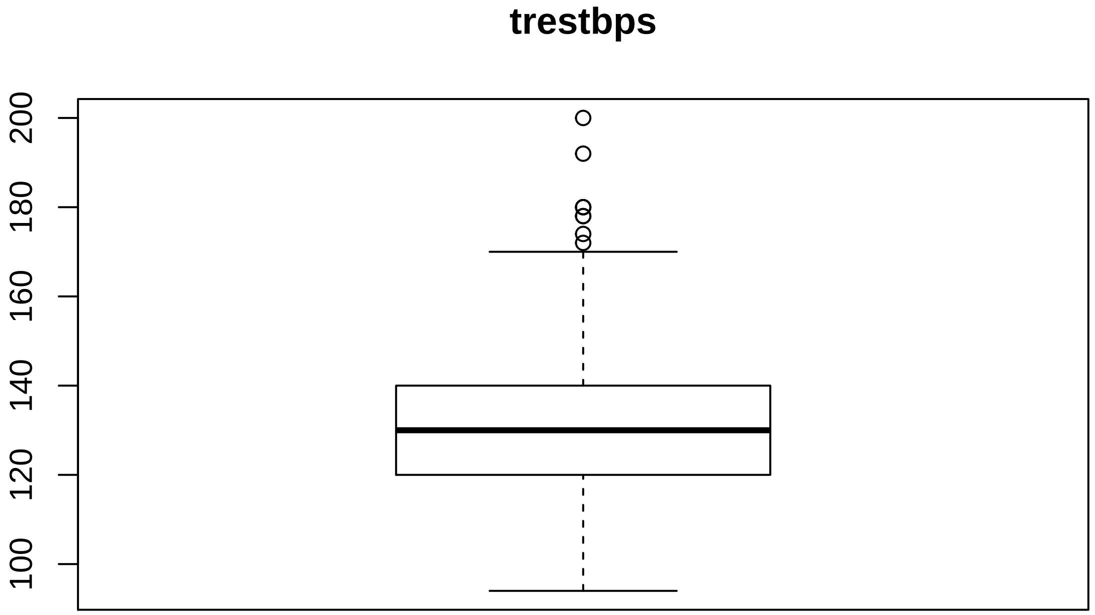

150
100
50
0
0
1
sex

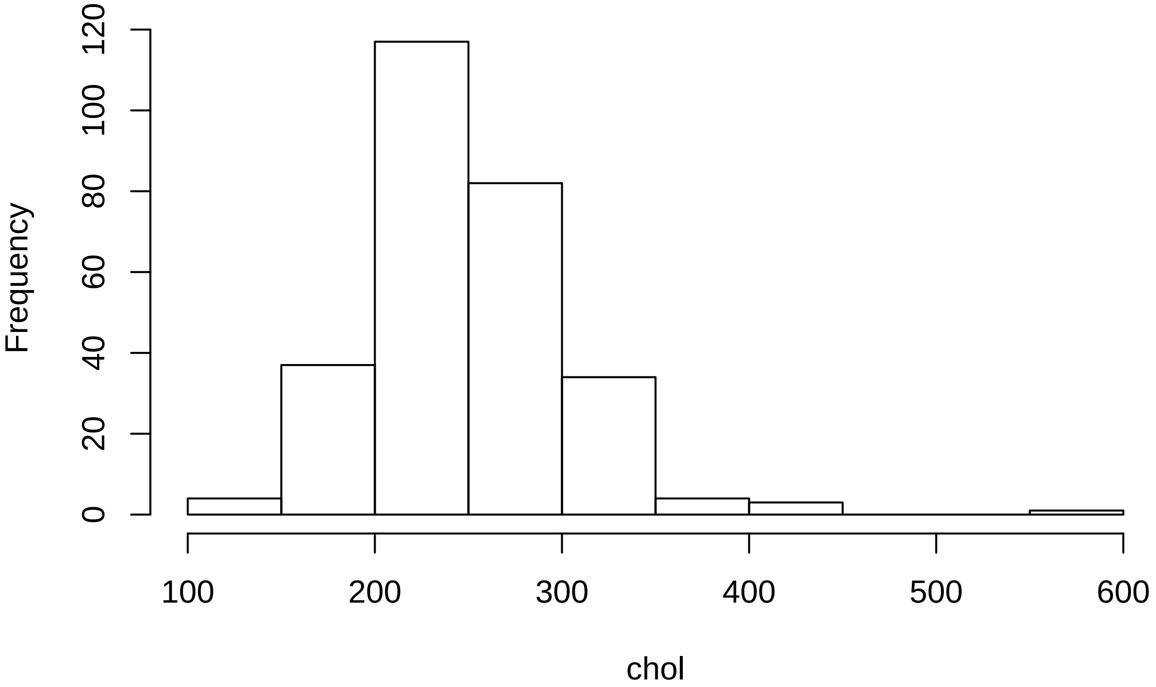

sex
1.0
0.8
0.6
0.4
0.2
0.0

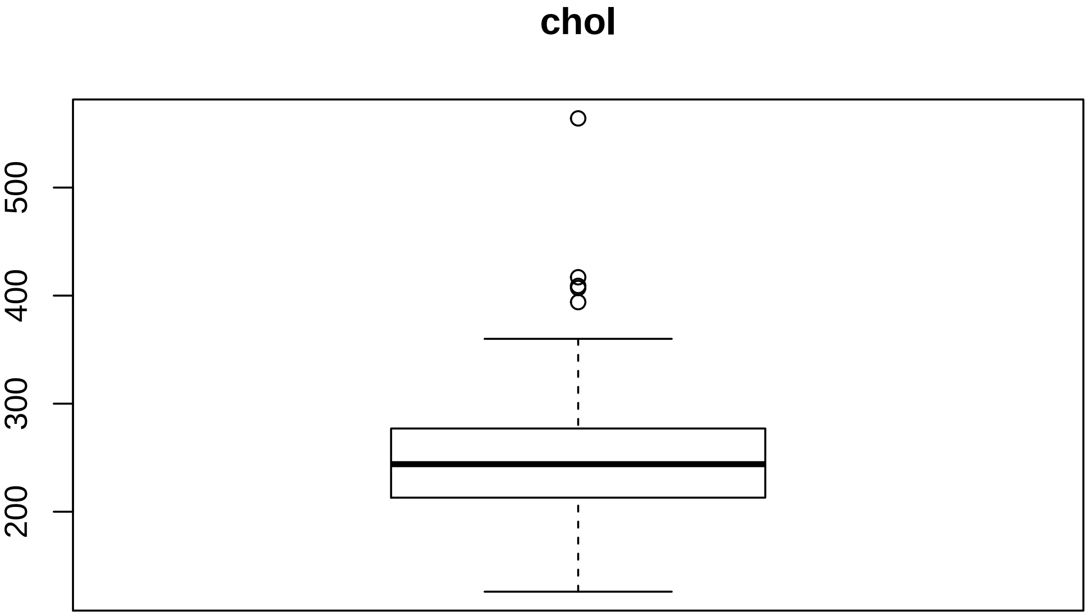

0
20
40
60
80
100
1
2
3
4
cp

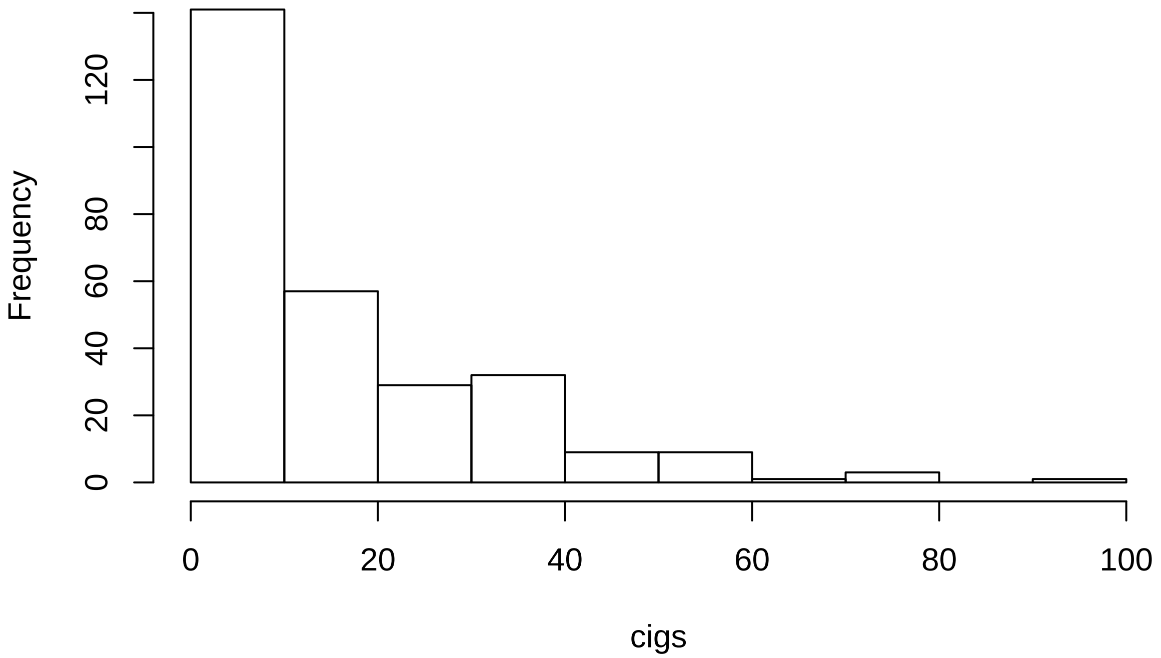

cp
1.0
1.5
2.0
2.5
3.0
3.5
4.0

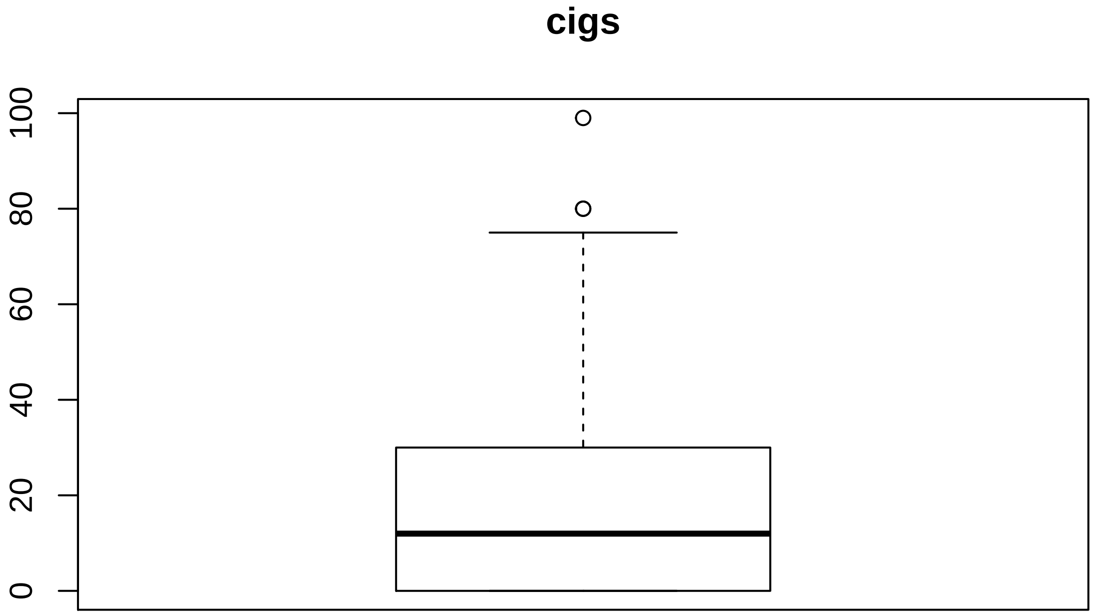

Frequency
0
10
20
30
40
50
60
100
120
140
160
180
200
trestbps

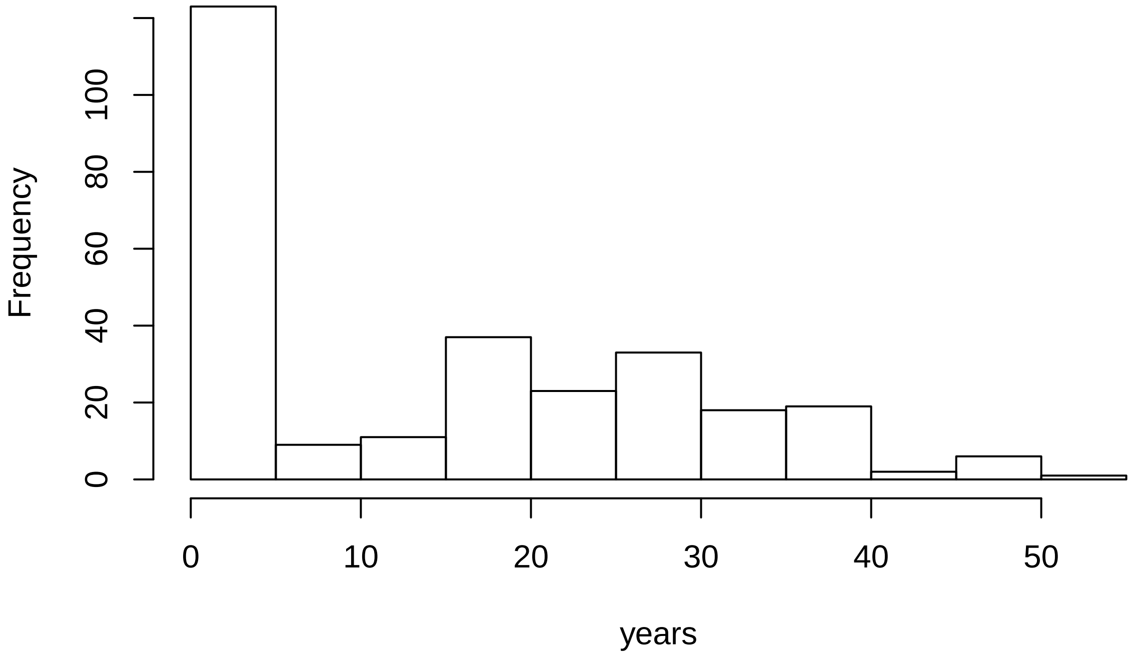

trestbps
100
120
140
160
180
200

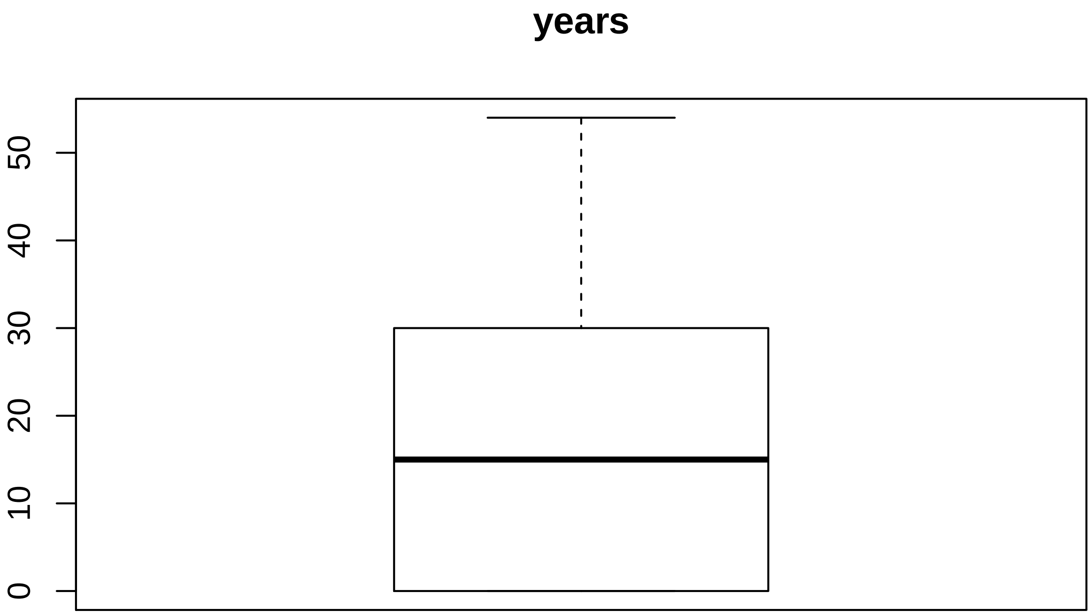

Frequency
0
20
40
60
80
100
120
100
200
300
400
500
600
chol

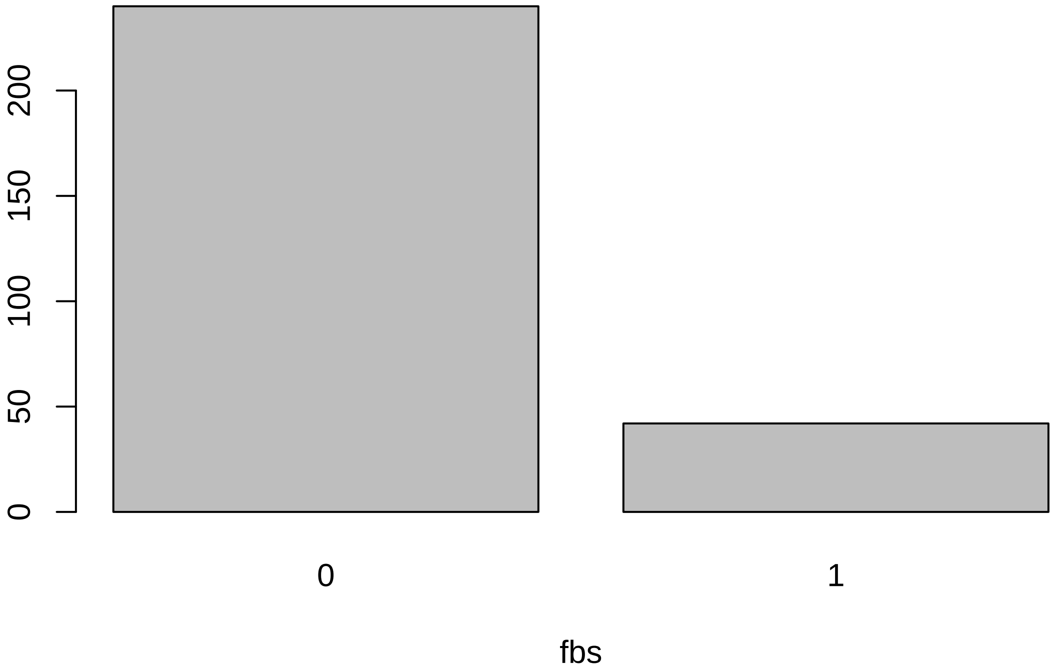

chol
200
300
400
500

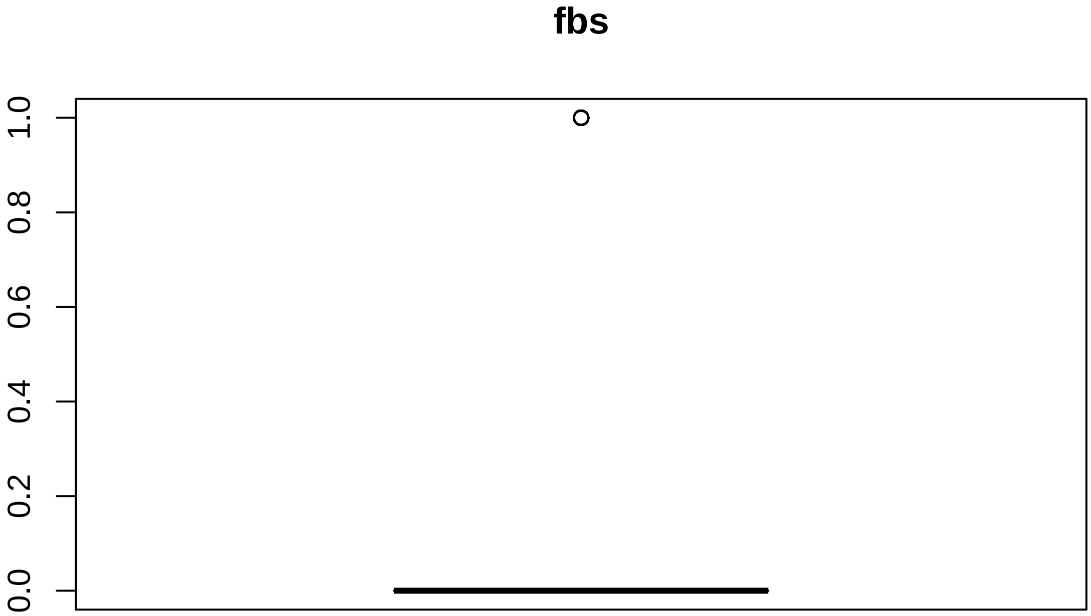

Frequency
0
20
40
60
80
120
0
20
40
60
80
100
cigs

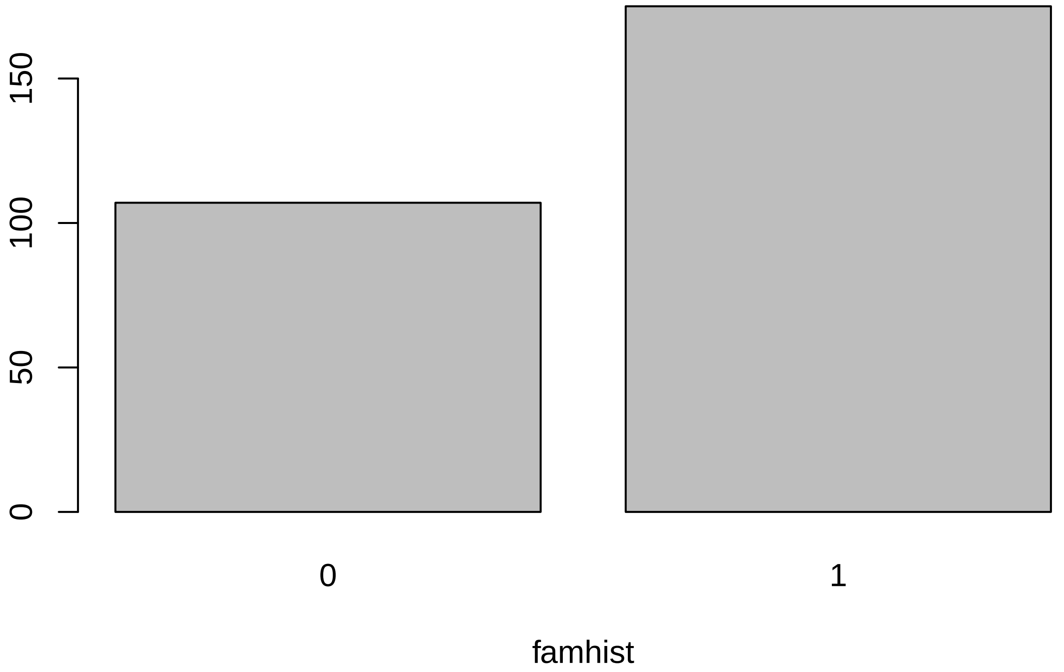

cigs
0
20
40
60
80
100

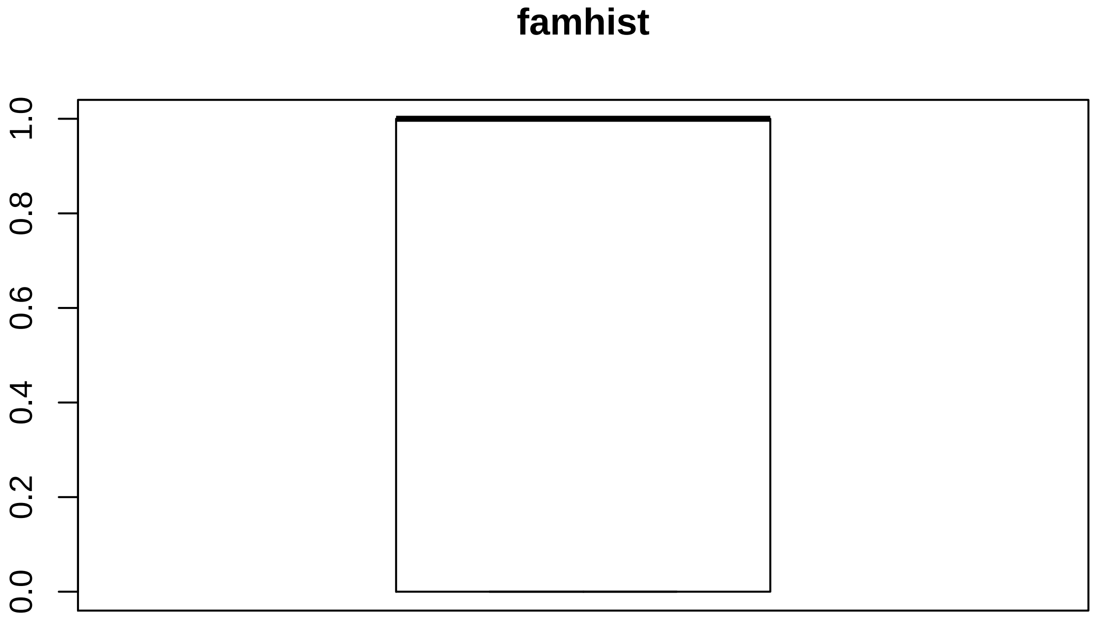

Frequency
0
20
40
60
80
100
0
10
20
30
40
50
years

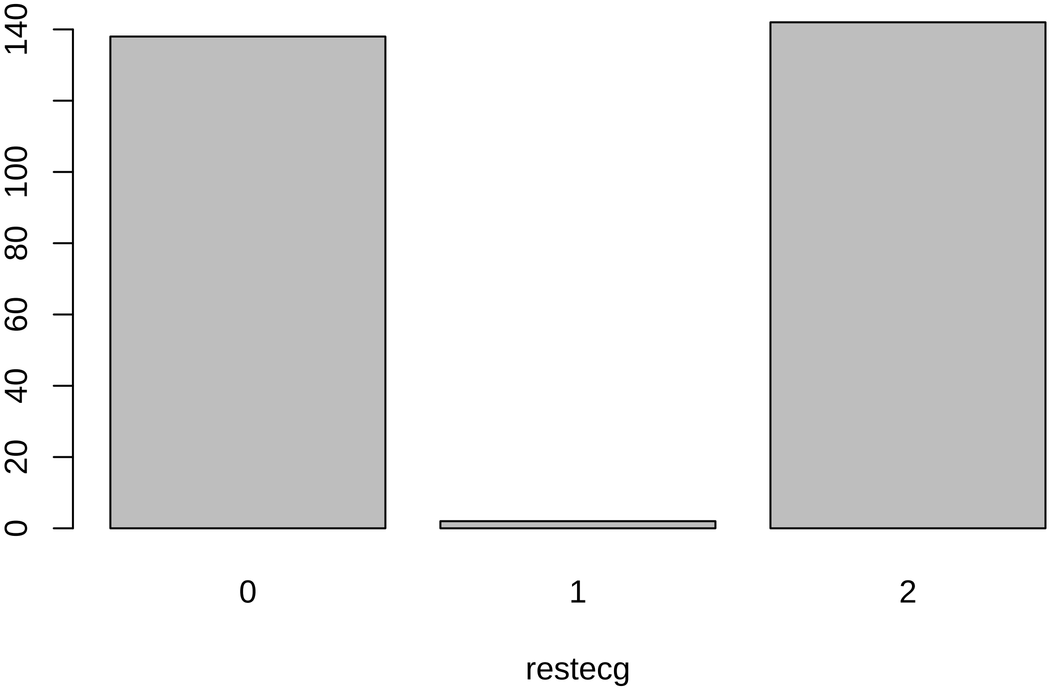

years
0
10
20
30
40
50

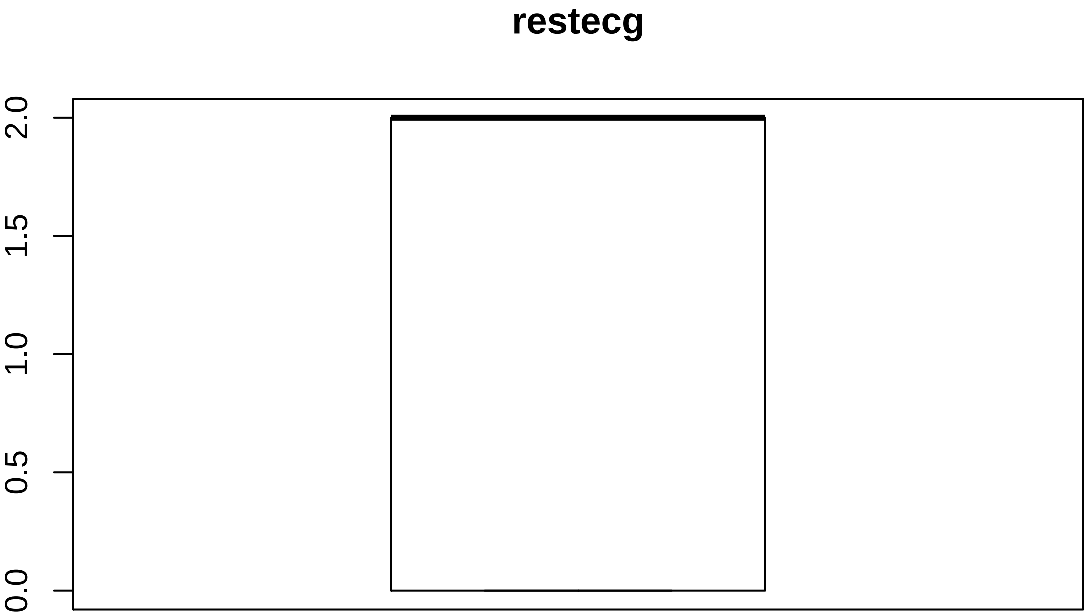

200
150
100
50
0
0
1
fbs

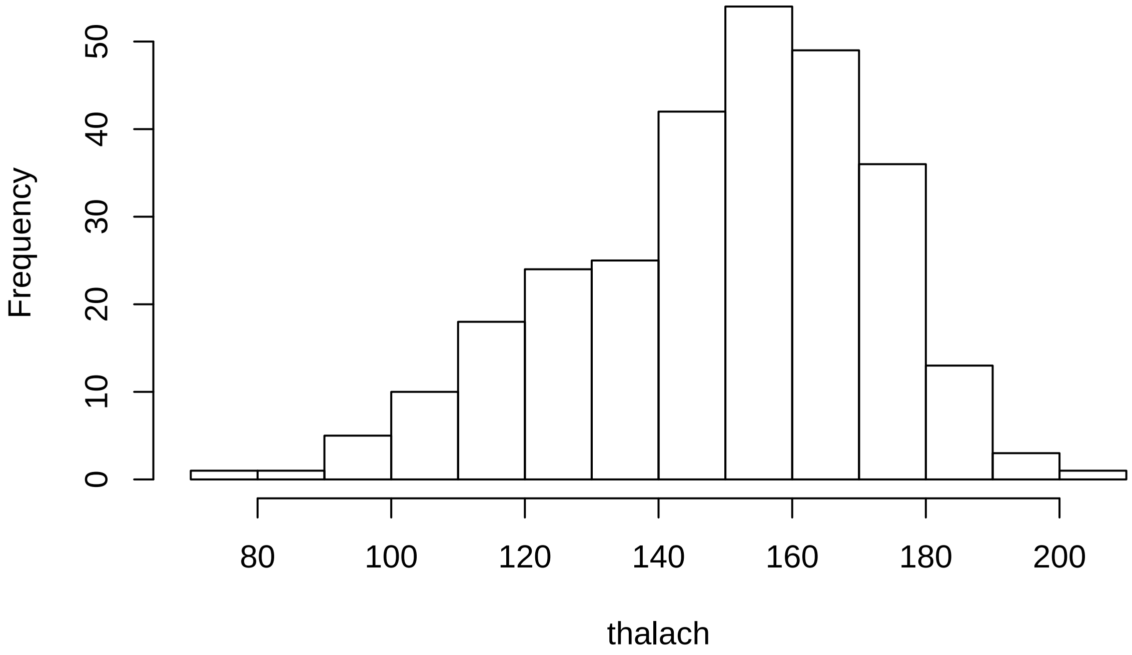

fbs
1.0
0.8
0.6
0.4
0.2
0.0

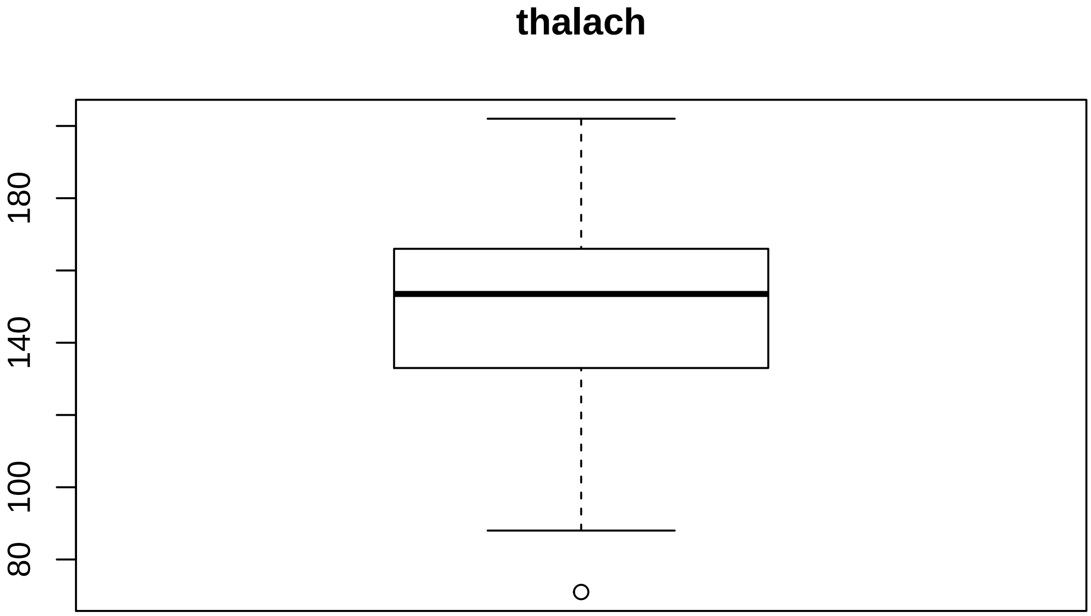

0
50
100
150
0
1
famhist

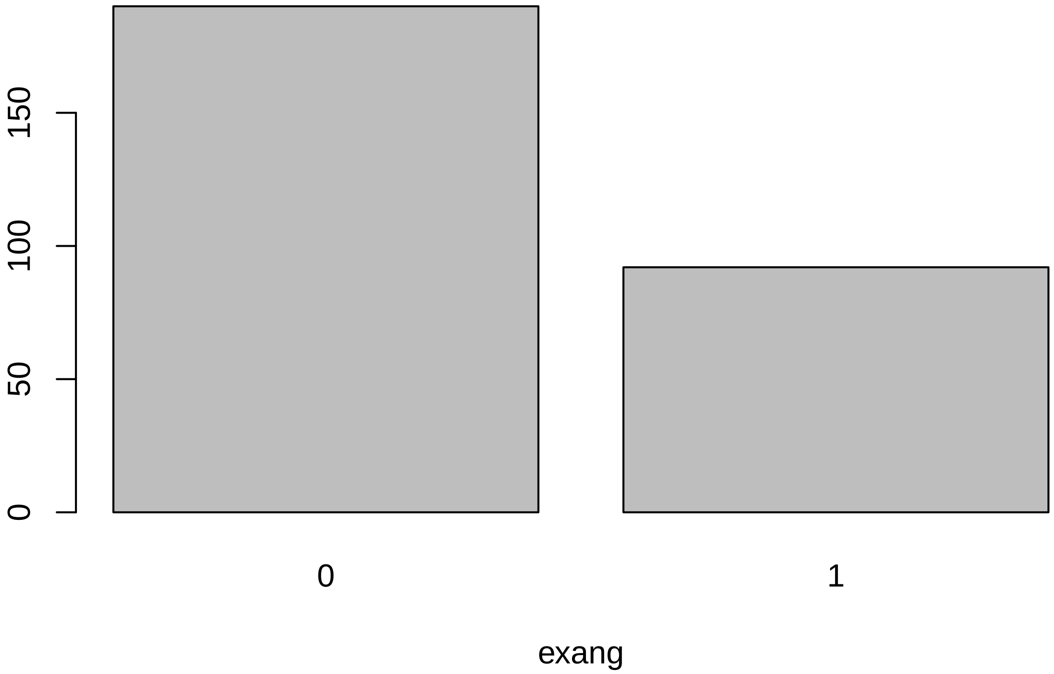

famhist
0.0
0.2
0.4
0.6
0.8
1.0

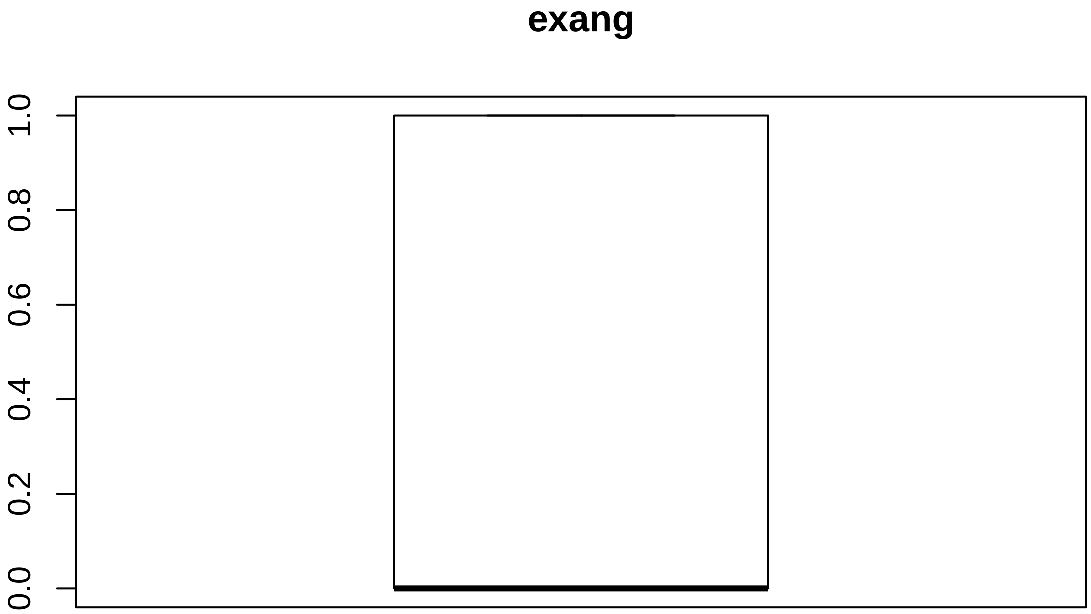

140
100
80
60
40
20
0
0
1
2
restecg

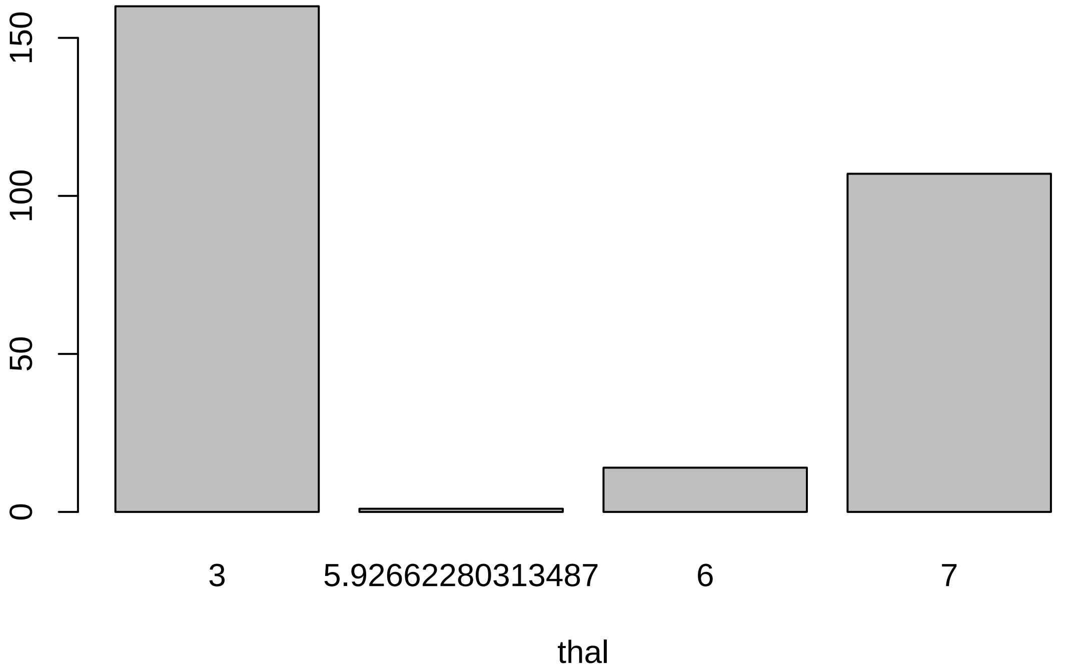

restecg
2.0
1.5
1.0
0.5
0.0

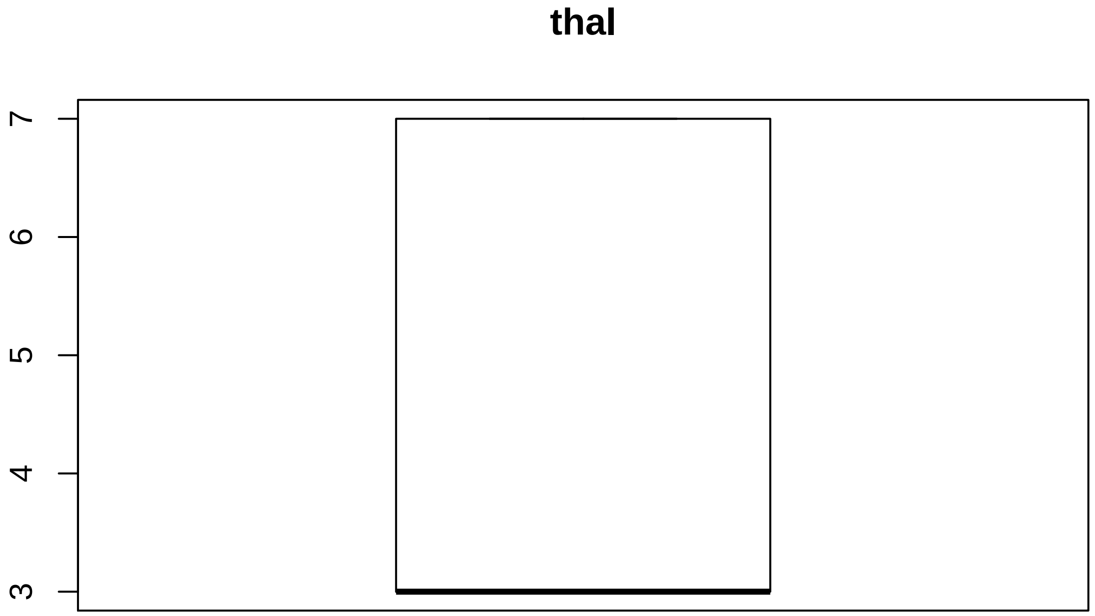

Frequency
0
10
20
30
40
50
80
100
120
140
160
180
200
thalach

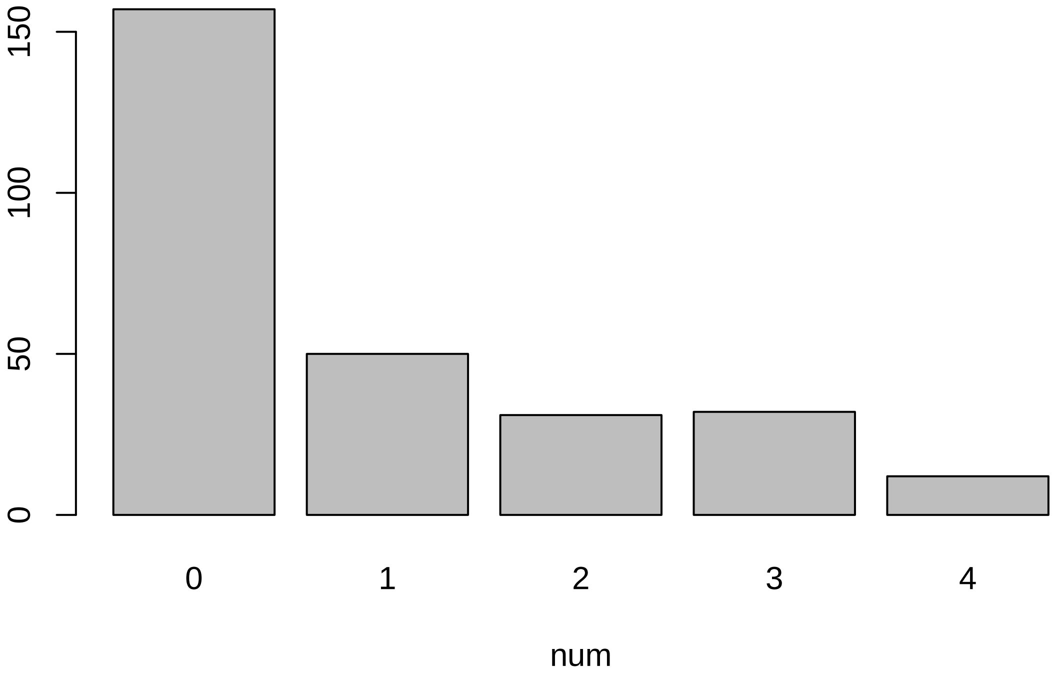

thalach
80
100
140
180

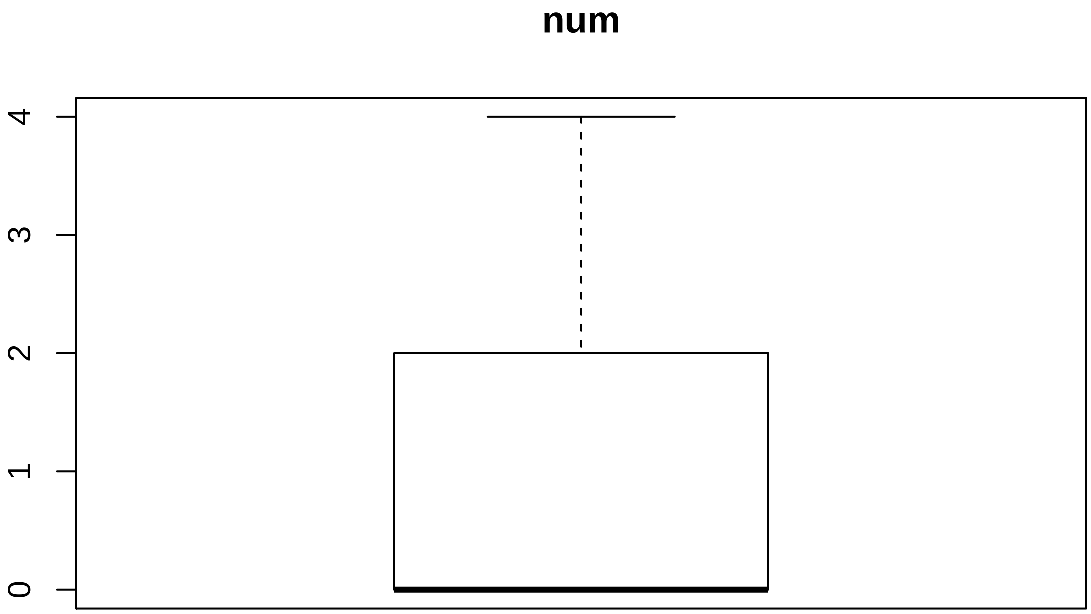

150
100
50
0
0
1
exang


exang
1.0
0.8
0.6
0.4
0.2
0.0


150
100
50
0
3
5.92662280313487
6
7
thal


thal
7
6
5
4
3

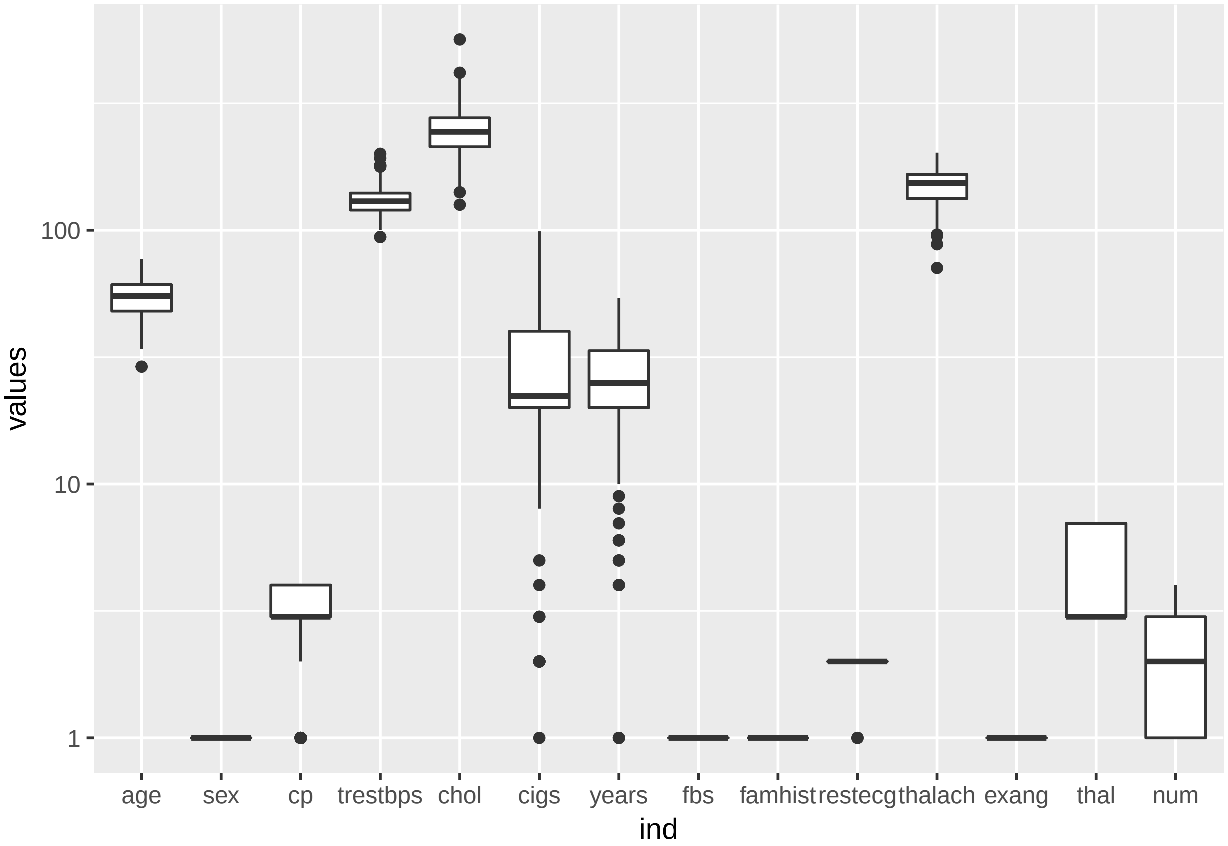




Looking at a log-scale boxplot of all variables, we can see some low outliers on trestbps, chol, and thalach.

```r
ggplot(stack(dt.nona), aes(x = ind, y = values)) +
  geom_boxplot() + scale_y_continuous(trans = 'log10')
```

```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1153 rows containing non-finite values (stat_boxplot).
```

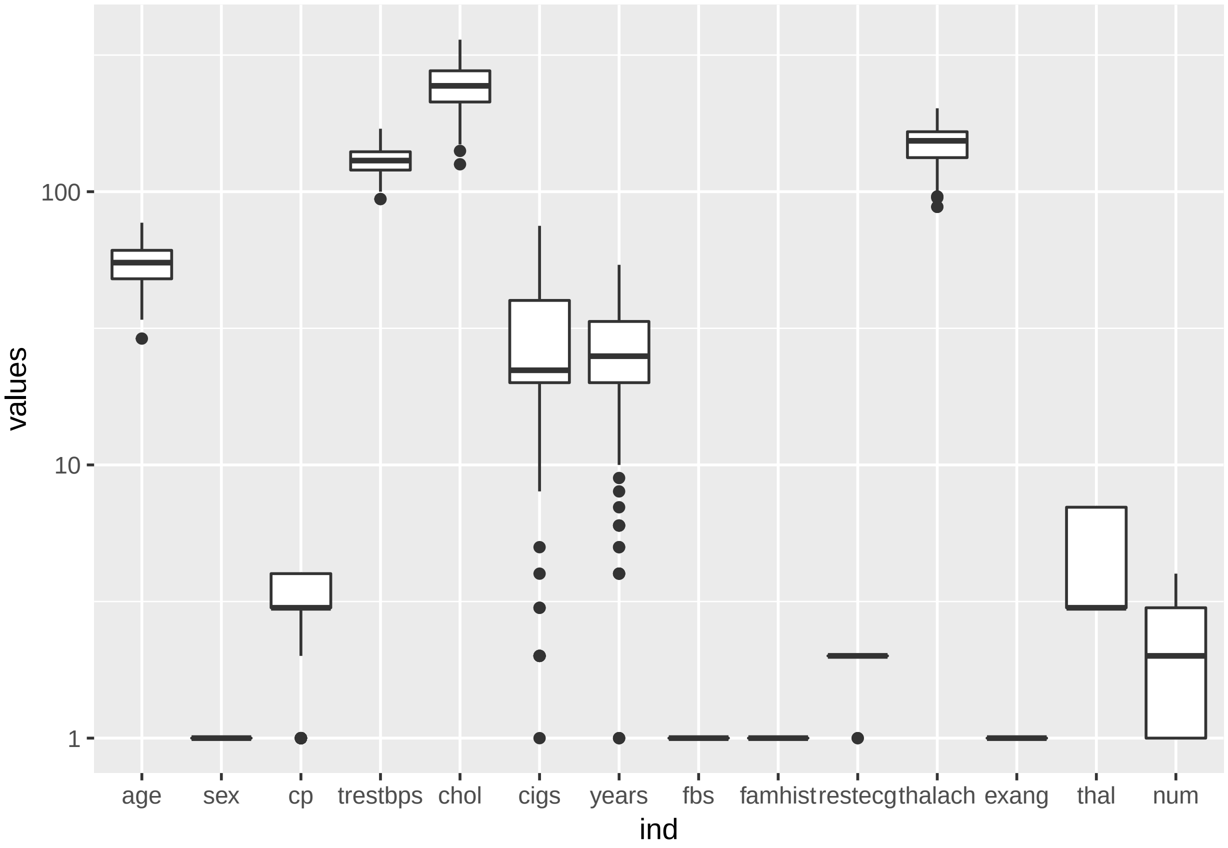


Outliers seem to be one low in thalach, a few high on cigs, and few high/low on cholestorol and trestbps. Looking carefully, it seems there are some lower cigs outliers we don't want to remove, because these are probably accurate data. We removed the upper ones because 100 cigarettes a day seems like an error in the data. The assumption is the other outliers may be expreiment/data errors.

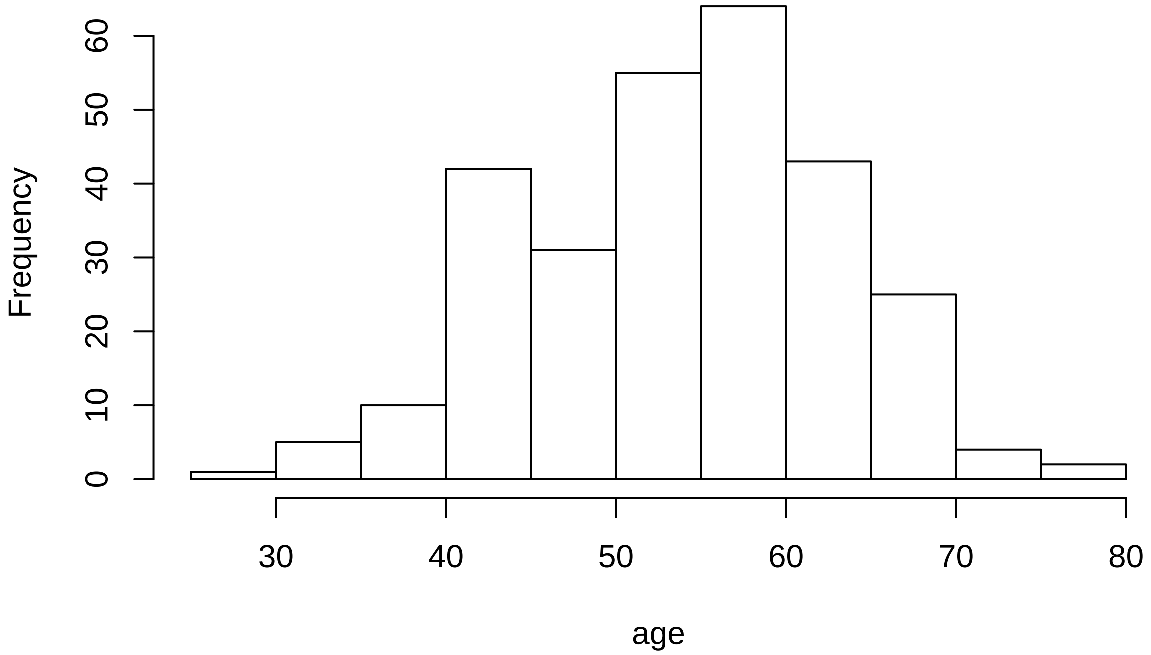
```
upper_outlier_cols <- c('trestbps', 'chol', 'cigs')
lower_outlier_cols <- c('thalach', 'chol', 'trestbps')
remove_outliers_up <- function(x) clamp(x, upper = boxplot.stats(x)$stats[5])
remove_outliers_low <- function(x) clamp(x, lower = boxplot.stats(x)$stats[1])

dt.nona[, upper_outlier_cols] <- dt.nona[, lapply(.SD, FUN = remove_outliers_up), .SDcols = upper_outli
dt.nona[, lower_outlier_cols] <- dt.nona[, lapply(.SD, FUN = remove_outliers_low), .SDcols = lower_outl

# you may have also noticed there is one weird 5.92... number in thal, this should be rounded to 6
dt.nona[, thal:=round(thal)]
# save data for later use
fn <- '~/Dropbox/MSDS/MSDS680_ncg_S8W1_18/week1/heart.disease.data.clean.csv'
fwrite(dt.nona, fn)
```

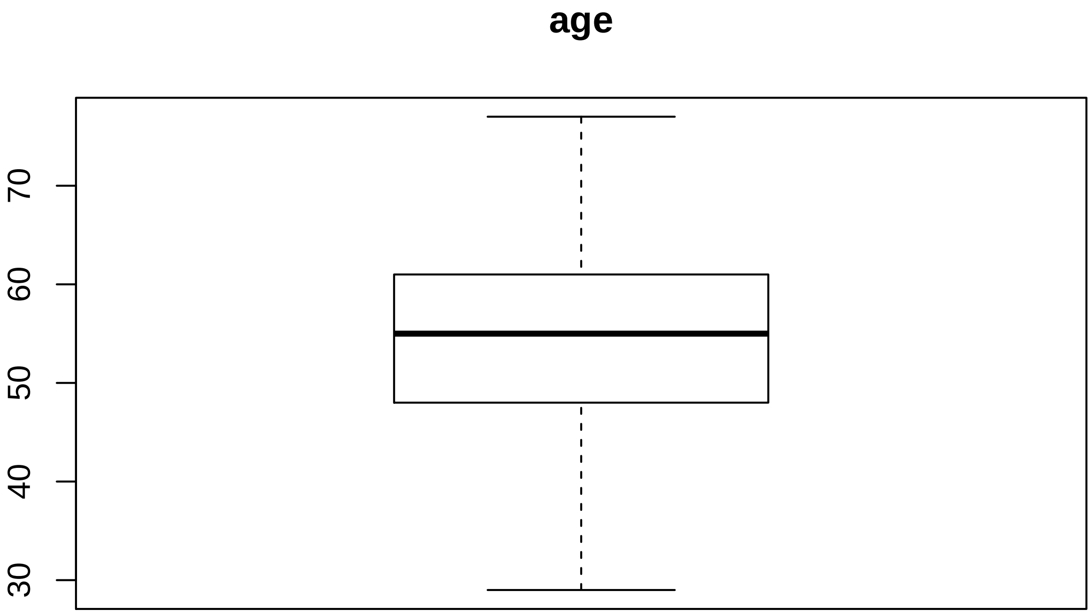
```
ggplot(stack(dt.nona), aes(x = ind, y = values)) +
  geom_boxplot() + scale_y_continuous(trans = 'log10')
```

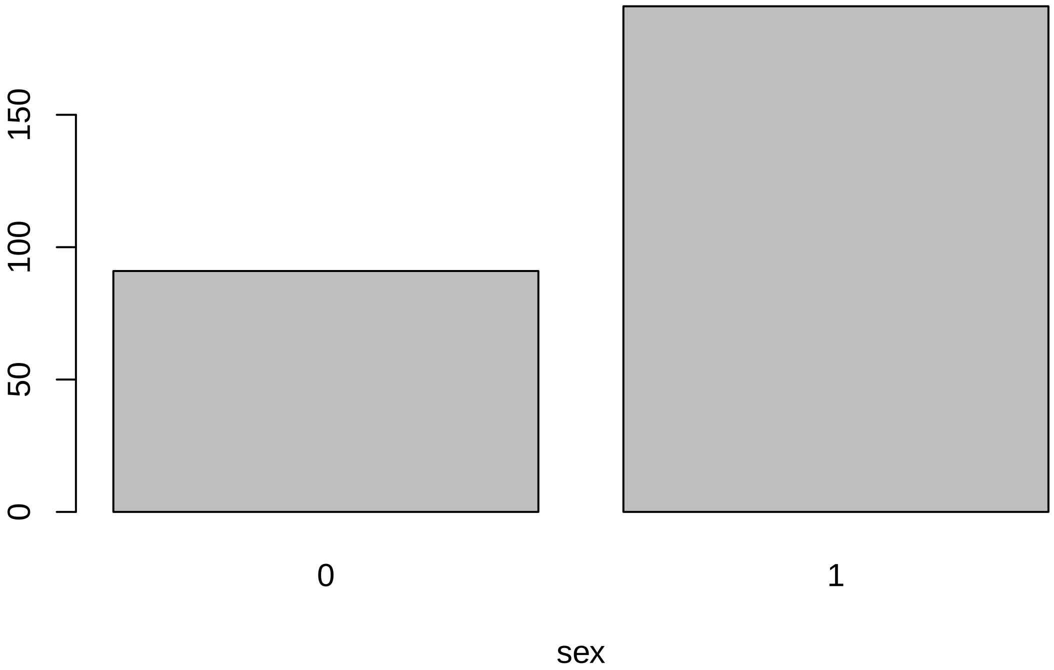
```
## Warning: Transformation introduced infinite values in continuous y-axis

## Warning: Removed 1153 rows containing non-finite values (stat_boxplot).
```

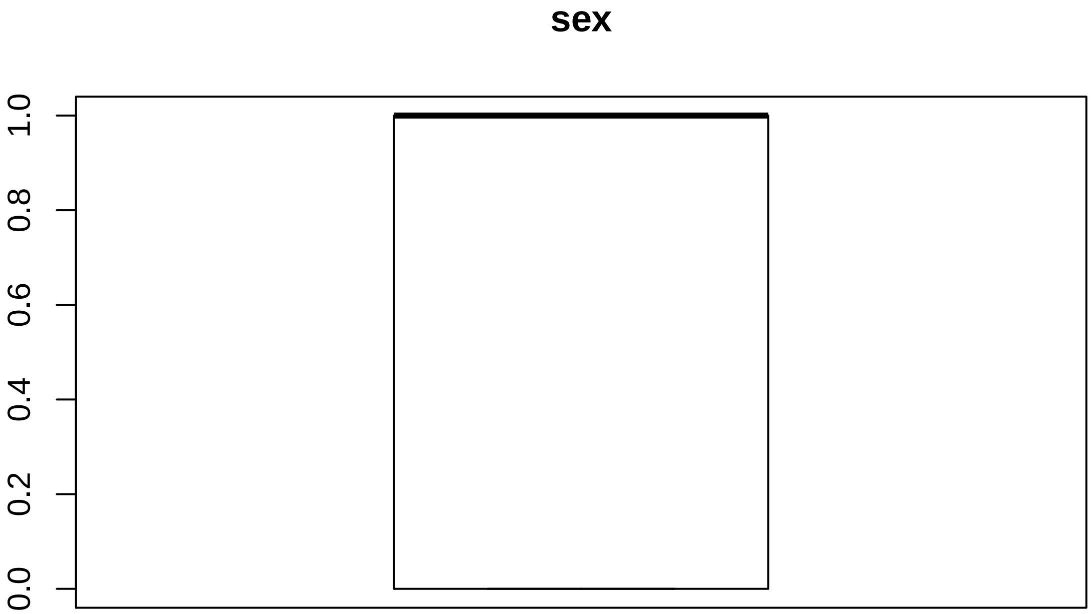


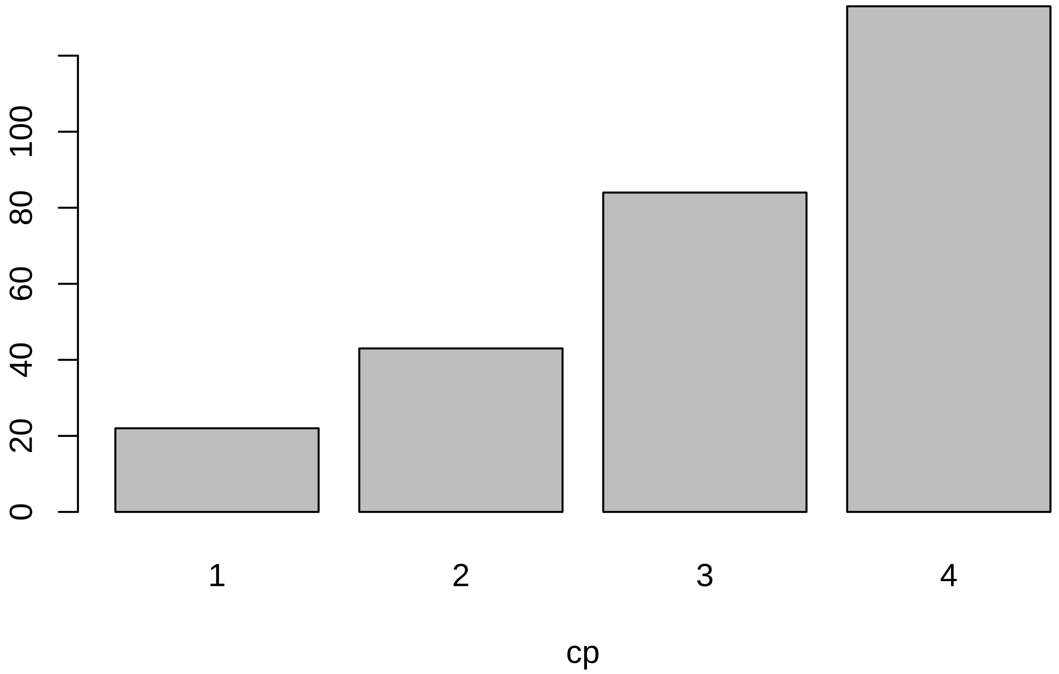
```
for (i in seq(dim(dt)[2])) {
  col.data <- dt.nona[, get(labels[i])]
  nlevs <- nlevels(as.factor(col.data))
  if (nlevs <= 10) {
    barplot(table(col.data), main = NULL, xlab = labels[i])
    # axis(1, at=seq(nlevs), labels=levels(as.factor(col.data)))
  } else {
    hist(as.numeric(col.data), main = NULL, xlab = labels[i])
  }
  cat('\n\n')
  boxplot(col.data, main = labels[i])
  cat('\n\n')
}
```

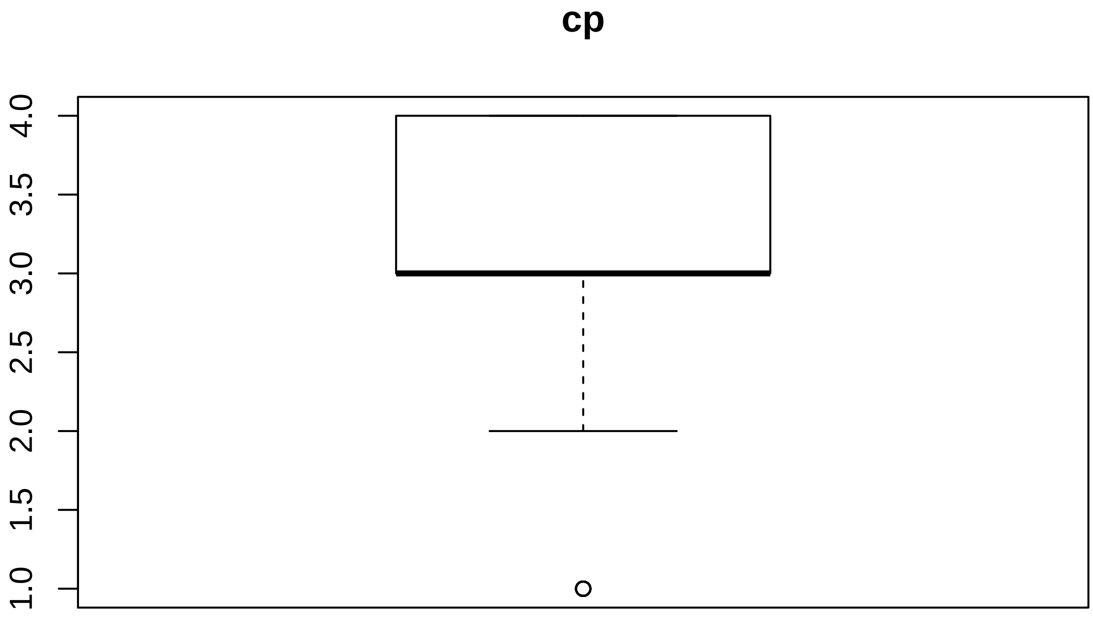

Frequency
0
10
20
30
40
50
60
30
40
50
60
70
80
age

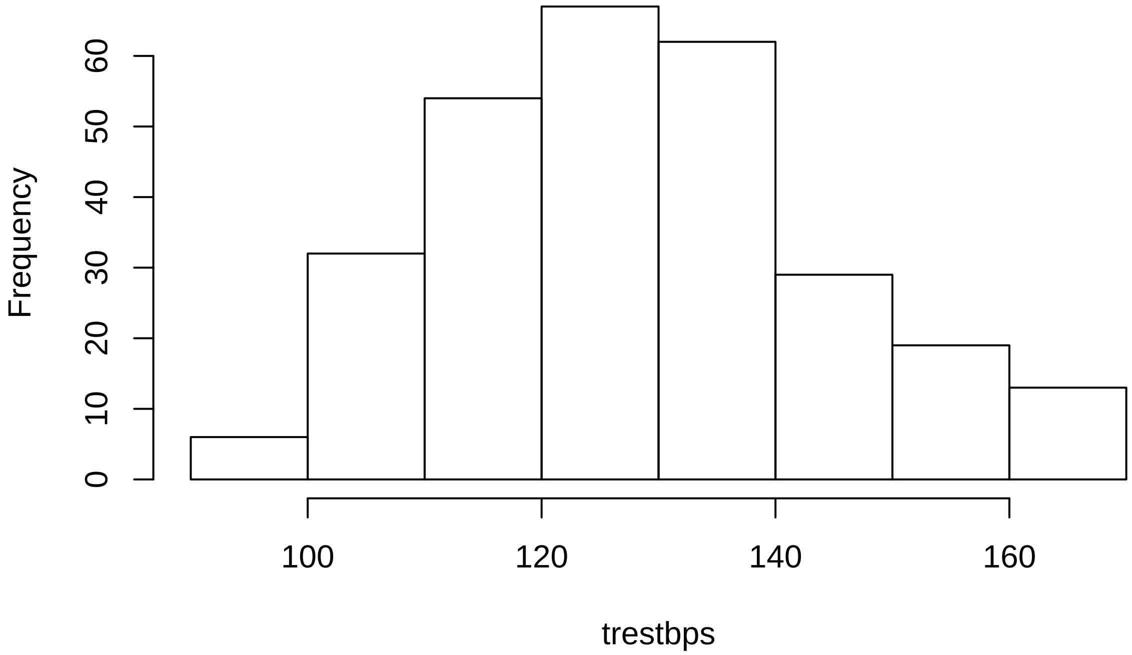

age
30
40
50
60
70

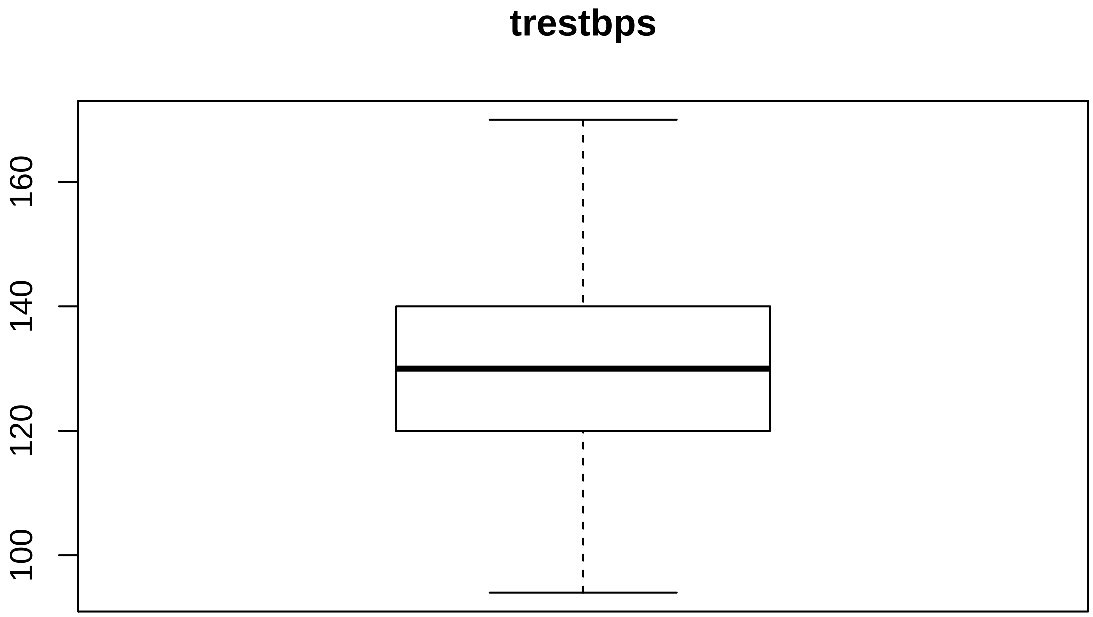

150
100
50
0
0
1
sex

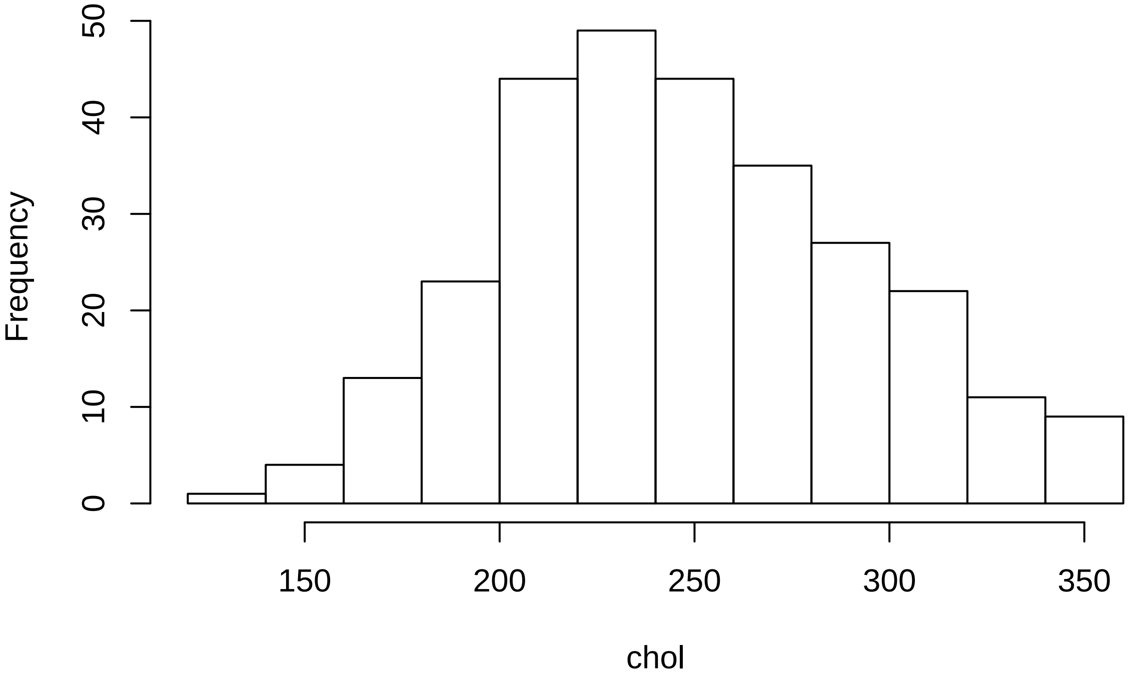

sex
1.0
0.8
0.6
0.4
0.2
0.0

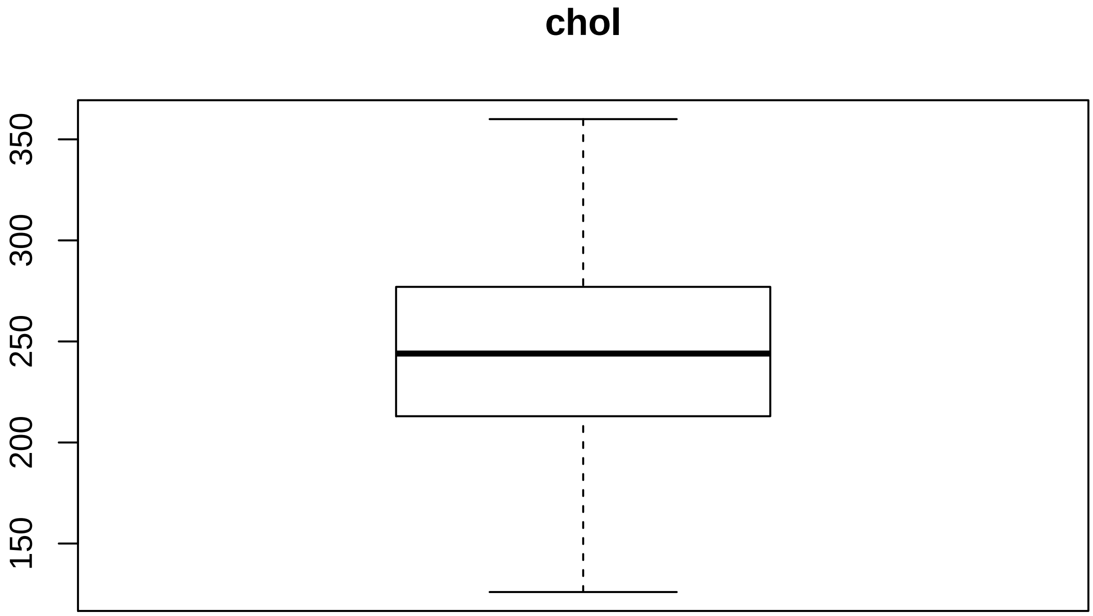

0
20
40
60
80
100
1
2
3
4
cp

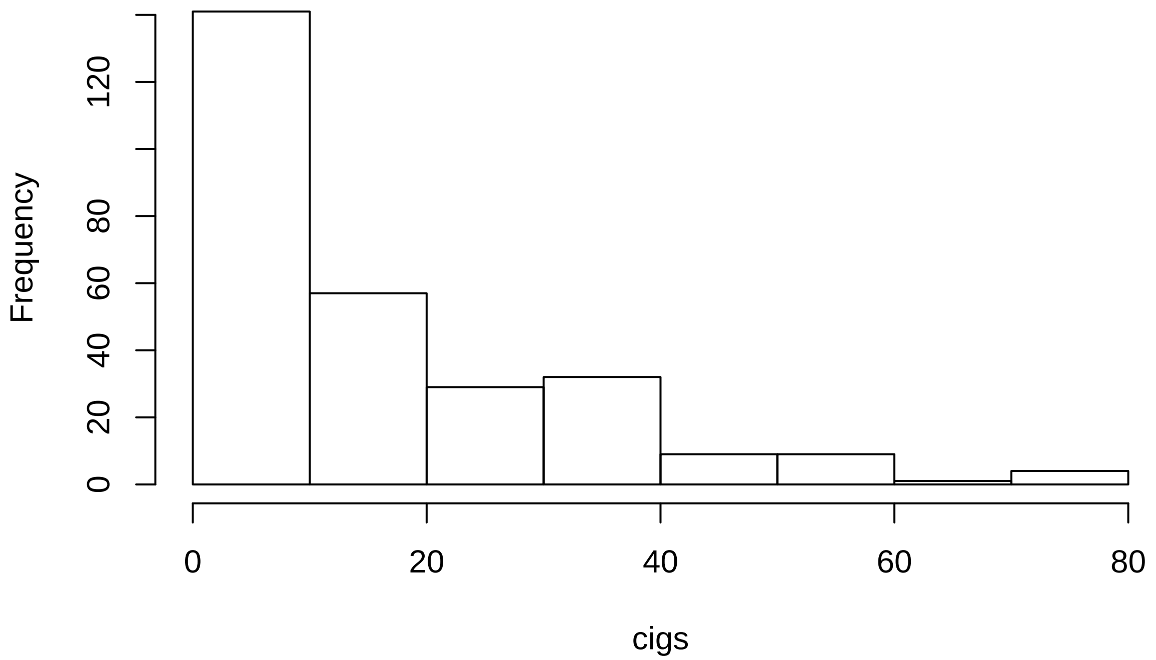

cp
1.0
1.5
2.0
2.5
3.0
3.5
4.0

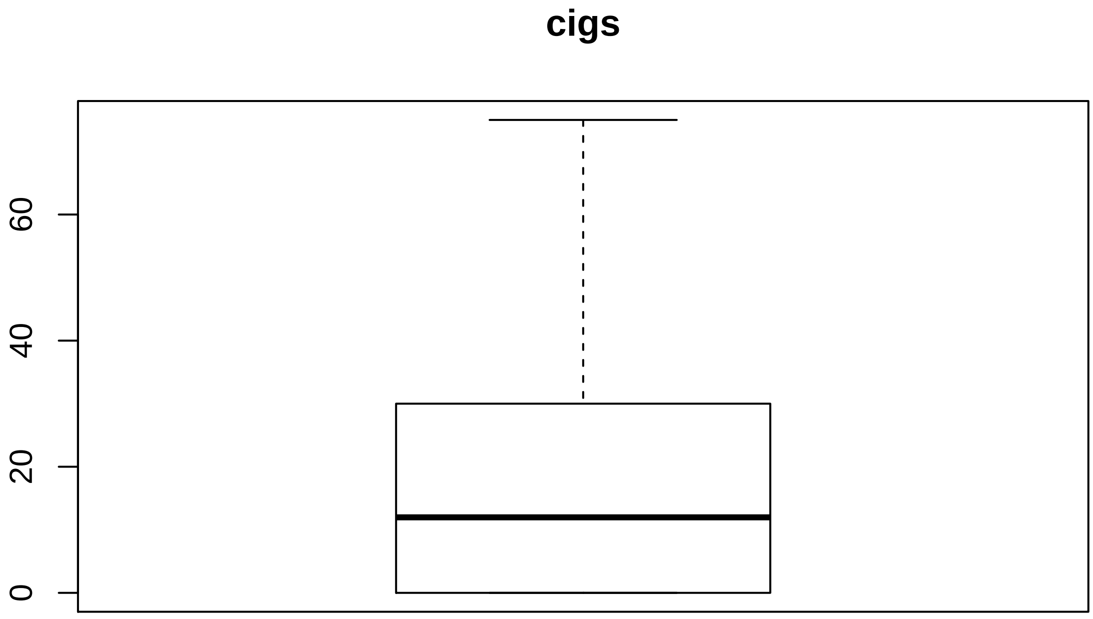

Frequency
0
10
20
30
40
50
60
100
120
140
160
trestbps

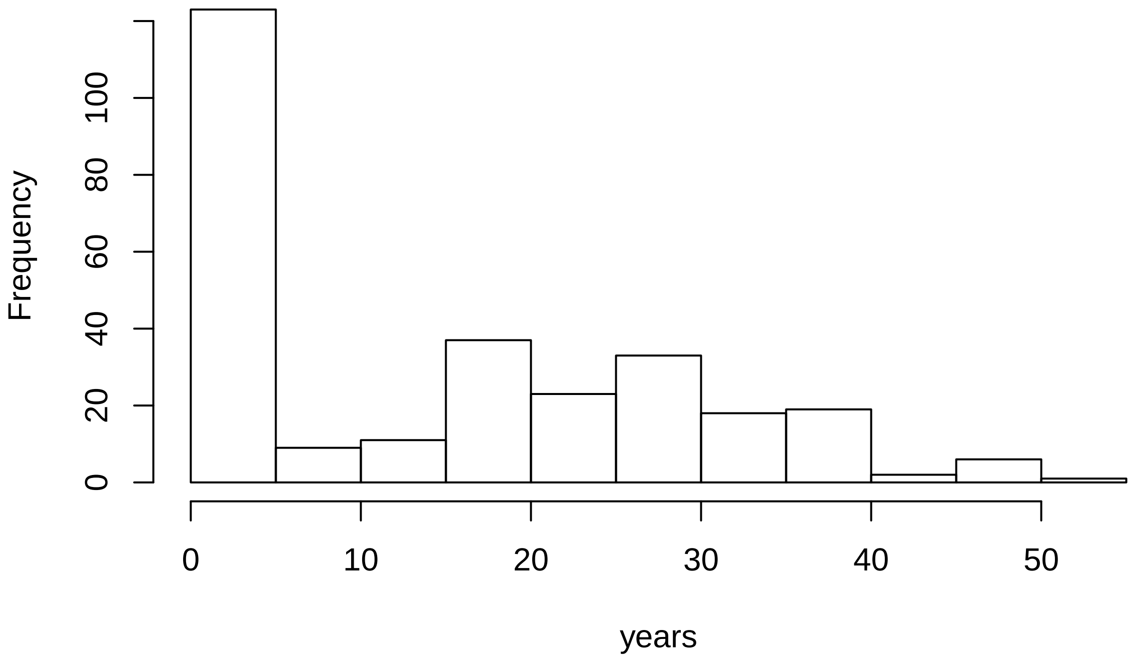

trestbps
100
120
140
160

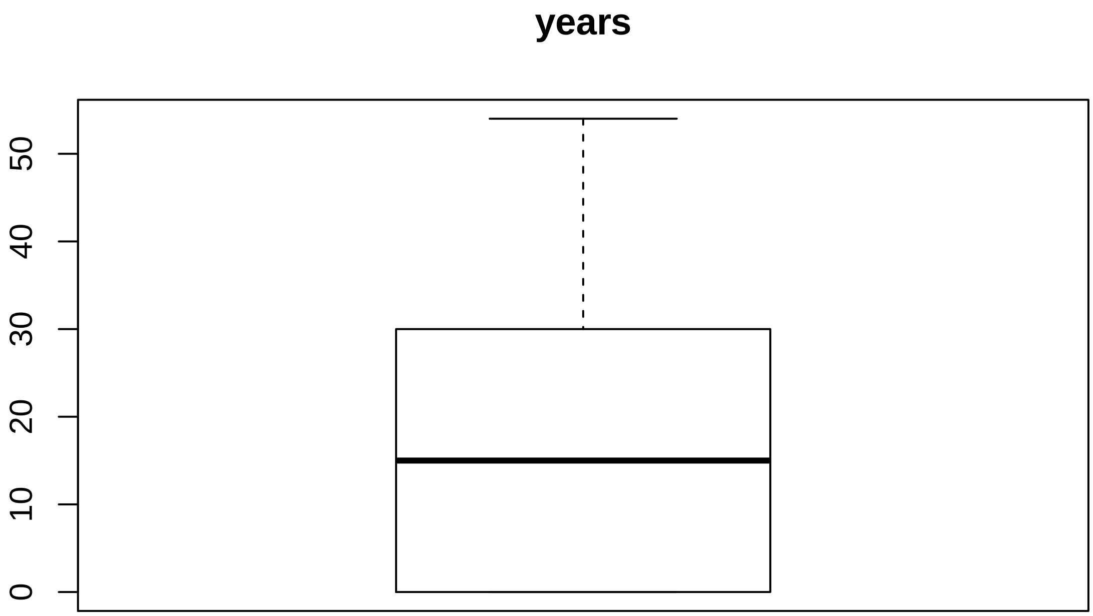

Frequency
0
10
20
30
40
50
150
200
250
300
350
chol

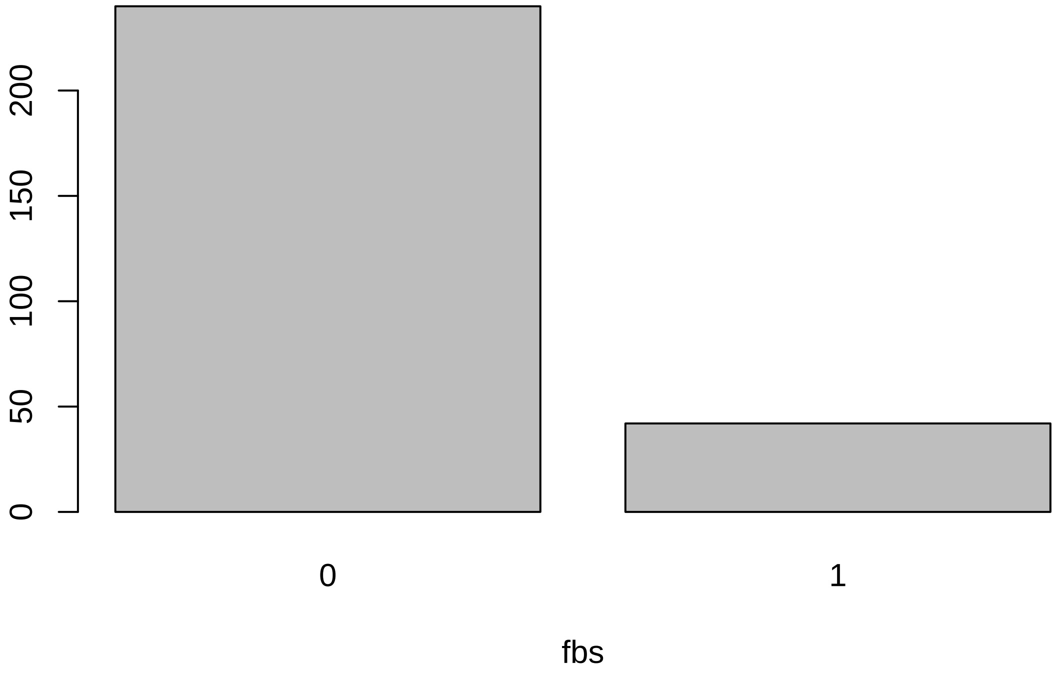

chol
150
200
250
300
350

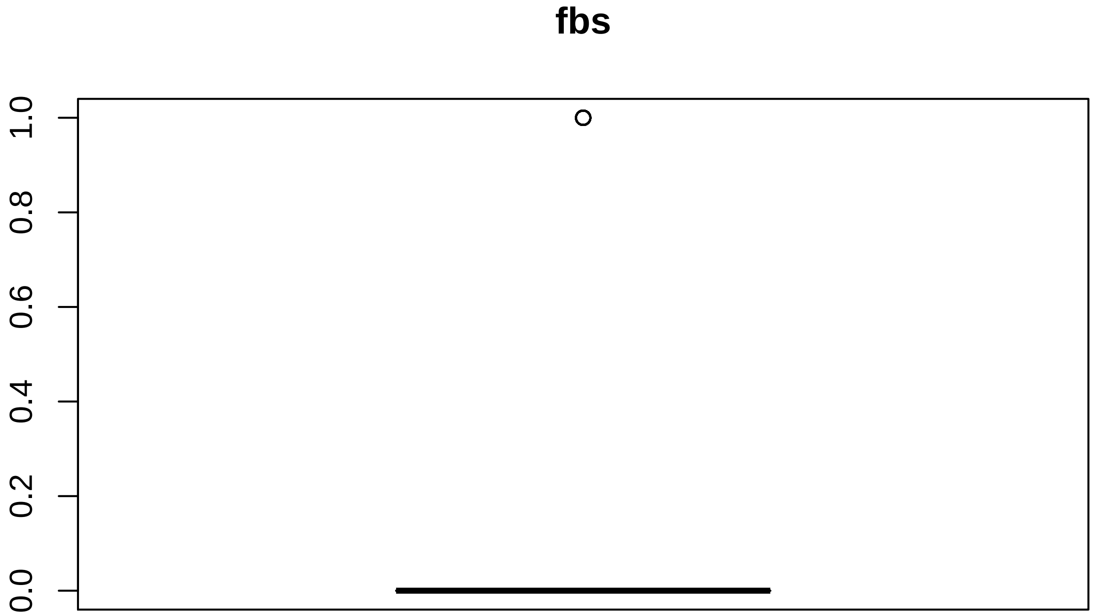

Frequency
0
20
40
60
80
120
0
20
40
60
80
cigs

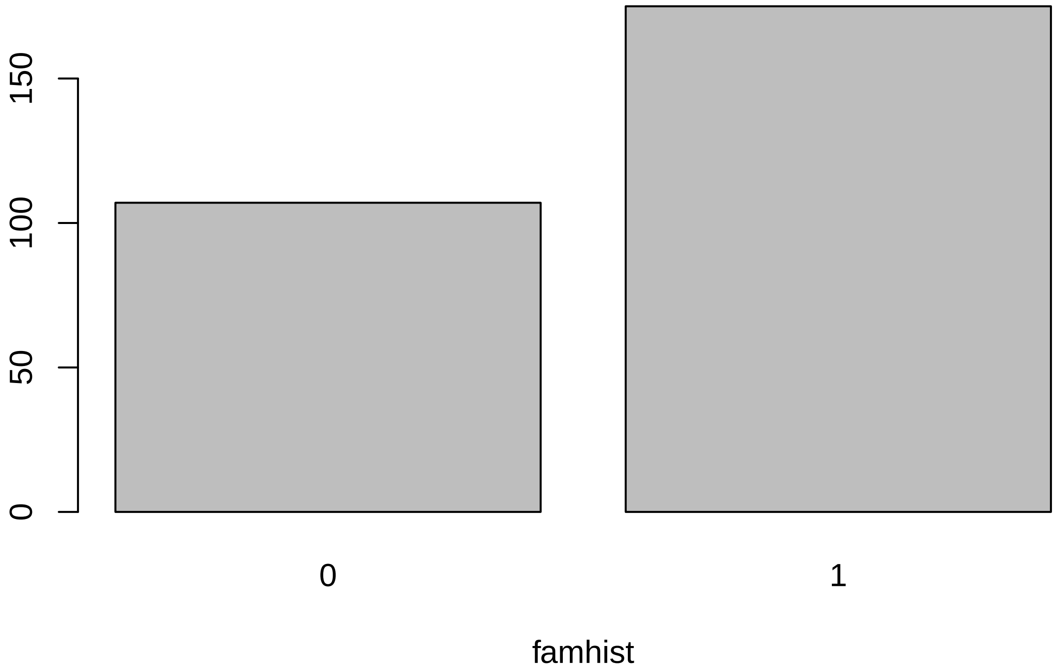

cigs
0
20
40
60

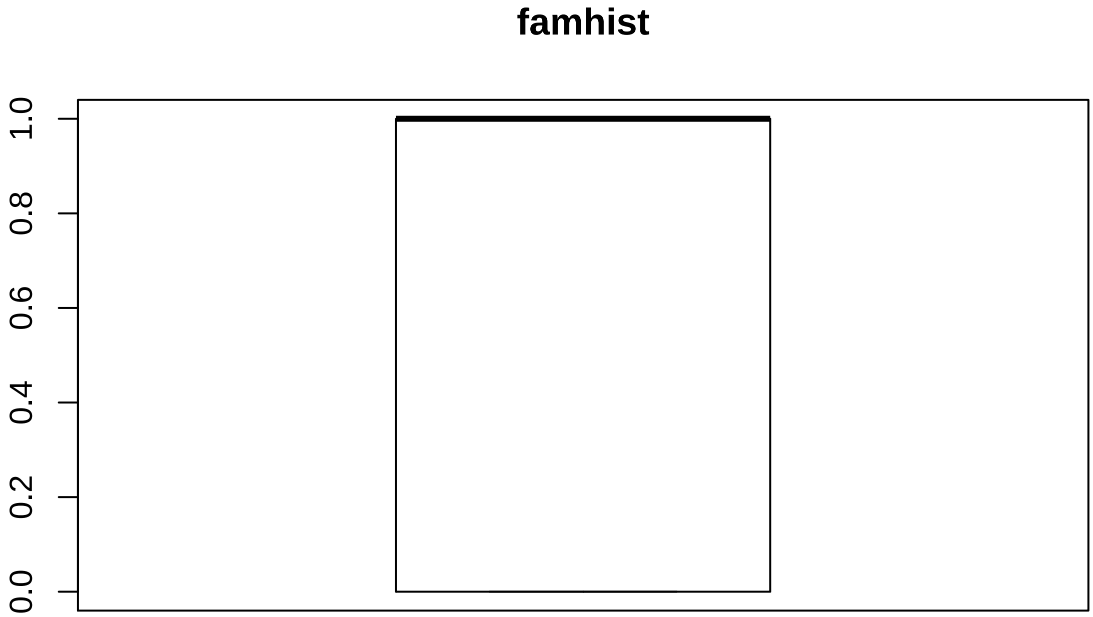

Frequency
0
20
40
60
80
100
0
10
20
30
40
50
years

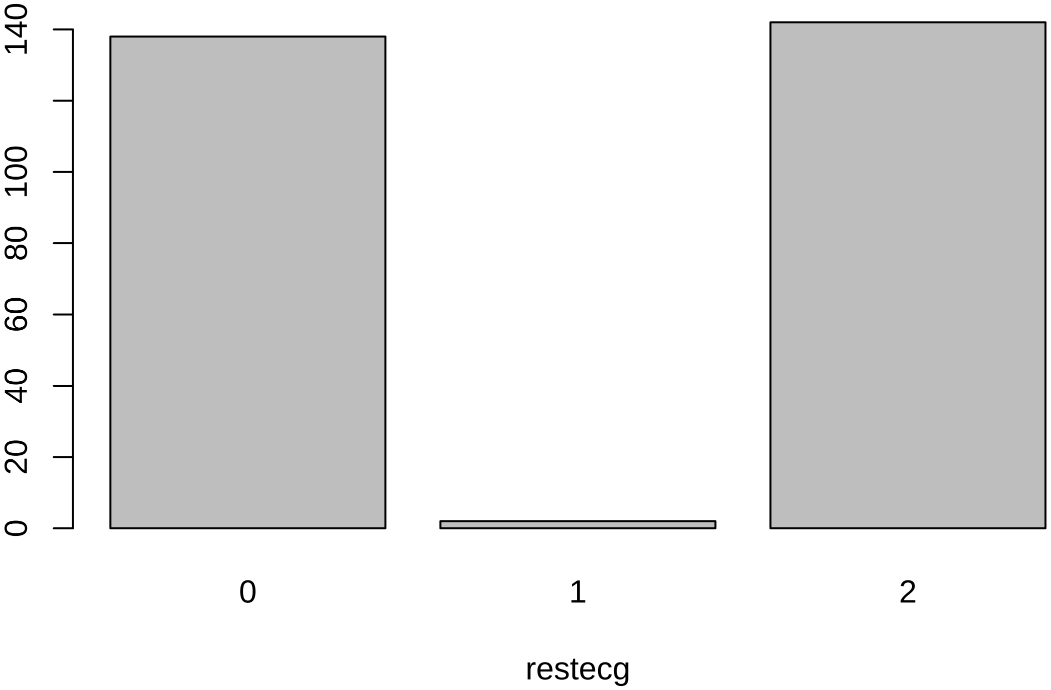

years
0
10
20
30
40
50

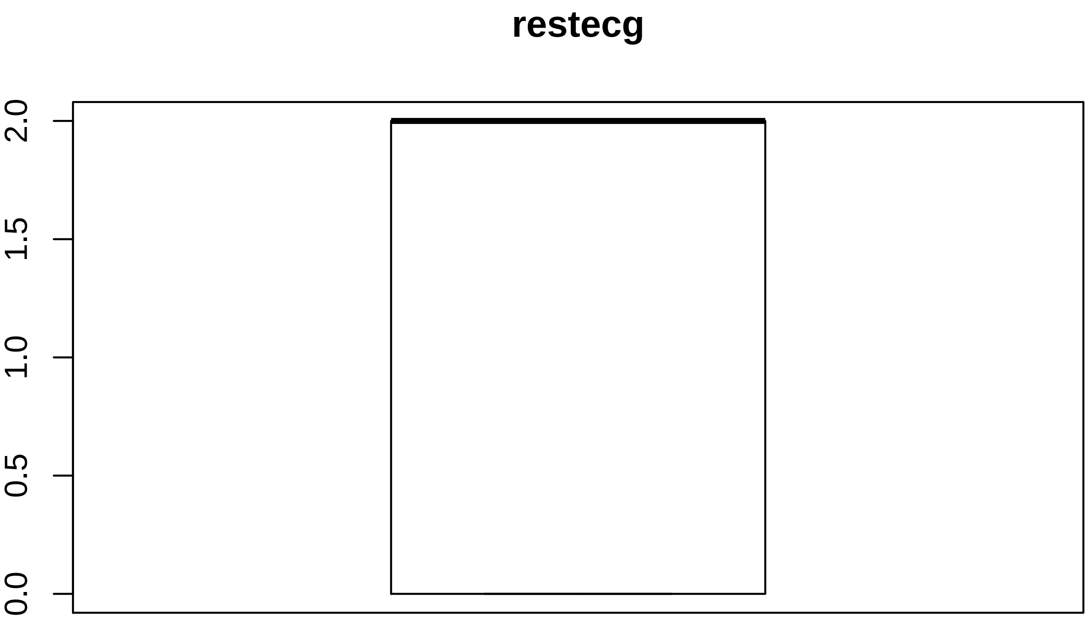

200
150
100
50
0
0
1
fbs

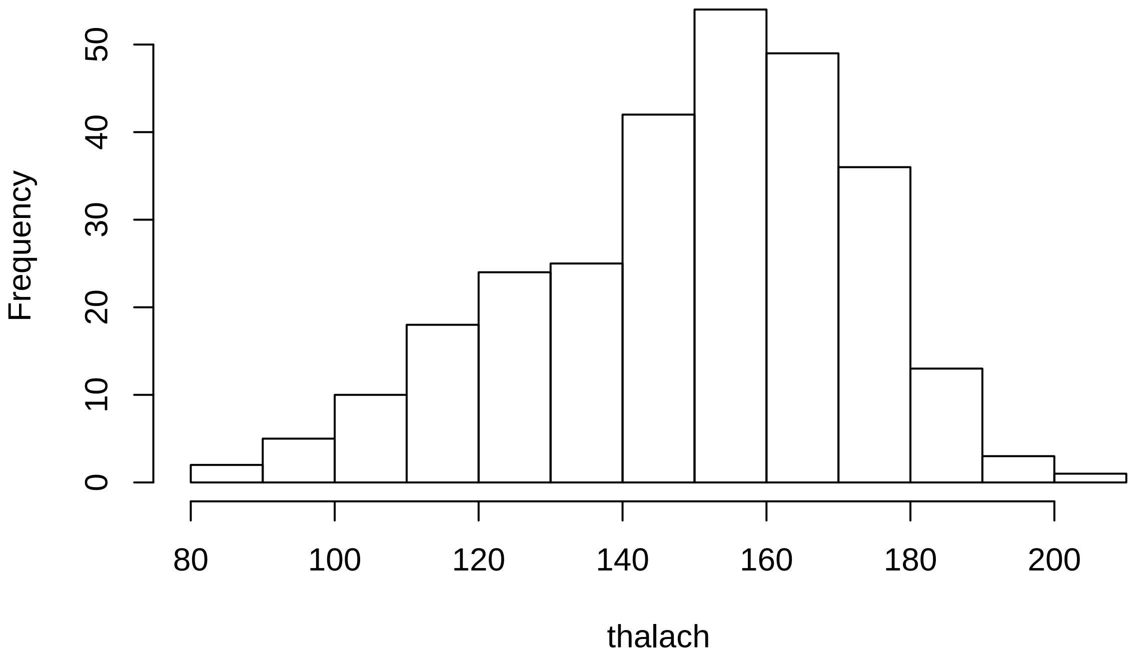

fbs
1.0
0.8
0.6
0.4
0.2
0.0

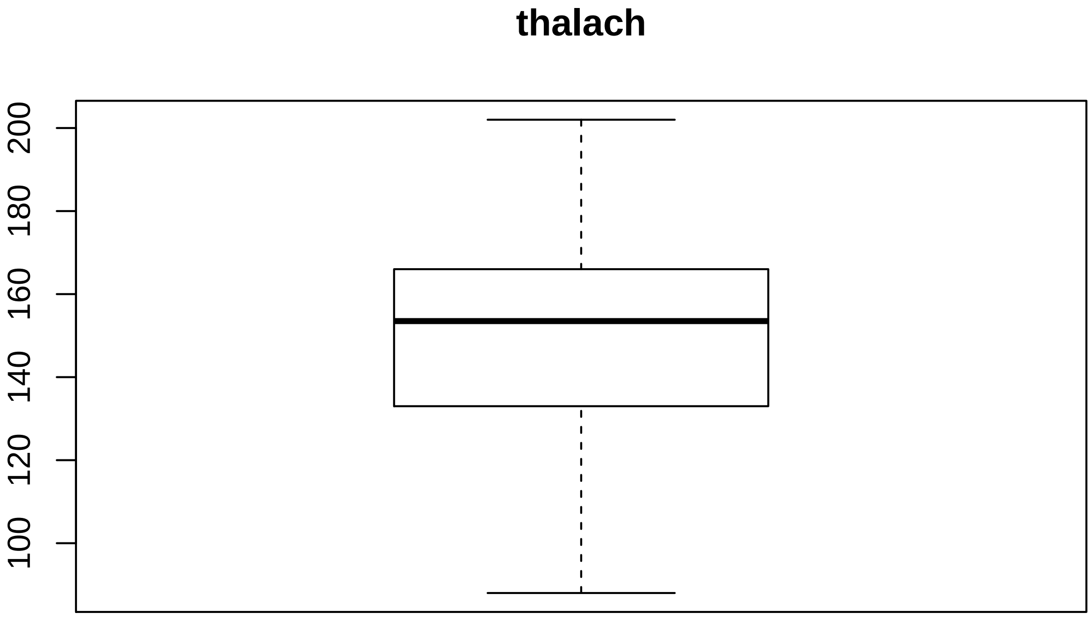

150
100
50
0
0
1
famhist

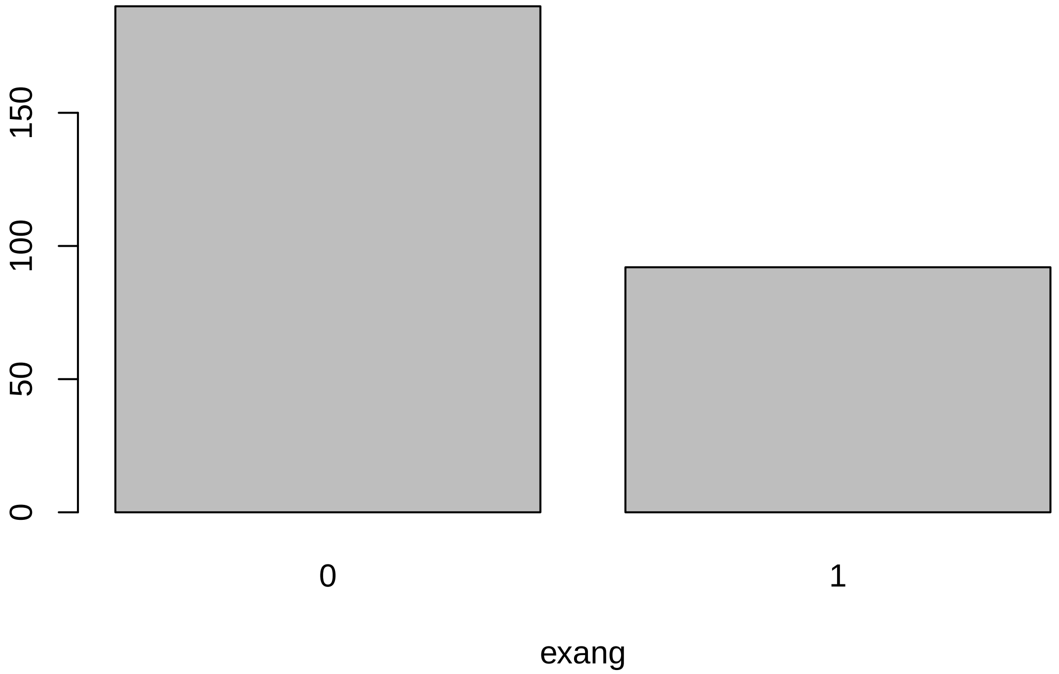

famhist
1.0
0.8
0.6
0.4
0.2
0.0

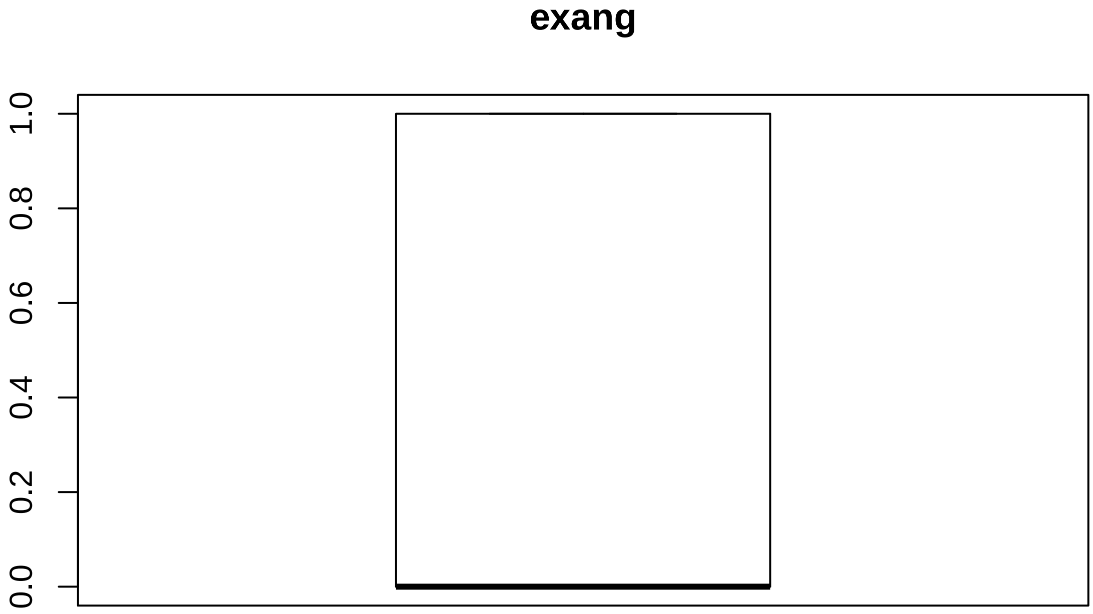

140
100
80
60
40
20
0
0
1
2
restecg

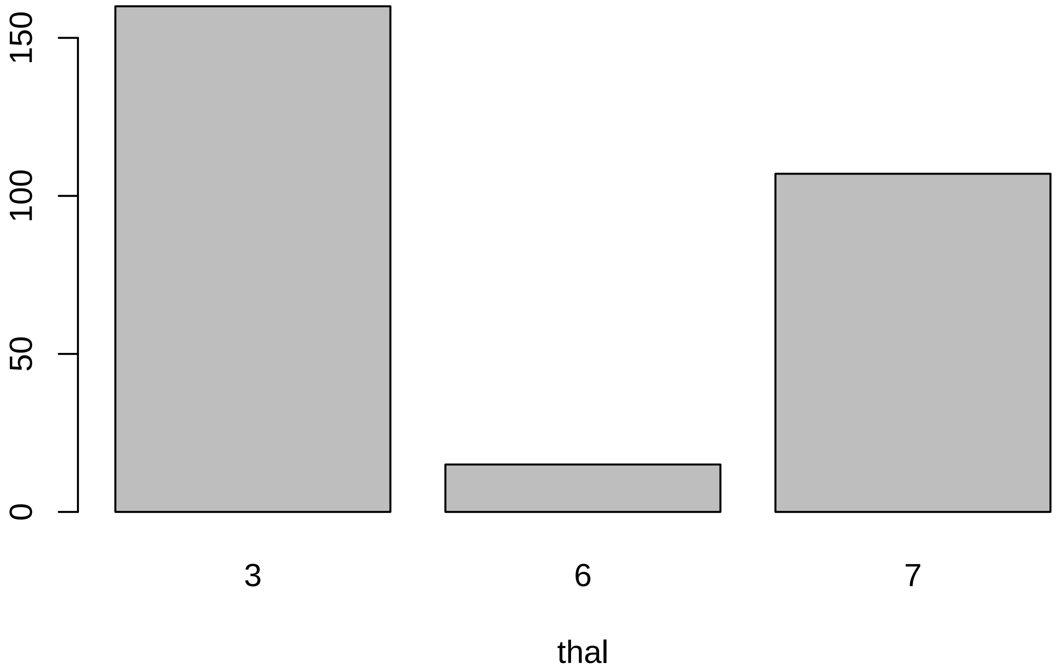

restecg
2.0
1.5
1.0
0.5
0.0

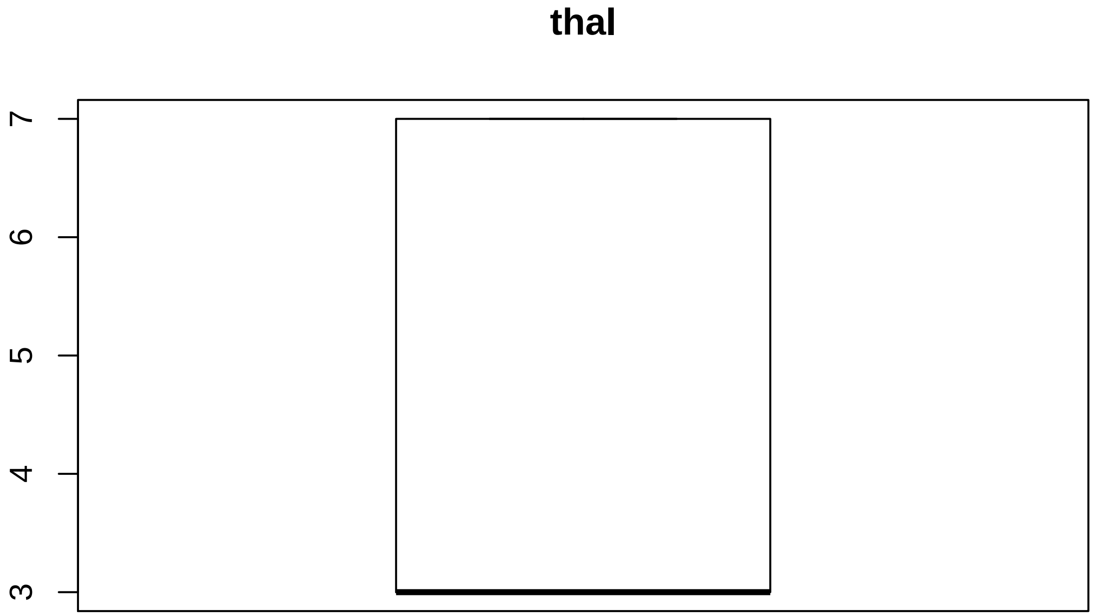

Frequency
0
10
20
30
40
50
80
100
120
140
160
180
200
thalach

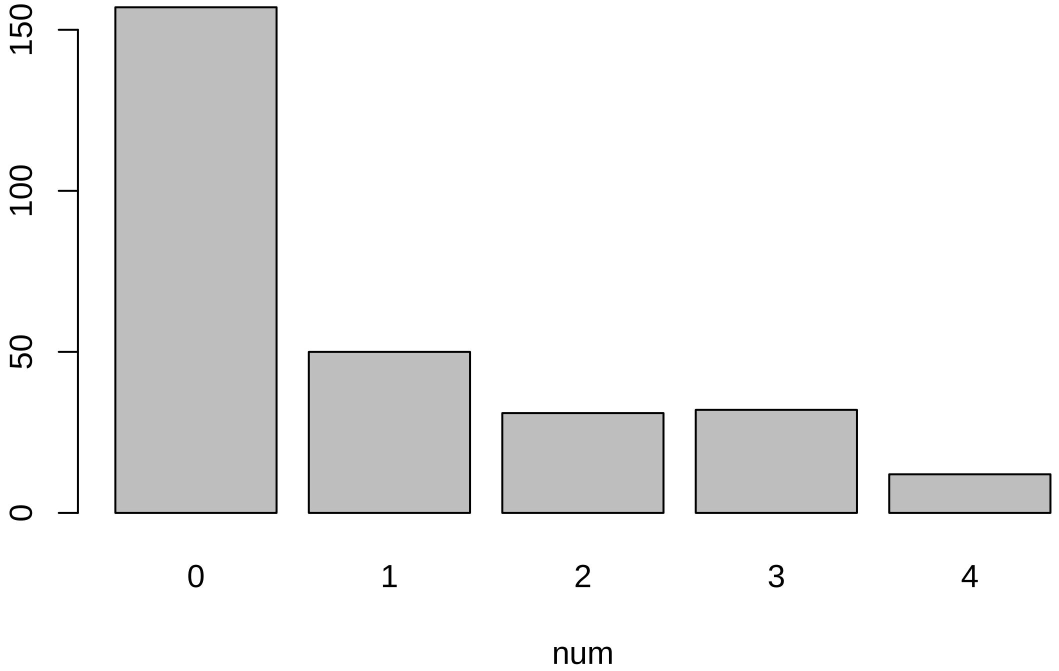

thalach
100
120
140
160
180
200

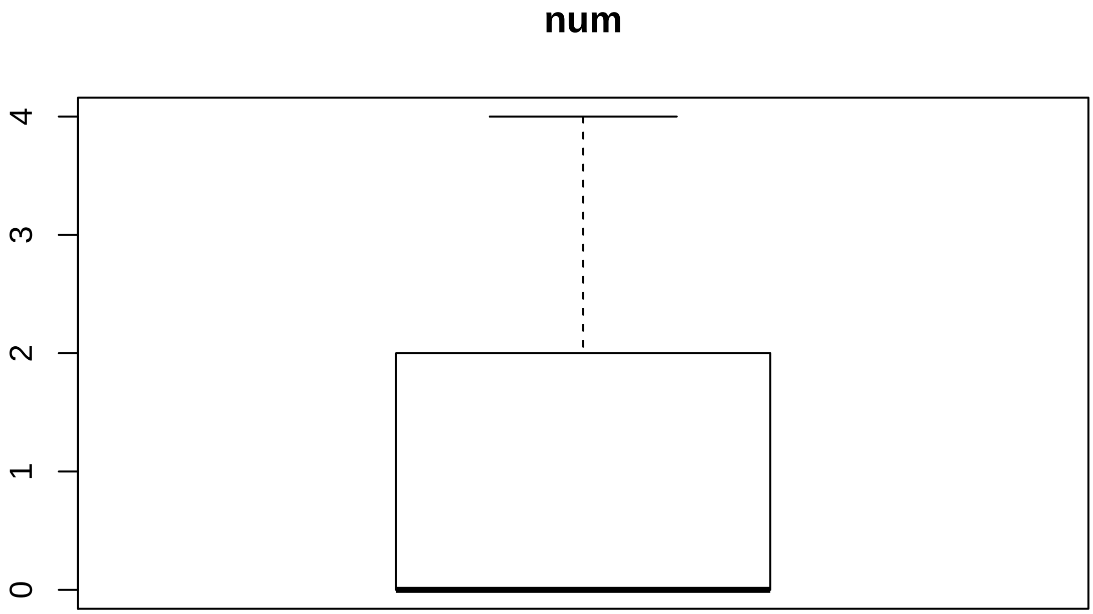

150
100
50
0
0
1
exang


exang
1.0
0.8
0.6
0.4
0.2
0.0


150
100
50
0
3
6
7
thal

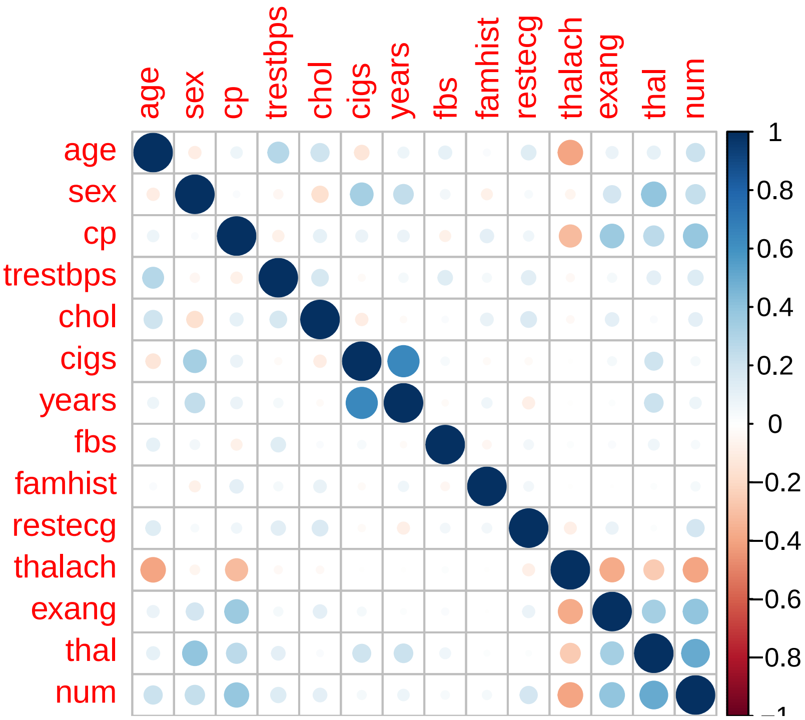

thal
7
6
5
4
3





We can see a number of things are positively correlated to the diagnosis, all except fasting blood sugar and cholesterol. Heart rate is the only thing negatively correlated, so the higher the heart rate (during exercise) the less chance of heart disease.

```
corrplot(cor(dt.nona), mar=c(3, 1, 1, 1))  # play around with margin parameters till it fits the screen
```


age
sex
cp
trestbps
chol
cigs
years
fbs
famhist
restecg
thalach
exang
thal
num
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1