Deliverable 1

Data Processing, Description, Validation and Profiling

Jordi Catafal, Lluís Cerdà, Judit Serna, Tomàs Serra

March 21, 2024

Contents

Data set information

This dataset, compiled in 1988, encompasses information from four distinct databases: Cleveland, Hungary, Switzerland, and Long Beach V. Comprising 76 attributes, inclusive of the predicted attribute, the dataset has been predominantly utilized in published experiments focusing on a subset of 14 key features. The critical "target" field denotes the percentage of heart attack risk in patients

Attribute Information

AGE: age in years SEX: (1 = male; 0 = female)

CP (Chest Pain Type):

--Value 0: typical angina (most serious)

--Value 1: atypical angina

--Value 2: non-anginal pain

--Value 3: asymptomatic (least serious)

TRESTBPS: resting blood pressure (in mm Hg on admission to the hospital)

CHOL: serum cholesterol in mg/dl

FBS: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) A fasting blood sugar level less than 100 mg/dL is normal. From 100 to 120 mg is considered prediabetes. If it is 125 mg/dL or higher on two separate tests, you have diabetes.

RESTECG (Resting Electrocardiographic Results):

--Value 0: normal

--Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

--Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

THALACH: maximum heart rate achieved

EXANG: exercise induced angina (1 = yes; 0 = no)

OLDPEAK: ST depression induced by exercise relative to rest SLOPE (the slope of the peak exercise ST segment):

--Value 0: upsloping

--Value 1: flat

--Value 2: downsloping

CA: number of major vessels (0-3) colored by fluoroscopy

THAL: 3 = normal; 6 = fixed defect; 7 = reversible defect

TARGET: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing

Load libraries

```
rm(list=ls())

library(AER)

## Loading required package: car

## Loading required package: carData

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

library(car)
library(FactoMineR)
```

Load data and take a look

We load the necessary packages and set working directory.

```
file_path = "C:/Users/judit/Documents/Universitat/ADEI/heart.csv"
df = read.csv(file_path, header = T)
head(df)

##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca
thal
## 1  52   1  0      125  212   0       1     168     0     1.0     2  2
3
## 2  53   1  0      140  203   1       0     155     1     3.1     0  0
3
```

```
## 3  70   1  0      145  174   0       1     125     1     2.6     0  0
3
## 4  61   1  0      148  203   0       1     161     0     0.0     2  1
3
## 5  62   0  0      138  294   1       1     106     0     1.9     1  3
2
## 6  58   0  0      100  248   0       0     122     0     1.0     1  0
2
##   target
## 1   0.23
## 2   0.37
## 3   0.24
## 4   0.28
## 5   0.21
## 6   0.78
```

Categorical data: change numeric value to categorical

In the attribute information section and as seen in the head of the dataset, categorical data is initially represented with numerical values. Below, we convert these numerical representations to their categorical counterparts.

Additionally, we illustrate the visualization of two variables: "sex" (categorical) using a plot, and "age" (numerical) using a histogram. We only do it for these two because we have many variables.

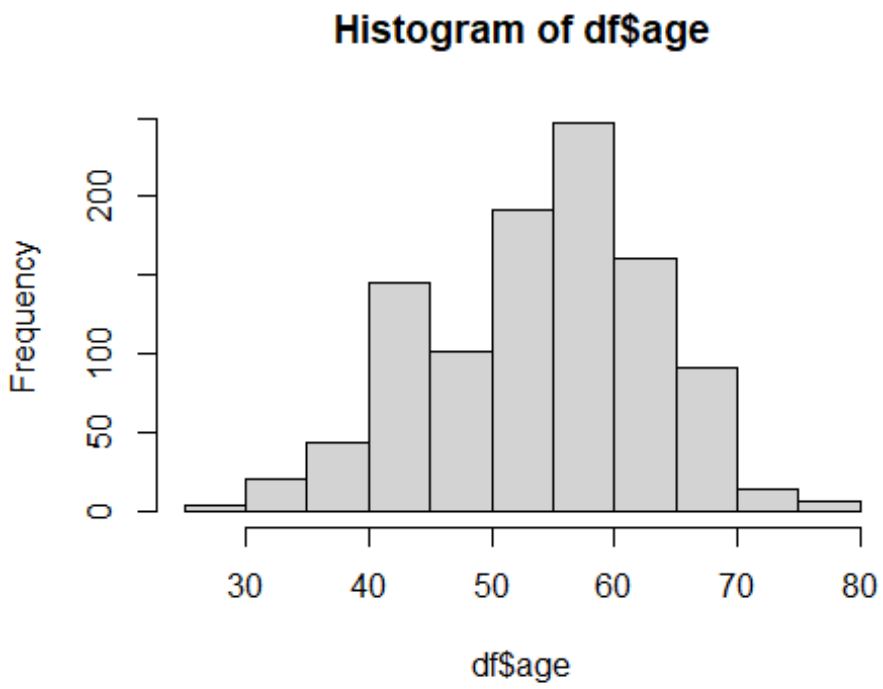
Furthermore, we define the data types as either numeric or categorical (factor).

```r
df$sex[which(df$sex==0)] <- "Female"
df$sex[which(df$sex==1)] <- "Male"
df$sex <- as.factor(df$sex)

df$cp[which(df$cp==0)] <- "Typical Angina"
df$cp[which(df$cp==1)] <- "Atypical Angina"
df$cp[which(df$cp==2)] <- "Non-anginal Pain"
df$cp[which(df$cp==3)] <- "Asymptomatic"
df$cp <- as.factor(df$cp)

df$fbs[which(df$fbs==0)] <- "<= 120 mg/dL"
df$fbs[which(df$fbs==1)] <- "> 120 mg/dL"
df$fbs <- as.factor(df$fbs)

df$restecg[which(df$restecg==0)] <- "Normal"
df$restecg[which(df$restecg==1)] <- "Abnormality"
df$restecg[which(df$restecg==2)] <- "Hypertrophy"
df$restecg <- as.factor(df$restecg)

df$exang[which(df$exang==0)] <- "No"
```

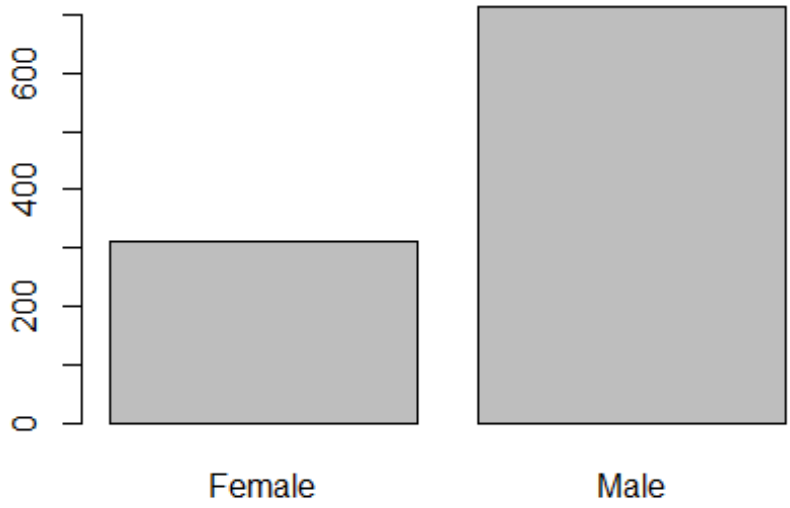
```r
df$exang[which(df$exang==1)] <- "Yes"
df$exang <- as.factor(df$exang)

df$slope[which(df$slope==0)] <- "Upsloping"
df$slope[which(df$slope==1)] <- "Flat"
df$slope[which(df$slope==2)] <- "Downsloping"
df$slope <- as.factor(df$slope)

df$thal[which(df$thal==0)] <- "Normal"
df$thal[which(df$thal==1)] <- "Fixed Defect"
df$thal[which(df$thal==2)] <- "Reversible Defect"
df$thal <- as.factor(df$thal)

head(df)
```

```
##   age    sex             cp trestbps chol          fbs     restecg
thalach
## 1  52   Male Typical Angina      125  212 <= 120 mg/dL Abnormality
168
## 2  53   Male Typical Angina      140  203  > 120 mg/dL      Normal
155
## 3  70   Male Typical Angina      145  174 <= 120 mg/dL Abnormality
125
## 4  61   Male Typical Angina      148  203 <= 120 mg/dL Abnormality
161
## 5  62 Female Typical Angina      138  294  > 120 mg/dL Abnormality
106
## 6  58 Female Typical Angina      100  248 <= 120 mg/dL      Normal
122
##   exang oldpeak       slope ca              thal target
## 1    No     1.0 Downsloping  2                 3   0.23
## 2   Yes     3.1   Upsloping  0                 3   0.37
## 3   Yes     2.6   Upsloping  0                 3   0.24
## 4    No     0.0 Downsloping  1                 3   0.28
## 5    No     1.9        Flat  3 Reversible Defect   0.21
## 6    No     1.0        Flat  0 Reversible Defect   0.78
```

```r
hist(df$age)
```




```r
plot(df$sex)
```




```r
df$age <- as.numeric(df$age)
df$trestbps <- as.numeric(df$trestbps)
```

```r
df$chol <- as.numeric(df$chol)
df$thalach <- as.numeric(df$thalach)
df$oldpeak <- as.numeric(df$oldpeak)

sapply(df, class)
```

```
##       age       sex        cp  trestbps      chol       fbs   restecg
thalach
## "numeric"  "factor"  "factor" "numeric" "numeric"  "factor"  "factor"
"numeric"
##     exang   oldpeak     slope        ca      thal    target
##  "factor" "numeric"  "factor" "integer"  "factor" "numeric"
```

Detection of errors

Upon reviewing the dataframe summary, we identify errors in the attributes "thal" and "ca". Specifically, for "thal", some individuals possess a value of 3, which does not align with the predefined categorical values. Similarly, for "ca", the maximum value observed is 4, whereas it should range from 0 to 3. We then change this values to NA to be treated later.

```r
summary(df)
```

```
##       age            sex                      cp         trestbps
##  Min.   :29.00   Female:312   Asymptomatic    : 77   Min.   : 94.0
##  1st Qu.:48.00   Male  :713   Atypical Angina :167   1st Qu.:120.0
##  Median :56.00                Non-anginal Pain:284   Median :130.0
##  Mean   :54.43                Typical Angina  :497   Mean   :131.6
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##       chol               fbs             restecg       thalach
exang
##  Min.   :126   <= 120 mg/dL:872   Abnormality:513   Min.   : 71.0   No
:680
##  1st Qu.:211   > 120 mg/dL :153   Hypertrophy: 15   1st Qu.:132.0
Yes:345
##  Median :240                      Normal     :497   Median :152.0
##  Mean   :246                                        Mean   :149.1
##  3rd Qu.:275                                        3rd Qu.:166.0
##  Max.   :564                                        Max.   :202.0
##     oldpeak              slope           ca
thal
##  Min.   :0.000   Downsloping:469   Min.   :0.0000   3
:410
##  1st Qu.:0.000   Flat       :482   1st Qu.:0.0000   Fixed Defect     :
64
##  Median :0.800   Upsloping  : 74   Median :0.0000   Normal           :
7
##  Mean   :1.072                     Mean   :0.7541   Reversible
```

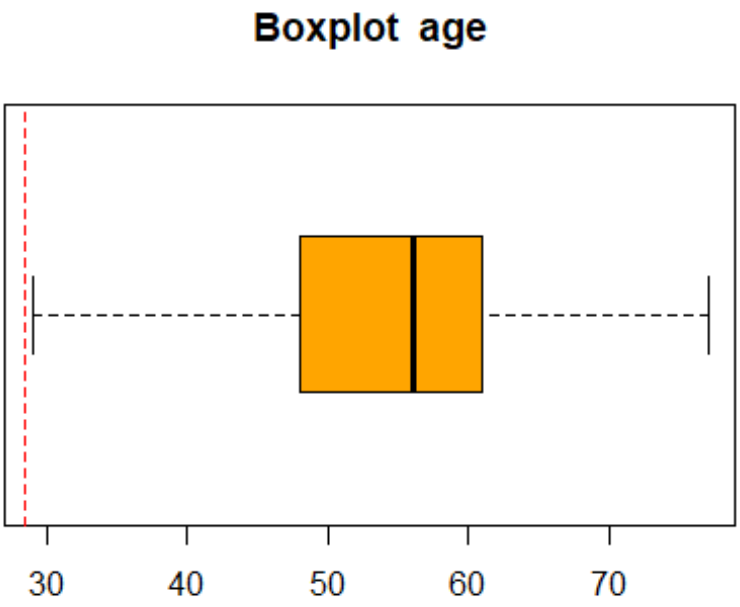
```
Defect:544
##  3rd Qu.:1.800                     3rd Qu.:1.0000
##  Max.   :6.200                     Max.   :4.0000
##      target
##  Min.   :0.1000
##  1st Qu.:0.2600
##  Median :0.7100
##  Mean   :0.5364
##  3rd Qu.:0.8100
##  Max.   :0.9000

df$thal[which(df$thal==3)] <- NA
df$ca[which(df$ca==4)] <- NA

miss_val = sum(is.na(df))
miss_val

## [1] 428

summary(df)

##       age            sex                      cp         trestbps
##  Min.   :29.00   Female:312   Asymptomatic    : 77   Min.   : 94.0
##  1st Qu.:48.00   Male  :713   Atypical Angina :167   1st Qu.:120.0
##  Median :56.00                Non-anginal Pain:284   Median :130.0
##  Mean   :54.43                Typical Angina  :497   Mean   :131.6
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##
##       chol               fbs             restecg       thalach
exang
##  Min.   :126   <= 120 mg/dL:872   Abnormality:513   Min.   : 71.0   No
:680
##  1st Qu.:211   > 120 mg/dL :153   Hypertrophy: 15   1st Qu.:132.0
Yes:345
##  Median :240                     Normal     :497   Median :152.0
##  Mean   :246                                       Mean   :149.1
##  3rd Qu.:275                                       3rd Qu.:166.0
##  Max.   :564                                       Max.   :202.0
##
##     oldpeak              slope           ca
thal
##  Min.   :0.000   Downsloping:469   Min.   :0.0000   3               :
0
##  1st Qu.:0.000   Flat       :482   1st Qu.:0.0000   Fixed Defect    :
64
##  Median :0.800   Upsloping  : 74   Median :0.0000   Normal          :
7
##  Mean   :1.072                     Mean   :0.6961   Reversible
Defect:544
##  3rd Qu.:1.800                     3rd Qu.:1.0000   NA's
```

```
:410
##  Max.   :6.200                     Max.   :3.0000
##                                    NA's   :18
##      target
##  Min.   :0.1000
##  1st Qu.:0.2600
##  Median :0.7100
##  Mean   :0.5364
##  3rd Qu.:0.8100
##  Max.   :0.9000
##
```

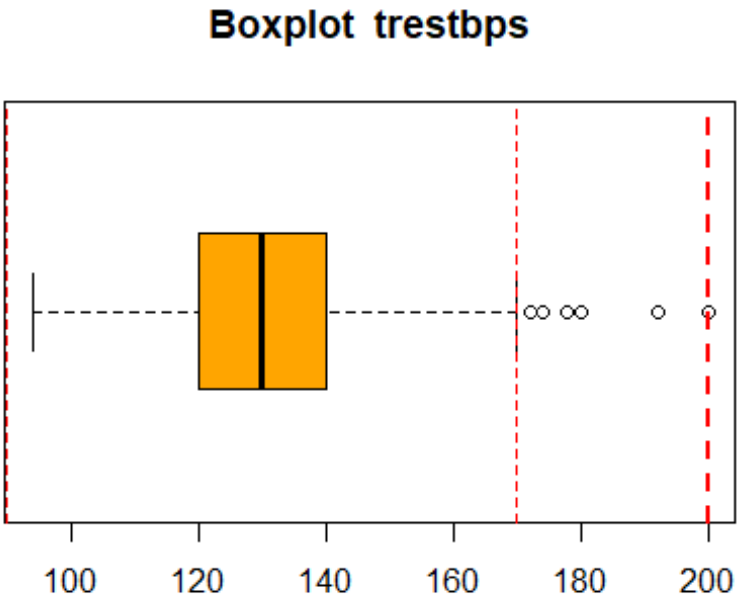
Univariate detection

Here, we plot, for each numeric variable, its boxplot to identify mild and extreme outliers.

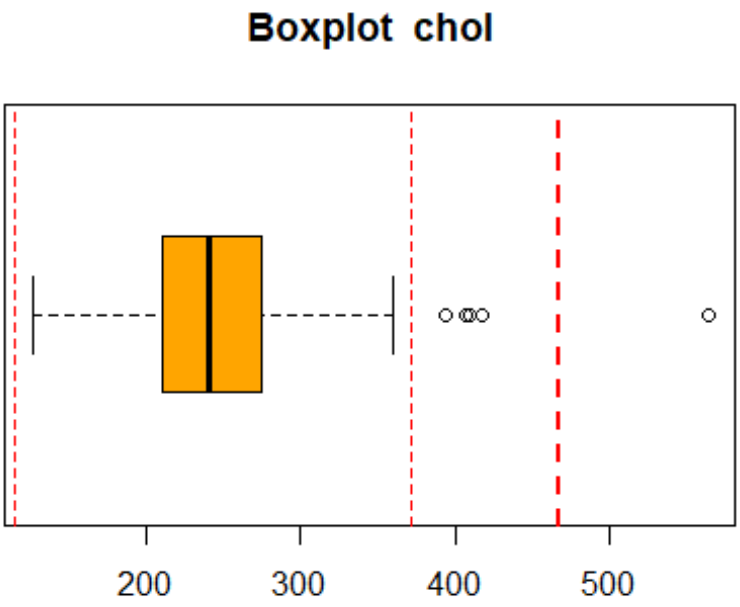
```
outliers <- function(column, name){

  sumlist <- summary(column)
  q1 <- sumlist[2]
  q3 <- sumlist[5]


  boxplot(column, main = paste("Boxplot ", name), col = "orange",
horizontal = T)

  # IQR calculation
  iqr <- q3 - q1

  # Mild inferior limit:
  mild_inf_lim <- sumlist[2]-1.5*iqr
  # Extreme inferior limit:
  extreme_inf_lim <- sumlist[2]-3*iqr
  abline(v=mild_inf_lim, col = "red", lty = 2)
  abline(v=extreme_inf_lim, col = "red",  lty = 2, lwd = 2)
  mild_sup_lim <- sumlist[5]+1.5*iqr
  extreme_sup_lim <- sumlist[5]+3*iqr
  abline(v=mild_sup_lim, col = "red", lty = 2)
  abline(v=extreme_sup_lim, col = "red",  lty = 2, lwd = 2)

}

outliers(df$age, "age")
```



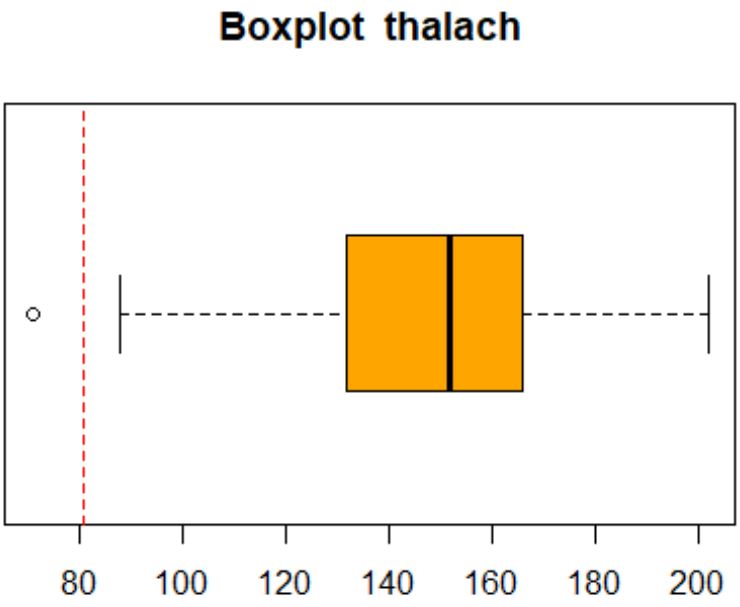

```
outliers(df$trestbps, "trestbps")
```



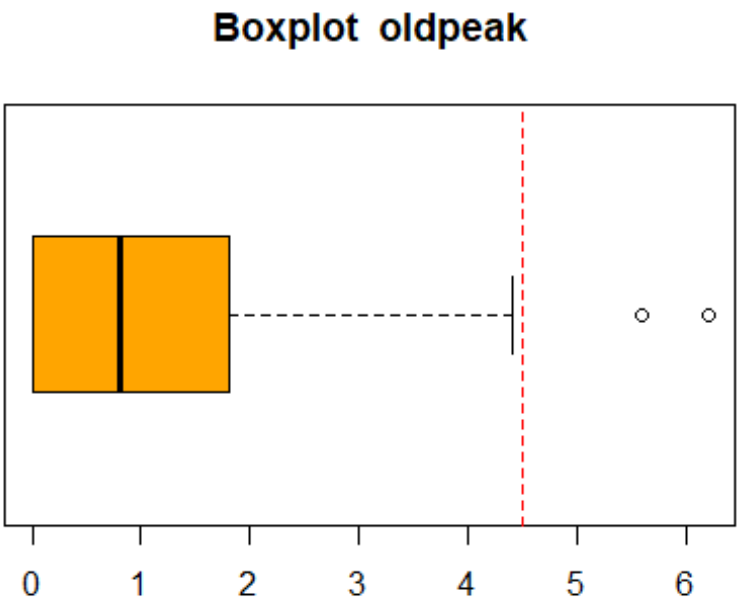

```
outliers(df$chol, "chol")
```



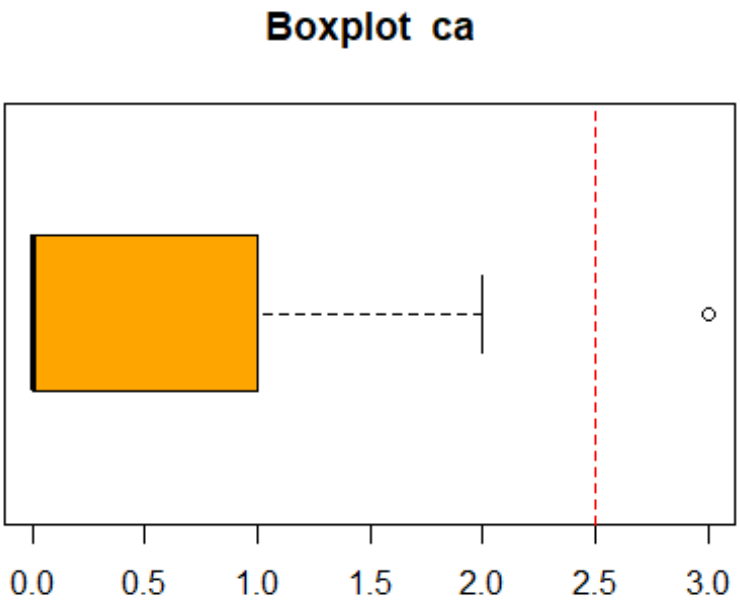

```r
outliers(df$thalach, "thalach")
```



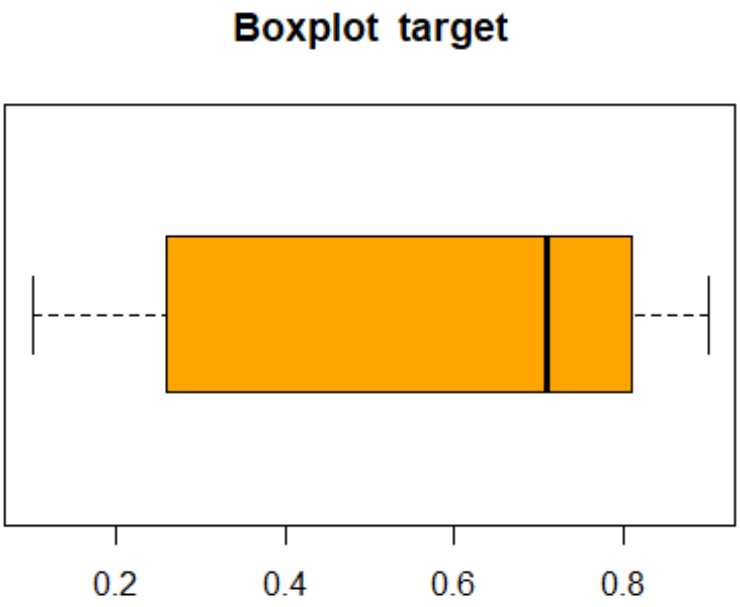

```r
outliers(df$oldpeak, "oldpeak")
```




```
outliers(df$ca, "ca")
```




```
outliers(df$target, "target")
```




```
FindMildOutliers <- function(data) {
  lowerq = quantile(data, na.rm = TRUE)[2]
  upperq = quantile(data, na.rm = TRUE)[4]
  iqr = upperq - lowerq #Or use IQR(data)
  # we identify extreme outliers
  extreme.threshold.upper = (iqr * 1.5) + upperq
  extreme.threshold.lower = lowerq - (iqr * 1.5)
  result <- which(data > extreme.threshold.upper | data <
extreme.threshold.lower)
}

FindExtremeOutliers <- function(data) {
  lowerq = quantile(data, na.rm = TRUE)[2]
  upperq = quantile(data, na.rm = TRUE)[4]
  iqr = upperq - lowerq #Or use IQR(data)
  # we identify extreme outliers
  extreme.threshold.upper = (iqr * 3) + upperq
  extreme.threshold.lower = lowerq - (iqr * 3)
  result <- which(data > extreme.threshold.upper | data <
extreme.threshold.lower)
}


FindMissingValues <- function(data) {
  result <- which(sum(is.na(data)) > 0)
}
```

Finding all Missing Values by Variables

The dataframe had no missing values. The numbers shown below are the changed errors by NA values, as we'll treat them as such.

```
miss <- colSums(is.na(df))
rank_miss <- sort(miss, decreasing = TRUE)
rank_miss
```

```
##     thal       ca      age      sex       cp trestbps     chol
fbs
##      410       18        0        0        0        0        0
0
##  restecg  thalach    exang  oldpeak    slope   target
##        0        0        0        0        0        0
```

Finding all Extreme Outliers by Variables

As seen in the boxplots, the only attribute with outliers is "chol".

```
df2 <- Filter(is.numeric, df)
pos <- lapply(df2, FindExtremeOutliers)
extr_out <- lengths(pos)
num_outliers = length(pos)
rank_extr <- sort(extr_out, decreasing = TRUE)
rank_extr
```

```
##     chol      age trestbps  thalach  oldpeak       ca   target
##        3        0        0        0        0        0        0
```

Finding all Mild Outliers by Variables

We find the exact number of mild outliers as seen in the boxplots, and rank them.

```
df2 <- Filter(is.numeric, df)
pos <- lapply(df2, FindMildOutliers)
mild_out <- lengths(pos)
rank_mild <- sort(mild_out, decreasing = TRUE)
rank_mild
```

```
##       ca trestbps     chol  oldpeak  thalach      age   target
##       69       30       16        7        4        0        0
```

Finding all Missing Values by Individuals

```
pos <- apply(df, 1, FindMissingValues)
pos <- which(pos > 0)
pos
```

```
##   [1]    1    2    3    4    8    9   12   14   18   26   33   34   39
40   42
##  [16]   44   48   50   52   53   54   55   56   57   59   66   70   71
72   74
##  [31]   75   78   81   84   90   93   98  102  103  107  108  109  112
```

```
113  114
##  [46]  116  117  118  121  123  125  129  135  141  146  148  149  152
153  154
##  [61]  155  156  157  158  159  164  166  172  175  176  177  180  181
183  186
##  [76]  187  189  190  192  193  197  200  203  204  209  210  211  213
217  219
##  [91]  222  226  230  231  233  236  238  239  242  243  247  248  253
254  255
## [106]  259  262  267  269  271  272  276  277  279  285  287  290  291
292  295
## [121]  298  302  304  306  311  312  313  315  317  319  323  327  335
336  337
## [136]  340  341  344  347  349  350  352  353  355  362  367  368  369
371  372
## [151]  381  382  383  384  385  392  394  395  397  398  401  405  408
415  418
## [166]  425  426  429  431  432  438  442  443  448  451  453  455  461
464  465
## [181]  466  472  475  476  477  478  480  481  482  483  486  487  488
493  494
## [196]  496  497  507  509  512  516  517  519  520  521  522  524  525
527  529
## [211]  531  539  540  541  546  550  551  553  555  556  557  570  572
575  577
## [226]  580  581  584  585  586  587  589  593  596  598  601  602  611
612  614
## [241]  615  621  622  625  626  627  628  631  635  642  643  644  646
651  654
## [256]  659  668  672  673  674  675  676  677  678  680  682  684  686
688  689
## [271]  691  693  695  696  698  700  702  705  706  707  710  711  712
717  720
## [286]  726  727  728  729  733  737  738  739  740  743  744  748  750
754  763
## [301]  765  766  767  768  769  774  777  778  782  785  786  787  788
790  791
## [316]  792  793  798  799  801  806  809  811  812  814  816  821  823
828  830
## [331]  832  833  834  838  841  843  845  850  853  854  856  860  863
864  873
## [346]  876  877  879  881  883  884  885  886  887  889  890  891  893
895  896
## [361]  897  900  909  910  911  913  916  917  920  921  922  923  926
928  929
## [376]  930  931  932  934  939  941  944  945  946  949  950  952  956
958  963
## [391]  964  967  969  971  972  975  976  980  982  984  985  986  989
992  994
## [406]  995  996  997  998 1000 1004 1006 1008 1010 1011 1013 1016 1018
```

```
1019 1022
## [421] 1025
```

Finding all Extreme Outliers by Individuals

```
df2 <- Filter(is.numeric, df)
posExtremeInd <- apply(df2, 1, FindExtremeOutliers)
posExtremeInd <- which(posExtremeInd > 0)
posExtremeInd
```

```
## [1] 159 193 465
```

Finding all Mild Outliers by Individuals

```
df2 <- Filter(is.numeric, df)
pos <- apply(df2, 1, FindMildOutliers)
pos <- which(pos > 0)
pos
```

```
##  [1]    7   12   63   98  108  114  124  151  155  159  168  180  193
210  212
## [16]  253  256  267  296  329  353  371  372  395  414  423  436  451
465  482
## [31]  486  494  543  547  553  578  579  585  611  642  662  666  675
686  743
## [46]  766  767  778  790  890  959  997 1014 1018
```

```
length(pos)
```

```
## [1] 54
```

Create variable adding the total number missing values, outliers and errors.

```
num_outliers_miss_errors = miss_val + num_outliers
num_outliers_miss_errors
```

```
## [1] 435
```

Imputation

Imputation of factor thal

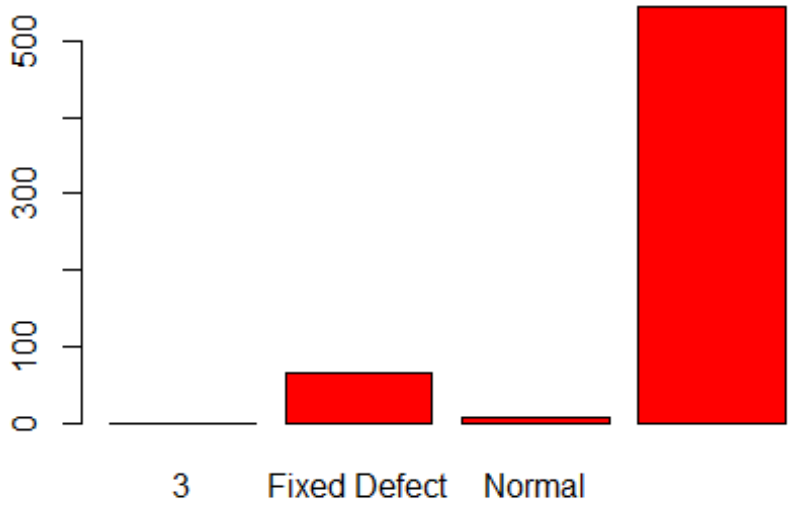
```
library(missMDA)
```

```
## Warning: package 'missMDA' was built under R version 4.3.3
```

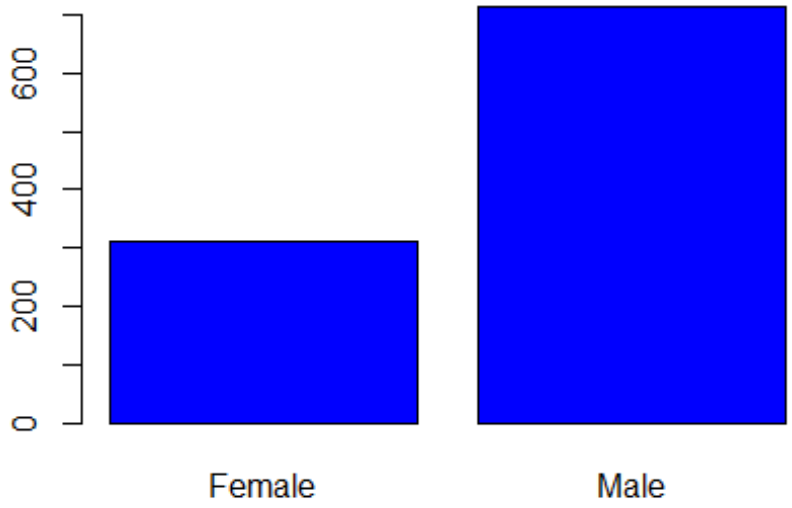
```
# Categorical imputation
f <- Filter(is.factor, df)
vars_dis = colnames(f)
summary(df[,vars_dis])
```

```
##      sex                      cp                 fbs             restecg
##  Female:312   Asymptomatic    : 77   <= 120 mg/dL:872
Abnormality:513
```

```
##  Male   :713   Atypical Angina :167   > 120 mg/dL :153   Hypertrophy:
15
##                Non-anginal Pain:284                      Normal
:497
##                Typical Angina  :497
##
##  exang              slope                     thal
##  No :680   Downsloping:469   3                 :  0
##  Yes:345   Flat       :482   Fixed Defect      : 64
##            Upsloping  : 74   Normal            :  7
##                              Reversible Defect:544
##                              NA's              :410

res.input<-imputeMCA(df[,vars_dis],method="EM")
summary(res.input$completeObs)

##      sex                     cp                fbs             restecg
##  Female:312   Asymptomatic    : 77   <= 120 mg/dL:872
Abnormality:513
##  Male   :713   Atypical Angina :167   > 120 mg/dL :153   Hypertrophy:
15
##                Non-anginal Pain:284                      Normal
:497
##                Typical Angina  :497
##  exang              slope                     thal
##  No :680   Downsloping:469   Fixed Defect      : 82
##  Yes:345   Flat       :482   Normal            :  7
##            Upsloping  : 74   Reversible Defect:936
##

# Validation
barplot(table(df$thal),col="red")
```




```
barplot(table(res.input$completeObs[,1]),col="blue")
```




```
df[,vars_dis] <- res.input$completeObs
summary(df)
```

```
##       age            sex                      cp          trestbps
##  Min.   :29.00   Female:312   Asymptomatic    : 77   Min.   : 94.0
##  1st Qu.:48.00   Male  :713   Atypical Angina :167   1st Qu.:120.0
##  Median :56.00                Non-anginal Pain:284   Median :130.0
##  Mean   :54.43                Typical Angina  :497   Mean   :131.6
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##
##       chol               fbs             restecg       thalach
exang
##  Min.   :126   <= 120 mg/dL:872   Abnormality:513   Min.   : 71.0   No
:680
##  1st Qu.:211   > 120 mg/dL :153   Hypertrophy: 15   1st Qu.:132.0
Yes:345
##  Median :240                      Normal     :497   Median :152.0
##  Mean   :246                                        Mean   :149.1
##  3rd Qu.:275                                        3rd Qu.:166.0
##  Max.   :564                                        Max.   :202.0
##
##     oldpeak              slope           ca
thal
##  Min.   :0.000   Downsloping:469   Min.   :0.0000   Fixed Defect     :
82
##  1st Qu.:0.000   Flat       :482   1st Qu.:0.0000   Normal           :
7
##  Median :0.800   Upsloping  : 74   Median :0.0000   Reversible
Defect:936
##  Mean   :1.072                     Mean   :0.6961
##  3rd Qu.:1.800                     3rd Qu.:1.0000
##  Max.   :6.200                     Max.   :3.0000
##                                    NA's   :18
##      target
##  Min.   :0.1000
##  1st Qu.:0.2600
##  Median :0.7100
##  Mean   :0.5364
##  3rd Qu.:0.8100
##  Max.   :0.9000
##
```

```
miss_val = sum(is.na(df))
miss_val
```

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

Imputation of numeric ca

```
library(class)
# Numeric imputation  only explanatory variables - never for target
n <- Filter(is.numeric, df)
summary(n)
```

```
##       age           trestbps          chol        thalach
oldpeak
##  Min.   :29.00   Min.   : 94.0   Min.   :126   Min.   : 71.0   Min.
:0.000
##  1st Qu.:48.00   1st Qu.:120.0   1st Qu.:211   1st Qu.:132.0   1st
Qu.:0.000
##  Median :56.00   Median :130.0   Median :240   Median :152.0   Median
:0.800
##  Mean   :54.43   Mean   :131.6   Mean   :246   Mean   :149.1   Mean
:1.072
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:275   3rd Qu.:166.0   3rd
Qu.:1.800
##  Max.   :77.00   Max.   :200.0   Max.   :564   Max.   :202.0   Max.
:6.200
##
##        ca             target
##  Min.   :0.0000   Min.   :0.1000
##  1st Qu.:0.0000   1st Qu.:0.2600
##  Median :0.0000   Median :0.7100
##  Mean   :0.6961   Mean   :0.5364
##  3rd Qu.:1.0000   3rd Qu.:0.8100
##  Max.   :3.0000   Max.   :0.9000
##  NA's   :18
```

```
vars_con = colnames(n)

fullvariables <- c(1,4,5,8,10,14)
aux <- df[,fullvariables]
dim(aux)
```

```
## [1] 1025    6
```

```
names(aux)
```

```
## [1] "age"      "trestbps" "chol"     "thalach"  "oldpeak"  "target"
```

```
summary(df)
```

```
##       age            sex                      cp         trestbps
##  Min.   :29.00   Female:312   Asymptomatic    : 77   Min.   : 94.0
##  1st Qu.:48.00   Male  :713   Atypical Angina :167   1st Qu.:120.0
##  Median :56.00                Non-anginal Pain:284   Median :130.0
##  Mean   :54.43                Typical Angina  :497   Mean   :131.6
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##
##       chol               fbs             restecg       thalach
exang
##  Min.   :126   <= 120 mg/dL:872   Abnormality:513   Min.   : 71.0   No
:680
##  1st Qu.:211   > 120 mg/dL :153   Hypertrophy: 15   1st Qu.:132.0
```

```
Yes:345
##  Median :240                      Normal     :497   Median :152.0
##  Mean   :246                                        Mean   :149.1
##  3rd Qu.:275                                        3rd Qu.:166.0
##  Max.   :564                                        Max.   :202.0
##
##     oldpeak              slope           ca
thal
##  Min.   :0.000   Downsloping:469   Min.   :0.0000   Fixed Defect     :
82
##  1st Qu.:0.000   Flat       :482   1st Qu.:0.0000   Normal           :
7
##  Median :0.800   Upsloping  : 74   Median :0.0000   Reversible
Defect:936
##  Mean   :1.072                     Mean   :0.6961
##  3rd Qu.:1.800                     3rd Qu.:1.0000
##  Max.   :6.200                     Max.   :3.0000
##                                    NA's   :18
##      target
##  Min.   :0.1000
##  1st Qu.:0.2600
##  Median :0.7100
##  Mean   :0.5364
##  3rd Qu.:0.8100
##  Max.   :0.9000
##
```

```
aux1 <- aux[!is.na(df$ca),]
dim(aux1)
```

```
## [1] 1007    6
```

```
aux2 <- aux[is.na(df$ca),]
dim(aux2)
```

```
## [1] 18  6
```

```
knn.ing = knn(aux1, aux2, df$ca[!is.na(df$ca)])

df$ca[is.na(df$ca)] <- as.numeric(as.character(knn.ing))

# Validation
summary(df)
```

```
##       age            sex                      cp         trestbps
##  Min.   :29.00   Female:312   Asymptomatic    : 77   Min.   : 94.0
##  1st Qu.:48.00   Male  :713   Atypical Angina :167   1st Qu.:120.0
##  Median :56.00                Non-anginal Pain:284   Median :130.0
##  Mean   :54.43                Typical Angina  :497   Mean   :131.6
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
```

```
##       chol                fbs              restecg       thalach
exang
##  Min.   :126   <= 120 mg/dL:872   Abnormality:513   Min.   : 71.0   No
:680
##  1st Qu.:211   > 120 mg/dL :153   Hypertrophy: 15   1st Qu.:132.0
Yes:345
##  Median :240                     Normal     :497   Median :152.0
##  Mean   :246                                       Mean   :149.1
##  3rd Qu.:275                                       3rd Qu.:166.0
##  Max.   :564                                       Max.   :202.0
##     oldpeak              slope           ca
thal
##  Min.   :0.000   Downsloping:469   Min.   :0.0000   Fixed Defect     :
82
##  1st Qu.:0.000   Flat       :482   1st Qu.:0.0000   Normal           :
7
##  Median :0.800   Upsloping  : 74   Median :0.0000   Reversible
Defect:936
##  Mean   :1.072                     Mean   :0.6878
##  3rd Qu.:1.800                     3rd Qu.:1.0000
##  Max.   :6.200                     Max.   :3.0000
##      target
##  Min.   :0.1000
##  1st Qu.:0.2600
##  Median :0.7100
##  Mean   :0.5364
##  3rd Qu.:0.8100
##  Max.   :0.9000
```

```
miss_val = sum(is.na(df))
miss_val
```

```
## [1] 0
```

Compute the correlation with all other variables. Rank these variables according the correlation

```
library(FactoMineR)
library(mvoutlier)
```

```
## Warning: package 'mvoutlier' was built under R version 4.3.3
```

```
## Loading required package: sgeostat
```

```
df2 <- Filter(is.numeric, df)
res <- cor(df2)
round(res, 2)
```

```
##             age trestbps  chol thalach oldpeak    ca target
## age        1.00     0.27  0.22   -0.39    0.21  0.37  -0.22
## trestbps   0.27     1.00  0.13   -0.04    0.19  0.10  -0.14
## chol       0.22     0.13  1.00   -0.02    0.06  0.14  -0.09
```

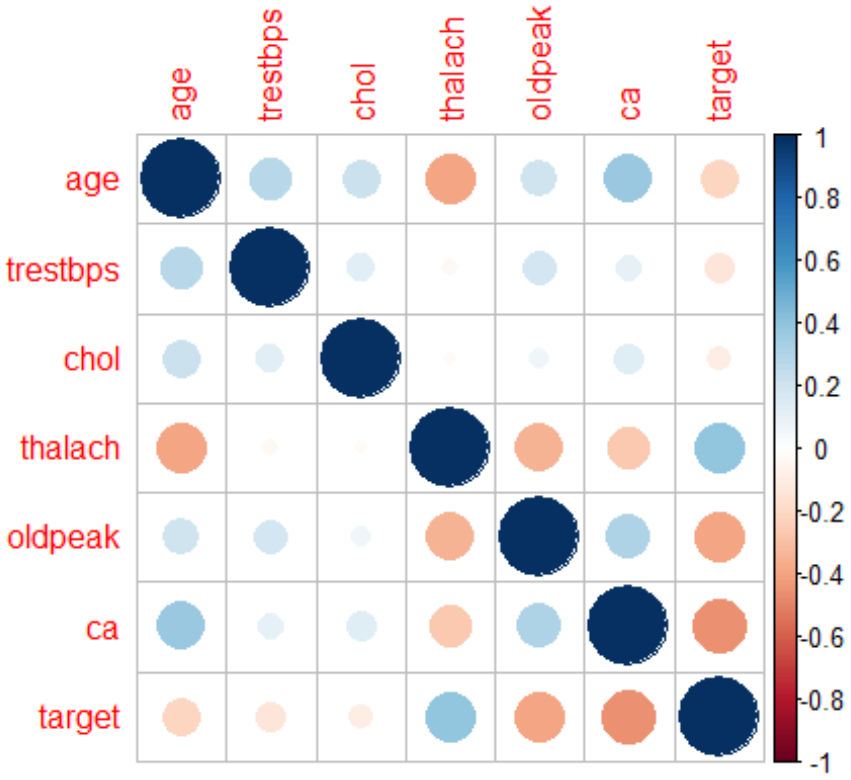
```
## thalach   -0.39    -0.04 -0.02    1.00   -0.35 -0.27   0.40
## oldpeak    0.21     0.19  0.06   -0.35    1.00  0.30  -0.40
## ca         0.37     0.10  0.14   -0.27    0.30  1.00  -0.46
## target    -0.22    -0.14 -0.09    0.40   -0.40 -0.46   1.00
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.92 loaded
```

```
corrplot(res)
```



Correlation variables and all other variables

Here, we only show one of the condes, as there are a lot of variables and the information conveyed by each is already interpreted in the text accompanying each "condes/catdes" call.

```
condes(df, 1)
```

```
##
## Link between the variable and the continuous variables (R-square)
##
================================================================================
========
##           correlation      p.value
## ca          0.3727176 3.963581e-35
```

```
## trestbps   0.2711214 9.961994e-19
## chol       0.2198225 1.107920e-12
## oldpeak    0.2081367 1.704543e-11
## target    -0.2159057 2.818627e-12
## thalach   -0.3902271 1.273827e-38
##
## Link between the variable and the categorical variable (1-way anova)
## ============================================
##                  R2      p.value
## slope   0.036974393 4.354455e-09
## restecg 0.033848438 2.281014e-08
## cp      0.036010239 3.672018e-08
## fbs     0.014699981 9.970347e-05
## sex     0.010658559 9.325674e-04
## exang   0.007772782 4.733034e-03
##
## Link between variable and the categories of the categorical variables
## ================================================================
##                        Estimate      p.value
## slope=Flat            1.3706123 2.277933e-08
## cp=Typical Angina     1.6445058 3.400069e-06
## fbs=> 120 mg/dL       1.5425998 9.970347e-05
## sex=Female            1.0172461 9.325674e-04
## exang=Yes             0.8459079 4.733034e-03
## restecg=Hypertrophy   3.7494584 1.660865e-02
## cp=Non-anginal Pain  -0.6225968 4.519044e-02
## exang=No             -0.8459079 4.733034e-03
## sex=Male             -1.0172461 9.325674e-04
## fbs=<= 120 mg/dL     -1.5425998 9.970347e-05
## cp=Atypical Angina   -2.8946935 6.007198e-07
## restecg=Normal       -0.3390728 3.650539e-07
## restecg=Abnormality  -3.4103856 1.421567e-08
## slope=Downsloping    -2.1915443 6.132279e-10

# age & sex -> There is a almosts nonexistent dependance between these
two variables, since p-value < 0.05
# age & cp -> There is a almosts nonexistent dependance between these two
variables, since p-value < 0.05
# age & trestbps -> There is slight positive correlation between the age
and trestbps,since p-value < 0.05
# age & chol -> There is slight positive correlation between the age and
chol,since p-value < 0.05
# age & fbs -> There is a slight dependance between these two variables,
since p-value < 0.05
# age & restecg -> There is a almosts nonexistent dependance between
these two variables, since p-value < 0.05
# age & thalach -> There is a near moderate negative correlation between
these variables, since p-value < 0.05
# age & exang -> Exang has no effect on the va  lues of age, p-value >
0.05
```

```
# age & oldpeak -> There is slight positive correlation between the age
and oldpeak ,since p-value < 0.05
# age & slope -> There is a almosts nonexistent dependance between these
two variables, since p-value < 0.05
# age & ca -> There is slight positive correlation between the age and
ca,since p-value < 0.05
# age & thal ->  There is a almosts nonexistent dependance between these
two variables, since p-value < 0.05
# age & target -> There is a slight negative correlation between these
variables, since p-value < 0.05


#Correlation with sex and all other variables
#catdes(df, 2)
# sex & age -> Age has a no effect in the value of sex, since p-value >
0,05
# sex & cp -> 34.93% of Female have Non-anginal Pain, 42.63% of Female
have Typical Angina,
#             and 4.17% are Asymptomatic.
#             8.98% of Male are Asymptomatic, 51.05% of Male have Typical
Angina, and 24.54%
#             of Male have Non-anginal Pain.
# sex & trestbps -> Trestbps has no effect in the value of sex, since p-
value > 0,05
# sex & chol -> Chol has a small to medium effect in the value of sex,
since p-value < 0,05
# sex & fbs -> Fbs is not significant.
# sex & restecg -> 3.53% of Female have Hypertrophy, and the rest are not
significant.
#                 0.56% of Male have Hypertrophy, and the rest are not
significant.
# sex & thalac -> Thalac is not significant.
# sex & exang -> 76.28% of Female have the value of NO, and the other
23.72% have a value of YES.
#                38.01% of Male have a value of YES, and the other 61.99%
have a value of NO.
# sex & oldpeak -> Oldpeak has no effect in the value of sex, since p-
value > 0,05
# sex & slope -> Slope is not significant.
# sex & ca -> Ca has a small effect in the value of sex, since p-value <
0,05
# sex & thal -> 80% of Female have Reversible Defect, 1.28% of Female
have Fixed Defect
#               and 16.99% have no value.
#               50.07% of Male have no value, 8.42% of Male have Fixed
Defect, 40.95% of
#               Male have Reversible Defect.
# sex & target -> Target has a medium effect in the value of sex, since
p-value < 0,05
```

```
# Correlation with cp and all other variable
#catdes(df, 3)
#cp & age -> Age has a small effect in the value of cp, since p-value < 0.05
#cp & sex -> 83.11% of Asymptomatic are Male and the other 16.88% are Female.
#            Sex has no effect on the people who has Atypical Angina.
#            61.62% of Non-anginal Pain are Male and the other 38.38% are Female.
#            73.24% of Typical Angina are Male and the other 26.76% are Female.
#cp & trestbps -> Trestbps has a medium effect in the value of cp, since p-value < 0.05.
#cp & chol -> Chol has no effect in cp's value, since p-value > 0.05.
#cp & fbs -> Fbs has no effect on the people who is Asymtomatic.
#            90.4% of Atypical Angina has fbs<=120 mg/dL and the other 9.58% has fbs>120 mg/dL.
#            19.36% of Non-anginal Pain has fbs > 120 mg/dL and the other 80.63% has fbs<=120 mg/dL.
#            Fbs has no effect on the people who has Typical Angina.
#cp & restecg -> restecg has no effect on the people who is Asymtomatic.
#               61.67% of Atypical Angina has abnormal restecg, 38.32% has normal restecg and the remaining Hypertrophy restecg has no effect.
#               41.19% of Non-anginal Pain has normal restecg,  57.74% has abnormal restecg and the remaining Hypertrophy restecg has no effect.
#               54.32% of Typical Angina has normal restecg, 43.26% has abnormal restecg and the remaining Hypertrophy restecg has no effect.
#cp & thalach -> Thalach has a large effect in the value of cp, since p-value < 0.05
#cp & exang -> 83.11% of Asymptomatic has no exang and the other 16.88% has exang.
#             92.81% of Atypical Angina has no exang and the other 7.18% has exang.
#             86.97% of Non-anginal Pain has no exang and the other 13.02% has exang.
#             56.94% of Typical Angina has exang and the other 43.06% has no exang.
#cp & oldpeak -> Oldpeak has a large effect in the value of cp, since p-value < 0.05
#cp & slope -> slope has no effect on the people who is Asymtomatic.
#             70.66% of Atypical Angina has downsloping slope, 24.55% has flat slope and the remaining Upsloping slope has no effect.
#             55.98% of Non-anginal Pain has downsloping slope, 38.73% has flat slope and the remaining Upsloping slope has no effect.
#             58.75% of Typical Angina has flat slope, 32.79% has downsloping slope and the remaining Upsloping slope has no effect.
#cp & ca -> Ca has a medium effect in the value of cp, since p-value < 0.05
#cp & thal -> thal has no effect on the people who is Asymtomatic.
```

```
#             thal has no effect on the people who has Atypical Angina.
#             95.07% of Non-anginal Pain has reversible defect thal,
3.87% has fixed defect and the remaining normal thal has no effect.
#             85.51% of Typical Angina has reversible defect thal, 13.68%
has fixed defect and the remaining normal thal has no effect.
#cp & target -> Target has a large effect in the value of cp, since p-
value < 0.05

#condes(df, 4)
# trestbps & age -> There is slight positive correlation between trestbps
and age, since p-value < 0.05
# trestbps & sex -> There is an almost nonexistent dependence between
these two variables, since p-value < 0.05
# trestbps & cp -> There is an almost nonexistent dependence between
these two variables, since p-value < 0.05
# trestbps & chol -> There is slight positive correlation between the
trestbps and chol,since p-value < 0.05
# trestbps & fbs -> There is a slight dependence between these two
variables, since p-value < 0.05
# trestbps & restecg -> There is an almost nonexistent dependence between
these two variables, since p-value < 0.05
# trestbps & thalach -> Talach has no effect on the values of trestbps,
p-value > 0.05
# trestbps & exang -> Exang has no effect on the values of trestbps, p-
value > 0.05
# trestbps & oldpeak -> There is slight positive correlation between
trestbps and oldpeak ,since p-value < 0.05
# trestbps & slope -> There is an almost nonexistent dependence between
these two variables, since p-value < 0.05
# trestbps & ca -> There is slight positive correlation between trestbps
and ca,since p-value < 0.05
# trestbps & thal ->  There is an almost nonexistent dependence between
these two variables, since p-value < 0.05
# trestbps & target -> There is a slight negative correlation between
these variables, since p-value < 0.05

# Correlation with chol and all other variable
#condes(df, 5)
#chol & age -> There is a slight positive correlation between the chol
and age,since p-value < 0.05
#chol & sex -> There is a almost nonexistent dependence between these two
variables, since p-value < 0.05
#chol & cp -> Only the category Typical Angina has a positive (or
significant) impact in the mean of chol
#chol & trestbps -> There is a non to a slight positive correlation
between these variables, since p-value < 0.05
#chol & fbs -> Fbs is not significant
#chol & restecg -> There is a almost nonexistent dependence between these
two variables, since p-value < 0.05
#chol & thalach -> Thalach is not significant
```

```
#chol & exang -> There is a almost nonexistent dependence between these
two variables, since p-value < 0.05
#chol & oldpeak -> There is a almost nonexistent correlation between
these two variables, since p-value < 0.05
#chol & slope -> Only the category Flat has a positive (or significant)
impact in the mean of chol
#chol & ca -> There is a almost nonexistent correlation between these two
variables, since p-value < 0.05
#chol & thal -> There is a almost nonexistent dependence between these
two variables, since p-value < 0.05
#chol & target -> There is a non to slight negative correlation between
these two variables, since p-value < 0.05


# Correlation with fbs and all other variable
#catdes(df, 6)
#fbs & age -> Age has a small effect in the value of fbs, since p-value <
0.05.
#fbs & sex -> Sex has no effect on people who has fbs <= 120 mg/dL
neither on fbs > 120 mg/dL.
#fbs & cp -> 17.32% of people who has fbs <= 120 mg/dL has Atypical
Angina cp, the 26.26% has Non-anginal Pain cp, and the people with
Typical Angina and Asymptomatic have no effect on fbs <= 120 mg/dL.
#           35.94% of people who has fbs > 120 mg/dL has Non-anginal Pain
cp, the 10.45% has Atypical Angina cp, and the people with Typical Angina
and Asymptomatic have no effect on fbs > 120 mg/dL.
#fbs & trestbps -> Trestbps has a small effect in the value of fbs, since
p-value < 0.05.
#fbs & chol -> Chol has no effect in values of fbs, since p-value > 0.05.
#fbs & restecg -> 51.83% of people who has fbs <= 120 mg/dL has
Abnormality restecg, the 46.44% has Normal restecg and the remaining
Hypertrophy restecg has no effect on fbs <= 120 mg/dL.
#                 60.13% of people who has fbs > 120 mg/dL has Normal
restecg, the 39.87% has Abnormality restecg and the remaining Hypertrophy
restecg has no effect on fbs > 120 mg/dL.
#fbs & thalach -> Thalach has no effect in values of fbs, since p-value >
0.05.
#fbs & exang -> Exang has no effect in values of fbs, since p-value >
0.05.
#fbs & oldpeak -> Oldpeak has no effect in values of fbs, since p-value >
0.05.
#fbs & slope -> 6.07% of people who has fbs <= 120 mg/dL has Upsloping
slope and the remaining Downsloping slope has no effect on fbs <= 120
mg/dL neither Flat slope.
#               13.72% of people who has fbs > 120 mg/dL has Upsloping
slope and the remaining Downsloping slope has no effect on fbs > 120
mg/dL neither Flat slope.
#fbs & ca -> Ca has a small effect in the value of fbs, since p-value <
0.05.
#fbs & thal -> 91.97% of people with fbs >= 120 mg/dL has Reversible
```

```
Defect thal, the 0.34% has Normal thal and the remaining 7.68% has Fixed
Defect thal.
#             21.57% of people with fbs >= 120 mg/dL has Fixed Defect
thal, the 2.61% has Normal thal and the remaining 75.82% has Reversible
Defect thal.
#fbs & target -> Target has no effect in values of fbs, since p-value >
0.05.

#catdes(df, 7)
#restecg & age ->  Age has a small effect in the value of restecg, since
p-value < 0.05.
#restecg & sex -> Sex has no effect on the people who have Abnormality.
#            73.33% of people with Hypertrophy are Female and the other
26.66% are Male.
#            Sex has no effect on the people considered "Normal".
#restecg & trestbps -> Trestbps has a small effect in the value of
restecg, since p-value < 0.05.
#restecg & chol -> Chol has a small effect in restecg value, since p-
value > 0.05.
#restecg & fbs -> 88.11% of people with Abnormality has fbs<=120 mg/dL
and the other 11.89% has fbs>120 mg/dL
#            Fbs has no effect on people who have Hypertrophy.
#            81.48% of people considered "Normal" has fbs<=120 mg/dL and
the other 18.52% has fbs>120 mg/dL
#restecg & thalach -> Thalach has a small effect in the value of restecg,
since p-value < 0.05.
#restecg & exang -> 70.76% of people with Abnormality have Exang=No and
the other 29.24% Exang=Yes.
#                   Exang has no effect on people who have Hypertrophy.
#                   62.37 of people considered "Normal" have Exang=No and
the other 37.62% Exang=Yes.
#restecg & oldpeak -> Oldpeak has a small effect in the value of restecg,
since p-value < 0.05
#restecg & slope -> 53.60% of people with Abnormality have Downsloping
and the other 40.74% Flat slope.
#                   26.66% of people with Hypertrophy have Upsloping and
the other 73.33% Flat slope.
#                   39.03% of people considered "Normal" have Downsloping
and 52.71% Flat slope.
#restecg & ca -> Ca has a small effect in the value of restecg, since p-
value < 0.05
#restecg & thal -> 7.79% of people with Abnormality have thal=Fixed
Defect and the rest has no effect with Abnormality values.
#                  thal has no effect on people who have Hyperthrophy
#                  thal has no effect on people considered "normal"
#restecg & target -> Target has a small effect in the value of restecg,
since p-value < 0.05


# Correlation with thalac and all other variable
```

```
#condes(df, 8)
#thalac & age -> There is a slight negative correlation between the
thalac and age,since p-value < 0.05
#thalac & sex -> Sex is not significant
#thalac & cp -> There is a almost nonexistent dependence between these
two variables, since p-value < 0.05
#thalac & trestbps -> Trestbps is not significant
#thalac & chol -> Thalach is not significant
#thalac & fbs -> Fbs is not significant
#thalac & restecg -> There is a almost nonexistent dependence between
these two variables, since p-value < 0.05
#thalac & exang -> There is a almost nonexistent dependence between these
two variables, since p-value < 0.05
#thalac & oldpeak -> There is a slight negative correlation between these
two variables, since p-value < 0.05
#thalac & slope -> There is a almost nonexistent dependence between these
two variables, since p-value < 0.05
#thalac & ca -> There is a slight negative correlation between these two
variables, since p-value < 0.05
#thalac & thal -> There is a almost nonexistent dependence between these
two variables, since p-value < 0.05
#thalac & target -> There is a non to slight negative correlation between
these two variables, since p-value < 0.05


# Correlation with exang and all other variable
#catdes(df, 9)
# exang & age -> Age has non to small effect in the value of exang, since
p-value > 0,05
# exang & sex -> Exang=NO:35.00% are Female, and the other 65.00% are
Male.
#                Exang=YES: 78.55% are Male, and the other 21.45% are
Female.
# exang & cp -> Exang=NO: 36.32% have Non-anginal Pain, 22.79% have
Atypical Angina, 9.41% are Asymptomatic, and
#               31.47% have Typical Angina.
#               Exang=YES: 82.03% have Typical Angina, 3.77% are
Asymptomatic, 3.48% have Atypical Angina, and
#               10.72% have Non-anginal Pain
# exang & trestbps -> Trestbps has no effect in the value of exang, since
p-value > 0,05
# exang & chol -> Chol has non to small effect in the value of exang,
since p-value < 0,05
# exang & fbs -> Fbs is not significant.
# exang & restecg -> Exang=NO:53.38% have Abnormality, 45.59% are Normal
and the rest is not significant.
#                    Exang=YES:54.2% are Normal, 43.48% have ABnormality
and the rest is not significant.
# exang & thalac -> Thalac has a large effect in the value of exang,
since p-value < 0,05
```

```
# exang & oldpeak -> Oldpeak has medium to large effect in the value of
exang, since p-value > 0,05
# exang & slope -> Exang=NO: 56.18% have Downsloping, 37.94% have Flat,
and 5.88% have Upsloping
#                  Exang=YES: 64.93% have Flat, 9.9% have Upsloping, and
25.22% have Downsloping
# exang & ca -> Ca has a small effect in the value of exang, since p-
value < 0,05
# exang & thal -> Exang=NO:92.5% have Reversible Defect, 7.05% have Fixed
Defect
#                 and the rest is not significant.
#                 Exang=YES: 15.07% have Fixed Defect, 83.77% have
Reversible Defect and the rest is not significant
# exang & target -> Target has a large effect in the value of exang,
since p-value < 0,05


# Correlation with oldpeak and all other variables
#condes(df, 10)
# oldpeak & age -> There is a slight positive correlation between the
oldpeak and age ,since p-value < 0.05
# oldpeak & sex -> There is almost nonexistent dependence between these
two variables, since p-value < 0.05
# oldpeak & cp -> There is a medium dependence between these two
variables, since p-value < 0.05
# oldpeak & trestbps -> There is slight positive correlation between the
oldpeak and trestbps,since p-value < 0.05
# oldpeak & chol -> There is slight positive correlation between the
oldpeak and chol,since p-value < 0.05
# oldpeak & fbs -> fbs has no effect on the values of oldpeak, since p-
value > 0.05
# oldpeak & restecg -> There is a small dependence between these two
variables, since p-value < 0.05
# oldpeak & thalach -> There is a near moderate negative correlation
between these variables, since p-value < 0.05
# oldpeak & exang -> There is a medium dependence between these two
variables, since p-value < 0.05
# oldpeak & slope -> There is a large dependence between these two
variables, since p-value < 0.05
# oldpeak & ca -> There is a slight positive correlation between the
oldpeak and ca,since p-value < 0.05
# oldpeak & thal ->  There is a medium dependence between these two
variables, since p-value < 0.05
# oldpeak & target -> There is a near moderate negative correlation
between these variables, since p-value < 0.05


# Correlation with slope and all other variable
#catdes(df, 11)
#slope & age -> Age has a small effect in values of slope, since p-value
```

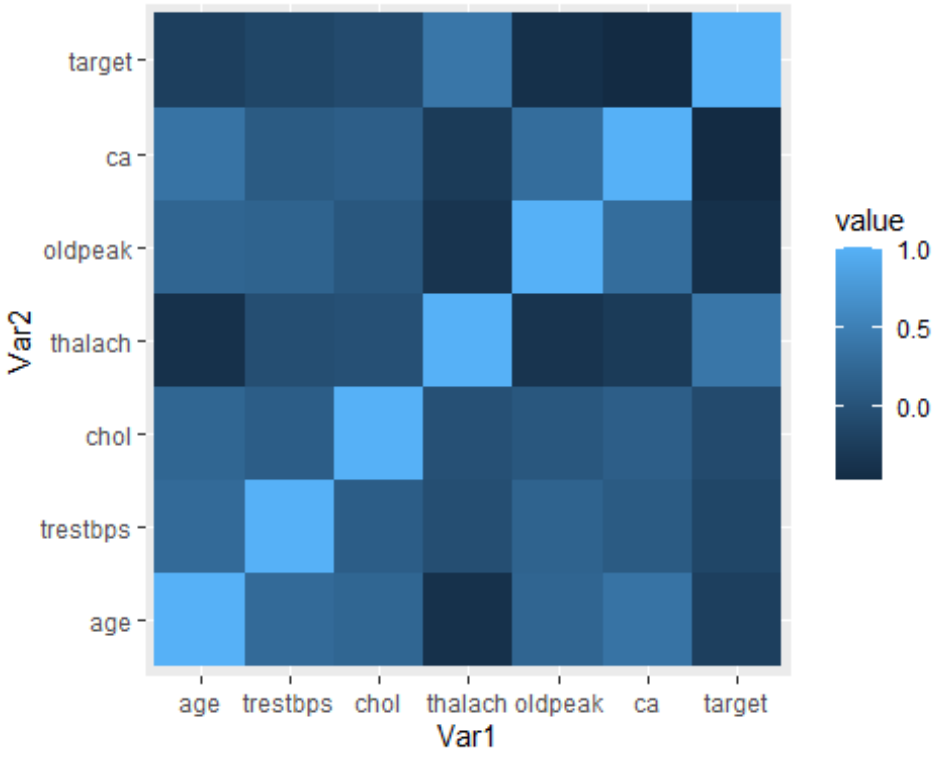
```
< 0.05 and eta2 < 0.06 and eta2 > 0.01.
#slope & sex -> Sex has no effect in values of slope, since p-value >
0.05.
#slope & cp -> 25.16% of Downsloping slope has Atypical Angina cp, 33.90%
has Non-anginal Pain, 34.75% has Typical Angina and the remaining
Asymptomatic values have no effect in Downsloping slope.
#             8.51% of Flat slope has Atypical Angina cp, 22.82% has Non-
anginal Pain, 60.58% has Typical Angina and the remaining Asymptomatic
values have no effect in Flat slope.
#             Cp has no effect in values of Upsloping slope.
#slope & trestbps -> Trestbps has a small effect in values of slope,
since p-value < 0.05 and eta2 < 0.06.
#slope & chol -> Chol has no effect in values of slope
#slope & fbs -> fbs has no effect in values of Downsloping slope.
#               fbs has no effect in values of Flat slope.
#               28.38% of Upsloping slope values have fbs > 120 mg/dL and
the remaining 71.62% have fbs <= 120 mg/dL.
#slope & restecg -> 58.64% of Downsloping slope values have Abnormality
restecg, 41.36% have Normal restecg and 0% have Hypertrophy restecg.
#                   54.36% of Flat slope values have Normal restecg,
43.36% have Abnormality restecg and 2.28% have Hypertrophy restecg.
#                   5.40% of Uplsoping values have Hypertrophy restecg,
and Normal restecg neither Abnormality restecg have effect in values of
Upsloping slope.
#slope & thalach -> thalach has a large effect in values of slope, since
p-value < 0.05 and eta2 > 0.14.
#slope & exang -> 81.45% of Downsloping slope values have no exang and
18.55% have exang.
#                 53.53% of Flat slope values have no exang and 46.47%
have exang.
#                 54.05% of Upsloping slope values have no exang and
45.94% have exang.
#slope & oldpeak -> oldpeak has a large effect in values of slope, since
p-value < 0.05 and eta2 > 0.14.
#slope & ca -> Ca has a small effect in values of slope, since p-value <
0.05 and eta2 < 0.06.
#slope & thal -> 97.87% of Downsloping slope values have Reversible
Defect thal, 1.49% have Fixed Defect thal and Normal thal has no effect
in Downsloping.
#               Thal has no effect in Flat slope values.
#               36.48% of Upsloping slope values have Reversible Defect
thal, 63.51% have Fixed Defect thal and Normal thal has no effect in
Upsloping.
#slope & target -> Target has a medium effect in values of slope, since
p-value < 0.05 and eta2 < 0.14 and eta2 > 0.06.


# Correlation with ca and all other variables
#condes(df, 12)
# ca & age -> There is a slight positive correlation between the ca and
```

```
age ,since p-value < 0.05
# ca & sex -> There is a small dependence between these two variables,
since p-value < 0.05
# ca & cp -> There is a small dependence between these two variables,
since p-value < 0.05
# ca & trestbps -> There is slight positive correlation between the ca
and trestbps,since p-value < 0.05
# ca & chol -> There is slight positive correlation between the ca and
chol,since p-value < 0.05
# ca & fbs -> There is a small dependence between these two variables,
since p-value < 0.05
# ca & restecg -> There is a small dependence between these two
variables, since p-value < 0.05
# ca & thalach -> There is a slight negative correlation between these
variables, since p-value < 0.05
# ca & exang -> There is a small dependence between these two variables,
since p-value < 0.05
# ca & oldpeak -> There is a slight positive correlation between the ca
and oldpeak,since p-value < 0.05
# ca & slope ->There is a small dependence between these two variables,
since p-value < 0.05
# ca & thal ->  There is a small dependence between these two variables,
since p-value < 0.05
# ca & target -> There is a near moderate negative correlation
correlation between these variables, since p-value < 0.05

#catdes(df, 13)
#thal & age ->  Age has no effect in the value of thal, since p-value >
0.05.
#thal & sex -> 89% of people with thal=Fixed Defect are Male and the
other 11% are Female.
#              Sex has no effect on the people with thal=Normal.
#              32.46% of people with thal=Reversible Defect are Female
and the other 67.54% are Male.
#thal & trestbps -> Trestbps has a small effect in the value of thal,
since p-value < 0.05.
#thal & chol -> Chol has a small effect in thal value, since p-value >
0.05.
#thal & fbs -> 33% of people with thal=Fixed Defect has fbs>=120 mg/dL
and the other 67% has fbs>120 mg/dL
#              57.14% of people with thal=Normal has fbs>=120 mg/dL and
the other 42.85% has fbs>120 mg/dL
#              12.63% of people with thal=Reversible Defect has fbs<=120
mg/dL and the other 87.36% has fbs>120 mg/dL
#thal & thalach -> Thalach has a small effect in the value of thal, since
p-value < 0.05.
#thal & exang -> 48% of people with thal=fixed defect have Exang=No and
the other 52% Exang=Yes.
#                Exang has no effect on people with thal=Normal
#                68.51% of people with thal=Reversible Defect have
```

```
Exang=No and the other 31.48% Exang=Yes.
#thal & oldpeak -> Oldpeak has a medium effect in the value of thal,
since p-value < 0.05
#thal & slope -> 7% of people with thal=Fixed Defect have Downsloping and
the other 47% Flat slope.
#                 slope has no effect on people with thal=Normal
#                 50% of people with thal=Reversible Defect have
Downsloping and 2.94% Flat slope.
#thal & ca -> ca has no effect in the value of thal, since p-value > 0.05
#thal & restecg -> 40% of people with thal=Fixed Defect have Abnormality
#                   And the rest has no effect with thal values.
#thal & target -> Target has a small effect in the value of thal, since
p-value < 0.05
```

Heat Map correlacions

```
df2 <- Filter(is.numeric, df)
df2 <- scale(df2)
cormat <- round(cor(df2),2)
library(reshape2)

## Warning: package 'reshape2' was built under R version 4.3.3

melted_cormat <- melt(cormat)

library(ggplot2)
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()
```



Profiling

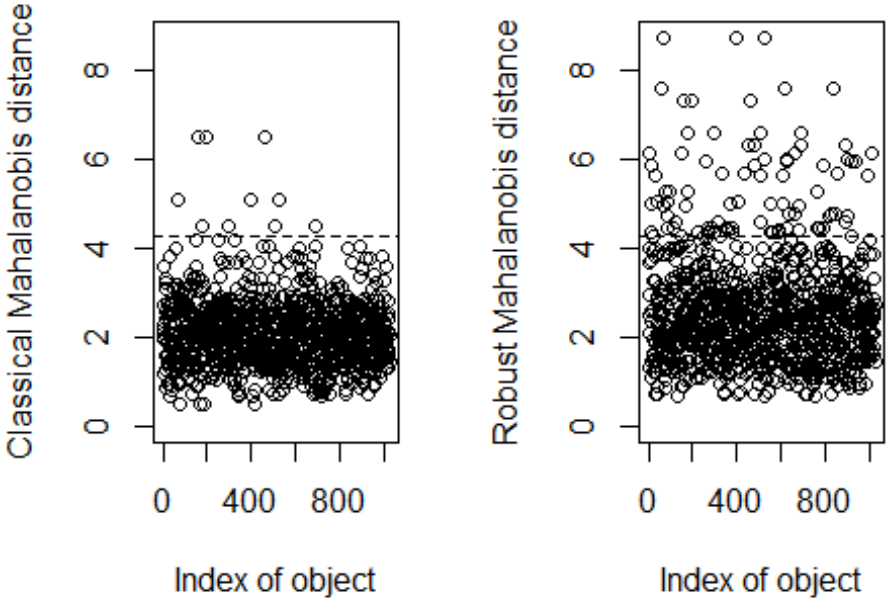
We now proceed to do the profiling. As we have a numeric target "target" and numeric and categorical explicative variables, we'll use the condes tool that provides us with information about the relationships between the indicated variables and the target.

```
# Continuous output:
library(FactoMineR)
summary(df$target)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1000  0.2600  0.7100  0.5364  0.8100  0.9000

res.condes <- condes(df, 14, proba = 0.50)
res.condes$quanti  # Global association to numeric variables

##          correlation      p.value
## thalach   0.39937441 1.564500e-40
## chol     -0.09378816 2.650143e-03
## trestbps -0.13566217 1.311535e-05
## age      -0.21590573 2.818627e-12
## oldpeak  -0.39601538 7.998564e-40
## ca       -0.45504567 1.573337e-53

res.condes$quali # Partial association of numeric variables to levels of
outcome factor

##                   R2      p.value
## cp       0.253890752 1.499596e-64
## exang    0.170849731 1.443650e-43
## slope    0.132703531 2.532984e-32
## sex      0.076369394 2.008368e-19
## restecg  0.035701748 8.550961e-09
## thal     0.022665084 8.169298e-06
## fbs      0.001373824 2.357714e-01

res.condes$category  # Partial association to significative levels in
factors

##                              Estimate      p.value
## exang=No                  0.124945652 1.443650e-43
## slope=Downsloping         0.124020359 7.009225e-33
## cp=Non-anginal Pain       0.084617565 4.199338e-24
## sex=Female                0.085785841 2.008368e-19
## cp=Atypical Angina        0.102081089 1.001772e-15
## restecg=Abnormality       0.098801024 2.329818e-09
## thal=Reversible Defect    0.058303056 3.008745e-06
## cp=Asymptomatic           0.018271306 1.491150e-02
## fbs=<= 120 mg/dL          0.014857288 2.357714e-01
## fbs=> 120 mg/dL          -0.014857288 2.357714e-01
```

```
## slope=Upsloping         -0.033533954 1.569582e-01
## restecg=Hypertrophy     -0.094438742 5.527308e-02
## thal=Fixed Defect       -0.100275222 1.303146e-06
## restecg=Normal          -0.004362283 3.705541e-08
## sex=Male                -0.085785841 2.008368e-19
## slope=Flat              -0.090486405 1.106168e-28
## exang=Yes               -0.124945652 1.443650e-43
## cp=Typical Angina       -0.204969960 9.699236e-66
```

Multivariate outliers

```
library(chemometrics)

## Warning: package 'chemometrics' was built under R version 4.3.3

## Loading required package: rpart

summary(df[,vars_con])

##       age           trestbps          chol        thalach
oldpeak
##  Min.   :29.00   Min.   : 94.0   Min.   :126   Min.   : 71.0   Min.
:0.000
##  1st Qu.:48.00   1st Qu.:120.0   1st Qu.:211   1st Qu.:132.0   1st
Qu.:0.000
##  Median :56.00   Median :130.0   Median :240   Median :152.0   Median
:0.800
##  Mean   :54.43   Mean   :131.6   Mean   :246   Mean   :149.1   Mean
:1.072
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:275   3rd Qu.:166.0   3rd
Qu.:1.800
##  Max.   :77.00   Max.   :200.0   Max.   :564   Max.   :202.0   Max.
:6.200
##        ca             target
##  Min.   :0.0000   Min.   :0.1000
##  1st Qu.:0.0000   1st Qu.:0.2600
##  Median :0.0000   Median :0.7100
##  Mean   :0.6878   Mean   :0.5364
##  3rd Qu.:1.0000   3rd Qu.:0.8100
##  Max.   :3.0000   Max.   :0.9000

mout<-Moutlier(df[,vars_con[1:5]],quantile = 0.9975, plot = TRUE)
```



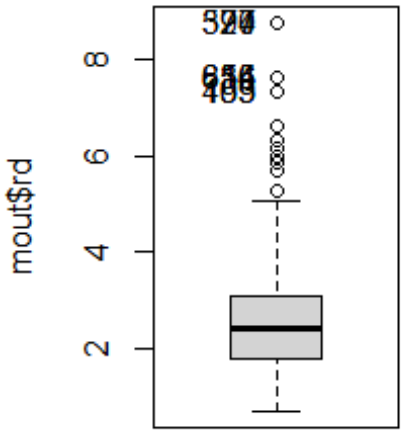

```
# Classical: Assumption of normality on the underlying generating
mechanism
# Robust: Median and absolute median deviations -> Not normal generating
mechanism

length(which(mout$rd>mout$cutoff))

## [1] 93

ll<-which(mout$rd>mout$cutoff)
Boxplot(mout$rd)

##  [1]  70 394 527  55  56 614 834 159 193 465

df[ll,c(vars_con)]

##      age trestbps chol thalach oldpeak ca target
## 7     58      114  318     140     4.4  3   0.37
## 10    54      122  286     116     3.2  2   0.23
## 12    43      132  341     136     3.0  0   0.21
## 14    51      140  298     122     4.2  3   0.37
## 23    45      104  208     148     3.0  0   0.77
## 30    55      180  327     117     3.4  0   0.26
## 36    46      150  231     147     3.6  0   0.39
## 55    55      140  217     111     5.6  0   0.22
## 56    55      140  217     111     5.6  0   0.35
## 70    62      160  164     145     6.2  3   0.32
## 71    59      170  326     140     3.4  0   0.32
```

```
## 78    63      140  187     144     4.0  2   0.28
## 83    46      150  231     147     3.6  0   0.37
## 89    62      140  268     160     3.6  2   0.35
## 90    68      144  193     141     3.4  2   0.28
## 93    63      140  187     144     4.0  2   0.22
## 114   57      110  335     143     3.0  1   0.14
## 125   61      120  260     140     3.6  1   0.31
## 151   58      114  318     140     4.4  3   0.12
## 152   54      192  283     195     0.0  1   0.15
## 159   67      115  564     160     1.6  0   0.83
## 166   59      170  326     140     3.4  0   0.11
## 176   56      200  288     133     4.0  2   0.12
## 181   63      140  187     144     4.0  2   0.26
## 193   67      115  564     160     1.6  0   0.89
## 220   54      122  286     116     3.2  2   0.26
## 247   54      192  283     195     0.0  1   0.20
## 259   38      120  231     182     3.8  0   0.30
## 266   45      104  208     148     3.0  0   0.77
## 268   67      120  237      71     1.0  0   0.13
## 269   58      132  224     173     3.2  2   0.20
## 285   58      132  224     173     3.2  2   0.15
## 295   56      200  288     133     4.0  2   0.23
## 297   67      120  237      71     1.0  0   0.33
## 311   61      120  260     140     3.6  1   0.21
## 327   54      192  283     195     0.0  1   0.33
## 333   37      130  250     187     3.5  0   0.89
## 353   57      110  335     143     3.0  1   0.21
## 357   59      164  176      90     1.0  2   0.19
## 371   43      132  341     136     3.0  0   0.34
## 379   67      120  237      71     1.0  0   0.14
## 382   58      132  224     173     3.2  2   0.14
## 394   62      160  164     145     6.2  3   0.22
## 410   46      150  231     147     3.6  0   0.28
## 434   37      130  250     187     3.5  0   0.70
## 451   63      150  407     154     4.0  3   0.23
## 465   67      115  564     160     1.6  0   0.86
## 482   63      150  407     154     4.0  3   0.39
## 483   51      140  298     122     4.2  3   0.26
## 507   61      120  260     140     3.6  1   0.18
## 509   56      200  288     133     4.0  2   0.36
## 510   55      180  327     117     3.4  0   0.36
## 527   62      160  164     145     6.2  3   0.26
## 529   59      178  270     145     4.2  0   0.85
## 552   54      122  286     116     3.2  2   0.23
## 553   43      132  341     136     3.0  0   0.38
## 560   67      120  237      71     1.0  0   0.37
## 588   59      164  176      90     1.0  2   0.16
## 590   54      122  286     116     3.2  2   0.26
## 610   55      180  327     117     3.4  0   0.39
## 611   43      132  341     136     3.0  0   0.30
```

```
## 614   55      140  217     111     5.6  0   0.33
## 625   59      178  270     145     4.2  0   0.79
## 627   58      132  224     173     3.2  2   0.16
## 628   38      120  231     182     3.8  0   0.22
## 634   61      138  166     125     3.6  1   0.18
## 661   61      138  166     125     3.6  1   0.11
## 662   58      114  318     140     4.4  3   0.36
## 682   59      170  326     140     3.4  0   0.39
## 683   59      164  176      90     1.0  2   0.16
## 686   63      150  407     154     4.0  3   0.13
## 689   56      200  288     133     4.0  2   0.27
## 765   63      140  187     144     4.0  2   0.39
## 766   57      110  335     143     3.0  1   0.35
## 767   57      110  335     143     3.0  1   0.27
## 768   68      144  193     141     3.4  2   0.28
## 788   51      140  298     122     4.2  3   0.14
## 793   68      144  193     141     3.4  2   0.33
## 813   62      140  268     160     3.6  2   0.25
## 822   62      140  268     160     3.6  2   0.39
## 825   61      138  166     125     3.6  1   0.27
## 834   55      140  217     111     5.6  0   0.40
## 848   61      138  166     125     3.6  1   0.28
## 852   37      130  250     187     3.5  0   0.80
## 887   61      120  260     140     3.6  1   0.33
## 890   63      150  407     154     4.0  3   0.19
## 897   59      178  270     145     4.2  0   0.80
## 903   62      140  268     160     3.6  2   0.35
## 914   45      104  208     148     3.0  0   0.83
## 920   38      120  231     182     3.8  0   0.21
## 934   38      120  231     182     3.8  0   0.28
## 987   55      180  327     117     3.4  0   0.20
## 1014  58      114  318     140     4.4  3   0.36
```

```
df$mout <- 0
df$mout[ ll ]<-1
df$mout <- factor( df$mout, labels=c( "NoMOut","YesMOut")) #We identify
the Mildoutliers if the row has the value "YesMOut" in the mout column.
table(df$mout)
```

```
##
##  NoMOut YesMOut
##     932      93
```


mout$rd
2
4
6
8