# Analysis-COPD

Lina

2024-05-29

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#################################################################
# Housekeeping Use for All Analyses #
#################################################################
date() # Current system time and date.
```

```
## [1] "Fri May 31 13:03:25 2024"
```

```
Sys.time() # Current system time and date (redundant).
```

```
## [1] "2024-05-31 13:03:25 +07"
```

```
R.version.string # R version and version release date.
```

```
## [1] "R version 4.2.3 (2023-03-15 ucrt)"
```

```
options(digits=6) # Confirm default digits.
options(scipen=999)# Suppress scientific notation.
options(width=60) # Confirm output width.
ls() # List all objects in the working # directory.
```

```
## character(0)
```

```
rm(list = ls()) # CAUTION: Remove all files in the #working directory. If this action is not desired, u
ls.str() # List all objects with finite detail.
getwd() # Identify the current working directory
```

```
## [1] "C:/Users/linan/Documents/GitHub/project/R-project/Regression-R"
```

```r
setwd("C:/Users/linan/Documents/GitHub/project/R-project/Regression-R") # Set to a new working directory
getwd()# Confirm the working directory.
```

```
## [1] "C:/Users/linan/Documents/GitHub/project/R-project/Regression-R"
```

```r
list.files()# List files at the PC directory
```

```
## [1] "COPD_student_dataset.csv"  "linear-regression.nb.html"
## [3] "linear-regression.Rmd"     "variable-exploration.pdf"
## [5] "variable-exploration.Rmd"
```

```r
.libPaths()# Library pathname
```

```
## [1] "C:/Users/linan/AppData/Local/R/win-library/4.2"
## [2] "C:/Program Files/R/R-4.2.3/library"
```

```r
.Library# Library pathname.
```

```
## [1] "C:/PROGRA~1/R/R-42~1.3/library"
```

```r
sessionInfo() # R version, locale, and packages.
```

```
## R version 4.2.3 (2023-03-15 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22631)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_Indonesia.utf8
## [2] LC_CTYPE=English_Indonesia.utf8
## [3] LC_MONETARY=English_Indonesia.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_Indonesia.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets
## [6] methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.2.3    fastmap_1.1.1     cli_3.6.1
##  [4] tools_4.2.3       htmltools_0.5.8   rstudioapi_0.16.0
##  [7] yaml_2.3.8        rmarkdown_2.26    knitr_1.45
## [10] xfun_0.40         digest_0.6.31     rlang_1.1.1
## [13] evaluate_0.23
```

```r
search()# Attached packages and objects.
```

```
## [1] ".GlobalEnv"        "package:stats"
## [3] "package:graphics"  "package:grDevices"
## [5] "package:utils"     "package:datasets"
## [7] "package:methods"   "Autoloads"
## [9] "package:base"
```

```
searchpaths() # Attached packages and objects.
```

```
## [1] ".GlobalEnv"
## [2] "C:/Program Files/R/R-4.2.3/library/stats"
## [3] "C:/Program Files/R/R-4.2.3/library/graphics"
## [4] "C:/Program Files/R/R-4.2.3/library/grDevices"
## [5] "C:/Program Files/R/R-4.2.3/library/utils"
## [6] "C:/Program Files/R/R-4.2.3/library/datasets"
## [7] "C:/Program Files/R/R-4.2.3/library/methods"
## [8] "Autoloads"
## [9] "C:/PROGRA~1/R/R-42~1.3/library/base"
```

```
######################################################################
```

```
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(Hmisc)
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(gmodels)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ---- tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5

## -- Conflicts ---------------------- tidyverse_conflicts() --
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x Hmisc::src()       masks dplyr::src()
## x Hmisc::summarize() masks dplyr::summarize()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
COPD.df <- read.table(file="COPD_student_dataset.csv", header=TRUE, dec=".", sep = ",")
```

```
str(COPD.df)
```

```
## 'data.frame':    101 obs. of  24 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID          : int  58 57 62 145 136 84 93 27 114 152 ...
##  $ AGE         : int  77 79 80 56 65 67 67 83 72 75 ...
##  $ PackHistory : num  60 50 11 60 68 26 50 90 50 6 ...
##  $ COPDSEVERITY: chr  "SEVERE" "MODERATE" "MODERATE" "VERY SEVERE" ...
##  $ MWT1        : int  120 165 201 210 204 216 214 214 231 226 ...
##  $ MWT2        : int  120 176 180 210 210 180 237 237 237 240 ...
##  $ MWT1Best    : int  120 176 201 210 210 216 237 237 237 240 ...
##  $ FEV1        : num  1.21 1.09 1.52 0.47 1.07 1.09 0.69 0.68 2.13 1.06 ...
##  $ FEV1PRED    : num  36 56 68 14 42 50 35 32 63 46 ...
##  $ FVC         : num  2.4 1.64 2.3 1.14 2.91 1.99 1.31 2.23 4.38 2.06 ...
##  $ FVCPRED     : int  98 65 86 27 98 60 48 77 80 75 ...
##  $ CAT         : int  25 12 22 28 32 29 29 22 25 31 ...
##  $ HAD         : num  8 21 18 26 18 21 30 2 6 20 ...
##  $ SGRQ        : num  69.5 44.2 44.1 62 75.6 ...
##  $ AGEquartiles: int  4 4 4 1 1 2 2 4 3 3 ...
##  $ copd        : int  3 2 2 4 3 2 3 3 2 3 ...
##  $ gender      : int  1 0 0 1 1 0 0 1 1 0 ...
##  $ smoking     : int  2 2 2 2 2 1 1 2 1 2 ...
##  $ Diabetes    : int  1 1 1 0 0 1 1 1 1 0 ...
##  $ muscular    : int  0 0 0 0 1 0 0 0 0 1 ...
##  $ hypertension: int  0 0 0 1 1 0 0 0 0 0 ...
##  $ AtrialFib   : int  1 1 1 1 0 1 1 1 1 0 ...
##  $ IHD         : int  0 1 0 0 0 0 0 0 0 0 ...
```

## Variables

Characters : Age, Gender, Pack History, Smoking Disease : COPDSeverity, CAT Walking ability : MWT1, MWT2, MWT1Best Lung function : FEV1, FEV1PRED, FVC, FVCPRED Anxiety&Depression : HAD QOL : SGRQ Comorbidities : Diabetes, Muscular, Hypertension, AtrialFib, IHD

numeric : Age, PackHistory, FEV, FEV1PRED, FVC, FVCPRED, CAT, HAD, MWT1, MWT2, MWT1Best, SGRQ factor : Gender, COPDseverity, copd, smoking, Diabetes, Muscular, Hypertension, AtrialFib, IHD

Change variable type :

```
COPD.df$AGE <- as.numeric(COPD.df$AGE)
COPD.df$MWT1 <- as.numeric(COPD.df$MWT1)
COPD.df$MWT2 <- as.numeric(COPD.df$MWT2)
```

```
COPD.df$MWT1Best<-as.numeric(COPD.df$MWT1Best)
COPD.df$FEV1PRED <- as.numeric(COPD.df$FEV1PRED)
COPD.df$FVCPRED <- as.numeric(COPD.df$FVCPRED)
COPD.df$CAT <- as.numeric(COPD.df$CAT)
```

Categorical

```
COPD.df$AGEquartiles <- as.factor(COPD.df$AGEquartiles)
COPD.df$copd <- as.factor(COPD.df$copd)
COPD.df$gender <- as.factor(COPD.df$gender)
COPD.df$Diabetes <- as.factor(COPD.df$Diabetes)
COPD.df$smoking <- as.factor(COPD.df$smoking)
COPD.df$muscular <- as.factor(COPD.df$muscular)
COPD.df$hypertension <- as.factor(COPD.df$hypertension)
COPD.df$AtrialFib <- as.factor(COPD.df$AtrialFib)
COPD.df$IHD <- as.factor(COPD.df$IHD)
```

```
str(COPD.df)
```

```
## 'data.frame':    101 obs. of  24 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID          : int  58 57 62 145 136 84 93 27 114 152 ...
##  $ AGE         : num  77 79 80 56 65 67 67 83 72 75 ...
##  $ PackHistory : num  60 50 11 60 68 26 50 90 50 6 ...
##  $ COPDSEVERITY: chr  "SEVERE" "MODERATE" "MODERATE" "VERY SEVERE" ...
##  $ MWT1        : num  120 165 201 210 204 216 214 214 231 226 ...
##  $ MWT2        : num  120 176 180 210 210 180 237 237 237 240 ...
##  $ MWT1Best    : num  120 176 201 210 210 216 237 237 237 240 ...
##  $ FEV1        : num  1.21 1.09 1.52 0.47 1.07 1.09 0.69 0.68 2.13 1.06 ...
##  $ FEV1PRED    : num  36 56 68 14 42 50 35 32 63 46 ...
##  $ FVC         : num  2.4 1.64 2.3 1.14 2.91 1.99 1.31 2.23 4.38 2.06 ...
##  $ FVCPRED     : num  98 65 86 27 98 60 48 77 80 75 ...
##  $ CAT         : num  25 12 22 28 32 29 29 22 25 31 ...
##  $ HAD         : num  8 21 18 26 18 21 30 2 6 20 ...
##  $ SGRQ        : num  69.5 44.2 44.1 62 75.6 ...
##  $ AGEquartiles: Factor w/ 4 levels "1","2","3","4": 4 4 4 1 1 2 2 4 3 3 ...
##  $ copd        : Factor w/ 4 levels "1","2","3","4": 3 2 2 4 3 2 3 3 2 3 ...
##  $ gender      : Factor w/ 2 levels "0","1": 2 1 1 2 2 1 1 2 2 1 ...
##  $ smoking     : Factor w/ 2 levels "1","2": 2 2 2 2 2 1 1 2 1 2 ...
##  $ Diabetes    : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 2 2 2 1 ...
##  $ muscular    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 2 ...
##  $ hypertension: Factor w/ 2 levels "0","1": 1 1 1 2 2 1 1 1 1 1 ...
##  $ AtrialFib   : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 2 2 2 1 ...
##  $ IHD         : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
```

```
describe(COPD.df)
```

```
## COPD.df
```

```
##
##  24  Variables      101  Observations
## ---------------------------------------------------------
## X
##        n  missing distinct     Info     Mean      Gmd
##      101        0      101        1       51       34
##      .05      .10      .25      .50      .75      .90
##        6       11       26       51       76       91
##      .95
##       96
##
## lowest :   1   2   3   4   5, highest:  97  98  99 100 101
## ---------------------------------------------------------
## ID
##        n  missing distinct     Info     Mean      Gmd
##      101        0       97        1    91.41    59.56
##      .05      .10      .25      .50      .75      .90
##       10       18       49       87      143      159
##      .95
##      164
##
## lowest :   1   2   3   6   8, highest: 165 166 167 168 169
## ---------------------------------------------------------
## AGE
##        n  missing distinct     Info     Mean      Gmd
##      101        0       33    0.998     70.1     8.73
##      .05      .10      .25      .50      .75      .90
##       55       60       65       71       75       79
##      .95
##       81
##
## lowest : 44 49 52 53 54, highest: 80 81 82 83 88
## ---------------------------------------------------------
## PackHistory
##        n  missing distinct     Info     Mean      Gmd
##      101        0       48    0.998     39.7    27.35
##      .05      .10      .25      .50      .75      .90
##        6       10       20       36       54       75
##      .95
##       90
##
## lowest :   1   3   5   6   8, highest:  90 100 103 105 109
## ---------------------------------------------------------
## COPDSEVERITY
##        n  missing distinct
##      101        0        4
##
## Value             MILD    MODERATE      SEVERE VERY SEVERE
## Frequency           23          43          27           8
## Proportion       0.228       0.426       0.267       0.079
## ---------------------------------------------------------
## MWT1
##        n  missing distinct     Info     Mean      Gmd
##       99        2       69        1    385.9    117.6
```

```
##      .05      .10      .25      .50      .75      .90
##    212.7    226.0    300.0    419.0    460.5    495.2
##      .95
##    510.1
##
## lowest : 120 165 201 204 210, highest: 511 522 558 576 688
## ------------------------------------------------------------
## MWT2
##        n  missing distinct     Info     Mean      Gmd
##      100        1       72        1    390.3    121.7
##      .05      .10      .25      .50      .75      .90
##    210.0    237.0    303.8    399.0    459.0    518.7
##      .95
##    541.1
##
## lowest : 120 176 180 210 230, highest: 563 575 577 582 699
## ------------------------------------------------------------
## MWT1Best
##        n  missing distinct     Info     Mean      Gmd
##      100        1       71        1    399.1    119.7
##      .05      .10      .25      .50      .75      .90
##    215.7    240.0    303.8    420.0    465.2    518.7
##      .95
##    540.9
##
## lowest : 120 176 201 210 216, highest: 558 575 577 582 699
## ------------------------------------------------------------
## FEV1
##        n  missing distinct     Info     Mean      Gmd
##      101        0       85        1    1.604   0.7645
##      .05      .10      .25      .50      .75      .90
##     0.68     0.73     1.10     1.60     1.96     2.70
##      .95
##     2.90
##
## lowest : 0.45 0.47 0.51 0.6  0.65, highest: 2.93 2.97 3.02 3.06 3.18
## ------------------------------------------------------------
## FEV1PRED
##        n  missing distinct     Info     Mean      Gmd
##      101        0       51    0.999    58.53    25.56
##      .05      .10      .25      .50      .75      .90
##       24       30       42       60       75       90
##      .95
##       93
##
## lowest : 3.29 3.39 14   17   24  , highest: 92   93   95   98   102
## ------------------------------------------------------------
## FVC
##        n  missing distinct     Info     Mean      Gmd
##      101        0       80        1    2.955    1.108
##      .05      .10      .25      .50      .75      .90
##     1.56     1.89     2.27     2.77     3.63     4.39
##      .95
##     4.70
```

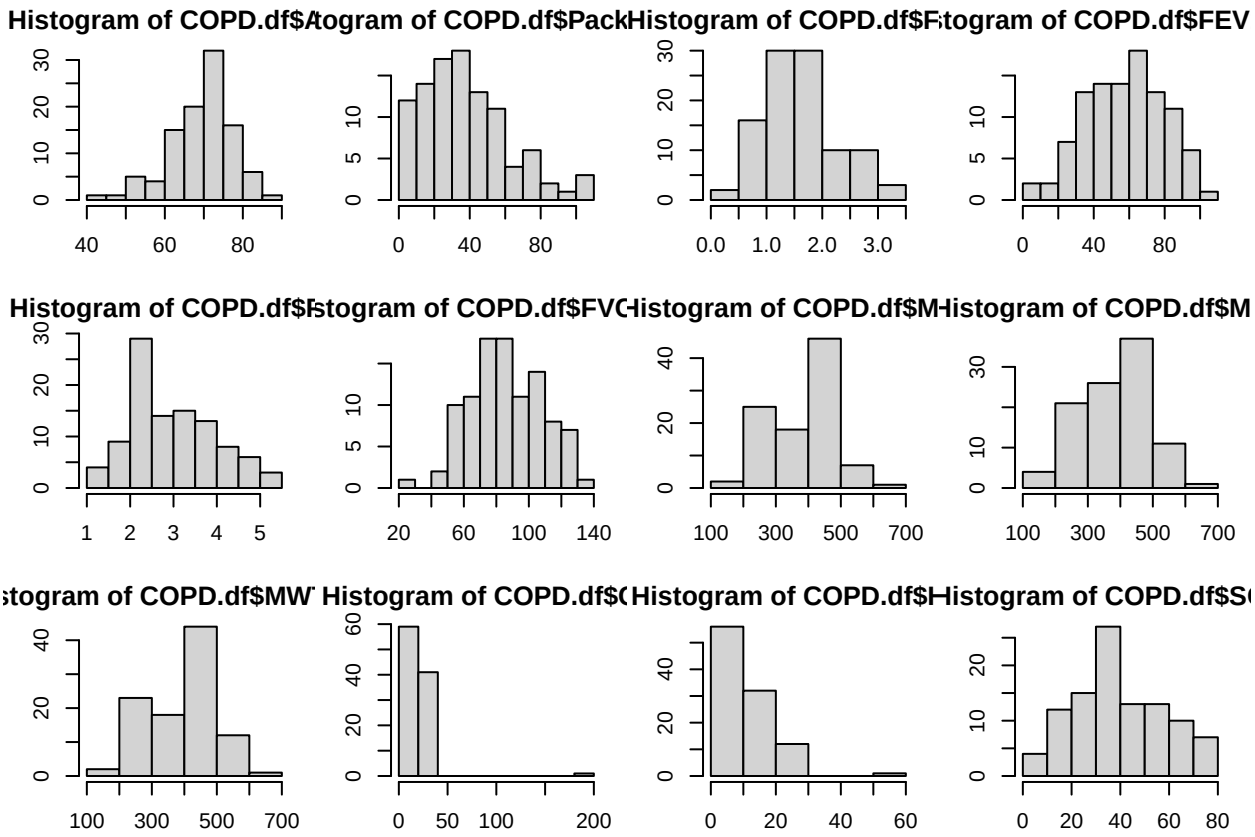
```
##
## lowest : 1.14 1.31 1.47 1.52 1.56, highest: 4.72 4.9  5.15 5.23 5.37
## ------------------------------------------------------------
## FVCPRED
##        n  missing distinct     Info     Mean      Gmd
##      101        0       57    0.999    86.44    24.92
##      .05      .10      .25      .50      .75      .90
##       53       60       71       84      103      118
##      .95
##      122
##
## lowest :  27  45  48  51  53, highest: 121 122 123 125 132
## ------------------------------------------------------------
## CAT
##        n  missing distinct     Info     Mean      Gmd
##      101        0       30    0.997    19.34    12.28
##      .05      .10      .25      .50      .75      .90
##        5        5       12       18       24       29
##      .95
##       30
##
## lowest :   3   4   5   6   7, highest:  29  30  31  32 188
## ------------------------------------------------------------
## HAD
##        n  missing distinct     Info     Mean      Gmd
##      101        0       28    0.997    11.18    8.984
##      .05      .10      .25      .50      .75      .90
##        1        2        6       10       15       22
##      .95
##       26
##
## lowest : 0    1    2    3    4   , highest: 23   26   29   30   56.2
## ------------------------------------------------------------
## SGRQ
##        n  missing distinct     Info     Mean      Gmd
##      101        0       89        1    40.19    20.88
##      .05      .10      .25      .50      .75      .90
##    10.92    16.29    28.41    38.21    55.23    67.56
##      .95
##    72.24
##
## lowest : 2     8.12  8.25  10.01 10.92
## highest: 72.56 73.82 75.56 76.5  77.44
## ------------------------------------------------------------
## AGEquartiles
##        n  missing distinct
##      101        0        4
##
## Value          1     2     3     4
## Frequency     26    24    28    23
## Proportion 0.257 0.238 0.277 0.228
## ------------------------------------------------------------
## copd
##        n  missing distinct
```

```
##      101        0        4
##
## Value          1     2     3     4
## Frequency     23    43    27     8
## Proportion 0.228 0.426 0.267 0.079
## ------------------------------------------------------------
## gender
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     36    65
## Proportion 0.356 0.644
## ------------------------------------------------------------
## smoking
##        n  missing distinct
##      101        0        2
##
## Value          1     2
## Frequency     16    85
## Proportion 0.158 0.842
## ------------------------------------------------------------
## Diabetes
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     80    21
## Proportion 0.792 0.208
## ------------------------------------------------------------
## muscular
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     82    19
## Proportion 0.812 0.188
## ------------------------------------------------------------
## hypertension
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     89    12
## Proportion 0.881 0.119
## ------------------------------------------------------------
## AtrialFib
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     81    20
## Proportion 0.802 0.198
## ------------------------------------------------------------
```

```
## IHD
##        n  missing distinct
##      101        0        2
##
## Value          0     1
## Frequency     92     9
## Proportion 0.911 0.089
## ------------------------------------------------------------
```

Histogram for numeric data Age, PackHistory, FEV, FEV1PRED, FVC, FVCPRED, CAT, HAD, MWT1, MWT2, MWT1Best, SGRQ

```
par(ask=TRUE)
par(mfrow=c(3,4))
par(mar = c(3, 3, 2, 1))
hist(COPD.df$AGE)
hist(COPD.df$PackHistory)
hist(COPD.df$FEV1)
hist(COPD.df$FEV1PRED)
hist(COPD.df$FVC)
hist(COPD.df$FVCPRED)
hist(COPD.df$MWT1)
hist(COPD.df$MWT2)
hist(COPD.df$MWT1Best)
hist(COPD.df$CAT)
hist(COPD.df$HAD)
hist(COPD.df$SGRQ)
```

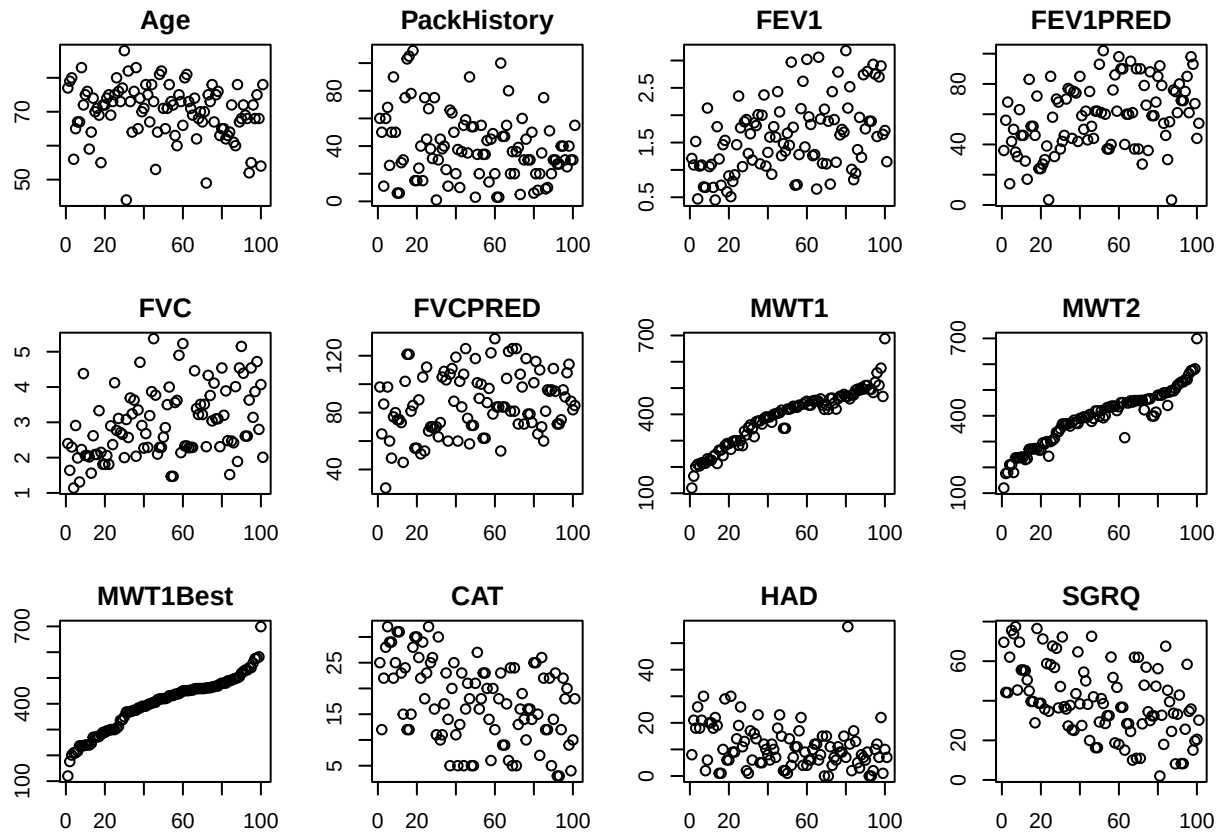



NB : spotted outlier in CAT variable

```
COPD.df$CAT[COPD.df$CAT>40]<-NA
```

QQ-plot

```
par(ask=TRUE)
par(mfrow=c(3,4))
par(mar = c(3, 3, 2, 1))
plot(COPD.df$AGE, main = "Age")
plot(COPD.df$PackHistory, main = "PackHistory")
plot(COPD.df$FEV1, main = "FEV1")
plot(COPD.df$FEV1PRED, main = "FEV1PRED")
plot(COPD.df$FVC, main="FVC")
plot(COPD.df$FVCPRED, main="FVCPRED")
plot(COPD.df$MWT1, main="MWT1")
plot(COPD.df$MWT2, main="MWT2")
plot(COPD.df$MWT1Best, main="MWT1Best")
plot(COPD.df$CAT, main = "CAT")
plot(COPD.df$HAD, main= "HAD")
plot(COPD.df$SGRQ, main = "SGRQ")
```




CrossTable for Categorical data factor : Gender, COPDseverity, copd, smoking, Diabetes, Muscular, Hypertension, AtrialFib, IHD

```
CrossTable(COPD.df$gender)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##            |         0 |         1 |
##            |-----------|-----------|
##            |        36 |        65 |
##            |     0.356 |     0.644 |
##            |-----------|-----------|
##
##
##
##
```

```r
CrossTable(COPD.df$copd)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##            |         1 |         2 |         3 |         4 |
##            |-----------|-----------|-----------|-----------|
##            |        23 |        43 |        27 |         8 |
##            |     0.228 |     0.426 |     0.267 |     0.079 |
##            |-----------|-----------|-----------|-----------|
##
##
##
##
```

```r
CrossTable(COPD.df$COPDSEVERITY)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##              |        MILD |    MODERATE |      SEVERE | VERY SEVERE |
##              |-------------|-------------|-------------|-------------|
##              |          23 |          43 |          27 |           8 |
##              |       0.228 |       0.426 |       0.267 |       0.079 |
##              |-------------|-------------|-------------|-------------|
##
##
##
##
```

```r
CrossTable(COPD.df$AGEquartiles)
```

```
##
##
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##            |         1 |         2 |         3 |         4 |
##            |-----------|-----------|-----------|-----------|
##            |        26 |        24 |        28 |        23 |
##            |     0.257 |     0.238 |     0.277 |     0.228 |
##            |-----------|-----------|-----------|-----------|
##
##
##
##
```

```
CrossTable(COPD.df$smoking)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##            |         1 |         2 |
##            |-----------|-----------|
##            |        16 |        85 |
##            |     0.158 |     0.842 |
##            |-----------|-----------|
##
##
##
##
```

```
CrossTable(COPD.df$Diabetes)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
```

```
##
##
## Total Observations in Table:  101
##
##
##           |         0 |         1 |
##           |-----------|-----------|
##           |        80 |        21 |
##           |     0.792 |     0.208 |
##           |-----------|-----------|
##
##
##
##
```

```
CrossTable(COPD.df$muscular)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##           |         0 |         1 |
##           |-----------|-----------|
##           |        82 |        19 |
##           |     0.812 |     0.188 |
##           |-----------|-----------|
##
##
##
##
```

```
CrossTable(COPD.df$hypertension)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
```

```
##           |         0 |         1 |
##           |-----------|-----------|
##           |        89 |        12 |
##           |     0.881 |     0.119 |
##           |-----------|-----------|
##
##
##
##
```

```
CrossTable(COPD.df$AtrialFib)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##           |         0 |         1 |
##           |-----------|-----------|
##           |        81 |        20 |
##           |     0.802 |     0.198 |
##           |-----------|-----------|
##
##
##
##
```

```
CrossTable(COPD.df$IHD)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##           |         0 |         1 |
##           |-----------|-----------|
##           |        92 |         9 |
##           |     0.911 |     0.089 |
##           |-----------|-----------|
```

```
##
##
##
##
```

```
summary(COPD.df)
```

```
##        X              ID              AGE
##  Min.   :  1   Min.   :  1.0   Min.   :44.0
##  1st Qu.: 26   1st Qu.: 49.0   1st Qu.:65.0
##  Median : 51   Median : 87.0   Median :71.0
##  Mean   : 51   Mean   : 91.4   Mean   :70.1
##  3rd Qu.: 76   3rd Qu.:143.0   3rd Qu.:75.0
##  Max.   :101   Max.   :169.0   Max.   :88.0
##
##   PackHistory    COPDSEVERITY            MWT1
##  Min.   :  1.0   Length:101         Min.   :120
##  1st Qu.: 20.0   Class :character   1st Qu.:300
##  Median : 36.0   Mode  :character   Median :419
##  Mean   : 39.7                      Mean   :386
##  3rd Qu.: 54.0                      3rd Qu.:460
##  Max.   :109.0                      Max.   :688
##                                     NA's   :2
##       MWT2         MWT1Best         FEV1
##  Min.   :120   Min.   :120   Min.   :0.45
##  1st Qu.:304   1st Qu.:304   1st Qu.:1.10
##  Median :399   Median :420   Median :1.60
##  Mean   :390   Mean   :399   Mean   :1.60
##  3rd Qu.:459   3rd Qu.:465   3rd Qu.:1.96
##  Max.   :699   Max.   :699   Max.   :3.18
##  NA's   :1     NA's   :1
##     FEV1PRED           FVC          FVCPRED
##  Min.   :  3.29   Min.   :1.14   Min.   : 27.0
##  1st Qu.: 42.00   1st Qu.:2.27   1st Qu.: 71.0
##  Median : 60.00   Median :2.77   Median : 84.0
##  Mean   : 58.53   Mean   :2.95   Mean   : 86.4
##  3rd Qu.: 75.00   3rd Qu.:3.63   3rd Qu.:103.0
##  Max.   :102.00   Max.   :5.37   Max.   :132.0
##
##       CAT             HAD             SGRQ      AGEquartiles
##  Min.   : 3.0   Min.   : 0.0   Min.   : 2.0   1:26
##  1st Qu.:12.0   1st Qu.: 6.0   1st Qu.:28.4   2:24
##  Median :18.0   Median :10.0   Median :38.2   3:28
##  Mean   :17.6   Mean   :11.2   Mean   :40.2   4:23
##  3rd Qu.:23.2   3rd Qu.:15.0   3rd Qu.:55.2
##  Max.   :32.0   Max.   :56.2   Max.   :77.4
##  NA's   :1
##  copd   gender smoking Diabetes muscular hypertension
##  1:23   0:36   1:16    0:80     0:82     0:89
##  2:43   1:65   2:85    1:21     1:19     1:12
##  3:27
##  4: 8
##
##
```

```
##
##  AtrialFib IHD
##  0:81      0:92
##  1:20      1: 9
##
##
##
##
##
```

2. Correlation matrix for all of the variables Age, PackHistory, FEV, FEV1PRED, FVC, FVCPRED, CAT, HAD, MWT1, MWT2, MWT1Best, SGRQ

```
my_data<-COPD.df[,c("AGE","PackHistory","FEV1","FEV1PRED","FVC","FVCPRED","MWT1","MWT2","MWT1Best","CAT
cor_matrix <- cor(my_data, use = "complete.obs")
```

```
cor_matrix
```

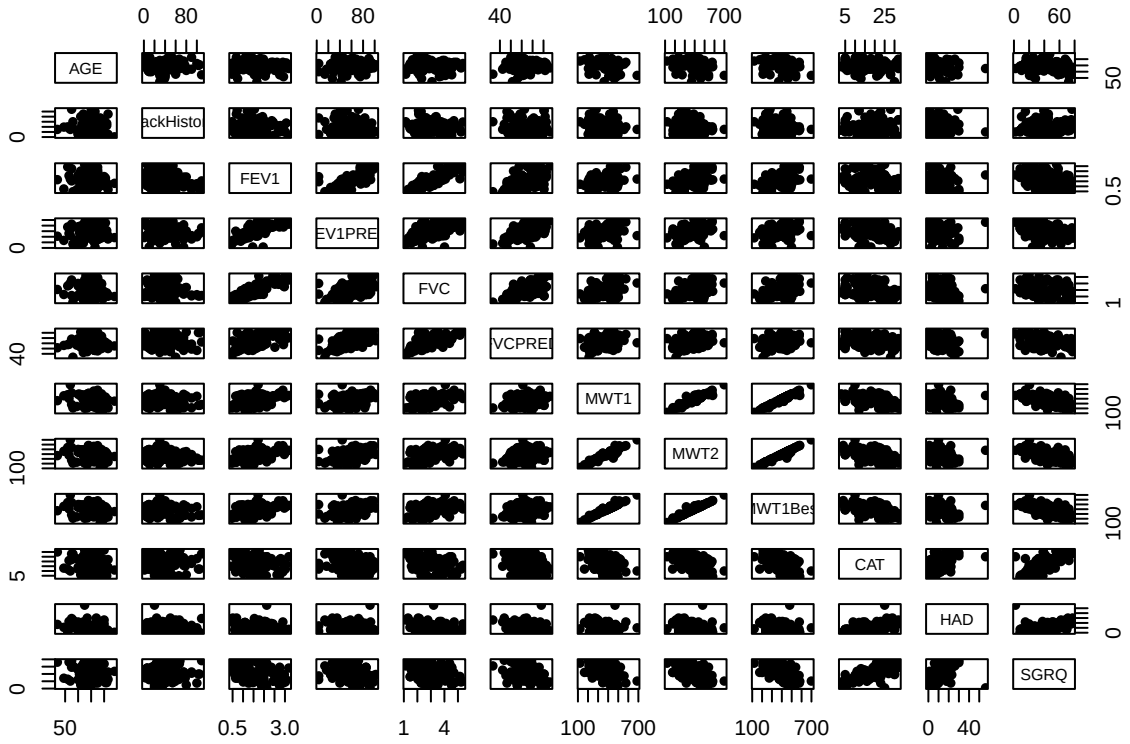
```
##                     AGE  PackHistory       FEV1  FEV1PRED
## AGE          1.00000000 -0.002327872 -0.0951062  0.063824
## PackHistory -0.00232787  1.000000000 -0.1139263 -0.103737
## FEV1        -0.09510622 -0.113926292  1.0000000  0.776589
## FEV1PRED     0.06382399 -0.103737169  0.7765886  1.000000
## FVC         -0.12242161 -0.092235437  0.8250070  0.535523
## FVCPRED      0.01639986 -0.000619901  0.5164061  0.632811
## MWT1        -0.25528262 -0.247814744  0.4572434  0.361608
## MWT2        -0.24416800 -0.275560704  0.4688137  0.404865
## MWT1Best    -0.23592582 -0.245479528  0.4666321  0.386294
## CAT         -0.08201584 -0.022211479 -0.2709537 -0.289304
## HAD         -0.20518297  0.036187701 -0.1606625 -0.114637
## SGRQ        -0.12654615  0.040245942 -0.3142075 -0.341055
##                     FVC      FVCPRED      MWT1      MWT2
## AGE         -0.1224216  0.016399858 -0.255283 -0.244168
## PackHistory -0.0922354 -0.000619901 -0.247815 -0.275561
## FEV1         0.8250070  0.516406104  0.457243  0.468814
## FEV1PRED     0.5355231  0.632810743  0.361608  0.404865
## FVC          1.0000000  0.624284281  0.454472  0.461754
## FVCPRED      0.6242843  1.000000000  0.257983  0.301001
## MWT1         0.4544721  0.257983250  1.000000  0.954355
## MWT2         0.4617537  0.301001068  0.954355  1.000000
## MWT1Best     0.4492602  0.259000224  0.982331  0.982224
## CAT         -0.2382556 -0.305299193 -0.426134 -0.490291
## HAD         -0.1498217 -0.161899699 -0.260132 -0.290703
## SGRQ        -0.2339245 -0.294294401 -0.507305 -0.522711
##                MWT1Best        CAT        HAD       SGRQ
## AGE         -0.235926 -0.0820158 -0.2051830 -0.1265461
## PackHistory -0.245480 -0.0222115  0.0361877  0.0402459
## FEV1         0.466632 -0.2709537 -0.1606625 -0.3142075
## FEV1PRED     0.386294 -0.2893041 -0.1146373 -0.3410551
## FVC          0.449260 -0.2382556 -0.1498217 -0.2339245
## FVCPRED      0.259000 -0.3052992 -0.1618997 -0.2942944
## MWT1         0.982331 -0.4261338 -0.2601325 -0.5073050
## MWT2         0.982224 -0.4902910 -0.2907034 -0.5227112
```

```
## MWT1Best     1.000000 -0.4741556 -0.2925360 -0.5372415
## CAT         -0.474156  1.0000000  0.5293911  0.7334832
## HAD         -0.292536  0.5293911  1.0000000  0.3897020
## SGRQ        -0.537241  0.7334832  0.3897020  1.0000000
```

```
round(cor_matrix,4)
```

```
##                 AGE PackHistory    FEV1 FEV1PRED     FVC
## AGE          1.0000     -0.0023 -0.0951   0.0638 -0.1224
## PackHistory -0.0023      1.0000 -0.1139  -0.1037 -0.0922
## FEV1        -0.0951     -0.1139  1.0000   0.7766  0.8250
## FEV1PRED     0.0638     -0.1037  0.7766   1.0000  0.5355
## FVC         -0.1224     -0.0922  0.8250   0.5355  1.0000
## FVCPRED      0.0164     -0.0006  0.5164   0.6328  0.6243
## MWT1        -0.2553     -0.2478  0.4572   0.3616  0.4545
## MWT2        -0.2442     -0.2756  0.4688   0.4049  0.4618
## MWT1Best    -0.2359     -0.2455  0.4666   0.3863  0.4493
## CAT         -0.0820     -0.0222 -0.2710  -0.2893 -0.2383
## HAD         -0.2052      0.0362 -0.1607  -0.1146 -0.1498
## SGRQ        -0.1265      0.0402 -0.3142  -0.3411 -0.2339
##             FVCPRED    MWT1    MWT2 MWT1Best     CAT
## AGE          0.0164 -0.2553 -0.2442  -0.2359 -0.0820
## PackHistory -0.0006 -0.2478 -0.2756  -0.2455 -0.0222
## FEV1         0.5164  0.4572  0.4688   0.4666 -0.2710
## FEV1PRED     0.6328  0.3616  0.4049   0.3863 -0.2893
## FVC          0.6243  0.4545  0.4618   0.4493 -0.2383
## FVCPRED      1.0000  0.2580  0.3010   0.2590 -0.3053
## MWT1         0.2580  1.0000  0.9544   0.9823 -0.4261
## MWT2         0.3010  0.9544  1.0000   0.9822 -0.4903
## MWT1Best     0.2590  0.9823  0.9822   1.0000 -0.4742
## CAT         -0.3053 -0.4261 -0.4903  -0.4742  1.0000
## HAD         -0.1619 -0.2601 -0.2907  -0.2925  0.5294
## SGRQ        -0.2943 -0.5073 -0.5227  -0.5372  0.7335
##                 HAD    SGRQ
## AGE         -0.2052 -0.1265
## PackHistory  0.0362  0.0402
## FEV1        -0.1607 -0.3142
## FEV1PRED    -0.1146 -0.3411
## FVC         -0.1498 -0.2339
## FVCPRED     -0.1619 -0.2943
## MWT1        -0.2601 -0.5073
## MWT2        -0.2907 -0.5227
## MWT1Best    -0.2925 -0.5372
## CAT          0.5294  0.7335
## HAD          1.0000  0.3897
## SGRQ         0.3897  1.0000
```

```
pairs(~AGE+PackHistory+FEV1+FEV1PRED+FVC+FVCPRED+MWT1+MWT2+MWT1Best+CAT+HAD+SGRQ, data=COPD.df, pch=16,
```




---

Using crosstable to examine categorical variables :

```
CrossTable(COPD.df$hypertension, COPD.df$IHD)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
##                      | COPD.df$IHD
## COPD.df$hypertension |         0 |         1 | Row Total |
## ---------------------|-----------|-----------|-----------|
##                    0 |        81 |         8 |        89 |
```

```
##                       |     0.000 |     0.001 |           |
##                       |     0.910 |     0.090 |     0.881 |
##                       |     0.880 |     0.889 |           |
##                       |     0.802 |     0.079 |           |
## ----------------------|-----------|-----------|-----------|
##                     1 |        11 |         1 |        12 |
##                       |     0.000 |     0.004 |           |
##                       |     0.917 |     0.083 |     0.119 |
##                       |     0.120 |     0.111 |           |
##                       |     0.109 |     0.010 |           |
## ----------------------|-----------|-----------|-----------|
##         Column Total  |        92 |         9 |       101 |
##                       |     0.911 |     0.089 |           |
## ----------------------|-----------|-----------|-----------|
##
##
```

Fit regression SGRQ (Quality of Life) with (FEV1)

```
copd_sgrq<- lm(SGRQ~FEV1, data=COPD.df)
```

```
summary(copd_sgrq)
```

```
##
## Call:
## lm(formula = SGRQ ~ FEV1, data = COPD.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33.86 -12.51  -2.01  12.14  36.13
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)    53.38       4.51   11.83 <0.0000000000000002
## FEV1           -8.23       2.60   -3.17               0.002
##
## (Intercept) ***
## FEV1        **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.5 on 99 degrees of freedom
## Multiple R-squared:  0.0921, Adjusted R-squared:  0.0829
## F-statistic:   10 on 1 and 99 DF,  p-value: 0.00204
```

```
confint(copd_sgrq)
```

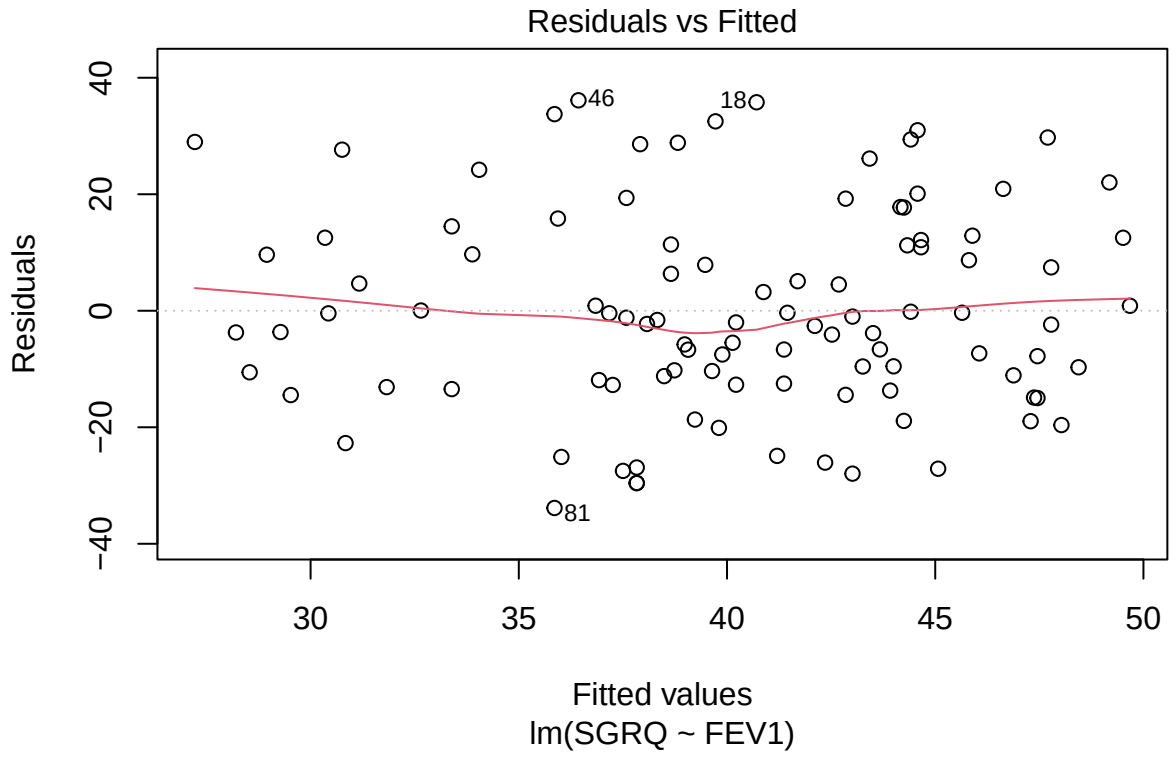
```
##                2.5 %   97.5 %
## (Intercept)  44.4250 62.33073
## FEV1        -13.3773 -3.07401
```

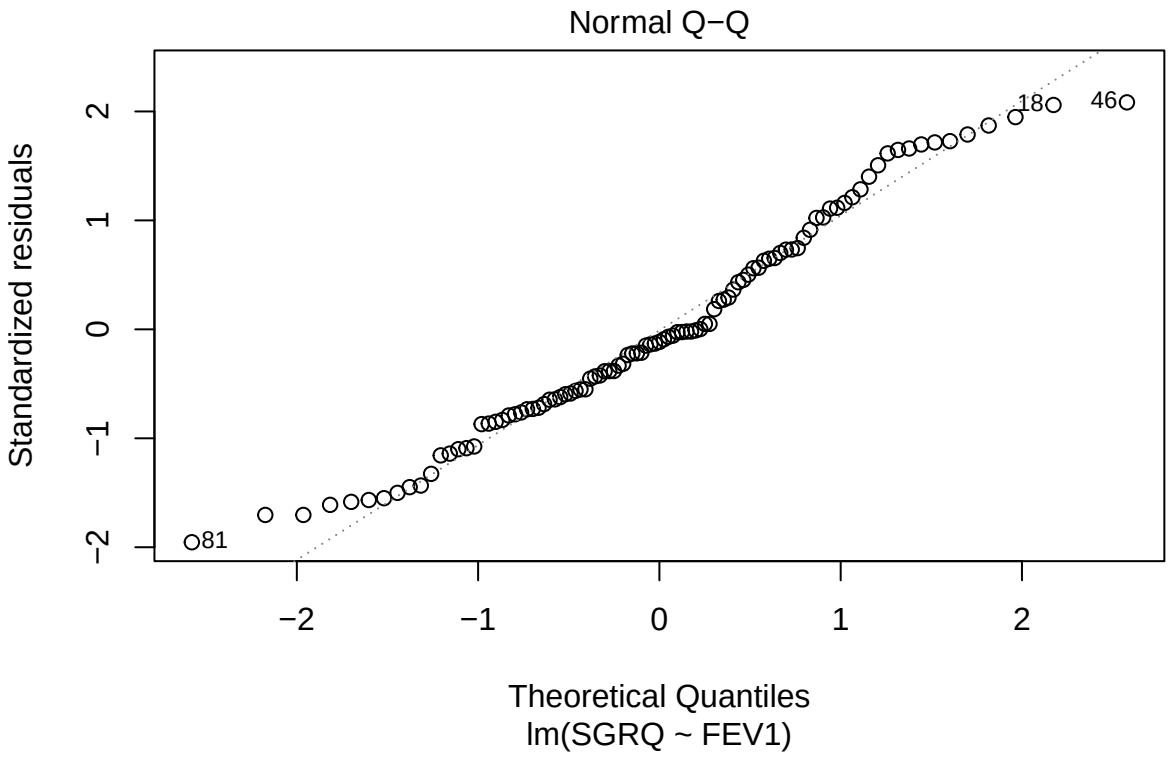
Check the model Fit the model :

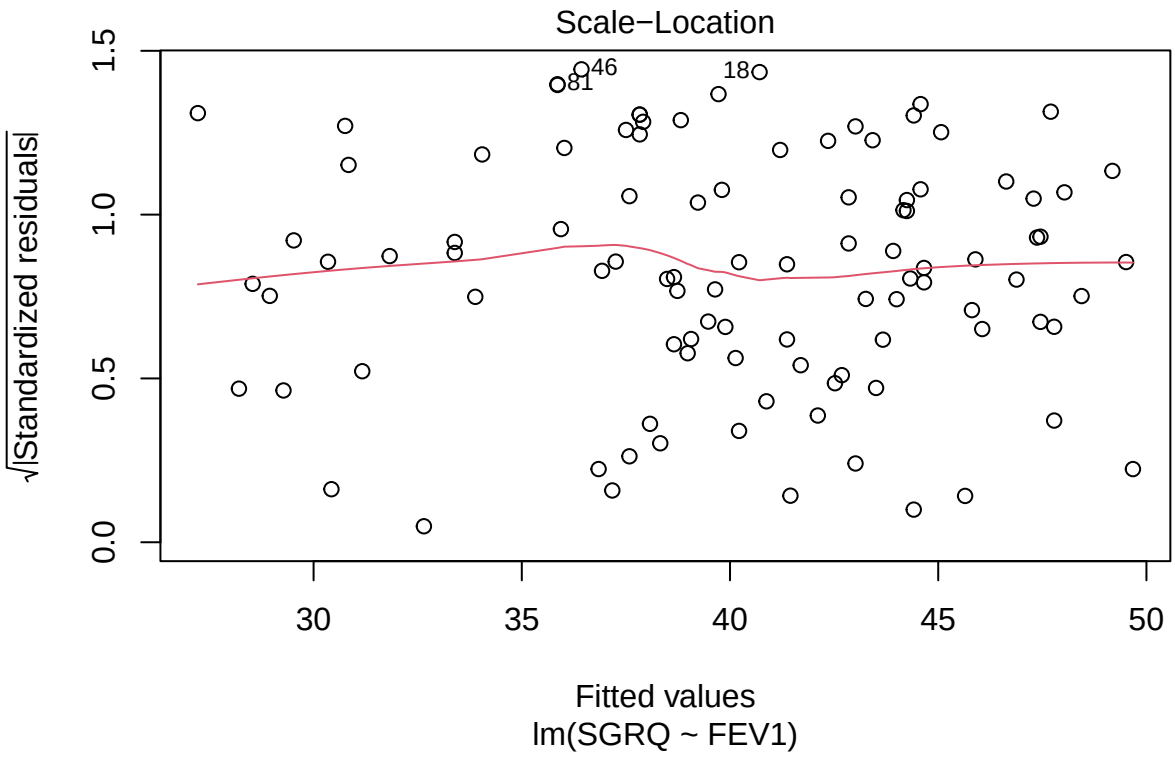
```
predictedValsgrq <- predict(copd_sgrq)
residualValsgrq <- residuals(copd_sgrq)
```

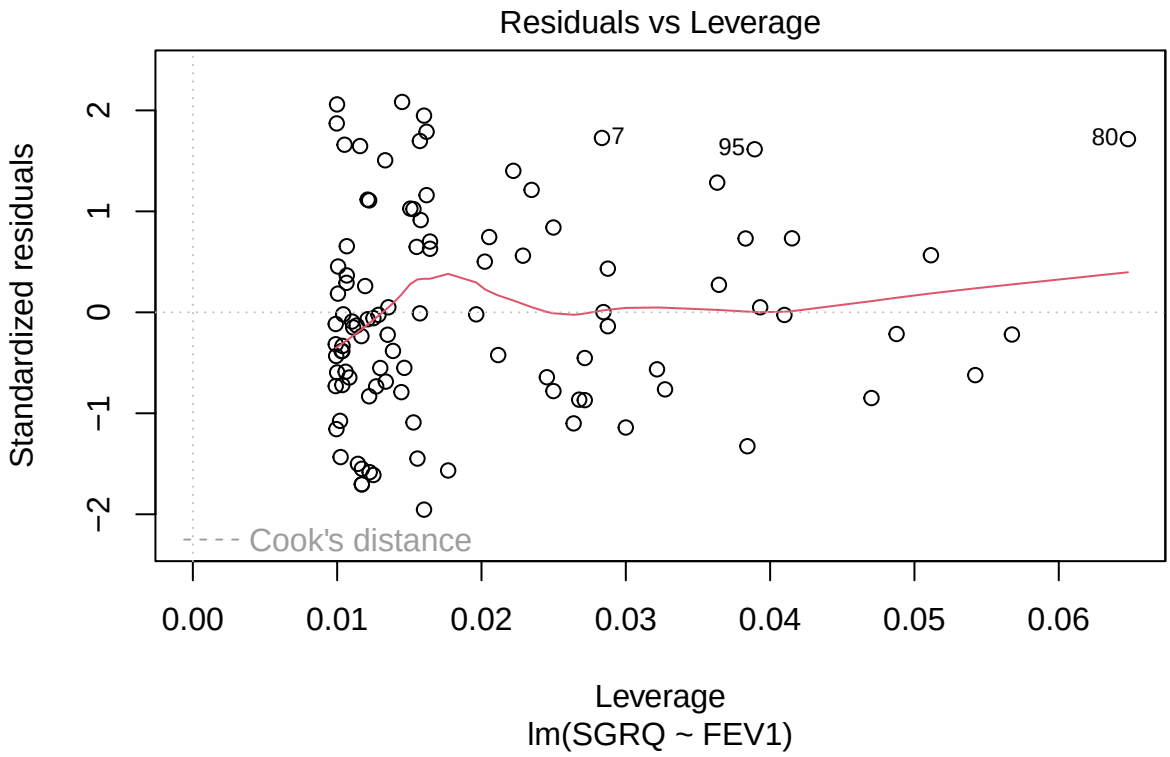
Check using plots :

```
plot(copd_sgrq)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(SGRQ ~ FEV1)
81
18
46
−2
−1
0
1
2

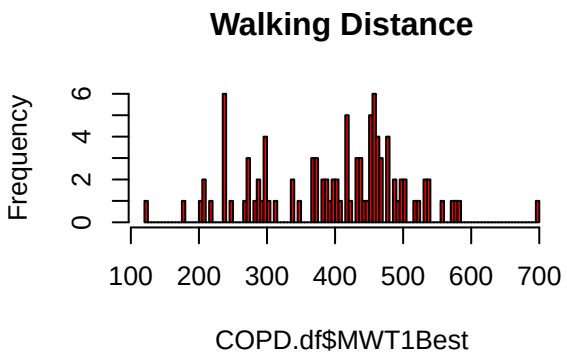

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
46
81
18
30
35
40
45
50
Fitted values
lm(SGRQ ~ FEV1)

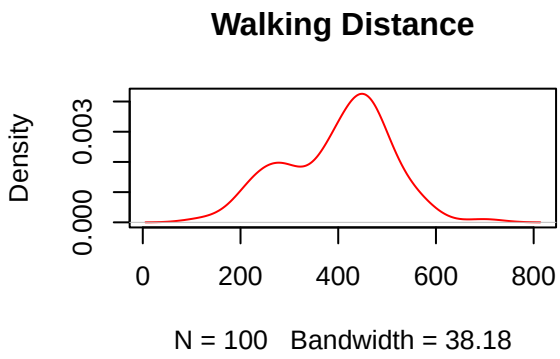


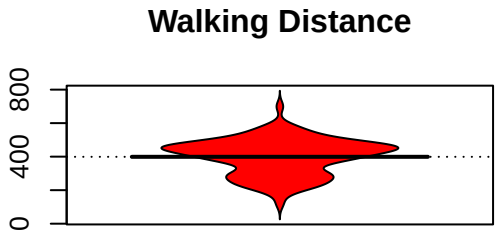

Lung function (FEV1), age (AGE), gender (gender), COPD severity (COPDSEVERITY) and presence of any comorbidity (comorbid) as the final predictor variables for your multivariable model to predict MWT1best

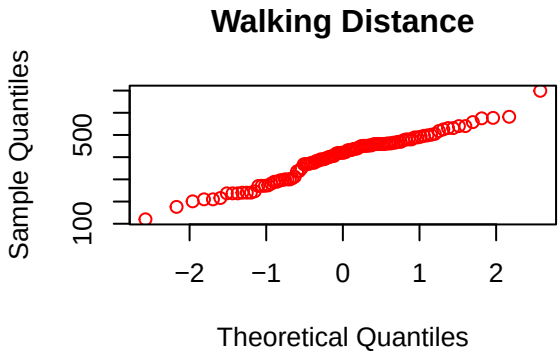
1. Check the normality using histogram, qqplot and shapiro-wilk (FEV, AGE to MWT1best)

```
par(ask=TRUE)
par(mfrow=c(2,2))

hist(COPD.df$MWT1Best, main="Walking Distance",col="red", breaks=200)

plot(density(COPD.df$MWT1Best, na.rm=TRUE),
     main="Walking Distance", col="red")

beanplot::beanplot(COPD.df$MWT1Best, main="Walking Distance", col="red", what=c(0.85,0.85,0.85,0), overa

qqnorm(COPD.df$MWT1Best,main="Walking Distance",col="red")
```

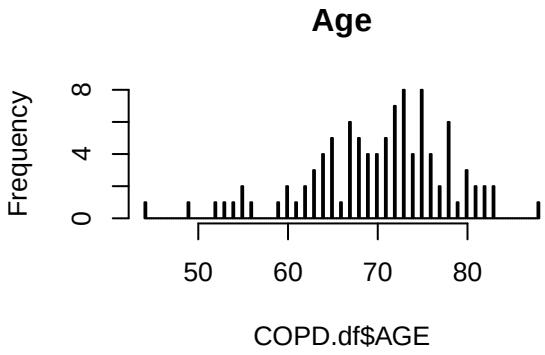


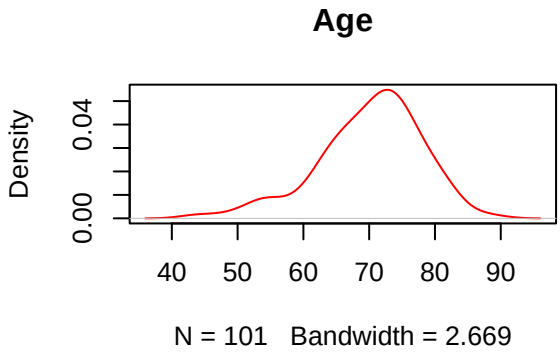



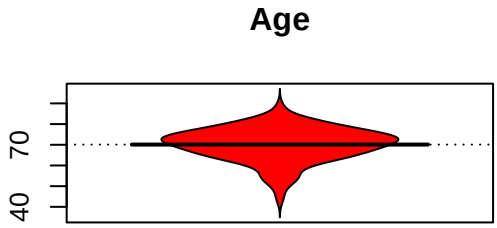



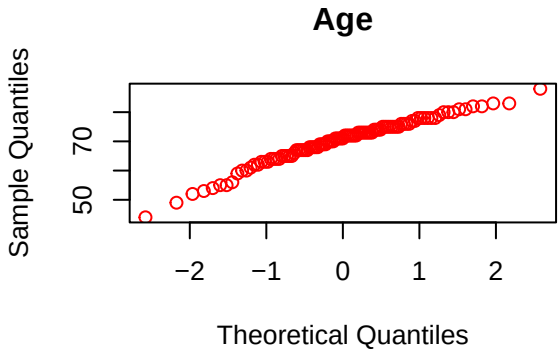




```r
par(ask=TRUE)
par(mfrow=c(2,2))

hist(COPD.df$AGE, main="Age",col="red", breaks=200)

plot(density(COPD.df$AGE, na.rm=TRUE),
     main="Age", col="red")

beanplot::beanplot(COPD.df$AGE, main="Age", col="red", what=c(0.85,0.85,0.85,0), overallline="mean", bo

qqnorm(COPD.df$AGE,main="Age",col="red")
```



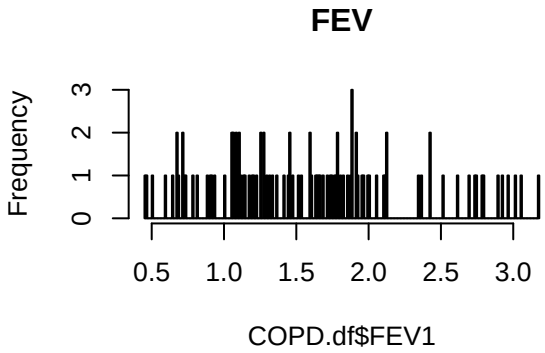



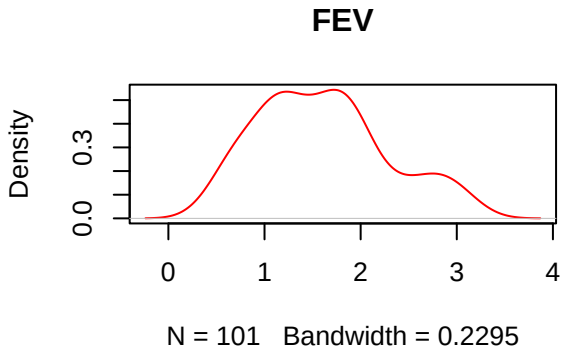



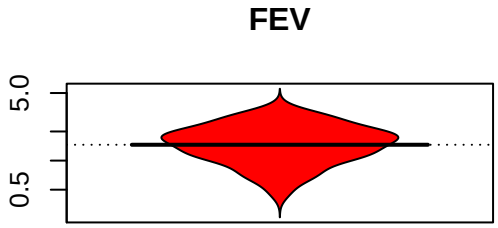

```
par(ask=TRUE)
par(mfrow=c(2,2))

hist(COPD.df$FEV1, main="FEV",col="red", breaks=200)

plot(density(COPD.df$FEV1, na.rm=TRUE),
     main="FEV", col="red")

beanplot::beanplot(COPD.df$FEV1, main="FEV", col="red", what=c(0.85,0.85,0.85,0), overallline="mean", bo
```

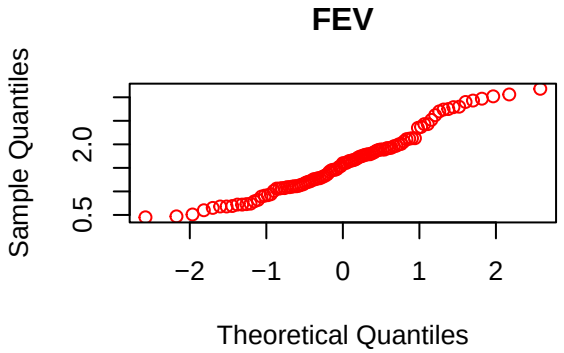
```
## log="y" selected
```

```
qqnorm(COPD.df$FEV1,main="FEV",col="red")
```




Saphiro Test

```
shapiro.test(COPD.df$MWT1Best)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  COPD.df$MWT1Best
## W = 0.9699, p-value = 0.0216
```

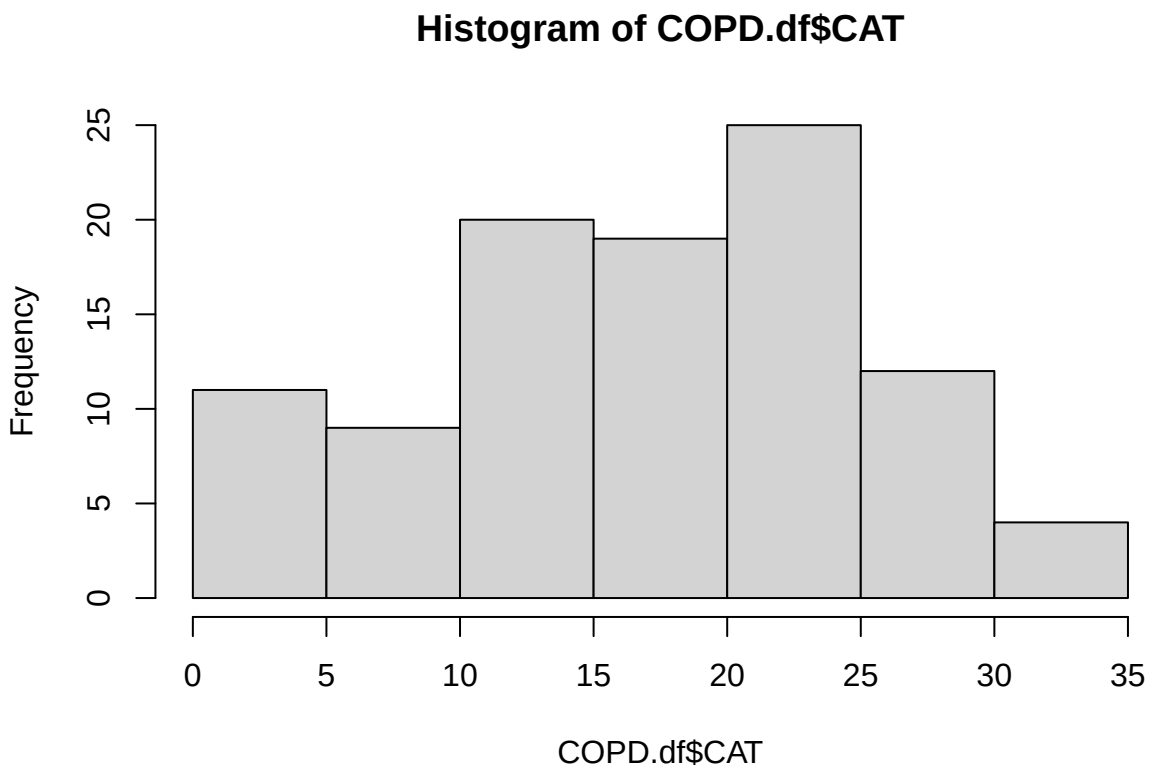
```
shapiro.test(COPD.df$AGE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  COPD.df$AGE
## W = 0.9677, p-value = 0.0139
```

```
shapiro.test(COPD.df$FEV1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  COPD.df$FEV1
## W = 0.9648, p-value = 0.00852
```

```r
hist(COPD.df$CAT)
```




There's possible outlier

```r
COPD.df$CAT[COPD.df$CAT>40] <- NA
```

2. Using crosstab

```r
CrossTable(COPD.df$gender, COPD.df$IHD)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  101
##
##
```

```
##                 | COPD.df$IHD
## COPD.df$gender |         0 |         1 | Row Total |
## ---------------|-----------|-----------|-----------|
##              0 |        34 |         2 |        36 |
##                |     0.044 |     0.455 |           |
##                |     0.944 |     0.056 |     0.356 |
##                |     0.370 |     0.222 |           |
##                |     0.337 |     0.020 |           |
## ---------------|-----------|-----------|-----------|
##              1 |        58 |         7 |        65 |
##                |     0.025 |     0.252 |           |
##                |     0.892 |     0.108 |     0.644 |
##                |     0.630 |     0.778 |           |
##                |     0.574 |     0.069 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |        92 |         9 |       101 |
##                |     0.911 |     0.089 |           |
## ---------------|-----------|-----------|-----------|
##
##
```