introduction

Albert Buchard

12/20/2016

How to keep your code clean

Coding convention

- Pick a naming convention and stick to it
- camelCase = “this is a nice style”
- snake_case = “this is ok too”
- Comment your code
- Look at the google style book to make sure your code is easilly readable by anyone
- https://google.github.io/styleguide/Rguide.xml
- they advice to use only “<-” and not “=” but i personally think it is pointless

Storage

Keep a README.md file at the root of your folder explaining where everything is, helping someone that knows nothing about your data to navigate your work. Keeping your work in the cloud, through services like dropbox, icloud, or google drive. The best would be github but it is not easy in the begining.

Folders

Keep your folder clean, with clear names in minuscules separated by “_" :

- data
  - raw
  - preprocessed
  - analysis
    - analysis_one …
- scripts
  - preprocessing: scripts that transforms the raw data in processed data
  - analysis: scripts that use preprocessed data and performs analysis on it
  - markdown: your markdown files
  - r_files: other R files, like utility functions
- media: here should go any ressources, presentations, images you produced or needed etc…
  - presentations
  - graphics
  - text
  - notes
- backups: you might need a backup folder when in doubt
  - data
  - script
  - media

Variables

R infer on its own the type of the variable that you want to create based on the input you give. All variables are at minimum a vector.

Atomic vector data type

```
# Character
a <- "This is a character vector"

# Numeric (integer)
a = 12
a <- 12

# Numeric (Float)
a <- 12.2

# Logical
a <- TRUE
a <- FALSE

print(a)
```

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

Combine atomic elements

Vectors needs to be of same type

```
a <- c(1,2,3)
print(a)
```

```
## [1] 1 2 3
```

```
a <- c(1,"a")
print(a)
```

```
## [1] "1" "a"
```

List can mix types

```
a <- list(1,2,3)
print(a)
```

```
## [[1]]
## [1] 1
##
## [[2]]
## [1] 2
##
## [[3]]
## [1] 3
```

```r
a <- list(1,"a")
print(a)
```

```
## [[1]]
## [1] 1
##
## [[2]]
## [1] "a"
```

Matrix 2*2 Needs to be of same type

```r
a <- matrix( c('a','a','b','c','b',2), nrow = 2, ncol = 3, byrow = T)
print(a)
```

```
##      [,1] [,2] [,3]
## [1,] "a"  "a"  "b"
## [2,] "c"  "b"  "2"
```

```r
a <- matrix( c('a','a','b','c','b',2), nrow = 2, ncol = 3, byrow = F)
print(a)
```

```
##      [,1] [,2] [,3]
## [1,] "a"  "b"  "b"
## [2,] "a"  "c"  "2"
```

Array N*N Needs to be of same type

```r
a <- array(c('green','yellow'),dim = c(3,3,2))
print(a)
```

```
## , , 1
##
##      [,1]     [,2]     [,3]
## [1,] "green"  "yellow" "green"
## [2,] "yellow" "green"  "yellow"
## [3,] "green"  "yellow" "green"
##
## , , 2
##
##      [,1]     [,2]     [,3]
## [1,] "yellow" "green"  "yellow"
## [2,] "green"  "yellow" "green"
## [3,] "yellow" "green"  "yellow"
```

Factor

For categorical variables

```r
# Create a vector.
apple_colors <- c('green','green','yellow','red','red','red','green')

# Create a factor object.
```

```r
factor_apple <- factor(apple_colors)

# Print the factor.
print(factor_apple)
```

```
## [1] green  green  yellow red    red    red    green
## Levels: green red yellow
```

```r
print(nlevels(factor_apple))
```

```
## [1] 3
```

```r
# Change names of the factors
levels(factor_apple) <- c("Kindof Green", "Kindof Red", "Kindof Yellow")
print(factor_apple)
```

```
## [1] Kindof Green  Kindof Green  Kindof Yellow Kindof Red    Kindof Red
## [6] Kindof Red    Kindof Green
## Levels: Kindof Green Kindof Red Kindof Yellow
```

Dataframe ++

```r
first_names = c("Melissa", "Sibylle", "Zoe", "Maria")
ages = c(23, 22, 24, 25)

df <- data.frame(first_name = first_names,
                 age = ages,
                 subject = as.character(c("Activity", "Motivation", "Fluid Intelligence", NA)))



print(df)
```

```
##   first_name age            subject
## 1    Melissa  23           Activity
## 2    Sibylle  22         Motivation
## 3        Zoe  24 Fluid Intelligence
## 4      Maria  25               <NA>
```

```r
print(df$age)
```

```
## [1] 23 22 24 25
```

Missing values are "NOT ASSIGNED" or "NA"

```r
print(df$subject)
```

```
## [1] Activity           Motivation         Fluid Intelligence
## [4] <NA>
## Levels: Activity Fluid Intelligence Motivation
```

```r
print(is.na(df$subject))
```

```
## [1] FALSE FALSE FALSE  TRUE
```

I dont want a factor, I want characters !

Sometimes you have ot use function such as apply or sapply, that performs simple loops on your data.

```
print(sapply(df[, 3], as.character))
```

```
## [1] "Activity"           "Motivation"         "Fluid Intelligence"
## [4] NA
```

```
df[, 3] <- sapply(df[, 3], as.character)

print(df)
```

```
##   first_name age            subject
## 1    Melissa  23           Activity
## 2    Sibylle  22         Motivation
## 3        Zoe  24 Fluid Intelligence
## 4      Maria  25               <NA>
```

Operators

Relational operators

```
print(12>23)
```

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

```
a = 12
print(a == 12)
```

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

```
print(a != 32)
```

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

```
print(a >= 11)
```

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

```
a = 32
```

Tests

```
if (a == 432) {
  print("a est egale a 12 !")
} else {
  print("pas egale a 12")
}
```

```
## [1] "pas egale a 12"
```

Logical operators

```
print((12>23)&&(12<23))
```

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

```
a <- 12
print((a>20)||(a==12))
```

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

```
print(!(a == 32))
```

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

```
!is.na(a)
```

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

Element wise logic

When a vector is tested against a vector of same length

```
a <- c(F, T, T)
b <- c(T, F, T)
print(a&b)
```

```
## [1] FALSE FALSE  TRUE
```

```
print(a|b)
```

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

Others

```
a <- 1:8
print(a)
```

```
## [1] 1 2 3 4 5 6 7 8
```

```
a <- rep("ce qui est repete", 4)
print(a)
```

```
## [1] "ce qui est repete" "ce qui est repete" "ce qui est repete"
## [4] "ce qui est repete"
```

Flow control statements

Flow controls statements are all thestatement of a language that will redirect the flow of execution of a program.

Conditional control

Sometime you want to execute something only if a condition is true. The most used is the "if/else if/else" statement.

```
a = c(F,F,F,T)

if (a[1]) {
  print("first")
} else if (a[2]) {
  print("second")
}  else if (a[3]) {
  print("third")
}  else if (a[4]) {
  print("quatrieme")
} else {
  print("invalid")
}
```

```
## [1] "quatrieme"
```

```
b <- c(F,F,F,T)

# TODO: write a statement that checks if b has any of its value equal to TRUE.
# If it does return all the indices of the TRUE values
# If not, say that you did not find a True value in any of the %SIZE% elements of b
# (hint:: ?any and ?which)
```

When you only want to check the value of *ONE* variable. Another way is to use the *switch* statement. It test the value of a variable against several possibilities, like so:

```
strangeName <- "Grabulas"
switch(strangeName,
       "BJ Gabbour" = {
         print("It was Gabbour all along !")
       },
       "Horticher" = {
         print("She's just everywhere")
       },
       "Grabulas" = {
         print("Run you fools !")
       })
```

```
## [1] "Run you fools !"
```

Loops

You often need to repeat some statement. That's what *for* and *while* are here for !

```
for (i in 1:NROW(df)) {
  person = df[i,]

  print(paste0(person$first_name, person$age))
}
```

```
## [1] "Melissa23"
```

```
## [1] "Sibylle22"
## [1] "Zoe24"
## [1] "Maria25"
```

```
i = 0
while(i<10) {
  print(i)
  i = i + 1
}
```

```
## [1] 0
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
```

Functions

Functions are the central concept to programming. You can think of it as a box containing a series of instructions. Usually they take input, perform change on it, and returns a value. Note that they do not always take input or return a value and act only a series on instructions that do not takes input and do not return anything, but for example will read or write some information on the disk, or setting some parameter.

A simple sum

```
sommeFunction = function  (x, y, z = 1) {
  # Do some action on the parameters
  sum = x+y+z

  # Return to sender the result of your computation
  return(sum)
}

somme = sommeFunction(1,2,3)
```

But of course R has a better function already built in !

```
sum(1,2,3,NA, na.rm = TRUE)
```

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

A function that generates participant ids

What if you want to set some default parameters ? Here is a more complex function that plays with strings to create random IDs for your subjects.

```
getRandomId = function(numberOfIds = 1, lenght=12, allowedCharacters = c(0:9, letters, LETTERS))
{
    # initialize vector
```

```r
    randomStrings = c(1:numberOfIds)

    # start the generation loop
    for (i in 1:numberOfIds)
    {
        randomStrings[i] <- paste(sample(allowedCharacters, lenght, replace=TRUE),
                                  collapse="")
    }

    # return the strings
    return(randomStrings)
}


# TODO Now generate 650 ids !
ids = getRandomId(650)

# letters and LETTERS are variables declared by default in R containing the minuscule and capital lette
print(c(0:9, letters, LETTERS))
```

```
##  [1] "0" "1" "2" "3" "4" "5" "6" "7" "8" "9" "a" "b" "c" "d" "e" "f" "g"
## [18] "h" "i" "j" "k" "l" "m" "n" "o" "p" "q" "r" "s" "t" "u" "v" "w" "x"
## [35] "y" "z" "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N" "O"
## [52] "P" "Q" "R" "S" "T" "U" "V" "W" "X" "Y" "Z"
```

```r
# Add the ids to the data
newDataFrame = data.frame(id=ids, sertARien=1:650)
```

Create your own function

Choose one between those three possible function, and create them: * A function that returns the product of two numbers such that aTimesb = product(a,b) * A function that adds a prefix to a string, such that prefixedString = prefix(prefixString, string) * A function that takes out the mean of each column of a data frame, and divides by the standard deviation (process called normalization)

Libraries

Install libraries

```r
install.packages("ggplot2")
install.packages("psych")
```

Load libraries

```r
library(ggplot2)
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
help(package= "ggplot2")
#vignette("ggplot2-specs", package = "ggplot2")
```

Explore your data

Load data

Usually you will load three type of data: * Excel files: .xslx * Comma separated values: .csv * R data file: .RData

You load them differently

```r
# Load a RData
setwd("~/Google Drive/Master Students/courses/introduction_a_r")
load("data/raw/data.RData")
pasdenom = read.csv("~/Google Drive/Master Students/courses/introduction_a_r/data/raw/data.csv")
class(pasdenom)
```

```
## [1] "data.frame"
```

```r
#or if you want to rename your data
renamedData = get(load("data/raw/data.RData"))
```

When in doubt, google it !

```r
# But.. how to load EXCEL FILES ?
# TODO Check stackoverflow / Google and load excel and csv file
# "load xlsx file R"
# "load csv file in R"
# can you copy paste ?? read.clipboard
```

Explore your data

Looking at your data before starting asking question is important to detect errors you might have made, wrong IDs, numbers of NA, wacky values… https://cran.r-project.org/web/packages/psych/vignettes/overview.pdf

```r
# View(data)

# Psych has a lot of tools for exploratory analysis
library(psych)
describe(data)
```

```
##              vars   n    mean     sd  median trimmed    mad     min
## age             1 650   41.22  13.96   41.00   41.23  17.79   18.00
## isOlder*        2 650     NaN     NA      NA     NaN     NA     Inf
## IQ              3 650   99.21  19.97   97.10   98.16  18.86   48.25
## responseTime    4 650 1945.46 390.79 1908.50 1922.20 409.20 1182.00
## performance     5 650    0.41   0.17    0.41    0.41   0.17    0.00
```

```
## id*             6 650     NaN     NA      NA     NaN     NA     Inf
## university      7 650    0.61   0.49    1.00    0.63   0.00    0.00
##                   max  range  skew kurtosis    se
## age             65.00   47.0 -0.02    -1.26  0.55
## isOlder*         -Inf   -Inf    NA       NA    NA
## IQ             175.85  127.6  0.61     0.73  0.78
## responseTime 3190.00 2008.0  0.47    -0.43 15.33
## performance     1.00    1.0  0.34     0.16  0.01
## id*              -Inf   -Inf    NA       NA    NA
## university       1.00    1.0 -0.43    -1.81  0.02
```

```r
describeBy(data, group = "isOlder")
```

```
## $`FALSE`
##              vars   n    mean     sd  median trimmed    mad     min
## age             1 592   39.09  12.77   39.00   39.11  16.31   18.00
## isOlder*        2 592     NaN     NA      NA     NaN     NA     Inf
## IQ              3 592  100.75  19.94   98.51   99.83  18.59   48.25
## responseTime    4 592 1883.95 348.13 1860.50 1864.23 376.58 1182.00
## performance     5 592    0.42   0.17    0.42    0.41   0.18    0.00
## id*             6 592     NaN     NA      NA     NaN     NA     Inf
## university      7 592    0.60   0.49    1.00    0.62   0.00    0.00
##                   max  range  skew kurtosis    se
## age             60.00   42.0 -0.01    -1.29  0.52
## isOlder*         -Inf   -Inf    NA       NA    NA
## IQ             175.85  127.6  0.56     0.75  0.82
## responseTime 3190.00 2008.0  0.51    -0.04 14.31
## performance     1.00    1.0  0.33     0.10  0.01
## id*              -Inf   -Inf    NA       NA    NA
## university       1.00    1.0 -0.41    -1.84  0.02
##
## $`TRUE`
##              vars  n    mean     sd  median trimmed    mad     min     max
## age             1 58   62.93   1.31   63.00   62.92   1.48   61.00   65.00
## isOlder*        2 58     NaN     NA      NA     NaN     NA     Inf    -Inf
## IQ              3 58   83.48  11.92   83.07   82.69  11.54   64.76  119.49
## responseTime    4 58 2573.36 204.34 2591.50 2576.29 208.31 2035.00 3084.00
## performance     5 58    0.35   0.13    0.37    0.36   0.12    0.05    0.69
## id*             6 58     NaN     NA      NA     NaN     NA     Inf    -Inf
## university      7 58    0.67   0.47    1.00    0.71   0.00    0.00    1.00
##                range  skew kurtosis    se
## age             4.00  0.08    -1.23  0.17
## isOlder*        -Inf    NA       NA    NA
## IQ             54.73  0.59     0.06  1.56
## responseTime 1049.00 -0.16    -0.15 26.83
## performance     0.64 -0.17    -0.40  0.02
## id*             -Inf    NA       NA    NA
## university      1.00 -0.72    -1.51  0.06
##
## attr(,"call")
## by.data.frame(data = x, INDICES = group, FUN = describe, type = type)
```

```r
describeData(data)
```

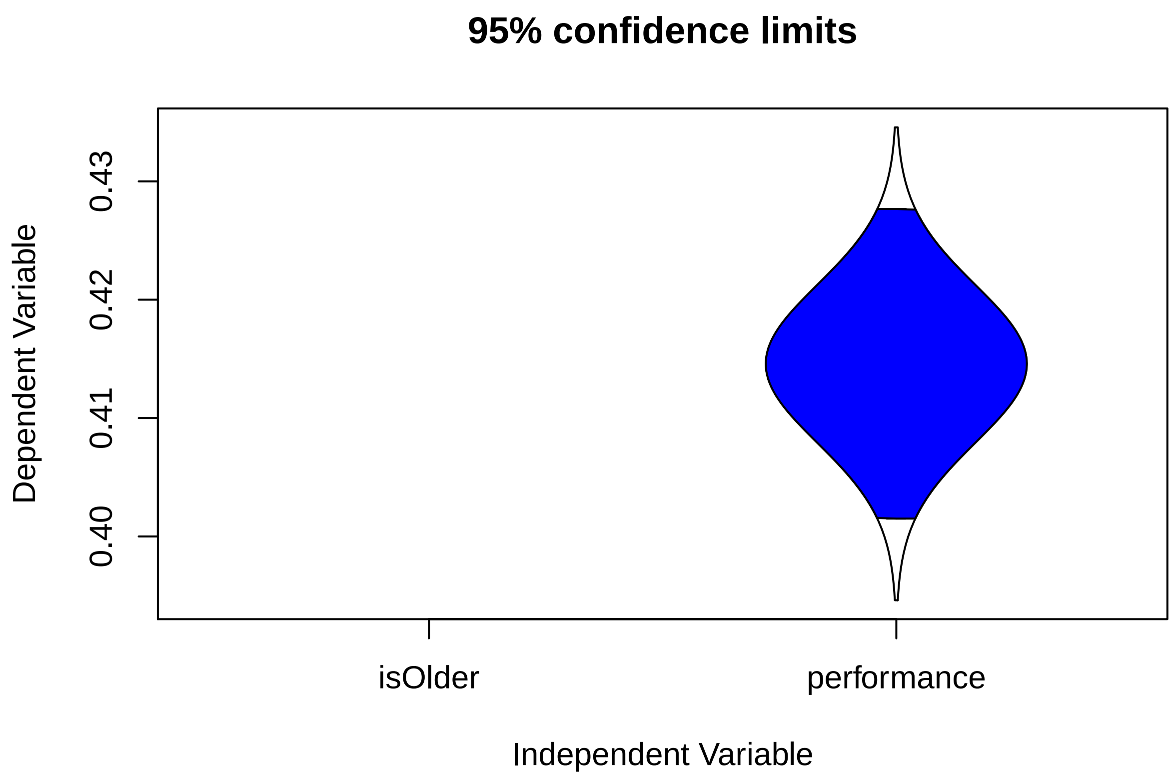
```
## n.obs =  650 of which  650   are complete cases.   Number of variables =  7  of which all are numeri
```

```
##               variable # n.obs type          H1          H2          H3
## age                 1   650    1          30          28          61
## isOlder*            2   650    2       FALSE       FALSE        TRUE
## IQ                  3   650    1   133.06719   133.40986    87.44439
## responseTime        4   650    1        1524        1432        2863
## performance         5   650    1   0.9158141   0.3465298   0.1488978
## id*                 6   650    3 vxFeC0Xgqc6i aUchqcWotUut 7nDl7aBhF3TD
## university          7   650    1           1           0           0
##                        H4          T1          T2          T3
## age                    27          30          44          39
## isOlder*            FALSE       FALSE       FALSE       FALSE
## IQ               95.40294   131.56426    97.63813    91.75452
## responseTime         1647        1451        2272        2047
## performance     0.4090058   0.5100192   0.5324680   0.1665098
## id*          5R2Mg2vnzCqT VNiTOX4LDyBP CpvqLy1Kmdu7 pXdzD0B52uCD
## university              0           0           1           0
##                        T4
## age                    58
## isOlder*            FALSE
## IQ               95.39839
## responseTime         2511
## performance     0.4886576
## id*          FiXUFOWB7poA
## university              1
```

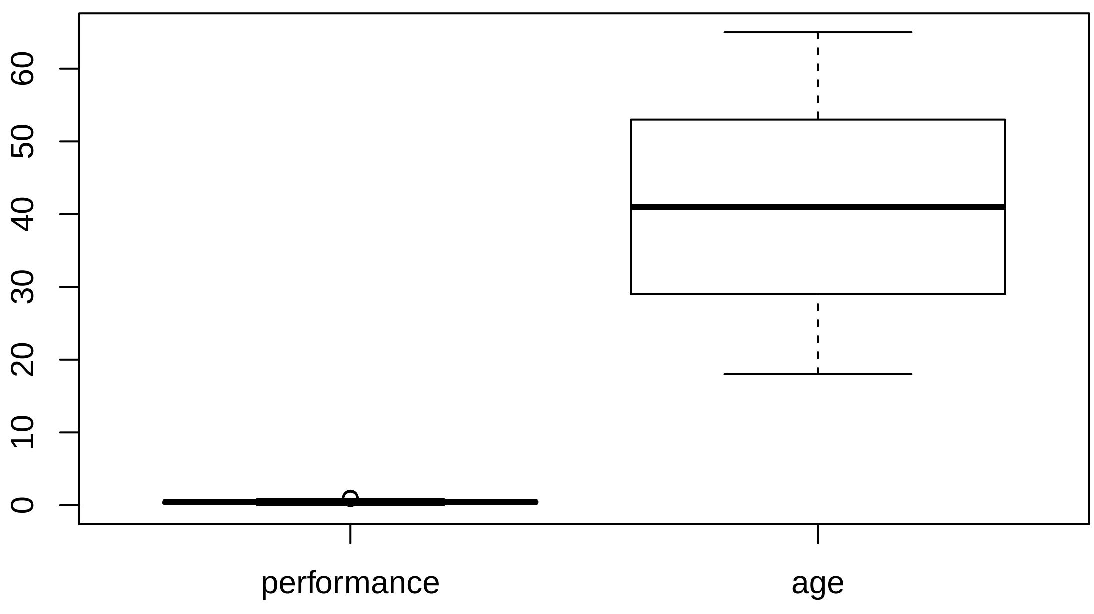
```r
head(data)
```

```
##   age isOlder        IQ responseTime performance           id university
## 1  30   FALSE 133.06719         1524   0.9158141 vxFeC0Xgqc6i          1
## 2  28   FALSE 133.40986         1432   0.3465298 aUchqcWotUut          0
## 3  61    TRUE  87.44439         2863   0.1488978 7nDl7aBhF3TD          0
## 4  27   FALSE  95.40294         1647   0.4090058 5R2Mg2vnzCqT          0
## 5  24   FALSE 125.69138         1484   0.6558282 0SWpcKdSUl6X          1
## 6  29   FALSE 113.91225         1706   0.5446137 iVNtPorZ5EuM          0
```

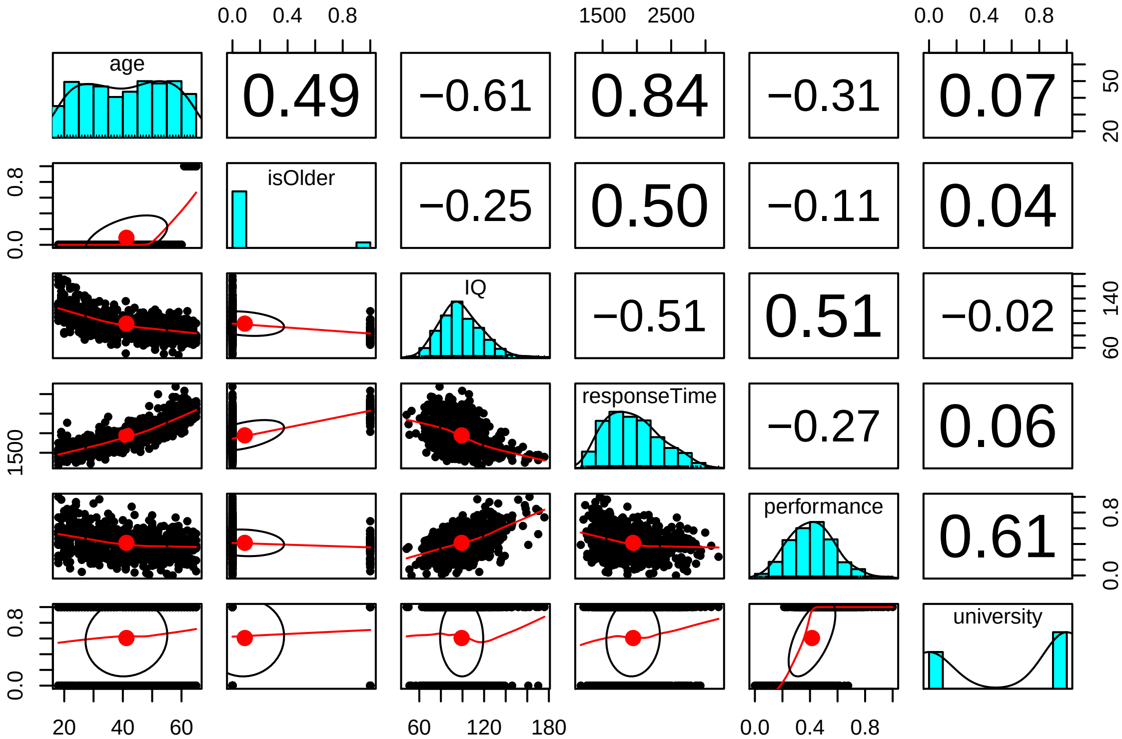
```r
# Some quick plots
error.bars(data[, c("isOlder", "performance")])
```



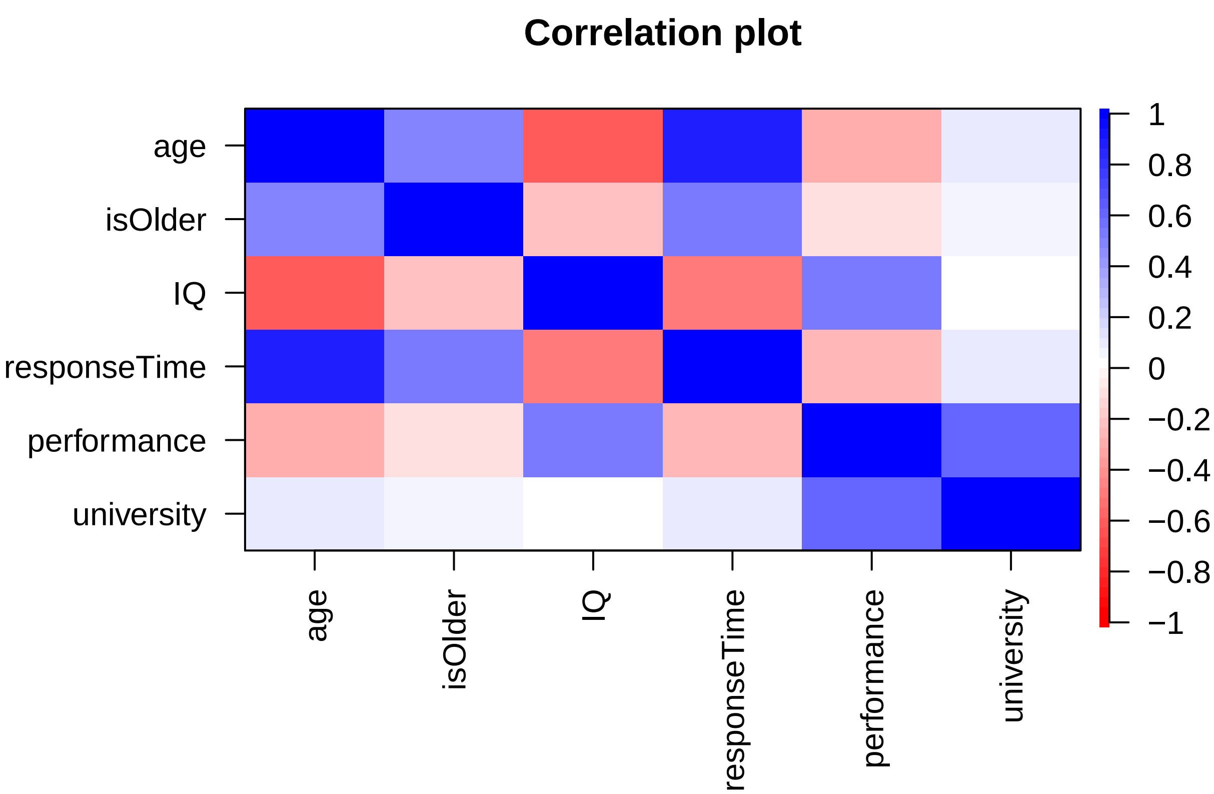

```r
boxplot(data[, c("performance", "age")])
```




```r
pairs.panels(data[-6])
```

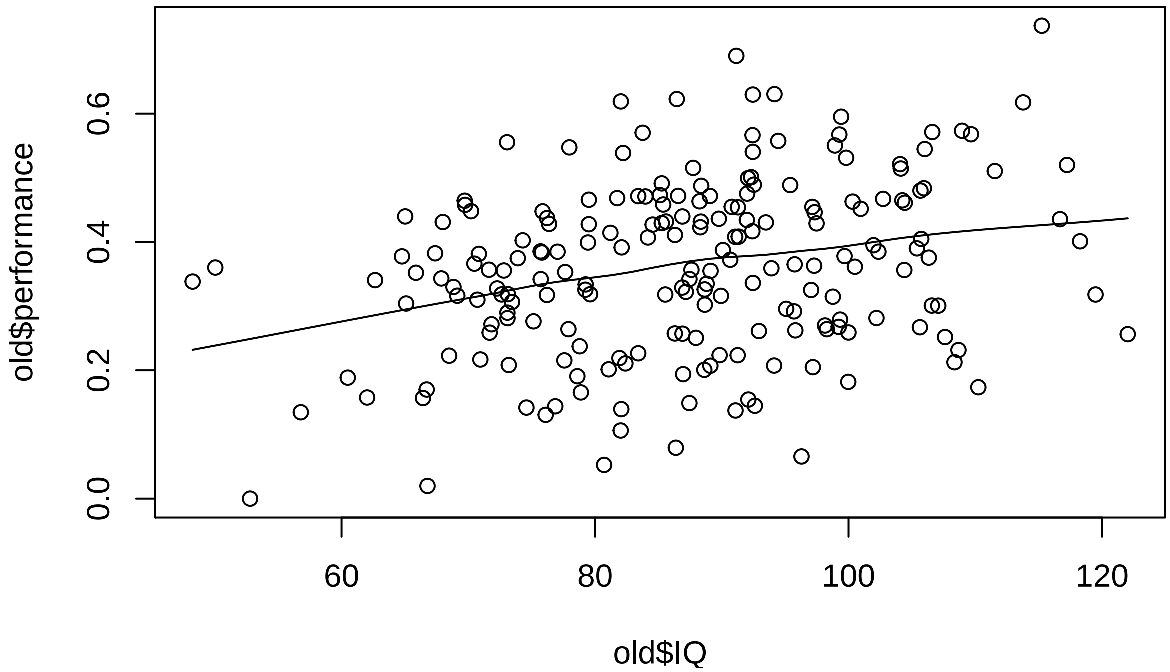



```r
corPlot(data.matrix(data[-6]))
```



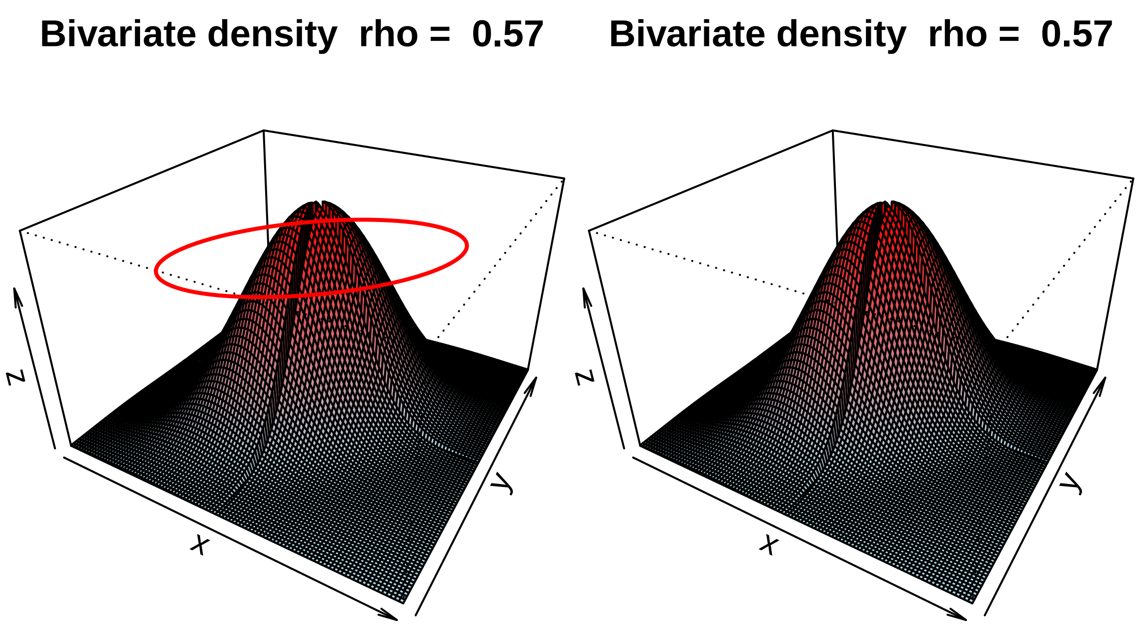

```r
# Look for outliers
# outlier(data[, c("performance", "IQ", "responseTime")],cex=.8)

# Filter your data
```

```r
dataScrubbed = scrub(data,3:4,min=c(88,1300), max=c(115, 1600), newvalue=NA)
old = subset(data,data$age>50)
scatter.smooth(old$IQ, old$performance)
```



Correlations

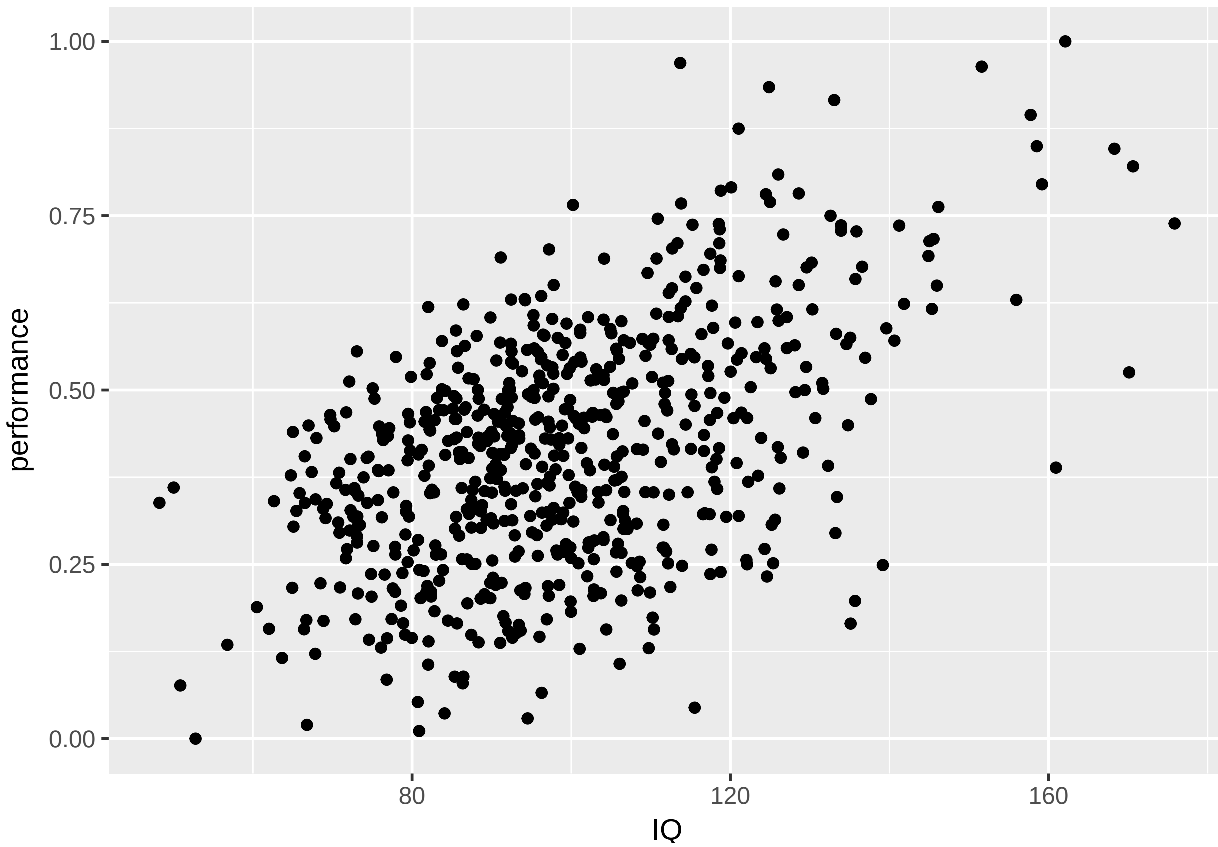
```r
# What is a correlation ?
draw.cor(expand=20,cuts=c(0,0),r = 0.57)
```



Plot Performance as a function of IQ. Performance = f(IQ).

```r
ggplot(data, aes(x=IQ, y=performance, color=isOlder)) + geom_point()
```

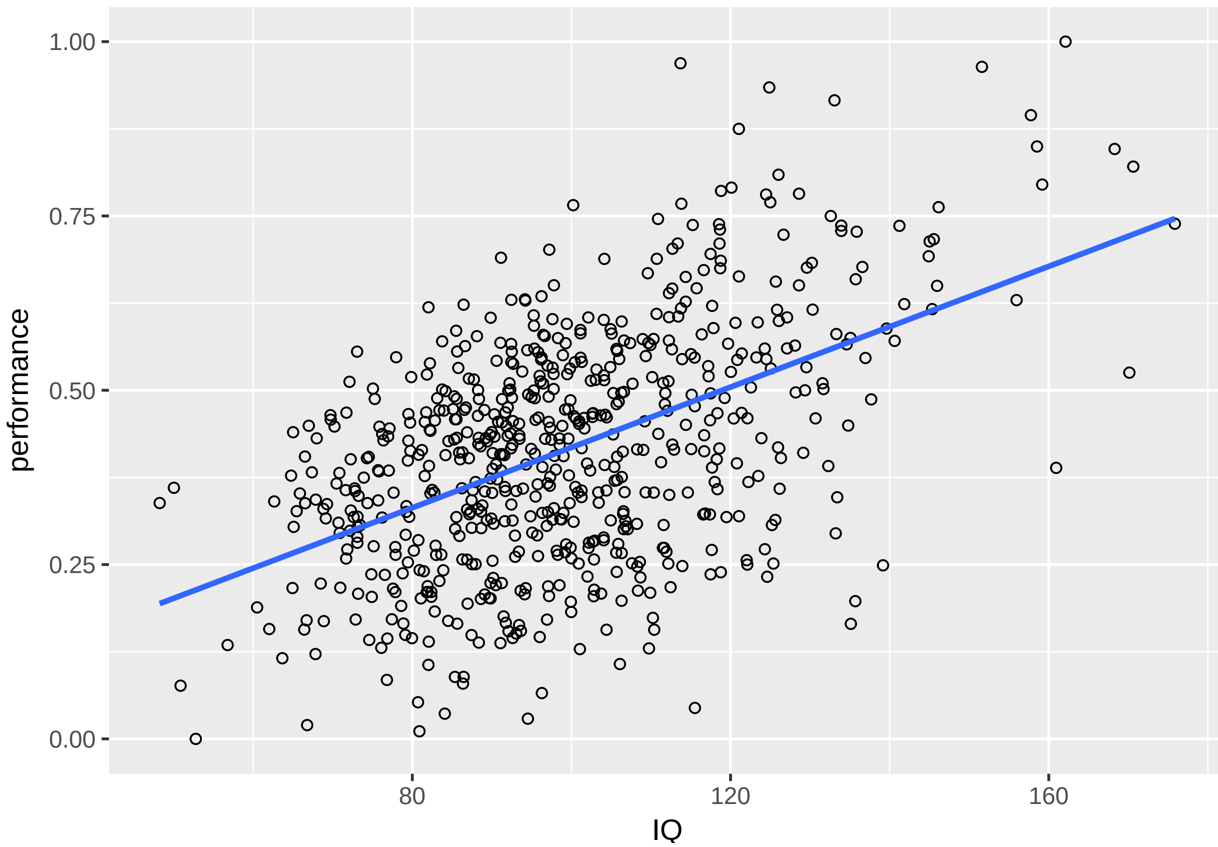

```r
ggplot(data, aes(x=IQ, y=performance)) + geom_point()
```



performance = beta*IQ + intercept + bruit

Now, what kind of relation do you think exist between the two?

```
ggplot(data, aes(x=IQ, y=performance)) +
    geom_point(shape=1) +     # Use hollow circles
    geom_smooth(method=lm,    # Add linear regression line
                se=FALSE)
```



Usually you want to know the interaction between multiple variables. For example the performance might be linked to the IQ, the Age, and/or the university !

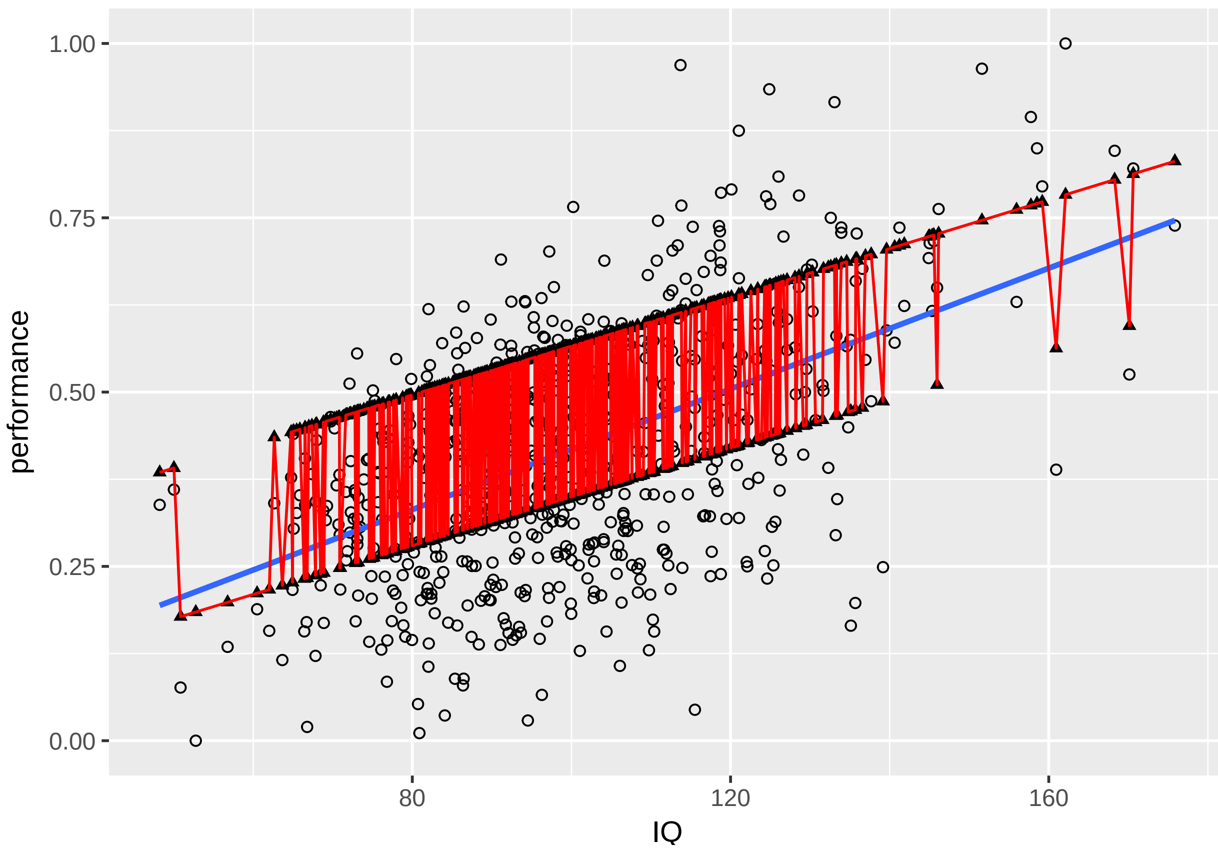
performance = beta$IQ + age + university + IQ$age + intercept + bruit

```r
# regression
regression = lm(formula = performance ~ IQ + age + university + IQ:age, data = data, na.action = na.omit
summary(regression)
```

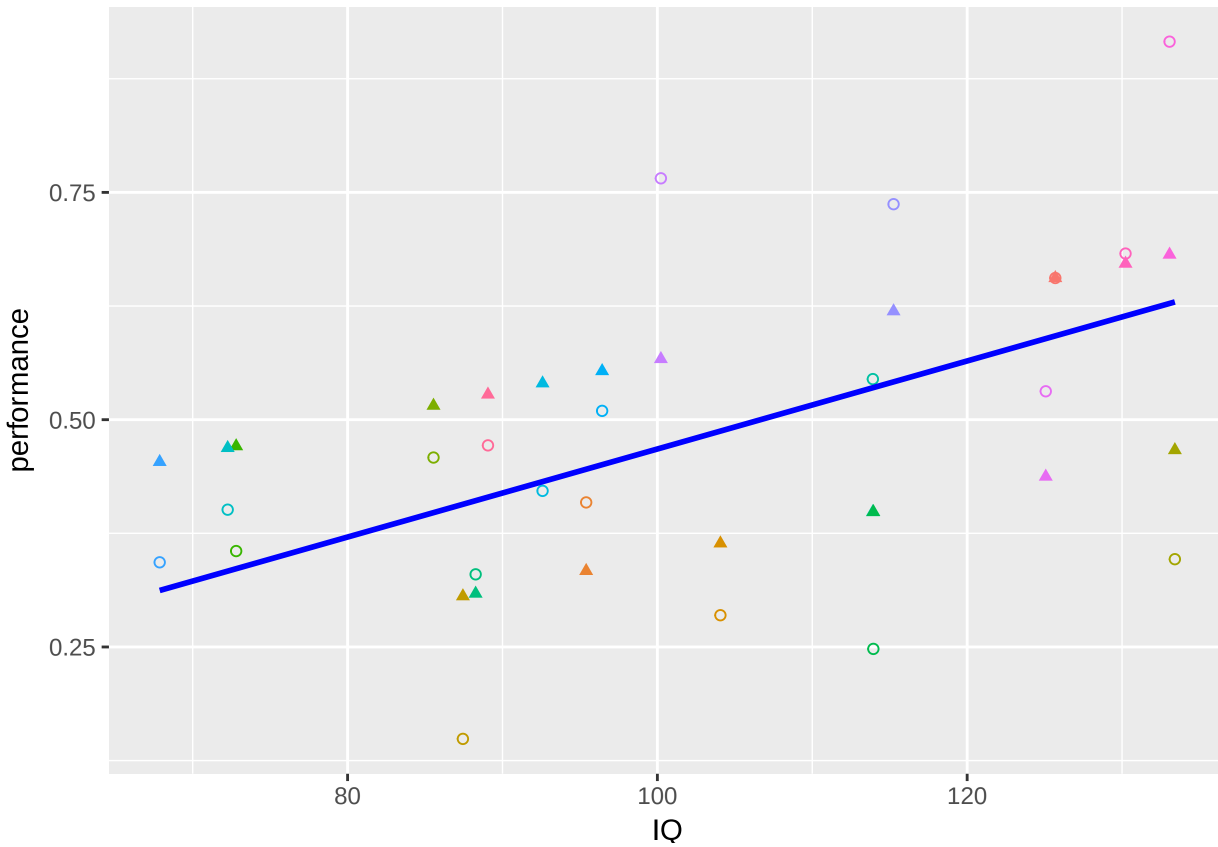
```
##
## Call:
## lm(formula = performance ~ IQ + age + university + IQ:age, data = data,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31793 -0.06466  0.00050  0.06549  0.40516
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.531e-01  6.511e-02  -2.351    0.019 *
## IQ           4.679e-03  5.885e-04   7.950 8.36e-15 ***
## age          8.188e-04  1.518e-03   0.539    0.590
## university   2.148e-01  8.225e-03  26.116  < 2e-16 ***
## IQ:age      -1.543e-05  1.513e-05  -1.020    0.308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1018 on 645 degrees of freedom
## Multiple R-squared:  0.6423, Adjusted R-squared:  0.6401
## F-statistic: 289.6 on 4 and 645 DF,  p-value: < 2.2e-16
```

```
data$prediction = 3.497e-03 * data$IQ + 2.162e-01 * data$university

ggplot(data, aes(x=IQ, y=performance)) +
    geom_point(shape=1) +    # Use hollow circles
    geom_smooth(method=lm,   # Add linear regression line
                se=FALSE)+
    geom_point(aes(x=IQ, y=prediction), shape=17)+
    geom_line(aes(x=IQ, y=prediction), color="red")
```




```
ggplot(data[1:20,], aes(x=IQ, y=performance, color=id)) +
    geom_point(shape=1)+    # Use hollow circles
    geom_smooth(method=lm,   # Add linear regression line
                se=FALSE, color="blue")+
    geom_point(aes(x=IQ, y=prediction, color=id), shape=17)+ theme(legend.position="none")
```




```r
# But is this model the good one ?
regression2 = lm(formula = performance ~ IQ  , data = data, na.action = na.omit)
summary(regression2)
```

```
##
## Call:
## lm(formula = performance ~ IQ, data = data, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44078 -0.10631  0.01573  0.10665  0.49161
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0148013  0.0290827  -0.509    0.611
## IQ           0.0043281  0.0002874  15.060   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1462 on 648 degrees of freedom
## Multiple R-squared:  0.2593, Adjusted R-squared:  0.2581
## F-statistic: 226.8 on 1 and 648 DF,  p-value: < 2.2e-16
```

```r
# Anova(model1,model2,test="Chisq") allows to compare models
anova(regression,regression2,test="Chisq")
```

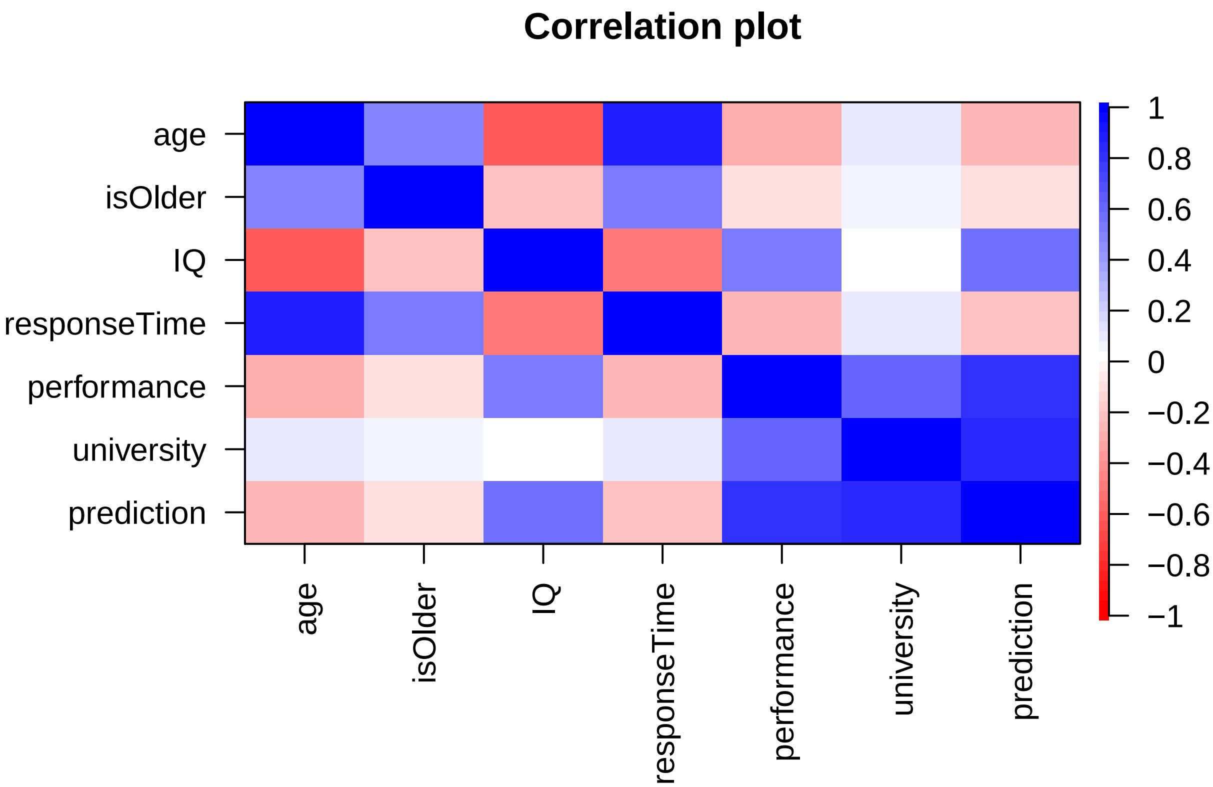
```
## Analysis of Variance Table
##
```

```
## Model 1: performance ~ IQ + age + university + IQ:age
## Model 2: performance ~ IQ
##   Res.Df    RSS Df Sum of Sq  Pr(>Chi)
## 1    645  6.690
## 2    648 13.855 -3   -7.1649 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

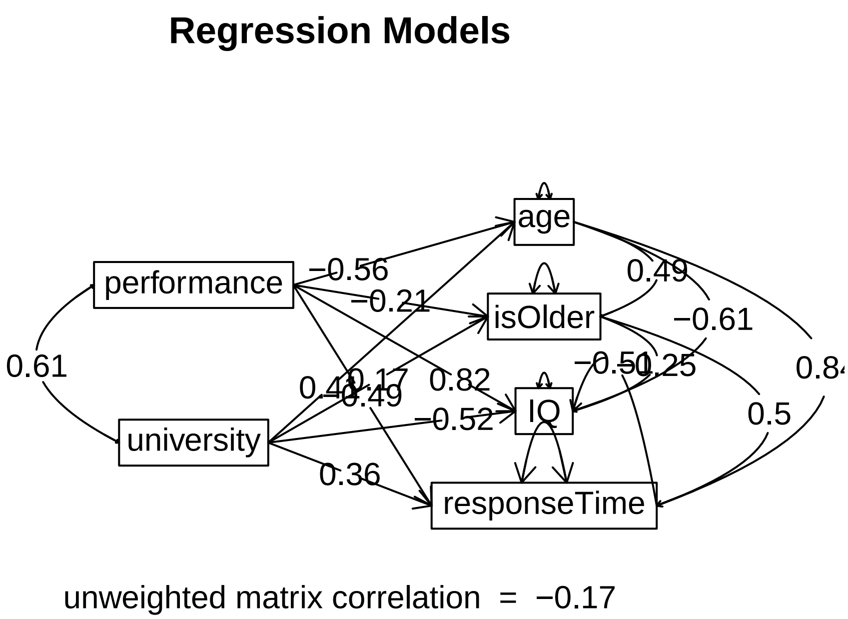
```r
# Other way to analyses correlation (usually through covariance matrix)
# correlation and analysis
corr.test(data.matrix(data[-6]))
```

```
## Call:corr.test(x = data.matrix(data[-6]))
## Correlation matrix
##                age isOlder    IQ responseTime performance university
## age           1.00    0.49 -0.61         0.84       -0.31       0.07
## isOlder       0.49    1.00 -0.25         0.50       -0.11       0.04
## IQ           -0.61   -0.25  1.00        -0.51        0.51      -0.02
## responseTime  0.84    0.50 -0.51         1.00       -0.27       0.06
## performance  -0.31   -0.11  0.51        -0.27        1.00       0.61
## university    0.07    0.04 -0.02         0.06        0.61       1.00
## prediction   -0.28   -0.10  0.54        -0.23        0.79       0.83
##              prediction
## age               -0.28
## isOlder           -0.10
## IQ                 0.54
## responseTime      -0.23
## performance        0.79
## university         0.83
## prediction         1.00
## Sample Size
## [1] 650
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##               age isOlder   IQ responseTime performance university
## age          0.00    0.00 0.00          0.0        0.00       0.23
## isOlder      0.00    0.00 0.00          0.0        0.03       0.56
## IQ           0.00    0.00 0.00          0.0        0.00       0.62
## responseTime 0.00    0.00 0.00          0.0        0.00       0.29
## performance  0.00    0.01 0.00          0.0        0.00       0.00
## university   0.06    0.28 0.62          0.1        0.00       0.00
## prediction   0.00    0.01 0.00          0.0        0.00       0.00
##              prediction
## age                0.00
## isOlder            0.05
## IQ                 0.00
## responseTime       0.00
## performance        0.00
## university         0.00
## prediction         0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

```r
corPlot(data.matrix(data[-6]))
```




```r
# Multivariate correlation predict y columns with x
setCor(y = 1:4,x=c(5,7),data=data)
```



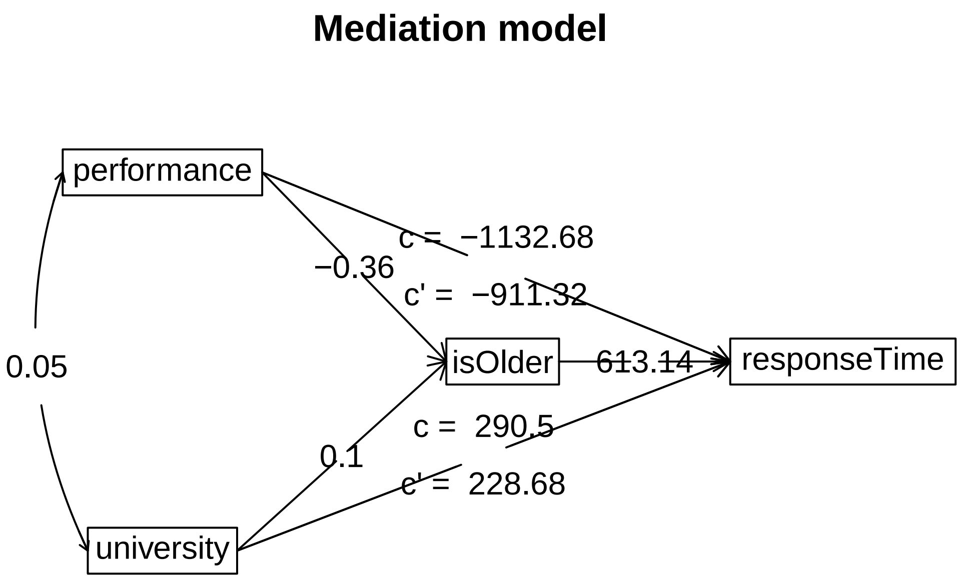

```
## Call: setCor(y = 1:4, x = c(5, 7), data = data)
##
## Multiple Regression from raw data
##
## Beta weights
##                age  isOlder    IQ responseTime
## performance -0.56    -0.21  0.82        -0.49
```

```
## university   0.41    0.17 -0.52         0.36
##
## Multiple R
##          age      isOlder           IQ responseTime
##         0.45         0.18         0.66         0.40
## multiple R2
##          age      isOlder           IQ responseTime
##        0.204        0.031        0.430        0.157
##
##  Unweighted multiple R
##          age      isOlder           IQ responseTime
##         0.21         0.09         0.29         0.19
##  Unweighted multiple R2
##          age      isOlder           IQ responseTime
##         0.05         0.01         0.09         0.04
##
##  SE of Beta weights
##              age isOlder   IQ responseTime
## performance 0.04    0.05 0.04         0.05
## university  0.04    0.05 0.04         0.05
##
##  t of Beta Weights
##                age isOlder     IQ responseTime
## performance -12.72   -4.41  22.09       -10.84
## university    9.40    3.55 -13.93         8.01
##
## Probability of t <
##             age isOlder IQ responseTime
## performance   0 1.2e-05  0      0.0e+00
## university    0 4.1e-04  0      5.6e-15
##
##  Shrunken R2
##          age      isOlder           IQ responseTime
##        0.202        0.028        0.428        0.155
##
## Standard Error of R2
##          age      isOlder           IQ responseTime
##        0.028        0.013        0.029        0.026
##
## F
##          age      isOlder           IQ responseTime
##        83.15        10.34       244.15        60.33
##
## Probability of F <
##          age      isOlder           IQ responseTime
##     0.00e+00     3.81e-05     0.00e+00     0.00e+00
##
##  degrees of freedom of regression
## [1]   2 647
##
## Various estimates of between set correlations
## Squared Canonical Correlations
## [1] 0.435 0.006
## Chisq of canonical correlations
```

```
## [1] 368.8   3.9
##
##  Average squared canonical correlation =  0.22
##  Cohen's Set Correlation R2 =  0.44
##  Shrunken Set Correlation R2 =  0.43
##  F and df of Cohen's Set Correlation  53.55 8 1280
## Unweighted correlation between the two sets =  -0.17
```

Mediation analysis

```r
mediate(y = 4, x = c(5,7), m = 2, data = data)
```

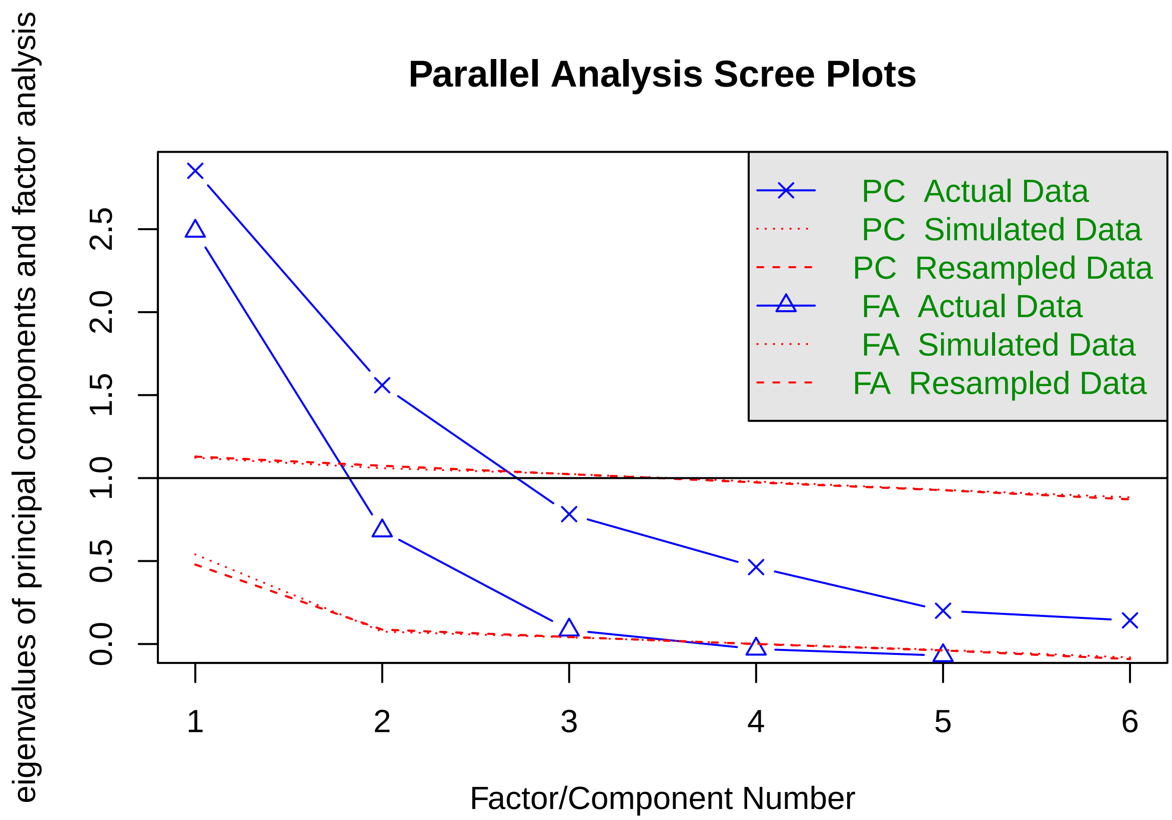



```
## Call: mediate(y = 4, x = c(5, 7), m = 2, data = data)
##
## The DV (Y) was  responseTime . The IV (X) was  performance university . The mediating variable(s) =
##
## Total Direct effect(c) of  performance  on  responseTime  =  -1132.68   S.E. =  104.52  t direct =  
## Direct effect (c') of  performance  on  responseTime  removing  isOlder  =  -911.32   S.E. =  93.14
## Indirect effect (ab) of  performance  on  responseTime  through  isOlder   =  -221.36
## Mean bootstrapped indirect effect =  -220.99  with standard error =  42.4  Lower CI =  -305.48    Up
##
## Total Direct effect(c) of  university  on  responseTime  =  290.5   S.E. =  36.29  t direct =  8.01
## Direct effect (c') of  university  on  NA  removing  isOlder  =  228.68   S.E. =  32.17  t direct =
## Indirect effect (ab) of  university  on  responseTime  through  isOlder   =  61.82
## Mean bootstrapped indirect effect =  -220.99  with standard error =  42.4  Lower CI =  29    Upper C
## R2 of model =  0.35
##  To see the longer output, specify short = FALSE in the print statement
##
##  Full output
##
##  Total effect estimates (c)
##              responseTime     se      t     Prob
## performance     -1132.68 104.52 -10.84 0.00e+00
## university        290.50  36.29   8.01 5.55e-15
```

```
##
## Direct effect estimates     (c')
##               responseTime    se     t     Prob
## performance       -911.32 93.14 -9.78 0.00e+00
## university         228.68 32.17  7.11 3.11e-12
##
##  'a'  effect estimates
##               isOlder   se     t     Prob
## performance   -0.36 0.08 -4.41 0.000012
## university     0.10 0.03  3.55 0.000414
##
##  'b'  effect estimates
##           responseTime    se     t Prob
## isOlder       613.14 44.09 13.91    0
##
##  'ab'  effect estimates
##               responseTime    boot    sd   lower   upper
## performance       -221.36 -220.99 42.40 -305.48 -140.89
## university          61.82   61.55 16.85   29.00   95.93
```

Dimensionality reduction

```r
# How many components should you expect ?
dataWithoutId = data[-c(6,8)]
fa.parallel(dataWithoutId)
```



```
## Parallel analysis suggests that the number of factors =  3  and the number of components =  2
```

```r
# Data loads on variable
principal(dataWithoutId, nfactors = 2)
```

```
## Principal Components Analysis
## Call: principal(r = dataWithoutId, nfactors = 2)
## Standardized loadings (pattern matrix) based upon correlation matrix
##                RC1   RC2   h2   u2 com
## age           0.92 -0.06 0.86 0.14 1.0
## isOlder       0.66  0.09 0.45 0.55 1.0
## IQ           -0.71  0.31 0.60 0.40 1.4
## responseTime  0.90 -0.03 0.81 0.19 1.0
## performance  -0.31  0.89 0.88 0.12 1.2
## university    0.17  0.88 0.81 0.19 1.1
##
##                        RC1  RC2
## SS loadings           2.74 1.67
## Proportion Var        0.46 0.28
## Cumulative Var        0.46 0.74
## Proportion Explained  0.62 0.38
## Cumulative Proportion 0.62 1.00
##
## Mean item complexity =  1.1
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.1
##  with the empirical chi square  207.44  with prob <  9.4e-44
##
## Fit based upon off diagonal values = 0.94
```

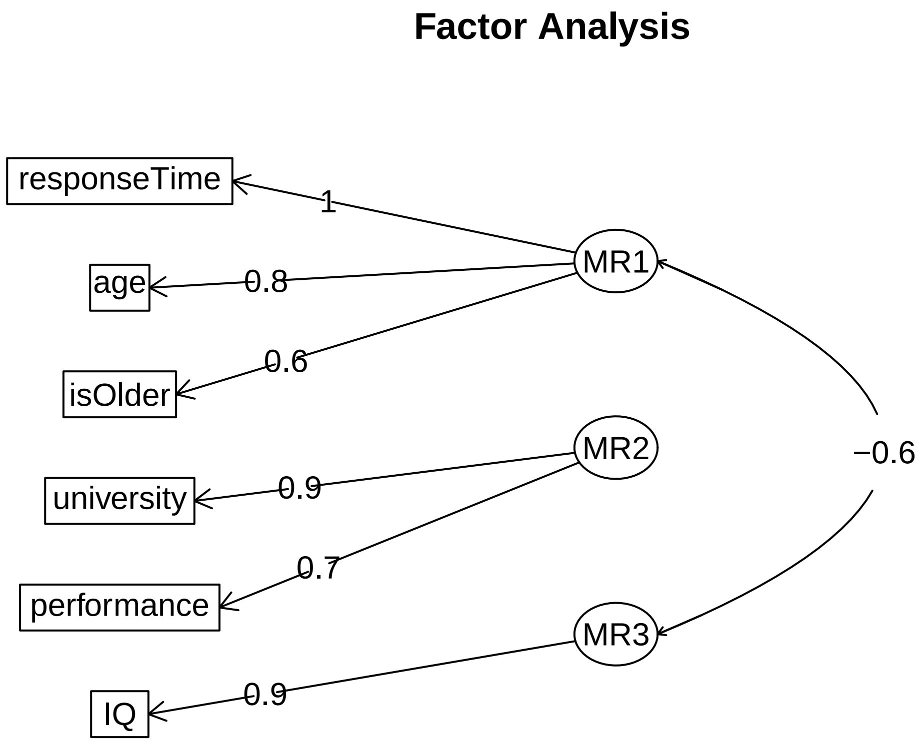
```r
# Latent variable loads on data (age, IQ, university)
factanal(dataWithoutId, factors = 3)
```

```
##
## Call:
## factanal(x = dataWithoutId, factors = 3)
##
## Uniquenesses:
##          age      isOlder           IQ responseTime  performance
##        0.148        0.700        0.068        0.142        0.005
##   university
##        0.464
##
## Loadings:
##              Factor1 Factor2 Factor3
## age           0.852          -0.354
## isOlder       0.544
## IQ           -0.347   0.112   0.894
## responseTime  0.900          -0.219
## performance  -0.172   0.902   0.390
## university            0.723
##
##                Factor1 Factor2 Factor3
## SS loadings      1.988   1.351   1.135
```

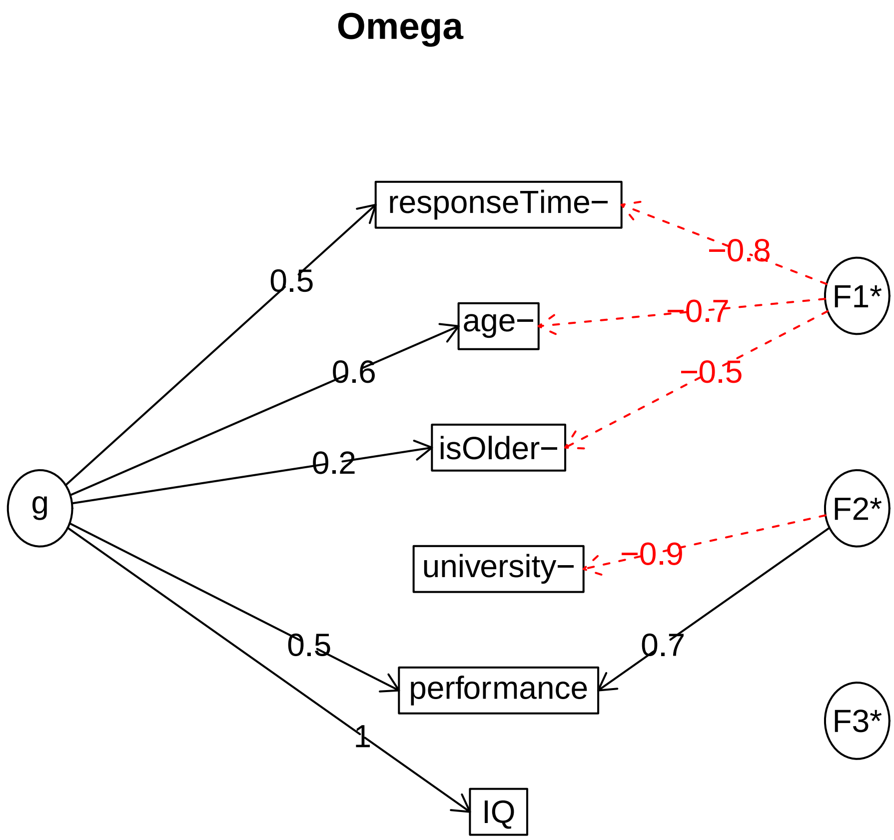
```
## Proportion Var   0.331   0.225   0.189
## Cumulative Var   0.331   0.556   0.746
##
## The degrees of freedom for the model is 0 and the fit was 6e-04
```

```r
fa.diagram(fa(dataWithoutId, nfactors = 3))
```

```
## Loading required namespace: GPArotation
```



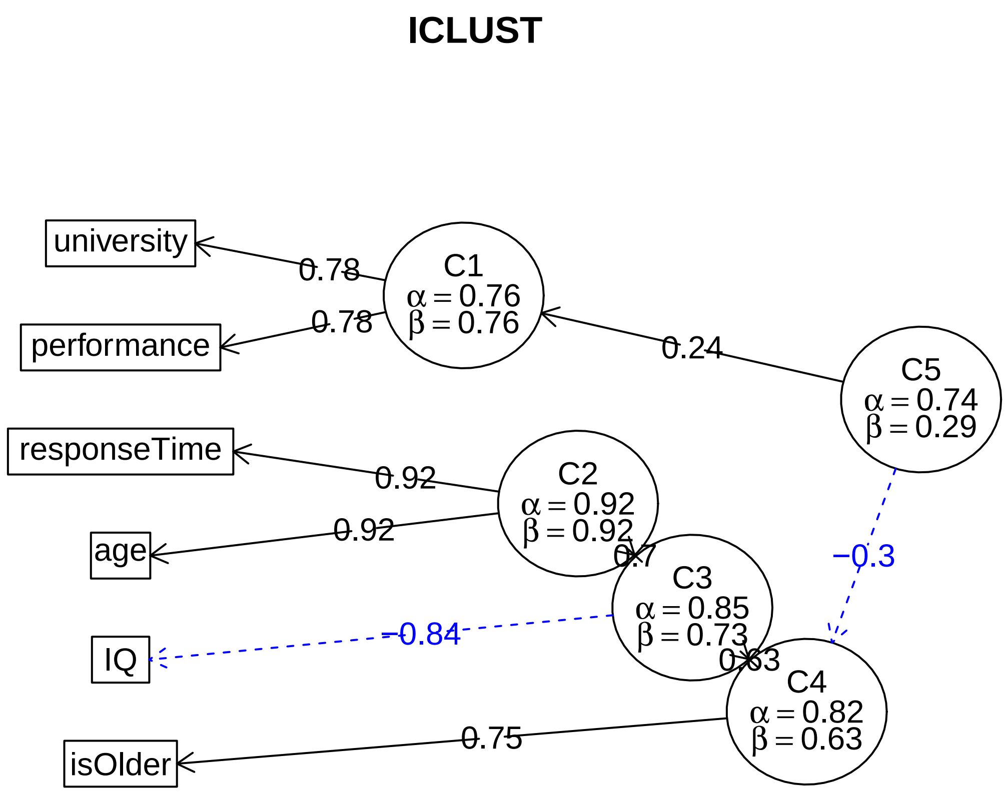

```r
# Hierarchical
#install.packages("GPArotation")
library(GPArotation)
omega(dataWithoutId)
```




```
## Omega
## Call: omega(m = dataWithoutId)
## Alpha:                 0.69
## G.6:                   0.82
## Omega Hierarchical:    0.57
## Omega H asymptotic:    0.64
## Omega Total            0.89
##
## Schmid Leiman Factor loadings greater than  0.2
##                   g   F1*   F2*   F3*   h2   u2   p2
## age-            0.61 -0.70              0.85 0.15 0.43
## isOlder-        0.23 -0.50              0.30 0.70 0.18
## IQ              0.95                    0.91 0.09 0.99
## responseTime-   0.50 -0.78              0.86 0.14 0.29
## performance     0.54        0.71        0.79 0.21 0.37
## university-            -0.87        0.75 0.25 0.00
##
## With eigenvalues of:
##    g  F1*  F2*  F3*
## 1.86 1.35 1.25 0.01
##
## general/max  1.38   max/min =   260.26
## mean percent general =  0.38    with sd =  0.34 and cv of  0.89
## Explained Common Variance of the general factor =  0.42
##
## The degrees of freedom are 0  and the fit is  0
## The number of observations was  650  with Chi Square =  1.06  with prob <  NA
```

```
## The root mean square of the residuals is  0
## The df corrected root mean square of the residuals is  NA
##
## Compare this with the adequacy of just a general factor and no group factors
## The degrees of freedom for just the general factor are 9  and the fit is  1.87
## The number of observations was  650  with Chi Square =  1207.7  with prob <  2.6e-254
## The root mean square of the residuals is  0.25
## The df corrected root mean square of the residuals is  0.33
##
## RMSEA index =  0.205  and the 90 % confidence intervals are  0.205 0.474
## BIC =  1149.41
##
## Measures of factor score adequacy
##                                                  g  F1*  F2*   F3*
## Correlation of scores with factors            0.95 0.92 0.92  0.08
## Multiple R square of scores with factors      0.91 0.84 0.84  0.01
## Minimum correlation of factor score estimates 0.82 0.69 0.68 -0.99
##
##  Total, General and Subset omega for each subset
##                                                  g  F1*  F2*  F3*
## Omega total for total scores and subscales    0.89 0.85 0.41 0.91
## Omega general for total scores and subscales  0.57 0.27 0.38 0.90
## Omega group for total scores and subscales    0.28 0.58 0.03 0.00
```

```
#Clusters
iclust(dataWithoutId)
```




```
## ICLUST (Item Cluster Analysis)
## Call: iclust(r.mat = dataWithoutId)
##
```

```
## Purified Alpha:
## [1] 0.74
##
## G6* reliability:
## [1] 0.62
##
## Original Beta:
## [1] 0.29
##
## Cluster size:
## [1] 6
##
## Item by Cluster Structure matrix:
##              [,1]
## age          -0.81
## isOlder      -0.44
## IQ            0.67
## responseTime -0.77
## performance   0.68
## university    0.26
##
## With eigenvalues of:
## [1] 2.4
##
## Purified scale intercorrelations
##  reliabilities on diagonal
##  correlations corrected for attenuation above diagonal:
##      [,1]
## [1,] 0.74
##
## Cluster fit =  0.68   Pattern fit =  0.89  RMSR =  0.21
```

```
# structural equatiom modelin
sem = esem(r = cor(dataWithoutId), varsX = c(5,6), varsY = 1:4, nfX = 2, nfY = 1,
n.obs = 650, plot = FALSE)
```

```
## The estimated weights for the factor scores are probably incorrect.  Try a different factor extracti
```

```
print(sem)
```

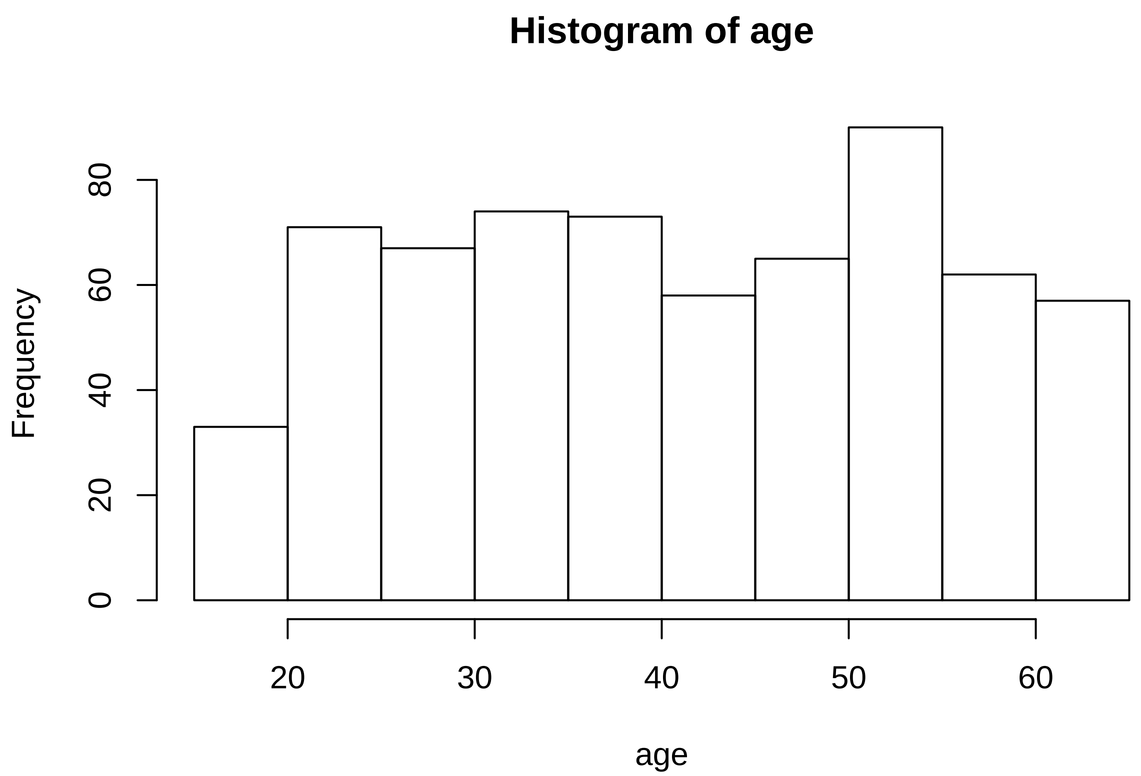
```
## Exploratory Structural Equation Modeling  Analysis using method =  minres
## Call: esem(r = cor(dataWithoutId), varsX = c(5, 6), varsY = 1:4, nfX = 2,
##     nfY = 1, n.obs = 650, plot = FALSE)
##
## For the 'X' set:
##                MR1  MR2
## IQ           -0.87 0.49
## responseTime  0.87 0.49
##
## For the 'Y' set:
##                MR1
## performance  1.00
## university   0.61
## age         -0.31
## isOlder     -0.11
```

```
##
## Correlations between the X and Y sets.
##       X1   X2    Y1
## X1  1.00 0.00 -0.49
## X2  0.00 1.00  0.12
## Y1 -0.49 0.12  1.00
##
## The degrees of freedom for the null model are  30  and the empirical chi square  function was  3543.5
## The degrees of freedom for the model are 0  and the empirical chi square function was  19.12
##   with prob <  NA
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  NA
##  with the empirical chi square  19.12  with prob <  NA
## The total number of observations was  650  with fitted Chi Square =  545.16  with prob <  NA
##
## Empirical BIC =  NA
## ESABIC =  NA
## Fit based upon off diagonal values = 0.99
## To see the item loadings for the X and Y sets combined, and the associated fa output, print with  sho
```

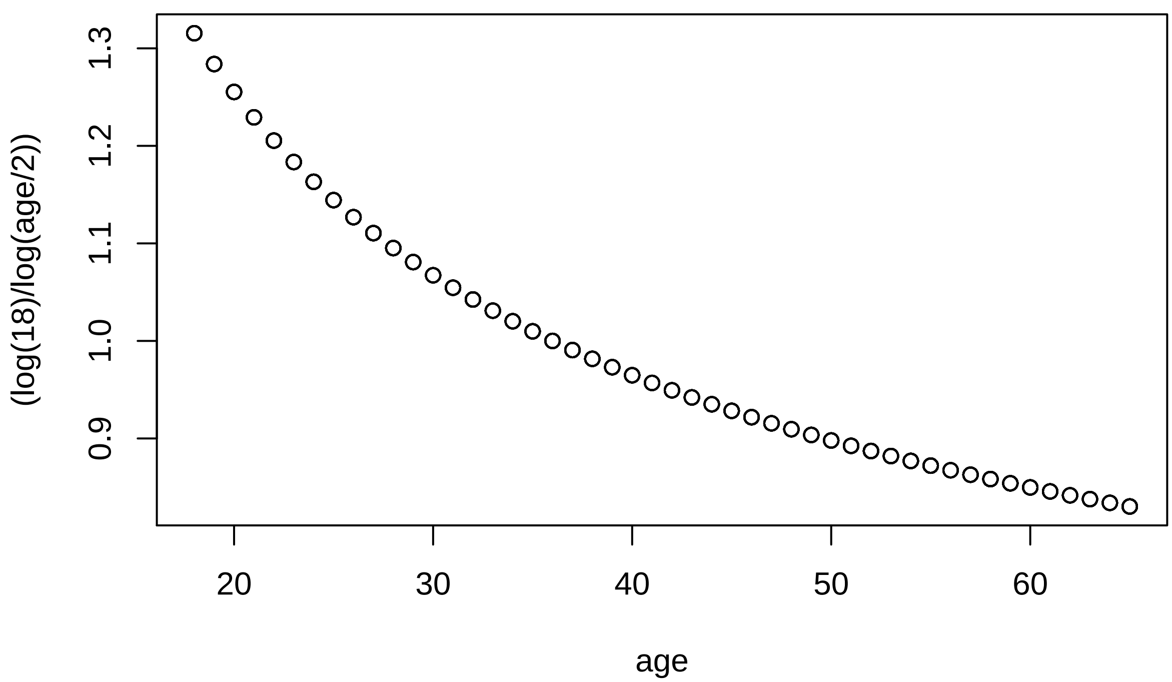
```
esem.diagram(sem)
```

How to generate fake data

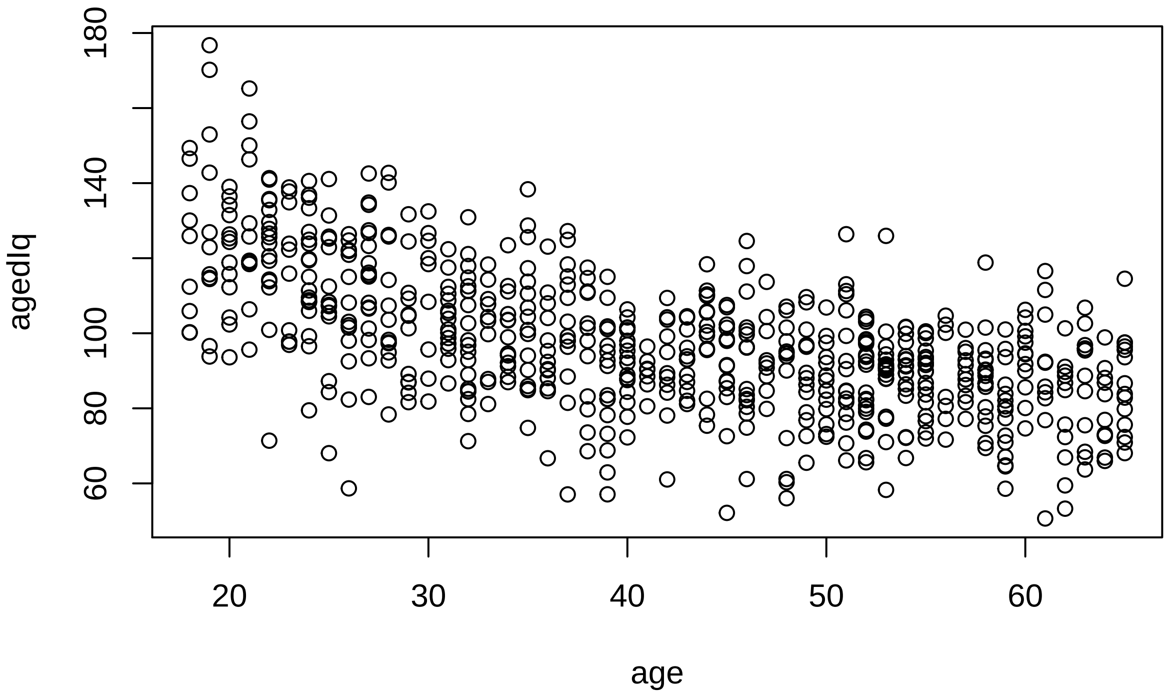
```
# Number of subjects
N = 650

# Sample from a list
age = sample(18:65, N, replace = T)

hist(age)
```




```
# Sample from gaussian distribution to generate IQ
baseIq = rnorm(N, 100, 15)
agedIq = baseIq * (log(18)/log(age/2))
data = data.frame(age = age, isOlder = age>60, IQ = agedIq)

plot(age, (log(18)/log(age/2)))
```

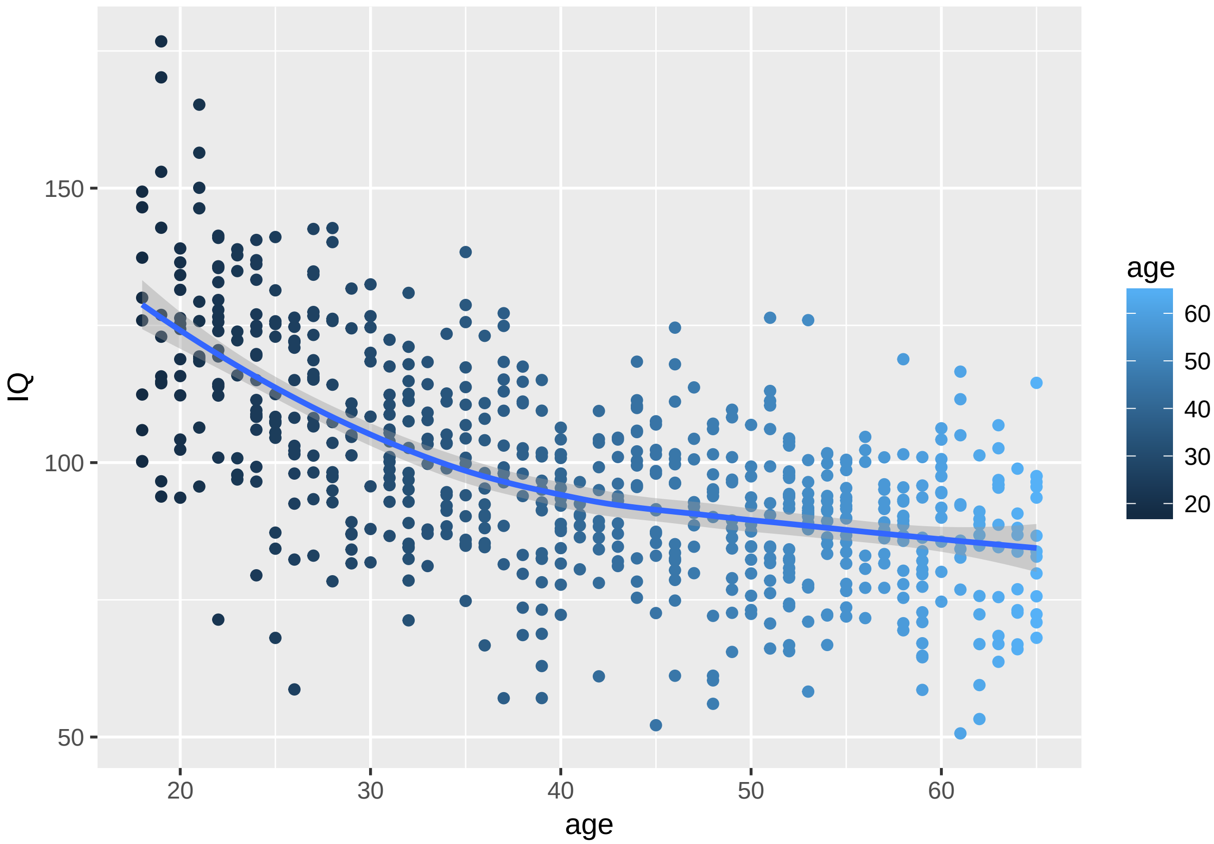


```r
plot(age,agedIq)
```

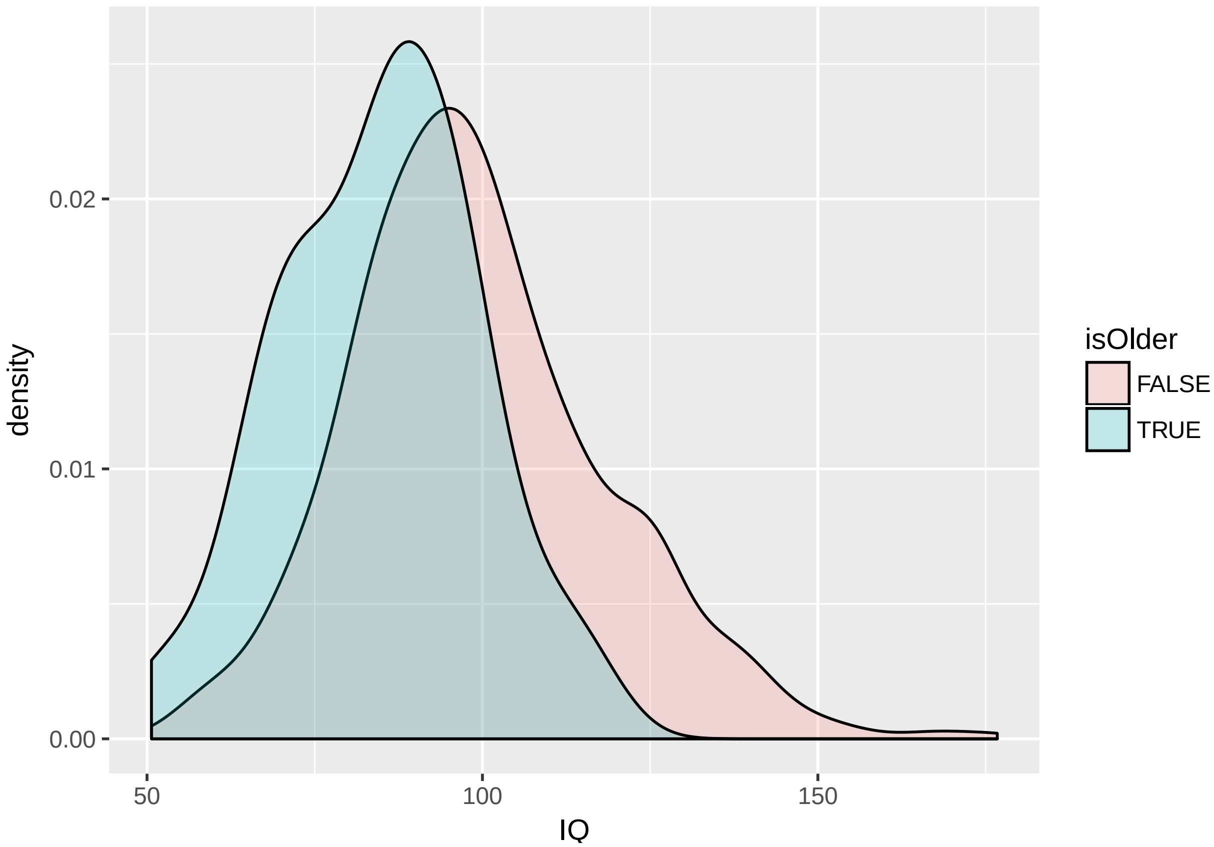



```r
# A slightly more beatifull way of plotting
# install.packages("ggplot")
library(ggplot2)

# Line plot + prediction error
ggplot(data, aes(x = age, y = IQ, color=age)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```




```
# Histograms
ggplot(data, aes(IQ, ..density.., fill = isOlder)) +
geom_density(alpha=0.2)
```



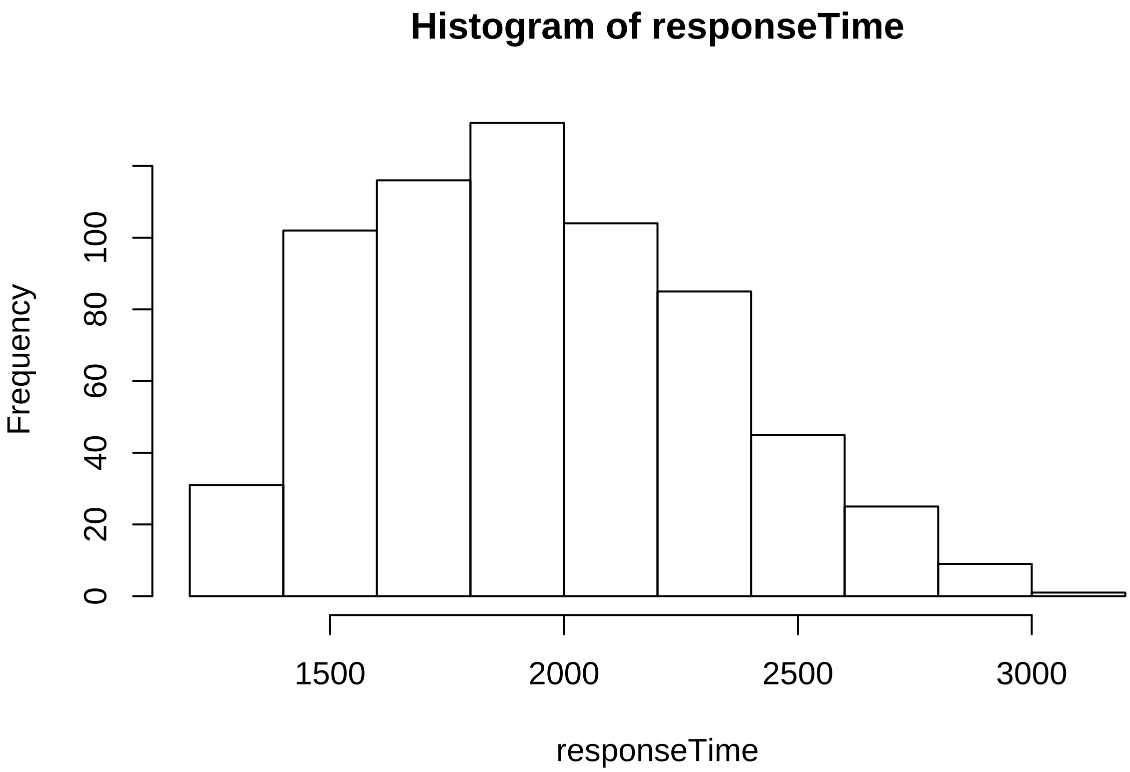

```r
# Sample from a custom distribution
# For example the Exponential
# P(x) = lambda * exp (-lambda*x)
# lambda = 1 / mean
# CDF = 1 - exp(-lambda*x) ==> x = - ln (1-CDF) / lambda
mean = 1500
lambda = 1 / mean
CDF = sample(0:100, N, replace = T) / 100
x = -log(1-CDF) / lambda
```

Complex distribution

Sample from from complex distribution such as Diffusion Model for Response Time Lets image you are slower with age !

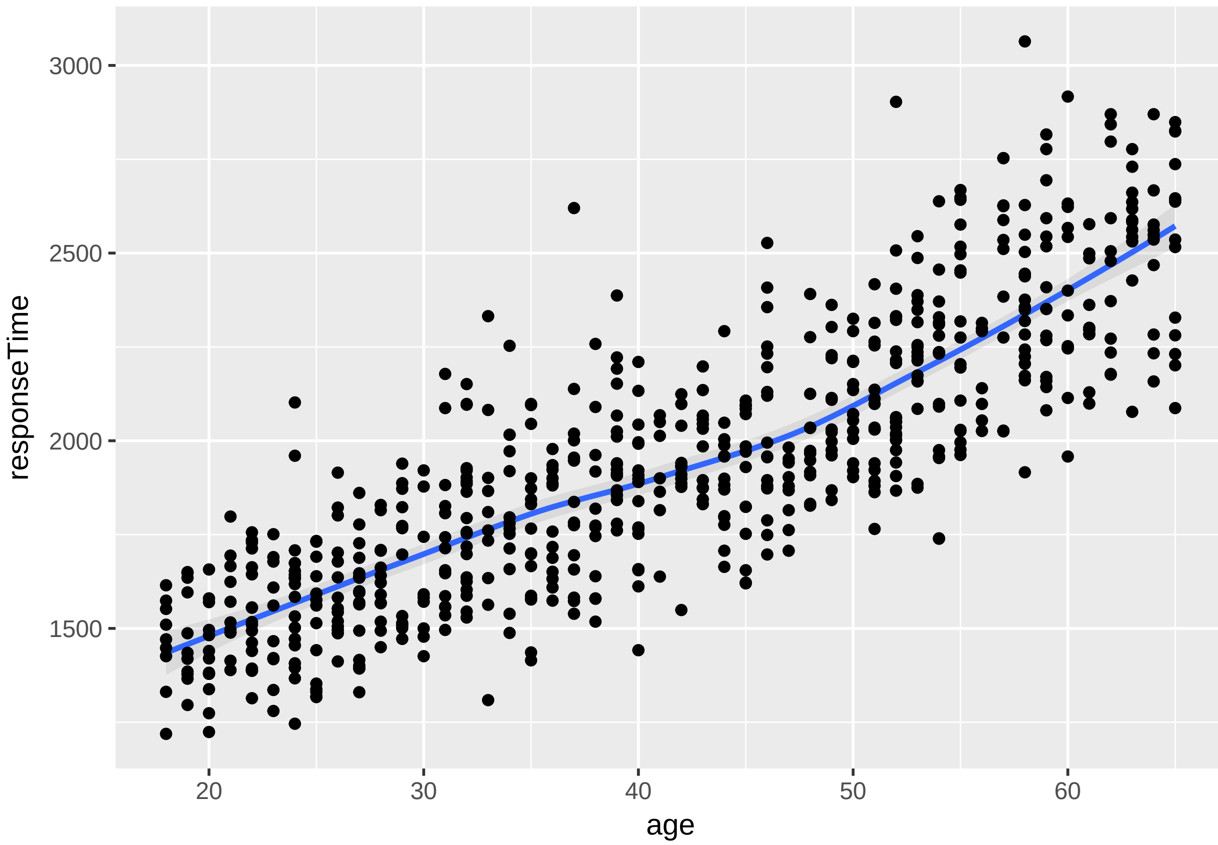
```r
# Here bias depends on age
responseTime = rep(NA, N)
currentAccumulators = rep(0,N)
bias = sample(c(-1,1), N, T)+rnorm(N,0,0.1)
threshold = 1000
for(i in 1:5000) {
  currentAccumulators = currentAccumulators + bias * abs(rnorm(N, 0,1-(0.008*age)))
  indices = which(abs(currentAccumulators) > threshold)
  newIndices = indices[!indices%in%which(!is.na(responseTime))]
  responseTime[newIndices] = i
}
```

```r
hist(responseTime)
```



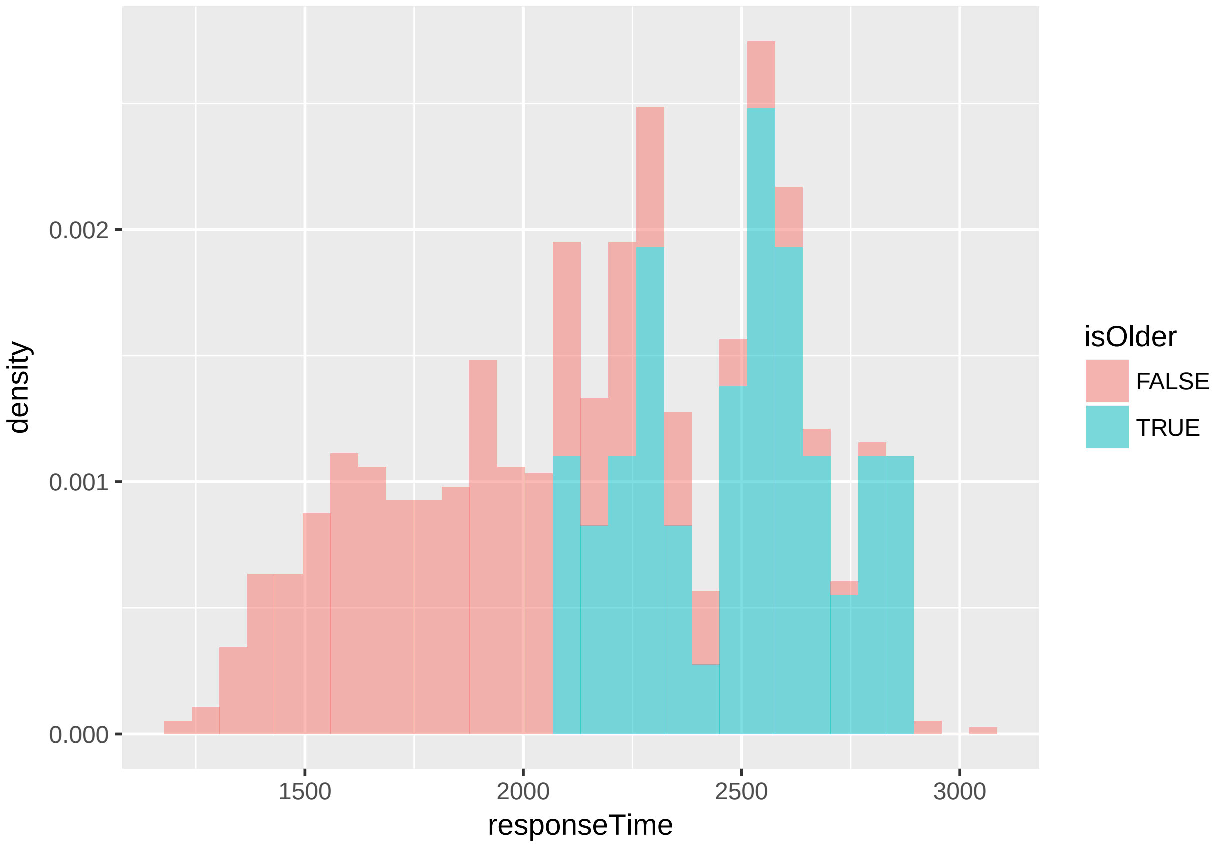

```r
data$responseTime = responseTime

ggplot(data, aes(age, responseTime)) + geom_smooth(alpha=0.2) + geom_point()
```

```
## `geom_smooth()` using method = 'loess'
```




```
ggplot(data, aes(responseTime, ..density.., fill=isOlder)) + geom_histogram(alpha=0.5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Create fake correlation in the data

For example imagine IQ is linked to performance and wether or not they went to university

```
# University
university = rep(0,N)
university[which(rnorm(N)>-0.25)] = 1

# Generate fake performance linked to IQ
performance = rnorm(N,60,15)*agedIq/100 + 30*university

# Scale between 0 and 1
performance = performance - min(performance)
performance = performance / max(performance)

data$performance = performance
data$id = getRandomId(N)
data$age = age
data$university =university
```

Save your data

```
# Set your working directory - your reference for the file system
setwd("~/Google Drive/Master Students/courses/introduction_a_r")
```

```r
# Save into your data/raw
save(data, file = "data/raw/data.RData")

# CSV format
write.csv(data, file = "data/raw/data.csv", row.names = FALSE)

# XLS - needs a library
library(WriteXLS)
WriteXLS(data, "data/raw/data.xls")
```

Example: Scaling function

```r
scale_by <- function (data, by = "minmax") {
        copied_data = data.frame(data)
        columns = names(data)

        switch(by,
               meanvar = {
                       center <- mean
                       spread <- sd
               },
               medianvar = {
                       center <- median
                       spread <- sd
               },
               minmax = {
                       center <- min
                       spread <- max
               })

        for (column in columns) {
                center_by = center(copied_data[[column]], na.rm = T)
                reduced_by = spread(copied_data[[column]], na.rm = T)
                copied_data[[column]] = (copied_data[[column]] - center_by) / reduced_by
        }
        return(copied_data)
}
```

Debugging

A good rule of thumb is that you will introduce a bug in your code every ten new lines. Besides real design issue in your code, bugs are usually due to overlooking certain extreme "use case" that you did not plan for, or because you did not pay attention to the values of your variables and content of your data.

Quick fix bugs: - Wrong variable name - Forgot a comma or parenthesis

Harder bugs: - Wrong format + Using a factor as a string or numeric + Some NA in the data + Some misformated strings or outlier values - Wrong data size - Silent bugs - return incorect values (+/- design issues)

Three steps to correct a bug: - Reproduce - Isolate : harder part - Correct

```
# Break on error
# browser()
# breakpoints
# print
```