Basic Concept of Statistics

Paolo Girardi and Livio Finos

15/10/2020

Contents

Ph.D. Course in Neuroscience
Calendar of the Basic Courses – Academic Year 2020-2021
Basic Concept of Statistics
Lesson 2 - Optional and preliminary course on use of R

Descriptive Statistics

In R some useful functions for the descriptive analysis are:

- plot(x, y): bivariate plot of x (on the x-axis) and y (on the y-axis);
- hist(x): histogram of the frequencies of x
- barplot(x): histogram of the values of x; use horiz=FALSE for horizontal bars
- dotchart(x): if x is a data frame, plots a Cleveland dot plot (stacked plots line-by- line and column-by-column)
- pie(x): circular pie-chart
- boxplot(x): box-and-whiskers plot
- stripplot(x): plot of the values of x on a line (an alternative to boxplot() for small sample sizes)
- mosaicplot(x): mosaic plot from frequencies in a contingency table
- qqnorm(x): quantiles of x with respect to the values expected under a normal law

Univariate Statistical Analysis with R

We import the dataset test.csv

```
test<-read.csv("test.csv",sep=";",header=T,dec=",")
head(test) #the first 6 rows
```

```
##   ID Age  BMI Gender     Education ACT SATV SATQ Stress Social
## 1  1  19 24.3      F     secondary  24  500  500      2      3
## 2  2  23 24.6      F     secondary  35  600  500      1      6
## 3  3  20 28.1      F     secondary  21  480  470      6      2
## 4  4  27 24.5      M        degree  26  550  520      1      3
## 5  5  33 24.1      M upper primary  31  600  550      5      2
## 6  6  26 23.1      M   post-degree  28  640  640      6      1
```

```
str(test)
```

```
## 'data.frame':    150 obs. of  10 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age      : int  19 23 20 27 33 26 30 19 23 40 ...
##  $ BMI      : num  24.3 24.6 28.1 24.5 24.1 23.1 23.2 21.9 27.3 24.1 ...
##  $ Gender   : Factor w/ 2 levels "F","M": 1 1 1 2 2 2 1 2 1 1 ...
##  $ Education: Factor w/ 6 levels "degree","lower primary",..: 5 5 5 1 6 3 3 5 1 3 ...
##  $ ACT      : int  24 35 21 26 31 28 36 22 22 35 ...
##  $ SATV     : int  500 600 480 550 600 640 610 520 400 730 ...
##  $ SATQ     : int  500 500 470 520 550 640 500 560 600 800 ...
##  $ Stress   : int  2 1 6 1 5 6 5 4 4 4 ...
##  $ Social   : int  3 6 2 3 2 1 5 2 6 5 ...
```

```
summary(test)
```

```
##        ID              Age             BMI        Gender         Education
##  Min.   :  1.00   Min.   :17.00   Min.   :19.00   F:94   degree       :55
##  1st Qu.: 38.25   1st Qu.:22.00   1st Qu.:22.80   M:56   lower primary: 3
##  Median : 75.50   Median :26.00   Median :23.90          post-degree  :43
##  Mean   : 75.50   Mean   :29.22   Mean   :23.94          primary      : 3
##  3rd Qu.:112.75   3rd Qu.:34.00   3rd Qu.:24.80          secondary    :39
##  Max.   :150.00   Max.   :65.00   Max.   :31.00          upper primary: 7
##       ACT             SATV            SATQ            Stress        Social
##  Min.   :15.00   Min.   :240.0   Min.   :250.0   Min.   :1.0   Min.   :1.000
##  1st Qu.:26.00   1st Qu.:542.5   1st Qu.:550.0   1st Qu.:2.0   1st Qu.:2.000
##  Median :29.00   Median :600.0   Median :605.0   Median :4.0   Median :3.000
##  Mean   :29.04   Mean   :608.6   Mean   :614.1   Mean   :3.6   Mean   :3.153
##  3rd Qu.:32.00   3rd Qu.:680.0   3rd Qu.:700.0   3rd Qu.:5.0   3rd Qu.:4.750
##  Max.   :36.00   Max.   :800.0   Max.   :800.0   Max.   :6.0   Max.   :6.000
```

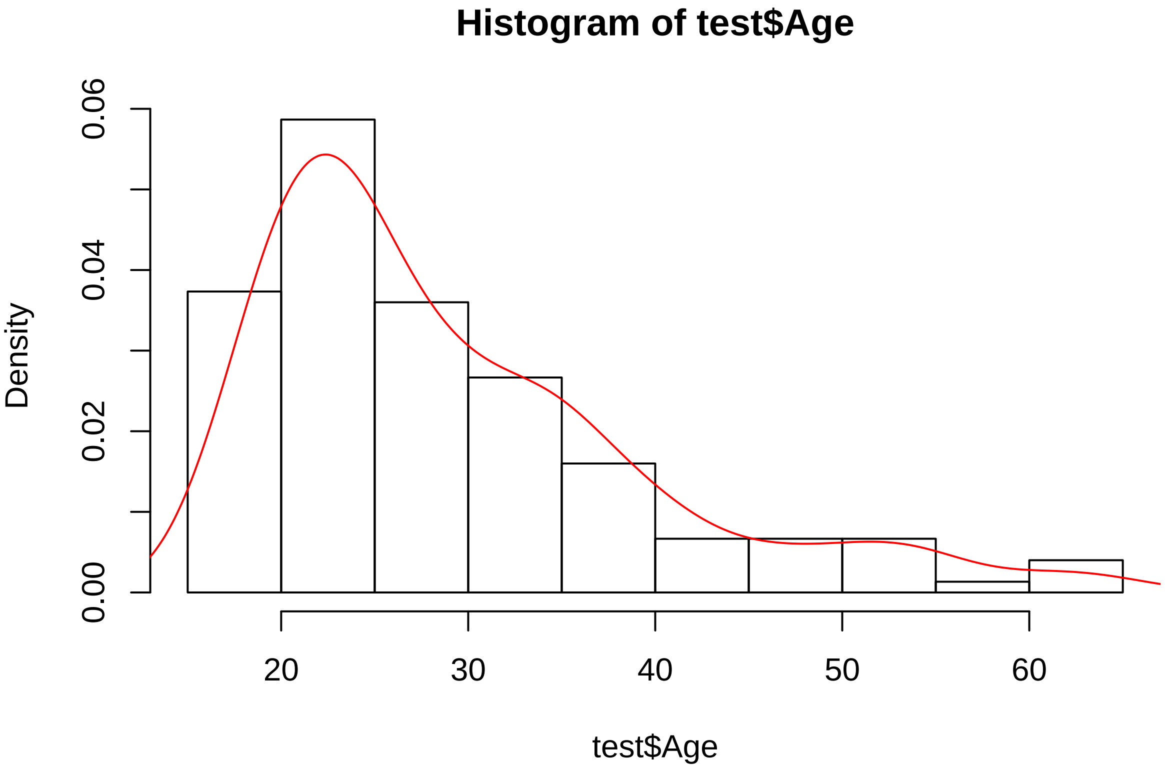
This dataset is formed by the first 150 subjects of a larger dataset. The dataset reported some information about the SAT and ACT test, performed on people during some job's selection.

Analysis of the Age variable

```
# A histogram with the density plot
summary(test$Age)
```

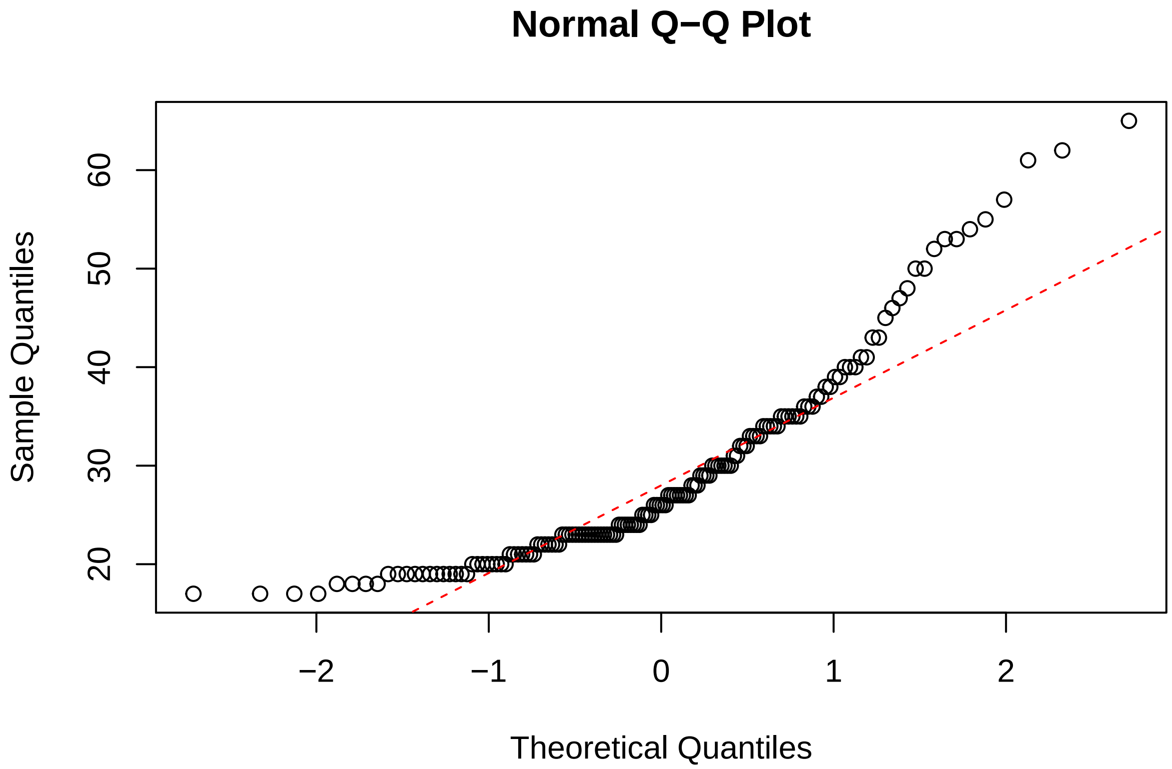
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   22.00   26.00   29.22   34.00   65.00
```

```
hist(test$Age,prob=T)
lines(density(test$Age),col=2)
```



The distribution is skewed, in particular few numbers after 40 years old. A qqplot can be used to visualise the distribution

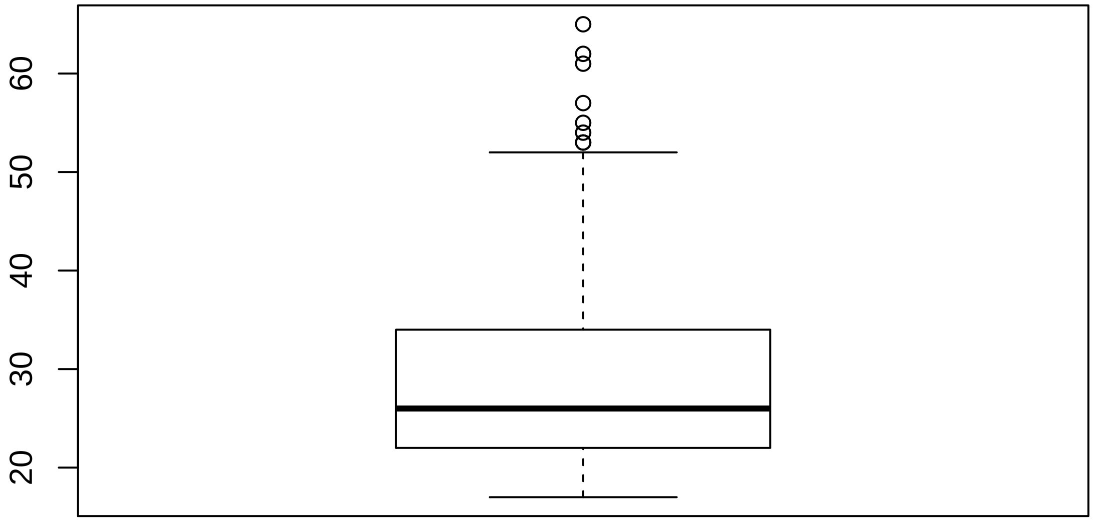
```
qqnorm(test$Age)
qqline(test$Age,col=2,lty=2)
```



The graph reports the comparison between the theoretical quantile of a Normal distribution and quantiles of the variabile Age. If the points follow the red line, a normal distribution can be assumed.

The function *boxplot()* performs (box and whiskers plot) as follows

```
boxplot(test$Age)
```



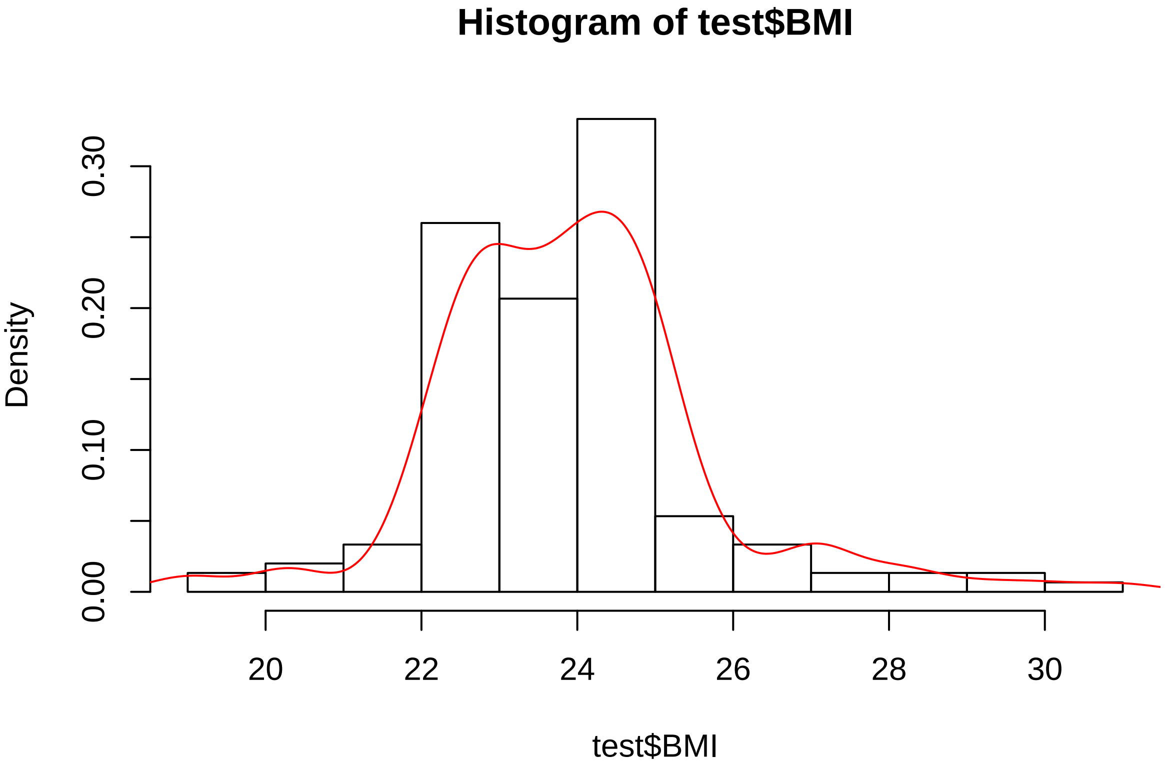
Analysis of the BMI variable

The BMI (Body Mass Index) is the ratio between weight/height.

```
# A histogram with the density plot
summary(test$BMI)
```

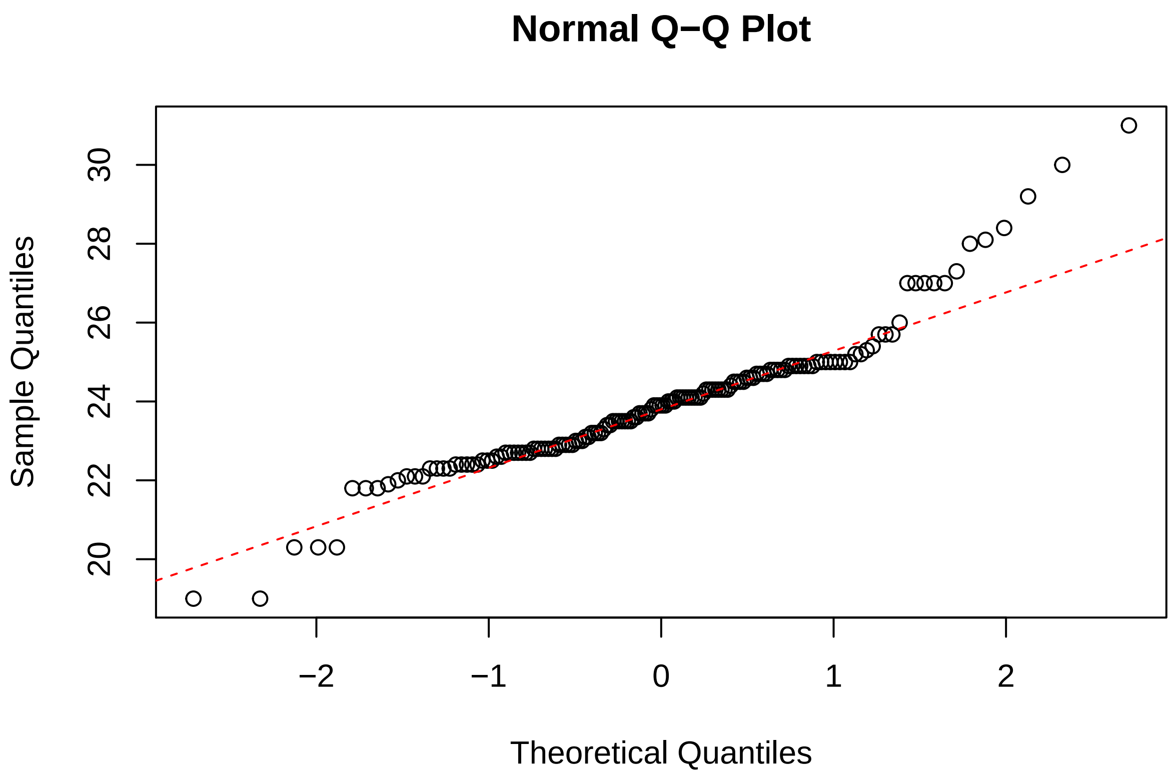
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    19.00    22.80    23.90    23.94    24.80    31.00
```

```
hist(test$BMI,prob=T)
lines(density(test$BMI),col=2)
```



The distribution looks simmetric but there is the presence of outliers (values to low and to high respect to the central cloud).

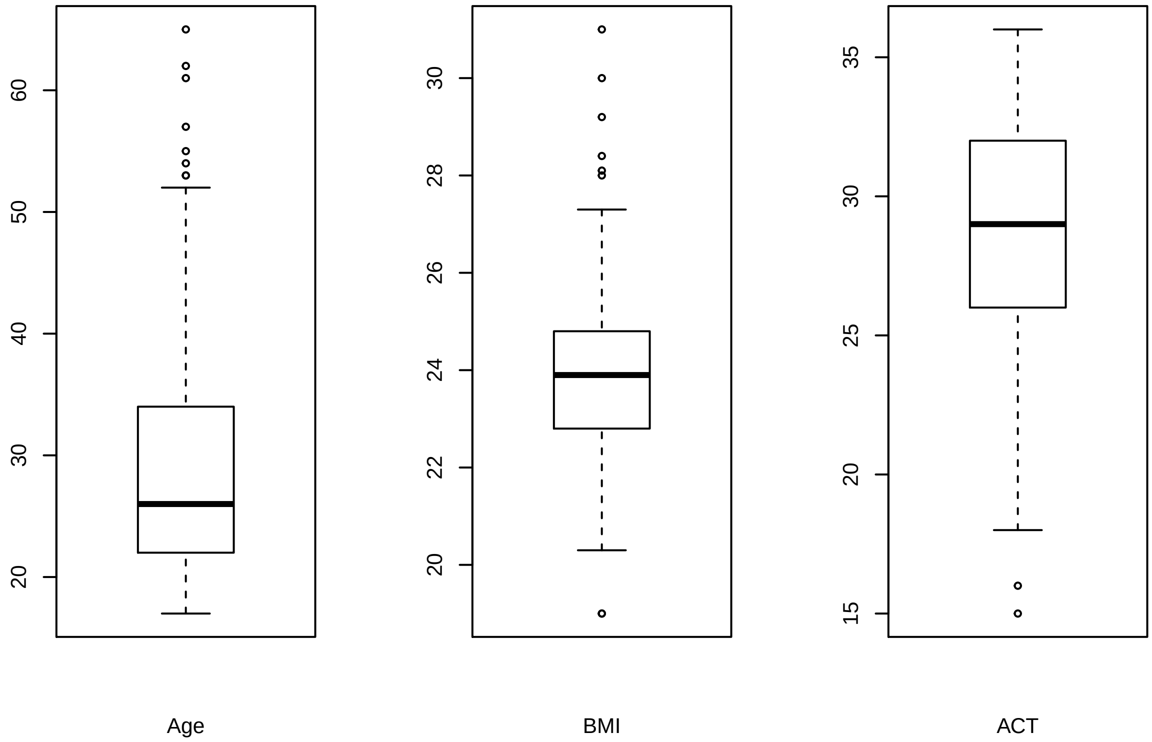
```
qqnorm(test$BMI)
qqline(test$BMI,col=2,lty=2)
```



The QQplot confirms the presence of anomalous values of BMI.

A unique plot with many boxplots.

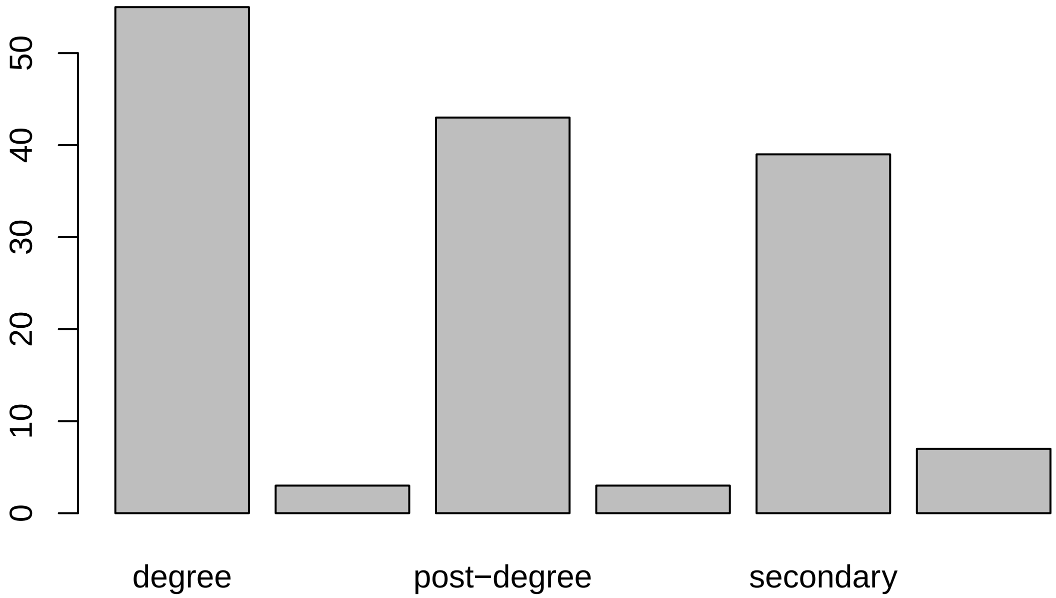
```
par(mfrow=c(1,3)) # 1  row 3 cols
boxplot(test$Age,xlab="Age")
boxplot(test$BMI,xlab="BMI")
boxplot(test$ACT,xlab="ACT")
```



```
par(mfrow=c(1,1))
```

Analysis of the Education variable

```
# A barplot with the frequency
barplot(table(test$Education))
```

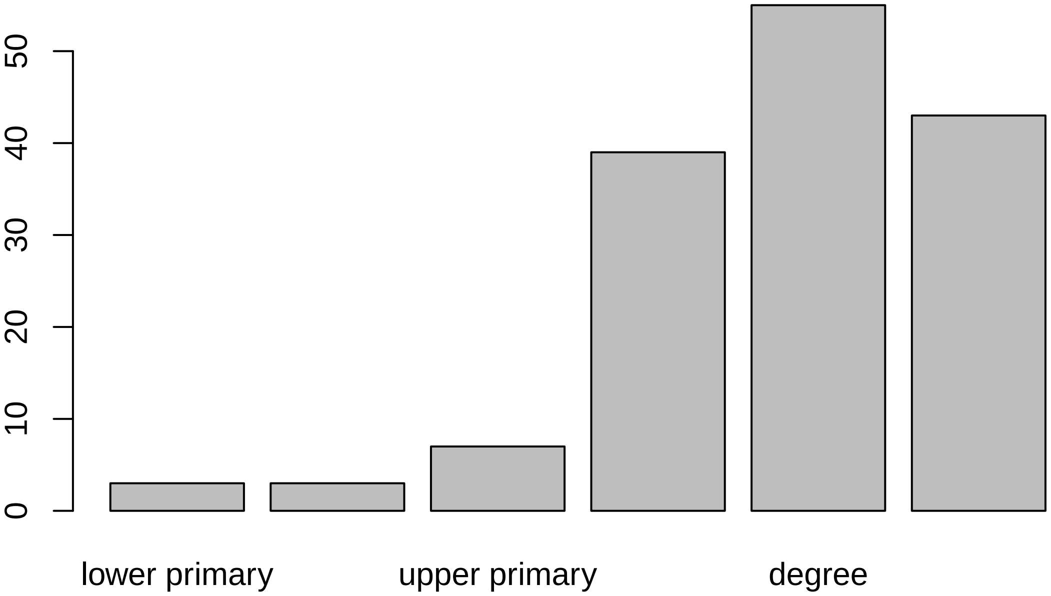


The barplot reports the frequency of each modality of the categorical variable. But this variable has an order. So we define the order as follows:

```
#here the levels
levels(test$Education)
```

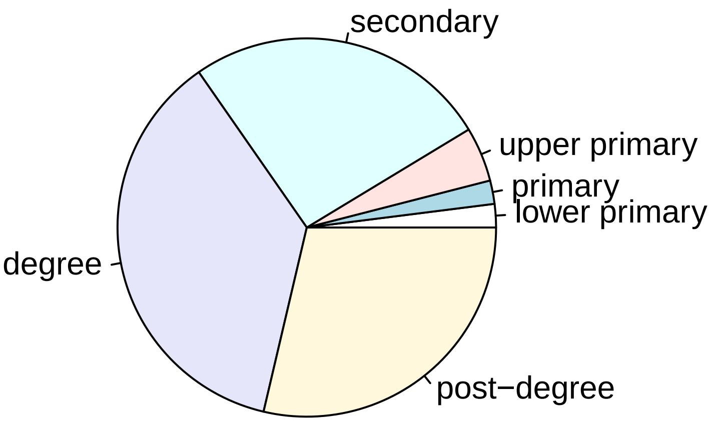
```
## [1] "degree"         "lower primary" "post-degree"   "primary"
## [5] "secondary"      "upper primary"
```

```
test$Education<-factor(test$Education,levels=
c("lower primary","primary","upper primary",
"secondary","degree","post-degree"),ordered =TRUE)
plot(test$Education) # here is ordered
```



Here a pie plot

```
pie(table(test$Education))
```



The function *table()* permits to obtain a frequency table

```
table(test$Education)
```

```
##
## lower primary       primary upper primary     secondary        degree
##             3             3             7            39            55
##   post-degree
##            43
```

```
#or a relative frequency table with the function prop.table()
prop.table(table(test$Education))
```

```
##
## lower primary       primary upper primary     secondary        degree
##    0.02000000    0.02000000    0.04666667    0.26000000    0.36666667
##   post-degree
##    0.28666667
```

Analysis of the Stress variable

The variable Stress is an integer values expressed on a likert scale (the common question:“How much are you stressed from 1 to 6?”). The likert scale is not numeric (variable on ratio scale), but it is an ordinal variable. With the command *factor()* R can set a factor, a categorical variable, even if it is formed by numbers.

```
is(test$Stress)
```

```
## [1] "integer"             "double"              "numeric"
## [4] "vector"              "data.frameRowLabels"
```

```
test$Stress<-factor(test$Stress)
table(test$Stress)
```

```
##
##  1  2  3  4  5  6
## 24 19 25 24 41 17
```

```
#or a relative frequency table with the function prop.table()
prop.table(table(test$Stress))
```

```
##
##         1         2         3         4         5         6
## 0.1600000 0.1266667 0.1666667 0.1600000 0.2733333 0.1133333
```

```
# the same for the variable social
test$Social<-factor(test$Social)
```

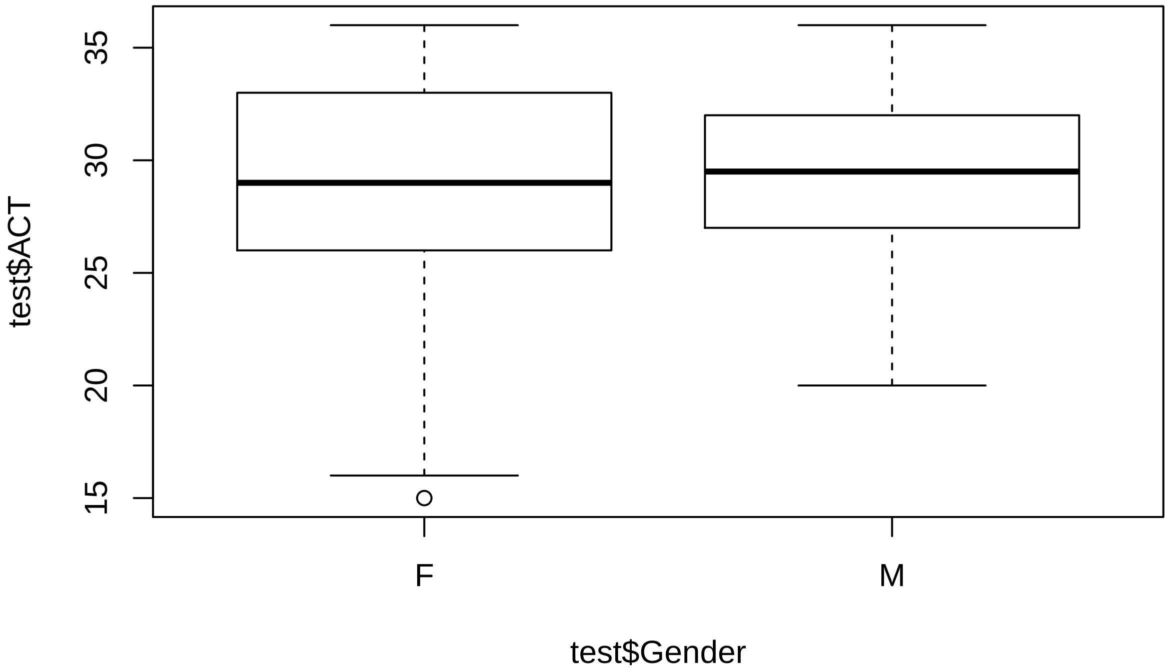
Bivariate Statistical Analysis with R

The dataset reported the results of 150 subjects on ACT e SAT tests.
Some variables influences the performances.

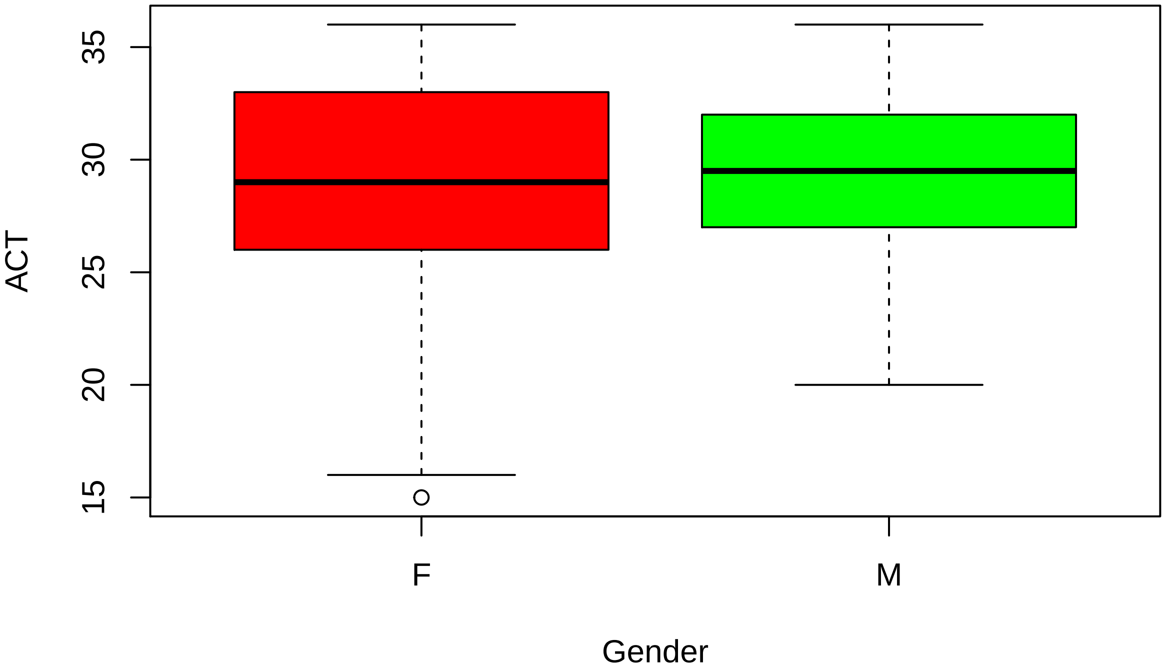
We try to reply to the question: “What are the factors that influenced the ACT, SATV and SATQ test?”

Quantitative vs qualitative variables

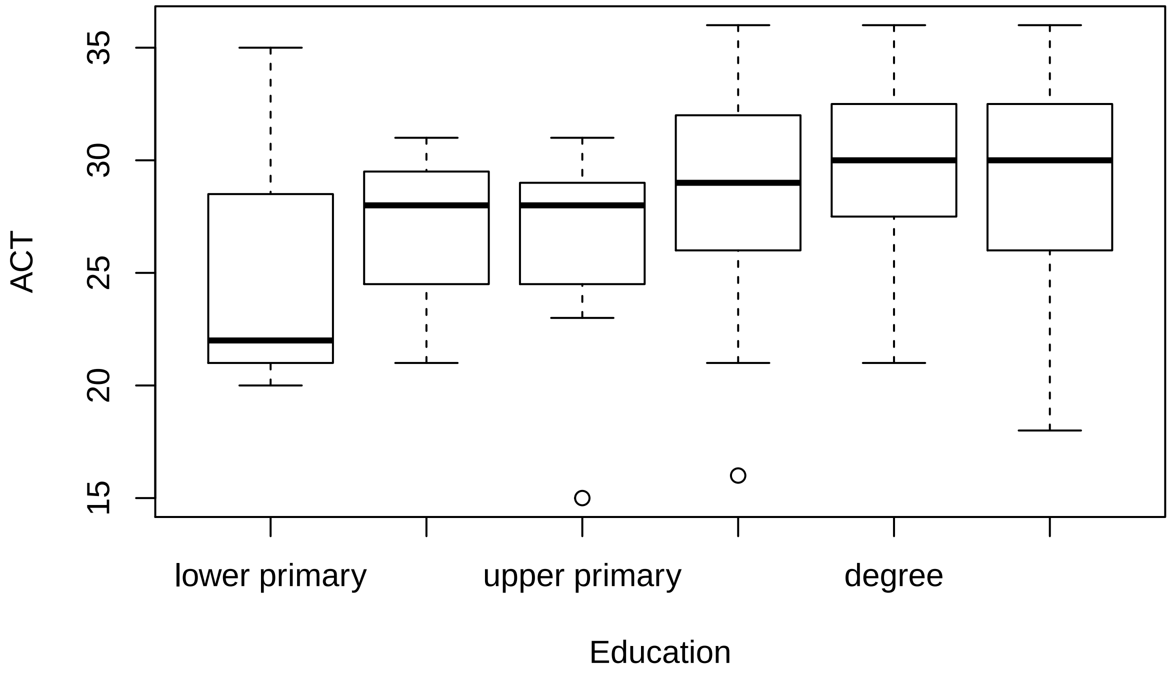
```
#ACT vs Gender and Education
boxplot(test$ACT~test$Gender)
```



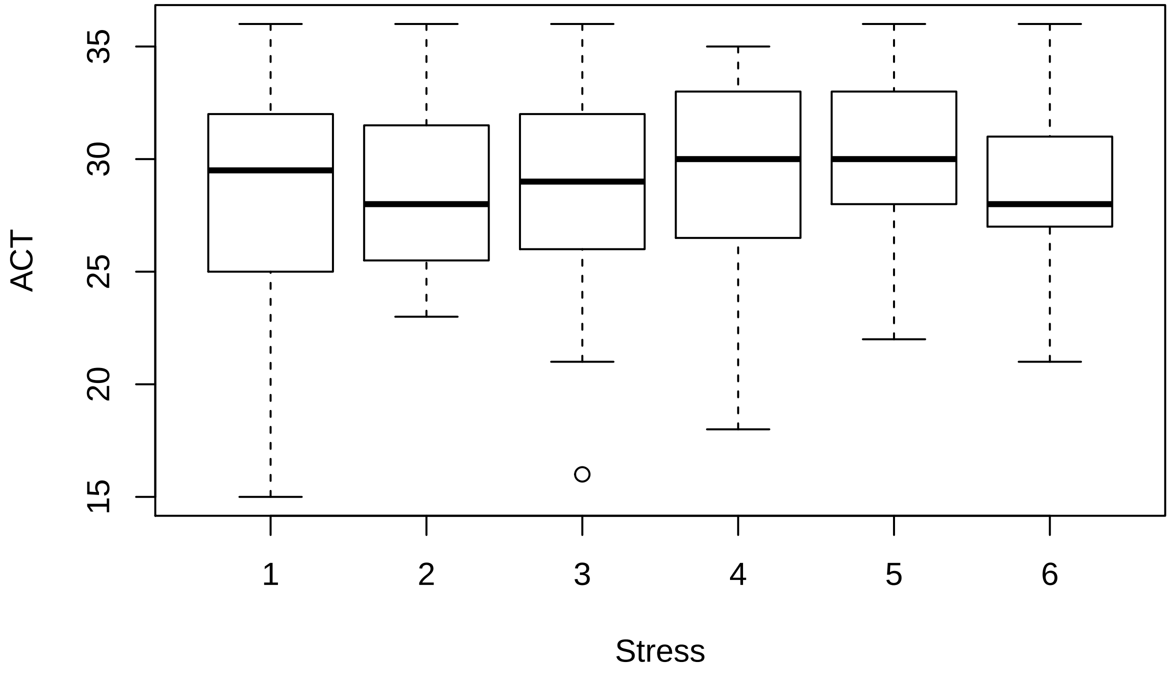
```
#change colour with col argument and labels
boxplot(test$ACT~test$Gender,col=c("red","green"),ylab="ACT",xlab="Gender")
```



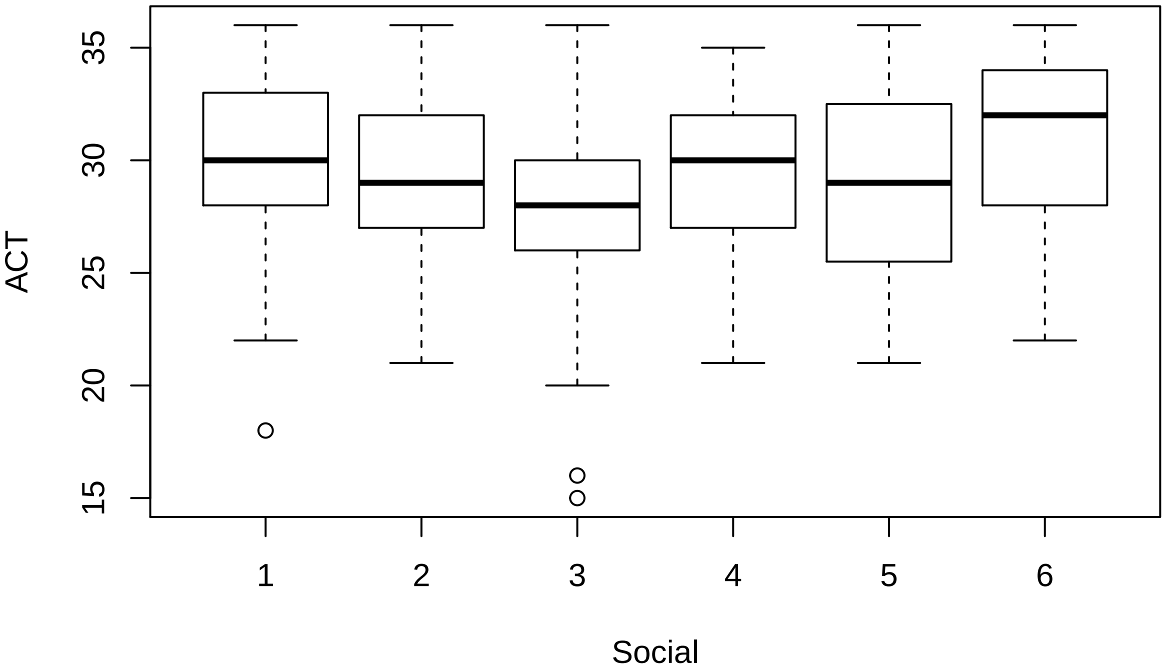
```
boxplot(test$ACT~test$Education,ylab="ACT",xlab="Education")
```



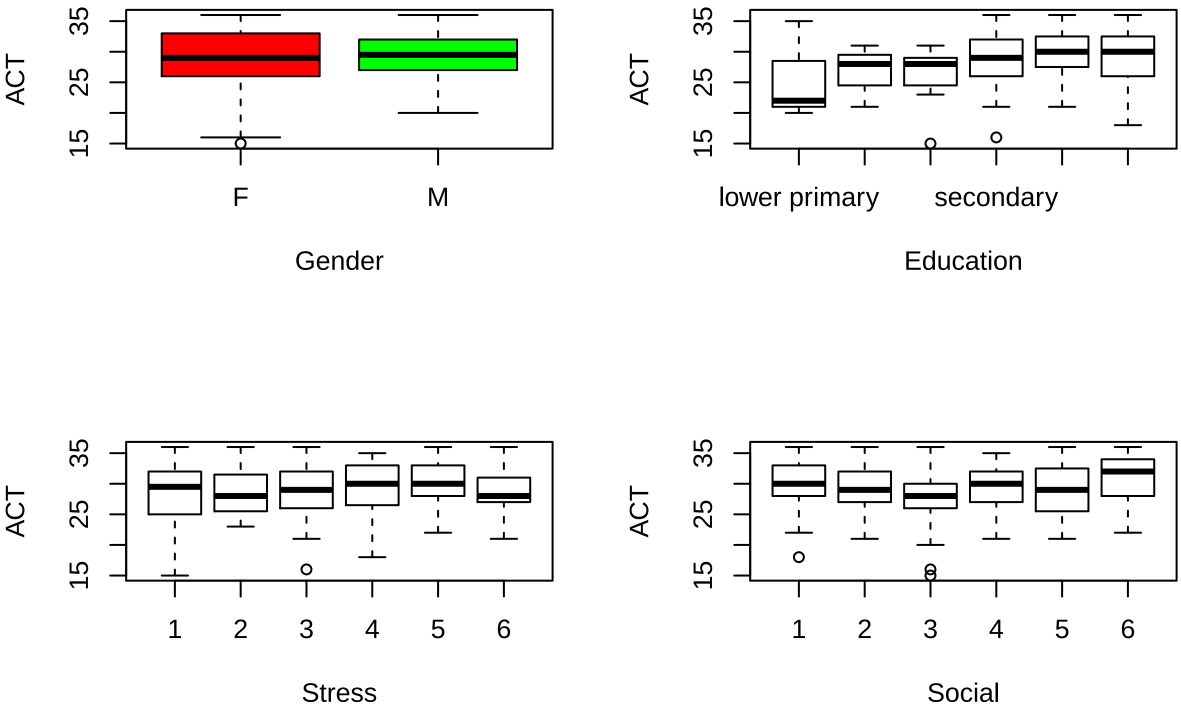
```
boxplot(test$ACT~test$Stress,ylab="ACT",xlab="Stress")
```



```
boxplot(test$ACT~test$Social,ylab="ACT",xlab="Social")
```



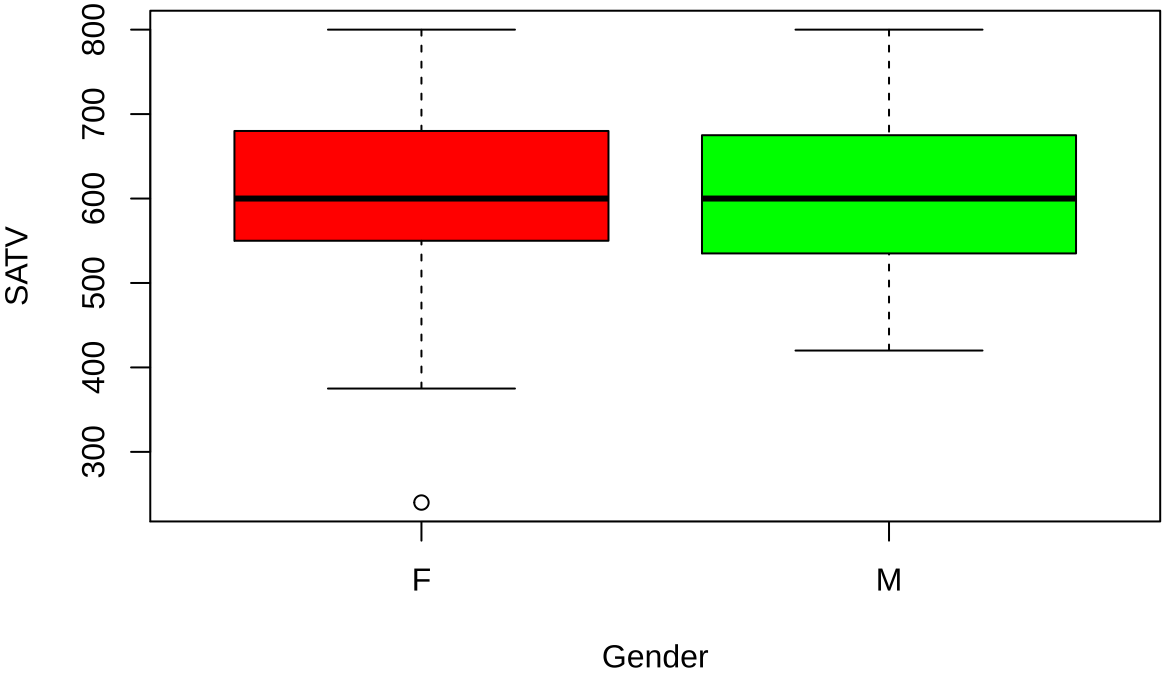
```
#all the plot in a unique figure
par(mfrow=c(2,2))
boxplot(test$ACT~test$Gender,col=c("red","green"),ylab="ACT",xlab="Gender")
boxplot(test$ACT~test$Education,ylab="ACT",xlab="Education")
boxplot(test$ACT~test$Stress,ylab="ACT",xlab="Stress")
boxplot(test$ACT~test$Social,ylab="ACT",xlab="Social")
```



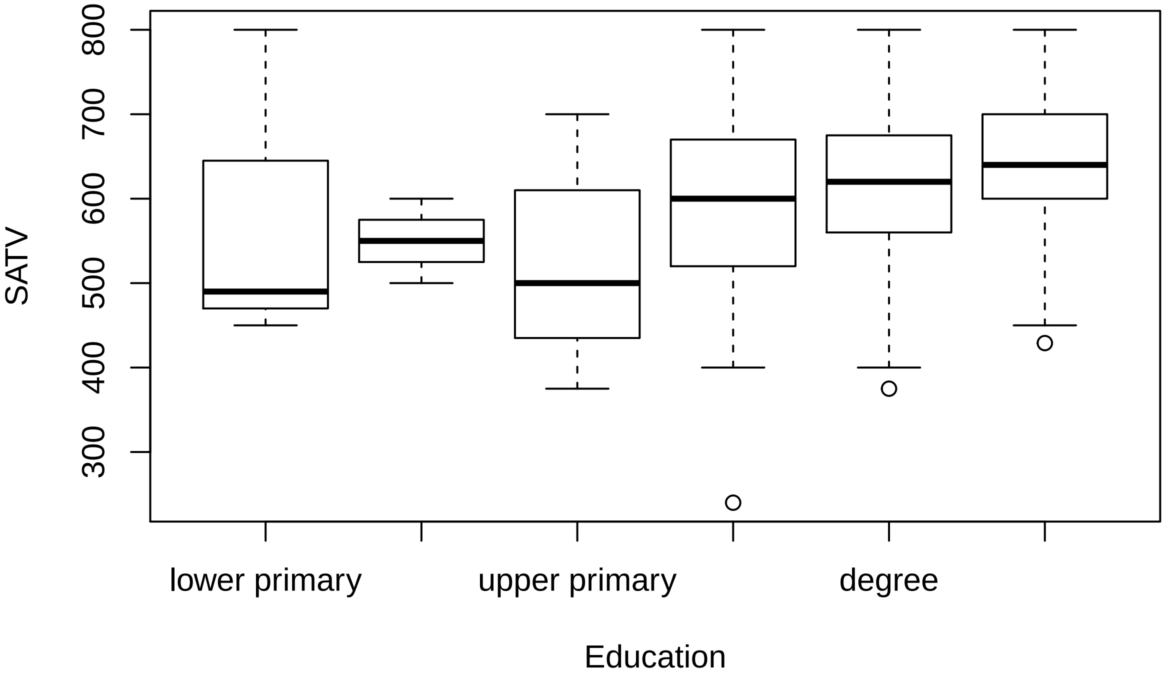
```
par(mfrow=c(1,1))
```

Here the statistical analysis for SATV and SATQ.
What are the comments on these charts?

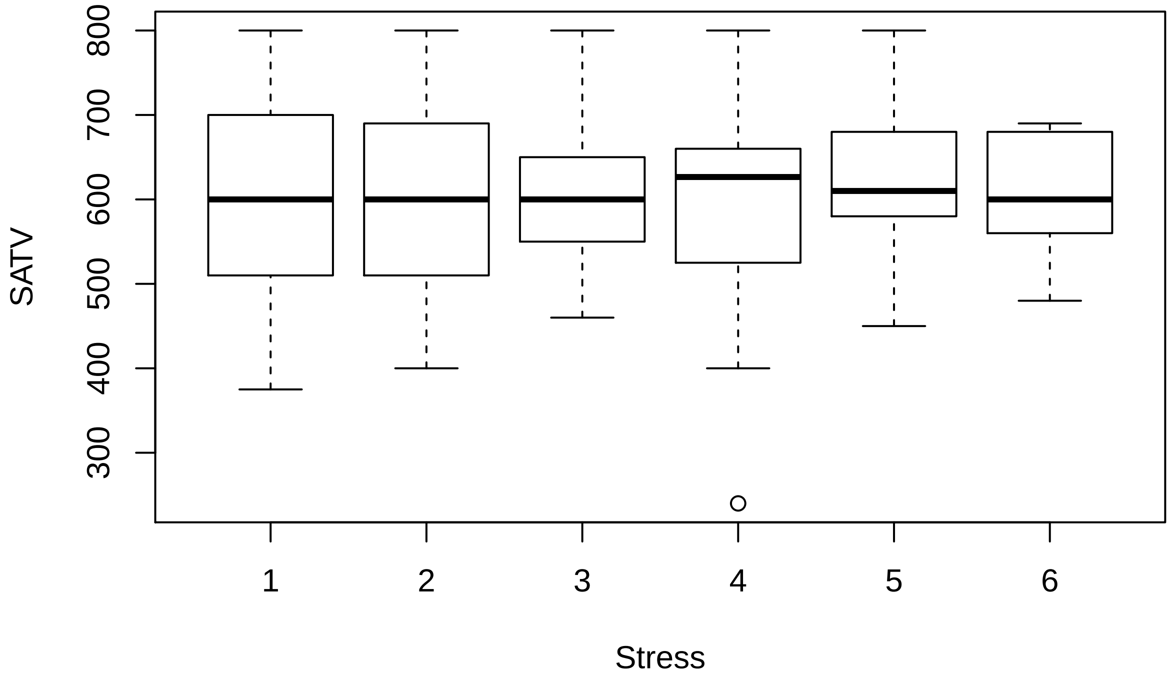
```
#SATV vs Gender and Education, Stress and Social
boxplot(test$SATV~test$Gender,col=c("red","green"),ylab="SATV",xlab="Gender")
```



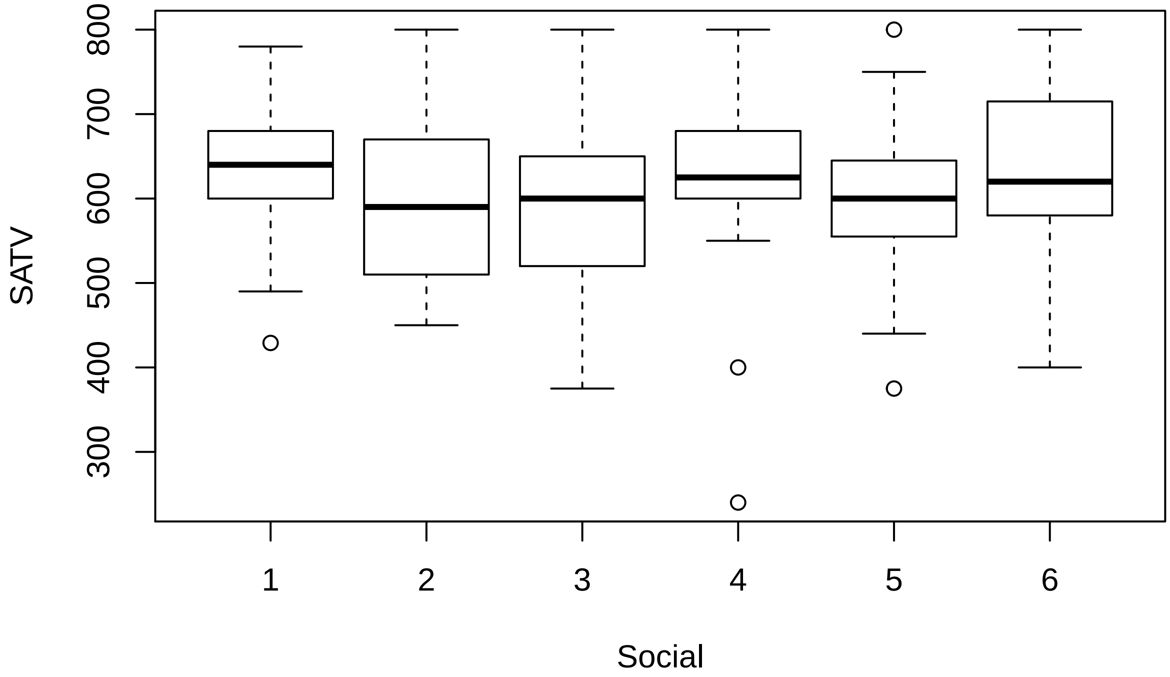
```
boxplot(test$SATV~test$Education,ylab="SATV",xlab="Education")
```



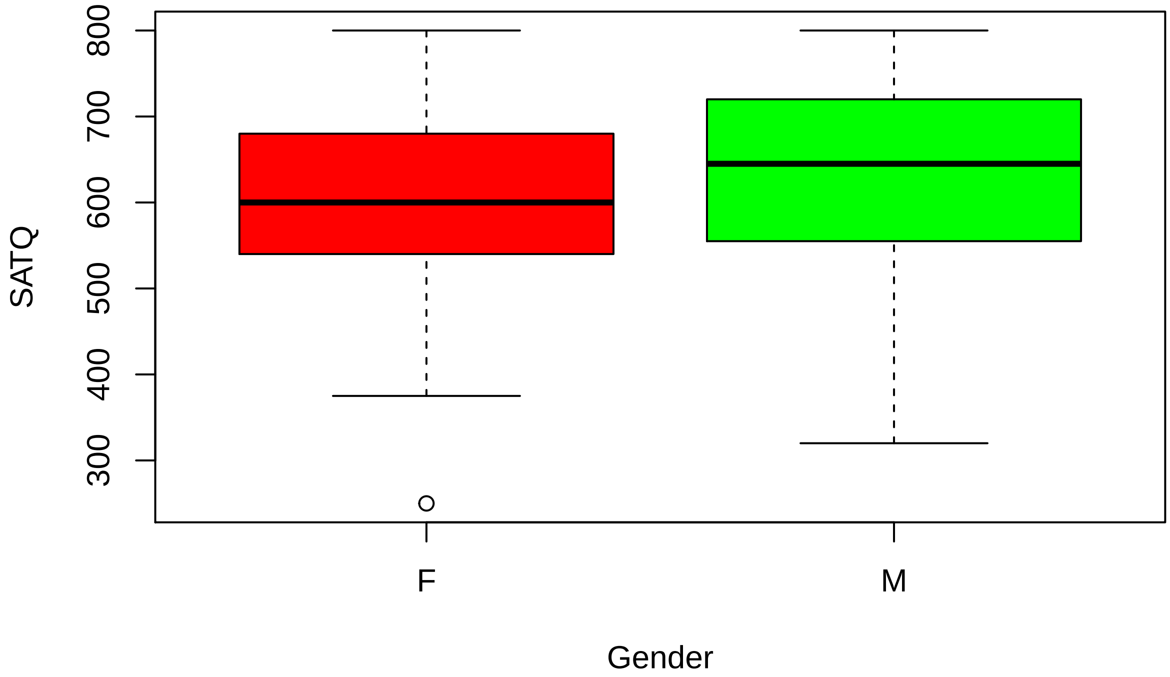
```
boxplot(test$SATV~test$Stress,ylab="SATV",xlab="Stress")
```



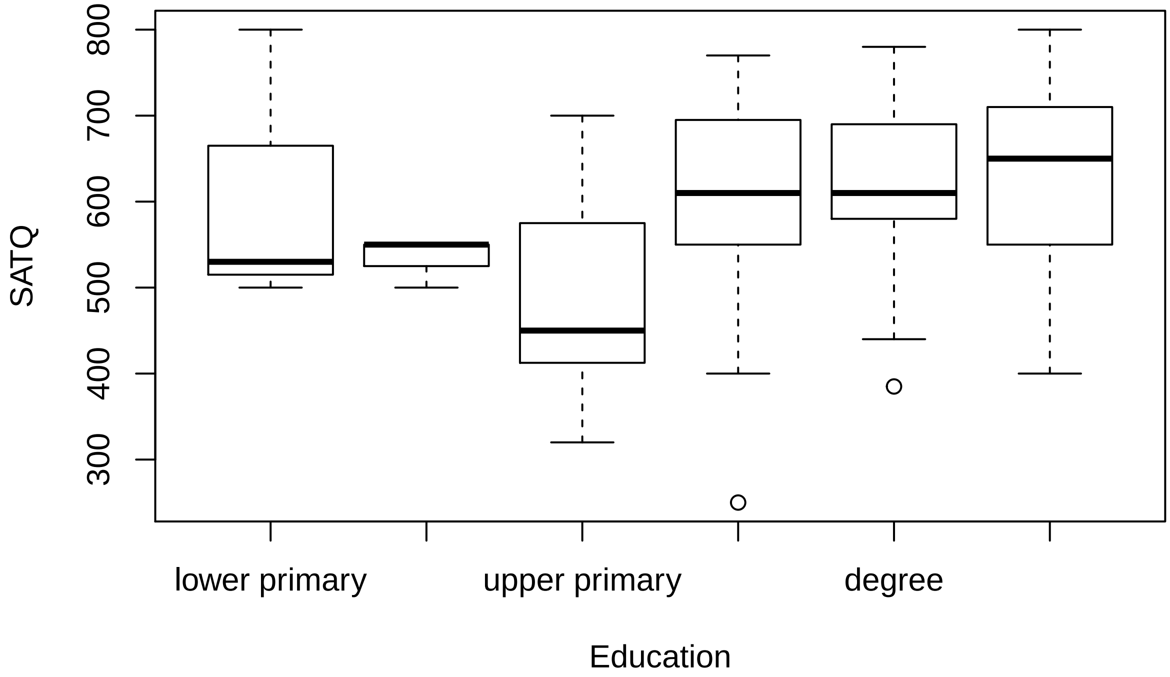
```
boxplot(test$SATV~test$Social,ylab="SATV",xlab="Social")
```



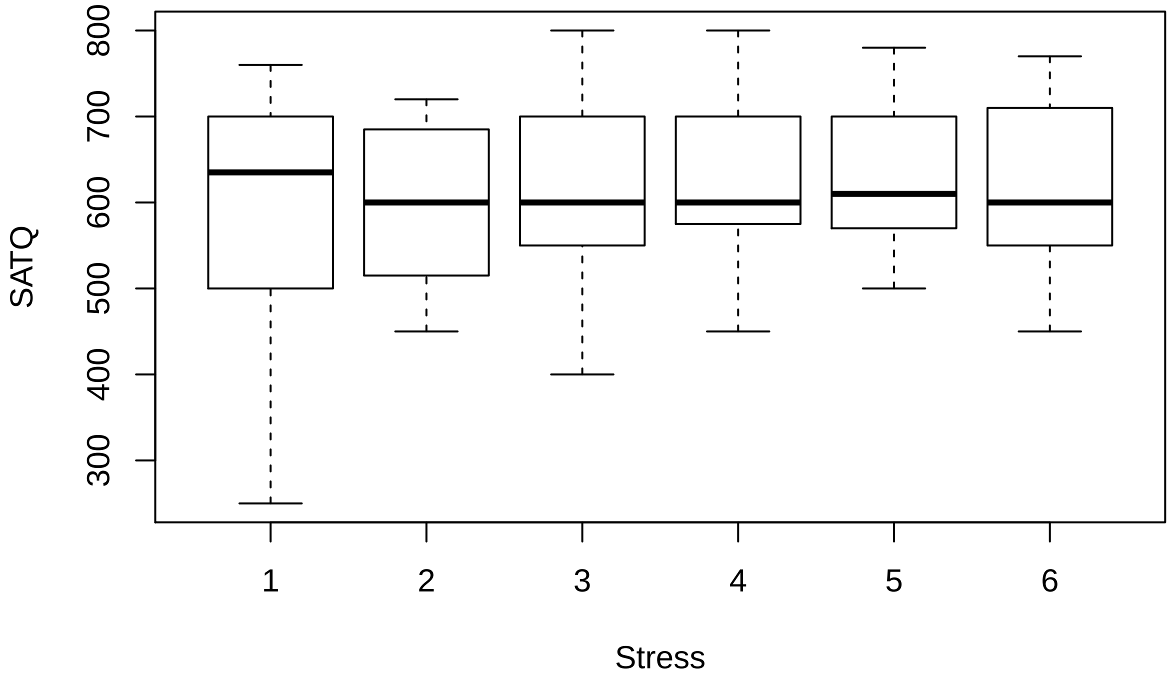
```
#SATQ vs Gender and Education, Stress and Social
boxplot(test$SATQ~test$Gender,col=c("red","green"),ylab="SATQ",xlab="Gender")
```



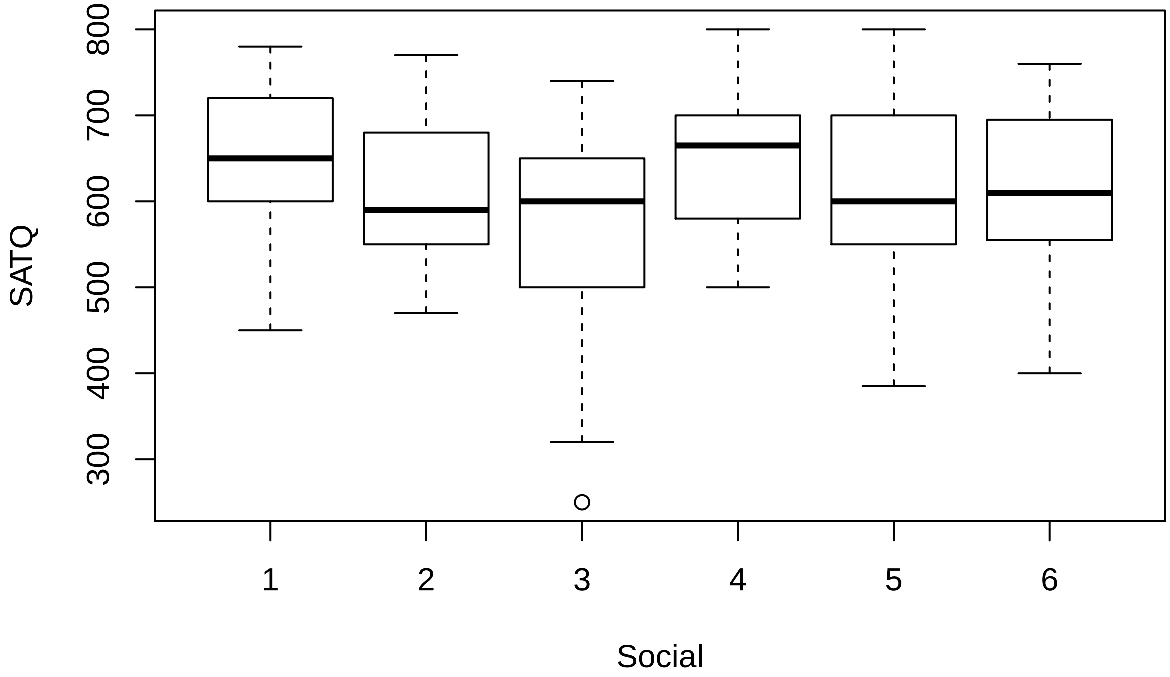
```
boxplot(test$SATQ~test$Education,ylab="SATQ",xlab="Education")
```



```
boxplot(test$SATQ~test$Stress,ylab="SATQ",xlab="Stress")
```



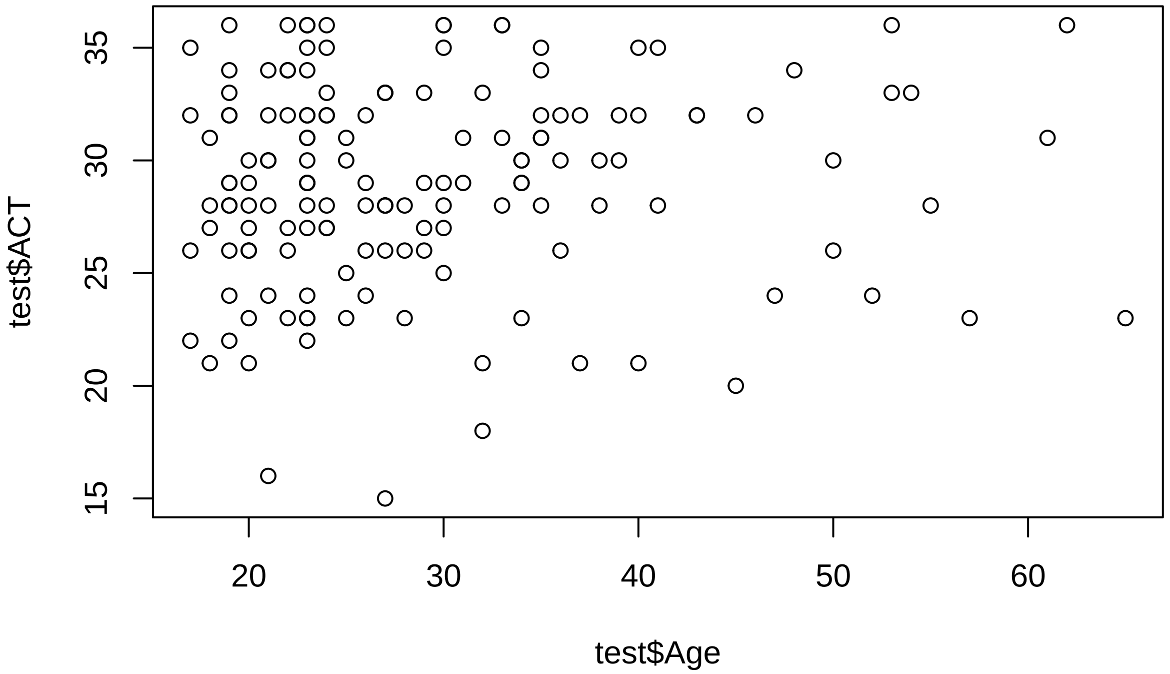
```
boxplot(test$SATQ~test$Social,ylab="SATQ",xlab="Social")
```



Quantitative vs Quantitative variables

Analysis of ACT vs. Age and BMI

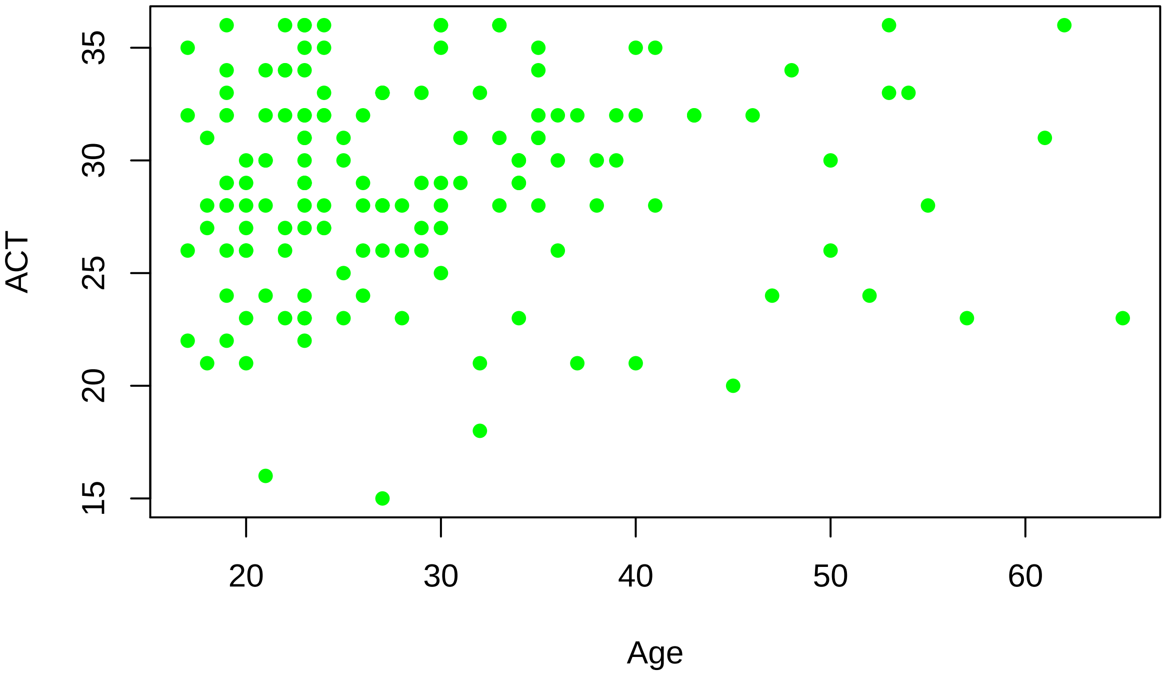
```
#ACT vs Age
plot(test$Age,test$ACT)
```



I can use some graphical parameter to have a better graph (please see ?plot or ?par for more options) :

- main, sub: title and subtitle
- xlab, ylab: label of the x and y axis
- xlim, ylim: limits of the x and y axis
- type: type of plot
- lty: type of lines
- pch:plot symbol
- cex: scale factor
- col: color of points etc.

```
#ACT vs Age
plot(test$Age,test$ACT,pch=16,col="green",xlab="Age",ylab="ACT")
```



Pearson correlation

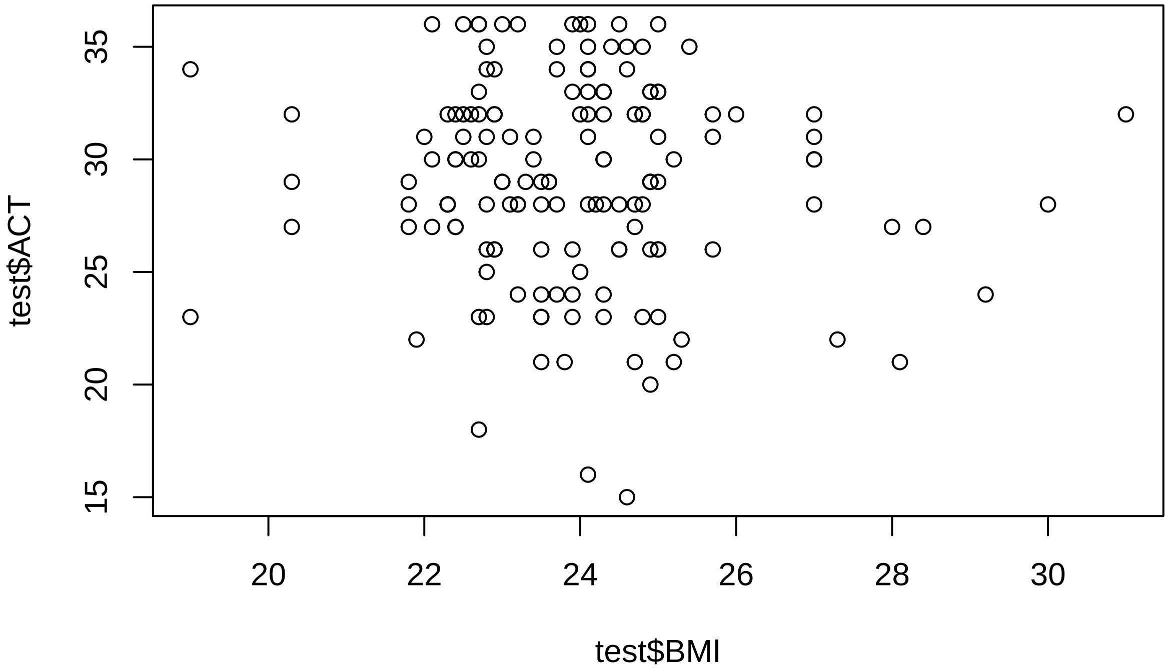
```
cor(test$Age,test$ACT) # pearson
```

```
## [1] 0.06821767
```

```
cor(test$Age,test$ACT,method="spearman") # spearman
```

```
## [1] 0.1033471
```

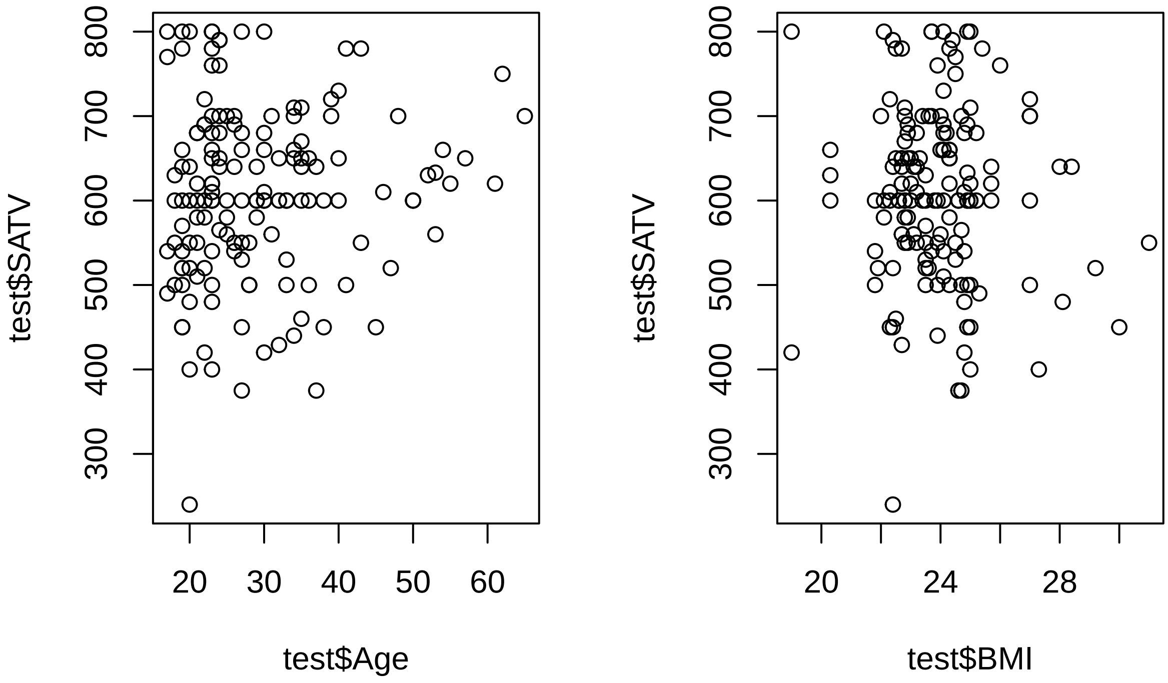
```
plot(test$BMI,test$ACT)
```



```
cor(test$BMI,test$ACT,method="spearman") # spearman
```

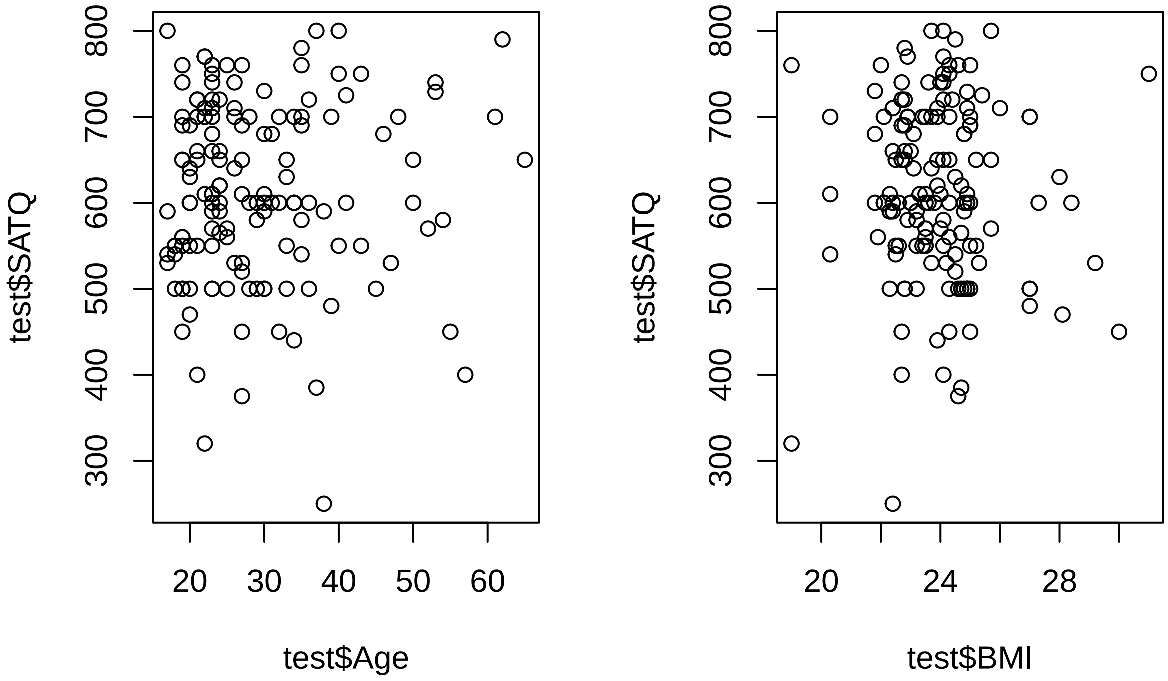
```
## [1] -0.0498391
```

```
par(mfrow=c(1,2))
plot(test$Age,test$SATV)
plot(test$BMI,test$SATV)
```



```
par(mfrow=c(1,1))
```

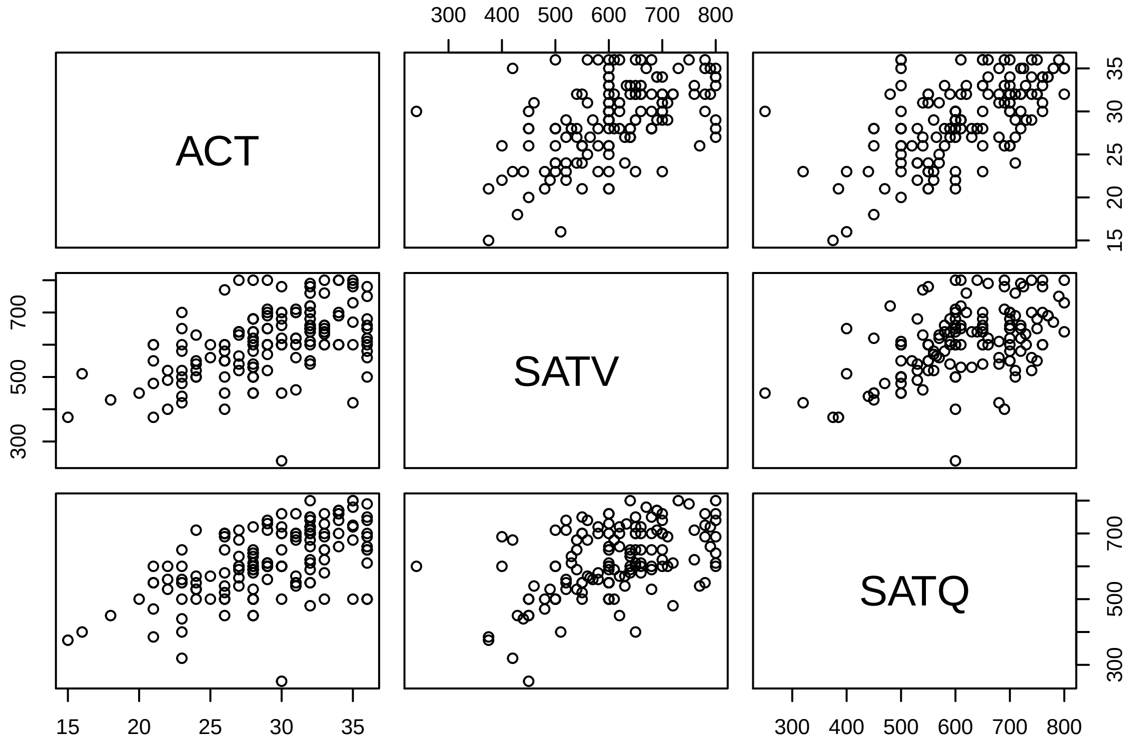
```
par(mfrow=c(1,2))
plot(test$Age,test$SATQ)
plot(test$BMI,test$SATQ)
```



```
par(mfrow=c(1,1))
```

Is there a correlation between test scores (ACT, SATQ, SATV)?

```
plot(test[,c("ACT","SATV","SATQ")])
```



```
cor(test[,c("ACT","SATV","SATQ")])
```

```
##            ACT      SATV      SATQ
## ACT  1.0000000 0.5146053 0.5728708
## SATV 0.5146053 1.0000000 0.5107873
## SATQ 0.5728708 0.5107873 1.0000000
```

```
cor(test[,c("ACT","SATV","SATQ")],method="spearman")
```

```
##            ACT      SATV      SATQ
## ACT  1.0000000 0.5280296 0.5644246
## SATV 0.5280296 1.0000000 0.5027085
## SATQ 0.5644246 0.5027085 1.0000000
```