Sessió 0

Principal characteristics of R and its programming:

- Open-source.
- Highly Active Community. Functions and packages are personal creations from users.
- Oriented to objects.
- Extremely comprehensive.
- It does not need a compiler to run code.
- Direction to machine learning.
- Compatibility with other Data Processing Technologies (Example: Use a spark cluster to process large datasets using R)
- R markdown to generate reports in any desired format.
- Operations directly on vectors, not too much looping.
- Data from APIs (and many other formats) can be easily pulled down.

Initial commands and basic descriptive statistics:

How to citate R?

```r
citation()
```

```
##
## To cite R in publications use:
##
##   R Core Team (2020). R: A language and environment for statistical
##   computing. R Foundation for Statistical Computing, Vienna, Austria.
##   URL https://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2020},
##     url = {https://www.R-project.org/},
##   }
```

```
##
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

Load package and data:

```r
# install.packages("car")
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```r
data(Davis)
attributes(Davis)
```

```
## $names
## [1] "sex"    "weight" "height" "repwt"  "repht" 
## 
## $class
## [1] "data.frame"
## 
## $row.names
##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12" 
##  [13] "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24" 
##  [25] "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36" 
##  [37] "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48" 
##  [49] "49"  "50"  "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60" 
##  [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"  "71"  "72" 
##  [73] "73"  "74"  "75"  "76"  "77"  "78"  "79"  "80"  "81"  "82"  "83"  "84" 
##  [85] "85"  "86"  "87"  "88"  "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96" 
##  [97] "97"  "98"  "99"  "100" "101" "102" "103" "104" "105" "106" "107" "108"
## [109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
## [121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
## [133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143" "144"
## [145] "145" "146" "147" "148" "149" "150" "151" "152" "153" "154" "155" "156"
## [157] "157" "158" "159" "160" "161" "162" "163" "164" "165" "166" "167" "168"
## [169] "169" "170" "171" "172" "173" "174" "175" "176" "177" "178" "179" "180"
## [181] "181" "182" "183" "184" "185" "186" "187" "188" "189" "190" "191" "192"
## [193] "193" "194" "195" "196" "197" "198" "199" "200"
```

```r
# Numeric Univariant Description
summary(Davis)
```

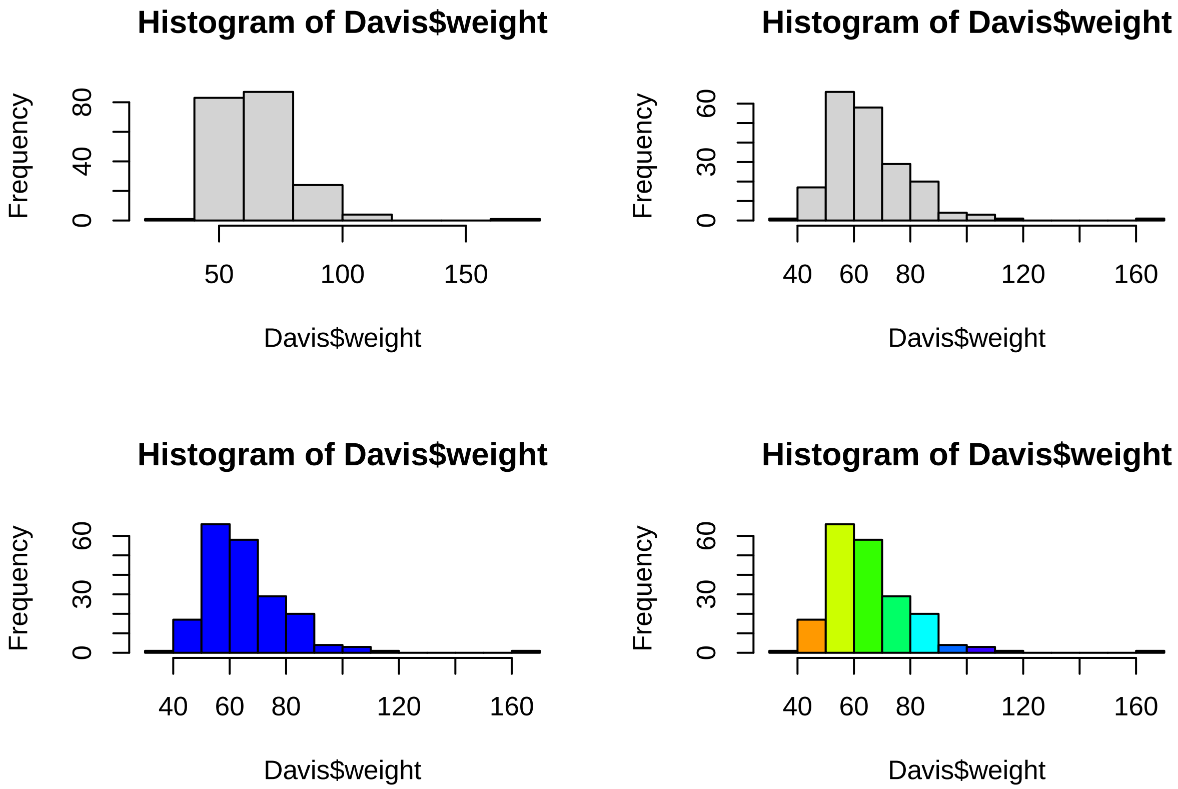
```
##  sex         weight          height          repwt            repht      
##  F:112   Min.   : 39.0   Min.   : 57.0   Min.   : 41.00   Min.   :148.0  
##  M: 88   1st Qu.: 55.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:160.5  
##          Median : 63.0   Median :169.5   Median : 63.00   Median :168.0  
##          Mean   : 65.8   Mean   :170.0   Mean   : 65.62   Mean   :168.5  
##          3rd Qu.: 74.0   3rd Qu.:177.2   3rd Qu.: 73.50   3rd Qu.:175.0  
##          Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0  
##                                          NA's   :17       NA's   :17     
```

```
# Rows and columns of data.frame Davis
dim(Davis)
```

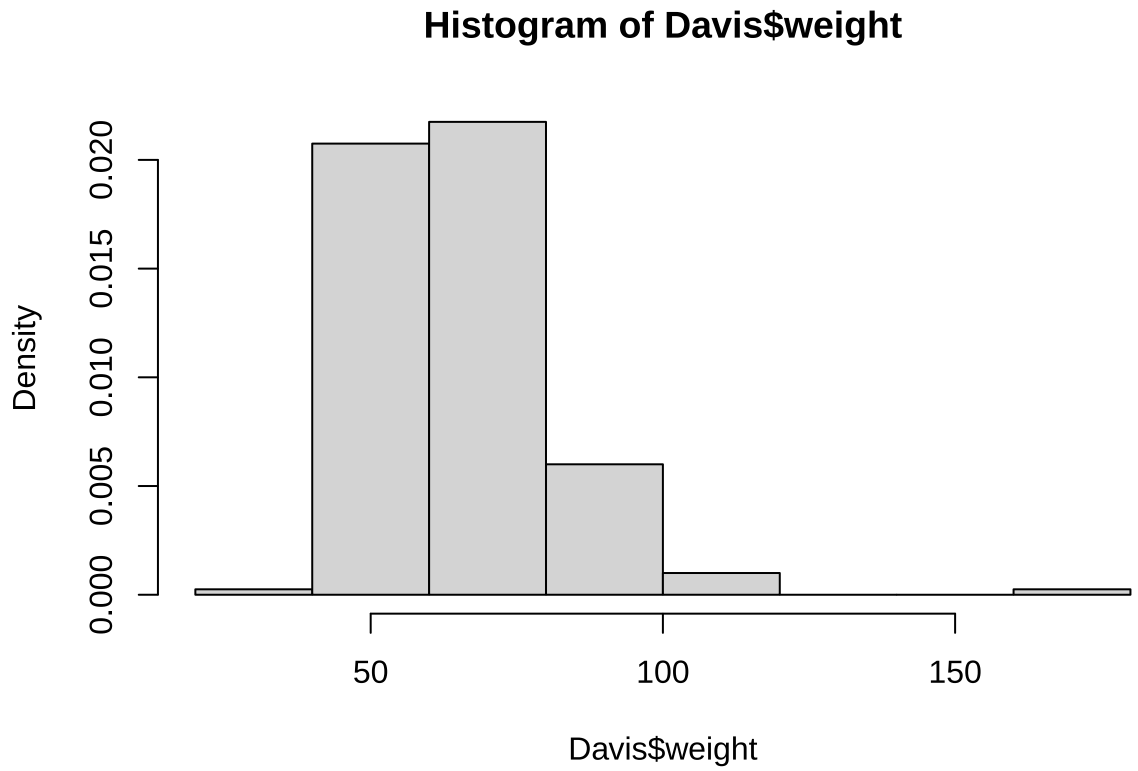
```
## [1] 200   5
```

Graphical description:

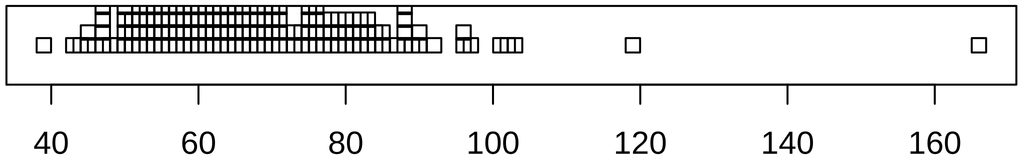
```
# Graphical Description of numeric data

# Histograms:
par(mfrow=c(2,2))
hist(Davis$weight)
hist(Davis$weight,10)
hist(Davis$weight,10,col="blue")
hist(Davis$weight,10,col=rainbow(10))
```

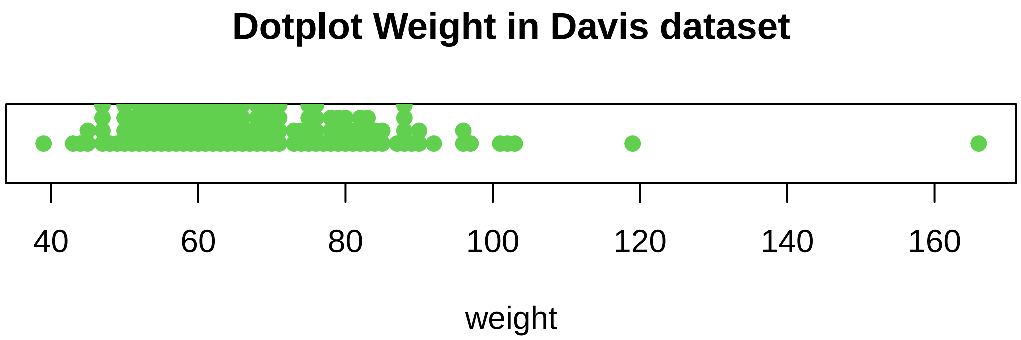


```
# Histogram with proportions:
par(mfrow=c(1,1))
hist(Davis$weight,freq=F)   # Proportions
```



```
# Dotplot:
par(mfrow=c(2,1))
stripchart(Davis$weight,method="stack")
stripchart(Davis$weight,method="stack",xlab="weight",pch=19,col=3,main="Dotplot Weight in Davis dataset
```





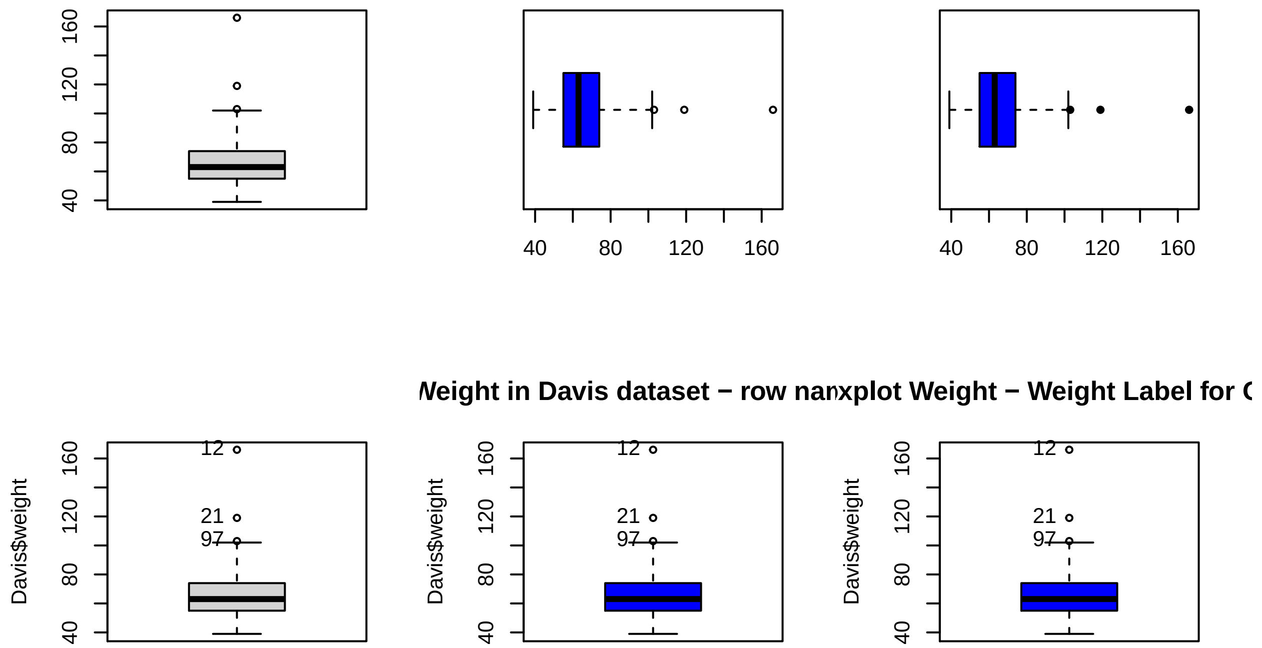
```
# Boxplots (two ways):
par(mfrow=c(2,3))
boxplot(Davis$weight)
boxplot(Davis$weight,col="blue",horizontal = TRUE)
boxplot(Davis$weight,col="blue",horizontal = TRUE, pch=19,labels=Davis$weight)

Boxplot(Davis$weight)
```

```
## [1] 12 21 97
```

```
Boxplot(Davis$weight,col="blue",main= "Weight in Davis dataset - row name Id")
```

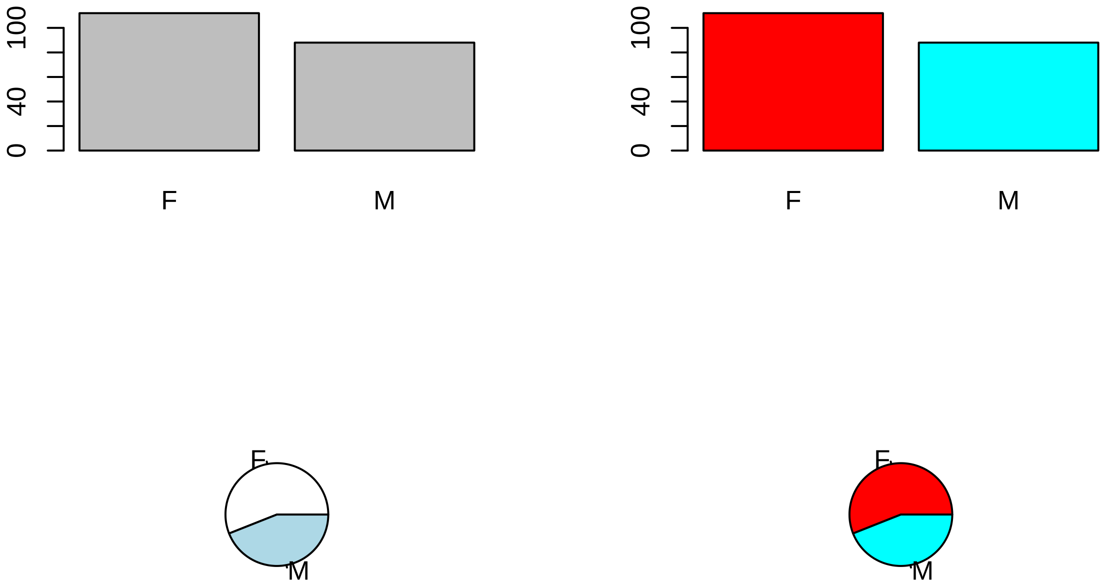
```
## [1] 12 21 97
```

```
Boxplot(Davis$weight,col="blue",main=" Boxplot Weight - Weight Label for Outliers",labels=Davis$weight)
```



```
## [1] 12 21 97
```

```
#Barplots and pie charts:
par(mfrow=c(2,2))
barplot(table(Davis$sex))
barplot(table(Davis$sex),col=rainbow(2))
pie(table(Davis$sex))
pie(table(Davis$sex),col=rainbow(2))
```



Description of variable factors:

```
table(Davis$sex)
```

```
##
##   F   M
## 112  88
```

```
margin.table(table(Davis$sex))
```

```
## [1] 200
```

```
prop.table(table(Davis$sex))
```

```
##
##    F    M
## 0.56 0.44
```

Ask for information, arguments and outputs of a function:

```
# ?boxplot
```

Other functions:

```r
# View(Davis)
head(Davis, n = 20) # n = 20 means  that the first 20 lines are printed in the R console
```

```
##    sex weight height repwt repht
## 1    M     77    182    77   180
## 2    F     58    161    51   159
## 3    F     53    161    54   158
## 4    M     68    177    70   175
## 5    F     59    157    59   155
## 6    M     76    170    76   165
## 7    M     76    167    77   165
## 8    M     69    186    73   180
## 9    M     71    178    71   175
## 10   M     65    171    64   170
## 11   M     70    175    75   174
## 12   F    166     57    56   163
## 13   F     51    161    52   158
## 14   F     64    168    64   165
## 15   F     52    163    57   160
## 16   F     65    166    66   165
## 17   M     92    187   101   185
## 18   F     62    168    62   165
## 19   M     76    197    75   200
## 20   F     61    175    61   171
```

```r
attach(Davis)
summary(weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    39.0    55.0    63.0    65.8    74.0   166.0
```

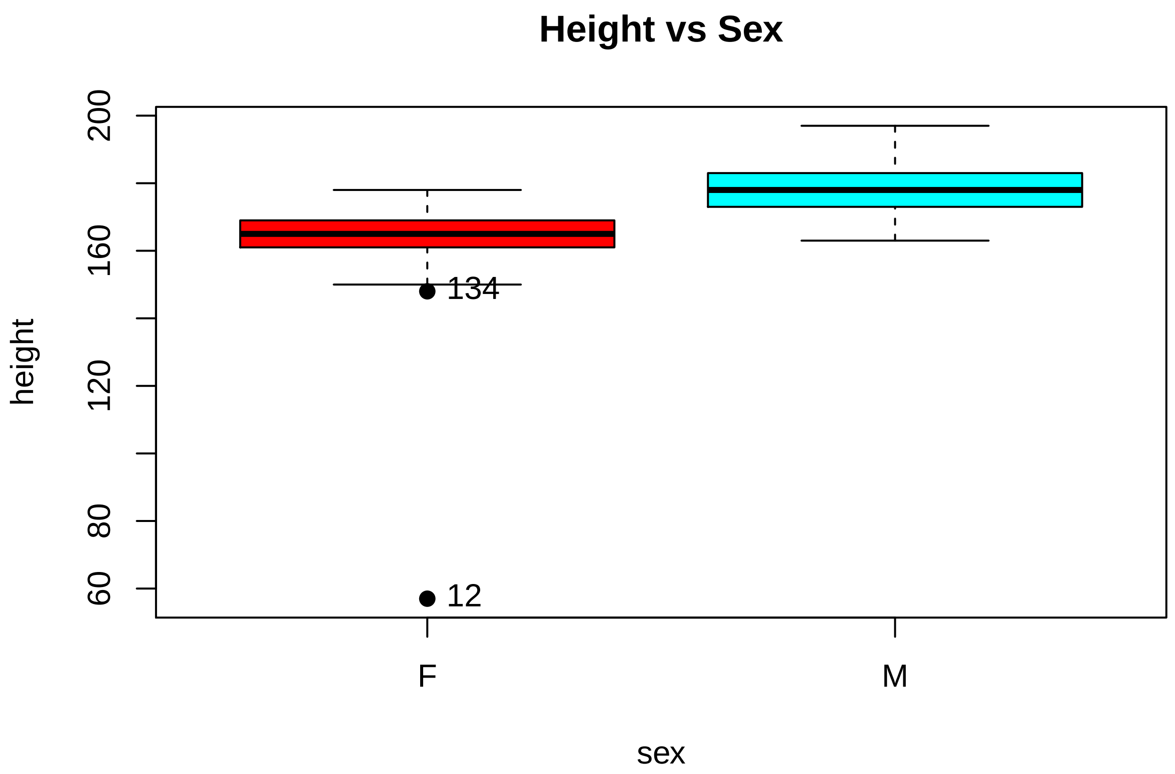
```r
detach(Davis)
# summary(weight) # Do not work

with(Davis,tapply(height,sex,summary))
```

```
## $F
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    57.0   161.0   165.0   163.7   169.0   178.0
##
## $M
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     163     173     178     178     183     197
```

```r
summary(Davis$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    57.0   164.0   169.5   170.0   177.2   197.0
```

```
with(Davis,Boxplot(height~sex,pch=19,col=rainbow(2),main="Height vs Sex"))
```



```
## [1] "12"  "134"
```

Set directory and load libraries to work with used cars data:

```
# setwd("C:/Users/lmontero/Dropbox/DOCENCIA/FIB-ADEI/PRACTICA/CarPrices/LABS")
setwd("E:/Docencia_UPC/GEI-ADEI/Lab 0") #Set working directory
# install.packages(c("car", "FactoMineR", "knitr"))
# Access to packages and its functions:
library(car)
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.0.5
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.0.5
```

Load Data and Select Sample:

```
# Lecture of DataFrames:
df1 <- read.table("audi.csv",header=T, sep=",")
df1$manufacturer <- "Audi"
df2 <- read.table("bmw.csv",header=T, sep=",")
df2$manufacturer <- "BMW"
df3 <- read.table("merc.csv",header=T, sep=",")
df3$manufacturer <- "Mercedes"
df4 <- read.table("vw.csv",header=T, sep=",")
df4$manufacturer <- "VW"

# Union by row:
df <- rbind(df1,df2,df3,df4)
dim(df)   # Size of data.frame
```

```
## [1] 49725    10
```

```
str(df) # Object class and description
```

```
## 'data.frame':    49725 obs. of  10 variables:
##  $ model       : chr  " A1" " A6" " A1" " A4" ...
##  $ year        : int  2017 2016 2016 2017 2019 2016 2016 2016 2015 2016 ...
##  $ price       : int  12500 16500 11000 16800 17300 13900 13250 11750 10200 12000 ...
##  $ transmission: chr  "Manual" "Automatic" "Manual" "Automatic" ...
##  $ mileage     : int  15735 36203 29946 25952 1998 32260 76788 75185 46112 22451 ...
##  $ fuelType    : chr  "Petrol" "Diesel" "Petrol" "Diesel" ...
##  $ tax         : int  150 20 30 145 145 30 30 20 20 30 ...
##  $ mpg         : num  55.4 64.2 55.4 67.3 49.6 58.9 61.4 70.6 60.1 55.4 ...
##  $ engineSize  : num  1.4 2 1.4 2 1 1.4 2 2 1.4 1.4 ...
##  $ manufacturer: chr  "Audi" "Audi" "Audi" "Audi" ...
```

```
names(df) # List of variable names
```

```
##  [1] "model"        "year"         "price"        "transmission" "mileage"
##  [6] "fuelType"     "tax"          "mpg"          "engineSize"   "manufacturer"
```

```
### Use birthday of 1 member of the group as random seed:
set.seed(12345)
# Random selection of x registers:
sam<-as.vector(sort(sample(1:nrow(df),1000)))
head(df)  # Take a look to the first rows/instances (6 rows)
```

```
##   model year price transmission mileage fuelType tax  mpg engineSize
## 1    A1 2017 12500       Manual   15735   Petrol 150 55.4        1.4
## 2    A6 2016 16500    Automatic   36203   Diesel  20 64.2        2.0
## 3    A1 2016 11000       Manual   29946   Petrol  30 55.4        1.4
## 4    A4 2017 16800    Automatic   25952   Diesel 145 67.3        2.0
## 5    A3 2019 17300       Manual    1998   Petrol 145 49.6        1.0
## 6    A1 2016 13900    Automatic   32260   Petrol  30 58.9        1.4
##   manufacturer
## 1         Audi
## 2         Audi
## 3         Audi
## 4         Audi
## 5         Audi
## 6         Audi
```

```
df<-df[sam,]  # Subset of rows _ It will be my sample
summary(df)
```

```
##     model                year          price        transmission
##  Length:1000        Min.   :2000   Min.   :  1495   Length:1000
##  Class :character   1st Qu.:2016   1st Qu.: 14277   Class :character
##  Mode  :character   Median :2017   Median : 19661   Mode  :character
##                     Mean   :2017   Mean   : 21562
##                     3rd Qu.:2019   3rd Qu.: 25996
##                     Max.   :2020   Max.   :139559
##     mileage         fuelType              tax             mpg
##  Min.   :     6   Length:1000        Min.   :  0.0   Min.   :  1.1
##  1st Qu.:  5711   Class :character   1st Qu.:125.0   1st Qu.: 44.8
##  Median : 17672   Mode  :character   Median :145.0   Median : 52.3
##  Mean   : 23971                      Mean   :127.6   Mean   : 53.5
##  3rd Qu.: 35902                      3rd Qu.:145.0   3rd Qu.: 61.4
##  Max.   :193000                      Max.   :570.0   Max.   :166.0
##    engineSize   manufacturer
##  Min.   :0.00   Length:1000
##  1st Qu.:1.50   Class :character
##  Median :2.00   Mode  :character
##  Mean   :1.95
##  3rd Qu.:2.00
##  Max.   :6.20
```

```
#Keep information in an .Rdata file:
save(list=c("df"),file="MyOldCars-Raw.RData")
```

R markdown offers the possibility to structure the document in different header levels:

Header 1

Header 2

Header 3

and simple text included into pargraphs.

To enumerate:

- Enumeration 1

- Enumeration 2

- ...

It also allows the user to write in **bold** and in *italica*.

In addition, equations in $LaTeX$ can be added when you are writing such as $2 \cdot x = 6$ and appart from text:

$$2 \cdot x = 6$$

More elements which can be included are:

- $LaTeX$ tables
- Images
- Links
- Bibliography
- $LaTeX$ matrices
- ...

More possibilities can be found visiting next reference:

10 R markdown possibilities