1) LESSON 7. STATISTICS IN R – PART 1

Class content:

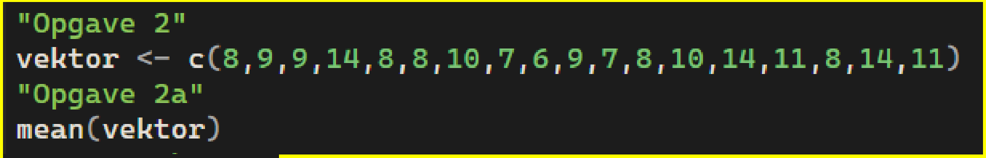
1) What is R and what is R Studio?
2) Installing R and R studio
3) Support materials
4) R components and layout
5) Opening the data in R
6) Descriptive statistics in R: summary functions and basic plots

1) Install R and check the version of the software you have installed. You can do that by typing R.Version() in the console.

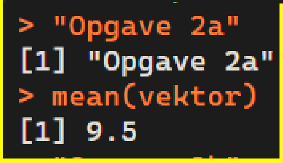
- Dette er gjort 😉

2) Create the following vector in R:
{8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11}
   a) For the data assigned to this vector, calculate the following:
      I. Mean
      II. Median
      III. Standard deviation
   b) Construct a histogram for the data
   c) Construct a boxplot for the data

i. Vi kan se at vi har følgende opskrivning og resultat fra kommandoen.

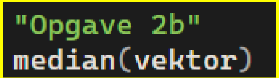


```
"Opgave 2"
vektor <- c(8,9,9,14,8,8,10,7,6,9,7,8,10,14,11,8,14,11)
"Opgave 2a"
mean(vektor)
```

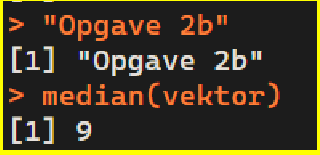


```
> "Opgave 2a"
[1] "Opgave 2a"
> mean(vektor)
[1] 9.5
```

ii. Vi har følgende opskrivning og resultat fra kommandoen.

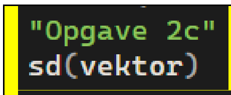


```
"Opgave 2b"
median(vektor)
```

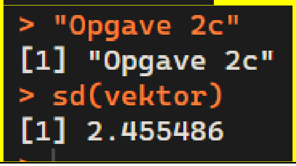


```
> "Opgave 2b"
[1] "Opgave 2b"
> median(vektor)
[1] 9
```

iii. Vi har følgende opskrivning og resultat fra kommandoen.



```
"Opgave 2c"
sd(vektor)
```



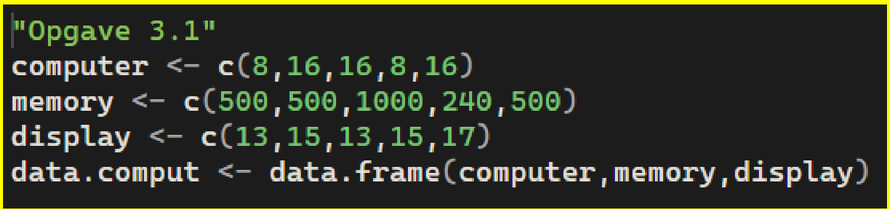
```
> "Opgave 2c"
[1] "Opgave 2c"
> sd(vektor)
[1] 2.455486
```

3) This exercise is divided in three steps. These steps are the following:

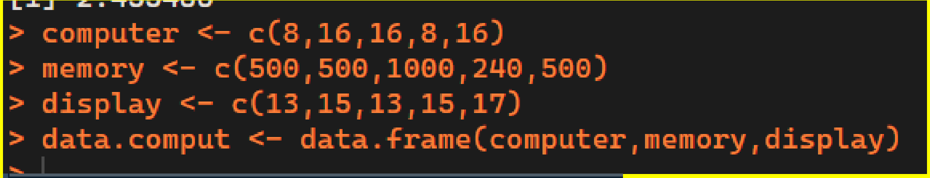
Step 1: Create a data frame in R called ***data.comput*** with data on 5 laptop computers regarding their memory, storage and display size. You know the following:

- Computer 1 has 8 GB RAM of memory, 500 GB storage drive and 13 inches display.
- Computer 2 has 16 GB RAM of memory, 500 GB storage drive and 15 inches display.
- Computer 3 has 16 GB RAM of memory, 1000 GB storage drive and 13 inches display.
- Computer 4 has 8 GB RAM of memory, 240 GB storage drive and 15 inches display.
- Computer 5 has 16 GB RAM of memory, 500 GB storage drive and 17 inches display.

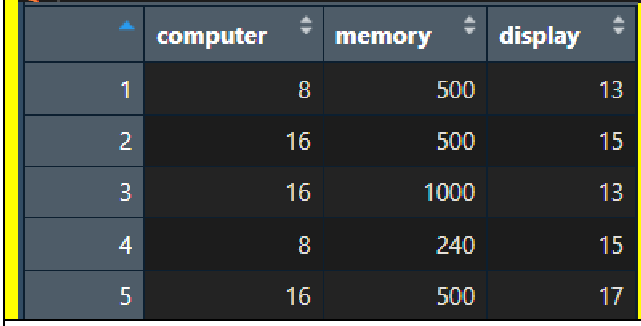
- Vi starter med at opskrive alle dataerne, hvor vi efterfølgende inddeler dem inde i selve rækker.



```
"Opgave 3.1"
computer <- c(8,16,16,8,16)
memory <- c(500,500,1000,240,500)
display <- c(13,15,13,15,17)
data.comput <- data.frame(computer,memory,display)
```



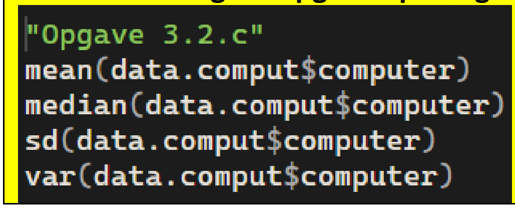
```
> computer <- c(8,16,16,8,16)
> memory <- c(500,500,1000,240,500)
> display <- c(13,15,13,15,17)
> data.comput <- data.frame(computer,memory,display)
```



	computer	memory	display
1	8	500	13
2	16	500	15
3	16	1000	13
4	8	240	15
5	16	500	17

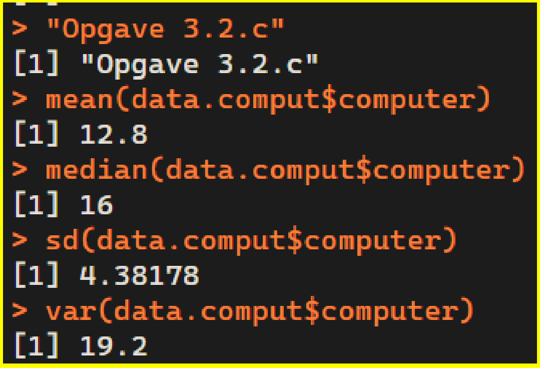
Step 2: Calculate the mean, median and standard deviation for the variables “memory”, “storage”, and “display”.

i. Vi udregner opgaven på følgende måde:



```
"Opgave 3.2.c"
mean(data.comput$computer)
median(data.comput$computer)
sd(data.comput$computer)
var(data.comput$computer)
```

- **Resultater fra kommandolinjen, bliver således for udregning af computersektionen!**



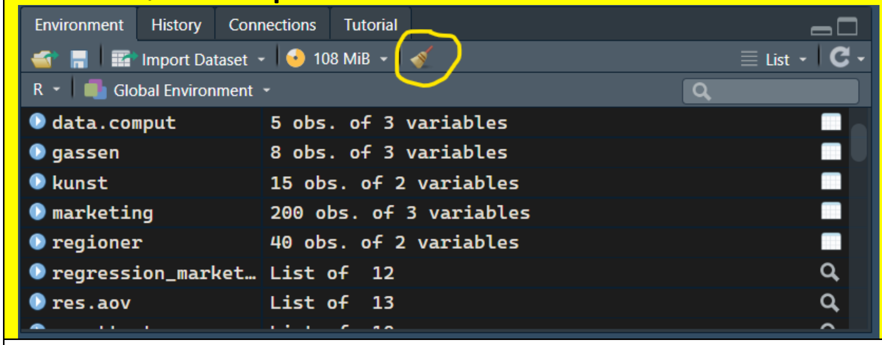
```
> "Opgave 3.2.c"
[1] "Opgave 3.2.c"
> mean(data.comput$computer)
[1] 12.8
> median(data.comput$computer)
[1] 16
> sd(data.comput$computer)
[1] 4.38178
> var(data.comput$computer)
[1] 19.2
```

- **Resultater fra kommandolinjen, bliver således for udregning af displaysektionen!**

- **Resultater fra kommandolinjen, bliver således for udregning af memory-sektionen!**
-

Step 3: Save the workspace (environment) containing the data frame ***data.comput*** in a work directory that is convenient to you. To practice how to open it again, close the R session and open the workspace again and see if you can easily recover the objects (i.e. data, values) of the previous session.

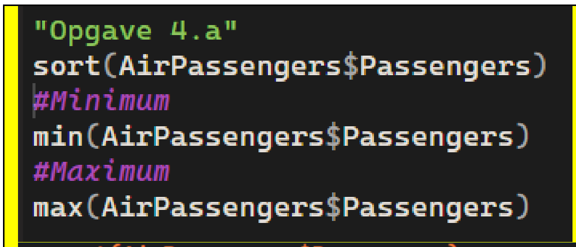
- **Det kan man teste selv, ellers kan man fjerne alle værdierne væk ved at bruge følgende ”børste-knap”.**



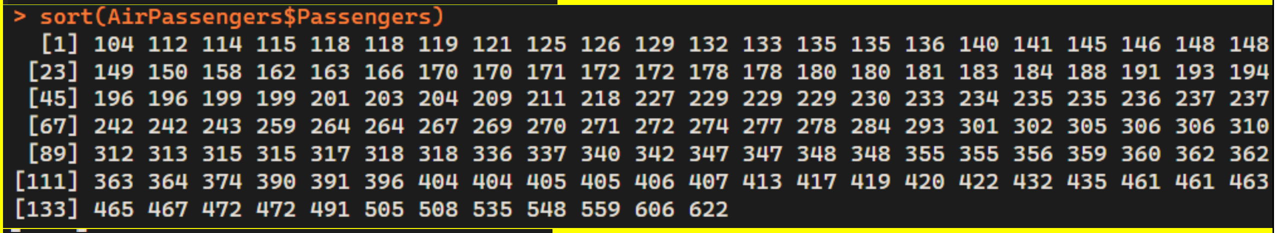
```
data.comput            5 obs. of 3 variables
gassen                 8 obs. of 3 variables
kunst                  15 obs. of 2 variables
marketing              200 obs. of 3 variables
regioner               40 obs. of 2 variables
regression_market…     List of  12
res.aov                List of  13
```

4) Import the dataset “Air_passengers.xlsx”, which contains data on the number of passengers that have flew in a specific airplane per month. Now do the following:
   a. Summarize the data. What is the minimum and maximum number of passengers who flew in this airplane?
   b. Make a histogram using the default *hist()* function. How would you describe the data distribution?
   c. Define the number of breaks and choose 5 breaks. The HELP tab can help you here.
   d. Change the number of breaks now to 20. Compare this histogram with the one obtained in item b.

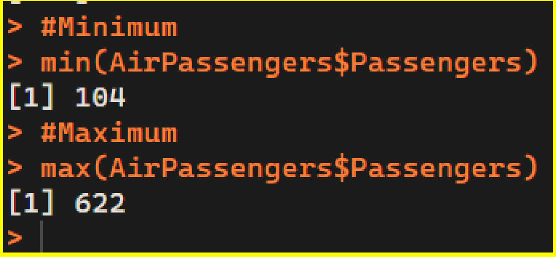
A. Vi har sorteret værdierne og derved fundet største og mindsteværdien.



```
"Opgave 4.a"
sort(AirPassengers$Passengers)
#Minimum
min(AirPassengers$Passengers)
#Maximum
max(AirPassengers$Passengers)
```

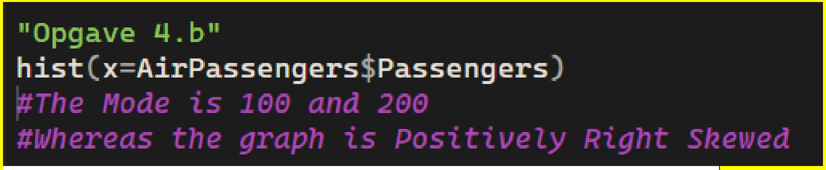


```
> sort(AirPassengers$Passengers)
  [1] 104 112 114 115 118 118 119 121 125 126 129 132 133 135 135 136 140 141 145 146 148 148
 [23] 149 150 158 162 163 166 170 170 171 172 172 178 178 180 180 181 183 184 188 191 193 194
 [45] 196 196 199 199 201 203 204 209 211 218 227 229 229 229 230 233 234 235 235 236 237 237
 [67] 242 242 243 259 264 264 267 269 270 271 272 274 277 278 284 293 301 302 305 306 306 310
 [89] 312 313 315 315 317 318 318 336 337 340 342 347 347 348 348 355 355 356 359 360 362 362
[111] 363 364 374 390 391 396 404 404 405 405 406 407 413 417 419 420 422 432 435 461 461 463
[133] 465 467 472 472 491 505 508 535 548 559 606 622
```

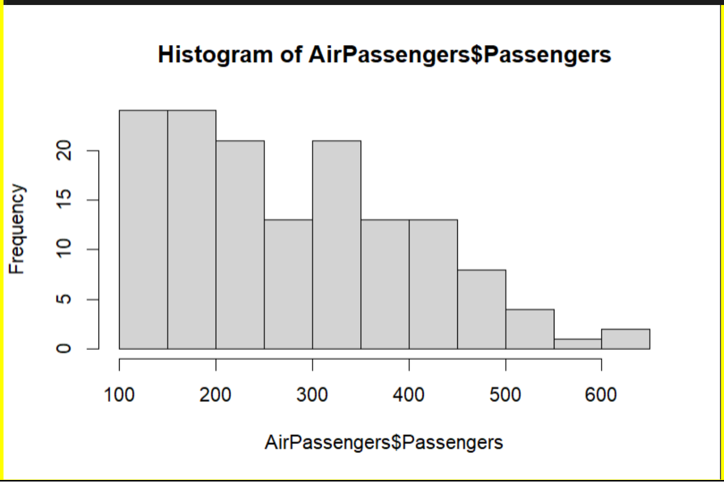


```
> #Minimum
> min(AirPassengers$Passengers)
[1] 104
> #Maximum
> max(AirPassengers$Passengers)
[1] 622
>
```

B. **Vi kan se, at vi har i tilfældet brugt hist(x=) funktionen.**



```
"Opgave 4.b"
hist(x=AirPassengers$Passengers)
#The Mode is 100 and 200
#Whereas the graph is Positively Right Skewed
```



5) Import the dataset "basketball.csv", which contains the scores obtained by three professional basketball players in the pre-season games. Make a boxplot for each of the players. When looking at the boxplots, who seems to be the best player? Can we be sure on this result?

- **Prøv, at løse denne opgave uden facitliste** 😉

You can also use the time in the class to check your answers on the point-giving activity (in ItsLearning) and ask the instructors in case there is any remaining question. :)