INTRODUCTION TO R

Bioinformàtica per a la Recerca Biomèdica
Ricardo Gonzalo Sanz
ricardo.gonzalo@vhir.org

1. Some data types in R

2. Getting data into R

3. Access to data

4. Plots with R

5. Install packages

6. Help!!

1. **Some data types in R**

2. **Getting data into R**

3. **Access to data**

4. **Plots with R**

5. **Install packages**

6. **Help!!**

- R store information in different types of objects.

- The most common are:

  1. **<u>Variables</u>**: A variable is a name for a value. We can create a new variable by assigning a value to it, using <-

```
width <- 5
width
## [1] 5
```

2. **<u>Vectors</u>**: a collection of numbers or characters:

```
myvec <- c(10,20,30,40,50)
myvec
## [1] 10 20 30 40 50

myvec + 1
## [1] 11 21 31 41 51

myvec + myvec
## [1]  20  40  60  80 100

food <- c("eggs", "beans", "bacon", "sausage")
```

3. **Dataframe**: is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column :

```
iris
      Species Sepal.Length Sepal.Width Petal.Length Petal.Width
       setosa          5.1         3.5          1.4         0.2
       setosa          4.9         3.0          1.4         0.2
       setosa          4.7         3.2          1.3         0.2
   versicolor          7.0         3.2          4.7         1.4
   versicolor          6.4         3.2          4.5         1.5
   versicolor          6.9         3.1          4.9         1.5
   versicolor          5.5         2.3          4.0         1.3
    virginica          6.3         3.3          6.0         2.5
    virginica          5.8         2.7          5.1         1.9
    virginica          7.1         3.0          5.9         2.1
    virginica          6.3         2.9          5.6         1.8
    virginica          6.5         3.0          5.8         2.2
```

4. <u>List</u>: is a generic vector containing other objects

```
n <- c(2, 3, 5)
s <- c("aa", "bb", "cc", "dd", "ee")
b <- c(TRUE, FALSE, TRUE, FALSE, FALSE)
x <- list(n, s, b, 3)   # x contains copies of n, s, b

x[2]

[[1]]
[1] "aa" "bb" "cc" "dd" "ee"
```

5. <u>Matrix</u>: is a collection of data elements arranged in a two-dimensional rectangular layout:

```
A <- matrix(c(2, 4, 3, 1, 5, 7), nrow=2, ncol=3,byrow = TRUE)

 A
[,1] [,2] [,3]
[1,]    2    4    3
[2,]    1    5    7
```
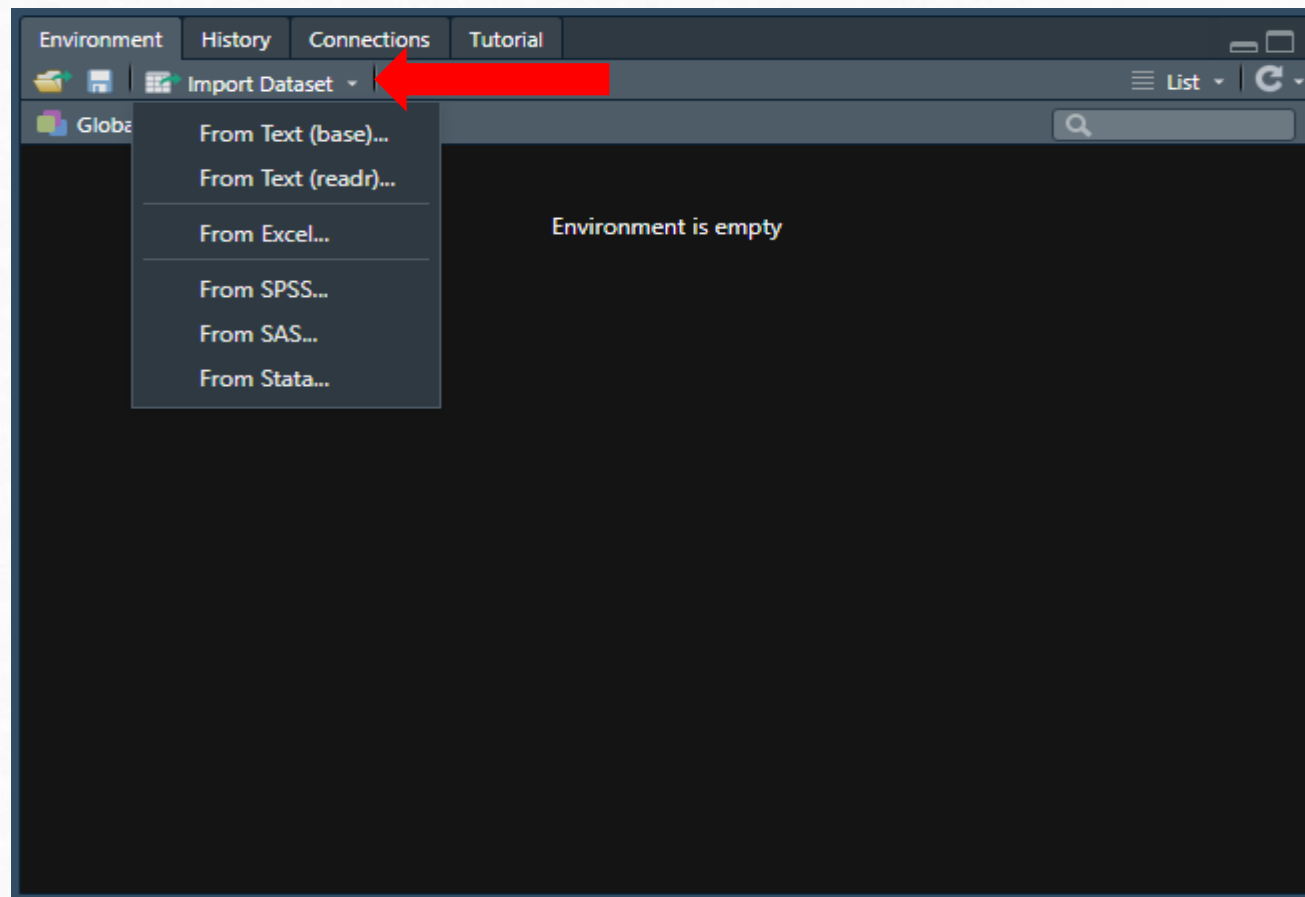
1. **Some data types in R**

2. **Getting data into R**

3. **Access to data**
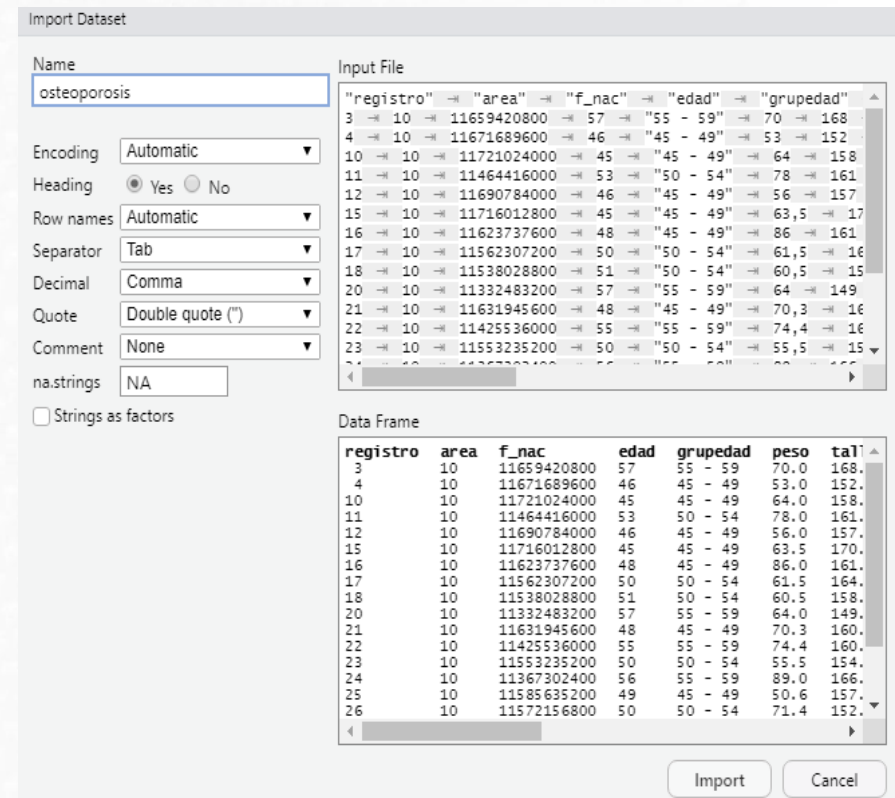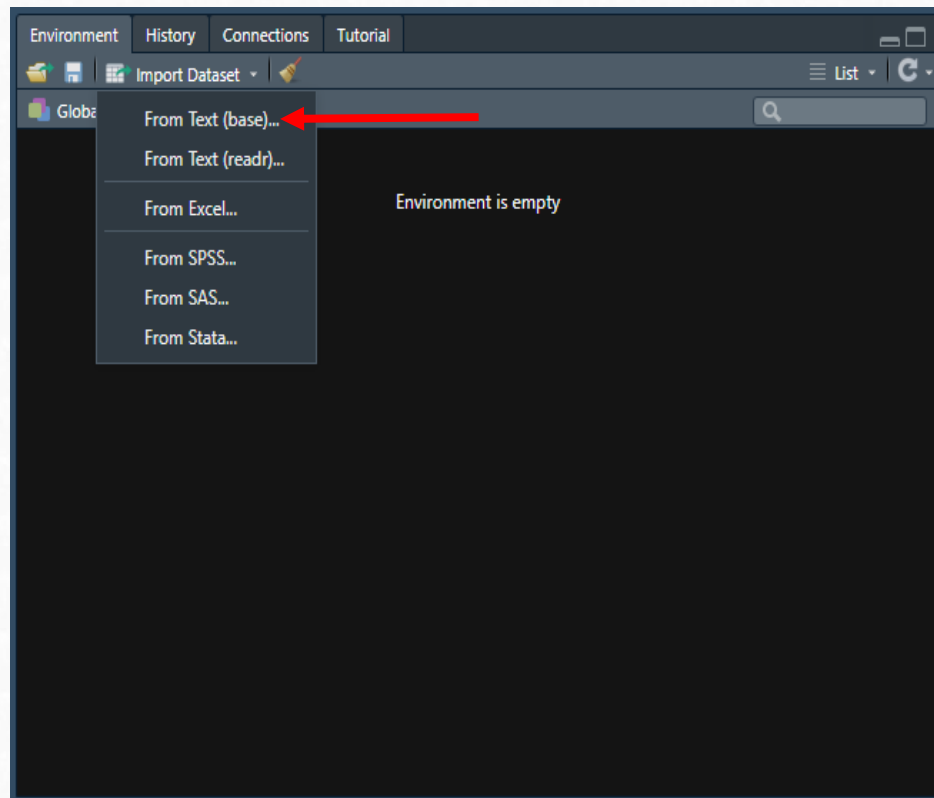
4. **Plots with R**

5. **Install packages**

6. **Help!!**

- The easiest way to get data into R is to click on the "Import Datasets" button.

- The easiest way to get data into R is to click on the "Import Datasets" button.

registro	area	f_nac	edad	grupedad	peso	talla	imc	bua	clasific	menarqui	edad_men	menop	tipo_men	nivel_ed
1	3	10	11659420800	57	55 - 59	70.0	168.0	24.80	69	OSTEOPENIA	12	99	NO	NO MENOPAUSIA/NO CONSTA	SECUNDARIOS
2	4	10	11671689600	46	45 - 49	53.0	152.0	22.94	73	OSTEOPENIA	13	99	NO	NO MENOPAUSIA/NO CONSTA	SECUNDARIOS
3	10	10	11721024000	45	45 - 49	64.0	158.0	25.64	81	NORMAL	14	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
4	11	10	11464416000	53	50 - 54	78.0	161.0	30.09	58	OSTEOPENIA	10	50	SI	NATURAL	PRIMARIOS
5	12	10	11690784000	46	45 - 49	56.0	157.0	22.72	89	NORMAL	13	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
6	15	10	11716012800	45	45 - 49	63.5	170.0	21.97	76	NORMAL	14	99	NO	NO MENOPAUSIA/NO CONSTA	SECUNDARIOS
7	16	10	11623737600	48	45 - 49	86.0	161.0	33.18	87	NORMAL	11	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
8	17	10	11562307200	50	50 - 54	61.5	164.0	22.87	74	NORMAL	10	99	NO	NO MENOPAUSIA/NO CONSTA	PRIMARIOS
9	18	10	11538028800	51	50 - 54	60.5	158.0	24.23	58	OSTEOPENIA	14	99	NO	NO MENOPAUSIA/NO CONSTA	SECUNDARIOS
10	20	10	11332483200	57	55 - 59	64.0	149.0	28.83	61	OSTEOPENIA	13	50	SI	AMBAS	PRIMARIOS
11	21	10	11631945600	48	45 - 49	70.3	160.0	27.46	67	OSTEOPENIA	12	48	SI	OVARIECTOMIA	SECUNDARIOS
12	22	10	11425536000	55	55 - 59	74.4	160.0	29.06	68	OSTEOPENIA	14	50	SI	NATURAL	PRIMARIOS
13	23	10	11553235200	50	50 - 54	55.5	154.5	23.25	73	OSTEOPENIA	11	48	SI	NATURAL	PRIMARIOS
14	24	10	11367302400	56	55 - 59	89.0	166.0	32.30	61	OSTEOPENIA	14	47	SI	NATURAL	PRIMARIOS
15	25	10	11585635200	49	45 - 49	50.6	157.0	20.53	68	OSTEOPENIA	14	40	SI	NATURAL	PRIMARIOS
16	26	10	11572156800	50	50 - 54	71.4	152.0	30.90	74	NORMAL	14	48	SI	AMBAS	PRIMARIOS
17	27	10	11590992000	49	45 - 49	78.0	157.0	31.64	62	OSTEOPENIA	12	46	SI	NATURAL	PRIMARIOS
18	28	10	11293516800	58	55 - 59	72.0	162.0	27.43	65	OSTEOPENIA	11	54	SI	NATURAL	PRIMARIOS
19	29	10	11215238400	61	60 - 64	68.0	155.5	28.12	65	OSTEOPENIA	14	50	SI	NATURAL	PRIMARIOS

What type of data is?

With R instructions:

```
osteoporosis2 <- read.csv2("osteoporosis.csv", sep = "\t", dec = ",", header = TRUE)
```

check the folder you are working on!

1. **Some data types in R**

2. **Getting data into R**

3. **Access to data**

4. **Plots with R**

5. **Install packages**

6. **Help!!**

- Once you have loaded the data, it is necessary to check it

- Different ways to do it

  - **head(*name of dataframe*)**: to see the first rows of a dataframe

  - **tail(*name of dataframe*)**: to see the last rows of a dataframe

  - **str(name of dataframe)**: to check the structure of the dataframe and their variables

  - **summary(name of dataframe)**: Little summary of the variables of the dataframe

```
head(osteoporosis)
  registro area        f_nac edad grupedad peso talla   imc bua   clasific menarqui edad_men menop                    tipo_men     nivel_ed
1        3   10 11659420800   57  55 - 59 70.0   168 24.80  69 OSTEOPENIA       12       99    NO NO MENOPAUSIA/NO CONSTA SECUNDARIOS
2        4   10 11671689600   46  45 - 49 53.0   152 22.94  73 OSTEOPENIA       13       99    NO NO MENOPAUSIA/NO CONSTA SECUNDARIOS
3       10   10 11721024000   45  45 - 49 64.0   158 25.64  81    NORMAL       14       99    NO NO MENOPAUSIA/NO CONSTA    PRIMARIOS
4       11   10 11464416000   53  50 - 54 78.0   161 30.09  58 OSTEOPENIA       10       50    SI                 NATURAL    PRIMARIOS
5       12   10 11690784000   46  45 - 49 56.0   157 22.72  89    NORMAL       13       99    NO NO MENOPAUSIA/NO CONSTA    PRIMARIOS
6       15   10 11716012800   45  45 - 49 63.5   170 21.97  76    NORMAL       14       99    NO NO MENOPAUSIA/NO CONSTA SECUNDARIOS
```

```
tail(osteoporosis)
     registro area        f_nac edad grupedad peso talla   imc bua   clasific menarqui edad_men menop tipo_men                     nivel_ed
995      1028   11 11190182400   63  60 - 64   71   161 27.39  57 OSTEOPENIA       14       48    SI NATURAL                     PRIMARIOS
996      1029   11 11287036800   60  60 - 64   64   158 25.64  69 OSTEOPENIA       10       40    SI   AMBAS                     SUPERIORES
997      1030   11 11066371200   67  65 - 69   68   157 27.59  75    NORMAL       11       55    SI NATURAL PRIMARIOS SIN FINALIZAR
998      1031   11 11289196800   59  55 - 59   72   153 30.76  67 OSTEOPENIA       12       56    SI NATURAL                     PRIMARIOS
999      1032   11 11137219200   64  60 - 64   80   152 34.63  55 OSTEOPENIA       14       50    SI NATURAL                     PRIMARIOS
1000     1033   11 11213164800   62  60 - 64   67   161 25.85  65 OSTEOPENIA       13       54    SI NATURAL                     SECUNDARIOS
```

```
str(osteoporosis)
'data.frame':        1000 obs. of  15 variables:
 $ registro: int   3 4 10 11 12 15 16 17 18 20 ...
 $ area    : int   10 10 10 10 10 10 10 10 10 10 ...
 $ f_nac   : chr   "11659420800" "11671689600" "11721024000" "11464416000" ...
 $ edad    : int   57 46 45 53 46 45 48 50 51 57 ...
 $ grupedad: chr   "55 - 59" "45 - 49" "45 - 49" "50 - 54" ...
 $ peso    : num   70 53 64 78 56 63.5 86 61.5 60.5 64 ...
 $ talla   : num   168 152 158 161 157 170 161 164 158 149 ...
 $ imc     : num   24.8 22.9 25.6 30.1 22.7 ...
 $ bua     : int   69 73 81 58 89 76 87 74 58 61 ...
 $ clasific: chr   "OSTEOPENIA" "OSTEOPENIA" "NORMAL" "OSTEOPENIA" ...
 $ menarqui: int   12 13 14 10 13 14 11 10 14 13 ...
 $ edad_men: int   99 99 99 50 99 99 99 99 99 50 ...
 $ menop   : chr   "NO" "NO" "NO" "SI" ...
 $ tipo_men: chr   "NO MENOPAUSIA/NO CONSTA" "NO MENOPAUSIA/NO CONSTA" "NO MENOPAUSIA/NO
CONSTA" "NATURAL" ...
 $ nivel_ed: chr   "SECUNDARIOS" "SECUNDARIOS" "PRIMARIOS" "PRIMARIOS" ...
```

3. Access to data

```
summary(osteoporosis)
    registro          area            f_nac                edad          grupedad             peso            talla              imc
 Min.   :   3.0   Min.   :10.00   Length:1000       Min.   :45.00   Length:1000       Min.   : 44.00   Min.   :138.0   Min.   :17.21
 1st Qu.: 280.8   1st Qu.:10.00   Class :character  1st Qu.:48.00   Class :character  1st Qu.: 60.50   1st Qu.:153.0   1st Qu.:24.80
 Median : 531.5   Median :11.00   Mode  :character  Median :52.00   Mode  :character  Median : 68.00   Median :157.0   Median :27.51
 Mean   : 529.9   Mean   :11.58                     Mean   :53.42                     Mean   : 69.12   Mean   :156.9   Mean   :28.11
 3rd Qu.: 781.2   3rd Qu.:13.00                     3rd Qu.:58.00                     3rd Qu.: 75.00   3rd Qu.:161.0   3rd Qu.:30.82
 Max.   :1033.0   Max.   :13.00                     Max.   :69.00                     Max.   :123.50   Max.   :180.0   Max.   :48.39
      bua           clasific          menarqui          edad_men          menop             tipo_men          nivel_ed
 Min.   : 11.0   Length:1000       Min.   : 8.00   Min.   :24.00   Length:1000       Length:1000       Length:1000
 1st Qu.: 62.0   Class :character  1st Qu.:12.00   1st Qu.:46.00   Class :character  Class :character  Class :character
 Median : 72.0   Mode  :character  Median :13.00   Median :51.00   Mode  :character  Mode  :character  Mode  :character
 Mean   : 73.3                     Mean   :12.71   Mean   :63.04
 3rd Qu.: 84.0                     3rd Qu.:14.00   3rd Qu.:99.00
 Max.   :136.0                     Max.   :17.00   Max.   :99.00
```

```
nrow(osteoporosis)
[1] 1000

ncol(osteoporosis)
[1] 15

colnames(osteoporosis)
 [1] "registro" "area"     "f_nac"    "edad"     "grupedad" "peso"     "talla"    "imc"
"bua"      "clasific" "menarqui" "edad_men" "menop"
[14] "tipo_men" "nivel_ed"
```

- Sometimes you are not interested in working with all the variables of cases of the dataset
- There are different ways to select them:

```
#select the first 3 rows and columns
osteoporosis[1:3, 1:3]
```

```
  registro area        f_nac
1        3   10 11659420800
2        4   10 11671689600
3       10   10 11721024000
```

```
#select the different rows and columns
osteoporosis[c(31:36, 115, 950), c(2, 4:6,13:15)]

    area edad grupedad peso menop                  tipo_men      nivel_ed
31    10   67  65 - 69 60.0  SI                     NATURAL      PRIMARIOS
32    10   50  50 - 54 70.0  SI                     NATURAL      PRIMARIOS
33    10   56  55 - 59 68.0  SI                     NATURAL      PRIMARIOS
34    10   67  65 - 69 63.0  SI                     NATURAL      PRIMARIOS
35    10   58  55 - 59 64.0  SI                     NATURAL      PRIMARIOS
36    10   57  55 - 59 75.0  SI                     NATURAL  SIN ESTUDIOS
115   10   45  45 - 49 57.0  NO NO MENOPAUSIA/NO CONSTA  SECUNDARIOS
950   11   62  60 - 64 75.5  SI                     NATURAL      PRIMARIOS


#select all the rows and by variable name and save the dataset with another name
osteoporosis2 <- osteoporosis[, c("edad", "grupedad", "tipo_men", "nivel_ed")]
head(osteoporosis2)

    edad grupedad                  tipo_men      nivel_ed
1    57  55 - 59 NO MENOPAUSIA/NO CONSTA SECUNDARIOS
2    46  45 - 49 NO MENOPAUSIA/NO CONSTA SECUNDARIOS
3    45  45 - 49 NO MENOPAUSIA/NO CONSTA   PRIMARIOS
4    53  50 - 54                 NATURAL   PRIMARIOS
5    46  45 - 49 NO MENOPAUSIA/NO CONSTA   PRIMARIOS
6    45  45 - 49 NO MENOPAUSIA/NO CONSTA SECUNDARIOS
```

- It is also possible to select some rows depending on logic expressions:

```
#select pacients older than 60 years
osteoporosis3 <- osteoporosis[which(osteoporosis$edad > 60), ]
head(osteoporosis3)

   registro area      f_nac edad grupedad peso  talla   imc bua      clasific menarqui edad_men menop tipo_men                  nivel_ed
19       29   10 11215238400   61  60 - 64 68.0 155.5 28.12  65    OSTEOPENIA       14       50    SI  NATURAL                  PRIMARIOS
23       34   10 10992758400   68  65 - 69 66.5 145.0 31.63  57    OSTEOPENIA       13       50    SI  NATURAL PRIMARIOS SIN FINALIZAR
24       35   10 10909382400   69  65 - 69 70.0 168.0 24.80  48 OSTEOPOROSIS       13       45    SI  NATURAL                  PRIMARIOS
27       38   10 11043907200   66  65 - 69 67.0 144.0 32.31  79       NORMAL       12       56    SI  NATURAL                  PRIMARIOS
28       39   10 10948089600   69  65 - 69 70.5 148.5 31.97  40 OSTEOPOROSIS       11       43    SI  NATURAL               SIN ESTUDIOS
29       40   10 11051251200   66  65 - 69 66.5 147.0 30.77  48 OSTEOPOROSIS       13       40    SI  NATURAL                  PRIMARIOS

osteoporosis4 <- subset(osteoporosis, edad > 60)
head(osteoporosis4)

   registro area      f_nac edad grupedad peso  talla   imc bua      clasific menarqui edad_men menop tipo_men                  nivel_ed
19       29   10 11215238400   61  60 - 64 68.0 155.5 28.12  65    OSTEOPENIA       14       50    SI  NATURAL                  PRIMARIOS
23       34   10 10992758400   68  65 - 69 66.5 145.0 31.63  57    OSTEOPENIA       13       50    SI  NATURAL PRIMARIOS SIN FINALIZAR
24       35   10 10909382400   69  65 - 69 70.0 168.0 24.80  48 OSTEOPOROSIS       13       45    SI  NATURAL                  PRIMARIOS
27       38   10 11043907200   66  65 - 69 67.0 144.0 32.31  79       NORMAL       12       56    SI  NATURAL                  PRIMARIOS
28       39   10 10948089600   69  65 - 69 70.5 148.5 31.97  40 OSTEOPOROSIS       11       43    SI  NATURAL               SIN ESTUDIOS
29       40   10 11051251200   66  65 - 69 66.5 147.0 30.77  48 OSTEOPOROSIS       13       40    SI  NATURAL                  PRIMARIOS
```

- We can combine different logic expressions

```
osteoporosis4 <- subset(osteoporosis, edad > 60 & nivel_ed == "PRIMARIOS")
head(osteoporosis4)
```

	registro	area	f_nac	edad	grupedad	peso	talla	imc	bua	clasific	menarqui	edad_men	menop	tipo_men	nivel_ed
19	29	10	11215238400	61	60 - 64	68.0	155.5	28.12	65	OSTEOPENIA	14	50	SI	NATURAL	PRIMARIOS
24	35	10	10909382400	69	65 - 69	70.0	168.0	24.80	48	OSTEOPOROSIS	13	45	SI	NATURAL	PRIMARIOS
27	38	10	11043907200	66	65 - 69	67.0	144.0	32.31	79	NORMAL	12	56	SI	NATURAL	PRIMARIOS
29	40	10	11051251200	66	65 - 69	66.5	147.0	30.77	48	OSTEOPOROSIS	13	40	SI	NATURAL	PRIMARIOS
31	45	10	11029651200	67	65 - 69	60.0	147.0	27.77	49	OSTEOPENIA	13	53	SI	NATURAL	PRIMARIOS
34	48	10	11034489600	67	65 - 69	63.0	157.0	25.56	66	OSTEOPENIA	13	50	SI	NATURAL	PRIMARIOS

- R is a powerful tool to plot your data

- Hadley Wickam (2009) introduced a modern (and perhaps easier) way to plot your data

- Extensions to ggplot2

  - GGally, ggrepel, …

Hadley Wickam book
http://moderngraphics11.pbworks.com/f/ggplot2-Book09hWickham.pdf
https://ggplot2-book.org/

STHDA (Statistical tools for high-throughput data analysis)
http://www.sthda.com/english/wiki/ggplot2-essentials
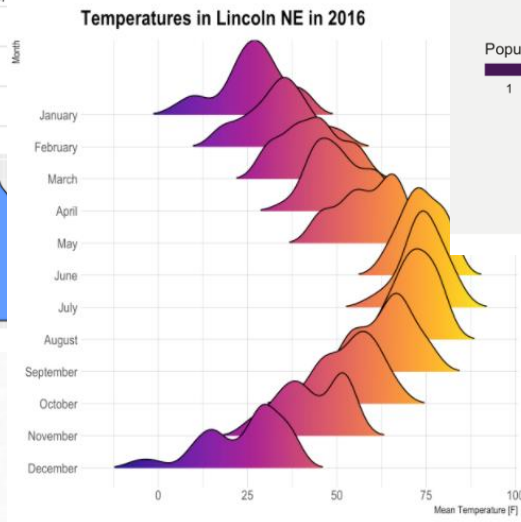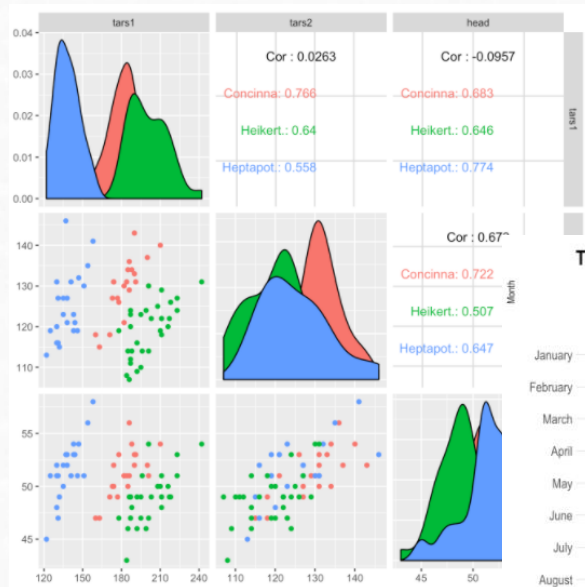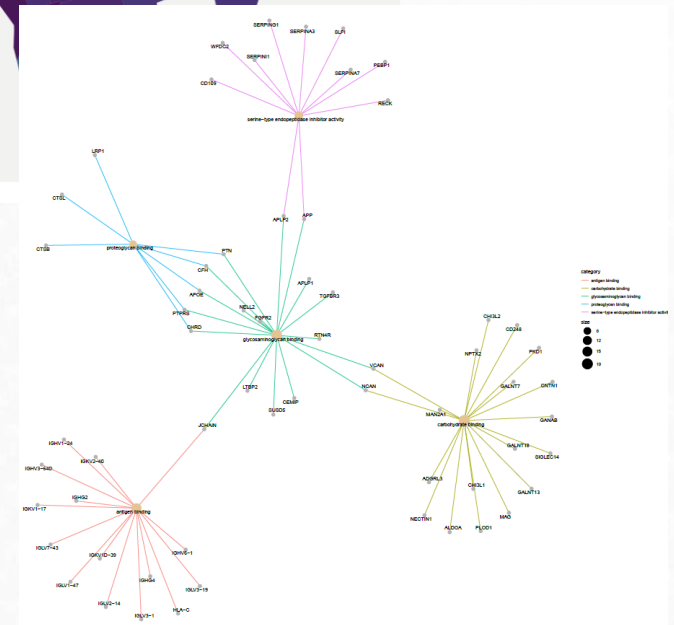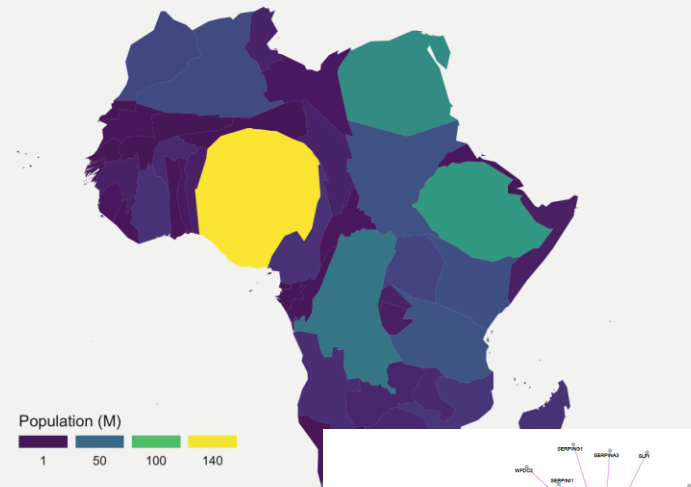
R Colors
http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

4. Plots with R

https://www.r-graph-gallery.com/

```r
library(ggplot2) #Remember to install the packages before call it

ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point()
```

```
library(ggplot2) #Remember to install the packages before call it

ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point()
```

```
library(ggplot2) #Remember to install the packages before call it

ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point()
```
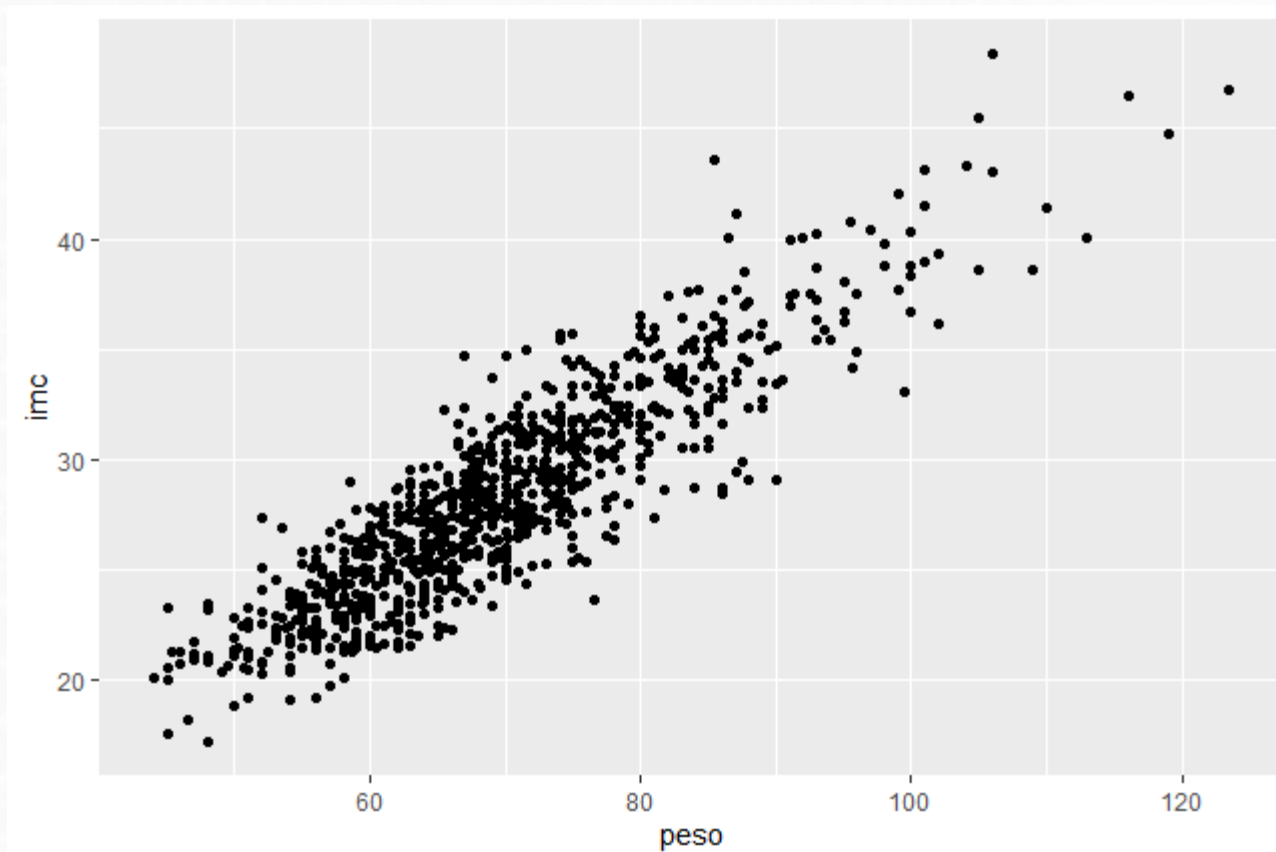
```
ggplot(osteoporosis, aes(x = peso, y = imc, color = grupedad)) +
    geom_point()
```
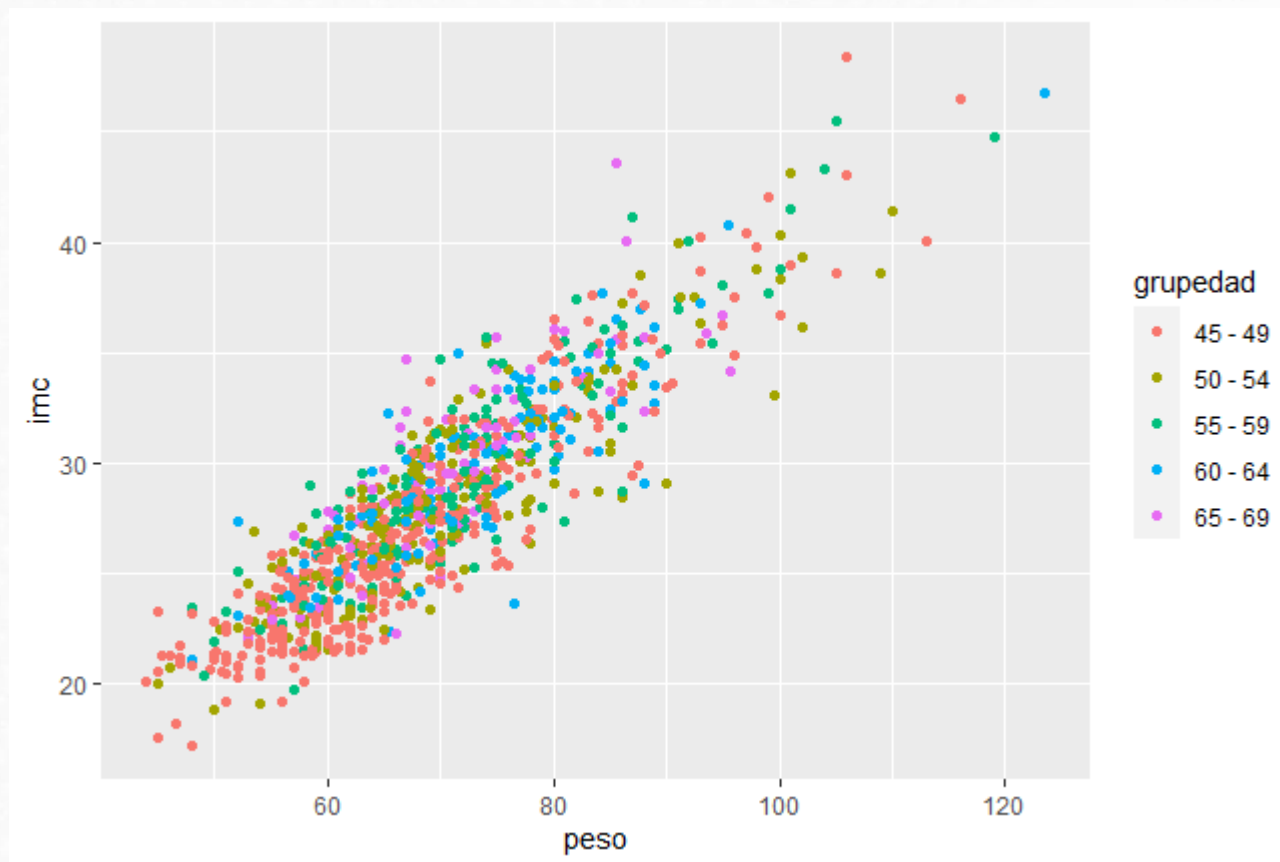
```
bp <- ggplot(osteoporosis, aes(x = grupedad, y = bua)) +
        geom_boxplot(fill='#A4A4A4', color="darkred")
```
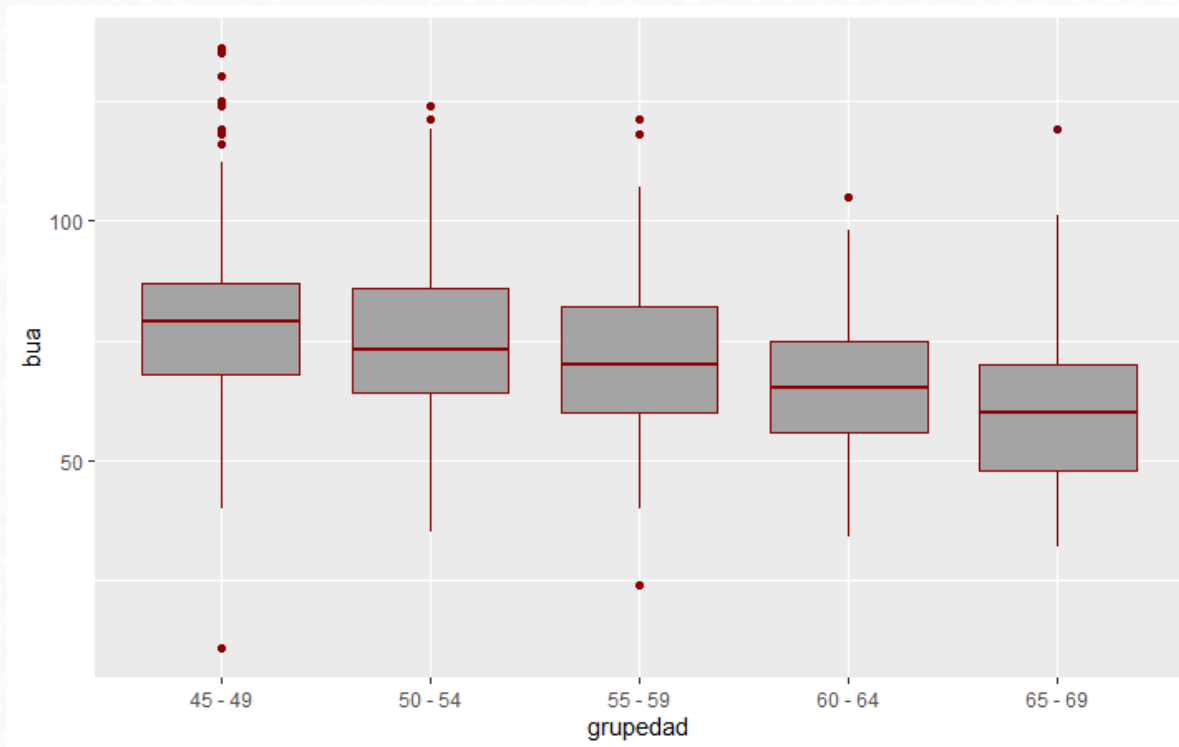
```
bp + geom_jitter(shape = 16, position = position_jitter(0.2)) +
        labs(x = "Age Group", y = "Women", title = "Osteo disease classified by age group")
```

1. **Some data types in R**

2. **Getting data into R**

3. **Access to data**

4. **Plots with R**

5. **Install packages**

6. **Help!!**

- Packages are extensions of the basic R functions

- It is necessary to install once in each computer when needed

- Call it each time you want to use it

- Two ways to install the packages:

  - install.packages("name of the package")

5. Install packages

- If package are hosted in Bioconductor repository:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")


BiocManager::install("airway")
```
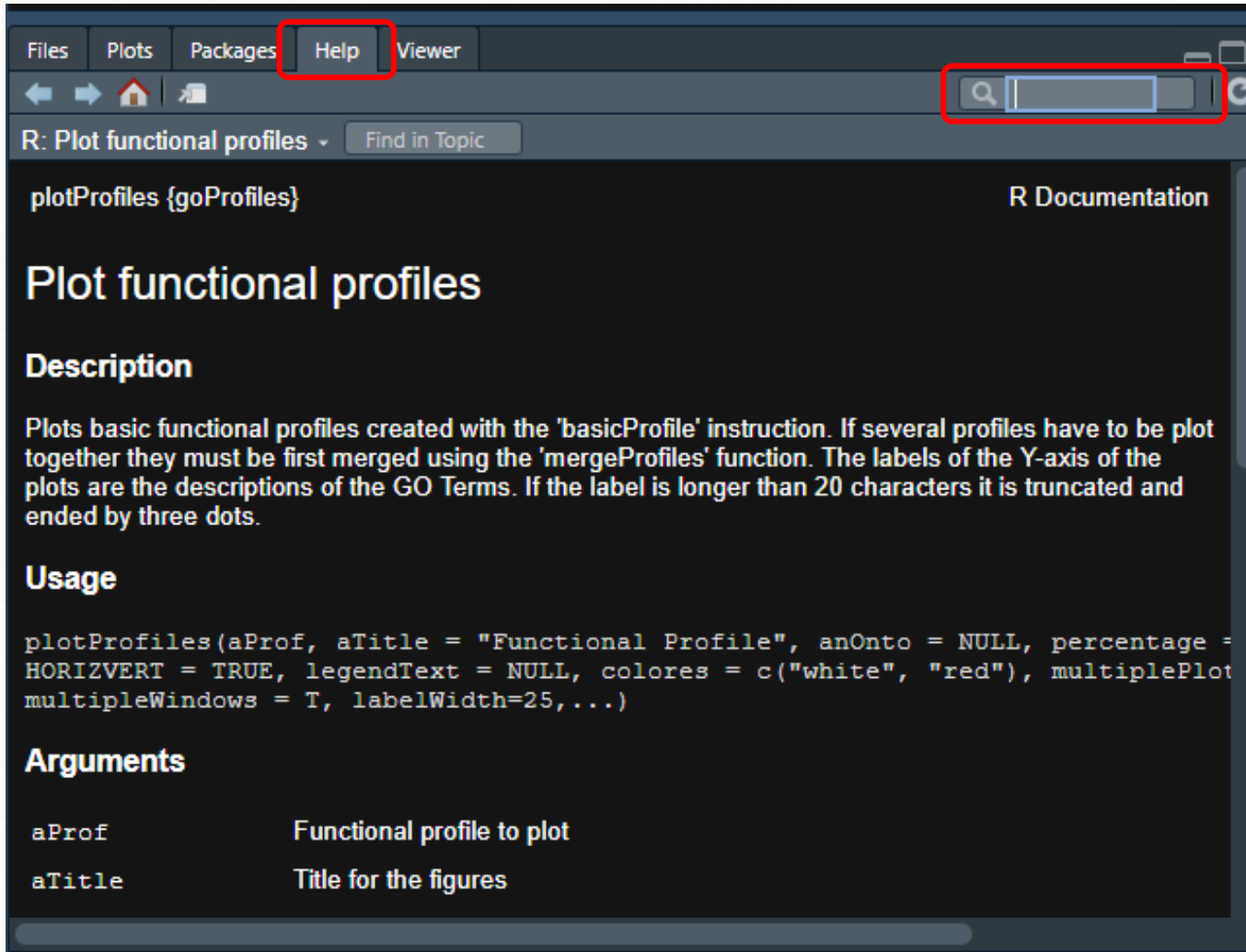
Bioconductor
airway
Rsamtools
GenomicFeatures
GenomicAlignments
BiocParallel
DESeq2
apeglm
genefilter
AnnotationDbi
org.Hs.eg.db
Gviz
ReportingTools
sva
RUVSeq
AnnotationHub
clusterProfiler
ReactomePA
topGO
GOstats
pasilla
IHW

Cran R
magrittr
dplyr
ggplot2
pheatmap
RColorBrewer
ggbeeswarm

1. **Some data types in R**

2. **Getting data into R**

3. **Access to data**

4. **Plots with R**

5. **Install packages**

6. **Help!!**

6. Help!!