Statistics with R: Exploratory Data Analysis II and Plots with R

Statistics with R - VHIO 2020

Ricardo Gonzalo Sanz
ricardo.gonzalo@vhir.org

Miriam Mota Foix
miriam.mota@vhir.org

06/10/2020

TABLE OF CONTENTS

TABLE OF CONTENTS

1. Elegant graphics for data analysis

Vall d'Hebron Institut de Recerca

- 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

1. Elegant graphics for data analysis

How ggplot2 works?

- It is based on the *Grammar of Graphics* (Wilkinson 2005)
- Grammar tells us that a graphic maps the data to the aesthetic attributes (colour, shape, size) of geometric objects (points, lines, bars).
- Plot may also include statistical transformations of the data and information about plot's coordinate system

1. Elegant graphics for data analysis

- Mapping components:
    - **Layer**: Geoms (what you actually see in the plot: points, lines,…), stats (summarise the data)
    - **Scales**: how we want to see the data (aesthetic). Color, shapes, legend, axes….
    - **Coord**: axes, gridlines
    - **Facet**: if you want to divide your data in different plots
    - **Theme**: Font size, background colors, …

1. Elegant graphics for data analysis

- How to install: `install.packages("ggplot2")`
- First steps. Three key components:
  - Data
  - Aesthetic mappings between variables
  - A least one layer. Usually created with a **geom** function

1. Elegant graphics for data analysis

- The data: (https://ggplot2.tidyverse.org/reference/mpg.html)

```
head(mpg)
```

```
# A tibble: 6 x 11
  manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
  <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compact
2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compact
3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compact
4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compact
5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compact
6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compact
```

A data frame with 234 rows and 11 variables:

manufacturer: manufacturer name

model: model name

displ: engine displacement, in litres

year: year of manufacture

"type" of car

cyl: number of cylinders

trans: type of transmission

drv: the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd

cty: city miles per gallon

hwy: highway miles per gallon

fl: fuel type

class

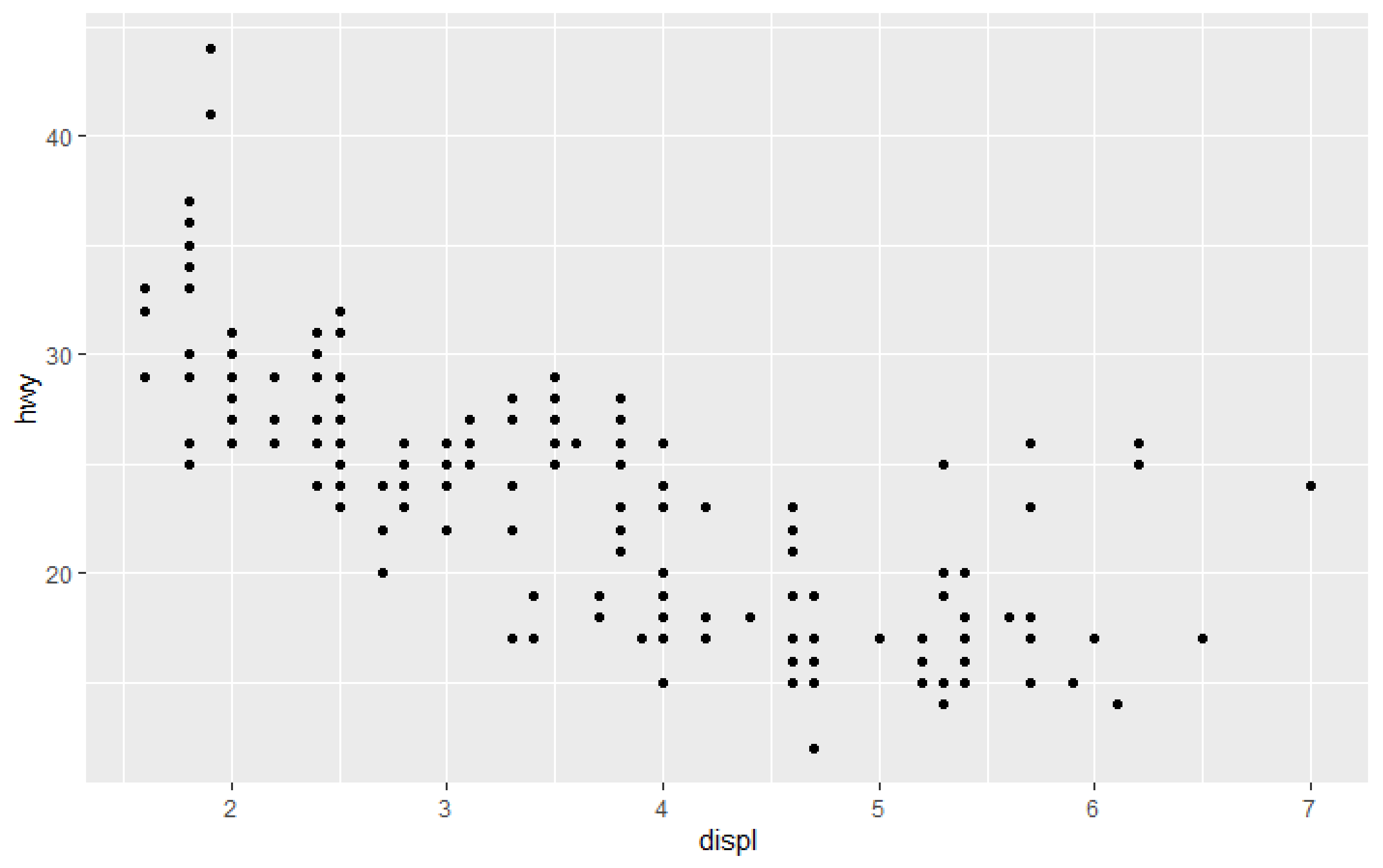
1. Elegant graphics for data analysis

- The plot:

```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point()
```

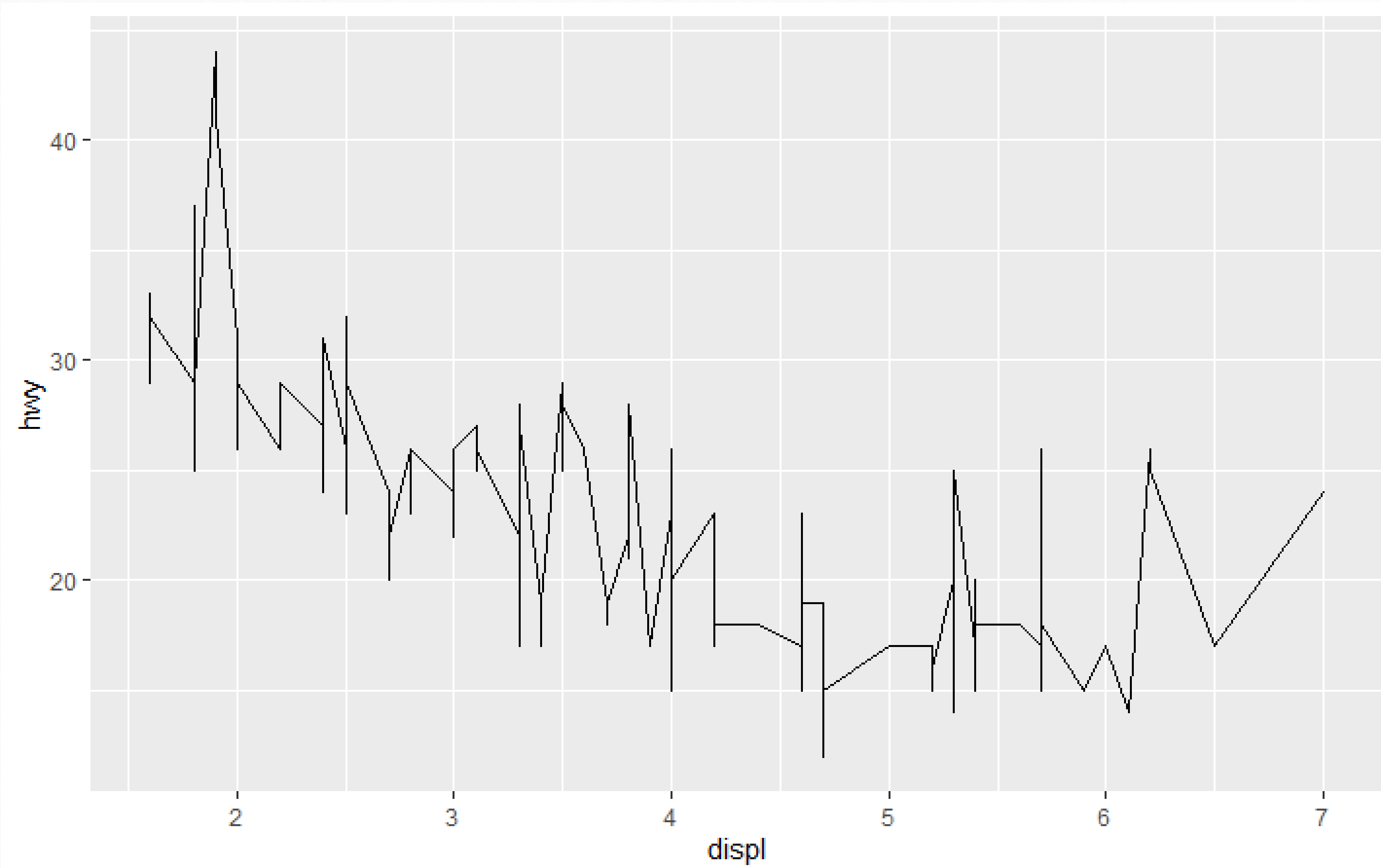
1. Elegant graphics for data analysis

```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point()
```



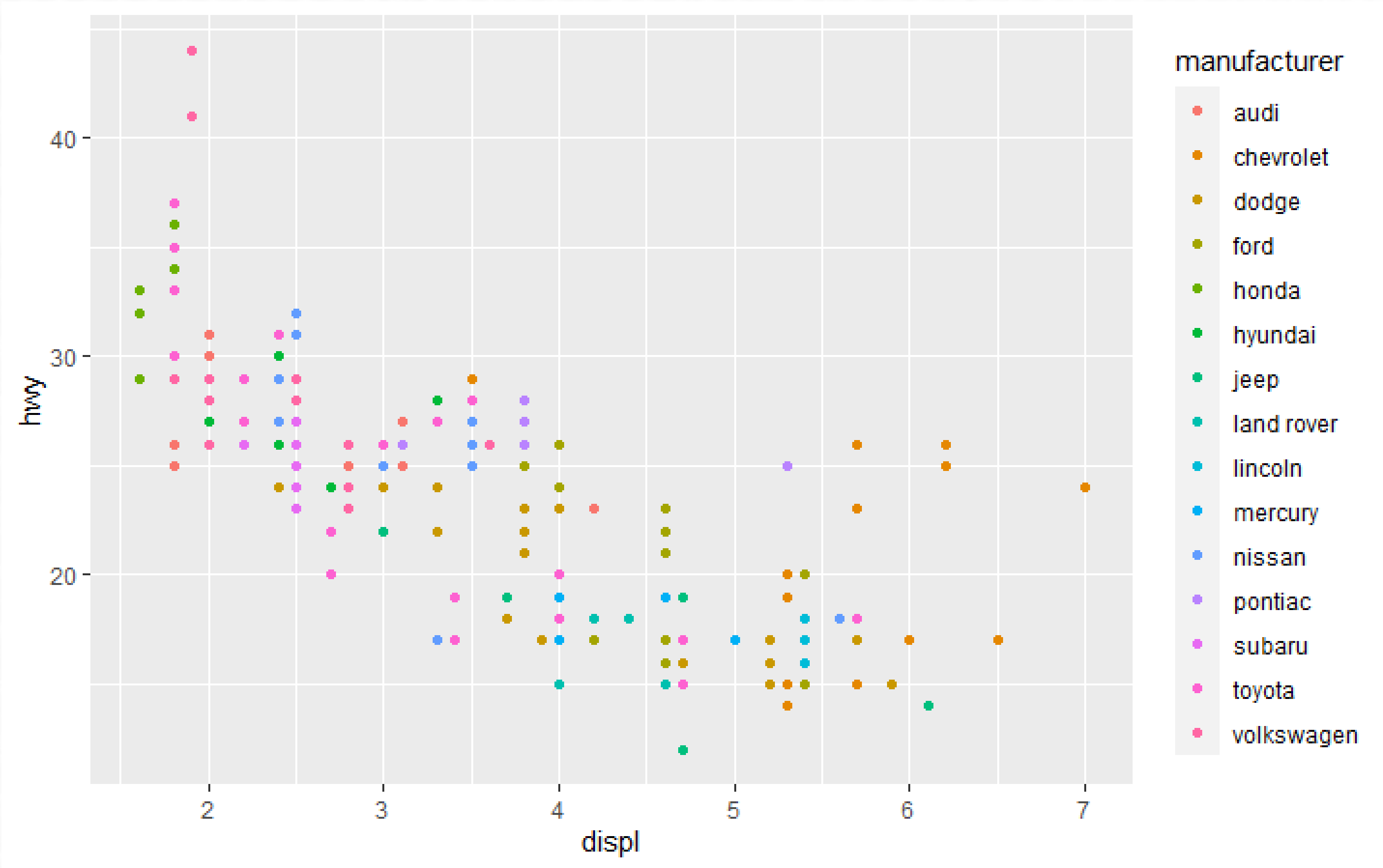
1. Elegant graphics for data analysis

```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_line()
```



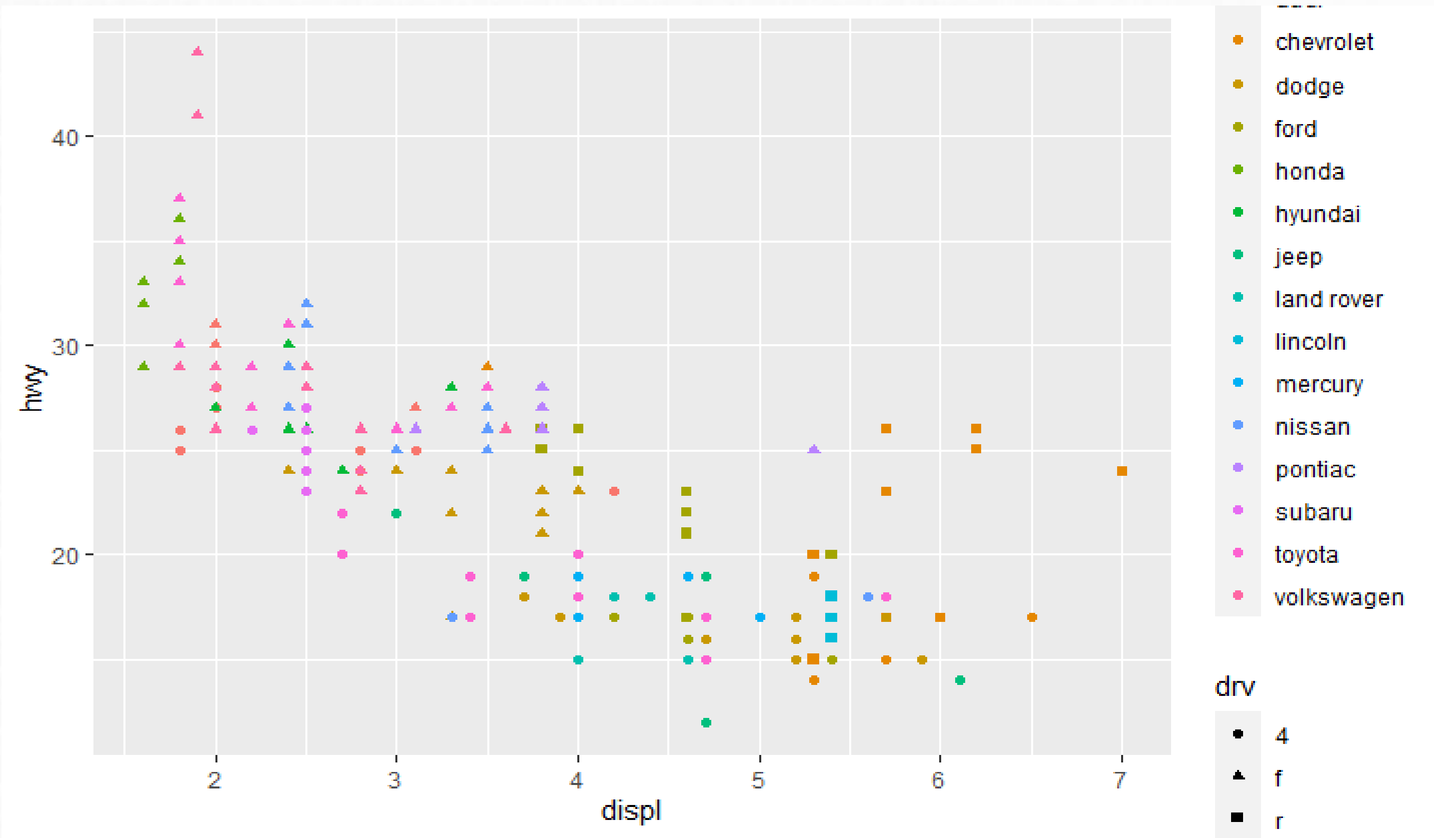
1. Elegant graphics for data analysis

```
ggplot(mpg, aes(x = displ, y = hwy, color = manufacturer)) +
  geom_point()
```

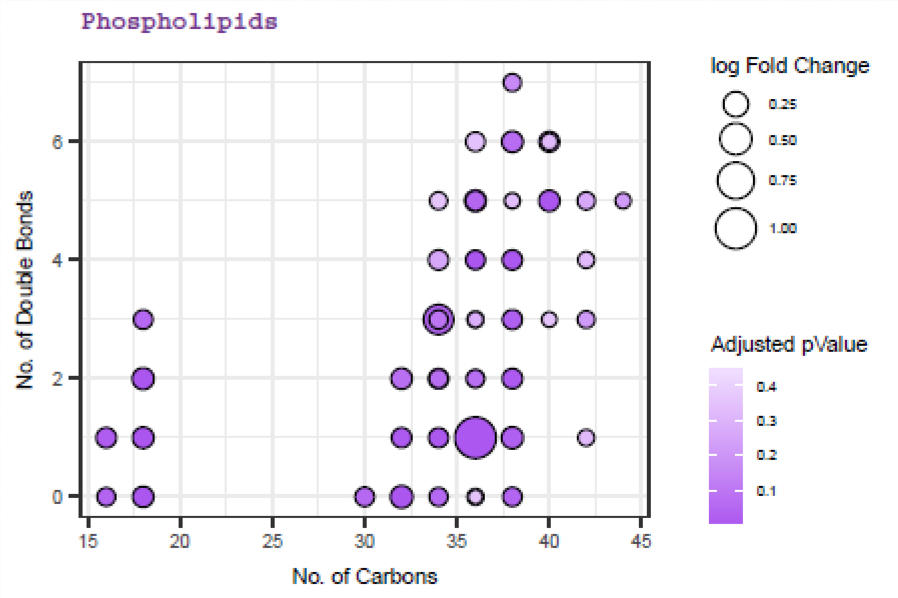


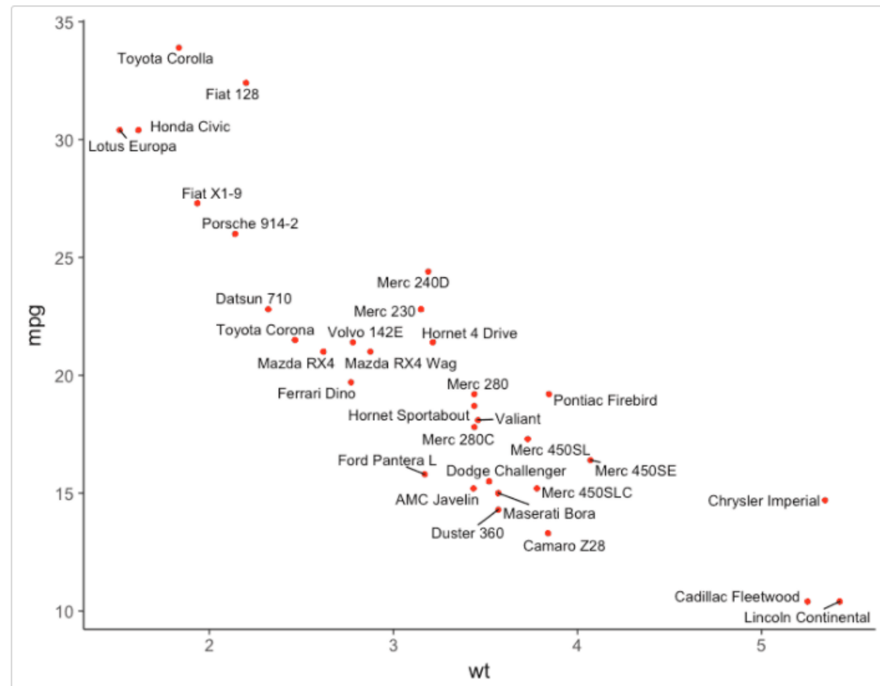
1. Elegant graphics for data analysis

```
ggplot(mpg, aes(x = displ, y = hwy, color = manufacturer, shape =  drv)) +
  geom_point()
```



1. Elegant graphics for data analysis





Vall d'Hebron
Institut de Recerca

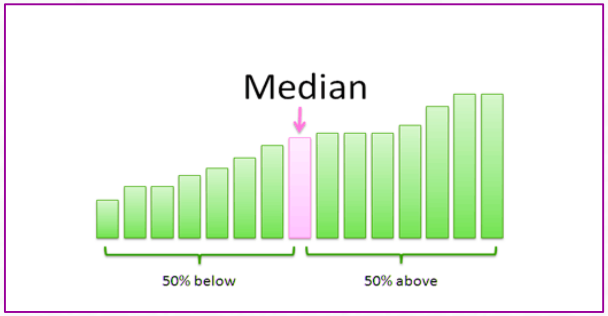
TABLE OF CONTENTS

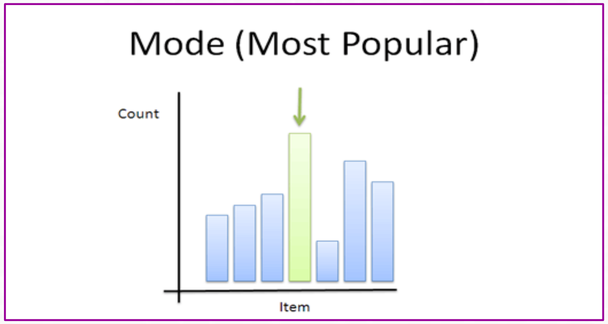
Last week we learned…

- We can analyse and describe each variable one by one:

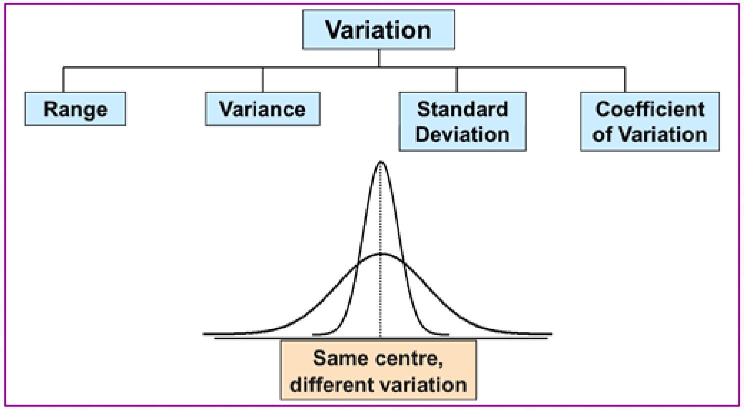
1. With some measures:

Measures of central tendency



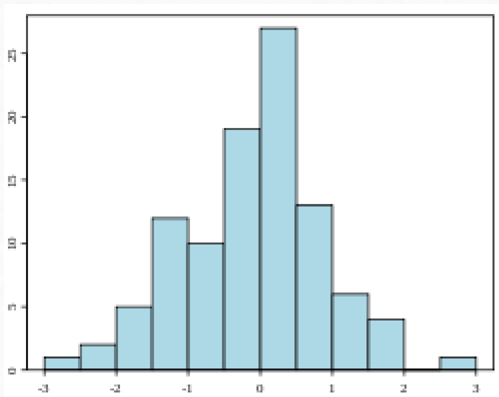


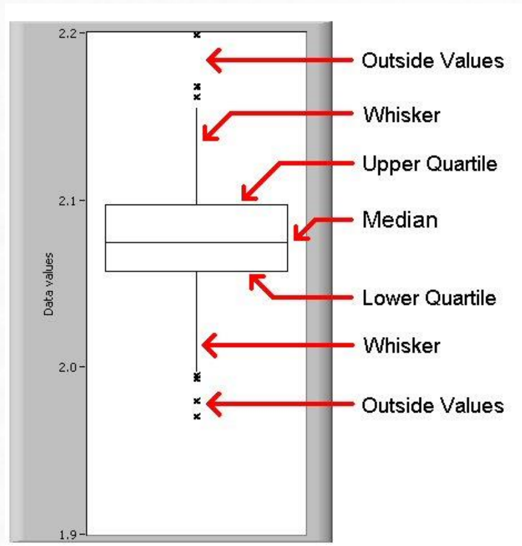
Measures of dispersion

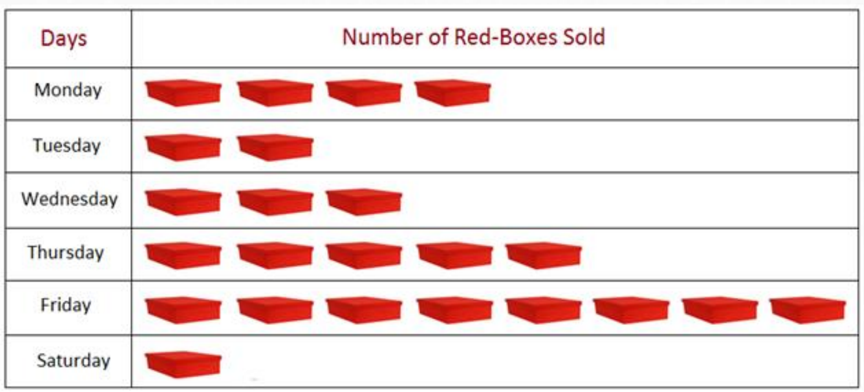


2. From univariate to Bivariate analysis

2. Some graphics







2. From univariate to Bivariate analysis

- In univariate analysis **only one** variable is analyzed each time

↓

the purpose of the analysis is **descriptive**

- If there are more than one variable in the dataset it could be interesting to guess if:
    - Does exist a relation between the two variables?
    - How important is this relation?
    - Which is the direction of the relation?

2. From univariate to Bivariate analysis

osteo

	registro	area	f_nac	edad	grupedad	peso	talla	bua	imc	clasific	me
1	3	10	11659420800	57	55 - 59	70.0	168.0	69	24.80159	OSTEOPENIA	
2	4	10	11671689600	46	45 - 49	53.0	152.0	73	22.93975	OSTEOPENIA	
3	10	10	11721024000	45	45 - 49	64.0	158.0	81	25.63692	NORMAL	
4	11	10	11464416000	53	50 - 54	78.0	161.0	58	30.09143	OSTEOPENIA	
5	12	10	11690784000	46	45 - 49	56.0	157.0	89	22.71897	NORMAL	
6	15	10	11716012800	45	45 - 49	63.5	170.0	76	21.97232	NORMAL	
7	16	10	11623737600	48	45 - 49	86.0	161.0	87	33.17773	NORMAL	
8	17	10	11562307200	50	50 - 54	61.5	164.0	74	22.86585	NORMAL	
9	18	10	11538028800	51	50 - 54	60.5	158.0	58	24.23490	OSTEOPENIA	
10	20	10	11332483200	57	55 - 59	64.0	149.0	61	28.82753	OSTEOPENIA	
11	21	10	11631945600	48	45 - 49	70.3	160.0	67	27.46094	OSTEOPENIA	
12	22	10	11425536000	55	55 - 59	74.4	160.0	68	29.06250	OSTEOPENIA	
13	23	10	11553235200	50	50 - 54	55.5	154.5	73	23.25070	OSTEOPENIA	
14	24	10	11367302400	56	55 - 59	89.0	166.0	61	32.29787	OSTEOPENIA	
15	25	10	11585635200	49	45 - 49	50.6	157.0	68	20.52822	OSTEOPENIA	
16	26	10	11572156800	50	50 - 54	71.4	152.0	74	30.90374	NORMAL	
17	27	10	11590992000	49	45 - 49	78.0	157.0	62	31.64429	OSTEOPENIA	
18	28	10	11293516800	58	55 - 59	72.0	162.0	65	27.43484	OSTEOPENIA	
19	29	10	11215238400	61	60 - 64	68.0	155.5	65	28.12212	OSTEOPENIA	
20	30	10	11405664000	55	55 - 59	75.0	161.0	92	28.93407	NORMAL	
21	31	10	11633155200	48	45 - 49	66.5	153.0	11	28.40788	OSTEOPOROSIS	
22	32	10	11287728000	59	55 - 59	101.0	156.0	82	41.50230	NORMAL	
23	34	10	10992758400	68	65 - 69	66.5	145.0	57	31.62901	OSTEOPENIA	
24	35	10	10909382400	69	65 - 69	70.0	168.0	48	24.80159	OSTEOPOROSIS	
25	36	10	11643868800	48	45 - 49	60.1	153.0	86	25.67389	NORMAL	
26	37	10	11551420800	50	50 - 54	67.0	159.0	105	26.50212	NORMAL	
27	38	10	11043907200	66	65 - 69	67.0	144.0	79	32.31096	NORMAL	
28	39	10	10948089600	69	65 - 69	70.5	148.5	40	31.96953	OSTEOPOROSIS	
29	40	10	11051251200	66	65 - 69	66.5	147.0	48	30.77421	OSTEOPOROSIS	
30	41	10	11333692800	57	55 - 59	58.5	142.0	80	29.01210	NORMAL	

2. From univariate to Bivariate analysis

osteo

	registro	area	f_nac	edad	grupedad	peso	talla	bua	imc	clasific	me
1	3	10	11659420800	57	55 - 59	70.0	168.0	69	24.80159	OSTEOPENIA	
2	4	10	11671689600	46	45 - 49	53.0	152.0	73	22.93975	OSTEOPENIA	
3	10	10	11721024000	45	45 - 49	64.0	158.0	81	25.63692	NORMAL	
4	11	10	11464416000	53	50 - 54	78.0	161.0	58	30.09143	OSTEOPENIA	
5	12	10	11690784000	46	45 - 49	56.0	157.0	89	22.71897	NORMAL	
6	15	10	11716012800	45	45 - 49	63.5	170.0	76	21.97232	NORMAL	
7	16	10	11623737600	48	45 - 49	86.0	161.0	87	33.17773	NORMAL	
8	17	10	11562307200	50	50 - 54	61.5	164.0	74	22.86585	NORMAL	
9	18	10	11538028800	51	50 - 54	60.5	158.0	58	24.23490	OSTEOPENIA	
10	20	10	11332483200	57	55 - 59	64.0	149.0	61	28.82753	OSTEOPENIA	
11	21	10	11631945600	48	45 - 49	70.3	160.0	67	27.46094	OSTEOPENIA	
12	22	10	11425536000	55	55 - 59	74.4	160.0	68	29.06250	OSTEOPENIA	
13	23	10	11553235200	50	50 - 54	55.5	154.5	73	23.25070	OSTEOPENIA	
14	24	10	11367302400	56	55 - 59	89.0	166.0	61	32.29787	OSTEOPENIA	
15	25	10	11585635200	49	45 - 49	50.6	157.0	68	20.52822	OSTEOPENIA	
16	26	10	11572156800	50	50 - 54	71.4	152.0	74	30.90374	NORMAL	
17	27	10	11590992000	49	45 - 49	78.0	157.0	62	31.64429	OSTEOPENIA	
18	28	10	11293516800	58	55 - 59	72.0	162.0	65	27.43484	OSTEOPENIA	
19	29	10	11215238400	61	60 - 64	68.0	155.5	65	28.12212	OSTEOPENIA	
20	30	10	11405664000	55	55 - 59	75.0	161.0	92	28.93407	NORMAL	
21	31	10	11633155200	48	45 - 49	66.5	153.0	11	28.40788	OSTEOPOROSIS	
22	32	10	11287728000	59	55 - 59	101.0	156.0	82	41.50230	NORMAL	
23	34	10	10992758400	68	65 - 69	66.5	145.0	57	31.62901	OSTEOPENIA	
24	35	10	10909382400	69	65 - 69	70.0	168.0	48	24.80159	OSTEOPOROSIS	
25	36	10	11643868800	48	45 - 49	60.1	153.0	86	25.67389	NORMAL	
26	37	10	11551420800	50	50 - 54	67.0	159.0	105	26.50212	NORMAL	
27	38	10	11043907200	66	65 - 69	67.0	144.0	79	32.31096	NORMAL	
28	39	10	10948089600	69	65 - 69	70.5	148.5	40	31.96953	OSTEOPOROSIS	
29	40	10	11051251200	66	65 - 69	66.5	147.0	48	30.77421	OSTEOPOROSIS	
30	41	10	11333692800	57	55 - 59	58.5	142.0	80	29.01210	NORMAL	

2. From univariate to Bivariate analysis

osteo

	registro	area	f_nac	edad	grupedad	peso	talla	bua	imc	clasific	me
1	3	10	11659420800	57	55 - 59	70.0	168.0	69	24.80159	OSTEOPENIA	
2	4	10	11671689600	46	45 - 49	53.0	152.0	73	22.93975	OSTEOPENIA	
3	10	10	11721024000	45	45 - 49	64.0	158.0	81	25.63692	NORMAL	
4	11	10	11464416000	53	50 - 54	78.0	161.0	58	30.09143	OSTEOPENIA	
5	12	10	11690784000	46	45 - 49	56.0	157.0	89	22.71897	NORMAL	
6	15	10	11716012800	45	45 - 49	63.5	170.0	76	21.97232	NORMAL	
7	16	10	11623737600	48	45 - 49	86.0	161.0	87	33.17773	NORMAL	
8	17	10	11562307200	50	50 - 54	61.5	164.0	74	22.86585	NORMAL	
9	18	10	11538028800	51	50 - 54	60.5	158.0	58	24.23490	OSTEOPENIA	
10	20	10	11332483200	57	55 - 59	64.0	149.0	61	28.82753	OSTEOPENIA	
11	21	10	11631945600	48	45 - 49	70.3	160.0	67	27.46094	OSTEOPENIA	
12	22	10	11425536000	55	55 - 59	74.4	160.0	68	29.06250	OSTEOPENIA	
13	23	10	11553235200	50	50 - 54	55.5	154.5	73	23.25070	OSTEOPENIA	
14	24	10	11367302400	56	55 - 59	89.0	166.0	61	32.29787	OSTEOPENIA	
15	25	10	11585635200	49	45 - 49	50.6	157.0	68	20.52822	OSTEOPENIA	
16	26	10	11572156800	50	50 - 54	71.4	152.0	74	30.90374	NORMAL	
17	27	10	11590992000	49	45 - 49	78.0	157.0	62	31.64429	OSTEOPENIA	
18	28	10	11293516800	58	55 - 59	72.0	162.0	65	27.43484	OSTEOPENIA	
19	29	10	11215238400	61	60 - 64	68.0	155.5	65	28.12212	OSTEOPENIA	
20	30	10	11405664000	55	55 - 59	75.0	161.0	92	28.93407	NORMAL	
21	31	10	11633155200	48	45 - 49	66.5	153.0	11	28.40788	OSTEOPOROSIS	
22	32	10	11287728000	59	55 - 59	101.0	156.0	82	41.50230	NORMAL	
23	34	10	10992758400	68	65 - 69	66.5	145.0	57	31.62901	OSTEOPENIA	
24	35	10	10909382400	69	65 - 69	70.0	168.0	48	24.80159	OSTEOPOROSIS	
25	36	10	11643868800	48	45 - 49	60.1	153.0	86	25.67389	NORMAL	
26	37	10	11551420800	50	50 - 54	67.0	159.0	105	26.50212	NORMAL	
27	38	10	11043907200	66	65 - 69	67.0	144.0	79	32.31096	NORMAL	
28	39	10	10948089600	69	65 - 69	70.5	148.5	40	31.96953	OSTEOPOROSIS	
29	40	10	11051251200	66	65 - 69	66.5	147.0	48	30.77421	OSTEOPOROSIS	
30	41	10	11333692800	57	55 - 59	58.5	142.0	80	29.01210	NORMAL	

Vall d'Hebron
Institut de Recerca

TABLE OF CONTENTS

Bivariate analysis

- Involves the analysis of **two** variables for the purpose of determining the empirical relationship between them.

⬇

easiest way is to measure how those two variables simultaneously change together

Bivariate analysis

- Involves the analysis of **two** variables for the purpose of determining the empirical relationship between them.

easiest way is to measure how those two variables simultaneously change together

- Major differentiating point between *univariate and bivariate* analysis (a part from the number of variables implicated) is that bivariate analysis goes beyond simply **descriptive**, since it study the **relationship** between the two variables.

Why bivariate analysis?

Let's begin by asking if:

People tend to marry other people of about the same age?

Our experience tells us "yes", but how good is the correspondence?

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

Sample of spousal ages of 10 White American Couples

Onlinestatbook (Authors: Rudy Guerra and David M. Lane)

Why bivariate analysis?

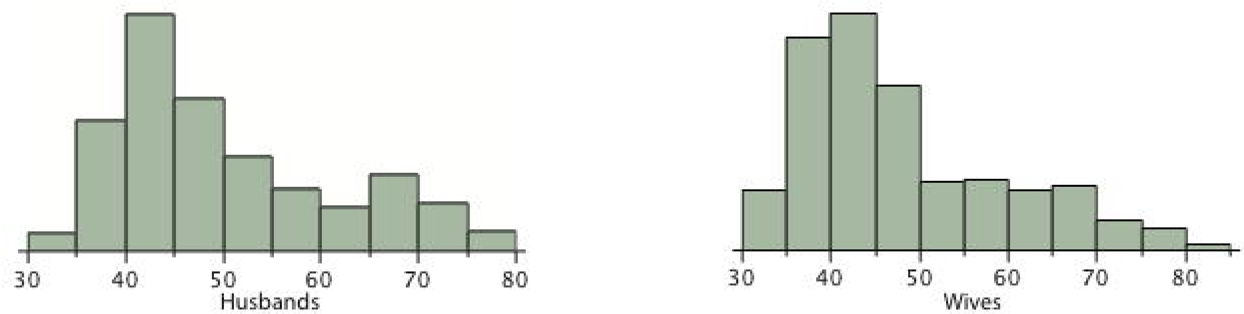


Figure 1. Histograms of spousal ages.

	Mean	Standard Deviation
Husbands	49	11
Wives	47	11

Onlinestatbook (Authors: Rudy Guerra and David M. Lane)

Why bivariate analysis?

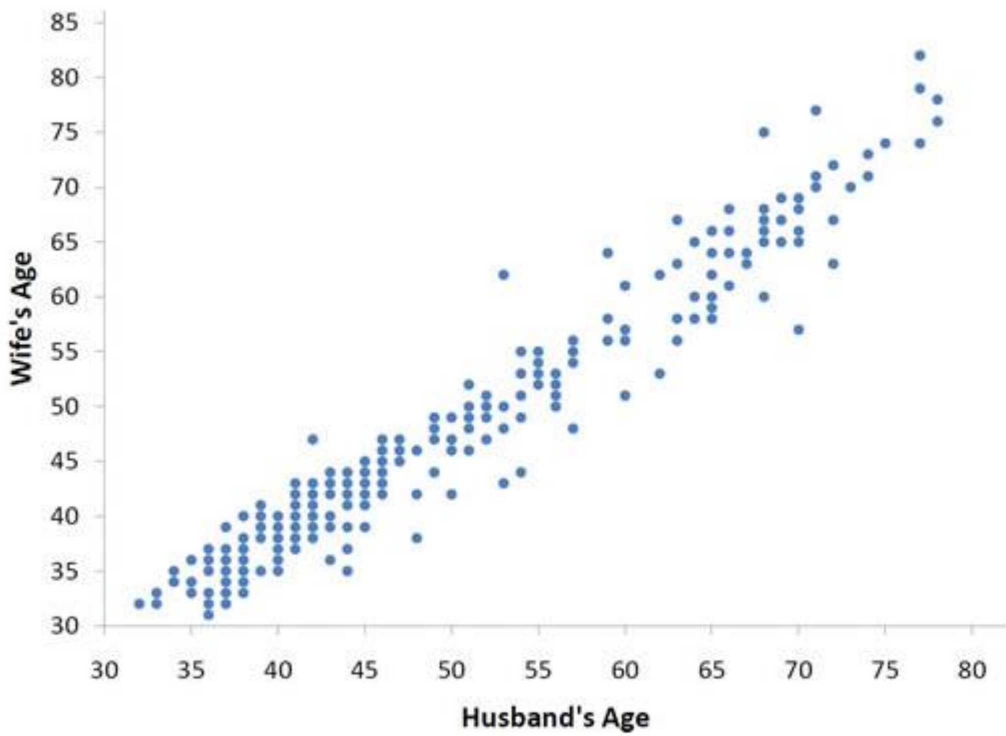


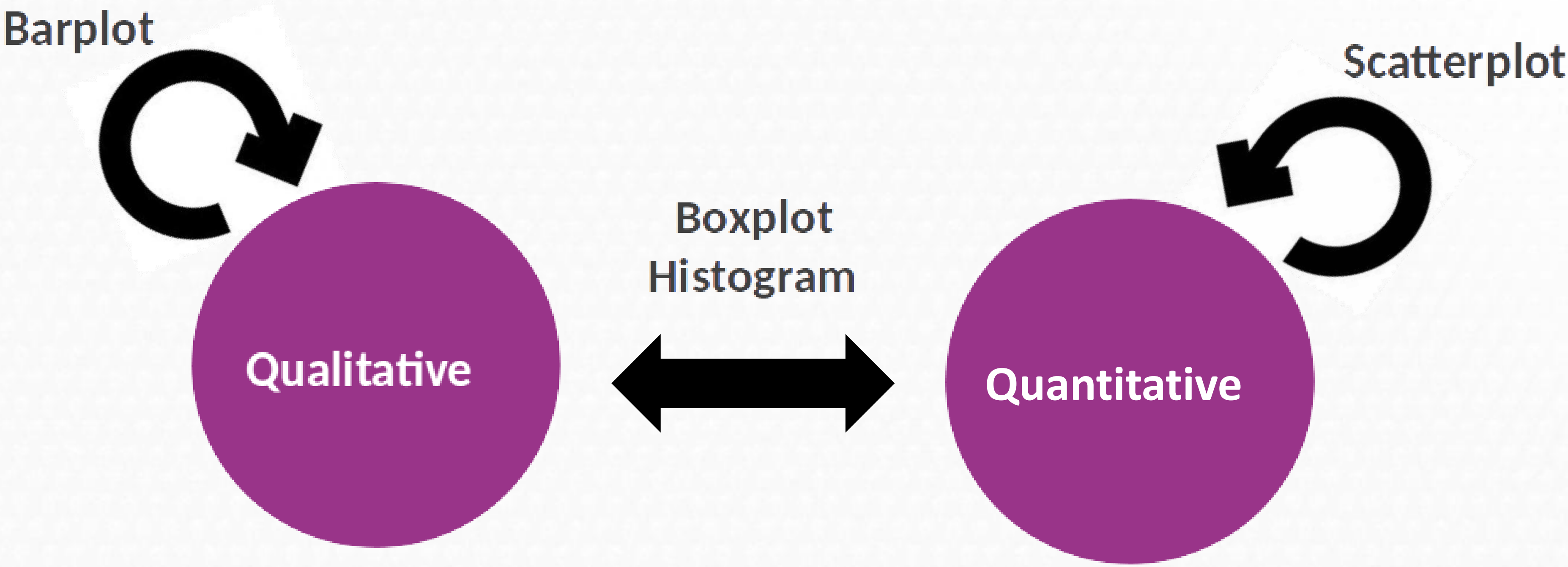
Figure 2. Scatter plot showing wife's age as a function of husband's age.

- The older the husband the older the wife.
- It is possible to know age of wives for an husband age.

Onlinestatbook (Authors: Rudy Guerra and David M. Lane)

3. Bivariate analysis

Some plots to study the relationship between two variables...



3. Bivariate analysis

3.1 Qualitative versus qualitative

osteo

	registro	area	f_nac	edad	grupedad	peso	talla	bua	imc	clasific	me
1	3	10	11659420800	57	55 - 59	70.0	168.0	69	24.80159	OSTEOPENIA	
2	4	10	11671689600	46	45 - 49	53.0	152.0	73	22.93975	OSTEOPENIA	
3	10	10	11721024000	45	45 - 49	64.0	158.0	81	25.63692	NORMAL	
4	11	10	11464416000	53	50 - 54	78.0	161.0	58	30.09143	OSTEOPENIA	
5	12	10	11690784000	46	45 - 49	56.0	157.0	89	22.71897	NORMAL	
6	15	10	11716012800	45	45 - 49	63.5	170.0	76	21.97232	NORMAL	
7	16	10	11623737600	48	45 - 49	86.0	161.0	87	33.17773	NORMAL	
8	17	10	11562307200	50	50 - 54	61.5	164.0	74	22.86585	NORMAL	
9	18	10	11538028800	51	50 - 54	60.5	158.0	58	24.23490	OSTEOPENIA	
10	20	10	11332483200	57	55 - 59	64.0	149.0	61	28.82753	OSTEOPENIA	
11	21	10	11631945600	48	45 - 49	70.3	160.0	67	27.46094	OSTEOPENIA	
12	22	10	11425536000	55	55 - 59	74.4	160.0	68	29.06250	OSTEOPENIA	
13	23	10	11553235200	50	50 - 54	55.5	154.5	73	23.25070	OSTEOPENIA	
14	24	10	11367302400	56	55 - 59	89.0	166.0	61	32.29787	OSTEOPENIA	
15	25	10	11585635200	49	45 - 49	50.6	157.0	68	20.52822	OSTEOPENIA	
16	26	10	11572156800	50	50 - 54	71.4	152.0	74	30.90374	NORMAL	
17	27	10	11590992000	49	45 - 49	78.0	157.0	62	31.64429	OSTEOPENIA	
18	28	10	11293516800	58	55 - 59	72.0	162.0	65	27.43484	OSTEOPENIA	
19	29	10	11215238400	61	60 - 64	68.0	155.5	65	28.12212	OSTEOPENIA	
20	30	10	11405664000	55	55 - 59	75.0	161.0	92	28.93407	NORMAL	
21	31	10	11633155200	48	45 - 49	66.5	153.0	11	28.40788	OSTEOPOROSIS	
22	32	10	11287728000	59	55 - 59	101.0	156.0	82	41.50230	NORMAL	
23	34	10	10992758400	68	65 - 69	66.5	145.0	57	31.62901	OSTEOPENIA	
24	35	10	10909382400	69	65 - 69	70.0	168.0	48	24.80159	OSTEOPOROSIS	
25	36	10	11643868800	48	45 - 49	60.1	153.0	86	25.67389	NORMAL	
26	37	10	11551420800	50	50 - 54	67.0	159.0	105	26.50212	NORMAL	
27	38	10	11043907200	66	65 - 69	67.0	144.0	79	32.31096	NORMAL	
28	39	10	10948089600	69	65 - 69	70.5	148.5	40	31.96953	OSTEOPOROSIS	
29	40	10	11051251200	66	65 - 69	66.5	147.0	48	30.77421	OSTEOPOROSIS	
30	41	10	11333692800	57	55 - 59	58.5	142.0	80	29.01210	NORMAL	

The way to study the relation will depend on the variable types:

- Two **qualitative** variables: contingency table

Used for organizing categorical variables and testing hypothesis with the chi-squared test for independence

3. Bivariate analysis

3.1 Qualitative versus qualitative

Vall d'Hebron Institut de Recerca

The way to study the relation will depend on the variable types:

- Two **qualitative** variables: contingency table

Used for organizing categorical variables and testing hypothesis with the chi-squared test for independence

- Count of individuals that simultaneously presents variable 1 (x) and variable 2 (y)

	y_1	y_1	$\cdots$	y_p	$n_{i.}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2p}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kp}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.p}$	N

Absolute

$$f_{ij} = \frac{n_{ij}}{N}$$

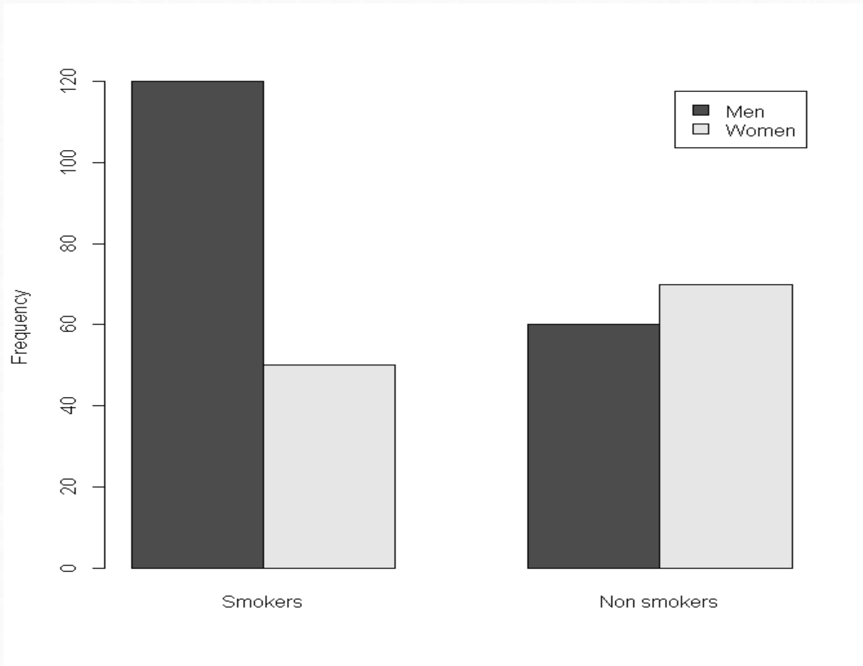
	y_1	y_1	$\cdots$	y_p	$f_{i.}$
x_1	f_{11}	f_{12}	$\cdots$	f_{1p}	$f_{1.}$
x_2	f_{21}	f_{22}	$\cdots$	f_{2p}	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	f_{k1}	f_{k2}	$\cdots$	f_{kp}	$f_{k.}$
$f_{.j}$	$f_{.1}$	$f_{.2}$	$\cdots$	$f_{.p}$	1

relative

An study wants to know if there are differences about smoking habits in men and women.

Gender	Smoking habits
1	1
2	1
1	0
1	0
1	0
1	1
2	1
…	…

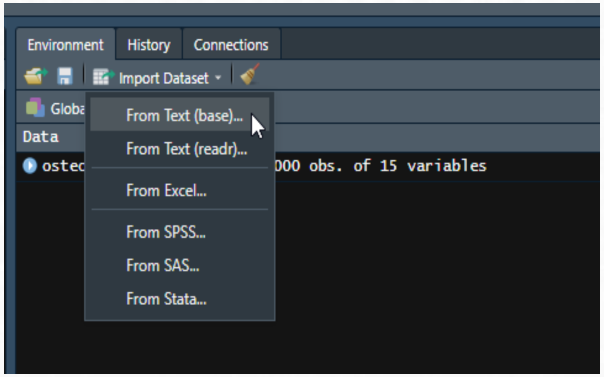
	Smokers	Non Smoking	Total
Men	120	60	180
Women	50	70	120
Total	170	130	300

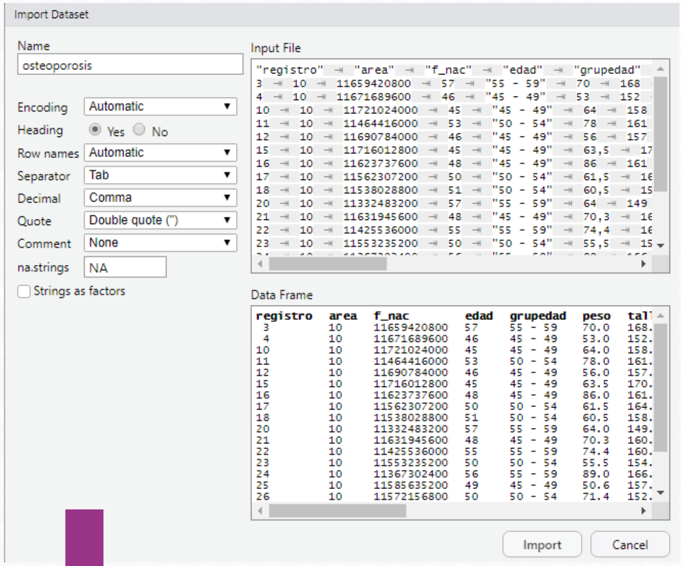


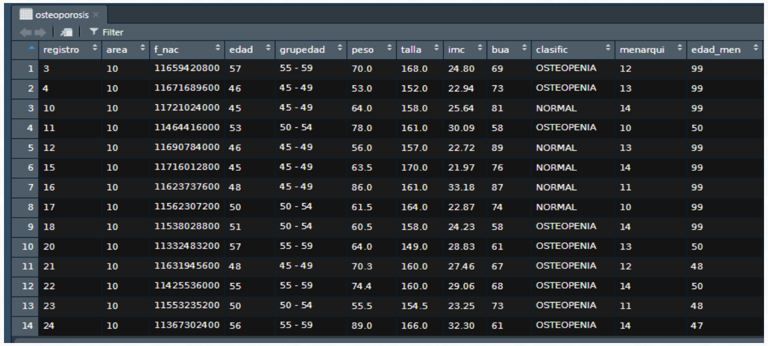
3. Bivariate analysis
3.1 Qualitative versus qualitative

Let´s do in R : Osteoporosis dataset (osteoporosis.txt)

Study if the group age (*grupedad*) of patients, influence in the illness type *(classific):*







	registro	area	f_nac	edad	grupedad	peso	talla	imc	bua	clasific	menarqui	edad_men
1	3	10	11659420800	57	55 - 59	70.0	168.0	24.80	69	OSTEOPENIA	12	99
2	4	10	11671689600	46	45 - 49	53.0	152.0	22.94	73	OSTEOPENIA	13	99
3	10	10	11721024000	45	45 - 49	64.0	158.0	25.64	81	NORMAL	14	99
4	11	10	11464416000	53	50 - 54	78.0	161.0	30.09	58	OSTEOPENIA	10	50
5	12	10	11690784000	46	45 - 49	56.0	157.0	22.72	89	NORMAL	13	99
6	15	10	11716012800	45	45 - 49	63.5	170.0	21.97	76	NORMAL	14	99
7	16	10	11623737600	48	45 - 49	86.0	161.0	33.18	87	NORMAL	11	99
8	17	10	11562307200	50	50 - 54	61.5	164.0	22.87	74	NORMAL	10	99
9	18	10	11538028800	51	50 - 54	60.5	158.0	24.23	58	OSTEOPENIA	14	99
10	20	10	11332483200	57	55 - 59	64.0	149.0	28.83	61	OSTEOPENIA	13	50
11	21	10	11631945600	48	45 - 49	70.3	160.0	27.46	67	OSTEOPENIA	12	48
12	22	10	11425536000	55	55 - 59	74.4	160.0	29.06	68	OSTEOPENIA	14	50
13	23	10	11553235200	50	50 - 54	55.5	154.5	23.25	73	OSTEOPENIA	11	48
14	24	10	11367302400	56	55 - 59	89.0	166.0	32.30	61	OSTEOPENIA	14	47

3. Bivariate analysis

3.1 Qualitative versus qualitative

```
table(osteoporosis$grupedad, osteoporosis$clasific)
```

```
          NORMAL OSTEOPENIA OSTEOPOROSIS
  45 - 49    233        138            7
  50 - 54    113        113            7
  55 - 59     67        100            9
  60 - 64     38         74           17
  65 - 69     18         42           24
```

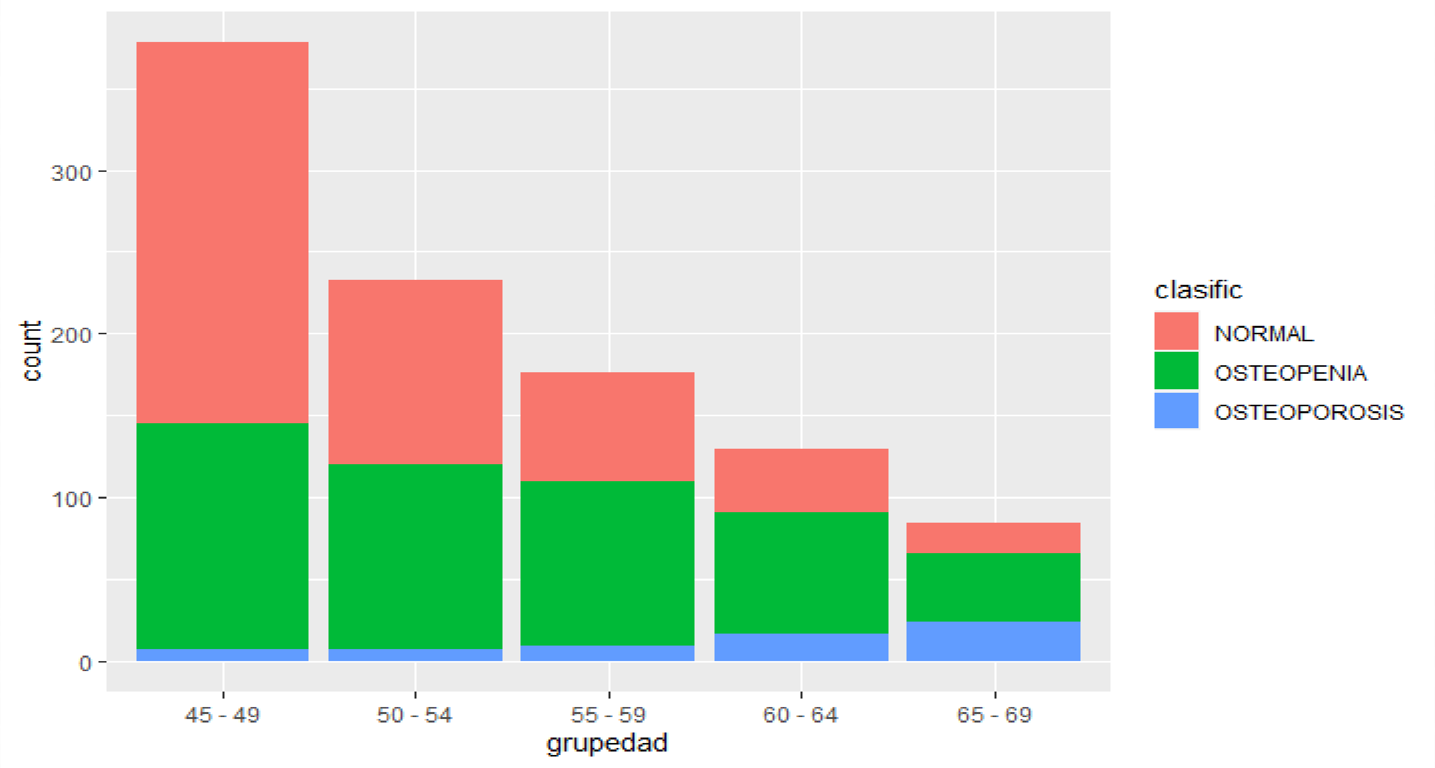
```
prop.table(table(osteoporosis$grupedad, osteoporosis$clasific))
```

```
          NORMAL OSTEOPENIA OSTEOPOROSIS
  45 - 49  0.233      0.138        0.007
  50 - 54  0.113      0.113        0.007
  55 - 59  0.067      0.100        0.009
  60 - 64  0.038      0.074        0.017
  65 - 69  0.018      0.042        0.024
```

3. Bivariate analysis
3.1 Qualitative versus qualitative

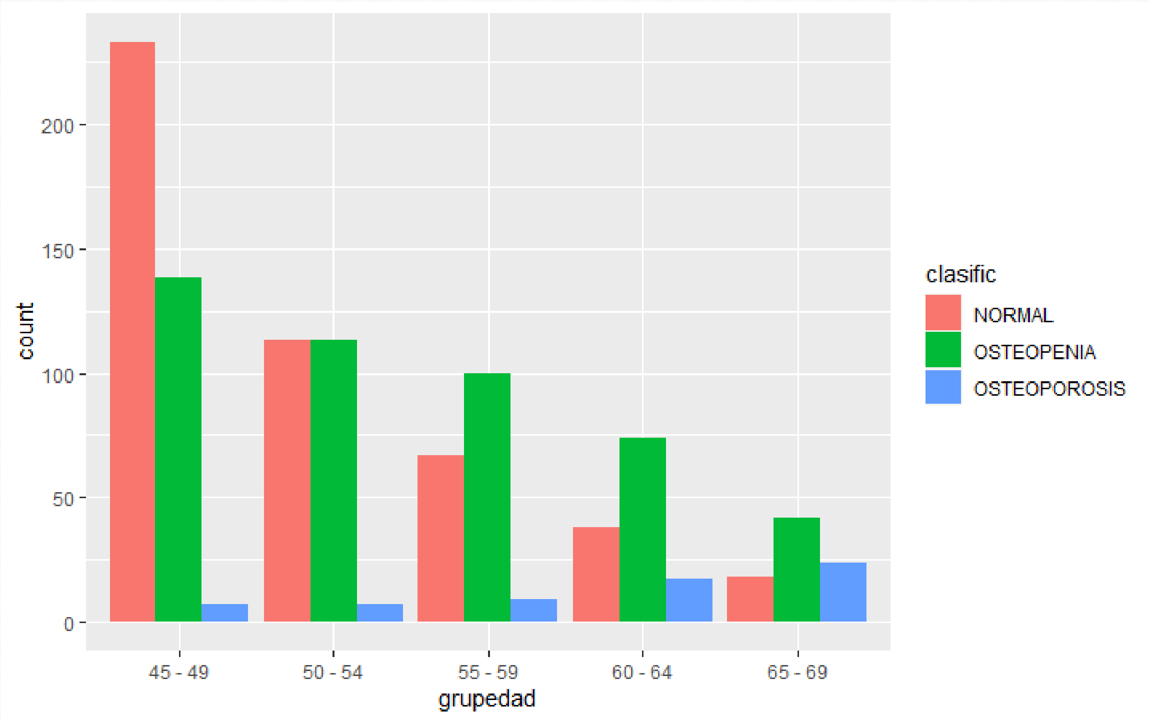
Barplot with R

```
ggplot(data = osteoporosis, aes(x = grupedad)) +
  geom_bar(aes(fill = clasific))
```



3. Bivariate analysis
3.1 Qualitative versus qualitative

```
ggplot(data = osteoporosis, aes(x = grupedad)) +
  geom_bar(aes(fill = clasific), position = "dodge")
```



Improving barplot

http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization

Change colors, legend position, labels and finally save it!

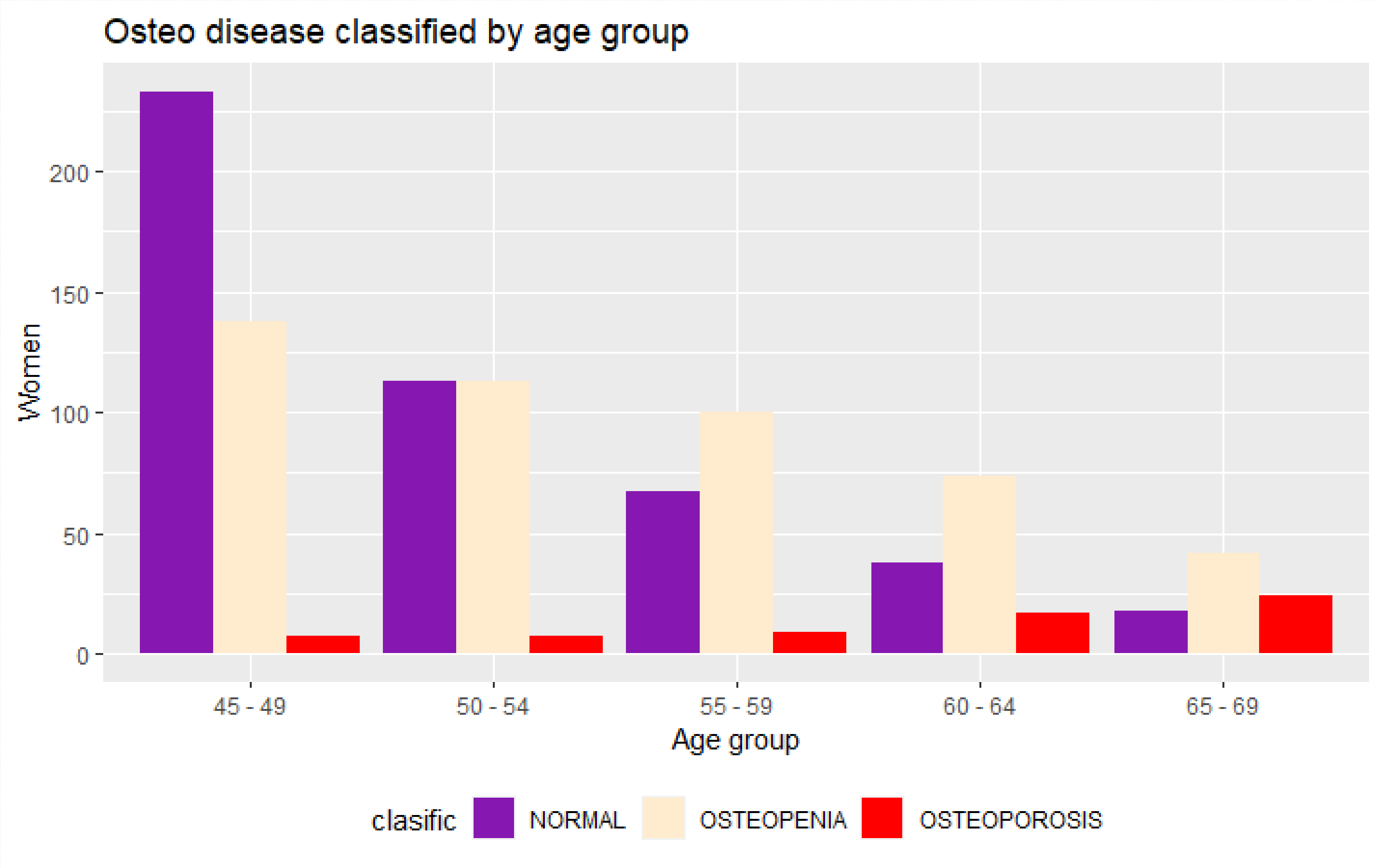
```
p + scale_fill_manual(values=c("#8618b1", "blanchedalmond", "red"))

p + theme(legend.position="bottom")

p + labs(x = "Age group", y = "Women", title = "Osteo disease classified by age group")

pdf("clasific_grupedad.pdf")
  p + labs(x = "Age group", y = "Women", title = "Osteo disease classified by age group")
dev.off()
```

Improving barplot



3. Bivariate analysis

3.1 Qualitative versus qualitative

Another way to introduce the data:

	Smokers	Non Smoking	Total
Men	120	60	180
Women	50	70	120
Total	170	130	300

```
tab <- matrix(data = c(120, 60, 50, 70), nrow = 2, ncol = 2, byrow = TRUE)
tab
```

```
     [,1] [,2]
[1,]  120   60
[2,]   50   70
```

```
colnames(tab) <- c("Smokers", "Nonsmokers")
rownames(tab) <- c("Men", "Women")
tab
```

```
      Smokers Nonsmokers
Men       120         60
Women      50         70
```

Another way to introduce the data:

	Smokers	Non Smoking	Total
Men	120	60	180
Women	50	70	120
Total	170	130	300

```
prop.table(tab)
```

```
        Smokers Nonsmokers
Men   0.4000000  0.2000000
Women 0.1666667  0.2333333
```

3. Bivariate analysis

3.2 Qualitative versus quantitative

Vall d'Hebron Institut de Recerca

osteo

	registro	area	f_nac	edad	grupedad	peso	talla	bua	imc	clasific	me
1	3	10	11659420800	57	55 - 59	70.0	168.0	69	24.80159	OSTEOPENIA	
2	4	10	11671689600	46	45 - 49	53.0	152.0	73	22.93975	OSTEOPENIA	
3	10	10	11721024000	45	45 - 49	64.0	158.0	81	25.63692	NORMAL	
4	11	10	11464416000	53	50 - 54	78.0	161.0	58	30.09143	OSTEOPENIA	
5	12	10	11690784000	46	45 - 49	56.0	157.0	89	22.71897	NORMAL	
6	15	10	11716012800	45	45 - 49	63.5	170.0	76	21.97232	NORMAL	
7	16	10	11623737600	48	45 - 49	86.0	161.0	87	33.17773	NORMAL	
8	17	10	11562307200	50	50 - 54	61.5	164.0	74	22.86585	NORMAL	
9	18	10	11538028800	51	50 - 54	60.5	158.0	58	24.23490	OSTEOPENIA	
10	20	10	11332483200	57	55 - 59	64.0	149.0	61	28.82753	OSTEOPENIA	
11	21	10	11631945600	48	45 - 49	70.3	160.0	67	27.46094	OSTEOPENIA	
12	22	10	11425536000	55	55 - 59	74.4	160.0	68	29.06250	OSTEOPENIA	
13	23	10	11553235200	50	50 - 54	55.5	154.5	73	23.25070	OSTEOPENIA	
14	24	10	11367302400	56	55 - 59	89.0	166.0	61	32.29787	OSTEOPENIA	
15	25	10	11585635200	49	45 - 49	50.6	157.0	68	20.52822	OSTEOPENIA	
16	26	10	11572156800	50	50 - 54	71.4	152.0	74	30.90374	NORMAL	
17	27	10	11590992000	49	45 - 49	78.0	157.0	62	31.64429	OSTEOPENIA	
18	28	10	11293516800	58	55 - 59	72.0	162.0	65	27.43484	OSTEOPENIA	
19	29	10	11215238400	61	60 - 64	68.0	155.5	65	28.12212	OSTEOPENIA	
20	30	10	11405664000	55	55 - 59	75.0	161.0	92	28.93407	NORMAL	
21	31	10	11633155200	48	45 - 49	66.5	153.0	11	28.40788	OSTEOPOROSIS	
22	32	10	11287728000	59	55 - 59	101.0	156.0	82	41.50230	NORMAL	
23	34	10	10992758400	68	65 - 69	66.5	145.0	57	31.62901	OSTEOPENIA	
24	35	10	10909382400	69	65 - 69	70.0	168.0	48	24.80159	OSTEOPOROSIS	
25	36	10	11643868800	48	45 - 49	60.1	153.0	86	25.67389	NORMAL	
26	37	10	11551420800	50	50 - 54	67.0	159.0	105	26.50212	NORMAL	
27	38	10	11043907200	66	65 - 69	67.0	144.0	79	32.31096	NORMAL	
28	39	10	10948089600	69	65 - 69	70.5	148.5	40	31.96953	OSTEOPOROSIS	
29	40	10	11051251200	66	65 - 69	66.5	147.0	48	30.77421	OSTEOPOROSIS	
30	41	10	11333692800	57	55 - 59	58.5	142.0	80	29.01210	NORMAL	

3.2 Qualitative versus quantitative

The way to study the relation will depend on the variable types:

- One **qualitative** variable and one **quantitative** variable: Table of statistics

Mean value of the variable in each category for each individual

Let´s do it in R:

Osteoporosis dataset

Study if bone density (*bua)* changes depending the age group

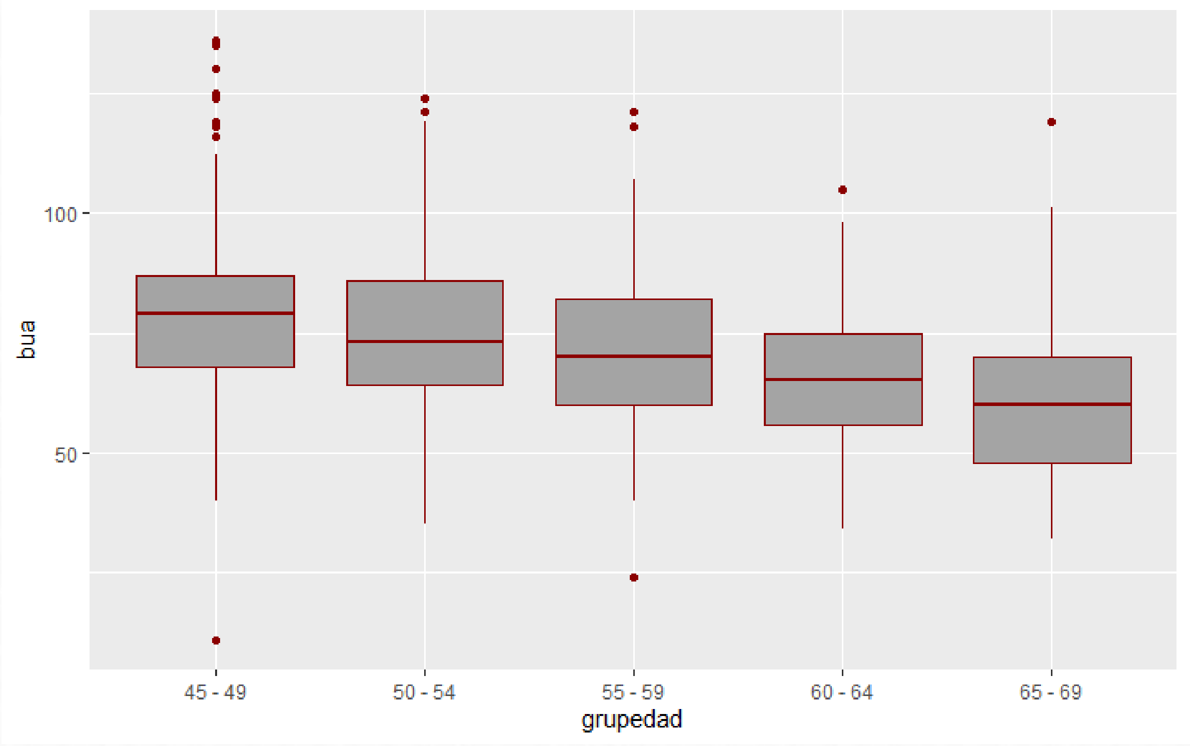
```
library(dplyr)
osteoporosis %>%
  group_by(grupedad) %>%
  summarize(mean(edad))
```

```
  45-49    50-54    55-59    60-64    65-69
78.75926 75.05150 71.43182 64.89147 60.66667
```

3.2 Qualitative versus quantitative

Study if bone density (*bua)* is related with the age group

```
bp <- ggplot(osteoporosis, aes(x = grupedad, y = bua)) +
  geom_boxplot(fill='#A4A4A4', color="darkred")
bp
```



Study if bone density (*bua)* is related with the age group

```
# Box plot with points
# 0.2 : degree of jitter in x direction
bp + geom_jitter(shape = 16, position = position_jitter(0.2)) +
          labs(x = "Age Group", y = "Women", title = "Osteo disease classified by age group")
```

3. Bivariate analysis

Exercise

Study the relationship between *menop* and group of illness (*classific*)

Study the relationship between *peso* iand group of illness (*classific*).

Vall d'Hebron Institut de Recerca

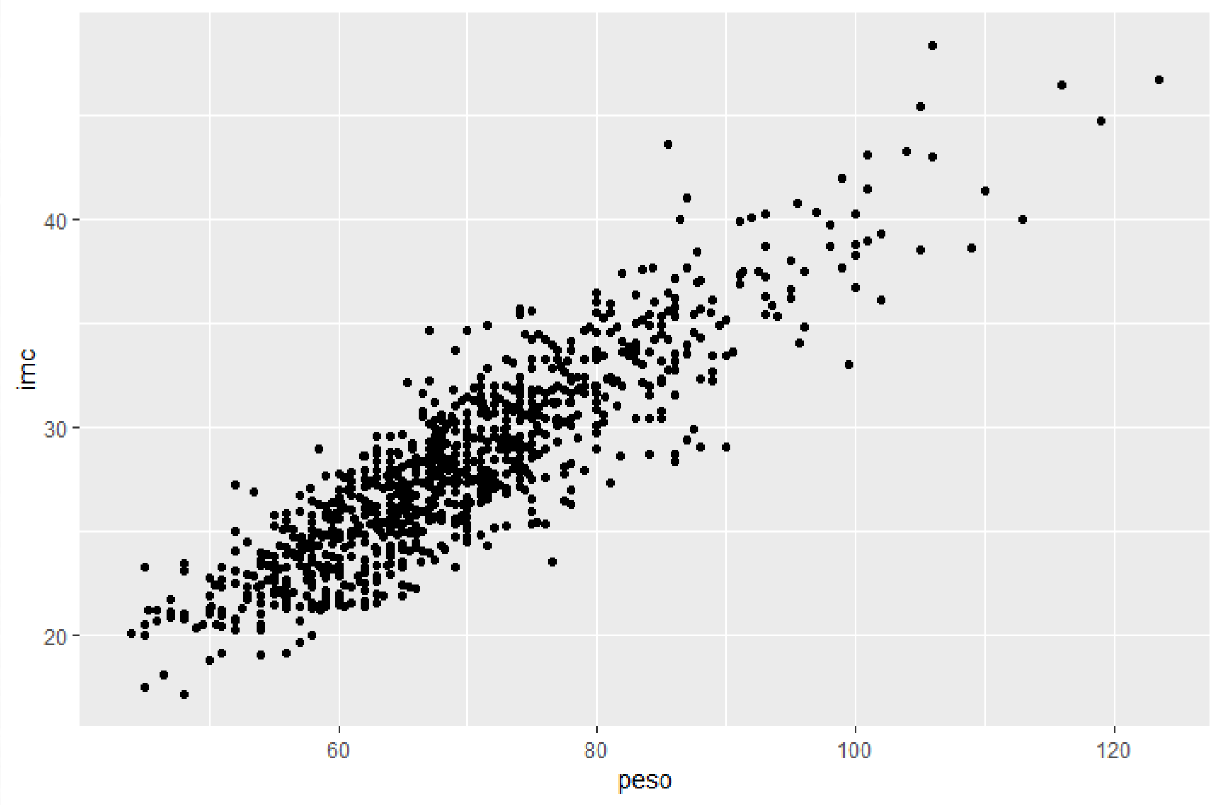
The way to study the relation will depend on the variable types:

- Two **quantitatives** variables:

osteo

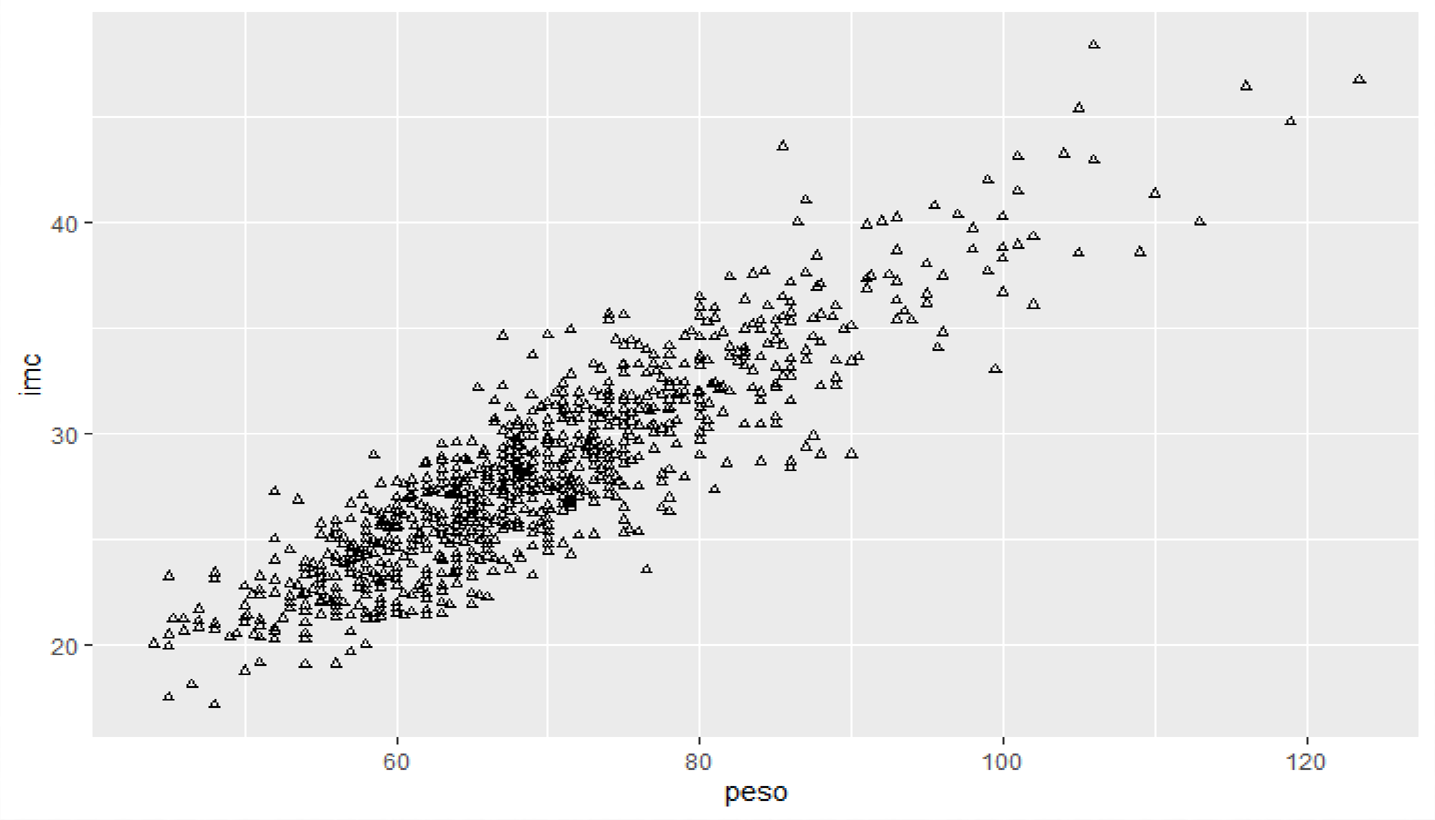
	registro	area	f_nac	edad	grupedad	peso	talla	bua	imc	clasific	me
1	3	10	11659420800	57	55 - 59	70.0	168.0	69	24.80159	OSTEOPENIA	
2	4	10	11671689600	46	45 - 49	53.0	152.0	73	22.93975	OSTEOPENIA	
3	10	10	11721024000	45	45 - 49	64.0	158.0	81	25.63692	NORMAL	
4	11	10	11464416000	53	50 - 54	78.0	161.0	58	30.09143	OSTEOPENIA	
5	12	10	11690784000	46	45 - 49	56.0	157.0	89	22.71897	NORMAL	
6	15	10	11716012800	45	45 - 49	63.5	170.0	76	21.97232	NORMAL	
7	16	10	11623737600	48	45 - 49	86.0	161.0	87	33.17773	NORMAL	
8	17	10	11562307200	50	50 - 54	61.5	164.0	74	22.86585	NORMAL	
9	18	10	11538028800	51	50 - 54	60.5	158.0	58	24.23490	OSTEOPENIA	
10	20	10	11332483200	57	55 - 59	64.0	149.0	61	28.82753	OSTEOPENIA	
11	21	10	11631945600	48	45 - 49	70.3	160.0	67	27.46094	OSTEOPENIA	
12	22	10	11425536000	55	55 - 59	74.4	160.0	68	29.06250	OSTEOPENIA	
13	23	10	11553235200	50	50 - 54	55.5	154.5	73	23.25070	OSTEOPENIA	
14	24	10	11367302400	56	55 - 59	89.0	166.0	61	32.29787	OSTEOPENIA	
15	25	10	11585635200	49	45 - 49	50.6	157.0	68	20.52822	OSTEOPENIA	
16	26	10	11572156800	50	50 - 54	71.4	152.0	74	30.90374	NORMAL	
17	27	10	11590992000	49	45 - 49	78.0	157.0	62	31.64429	OSTEOPENIA	
18	28	10	11293516800	58	55 - 59	72.0	162.0	65	27.43484	OSTEOPENIA	
19	29	10	11215238400	61	60 - 64	68.0	155.5	65	28.12212	OSTEOPENIA	
20	30	10	11405664000	55	55 - 59	75.0	161.0	92	28.93407	NORMAL	
21	31	10	11633155200	48	45 - 49	66.5	153.0	11	28.40788	OSTEOPOROSIS	
22	32	10	11287728000	59	55 - 59	101.0	156.0	82	41.50230	NORMAL	
23	34	10	10992758400	68	65 - 69	66.5	145.0	57	31.62901	OSTEOPENIA	
24	35	10	10909382400	69	65 - 69	70.0	168.0	48	24.80159	OSTEOPOROSIS	
25	36	10	11643868800	48	45 - 49	60.1	153.0	86	25.67389	NORMAL	
26	37	10	11551420800	50	50 - 54	67.0	159.0	105	26.50212	NORMAL	
27	38	10	11043907200	66	65 - 69	67.0	144.0	79	32.31096	NORMAL	
28	39	10	10948089600	69	65 - 69	70.5	148.5	40	31.96953	OSTEOPOROSIS	
29	40	10	11051251200	66	65 - 69	66.5	147.0	48	30.77421	OSTEOPOROSIS	
30	41	10	11333692800	57	55 - 59	58.5	142.0	80	29.01210	NORMAL	

```
# Basic scatter plot
ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point()
```

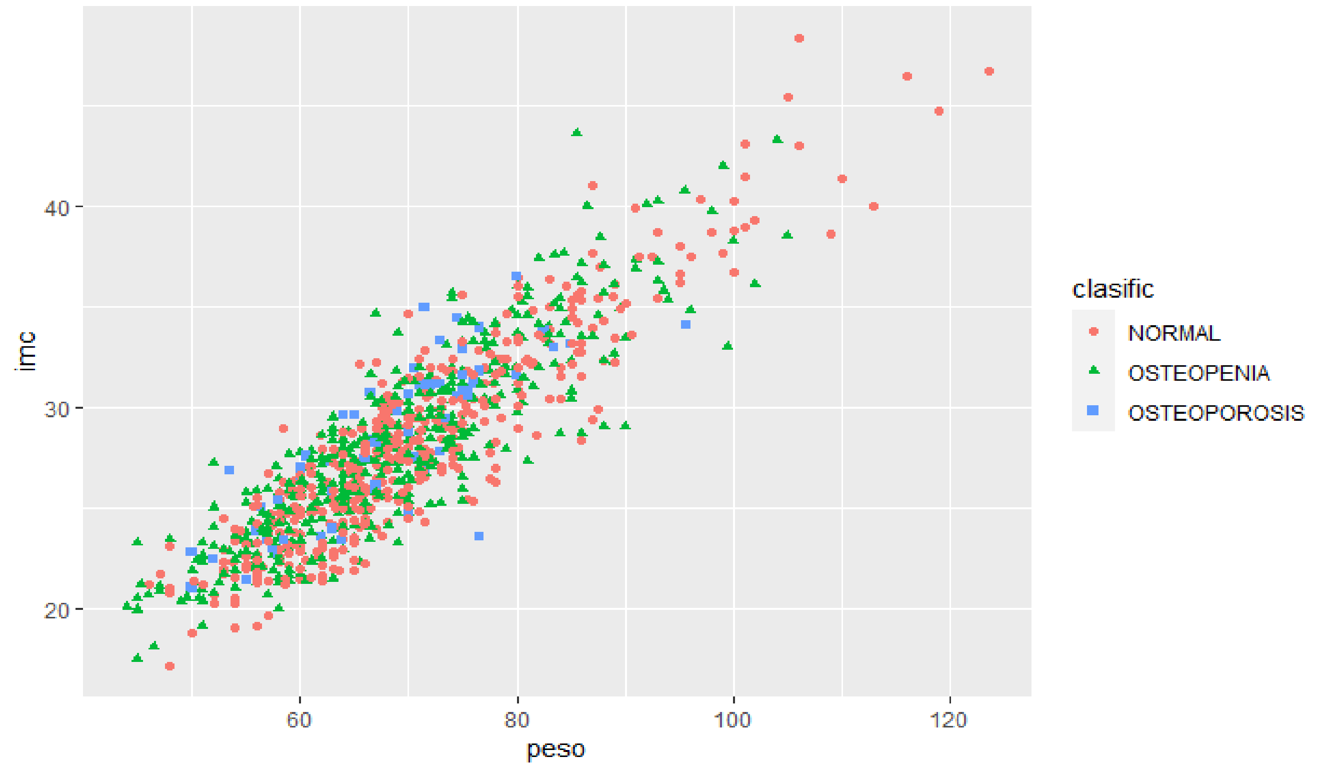


http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization

```
# Change the point size, and shape
ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point(size = 1, shape = 1)
```



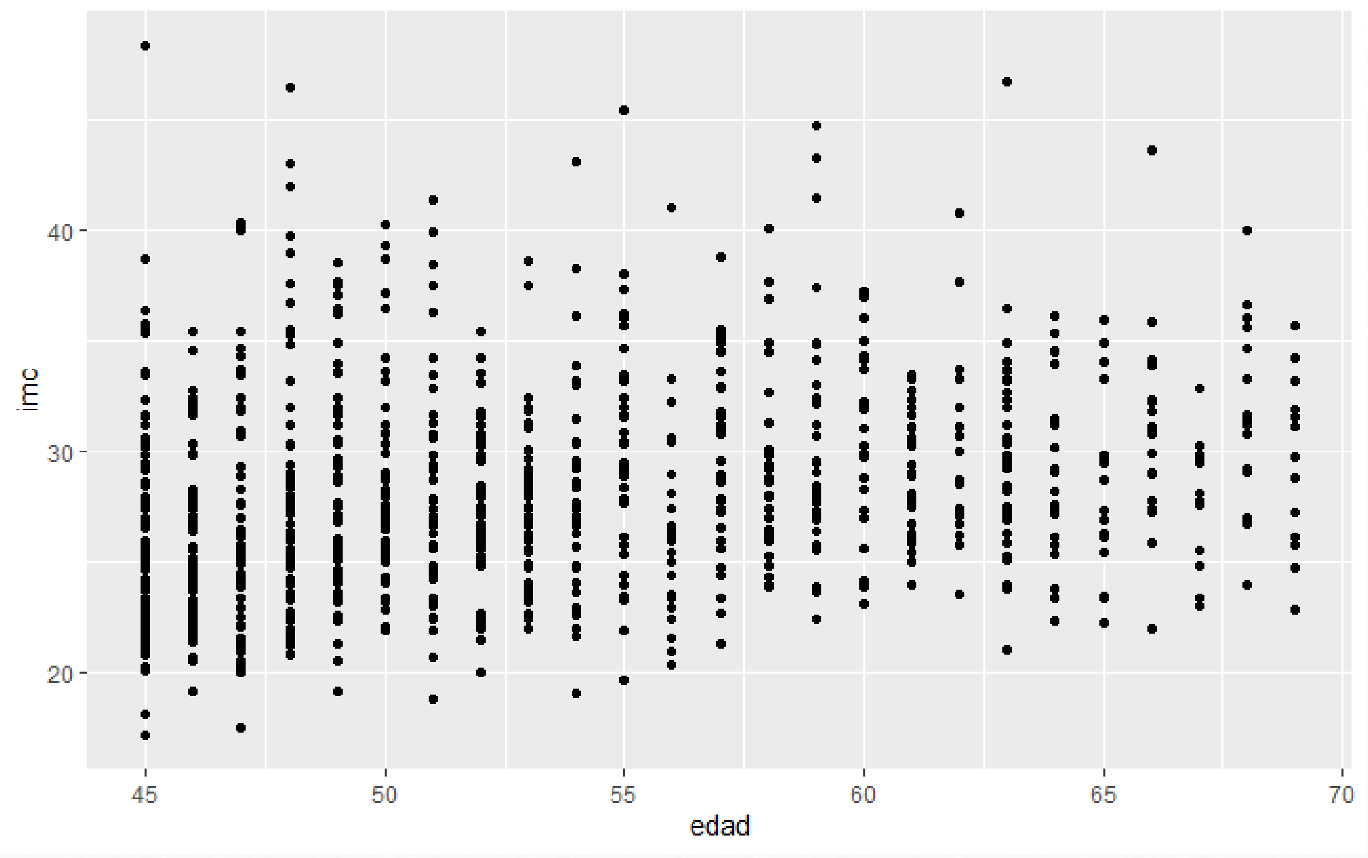
```
# Color the points depending of another variable
ggplot(osteoporosis, aes(x = peso, y = imc, color = clasific, shape = clasific)) +
  geom_point()
```



Vall d'Hebron Institut de Recerca

But not always the correlation is good!

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

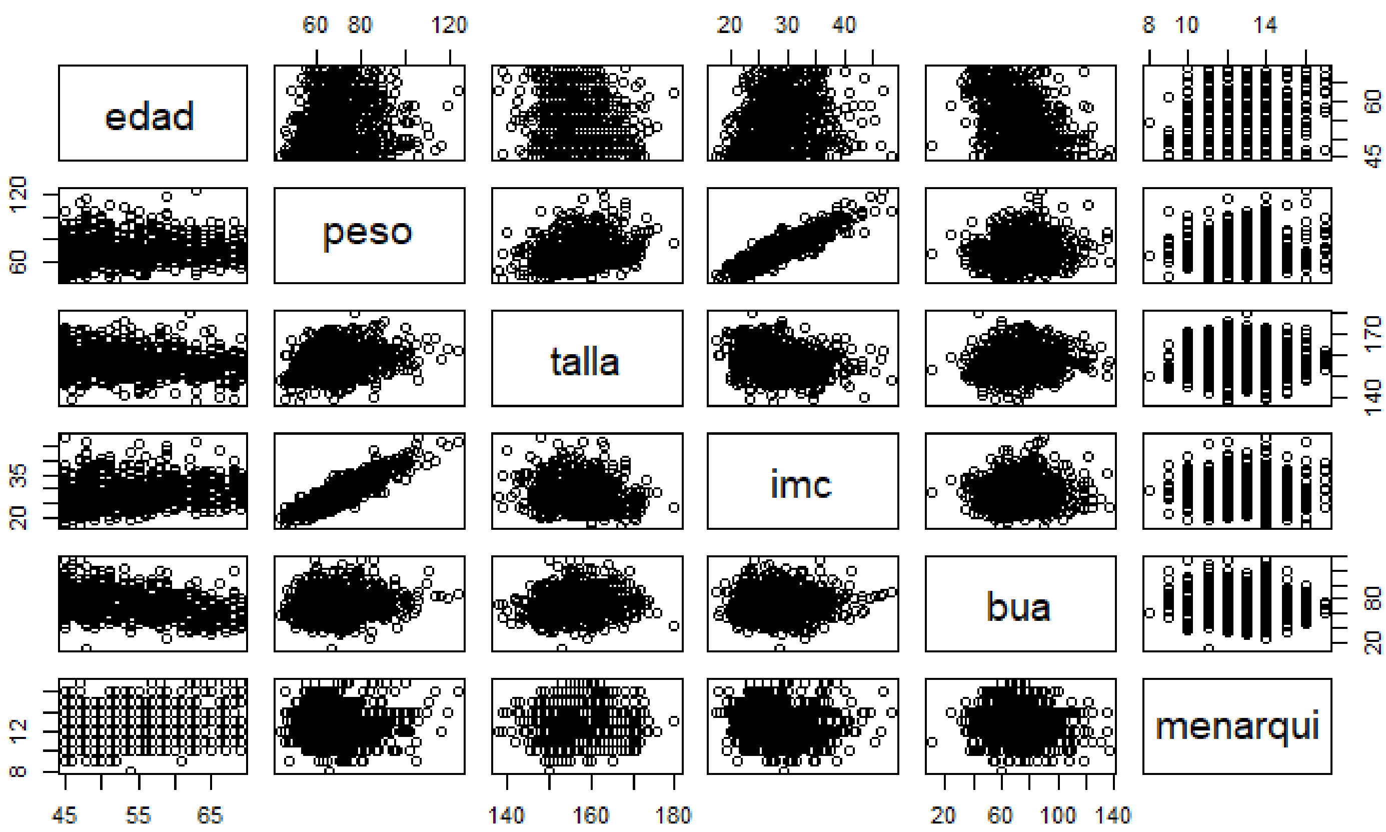


Whatever you compute Always look at the data!

3. Bivariate analysis

3.3 Quantitative versus Quantitative

```
pairs(osteoporosis[, c("edad", "peso",  "talla", "imc", "bua", "menarqui")])
```



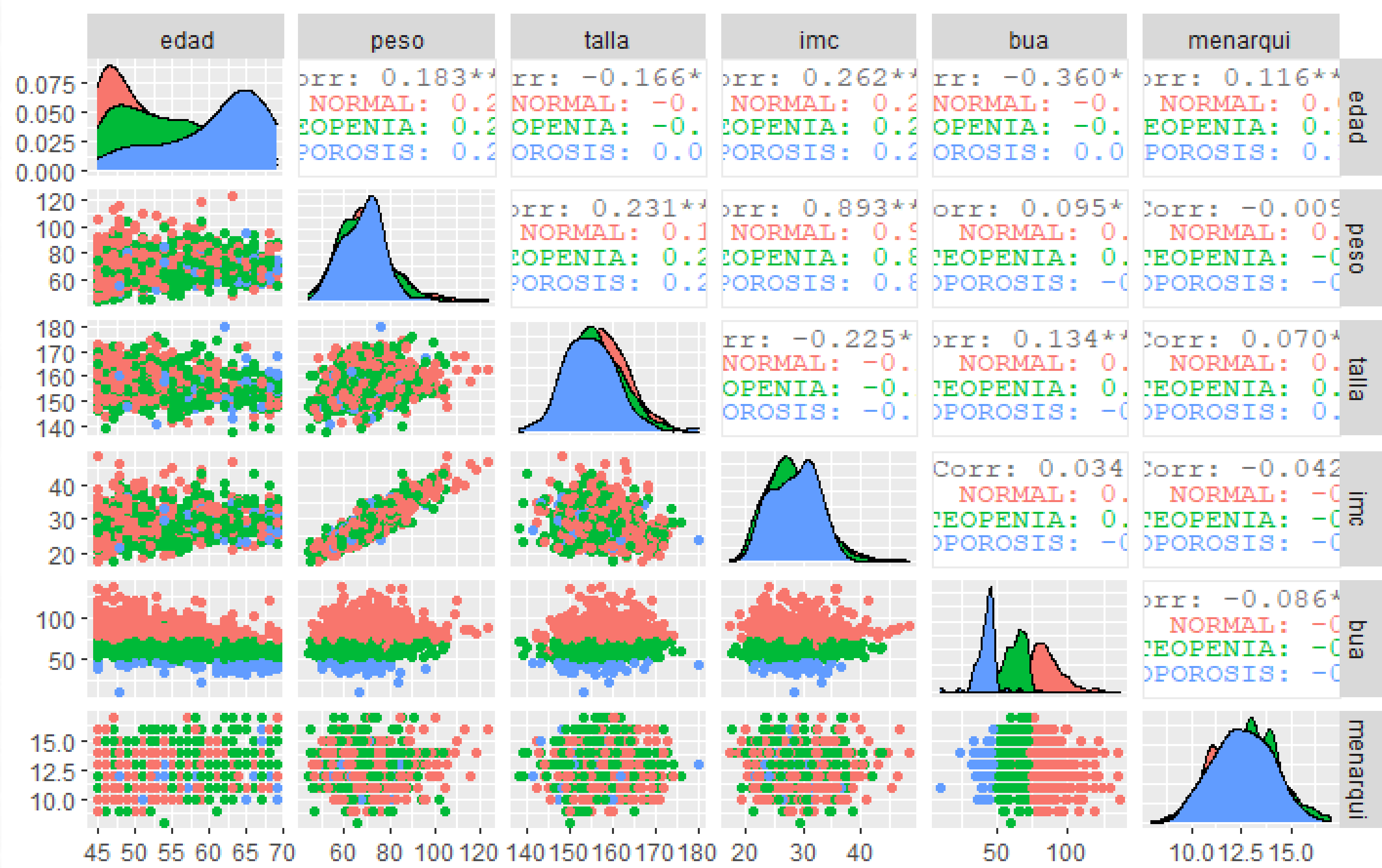
```
library(GGally)

ggpairs(osteoporosis, columns = c("edad", "peso",  "talla", "imc", "bua", "menarqui"),
ggplot2::aes(colour = clasific))
```

3. Bivariate analysis
3.3 Quantitative versus Quantitative

```
library(GGally)

ggpairs(osteoporosis, columns = c("edad", "peso",  "talla", "imc", "bua", "menarqui"),
ggplot2::aes(colour = clasific))
```



Vall d'Hebron
Institut de Recerca

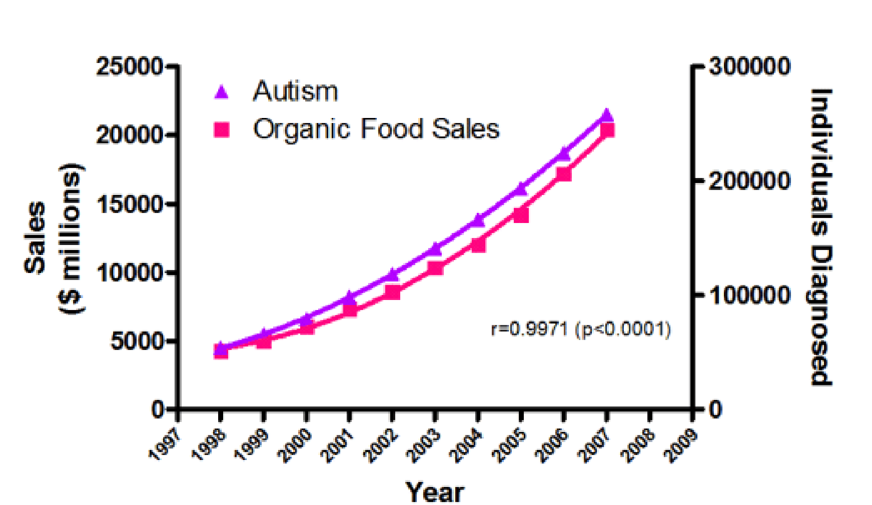
TABLE OF CONTENTS

Main characteristics of correlation analysis:

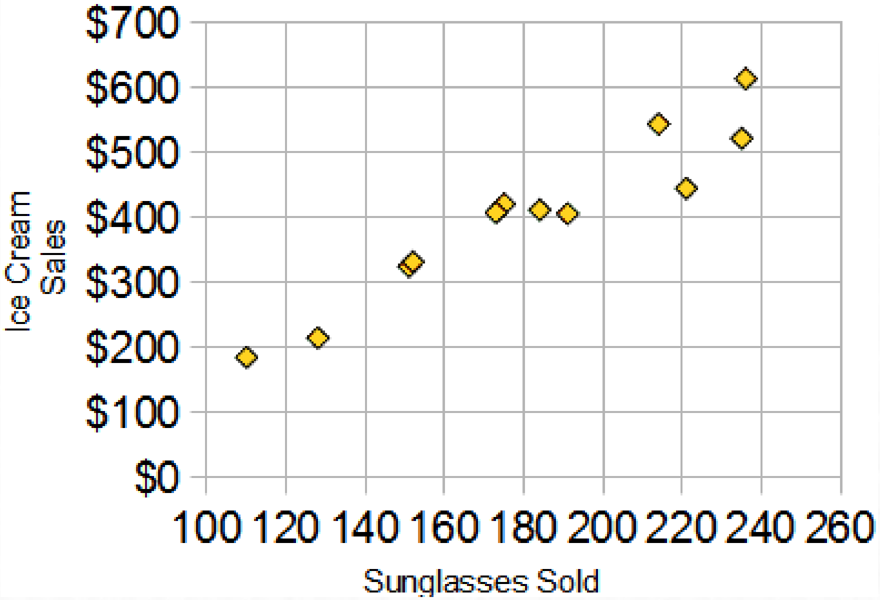
- Correlation analysis allows
  - To study the **way** of relation between the two variables
  - **To quantify** the intensity of relation
- Correlation is **not causation** ➡ one thing does not causes the other
- In the correlation analysis, the two variables have the **same weight**
- The **correlation coefficient** measures the strength of a **linear** relation

Main characteristics of correlation analysis:

Correlation is not causation



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

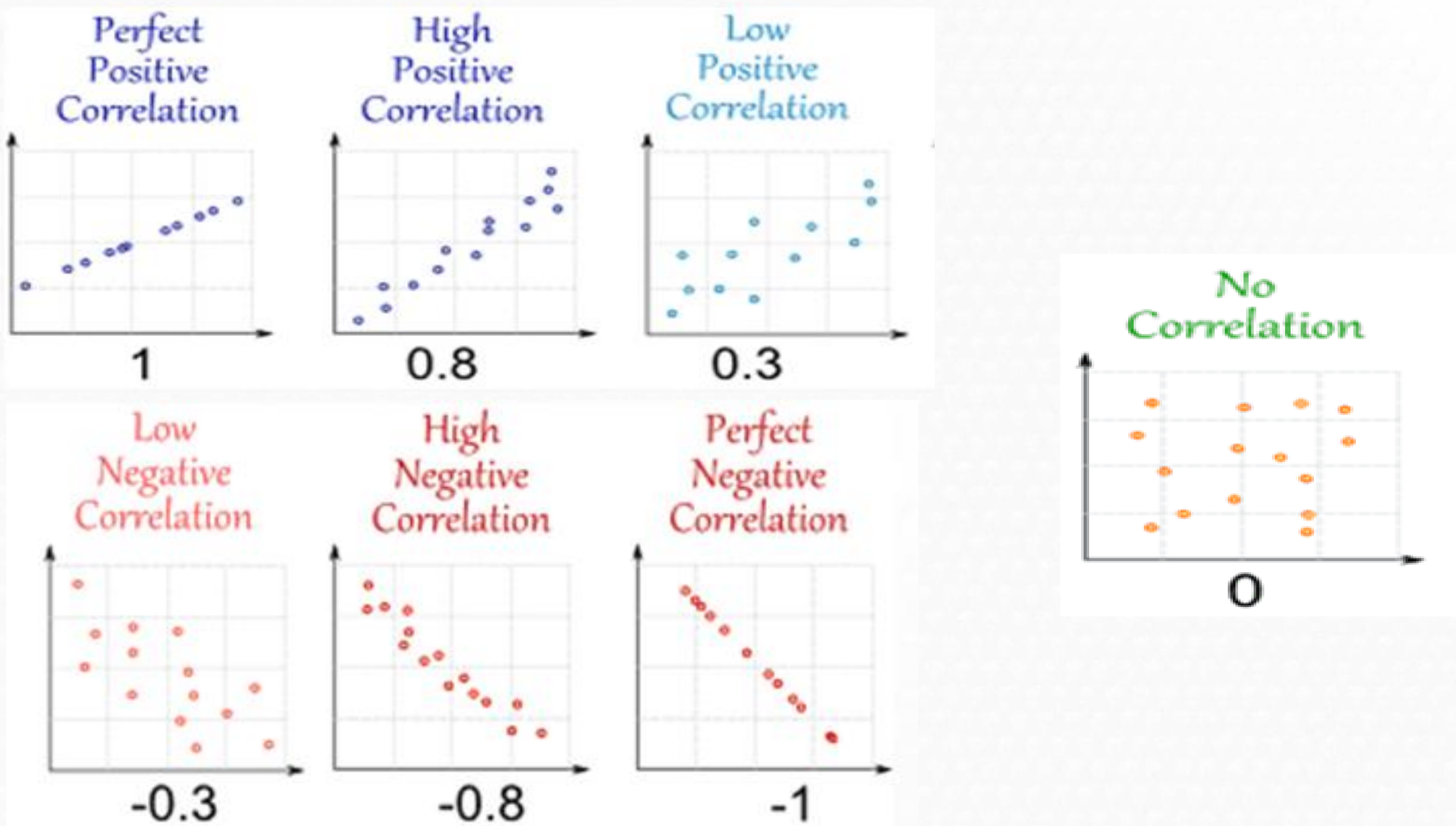


Pearson correlation coefficient

- Measures linear correlation between two variables
- It is represented by letter **r**. It has no dimensions (no units)
- Values go from **-1** to **+1**
  - **r=0** indicates no linear relation between the variables
  - **r>0** indicates direct relation between the variables
  - **r<0** indicates indirect relation between the variables
  - **r=1/-1** indicates a perfect relation between the variables

4. Correlation
2. Types of correlation

Pearson correlation coefficient. Examples

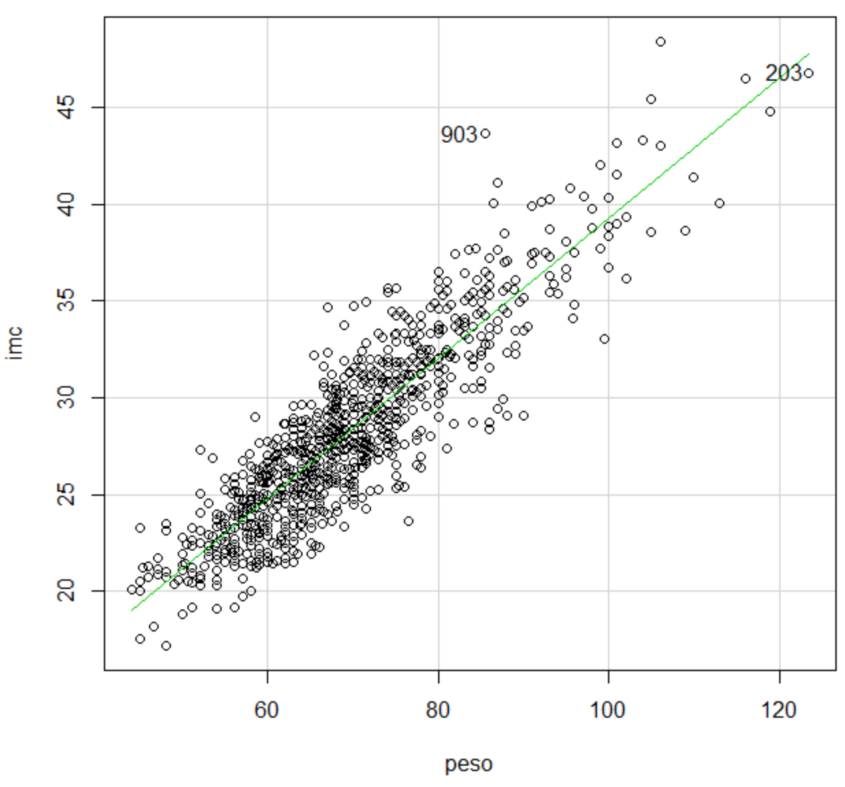


4. Correlation

2. Types of correlation

Study the relationship between *peso* and *body mass index (imc)*:

```
          imc      peso
imc  1.0000000 0.8927967
peso 0.8927967 1.0000000
```



Pearson correlation coefficient. How to in R?

Bone density and **age** are correlated?

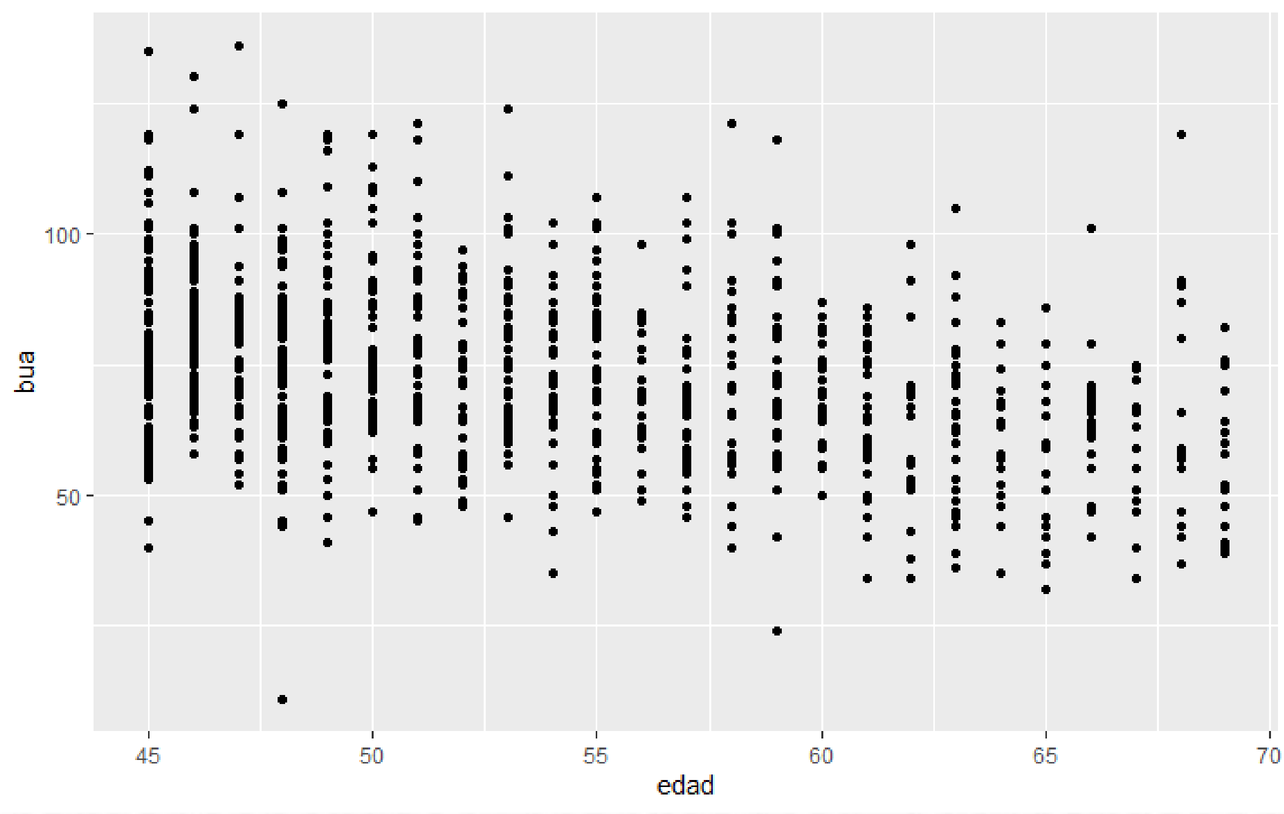
```
cor(osteoporosis$bua, osteoporosis$edad, method = "pearson")

[1] -0.3601883
```

Don’t forget to look the graphic!!

Pearson correlation coefficient. How to in R-commander?

```
ggplot(osteoporosis, aes(x = edad, y = bua)) +
  geom_point()
```



Pearson correlation coefficient. How to in R?

Exercise 1. Do you think that exists a relationship between *peso* and *talla*? What type of relationship? Show a scatterplot of the values.

Non Parametric correlation: Spearman correlation coefficient

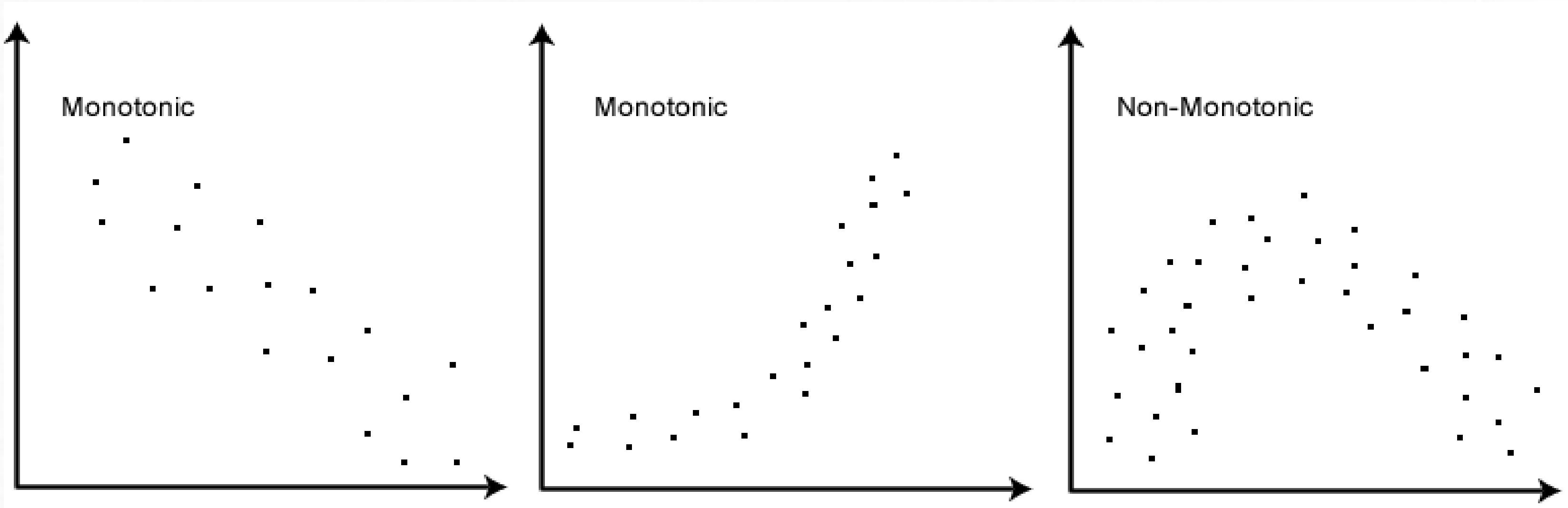
- Pearson correlation coefficient is severely affected by **outliers** and if the relation **is not linear**

Better to use **Spearman** correlation coefficient (use the ranks between the numbers instead the values) to calculate the correlation coefficient

- Evaluates the **monotonic** relationship between the variables (not the **linear** relationship as Pearson does).

The variables tend to change together but not necessarily at a constant rate

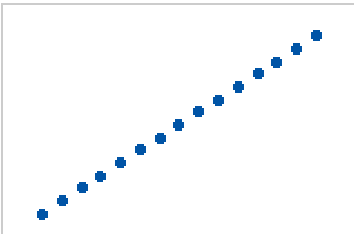
Non Parametric correlation: Spearman correlation coefficient

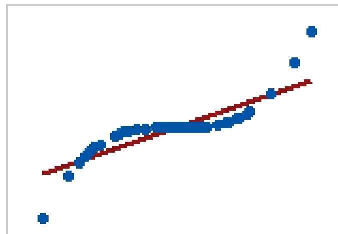


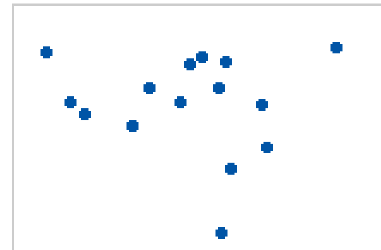
4. Correlation
2. Types of correlation

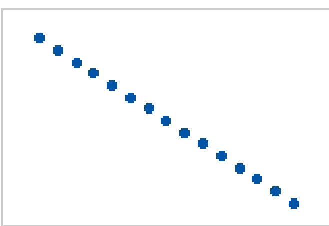
Vall d'Hebron Institut de Recerca

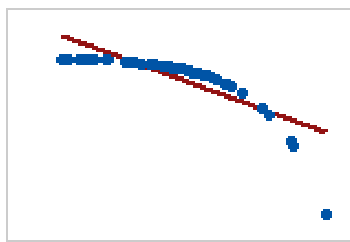
Comparison of Pearson and Spearman coefficients.

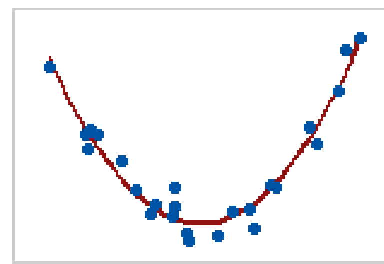

Pearson = +1, Spearman = +1


Pearson = +0.851, Spearman = +1


Pearson = -0.093, Spearman = -0.093


Pearson = -1, Spearman = -1


Pearson = -0.799, Spearman = -1


Coefficient of 0

Always examine a scatterplot to determine the form of the relationship

Spearman correlation coefficient. How to in R?

```
cor(osteoporosis$bua, osteoporosis$edad, method = "spearman")

[1] -0.3540295
```

Correlation matrix

```
cor(osteoporosis[, c("edad", "peso",  "talla", "imc", "bua", "menarqui")])
```

```
               edad         peso       talla         imc         bua     menarqui
edad      1.0000000  0.182629245 -0.16635268  0.26173285 -0.36018834  0.115901253
peso      0.1826292  1.000000000  0.23110585  0.89278635  0.09467837 -0.008526465
talla    -0.1663527  0.231105848  1.00000000 -0.22546438  0.13350207  0.070002843
imc       0.2617329  0.892786346 -0.22546438  1.00000000  0.03415938 -0.041607661
bua      -0.3601883  0.094678365  0.13350207  0.03415938  1.00000000 -0.085935539
menarqui  0.1159013 -0.008526465  0.07000284 -0.04160766 -0.08593554  1.000000000
```

4. Correlation. Exercises

Exercise 2. An hypothetic study, published last year that exists a relation between ***age*** and ***systolic blood pressure (sbp)*?** Do you think is it true? Show a scatterplot of the values? If not, find another variable in the dataset that has a good correlation with ***systolic blood pressure***.

Use dataset Framingham250.csv