Statistics Course with R - Day 3

Ricardo Gonzalo Sanz

6/10/2020

Contents

Elegant Graphics for data analysis

```
#install the package
#install.packages(ggplot2)
#load the package
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

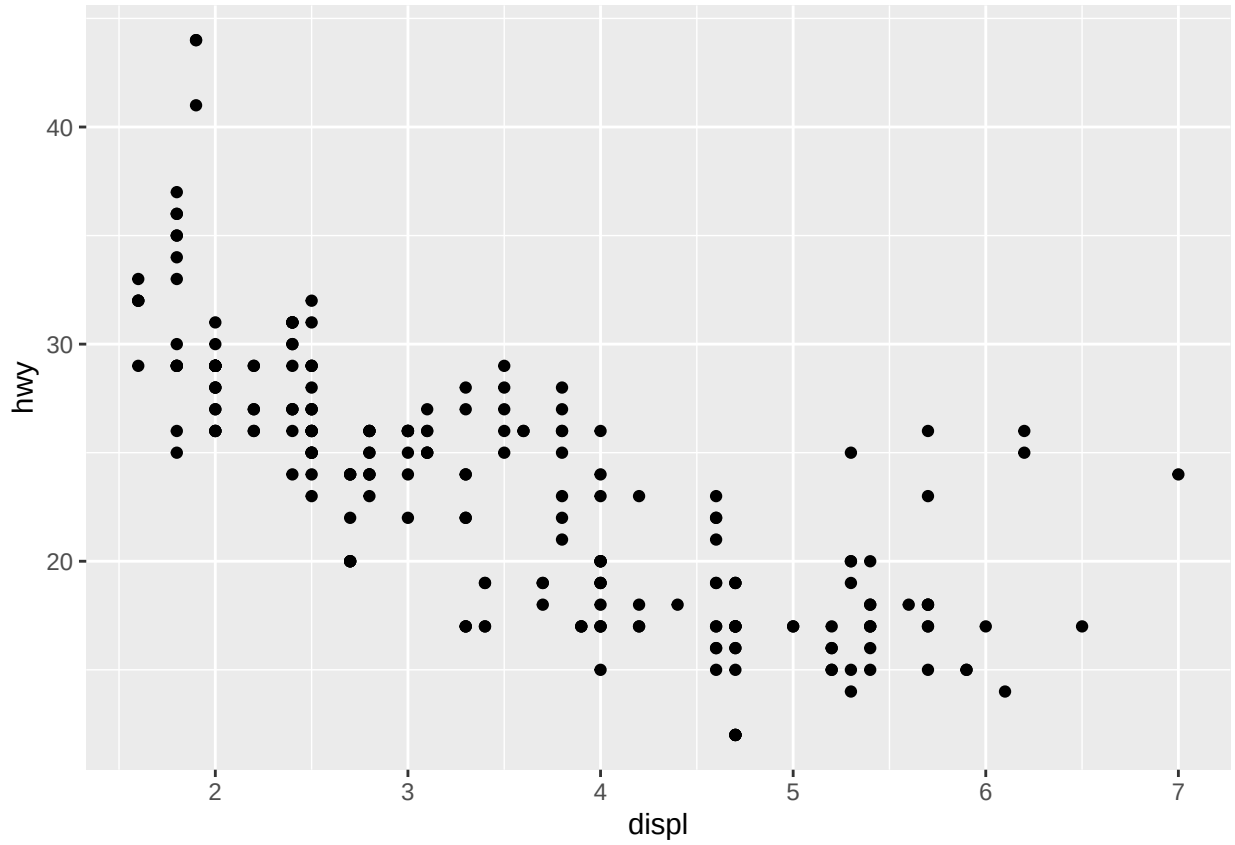
```
#see the data
head(mpg)
```

```
## # 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     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

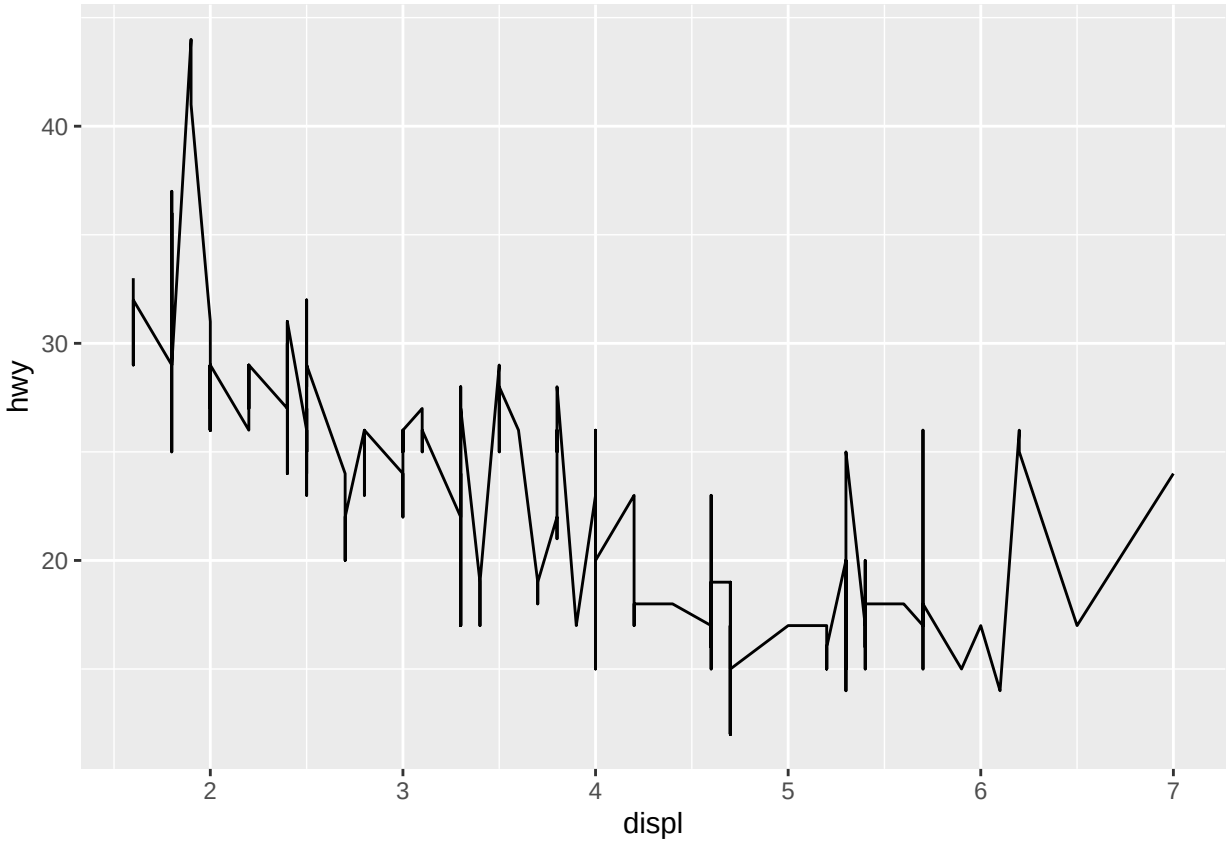
```
colnames(mpg)
```

```
##  [1] "manufacturer" "model"        "displ"        "year"         "cyl"
##  [6] "trans"        "drv"          "cty"          "hwy"          "fl"
## [11] "class"
```

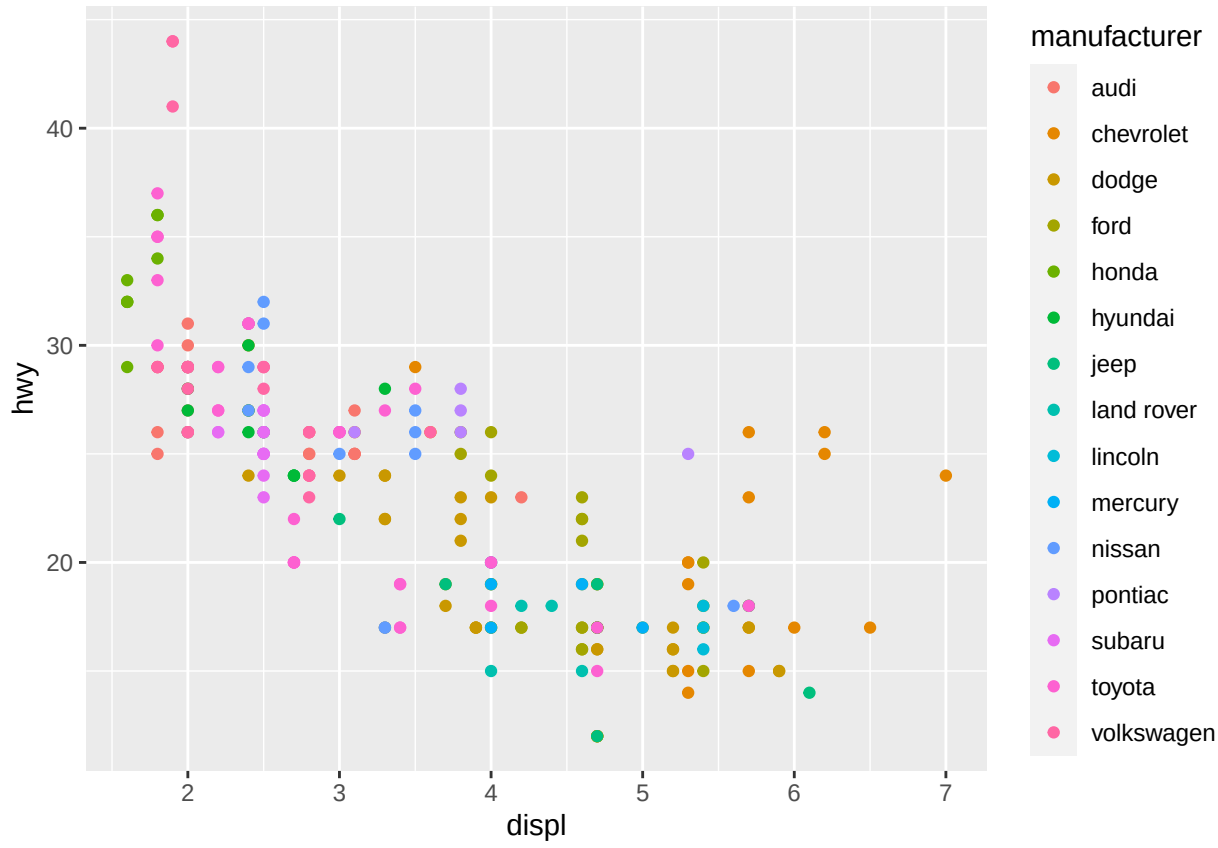
```
#do the basic plot
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point()
```



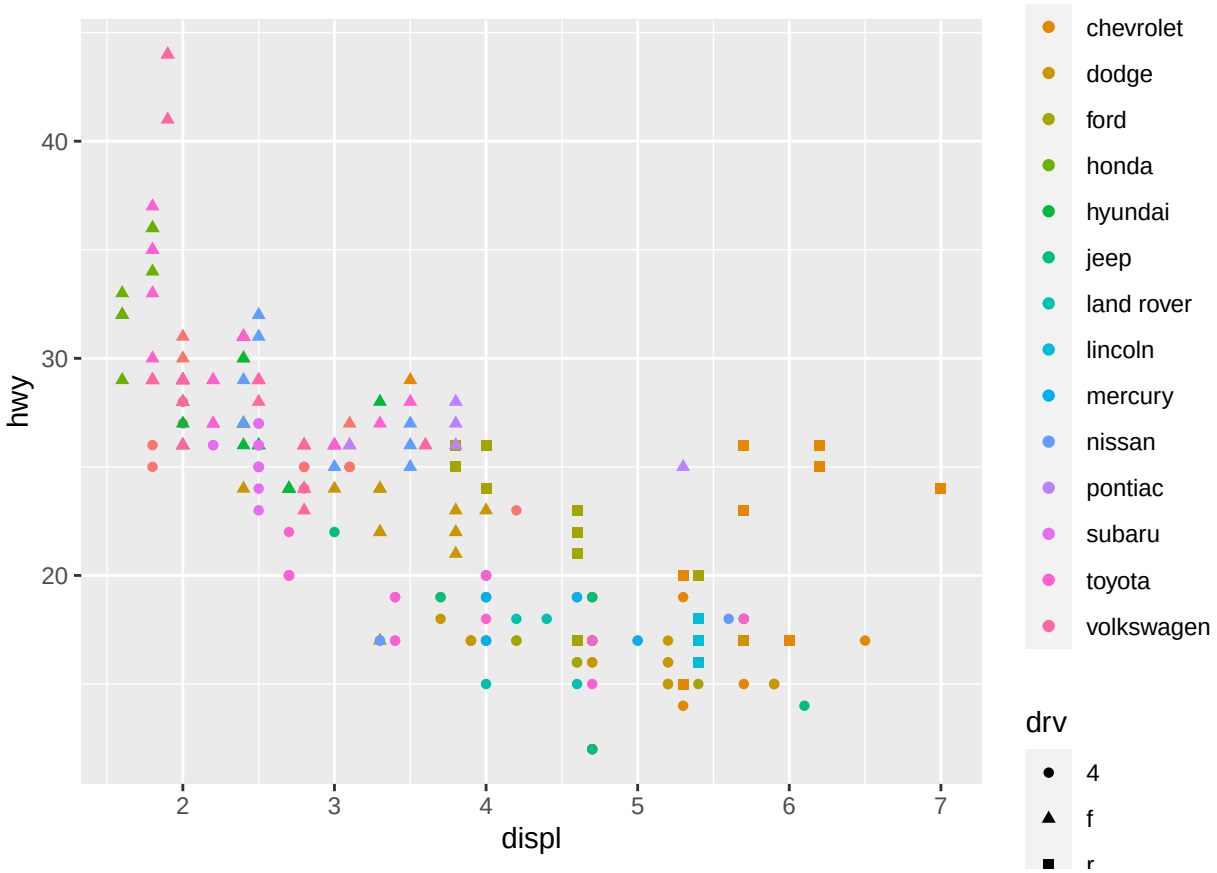

```
#change to lines
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_line()
```




```
#Improving the graphic (color by manufacturer)
ggplot(mpg, aes(x = displ, y = hwy, color = manufacturer)) +
  geom_point()
```




```
#Improving the graphic II (add shape by drv)
ggplot(mpg, aes(x = displ, y = hwy, color = manufacturer, shape =  drv)) +
  geom_point()
```



Bivariate Analysis

Qualitative versus qualitative

```
#load the data
osteoporosis <- read.csv2("osteoporosis.txt", sep = "\t", header = TRUE, dec = ",")

#see the data is correctly loaded
head(osteoporosis)
```

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

```
## 6       14       99    NO NO MENOPAUSIA/NO CONSTA SECUNDARIOS
```

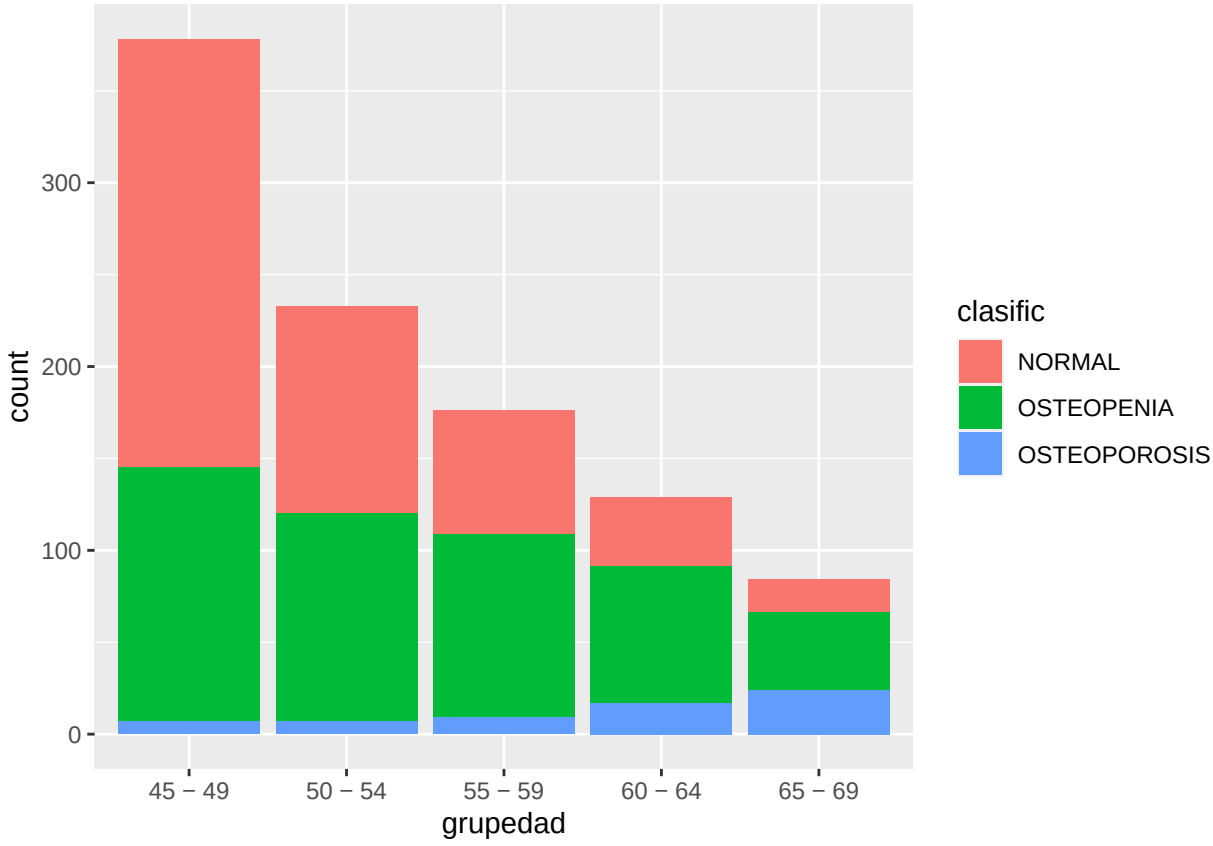
```r
#contingency table
table(osteoporosis$grupedad, osteoporosis$clasific)
```

```
##
##           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
```

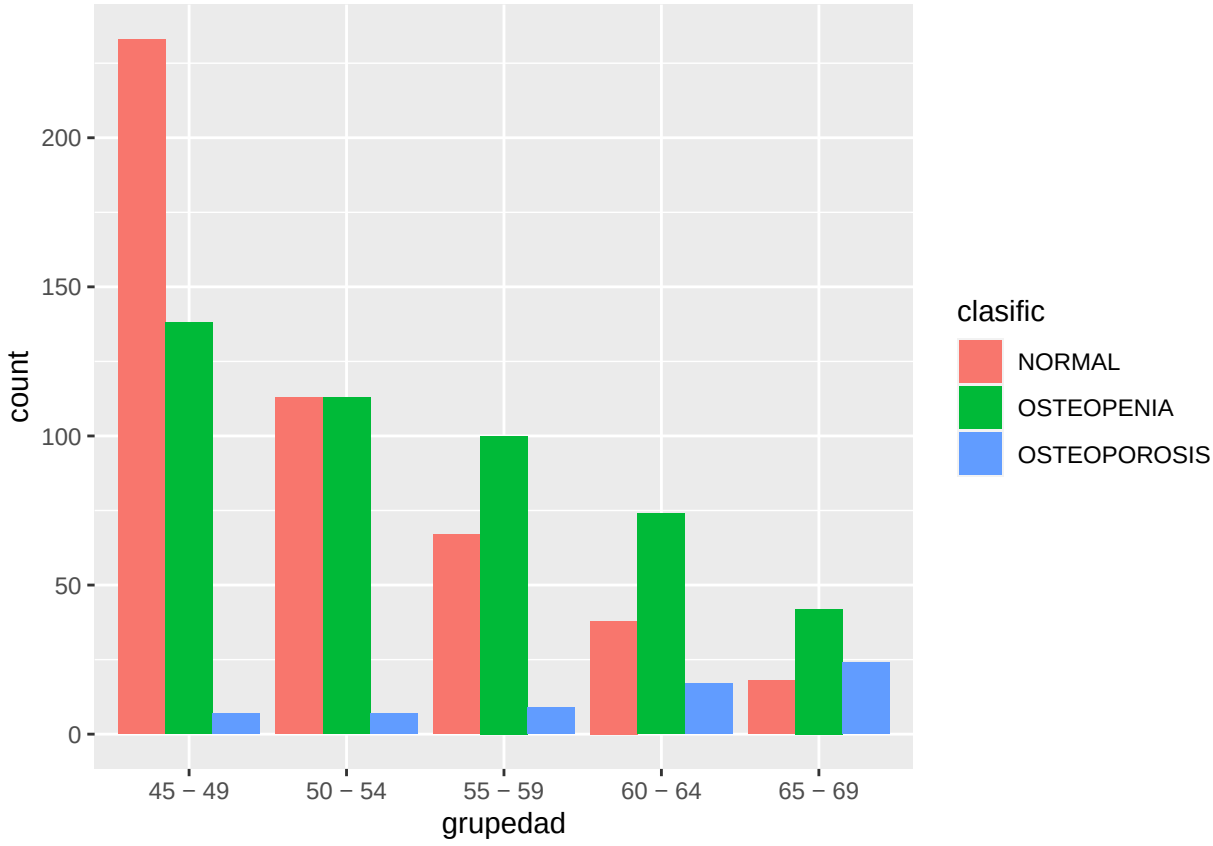
```r
#contingency table in %
prop.table(table(osteoporosis$grupedad, osteoporosis$clasific))
```

```
##
##           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
```

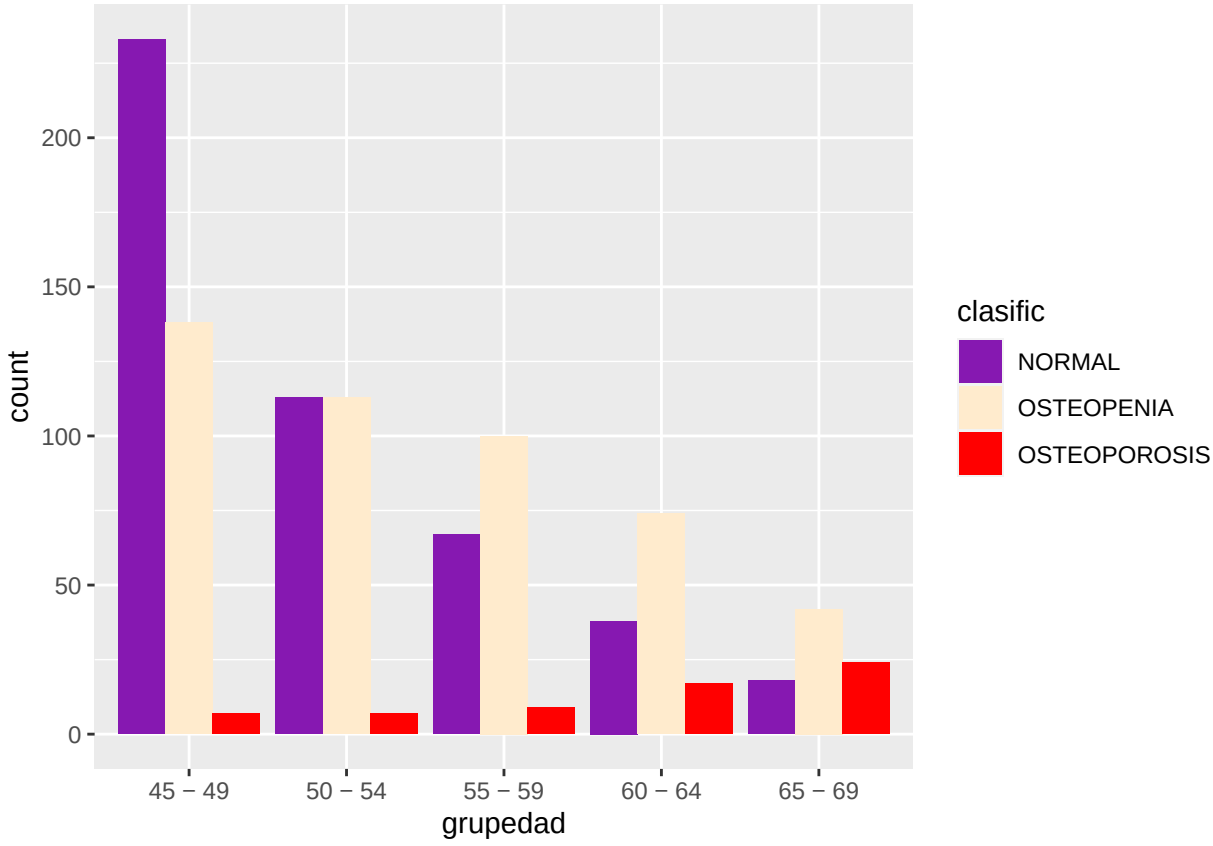
```r
#plot the data: stacket barplot
ggplot(data = osteoporosis, aes(x = grupedad)) +
  geom_bar(aes(fill = clasific))
```



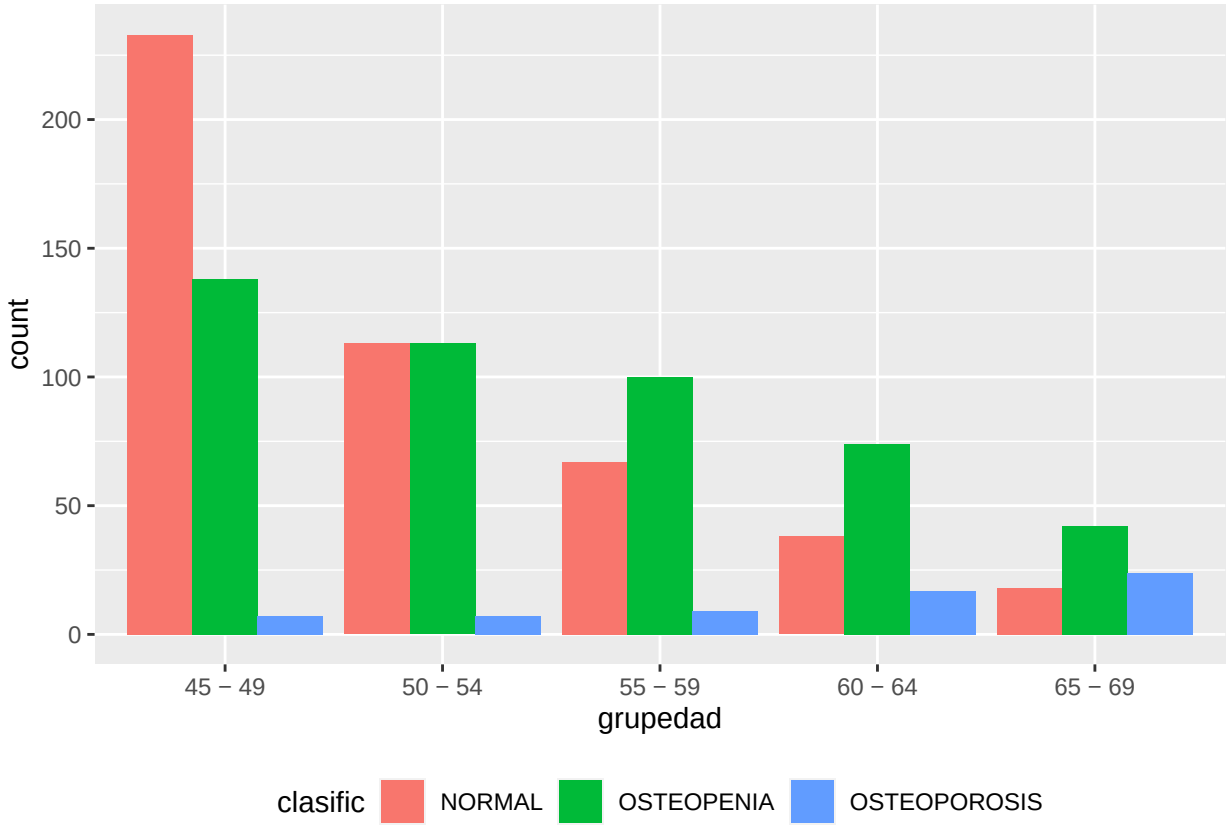

```
#plot the data II: bars side by side
ggplot(data = osteoporosis, aes(x = grupedad)) +
  geom_bar(aes(fill = clasific), position = "dodge")
```




```
#Change colors, legend position, labels and finally save it!
p <- ggplot(data = osteoporosis, aes(x = grupedad)) +
  geom_bar(aes(fill = clasific), position = "dodge")
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()
```

```
## pdf
##   2
```

Another to introduce the data

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

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

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

tab
```

```
##       Smokers Nonsmokers
```

```
## Men       120          60
## Women      50          70
```

```
#Look in %
prop.table(tab)
```

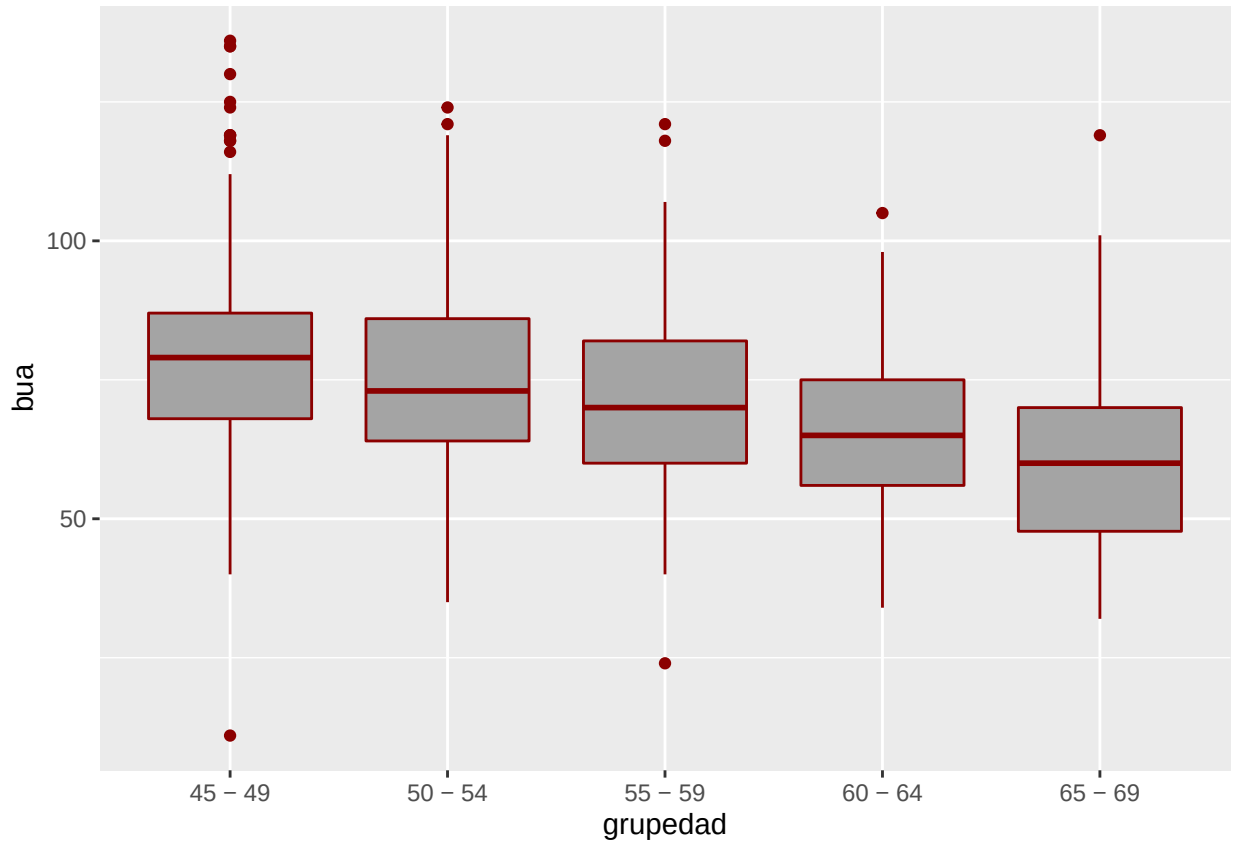
```
##          Smokers Nonsmokers
## Men   0.4000000  0.2000000
## Women 0.1666667  0.2333333
```

Qualitative versus quantitative

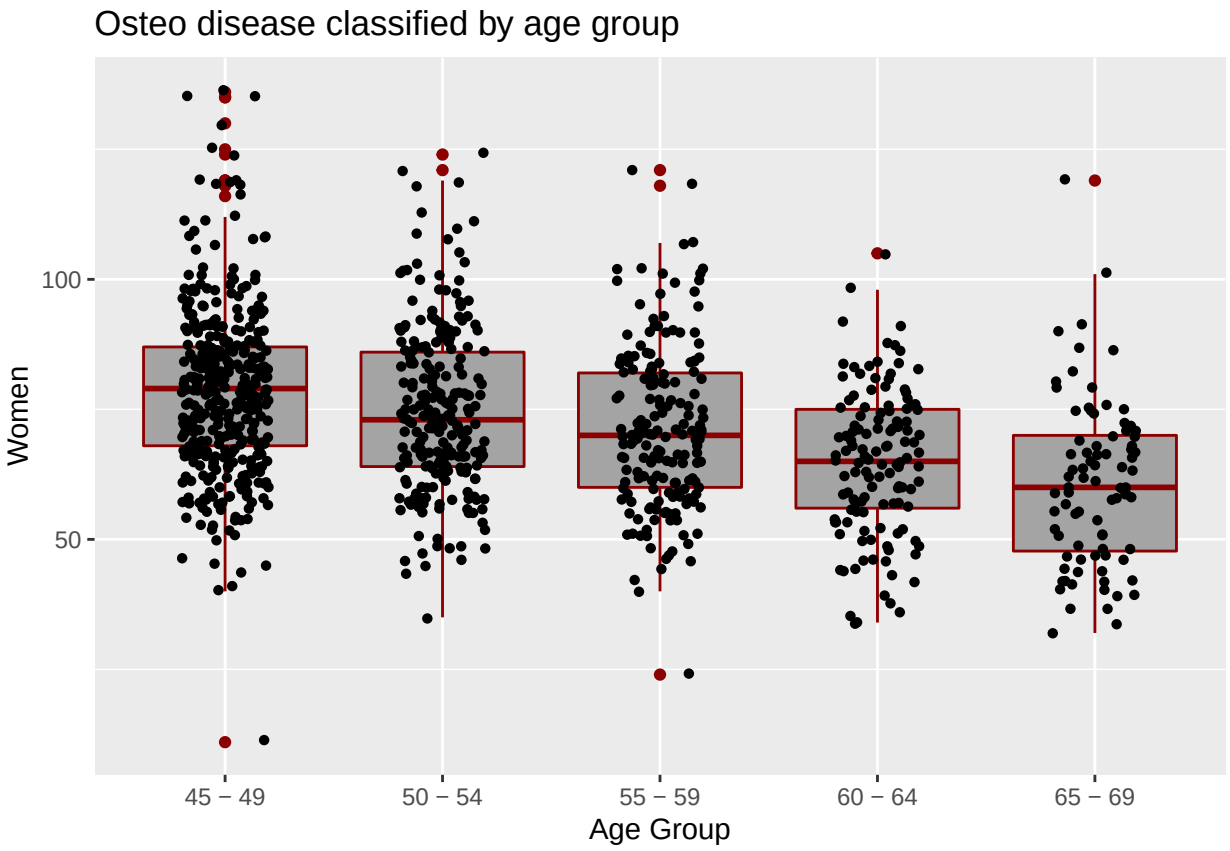
```
#Table of statistics
with(osteoporosis, tapply(bua, list(grupedad), mean, na.rm=TRUE))
```

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

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

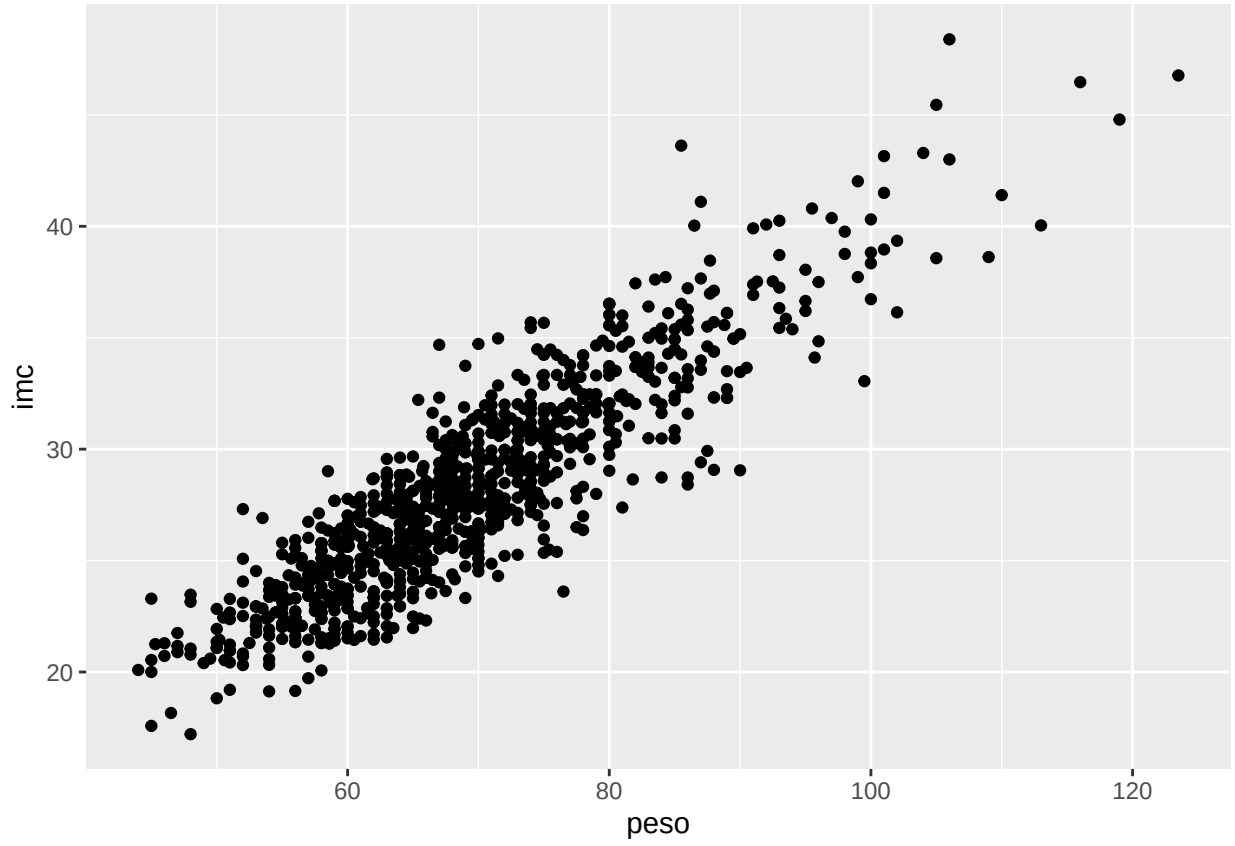


```r
# Box plot with points
# 0.2 : degree of jitter in x direction
bp + geom_jitter(shape = 16, position = position_jitter(0.2)) +
   labs(x = "Age Group", y = "Women", title = "Osteo disease classified by age group")
```

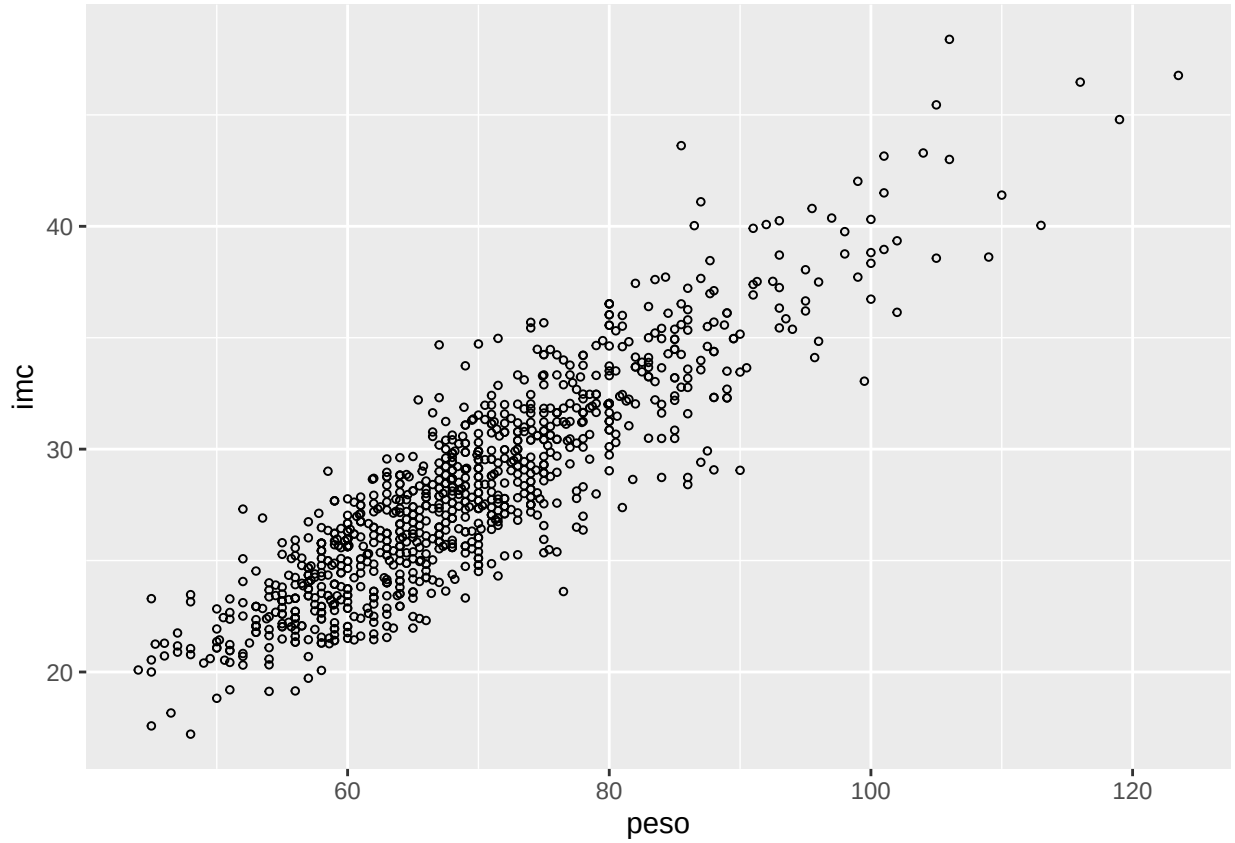


Quantitative versus quantitative

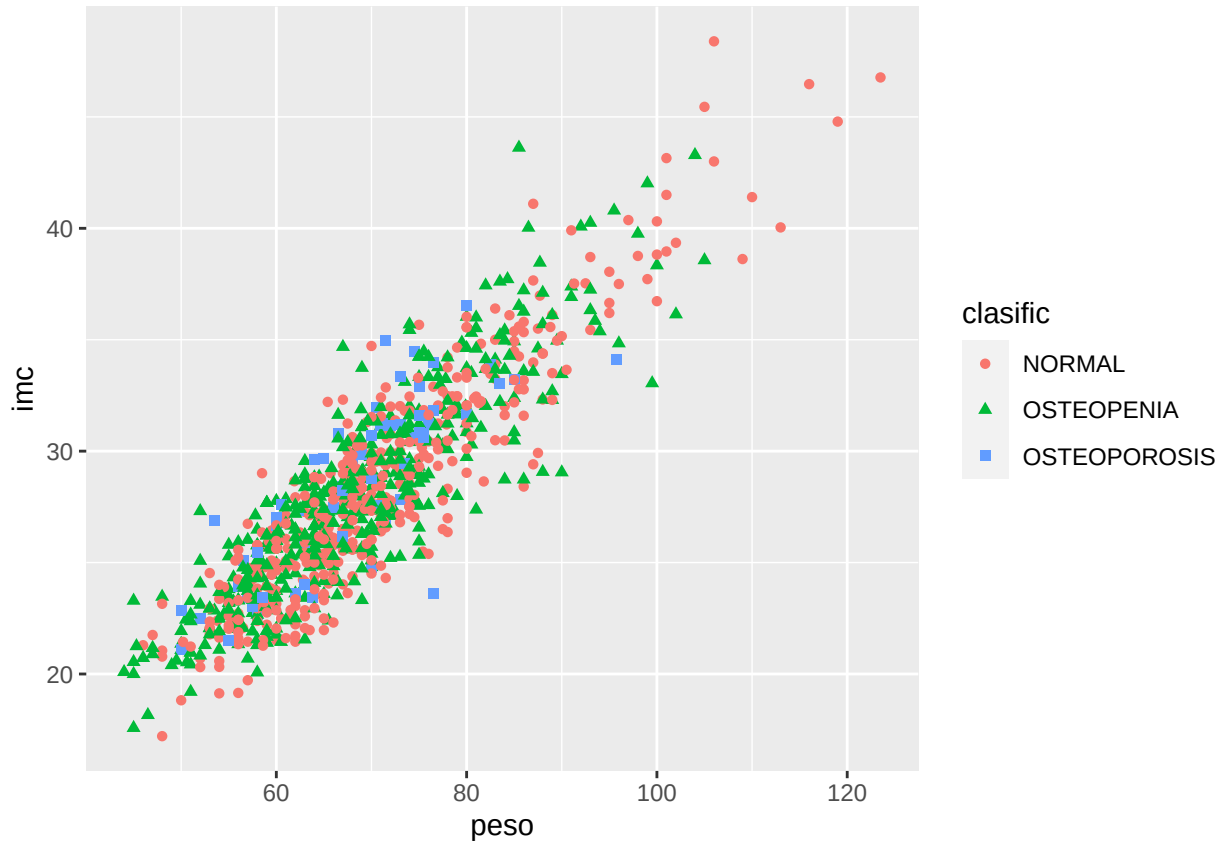
```r
# Basic scatter plot
ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point()
```



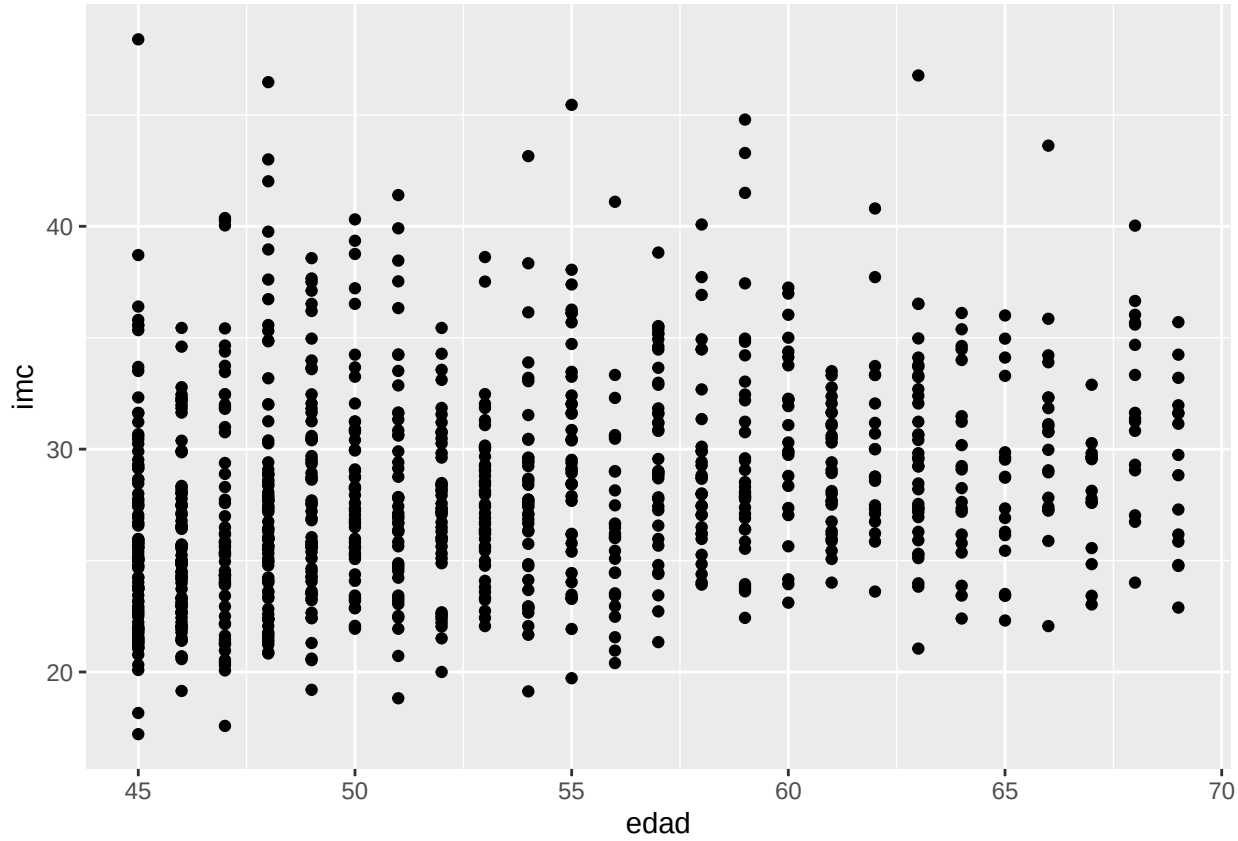

```
# Change the point size, and shape
ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point(size = 1, shape = 1)
```



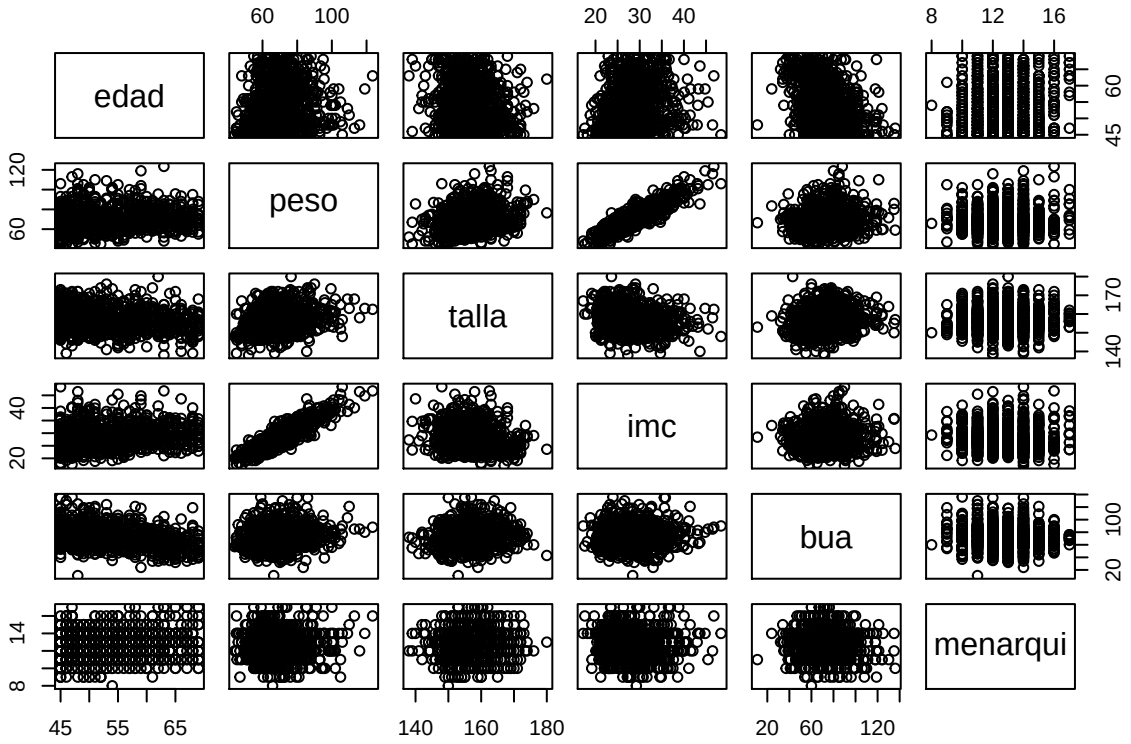

```
# Color the points depending of another variable
ggplot(osteoporosis, aes(x = peso, y = imc, color = clasific, shape = clasific)) +
  geom_point()
```




```
#not always the correlation is good
ggplot(osteoporosis, aes(x = edad, y = imc)) +
  geom_point()
```




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

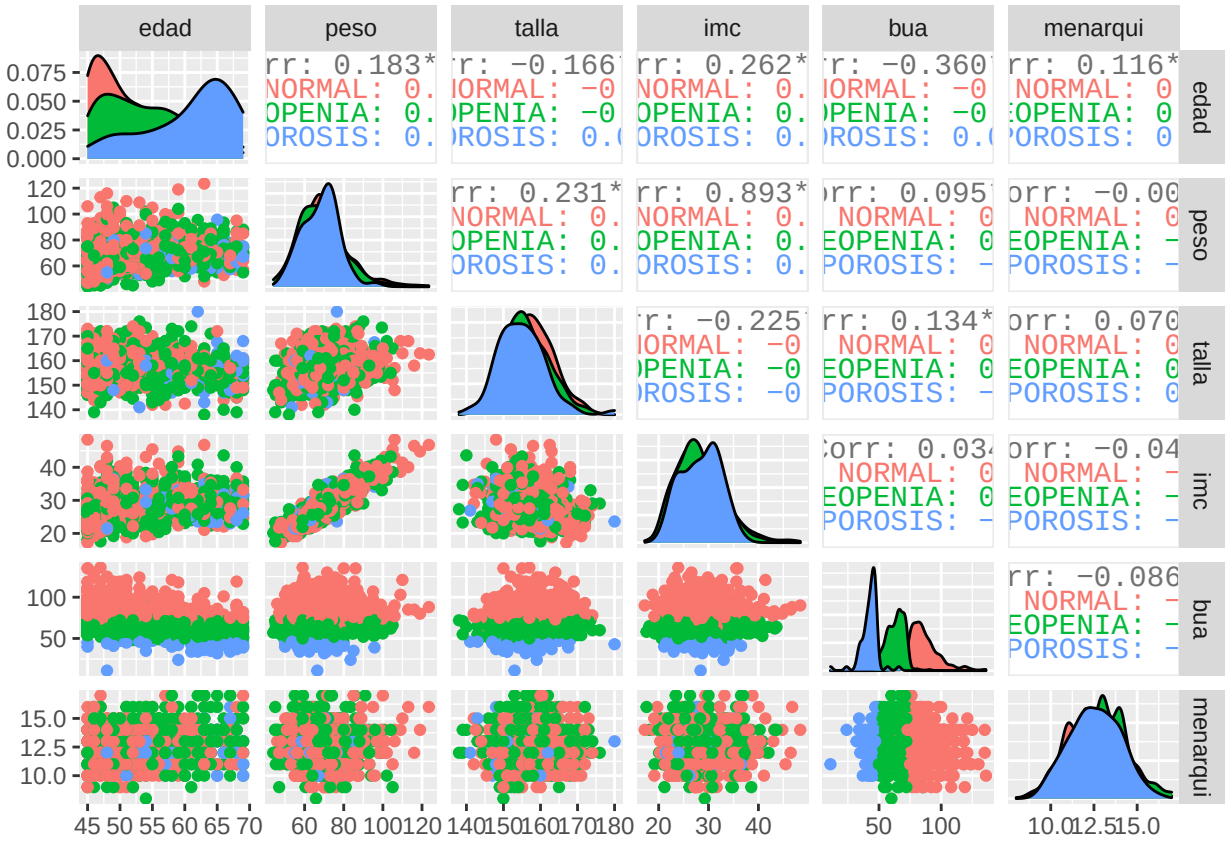



```
#with ggplots
#install.packages(GGally)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.0.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

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



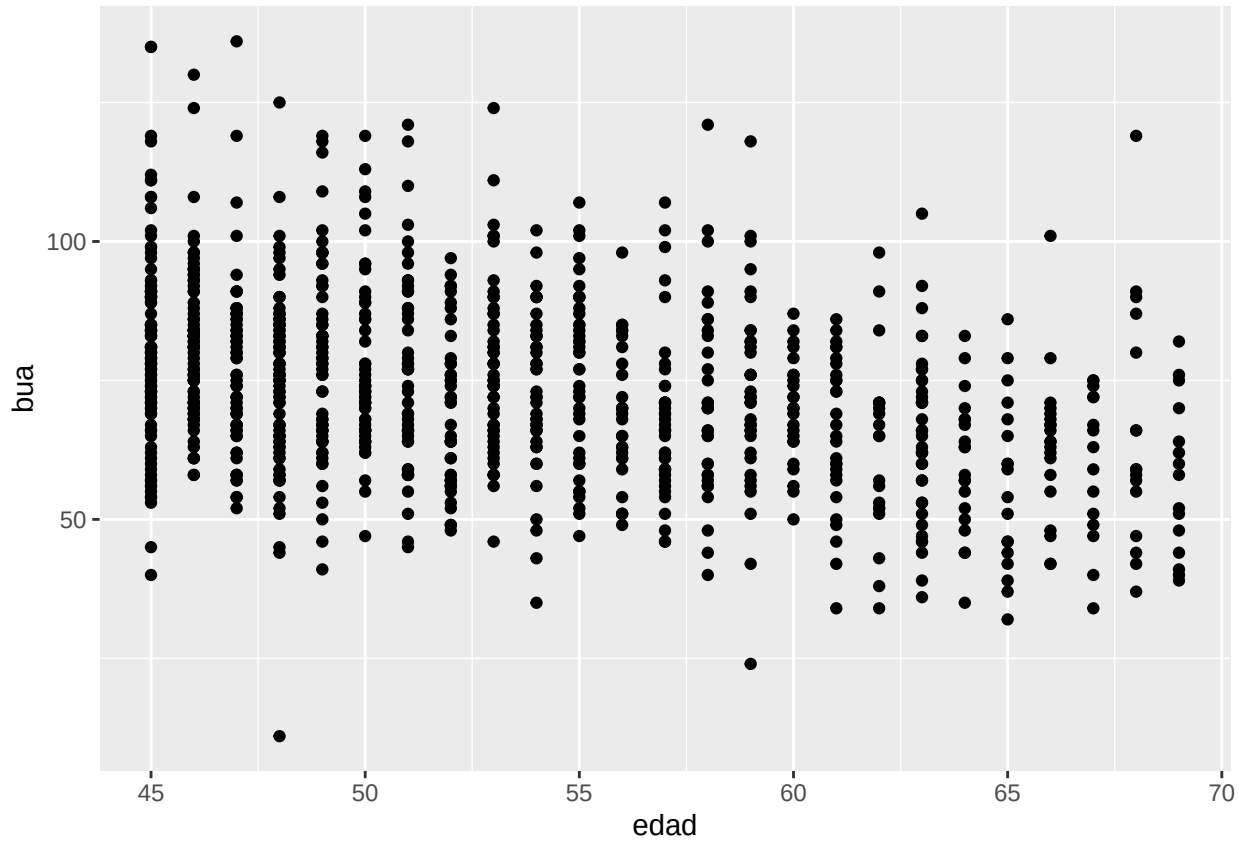
Correlation

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

```
## [1] -0.3601883
```

```
#the plot

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




```r
#Spearman correlation
cor(osteoporosis$bua, osteoporosis$edad, method = "spearman")
```

```
## [1] -0.3540295
```