Statistics Course with R - Day 3

UEB

27/04/2021

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 4.0.3
```

```
#see the data (we'll take data from package)
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"
```

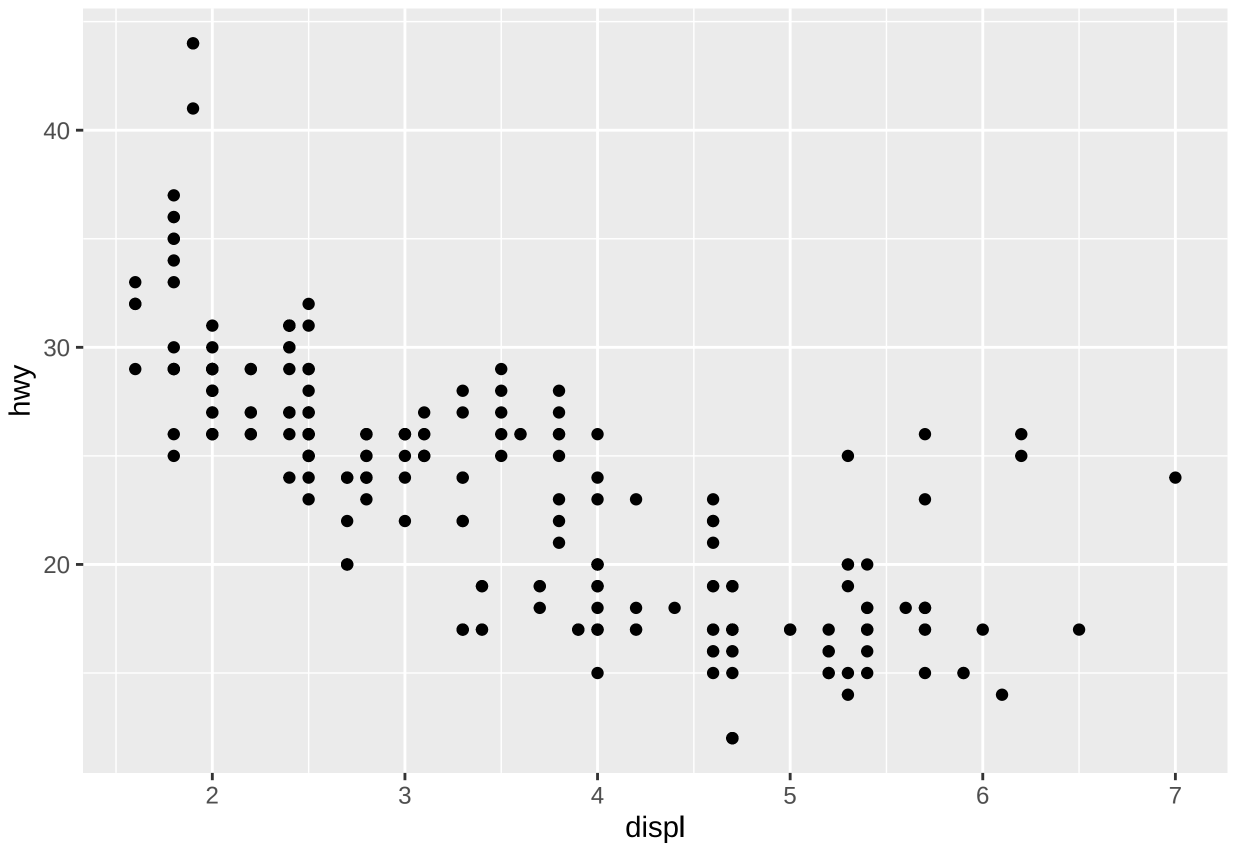
```r
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

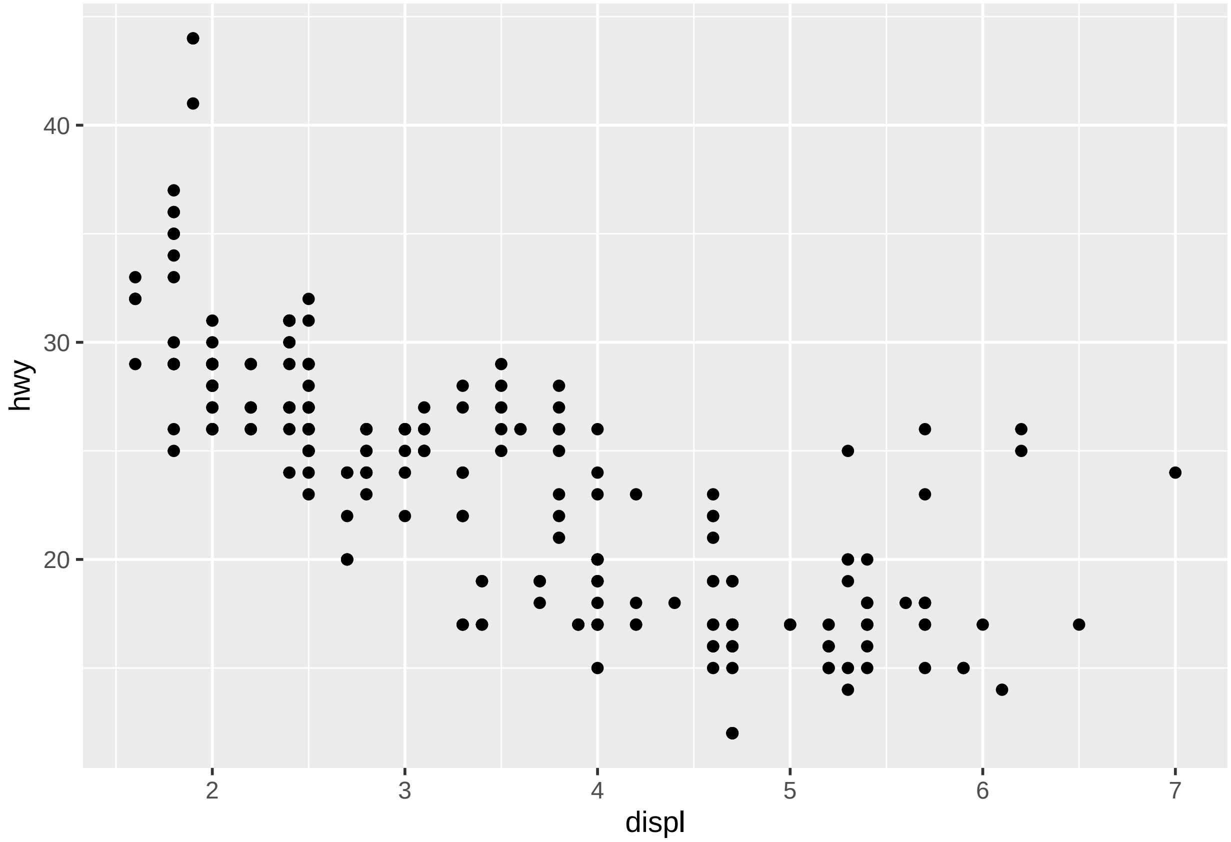
```r
summary(mpg)
```

```
##  manufacturer          model               displ            year     
##  Length:234         Length:234         Min.   :1.600   Min.   :1999  
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999  
##  Mode  :character   Mode  :character   Median :3.300   Median :2004  
##                                        Mean   :3.472   Mean   :2004  
##                                        3rd Qu.:4.600   3rd Qu.:2008  
##                                        Max.   :7.000   Max.   :2008  
##       cyl           trans               drv                 cty       
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00  
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00  
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00  
##  Mean   :5.889                                         Mean   :16.86  
##  3rd Qu.:8.000                                         3rd Qu.:19.00  
##  Max.   :8.000                                         Max.   :35.00  
##       hwy             fl               class          
##  Min.   :12.00   Length:234         Length:234        
##  1st Qu.:18.00   Class :character   Class :character  
##  Median :24.00   Mode  :character   Mode  :character  
##  Mean   :23.44                                        
##  3rd Qu.:27.00                                        
##  Max.   :44.00                                        
```

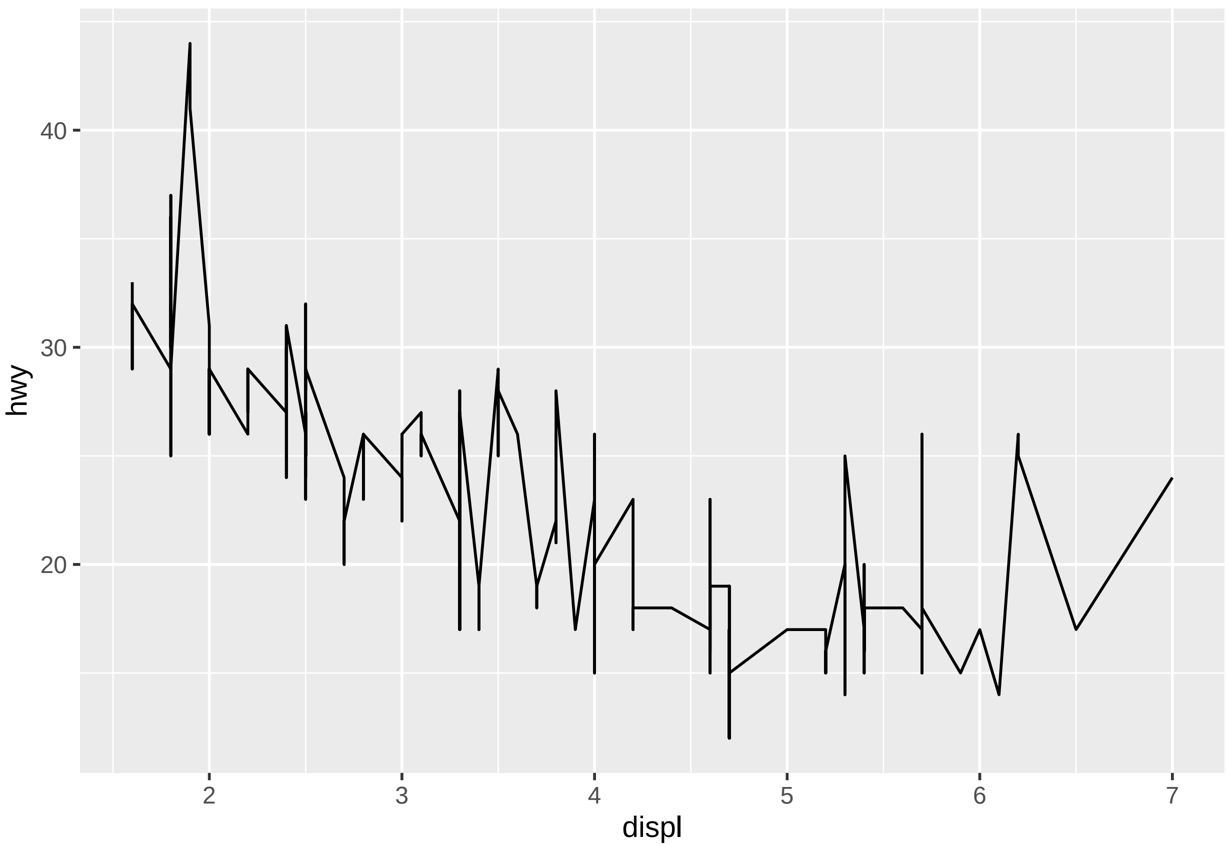
```r
#do the basic plot
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point()
```



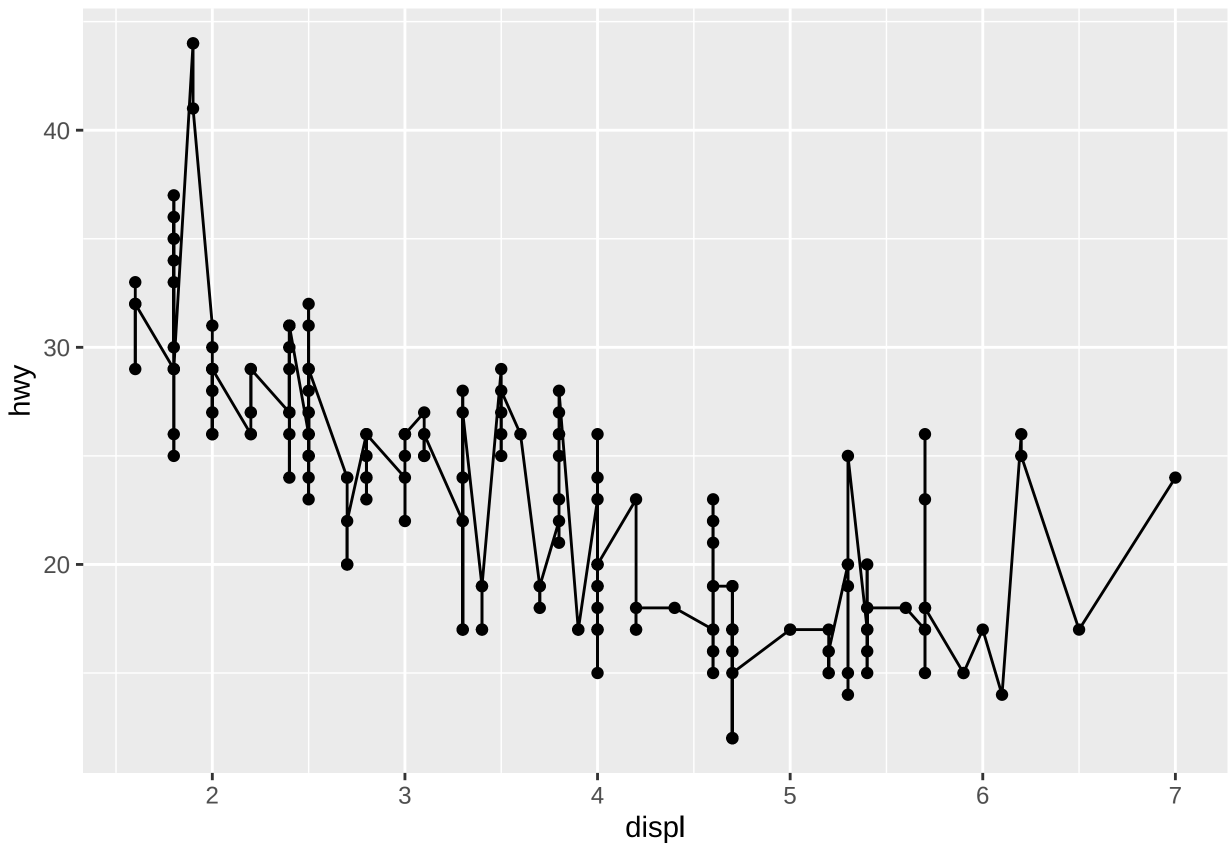
```
#it can be assigned to an object too
p <- ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point()
p
```



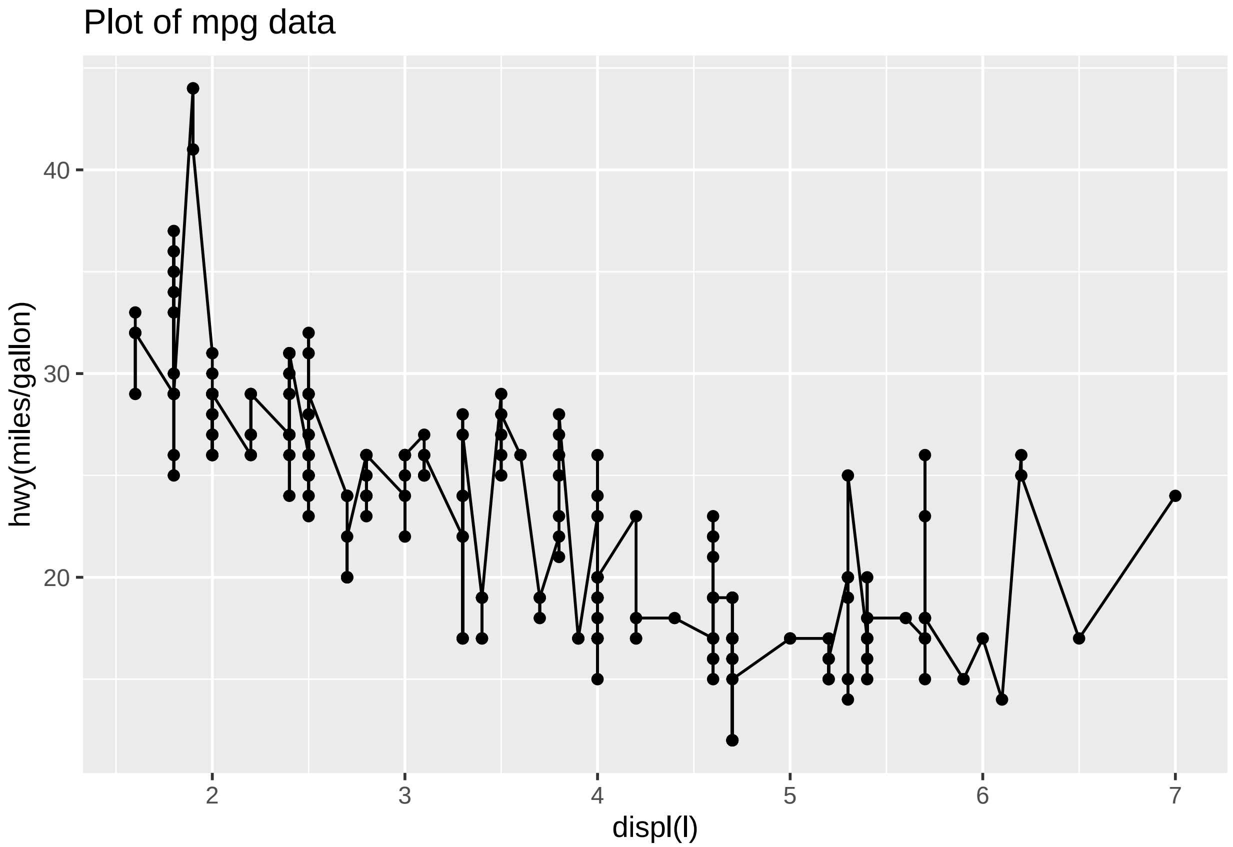
```
#change geom to lines
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_line()
```



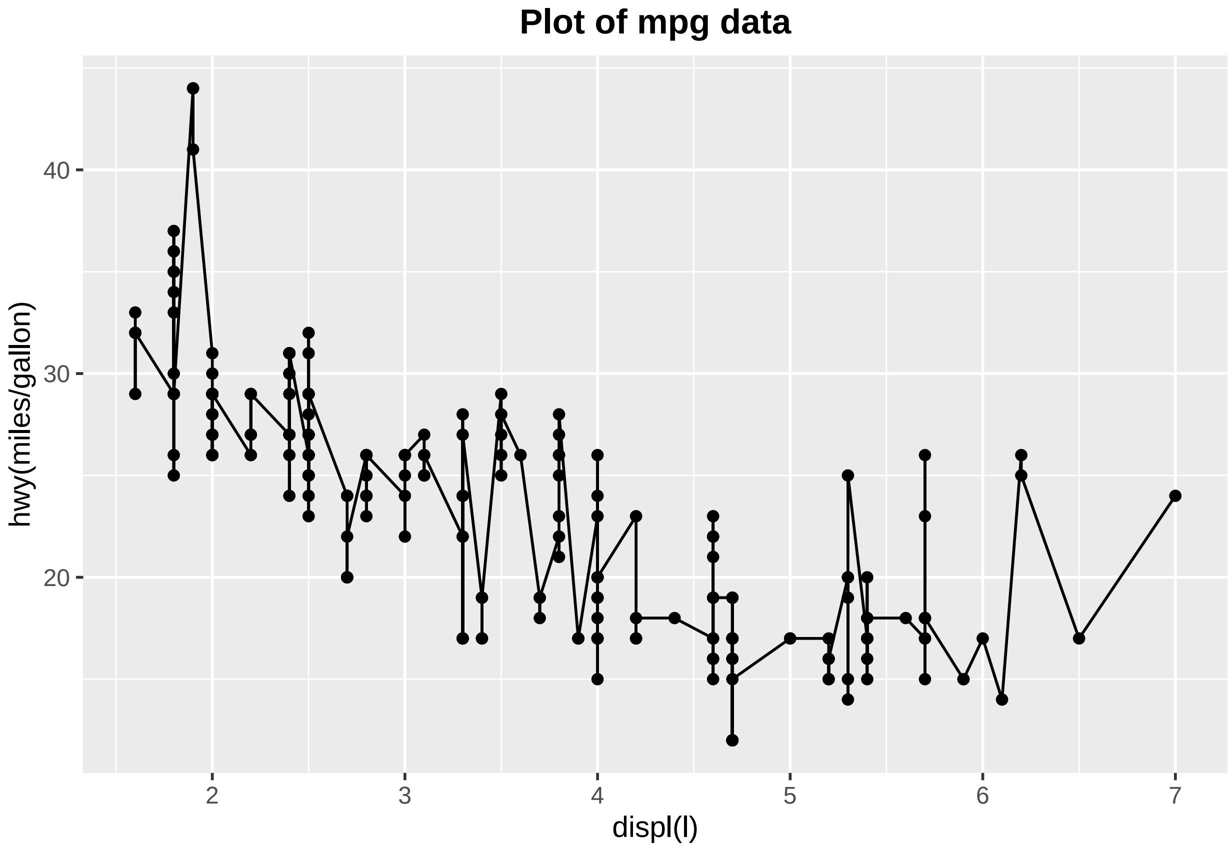
```
#add layers
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  geom_line()
```



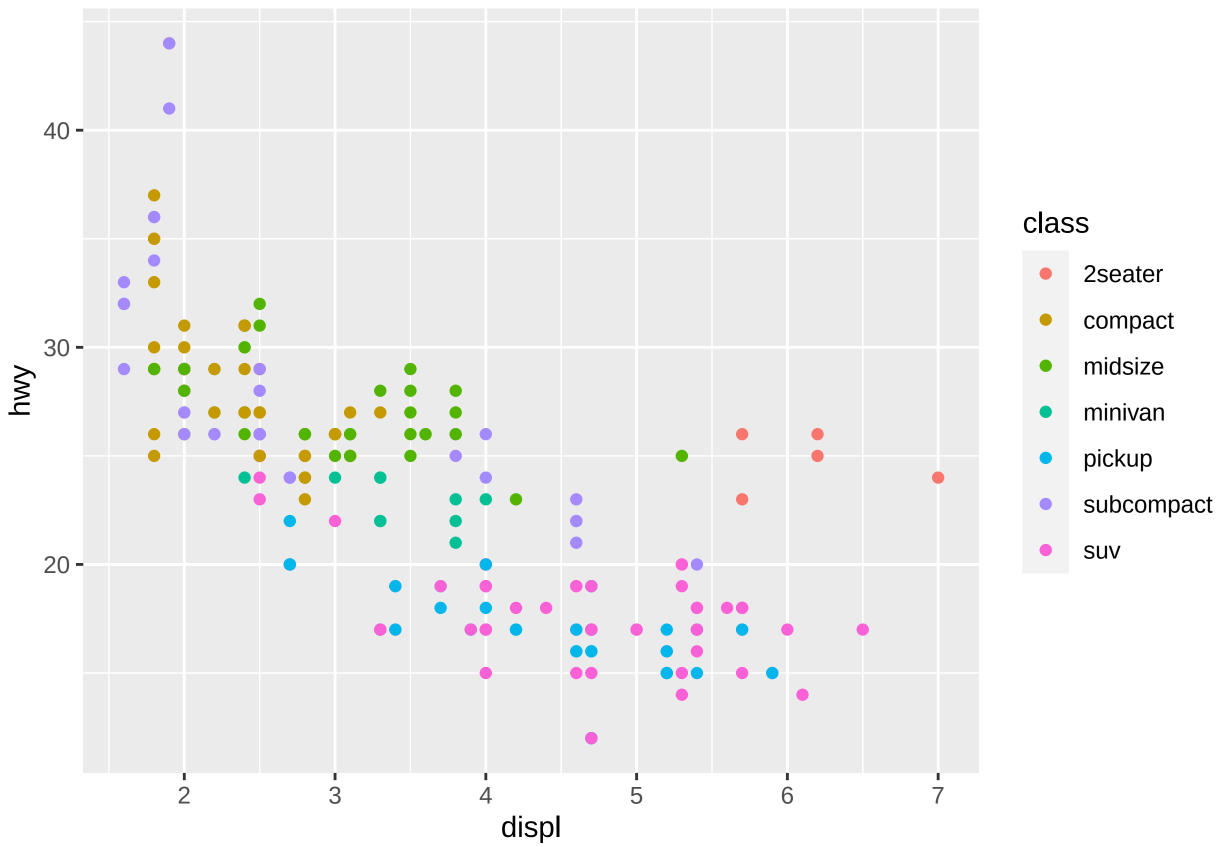
```r
#add layers (title)
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  geom_line() +
  labs(title="Plot of mpg data", x="displ(l)", y="hwy(miles/gallon)")
```



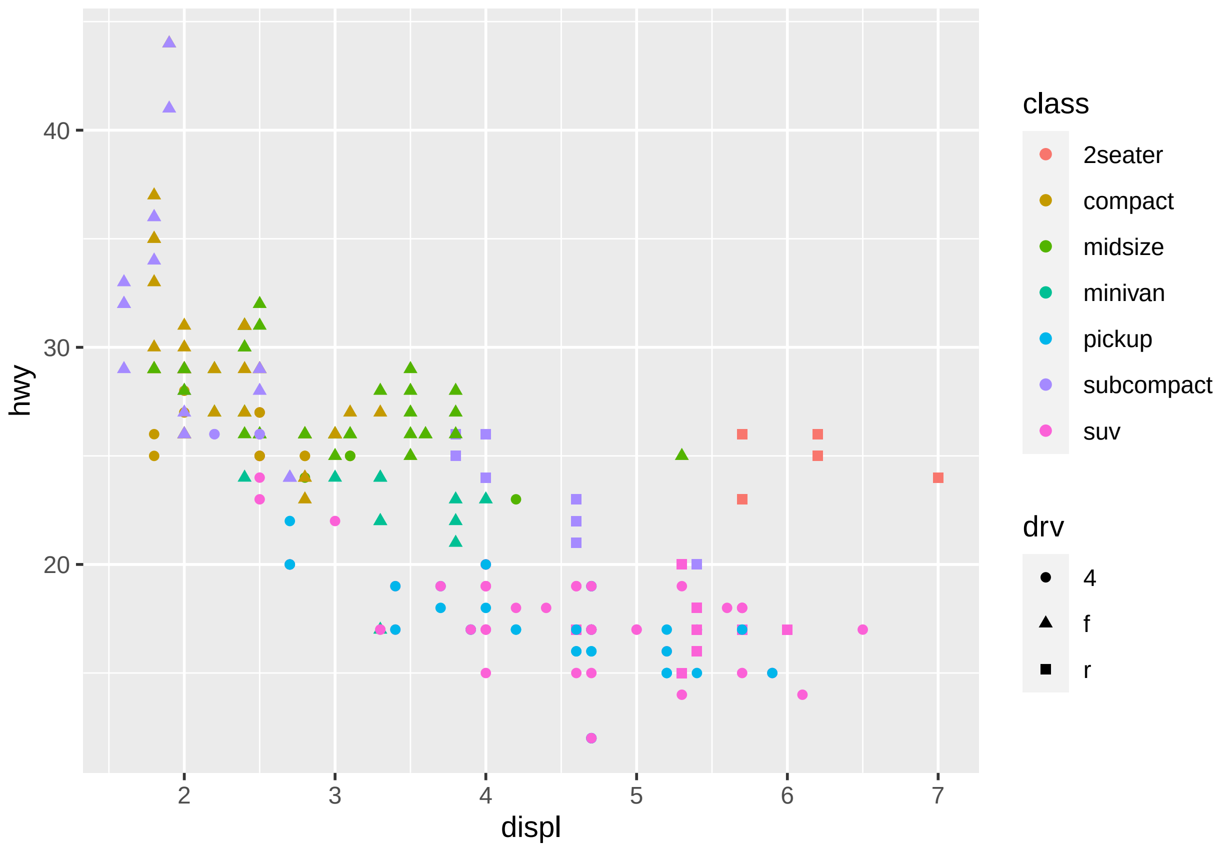
```r
#formatting labels
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  geom_line() +
  labs(title="Plot of mpg data", x="displ(l)", y="hwy(miles/gallon)") +
  theme(plot.title=element_text(face="bold", hjust=0.5))
```



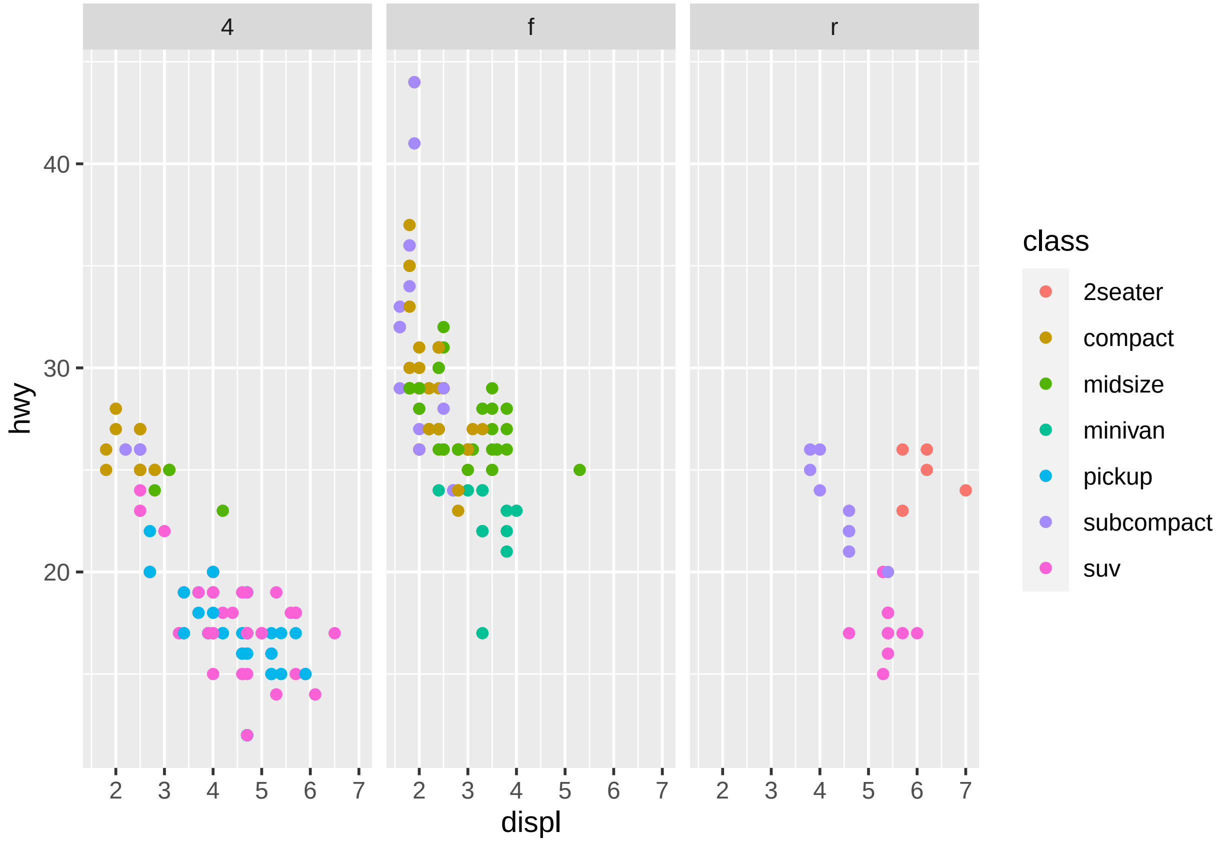
```
#Playing with aes (color by class)
ggplot(mpg, aes(x = displ, y = hwy, color = class)) +
  geom_point()
```



```
#Playing with aes (add shape by drv)
ggplot(mpg, aes(x = displ, y = hwy, color = class, shape =  drv)) +
  geom_point()
```



```r
#Facets
ggplot(mpg, aes(x = displ, y = hwy, color = class)) +
  geom_point() +
  facet_grid(. ~ drv)
```



Descriptive statistics for univariate data

Quantitative variables

- Numeric summaries

```
mean(mpg$displ)
```

```
## [1] 3.471795
```

```
median(mpg$displ)
```

```
## [1] 3.3
```

```
sd(mpg$displ)
```

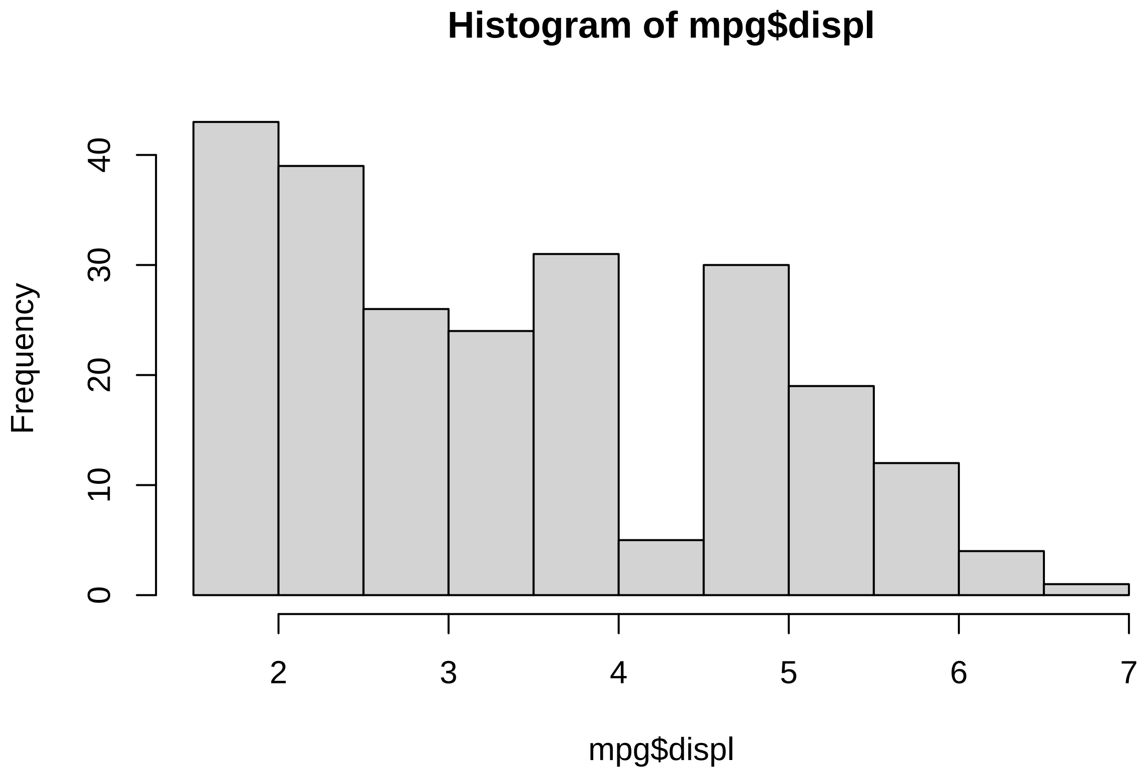
```
## [1] 1.291959
```

```
#using summary function
summary(mpg$displ)
```

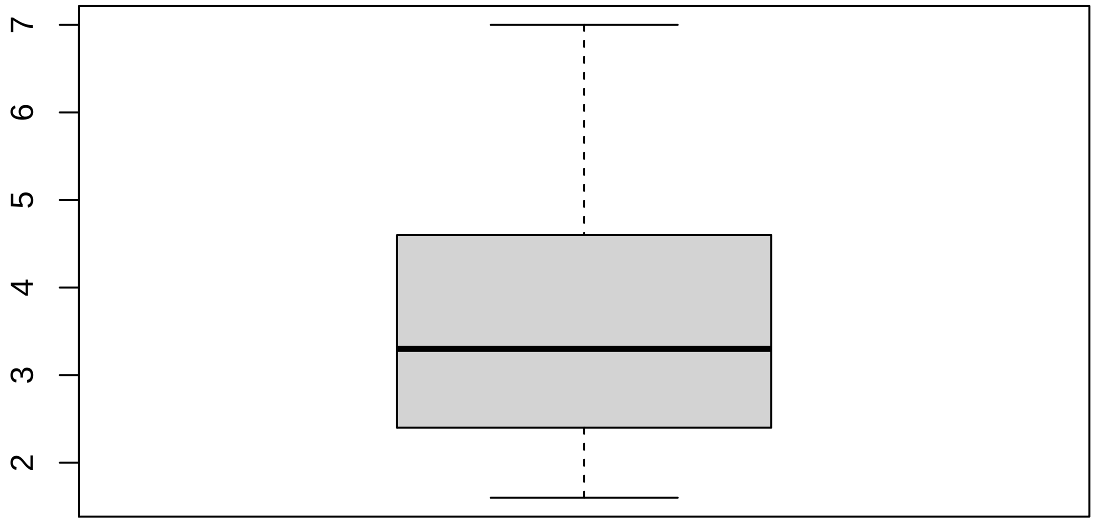
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.600   2.400   3.300   3.472   4.600   7.000
```

- Graphic summaries

```
hist(mpg$displ)
```



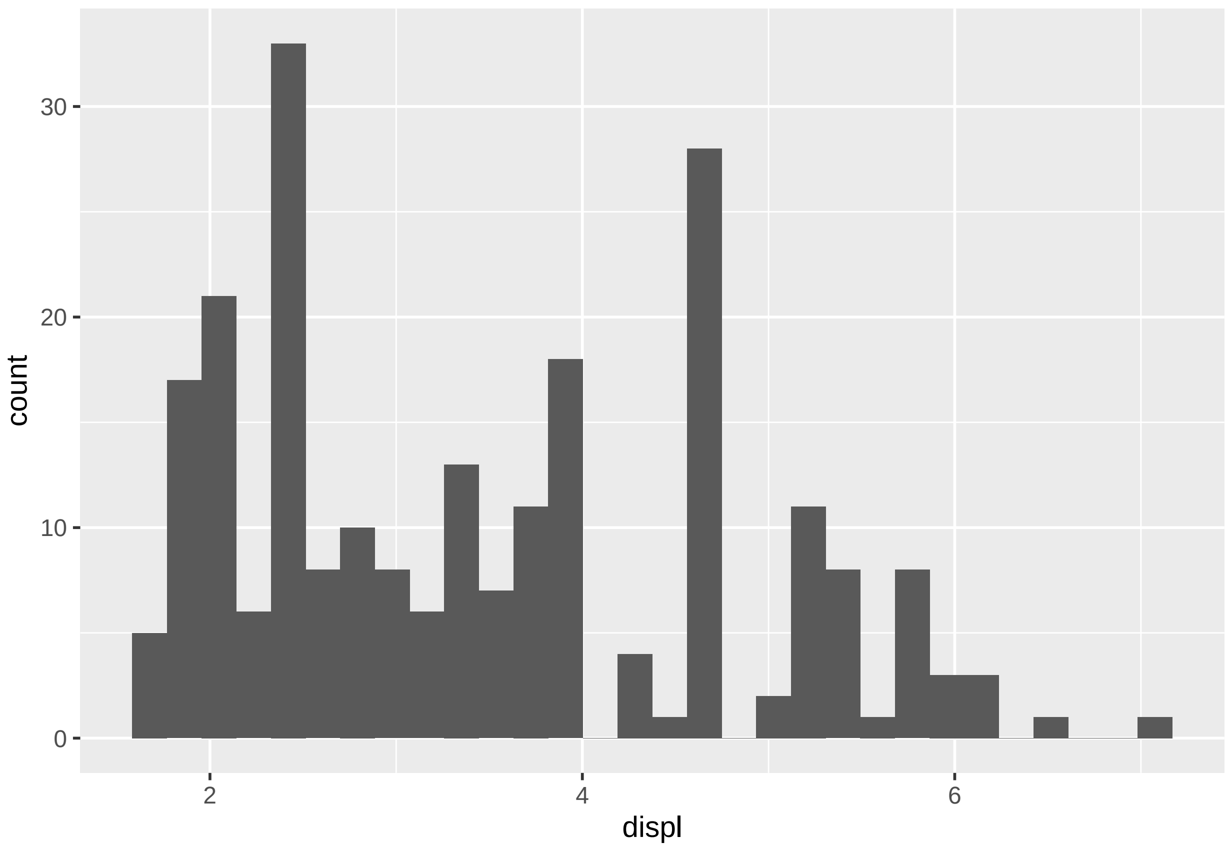
```
boxplot(mpg$displ)
```



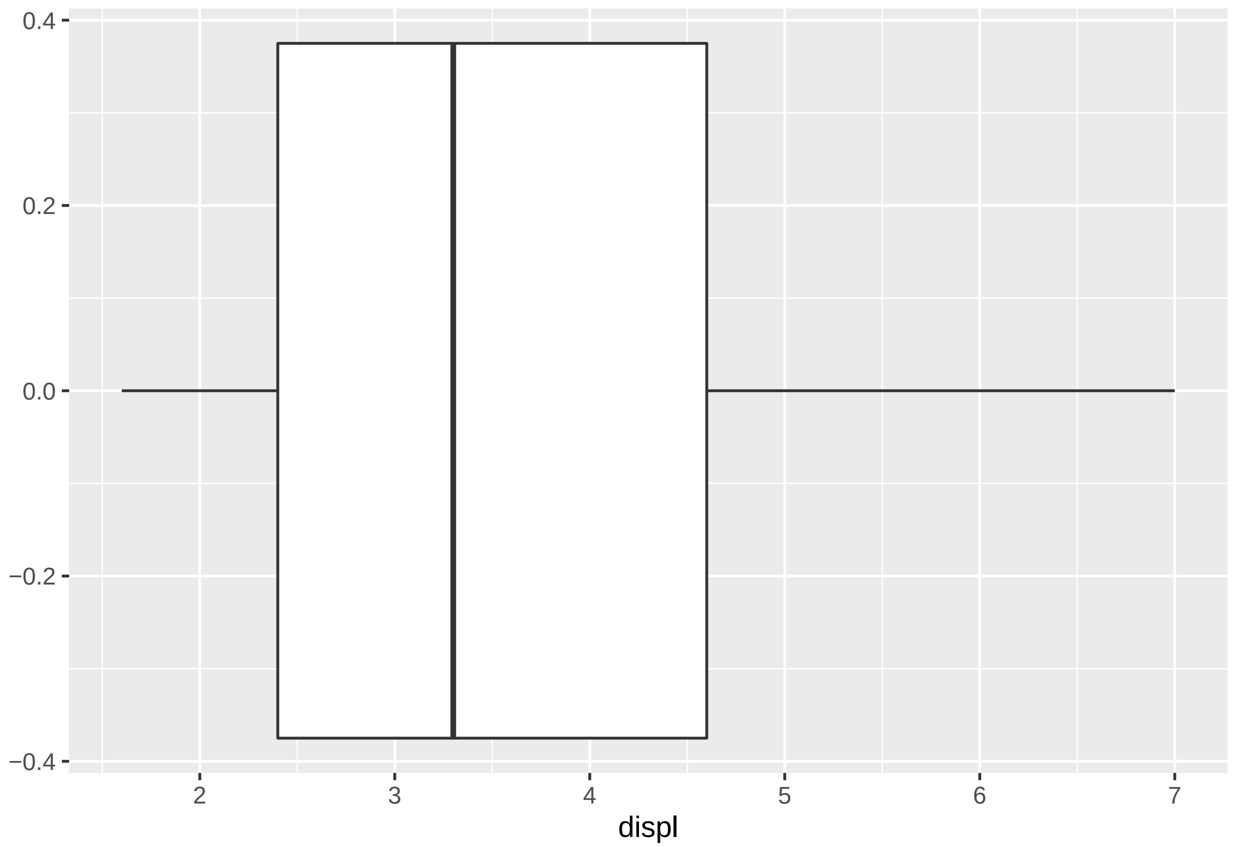
With ggplot

```
ggplot(mpg, aes(displ)) +
  geom_histogram()
```

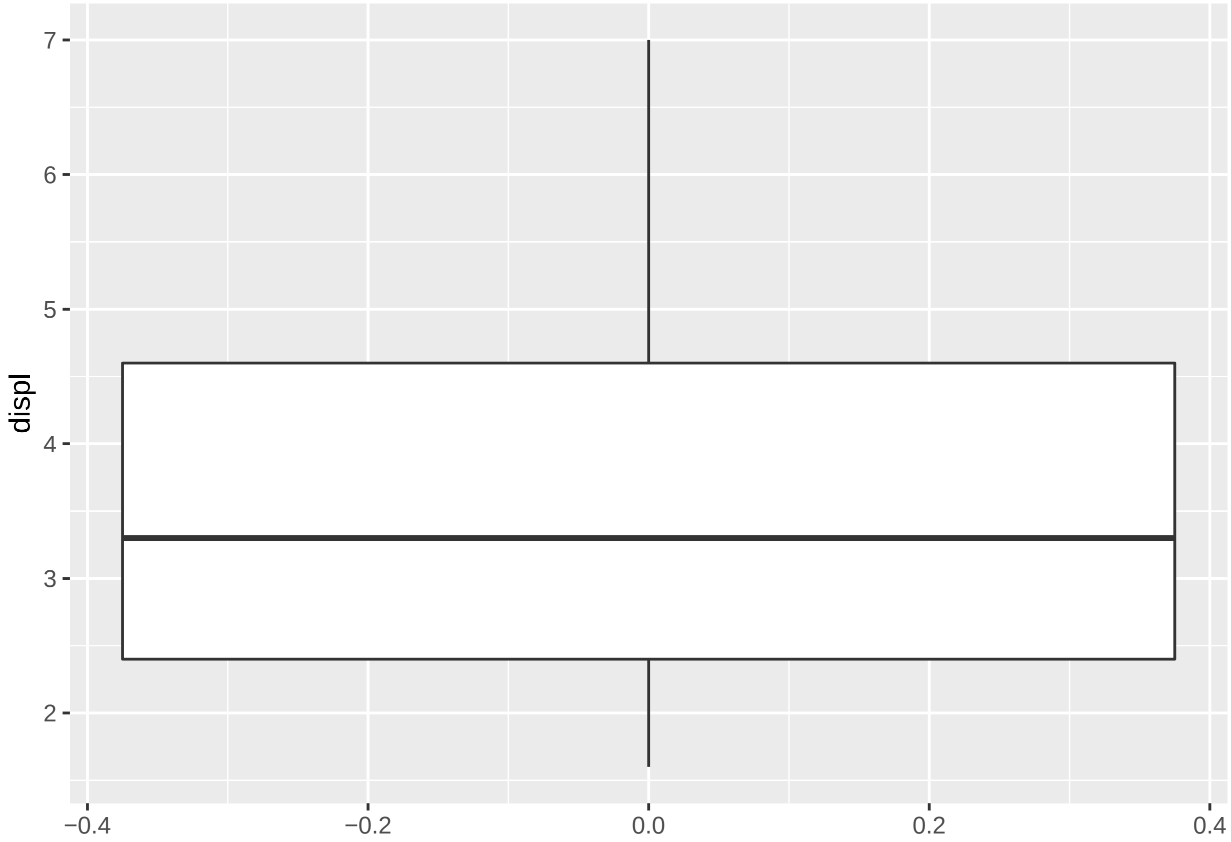
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(mpg, aes(displ)) +
  geom_boxplot()
```



```
ggplot(mpg, aes(displ)) +
  geom_boxplot() +
  coord_flip()
```



Qualitative variables

- Numeric summaries

```
#absolute frequencies
table(mpg$class)
```

```
##
##    2seater    compact    midsize    minivan     pickup subcompact        suv
##          5         47         41         11         33         35         62
```

```
#relative frequencies
prop.table(table(mpg$class))
```

```
##
##    2seater    compact    midsize    minivan     pickup subcompact        suv
## 0.02136752 0.20085470 0.17521368 0.04700855 0.14102564 0.14957265 0.26495726
```

```
# install.packages("gmodels")
library(gmodels)
```

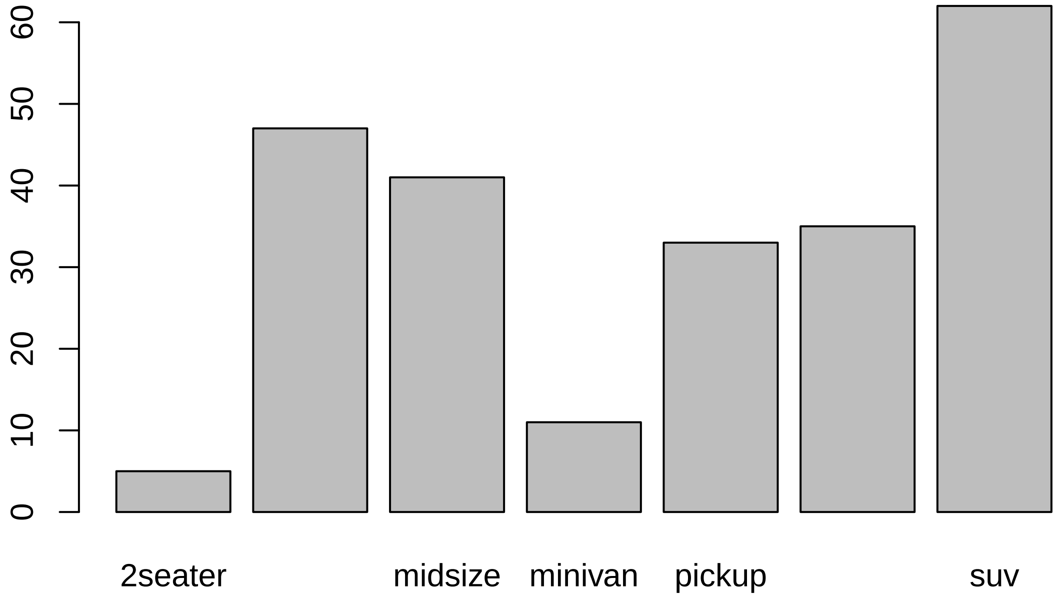
```
## Warning: package 'gmodels' was built under R version 4.0.5
```

```
CrossTable(mpg$class)
```

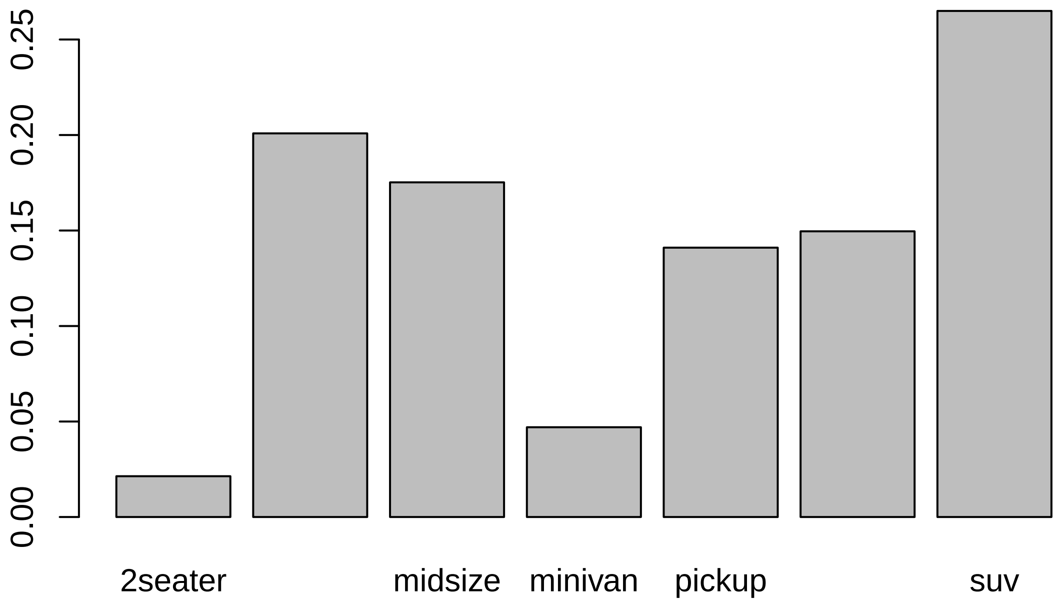
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  234
##
##
##            |    2seater |    compact |    midsize |    minivan |     pickup |
##            |------------|------------|------------|------------|------------|
##            |          5 |         47 |         41 |         11 |         33 |
##            |      0.021 |      0.201 |      0.175 |      0.047 |      0.141 |
##            |------------|------------|------------|------------|------------|
##
##
##            | subcompact |        suv |
##            |------------|------------|
##            |         35 |         62 |
##            |      0.150 |      0.265 |
##            |------------|------------|
##
##
##
##
```

- Graphic summaries

```
barplot(table(mpg$class))
```

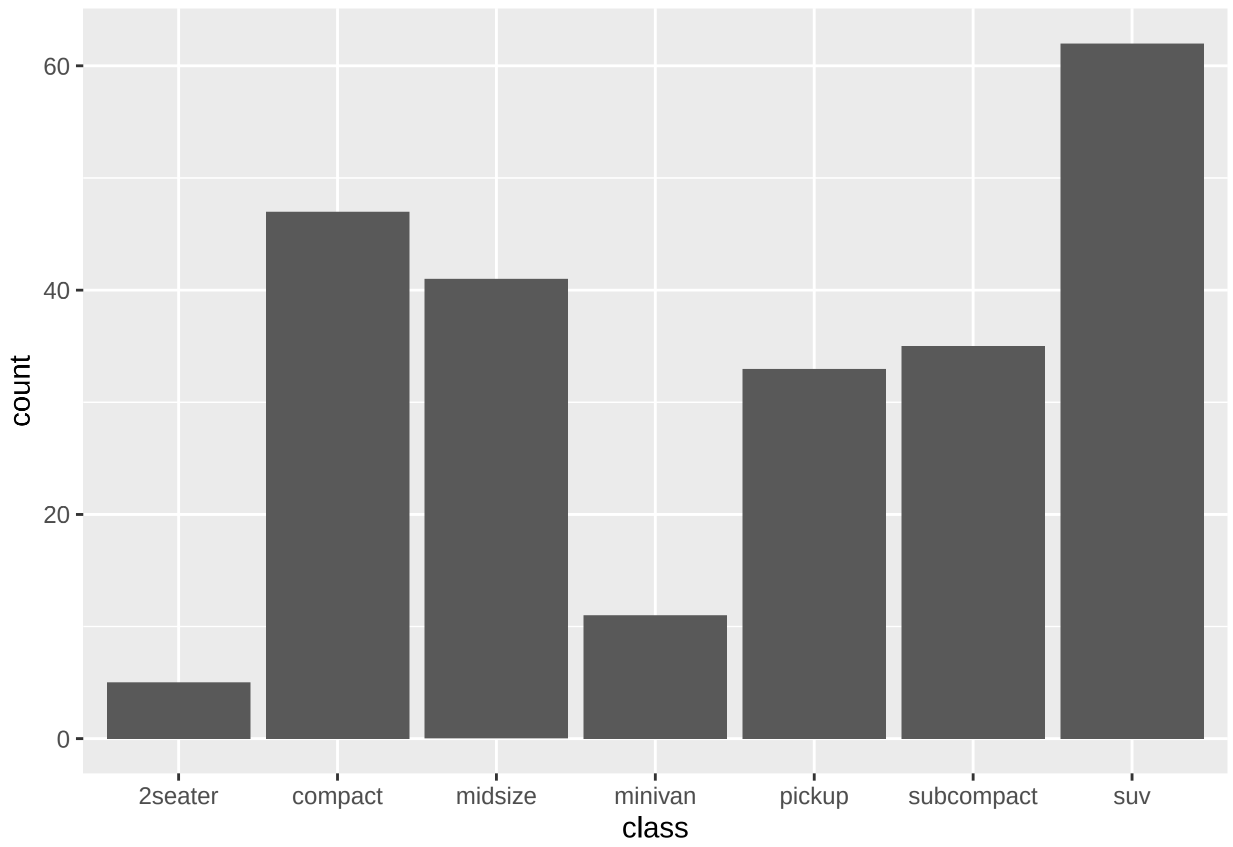


```
barplot(prop.table(table(mpg$class)))
```

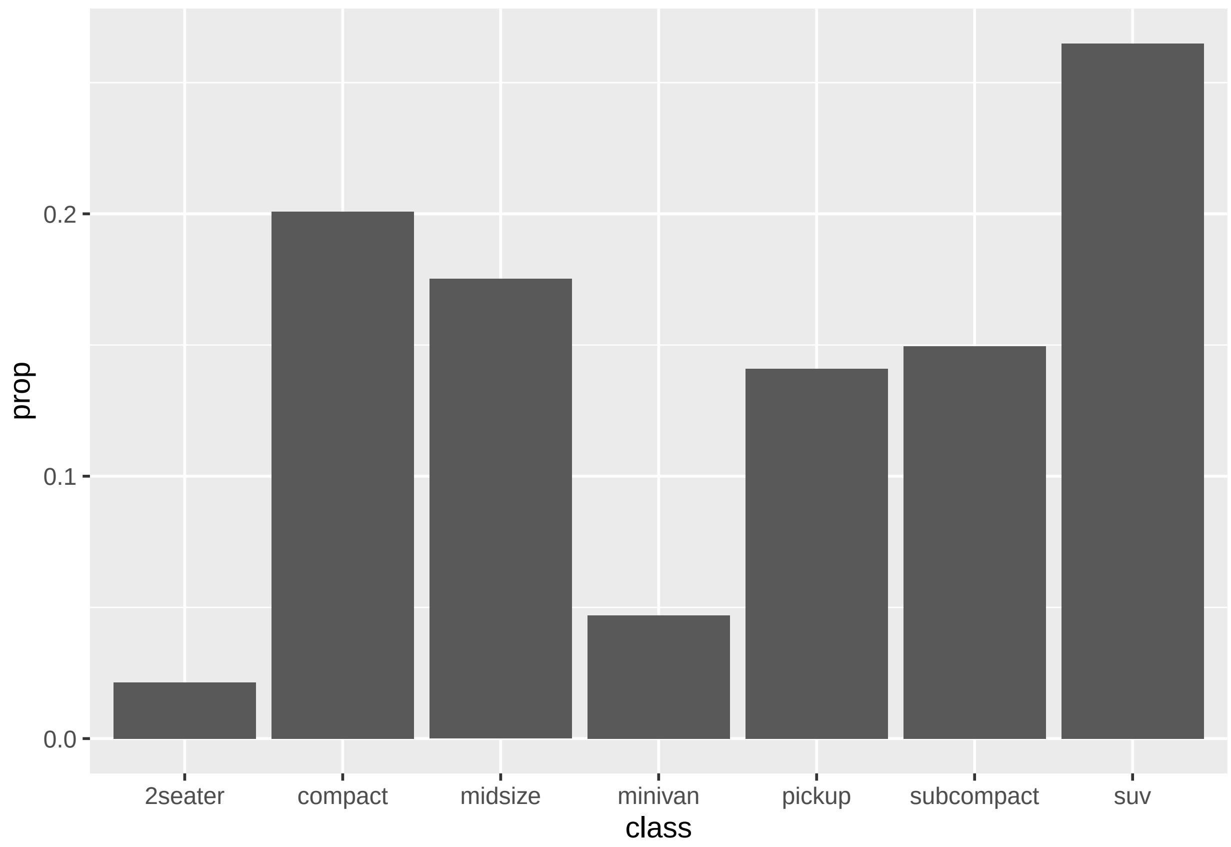


With ggplot2

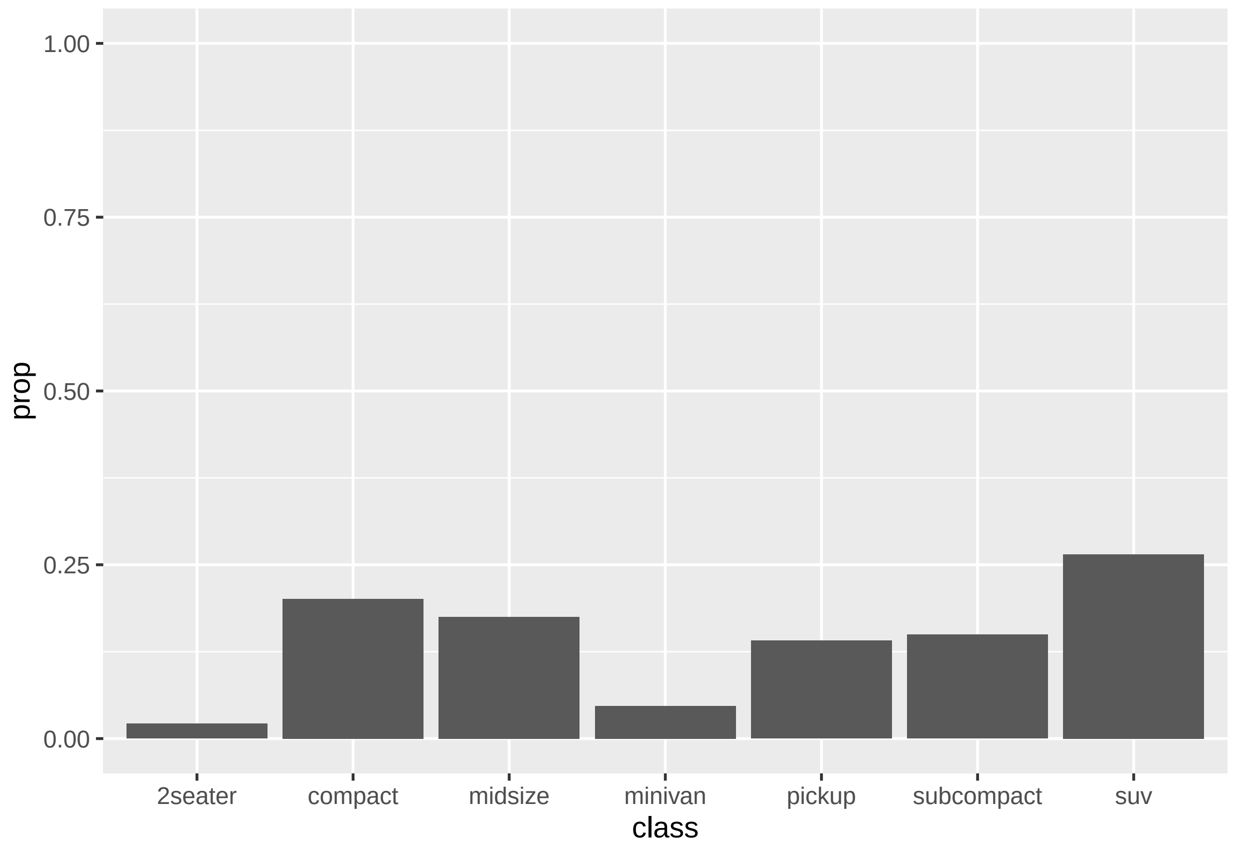
```
ggplot(mpg, aes(class)) +
  geom_bar()
```



```
#for relative frequency
ggplot(mpg, aes(class)) +
  geom_bar(aes(y=..prop.., group=1))
```



```
ggplot(mpg, aes(class)) +
  geom_bar(aes(y=..prop.., group=1)) +
  scale_y_continuous(limits=c(0,1))
```



Bivariate Analysis

```
#load the data
osteoporosis <- read.csv2("osteoporosis.csv", 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
```

```
#overview of data
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" "NATU
##  $ nivel_ed: chr  "SECUNDARIOS" "SECUNDARIOS" "PRIMARIOS" "PRIMARIOS" ...
```

Qualitative versus qualitative

- Numeric bivariate analysis

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

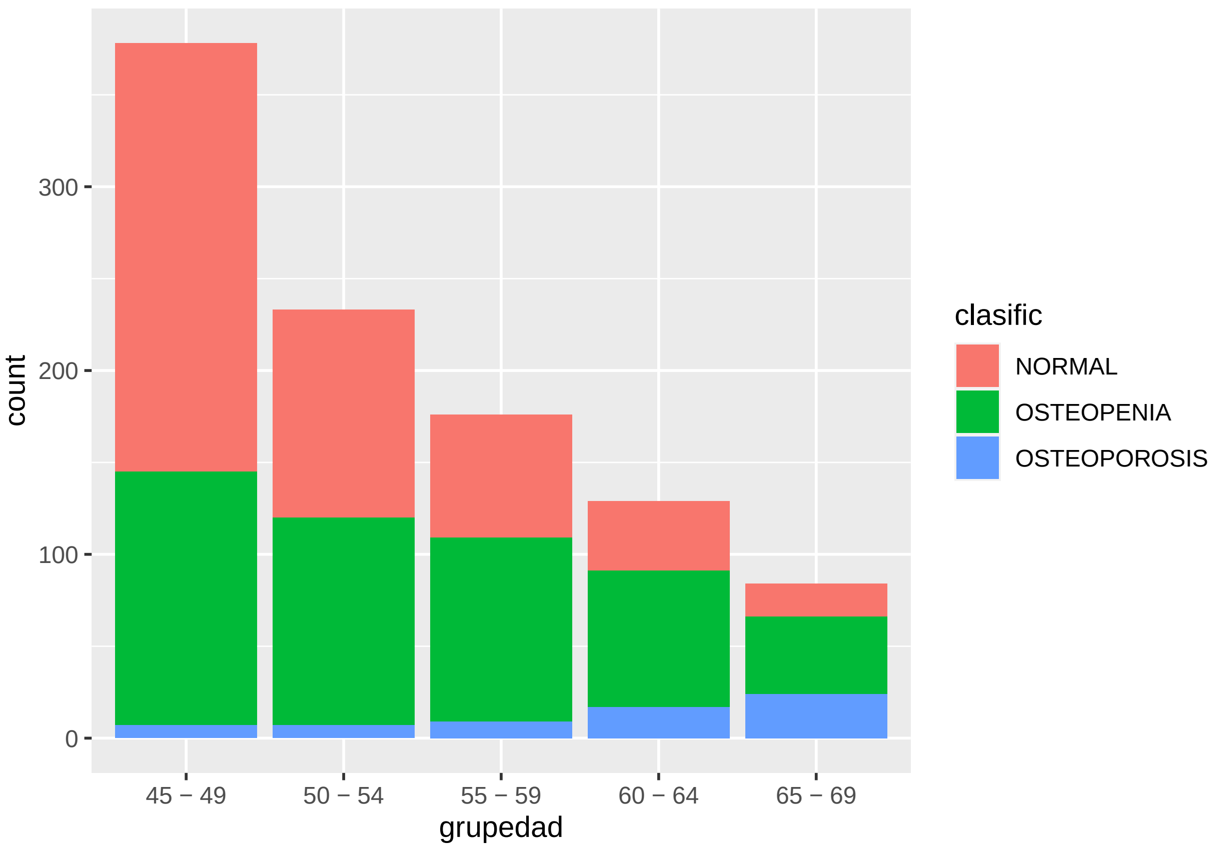
```
##
##             NORMAL OSTEOPENIA OSTEOPOROSIS
##   45 - 49    233        138            7
##   50 - 54    113        113            7
##   55 - 59     67        100            9
##   60 - 64     38         74           17
##   65 - 69     18         42           24
```

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

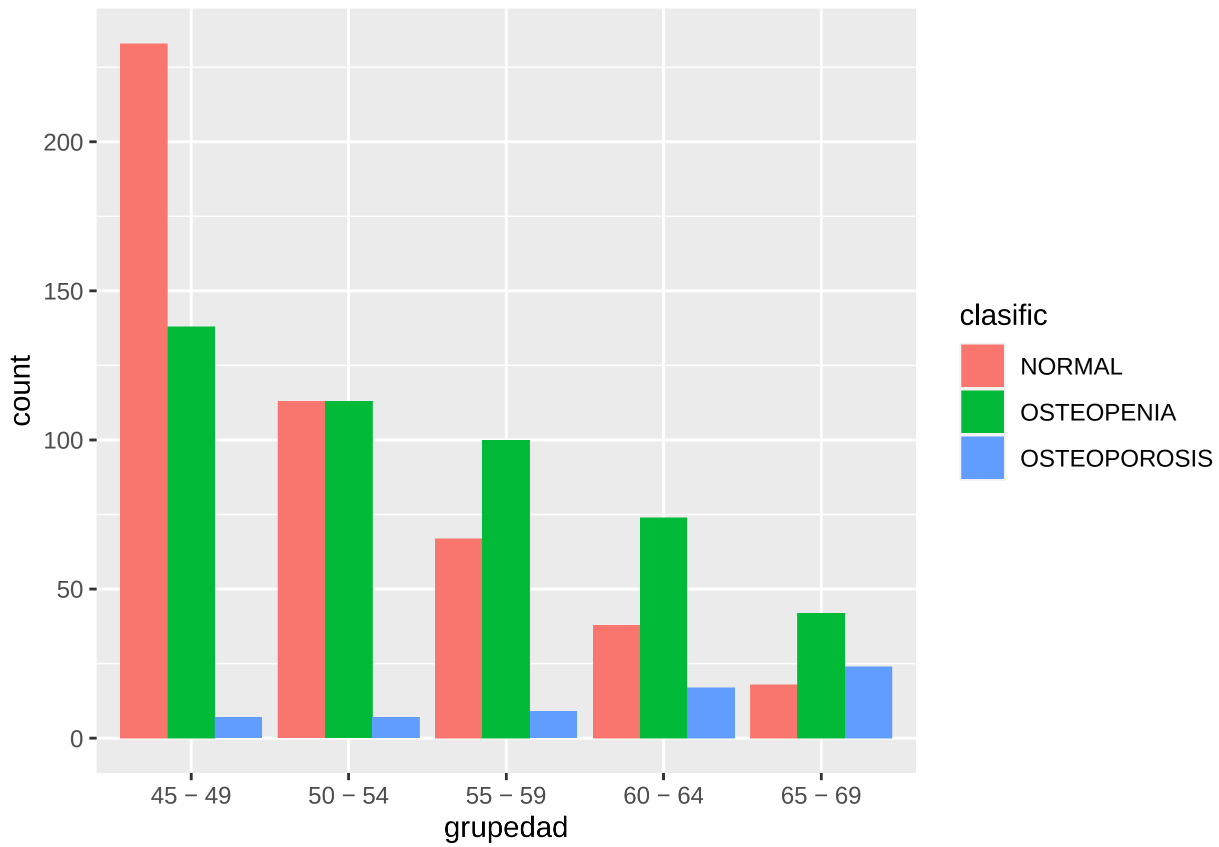
```
##
##             NORMAL OSTEOPENIA OSTEOPOROSIS
##   45 - 49   0.233      0.138        0.007
##   50 - 54   0.113      0.113        0.007
##   55 - 59   0.067      0.100        0.009
##   60 - 64   0.038      0.074        0.017
##   65 - 69   0.018      0.042        0.024
```

- Graphic analysis

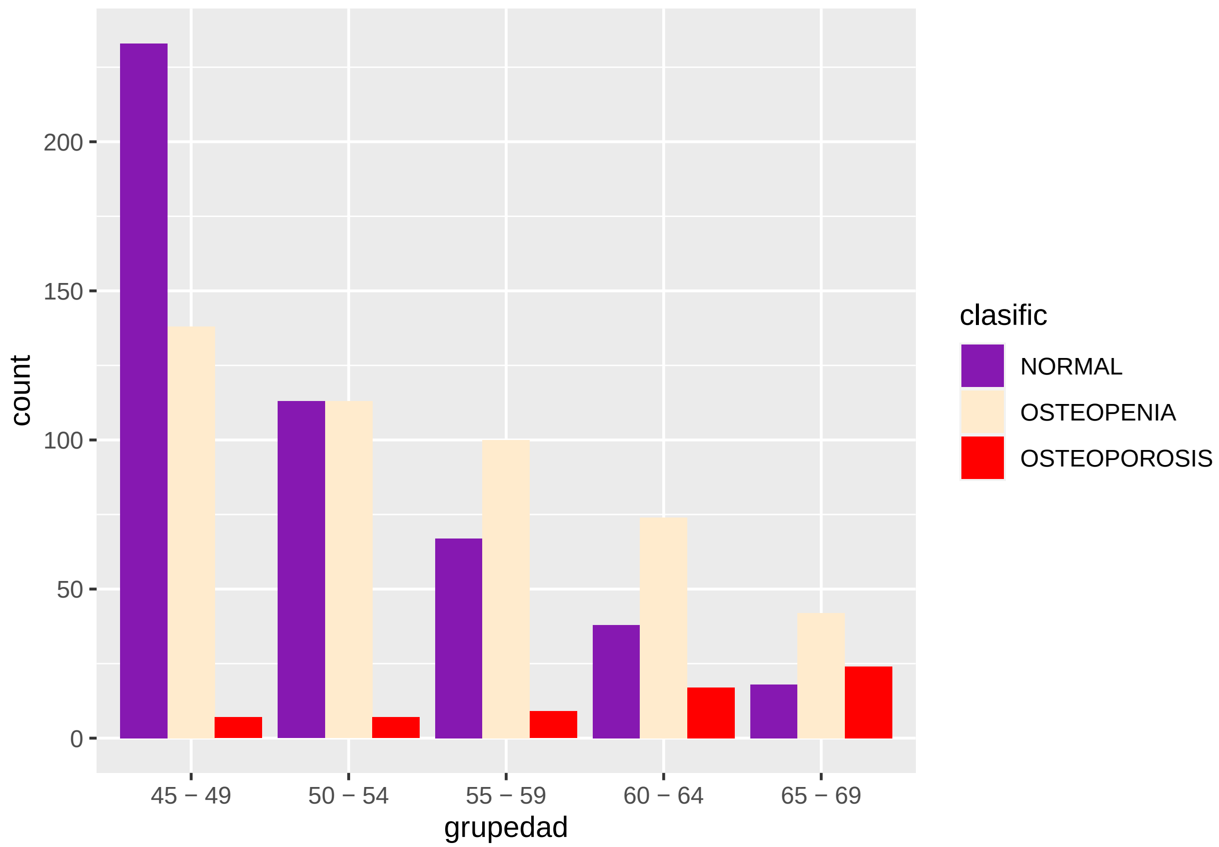
```
#plot the data: stacked barplot
ggplot(data = osteoporosis, aes(x = grupedad)) +
  geom_bar(aes(fill = clasific))
```



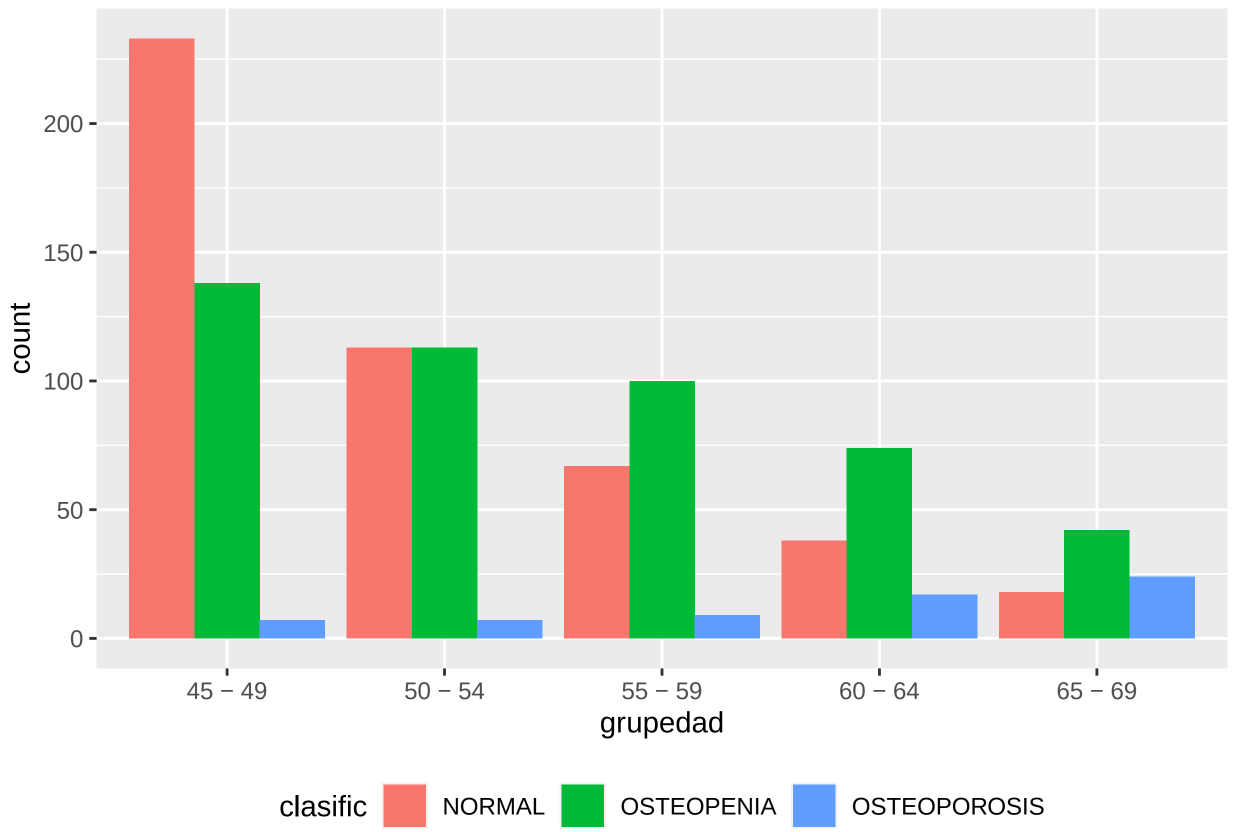
```
#plot the data II: bars side by side
ggplot(data = osteoporosis, aes(x = grupedad)) +
  geom_bar(aes(fill = clasific), position = "dodge")
```



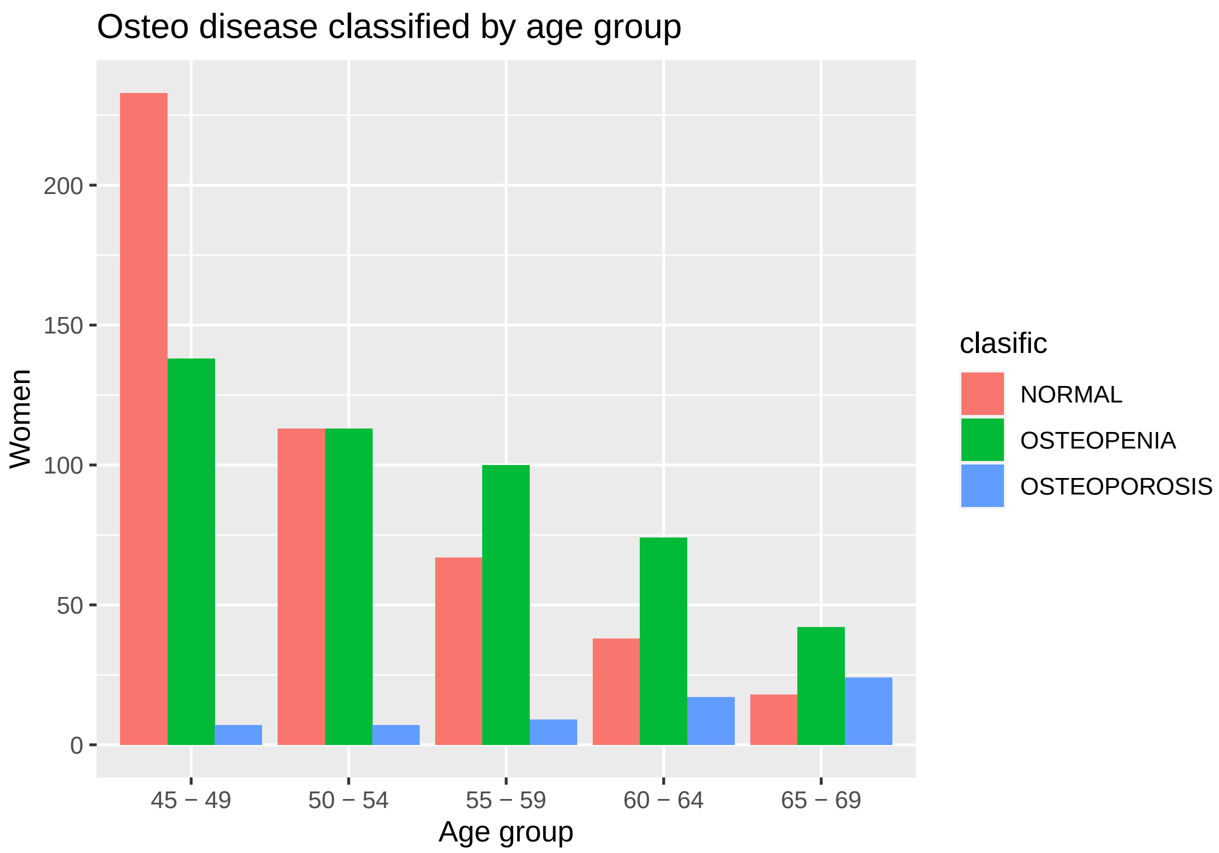
```
#Change colors, legend position, labels and finally save it!
p <- ggplot(data = osteoporosis, aes(x = grupedad)) +
  geom_bar(aes(fill = clasific), position = "dodge")
p + scale_fill_manual(values=c("#8618b1", "blanchedalmond", "red"))
```



```r
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
```

Qualitative versus quantitative

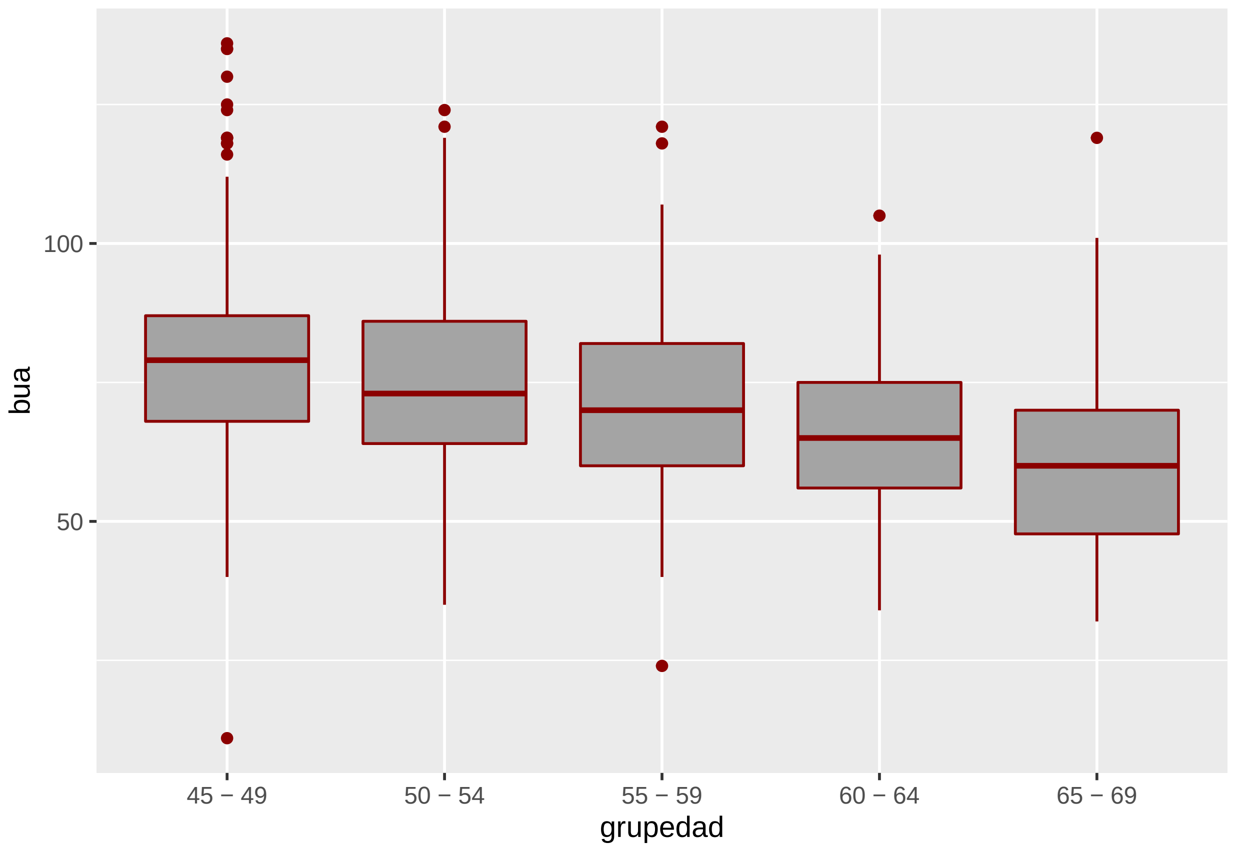
- Numeric analysis

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

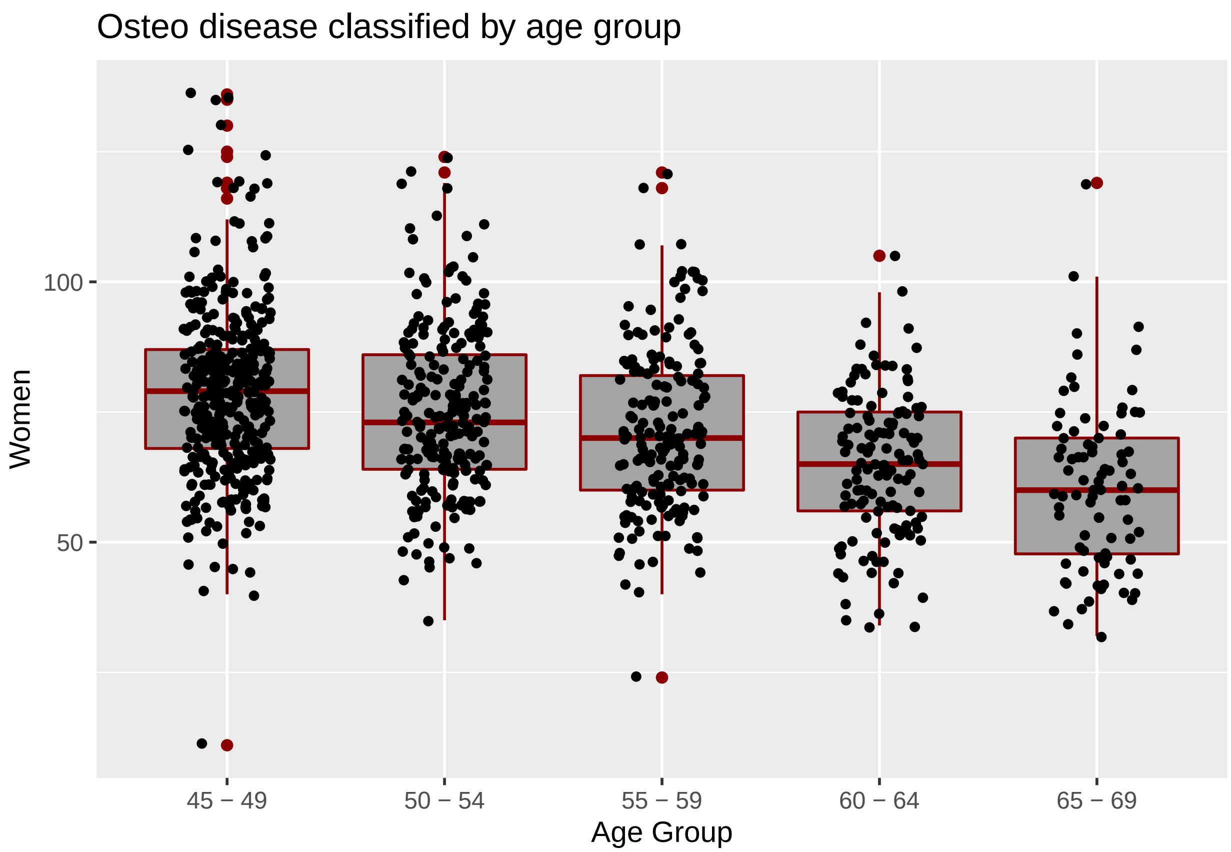
```
##  45 - 49  50 - 54  55 - 59  60 - 64  65 - 69
## 78.75926 75.05150 71.43182 64.89147 60.66667
```

- Graphic analysis

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



```
# 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")
```



Exercise solution

Study the relationship between menop and group of illness (clasific)

```
#explore variables
head(osteoporosis[,c("menop", "clasific")])
```

```
##   menop   clasific
## 1    NO OSTEOPENIA
## 2    NO OSTEOPENIA
## 3    NO     NORMAL
## 4    SI OSTEOPENIA
## 5    NO     NORMAL
## 6    NO     NORMAL
```

```
str(osteoporosis[,c("menop", "clasific")])
```

```
## 'data.frame':    1000 obs. of  2 variables:
##  $ menop   : chr  "NO" "NO" "NO" "SI" ...
##  $ clasific: chr  "OSTEOPENIA" "OSTEOPENIA" "NORMAL" "OSTEOPENIA" ...
```

```
#Numeric summaries for two categorical variables: contigency table
table(osteoporosis$menop, osteoporosis$clasific)
```

```
##
##      NORMAL OSTEOPENIA OSTEOPOROSIS
##   NO    189        108            6
##   SI    280        359           58
```

```
addmargins(table(osteoporosis$menop, osteoporosis$clasific))
```

```
##
##       NORMAL OSTEOPENIA OSTEOPOROSIS  Sum
##   NO     189        108            6  303
##   SI     280        359           58  697
##   Sum    469        467           64 1000
```

```
### proportions with respect to total
prop.table(table(osteoporosis$menop, osteoporosis$clasific))
```

```
##
##      NORMAL OSTEOPENIA OSTEOPOROSIS
##   NO  0.189      0.108        0.006
##   SI  0.280      0.359        0.058
```

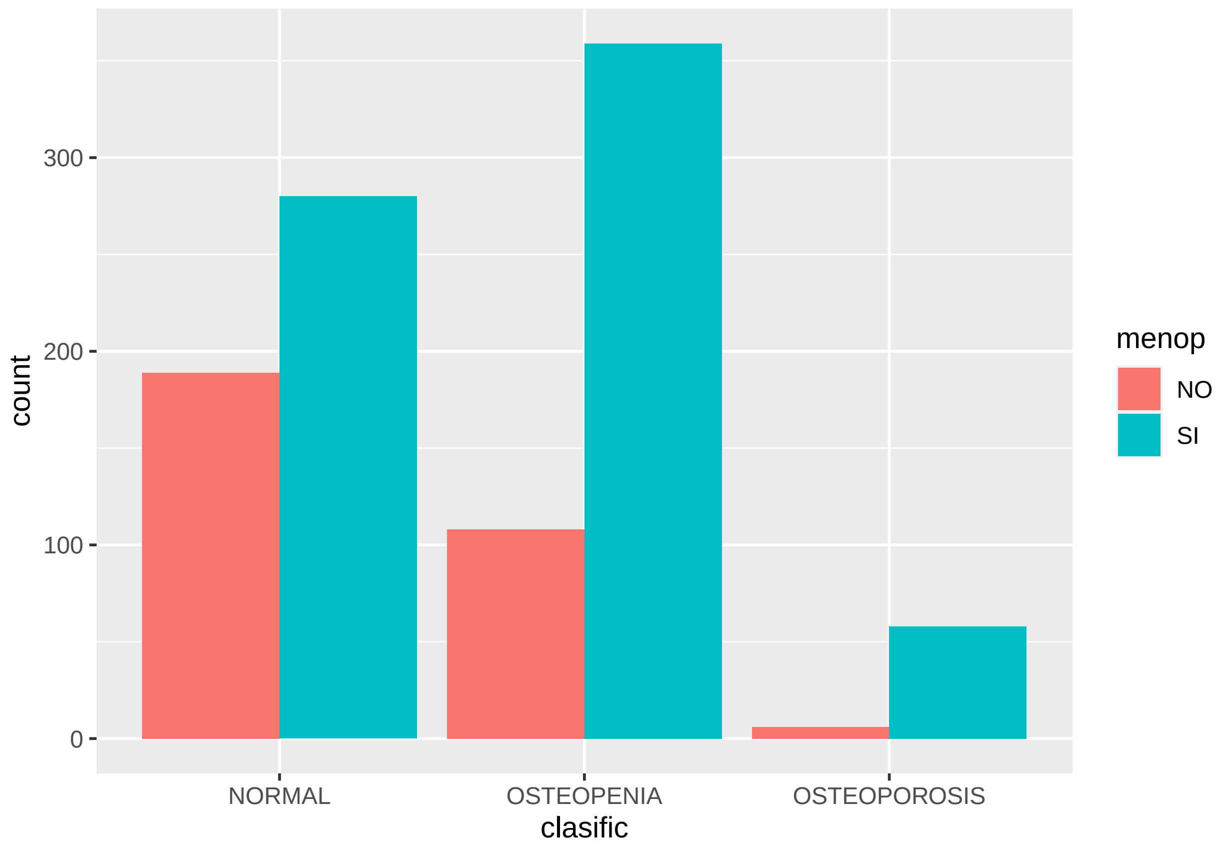
```
###relative frequencies with respect to rows
prop.table(table(osteoporosis$menop, osteoporosis$clasific), margin=1)
```

```
##
##          NORMAL OSTEOPENIA OSTEOPOROSIS
##   NO 0.62376238 0.35643564   0.01980198
##   SI 0.40172166 0.51506456   0.08321377
```

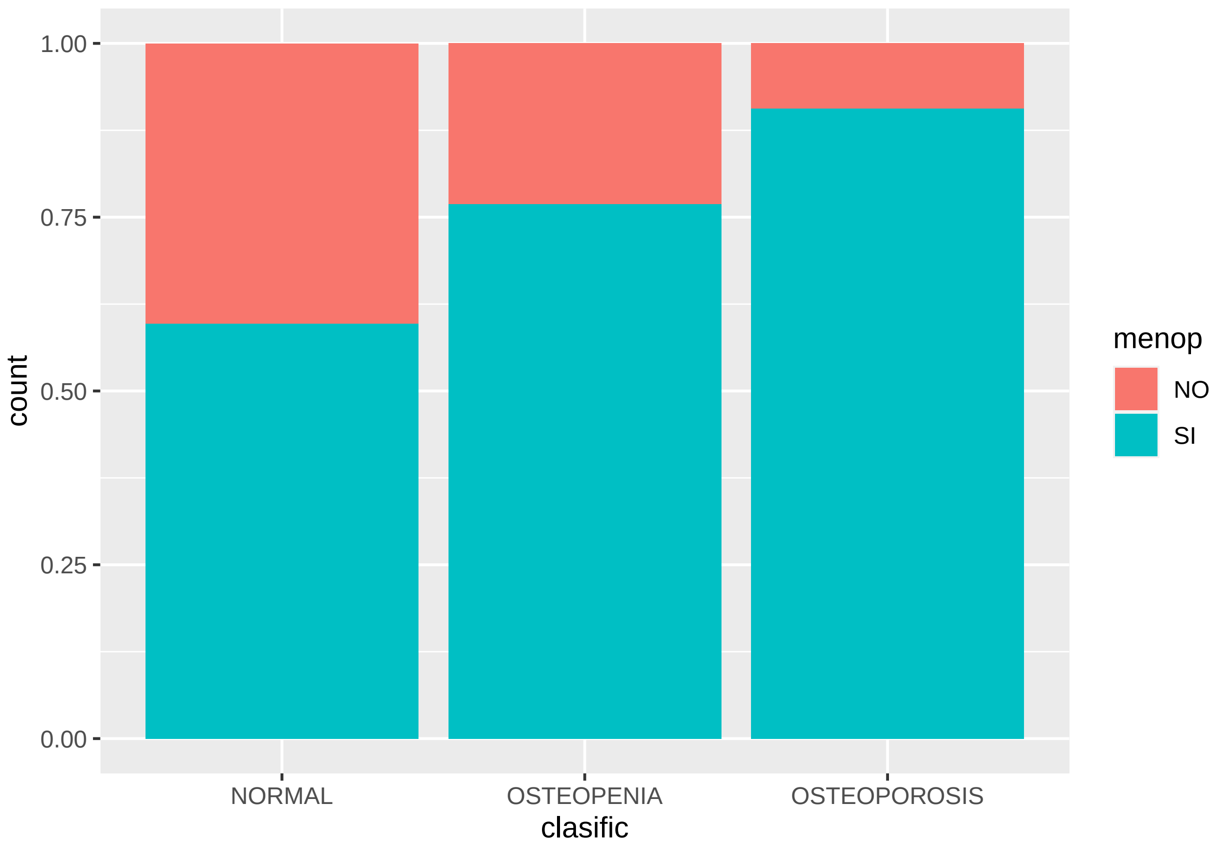
```
###relative frequencies with respect to columns
prop.table(table(osteoporosis$menop, osteoporosis$clasific), margin=2)
```

```
##
##         NORMAL OSTEOPENIA OSTEOPOROSIS
##   NO 0.4029851  0.2312634    0.0937500
##   SI 0.5970149  0.7687366    0.9062500
```

```
#Graphic summaries for two categorical variables: barplot
ggplot(data = osteoporosis, aes(x = clasific)) +
  geom_bar(aes(fill = menop), position = "dodge")
```



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



Study if peso is different in each group of illness (clasific).

```
#explore variables
head(osteoporosis[,c("peso", "clasific")])
```

```
##   peso   clasific
## 1 70.0 OSTEOPENIA
## 2 53.0 OSTEOPENIA
## 3 64.0     NORMAL
## 4 78.0 OSTEOPENIA
## 5 56.0     NORMAL
## 6 63.5     NORMAL
```

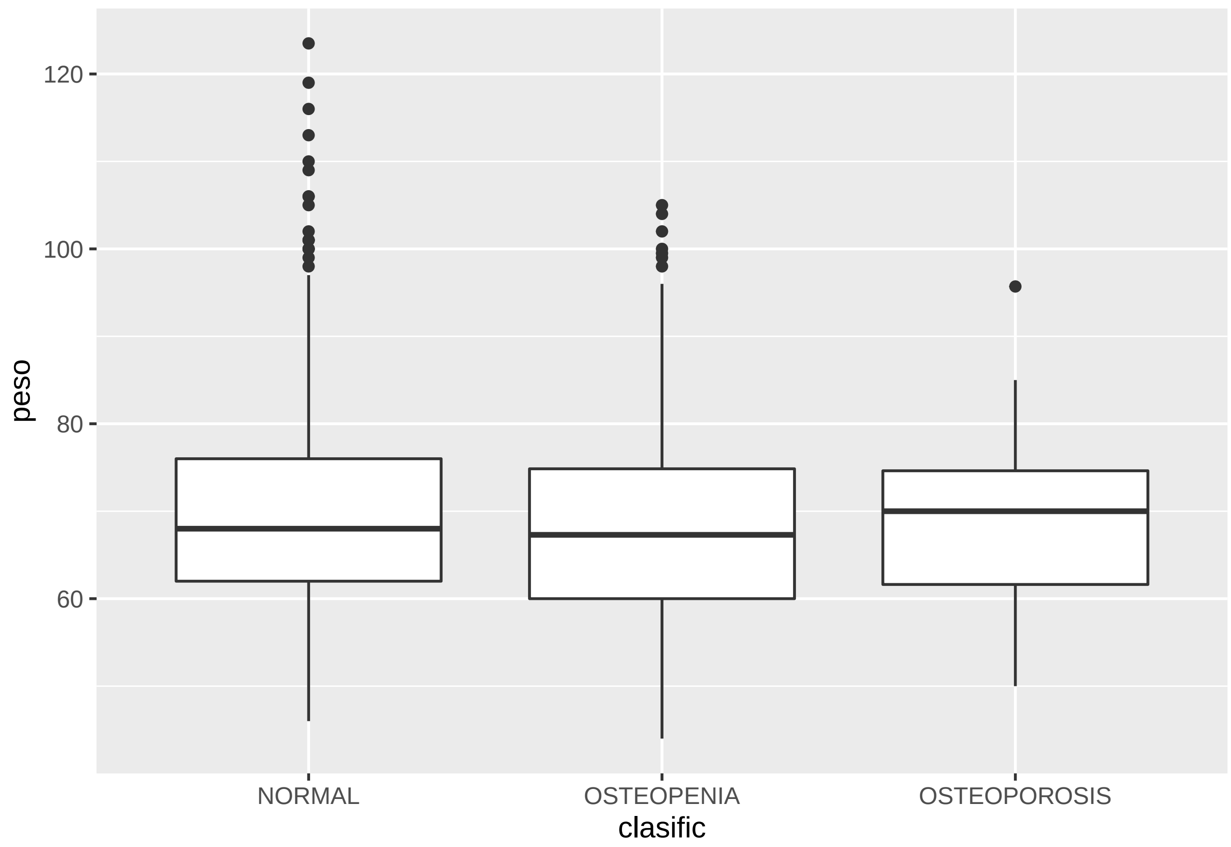
```
str(osteoporosis[,c("peso", "clasific")])
```

```
## 'data.frame':    1000 obs. of  2 variables:
##  $ peso    : num  70 53 64 78 56 63.5 86 61.5 60.5 64 ...
##  $ clasific: chr  "OSTEOPENIA" "OSTEOPENIA" "NORMAL" "OSTEOPENIA" ...
```

```
#Numeric summaries for one categorical and one continuous variables: table of statistics
with(osteoporosis, tapply(peso, list(clasific), mean, na.rm=TRUE))
```

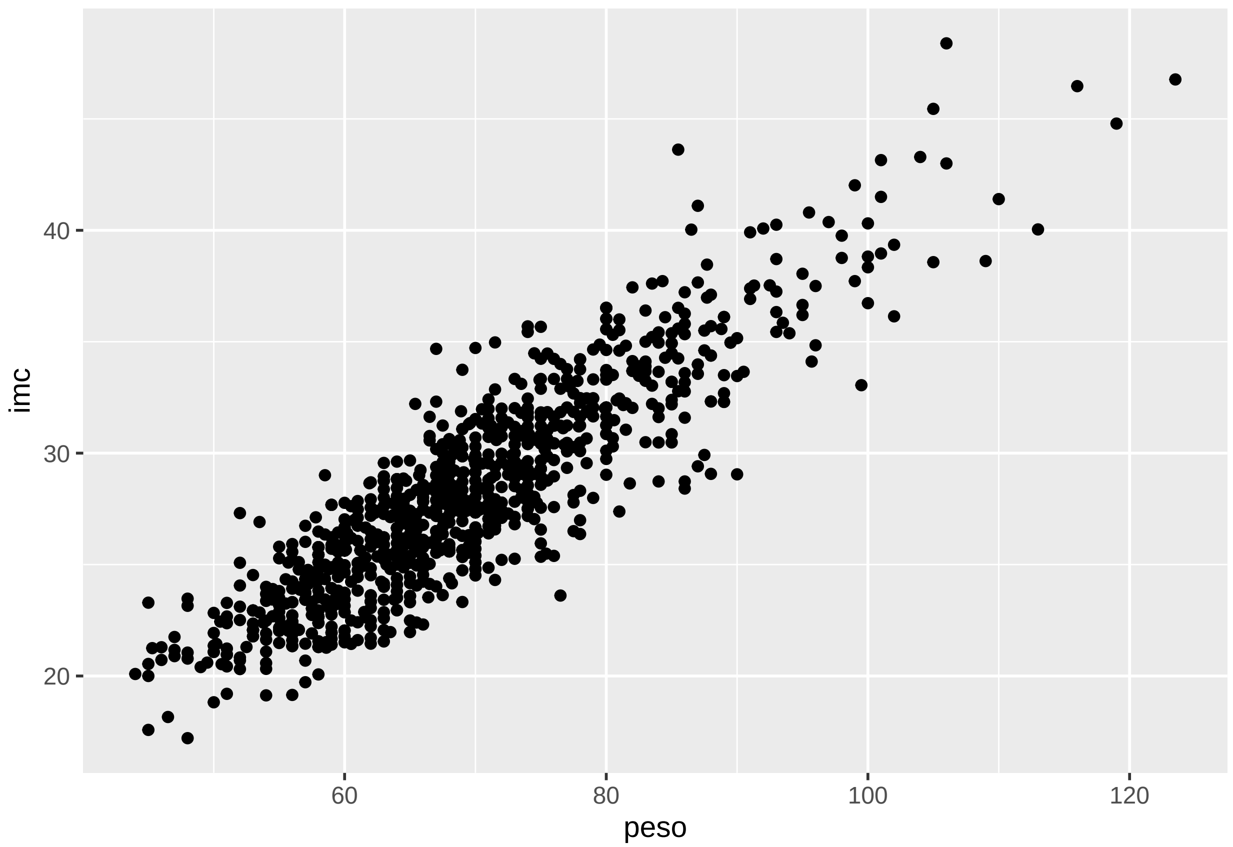
```
##       NORMAL   OSTEOPENIA OSTEOPOROSIS
##     70.33284     68.03041     68.22656
```

```
#Graphic summaries for one categorical and one continuous variables: grouped boxplot
ggplot(osteoporosis, aes(x = clasific, y = peso)) +
  geom_boxplot()
```

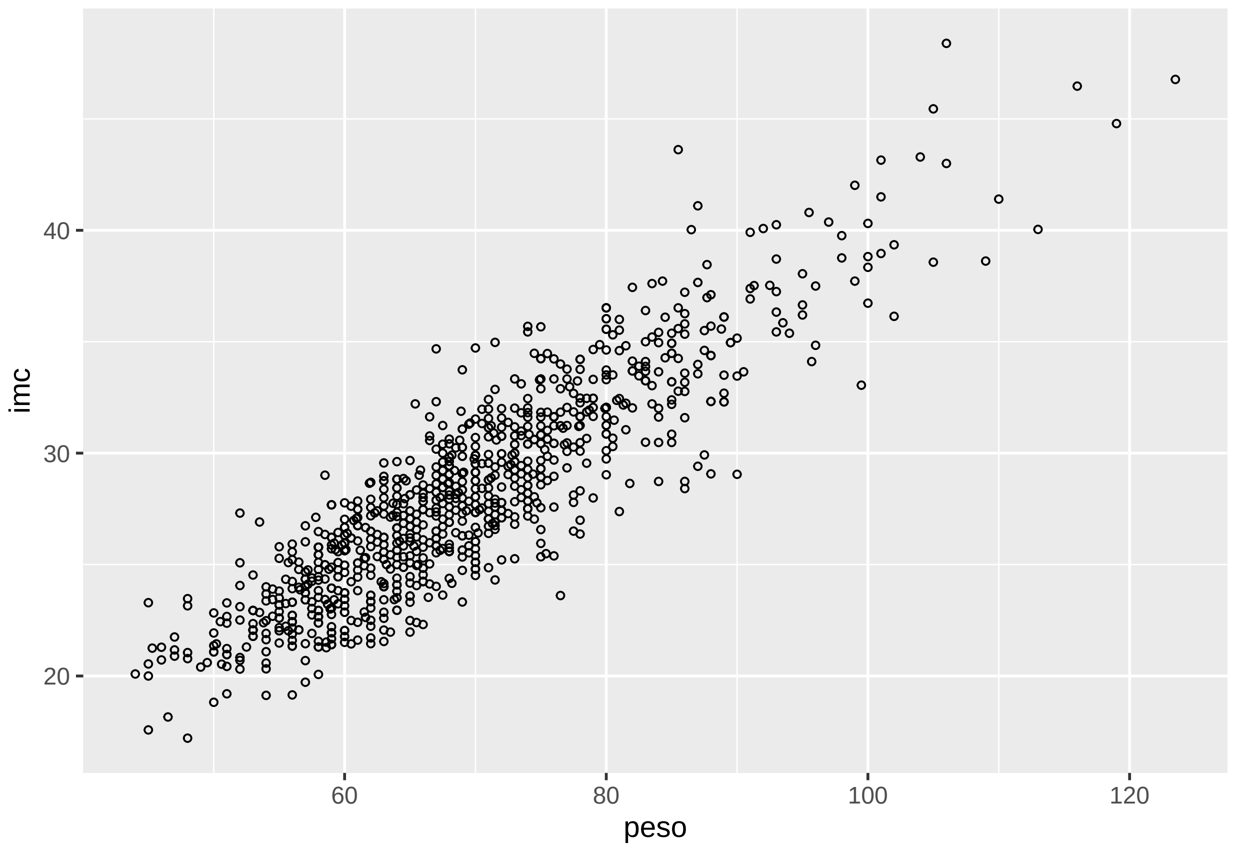


Quantitative versus quantitative

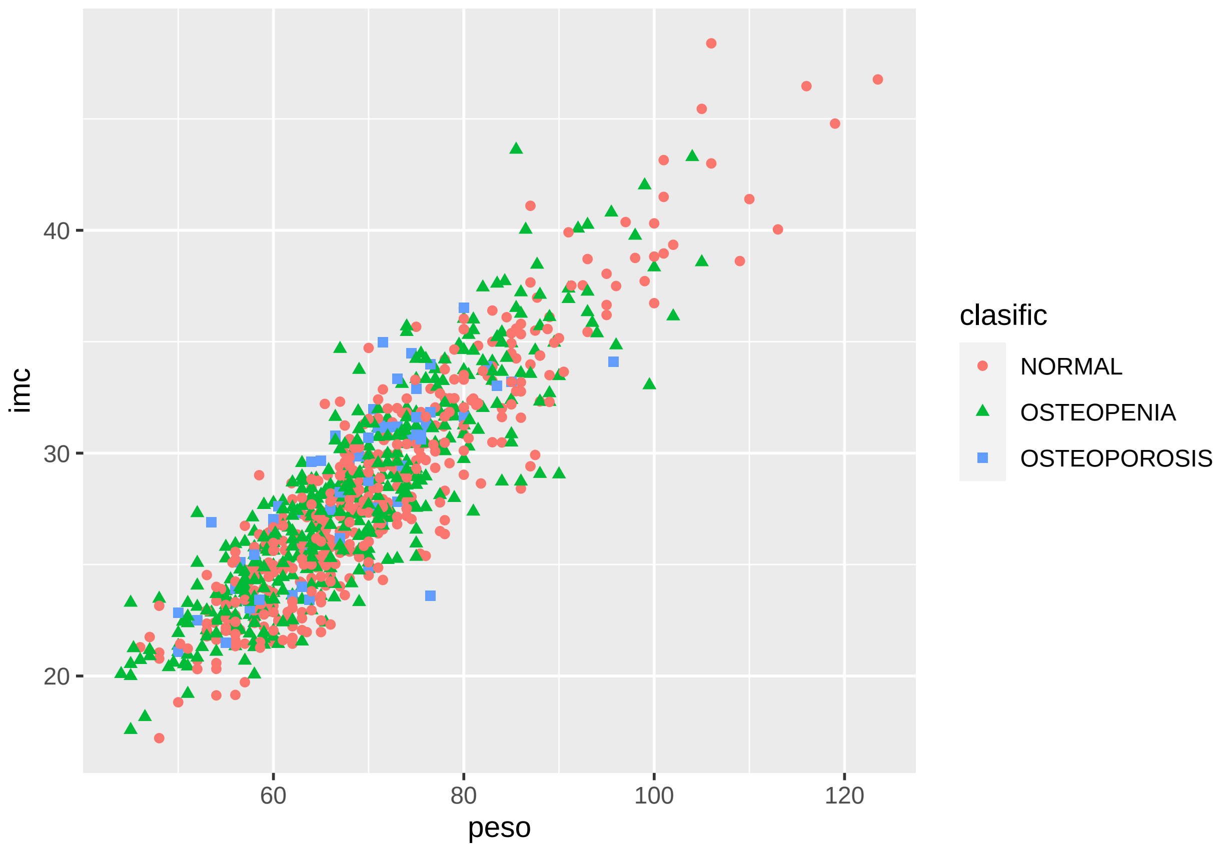
```
# Basic scatter plot
ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point()
```



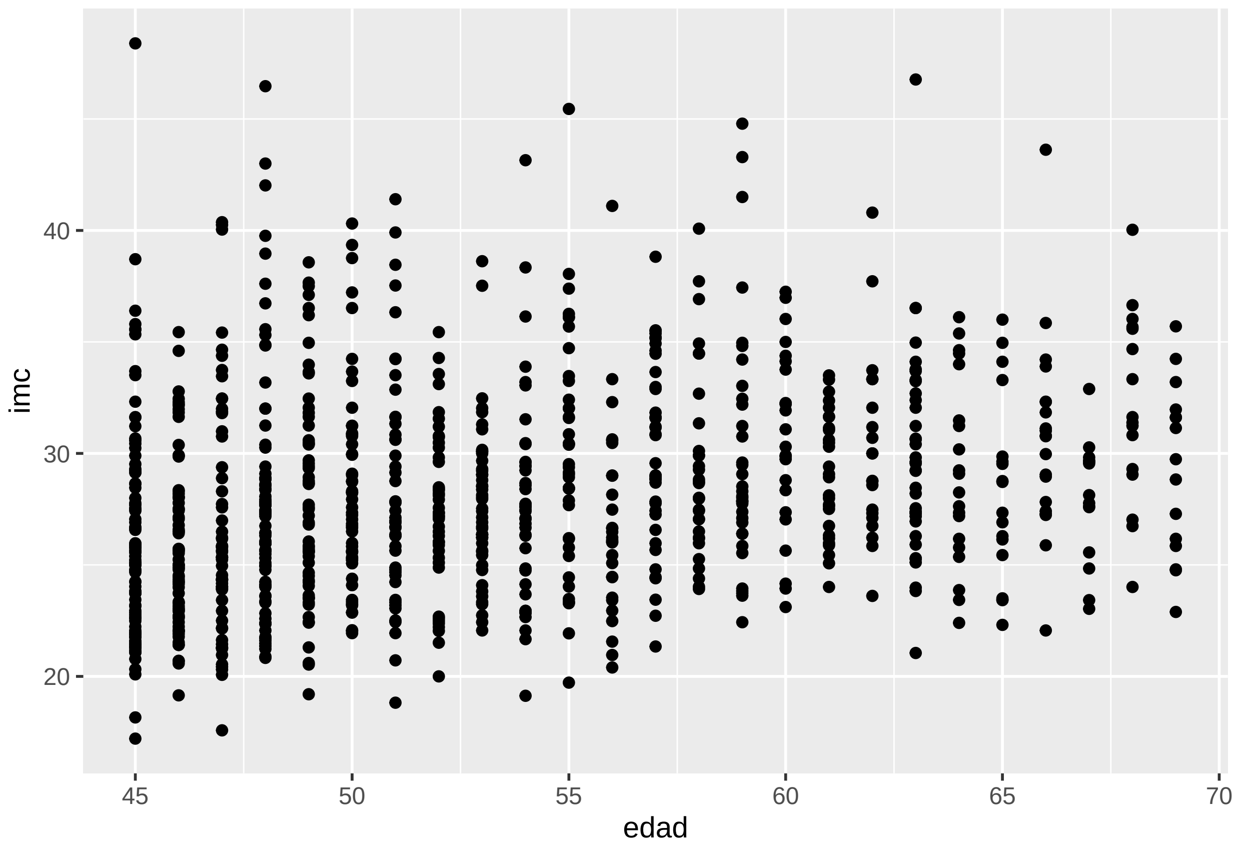
```r
# Change the point size, and shape
ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point(size = 1, shape = 1)
```



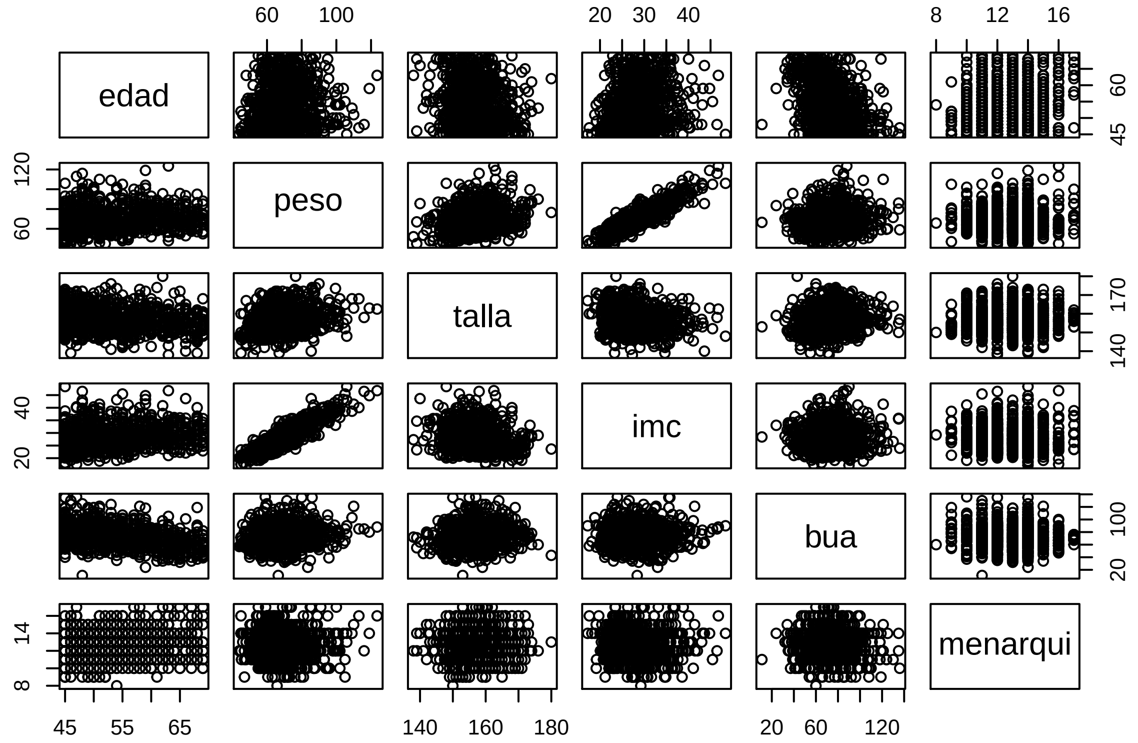
```r
# Color the points depending of another variable
ggplot(osteoporosis, aes(x = peso, y = imc, color = clasific, shape = clasific)) +
  geom_point()
```



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



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

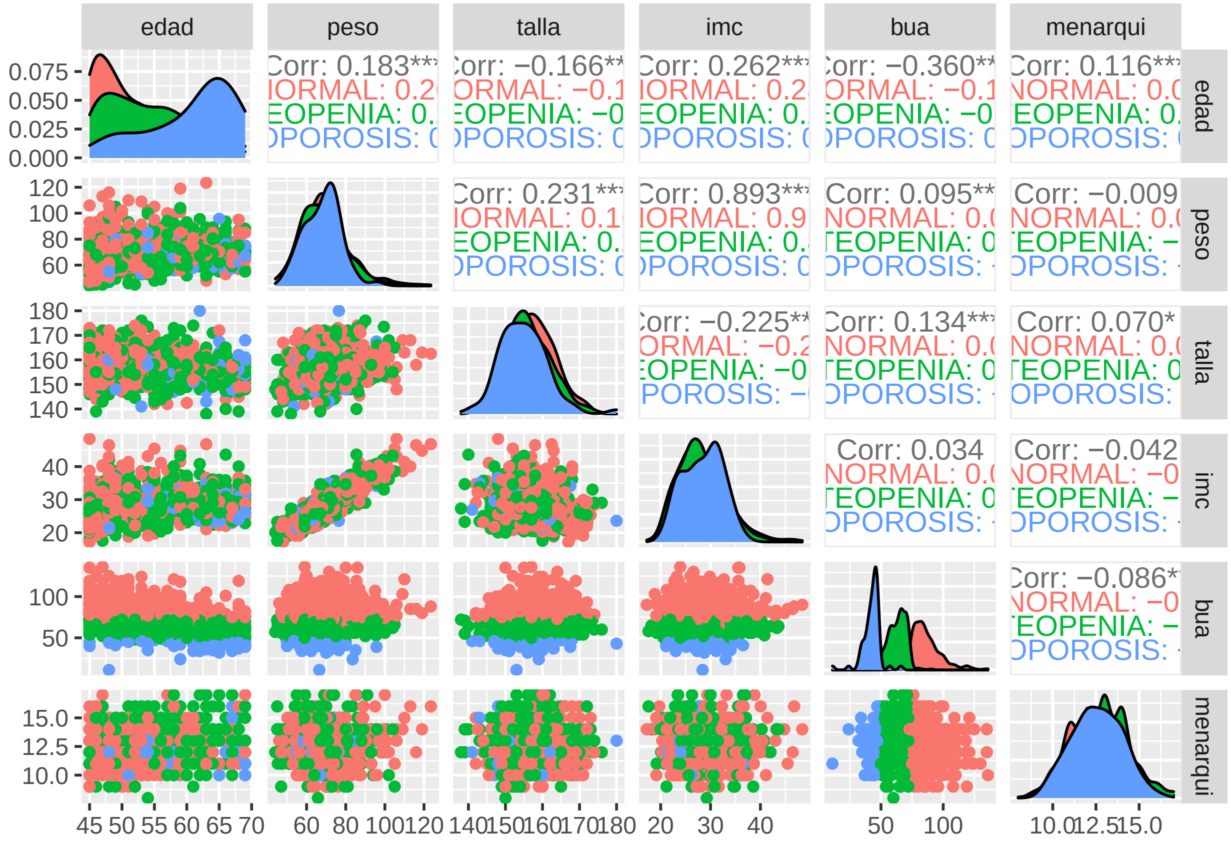


```
#with ggplots
# install.packages("GGally")
library(GGally)
```

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

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

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



Correlation

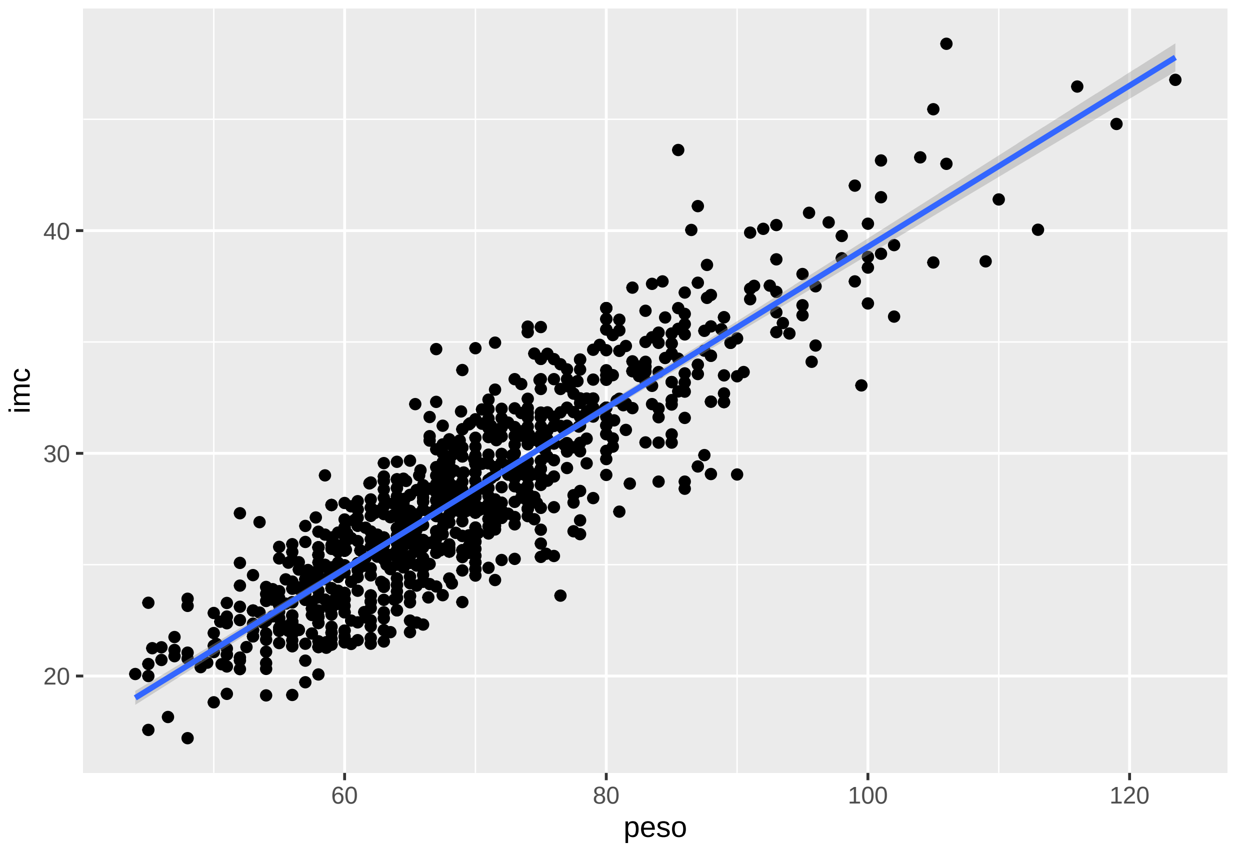
Pearson correlation between imc and peso

```
#Pearson correlation
cor(osteoporosis$imc, osteoporosis$peso, method = "pearson")
```

```
## [1] 0.8927863
```

```
#the plot
ggplot(osteoporosis, aes(x = peso, y = imc)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

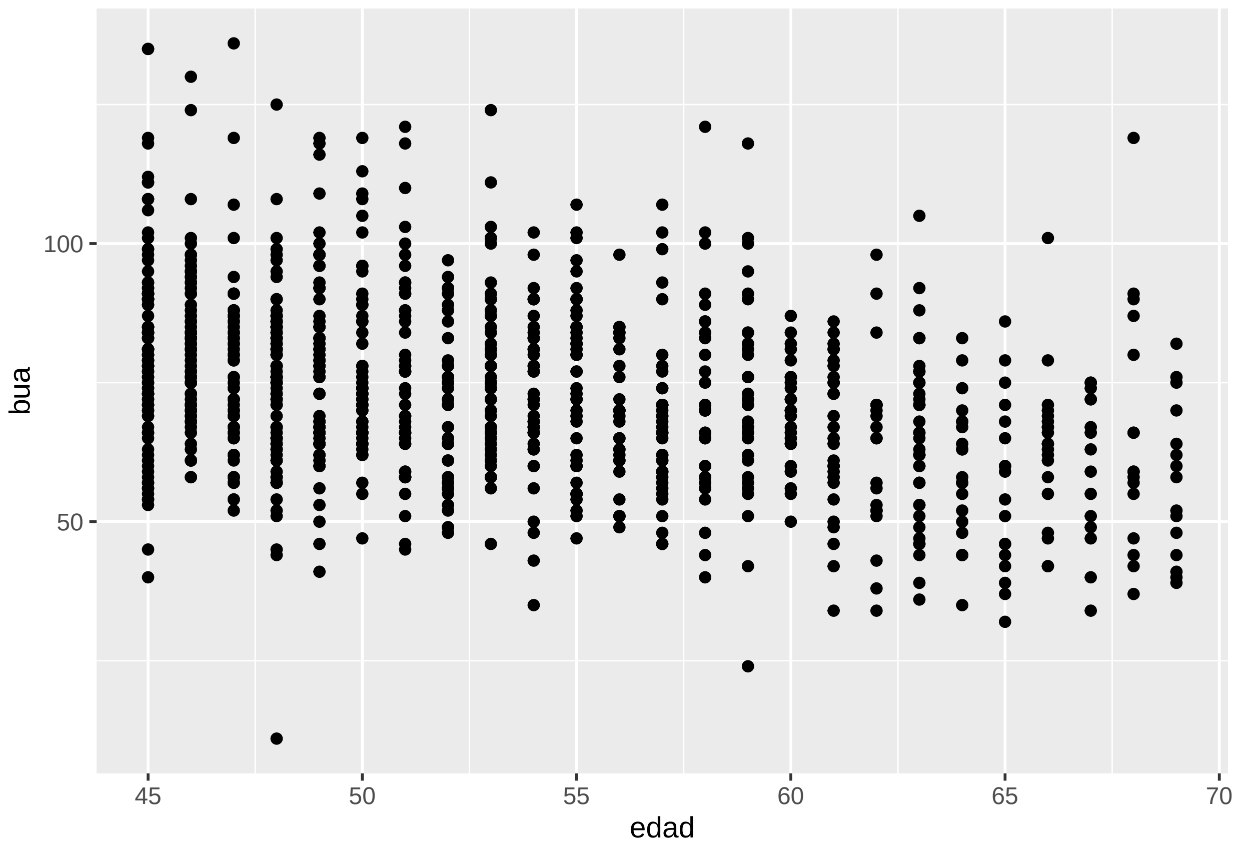


Pearson correlation between bua and edad

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

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

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



Spearman correlation between bua and edad

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

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