Lab 3: Exploratory data analysis

Benjamin Lacroix

February 11, 2020

1 Introduction

R has two main graphics packages:

- graphics: this is the standard R package. It is usually possible to produce quick and dirty plots with a single short command, which is useful for EDA. Its also possible to refine the figures with a bit more effort.
- ggplot2: based on the ideas of 'Grammar of Graphics'. The idea is to map data dimensions to attributes of geometric objects e.g. a variable might be mapped to the height of a bar. It is usual to add objects to an initial object created by calling ggplot.

There are a number of other packages with specialised plotting functions, for example the correlation matrix visualisation seen in lectures is produced by a dedicated package.

The first section of this lab is to reproduce what we saw in the lecture on the mtcars data set. The second section will be about producing your own analysis on the dataset 'diamonds' that comes with the ggplot2 library

2 The 'mtcars' dataset

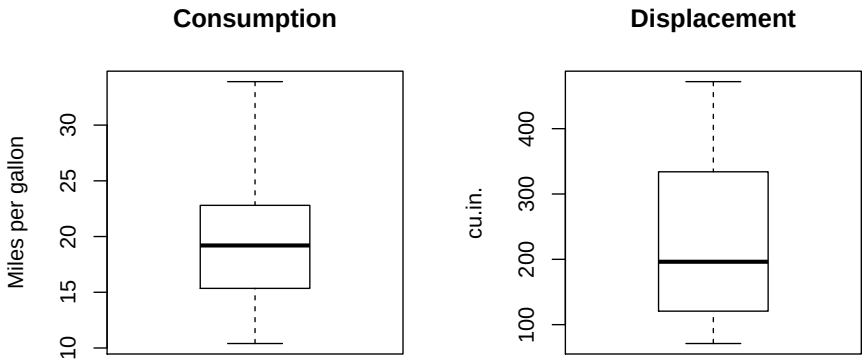
Load the mtcars dataset in R

```
data("mtcars")
?mtcars # to look at the descriptin of the variables
```

2.1 Boxplots

Use the 'graphics' package boxplot() function to do a box plot over all the numeric variables

```
# The following command allows you to group multiple plots.
# Here, we create a grid of 1 row and 2 column
par(mfrow=(c(1,2)))
boxplot(mtcars$mpg, ylab = "Miles per gallon", main = "Consumption")
boxplot(mtcars$disp, ylab = "cu.in.", main = "Displacement")
```



Tasks:

- Add a row with the boxplots for drat and qsec
- Do the same with histograms instead of boxplots

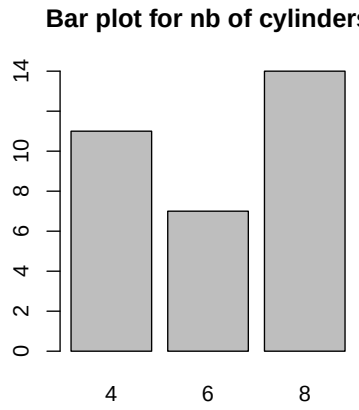
2.2 Tabulating variables

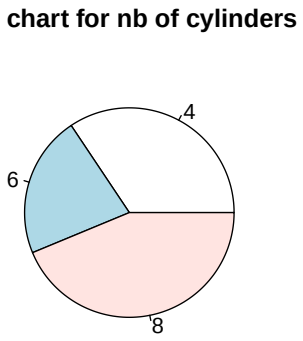
Here we use the 'table' function to count the number of occurences of each balue in a given vector (variable). These statistics can be visualised using barplot and pie charts

```
cyl.table <- table(mtcars$cyl)
cyl.table

##
##  4  6  8
## 11  7 14

par(mfrow=(c(1,2)))
barplot(cyl.table, main = "Bar plot for nb of cylinders")
pie(cyl.table, main = "Pie chart for nb of cylinders")
```



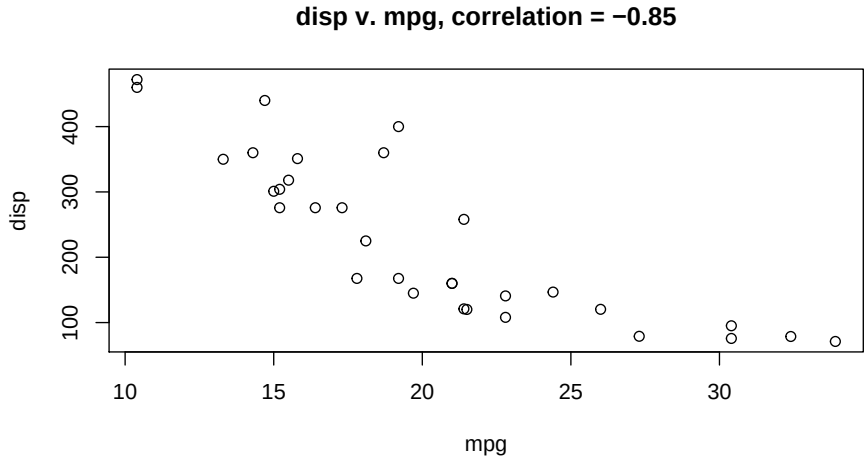


Tasks: Do the same with the 'gear' variable

2.3 Scatter plots

Scatter plots are used to visualise the relationship between two NUMERIC variables. This relationship can be evalauted by the correlation coeffficient using the function cor() in R

```
plot(x=mtcars$mpg, y = mtcars$disp, ylab = "disp", xlab = "mpg",
     main = paste0("disp v. mpg, correlation = ",round(cor(mtcars$mpg,mtcars$disp),2)))
```



Tasks:

- Pick another pair of variables and do the same plot as above. Include as I did in the title the correlation coefficient
- Use the function pairs() to plot the scatterplot of every pair of variables in mtcars

We can also evalaute the correlation of each pair of variables using the function cor() over the whole dataset

```
cor(mtcars)
```

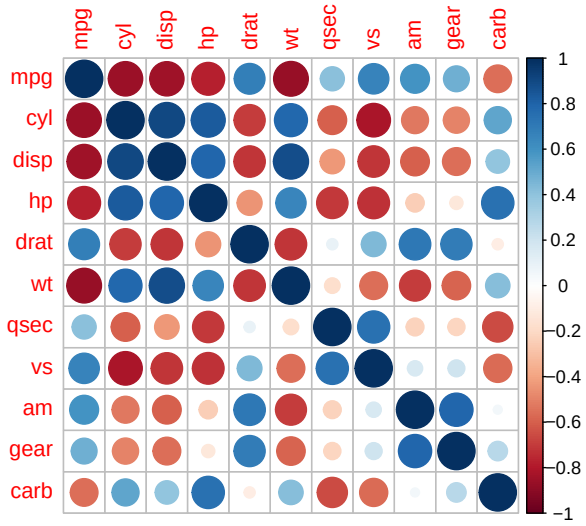
```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

and visualise these correlation using the function corrplot() from the library 'corrplot'

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

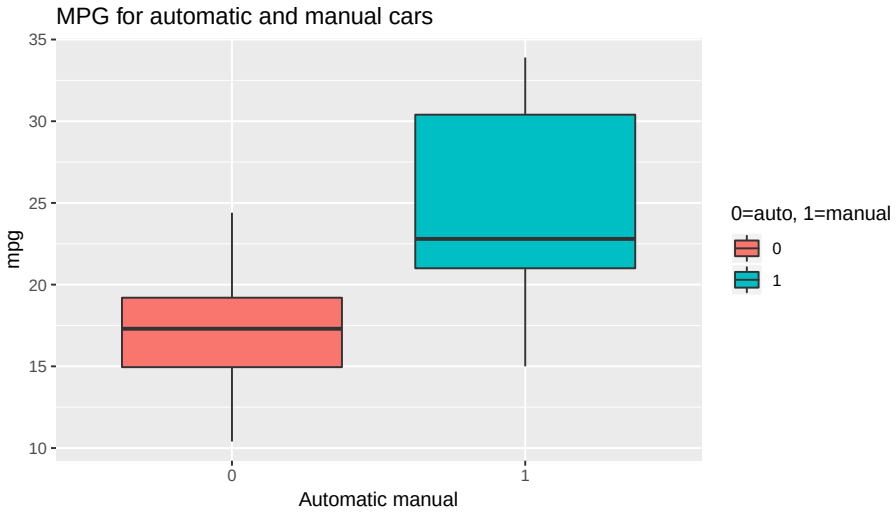
```
corrplot(cor(mtcars))
```



2.4 Bivariate analysis for categorical and numerical variables

As we saw in the lecture, we can visualise the relation between a categorical and a numerical variable using boxplots:

```
library(ggplot2)
ggplot(mtcars, aes(factor(am),mpg, fill = factor(am))) + geom_boxplot() +
  ggtitle("MPG for automatic and manual cars") + xlab("Automatic manual") +
  labs( fill = "0=auto, 1=manual")
```



We can see here that the consumption is much higher on manual cars than in automatic cars.

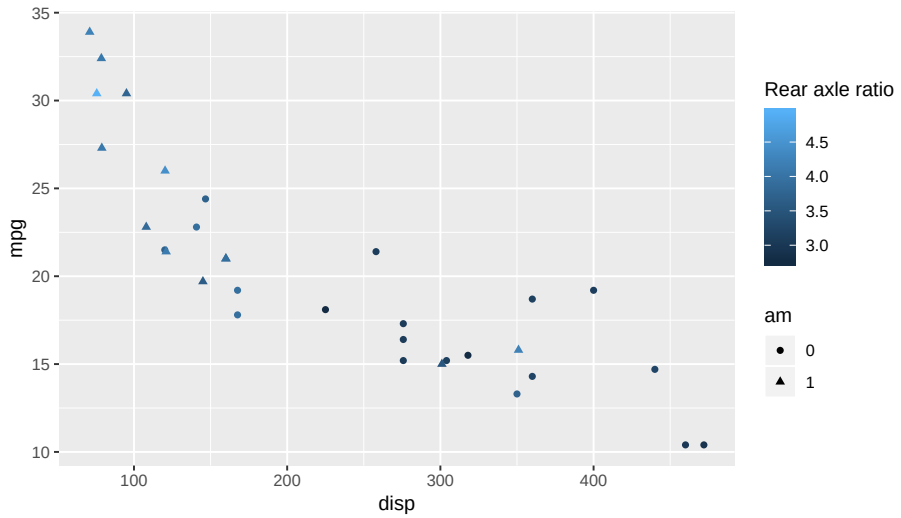
Tasks: Create the boxplots of the car consumption (mpg) grouped by the number of cylinders

2.5 Multivariate visualisations

ggplot allows us to add colours and shape to our scatter plots in a fairly straight forward way. (You can do it with 'graphics' and the function plot(), but doesn't look as good)

This will be quite usefull when we will look at classification problems.

```
ggplot(mtcars, aes(x = disp, y = mpg, color = drat, shape = factor(am))) + geom_point() +
  labs(color = "Rear axle ratio", shape = "am")
```



Note that the colour can be a continuous variable, but it can also be a factor. The shape must always be a factor.

3 Your turn on the 'diamonds' dataset

The diamonds dataset in included in the ggplot2 library:

```
library(ggplot2)
data("diamonds")
```

Tasks:

- Have an initial look at the diamonds dataset using the functions we saw las week (str, summary and head). You will notice that the dataset is quite large (> 50000 rows) so some operations might take few seconds to complete.

- Calculate and plot the correlation between all the variables (be careful, you can only calculate the correlation coeeficient on numeric variables)
- Using boxplots and scatter plots, visualise the relationship between the price and all the other variables.
- carat seems to be the variable that has the most influence on the price, can we augment it with one of the variable in a scatter plot.