Data Science and R – Lab 11

Working directories, reading files and plotting with `ggplot2`

`0. Loading ggplot2 and setting working directory`

Make sure that ggplot2 (is installed and) loaded in RStudio

```
library(ggplot2)
```

Check what is your working directory by typing `getwd()`. Change this directory to the Downloads folder by typing

- `setwd("C:/Users/<username>/ **Downloads**")` (Windows) OR
- `setwd("/Users/<username>/**Downloads**")` (macOS)

where `<username>` is your username on your computer. If you cannot locate the path to your Downloads folder, just set your working directory to `"C:/"`.

Open an R script file and save it as `lab11.Rmd`, preferably in your working directory. If you open a markdown file, please make sure that it is saved in your working directory. RStudio often does this automatically for you.

`1.Computers from (before?) the time you were born (1990s)`

Click on the link below and open it in a Web browser

https://raw.githubusercontent.com/vincentarelbundock/Rdatasets/master/csv/Ecdat/Computers.csv

Right click on the file, select "Save As" and choose your working directory as the location to save. The file will be saved as `Computers.csv`

This is a dataset from the 90s about computer prices. Let's dig deeper.

a) Read this file into a data frame called `computers` using `read.csv("Computers.csv")`. Check the variables in `computers` using `head()`, `str()` or `View()`

b) Plot a scatter plot of the `speed+hd+ram` on the x-axis and the `price` on the y-axis. Use `ggplot()` with `aes()` and `geom_points()`. Do you see any general pattern between the two variables?

---

c) Transform the x-axis to its `log` using `scale_x_continuous()` and add it. Reduce point opacity by supplying `geom_point(alpha=0.5)`.

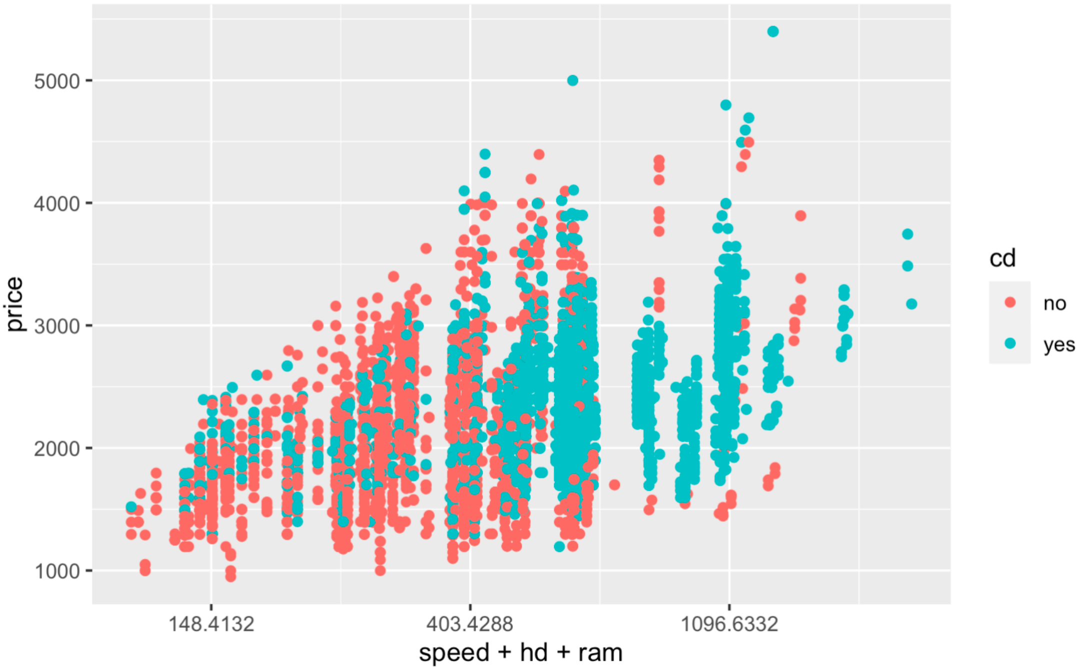
---

d) Add a regression line using the `geom_smooth()` function with `method=lm`. Is price generally higher when speed combined with hard disk and memory is higher?

---

e) We now want to study the effect of having a CD drive on top of the speed, hard drive and memory. Split the data by `colour` along the `cd` variable (add within `aes`) and remove the `geom_smooth()` (See plot below). Are computers with CD drives generally more expensive, have higher speed, HD and RAM?

f) Finally, split the data by colour but this time with `screen` size. What can you conclude about the screen size and the price of the computers?



2. Pigeon racing dataset

Use `read.csv()` to directly create a data frame `pigeon` from this URL (right click, copy the link address and paste it as the first parameter of `read-csv)`:

https://github.com/joanby/python-ml-course/raw/master/datasets/pigeon-race/pigeon-racing.csv

a) How many observations (rows) and variables (columns) does pigeon have?

b) Plot a scatter plot with the position (Pos) on the x-axis and Speed on the y-axis using ggplot(). What do you observe?

c) Split the data by `Sex` using colour. Is one sex particularly faster than the other or what does the plot suggest?

3. SAT scores by State

Download this dataset into `sat` using `read.csv()` with the appropriate parameters after viewing it:

http://www.randomservices.org/random/data/SATbyState.txt

a) Plot a scatter plot of the `Verbal` test scores on the x-axis and `Math` scores on the y-axis. What is the relationship between the Math and Verbal scores?

____

b) Do you see any outliers in this data? What could be the reasons for it?

____

c) Split the data by the participation `Rate` using `size`. What can you say about the participation rate of states with low scores in Maths or Verbal tests?

____

4. SAT score by year

Inspect and download the SAT score by year and gender from here into `sat2`:

http://www.randomservices.org/random/data/SATbyYear.txt

a) Plot the year on the x-axis and the average verbal scores (`AVerbal`) on the y-axis. What do you see of the trend of the average SAT score for the verbal test?

____

b) Split the data by the female verbal score (`FVerbal`) using `size`. Are the female verbal scores also following the trend of the average verbal scores over the years? What about the male verbal scores?

____

c) Plot the year on the x-axis and the average Math scores (`AMath`) on the y-axis. What do you see of the trend of the average SAT score for the Math test?

____

d) Again we want to see if the average Math scores is reflected in both genders respectively over time. Split the data by the female Math score (`FMath`) using `size`. Are the female Math scores also following the trend of the average Math scores over the years? What about the male Math scores (`MMath`)?

____