Notebooks, I/O and Exploratory Data Analysis (EDA)

Mark Asch - IMU/VLP/CSU

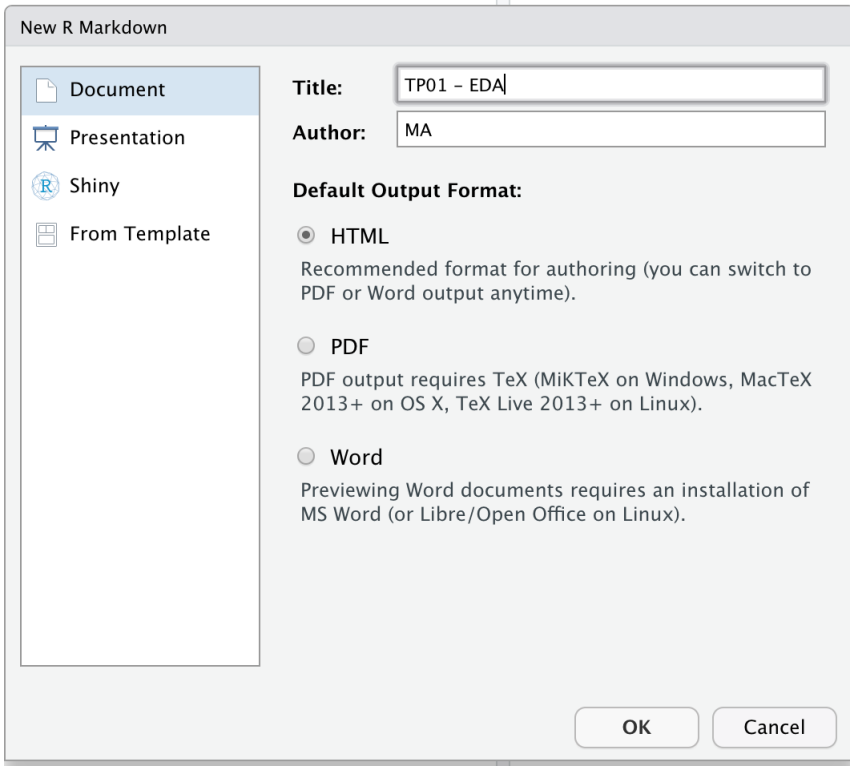
2023

Intro to ML

Creation of a Notebook in RStudio

- Open a new notebook :

  $\Rightarrow$ File $\rightarrow$
  $\Rightarrow$ New file $\rightarrow$
  $\Rightarrow$ R markdown... ou R Notebook



  $\Rightarrow$ Insert text, code, results, commentaries and conclusions.

  $\rightarrow$ Code $\rightarrow$ Insert Chunk

$\Rightarrow$ Execution : (Code $\rightarrow$) Run $\rightarrow$ Run current chunk

$\Rightarrow$ Visualize/Output : Knit $\rightarrow$ Knit to HTML (Knit to PDF)

Structure of a Notebook

- blocks of
  - ⇒ text
  - ⇒ code
  - ⇒ graphics
- instructions/recommendations :
  - ⇒ divide into blocks so that each block only produces a SINGLE output (text or graphic)
  - ⇒ avoid multiple outputs...
  - ⇒ comment each block with markdown (see below)
  - ⇒ comment each result!

Markdown

- titles: #, ##, etc.
- lists: -, 1., a., etc.
- text format:
  - ⇒ *italics*
  - ⇒ **bold**
- symbols, formulas, equations : use LaTeX...
  - ⇒ $ Ax = b$
  - ⇒ $$ I = \int_a^b f(x) dx $$
- verbatim (keywords, etc) : ‘code‘

Input/Output of Data (I/O)

- many databases are included in R, and/or in R libraries

```
> data() # for the complete list
> data("iris") # load the  iris data
> head (iris) # the first 6  lines
```

- a large number available from the UCI archives: https://archive.ics.uci.edu/

- load and save data in an R session:

```
> save(x, y, z, file = "data.Rdata")
> save(data, file = "data.Rdata")
> load("data.Rdata")
```

Dataframes

- create a dataframe (R data structure) :

```
> subject_name <- c("J. Du", "A. Du","P. Ba")
> temperature <- c(36.8, 37.8, 39.5)
> status_COVID <- c(FALSE, FALSE, TRUE)
> gender <- factor(c("M", "F", "M"))
> blood <- factor(c("O", "AB", "A"),
            levels=c("A", "B", "AB", "O"))
# Create the dataframe
pt_data <- data.frame(subject_name, temperature,
        status_COVID , gender, blood,
        stringsAsFactors=FALSE)
# Print the dataframe
> pt_data
 subject_name  temperature status_COVID gender bloo
1 J. Du        36.8         FALSE       M      O
2 A. Du        37.8         FALSE       F     AB
3 P. Ba        39.5         TRUE        M      A
```

File I/O

- load a CSV file (exported from a spreadsheet/Excel) :

```
> my_data <- read.csv("data.csv", header=TRUE)
```

- if the separator is a «;» : read.csv2(...)
- save a model for later use:

```
final_model <- ...
# save model on disk
saveRDS(final_model, "./final_model.rds")
# later on...
# load the model
super_model <- readRDS("./final_model.rds")
print(super_model)
# predictions of "new data"
previsions <- predict(super_model, ... )
```

Exploratory Data Analysis

✔ A first, crucial step in the process of «data science»

✔ There is no hypothesis, no model - we explore and try to understand the problem!

✔ The tools of EDA are:

- plots
- graphics
- summary statistics

✔ The methodology:

- pass systematically through all the data
- calculate all the summary statistics: mean, minumum, maximum, quartiles, outliers
- plot all distributions of all the variables ("box plots")
- plot all time series
- try changes of variables
- look at all the relations two-by-two («scatterplots»)

- See also the document learn_stat.pdf