An introduction to R software environment

Prelude

Edmondo Di Giuseppe

Institute of Bioeconomy, CNR-National Research Council of Italy

RISIS Winter School

Tools and methods for analysing complex Science, Technology and Innovation (STI) systems: A gentle introduction to Network Science (NS), Spatial Models (SM) and Machine Learning (ML)

ONLINE PLATFORM, February 15-26, 2021

1. What is R
   - That is the question!
   - Popularity
   - Objects oriented
2. What foR
   - Native foR statistical modeling
3. R at work
   - Work with data
   - Visualizing
4. Editor foR
5. Lesson Plan

is a free software environment for data analysis, graphics and statistical computing

The **R** name is based on the (first) names of the first two developers **R**obert Gentleman and **R**oss Ihaka, playing on the name of the **S** language, from which R is derived.

early 90s **Robert Gentleman** and **Ross Ihaka** start developing R as a derivation of *S language* at the Department of Statistics of the University of Auckland

1995 R is available under **General Public License** (GPL)

1997 **R core** group of programming developers is founded

2002 **R foundation for Statistical computing** is established in Vienna

National Research Council of Italy
Institute of BioEconomy

R

is a free software environment for data analysis, graphics and statistical computing

The **R** name is based on the (first) names of the first two developers **R**obert Gentleman and **R**oss Ihaka, playing on the name of the **S** language, from which R is derived.

⇓

early 90s **Robert Gentleman** and **Ross Ihaka** start developing R as a derivation of *S language* at the Department of Statistics of the University of Auckland

1995 R is available under **General Public License** (GPL)

1997 **R core** group of programming developers is founded

2002 **R foundation for Statistical computing** is established in Vienna

Outline

1 What is R
- That is the question!
- Popularity
- Objects oriented

2 What foR
- Native foR statistical modeling

3 R at work
- Work with data
- Visualizing

4 Editor foR

5 Lesson Plan

What is really R?

- a tool for data mining and statistical modeling?
- a programming language for data mining and statistical modeling?

⇓

is a dialect

What is really R?

- a tool for data mining and statistical modeling?
- a programming language for data mining and statistical modeling?

⇩

R is a dialect

Outline

1 What is R
- That is the question!
- Popularity
- Objects oriented

2 What foR
- Native foR statistical modeling

3 R at work
- Work with data
- Visualizing

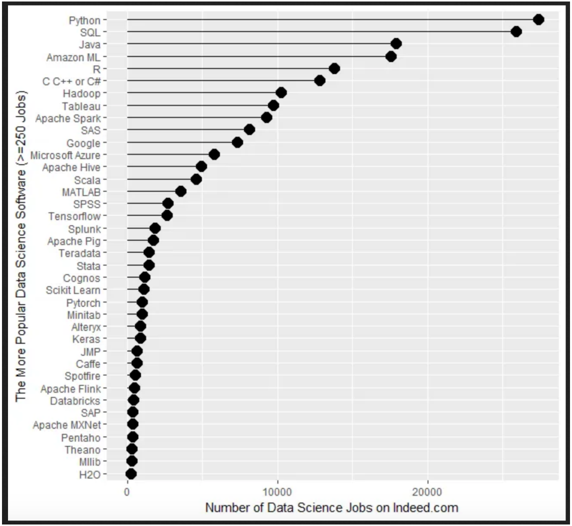
4 Editor foR

5 Lesson Plan

National Research Council of Italy
Institute of BioEconomy

The Popularity of Data Science Software

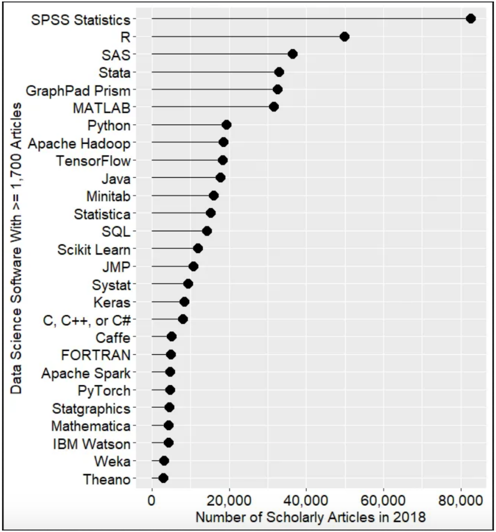
r4stats.com



National Research Council of Italy
Institute of BioEconomy

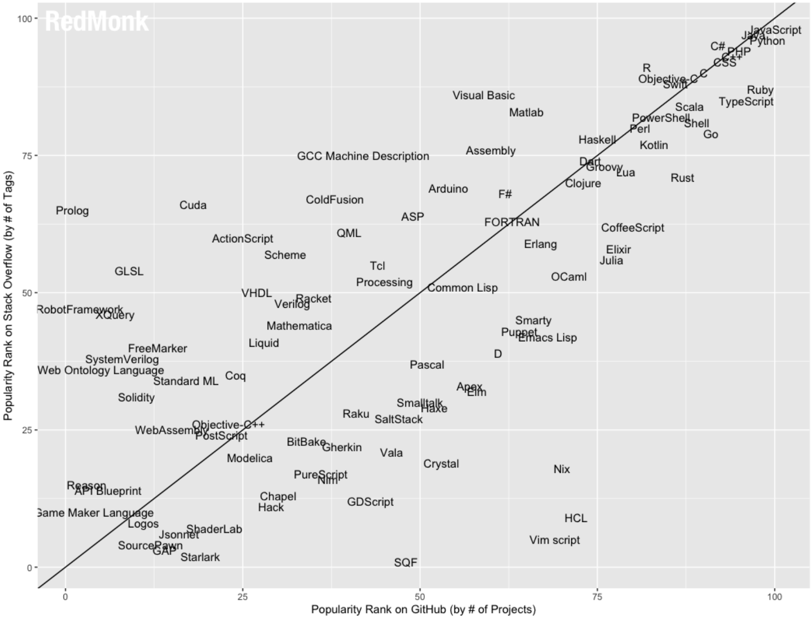
The Popularity of Data Science Software

r4stats.com



National Research Council of Italy
Institute of BioEconomy

RedMonk Q320 Programming Language Rankings



R’s pros and cons (very short list)

Pros

17058 available packages (retrieved from CRAN Feb 9, 2021)

2138 open projects on R-forge

large community

Cons

limits in processing extremely large files (R generally processes data in-memory).

National Research Council of Italy
Institute of BioEconomy

Outline

1 What is R
- That is the question!
- Popularity
- Objects oriented

2 What foR
- Native foR statistical modeling

3 R at work
- Work with data
- Visualizing

4 Editor foR

5 Lesson Plan

National Research Council of Italy
Institute of BioEconomy

R is an interpreted language

↓

It means that **CODE** entered into the **R CONSOLE** (or run as R script in batch mode) is executed by the **interpreter**, a program within the R system.

Object oriented programming

Data are stored in objects.

vector

matrix

array

factors

dataframe

list

...

Every object has a **type** (double, complex, etc.) and is a member of a **class** (numeric, character, etc.).

National Research Council of Italy
Institute of BioEconomy

R code is composed of

⬇

a series of expressions

- arithmetic expressions
- assignment statements
- conditional statements
- loops
- ...

functions

- pre-loaded
- via packages to be installed and loaded
- own

National Research Council of Italy
Institute of BioEconomy

Outline

1. What is R
   - That is the question!
   - Popularity
   - Objects oriented
2. What foR
   - Native foR statistical modeling
3. R at work
   - Work with data
   - Visualizing
4. Editor foR
5. Lesson Plan

National Research Council of Italy
Institute of BioEconomy

Statistical modeling

Most statistical theory focuses on `data modeling`, `prediction` and `inference`.

Packages for classical and advanced statistical models:

- Time series analysis
- Regression and Classification
- Principal Component Analysis
- Spatio-temporal Analysis
- Bayesian modeling
- ...

Among them:

- Network Science (NS)
- Spatial Models (SM)
- Machine Learning (ML)

Outline

1 What is R
- That is the question!
- Popularity
- Objects oriented

2 What foR
- Native foR statistical modeling

3 R at work
- Work with data
- Visualizing

4 Editor foR

5 Lesson Plan

Data import

Supported data files

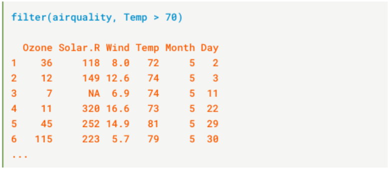
- .xls
- .csv
- .txt
- NetCDF
- .....

Connection to databases

- SAS
- Microsoft Access
- MySQL
- STATA
- .....

Data manipulation: `dplyr` package

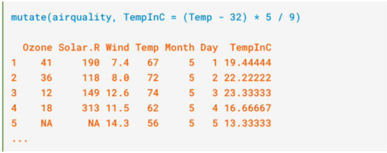
Data manipulation is the process of **arranging data** in order to make data analysis process, such as visualizing and modeling.



```
filter(airquality, Temp > 70)

  Ozone Solar.R Wind Temp Month Day
1    36     118  8.0   72     5   2
2    12     149 12.6   74     5   3
3     7      NA  6.9   74     5  11
4    11     320 16.6   73     5  22
5    45     252 14.9   81     5  29
6   115     223  5.7   79     5  30
...
```

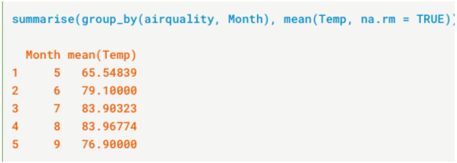
(a)



```
mutate(airquality, TempInC = (Temp - 32) * 5 / 9)

  Ozone Solar.R Wind Temp Month Day  TempInC
1    41     190  7.4   67     5   1 19.44444
2    36     118  8.0   72     5   2 22.22222
3    12     149 12.6   74     5   3 23.33333
4    18     313 11.5   62     5   4 16.66667
5    NA      NA 14.3   56     5   5 13.33333
...
```

(b)



```
summarise(group_by(airquality, Month), mean(Temp, na.rm = TRUE))

  Month mean(Temp)
1     5   65.54839
2     6   79.10000
3     7   83.90323
4     8   83.96774
5     9   76.90000
```

(c)

National Research Council of Italy
Institute of BioEconomy

Outline

1 What is R
- That is the question!
- Popularity
- Objects oriented

2 What foR
- Native foR statistical modeling

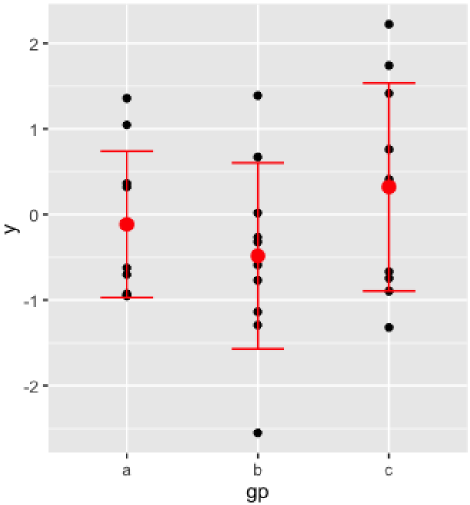
3 R at work
- Work with data
- Visualizing

4 Editor foR

5 Lesson Plan

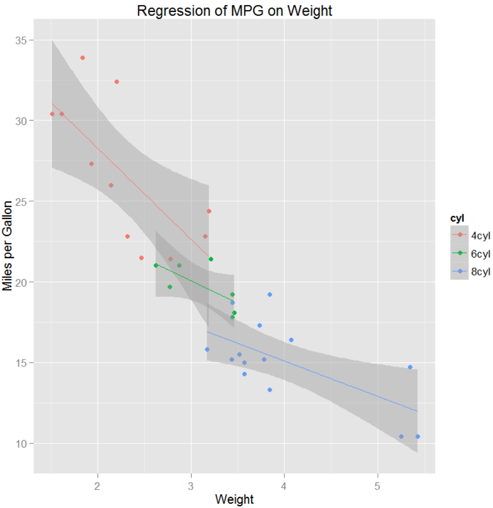
National Research Council of Italy
Institute of BioEconomy

Visualizing data: `ggplot2` package



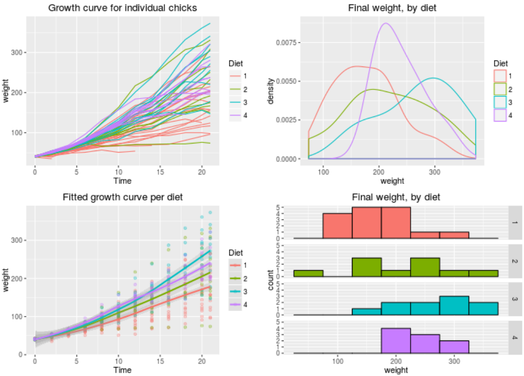
National Research Council of Italy
Institute of BioEconomy

Visualizing data: `ggplot2` package

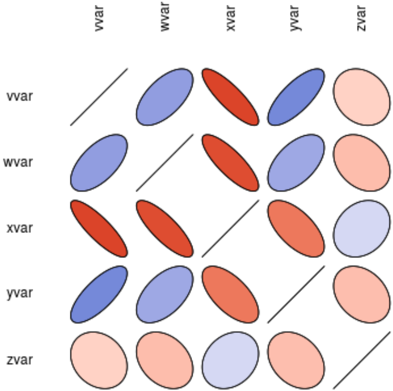


National Research Council of Italy
Institute of BioEconomy

Visualizing data: `ggplot2` package

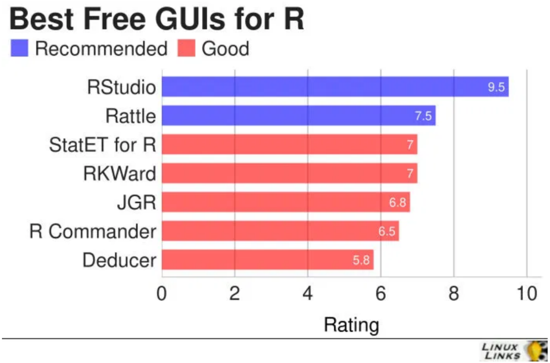


Visualizing data: `ggplot2` package



Graphical User Interfaces for R

LinuxLinks



National Research Council of Italy
Institute of BioEconomy

Monday 15th of February 2021 - afternoon

First part (15.15 - 16.00)

- Presentation of **RStudio** editor and its utilities
- Introduction to basic **R syntax**
- Installation of **R packages** and use of functions
- Comfort break (15 minutes)

Second part (16.15 - 17.00)

- **if else** statement and **loop**
- Writing **own functions**
- Comfort break (5 minutes)

Last part (17.05 - 17.30)

- Grouping and summarizing data: the **dplyr** package

National Research Council of Italy
Institute of BioEconomy