UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Superior d'Enginyeries Industrial,
Aeroespacial i Audiovisual de Terrassa

MUEI, MUEA, and MASE

Advanced Engineering Data Analysis (AEDA)
General Information

Prof. Daniel Fernández

daniel.fernandez.martinez@upc.edu

Professor

Daniel Fernández (**coordinator**)

daniel.fernandez.martinez@upc.edu

My background

- B.S. Statistics . UPC. 1998
- B.S. Computer software. UPC. 2000
- M.S. in statistics and probability. Mathematics Research Centre. 2005
- Ph.D. in Statistics - Victoria University of Wellington (VUW). 2015
- Post Doctoral Fellow in Statistics. VUW. 2015
- Moore-Sloan Post Doctoral Associate in Statistics. NYU. 2016
- Professor, Department of Epidemiology and Biostatistics, SUNY, 2017
- Currently, Serra-Húnter lecturer professor. UPC.

Class method

- Virtual campus: Atenea, https://atenea.upc.edu/

Class method

- Virtual campus: Atenea, https://atenea.upc.edu/
- Mondays 17-19h. Theory (Room: TR5-0.3)

Class method

- Virtual campus: Atenea, https://atenea.upc.edu/
- Mondays 17-19h. Theory (Room: TR5-0.3)
- Tuesdays 15-17h. Lab/Practice (Room: TR5-0.3)

Do you have laptops?

Goal

The purpose of this course is:

- To introduce the most common and well-known multivariate statistical methods to non-mathematicians.
- It is not intended to be a comprehensive course (mathematically speaking). However, it is important not to take the statistical methods a black box.
- The intention is to keep the details to a minimum while serving as a practical guide that illustrates the possibilities of multivariate statistical analysis.
- In other words, it is a course to "get you going" in a particular area of statistical methods.

Preliminary Knowledge

It is assumed that you (the students) have a working knowledge of:

- Elementary statistics
  - Basic statistics: summary statistics (mean, median,…), Normal distribution, CI, Hypothesis testing,…
  - linear regression
  - EDA (visualization)

- Some facility with algebra is also required to follow the equations in certain parts of the text. Understanding the theory of multivariate methods requires some matrix algebra. However, the amount needed is not great.

Preliminary ideas

- We will learn a comprehensive set of multivariate methods. The important thing you do not have to memorize the techniques and their assumptions. **You need to know what they do conceptually.**

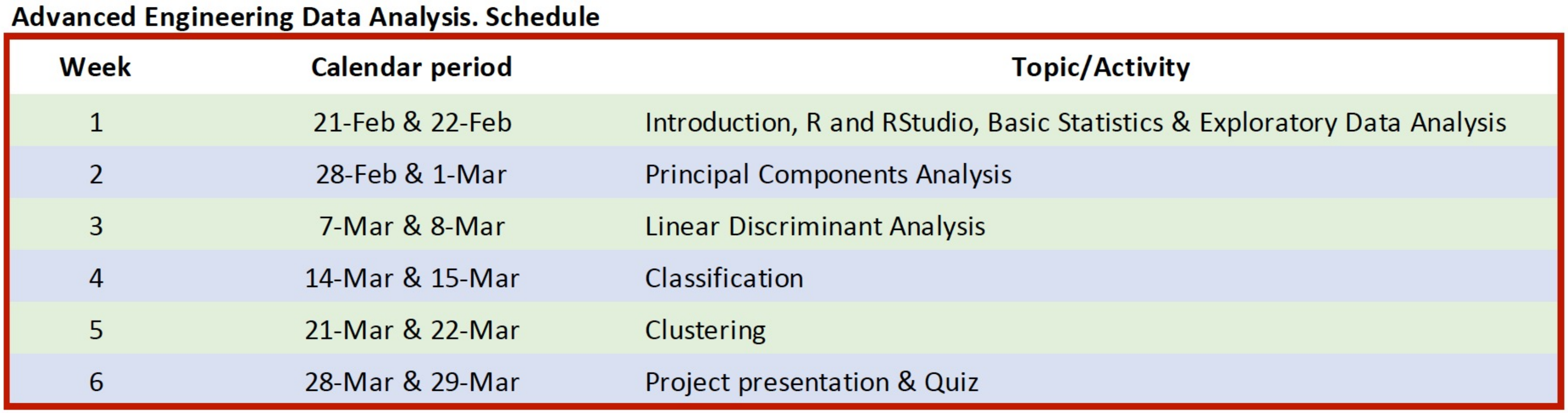
- This course should be helpful to **train your brain** in a way that if you have data and you want to analyze it, you are can go to a book/slides/etc. and find what you want.

- Think about statistics techniques as a **toolbox**, and you use the one you consider more necessary in each moment.

Thought about models

- All models and techniques **are wrong**.
- Nothing can simulate the real data to perfection, but we always look for the best model (i.e., the model with less uncertainty), but **uncertainty is always there**.
- In Physics is possible, because they are laws, However, we **play with uncertainty in Statistics**. The good thing about statistics is that we can measure the error (95% of CI, for example)
- "*I think it is much more interesting to live with uncertainty than to live with answers that might be wrong*" –Richard Feynman (https://en.wikipedia.org/wiki/Richard_Feynman).

Tentative schedule



Advanced Engineering Data Analysis. Schedule

Week	Calendar period	Topic/Activity
1	21-Feb & 22-Feb	Introduction, R and RStudio, Basic Statistics & Exploratory Data Analysis
2	28-Feb & 1-Mar	Principal Components Analysis
3	7-Mar & 8-Mar	Linear Discriminant Analysis
4	14-Mar & 15-Mar	Classification
5	21-Mar & 22-Mar	Clustering
6	28-Mar & 29-Mar	Project presentation & Quiz

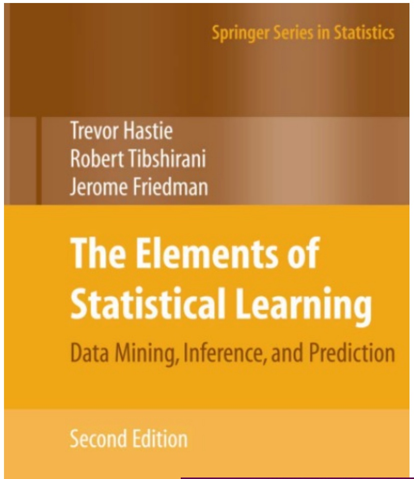
Class description

- Mondays 17-19h. Theory
  - We will show the methods with examples.

- Tuesdays 15-17h. Lab Practice
  - Goal: practice.
  - We will deliver examples and lab practice (**not evaluable**).
  - Lab practice will be solved **individually** (COVID-19)
  - Questions can be posed during the two hours.
  - Important: we **will not respond** to questions regarding the lab practice after the end of the class.

Note: Lab practice is not the same as the evaluable exercises you will be doing in groups (3-4 people) throughout the course

Basic Bibliography



Springer Series in Statistics

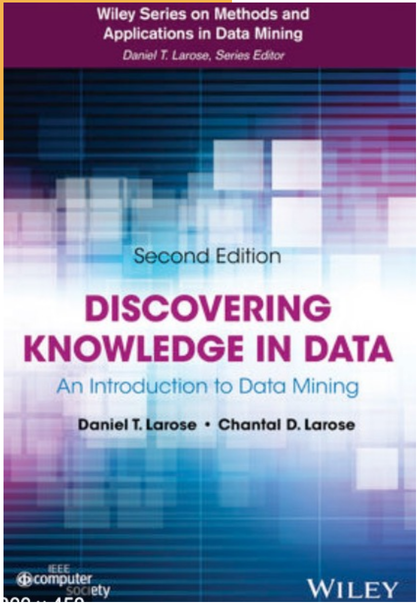
Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer



Wiley Series on Methods and Applications in Data Mining

Daniel T. Larose, Series Editor

Second Edition

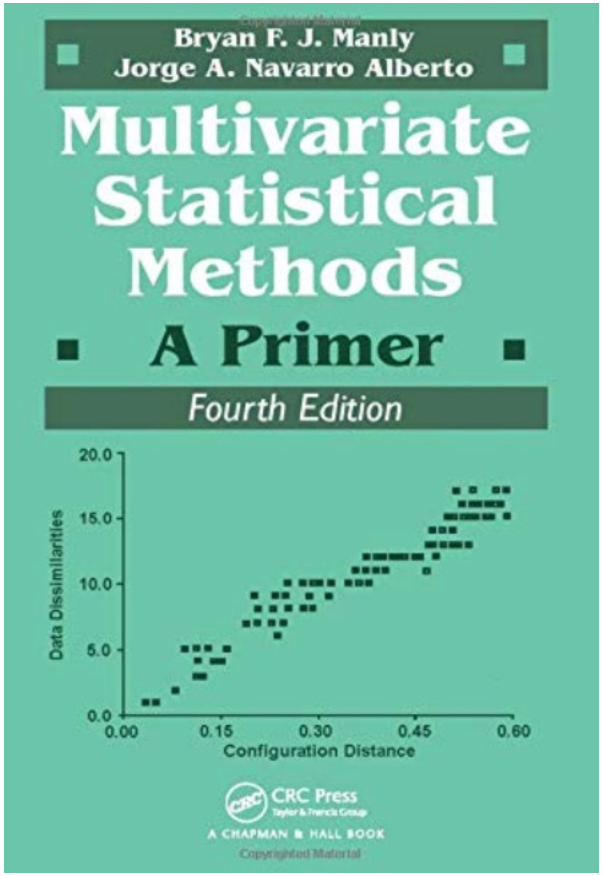
DISCOVERING KNOWLEDGE IN DATA

An Introduction to Data Mining

Daniel T. Larose • Chantal D. Larose

IEEE computer society

WILEY



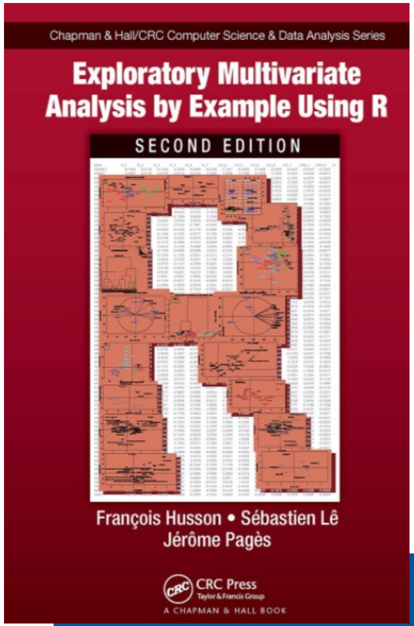
Bryan F. J. Manly
Jorge A. Navarro Alberto

Multivariate Statistical Methods

A Primer

Fourth Edition

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



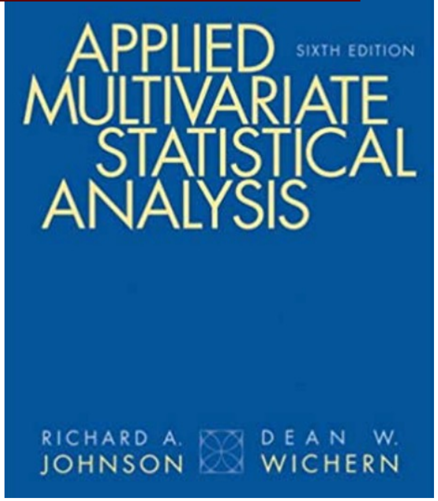
Chapman & Hall/CRC Computer Science & Data Analysis Series

Exploratory Multivariate Analysis by Example Using R

SECOND EDITION

François Husson • Sébastien Lê
Jérôme Pagès

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



APPLIED MULTIVARIATE STATISTICAL ANALYSIS

SIXTH EDITION

RICHARD A. JOHNSON
DEAN W. WICHERN

Course grading

Academic assessment will be based on the grades obtained in:

1. Exercises – **30%**

2. A exam (Quiz) – **30%**

3. A project – **40%**

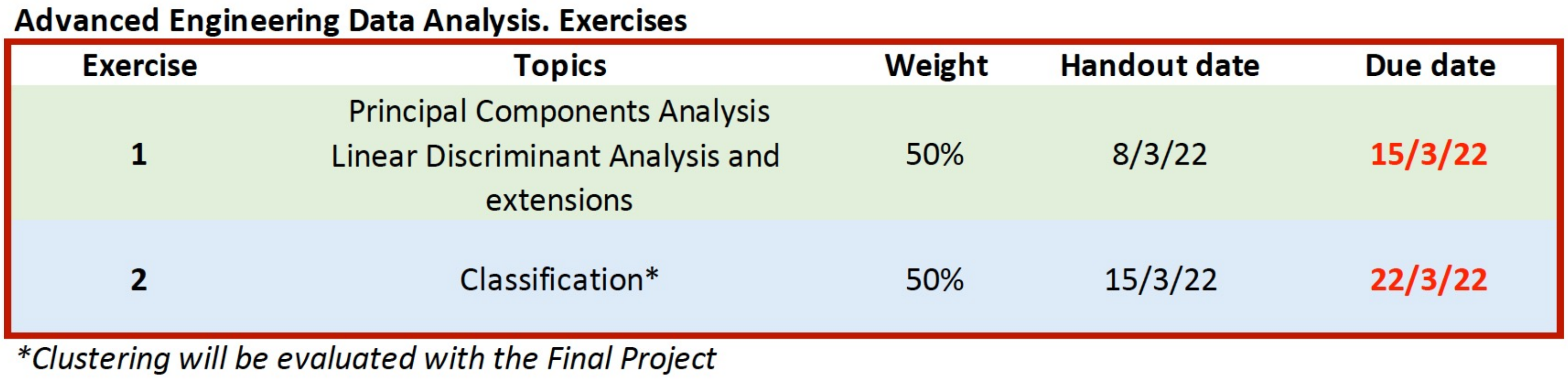
Exercises & Quiz

- **Exercises**
  - They will be 2 assignments corresponding to the different units of teaching (throughout the course).
  - They will be done in **3-4 people**
  - Submitted as a report (**PDF format**) to Atenea.
  - The report must have **2 sheets at most (four pages)**,
  - Written in 11-12 font size.

- **Quiz**
  - It will take place the last day of the course -- **Mar 29th**
  - It can have methodological, R and practical questions.

Exercises

- **Exercises**



Advanced Engineering Data Analysis. Exercises

Exercise	Topics	Weight	Handout date	Due date
1	Principal Components Analysis Linear Discriminant Analysis and extensions	50%	8/3/22	**15/3/22**
2	Classification*	50%	15/3/22	**22/3/22**

**Clustering will be evaluated with the Final Project*

Course project

- The project will be done on a **real data set**
- **Oral presentation** on second-to-last day of class (**Mar 28th**)
- Form a **3-4-person group** (same group as exercises. You are currently 12, that means 3 groups of 4 or 4 groups of 3, for instance)

- It must be done using R RStudio rmarkdown

- Basic steps of the work to do:
  - Choose a "real-world" domain and define the problem
  - Implement and test methods
  - Write a report
  - Present it orally

Instructions: **Project_instructions.pdf** document in Atenea

Examples of data sets repositories

- UC Irvine Machine Learning Repository
  http://archive.ics.uci.edu/ml/index.php

- Kaggle, with thousands of data sets in Business, Computer Science, Earth and Nature, Health, among other fields.
  https://www.kaggle.com/datasets

- IFCS Data Challenging Repository
  https://ifcs.boku.ac.at/repository/challenge2/

- Google public datasets https://cloud.google.com/bigquery/public-data/

- Eurostat database: https://ec.europa.eu/eurostat/data/database

Software

Software

- We will be using R and Rstudio.

- We recommend to start visiting the following webs:
  - The R Project for Statistical Computing: https://www.r-project.org/
  - The Comprehensive R Archive Network: https://cran.r-project.org/

- Reading the manuals included in the installation:

  https://cran.r-project.org/manuals.html
- Reading other documents:

  https://cran.r-project.org/other-docs.html
- And especially, this short introduction:

  https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

R and R studio. Installation

- R

It depends on the operating system, but we can consult in:

http://cran.r-project.org/bin

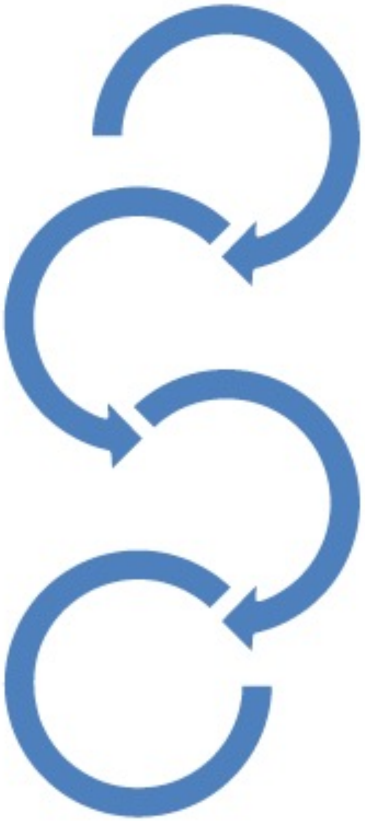
- Rstudio

R-Studio is one of the platforms from which you can run R as well as manage its packages, results, files, etc... https://www.rstudio.com/

It depends on the operating system, but we can consult in:

https://www.rstudio.com/products/rstudio/download/#download

Data analytics and R



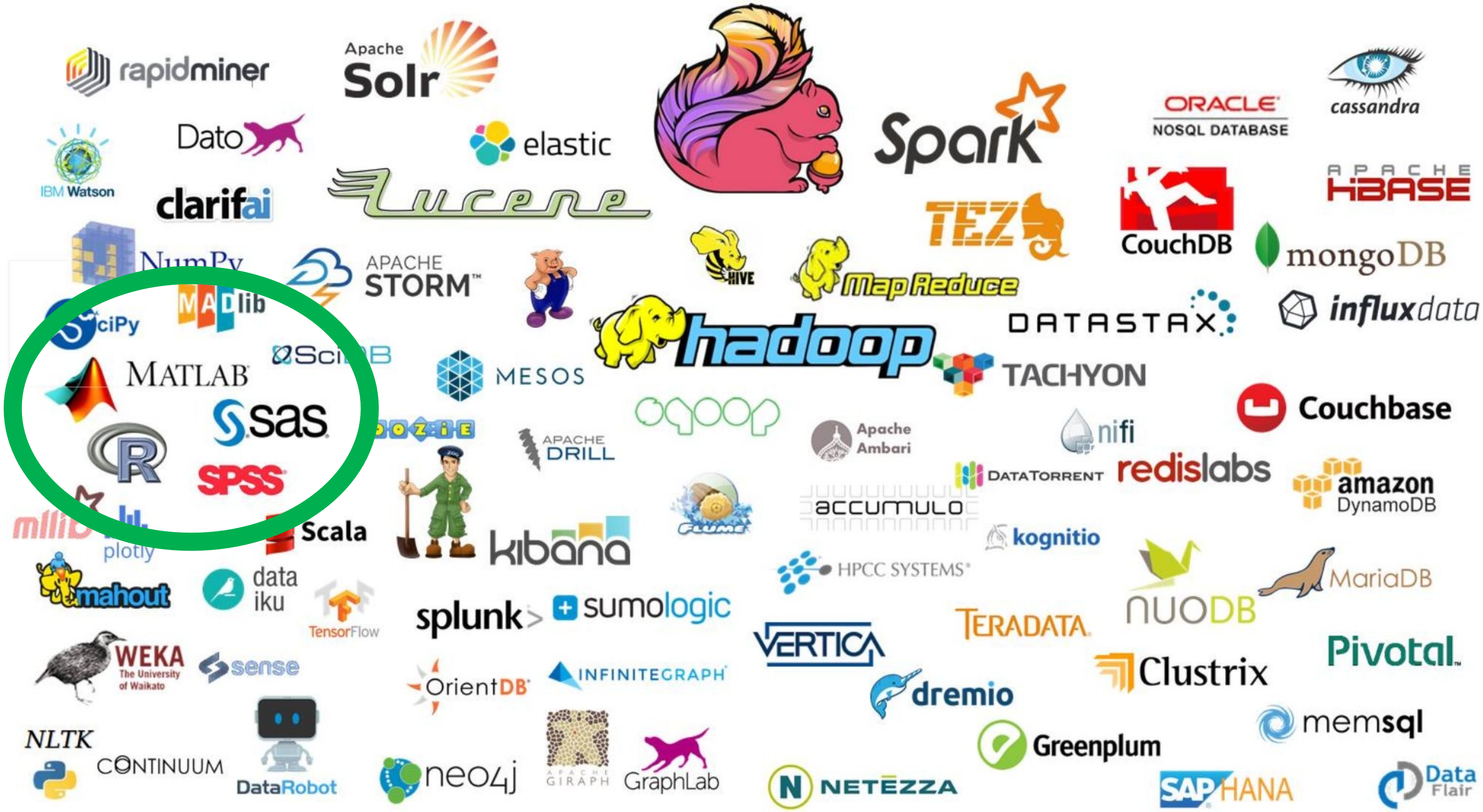
IMPORT

CLEANING & TRANSFORM

VISUALIZE & MODELLING

COMMUNICATE

What's R in the universe



R is increasing a lot

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

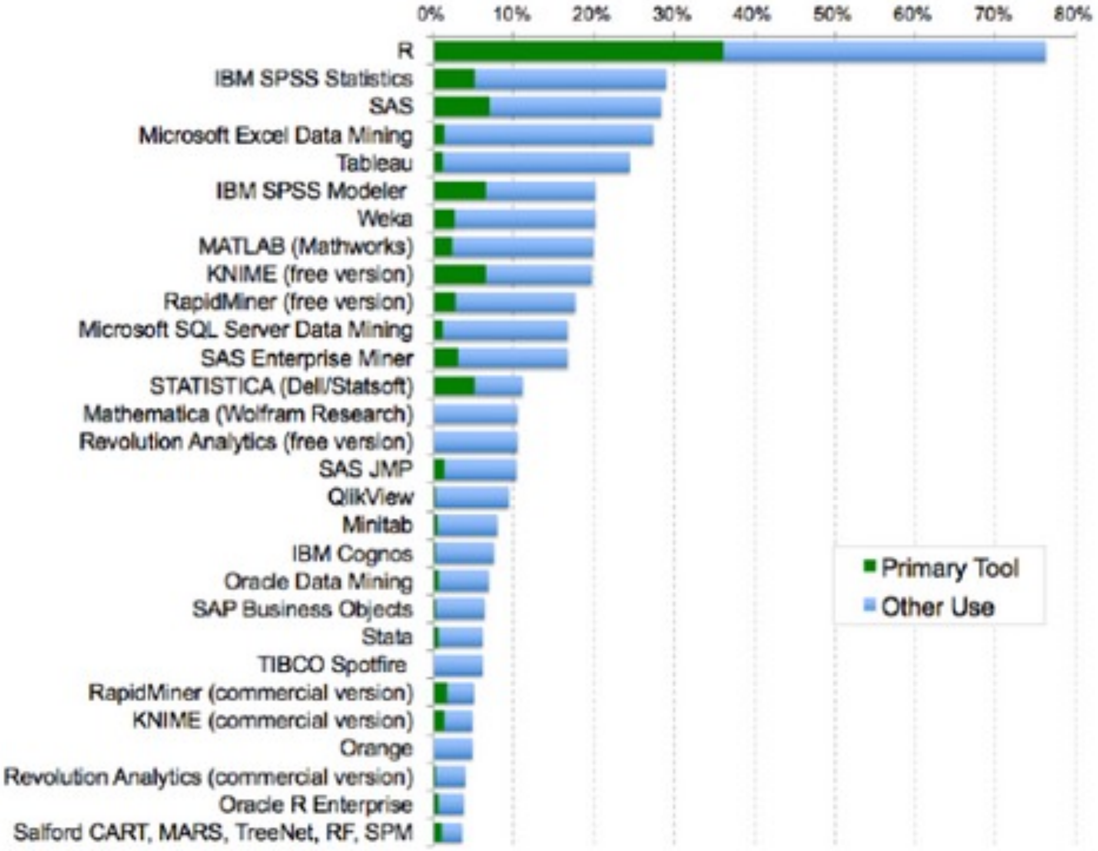


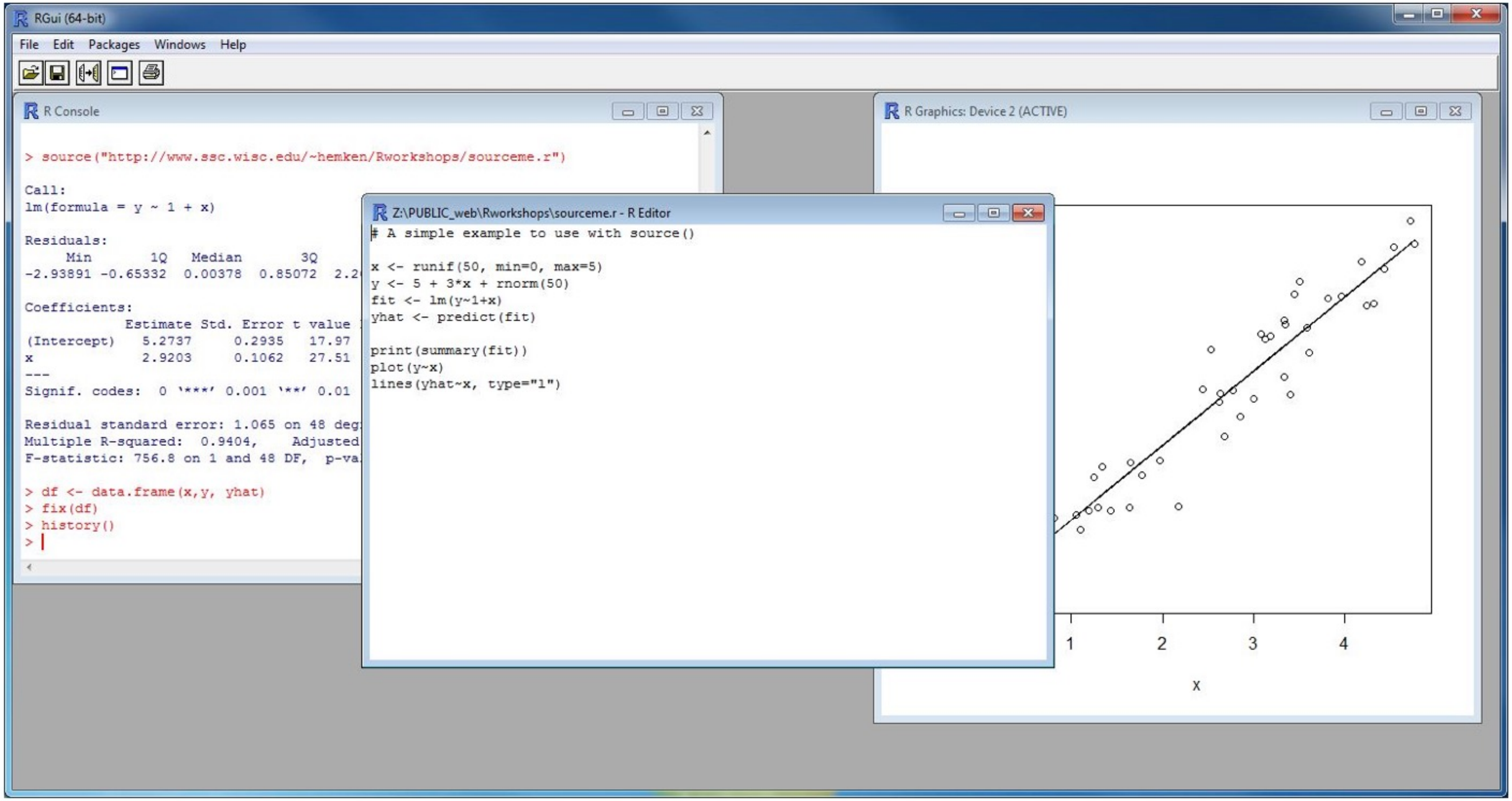
Figure 6a. Analytics tools used by respondents to the 2015 Rexer Analytics Survey. In this view, each respondent was free to check multiple tools.

R vs. Rstudio

R and RStudio are two different programs

- R is the program that calculates. It is **open-source**, **collaborative** and (**initially**) focused on **statistical computing**.
- R is also the language in which we write the commands, as it is also the **programming language**.
- R is w**idely spread** because:
  - There are many forums on the internet where users raise / resolve their doubts and / or proposals.
  - there are many libraries for R in continuous development and that allow to use it in a much faster and more efficient way.

R. Interface


RGui (64-bit)
File
Edit
Packages
Windows
Help
R Console
Z:\PUBLIC_web\Rworkshops\sourceme.r - R Editor
R Graphics: Device 2 (ACTIVE)
1
2
3
4
x

```
> source("http://www.ssc.wisc.edu/~hemken/Rworkshops/sourceme.r")

Call:
lm(formula = y ~ 1 + x)

Residuals:
     Min       1Q   Median       3Q
-2.93891 -0.65332  0.00378  0.85072  2.2

Coefficients:
            Estimate Std. Error t value
(Intercept)   5.2737     0.2935   17.97
x             2.9203     0.1062   27.51
---
Signif. codes:  0 '***' 0.001 '**' 0.01

Residual standard error: 1.065 on 48 deg
Multiple R-squared:  0.9404,    Adjusted
F-statistic: 756.8 on 1 and 48 DF,  p-va

> df <- data.frame(x,y, yhat)
> fix(df)
> history()
>
```

```
# A simple example to use with source()

x <- runif(50, min=0, max=5)
y <- 5 + 3*x + rnorm(50)
fit <- lm(y~1+x)
yhat <- predict(fit)

print(summary(fit))
plot(y~x)
lines(yhat~x, type="l")
```

R Console interface

R vs. Rstudio

R and RStudio are two different programs

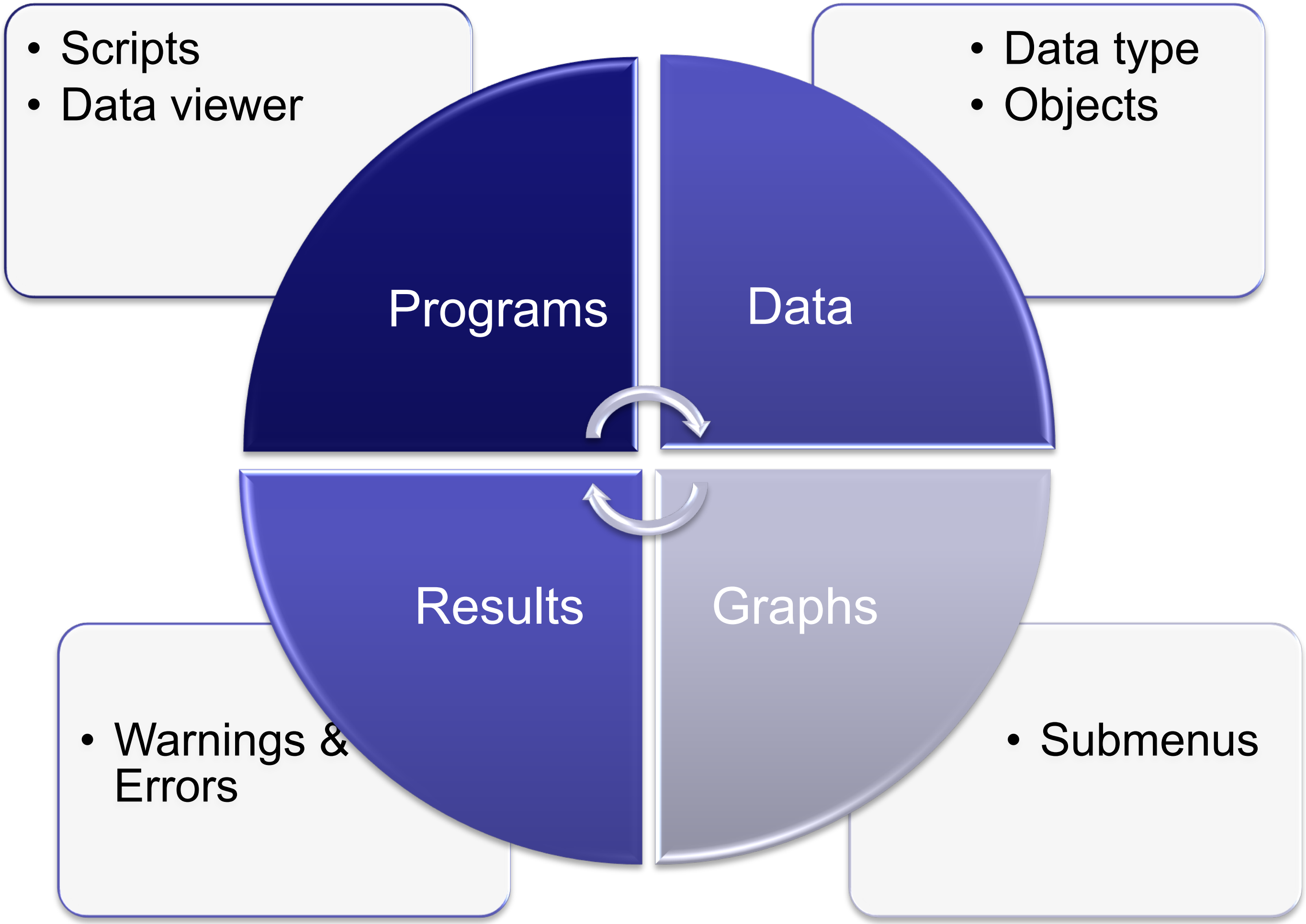
- RStudio is the **interface** where we will be working, as it offers us some **comforts**.

- It allows us to create **scripts** in a more agile way and in a much more **pleasant environment** than using R.

- It also makes it **easier** to install and uninstall libraries, load and view databases, etc.

- It offers **other possibilities** beyond R such as creating web pages, pdf's or word files with integrated R, which goes beyond this course.

R studio



- Scripts
- Data viewer

- Data type
- Objects

Programs

Data

Results

Graphs

- Warnings & Errors

- Submenus

RStudio. Interface

Toolbar

Save scripts, settings,…

Script

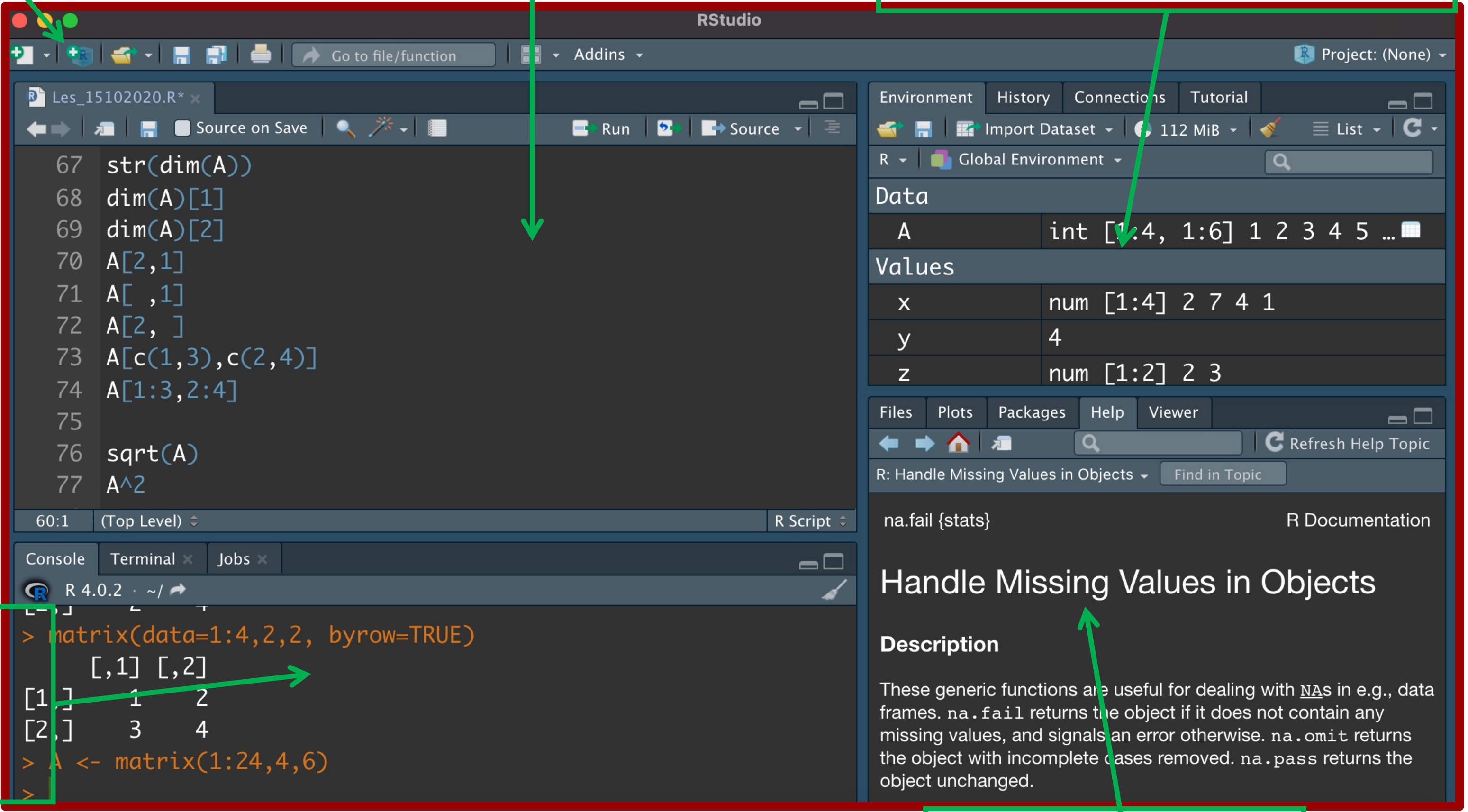
Text file where the analysis instructions are saved

Environment

Set of objects in memory

History

Set of executed Instructions



```
str(dim(A))
dim(A)[1]
dim(A)[2]
A[2,1]
A[ ,1]
A[2, ]
A[c(1,3),c(2,4)]
A[1:3,2:4]

sqrt(A)
A^2
```

```
> matrix(data=1:4,2,2, byrow=TRUE)
     [,1] [,2]
[1,]    1    2
[2,]    3    4
> A <- matrix(1:24,4,6)
>
```

Data	
A	int [1:4, 1:6] 1 2 3 4 5 …
Values	
x	num [1:4] 2 7 4 1
y	4
z	num [1:2] 2 3

na.fail {stats} R Documentation

Handle Missing Values in Objects

Description

These generic functions are useful for dealing with NAs in e.g., data frames. `na.fail` returns the object if it does not contain any missing values, and signals an error otherwise. `na.omit` returns the object with incomplete cases removed. `na.pass` returns the object unchanged.

Console

Commands and results

Explorer, Graphics Packages, Help, & Viewer

Tutorials in R

- We recommend to start visiting the following webs:
  - The R Project for Statistical Computing: https://www.r-project.org/
  - The Comprehensive R Archive Network: https://cran.r-project.org/
- Reading the manuals included in the installation:

https://cran.r-project.org/manuals.html

- Reading other documents: https://cran.r-project.org/other-docs.html
- And especially, these tutorials:

  https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

  https://www.theanalysisfactor.com/r/