



[EDA, info](#)

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources

Exploratory data analysis

Info

Vladimir Batagelj

IMFM Ljubljana, IAM UP Koper, NRU HSE Moscow

Master's programme

Applied Statistics with Social Network Analysis

International Laboratory for Applied Network Research

NRU HSE, Moscow 2021



Outline

EDA, info

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources

- 1 Some links
- 2 Program
- 3 Requirements
- 4 Time-table
- 5 Projekt R
- 6 Sources



Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si

Current version of slides (November 9, 2021 at 13:55): [slides PDF](#)



Some links

[EDA, info](#)

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources

Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si

EDA on wiki:

<http://vlawiki.fmf.uni-lj.si/doku.php?id=ru:hse:eda>

Master's program: <https://www.hse.ru/en/ma/sna/>



Program

EDA, info

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources

13. Exploratory Data Analysis. (4 cr.)

Prerequisites: Two statistics courses at the graduate level, or consent of instructor.

Numerical and graphical techniques for summarizing and displaying data. Exploration versus confirmation. Connections with conventional statistical analysis and data mining.

Applications to large data sets.

- Tukey, J.W., 1977. Exploratory data analysis. (for historical context in this area)
- Bock, H.H. and Diday, E. eds., 2000. Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer.
- Martinez, W.L., Martinez, A. and Solka, J., 2010. Exploratory data analysis with MATLAB. CRC Press.

EDA, info

V. Batagelj

Some links

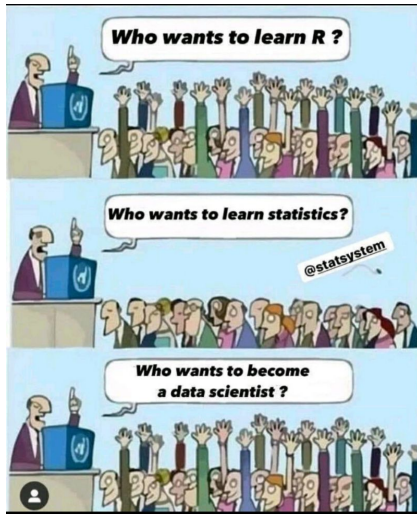
Program

Requirements

Time-table

Projekt R

Sources





Requirements

[EDA, info](#)

V. Batagelj

[Some links](#)

[Program](#)

[Requirements](#)

[Time-table](#)

[Projekt R](#)

[Sources](#)

Home projects:

- 1 find, download, fuse and clean a data set; save it in CSV or JSON; basic analyses
- 2 collect the data from the web site; save them; basic analyses
- 3 analysis of a large data set: at least 10 000 units with mixed variables

January 15, 2022



Time-table

EDA, info

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources

The lectures will be on Zoom

Vladimir Batagelj is inviting you to a scheduled Zoom meeting.

Topic: EDA

Time: This is a recurring meeting Meet anytime

Join Zoom Meeting

<https://us06web.zoom.us/j/96422044891?pwd=ckhJVzR0LzI0QXZDTHNKYWlJekJOUT09>

Meeting ID: 964 2204 4891

Passcode: 288814

from 18:30 till 19:50 and from 20:00 till 21:20 (Moscow time) on the following days

November 8, 9, 11, 15, 16, 18, 19, 22, 23, 25



Programming language and environment R

EDA, info

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources



Ross Ihaka and Robert Gentleman at DSC 2001

In this course we shall use the programming language/environment R.

R was developed in mid nineties by Robert Gentleman and Ross Ihaka from the Auckland University in New Zealand. It started as an open code version of S – a programming language for statistics. S was developed in 1976 by John Chambers and collaborators from Bell Laboratories. The commercial version of S is known as S-plus.

Project R was joined by many statisticians all around the world and gradually R became a language in which most of new statistical methods are developed and published.



Why R ?

EDA, info

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources

- R is **free** – we can use it for free; we can install it on our home PC; it runs on all main OSs: Windows, Linux / Unix and Mac.
- R is **open code** – we can inspect all its code: learning, security, adaptations. Project collaborators created over 18366 (October 2021) packages – program libraries for solving specific data analysis problems (**CRAN/Contributed**, **R-Forge**, **GitHub**).
- provides procedures for high quality *visualization* of data and results.
- R evolved as a programming language for statistics, but has also many *applications in related fields*: decision support, finance, biochemistry, etc.



Alternatives

[EDA, info](#)

V. Batagelj

[Some links](#)

[Program](#)

[Requirements](#)

[Time-table](#)

[Projekt R](#)

[Sources](#)

Computer scientists (data mining) prefer **Python** instead of R. Packages: Pandas, Numpy, SciPy, NLTK, Scikit-Learn, Orange, Mathplotlib, Plotly, etc.

In the future the new programming language **Julia** could replace R.

To document the analysis the R's **Markdown** is often used or some other type of notebooks. In data analysis are quite popular **Jupyter** notebooks based on Python (**Anaconda Python distribution**) : **R** and **Julia**.

For collaborative projects we can use **wikis** or **GitHub**.



Sources

EDA, info

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources

- Projekt R: <http://www.r-project.org/>
- CRAN (The Comprehensive R Archive Network):
<http://cran.at.r-project.org/>
- RStudio: <https://www.rstudio.com/>
- The R Journal: <http://cran.r-project.org/doc/Rnews/>
- conference UseR! :
<http://www.r-project.org/conferences.html>
- [the R graph gallery](#)
- Reference Cards: [R1](#), [R2](#), [RStudio](#)



Books

[EDA, info](#)

V. Batagelj

Some links

Program

Requirements

Time-table

Projekt R

Sources

- [Use R!](#), Springer
- [The R Series](#), Chapman & Hall/CRC
- [O'Reilly](#)
- [Manning](#) / Data Science
- Chambers, John: Software for Data Analysis: Programming with R. Springer 2008.
- Wickham, Hadley, Grolemund, Garrett: R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly 2017. [WWW](#)
- Wickham, Hadley: ggplot2: Elegant Graphics for Data Analysis, 2nd ed., Springer 2016.
- McNicholas, Paul D., Tait, Peter: Data Science with Julia. Chapman and Hall/CRC 2019.