Exploratory data analysis

Info

Vladimir Batagelj

IMFM Ljubljana and IAM UP Koper

Master's programme

Applied Statistics with Social Network Analysis

International Laboratory for Applied Network Research

NRU HSE, Moscow 2017

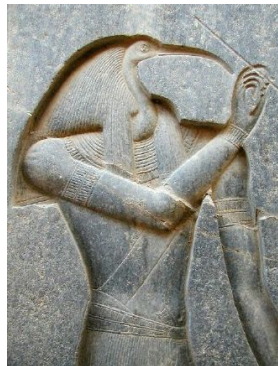# Outline

EDA, info

V. Batagelj

Teachers
Program
Requirements
Time-table
Projekt R
Sources

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

Current version of slides (November 7, 2017 at 20 : 37): slides PDF

Teachers

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

Daria Maltseva: `d_malceva@mail.ru`

EDA on wiki:
`http://vladowiki.fmf.uni-lj.si/doku.php?id=ru:hse:eda`

Master's programme: `https://www.hse.ru/en/ma/sna/`

Program

EDA, info

V. Batagelj

Teachers

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.

Requirements

Home project:

1. download and fuse data; save them in CSV and 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

Time-table

The lectures will be from 18h to 21h:

November 7, 8, 9, 10, 11
November 12 - off
November 13, 14, 15, 16

On November 17th, we move to the training center.
Lectures from 3 pm until 9 pm, with an hour break for dinner and
some breaks for coffee.

Programming language and environment R

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 John Chambers and collaborators from Bell Laboratories. The commercial version of S is known as S-plus.

Project R was joined by many statistitians 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

Teachers
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, addaptations; project collaborators created over 11750 (November 2017) packages – program libraries for solving specific data analysis problems (CRAN/Contributed, R-Forge).

- 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.

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
- Ashlee Vance: Data Analysts Captivated by R's Power. The New York Times, 6. jan. 2009