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 2020

Outline

1 Some links

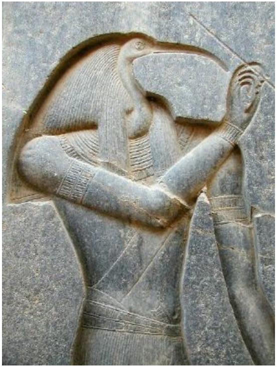
2 Program

3 Requirements

4 Time-table

5 Projekt R

6 Sources



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

Current version of slides (November 10, 2020 at 21 : 18): slides PDF

Some links

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

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

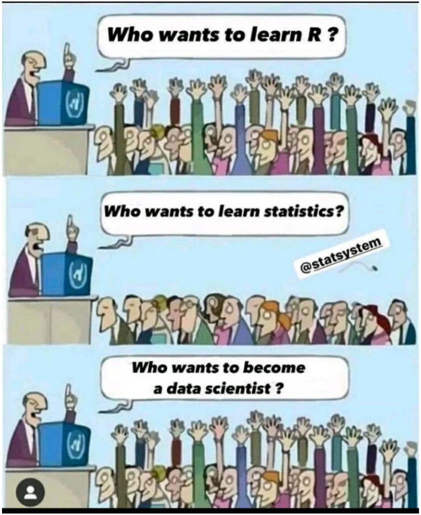
Master's program: **`https://www.hse.ru/en/ma/sna/`**

Program

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.


Who wants to learn R ?
Who wants to learn statistics?
@statsystem
Who wants to become a data scientist ?

Requirements

Home projects:

1. download, fuse and clean the data; save them 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

Time-table

The lectures will be on Zoom from 18:30 till 19:50 and from 20:00 till 21:20 (Moscow time) on the following days

October 5, 6, 8, 13, 15, 20, 22, 23, 27, 29

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 by 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 ?

- 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 16387 (October 2020) 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

Computer scientists (data mining) prefer Python instead of R. Packages: Pandas, Numpy, SciPy, NLTK, Scikt-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

- 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

- 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.