Syllabus

Data Analysis 1a: Foundation of Data management in R

- **Instructor:** Gergely Daroczi
- **Credits:** 2 (4 ECTS)
- **Term:** Fall, 2017/2018
- **Course level:** MA
- **Prerequisites:** -

Course description

This is an introductory course on how to use the R programming language and software environment for data manipulations and munging, exploratory data analysis and data visualizations.

Learning outcomes

Students will be familiar to the R ecosystem and learn how to use R for the most common data analysis tasks, including loading, cleaning, transforming, summarizing and visualizing data.

Reading list

Class materials hosted at https://github.com/daroczig/CEU-R-lab

Technical Prerequisites

If you have not filled in the Introduce Yourself survey, please do so now.

Although the required software is already installed on the computers in the School Lab, but if you plan to use your own laptop, please make sure to install the below items before attending the first class:

- R
- RStudio Desktop with Open Source License
- git

More detailed instructions and the most recent version of this document can be found on the class GitHub page. Open a GitHub ticket in case of any question.

Assessment

Assessment is through in-class quizzes (20%), exam (50%) and homeworks (30%). The weekly quizzes and the final in-class exam use the datacamp.com platform, the weekly homeworks will alternate between individual programming exercises (2nd and 4th week) hosted on Datacamp and group assignments (3rd and 6th week) with free choice of tools.

Grading policy

- Students may not miss more than 2 sessions. Failing to do so will yield an automatic Fail grade.
- To pass, students will need to get at least 50% of the overall grade AND at least 50% of the exam. Failure to do so, will yield a Fail grade

Course schedule and materials for each session

Week 1 (200 min)

General Introduction into the R Ecosystem (50 mins):

- Downloading and installing R
- History of R, R packages, CRAN
- R community, R-bloggers, StackOverflow, Coursera, DataCamp
- R User Groups & meetups

Demonstration of a Data Analysis Project in R (50 mins):

- hotel price/stars dataset

Brief Overview on R Coding Tools (25 mins):

- RStudio
- git, GitHub

R Syntax Basics (45 mins):

- Constants, operators, functions, variables
- Random numbers
- Vectors and vector indexing
- Simple descriptive stats
- Loops
- Conditional expressions

The Power of R (30 mins):

- Applying PCA on an image for outlier-detection
- Visualizing MDS on a distance matrix

Reference Datasets:

- NASA Image
- Distance Matrix of European Cities

Week 2 (200 min)

A Systematic Introduction into Data Types (50 mins)

- Levels of measurement (nominal, ordinal, interval, ratio scale)
- Vector types
- data.frame objects, rows and columns, indexing
- Characteristics of tidy data

Basic Data Transformations (50 mins):

- Create new variables in a data.frame
- Filter rows and columns
- Merging datasets

Introduction to data.table for More Complex Data Transformations (100 mins):

- Filtering and ordering data
- Summaries and aggregates
- New variables
- Relational data
- Joins on Keys
- Introduction into fuzzy joins
- Transforming wide and long tables

Reference Datasets:

- Weight and Height of Students

Week 3 (200 min)

EDA - Univariate Descriptive Statistics + crosstabs + correlation + ANOVA

EDA - First Steps with Data Visualization:

- Why not Use Pie Charts
- Plots outside of Excel: dotchart and violinplot examples
- The Grammar of Graphics in R with ggplot2
- Using labels for variable names

Week 4 (200 min)

Introduction to Non-tabular Data Types:

- Time-series
- Spatial data
- Network data

Big Data Problems:

- What is Big Data
- 4V: volume, variety, velocity, veracity

Data Transformations:

- Converting Numeric Variables into Factors
- Date Operations
- String Parsing
- Geocoding

Dirty Data Problems:

- missing values
- data imputation
- duplicates

- forms of data dates
- outliers
- spelling

Week 5 (200 min)

Data Sources:

- sqlite examples for relational databases
- Loading SPSS and SAS files
- Reading from Excel and Google Spreadsheets
- API and web scraping examples

Case Study: Who are the better CEO, men or women?

- Learning about data problems
- Managing time features
- Joing datasets
- EDA
- Cross tabulation

Reference Datasets:

- Bickel et al, 1975
- SQlite database

Week 6 (200 min)

In-class exam (90 min)

Dynamic Reports and Reproducible Research:

- Introduction to markdown
- First steps with knitr
- Markdown in R with pander
- Chunk options and document formats in rmarkdown and kintr