Syllabus

Data Analysis 1: Exploration (BA track)

Department of Economics and Business
CENTRAL EUROPEAN UNIVERSITY

- **Instructor:** Gábor Békés (bekesg@ceu.edu Office hours: Wednesday 10.00-11 and 17.00-18.00 by appointment) and Gergely Daróczi (daroczig@rapporter.net, office hours: Monday 19.10-20.00 pm by appointment)
- **Credits:** 2 (4 ECTS)
- **Term:** Fall 2018-2019
- **Course level:** [MA/MSc]
- **Prerequisites:** Presession in Mathematics and Informatics for Business Analytics
- **Course drop:** Course can be dropped free of charge 24 hours after the first session. After this date drop is possible until the course is halfway over (late drop fee applies). No changes are allowed past that date.

Course description

About 80% of data science tasks are composed of managing data, from understanding and altering features of the dataset and variables, to combining various datasets. This course introduces the critical tasks of data collection and data wrangling, presentation and understanding of descriptive statistics and basics of visualization.

One half of the course focuses on classic statistics methods and their applications, such as data collection and sampling, generalization from the sample to the population and hypothesis testing.

The other half of the course serves as 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

By successfully completing the course the students will be able to:

- Understand key issues of data gathering and manipulation
- Successfully formulate research questions that are answerable by empirical analysis;
- Produce meaningful descriptive statistics and informative graphs;
- Become 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

For the lectures,

Békés - Kézdi: Data Analysis for Business, Economics and Policy (under preparation) - available as handouts; Chapters 1-6

For the seminar
Class materials hosted at https://github.com/daroczig/CEU-R-lab
Optional:

David Salsburg Lady tasting Tea - How Statistics revolutionized science in the twentieth century

Hans Rosling, Factfulness: Ten Reasons We're Wrong About the World--and Why Things Are Better Than You Think

Assessment

- Start-of-the-class Quizzes (10%)
- Assignments (40%)
- Closed book exam (50%)

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

Technical/laptop requirement

- Students need to use their own laptops (with R and RStudio installed) during seminars.

Course schedule

Lectures

1. Origins of data (data table, data quality, survey, scraping, sampling, ethics)
2. Preparing data for analysis (tidy data, source of variation, variable types, missing data, data cleaning)
3. Describing variables (probability, distributions, extreme values, summary stats)
4. Comparison and correlation (conditional probability, conditional distribution, conditional expectation, visual comparisons, correlation, quick intro to linear regression)
5. Generalizing from a dataset (repeated samples, confidence interval, standard error estimation via bootstrap and formula, external validity)
6. Testing hypotheses (null and alternative hypotheses, t-test, false positives / false negatives, p-value, testing multiple hypotheses)

Seminars

1. R ecosystem, basic syntax, vectors, functions
2. Introduction to data frames and data.table, column types
3. Data transformations
4. Data visualization with ggplot2

5. Sampling, simulations, more exploratory data analysis
6. Hypothesis testing