Syllabus

Coding 1: Data Management and Analysis with R

- **Instructor:** Agoston Reguly

  Email: reguly_agoston@phd.ceu.edu
  Office hours: Wednesday 11:00-12:00 (online or in-person, Budapest campus)

- **Credits:** 2 (4 ECTS)
- **Term:** Fall 2021-2022
- **Module:** MS in Business Analytics
- **Course level:** [MA/MS]
- **Prerequisites:** Mathematics and Informatics Pre-session 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.

1. COURSE DESCRIPTION

The course serves as an introduction to the R programming language and software environment for data exploration, data munging, data analysis, and data visualization. As such it complements Data Analysis 1 and 2 which focus on the statistical methods, while this class focuses on how to program these methods in R.

2. LEARNING OUTCOMES

- Produce meaningful descriptive statistics and informative graphs
- Become familiar with the R ecosystem
- Learn how to use R for the most common data tasks: loading, cleaning, transforming, summarizing, and visualizing data

Other outcomes. The course will also help develop skills in the following areas:

Learning Area	Learning Outcome
Critical Thinking	Evaluate and create proper program codes which are reusable.
Quantitative Reasoning	Reasonable code which optimizes cognitive and computational effort.
Technology Skills	Become familiar with R ecosystem, use of R Markdown.
Interpersonal Communication Skills	Understand coding and informatic language, be able to communicate on these topics both online and offline.
Management Knowledge and Skills	Be able to create such reports which serves as a pillar for the management's decision.

Cultural Sensitivity and Diversity	Appreciate the different practices and solutions in coding.
Ethics and Social Responsibility	Create easily readable and reliable codes which can be used and reproduced by others as well.

3. READING LIST

Required:

- Class materials will be hosted on github.com – link will be shared at the first class.

Optional readings:

- Hadley W. - Garrett G. [**HG**] (2017): R for Data Science. Online version here.
- Kieran H. [**KH**] (2019): Data Visualization. Online version here.
- James, Witten, Hastie & Tibshirani [**JWHT**] (2013): Introduction to Statistical Learning. Online version and supplementary codes are available here.

Databases. The CEU Library boasts a range of databases covering financial and company data, market and industry reports, global news and more. For a full list of databases visit the CEU Library.

- Refinitiv (Thomson Reuters) Eikon for Students + Datastream/Thomson ONE
  - Eikon: Platform used by finance practitioners including market traders to monitor and analyze financial information. Information, analytics and news on all major financial markets including real-time pricing data, financial research, global financial news and commentary, financial estimates, fundamentals analysis, visual analysis through charting. Import/export from Excel.
  - Datastream: Range of economic, securities and company financial data. Excel add-in.
  - Thomson ONE: Global overviews on 55,000 public companies, one million private companies. Reuters News, ownership, deals, private equity, key ratios, company filings, officers and directors. Investext analyst reports, active and historical research from 1,600 independent research firms, brokerages, investment banks.
- Standard & Poor's Capital IQ
  - Web and Excel-based platform combining deep global company information, credit ratings and research, and market research with powerful tools for risk assessments. Real-time and historical information on markets, industries, companies, transactions and people. Tearsheet data.
- Lexis Nexis Academic
  - Global database of news, business, legal and other sources. Full text of 350 newspapers, 300 magazines and journals, 600 newsletters. Wire services including Associated Press, Business Wire and PR Newswire. Company financial information, market research, industry reports.

4. TEACHING METHODS AND LEARNING ACTIVITIES

The course will involve a mix of

- Lectures to present basic ideas in R.
- Live-coding classes to show and practice coding in R.
- Seminar-type classroom to solve coding problems.
- Homework to practice through problems and deepen the student's knowledge.

5. ASSESSMENT

- Assignments (50%)
    a. 6 short assignments through the semester (35%)
        i. After 1-5 and 10th session, there will be short assignment (homework) on the learned material. One need to upload the solution to their github repo, where I will check the solutions.
        ii. Deadline: before next class.
    b. Team project (15%): write a descriptive report for your dataset.
        i. Based on the created dataset from Data Analysis 1.
        ii. Deadline: 31 October Sunday 23.55 - on a github repo (can decide you create a new one or put it into one member's repo.)

- Take home examination (50%)
    a. Carry out a complete data analysis project from scratch.
        i. Joint work with Data Analysis 2.
        ii. Submission deadline: TBA.

- Demonstrate useful resources/materials (up to 10%)
    a. Reference your resource/material on slack channel
    b. Demonstrate it in class (1-3 min)
    c. Package with functions - 5%
    d. Alternative solution to problems - 2.5%

This sums to 110%, so you can lose 10% without it affecting your grade at all. There will be no final exam.

6. TECHNICAL/LAPTOP REQUIREMENT

You will need your laptop with RStudio and R installed in order to participate in class.

7. TOPIC OUTLINE AND SCHEDULE

Session	Topics	Readings
October 4 (100 min)	Introduction: Interface, tidyverse, ggplot, tryout of Rmarkdown, basics of R.	HG - Ch 1, 2, 4, 16
October 13 (200 min)	Data import and export and merging with tidy approach (readr, tibble and github). Data types and manipulations with them (numeric, string, factors and dates/times) Data cleaning, munging (dplyr): filtering and removing observations.	HG - 3, 7, 8, 9, 10, 11, 14, https://happygitwithr.com/

October 20 (200 min)	Data exploration (summary, ggplot): creating summary statistics, graphs and simple hypothesis testing. Getting know Rmarkdown: creating pdf/html/beamer. Creating tidy tables and graphs. Automatize reports with Rmarkdown.	HG – 5, 21, 22, 23, 24
October 27 (100 min)	Loops and conditionals: Writing conditional statements and loops. Writing a simple bootstrapping routine.	HG – 17, JWHT – Ch 5.2-3
November 2 (100 min)	Functions, error handling and random numbers: writing your own descriptive statistics function.	HG - 15
November 8 (100 min)	Cross-section analysis (lm,feols): running and understanding the output of a linear model in R. Variable transformations: log, interactions, splines.	JWHT - Ch 2.3, 3
November 15 (100 min)	Advanced methods in modelling: predictions, confidence intervals, hypothesis testing.	
November 24 (100 min)	Non-linear binary models: probit, logit models and their descriptives, margins.	-
November 29 (100 min)	Manipulating dates/times variables. Time series models: running and understanding the output. Model diagnostics.	-
December 8 (100 min)	Reporting models in Rmarkdown and basic spatial data and its visualization. Recap.	-

Note that readings are covering more and sometimes less than what we are actually covering in class. All necessary information will be given at the seminars.

8. SHORT BIO OF THE INSTRUCTOR

Ágoston Reguly is finishing his PhD in Economics at the Department of Economics and Business at Central European University. His main research interests are in econometrics: policy evaluation with machine learning techniques and optimal survey designs. Currently, he is working on a paper with causal trees and bandwidth selection. Also, he have taught multiple courses at CEU such as data analysis or mathematics. Before CEU he has worked for more than three years at the Hungarian Government Debt Management Agency.