Fundamentals of Data Analysis

Lecturer: Petra Kralj Novak (NovakPe@ceu.edu)
Teaching Assistant: Sandeep Chowbhary (Chowdhary_Sandeep@phd.ceu.edu)

US Credits: 2; ECTS Credits: 4
Academic Year 2021-2022 Winter Term

Mandatory for 1st year students of QSS with *Introduction to Programming in Python* as a pre-requisite. The course is one of the introductory courses of the module Programming and Data Analysis. Some later courses of the same module heavily build on the knowledge gained here, especially Introduction to Machine Learning and Data Mining, the *Data science review project* and the *Data science advanced project*.

Schedule
Classes: Fridays 10:40-12:50
Office hours: Friday 13:50 – 14:50 room D-304 (online students should write an e-mail).

Classroom
QS D-106

Zoom link:
https://ceu-edu.zoom.us/j/95111955565?pwd=Y1BUeXF1ZVpxRGFLZzZPRTN0Zkk1dz09

The aim of the course
After completing this course, the student will be able to perform basic data analysis tasks in an effective and efficient manner, and will be aware of common pitfalls and how to avoid them. The student will gain hands-on expertise with data collection, data cleaning and preprocessing. The student will be able to assess data quality, and use graphics to describe and summarize data. Furthermore, the student will be able to generalize from data and assess if the observations are significant.

Learning outcomes
Students will learn to collect appropriate data, and assess whether the available data can help answer a research question. Students will acquire the necessary skills to collect, clean and organize data. They will know the common data formats and protocols. They will be able to explore and summarize the data by using descriptive statistics and basic visualization. After taking this course, students should be able to write complex scripts for their data analytics projects going beyond the codes covered at class. The students will further develop their programming skills and be able to independently use and understand contemporary data analytics Python libraries.

Detailed content
At the beginning of each class, a theoretical concept will be explained and discussed. The second part of the class will be devoted to hands-on tasks illustrating the new concepts on both synthetic and real data. The focus will be on best practices and common pitfalls in data analytics. Help and advice from the instructor and TA will be provided. Homework assignments will consolidate the learning material and further enhance the student's coding skills.

Week 1: **Introduction to Object-oriented programming**

- Differences between procedural programming and object-oriented programming
- Defining a class, creating an object
- Member attributes
- Constructors
- Member methods
- Operator functions

Week 2: **Object-oriented programming – Part II**

- Class inheritance
- Naming conventions
- Abstract classes
- Iterators and Generators

Week 3: **Working with Pandas**

- The Pandas library intro
- Series, data frames, data types
- Indexing
- Filtering, aggregating
- Missing values
- Index alignment

Week 4: **The origins of data**

- Variable types
- Storing data: Txt, XML, JSON, html, Csv
- Big data
- Good practices of data collection
- Ethical and legal principles
- Repositories of existing data

Week 5: **Exploratory data analysis**

- Sampling
- The use of exploratory data analysis
- Frequencies and probabilities
- Distributions and their visualizations
- Extreme values
- Summary statistics and their visualization

Week 6: **Covariance, correlation**

- Independent and dependent variables
- Conditional distribution, conditional expectation
- Dependence, covariance, correlation

Week 7: **Partial exam**

Week 8: **Generalizing from data: Part I**

- Repeated samples, sampling distribution

Week 9: **Generalizing from data: Part II**

- Confidence interval
- Standard error

Week 10: **Testing Hypotheses**

- Intuition
- Null hypothesis, alternative hypothesis
- Hands on: T-test, Welch's t-test

Week 11: **Non-parametric statistics**

- Sing test, Wilcoxon signed rank test, Mann-Whitney
- Hands on examples

Week 12: **Final project presentations**

Recommended reading

Part of the course and some notebooks are based on Part I of the book
Békés, Gábor, and Gábor Kézdi. *Data Analysis for Business, Economics, and Policy*. Cambridge University Press, 2021.

Part of the course is based on the book Theodore Petrou. *Pandas Cookbook: Recipes for Scientific Computing, Time Series Analysis and Data Visualization using Python*. Packt Publishing Ltd, 2017.

All needed material will be uploaded to the Moodle site of the course.

Jupyter Notebook environment

Jupyter Notebook in the environment of Anaconda, freely downloadable here (download the version, which contains Python 3.8):
https://www.anaconda.com/download/
Setup instructions for Gabor's data analysis notebooks is available here:
https://gabors-data-analysis.com/howto-python/
and the book-related code can be found at the repository:
https://github.com/gabors-data-analysis

Assignments
Each Friday, a homework will be announced. The deadline of the submission is the following Thursday, 9.00am. Late assignments will be evaluated but are not worth homework points.

Assessment
Attendance is mandatory for at least 70 percent of the lectures. Absence from more than 30 percent of the classes automatically leads to failure.
A student's grade in this course will be a weighted average of his/her performance on the homework assignments and the partial exam and the final project. The weights are as follows:

- Performance on the homework assignments: 20%;
- Performance on the midterm exam: 40%;
- Performance on the final project: 40%.

Furthermore, there is a threshold of at least 50% of points on each of the three parts.