UGST4039 | Fundamentals of Data Analysis | Winter 2024

Instructor: Elisa Omodei (omodeie@ceu.edu; office: B-302)

Teaching Assistant: Leonardo Di Gaetano (Di-Gaetano_Leonardo@phd.ceu.edu)

Office Hours: TBD based on the students' schedule

No. of Credits: 3 US Credits / 6 ECTS

Term: Winter

Course type: Bachelor's course, BA/BSc in Data Science and Society mandatory course

Prerequisites: Introduction to Programming in Python

Brief course description

After completing this course, students 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. Students will gain hands-on expertise with data cleaning and preprocessing. They will be able to assess data quality, and use graphics to describe and summarize data. Furthermore, students will be able to generalize from data and assess if the observations are significant.

Learning outcomes

Students will learn to retrieve, clean and organize data, and assess whether the available data can help answer a research question. 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, and to perform basic statistical inference and correctly interpret its results. After taking this course, students should be able to write complex scripts for their data analysis projects going beyond the codes covered at class. They will further develop their programming skills and be able to independently use and understand contemporary data analysis Python libraries.

Learning activities and teaching methods

Each week will feature a two-hour block taught by the instructor on Monday followed by a one hour block later in the week taught by the TA on Thursday. The first block will consist of theoretical lectures and some practical examples, whereas the second block will consist of hands-on tutorials where the students will learn how to put into practice the concepts and methods introduced in the lectures.

Detailed Content

- Week 1: Working with data (data collection, data formats, data repositories)
- Week 2: Introduction to Pandas, Python's data analysis library
- Week 3: Data cleaning and preparation
- Week 4: Descriptive statistics
- Week 5: Plotting and Visualization
- Week 6: Modeling and Fitting Distributions (Part 1)
- Week 7: *Students presentations and feedback session*
- Week 8: Modeling and Fitting Distributions (Part 2)
- Week 9: Confidence intervals and error bars
- Week 10: Hypothesis about the mean of normal populations
- Week 11: Relationship between variables and Linear Regression
- Week 12: *Students presentations and feedback session*

Attendance policy

Regular in-person class attendance is a precondition for course completion.

- Students who miss more than **three 1 hour classes unexcused** cannot receive a passing grade.
- Students who miss more than twelve 1 hour classes in total (including both excused and unexcused absences) cannot receive a passing grade.

In justified cases, permission for absence should be requested in advance:

- Students must fill out the excused absence form (the link can be found in the Student Handbook).
- A member of the BA team will reply to the student and confirm whether the absence has been approved.
- **The student is then responsible for writing to the instructor and the TA (with absenceBA@ceu.edu copied in).**
- In the event of an unexpected last-minute ability to attend class, students should make a good faith effort to communicate this to their professors in advance of the class or as soon as possible. Students should inform both absenceBA@ceu.edu and the instructor and the TA via email.

For singular, non-excused absence (e.g. personal appointments, non-urgent family reasons), students do not need to fill out the form, but must inform the instructor and the TA directly. There is no need to provide reasons for non-excused absences.

Assessment, Evaluation and Grading Criteria

Course assessment will be performed by means of a group project that will consist of the exploration, in python, of an open data set through an appropriate selection of the exploratory data analysis and statistical inference methods introduced in class.

The project is to be fulfilled in groups of 3-4 students, making it clear which part of the project was carried out by which student.

The project will consist of the following assignments, each determining a given percentage of the final grade as indicated in parenthesis:

- **Monday, February 19th**: Oral presentation on the exploratory data analysis part of the project (15% of the final grade)
- Written report (max 4 pages; a template will be provided) describing the exploratory data analysis (30% of the final grade), due by **Sunday, February 25th.** The code used to produce the results, tables and figures included in the report is also to be submitted as an appendix.
- **Monday, March 25th**: Oral presentation on the statistical inference part of the project (20% of the final grade)
- Final written report (max 8 pages; a template will be provided) (35% of the final grade), due by **Sunday, April 7th.** The code used to produce the results, tables and figures included in the report is also to be submitted as an appendix.

The use of AI tools (e.g., ChatGPT) is strongly discouraged in these early stages of learning. The students can be asked by the instructor at any time to explain their own submitted code. Failure to convincingly justify one's own code will result in the removal of the points assigned to the corresponding parts of the project.

Grading criteria

Oral presentations

- Presentation structure and content organization (clarity of research questions and hypotheses; description and justification of chosen data and methods; interpretation of results and appropriateness of drawn conclusions; identification of limitations)
- Slide design (text readability, use of visuals, consistent style, etc.)
- Clarity of exposition and ability to engage with the audience
- Mastering of the material (during both the presentation and the Q&A)
- Time management

Written reports

- Contextualization and motivation
- Clarity of research questions and hypotheses
- Description and justification of data and methods (including assumption checks)
- Interpretation of results and appropriateness of drawn conclusions
- Identification of limitations and opportunities for future work
- Report structure and respect of the page limit
- Quality of writing
- Incorporation of feedback received during the oral presentation