UGST4039 | Fundamentals of Data Analysis | Winter 2024

Course Title: Fundamentals of Data Analysis

Instructor(s): Leonardo Di Gaetano (Di-Gaetano_Leonardo@phd.ceu.edu, office D305)

University, department, level, year of studies : Bachelor's course, BA/BSc in Data Science and Society , 1st and 2nd year

Course type: mandatory course

Other related courses in the curriculum, prerequisites: Introduction to Programming in Python

Course Description This course is designed as an hands-on introduction to data analysis, emphasizing practical skills in data handling, cleaning, and visualization using computer-based tools. Over a period of 12 weeks, participants will engage in a structured curriculum that combines lectures, seminars, and practical sessions to explore the fundamentals of data analysis. The course covers a wide range of topics, including data collection methods, data formats, data repositories, and the use of Python's Pandas library for data analysis. Special focus is given to the process of data cleaning and preprocessing to ensure data quality, as well as the application of statistical methods to describe, summarize, and infer from data sets.

Participants will benefit from a computer-based, interactive learning environment where they can apply theoretical concepts in real-world scenarios. Each week introduces a new aspect of data analysis, from basic data handling to advanced techniques in statistical analysis and linear regression. The course format encourages active participation through group work, student presentations, and feedback sessions, allowing students to consolidate their learning and apply it to practical data analysis tasks. By the end of the course, participants will have developed a comprehensive skill set that prepares them for further studies in data science or immediate application in their professional lives.

Course Aims: The purpose of this course is to equip students with a comprehensive understanding of data analysis fundamentals and to introduce intermediate techniques that prepare them for advanced study and practical application in the field of data science. This course aims to bridge the gap between basic data handling and more sophisticated statistical analysis methods, ensuring a solid foundation for further exploration and professional development.

Learning outcomes:

After completing this course, students will be able to:

- Retrieve, clean, and organize data to determine its utility in answering research questions.

- Identify and work with common data formats and protocols, ensuring compatibility and efficiency in data analysis processes.
- Utilize descriptive statistics and basic visualization techniques to explore and summarize data, providing clear insights into its underlying patterns and trends.
- Perform basic statistical inference and correctly interpret the results, facilitating informed decisions and robust analyses.
- Write complex scripts for data analysis projects, extending beyond the examples and codes covered in class, demonstrating an advanced understanding of programming concepts.
- Independently use and understand contemporary data analysis libraries in Python, further developing their programming skills and enhancing their ability to tackle complex data analysis challenges.

Mapping Your Course

Detailed Content

Week #	Title/Major Topics	Learning Outcomes	Assessment/Assignment	Teaching/Learning Methodologies	Prepare/Engage/Consolidate	Resources
1	Data Collection, Formats, Repositories	Understand different data collection methods, familiarize with data formats, identify appropriate data repositories.	Quiz on data formats and repositories	Lecture, Seminar	Prepare: Read introductory materials on data science. Engage: Class discussion. Consolidate: Quiz.	Mandatory: Articles on data collection methods. Optional: Video tutorials on using repositories.

2	Introduction to Pandas	Gain basic proficiency in Pandas for data manipulation and analysis.	Homework: Data manipulation task using Pandas	Lecture, Practical session	Prepare: Tutorial on Pandas basics. Engage: Hands-on session. Consolidate: In class discussion to wrap up (10 minutes). Homework.	Mandatory: Pandas documentation. Recommended: Online Pandas tutorials.
3	Data Cleaning and Preparation	Learn techniques for cleaning and preparing data for analysis.	Project: Clean and prepare a dataset	Seminar, Group work	Prepare: Read about data cleaning techniques. Engage: Group activity. Consolidate: In class discussion to wrap up (10 minutes).	Mandatory: Textbook chapter on data preparation. Optional: Cleaning data case studies.
4	Descriptive Statistics	Understand and apply measures of central tendency and variability.	Assignment: Compute descriptive statistics for a dataset	Lecture, Practical session	Prepare: Study statistical measures. Engage: In-class exercises. Consolidate: Video on visual intuition on descriptive statistics to wrap up (10 minutes).	Recommended: Statistics textbooks. Optional: Statistics software tutorials.
5	Plotting and Visualization	Develop skills in data visualization using various chart types.	Project: Create a data visualization portfolio	Seminar, Practical session	Prepare: Review visualization principles. Engage: Visualization workshop. Consolidate: Plotting together, in class activity (10 minutes).	Mandatory: Visualization tool documentation. Recommended: Visualization design principles.
6	Distributions	Understand different types	Quiz on distribution types and applications	Lecture, Seminar	Prepare: Read about probability distributions.	Recommended: Probability and

		of distributions and their applications.			Engage: Discussion. Consolidate: Quiz.	statistics textbooks.
7	Students Presentations and Feedback Session	Enhance presentation skills and receive feedback on progress.	Presentation: Individual or group project	Student presentation, Feedback session	Prepare: Develop presentation. Engage: Presentation and peer review. Consolidate: Feedback.	Optional: Presentation skills workshops.
8	Relationships between Variables	Learn about correlation and causation, and how to analyze relationships between variables.	Assignment: Analyze relationships in a given dataset	Lecture, Practical session	Prepare: Study correlation and regression. Engage: Data analysis exercises. Consolidate: Assignment.	Mandatory: Textbook chapters on correlation and regression.
9	Confidence Intervals and Error Bars	Understand the concepts of confidence intervals and error bars in data analysis.	Quiz on confidence intervals and error bars	Lecture, Seminar	Prepare: Review statistical inference concepts. Engage: Problem-solving session. Consolidate: Quiz.	Recommended: Articles on statistical inference.
10	Hypothesis about the Mean of Normal Populations	Learn hypothesis testing for the mean of normal populations.	Project: Conduct hypothesis testing on a dataset	Seminar, Group work	Prepare: Read about hypothesis testing. Engage: Group discussion. Consolidate: In class discussion to wrap up (10 minutes)..	Mandatory: Statistical hypothesis testing textbooks. Optional: Case studies.

11	Linear Regression	Acquire knowledge on linear regression analysis and its applications.	Assignment: Perform a linear regression analysis	Lecture, Practical session	Prepare: Study linear regression models. Engage: Regression analysis exercises. Consolidate: In class discussion to wrap up (10 minutes)..	Mandatory: Textbook on regression analysis. Recommended: Regression software tutorials.
12	Students Presentations and Feedback Session	Apply knowledge gained throughout the course in presentations, receive final feedback.	Presentation: Final project	Student presentation, Feedback session	Prepare: Finalize projects. Engage: Final presentations and peer review. Consolidate: Final feedback.	Optional: Articles on effective communication and feedback.

Report Evaluation Criteria

Criteria	Percentage	Description
Introduction and Background	10%	Clarity in project objectives and dataset choice justification.
Data Collection and Cleaning	15%	Methodology and choices for data cleaning, and preprocessing.
Plots and Visualizations	15%	Quality and effectiveness of visual data representations.
Preliminary Analysis Methodology	15%	Depth and originality in analysis methods applied.
Results Interpretation	15%	Insightfulness in presenting and interpreting results.
Coverage of Course Topics	15%	Application breadth of Python functions and analytical methods taught.
Organization, Structure, and Code Quality of the Jupyter Notebook	15%	Logical structuring, code efficiency, and readability of the notebook.

Oral Presentation Evaluation Criteria

Criteria	Percentage	Description
Data Collection and Cleaning	10%	Methodology and choices for data cleaning, and preprocessing.
Plots and Visualizations	10%	Quality and effectiveness of visual data representations.
Preliminary Analysis Methodology	10%	Description and justification of chosen data and methods;
Results Interpretation	10%	Insightfulness in presenting and interpreting results.

Coverage of Course Topics	10%	Application breadth of Python functions and analytical methods taught.
Presentation structure, narrative and content organization	15%	Clarity of research questions, hypotheses, conclusions and limitations
Slide design	5%	Text readability, use of visuals, consistent style, etc.
Oral exposition	10%	Clarity of exposition and ability to engage with the audience
Q&A	10%	Ability to answer to questions and critics to your analysis
Time management	10%	Being able to present the project in time (10 min)