STAT 1281/2280: Data Science with Python

(Summer 2025)

Basic Information

- Instructor: Junshu Bao, Teaching Associate Professor, Department of Statistics
  - Email: jub69@pitt.edu
  - Webpage: https://sites.pitt.edu/~jub69/resources.html
  - Office Hours: Monday 4 PM – 5 PM
  - Zoom link for OH: https://pitt.zoom.us/j/7550746319
- Teaching Assistant: Marshall Honaker
  - Email: mah733@pitt.edu
  - Office Hours: Tuesday 1 PM – 2 PM
  - Zoom link for OH: https://pitt.zoom.us/j/5884917250
- Meeting time: M/T/W/TH 10 AM – 11:15 AM

Prerequisites

- Statistics knowledge at the level of STAT 1000 or above.
- Students who have had STAT 1221, STAT 1151, and STAT 1152 will be better prepared for the topics about machine learning.
- No prior knowledge of programming is required.

Course Overview

This course covers popular topics and widely used tools and techniques in data science. You will learn how to import data, tidy and transform it, visualize it, and how to join data sets. You will also learn about training, tuning, and testing various machine learning models, and ultimately generating reproducible reports.

Python as well as a collection of powerful, open-source tools will be explored and experienced within the context of solving data science problems:

- Jupyter notebook (creating reports)
- numpy (data structure)
- pandas (data wrangling)
- matplotlib & seaborn (data visualization)
- scikit-learn (data preprocessing & machine learning)
- tensorflow & scipy (machine learning)

Responsible uses and understanding of both the capabilities and limitations of data science tools will be stressed throughout the course.

Learning Objectives

- Workflow of Data Science
- Data Wrangling and Visualization
- Machine Learning Models
- Python and Jupyter Notebook

By the end of the course, you should be able to get the data, explore it, formulate a research question, use tools and techniques in data science to explore the answer to the question, and share your findings.

Textbooks

- *Python Data Science Handbook* by Jake VanderPlas
- *Machine Learning with Python Cookbook* by Chris Albon
- *Introduction to Machine Learning with Python* by Andreas C. Müller and Sarah Guido
- *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* by Aurelien Geron
- *A Whirlwind Tour of Python* by Jake Vanderplas

These books are either free or available in our Pitt library system.

Course Grades:

Grade	Percentage	Grade	Percentage
A+	[99%,100%]	C+	[77%,80%)
A	[93%,97%)	C	[73%,77%)
A-	[90%,93%)	C-	[70%,73%)
B+	[87%,90%)	D+	[67%,70%)
B	[83%,87%)	D	[63%,67%)
B-	[80%,83%)	D-	[60%,63%)
		F	[0,60%)

Course Management System: Canvas

- Lecture notes
- Reading material
- Homework assignments and exams
- Data sets

Grading Components

1. Homework Assignments (40%)

- Homework will be assigned weekly.
- Homework questions require coding in Python.
- Homework assignments may be done in collaboration with other students. However, the final product must be written by you, in your own words. This applies to Python code too – share ideas but write your own code.
- Anything handed in after the due date and time will be applied a late penalty, unless cleared well in advance. A 20% penalty will be applied if it is submitted within 24 hours of the due date and time. No late submission will be accepted after 24 hours.
- I expect that you will start soon after receiving the assignment. The assignments are definitely not designed to be one-night jobs. You must show your work for full credit.
- Homework is completed using Jupyter notebook and you are required to upload a knitted HTML, Word, or PDF file to Canvas.

2. Quizzes (20%)

- There will be four quizzes throughout the semester.
- Quiz questions may require coding in Python.

3. Exams (40%)

- Exam 1: 20%
- Exam II: 20%

Course Outline

The course material will be organized as follows:

- Module 1: Python Basics
  - The Very Basics
  - Basic Python Semantics
  - Data structure: Vectors, Matrices, and Arrays
  - Data Import
  - Conditionals, Loops, and Functions
- Module 2: Data Management
  - Data Wrangling
  - Data Preprocessing/Transformation
    - Handling Numerical Data
    - Handling Categorical Data
- Module 3: Data Visualization
  - Graphing Basics
  - Scatterplots, Density Plots, and Histograms
  - More Graphing Details

- Module 4: Supervised Learning
  - Introduction and Primer
  - Linear Models for Regression
  - Linear Models for Classification
  - K-Nearest Neighbors
  - Decision Trees
  - Random Forests and Ensemble Learning
  - Naive Bayes
  - Support Vector Machines
  - Neural Networks

- Module 5: Unsupervised Learning
  - Dimension Reduction
  - Clustering
    - K-Means Clustering
    - DBSCAN
    - Agglomerative Clustering
    - Hierarchical Clustering

University Policies:

Academic Integrity

Students in this class will be expected to comply with the University of Pittsburgh's Policy on Academic Integrity. Any student suspected of violating this obligation for any reason during the semester will be required to participate in the procedural process, initiated at the instructor level, as outlined in the University Guidelines on Academic Integrity. This may include, but is not limited to, the confiscation of the examination of any individual suspected of violating University Policy. Furthermore, no student may bring any unauthorized materials to an exam, including dictionaries and programmable calculators.

To learn more about Academic Integrity, visit the Academic Integrity Guide for an overview of the topic. For hands-on practice, complete the Understanding and Avoiding Plagiarism tutorial.

Copyright Notice

Statement on Classroom Recording

To ensure the free and open discussion of ideas, students may not record classroom lectures, discussion and/or activities without the advance written permission of the instructor, and any such recording properly approved in advance can be used solely for the student's own private use.

Diversity and Inclusion

The University of Pittsburgh does not tolerate any form of discrimination, harassment, or retaliation based on disability, race, color, religion, national origin, ancestry, genetic information, marital status, familial status, sex, age, sexual orientation, veteran status or gender identity or other factors as stated in the University's Title IX policy. The University is committed to taking prompt action to end a hostile environment that interferes with the University's mission. For more information about policies, procedures, and practices, see: http://diversity.pitt.edu/affirmative-action/policies-procedures-and-practices.

I ask that everyone in the class strive to help ensure that other members of this class can learn in a supportive and respectful environment. If there are instances of the aforementioned issues, please contact the Title IX Coordinator, by calling 412-648-7860, or e-mailing titleixcoordinator@pitt.edu. Reports can also be filed online: https://www.diversity.pitt.edu/make-report/report-form. You may also choose to report this to a faculty/staff member; they are required to communicate this to the University's Office of Diversity and Inclusion. If you wish to maintain complete confidentiality, you may also contact the University Counseling Center (412-648-7930).

Disability Services

If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact both your instructor and Disability Resources and Services (DRS), 140 William Pitt Union, (412) 648- 7890, drsrecep@pitt.edu, (412) 228-5347 for P3 ASL users, as early as possible in the term. DRS will verify your disability and determine reasonable accommodations for this course.

Accessibility

The Canvas LMS platform was built using the most modern HTML and CSS technologies and is committed to W3C's Web Accessibility Initiative and Section 508 guidelines. Specific details regarding individual feature compliance are documented and updated regularly.