Data Science Applications and Analysis

PSTAT 100

Spring 2021

Instructor:	Trevor Ruiz (tdr@ucsb.edu)
Teaching assistants:	TBD
Office hours:	TBD

Course information

Description (from catalog)

Overview and use of data science tools in Python for data retrieval, analysis, visualization, reproducible research and automated report generation. Case studies will illustrate practical use of these tools. This new course will focus on concepts that are relevant for data science by using some of the popular software tools in this area. Doing data science is more than using isolated methods. Creatively using a collection of concepts and domain knowledge is emphasized to clean, transform, analyze, and present data. Concepts in data ethics and privacy will also be discussed. Case studies will illustrate real usage scenarios. Prerequisites: Probability and Statistics I (PSTAT 120A), Linear Algebra (MATH 4A), and prior experience with Python or another programming language (CMPSC 8 or equivalent). Credit units: 4.

Audience and goals

This course is a hands-on introduction to data science intended for intermediate-level students from any discipline with some exposure to probability and basic computing skills, but few or no upper-division courses in statistics or computer science. The course introduces central concepts in statistics – such as sampling variation, uncertainty, and inference – through an applied and computational lens alongside techniques for data exploration and analysis. Course activities model standard data science workflow practices by example, and successful students acquire programming skills, project management skills, and subject exposure that will serve them well in upper-division courses as well as in independent research or projects.

Format

Prerecorded lecture and lab segments will be delivered asynchronously, and lab sections will be held during scheduled times as office hours via Zoom. Asynchronous communication will be facilitated on Nectir, and the course GauchoSpace page will link to all course content and resources. The class will progress according to the following weekly schedule.

- ***Mondays*** at 9am: reading, lectures, and lab released with weekly announcement.
- ***Fridays*** at 5pm: lab activity due; homeworks released/due biweekly.

Materials

Readings for the course will draw on multiple sources, including in particular the Python Data Science Handbook and Berkeley's Data 8 Inferential Thinking and Data 100 Principles and Techniques of Data Science textbooks, all available online. Computing will be hosted on an LSIT server (link to be provided).

Tentative schedule

The tentative weekly lecture schedule is indicated below and subject to change based on the progress of the class.

Week	Lecture Topic(s)	Lab	Assessments
1	Data science lifecycle	Jupyter notebooks	
2	Sampling and inference	Summary statistics and simulation	
3	Data wrangling and tidy data	Pandas	HW1 due
4	Elements of data visualization	Plot types and aesthetics	
5	Exploratory analysis I	Density estimation	HW2 due
6	Exploratory analysis II	Dimension reduction	
7	Statistical models I	Simple linear regression	HW3 due
8	Statistical models II	Multiple regression	
9	Classification	Project workflow	HW4 due
10	TBD	TBD	
11	None (finals week)	None	Project due

Learning outcomes

In this course, students will:

1. Simulate, retrieve, organize, summarize, and visualize, and model data using scientific computing tools in Python.

2. Practice critical thinking about the relationship between data collection and scope of inference, and assess the plausibility of assumptions required to meaningfully model real data.

3. Use appropriate programming style, conventions, and practices to write readable, organized, and reproducible codes.

4. Demonstrate good data science workflow and communication practices through completing a collaborative data analysis project and preparing a written summary of results.

Assessments

Your attainment of course learning outcomes will be measured by the following assessments, with the relative weighting for final grade calculations indicated in parentheses. All assessments within each category are given equal weight.

- **Labs** (40%). Labs will be given weekly. These are structured coding assignments with small exercises throughout that introduce the programming skills needed to complete homework assignments. The TAs will review the labs in posted videos to help you complete the exercises, and you will be responsible for turning in the completed lab by the end of the week (Friday at 5pm PST). Submissions will be graded out of 10 points each.

- **Homeworks** (40%). Homeworks will be assigned biweekly. These are fairly involved assignments in which you'll apply concepts and techniques from the lectures and programming skills from the labs to real and simulated data sets. Collaboration is encouraged, and group submissions will be allowed for groups of at most 3 students. Homeworks will be graded out of 50 points each.

- **Project** (20%). There will be one final project consisting of an open-ended data analysis that you will complete with a partner or in a group of 3. Details will be released around Week 8, and the project will be due by the end of the scheduled final exam time for the course. The project will be graded out of 30 points.

Course Policies

Communication

There are four means of communication with the instructor, TAs, and other students: Nectir, office hours, email, and (Zoom) appointments. Please use them in that order of priority; email and appointments should not be used to discuss course material.

1. **Nectir**. Consider Nectir as your primary communication resource for the course — this will be our virtual classroom and your way to stay connected with the instructor, the TAs, and your classmates throughout the term. You can start and participate in threaded conversations in the group chat, create discussions for specific purposes as you see fit (*e.g.*, forming a study group), and exchange direct messages with anyone in the class. The instructor and TAs will monitor each page as well as their direct messages daily during online hours, so posts and messages shared and sent via Nectir are the fastest way to interact with the group and resolve questions. You are encouraged to participate actively — the instructor and TAs will rely on Nectir conversations to get to know each of you and gauge how the class is doing, and your fellow students will benefit from your engagement and contributions.

2. **Office hours**. Office hours will be offered once weekly on Zoom by the instructor. These are opportunities to interact informally in real time and discuss course material or assignments. While the format is flexible, students will be encouraged to suggest discussion topics via the chat upon entry.

3. **Email**. Please use email with discernment for simple communication regarding personal matters (*e.g.*, needs for special accommodations due to medical or other emergencies). Please refrain from communicating about course material via email. A response is guaranteed within 48 weekday hours (so if you email on Friday afternoon, you may not receive a reply until Tuesday afternoon). In light of this response policy, bear in mind that you are likely to receive replies to messages or posts in Nectir much faster than replies to email. If your message is time-sensitive, please indicate so in the subject and we will do our best to respond promptly.

4. **Appointment**. You can schedule 20-minute Zoom appointments with the instructor via Calendly as needed. If you schedule an appointment, you will be prompted to indicate what you wish to discuss. Due to the course enrollment appointments will not be granted for private instruction/review/discussion of course material; any such appointments will be cancelled.

Expected time commitment

The course is 4 credit units; each credit unit corresponds to an approximate time commitment of 3 hours. So, expect to allocate 12 hours per week to the course on average. Bear in mind that homework assignments will be labor-intensive, so you may find yourself spending only a few hours (say 6) one week and many more the following week (say 18). If you find yourself spending considerably more than 12 hours on the course on a regular basis, please let the instructor or TAs know so that we can help you balance the workload.

Grades

Your overall grade in the course will be calculated as the weighted average of the proportions of total possible points in each assessment category according to the weightings indicated in the Assessments section and reported as a percentage rounded to two decimal places. Tentatively, letter grades will be assigned according to the rubric below – a curve is possible, but not guaranteed.

A	93% – 100%
A-	90% – 92.99%
B+	87% – 89.99%
B	83% – 86.99%
B-	80% – 82.99%
C+	77% – 79.99%
C	73% – 76.99%
C-	70% – 72.99%
D+	67% – 69.99%
D	60% – 66.99%
F	0% – 59.99%

You can keep track of your marks on individual assessments, your marks in each assessment category, and your overall grade in the Gauchospace gradebook. Please notify the instructor or TAs of any errors in grade *entry*; please do not attempt to negotiate the grades themselves. If at the end of the course you believe your grade was unfairly assigned, you are entitled to contest it according to the procedure outlined here in the UCSB General Catalog.

Conduct

Please be especially mindful of maintaining respectful and kind communication. Bear in mind that this is much more difficult with written communication, and consider carefully how your words might be received by others. Just as in a classroom, you are expected to uphold the UCSB student code of conduct in your online behavior. You can find the student code of conduct on the Office of Student Conduct website from this page. If you are uncomfortable with the online conduct of another participant for any reason, please notify the instructor or TAs.

Academic integrity

Please maintain integrity in learning. Your work in the course must be your own. Any form of plagiarism, cheating, misrepresentation of individual effort on assignments and assessments, falsification of information or documents, or misuse of course materials compromises your own learning experience, that of your peers, and undermines the integrity of the UCSB community. Any evidence of dishonest conduct will be discussed with the student(s) involved and reported to the Office of Student Conduct. Depending on the nature of the evidence and the violation, penalty in the course may range from loss of credit to automatic failure. For a definition and examples of dishonesty, a discussion of what constitutes an appropriate response from faculty, and an explanation of the reporting and investigation process, see the OSC page on academic integrity.

Late work

Homeworks submitted within 48 hours of the deadline will be evaluated for 75% credit; homeworks submitted more than 48 hours late will not be accepted.

Every student can submit two late labs without penalty. Otherwise, labs submitted within 48 hours of the deadline will be evaluated for 50% credit, and no submissions will be accepted more than 48 hours late.

No late project submissions will be accepted.

Extensions due to personal circumstances will be considered but should be arranged in advance of relevant deadlines.

Accommodations

Reasonable accommodations will be made for any student with a qualifying disability. Such requests should be made through the Disabled Students Program (DSP). More information, instructions on how to access accommodations, and information on related resources can be found on DSP website. Remote learning may present unique accommodation needs requiring additional flexibility; students receiving accommodation via DSP are invited to discuss this with the instructor if desired.

Student evaluation of teaching

Toward the end of the term you will be given an opportunity to provide feedback about the course via ESCI. Your suggestions and assessments are essential to improving the course, so please take the time to fill out the evaluations thoughtfully.