Data Science Concepts & Analysis

PSTAT 100

Winter 2022

Instructor:	Trevor Ruiz (tdr@ucsb.edu)
Teaching assistants:	Chao Zhang (czhang@pstat.ucsb.edu)
	TBA
Learning assistants:	Annie Huang (anniehuang@ucsb.edu)
	Dante Coletta (dcoletta@ucsb.edu)
	Priyanka Banerjee (pbanerjee@ucsb.edu)
Office hours:	TBA

Course information

Description (from catalog)

Overview of data science key concepts and the use of tools for data retrieval, analysis, visualization, and reproducible research. Topics include an introduction to inference and prediction, principles of measurement, missing data, and notions of causality, statistical "traps", and concepts in data ethics and privacy. Case studies will illustrate the importance of domain knowledge. Prerequisites: Probability and Statistics I (PSTAT 120A), Linear Algebra (MATH 4A), and prior experience with Python or another programming language (CMPSC 9 or CMPSC 16). 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 – in an applied setting together with 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 will be delivered asynchronously. (Lab sections may resume in person later in the quarter depending on university policy.) 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: release date for reading, lectures, and lab; weekly announcement.
- ***Fridays*** at 5pm: due date for all assessments; labs due weekly; homeworks due biweekly.

Materials

Readings for the course will draw on multiple sources, including in particular the Python Data Science Handbook and Berkeley's Data 100 Principles and Techniques of Data Science textbooks, both available online. Computing will be hosted at pstat100.lsit.ucsb.edu.

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
0	Course introduction	NumPy review	
1	Principles of tidy data	Pandas	
2	Data collection and sampling	Simulated sampling designs	HW1 due
3	Data visualization	Data viz in Altair	
4	Exploratory analysis I	Smoothing	HW2 due
5	Exploratory analysis II		
6	Statistical models I	Principal components	HW3 due
7	Statistical models II	Regression	Project plan due
8	Case study	Report preparation	HW4 due
9	Closing		
10	Finals week		Project report due

Learning outcomes

In this course, students will:

1. Establish foundational data science workflow practices: critical assessment of data quality and sampling design; inspection and tidying of raw data; exploratory, descriptive, visual, and inferential analysis; and interpretation and communication of results.

2. Import, organize, summarize, and visualize, and model data using scientific computing tools in Python.

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

4. Carry out case studies using real data sets and practice clear communication and proper interpretation 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 in most weeks. These are structured coding assignments with small exercises throughout that introduce the programming skills needed to complete homework assignments. 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 data sets in order to reproduce an analysis and answer substantive questions. 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 course project consisting of an open-ended data analysis that you will complete with a partner or in a small group. You will prepare an interim report around the midpoint of the term identifying a dataset and project plan, and prepare a final report of work and findings toward the end of the quarter. Each report will be graded out of 50 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, 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 on a weekly basis via Zoom by both the instructor and the TAs. These are opportunities to interact informally in real time and discuss course material or assignments.

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.

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 6-8) one week and many more the following week (say 15-18). 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; letter grades will be assigned based on this overall score.
You can keep track of your marks on individual assessments in Gradescope. Please notify the instructor or TAs of any errors in grade entry or discrepancies in assessment; otherwise, 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. You are encouraged to collaborate in this course, but all submitted work 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

Late work will not be accepted beyond 72 hours after the deadline for any assignment without prior approval. There is a one-hour grace period on all submission deadlines. After that, work submitted within 72 hours of the deadline will be evaluated for 75% credit.

Every student can submit two late assignments within 72 hours of the deadline without penalty. After the first two late submissions, the 75% penalty will be applied.

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.

In addition, content-specific feedback will be collected for the Central Coast Data Science Partnership (CCDSP) at the end of the quarter. This information will be used to assess learning outcomes, understand student demographics, and plan further course development; your input on the CCDSP survey is especially valuable, and a small amount of course credit will be offered for completion of this survey.