Statistical Data Science

PSTAT 134/234

Fall 2024

Instructor:	Dr. Katie Coburn (katie_m_coburn@ucsb.edu)
Teaching Assistants:	Mengye Liu (mengye@ucsb.edu)
	Lihao Xiao (lihao@ucsb.edu)
Lectures:	TR 3:30 - 4:45 pm CHEM 1171
Sections:	W 10:00 - 10:50 am PHELP 1525 Liu
	W 11:00 - 11:50 am PHELP 1525 Liu
	W 12:00 - 12:50 pm PHELP 1525 Xiao
	W 1:00 - 1:50 pm PHELP 1525 Xiao

Course information

Description

Applications of advanced data science tools for data retrieval, statistical analysis and machine learning, optimization, and visualization. Multiple case studies will illustrate the practical use of these tools. Prerequisites: PSTAT 131 or 231 or CS 165B and CS 9 (preferred) or CS 16, all with a minimum grade of C or better. Credit units: 4.

This course is taught at two levels, one aimed at undergraduates (134 level) and one at graduate students (234 level). Lectures are given at a level intended to be accessible and relevant to all students in the course. Students taking the course at the 234 level will be assigned **additional homework questions**. The final project assignment is the same for both levels.

Attendance

Lectures and lab sections will be conducted in person only. We will not record lectures or labs. You are strongly encouraged to attend **all** lectures and lab sections; choosing not to attend will result in your missing the (potentially very important) content that was presented or discussed.

If you choose not to attend a lecture or section, you are still expected to complete the assignments and readings. If you are unable to attend for any **university-sanctioned reason**, please inform the instructor and your TA(s) as soon as possible.

Format

Lectures and sections will be provided **in person only**. You are expected to attend every lecture and section. Lecture slides will be posted on Canvas before the corresponding class. The class will progress according to a weekly schedule (provided below). Both lecture periods each week will consist of slides that cover related concepts, sometimes including live coding to demonstrate how to apply the concepts.

Section each week will be dedicated to completing an R- and Python-based lab. Section attendance is recorded and is **very strongly** recommended, since doing so will enable you to learn and practice your coding skills in an environment where you can easily obtain help from the TA(s) and your peers.

Should you need to switch sections, please ask the TA of the section you want to switch to if there is room for you to attend. If the TA allows, you can then begin attending the new section while remaining officially enrolled in your previous section. **Sections will be held beginning the week of September 30th**.

Required Materials

The required materials used for this class are as follows. Note that all are freely available online and do not require purchase.

- **Required Books:**
  - *R for Data Science* Second Edition, by H. Wickham, M. Cetinkaya-Rundel, G. Grolemund. Available: https://r4ds.hadley.nz/
  - *Python Data Science Handbook*, by J. VanderPlas. Available here
- **Required Software:**
  - The R statistical environment version 4.4.1, available here: https://www.r-project.org/
  - The RStudio IDE (integrated development environment) Desktop version. Available: https://www.rstudio.com/
  - The Python programming language version 3.12 or higher, available here: https://www.python.org/downloads/

If you prefer, rather than installing the required software manually onto your personal computer, you can utilize our course JupyterHub at the following link: https://github.com/katiecoburn/pstat-134-234. More information about this is available on Canvas.

Supplemental Materials

The following are materials that are **not required** for this course, but are **recommended**. Do not feel expected to read all of these; instead I encourage you to use them as reference manuals and look subjects up as necessary. You will not be quizzed on any readings from these books unless directly specified.

- *Python for Data Analysis* Third Edition, by W. McKinney. Available: https://wesmckinney.com/book/
- *The StatQuest Illustrated Guide to Machine Learning*, by J. Starmer. Available on Amazon: $27 for paperback. Based on videos available free at https://www.youtube.com/@statquest
- *Supervised Machine Learning for Text Analysis in R*, by E. Hvitfeldt and J. Silge. Available: https://smltar.com/
- *Feature Engineering and Selection*, by M. Kuhn and K. Johnson. Available: http://www.feat.engineering/
- *R Markdown Cookbook*, by Y. Xie, C. Dervieux and E. Riederer. Available: https://bookdown.org/yihui/rmarkdown-cookbook/
- *Tidy Modeling with R*, by M. Kuhn and J. Silge. Available: https://www.tmwr.org/
- *Text Mining with R: A Tidy Approach*, by J. Silge and D. Robinson. Available: https://www.tidytextmining.com/
- *Natural Language Processing with Python*, by S. Bird, E. Klein, and E. Loper. Available: https://www.nltk.org/book/

Learning Outcomes

Upon completion of this course, students should be able to:

1. Demonstrate the ability to use appropriate statistical methodologies for real-world data analysis settings;

2. Apply a variety of technological tools, such as statistical software and computer programming languages, for the purpose of statistical data analysis;
3. Communicate data science concepts and reasoning in written reports and oral presentations using both technical and non-technical language;
4. Participate effectively in teams to accomplish the goals of a data science project.

Assessments

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

Collaboration and open science are fundamental aspects of data science. You can (and in fact are encouraged) to use the Internet, StackOverflow, etc. as a resource when coding. However, you **should not** directly copy code or answers. Your code should be well documented; include comments describing each line. If you use a source for help coding, you **must** cite it. You **are permitted** to collaborate/work in groups on the homework if you choose; if you choose to do so, you should list the names of the students you worked with at the beginning of your submitted homework assignment.

You are permitted to make use of multimodal large language models (like ChatGPT) when writing the **code** portions of assignments if you choose. However, you should be aware that the work of such models is often flawed at best, if not outright inaccurate, and be prepared to double- and triple-check your work before submission. If you do use ChatGPT for **any** part of your code, you absolutely **must cite** the use of ChatGPT, specifically indicating exactly which lines it was used for and what prompt(s) it was given. If we detect your use of ChatGPT **without** such a citation, it will be considered plagiarism and handled accordingly. You are **not permitted** to use ChatGPT to write the text portions of assignments; if we detect such use, it will also be considered plagiarism.

- **Homework** (30%). Homework is assigned approximately every two weeks (this may vary) and will always be due by Sundays at 11:59 PM PDT. It will be **graded for accuracy** and solutions for the homework assignments will be posted approximately one week after each deadline. The graduate students will need to complete additional questions, but all assignments are ultimately rescaled to be out of 50 points. There are a total of **five** homework assignments and none will be dropped.
- **Labs** (20%). Section attendance is mandatory, and will be measured through the completion of a lab. Upon participating and completing the labs, you will receive credit for section attendance. You may miss a total of **two** sections without penalty. Doing the labs will be greatly beneficial for your homework, quiz, and project grades; with that in mind, you are strongly encouraged to attend all lab sections and follow along with the materials for each lab.
- **Quizzes** (20%) Each week, beginning the week of **October 7th** and continuing through week 9, there will be a short quiz posted on Canvas. Note that there will be no quiz during Thanksgiving week. There will be a total of **six quizzes**, with **no make-up** quizzes. Your lowest quiz score will be dropped.

  You will have a 24-hour window to take each quiz; that window will begin at 8:00 AM Thursday and close at 8:00 AM Friday. Within that window, each quiz will have a 15-minute time limit, which will begin once you open the assignment. (DSP accommodations will be built in via Canvas; make sure you submit an instructor letter for any necessary accommodations.) The quizzes will consist of a few questions similar to the homework; they may be conceptual questions, they may ask you to interpret some results, or they may ask you to perform some analyses. They are meant to serve as academic markers to provide an ongoing assessment of your understanding and to flag areas that may require more attention.

  If you do not remember to log in and complete the quiz during that designated 24-hour window, you will receive a zero for that quiz, with no exceptions.

- **Final Project** (30%). The final project is an important portion of the class and allows students to demonstrate their understanding of the material by working with a data set of their choice. Students should begin thinking about a project, forming groups of four (4), and looking for a data set immediately; a page on Canvas provides a list of possible resources. You will receive a considerable amount of guidance about the project over the course of the quarter.

  During the final exam time for the course (**Thursday, December 12th from 4:00 to 7:00 PM**), all groups will be required to give a short five-minute presentation on their project, in the form of either a poster or PowerPoint-style slideshow, followed by answering three minutes of questions. Your group will also need to submit a short (3-5 paragraph) report on your project, in the form of a Markdown or Quarto file or a Jupyter notebook.

  The project will require work and cannot be satisfactorily completed in only a day (or even a week). You are **strongly advised** to work on it in stages throughout the quarter. See Canvas for more detailed instructions and a rubric.

Tentative Course Schedule

The weekly schedule is indicated below. The topics and reading are subject to change based on the progress of the class.

	Topic	**Assignment**	**Project Stage**
Week 1	Intro to data science	Data Memo	Pick topic
Week 2	Application programming interfaces (APIs)	Homework 1	Find data
Week 3	Web scraping		Tidy data
Week 4	Data cleaning	Homework 2	EDA
Week 5	Data manipulation & feature engineering		
Week 6	Data visualization	Homework 3	Run models
Week 7	Natural language processing (NLP)	Homework 4	
Week 8	Recommender systems		Write-up
Week 9	Image classification	Homework 5	Edits
Week 10	Effective communication	Final Project	Final draft

Time Commitment

The course is 4 credit units; each credit unit corresponds to an approximate time commitment of 3 hours. You should expect to allocate 12 hours per week to the course. If you find yourself spending considerably more time on the course on a regular basis, please let the instructor or TAs know so that we can help you balance the workload.

A suggested allocation of this time is as follows:

- Reading and class time: 3 hours (25%)
- Homework: 4.5 hours (37.5%)
- Sections: 1 hour (8.3%)
- Final project: 4.5 hours (37.5%)

Course Policies

Communication

There are four means of communication with other students or the instructional team: during/after class, office hours, email, and individual appointments. **Please use them in that order**. Note that I can be slow at responding to email, which is why it is low on the list; to compensate, however, I offer several office hours weekly and am always willing to discuss problems or answer questions in person/during class.

1. **During/after class**. The easiest guaranteed way to contact me is to come up after class and say hello, or to raise your hand and ask a question during class.
2. **Office hours**. Office hours are offered at a minimum of twice weekly. I hold one session of 4 hours a week, and the other(s) are held by various combinations of TAs or ULAs. These are opportunities to interact informally, ask questions, and discuss course material or assignments.
3. **Email**. Please use email with discernment for simple communication. A response is guaranteed within 48 weekday hours (so if you email on Friday afternoon, you may not receive a reply until Tuesday afternoon). If your message is time-sensitive, please indicate so in the subject, and I will do my best to respond promptly.
4. **Appointment**. You can schedule individual 20-minute appointments with me as needed. These appointments may be either on Zoom or in person. This mode of communication is best suited to more complex or nuanced communication regarding personal matters. If you schedule an appointment, you will be prompted to indicate what you wish to discuss.

Extra Credit

If the class reaches a 90% submission rate of course evaluations at the end of the quarter, the **entire class** will receive 5 free points on the final project.

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 according to the rubric below.

A+	100%
A	94% – 99.99%
A-	90% – 93.99%
B+	87% – 89.99%
B	84% – 86.99%
B-	80% – 83.99%
C+	77% – 79.99%
C	74% – 76.99%
C-	70% – 73.99%
D+	67% – 69.99%
D	64% – 66.99%
D-	60% – 63.99%
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 Canvas 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.

Deadlines

You receive two free late homework submissions without penalty. This policy applies only to homeworks. When you wish to use one of these late submissions, simply submit within one week of the original deadline, and **add a note** at the beginning of the assignment explaining that it is a free late submission.

Non-exempted homeworks submitted within one week of the deadline will be awarded 50% credit. **No credit** will be awarded for homework turned in more than one week late; please plan ahead and submit your work on time. No late quizzes will be accepted.

The final project deadline is firm and no late submissions will be accepted.

Extensions

Extensions may be granted based on individual circumstances at the instructor's discretion.

Conduct

Please be mindful of maintaining respectful and kind communication. You are expected to uphold the UCSB student code of conduct in your behavior when in class, in section, or interacting with other students or the instructional team. You can find the student code of conduct on the Office of Student Conduct website from this page. If you are uncomfortable with the conduct of another individual 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.

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 Evaluations

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

Student Resources

Any students in need are encouraged to make use of the following resources.

- Financial Crisis Response Team
  https://food.ucsb.edu/about/committees/financial-crisis-response-team

- Food Security and Basic Needs (Food, Housing, Technology) Advising Center
  https://food.ucsb.edu/resources/basic-needs-advocates
- Undocumented Student Services
  http://www.sa.ucsb.edu/dreamscholars/home
- Campus Advocacy, Resources, and Education (CARE)
  https://care.ucsb.edu/
  24/7 Confidential Phone: (805) 893-4613
- The Trevor Project
  https://www.thetrevorproject.org/
- Counseling and Psychological Services (CAPS)
  https://caps.sa.ucsb.edu/
  24/7 Counselors: (805) 893-4411, press 2