

Instructor: Ethan P. Marzban (he/him)

Course Site: [ucsb-pstat100.github.io](https://ucsb-pstat100.github.io)

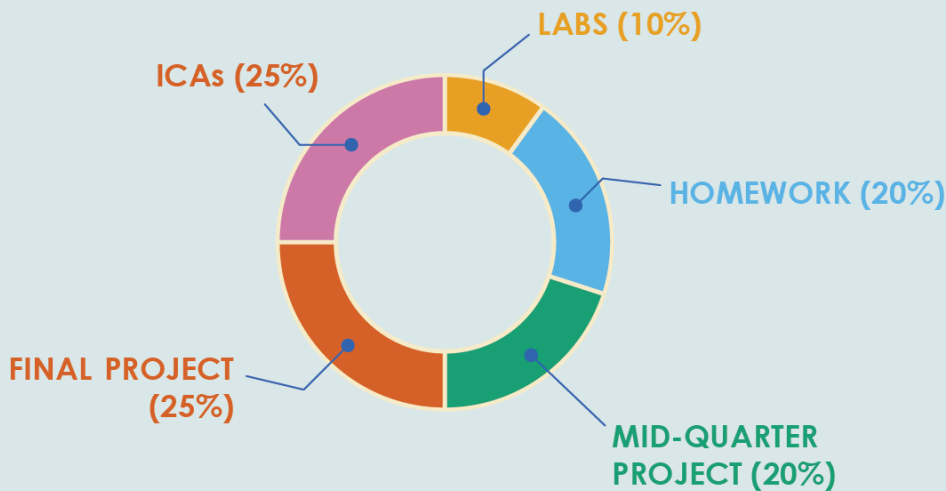
## WEEKLY MEETINGS

M, T, W, R: 12:30 - 1:35pm in ILP 2207

## SECTIONS AND TEACHING ASSISTANTS

• T, R 2 - 2:50pm in PHELPS 1517: Erika McPhillips  
 • T, R 3 - 3:50pm in PHELPS 1517: Erika McPhillips

## GRADE BREAKDOWN



## Approximate Grade Cutoffs

- A+ 100%
- A [95.0000, 100.0000)
- A- [90.0000, 95.0000)
- B+ [86.3333, 90.0000)
- B [83.3333, 86.3333)
- B- [80.0000, 83.3333)
- C+ [76.0000, 80.3333)
- C [73.3333, 76.3333)
- C- [70.0000, 73.3333)
- D+ [66.0000, 70.3333)
- D [63.3333, 66.3333)
- D- [60.0000, 63.3333)
- F < 60%

Cutoffs may be adjusted at the end of the quarter, depending on class performance.

## OVERVIEW OF REQUIRED ASSIGNMENTS

**WE WANT YOU TO SUCCEED!** To that end, we have designed a wide array of different assignment types which will be used to periodically check-in with you and your understanding of the course. These assignments are:



## LAB ASSIGNMENTS

Designed to help you practice the computational aspects of Data Science. Each Lab will have an associated assignment.



## HOMEWORK ASSIGNMENTS

Designed to probe your theoretical understanding of Data Science. Two homework assignments will be assigned in total.



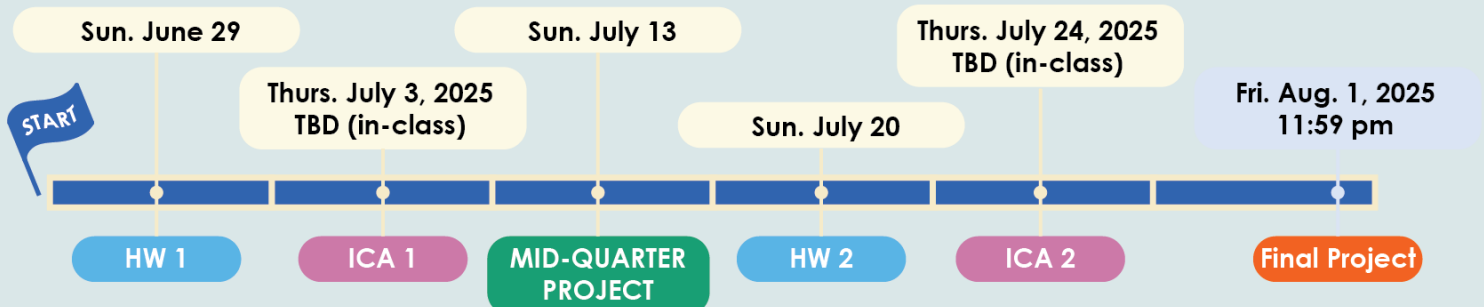
## PROJECTS

Designed to help you apply course concepts to real-world data, and produce presentable results.



## IN-CLASS ASSESSMENTS

Designed to assess your retention of course concepts in a timed setting. There will be two ICAs this quarter.



## PREREQUISITES



**PROBABILITY (PSTAT 120A)**



**LINEAR ALGEBRA (MATH 4A)**



**PRIOR PROGRAMMING EXPERIENCE**  
(eg. CMPSC 9 OR CMPSC 16)

As we will soon see, Data Science lies at the intersection of many fields; most prominently statistics, mathematics, and computer science. Consequently, a strong introduction to all three of these fields will set you up for great success as a Data Scientist! Each of these prerequisites touch upon one of these fields

Additionally, we will be programming in the language “R”, but you are not expected to have any prior experience with R specifically.

## LEARNING OUTCOMES

*At the end of this course, we hope you will be able to:*

- 1 critically assess data quality and sampling design
- 2 retrieve, inspect, and clean raw data
- 3 understand the basics of exploratory, descriptive, visual, and inferential techniques
- 4 interpret and communicate results in context

## SOME POTENTIALLY USEFUL RESOURCES



**DISABLED STUDENTS PROGRAM (DSP):** <https://dsp.sa.ucsb.edu/>

If you require accommodations for lectures, in-class assessments, or assignments in general, DSP is a great office to consult in order to get you those accommodations. The DSP office also handles short-term accommodations, for example due to a sprained wrist.



**CAMPUS LEARNING ASSISTANCE SERVICES (CLAS):** <https://clas.sa.ucsb.edu/>

Though CLAS does not have a designated resource for PSTAT 100, they do offer extensive support relating to many PSTAT classes (including PSTAT 120A), and can be a great resource to get additional help from outside the course staff.



**COUNSELING AND PSYCHOLOGICAL SERVICES (CAPS):** <https://caps.sa.ucsb.edu/>

College can be stressful! CAPS is here to help you navigate that stress, and try and help you navigate the complexities of being a student. In addition to various counseling services, they also have a selection of “egg-chairs” which can provide a nice physical break from the daily grind.



**TECHNOLOGY NEEDS:** <https://basicneeds.ucsb.edu/resources/technology-resources>

As a part of this course, you will be asked to program using the language “R”. This will necessitate that you have access to a laptop, which we understand is not the case for everyone. Please consult the link above for information on how to obtain a loaner laptop and other technological needs.



### LATE SUBMISSIONS

We understand that life happens! To that end, you are one late homework, which must be submitted within 24 hours of the original deadline, and two late labs, which must be submitted within 24 hours of the original deadline



### AI POLICY

The use of AI (Artificial Intelligence) on Labs, Homework, and Projects is permitted so long as its use is properly cited. Failure to disclose the use of AI will be treated as Academic Misconduct. The use of AI on ICAs is strictly prohibited.



### EMAIL POLICY

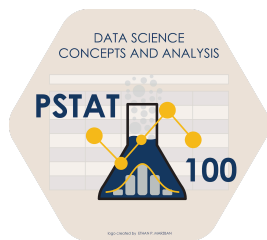
Please refrain from emailing the course staff (except in cases of extreme emergency). Instead, please communicate through EDStem, or the inquiry form linked on the course website. Thank you for your cooperation!



### COLLABORATION POLICY

Collaboration on Homework and Lab is encouraged! Collaboration on Projects will be required. Collaboration on ICAs is strictly prohibited, and punishable in accordance with policies on Academic Misconduct. Please don't cheat!

**FOR ADDITIONAL INFORMATION ON POLICIES, COURSE STAFF CONTACT INFORMATION, AND A WEEKLY BREAKDOWN OF TOPICS AND LECTURES, PLEASE CONSULT THE MAIN COURSE SITE. THANKS!**



# PSTAT 100: Data Science Concepts and Analysis

## Summer Session A, 2025

### Course Policies and Syllabus

Welcome to PSTAT 100: Data Science Concepts and Analysis! I am very excited to join you as your instructor this quarter. Our journey together will take us through the basics of Data Science, and aims to prepare you for your future endeavors in the field, whether they be in classes, industry, or academia. Here's to a great quarter!

– Ethan

### LECTURE INFORMATION

M, T, W, Th: 12:30 - 1:35 pm in ILP 2207

### SECTION INFORMATION

- T, R from 2 - 2:50 pm in PHELPS 1517
- T, R from 3 - 3:50 pm in PHELPS 1517

### COURSE STAFF

#### Instructor

Ethan P. Marzban (he/him)  
[epmarzban@pstat.ucsb.edu](mailto:epmarzban@pstat.ucsb.edu)

#### Teaching Assistant

Erika McPhillips (she/her)  
[elm00@pstat.ucsb.edu](mailto:elm00@pstat.ucsb.edu)

*Please reserve email for emergencies only; thank you! (More Information Below)*

### COURSE DESCRIPTION

The official description of this course, from the UCSB Course Catalog, is:

*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. Credit units: 4.*

Indeed, this course is designed to be a hands-on introduction to Data Science for intermediate-level students with some exposure to probability and basic computation skills, but with few or no upper-division courses in statistics.

### PREREQUISITES

- **PSTAT 120A** (Probability at a calculus-based level)
- **Math 4A** (a first pass at Linear Algebra)
- **Prior Experience with a programming language** (e.g. Python through CMPSC 9 or CMPSC 16, or R through PSTAT 10)

**Note on Programming:** the primary programming language of this course will be R, though you are not expected to necessarily have prior experience coding with R (so long as you have experience coding in another language, e.g. Python, Julia, etc.)

## 📖 TEXTBOOKS/READINGS

There are four main textbooks for this class (all of which are freely available online; links provided below); readings will be regularly assigned from them:

- *Modern Data Science with R (2e)*, by Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton.
- *Introduction to Modern Statistics*, 2nd Ed., by Mine Çetinkaya-Rundel and Johanna Hardin
- *R for Data Science (2e)*, by Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund
- *An Introduction to Statistical Learning with Applications in R*, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

We will also read through a handful of articles; these will all appear in the Course Schedule.

### Additional (Optional) Texts:

- *Learning Data Science*, by Sam Lau, Joey Gonzalez, and Deb Nolan.  
A great comprehensive resource, but written in Python (so be aware!)
- *An Introduction to R*, by Alex Douglas, Deon Roos, Francesca Mancini, Ana Couto & David Lusseau  
A good textbook for those of you new to programming in R

## 💡 LEARNING OUTCOMES

By the end of this course, you should be able to:

- critically assess data quality and sampling design
- retrieve, inspect, and clean raw data
- understand the basics of exploratory, descriptive, visual, and inferential techniques
- interpret and communicate results in context

## 📁 ASSESSMENTS

- **Labs:** shorter, structured coding assignments designed to introduce programming concepts and skills. There will be a Lab assignment associated with every Discussion Section; labs will be due (TBD).
- **Homeworks:** slightly larger in scope than labs; will contain both programming and theoretical/conceptual concepts. We will have a total of 2 homework assignments released throughout the quarter. Homeworks are designed to probe your theoretical understanding of concepts that arise in the field of Data Science, as well as prepare you for the projects by asking you to address more open-ended questions.

- **Mid-Quarter Project:** more open-ended than homeworks and labs, and is designed to more closely simulate real-world data science projects and endeavors. Details on the mid-quarter project will be released in the later part of Week 1.
- **Final Project:** a final, comprehensive project that will be due on Friday August 1, 2025. You will be required to work in teams on the Final Project.

**The final project will be due by 11:59pm on Friday, August 1, 2025.**

- **2 In-Class Assessments (ICAs)** will be administered (see below for dates). I hesitate to call these “exams” because they are not intended to be as high-stakes as exams, however they are designed to test your retention of course material. More information will be released as we approach the date of the first ICA.

- In-Class Assessment 1: **Thursday, July 3, 2025 during lecture, at TBD**
- In-Class Assessment 2: **Thursday, July 24, 2025 during lecture, at TBD**

**There will be no ICAs offered at alternate times, for any reasons. So, please plan accordingly!**

## Final Course Grades

Your final grade in the course will be computed according to the following scheme:

| Assessment           | Weight           |
|----------------------|------------------|
| Labs                 | 10%              |
| Homework             | 20% (10% each)   |
| Mid-Quarter Project  | 20%              |
| Final Project        | 25%              |
| In-Class Assessments | 25% (12.5% each) |

Your final letter grade will be issued according to the following scheme (note that I am using interval notation:  $[a, b)$  means all numbers greater than or equal to  $a$  but less than (and not equal to)  $b$ ) :

| Grade | Course Percentage   |
|-------|---------------------|
| A+    | 100%                |
| A     | [95.0000, 100.0000) |
| A-    | [90.0000, 95.0000)  |
| B+    | [86.3333, 90.0000)  |
| B     | [83.3333, 86.3333)  |
| B-    | [80.0000, 83.3333)  |

| Grade | Course Percentage  |
|-------|--------------------|
| C+    | [76.0000, 80.3333) |
| C     | [73.3333, 76.3333) |
| C-    | [70.0000, 73.3333) |
| D+    | [66.0000, 70.3333) |
| D     | [63.3333, 66.3333) |
| D-    | [60.0000, 63.3333) |
| F     | < 60%              |

I have elected to adopt an uncurved grading scheme to eliminate any sense of “competition” among students; I highly encourage you all to collaborate with and uplift each other. Having said that, I will certainly consider adjusting the cutoffs at the end of the quarter if necessary.

## POLICIES

### Late Submissions

I understand that life happens! To that effect, I am allowing:

- One late homework, which must be submitted within 24 hours of the original deadline
- Two late labs, which must be submitted within 24 hours of the original deadline

Given that the projects will be submitted in groups, **no late submissions for the projects will be accepted** for any reason. Additionally, as stated above, we will not be offering any make-ups for the ICAs; furthermore, failure to take both ICAs will result in a grade of “F” being administered.

### Communication

There are two primary means of communication outside of scheduled class meetings: office hours and an EdStem Discussion Forum (please see Canvas for a join link; for security purposes, we are only allowing currently-enrolled students join the discussion forum). If you are unsure of how to reach out to appropriate parties, please consult the following table:

| Topic                         | Redirect to...         |
|-------------------------------|------------------------|
| Troubleshooting code          | EdStem                 |
| Checking answers              | Office hours or EdStem |
| Clarifying assignment content | Office hours or EdStem |

|                               |                    |
|-------------------------------|--------------------|
| Assignment submission         | Gradescope         |
| Re-evaluation request         | Gradescope         |
| Question about missing grades | Fill Out This Form |

Some additional comments:

- Please note that we (the course staff) request you **refrain from emailing us except in case of extreme emergency** (it is up to you to decide what is an ‘emergency’). Please bring all of your questions to the course staff during either Office Hours or after Lecture/Section. Thank you!
- If you have questions or concerns about missing grades, please use the form linked in the table above. You are allowed to submit the form multiple times, but we ask that you please wait at least 48 business hours before submitting follow-ups. Thank you!

### Collaboration and Academic Integrity

Data Science (as we will see) is an inherently collaborative field. Indeed, collaboration is *required* for both projects; collaboration is also encouraged for homework and lab assignments. However, there are limitations to collaboration:

- collaboration on the ICAs is strictly prohibited
- do not copy other people’s work and try to pass it off as your own
- if/when you work in groups, include the names of **all** group members on the assignment

Failure to abide by these principles will be treated as academic misconduct. Anyone found guilty of academic misconduct will be reported to the Academic Senate, and will receive at minimum a failing grade on the assignment in question; further actions may also include failing the course, and marks being made on permanent records. Depending on the severity of the infraction, expulsion is also a possibility.

Basically, don’t cheat- please! If you’re ever struggling with course material, please come talk to me or the TA’s. We are truly here for you, and want only the best for you.

### Section Switching

Lab Sections take place in special “Collaborate Classrooms” which are equipped with laptops. There are only a fixed number of seats and laptops in these classrooms, meaning we cannot under any circumstance over-enroll sections. Therefore, if you want to switch sections unofficially (we do not have the ability to switch your official enrollment through GOLD), please follow the steps at this link. Any requests to switch sections that do not adhere to the guidelines posted at that link will be ignored.



## AI Policy

It is undeniable that the recent advances in Generative AI (GenAI) and Large Language Models (LLMs) like ChatGPT have reshaped the educational landscape. Indeed, when utilized properly, they can be an incredibly useful power. With that said, I would like to establish some clear ground rules with regards to the use of GenAI in this class:

- The use of AI on ICAs is strictly prohibited.
- The use of AI on other assignments is discouraged, but not prohibited - I only ask that if you use AI (for a HW, Lab, or Project) that you **please cite it**.
  - If it is found that you used AI without citing its use, this will be treated as academic misconduct.

### Important

Please be careful when using Generative AI. It is still a relatively new innovation, and can not only occasionally produce inaccurate answers but can also produce unethical answers. Read all terms and conditions carefully, and ensure you understand (a) how the tool will give you an answer, and (b) what the tool will do with your data.

## DISABLED STUDENTS PROGRAM (DSP)

If you have a disability, or otherwise require accommodations for the ICAs please reach out to the Disabled Students Program (DSP) ASAP to ensure your request(s) for accommodation can be processed. We ask that all requests be logged at least a week in advance, to ensure the system enough time to process. Please note that we cannot grant any requests for accommodations unless they come to us from DSP directly.

## TECHNOLOGY NEEDS

As a part of this course, you will be required to program in **R**. Though the Lab Sections take place in specially designed classrooms that come equipped with computers, your homework and quizzes may cover **R**-related questions, which means we expect you to have access to a laptop capable of connecting to the internet. If you do not currently possess such a laptop, please check out UCSB's Basic Needs Resource page on Technology Resources to try and acquire one.

## ★ DISCLAIMER

The instructor reserves the right to modify this syllabus if he deems such modifications academically advisable. Such modifications, should they occur, will be announced publicly.

## FACULTY MENTOR

The faculty mentor for this course is Dr. Jack Miller. They can be reached at [jbmillar@pstat.ucsb.edu](mailto:jbmillar@pstat.ucsb.edu). Please note that Dr. Miller will not be able to authorize regrades or accommodations/extensions for the course; for such inquiries, please utilize the communications channels listed above. Thank you!

## SOME GENERAL TIPS FOR SUCCESS

### ***Form study groups***

Data Science is not meant to be a lonely field! There is much we can learn from one another, and it can be an incredibly enlightening experience to discuss problems and ideas with one another. (Just make sure you don't violate any of the Academic Integrity points listed above)

### ***Start things early!***

Make sure you're giving yourself enough time to complete the homework assignments, labs, projects, and make sure to leave plenty of time to study for the ICAs. I'd recommend creating a weekly schedule for yourself, and allocating time each day for PSTAT 100 material (whether that be working on an assignment, reading lecture slides, or coming to Office Hours).

### ***Practice Makes Progress***

The best way to start learning Data Science is to start *doing* Data Science. The various textbooks and resources linked on the Course Website come equipped with additional practice problems and exercises which I highly recommend you work through.

### ***Attend office hours (TA and Instructor) regularly***

Even if you don't have a specific question, you're always more than welcome to sit in on Office Hours and listen to other people's questions. (Sometimes, doing so will help you formulate your own questions!)

### ***Attend lectures and Discussion Sections.***

It's true that we do not have an attendance policy, but please don't let yourselves fall behind on attendance. Studies show that regular exposure is the best way to learn material, and there really is no substitute for going to Section and Lecture. Also, while you're in Lecture, **take your own notes!** Even the act of writing things down and having to synthesize what you think is important information can help you process and learn the material in real time.

### ***Don't be too hard on yourself!***

Though a *little* stress can be a good motivating factor for some, please don't stress yourself out too much. Your performance in this course is **not** an evaluation of who you are as a person!

## 📅 SCHEDULE AT-A-GLANCE

Please note that this schedule is only tentative, and may change at the discretion of the instructor (and also according to how quickly we run through the material this quarter). A full (and updated) version of this schedule will be maintained on the Course Website.

| Week | Date                 | Lecture Topic                         | Anything Due?                  |
|------|----------------------|---------------------------------------|--------------------------------|
| 1    | Mon, June 23, 2025   | Introduction to Data                  |                                |
|      | Tues, June 24, 2025  | Data Structures / Tidy Data           |                                |
|      | Wed, June 25, 2025   | Visualizations, Part I                |                                |
|      | Thurs, June 26, 2025 | Visualizations, Part II               |                                |
|      | Sun, June 29, 2025   |                                       | <b>HW01 Due</b>                |
| 2    | Mon, June 30, 2025   | Geometry of Data                      |                                |
|      | Tues, July 1, 2025   | Principal Components Analysis         |                                |
|      | Wed, July 2, 2025    | Review / Catch-Up                     |                                |
|      | Thurs, July 3, 2025  | <b>In-Class Assessment 01</b>         |                                |
| 3    | Mon, July 7, 2025    | Sampling / Missing Data               |                                |
|      | Tues, July 8, 2025   | Study Design / Sampling Distributions |                                |
|      | Wed, July 9, 2025    | Estimation / Confidence Intervals     |                                |
|      | Thurs, July 10, 2025 | Hypothesis Testing, Part I            |                                |
|      | Sun, July 13, 2025   |                                       | <b>Mid-Quarter Project Due</b> |
| 4    | Mon, July 14, 2025   | Hypothesis Testing, Part II           |                                |
|      | Tues, July 15, 2025  | Introduction to Statistical Modeling  |                                |
|      | Wed, July 16, 2025   | Regression, Part I                    |                                |
|      | Thurs, July 17, 2025 | Regression, Part II                   |                                |
|      | Sun, July 20, 2025   |                                       | <b>HW02 Due</b>                |
| 5    | Mon, July 21, 2025   | Regression, Part III                  |                                |
|      | Tues, July 22, 2025  | Classification                        |                                |
|      | Wed, July 23, 2025   | Review / Catch-Up                     |                                |
|      | Thurs, July 24, 2026 | <b>In-Class Assessment 02</b>         |                                |
| 6    | Mon, July 28, 2025   | Data Ethics                           |                                |
|      | Tues, July 29, 2025  | Causal Inference                      |                                |
|      | Wed, July 30, 2025   | Neural Networks                       |                                |
|      | Thurs, July 31, 2025 | Closing Remarks                       |                                |
|      | Fri, August 1, 2025  |                                       | <b>Final Project Due</b>       |