COGS 108: Data Science in Practice
Winter 2020
TTh 8-9:20 (Galbraith Hall 242)

Instructor: Shannon Ellis (sellis@ucsd.edu)

Office Hours:

Date & Time	Location	Instructional Staff
Mon 1-2 PM	CSB 227	Matthew Feigelis
Wed 2-4 PM	CSB 243	Professor Ellis
F 11AM-12PM	SSRB 100	Sam Lau
Fri 1:30-2:30PM	Sun God Lounge	Shreenivas Venkataramanan

Instructional Staff & Section Times

	Day	Time	Location	Staff (role)
A01	Mon	2 PM	MANDEB-150	Matt (TA) & Emily (IA)
A02	Mon	3 PM	MANDEB-150	Matt (TA) & Emily (IA)
A03	Wed	12 PM	PCYNH 121	Sam (TA) & John (IA)
A04	Wed	1 PM	PCYNH 121	Sam (TA) & Amir (IA)
A05	Wed	2 PM	PCYNH 121	Amir (IA) & John (IA)
A06	Fri	11 AM	PCYNH 121	Shreenivas (TA) & Josh (IA)
A07	Fri	12 PM	PCYNH 121	Shreenivas (TA) & Josh (IA)

Course GitHub: https://github.com/COGS108/
DataHub: datahub.ucsd.edu
Course Piazza*: piazza.com/ucsd/winter2020/cogs108
Course podcast/screencast: https://podcast.ucsd.edu/?q=COGS108
Course Feedback (anonymous): Google Form

*You will be able to post anonymously on Piazza; however, you will only be anonymous to your classmates. Your instructional staff will be able to see who you are.

COURSE OBJECTIVES

- Formulate a plan for and complete a data science project from start (question) to finish (communication)
- Explain and carry out descriptive, exploratory, inferential, and predictive analyses in Python
- Communicate results concisely and effectively in reports and presentations
- Identify and explain how to approach an unfamiliar data science task

COURSE MATERIALS

- All students will need access to an iClicker*
- There is no textbook
- All materials will be provided on GitHub

CLASS TECHNOLOGY

- Python (>= 3.6; Anaconda distribution)
- Jupyter Notebooks
- git and GitHub (option to use SourceTree or other GUI)

*You *will* need a clicker, of this brand, as no other brand will work with the system we are using. You must register your clicker on Canvas & bring it to lecture. If you previously registered one on Canvas & are using the same clicker in this class, you do *not* have to register it again. If you previously registered on TritonEd, you DO have to register again on Canvas. If you would rather use the REEF app, you are free to do so; however, note that if the Wi-Fi is down, the app will not work and you will not get credit for those responses.

COURSE OVERVIEW

This class is a hands-on practical programming course intended to get you experience working on data science projects. In COGS 9 (Introduction to Data Science) you (may have) learned why data and data science are important. This class goes beyond appreciation for what *can* be done to actually *doing it*. Often the best way to learn something is to do it yourself. Often, this process will involve attempting to do something, doing it wrong, learning from your mistakes, and then succeeding. That's part of the data science process. This course is all about the *practice* of data science.

In focusing on the application, there is theory that won't be discussed and mathematical proofs that won't be done. That is by design. In particular:

1. There are entire courses dedicated to each of the topics we'll cover. To have time to *do* anything, we can't teach all the details in a single course.
2. Experts in each of these domains are out there and excited to teach you the nitty gritty about each topic.
3. My expertise is *not* machine learning. It's data science, education, human genetics, and the intuition behind data analysis.
4. We're promoting data literacy. We believe that everyone who is data literate is at an advantage as they go out into the modern world. Data literacy is not limited to those who are computational gurus or math prodigies. You do not have to be either of those to excel at this course.

In this course, you will try many methods. Every so often, you'll even be asked to implement a technique that has *not* been explicitly taught. Again, this is by design. As a data scientist, you'll regularly be asked to step outside of your comfort zone and into something new. Our goal is to get you as comfortable as possible in that space now. We want to provide you with a technical and a data science mindset that will allow you to ask the right questions for the problem at hand *and* set off alarm bells when something in your dataset or analysis is "off."

GRADING & ATTENDANCE

Grading:

	% of Total Grade
(6) Assignments	45
Class Participation (9) Weekly Quizzes OR Lecture Attendance	10
(1) Project Proposal	10
(1) Project Check-In	2
(1) Project Survey	3
(1) Final Project	25
(2) Guest Lectures	5

Final exam date: No final exam (or any exams, for that matter). Only a final project deadline, due **Thursday March 19th** of finals week by 11:59 PM.

Your letter grade will be determined using the standard grading scale. Grades are *not* rounded up.

Grades

All grades will be released on Canvas. We will try to send out automated alerts if we do not receive a submission and/or if it fails to be processed, but *ultimately it is your responsibility to check your grades and get in touch if any are missing or you think there is a problem*.

Assignment Regrades

We will work hard to grade everyone fairly and return assignments quickly. But, we know you also work hard and want you to receive the grade you've earned. Occasionally, grading mistakes do happen, and it's important to us to correct them.

If you think there is a mistake in your grade, request a regrade *within 72 hours of your receipt of the grade*. This message should include evidence of why you think your answer was correct (i.e. a specific reference to something said in lecture) and should point to the specific part of the assignment in question.

Note that points will not be rewarded if you fail to follow instructions. For example, if the instructions say to name the variable `orange` and you name it `ornage`, you will not be rewarded credit upon regrade. This is because (1) following instructions and being detail-oriented is important and (2) there are hundreds of students taking the course this quarter. It would be an unfair burden to place on TAs if we didn't have this policy.

Lecture Attendance

Our goal is to make lecture and discussion section worth your while to attend. However, most days it will be up to you whether or not you show up.

Attendance is only *required* on guest lecture days - these will be announced at least a week in advance in lecture and on Piazza. Each guest lecture is worth 2.5% of your grade.

For all other lecture days, there will be a flexible attendance policy. Starting with the second lecture, for all non-guest lecture days, if you attend at least 75% of the lectures (where attendance means you answered 50% of that day's iclicker questions), you will receive full credit for 10% of your grade. If you do not, your weekly quiz grade will be used for this 10% of your grade.

Weekly Quizzes

There will be weekly quizzes on Canvas covering the previous week's lecture content. Quizzes will be released on Friday, the week before they are due. The first weekly quiz will be due Friday at 11:59 PM of week 2. You will have one attempt and 15 minutes to complete the quiz. Quizzes will typically have approximately 10 questions. If you do not attend 75% of the lectures, your average across these quizzes will make up your lecture participation grade (with your lowest weekly score quiz being dropped).

If you *do* receive lecture attendance credit and decide to *also* do the weekly quizzes, if your average across the weekly quizzes is higher than your lowest assignment score, your quiz average will replace your lowest assignment score. This is only available for those who receive lecture attendance.

Discussion Section Attendance

Attendance is not required. But, to get the most out of this course and get the most amount of practice possible, we suggest attending section regularly.

You'll want to try to attend the section to which you're assigned; however, if something comes up one week or you can't make your assigned section, you are free to attend another section. If any one section becomes too crowded, we will return to and adjust this policy, as needed.

If you're struggling on assignments, you are strongly encouraged to attend section. (You're also encouraged to attend simply to maximize how much you learn from this course.) However, attendance is not required. Attendance will be taken during section because if you attend and participate in eight weeks of discussion section, you will have the opportunity to replace your lowest assignment score with full credit for that assignment. (Note: There are two weeks where Monday sections will not meet due to holidays. Students in the Monday sections should plan to attend a Wed or Fri section on those weeks.)

ASSIGNMENTS

Assignments are hands-on in this course. They will be completed individually in Jupyter Notebooks and released and submitted on datahub.

The *practice* of data science involves writing code to answer questions and accomplish tasks. Thus, to get practice, your assignments will require you to use Python to do just that. Not everything will be explicitly mapped out step-by-step for you. This is intentional. Figuring things out when it's not entirely clear what to do next is part of the practice here. You'll attempt things that won't work and become comfortable with this. You'll get stuck and work to get unstuck. Not quite knowing exactly what's going on at all times is part of the process. And, to be honest, part of the job of being a data scientist.

That said, the first two assignments will be the simplest assignments and aim to get you up to speed in Python. If the first two assignments are particularly difficult for you, that's ok. But, it's then up to you to determine if you want to put in the work to make it through the rest of the quarter. Assignments will take more time and be more difficult starting with the third assignment.

As assignments become more difficult, we don't want you to get or feel *totally* lost. If you've thought long and hard, gone down a long rabbithole on Stack Overflow, and can't even get a sense of what the next step may be, take a step away. Take a break. Then, come back and see if you can't solve it with a refreshed mind. If you're *still* totally stuck, ask on Piazza, talk to a classmate, and/or attend office hours for help.

With regards to asking questions of instructional staff, we're here to help you, but there are way more students than there are instructors. So, help each other. Ask one another first. It's awesome that we all have different backgrounds and experiences - let's use that to our advantage. In fact, this is how the best data science gets done. Diverse minds solving a problem invariably improves the solution. Also, teaching something to someone else is the best way to determine if you really know something. So, it's win-win. The person who's stuck gets unstuck and the person who helped is more sure in their knowledge. Help one another! Section and office hours are meant to be collaborative.

Also, your instructional staff may not know the answer to everything off the top of their head right away. There is an entire field of data science topics and programming out there - we'll do our best to help and show you where to look and how to figure out the answer (or steer you in a different direction if your approach is going to lead you in a messy and intractable direction), but know that we may not have all the answers.

Deadlines

Assignments will be submitted individually on datahub. We'll talk about the details for submission in class. Assignments will always be released at least a week before the assignment due date. On weeks with assignment deadlines, they *will always be due Friday at 11:59 PM of the week specified (see Course Outline below)*.

Check to ensure that your file shows up under "Submitted assignments" on datahub after you click submit. If the file is the incorrect file, corrupted, or otherwise unreadable, we cannot grade it and we will mark your assignment as late.

Late assignments earn fractional credit (75% within one week late; no late assignments accepted after one week).

Feedback & Grades

It is your responsibility to ensure that we receive a submission from you on datahub *and* that you submit the correct file (a Jupyter notebook with the .ipynb extension with the same file name as the assignment) for each assignment. If you identify that a mistake has been made, it is your responsibility to get in touch on Piazza should a problem arise. You will receive individualized feedback via email with your grade and feedback about a week after each assignment is due.

Assignment Questions on Piazza

Piazza will be used for all general questions. For example, if you are confused by what a question is asking or are unsure where to start to look for the answer and need direction, Piazza is the place to go. However, when asking or answering questions on Piazza, code that answers assignment questions should **not** be provided. Instead, answer with suggestions as to what topics/ideas/lectures to look into or vague pseudocode that helps move the person

asking the question in the right direction. For general programming questions (unrelated to the assignment answers), feel free to share minimal code segments.

COURSE PROJECT

Your course project will be completed in a group of 4-5 people, *even if you don't want to work with a group.* No exceptions. The reality of data science is that you will have to work with others. You'll need to work together to communicate effectively, manage time, organize your projects, and accomplish a goal. People will have different knowledge and skills sets. It is your job as a group to work together to figure out how to maximize each group member's skills to make sure that your differences are helpful to accomplishing your goal, rather than a hindrance. For example, some of you will find the programming aspects of the class assignments *very easy*, while others will struggle. Alternatively, some of you may find experimental research and hypothesis testing intuitive, while others find it confusing and frustrating. It is best for your project if you choose a team with a mix of background and experience.

Project Proposal

You will have to submit a group project proposal by the end of week four (see course outline below) on GitHub. This means that you will have to have to have met as a group by then, have determined what question you want to ask in your project, started to identify the data you'll need/the data you'll use to answer your group's question of interest, have laid out a plan for your project's completion going forward. You will receive feedback on your project proposal to help guide your final project. However, we strongly encourage you to chat with the instructional staff throughout the quarter as you work on your project to elicit more feedback and ensure you're going in the right direction.

Project Check-In

This will be a minimal check-in where you as a group will assess how your project is going, identifying what you've accomplished and where you are relative to what you planned. This will be graded for completion, not whether you have perfectly stuck to your group's plan. Check-In will be submitted on GitHub.

Group Project Survey

Every individual in the class will assess how working with their group as a whole and each individual in the group throughout the quarter has been via a short survey. Link to survey will be provided to students. Surveys will be completed individually.

Final Project

The final project will be a full, detailed data science report in the form of a Jupyter notebook that carries out an analysis from start to finish. This report will answer the data science question your group has chosen to answer. The topic will be up to you and more details will be provided in class, but generally this report will include (1) background research and ethical considerations, (2) your data science question(s) and hypothesis/hypotheses, (3) data & data wrangling, (4) a descriptive & an exploratory data analysis, (5) your full analysis, (6) your results, and your (7) conclusion(s). Along with your report, each individual will submit feedback about the team and each individual members' contributions to the project. Details about this will be sent out during week 5.

Final Project Extra Credit

Being an effective data scientist requires effective communication. The report you submit will demonstrate your ability to communicate in the written form; however, oral communication is equally important. Extra credit on the

final project is optional and can be earned in one of two ways: (1) A 3-5 minute video that communicates your group project's question, analysis, and results. This can be a filmed presentation or a video that is more creative. Most importantly, it must effectively communicate your team's project. Videos will be submitted on Canvas. (2) A 5 minute, group, in-person presentation Monday of Finals week (3/16). Time slots will be scheduled toward the end of the quarter. All members who choose to participate in the extra credit must be present on this day and at the scheduled time.

DISCUSSION SECTIONS

Section will be used to review material from lecture by getting hands-on programming experience. You will be given tips for working in Python, guided through Jupyter notebooks to clarify topics presented in class, and will be given time to get additional practice. There will be information covered in section that are not covered in lecture and that will be needed (or at least very helpful) for the assignments. See course outline below for what general topic will be covered in section each week.

Additionally, there will be time to form final project groups in discussion section during week 3. If you do not have a group, you should attend section that week.

Optional Readings:
There are no required readings for this course; however, if you're interested in learning more and reading about data science topics, we recommend the following:

- Donoho D, *50 Years of Data Science*
- Wickham H*, Tidy Data*
- Woo K & Broman K, *Data in Spreadsheets*
- Tukey JW, *Exploratory Data Analysis*
- Grus J*, Data Science from Scratch*

OTHER GOOD STUFF

Class Conduct

In all interactions in this class, you are expected to be respectful. This includes following the UC San Diego principles of community .

This class will be a welcoming, inclusive, and harassment-free experience for everyone, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, ethnicity, religion (or lack thereof), political beliefs/leanings, or technology choices.

At all times, you should be considerate and respectful. Always refrain from demeaning, discriminatory, or harassing behavior and speech. Last of all, take care of each other.

If you have a concern, please speak with Prof Ellis, your TAs, or IAs. If you are uncomfortable doing so, that's ok! The OPHD (Office for the Prevention of Sexual Harassment and Discrimination) andr CARE (confidential advocacy and education office for sexual violence and gender-based violence) are wonderful resources on campus.

Academic Integrity

Don't cheat.

You are encouraged to (and at times will have to) work together and help one another. However, you are personally responsible for the work you submit. For assignments, it is also your responsibility to ensure you understand everything your group has submitted and to make sure the correct file has been uploaded, that the upload is uncorrupted, and that it renders correctly. Projects may include ideas and code from other sources—but these other sources must be documented with clear attribution. Please review academic integrity policies here.

Know that a third of the class typically feels overwhelmed at the start of the quarter. That said, the average is quite high in this course typically (A-). So, while we anticipate you all doing well in this course, if you are feeling lost or overwhelmed, that's ok! Should that occur, we recommend: (1) asking questions in class, (2) attending office hours and/or (3) asking for help on Piazza.

Cheating and plagiarism have been and will be strongly penalized. If, for whatever reason, datahub is down or something else prohibits you from being able to turn in an assignment on time, immediately contact me by emailing your assignment by email (sellis@ucsd.edu), or else it will be graded as late.

Disability Access

Students requesting accommodations due to a disability must provide a current Authorization for Accommodation (AFA) letter. These letters are issued by the Office for Students with Disabilities (OSD), which is located in *University Center 202* behind Center Hall. Please make arrangements to contact Prof Ellis privately to arrange accommodations. If you are struggling to get a meeting with OSD, you can let Prof Ellis know and she's likely able to help accommodate while you work to get official documentation.
Contacting the OSD can help you further:
858.534.4382 (phone)
osd@ucsd.edu (email)
http://disabilities.ucsd.edu

How to Get Your Question(s) Answered and/or Provide Feedback

It's *great* that we have so many ways to communicate, but it can get tricky to figure out who to contact or where your question belongs or when to expect a response. These guidelines are to help you get your question answered as quickly as possible *and* to ensure that we're able to get to everyone's questions.

That said, to ensure that we're respecting their time, TAs and IAs have been instructed they're only obligated to answer questions between normal working hours (M-F 9am-5pm). However, I *know* that's not when you may be doing your work. So, please feel free to post whenever is best for you while knowing that if you post late at night or on a weekend, you may not get a response until the next day. As such, do your best not to wait until the last minute to ask a question.

If you have…

- **Questions about course content**: these are awesome! We want everyone to see them and have their questions answered too…so post these to Piazza!
- **A technical assignment question**: Come to office hours (or post to Piazza). Answering technical questions is often best accomplished in person where we can discuss the question and talk through ideas. However, if that is not possible, post your question to Piazza. Be as specific as you can in the question you ask. And, for

those answering, help your classmates as much as you can *without* just giving the answer. Help guide them, point them in a direction, provide pseudo code, but do **not** provide code that answers assignment questions.

- **Been stuck on something for a while (>30min) and aren't even really sure where to start**: Programming can be frustrating and it may not always be obvious what is going wrong or why something isn't working. That's ok - we've all been there! IF you are stuck, you can and should reach out for help, even if you aren't exactly sure what your specific question is. To determine *when* to reach out, consider the **2-hour rule**. This rule states that if you are stuck, work on that problem for an hour. Then, take a 30 minute break and do something else. When you come back after your break, try for another 30 minutes or so to solve your problem. If you are still completely stuck, stop and contact us (office hours, post on Piazza). If you don't have a specific question, include the information you have (what you're stuck on, the code you've been trying that hasn't been happening, and/or the error messages you've been getting).
- **Questions about course logistics**: First, check the syllabus. If the answer is not there, ask a classmate. If you still are unsure, post on Piazza
- **Questions about a grade**: For programming assignments, reply to the COGS 108 email directly; For project-related regrades, post a note to instructors on Piazza and select the 'regrades' tag. Include specifics as to why you feel you mistakenly/unfairly lost points in that post.
- **A specific section-related question**: send a direct message on Piazza to your TA/IA
- **Something super cool to share related to class**: feel free to email Prof Ellis or come to office hours. Be sure to include COGS108 in the email subject line and your full name in your message.
- **Something you want to talk about in-depth**: meet in person during office hours or schedule a time to meet by email. Be sure to include COGS108 in the email subject line.
- **Some feedback *about the course* you want to share anonymously:** If you've been offended by an example in class, really liked or disliked a lesson, or wish there were something covered in class that wasn't but would rather not share this publicly, etc., please fill out the anonymous Google Form*

*This form can be taken down at any time if it's not being used for its intended purpose; however, you all will be notified should that happen.

COURSE OUTLINE

Week	Topic(s)	Discussion Section	Due (Fri @ 11:59 PM)
1	Data Science, Ethics, & Version Control		--
2	Programming & Python	intro + datahub + git (A1)	--
3	Data, Parameterization & Data Wrangling	project groups/proposals	A1 - git + python (1/24)
4	Data Intuition & Data Visualization	pandas I (A2)	Project Proposal* (1/31)
5	Descriptive & Exploratory Data Analysis	pandas II (A2)	A2 - pandas (2/7)
6	Inference	dataviz/seaborn (A3)	A3 - Data Exploration (2/14)
7	Machine Learning	web pages + python (A4)	A4 - Data Privacy Project Check-In* (2/21)
8	Text Analysis	inference (A5)	A5 - Data Analysis (2/28)
9	Geospatial Analysis	text data (A6)	A6 - NLP (3/6)

10	Future of Data Science	advice	Group Work Survey (3/13)

Final Project*: due **Thursday March 19th** of finals week by 11:59 PM

*indicates group submission. All other assignments/surveys are completed & submitted individually.