Stat 231 – Data Science - Syllabus
Spring 2024 - Section 01

Meetings: MW 8:30-9:50 AM; Webster 102
Professor: Amy Wagaman — Email: awagaman@amherst.edu
Office: 306 Seeley Mudd — Please call me Professor or Professor Wagaman
Webpage: Moodle site
Office Hours: MF 2-3 pm, WTh 3-4 pm, or by appt.

Course Description:
Computational data analysis is an essential part of modern statistics and data science. This course provides a practical foundation for students to think with data by participating in the entire data analysis cycle. Students will generate statistical questions and then address them through data acquisition, cleaning, transforming, modeling, and interpretation. This course will introduce students to tools for data management, wrangling, and databases that are common in data science and will apply those tools to real-world applications. Students will undertake practical analyses of large, complex, and messy data sets leveraging modern computing tools.

Course Objectives\Learning Goals:
After successfully completing this course, you will be able to:

- Understand key data science terms and ideas, including ethical implications
- Apply the data analysis cycle – starting from questions, performing analysis, and communicating results
- Analyze data reproducibly with appropriate software
- Generate appropriate data visualizations with an understanding of how graphics are built (including both static and interactive graphics)
- Perform data wrangling with an understanding of the tidyverse
- Use version control (Git) to aid in project workflow
- Practice communicating about data science via group work, assignments, and projects
- Support fellow student learners via group work and improve your own group work skills

Pre-requisites and necessary background: STAT 135 (or equivalent, STAT 111, STAT 136, PSYC 122, AP Stat 5, etc.) **and** COSC 111, or consent of the instructor.

Required Resources:

1. Textbook: ***Modern Data Science with R***. Second Edition. Baumer, Benjamin S., Daniel T. Kaplan, and Nicholas J. Horton. *Modern data science with R*. CRC Press, 2021. The textbook is freely available online (link on Moodle as well): https://mdsr-book.github.io/mdsr2e/

2. Other: You will also need a computer, access to the internet, and either access to the R\RStudio server (which is automatically setup via your Amherst login) or R /RStudio/Posit on your personal machine, appropriate access to Git/GitHub, and access to Moodle, Zoom (for our remote times), and Gradescope. Both Zoom and Gradescope are accessible through Moodle. Instructions for installing R/Rstudio/Posit and getting set up for class with Git/GitHub are provided separately.

I ***strongly encourage*** you to obtain a copy of the text in a format that works for you, and to consider installing R/RStudio and Git/Github on your own computer for use in class.

Course Structure:
The course is divided into 4 major parts. Each part takes a few weeks to cover and is associated with a different course project. Class time is generally reserved for looking at examples, working through labs for practice with concepts, and working with groupmates on projects.

The major parts are:

1. Data Visualization, Wrangling, Reproducibility, and Ethics – Think of this as laying the foundation for what comes next, building up your skills working with data. The associated project is an individual one where you will apply the data analysis cycle. This will take roughly the first 4-5 weeks of class (overlaps slightly with the next).
2. Interactive Data Visualization / Building a Data Product – You'll focus on using your skills from part 1 in a group project to build a Shiny app to answer a question of interest to you. This will practice the data analysis cycle and show you how to build a data product (the app). This runs from roughly week 5 to week 8 of the course.
3. Selected Data Science Topics – We'll introduce a few selected data science topics – clustering, network analysis, spatial analysis, etc. during roughly weeks 6 through 11 of the course. These will allow you to pick an area of focus for the final course project.
4. Final Course Project – This project will involve you practicing the data analysis cycle with a question(s) that will allow you to explore one of the topics from part 3 more in depth. The remaining weeks of the course are devoted to work and presentations on this project. This project may be individual or group (TBD by me based on class size).

As we learn key data science terms and concepts, we will practice applying what we have learned via activities and problems and learn to implement relevant analyses in the software R in labs and in assignments.

Class time will be devoted to a mixture of short lectures, sharing of examples beyond the textbook, explanation of the associated code, and lab or other activities designed to practice the concepts. Class activities will often involve group work and I encourage you to get to know your peers to facilitate learning from one another and forming study/homework groups, as appropriate.

Course Outline:

Note that this is a tentative outline and is subject to change. It is provided to give you a sense of the course and rough timelines. Refer to Moodle for up to date scheduling and relevant changes, as well as more specific weekly reading assignments. Homework comes in 2 "types" described below in detail, just referred to as Hmk in the outline. An * next to a Hmk means it is Prep only, while ** means the associated Practice covers 1.5 weeks. (See Homework assignments section for details).

Week	**Topic(s)**	**Chapter(s)**	**Assignments Due**
0	Intro to R/RStudio, Regression	App B, App E	-
1	Intro to Data Science, GitHub	1, App D, 8.1, 8.8	Hmk0
2	Visualization	2, 3, 8.2	Hmk1; 1st Proj. Proposal
3	Data Wrangling and Tidy Data	4, 5, 6 (skip 6.4), 8.3, 8.4	Hmk2
4	Data Intake; Algorithmic Thinking	6.4, 8.5-8.7, 8.9-8.10	Hmk3
5	Shiny	14	Hmk4*; 1st Project; 2nd Proj. Proposal
6	Clustering, Text as Data	12.1, 19	Hmk5
7	Project Work	-	Pre-reflection; Data set
	Spring Break		
8	Networks, Project Presentations	20	Hmk6*; 2nd Project
9	Spatial Data	17	Hmk7**; Reflection
10	Iteration and Simulation	7, 13	Hmk8*; 3rd Project Proposal
	April Break (No class M)		
11	SQL and Project Work	15	Hmk9**
12	Project Work	-	-
13	Project Presentations	-	3rd Project

A final reflection activity will be assigned after the third project is submitted and will need to be completed during reading period or in the first two days of finals week (deadline will be posted closer to time). All other coursework will be completed before reading period.

Major Assignments - Expectations and Policy

Attendance/Participation:

- Attendance in class is expected. We will move quickly through material. Attendance will help your conceptual understanding via examples, activities, and includes practice with the software R and RStudio. Attendance is also vital for days that project work is indicated.
- Participation is just as important as attendance. You should be fully present and engaged in class, especially during group work, where everyone is expected to contribute.
- Repeatedly missing class (without an excuse), or repeatedly working on non-course related material in class are grounds for losing all participation points.

- Finally, participation also includes miscellaneous other activities, such as asking questions for clarification based on the reading, working with your peers for activities (aside from the project(s)), and completing evaluations of group work.
- I reserve the ability to adjust participation points if serious issues are raised in the evaluation of group work. If severe problems occur, individuals may be required to complete group projects individually, with a penalty (at my discretion) incurred.
- In the event that you miss a class, you are responsible for making up the material, such as getting notes from a classmate, or working through the posted activity on your own. If you have questions after working through an activity as best you can, you can discuss it with me during office hours or by setting up a separate appointment. This usually involves contacting me to learn what you need to make up. I strongly encourage you to email me as soon as you know you will miss a class.
- Your GitHub repo for the class should provide evidence of your participation and may be viewed by me at the end of the semester as such. For example, it should have copies of all your completed assignments and labs in an organized fashion.
- Final participation points are assigned at my discretion.
- Please turn cell phones **OFF or to silent** when you are in class!

Reading:

We cover a good bit of the textbook. It's a lot of material to get through. Reading assignments are provided on Moodle so that you can read the material **before** discussing the topics and our examples in class. Taking notes on the reading is strongly encouraged! Doing the reading will also help you tackle the homework – particularly the "Prep" component which is due every week on Sunday evening before class Monday.

I encourage you to refer back to the reading for clarification and examples as we continue working through the material. For example, we often have an activity after going over examples related to the reading topics. If you read before class, complete the "Prep" assignment, listen to lecture/examples and work through our lab for practice, and then try the "Practice" assignment, you can refer back to the reading and your notes if you have questions.

Homework Assignments:

- There are two types of homework assignments in our class. The first are "Prep" assignments and the second are "Practice" assignments.
- Prep assignments are designed to make sure you have completed the assigned reading and are *prepared* to participate in the class activities for the week. They are straightforward assignments from the reading.
- **Prep assignments are to be completed individually ONLY.**
- Prep assignments are due Sunday by midnight on their associated week to Gradescope and are graded for completion, unless otherwise stated (one may be due on a different day due to Spring Break).
- Tentatively, there are 9 scheduled Prep assignments with content.
- Practice assignments are designed to have you further practice concepts that we engaged with in class. You can usually start on them as soon as you are comfortable with the material, which may be right after completing the associated Prep

assignment, or you may want to wait until after Monday's activities. Group work is permitted and encouraged on practice assignments.

- Practice assignments are due Thursday by midnight on their associated week to Gradescope and are graded for correctness (including code readability). I reserve the right to extend this deadline to Friday (or other) evening, and if applied, this will be sent as notification to the class with a deadline change in Gradescope.
- Tentatively, there are 6 scheduled Practice assignments with content, plus one to be sure you can submit to Gradescope (Practice0). Two of the practice assignments cover 1.5 weeks instead of 1 and are due at the end of the second week to balance assignments with project due dates.
- For questions where explanations are required, you are expected to write FULL SENTENCES. We want to practice communicating about data science and statistics at all times.
- You must submit a .pdf for your submission for each assignment. I.E. compile the RMarkdown files to **.pdf**. (Otherwise, a scan of each is needed.) This should be straightforward following the templates and instructions for compiling.
- ***Tentative*** homework due weeks are in the schedule. Exact dates and any changes in the schedule will be set in Moodle and in Gradescope.
- There are no extensions possible for Prep assignments. For Practice assignments, I ***may*** offer a 24 hour late window, visible in Gradescope as a late deadline. If offered, homework turned in during the late window will be marked down 20%. All other late homework will be recorded as a 0. Submitting an assignment during this window acknowledges the penalty (you don't explicitly need to ask to submit it late).
- Both Prep and Practice assignments are worth differing numbers of points based on length and problem difficulty/work involved. Prep assignments are worth less than the corresponding Practice assignment.
- Start on assignments early so you can make use of resources to support your learning.
- **Remember that Prep assignments are to be completed individually**.
- For Practice assignments, you are encouraged to work with your fellow students. If you receive help from someone on a problem, you need to make a note of it on the top page of your homework (see template). However, there is a fine line between working together, getting help, and "copying"; the work you submit must be your own - **you should understand and be able to explain all parts of your submission**. Be sure you can explain what you have written. It doesn't help you learn if you copy a phrase or piece of code without knowing what it means or does.
- Under truly exceptional circumstances, I may grant a homework extension with a minimum of 24 hours prior notice via email. You should NOT expect to receive an extension just because you sent an email.

Typical Course Week:

- On Friday and the weekend, you should read the assigned material for the week and work on the associated Prep assignment (assuming there is one).
- Prep assignments are due Sunday by midnight. You can revisit reading as desired before class sessions on Monday and Wednesday.

- Class sessions on Monday and Wednesday will practice key ideas, and may involve short lectures/examples, lab activities, or discussions.
- Attend office hours as needed for questions or just to work in the company of others.
- Work on Practice assignments as soon as you are comfortable. I recommend that you read them over as soon as they are posted so you get a sense of what you will need to do.
- Practice assignments are due Thursday by midnight.
- When projects are running, some class periods may be devoted to project work, and there will be weeks with little (or no) reading, to allow you time to work on the projects.

Projects:

Projects are a major component of this class. Three are tentatively scheduled, one individual, one group, and one TBD. Groups will be assigned by me, and you will have opportunities for reflection on group interaction and performance. Projects are designed to help you practice the data analysis cycle and build your group work skills. More details about the projects will be announced as they are introduced (our first, the individual one, comes in just our second week of class!).

"Labs" and Other Activities:

We will use the R statistical environment with the Posit/RStudio interface and .Qmd/RMarkdown (.Rmd) files for structure throughout the course. Posit/.Qmd is a new update to the RStudio/.Rmd system. Most .Rmds you see can also run as .Qmds but .Qmds offer some more flexible options. I will be working to convert past course materials to .Qmds, but you may occasionally still see a .Rmd file.

Whatever format the files are in, you will need to bring your laptops to class (default: bring it). The link to the web server can be found on Moodle or just type in: https://r.amherst.edu/. You can also install R/RStudio/Posit on your own machine (instructions provided separately). We will use Git and GitHub for version control (more information provided separately). Again, it is encouraged to get the software on your own machine (especially if you intend to major in statistics).

Please contact IT (askit@amherst.edu) and cc me on the email if you experience issues with the RStudio/Posit server hosted by the College.

Coding is a big part of this course. You will be working to build your coding skills as we tackle various data science tasks. This will include learning coding etiquette (coding style), how to comment code, and using version control software. I will help with commands whenever asked, but you will also develop skill in figuring out how to answer coding questions yourself.

If you do not complete a lab during allocated class time, you should complete it later on your own. At the end of the semester, your GitHub repo should contain a completed pdf (and its associated .Rmd) of each of the class lab activities, from your code.

A host of R help is available to you. There are reference books available via the library, online help via a variety of websites and blogs, your fellow students, the Statistics and Data Science Fellows, and me. Please do not hesitate to ask me for assistance!

Grading Policy:	Class Participation:	10%
	Homework – Prep:	10%
	Homework – Practice:	20%
	Individual Project:	15%
	Group Project I (Shiny):	20%
	Final Project:	25%

Academic Integrity / Intellectual Responsibility

Be sure to familiarize yourself with the Amherst College Honor Code. Specific items to note for our class include:

On the use of generative AI for assignments, including coding – This course is designed to build fundamental knowledge for you in terms of the data analysis process. It will engage your intellect - helping you approach material/challenges and tackle it with your own knowledge and critical thinking skills. To support your growth this semester, the use of generative AI tools (artificial intelligence or machine learning tools such as ChatGPT or Github Copilot) for assignments or other course activities is **prohibited**, unless the assignment specifically directs you to use them.

In the event that an assignment directs you to use a generative AI tool, you must properly document and credit use of the tool. This would include listing the tool, date of query, the query, and description of where the result is used in your work.

Policies regarding generative AI may differ from instructor to instructor.

For assignments – the work you turn in must be your own.

For Prep assignments, you cannot work with other individuals. Your resources for those assignments are the textbook, me, and course examples.

For Practice assignments, you may work with other students in the class as you like, but must give credit where credit is due (details below). You may also get assistance from the SDS Fellows for these assignments.

You are expected to show appropriate supporting work for all questions. Solutions without appropriate supporting work will not receive credit.

Details: For Practice assignments, if you work with fellow students, you cannot copy a solution from another student. Instead, they may explain the problem to you or show you their work for part of a problem (e.g. their code), but you need to write your own solution to turn in. If you receive assistance like this, you need to list who you got help from (you do not need to list me if you come see me in office hours, listing the SDS Fellows is also optional). In particular, writing in assignments should NOT be similar, but you might discuss ideas with your classmates before completing questions that involve a decent bit of writing.

Use of the internet (or any other medium) to find copies of solutions to the problems (from any source) is strictly prohibited. You may not reference solutions provided to previous sections of the course by any instructor. You CAN use the internet to get help on what a particular function does, or to get details about how to address some issue you are encountering with R. Note the balance here – ideally you will use available resources to learn and tackle assignments, but you shouldn't be trying to find a complete solution somewhere to just copy.

Solutions should be presented in a well-organized, proofread, sentence format. Bullet point answers are not appropriate for our course. For assignments, spelling, grammar, etc. all count. You are strongly encouraged to proofread your submissions in advance of submission. RStudio has a spell check feature you should get used to using!

Coding readability and organization counts in this class! Be sure to review your code for style and readability before submitting assignments as well.

For group work – all group work should be equally shared and participation credit will be given to all group members. You should come prepared to every class to support your group's work for lab activities (which may be random groups or just who you are sitting with) and for the projects (groups assigned by me).

For the group project(s) – the group work should be equally shared and credit for the presentations and final deliverables will be given to all group members. All members are expected to be involved in ALL project stages. It is not appropriate for one person to do all the data intake, another to do the wrangling, and for another to be the sole author of the report, etc. Reflections of group work will be completed. Should reflections reveal extreme discrepancies in workload, I may make appropriate adjustments to project scores. You are expected to be honest about your contributions. Ideally, let me know immediately if there are issues with your group dynamics, so that I can assist.

If you are ever unsure about acceptable levels of collaboration, please ask! This will help you avoid unintentional violations of policy.

Class repo – A Github repo will be provided with all our class materials in it, and as the semester goes, it will contain solutions to many of those activities as well. It is a violation of the Honor Code to share this repo with students in future sections of the course. Just as you are not permitted to access past materials/solutions from previous course sections, you are not permitted to share this semester's materials with students in the course in future semesters (particularly in regards to solutions to assignments).

Communication

Please let me know how you want to be referred to both in terms of your name (NameCoach is good!) and your pronouns (he/him, she/her, they/them, etc.). Additionally, how you identify in terms of your gender, race, class, sexuality, religion, and dis/ability, among all aspects of your identity, is your choice whether to disclose (e.g., should it come up in classroom conversation about our experiences and perspectives) and should be self-identified, not presumed or imposed. I will do my best to address and refer to all students accordingly, and I ask you to do the same for all of your fellow classmates.

If you need to reach out and communicate with me, please email me at awagaman@amherst.edu. Please use a reasonable subject line (i.e. "Hmk 1 Question", not "HELP") and refer to me as Professor or Professor Wagaman in the email (and in class). I often sign emails with just my initials, ASW, to save time.

Please do not email me with questions that are easily found in the syllabus or on Moodle or in Gradescope (e.g. When is this assignment due?). Please do reach out about personal, academic, and intellectual concerns or questions. I will do my best to respond to emails within 24 hours. Do not expect replies for emails sent after 6 pm until the next morning (though you may often get them).

As a diverse community, it is important that we agree to conduct ourselves in a professional manner and that we work together to foster and preserve a classroom environment in which we can respectfully discuss and deliberate controversial questions as they arise. I will make every reasonable attempt to create an atmosphere in which each student feels comfortable voicing their argument without fear of being personally attacked, mocked, demeaned, or devalued.

Any behavior (including harassment, sexual harassment, and racially and/or culturally derogatory language) that threatens this atmosphere will not be tolerated. Please alert me immediately if you feel threatened, dismissed, or silenced at any point during our semester together and/or if your engagement in discussion has been in some way hindered by the learning environment.

Inclusion and Accessibility

I strive to make this course welcoming to all students. If you would like to discuss your learning needs with me, please schedule a meeting. I look forward to working with you to understand and support your academic success.

Your mental and physical health are foundational to your overall success and take priority over academic performance. We can work together to make adjustments to the course so that you can take care of personal needs while also demonstrating your knowledge of the material. Please let me know as soon as something arises so we can work together to ensure your success in the course. I do not expect you to give me all the details of your personal life—just make me aware that there is a situation. Additional support services are available that you should make use of, including your Class Dean, the Counseling Center, and student resource centers.

In particular, if you have a documented disability that requires accommodations, you will need to register with Accessibility Services for coordination of your academic accommodations. You can reach them via email at accessibility@amherst.edu, or via phone at 413-542-2337. Once you have your accommodations in place, be sure the ones you want to use for our class are in AC Data appropriately, contact me and I will be glad to meet with you privately during my office hours or at another agreed upon time to discuss the best implementation of your accommodations. Please arrange this meeting early in the semester!

We will learn to produce data products this semester (such as the Shiny app). It is extremely important that data products produced be accessible. Imagine making an app for someone to use only to find out they can't distinguish between the line types in a plot because the lines are too thin for them to easily see. We will discuss this as we work on various projects in class, and you may find it useful to refer to this webpage created by Amherst's ATS, which focuses on creating accessible documents.

In particular, to help keep materials accessible for each other this semester, please use at minimum, a 12 point font for all communication (including email) and submissions (this is the default in our files), and for your computer screen when sharing/discussing code with a classmate or me.

In addition, be sure to consider your color choices. To start, we'll use default palettes in R, but as we learn to customize those, the color choices become up to you. Finally, some of you know you can customize your RStudio interface (change colors, etc. to what you like). Be aware that a color scheme that works for you may not work for someone else, and if they ask you to make adjustments on their behalf, please do so (so long as your own ability to participate is preserved). Imagine if you are assigned to do a group activity, where one person can't really read the screen of the person who is taking notes/coding for the group because the color contrast is too low or they don't read a white text on black background well. We want everyone to be able to fully participate in class activities.

Campus Emergencies – Some Notes

- One reason I take attendance is to make sure everyone is accounted for in an emergency (such as firefighters working to clear the building). If you know you will miss class in advance, please be in touch with me – it also helps me keep you up to date on the material!

- Classrooms on campus can be locked without requiring a key. The door has a mechanism on the inside to lock it. If you'd like to be shown how this works, just let me know.
- Emergency gathering point – if there is a fire drill, real fire, etc. such that we have to evacuate the classroom, the gathering point for the class will be out by the War Memorial.

Helpful Hints for the semester:

- **Be prepared and attend class!** We move quickly through the material, and it is best not to fall behind**.**
- **Don't wait until the last minute**. If you wait until the last minute to do your work, you will be swamped, confused, sleep-deprived, and you will hate data science and statistics. Coding well takes longer than you might think!
- **Do not fear office hours!** You don't need an appointment to chat. Feel free to come just to work in the company of your peers. I won't "check" your answers, but if you want to work through a problem similar to one on the homework with me, that's perfectly fine.
- **Ask for help as soon as you are having problems**. The material builds on itself. Don't wait to ask for help. The foundational material from the first few weeks is very important for the later parts of the course – you'll be using it continuously.
- **Communication is very important in data science and statistics.** After all, what good is having the answer if you can't tell others (in plain English) what it is?
- **Use the TEXTBOOK!** Read it, re-read it. Take notes from it.
- **Check Moodle!** Always check the course Moodle site for updates on deadlines, etc.
- **Get to know and talk to your classmates!** Use the opportunities for group work in class to get to know your fellow students.
- **Pull the course content Git repo often!** This is how you'll get our examples, assignments, and labs. (We'll get Git set up in a separate assignment).