Fundamentals of Statistical Data Science

STA 141A

This class is about communication and computational reasoning. The goal is to provide you with skills necessary to modern statisticians and data scientists. The major themes of the class are:

- Creating graphics to gain insights from data.
- Writing about and interpreting results.
- Reasoning about problems that are open-ended or do not have a clear-cut solution.
- Computing with the R programming language.

Topics

For the first half of the quarter, we'll focus on techniques to explore and summarize data. In the second half of the quarter, we'll move on to cleaning messy data sets and implementing statistical algorithms.

Week	Topics
0	R syntax; RStudio; File paths
1	Data frames; Data types; ggplot2
2	Exploratory data analysis (EDA)
3	Designing graphics; Reproducibility
4	EDA; Spatial data; ggmap
5	Shapefiles; Writing functions; Debugging
6	Text files; String processing; Regular expressions
7	Regular expressions; Date processing
8	Tidy data; Relational data
9	Statistical learning; Efficient code
10	Resampling methods

There is no required textbook, but a list of references is posted on Piazza (the class forum).

Waitlist

This class has a long waitlist. If you decide you don't want to take the class, please drop immediately to make room for others. Note: the drop deadline is October 9th.

The Statistics Department PTA policy is at statistics.ucdavis.edu/courses/pta-policy. If you have any questions adding or dropping, contact Kim McMullen (stat-advising@ucdavis.edu).

Contacts

Name	@ucdavis.edu	Role
Nick Ulle	naulle	Instructor
Ken Wang	kenwang	TA
Patrick Vacek	prvacek	TA

We will only use email for private matters (grading, emergencies, etc.). Please **do not** email us about class material – post on Piazza instead.

Office hours are posted on Piazza.

Piazza

Piazza (www.piazza.com) is the class' online forum. You should have already received an email inviting you to Piazza. If you did not, or have any problems accessing Piazza, please email me or a TA as soon as possible. Note: the Piazza access code is `"sta141a"`.

Post your questions about class material on Piazza. Anyone in the class can answer your question, so you're likely to get an answer quickly. When you use Piazza:

- Be polite and respectful to others.
- Search before you post. Your question may have already been asked and answered.
- When you post a question, explain the context and give an example of what you mean.

All announcements will be posted on Piazza. Assignments and files will be posted on Canvas.

Grade Breakdown

Participation (15%) Participate in lecture, discussion, office hours, and Piazza. Asking questions is the best way to get help. Answering questions is a great way to verify that you understand the material.

Assignments (85%) There will be 2 short assignments followed by 4 long assignments.

We will grade assignments against a rubric that measures the quality of your writing, graphics, and code. The rubric will be posted before or with the first assignment.

No assignment grades will be dropped. If you can't turn in an assignment on time and have a legitimate excuse (such as medical emergency), we can arrange an extension or alternative.

Assignment Guidelines

- Write your solutions in report format, as if you are a professional data scientist or data journalist. In your writing, focus on conjectures and discoveries about the data rather than on programming details.
- Include all of your code in an appendix (at the end of your report). We will read and grade your code. Use a consistent style, and organize your code with comments and white space. There are many style guides online for R code.
- The goal of data science is to understand and gain new insights from data. Sometimes the best way to do this is to use a statistical model. However, do not use arbitrary models just to make your results seem more "statistical".

Academic Honesty

Even professional programmers talk to their coworkers and use references to help solve programming problems. So I encourage you to:

- Discuss the problems with your classmates.
- Search for references online and in books.
- Adapt short pieces of code (≤ 10 lines) you find on Piazza or online. When you do this, you must **cite the source**. For Piazza, cite the post number. For other sources, cite the title, author, and URL.

That said, **all writing and graphics must be your own work. A large majority of your code must be your own work.** If you're unsure whether something is okay, please ask!

The university code of academic conduct (sja.ucdavis.edu/files/cac.pdf) applies to this class. In particular:

> Students are responsible to know what constitutes cheating. Ignorance is not an excuse.