SDS 322E: Elements of Data Science
MW 1:00PM — 2:30PM, Room: UTC 4.134

Instructor

Dr. Antonio Linero, WEL 5.244

Office Hours: 2:00pm — 4:00pm Thursday, or by appointment.

Contact: antonio.linero@austin.utexas.edu

Teaching Assistant

Brandon Carter

Office Hours: 10:00am — 11:00am Tuesday and 9:00am — 10:00am Thursday, or by appointment.

Contact: carterjb@utexas.edu

Purpose and Content

This course is intended to be applied and hands-on. Through programming, discussions, and presentations, you will learn how to apply concepts, interpret analyses, and communicate results. You will develop crucial skills such as critical thinking, problem solving, and working collaboratively as well as independently.

Data Science is a dynamic and universal discipline. It is an umbrella term that covers many individual skills: data management, wrangling and manipulation, data workflow and programming, data visualization and communication, collaboration and data ethics, also including applied statistics and machine learning. In our increasingly data-driven world, this course will provide you with skills needed to solve problems, make predictions, and provide answers to research questions in the discipline of your choice.

You will work primarily in `R` (an open-source programming language and statistical computing environment) via RStudio, an integrated desktop environment (IDE). Though you will learn and use many programming concepts, this is not a computer science course: instead, we will be using these tools in an applied fashion to solve data problems.

What do we learn?

Below is a general outline:

1. In the first part of the course, we become familiar with `R`:
   - We go over programming topics that are fundamental to data science (control flow, data structures, iteration, vectorization, functions).
   - The notebook environment will be used throughout the course: with it, you will create RMarkdown documents which can be compiled to static documents/reports in PDF, HTML, and DOCX formats from R Studio (embedding code, text/markdown, media, math, etc).

2. Next, you will dive into a cutting-edge set of tools for data visualization and wrangling:
    - You will learn the `tidyverse` which includes a suite of industry-standard packages (`ggplot2`, `dplyr`, and `tidyr`) used for manipulating, visualizing, and interpreting patterns in data.
    - You will learn how to clean messy data, how to visualize the distribution of variables, how to plot relationships among variables, how to filter and select relevant information from datasets, and how to easily generate summary statistics. We will cover how to reshape datasets for different purposes, how to pull data from various sources, and how to join different datasets/tables together in the presence of complicating factors such as duplicate entries.
    - We will also explore how to process text data and use regular expressions to mine this type of data for insights.
3. In the third part of the course we will put these newly acquired tools to use in analyzing data, diving into exploratory data analysis and prediction. We will prioritize application over theory: Our main goal is for us to gain facility with important techniques (running them, interpreting the outputs, gaining intuition), to be able to judge when such techniques are appropriate, and to interpret research that makes use of them.
    - With exploratory data analysis, we take on unsupervised approaches to classification such as cluster analysis (grouping similar observations), as well as dimensionality reduction techniques (distilling a big set of variables down into a just a few that capture the most important trends).
    - We will also introduce supervised approaches to classification such as regression and tree-based methods. We then turn to assessing classification performance and then apply these same concepts to assessing how well our classification procedure would work on new observations, a process known as cross-validation.

Why `R`?

The R community is huge and supportive; its resources are amazing and abundant. There are thousands of packages, developed and vetted by researchers and statisticians all over the world, implementing cutting-edge algorithms, and statistical techniques (all free, well-documented, and open-source). This makes `R` extremely versatile. You can also generate high-quality reports, books, websites, slideshows, videos, animated GIFs, GitHub repositories, all from the comfort of RStudio.

Prerequisites

There are no strict prerequisites for this course besides basic numeracy and computer literacy.

Course Organization

We will meet twice a week for lectures and once a week for labs. To initiate interest in the key concepts before lectures, you will have reading assignments, and you can share your thoughts and ask questions through the discussion boards on Canvas. Worksheets will be used during class to practice concepts with examples and programming exercises. During labs, you will work on a short assignment within a group but you will submit individually.

Communication

I will frequently post information to the class using Canvas announcements. Please check your notifications settings to ensure that you receive immediate email versions of the announcements. For more personal question/feedback, feel free to email me at antonio.linero@austin.utexas.edu or through Canvas messaging. I will get back to you within 2 business days.

- In the email subject, include the code of the course (SDS 322E) with the reason for emailing: questions about an assignment (e.g., HW 1, Lab 2, Project 1, . . .), time conflict, sickness, etc.
- Questions about assignments that are due within 48 hours might not be answered on time so plan ahead! Don't expect answers during the weekend or after 5pm.

Course Materials

Textbook The textbook we will be using for this course is available for free at `https://rafalab.github.io/dsbook/`.

Technology Requirements This is a computing class: as such, lots of class time will be spent working through coding exercises; being able to follow along on your computer will be extremely important.

Download and install the statistical software package `R` (www.r-project.org) and the user-friendly interface RStudio (www.rstudio.com). *If you have used R before, make sure to update your version to the last available.*

You can also access RStudio via the server from the university https://edupod.cns.utexas.edu/. You will gain access to the server during the first week of class.

Textbooks There will be assigned readings for this course from excellent freely-available online textbooks.

- **`R` For Data Science:** http://r4ds.had.co.nz. Written in part by the infamous Hadley Wickham, this book is the foundational text of tidyverse. It is well written, concise, and easy to follow.
- **Introduction to Data Science: Data Analysis and Prediction Algorithms with `R`:** https://rafalab.github.io/dsbook/. This is an extremely good introduction to data science that we will rely upon to zoom in on specific topics.

Assignments and Grading

This course will have 12 graded homeworks, 13 labs, and 2 graded projects. Unless otherwise noted, a homework or project will be due at 11:59pm every Sunday, and labs will be due at 11:59pm every Wednesday. Both homeworks and projects must be submitted as pdf files on Canvas. Homeworks are worth 10 points each, projects are worth 150 points each, and labs are worth 5 points each. The two lowest-scoring homeworks will be dropped, so that a maximum of 100 points can be obtained from the homeworks.

In summary, there are 465 possible points in this course. **There are no traditional exams in this class and there is no final.** The class will use ± grading. The following minimum grades will be guaranteed:

- 432 points (93%): A
- 418 points (90%): A-
- 404 points (87%): B+
- 385 points (83%): B
- 372 points (80%): B-
- 358 points (77%): C+
- 339 points (73%): C
- 325 points (70%): C-
- 311 points (67%): D+
- 292 points (63%): D
- 279 points (60%): D-

While I reserve the right to lower the gradepoint cutoffs, I have never done so for this course in the past.

Grading

Deadlines for all assignments are strict. Please keep track of deadlines with the due dates on Canvas. There is a **10% penalty** for work turned in within 24 hours of the deadline, and a **20% penalty** for work turned in within 48 hours of the deadline; **after this, late work will not receive credit**.

Of course, I will be flexible in the event of extenuating circumstances, provided that any relevant documentation is provided. However, requests for accommodation should be requested promptly rather than after-the-fact. Make-ups for graded assignments are guaranteed only under the following circumstances:

- You are away from UT as part of a UT-sponsored activity, including athletics.
- The deadline is in conflict with a religious observance.
- You provide documentation for an illness or serious emergency.

Regrade policy Always review your feedback. If you think there may be an error in grading, we will be accepting regrade requests within a two-week limit. Straight-forward clerical errors will be taken care of directly, but if you have a more complicated regrade request, keep in mind that the entire assignment will be subject to re-grading, so your grade may go up or down as a result. Keep track of your progress on Canvas.

Academic Accommodations

I respect and welcome students of all backgrounds, identities, and abilities. I am committed to creating an effective learning environment for all students, but I can only do so if you discuss your needs with me as early as possible. Therefore:

- Please request a meeting as soon as possible to discuss any accommodations.
- Please notify me as soon as possible if the material being presented in class is not accessible.

Any student who requires academic accommodations should contact Services for Students with Disabilities as soon as possible to request an official letter outlining authorized accommodations. For more information, visit `http://ddce.utexas.edu/disability/about/`.

If You Need Help

All of us benefit from support during times of struggle. You are not alone. If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, I strongly encourage you to seek support.

CMHC Crisis Line is a confidential service that offers an opportunity for UT-Austin students to talk with trained counselors about urgent concerns. A counselor is available every day of the year, including holidays. You can call when you want, at your convenience. Telephone counselors will spend time addressing your immediate concerns. CMHC 24/7 Crisis Line: 512-471-CALL (2255).

BeVocal is a university-wide initiative to promote the idea that individual Longhorns have the power to prevent high-risk behavior and harm. At UT Austin all Longhorns have the power to intervene and reduce harm. To learn more about BeVocal and how you can help to build a culture of care on campus, go to: `https://wellnessnetwork.utexas.edu/BeVocal/index.html`

Tutoring is available from Sanger Learning Center (`https://ugs.utexas.edu/slc`) or from the Longhorn Center for Academic Excellence (`https://diversity.utexas.edu/academiccenter/academic-support/`). However, before you consider getting outside support for this class, please consider coming to ask for help during student hours.

What if I Am Feeling Sick?

Please do not come to in-person meetings if you are at all sick. I will be flexible with absences by providing make-up assignments for missed participation credit as long as you communicate with me about your absence within a reasonable time of the missed assignment AND there isn't a consistent pattern of requesting to miss class or turn in assignments late.

Please follow the university's recommendations regarding COVID-19:

- Adhere to current university guidance on COVID-19 safety, including when to get tested after exposure, masking policy, quarantine protocol, etc.
- UT encourages all students to be up-to-date on their vaccines/boosters. See the above link for information on getting a vaccine/booster through UT.

Diversity and Inclusion

In accordance with federal and state law, UT Austin prohibits unlawful discrimination, including harassment, on the basis of race, skin color, religion, national origin, gender, gender identity, gender expression, sexual orientation, age, disability, citizenship, and veteran status.

In a perfect world, data science would be objective. However, some studies are subjective and were historically built on a small subset of privileged voices. I acknowledge that some examples for this course may have biases due to the lens with which it was written. Integrating a diverse set of studies is important for a more comprehensive understanding of data science. Furthermore, I would like to create a learning environment that supports a diversity of thoughts, perspectives and experiences, and honors your identities. If something was said in class that made you feel uncomfortable, please talk to me about it. Learn about The Division of Diversity and Community Engagement at UT Austin: https://diversity.utexas.edu/about-ddce/

UT Austin provides upon request appropriate academic accommodations for qualified students with disabilities. Regardless of whether or not you plan to use your accommodations, bring me your letter within 2 weeks of receiving it. For more information, contact the Office of the Dean of Students at 471-6259.

Academic Dishonesty

This course is built upon the idea that team-based learning is an important and powerful way to learn. I encourage you to study and work on assignments together. However, **any work you turn in must be your own**. Everything turned in under your name must be from your brain. Simply copying another student's answers is always unacceptable. **Students who violate the academic honesty expectations for this class will be penalized, up to receiving a failing grade for the class and being reported to the Office of the Dean of Students for academic dishonesty.**

Students should be aware that project assignments will be submitted to the plagiarism-detection tool Turnitin, which is intended to address plagiarism and improper citation. The software works by cross-referencing submitted materials with an archived database of journals, essay, newspaper articles, books, and other published work. Other methods may be used to determine the originality of the paper, and this software is not intended to replace or substitute for my judgment regarding detection of plagiarism.

Under Title IX, I Am a Mandatory Reporter

Title IX covers sex discrimination liabilities for the university under federal law. It requires me to report any claims of sexual harassment, abuse, discrimination, and so forth (including vague hearsay), **regardless of whether the person confiding in me wants me to report or not.** If you need help I'm of course willing to do what I can (or, if I can't, direct you to those who can).

SHARING OF COURSE MATERIALS IS PROHIBITED

No materials used in this class, including, but not limited to, lecture slides, assessments, and other class materials, may be shared online or with anyone outside of the class unless you have my explicit, written permission. Unauthorized sharing of materials promotes cheating. It is a violation of the University's Student Honor Code and an act of academic dishonesty. I am well aware of the sites used for sharing materials, and any materials found online that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in sanctions, including failure in the course.

IMPORTANT SAFETY INFORMATION

If you have concerns about the safety or behavior of fellow students, TAs, or instructors, call the Behavior Concerns Advice Line (BCAL) at 512-232-5050. Your call can be anonymous. If something doesn't feel right, it probably isn't. Trust your instincts and share your concerns.

Tentative Course Schedule

Unit	Week	Day	Date	Lecture Topic	Lab Tue 11:59pm	Homework Sun 11:59pm
Tools + Visualization	1	M W	8/22 8/24	Introduction Introduction to R	—	—
	2	M W	8/29 8/31	R Programming Productivity Tools	Lab 1	HW 1
	3	M W	9/5 9/7	**Labor Day** Programming Concepts	Lab 2	HW 2
	4	M W	9/12 9/14	Programming Concepts Programming Concepts	Lab 3	HW 3
	5	M W	9/19 9/21	Visualization 1 Visualization 2	Lab 4	HW 4
	6	M W	9/26 9/28	Visualization 3 Visualization 4	Lab 5	HW 5
Data Wrangling	7	M W	10/3 10/5	Wrangling 1 Wrangling 2	Lab 6	HW 6
	8		10/10 10/12	Wrangling 3 Wrangling 4	Lab 7	HW 7
	9		10/17 10/19	Wrangling 5 Text Manipulation 1	Lab 8	HW 8
	10		10/24 10/26	Text Manipulation 2 Text Manipulation 3	Lab 9	**Project 1 Due**
EDA + Classification + Prediction	11		10/31 11/2	EDA 1 EDA 2	Lab 10	HW 9
	12		11/7 11/9	EDA 3 EDA 4	Lab 11	HW 10
	13		11/14 11/16	Classification and Prediction 1 Classification and Prediction 2	Lab 12	HW 11
	FB		11/21 11/23	**Fall Break**	—	—
	14		11/28 11/30	Classification and Prediction 3 Classification and Prediction 4	Lab 13	HW 12
	15		12/5	Classification and Prediction 5 **No Class**	—	
	16			**No Final**		