SDS 322E

Spring 2026

Course Title

ELEMENTS OF DATA SCIENCE

Course Information

Instructional Mode: Face-to-face
Meeting Times: Mon/Wed/Fri 10:00 AM - 11:00 AM; Thu 10:00 AM - 11:00 AM
Meeting Location: UTC 3.122; FAC 101B
Unique Number: 60415

Additional Sections: Unique Number 60420 Lab Thu 11:00AM - 12:00 PM

Instructor

Shuangjie Zhang
Email: sz23328@eid.utexas.edu

Office Hours and Location

Instructor Office Location: WEL 3.226G

Instructor Office Hours: Friday 11:45-1:45pm in WEL 3.226G

Graduate TA Office Hours: Th 2pm-4pm TBD

Welcome Message

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. Besides learning to program with R, we will use tools such as online repositories to collaborate on projects (e.g., GitHub). You will develop crucial skills such as critical thinking, analytic thinking, problem solving, and working collaboratively and independently.

Data Science is a dynamic and universal discipline. It is an umbrella term that covers many individual skills: data management, data wrangling, data workflow, data visualization, programming, communication, collaboration and data ethics, but 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 development 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 will we learn?

Below is a general outline of the four units of this course:

Unit 1

In the first part of the course, we become familiar with R:

- We go over programming and data science topics that are fundamental to data science (tidy data, summarizing data in 1- and 2-dimensions, basic visualization, hypotheses, informative tools, systems thinking).
- The notebook environment will be used throughout the course: with it, you will create R Markdown documents which can be compiled to static documents/reports in PDF, HTML, and DOCX formats from RStudio (embedding code, text/markdown, media, math, …).

Unit 2

In the second part, you will dive into a cutting-edge set of tools for data wrangling and tidying:

- Tidyverse 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.

Unit 3

In the third part of the course, we put these newly acquired tools to use in analyzing data, diving into classification 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.

- We will consider 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 apply these same concepts to assessing how well our classification procedure would work on new observations, a process known as cross-validation.

Unit 4

Finally, in the last part of the course we expand our horizons with a variety of extended topics that you may explore after the course is over. Topics may include causal inference and building dashboards.

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.

Catalog Description

Explore data science tools and examine data wrangling; exploratory data analysis and data visualization; markdown and data workflow; simulation-based inference; and classification methods. R programming is emphasized and Python programming is introduced.

Overview of the Class

We will meet three times 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. Exercises will be used during class to practice concepts with examples and programming exercises. During labs, you will work on a group assignment. A typical week looks as follows:

Typical weekly schedule

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
Lecture 10-11am		Lecture 10-11am	Lab 10am-11am, 11am-12pm	Lecture 10-11am	Homework due (if assigned)
			Pre-Lab Quiz due 12:01pm (if applicable) Lab (due at the end of the day at 11:59pm)	OH: 11:45am-1:45 pm (WEL 3.226G)	

Office Hours

Student office hours are times outside of scheduled class time for students to ask in-depth questions, explore any topic of interest that was not fully addressed in class, or discuss topics that need to be revisited. Meeting during instructor office hours is also a way for me to get to know you, understand your study habits and learn about your career aspirations.

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 the address above or through Canvas messaging.

Email Etiquette:

- 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, …
- 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 outside of normal business hours.

Textbooks

There will be assigned readings for this course from the following books, all of which are freely available online:

- *R Programming for Data Science* (Peng) - This books covers the foundations of the R programming language and how it can be used for both data analysis and for more general programming. A PDF copy of this book can be downloaded from the Canvas course site.
- *Exploratory Data Analysis with R* (Peng) - This book covers visualization principles and R functions. A PDF copyof the book can be downloaded from the Canvas course site.
- *The Art of Data Science* (Peng & Matsui) - This books covers more of the theoretical and general aspects of *doing*data analysis. A PDF copy of this book can be downloaded from the Canvas course site.
- *Tidyverse Skills for Data Science in R* (Wright, Ellis, Hicks & Peng) - This book focuses on the Tidyverse (including Tidymodels) collections of packages and how they can be used for data analysis. There is no PDF version of this book.

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 and the user-friendly interface RStudio (both are free):

Download R from cloud.r-project.org

Download RStudio Desktop from posit.co

If you have used R before, make sure to update your version to the latest available.

You can 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*.

Slides/Notes

You will be provided lecture slides and other materials on Canvas.

Lecture Recordings

This class is using the Lectures Online recording system. This system records the audio and video material presented in class for you to review after class. Links for the recordings will appear in the Lectures Online tab on the Canvas page for this class. You will find this tab along the left side navigation in Canvas.

To review a recording, simply click on the Lectures Online navigation tab and follow the instructions presented to you on the page. You can learn more about how to use the Lectures Online system at http://sites.la.utexas.edu/lecturesonline/students/how-to-access-recordings/.

You can find additional information about Lectures Online at: https://sites.la.utexas.edu/lecturesonline/.

Tentative Schedule

Tentative Course Schedule for Spring 2026

Tentative Schedule of Topics (Spring 2026)

Unit	Week	Date	Lecture Topic	Lab TH 10a/11a	RP/Others F 11:59pm	HW Sa 11:59pm
Unit 1: Analytic Thinking and Visualization	1	1/12	Introduction Analysis Process; R Language	Lab 0	Intro Survey	–
	2	1/19	Summarizing data Information and informative tools	Lab 1		HW1
	3	1/26	Visualizing Data I Systems thinking in analysis	Lab 2		–
	4	2/2	Visualizing Data II Systems/skeptical thinking	Lab 3		HW2
Unit 2: Data Wrangling	5	2/9	Data transformation Reshaping data	Lab 4		HW3
	6	2/16	Joining/Merging Tidying	Lab 5		HW4
	7	2/23	tidyr; Regular expressions Missing data	Lab 6		HW5
Unit 3: Modeling and Analysis	8	3/2	Tidytext; Multi-dimensional data	Work on Project		Project 1
	9	3/9	Dimension reduction PCA	Lab 7		–
		3/16	**Spring Break**			
	10	3/23	Statistical learning Linear regression	Lab 8		–
	11	3/30	Logistic Regression Non-parametric methods	Lab 9		HW6
Unit 4: Expanding Horizons	12	4/6	Prediction metrics Tidymodels I	Lab 10		HW7
	13	4/13	Tidymodels II Causal Inference I	Lab 11		HW8
	14	4/20	Causal Inference II Dashboards	Lab 12		–
	15	4/27	TBD			–
No Final Exam				Project 2 (4/30)		–

Prerequisites for the Course

Prerequisite: An introductory statistics course. Only one of the following may be counted: Statistics and Data Sciences 313; 322E; 348.

Learning Outcomes

At the conclusion of this course, students will be able to:

- Write R code and develop data analyses using R markdown documents

- Use Tidyverse tools to wrangle, tidy, process, clean, and visualize data

- Develop unsupervised and supervised machine learning models to quantify relationships in high-dimensional data

Grading Policy

Deadlines for all assignments are strict. Please keep track of deadlines with the due dates and road map on Canvas. We need to have deadlines in order to organize the grading and provide feedback to all students in a timely manner. If unforeseen circumstances prevent you from submitting on time, communicate with me immediately, preferably prior to the deadline or within a reasonable time of the missed deadline.

Make-ups for any 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. Final grades will be assigned according to the overall percentage ranges listed below. Please note that grade cutoffs are firm and there is no curve:

Table describing grad ranges

A	**A-**	**B+**	**B**	**B-**	**C+**	**C**	**C-**	**D+**	**D**	**D-**
>= 93.0%	90.0–92.9%	87.0–89.9%	83.0–86.9%	80.0–82.9%	77.0–79.9%	73.0–76.9%	70.0–72.9%	67.0–69.9%	63.0–66.9%	60.0–62.9%

Generative Artificial Intelligence

The University encourages all students to engage with AI responsibly and to recognize the important limitations of using generative AI for learning. The University offers specific guidance for responsible adoption of AI for learning and the limitations of using

generative AI for learning that all students should review regularly. In this course, students are 100% responsible for all academic work they produce.

The use of generative artificial intelligence tools (or Large Language Models [LLMs]) such as CoPilot, UT Spark or ChatGPT in this class is strictly prohibited. This includes using AI to generate ideas, outline an approach, answer questions, solve problems, or create original language. All work in this course must be your own or created in group work, where allowed.

Overview of all Major Course Requirements and Assignments

Weekly Homework Assignments 25%: There will be 8 homework assignments for this course, due on Gradescope by Friday, 11:59pm the week they are due. You will submit a knitted .Rmd file as a .pdf (we will cover the details in class). The lowest homework assignment will be dropped. Because homework solutions may be discussed during the next meeting, no late submission will be accepted.

Weekly Lab Assignments 25%: There will be 12 lab assignments to be completed with your assigned group during your scheduled lab hour. The labs are relatively low-stakes and designed to give you additional practice but also learning new concepts. Some will be more challenging, so it is always a good idea to organize your notes beforehand. The two lowest lab assignments will be dropped.

Projects 45%: The research project is the culminating experience of the course! There will be three different parts throughout the course:

1. Design: Specifying a data science question and identifying a dataset
2. Wrangling and Visualization of your data
3. Modeling and formal analysis

For each report you will produce a written document that will be peer-assessed. You may collect data yourself or use data sources available online. Links to possible data sources will be posted on Canvas as the semester progresses.

Others 5%: Other assignments will provide opportunities for personal and professional development or provide feedback to course staff.

Required Course Materials

Exploratory Data Analysis with R
Authors: Roger D. Peng
Publisher: Leanpub
This book covers visualization principles and R functions. A PDF copy of the book can be downloaded from the Canvas course site.

The Art of Data Science
Authors: Elizabeth C. Matsui and Roger D. Peng
This books covers more of the theoretical and general aspects of *doing* data analysis. A PDF copy of this book can be downloaded from the Canvas course site.

Tidyverse Skills for Data Science in R
Authors: Carrie Wright, Shannon Ellis, Stephanie HIcks, and Roger D. Peng
This book focuses on the Tidyverse (including Tidymodels) collections of packages and how they can be used for data analysis. There is no PDF version of this book. You can find the book online at http://jhudatascience.org/tidyversecourse/

R Programming for Data Science
Authors: Roger D. Peng
Publisher: Leanpub
This books covers the foundations of the R programming language and how it can be used for both data analysis and for more general programming. A PDF copy of this book can be downloaded from the Canvas course site.

A Notice of Academic Accommodations from Disability and Access (D&A)

If you are a student with a disability, or think you may have a disability, and need accommodations please contact Disability and Access (D&A). You may refer to D&A's website for contact and more information: http://disability.utexas.edu/. If you are already registered with D&A, please deliver your accommodation letter to me as early as possible in the semester so we can discuss your approved accommodations.

University Policies and Resources for Students Canvas Page

This Canvas page is a supplement to all UT syllabi and contains University policies and resources that you can refer to as you engage with and navigate your courses and the university.

Canvas Assignments

Due Date	Assignment Name	Points
	Class GroupMe	0
1/15/26	Lab 0: Software and Markdown	6
1/16/26	Introduction Survey	10
1/22/26	Lab 1: Pre-Lab Quiz	5
1/22/26	Lab 1: Intro Lab	22
1/24/26	Homework 1	20
1/29/26	Lab 2: Pre-Lab Quiz	5
1/29/26	Lab 2: ggplot2	20
2/5/26	Lab 3: Pre-Lab Quiz	5
2/5/26	Lab 3: Visualizations	20
2/7/26	Homework 2	22
2/12/26	Lab 4: Pre-Lab Quiz	5
2/12/26	Lab 4: Data wrangling with dplyr	20
2/14/26	Homework 3	20
2/19/26	Lab 5: Joining and Merging Data	20
2/21/26	Homework 4	20
2/26/26	Lab 6: Pre-Lab Quiz	5
2/26/26	Lab 6: Tidying data	20
2/28/26	Homework 5	20

Due Date	Assignment Name	Points
3/7/26	Project 1	100
3/12/26	Lab 7: Pre-Lab Quiz	5
3/12/26	Lab 7 - Text mining	20
3/26/26	Lab 8 - Clustering and PCA	20
4/2/26	Lab 9 - Predictions with Linear Regression	20
4/4/26	Homework 6	20
4/9/26	Lab 10 - Classification	20
4/11/26	Homework 7	20
4/16/26	Lab 11: Pre-Lab Quiz	4
4/16/26	Lab 11: Predicting Air Pollution in the U.S.	22
4/18/26	Homework 8	20
4/21/26	Project 2 Preliminary Report	10
4/23/26	Lab 12: Propensity Scores	20
4/30/26	Project 2 Final Report	100

Final Exam Date and Time

The Final Exam Schedule is available here.

Class Canvas Site Weblink

https://utexas.instructure.com/courses/1444537/