CSE 2600 Fall 2024

Introduction to Data Science and Engineering

Instructor: Matt Lamoureux

Instructor email: `matthew.lamoureux@uconn.edu`

Office: Science 1 MZ1 120

Office Hours: Mon 1-2 in SCI1 office, Wed 10-11 in ITE lobby.

TA info: TBD

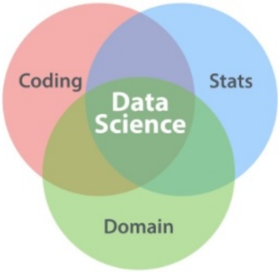
Text: *An Introduction to Statistical Learning with Applications in Python*,
by G. James, D. Witten, T. Hastie, R. Tibshirani, & J. Taylor.
Available for download at `www.statlearning.com`

Supplies: Access to Python, Jupyter Notebooks, and a scientific calculator

Prerequisites: CSE 2050 (Data Structures).

Learning Objectives: Prepare for career readiness and/or research in data science.

(1) (the foundation) Understand the role and importance of data science in various domains.
(2) (the engineering) Preprocess, clean, and manipulate large datasets.
(3) (the science) Apply statistical methods and machine learning algorithms to derive insights.
(4) (the communication) Visualize data and communicate analysis effectively.



Source: `https://www.youtube.com/watch?v=r2I3IDKwyMw`

Alignment to Career Readiness Competencies: See `www.naceweb.org` for more detail.

(1) *Teamwork*: Conduct a data exploration project with peers and report on findings.
(2) *Communication*: Present on project work that would speak effectively to both technical and non-technical audiences.
(3) *Critical Thinking*: Participate in real-time predictive modeling activities during Friday classes.
(4) *Equity & Inclusion*: Discuss the ethical considerations in practicing data science.
(5) *Technology*: Leverage Github to share project work effectively.

Academic Integrity:

Integrity is a crucial part of the academic experience. You must observe the University's Academic Integrity Policy as found in the Student Code. Cheating can result in one or more of the following: a score of zero on the assignment; a grade of F in the course; expulsion from the university and/or any subsidiary programs.

Summary of Grades:

Participation	Friday activities	10%
Homework	four assignments	20%
Midterm Assessment	tentatively 10/9	30%
Final Project Report	submitted by 12/6	20%
Final Project Presentation	beginning end of Nov	20%

Participation:

- Because this course is fast-paced, attendance in lectures will be critical to achieving learning outcomes. Take notes that can help with future work, and ask questions as you encounter new and/or confusing material.
- Each Friday class will be dedicated to lab-style coding work done in real time. These activities are meant to better solidify the concepts seen in lectures.
- There will be weekly assessments submitted on HuskyCT to help us (student and instructor) learn what concepts have or haven't been mastered. These are not meant to be intimidating, and should not take more than 10 minutes to complete. These will be due Fridays at noon, and you can miss 3 without penalty—if you don't send me any emails requesting extensions.

Homework:

- There will be 4 assignments that must be submitted via one PDF on HuskyCT.
- I am happy to help! Feel free to attend office hours or reach out with questions.
- Deadlines will be available on HuskyCT and generally at midnight. You can submit an assignment late only if: (1) grading has not yet begun and (2) the solutions have not yet been posted. There is no need to email me. To be safe, start each assignment early and submit whatever work is completed by the deadline.
- It is **highly** recommended that students form study groups and work together on homework assignments. Please be sure to include the names of your study group members on your homework submission, as a form of acknowledging their contribution.

Midterm Assessment:

- Multiple choice questions ($\sim$25%) based on information from the lecture slides will assess fact recognition, important definitions, etc. Some of these questions may be directly taken from hints that I've provided during class.
- Free response questions ($\sim$75%) will allow you to apply methods from class and are likely to be formatted similarly to homework questions.
- Requests for accommodations (through CSD) are welcomed, and these should be arranged at the beginning of the semester, or at least a week before an exam.
- If you cannot attend the exam for any reason, notify me at least a week in advance. Regarding exceptions: I reserve the right to decide what constitutes a last-minute emergency.

Final Project:

- You will collaborate in groups of 2 or 3 on a topic of your choice. Once groups are finalized, you will need to submit a short abstract describing your project proposal on HuskyCT.
- Both the live presentation (in class) and the written report will allow you to demonstrate your ability to communicate data analysis effectively.
- A more formal rubric will be given after the midterm exam.

Grading Scheme: The usual one.

Points	**Grade**
93-100	A
90-92	A-
87-89	B+
83-86	B
80-82	B-
77-79	C+
73-76	C
70-72	C-
67-69	D+
63-66	D
60-62	D-
00-59	F

Tentative Schedule: Subject to change.

Week	**Date**	**ISLP**	**Topic**
1	8/26		Data Objects and Attributes
			Measures of Central Tendency
2	9/2		Data Visualizations
3	9/9	Ch 2	Foundations of Modeling
4	9/16	Ch 3	Simple Linear Regression
			Binning & Normalization
5	9/23	Ch 3	Multiple Linear Regression
			Correlation
6	9/30	Ch 4	Classification
			Probability
7	10/7		Review
			Midterm Exam
			Ethics
8	10/14	Ch 5	Resampling
		Ch 8	Bagging
9	10/21	Ch 8	Boosting
			Model Comparison
10	10/28	Ch 12	Principal Components Analysis
		Ch 12	Clustering
11	11/4		TBD
12	11/11		TBD
13	11/18		Databases & SQL
14	12/2		**Project Presentations**