CMSC 462 — Introduction to Data Science

Semester: Fall 2025
Lecture Time: Monday & Wednesday, 10:00–11:15 AM
Location: Janet & Walter Sondheim 205
Instructor: Dr. Justin Brooks
Office Hours: By appointment

Course Overview

Data science sits at the intersection of statistics, computer science, and domain knowledge. In today's world, datasets are large, messy, and ever-changing. Tools powered by artificial intelligence can write code, generate models, and summarize data. But without a solid understanding of the fundamental concepts, these tools can be misused or misinterpreted.

This course is designed to help students build the *conceptual core* of data science. Rather than focusing on specific tools or packages, we will explore how to think like a data scientist: how to frame a research question, examine a dataset, clean and wrangle it, choose appropriate modeling strategies, evaluate results, and communicate findings.

A unique aspect of this course is the inclusion of time series analysis and signal processing. These topics are often overlooked in introductory courses, yet they are essential in real-world contexts like forecasting, behavioral analytics, financial modeling, and healthcare monitoring.

Learning Goals

By the end of this course, students will:

- Be able to receive a dataset and research question and analyze them appropriately.
- Understand the conceptual building blocks of data science: data types, visualization, modeling, and evaluation.
- Recognize the difference between correlation and causation.
- Apply foundational techniques in time series and signal processing.
- Make informed decisions in messy, ambiguous analytic scenarios.
- Communicate their methods and findings clearly.

Why This Course is Structured This Way

The weekly topics in this course are ordered to mirror a real data science workflow:

1. **Weeks 1–5**: We begin with foundational thinking—what data science is, what kinds of data exist, and how to explore them through multiple layers of exploratory data analysis. Students build toward a real-world workflow by layering EDA concepts, data structure handling, and data “wrangling”.

2. **Weeks 6–7**: We introduce statistical reasoning—causality vs. correlation and the foundations of probability and inference.

3. **Week 8**: The first exam tests conceptual mastery of EDA, wrangling, causality, and inference.

4. **Weeks 9–13**: We dive into modeling—regression, classification, model evaluation, unsupervised learning, and time series/signal processing.

5. **Week 14**: The second exam checks students' grasp of modeling and time-based analytics.

6. **Weeks 15–16**: Final project presentations and summative evaluation, where students demonstrate the full data science workflow—from exploration to modeling to insight.

The structure is intentional: build core understanding, introduce statistical judgment, and then develop deeper modeling capabilities within a practical workflow framework.

Course Materials

There is no required textbook. Selected readings, datasets, and tools will be provided through the course. Examples include open-source notebooks, tutorials, and current research.

Attendance and Format

This course is **in-person only**. There is no remote option, and **no recordings** will be made. Students are responsible for catching up on missed material through classmates. This will be discussed during the first week of class.

Assignments and Grading

Component	**Weight**
Exam 1: EDA, Causality & Inference	20%
Exam 2: Modeling & Time Series Concepts	20%
Final Project	30%
Homework	20%
Participation	10%

- **Homework** will consist of conceptual exercises and light coding tasks, designed to reinforce in-class material. It will not be heavily graded but is essential for learning. Homework assignments will be provided on Monday and are *due 5pm ET the Friday* of that week.

  Planned Homework Topics & Timing:

  - **HW 1** (Week 2): Data types, basic structures, variable interpretation
  - **HW 2** (Week 3): EDA: descriptive stats, summaries, visualizations
  - **HW 3** (Week 5): Data wrangling: reshaping, missingness, joins
  - **HW 4** (Week 6): Causality vs. correlation + probability basics
  - **HW 5** (Week 7): Hypothesis testing + interpretation
  - **HW 6** (Week 10): Modeling I: regression/classification evaluation
  - **HW 7** (Week 12): Unsupervised learning: clustering, PCA

  - **HW 8** (Week 13): Time series concepts + workflow reflection (pre-final project)

- **Exams** will be written and subjective, focused on conceptual reasoning and clear articulation of ideas.
- **Final Project** will involve analysis of a real-world dataset, applying the full end-to-end workflow.
- **No group work** is allowed.

Use of AI Tools

We recognize that tools like **ChatGPT** and **Claude (Anthropic)** are part of modern work. We do not encourage or discourage the use of these tools. However, all assignments and projects must include a short disclosure paragraph indicating:

- What tool(s) were used
- For what part(s)
- How the output was used or modified

This policy is for transparency only. **AI use does not affect grades.**

Accommodations, Inclusion, and Integrity

We adhere to all UMBC policies on academic integrity, disability accommodations, and classroom inclusion. If you need accommodations, please contact SDS and arrange to meet with the instructor. Academic dishonesty, including plagiarism and uncredited code copying, will not be tolerated.

Final Note

This course emphasizes **thinking clearly about data**. The tools may change, the platforms may evolve, and AI may become more capable—but the ability to interpret, question, and explain your analysis will remain invaluable. That is the true goal of this class.

Weekly Breakdown & Topics

The weekly topics in this course are ordered to mirror a real data science workflow, with each topic typically covered over one or two lectures depending on complexity. This breakdown maintains the conceptual structure outlined earlier while aligning all lecture counts to the 16-week semester.

Week	Lecture(s)	Topic(s)
1	Lecture 1	Intro to Data Science – What is DS, course overview, tools, mindset
2	Lecture 2	Data Types & Structures – Variables, types, basic formats
3	Lecture 3, 4	Exploratory Data Analysis (EDA) – Missing data, Descriptive stats, visualizations
4	Lecture 5, 6	Causality vs. Correlation – Simpson's Paradox; Data Sources & Collection
5	Lecture 7, 8	Data Wrangling I & II – Cleaning, reshaping, transforming
6	Lecture 9, 10	Probability & Inference – Random variables, CI, p-values
7	Lecture 11, 12	Hypothesis Testing – z, t, chi-squared tests, practice problems
8	Lecture 13, 14	Midterm 1 Review & Exam (EDA, Wrangling, Causality, Inference)
9	Lecture 15, 16	Regression I & II – Linear modeling, assumptions, diagnostics
10	Lecture 17, 18	Classification I & II – Logistic regression, evaluation

11	Lecture 19, 20	Model Evaluation & Feature Engineering – Overfitting, transformation
12	Lecture 21, 22	Unsupervised Learning – Clustering (K-means), PCA, t-SNE
13	Lecture 23, 24	Time Series & Signal Processing – Trends, seasonality, FFT
14	Lecture 25	Midterm 2 Review & Exam (Modeling & Time Series)
15	Lecture 26, 27	Capstone Project Work Time / Feedback Sessions
16	Lecture 28	Final Project Submissions/ Wrap-Up