ECON 122: Data Science and Statistical Learning

Claremont McKenna College, Fall 2025

Instructor: Prof. Michael Gelman (mgelman@cmc.edu)

Office Hours: Mo/We 1:00-2:00 PM, Bauer 216

Class Schedule:

- **CM-01:** Mo/We, 09:35 AM - 10:50 AM, BC35
- **CM-02:** Mo/We, 11:10 AM - 12:25 PM, BC35

Course Tutors:

- Joshua Zhou: BC 24, Mo 08:00-10:00 PM
- Asher Frye: BC 22, We 08:00-10:00 PM

Course Website: A GitHub page with course materials will be maintained.

Course Summary

This course provides a comprehensive introduction to the fundamental concepts and techniques of data science with a strong emphasis on machine learning models. Using the R programming language and the Tidyverse ecosystem, students will learn the entire data science pipeline, from data acquisition and cleaning to exploratory data analysis, visualization, and model building. The course will cover both supervised and unsupervised learning algorithms, including linear and logistic regression, decision trees, and clustering. The primary goal is to equip students with the practical skills needed to analyze real-world datasets and communicate data-driven insights effectively.

Course Objectives

Upon successful completion of this course, students will be able to:

- Navigate the R programming environment and use core Tidyverse packages for data manipulation and visualization.
- Apply principles of exploratory data analysis to understand and summarize datasets.
- Formulate and implement supervised machine learning models for regression and classification tasks.
- Utilize unsupervised learning techniques to discover patterns in data.
- Critically evaluate and compare the performance of different machine learning models.
- Effectively communicate the results of a data analysis project through written reports and presentations.
- Work collaboratively on data science projects using version control.

Prerequisites

- ECON 50
- ECON 120 Statistics (or equivalent)
- Another college-level statistics course with permission of the instructor
- CSCI 040 Computing for the Web (or equivalent)

Required Materials

- **Primary Textbook:** *R for Data Science* by Hadley Wickham & Garrett Grolemund. Available for free online.
- **Supplemental Text:** *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. This book provides a more in-depth theoretical foundation. A free PDF is available online.
- **Software:** R, RStudio (Desktop Edition), Git, and a GitHub account. All are free and open-source.

Evaluation Structure

- **Overall Weighting:**
  - Class Attendance/Participation: 5%
  - Problem Sets: 15%
  - Team Projects: 15%
  - Midterm 1: 20%
  - Midterm 2: 20%
  - Final Project: 25%
- **Exams:**
  - Midterm 1: September 29
  - Midterm 2: November 24
- **Problem Sets:**
  - Assigned regularly on GitHub.
  - Collaboration is allowed, but submissions must be individual.
  - Grading is based on a "good-faith effort" (2 points), insufficient effort (1 point), or no submission (0 points).
  - Late assignments are not accepted, but the lowest grade will be dropped.
- **Team Projects:** Two projects assigned during the semester, with collaboration on GitHub.
- **Final Project:** A group project (3-4 students) with a clear question and answer, using outside data and course skills.

Policies

- **Disability:** Students with a disability must notify the instructor within the first two weeks of the semester. Alternative exam arrangements must be coordinated with the Dean of Students at least two weeks prior to the exam.
- **Re-grading:** Concerns about grading must be raised within 7 days of receiving the graded item. The entire assignment or exam will be re-graded, which could result in a lower score.

- **Academic Dishonesty:** Any academic dishonesty will be taken extremely seriously and reported to the Academic Standards Committee.
- **Absence:** There are no make-up exams. Missed exams due to a documented emergency will be handled on a case-by-case basis.

Course Topic Outline

- **Part I: Foundations of Data Science**
  - Foundations & Reproducibility: 1 week
  - Data Wrangling: 1 week
  - Exploratory Data Analysis (EDA): 1 week
  - Advanced Data Wrangling & Feature Engineering: 1 week
- **Part II: Supervised Learning**
  - Linear Regression: 1 week
  - Logistic Regression: 1 week
  - Model Evaluation: 1 week
  - Resampling Methods: 1 week
  - Decision Trees: 1 week
- **Part III: Unsupervised and Ensemble Methods**
  - Unsupervised Learning - Clustering: 1 week
  - Ensemble Methods: 1 week
  - Final Project & Conclusion: 1 week