Data Science and Statistical Learning

Claremont McKenna College, Fall 2023

ECON 122 CM-01: Mo/We, 09:35 AM - 10:50 AM, BC35

ECON 122 CM-02: Mo/We, 11:10 AM - 12:25 PM, BC35

Instructor: Prof. Michael Gelman (mgelman@cmc.edu)
Office Hours: Mo/We 1:00-2:00 PM Bauer 216

Course Tutors:

Jen Lim	BC 24, Mo 08:00-10:00 PM
Abizer Mamnoon	BC 22, We 08:00-10:00 PM

Website: I will maintain a GitHub page with course materials.

Course Summary:
Economists traditionally use datasets compiled by statistical agencies (macroeconomic data) or collected via survey (microeconomic data). With the rise in internet and mobile phone usage, many datasets that were not initially designed for economic analysis (Yelp, Zillow, Twitter, etc.) can be readily captured and transformed for use in economic analysis. This course will cover methods to collect, clean, and transform data from traditional and non-traditional sources. In order to analyze this data, we will cover supervised statistical learning (decision tree, neural networks) as well as unsupervised learning (clustering) methods.

Course Objectives:
At the completion of this course, students will be able to:

- collect data from various sources
- clean and transform data for the purposes of analysis
- understand different statistical models used to analyze data and make predictions
- apply various statistical models to better understand and to make predictions with our data

Prerequisite:

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

Textbook:
Modern Data Science with R by Baumer, Kaplan and Horton (1st Edition).

Technology:
We will use R and RStudio for all the exercises, problem sets, and projects throughout the course.

Evaluation Structure:

Exams:
Midterm 1 is 10/11.
Midterm 2 is 11/29.

Problem Sets:
Problem sets will be assigned regularly from GitHub. You may work with others but need to write up and submit your assignment on your own. If you put in a good-faith effort to get the correct answers, you will get full credit for that submission. Points are assigned as follows:

2 points: The submission is a "good faith effort."
1 point: Assignment submitted, but there is insufficient effort. Many questions, or important questions, were skipped. The work is sloppy, unclear, or inadequate to derive the answers given.
0 points: No assignment submitted.

No late assignments are accepted but the lowest grade will be dropped at the end of the semester.

Team Projects:
There will be a few team project assigned throughout the semester. You will work with 2-3 other members and collaborate using GitHub.

Final Project:
You will work in groups of 2-3 students on a project that meets the following criteria:

- It must have a clear question and answer
- It must use data not provided in this course
- It must use the skills and models learning in the course

Overall Weighting:

Class Attendance/Participation	5%
Problem Sets	15%
Team Projects	15%
Midterm 1	20%
Midterm 2	20%
Final Project	25%

Policies:

- **Disability:** If you have any problems with the terms of this syllabus due to disability, you must notify me within the first two weeks of the semester. If you need alternative arrangements for exams I must be contacted by the Dean of Students, but it is also your responsibility to coordinate with me about the time and place of your exam at least two weeks before that exam.
- **Re-grading:** If you have concerns about the grading of a question on an assignment or exam, raise the issue within **7** days of receiving the graded assignment or exam. I will then re-grade the entire assignment or exam. *Keep in mind that this means there is a possibility of walking away from the re-grade with a lower score.*
- **Academic Dishonesty:** Please be aware that any incidence of academic dishonesty (plagiarism, cheating, etc.) will be taken **extremely** seriously. All cases will be reported to the Academic Standards Committee immediately.
- **Absence:** There will be no make up exams. If an exam is missed due to a legitimate and documented emergency, I will coordinate with the student to come up with a solution. I reserve the right to approve all such cases.

Course Topic Outline:

- Intro (1 week)
- Data visualization (2 weeks)
- Cleaning and reshaping data (2 weeks)
- Importing and merging data (2 week)
- Statistical learning overview (1 week)
- Evaluating models (2 weeks)
- Supervised learning (2 weeks)
- Unsupervised learning (2 weeks)