Data Science for Economists

ECON 520

Instructor Info

Pedro H. C. Sant'Anna

Office Hrs: Wed, 10 am–11 am, by appointment.

Rich 331, or via Zoom at https://emory.zoom.us/my/pedrosantanna

http://psantanna.com

pedro.santanna@emory.edu

Course Info

Prereq: Econ 526 and Econ 725.

Tues & Thurs

11:30 am–12.45 pm

North Decatur Building 111

TA Info

Katie Leinenbach

Office Hrs: Tues, 5 pm–6 pm

Randall Rollins, R400A-9

Course Description

Data science is a rapidly developing field that combines statistics, econometrics, AI, and machine learning with data management, data wrangling, and data visualization to extract signals from data and inform decision-makers. Nearly every major company uses data science to optimize its services: Netflix uses it to recommend new shows to its viewers, Amazon uses it to optimize inventory and pricing, Microsoft uses it to design rewards within Xbox, to decide how much computing components to buy (and when), and to develop new products such as Co-pilot. This class will give students an overview of the data science workflow, from collecting data to drawing insights from which a decision-maker can make informed decisions. Along the way, we will broadly cover a variety of advances in data collection, data storage, visualization, machine learning, and econometrics topics, as well as reinforcing good programming practices. The primary goal of this course is to provide a set of skills that will allow you to navigate the data-science world and be ready for jobs that require those skills.

Learning Objectives

- Becoming familiar with data-science workflow
- Improve your data wrangling and data visualization skills
- Better understand supervised and unsupervised machine learning procedures
- Have a working knowledge of linear and binary regression models and their use in causal modeling

Course Materials

We will borrow material from a few textbooks and online materials, as no "ideal" textbook is available. As a result, we will not follow a textbook in a "chapter-by-chapter" manner.

Slides, codes, and datasets will be posted on Canvas.

All lectures will be recorded to the Zoom Cloud and links will be posted on Canvas. The instructor holds copyrights of his instructional recordings, and they should not be downloaded for personal or commercial use and/or posted or shared anywhere else on the internet.

Main textbooks

- Békés, Gábor & Gábor Kézdi, *Data Analysis for Business, Economics, and Politics*. Cambridge University Press. 2021. ("DA")
- James, Gareth, & Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor, *An Introduction to Statistical Learning, with Applications in Python*. Springer, 2023. Freely available at https://www.statlearning.com/. ("ISL")

Other textbooks

- Taddy, Matt, *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*. 1st Edition, McGraw-Hill, 2019.
- Taddy, Matt & Leslie Hendrix & Matthew Harding, *Modern Business Analytics: Practical Data Science for Decision Making*, 1st Edition, McGraw-Hill, 2023. ("MBA")

Other Online Materials

- Grant McDermott notes on "Data Science for Economists", freely available at https://github.com/uo-ec607/lectures.
- Tyler Ransom notes on "Data Science for Economists", freely available at https://github.com/pedrohcgs/DScourseS24_Ransom.
- Michael Knaus notes on "Causal Machine Learning" freely available at https://github.com/pedrohcgs/causalML-teaching.

Overall Grading

The final grade will be determined by a weighted average of scores from 7 problem sets (50%), 7 quizzes (25%), and a course project (25%).

The map from numerical to letter grade is the following:

Grade	Lower Limit	Upper Limit
A	95	100
A-	90	94.999
B+	85	89.999
B	80	84.999
B-	70	70.999
C+	65	69.999
C	60	56.999
C-	55	59.999
D+	50	54.999
D	40	49.999
F	0	39.999

Problem Sets

- You must write and submit your own problem set and computer code, although I encourage you to collaborate with your classmates. I DO NOT want to see a bunch of copies of identical code. I DO want to see each of you learning how to code these problems so that you can do it on your own.
- Written solutions must be submitted as PDF documents or Jupyter Notebooks.
- Problem Sets will involve both theoretical and practical questions. They will also involve coding, which can be done in `R` or `Python`.
- Problem sets will be due as announced on the day it is posted in Canvas. Late problem sets will not receive any credit. Partially completed problem sets will receive partial credit.
- Only the six (out of 7) highest graded problem sets will be used to calculate the average that will be weighted by 50%. In other words, the problem set with the lowest grade will have zero weights.

Quizzes

- Quizzes will be admitted on the same day as the problem sets are due and will cover the material of the problem set.
- They will be short, lasting around 20 minutes, unless otherwise stated. They will involve short questions related to theory or codes. They will take place at the beginning of the class.
- They are closed books, and computer usage is not allowed. You may bring **one** page to consult during the quiz.
- Only the six (out of 7) highest graded quizzes will be used to calculate the average that will be weighted by 25%.In other words, the quiz with the lowest grade will have zero weights.
- Failure to be physically present in the quiz will mean the corresponding quiz grade will be zero (0), unless the student has been excused by the instructor. Approval must be obtained at most one day after the problem set related to the quiz has been posted.

Project

Since this course is hands-on, we want you to have a final project summarizing all the skills you have learned here. Here are some details about the project:

- You should form a team of two students so the project is collaborative.
- You should collect data on and analyze a research question of your choosing using methods taught in this course.
- Write up a 10-page (12pt font, double spaced, excluding References, Figures, and Tables) summary of your findings, including a discussion about what prior studies of the same topic have found and citations to prior studies.

- Turn in the written summary report and a GitHub repository containing all materials required to reproduce the results.
- Summary report should be written in LaTeX (or Markdown) and turned in as a PDF (source code for the summary report should also be included in your GitHub repository).
- An example of what the final product should look like is here, with LaTeX source code here and BibTeX source code here.
- The project will be graded according to the following guidelines:

Category / Description	Points Earned	Points Possible
Code to reproduce results		
README file with directions on how to replicate the results		20
Automated compilation of figures / tables / other output		10
Exhibits good programming practices		10
Sanely organized		10
Subtotal		*50*
Written Report		
Exposition clarity (excluding References, Tables, and Figures)		15
At least 1 **properly formatted** equation		5
At least three references in the bibliography (with BibTex file)		5
Properly include in-text citations (using `\citet{bibkey}`)		5
At least 1 table and 1 graphical visualization		10
Compiled in LaTeX or Rmarkdown		5
Source code of report included with other code		5
Subtotal		*50*
Grand total		**100**

AI Policy

You are allowed and encouraged to use Artificial Intelligence (AI) resources in all aspects of this course. This includes, but is not limited to, using AI to help you write code, to help you debug code, to help you find answers to questions, and to help you find data. You are also allowed and encouraged to use AI to help you with your problem sets and final project.

Having said that, all coding, projects, and/or analytical answers incorporating any form of AI assistance must include a clear disclaimer stating the specific tool used and the extent of its contribution. Improper citations violate the Laney Graduate School Honor Code.

Attendance

To encourage student presence in the classroom, attendance will be taken each class (except when a quiz is scheduled), and those present on that day will earn 0.07 bonus points that will be added to the final average before a grade is assigned. The 0.07 bonus point is per class attended. Attendance records will be held up to date on Canvas.

Attendance may be recorded using PollEverywhere at a randomly chosen point in class. If the total number of respondents is larger than the number of people physically present in the classroom at that point, no bonus points will be awarded that day.

Make-up Policies

Emory College of Arts and Sciences does not have an attendance policy and, therefore, does not provide absence excuses. In the event of a catastrophic (and documented) occurrence that necessitates an absence from a problem set or quiz (only), the student should immediately seek help from the Office of Undergraduate Education (OUE), `http://college.emory.edu/oue/documents/student-absence-faculty-guidelines.pdf`.

The Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act of 1996 (HIPPA) regulations (U.S. Department of Health and Human Services and U.S. Department of Education) dictate that students do not have to provide medical documentation or disclose personal/medical issues with professors. However, the OUE class deans and academic advisers may collect this documentation and could provide verification to professors upon students' requests (see 'Situations where OUE will contact/work with instructors on a student's behalf' in the link above). This must be done within 48 hours after missing the assignment or quiz. If approval is granted by the instructor, the weight of the student's scores for the missed quiz/problem set assessment will be transferred to the next chronological quiz/problem set. If a letter (or e-mail) from the OUE is not received by the instructor within 48 hours, or approval is not explicitly obtained from the instructor, after a missed assessment, the missed assessment will receive a score of zero (0) points.

Student self-service absence form cannot be used as the basis to transfer the student's scores for the missed assessment to the next chronologically assessment. If the student is facing the following circumstances:

- Short-Term Illness (which represents 3 days or less of missed class),
- Longer-Term Illness (representing four calendar days or more of missed class),
- Quarantine,
- Short-Term Personal Emergency,
- Religious Holiday Observance,
- Student Athlete Absence

Please contact the instructor immediately before the intended missed end-of-chapter assessment.

Academic Integrity

Students are expected to adhere to the Emory College Honor Code as well as its Conduct Code, see https://conduct.emory.edu/. Specifically, the honor code is in effect throughout the semester. By taking this course, you affirm that it is a violation of the code to cheat on assignments, quizzes, and course projects, to plagiarize, to deviate from the teacher's instructions about collaboration on work that is submitted for grades, to give false information to a faculty member, and to undertake any other form of academic misconduct. You also affirm that if you witness others violating the code you have a duty to report them to the honor council.

Students caught cheating during any written examination will be asked to leave the classroom immediately and will earn an automatic grade of zero (0) points for said examination. The instructor will also report the incident for further disciplinary action. Please refer to the Laney Graduate School Honor Code. or Emory Undergraduate Code of Conduct for further details.

Special Circumstances

Students requiring any type of special classroom/testing accommodation for a disability, religious belief, scheduling conflict, or other impairment that might affect his or her successful completion of this course must personally present the requested remedy or other adjustment in written form (signed and dated) to the instructor, i.e., supporting memorandum of accommodation from the Department of Accessibility Services, https://accessibility.emory.edu/index.html. Requests for accommodations must be received and authorized by the instructor in written form no less than two weeks in advance of need. No accommodation should be assumed unless so authorized. In the event of needs identified later in the course, or for which an adjustment cannot be made on a timely basis, a grade of "I," Incomplete for the course, will be given to accommodate the unanticipated request.

Online Etiquette

All student questions posted on the course Canvas Discussion page (or any other forum related to this class that has been approved by the instructor) will be answered within 48 hours. Students are reminded that, as with any form of written communication, it is important to be mindful of one language and tone. The instructor reserves the right to erase any question or comment the instructor perceives as offensive, unnecessary, or can be interpreted as sarcastic, mocking, rude, or condescending by anyone in the class.

Diversity and Inclusivity Statement

I consider this classroom to be a place where you will be treated with respect, and I welcome individuals of all ages, backgrounds, beliefs, ethnicities, genders, gender identities, gender expressions, national origins, religious affiliations, sexual orientations, ability - and other visible and non-visible differences. All members of this class are expected to contribute to a respectful, welcoming, and inclusive environment for every other class member.

Class Schedule

Our schedule is ambitious and may be adjusted over the course. In addition, as I (the instructor) expect twins in the first three weeks of the course, I may miss some classes. If that happens, recordings will be posted.

Part 1: Introduction to Data-Science

Week 1: Jan 15-19	What is Data-Science and a Course Overview
Week 2: Jan 22 - 26	Data, Data Types, and Data Science Tools: Part I - Git, GitHub, and coding best practices - Where data come from
Week 3: Jan 29 - Feb 2	Data, Data Types, and Data Science Tools: Part II - APIs and Web scraping - Survey and Experiments - SQL and Big Data Tools

Part 2: Exploratory Data Analysis

Week 4: Feb 5 - 9	Wrangling Data & Exploratory Data Analysis: Part I
Week 5: Feb 12 - 16	Wrangling Data & Exploratory Data Analysis: Part II
Week 6: Feb 19 - 23	Modeling continuous and discrete variables

Part 3: Statistical Supervised Learning

Week 7: Feb 26 - Mar 1	Introduction to Inference with a focus on A/B tests
Week 8: Mar 4 - 8	Introduction to Optimization
Week 9: Mar 11 -15	SPRING BREAK
Week 10: Mar 18 - 22	Writing and optimizing functions • Debugging strategies and simulations
Week 11: Mar 25 - 29	Linear Regression Models
Week 12: Apr 1 - 5	Logistic Regression Models

Part 4: Introduction to Machine Learning

Week 13: Apr 8 - 12	Supervised Machine Learning
Week 14: Apr 15 - 19	Unsupervised Machine Learning
Week 15: Apr 22 - 26	Causal Inference with Machine Learning