ACE 592 SAE:
Data Science for Applied Economics

Spring 2021

Class Meets: TR, 4:00 - 5:20
Instructor: Professor Hutchins
Contact: jhtchns2@illinois.edu

Office: 431 Mumford
Office Hours: MW, 2-3pm
(or by appointment)

TA: Rocío Valdebenito
Contact: riv2@illinois.edu

Office:
Office Hours: TR, 10am-11am

Course Description

The ability to obtain, process, and analyze data using coding and algorithms has become an absolutely essential skill for anybody engaged in economic analysis. In the digital age, there is more data available than ever before on human behavior: from analyzing an elected official's opinion from Twitter to identifying a farmer's crop choices through satellite images. For those engaged in academic research in economics, these new data sources have drastically expanded the number of research questions that can be answered and revolutionized how we answer classic questions in the field. For those working outside of academia, the ability to process and analyze large datasets to provide insight has become an essential skill for producing useful analysis on a problem or question.

The goal of this course is to teach students in applied economics how to use data science tools and workflow for answering questions in economics. Using the python programming language and git version control, we will cover obtaining data via scraping and APIs, processing and cleaning data using python, and analyzing data via data visualization and basic machine learning techniques. The course will broadly cover the basics of text, spatial, and numeric data with an emphasis on their uses in analyzing economic questions and conducting research.

Data science can be broadly defined as "an approach to data analysis with a foundation in code and algorithms." The main goals of data science include drawing useful conclusions from large and diverse datasets through exploration, prediction, and inference. Data science is fundamentally different than most previous approaches to data analysis because:

- It uses **diverse data sources**, such as text, image, spatial, or numeric.

- There is an emphasis on **work flow** that uses coding and version control to maximize reproducibility and transparency.
- **New tools and approaches** for answering questions about data.

Course Objectives

In light of this, by the end of the course the student will be able to:

1. Obtain and process text, image, and numeric data using Python.
2. Analyze data using basic data visualization and machine learning in Python.
3. Construct a git repository and collaborate on a research project on Github.
4. Document code and communicate results using Jupyter notebooks.

To achieve these objectives, we will learn three main tools:

- **Python**, an object-oriented and general purpose programming language that can be used for reading, processing, and analyzing data.
- **Git**, a version control software for documenting and collaboration, as well as its online platform Github.
- **Jupyter notebooks**, a development environment for interactive programming that supports markdown.

Course Delivery

Classes will take place synchronously on Zoom on Tuesdays and Thursdays, 4-5pm.

The TA session will also take place synchronously on Zoom on Friday at 10am.

Prerequisites

There is no formal prerequisite for this course, but to be the most successful in this course the student *should be at least a masters student, should have a baseline knowledge of econometrics and statistics, and is exposed to programming (e.g. STATA, SAS, Matlab)*, though **no prior programming knowledge of Python is assumed for this course.**

Since much of this course is learning how to write code, students should note that **to learn to code, the best way is to do it, a lot.** Part of learning how to code will also be learning what resources exist to help you overcome problems, so while the lectures and homeworks exist as guides **how much you learn is a function of how much effort you put into practicing and supplementing your existing knowledge.**

Grading and Assignments

Grading will be based on two components:

- **Three assignments**, available starting at the beginning of the course and due throughout the semester (20 points each).
  - Submissions must be in the form of **Jupyter Notebook outputted as an HTML file**.
  - Group work is encouraged, but write ups **must be individual.**
- One **final analysis project and presentation**, done in groups (30 points final project + 10 points presentation).
  - The project must answer an applied economics question using data.
  - Groups must be formed by **Feb 18**, and the topic has to be approved by me by **March 18**.
  - Submission will be 1) a presentation done at the end of the semester and 2) a Github repository for your group's project.

Grading Rubric

Assignment	Points	Due Date
Homework 1	20	March 11
Homework 2	20	April 1
Homework 3	20	April 22
Final Presentation	10	May 11- May 13
Final Project	30	May 13
Total	100	

Resources

There is no textbook for this course, as most of the learning in this course is self-guided to some extent. As you come across problems in solving assignments and the like, these resources can help you overcome issues.

Big Picture Resources

- DataCamp, which all of you are signed up for.
- Coding for Data course

- Python for Data Science
- Inferential Thinking, a Berkeley course on data science:
- Fundamentals of Data Visualization
- Python Graph Gallery

Specific Resources

- Stack Overflow for very specific, coding problems.
- DataCamp Cheat Sheets, for git, Jupyter, and various python operations.

Course Schedule

Subject to change given needs of the semester in these trying times.

Week	Module	Main Topics
1	Introduction	- Python, git, and Jupyter basics
2		- Pandas, numpy, matplotlib
3		- requests, API basics
4	Text as Data	- Python processing text
5		- HTML parsing and scraping, basic NLP
6		- **Homework 1**
7	Images as Data	- Image editing, spatial data types
8		- Spatial statistics, basic mapping
9		- **Homework 2**
10	Numbers as Data	- Advanced pandas and numpy operations
11		- Distributed programming, scaling and parallelizing;
12		- **Homework 3**
13	Data Analysis	- Advanced visualization
14		- Unsupervised learning, feature generation
15		- Supervised learning, cross-validation
16	Presentations	