Data Skills for Computational Social Science

Chase Coleman, Spencer Lyon, and Thomas Sargent

Fall 2021

E-mail: cc7768@gmail.com
E-mail: spencerlyon2@gmail.com
E-mail: ts43@nyu.edu
Office Hours: By appointment

Web: **http://www.chasegcoleman.com**
Web: **http://www.spencerlyon.com**
Web: **http://www.tomsargent.com/**
Class Hours: Thursday 18:00-20:50

Course Description

This course teaches the foundational skills necessary to do modern data analytics using the Python programming language. We assume that students have previously worked with Python. We will add to existing Python skills and teach the core scientific and data-specific libraries (numpy, scipy, matplotlib, and pandas). We will use these skills to analyze a variety of social science datasets and answer research and business questions.

Course Materials

- **QuantEcon Datascience** Lectures from the QuantEcon datascience sequence at **https://datascience.quantecon.org**
- **Python Data Science Handbook** by Jake Vanderplas **https://jakevdp.github.io/PythonDataScienceHandbook/**
- **Python for Data Analysis, 2nd Edition** by Wes McKinney

Prerequisites

Students should have prior experience using the Python programming language before starting this course. Ideally, students will have completed the Summer "Pre-Course" that teaches these skills. Throughout this course we will leverage programming skills such as control flow constructs (`if/else`, `for/while`), defining custom functions (`def`), and finding help on existing functions (`?` in Jupyter environments and `help` elsewhere).

Althoughy a course in probability or statistics is not a prerequisite, students will find some knowledge of these topics to be helpful.

Students are required to have access to a personal computer (laptop) that can be brought with them to lectures. Computers will be used by students in each class session and those without a laptop will not get the necessary in-class practice in order to master the concepts we study.

Course Objectives

The key objective in this course is for students to use Python to do meaningful data analysis using social science data sets. Success in this course can be described as the student's ability to do the following:

- Read, write, and understand basic programs written in the Python programming language
- Import, clean, combine, and summarize datasets from a variety of sources
- Construct informative visualizations of raw data and model results
- Implement basic data engineering best practices such as optimizing organization and structure of datasets, using effective storage formats for a given task, and automating repetitive extract-transform-load (ETL) processes

Course Structure

Class Structure

This course will meet once a week for 3 hours.

Class will be treated as a mixture of lecture time and lab time. Students should bring, and expect to use, their laptops every time the class meets.

Assessments

This course will use a mixture of homework assignments, in-class quizzes, exams, and a final project to evaluate students.

Homework: At the beginning of the course, homework will be assigned almost every week. Later in the course, there will be less frequent assignments in order to assure that you have time to work on your class project. Your two lowest homework grades will be dropped.
In-class participation: As the class will be offered virtually, being an active participant in lectures and discussion requires effort. We recognize and appreciate this effort and will reward it accordingly.
Exams: There will be 1 take-home exam.
Project: There will be a class project aimed at helping you apply the tools that you have learned to a "real-world problem."

Other than for the exam, we highly encourage students to work together. We have found that groups of 3-4 seem to work best. We believe that collaborative work is the best way to learn the type of material that we cover. We advise students not to rely on others to do work that you do not understand.

Grading Policy

The assignments just described will be the main inputs to the grade for the course. Assignments will be weighted evenly within groups and overall according to the following decision rule:

- Homework assignments: 25%
- In-class participation: 15%
- Tests: 20%
- Project: 40%

This weighting reflects our opinion that the most important skills to be acquired in this class are communicated by one's ability successfully to apply the tools that you learn to an interesting question in the social sciences.

Schedule and weekly learning goals

The schedule is tentative and subject to change. Several of the modules below will occupy more than one week. The learning goals target key concepts to be mastered after each module. Successive modules build on early modules.

Part 1: Introduction to Pandas

Sources and tools:

- Class notes
- `pandas` package: https://pandas.pydata.org/
- https://python.quantecon.org/pandas.html
- https://datascience.quantecon.org/pandas/intro.html
- `pandas.DataFrame`
- `pandas.Series`
- Chapter 5 of Python for Data Analysis

Topics to be mastered:

- Pandas datatypes: DataFrame and Series
- Basic operations with DataFrames: summary statistics, aggregations, transformations, data selection
- Sorting and ranking
- Value counts
- Function application and mapping
- Duplicate labels
- Basic visualization using the `plot` method

Part 2: Organizing Data With Pandas, I

Sources and tools:

- Class notes
- `pandas.Index`
- https://datascience.quantecon.org/pandas/the_index.html
- https://datascience.quantecon.org/pandas/storage_formats.html
- Chapter 6 of Python for Data Analysis

Topics to be mastered:

- Understanding the `Index` in pandas
- Storage formats
- Reindexing
- Stacking and melting
- Hierarchical indexing

Part 3: Organizing Data With Pandas, II

Sources and tools:

- Class notes
- https://datascience.quantecon.org/pandas/data_clean.html
- https://datascience.quantecon.org/pandas/reshape.html
- Chapter 7 and 8 of Python for Data Analysis

Topics to be mastered:

- Cleaning, reshaping, and merging datasets
- `merge`, `join` and `combine`
- Stacking and melting
- Handling missing data
- Discretization and binning
- Random sampling
- String manipulation

Part 4: Grouped Operations with Pandas, I

Sources and tools:

- Class notes
- `pandas.groupby`
- https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html#
- https://datascience.quantecon.org/pandas/groupby.html
- Chapter 10 of Python for Data Analysis

Topics to be mastered:

- `groupby` method with built-in methods
- Groupby mechanics
- Custom grouped functions

Part 5: Grouped Operations with Pandas, II

Sources and tools:

- Class notes
- `pandas.groupby`
- https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html#
- https://datascience.quantecon.org/pandas/groupby.html
- Chapter 10 of Python for Data Analysis

Topics to be mastered:

- Aggregation with multiple function application
- General split-apply-combine
- `transform`
- `apply`

Part 6: Grouped Operations with Pandas, III

Sources and tools:

- Class notes
- https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html#
- https://datascience.quantecon.org/pandas/groupby.html
- Chapter 12 of Python for Data Analysis

Topics to be mastered:

- Pivot tables and cross-tabulation
- Method chaining with `pipe`
- Categorical data

Part 7: Time Series with Pandas, I

Sources and tools:

- Class notes
- https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html
- https://datascience.quantecon.org/pandas/timeseries.html
- Chapter 11 of Python for Data Analysis

Topics to be mastered:

- Rolling-window operations
- Resampling frequency of observations
- Doing arithmetic with dates, date ranges, periods, and TimeDeltas

Part 8: Time Series with Pandas, II

Sources and tools:

- Class notes
- https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html
- https://datascience.quantecon.org/pandas/timeseries.html
- Chapter 11 of Python for Data Analysis

Topics to be mastered:

- Upsampling and interpolation
- Downsampling
- Handling time zones

Part 9: Data Visualization, I

Sources and tools:

- Class notes
- https://datascience.quantecon.org/pandas/matplotlib.html
- https://seaborn.pydata.org/
- https://plot.ly/python/
- https://altair-viz.github.io/
- Chapter 9 of Python for Data Analysis
- https://datascience.quantecon.org/applications/maps.html

Topics to be mastered:

- Intermediate `matplotlib`
- Statistical visualization with `seaborn`
- Widgets

Part 10: Data Visualization, II

Sources and tools:

- Class notes
- Chapter 9 of Python for Data Analysis
- https://datascience.quantecon.org/applications/maps.html

Topics to be mastered:

- Interactive web-based visualizations, and dashboards using `plotly` and `altair` — As an example of what could be done, see Mike Waugh's webpage

Part 11: Data Harvesting

Sources and tools:

- Class notes
- https://scrapy.org/
- https://camelot-py.readthedocs.io/en/master/
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Topics to be mastered:

- Integrating with Web APIs
- Scraping data from websites without an api (`scrapy`)
- Extracting data from PDFs (`camelot`)

Part 12: Data Engineering

Sources and tools:

- Class notes
- https://airflow.apache.org/
- https://www.sqlalchemy.org/

Topics to be mastered:

- Basic introduction to databases (using SQLite through `sqlalchemy`)
- Automation and data pipelines using Apache Airflow
- We will illustrate these tools by creating an automatically updating database on one of a few potential topics. Our choice of topic will depend on class interest.

Part 13: Case studies, I

Sources and tools:

- http://www.tomsargent.com/research/ReadMe_Pub.pdf
- https://datascience.quantecon.org/applications/
- Chapter 14 of Python for Data Analysis
- https://datascience.quantecon.org/applications/recidivism.html

Topics to be mastered:

- Combine the tools learned in this class to generate automatically updated databases and visualizations, covering topics such as
  - Inequality data
  - U.S. bond data and term structure of interest rates; see Hall, Payne, Sargent bond dataset

Part 14: Case studies, II

Sources and tools:

- https://datascience.quantecon.org/applications/
- Chapter 14 of Python for Data Analysis

Topics to be mastered:

- Examples from *The Great Reversal* by Thomas Phillipon

Part 15: Case studies, III

Topics to be mastered:

- Student presentations of class projects

Course Policies

Professional Behavior

Attend class. They say "eighty percent of success is just showing up." We have found that those who show up perform systematically better.

Arrive to class on time and stay until the end of class. Chronically arriving late or leaving class early is unprofessional and disruptive to the rest of the class.

We understand that the electronic recording of notes will be important for class and so computers will be allowed in class. Please refrain from using computers for anything but activities related to the class. Phones are prohibited as they are rarely useful for anything in the course. Eating and drinking are allowed in class but please refrain from it affecting the course. Try not to eat your lunch in class as the classes are typically active.