Working with Big Data - Syllabus

Instructor: Daniel Jaramillo (jaramillocalderondc@gmail.com)
Office Hours: By appointment
Class Hours: 9:00 AM - 12:00 PM
Class Room: ALRD 121

Learning Materials

Please bring the following items to every class

- Laptop
- Laptop charger
- Paper notebook (or tablet)
- Writing utensil (or stylus for your tablet)

Course Overview

With the rapid proliferation of larger and more complicated datasets, wrangling, analyzing, characterizing, interpreting, and generally deriving value from data are becoming increasingly important skills. In this course, students will learn the basics of computational programming and data analysis by applying that knowledge to computer-based coding exercises, real data sets, and their own project. When relevant, we will examine the material we are learning in a big data context.

This course roughly follows (and at points, borrows examples and code from) the material laid out in QuantEcon's lecture notes "Introduction to Economic Modeling and Data Science" with some deviations in topics, course materials, and mathetmatical sophistication. In particular, we do not cover the economic-specfic material nor do we assume that students have the same mathematical pre-requisites. The necessary math will be taught as part of this course. While this course is self-contained and requires no economic knowledge, QuantEcon is a valuable alternative reference for the topics covered in this class and more advanced topics.

Learning Objectives

Students will begin by learning the fundamentals of coding in Python (a popular open-source programming language) using Jupyter notebooks. Simultaneously, they will learn to use GitHub for code collaboration and basic version control. Afterwards, we provide a hands-on introduction to various topics including

- Data visualization and mapping
- Basic mathematics for data analysis
- Data manipulation using the Pandas Python library
- Scientific computing using the NumPy Python library

With a working knowledge of these tools, students will be ready to learn about and implement various regression and classification techniques such as

- Linear Regression
- Lasso Regression
- Regression Trees

- Logistic Regression
- Classification Trees

Course Structure

This course consists of 13 3-hour classes (3 credit course equivalent). Roughly the first half of each class will be dedicated to lecture. After lecture, there will be a short break. The remainder of the class period will be spent working on a lab assignment or the final project. Students will generally finish the lab assignments in class. If lab assignments are not finished in class, students must submit them **within 48 hours of the end of class**.

Students will also work on a group project both in and out of class. On the final exam day, students will present their project instead of taking a final exam. More details on that final project and presentation will be shared in class.

Course materials will be provided on GitHub. Canvas will only be used for assignment submission and keeping track of grades.

Grading

- 50% In-class lab assignments
- 35% Project
- 10% Project Presentation
- 5% Project Proposal

Detailed Timeline

Note that the following is an ambitious timeline for the class and is subject to change.

Day 1 (July 16)

Lecture: Big Data, course description, introduction to GitHub, Jupyter, & Python
Lab: Sign up for GitHub, access Jupyter Open, create & clone repositories

Day 2 (July 17)

Lecture: Python fundamentals
Lab Assignment 1: Python fundamentals

Day 3 (July 18)

Lecture: Collections
Lab Assignment 2: Collections & Control flow

Day 4 (July 19)

Lecture: Control Flow
Lab: Introduction to project, getting into groups, start project

Day 5 (July 22)

Lecture: Functions
Lab Assignment 3: Functions

Day 6 (July 23)

Lecture: Introduction to data & Pandas
Lab Assignment 4: Data & Pandas

Day 7 (July 24)

Lecture: More data & Pandas
Lab: Work on group project

Day 8 (July 25)

Lecture: Introduction to arrays, matrix algebra, & NumPy
Lab Assignment 5: Arrays, matrix algebra, & NumPy

Day 9 (July 26)

Lecture: Basic plotting & mapping
Lab: Working on project in groups. You should be collecting data, figuring out topic(s)/story, thinking of visualizations

Day 10 (July 29)

Lecture: Introduction to regression
Lab Assignment 6: Regression

Day 11 (July 30)

Lecture: More regression
Lab: Working on Projects in groups. Should be refining descriptive statistics and figures and begin writing up some text in Jupyter Notebook.

Day 12 (July 31)

Lecture: Introduction to Classification
Lab: Working on project in groups. You should be integrating figures, descriptive statistics, and writing into Jupyter Notebook. You should also start thinking of which classification and/or regression models you will run.

Day 13 (August 1)

Lecture: Classification Continued
Lab: Working on project in groups. You should be running your analyses and thinking about how to frame them.

Day 14 (August 2)

When you get to class, you will received an aditional hour of time to work on your project or presentation. At the end of that time, your project must be submitted to me on Canvas. Then, each group will present their project for the remaining two hours.