Data Analytics in Python

Prof Hardeep Johar
hj2203@columbia.edu

The Course
The goal of this course is to give you a working knowledge of how to use Python to extract knowledge and information from data. At the end of this course you will be competent in using Python libraries to work with and analyze offline as well as online data. We will work with python packages for statistical analysis, machine learning, natural language processing (text mining) and the analysis of networks. We will also examine techniques for data visualization and will (briefly) review how to present results using lightweight apps like flask.

This is an intensive hands on course so be prepared for a lot of work and a significant time commitment. But your reward - proficiency in data analytics skills - will be substantial!

Prerequisites
B8136 (Introduction to Programming) or B8126 (Web App Programming). Anyone who has taken college level programming classes or has prior experience with programming can apply for a waiver.

Do I need to be a math and stats whiz to take this class?

No. The best part of Python is that you can leave the math and stats stuff to programs. All you need to bring to the class is your creativity and a basic understanding of programming. Analyzing data is all about asking the right questions and then presenting your results in the best way possible to get your message across. As we work with the various libraries, we’ll get a sense of when to use what sort of analysis. But we will leave all the grungy stuff to Python.
Topics:
1. Review of web scraping and APIs
2. Data analysis and visualization
3. Mining text data
4. Analysis of networks
5. Machine learning
6. Presentation (flask)

Evaluation and learning components

**Home assignments:** We’ll have a few home assignments as well. Like the quizzes, assignments are not meant to be diagnostic but rather to help you practice and learn so they will be very lightly graded. You can consult with others, ask me questions, use google or duckduckgo for help, but do try them on your own first. Because it is important that you do the assignments, I will accept late assignments. But, any assignment submitted after the due date risks a 30% reduction in score.

**Project:** There is no better way to learn something than to go out and use it so start thinking about a data set you’d like to analyze. Final submission will include a design report, Python code, and an in-class “speed-date” presentation and demonstration. A significant part of your project grade will come from how other students rate your work.

**Participation:** Demonstrate engagement in the course by asking questions. I’ll respond to every question, either online or, if the response is of general interest, in the classroom.