Data Analytics in Python

Hardeep Johar (hj2203@columbia.edu)

Tuesdays 2:00 - 5:15. Kravis 820

Course Description

The collection, interpretation, and analysis of data has always been a central pillar of business decision making. Historically, this has followed a two step process, statisticians gather data, organize it, run analytics and prepare reports. At some future point, a decision maker examines these reports, interprets the results and makes decisions. However, with the advent of powerful and inexpensive computing platforms, the collection and analysis of data has moved into the continuous decision making cycle itself, with decisions being constantly updated as new data is instantly analyzed and acted upon. Consequently, decision makers can no longer isolate themselves from the grungy side of data and they need to know where the data originated, how it was transformed, what is the nature, the strengths and the limitations of the analytical techniques used. Today, to be effective, decision makers need an intuitive understanding of the statistics, the math, and the programming that underlie this “live” analytical and decision making process.

The objective of this course is to give you an understanding of the analytical side of the decision making cycle, focusing on programming as the element that “glues” the collection, transformation, visualization, and analysis of data. We will see how to get data from common sources (APIs, web scraping), examine the rudiments of data visualization (charts, maps), and get an intuitive understanding of the types of analytical tools in use today (machine learning, deep learning, analysis of networks, analyzing natural language texts).

With its extensive collection of libraries, Python is fast becoming the platform of choice for data analytics so Python will be our language for this course. The course is very hands on, and you should expect a lot of programming work, all of it fairly intense. A basic understanding of how to write programs in Python is therefore a must for this class. But, the primary takeaway from the course is not the programming but rather an understanding of the mechanics, the vocabulary, and the techniques in data analytics. Even if you find programming a frustrating and head banging exercise, you can get a lot out of the class (if you’re willing to suffer a bit!).

Prerequisites

Prior exposure to some programming language is helpful and you should have taken B8154: Python for MBAs or cleared the waiver exam. I encourage you to explore online

Python programming courses before the start of the semester (bearing in mind that we'll be working with Python 3.9 and not 2.7). The better prepared you are with Python at the start of this course, the more you'll get out of it.

Evaluation components

1. Assignments

Expect a programming assignment every week. Assignments are not meant to be evaluative (though, unfortunately, they will be evaluated). Think of them as learning mechanisms and make a good faith attempt to do them on your own. You may take the help of other students, the TA, or see me during my office hours but you should complete them yourself. Assignments will be lightly graded so turning them in by the due date is definitely to your advantage. Late assignments will be penalized 25% if one week late and 50% after that.

2. 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 (brief) report, Python code, and an in-class presentation conducted in a speed date format. A significant part of your project grade will come from how other students rate your work so your focus should not necessarily be on sophisticated analysis but on presenting your work in a way that is easy to understand while highlighting the key takeaways..

3. Quizzes

There will be several short quizzes - during class. The purpose of the quizzes is to quickly check recall and, generally, if you've been paying attention in class and doing your homework in a timely fashion you shouldn't have to worry too much about them.

4. Exam

There will be a take home final exam

Schedule

Module 1: Python review

Module 2: Getting data (APIs, web scraping)

Module 3: Python toolbox (numpy and pandas)

Module 4: Data visualization (maps, charts, interactive charts)

Module 5: Introduction to Machine learning

Module 6: Text analytics

Module 7: Deep learning (very basic introduction!)

Module 8: Network analysis (time permitting)

Frequently asked questions

Q. What sort of computing background do I need to bring to the class?

A. You need some prior programming exposure. If you've taken B8154: Python for MBAs or if you can pass the advanced qualifying exam, you should be more than fine. Though this is an intensive programming class, the goal is not to create super programming geeks (though that will be good!) but rather to give you a sense for what is possible in the analytics world.

Q. The project. Could you give us some more information on what is expected?

A. Unfortunately I can't give you a whole lot of guidance in selecting a project (but will be happy to guide you once you've chosen a data set). Every project is different. Some have a heavy analytical component, while others focus more on finding interesting patterns in the data or building an interactive interface to data. The best suggestion I can give is that you find an area that interests you, look for a large data set in that area, and then analyze the heck out of it. If you absolutely must see a sample, then here is one

Q. Mac, or Windows or Linux?

A. Either is fine but, if you have the choice, then please use a Mac or a Linux machine because, sometimes, Windows just doesn't like to install tricky libraries. In particular, if you have a Mac and are using some sort of Windows emulator then please use Mac OS-X and not the Windows emulator. The double redirection will make everything a lot slower and you'll have to deal with installation quirks. But, if you are a Windows user, don't worry - we'll make it work.

Q. What software will we need?

A. We will use Anaconda Python (current version 3.9)

Q. What format will the class material be distributed in?

A. We will use Jupyter notebooks (Jupyter is automatically installed when you install Anaconda Python) and all class material will be made available in the form of notebooks before class. In class, you can follow along by running the code in your notebooks and work on the included small practice problems. The notebooks will run on google colab if you prefer not to install anything locally (though the text may render differently).

Texts

There is no text for this class. The following will be helpful if you want to go above and beyond the material covered in the course:

- *Learning Python, 5th Edition Powerful Object-Oriented Programming*, Mark Lutz. O'Reilly Media, 2013.
- *Web Scraping with Python: Collecting More Data from the Modern Web* by Ryan Mitchell. 2nd Edition. O'Reilly Media May 2018. ISBN: 978-1491985571
- *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter 3rd Edition.* O'Reilly Media September 2022. ISBN: 978-1098104030

Online resources

Python documentation: *http://docs.python.org/3.9/index.html*

Python tutorial: https://docs.python.org/3.9/tutorial/

Python Regular Expressions https://docs.python.org/3.9/library/re.html

BeautifulSoup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Pandas: https://pandas.pydata.org/pandas-docs/stable/

NLTK: https://www.nltk.org

Big Data http://www.mmds.org/

SQL: https://www.w3schools.com/sql/

MySQL: https://www.mysqltutorial.org