STA 5106 Computational Methods in Statistics I

Department of Statistics
Florida State University

Class 4
September 5, 2019

What is Python?

- Python is a widely used general-purpose, high-level programming language.

- As of Jan 2015, there are more than 54,000 Python packages offering a wide range of functionality, including:
  - graphical user interfaces, web frameworks, multimedia, databases, networking and communications
  - test frameworks, automation and web scraping, documentation tools, system administration
  - **scientific computing, text processing, image processing**

Which Version to Use?

- Two different versions:
  - Python 2.0 (released on 10/16/2000) is most commonly-used.
  - Python 3.0 (released on 12/03/2008) is a major, backwards-incompatible release.

- We will use 3.0 in this class.

Development Environment

- Most Python implementations (including CPython) can function as a command line interpreter. That is, Python acts as a shell.

- Other shells add capabilities beyond those in the basic interpreter such as **iPython.**

- Python IDE (integrated development environments):
  - Canopy
  - IDLE
  - PyCharm
  - **Spyder (in Anaconda)**
  - … (over 20)

Why use Python for Data Analysis?

- Is data analysis all about numerics and filtering, and maybe plotting?

- Of course NOT. In the real world: data is messy, and in many cases, the majority of the work in a data analysis project is retrieving the data, parsing it, and so on.

- General-purpose scripting languages, such as Python, have much better language and library support than any of the data-specific languages such as MATLAB and R.

Python for Scientific Computing

- The **NumPy** library provides a solid MATLAB-like matrix data structure, with efficient matrix and vector operations.
- The **Scipy** library includes a very large collection of numerical, statistical, and optimization algorithms.
- The **Pandas** provides R-style Data Frame objects (using NumPy arrays underneath to ensure fast computation), along with a wealth of tools for manipulating them.

Python for Scientific Computing

- Python has a huge number of well-known libraries for the messier parts of analysis. For example, **Beautiful Soup** is best-of-breed for quickly scraping and parsing real-world HTML.

- Together with **iPython**, these libraries make Python a useful and popular tool for data analysis.

Python in This Class

- **Anaconda** at **https://www.anaconda.com/download**.
- Free enterprise-ready Python distribution for large-scale data processing, predictive analytics, and scientific computing.
- Most popular Python platform with over 4.5M users.
- 1000+ packages including Numpy, SciPy, Pandas, iPython, …
- Cross platform on Linux, Windows, Mac.
- Easily switch between Python 2.7, 3.6.
- My favorite IDE: Spyder (very much like Matlab)

Important References

- **NumPy for Matlab Users:** Great intro tool for users who are familiar with Matlab. It is in a clear and concise form.

- **MATLAB commands in numerical Python (NumPy):** Comparison of commonly-used commands for Matlab, R and Python. It is a more complete version.

- **Google for help on everything.**