MSDS 7333 Quantifying the World

Brief Review of Python Pandas Library

R is Like Batman

- Detective work
- Intelligence
- Cunning
- Usage of tools
- More brain than muscles

Python is Like Superman

- Muscle Power

- Super strength

- Elegance

- Wide range

- More muscles than brain

R vs. Python

R for analysis, data visualization, and modeling	Python for data preparation, data munging
Offers great flexibility for analysis	Great flexibility and ability to extract information from free text, websites, and social media sites
R makes it easy to think while doing analysis	Good with mining images and preparing data for analysis
Constant upgrades and enhancements due to active R community	Can handle large volumes of data better than R
Exceptional data visualization tools	

Really, it's R **and** Python

- "Python and R together complete the toolkit for a data scientist"

- "For a serious data scientist, it is a good idea to have some functional knowledge of both R and Python"

[Free books to learn R and Python](#)

Opening Python

- Double click on Anaconda symbol on desktop

  - Programs > Anaconda

  - Applications > double click on Anaconda

- Launch iPython Notebook

- For a challenge – launch Spyder (RStudio for Python)

- We will be using Spyder in this module

Spyder Editor and Variable Examples

```python
# -*- coding: utf-8 -*-
"""
Spyder Editor

This is a temporary script file.
"""

# this is a comment
# place this at the top of a python file to enable running from the terminal as >> ./filename.py
# ! /usr/bin/python
# otherwise you can run with >> python filename.py {or} >> python27 filename.py

int_val = 8
long_val = 2342342325L
float_val = 2.0
bool_val = True

print "Variable Type Examples:"
print type(int_val)
print type(long_val)
print type(float_val)
print type(bool_val)

# testing for the type of a variable
print isinstance(float_val,float)
print isinstance(float_val, int)

```

Python Structures

- Strings

- Lists

- Tuples

- Dictionaries

- Comprehensions

- Loops

- Stacks and Queues

- Conditionals

- Functions

Python Data Analysis Library (Pandas)

- A set of labeled array data structures, the primary of which are Series and DataFrame

- Index objects enabling both simple axis indexing and multi-level / hierarchical axis indexing

- An integrated group by engine for aggregating and transforming data sets

- Date range generation (date_range) and custom date offsets enabling the implementation of customized frequencies

- Input/Output tools: loading tabular data from flat files (CSV, delimited, Excel 2003), and saving and loading pandas objects from the fast and efficient PyTables/HDF5 format

- Memory-efficient "sparse" versions of the standard data structures for storing data that is mostly missing or mostly constant (some fixed value)

- Moving window statistics (rolling mean, rolling standard deviation, etc.)

- Static and moving window linear and panel regression

- More in Chapter 5 of Python for Data Analysis (McKinney text)