Day 3: Getting started with pandas and NumPy I

Raymond Yee

January 29, 2013 (`http://bit.ly/wwod1303`)

Goals Today

- Making sure everyone well on his/her way for Day 2 HW
- Getting people set up with using git/github
- Making sure we can all communicate on Piazza
- Discussing the importance of questions and answers
- Orienting ourselves for working through early chapters of `PfDA`

Please make sure you sign in for today – please say hi to me when you sign sheet

Git/ Github

If you don't have git set up:

`https://help.github.com/articles/set-up-git`

The git repo for class:

`<https://github.com/rdhyee/working-open-data>`

To clone:

```
git clone git@github.com:rdhyee/working-open-data.git
```

Help documentation at github:

`https://help.github.com/`

pydata-book on github

```
https://github.com/pydata/pydata-book
```

```
git clone git://github.com/pydata/pydata-book.git
```

bSpace

I will use bSpace primarily for its

- gradebook
- maybe as a dropbox

Other uses for bSpace?

Piazza

`https://piazza.com/berkeley/spring2013/info290t/home`

Why use?

- hopefully as a good place for Q&A
- the polling looks good

cf stackoverflow questions tagged pandas

What have people's previous experiences with Piazza been?

Thanks to Gilbert for posting Reading List / Syllabus?.

Let's check in with current schedule:

`http://htmlpreview.github.com/?https://github.com/rdhyee/working-open-data/blob/master/lectures/day01.html#(8)`

Let's use Piazza in class

- Post an answer to What's a data set that intrigues you?
- Answer Poll: How difficult did you find the homework from

Newbie-style Questions from Class

Any advice for Python beginners?

Will the class be structured such that still improving their grasp of handling data can keep up?

Content-style Questions from Class

What amount of analysis will we be doing? Project work?

Will we learn about creating our own datasets from public info that isn't well organized?

To what extent will this course cover the math and statistics behind working with open data?

What kinds of analysis do we do on the data? Does this course dive into statistical models for data analysis?

Is it a good idea to learn the tools for Python in conjunction with a statistics foundation?

Meta-style Questions from Class

Would you adjust course material based on the feedback of the class?

Are we allowed to take lunch breaks?

How many hours will problem sets take?

Why are we using bspace rather than Piazza?

Will this class take a lot of my free time?

How will the project team be formed?

What kind of projects are you expecting? When should we start thinking about them?

Questions for the Instructor from Class

What was your first open data project?

What is the coolest use of open data you've seen?

What is the most complicated example of Python programming you've done?

What is your background and interest in using open data?

What work have you done with open data?

Can you describe some favorite professional experiences with open data?

Can you describe some favorite projects from previous offerings of this course?

Who is the next "Todd Park" in the open data movement?

Do you have any experience with institutions that did not want to open their data but were convinced otherwise and why?

Getting these questions answered

I will work in answers as we go, but if you want something answered, please post in Piazza.

Updating ipython, pandas, etc. using enpkg

Last week, I walked some of you through using `easy_install` to update iPython. A possibly better way is to use `enpkg`.

Useful instructions for using `enpkg` – I used it update numpy and pandas:

`https://www.enthought.com/products/update.php`

```
enpkg -s ipython

sudo enpkg ipython 0.13.1

enpkg matplotlib 1.2.0

sudo enpkg pandas 0.10.0
```

Conceptual Overview of book

`http://pandas.pydata.org/pandas-docs/stable/overview.html`:

pandas consists of the following things [highlights]:

- A set of labeled array data structures, the primary of which are Series/TimeSeries 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
- 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.

I want all of us to be able to type / copy&paste code from book and confirm to ourselves that we understand what's going on and to **note when we don't understand something**.

In-class walk-through of what I mean by working through book. In

Big concepts

- pandas is the main library – built on numpy (array)
- Series, DataFrame
- matplotlib to do some plotting

Where we are going for Day 4

Next time, I will assign next problem set.

- set of exercise to deepen your understanding of Chap 2, 4, 5 (first pass)
- use that understanding to analyze census data
- take a walk through more of census data, data.gov, local gov't data

RY's to do list

- grade homework 1
- scheduling guest speakers
- maybe I will start making little videos. . .

In Class Activity

Break up into groups that can work together. I suspect we'll have 2 large groups:

- those who want to work on Day 2 HW
- those who want to work through Chap 2 of `PfDA`

What do people want help with?

Assignments / Homework

Reminder: Day 2 homework

Due tomorrow (Wed, noon) – send to `yee@berkeley.edu`

Piazza participation

- Post an answer to What's a data set that intrigues you?
- Answer Poll: How difficult did you find the homework from Day 2?
- Answer Poll: What have you done with stackoverflow?
- Optional, but highly recommended: post any questions you want help with or are interested to discuss with your classmates or instructors I hope you also feel free to answer questions too.

Readings for Day 4

- `PfDA`, Chapter 2