

Data analysis with Python and pandas

James Polera (@uncryptic)
james@uncryptic.com

About me

- ✳ My name is James Polera, I'm the IT manager for a mid-size law firm in Union county, and part of the sister technology company (they do .NET primarily). I also own a consulting company where I do mostly Python/Django apps (in my spare time, of which I have none).
- ✳ I've been working in IT professionally since 2000 in both Sysadmin and Developer roles (often at the SAME TIME).
- ✳ I'm an autodidactic polyglot (but Python is my “go to” language)

Things we will do

- ✳ Crash course in pandas
- ✳ Introduction to the included data structures
- ✳ Some basic data processing
- ✳ Extol the virtues of virtualenv (but not in a pushy way)
- ✳ Use the fabulous IPython HTML Notebook

Things we will not do

- * Any advanced statistics (it's beyond the scope of this talk and the knowledge of your speaker).
- * Go over installing modules via pip.
- * Cover *all* of pandas. There's a lot to it.
- * Be apprehensive. If you think I'm wrong about anything, tell me. I'd rather know if there is a better way to do things.

What is pandas?

- ✳ A library for doing data analysis in Python
- ✳ Project lead: Wes McKinney
- ✳ <http://pandas.pydata.org>
- ✳ Built on top of NumPy (<http://www.numpy.org>)

Notable features

- ✳ Data alignment (think relational database tables)
- ✳ Data grouping
- ✳ Support for various file formats (csv, xls, HDF5, SQL databases)

Notable features

- ✳ Ability to add and remove columns on the fly
- ✳ Integration with matplotlib
- ✳ Intelligent merging and joining of datasets

A little bit about NumPy

- * pandas leverages some great work from the NumPy project and builds on it.
- * The ndarray data structure and NumPy's broadcasting abilities are heavily used in pandas.

What's ndarray?

- * ndarray is an N-dimensional (i.e. multi dimensional) array
- * Supports what the NumPy project calls “broadcasting”
- * Let's take a look

pandas data structures

- ✳ Series
 - ✳ A Series is a one dimensional labeled array. It can hold any Python datatype

pandas data structures

- ✳ DataFrame
 - ✳ A 2-dimensional labeled data structure with columns of potentially different types.
 - ✳ It's like a spreadsheet or a database table.
 - ✳ It's one of the coolest things about pandas.

pandas data structures

- * Panel
- * A Panel is a container for 3 dimensional data.
- * We're not going to cover it in this talk, but I mention it here as an exercise for you to follow up on.

Tonight's dataset (roughly) brought to you by

- * This guy:  (@chrisbaglieri)
- * Chris is a software engineer who was a PUG/IP regular back in 2011 before moving out of the area.
- * He wrote a Ruby gem called “quake”
<https://github.com/chrisbaglieri/quake>
- * Check it out!

Processing data: Pure Python

- ✳ From the Python standard library
- ✳ The csv module

Processing data: Pure Python

- ✳ From the Python standard library
 - ✳ `collections.namedtuple`
 - ✳ `namedtuple` makes it easy to give meaning to the items in a tuple, making them behave more like an object with getters.

Virtues

- ✳ Words to live by: “*...the three great virtues of a programmer: laziness, impatience, and hubris.*” - **Larry Wall (creator of the Perl programming language)**

Processing data: pandas

- ✳ This *is* why you came here, right?
- ✳ Let's take a look at that dataset again.

What do I need to install?

ipython==0.13.1
matplotlib==1.2.0
numpy==1.6.2
pandas==0.10.1
python-dateutil==2.1
pytz==2012j
pyzmq==2.2.0.1
tornado==2.4.1
wsgiref==0.1.2

Resources

<http://pandas.pydata.org>

<http://earthquake.usgs.gov/earthquakes/feed/>

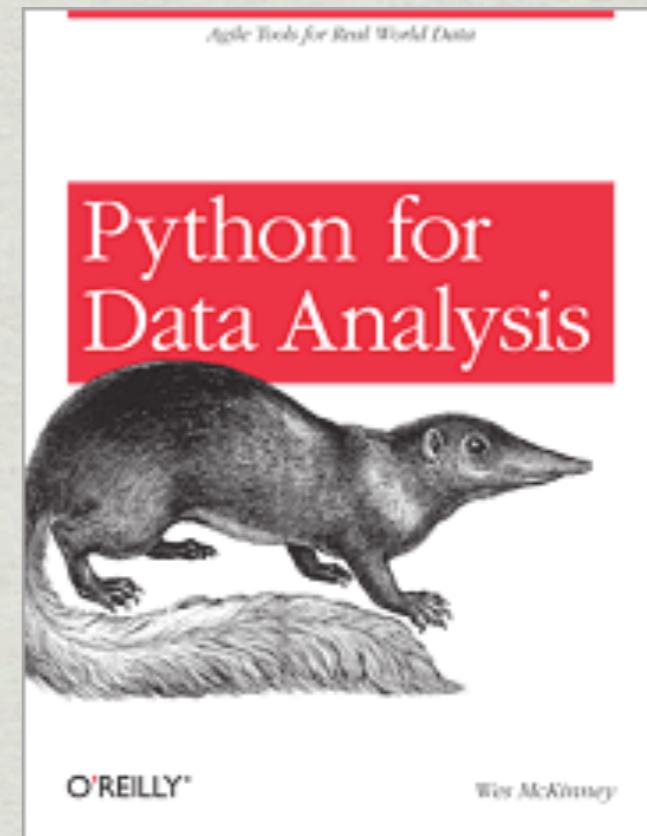
[https://github.com/polera/data analysis python pandas](https://github.com/polera/data_analysis_python_pandas)

Other Talks (by other people):

- from PyData NYC 2012

<http://vimeo.com/search?q=python+pandas>

https://en.wikipedia.org/wiki/Hierarchical_Data_Format



Thank you

- ✳ Comments and questions welcome.
- ✳ You can reach me at james@uncryptic.com or on Twitter @uncryptic.