

Snake Eyes

Seeing your data with Python

To Do

- Final project
- Class meetup
- Last module's homework
- Moving to Python
- This module's homework

Final Project

- Posted on Blackboard page
- Create a public visualization
- Use data relevant to a current policy, business, or justice issue
 - Find data
 - Get sign-off on project
 - Clean/transform data
 - Create visualization
 - Write about its importance
 - Get it up on our site

Final Project

- Consider this a portfolio piece
- Will stay up either as long as I can keep it or as long as you want
- Min 1 month public
- Initial proposal due 11/15
- Final project due 12/13

Initial Proposal

- I send back ~1/2 for further work
- Should include a link to the data (if possible)
- One paragraph on what the data is and what you're hoping to highlight
- One paragraph on data cleaning/munging that will be required and additional data sets needed
- One paragraph on what types of visualizations/tech you think you'll use

Suggested Data Sources

- UN <http://data.un.org/>
- World Bank <http://datacatalog.worldbank.org/>
- NYC open data <https://nycopendata.socrata.com/>
- NYS open data <https://data.ny.gov/>
- Assorted interesting data sets: <http://tinyletter.com/data-is-plural/archive>
- Anywhere else (just run it by me)

Suggestions

- Don't just do a pie chart
- Don't do a choropleth with a multi-color scale
- Please find something you're interested in
- Please do something less ambitious, better, unless you're really excited about the topic
- Please use data you can make public

Class Meetup

- Let's get water/coffee/tea/beer/food on Friday, November 18
- Will meet at Pershing Square right by Grand Central Terminal
- Come if you can
- I'm around for coffee otherwise

Last Module's Homework

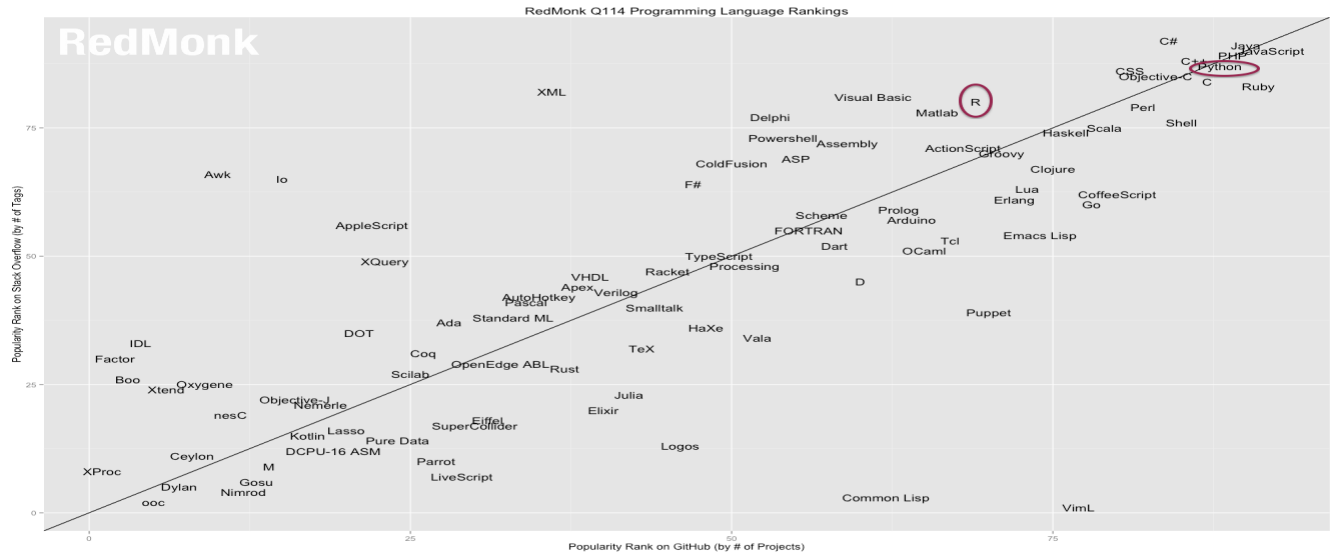
- Will go over in the app

Q1: As a researcher, you frequently compare mortality rates from particular causes across different States. You need a visualization that will let you see (for 2010 only) the crude mortality rate, across all States, from one cause (for example, Neoplasms, which are effectively cancers). Create a visualization that allows you to rank States by crude mortality for each cause of death.

Q2: Often you are asked whether particular States are improving their mortality rates (per cause) faster than, or slower than, the national average. Create a visualization that lets your clients see this for themselves for one cause of death at the time. Keep in mind that the national average should be weighted by the national population.

Moving to Python

- Switching from R to Python for this module
- Python is a great general purpose language, very popular



<http://redmonk.com/sograzy/2014/01/22/language-rankings-1-14/>

Visualization/Data Exploration is not Python's strength

- Not primarily a Read-Eval-Print Loop (REPL) environment
- Primarily viz tool is `matplotlib`: much lower level than `ggplot2`
- Much poorer set of baseline tools to analyze data

So Why Python?

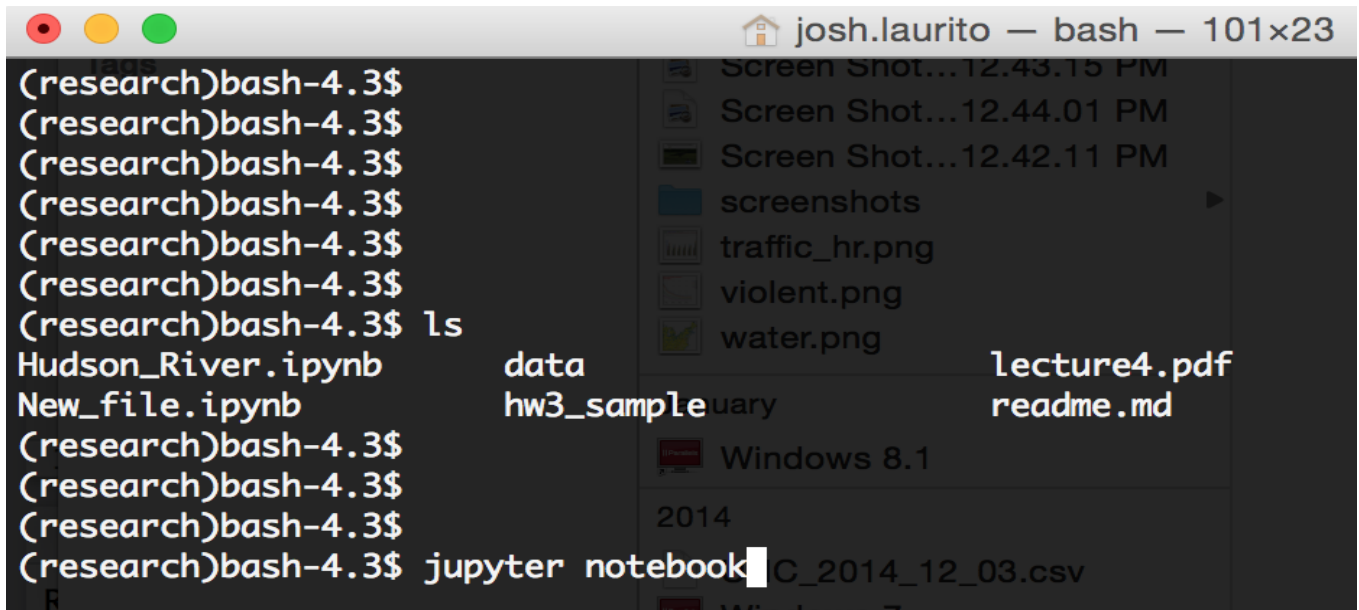
- Namespacing for multiple libraries
- More libraries for working with other languages/web
- Considered to be better when dealing with big data sets (debatable)
- I use Python primarily, but still use R for some problems
- Makes you more hire-able in some industries

Ways We Can Improve Python

- Use **Jupyter Notebook** to create REPL environment (link at) <https://jupyter.readthedocs.io/en/latest/install.html>
- Sister project **Anaconda** sort of acts as CRAN <http://continuum.io/downloads>
- Both from Continuum Analytics: US Government funded via DARPA
- NEITHER is required for this module's work: jupyter strongly encouraged

Keys to Using Jupyter

- Navigate to correct directory and run `jupyter notebook`

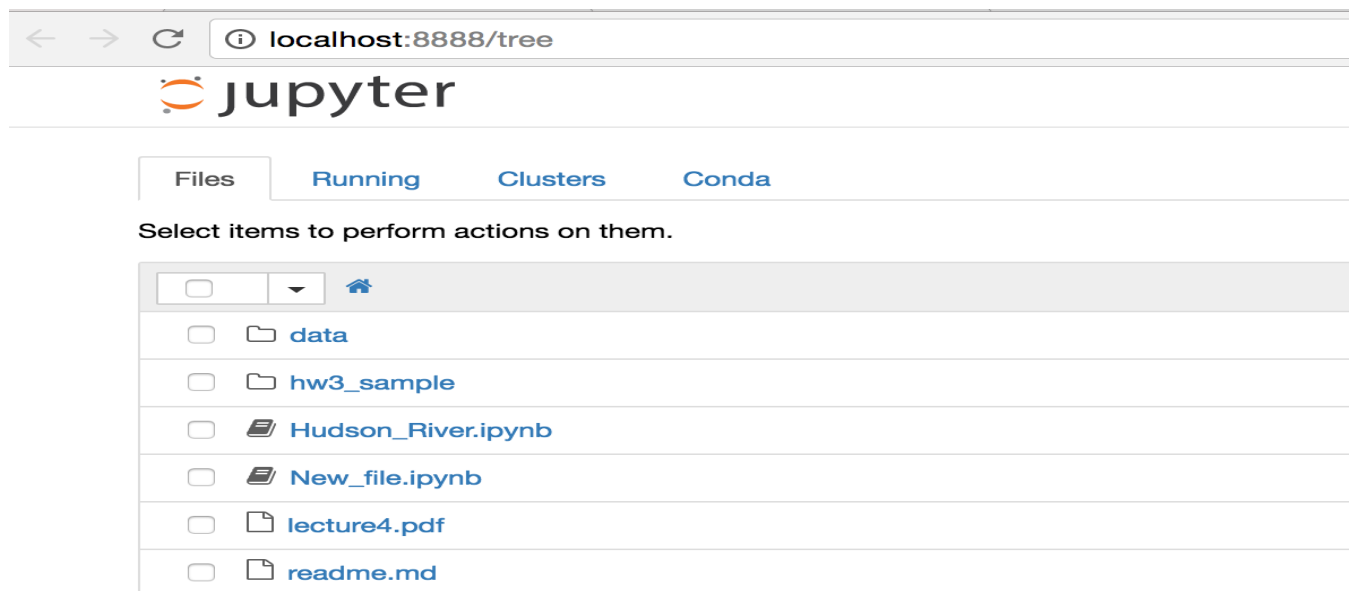


A terminal window titled "josh.laurito — bash — 101x23" showing a series of commands and their outputs. The prompt is "(research)bash-4.3\$". The commands entered are "ls", "Hudson_River.ipynb", "New_file.ipynb", and "jupyter notebook". The output of "ls" shows a directory listing with files like "traffic_hr.png", "violent.png", "water.png", "lecture4.pdf", and "readme.md". The "jupyter notebook" command is currently being executed, with a cursor at the end of the line.

```
(research)bash-4.3$  
(research)bash-4.3$  
(research)bash-4.3$  
(research)bash-4.3$  
(research)bash-4.3$  
(research)bash-4.3$  
(research)bash-4.3$ ls  
Hudson_River.ipynb      data  
New_file.ipynb          hw3_sample  
(research)bash-4.3$  
(research)bash-4.3$  
(research)bash-4.3$  
(research)bash-4.3$ jupyter notebook
```

Keys to Using iPython

- This will open a notebook viewer



Keys to Using iPython

- Once you have typed in code, run it line-by-line, like R
- Visualization libraries will appear in-line!

Libraries to Assist With Graphing

- **pandas** built in graphing <http://pandas.pydata.org/pandas-docs/stable/visualization.html>
- **Seaborn** <http://web.stanford.edu/~mwaskom/software/seaborn/>
- **bokeh** <https://github.com/bokeh/bokeh>

This module's homework

- Hudson River Water Pollution
- Data from Riverkeeper <http://www.riverkeeper.org/>

Good Luck!

