Python for Data Science

Sarah Wait Zaranek Ph.D.
Curoverse,Inc.

Boston Code Camp 26

Thank You to the Sponsers

Boston Code Camp 26 - Thanks to our Sponsors!

- Platinum
  - Microsoft
- Gold
  - HealthcareSource Quality Talent Suite™
- Silver
  - BLUEMETAL
  - RGood Software
  - Robert Half Technology
- Bronze
  - Robert Hurlbut Consulting Services
  - Infragistics Design / Develop / Experience
  - Syncfusion
  - slalom

- In-Kind Donations
  - GrapeCity
  - JetBRAINS
  - Manning User Group Program Member
  - Wiley
  - steema software

Goals for this session

- Gain familarity with Python & Python libraries used for Data Science
- Walkthrough "real world" example
- Provide examples to explore, modify, and expand (GitHub)
- Resources to learn more
- Have some fun

github.com/swzCuroverse/BostonCodeCamp26

Why Consider Python?

- Ease of use (high productivity)
- Strong ecosystem of supporting libaries
- Supports multiple programming paradigms : Procedural, OO, Functional...
- Widely used general purpose language

Setup

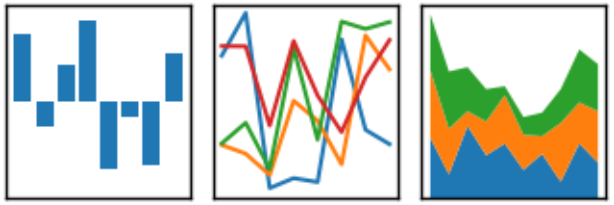
- Running Linux (Ubuntu, Xenial Xerus)
- Python (3.5, 64 bit) in a Jupyter Notebook
- Anaconda Distribution (700+ libraries)

Useful Libraries

- pandas
- NumPy
- matplotlib
- Seaborn
- scikit-learn

pandas

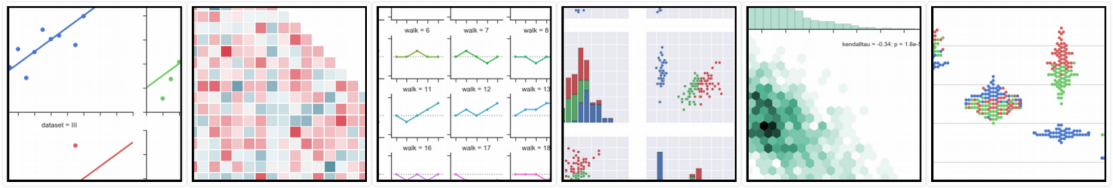
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



NumPy

matplotlib

Seaborn: statistical data visualization

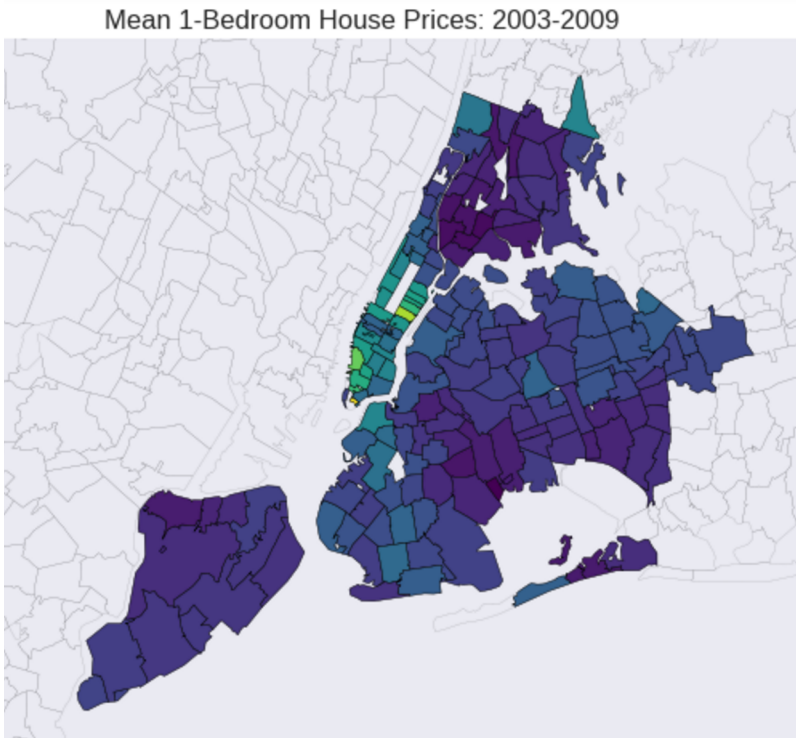


scikit learn

Demo : Housing Prices

- Examine NYC housing prices
- Use Python to
  - Import from text files (Excel and csv)
  - Perform data munging (cleaning, filtering, merging)
  - Calculate descriptive statistics and model data
  - Visualize data (charts, graphs and maps)

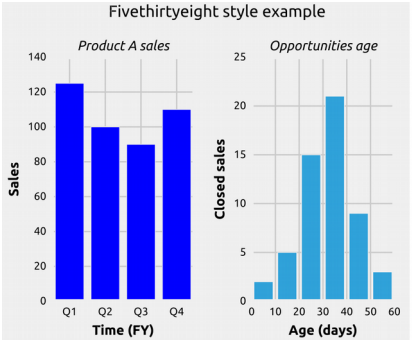
NYC OpenData



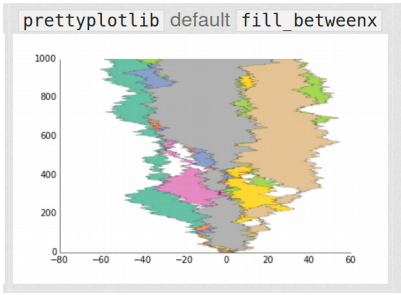
Customizing Graphs

- Colormaps
  - matplotlib.org/examples/color/colormaps_reference.html
- Named Colors
  - matplotlib.org/examples/color/named_colors.html

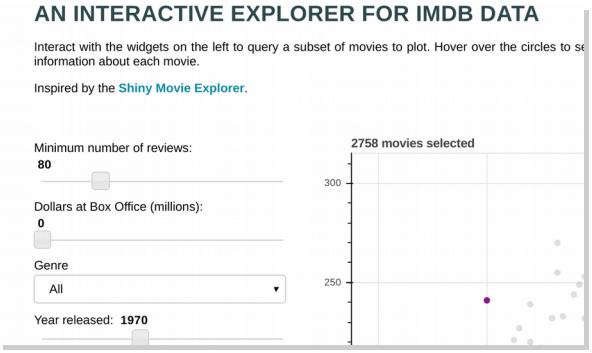
- Style Sheets
  - matplotlib.org/users/style_sheets.html
  - www.futurile.net/2016/03/31/colormaps-in-matplotlib/



- Custom Libraries (prettyplotlib)
  - blog.olgabotvinnik.com/prettyplotlib/



Other Capabilites



- Building Interactive Graphics
  - Bokeh
- Performance, Parallel and Large Data
  - Cython (Python to C)
  - `multiprocessing` (Python Standard Library)
  - Dumbo (MapReduce with Python)
  - Spark Python APIs

Questions ?