Pandas for Data Analysis

Phan Hoang / @pphoang17

StatsModels
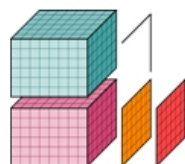Statistics in Python
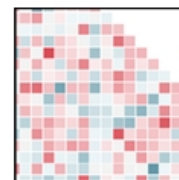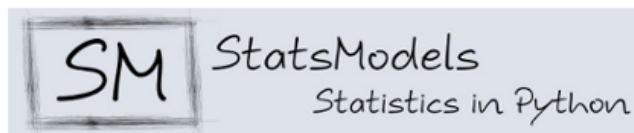
Seaborn

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

xarray

scikit learn

scikit-image
image processing in python

NumPy

SciPy

matplotlib

jupyter

python™

IP[y]:
IPython

What's Pandas ?

- Open-source library, build on top of Numpy
- **NUMFocus** project: https://www.numfocus.org
- **Pan**el **Da**ta **S**ystem
- D.S tools for data manipulation and analysis

Download sample data at:

http://bit.ly/2vqB1hg

Top Techniques ??

- Basic Usage

- Tips and Tricks

- Optimize Pandas Performance

- Data Munging and Visualization

Basic Usage

- Data Structure

- Basic Stats

- Indexing

- Handing Missing Values

- Plotting

Indexing

- Identification

- Selection

- Advanced: Alignment

Function Application

(applymap, apply, map, ..)

Data Munging and Visualization

(group, groupby, merge, join, concat, ..)

Performance

- Avoid loops
- Build-in method / Vectorization
    - Pandas
    - Numpy
- Masking
- Data Type (category)
- ...

Masking

(boolean indexing, query, where, all, any, ..)

Categories

(data type)

Questions ?

github.com/huyhoang17

References

- 10m with Pandas: http://pandas.pydata.org/pandas-docs/stable/10min.html

- Chris Albon Notes: https://chrisalbon.com/

- Intro to Pandas data structure:

  http://www.gregreda.com/2013/10/26/intro-to-pandas-data-structures

- Scipy Lecture Note: http://www.scipy-lectures.org/index.html

- DS Ipython Notebooks:

  https://github.com/donnemartin/data-science-ipython-notebooks

- https://www.datascience.com/blog/straightening-loops-how-to-vectorize-data-aggrega
  tion-with-pandas-and-numpy/

- https://penandpants.com/2014/09/05/performance-of-pandas-series-vs-numpy-arrays

- https://www.linkedin.com/pulse/doing-data-analysis-science-python-pandas-ali-syed