Data Analysis in Python with Pandas

Jonathan Wood
@JWood

Wintellect Core Services

Microsoft Regional Director

Consulting

Custom software application development and architecture

Instructor Led Training

Microsoft's #1 training vendor for DevDiv for 14 years

On-Demand Training

World class, subscription based online training

2011 MICROSOFT. PARTNER OF THE YEAR WINNER Microsoft Partner Network

Microsoft GOLD CERTIFIED Partner

Xamarin PREMIER CONSULTING PARTNER

Xamarin Certified Mobile Developer

Agenda

- Introducing and installing Jupyter and Pandas
- Introduce and go over Numpy basics
- Data analysis project
  - Loading data
  - Descriptive statistics
  - Missing and duplicate data
  - Column operations
  - Visualizations

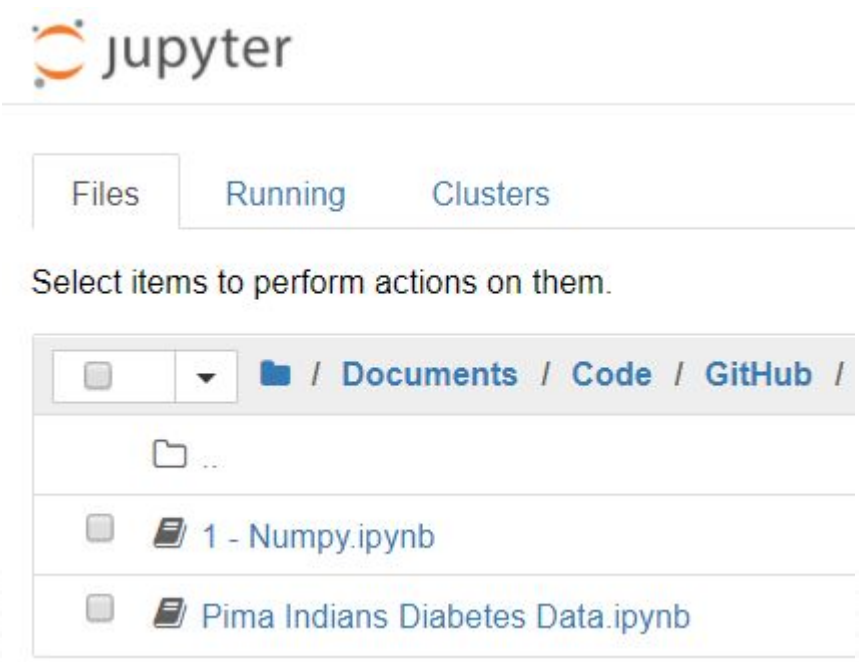
Poll

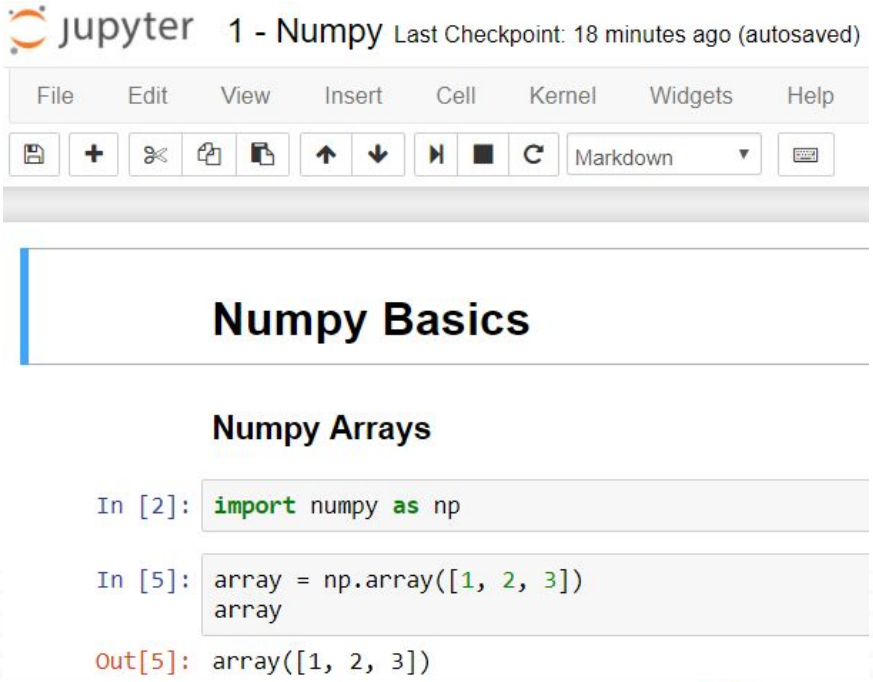
How much Python experience do you have?

Jupyter

Web based platform for data and scientific computing.

Can mix code, text, and visualizations.





Demo

Installing Python and Jupyter

Numpy

Basis for a lot of data science related libraries.

Features include:

- Handle large and multidimensional arrays.
- Random generation methods.
- Library of mathematical functions.

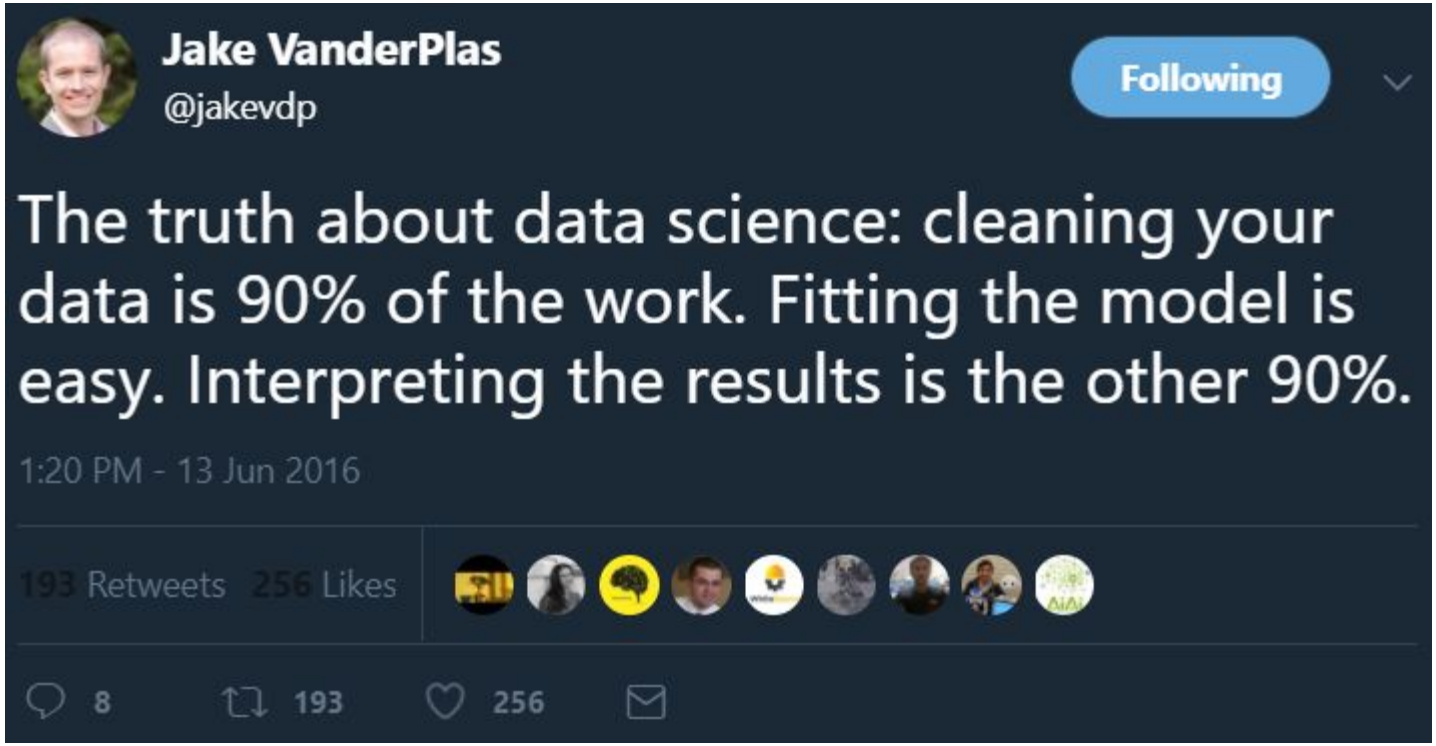
Demo

Basics of Numpy

Poll

Have you ever heard or worked with Pandas before?

What is Pandas?



What is Pandas

Open source tool for fast and easy data analysis in Python.

Built on top of Numpy.

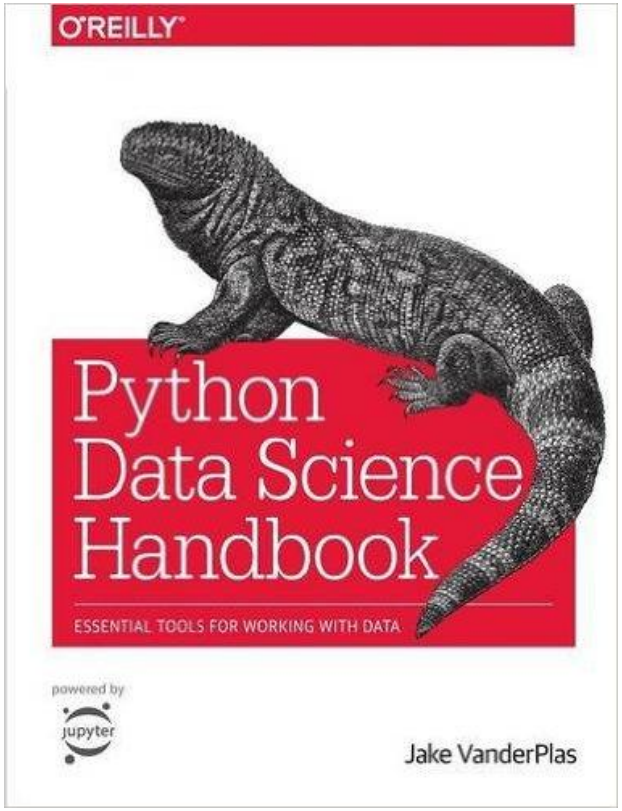
Pandas provides:

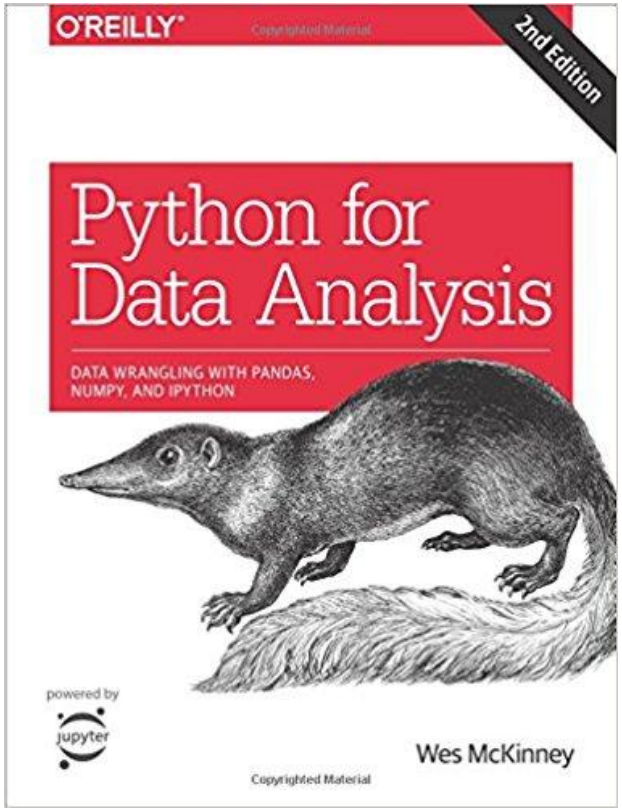
- Loading of external data from many formats
- Easy data manipulation
- Integrates matplotlib for visualizations

Demo

Data Analysis with Pandas

Additional Resources





Call to Action

Find interesting data and play around with it by doing analysis.

UCI Machine Learning Repository

Datasets Subreddit

Kaggle Datasets

GitHub Awesome Public Datasets

Questions?

Notebooks and slides will be on GitHub.

Feel free to reach out on Twitter – @JWood

Have fun answering questions with data!