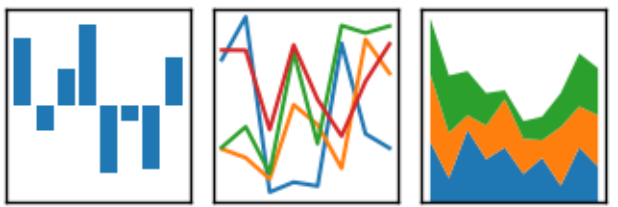
Analysis with Jupyter and Pandas

pandas

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



Paul Rodgers @prodg
Cloud RI

Who am I?

- Sr. Data Analyst at Virgin Pulse
- Early adopter
- Booted a PDP-8 and a Raspberry Pi (a few months apart)
- Generalist
  - Database architect for early web app (MySQL/mod_perl)
  - ETL veteran (without SSIS or Informatica etc.)
  - Operations automation
- Evangelist

What is Jupyter?

- Interactive code execution environment
- Tells a story
  - Allows the use of data, code and rich content
  - Enables the author to create a narrative
  - Engages the audience
  - Increases comprehension
  - Memorialize all aspects of the project

Examples

A notebook served by nbviewer.jupyter.org

The Waiting Time Paradox, or, Why Is My Bus Always Late?¶

A notebook served directly by github's viewer

LA Times article on the cost of legal settlements

Paul Rodgers @prodg
Cloud RI

What is Pandas?

- Programmable two dimensional tabular data management tool (Excel optimized beyond description)
- Similar to R dataframes
- Leverages Numpy (fast array math)
- Top notch CSV importer
- RAM based
- Rich data manipulation, categorization and period tools
- Database style joins

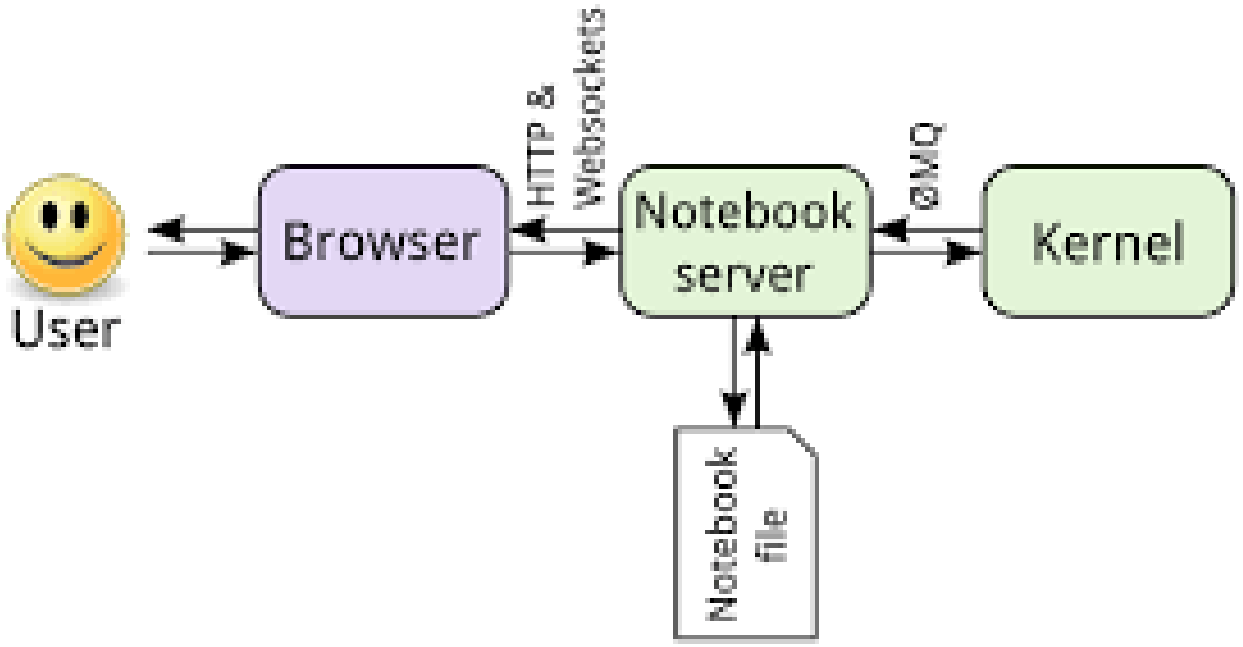
Who's using Jupyter?

- Academics
  - Paul Romer, Nobel Economist 2018
- Journalists
  - Los Angeles Times Data Desk (github.com/datadesk)
- Data Scientists
- Netflix!!
  - Papermill, nteract, Commuter

Who's using Pandas?

- Financial Analysts (Pandas was born here)
  - Time series and period savvy
- Data Scientists
- Me and (hopefully) you!

Architecture



Jupyter Extensions

- Notebook Extensions
  - Code folding
  - Snippets
    - Access to json library of code snippets
  - Freeze
    - Protect cells
  - Skip Traceback
    - Folding for error messages

Paul Rodgers @prodg
Cloud RI

Jupyter Ecosystem

- Nbviewer
  - append your notebook's name to a url
  - http://nbviewer.jupyter.org/github/prodg1974/Jupyter-intro/blob/master/Jupyter-intro.ipynb
- Jupyter Lab
  - New multitab interface
  - Improved interaction with the notebook (drag cells etc.)

Jupyter Ecosystem (continued)

- Jupyter Hub
  - Multiuser version for business teams, classroom, research labs
- Papermill
  - A tool for parameterizing, executing, and analyzing Jupyter Notebooks
- Interact
  - UI widgets

Hands On – The Data

- Sourcing data from retrosheet.org
- Compiled chadwick tools from source
- Convert raw (basically key,value) event files to 36 columns
- 191,196 events in 2017

Hands On – The Techniques

- Shell interaction and importing
- Intro to Pandas critical elements
  - Series
  - Dataframe
  - Index
- Poking around to understand the data
- Restricting output by columns and rows

Hands On – The Techniques (cont)

- Using python to build filters
- Querying the dataframe
- Merging data sets
- Grouping and aggregation

This environment

- Python 3.7.0
- Pandas 0.23.4
- Jupyter 4.4.0
- Ipython 6.5.0
- Cookiecutter
  - https://github.com/drivendata/cookiecutter-data-science
- Simple-salesforce 0.74.2

Basic system interaction

- IPython Magics
  - %history
- Shell interaction
  - ! cmd execution
  - %% shell command ‘stack’
- Shell output assignment
  - Variable = !command

Passing data to/from the shell

- Assignment to python variable
- Passing python output to the shell
  - {variable}
  - {command result}

Paul Rodgers @prodg
Cloud RI

Building our Data Set

- Looping shell commands with python
- Reading a csv file into a python dictionary
- Exploring with attributes
  - Shape
  - columns
- Exploring with methods
  - .head( )
  - .describe( )
  - .info( )

Pandas fundamental objects

- Series
  - .value_counts( )
- Dataframe
- Index

Paul Rodgers @prodg
Cloud RI

Filtering data with booleans

- Filter by column values
- Combining filters
- Restricting columns returned
- Working with strings

Paul Rodgers @prodg
Cloud RI

Grouping and Aggregating

- The group object
- The split, apply, combine concept
- Cross tabs

Paul Rodgers @prodg
Cloud RI