The Humble Data Frame

Becky Sweger
data engineer
@bendystraw

Data Engineer??

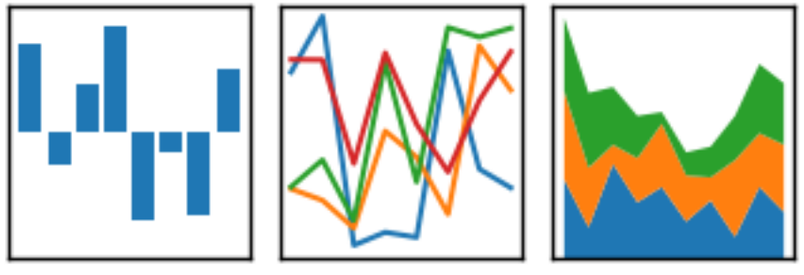
Pipeline Patterns

- Spreadsheets
- Database(less)
- Glue

A word about...

pandas

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



Spreadsheets

tables

❤️

code

demo

spreadsheet as data frame

Database-less

munging

❤️

vectors

demo

set-based data munging without a database

Glue

everyone

❤️

data frames

data conversion

demo

The Cons

- Data transfer
- Memory constrained

Data Transfer

- Ask: what are you solving for?
- Operate close to your data
- Compress

“Have 5-10 times as much RAM as the size of your dataset”

- Wes McKinney, creator of Pandas

Memory

- Read smarter (if you can)
- To the cloud!
- Divide and conquer
- Column store (e.g., Apache Arrow)

More Info

- Useful Pandas snippets (my cheatsheet): https://gist.github.com/bsweger/e5817488d161f37dcbd2

- Data Munging with Python and Pandas: https://github.com/bsweger/pandas-munging

- Comparison of Dask and Spark: http://dask.pydata.org/en/latest/spark.html

- Apache Arrow and the 10 things I hate about Pandas: http://wesmckinney.com/blog/apache-arrow-pandas-internals/