Topic Outline: Cleaning & Shaping Data

Revised: April 3, 2016

Materials

- Today's handouts: this outline
- Posted on *Topic outlines & links* page of website

Plan for second half of the course

Comments and suggestions welcome, but here's the current plan. The topics at the end are maybes, they depend on interest and time.

- Advanced data management with Pandas
  - Data cleaning: fixing numbers treated as strings, selecting rows and columns, Boolean selection
  - Data shaping: switching rows and columns, pivoting, indexing
  - Merging datasets: combining information from different dataframes/worksheets
  - Summarizing data: statistics (mean, standard deviation), grouping observations
- Updating and installing Python and packages
  - Conda and Pip
- Advanced graphics
  - Packages: Seaborn and Plotly
  - Applications: maps, animations
- Web scraping
  - Grabbing data from websites that are not designed for it
  - Package: Beautiful Soup
  - Problem: not easy to do if you don't understand the language of websites (html)
- Distribution, dependence, and dynamics
  - The mean isn't enough, we want to know the range of outcomes.
  - Distribution is about outcomes: equity returns, options, individual incomes, many more. Graphical methods for identifying the "long tail."
  - Dependence is about the relation between two variables, often summarized by correlation.
  - Dynamics is about dependence over time: Are good times followed by good times, or the opposite? Examples: equity returns, economic growth, bond ratings.

Approach to second half

- Cover material you find useful
- A menu for you to choose from
- **Ask for help** if you're stuck, either in class or if you use similar methods for your project
- Work small topics and applications into the flow

Pandas 2: Data cleaning

- Setup
  - Download IPython notebook
    https://github.com/DaveBackus/Data_Bootcamp/blob/master/Code/IPython/bootcamp_pandas-clean.ipynb
    and save Raw file in your `Data_Bootcamp` directory
    In short: GitHub ⇒ Code ⇒ IPython ⇒ `bootcamp_pandas-clean.ipynb` ⇒ Raw
  - Open in Jupyter
- **Want operator**
  - Keep in mind what we want, then figure out how to get there
  - Problems: numbers contain dollar signs or commas, interpreted strings; dataset contains a lot of stuff we don't want; rows and columns and flipped.
- String methods
  - Fixing oddities in data: dollar signs, commas, etc.
- Missing values
  - Identify missing values, then Pandas will automatically work around them
- Selecting variables (columns) and observations (rows)
  - Various things we don't use much
- Boolean selection
  - Things we use
  - Comparisons, boolean selection, the `isin` method

Pandas 3: Data shaping

- Setup
  - Download IPython notebook
    https://github.com/DaveBackus/Data_Bootcamp/blob/master/Code/IPython/bootcamp_pandas-shape.ipynb
    and save Raw file in your `Data_Bootcamp` directory
    In short: GitHub ⇒ Code ⇒ IPython ⇒ `bootcamp_pandas-shape.ipynb` ⇒ Raw

  - Open in Jupyter
- Want operator
  - Problem: Rows and columns are switched, or worse
- Setting and resetting the index
- Multi-indexes
- Switching rows and columns, pivoting
- Stacking and unstacking: essentially the components of pivoting

After class

Same as last time:

- Required
  - Submit Revised Project Ideas
- Recommended
  - Skim Project Guide
  - Skim Project Examples
  - Bounce around project ideas with friends