Data Bootcamp: Project Examples

Revised: September 6, 2016

As always, the pdf of this document comes with links.

If you're having trouble finding a project idea, here are some you could use or adapt. Feel free to modify them in any way you find interesting.

Prepackaged projects

These are relatively well defined projects that extend things we've done in class.

- **Demographics.** Delve deeper into the population dynamics of a country you find interesting, or compare two or three countries. You could start with our demography notebook, where we explore UN data on the age distribution of the population, life expectancy, mortality (deaths), and fertility (births).

  One example: Look at the data for China, see where it leads you.

  Another one: Anne Case and Angus Deaton generated a lot of discussion, and some controversy, when they documented an increase in mortality of middle-aged white men in the US and their causes of death, including suicide and drug overdose. Links: New York Times, more Times, Case-Deaton paper (look at the graphs), Gelman-Auerbach paper (ditto). Can you reproduce their work? Extend it? Take it in a different direction?

- **Current economic conditions.** Describe current business conditions in the US. (Most banks produce this kind of thing, you should be able to find examples.) This might include:
  - A list of indicators and why you think they're useful.
  - An assessment of the indicators: either the correlation with industrial production or some other quantitative measure of how closely it tracks the economy.
  - A summary of what your indicators overall suggest about the current state of the US economy.

  Chapter 11 of the Global Economy book has a nice overview of what's involved and FRED codes for some popular indicators. You can find more indicators — and descriptions — on the Bloomberg economic calendar.

  A variant: come up with a cool graphic. We tried a heatmap, but it didn't wow us. Another idea is a radar chart, like this one from the Atlanta Fed. This requires some serious coding, but the results would be striking.

- **Emerging market opportunities.** An assessment of (long-term) business conditions in emerging market economies. Here are a couple examples from the Global Economy course: Foxconn's Next Frontier and Opportunity in Ghana. In each one, we take a business proposition (start business X) and assess the relevant aspects of the business

climate in two or three countries. The World Bank's Open Data Blog is a good source of similar policy-related ideas.

Key data sources for this kind of work:

- The World Bank's Open Data Portal is a good place to start. A large collection of social, political and economic data for a broad range of countries.
- The World Bank also has (separate) databases related to Doing Business and Governance Indicators.
- The Penn World Table has data on GDP, employment and hours worked, and capital.
- The World Economic Forum does not produce its own data, but their Global Competitiveness Report has a good summary of information from other sources.

Random ideas

All of these need some work, but they seem promising to us:

- **Energy prices.** They've dropped a lot, but where will they go from here? We can think of this in terms of supply and demand. Demand depends primarily on growth: higher growth means higher demand. If China grows more slowly, theyll demand less. Supply depends on available resources and technologies to develop new sources — fracking, for example. These links focus on supply, which seems to be the central issue right now: Alphaville, Econbrowser.
- **Carbon and energy intensity.** How much energy do countries use? How much is due to GDP (rich countries use more energy)? How much to other things? Can we document the sources of differences in a useful way? Here are a couple blog posts: Conversable, Econbrowser.
- **How people spend their time.** The American Time Use Study (ATUS) tells us how people spend every hour of the day. You can imagine marketing interest in how much time people spend watching TV, accessing the internet, etc. Many other countries have similar data. Links: survey, Adweek, Business Insider academic paper, Nielsen.
- **Hong Kong/Shanghai share prices ("Hang Seng AH Premium").** Shares of some of the same companies are traded in both locations, but at different prices. Document this over time, showing how the discount on HK shares has varied over time. Why do you think that is? Some links: report 1, report 2, FT, Schoenholtz.
- **Deflation odds.** This is moderately technical, but nevertheless interesting. Inflation-protected government bonds (TIPS) have coupon and principal that are ties to the consumer price index, but theres a lower bound: coupon and principal can never go below their issue values. So if theres deflation (negative inflation) over the life of the bond, the payments dont go down. That gives the bonds an option-like feature that can be used to infer the (risk-adjusted) probability of deflation. More at the Atlanta Fed: blog post, paper.

Data and package projects

Here the idea is to describe a dataset or a package. Why stop there? Because enough is enough! But if you do this and still have time for more, feel free to keep going. And **if you get stuck, ask for help**.

- **Describe a package.** Take a package we haven't used and write a tutorial that explains to others how to install and use it. Some examples that have come up in class:
  - Beautiful Soup "scrapes" websites. The term refers to grabbing data that is posted online but not in a user-friendly format. We can still grab it with the right tools and a little understanding of the structure of websites.
  - Bokeh does interactive web graphics. Spencer's a big fan. One option is to take some graphics we've done in class and redo them in Bokeh, illustrating its capabilities.
  - Zipline is an "algorithmic trading library."
  - TA-lib "performs technical analysis with financial market data."
- **Describe a dataset.** Take a dataset you find interesting. Write code that reads it into Pandas and explain to others how the data might be used. See the course data page for examples. If this sounds insufficiently ambitious, keep in mind that some of these datasets are pretty complicated.

  We have a growing collection of examples, which could serve as your starting point:
  - MEPS. The Medical Expenditure Panel Survey covers medical expenditures for individuals. This short report illustrates its potential. Figure 1, specifically, shows that 22 percent of total expenditures come from 1 percent of individuals. (Think about that for a minute.) Brian LeBlanc has reproduced this figure from the raw data in an IPython notebook.
  - SCF. The Fed's Survey of Consumer Finances collects data every three to five years on household income and (especially) assets and liabilities. It's a great resource for things like student loan debt, the wealth distribution, and many other things. Here's a nice application to credit card debt and interest rates. Brian LeBlanc's notebook shows us the Python code.
  - Airbnb. Chase likes this collection of Airbnb data and is trying to figure out what to do with it. If you have ideas, get in touch.
  - College Scorecard. The federal government has posted an enormous collection of material online about US colleges: graduation rates, financial aid, salaries after graduation. Links: the data page, a 538 piece using the data, a Python program to read in one of the datasets, and a readable overview.

  If you have your own ideas, all the better. Or skim the programs in our sources and apps collection. Or our Lab repository. (Some of them are works in progress.)