# Capstone 1 Ideas

## Idea 1: What do you do with a PhD in science?

### Dataset:

* National Science Foundation’s Survey of Doctorate Recipients :
  + https://www.nsf.gov/statistics/srvydoctoratework/#sd&tools&micro&profiles&tabs-1
* Sponsored by the National Center for Science and Engineering Statistics and by the National Institutes of Health

### Description:

* Biennial Survey conducted since 1973
* Provides Demographic, education, and Career History Information about individuals with a research doctoral degree in science, engineering, or health (SEH) from US institutions
* Data available from 1993
* Key variables:
  + Demographics (e.g., age, race, sex, ethnicity, citizenship)
  + Educational history
  + Employment Status
  + Field of degree
  + Occupation

### Problems/Questions:

* Which doctorates earn the most money?
* Which doctorates stay in academia the most? Industry? Government? NGOs? Etc.
* Which doctorates live the longest?
* Which doctorates are healthiest?
* Which doctorates have the most children?
* Which doctorates have the most prior education?
* How has the demographics of doctorate earners changed with time?
* What are the main ways in which doctorate earners differ from non-doctorate earners? (Would need data about non-doctorate earners for this)

### Notes/Concerns:

* Lots of numerical data here, so that looks good. Too much?
* Data rich
* Also have access to National survey of college graduates – compare to regular graduates?

### Idea 2: Business owners

Dataset:

* U.S. Census Bureau’s Survey of Business Owners and Self-Employed Persons:
  + https://www.census.gov/programs-surveys/sbo/about.html

Description:

* “Provides the only comprehensive, regularly collected source of information on selected economic and demographic characteristics for businesses and business owners by gender, ethnicity, race, and veteran status.”
* Collected every 5 years since 1972
* all nonfarm businesses filing Internal Revenue Service tax forms
* Business ownership is defined as having 51 percent or more of the stock or equity in the business and is categorized by:
  + Gender: Male; female; or equally male/female
  + Ethnicity: Hispanic; equally Hispanic/non-Hispanic; non-Hispanic
  + Race: White; Black or African American; American Indian or Alaska Native; Asian; Native Hawaiian or Other Pacific Islander; some other race; minority; equally minority/nonminority; nonminority
  + Veteran status: Veteran; equally veteran/nonveteran; nonveteran
  + Publicly held and other firms not classifiable by gender, ethnicity, race, and veteran status
* Characteristics of Business Owners:
  + How Initially Acquired Business
  + Year Acquired Ownership of Business
  + Primary Function(s) in Business
  + Average Hours Per Week Spent Working
  + This Business Primary Source of Income
  + Prior Experience Owning a Business
  + Highest Level of Education Completed
  + Age of the Owner in 2012
  + Owner Born a Citizen of the United States
  + Service-Disabled and Other Veteran Characteristics

Questions/Problems:

* Has the education of business owners increased with time?
* Has the percentage of the economy made up by businesses grown or shrunk?
* Has the number of small businesses grown or shrunk?
* What are the main sectors that small business owners work in?
* How old are most entrepreneurs when they start their first business?
* How does the success of second businesses compare to first businesses?
* Are firms with more owners more successful than firms with fewer owners? Older vs younger?

Notes/Concerns: Lots of data in tables that I’m able to export; many, many tables

### Idea 3: financial well-being

Dataset:

* Consumer Financial Protection Bureau’s National Financial Well-Being Survey:
  + https://www.consumerfinance.gov/data-research/financial-well-being-survey-data/

Description:

* 6,000 responses to the agency’s 10-question Financial Well-Being Scale
* additional demographic and financial information
* Includes information on:
  + Income and employment
  + Savings and safety nets
  + Past financial experiences
  + Financial behaviors, skills, and attitudes

Questions/Problems:

* What are the main factors contributing to financial well-being? Top 5
* What are best practices for financial well-being for someone just graduating university? Mid-life? Retiring?
* What financial attitudes are best for becoming wealthy?
* How does having children impact one’s financial attitudes, behaviors?
* Does education improve one’s financial well-being and attitude towards money
* What are the main ways men and women differ when it comes to spending habits?
* How does greater understanding of finance (stocks, mutual funds, volatility, inflation, compound interest, etc.) affect financial behaviour?

Notes/Concerns:

* Have stub code to read the survey data into Python, R, etc.
* Survey Data File is **all** numbers; little string or categorical data which could be a downside, I think; in other words, survey is small; on 2nd thought, might not be an issue because the numbers stand in for verbal responses
* Comes with a user’s guide, which is a plus
* Many, many columns, a big plus

## Extra:

#### Two thousand billionaires

Dataset:

* Peterson Institute For International Economics: <https://piie.com/publications/working-papers/origins-superrich-billionaire-characteristics-database?ResearchID=2917>
* Forbes: <http://www.forbes.com/billionaires/list/>

Description:

* This dataset contains biographical information about every billionaire in the world for the years 1996, 2001, and 2014: gender, age, country of origin, industry, company type, source of wealth, and so on

Question/Problem to Analyze:

* What factors contribute most to becoming a billionaire? What are the commonalities among billionaires?
* What factors distinguish American billionaires from European ones? What factors distinguish software billionaires from other ones? Young billionaires from old ones?
* How has the composition of billionaires changed over time?
* What distinguished billionaires from mere-millionaires? (would require a dataset of millionaires, which would be vastly larger than the fewer than 2,000 billionaires in the world, no? )

Notes/Concerns:

* Already a complete working paper on this topic. Would not be doing original work. Would simply be reproducing the author’s work, no?
* This project looks like one that only requires Excel, specifically the use of Pivot Tables. It seems like a classic business intelligence problem, not a data science one. I know the line is often blurry, but still, shouldn’t there be number crunching of some sort to reveal patterns?