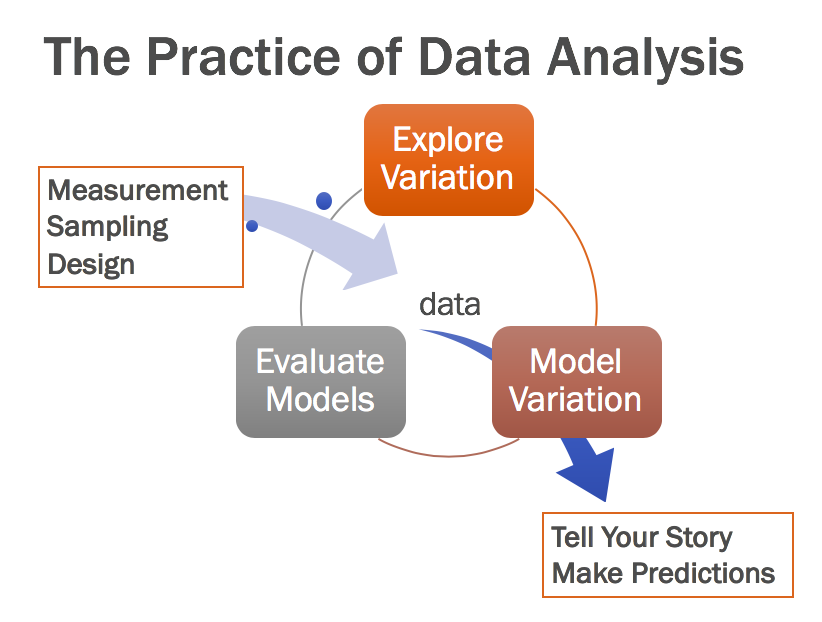
Week 2.2 - Exploring Variation

* Questions from Reading
  + http://tinyurl.com/datacampsandbox
  + library(okcupiddata)
  + head(profiles)
* Class in three parts: Exploring variation, Modeling Variation, Evaluating Model
  + 
  + What did we do 2.1? Just the data part (measurement, sampling, we didn’t do design but that stuff is involved there)
    - Take a look at profiles (from okcupid) -- how is this tidy data? What are the observations? What are the cases? What is the type of observation?
  + Today (2.2), Exploring variation
* Exploring Variation in Science
  + Science in school (5 steps of scientific method)
  + Science in real life (theory space and data space)
  + Think back to Distribution Triad (aka Westvaco): theory space is about DGP and data space is about sample
* Let’s try this “DGP vs. sample” (theory space versus sample space) idea <https://www.okcupid.com/> - free online dating app
  + <https://theblog.okcupid.com/how-to-make-a-great-dating-profile-e6427e94af52>
  + <http://tinyurl.com/datacampsandbox>
  + How to look at data
    - Str
    - Head
  + Which variables can we make a histogram for? (What’s special about these variables?)
  + Let’s take age
    - gf\_histogram
    - **How would you describe this distribution?**
    - **Why does this distribution look like this?**
    - **Is this sample or population?**
    - Favstats
    - Filter out people over 100
    - Gf\_histogram on filtered data
    - What happens when we change bin size?
    - **Distribution triad - where did this sample come from? What is DGP/population?** Is it people on okcupid or people trying to date or dating aged adults? Distribution does not look normal… could it have come from normal population?
  + Try with height
    - How would you describe this distribution?
    - Why does this distribution look like this?
    - Is this sample or population?
    - Favstats
    - Filter
    - Distribution triad - where did this sample come from? What is DGP/population? Could it have come from normal population?
  + Try with body\_type
    - How would you describe this distribution?
    - Why does this distribution look like this?
    - Is this sample or population?
    - Histogram? Favstats? Tally? When do we use histogram? When do we use tally, or bar? What happens when we run favstats?
    - Distribution triad - where did this sample come from? What is DGP/population?
  + Lesson summary:
    - always look at your data
    - always wonder -- where did it come from (DGP)? Back to science -- going back and forth between looking at data and thinking about DGP.

Extra R Practice

* Age in days (rather than years)
* How would we look at just a few people’s Age and Height and Sexual preference
* How would you figure out what percentage of this sample is male?
* Divide data up into rich, middle, and poor
* How to look at first 10 rows of profiles
* Filter for only people who provided their income
* Filter for unmarried people
* What are some interesting research questions you can ask with this dataframe?
* Put a title, change x label, change y label
* Interpret 5 number summary
* Small versus large sample