Problem Set 2, Data Science Tools

Kevin Kontchou

January 2018

Data Science Tools

1 Measurements

Measurement is how "insights" or "policies" are constructed, because one cannot construct a policy without first measuring what the policy is meant to accomplish.

2 Statistical Programming Languages

The three main statistical programming languages discussed in this class are R, Python, and Julia

3 Web Scraping

With the proliferation of the internet, data is being collected all the time and stored in publicly accessible places (we call these webpages). Thus, one of the tools in a data scientist's toolbox should be the ability to leverage this information to better inform the objective at hand

4 Handling Large Data Sets

In some cases, you might encounter data sets that are too big to fit on a single hard drive (or are too big to fit in R/Julia/Python's memory).If you try to load into R a data set that is larger than your machine's RAM capacity, your machine will freeze and you'll have to unplug it if you want to use it again.What do you do when you can't open all of your data? Depending on how the data is stored, you may be able to split the files up into manageable chunksBut that's not a viable longterm solution if your data set gets updated, or if you want to compute summary statistics on the full set of data, or if you want to do other operations on it, like create a new variable.

5 Visualization

Visualization is an important tool for data scientists. Visualization allows humans to see in multiple dimensions what the data looks like, spot outliers, and otherwise perform "sanity checks" on the data.

6 Modeling

Now that you've collected your data, cleaned it up, and visualized it, you're ready to start doing some statistical modeling.