ECON 5253: PS2

William Lorton

February 9, 2021

Data Scientist Tools List

1. Measurement
    - Must know how to appropriately measure the components involved in your task.
    - Often involves giving a quantitative interpretation to a problem that began with only a qualitative dimension.
    - Leads to concrete policy creation.
2. Statistical programming languages
    - Main DS languages are R, Python, and Julia.
    - Stata, SAS, SPSS, Matlab, and JavaScript are also used.
    - R and Python are the most popular currently (Julia is the relatively new up-and-comer).
    - Note the difference between scripted languages (e.g. those listed above) and faster compiled languages like C++, C, and FORTRAN.
3. Data visualization
    - Communicate insights to those with and without intimate knowledge of the data.
    - Identify outliers, check for correctness, see data from new perspectives, etc.
    - R: ggplot2 package; Python: matplotlib package; Julia: Plots.jl package.
    - Lots of companies use interactive software like Tableau as well.
4. Big Data management software
    - For when data is too large for your hard drive(s) and/or overloads your computer's memory when using R/Python/Julia/Matlab/etc.
    - Use Resilient Distributed Datasets (RDDs) via Hadoop or Spark; divides data into smaller bits and executes actions on it simultaneously via a cluster of computers.
    - Can do variety of typical commands: subsetting, merging, summary statistics generation, etc.

    - Also helpful to know SQL (a database management language) when dealing with very large datasets.
5. Data collection tools
    - For example, know how to web scrape (lift data off of public web pages very quickly).
    - Use an Application Program Interface (API) if available and/or parse HTML file text.
6. Modeling
    - Combine knowledge of statistics and statistical programming languages to test theories, predict outcomes, and establish causal relationships.
    - Of course, must have already collected, cleaned, and (usually) visualized the data.