Problem Set 2 - ECON 5253

Thomas Mondry

February 6, 2022

1 The data scientist's tools

I understand data science as a process for using data to extract insights; the tools used by data scientists can be organized by the role they play in this process.

- Tools for working with data
  - Data acquisition: Web scraping, usually either through manual HTML scraping or APIs
  - Big data management: File splitting (useful in limited situations), computing clusters & RDDs (using software such as Hadoop or Spark), database transformation tools (i.e. SQL)
- Tools for getting insights from data
  - Data visualization: Packages built to rurn natively in programming languages (e.g. R's ggplot2, Python's matplotlib, Julia's Plots.jl), or third party visualization software (e.g. Tableau, Power BI)
  - Project formulation & measurement: awareness of what is contained in the data & quantifiable parameter(s) of interest
  - Modeling:
    - Goals: testing theories, predicting behavior, extracting causality (especially useful in econometric applications)
    - Programming languages: R and Python (most common), Julia (relatively new & powerful)
    - Model types: regression, classification, clustering