PS2 Yarberry

Megan N. Yarberry

January 28, 2020

1 The Main Tools of a Data Scientist

1. **Measurement** - having a solid understanding of what data is to help achieve direct of what needs to be accomplished
2. **Statistical Programming Languages** - the three main statistical programming languages are R, Python and Julia.
3. **Web Scraping** - the ability to scrap data off from the web two different ways either by a program interface called API or through Parsing text from HTML files to extract the data
4. **Handling Large Data Sets** - Most laptops and desktop computers are not strong enough to handle large data sets as a result there is two ways to dealing with the data.
   (a) **RDDs** - Resilient Distribution Datasets using a cluster of computers and software and you break it up into manageable chunks and executes actions on those chunks in parallel.
   (b) **SQL** - is the most common database language to transform data into a more usable form for statistical software to use.
5. **Visualization** - one of the most important tools allows us to project the data in an visual and organized manner compared to tables and numbers. Programs that allow us to do this is ggolot2 for R, matplotlib for Python, Plots.jl for Julia and Tableau.
6. **Modeling** - once the data is collected, cleaned and visualized you can now model the data to test theories, predict behaviors and explain behavior.