BIG DATA, MACHINE LEARNING, AND THEIR REAL WORLD APPLICATIONS

Jonathan Wayne Korn

2021-04-14

Outline

Week 01, 06/28 - 07/04, Day 3

Wednesday* (Pre-Session Readings and Assignments)*:

- Access the supporting script(s)/file(s) in the following github repo you cloned yesterday to prepare your tools for the course @ 04_prepare.tools and review the document labeled prepare_tools_guidelines. *(… attempt to prepare your tools, we will be reviewing the process in class so any mistakes we can correct.)*
- Watch the following video @ R or Python: Which Should You Learn in 2020? *(19 minutes)* and learn about the iconic battle between R and Python programming languages for dominance in data science. *(Which language you think will be the winner?) (… be prepared to discuss in class.)*

Wednesday* (morning session ~ 9:10am to 11:00am)*:

- R and Python programming discussion.
- Preparing the Tools *(R and Python, Anaconda, Github)*
  - Importing Data:
    - Data Objects *(local/global variables, lists, vectors, matrices, dataframes)*
    - File Types *(cvs, xlsx, SAV, etc.)*
    - APIs *(Discussion on Connecting to API)* at least one project will require to access an api *(i.e. quantmod ~ query stock data)*

Wednesday* (afternoon session ~ 1:10pm to 3:00pm)*:

- Introduction to Processing Data in R and Python ~ *Structured*
  - Subset Variables
  - Data Type Conversions
  - Imputing Missing/NA Values
  - Imputing Outliers
  - Data Normalization Techniques
  - Optional Balancing of the Data *(Randomize Sampling, Automated NoiseFilters)*

Wednesday* (Post-Session Readings and Assignments)*:

- Access Assignment #1 documentation @ folder labeled assignment1
  - Make sure to use the support scripts for R and python @ r_learning and py_learning