Principles of Quality Data Analysis

1. **Write a data analysis plan before conducting the analysis.**
   1. What datasets are being used?
   2. What variables within those datasets are important?
   3. What are the variable definitions?
   4. What outcome/final variables are going to be created?
   5. What summary statistics or graphs are going to be output?
2. **Write a readme in the project directory.**
   1. Who – are analysts and the PI?
   2. What – are you doing? Include a quick project summary for context. Which scripts do which tasks?
   3. Where – are the locations of the data, the scripts, and supporting documents?
   4. When – did this project start and when was it last updated?
   5. Consider – A project map: Raw Data -> Processing/Filtering Scripts -> Intermediary Data -> Further Processing/Analysis -> Output
3. **Use clean, consistent programming syntax.**
   1. Use consistent and proper indentation to maximize readability. Variable lists in KEEP or SELECT statements should be one line per variable.
   2. Use consistent capitalization, underscoring, camel casing, and hyphenation.
   3. Use consistent and descriptive variable AND dataset names (Instead of “RO1” use “AllSimDates” or instead of “FROM\_DT” use “Delivery\_DT”.
4. **Organize programs in a clear, clean and consistent manner.**
   1. Start each script with:
      1. The programmer’s name, date and purpose of the program.
      2. List all datasets to be read in and where they came from if not raw data.
      3. List all datasets to be created/saved to be used later.
      4. List all important variables in all datasets.
      5. Defining options and libraries.
   2. Clearly comment all major sections of programs with their purpose and intent.
   3. Wherever possible, assign long lists of things to variables/formats (codes, variable names, etc).
5. **Organize folders clearly and cleanly, with obvious names. Create a new folder and set of subfolders for each project – keep the data, documentation, and programs separate.**
   1. data
      1. /derived\_data
      2. /raw\_data
   2. scripts
      1. /sas
      2. /R
      3. /functions (or macros)
   3. reports
      1. /manuscripts
      2. /tables and figures
   4. documentation and plans
6. **Keep meticulous notes.**
   1. For each project, keep a ‘Notes and Meeting Minutes’ word document with dates, names, keywords, and a summary/action items from each meeting so you can ctrl-f items later.
7. **Stay on top of version control**.
   1. Either use Git or archive folders + dated versions. Git is strongly preferred.

Daily Workflow Tips

1. **Separate work sessions into 1-2 hour intervals with breaks in between for maximum productivity.**
   1. Spend the first 15-30 minutes reviewing notes, planning, and if modifying code, reviewing the relevant sections to be modified.