Best Practices for Data Cleaning Using R and RStudio

Some good general principles for working in R include: keeping an eye on file paths, remembering to check the working directory, and verifying which project space you're working in. For more information on installing R and RStudio, see this guide: https://subjectguides.lib.neu.edu/HowToInstallRRStudio

Important Vocabulary

- **R:** A computer programming language.. You can use R without having to download RStudio.
- **RStudio:** An integrated development environment (IDE) for programming, usually in R. You need to download R to use RStudio.
- **R Markdown:** A file format for making dynamic documents with R. An R Markdown document is written in markdown (an easy-to-write plain text format) and contains chunks of embedded R code.
- **R Notebooks:** An R Markdown document with chunks that can be executed independently and interactively.
- **ggplot2:** an open-source data visualization package meant for use with the programming language R. The software is free to download and install on personal computers.

DITI encourages you to use projects to keep your work organized. Working in a project will help you make sure you know where your working directory is, so you will be able to construct accurate file paths. Many difficulties that might arise when using R and RStudio emerge from forgetting to open your project or losing track of your working directory. Even if you start by opening your project, it's still a good idea to double-check your working directory by using the getwd() function. Here is a great resource on using projects in RStudio.

A Note on Organizing Your Work

Take some time thinking about how you want to set your data up. Things can get messy if you don't create clearly named, unique folders and save data to the right location on your computers. You should also have well-documented code that you can use consistently from one dataset to another instead of having to write new code at the start of each session in RStudio.

DITI suggests making sure there is a **general project folder with all of your coding work** saved somewhere you can find it, holding all of the projects you work on using R. Be clear with how you name your subfolders and datafiles, as you will have to find them later when setting up your working directory in RStudio.

Best practices for data cleaning

- **Read the documentation for any functions you are using** to make sure that you're using them appropriately.
- **Carefully familiarize yourself with the dataset** and make sure that you are considering its structure as you are manipulating it in R**.**
- **Saving:**
    - **Regularly save** your projects, code, and data.
    - Always **save an unmodified version** of your data before you begin making any changes.
- **Documentation:**
    - **Document all changes** that you make to your data. Different kinds of documentation will be necessary at different levels—you should be including comments in your code, keeping notes to yourself as you're working, and writing more detailed documentation as necessary so that you are prepared to share the results of your research.
    - **Establish consistent conventions for naming variables** and keep track of when you overwrite them or switch to a new data source**.** You call variables into existence because you think you will use them frequently or will need them later. So, failing to keep track means that you might lose a lot of ground. You should make sure that your variable names are clear, brief, and specific.
    - **Make a plan to address irregularities in the data** by developing a documentation system for error-management and instructions for troubleshooting errors.
- **Review your results** after you make any global changes.