Data Management in Stata
University of Kansas

1 Introduction

In this workshop, we will walk through the application of data management techniques in a research environment. The focus is on best practices for downloading, transforming, merging, and saving data.

Although it may not be the most fun part of a data analysis project, your future collaborators (including yourself!) will be grateful for the time spent carefully arranging and documenting your process.

2 Record Data Selections

The data come from the WQP with the following selections:

- State: Kansas
- Data Source: NWIS USGS and WQZ (EPA)
- Date Range: 01/01/2020 - 12/30/2025
- Data Profile: Sample Results (Biological)
- File Format: CSV

3 Setting up your Folders in File Explorer

Folders should be set up for each part of the project. For a data analysis project, your file system should be something similar to the following:

- Data
- Working Data
- Do Files
- Literature
- Output

It is important that your working data is separate from your raw data. You want to always have access to the original files sent to you, in case something happens to the data you are working with, or you need to restart. The "Data" folder is used for your raw data, and "Working Data" will be for your working data.

The "Output" folder will be for any graphs or figures you create, and the "Literature" folder for any studies, guides, or documents that are relevant to your project.

4 Set Up Your Stata Environment

Set your working directory using the *cd* command or the file menu in the top left. Similarly, you can import your data using the *import* command or the file menu. **It is important that even if you choose to use the file menu to load your data, that you copy the line of code into your do file.**

5 Save, Save, Save!

Make sure to save your data near the top of the do file if you imported data, as well as at meaningful transitions in the function of the code. For this workshop, this is done at each stage of data processing. You may find that this is more or less than what you need, **but the minimum is once, at the end of the do file.**

Saving in these instances will be routed to the "Working Data" folder.

6 Modifying Data

Good, reproducible code should run cleanly from top to bottom, calling in data, transforming it, and saving it without any assistance. To do so, you should track all changes to the data in the do file. Common modifications include dropping variables, renaming and labeling variables, reshaping data, and merging datasets.