1. Data Wrangling Introduction

Overview

Questions

- Why are Jupyter notebooks useful for cleaning and wrangling data?

Objectives

- Understand how Jupyter notebooks provide a method of documenting the steps involved in cleaning and wrangling data.

Why Pandas and Jupyter Notebooks

- Reproducibility and transparency in scientific research are essential for building trust and confidence in the results, promoting scientific integrity and advancing knowledge.

- Python and the Pandas library a source of powerful functions for cleaning, formatting and analyzing data in a clear and systematic way.

- Jupyter notebooks provide a solution for documenting and sharing Notebooks

    - You can weave in code, images, videos, latex, JavaScript widgets, etc.

- Python, Pandas and Jupyter notebooks promote transparency and reproducibility in data analysis.

What this Workshop Does not Cover

- How to write basic python code

- How to set up a notebook environment

- How to run other programming languages e.g. R in jupyter notebooks

Jupyter Notebooks

- Jupyter notebooks provide an interactive and user-friendly environment for data analysis and visualization
- They allow for clear documentation of the analysis process, including the code, results, and narrative
- They enable reproducibility and facilitate collaboration
- They support multiple programming languages, including Python, R, and Julia, making them versatile for a wide range of data science tasks
- They are widely used in data science and machine learning, making it easy to find resources and help online.

Lesson Logistics

- We will be using SAGE3 to run each lesson and complete the exercises
- Participants will form groups of 4 and work together in separate `Boards`
- Binder allows us to run a Jupyter Notebook instance from the link included in each lesson.

Using SAGE3 and Binder: A Brief Demo