Module 5: What Tools Are Used to Manipulate Data?

Background

This module follows up on the prior module on data tables and databases to explore how software systems and tools support and facilitate the data work that we do.

Additional Reading for Facilitators

- Chopra, Bhavya, et al. "CoWrangler: Recommender System for Data-Wrangling Scripts." *Companion of the 2023 International Conference on Management of Data*, Association for Computing Machinery, 2023, pp. 147–50, https://doi.org/10.1145/3555041.3589722.
- Kandel, Sean, et al. "Wrangler: Interactive Visual Specification of Data Transformation Scripts." *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11*, ACM Press, 2011, p. 3363, https://doi.org/10.1145/1978942.1979444.

Module Motivation

This module explores the affordances (benefits) and drawbacks of different tools available to learners. We examine what we need from a data tool, why some tools are better suited to different kinds of data processing work, and how to identify and learn to use potential new tools.

Learning Goals

Students will be able to:

1. Identify and delineate key characteristics between data processing tools (such as a local Excel document vs. an Excel document saved to the cloud vs. Google Sheets)
2. Identify key differences between programming-based data tools vs. spreadsheet tools and decide which is more appropriate for the task at hand

Warm-Up: Data Tools (Slide 1)

Purpose:

Check in with learners' associations with data tools or data software.

Discussion Topics:

- **Storage systems**: Cloud storage vs. local storage
- **Product variations**: Different companies' solutions
- **Security concepts**: Encryption and access controls
- **Access considerations**: Who can potentially see file contents?
    - Co-workers
    - Clients
    - Product companies
- **Practical factors**:
    - Licensing fees
    - Availability across different operating systems or devices

Facilitation Notes:

- Use this as an opportunity to identify tools already in use within the organization
- Help learners begin categorizing tools by key characteristics
- Encourage sharing of experiences with different tools

Discussion 1: Spreadsheets vs. Programming Tools (Slides 3-11)

Purpose:

Open-ended exploration of tools used in the organization and their classification as "spreadsheet" or "programming" tools.

Key Framework:

Primary distinction: Tools where you interact with data via GUI vs. those where you directly program data transformations.

Important Notes:

- **Overlap exists**: Some tools fit both categories (e.g., you can program in Excel)
- **Both are useful**: The goal is to determine when one may be more useful than the other
- **Context matters**: Consider dataset size, data type, and required transformations

Discussion Considerations:

When to Use Spreadsheet Tools:

- Smaller datasets
- Simple data transformations
- Visual data exploration
- Collaborative work requiring accessibility
- Quick prototyping and ad-hoc analysis

When to Use Programming Tools:

- Large datasets
- Complex data transformations
- Reproducible workflows
- Automated processes
- Advanced statistical analysis

Activity Structure:

Step 1: Tool Inventory

Review tools currently used in the organization and categorize them.

Step 2: Project Reflection

Step through different past projects and discuss:

- What tools were originally selected?
- Given current knowledge, would you keep using those tools?
- What tools would you choose now and why?

Time Management:

From DataWorks experience: This discussion easily filled the entire session. Plan accordingly and be prepared to facilitate deep exploration of tool selection rationale.

Key Learning Points:

Dataset Size Considerations:

- **Small datasets** (hundreds of rows): Spreadsheets often sufficient
- **Medium datasets** (thousands of rows): Either tool may work, depending on complexity
- **Large datasets** (millions of rows): Programming tools typically necessary

Data Type Considerations:

- **Structured, tabular data**: Both tools work well
- **Unstructured data**: Programming tools often better
- **Mixed data types**: Programming tools provide more flexibility

Standardization and Transformation Needs:

- **Simple transformations**: Spreadsheets may be adequate
- **Complex, multi-step processes**: Programming tools offer better control
- **Reproducible workflows**: Programming tools provide documentation advantages

Discussion Questions:

1. What makes a tool "user-friendly" vs. "powerful"?
2. How do we balance ease of use with functionality needs?
3. When is it worth investing time to learn a new tool?
4. How do collaboration needs influence tool selection?

Reflection & Journal Entry

Assignment:

Practice deciding whether spreadsheets are the right tool for different kinds of potential projects.

Context for Facilitators:

At DataWorks, learners complete or will shortly complete:

- Introduction to Python course
- Data Analysis with Microsoft Excel course

Note: These courses are available on the DataWorks website and provide frame of reference for this exercise.

Reflection Prompts:

- Consider a recent or upcoming data project
- Evaluate tool options based on:
    - Dataset characteristics
    - Required transformations
    - Collaboration needs
    - Timeline constraints
    - Skill levels of team members
- Justify your tool selection with specific reasoning

Learning Outcomes:

- Practice systematic tool evaluation
- Apply decision-making framework to real scenarios
- Develop confidence in tool selection rationale

Optional Extension:

If learners have varying levels of experience with different tools, consider pairing them to discuss their reasoning and learn from each other's perspectives.