Week 5 Discussion

Made by: Sirui Tao, Samuel Lee, Parth Sandeep

Today’s Content

- Paper 1: Tidy Data
- Paper 2: Data organization in spreadsheets
- Q&A

Reading 1 - Tidy Data

Tidy Data by Hadley Wickham

- 80% of time are spent on data cleaning and preparation
- Tidy data: Easier analysis
- Rows and columns
- Qualitative (String,…) v.s. Quantitative (int, float, …) values

Principles of Tidy Data

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

country	year	rate
Afghanistan	1999	**745** / 19987071
Afghanistan	2000	**2666** / 20595360
Brazil	1999	**37737** / 172006362
Brazil	2000	**80488** / 174504898
China	1999	**212258** / 1272915272
China	2000	**213766** / 1280428583

table3

country	year	cases	population
Afghanistan	1999	**745**	19987071
Afghanistan	2000	**2666**	20595360
Brazil	1999	**37737**	172006362
Brazil	2000	**80488**	174504898
China	1999	**212258**	1272915272
China	2000	**213766**	1280428583

Source: R for Data Science by Hadley Wickham

Common Mistakes

1. Column headers are values, not variable names.
2. Multiple variables are stored in one column.
3. Variables are stored in both rows and columns.
4. Multiple types of observational units are stored in the same table.
5. A single observational unit is stored in multiple tables.

Practice

Why this dataset is messy? How can you improve it?

Order ID	Category	Amount
CA-2011-167199	Binders | Art | Phones | Fasteners | Paper	609.98 | 5.48 | 391.98 | 755.96 | 31.12
CA-2011-149020	Office Supplies | Furniture	2.98 | 51.94
CA-2011-131905	Office Supplies | Technology | Technology	7.2 | 42.0186 | 42.035
CA-2011-127614	Accessories | Tables | Binders	234.45 | 1256.22 | 17.46

Cleaned

Order ID	Category	Amount
CA-2011-167199	Binders	609.98
CA-2011-167199	Art	5.48
CA-2011-167199	Phones	391.98
CA-2011-167199	Fasteners	755.96
CA-2011-167199	Paper	31.12
CA-2011-149020	Office Supplies	2.98
CA-2011-149020	Furniture	51.94
CA-2011-131905	Office Supplies	7.2
CA-2011-131905	Technology	42.0186
CA-2011-131905	Technology	42.035
CA-2011-127614	Accessories	234.45
CA-2011-127614	Tables	1256.22
CA-2011-127614	Binders	17.46

Discussion questions

1. What are some common challenges you might face when organizing data for analysis, and how can tidy data principles help address these challenges?

2. Can you think of a situation where tidy data principles might not be the best approach for organizing data?

Question 1

What is one rule of 'tidy data' you found most interesting, and why?

Reading 2 - Data Organization In Spreadsheets

Data Organization In Spreadsheets

This guide outlines key practices for organizing spreadsheet data effectively to minimize errors and improve analysis.

- Formatting
- Consistency
- Data structure
- Storage recommendations

Common Spreadsheet Mistakes

- Using inconsistent formats and identifiers.
- Leaving cells empty and merging cells.

	A	B	C
1	Date	Assay date	Weight
2		12/9/05	54.9
3		12/9/05	45.3
4	12/6/2005	e	47
5		e	45.7
6		e	52.9
7		1/11/2006	46.1
8		1/11/2006	38.6

Consistency & Formatting

- Maintain consistency across data entries.
- Use YYYY-MM-DD format for dates.
- Avoid empty cells; mark as 'N/A' if needed.
- Ensure each cell contains a single piece of information.

Data Organization

- Organize data in a single rectangular format.
- Create a comprehensive data dictionary.
- Avoid using color or highlighting as data indicators.

Case Studies and Examples

Example 1: A research team standardizes data entry formats across multiple spreadsheets, significantly reducing analysis errors.

Example 2: Implementing a 'N/A' policy for empty cells in a sales database helps in accurately identifying trends and gaps in the sales process.

Case Study: A company's finance department adopts single information per cell rule, enhancing the accuracy of financial reporting and budgeting processes.

Conclusion

- Ensure consistency for data integrity.
- Standardize entries; placeholders for gaps.
- Single data point per cell to improve analysis and decisions.

Data Management & Backup

- Employ consistent naming conventions.
- Implement regular data backup practices.
- Use data validation to prevent entry errors.
- Save data in plain text formats for accessibility.

Discussion questions

1. How can the principles of data organization in spreadsheets prevent common analysis errors?

2. Consider a scenario where poorly organized spreadsheet data led to significant errors in analysis or decision-making. How could these errors have been prevented with better data organization practices?

Question 2

Any feedback? Suggestions for improvement?

Logistics

What is for this week

- Fri, Feb 09 **Quiz 3 due**

What is for next week

- Mon, Feb 12 **Assignment 1 due**
- Fri, Feb 16 **Final Project Part 1 due**

Any questions from class, quiz 2, or final project

- Feel free to just raise your hand or come up and ask in private

- Other people can leave early or discuss with your teammates about the final project if you want to