

Data Organization in Spreadsheets (Broman and Woo, 2017)

1. Be consistent

- Consistent codes for categorical variables
- Consistent fixed code for any missing values
- Consistent variable names
- Consistent subject identifiers
- Consistent data layout in multiple files
- Be careful about extra spaces within cells

2. Choose good names for things

- **Choose short, unique and meaningful names.**
- Avoid spaces, they make programming harder.

3. Write dates as YYYY-MM-DD

4. No empty cells

- **OK to have empty cells for “missing” values.**
- It is common for cells to be left blank when a single value was meant to be repeated multiple times. Please do not do this! It is additional work for the analyst to determine the implicit values for these cells.
- Do not merge cells.

5. Put just one thing in a cell

6. Make it a rectangle

- Single big rectangle with rows corresponding to subjects and columns corresponding to variables.
- The first row should contain variable names.
- **Do not use more than one row for the variable names.**
- If there is not a reasonable way to create a single rectangle, you may need additional sheets.

7. Create a data dictionary

8. No calculations in the raw data file

9. Do not use font color or highlighting as data

10. Make backups

- Keep an original version in a secure location!

11. Use data validation to avoid errors