The Data Sharing Champion

Tabular Data

A Guide for Formatting and Sharing Tabular Data

The ASAP Open Science Team
Dana Lewis, PhD
Matthew Lewis, PhD
Devin Snyder, PhD
Robert Thibault, PhD

Formatting & Sharing Tabular Data

The purpose of this document is to provide guidance on how to properly format and share tabular data to increase accessibility and interoperability.

The ASAP Open Science Policy requires that the data underlying all results reported in a manuscript be deposited in a publicly accessible repository. This must be done no later than time of publication and the dataset must be cited in the publication with a persistent identifier.

To assist ASAP grantees in sharing usable tabular data and to ensure that they meet the ASAP data-sharing standards outlined in this document, **the Open Science Team will review all ASAP-affiliated Zenodo dataset uploads.**

If you have additional questions, email the Open Science Team at **openscience@parkinsonsroadmap.org** with the subject title: "Formatting Tabular Datasets Question:...".

Table of Contents

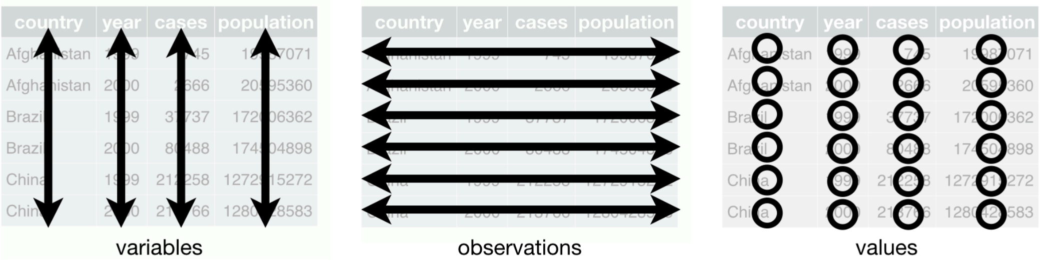
What is tabular data?

Tabular data is any data organized in rows and columns in a table-like format. Tabular data is often **source or analysis-ready data,** data that is used to produce all results presented in a manuscript (e.g., figure panels, tables, graphs, and numbers). This includes image quantification data.

Tabular data that only includes summary data (e.g., means and standard deviations) can be shared, but are not sufficient to meet the ASAP Open Science Policy requirements.

How should I format my tabular data?

Tabular data can be organized and represented in a variety of ways, however, to increase reusability and interoperability, **tabular data should be shared in 'tidy' format.**



from R for Data Science

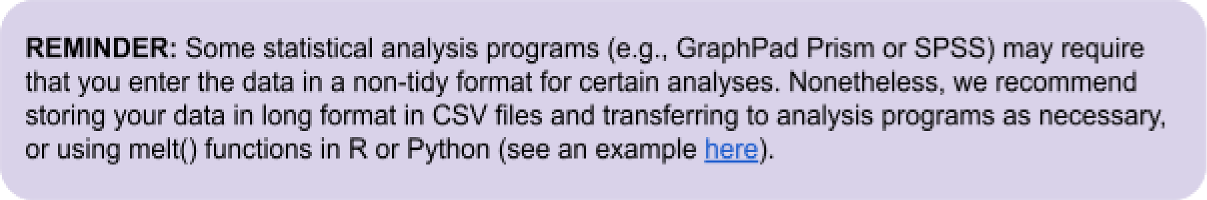
There are **three rules** to follow when making a dataset tidy:

1. **Each row is a single observation**
   - The first row (i.e., header row) contains variable names
   - Do not use more than one row for a header to improve machine-readability

2. **Each column is a single variable**
   - All measurements in the same column must use the same units of measurement.
   - There should be one column for each type of information; different columns should not contain measurements of the same variable.

3. **Each cell contains a single value**

  - Cells should not contain units
    - Units can be in the header row or in the readme file

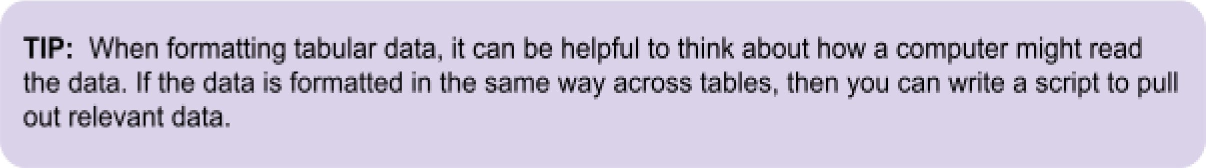


REMINDER: Some statistical analysis programs (e.g., GraphPad Prism or SPSS) may require that you enter the data in a non-tidy format for certain analyses. Nonetheless, we recommend storing your data in long format in CSV files and transferring to analysis programs as necessary, or using melt() functions in R or Python (see an example here).

Tabular Data Best Practices

Formatting and Organization

- Do **not** use formatting (e.g., merged cells, colors, or highlighting) to convey information.
- Each table should be its own file or sheet within a file. Do not include multiple tables in the same sheet.
- Files should be saved in a standard, non-proprietary format, like CSV.
- Files should include only data values; no calculations or graphs should be included.
  - Summary data (e.g., means and standard deviations) can be shared but should not be included in the same table as the individual data points.



TIP: When formatting tabular data, it can be helpful to think about how a computer might read the data. If the data is formatted in the same way across tables, then you can write a script to pull out relevant data.

Quality Control

- Perform a set of quality control checks on your spreadsheet before depositing to a repository. This may include:
  - Checking that Microsoft Excel hasn't changed numbers to dates, dates to numbers, or gene names to dates.
  - Check that zeros and missing values are correctly notated.
  - Check summary statistics (including min and max values) for any obvious errors

Naming Conventions

- Use consistent naming conventions for variables and subject IDs across files.
  - We recommend utilizing underscores instead of spaces in variable/column names.
- Use YYYY-MM-DD format for dates.
- Carefully consider how to approach missing data.
  - Blank cells, zeros, and NA are all treated differently by programs like R, Python, SQL, and Excel.

    - We recommend using a common label for missing data like NA and indicating how missing values are coded in the readme file. Do not leave any cells blank.
- Use descriptive names to name files so that it is easy to keep track of what data they contain. Name files in a consistent manner. Explain this naming convention and the relationship between files in the readme file.

Why does ASAP require tabular data in tidy format?

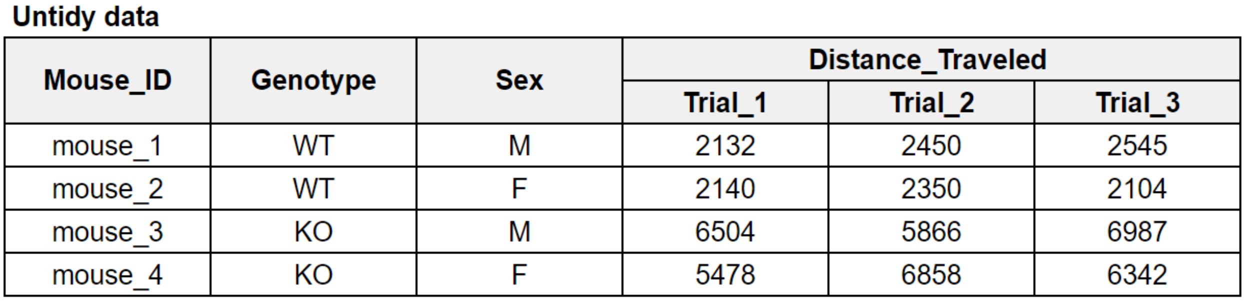
While data sharing is increasingly common across the scientific community, much of the data are poorly formatted and thus not reusable. Therefore, the ASAP Open Science Policy requires that the data underlying all results reported in a manuscript be deposited in a publicly accessible repository no later than time of publication and cited in the publication with their persistent identifier. **This includes source or analysis-ready data, which are often in tabular format.**

Sharing tabular data in 'tidy' format has many benefits including:

- Increased machine-readability and interoperability for subsequent reuse or meta-analysis of datasets
- Provides a consistent and easy workflow for statistical analysis
- Makes errors more detectable (see Spreadsheet Horror Stories)

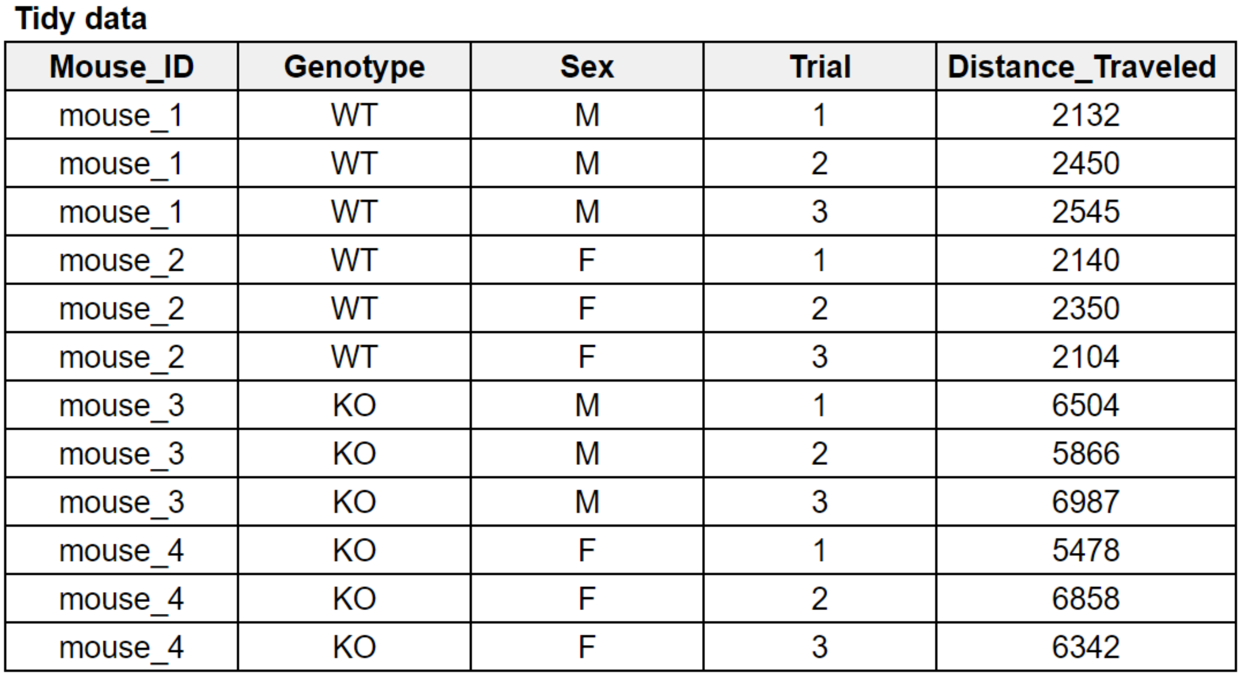
Examples of tidy and untidy data

Data organization: This data is untidy in several ways: it has multiple header rows, merged cells, and each row contains multiple observations. When using repeated measures, each observation should receive its own row (instead of all observations from a single subject listed on the same row). **To make this dataset tidy, use a single header row, and use one row for each observation.**



Untidy data

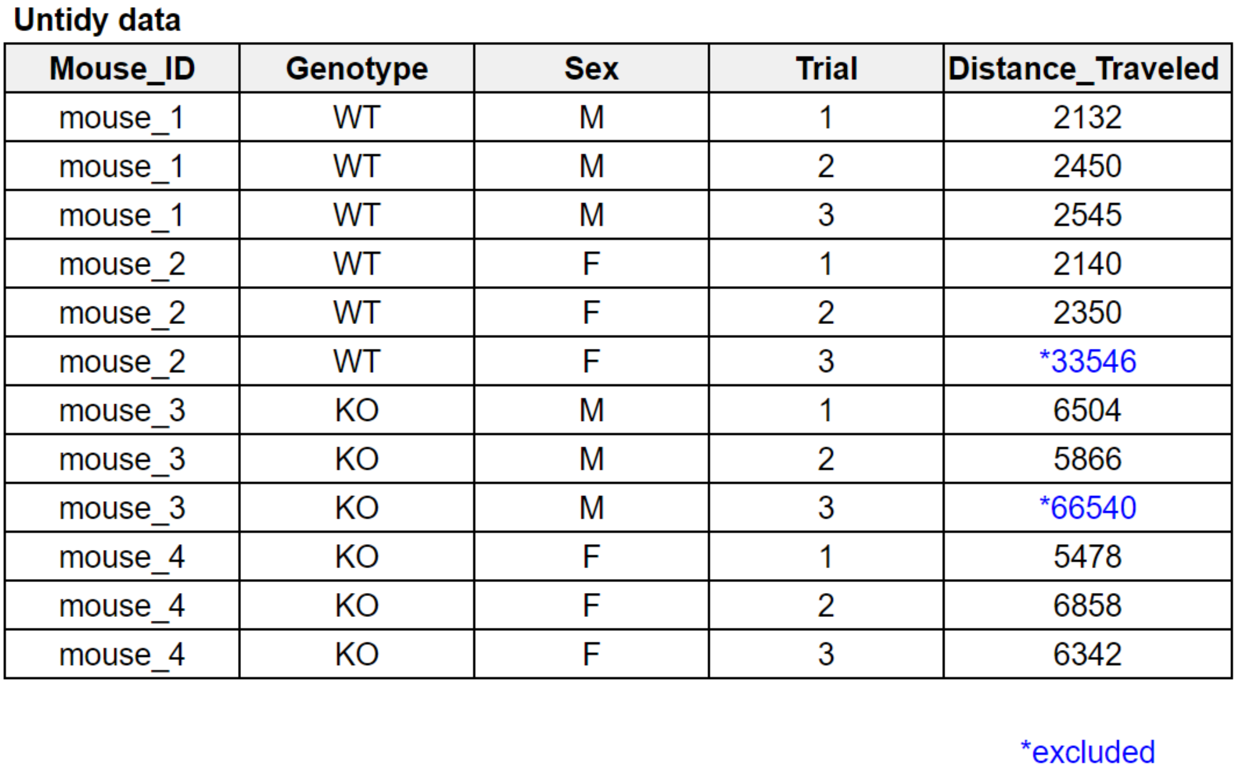
Mouse_ID	Genotype	Sex	Distance_Traveled		
			Trial_1	Trial_2	Trial_3
mouse_1	WT	M	2132	2450	2545
mouse_2	WT	F	2140	2350	2104
mouse_3	KO	M	6504	5866	6987
mouse_4	KO	F	5478	6858	6342



Tidy data

Mouse_ID	Genotype	Sex	Trial	Distance_Traveled
mouse_1	WT	M	1	2132
mouse_1	WT	M	2	2450
mouse_1	WT	M	3	2545
mouse_2	WT	F	1	2140
mouse_2	WT	F	2	2350
mouse_2	WT	F	3	2104
mouse_3	KO	M	1	6504
mouse_3	KO	M	2	5866
mouse_3	KO	M	3	6987
mouse_4	KO	F	1	5478
mouse_4	KO	F	2	6858
mouse_4	KO	F	3	6342

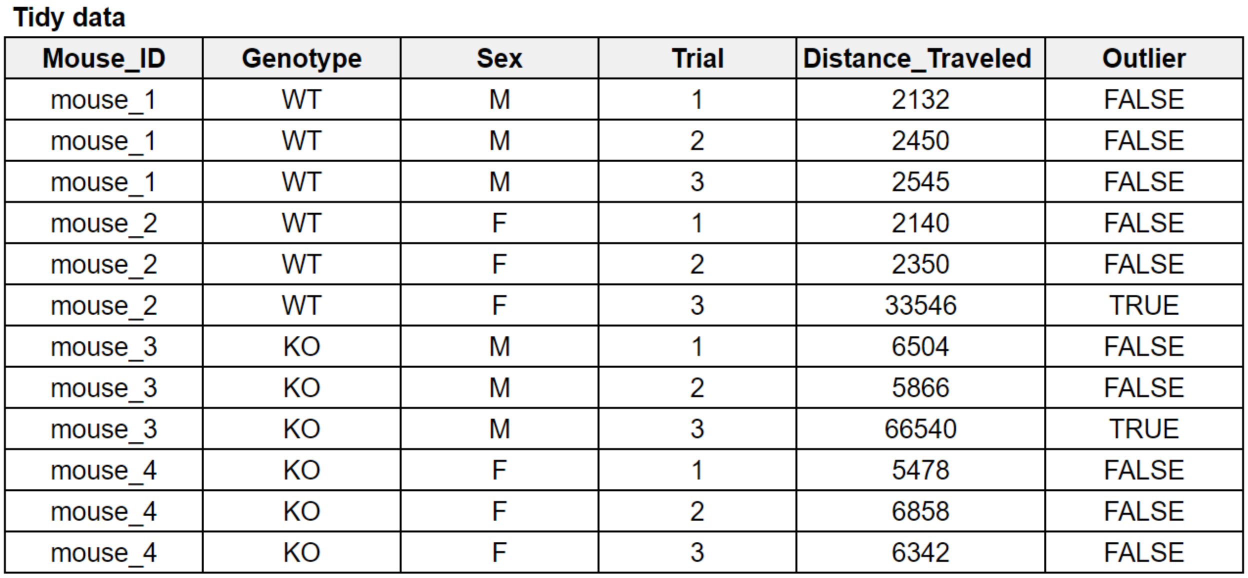
Using formatting and colors: This data is untidy because it uses formatting and colors to indicate data that should be excluded as outliers. **To make this dataset tidy, add a column to indicate the data points that were excluded.**



Untidy data

Mouse_ID	Genotype	Sex	Trial	Distance_Traveled
mouse_1	WT	M	1	2132
mouse_1	WT	M	2	2450
mouse_1	WT	M	3	2545
mouse_2	WT	F	1	2140
mouse_2	WT	F	2	2350
mouse_2	WT	F	3	*33546
mouse_3	KO	M	1	6504
mouse_3	KO	M	2	5866
mouse_3	KO	M	3	*66540
mouse_4	KO	F	1	5478
mouse_4	KO	F	2	6858
mouse_4	KO	F	3	6342

*excluded

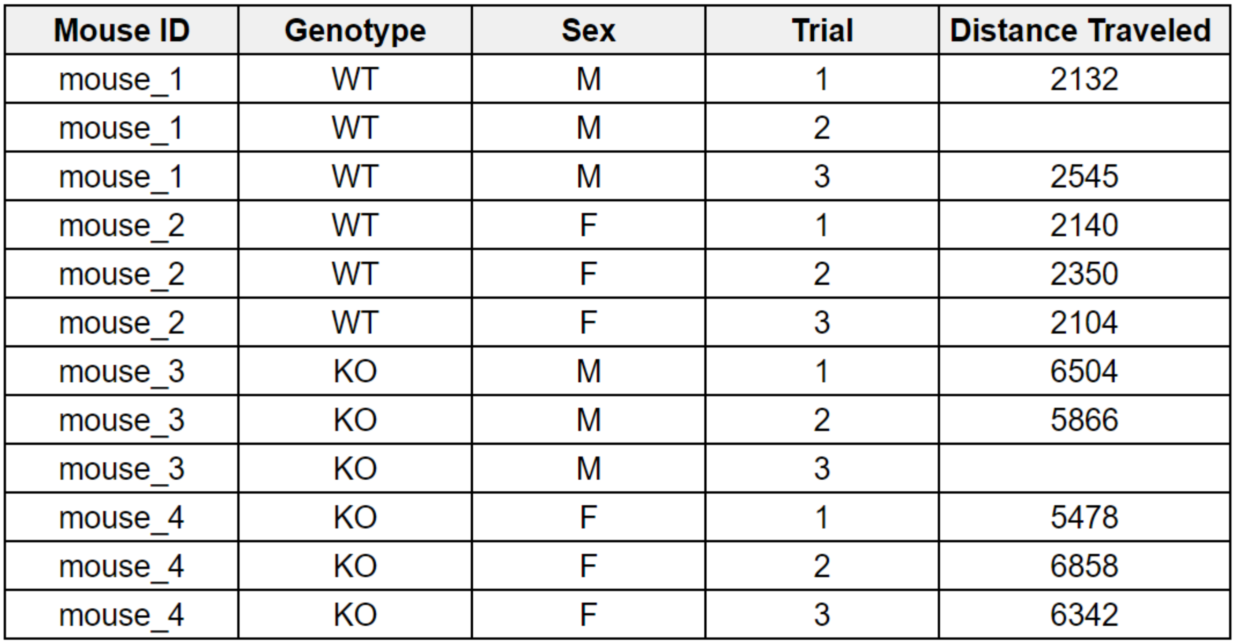


Tidy data

Mouse_ID	Genotype	Sex	Trial	Distance_Traveled	Outlier
mouse_1	WT	M	1	2132	FALSE
mouse_1	WT	M	2	2450	FALSE
mouse_1	WT	M	3	2545	FALSE
mouse_2	WT	F	1	2140	FALSE
mouse_2	WT	F	2	2350	FALSE
mouse_2	WT	F	3	33546	TRUE
mouse_3	KO	M	1	6504	FALSE
mouse_3	KO	M	2	5866	FALSE
mouse_3	KO	M	3	66540	TRUE
mouse_4	KO	F	1	5478	FALSE
mouse_4	KO	F	2	6858	FALSE
mouse_4	KO	F	3	6342	FALSE

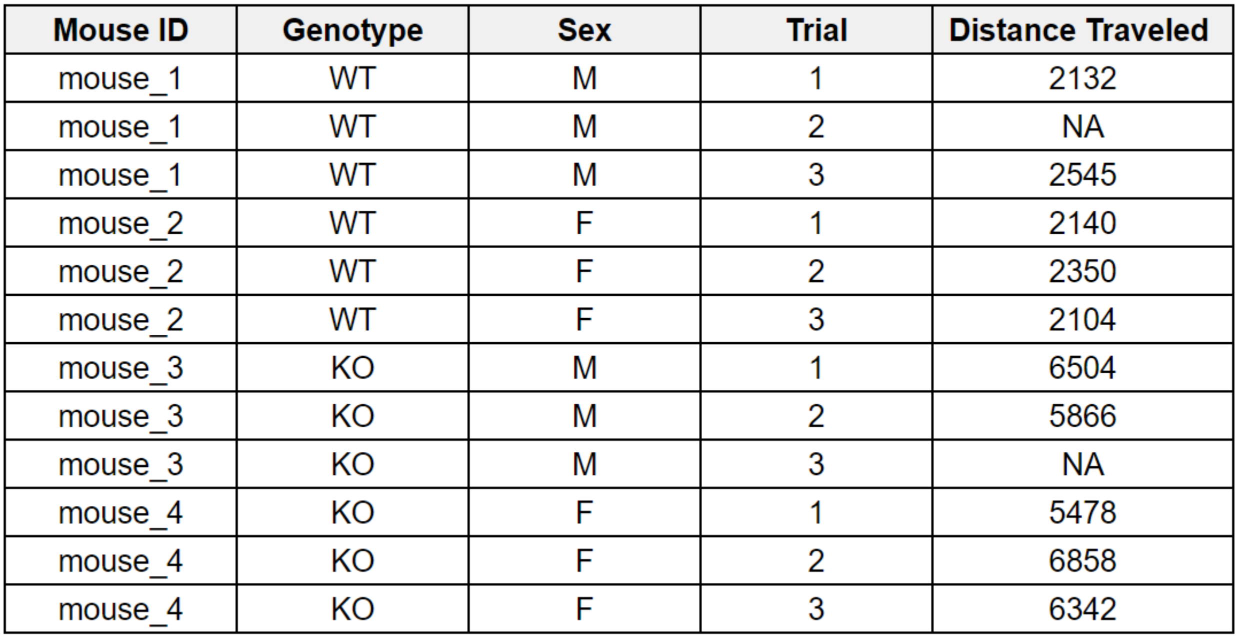
Blank or missing data points: This data is untidy because it has blank cells. They should be filled in with a consistent code, such as NA. **To make this dataset tidy, use a code like NA to represent missing data.**

Untidy data



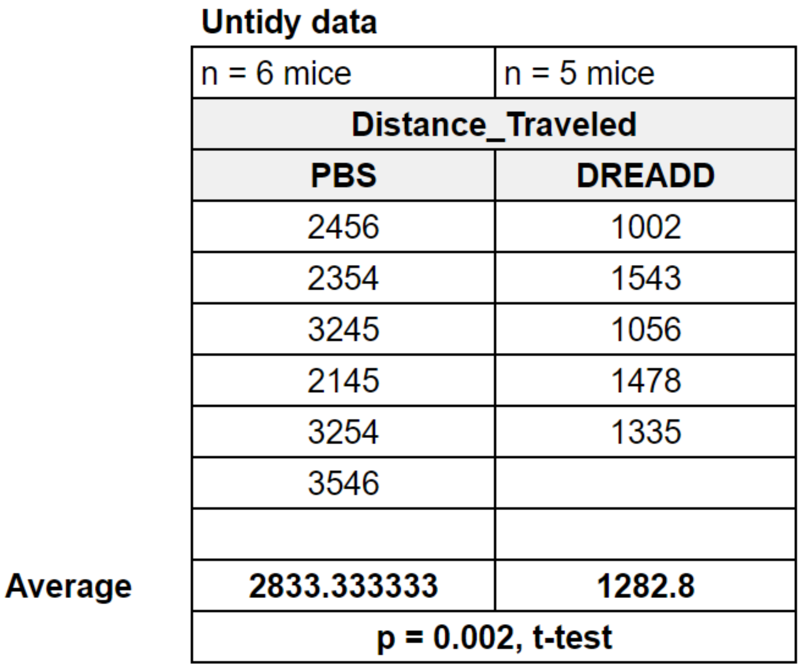
Mouse ID	Genotype	Sex	Trial	Distance Traveled
mouse_1	WT	M	1	2132
mouse_1	WT	M	2	
mouse_1	WT	M	3	2545
mouse_2	WT	F	1	2140
mouse_2	WT	F	2	2350
mouse_2	WT	F	3	2104
mouse_3	KO	M	1	6504
mouse_3	KO	M	2	5866
mouse_3	KO	M	3	
mouse_4	KO	F	1	5478
mouse_4	KO	F	2	6858
mouse_4	KO	F	3	6342

Tidy data



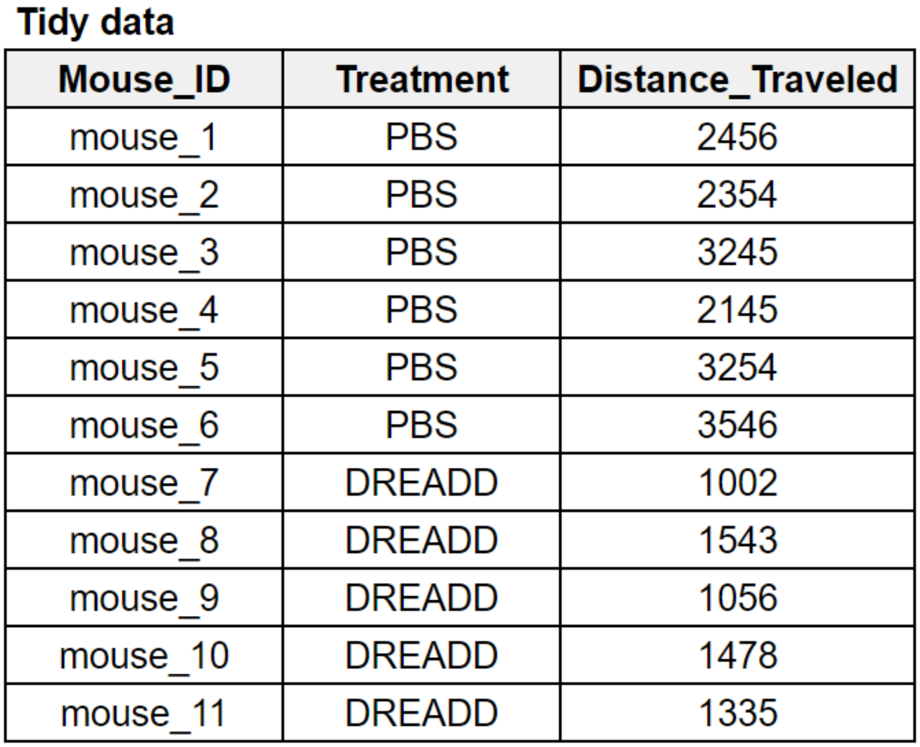
Mouse ID	Genotype	Sex	Trial	Distance Traveled
mouse_1	WT	M	1	2132
mouse_1	WT	M	2	NA
mouse_1	WT	M	3	2545
mouse_2	WT	F	1	2140
mouse_2	WT	F	2	2350
mouse_2	WT	F	3	2104
mouse_3	KO	M	1	6504
mouse_3	KO	M	2	5866
mouse_3	KO	M	3	NA
mouse_4	KO	F	1	5478
mouse_4	KO	F	2	6858
mouse_4	KO	F	3	6342

Summary Statistics: This data is untidy in several ways: it includes average values, statistics, and Ns in the table, it uses merged headers, and each row includes data from multiple subjects. The average values, statistics, and Ns should be included in the Results section or Figure legends of a manuscript. **To make this dataset tidy, remove the summary statistics and Ns, use a single header row, and use one row for each observation.**



Untidy data

	n = 6 mice	n = 5 mice
	Distance_Traveled	
	PBS	**DREADD**
	2456	1002
	2354	1543
	3245	1056
	2145	1478
	3254	1335
	3546	
Average	**2833.333333**	**1282.8**
	p = 0.002, t-test	



Tidy data

Mouse_ID	Treatment	Distance_Traveled
mouse_1	PBS	2456
mouse_2	PBS	2354
mouse_3	PBS	3245
mouse_4	PBS	2145
mouse_5	PBS	3254
mouse_6	PBS	3546
mouse_7	DREADD	1002
mouse_8	DREADD	1543
mouse_9	DREADD	1056
mouse_10	DREADD	1478
mouse_11	DREADD	1335