# Tidy Data

Getting and Cleaning Data
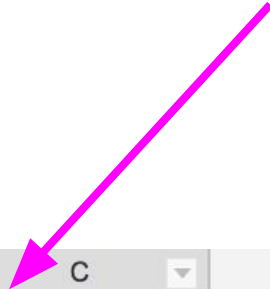
# ROW

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

# COLUMN

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

There are 7 different **variables** in this spreadsheet.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

For each variable, we see there are 4 different **observations**.

|   | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

For each variable and observation, there are **values**.

**Demographic Survey Data**

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

Two different **types** of data, contained in two different spreadsheets

**Doctor's Office Measurements Data**

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Height_inches | Weight_lbs | Insulin | Glucose |
| 2 | 1004 | Smith | Jane | 65 | 180 | 0.60 | 163 |
| 3 | 4587 | Nayef | Mohammed | 75 | 215 | 1.46 | 150 |
| 4 | 1727 | Doe | Janice | 62 | 124 | 0.72 | 177 |
| 5 | 6879 | Jordan | Alex | 77 | 160 | 1.23 | 205 |

# Wide data

| LastName | Sex | City | Occupation |
|----------|-----|------|------------|
| Smith | F | Frederick | Welder |
| Nayef | M | Upper Darby | Nurse |
| Doe | F | San Diego | Doctor |
| Jordan | M | Birmingham | Teacher |

In **wide format**, each observation is a row, each variable is a column, and each cell contains a value.

In **long format**, each observation has multiple rows, one for each variable measured. In this format, the values are all in the same column, one per row.

# Long data

| LastName (Observation) | Variable | Value |
|------------------------|----------|-------|
| Smith | Sex | F |
| Nayef | Sex | M |
| Doe | Sex | F |
| Jordan | Sex | M |
| Smith | City | Frederick |
| Nayef | City | Upper Darby |
| Doe | City | San Diego |
| Jordan | City | Birmingham |
| Smith | Occupation | Welder |
| Nayef | Occupation | Nurse |
| Doe | Occupation | Doctor |
| Jordan | Occupation | Teacher |

1. Each variable you measure should be in a
   **single column**

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 100... | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 458... | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 172... | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 687... | Jordan | Alex | male | Birmingham | AL | Teacher |

# 2. Every observation of a variable should be in a **different row**

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

# 3. There should be one spreadsheet for each type of data

**Demographic Survey Data**

|   | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

**Doctor's Office Measurements Data**

|   | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Height_inches | Weight_lbs | Insulin | Glucose |
| 2 | 1004 | Smith | Jane | 65 | 180 | 0.60 | 163 |
| 3 | 4587 | Nayef | Mohammed | 75 | 215 | 1.46 | 150 |
| 4 | 1727 | Doe | Janice | 62 | 124 | 0.72 | 177 |
| 5 | 6879 | Jordan | Alex | 77 | 160 | 1.23 | 205 |

# 4. If you have multiple spreadsheets, they should include a column in each spreadsheet with the same column label that **allows them to be joined or merged**

Demographic Survey Data

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

Doctor's Office Measurements Data

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Height_inches | Weight_lbs | Insulin | Glucose |
| 2 | 1004 | Smith | Jane | 65 | 180 | 0.60 | 163 |
| 3 | 4587 | Nayef | Mohammed | 75 | 215 | 1.46 | 150 |
| 4 | 1727 | Doe | Janice | 62 | 124 | 0.72 | 177 |
| 5 | 6879 | Jordan | Alex | 77 | 160 | 1.23 | 205 |

# Rules for Tidy Spreadsheets

1. Be consistent

2. Choose good names for things

3. Write dates as YYYY-MM-DD

4. No empty cells

5. Put just one thing in a cell

6. Don't use font color or highlighting as data

7. Save the data as plain text files

# 1. Be Consistent!

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | ID | LastName | FirstName | Sex | City | State | Occupation |
| 2 | 1004 | Smith | Jane | female | Frederick | MD | Welder |
| 3 | 4587 | Nayef | Mohammed | male | Upper Darby | PA | Nurse |
| 4 | 1727 | Doe | Janice | female | San Diego | CA | Doctor |
| 5 | 6879 | Jordan | Alex | male | Birmingham | AL | Teacher |

Keep exactly the same variable names across spreadsheets.

In these data, sex is always specified as "female" or "male." Pick a way to code your variables and stick to it.

## 2. Choose good names for things

| | **Do this…** | **Not This!** |
|---|---|---|
| Avoid Extra Spaces | 'male' | 'male ' |
| Use underscores not spaces | doctor_visit_v1 | Doctor Visit 1 |
| Choose meaningful names | doctor_visit_v1 | "F1" |

# 3. Write dates as YYYY-MM-DD

|  | **Do this…** | **Not This!** |
|---|---|---|
| Use 'ISO 8601' standard | 2018-02-27 | 2/27 *or* 2_27_2018 *or* Feb 27 |

Broman KW, Woo KH. (2017) Data organization in spreadsheets. *PeerJ Preprints* 5:e3183v1
https://doi.org/10.7287/peerj.preprints.3183v1

# 4. No empty cells



**A**

| | A | B | C |
|---|---|---|---|
| 1 | id | date | glucose |
| 2 | 101 | 2015-06-14 | 149.3 |
| 3 | 102 | | 95.3 |
| 4 | 103 | 2015-06-18 | 97.5 |
| 5 | 104 | | 117.0 |
| 6 | 105 | | 108.0 |
| 7 | 106 | 2015-06-20 | 149.0 |
| 8 | 107 | | 169.4 |

**B**

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | 1 min | | | 5 min | | | | |
| 2 | strain | normal | | mutant | | normal | | mutant | |
| 3 | A | 147 | 139 | 166 | 179 | 334 | 354 | 451 | 474 |
| 4 | B | 246 | 240 | 178 | 172 | 514 | 611 | 412 | 447 |

Broman KW, Woo KH. (2017) Data organization in spreadsheets. *PeerJ Preprints*5:e3183v1 https://doi.org/10.7287/peerj.preprints.3183v1

# 5. Put just one thing in a cell

# 6. Don't use font color or highlighting as data

**A**

|   | A | B | C |
|---|---|---|---|
| 1 | id | date | glucose |
| 2 | 101 | 2015-06-14 | 149.3 |
| 3 | 102 | 2015-06-14 | 95.3 |
| 4 | 103 | 2015-06-18 | 97.5 |
| 5 | 104 | 2015-06-18 | 1.1 |
| 6 | 105 | 2015-06-18 | 108.0 |
| 7 | 106 | 2015-06-20 | 149.0 |
| 8 | 107 | 2015-06-20 | 169.4 |

**B**

|   | A | B | C | D |
|---|---|---|---|---|
| 1 | id | date | glucose | outlier |
| 2 | 101 | 2015-06-14 | 149.3 | FALSE |
| 3 | 102 | 2015-06-14 | 95.3 | FALSE |
| 4 | 103 | 2015-06-18 | 97.5 | FALSE |
| 5 | 104 | 2015-06-18 | 1.1 | TRUE |
| 6 | 105 | 2015-06-18 | 108.0 | FALSE |
| 7 | 106 | 2015-06-20 | 149.0 | FALSE |
| 8 | 107 | 2015-06-20 | 169.4 | FALSE |

| When.. | Be sure to… | So Do this… | Avoid this… | Why? |
|---|---|---|---|---|
| Naming variables (aka assigning column headers) | Use meaningful variable names | `AgeAtDiagnosis` | `ADx` | `ADx` is an unclear and uninformative abbreviation |
| Naming variables | Avoid spacing in column headers | `AgeAtDiagnosis` | `Age At Diagnosis` | Spacing in variable names makes the analyst's life more difficult |
| Naming variables | Use consistent capitalization | `AgeAtDiagnosis` | Using both `AgeAtDiagnosis` and `ageatdiagnosis` | Using consistent column names across tables/spreadsheets simplifies any merging the statistician may have to do. |
| Naming variables | Avoid using separators, but if it's necessary, use an underscore (`_`) | `IGF1` (or `IGF_1`) | `IGF.1`, `IGF-1`, `IGF/1`, `IGF,1` | Separators (commas, periods, hyphens, slashes, spaces etc.) often have different meanings in coding languages than they do in text. Avoiding them avoids error. |
| Coding variables | Avoid unnecessary spaces | 'male' | 'male ' | That extra space after 'male ' makes it different from 'male' without a space. |
| Coding variables | Be consistent! | 'male' | 'Male','male', and 'M', | In the eyes of the statistician, 'Male','male', and 'M' could be incorrectly perceived as three different values. |
| Coding variables | Be careful of spelling errors | 'male' | 'maale' | That extra 'a' makes these two different categories. |
| Coding date and time | Use ISO 8601 coding | 'YYYY-MM-DD' | 'MM/DD/YY` and `Month Day, Year` | Consistency simplifies the analyst's life, and YYYY-MM-DD will not be misconstrued if opened in Excel. |
| Coding missing data | Not leave any cells blank and use a consistent value | 'NA' | '0', '-9', red-highlighted blank cells, '.', '-', … | Each cell should be filled with a consistent value. Pick a way to denote missingness (ideally 'NA') and stick with it. Avoid using numbers or punctuation to denote missing data. |
| Entering data | Stick to text and numbers | Convey all information with direct text/numerical entry | Using cell highlighting or font color to convey information | Your analyst may not use the same platform for analysis as you used for data entry, so avoiding font color and cell highlighting will minimize issues. |
| Generating an Excel file | Save the data in an appropriate format | Use one worksheet per table and save as CSV or text files | Multiple worksheets | Statisticians require this format to import your data onto other platforms. |
| Entering Data | Avoid entering unnecessary lines of text at the start | Start your first row with variable names | Adding lines of text | This violates the rules of tidy data and makes processing more difficult. Include this information in the "Code book" instead. |
| Opening files in Excel | Know and avoid its pitfalls | Consistently include one value per cell and be careful of date and time data. | Using macros, splitting cells, and merging cells | These formats are not amenable to data analysis on other platforms. |

# Tidy data = rectangular data

**A**

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | id | sex | glucose | insulin | triglyc |
| 2 | 101 | Male | 134.1 | 0.60 | 273.4 |
| 3 | 102 | Female | 120.0 | 1.18 | 243.6 |
| 4 | 103 | Male | 124.8 | 1.23 | 297.6 |
| 5 | 104 | Male | 83.1 | 1.16 | 142.4 |
| 6 | 105 | Male | 105.2 | 0.73 | 215.7 |

# Summarizing: Tidy Data

Getting and Cleaning Data