L02 - Tidy Data

Presenter: Olivia Beck
Content Credit: Matthew Beckman, Hadley Wickham

May 17, 2023

Tidy Data

> "Happy families are all alike; every unhappy family is unhappy in its own way." — Leo Tolstoy

> "Tidy datasets are all alike, but every messy dataset is messy in its own way." — Hadley Wickham

- Key ideas:
  - *Cases = Rows*
  - *Variables = Columns*
- How should we define **case**?
- How do we identify **variables**?
- Advantages and Disadvantages

Vocabulary

Variable

- In data science, the word variable has a different meaning than in mathematics.
  - *In algebra, a variable is an unknown quantity.*
  - *In data, a variable is known; it represents a feature that has been measured or observed. "Variable" refers to a specific quantity or quality that can vary from one case to another.*
- Types of variables
  - *quantitative : a number*
  - *categorical (R calls these factors): tells which category or group a case falls into*
  - *all non-numerical values are categorical, but not all numerical values are quantitative*
    - e.g. zip code, IP address, dates

Cases

- Unit of observation or analysis
  - *this is extremly context specific*

What is Tidy Data

- Being neat is **not** what makes data tidy!

There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation/case must have its own row.
3. Each value must have its own cell.

It is your job as the researcher to define the variables, observations, and values.

- The "tidyness" of the data set depends on the research question. It is not an inherent property to the data set itself.
- When data are in tidy form, it's often straightforward to transform the data into arrangements that are useful for answering interesting questions.

Example of Untidy data

	A	B	E	F	G	H	I	J
1			City of Minneapolis Statistics					
2			General Election November 5, 2013					
3	Ward	Precinct	Voters Registering by Absentee	Total Registrations	Voters at Polls	Absentee Voters	Total Ballots Cast	Total Turnout
4	City-Wide Total		708	6,634	75,145	4,954	80,099	33.38%
5								
6	1	1	3	28	492	27	519	27.23%
7	1	2	1	44	836	56	892	31.71%
8	1	3	0	40	905	19	924	38.87%
9	1	4	5	29	768	26	794	36.62%
10	1	5	0	31	683	31	714	37.46%
11	1	6	0	69	739	20	759	32.62%
12	1	7	0	47	291	8	299	15.79%
13	1	8	0	43	415	5	420	30.55%
14	1	9	0	42	596	25	621	25.42%
15	Ward 1 Subtotal		9	373	5,725	217	5,942	30.93%
16								
17	2	1	1	63	1,011	39	1,050	36.42%
18	2	2	5	44	679	37	716	50.39%
19	2	3	4	48	324	18	342	18.88%
20	2	4	0	53	117	3	120	7.34%
21	2	5	2	50	495	26	521	25.49%
22	2	6	1	36	433	19	452	39.10%
23	2	7	0	39	138	7	145	13.78%
24	2	8	1	50	1,206	36	1,242	47.90%
25	2	9	2	39	351	16	367	30.56%
26	2	10	0	87	196	5	201	6.91%
27	Ward 2 Subtotal		16	509	4,950	206	5,156	27.56%
28								
29	3	1	0	52	165	1	166	7.04%

Example of Tidy Data

ward	precinct	registered	voters	absentee	total.turnout
1	1	28	492	27	0.2723
1	4	29	768	26	0.3662
1	7	47	291	8	0.1579
2	1	63	1011	39	0.3642
2	4	53	117	3	0.0734
2	7	39	138	7	0.1378

... and so on for 117 rows altogether.

- Disadvantages
  - *tidy data can be hard for human to quickly interpret*
  - *often not the ideal form for creating graphics*
- Advantages
  - *clear definitions*
  - *tidy data can easily be wrangled to a useful form for interpretation and visualization*

Tidy Data Example

From https://r4ds.had.co.nz/tidy-data.html

You can represent the same underlying data in multiple ways. The example below shows the same data organised in four different ways. Each dataset shows the same values of four variables country, year, population, and cases, but each dataset organises the values in a different way.

Which ones of these is tidy?

Option 1

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
table1
```

```
## # A tibble: 6 × 4
##   country      year  cases population
##   <chr>       <dbl>  <dbl>      <dbl>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

Option 2

```
table2
```

```
## # A tibble: 12 × 4
##    country      year type            count
##    <chr>       <dbl> <chr>           <dbl>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases           37737
##  6 Brazil       1999 population  172006362
##  7 Brazil       2000 cases           80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases          212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases          213766
## 12 China        2000 population 1280428583
```

Option 3

```
table3
```

```
## # A tibble: 6 × 3
##   country      year rate
##   <chr>       <dbl> <chr>
## 1 Afghanistan  1999 745/19987071
## 2 Afghanistan  2000 2666/20595360
## 3 Brazil       1999 37737/172006362
## 4 Brazil       2000 80488/174504898
## 5 China        1999 212258/1272915272
## 6 China        2000 213766/1280428583
```

Option 4

```
table4a
```

```
## # A tibble: 3 × 3
##   country     `1999` `2000`
##   <chr>        <dbl>  <dbl>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

```
table4b
```

```
## # A tibble: 3 × 3
##   country         `1999`     `2000`
##   <chr>            <dbl>      <dbl>
## 1 Afghanistan   19987071   20595360
## 2 Brazil       172006362  174504898
## 3 China       1272915272 1280428583
```

Example Continted

Table 1!

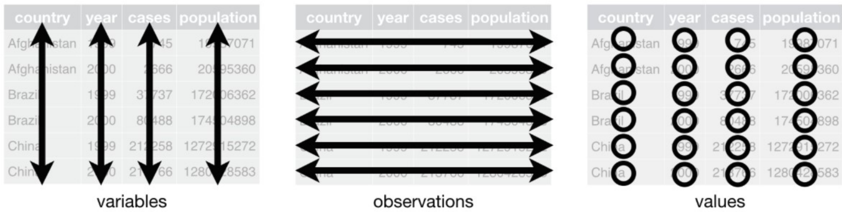


Figure 12.1: Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells.

- Note that all tables contain the same information, just represented differently. Thus, we can transform Tables 2, 3, 4a/4b into Table 1, and vice versa.

Table 2 to Table 1

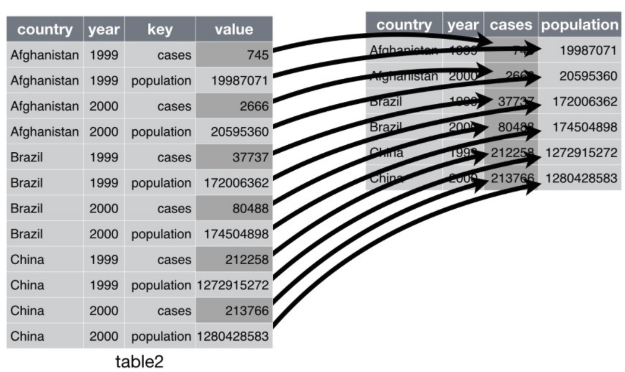


Table 3 to Table 1

Table 4 to Table 1

Make each table tidy individually, then combine the two tables.

(you don't need to be able to interpret this code right now, just look at the end tables along the way.)

```
table4a.temp <-
  table4a %>%
  pivot_longer(!country, names_to = "year", values_to = "cases")

table4b.temp <-
  table4b %>%
  pivot_longer(!country, names_to = "year", values_to = "population")

left_join(table4a.temp, table4b.temp)
```

```
## Joining with `by = join_by(country, year)`
```

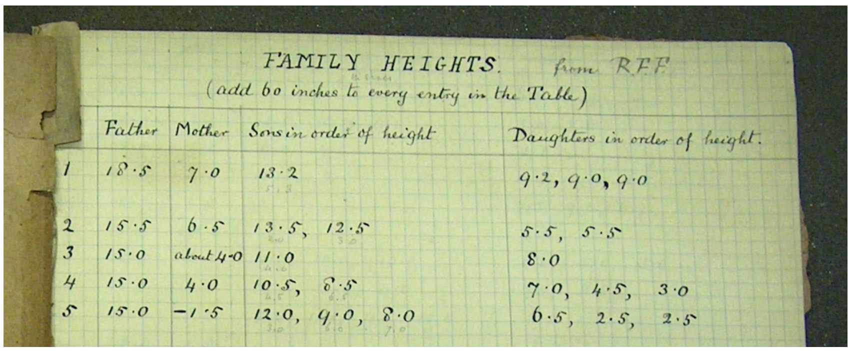
```
## # A tibble: 6 × 4
##   country     year   cases population
##   <chr>       <chr>  <dbl>      <dbl>
## 1 Afghanistan 1999     745   19987071
## 2 Afghanistan 2000    2666   20595360
## 3 Brazil      1999   37737  172006362
## 4 Brazil      2000   80488  174504898
## 5 China       1999  212258 1272915272
## 6 China       2000  213766 1280428583
```

Galton Data

In the 1880s, Francis Galton started to make a mathematical theory of evolution.

Here's part of a page from his lab notebook. Discuss the following in groups:

- What might he investigate with these data (e.g., **Research Question**)?
- Are these data **tidy** according to our definition?
- What are the **cases**?
- What are the **variables**?
- How many **rows** of data should the result have?
- How many **columns** of data should the result have? What is the data type of each column?
- What are some additional variables (not yet shown) that might be of interest? How would you recommend showing that information in the data table?



FAMILY HEIGHTS. from R.F.F.
(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5

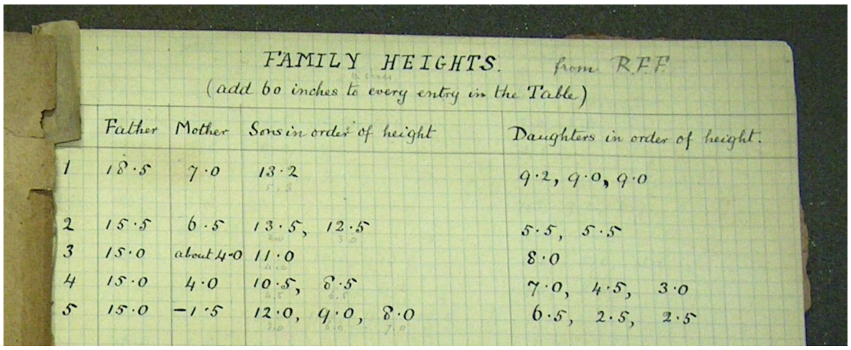
A page from Francis Galton's notebook.

Activity 01: Tidy Data

Work to put these tables in tidy form

- Work with your partner
- As a team, you will put two different data sets into “tidy” form.
- **See Canvas for details**
  - *View-only source data is provided*
  - *use any software you like*
  - *must submit a CSV to Canvas*
  - *do not use spaces in your file names*
- Tip: **Sketch things out together on paper before you do anything in the computer**

Table 1: Galton’s Height measurements data



FAMILY HEIGHTS. from R.F.F.

(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5

A page from Francis Galton’s notebook.

Table 2: Presidents

US Presidents

Name	Start Date	End date	VP
Adams, John	March 4, 1797	March 4, 1801	Thomas Jefferson
James Madison	March 4, 1809	March 4, 1817	George Clinton (03/04/1809 - 04/20/1812) Elbridge Gerry (03/04/1813 - 11/23/1814)
Martin VanBuren	4th March, 1837	4th March 1841	Richard Mentor Johnson
William Henry Harrison	03/04/1841	04/04/1841	John Tyler
John Tyler	April 4th, 1841	March 4th, 1845	Vacant throughout entire presidency

Key:

= Federalist

= Democratic-Rebuplican

= Democrat

= Whig

Code Books

What is a code book?

- A **codebook** describes the contents, structure, and layout of a data collection.
- A well-documented codebook contains information intended to be complete and self-explanatory for each variable in a data file
- https://www.icpsr.umich.edu/web/ICPSR/cms/1983
- Federal Elections Comission
  - *https://www.fec.gov/data/browse-data/?tab=bulk-data*

References

- https://dtkaplan.github.io/DataComputingEbook/chap-tidy-data.html#chap:tidy-data
- https://r4ds.had.co.nz/tidy-data.html
- https://www.icpsr.umich.edu/web/ICPSR/cms/1983