

ISA 401: Business Intelligence & Data Visualization

09: Tidy Data

Fadel M. Megahed, PhD

Associate Professor
Department of Information Systems and Analytics
Farmer School of Business
Miami University

Twitter: [FadelMegahed](#)
GitHub: [fmegahed](#)
Email: fmegahed@miamioh.edu
Office Hours: Automated Scheduler for Virtual Office Hours

Spring 2022

Quick Refresher from Last Week

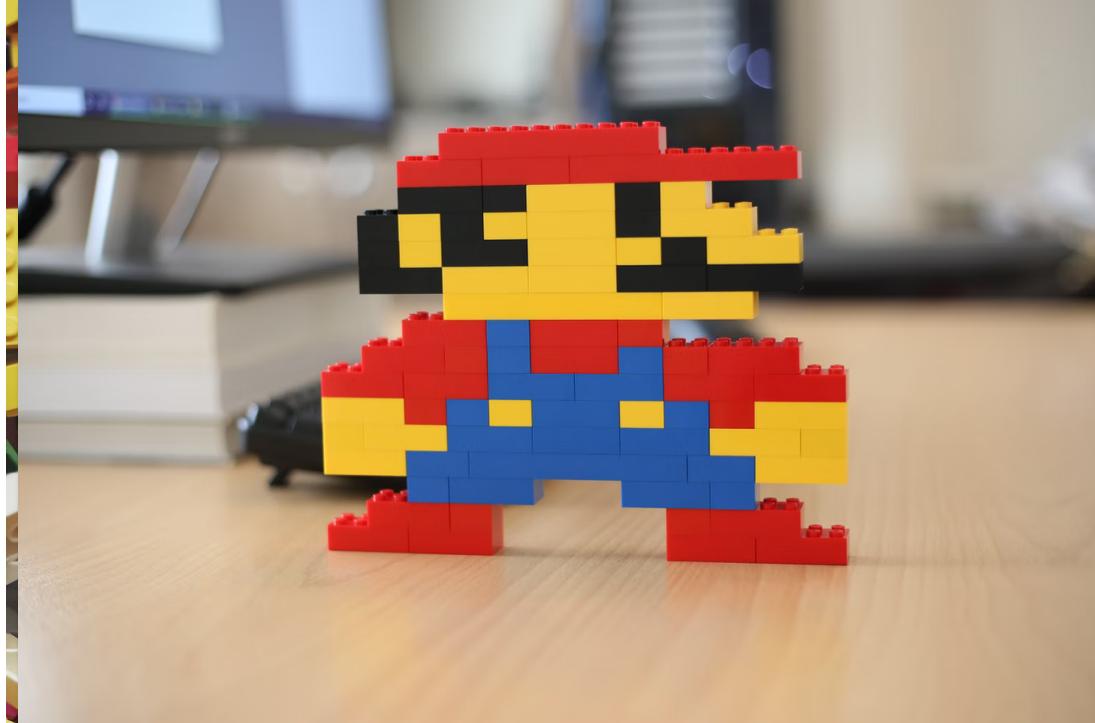
- Describe what is an API
- Download data using APIs

Learning Objectives for Today's Class

- Define tidy data
- Perform pivot and rectangling operations in 

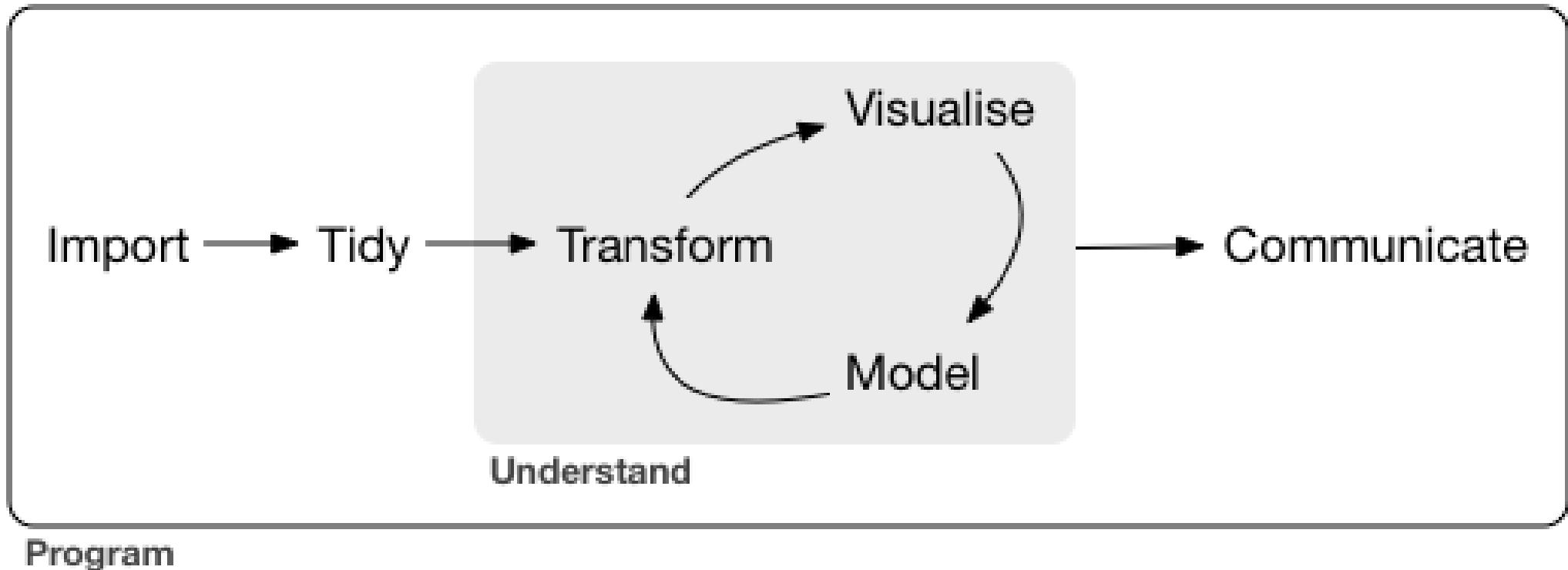
Tidy Data





Sources: All three images are obtained from Upsplash

The R for Data Science Workflow



The Rationale for Tidy Data

- The **tidy framework** provides a **consistent way to organize your data** in .
- Getting your data into this format requires some **upfront work, but that work pays off in the long term.**
- Once you have tidy data and the tidy tools provided by packages in the **tidyverse**, you will spend **much less time munging data from one representation to another, allowing you to spend more time on the analytic questions at hand.**

“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its structure.”

-HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

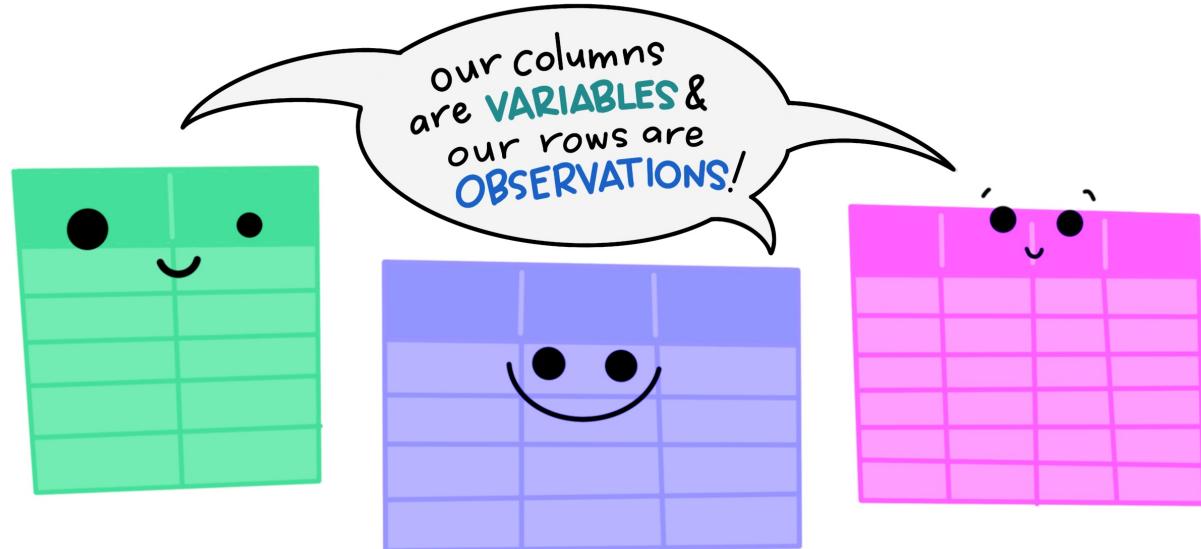
each column a variable

each row an observation

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

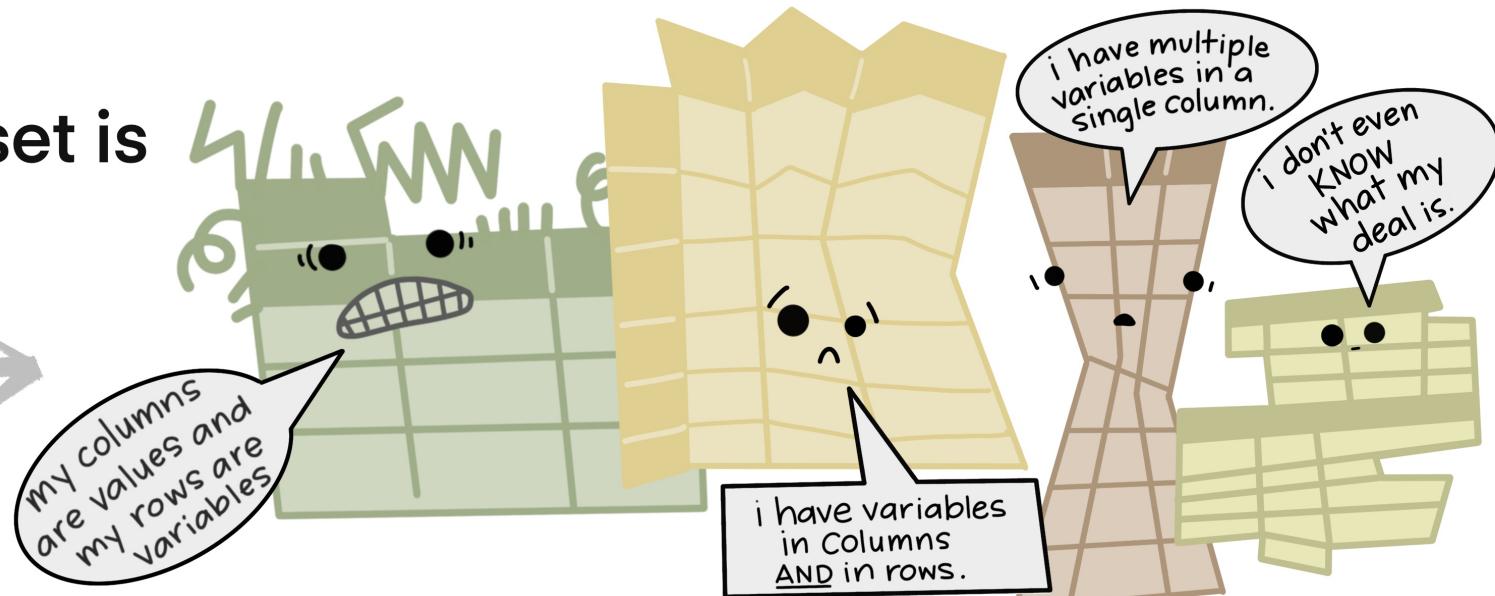
Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

The standard structure of
tidy data means that
“tidy datasets are all alike...”



“...but every messy dataset is
messy in its own way.”

-HADLEY WICKHAM



tidy data \neq clean data

The `movies` data is tidy but not clean.

```
movies <- as_tibble(jsonlite::read_json(  
  "https://vega.github.io/vega-editor/app/data/movies.json",  
  simplifyVector = TRUE))  
  
movies %>%  
  relocate(Release_Date, US_DVD_Sales) %>% # move cols to front  
  slice(37:39, 268:269) %>% # filter specific row numbers  
  print(width = 80) # print nicely
```

```
## # A tibble: 5 x 16  
##   Release_Date US_DVD_Sales Title      US_Gross Worldwide_Gross Production_Budg~  
##   <chr>          <int> <chr>      <int>        <dbl>            <int>  
## 1 9-Mar-94          NA Four Wedd~  52700832     242895809        4500000  
## 2 18-Oct-06          NA 51 Birch ~    84689       84689        350000  
## 3 1963-01-01          NA 55 Days a~ 10000000 100000000       170000000  
## 4 <NA>                NA Drei           0            0        7200000  
## 5 16-Jan-98          NA The Dress    16556      16556        2650000  
## # ... with 10 more variables: MPAA_Rating <chr>, Running_Time_min <int>,  
## #   Distributor <chr>, Source <chr>, Major_Genre <chr>, Creative_Type <chr>,  
## #   Director <chr>, Rotten_Tomatoes_Rating <int>, IMDB_Rating <dbl>,
```

05 : 00

Non-graded Activity: Tidy or Not?

Activity

table1

table2

table3

table4a

table4b

- In the next five panels, there are five tables all displaying the number of TB cases documented by the World Health Organization in Afghanistan, Brazil, and China between 1999 and 2000.
- The data contains values associated with four variables (country, year, cases, and population), but each table organizes the values in a different layout.
- **Based on the information in the previous slide, please document which of the table(s) is(are) tidy and if not, which rules are violated.**
- **Discuss your answer with your neighboring colleague.**

Note that you have a total of five minutes for this non-graded activity.

Non-graded Activity: Tidy or Not?

Activity	table1	table2	table3	table4a	table4b
----------	--------	--------	--------	---------	---------

table1 from the tidyverse package is printed below:

```
## # A tibble: 6 x 4
##   country     year   cases population
##   <chr>      <int>  <int>       <int>
## 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
```

tidy/not-tidy:

data observations? data variables?

rules broken (if any):

Non-graded Activity: Tidy or Not?

Activity

table1

table2

table3

table4a

table4b

```
## table2 from the tidyverse package is printed below:
```

```
## # A tibble: 12 x 4
##   country     year type       count
##   <chr>      <int> <chr>     <int>
## 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
```

tidy/not-tidy: & rules broken (if any):

Non-graded Activity: Tidy or Not?

Activity

table1

table2

table3

table4a

table4b

```
## table3 from the tidyverse package is printed below:
```

```
## # A tibble: 6 x 3
##   country     year    rate
## * <chr>       <int> <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
```

tidy/not-tidy:

data observations? data variables?

rules broken (if any):

Non-graded Activity: Tidy or Not?

Activity

table1

table2

table3

table4a

table4b

```
## table4a from the tidyr package is printed below:
```

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

tidy/not-tidy:

data observations? data variables?

rules broken (if any):

Non-graded Activity: Tidy or Not?

Activity

table1

table2

table3

table4a

table4b

```
## table4b from the tidyr package is printed below:
```

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

tidy/not-tidy:

data observations? data variables?

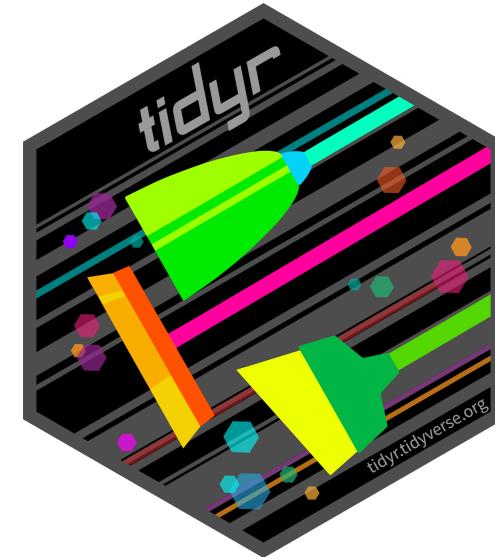
rules broken (if any):

Getting Data into Tidy Format

Key Functions from the `tidyverse`



type	function()	function()
pivoting	<code>pivot_longer()</code>	<code>pivot_wider()</code>
splitting/combining	<code>separate()</code>	<code>unite()</code>
nesting/unnesting	<code>nest()</code>	<code>unnest()</code>
missing	<code>complete()</code>	<code>fill()</code>



Wide Vs Long Data

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

pivot_() to Transform Wide from/to Long

wide

	id	x	y	z
1		a	c	e
2		b	d	f

The pivot_longer() Function

```
pivot_longer(wide,  
            cols = x:z,  
            names_to = "key",  
            values_to = "val")
```

returns

	id	key	val
1	1	x	a
	1	y	c
	1	z	e
2	2	x	b
	2	y	d
	2	z	f

pivot_longer() for table4a [1]

To tidy a dataset like this, we need to pivot the **offending columns into a new pair of variables**.

To describe that operation we need **three parameters**:

- The set of columns whose names are values, not variables. In this example, those are the columns **1999** and **2000**.
- The name of the variable to move the column names to. Here it is **year**.
- The name of the variable to move the column values to. Here it's **cases**.

pivot_longer() for table4a [2]

country	year	cases	country	1999	2000
Afghanistan	1999	745	Afghanistan	745	2666
Afghanistan	2000	2666	Brazil	37737	80488
Brazil	1999	37737	China	212258	213766
Brazil	2000	80488			
China	1999	212258			
China	2000	213766			

table4

The diagram illustrates the process of pivoting the original data frame 'table4a' into a longer, tidier form. It shows two tables side-by-side. The left table, 'table4a', contains three rows for each country (Afghanistan, Brazil, China) across two years (1999 and 2000). The right table, 'table4', is the result of pivoting, where the 'cases' values are now stacked under their respective country and year headers. Arrows indicate the mapping from 'table4a' to 'table4'. The 'table4' table also includes a label 'table4' below it.

Pivoting table4a into a longer, tidy form

pivot_longer() for table4a [3]

```
table4a %>%  
  pivot_longer(c(`1999`, `2000`), names_to = "year", values_to = "cases")
```

```
## # A tibble: 6 x 3  
##   country     year   cases  
##   <chr>       <chr>   <int>  
## 1 Afghanistan 1999     745  
## 2 Afghanistan 2000    2666  
## 3 Brazil      1999  37737  
## 4 Brazil      2000  80488  
## 5 China       1999 212258  
## 6 China       2000 213766
```

The pivot_wider() Function

pivot_wider() for table2 [1]

- `pivot_wider()` is the opposite of `pivot_longer()`.
- You use it when an observation is scattered across multiple rows.
- For example, take `table2`: an observation is a country in a year, but each observation is spread across two rows.

```
head(table2, n = 3)
```

```
## # A tibble: 3 x 4
##   country     year type     count
##   <chr>       <int> <chr>     <int>
## 1 Afghanistan 1999 cases     745
## 2 Afghanistan 1999 population 19987071
## 3 Afghanistan 2000 cases     2666
```

pivot_wider() for table2 [2]

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

table2

pivot_wider() for table2 [3]

```
table2 %>%  
  pivot_wider(names_from = type, values_from = count)
```

```
## # A tibble: 6 x 4  
##   country     year   cases population  
##   <chr>       <int>  <int>      <int>  
## 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
```

seperate() for table3 [1]

```
## # A tibble: 6 x 3
##   country     year    rate
## * <chr>       <int> <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
```

table3 has a different problem:

- we have one column (`rate`) that contains two variables (`cases` and `population`).
- To fix this problem, we'll need the `separate()` function.

seperate() for table3 [2]

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

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

table3

seperate() for table3 [3]

```
table3 %>%  
  separate(rate, into = c("cases", "population"), convert = TRUE)
```

```
## # A tibble: 6 x 4  
##   country     year   cases population  
##   <chr>       <int>   <int>      <int>  
## 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
```

05 : 00

Non-graded Class Activity

Activity

Your Solution

In this five minute non-graded activity, please do the following

- Go to <https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>
- Download the data for **Deaths** by clicking on the tab to the right of the page.
- **Tidy this data based on the information you have learned in today's class.**

05 : 00

Non-graded Class Activity

Activity

Your Solution

In your RStudio Session, please read the data, load the required packages and write the code needed to transform the `deaths` data into a tidy format.

Recap

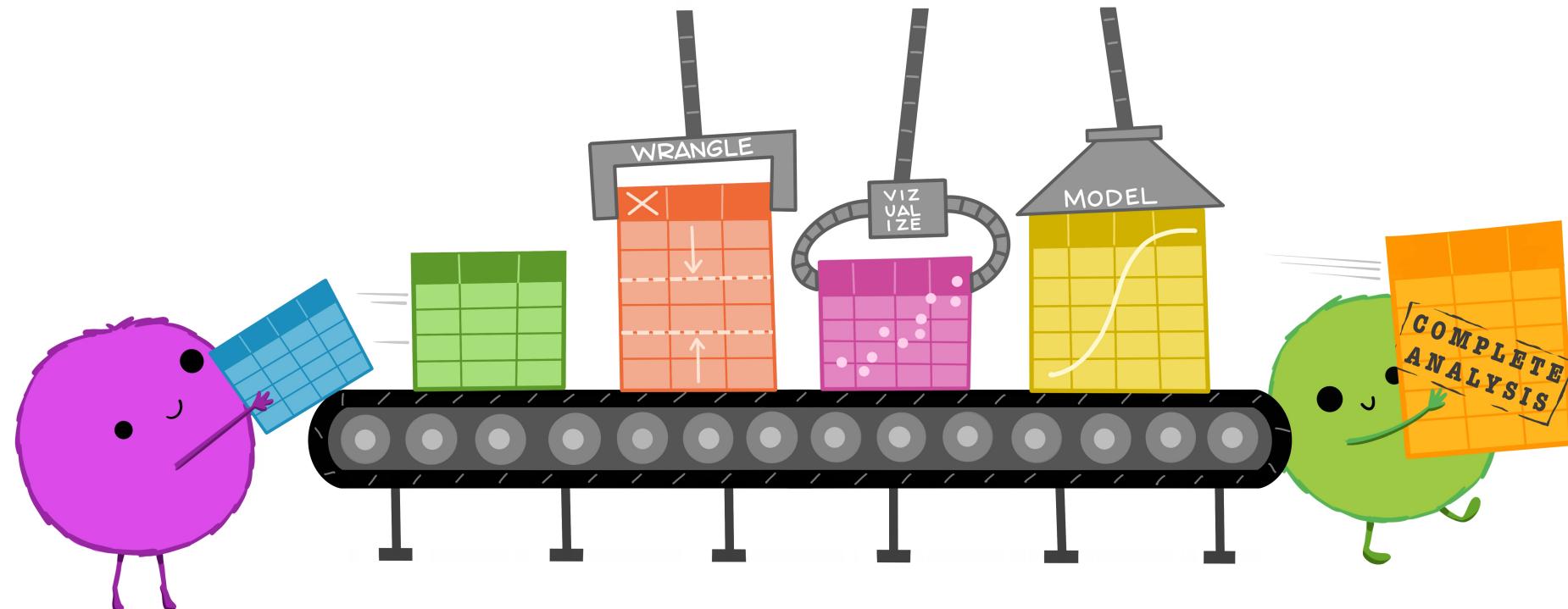
Summary of Main Points

By now, you should be able to do the following:

- Define tidy data
- Perform pivot and rectangling operations in 

Advantages of Tidy Data

- one set of consistent tools for different datasets
- easier for automation and iteration



Supplementary Reading