Data Wrangling

Lab 3

“Data scientists, according to interviews and expert estimates, spend from **50 percent to 80 percent of their time** mired in the mundane labor of collecting and preparing data, before it can be explored for useful information.”

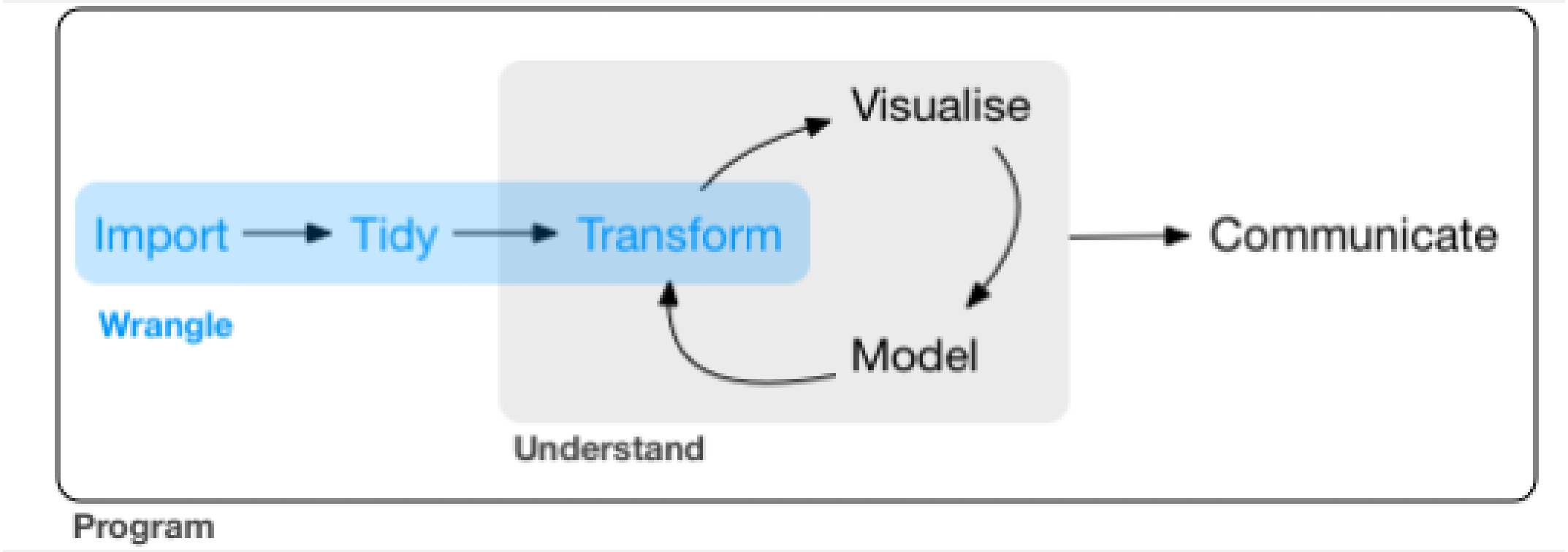
- *New York Times* 2016

Agenda

- Common problems in data wrangling – table contents and formatting
- Long, wide, and unusual data formats
- Thoughts on the tidyverse, tidy data
- Computer memory and “big” data

- Indexing, logical indexing in base R
- Factor/categorical variables

The data analysis pipeline



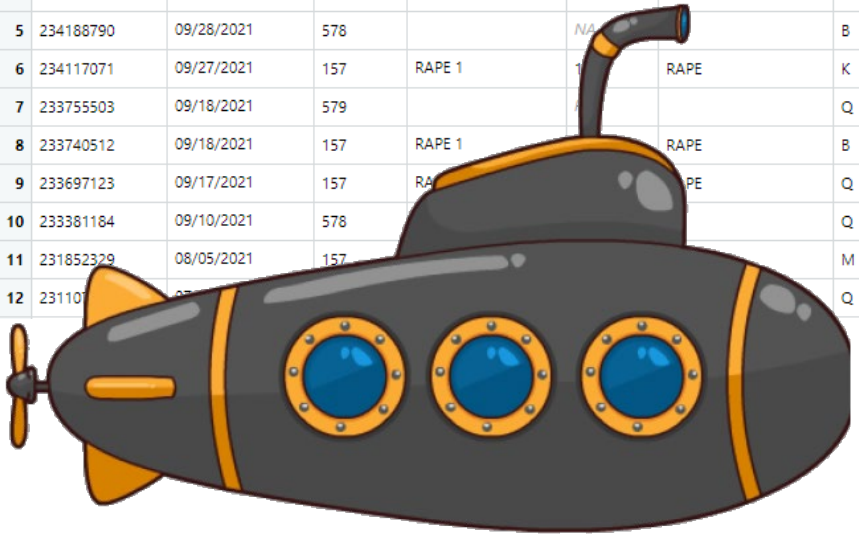
Hadley Wickham's R for Data Science (mostly tidyverse)

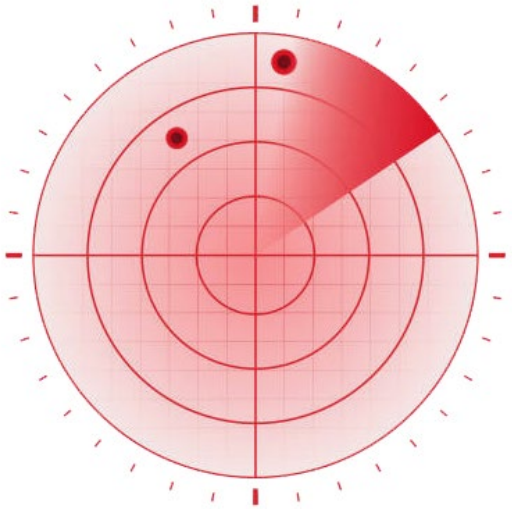
Any issues with this data frame?

	ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	ARREST_BORO	AGE_GROUP
1	238013474	12/18/2021	157	RAPE 1	104	RAPE	Q	18-24
2	236943583	11/25/2021	263	ARSON 2,3,4	114	ARSON	K	25-44
3	234938876	10/14/2021	594	OBSCENITY 1	116	SEX CRIMES	K	25-44
4	234788259	10/11/2021	263	ARSON 2,3,4	114	ARSON	B	18-24
5	234188790	09/28/2021	578		*NA*		B	25-44
6	234117071	09/27/2021	157	RAPE 1	104	RAPE	K	25-44
7	233755503	09/18/2021	579		*NA*		Q	18-24
8	233740512	09/18/2021	157	RAPE 1	104	RAPE	B	25-44
9	233697123	09/17/2021	157	RAPE 1	104	RAPE	Q	25-44
10	233381184	09/10/2021	578		*NA*		Q	25-44
11	231852329	08/05/2021	157	RAPE 1	104	RAPE	M	25-44
12	231107433	07/20/2021	155	RAPE 2	104	RAPE	Q	45-64

Key tools for navigating datasets

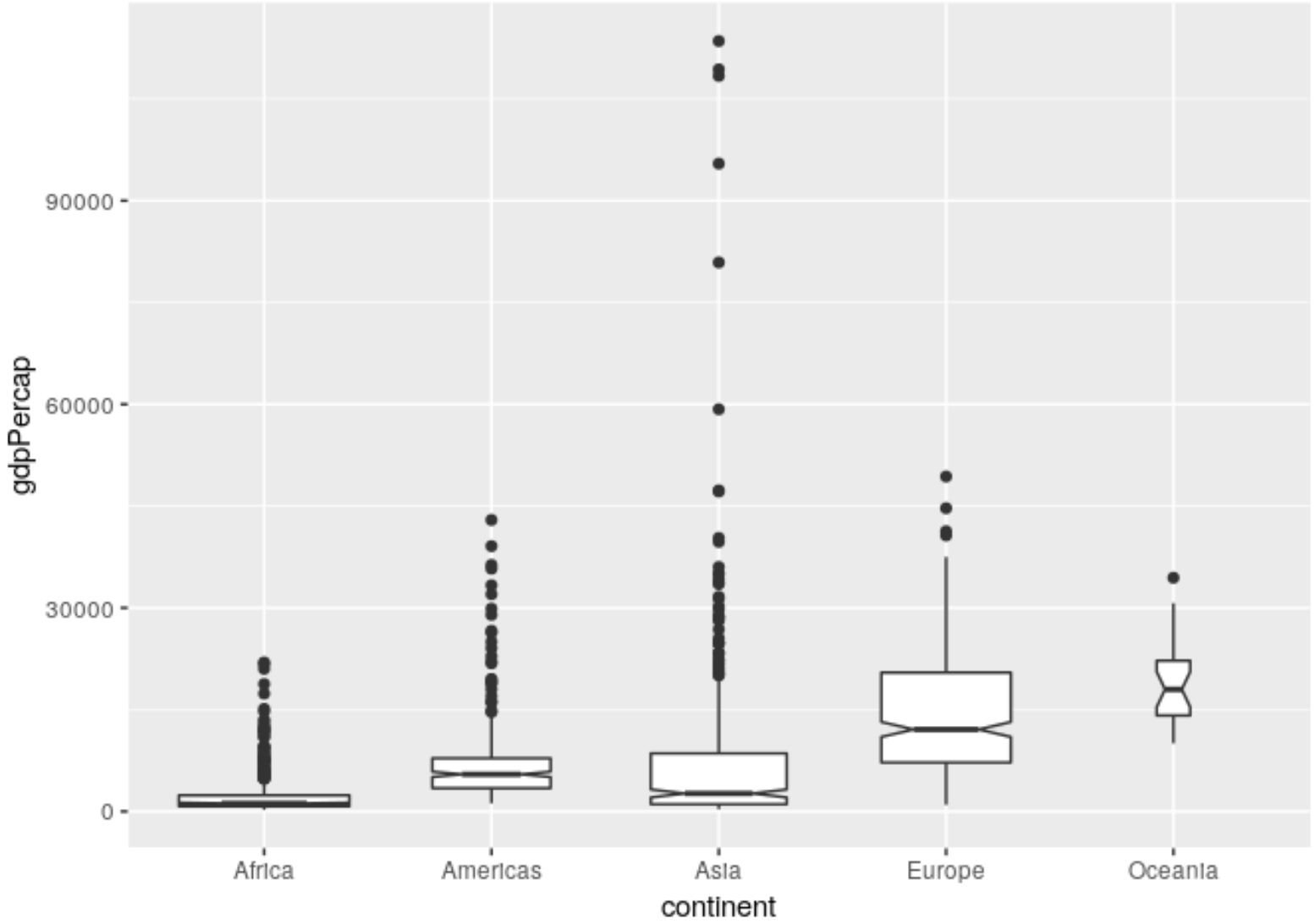
	ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	ARREST_BORO	AGE_GROUP
1	238013474	12/18/2021	157	RAPE 1	104	RAPE	Q	18-24
2	236943583	11/25/2021	263	ARSON 2,3,4	114	ARSON	K	25-44
3	234938876	10/14/2021	594	OBSCENITY 1	116	SEX CRIMES	K	25-44
4	234788259	10/11/2021	263	ARSON 2,3,4	114	ARSON	B	18-24
5	234188790	09/28/2021	578		NA		B	25-44
6	234117071	09/27/2021	157	RAPE 1	1	RAPE	K	25-44
7	233755503	09/18/2021	579				Q	18-24
8	233740512	09/18/2021	157	RAPE 1		RAPE	B	25-44
9	233697123	09/17/2021	157	RA		PE	Q	25-44
10	233381184	09/10/2021	578				Q	25-44
11	231852329	08/05/2021	157				M	25-44
12	23110						Q	45-64



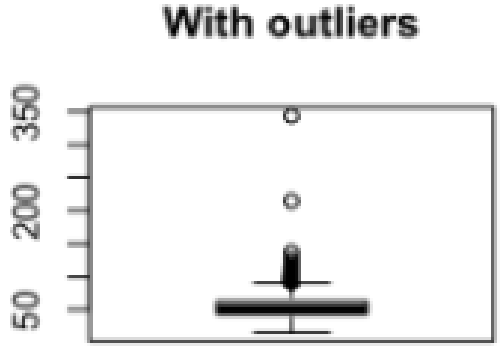


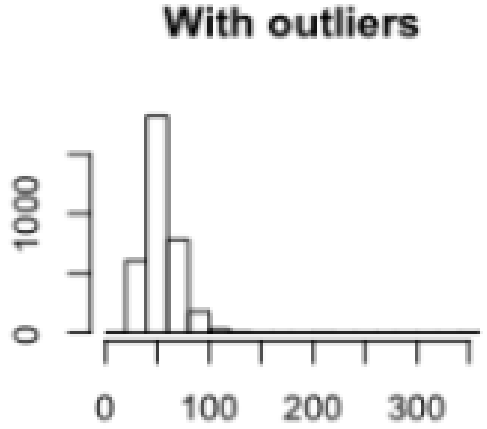
1. Plotting “on-the-fly”
2. Logical queries
3. Error messages

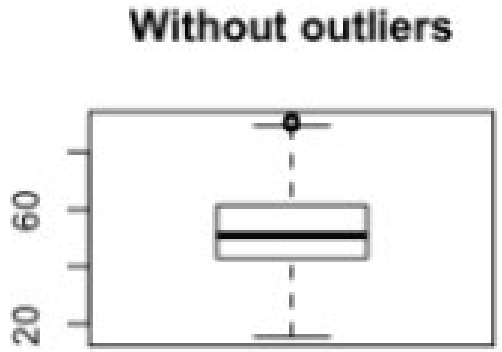
Anything wrong with this plot?



Outliers – what to do?







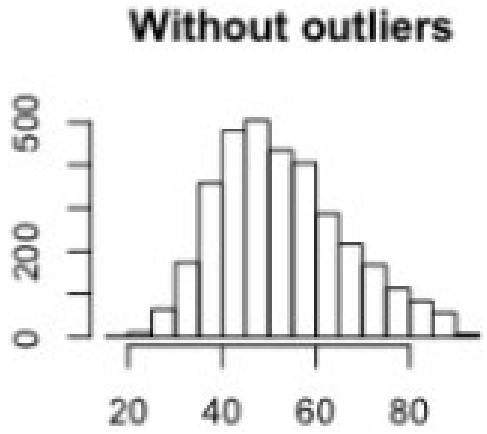


Table formatting

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

cases

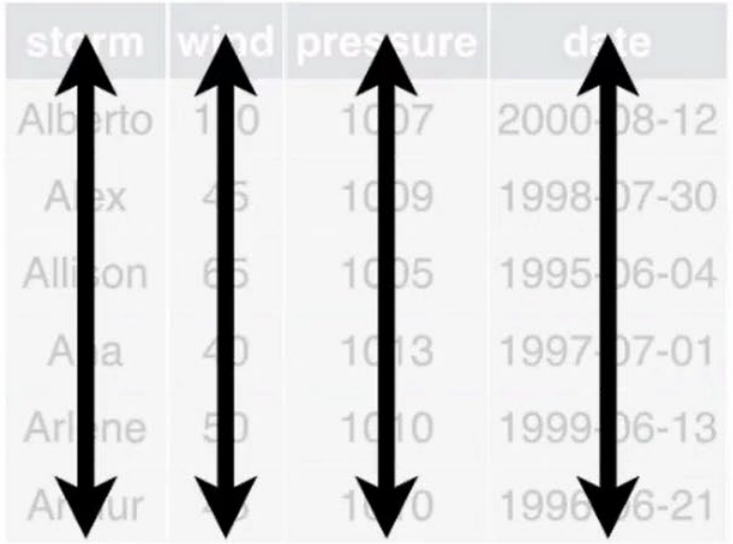
Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

pollution

city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

Table formatting

storms



storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

- Storm name
- Wind Speed (mph)
- Air Pressure
- Date

cases

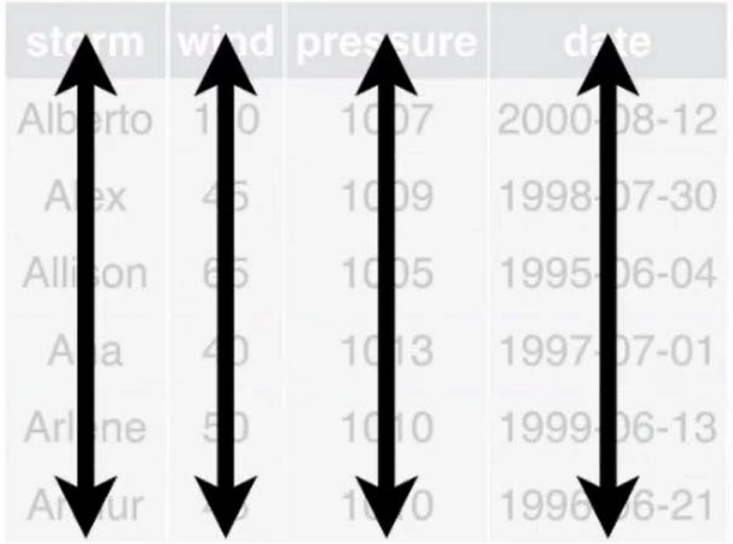
Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

pollution

city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

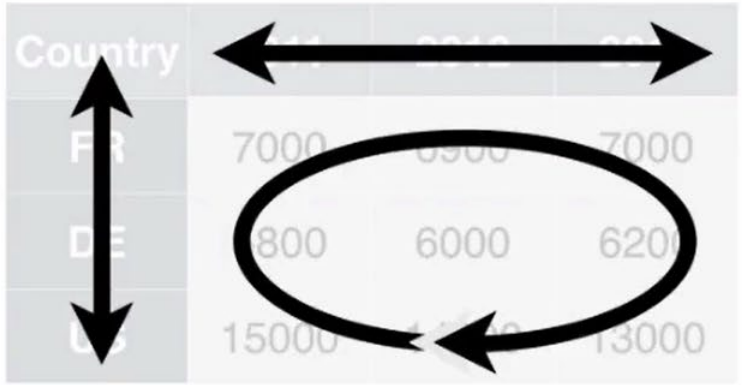
Table formatting

storms



- Storm name
- Wind Speed (mph)
- Air Pressure
- Date

cases



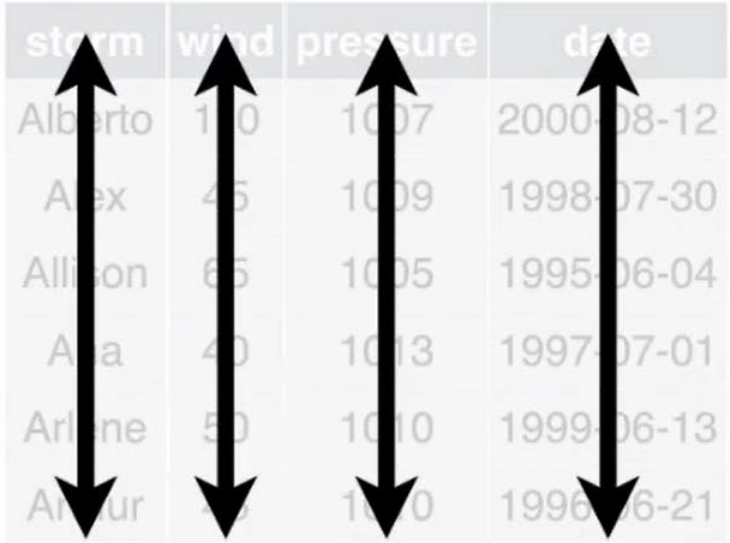
- Country
- Year
- Count

pollution

city	particle size	amount ($\mu g/m^3$)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

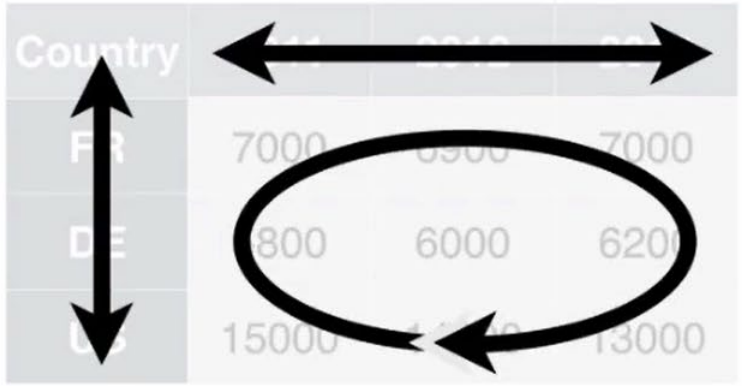
Table formatting

storms



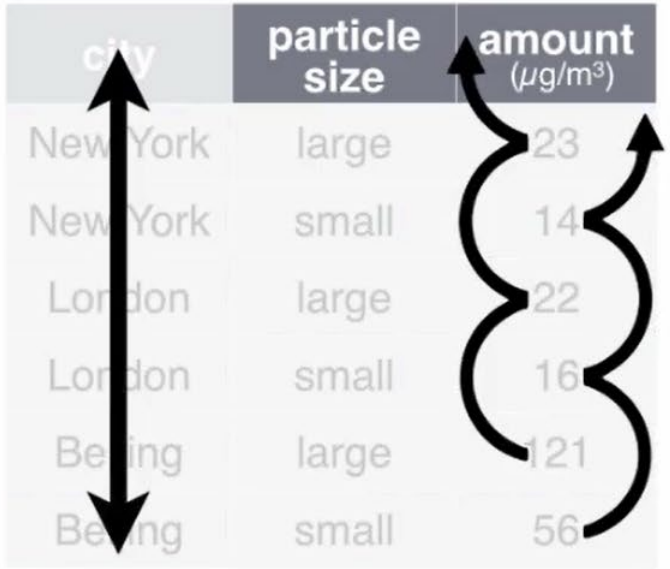
- Storm name
- Wind Speed (mph)
- Air Pressure
- Date

cases



- Country
- Year
- Count

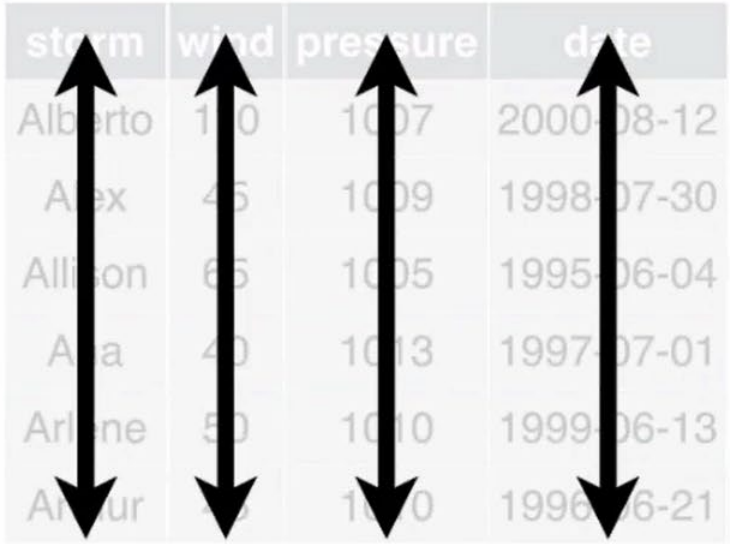
pollution



- City
- Amount of large particles
- Amount of small particles

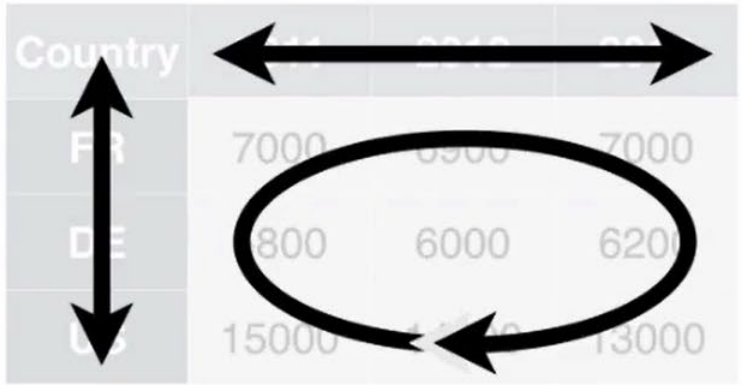
Table formatting

storms



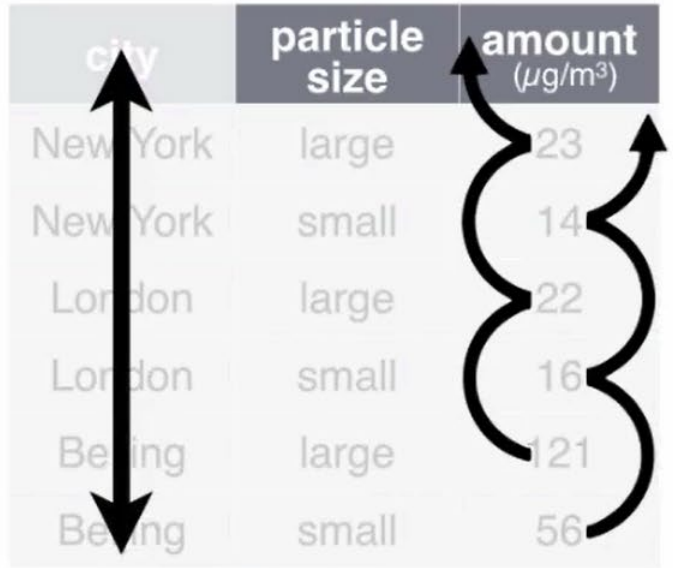
```
storms$storm
storms$wind
storms$pressure
storms$date
```

cases



```
cases$country
names(cases)[-1]
unlist(cases[1:3, 2:4])
```

pollution




```
pollution$city[1,3,5]
pollution$amount[1,3,5]
pollution$amount[2,4,6]
```

Three principles of tidy data

storms

storm	wind	pressure	date
Alberto	110	1007	2000-08-12
Alex	45	1009	1998-07-30
Allison	65	1005	1995-06-04
Ana	40	1013	1997-07-01
Arlene	50	1010	1999-06-13
Arthur	45	1010	1996-06-21

1. Each **variable** is saved in its own **column**
2. Each **observation** is saved in its own **row**
3. Each “type” of observation is stored in a **single table** (e.g. storms)

Enter the tidyverse

R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

Consider [criticisms of tidyverse]

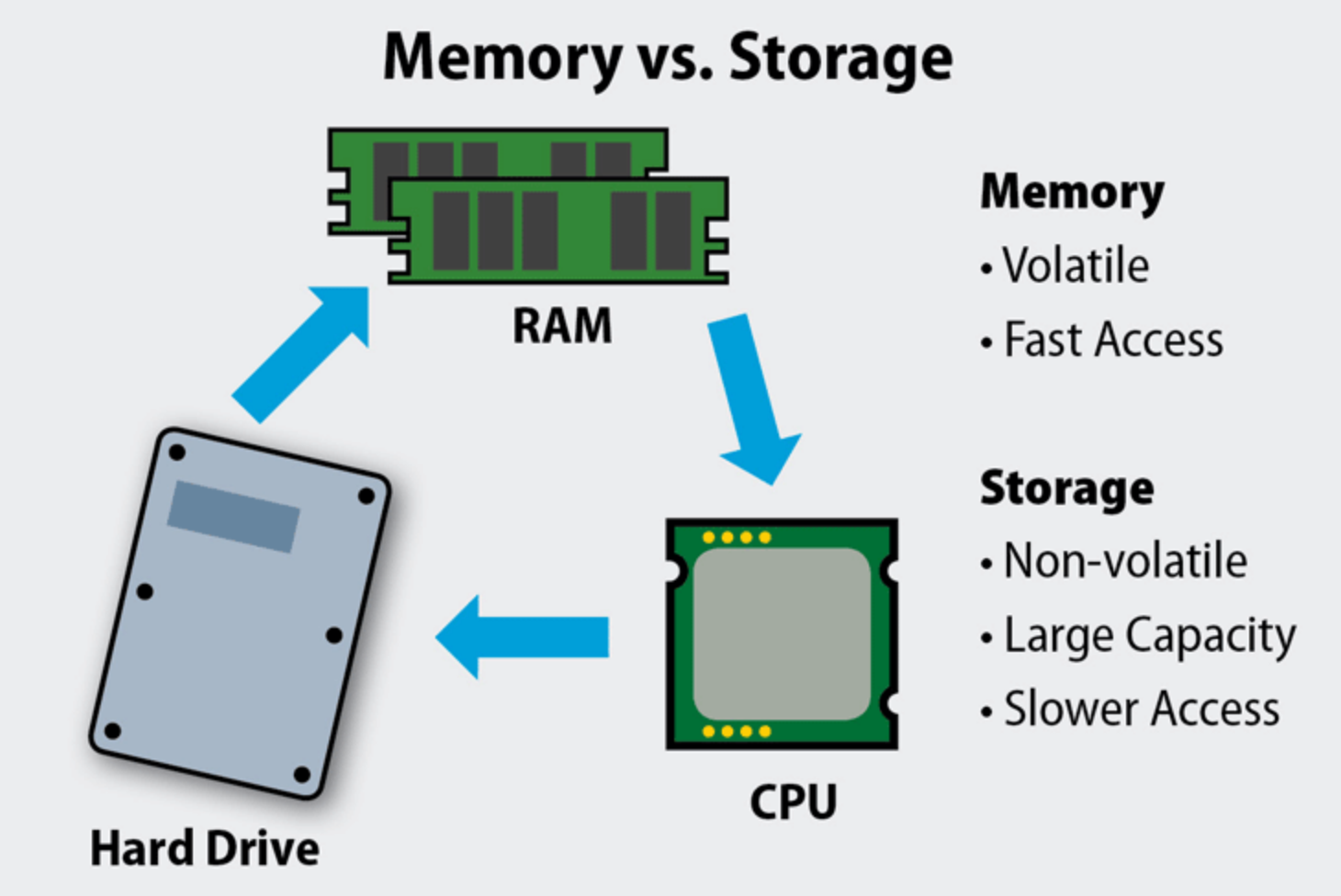
Tidyverse and beyond resources

RStudio cheat sheets including tidyverse packages

Good overview of data wrangling using tidyverse

Hadley Wickham's R for Data Science

Computer memory and working with data



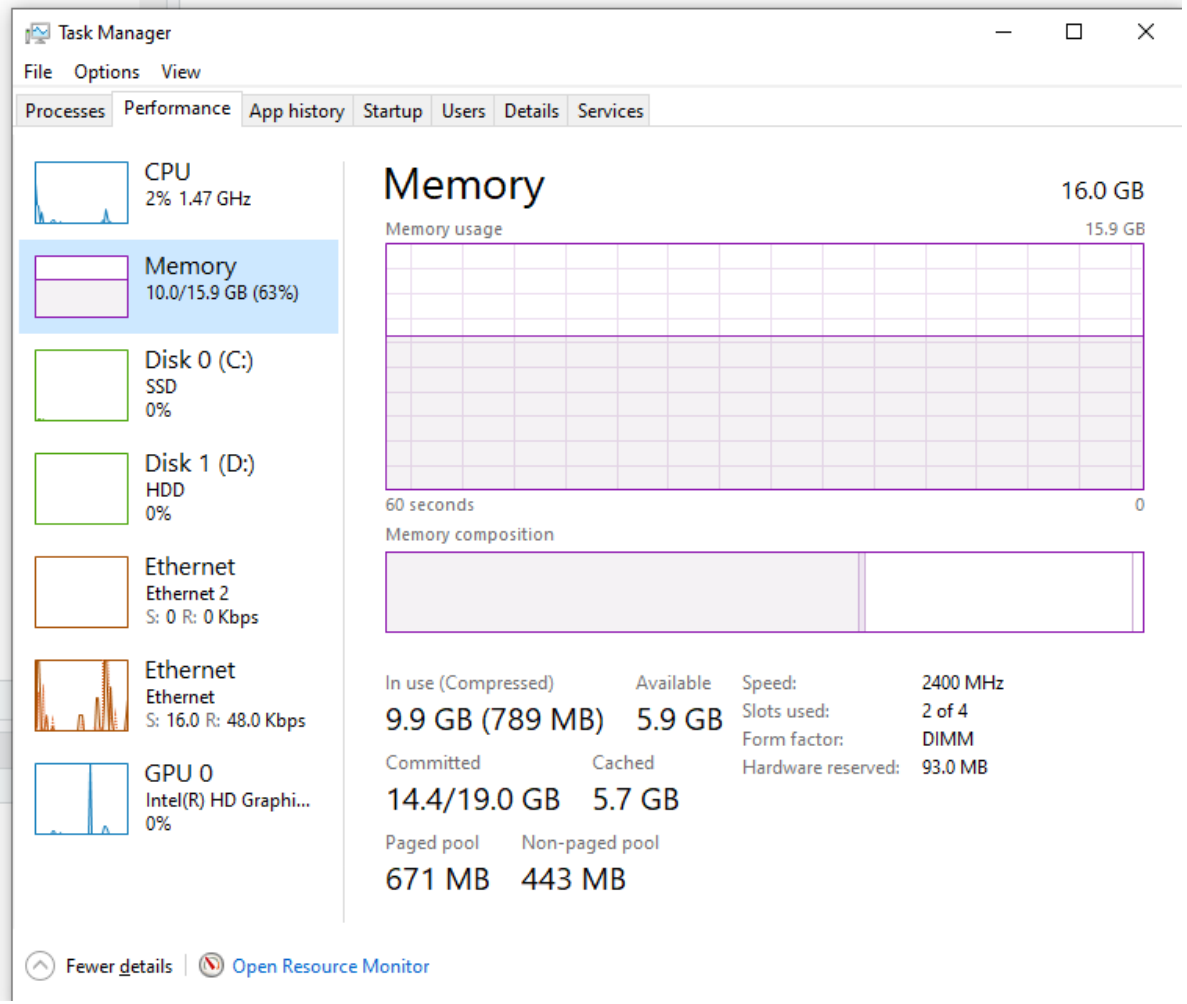
“Medium” data:
< 2 GB

“Large” data:
2-10 GB

“Big” data:
10+ GB

Computer memory

arrests	155507 obs. of 15 variables
pricey	39416 obs. of 10 variables
tb	121149 obs. of 5 variables



Rstudio with
~300,000 rows of data

Google Chrome with
10 tabs open