R4DS

Cohort 4

Wed 6:00 – 7:00 US Central

Twitter: @Rspjut

5-MINUTE ICE BREAKER

Any home projects you are putting off?

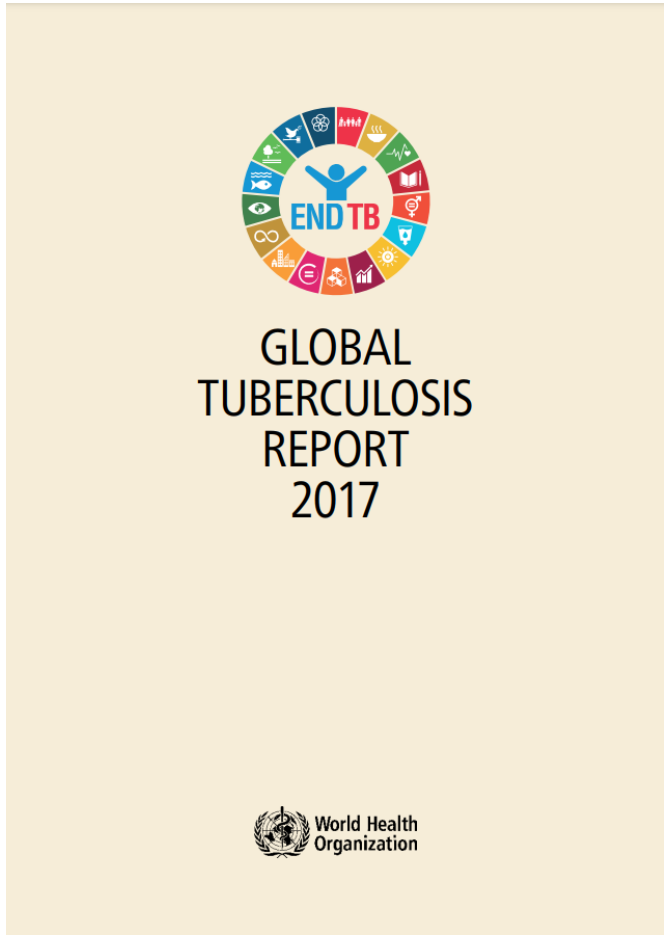
AGENDA

- 5-Minute Ice breaker
- Quick Housekeeping Reminders
- Tidy Data Case Study
- Chapter 13 – Relational Data
- Next Week
- Getting Help

QUICK HOUSEKEEPING REMINDERS

- Video camera is optional, but encouraged.
- I purposely err on the side of going fast. Slowing me down does not hurt my feelings.
- Take time to learn the theory (Grammar of Graphics, Tidy Data whitepaper, Relational Database theory, Appropriate Visualization Types, etc.).
- Please do the chapter exercises. Second-best learning opportunity!
- Please plan on teaching one of the lessons. Best learning opportunity!

TIDY DATA: CASE STUDY



who %>% View()

	country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m
13	Afghanistan	AF	AFG	1992	NA	NA	NA	
14	Afghanistan	AF	AFG	1993	NA	NA	NA	
15	Afghanistan	AF	AFG	1994	NA	NA	NA	
16	Afghanistan	AF	AFG	1995	NA	NA	NA	
17	Afghanistan	AF	AFG	1996	NA	NA	NA	
18	Afghanistan	AF	AFG	1997	0	10	6	
19	Afghanistan	AF	AFG	1998	30	129	128	
20	Afghanistan	AF	AFG	1999	8	55	55	
21	Afghanistan	AF	AFG	2000	52	228	183	
22	Afghanistan	AF	AFG	2001	129	379	349	
23	Afghanistan	AF	AFG	2002	90	476	481	
24	Afghanistan	AF	AFG	2003	127	511	436	

Info Encoded in Column Headers

New (if the TB cases are new or old; all of these are new)

Type of TB (rel, ep, sn, sp)

Patient Sex (m, f)

Age Group (014 = 0 to 14 years, etc.)

TIDY DATA: CASE STUDY

Objective: Convert columns into rows and remove NA values.

Command: pivot_longer, values_drop_na = TRUE

	country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m
14	Afghanistan	AF	AFG	1993	NA	NA	NA	
15	Afghanistan	AF	AFG	1994	NA	NA	NA	
16	Afghanistan	AF	AFG	1995	NA	NA	NA	
17	Afghanistan	AF	AFG	1996	NA	NA	NA	
18	Afghanistan	AF	AFG	1997	0	10	6	
19	Afghanistan	AF	AFG	1998	30	129	128	
20	Afghanistan	AF	AFG	1999	8	55	55	
21	Afghanistan	AF	AFG	2000	52	228	183	
22	Afghanistan	AF	AFG	2001	129	379	349	
23	Afghanistan	AF	AFG	2002	90	476	481	
24	Afghanistan	AF	AFG	2003	127	511	436	

```
who %>%
  pivot_longer(5:60,
               names_to = "key",
               values_to = "cases",
               values_drop_na = TRUE)
```

Filter

	country	iso2	iso3	year	key	cases
1	Afghanistan	AF	AFG	1997	new_sp_m014	0
2	Afghanistan	AF	AFG	1997	new_sp_m1524	10
3	Afghanistan	AF	AFG	1997	new_sp_m2534	6
4	Afghanistan	AF	AFG	1997	new_sp_m3544	3
5	Afghanistan	AF	AFG	1997	new_sp_m4554	5
6	Afghanistan	AF	AFG	1997	new_sp_m5564	2
7	Afghanistan	AF	AFG	1997	new_sp_m65	0
8	Afghanistan	AF	AFG	1997	new_sp_f014	5
9	Afghanistan	AF	AFG	1997	new_sp_f1524	38
10	Afghanistan	AF	AFG	1997	new_sp_f2534	36
11	Afghanistan	AF	AFG	1997	new_sp_f3544	14

TIDY DATA: CASE STUDY

Objective: Account for inconsistent column naming (new_ep vs newrel)

Command: mutate, str_replace

	country	iso2	iso3	year	key	cases
76028	Zimbabwe	ZW	ZWE	2012	new_ep_f2534	710
76029	Zimbabwe	ZW	ZWE	2012	new_ep_f3544	579
76030	Zimbabwe	ZW	ZWE	2012	new_ep_f4554	228
76031	Zimbabwe	ZW	ZWE	2012	new_ep_f5564	140
76032	Zimbabwe	ZW	ZWE	2012	new_ep_f65	143
76033	Zimbabwe	ZW	ZWE	2013	newrel_m014	1315
76034	Zimbabwe	ZW	ZWE	2013	newrel_m1524	1642
76035	Zimbabwe	ZW	ZWE	2013	newrel_m2534	5331
76036	Zimbabwe	ZW	ZWE	2013	newrel_m3544	5363
76037	Zimbabwe	ZW	ZWE	2013	newrel_m4554	2349
76038	Zimbabwe	ZW	ZWE	2013	newrel_m5564	1206
76039	Zimbabwe	ZW	ZWE	2013	newrel_m65	1208
76040	Zimbabwe	ZW	ZWE	2013	newrel_f014	1252
76041	Zimbabwe	ZW	ZWE	2013	newrel_f1524	2069

```
who %>%
  pivot_longer(5:60,
               names_to = "key",
               values_to = "cases",
               values_drop_na = T) %>%
  mutate(key = str_replace(key, "newrel","new_rel"))
```

	country	iso2	iso3	year	key	cases
76028	Zimbabwe	ZW	ZWE	2012	new_ep_f2534	710
76029	Zimbabwe	ZW	ZWE	2012	new_ep_f3544	579
76030	Zimbabwe	ZW	ZWE	2012	new_ep_f4554	228
76031	Zimbabwe	ZW	ZWE	2012	new_ep_f5564	140
76032	Zimbabwe	ZW	ZWE	2012	new_ep_f65	143
76033	Zimbabwe	ZW	ZWE	2013	new_rel_m014	1315
76034	Zimbabwe	ZW	ZWE	2013	new_rel_m1524	1642
76035	Zimbabwe	ZW	ZWE	2013	new_rel_m2534	5331
76036	Zimbabwe	ZW	ZWE	2013	new_rel_m3544	5363
76037	Zimbabwe	ZW	ZWE	2013	new_rel_m4554	2349
76038	Zimbabwe	ZW	ZWE	2013	new_rel_m5564	1206
76039	Zimbabwe	ZW	ZWE	2013	new_rel_m65	1208
76040	Zimbabwe	ZW	ZWE	2013	new_rel_f014	1252
76041	Zimbabwe	ZW	ZWE	2013	new_rel_f1524	2069

TIDY DATA: CASE STUDY

Objective: Decode information in the “key” column

Command: separate by delimiter, separate by position

	country	iso2	iso3	year	key	cases
1	Afghanistan	AF	AFG	1997	new_sp_m014	0
2	Afghanistan	AF	AFG	1997	new_sp_m1524	10
3	Afghanistan	AF	AFG	1997	new_sp_m2534	6
4	Afghanistan	AF	AFG	1997	new_sp_m3544	3
5	Afghanistan	AF	AFG	1997	new_sp_m4554	5
6	Afghanistan	AF	AFG	1997	new_sp_m5564	2

```
who %>%
  pivot_longer(5:60,
               names_to = "key",
               values_to = "cases",
               values_drop_na = TRUE) %>%
  mutate(key = str_replace(key, "newrel","new_rel")) %>%
  separate(key,
           into = c("new","diag", "gage")) %>%
  separate(gage,
           into = c("gender", "agegroup"),
           sep = 1)
```

	country	iso2	iso3	year	new	diag	gender	agegroup	cases
1	Afghanistan	AF	AFG	1997	new	sp	m	014	0
2	Afghanistan	AF	AFG	1997	new	sp	m	1524	10
3	Afghanistan	AF	AFG	1997	new	sp	m	2534	6
4	Afghanistan	AF	AFG	1997	new	sp	m	3544	3
5	Afghanistan	AF	AFG	1997	new	sp	m	4554	5
6	Afghanistan	AF	AFG	1997	new	sp	m	5564	2

TIDY DATA: CASE STUDY

Objective: Remove unneeded columns and filter to 2008 – 2010 data

Command: select, filter

	country	iso2	iso3	year	new	diag	gender	agegroup	cases
1	Afghanistan	AF	AFG	1997	new	sp	m	014	0
2	Afghanistan	AF	AFG	1997	new	sp	m	1524	10
3	Afghanistan	AF	AFG	1997	new	sp	m	2534	6
4	Afghanistan	AF	AFG	1997	new	sp	m	3544	3
5	Afghanistan	AF	AFG	1997	new	sp	m	4554	5
6	Afghanistan	AF	AFG	1997	new	sp	m	5564	2

```
who %>%
  pivot_longer(5:60,
               names_to = "key",
               values_to = "cases",
               values_drop_na = TRUE) %>%
  mutate(key = str_replace(key, "newrel","new_rel")) %>%
  separate(key,
           into = c("new","diag", "gage")) %>%
  separate(gage,
           into = c("gender", "agegroup"),
           sep = 1) %>%
  select(-new, -iso2, -iso3) %>%
  filter(year %in% c('2008','2009','2010'))
```

	country	year	diag	gender	agegroup	cases
1	Afghanistan	2008	sp	m	014	187
2	Afghanistan	2008	sp	m	1524	941
3	Afghanistan	2008	sp	m	2534	773
4	Afghanistan	2008	sp	m	3544	545
5	Afghanistan	2008	sp	m	4554	570
6	Afghanistan	2008	sp	m	5564	630

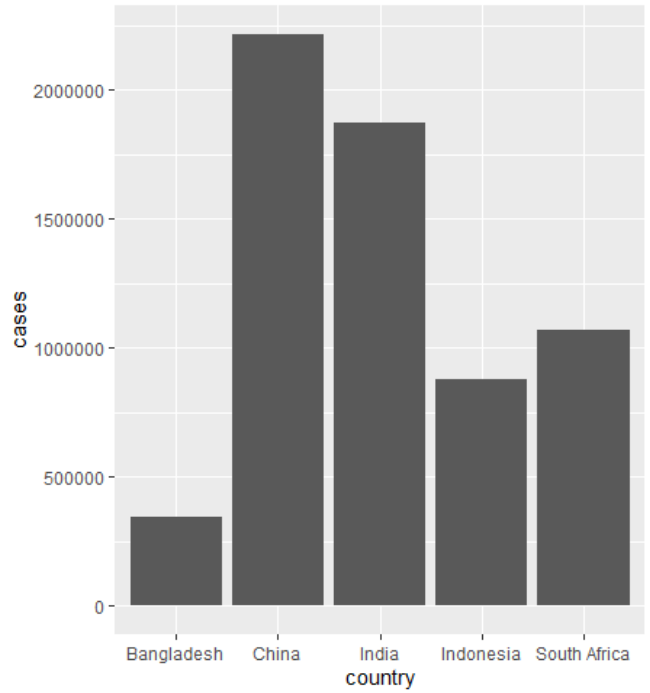
TIDY DATA: CASE STUDY

Objective: Plot the top 5 countries with highest cases

Command: group_by, summarize, arrange, top_n, ggplot + geom_bar

	country	year	diag	gender	agegroup	cases
1	Afghanistan	2008	sp	m	014	187
2	Afghanistan	2008	sp	m	1524	941
3	Afghanistan	2008	sp	m	2534	773
4	Afghanistan	2008	sp	m	3544	545
5	Afghanistan	2008	sp	m	4554	570
6	Afghanistan	2008	sp	m	5564	630

```
who %>%
  pivot_longer(5:60,
               names_to = "key",
               values_to = "cases",
               values_drop_na = TRUE) %>%
  mutate(key = str_replace(key, "newrel","new_rel")) %>%
  separate(key,
           into = c("new","diag", "gage")) %>%
  separate(gage,
           into = c("gender", "agegroup"),
           sep = 1) %>%
  select(-new, -iso2, -iso3) %>%
  filter(year %in% c('2008','2009','2010'))
  group_by(country) %>%
  summarize(cases = sum(cases)) %>%
  arrange(desc(cases)) %>%
  top_n(5) %>%
  ggplot() + geom_bar(aes(country, cases), stat = "identity")
```



RELATIONAL DATA : THEORY AND PRACTICE

Illustrations from https://www.geeksforgeeks.org

UN-NORMALIZED TABLE (ZERO NORMAL FORM)

STUD_NAME	STUD_PHONE	STUD_STATE	STUD_COUNTRY
RAMESH	971-627-1721, 987-171-7178	HARYANA	INDIA
RAMESH	989-829-7281	PUNJAB	INDIA

FIRST NORMAL FORM

STUD_NAME	STUD_PHONE	STUD_STATE	STUD_COUNTRY
RAMESH	971-627-1721	HARYANA	INDIA
RAMESH	987-171-7178	HARYANA	INDIA
RAMESH	989-829-7281	PUNJAB	INDIA

- Each table cell should contain a single value.
- Each record (all values combined) needs to be unique.

RELATIONAL DATA : THEORY AND PRACTICE

Illustrations from https://www.geeksforgeeks.org

FIRST NORMAL FORM

STUD_NAME	STUD_PHONE	STUD_STATE	STUD_COUNTRY
RAMESH	971-627-1721	HARYANA	INDIA
RAMESH	987-171-7178	HARYANA	INDIA
RAMESH	989-829-7281	PUNJAB	INDIA

SECOND NORMAL FORM

STUD_NO	STUD_NAME	STUD_STATE	STUD_COUNTRY
1	RAMESH	HARYANA	INDIA
2	RAMESH	PUNJAB	INDIA

STUD_NO	STUD_PHONE
1	971-627-1721
1	987-171-7178
2	989-829-7281

- Be in First Normal Form
- Single Column Primary Key

RELATIONAL DATA : THEORY AND PRACTICE

Illustrations from https://www.geeksforgeeks.org

SECOND NORMAL FORM

STUD_NO	STUD_NAME	STUD_STATE	STUD_COUNTRY
1	RAMESH	HARYANA	INDIA
2	RAMESH	PUNJAB	INDIA

STUD_NO	STUD_PHONE
1	971-627-1721
1	987-171-7178
2	989-829-7281

- Be in Second Normal Form
- Have no transitive functional dependencies

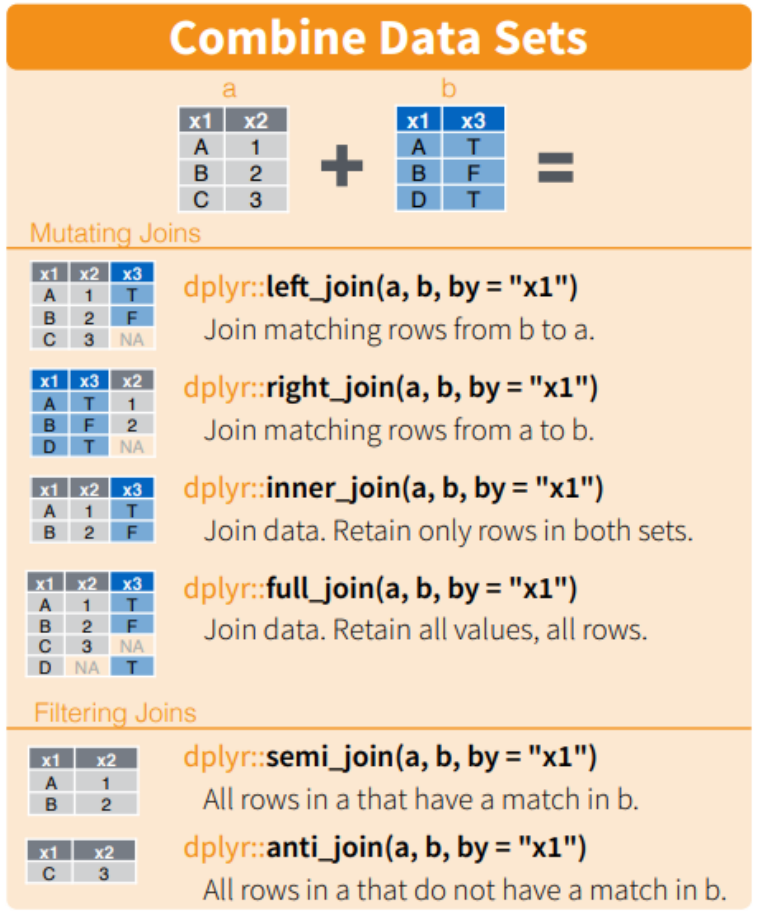
THIRD NORMAL FORM

STUD_NO	STUD_NAME	STUD_STATE
1	RAMESH	HARYANA
2	RAMESH	PUNJAB

STUD_NO	STUD_PHONE
1	971-627-1721
1	987-171-7178
2	989-829-7281

STATE	COUNTRY
HARYANA	INDIA
PUNJAB	INDIA

RELATIONAL DATA : JOINS




```r
library(tidyverse)
left_tibble <- tibble(ID = 1:2,
                      Col1 = c("a1", "a2"))
right_tibble <- tibble(ID = 2:3,
                       Col1 = c("b1", "b2"))

left_tibble
right_tibble
```

left_tibble

	ID	Col1
1	1	a1
2	2	a2

right_tibble

	ID	Col1
1	2	b1
2	3	b2

NEXT WEEK…

- Chapter 14 – Strings

GETTING HELP

- Ask questions during our call
- Google
- Stack Overflow
- Slack
- Office Hours r4ds.io/calendar
- Twitter #rstats
- r4ds answer keys: Jeff Arnold (preferred) or Bryan Shalloway (also good)
- Cheatsheets


RStudio