

Data Visualization with R

Workshop Part 2



Tuberculosis incidence

The TB data is from the WHO.

Show 10 entries

Search:

	country	iso3	year	new_sp_m04	new_sp_m514	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_n
1	Australia	AUS	2009	0	3	3	30	37	
2	Australia	AUS	2010	0	2	2	42	33	
3	Australia	AUS	2011	is the data in tidy form?		2	no	38	44
4	Australia	AUS	2012	2		1	3	26	40
5	Australia	AUS	2013	what are the variables?		country, year, sex, age			
6	Australia	AUS	2014	how many variables are there?		four (five is ok)			
7	Australia	AUS	2015						
8	Australia	AUS	2016						
9	Australia	AUS	2017						
10	Australia	AUS	2018						

Reshape your data into tidy form so that it is easy, and clear how the variables are mapped into elements of the plot.

Showing 1 to 10 of 20 entries

Previous

1

2

Next

Tidy data

Show 10 ▾ entries

Search:

	country	year	age_group	sex	count
1	Australia	1997	15-24	m	is count a variable?
2	Australia	1997	25-34	m	24
3	Australia	1997	35-44	m	no
4	Australia	1997	45-54	m	13
5	Australia	1997	55-64	m	17
6	Australia	1997	65-	m	28
7	Australia	1997	15-24	f	10
8	Australia	1997	25-34	f	15
9	Australia	1997	35-44	f	9
10	Australia	1997	45-54	f	5

Showing 1 to 10 of 192 entries

Previous

1

2

3

4

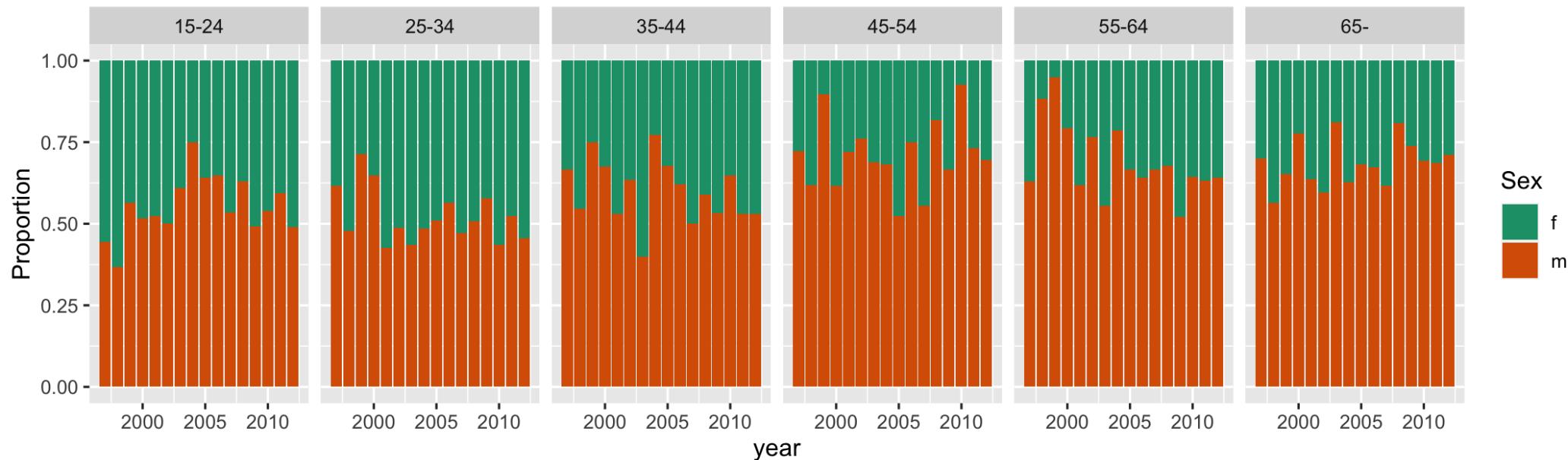
5

...

20

Next

```
ggplot(tb_oz, aes(x = year, y = count, fill = sex)) +  
  geom_bar(stat = "identity", position = "fill") +  
  facet_wrap(~age_group, ncol = 6) +  
  scale_fill_brewer(name = "Sex", palette = "Dark2") +  
  ylab("Proportion")
```



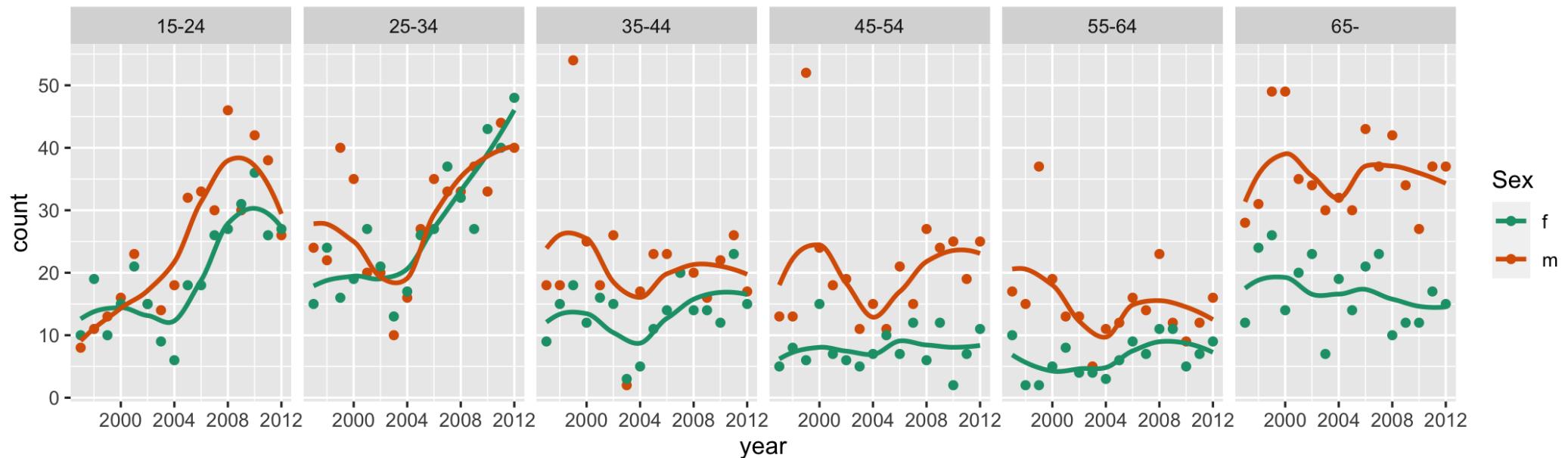
What do we learn about the data?

1. Generally, proportion is higher for males
2. Relative proportion is higher for older males
3. Quite variable proportions from year to year

What does this plot hide from us?

Information about counts is lost,
thus reliability of the proportions is
lost

```
ggplot(tb_oz, aes(x = year, y = count, colour = sex)) +  
  geom_point() +  
  geom_smooth(se = F) +  
  facet_wrap(~age_group, ncol = 6) +  
  scale_colour_brewer(name = "Sex", palette = "Dark2")
```



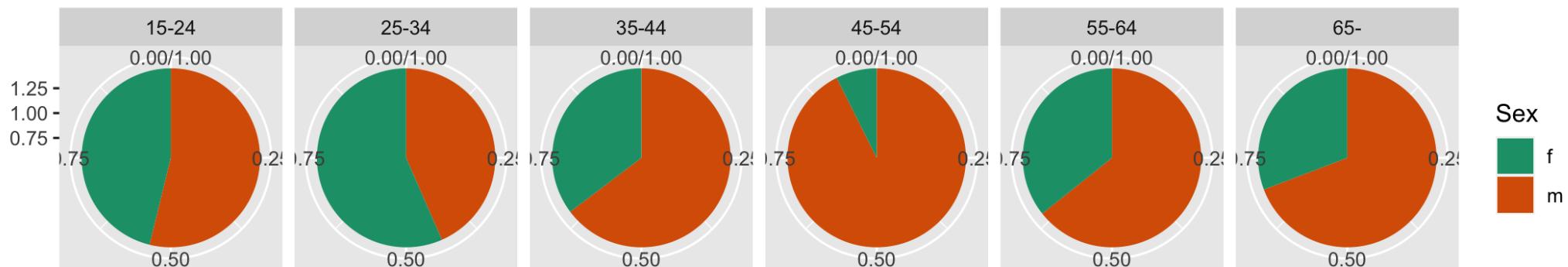
What do we learn?

Generally, counts are quite varied from year to year, but relatively stable

Increasing trend in counts for both males and females under 35

Counts for males almost always higher than females

```
tb_oz %>%
  filter(year == 2010) %>%
  ggplot(aes(x = 1, y = count, fill = sex)) +
  geom_bar(stat = "identity", position = "fill") +
  facet_wrap(~age_group, ncol = 6) +
  scale_fill_brewer(name = "Sex", palette = "Dark2") +
  xlab("") + ylab("") +
  coord_polar(theta = "y")
```



What do we learn?

In 2010,

there were almost no 45-54 year old women with TB

there were more 24-35 year old women with TB than men

generally more males than females had TB

How many plots should you usually do?

Lots! In order to understand your data, look at it in many different ways. Like you might do to explore some new object.



`</> Open part2-exercise-01.Rmd`

15 : 00

Session Information

```
devtools::session_info()
```

```
## - Session info -
##   setting  value
##   version  R version 4.1.2 (2021-11-01)
##   os        macOS Big Sur 11.5.1
##   system   aarch64, darwin20
##   ui        X11
##   language (EN)
##   collate  en_AU.UTF-8
##   ctype    en_AU.UTF-8
##   tz       Australia/Melbourne
##   date     2022-02-20
##   pandoc   2.16.2 @ /usr/local/bin/ (via rmarkdown)
##
```

These slides are licensed under

