

Data Visualization with R

Workshop Day 2

Review of Grammar of Graphics

Presented by Di Cook

Department of Econometrics and Business Statistics



MONASH University

29th July 2020 @ Statistical Society of Australia | Zoom

demo data for today



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_m
1	Australia	AUS	1980						
2	Australia	AUS	1981						
3	Australia	AUS	1982						
4	Australia	AUS	1983						
5	Australia	AUS	1984						
6	Australia	AUS	1985						
7	Australia	AUS	1986						
8	Australia	AUS	1987						
9	Australia	AUS	1988						
10	Australia	AUS	1989						

Showing 1 to 10 of 78 entries

Previous

1

2

3

4

5

...

8

Next

- Is the data in tidy form?
- What are the variables in this data?
- How many variables are there? country (name, iso3), year, sex, age

Tidying the data

Show 10 entries

 Search:

	country	year	age_group	sex	count
1	Australia	1997	15-24	m	8
2	Australia	1997	25-34	m	24
3	Australia	1997	35-44	m	18
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

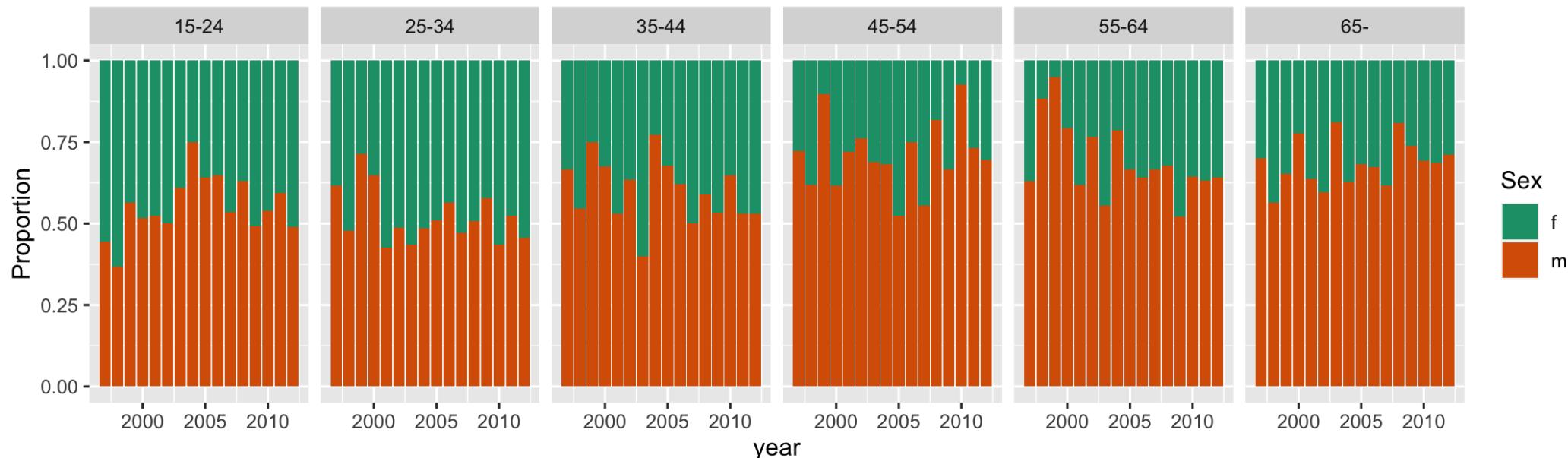
5

...

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")
```

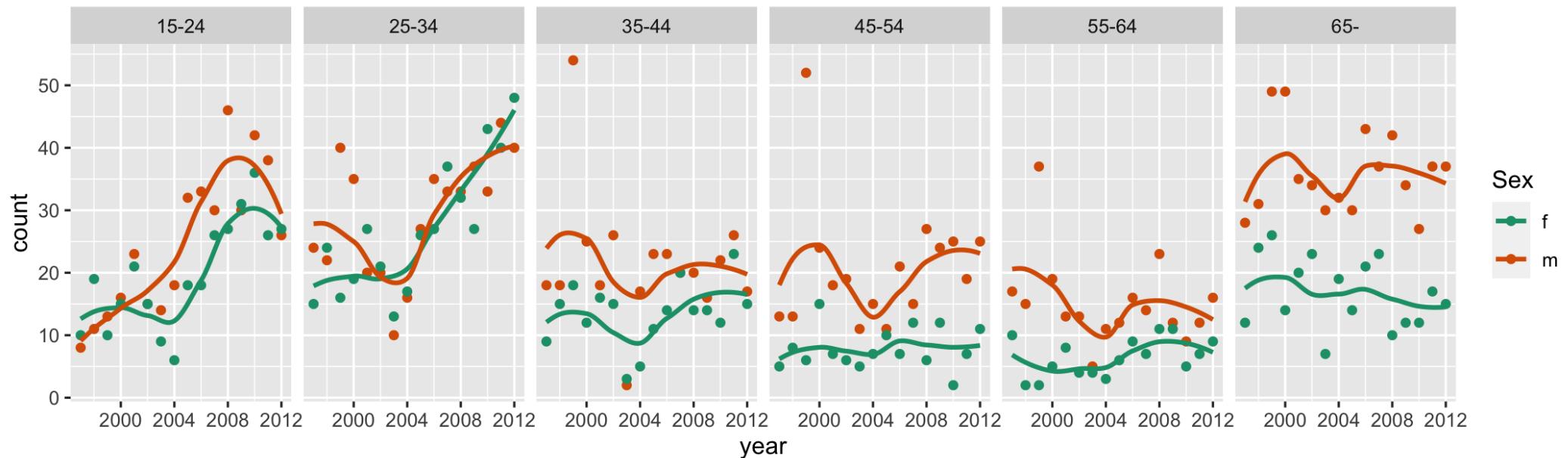


What do we learn?

- Generally, proportion is higher for males
- Relative proportion is higher for older males
- Quite variable proportions from year to year

What don't we learn from this plot?

```
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





</> Open day2-exercise-01.Rmd

15:00

Session Information

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

```
## - Session info --
##   setting  value
##   version  R version 4.0.1 (2020-06-06)
##   os        macOS Catalina 10.15.6
##   system   x86_64, darwin17.0
##   ui        X11
##   language (EN)
##   collate  en_AU.UTF-8
##   ctype    en_AU.UTF-8
##   tz       Australia/Melbourne
##   date     2020-07-29
##
## - Packages --
##   package      * version    date lib source

```

These slides are licensed under

