

Data Visualization with R

Workshop Day 2

Determining the best plot design

Presented by Di Cook

Department of Econometrics and Business Statistics



MONASH University

12th Nov 2020 @ Statistical Society of Australia | Zoom

Let's play a game:
Which plot wears it better?



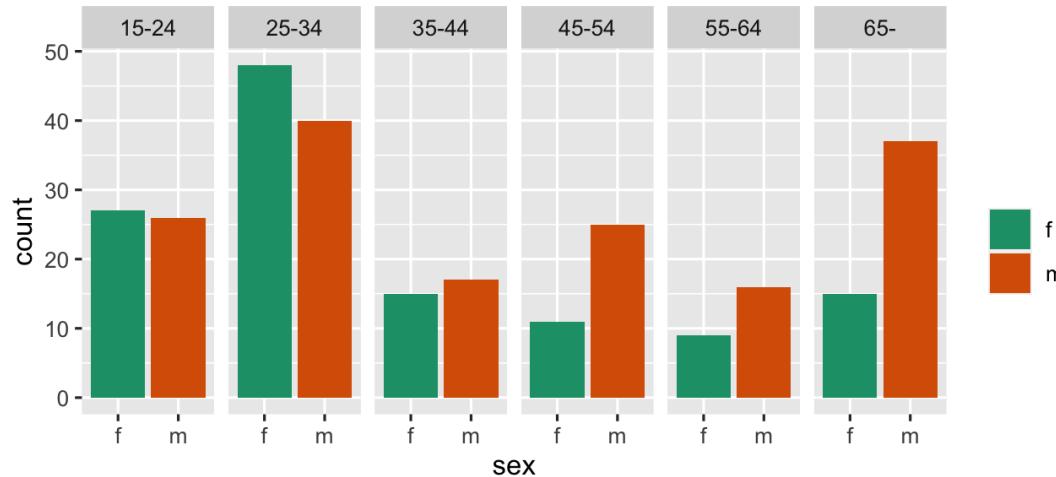
On the next slide we have made **two different plots** of 2012 TB incidence in Australia, based on two variables:

```
## # A tibble: 5 x 3
##   sex    age_group count
##   <chr> <fct>     <dbl>
## 1 m     15-24      26
## 2 m     25-34      40
## 3 m     35-44      17
## 4 m     45-54      25
## 5 m     55-64      16
```

- In arrangement A, separate plots are made for age, and sex is mapped to the x axis.
- Conversely, in arrangement B, separate plots are made for sex, and age is mapped to the x axis.

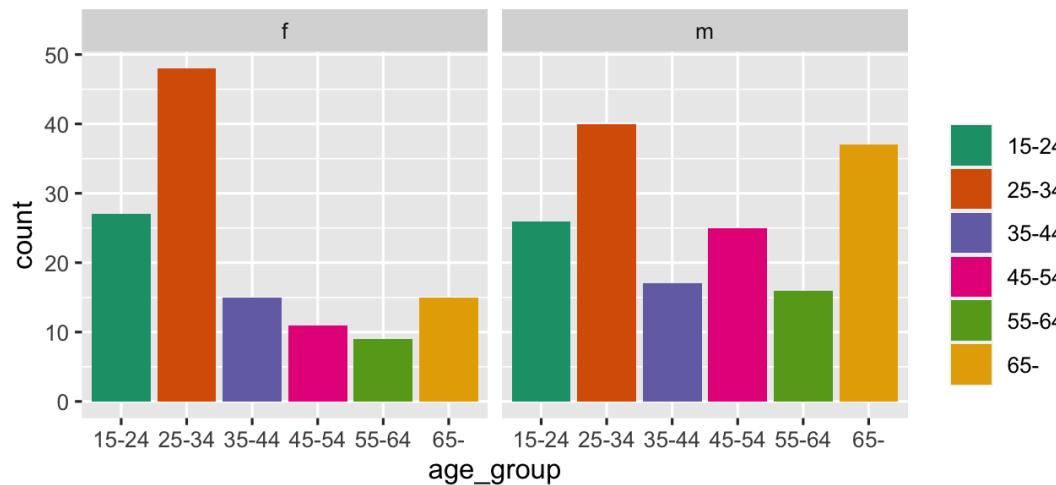
If you were to answer the question: **At which age(s) are the counts for males and females relatively the same?** Which plot makes this easier?

Arrangement A



We've got two different rearrangements of the same information. **At which age(s) are the counts for males and females relatively the same?** Which plot makes this easier?

Arrangement B

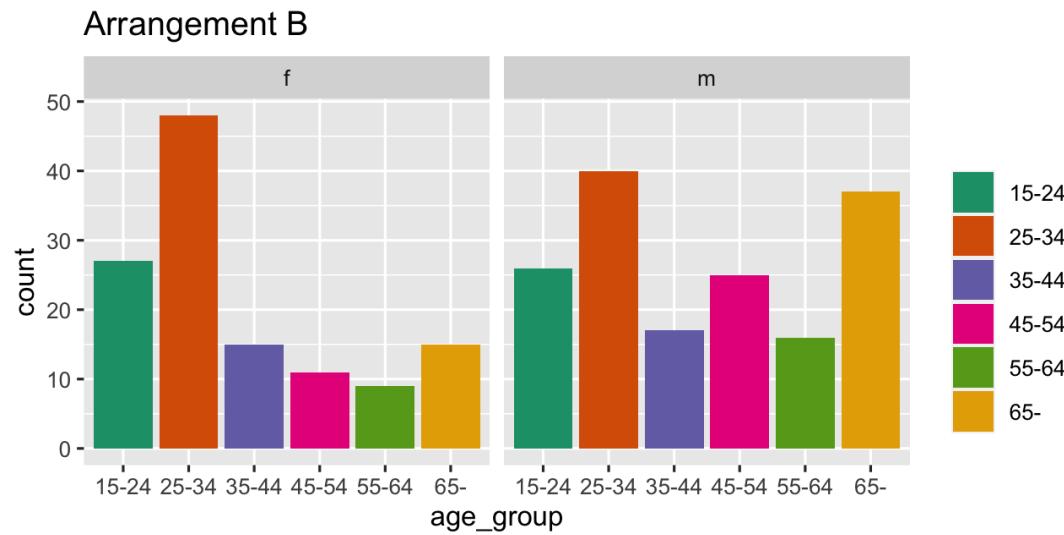
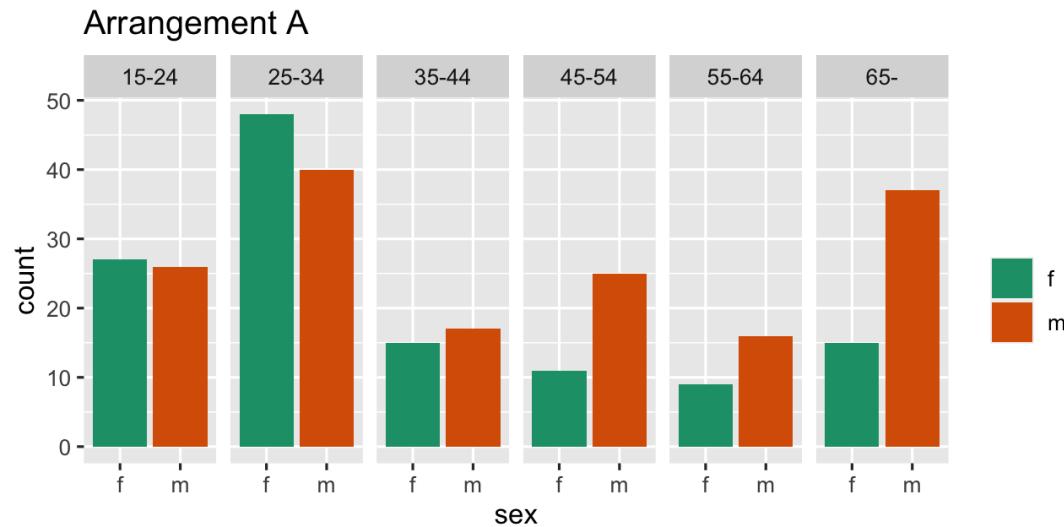


What do we learn? That is different from each? What's the focus of each? What's easy, what's harder?

00 : 30

Try to write out a question that would be easier to answer from arrangement B.

00 : 30



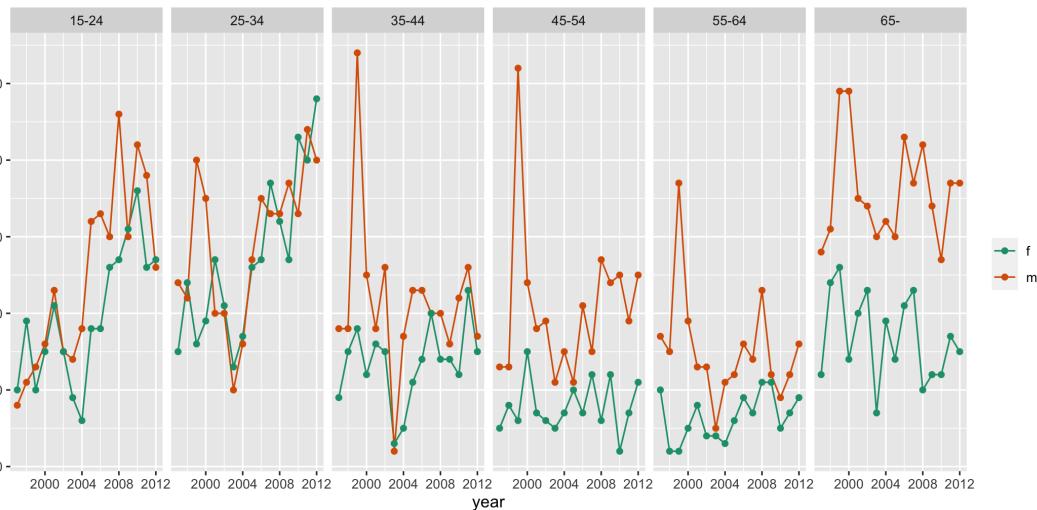
On the next slide we have made **two different plots** of TB incidence in the Australia, based on three variables:

```
## # A tibble: 5 x 4
##   year sex   age_group count
##   <dbl> <chr> <fct>     <dbl>
## 1 1997 m    15-24      8
## 2 1997 m    25-34     24
## 3 1997 m    35-44     18
## 4 1997 m    45-54     13
## 5 1997 m    55-64     17
```

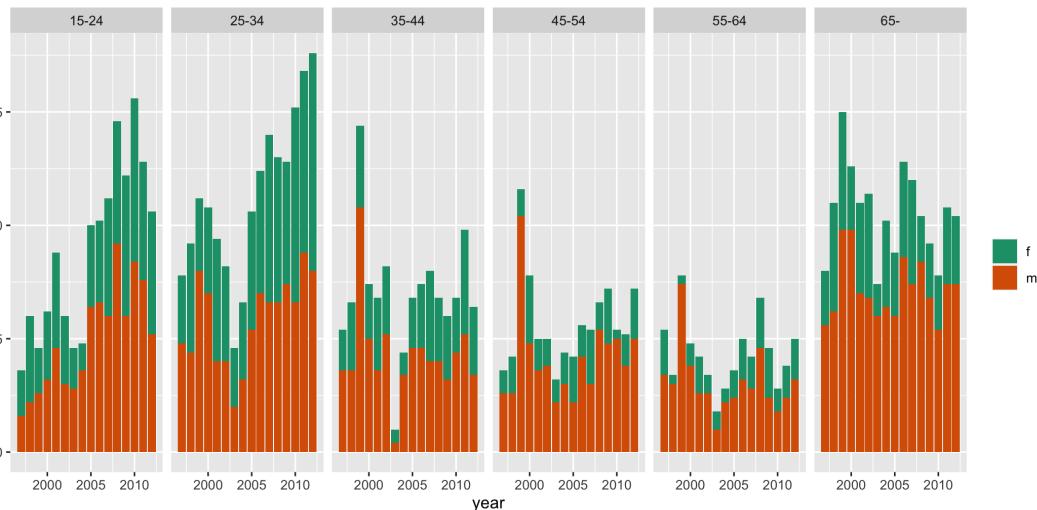
- In plot type A, a line plot of counts is drawn separately by age and sex, and year is mapped to the x axis.
- Conversely, in plot type B, counts for sex, and age are stacked into a bar chart, separately by age and sex, and year is mapped to the x axis

If you were to answer the question: **The trend in incidence over years for females is generally decreasing?** Which plot makes this easier?

Type A



Type B



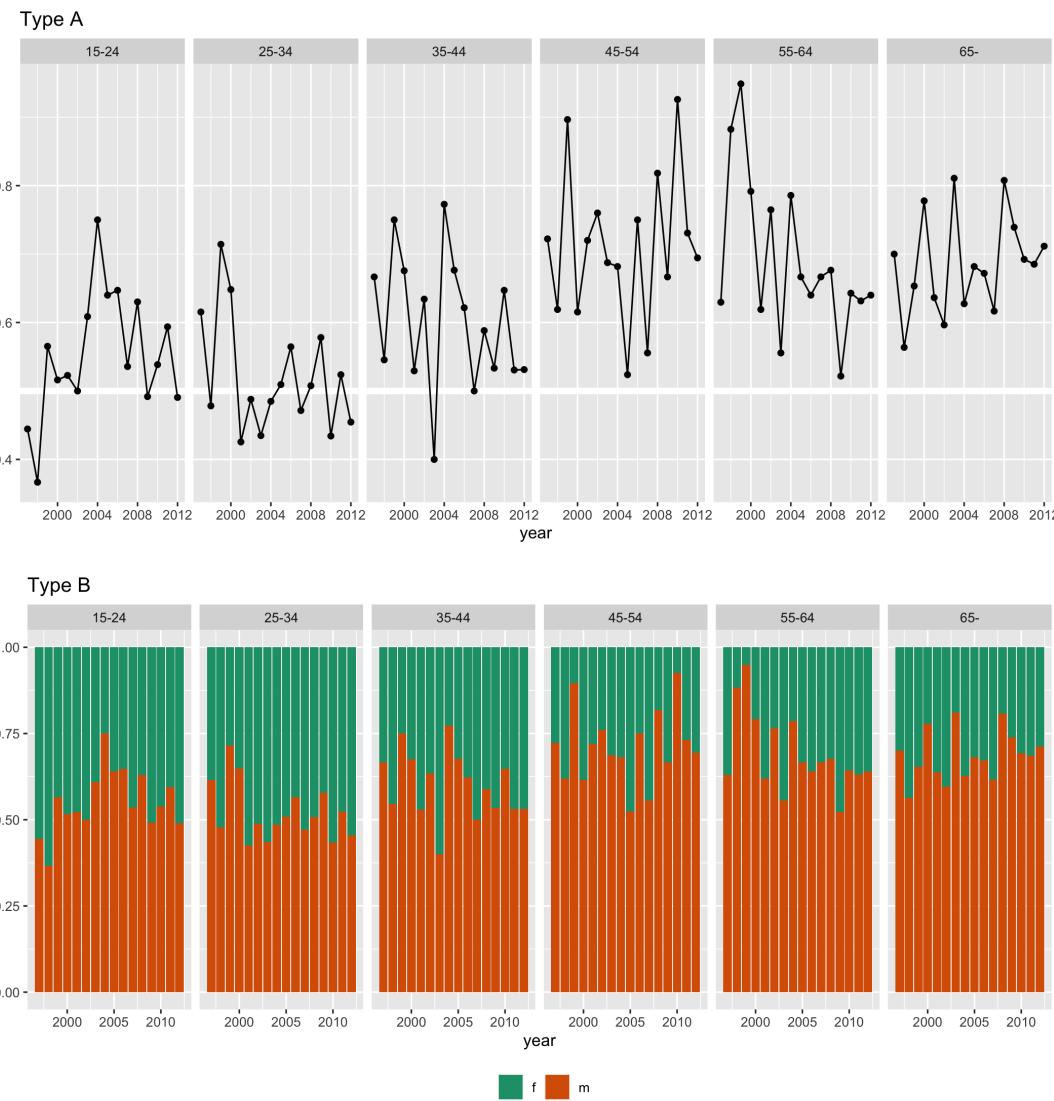
Which type of plot makes it easier to answer: **The trend in incidence over years for females is generally flat?**

What are the pros and cons of each way of displaying the same information? Should specific limits on axes be made?

00 : 30

The following plots focus on proportion of males vs females. Plot A computes the proportion and displays this as a line plot. Plot B uses a 100% chart of stacked bars for females and males. What are the strengths and weaknesses of each?

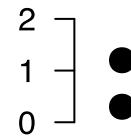
00 : 30



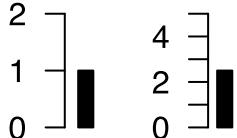
Perceptual principles

- Hierarchy of mappings
- Pre-attentive: some elements are noticed before you even realise it.
- Color palettes: qualitative, sequential, diverging, *palindrome*.
- Proximity: Place elements for primary comparison close together.
- Change blindness: When focus is interrupted differences may not be noticed.

POSITION
COMMON SCALE



POSITION
NON-ALIGNED SCALE



LENGTH



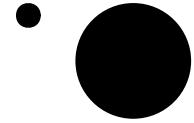
DIRECTION



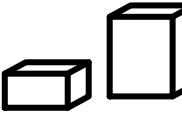
ANGLE



AREA



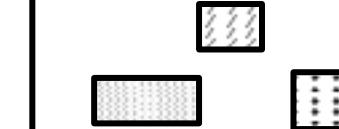
VOLUME



CURVATURE



TEXTURE



SHAPE



Hierarchy of mappings

1. Position - common scale (BEST)
2. Position - nonaligned scale
3. Length, direction, angle
4. Area
5. Volume, curvature
6. Shading, color (WORST)

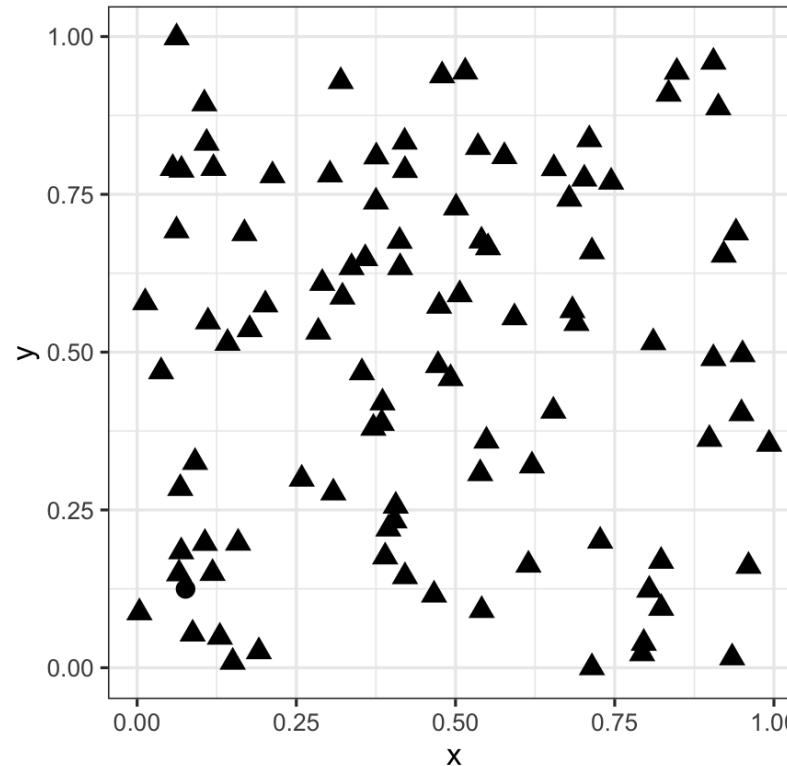
(Cleveland, 1984; Heer and Bostock, 2009)

1. scatterplot, barchart
2. side-by-side boxplot, stacked barchart
3. piechart, rose plot, gauge plot, donut, wind direction map, starplot
4. treemap, bubble chart, mosaicplot
5. chernoff face
6. choropleth map

Try to come up with a plot type for one of the mappings.

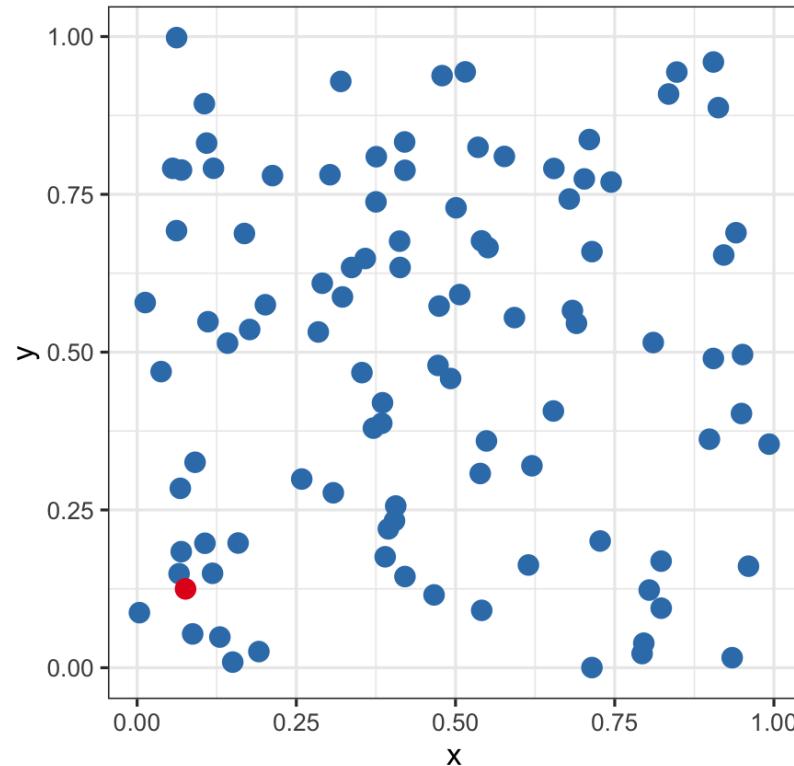
Pre-attentive

Can you find the odd one out?



Pre-attentive

Is it easier now?



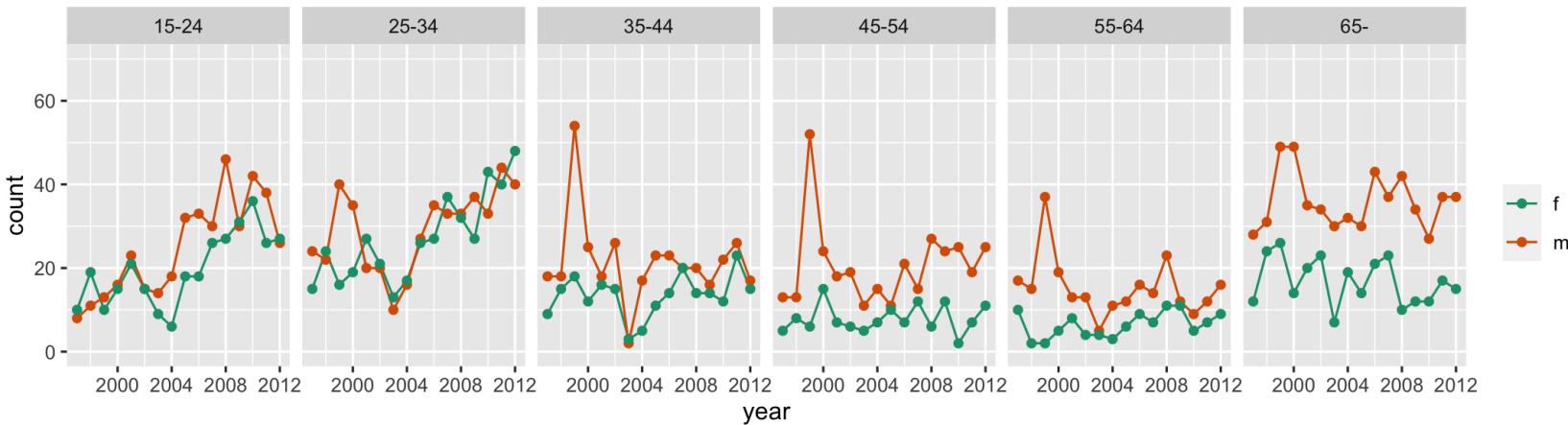
Proximity

Place elements that you want to compare close to each other. If there are multiple comparisons to make, you need to decide which one is most important.

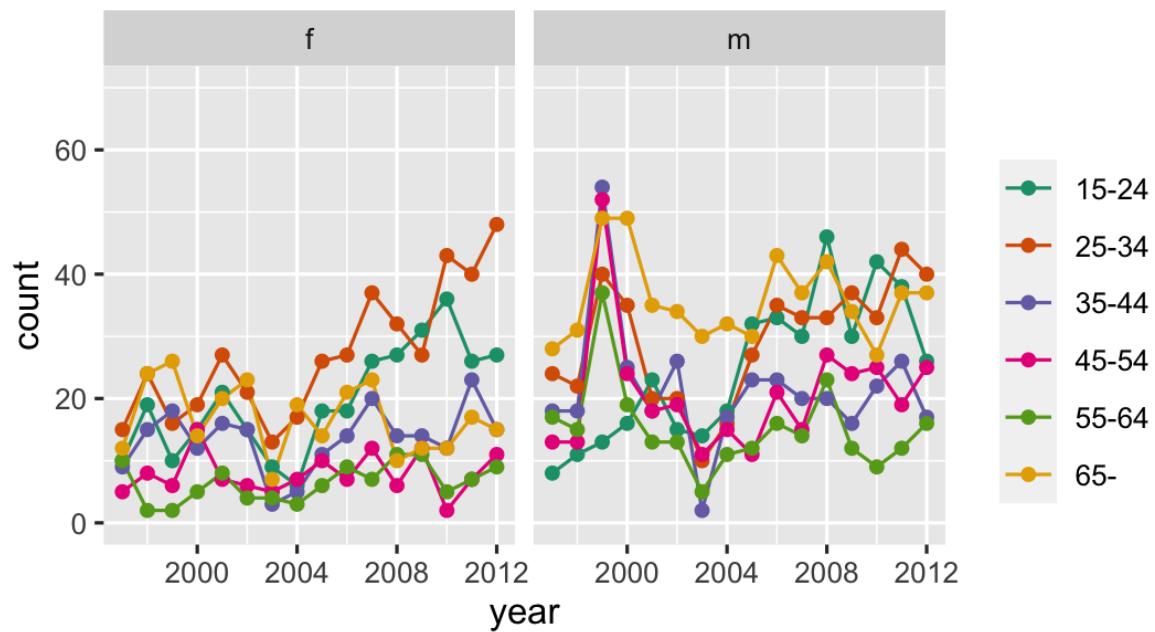
```
ggplot(tb_oz, aes(x = year, y = count, colour = sex)) +  
  geom_line() + geom_point() +  
  facet_wrap(~age_group, ncol = 6) +  
  ylim(c(0, 70)) +  
  scale_colour_brewer(name = "", palette = "Dark2") +  
  ggttitle("Arrangement A")
```

```
ggplot(tb_oz, aes(x = year, y = count, colour = age_group)) +  
  geom_line() + geom_point() +  
  facet_wrap(~sex, ncol = 2) +  
  ylim(c(0, 70)) +  
  scale_colour_brewer(name = "", palette = "Dark2") +  
  ggttitle("Arrangement B")
```

Arrangement A

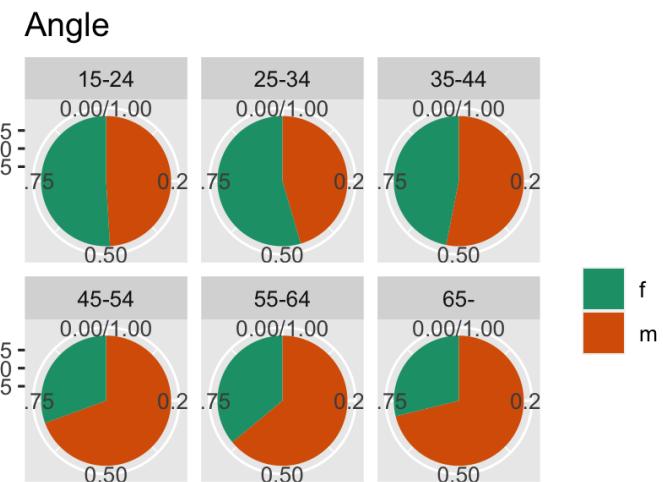
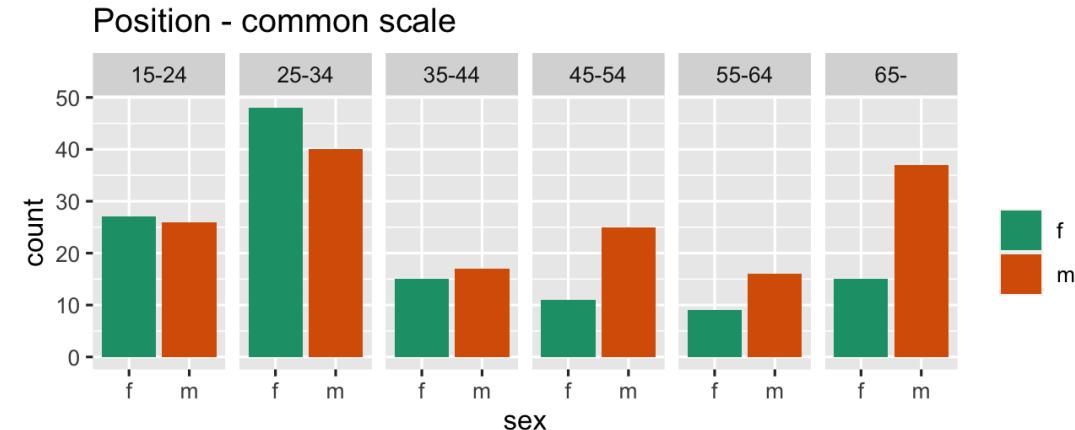


Arrangement B



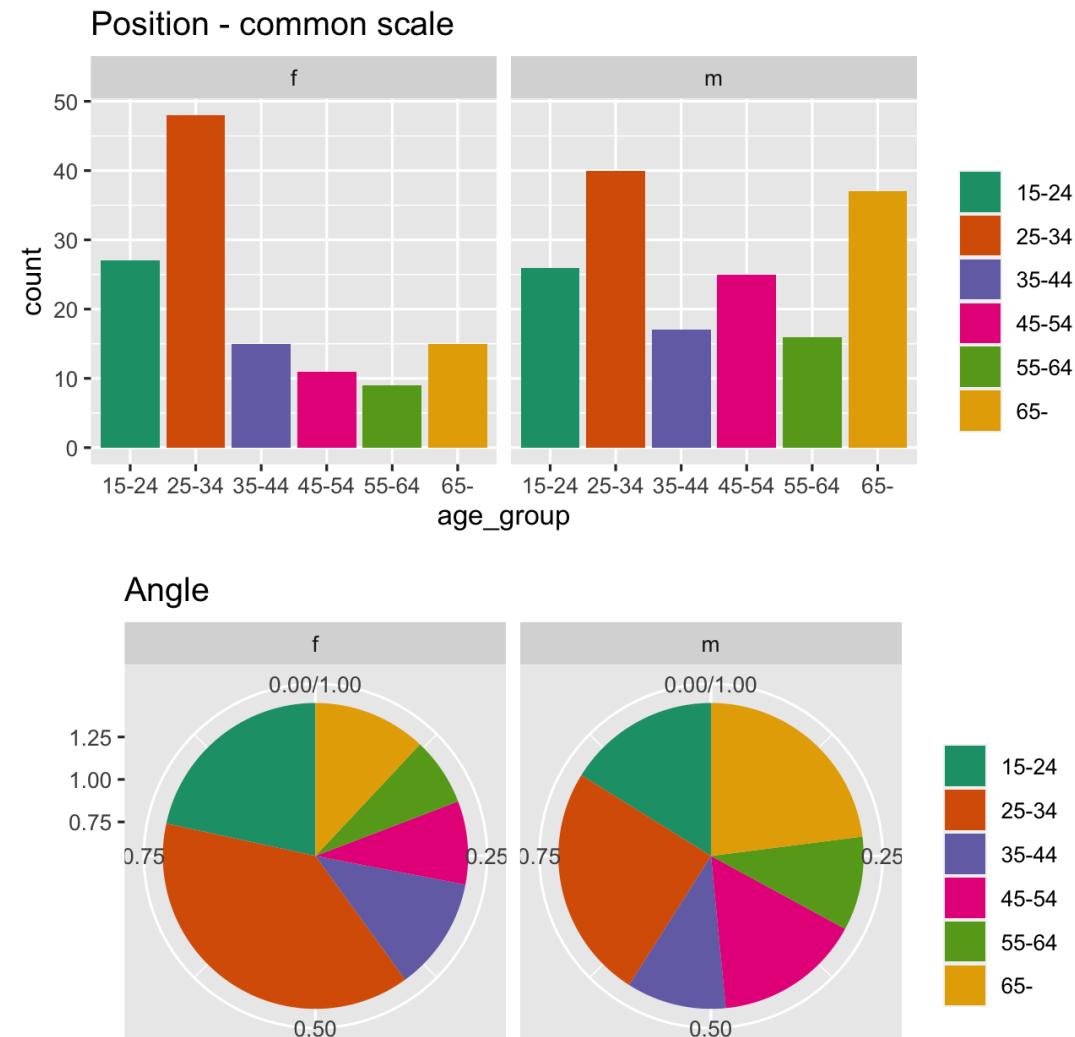
Mapping and proximity

Same proximity is used, but different geoms. Is one better than the other to determine the relative ratios of males to females by age?



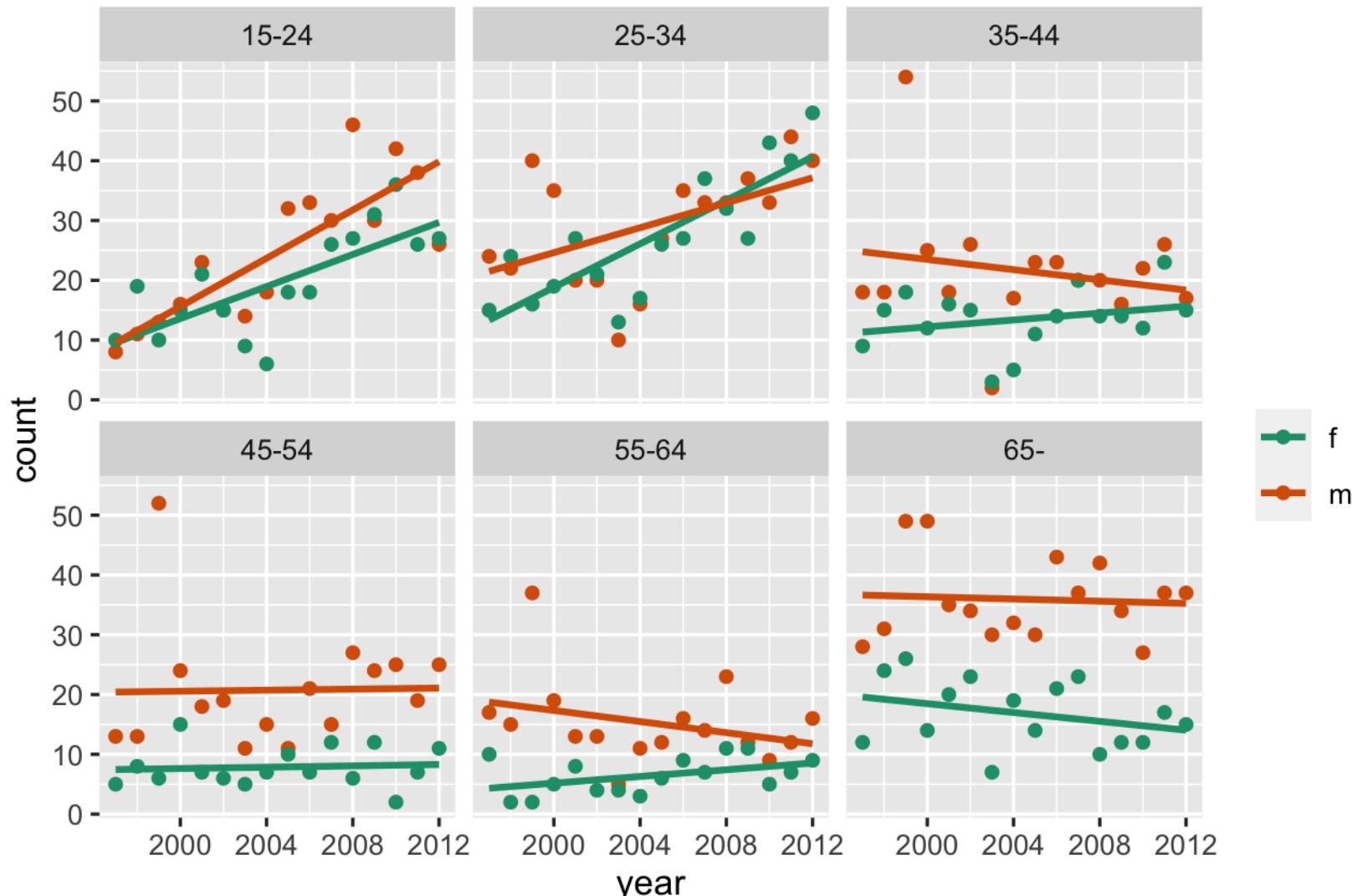
Mapping and proximity

Same proximity is used, but different geoms. Is one better than the other to determine the relative ratios of ages by sex?



Change blindness

Which has the steeper slope, 15-24 or 25-34 males?

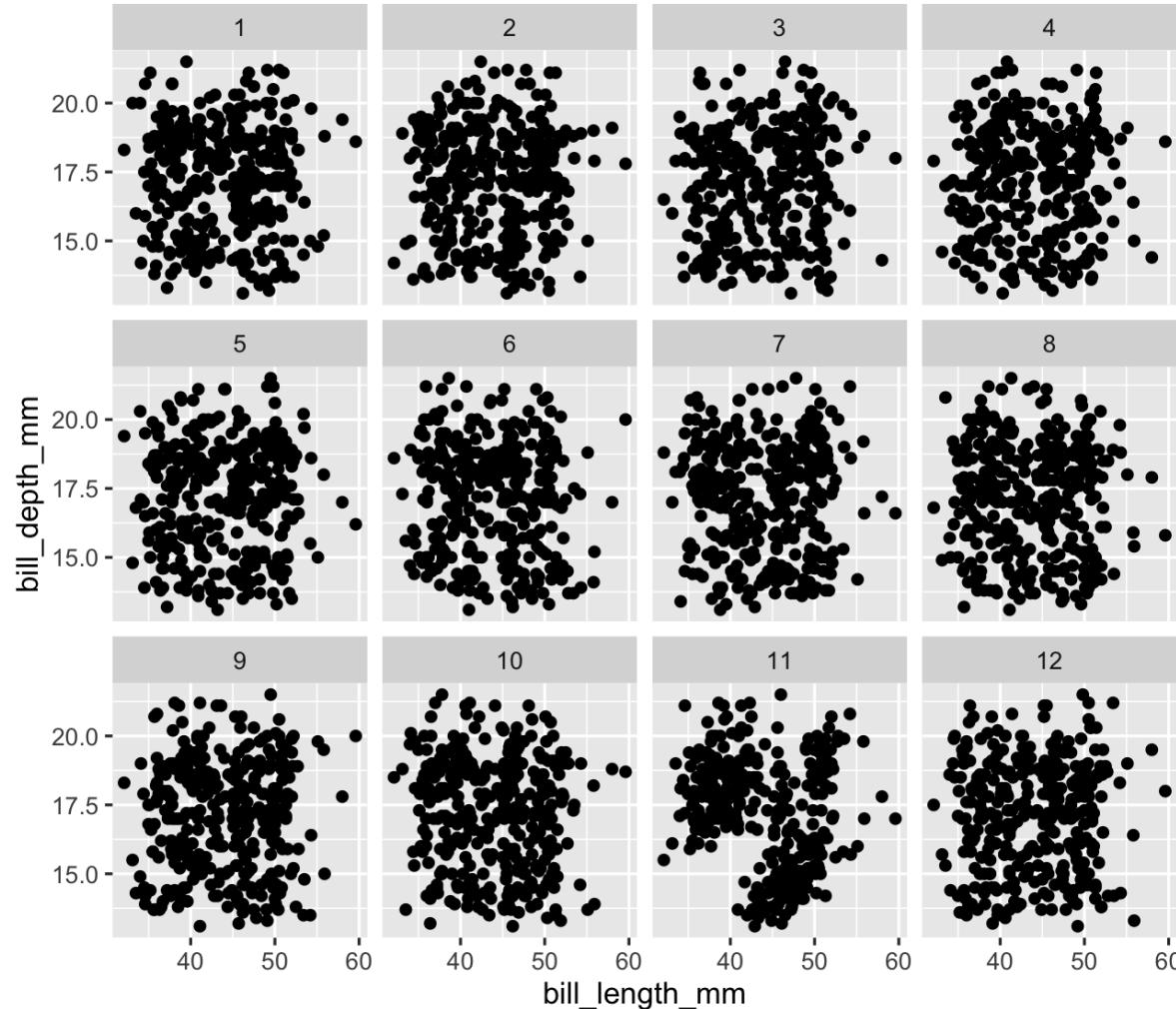


Change blindness

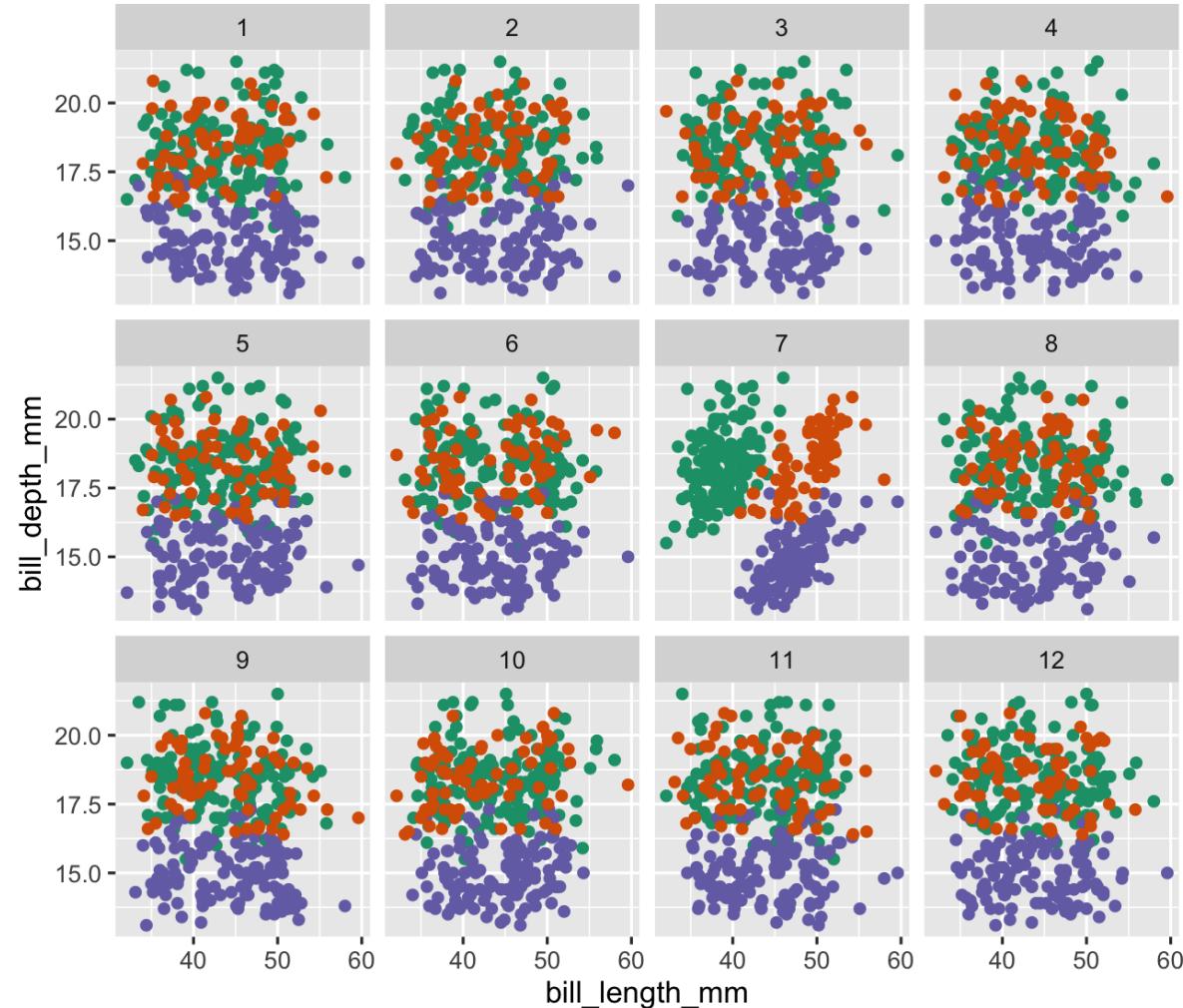
Which has the steeper slope, 15-24 or 25-34 males?

Making comparisons across plots requires the eye to jump from one focal point to another. It may result in not noticing differences.

Which one is different?



Which one is different?



Testing infrastructure

Both of these were quite easy. The testing procedure is called a lineup protocol:

1. Based on the grammar description of the plot, determine a null generating method (eg permute, simulate)
2. Generate many null plots, and embed your data plot randomly among them
3. Show to a good number of observers (two sample problem) and ask them to pick the plot that is different. (Crowd-sourcing can help.)
4. The plot type/style that has the larger proportion of observers detecting the data plot is the better design.

Resources

- Fundamentals of Data Visualization, Claus O. Wilke
- Hofmann, H., Follett, L., Majumder, M. and Cook, D. (2012) Graphical Tests for Power Comparison of Competing Designs,
<http://doi.ieeecomputersociety.org/10.1109/TVCG.2012.230>.
- Wickham, H., Cook, D., Hofmann, H. and Buja, A. (2010) Graphical Inference for Infovis, <http://doi.ieeecomputersociety.org/10.1109/TVCG.2010.161>.



</> Open day2-exercise-04.Rmd

15:00

Session Information

```
## - Session info --
##   setting  value
##   version  R version 4.0.1 (2020-06-06)
##   os        macOS Catalina 10.15.7
##   system   x86_64, darwin17.0
##   ui        X11
##   language (EN)
##   collate  en_AU.UTF-8
##   ctype    en_AU.UTF-8
##   tz       Australia/Sydney
##   date     2020-11-08
##
## - Packages --
##   package      * version  date     lib
##   anicon        0.1.0    2020-06-19 [1]
```

These slides are licensed under

