ETC1010: Data Modelling and Computing

Week of Data Visualisation: Lecture 3
Dr. Nicholas Tierney & Professor Di Cook
EBS, Monash U.
2019-08-14
Growing area of research by Carol Dweck of Stanford

- Reframe success + failure as opportunities for growth
- Growth and fixed mindsets

Understanding Learning
<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;I'll never understand&quot;</td>
<td>&quot;Just don't get programming&quot;</td>
</tr>
<tr>
<td>&quot;I understand more than I did yesterday&quot;</td>
<td>&quot;I can learn how to program&quot;</td>
</tr>
<tr>
<td>&quot;Compared to this last week, I've learnt quite a bit!&quot;</td>
<td>&quot;I'm not a maths person&quot;</td>
</tr>
</tbody>
</table>
Using different geometries

Mapping of variables from the data to graphical elements

Going from tidy data to a data plot, using a grammar

Overview for today
The case notifications data is tidied here, with only counts for Australia.

Data is tidied here, with only counts for Australia.

Example: Tuberculosis data from WHO.

<table>
<thead>
<tr>
<th>country</th>
<th>iso3</th>
<th>year</th>
<th>count</th>
<th>gender</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>AUS</td>
<td>1997</td>
<td>197</td>
<td>M</td>
<td>6</td>
</tr>
<tr>
<td>Australia</td>
<td>AUS</td>
<td>1998</td>
<td>198</td>
<td>M</td>
<td>7</td>
</tr>
<tr>
<td>Australia</td>
<td>AUS</td>
<td>1999</td>
<td>199</td>
<td>M</td>
<td>8</td>
</tr>
<tr>
<td>Australia</td>
<td>AUS</td>
<td>2000</td>
<td>200</td>
<td>M</td>
<td>9</td>
</tr>
<tr>
<td>Australia</td>
<td>AUS</td>
<td>2001</td>
<td>201</td>
<td>M</td>
<td>10</td>
</tr>
</tbody>
</table>
The "100% charts"

ggplot(tb.au, aes(x = year, y = count, fill = gender)) + 
geom_bar(stat = "identity", position = "fill") + 
facet_grid(~ age) + 
scale_fill_brewer(palette = "Dark2")
Let's unpack a bit.
"The simple graph has brought more information to the data analyst's mind than any other device."
— John Tukey

Data Visualization
Data Visualisation

- The creation and study of the visual representation of data.
- Many tools for visualizing data (R is one of them).
- Many approaches/systems within R for making data visualizations (ggplot2 is one of them, and that’s what we’re going to use).
ggplot2 is a tool that enables us to concisely describe the components of a grammar of graphics. It is inspired by the book Grammar of Graphics. The gg in "ggplot2" stands for grammar of package ggplot2 is tidyverse's data visualization tool. Source: ggplot2byLelandWilkinson†
Our first ggplot!
Our first ggplot!
Our first 88 plots (what's the data again?)

Country 1997
Year count gender age
Australia AUS 1998 18m
Australia AUS 1999 13m
Australia AUS 2000 16m
Australia AUS 2001 15m
Australia AUS 2002 15w
Australia AUS 2003 14m
Australia AUS 2004 18w
Australia AUS 2005 32w
Australia AUS 2006 33m

Your logging plant!
(was that?)
Our first ggplot!
library(ggplot2)

ggplot(tb_au, aes(x=year, y=count, fill=gender)) + geom_col()
Our first ggplot!

```r
library(ggplot2)
ggplot(tb_au, aes(x = year, y = count, fill = gender)) + geom_col(position = "fill")
```
library(ggplot2)
ggplot(tb_au, aes(x = year, y = count, fill = gender)) + geom_col(position = "fill") + scale_fill_brewer(palette = "Dark2")
Our first ggplot!
What do we learn

The "100% charts"
What do we learn?

- These proportions tend to be higher in the older age groups. For all years.
- Having TB is higher than females across (almost) all ages, and years. The proportion of males focus is on proportion in each category.
ggplot() is the main function

Code structure of ggplot

Structure of code for plots can often be summarized as

- Plots are constructed in layers

```r
ggplot(data = [dataset],
      mapping = aes(x = [x-variable],
                     y = [y-variable]),

      geom_xxx())

other options
```
For help with the ggplot2, see ggplot2.tidyverse.org

To use ggplot2 functions, first load tidyverse

How to use ggplot
Let's look at some more options to emphasise different features.
ggplot(tb_au, aes(x = year, y = count, fill = gender)) +
  geom_col(position = "fill") +
scale_fill_brewer(palette = "Dark2") +
facet_wrap(~ age)

27/46
ggplot(tb_au, aes(x = year, y = count, fill = gender)) + geom_col(position = "fill") + scale_fill_brewer(palette = "Dark2") + facet_grid(~ age)

Emphasizing different features with ggplot2
ggplot(tb_au, aes(x = year, y = count, fill = gender)) + geom_col() + scale_fill_brewer(palette = "Dark2") + facet_grid(~ age)
Recent years:

- Incidence has been increasing among younger age groups.
- Doubling or tripling other years.
- 1999 saw a bit of an outreach, in most age groups, with numbers (45-64).
- Different across ages, and years, counts tend to be lower in middle age.
- Focus is on counts in each category.

What do we learn?
ggplot(tb_au, aes(x = year, y = count, fill = gender)) + geom_col(position = "dodge") + scale_fill_brewer(palette = "Dark2") + facet_grid(~ age)
What do we learn?

- There is similar incidence between males and females in younger age groups.
- Incidence among males relative to females is from middle age on.
- Focus is on counts by gender, predominantly male incidence.
- `position = "dodge"` is used in `geom_col`.
Sear the bar charts

```r
ggplot(tb_au, aes(x = year, y = count, fill = gender)) + geom_col() + scale_fill_brewer(palette = "Dark2") + facet_grid(gender ~ age)
```
but seems to be have stabilized for males.
Growing incidence in the 25-34 age group is still affecting females.

- 1999 outbreak mostly affected males.
- Easier to focus separately on males and females.
- Note `facet-grid vs facet-wrap` gender as well as age

`facet-grid(gender ~ age) + facet.by`
Pie charts?

```r
tibble敖(tb_au,
    aes(x = year, y = count, fill = gender))
  + geom_col() +
  scale_fill_brewer(palette = "Dark2")
  + facet_grid(gender ~ age) +
  coord_polar() +
  theme(axis.text = element_blank())
```
What do we learn?

- Emphasizes the middle years as low incidence.
- Rather than the default Cartesian coordinates, `coord_polar()` + plot is made in polar coordinates.
- Bar charts in polar coordinates produce Rose charts.
ggplot(tb_au, aes(x = 1, y = count, fill = factor(year))) + geom_col(position = "fill") + facet_grid(gender ~ age)
What do we see in the code?

(That gives us the rainbow chart)

Year is now mapped to colour (that’s what gives us the rainbow to x; that makes a single stacked bar chart.)

Notice how the mappings are different: a single number is mapped to colour.

A single stacked bar, in each facet.

Year is mapped to colour.

Notice how the mappings are different: a single number is mapped to colour.
Pretty chart but not easy to interpret.

What do we learn?
```
print("(actual) Pie charts")
```
What is different in the code?

- coord_polar(theta="y")

is using the `y` variable to do the angles for the polar coordinates to give a pie chart.
What do we learn?

... comparisons across age groups.

Pretty chart but not easy to interpret, or make...
Using named plots, e.g., pie chart, bar

The various looks of David Bowie

Why?
Your Turn:

- Use the rest of the lab time to coordinate with your group on the first assignment.
- Take the lab quiz.
- Do the lab exercises.
References

- Mine Gethintya-Rundel’s Introduction to Ggplot using Star Wars.
- Garret Aden Buie’s Gentle Introduction to Ggplot2
- Data made available from WHO Chapter 3 of R for Data Science