MONASH University

Elegant graphics for data analysis with R

Presenter: *Emi Tanaka*

Department of Econometrics and Business Statistics

emi.tanaka@monash.edu @statsgen

Thu 2nd Sep 2021 | UOW Data and Decision Science Network

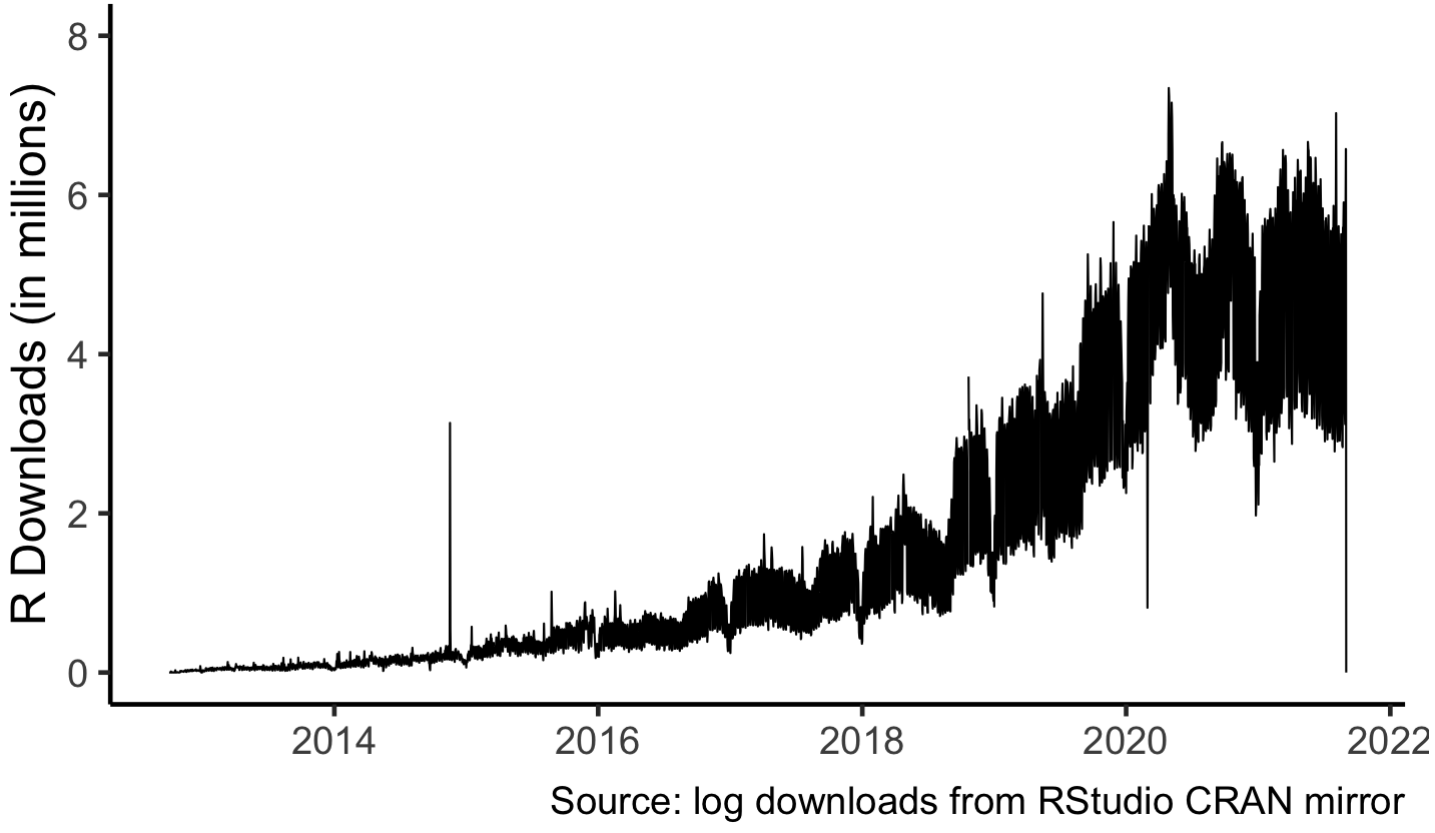
Emi Tanaka

Monash University
emi.tanaka@monash.edu
@statsgen

- I've been using R on-and-off for over 15 years
- Even as a long time R user, ***R has changed rapidly*** in the past decade
- Most of what I am talking to you about today is something I *learnt in the last 5 years*, so it's never too late to start learning R!

The number of R users is growing

R



R is one of the top programming languages

Note: unlike other languages, R is *not* a general purpose language but a programming language specifically for statistics (and data science)

R is *open source*

- The code within a function is visible

```
x <- c(17, 20, 21, 22, 24, 21, 19, 21, 19, 21, 13, 15, 24, 16, 17,
       21, 15, 24, 20, 17, 22, 20, 24, 14, 14, 22, 22, 16, 17, 19, 18,
       18, 21, 17, 24, 24, 24, 16, 25, 24)

fivenum(x)

## [1] 13 17 20 22 25
```

- You can inspect the code within a function by just typing its name:

```
fivenum

## function (x, na.rm = TRUE)
## {
##     xna <- is.na(x)
##     if (any(xna)) {
##         if (na.rm)
##             x <- x[!xna]
##         else return(rep.int(NA, 5))
```

Lastest statistical developments are often implemented in R first

There are over **18,100** contributed R-packages

Power of programming

- Learning to program is a steep learning curve!
- We throw our first year undergraduates into the deep end and they come out okay!
- **Investing in learning to program is worthwhile**
- Learning statistics in a point-and-click type of software, using tools that is hard to track actions (e.g. Excel), etc will *limit your growth*
- The world can rapidly change... you'll want the ability to use the latest and best statistical practice

Grammar of graphics

Wickham (2016) ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*

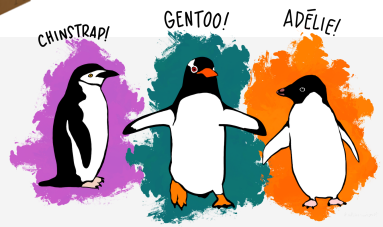
Basic structure of ggplot

```
ggplot(data = <data>, mapping = aes(<mappings>)) +
    <layer>()
```

1. **data** as `data.frame` (or `tibble`),
2. a set of **aesthetic** mappings between variables in the data and visual properties, and
3. at least one **layer** which describes how to render each observation.

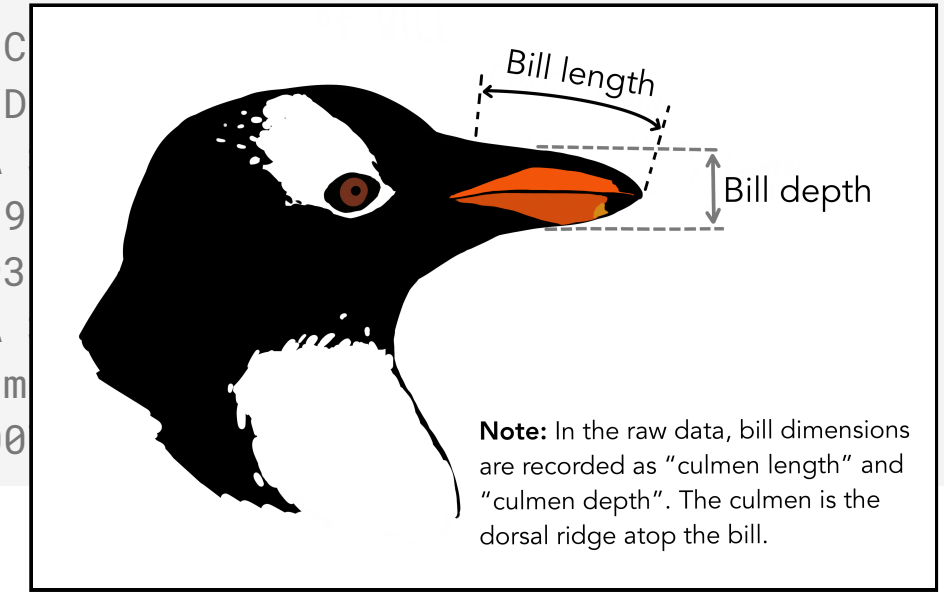
Palmer penguins

`penguins` data is from the `palmerpenguins`




```
library(palmerpenguins)
str(penguins)
```

```
## tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","C
##  $ island           : Factor w/ 3 levels "Biscoe","D
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA
##  $ sex              : Factor w/ 2 levels "female","m
##  $ year             : int [1:344] 2007 2007 2007 200
```



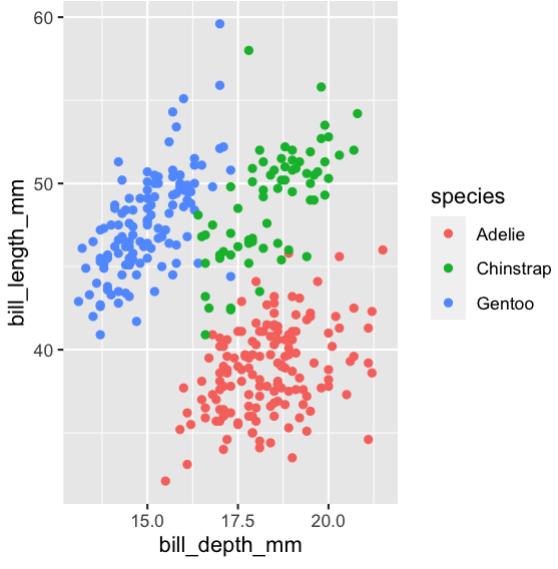
Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/

Gorman KB, Williams TD, Fraser WR (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE 9(3):e90081.

Aesthetic mappings

aesthestic = column in data

```
ggplot(data = penguins,
       mapping = aes(x = bill_depth_mm, y = bill_length_mm, color = species)) +
  geom_point()
```



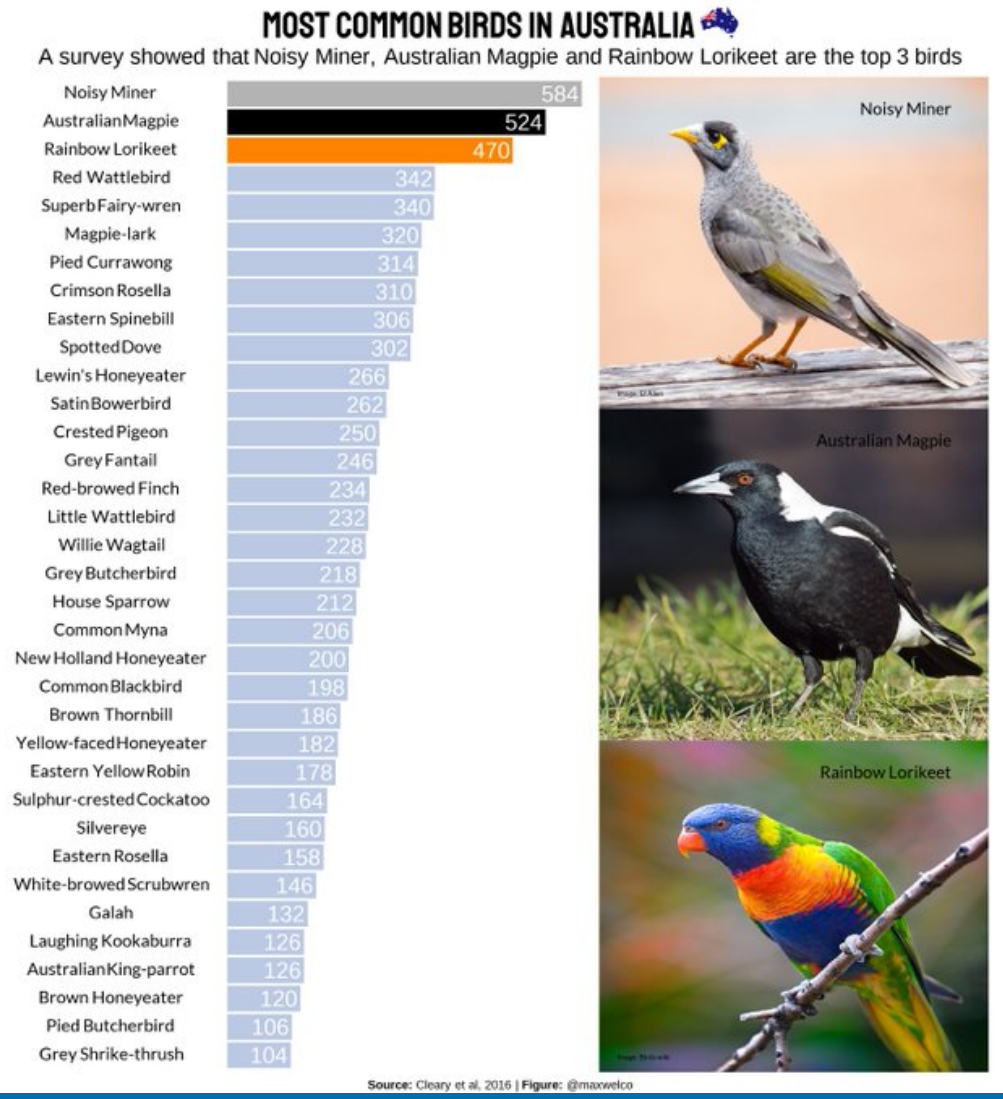

```
aes(x = bill_depth_mm, y = bill_length_mm, color = species)
```

year	sex	body_mass_g	flipper_length_mm	bill_depth_mm	bill_length_mm	island	species

- `bill_depth_mm` is mapped to the `x` coordinate
- `bill_length_mm` is mapped to the `y` coordinate
- `species` is mapped to the `color`

Bird bath survey in Australia

- Clearly et al. (2016) study collected data from 2,500 citizen scientists on bathing birds all over Australia
- The data accompanying the article is available here
- The cleaned up version of this data is available here
- The plot on the right is made by Maxwel C. Oliveira using this data



Wrangling data

data wrangling plot setup

`bird_baths` data

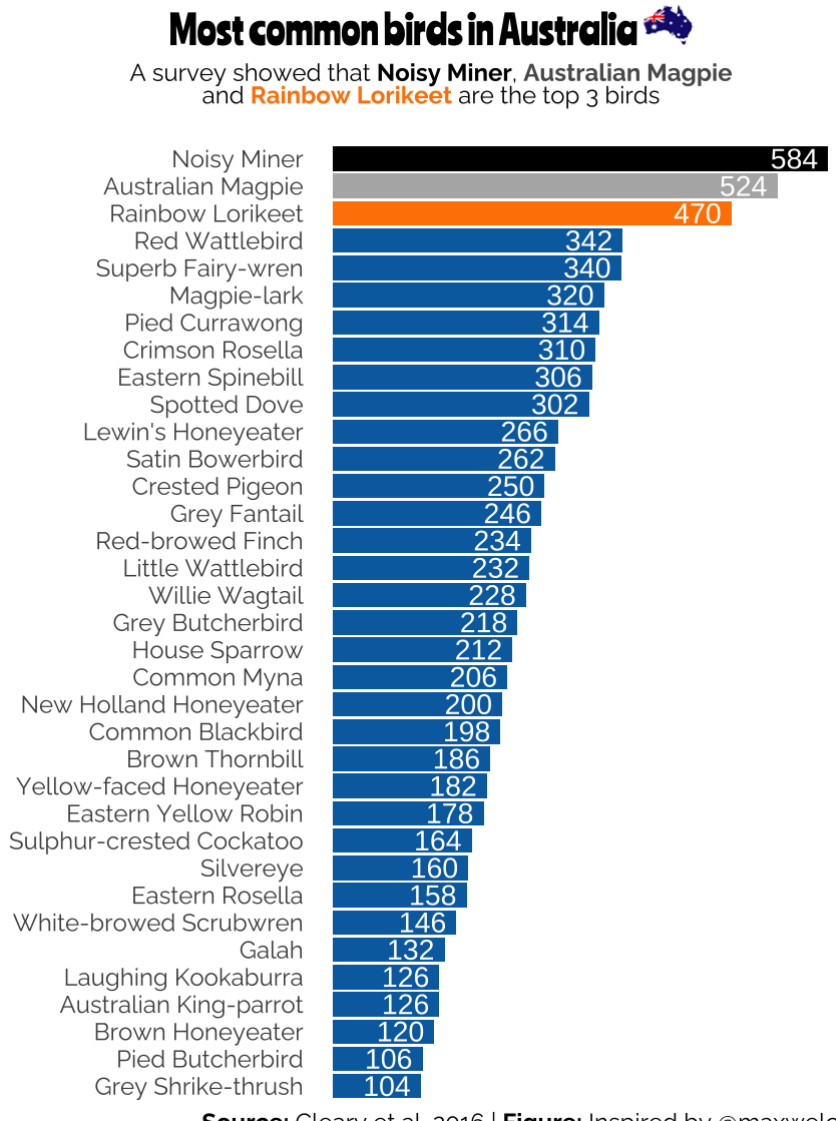
```
## 'data.frame':    161057 obs. of  5 variables:
##  $ survey_year: int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
##  $ urban_rural: chr  "Urban" "Urban" "Urban" "Urban" ...
##  $ bioregions : chr  "South Eastern Queensland" "South Eastern Queensland" "South Eastern Que
##  $ bird_type  : chr  "Bassian Thrush" "Chestnut-breasted Mannikin" "Wild Duck" "Willie Wagtai
##  $ bird_count : int  0 0 0 0 0 0 0 0 0 0 ...
```

`bird_df` data

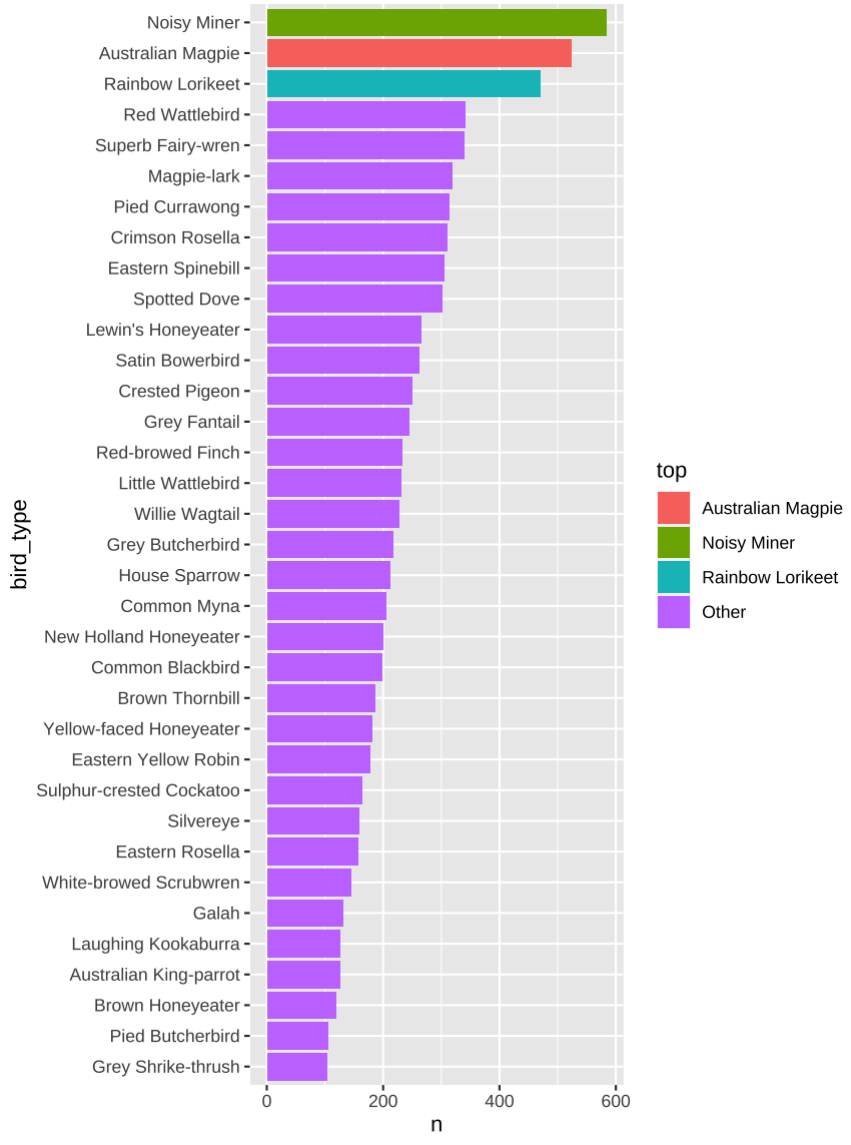
```
## # A tibble: 6 × 3
##   bird_type                n top
##   <fct>                <int> <fct>
## 1 Australian King-parrot   126 Other
## 2 Australian Magpie        524 Australian Magpie
## 3 Brown Honeyeater         120 Other
## 4 Brown Thornbill          186 Other
## 5 Common Blackbird         198 Other
## 6 Common Myna              206 Other
```

Creating reproducible publication-ready plots

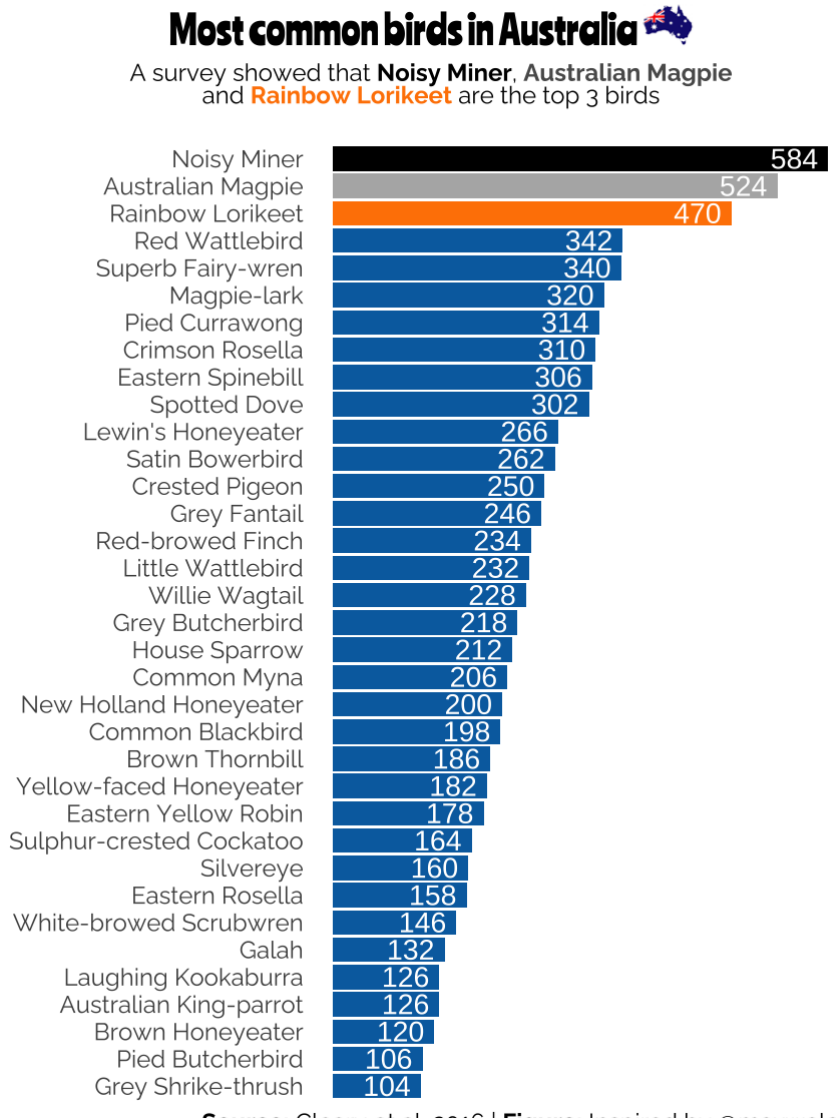
```
ggplot(bird_df, aes(x = n, y = bird_type, fill = top)) +
  geom_col() +
  scale_fill_manual(values = c("gray70", "black", "#FF8300",
  labs(x = NULL,
       y = NULL,
       title = "Most common birds in Australia <img src='htt
       subtitle = "A survey showed that **Noisy Miner**, <b
       caption = "**Source:** Cleary et al, 2016 | **Figure:
  geom_text(aes(label = n),
            nudge_x = -40,
            color = "white",
            size = 5) +
  theme(text = element_text(family = "Raleway"),
        legend.position = "none",
        axis.text.y = element_text(size = 12),
        axis.text.x = element_blank(),
        panel.background = element_blank(),
        axis.ticks.length = unit(0, "mm"),
        plot.title = element_markdown(size = 20,
                                      family = "Ranchers",
                                      hjust = 0.5),
        plot.subtitle = element_markdown(size = 12,
                                         hjust = 0.5),
        plot.caption = element_markdown(size = 12, hjust = 1
        plot.title.position = "plot")
```



Before



After



Source: Cleary et al, 2016 | **Figure:** Inspired by @maxwelco

Resources & Community for Learning R

- Australian Local R-Ladies Chapters:
  - R-Ladies Melbourne
  - R-Ladies Sydney
  - R-Ladies Canberra
  - R-Ladies Perth
  - R-Ladies Adelaide
  - R-Ladies Brisbane
- R for data science community
- Tidy Tuesday
- Big Book of R

Online Data Visualisation with R Workshops

Statistical Society of Australia

Coming in early December:

- **Data Visualisation with R** hosted by the Statistical Society of Australia NSW Branch

  Tentatively scheduled for full day on Mon 6th Dec

- **Advanced Data Visualisation with R** hosted by the Statistical Society of Australia Canberra Branch

  Tentatively scheduled for half days on Wed-Thu 8th-9th Dec

by Dr Emi Tanaka and Professor Dianne Cook

SSA Events

For full list of events, check out: 🔗 https://www.statsoc.org.au/Events-listing

MONASH University

You can find this slide at

emitanaka.org/plotting-with-R

and for those interested in the code,

github.com/emitanaka/plotting-with-R

Presenter: *Emi Tanaka*

Department of Econometrics and Business Statistics

emi.tanaka@monash.edu @statsgen

Thu 2nd Sep 2021 | UOW Data and Decision Science Network