Class 1: An introduction to data visualisation

Andrew Parnell
andrew.parnell@mu.ie

PRESS RECORD https://andrewcparnell.github.io/dataviz_course

Let's get started

- About me
- Tell me:
  - who you are,
  - where you are from,
  - your previous experience in working with R and visualisation,
  - what you are working on,
  - what you want to get out of the course,
- Timetable for the course
- Pre-requisites

How this course works

- This course lives on GitHub, which means anyone can see the slides, code, etc, and make comments on it
- The timetable document (`index.html`) provides links to all the pdf slides and practicals
- The slides and the practicals are all written in `Rmarkdown` format, which means you can load them up in Rstudio and see how everything was created
- Let me know if you spot mistakes, as these can be easily updated on the GitHub page
- There is a `dataviz_course.Rproj` R project file from which you should be able to run all the code

Copyright statement

All the non-GitHub materials provided in the Introduction to Data Visualisation are copyright of Andrew Parnell and Catherine Hurley.

This means:

- As a user (the student) you have permission (licence) to access the materials to aid and support your individual studies.
- You are not permitted to copy or distribute any materials without the relevant permission
- As faculty we may reserve the right to remove a user in the event of any possible infringement

Course format and other details

- Lectures will take place in the morning via Zoom, practical classes in the afternoon
- In the practical classes I will go round the room asking people how they are getting on
- If you want to send me a private message use Slack
- Please ask lots of questions, but **MUTE YOUR MICROPHONE** when not asking them
- Some good resources:
  - Data Visualisation (chapter) by Hadley Wickham
  - ggplot2 reference guide
  - The psychology of data visualisation by Michael Friendly
  - Philosphy of visualisation by Hadley Wickhame

Why visualise data?

```
library(datasauRus)
datasaurus_four <- datasaurus_dozen %>%
  filter(str_detect(
    dataset,
    "(dino|bullseye|star|x_shape)"
  ))
datasaurus_four %>% summary()
```

```
##    dataset                  x                y
##  Length:568         Min.   :19.29   Min.   : 2.949
##  Class :character   1st Qu.:40.93   1st Qu.:23.429
##  Mode  :character   Median :52.96   Median :45.390
##                     Mean   :54.26   Mean   :47.836
##                     3rd Qu.:67.37   3rd Qu.:71.101
##                     Max.   :98.21   Max.   :99.487
```

Summarise as a table...

```
datasaurus_four %>%
  group_by(dataset) %>%
  summarize(
    mean_x = mean(x),
    mean_y = mean(y),
    std_dev_x = sd(x),
    std_dev_y = sd(y),
    corr_x_y = cor(x, y)
  )
```

```
## # A tibble: 4 x 6
##   dataset  mean_x mean_y std_dev_x std
##   <chr>     <dbl>  <dbl>     <dbl>
## 1 bullseye   54.3   47.8      16.8
## 2 dino       54.3   47.8      16.8
## 3 star       54.3   47.8      16.8
## 4 x_shape    54.3   47.8      16.8
```

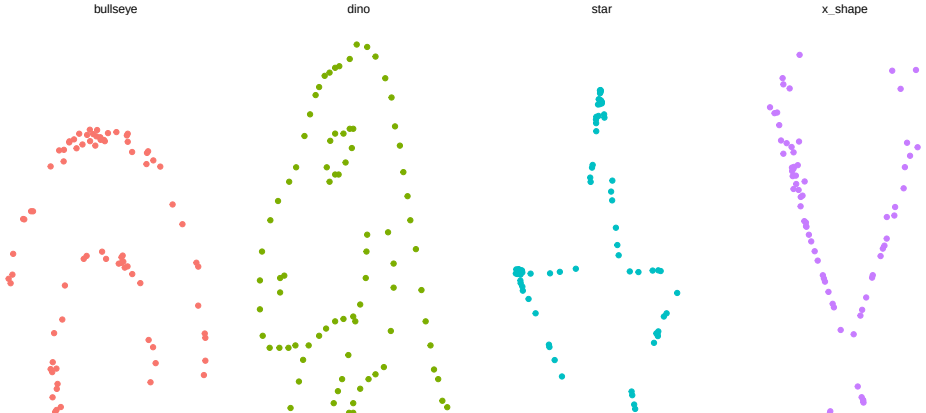
Do some linear regressions?

```
datasaurus_four %>%
    group_by(dataset) %>%
    summarize(
      intercept = lm(y ~ x)$coefficients[1],
      slope = lm(y ~ x)$coefficients[2]
    )
```

```
## # A tibble: 4 x 3
##   dataset  intercept  slope
##   <chr>        <dbl>  <dbl>
## 1 bullseye      53.8 -0.110
## 2 dino          53.5 -0.104
## 3 star          53.3 -0.101
## 4 x_shape       53.6 -0.105
```

. . . or as a plot

```
ggplot(datasaurus_four, aes(x = x, y = y, colour = dataset)) +
  geom_point() +
  theme_void() +
  theme(legend.position = "none") +
  facet_wrap(~dataset, ncol = 4)
```



Introduction to example data sets used in the course:

- All used data sets in this course are either in the `data` directory or come from packages
- The `data_desciption.txt` file in the data directory contains a list of all the fields and references
- Look at the help file for the data sets that come from packages to see the full list of fields

palmerpenguins

```
library(palmerpenguins)
penguins %>% glimpse()
```

```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1,
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1,
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475,
## $ sex               <fct> male, female, female, NA, female, male, female, male
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007
```

Swiss Willow Tits

3 replicate measurements on whether Swiss Willow Tits were found with covariates on forest cover and elevation

```
##   rep.1 rep.2 rep.3 c.2 c.3 elev forest dur.1 dur.2 dur.3 length alt
## 1     0     0     0   0   0  420      3   240    58    73    6.2 Low
## 2     0     0     0   0   0  450     21   160    39    62    5.1 Low
## 3     0     0     0   0   0 1050     32   120    47    74    4.3 Med
## 4     0     0     0   0   0 1110     35   180    44    71    5.4 Med
## 5     0     0     0   0   0  510      2   210    56    73    3.6 Low
## 6     0     0     0   0   0  630     60   150    56    73    6.1 Low
```

- How do the covariates affect the chance of finding the birds?
- Are these effects linear?
- What do we do with the missing data?

Palaeoclimate pollen data

A set of modern pollen counts and their associated climates. The variables are: GDD5 (Growing degree days about 5C), MTCO (Mean temperature of the coldest month), pollen counts of taxa (Abies - Graminaea).

```
##   GDD5 MTCO Abies Alnus Betula Picea Pinus.D Quercus.D Gramineae
## 1 1874 -7.9     0    50    158     7     721        22         0
## 2 1623 -5.5     0    38     28   302     537        19         0
## 3 1475 -4.7     0   276    183   110     136         0         0
## 4 1360 -8.8     0   111    354   141     364         0         0
## 5 1295 -6.9     0    91     50   151     708         0         0
## 6 1539 -7.8     0    51    194    82     673         0         0
```

- How are pollen species affected by these climate variables?
- Are these effects linear?
- Are their relationships between the pollen taxa?

A checklist for data visualisation

- What is the message you are trying to convey?
- What medium will the visualisation be displayed in (paper/poster/screen/interactive/...)?
- How much space do you have?
- How much explanation can you give to accompany the visualisation?
- What size will the visualisation be?
- Will colour and transparency be allowed?
- How long will people spend looking at the visualisation?

Reminder of basic data types and their influence on visualisation tools

Lots of different categorisations of data but the most important ones are:

- Categorical data (e.g. names)
- Ordinal data (e.g. agreement levels)
- Continuous data (e.g. height in cm)

Visualisations usually involve multiple variables, often of different types

The data type will often strongly guide the choice of visualisation

Some basic plot types: 1 Bar charts

Some basic plot types: 2 Histograms

Some basic plot types: 3 Boxplots

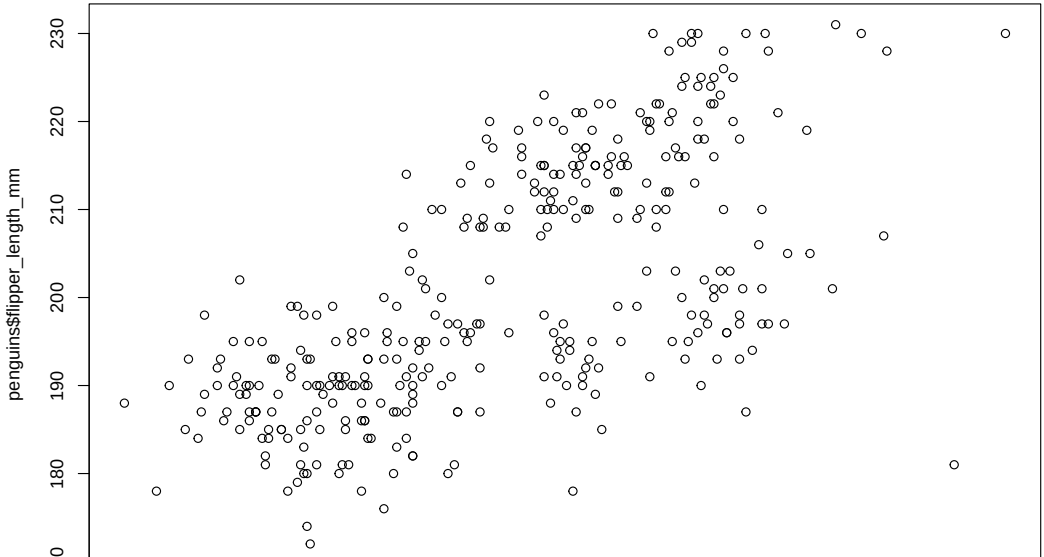
Some basic plot types: 4 Scatter plots

10 basic rules for plotting

1. Avoid ink that isn't representing the data
2. Avoid outlines; try to fill things in if you can. Use transparency
3. Don't use 3D visualisations
4. Don't use pie charts. Ever
5. Write clear and informative captions/titles appropriate for the medium
6. Try to label interesting features of a plot directly
7. Use colour carefully; try to avoid too many colours
8. If you must have a legend then think carefully about the labelling
9. Use small number of multiple figures (facets)
10. Think about the units of the axes - should a plot be square or rectangular?

Exercise: try and list 5 things that are bad about the following plot

```
plot(penguins$bill_length_mm, penguins$flipper_length_mm)
```



Summary

- X
- Y
- Z