University of Glasgow | College of Medical, Veterinary & Life Sciences

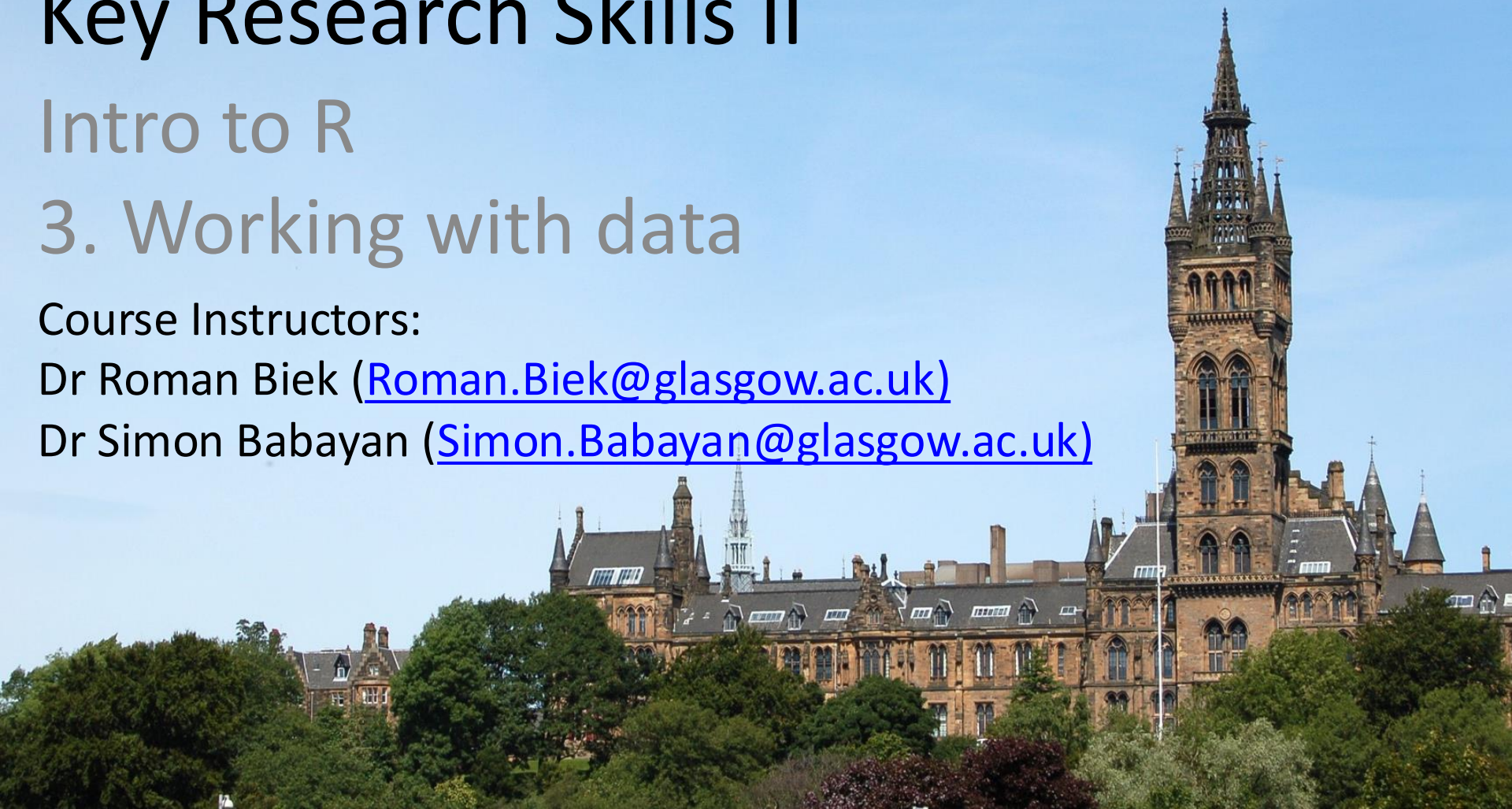
School of Biodiversity, One Health, and Veterinary Medicine

Key Research Skills II

Intro to R

3. Working with data

Course Instructors:
Dr Roman Biek (Roman.Biek@glasgow.ac.uk)
Dr Simon Babayan (Simon.Babayan@glasgow.ac.uk)



Review of second day

Manipulating data with dplyr

select()

slice()

filter()

mutate()

arrange()

Review of second day

Plotting with ggplot2

name of data object

aesthetics mapping (which variables)

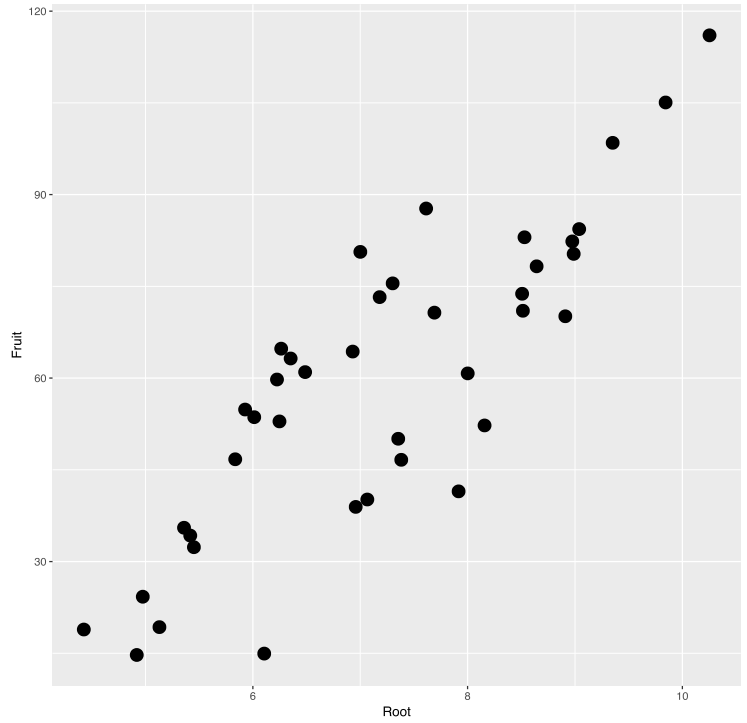
```
ggplot(compensation, aes(x = Root, y = Fruit)) +
    geom_point(size = 5)
```

geom (type of graph)

options

Many different geoms available e.g.

```
geom_line()
geom_boxplot()
geom_histogram()
geom_bar()
```

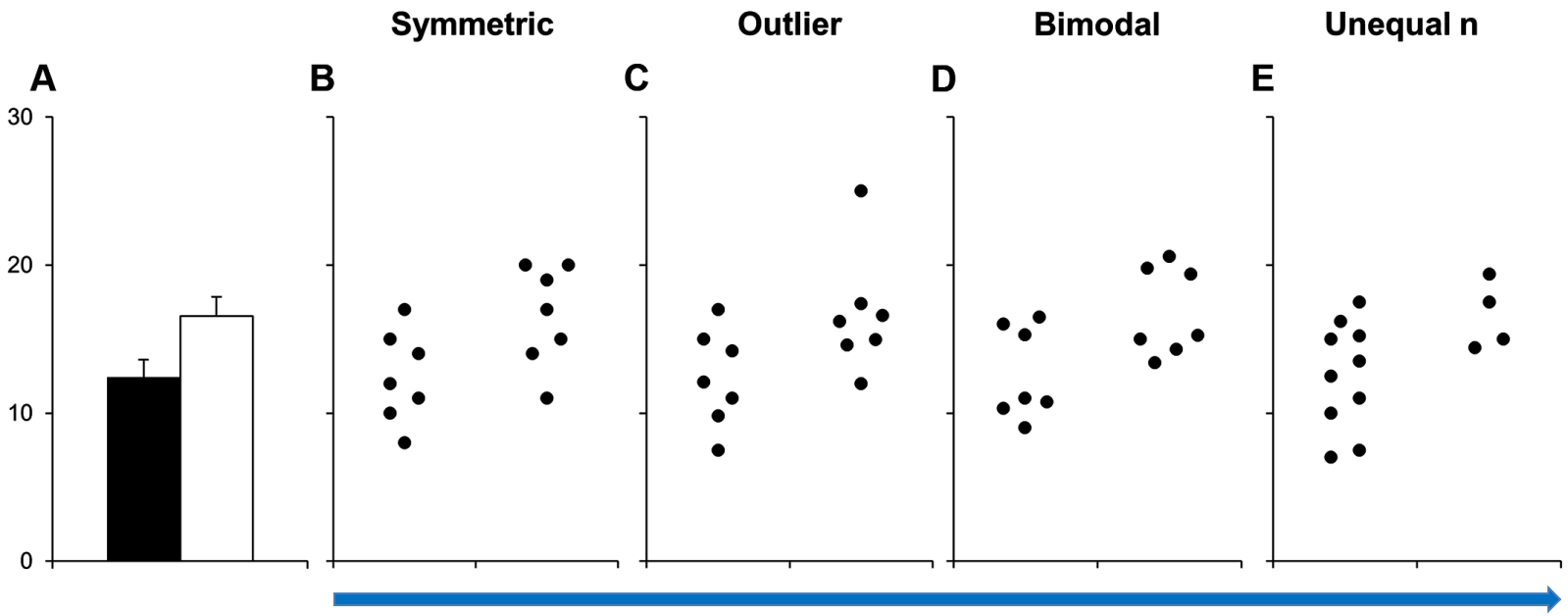


Bar plots under criticism

Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm

Tracey L. Weissgerber, Natasa M. Milic, Stacey J. Winham, Vesna D. Garovic

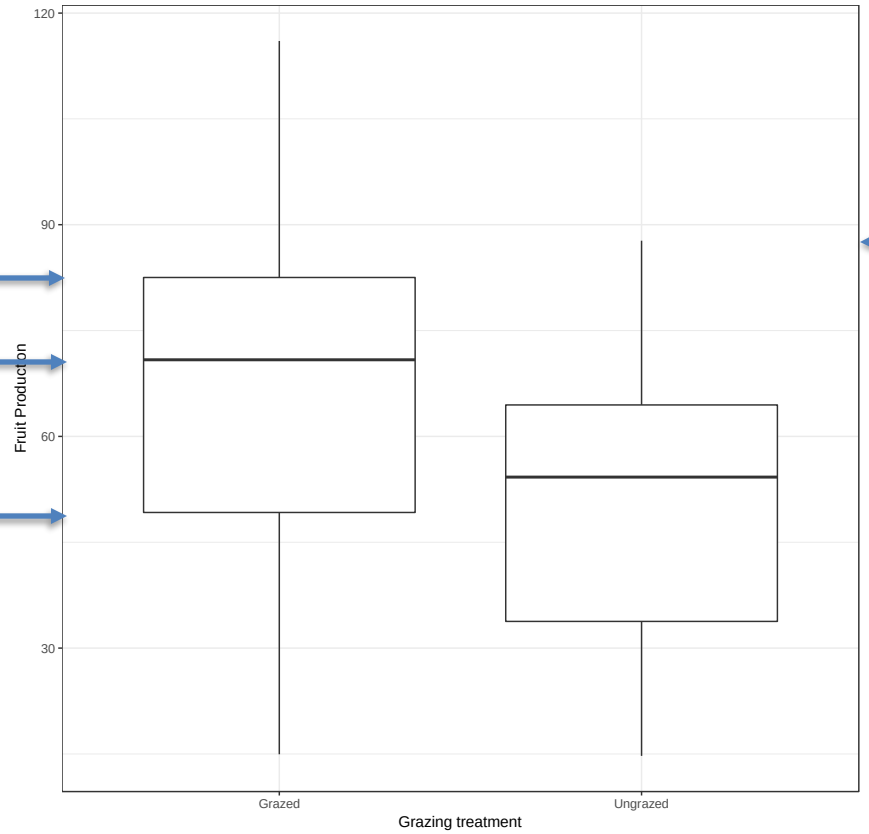
Published: April 22, 2015 • http://dx.doi.org/10.1371/journal.pbio.1002128



All of these observations result in the same bar plot and error bars on the left

Alternative to barplots: boxplots

```
ggplot(compensation, aes(x = Grazing, y = Fruit)) +
    geom_boxplot() +
    geom_point(size = 4, colour = 'lightgrey', alpha = 0.5) +
    xlab("Grazing treatment") +
    ylab("Fruit Production") +
    theme_bw()
```



Third quartile (75th percentile)

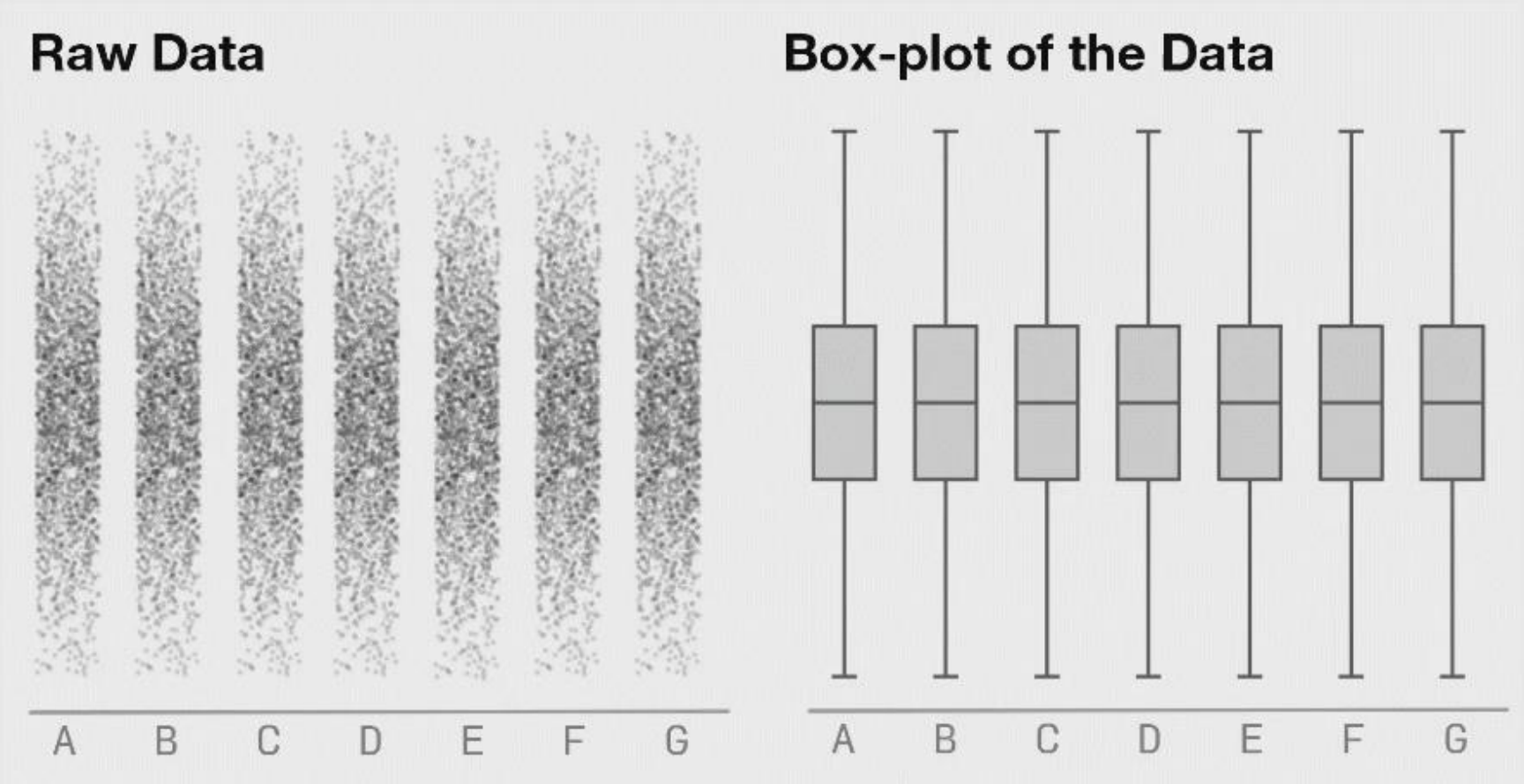
median

First quartile (25th percentile)

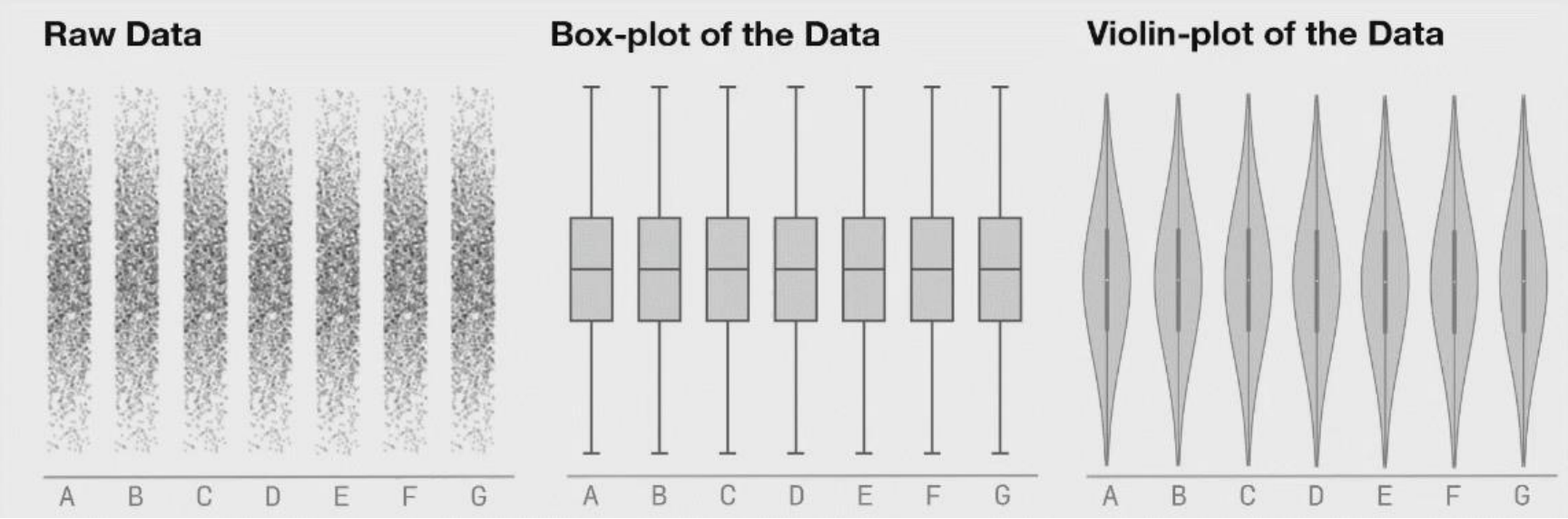
"whiskers" largest value up to 1.5x the interquartile range

Beyond that, values are flagged up as outliers

Even boxplots are not showing the full picture



Ideally, choose method that shows distribution



Saving graphs

Graphs in plots window in R Studio might look different from figure in html report => ignore plots window and optimise code for report

In html reports, figures are saved at default size of 5x7 inches

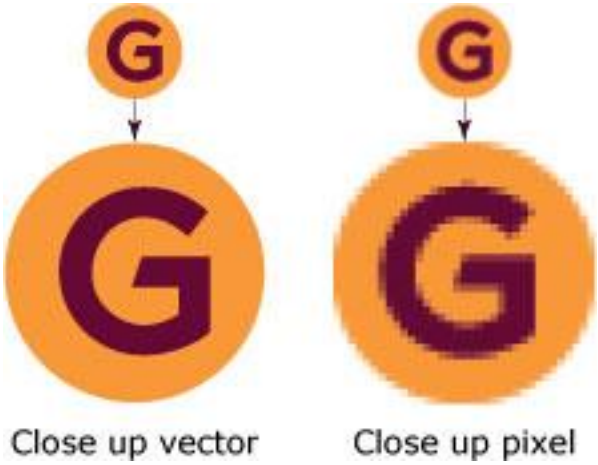
To create standalone figure (usually not part of your script) you can use

```
ggsave(file="myFigure.pdf")
ggsave(file="myFigure.png")
```

Extension specifies what kind of file you want

possible to change size, resolution, file type, etc

Vector- vs. raster based images



- Saving graphs in vector-based formats (pdf, svg, postscript or eps) preserves quality, even at high resolution
- Raster-based file formats (png, tiff, jpeg, bitmap) can result in pixilated images

Recommended reading on graphics: Section V in 'Practical computing for biologists' by Haddock & Dunn (Sinnauer Associates, 2011)

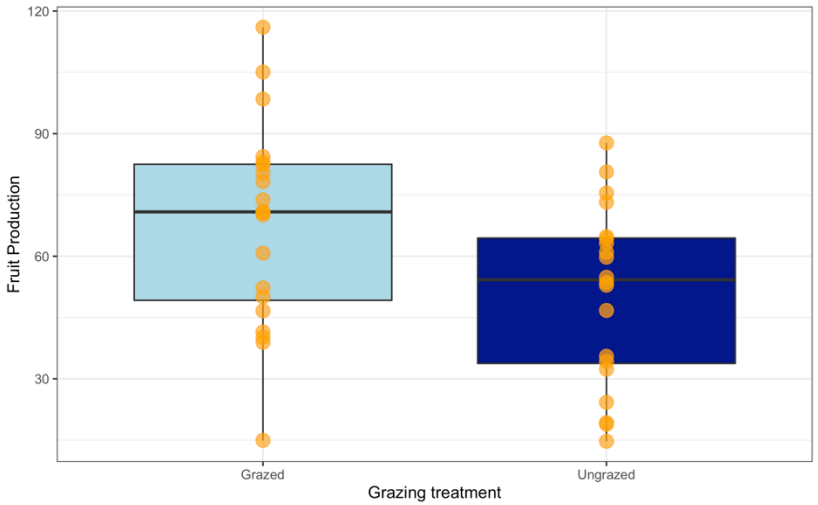
Colours

- Endless opportunities but probably best to keep things simple
- Can specify colours by name (e.g. “black”) or numbers 1 through 8
- ggplot makes colour choices for you
- See *R.Colours.pdf* on Moodle for full list
- Check out package and website ‘colorspace’ http://colorspace.r-forge.r-project.org - provides useful colour palettes including options that are colour-blind safe

Assigning colours

- Fill = for boxplots, barplots, dotplots
- Colour = for lines, points
- Example:

```
ggplot(compensation, aes(x = Grazing, y = Fruit)) +
    geom_boxplot(fill = c("lightblue", "darkblue")) +
    geom_point(size = 4, colour = 'orange', alpha = 0.7) +
    xlab("Grazing treatment") +
    ylab("Fruit Production") +
    theme_bw()
```



Advancing skills with ggplot2

- Skip forward to Chapter 8 of the book
  - Explains how to customize plots with scales and themes

- Online courses - type “free online ggplot2 courses” in your search engine

- Look at cheatsheet in R Studio (Help > Cheatsheets)

Fundamentals of data visualization

Great book by Claus Wilke, freely available online

https://serialmentor.com/dataviz/

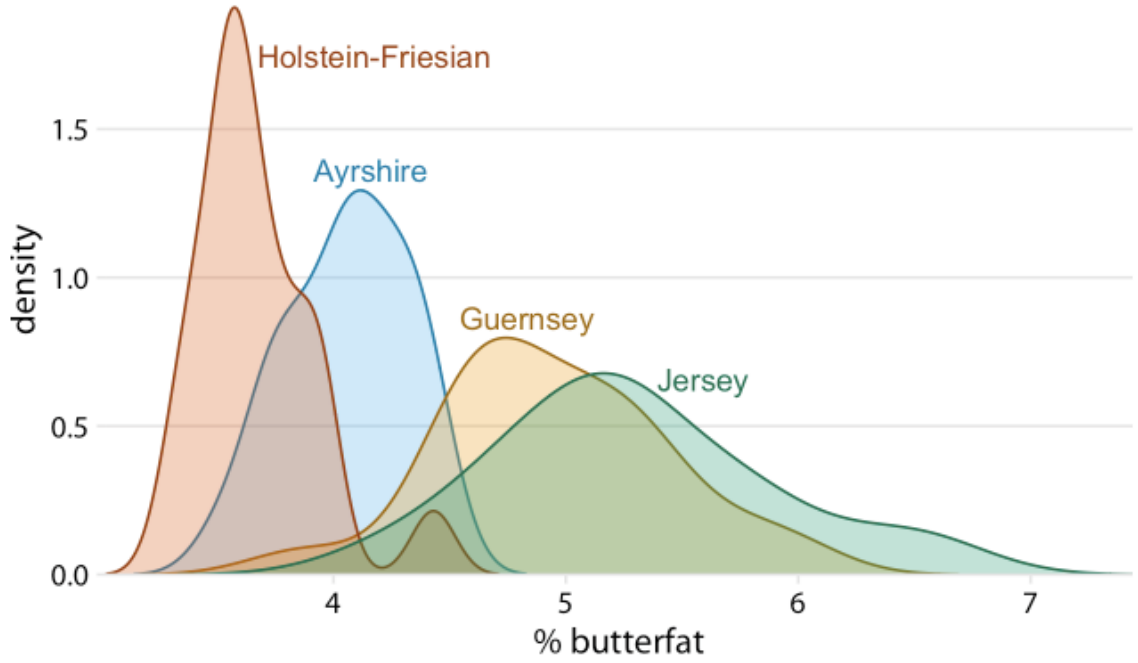


Figure 7.11: Density estimates of the butterfat percentage in the milk of four cattle breeds. Data Source: Canadian Record of Performance for Purebred Dairy Cattle

Different classes of R objects

- So far, you've encountered
  - **Data frame**: two dimensions (rows/columns), can be mix of numbers and text, so most flexible
  - **Tibble**: modern form of data frame (by H Wickham)
- Many other classes exist e.g.
  - **Matrix**: also two dimensions but only allows one type of data (text or numbers) => special type of **array**, which can have one, two, or more dimensions
- Use `class()` to interrogate what your object is

Compiling html reports

- Check your report before submitting
- Find a formatting style that works for you and stick to it consistently throughout the script
- Ideally, formatting should work well within both the script and the report e.g.

```
# ------------------------------------------------------------
#' **ENTERING AND EXPLORING THE DATA**
# ------------------------------------------------------------
```

Bold text in report

Only shows up in the script

Keeping up to date about R

Useful: https://www.r-bloggers.com/

Option to subscribe to daily updates

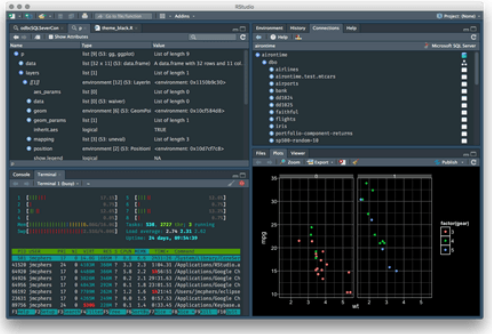
[R-bloggers] RStudio v1.1 Released (and 6 more aRticles)

- **RStudio v1.1 Released**
- **Genome-wide association studies in R**
- **Celebrating 20 years of CRAN**
- **[summer Astrostat school] room with a view [jatp]**
- **How to sample from multidimensional distributions using Gibbs sampling?**
- **mea culpa!**
- **A step change in managing your calendar, without social media**

RStudio v1.1 Released

Posted: 09 Oct 2017 10:30 AM PDT

(This article was first published on **RStudio Blog**, and kindly contributed to **R-bloggers**)



another great resource

https://stackoverflow.com/

stackoverflow

Today's exercise: bird ticks dataset

Download instructions and dataset on Moodle.

Consolidating familiar concepts while adding new tool for manipulating data and visualising data:

- redefining data columns or adding new ones
- ordering
- creating subsets
- dealing with missing values
- indexing
- more plotting

Assignments

Finish today's exercise.

By **Monday (14th Oct, midnight)** upload your

- R script (.R) or notebook (.Rmd)
- html report

Break(s)

- Take a break now! … or whenever you need one.