THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

From nothing to something much more with RStudio

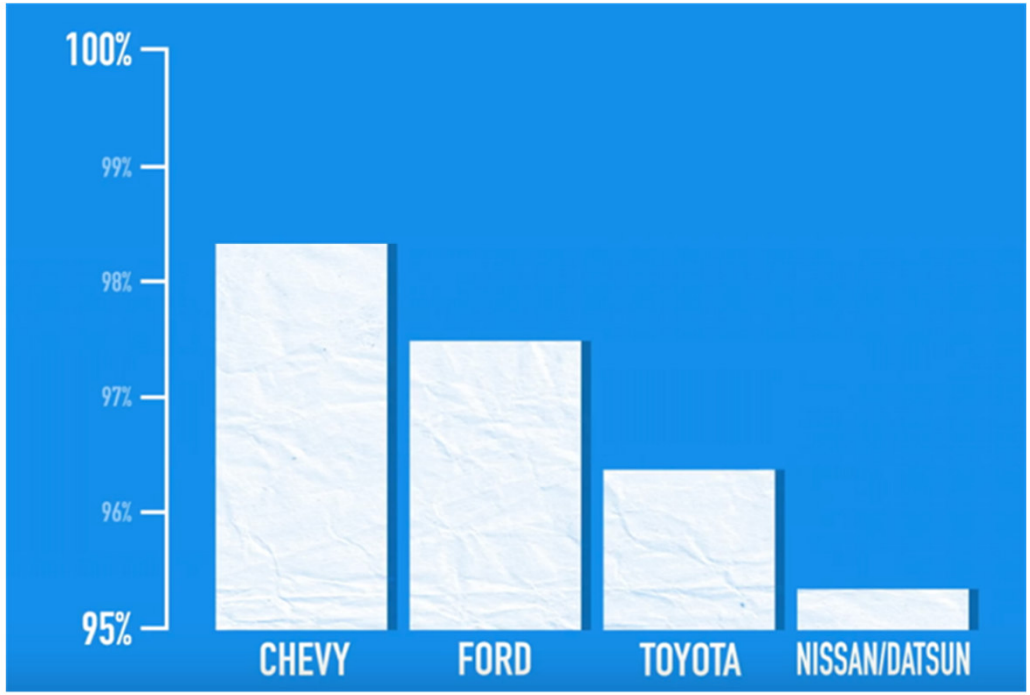
Lucia Michielin,
Digital Skills Training Manager

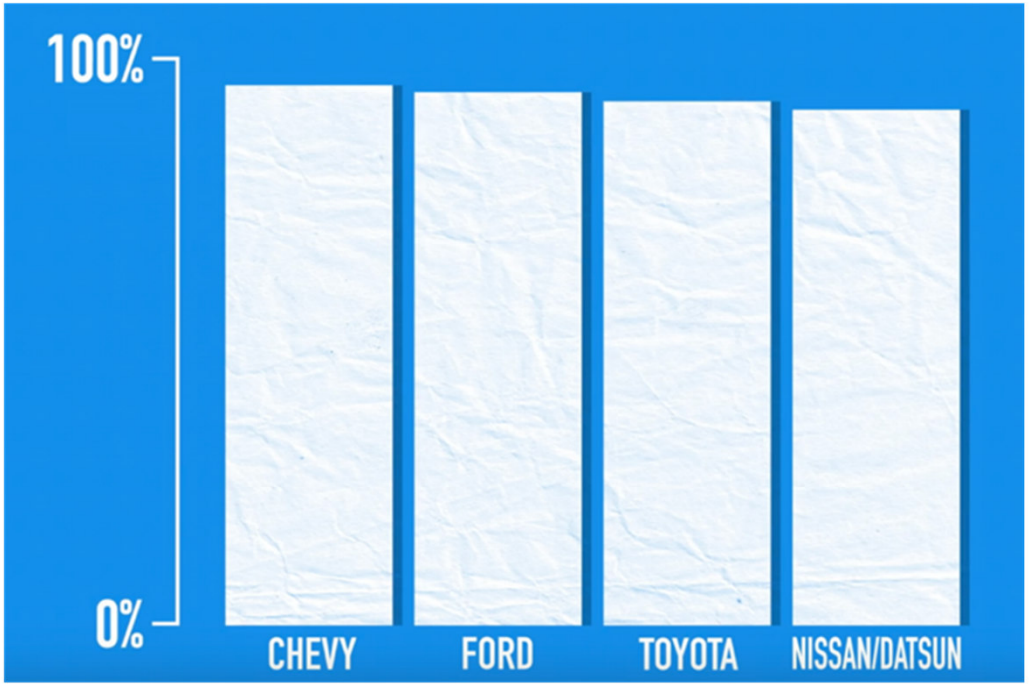
Aislinn Keogh,
CDCS Training Fellow

Data Visualisation - Telling a story

- Building a graph is like telling a story
- Good story telling/bad story telling
- 20 second rule
- Reading Suggestion
  - **How Charts Lie** by Alberto Cairo (also the Truthful Art and The Functional Art)
  - **Data Visualisation** by Andy Kirk
  - **Am I Overthinking This?** Michelle Rial

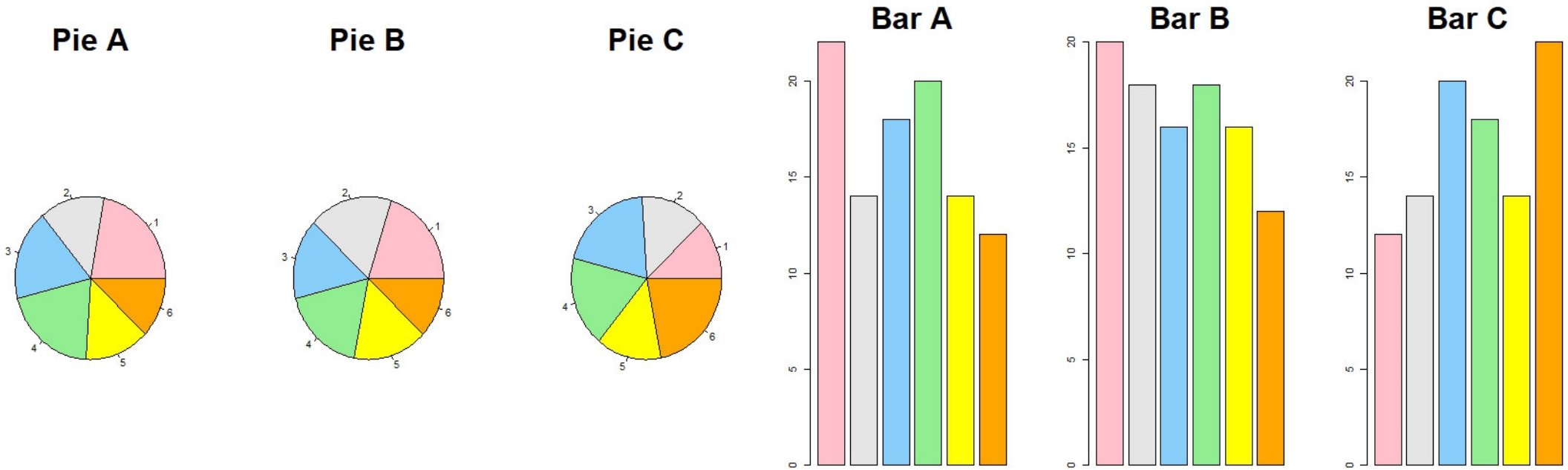
How Charts Lie





From Cases and Tools in Biotechnology Management
Author: Trent Tucker

How Charts Lie



Can you easily rank the blue vs the green changes?

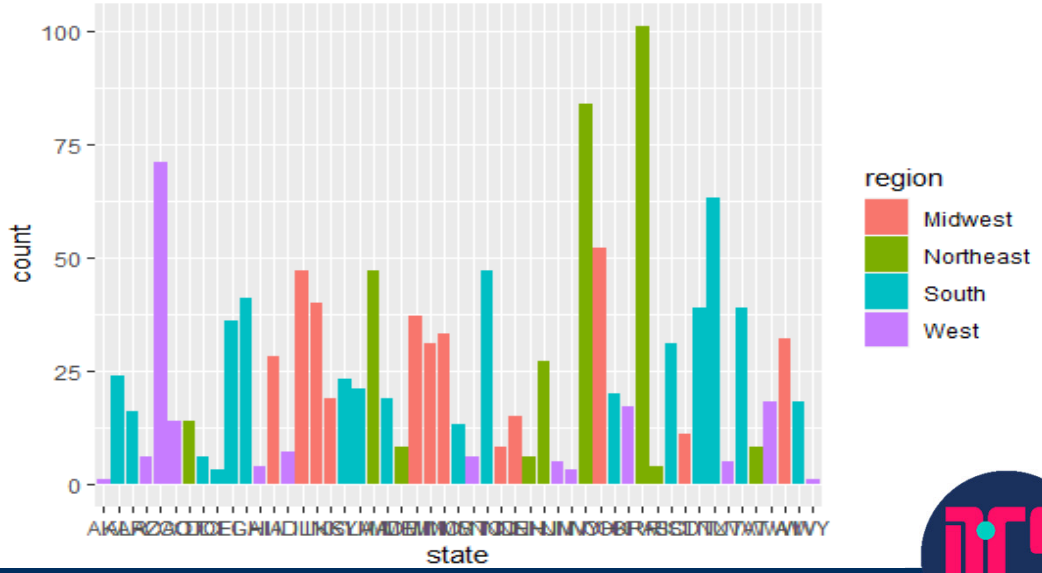
Before plotting…think about what you want to convey

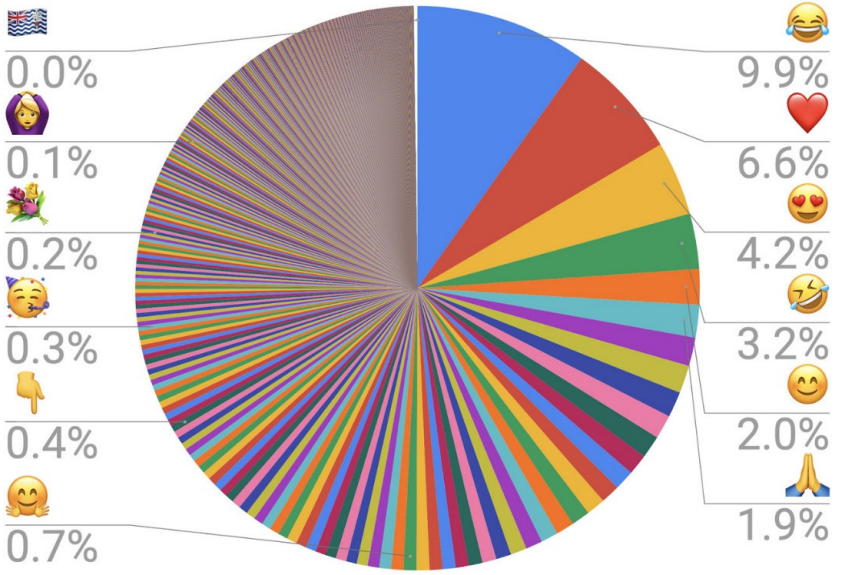
https://datavizcatalogue.com/

- Analyse one variable (is it constant, does it have one or more peaks)?
- Tell a story (what do you think your data can tell)?
- Show a relation between two variables?
- Suggest a trend?
- Show a change (across time, across space, etc)?

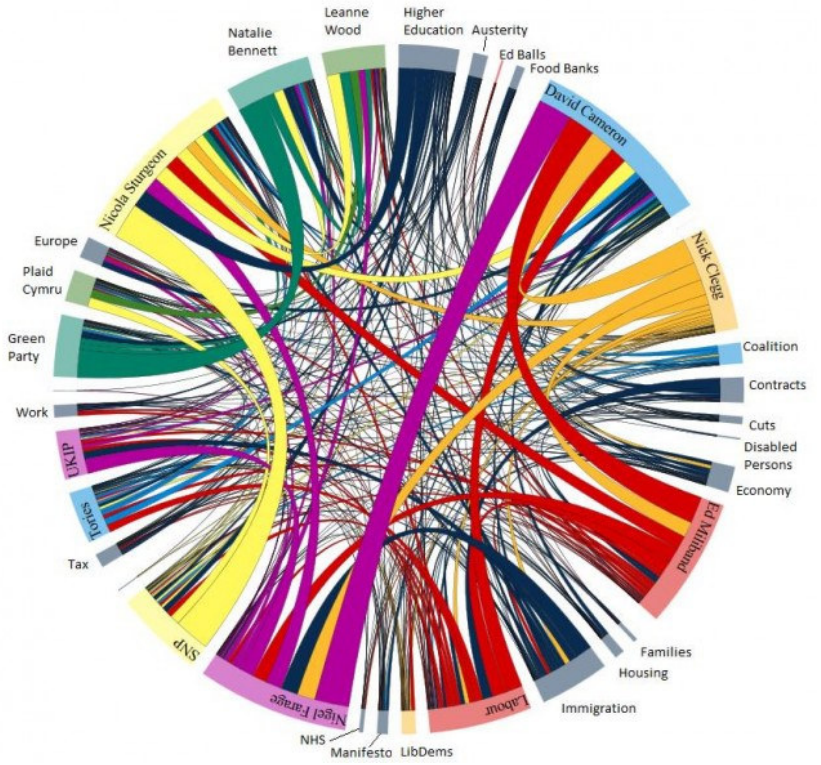
Consider the medium

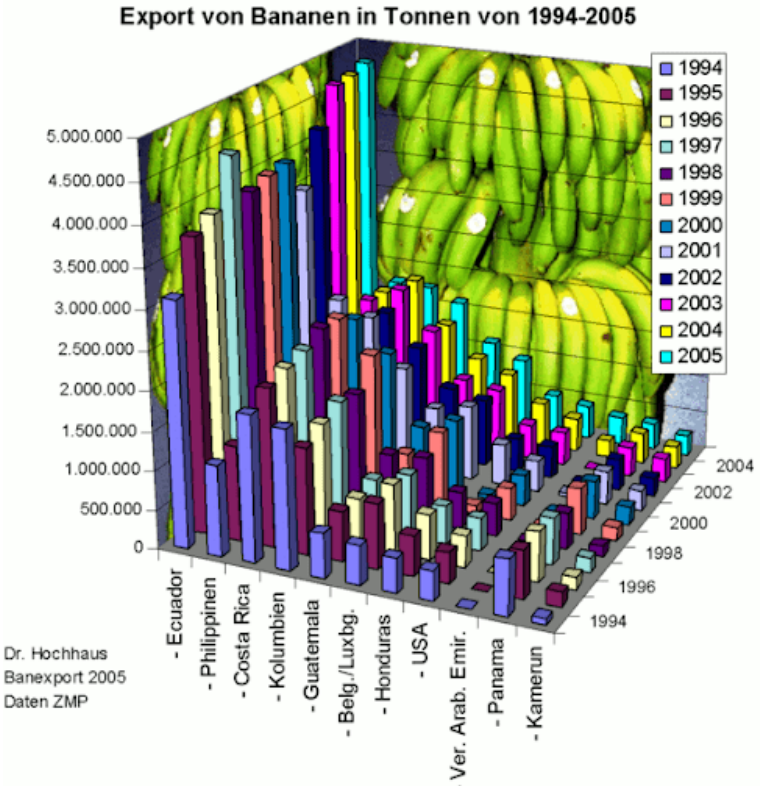
- Will it be printed? Visualized online…?
- How big will it be?
- How can I improve the understanding?





Avoid showing off



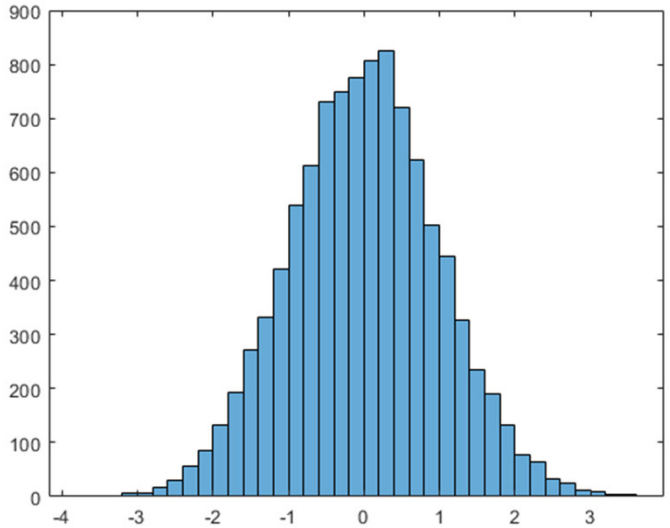


https://rafalab.github.io/dsbook/data-visualization-principles.html

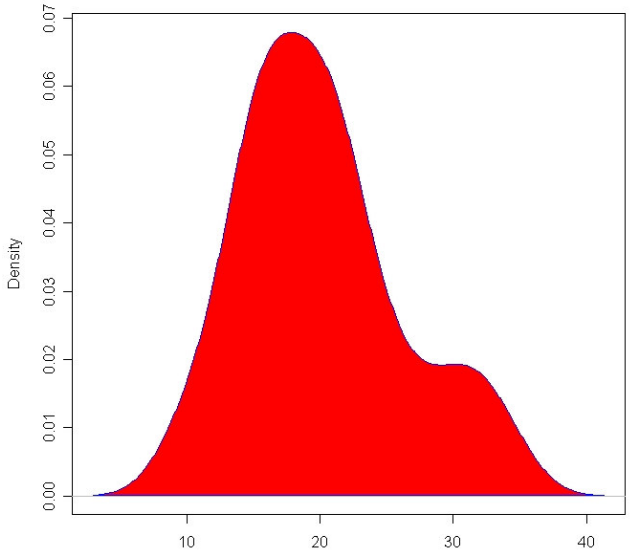
DataVis– the Coco Chanel Approach

- Don’t try to fit too much in the same visualisation
- Is what you were trying to tell with the graph still the focus of the visualisation
- Limit the non data frills to the max (no graph would be more readable with a textured background)
- Can your visualisation be subdivided in more graphs each showing a specific aspect
- 3 is “normally” the max number of variables you can think to plot (x,y, colour/shape)
- Colours are powerful use them with parsimony (also consider colour-blindness)

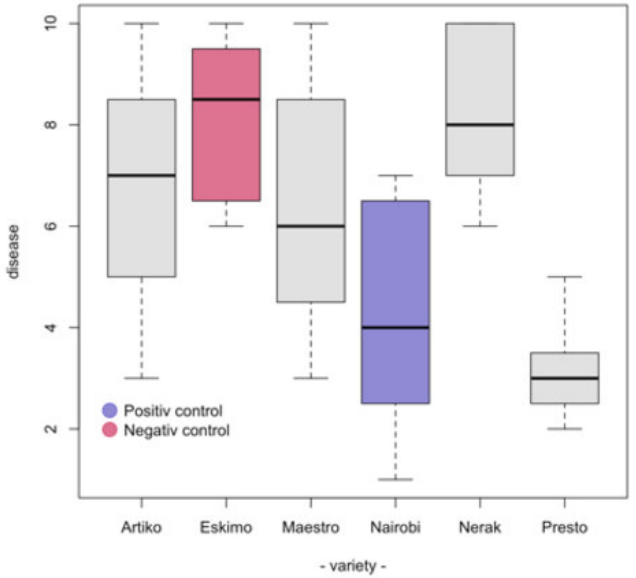
Distribution



Histogram

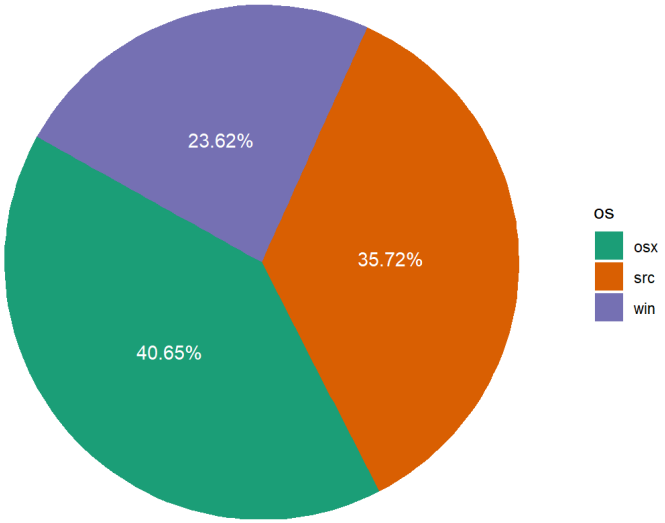


Density Plot

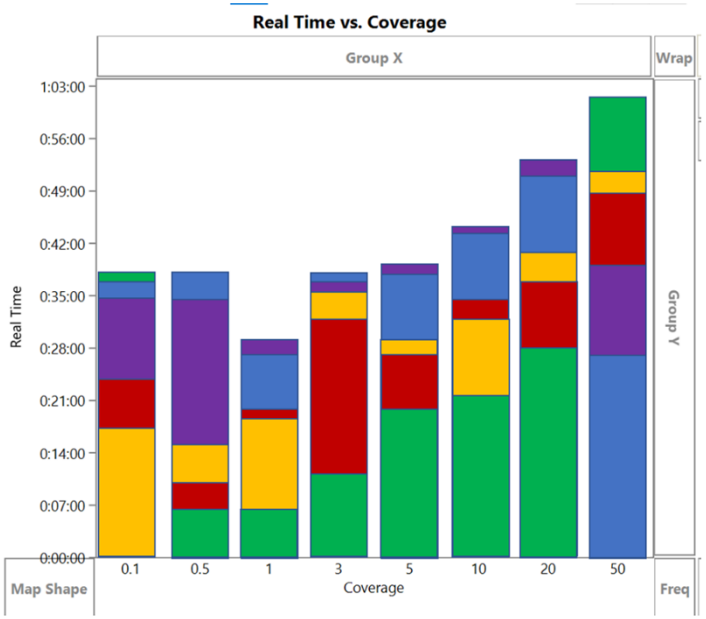


Boxplot

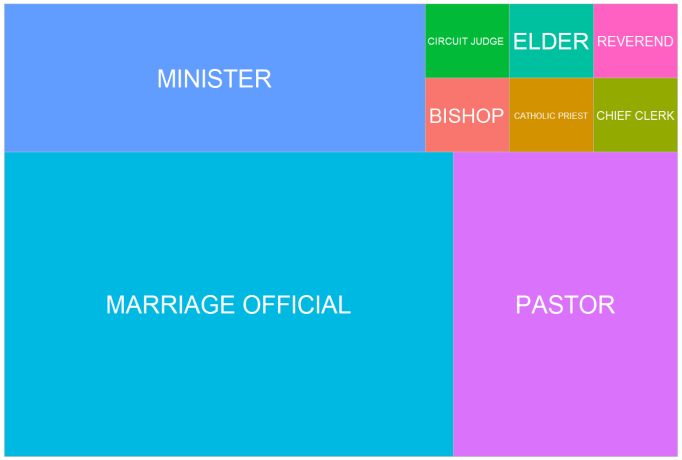
Ratio



Piechart

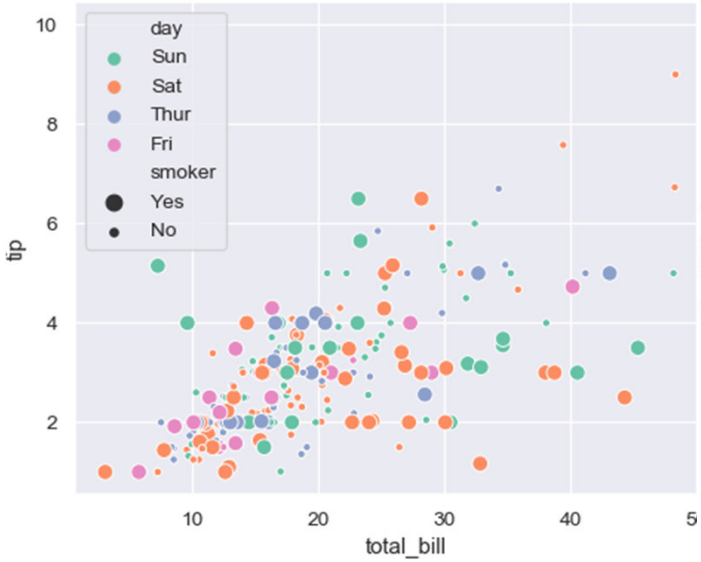


Stacked bar graph

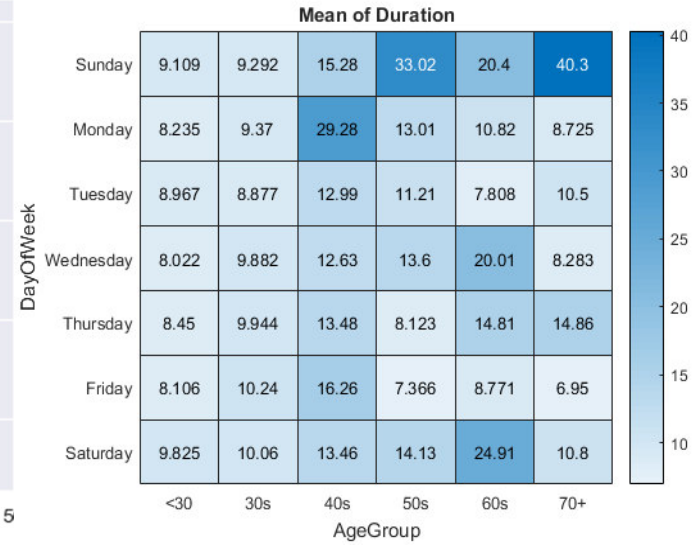


Treemap

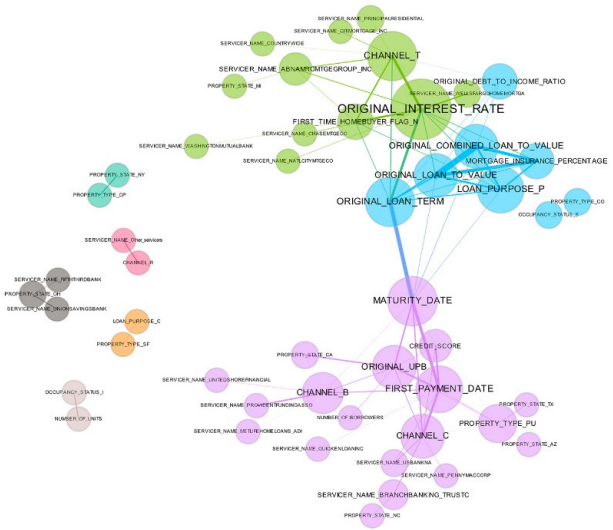
Relations



Scatter plot

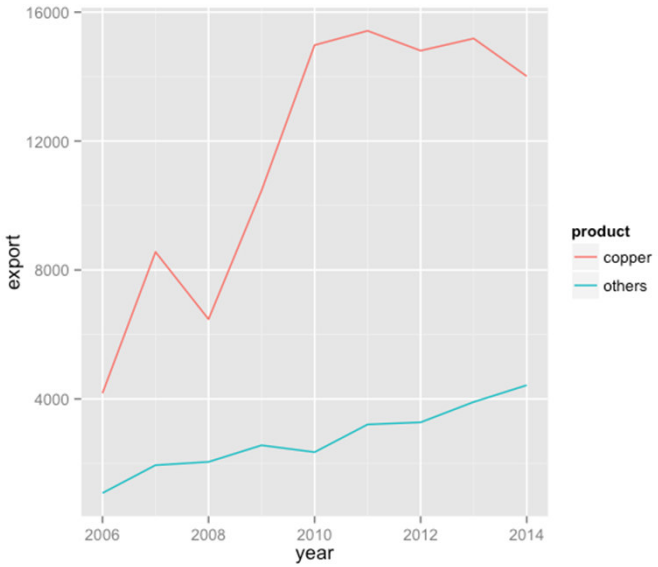


Heatmap

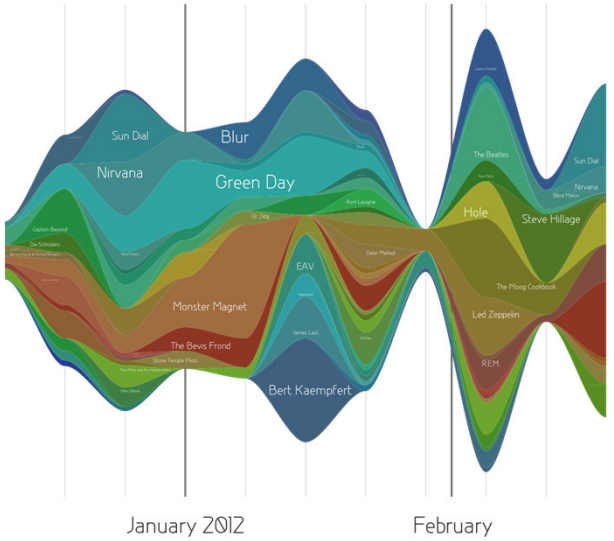


Network diagram

Change



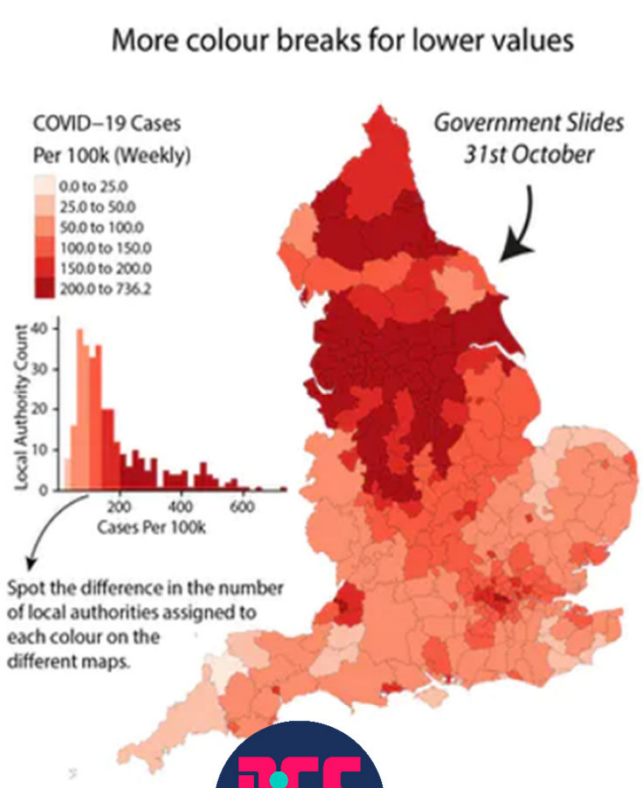
Line graph

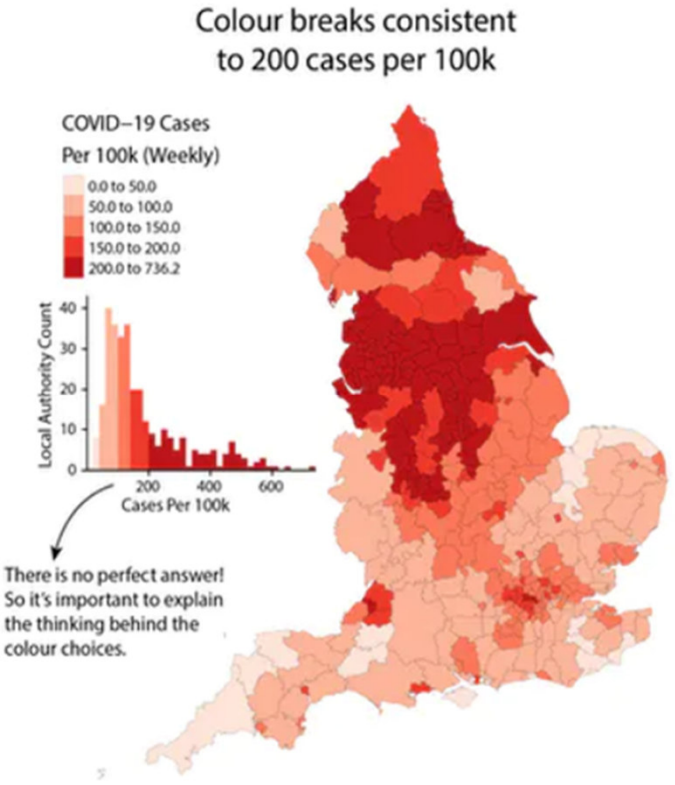


Stream graph

Using Colour

Be consistent
Not just with colours but where you put the breaks



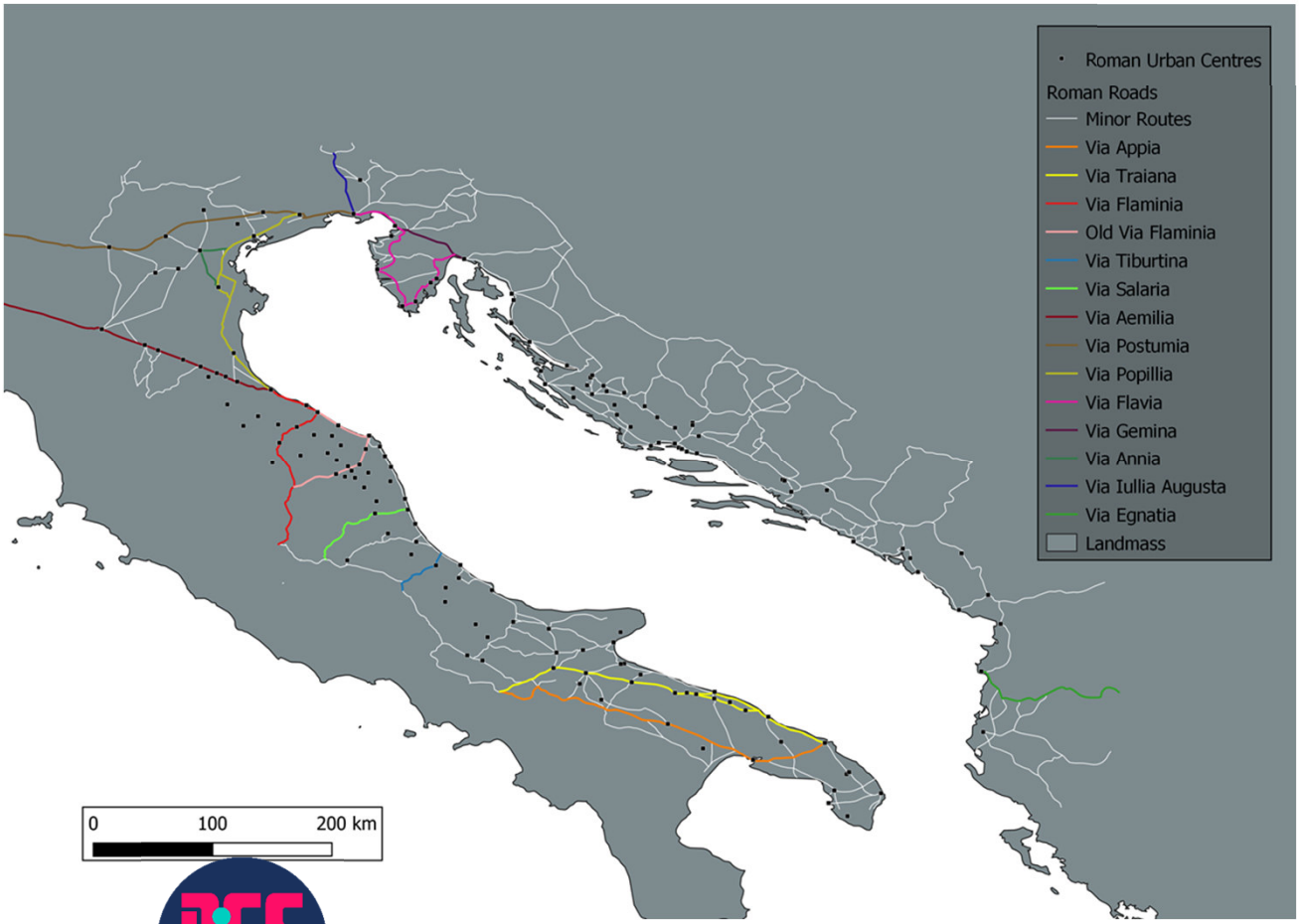


Prof. James Cheshire UCL
https://theconversation.com/next-slide-please-data-visualisation-expert-on-whats-wrong-with-the-uk-governments-coronavirus-charts-149329

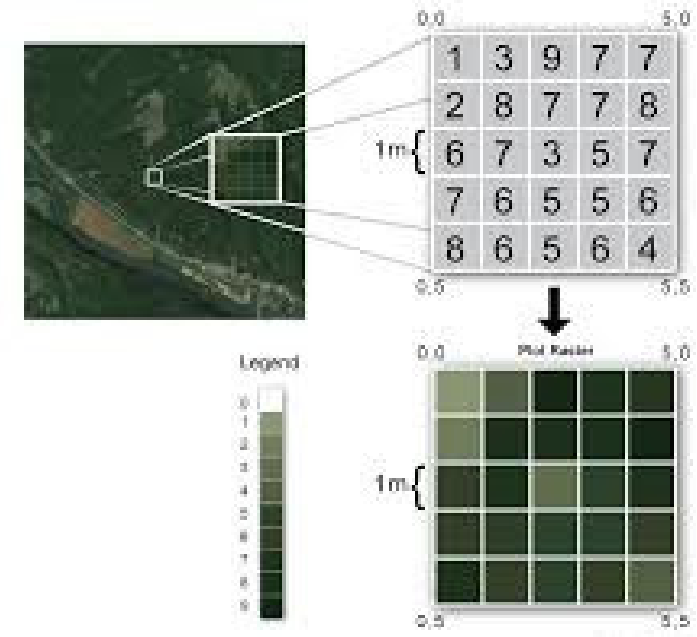
Geographical Data

Vectors

- Generalised representation of the real world
- Composed of XY coordinates (Lon Lat) = vertices
- Points, lines or polygons
- Normally in .shp (shapefile) format

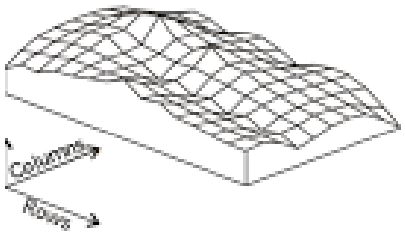


Geographical Data



Raster

- Representation by dividing a surface into a regular grid array
- Each cell has an associated value
- Same size and shape for each cell
- No. Of cells across an area determines spatial resolution
- Normally in tiff. format (but you can create one using vectors)



Rows	0	1	2	3	4	5	6	7	8	9
0	50.0	62.5	70.7	74.3	75.0	75.1	73.3	67.7	55.7	45.0
1	62.7	67.4	73.5	77.0	77.7	78.6	76.6	71.0	62.0	55.0
2	60.9	70.9	77.5	84.3	84.5	88.2	82.0	74.7	67.9	63.4
3	70.0	75.3	78.9	86.1	86.9	97.5	86.6	76.7	67.5	61.5
4	71.6	75.4	76.6	75.4	77.3	89.9	89.8	76.6	64.7	55.0
5	76.3	75.0	73.7	71.8	74.5	83.2	81.8	74.9	67.2	61.6
6	79.8	76.8	71.2	61.0	72.5	79.0	77.5	72.9	67.6	63.8
7	80.3	79.1	70.1	67.3	74.4	90.0	78.0	70.9	63.5	57.3
8	76.5	73.4	65.8	58.2	71.7	77.5	74.3	68.6	58.8	48.0
9	80.0	71.5	65.6	63.2	69.7	76.2	71.1	67.1	61.2	56.0
10	70.6	68.2	62.8	53.0	65.0	69.6	67.7	65.4	62.6	60.4
11	69.3	66.6	61.9	58.6	62.2	65.6	64.5	62.9	60.9	59.1
12	65.7	64.8	59.3	50.0	57.9	60.4	60.8	59.7	56.5	53.3
13	63.3	63.1	58.9	53.8	55.5	59.7	58.5	56.2	48.0	39.3
14	62.7	64.7	58.2	44.3	48.7	55.0	56.5	53.7	40.2	25.0

Vector

- Discrete data (e.g. POI, roads, admin boundary)
- For visualisation -
    - Scaling up and down without compromising quality
- Multiple data per object

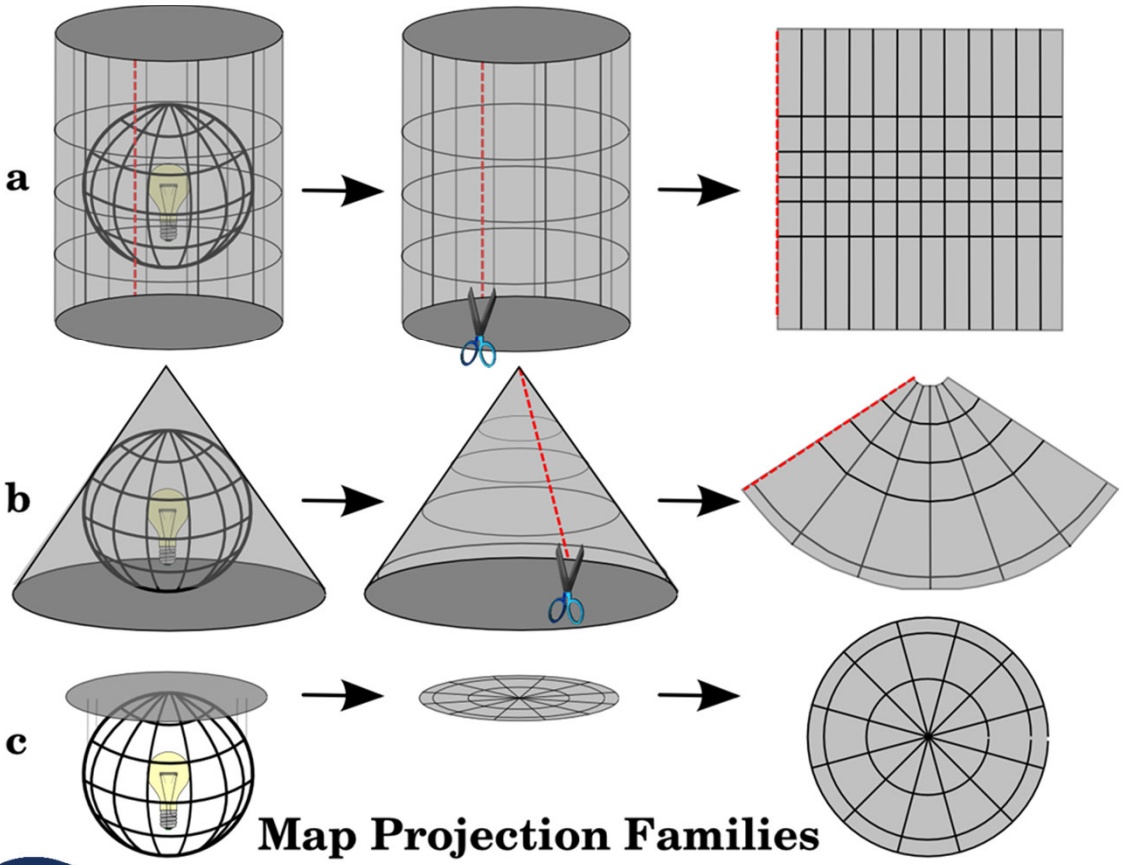
Raster

- Continuous data (e.g. temperature, elevation, CO2 level)
- For mathematical computation -
    - Easier with continuous data
- One attribute per cell

Geographical Data

Projections and Coordinate Reference Systems (CRS)

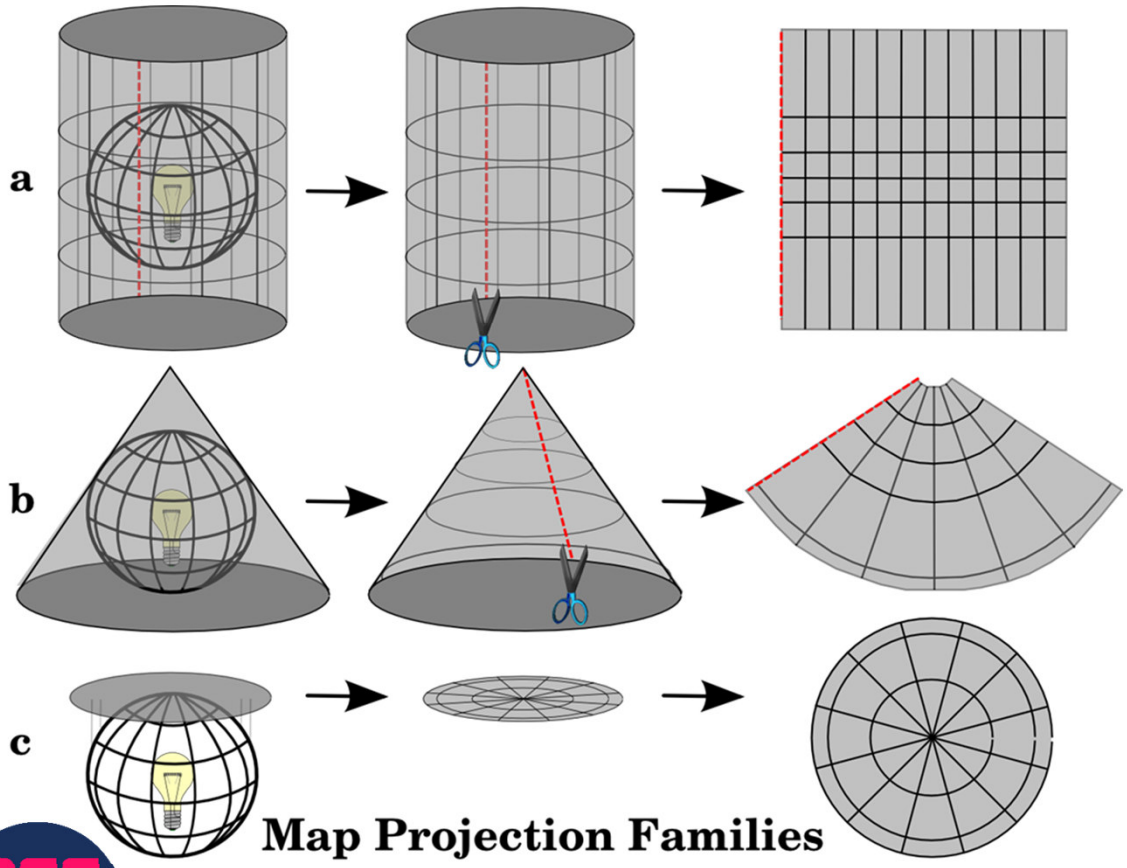
- Maps, whether physical or digital, generally portray a three dimensional space (the planet) on a 2 dimensional medium
- CRS are the systems that, with the use of coordinates, define how the 3D data is projected onto the 2D output



Geographical Data

Projections and Coordinate Reference Systems (CRS)

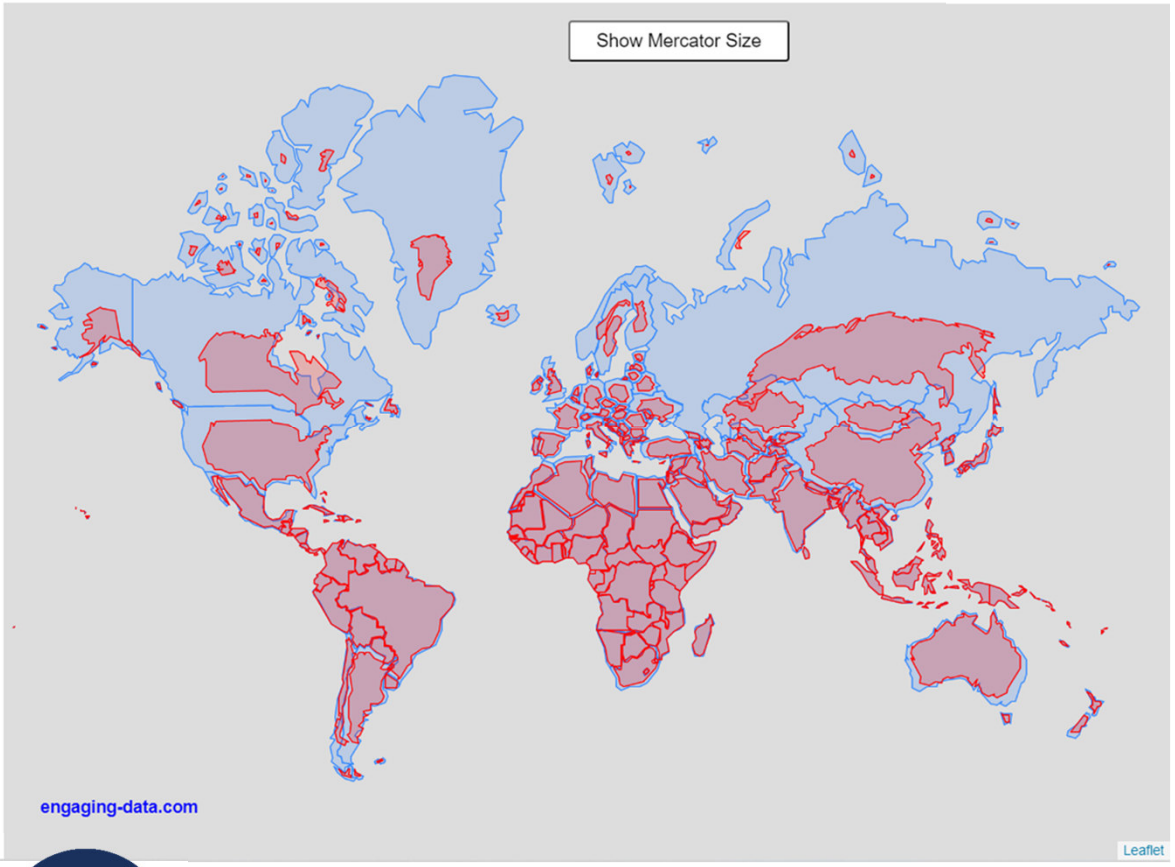
- There are three main map projection families
  - A) Cylindrical
  - B) Conical
  - C) Planar
- Each has advantages and disadvantages
- None actually change the data itself, simply the way it is presented/projected



Geographical Data

Projections

- A common projection is Mercator
- No projections are entirely accurate, they are representations of reality, not reality itself
- In the UK, the projection is EPSG:27700



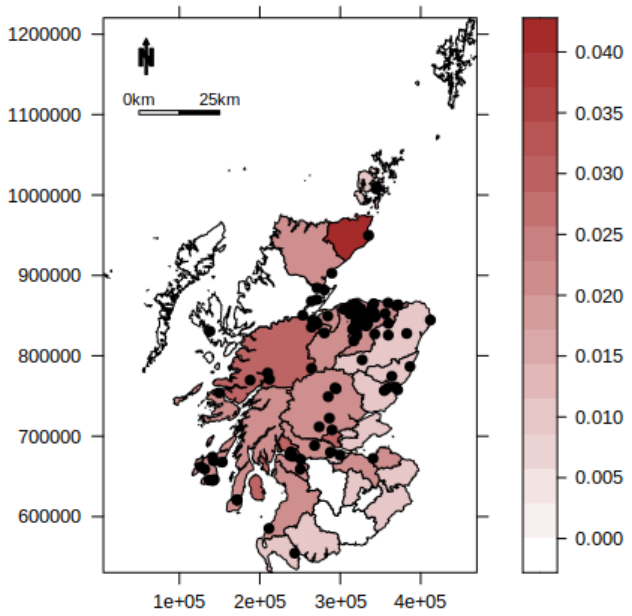
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

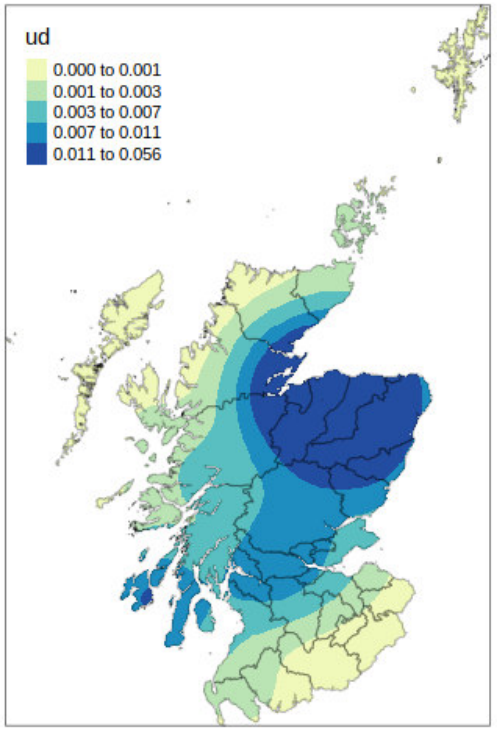
DATA CULTURE SOCIETY

Geographical Data

Geospatial Data in R

- You won't need to worry too much about vectors, rasters and CRS in this class
- We will mainly be working with vectors and sticking to standard CRS
- It is important to understand some of the differences and issues involved





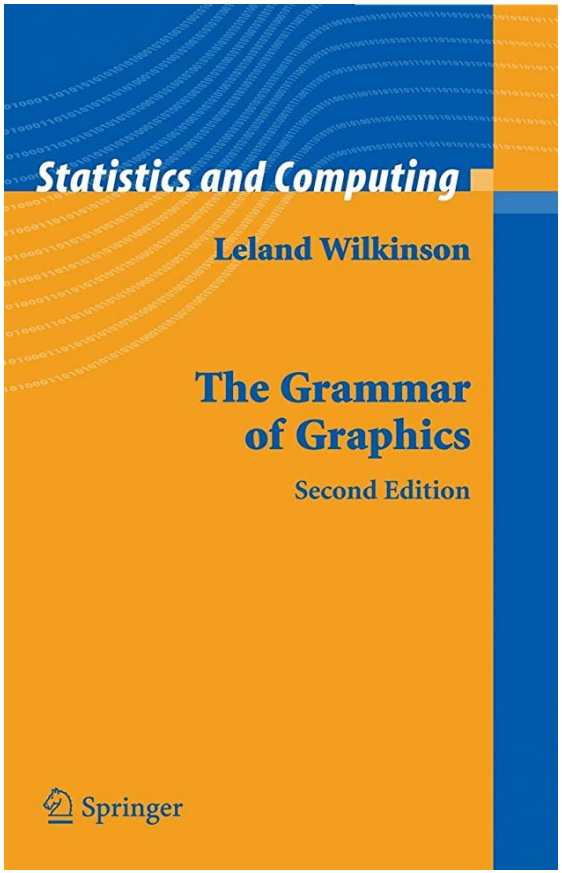
DATA CULTURE SOCIETY

The Grammar of Graphics

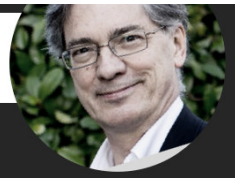
Sentences are **elegant compositions** of carefully chosen grammatical **elements** that convey **precise** and clear messages

Visualisations are elegant mapping of **data** onto the right visual **encondings** to tell a story

By Leland Wilkinson



Grammar of Graphics by L. Wilkinson



Describes non-data ink. Design elements!

The plotting space you are using

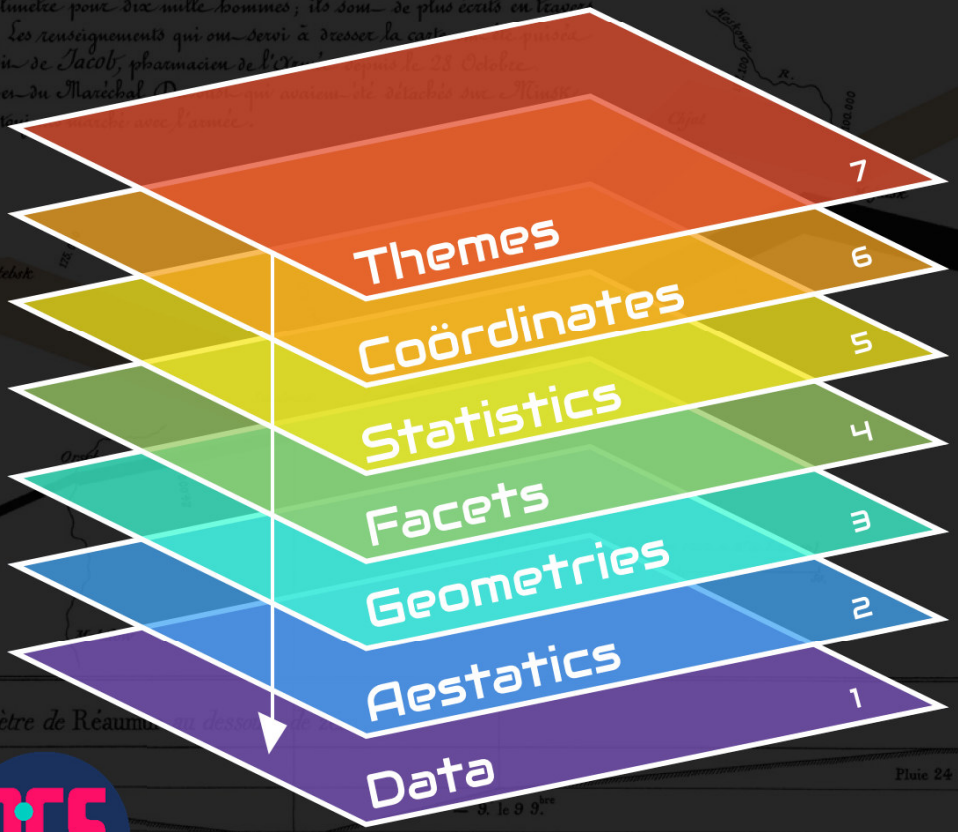
Statistical models & summaries

Rows and columns of sub-plots

Shapes used to represent your data

The scales on which the data is mapped

The actual variables to be plotted



www.cdcs.ed.ac.uk

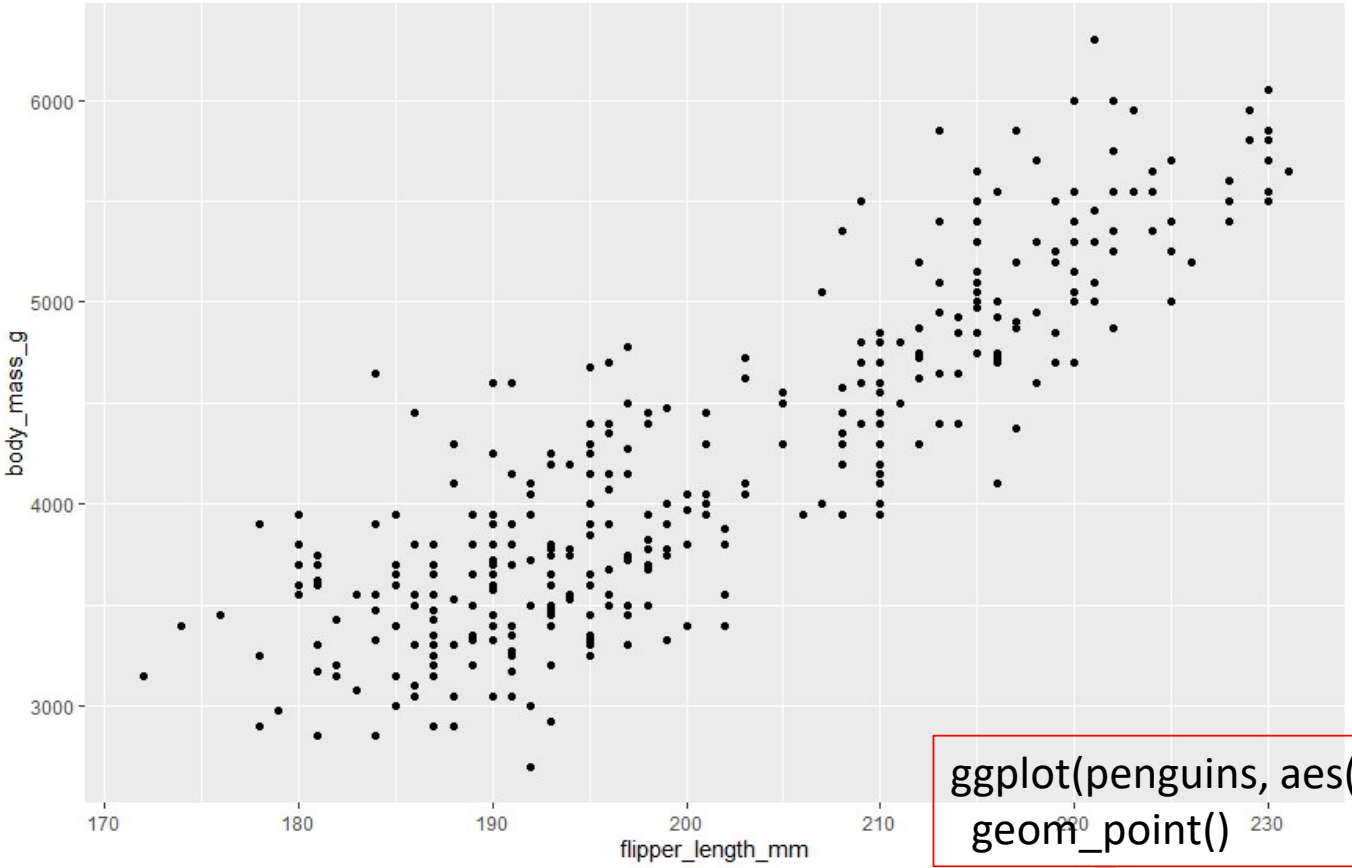
The data

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
	<fct>	<fct>	<dbl>	<dbl>	<int>	<int>	<fct>	<int>
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18	195	3250	female	2007
4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
10	Adelie	Torgersen	42	20.2	190	4250	NA	2007
11	Adelie	Torgersen	37.8	17.1	186	3300	NA	2007
12	Adelie	Torgersen	37.8	17.3	180	3700	NA	2007
13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007
14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007
15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007
16	Adelie	Torgersen	36.6	17.8	185	3700	female	2007
17	Adelie	Torgersen	38.7	19	195	3450	female	2007
18	Adelie	Torgersen	42.5	20.7	197	4500	male	2007
19	Adelie	Torgersen	34.4	18.4	184	3325	female	2007
20	Adelie	Torgersen	46	21.5	194	4200	male	2007

Level 1: The Dataset I am using

- If I want to plot the same type of chart with different data all you have to change is the data reference in the code
- In our example will be the Palmer Penguins

Aesthetics and geometries

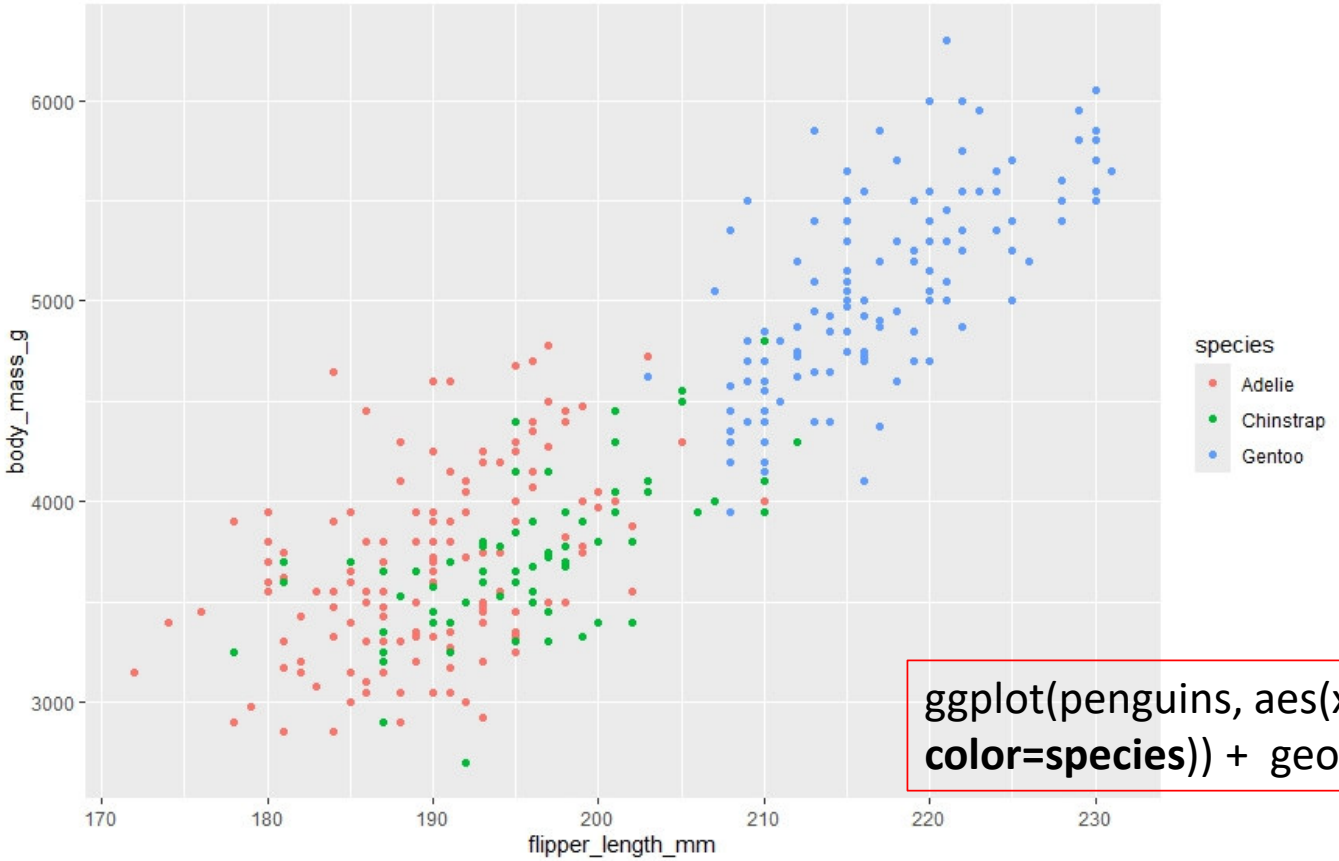


Level 2 and 3

- **Aesthetics** identify which variables I want to work on (depending on the type of chart you are going to work on one or more variables)
- **Geometries** identify the type of chart that I want to produce.

```
ggplot(penguins, aes(x=flipper_length_mm, y=body_mass_g)) +
  geom_point()
```

Aesthetics

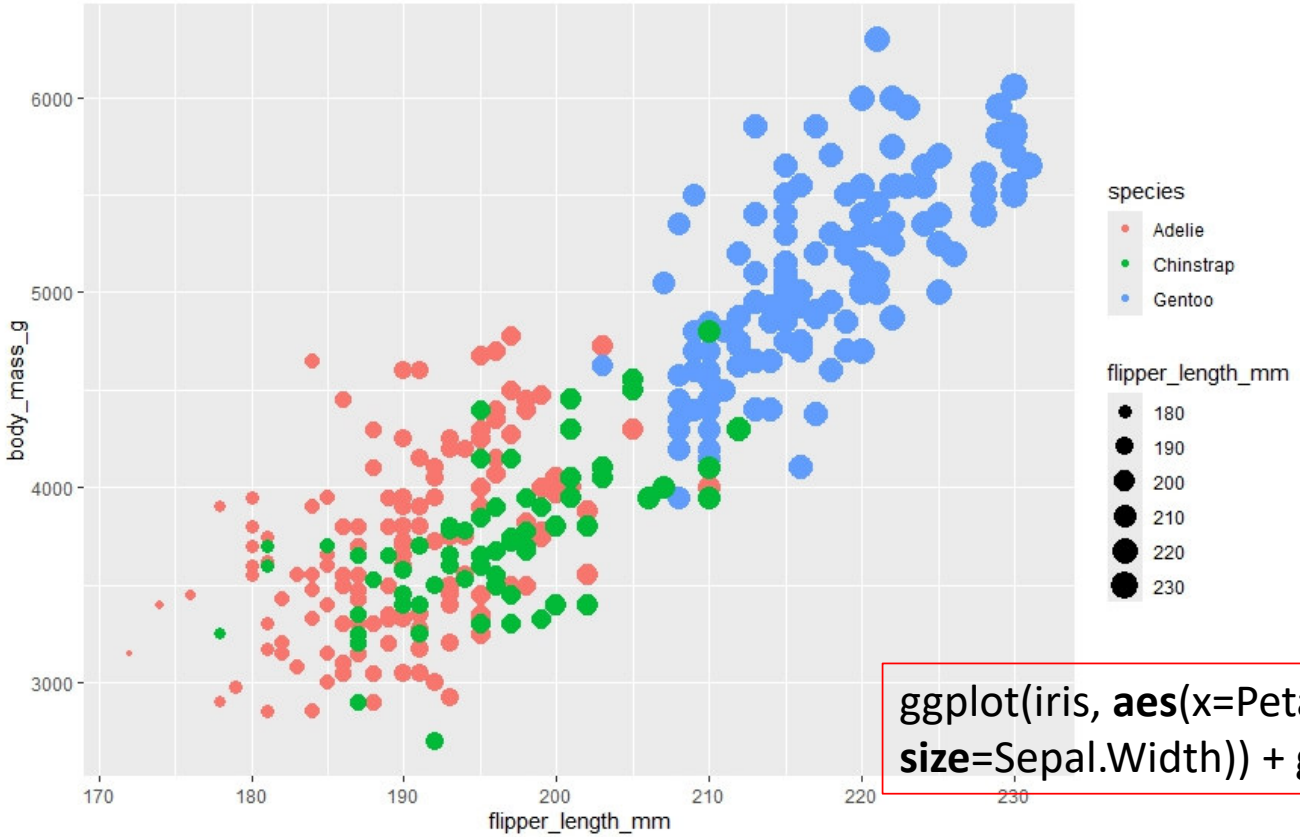


Level 2

- **Aesthetics** Add a third colour coded variable. Since it is still linked to the aesthetics I need to set it within that part of the **code**

```
ggplot(penguins, aes(x=flipper_length_mm, y=body_mass_g,
color=species)) +  geom_point()
```

Aesthetics

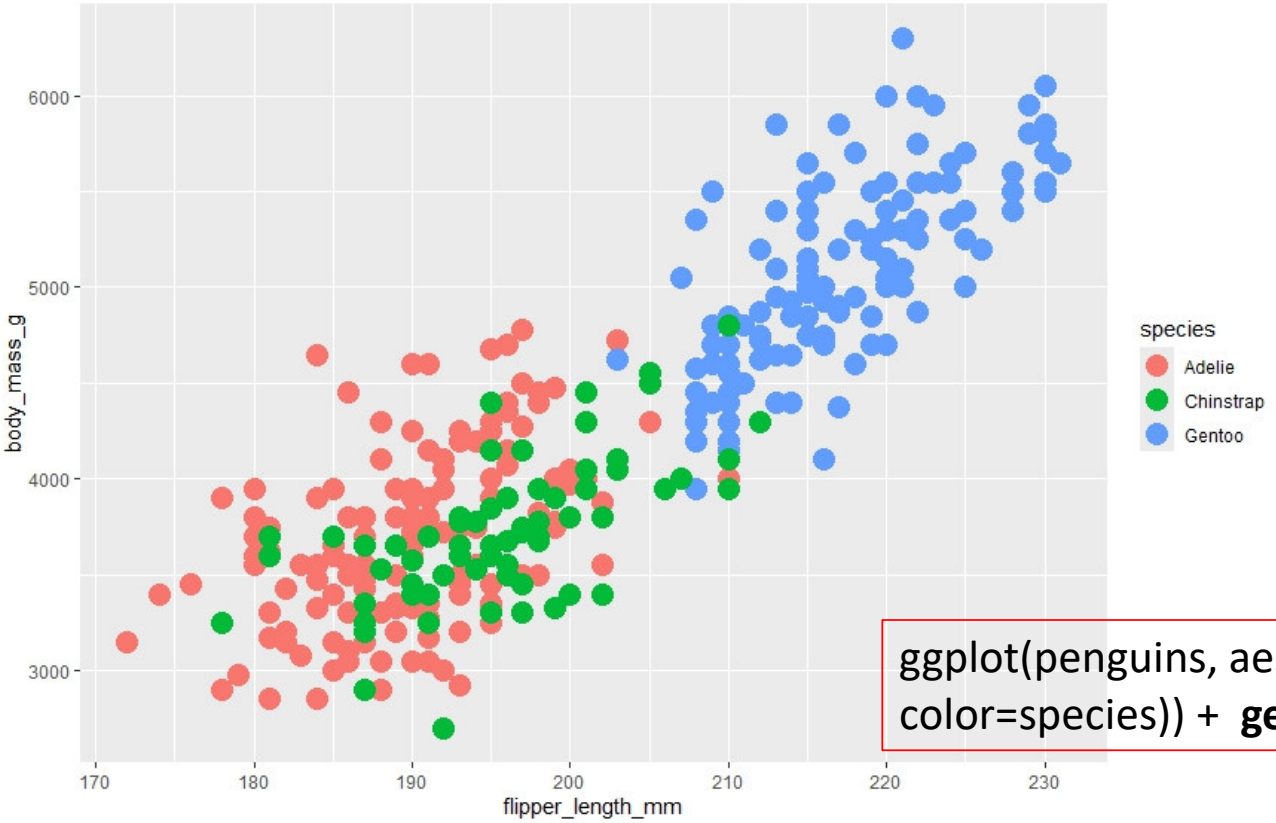


Level 2

- **Aesthetics** Add a fourth size coded variable. Since it is still linked to the aesthetics I need to set it within that part of the **code**

```
ggplot(iris, aes(x=Petal.Length, y=Petal.Width, color=Species,
size=Sepal.Width)) + geom_point()
```

Geometry

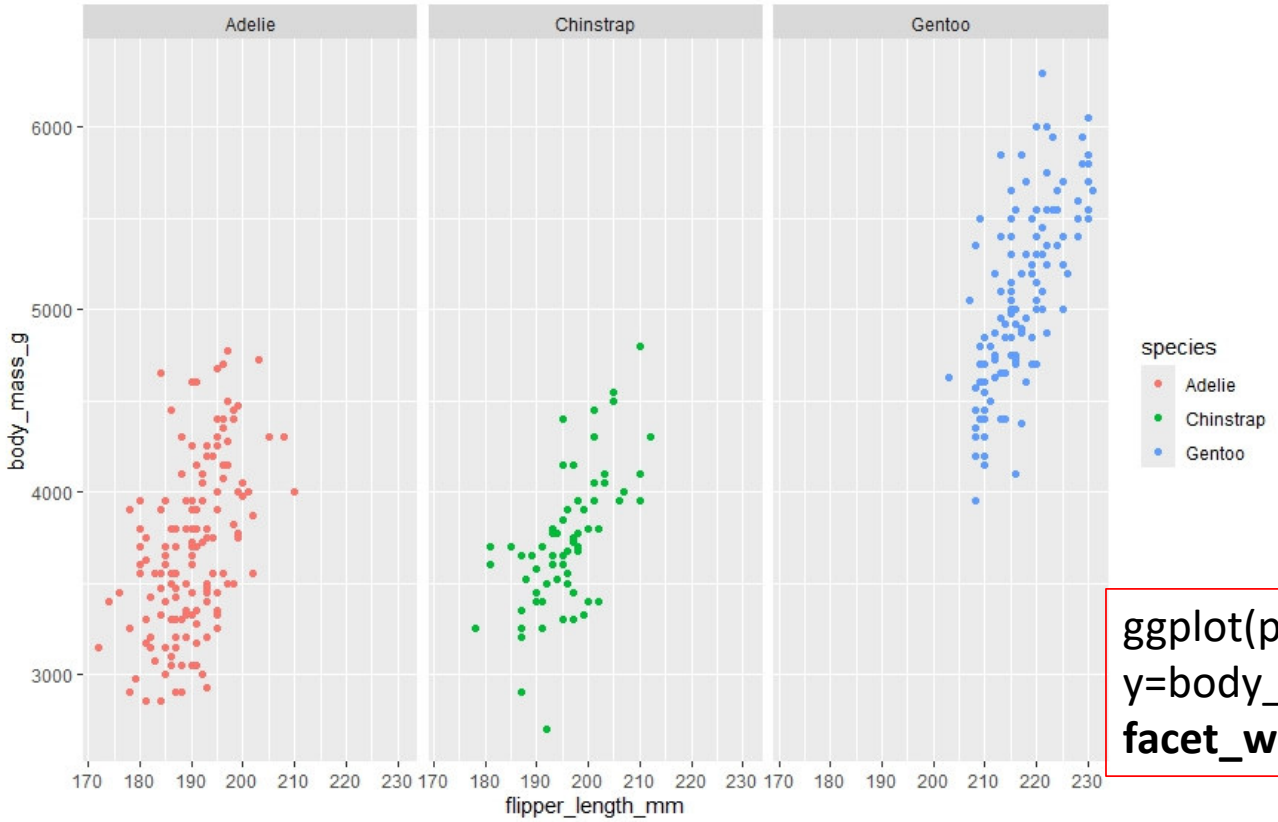


Level 3

- **Geometry.** If I just want all the dots to be bigger I shall set the value in the geometry part of the **code.**

```
ggplot(penguins, aes(x=flipper_length_mm, y=body_mass_g, color=species)) + geom_point(size=5)
```

Facet

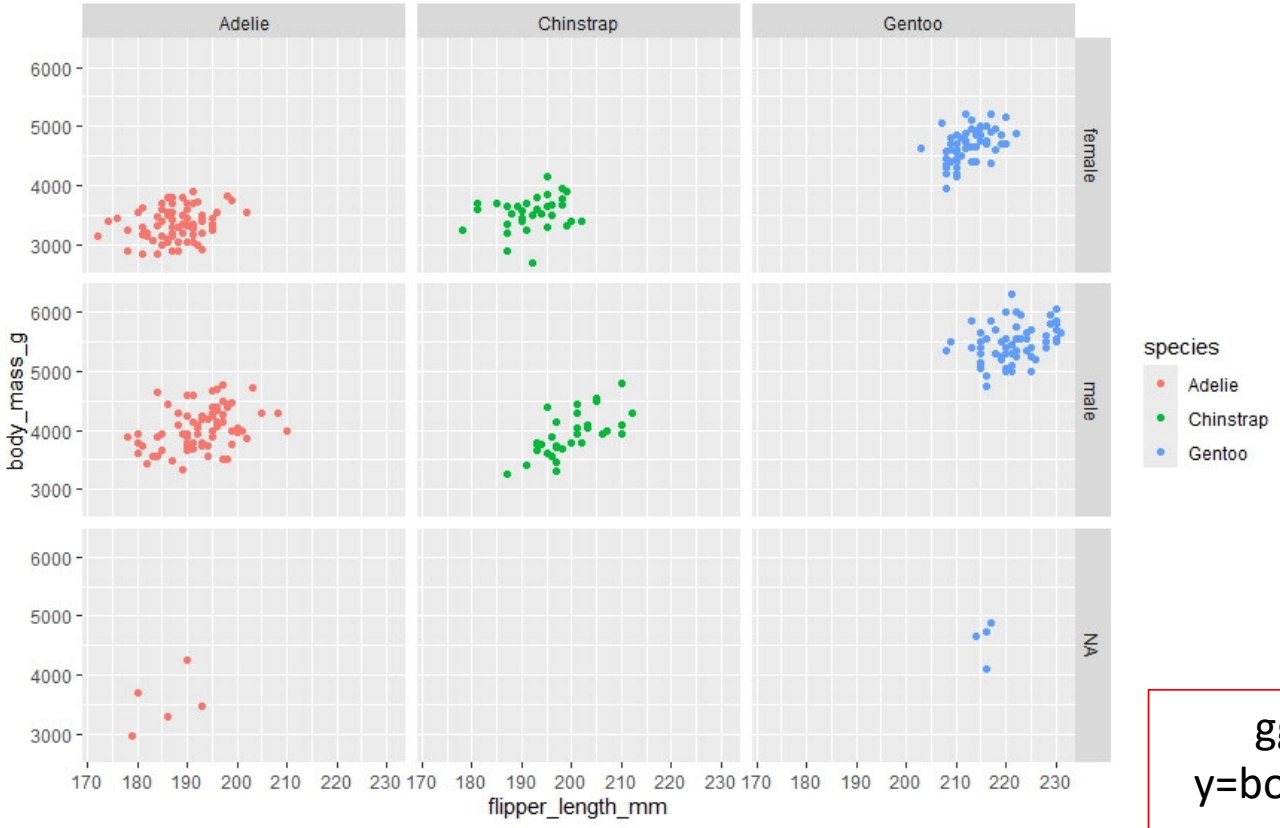


Level 4 Subplot

- **Facet.** If I just want to visualise the results on separate subplot you need to add this new level of information

```
ggplot(penguins, aes(x=flipper_length_mm, y=body_mass_g, color=species)) + geom_point()+
facet_wrap(~species)
```

Facet

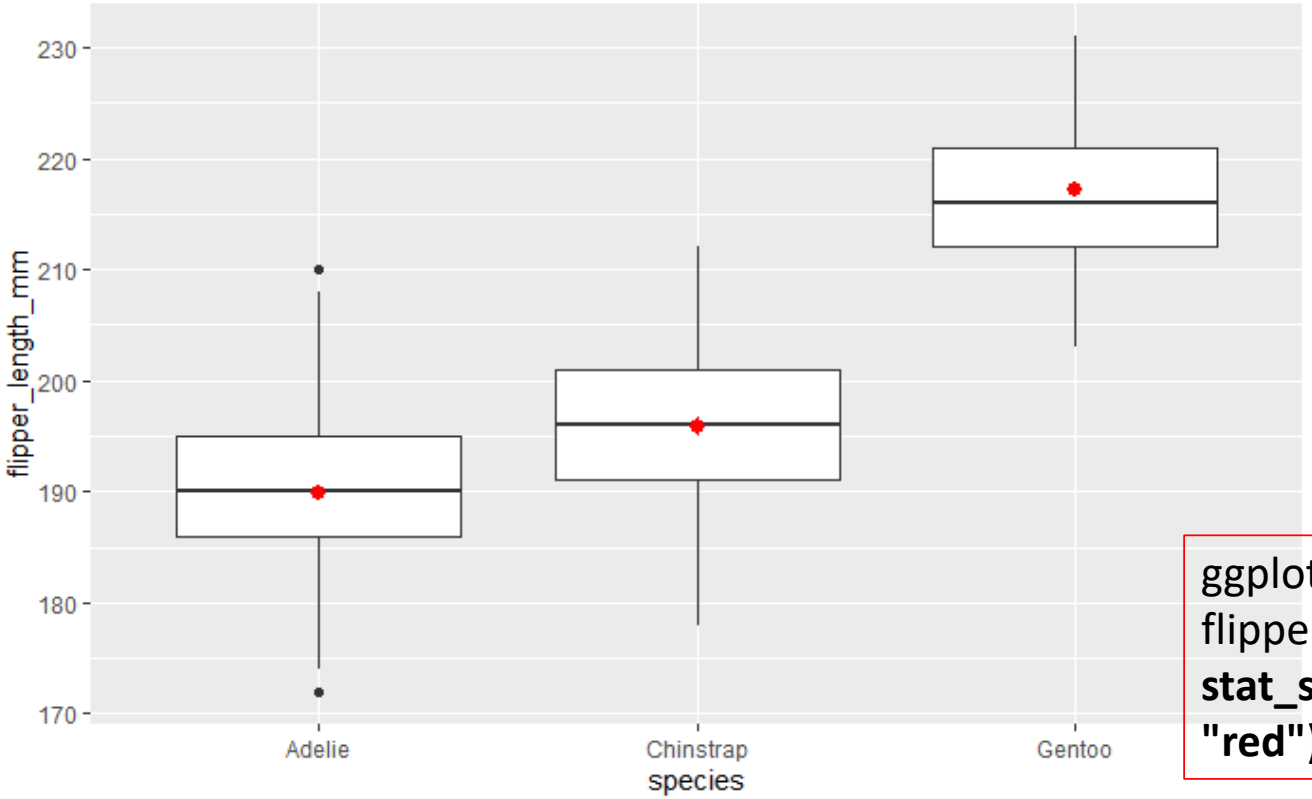


Level 4 Subplot

- **Facet.** If I just want to visualise the results on separate subplot you need to add this new level of information
- **If you want to subplot across 2 different variables use facet_grid**

ggplot(penguins, aes(x=flipper_length_mm, y=body_mass_g, color=species)) + geom_point()+ **facet_grid(sex~species)**

Statistics

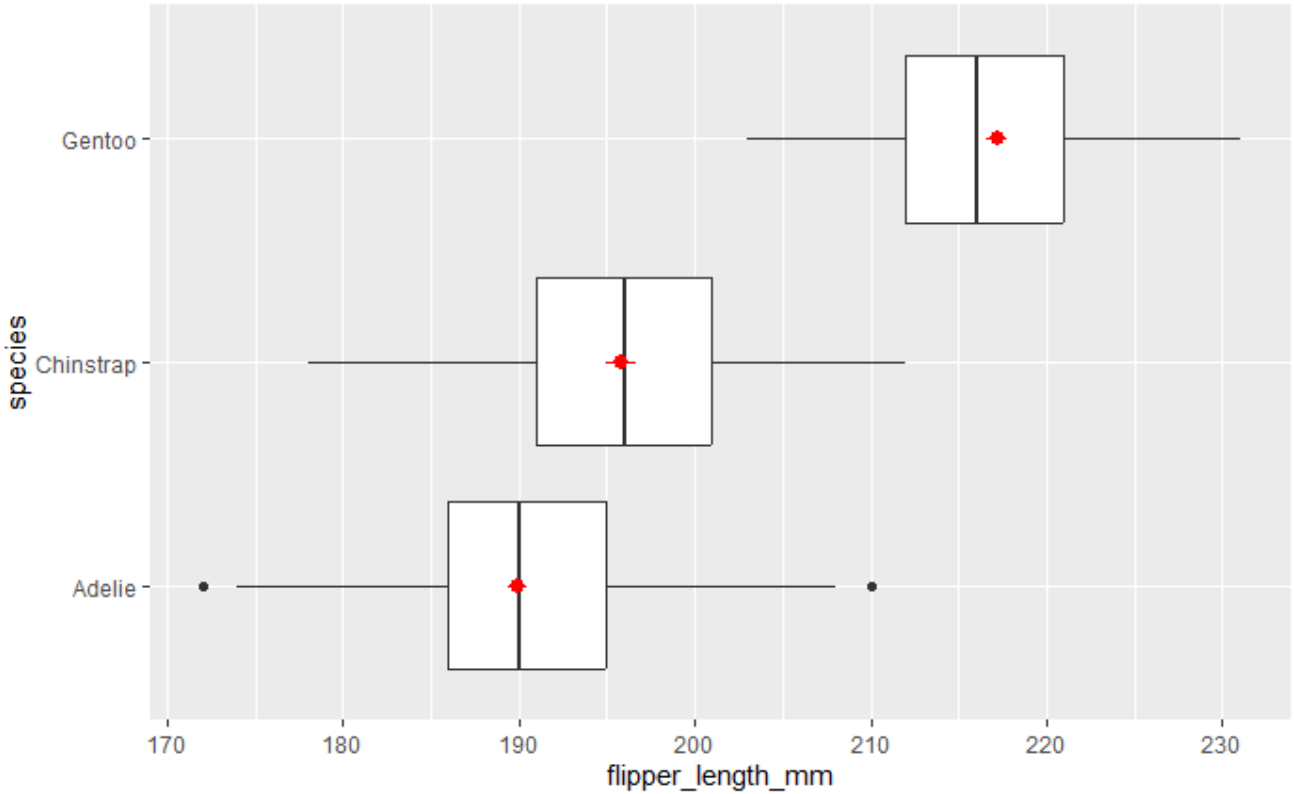


Level 5 Summarizing Stats

- **Instead** of the data if you want to plot the summarising of those data you do so using the **stats level**

```
ggplot(data = penguins,  aes(x = species, y = flipper_length_mm))+  geom_boxplot()+
stat_summary(fun.data = mean_se,    color = "red")
```

Coordinates

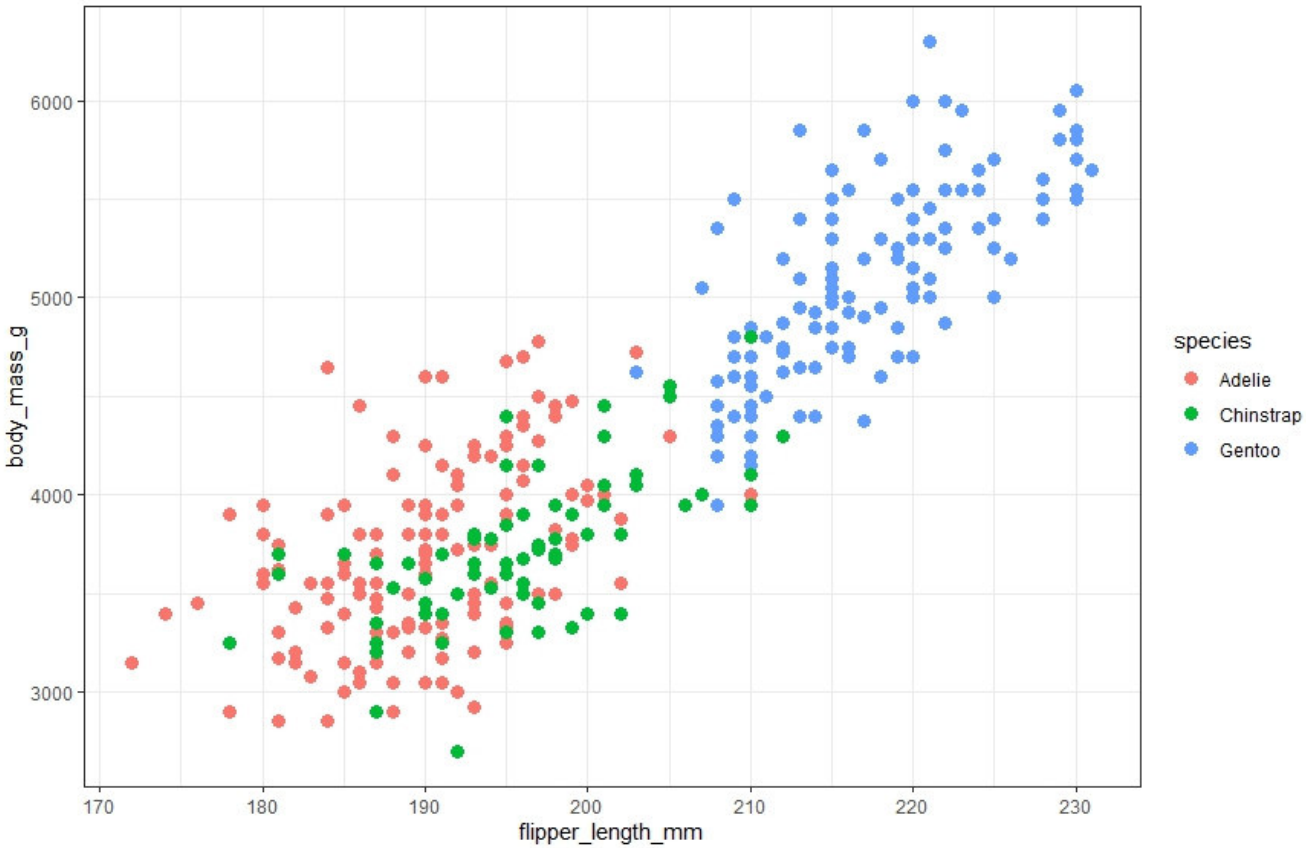


Level 6 Coordinates

- On this level you can set the attributes of the coordinates, change the scale or transform

```
ggplot(data = penguins,  aes(x = species, y = flipper_length_mm))+  geom_boxplot()+ stat_summary(fun.data = mean_se,  color = "red")+ coord_flip()
```

Themes

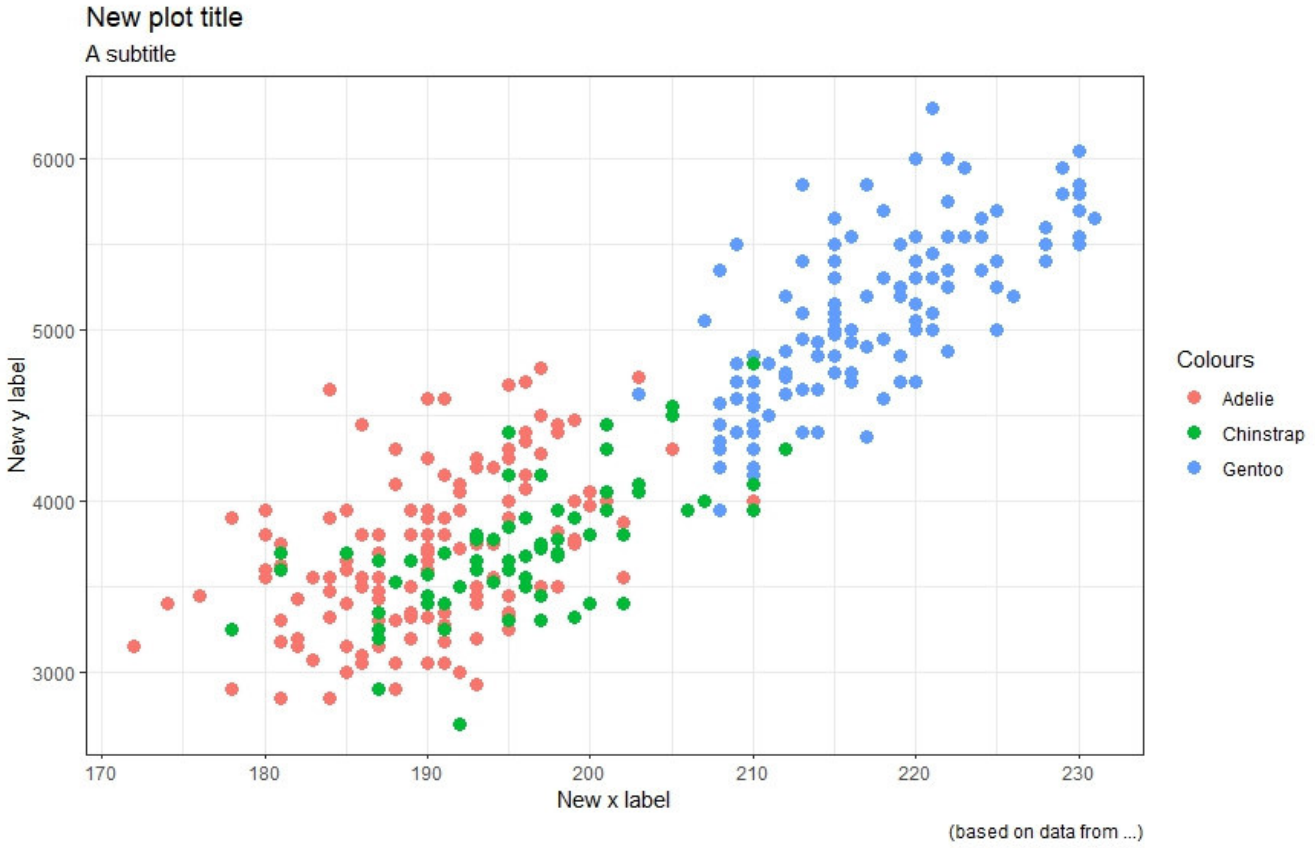


Level 7 Global settings

- **Global settings** of the charts can be changed in the theme part of the code. All non data

```
ggplot(penguins,
aes(x=flipper_length_mm,
y=body_mass_g, color=species)) +
geom_point(size=3)+  theme_bw()
```

Labels



Level 7 labels

- A peculiar setting you want to pay attention to are the lables

```
ggplot(penguins,
aes(x=flipper_length_mm,
y=body_mass_g, color=speciLablees))
+  geom_point(size=3)+
theme_bw()+  labs(title = "New plot
title", subtitle = "A subtitle", caption
= "(based on data from ...)", x =
"New x label", y= "New y label",
color = "Colours")
```


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY

TIME FOR R