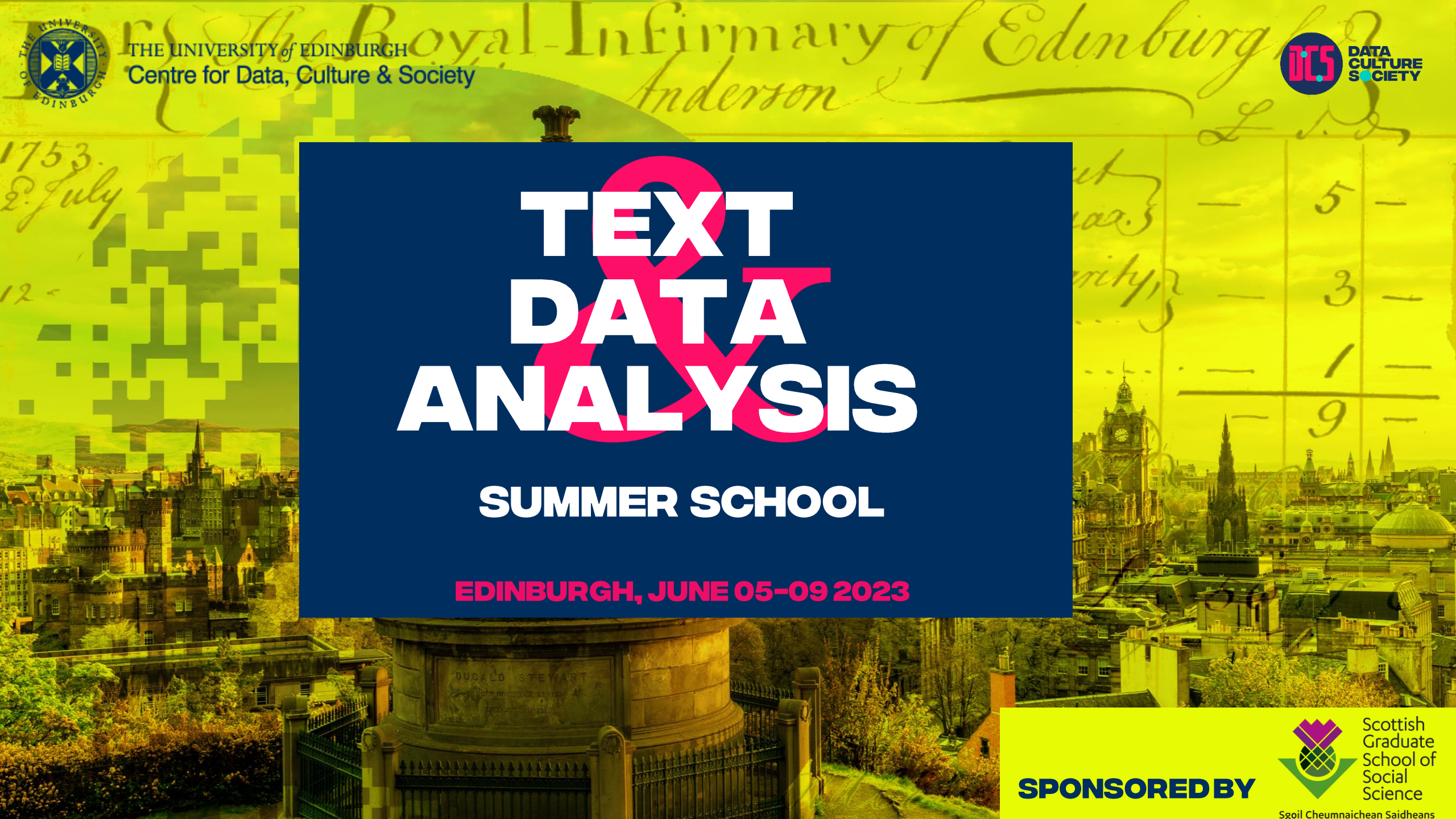

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TEXT & DATA ANALYSIS
SUMMER SCHOOL
EDINBURGH, JUNE 05-09 2023
SPONSORED BY
Scottish Graduate School of Social Science
Sgoil Cheumnaichean Saidheans


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TODAY'S SCHEDULE
Seminar: Combining Computational Research Methods for Text Analysis and Visualisation
Hands-on session 1: Data Visualisation 1
Hands-on session 2: Data Visualisation 2
Conclusion: Summing Up and Next Steps

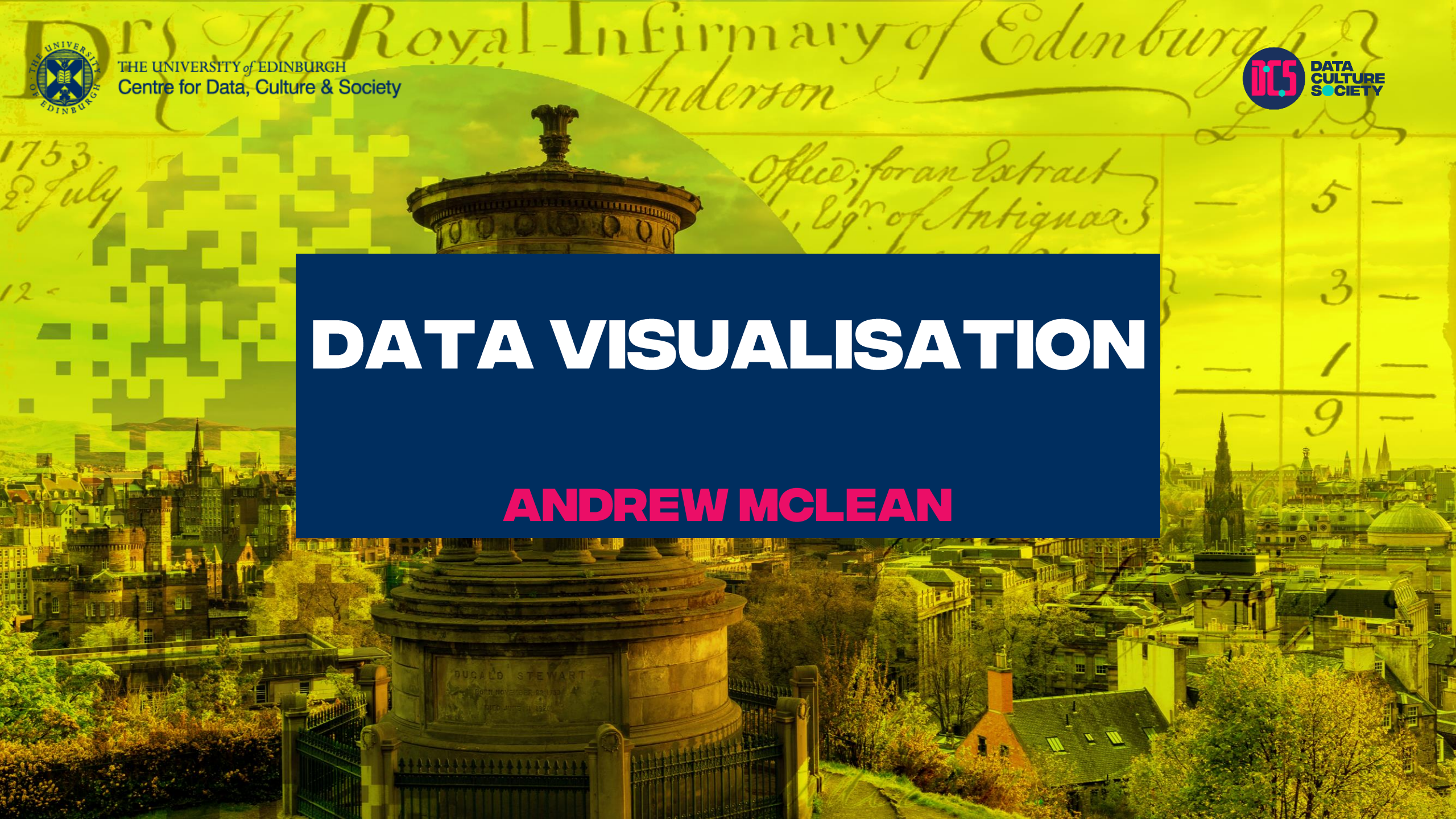
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

COMBINING COMPUTATIONAL RESEARCH METHODS FOR TEXT ANALYSIS AND VISUALISATION

Dr Pedro Jacobetty,

Research fellow at the University of Edinburgh


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
COFFEE BEAK
WE ARE GOING TO RESTART AT 11:00



DATA VISUALISATION

ANDREW MCLEAN

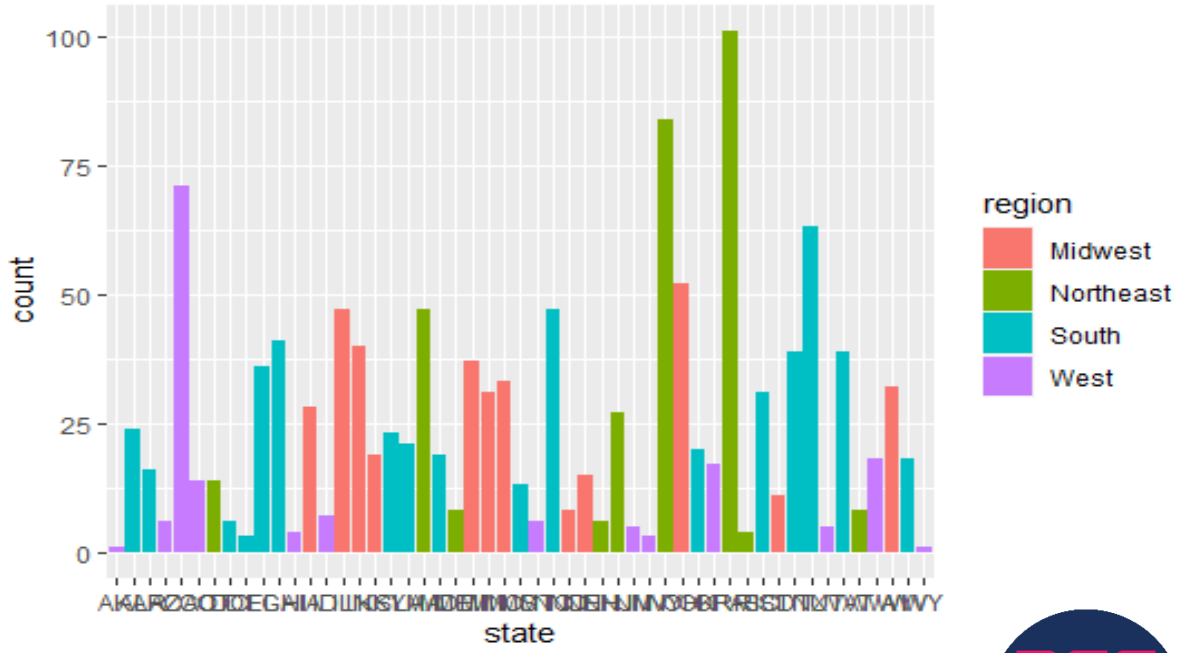
BEFORE PLOTTING…THINK TO WHAT YOU WANT TO CONVEY

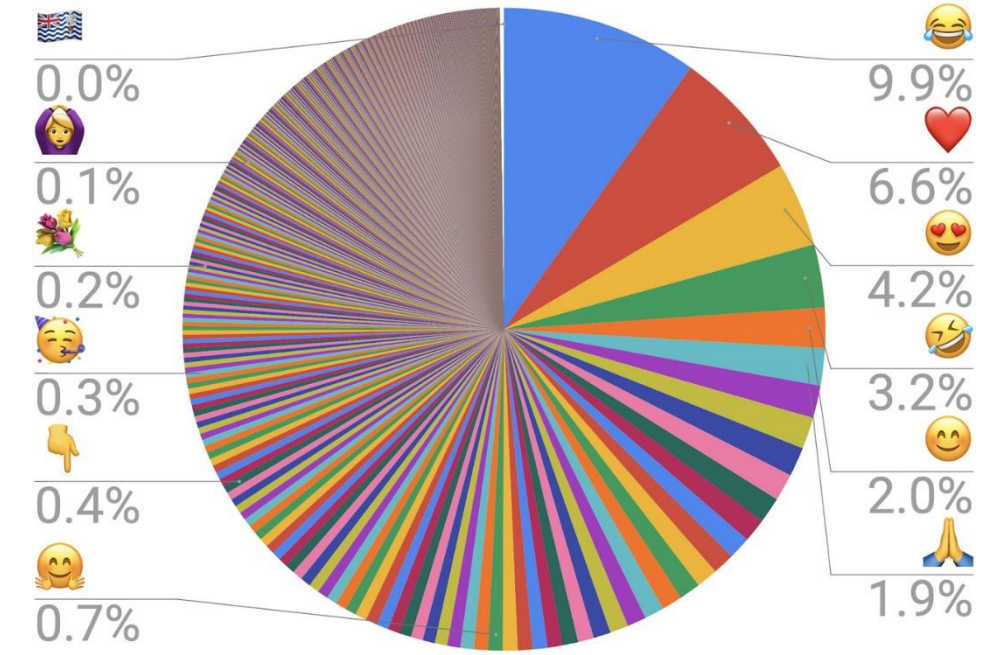
https://datavizcatalogue.com/

- 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)?

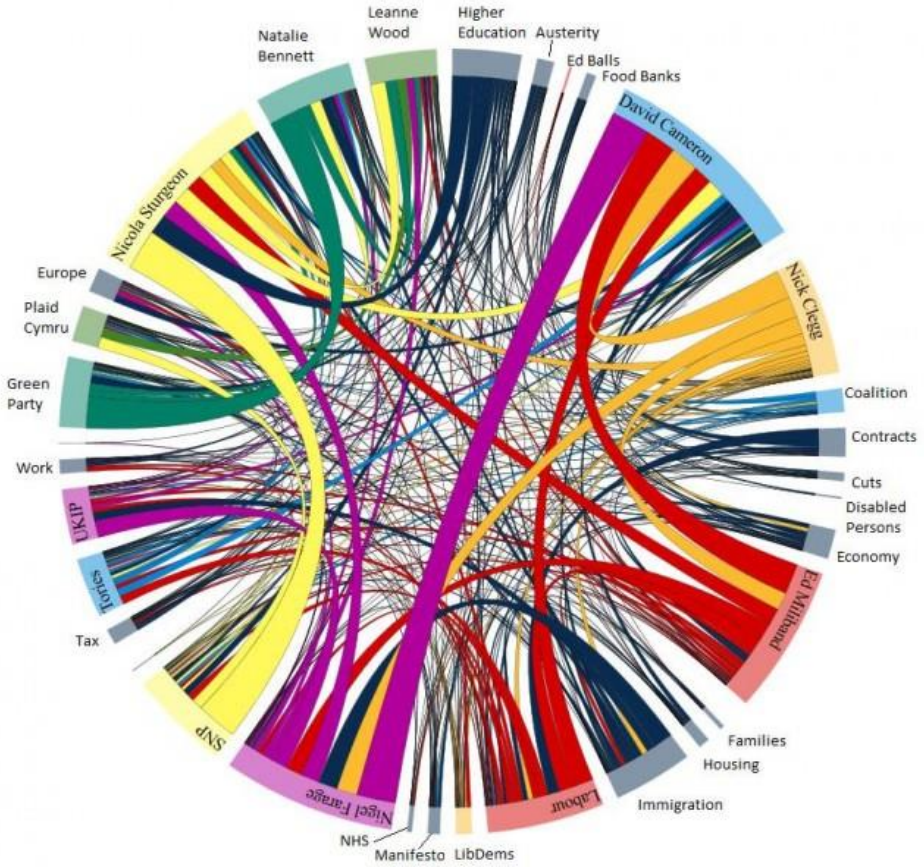
CONSIDER THE MEDIUM

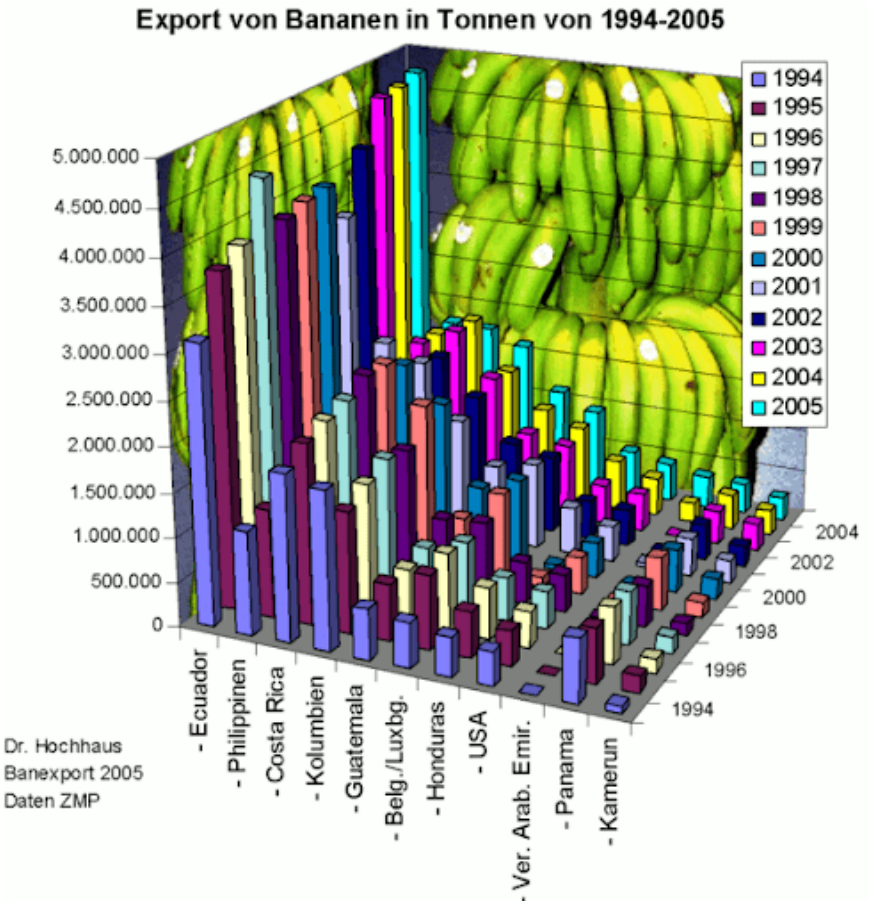
- Will it be printed? Visualized online…?
- How big will it be?
- How can I improve the understanding?





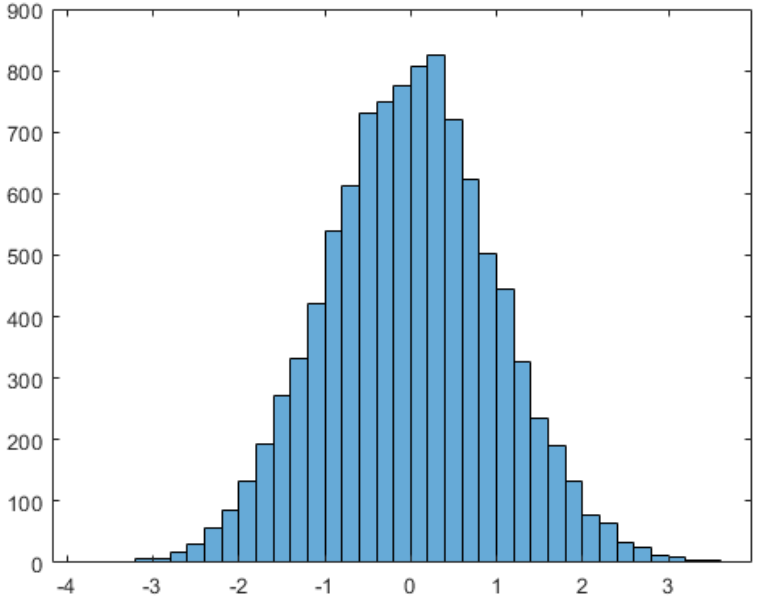
AVOID SHOWING OFF

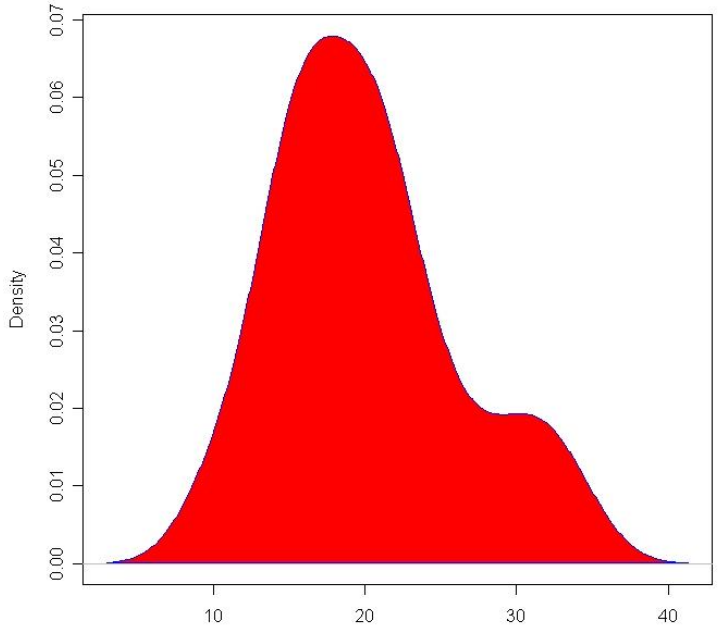


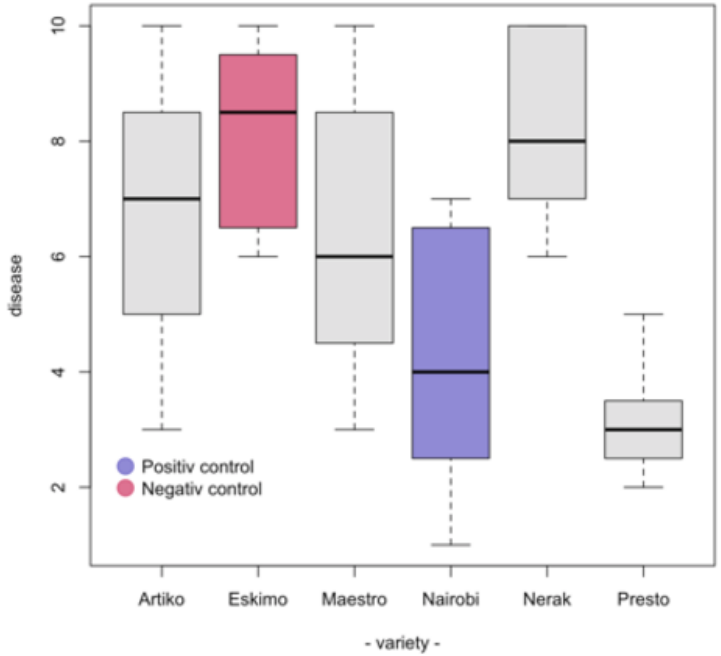


https://rafalab.github.io/dsbook/data-visualization-principles.html

DISTRIBUTION





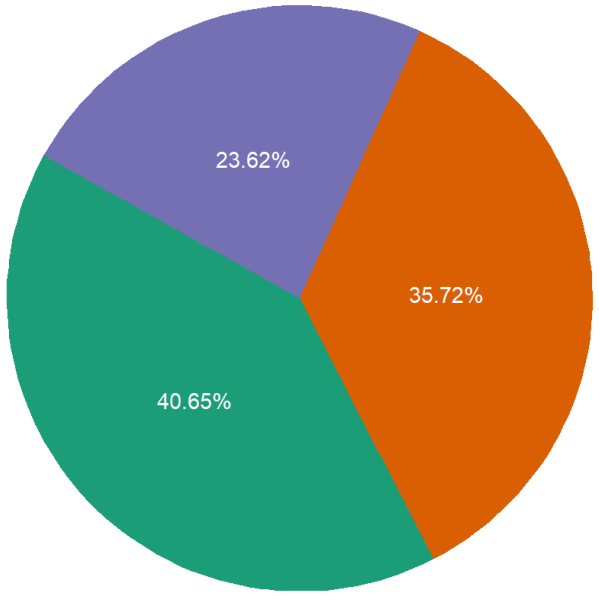


Histogram

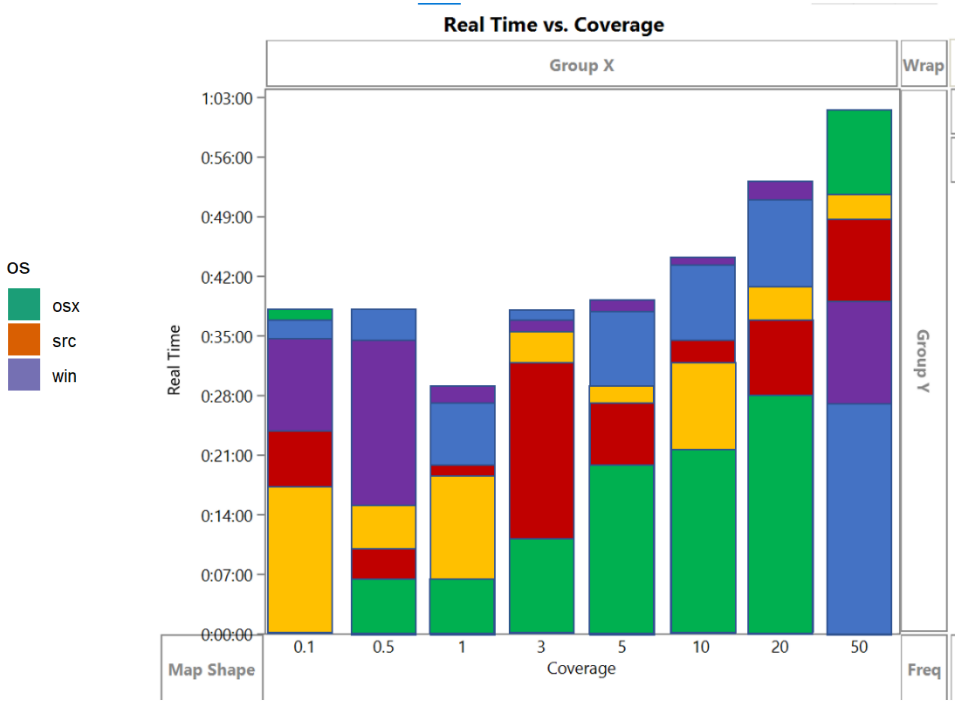
Density Plot

Boxplot

RATIO



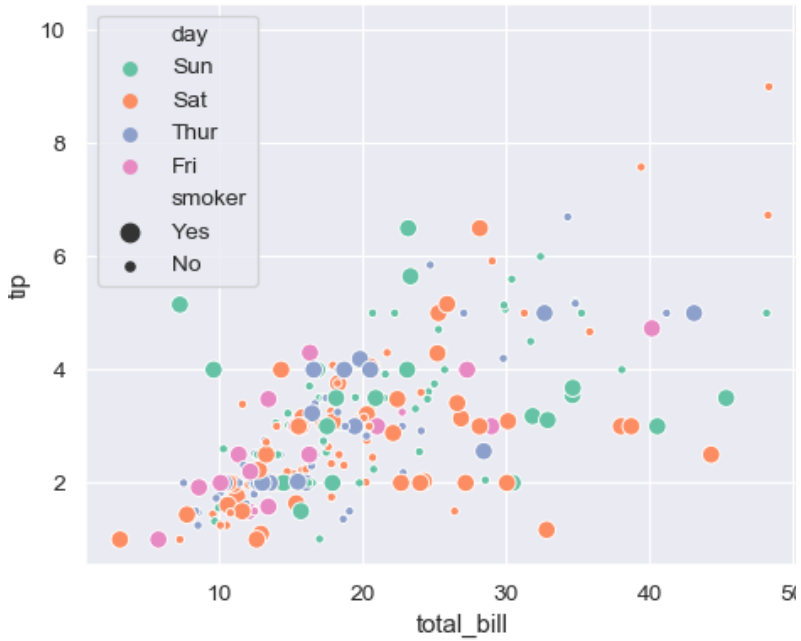
Piechart



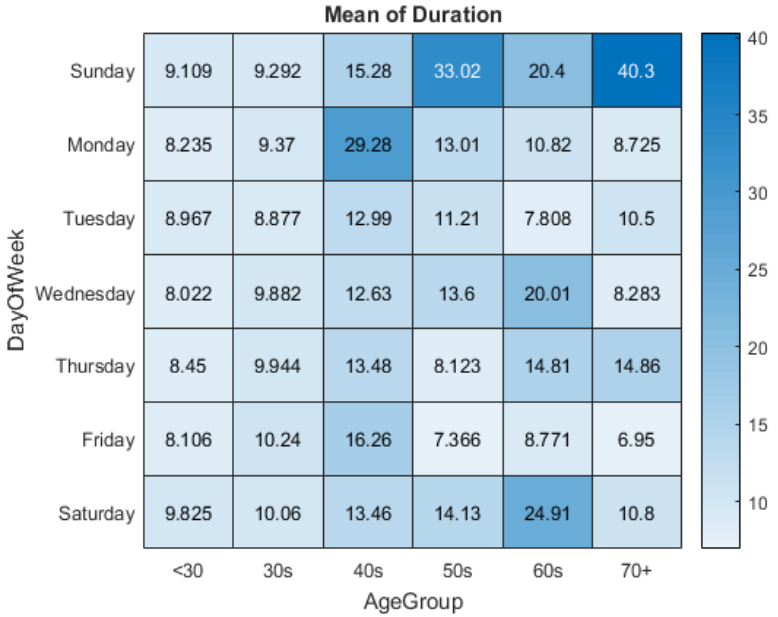
Stacked bar graph

Treemap

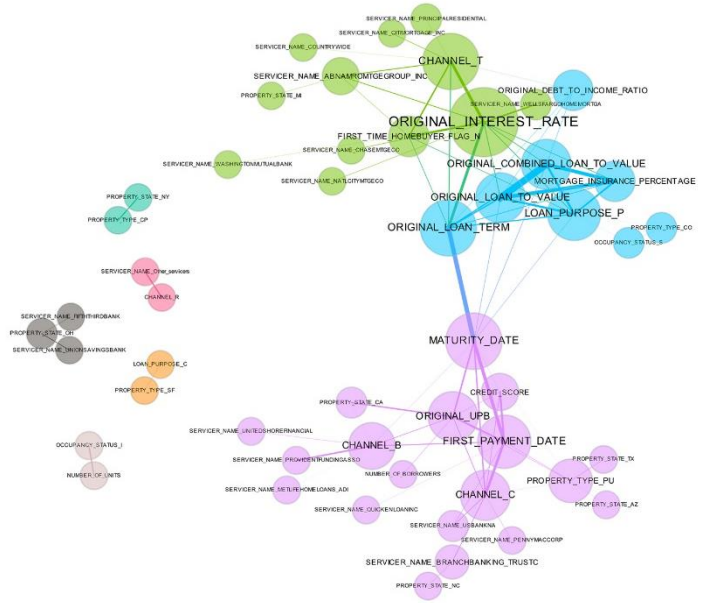
RELATIONS



Scatter plot

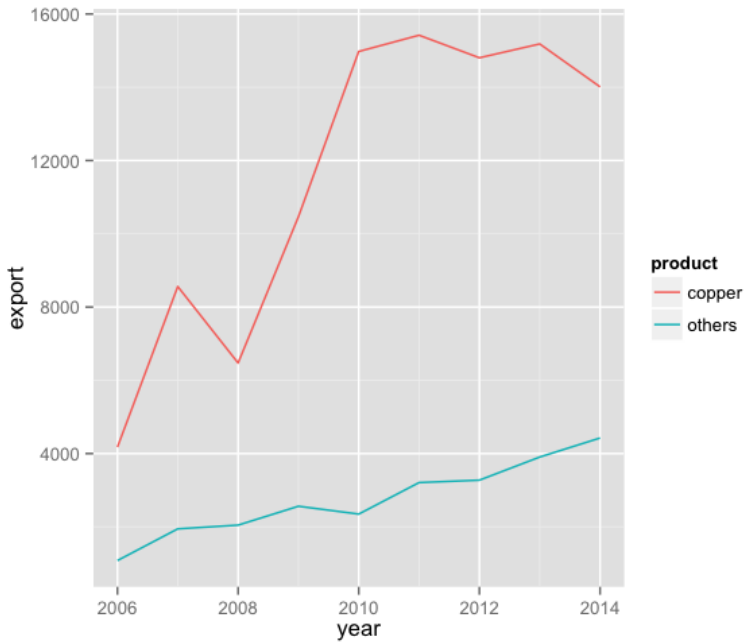


Heatmap

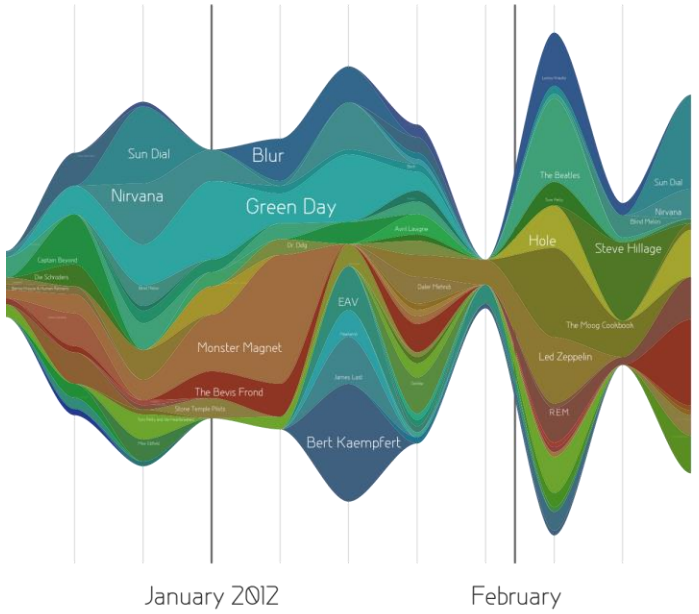


Network diagram

CHANGE

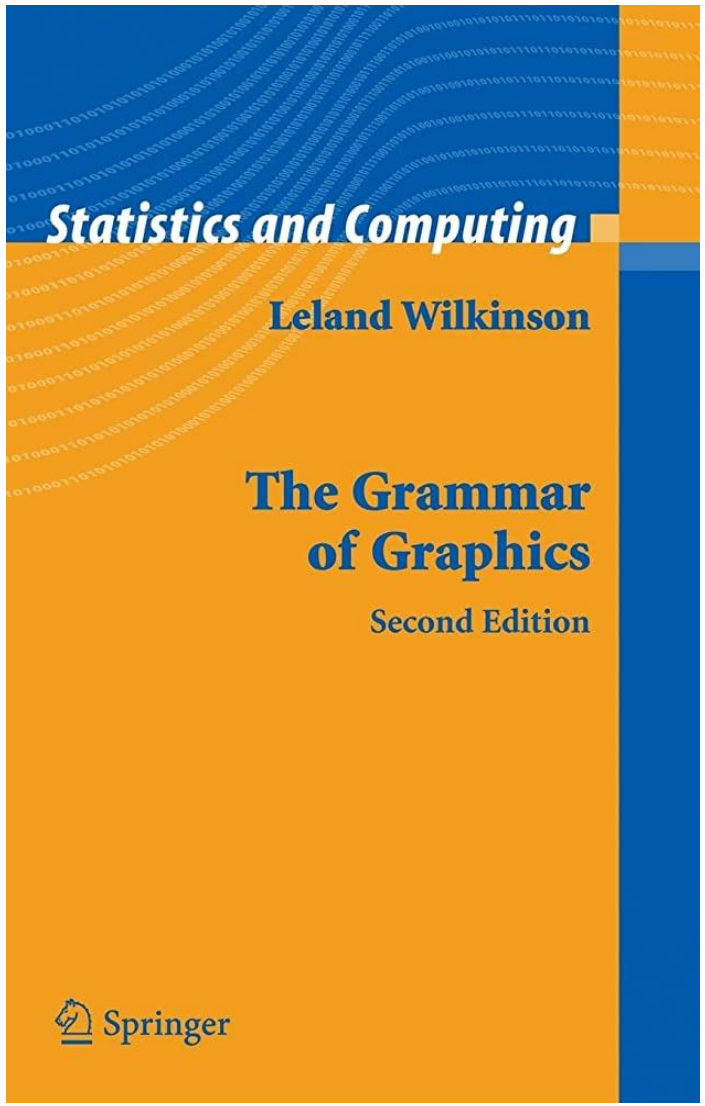


Line graph



Stream graph

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

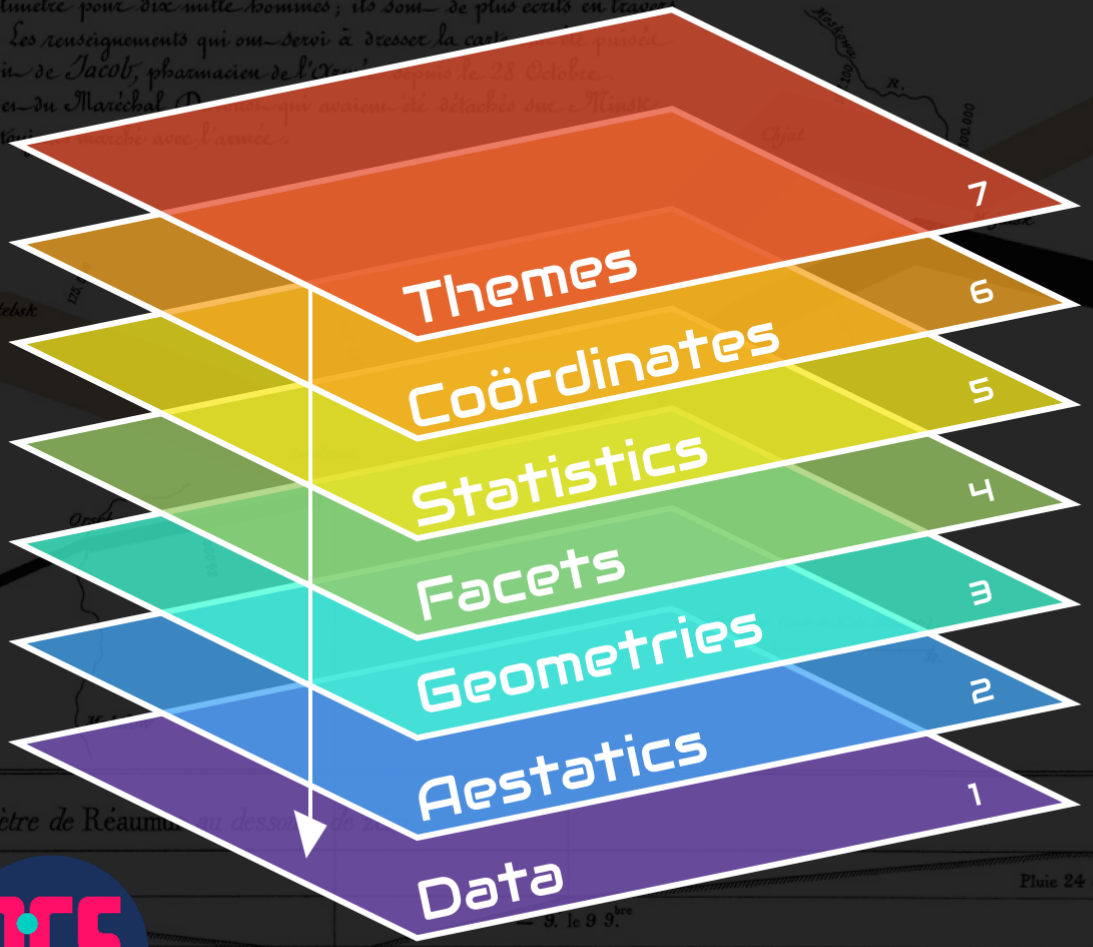
Statistical models & summaries

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



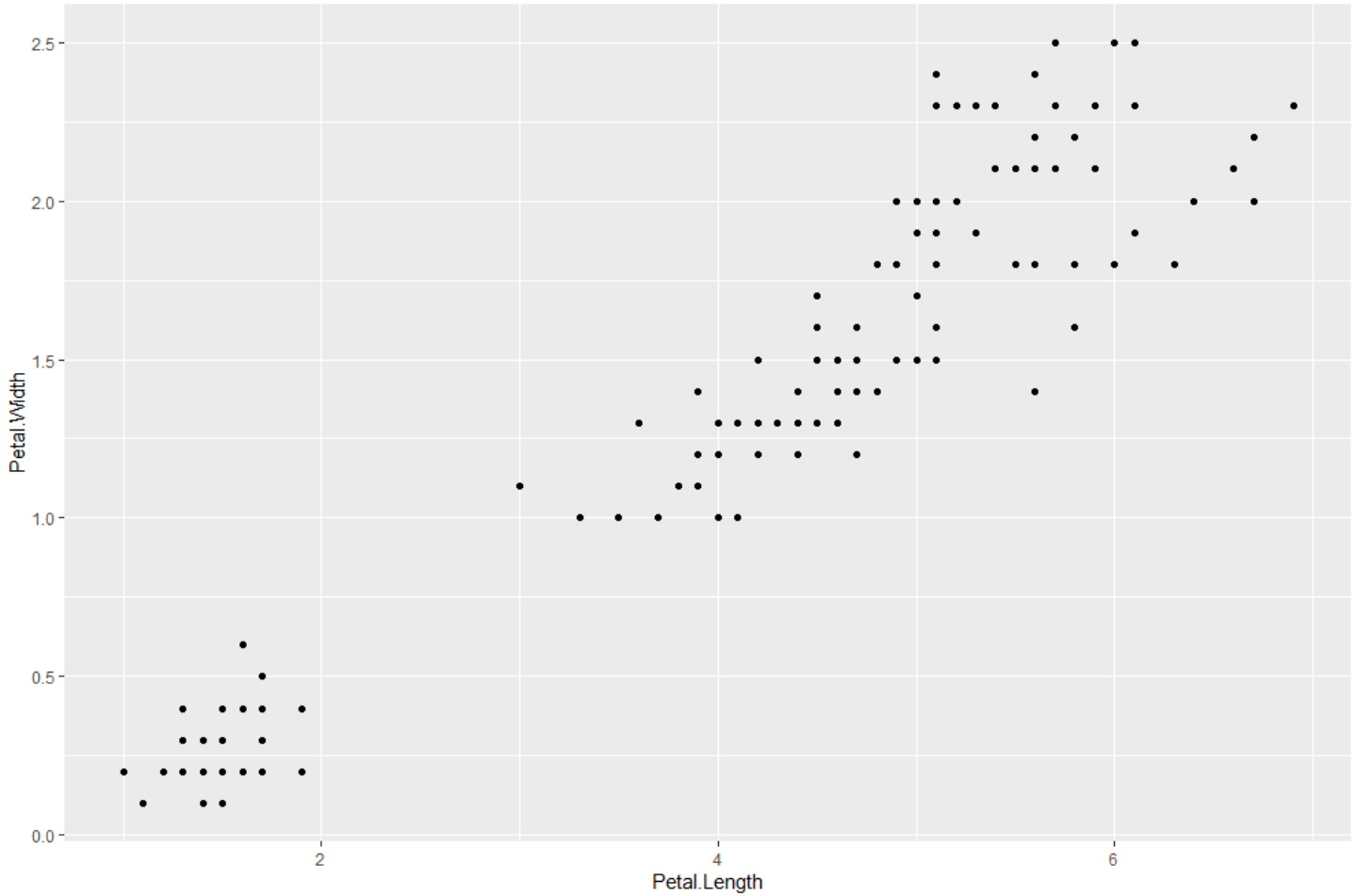
THE DATA

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa

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

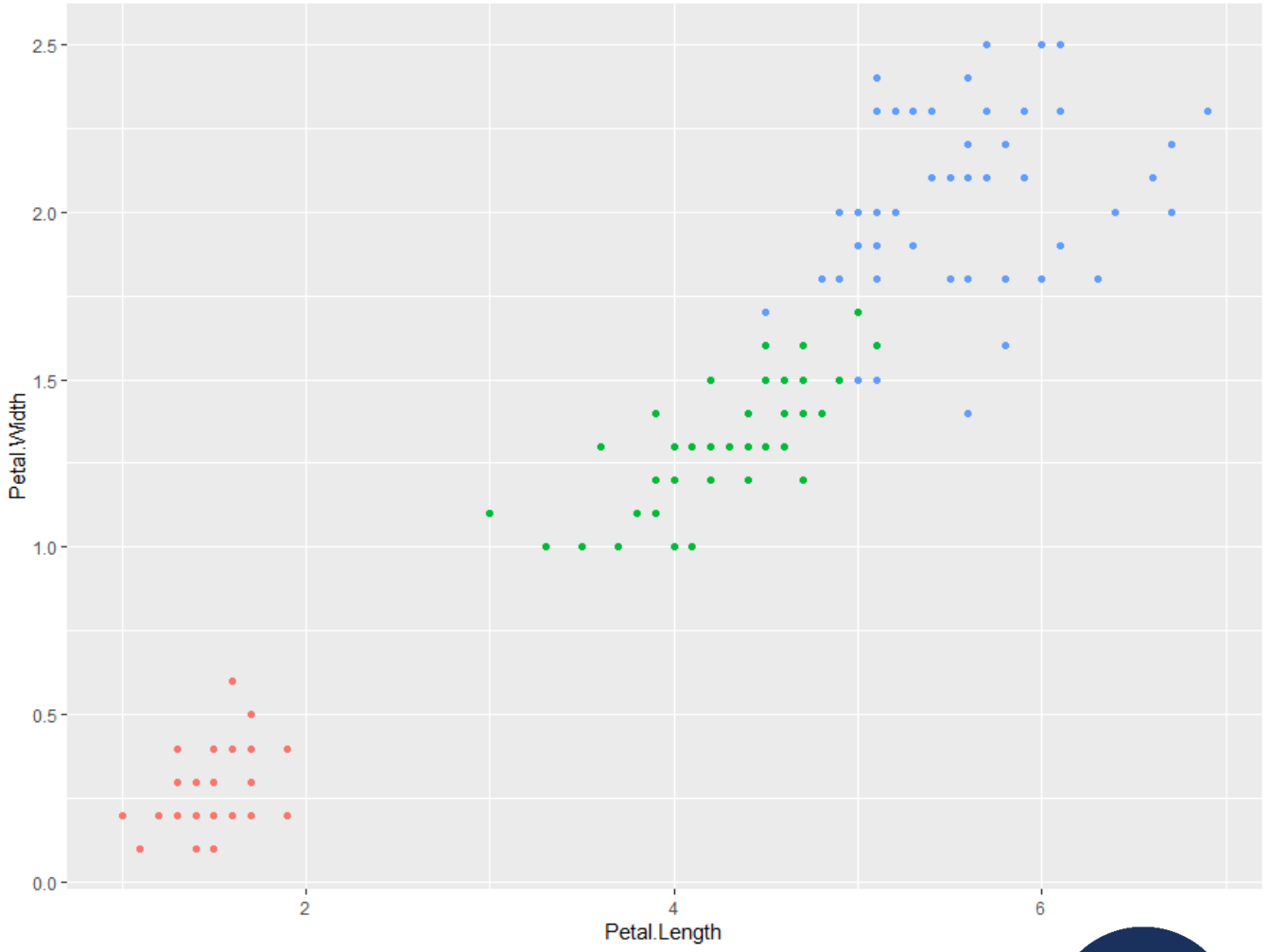
AESTHETIC AND GEOMETRIES



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.

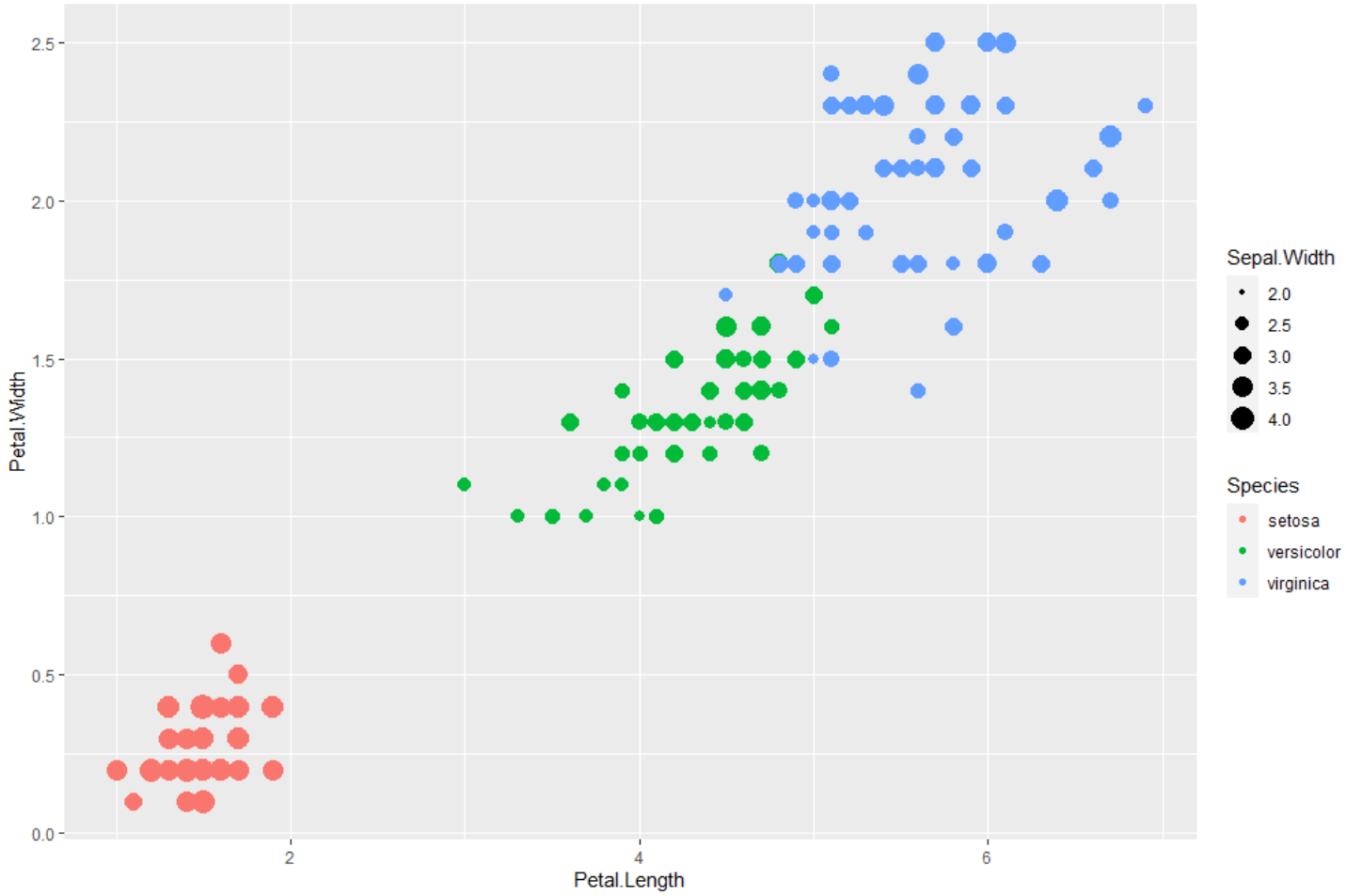
AESTHETIC



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**

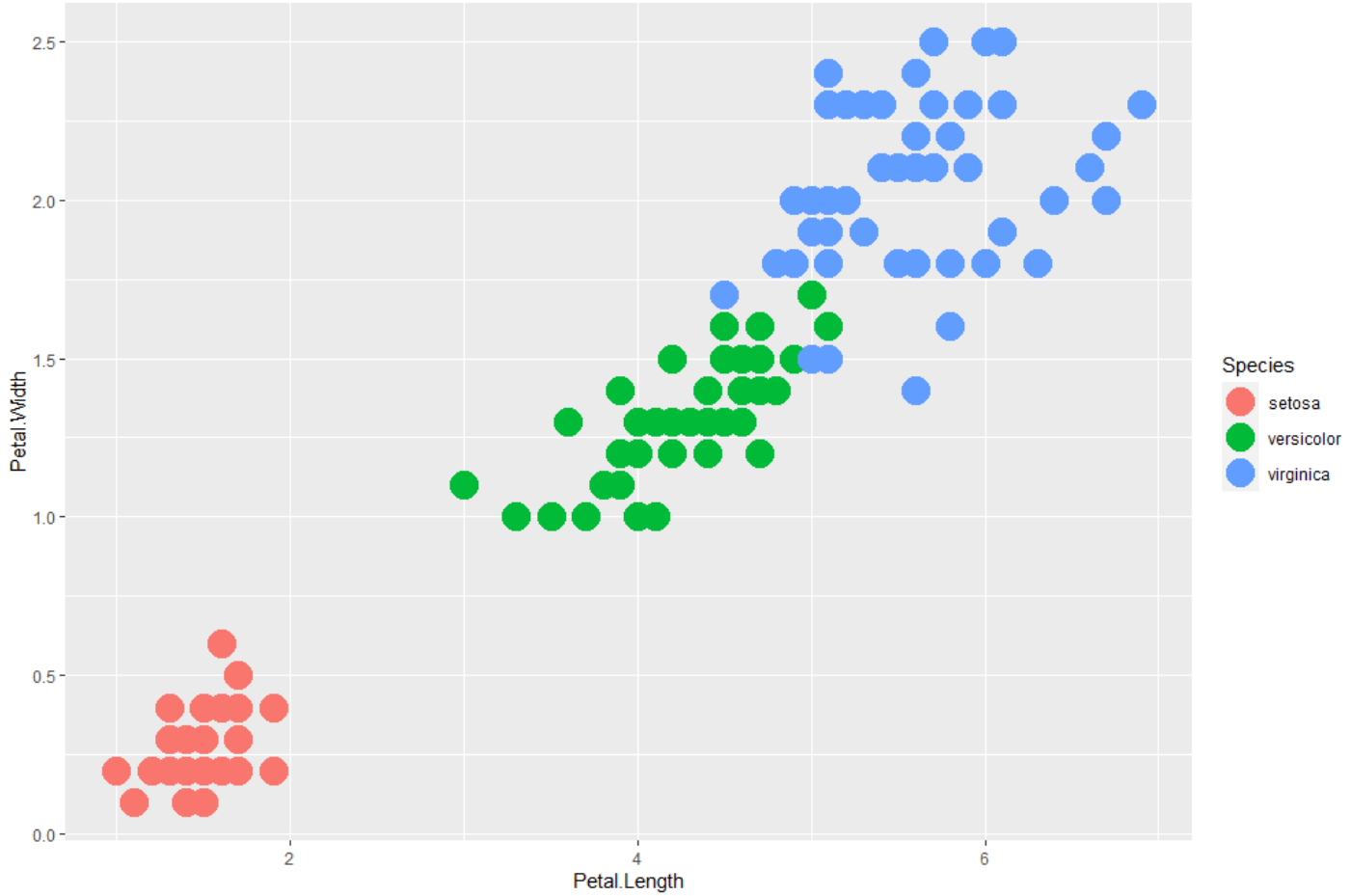
AESTHETIC



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**

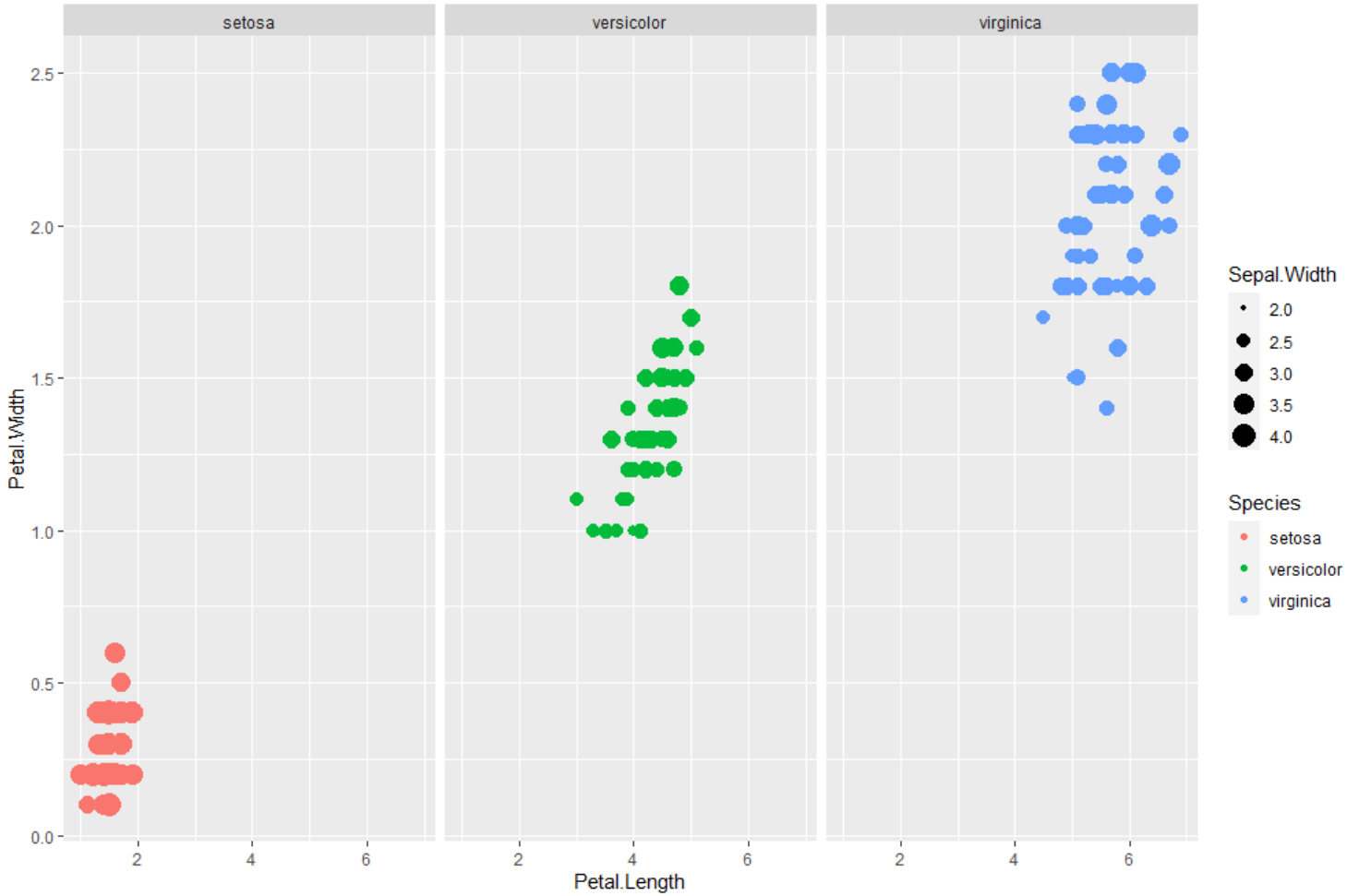
GEOMETRY



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.**

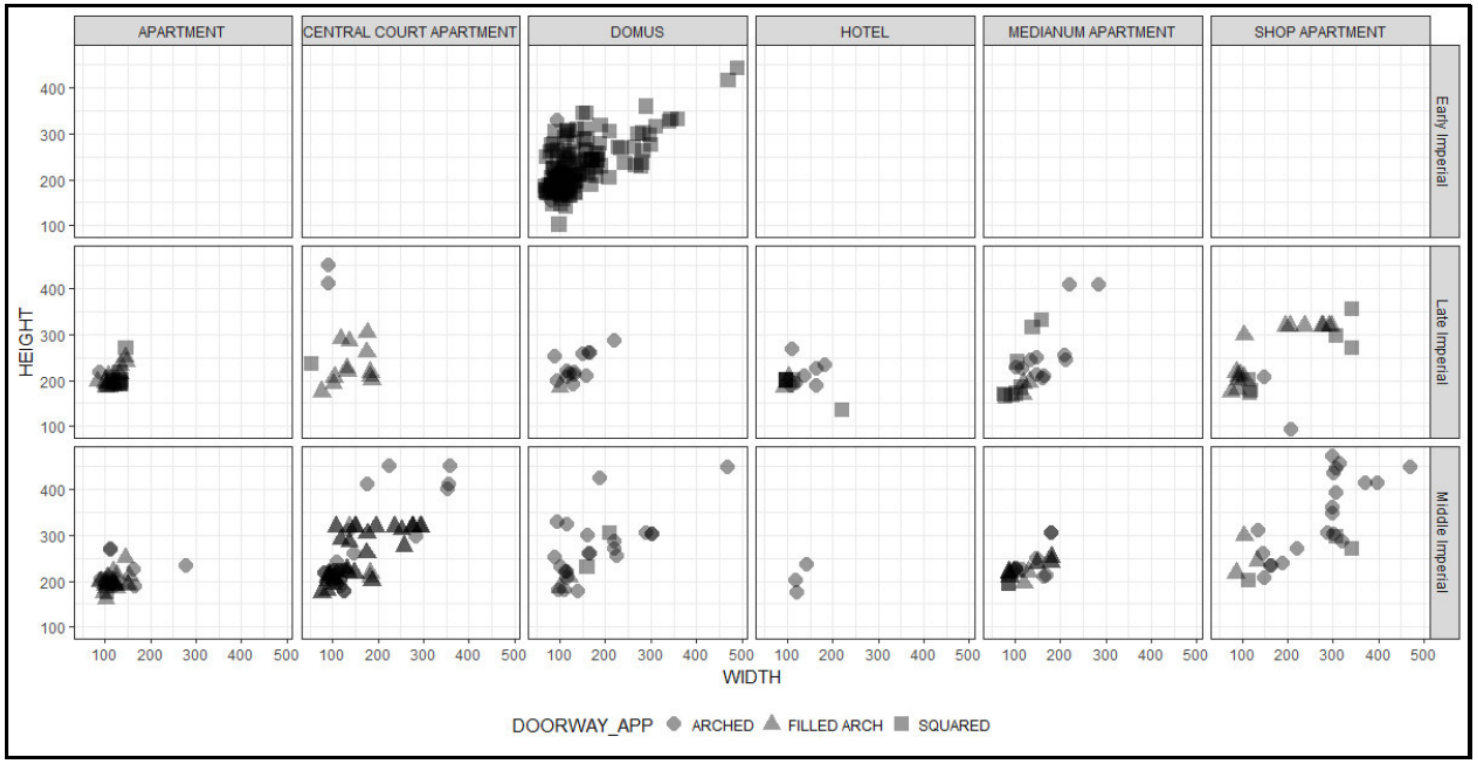
FACET



Level 4 Subplot

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

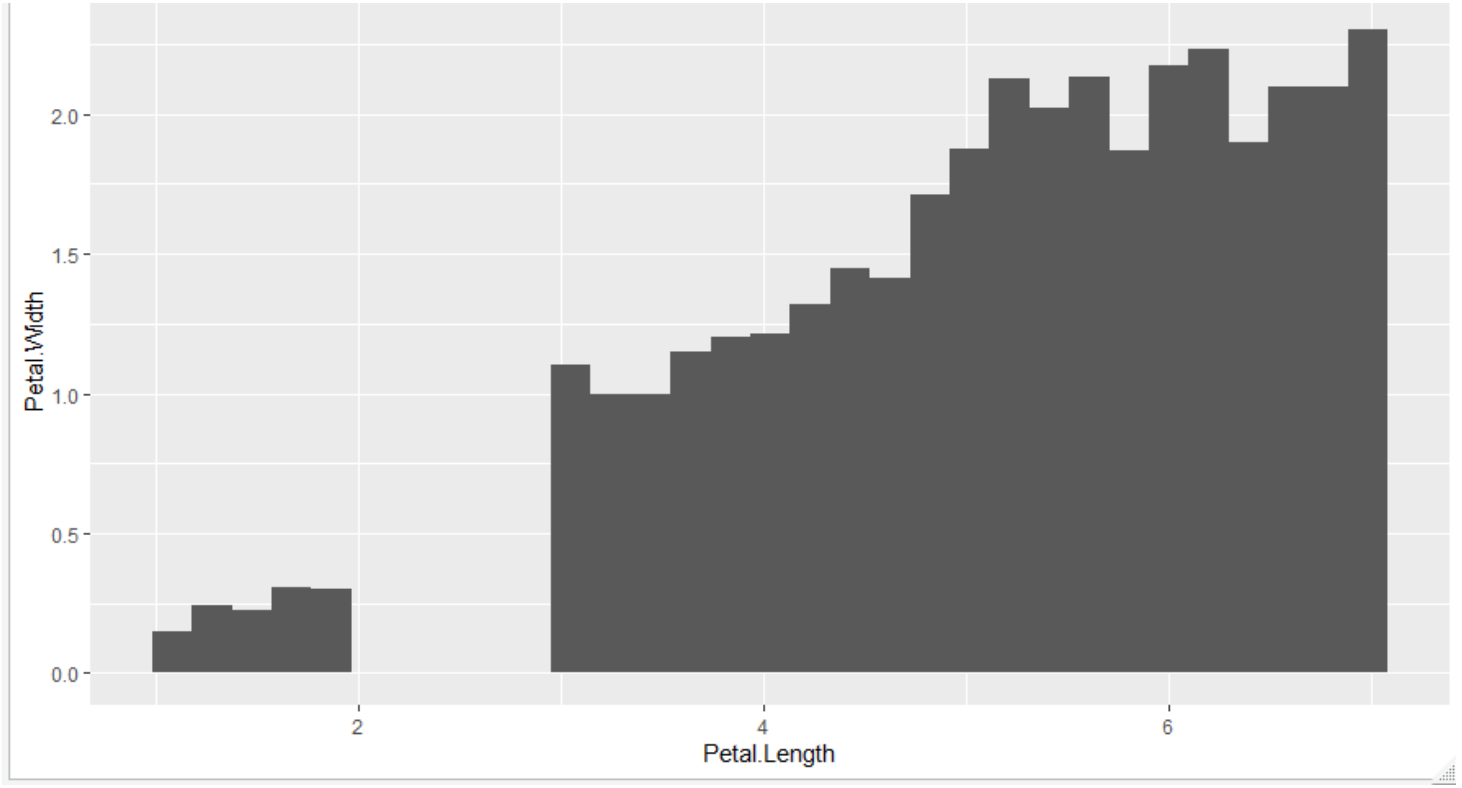
FACET



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**

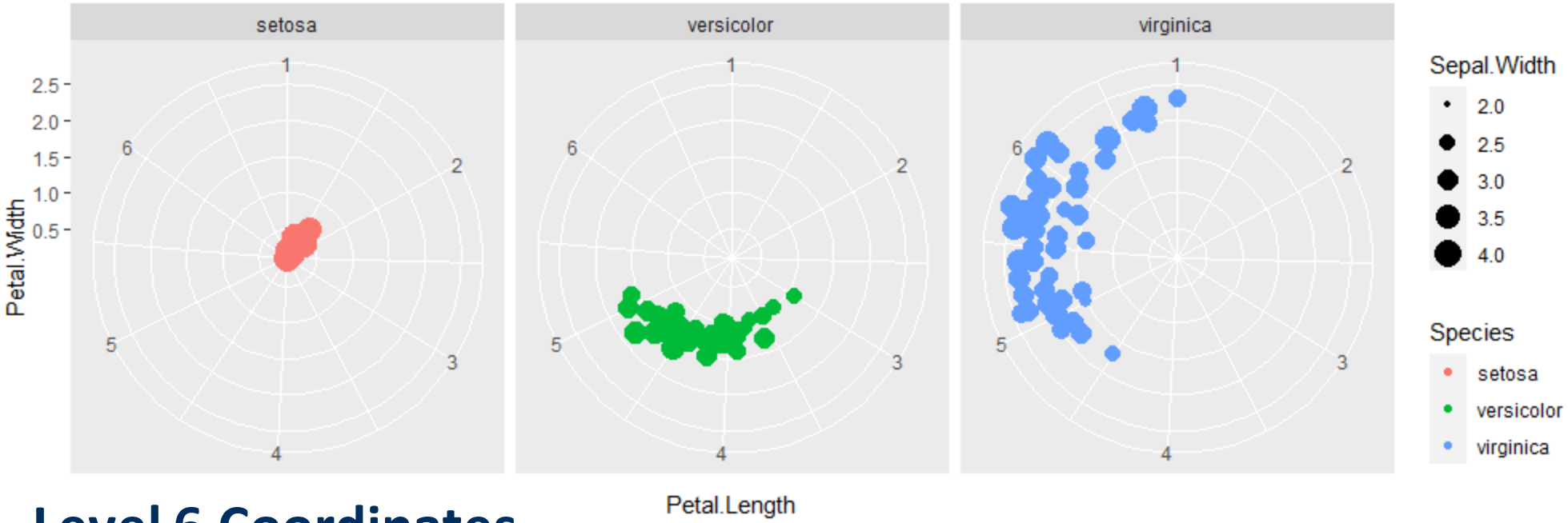
STATISTICS



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**

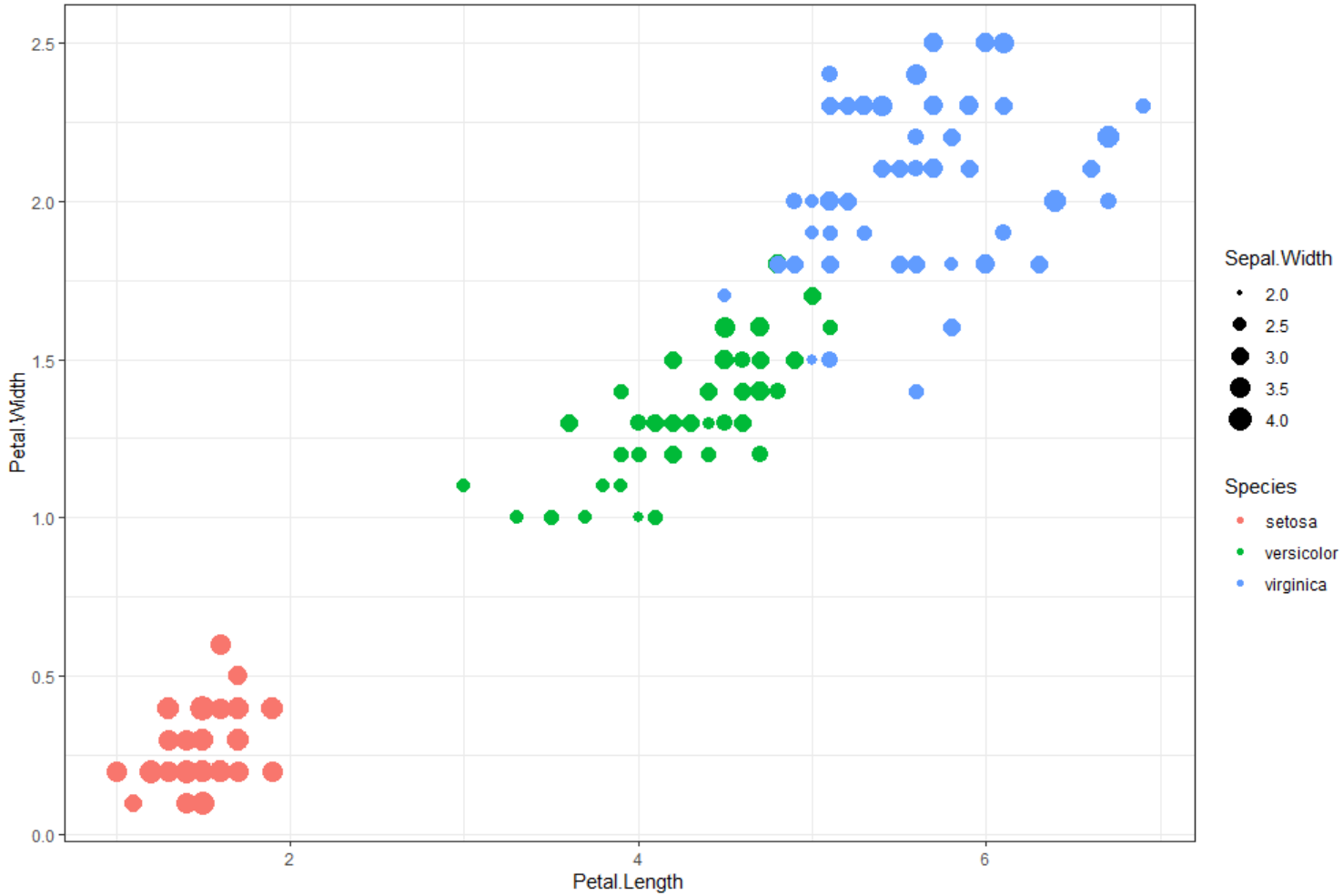
COORDINATES



Level 6 Coordinates

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

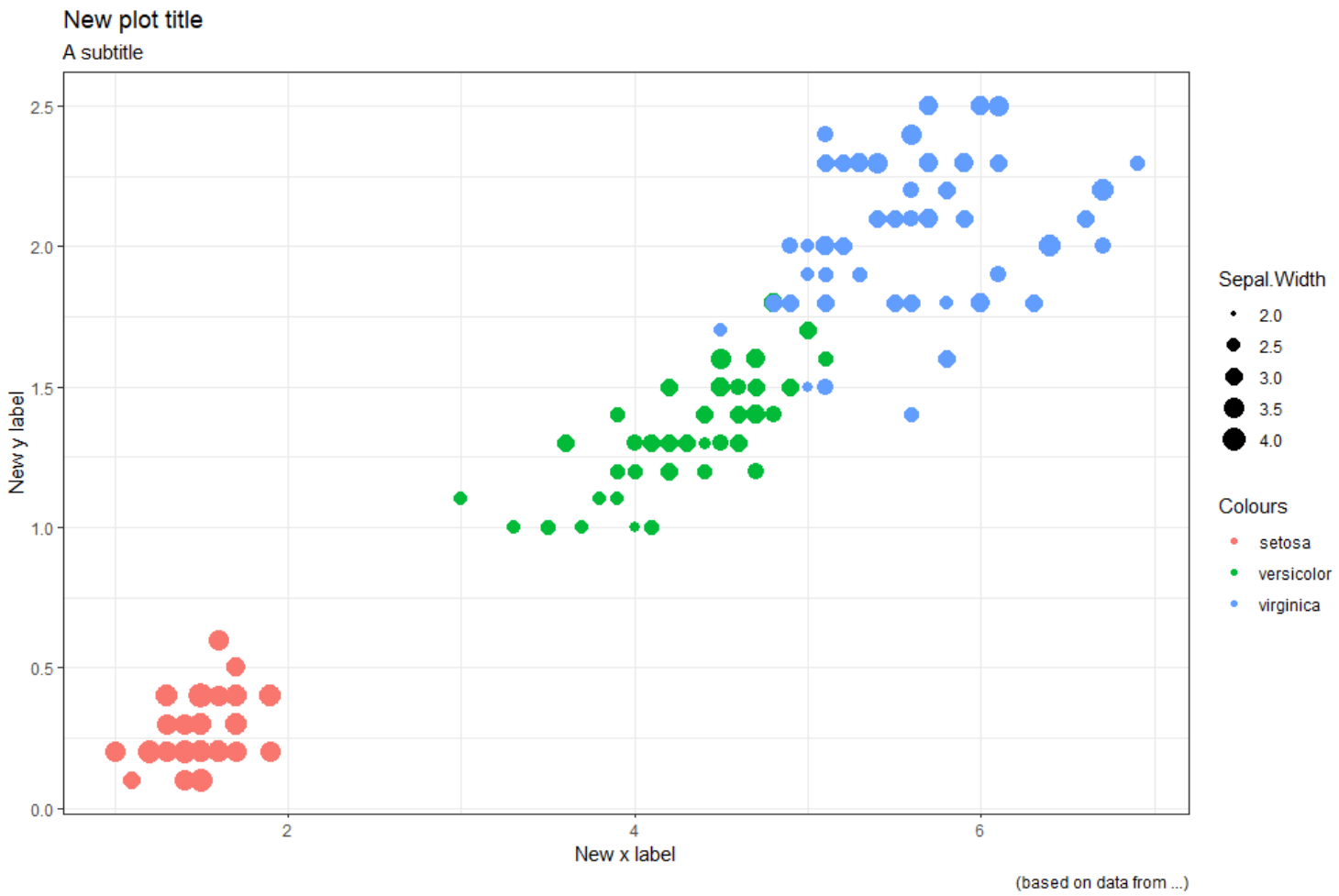
THEMES



Level 7 Global settings

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

LABELS



Level 7 labels

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


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

TIME FOR R

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

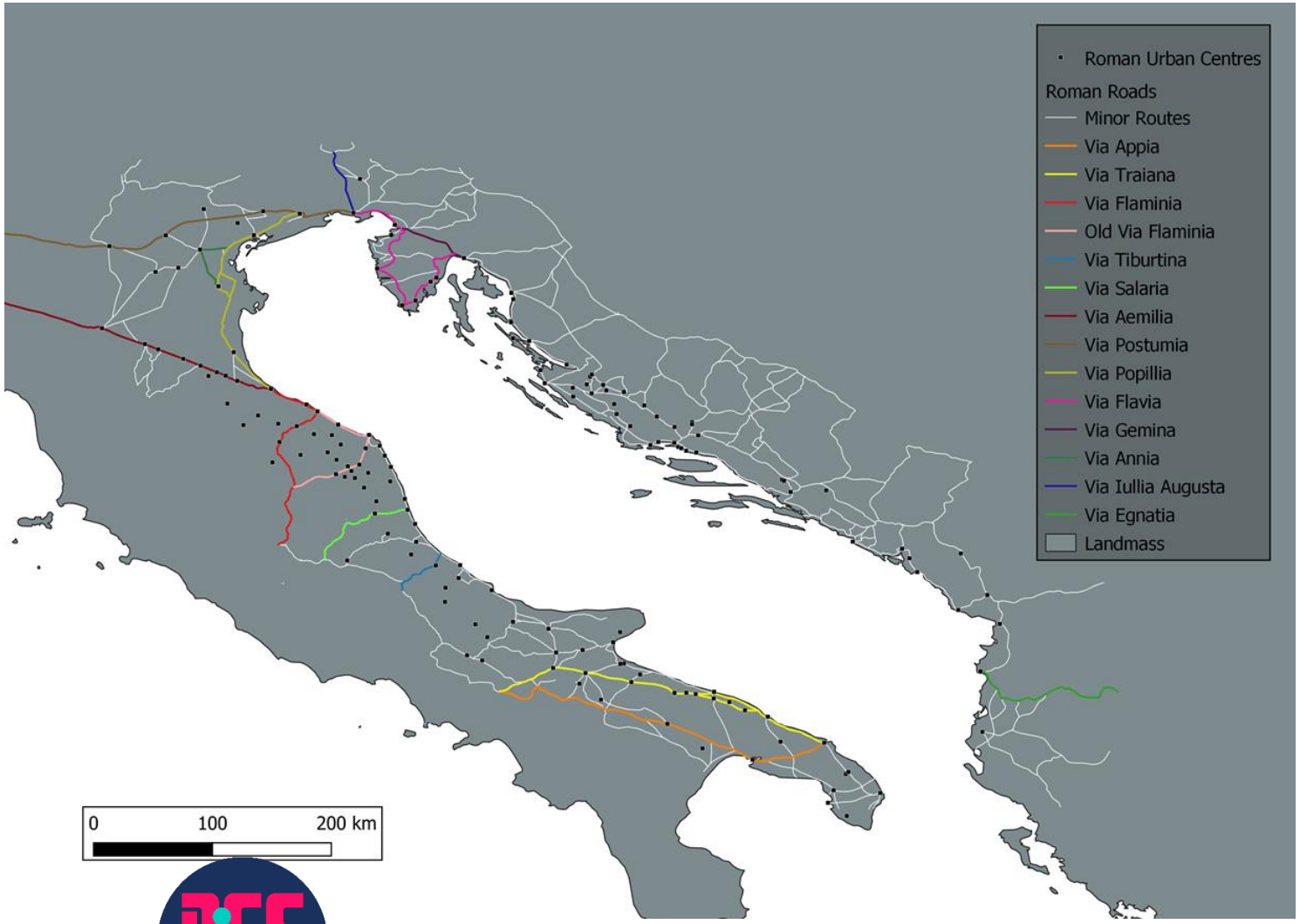
LUNCH BEAK

WE ARE GOING TO RESTART AT 13:30

GEOGRAPHICAL DATA

Vectors

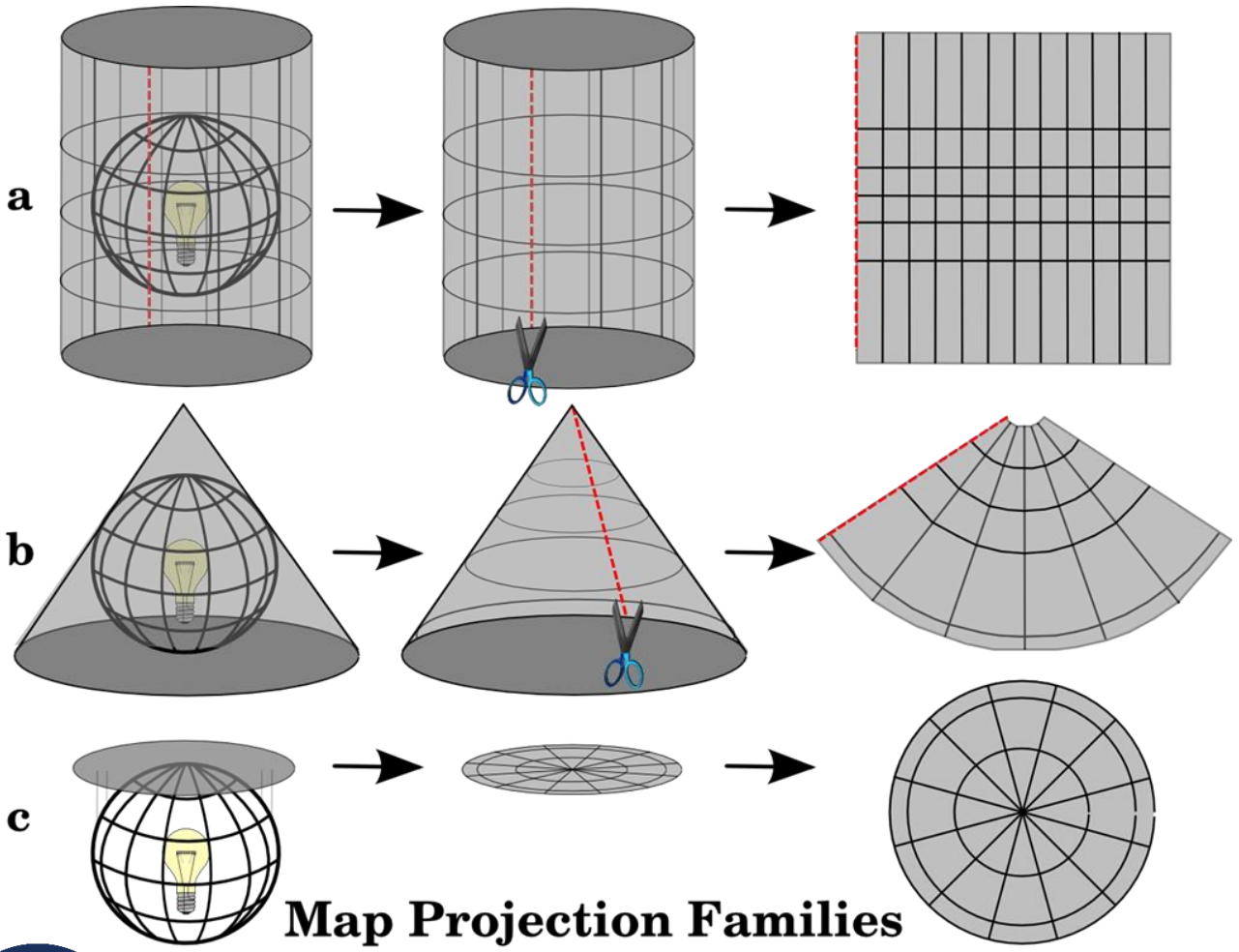
- Generalised representation of the real world
- Points, lines or polygons
- Normally in .shp (shapefile) format



GEOGRAPHICAL DATA

Projections and Coordinate Reference Systems (CRS)

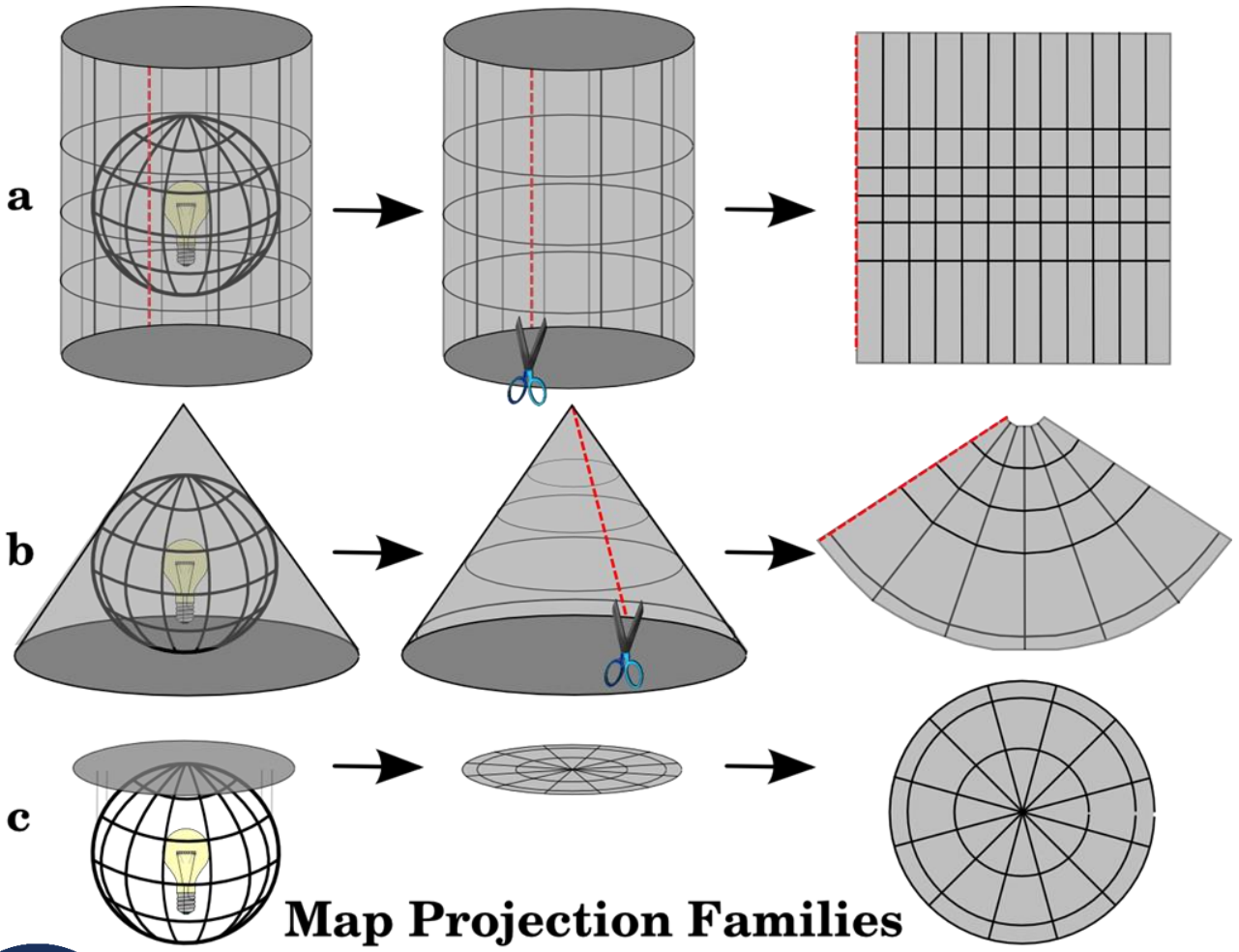
- Maps, whether physical or digital, generally portray a three dimensional space (the planet) on a 2 dimensional medium
- CRS are the systems that, with the use of coordinates, define how the 3D data is projected onto the 2D output



GEOGRAPHICAL DATA

Projections and Coordinate Reference Systems (CRS)

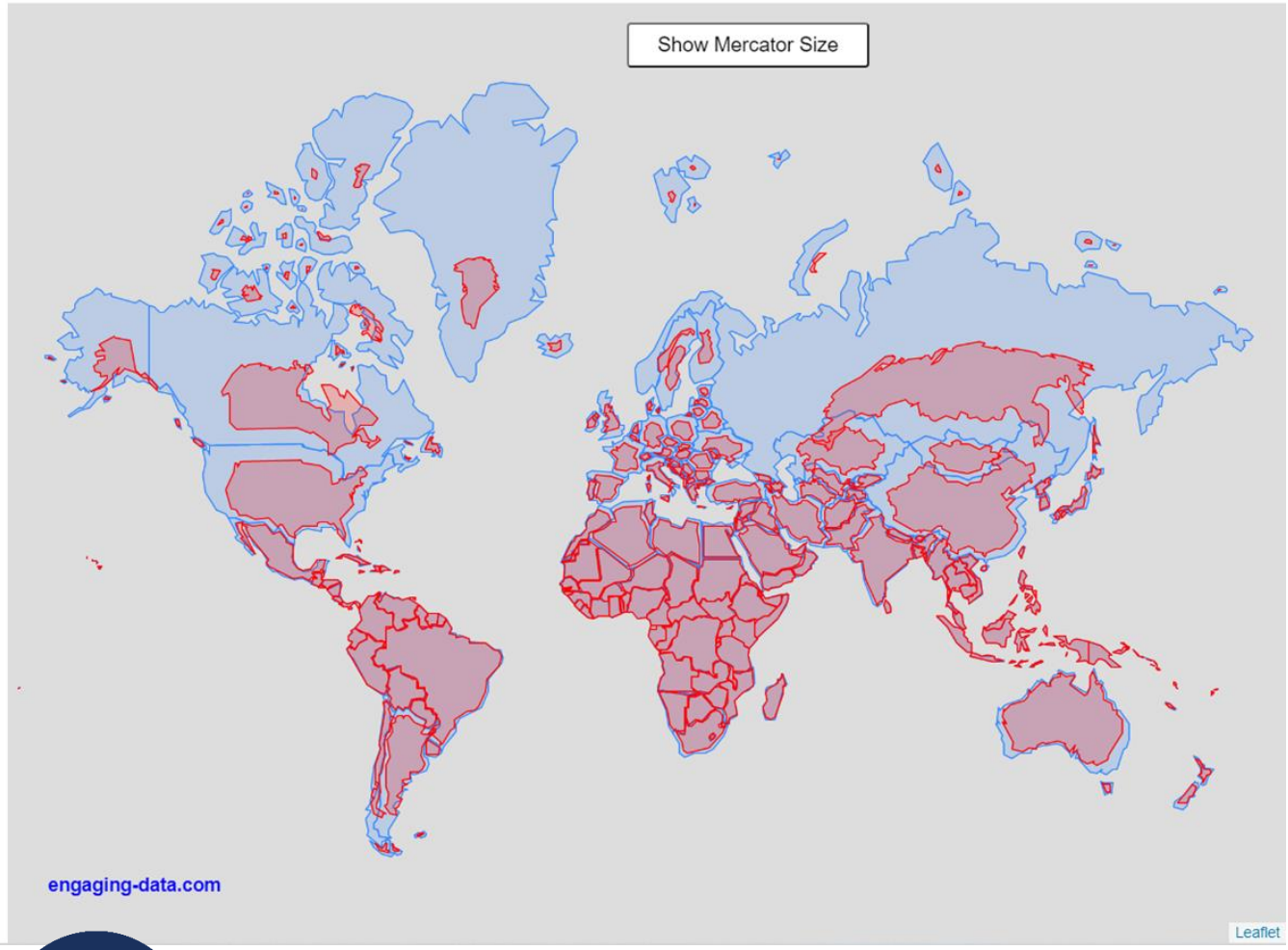
- There are three main map projection families
  - A) cylindrical
  - B) Conical
  - C)Planar
- Each has advantages and disadvantages
- None actually change the data itself, simply the way it is presented/projected



GEOGRAPHICAL DATA

Projections

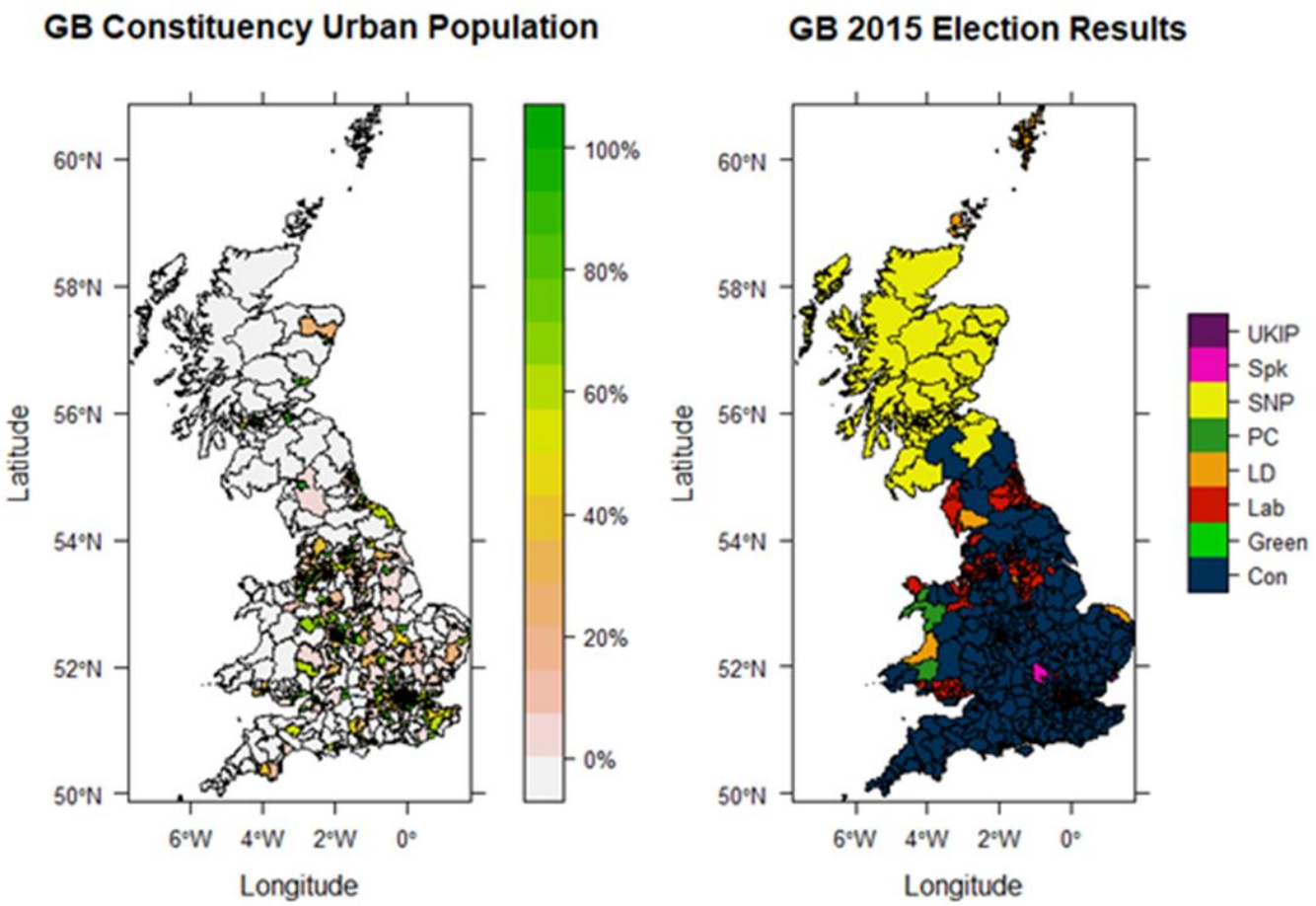
- A common projection is Mercator
- No projections are entirely accurate, they are representations of reality, not reality itself



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




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

TIME FOR R


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
COFFEE BEAK
WE ARE GOING TO RESTART AT 15:30

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

NEXT STEPS

THE CDCS TEAM

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

OUR RESEARCH QUESTIONS:

- What did we learn about the cost of living crisis in Scotland?
- What patterns did our data reveal?

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

HOW IS THE COST OF LIVING CRISIS PORTRAYED BY THE SCOTTISH AND UK GOVERNMENTS?

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

WHAT ARE COMMON THEMES IN HOW THE COST OF LIVING CRISIS IS PRESENTED?

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

IS THE COST OF LIVING CRISIS GETTING BETTER IN 2023?

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

ARE CERTAIN VARIABLES BETTER INDICATORS THAN OTHERS FOR THE IMPACT OF THE COST OF LIVING CRISIS?

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

HAS LOCATION PLAYED A ROLE IN HOW THE COST OF LIVING CRISIS AFFECTS LOCAL AUTHORITIES?

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

THE PROBLEMS OF WORKING WITH REAL WORLD DATA

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

BYOD HIGHLIGHTS

ATTENDEES

NEXT STEPS?

- Data Carpentries: https://datacarpentry.org/
- CDCS Training Programme: https://www.cdcs.ed.ac.uk/training/training-pathways
- Stack Exchange: https://stats.stackexchange.com/questions/tagged/r

Data Upskilling short courses

- Courses include:
- Data Carpentry courses – learn to organise, clean and analyse data (entry level – no prior coding experience necessary):
  - Data Cleaning and Organising with Python
  - Data Cleaning and Organising with R
  - R, Regular Expressions, SQL
  - Introduction to Statistics in R
- Programming Skills
- Software Development
- Leading Technology and Innovation in Organisations
- Introduction to Data Ethics for Business

MORE INFORMATION: WEB SCRAPING

- https://www.datacamp.com/tutorial/r-web-scraping-rvest
- https://rvest.tidyverse.org/
- https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/
- https://cran.r-project.org/web/packages/rvest/vignettes/rvest.html
- https://www.youtube.com/watch?v=Dkm1d4uMp34
- https://jtr13.github.io/cc19/web-scraping-using-rvest.html
- https://cran.r-project.org/web/packages/rvest/rvest.pdf

MORE INFORMATION: TEXT ANALYSIS

- https://programminghistorian.org/en/lessons/topic-modeling-and-mallet
- https://www.tidytextmining.com/
- https://www.matthewjockers.net/text-analysis-with-r-for-students-of-literature/
- https://rpubs.com/tsholliger/301914
- https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
- https://tedunderwood.com/2012/08/14/where-to-start-with-text-mining/
- https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/
- https://www.tidytextmining.com/topicmodeling.html
- https://ladal.edu.au/topicmodels.html
- https://towardsdatascience.com/beginners-guide-to-lda-topic-modelling-with-r-e57a5a8e7a25
- http://dirichlet.net/pdf/wallach06topic.pdf

MORE INFORMATION: SENTIMENT ANALYSIS

- https://www.matthewjockers.net/2015/02/02/syuzhet/
- https://tedunderwood.com/2015/03/24/why-its-hard-for-syuzhet-to-be-right-or-wrong-yet/
- https://www.wiley.com/en-us/Sentiment+Analysis%3A+A+Practitioner%27s+Guide+to+Emotion+Detection+for+Computer+Vision+and+Natural+Language+Processing-p-9781118745585
- https://www.tidytextmining.com/sentiment.html

MORE INFORMATION: DATA WRANGLING

- https://www.researchgate.net/publication/309343107_Ten_Simple_Rules_for_Digital_Data_Storage
- https://r4ds.had.co.nz/workflow-projects.html
- https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

MORE INFORMATION: REGRESSION AND NHT

- Brown, VA. (2021). An Introduction to Linear Mixed-Effects Modeling in R. Advances in Methods and Practices in Psychological Science. 4(1). doi:10.1177/2515245920960351
- http://www.sthda.com/english/wiki/one-sample-t-test-in-r
- https://statistics.berkeley.edu/computing/r-t-tests
- https://www.statmethods.net/stats/ttest.html
- https://www.youtube.com/watch?v=pTmLQvMM-1M
- https://www.youtube.com/watch?v=NF5_btOaCig
- https://www.linkedin.com/learning/r-statistics-essential-training/comparing-means-with-the-t-test?autoplay=true&u=50251009
- http://www.sthda.com/english/wiki/one-way-anova-test-in-r
- https://www.datanovia.com/en/lessons/anova-in-r/
- https://www.r-bloggers.com/performing-anova-test-in-r-results-and-interpretation/
- https://www.scribbr.com/statistics/anova-in-r/
- https://www.youtube.com/watch?v=NF5_btOaCig
- https://www.youtube.com/watch?v=fT2No3Io72g
- https://www.linkedin.com/learning/r-statistics-essential-training/comparing-means-with-a-two-factor-anova?autoplay=true&u=50251009
- https://learningstatisticswithr.com/
- Field, Z., Miles, J., & Field, A. (2012). Discovering statistics using R. *Discovering Statistics Using R*, 1-992. (Chapter 7).

MORE INFORMATION: PCA AND CLUSTERING

https://setosa.io/ev/principal-component-analysis/
https://www.youtube.com/watch?v=FgakZw6K1QQ

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

https://towardsdatascience.com/principal-component-analysis-pca-101-using-r-361f4c53a9ff

https://www.linkedin.com/learning/r-statistics-essential-training/conducting-a-principal-components-factor-analysis?autoplay=true&u=50251009

https://www.youtube.com/watch?v=0Jp4gsfOLMs

https://www.datacamp.com/tutorial/pca-analysis-r

MORE INFORMATION: DATA VISUALISATION

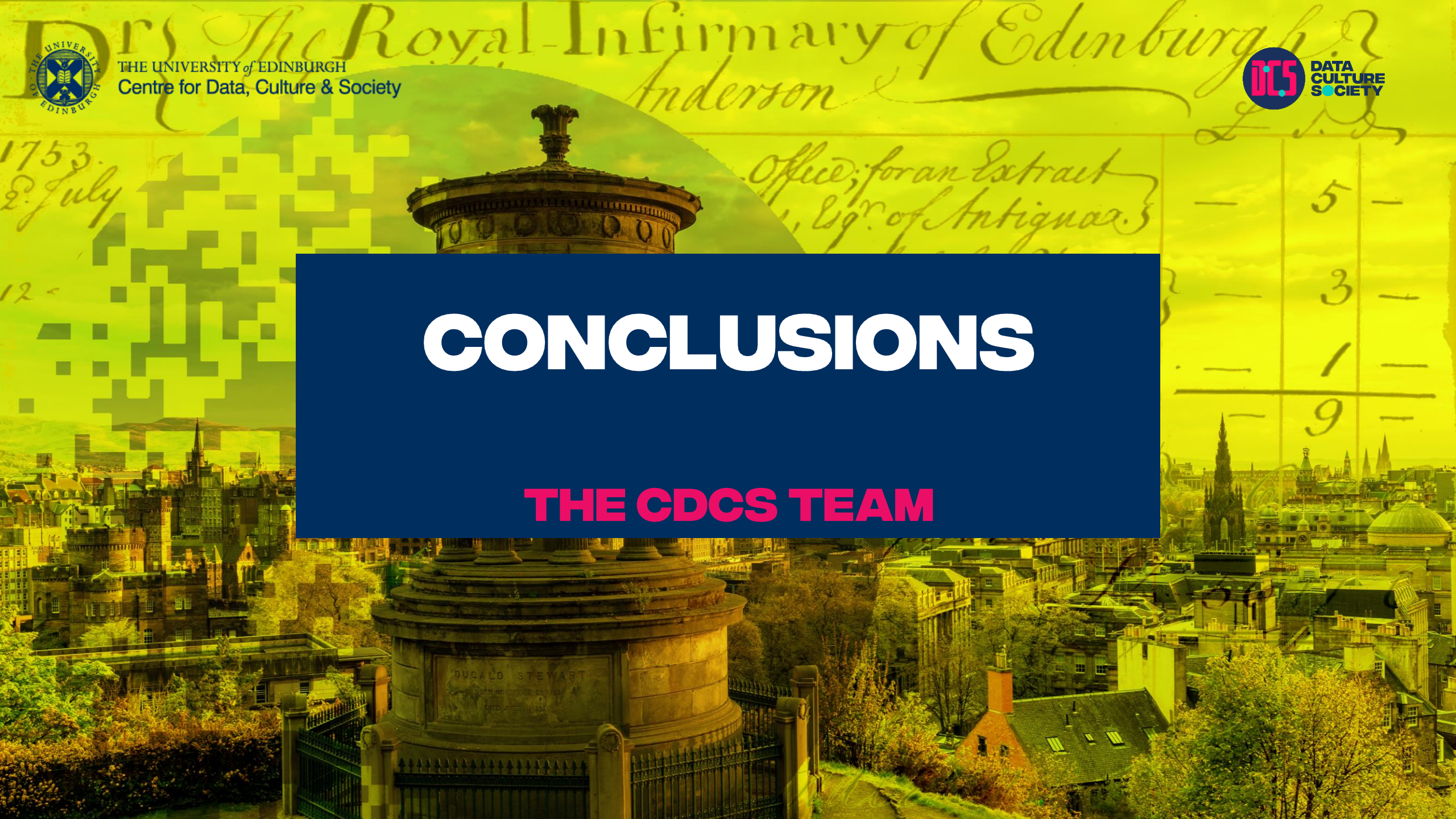
- https://www.linkedin.com/learning/data-visualization-in-r-with-ggplot2
- https://www.linkedin.com/learning/r-essential-training-wrangling-and-visualizing-data
- https://r-graph-gallery.com/
- https://www.safe.com/what-is/spatial-data/
- https://docs.qgis.org/3.16/en/docs/gentle_gis_introduction/coordinate_reference_systems.html
- https://www.youtube.com/watch?v=D3tdW9l1690
- https://engaging-data.com/country-sizes-mercator/
- https://thetruesize.com/#/aboutModal?borders=1~!MTU0MzY5MjU.NTEzMTkzNw*MzIyNDYwNjQ(NTI0NjA2NA

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

CONCLUSIONS

THE CDCS TEAM



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

THANK YOU!

THE CDCS TEAM

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

DINNER

PIZZA POSTO

16 Nicholson Street

Edinburgh EH8 9DH

www.cdcs.ed.ac.uk