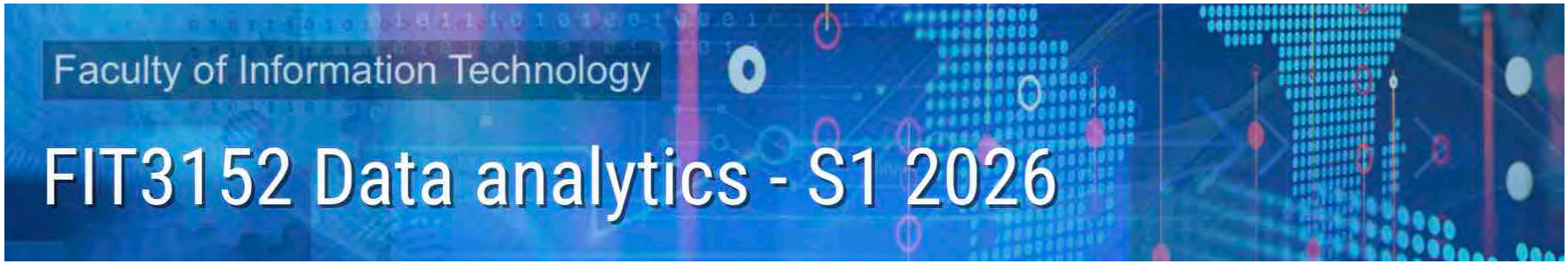


Lecture 2

- Visualising data
- Getting to know a data set
- Graphing data in R
- Assignment 1

Presenter: Dr John Betts

Week-by-week outline

Clayton seminar is Tuesday 12:00 – 2:00 pm.

Applied sessions begin Week 2, follows seminar by a week.

Week Starting	Seminar	Topic	App Ses	A1	A2	Q/P	A3	Due Date
2/3/2026	1	Introduction to Data Science, R, review of basic statistics	-					
9/3/2026	2	Data visualisation	S1					
16/3/2026	3	Data manipulation	S2					
23/3/2026	4	Regression modelling	S3					
30/3/2026	5	Clustering	S4					
6/4/2026	-	Mid-semester Break						
13/4/2026	6	Classification using decision trees	S5					17/4/2026
20/4/2026	7	Improving and evaluating classifiers. Naïve Bayes classification	S6					
27/4/2026	8	Ensemble methods, Artificial Neural Networks	S7					
4/5/2026	9	Network analysis	S8					
11/5/2026	10	Introduction to text analysis	S9					15/5/2026
18/5/2026	11	Text analysis applications	Quiz/Prac					22/5/2026
25/5/2026	12	Text Network Analysis, Review of the unit, Assignment 3	S10,11,12					
1/6/2026		SWOT VAC	-					
8/6/2026		EXAM PERIOD	-					12/6/2026

Assessment details

Assignment 1, Due 17th April, Weighting 25%

- Covers data manipulation, visualisation, and data analysis using a variety of techniques. Submission is a written report and short video explaining the key findings of your research.

Assignment 2, Due 15th May, Weighting 20%

- Covers machine learning/artificial intelligence models using R. Submission is a written report and short video.

Assignment 3, Due 12th June, Weighting 25%

- Covers text analysis, networks and clustering using R. Submission is a written report and short video.

Quiz + Practical Activity, Week 11 (Due 22nd May), Weighting 30%

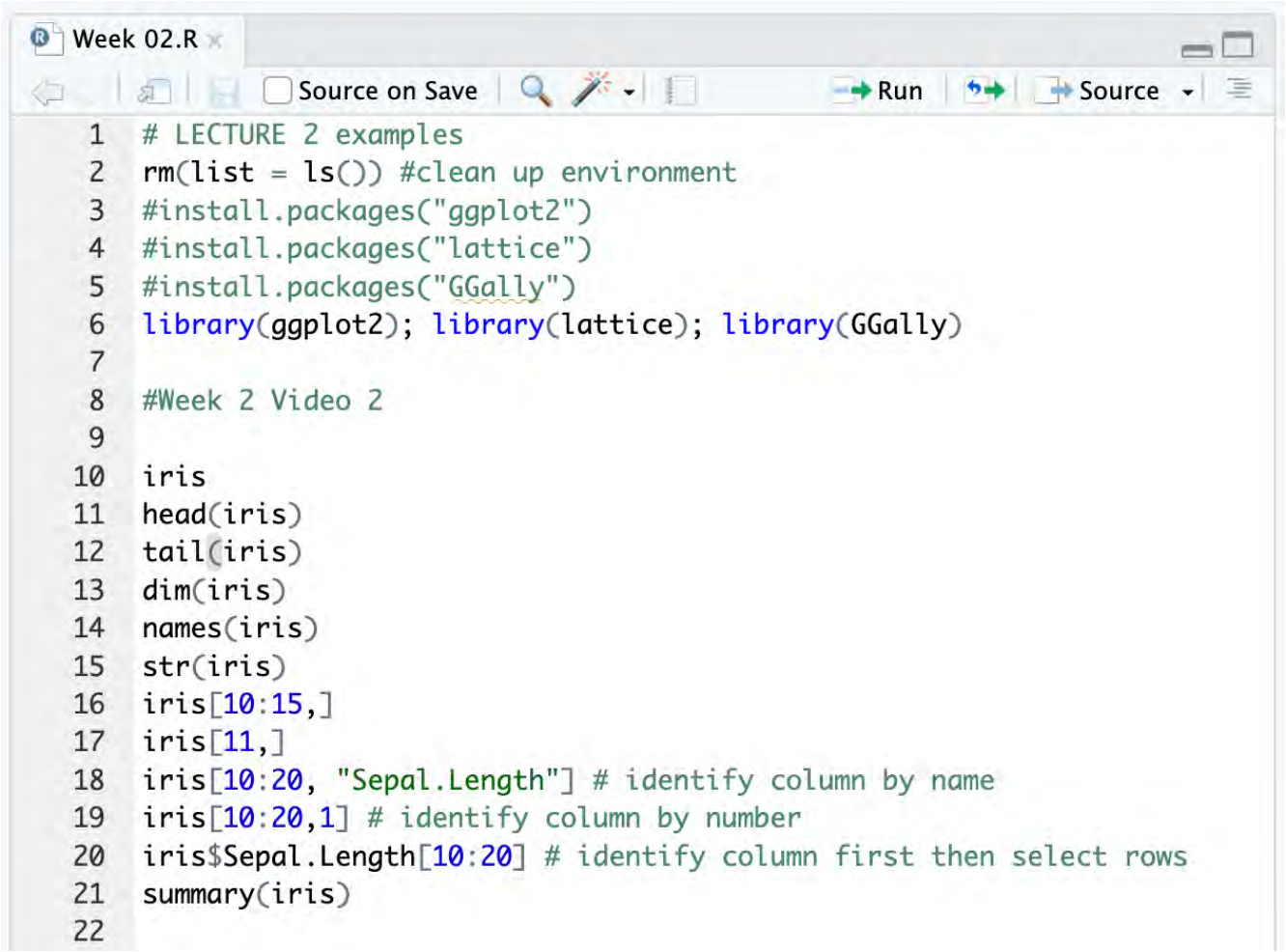
- You will do practical activities and quiz style questions under supervision during your applied session. Content will cover topics from Weeks 1 – 9.

Note: Scripts

Scripts allow you to save your working from session to session.

- Use them to automate environment settings etc.
- Create a new script: File > New File > R Script
- Save with a filename
- Use “Source” to evaluate on the fly
- Note: # comments, pre-emptive text
- Next slide shows previous example as a script…

Scripts: example from today's notes



Visualising data

Why make a visual display of your data or results of your analysis?

Visualising data

This topic explores data visualisation for analytics and creating graphics in R.

- By the end of this topic, you'll be able to:
- Understand the importance and purpose of data visualisation.
- Generate plots using base graphics and more advanced plots in R using the ggplot2 package.
- Interpret and communicate insights from visual representations of data.
- Explore the structure of a dataset through visual summaries.

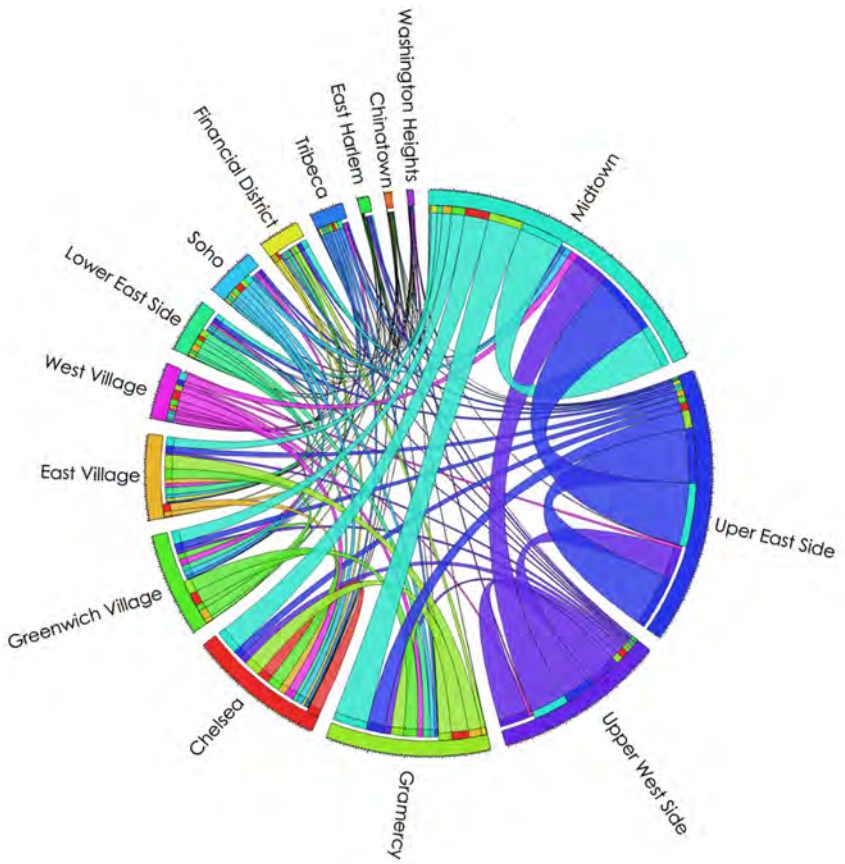
Visualising data

Some examples of data graphics follow. For each image think about:

- What information is being conveyed? *What message is being told by the data?*
- How is information being conveyed? What is the main device used: size, shape, colour, position...?
- How many dimensions (*number of variables associated with each data point*) are presented?
- How is space used?

New York taxi trips by neighbourhood

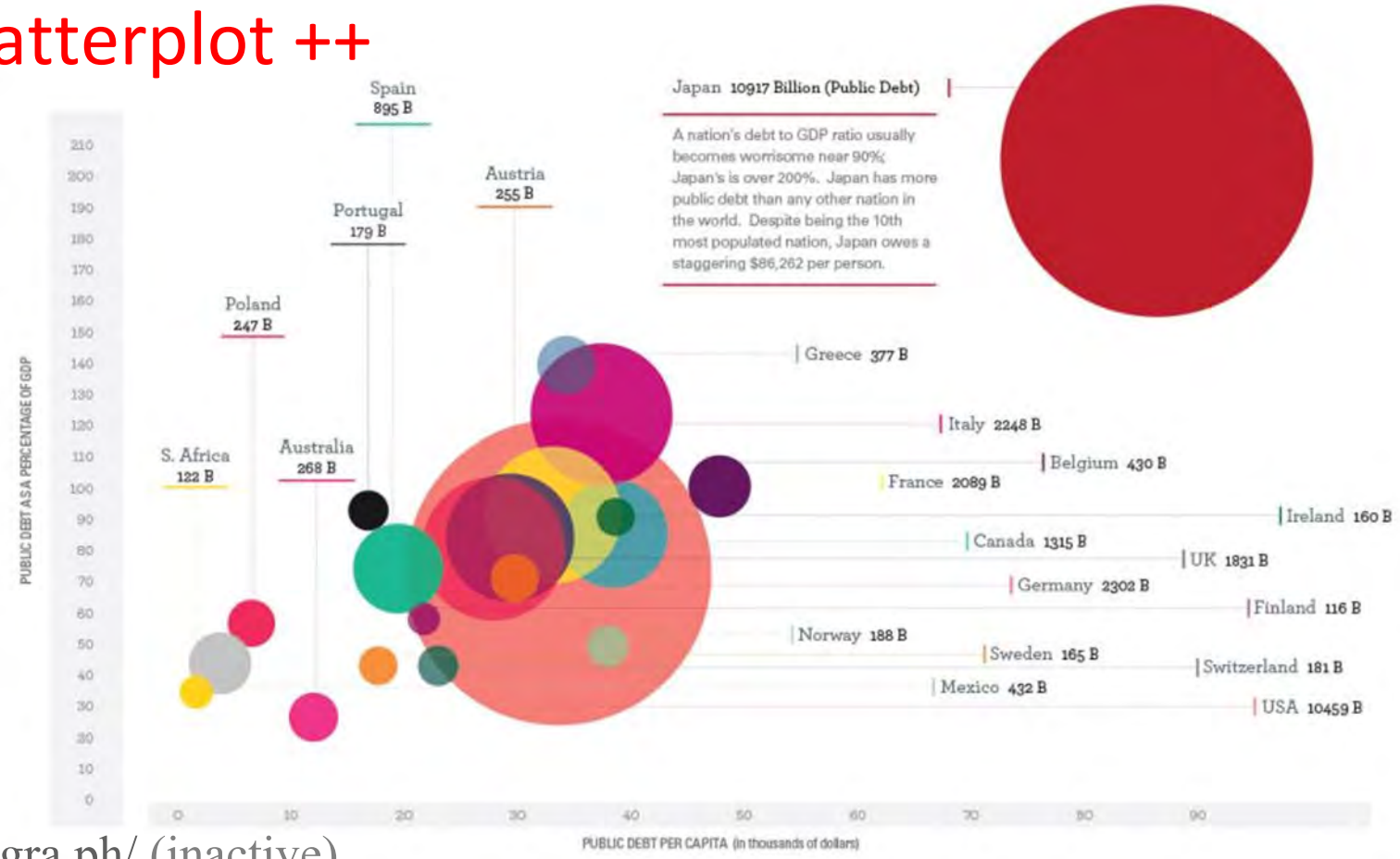
Network



binaryspark.com/

Debt crisis: Japan

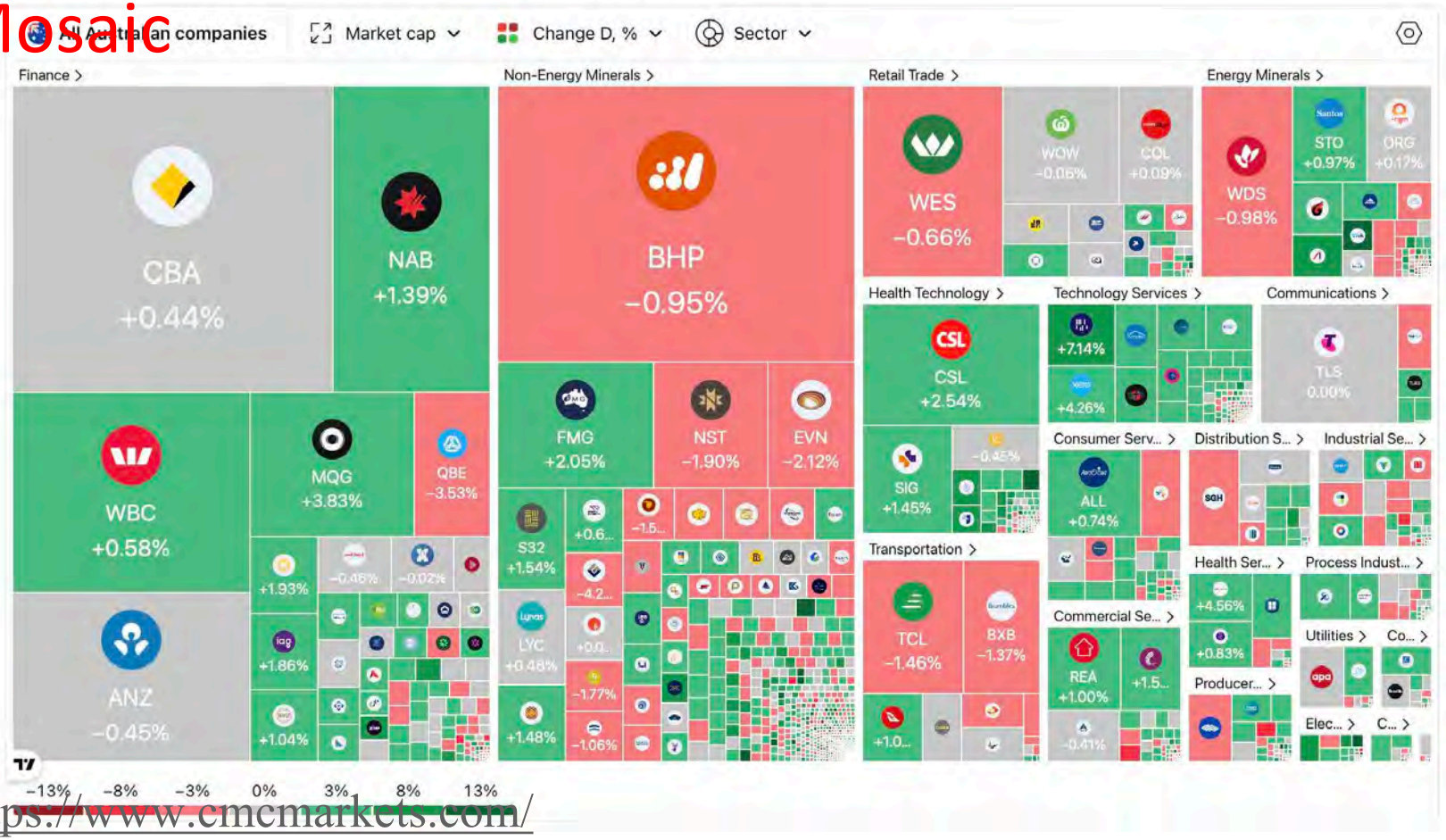
Scatterplot ++



infogra.ph/ (inactive)

ASX Share price information

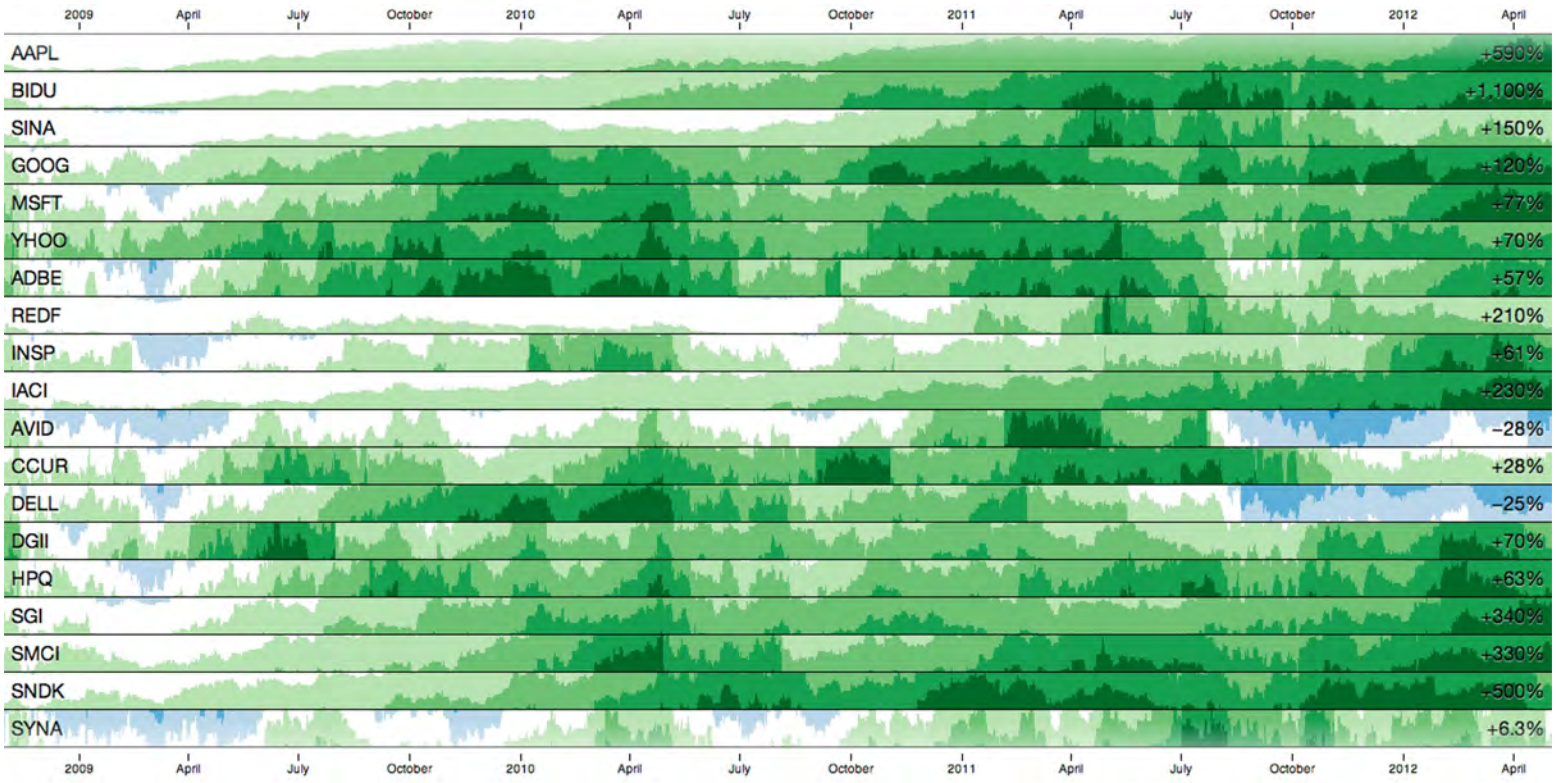
Mosaic



https://www.cmcmarkets.com/

Share price/volume over time

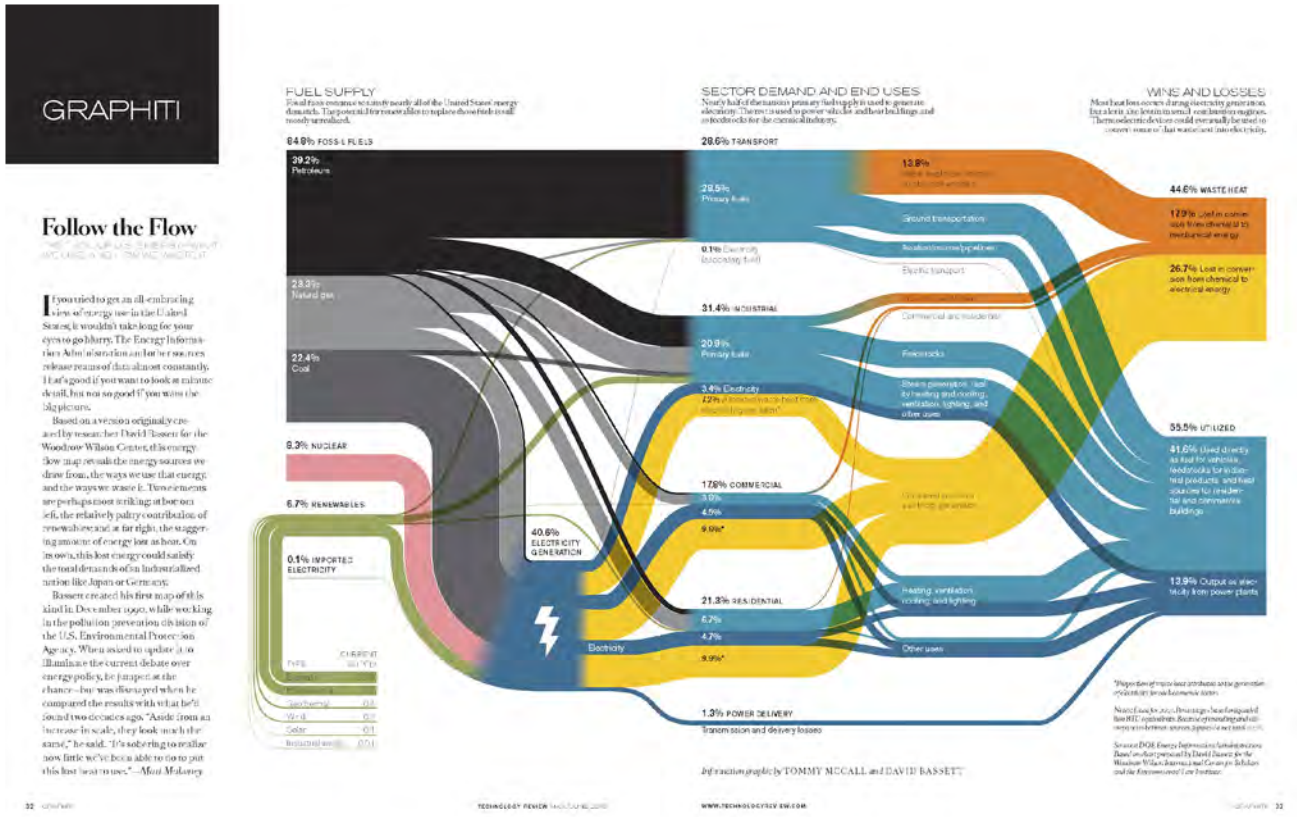
Horizon



bost.ocks.org/

US energy production/consumption

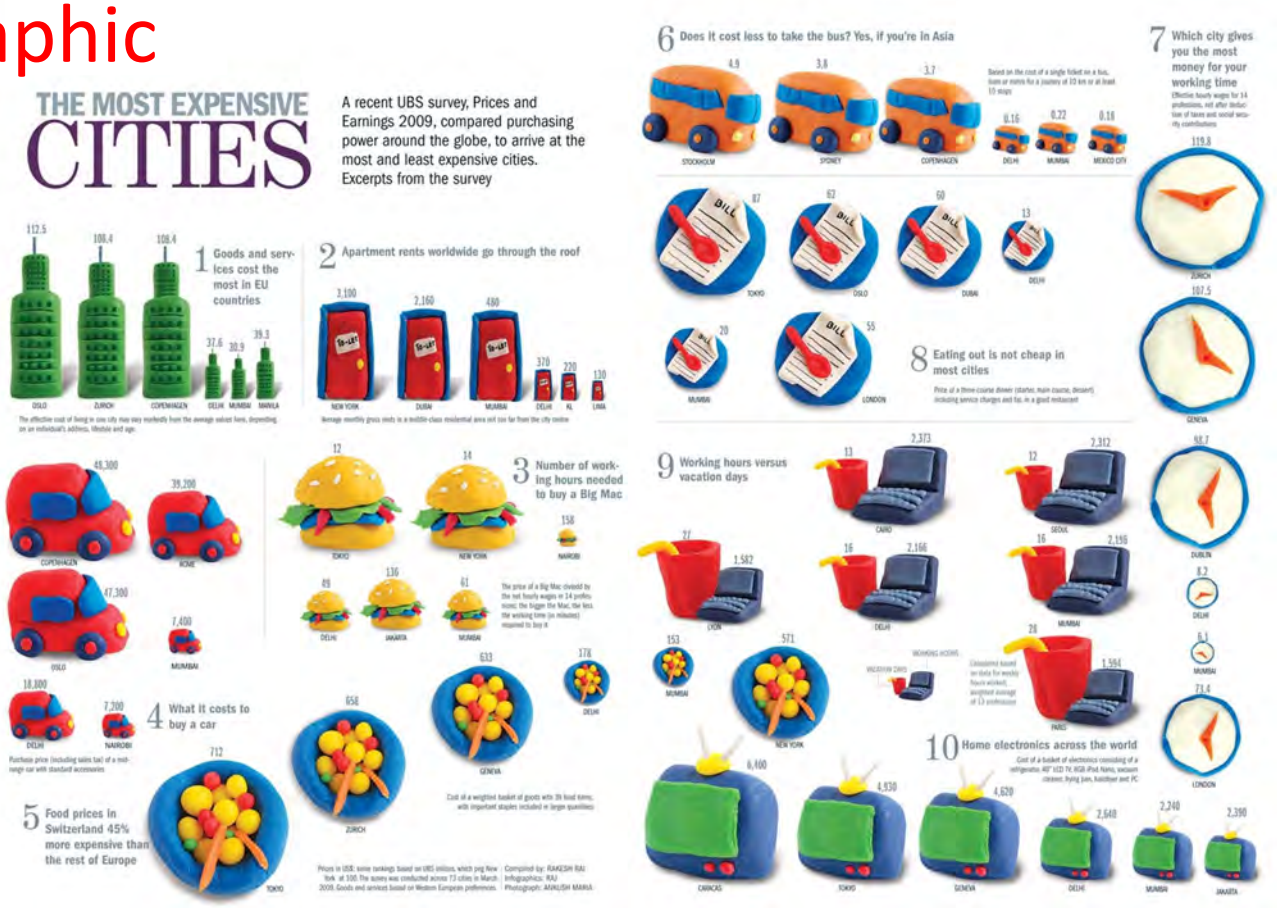
Sankey Diagram



visual.ly/

Most expensive cities

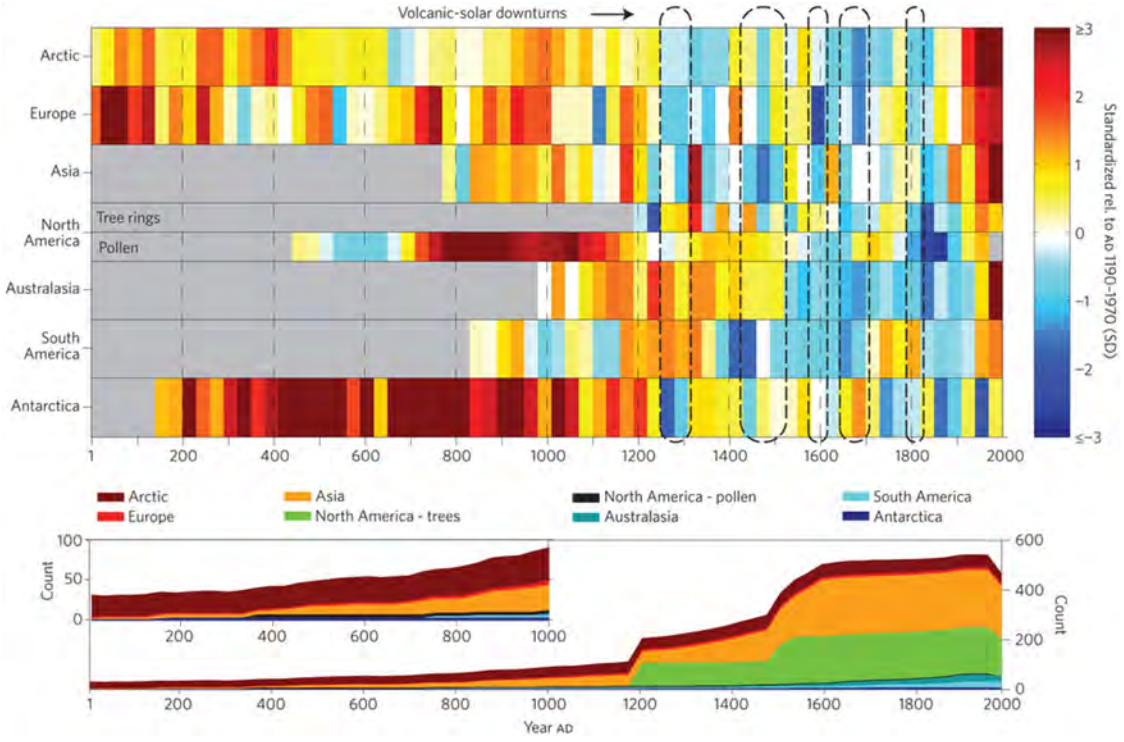
Infographic



infographiclist.com/ (inactive)

Climate change

Continental-scale temperature variability during the past two millennia



nature.com/

Climate change

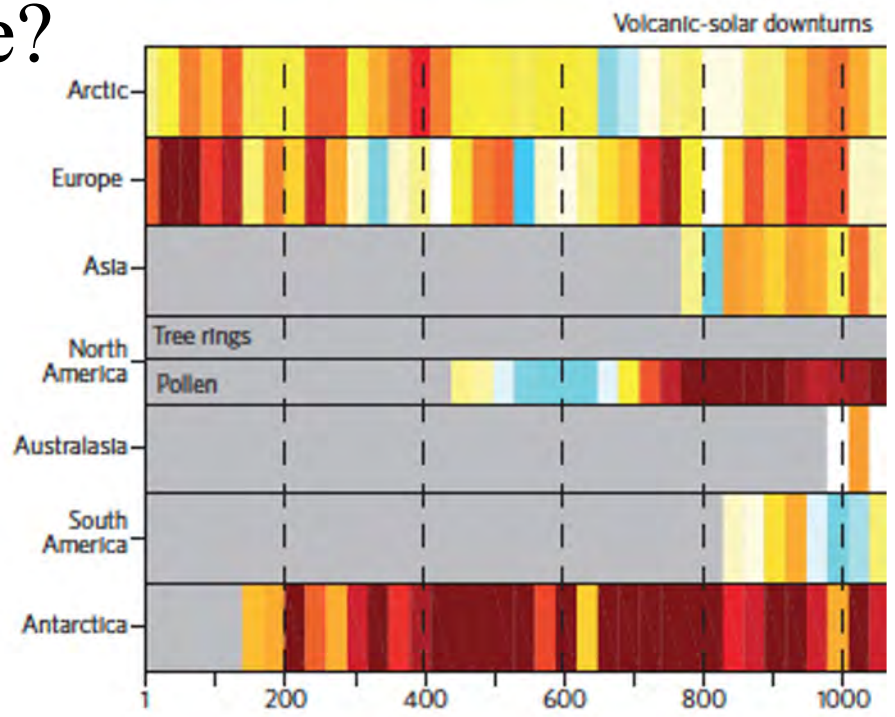
...

The ‘2k Network’ of the IGBP Past Global Changes (PAGES) project aims to produce a global array of regional climate reconstructions for the past 2000 years. ... Nine PAGES 2k working groups represent eight continental-scale regions and the oceans. Regional representation brings critical expert knowledge of individual proxy data sets, which is essential for improving palaeoclimate reconstructions. The PAGES 2k Network is coordinated with the National Oceanic and Atmospheric Administration (NOAA) World Data Center for Paleoclimatology to establish a benchmark database of proxy climate records for the past two millennia ...

Question 1 (for discussion in our seminar)

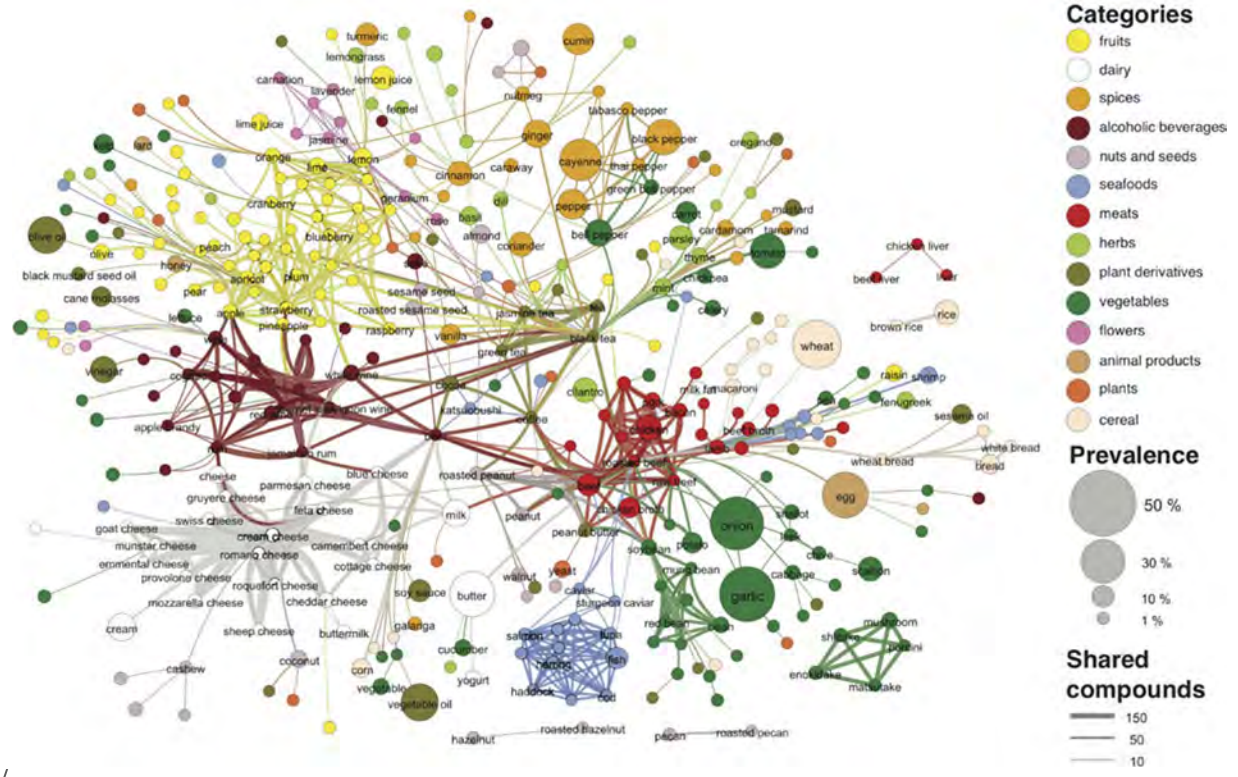
How many dimensions (variables + factors) are presented in the figure?

A. 1
B. 2
C. 3
D. 4
E. More than 4



Food networks

Flavor network and the principles of food pairing

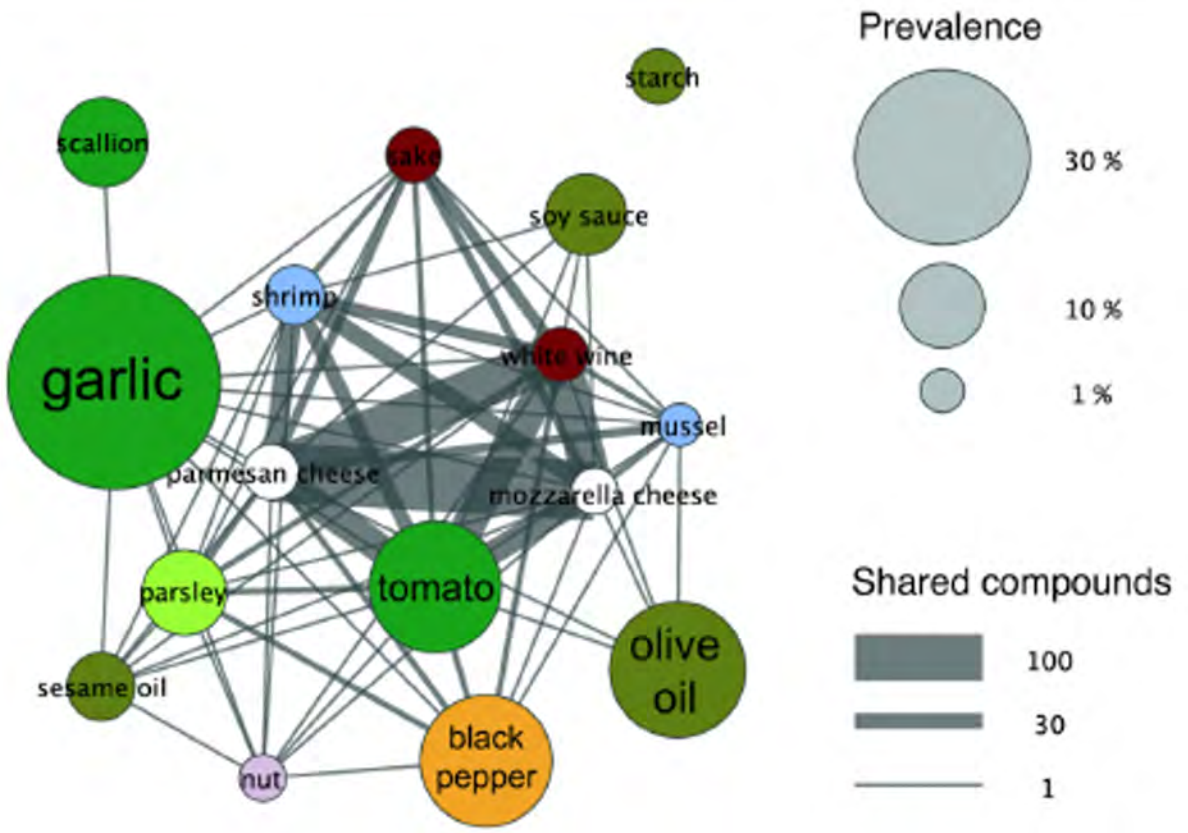


nature.com/

Question 2

How many dimensions are displayed in this figure?

A. 1
B. 2
C. 3
D. 4
E. More than 4



What data is required to store all the information contained in the graph?

Inspiration:

What type of graphic do you want to create?

What data do you have, and what story do you want the graphic to tell?

Some starting points:

The Visualization Zoo…

A tour through the visualization zoo

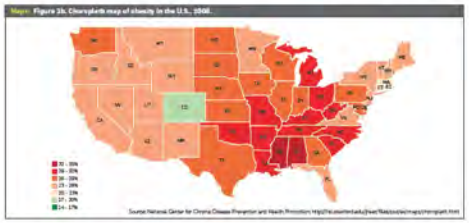
http://dl.acm.org/citation.cfm?id=1743567

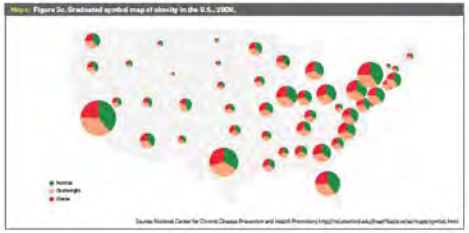
Identifies the major graphic families and their subtypes.

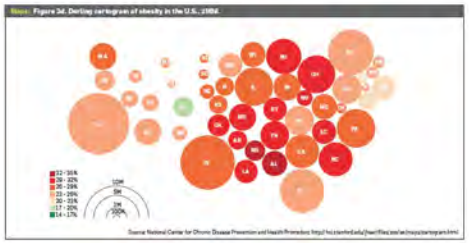
- Maps
- Time Series
- Statistical distributions
- Hierarchies
- Networks

The Visualization Zoo…

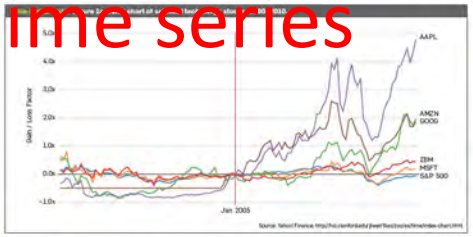
Maps

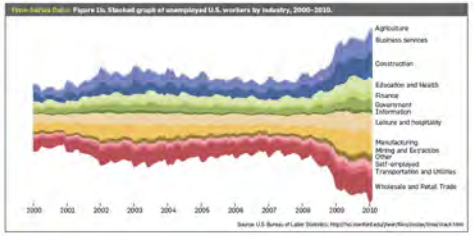


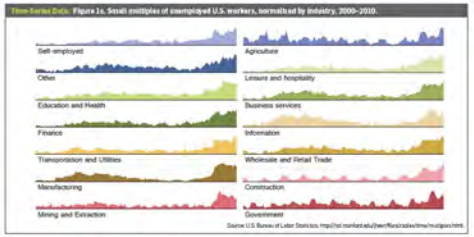


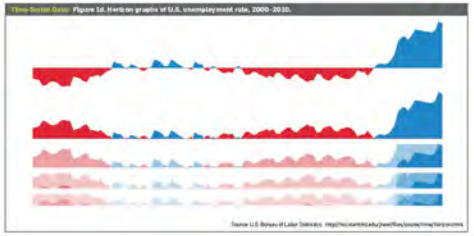


Time series

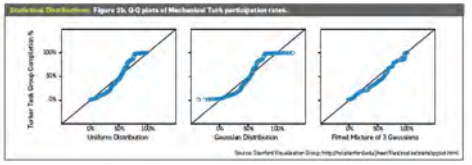


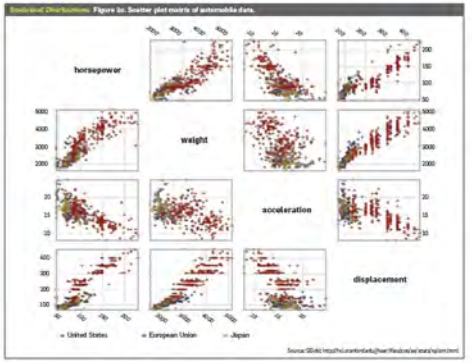


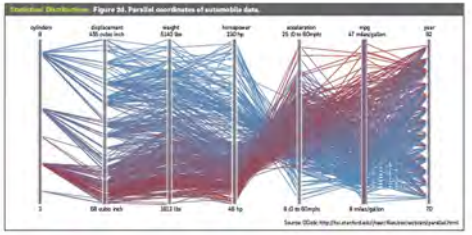




Statistical

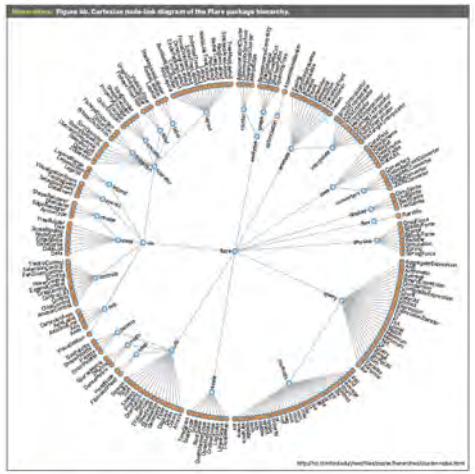


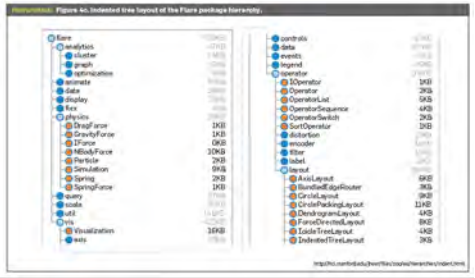


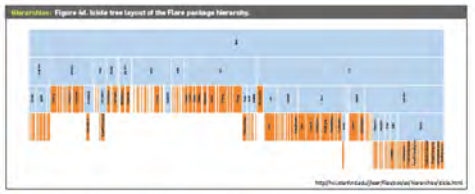


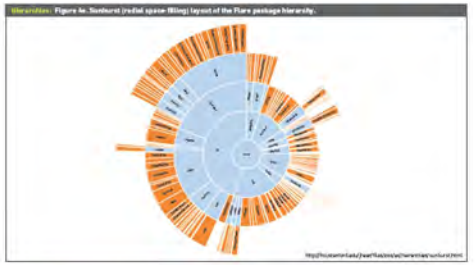
The Visualization Zoo…

Hierarchies

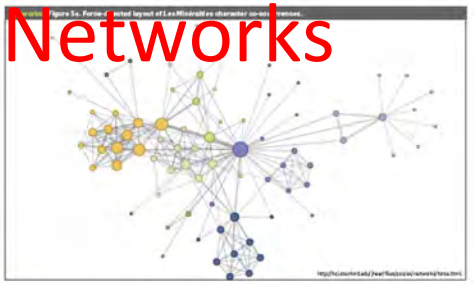


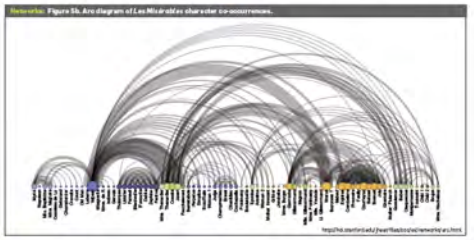


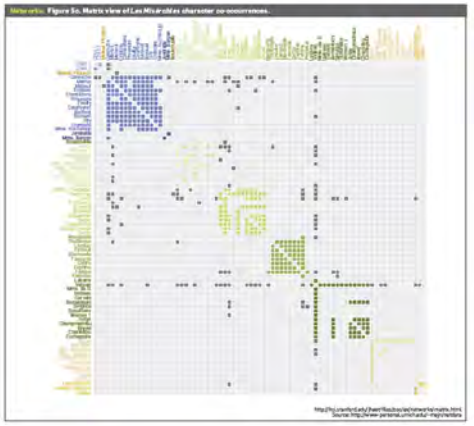




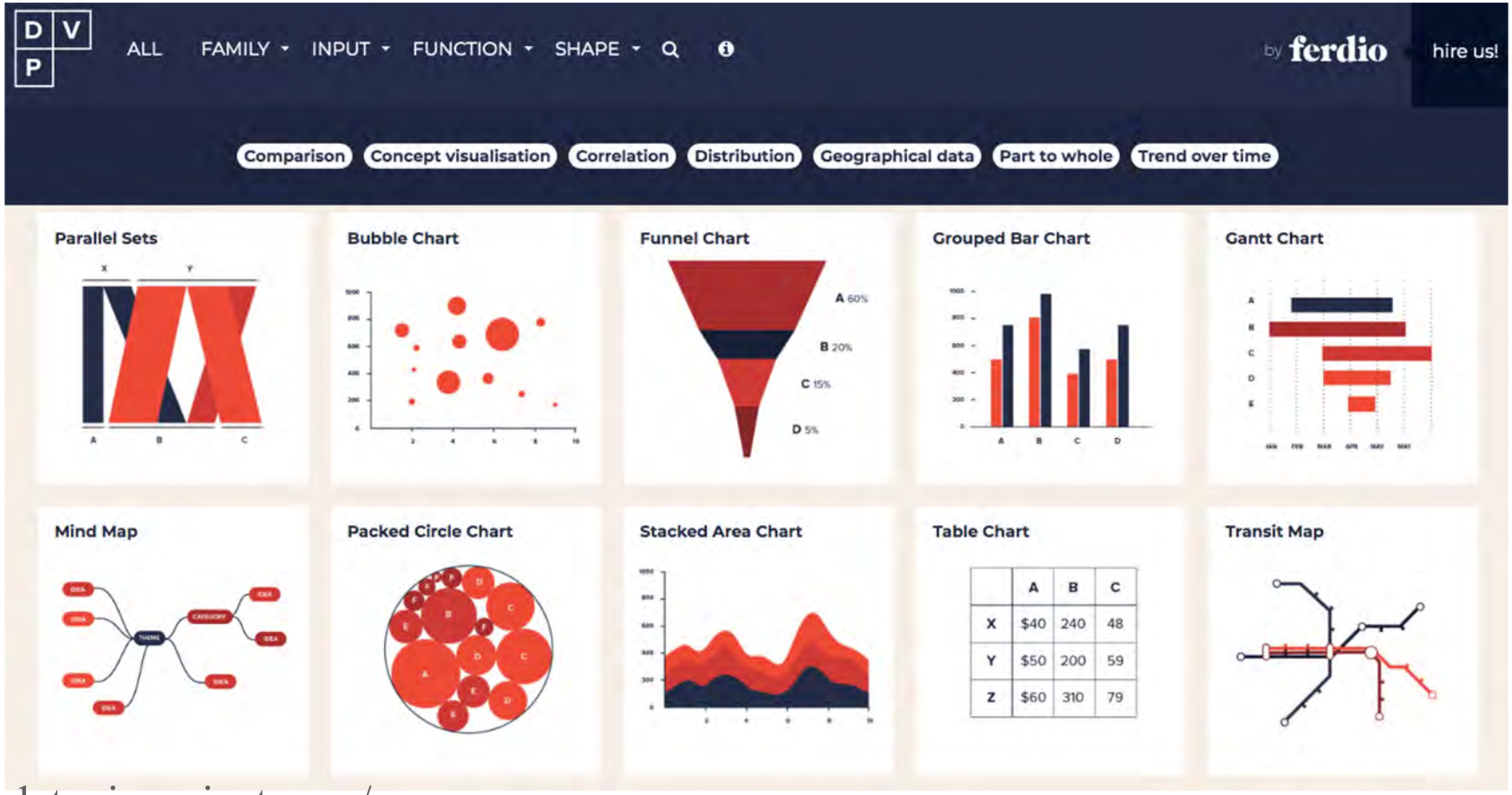
Networks







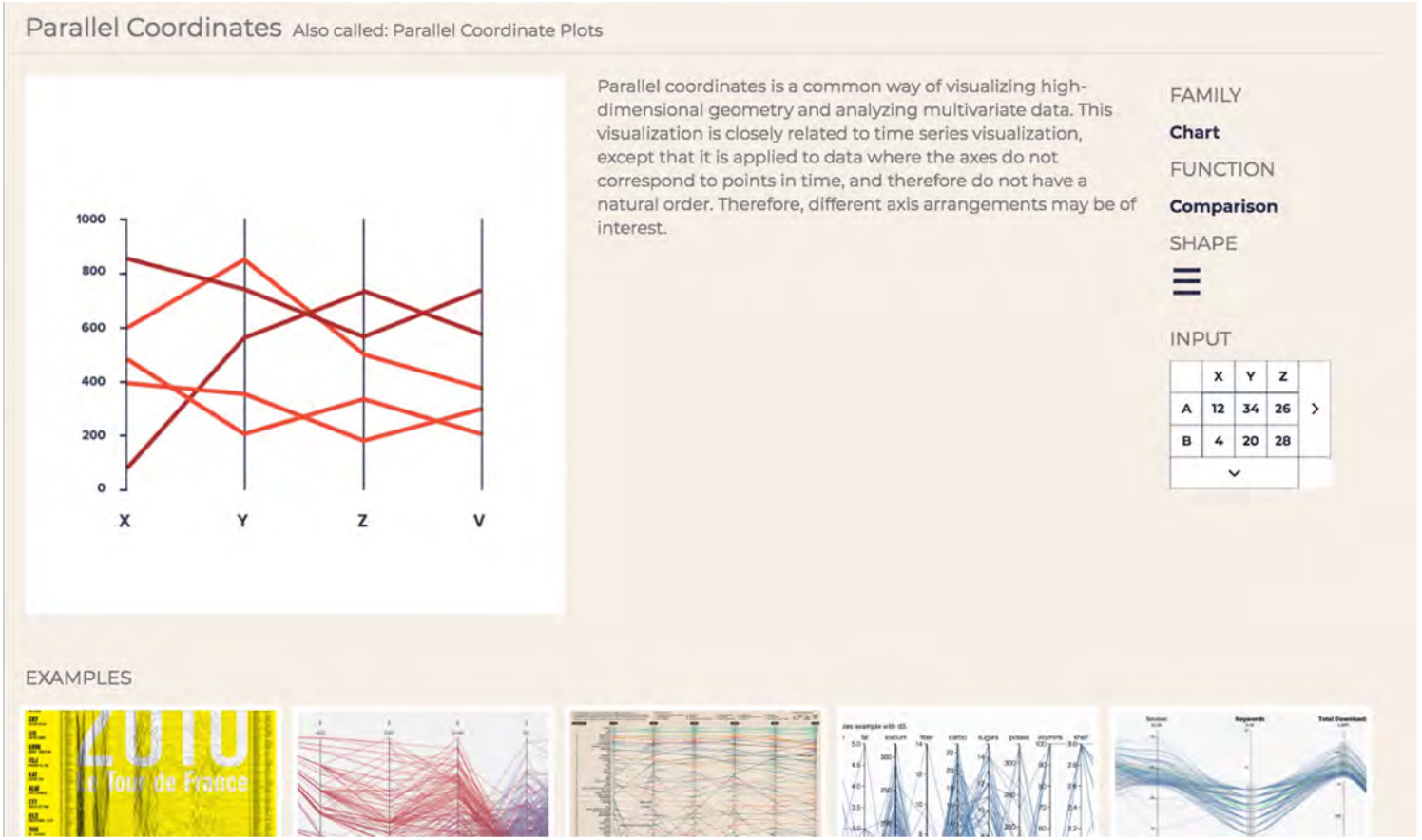
Data Viz Project



	A	B	C
X	$40	240	48
Y	$50	200	59
Z	$60	310	79

datavizproject.com/

Data Viz Project

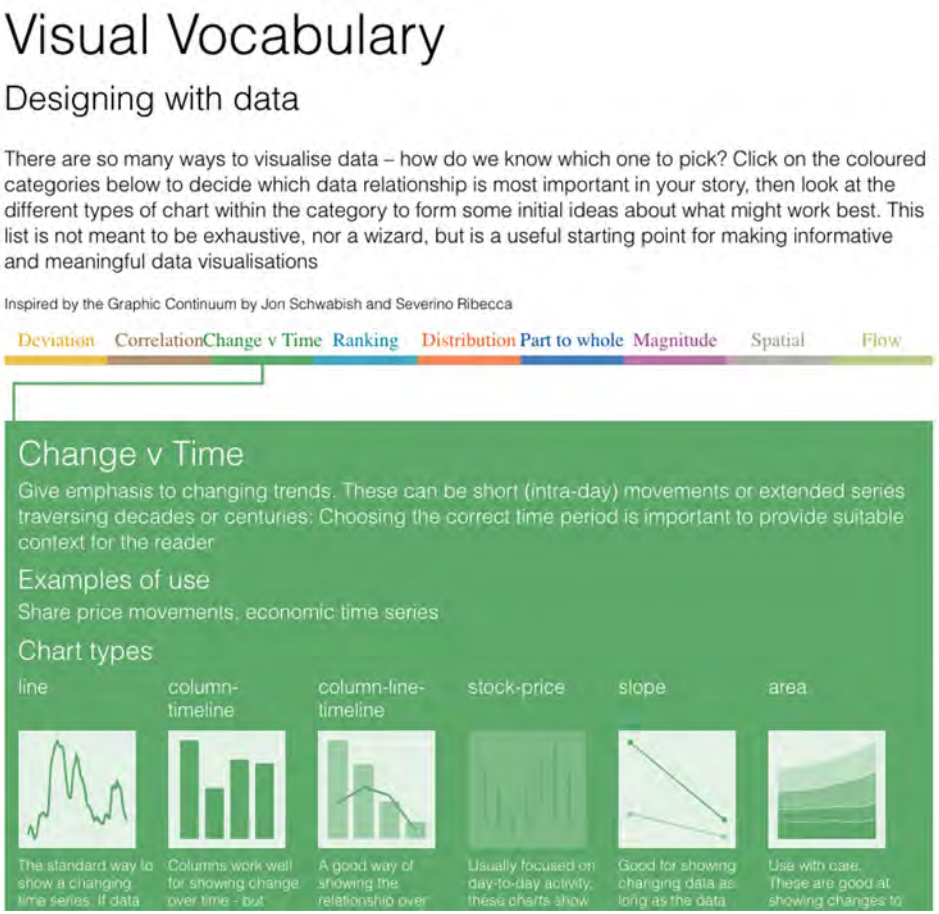


datavizproject.com/

FT: visual vocabulary

github.com/ft-interactive/

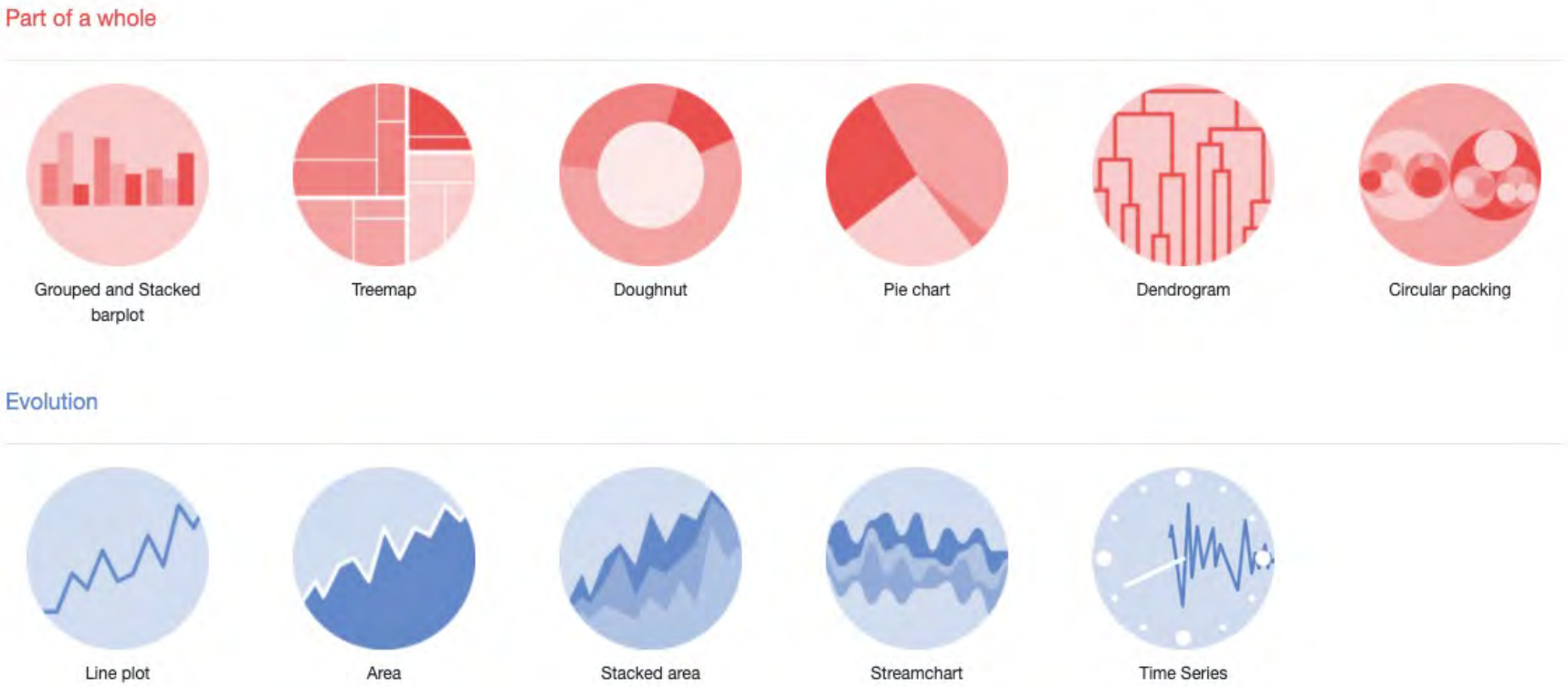
Visual vocabulary interactive



ft-interactive.github.io/

The R Graph Gallery

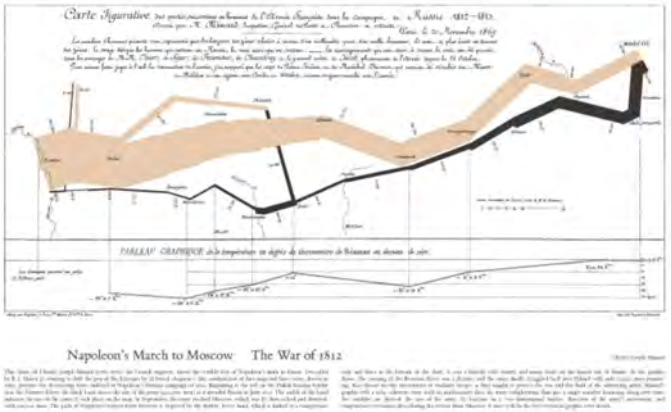
Has lots of graph styles on display with reproduceable code. www.r-graph-gallery.com/

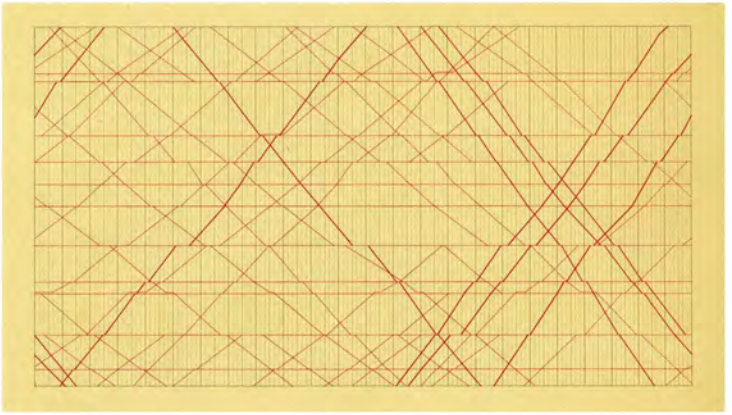


Edward Tufte…

One source of inspiration is Edward Tufte:

- Read: Tufte, E. The visual display of quantitative information, Graphics Press (via Monash Library).
- A strong advocate for good information design.





https://www.edwardtufte.com/tufte/

https://medium.com/

TED talks on data science

Playlist on data and data science: Making sense of too much data

https://www.ted.com/playlists/56/making_sense_of_too_much_data

In particular: Hans Rosling, David McCandless, Deb Roy, Nate Silver, Mona Chalabi, Jennifer Golbeck – but all worth watching...

MONA CHALABI

3 ways to spot a bad statistic

TOMMY MCCALL

The simple genius of a good graphic

HANS ROSLING

The best stats you've ever seen

Getting to know a data set

Outline

In the following slides we'll cover:

- The Iris data
- Examining sections of data frame
- Data types
- Data frame dimension and structure
- Selecting elements, rows or columns
- 5-point summary

Edgar Anderson's Iris data

50 samples from 3 species:

- Iris setosa, – virginica, – versicolor

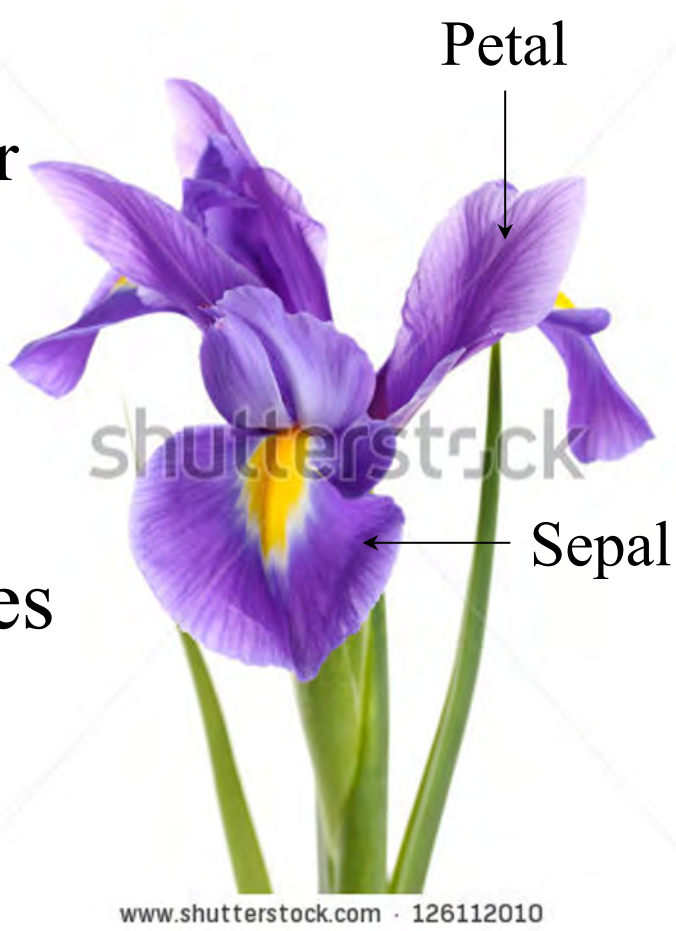
Four features measured:

- Sepal width and length
- Petal width and length

Is it possible to distinguish species using physical measurements?

- Data is packaged with R: "iris"

wikipedia.org/



Print (and data types)

> iris *# = prints out the data set. Ok for small data sets*

	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
...					

Row numbers

Numeric data

Factor

Print head and tail

> head(iris)

```
  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
```

> tail(iris)

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
145          6.7         3.3          5.7         2.5 virginica
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

Dimension, column names, structure

```
> dim(iris)
[1] 150 5
> names(iris)
[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"
"Species"
> str(iris)
'data.frame': 150 obs. of 5 variables:
$ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
$ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
$ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
$ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
$ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1
1 1 1 1 ...
```

Question 3

How many dimensions in the Iris data?

A. 1
B. 2
C. 3
D. 4
E. More than 4

Selection of rows and/or columns

Use this syntax: DataFrame[rows,columns].
Blank means select all rows/columns.

```
> iris[10:15,] # multiple rows
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
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
```

```
> iris[11,] # single row
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
11          5.4         3.7          1.5         0.2  setosa
```

Part of a single column

```
> iris[10:20, "Sepal.Length"] # identify column by name
[1] 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1 5.7 5.1
> # or
> iris[10:20,1] # identify column by number
> [1] 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1 5.7 5.1
> # or
> iris$Sepal.Length[10:20] # identify column first then
  select rows
> [1] 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1 5.7 5.1
```

Summary

Create a mean + 5-point summary of each numerical column, and list of levels and counts for factors.

```
> summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

The real irises

Which species is easiest to differentiate?

A. versicolor

B. virginica

C. setosa

D. Too hard to tell.

Class activity

The data set ‘mpg’ is contained in the ggplot2 package. Let’s get to know it (how many dimensions, types of variables, range etc.) without any graphics.

```
> ?mpg # information about the data
> head(mpg)
> str(mpg)
> summary(mpg)
> tail(mpg)
> unique(mpg$column) #particular columns
```

- See worksheet (MPG Summary) on Moodle

Class activity

```
> str(mpg)
Classes 'tbl_df', 'tbl' and 'data.frame':      234 obs. of  11 variables:
 $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
 $ model       : chr  "a4" "a4" "a4" "a4" ...
 $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
 $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
 $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
 $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
 $ drv         : chr  "f" "f" "f" "f" ...
 $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
 $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
 $ fl          : chr  "p" "p" "p" "p" ...
 $ class       : chr  "compact" "compact" "compact" "compact" ...

> head(mpg)
# A tibble: 6 x 11
  manufacturer model displ  year   cyl      trans   drv   cty   hwy
         <chr> <chr> <dbl> <int> <int>      <chr> <chr> <int> <int>
1         audi    a4   1.8  1999     4   auto(l5)     f    18    29
2         audi    a4   1.8  1999     4 manual(m5)     f    21    29
3         audi    a4   2.0  2008     4 manual(m6)     f    20    31
4         audi    a4   2.0  2008     4   auto(av)     f    21    30
5         audi    a4   2.8  1999     6   auto(l5)     f    16    26
6         audi    a4   2.8  1999     6 manual(m5)     f    18    26
# ... with 2 more variables: fl <chr>, class <chr>
```

Class activity

```
> summary(mpg)
 manufacturer          model               displ            year
 Length:234         Length:234         Min.   :1.600   Min.   :1999
 Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999
 Mode  :character   Mode  :character   Median :3.300   Median :2004
                                       Mean   :3.472   Mean   :2004
                                       3rd Qu.:4.600   3rd Qu.:2008
                                       Max.   :7.000   Max.   :2008
      cyl            trans               drv
 Min.   :4.000   Length:234         Length:234
 1st Qu.:4.000   Class :character   Class :character
 Median :6.000   Mode  :character   Mode  :character
 Mean   :5.889
 3rd Qu.:8.000
 Max.   :8.000
      cty             hwy              fl
 Min.   : 9.00   Min.   :12.00   Length:234
 1st Qu.:14.00   1st Qu.:18.00   Class :character
 Median :17.00   Median :24.00   Mode  :character
 Mean   :16.86   Mean   :23.44
 3rd Qu.:19.00   3rd Qu.:27.00
 Max.   :35.00   Max.   :44.00
    class
 Length:234
 Class :character
 Mode  :character
```

Class activity

```
> tail(mpg)
# A tibble: 6 x 11
  manufacturer  model displ  year   cyl      trans   drv   cty   hwy
         <chr>  <chr> <dbl> <int> <int>      <chr> <chr> <int> <int>
1   volkswagen passat   1.8  1999     4   auto(l5)     f    18    29
2   volkswagen passat   2.0  2008     4   auto(s6)     f    19    28
3   volkswagen passat   2.0  2008     4 manual(m6)     f    21    29
4   volkswagen passat   2.8  1999     6   auto(l5)     f    16    26
5   volkswagen passat   2.8  1999     6 manual(m5)     f    18    26
6   volkswagen passat   3.6  2008     6   auto(s6)     f    17    26
# ... with 2 more variables: fl <chr>, class <chr>

> unique(mpg$manufacturer)
 [1] "audi"       "chevrolet"  "dodge"      "ford"       "honda"
 [6] "hyundai"    "jeep"       "land rover" "lincoln"    "mercury"
[11] "nissan"     "pontiac"    "subaru"     "toyota"     "volkswagen"
```

Why do we need to view summaries? Can't we just plot graphs straight away?

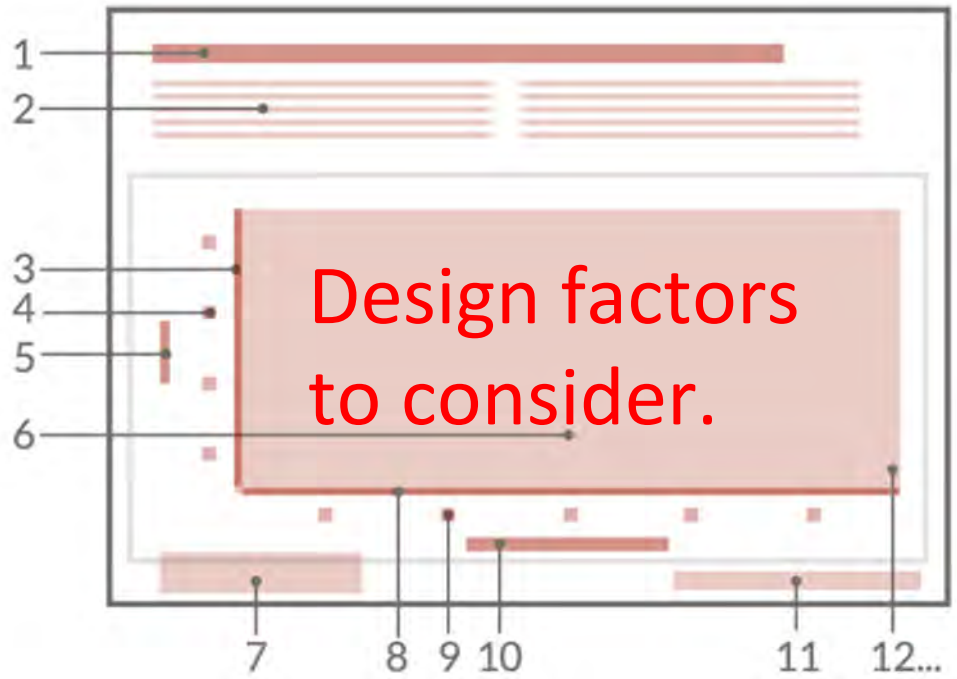
Graphing your data in R

Outline

In the following slides we'll cover:

- Graphing your data in R (base graphics)
- Visualising more variables (base + lattice)
- Presentation quality graphics (ggplot2)
- Displaying data compactly
- Elements of good visual display

Elements of a figure



Typical elements: title (1), subtitle (2), y-axis (3), label (4), name (5), data area (6), legend (7), X-axis (8), label (9), and name (10), sources (11). Further elements: annotations/lines/symbols (12).

Thomas Rahlf: Data Visualisation with R

Base graphics

These are the graphic functions built into the basic R installation.

- High level graphic functions create new graphs with axis, labels and titles.
- Low level graphic functions then annotate plots with points, lines and text.

Useful references:

- A Tiny Handbook of R, Chapter 3.
- Also, Exploratory Analysis with R, Chapter 9:

  https://bookdown.org/rdpeng/exdata/the-base-plotting-system-1.html

Base graphics: high level functions

Common plots for the main area of the graph:

> plot *# Scatterplot*

> pairs *# Scatterplot matrix*

> hist *# Histogram*

> stem *# Stem-and-leaf plot*

> boxplot *# Box-and-whisker plot*

> barplot *# Bar plot*

> dotchart *# Dot plot*

- See ATHR page 49, also Peng, R., Exploratory Data Analysis with R.

Base graphics: low level functions

Some low-level plotting functions include:

> lines *# Draw lines between given coordinates*
> text *# Draw text at given coordinates*
> abline *# Line y = ax + b, horizontal or vertical*
> axis *# Add an axis*
> arrows *# Draw arrows*
> grid *# Add a rectangular grid*
> legend *# Add a legend (a key)*

- See ATHR page 50, also Peng, R.

Base graphics: graphics parameters

Some low-level additions/controls include:

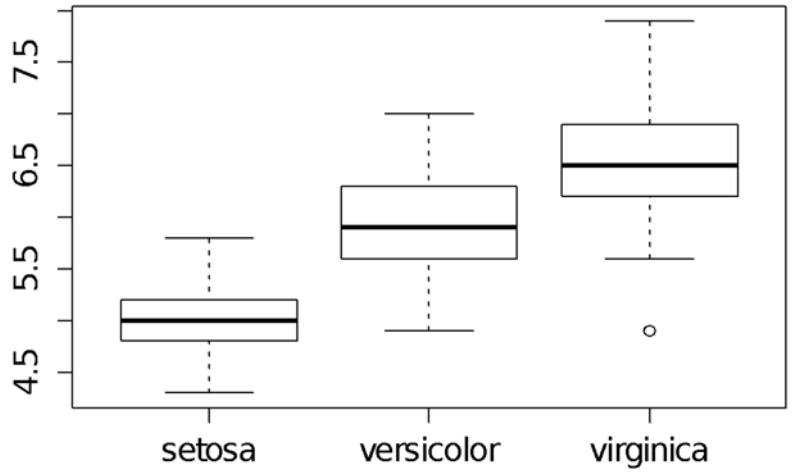
> main *# Title of the plot*
>
> ylab, xlab *# Labels for the y-axis and x-axis*
>
> type *# Plot type (points, lines, both, ...),*
>
> pch *# Plot character (circles, dots, , symbols, ...)*
>
> lty *# Line type (solid, dots, dashes, ...)*
>
> lwd *# Line width*
>
> col *# Colour of plot characters… and many others*

- See ATHR page 50, also Peng, R.

Boxplot

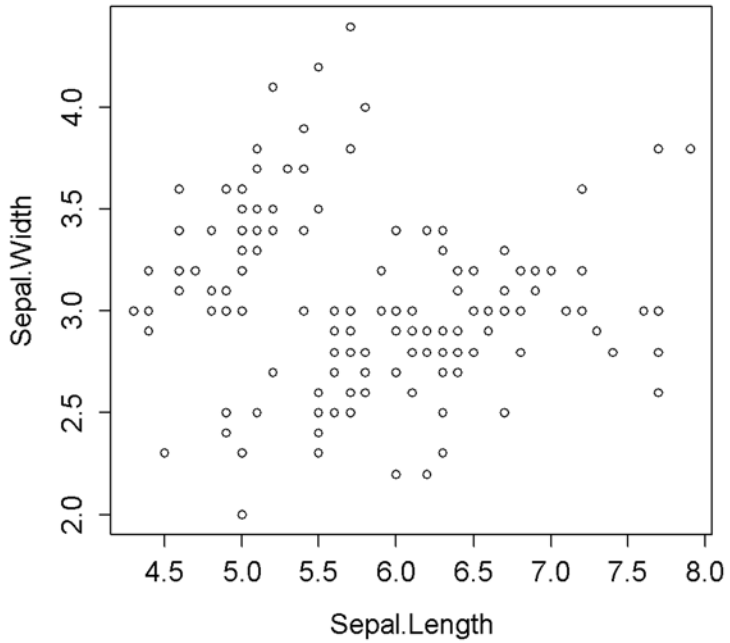
Each variable can be viewed as a boxplot distinguished by level:

```
> boxplot(Sepal.Length ~ Species, data = iris)
> # note ~ indicates grouping variable
```



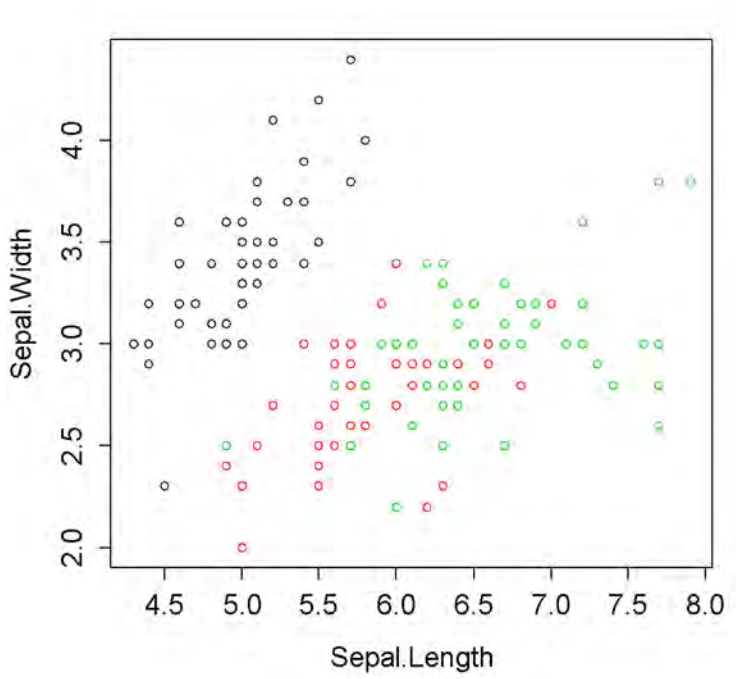
Scatterplot

```
> with(iris, plot(Sepal.Length, Sepal.Width))
> # using 'with' simplifies column names etc.
> # another alternative is to "attach" the data frame
```



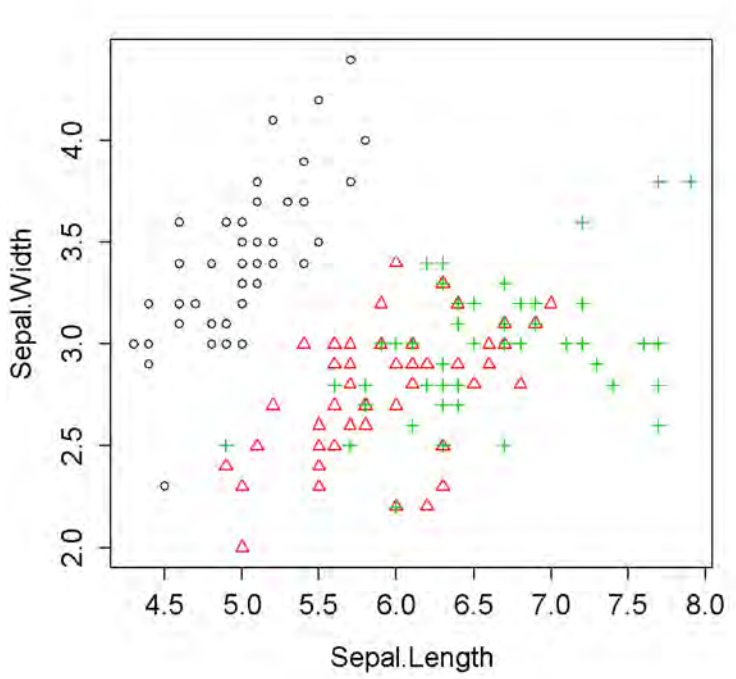
Scatterplot + colour

> with(iris, plot(Sepal.Length, Sepal.Width, col = Species))



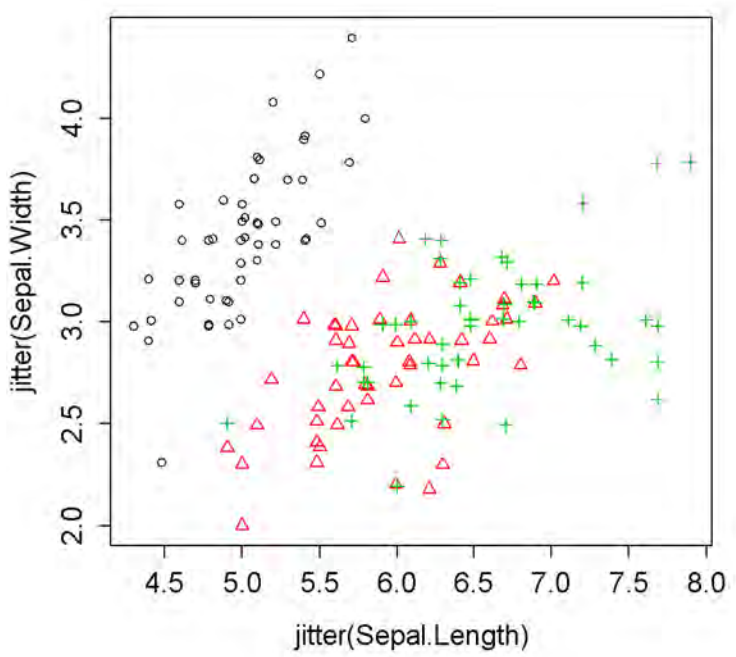
Scatterplot + plot symbol

> with(iris, plot(Sepal.Length, Sepal.Width, col = Species, pch=as.numeric(Species)))



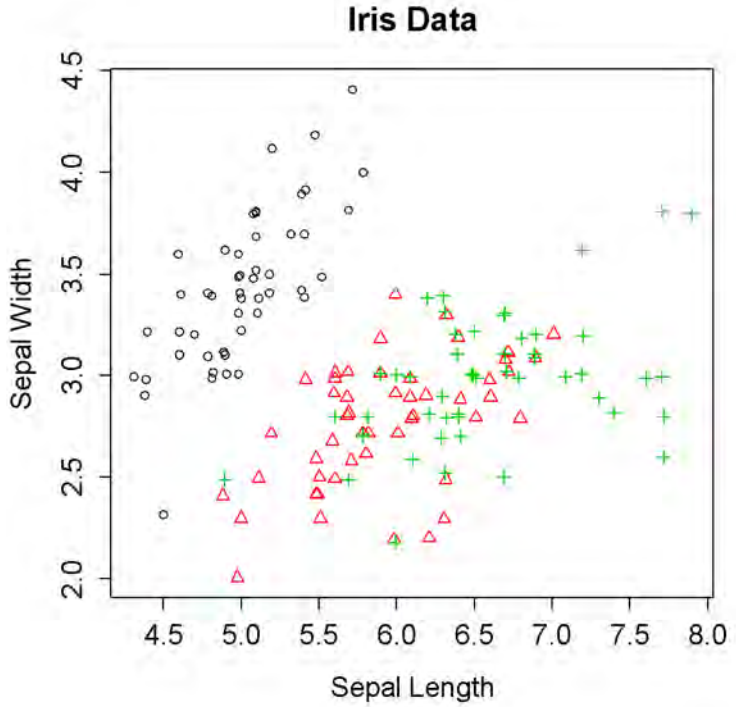
Scatterplot + jitter

```
> with(iris, plot(jitter(Sepal.Length), jitter(Sepal.Width),
  col = Species, pch=as.numeric(Species)))
> # jittering reveals some of the overlapping data points
```



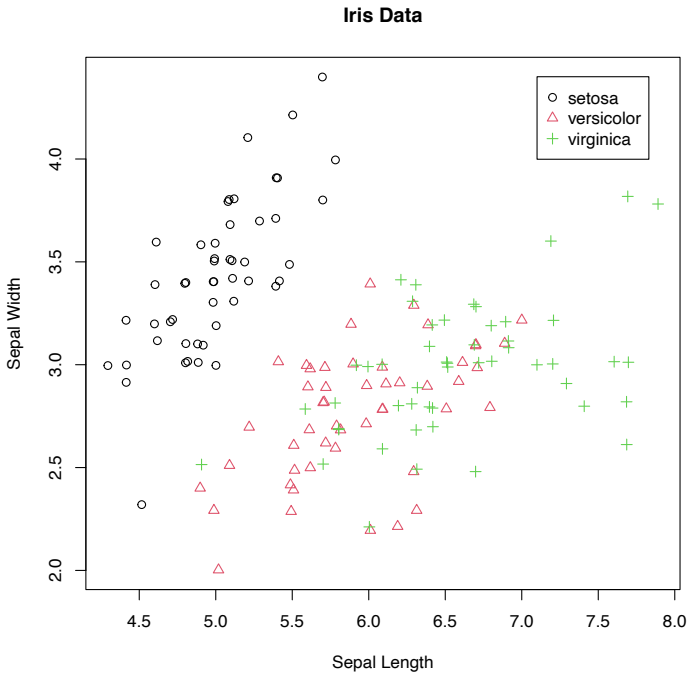
Scatterplot + labels

```
> with(iris, plot(jitter(Sepal.Length), jitter(Sepal.Width),
  col = Species, pch=as.numeric(Species), main = ("Iris
  Data"), xlab = "Sepal Length", ylab = ("Sepal Width")))
```



Scatterplot + legend

```
> # Follow the plot command with:
> with(iris, legend(7.1, 4.4, as.vector(unique(Species)),
  pch=unique(Species), col = unique(Species)))
```



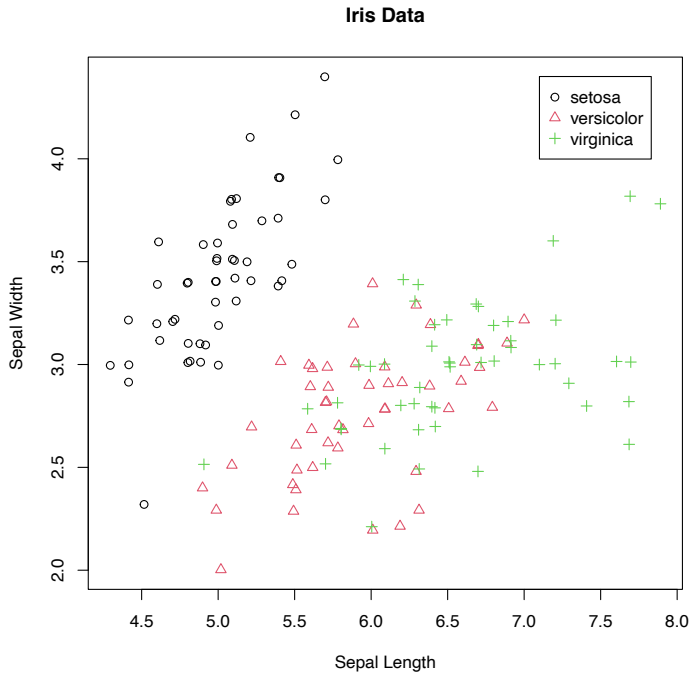
Complete plot command

```
> with(iris, plot(jitter(Sepal.Length), jitter(Sepal.Width),
  col = Species, pch=as.numeric(Species), main = ("Iris
  Data"), xlab = "Sepal Length", ylab = ("Sepal Width")))
> with(iris, legend(7.1, 4.4, as.vector(unique(Species)),
  pch=unique(Species), col = unique(Species)))
```

Question 4

Which species is easiest to differentiate based on sepal size and shape?

A. versicolor
B. virginica
C. setosa
D. Too hard to tell.



Saving graphics

Diverting graphics from RStudio to a file:

- The code below opens a file, diverts the output from RStudio to a named file (of type jpg in this case) and saves it in the working directory.

```
> jpeg("filename.jpg")
> plot(x,y) # put your plotting commands here
> dev.off()
```

- Another method is to use "Export" command under the plot tile in the "help/display" window in Rstudio.

The lattice package

Reading notes only = *

The lattice package has multi-panel graphing functions conditioned on variables, including:

> xyplot *# Multi-panel conditioning scatterplot*
> barchart *# Bar plot*
> dotplot *# Dot plot*
> splom *# Scatterplot matrix*
> bwplot *# Box-and-whisker plot*
> histogram *# Histogram*
> densityplot *# Smoothed histogram*

- See ATHR page 54

lattice

The lattice package comes with the base installation of R.

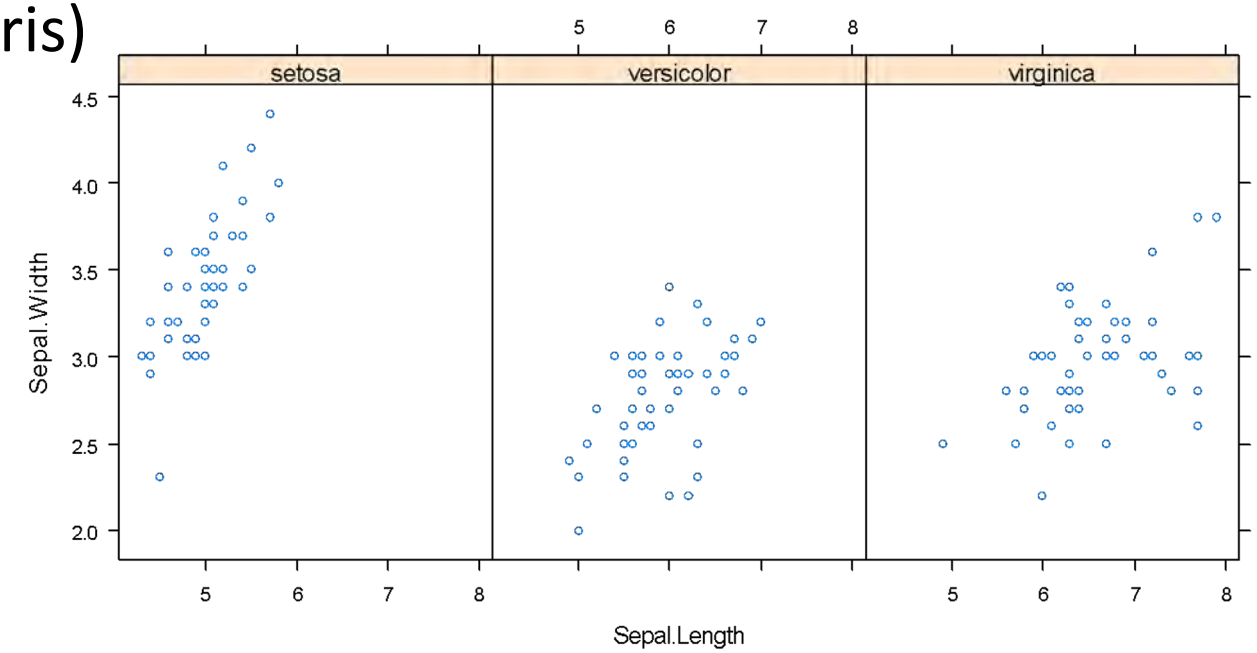
To run add it to the library of packages in the current environment:

```
> library(lattice)
```

xyplot

Conditioning on species:

- Syntax: xyplot(y ~ x | g) : *plot y on x grouped by g*
  - > xyplot(Sepal.Width ~ Sepal.Length | Species, data = iris)

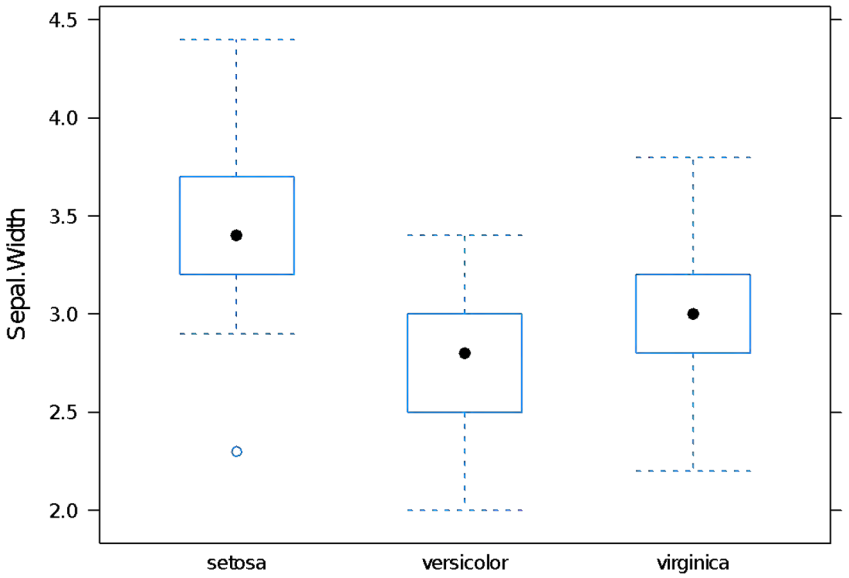


bwplot

Conditioning on species:

- Syntax: bwplot(y ~ g) : *plot y grouped by g*

  > bwplot(Sepal.Width ~ Species, data = iris)



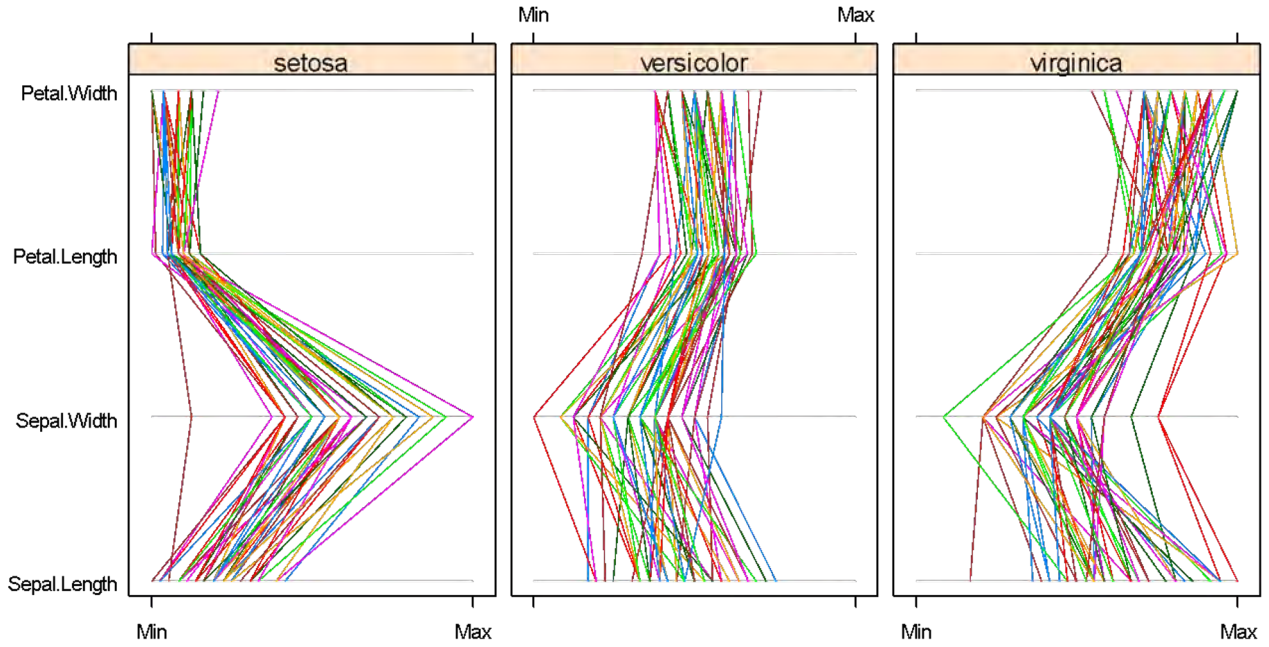
Parallel coordinates

The main reason we still use lattice

Each data point plotted across 4 numeric variables

- Syntax: parallelplot(~y|g): *plot columns y grouped by g*

  > parallelplot(~iris[1:4] | Species, data = iris)



Presentation quality graphs

ggplot2

- The most commonly used packages for display quality graphics.
- Written by Hadley Wickham and Winston Chang, it is an implementation of *The Grammar of Graphics* by Leland Wilkinson and views a graphic as being made up of data points + scales + annotations + statistical summaries… in a structured way, a grammar. See:

http://vita.had.co.nz/papers/layered-grammar.pdf

ggplot2: graphic objects

Some main classes of graphic objects:

- Geoms (geometric objects: think of as type of plot)
- Statistics (summaries, data transformations)
- Scales/coordinate systems
- Faceting (conditional grouping of subsets of data)
- Position adjustments (jitter etc.)
- Annotation
- Aesthetics (colours, line styles etc.)

ggplot2

To install package and add to library:

```
> install.packages("ggplot2")
> library(ggplot2)
```

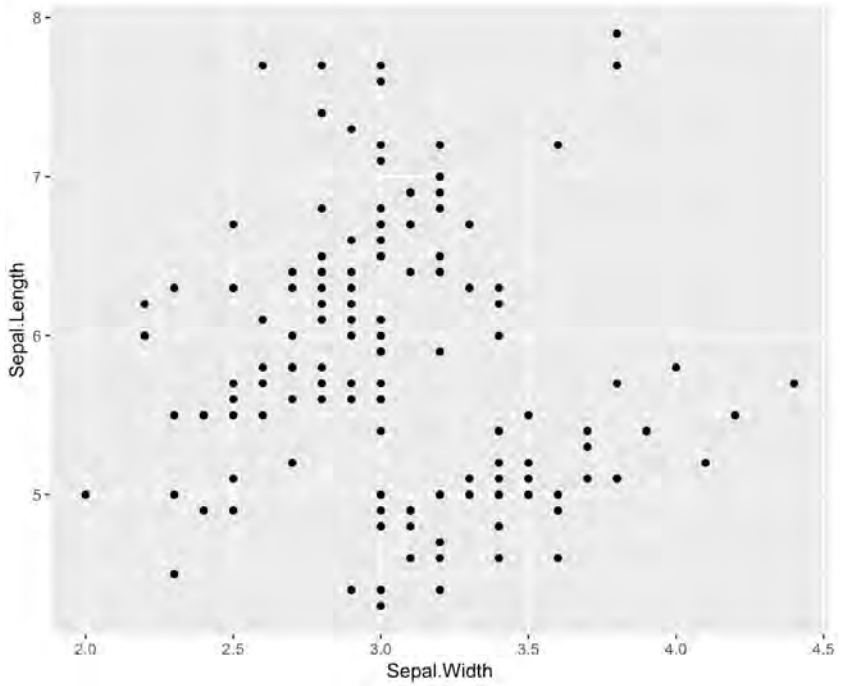
?ggplot (from R help)

- `ggplot() is used to construct the initial plot object, and is almost always followed by a plus sign (+) to add components to the plot.`
- `There are three common patterns used to invoke ggplot():`
- `ggplot(data = df, mapping = aes(x, y, other aesthetics))`
- `ggplot(data = df)`
- `ggplot()`

ggplot – basic plot

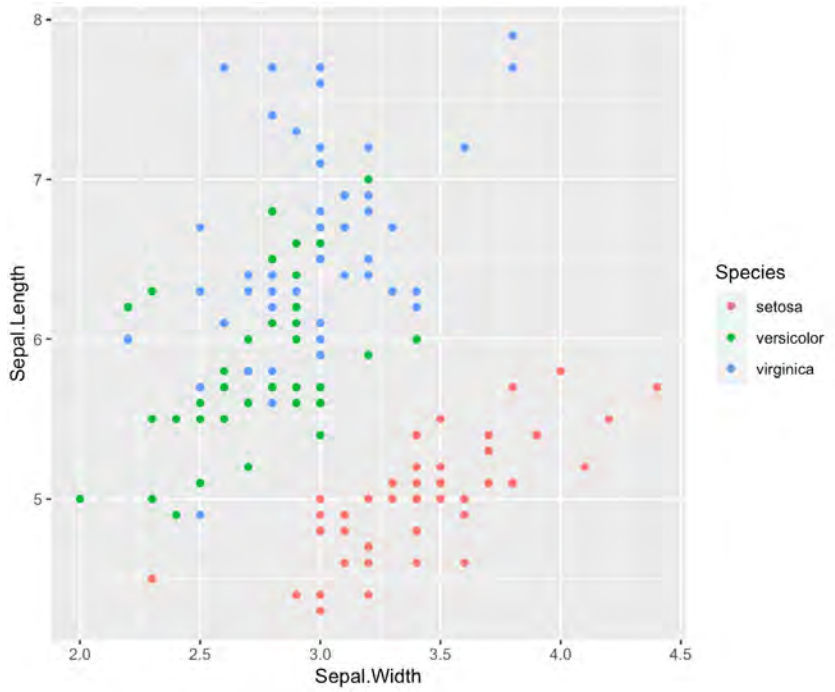
ggplot is the basic plotting function

> ggplot(data = iris, aes(x = Sepal.Width, y = Sepal.Length)) + geom_point())



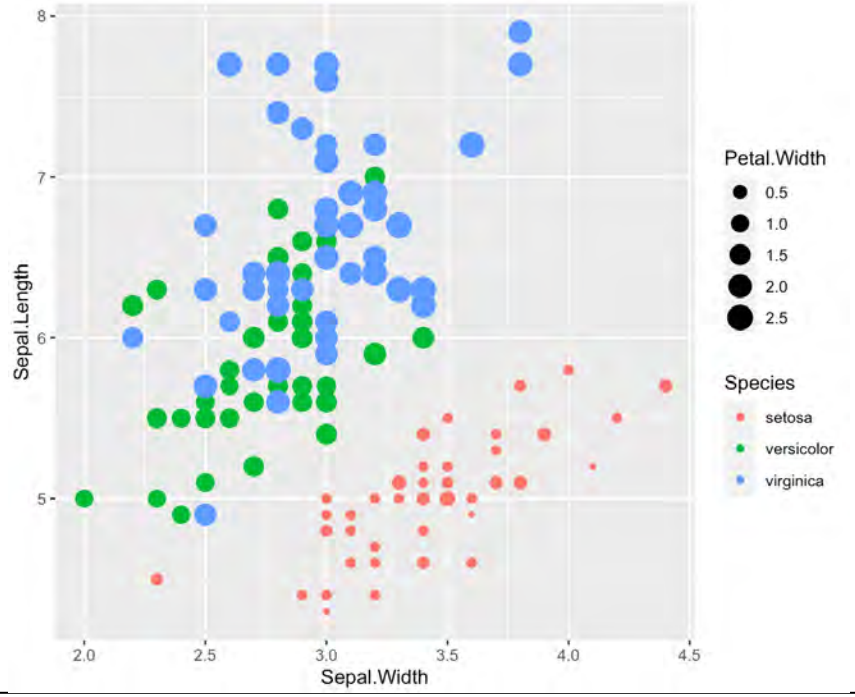
ggplot + colour

> ggplot(data = iris, aes(x = Sepal.Width, y = Sepal.Length, colour = Species)) + geom_point()



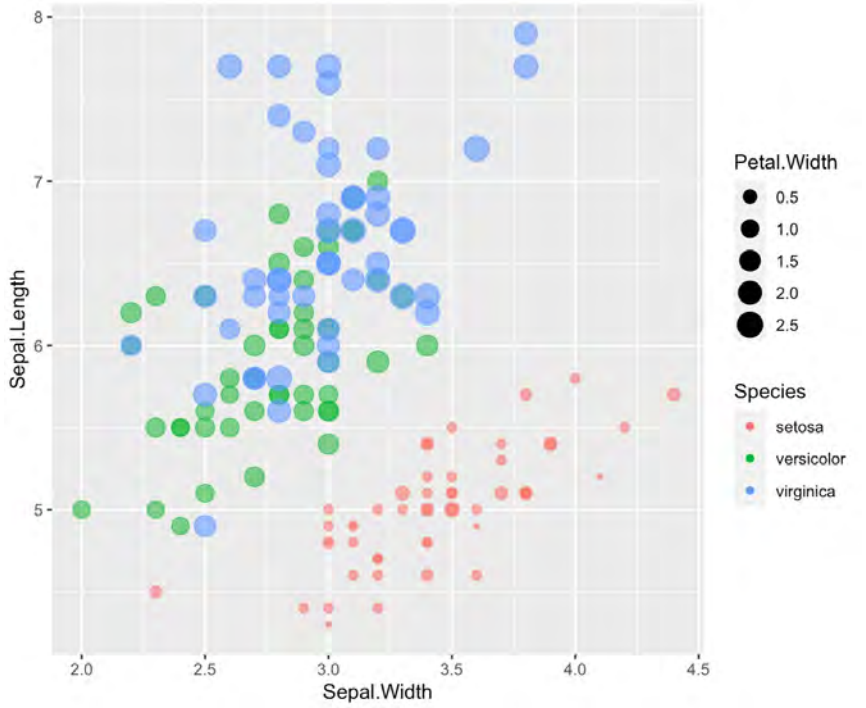
ggplot + colour + size

> ggplot(data = iris, aes(x = Sepal.Width, y = Sepal.Length, colour = Species, size = Petal.Width)) + geom_point()



ggplot + colour + size + alpha

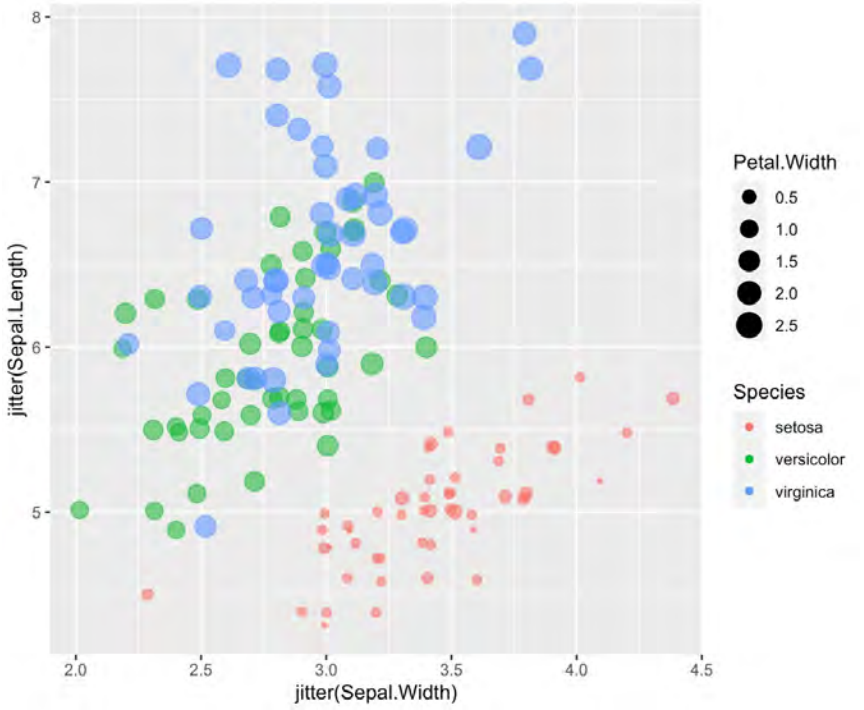
> ggplot(data = iris, aes(x = Sepal.Width, y = Sepal.Length, colour = Species, size = Petal.Width, alpha = I(0.6))) + geom_point()



ggplot + colour + size + alpha + jitter

> ggplot(data = iris, aes(x = jitter(Sepal.Width), y = jitter(Sepal.Length), colour = Species, size = Petal.Width, alpha = I(0.6))) + geom_point()

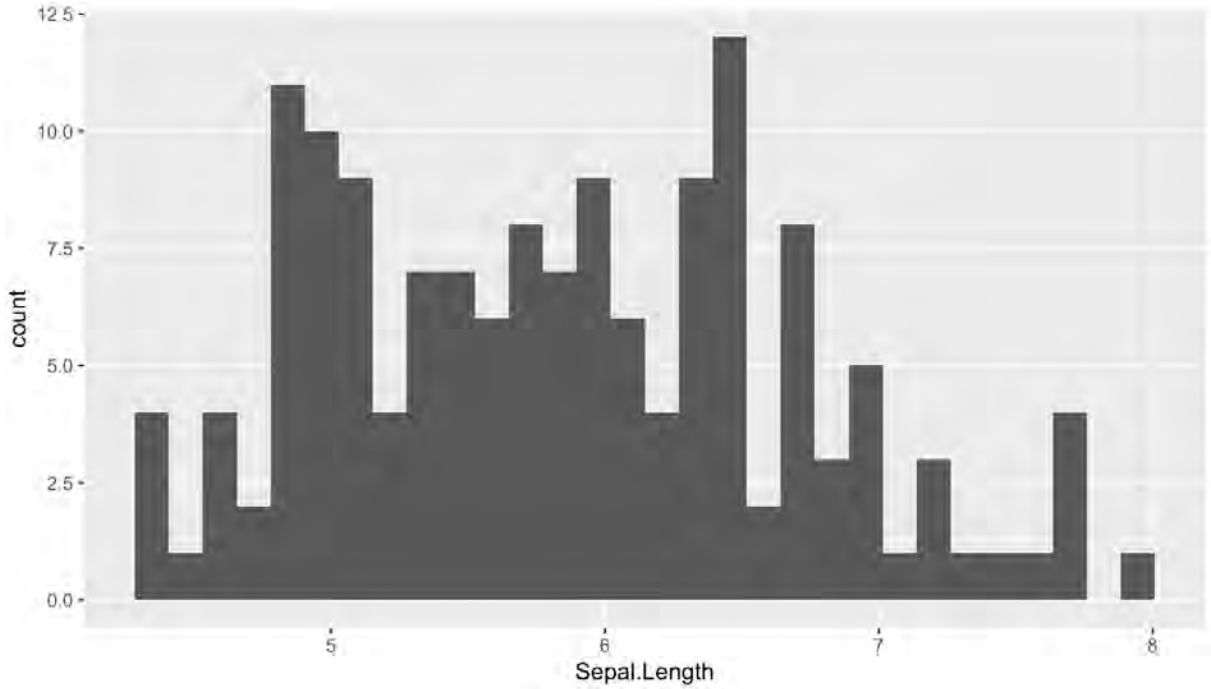
alpha and jitter help reveal overlapping data points



Histogram – basic plot

Not very useful to distinguish between species

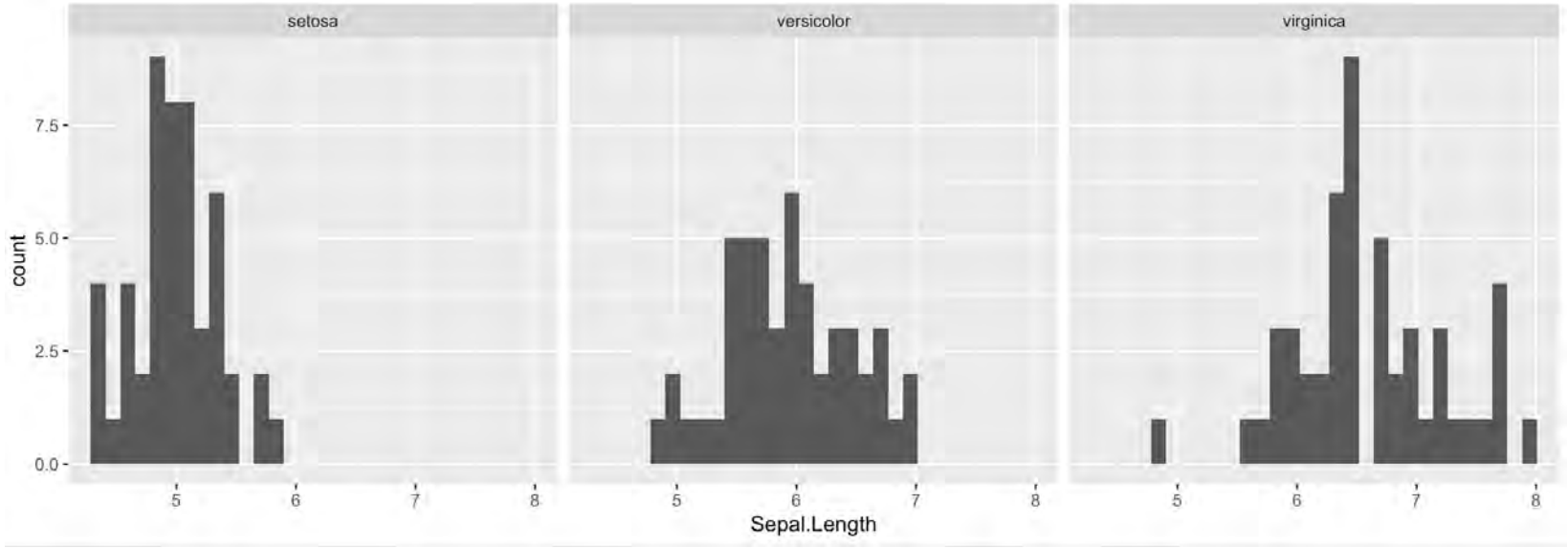
```
> ggplot(data = iris, aes(x = Sepal.Length)) +
  geom_histogram()
```



Histogram – faceting by species

We can now see sepal length for each species

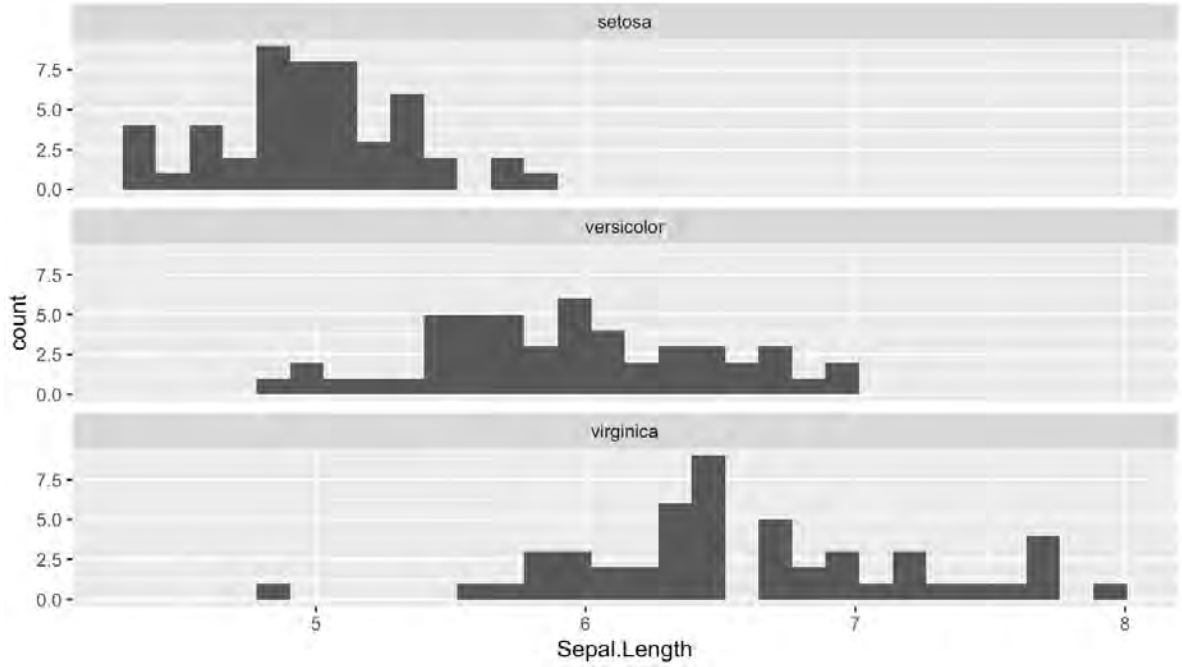
```
> ggplot(data = iris, aes(x = Sepal.Length)) +
  geom_histogram() + facet_wrap(~Species)
```



Histogram – faceting by species

Alternative format

```
> ggplot(data = iris, aes(x = Sepal.Length)) +
  geom_histogram() + facet_wrap(~Species, nrow = 3)
```



Creating plots by name

To improve your graphs first define them by name (as a graph object)

- You can progressively add features.
- Use a script to make this process easier.

For the previous plot:

```
> g = ggplot(data = iris, aes(x = Sepal.Length))
> g = g + geom_histogram()
> g = g + facet_wrap(~Species, nrow = 3)
> g # to display graph
```

ggplot2: grammar

Graphs are constructed first with a

- Geom, which specifies the type of plot and the data

Following this, aesthetic elements are added

- Statistics (summaries, data transformations)
- Scales/coordinate systems
- Faceting (conditional grouping of subsets of data)
- Position adjustments (jitter etc.)
- Annotation
- Aesthetics

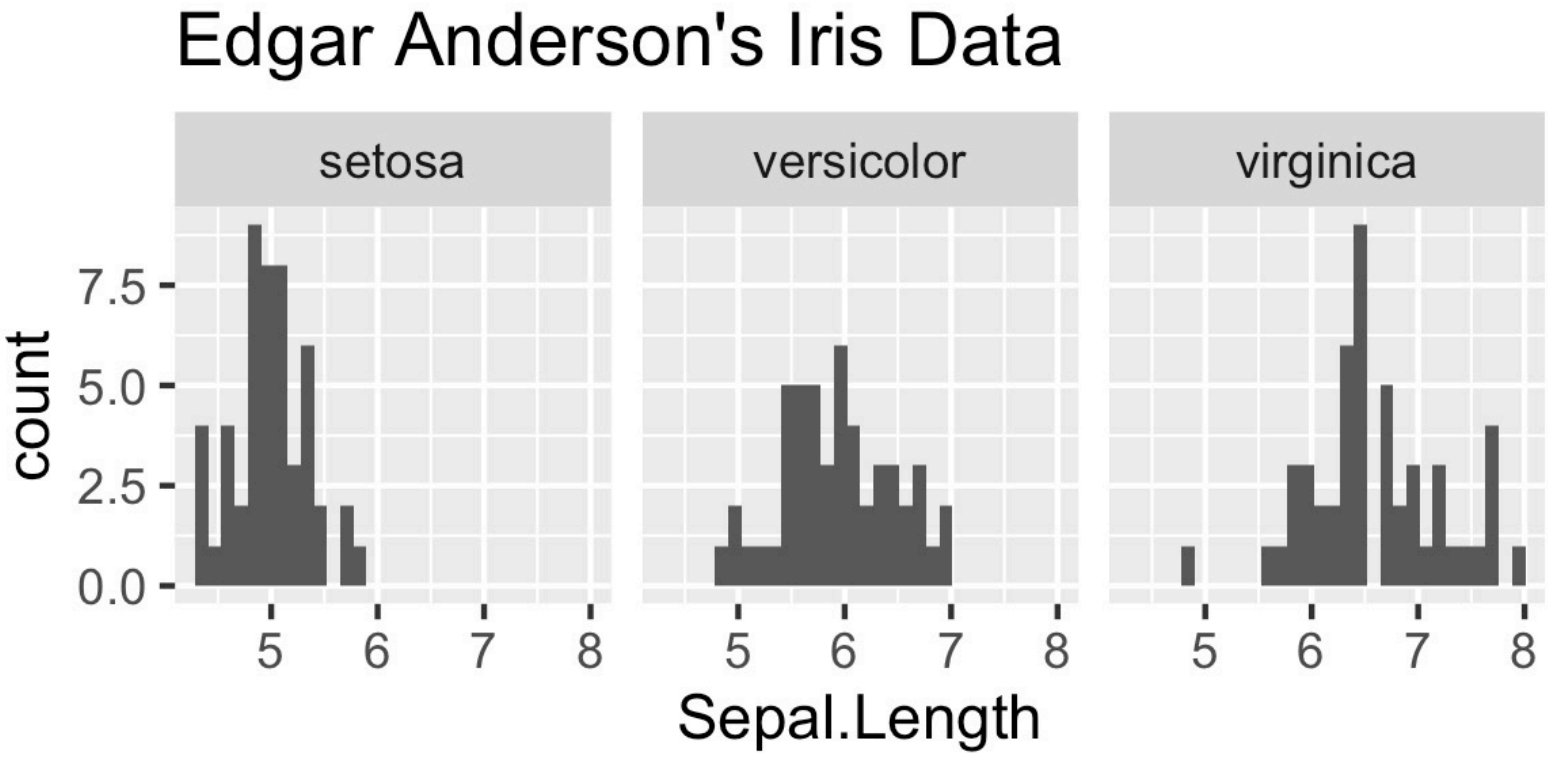
Adding a title and saving

To add a title, and save:

```
> g = ggplot(data = iris, aes(x = Sepal.Length))
> g = g  + geom_histogram()
> g = g  + facet_wrap(~Species)
> g = g + ggtitle("Edgar Anderson's Iris Data")
> g # to display graph
> ggsave("EAI.jpg", g, width = 10, height = 5, units = "cm")
```

Adding a title and saving

Finished plot



Extension: displaying data compactly

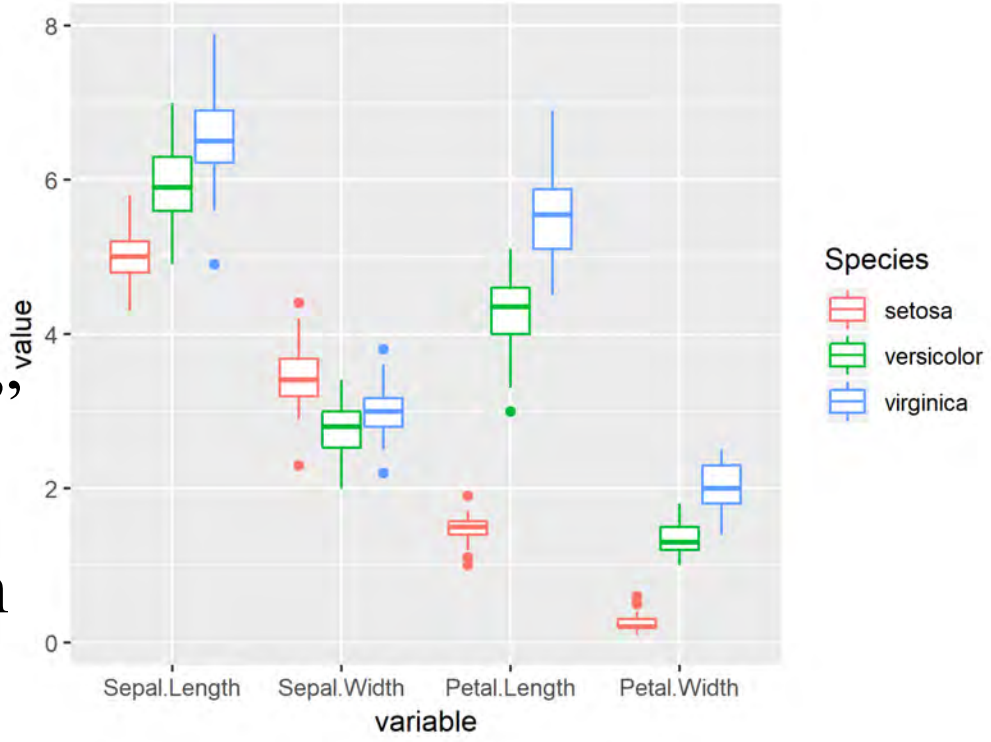
Three examples of graphics to display attributes by multiple factors compactly follow:

- Side-by-side boxplots
- Heatmaps
- Correlation Matrix

- We haven't covered the the data formatting necessary to plot these yet. (We'll do this next week.)

Side-by-side boxplots

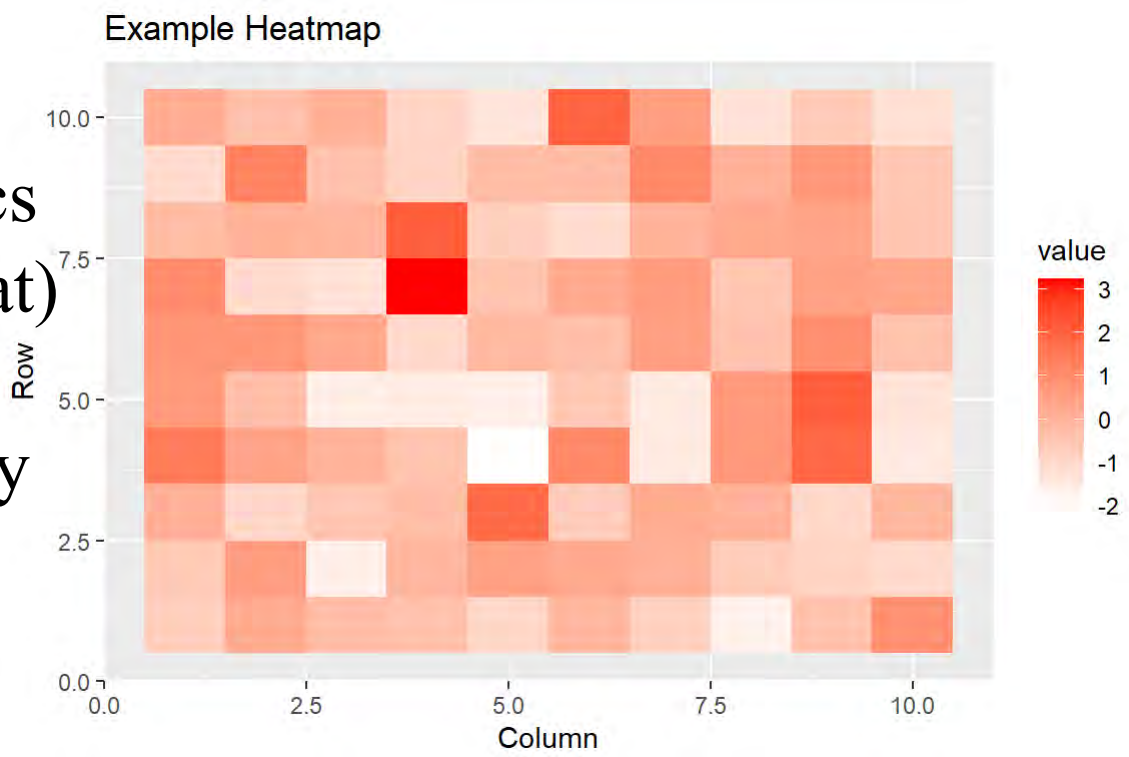
- Using ggplot2.
- Factor levels of each variable shown side-by-side.
- Data needs to be in a "long" format (then grouped by variables/factors) - more on this in next week's lecture.



https://statisticsglobe.com/draw-multiple-boxplots-in-one-graph-in-r

Heatmaps

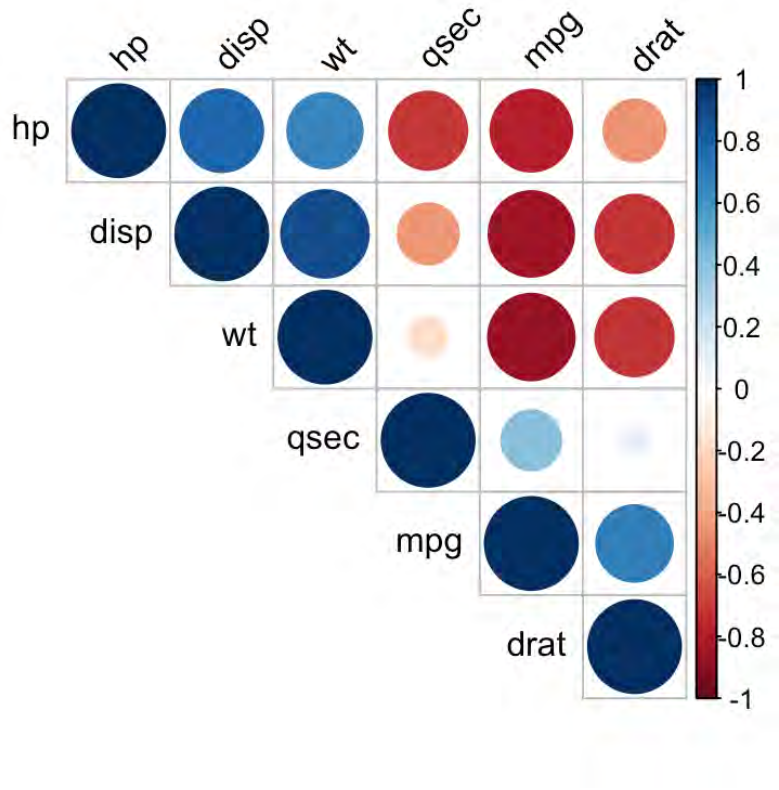
- Values of a variable by two (x,y) factors.
- Can use base graphics (data in matrix format)
- Or ggplot2 (data in "long" format). Many other packages too. Example uses plotly.



https://rpubs.com/lumumba99/1026665

Correlation matrix

- Displays pair-wise correlation for several variables.
- Calculate correlation first. This plot uses corrplot package.
- Colour and size are both based on correlation coefficient in this example.



https://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide...

Elements of good visual display

Elements of good visual display 1

From Edward Tufte (Vis Disp Quant Inf),
Graphical displays should:

- show the data,
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else,
- avoid distorting what the data have to say,
- present many numbers in a small space,
- make large data sets coherent, cont…

Elements of good visual display 2

- encourage the eye to compare different pieces of data,
- reveal the data at several levels of detail, from a broad overview to the fine structure,
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration, and
- be closely integrated with the statistical and verbal descriptions of a data set.
- Graphics reveal data. Indeed, graphics can be more precise and revealing than conventional statistical computations.

Elements of good visual display 3

See also:

Five principles of good graphs

https://scc.ms.unimelb.edu.au/

Statistician Heal Thyself: Have We Lost the Plot?

https://www.tandfonline.com/ (use institutional login)

- Chapter 6, A checklist for good graphical practice.
- These are included in the applied session notes for you to discuss in class.

Assignment 1

Assignment 1

MONASH University

Faculty of Information Technology

FIT3152 Data analytics – 2026: Assignment 1

Your task	• Analyse the country level predictors of confidence in **social organisations and how these change over time** using data from the World Values Survey. • This is an individual assignment.
Value	• This assignment is worth **25%** of your total marks for the unit. • It has **40** marks in total.
Suggested Length	• 8 – 10 A4 pages, approximately 2,000 words (for your report) + extra pages as appendix for your R script and report on how Generative AI used, if required. • Font size 11 or 12pt, single spacing.

Assignment 1

Due Date	**11.55pm Friday 17th April 2026**
Submission	• Submit a single PDF file **and** single video file on Moodle. • Note that submission of a video report is a hurdle requirement. • Use the naming convention: *FirstnameSecondnameID.{pdf, mp4, mov etc.}* • Turnitin will be used for similarity checking of all written submissions.
Generative AI Use	• In this assessment, you can use generative artificial intelligence (AI) in order to search for R functions and examples to perform tasks that you specify only. Any use of generative AI must be appropriately acknowledged (see Learn HQ).
Late Penalties	• 5% (**2** mark) deduction per calendar day for up to one week. • Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.

Assignment 1

Instructions

Address each of the research questions below and report the results of your analysis and your interpretation of those results. Report any assumptions you've made in your analysis. Include your R code as an appendix. Your R code must be machine readable text as the university requires all student submissions to be processed by plagiarism detection software. See information on Generative AI below.

There are two options for compiling your written report:
(1) You can create your report using any word processor with your R code pasted in as machine-readable text as an appendix, and save as a pdf, or
(2) As an R Markdown document that contains the R code with results and discussion interleaved. Render this as a HTML file and save as a pdf.

Your video report should be less than 100MB in size. You may need to reduce the resolution of your original recording to achieve this. Use a standard file format such as .mp4, or mov for submission.

Assignment 1

Software

It is expected that you will use R for your data analysis, graphics and tables. You are free to use any R packages you need but must document these in your report and include in your R code.

Use of Generative AI

AI & Generative AI tools may be used in GUIDED ways within this assessment/task as per the guidelines provided.

In this assessment, you can use generative artificial intelligence (AI) in order to search for R functions and examples to perform tasks that you specify only. Any use of generative AI must be appropriately acknowledged (see Learn HQ).

If you do use Generative AI for your assignment, then you must include the statement "Generative AI was used in this assignment." in the introductory/first paragraph of your report. You must also include the following information as an appendix in your report: (1) the technology you used (e.g. ChatGPT), (2) the information that was generated (e.g. R code fragments), (3) the prompts used (i.e. the questions you asked), and (4) how the output was used in your work.

If you did not use generative AI in your assignment, then include the statement "Generative AI was not used in this assignment." in the introductory/first paragraph of your report.

Assignment 1

Questions

The World Values Survey (WVS) is an international research program that studies the social, political, economic, religious and cultural attitudes and values of people around the world. You can read more here: https://www.worldvaluessurvey.org/WVSContents.jsp.

For this assignment you will analyse data collected over Waves 1 - 7, from 1981 to 2022. The aim of this assignment is to understand country-level differences in participant responses and the predictors of confidence in social organisations, and how these responses and predictors of confidence have changed over time.

Social organisations include aspects of society such as religion, armed forces, the press, television, trade unions, police, the courts, government, banks, and international and environmental organisations etc. They are indicated in your data by column names having the prefix "C". Predictor variables (**attributes**) include personal information such as age and gender, happiness indicators, attitudes and values towards others, political and social views and participation.

Each student will be assigned a **different** subset of organisations and attributes to study. Your task is to analyse **all** the survey data assigned to you, with a **focus** on the country you have been allocated.

Assignment 1

1. Descriptive analysis. (5 Marks)

(a) Describe the data overall, including things such as dimension, data types, distribution of numerical responses, variety of non-numerical (text) responses, missing values, and anything else of interest or relevance.

Assignment 1

2. Focus country vs all other countries as a group (independent of time). (13 Marks)

For Question 2 ignore the effect of time. That is, do not separate your data by years or waves when answering the questions below.

(a) Identify your focus country from the accompanying list (**WVSFocusCountry.pdf**). How do participant responses for your focus country differ from the other countries in the survey (treating them as a group)?

(b) How well do participant responses (attributes) predict confidence in social organisations in your **focus country**? Which attributes seem to be the best predictors? Confidence in which social organisations can be more reliably predicted? Explain your reasoning.

(c) Repeat Question 2(b) for the **other countries** as a group. Which attributes are the strongest predictors? Confidence in which social organisations can be more reliably predicted? How do these results compare to those of your focus country?

Assignment 1

3. Focus country vs all other countries as a group (over time). (12 Marks)

For Question 3 study the effect of time by separating your data by years or waves when answering the questions below.

(a) How do participant responses for your **focus country** vary over time (using either years or successive waves)? Describe these changes over time and comment on whether they are significant or not. Perform the same analysis for the **other countries** (as a group) and compare the results with your focus country. What are the major differences between the two groups? Create a graphic enabling a reader to compare results (focus *vs* other countries) over time, for the most interesting results. Describe your reasoning for the design of the graphic.

(b) How does the ability of participant responses (attributes) to predict confidence in social organisations in your **focus country** change over time? Do the important attributes for predicting confidence change over time? Perform the same analysis for the **other countries** (as a group) and compare the results. What are the major differences between the two groups? Create a graphic enabling a reader to compare results (focus *vs* other countries) over time, for the strongest predictors. Describe your reasoning for the design of the graphic.

Assignment 1

4. Video Presentation: (Submission Hurdle and 4 Marks)

Record a short presentation using your smartphone, Zoom, or similar method. Your presentation should be approximately 5 minutes in length and summarise your main findings for Sections 1 – 3, as well as describing how you conducted your research, any assumptions made, and how you designed your graphics.

5 Overall considerations (6 Marks)

This includes: the quality and clarity of your reasoning and assumptions; the strength of support for your findings; the quality of your writing in general and communication of results;-the quality of your graphics throughout; the quality of your R coding.

Assignment 1

Data

The data for this assignment is a reduced version of the World Values Survey Waves 1 -7 data. The filename is "WVSExtract.csv". The data includes ordinal data coded on a numerical scale. For this assignment assume it is reasonable to treat these responses as numerical.

Create your individual data as follows:

```
rm(list = ls())
set.seed(12345678) # Your Student Number
VCData = read.csv("WVSExtract.csv")
VC = VCData[sample(1:nrow(VCData),100000, replace=FALSE),]
VC = VC[,c(1:3,sort(sample(4:50,25,replace=FALSE)),
sort(sample(51:65,8,replace=FALSE)))]
#write.csv(VC, "FIT3152A1Data_YourName.csv", row.names = FALSE)
```

You can save the “VC” file you created by uncommenting the last line above. You can then delete the WVSExtract.csv file you downloaded.

Locate your focus country using the accompanying document WVSFocusCountry.pdf. A list matching country names with three letter code is in WVSCountryCodes.pdf.

Assignment 1

Data fields and brief descriptor

Most fields are on integer scales over varying range. The convention is that larger numbers generally indicate greater agreement with statement or frequency of occurrence. Some exceptions given below. Fields in bold indicate confidence in social organisations.

You can access more detail on each field in your data from the *WVS-7 Master Questionnaire 2017-2020 English.pdf*, linked from https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp. Use the question ID given in the **WVS Wave 7 Reference** in the table below.

Column Name	**Original Descriptor**	**WVS Wave 7 Reference**
Wave	Chronology of EVS-WVS waves	A_WAVE
Country	ISO 3166-1 alpha-3 country code	B_COUNTRY_ALPHA
Year	Year survey	A_YEAR
ILFam	Important in life: Family	Q1
ILFriends	Important in life: Friends	Q2
ILLeisure	Important in life: Leisure time	Q3
ILPolitics	Important in life: Politics	Q4
ILWork	Important in life: Work	Q5

Assignment 1

CChurches	Confidence: Churches	Q64
CArmedForces	Confidence: Armed Forces	Q65
CPress	Confidence: The Press	Q66
CUnions	Confidence: Labour Unions	Q68
CPolice	Confidence: The Police	Q69
CParliament	Confidence: Parliament	Q73
CCivilService	Confidence: The Civil Services	Q74
CTelevision	Confidence: Television	Q67
CGovernment	Confidence: The Government	Q71
CPolParties	Confidence: The Political Parties	Q72
CMajComp	Confidence: Major Companies	Q77
CEnvProt	Confidence: The Environmental Protection Movement	Q79
CWomensMvt	Confidence: The Womens' Movement	Q80
CCourts	Confidence: Justice System/Courts	Q70
CEU	Confidence: The European Union	Q82_EU

Assignment 1

JB EDA V00.R | VC

Filter

	Wave	Country	Year	ILFam	ILWork	ICQIndependence	ICQHardWork	ICQTolerance	ICQDetermination	ICQFaith	ACTReligion	ACTSport	ACTUnions
149999	5	JOR	2007	1	1	1	1	0	0	1	0	0	0
57968	5	UKR	2006	1	2	1	0	1	0	1	2	1	0
34203	7	AUS	2018	1	2	1	1	1	1	0	1	0	0
128452	6	BLR	2011	1	1	0	1	1	1	0	0	0	0
196160	7	NLD	2022	1	2	1	1	1	0	0	1	2	0
54028	7	USA	2017	2	1	0	1	1	1	0	0	0	0
45965	5	HUN	2009	2	2	1	0	1	1	0	0	0	0
102378	3	GEO	1996	1	1	1	1	1	0	0	0	0	0
84273	4	SGP	2002	1	3	1	0	1	0	1	-4	-4	-4
92075	5	FIN	2005	1	1	1	0	1	1	0	2	2	0
180253	6	MEX	2012	1	1	1	0	1	1	1	0	0	0
111049	3	NZL	1998	1	2	0	0	1	1	0	1	2	0
135207	5	GBR	2005	1	1	1	1	1	1	0	2	1	0
146977	7	MYS	2018	1	1	0	0	1	0	1	2	0	2
59242	7	CHN	2018	1	1	0	1	1	0	0	0	0	0
149092	5	CHE	2007	1	1	0	1	0	0	0	0	0	0
39163	4	ZAF	2001	1	1	0	0	1	0	1	-4	-4	-4
79507	7	GBR	2022	1	1	0	0	1	0	0	1	0	0
107645	3	SVK	1998	1	1	0	1	0	0	1	2	0	0
115588	7	ROU	2018	1	1	0	0	0	1	0	0	0	0
180425	6	ARM	2011	1	1	0	1	1	0	0	0	0	0
84826	2	NGA	1990	1	1	0	1	0	0	1	-4	-4	-4
53654	5	USA	2006	1	1	0	1	1	0	1	0	0	1
65630	5	UKR	2006	1	2	0	1	1	0	0	0	0	0
139335	5	GEO	2009	1	1	1	1	0	0	1	0	0	1
160800	6	MAR	2011	2	2	0	1	0	0	0	0	0	0
96936	3	FIN	1996	1	1	1	0	0	0	0	2	0	0

Summary

This week we've covered:

- Visualising data
- Getting to know a data set
- Graphing data in R
- Assignment 1

References

Books – online from the Monash Library

- Wickham, H., ggplot2 elegant graphics for data analysis, Springer
- Wilkinson, L., and Wills, G., The grammar of graphics, Springer
- Rahlf, T., Data visualisation with R, Springer.

Wickham, H., R for data science https://r4ds.hadley.nz/

Chang, W., R Graphics Cookbook r-graphics.org/

Peng, R., Exploratory Data Analysis with R https://bookdown.org/

Wickham, H., Layered grammar of graphics http://vita.had.co.nz/

Five principles of good graphs https://scc.ms.unimelb.edu.au/

Gordon & Finch, Statistician Heal Thyself: Have We Lost the Plot? https://www.tandfonline.com/

Heer, et. al., A tour through the visualization zoo https://dl.acm.org/doi/

ggplot2 Cheat Sheet https://github.com/rstudio/cheatsheets/

Reference: *ggplot2*

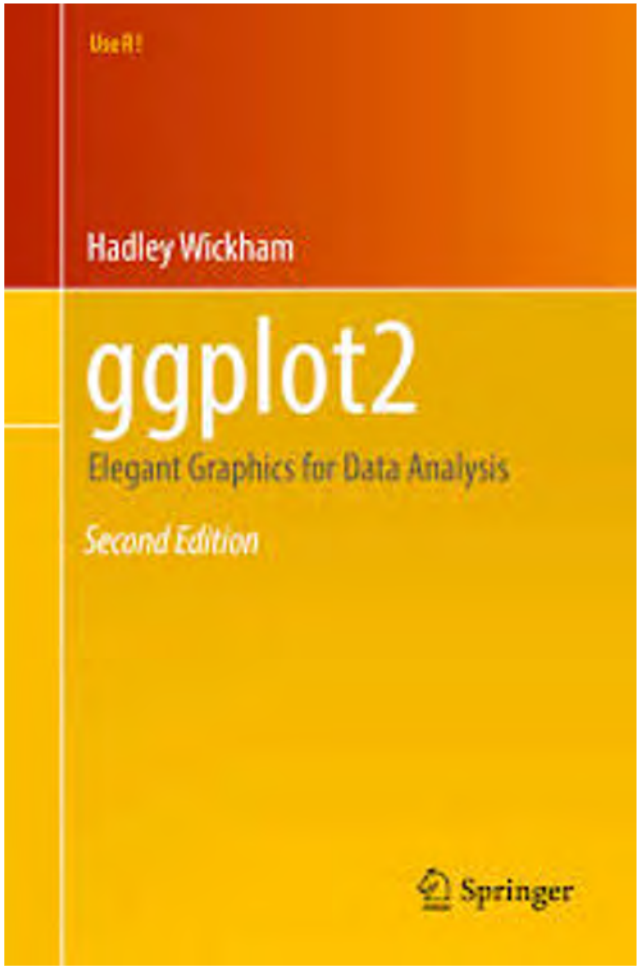
Access book via Monash library, or help from links below:

- ggplot2 is a plotting system for R, based on the grammar of graphics.

https://ggplot2.tidyverse.org/

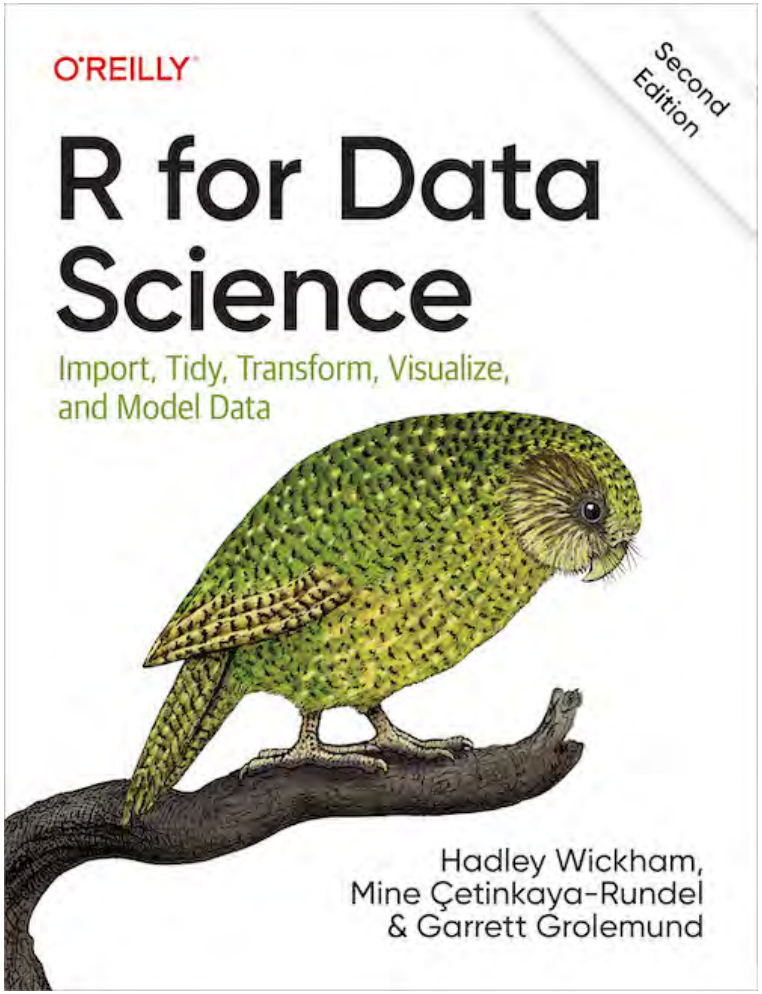
- Online help links from main page and is a useful reference. Many examples with code are given.

https://ggplot2.tidyverse.org/reference/



Reference: *R for Data Science*

- A physical and web-based book by the author of ggplot2, Hadley Wickham, and others:
  https://r4ds.hadley.nz/
- The book takes you through all aspects of the data science workflow.
- Chapter 1 is on ggplot2, including syntax and incrementally building up complex graphs.



Reference: *R Graphics Cookbook*

The R Graphics Cookbook has 150 recipes for graph drawing.

It covers ggplot and base graphics.

It explains the reasoning behind the recipe.

Available online for free at

r-graphics.org/