

FIT3152 Data analytics – Lecture 3

Data Manipulation in R

- Making tables and summaries, Working with factors,
- Transforming data, *dplyr* package.

Visualising data...

- Quick follow up of last week's lecture

R Tips

- Scripts, R Markdown, Notebooks, User-defined functions

Assignment 1

Consultations

Consultations have commenced.

Most are on Zoom.

Check Moodle for days/times:

<https://lms.monash.edu/course/view.php?id=153815§ion=2>

Unit outline (week-by-week)

Clayton lecture is Wednesday 11:00am – 1:00pm (AEDT).
Tutorials begin Week 2 and follow lecture by a week.

Week Starting	Lecture	Topic	Tutorial	A1 25	A2 30	Q/P 25	A3 20	Due	
27/2/23	1	Intro to Data Science, review of basic statistics using R	...						
6/3/23	2	Exploring data using graphics in R	T1						
13/3/23	3	Data manipulation in R	T2						
20/3/23	4	Regression modelling	T3						
27/3/23	5	Clustering	T4						
3/4/23	6	Data Science methodologies, dirty/clean/tidy data	T5						
10/4/23	-	Mid-semester Break	-	-	-	-	-	-	
17/4/23	7	Classification using decision trees	T6					17/4/23	Mo
24/4/23	8	Naïve Bayes, evaluating classifiers	T7						
1/5/23	9	Ensemble methods, artificial neural networks	T8						
8/5/23	10	Text analysis	T9					12/5/23	Fr
15/5/23	11	Network analysis	T10					19/5/23	Fr
22/5/23	12	Review of course	T11						
29/5/23		SWOT VAC							
5/6/23		EXAM PERIOD						9/6/23	Fr

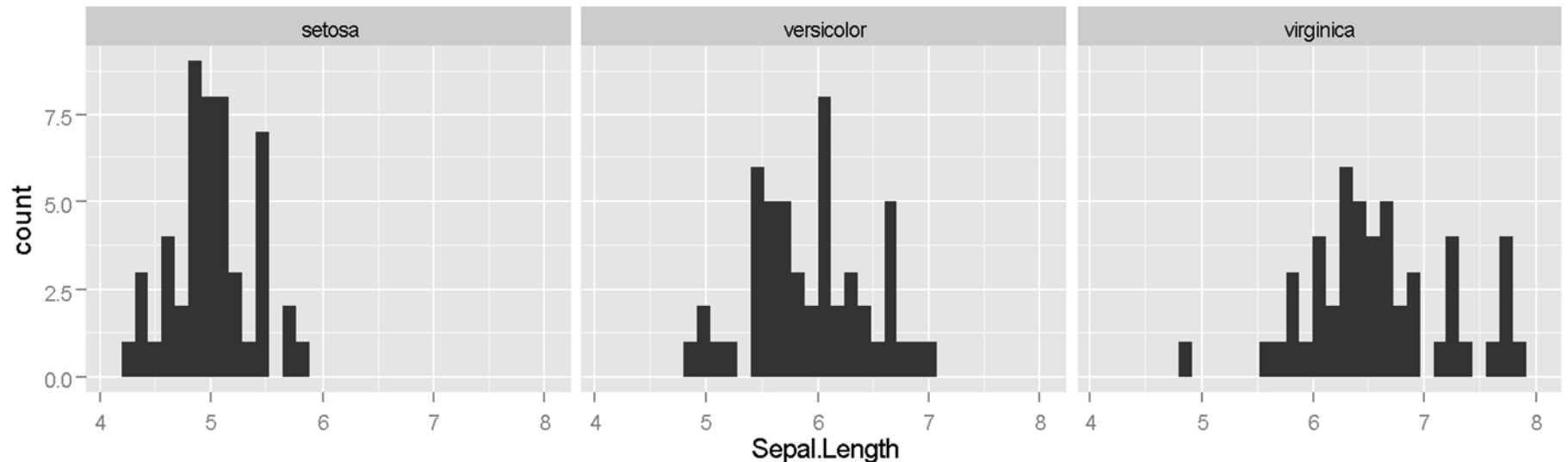
Brief review of Lecture 2...

Visualising data

- The number of dimensions in a data set
- Major families of graph types: time series, statistical distributions, maps, hierarchies, networks.
- Plotting the Iris data: basic scatterplot and increasing the number of dimensions presented.
- lattice and ggplot2 packages, and the Grammar of Graphics approach.

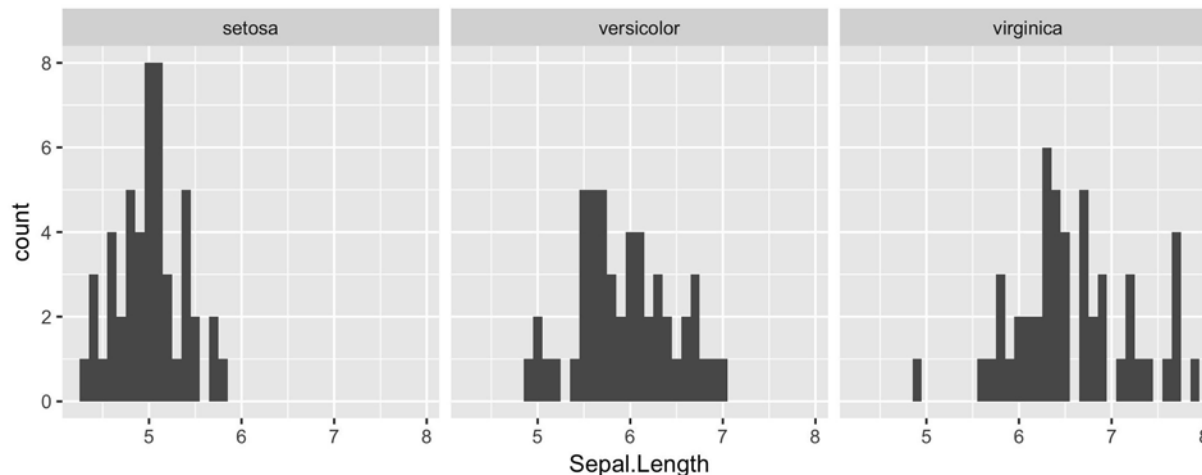
Syntax: histogram + facet_wrap

```
> qplot(Sepal.Length, data = iris, geom = "histogram",  
  facets = Species ~ .) + facet_wrap(~ Species, ncol = 3)
```



Using the grammar approach...

- > `m = ggplot(iris, aes(x = Sepal.Length))` #data
- > `m = m + geom_histogram(binwidth = 0.1)` #graph type
- > `m = m + facet_wrap(~Species, ncol = 3)` #grouping var
- > `ggsave("irissepallen.jpg", m, width = 20, height = 8, units = "cm")`



MPG example

Recall the mpg data set.

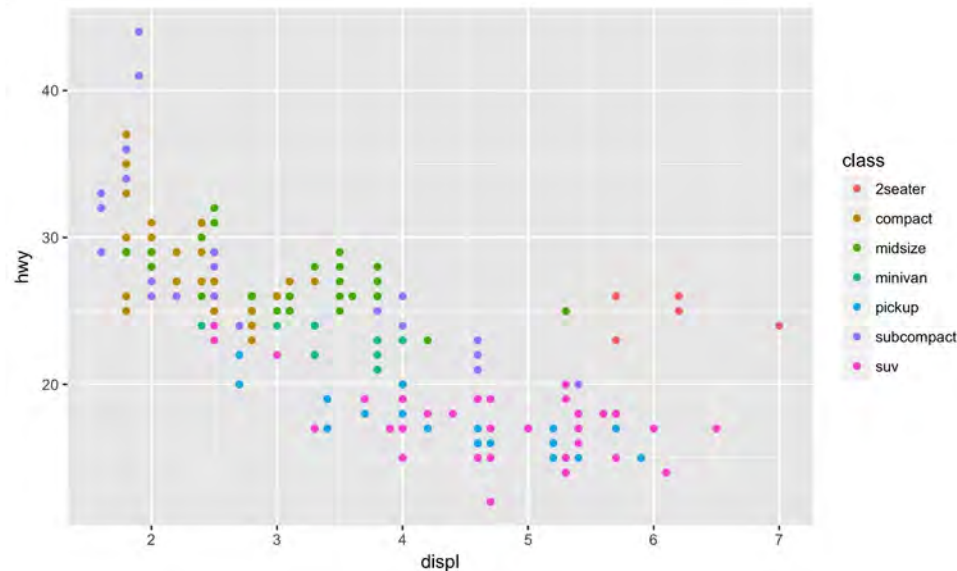
```
> head(mpg)
# A tibble: 6 x 11
  manufacturer model displ  year  cyl    trans  drv  cty   hwy fl    class
    <chr>    <chr> <dbl> <int> <int>    <chr> <chr> <int> <int> <chr>   <chr>
1      audi     a4   1.8  1999     4 auto(l5)   f    18    29 p compact
2      audi     a4   1.8  1999     4 manual(m5)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

Investigate the relationship between fuel consumption and engine displacement.

Basic plot

Using a grammar of graphics approach (from R4DS)

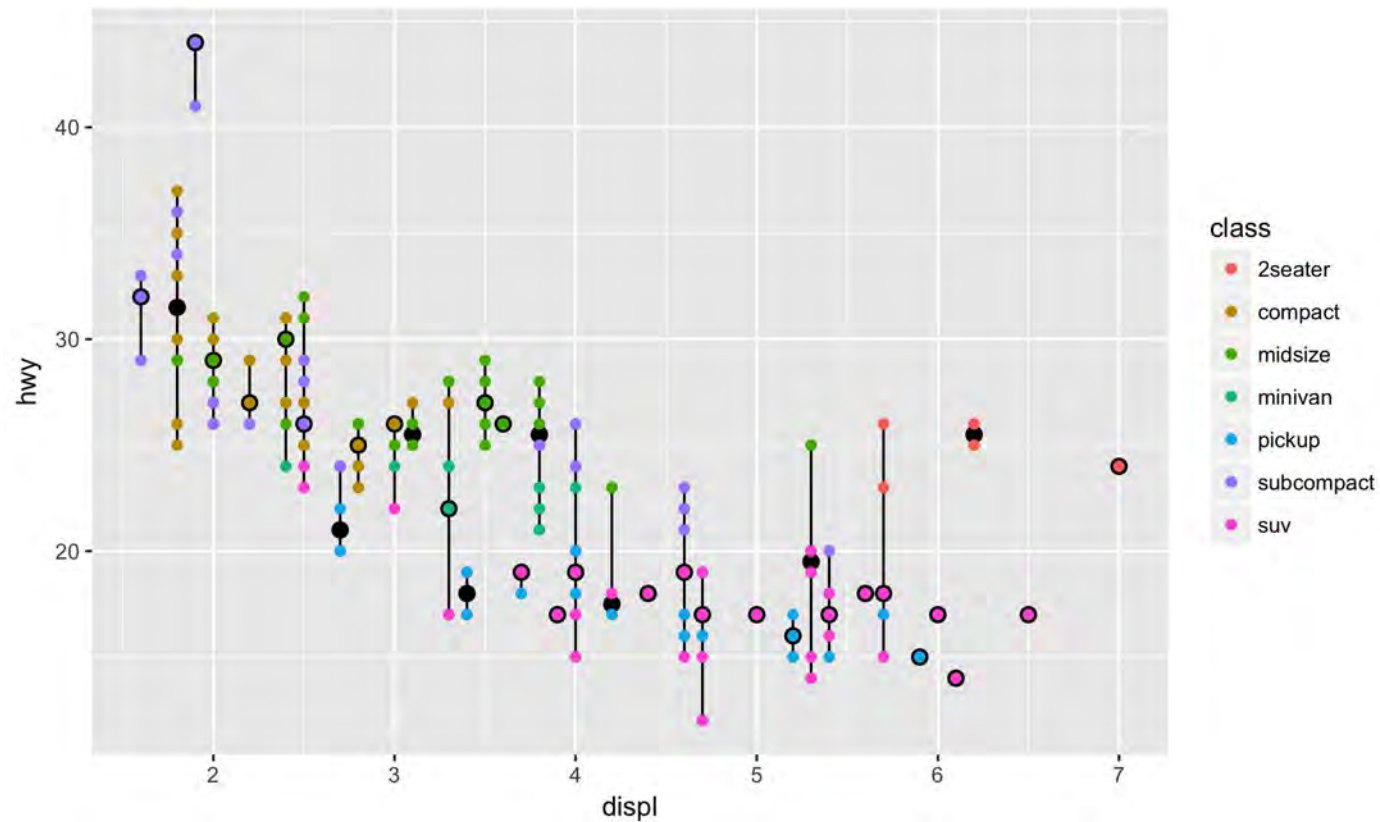
- > `g = ggplot(data = mpg)`
- > `g = g + geom_point(mapping = aes(x = displ, y = hwy, color = class))`
- > `g`



Underplotting (min, median, max)

- > d <- ggplot(mpg, aes(displ, hwy, color = class)) +
geom_point()
- > d = d + stat_summary(mapping = aes(x = displ, y =
hwy), fun.min = min, fun.max = max, fun = median,
orientation = "x", colour = "black")
- > d = d + geom_point(mapping = aes(x = displ, y = hwy,
color = class)) **# overplots original points**
- > d
- > ggsave("hwyvdispl.jpg", d, width = 20, height = 12,
units = "cm")

The plot. What improvements would you make?



Improving: axes, title, legend

- > d = d + theme(axis.text = element_text(size = 8))
- > d = d + theme(axis.title = element_text(size = 10))
- > # could by axis.text.x or .y etc. to adjust separately
- > d = d + xlab("Engine Displacement (litres)")
- > d = d + ylab("Highway Fuel Consumption (mpg)")
- > d = d + ylim(10,50) + xlim(1,7) ←
- > d = d + theme(plot.title = element_text(size = 14))
- > d = d + theme(plot.title = element_text(hjust = 0.5))
- > d = d + ggtitle("Highway ... and Class")
- > d = d + theme(legend.position = c(0.91, 0.76),
legend.key.height= unit(0.5, 'cm'))

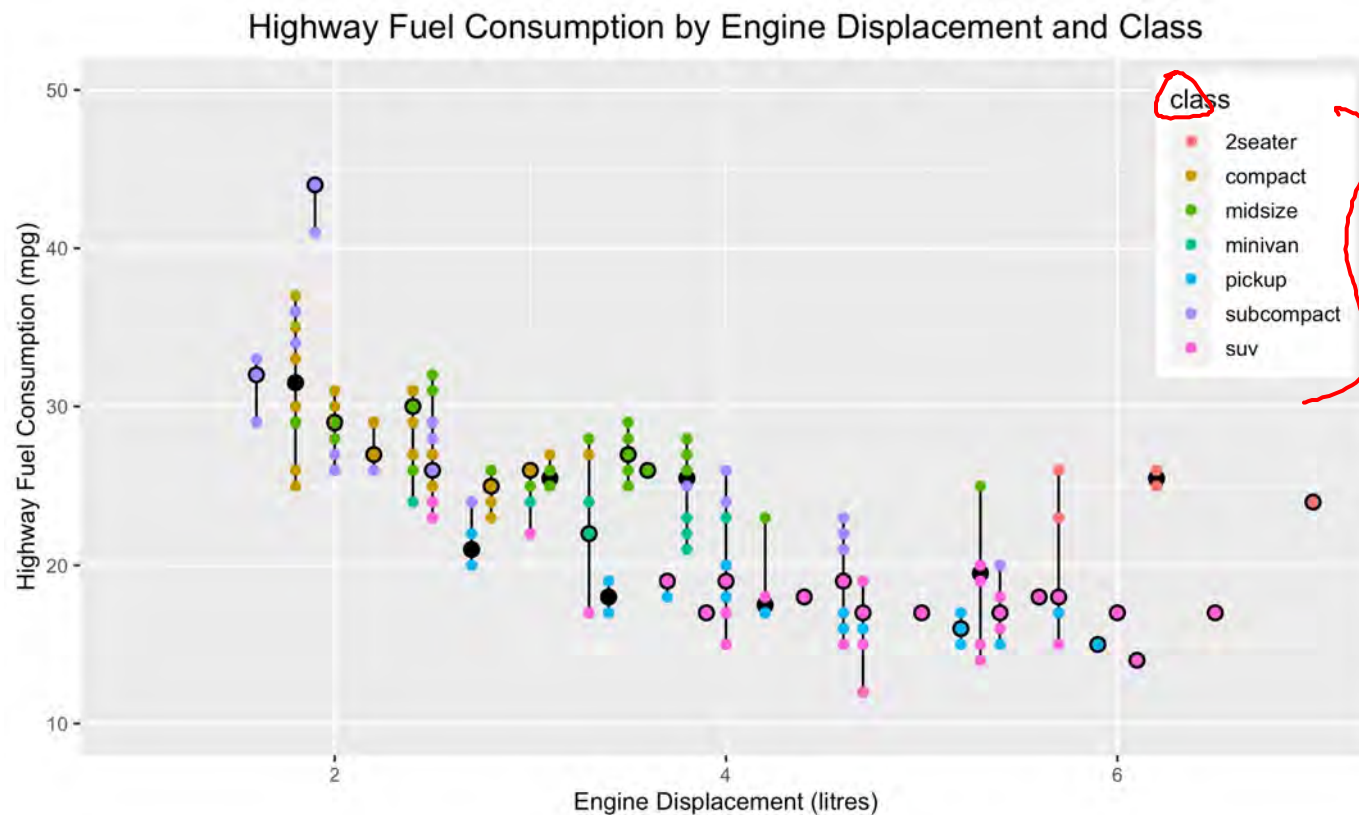
Size

Labels

Title

Position
Legend

Making incremental improvements by trial and error

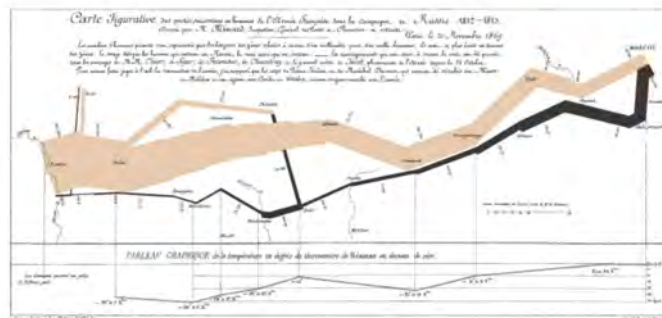


better
labels

Better graphics

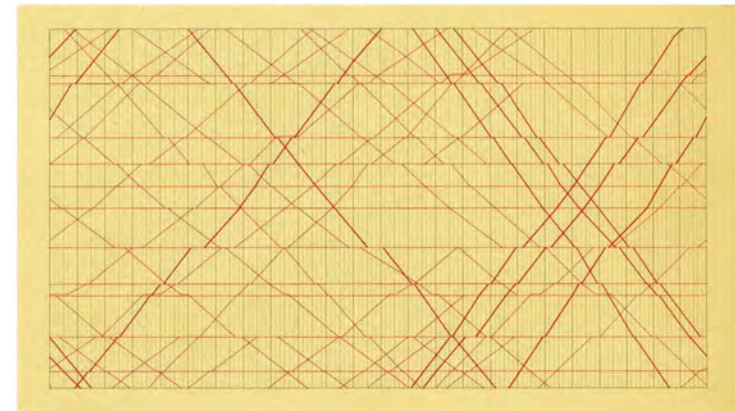
One source of inspiration is Edward Tufte:

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



Napoleon's March to Moscow The War of 1812

The issue of human population density across the landscape remains an important topic in human geography. In the past, it has been argued that the density of human settlements is a function of the availability of land resources. However, more recent research has shown that the density of human settlements is also a function of the availability of water resources. This is because water is a key resource for human settlements, and its availability can limit the density of human settlements. For example, in arid regions, the density of human settlements is often low because of the lack of water resources. In contrast, in regions with abundant water resources, the density of human settlements is often high. This is because water is a key resource for human settlements, and its availability can limit the density of human settlements.



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

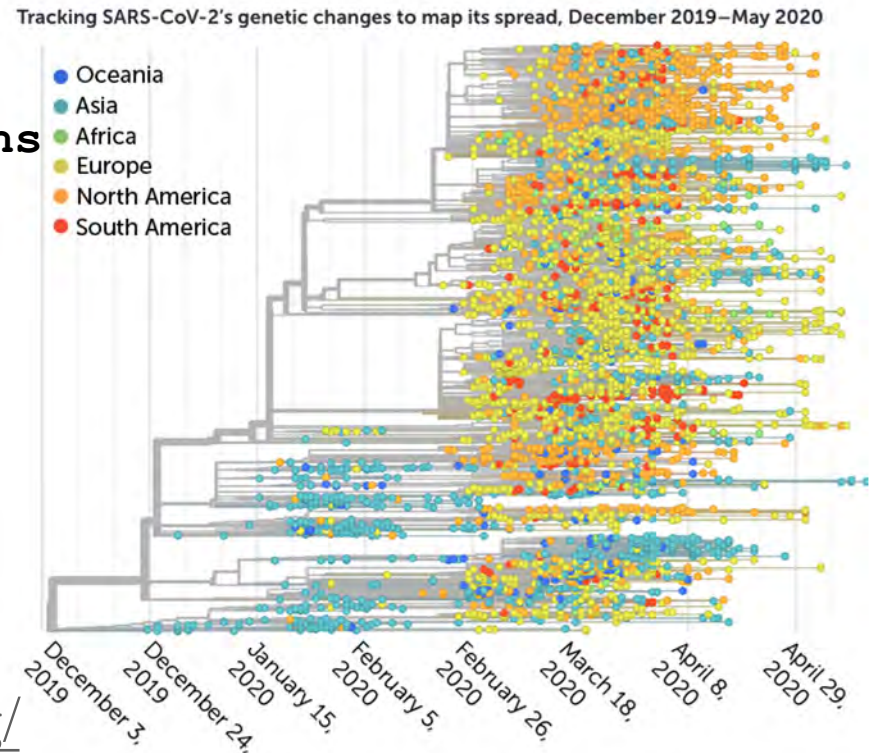
<https://medium.com/>

Review questions

Question 1

The figure is from the _____ graph family?

- A. Time Series
- B. Statistical Distributions
- C. Maps
- D. Hierarchies
- E. Networks

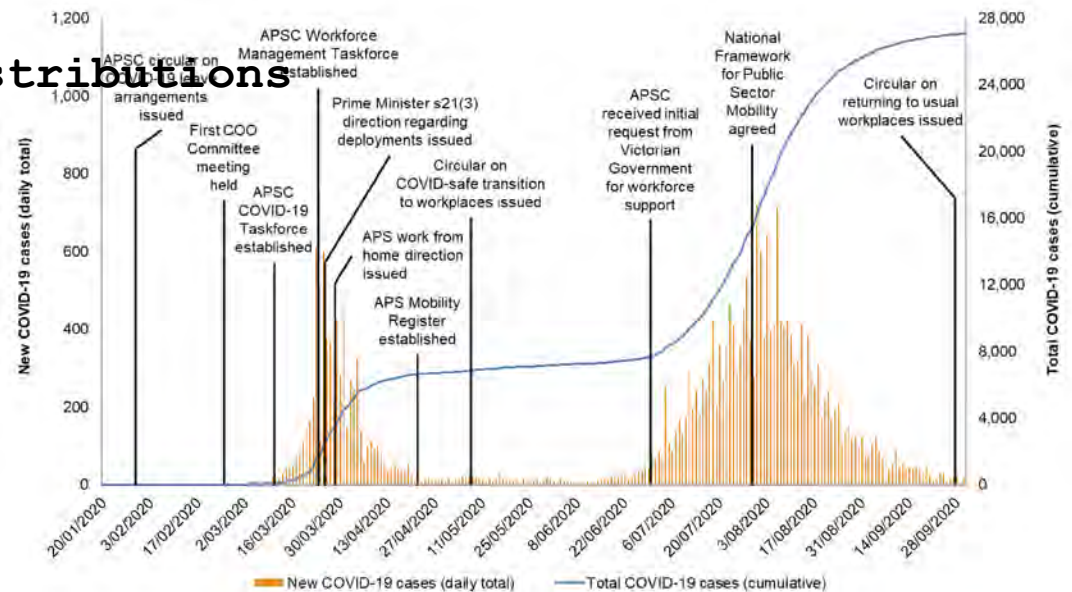


Source: <https://www.sciencenews.org/>

Question 2

The figure is from the _____ graph family?

- A. Time Series
- B. Statistical Distributions
- C. Maps
- D. Hierarchies
- E. Networks



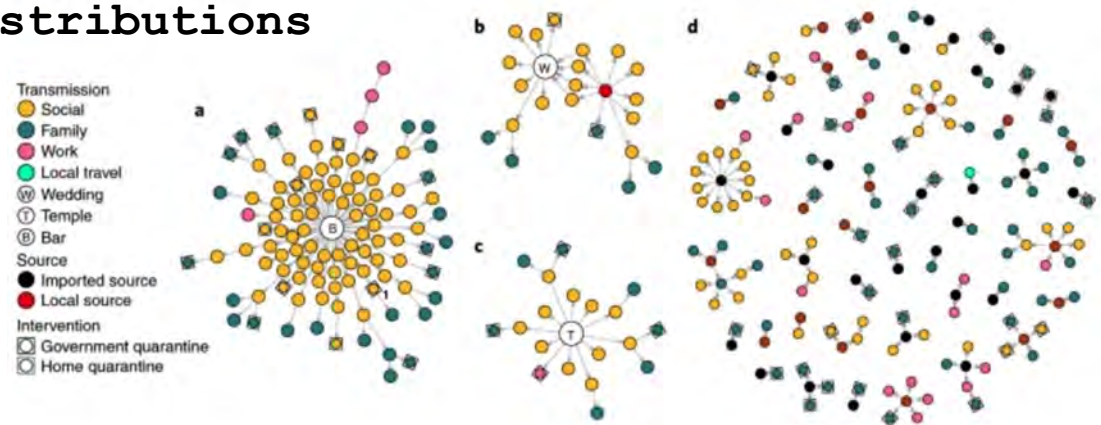
Source: <https://www.anao.gov.au/>

Question 3

The figure is from the _____ graph family?

- A. Time Series
- B. Statistical Distributions
- C. Maps
- D. Hierarchies
- E. Networks

Fig. 2: Chains of SARS-CoV-2 transmission in Hong Kong initiated by local or imported cases.



Source: <https://www.nature.com/>

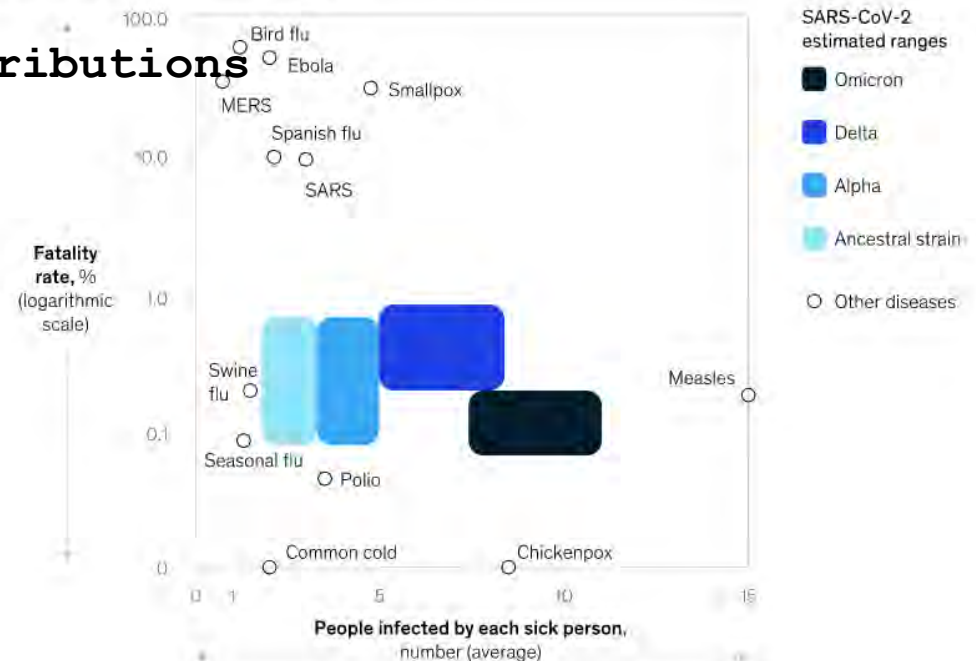
Question 4

The figure is from the _____ graph family?

- A. Time Series
- B. Statistical Distributions
- C. Maps
- D. Hierarchies
- E. Networks

Omicron is more infectious than other common viruses, and less fatal than Delta.

Disease fatality and infection rates¹



<https://www.mckinsey.com/>

Some R tips

Scripts:

- Very important: learn how to use these now if you've not done so already.

RMarkdown:

- Useful if you're doing a job that requires a lot of routine reporting, but not essential. Also, Notebooks.

User-Defined Functions

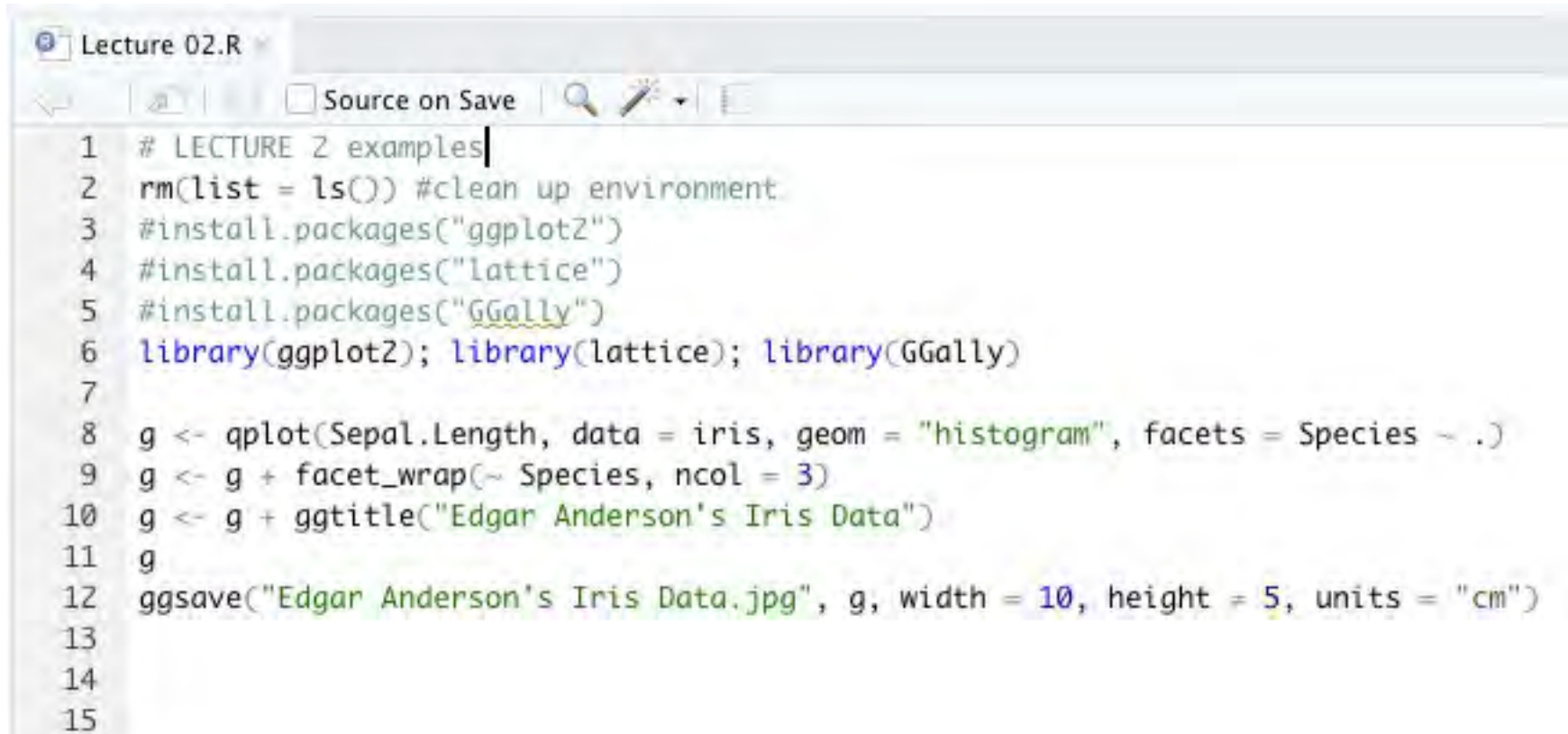
- Useful, and they improve your R code. See also anonymous functions defined on the fly.

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 example from last lecture as a script...

Scripts



```
Lecture 02.R
Source on Save

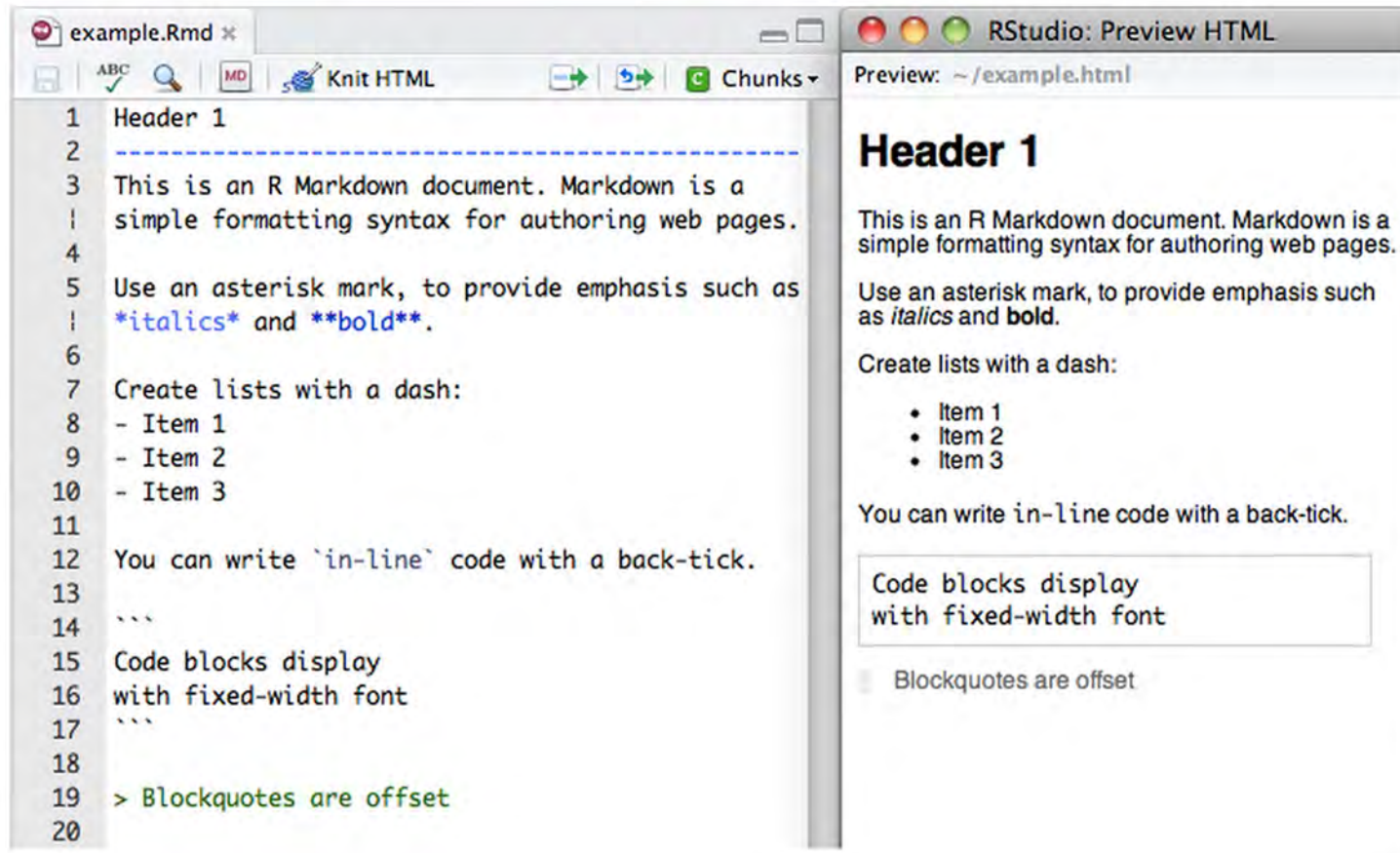
1 # LECTURE 2 examples
2 rm(list = ls()) #clean up environment
3 #install.packages("ggplot2")
4 #install.packages("lattice")
5 #install.packages("GGally")
6 library(ggplot2); library(lattice); library(GGally)
7
8 g <- qplot(Sepal.Length, data = iris, geom = "histogram", facets = Species ~ .)
9 g <- g + facet_wrap(~ Species, ncol = 3)
10 g <- g + ggtitle("Edgar Anderson's Iris Data")
11 g
12 ggsave("Edgar Anderson's Iris Data.jpg", g, width = 10, height = 5, units = "cm")
13
14
15
```

R Markdown

Is a package that enables the creation of HTML and PDF documents etc. based on your R session. You may choose to use it, but it is optional.

- You can embed R code and graphics.
- You can get started with R Markdown by creating a new R Markdown file in R Studio (the required files will be automatically installed).
- <http://rmarkdown.rstudio.com/>

R Markdown



- <http://rmarkdown.rstudio.com/>

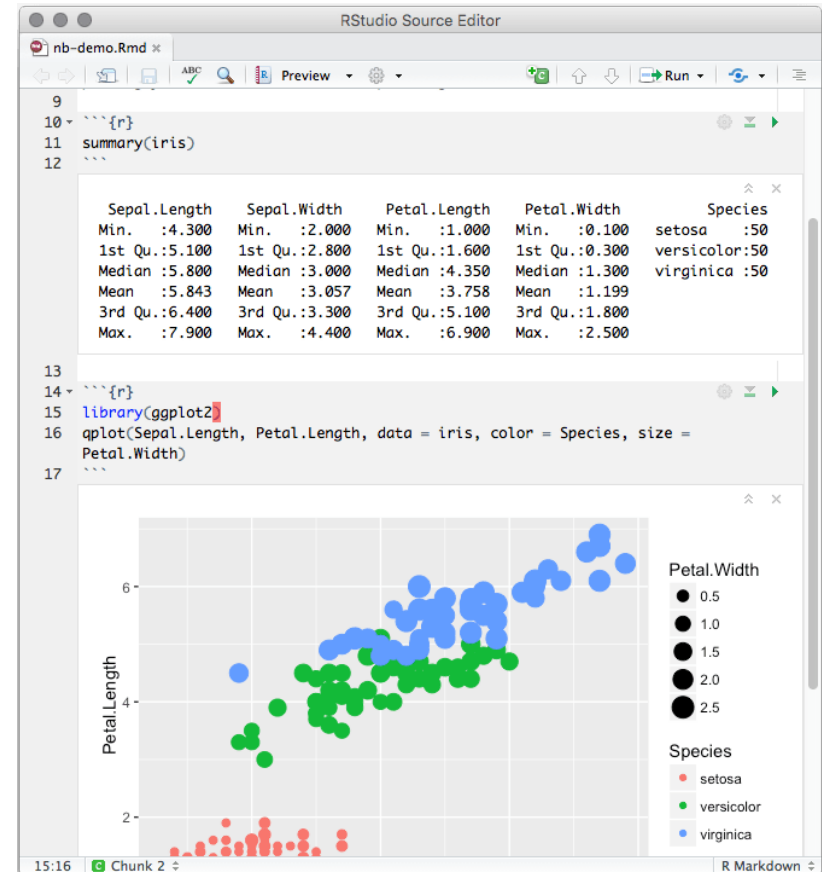
R Notebooks

Mentioned last week...

- These are HTML documents that enable the interleaving of text and chunks of executable code.
- File > New File > R Notebook

See:

<https://rmarkdown.rstudio.com/>



Source: <https://bookdown.org/>

Creating user-defined functions

It is possible to create named, user-defined, functions that can be saved between sessions using a script (see ATHR pp. 40 – 41).

Syntax:

```
> my_function <- function(arg1, arg2, ...) {  
>   object <- Calculations(arg1, arg2, ...)  
>   Return(object)  
> }
```

Creating user-defined functions

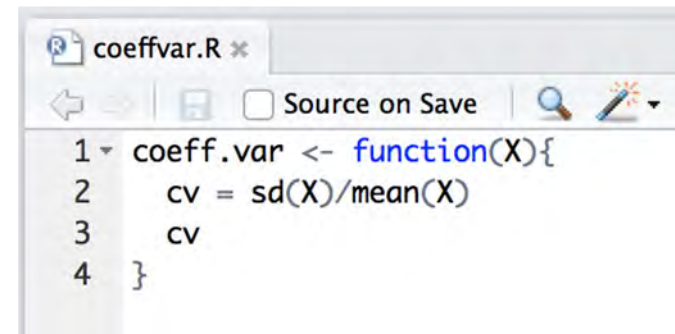
Example:

```
> coeff.var <- function(X){  
>   cv = sd(X)/mean(X)  
>   cv}  
  
> Y = c(1, 2, 3, 4, 5, 6)  
> coeff.var(Y)  
[1] 0.5345225
```

Saving and re-using functions

In Rstudio:

- Create a new R script,
- Write function in script editor,
- Save as (filename.R)

A screenshot of the RStudio script editor window. The title bar shows 'coeffvar.R'. The editor contains the following R code:

```
1 coeff.var <- function(X){  
2   cv = sd(X)/mean(X)  
3   cv  
4 }
```

To run function in a new session of R studio:

- Open and run script: code > source file (filename.R)

Assignment 1

Assignment 1: Summary

FIT3152 Data analytics – 2023: Assignment 1

Your task	<ul style="list-style-type: none">Analyse the country level predictors of pro-social behaviours to reduce the spread of COVID-19 during the early stages of the pandemic.This is an individual assignment.
Value	<ul style="list-style-type: none">This assignment is worth 25% of your total marks for the unit.It has 40 marks in total.
Suggested Length	<ul style="list-style-type: none">8 – 10 A4 pages (for your report) + extra pages as appendix (for your R script and clustering table).Font size 11 or 12pt, single spacing.
Due Date	11.55pm Monday 17th April 2023
Submission	<ul style="list-style-type: none">Submit a single PDF file and single video file on Moodle.Use the naming convention: <i>FirstnameSecondnameID.{pdf, mp4, mov etc.}</i>Turnitin will be used for similarity checking of all written submissions.
Generative AI Use	<ul style="list-style-type: none">In this assessment, you must not use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task.
Late Penalties	<ul style="list-style-type: none">10% (4 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

Instructions

Address each of the research questions below and report the results of your analysis and your interpretation of those results.

You are expected to include at least one high quality multivariate graphic summarising key results. You may also include other simpler graphs and tables. Report any assumptions you've made in modelling and 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.

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 Markup document that contains the R code with the discussion/text interleaved. Render this as an 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

Software

It is expected that you will use R for your data analysis and 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. You may use other software, such as Excel, to create the table of clustering data for Question 3(a)

Assignment 1: Questions:

Questions

During the early stages of the COVID-19 pandemic, researchers surveyed participants around the globe. A baseline study was conducted with the aim of identifying the most important predictors of pro-social COVID-19 behaviours, that is, actions that would reduce the spread of the virus. You can read a more detailed description of the research and results in Van Lissa (2022), see references.

The aim of this assignment is to understand country-level differences in predictors of pro-social behaviours, reported by participants as: “I am willing to:

- help others who suffer from coronavirus.” (c19ProSo01)
- make donations to help others that suffer from coronavirus.” (c19ProSo02)
- protect vulnerable groups from coronavirus even at my own expense.” (c19ProSo03)
- make personal sacrifices to prevent the spread of coronavirus.” (c19ProSo04)

Your task is to analyse the baseline survey data overall, with a focus on the country you have been assigned. You may make use of any additional data you require to answer the following questions.

Assignment 1: Questions 1 & 2

1. Descriptive analysis and pre-processing. (6 Marks)

- (a) Describe the data overall, including things such as dimension, data types, distribution of numerical attributes, variety of non-numerical (text) attributes, missing values, and anything else of interest or relevance.
- (b) Comment on any pre-processing or data manipulation required for the following analysis.

2. Focus country vs all other countries as a group. (12 Marks)

- (a) Identify your focus country from the accompanying list (FocusCountryByID.pdf). How do participant responses for your focus country differ from the other countries in the survey as a group?
- (b) How well do participant responses (attributes) predict pro-social attitudes (**c19ProSo01, 2, 3 and 4**) for your focus country? Which attributes seem to be the best predictors? Explain your reasoning.
- (c) Repeat Question 2(b) for the other countries as a group. Which attributes are the strongest predictors? How do these attributes compare to those of your focus country?

Assignment 1: Question 3

3. Focus country vs cluster of similar countries. (10 Marks)

(a) Using several social, economic, health, political or other indicators, identify between 3 and 7 countries (in the baseline data) that are similar to your focus country using clustering. Van Lissa (2022) refers to several indicators you might consider, among others. Some of these are listed in the references, but these are not exhaustive. State the indicators used and describe how you calculated/identified similar countries. Copy and paste the table of values you used for your clustering into your report as an Appendix.

(b) How well do participant responses predict pro-social attitudes (**c19ProSo01, 2, 3 and 4**) for this cluster of similar countries? Which attributes are the strongest predictors? How do these attributes compare to those of your focus country? Comment on the similarity and/or difference between your results for this question and Question 2(c). That is, does the group of all other countries 2(c), or the cluster of similar countries 3(b) give a better match to the important attributes for predicting pro-social attitudes in your focus country? Discuss.

Assignment 1: Presentation/Overall

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

Record a short presentation using your smart phone, 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 and any assumptions made. Pay particular emphasis to your results in Questions 2(c) and 3(b)

5 **Overall considerations (8 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, including at least one high-quality multivariate graphic; the quality of your R coding.

Assignment 1: Data generation

Data

The data for this assignment is a reduced version of that collected for the PsyCorona baseline study, Van Lissa et al. (2022). The filename is “PsyCoronaBaselineExtract.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) # XXXXXXXX = your student ID  
cvbase = read.csv("PsyCoronaBaselineExtract.csv")  
cvbase <- cvbase[sample(nrow(cvbase), 40000), ] # 40000 rows
```

Locate your focus country using the accompanying document FocusCountryByID.pdf.

Assignment 1: Selected references

References and web links

C. J. Van Lissa, et al., (2022) Using machine learning to identify important predictors of COVID-19 infection prevention behaviors during the early phase of the pandemic. Patterns 3, 100482.

<https://doi.org/10.1016/j.patter.2022.100482>

The World Bank Data Collections (and Governance Indicators)

<https://datacatalog.worldbank.org/collections>

<http://info.worldbank.org/governance/wgi/>

Organisation for Economic Co-operation and Development (OECD)Data

<https://data.oecd.org/>

Global Health Security Index: Reports and Data

<https://www.ghsindex.org/report-model/>

World Health Organization

<https://www.who.int/>

Assignment 1: Code Book extract

Data fields and brief descriptor (note AD = Agree/Disagree). See BaselineCodebookExtract for full description.)

Concept	Variable Name	Label
Affect	affAnx	How did you feel over the last week? - Anxious
	affCalm	...Calm
	affContent	...Content
	affBor	...Bored
	affEnerg	...Energetic
	affDepr	...Depressed
	affExc	...Excited
	affNerv	...Nervous
	affExh	...Exhausted
	affInsp	...Inspired
	affRel	...Relaxed
Likelihood	PLRAC19	How likely is it that... in the next few months? - You will get infected with coronavirus.
	PLRAEco	... Your personal situation will get worse due to economic consequences of coronavirus.

Data manipulation

Summarizing data by groups

Data grouped by factors:

- Applying a function to a single column
- Applying a function to a group of columns

Why do we need to do this?

- To simplify the data, making comparisons easier
- Reduce data complexity, enabling further analysis

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”

http://en.wikipedia.org/wiki/Iris_flower_data_set



www.shutterstock.com · 126112010

Print

```
> iris # = print(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
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
...					

Two challenges

(1) Easy!

- Create a table of column means grouped by species.

(2) Harder!

- Create a CSV file containing the correlation between sepal length and sepal width, and petal length and petal width for each species.

High level view

Data analysis is easier if you have a high-level view of the data:

- 4 columns + 1 factor (Species)
- Two pairs of related columns: sepals & petals

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Setosa
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Virginica
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Versicolor

Challenge 1. Function: aggregate

The ‘aggregate’ function creates a table by applying a function to data in individual columns grouped by a factor (or factors). To calculate averages:

- Note: columns referred to by their index [(number)] for compactness

```
> aggregate(iris[1:4], iris[5], mean)
```

	Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	setosa	5.01	3.43	1.46	0.246
2	versicolor	5.94	2.77	4.26	1.326
3	virginica	6.59	2.97	5.55	2.026

?aggregate



- Description

`aggregate(x, ...)` : Splits the data into subsets, computes summary statistics for each, and returns the result in a convenient form.

- Usage

`aggregate(x, by, FUN, ..., simplify = TRUE)`

- Arguments

X : An R object.

By : List of grouping elements

FUN : Function to compute the summary statistics

Simplify : Indicates whether results should be simplified to a vector or matrix if possible.

Challenge 2. Function: by

declare function

‘by’ enables a function to be applied across individual or multiple columns of a data frame grouped by a factor or factors.

- To calculate the correlation of sepal length and width

> by(iris, iris[5], function(df) cor(df\$Sepal.Length,
df\$Sepal.Width))

Species: setosa

[1] 0.743

Species: versicolor

[1] 0.526

Species: virginica

[1] 0.457

List

*df =
temp data
frame*

?by

- Description

Apply a Function to a Data Frame Split by Factors

- Usage

`by(data, INDICES, FUN, ..., simplify = TRUE)`

- Arguments

Data : an R object, normally a data frame, possibly a matrix.

INDICES : a factor or a list of factors, each of length `nrow(data)`.

FUN : a function to be applied to data frame subsets of data...

?by: applying the cor function

Looking more closely at the way correlation is calculated:

Data frame Column of factors Declaring a new anonymous function on the fly. Parameter is temporary data frame (df) created for each factor.

```
> by(iris, iris[5], function(df) cor(df$Sepal.Length,  
  df$Sepal.Width))
```

Values in temp data frame passed to cor function

Anonymous functions

If a function is only to be used once, it can be defined when it is used. These are anonymous functions (having no name) see ATHR p.41.

See previous slide for an example:

```
> by(iris, iris[5], function(df) cor(df$Sepal.Length,  
  df$Sepal.Width))
```

...

Changing earlier example to a more compact notation, using column indexes.

From:

```
> by(iris, iris[5], function(df) cor(df$Sepal.Length,  
  df$Sepal.Width))
```

To:

```
> by(iris, iris[5], function(df) cor(df[1], df[2]))
```

Function: as.table

This function converts the output format of a function from a list to a table

```
> as.table(by(iris, iris[5], function(df) cor(df[1], df[2])))
```

Species

setosa	versicolor	virginica
0.743	0.526	0.457

Function: as.data.frame

This function converts “coerces” the output of a table into a data frame

```
> Sepal.cor <- as.data.frame(as.table(by(iris, iris[5],  
function(df) cor(df[1], df[2]))))
```

```
> Sepal.cor
```

	Species	Freq
1	setosa	0.743
2	versicolor	0.526
3	virginica	0.457

always

Function: colnames

This function assigns new column names to a data frame.

```
> colnames(Sepal.cor) <- c("Species", "Sepal.cor")
```

```
> Sepal.cor
```

	Species	Sepal.cor
1	setosa	0.743
2	versicolor	0.526
3	virginica	0.457

Now for petals...

Repeating the previous code for petals...

```
> Petal.cor <- as.data.frame(as.table(by(iris, iris[5],  
  function(df) cor(df[3], df[4]))))  
> colnames(Petal.cor) <- c("Species", "Petal.cor")  
> Petal.cor
```

	Species	Petal.cor
1	setosa	0.332
2	versicolor	0.787
3	virginica	0.322

iris (row, col)

Merging data frames (and saving)

Using a common column – “Species” – and rounding data. *Note: we could have used cbind to combine data frames since they have same format.*

- > iris.cor <- merge(Sepal.cor, Petal.cor, by = "Species")
- > iris.cor[,2] = round(iris.cor[,2], digits = 3)
- > iris.cor[,3] = round(iris.cor[,3], digits = 3)
- > write.csv(iris.cor, file = "Iris.cor.csv",
row.names=FALSE)

| Rounding

Set		
Var		
Var		

The saved file

SepalPetalcor.csv

Species	Sepal.cor	Petal.cor
setosa	0.743	0.332
versicolor	0.526	0.787
virginica	0.457	0.322

This gives a much more compact presentation of the main correlations compared to the table created last lecture using:

```
> by(iris[1:4], factor(iris$Species), cor)
```



Correlation matrix – by... factor

From last week: Pairwise correlation by species

> by(iris[1:4], factor(iris\$Species), cor)

```
factor(iris$Species): setosa
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      1.0000000    0.7425467    0.2671758    0.2780984
Sepal.Width      0.7425467    1.0000000    0.1777000    0.2327520
Petal.Length      0.2671758    0.1777000    1.0000000    0.3316300
Petal.Width      0.2780984    0.2327520    0.3316300    1.0000000
-----
factor(iris$Species): versicolor
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      1.0000000    0.5259107    0.7540490    0.5464611
Sepal.Width      0.5259107    1.0000000    0.5605221    0.6639987
Petal.Length      0.7540490    0.5605221    1.0000000    0.7866681
Petal.Width      0.5464611    0.6639987    0.7866681    1.0000000
-----
factor(iris$Species): virginica
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length      1.0000000    0.4572278    0.8642247    0.2811077
Sepal.Width      0.4572278    1.0000000    0.4010446    0.5377280
Petal.Length      0.8642247    0.4010446    1.0000000    0.3221082
Petal.Width      0.2811077    0.5377280    0.3221082    1.0000000
```

dplyr



If some of the manipulation we've done so far looks a bit intimidating, you might want to try the 'dplyr' package. It:

- is a *Grammar of Data* manipulation,
- provides a consistent set of verbs to simplify the most common data manipulation challenges.
- <https://dplyr.tidyverse.org/>
- See Chapter 5, Data Transformation in R for Data Science <https://r4ds.had.co.nz/>

dplyr



dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges:

- `mutate()` adds new variables that are functions of existing variables
- `select()` picks variables based on their names.
- `filter()` picks cases based on their values.
- `summarise()` reduces multiple values down to a single summary.
- `arrange()` changes the ordering of the rows.

These all combine naturally with `group_by()` which allows you to perform any operation “by group”. You can learn more about them in `vignette("dplyr")`. As well as these single-table verbs, dplyr also provides a variety of two-table verbs, which you can learn about in `vignette("two-table")`.

From: <https://dplyr.tidyverse.org/>

dplyr



Quick start:

- Use pipes, `%>%`, to connect data to a grouping variable, and then apply a function.
- For example, to find average sepal length by species::

```
> iris %>% group_by(Species) %>%  
summarise(Ave.Sepal.len = mean(Sepal.Length))
```

```
# A tibble: 3 × 2  
  Species Ave.Sepal.len  
  <fct>      <dbl>  
1 setosa      5.01  
2 versicolor  5.94  
3 virginica   6.59
```

dplyr



Tibbles...

- Dplyr creates tibbles instead of data frames. To get an overview of the difference between these, see:
- <https://r4ds.had.co.nz/tibbles.html>
- You can convert a tibble to a data frame if preferred using:
 - > `NewDataFrame = as.data.frame(TibbleName)`

dplyr (Challenge 1)



- For Challenge 1, find column means by species:
> iris %>% group_by(Species) %>% summarise(ASL = mean(Sepal.Length), ASW = mean(Sepal.Width), APL = mean(Petal.Length), APW = mean(Petal.Width))

```
# A tibble: 3 × 5
  Species      ASL      ASW      APL      APW
  <fct>      <dbl> <dbl> <dbl> <dbl>
1 setosa      5.01   3.43   1.46  0.246
2 versicolor  5.94   2.77   4.26  1.33
3 virginica   6.59   2.97   5.55  2.03
```


dplyr (Challenge 2)



- For Challenge 2, find the correlation between sepal length and width, and petal length and width by species – first step is shown below:
> iris %>% group_by(Species) %>% summarise(Sepal.cor = cor(Sepal.Length, Sepal.Width))

```
# A tibble: 3 × 2
  Species      Sepal.cor
  <fct>        <dbl>
1 setosa      0.743
2 versicolor 0.526
3 virginica   0.457
```

Two more challenges

(3) Easy!

- Examine the difference between the aspect ratios (Length / Width) for sepals and petals between the different species.

(4) Harder!

- Report the data for the flower having the longest petal in each species.

Challenge 3: Add/remove columns

By default, R will add a new column to a data frame if the output of a column operation is specified as a new column.

Alternatively, the `cbind` function can be used to append a vector or data frame by columns.

This lets us store the results of row operations, including factor generation.

Making new columns

Add two columns containing the aspect ratio (length/width) for sepals and petals:

- > `niris <- iris # creating a new data frame`
- > `niris$Sepal.ar <- niris$Sepal.Length/niris$Sepal.Width`
add new column
- > `niris$Petal.ar <- niris$Petal.Length/niris$Petal.Width`
add new column
- > `head(niris)`

The augmented data frame: niris

```
> head(niris)
```

new
cols added

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Sepal.ar	Petal.ar
1	5.1	3.5	1.4	0.2	setosa	1.46	7.00
2	4.9	3.0	1.4	0.2	setosa	1.63	7.00
3	4.7	3.2	1.3	0.2	setosa	1.47	6.50
4	4.6	3.1	1.5	0.2	setosa	1.48	7.50
5	5.0	3.6	1.4	0.2	setosa	1.39	7.00
6	5.4	3.9	1.7	0.4	setosa	1.38	4.25

Deleting columns

This is easy – but cannot be undone!

To remove a single column, do it by name.

To remove the first column:

```
> niris$Sepal.Length <- NULL
```

Tedious for multiple columns. A quicker but potentially dangerous way to remove first 4 columns:

```
> niris <- niris[,c(5:7)] # reassign cols 5:7 on to self!
```

After removing columns:

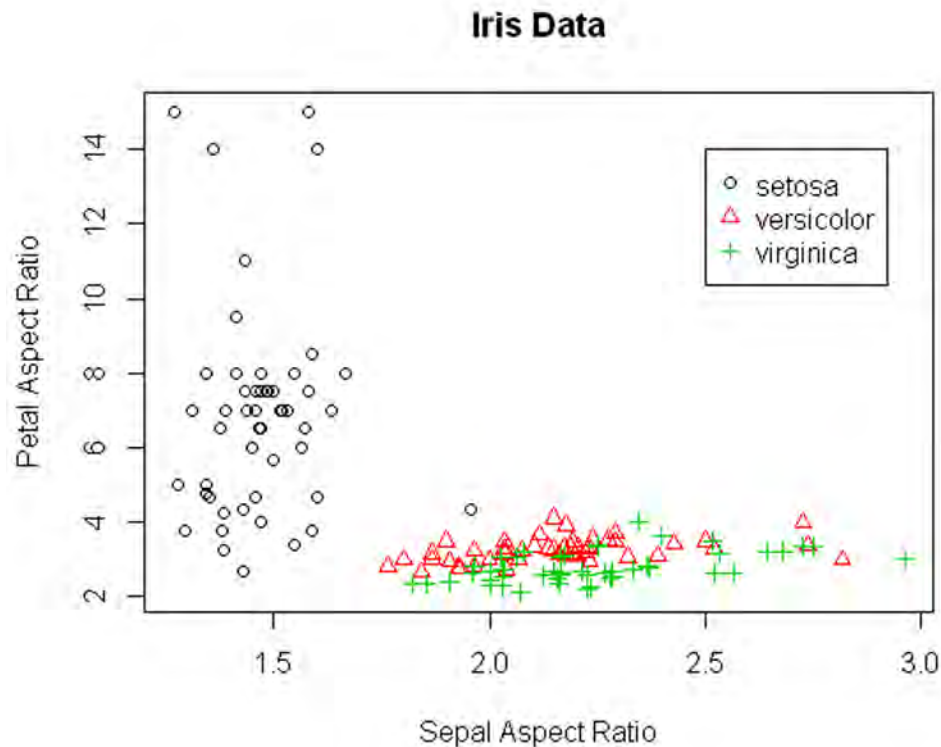
```
> head(niris)
```

	Species	Sepal.ar	Petal.ar
1	setosa	1.46	7.00
2	setosa	1.63	7.00
3	setosa	1.47	6.50
4	setosa	1.48	7.50
5	setosa	1.39	7.00
6	setosa	1.38	4.25

Scatterplot



Petal vs Sepal aspect ratio (Length / Width)





Code for scatterplot on previous slide:

- > with(niris, plot(Sepal.ar, Petal.ar, col = Species, pch=as.numeric(Species), main = ("Iris Data"), xlab = "Sepal Aspect Ratio", ylab = ("Petal Aspect Ratio")))
- > with(niris, legend(2.5, 14, as.vector(unique(Species)), pch=unique(Species), col = unique(Species)))

Challenge 4: using dplyr



This task shows off how useful dplyr is:

To find the flower having the longest petal in each species:

```
> iris %>% group_by(Species) %>% top_n(1, Petal.Length)
```

```
# A tibble: 4 × 5
```

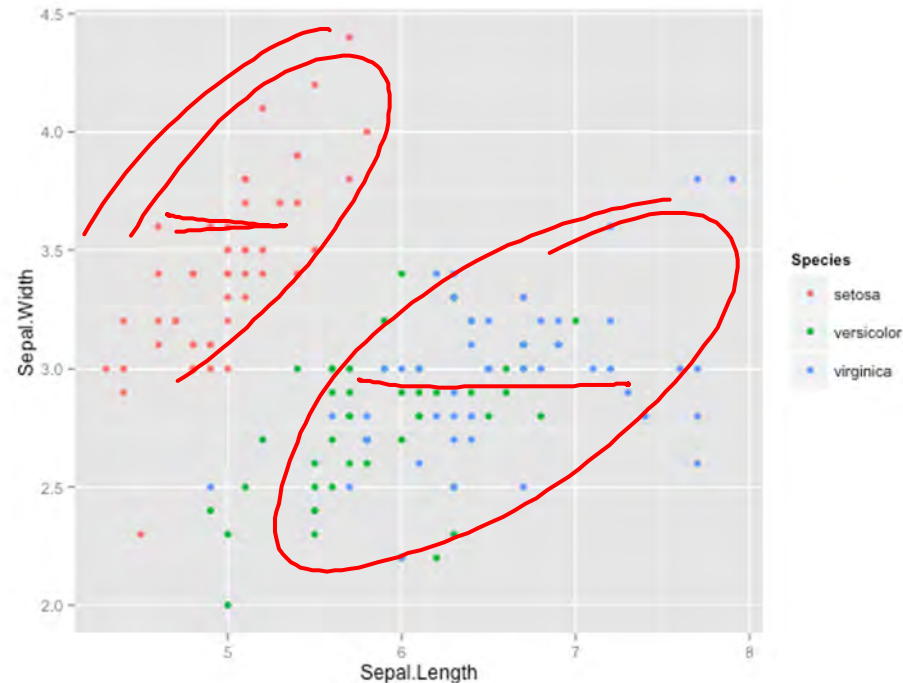
```
# Groups:   Species [3]
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
1	4.8	3.4	1.9	0.2	setosa
2	5.1	3.8	1.9	0.4	setosa
3	6	2.7	5.1	1.6	versicolor
4	7.7	2.6	6.9	2.3	virginica

Results show two I.setosa flowers having equally long petals.

Challenge 5: recoding and indexing

Does Iris setosa have an average sepal width greater than I.versicolor and virginica combined?



Challenge 5: recoding

factor

To compare *I.setosa* against the other two species, we need to create a new index as a column that groups *I.versicolor* and *virginica*.

- Note: use the function “recode” from the “car” package
 - > `niris = iris # clone iris data`
 - > `install.packages("car")`
 - > `library(car)`

Challenge 5: creation of new factor

```
> ...  
> niris$vvs = recode(niris$Species," 'versicolor' =  
  '0';'virginica' = '0';'setosa' = '1' ")  
> print(niris[c(1,51,101),]) # as a check
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	vvs
1	5.1	3.5	1.4	0.2	setosa	1
51	7.0	3.2	4.7	1.4	versicolor	0
101	6.3	3.3	6.0	2.5	virginica	0

Challenge 5: t-Test

```
> ... test group  
> t.test(niris$Sepal.Width~niris$vvs, alternative = "less")
```

```
Welch Two Sample t-test  
data:  niris$Sepal.Width by niris$vvs  
t = -8.8121, df = 87.596, p-value = 5.177e-14  
alternative hypothesis: true difference in means is less than 0  
95 percent confidence interval:  
-Inf -0.451108  
sample estimates:  
mean in group 0 mean in group 1  
2.872 3.428
```

Challenge 5: Data frames as subsets

Or we could have made two new data frames from the original iris data, one for I.setosa, and one combining I.versicolor and I.virginica.

- Note: use of logical operators “==” (is equal to), and “%in%” (is contained in)...

new data frame

```
> iris.set = iris[iris$Species == "setosa",]  
> iris.ver.vir = iris[(iris$Species %in%  
  c("virginica", "versicolor")),]
```


Summary

Summarizing data using factors using

- Base functions: aggregate, by
- dplyr package functions: group_by, summarise.

Creating and removing columns

Searching, indexing and combining rows

- dplyr functions: group_by, top_n.
- Base functions: as.table, as.data.frame, colnames, rbind, cbind, logical operators “==” and “%in%”.

Answers to the review questions

1. D: Hierarchy (Phylogenetic tree)
2. A: Time Series (Daily infections)
3. E: Network (Transmission network)
4. B: Statistical (Scatter (Bubble) plot)

References

Books – online from the Monash Library

- Spector, P., Data manipulation with R.
- Wickham, H., ggplot2: elegant graphics for data analysis.

dplyr Cheat Sheet <https://github.com/rstudio/cheatsheets>

R Reference card (Tom Short) available from contributed documentation on CRAN site.

<http://cran.r-project.org/>