Lecture 4

Week 2, Nov 5th 2020

Introduction

- rsakai@ic.ac.uk
- PhD in Data Visualisation in Bioinformatics

Lecture 4

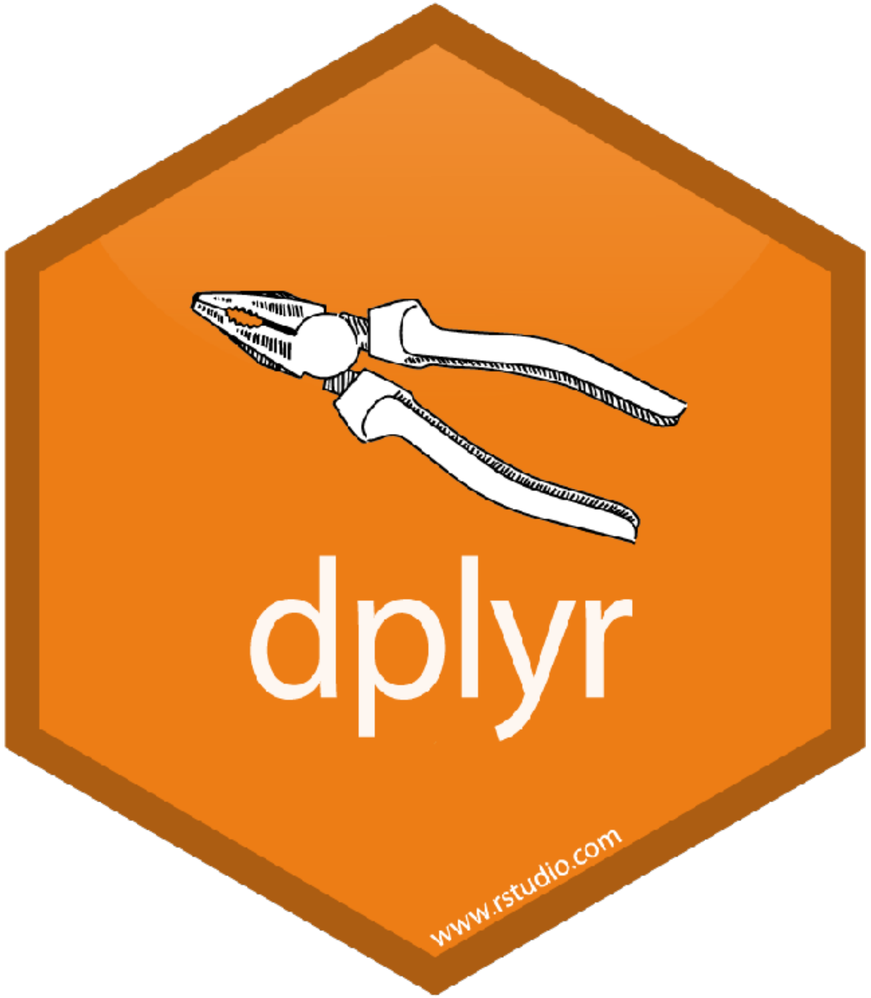
- Introduction to practical sessions
- Group assignment
- Introduction to R Markdown
- Introduction to visual analytics in R
- Introduction to tidyverse

Lecture 7

- Introduction to ggplot2
  - geometric objects
  - layered grammar of graphics
  - statistical transformation
  - position adjustment

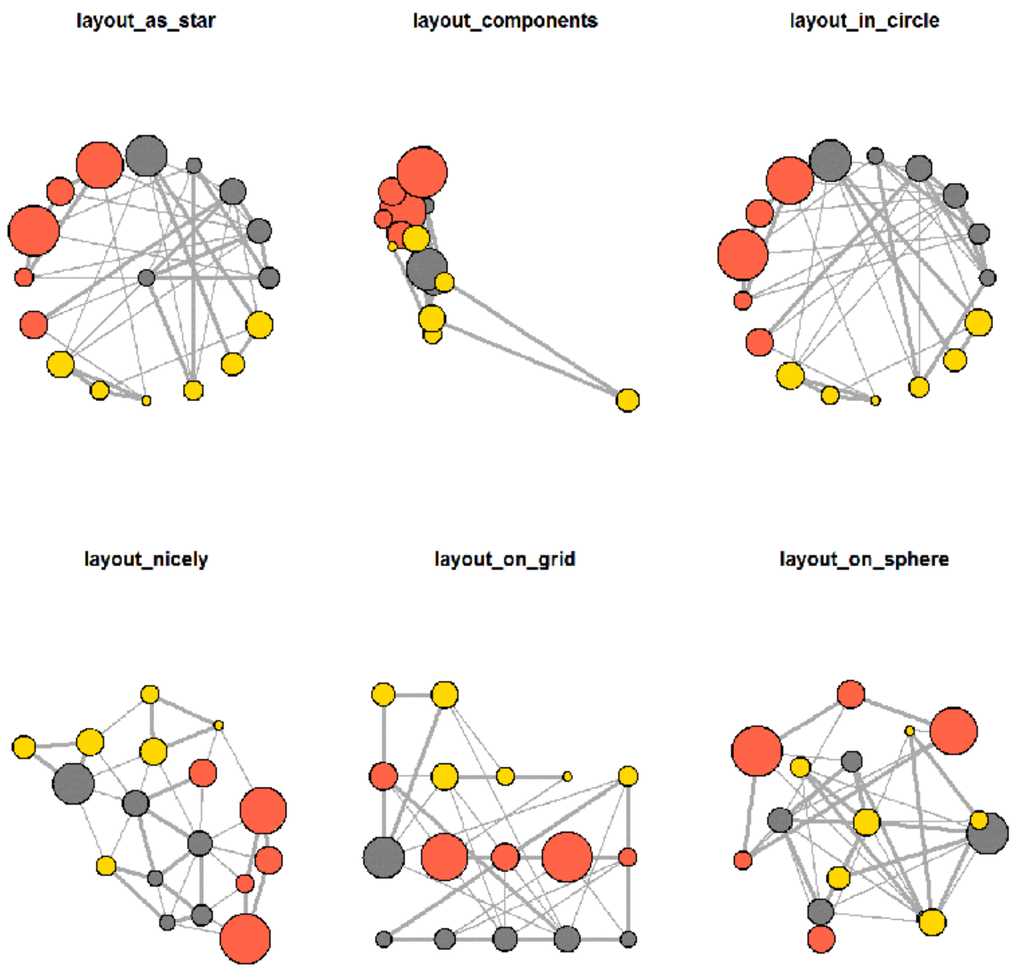
Lecture 8

- Data transformation
  - `dplyr` package
- Presentation of graphs
  - customising outputs to communicate
  - Importing data
  - Exporting images
  - Labels
  - Scales



Lecture 9

- Visualisation techniques
  - Choropleth
  - Other techniques
- Network visualisation



Lecture 10

- Interactive Data Visualization and Beyond
  - Plotly
  - Shiny
  - Other topics

- Python
  - Altair

Setting the expectations

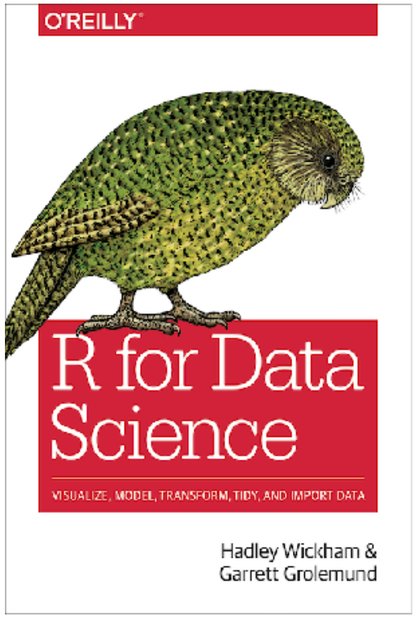
- Introduction to visual analytics in R
- Ability to:
  - apply learning/theories from lectures
  - apply baic data transformations
  - create static data visualisation to understand the data
  - refine visual outputs for communication
  - use visualisation techniques (e.g. interactive) for visual analytics
- Vigilant and resourceful analysts in R
  - developping metal models for visual analytics
  - You will have access to ggplot2 and dplyr cheatsheets

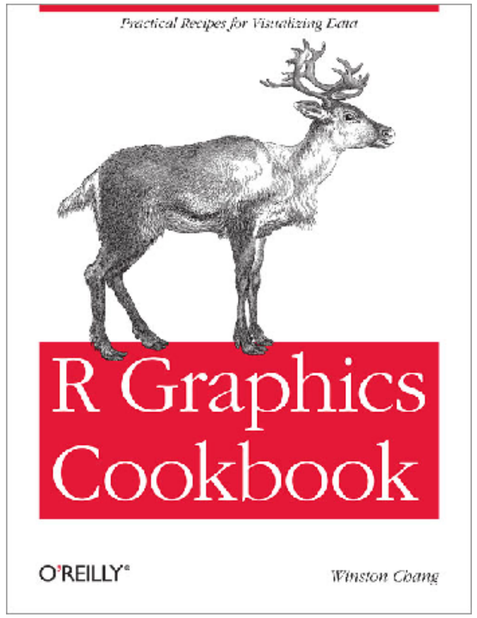
Practical sessions

- Lecture structure:
  - 45 min - lecture
  - 5 min - break
  - 45 min - lecture
  - 5 min - break
  - 20 min - in-class exercise / breakout sessions
- In-clsass excercise
  - Each in-class exercise counts for 2% towards your overall grade
  - Deadline by midnight of the day of lecture
  - Exercises are made available on the Hub, prior to each lecture
  - Late assignment will receive 0% unless there are exception circumstances and/or sufficient prior notice.
  - Knit the RMarkdown and hand in the HTML output.

Recommended readings

- Hadley Wickham and Garrett Grolemund, R for Data Science: Import, Tidy, Transform, Visualize, and Model Data (Sebastopol, California: O'Reilly Media, 2017), http://r4ds.had.co.nz/
- Winston Chang, R Graphics Cookbook 2e. O'Reilly. https://r-graphics.org





Your turn!

Poll: How familiar are you with R?

What is your programming language of choice?

Module Assignment

BUSI97273

Introduction

- Take the Airbnb data from Inside Airbnb, explore it, analyse it, and tell a nuanced story about it using visualisation.
  - The website curates and provides publicly available information about Airbnb's listings in cities around the world.
- Work in a team of 3 - 5 (already assigned)
  - We would like this project to be as useful for you and your future career as possible
  - You will hopefully want to show off your final project in a portfolio or during job interviews
- Task: each team will compile and submit:
  1. An executive summary of your findings and key messages (250 words, 3 figures max)
  2. A recorded presentation detailing your analysis process and justifications for visualisation design (video of 8~10 mins, but no longer than 10 mins)
  3. An HTML output of R Markdown/Jupyter notebook to show the process and how to reproduce your key figures from the downloaded data.

Instructions

- Download the dataset
  - http://insideairbnb.com/get-the-data.html
  - The data set is large and will probably not open well in Excel, so you will need to load the CSV file into R or another platform of your choice (e.g. Python). You can use Tableau, too.
  - It is up to your team to decide which datasets to use (choice of city/cities, and calendar/reviews/listings data)
- Data wrangling and transformation
  - Keep a record of your process so that the analysis is reproducible
- Find a story
  - Explore the story and make sure it is true and insightful
  - You can make a story around a business argument. It is up to you to decide your audience.
- Create visualisation and refine
  - You must present at least 3 different chart types (i.e. don't just make 3 scatter plots) in executive summary and markdown deliverables
  - Consider Why, How, and What of data visualisation from the lecture

Deliverable 1: Executive summary

- The goal of an executive summary is to **communicate** the key findings concisely with visualisation
- 250 words and 3 static figures (max)
  - Make sure each figure has a title and axes are well-labeled
  - You may add a figure legend for each figure. The text in figure legend is not included in the word count limit (250 words), but be concise.
- Include your group ID in the header and student names in the footer
- File naming convention: BUSI97273Visualisation_Group#-Summary.pdf

Deliverable 2: Recorded presentation

- The goal of a recorded presentation is to **explain** your analysis process
  - Explain how you used visualisation to find new insights. Tell us an interesting story.
  - Explain why, how, and what of data visualisation for your key selected figures
  - Explain how you refined selected visualisations for communication. We like to see the process.
  - Justify your design choices where applicable.
  - If interactive visualisation is used, explain how interactivity was used in analysis
- 8~10 min presentation (upload recording to YouTube or Vimeo)
  - You may choose to be in the video in person or just do a voice over against a series of presented visuals, or a slide deck that you run through with narration. Or, you could record a zoom call and trim the video.
  - Each member should appear across the recording.
  - You may do a recording from camera phones, it doesn't need to have high production value. Just make sure the key parts of your presentation are visible to the eye and audible to the ear.
- Provide the URL and name the title as BUSI97273Visualisation_Group#-Presentation
  - Please also provide the PDF or powerpoint slides used for the presentation. This is for the reviewer who is marking, and you will be evaluated based on what is in the video.

Deliverable 3: R Markdown

- The goal of an R Markdown is **reproducibility***
  - Select up to 3 key visualisations to show the process of createing the representation from raw data to visual outputs.
  - Create an R Markdown document with HTML template and explain how you produced key visualisations. Alternatively, you may use Jupyter notebook and export the HTML output.
  - Make sure to comment on each step, so other analysts (or future self) can understand what each step does
  - You only need to submit the output HTML file.
  - You are encouraged to include exploratory and intermediate data visualisations that are used in your analysis leading to your key selected figures. Make sure you capture your insights and thoughts in documentation.
  - File naming convention: BUSI97273Visualisation_Group#-KeyFigure.html

*Reproducibility is a key objective but not the only objective of this deliverable.

Submission details

- Deliverables (as described in previous slides) and file naming convention:
  1. An executive summary (BUSI97273Visualisation_Group#-Summary.pdf)
  2. A recorded presentation and slides (PDF or pptx). Provide the URL and name the title as BUSI97273Visualisation_Group#-Presentation
  3. An R Markdown file (BUSI97273Visualisation_Group#-KeyFigure.html)
- Upload your files to the Hub by Monday, 7th December no later than 24:00
  - Marking is intended to be completed and grades returned by Thursday, 26th February.
- This assignment contributes 25% towards your overall module grade
  - Late assignment will receive 0% unless there are exception circumstances and/or sufficient prior notice.

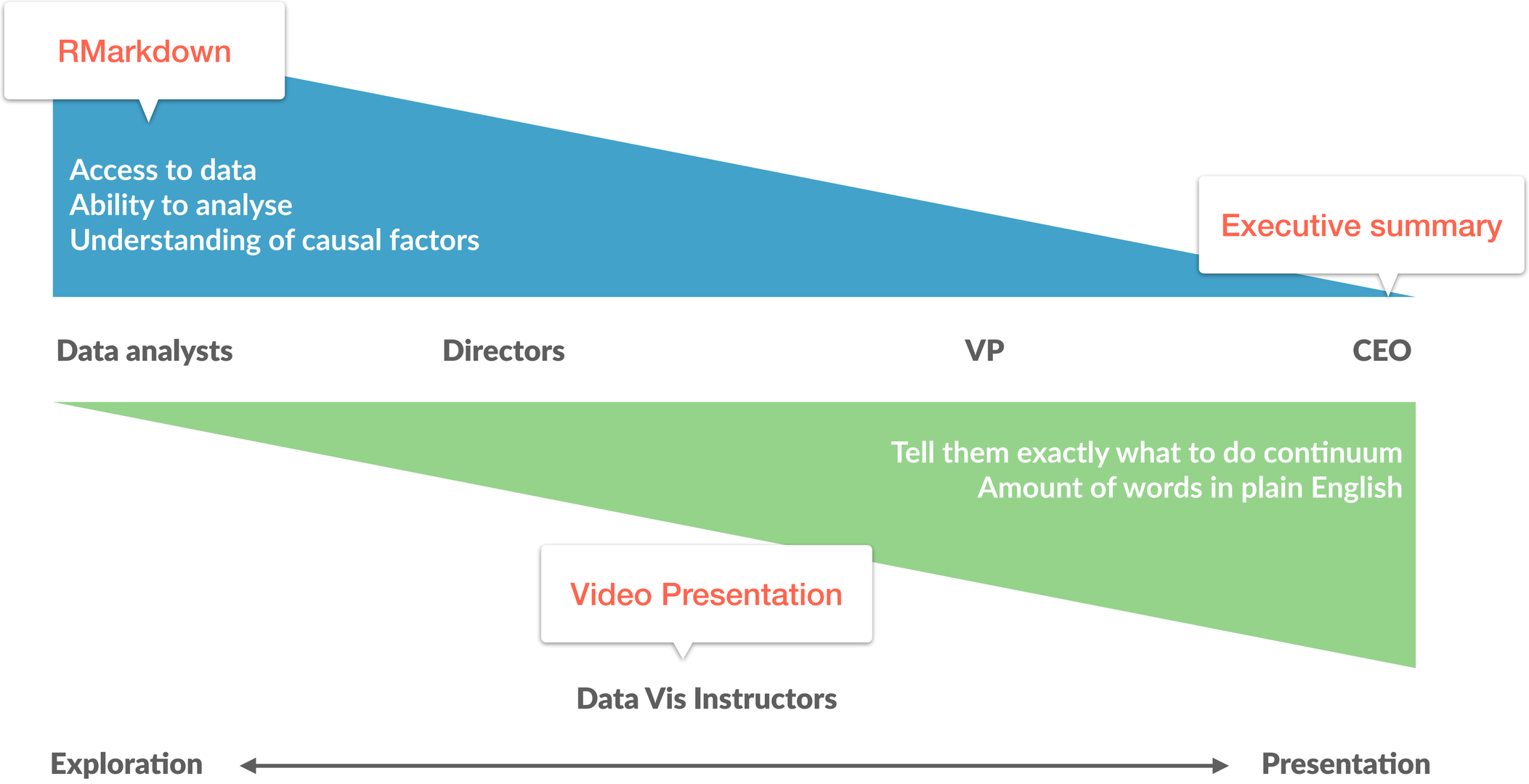
Useful information

- Grading
  - Presentation 50%
  - Executive summary 25%
  - R markdown/Jupyter notebook 25%
- Example workflow
  - Discuss/brainstorm questions you may want to address with the data
  - Perform exploratory data analysis to summarise the main characteristics of the datasets of your choice. You may use a dataset from a city or compare between multiple cities.
  - Evaluate if additional data or data transformation is required to address your questions
  - Visualise the data and try different visualisation techniques and encodings first, then evaluate which visualisation works best
  - Refine the key visualisation for communication and story telling

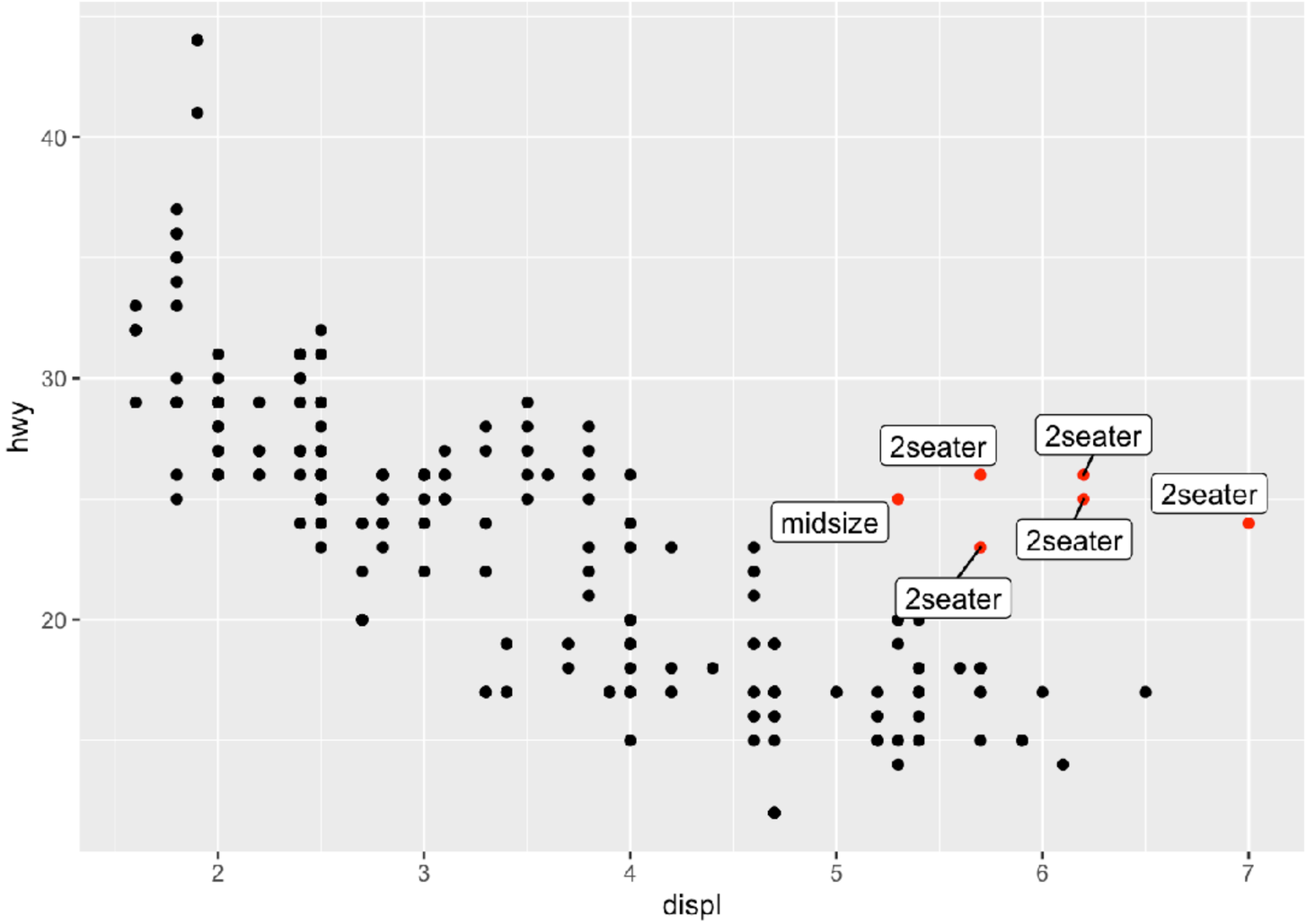
Useful information

- Tips
  - Clearly define your target audience. Who is your analysis for? How does he/she benefit from your insights?
  - Use colour consistently, (and perhaps sparingly)
    - Presentation that has a cohesive look-and-feel
    - Use the consistent colour palettes
  - Use features of Powerpoint to engage the audience
    - Consider using transition to draw the auidence's attention (e.g. preattentive attributes)
    - You can use Powerpoint to draw legends, labels, and annotations on top of graph

designed for whom?



Storytelling



R Markdown

unified authoring framework for data science

R Markdown

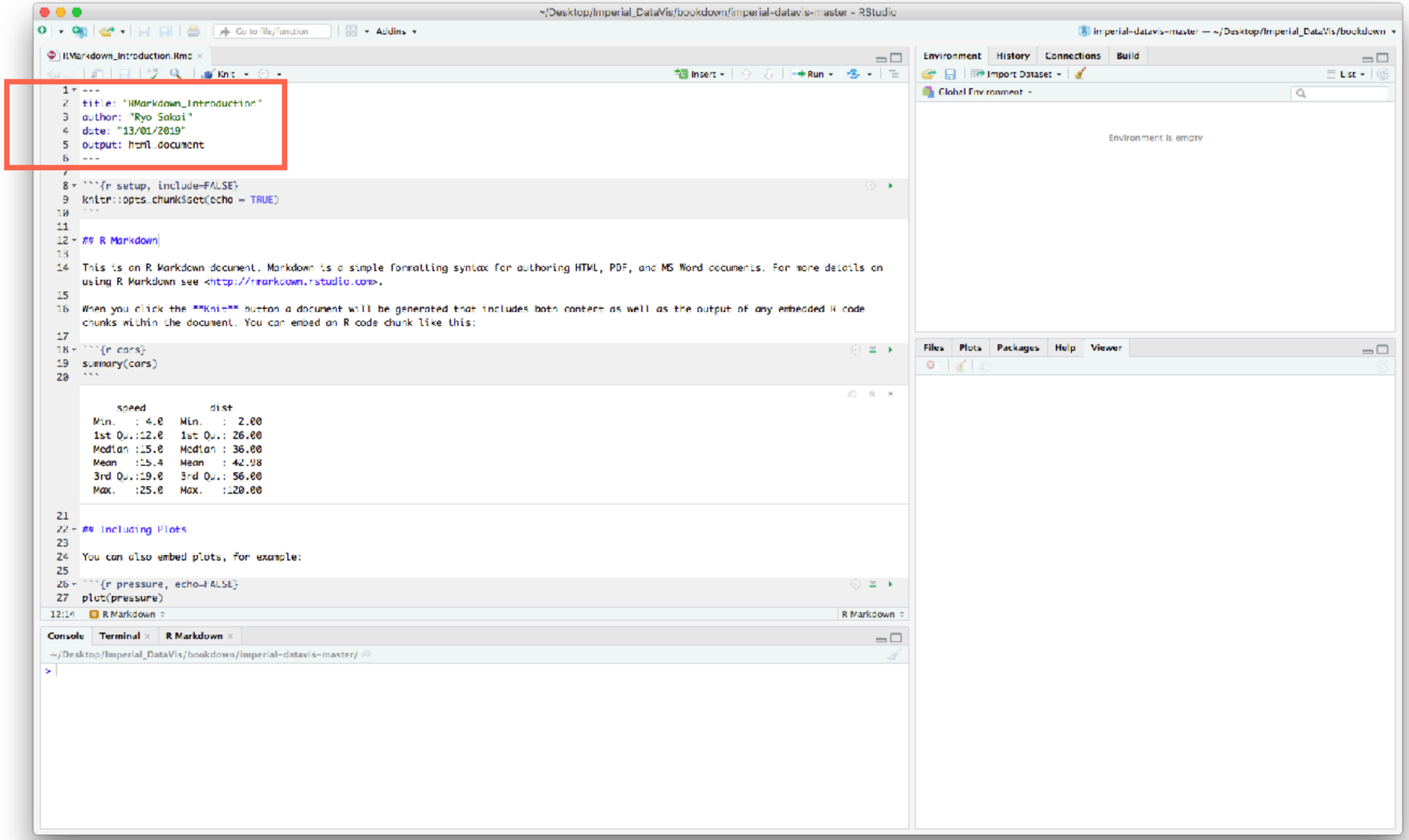
- an unified framework for data science:
  - combining your codes, its results (visualisation) and your prose commentary
  - designed for communication, collaboration and doing data science
- comes with RStudio, otherwise with **rmarkdown** package
- notebook interface, R Notebook
- outputs formats: HTML, PDF, Word, eBook, blog etc…

R Markdown

YAML header

settings for the whole document

"YAML Ain't Markup Language"



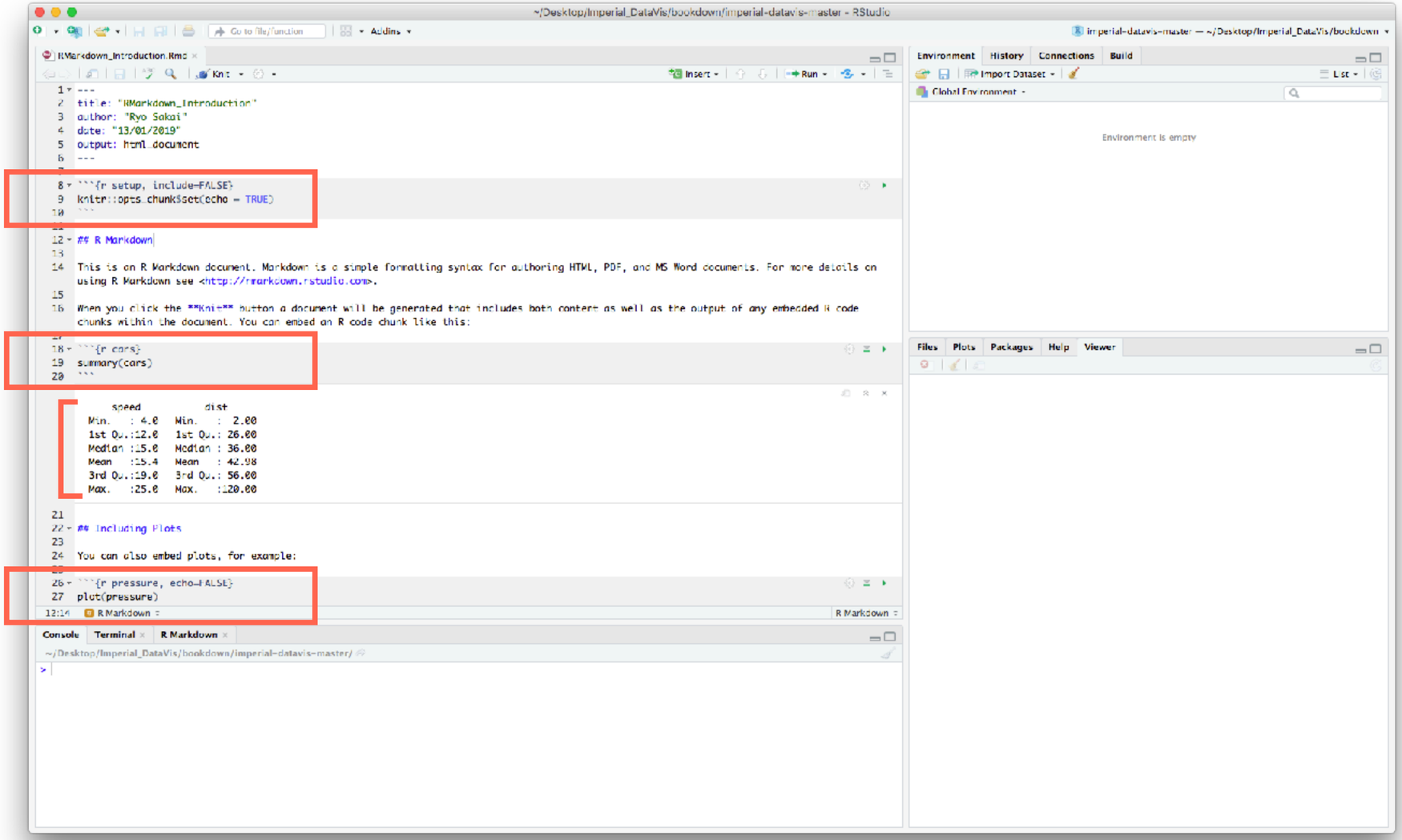
R Markdown

R code chunk

R code chunk

R code output

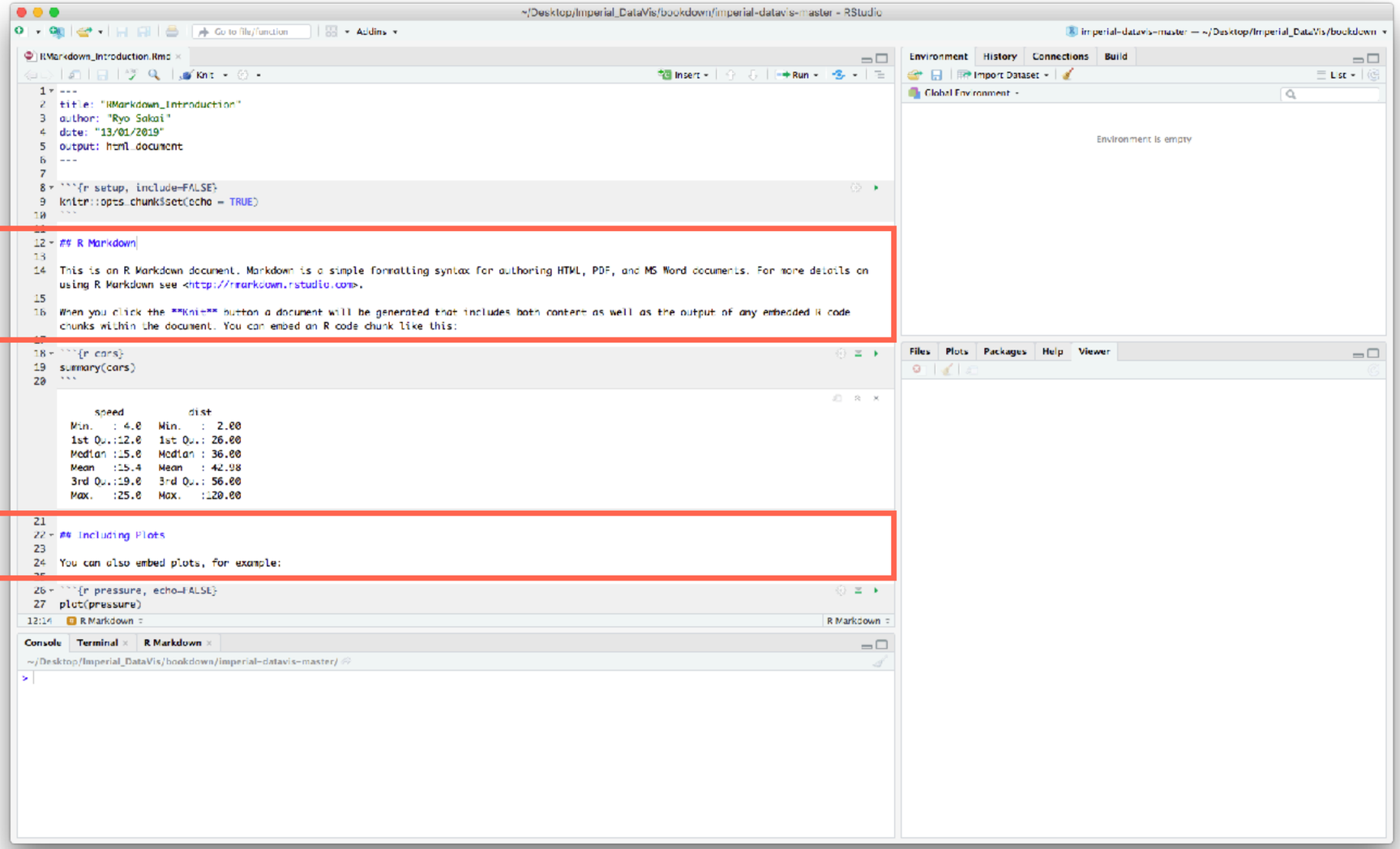
R code chunk



R Markdown

Markdown

Your notes/prose commentary



Your turn!

1. Create a new R Markdown document with ***File>New File>R Markdown…***
    1. Knit by clicking the "knit button" 

    2. Knit by keyboard short cut: **Cmd/Ctrl + Shift + K**
    3. Practice the keyboard shortcut to insert a code chunk: **Cmd/Ctrl + Alt + I**
    4. Execute a code chuch with the keyboard shortcut: **Cmd/Ctrl + Shift + Enter**
2. Create new R Markdown files to test 3 built-in formats:
    1. HTML
    2. PDF
    3. WORD
3. Optional further reading
    1. https://r4ds.had.co.nz/r-markdown.html

Text format with Markdown

- a lightweight set of conventions for formatting plain text files
- Go to ***Help > Markdown Quick Reference*** on RStudio

```
Text formatting
------------------------------------------------------------

*italic*  or _italic_
**bold**   __bold__
`code`
superscript^2^ and subscript~2~

Headings
------------------------------------------------------------

# 1st Level Header

## 2nd Level Header

### 3rd Level Header

Lists
------------------------------------------------------------

*   Bulleted list item 1
```

Code chunks

- Shortcut to insert code chunk: **Cmd/Ctrl + Alt + I**
- Codes are surrounded by `` ```{r} `` and `` ``` ``
- Shortcut to run code chunk: **Cmd/Ctrl + Shift + Enter**
- Options to set chunk name by `` ```{r by-name} ``
- Chunk options:

Option	Run code	Show code	Output	Plots	Message	Warnings
`eval = FALSE`	X		X	X	X	X
`include = FALSE`		X	X	X	X	X
`echo = FALSE`		X				
`results = “hide”`			X			
`fig.show = “hide”`				X		
`message = FALSE`					X	
`warning = FALSE`						X

Your turn!

1. Type the following code and knit your R Markdown.

```
```{r setup, include = FALSE}
library(ggplot2)
library(dplyr)

smaller <- diamonds %>%
  filter(carat <= 2.5)
```

We have data about `r nrow(diamonds)` diamonds. Only `r nrow(diamonds) -
nrow(smaller)` are larger than 2.5 carats. The distribution of the remainder is
shown below:

```{r, echo = FALSE}
smaller %>%
  ggplot(aes(carat)) +
  geom_freqpoly(binwidth = 0.01)
```
```

Table output

- **knitr::kable** output nice tables
- For more functionalities and styling options, check:
  - **kableExtra** package
  - https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html

```r
library(knitr)
library(kableExtra)

dt <- mtcars[1:5, 1:6]
kable(dt, caption = "Motor Trend Car Road Tests - mtcars") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

Motor Trend Car Road Tests - mtcars

	mpg	cyl	disp	hp	drat	wt
Mazda RX4	21.0	6	160	110	3.90	2.620
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875
Datsun 710	22.8	4	108	93	3.85	2.320
Hornet 4 Drive	21.4	6	258	110	3.08	3.215
Hornet Sportabout	18.7	8	360	175	3.15	3.440

More information

- R Markdown cheat sheet [link]
- R Markdown: The definitive guide by Yihui Xie [link]
- RStudio website [link]

Introduction to `ggplot2`

`ggplot2`

- R has several systems for making graphs (Base R, lattice, etc…)
- **`ggplot2`** is perhaps the most elegant and versatile
- Implemented based on the *grammar of graphics*
- part of **`tidyverse`** package

`tidyverse` package

- a collection of R packages designed for data science

```
# Install if you have not install, or to update the package
install.packages("tidyverse")

# Load the package
library(tidyverse)

# We will use this dataset for this section
mpg
```

Your turn!

1. Look up the documentation on **`mpg`** dataset
   - **`mpg`** is a dataset that comes with ggplot2 package
     - Make sure you load the ggplot2 library before looking up
     - You can run either **`library(ggplot2)`**, or **`library(tidyverse)`**

Solution

```
# Check the dataset:
?mpg
```

Solution

Imperial_oncampus_2020 - RStudio

Lecture_4.R

```r
library(tidyverse)

# Look up documentation on mpg
?mpg
```

Console

```
> ?mpg
>
>
>
>
>
>
>
>
```

R: Fuel economy data from 1999 to 2008 for 38 popular models of...

mpg {ggplot2} R Documentation

Fuel economy data from 1999 to 2008 for 38 popular models of cars

Description

This dataset contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

Usage

```
mpg
```

Format

A data frame with 234 rows and 11 variables:

manufacturer

manufacturer name

model

model name

displ

engine displacement, in litres

year

year of manufacture

`mpg` dataset

- *Do cars with big engines use more fuel than cars with small engines?*
  - What do you think the relationship between engine size and fuel efficiency is like?
  - Is it a linear or non-linear function?
- Your turn!
  - Look up which variable is for car's engine size in **`mpg`** dataset.
  - Look up which variable is for car's fuel efficiency in **`mpg`** dataset.

`displ` variable for car's engine size in

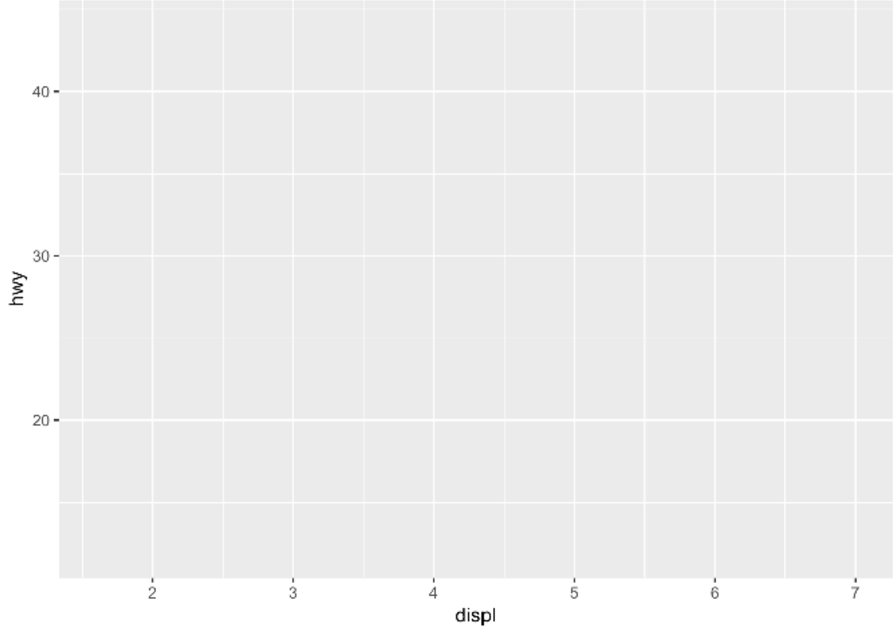
`hwy` variable for car's fuel efficiency on the highway, in miles per gallon

`ggplot2::qplot()`

- equivalent to the base `plot()` function
- a convenient wrapper for creating a number of different types of plots using a consistent calling scheme

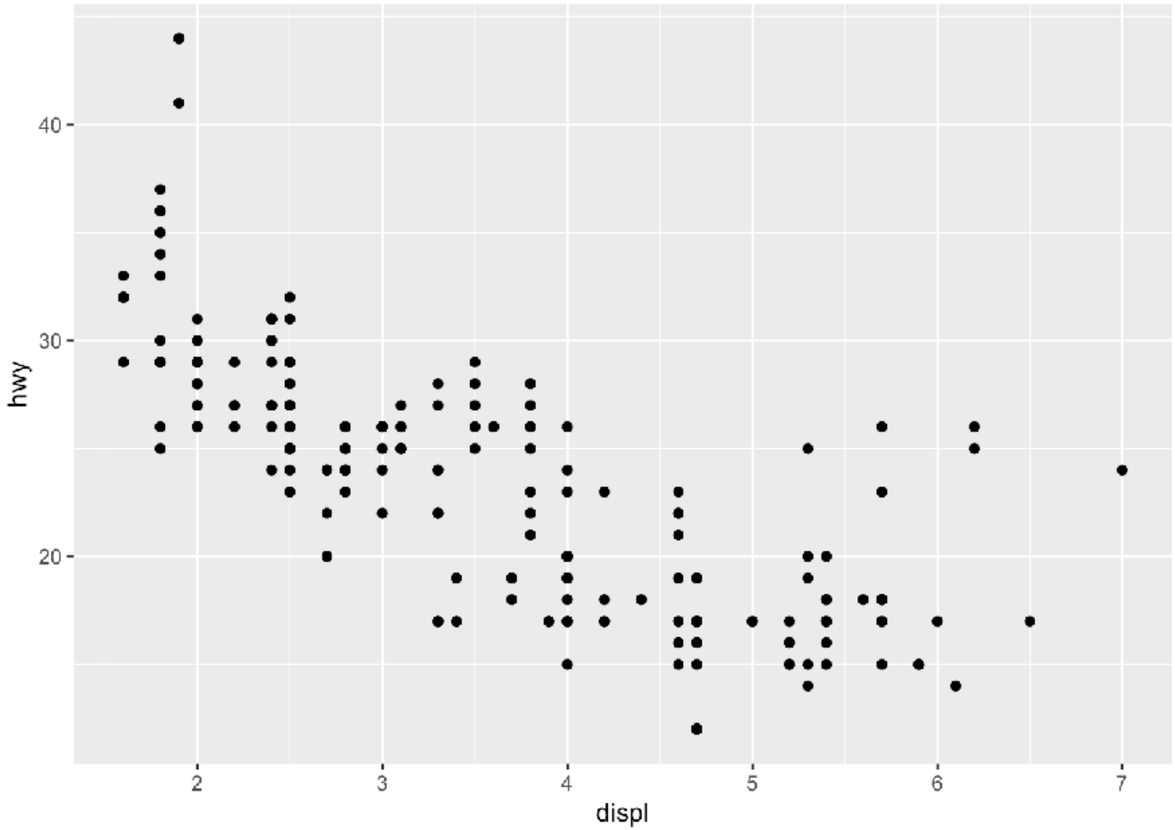
Canvas

```r
# left
ggplot(data = mpg)

# right
ggplot(data = mpg, mapping = aes(x = displ, y = hwy))
```

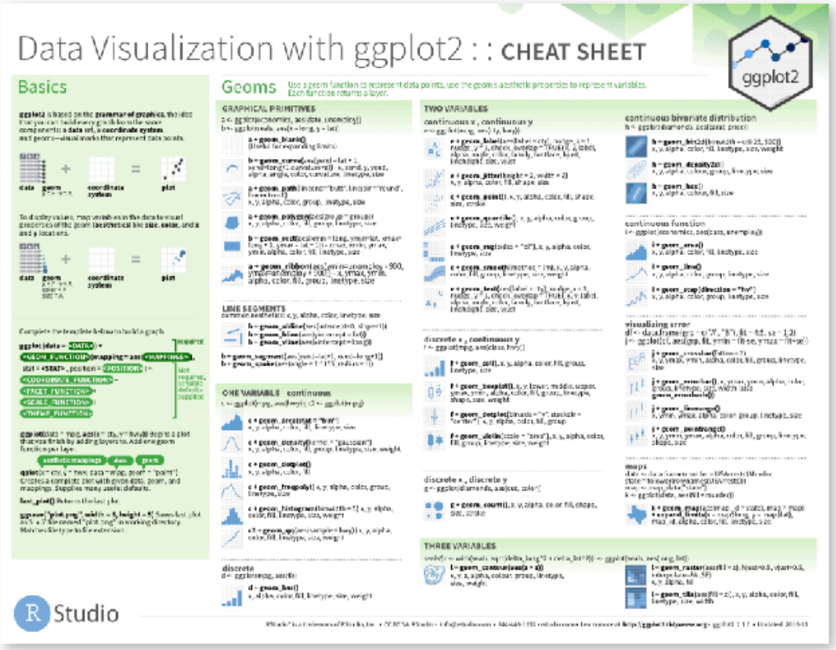
First scatter plot

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point()
```



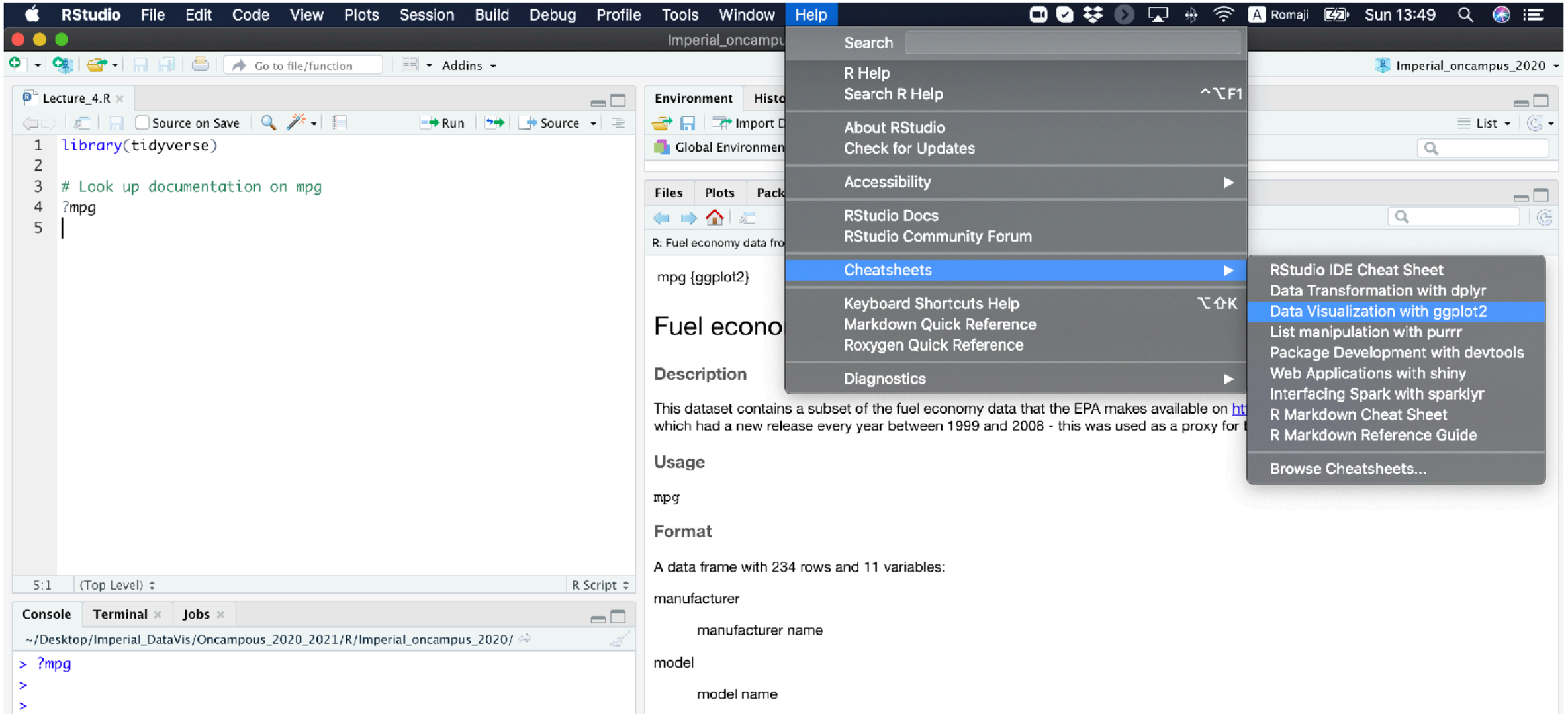
Basic template

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```



ggplot2 cheat sheet

Cheatsheets on RStudio



ggplot object

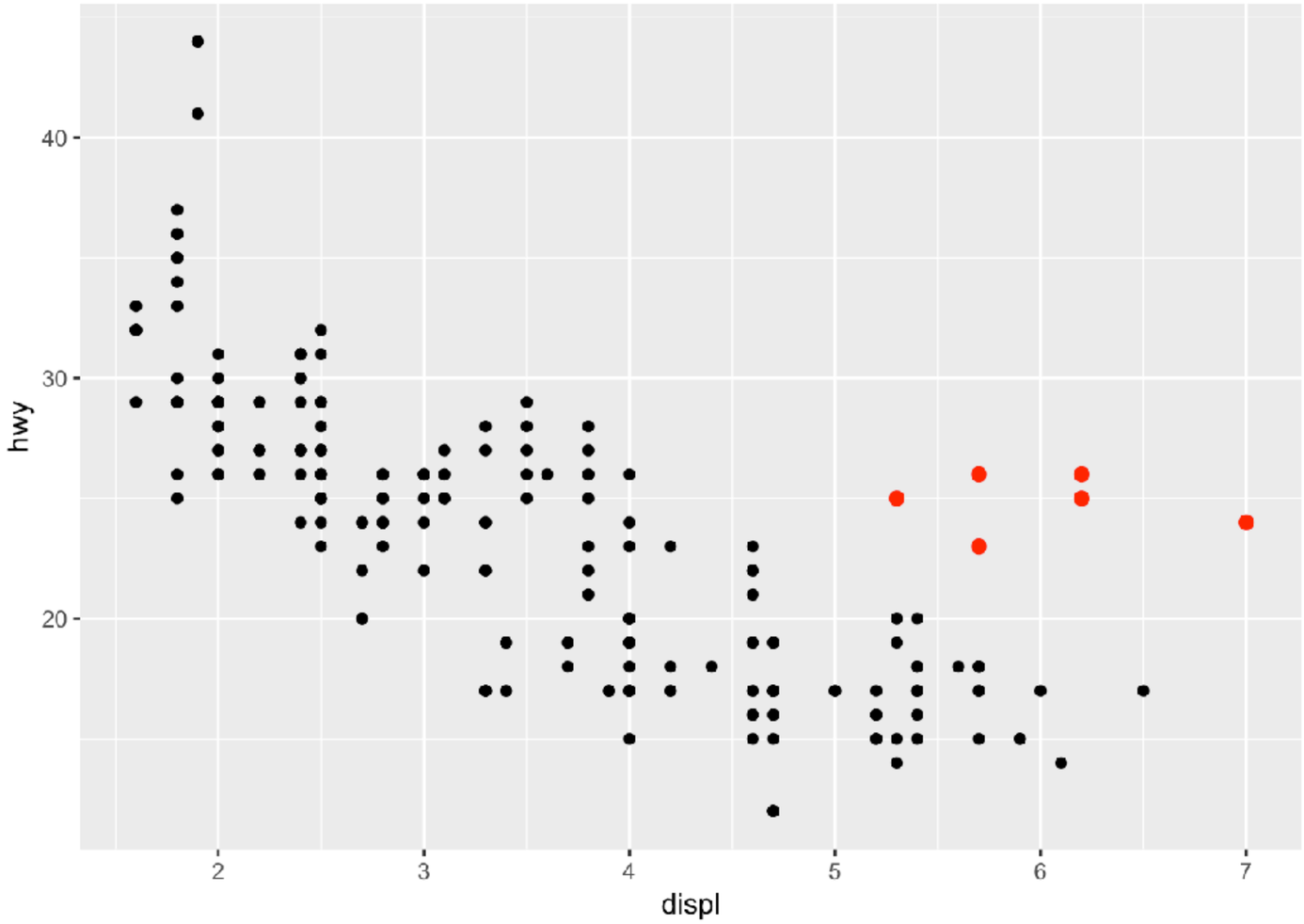
```r
ggplot(data = mpg)
# or (more details in lecture 6)
mpg %>% ggplot()


p <- ggplot(data = mpg)

class(p)
# [1] "gg"     "ggplot"
```

Your turn!

1. Run `ggplot(data = mpg)`. What do you see?
2. How many rows are in `mpg`? How many columns/variables?
3. What does the `drv` variable describe?
4. Make a scatter plot of `hwy` on the x-axis, `cyl` on y-axis.
5. Make a scatter plot of `class` vs. `drv`. Why is this plot not very useful?

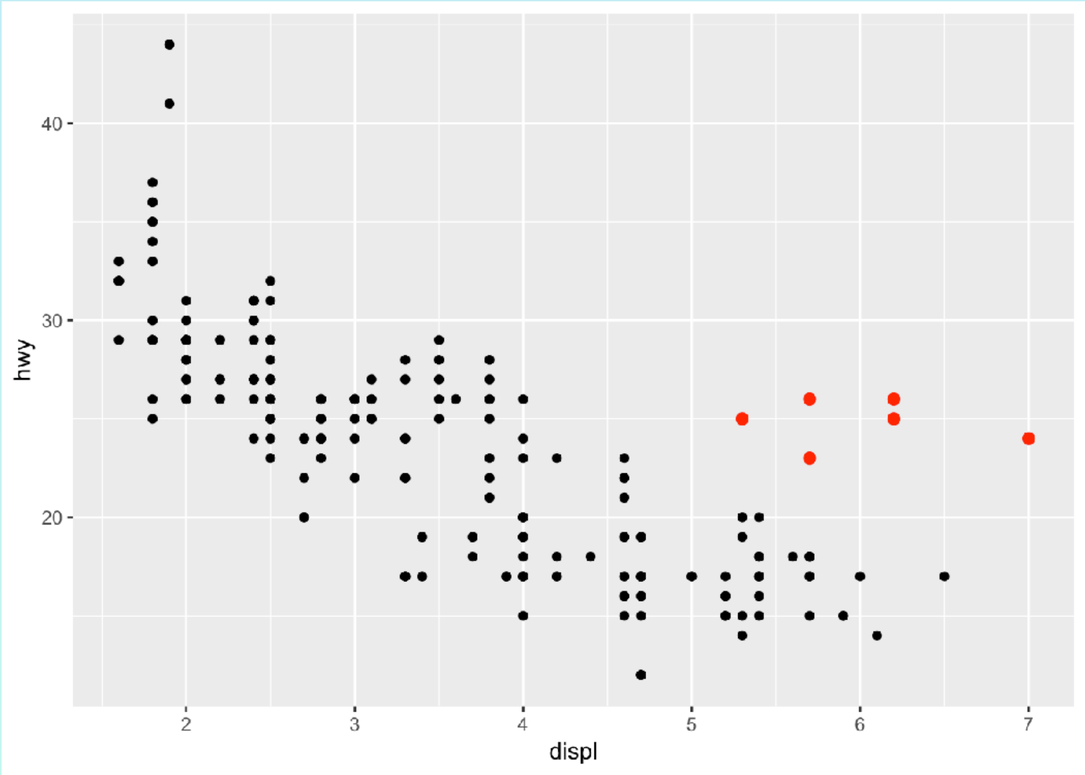
Introduction to Visual Analytics

Exploratory analysis



Your turn!

1. Think about why this may be. Come up with hypotheses.
2. Look up `?mpg` to see which variable may help you address your queries.




```
table(mpg$class)
```

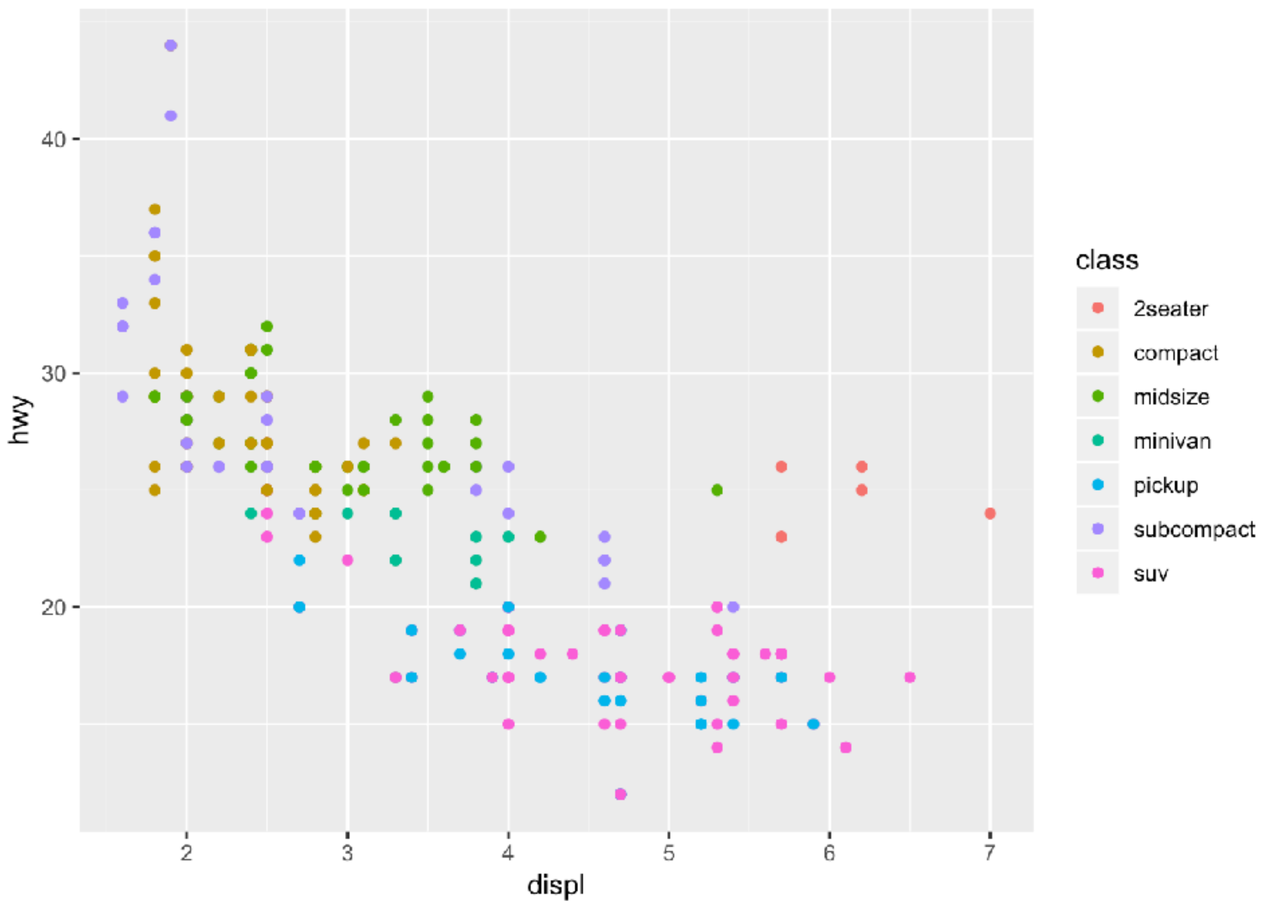
Var1	Freq
2seater	5
compact	47
midsize	41
minivan	11
pickup	33
subcompact	35
suv	62

Datasets

```r
library(tidyverse)

# if you have the library loaded, you can use the dataset directly
mpg

# or you can assigned it to a variable
df <- mpg

# or you can use data() function
data(mpg)
```

Categorical data

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



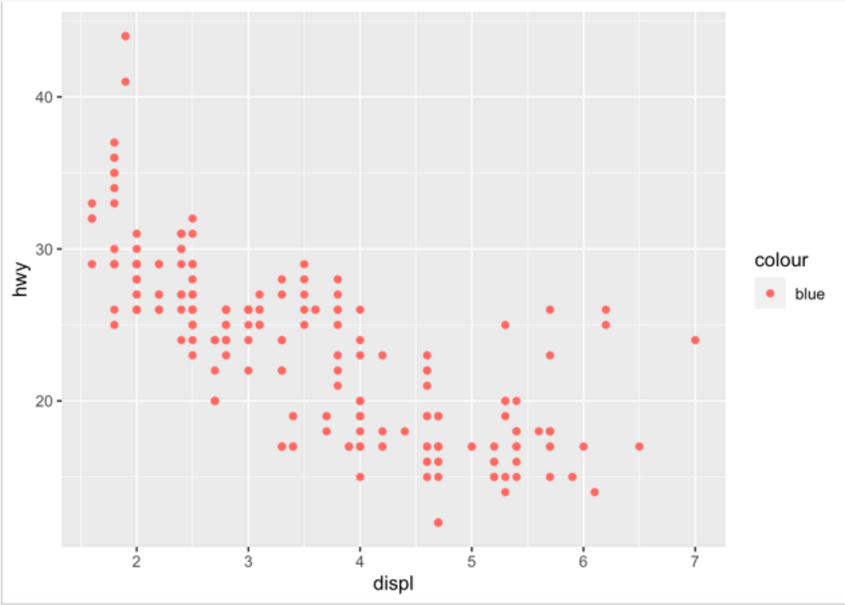
Your turn!

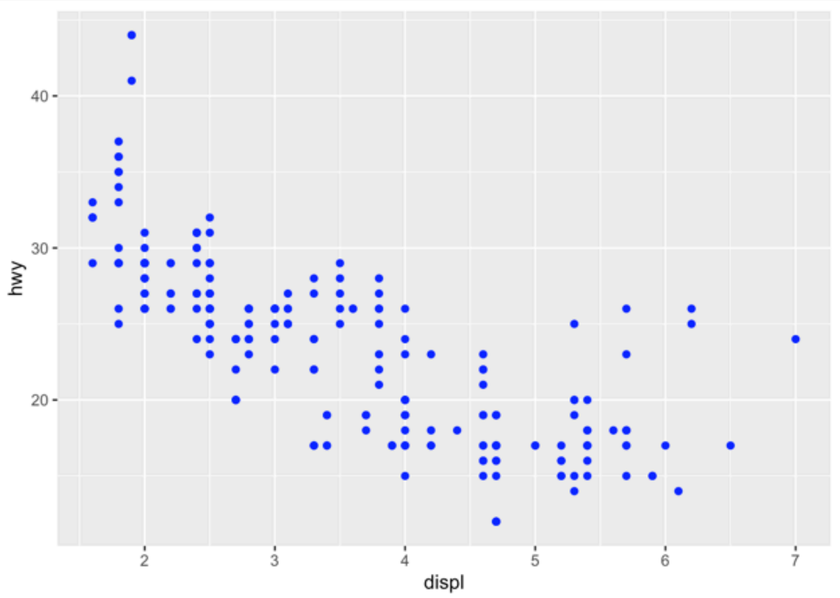
1. Run the following code. Why are the points not blue?

```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, color =“blue”))
```

2. What happens if you map an aesthetic something other than a variable name, like `aes(color = displ < 5)`?

Solution






```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color ="blue"))
```

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color ="blue")
```

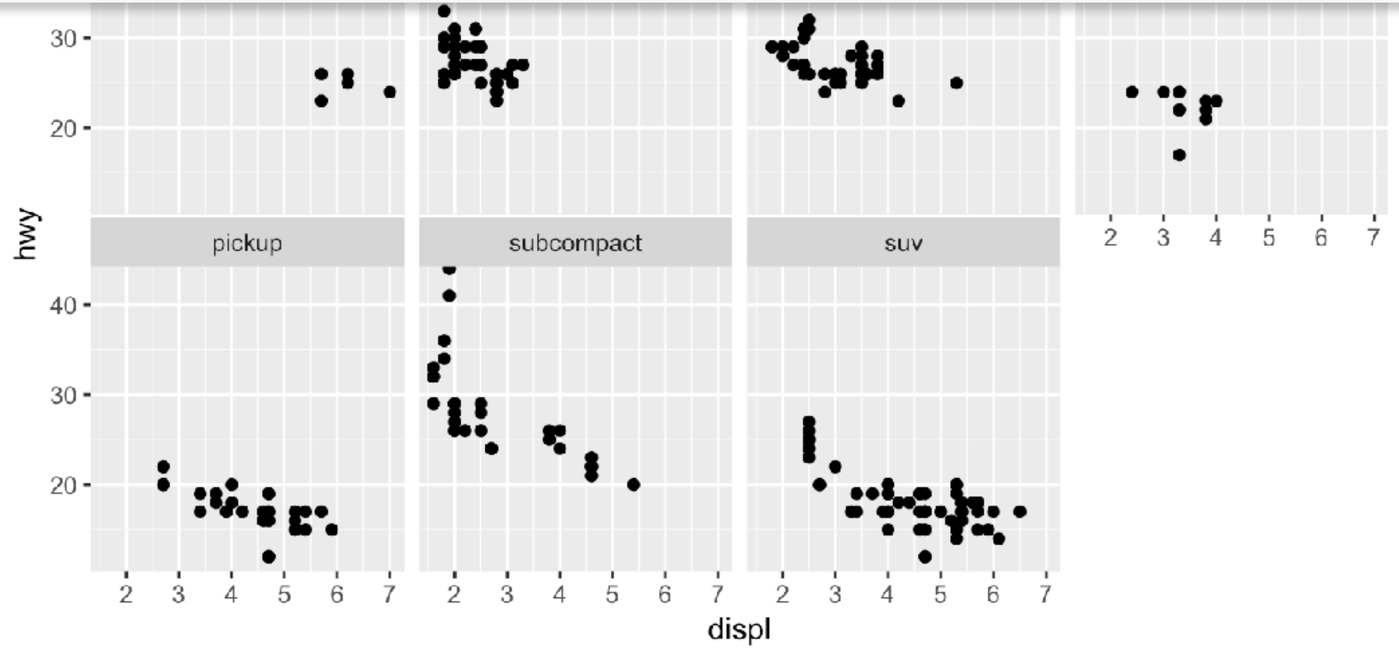
Facets

- partition the data and create small multiple plots side by side
- facet by categorical variable
- **`ggplot2`** comes with 2 useful faceting functions:
  - **`facet_wrap()`**
  - **`facet_grid()`**

Facets

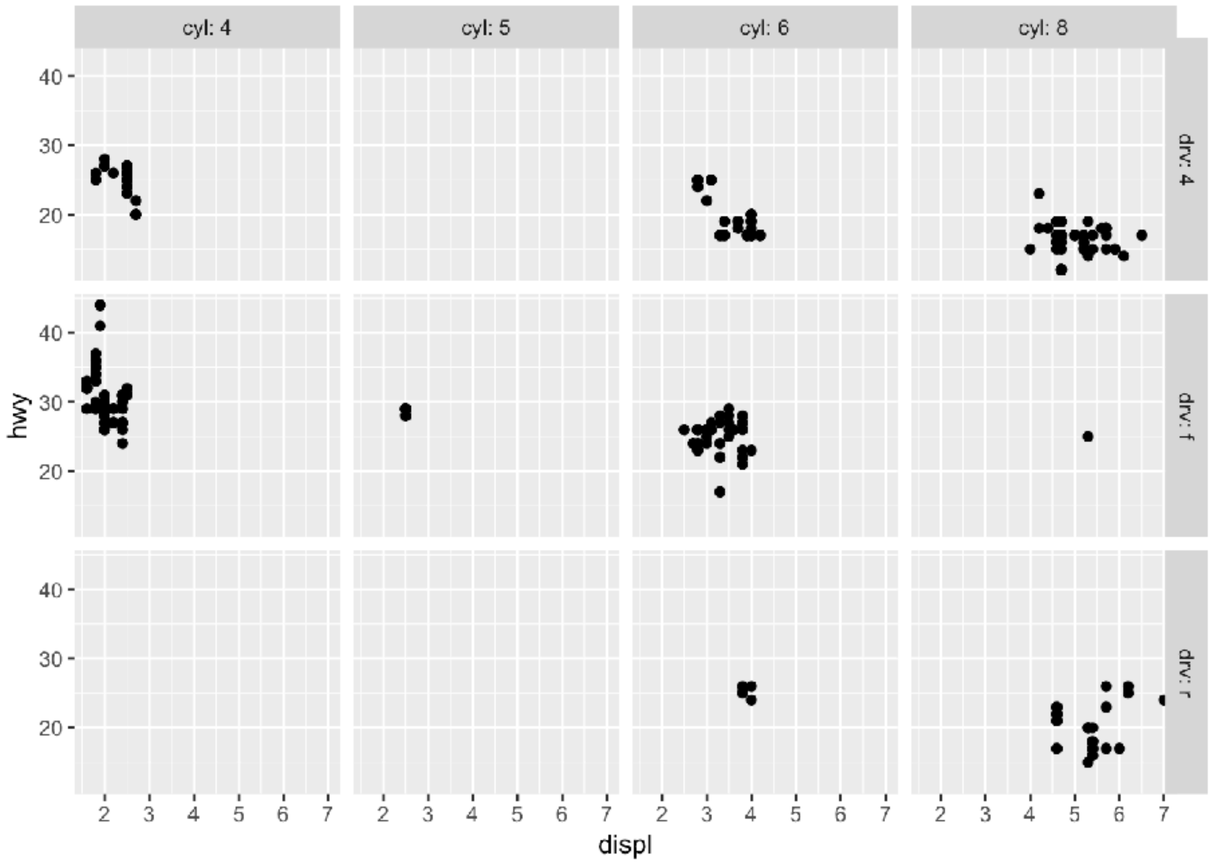
```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(vars(class), nrow = 2)
```



Facets

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ cyl, labeller = label_both)
```



Facets

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ cyl, labeller = label_both)
```

```
p <- ggplot(mpg, aes(displ, cty)) + geom_point()

# Use vars() to supply variables from the dataset:
p + facet_grid(rows = vars(drv))
p + facet_grid(cols = vars(cyl))
p + facet_grid(vars(drv), vars(cyl))

# The historical formula interface is also available:
p + facet_grid(. ~ cyl)
p + facet_grid(drv ~ .)
p + facet_grid(drv ~ cyl)
```

Your turn!

1. What happens if you facet on a continuous variable?
2. What does the following code produce? What does . do in the formula?

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(drv ~ ., labeller = label_both)
```

```
ggplot(data = mpg)+
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_grid(. ~ cyl, labeller = label_both)
```

Lecture 4 - Summary

- Introduction to practical sessions
- Group assignment
- Introduction to R Markdown
- Introduction to visual analytics in R
- Introduction to tidyverse

Lecture 7 - Next lecture

- Introduction to ggplot2
  - geometric objects
  - layered grammar of graphics
  - statistical transformation
  - position adjustment

In-class exercise

- Instruction:
  - Go to Insendi and download the markdown:

  - Work together with your classmates in the breakout room
  - If you have a question, send a message to the instructor
    - You may be pulled out of breakout room if there is a common question
    - Also, check the forum to see answers to FAQs
  - Submit the HTML output indivually, via Insendi by the end of the day.