

Data Science in the tidyverse



Charlotte Wickham

cwick.co.nz

cwickham@gmail.com

 @cvwickham



Data Science in the tidyverse by [Charlotte Wickham](#) is licensed under a [Creative Commons Attribution 4.0 International License](#). Based on a work at <https://github.com/rstudio/master-the-tidyverse>

Introduction

HELLO

my name is

Charlotte

HELLO

my name is

Tonya

HELLO

my name is

Aaron

Your Turn

Introduce yourselves to your neighbours:

Who are you?

What do you do with data?

How would you describe your R experience?

No sticky note: "I'm happily working on it"



Blue sticky note: "I'm all done and ready to move on"



Orange sticky note: "I'm stuck, can someone help me?"



Alternatively, flag one of us down

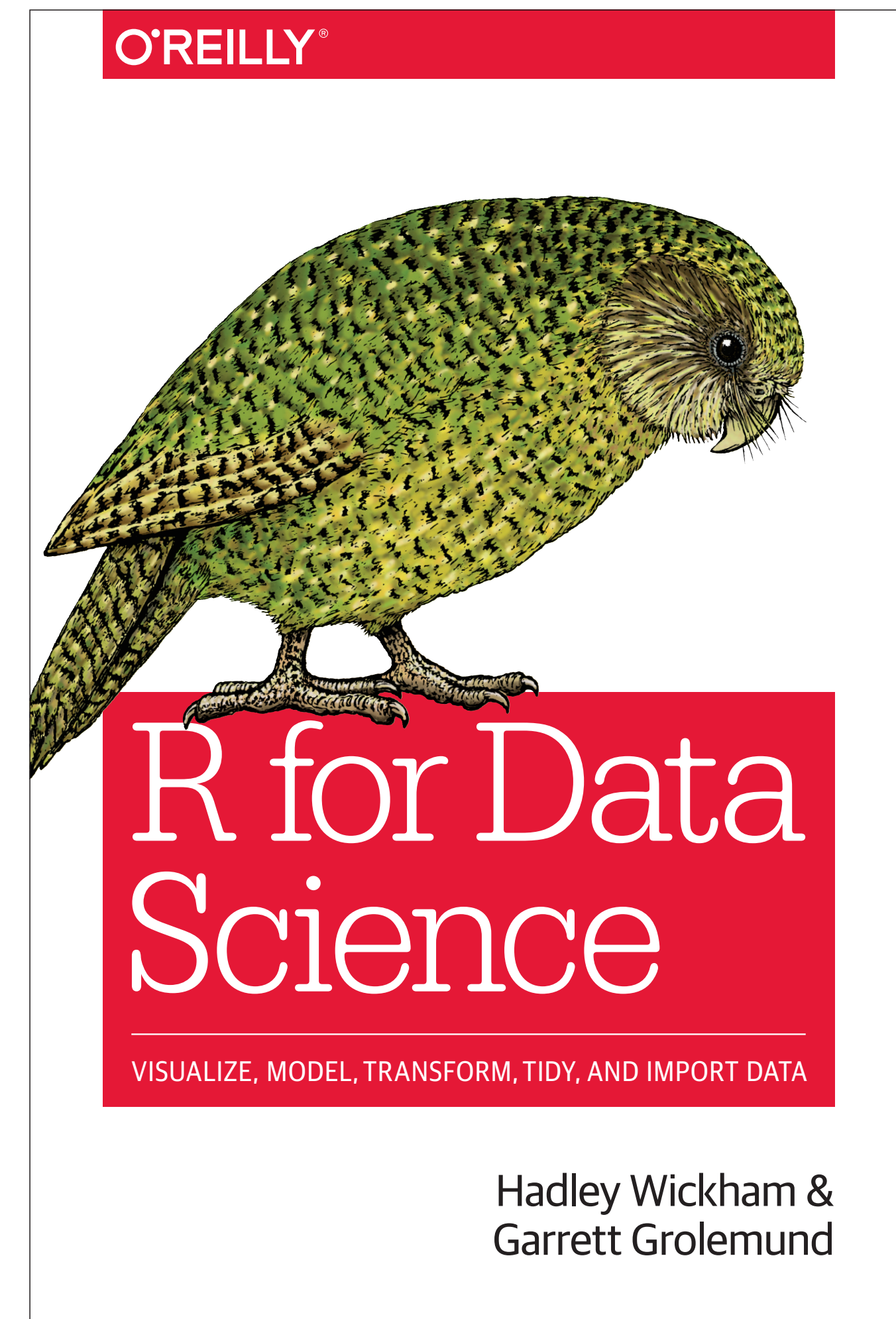
Hopefully, color-blind friendly, let me know if not.

This class is heavily based on

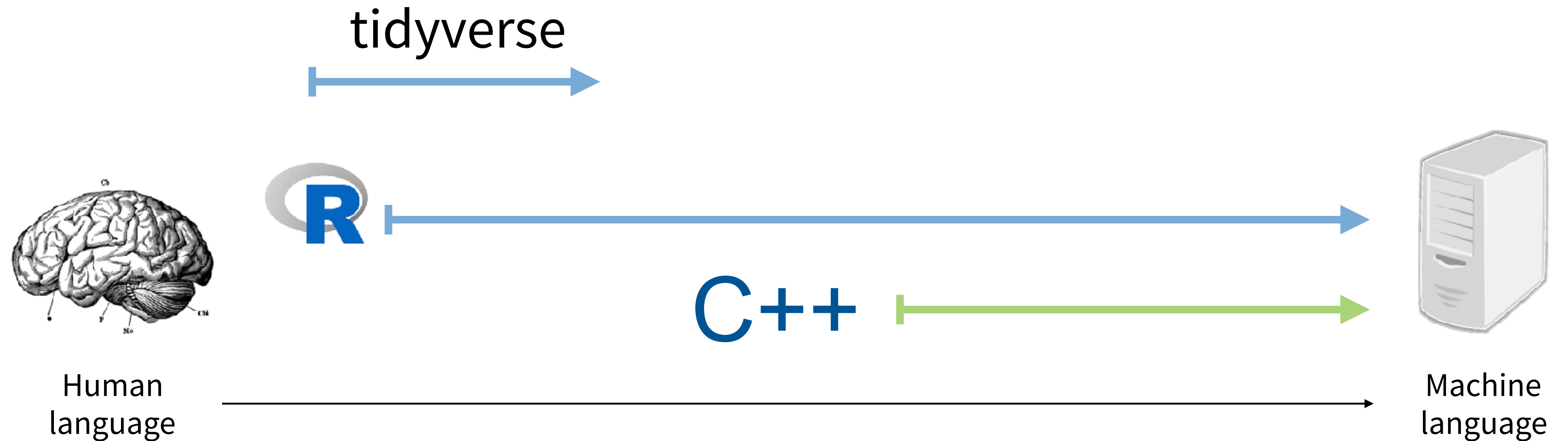
R for Data Science

<http://r4ds.had.co.nz/>

Links to the relevant
sections of the book

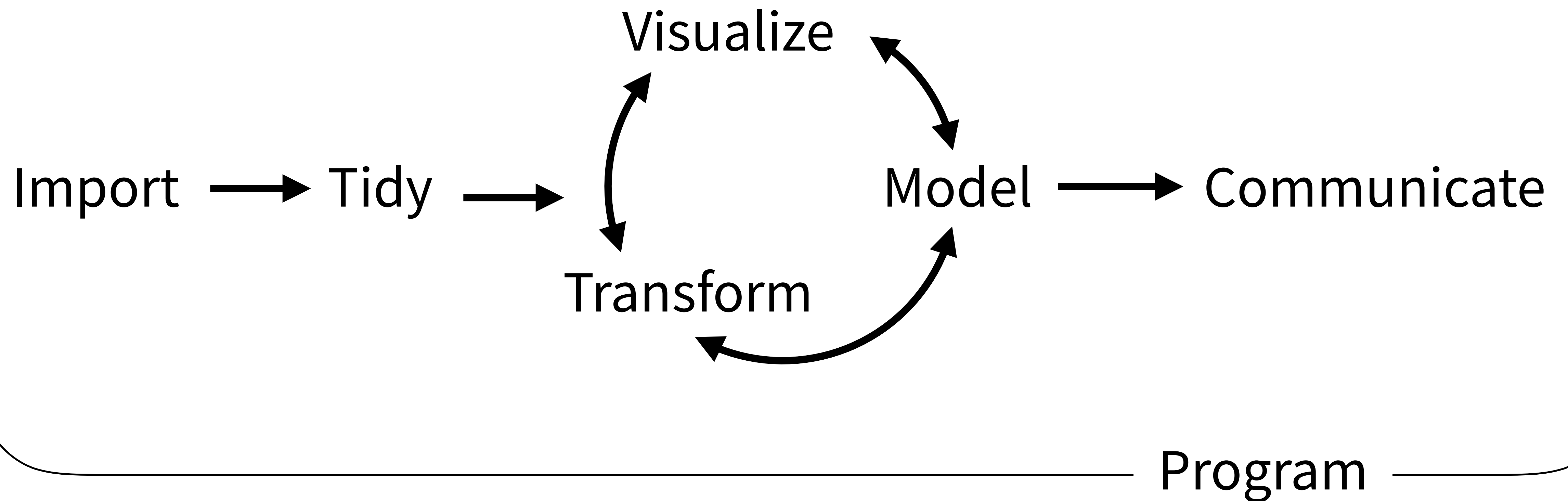


R - A computer language for scientists

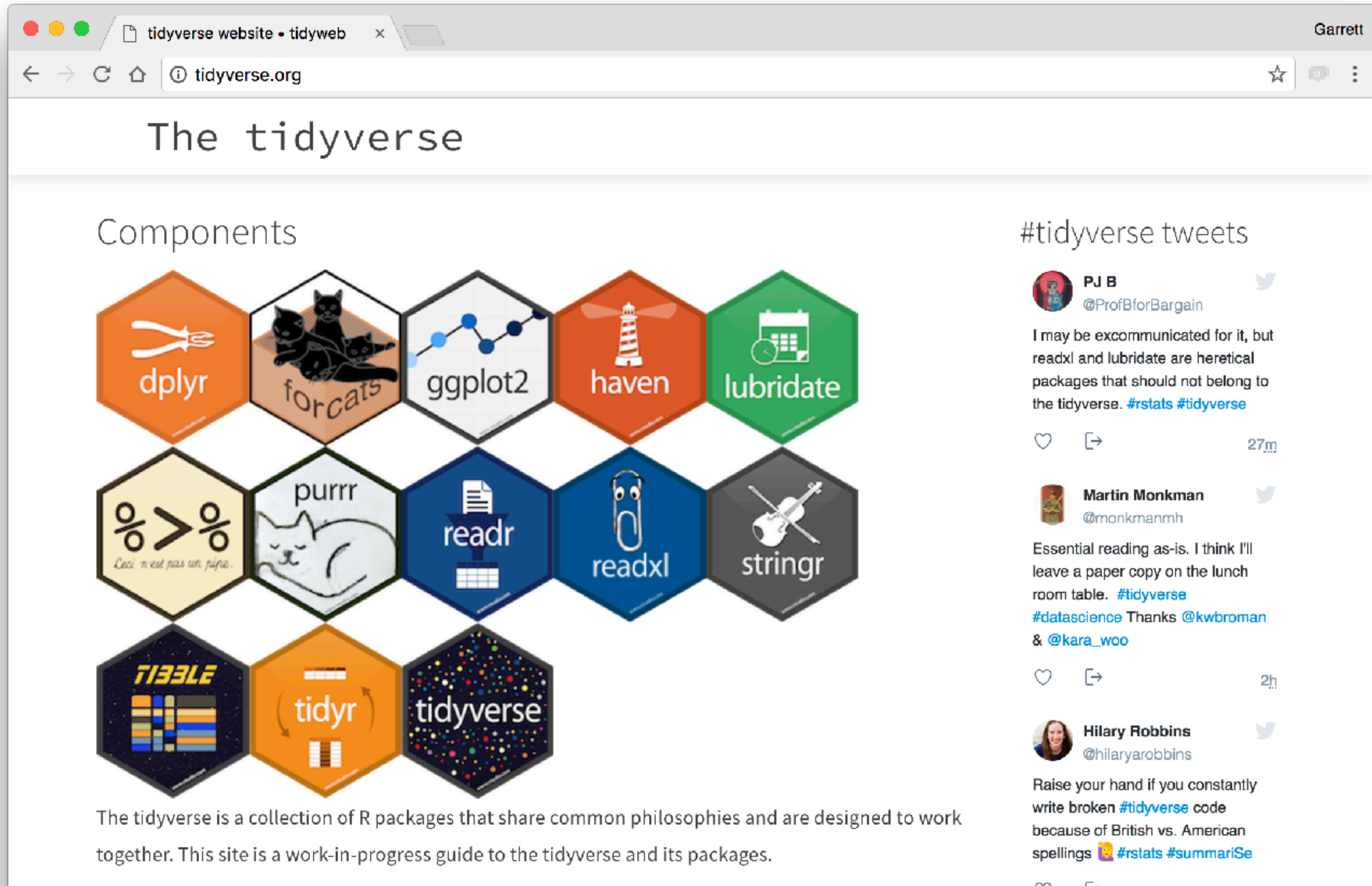


You spend less time thinking about code, and more time thinking about **data analysis**.

(Applied) Data Science



tidyverse.org



The screenshot shows the tidyverse.org website in a web browser. The browser's address bar displays 'tidyverse.org'. The page title is 'The tidyverse'. Below the title, the 'Components' section features a grid of 14 hexagonal logos for various R packages: dplyr, forcats, ggplot2, haven, lubridate, %>% (with the text 'Like: in real life we use pipes'), purrr, readr, readxl, stringr, tibble, tidyr, and tidyverse. To the right of the components is a sidebar titled '#tidyverse tweets' containing three tweets. The first tweet is from PJ B (@ProfBforBargain) discussing the exclusion of readxl and lubridate. The second is from Martin Monkman (@monkmanmh) mentioning a paper copy on a lunch table. The third is from Hilary Robbins (@hilaryarobbins) about writing broken code due to British vs. American spellings. The browser's user interface includes standard navigation buttons, a star icon for bookmarks, and a user profile icon labeled 'Garrett'.

The tidyverse

Components

dplyr forcats ggplot2 haven lubridate

%>%
Like: in real life we use pipes.

purrr readr readxl stringr

tibble tidyr tidyverse

#tidyverse tweets

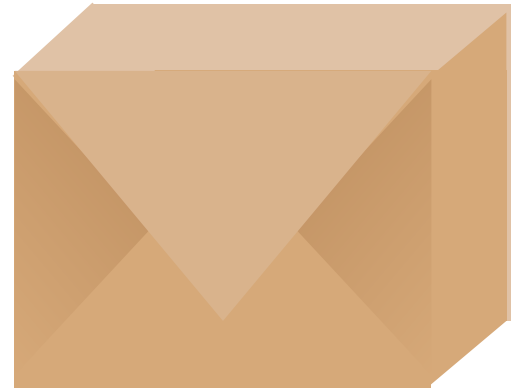
PJ B
@ProfBforBargain
I may be excommunicated for it, but readxl and lubridate are heretical packages that should not belong to the tidyverse. #rstats #tidyverse
27m

Martin Monkman
@monkmanmh
Essential reading as-is. I think I'll leave a paper copy on the lunch room table. #tidyverse #datascience Thanks @kwbroman & @kara_woo
2h

Hilary Robbins
@hilaryarobbins
Raise your hand if you constantly write broken #tidyverse code because of British vs. American spellings 🇬🇧 #rstats #summarise
5m

The tidyverse is a collection of R packages that share common philosophies and are designed to work together. This site is a work-in-progress guide to the tidyverse and its packages.

tidyverse



An R package that serves as a short cut for installing and loading the components of the tidyverse.

```
library("tidyverse")
```

```
install.packages("tidyverse")
```

does the equivalent of

```
install.packages("ggplot2")  
install.packages("dplyr")  
install.packages("tidyr")  
install.packages("readr")  
install.packages("purrr")  
install.packages("tibble")  
install.packages("stringr")  
install.packages("forcats")  
install.packages("lubridate")  
install.packages("hms")  
install.packages("DBI")  
install.packages("haven")  
install.packages("httr")  
install.packages("jsonlite")  
install.packages("readxl")  
install.packages("rvest")  
install.packages("xml2")  
install.packages("modelr")  
install.packages("broom")
```

```
install.packages("tidyverse")
```

does the equivalent of

```
install.packages("ggplot2")  
install.packages("dplyr")  
install.packages("tidyr")  
install.packages("readr")  
install.packages("purrr")  
install.packages("tibble")  
install.packages("stringr")  
install.packages("forcats")  
install.packages("lubridate")  
install.packages("hms")  
install.packages("DBI")  
install.packages("haven")  
install.packages("httr")  
install.packages("jsonlite")  
install.packages("readxl")  
install.packages("rvest")  
install.packages("xml2")  
install.packages("modelr")  
install.packages("broom")
```

```
library("tidyverse")
```

does the equivalent of

```
library("ggplot2")  
library("dplyr")  
library("tidyr")  
library("readr")  
library("purrr")  
library("tibble")  
library("stringr")  
library("forcats")
```

Day 1

**Introduction and
Visualize Data**

9:00 - 10:30

Morning Break

10:30 - 11:00

**Visualize Data/
Transform Data**

11:00 - 12:30

Lunch

12:30 - 1:30

Transform Data

1:30 - 3:00

Afternoon Break

3:00 - 3:30

**Tidy Data/
Case Study**

3:30 - 5:00

Day 2

Data Types	9:00 - 10:30
Morning Break	10:30 - 11:00
Iteration	11:00 - 12:30
Lunch	12:30 - 1:30
Modelling	1:30 - 3:00
Afternoon Break	3:00 - 3:30
Organization with list columns	3:30 - 5:00

Getting Started

Your Turn

Instructions with screenshots at bit.ly/rstudio18-setup

First, if you haven't already

- Visit <https://rstudio.cloud/project/10871>
- Log In / Sign Up
- "Save a copy" of the project
- Open project/data-science-in-the-tidyverse.Rproj

When you have your copy of the project, let us know by **putting** up the **Blue** post-it.

Then open 00-Getting-started.Rmd and take a look around!

rstudio.cloud

A bit like RStudio Server, but hosted for you.

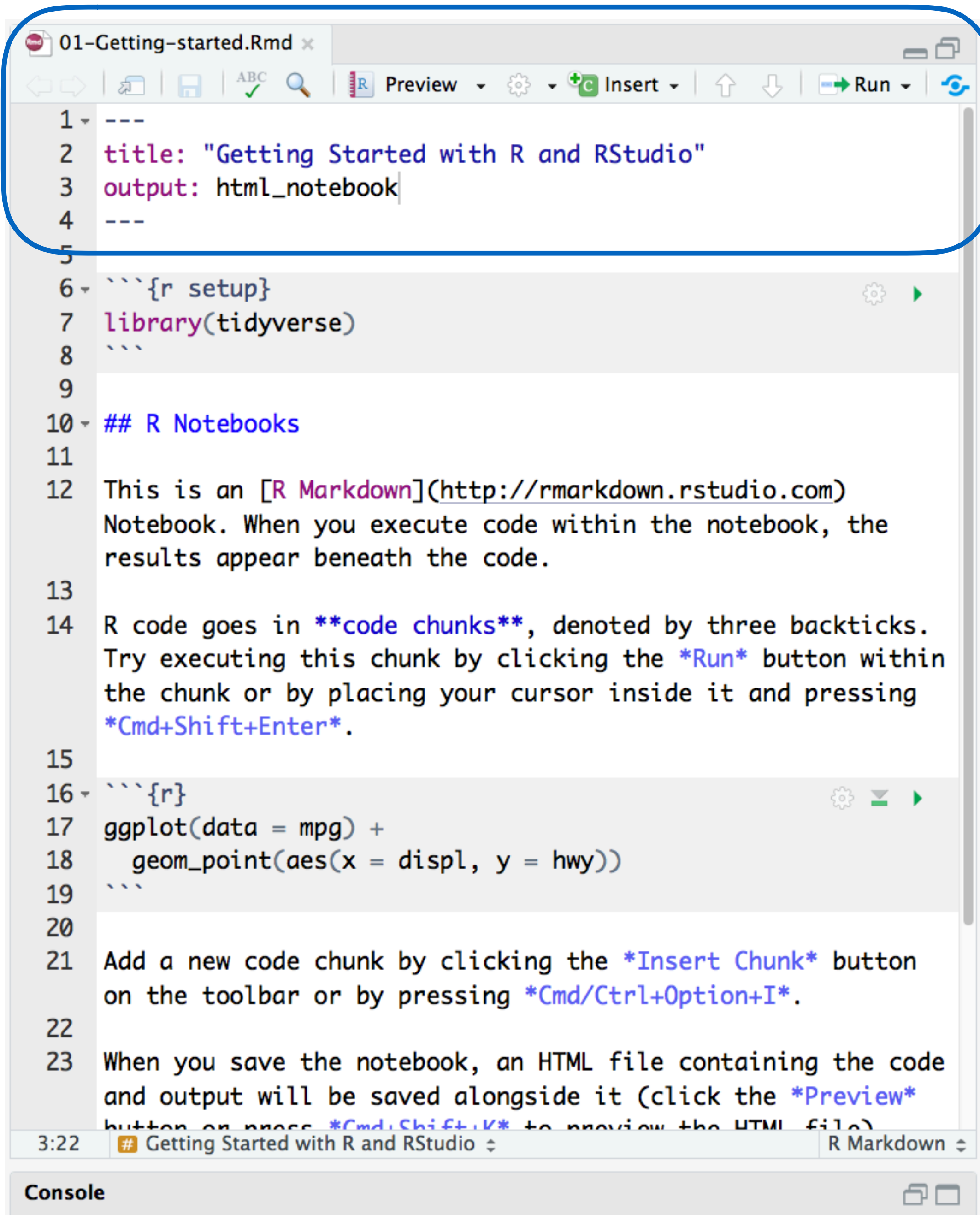
Currently in *alpha*.

If you navigate away, go to rstudio.cloud, and you'll see your project in your Workspace.

R notebooks

An authoring format for
Data Science

00-Getting-started.Rmd is
an R notebook



```
1 ---
2 title: "Getting Started with R and RStudio"
3 output: html_notebook
4 ---
5
6 ```{r setup}
7 library(tidyverse)
8 ```
9
10 ## R Notebooks
11
12 This is an [R Markdown](http://rmarkdown.rstudio.com)
13 Notebook. When you execute code within the notebook, the
14 results appear beneath the code.
15
16 R code goes in code chunks, denoted by three backticks.
17 Try executing this chunk by clicking the Run button within
18 the chunk or by placing your cursor inside it and pressing
19 Cmd+Shift+Enter.
20
21 ```{r}
22 ggplot(data = mpg) +
23   geom_point(aes(x = displ, y = hwy))
24 ```
25
26 Add a new code chunk by clicking the Insert Chunk button
27 on the toolbar or by pressing Cmd/Ctrl+Option+I.
28
29 When you save the notebook, an HTML file containing the code
30 and output will be saved alongside it (click the Preview
31 button on the toolbar or press Cmd+Shift+K to preview the HTML file).
```

The screenshot shows an R Markdown notebook titled "01-Getting-started.Rmd". The editor has a toolbar with buttons for navigation, saving, previewing, and running code. The notebook content includes a YAML header, two code chunks, and several paragraphs of text explaining R Markdown syntax and usage. Two code chunks are highlighted with blue rounded rectangles: the first contains R setup code for the tidyverse, and the second contains a ggplot2 command to create a scatter plot. The status bar at the bottom shows the time as 3:22 and the current file name.

```
1 ---
2 title: "Getting Started with R and RStudio"
3 output: html_notebook
4 ---
5
6 ```{r setup}
7 library(tidyverse)
8 ```
9
10 ## R Notebooks
11
12 This is an [R Markdown](http://rmarkdown.rstudio.com)
13 Notebook. When you execute code within the notebook, the
14 results appear beneath the code.
15
16 R code goes in code chunks, denoted by three backticks.
17 Try executing this chunk by clicking the Run button within
18 the chunk or by placing your cursor inside it and pressing
19 Cmd+Shift+Enter.
20
21 ```{r}
22 ggplot(data = mpg) +
23   geom_point(aes(x = displ, y = hwy))
24 ```
25
26 Add a new code chunk by clicking the Insert Chunk button
27 on the toolbar or by pressing Cmd/Ctrl+Option+I.
28
29 When you save the notebook, an HTML file containing the code
30 and output will be saved alongside it (click the Preview
31 button or press Cmd+Shift+K to preview the HTML file).
```

3:22 # Getting Started with R and RStudio R Markdown

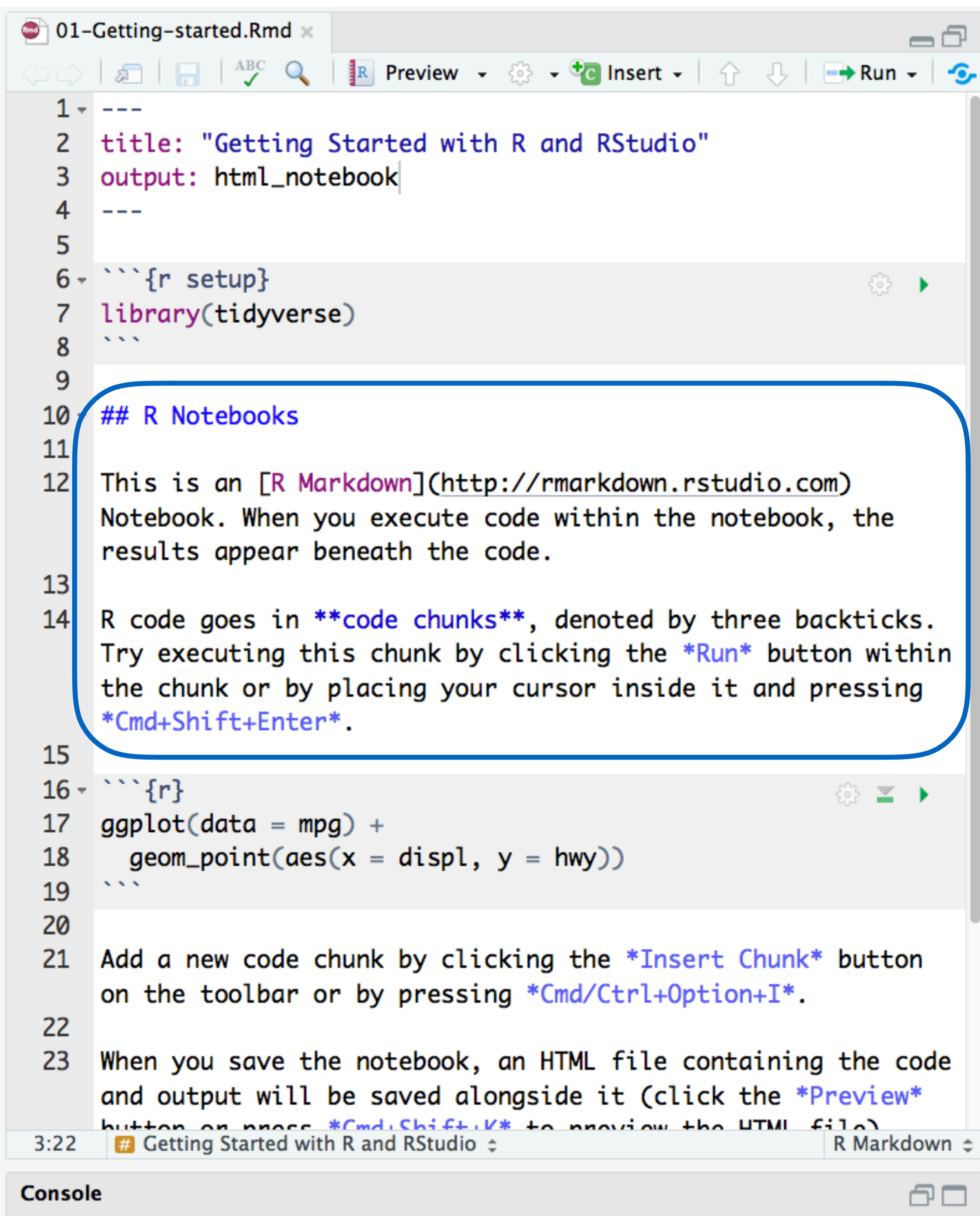
R notebooks

An authoring format for
Data Science

00-Getting-started.Rmd is
an R notebook

Integrates:

- Code



The screenshot shows an R Markdown notebook titled "01-Getting-started.Rmd". The editor has a toolbar with buttons for navigation, saving, previewing, and running code. The notebook content is as follows:

```
1 ---
2 title: "Getting Started with R and RStudio"
3 output: html_notebook
4 ---
5
6 ```{r setup}
7 library(tidyverse)
8 ```
9
10 ## R Notebooks
11
12 This is an [R Markdown](http://rmarkdown.rstudio.com)
13 Notebook. When you execute code within the notebook, the
14 results appear beneath the code.
15
16 R code goes in code chunks, denoted by three backticks.
17 Try executing this chunk by clicking the Run button within
18 the chunk or by placing your cursor inside it and pressing
19 Cmd+Shift+Enter.
20
21 ```{r}
22 ggplot(data = mpg) +
23   geom_point(aes(x = displ, y = hwy))
24 ```
25
26 Add a new code chunk by clicking the Insert Chunk button
27 on the toolbar or by pressing Cmd/Ctrl+Option+I.
28
29 When you save the notebook, an HTML file containing the code
30 and output will be saved alongside it (click the Preview
31 button or press Cmd+Shift+K to preview the HTML file).
```

The status bar at the bottom shows the time 3:22, the file name "# Getting Started with R and RStudio", and the current mode "R Markdown".

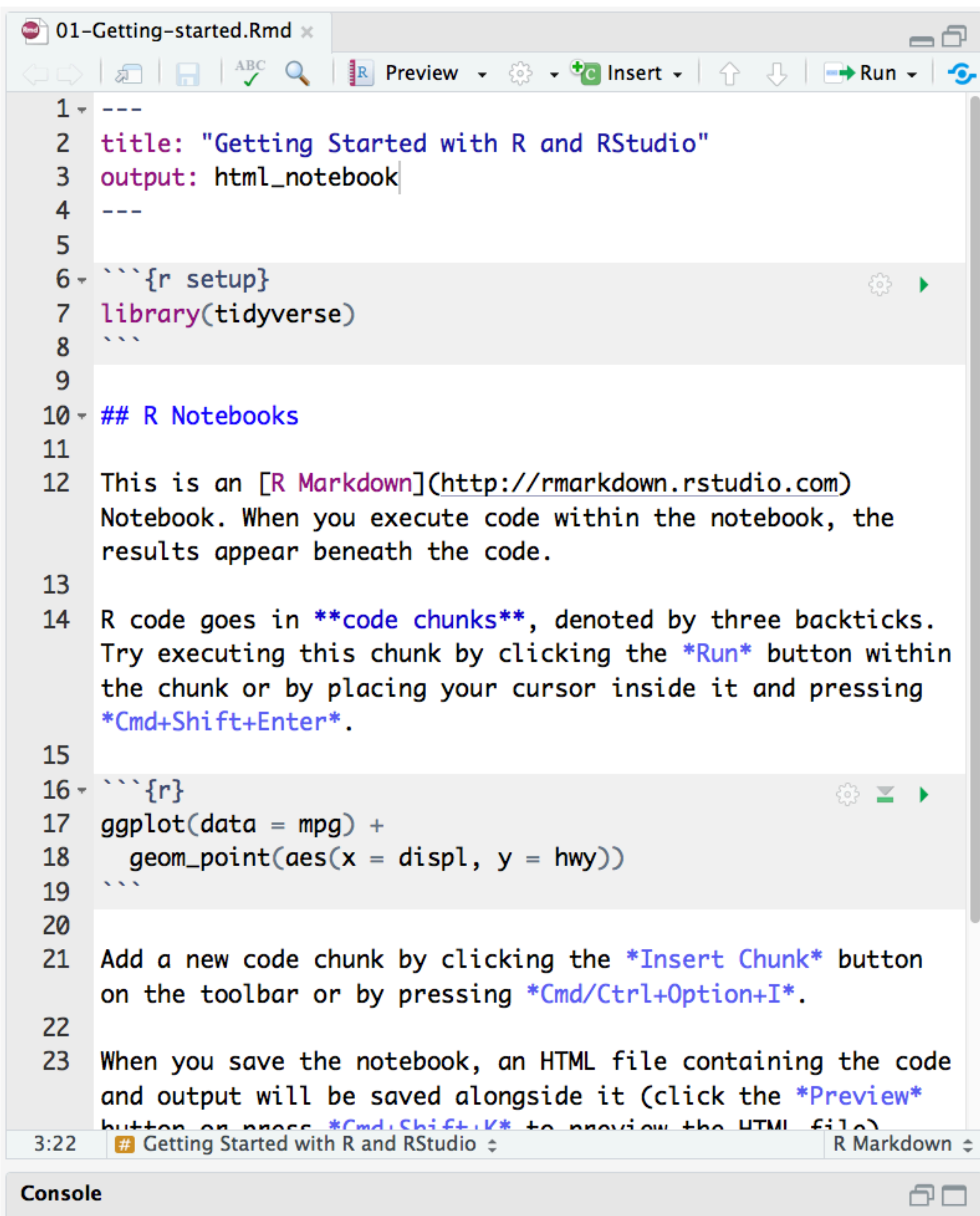
R notebooks

An authoring format for
Data Science

00-Getting-started.Rmd is
an R notebook

Integrates:

- Code
- Text



The screenshot shows an R Markdown notebook titled "01-Getting-started.Rmd". The editor has a toolbar with buttons for navigation, saving, previewing, inserting, and running. The notebook content includes a title, output format, a code chunk for setting up the tidyverse library, a section header "## R Notebooks", and several paragraphs of text explaining R Markdown syntax and usage. There are two code chunks: one for library setup and another for a ggplot2 plot. The status bar at the bottom shows the time 3:22, the file name, and the current mode (R Markdown).

```
1 ---
2 title: "Getting Started with R and RStudio"
3 output: html_notebook
4 ---
5
6 ```{r setup}
7 library(tidyverse)
8 ```
9
10 ## R Notebooks
11
12 This is an [R Markdown](http://rmarkdown.rstudio.com)
13 Notebook. When you execute code within the notebook, the
14 results appear beneath the code.
15
16 R code goes in code chunks, denoted by three backticks.
17 Try executing this chunk by clicking the Run button within
18 the chunk or by placing your cursor inside it and pressing
19 Cmd+Shift+Enter.
20
21 ```{r}
22 ggplot(data = mpg) +
23   geom_point(aes(x = displ, y = hwy))
24 ```
25
26 Add a new code chunk by clicking the Insert Chunk button
27 on the toolbar or by pressing Cmd/Ctrl+Option+I.
28
29 When you save the notebook, an HTML file containing the code
30 and output will be saved alongside it (click the Preview
31 button on the toolbar or press Cmd+Shift+K to preview the HTML file).
```

R notebooks

An authoring format for
Data Science

00-Getting-started.Rmd is
an R notebook

Integrates:

- Code
- Text
- Output

Your Turn

Read the instructions.

Run the code by
hitting the play
button,
or using the keyboard
shortcut.

```
01-Getting-started.Rmd x
1 ---
2 title: "Getting Started with R and RStudio"
3 output: html_notebook
4 ---
5
6 ```{r setup}
7 library(tidyverse)
8 ```
9
10 ## R Notebooks
11
12 This is an [R Markdown](http://rmarkdown.rstudio.com)
13 Notebook. When you execute code within the notebook, the
14 results appear beneath the code.
15
16 R code goes in code chunks, denoted by three backticks.
17 Try executing this chunk by clicking the Run button within
18 the chunk or by placing your cursor inside it and pressing
19 Cmd+Shift+Enter.
20
21 ```{r}
22 ggplot(data = mpg) +
23   geom_point(aes(x = displ, y = hwy))
24 ```
25
26 Add a new code chunk by clicking the Insert Chunk button
27 on the toolbar or by pressing Cmd/Ctrl+Option+I.
28
29 When you save the notebook, an HTML file containing the code
30 and output will be saved alongside it (click the Preview
31 button on press Cmd+Shift+K to preview the HTML file).
```

3:22 # Getting Started with R and RStudio R Markdown

Console

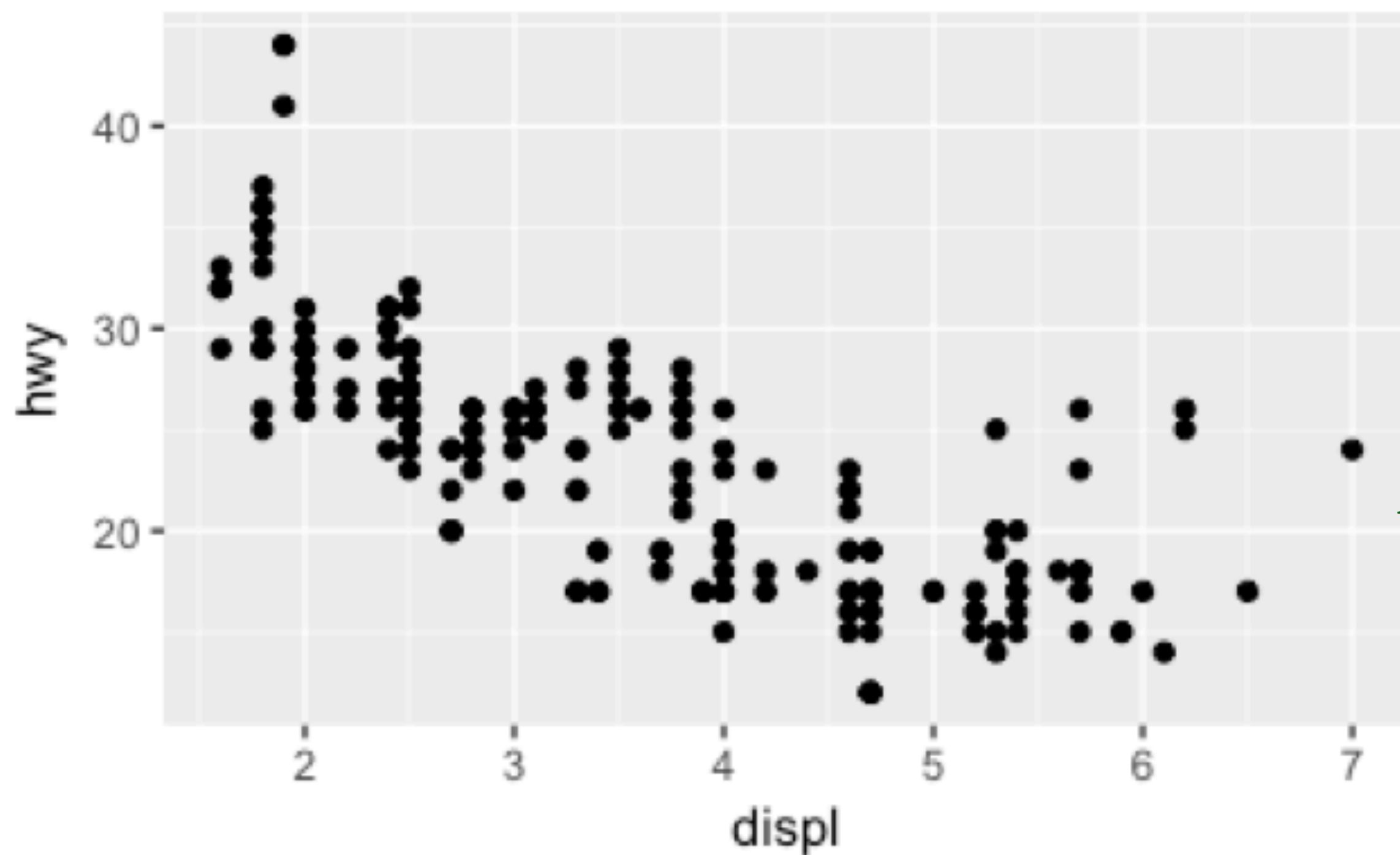
```
```{r}
ggplot(data = mpg) +
 geom_point(aes(x = displ, y = hwy))
```
```



Click to run code
in chunk



Click to run all
code chunks
above



Code result

R Notebooks

An easy way to combine R code and narrative

Useful for us:

- I'll provide starter code
- You can complete "Your Turns"
- At the end, a useful record for you

Your Turn

Open 01-Visualize.Rmd



I'm working on it



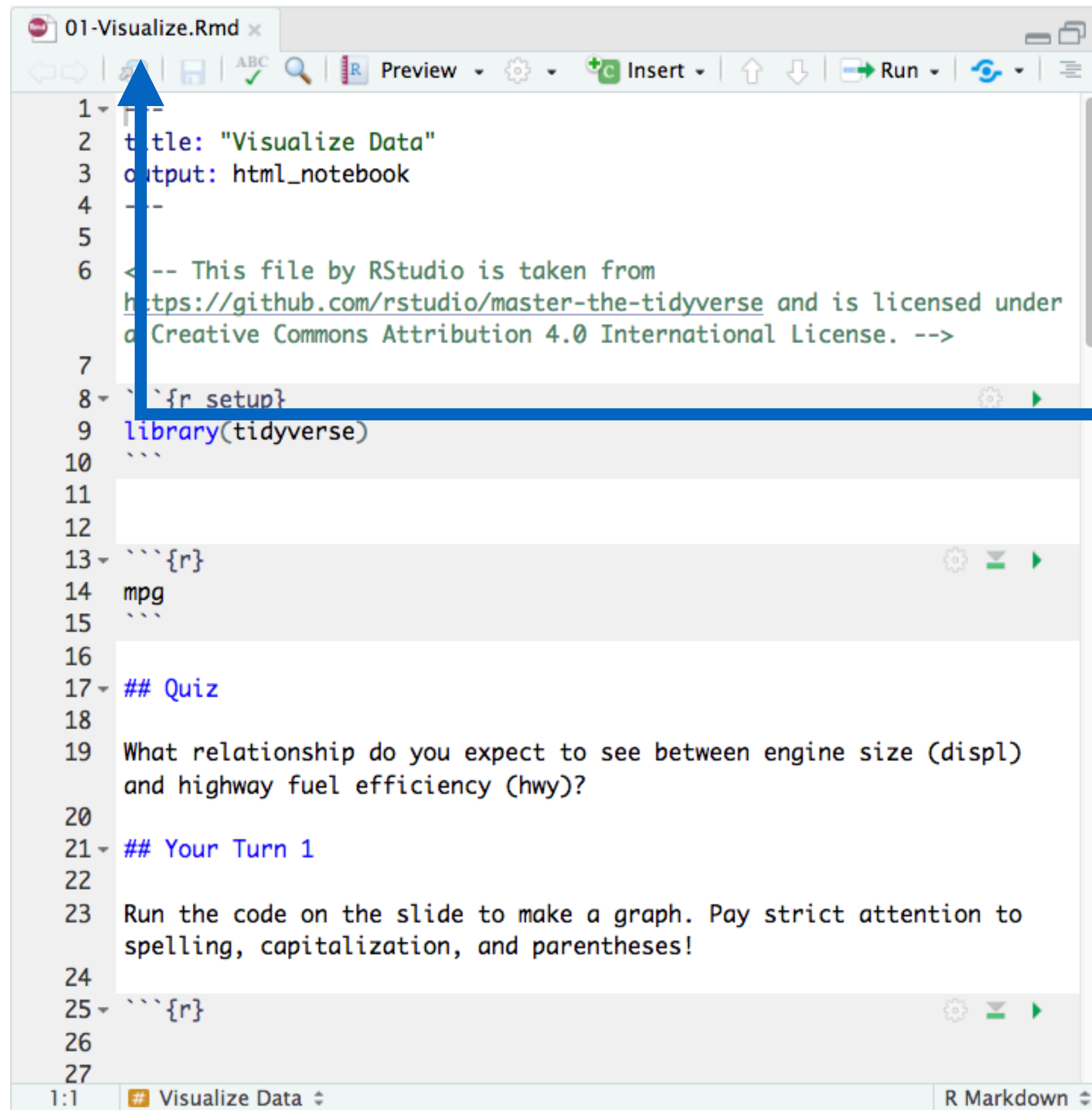
I'm stuck!



I'm done!

If you get lost or need to restart

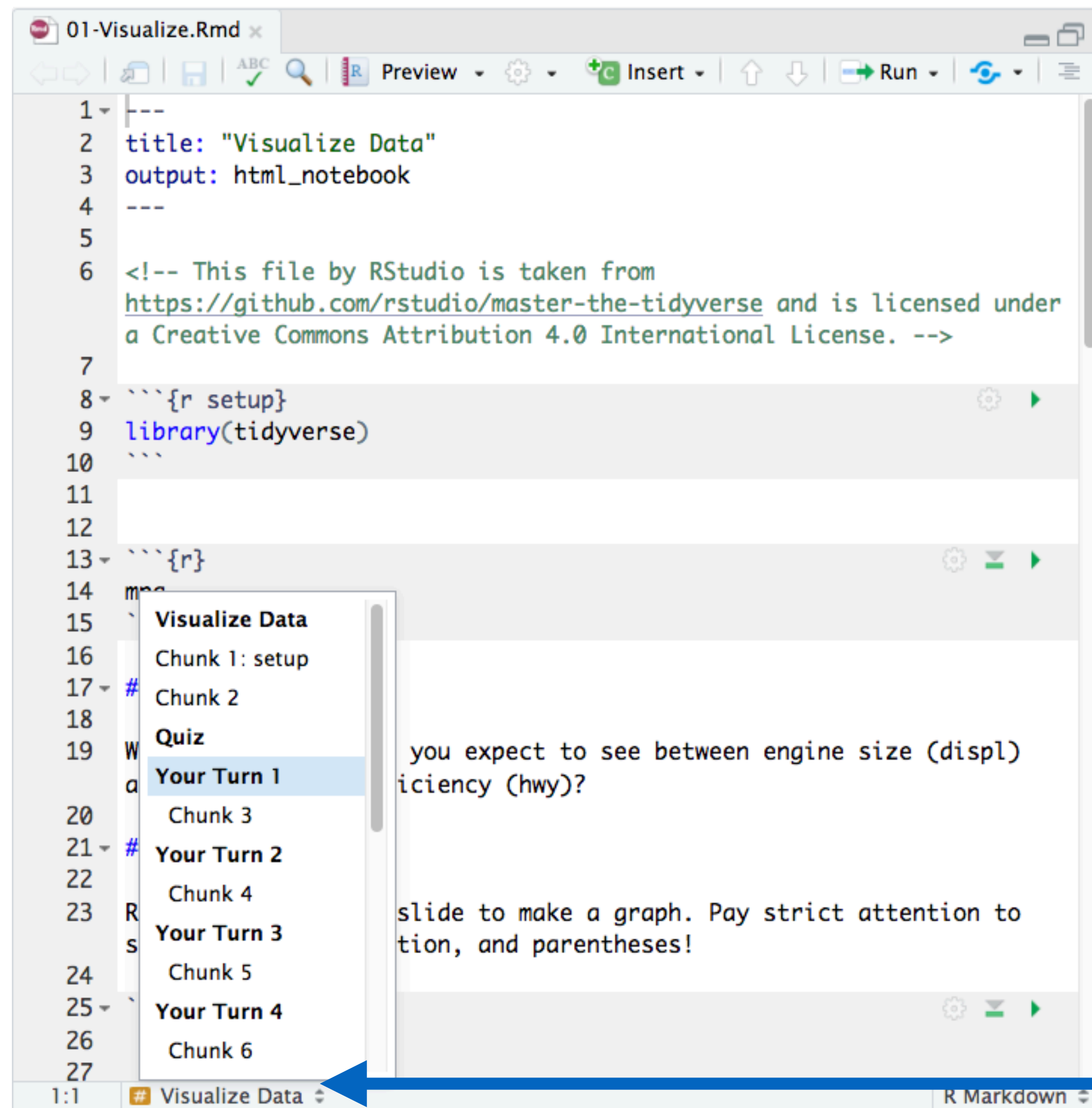
Check you are in the
right file

A screenshot of the RStudio R Markdown editor window. The title bar shows '01-Visualize.Rmd'. The editor contains R Markdown code with line numbers 1 through 27. A blue arrow points from the text 'Check you are in the right file' to the title 'Visualize Data' on line 2. The code includes a title, output format, a license notice, R code chunks for library setup and data loading, and a quiz section.

```
1 title: "Visualize Data"
2 output: html_notebook
3 ---
4
5
6 <-- This file by RStudio is taken from
7 https://github.com/rstudio/master-the-tidyverse and is licensed under
8 a Creative Commons Attribution 4.0 International License. -->
9
10 {r setup}
11 library(tidyverse)
12
13 {r}
14 mpg
15
16
17 ## Quiz
18
19 What relationship do you expect to see between engine size (displ)
20 and highway fuel efficiency (hwy)?
21
22 ## Your Turn 1
23
24 Run the code on the slide to make a graph. Pay strict attention to
25 spelling, capitalization, and parentheses!
26
27 {r}
```

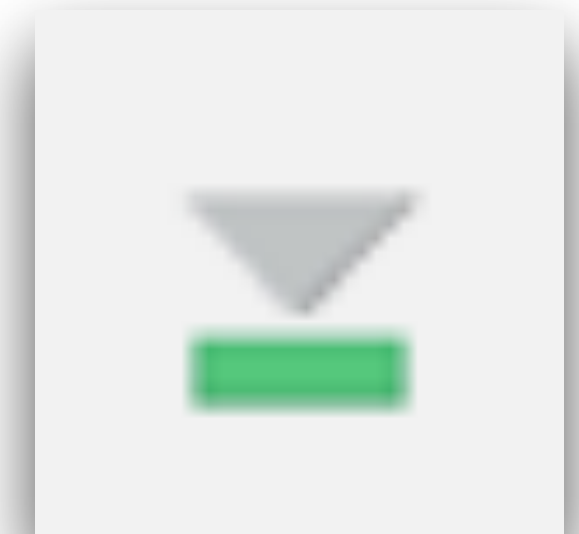

If you get lost or need to restart

Use the section
browser to quickly
navigate to the
right *Your Turn*

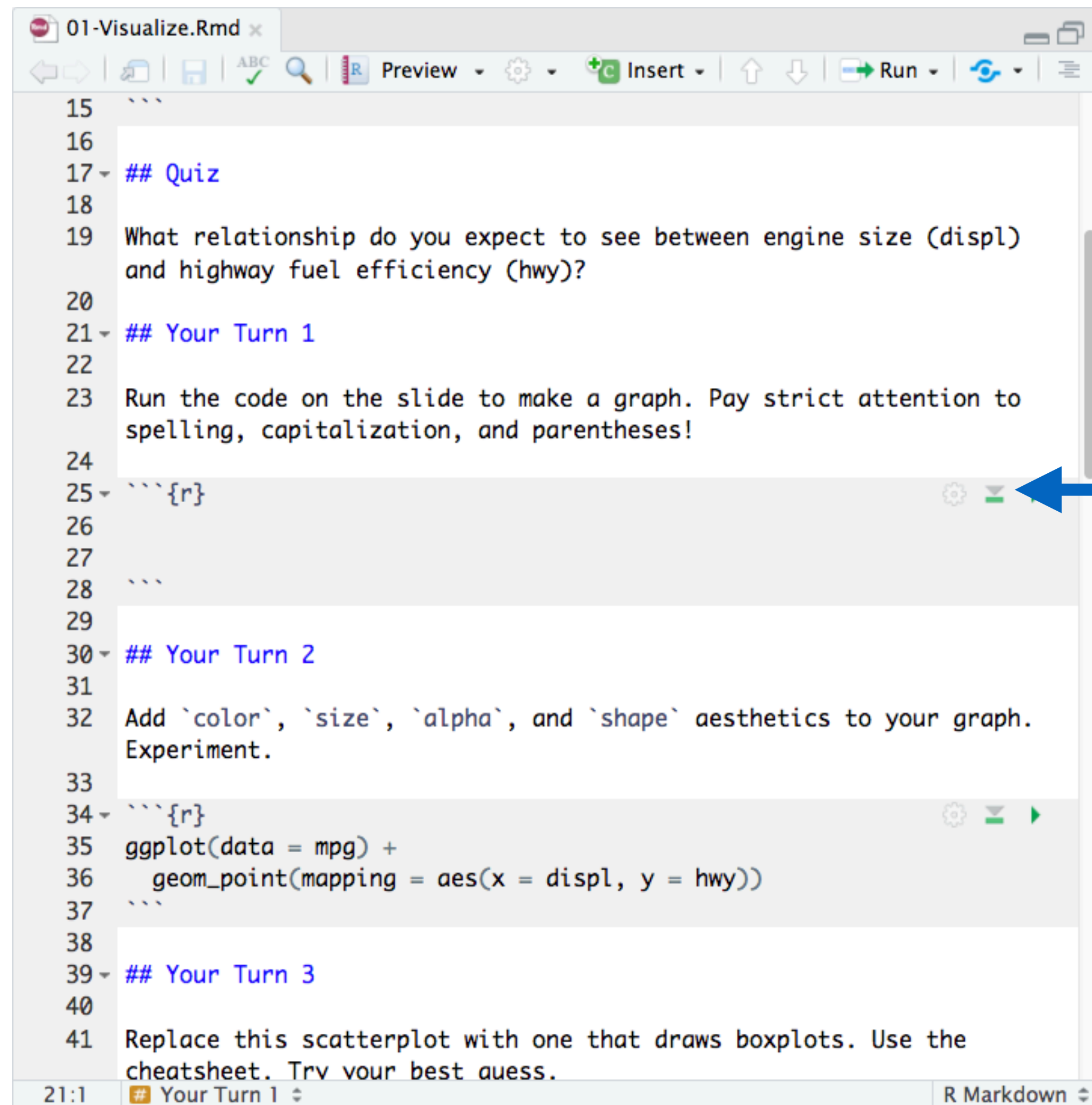


If you get lost or need to restart

Click to run all
chunks before this
one.



You should be ready
to go.



```
15  ```
16
17  ## Quiz
18
19  What relationship do you expect to see between engine size (displ)
20  and highway fuel efficiency (hwy)?
21
22  ## Your Turn 1
23
24  Run the code on the slide to make a graph. Pay strict attention to
25  spelling, capitalization, and parentheses!
26
27  ```{r}
28
29
30  ## Your Turn 2
31
32  Add `color`, `size`, `alpha`, and `shape` aesthetics to your graph.
33  Experiment.
34
35  ```{r}
36  ggplot(data = mpg) +
37    geom_point(mapping = aes(x = displ, y = hwy))
38
39  ## Your Turn 3
40
41  Replace this scatterplot with one that draws boxplots. Use the
42  cheatsheet. Try your best guess.
```