

# Data Science in the tidyverse



Hadley Wickham

hadley.nz

🐦 @hadleywickham



This material is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/). Based on a work at <https://github.com/cwickham/data-science-in-tidyverse>

# Introduction

**HELLO**

my name is

**Hadley**



# 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"



**Green** 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.

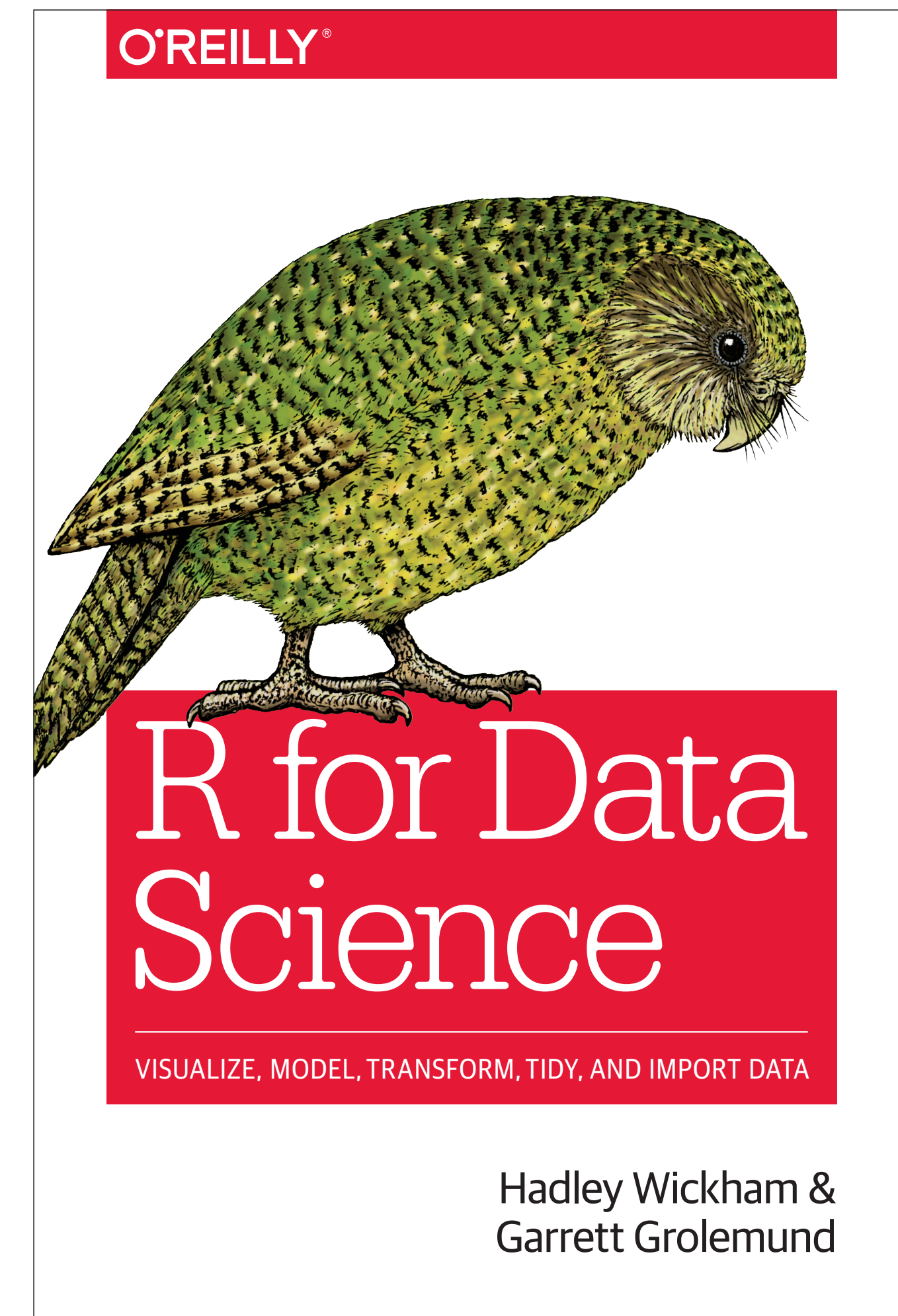


**This class is heavily based on**

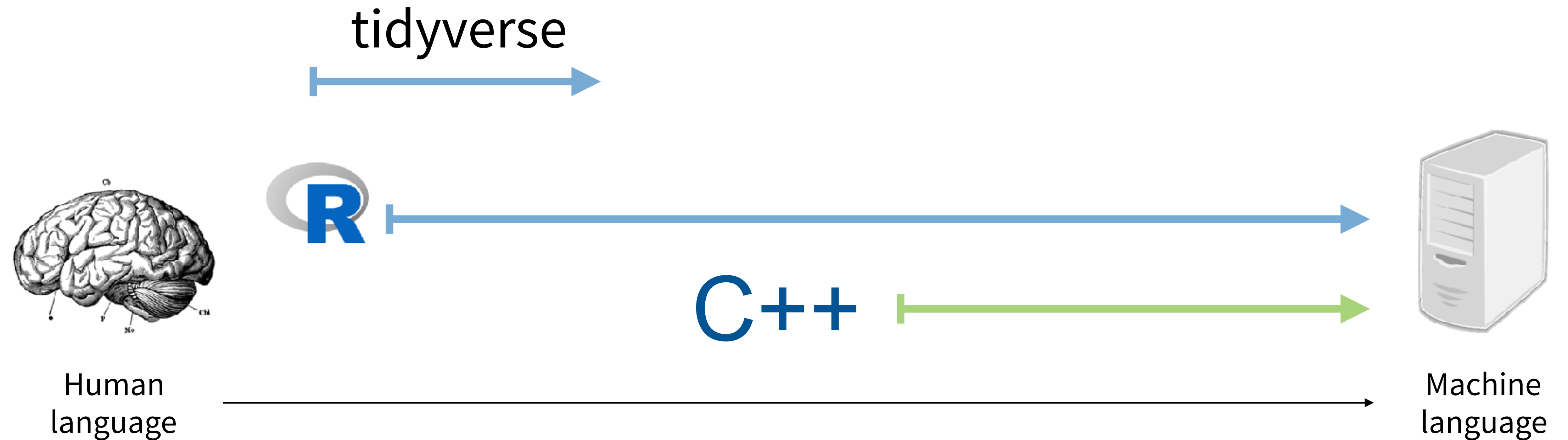
R for Data Science

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

Links to the relevant  
sections of the book



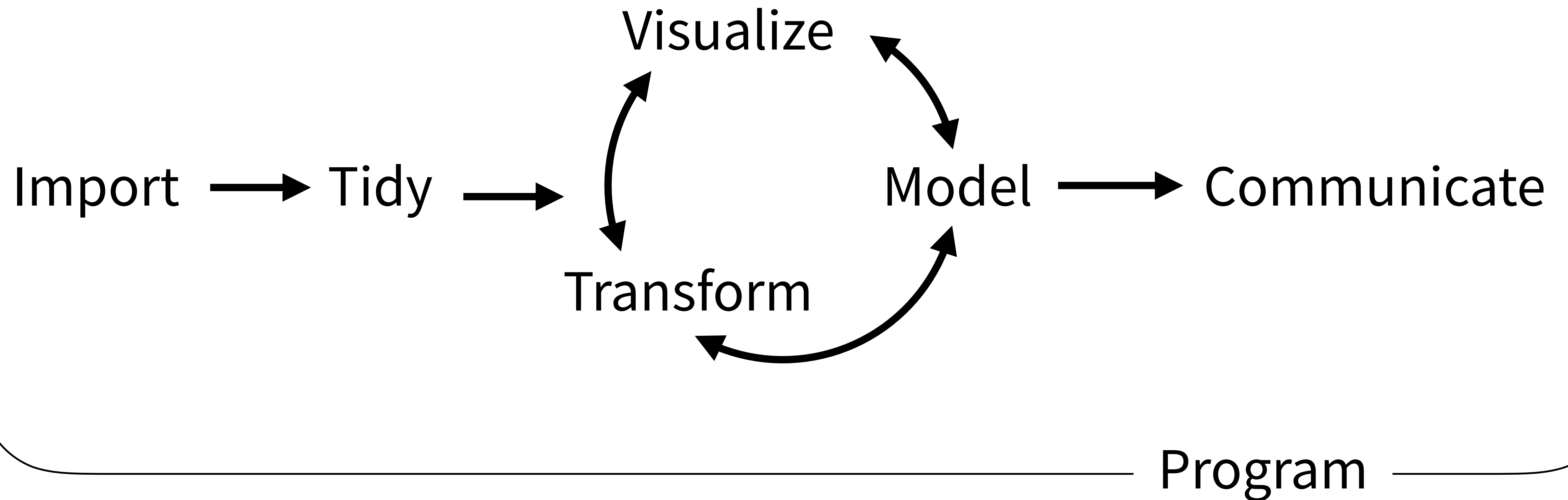
# **R** - A programming language for data

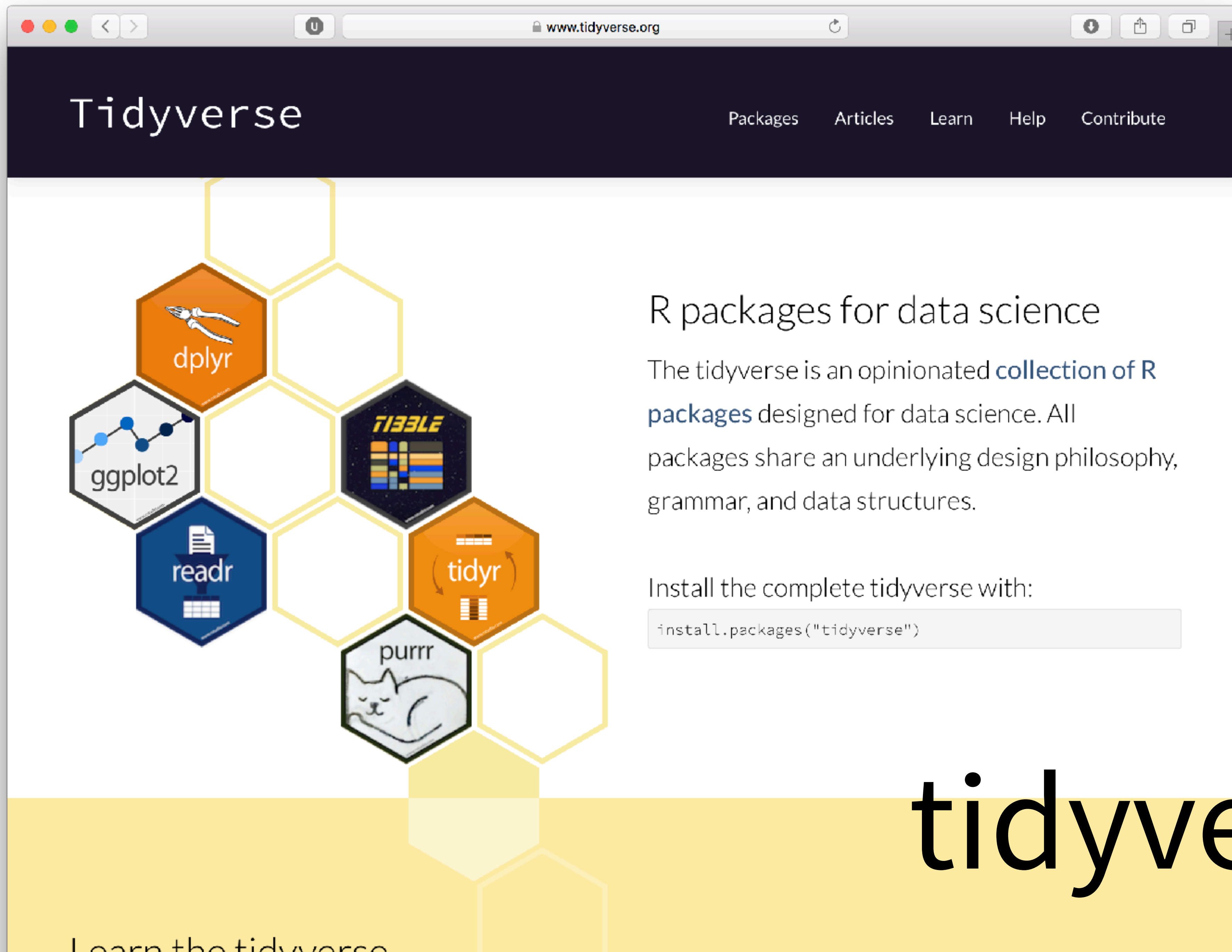


You spend less time thinking about code, and more time thinking about **data analysis**.



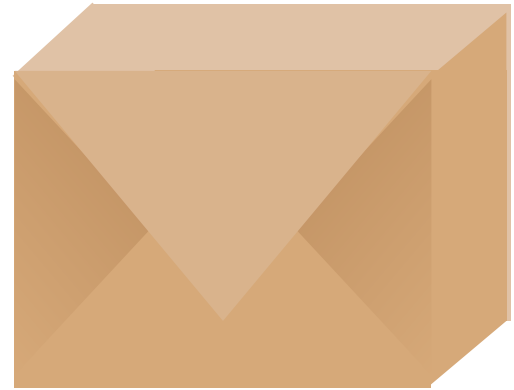
# (Applied) Data Science





tidyverse.org

# 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")
```

# Option 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

# Option 2

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

# Getting Started



# Your Turn

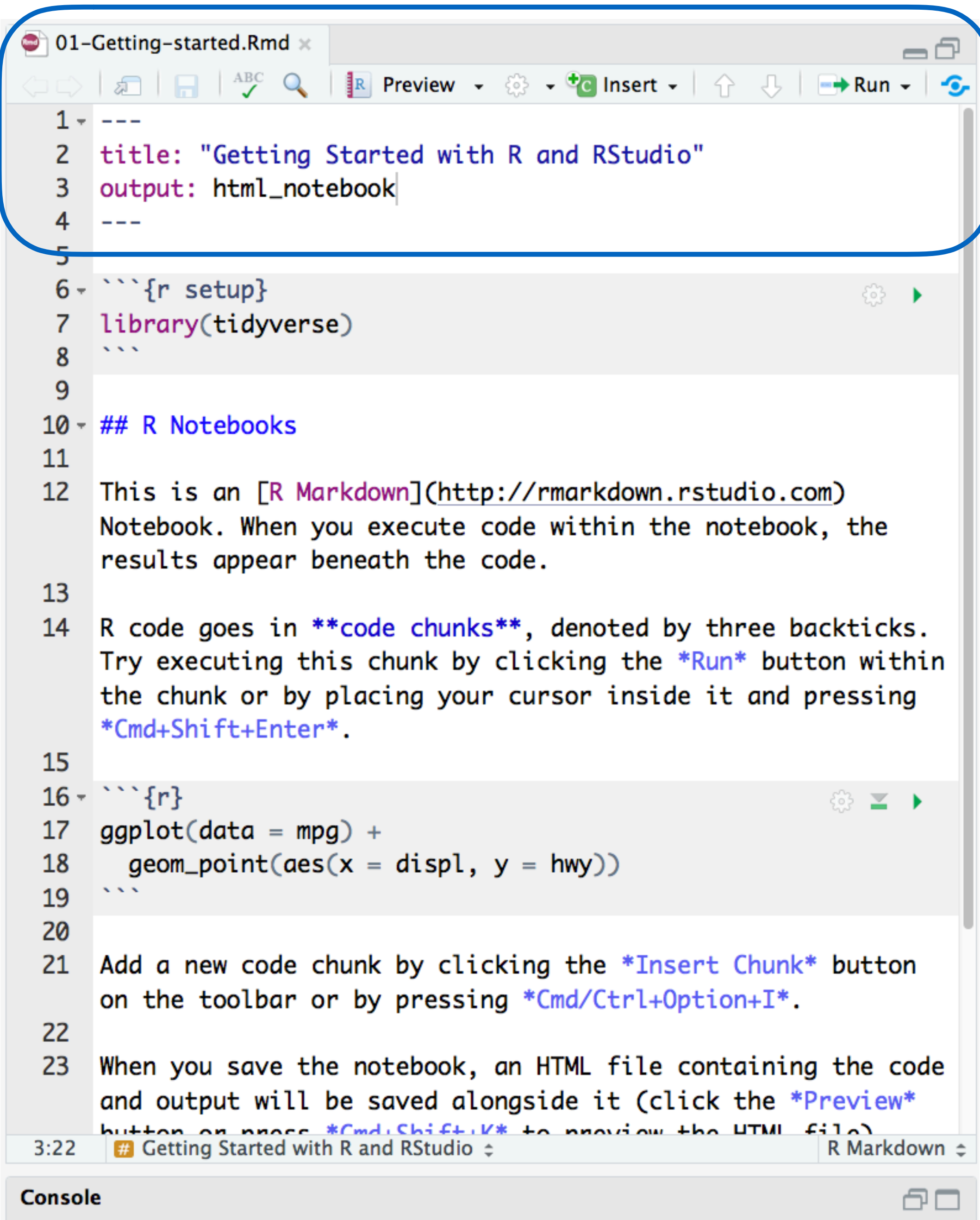
Open data-science-in-the-tidyverse.Rproj

**Then** open 00-Getting-started.Rmd and follow the instructions!

# R notebooks

An authoring format for  
Data Science

00-Getting-started.Rmd is  
an R notebook

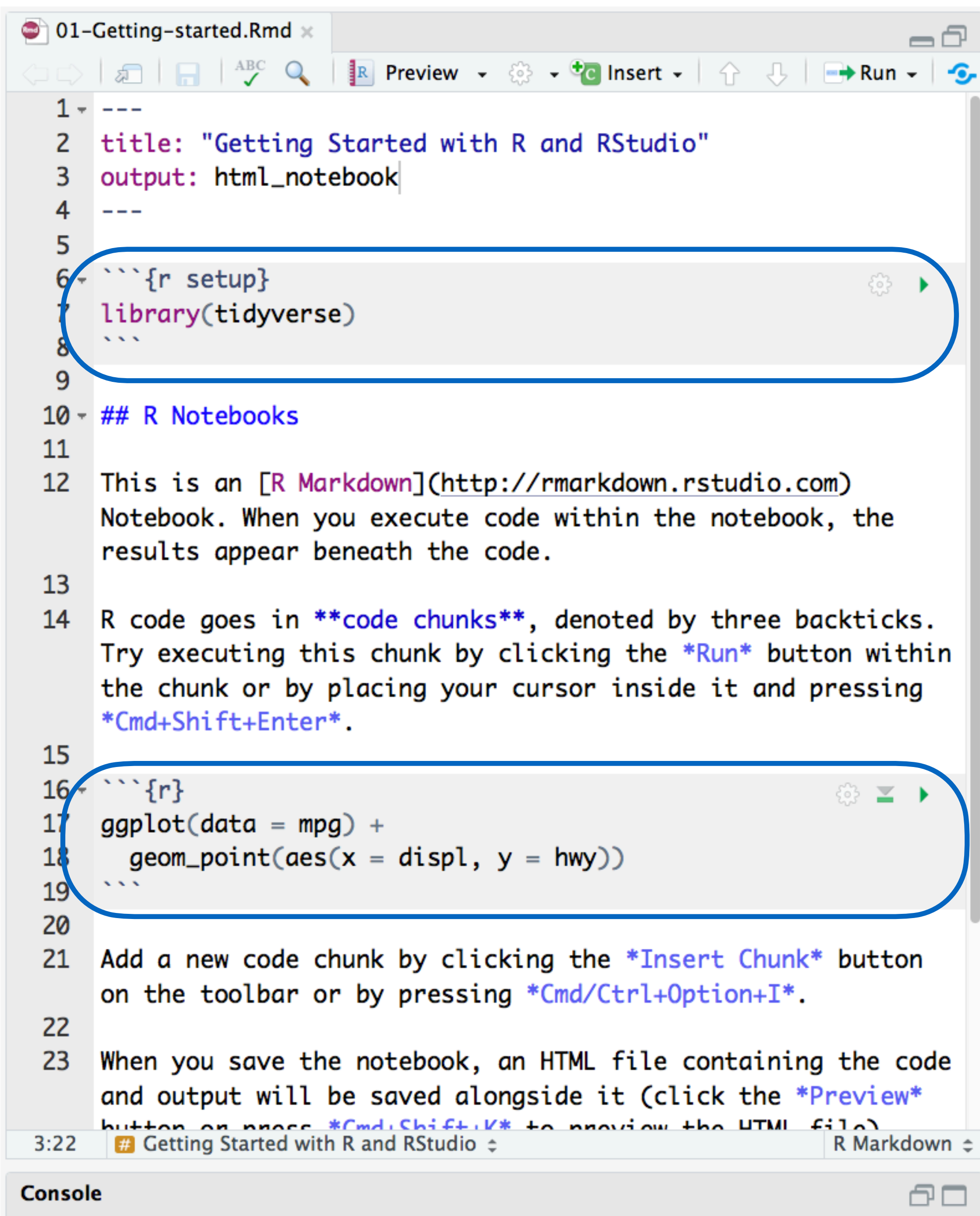


The screenshot shows the RStudio interface with a file named '01-Getting-started.Rmd'. The editor displays an R Markdown document with the following content:

```
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 bottom of the window shows a status bar with the time '3:22', the file name '# Getting Started with R and RStudio', and the document type 'R Markdown'.





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 title, output format, and two code chunks. The first code chunk is enclosed in a blue rounded rectangle and contains R code to set up the tidyverse. The second code chunk is also enclosed in a blue rounded rectangle and contains R code to create a scatter plot using ggplot2. The notebook text explains that R code is executed within code chunks and that the results appear below the code. It also provides instructions on how to add new code chunks and how to preview the HTML output.

```
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

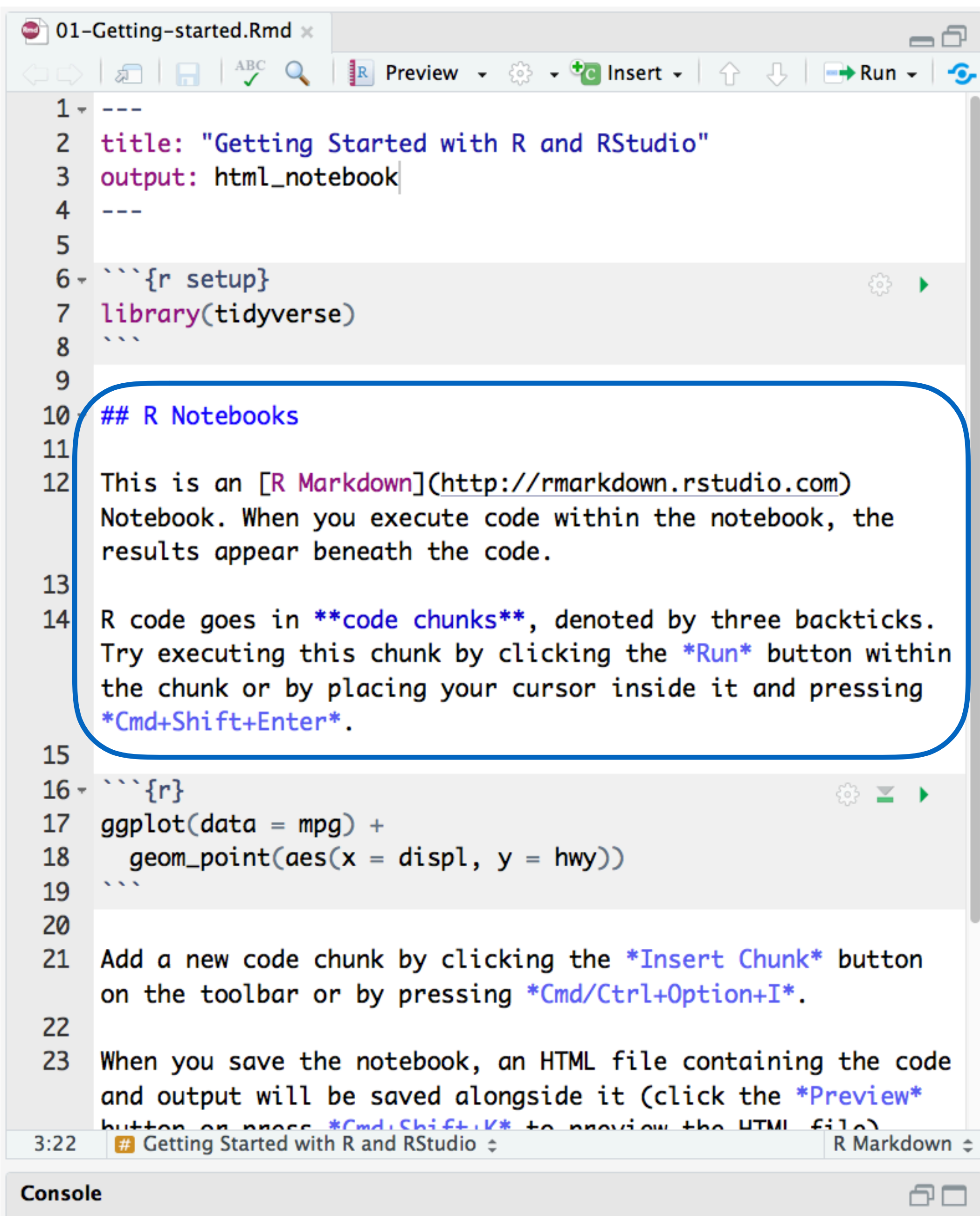
# R notebooks

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 file type "R Markdown". A "Console" panel is visible at the very bottom.

# R notebooks

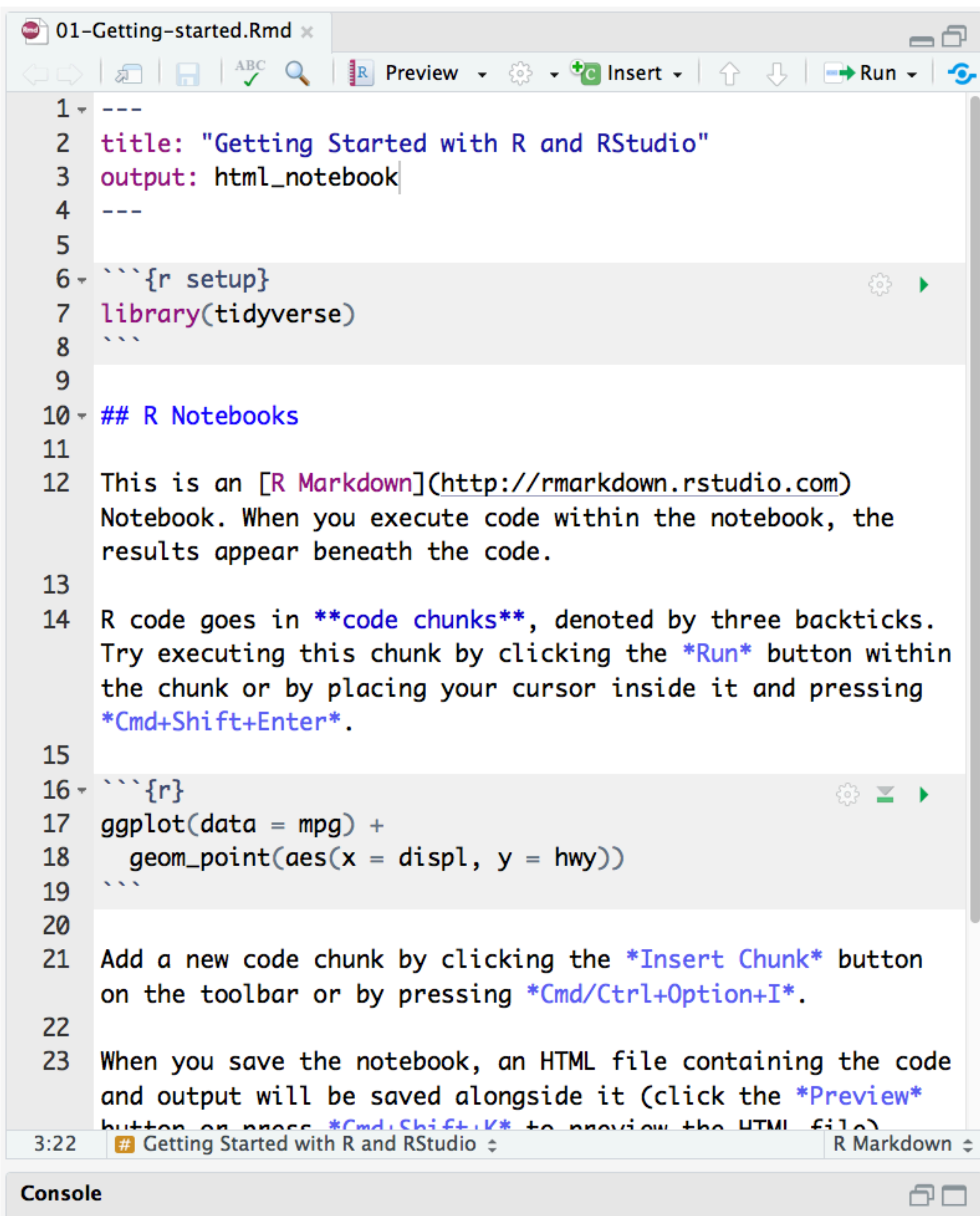
An authoring format for  
Data Science

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 chunks, and running code. The notebook content includes a title, output format, a code chunk for setting up the tidyverse library, a section header "## R Notebooks", and explanatory text about R Markdown and code chunks. A second code chunk contains a ggplot2 command. 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

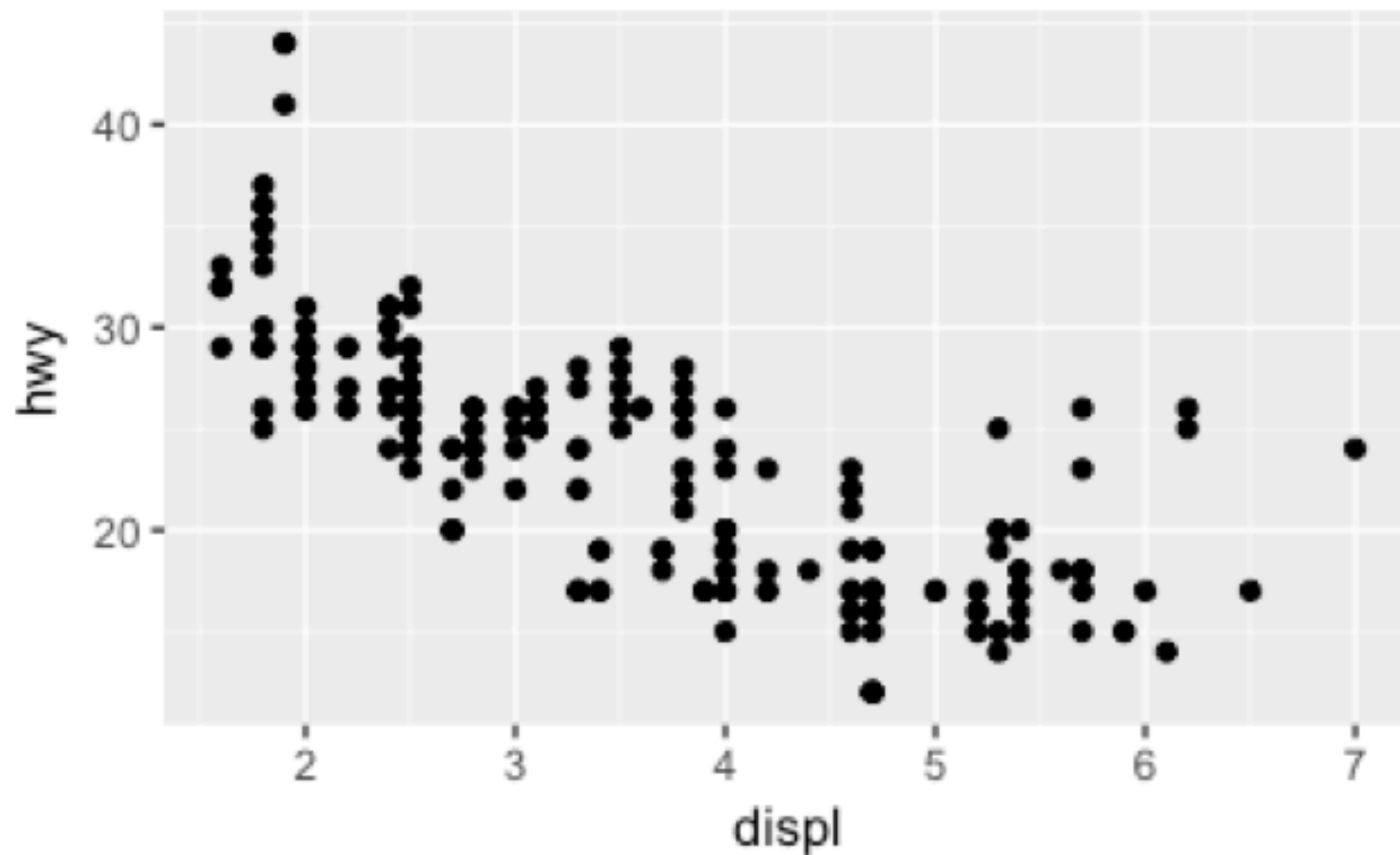
```
```{r}
ggplot(data = mpg) +
  geom_point(aes(x = displ, y = hwy))
```
```



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