Data Science in the tidyverse

Hadley Wickham

hadley.nz

@hadleywickham

Introduction

HELLO

my name is

Hadley

Your Turn

Introduce yourselves to your neighbours:

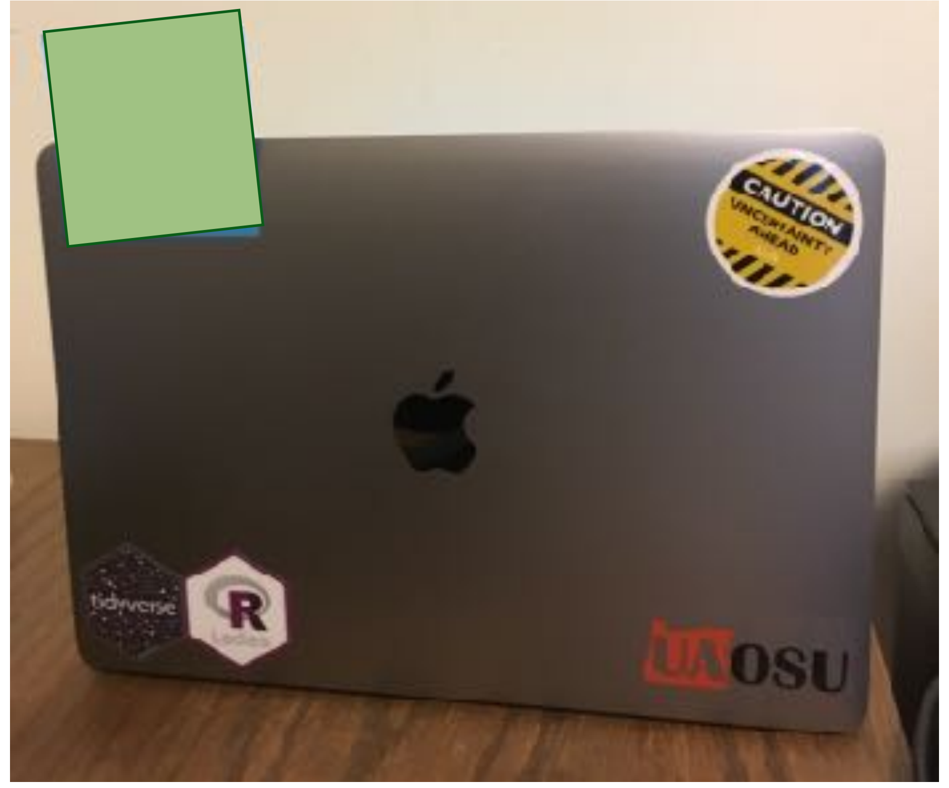
Who are you?

What do you do with data?

How would you describe your R experience?

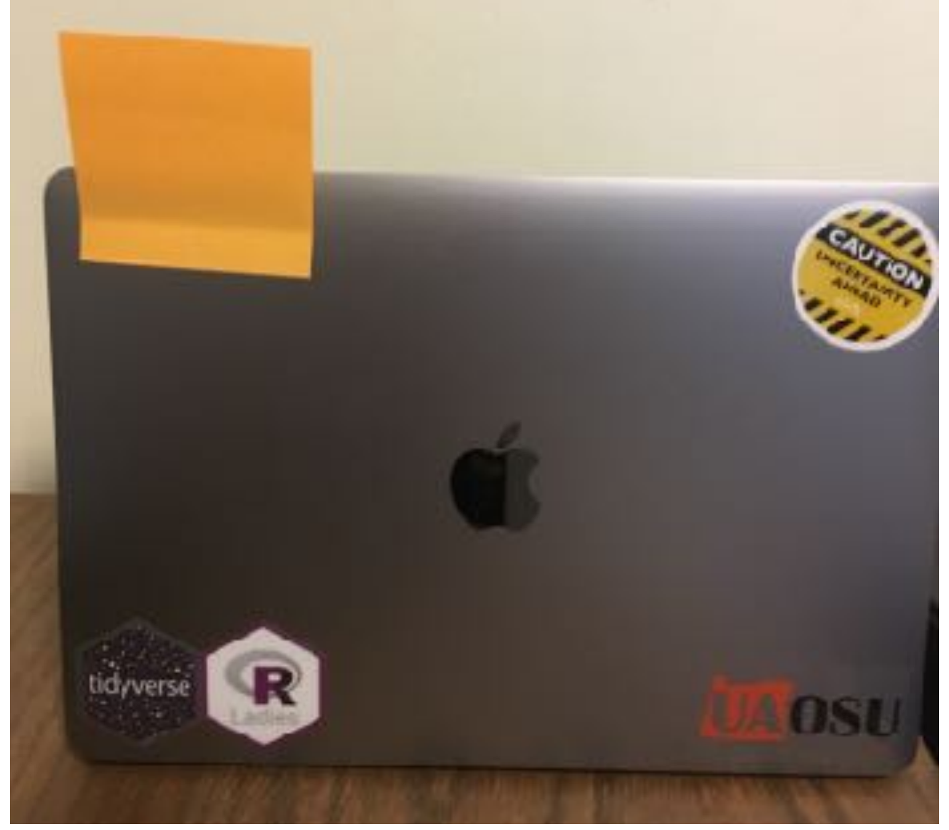
No sticky note: "I'm happily working on it"

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



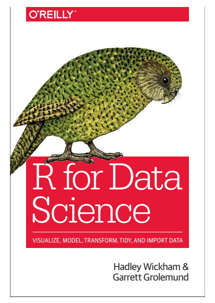
Hopefully, color-blind friendly, let me know if not.

This class is heavily based on

R for Data Science

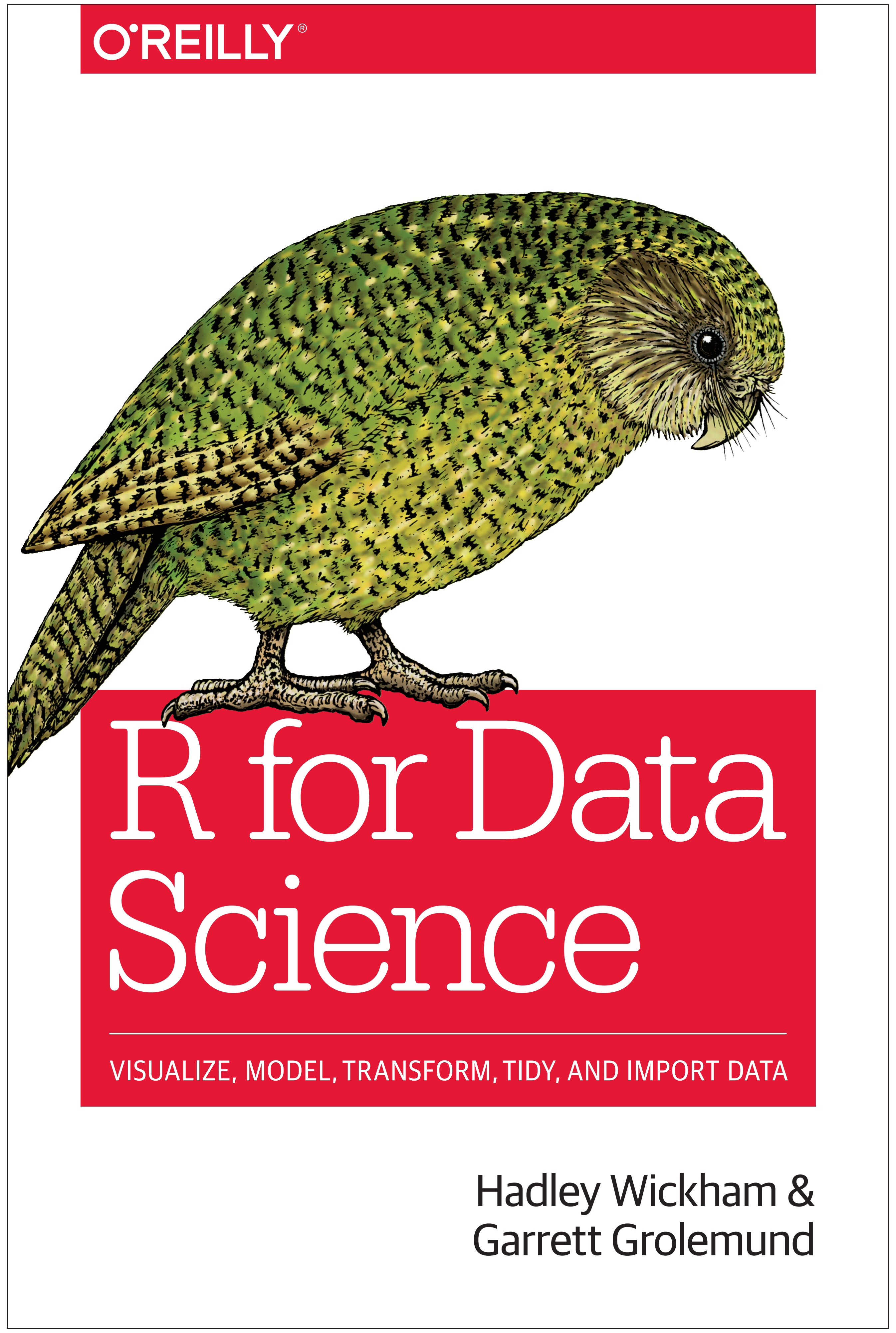
http://r4ds.had.co.nz/

Links to the relevant sections of the book

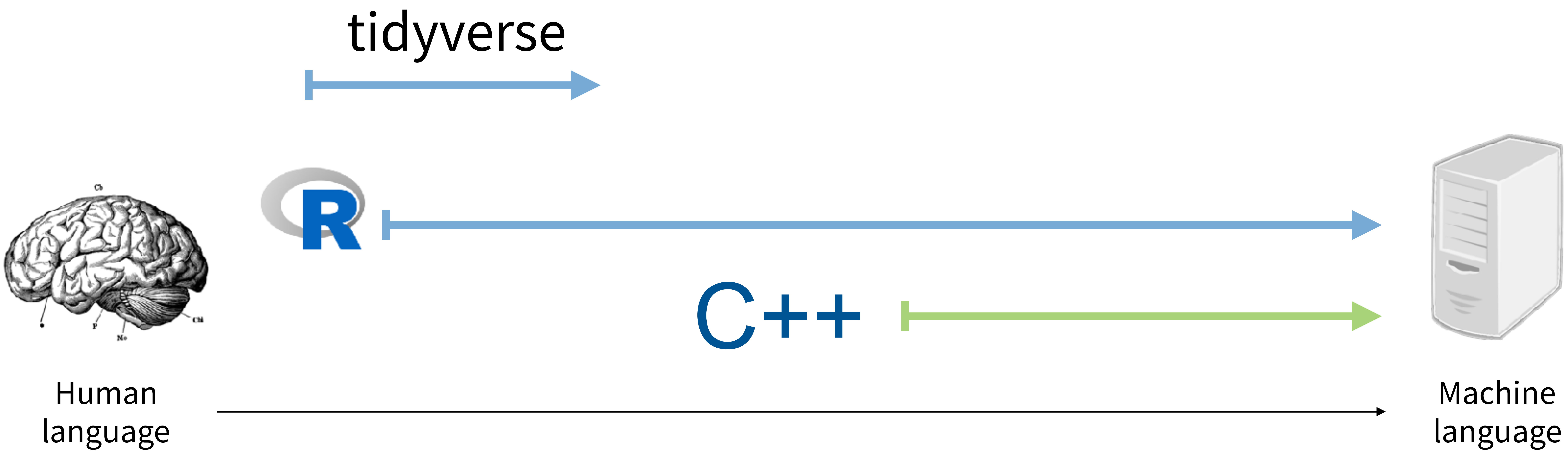


In R4DS

Introduction



R - A programming language for data



tidyverse

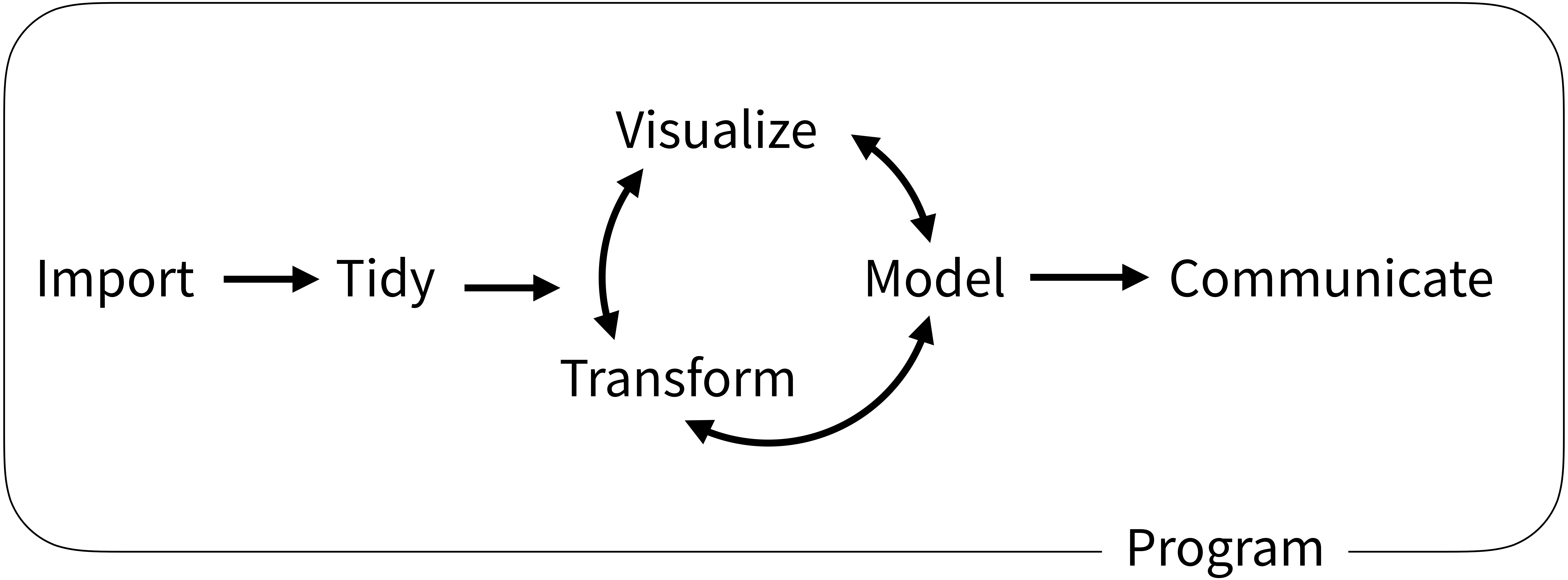
C++

Human language

Machine language

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

(Applied) Data Science


Import

Tidy

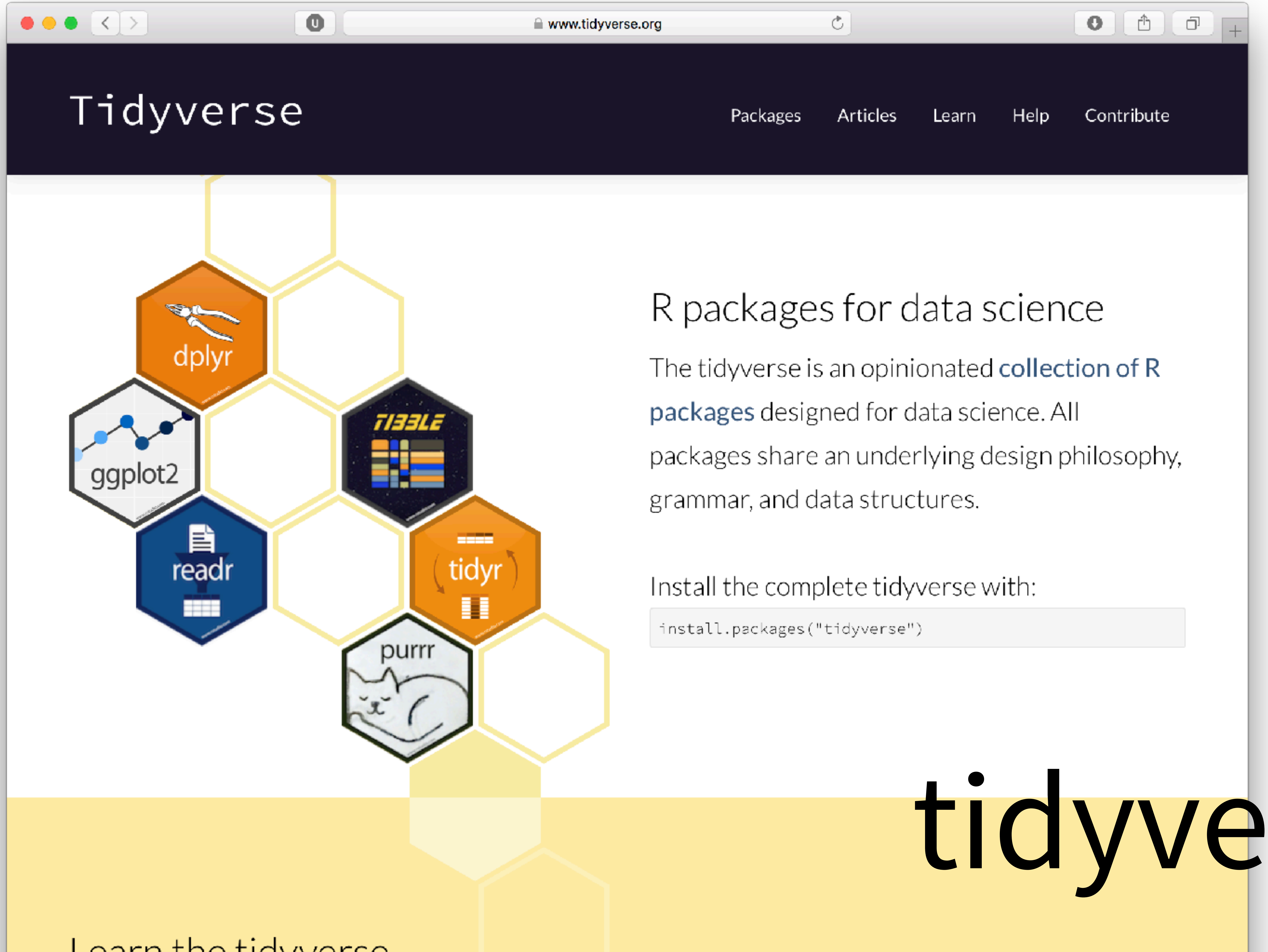
Visualize

Transform

Model

Communicate

Program



R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

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

Learn the tidyverse

tidyverse.org

tidyverse

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

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

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

does the equivalent of

```r
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

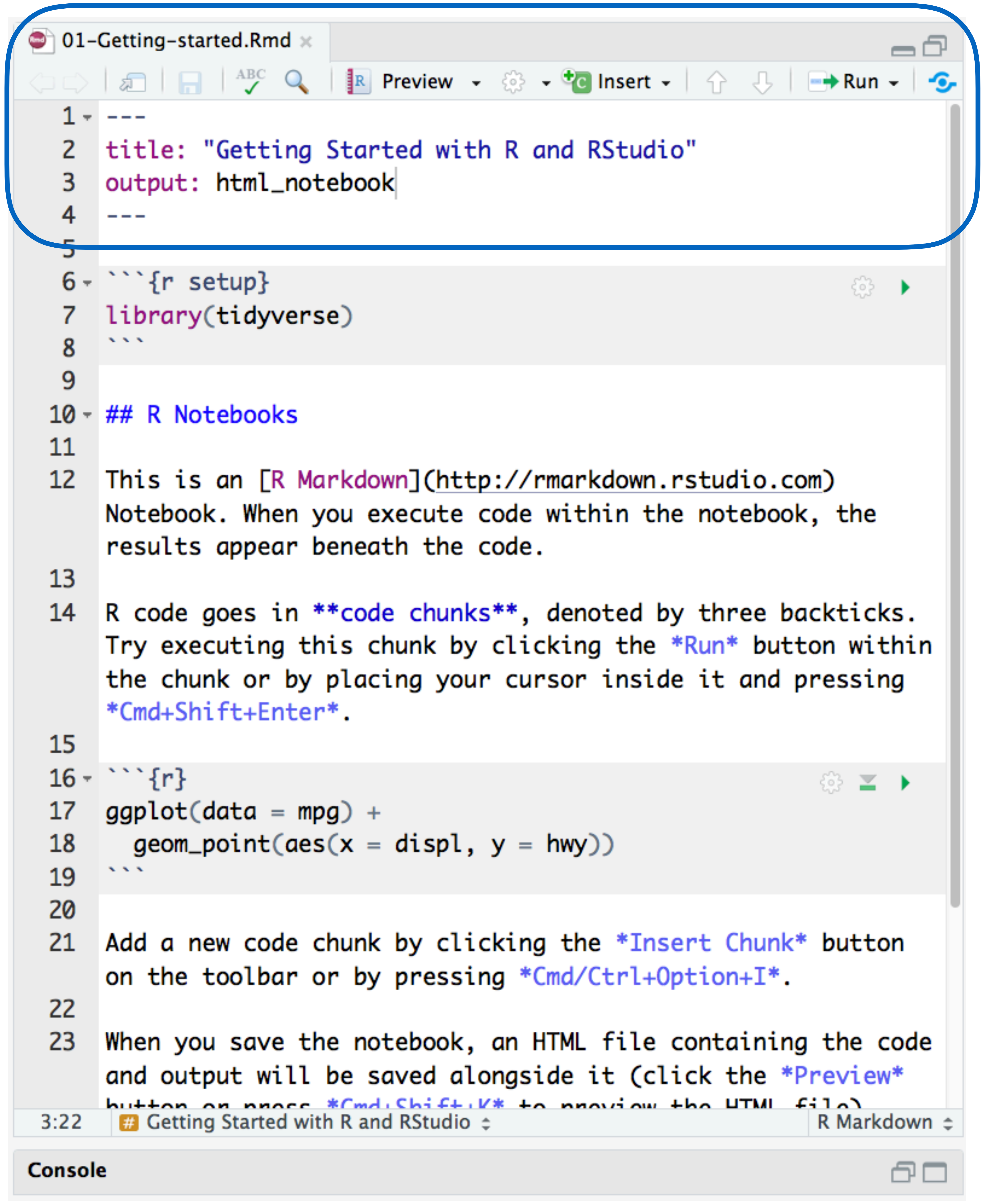
Data Types	9:00 - 10:30
Morning Break	10:30 - 11:00
Iteration	11:00 - 12:30
Lunch	12:30 - 1:30
Modelling	1:30 - 3:00
Afternoon Break	3:00 - 3:30
Organization with list columns	3:30 - 5:00

Getting Started

Your Turn

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

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

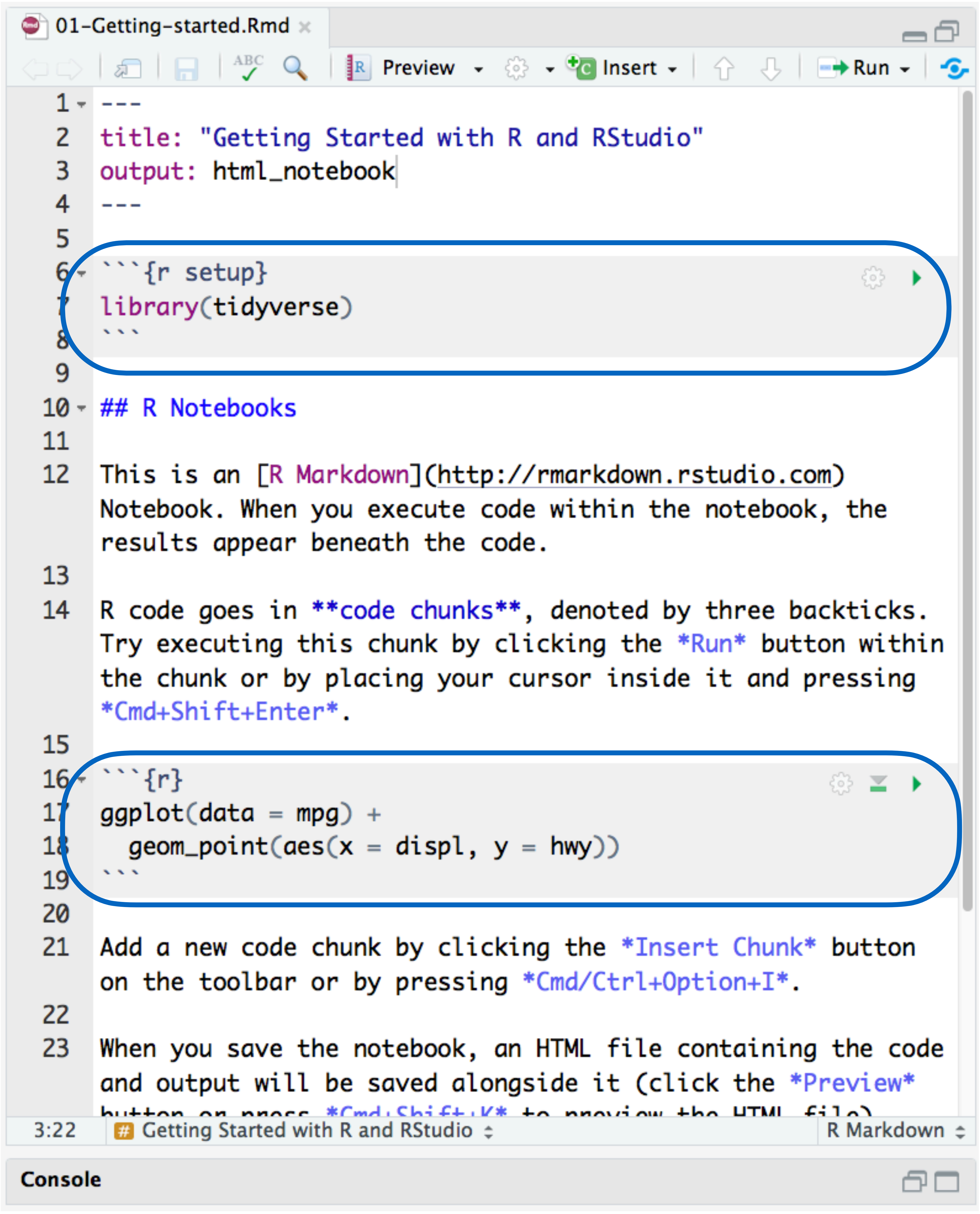


~~~
---
title: "Getting Started with R and RStudio"
output: html_notebook
---

```{r setup}
library(tidyverse)
```

## R Notebooks

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

R code goes in **code chunks**, denoted by three backticks. Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

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

Add a new code chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd/Ctrl+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button 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



~~~
---
title: "Getting Started with R and RStudio"
output: html_notebook
---

```{r setup}
library(tidyverse)
```

## R Notebooks

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

R code goes in **code chunks**, denoted by three backticks. Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

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

Add a new code chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd/Ctrl+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file)
~~~

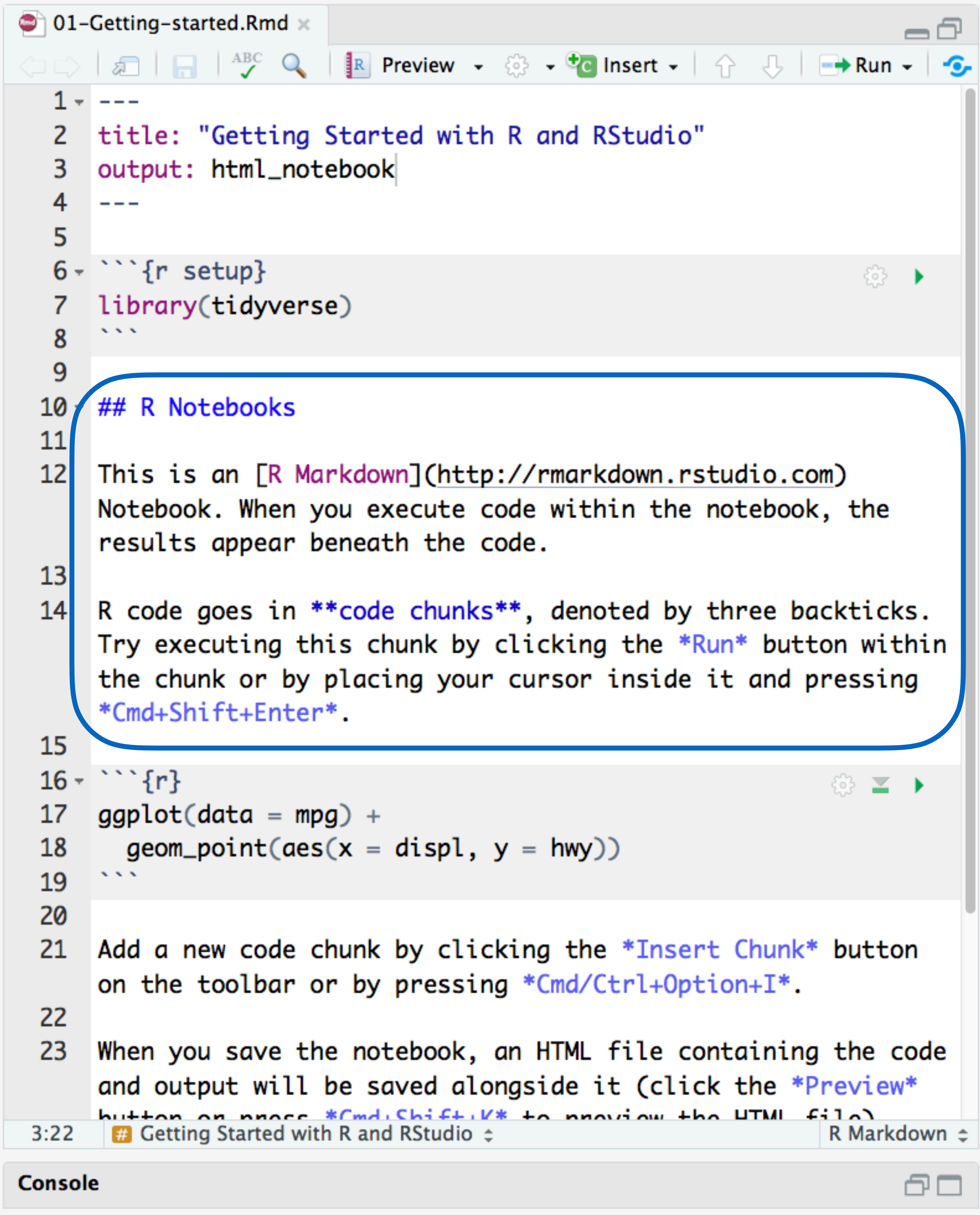
R notebooks

An authoring format for Data Science

00-Getting-started.Rmd is an R notebook

Integrates:

- Code



~~~
---
title: "Getting Started with R and RStudio"
output: html_notebook
---

```{r setup}
library(tidyverse)
~~~

R Notebooks

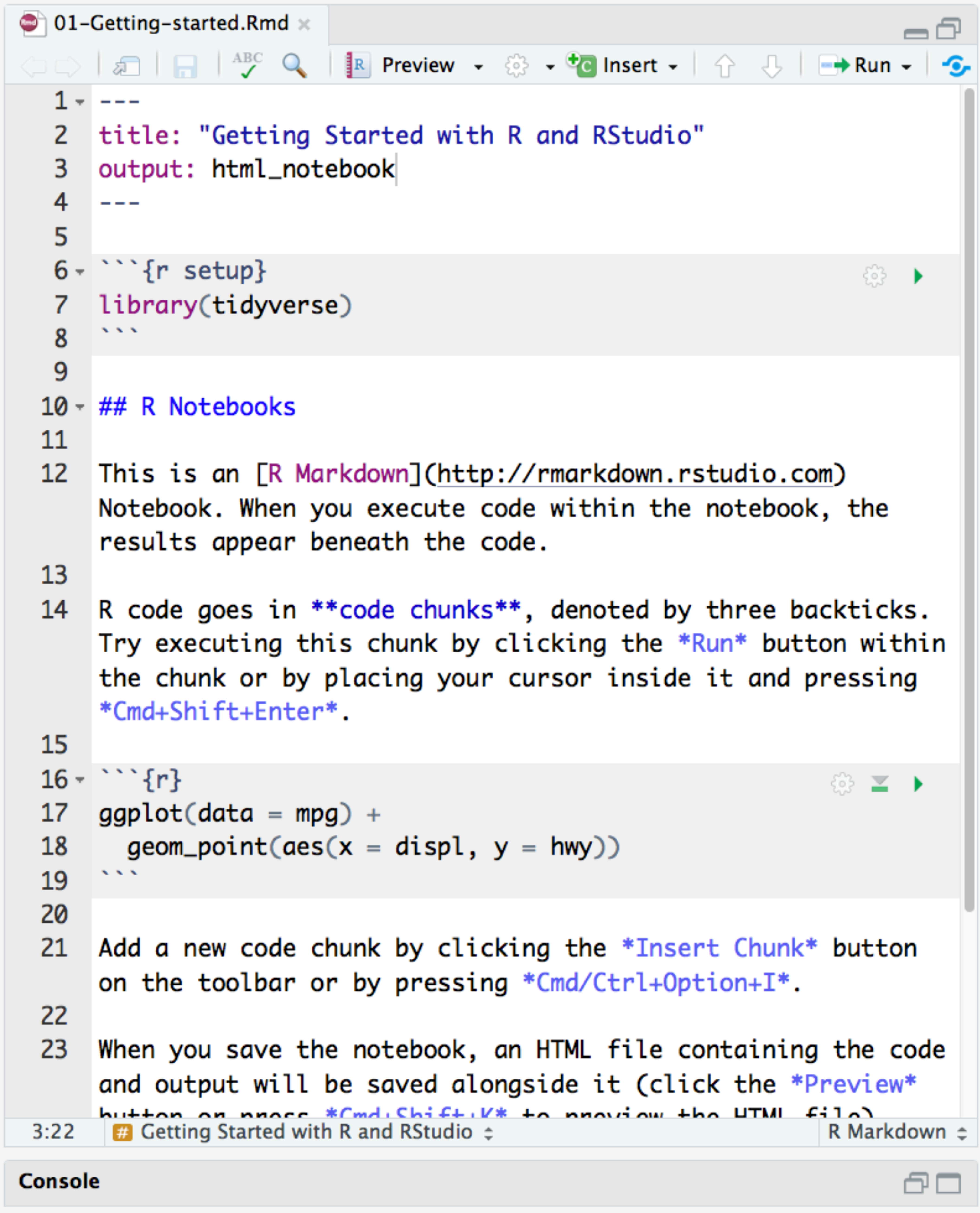
This is an [R Markdown](http://rmarkdown.rstudio.com)
Notebook. When you execute code within the notebook, the
results appear beneath the code.

R code goes in **code chunks**, denoted by three backticks.
Try executing this chunk by clicking the *Run* button within
the chunk or by placing your cursor inside it and pressing
Cmd+Shift+Enter.

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

Add a new code chunk by clicking the *Insert Chunk* button
on the toolbar or by pressing *Cmd/Ctrl+Option+I*.

When you save the notebook, an HTML file containing the code
and output will be saved alongside it (click the *Preview*
button 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
```



~~~
---
title: "Getting Started with R and RStudio"
output: html_notebook
---

```{r setup}
library(tidyverse)
~~~

## R Notebooks

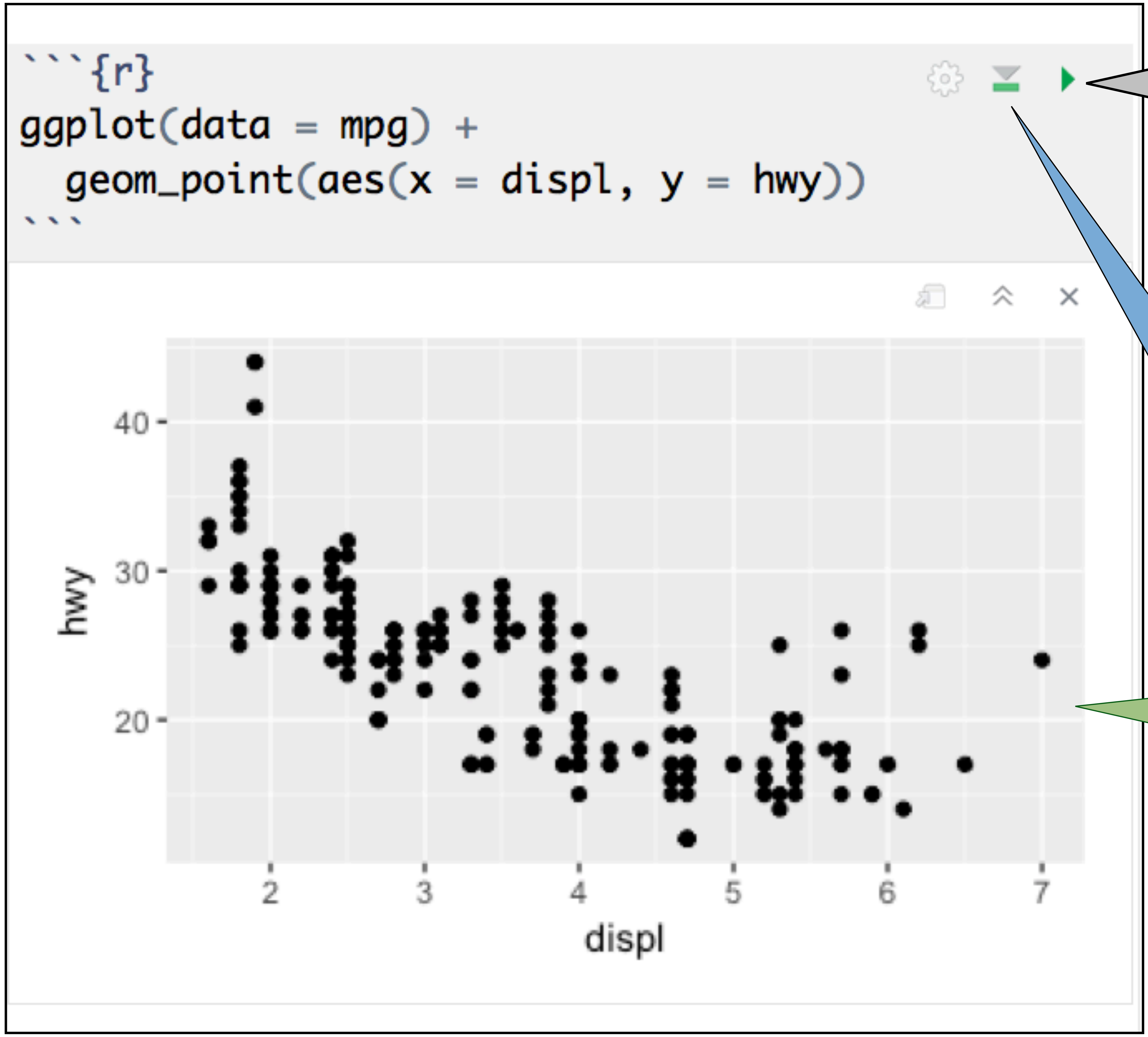
This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

R code goes in **code chunks**, denoted by three backticks. Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

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

Add a new code chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd/Ctrl+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button 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