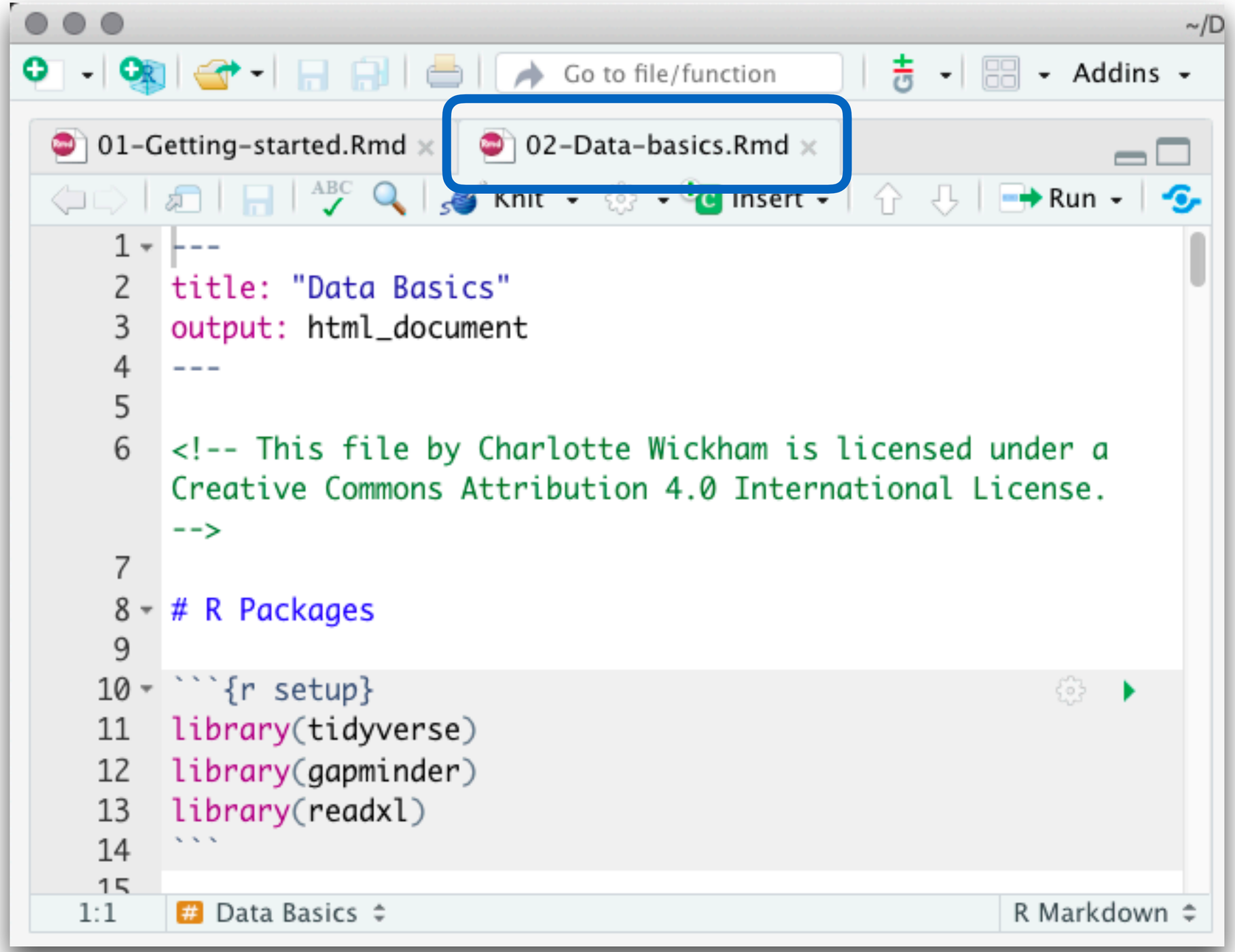
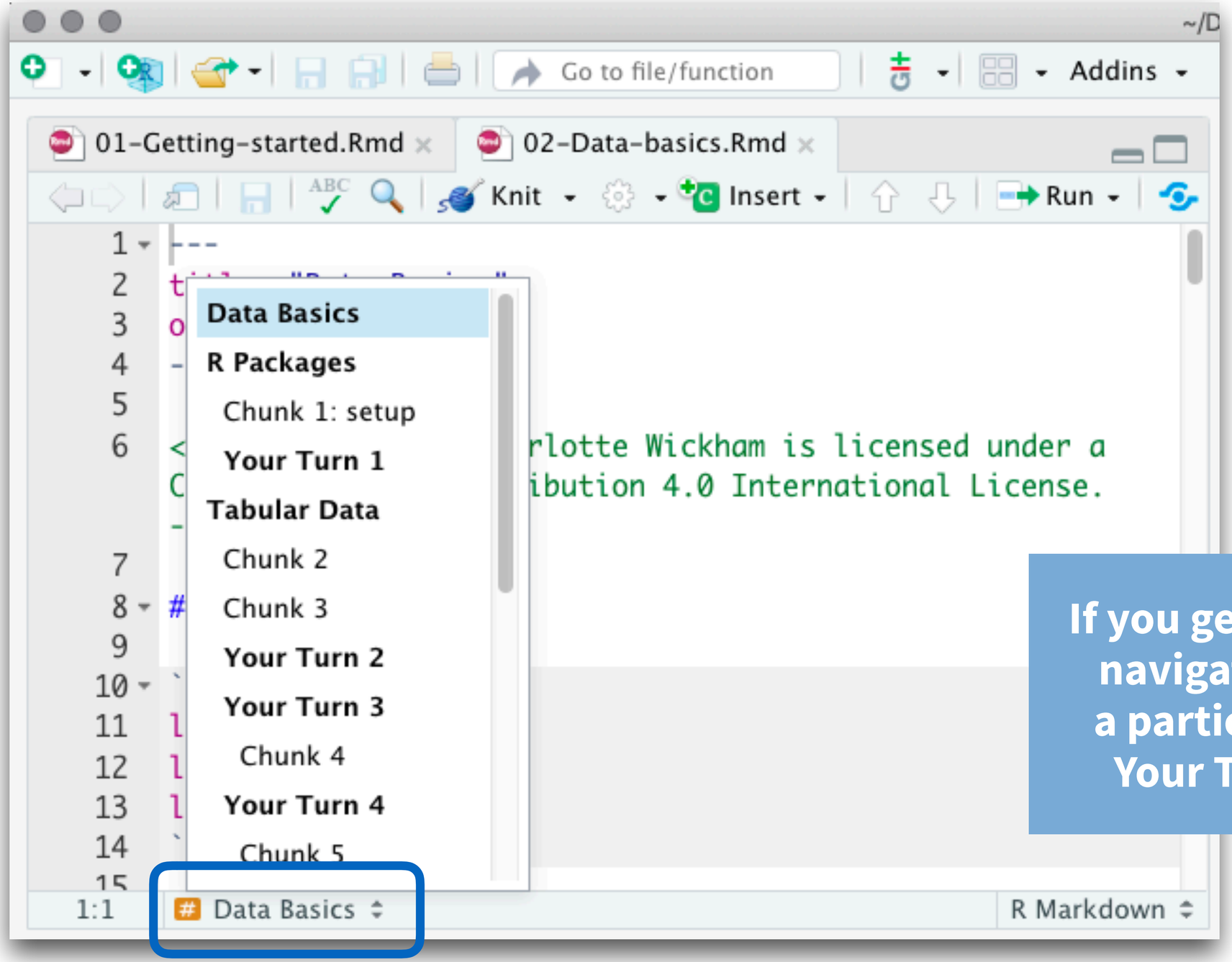
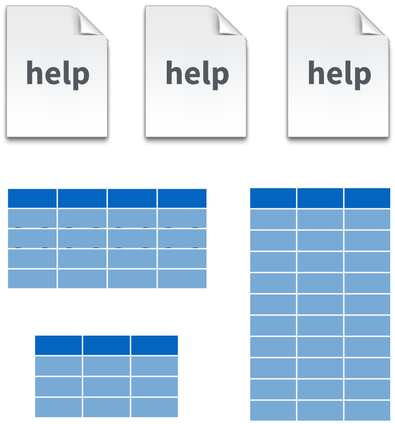
Data Basics


~/D
Go to file/function
Addins
01-Getting-started.Rmd
02-Data-basics.Rmd
Knit
Insert
Run
1 ---
2 title: "Data Basics"
3 output: html_document
4 ---
5
6 <!-- This file by Charlotte Wickham is licensed under a
Creative Commons Attribution 4.0 International License.
-->
7
8 # R Packages
9
10 ```{r setup}
11 library(tidyverse)
12 library(gapminder)
13 library(readxl)
14 ```
1:1
Data Basics
R Markdown

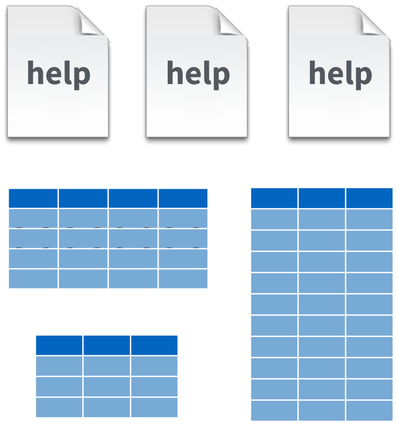


If you get lost: navigate to a particular Your Turn

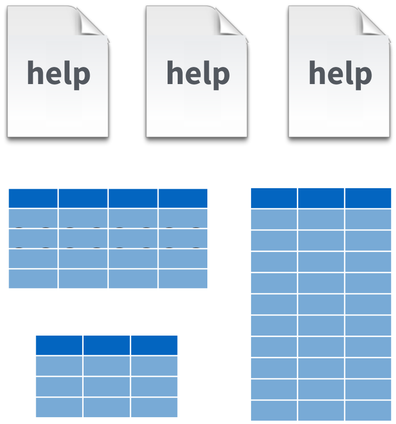
R Packages



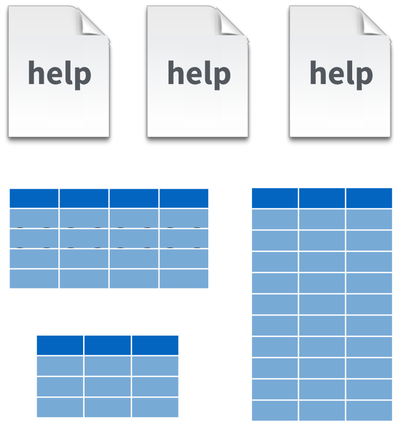
function1()
function2()
function3()
function4()



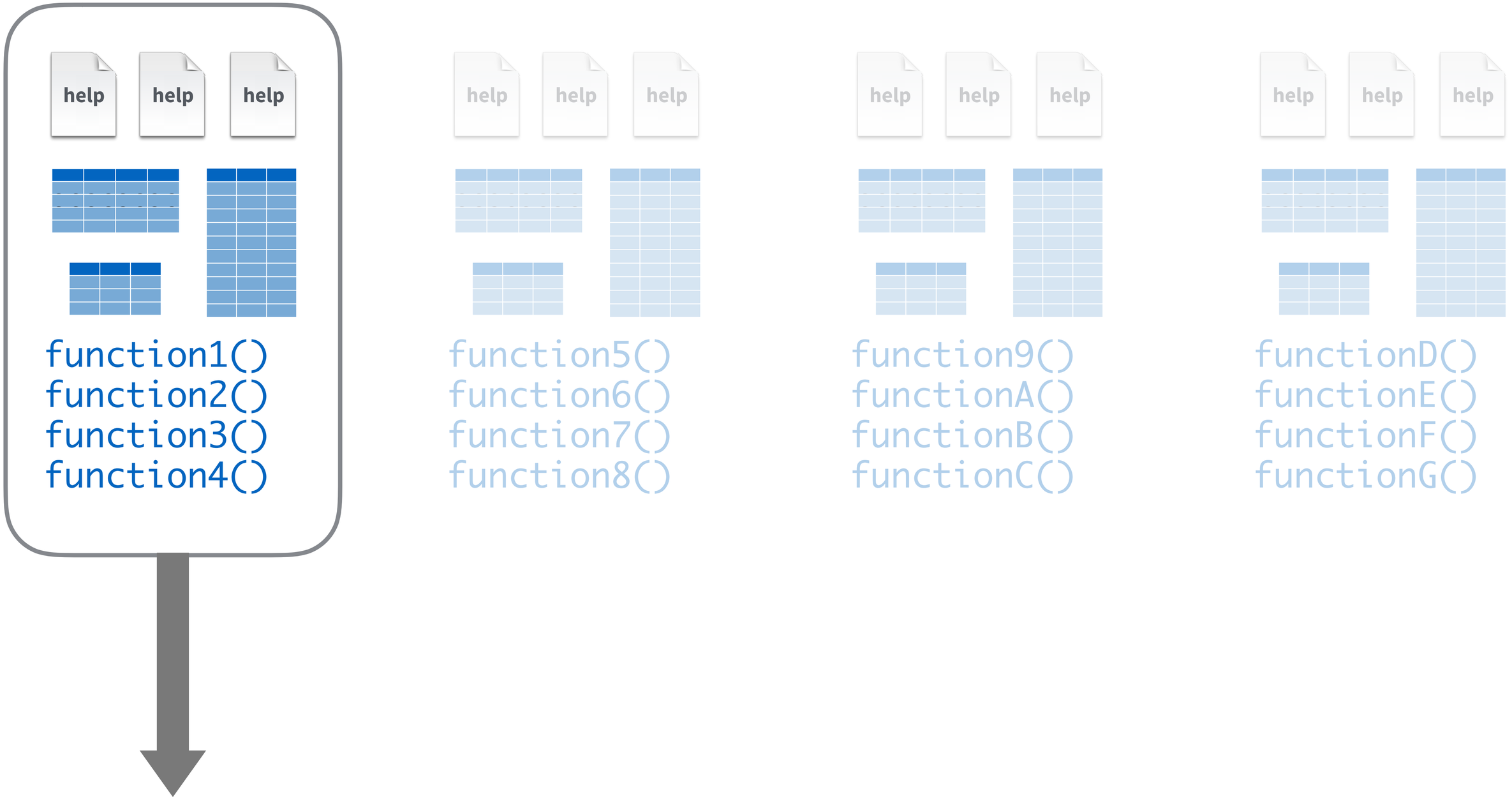
function5()
function6()
function7()
function8()



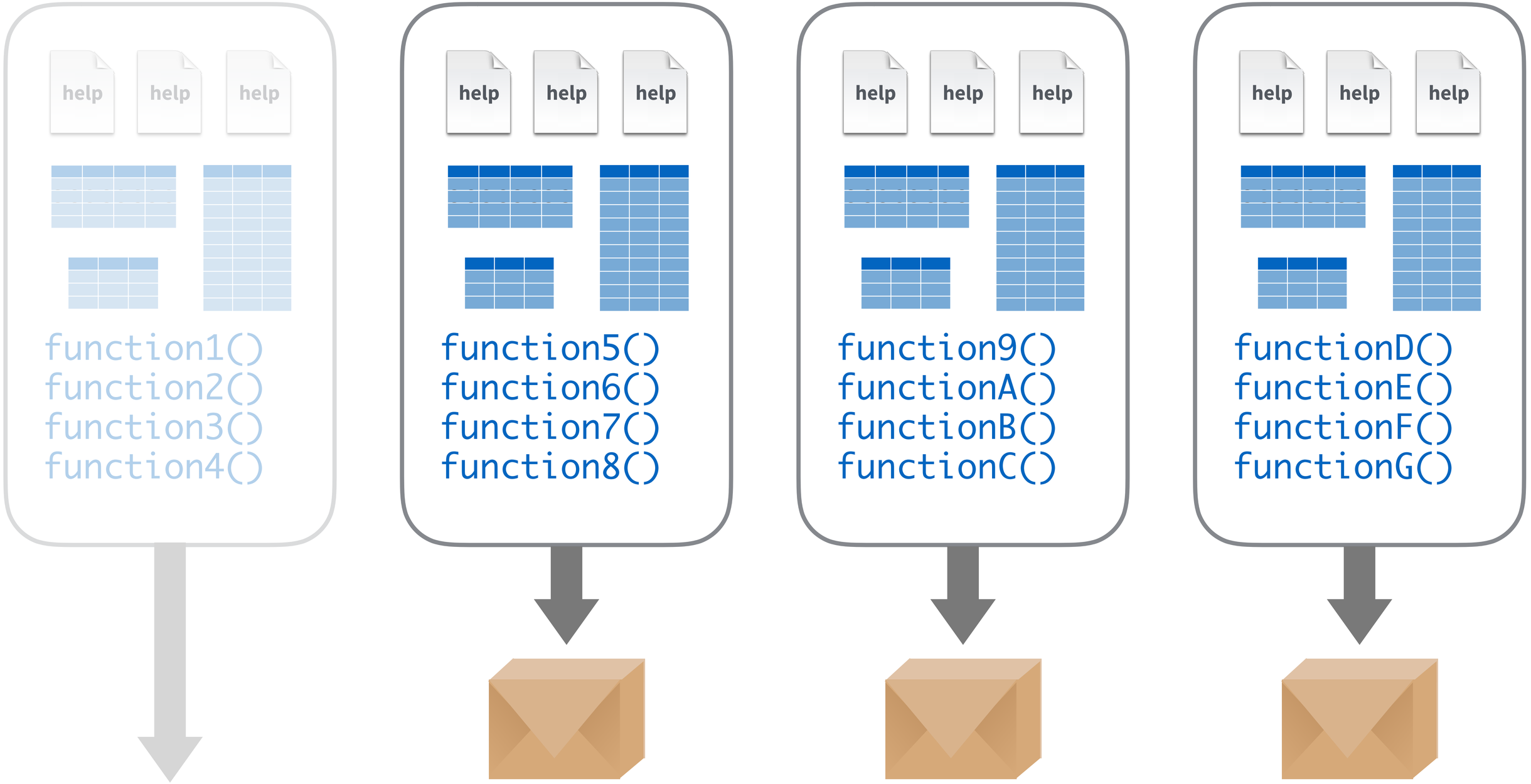
function9()
functionA()
functionB()
functionC()



functionD()
functionE()
functionF()
functionG()


help
help
help
function1()
function2()
function3()
function4()
help
help
help
function5()
function6()
function7()
function8()
help
help
help
function9()
functionA()
functionB()
functionC()
help
help
help
functionD()
functionE()
functionF()
functionG()

Base R



Base R

R Packages

Using packages

1

```
install.packages("foo")
```

Downloads files to computer

1 x per computer

2

```
library("foo")
```

Loads package

1 x per R Session

We did this for you on rstudio.cloud

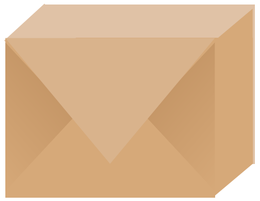
Your Turn 1

With your neighbor:

What R packages are being loaded in the first chunk of 02-Data-Basics.Rmd?

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

tidyverse



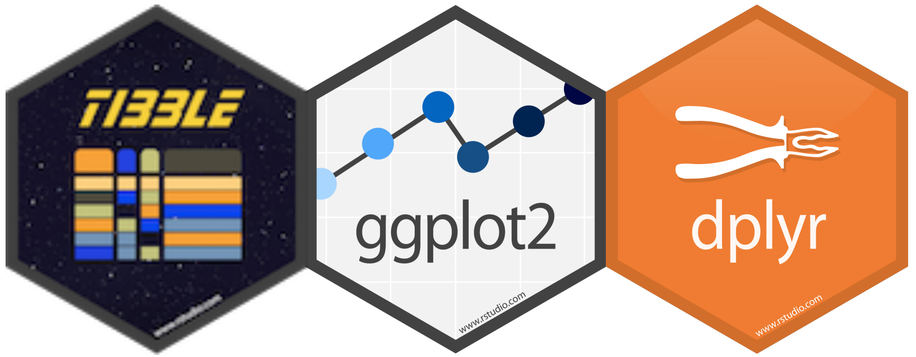
An R package that serves as a short cut for installing and loading the components of the tidyverse.

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

What is the tidyverse?

"...the tidyverse makes data science faster, easier and more fun ..."

"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.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

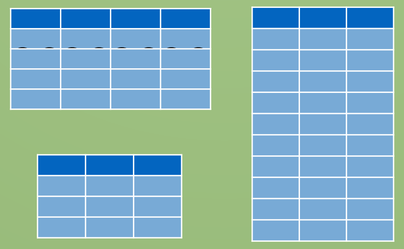
```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("stringr")
install.packages("forcats")
install.packages("lubridate")
install.packages("hms")
install.packages("DBI")
install.packages("haven")
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

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("stringr")
install.packages("forcats")
```

Tabular Data



Data frames and tibbles

The most common kind of data objects, for rectangular data

Data frames - a type of object native to R

Tibbles - a.k.a `tbl` - a type of data frame common in the tidyverse

Tibbles have slightly different default behaviour than data frames, but in R markdown you mostly won't notice a difference.

Your Turn 2

Take a look at the mpg dataset in two ways:

1. Run `mpg` in the code chunk

2. Type `mpg` on the Console and hit Enter

What do you notice about the difference in the way they are displayed?

`mpg` in an R markdown code chunk:

1. Run the code in this chunk:

```{r}
mpg
```

manufacturer <chr>	model <chr>	displ <dbl>	year <int>	cyl <int>
audi	a4	1.8	1999	4
audi	a4	1.8	1999	4
audi	a4	2.0	2008	4
audi	a4	2.0	2008	4
audi	a4	2.8	1999	6
audi	a4	2.8	1999	6
audi	a4	3.1	2008	6
audi	a4 quattro	1.8	1999	4
audi	a4 quattro	1.8	1999	4
audi	a4 quattro	2.0	2008	4

1-10 of 234 rows | 1-5 of 1... Previous 1 2 3 4 5 ... 24 Next

`mpg` in the Console:

```
> mpg
# A tibble: 234 x 11
   manufacturer model displ  year   cyl trans drv     cty   hwy fl
   <chr>        <chr> <dbl> <int> <int> <chr> <chr> <int> <int> <chr>
 1 audi         a4      1.8  1999     4 auto… f        18    29 p
 2 audi         a4      1.8  1999     4 manu… f        21    29 p
 3 audi         a4      2    2008     4 manu… f        20    31 p
 4 audi         a4      2    2008     4 auto… f        21    30 p
 5 audi         a4      2.8  1999     6 auto… f        16    26 p
 6 audi         a4      2.8  1999     6 manu… f        18    26 p
 7 audi         a4      3.1  2008     6 auto… f        18    27 p
 8 audi         a4 q…   1.8  1999     4 manu… 4        18    26 p
 9 audi         a4 q…   1.8  1999     4 auto… 4        16    25 p
10 audi         a4 q…   2    2008     4 manu… 4        20    28 p
# … with 224 more rows, and 1 more variable: class <chr>
>
```

Your Turn 3

Run the code in the chunk line by line with shortcut Crtl/Cmd + Enter

```
dim(x = mpg)
```

```
names(x = mpg)
```

```
glimpse(x = mpg)
```

```
View(x = mpg)
```

What do each of these functions do?

Getting an overview of data

```
dim(x = mpg)      # Dimensions of data
names(x = mpg)    # Variable names
glimpse(x = mpg) # Nice overview
View(x = mpg)     # Open Viewer pane
```

Your Turn 4

Write code in the empty chunks to find:

- The number of rows in `gapminder`
- The names of the variables in `gapminder`

```
dim(x = gapminder)
```

```
names(x = gapminder)
```

?

for help on data

mpg and gapminder are **built-in** datasets, they come with a package.

You can also use:

```
?data_name
```

to get more info on built-in data

Your Turn 5

Try

```
?mpg
```

What is this data?

Vector Data

Vectors

In R vectors are 1-dimensional arrays, that hold data all of the same type.

They can be constructed with `c()`

```
c(1, 3, 2, 1, 1)
```

But, you'll usually want to assign them to something

```
my_numbers <- c(1, 3, 2, 1, 1)
```

Basic data types

Integer	Whole numbers	`c(1L, 2L, 3L, 4L)`
Double	Numbers	`c(1, 2, 3, 4)`
Character	Text	`c("1", "2", "3", "4")`
Logical	True or False	`c(TRUE, FALSE, FALSE, TRUE)`

Your Turn 6

Take another look at mpg.

What type of data is in each column?

```
78 ```{r}
79 mpg
80 ```
```

manufacturer <chr>	model <chr>	displ <dbl>	year <int>	cyl <int>	trans <chr>	drv <chr>	cty <int>	hwy <int>	fl <chr>

Integer	<int>	`year`: 1999, 2008
Double	<dbl>	`displ`: 1.8, 2, 3.1
Character	<chr>	`model`: "a4", "camry"
Logical	<lgl>	

Importing Data

readr

Simple, consistent functions for working with (mostly) plain text data.

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

readxl

Simple, consistent functions for working Excel data

```
# install.packages("tidyverse")
library(readxl)
```

haven

Simple, consistent functions for working with SAS, SPSS and Stata data

```
# install.packages("tidyverse")
library(haven)
```

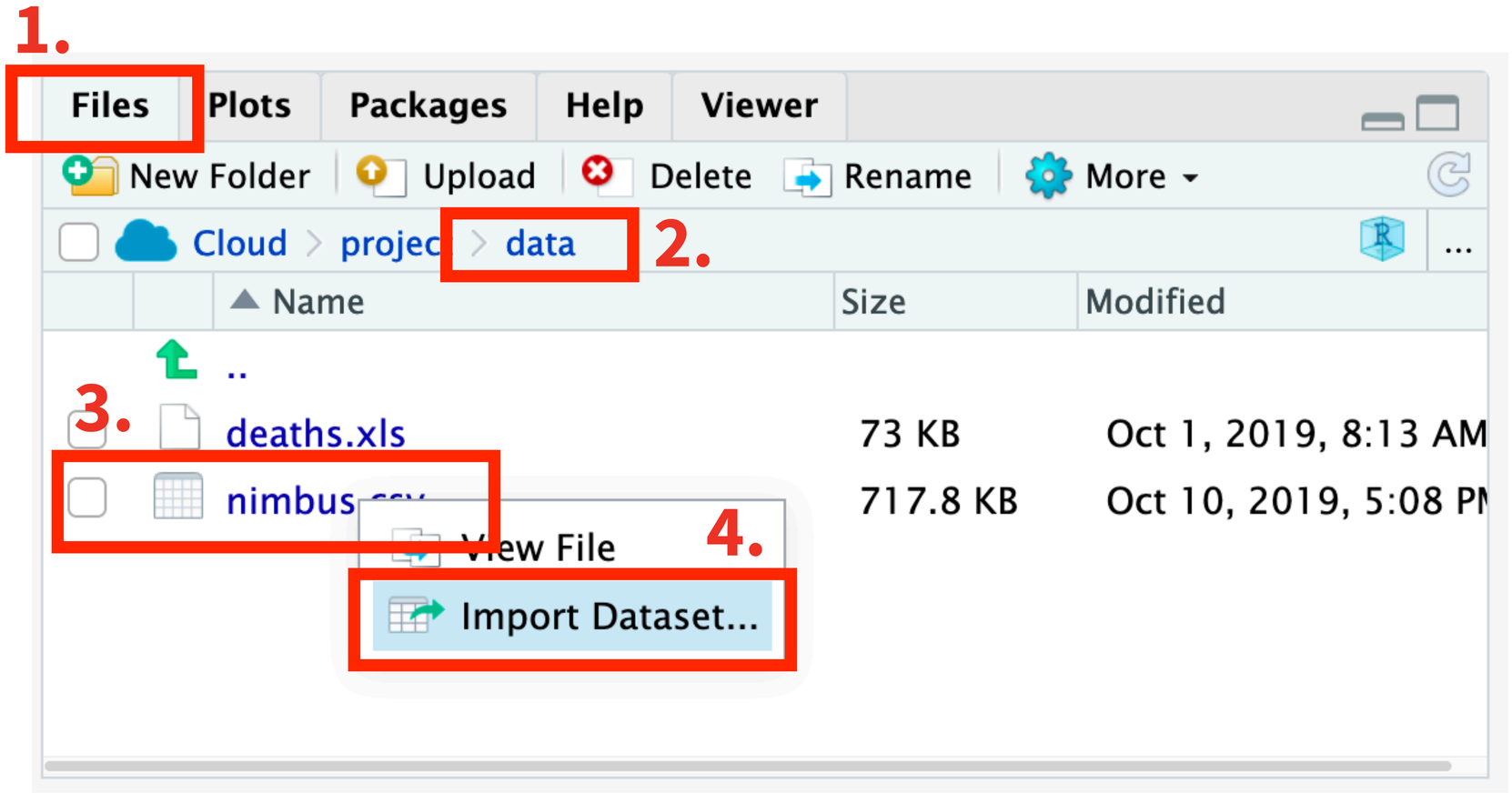
Import Dataset...

1. In the Files pane

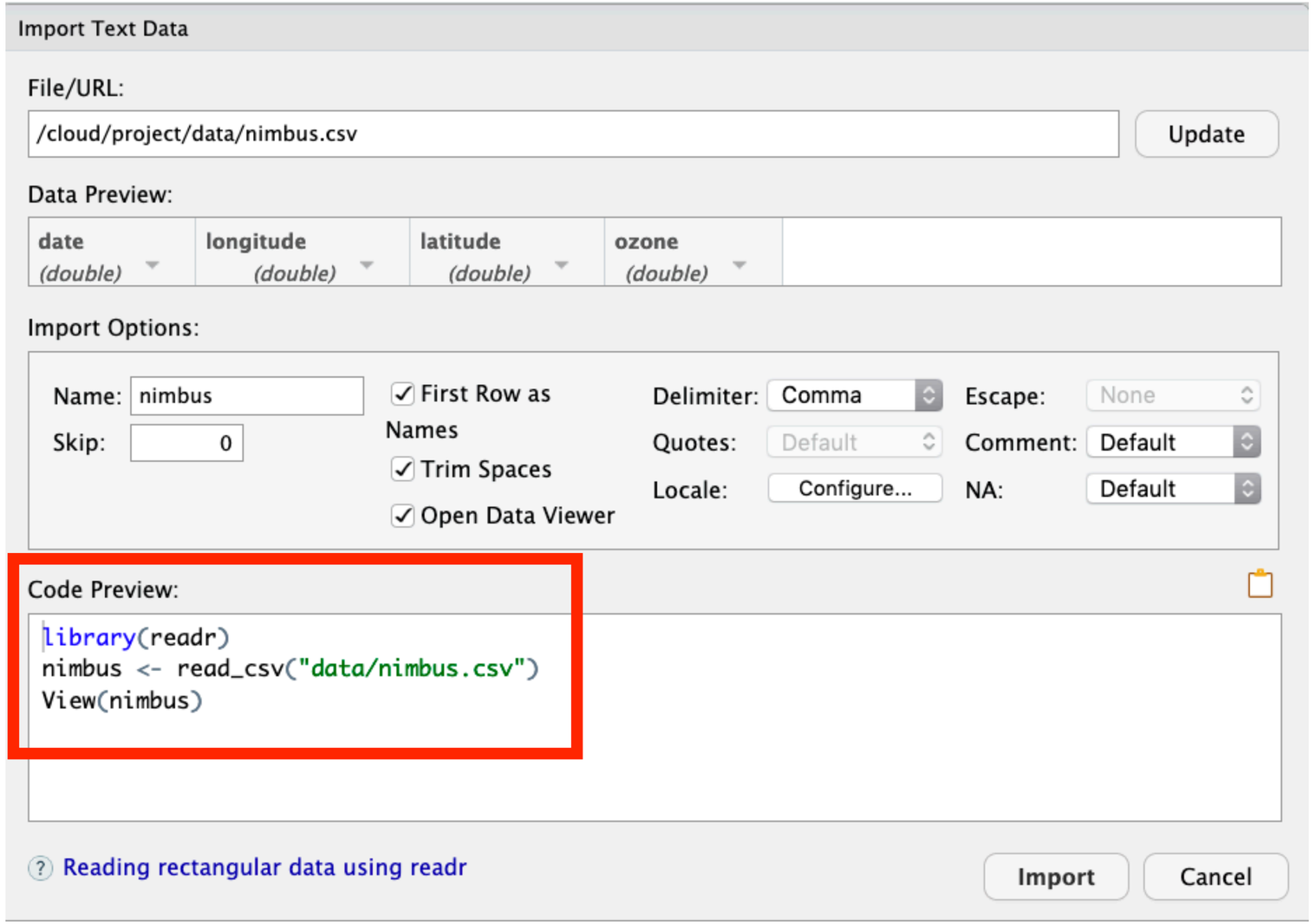
2. Navigate to the data folder

3. Click on nimbus.csv

4. Import Dataset...

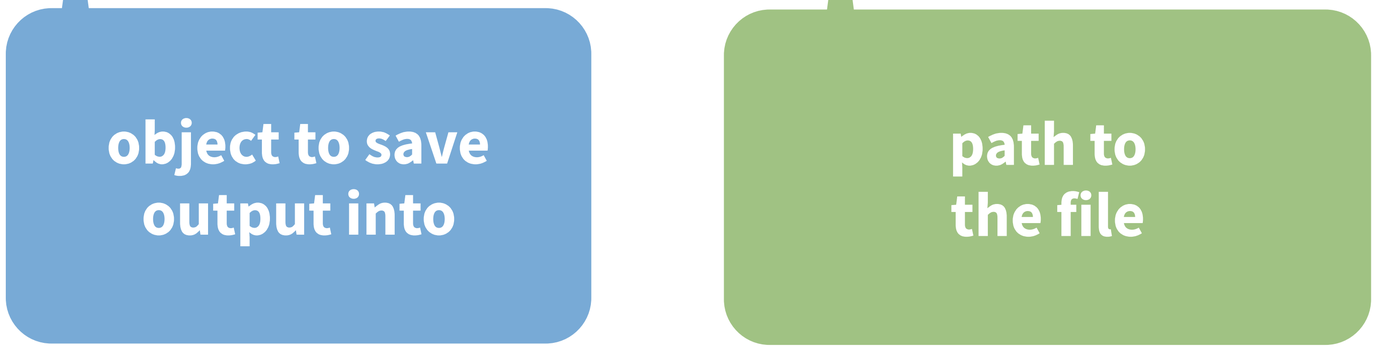


Import Dataset...



readr

```
nimbus <- read_csv("data/nimbus.csv")
```



Your Turn 7

What code do you need to read in deaths.xls?

Use the Import Data tool to help generate the code.

Challenge: Can you see a problem in the imported data? Can you Import again and fix it?

readxl

```
library(readxl)
deaths <- read_excel("data/deaths.xls", skip = 4)
```

readr

```
df <- read_csv("path/to/file.csv", ...)
```

haven

```
df <- read_spss("path/to/file.sav", ...)
```

readxl

```
df <- read_excel("path/to/file.xls", ...)
```

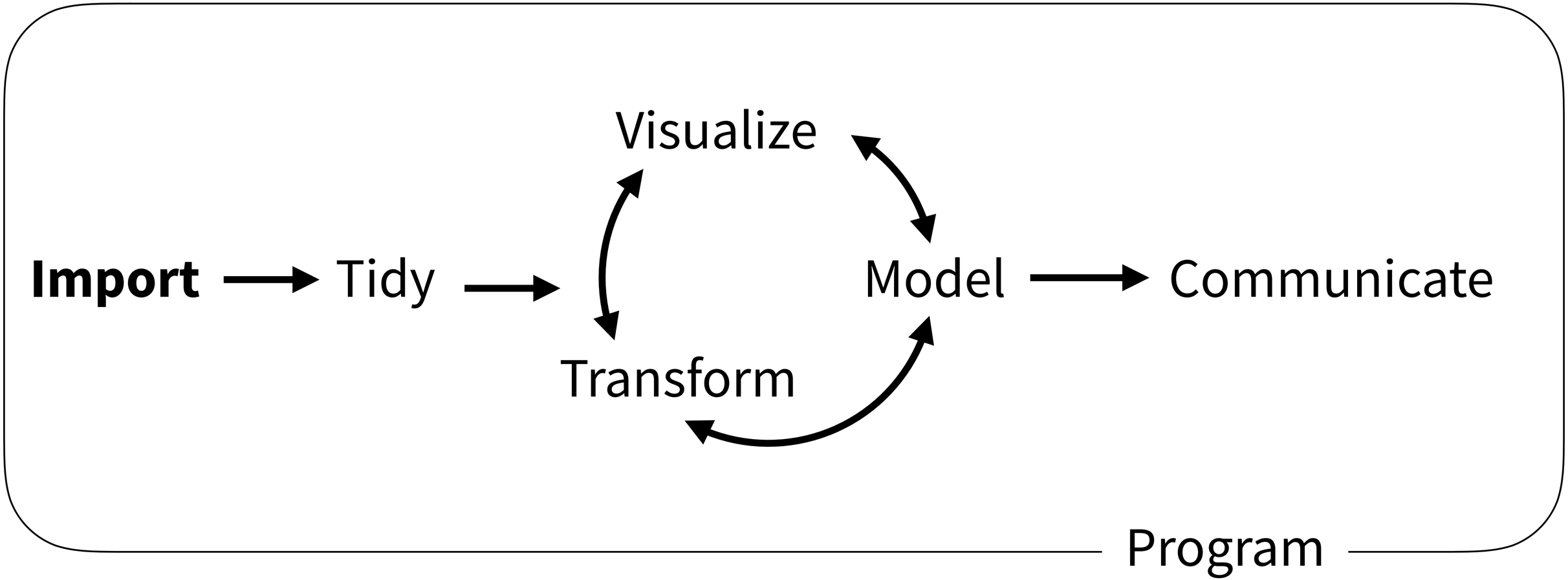
Import functions in the tidyverse have consistent syntax

Other types of data

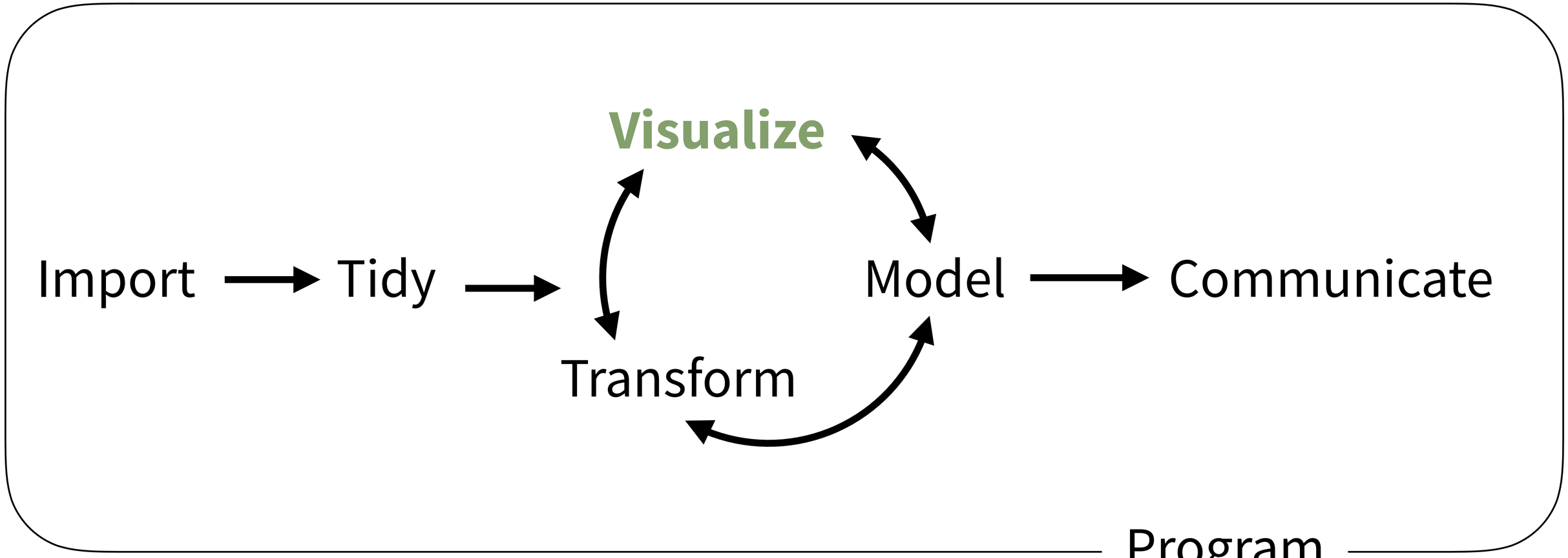
package	accesses
jsonlite	json
xml2	xml
httr	web API’s
rvest	web pages (web scraping)
DBI	databases
sparklyr	data loaded into spark

Wrapping Up

(Applied) **Data Science**



(Applied) **Data Science**



Up next...

Your Turn over lunch...

The mpg dataset has measurements on cars highway fuel efficiency (hwy) and their engine size (displ).

What relationship would you expect to see between highway fuel efficiency and engine size?