

Data Basics

The screenshot shows the RStudio interface with two R Markdown files open:

- 01-Getting-started.Rmd** (tab is not selected)
- 02-Data-basics.Rmd** (tab is selected and highlighted with a blue border)

The code editor displays the following content for the selected file:

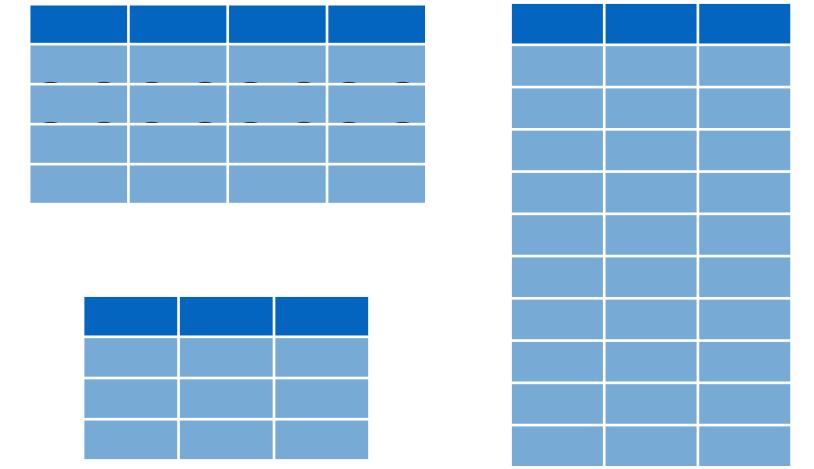
```
1 ---  
2 title: "Data Basics"  
3 output: html_document  
4 ---  
5  
6 <!-- This file by Charlotte Wickham is licensed under a  
7 Creative Commons Attribution 4.0 International License.  
8 -->  
9  
10 # R Packages  
11  
12 ```{r setup}  
13 library(tidyverse)  
14 library(gapminder)  
15 library(readxl)  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
999  
1000  
1000
```

The screenshot shows the RStudio interface. The top menu bar includes 'File', 'Edit', 'View', 'Code', 'Tools', 'Help', and 'Addins'. The toolbar contains icons for new file, new R file, open file, save, print, 'Go to file/function', a plus sign, a grid, and 'Addins'. Below the toolbar are two tabs: '01-Getting-started.Rmd' and '02-Data-basics.Rmd'. The main code editor area displays the following R Markdown code:

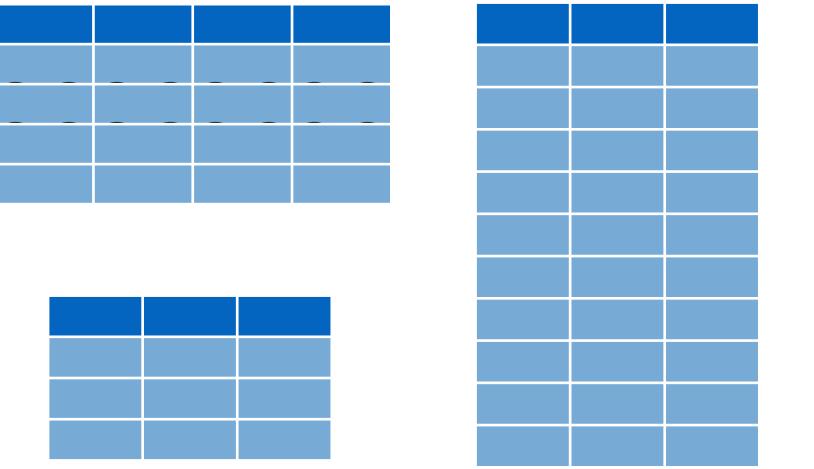
```
1 ---  
2 Data Basics  
3  
4 - R Packages  
5   Chunk 1: setup  
6 < Your Turn 1  
C  
- Tabular Data  
-  
7   Chunk 2  
8 # Chunk 3  
9 Your Turn 2  
10 Your Turn 3  
11 l  
12 l Your Turn 4  
13 l  
14   Chunk 5  
15  
1:1 # Data Basics R Markdown
```

The code editor has a vertical scrollbar on the right. To the right of the code editor, there is a large text block containing the text: "Hadley Wickham is licensed under a Creative Commons Attribution 4.0 International License." Below this text is a blue callout box with white text that reads: "If you get lost: navigate to a particular Your Turn". A blue rectangle highlights the "# Data Basics" part of the status bar at the bottom of the code editor.

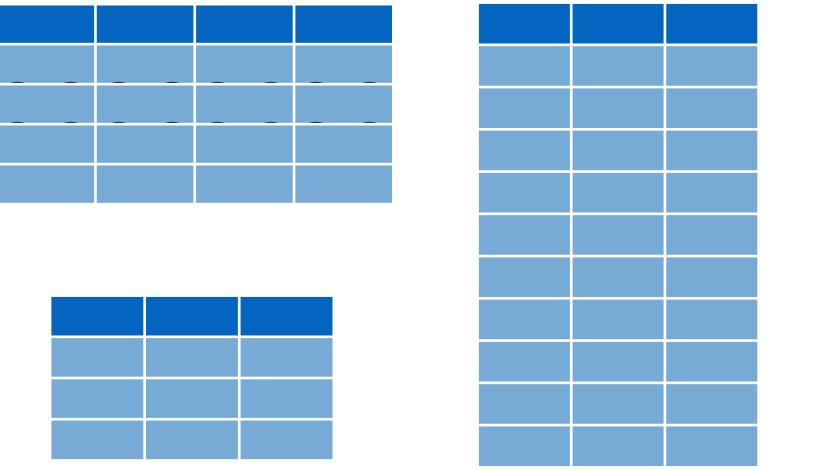
R Packages



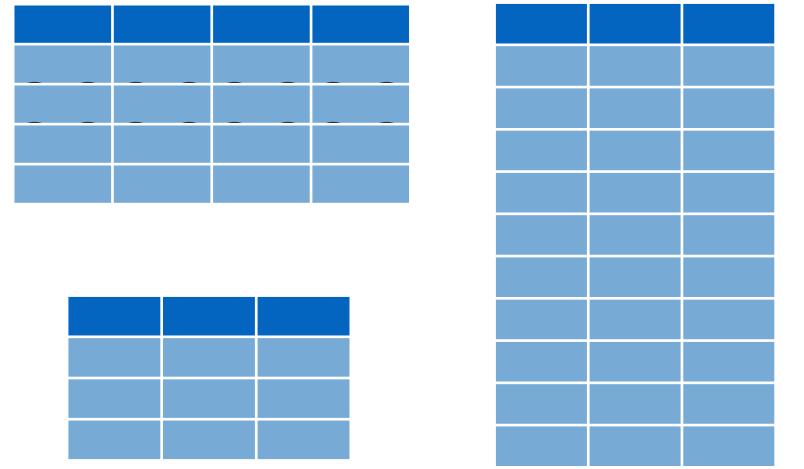
function1()
function2()
function3()
function4()



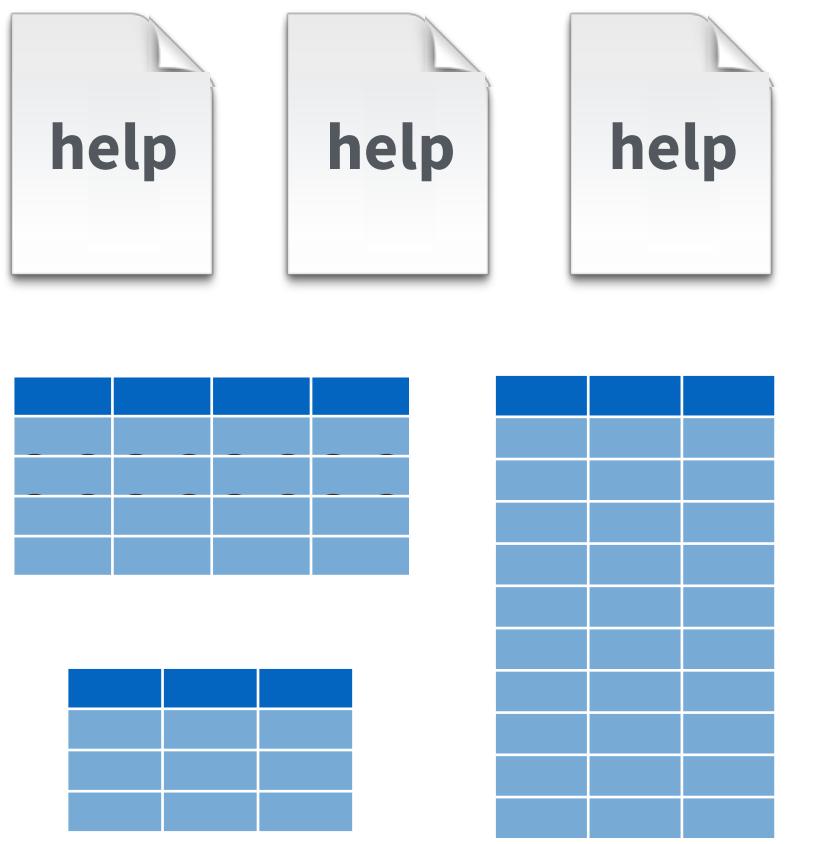
function5()
function6()
function7()
function8()



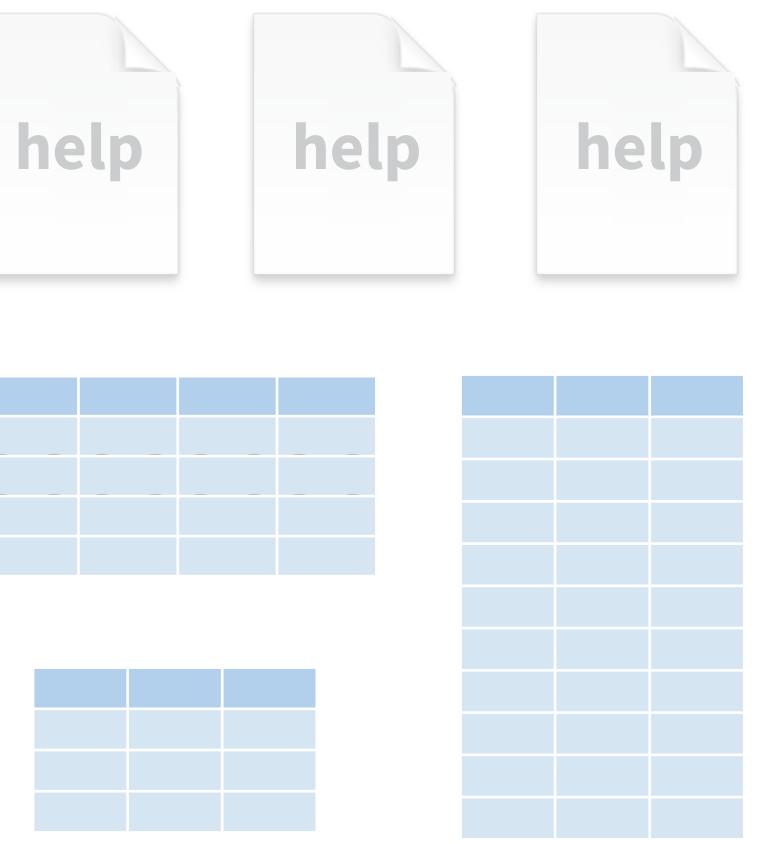
function9()
functionA()
functionB()
functionC()



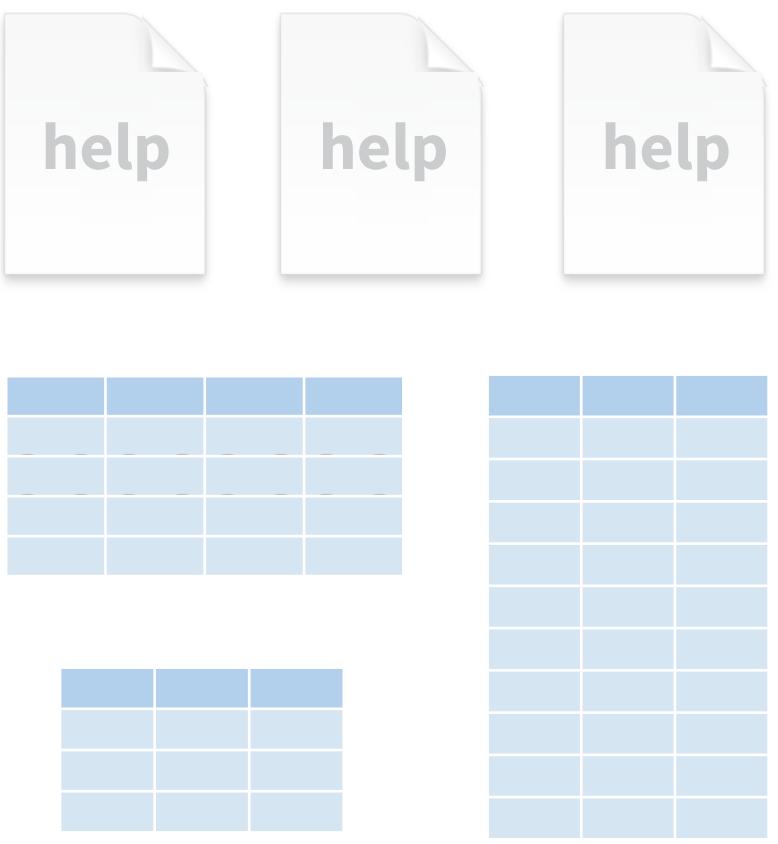
functionD()
functionE()
functionF()
functionG()



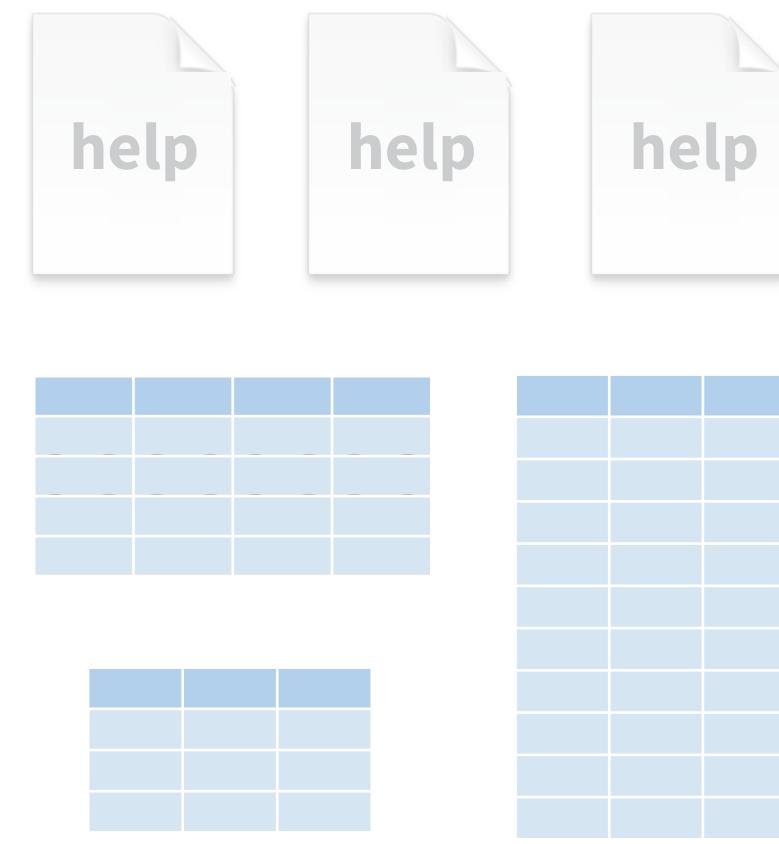
function1()
function2()
function3()
function4()



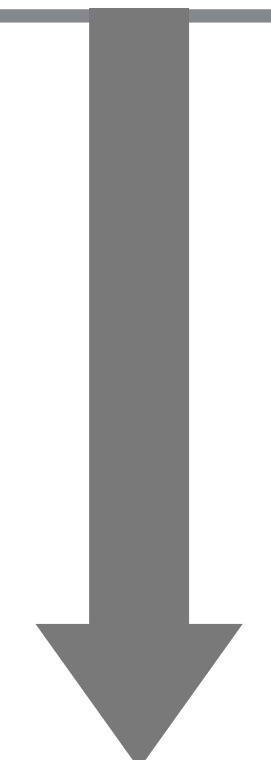
function5()
function6()
function7()
function8()



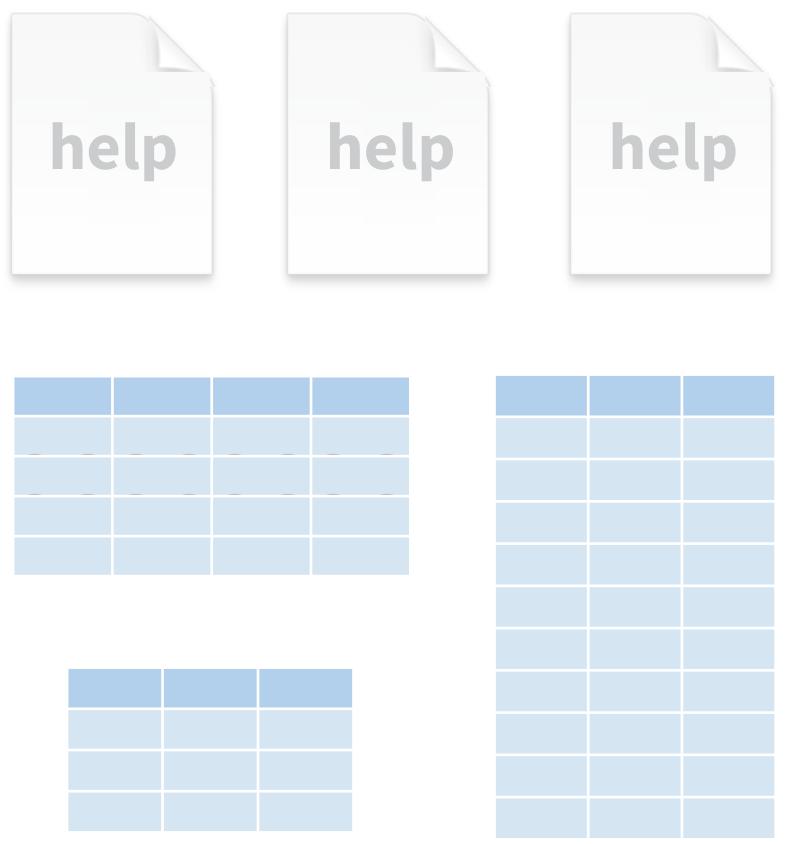
function9()
functionA()
functionB()
functionC()



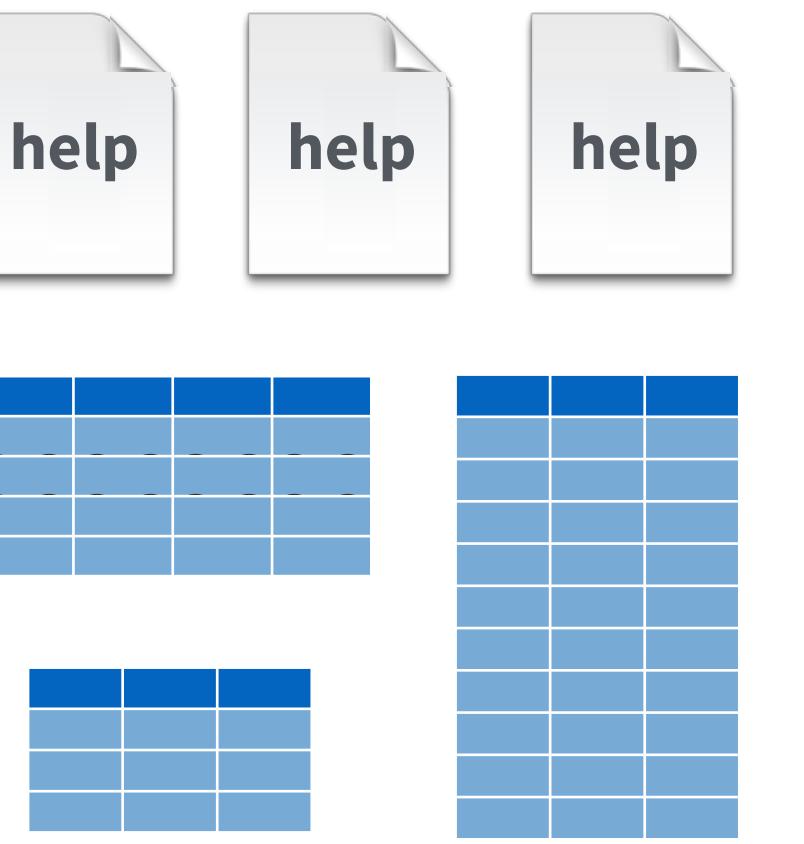
functionD()
functionE()
functionF()
functionG()



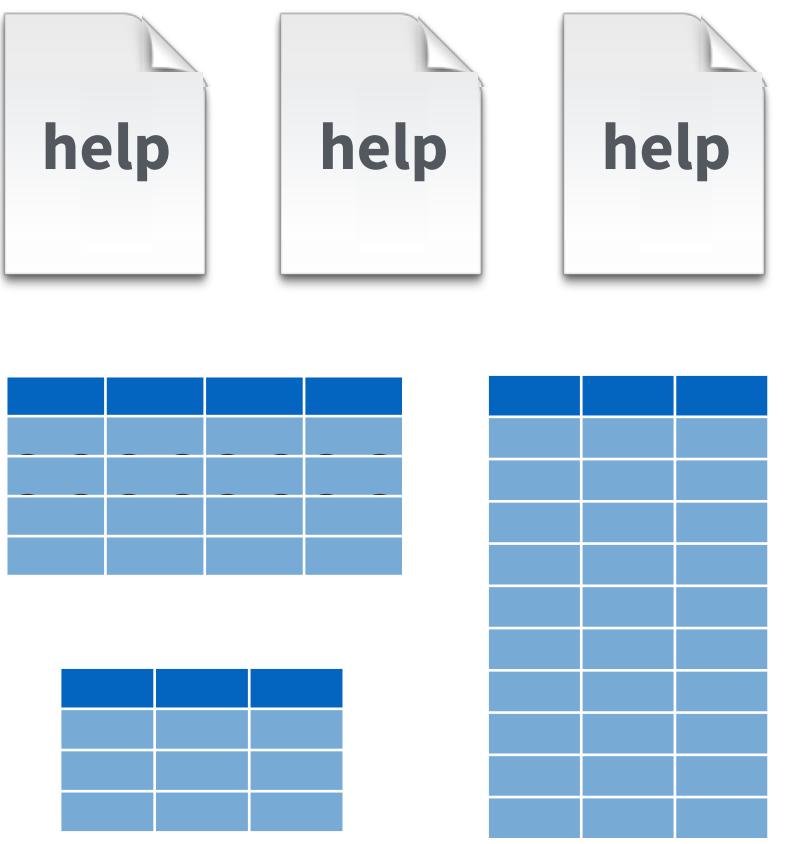
Base R



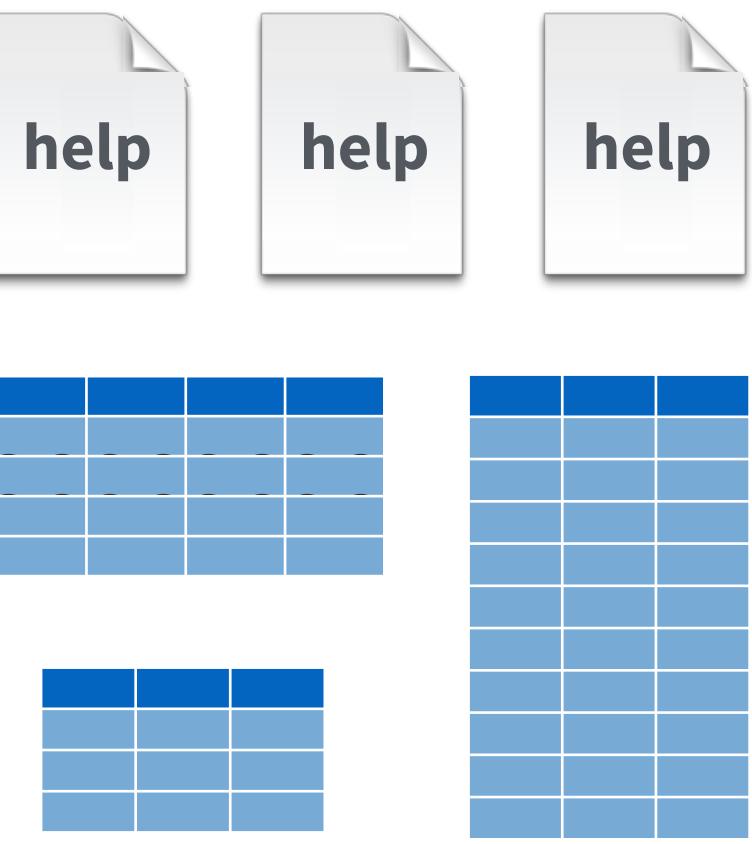
function1()
function2()
function3()
function4()



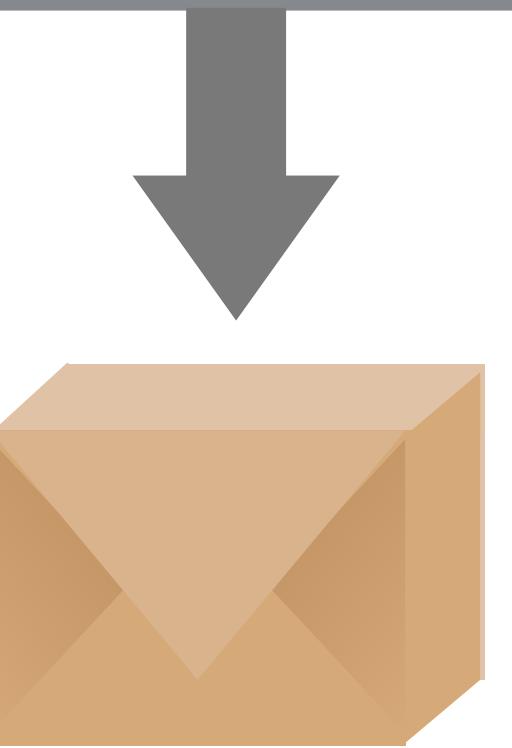
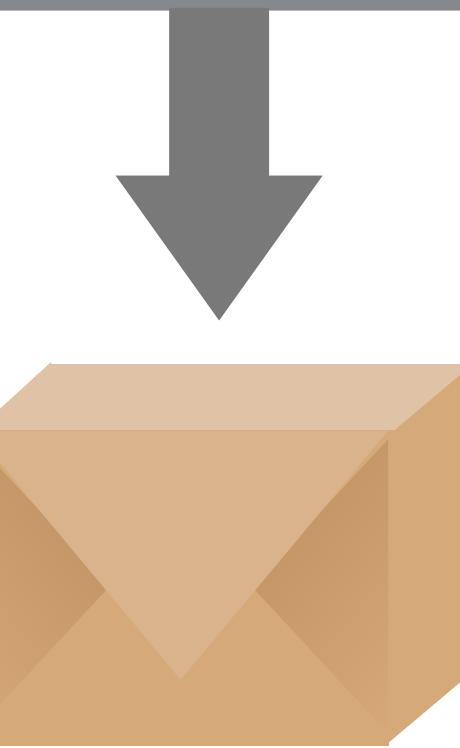
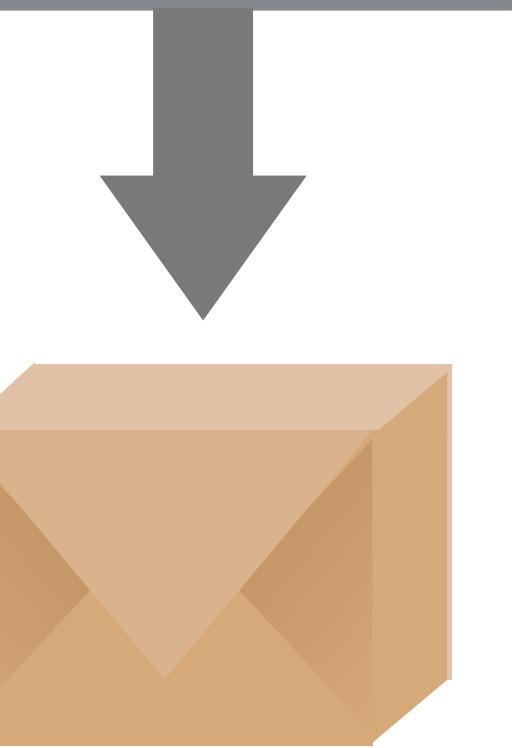
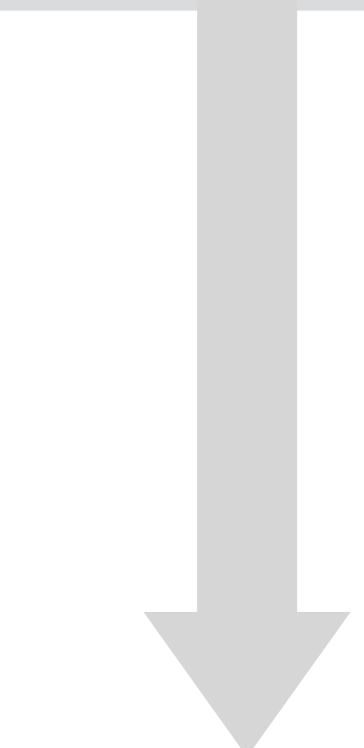
function5()
function6()
function7()
function8()



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



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



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](#)

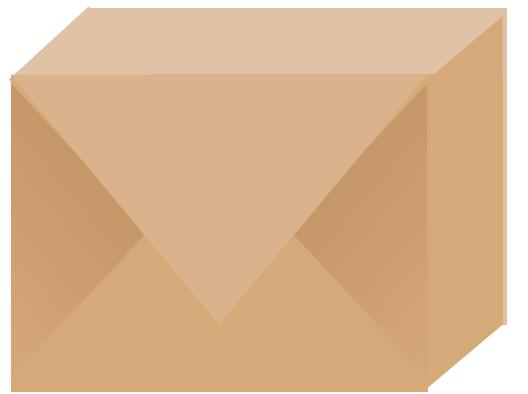
Your Turn 1

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.

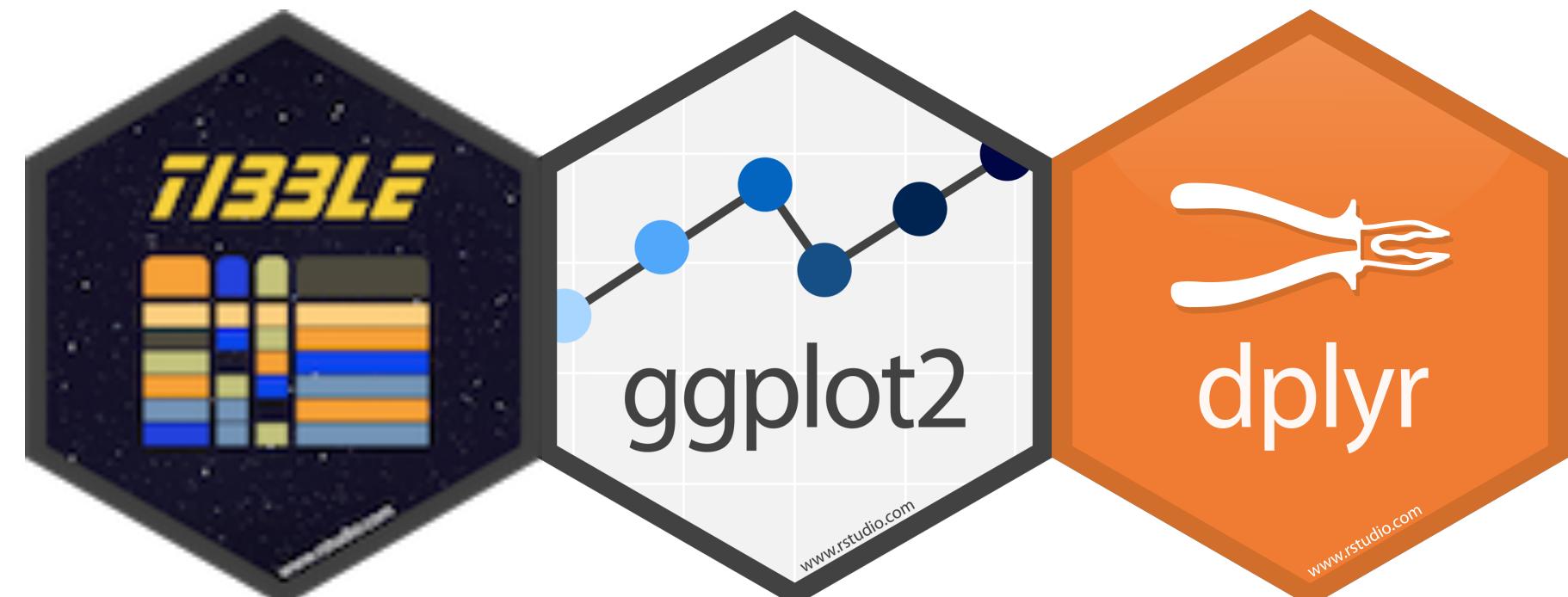
```
library("tidyverse")
```



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

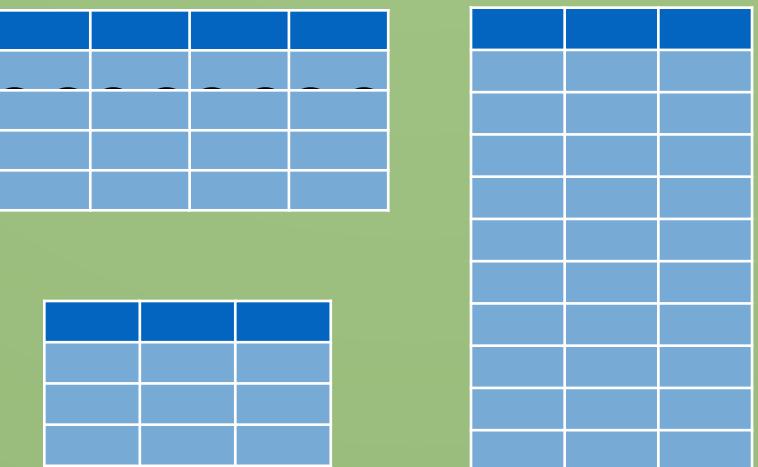
```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("stringr")
install.packages("forcats")
install.packages("lubridate")
install.packages("hms")
install.packages("DBI")
install.packages("haven")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

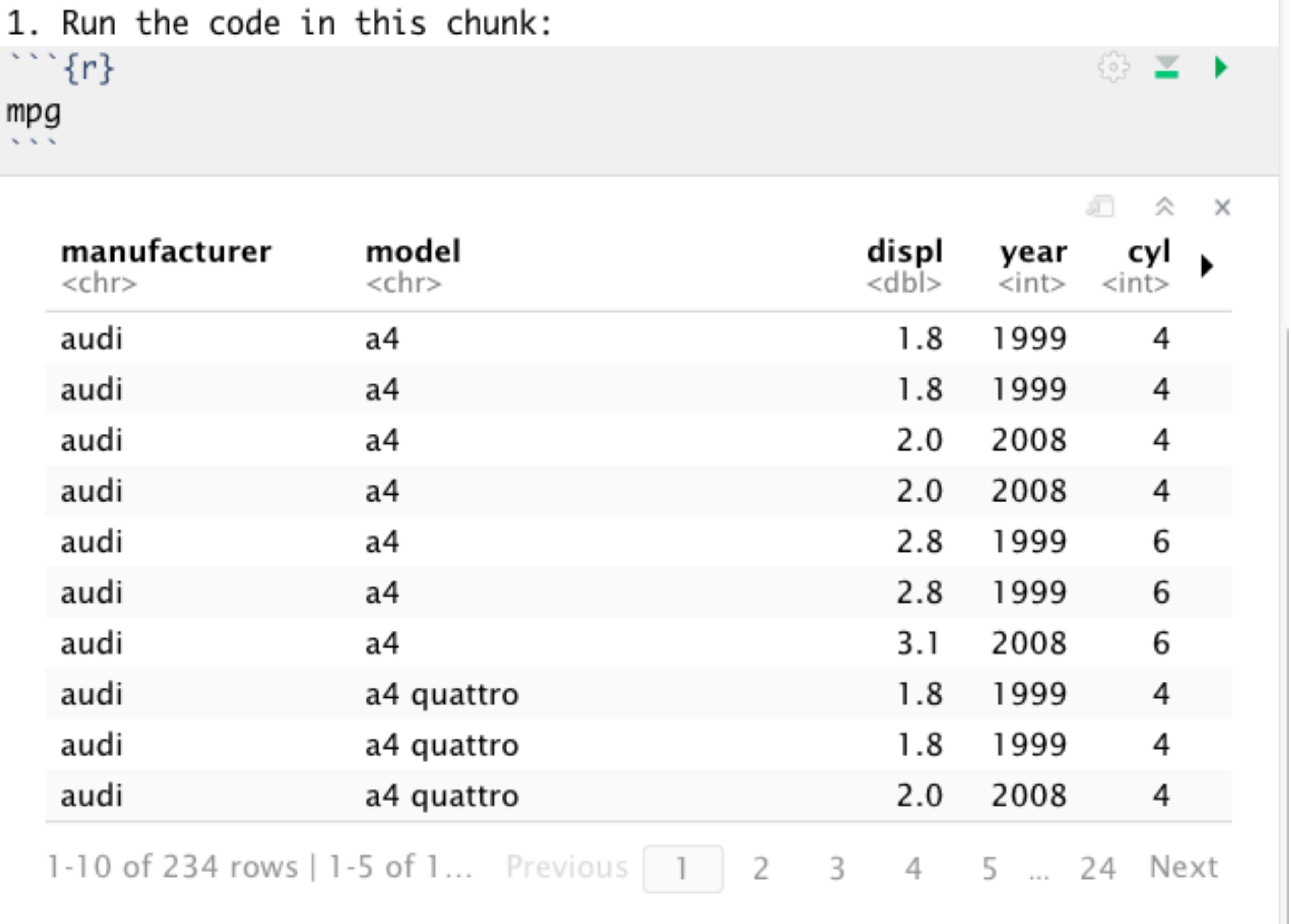
1. Run `mpg` in the code chunk
2. Type `mpg` on the Console and hit Enter

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 | model | displ | year | cyl |
|--------------|------------|-------|-------|-------|
| <chr> | <chr> | <dbl> | <int> | <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 | <code>c(1L, 2L, 3L, 4L)</code> |
| Double | Numbers | <code>c(1, 2, 3, 4)</code> |
| Character | Text | <code>c("1", "2", "3", "4")</code> |
| Logical | True or False | <code>c(TRUE, FALSE, FALSE, TRUE)</code> |

Your Turn 6

Take another look at mpg.

What type of data is in each column?

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

| manufacturer | model | displ | year | cyl | trans | drv | cty | hwy | fl |
|--------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| <chr> | <chr> | <dbl> | <int> | <int> | <chr> | <chr> | <int> | <int> | <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)
```

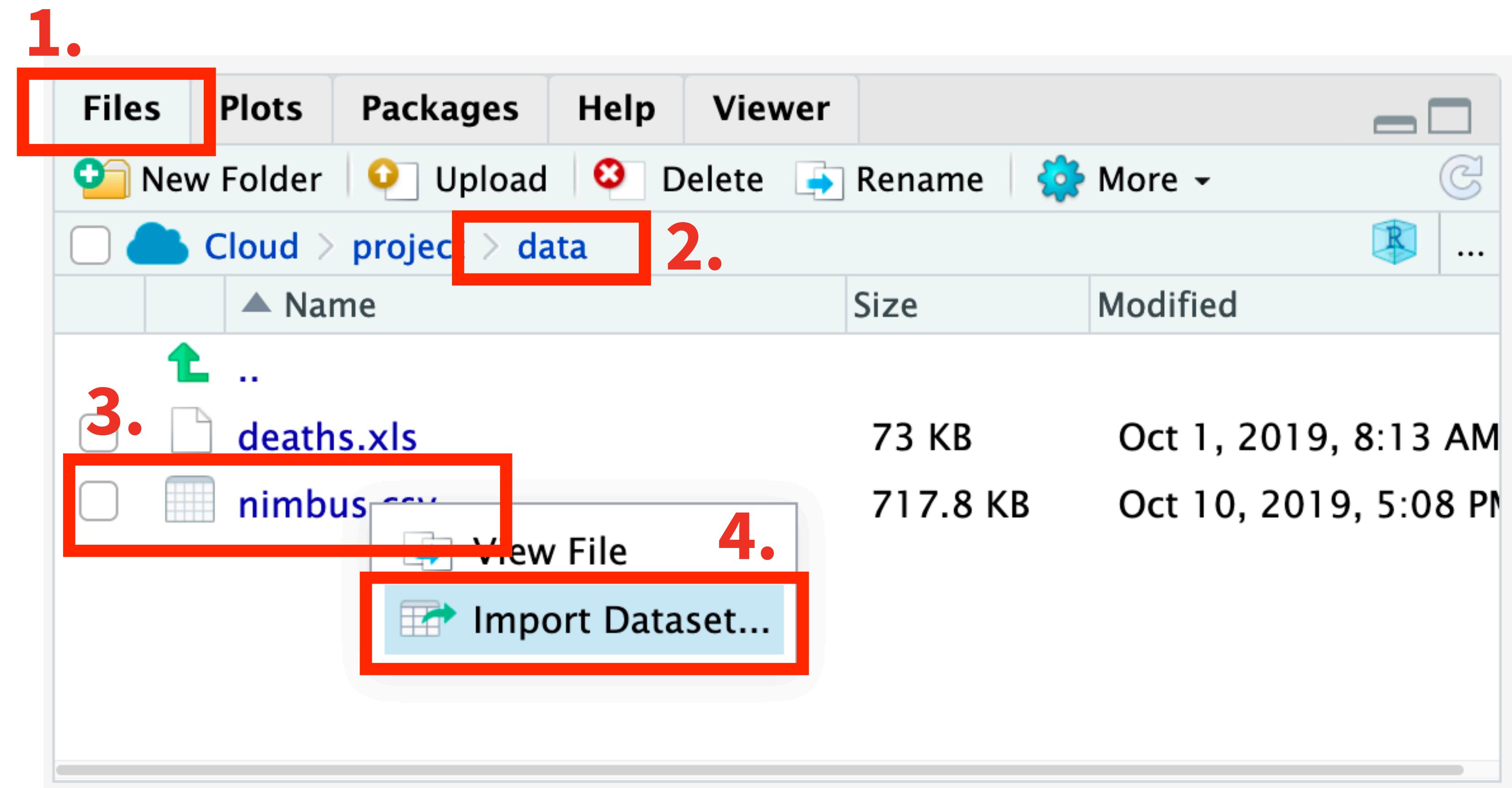
Import Dataset...

1. In the Files pane

2. Navigate to the data folder

3. Click on nimbus.csv

4. Import Dataset...



Import Dataset...

Import Text Data

File/URL:

/cloud/project/data/nimbus.csv Update

Data Preview:

| date
(double) | longitude
(double) | latitude
(double) | ozone
(double) | |
|------------------|-----------------------|----------------------|-------------------|--|
|------------------|-----------------------|----------------------|-------------------|--|

Import Options:

| | | | |
|---|--|---|---|
| Name: <input type="text" value="nimbus"/> | <input checked="" type="checkbox"/> First Row as Names | Delimiter: <input type="button" value="Comma"/> | Escape: <input type="button" value="None"/> |
| Skip: <input type="text" value="0"/> | <input checked="" type="checkbox"/> Trim Spaces | Quotes: <input type="button" value="Default"/> | Comment: <input type="button" value="Default"/> |
| | <input checked="" type="checkbox"/> Open Data Viewer | Locale: <input type="button" value="Configure..."/> | NA: <input type="button" value="Default"/> |

Code Preview:

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

Copy

[? Reading rectangular data using readr](#) Import Cancel

readr

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

object to save
output into

path to
the file

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", ...)
```

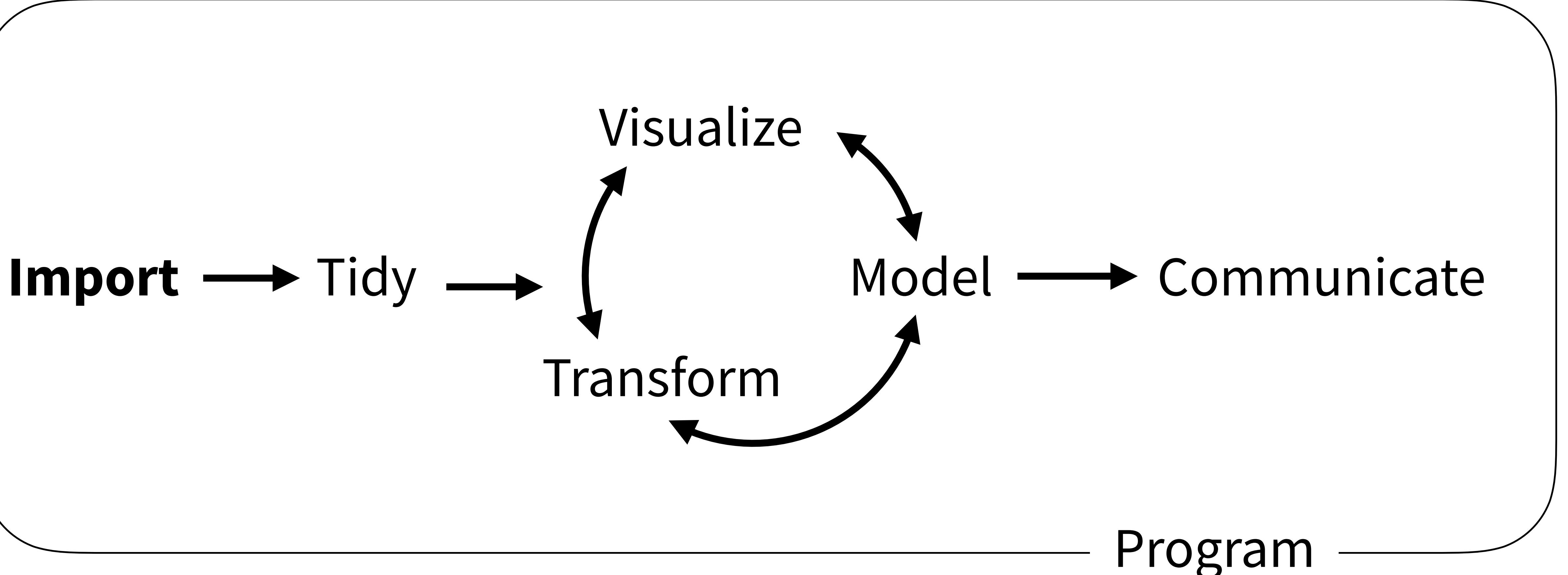
Import
functions in
the tidyverse
have
consistent
syntax

Other types of data

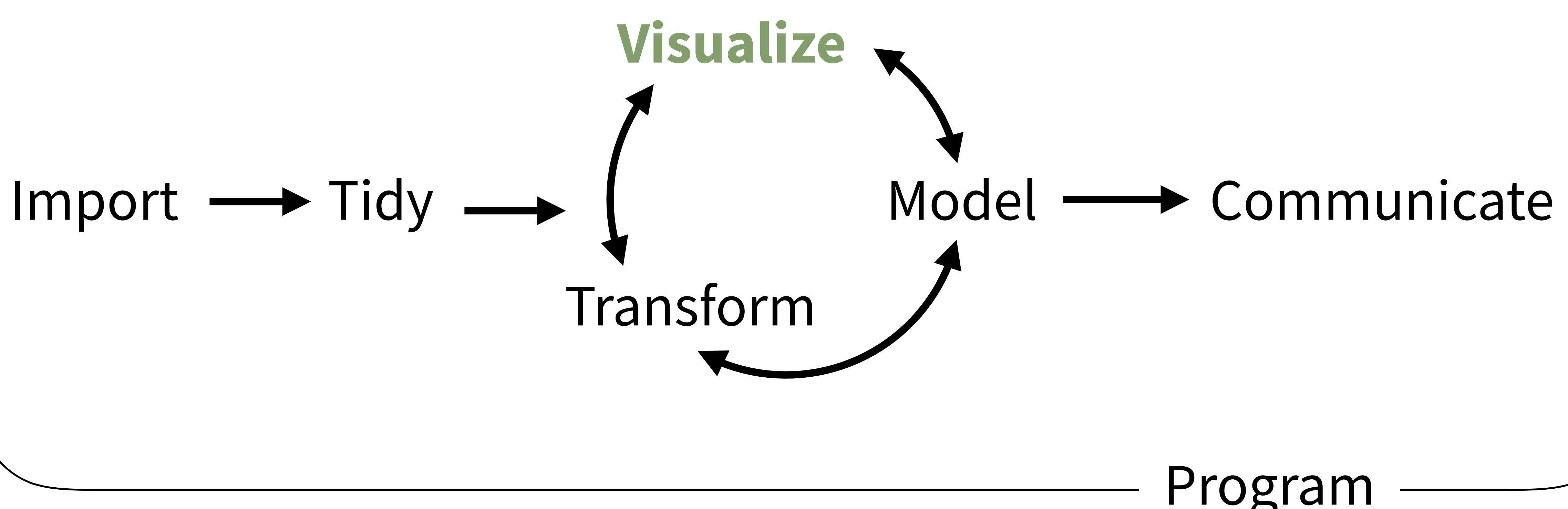
| package | accesses |
|----------|--------------------------|
| jsonlite | json |
| xml2 | xml |
| httr | web API's |
| rvest | web pages (web scraping) |
| DBI | databases |
| sparklyr | data loaded into spark |

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?