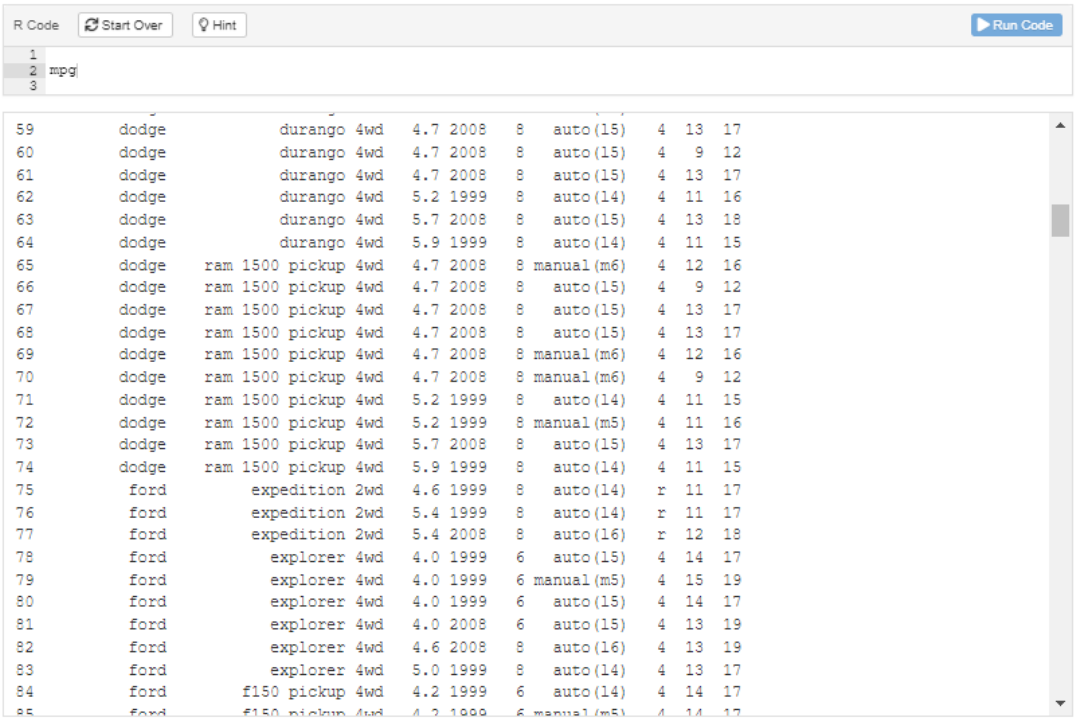
Eddie Aguilar

Data frames

✓ What is a data frame?

A **data frame** is a rectangular collection of values, usually organized so that variables appear in the columns and observations appear in rows.

Here is an example: the `mpg` data frame contains observations collected by the US Environmental Protection Agency on 38 models of cars. To see the `mpg` data frame, type `mpg` in the code chunk below and then click "Run Code".



59	dodge	durango 4wd	4.7	2008	8	auto(15)	4	13	17
60	dodge	durango 4wd	4.7	2008	8	auto(15)	4	9	12
61	dodge	durango 4wd	4.7	2008	8	auto(15)	4	13	17
62	dodge	durango 4wd	5.2	1999	8	auto(14)	4	11	16
63	dodge	durango 4wd	5.7	2008	8	auto(15)	4	13	18
64	dodge	durango 4wd	5.9	1999	8	auto(14)	4	11	15
65	dodge	ram 1500 pickup 4wd	4.7	2008	8	manual(m6)	4	12	16
66	dodge	ram 1500 pickup 4wd	4.7	2008	8	auto(15)	4	9	12
67	dodge	ram 1500 pickup 4wd	4.7	2008	8	auto(15)	4	13	17
68	dodge	ram 1500 pickup 4wd	4.7	2008	8	auto(15)	4	13	17
69	dodge	ram 1500 pickup 4wd	4.7	2008	8	manual(m6)	4	12	16
70	dodge	ram 1500 pickup 4wd	4.7	2008	8	manual(m6)	4	9	12
71	dodge	ram 1500 pickup 4wd	5.2	1999	8	auto(14)	4	11	15
72	dodge	ram 1500 pickup 4wd	5.2	1999	8	manual(m5)	4	11	16
73	dodge	ram 1500 pickup 4wd	5.7	2008	8	auto(15)	4	13	17
74	dodge	ram 1500 pickup 4wd	5.9	1999	8	auto(14)	4	11	15
75	ford	expedition 2wd	4.6	1999	8	auto(14)	r	11	17
76	ford	expedition 2wd	5.4	1999	8	auto(14)	r	11	17
77	ford	expedition 2wd	5.4	2008	8	auto(16)	r	12	18
78	ford	explorer 4wd	4.0	1999	6	auto(15)	4	14	17
79	ford	explorer 4wd	4.0	1999	6	manual(m5)	4	15	19
80	ford	explorer 4wd	4.0	1999	6	auto(15)	4	14	17
81	ford	explorer 4wd	4.0	2008	6	auto(15)	4	13	19
82	ford	explorer 4wd	4.6	2008	8	auto(16)	4	13	19
83	ford	explorer 4wd	5.0	1999	8	auto(14)	4	13	17
84	ford	f150 pickup 4wd	4.2	1999	6	auto(14)	4	14	17
85	ford	f150 pickup 4wd	4.2	1999	6	manual(m5)	4	14	17

A note about mpg

The code above worked because I've already loaded the ggplot2 package for you in this tutorial: `mpg` comes in the ggplot2 package. If you would like to look at `mpg` on your own computer, you will need to first load ggplot2. You can do that in two steps:

1. Run `install.packages('ggplot2')` to install ggplot2 if you do not yet have it.
2. Load ggplot2 with the `library(ggplot2)` command

After that, you will be able to access any object in ggplot2—including `mpg`—until you close R.

Did you notice how much information was inside `mpg`? Me too. Sometimes the contents of a data frame do not fit on a single screen, which makes them difficult to inspect. We'll look at an alternative to using and examining data frames soon. But first let's get some help…

Help pages

✓ How to open a help page

You can learn more about `mpg` by opening its help page. The help page will explain where the `mpg` dataset comes from and what each variable in `mpg` describes. To open the help page, type `?mpg` in the code chunk below and then click "Run Code".

R Code | Start Over | Hint | Run Code

```
1 ?mpg
2
3
```

✓ ? syntax

You can open a help page for any object that comes with R or with an R package. To open the help page, type a `?` before the object's name and then run the command, as you did with `?mpg`. This technique works for functions, packages, and more.

Notice that objects created by you or your colleagues will not have a help page (unless you make one).

✓ Exercises

Use the code chunk below to answer the following questions.

R Code | Start Over | Run Code

```
1 ?mpg
2
3 cars
4
5
```

```
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
7     10   18
8     10   26
9     10   34
10    11   17
11    11   28
12    12   14
13    12   20
14    12   24
15    12   28
16    13   26
17    13   34
18    13   34
19    13   46
20    14   26
21    14   36
22    14   60
23    14   80
```

Quiz

What does the `drv` variable of `mpg` describe? Read the help for `?mpg` to find out.

✗ Whether or not the vehicle has driver side airbags

✗ Whether a car is automatic or manual transmission

✗ The number of cylinders in the car's engine

✓ Something else

Correct!

`drv` describes the type of drivetrain in a car: front wheel drive, rear wheel drive, or four wheel drive.

How many rows are in the data frame named `cars`?

✗ 2

✗ 25

✓ 50

✗ 100

Correct!

How many columns are in the data frame named `cars`?

✗ 1

✓ 2

✗ 4

✗ more than four

Correct!

Tibbles

✓ What is a tibble?

The `flights` data frame in the nycflights13 package is an example of a *tibble*. Tibbles are a data frames with some extra properties.

To see what I mean, use the code chunk below to print the contents of `flights`.

R Code | Start Over | Hint | Run Code

```
1 flights
2
3
```

```
# A tibble: 336,776 × 19
    year month   day dep_time sched_de…¹ dep_d…² arr_t…³ sched…⁴ arr_d…⁵ carrier
   <int> <int> <int>    <int>      <int>   <dbl>   <int>   <int>   <dbl> <chr>
 1  2013     1     1      517        515       2     830     819      11 UA
 2  2013     1     1      533        529       4     850     830      20 UA
 3  2013     1     1      542        540       2     923     850      33 AA
 4  2013     1     1      544        545      -1    1004    1022     -18 B6
 5  2013     1     1      554        600      -6     812     837     -25 DL
 6  2013     1     1      554        558      -4     740     728      12 UA
 7  2013     1     1      555        600      -5     913     854      19 B6
 8  2013     1     1      557        600      -3     709     723     -14 EV
 9  2013     1     1      557        600      -3     838     846      -8 B6
10  2013     1     1      558        600      -2     753     745       8 AA
# … with 336,766 more rows, 9 more variables: flight <int>, tailnum <chr>,
#   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
#   minute <dbl>, time_hour <dttm>, and abbreviated variable names
#   ¹sched_dep_time, ²dep_delay, ³arr_time, ⁴sched_arr_time, ⁵arr_delay
```

Good Job. `flights` describes every flight that departed from New York City in 2013. The data comes from the US Bureau of Transportation Statistics, and is documented in `?flights`.

The tibble display

You might notice that `flights` looks a little differently than `mpg`. `flights` shows only the first few rows of the data frame and only the columns that fit on one screen.

`flights` prints differently because it's a **tibble**. Tibbles are data frames that are slightly tweaked to be more user-friendly. For example, R doesn't try to show you all of a tibble at once (but it will try to show you all of a data frame that is not a tibble).

You can use `as_tibble()` to return a tibble version of any data frame. For example, this would return a tibble version of `mpg`: `as_tibble(mpg)`.

Previous Topic | Next Topic

```
##  2  2013     1     1      533        529       4     850     830      20 UA
##  3  2013     1     1      542        540       2     923     850      33 AA
##  4  2013     1     1      544        545      -1    1004    1022     -18 B6
##  5  2013     1     1      554        600      -6     812     837     -25 DL
##  6  2013     1     1      554        558      -4     740     728      12 UA
##  7  2013     1     1      555        600      -5     913     854      19 B6
##  8  2013     1     1      557        600      -3     709     723     -14 EV
##  9  2013     1     1      557        600      -3     838     846      -8 B6
## 10  2013     1     1      558        600      -2     753     745       8 AA
## # … with 336,766 more rows, 9 more variables: flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>, and abbreviated variable names
## #   ¹sched_dep_time, ²dep_delay, ³arr_time, ⁴sched_arr_time, ⁵arr_delay
```

Did you notice that a row of three (or four) letter abbreviations appears under the column names of `flights`? These abbreviations describe the *type* of data that is stored in each column of `flights`:

- `int` stands for integers.
- `dbl` stands for doubles, or real numbers.
- `chr` stands for character vectors, or strings.
- `dttm` stands for date-times (a date + a time).

There are three other common types of variables that aren't used in this dataset but are used in other datasets:

- `lgl` stands for logical, vectors that contain only `TRUE` or `FALSE`.
- `fctr` stands for factors, which R uses to represent categorical variables with fixed possible values.
- `date` stands for dates.

This row of data types is unique to tibbles and is one of the ways that tibbles try to be more user-friendly than data frames.

✓ Test your knowledge

Which types of variables does `flights` contain? Check all that apply.

✓ integers

✓ doubles

✗ factors

✓ characters

Great Job!

Congratulations

You've met R's basic table structures—data frames and tibbles; and you have learned how to inspect their contents. When you are ready, go