Data Import

David Gerard

2019-02-08

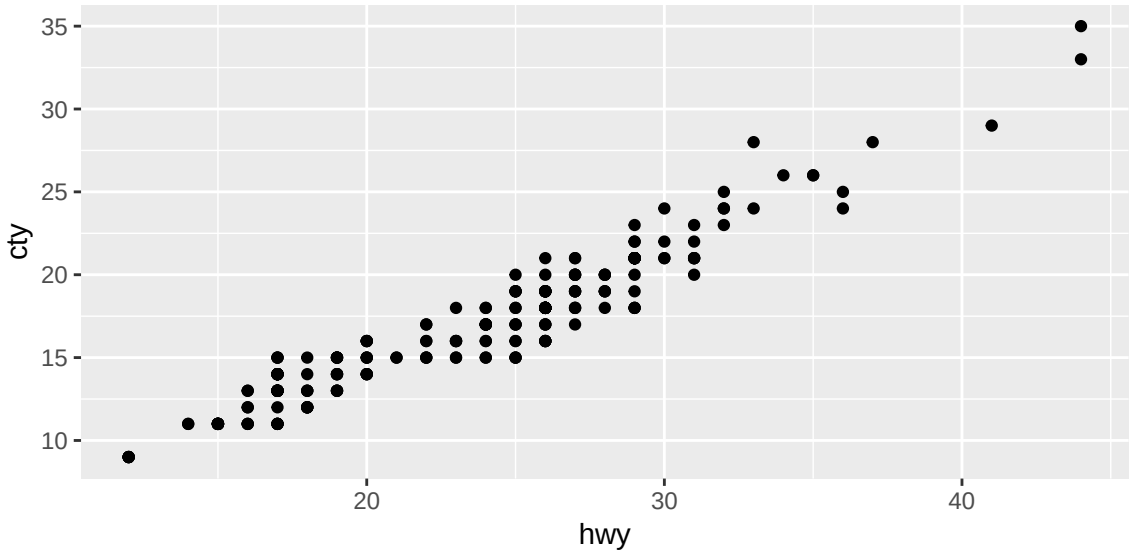
Learning Objectives

- Import data from CSV's,
- Working Directories
- Chapter 11 of RDS

Working Directories

- The working directory is where R will look for and save things by default.
- When you specify to save a figure, save a file, or load some data, it will be with respect to the working directory.
- You can see where the current working directory is by `getwd()`, or by looking at the top of the console in RStudio.
- You can change the working directory by Session > Set Working Directory > Choose Directory. Or by CONTROL + SHIFT + HA. Or you can use the `setwd()` command.
- A shortcut is to set the working directory to your sourcefile location with Session > Set Working Directory > To Source File Location.
- When you read and write files/figures, you can then specify the path from the position of the working directory.
- Suppose we want to save the following figure:

```
suppressPackageStartupMessages(library(tidyverse))
data("mpg")
pl <- ggplot(mpg, aes(x = hwy, y = cty)) +
  geom_point()
pl
```



- To save `pl` in the current folder, we would use:

```
ggsave(filename = "./my_saved_plot.pdf", plot = pl)
```

- The "`.`" means "the current folder".
- To save `pl` in the folder one level up we would use:

```
ggsave(filename = "../my_saved_plot.pdf", plot = pl)
```

- The "`..`" means "go one level up".
- If we are in the analysis folder, and we want to save `pl` in the output folder, we would use:

```
  ggsave(filename = "../output/my_saved_plot.pdf", plot = pl)
```

- If we have a subfolder called "fig" within out current folder. We could save `pl` in "fig" with

```
  ggsave(filename = "./fig/my_saved_plot.pdf", plot = pl)
```

- **NEVER USE ABSOLUTE PATHS**. For example, you should never start the path from "C" if you use Windows. This makes your code non-transferable to other users.

readr

- A lot of datasets come in comma-separated or tab-separated formats. For example, These are the first few rows of hate_crimes2.csv:

```
state,median_house_inc,share_unemp_seas,share_pop_metro,share_pop_hs,share_non_citizen,share_white_
Alabama,42278,0.06,0.64,0.821,0.02,0.12,0.472,0.35,0.63,0.125838926,1.806410489
Alaska,67629,0.064,0.63,0.914,0.04,0.06,0.422,0.42,0.53,0.143740118,1.656700109
Arizona,49254,0.063,0.9,0.842,0.1,0.09,0.455,0.49,0.5,0.225319954,3.413927994
```

```
Arkansas,44922,0.052,0.69,0.824,0.04,0.12,0.458,0.26,0.6,0.069060773,0.869208872
California,60487,0.059,0.97,0.806,0.13,0.09,0.471,0.61,0.33,0.255805361,2.397985899
Colorado,60940,0.04,0.8,0.893,0.06,0.07,0.457,0.31,0.44,0.390523301,2.804688765
Connecticut,70161,0.052,0.94,0.886,0.06,0.06,0.486,0.3,0.41,0.335392269,3.772701469
Delaware,57522,0.049,0.9,0.874,0.05,0.08,0.44,0.37,0.42,0.322754169,1.469979563
District of Columbia,68277,0.067,1,0.871,0.11,0.04,0.532,0.63,0.04,1.52230172,10.95347971
```

- In the file, each column is separated by a comma. Each row is separated by a new line.
- To read a CSV (comma-separated values) file into R, use the `read_csv()` function from the readr package.
- The readr package is apart of the tidyverse, and so it is automatically loaded when you load the tidyverse.

```
library(tidyverse)
hate_crimes <- read_csv(file = "../../data/hate_crimes1.csv")
```

```
## Parsed with column specification:
## cols(
##   `state median_house_inc    share_unemp_seas    share_pop_metro share_pop_hs    share_non_citiz
## )
```

- Use `read_tsv()` if columns are separted by tabs.
- Use `read_csv2()` if columns are separated by semi-colans.
- Other file formats are listed in RDS
- You want to import data directly from Excel? Don't.
  - First export the Excel spreadsheet as a CSV. Then read the CSV file into R.
- You are using colors to represent meaningful information in Excel? Don't.
  - Edit the data so that the information is encoded by a new variable.

Special Considerations

- **Always check your data immediately after importing it**.
  - Check that the types are correct for each of the variables.
  - Check that the missing data were coded correctly.
- Sometimes the files code missing data other than `NA`. For example, it's common to use periods `.`, or in some genomic settings they use `-9` as missing.
- R won't know how to handle this without you telling it, so you'll have to know what the missing data encoding is and specify it with the `na` argument in `read_csv()`.
- readr will try to guess the type for each column (double, integer, character, logic, etc). Sometimes it guesses wrong. If it seems to be guessing wong, use the `col_types` to explicitly specify the column types.
- Sometimes there are comments at the start of a data file. You can skip the first few lines before starting to read data with the `skip` argument.
- If the comments begin with a special character, you can use the `comment` argument.