Data basics

Richard Layton

2017-09-05

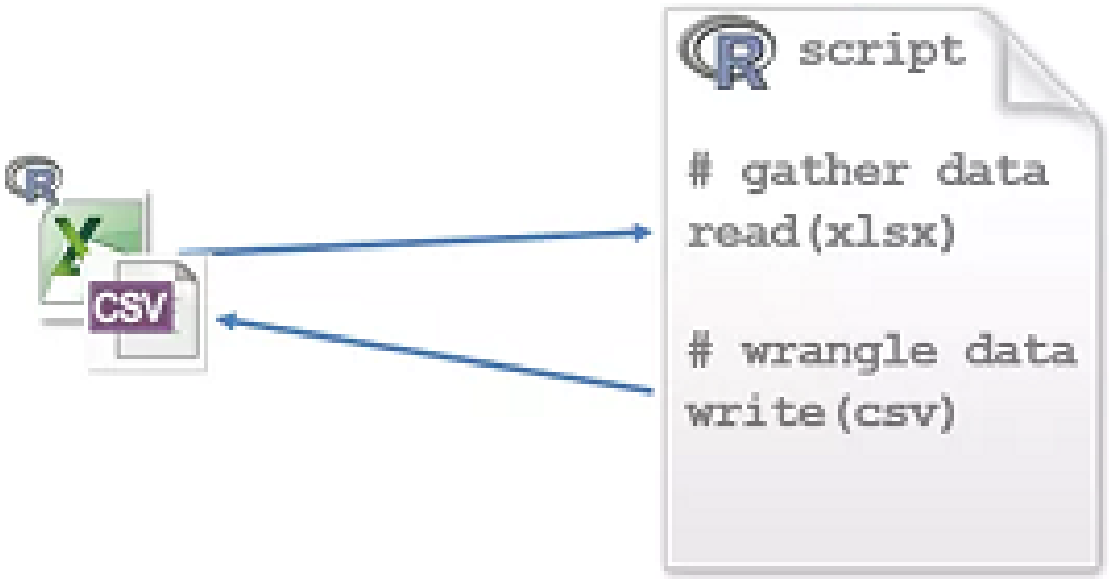
Today's R packages

Package	For
tidyverse	ggplot2, dplyr, readr, tidyr
ggplot2	Graphing data
dplyr	Data manipulation
readr	Read rectangular data files (like csv or tsv)
tidyr	set of functions that help you get to tidy data
readxl	Read .xls and .xslx files
VIM	Examine missing values in a data frame

Learning objectives

After working through the data basics tutorial, you should be able to

- Read data from an Excel file
- Read and write CSV data files
- Obtain and manipulate data sets included with R and R packages



The readxl package

```
library(readxl)
my_data <- read_excel(
        path = "data/my_file_name.xlsx"
     , sheet = "sheet_tab_name"
     )
```

- To read .xls or .xlsx files
- Assumes data is tidy or nearly tidy

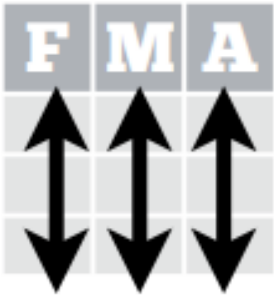
Row 1 has the names of the variables

Row 2 starts the data

	A	B	C
1	year	country	production
2	1977	Japan	17.7
3	1978	Japan	19
4	1979	Japan	19.9
5	1980	Japan	24.3
6	1977	USA	30
7	1978	USA	29.1
8	1979	USA	27.2

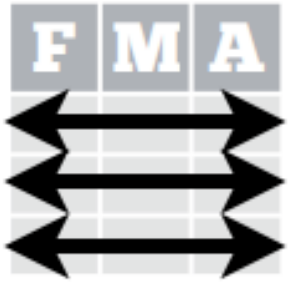
What is tidy data?

In a tidy data set:



Each **variable** is saved in its own **column**

&



Each **observation** is saved in its own **row**

Source: data-wrangling-cheatsheet, https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf

When your data is not tidy

Row 1 has merged cells with data

Row 2 has more variable name information

Row 3 starts the data

	A	B	C	D	E
1		Rural		Urban	
2	Group	Men	Women	Men	Women
3	50-54	11.7	8.7	15.4	8.4
4	55-59	18.1	11.7	24.3	13.6
5	60-64	26.9	20.3	37	19.3
6	65-69	41	30.9	54.6	35.1
7	70-74	66	54.3	71.1	50

When your data is not tidy

```
wide_data <- read_excel("data/VADeaths.xlsx")
```

X__1	Rural	X__2	Urban
Group	Men	Women	Men
50-54	11.7	8.6999999999999993	15.4
55-59	18.100000000000001	11.7	24.3
60-64	26.9	20.3	37
65-69	41	30.9	54.6
70-74	66	54.3	71.099999999999

We can skip lines, but might lose information

```
wide_data <- read_excel("data/VADeaths.xlsx", skip = 1)
```

Group	Men	Women	Men___1	Women___1
50-54	11.7	8.7	15.4	8.4
55-59	18.1	11.7	24.3	13.6
60-64	26.9	20.3	37.0	19.3
65-69	41.0	30.9	54.6	35.1
70-74	66.0	54.3	71.1	50.0

Reading and reshaping data like this is a topic for another day