Chapter8

Ayush Kumar Shah

9/23/2020

Chapter 8 Data import with readr

- readr is imported with tidyverse
- functions in readr
  - read_csv()
  - read_tsv()
  - read_delim() (any delimiter)
  - read_fwf() (fixed width file)

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ ti
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Rad_csv

read_csv different from read.csv of R.

```
# heights <- read_csv("data/heights.csv")
#> Parsed with column specification:
#> cols(
#> earn = col_double(),
#> height = col_double(),
#> sex = col_character(),
#> ed = col_integer(),
#> age = col_integer(),
#> race = col_character()
#> )
```

Inline tables

```
read_csv("a,b,c
1,2,3
4,5,6")
```

```
## # A tibble: 2 x 3
##       a     b     c
##   <dbl> <dbl> <dbl>
## 1     1     2     3
## 2     4     5     6
```

Documenting the data

```
read_csv("The first line of metadata
The second line of metadata
x,y,z
1,2,3", skip = 2)
```

```
## # A tibble: 1 x 3
##       x     y     z
##   <dbl> <dbl> <dbl>
## 1     1     2     3
```

```
read_csv("# A comment I want to skip
x,y,z
1,2,3", comment = "#")
```

```
## # A tibble: 1 x 3
##       x     y     z
##   <dbl> <dbl> <dbl>
## 1     1     2     3
```

Give default names to columns

```
read_csv("1,2,3\n4,5,6", col_names = FALSE)
```

```
## # A tibble: 2 x 3
##      X1    X2    X3
##   <dbl> <dbl> <dbl>
## 1     1     2     3
## 2     4     5     6
```

Pass col names

```
read_csv("1,2,3\n4,5,6", col_names = c("x", "y", "z"))
```

```
## # A tibble: 2 x 3
##       x     y     z
##   <dbl> <dbl> <dbl>
## 1     1     2     3
## 2     4     5     6
```

Missing data

```
read_csv("a,b,c\n1,2,.", na = ".")
```

```
## # A tibble: 1 x 3
##       a     b c
##   <dbl> <dbl> <lgl>
## 1     1     2 NA
```

Advantages compared to R's read.csv

- Faster (10x), progress bar.
- produce tibbles.
- more reproducible across all type of systems.