Importing and Transforming Data Exercises

Ai Yukino

Import packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

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

```
library(readxl)
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

Exercise 1

List all example files available with the readr library.

```
readr_example()
```

```
##  [1] "challenge.csv"         "chickens.csv"          "epa78.txt"
##  [4] "example.log"           "fwf-sample.txt"        "massey-rating.txt"
##  [7] "mtcars.csv"            "mtcars.csv.bz2"        "mtcars.csv.zip"
## [10] "whitespace-sample.txt"
```

Exercise 2

Read the `mtcars.csv` file.

```r
file_path <- readr_example("mtcars.csv")
read_csv(file = file_path)
```

```
## # A tibble: 32 x 11
##      mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  21       6  160    110  3.9   2.62  16.5     0     1     4     4
##  2  21       6  160    110  3.9   2.88  17.0     0     1     4     4
##  3  22.8     4  108     93  3.85  2.32  18.6     1     1     4     1
##  4  21.4     6  258    110  3.08  3.22  19.4     1     0     3     1
##  5  18.7     8  360    175  3.15  3.44  17.0     0     0     3     2
##  6  18.1     6  225    105  2.76  3.46  20.2     1     0     3     1
##  7  14.3     8  360    245  3.21  3.57  15.8     0     0     3     4
##  8  24.4     4  147.    62  3.69  3.19  20       1     0     4     2
##  9  22.8     4  141.    95  3.92  3.15  22.9     1     0     4     2
## 10  19.2     6  168.   123  3.92  3.44  18.3     1     0     4     4
## # ... with 22 more rows
```

Exercise 3

Read the first 10 lines from the `mtcars.csv` file.

```
file_path <- readr_example("mtcars.csv")
read_csv(file_path, n_max = 10)
```

```
## # A tibble: 10 x 11
##      mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  21       6  160    110  3.9   2.62  16.5     0     1     4     4
##  2  21       6  160    110  3.9   2.88  17.0     0     1     4     4
##  3  22.8     4  108     93  3.85  2.32  18.6     1     1     4     1
##  4  21.4     6  258    110  3.08  3.22  19.4     1     0     3     1
##  5  18.7     8  360    175  3.15  3.44  17.0     0     0     3     2
##  6  18.1     6  225    105  2.76  3.46  20.2     1     0     3     1
##  7  14.3     8  360    245  3.21  3.57  15.8     0     0     3     4
##  8  24.4     4  147.    62  3.69  3.19  20       1     0     4     2
##  9  22.8     4  141.    95  3.92  3.15  22.9     1     0     4     2
## 10  19.2     6  168.   123  3.92  3.44  18.3     1     0     4     4
```

Exercise 4

Read the `example.log` file.

```
file_path <- readr_example("example.log")
read_csv(file_path)
```

```
## # A tibble: 1 x 1
##   `172.21.13.45 - Microsoft\\JohnDoe [08/Apr/2001:17:39:04 -0800] "GET /script~`
##   <chr>
## 1 "127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] \"GET /apache_pb.gif HTTP/1.0~
```

Exercise 5

List all sheets in `readxl_example("datasets.xlsx")`.

```
file_path <- readxl_example("datasets.xlsx")
excel_sheets(file_path)
```

```
## [1] "iris"     "mtcars"   "chickwts" "quakes"
```

Exercise 6

Read data from the last sheet.

```
file_path <- readxl_example("datasets.xlsx")
read_xlsx(file_path, sheet = "quakes")
```

```
## # A tibble: 1,000 x 5
##      lat  long depth   mag stations
##    <dbl> <dbl> <dbl> <dbl>    <dbl>
##  1 -20.4  182.   562   4.8       41
##  2 -20.6  181.   650   4.2       15
##  3 -26    184.    42   5.4       43
##  4 -18.0  182.   626   4.1       19
##  5 -20.4  182.   649   4         11
##  6 -19.7  184.   195   4         12
##  7 -11.7  166.    82   4.8       43
##  8 -28.1  182.   194   4.4       15
##  9 -28.7  182.   211   4.7       35
## 10 -17.5  180.   622   4.3       19
## # ... with 990 more rows
```

Exercise 7

Load the `dplyr` package. Install and load the `AER` package and run the command `data("Fertility")` which loads the dataset Fertility to your workspace. Take a `glimpse()` at the data.

```
data("Fertility")
glimpse(Fertility)
```

```
## Rows: 254,654
## Columns: 8
## $ morekids <fct> no, no, no, no, no, no, no, no, no, no, yes, no, no, no, no, ~
## $ gender1  <fct> male, female, male, male, female, male, female, male, female,~
## $ gender2  <fct> female, male, female, female, female, female, male, male, mal~
## $ age      <int> 27, 30, 27, 35, 30, 26, 29, 33, 29, 27, 28, 28, 35, 34, 32, 2~
## $ afam     <fct> no, no, no, yes, no, no, no, no, no, no, no, no, no, no, no, ~
## $ hispanic <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, no, no, n~
## $ other    <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, no, no, n~
## $ work     <int> 0, 30, 0, 0, 22, 40, 0, 52, 0, 0, 0, 52, 52, 52, 8, 7, 0, 40,~
```

Exercise 8

Select rows 35 to 50 and print to console its age and work entry.

```
Fertility %>%
  {.[c(35, 50),]} %>%
  select(age, work)
```

```
##    age work
## 35  28   20
## 50  29    0
```

Exercise 9

Select the last row in the dataset and print to console.

```
tail(Fertility, 1)
```

```
##        morekids gender1 gender2 age afam hispanic other work
## 254654      yes  female  female  35   no       no    no    0
```

Exercise 10

Count how many women proceeded to have a third child.

```
Fertility %>%
  filter(morekids == "yes") %>%
  nrow()
```

```
## [1] 96912
```

Exercise 11

There are four possible gender combinations for the first two children. Which is the most common?

```
ff <- Fertility %>%
  filter(gender1 == "female", gender2 == "female") %>%
  nrow()
fm <- Fertility %>%
  filter(gender1 == "female", gender2 == "male") %>%
  nrow()
mm <- Fertility %>%
  filter(gender1 == "male", gender2 == "male") %>%
  nrow()
mf <- Fertility %>%
  filter(gender1 == "male", gender2 == "female") %>%
  nrow()

tb <- tibble(gender_pair = c("female-female", "female-male",
                             "male-male", "male-female"),
       count = c(ff, fm, mm, mf))
tb
```

```
## # A tibble: 4 x 2
##   gender_pair   count
##   <chr>         <int>
## 1 female-female 60946
## 2 female-male   62724
## 3 male-male     67799
## 4 male-female   63185
```

So the "male-male" pair is most common.