Tutorial and Laboratories

Week 5

The purpose of this week's tutorial and lab exercises is to introduce you to a special kind of data frames called `tibbles`. Also, to get familiar with different ways to import and to handle text files using the `readr` package.

The objectives of this tutorial and lab are to:

1- Get you familiar with creating, accessing and operating on 'tibbles'.
2- Get yourself familiar with reading and writing data with functions from the `readr` package

Tibbles

Tibbles are a modern way to create and to handle data frames. Tibbles incorporate features of the old data frames that resist the test of time with new added features to make reading, accessing and handling the big datasets are much easier and effective.

Note: To work with creating, reading or writing tibbles, you need to install the package called *"tidyverse"*. This can be done by installing and calling this package as following:

```
install.packages("tidyverse")
library(tidyverse)
```

You need to install the packages only once. Next time you just need to call it using the library(package name) command.

Example

To read a data frame as a tibble, we use `as_tibble(the name of the data frame)` function, for example:

```
books_df <- data.frame(author=c("Reda", "John", "Edward", "Ben"),
+ year=c(2104, 2016, 2005, 2019),
+ publisher=c("Wiley", "Springer", "Sage", "International Books"))

# convert the data frame as a tibble
books_tbl <- as_tibble(book_df)
book_tbl
```

Exercise 1 (Tibbles)

```
Anything_df <- data.frame(
  A = 1:1000,
  B = A * 2 + rnorm(length(A))
)

Anything_tbl <- tibble(
  A = 1:1000,
  B = A * 2 + rnorm(length(A))
)
```

Given the above tibble, practice referring to non-syntactic names in the following data frame:

1. What has happened in creating both of them?
2. Extract the variables called A and B from the tibble.
3. On the tibble, create a new column called C which is B divided by A.
4. On the tibble, renaming the columns to First, Second and Third.

Importing data

R provides base functions to read flat files with different formats such as (comma-delimited, tab-delimited, semi-colon delimited, plain texts with multi-lines). To read a CSV file, you can use read.csv from the base R packages or \`read_csv\` from the \`readr\` package, which reads the contents of a file into a tibble. Please refer to slide #18 from week5's lecture for the different function instances to read and to write data into files.

Exercise 2 (read_csv)

Identify what is wrong with each of the following inline CSV files. What happens when you run each of them?

```
read_csv("a,b\n1,2,3\n4,5,6")
read_csv("a,b,c\n1,2\n1,2,3,4")
read_csv("a,b\n\"1")
read_csv("a,b\n1,2\na,b")
read_csv("a;b\n1;3")
```

Parsing variables

The `readr` package provides a way to guess the types of the columns as well as to parse the character contents into more specific type such as converting the "3 September, 2019" string to a proper formatted date type, as following:

```
parse_date("3 September, 2019", format="%d %B, %Y")
```

Exercise 3 (Parsing date)

Generate the correct format string to parse each of the following dates and times:

Note: Please refer to section (11.3.4) from (https://r4ds.had.co.nz/) for more details about parsing the different formats of date and datetime.

```
d1 <- "January 1, 2010"
d2 <- "2015-Mar-07"
d3 <- "06-Jun-2017"
d4 <- c("August 19 (2015)", "July 1 (2015)")
d5 <- "12/30/14" # Dec 30, 2014
t1 <- "1705"
t2 <- "11:15:10.12 PM"
```

Exercise 4 (Handling built-in dataset – Unsupervised activity)

Load the built-in `*airquality*` dataset. This data set includes daily air quality measurements in New York from May to September 1973 over a period of 5 months.

The variables of this dataset are described in the following table:

Variable	Type	Unit	Description
Ozone	numeric	parts per billion	mean Ozone concentration
Solar.R	numeric		Solar radiation
Wind	numeric	miles per hour	average wind speed
Temp	numeric	Fahrenheit	maximum daily temperature
Month	numeric		Month of observation
Day	numeric		day of month

1. Convert the data to a `tibble`
2. Find the observations, which include NA values
3. Add a new column, which states the date of each observation