STAT346: Statistical Data Science I

HW#4 – Due: November 25, 2020 by 6 p.m.

November 11, 2020

Instruction: Answer to the following questions and write your report using R Markdown. You should submit two files, through KU Black Board system (https://kulms.korea.ac.kr), which should have the following naming format:

- `stat346_hw4_your_id.rmd`
- `stat346_hw4_your_id.pdf` or `stat346_hw4_your_id.docx`

---

1. By using `gather` and `spread` from `tidyr` package, generate your code to convert the following data frame to wide format:

ID	grp	sex	meanL	sdL	meanR	sdR
1	A	F	0.22	0.11	0.34	0.08
2	A	M	0.47	0.33	0.57	0.33
3	B	F	0.33	0.11	0.40	0.07
4	B	M	0.55	0.31	0.65	0.27

   The result should look like the following display.

ID	F.meanL	F.meanR	F.sdL	F.sdR	M.meanL	M.meanR	M.sdL	M.sdR
1	0.22	0.33	0.11	0.08	0.47	0.57	0.33	0.33
2	0.33	0.40	0.11	0.07	0.55	0.65	0.31	0.27

---

2. In the `Marriage` data set included in `mosaicData`, the `appdate`, `ceremonydate`, and `dob` variables are encoded as factors, even though they are dates. Use the `lubridate` package to convert those three columns into a date format.

```
library(mosaic)
Marriage %>% select(appdate,ceremonydate,dob) %>%
  glimpse()
```

```
## Rows: 98
## Columns: 3
## $ appdate      <date> 1996-10-29, 1996-11-12, 1996-11-19, 1996-12-02, 1996-...
```

```
## $ ceremonydate <date> 1996-11-09, 1996-11-12, 1996-11-27, 1996-12-07, 1996-...
## $ dob          <date> 2064-04-11, 2064-08-06, 2062-02-20, 2056-05-20, 2066-...
```

---

3. Click **this link** to find a dataset,"China-Global-Investment-Tracker-2019-Spring-FINAL.xlsx", containing information about investments made by Chinese companies and government entities outside of China. This is an Excel spreadsheet. Use the `readxl` package to load the spreadsheet into R. Then use `dplyr` and `tidyr` to answer the following questions.

```
library(readxl)
data <- read_excel("China-Global-Investment-Tracker-2019-Spring-FINAL.xlsx",skip=5)

# Change column names
colnames(data) <- data %>% colnames() %>%
  str_replace_all(" ","_") %>% str_to_lower()

glimpse(data)
```

```
## Rows: 1,571
## Columns: 12
## $ year                 <dbl> 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005...
## $ month                <chr> "January", "January", "February", "March", "Ap...
## $ investor             <chr> "Minmetals", "China Academy of Sciences", "Min...
## $ quantity_in_millions <dbl> 500, 1740, 550, 670, 130, 120, 100, 4200, 1420...
## $ share_size           <chr> NA, NA, "0.5", "0.85", "0.17", "0.4", "1", "0....
## $ transaction_party    <chr> "Cubapetroleo", "IBM", "Codelco", "Highlands P...
## $ sector               <chr> "Metals", "Technology", "Metals", "Metals", "E...
## $ subsector            <chr> NA, NA, "Copper", "Steel", "Oil", "Oil", "Auto...
## $ country              <chr> "Cuba", "USA", "Chile", "Papua New Guinea", "C...
## $ region               <chr> "North America", "USA", "South America", "East...
## $ bri                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ greenfield           <chr> "G", NA, "G", "G", NA, "G", NA, NA, NA, "G", N...
```

(a) Use `dplyr` to write a simple check that every country is only listed as belonging to a single region. If the check fails for any country, provide details.

(b) Construct a table with the regions in the columns and the sectors in the rows, showing the proportion of the total investment for each region that was made in each sector. Shorten the column names and truncate the proportions so that the table fits on the screen without wrapping. Which sector most commonly receives the greatest share of investment? What are the exceptions to this finding?

(c) Construct a table showing the mean and standard deviation of the investment amounts in each sector, sorted by the means. Construct the same table for the log investment amounts. Describe the overall relationship between the means and standard deviations. Identify an exception to the overall pattern and explain it in more detail.

(d) Construct a table showing the total investment per `sector` (column) per `year` (row). State which sectors contributed the most to investment growth from 2005-2012, and which contributed the most from 2013-2015.

4. In this assignment, we will be working with the infant mortality data set, found here:

   `http://johnmuschelli.com/intro_to_r/data/indicatordeadkids35.csv`

   The packages listed below are simply suggestions, but please edit this list as you see fit.

```
library(tidyverse)
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
```

   (a) Read the data using `read_csv` and name it `mort`. Rename the first column to `country` using the `rename` command in `dplyr`. Create an object year variable by extracting column names (using `colnames`) and make it to an integer `as.integer`), excluding the first column either with string manipulations or bracket subsetting or subsetting with `is.na`.

   (b) Reshape the data so that there is a variable named `year` corresponding to `year` (key) and a column of the mortalities named `mortality` (value), using the `tidyr` package and its `gather` function. Name the output `long` and make `year` a numeric variable.

   (c) Read in this the tab-delim file and call it `pop`:

   `http://johnmuschelli.com/intro_to_r/data/country_pop.txt`.

   Use `read_tsv`. The file contains population information on each country. Rename the second column to `country` and the column `% of world population` to `percent`.

   (d) Determine the population of each country in `pop` using `arrange`. Get the order of the countries based on this (first is the highest population), and extract that column and call it `pop_levels`. Make a variable in the long data set named `sorted` that is the `country` variable coded as a factor based on `pop_levels`.

   (e) Parts (i)–(iii) below are only broken up here for clarity, but all three components can be addressed in one chunk of code/as one function, using `%>%` as necessary.

      (i) Subset `long` based on years 1975-2010, including 1975 and 2010 and call this `long_sub` using `&` or the `between` function.

      (ii) Further subset `long_sub` for the following countries using `filter` and the `%in%` operator on the sorted country factor (`sorted`): `c("Venezuela", "Bahrain", "Estonia", "Iran", "Thailand", "Chile", "Western Sahara", "Azerbaijan", "Argentina", "Haiti")`.

      (iii) Lastly, remove missing rows for `mortality` using `filter` and `is.na`.

   Be sure to assign your final object created from (i) through (iii) as `long_sub` so you can use it in the following.

   (f) Plotting: create "spaghetti"-line plots for the countries in `long_sub`, using different colors for different countries, using `sorted`. The x-axis should be `year`, and the y-axis should be `mortality`. Make the plot using (i) `qplot` and (ii) `ggplot`.