Homework 3

Due Wednesday September 19, 2019

2019-09-18

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about R and version control, munging and 'tidying' data, good programming practice and finally some basic programming building blocs. To begin the homework, we will for the rest of the course, start by loading data and then creating tidy data sets.

Problem 1

Work through the "Getting and Cleaning Data" lesson parts 3 and 4.

From the R command prompt:

```
library(swirl)
swirl()
```

Problem 2

Create a new R Markdown file within your local GitHub repo folder (file–>new–>R Markdown–>save as).

The filename should be: HW3_lastname, i.e. for me it would be HW3_Settlage

You will use this new R Markdown file to solve the following problems.

Problem 3

Redo Problem 4 parts a-d from last time using the tidyverse functions and piping.

Load packages

```
library(tidyverse)
```

a. Sensory data from five operators.

http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat

```
## import and format dataset
u1 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
d1 <- scan(u1, skip = 2)[-seq(from = 1, to = 145, by = 16)] %>%
      cbind(rep(1:10, each = 15), rep(1:5, times = 30), .) %>%
      "colnames<-" (c("item", "operate", "value")) %>%
      tbl_df(.)

## Display firt 5 rows
knitr::kable(head(d1, n = 5),
             caption = "First 5 rows")
```

Table 1: First 5 rows

item	operate	value
1	1	4.3
1	2	4.9
1	3	3.3
1	4	5.3
1	5	4.4

```
## summarize by item
knitr::kable(
  d1 %>%
  group_by(item) %>%
  summarise(
    mininum = min(value),
    median = median(value),
    mean = mean(value),
    maximum = max(value),
    "standard deviation" = sd(value)
    ),
  caption = "Summary Statistics of Sensory Value by Item"
)
```

Table 2: Summary Statistics of Sensory Value by Item

item	mininum	median	mean	maximum	standard deviation
1	3.3	4.4	4.466667	5.7	0.7788881
2	4.2	5.3	5.313333	6.3	0.7130084
3	1.3	2.6	2.773333	4.6	0.8413142
4	5.9	6.9	6.880000	8.2	0.6646159
5	4.9	5.9	5.920000	7.0	0.5002856
6	1.1	2.1	2.393333	4.0	0.8180697
7	0.7	1.2	1.406667	3.1	0.6419464
8	3.0	4.6	4.426667	4.9	0.5737927
9	6.7	8.8	8.466667	9.4	0.7612646
10	2.8	4.8	4.520000	5.5	0.8247943

```
## summarize by operate
knitr::kable(
  d1 %>%
  group_by(operate) %>%
  summarise(
    mininum = min(value),
    median = median(value),
    mean = mean(value),
    maximum = max(value),
    "standard deviation" = sd(value)
    ),
  caption = "Summary Statistics of Sensory Value by Operate"
)
```

Table 3: Summary Statistics of Sensory Value by Operate

operate	mininum	median	mean	maximum	standard deviation
1	0.9	4.55	4.593333	9.0	2.239140
2	1.5	4.95	5.063333	9.2	2.045429
3	0.8	4.15	4.166667	9.0	2.098494
4	0.9	5.40	5.193333	9.4	2.132334
5	0.7	4.60	4.266667	8.8	2.143206

b. Gold Medal performance for Olympic Men's Long Jump, year is coded as 1900=0.

http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat

```
## import and format dataset
u2 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"

d2.names <- readLines(u2, n = 1) %>%
            strsplit(., split = " ") %>%
            unlist(.)
d2.names <- c(d2.names[1], paste(d2.names[2], d2.names[3]))

d2 <- scan(u2, skip = 1) %>%
      as.numeric(.) %>%
      matrix(., byrow = T, ncol = 2) %>%
      "colnames<-" (d2.names) %>%
      tbl_df(.) %>%
      mutate(., Year = Year + 1900) %>%
      arrange(., Year)

## first 5 rows of dataset
knitr::kable(head(d2, n = 5),
             caption = "First 5 rows")
```

Table 4: First 5 rows

Year	Long Jump
1896	249.75
1900	282.88
1904	289.00
1908	294.50
1912	299.25

```
## Summary
knitr::kable(summary(d2[, 2]), align = "l")
```

Long Jump
Min. :249.8
1st Qu.:295.4
Median :308.1
Mean :310.3
3rd Qu.:327.5

Long Jump
Max. :350.5

c. Brain weight (g) and body weight (kg) for 62 species.

http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat

```
## 1. Overview of the dataset.
u3 <-
  "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"

d3.names <- readLines(u3)[1]  %>%
            strsplit(., split = " ") %>%
            unlist(.)  %>%
            unique(.)
d3.names <- c(paste(d3.names[1], d3.names[2]), paste(d3.names[3], d3.names[2]))



d3 <- readLines(u3)[2:22] %>%
      strsplit(., split = " ") %>%
      unlist(.) %>%
      as.numeric(.) %>%
      matrix(., byrow = T, ncol = 2) %>%
      "colnames<-" (d3.names)

## 3. Display the data and summry statistics
knitr::kable(head(d3, n = 5),
             caption = "First 5 rows")
```

Table 6: First 5 rows

Body Wt	Brain Wt
3.385	44.5
521.000	655.0
2.500	12.1
0.480	15.5
0.785	3.5

```
knitr::kable(summary(d3))
```

Body Wt	Brain Wt
Min. : 0.005	Min. : 0.10
1st Qu.: 0.600	1st Qu.: 4.25
Median : 3.342	Median : 17.25
Mean : 198.790	Mean : 283.13
3rd Qu.: 48.203	3rd Qu.: 166.00
Max. :6654.000	Max. :5712.00

d. Triplicate measurements of tomato yield for two varieties of tomatos at three planting densities.

http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat

```
## 1. Overview of the dataset.
u4 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"

## 2. Clean the data
d4 <- readLines(u4)

d4.r3 <- unlist(strsplit(d4[3], split = "\\s+"))
d4.r4 <- unlist(strsplit(d4[4], split = "\\s+"))

d4.variety <- c(d4.r3[1], d4.r4[1]) %>%
              rep(., each = 9)

d4.density <- as.numeric(unlist(strsplit(d4[2], split = "\\s+")))[-1] %>%
              rep(., times = 2, each = 3)


d4.yield <- as.numeric(unlist(strsplit(d4.r3[2:4], split = ",")),
                       unlist(strsplit(d4.r4[2:4], split = ",")))

d4 <- data.frame(d4.variety, d4.density, d4.yield) %>%
      tbl_df() %>%
      "colnames<-" (c("Variety", "Density", "Yield"))

##3. Display data and summary statistics
knitr::kable(head(d4, n = 5),
             caption = "First 5 rows")
```

Table 8: First 5 rows

Variety	Density	Yield
Ife#1	10000	16.1
Ife#1	10000	15.3
Ife#1	10000	17.5
Ife#1	20000	16.6
Ife#1	20000	19.2

```
knitr::kable(
  d4 %>%
  group_by(Variety) %>%
  summarise(
        mininum = min(Yield),
        median = median(Yield),
        mean = mean(Yield),
        maximum = max(Yield),
        "standard deviation" = sd(Yield)
      ),
  caption = "Summary Statistics of Yield by Variety"
)
```

Table 9: Summary Statistics of Yield by Variety

Variety	mininum	median	mean	maximum	standard deviation
Ife#1	15.3	18	18.11111	21	1.985223
PusaEarlyDwarf	15.3	18	18.11111	21	1.985223

```
knitr::kable(
  d4 %>%
  group_by(Density) %>%
  summarise(
        mininum = min(Yield),
        median = median(Yield),
        mean = mean(Yield),
        maximum = max(Yield),
        "standard deviation" = sd(Yield)
      ),
  caption = "Summary Statistics of Yield by Density"
)
```

Table 10: Summary Statistics of Yield by Density

Density	mininum	median	mean	maximum	standard deviation
10000	15.3	16.1	16.30000	17.5	0.995992
20000	16.6	18.5	18.10000	19.2	1.203329
30000	18.0	20.8	19.93333	21.0	1.500222

Problem 4

Finish this homework by pushing your changes to your repo. In general, your workflow for this should be:

1. In terminal: git pull – to make sure you have the most recent local repo
2. In terminal: do some work
3. In terminal: git add – check files you want to commit
4. In terminal: git commit – make message INFORMATIVE and USEFUL
5. In terminal: git push – this pushes your local changes to the repo

If you have difficulty with steps 1-5, git is not correctly or completely setup.

Only submit the .Rmd and .pdf solution files. Names should be formatted HW3_lastname_firstname.Rmd and HW3_lastname_firstname.pdf

Optional preperation for next class:

TBD – could be something sent as a class message