HW2_Jingbin

Jingbin Xu

8/30/2020

Problem 3

Version control helps me collaborate with other team members, and we could reverse back if any mistakes or changes happened.

Problem 4

a. Sensory data from five operators

We are looking at the sensory data from Wu and Hamada's textbook: http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat.

First, we will get the data from the link above:

```
# getting "<http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat>"

url_sensory <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
#sensory_data_raw <- fread(url_sensory, header = F, fill = T, data.table = F)

# Save data as RDS format to protect against the website going down

#saveRDS(sensory_data_raw, "sensory_data_raw.RDS")
sensory_data_raw <- readRDS("sensory_data_raw.RDS")
```

Next, we proceed to data clearning with base R function.

```
# Data cleaning to fix the NA and creat columns
# Romove first two row
# dim(sensory_data_raw)
sensory_data <- sensory_data_raw[-1:-2, ]
sensory_data$V1<- as.numeric(sensory_data$V1)

# Indexing the dimension of the dataset
nrow <- dim(sensory_data)[1]
ncol <- dim(sensory_data)[2]

# Remove the NA values in the dataset
for (i in 1:nrow) {
  if (is.na(sensory_data[i,ncol])){
    sensory_data[i,1:ncol] <- c(sensory_data[i-1,1],sensory_data[i,1:ncol-1])
  }
}
# Rename the columns by item and operator

colnames(sensory_data) <- c("Item", "Operator1", "Operator2", "Operator3",
```

```
                              "Operator4", "Operator5")
# Rename the level of the items
sensory_data$Item <- as.factor(sensory_data$Item)
levels(sensory_data$Item) <- c("One", "Two", "Three", "Four", "Five",
                               "Six", "Seven", "Eight", "Nine", "Ten")
```

Then, by using tidyverse, we could group data by item and operator.

```
# Explore the datset by item, operator, value
sensory_data_tv <- sensory_data %>%
  gather(key = "Operator", value = "Value",
         "Operator1",
         "Operator2",
         "Operator3",
         "Operator4",
         "Operator5")

# Explore the mean by item and operator
summary_sensory <- sensory_data_tv %>%
  group_by(Item) %>%
  summarize(Avg = mean(Value, 2)) %>%
  arrange(Item)
```

Finally, we could present the table.

Table 1: Sensory data summary by Base R

Item	Operator1	Operator2	Operator3	Operator4	Operator5
One : 3	Min. :0.900	Min. :1.500	Min. :0.800	Min. :0.900	Min. :0.700
Two : 3	1st Qu.:2.850	1st Qu.:3.450	1st Qu.:2.650	1st Qu.:3.925	1st Qu.:2.250
Three : 3	Median :4.550	Median :4.950	Median :4.150	Median :5.400	Median :4.600
Four : 3	Mean :4.593	Mean :5.063	Mean :4.167	Mean :5.193	Mean :4.267
Five : 3	3rd Qu.:5.950	3rd Qu.:6.225	3rd Qu.:5.400	3rd Qu.:6.275	3rd Qu.:5.800
Six : 3	Max. :9.000	Max. :9.200	Max. :9.000	Max. :9.400	Max. :8.800
(Other):12	NA	NA	NA	NA	NA

Table 2: Sensory data summary by Tidyverse

Item	Avg
One	4.4
Two	5.3
Three	2.6
Four	6.9
Five	5.9
Six	2.1
Seven	1.2
Eight	4.6
Nine	8.8
Ten	4.8

b. Gold Mmdal performance for Olympic men's long jump

We are looking at the sensory data from Wu and Hamada's textbook: http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat

First we will get the data from the link above:

```
# getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
url_lj <- "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
longjump_data_raw <- fread(url_lj, fill=TRUE, skip = 1, data.table = FALSE)
colnames(longjump_data_raw) <- c("year1", "jump1",
                                 "year2", "jump2",
                                 "year3", "jump3",
                                 "year4", "jump4")
saveRDS(longjump_data_raw, "longjump_data_raw.RDS")
longjump_data_raw <- readRDS("longjump_data_raw.RDS")
```

Then, we clean the data using baseR function.

```
# Get the column year of the jump data
longjump_data_year <- c(longjump_data_raw$year1,
                        longjump_data_raw$year2,
                        longjump_data_raw$year3,
                        longjump_data_raw$year4)
# Get the column jump
longjump_data_jump <- c(longjump_data_raw$jump1,
                        longjump_data_raw$jump2,
                        longjump_data_raw$jump3,
                        longjump_data_raw$jump4)
# Combine the column of year and jump, remove NA rows using na.omit
lj_data <- na.omit(as.data.frame(cbind(Year = longjump_data_year+1900,
                                       Jump = longjump_data_jump)))
```

Next, we clean the data using tidyverse.

```
# Compute the average jump prior to 1950
lj_data_1950_less <- lj_data %>% filter(Year < 1950) %>%
                                 summarise(Avg = mean(Jump))
# Compute the average jump later than 1950
lj_data_1950_more <- lj_data %>% filter(Year > 1950) %>%
                                 summarise(Avg = mean(Jump))
# Summary statistics
summary_1950 <- as.data.frame(cbind(Year = c("< 1950", "> 1950"),
                                    Junp = c(round(lj_data_1950_less[1, ],2),
                                             round(lj_data_1950_more[1, ],2))))
```

We present the summarize table and have following findings.

- The plot shows as jump distance increasing over time. There may exists a linear trend.

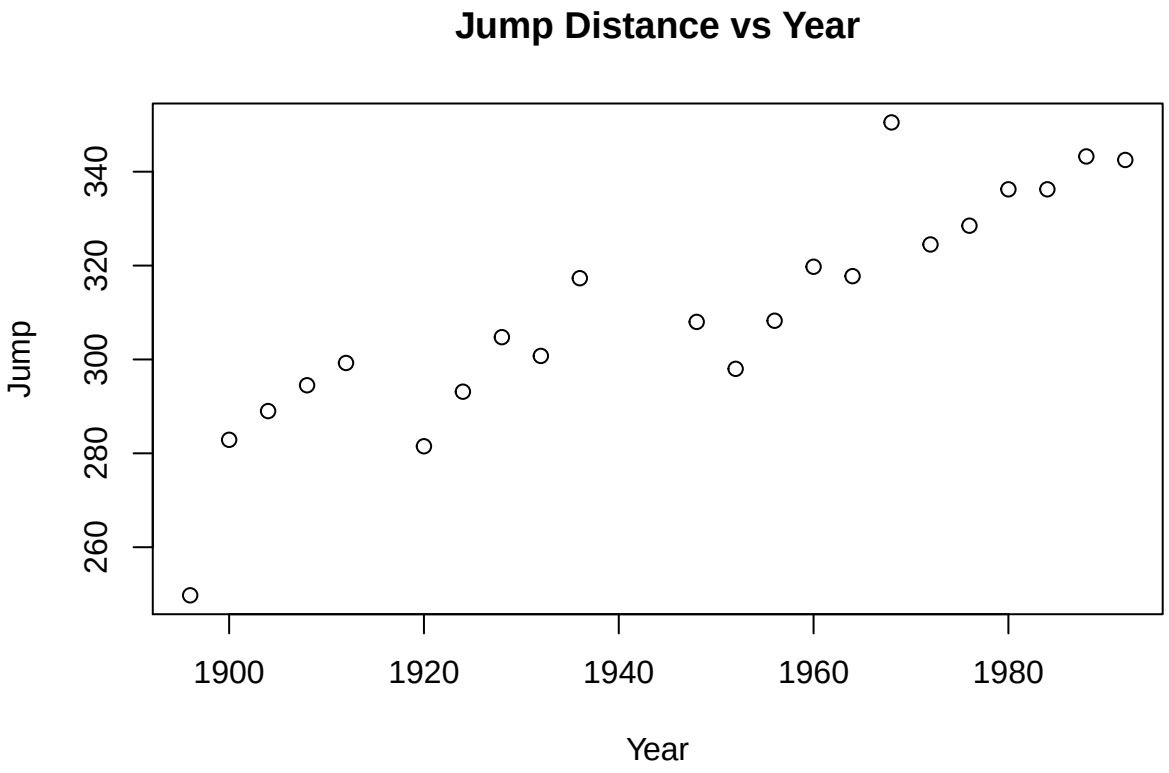
Table 3: Jump data summary by Base R

Year	Jump
Min. :1896	Min. :249.8
1st Qu.:1921	1st Qu.:295.4
Median :1950	Median :308.1
Mean :1945	Mean :310.3

Year	Jump
3rd Qu.:1971	3rd Qu.:327.5
Max. :1992	Max. :350.5

Table 4: Jump data summary by Tidyverse

Year	Junp
< 1950	292.8
> 1950	327.77



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

We obtained the data for brain weight and boday weight for 62 species from the following link:

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

```
# getting data
url_brain <- "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
brain_data_raw <- fread(url_brain, header = T, fill = T)
saveRDS(brain_data_raw, "brain_data_raw.RDS")
brain_data_raw <- readRDS("brain_data_raw.RDS")
```

Next, we proceed to formulate new data using base R function.

```
# Get the body weight column
body <- as.vector(unlist(c(brain_data_raw[, 1],
```

```
                            brain_data_raw[, 3],
                            brain_data_raw[, 5])))
# Get the brain weight column
brain <- as.vector(unlist(c(brain_data_raw[, 2],
                            brain_data_raw[, 4],
                            brain_data_raw[, 6])))
# Combine the dataset
b_data <- as.data.frame(cbind(Body = body,
                              Brain = brain))
```

Then, using tidyverse function to manipulate our data

```
# Top 5 brain weight
top_5_brw <- b_data %>% top_n(5, Brain) %>%
                        arrange(desc(Brain))
# Top 5 body weight
top_5_bdw <- b_data %>% top_n(5, Body) %>%
                        arrange(desc(Body))
```

We could summarize the folloing table and have following finding:

- By boxplots, we find that outliers exist in the dataset, we need to further examine the outliers.

Table 5: Brain and body data summary by Base R

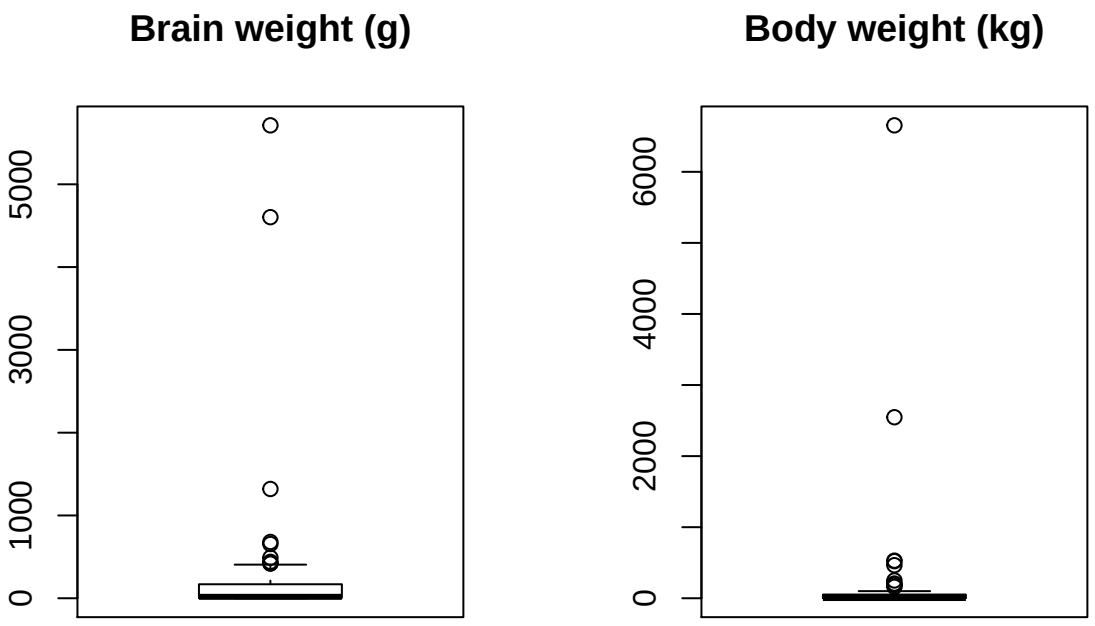
Body	Brain
Min. : 0.005	Min. : 0.10
1st Qu.: 0.600	1st Qu.: 4.25
Median : 3.342	Median : 17.25
Mean : 198.790	Mean : 283.13
3rd Qu.: 48.203	3rd Qu.: 166.00
Max. :6654.000	Max. :5712.00
NA's :1	NA's :1

Table 6: Top 5 body weight data summary by Tidyverse

Body	Brain
6654	5712
2547	4603
529	680
521	655
465	423

Table 7: Top 5 brain weight data summary by Tidyverse

Body	Brain
6654	5712
2547	4603
62	1320
529	680
521	655



d. Triplicate measurements of tomato yield

By the following website link, we obtained our dataset for triplicate measurements of tomato yield. http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat

First, let's import the web data.

```
# getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"
url_to <- "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"
to_data_raw <- fread(url_to, header = F)
saveRDS(to_data_raw, "to_data_raw.RDS")
to_data_raw <- readRDS("to_data_raw.RDS")
```

```
## Convert columns and rows
# group 1: lfe
yield_10_g1 <- as.numeric(unlist(str_split(to_data_raw[1,2], ","), recursive = T))
yield_20_g1 <- as.numeric(unlist(str_split(to_data_raw[1,3], ","), recursive = T))
yield_30_g1 <- as.numeric(unlist(str_split(to_data_raw[1,4], ","), recursive = T))

# group 2: pusa
yield_10_g2 <- as.numeric(unlist(str_split(to_data_raw[2,2], ","), recursive = T))[1:3]
yield_20_g2 <- as.numeric(unlist(str_split(to_data_raw[2,3], ","), recursive = T))
yield_30_g2 <- as.numeric(unlist(str_split(to_data_raw[2,4], ","), recursive = T))

# combine data to dataframe
to_data <- data.frame(group1 = c(yield_10_g1, yield_20_g1, yield_30_g1),
                      group2 = c(yield_10_g2, yield_20_g2, yield_30_g2),
                      density = as.factor(sort(rep(c(10,20,30), 3))))
```

```
# Compute the avg of measure for group 1 by density
to_data_group1 <- to_data %>% group_by(density) %>%
  summarise(Avg = mean(group1,2))
# Compute the avf of measure for group 2 by density
to_data_group2 <- to_data %>% group_by(density) %>%
```

```
  summarise(Avg = mean(group2,2))
# Combine table
summary_dens <- as.data.frame(cbind(Group1 = to_data_group1,
                                    Group2 = to_data_group2))
```

Then we finalize with the summary table. And have the following finding:

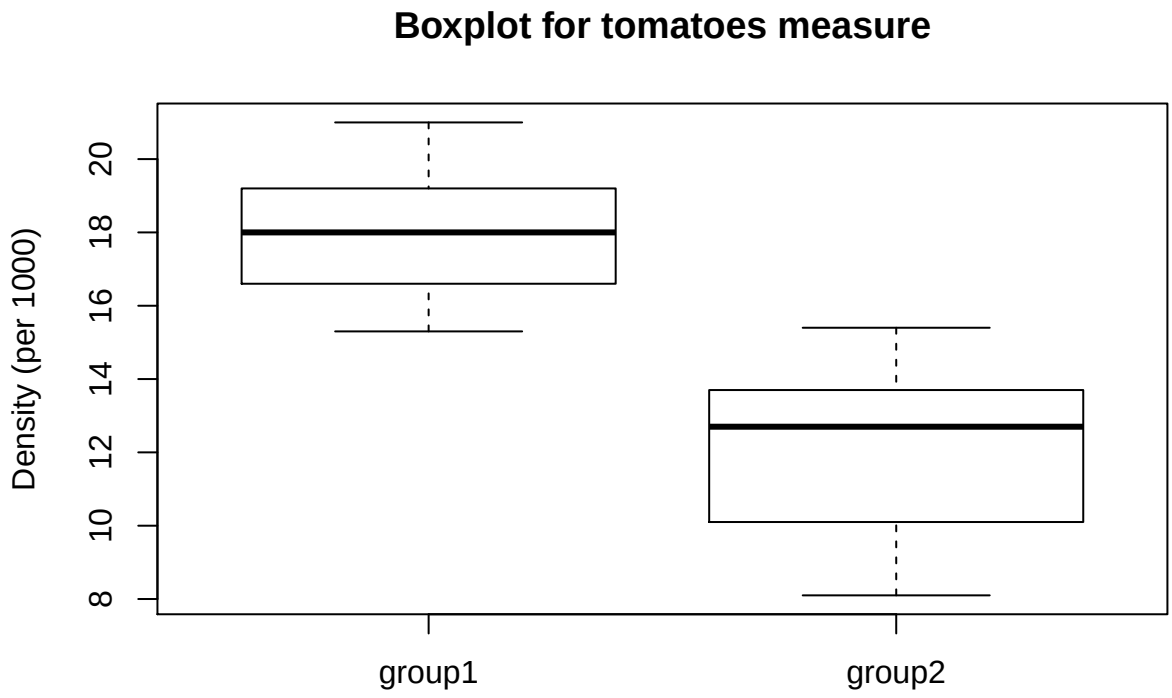
- Overall, group 1 has relatively higher measure of tomatoes than group 2.

Table 8: Tomato measure data summary by Base R

group1	group2	density
Min. :15.30	Min. : 8.10	10:3
1st Qu.:16.60	1st Qu.:10.10	20:3
Median :18.00	Median :12.70	30:3
Mean :18.11	Mean :12.02	NA
3rd Qu.:19.20	3rd Qu.:13.70	NA
Max. :21.00	Max. :15.40	NA

Table 9: Tomato measure data summary by Tidyverse

Group1.density	Group1.Avg	Group2.density	Group2.Avg
10	16.1	10	8.6
20	18.5	20	12.7
30	20.8	30	14.4



Problem 5

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

1. git pull – to make sure you have the most recent repo
2. In R: do some work
3. git add – this tells git to track new files
4. git commit – make message INFORMATIVE and USEFUL
5. 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. See me for help.

Only submit the .Rmd and .pdf solution files. Names should be formatted HW2_lastname.Rmd and HW2_lastname.pdf