HW2-906338789

Xueying Liu

9/10/2020

Problem 3

I think I will definitely use it since I am always revising my codes with with new idea, so it is useful when I make a mistake and want to revert back and campare the old version with the new one. It can also be helpful by sharing it with other people so that people can work together on the same problem.

Problem 4

a. Sensory data from five operators

```
## getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
url="https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
sensory_data_raw=fread(url,header = TRUE,fill=TRUE,skip="Item",data.table = FALSE)
saveRDS(sensory_data_raw,"sensory_data_raw.RDS")
sensory_data_raw=readRDS("sensory_data_raw.RDS")
## filling the first column with Item number
for(i in 0:9)
{
  sensory_data_raw[(3*i+2):(3*i+3),]=c(i+1,sensory_data_raw[(3*i+2):(3*i+3),])
  }
```

To tidy the data, we need to push operator into a column.

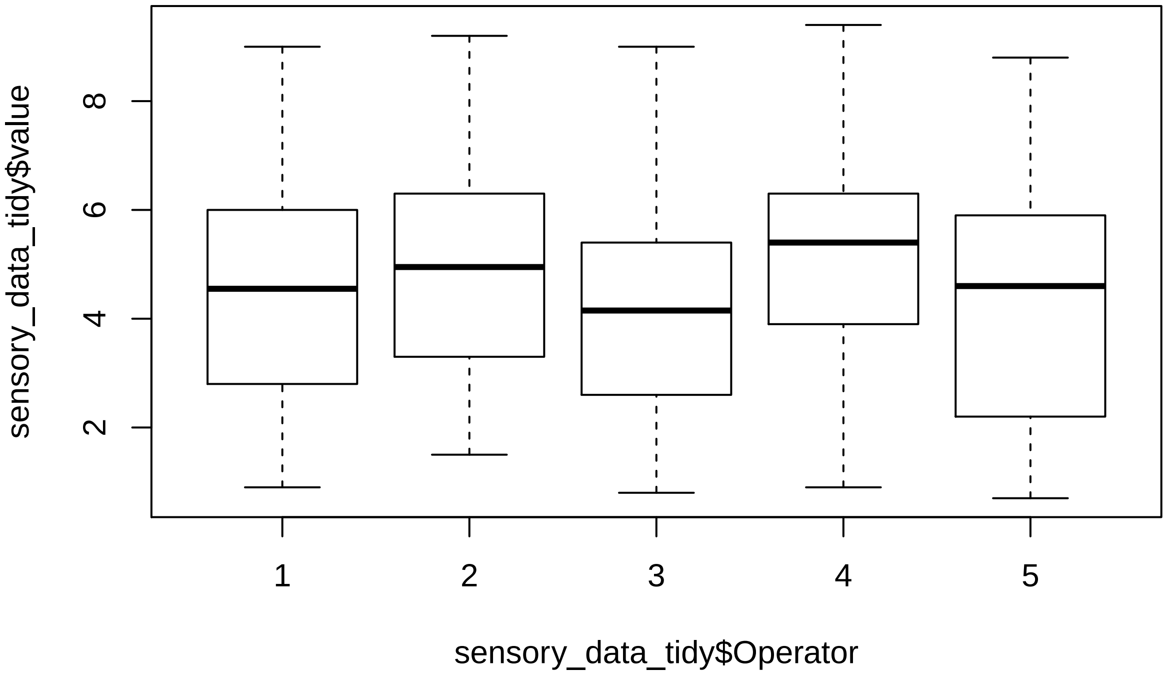
```
sensory_data_tidy=data.frame(rep(sensory_data_raw$Item,5),stack(sensory_data_raw[,-1]))
colnames(sensory_data_tidy)=c("Item","value","Operator")
head(sensory_data_tidy)
```

```
##   Item value Operator
## 1    1   4.3        1
## 2    1   4.3        1
## 3    1   4.1        1
## 4    2   6.0        1
## 5    2   4.9        1
## 6    2   6.0        1
```

We have converted the dataframes to tidy data frames using the base function. Here is a summary and boxplot of the data:

Item	value	Operator
Min. : 1.0	Min. :0.700	1:30
1st Qu.: 3.0	1st Qu.:3.025	2:30
Median : 5.5	Median :4.700	3:30

Item	value	Operator
Mean : 5.5	Mean :4.657	4:30
3rd Qu.: 8.0	3rd Qu.:6.000	5:30
Max. :10.0	Max. :9.400	NA



Then we choose to use *tidyverse*() function to tidy the raw data:

```
# stack and fix column names using tidyverse
sensory_data_tv=sensory_data_raw %>%
                     gather(key="operator",value="value",2:6)
head(sensory_data_tv)
```

```
##   Item operator value
## 1    1        1   4.3
## 2    1        1   4.3
## 3    1        1   4.1
## 4    2        1   6.0
## 5    2        1   4.9
## 6    2        1   6.0
```

b. Gold Medal performance for Olympic Men's Long Jump

```
## getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
url="https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
gold_data_raw=fread(url,data.table = FALSE)
```

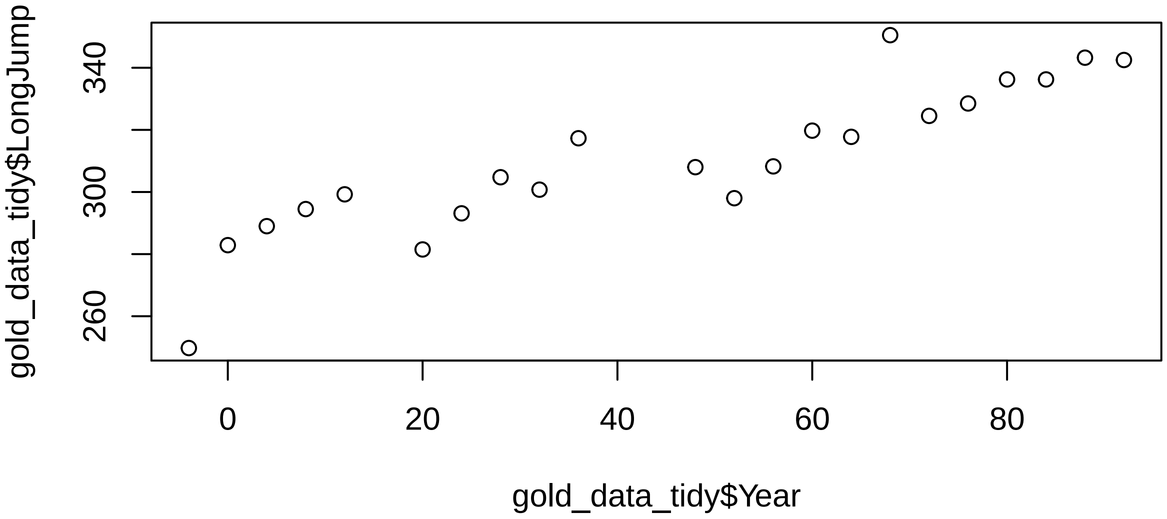
```r
saveRDS(gold_data_raw,"gold_data_raw.RDS")
gold_data_raw=readRDS("gold_data_raw.RDS")
gold_data_raw=gold_data_raw[,1:8]
colnames(gold_data_raw)=c("Year","LongJump","Year","LongJump","Year","LongJump","Year","LongJump")
```

```r
gold_data_tidy=data.frame(rbind(gold_data_raw[,1:2],gold_data_raw[,3:4]
                                ,gold_data_raw[,5:6],gold_data_raw[,7:8]))
## Drop the raws with missing value
gold_data_tidy=DropNA(gold_data_tidy)
head(gold_data_tidy)
```

```
##   Year LongJump
## 1   -4   249.75
## 2    0   282.88
## 3    4   289.00
## 4    8   294.50
## 5   12   299.25
## 6   20   281.50
```

We have converted the dataframes to tidy data frames using the base function. Here is a summary and plot of the data:

Year	LongJump
Min. :-4.00	Min. :249.8
1st Qu.:21.00	1st Qu.:295.4
Median :50.00	Median :308.1
Mean :45.45	Mean :310.3
3rd Qu.:71.00	3rd Qu.:327.5
Max. :92.00	Max. :350.5



Then we choose to use *tidyverse*() function to tidy the raw data:

```
# stack and fix column names using tidyverse
gold_data_tv=data.frame(gather(gold_data_raw,key = "year",value="year",1,3,5,7)[,6],
                        gather(gold_data_raw,key="LongJump",value="LongJump",2,4,6,8)[,6])
colnames(gold_data_tv)=c("Year","LongJump")
head(gold_data_tv)
```

```
##   Year LongJump
## 1   -4   249.75
## 2    0   282.88
## 3    4   289.00
## 4    8   294.50
## 5   12   299.25
## 6   20   281.50
```

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

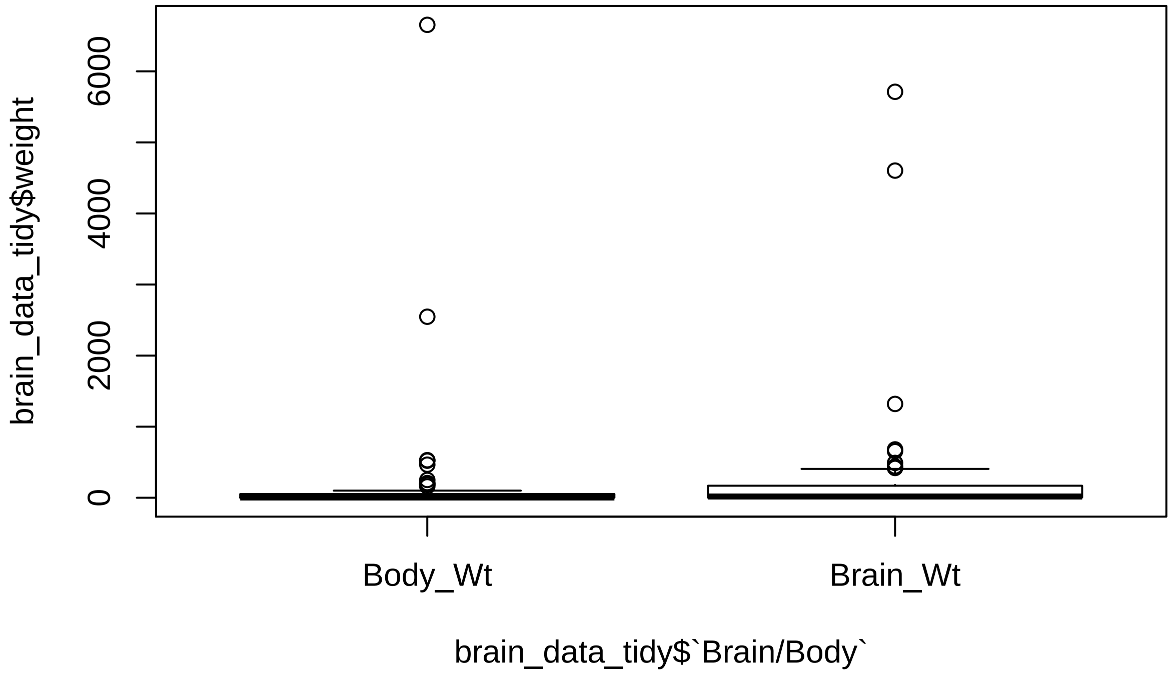
```
## getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
url="https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
brain_data_raw=fread(url,data.table = FALSE,header =TRUE,fill=TRUE)
saveRDS(brain_data_raw,"brain_data_raw.RDS")
brain_data_raw=readRDS("brain_data_raw.RDS")
colnames(brain_data_raw)=c(rep(c("Body_Wt","Brain_Wt"),3))
```

```
brain_data_rbind=DropNA(data.frame(rbind(brain_data_raw[,1:2],
                                         brain_data_raw[,3:4],brain_data_raw[,5:6])))
brain_data_tidy=data.frame(stack(brain_data_rbind))
colnames(brain_data_tidy)=c("weight","Brain/Body")
head(brain_data_tidy)
```

```
##    weight Brain/Body
## 1   3.385    Body_Wt
## 2   0.480    Body_Wt
## 3   1.350    Body_Wt
## 4 465.000    Body_Wt
## 5  36.330    Body_Wt
## 6  27.660    Body_Wt
```

We have converted the dataframes to tidy data frames using the base function. Here is a summary of the data:

weight	Brain/Body
Min. : 0.005	Body_Wt :62
1st Qu.: 1.387	Brain_Wt:62
Median : 7.450	NA
Mean : 240.962	NA
3rd Qu.: 98.650	NA
Max. :6654.000	NA



Then we choose to use *tidyverse*() function to tidy the raw data:

```
# stack and fix column names using tidyverse
brain_data_tv=gather(brain_data_rbind,key="Brain/Body",value="value",Body_Wt:Brain_Wt)
head(brain_data_tv)
```

```
##   Brain/Body   value
## 1    Body_Wt   3.385
## 2    Body_Wt   0.480
## 3    Body_Wt   1.350
## 4    Body_Wt 465.000
## 5    Body_Wt  36.330
## 6    Body_Wt  27.660
```

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

```
## getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"
url="https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"
tomato_data_raw=fread(url,data.table = FALSE,skip="1000")
saveRDS(tomato_data_raw,"tomato_data_raw.RDS")
tomato_data_raw=readRDS("tomato_data_raw.RDS")
```

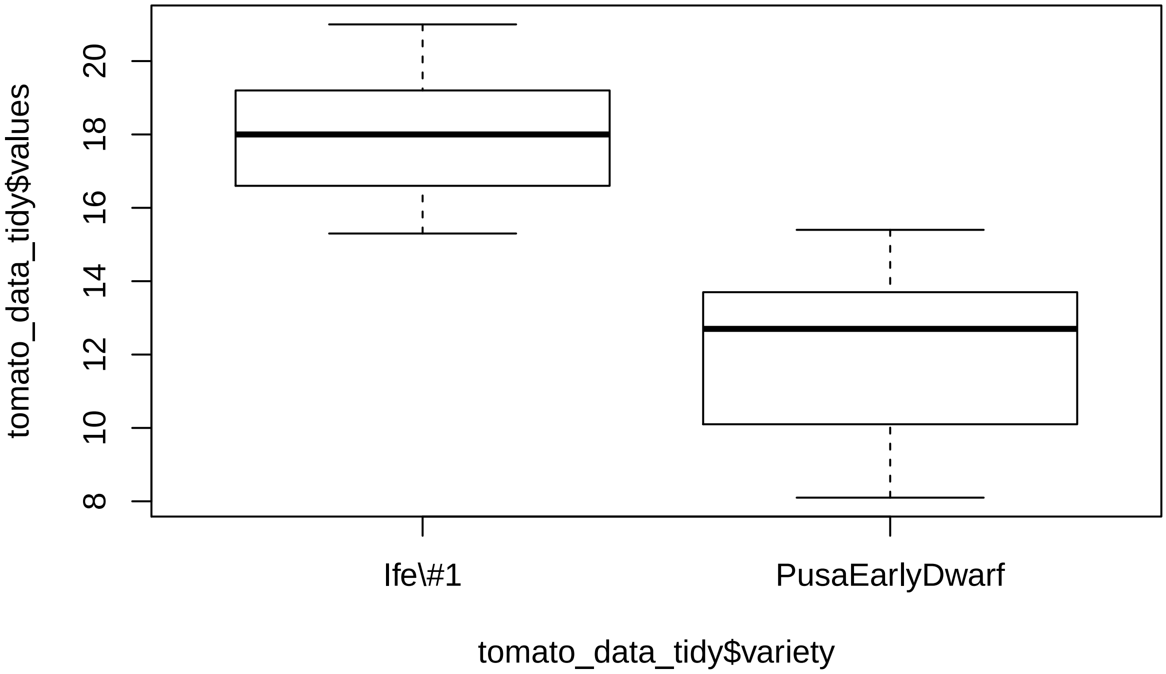
```
tomato_data=data.frame(varity=rep(tomato_data_raw[1:2,1],3),stack(tomato_data_raw[,-1]))
colnames(tomato_data)=c("variety","value","planting_density")
value=do.call("rbind", strsplit(tomato_data$value, ","))
value=data.frame(apply(value,2,as.numeric))
```

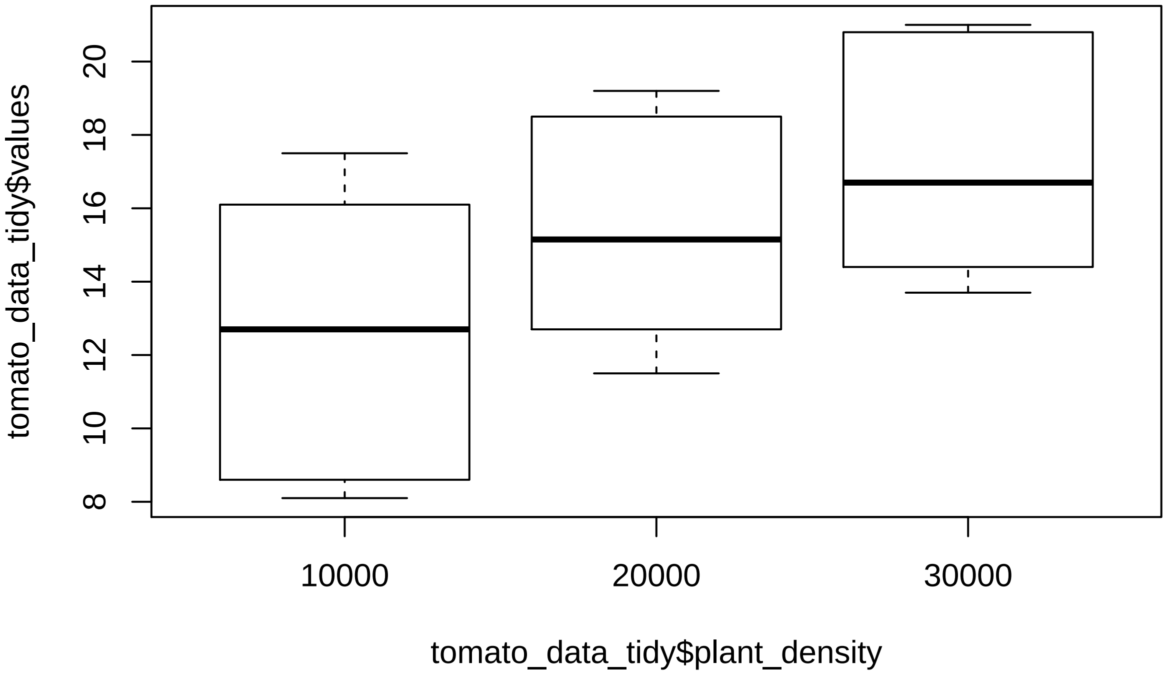
```
colnames(value) = c("value1","value2","value3")
tomato_data_split=data.frame(cbind(tomato_data$variety,value,tomato_data$planting_density))
tomato_data_tidy=data.frame(variety=rep(tomato_data_split[,1],3),stack(tomato_data_split),
                             plant_density=rep(tomato_data_split[,5],3))[,-3]
head(tomato_data_tidy)
```

```
##          variety values plant_density
## 1        Ife\\#1   16.1         10000
## 2 PusaEarlyDwarf    8.1         10000
## 3        Ife\\#1   16.6         20000
## 4 PusaEarlyDwarf   12.7         20000
## 5        Ife\\#1   20.8         30000
## 6 PusaEarlyDwarf   14.4         30000
```

We have converted the dataframes to tidy data frames using the base function. Here is a summary and boxplots of the data:

variety	values	plant_density
Ife#1 :9	Min. : 8.10	10000:6
PusaEarlyDwarf:9	1st Qu.:12.95	20000:6
NA	Median :15.35	30000:6
NA	Mean :15.07	NA
NA	3rd Qu.:17.88	NA
NA	Max. :21.00	NA





Then we choose to use *tidyverse*() function to tidy the raw data:

```
# stack and fix column names using tidyverse
tomato_data_tv=gather(tomato_data_split,key="value",value="value",value1:value3)[,-3]
head(tomato_data_tv)
```

```
##   tomato_data.variety tomato_data.planting_density value
## 1             Ife\\#1                        10000  16.1
## 2      PusaEarlyDwarf                        10000   8.1
## 3             Ife\\#1                        20000  16.6
## 4      PusaEarlyDwarf                        20000  12.7
## 5             Ife\\#1                        30000  20.8
## 6      PusaEarlyDwarf                        30000  14.4
```