STAT5014 Homewoek 2

Yixin Chen

09/13/2020

Problem 3

Version control will help me keep track of any changes I made to the project. It will be needed in the final project of this course. Because there will be multiple people working on a project and we need to review the history of the code during the project.

Problem 4

a. Sensory data from five operators.
http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat

First, import the data.

```
url1 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
sensory_raw <- fread(url1,fill = T, skip = 1, data.table = F)
saveRDS(sensory_raw, "sensory.RDS")
sensory_raw <- readRDS("sensory.RDS")
```

Second, clean the data with base R.

```
sensory_tidy_br <- data.frame(item = rep(seq(1,10), each=15),
                              operator = rep(c(1:5),30),
                              n = NA)
temp_n<-matrix(NA,15,10)
for (i in 1:10) {
  temp_n[,i] <- as.numeric(c(sensory_raw[3*i-2,-1], sensory_raw[3*i-1,-6],
                             sensory_raw[3*i,-6]))
}
sensory_tidy_br$n <- as.vector(temp_n)
```

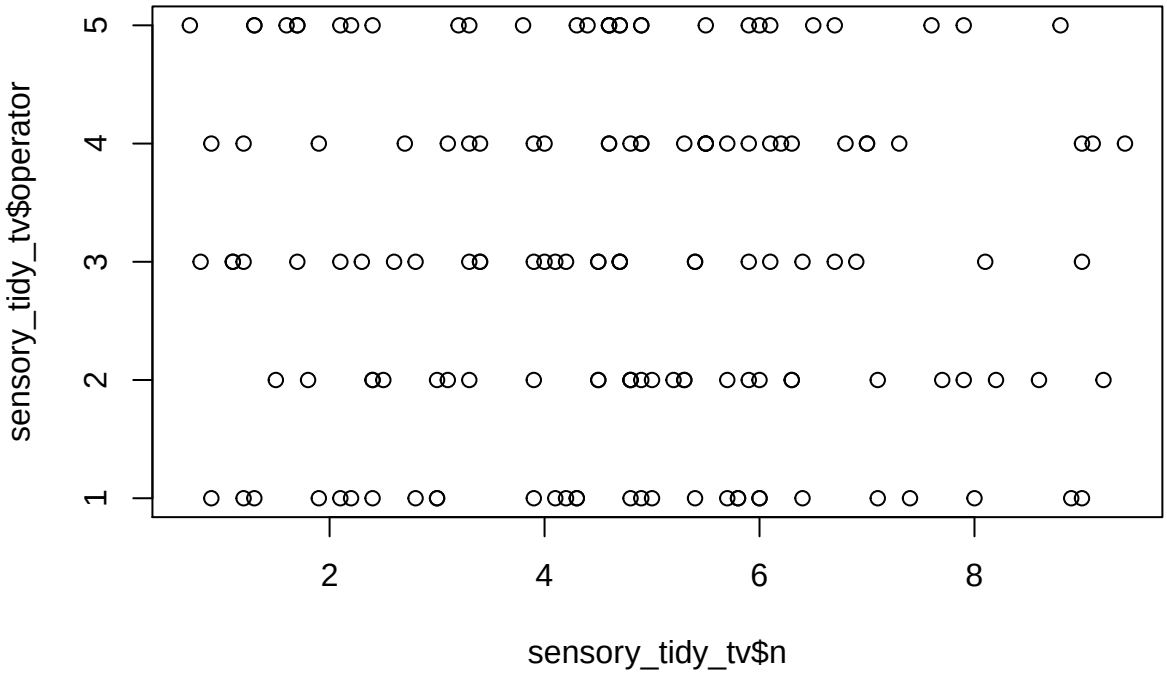
Third, clean the data with tidyverse

```
  temp<-sensory_raw
  colnames(temp) <- c("Item","op1","op2","op3","op4","op5")
  temp[is.na(temp)]<-0
  temp1<-temp %>% filter(op5 == 0)
  temp2<-temp %>% filter(op5 != 0)
  temp1[,-1]<-temp1[,-6]
  temp1[,1]<-rep(c(1:10),each =2)
  sensory_tidy_tv<-bind_rows(temp1,temp2)
  sensory_tidy_tv <- sensory_tidy_tv %>%
                          gather(key = operator, value = "n",-1) %>%
                          arrange(Item)
  sensory_tidy_tv$operator<-rep(rep(c(1:5),each =3),10)
```

Here is the summary for sensory data:

Item	operator	n
Min. : 1.0	Min. :1	Min. :0.700
1st Qu.: 3.0	1st Qu.:2	1st Qu.:3.025
Median : 5.5	Median :3	Median :4.700
Mean : 5.5	Mean :3	Mean :4.657
3rd Qu.: 8.0	3rd Qu.:4	3rd Qu.:6.000
Max. :10.0	Max. :5	Max. :9.400

Here is a scatter plot for n versus Operator:



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

First, import the data.

```
url2 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
longj_raw <- fread(url2,skip = 1,fill = T)
saveRDS(longj_raw, "longj.RDS")
longj_raw <- readRDS("longj.RDS")
```

Second, clean the data with base R.

```
longj_tidy_br <- data.frame(Year = c(longj_raw$V1, longj_raw$V3, longj_raw$V5, longj_raw$V7),
                            LongJump = c(longj_raw$V2, longj_raw$V4, longj_raw$V6, longj_raw$V8))
longj_tidy_br <- longj_tidy_br[1:22,]
```

Third, clean the data with tidyverse.

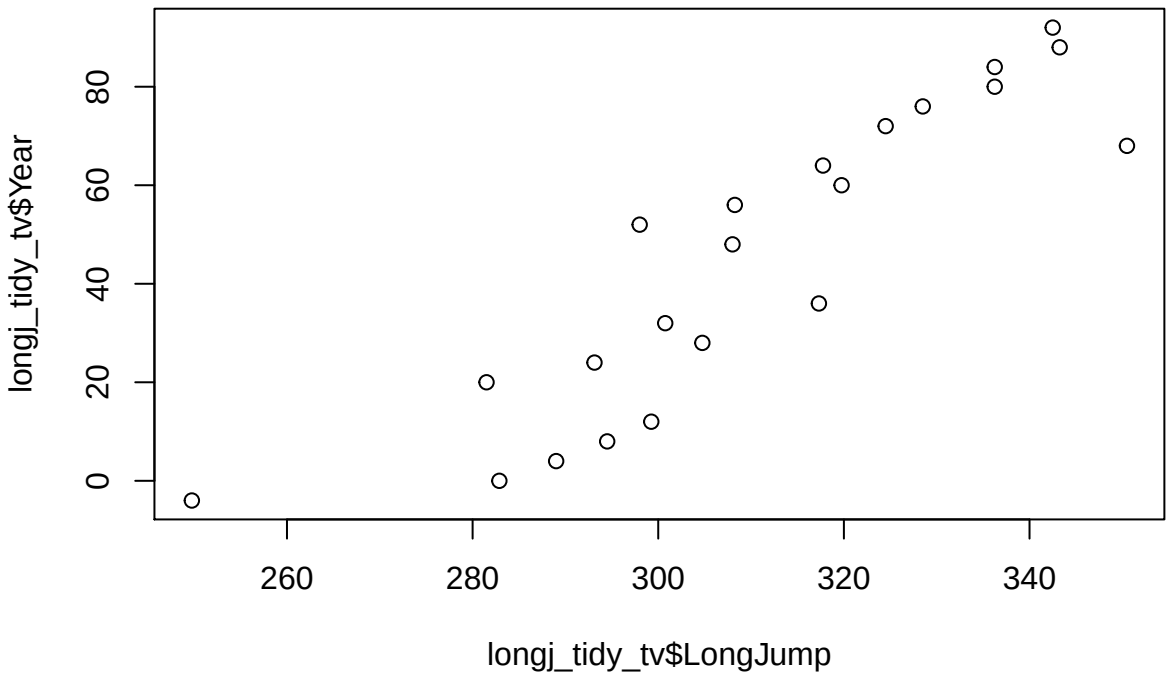
```
#colnames(longj_raw) <- rep(c("Year","LongJump"),4)
year <- longj_raw %>%
```

```
                  select(V1,V3,V5,V7) %>%
                  gather("V",value="Year",1,2,3,4) %>%
                  select(Year) %>%
                  slice(1:22)
longjump <- longj_raw %>%
                  select(V2,V4,V6,V8) %>%
                  gather("V",value="LongJump",1,2,3,4) %>%
                  select(LongJump) %>%
                  slice(1:22)
longj_tidy_tv <- bind_cols(year, longjump)
```

Here is the summary for long jump 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

Here is a scatter plot for long jump data:



c. Brain weight (g) and body weight (kg) for 62 species.
http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat

First, import the data.

```
url3 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
bbweight_raw <- fread(url3,skip = 1,fill = T)
saveRDS(bbweight_raw, "bbweight.RDS")
bbweight_raw <- readRDS("bbweight.RDS")
```

Second, clean the data with base R.

```
bbweight_tidy_br <- data.frame(BodyWt = c(bbweight_raw$V1, bbweight_raw$V3, bbweight_raw$V5),
                          BrainWt = c(bbweight_raw$V2, bbweight_raw$V4, bbweight_raw$V6))
bbweight_tidy_br <- bbweight_tidy_br[1:62,]
```

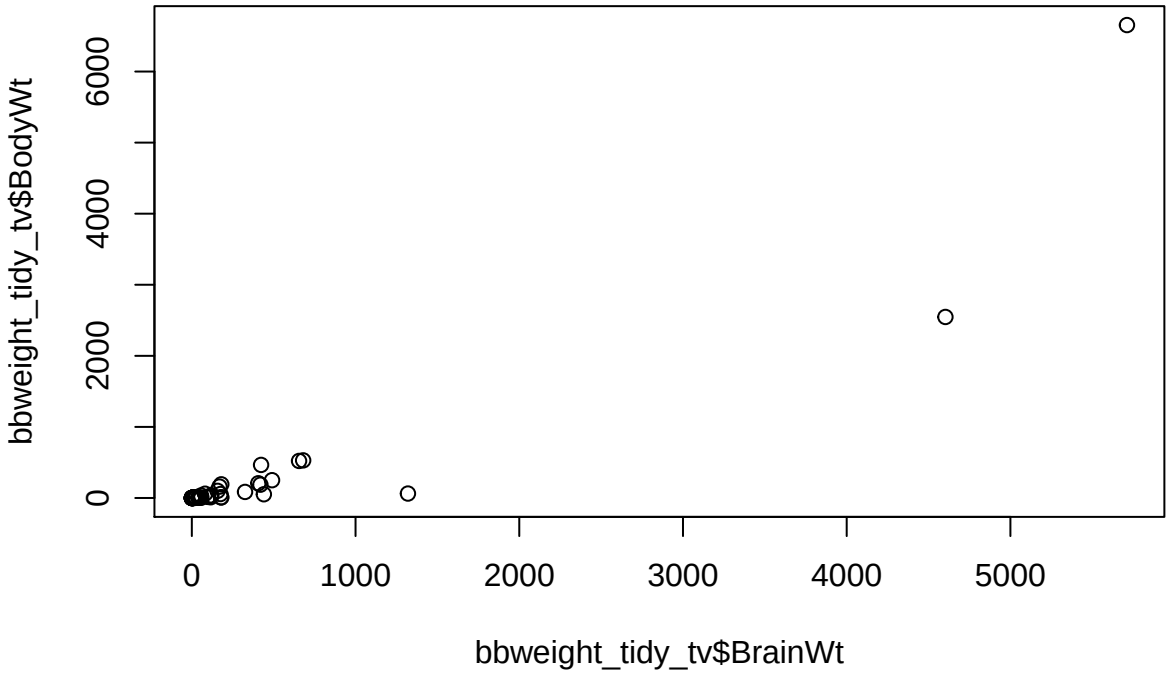
Third, clean the data with tidyverse.

```
#colnames(longj_raw) <- rep(c("Year","LongJump"),4)
BodyWt <- bbweight_raw %>%
                  select(V1,V3,V5) %>%
                  gather("V",value="BodyWt",1,2,3) %>%
                  select(BodyWt) %>%
                  slice(1:62)
BrainWt <- bbweight_raw %>%
                  select(V2,V4,V6) %>%
                  gather("V",value="BrainWt",1,2,3) %>%
                  select(BrainWt) %>%
                  slice(1:62)
bbweight_tidy_tv <- bind_cols(BodyWt, BrainWt)
```

Here is the summary for Brain Body Weight data:

BodyWt	BrainWt
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.: 48.202	3rd Qu.: 166.00
Max. :6654.000	Max. :5712.00

Here is a scatter plot for long jump data:



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

First, import the data.

```
url4 <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"
tomato_raw <- fread(url4,skip = 1)
```

```
## Warning in fread(url4, skip = 1): Detected 3 column names but the data has 4
## columns (i.e. invalid file). Added 1 extra default column name for the first
## column which is guessed to be row names or an index. Use setnames() afterwards
## if this guess is not correct, or fix the file write command that created the
## file to create a valid file.
```

```
saveRDS(tomato_raw, "tomato.RDS")
tomato_raw <- readRDS("tomato.RDS")
```

Second, clean the data with base R.

```
colnames(tomato_raw)<-c("Variety","d10k","d20k","d30k")
d10k<- strsplit(as.character(tomato_raw$d10k),',')
d20k<- strsplit(as.character(tomato_raw$d20k),',')
d30k<- strsplit(as.character(tomato_raw$d30k),',')
Ife<- c(as.numeric(d10k[[1]]), as.numeric(d20k[[1]]),as.numeric(d30k[[1]]))
ped<- c(as.numeric(d10k[[2]]), as.numeric(d20k[[2]]),as.numeric(d30k[[2]]))
tomato_tidy_br <- data.frame(Variety = rep(c("Ife#1","PusaEarlyDwarf"), each = 9),
                             Density = rep(rep(c(10000,20000,30000),each = 3),2),
                             Yield = c(Ife,ped))
```

Third, clean the data with tidyverse.

```
d10k <- tomato_raw %>%
                  separate(d10k, into = c("r1_10k","r2_10k","r3_10k") ,sep = ",") %>%
                  select(Variety,r1_10k,r2_10k,r3_10k) %>%
                  gather(key = "Density", value = "Yield",2:4)
```

```
## Warning: Expected 3 pieces. Additional pieces discarded in 1 rows [2].
```

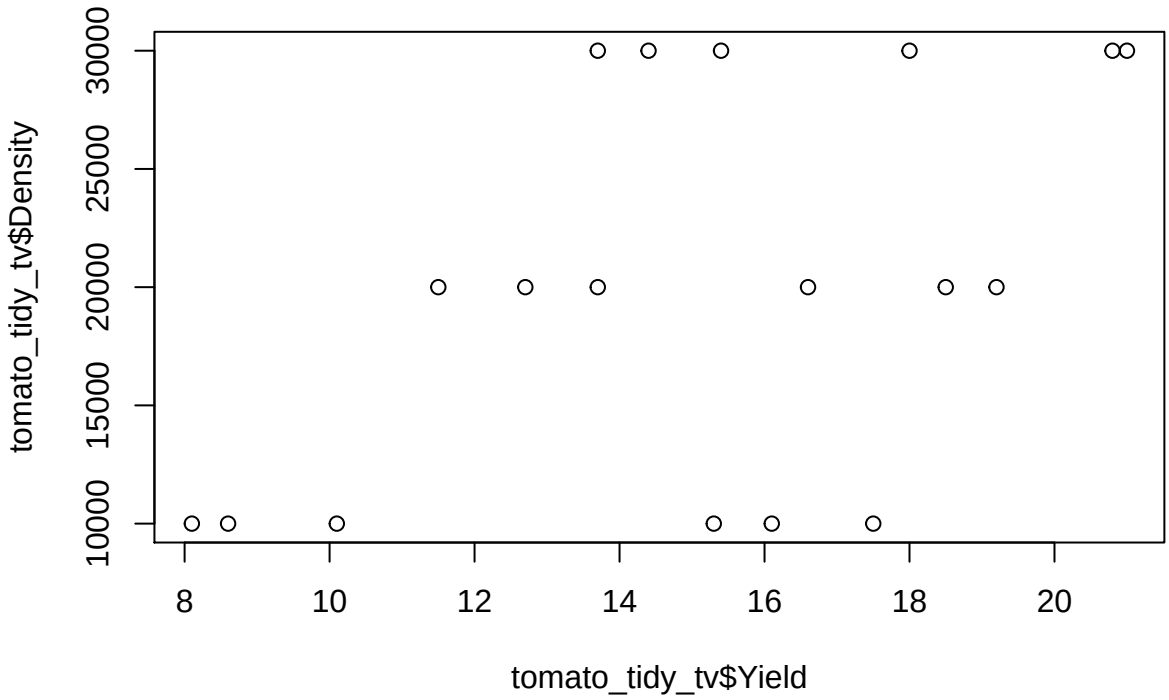
```
d20k <- tomato_raw %>%
                  separate(d20k, into = c("r1_20k","r2_20k","r3_20k") ,sep = ",") %>%
                  select(Variety,r1_20k,r2_20k,r3_20k) %>%
                  gather(key = "Density", value = "Yield",2:4)

d30k <- tomato_raw %>%
                  separate(d30k, into = c("r1_30k","r2_30k","r3_30k") ,sep = ",") %>%
                  select(Variety,r1_30k,r2_30k,r3_30k) %>%
                  gather(key = "Density", value = "Yield",2:4)
tomato_tidy_tv <- bind_rows(d10k,d20k,d30k)
tomato_tidy_tv$Density <- rep(c(10000,20000,30000), each = 6)
tomato_tidy_tv <- tomato_tidy_tv %>% arrange(Variety)
```

Here is the summary for tomato data:

Variety	Density	Yield
Length:18	Min. :10000	Length:18
Class :character	1st Qu.:10000	Class :character
Mode :character	Median :20000	Mode :character
NA	Mean :20000	NA

Variety	Density	Yield
NA	3rd Qu.:30000	NA
NA	Max. :30000	NA

Here is a scatter plot for Yield versus Density:



##

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