

# HW2\_yye1997

Youhui Ye

8/30/2020

## Problem 3

From my point of view, I will definitely use version control on my future's programming. Even though I finish projects on my own, it give me chance to make mistakes and to test a new feature. Needless to say, it allows us to develop different versions when we are cooperating with others.

## Problem 4

a.

First, we need to get the data from the link above:

```
## getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
sensory_data_raw <- read.csv("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat")
saveRDS(sensory_data_raw, "sensory_data_raw.RDS")
sensory_data_raw <- readRDS("sensory_data_raw.RDS")
```

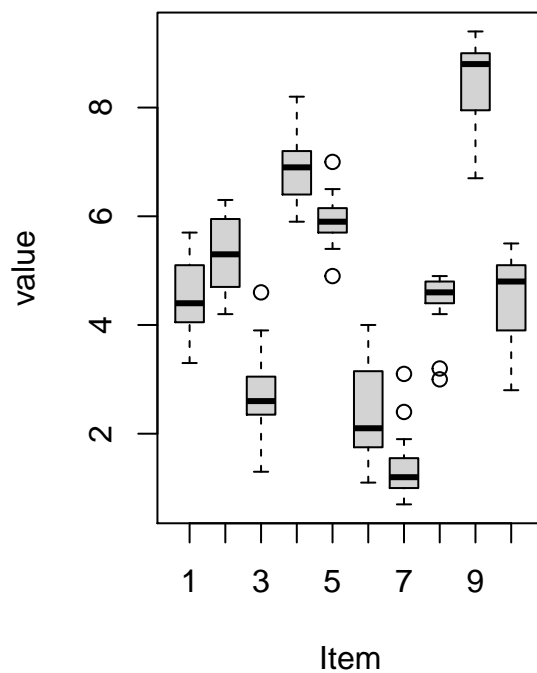
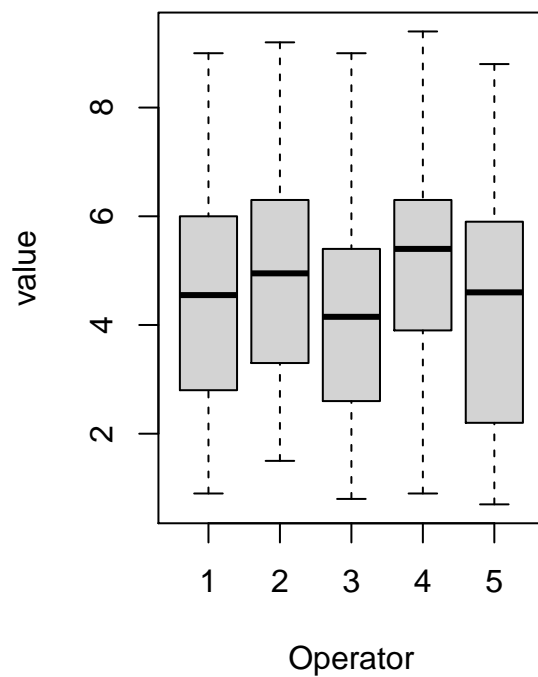
Need to tidy the data, basic issues are one cell has multiple values and operator, item are not columns.

```
## split every single value in one cell
df <- data.frame(matrix(ncol = 6, nrow = 31))
for (i in 1:6) {
  df[,i] <- sapply(strsplit(as.character(sensory_data_raw$Operator), " "), `[,`, i)
}
## modify column names
colnames(df) <- df[1,]
df <- df[-1,]
## insert item indices
for (i in 1:30) {
  if (is.na(df[i,6])){
    df[i,1:6] <- c(df[i-1,1],df[i,1:5])
  }
}
sensory_data_br <- data.frame(as.factor(rep(rep(1:10, each = 3), 5)), stack(df[,2:6]))

colnames(sensory_data_br) <- c("Item", "value", "Operator")
## convert value to numeric type
sensory_data_br$value <- as.numeric(sensory_data_br$value)
```

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

Item	value	Operator
1 :15	Min. :0.700	1:30
2 :15	1st Qu.:3.025	2:30
3 :15	Median :4.700	3:30
4 :15	Mean :4.657	4:30
5 :15	3rd Qu.:6.000	5:30
6 :15	Max. :9.400	NA
(Other):60	NA	NA



Now, we use tidyverse to tidy data frames again.

```
## using tidyverse
sensory_data_tv <- sensory_data_raw %>%
  separate(col = "Operator", into = c("Item", as.character(1:5)), sep = " ", fill = "left") %>%
  fill("Item") %>%
  slice(2:n()) %>%
  gather(key = "Operator", value = "value", -Item)
```

b.

First, we need to get the data from the link above:

```
## getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
# LongJump_data_raw <- fread("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat")
# names(LongJump_data_raw) <- make.unique(names(LongJump_data_raw))
```

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

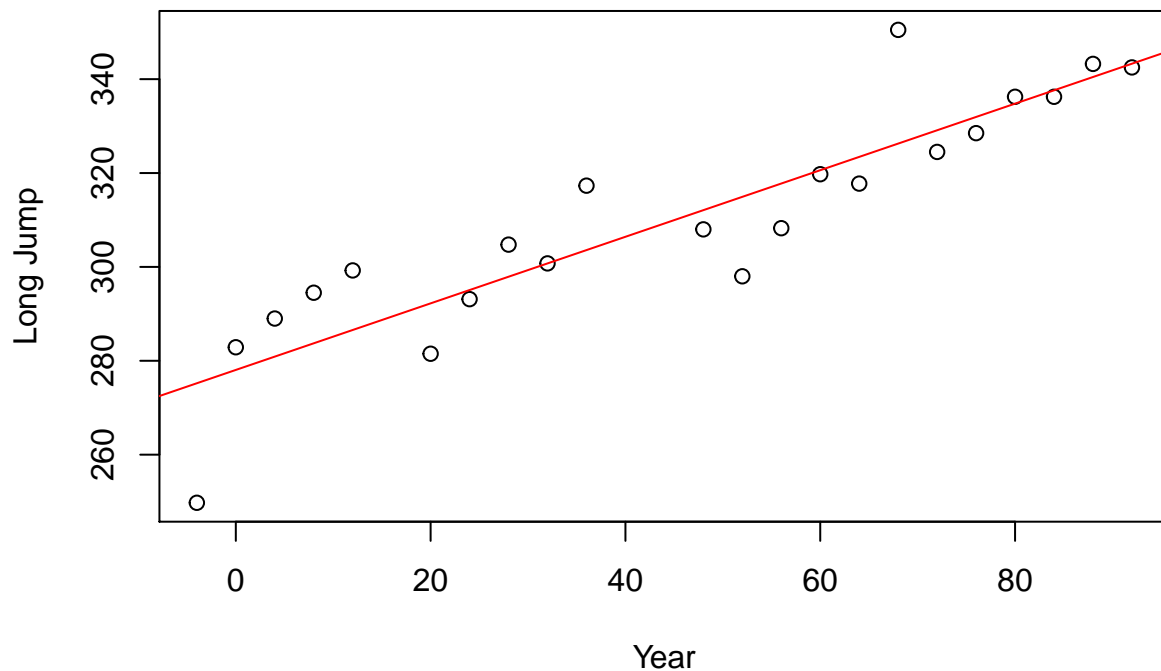
Need to tidy the data, basic issues are “LongJump” was regarded as two separated words and 2 variables were split into several parts.

```
## stack data and fix columns
LongJump_data_tidy_br <- data.frame(stack(LongJump_data_raw[,c(1,3,5,7)]),
                                   stack(LongJump_data_raw[,c(2,4,6,8)]))
LongJump_data_tidy_br <- LongJump_data_tidy_br[-c(23,24),c(1,3)]
colnames(LongJump_data_tidy_br) <- c("Year", "LongJump")
```

We have converted the data frames to tidy data frames using the base functions. Here is a summary 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

Also, the scatter plot and the fitted line show a positive relationship between 2 variables.



Now, we use tidyverse to clean and tidy data again.

```
## stack and fix column names using tidyverse
## making new names for the data set
colnames(LongJump_data_raw) <- paste0(c("Year", "LongJump"), rep(1:6, each=2))
LongJump_data_tv <- LongJump_data_raw %>%
  melt(measure=paste0("Year", "LongJump"),
       value.name=c("Year", "LongJump"), na.rm = TRUE) %>%
  select(-variable)
```

C.

First, we need to get the data from the link above:

```
## getting "https://https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
# bbw_data_raw <- fread("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat")
# saveRDS(bbw_data_raw, "bbw_data_raw.RDS")
bbw_data_raw <- readRDS("bbw_data_raw.RDS")
```

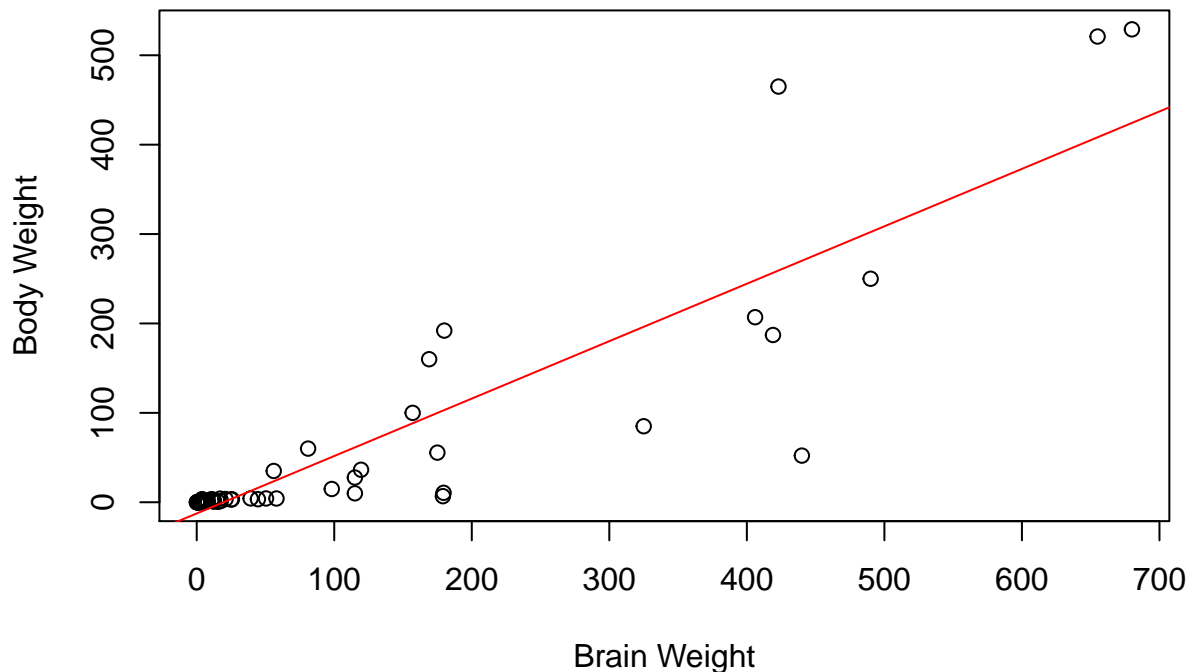
Need to tidy the data, basic issues are the same as the last one.

```
bbw_data_tidy_br <- data.frame(stack(bbw_data_raw[, c(1, 3, 5)]),
                              stack(bbw_data_raw[, c(2, 4, 6)]))
bbw_data_tidy_br <- bbw_data_tidy_br[complete.cases(bbw_data_tidy_br), c(1, 3)]
colnames(bbw_data_tidy_br) <- c("BodyWt", "BrainWt")
```

Here is a summary table of body and brain weight.

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

Also, the scatter plot and the fitted line show a positive relationship between 2 variables. There seem to be 2 outliers,



Now, we use tidyverse package to tidy this data set again.

```
colnames(bbw_data_raw) <- paste0(c("BodyWt", "BrainWt"), rep(1:6, each=2))
bbw_data_tv <- bbw_data_raw %>%
  melt( measure=patterns("^BodyWt", "^BrainWt"),
        value.name=c("BodyWt", "BrainWt"), na.rm = TRUE) %>%
  select(-variable)
```

d.

First, we need to get the data from the link above:

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

Need to tidy the data, basic issues are densities are columns and one cell contains multiple values.

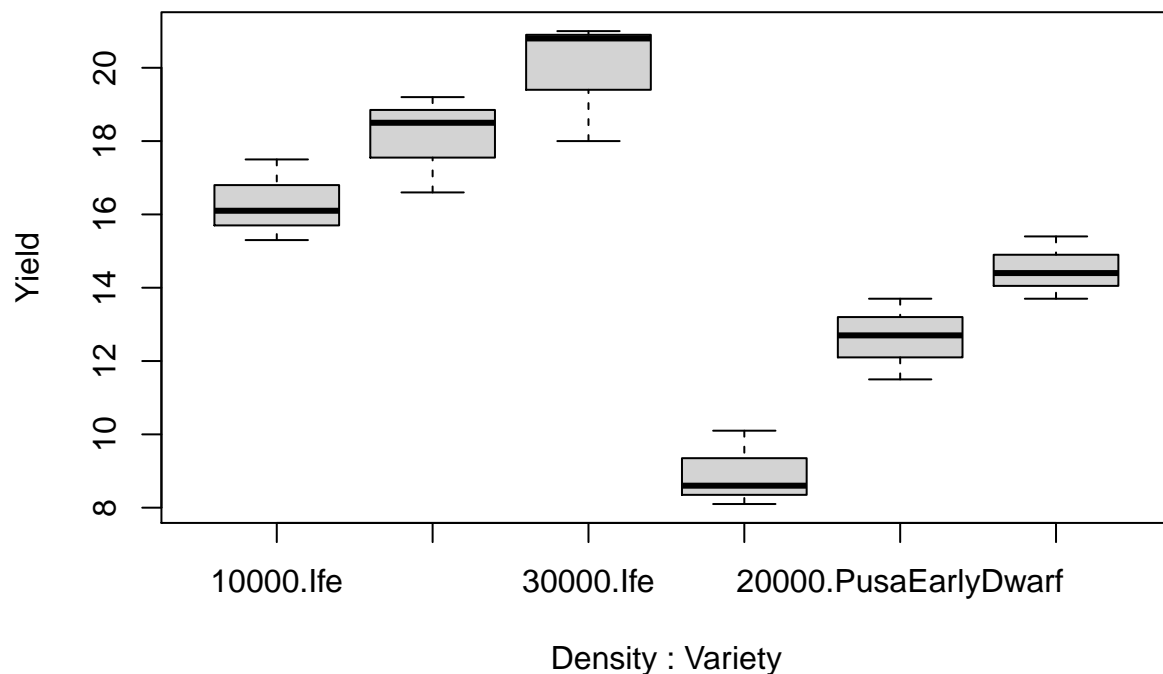
```
## creating new variables
Density <- rep(c("10000", "20000", "30000"), each = 3, times = 2)
Variety <- rep(c("Ife", "PusaEarlyDwarf"), each = 9)
Yield <- rep(0, 18)
for (i in 1:2) {
  for (j in 2:4) {
    Yield[((i-1)*9 + (j-2)*3 + 1): ((i-1)*9 + (j-2)*3 + 3)] <- +
      as.numeric(unlist(strsplit(as.character(ty_data_raw[i,j,with=FALSE]), ',')))
  }
}
```

```
}
ty_data_br <- data.frame(Variety, Density, Yield)
```

Here is a summary table of tomato yield.

Variety	Density	Yield
Length:18	Length:18	Min. : 8.10
Class :character	Class :character	1st Qu.:12.95
Mode :character	Mode :character	Median :15.35
NA	NA	Mean :15.07
NA	NA	3rd Qu.:17.88
NA	NA	Max. :21.00

Also, the boxplot by density and variety shows apparent trends.



Now, we use tidyverse package to tidy this data set again.

```
ty_data_tv <- ty_data_raw %>%
  separate_rows("10000") %>%
  separate_rows("20000", "30000") %>%
  gather(key = "Density", value = "Yield", "10000":"30000") %>%
  distinct() %>%
  na_if("") %>%
  drop_na()
```