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:

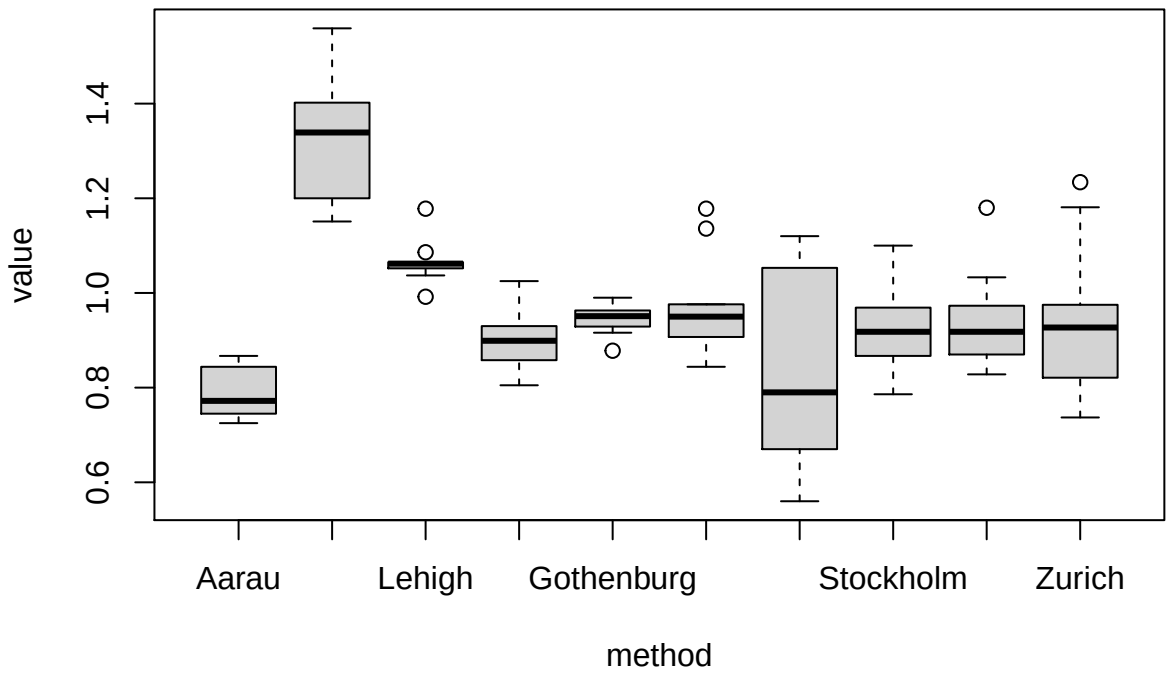
```
## getting "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/fullgirder.dat"
# girder_data_raw <- fread("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/fullgirder.dat")
# saveRDS(girder_data_raw, "girder_data_raw.RDS")
girder_data_raw <- readRDS("girder_data_raw.RDS")
```

Need to tidy the data, basic issue is sites are columns, need to push them into a column.

```
## stack data and fix columns
girder_data_tidy_br <- data.frame(girder=rep(girder_data_raw$Girder,2),
                                  stack(girder_data_raw[,-1]))
colnames(girder_data_tidy_br) <- c("girder", "value", "method")
```

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

girder	value	method
Length:90	Min. :0.5600	Aarau : 9
Class :character	1st Qu.:0.8582	Karlsruhe : 9
Mode :character	Median :0.9310	Lehigh : 9
NA	Mean :0.9685	Cardiff : 9
NA	3rd Qu.:1.0617	Gothenburg: 9
NA	Max. :1.5590	Osaka : 9
NA	NA	(Other) :36



Now, we use tidyverse to tidy data frames again.

```
## stack and fix column names using tidyverse
girder_data_tidy_tv <- girder_data_raw %>% gather(key="method", value="value", Aarau:Zurich)
```

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")
```

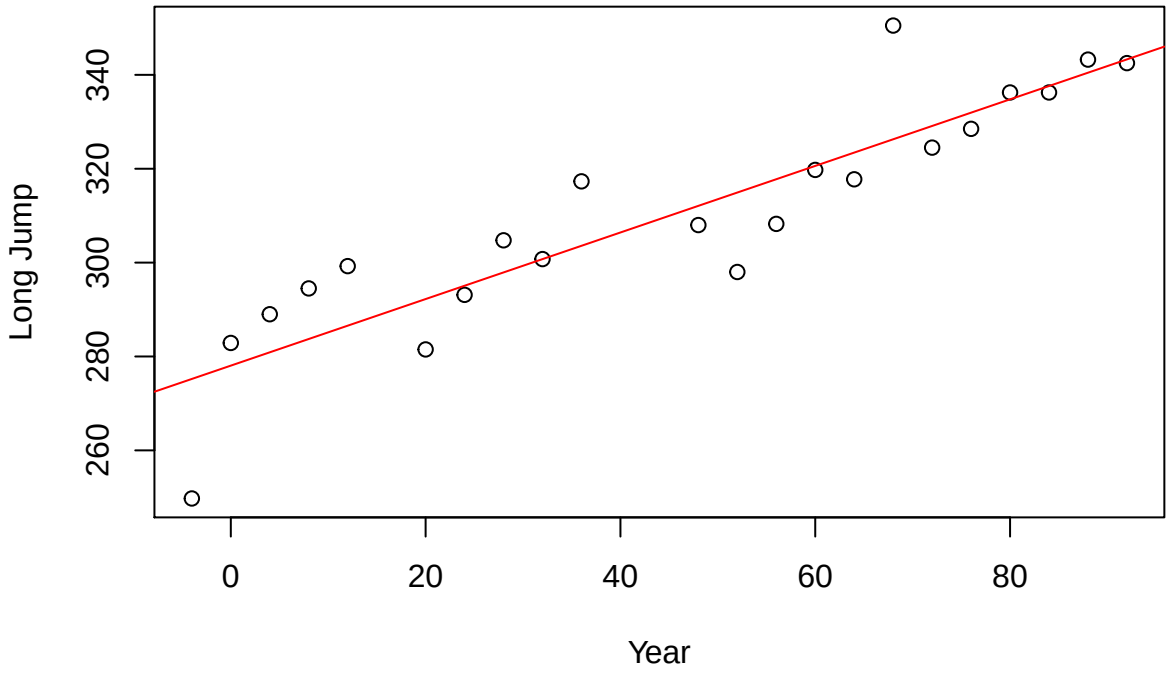
Need to tidy the data, basic issues are "LongJump" was regarded as two separated words and 2 variables were split into several parts.

```
## 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 relationshup 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=patterns("^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")
```

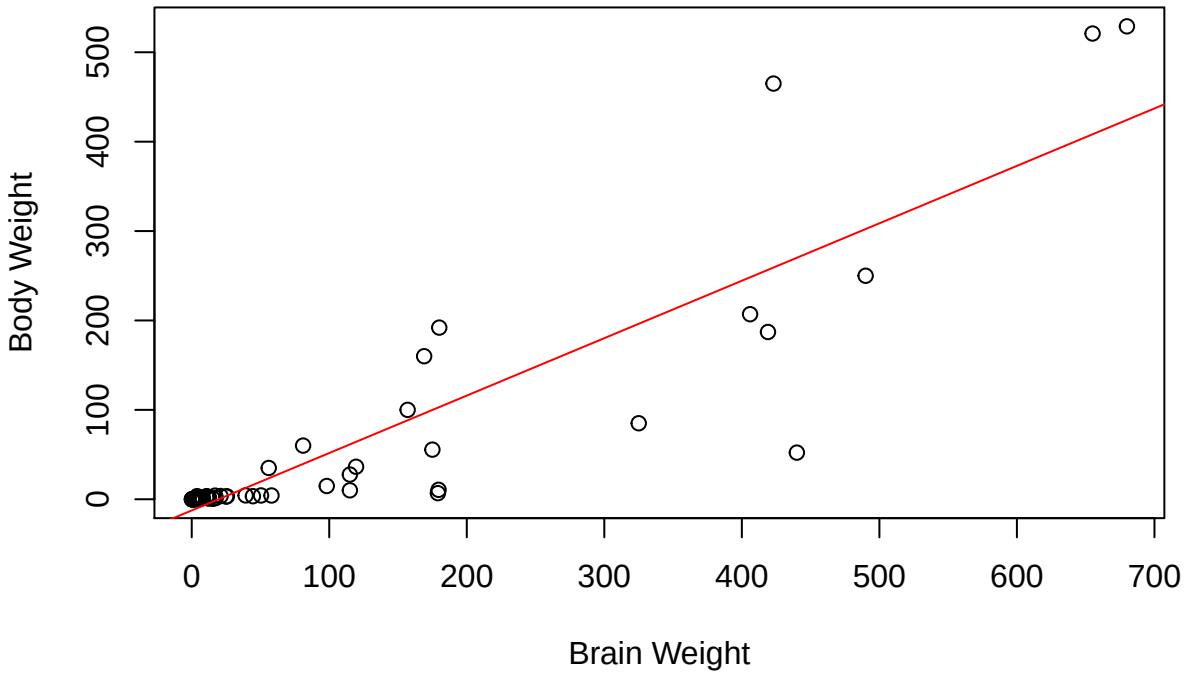
Need to tidy the data, basic issues are the same as the last one.

```
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.

```r
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:

```r
## 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.

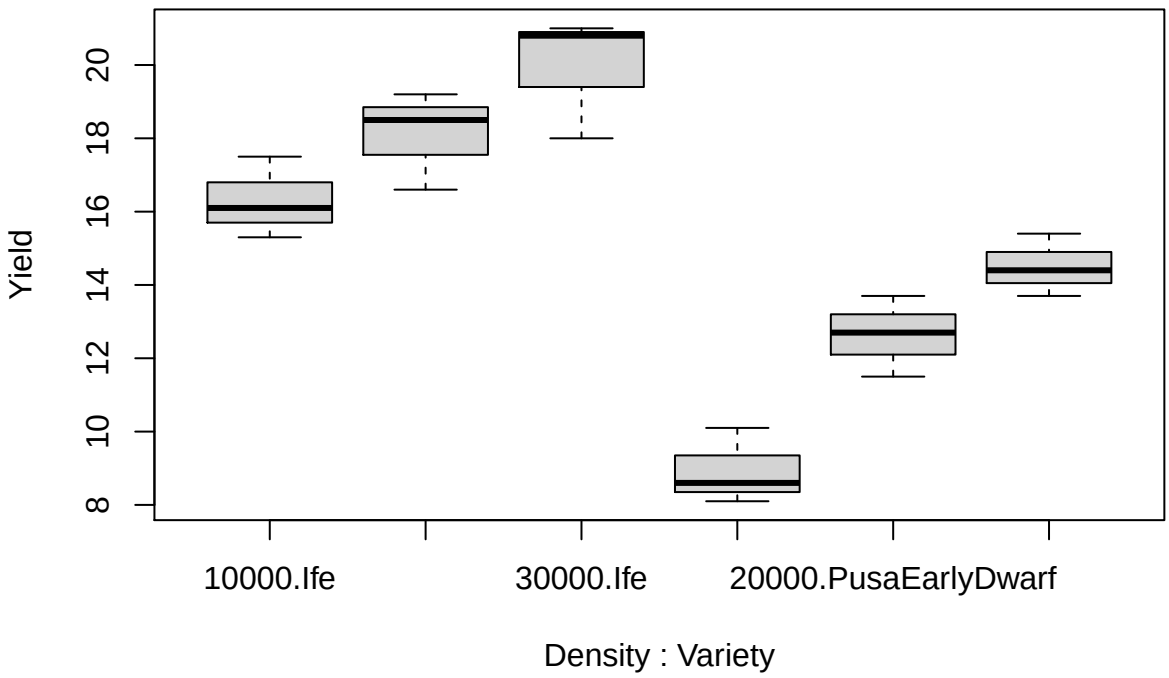
```r
## creating new variables
Density <- rep(c("10000","20000","30000"), each = 3,times = 2)
Variety <- rep(c("Ife", "PusaEarlyDwarf"), each = 9)
Yield <- rep(0,18)
for (i in 1:2) {
  for (j in 2:4) {
    Yield[((i-1)*9 + (j-2)*3 + 1): ((i-1)*9 + (j-2)*3 + 3)] <- +
      as.numeric(unlist(strsplit(as.character(ty_data_raw[i,j,with=FALSE]),',')))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

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()
```