Homework4

Anbin Rhee

10/20/2021

Problem 1

If I need a bit more work with the tidyverse, I will check out Rstudio.cloud Primer titled “Tidy your data”.

Problem 2

I used R Markdown file for this homework.

Problem 3

Part A

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(reshape)
```

```
##
## Attaching package: 'reshape'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

```
## The following object is masked from 'package:dplyr':
##
##     rename
```

```
PartA <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/ThicknessGauge.dat",header =
colnames(PartA) <- c("part","1.1","1.2","2.1","2.2","3.1","3.2")
PartA <- melt(PartA, id.vars = "part")
PartA <- separate(data = PartA, col = 'variable',into = c("operator", "measurement"))
```

I imported the data and renamed the columns as “1.1”, “1.2”, “2.1”, “2.2”, “3.1”, and “3.2”. Then, I rearranged the data in order to the observations to be distinguished and separated the observations by operator and measurement.

```
PartA$part <- factor(PartA$part)
PartA$operator <- factor(PartA$operator)
PartA$measurement <- factor(PartA$measurement)
summary(PartA)
```

```
##       part    operator measurement     value
##  1      : 6   1:20     1:30        Min.   :0.9520
##  2      : 6   2:20     2:30        1st Qu.:0.9550
##  3      : 6   3:20                 Median :0.9570
##  4      : 6                        Mean   :0.9561
##  5      : 6                        3rd Qu.:0.9570
##  6      : 6                        Max.   :0.9580
##  (Other):24
```

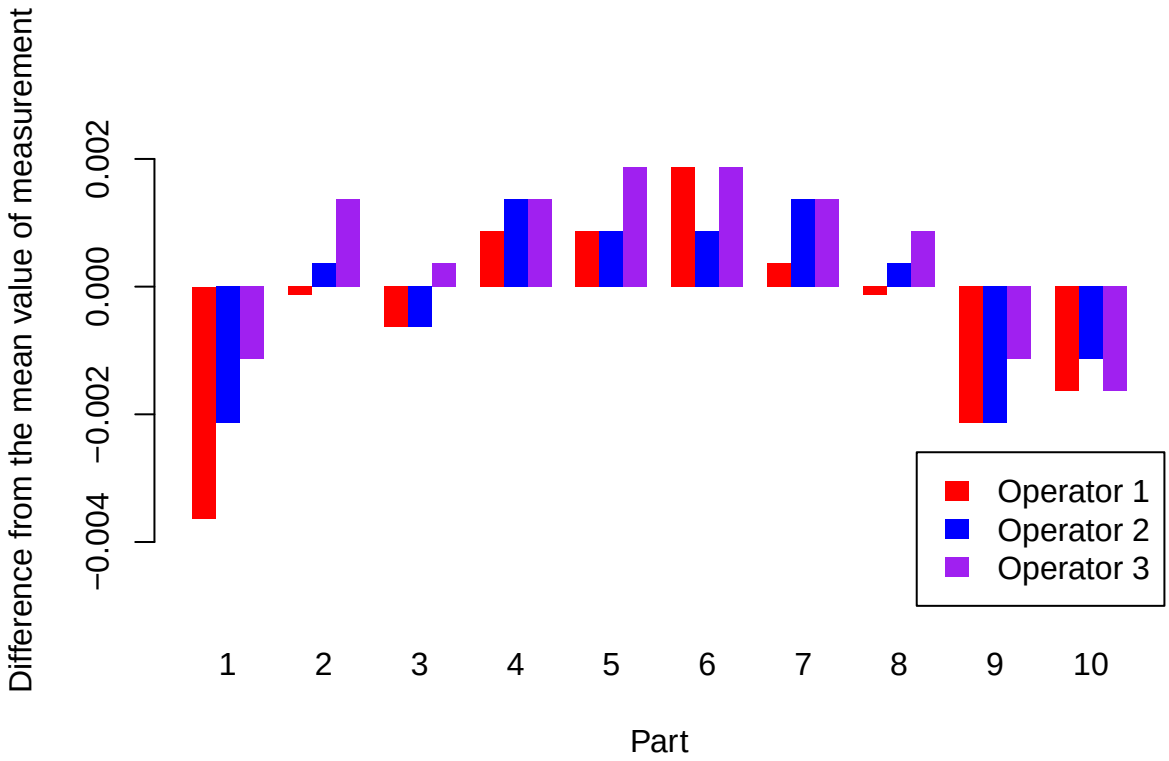
```
knitr::kable(PartA, caption = "Measurements of the part's wall thickness")
```

Table 1: Measurements of the part's wall thickness

part	operator	measurement	value
1	1	1	0.953
2	1	1	0.956
3	1	1	0.956
4	1	1	0.957
5	1	1	0.957
6	1	1	0.958
7	1	1	0.957
8	1	1	0.957
9	1	1	0.954
10	1	1	0.954
1	1	2	0.952
2	1	2	0.956
3	1	2	0.955
4	1	2	0.957
5	1	2	0.957
6	1	2	0.958
7	1	2	0.956
8	1	2	0.955
9	1	2	0.954
10	1	2	0.955
1	2	1	0.954
2	2	1	0.956
3	2	1	0.956
4	2	1	0.958
5	2	1	0.957
6	2	1	0.957
7	2	1	0.958
8	2	1	0.957
9	2	1	0.954
10	2	1	0.956
1	2	2	0.954
2	2	2	0.957

part	operator	measurement	value
3	2	2	0.955
4	2	2	0.957
5	2	2	0.957
6	2	2	0.957
7	2	2	0.957
8	2	2	0.956
9	2	2	0.954
10	2	2	0.954
1	3	1	0.954
2	3	1	0.958
3	3	1	0.957
4	3	1	0.957
5	3	1	0.958
6	3	1	0.958
7	3	1	0.958
8	3	1	0.957
9	3	1	0.955
10	3	1	0.954
1	3	2	0.956
2	3	2	0.957
3	3	2	0.956
4	3	2	0.958
5	3	2	0.958
6	3	2	0.958
7	3	2	0.957
8	3	2	0.957
9	3	2	0.955
10	3	2	0.955

```
PartAplot <- aggregate(x = PartA$value, by = list(PartA$part, PartA$operator), FUN = mean)
PartAplot['difference'] <- PartAplot$x - mean(PartAplot$x)
par(xpd=TRUE)
barplot(difference ~ Group.2 + Group.1,
        data = PartAplot,
        beside = TRUE,
        xlab = "Part",
        ylab = "Difference from the mean value of measurement",
        col = c("red", "blue", "purple"),
        ylim = c(-0.005,0.003),
        border = NA)
legend("bottomright", c("Operator 1", "Operator 2","Operator 3"),
       fill = c("red", "blue", "purple"), border = NA)
```



I showed the data as table form and drew a plot which shows the difference from the mean value of measurements.

Part B

```
PartB <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat", hea

# rename the columns
colnames(PartB) <- rep(c("BodyWeight", "BrainWeight"), 3)
# rearrange data frame to 2 columns
PartB <- rbind(PartB[,1:2], PartB[,3:4], PartB[1:20,5:6])
```

I imported the data of Part B and renamed the columns as “Body Weight” and “Brain Weight”. Then, I rearranged the data.

```
summary(PartB)
```

```
##    BodyWeight        BrainWeight
##  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
```

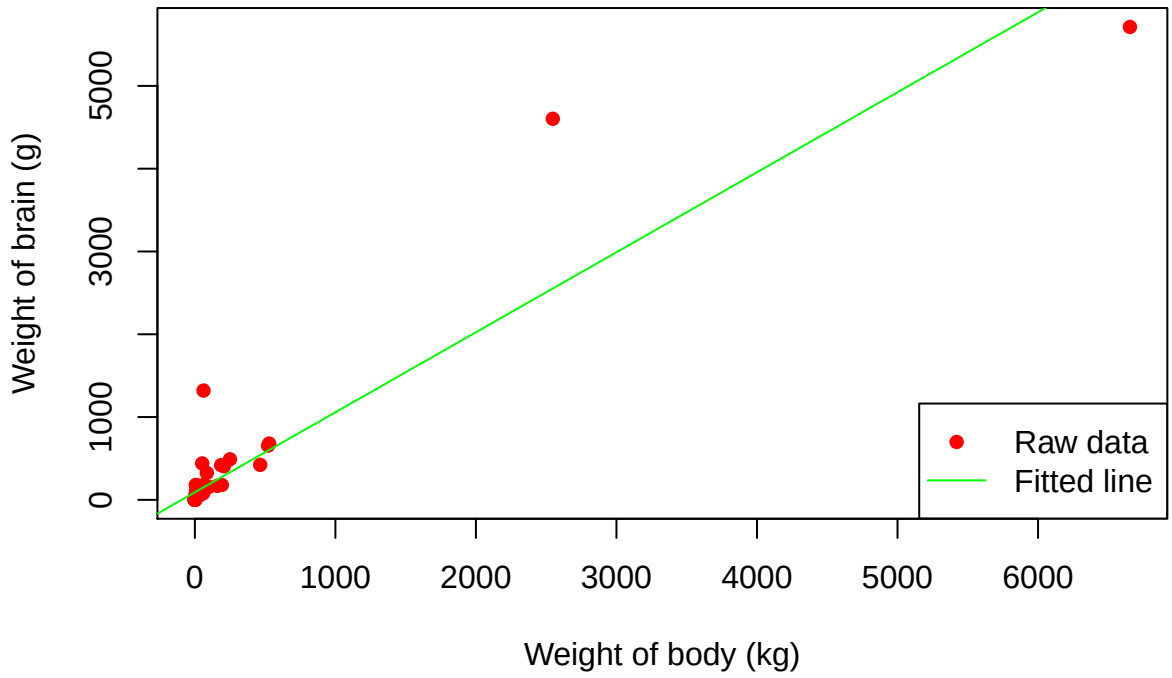
```
knitr::kable(PartB, caption = "Body and brain weight")
```

Table 2: Body and brain weight

BodyWeight	BrainWeight
3.385	44.50
0.480	15.50
1.350	8.10
465.000	423.00
36.330	119.50
27.660	115.00
14.830	98.20
1.040	5.50
4.190	58.00
0.425	6.40
0.101	4.00
0.920	5.70
1.000	6.60
0.005	0.10
0.060	1.00
3.500	10.80
2.000	12.30
1.700	6.30
2547.000	4603.00
0.023	0.30
187.100	419.00
521.000	655.00
0.785	3.50
10.000	115.00
3.300	25.60
0.200	5.00
1.410	17.50
529.000	680.00
207.000	406.00
85.000	325.00
0.750	12.30
62.000	1320.00
6654.000	5712.00
3.500	3.90
6.800	179.00
35.000	56.00
4.050	17.00
0.120	1.00
0.023	0.40
0.010	0.30
1.400	12.50
250.000	490.00
2.500	12.10
55.500	175.00
100.000	157.00
52.160	440.00
10.550	179.50
0.550	2.40
60.000	81.00
3.600	21.00

BodyWeight	BrainWeight
4.288	39.20
0.280	1.90
0.075	1.20
0.122	3.00
0.048	0.33
192.000	180.00
3.000	25.00
160.000	169.00
0.900	2.60
1.620	11.40
0.104	2.50
4.235	50.40

```
# scatter plot and fitted simple linear model
plot(x = PartB$BodyWeight, y = PartB$BrainWeight,
     col = "red", pch = 16,
     xlab = 'Weight of body (kg)',
     ylab = 'Weight of brain (g)')
abline(lm(BrainWeight ~ BodyWeight, PartB),
       col = "green")
legend(x = "bottomright", legend = c("Raw data", "Fitted line"),
       col = c("red","green"), lty = c(0,1), pch = c(16,NA))
```



The summary of the data is presented in the table and the scatter plot between Body Weight and Brain Weight is also presented.

Part C

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following object is masked from 'package:reshape':
##
##     melt
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
PartC <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat",header = 
colnames(PartC) <- rep(c("year", "long jump"), 4)
PartC <- rbind(PartC[,1:2], PartC[,3:4], PartC[,5:6], PartC[1:4, 7:8])
```

I imported data of Part C and renamed the columns as "year" and "long jump".

```
PartC$year <- PartC$year + 1900
# show the table of data (first 6 observations)
knitr::kable(PartC, caption = "Gold Medal performance for Olympic Men's Long Jump")
```

Table 3: Gold Medal performance for Olympic Men's Long Jump

year	long jump
1896	249.75
1900	282.88
1904	289.00
1908	294.50
1912	299.25
1920	281.50
1924	293.13
1928	304.75
1932	300.75
1936	317.31
1948	308.00
1952	298.00
1956	308.25
1960	319.75
1964	317.75
1968	350.50
1972	324.50
1976	328.50
1980	336.25
1984	336.25
1988	343.25
1992	342.50

```
summary(PartC)
```

```
##       year        long jump
##  Min.   :1896   Min.   :249.8
##  1st Qu.:1921   1st Qu.:295.4
##  Median :1950   Median :308.1
##  Mean   :1945   Mean   :310.3
##  3rd Qu.:1971   3rd Qu.:327.5
```

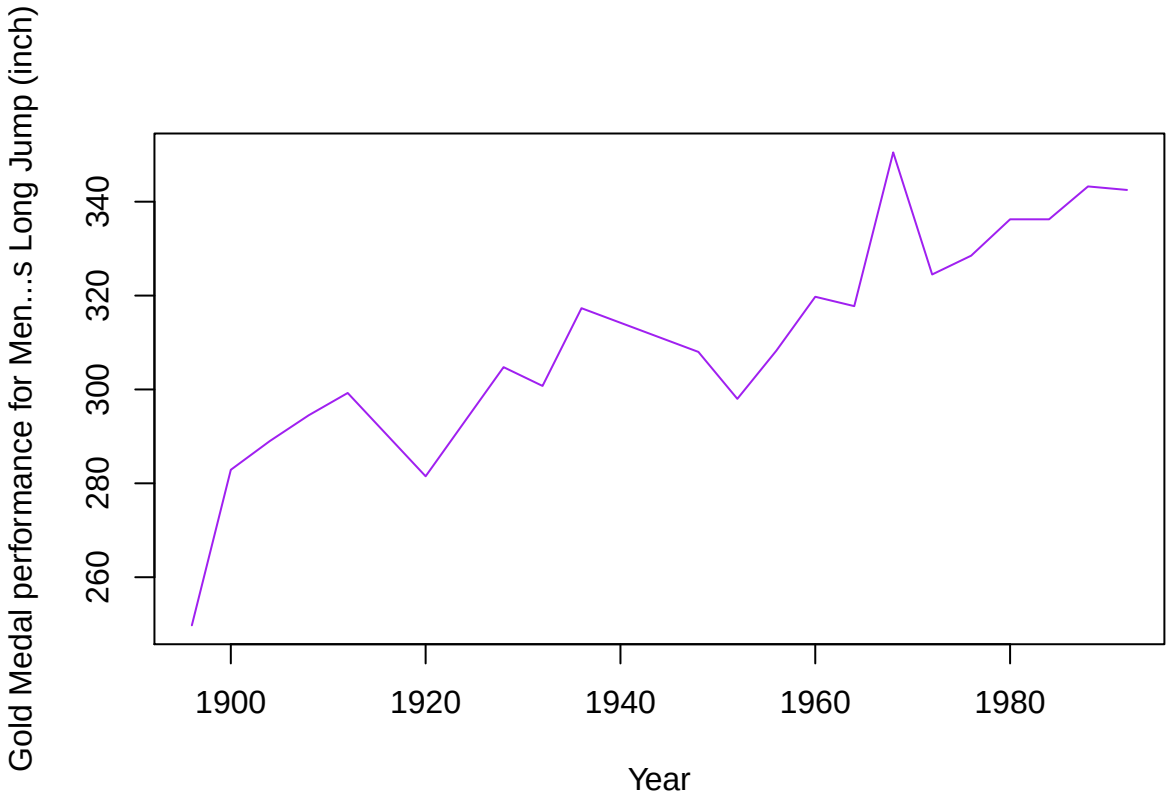
```
##  Max.   :1992   Max.   :350.5
```

```
# scatter plot and fitted simple linear model
plot(PartC, col = "purple", lwd = 1,
     type = 'l',
     xlab = 'Year',
     ylab = 'Gold Medal performance for Men's Long Jump (inch)')
```

```
## Warning in title(...): conversion failure on 'Gold Medal performance for Men's
## Long Jump (inch)' in 'mbcsToSbcs': dot substituted for <e2>
```

```
## Warning in title(...): conversion failure on 'Gold Medal performance for Men's
## Long Jump (inch)' in 'mbcsToSbcs': dot substituted for <80>
```

```
## Warning in title(...): conversion failure on 'Gold Medal performance for Men's
## Long Jump (inch)' in 'mbcsToSbcs': dot substituted for <99>
```



The summary of the data of Part C is presented in table above. I also drew a scatter plot between year and gold medal performance for men's long jump.

Part D

```
PartD <- fread("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat",header = FALSE, skip
colnames(PartD) <- c("category", "10000", "20000", "30000")
PartD <- separate(data = PartD, col = '10000',into = c("10000.1", "10000.2", "10000.3"),
                  remove = TRUE, sep = ',')
```

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

```
PartD <- separate(data = PartD, col = '20000',into = c("20000.1", "20000.2", "20000.3"),
                  remove = TRUE, sep = ',')
```

```
PartD <- separate(data = PartD, col = '30000',into = c("30000.1", "30000.2", "30000.3"),
                remove = TRUE, sep = ',')
```

I imported the data in Part D and renamed the columns as "category", "10000", "20000", and "30000". Then, separated the columns by three.

```
PartD <- melt(PartD, id.vars = "category")
# separate columns to Planting Density and measurement
PartD <- separate(data = PartD, col = 'variable',
                into = c("PlantingDensity", "measurement"),remove = TRUE)
PartD$category <- factor(PartD$category)
PartD$PlantingDensity <- factor(PartD$PlantingDensity)
PartD$measurement <- factor(PartD$measurement)
PartD$value <- as.numeric(PartD$value)
knitr::kable(head(PartD), caption = "Measurements of tomato yield")
```

Table 4: Measurements of tomato yield

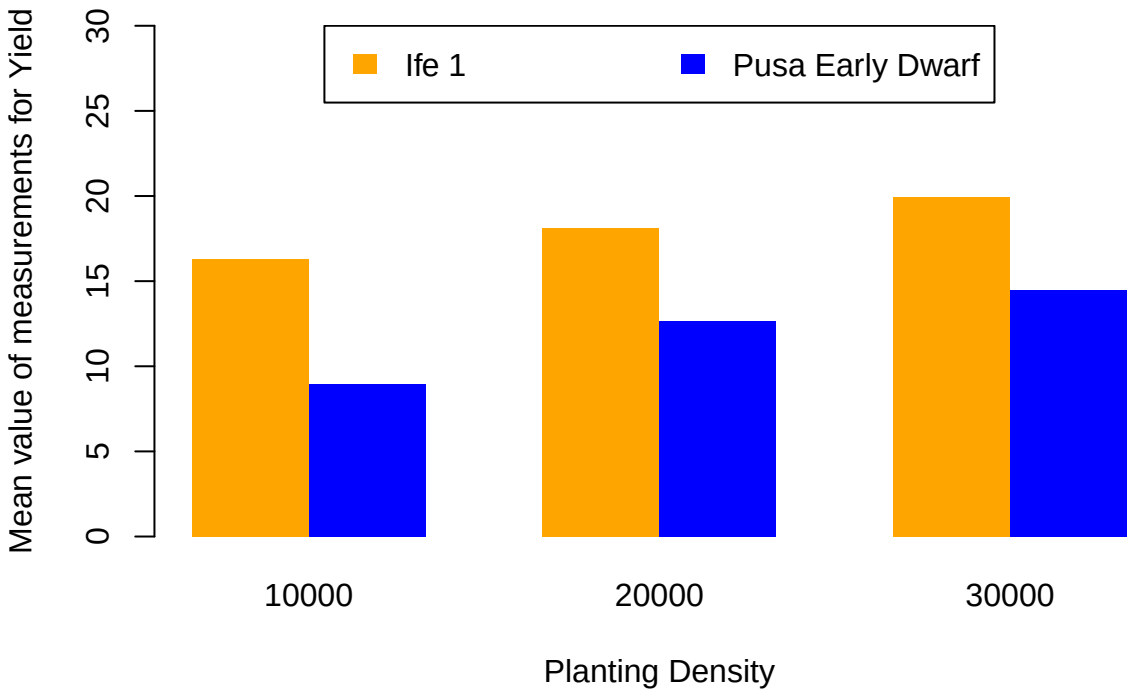
category	PlantingDensity	measurement	value
Ife#1	10000	1	16.1
PusaEarlyDwarf	10000	1	8.1
Ife#1	10000	2	15.3
PusaEarlyDwarf	10000	2	8.6
Ife#1	10000	3	17.5
PusaEarlyDwarf	10000	3	10.1

```
knitr::kable(summary(PartD), caption="Summary of variables")
```

Table 5: Summary of variables

category	PlantingDensity	measurement	value
Ife#1 :9	10000:6	1:6	Min. : 8.10
PusaEarlyDwarf:9	20000:6	2:6	1st Qu.:12.95
NA	30000:6	3:6	Median :15.35
NA	NA	NA	Mean :15.07
NA	NA	NA	3rd Qu.:17.88
NA	NA	NA	Max. :21.00

```
PartDplot <- aggregate(x = PartD$value, by = list(PartD$category, PartD$PlantingDensity), FUN = mean)
# plot the yield by category and Planting Density
barplot(x ~ Group.1 + Group.2,
        data = PartDplot,
        beside = TRUE,
        col = c("orange", "blue"),
        xlab = "Planting Density",
        ylab = "Mean value of measurements for Yield",
        ylim = c(0,30),
        border = NA)
legend("top", c("Ife 1", "Pusa Early Dwarf"),
       fill = c("orange", "blue"),
       border = NA, horiz = TRUE)
```



I showed the summary of the data by table and drew a plot which shows the mean value of measurements for tomato yield.

Part E

```
library(ggplot2)
PartE <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LarvaeControl.dat",header =
PartE <- PartE[,colSums(is.na(PartE))<nrow(PartE)]
colnames(PartE) <- c("Block","Age1.Treatment1","Age1.Treatment2","Age1.Treatment3","Age1.Treatment4","Ag
PartE <- melt(as.data.table(PartE), id.vars = "Block")
PartE <- separate(data = PartE, col = 'variable', into = c("age", "treatment"), remove = TRUE)
PartE$Block <- factor(PartE$Block)
PartE$age <- factor(PartE$age)
PartE$treatment <- factor(PartE$treatment)
knitr::kable(PartE, caption = "Larvae counts at two ages")
```

Table 6: Larvae counts at two ages

Block	age	treatment	value
1	Age1	Treatment1	13
2	Age1	Treatment1	29
3	Age1	Treatment1	5
4	Age1	Treatment1	5
5	Age1	Treatment1	0
6	Age1	Treatment1	1
7	Age1	Treatment1	1
8	Age1	Treatment1	4
1	Age1	Treatment2	16
2	Age1	Treatment2	12
3	Age1	Treatment2	4
4	Age1	Treatment2	12

Block	age	treatment	value
5	Age1	Treatment2	2
6	Age1	Treatment2	1
7	Age1	Treatment2	3
8	Age1	Treatment2	4
1	Age1	Treatment3	13
2	Age1	Treatment3	23
3	Age1	Treatment3	4
4	Age1	Treatment3	1
5	Age1	Treatment3	2
6	Age1	Treatment3	1
7	Age1	Treatment3	1
8	Age1	Treatment3	7
1	Age1	Treatment4	20
2	Age1	Treatment4	15
3	Age1	Treatment4	1
4	Age1	Treatment4	5
5	Age1	Treatment4	2
6	Age1	Treatment4	3
7	Age1	Treatment4	0
8	Age1	Treatment4	3
1	Age1	Treatment5	16
2	Age1	Treatment5	17
3	Age1	Treatment5	2
4	Age1	Treatment5	3
5	Age1	Treatment5	0
6	Age1	Treatment5	5
7	Age1	Treatment5	1
8	Age1	Treatment5	1
1	Age2	Treatment1	28
2	Age2	Treatment1	61
3	Age2	Treatment1	7
4	Age2	Treatment1	14
5	Age2	Treatment1	3
6	Age2	Treatment1	7
7	Age2	Treatment1	10
8	Age2	Treatment1	13
1	Age2	Treatment2	12
2	Age2	Treatment2	49
3	Age2	Treatment2	2
4	Age2	Treatment2	5
5	Age2	Treatment2	3
6	Age2	Treatment2	6
7	Age2	Treatment2	5
8	Age2	Treatment2	11
1	Age2	Treatment3	40
2	Age2	Treatment3	48
3	Age2	Treatment3	4
4	Age2	Treatment3	14
5	Age2	Treatment3	2
6	Age2	Treatment3	7
7	Age2	Treatment3	8
8	Age2	Treatment3	10

Block	age	treatment	value
1	Age2	Treatment4	31
2	Age2	Treatment4	44
3	Age2	Treatment4	5
4	Age2	Treatment4	9
5	Age2	Treatment4	7
6	Age2	Treatment4	7
7	Age2	Treatment4	3
8	Age2	Treatment4	12
1	Age2	Treatment5	22
2	Age2	Treatment5	45
3	Age2	Treatment5	2
4	Age2	Treatment5	8
5	Age2	Treatment5	0
6	Age2	Treatment5	4
7	Age2	Treatment5	6
8	Age2	Treatment5	8

```
knitr::kable(summary(PartE), caption="Summary of variables")
```

Table 7: Summary of variables

Block	age	treatment	value
1 :10	Age1:40	Treatment1:16	Min. : 0.00
2 :10	Age2:40	Treatment2:16	1st Qu.: 2.75
3 :10	NA	Treatment3:16	Median : 5.50
4 :10	NA	Treatment4:16	Mean :10.50
5 :10	NA	Treatment5:16	3rd Qu.:13.00
6 :10	NA	NA	Max. :61.00
(Other):20	NA	NA	NA

```
ggplot(PartE, aes(y = value, x = Block,
                color = treatment,
                shape = age))+
  geom_point(size = 4)+
  ylab("Larvae counts")
```

I imported the data of Part E and renamed the columns as "Block", "Age1.Treatment1", "Age1.Treatment2", "Age1.Treatment3", "Age1.Treatment4", "Age1.Treatment5", "Age2.Treatment1", "Age2.Treatment2", "Age2.Treatment3", "Age2.Treatment4", "Age2.Treatment5". Also the summaries of the data are presented in tables above and drew a plot between Block and Larvae counts.

Problem 5

I knitted this document to pdf and pushed to GitHub.

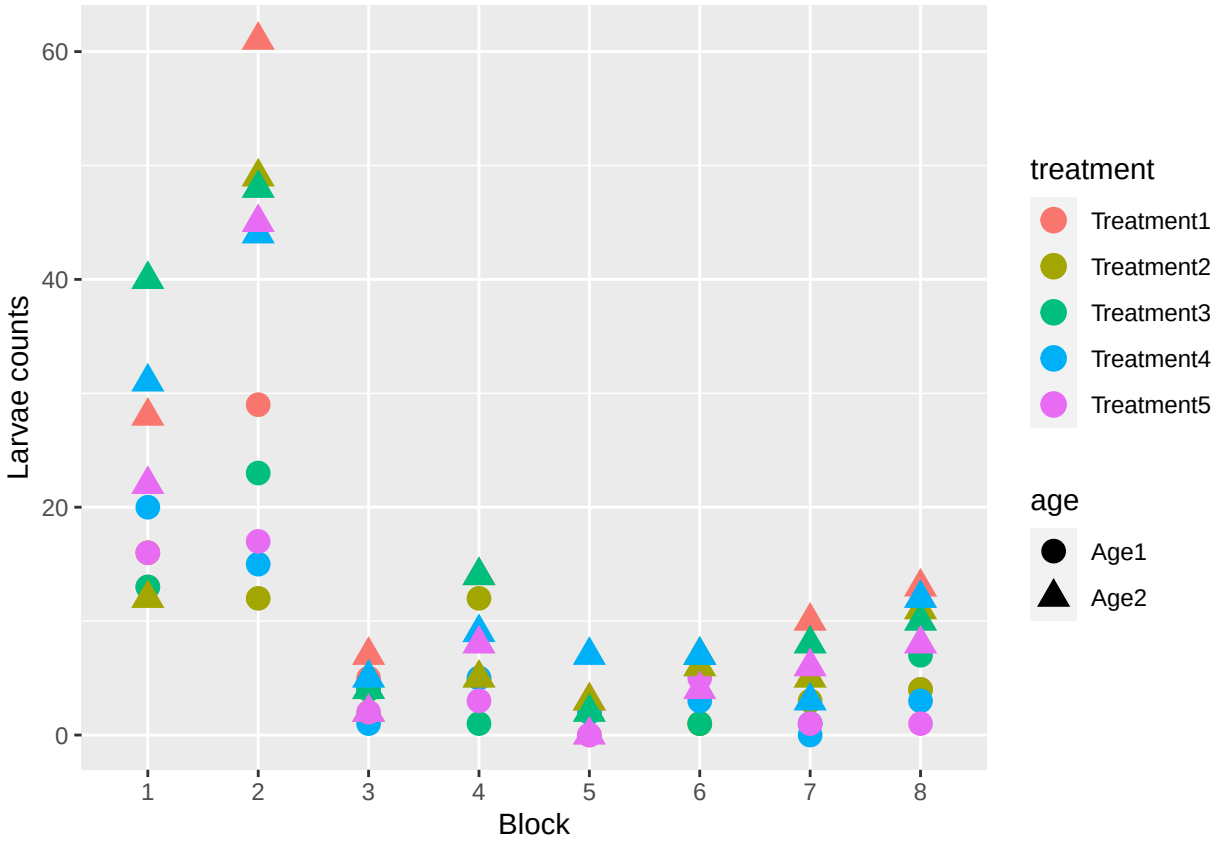


Figure 1: Larvae counts in different blocks