# HW4_boyuj

Boyu Jiang

10/16/2021

## Part A

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

# import data
pa <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/ThicknessGauge.dat",
                    header = FALSE, skip = 2, sep=" ")

# rename the columns
colnames(pa) <- c("part",
                  "operator1.1st","operator1.2nd",
                  "operator2.1st","operator2.2nd",
                  "operator3.1st","operator3.2nd")

# rearrange the data frame so that observations are distinguished by operator and measurement
pa <- melt(pa, id.vars = "part")

# separate operator and measurement into 2 columns
pa <- separate(data = pa, col = 'variable',
               into = c("operator", "measurement"))
pa$part <- factor(pa$part)
pa$operator <- factor(pa$operator)
pa$measurement <- factor(pa$measurement)

# show the table of data (first 6 observations)
knitr::kable(head(pa), caption = "Measurements of the part's wall thickness (partial)")
```

Table 1: Measurements of the part's wall thickness (partial)

| part | operator | measurement | value |
|---|---|---|---|
| 1 | operator1 | 1st | 0.953 |
| 2 | operator1 | 1st | 0.956 |
| 3 | operator1 | 1st | 0.956 |
| 4 | operator1 | 1st | 0.957 |
| 5 | operator1 | 1st | 0.957 |
| 6 | operator1 | 1st | 0.958 |

```r
# show the summary table of data
knitr::kable(summary(pa), caption="Summary of variables")
```

Table 2: Summary of variables

| part | operator | measurement | value |
|---|---|---|---|
| 1 : 6 | operator1:20 | 1st:30 | Min. :0.9520 |
| 2 : 6 | operator2:20 | 2nd:30 | 1st Qu.:0.9550 |
| 3 : 6 | operator3:20 | NA | Median :0.9570 |
| 4 : 6 | NA | NA | Mean :0.9561 |
| 5 : 6 | NA | NA | 3rd Qu.:0.9570 |
| 6 : 6 | NA | NA | Max. :0.9580 |
| (Other):24 | NA | NA | NA |

```r
# merge part and operator, compute the mean of 2-time measurement values
paplot <- aggregate(x = pa$value, by = list(pa$part, pa$operator), FUN = mean)

# plot the difference between each operator's measurement and mean value
paplot['difference'] <- paplot$x - mean(paplot$x)

barplot(difference ~ Group.2 + Group.1,
        data = paplot,
        beside = TRUE,
        xlab = "Part",
        ylab = "Difference from the mean value of measurement",
        col = c("skyblue2", "chocolate", "green"),
        ylim = c(-0.004,0.002),
        border = NA)
legend("bottom", c("Operator 1", "Operator 2","Operator 3"),
       fill = c("skyblue2", "chocolate", "green"),
       border = NA, horiz = TRUE)
```

## Part B

```r
library(dplyr)
library(tidyr)
library(reshape)

# import data
pb <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat",
                    header = FALSE, skip = 1, sep = " ")

# rename the columns
colnames(pb) <- rep(c("BodyWt", "BrainWt"), 3)

# rearrange data frame to 2 columns
pb <- rbind(pb[,1:2], pb[,3:4], pb[1:20,5:6])
```

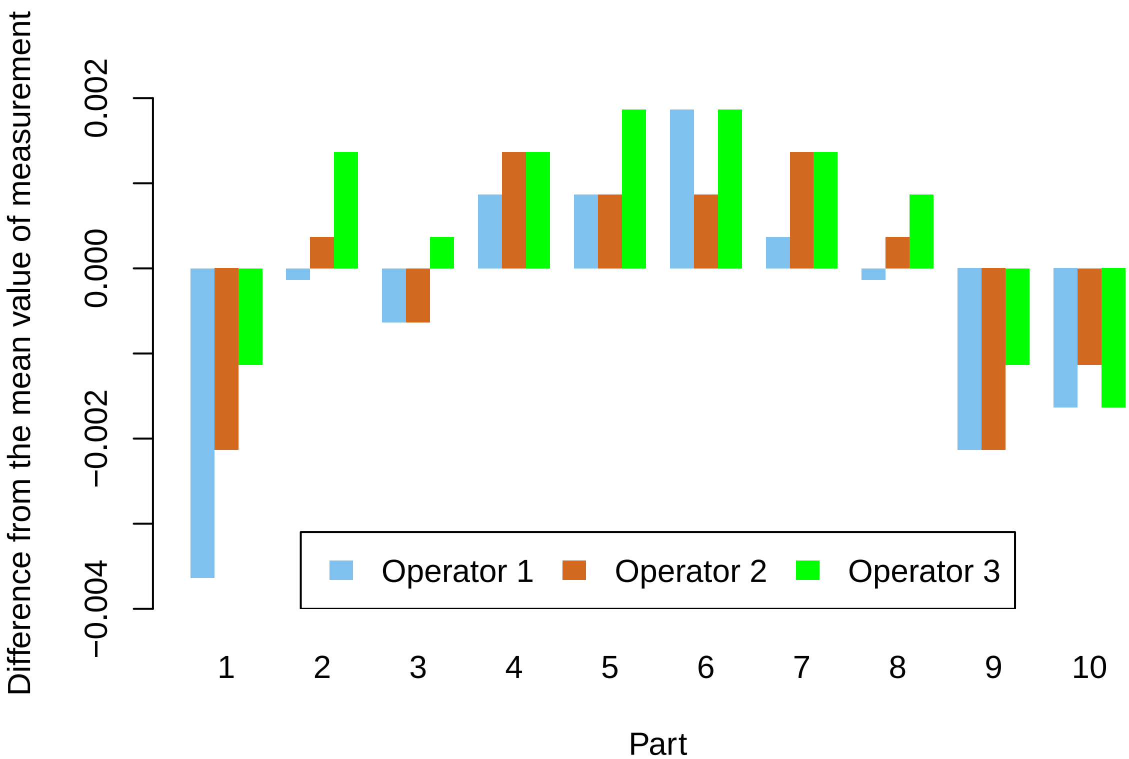


Figure 1: Measurements of wall thickness by three operators

```r
# show the table of data (first 6 observations)
knitr::kable(head(pb), caption = "Body and brain weight (partial)")
```

Table 3: Body and brain weight (partial)

| BodyWt | BrainWt |
|---|---|
| 3.385 | 44.5 |
| 0.480 | 15.5 |
| 1.350 | 8.1 |
| 465.000 | 423.0 |
| 36.330 | 119.5 |
| 27.660 | 115.0 |

```r
# show the summary table of data
knitr::kable(summary(pb), caption="Summary of variables")
```

Table 4: Summary of variables

| 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 |

```r
# scatter plot and fitted simple linear model
plot(x = pb$BodyWt, y = pb$BrainWt,
     col = "blue", pch = 16,
     xlab = 'Body Weight (kg)',
     ylab = 'Brain Weight (g)')
abline(lm(BrainWt ~ BodyWt, pb),
       col = "red")
legend(x = "topleft", legend = c("Raw data", "Regression line"),
       col = c("blue","red"), lty = c(0,1), pch = c(16,NA))
```

## Part C

```r
library(dplyr)
library(tidyr)
library(reshape)
library(data.table)

# import data
pc <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat",
                 header = FALSE, skip = 1, sep = " ")
```

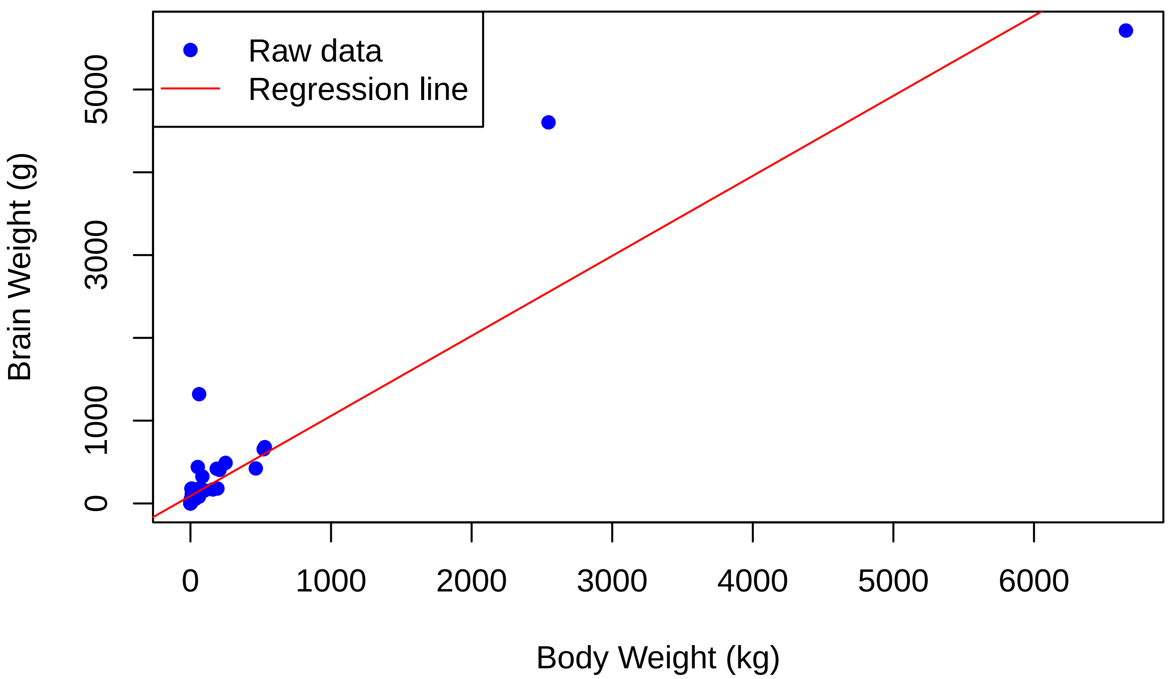


Figure 2: Body weight and brain weight

```r
# rename the columns
colnames(pc) <- rep(c("year", "long jump"), 4)

# rearrange data frame to 2 columns
pc <- rbind(pc[,1:2], pc[,3:4], pc[,5:6], pc[1:4, 7:8])
pc$year <- pc$year + 1900

# show the table of data (first 6 observations)
knitr::kable(head(pc), caption = "Gold Medal performance for Olympic Men's Long Jump (partial)")
```

Table 5: Gold Medal performance for Olympic Men's Long Jump (partial)

| year | long jump |
|---|---|
| 1896 | 249.75 |
| 1900 | 282.88 |
| 1904 | 289.00 |
| 1908 | 294.50 |
| 1912 | 299.25 |
| 1920 | 281.50 |

```r
# show the summary table of data
knitr::kable(summary(pc), caption="Summary of variables")
```

Table 6: Summary of variables

| 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 |

```r
# scatter plot and fitted simple linear model
plot(pc, col = "red", lwd = 3,
     type = 'l',
     xlab = 'Year',
     ylab = 'Gold Medal performance for Men's Long Jump (inch)')
```

## Part D

```r
library(dplyr)
library(tidyr)
library(reshape)
```

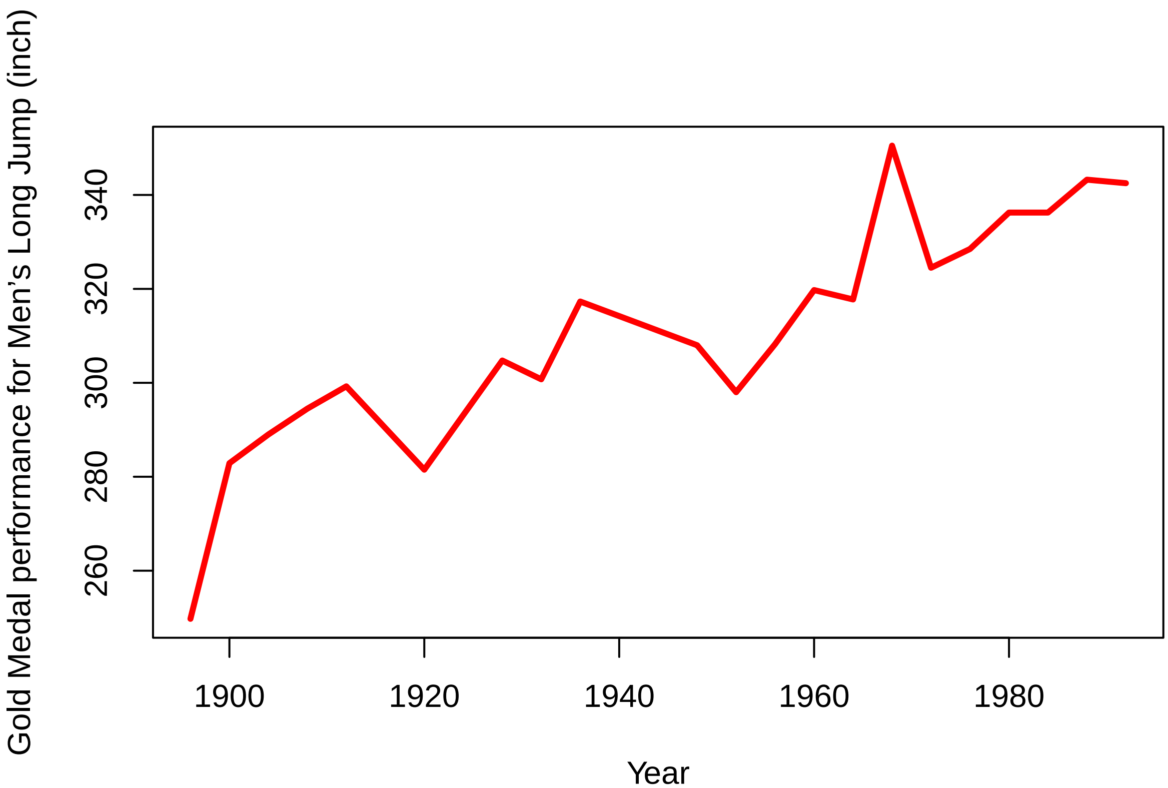


Figure 3: Gold Medal performance for Olympic Men's Long Jump

```r
library(data.table)

# import data
pd <- fread("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat",
                    header = FALSE, skip = 0, sep = " ", sep2 = ",")

# rename the columns
colnames(pd) <- c("category", "10k", "20k", "30k")

# separate columns
pd <- separate(data = pd, col = '10k',
               into = c("10k.1", "10k.2", "10k.3"),
               remove = TRUE, sep = ',')
pd <- separate(data = pd, col = '20k',
               into = c("20k.1", "20k.2", "20k.3"),
               remove = TRUE, sep = ',')
pd <- separate(data = pd, col = '30k',
               into = c("30k.1", "30k.2", "30k.3"),
               remove = TRUE, sep = ',')

# melt tomato categories so that observations are distinguished by Planting Density and measurement
pd <- melt(pd, id.vars = "category")

# separate columns to Planting Density and measurement
pd <- separate(data = pd, col = 'variable',
               into = c("PlantingDensity", "measurement"),
               remove = TRUE)

pd$category <- factor(pd$category)
pd$PlantingDensity <- factor(pd$PlantingDensity)
pd$measurement <- factor(pd$measurement)
pd$value <- as.numeric(pd$value)

# show the table of data (first 6 observations)
knitr::kable(head(pd), caption = "Measurements of tomato yield (partial)")
```

Table 7: Measurements of tomato yield (partial)

| category | PlantingDensity | measurement | value |
|---|---|---|---|
| Ife#1 | 10k | 1 | 16.1 |
| PusaEarlyDwarf | 10k | 1 | 8.1 |
| Ife#1 | 10k | 2 | 15.3 |
| PusaEarlyDwarf | 10k | 2 | 8.6 |
| Ife#1 | 10k | 3 | 17.5 |
| PusaEarlyDwarf | 10k | 3 | 10.1 |

```r
# show the summary table of data
knitr::kable(summary(pd), caption="Summary of variables")
```

Table 8: Summary of variables

| category | PlantingDensity | measurement | value |
|---|---|---|---|
| Ife#1 :9 | 10k:6 | 1:6 | Min. : 8.10 |
| PusaEarlyDwarf:9 | 20k:6 | 2:6 | 1st Qu.:12.95 |
| NA | 30k: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 |

```
# merge category and Planting Density, compute the mean of 3-time measurement values
pdplot <- aggregate(x = pd$value, by = list(pd$category, pd$PlantingDensity), FUN = mean)

# plot the yield by category and Planting Density
barplot(x ~ Group.1 + Group.2,
        data = pdplot,
        beside = TRUE,
        col = c("skyblue2", "chocolate"),
        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("skyblue2", "chocolate"),
       border = NA, horiz = TRUE)
```

# Part E

```
library(dplyr)
library(tidyr)
library(reshape)
library(data.table)
library(ggplot2)

# import data
pe <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LarvaeControl.dat",
                    header = FALSE, skip = 3, sep = " ")

pe <- pe[,colSums(is.na(pe))<nrow(pe)]

# rename the columns
colnames(pe) <- c("Block","Age1.Treatment1","Age1.Treatment2","Age1.Treatment3","Age1.Treatment4","Age1


# melt block so that observations are distinguished by age and treatment
pe <- melt(as.data.table(pe), id.vars = "Block")

# separate columns to Planting Density and measurement
pe <- separate(data = pe, col = 'variable',
```

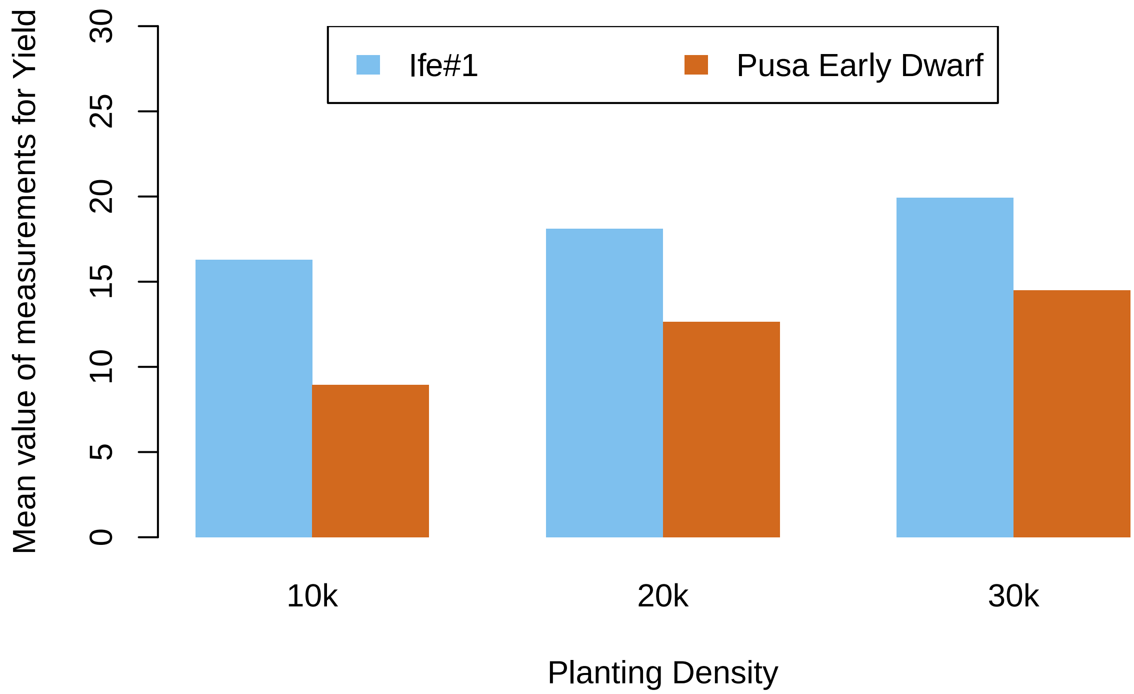


Figure 4: Measurements of tomato yield

```r
                into = c("age", "treatment"),
                remove = TRUE)

pe$Block <- factor(pe$Block)
pe$age <- factor(pe$age)
pe$treatment <- factor(pe$treatment)

# show the table of data (first 6 observations)
knitr::kable(head(pe), caption = "Larvae counts at two ages (partial)")
```

Table 9: Larvae counts at two ages (partial)

| 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 |

```r
# show the summary table of data
knitr::kable(summary(pe), caption="Summary of variables")
```

Table 10: 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 |

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

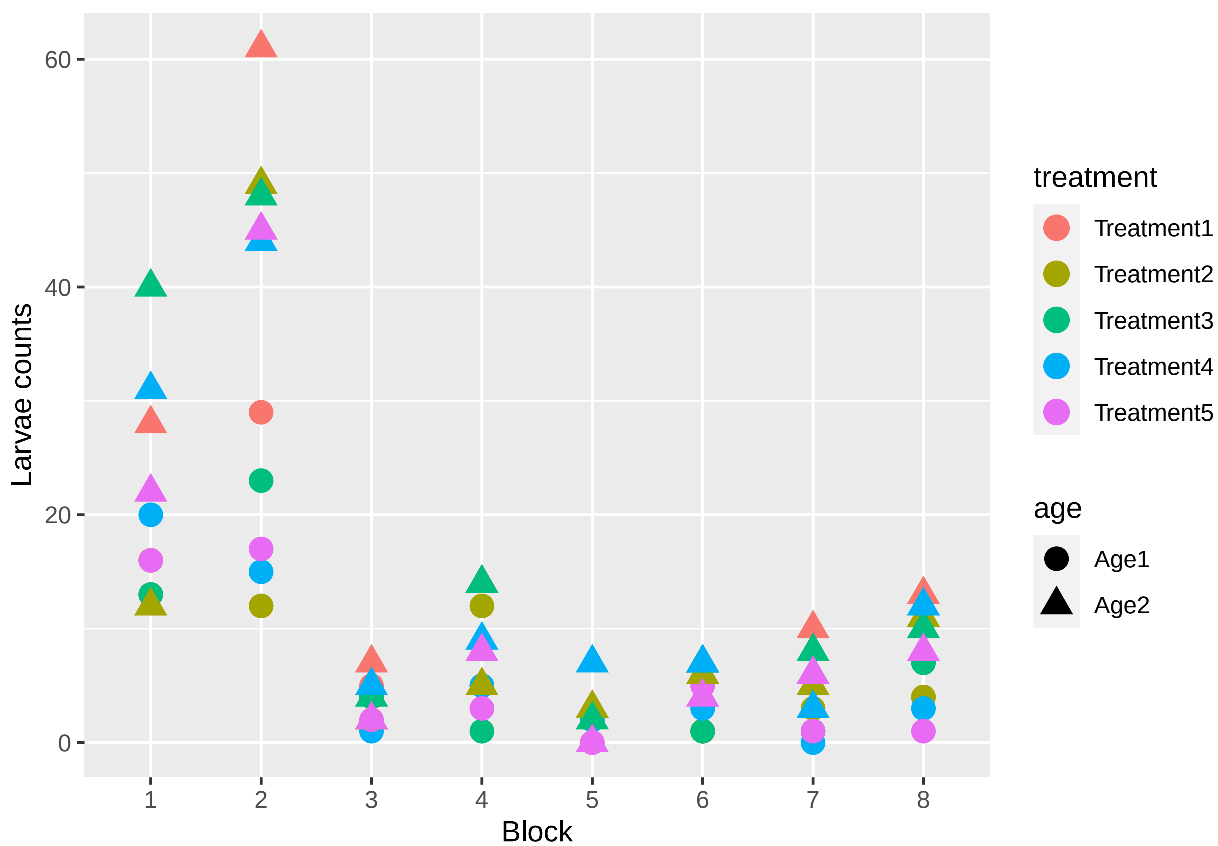


Figure 5: Larvae counts in different blocks