HW4_ssoon

Samuel Soon

10/11/2021

A

```
parta  <- read.delim("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/ThicknessGauge.dat",
                     header = TRUE, sep="\t")

# Use tidyr to separate the block into individual columns for tidying

parta <- (parta %>% separate(col=X,
                             into = c("Part", "O1_1", "O1_2", "O2_1", "O2_2", "O3_1", "O3_2"),
                             convert=TRUE, sep = " "))[2:nrow(parta),]

# Remove a unnecessary empty column

parta <- subset(parta, select = -c(Operator))

# Melt the dataframe so that observations are distinguished by operator and measurement

parta <- melt(parta, id.vars = "Part", value.name = "Thickness")

# Separate Operator and Measurement into 2 variables

parta <- parta %>% tidyr::extract(variable, into=c("Operator", "Measurement"),
                                  regex='([^_]+)_([^_]+)', convert=TRUE)

kable(head(parta),caption="Duplicate Measurements of Wall Parts Taken By Operators")
```

Table 1: Duplicate Measurements of Wall Parts Taken By Operators

Part	Operator	Measurement	Thickness
1	O1	1	0.953
2	O1	1	0.956
3	O1	1	0.956
4	O1	1	0.957
5	O1	1	0.957
6	O1	1	0.958

```
kable(summary(parta), caption="Summary")
```

Table 2: Summary

Part	Operator	Measurement	Thickness
Length:60	Length:60	Min. :1.0	Min. :0.9520
Class :character	Class :character	1st Qu.:1.0	1st Qu.:0.9550
Mode :character	Mode :character	Median :1.5	Median :0.9570
NA	NA	Mean :1.5	Mean :0.9561
NA	NA	3rd Qu.:2.0	3rd Qu.:0.9570
NA	NA	Max. :2.0	Max. :0.9580

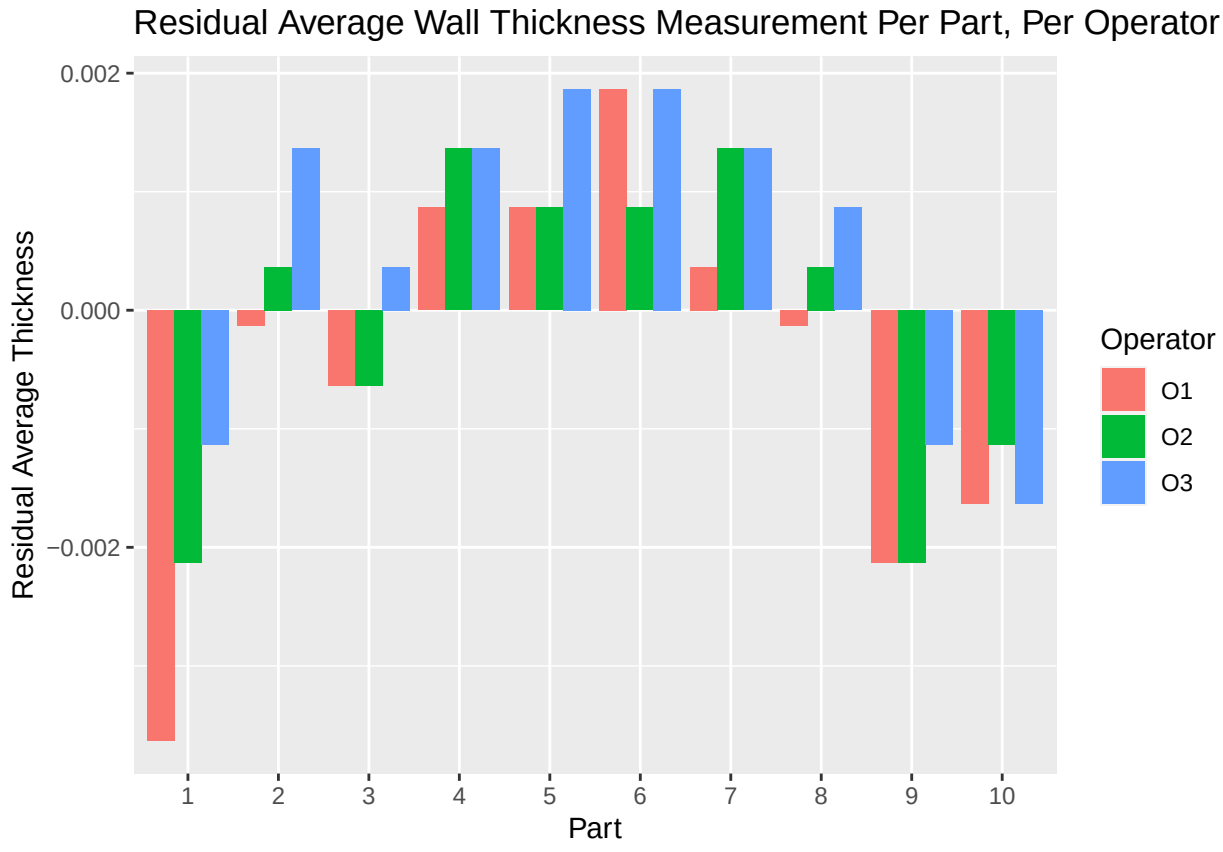
```
sum <- aggregate(parta$Thickness, list(parta$Part, parta$Operator), FUN=mean)

sum$x = sum$x - mean(sum$x)

sum$Group.1 <- as.factor(as.numeric(sum$Group.1))

# Chose to plot a graph that showed how each wall measurement differed from average observed wall thick

ggplot(sum,aes(x = Group.1,y = x,group=Group.2 ,fill = Group.2)) +
  geom_bar(stat = "identity",position = "dodge") +
  labs(x="Part", y="Residual Average Thickness", fill = "Operator",title="Residual Average Wall Thickne
```



B

```
partb  <- read.delim("http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat",
                     header = TRUE, sep="\t")

# Separate data into individual columns

partb <- partb %>% separate(col=Body.Wt.Brain.Wt.Body.Wt.Brain.Wt.Body.Wt.Brain.Wt,
                            into = c("1","2","3","4","5","6"), convert=TRUE, sep = " ")

# Collapse the repeating columns into single columns containing the information we need

body <- as.vector(as.matrix(partb[,c(TRUE,FALSE)]))
brain <- as.vector(as.matrix(partb[,c(FALSE, TRUE)]))

# Remove NA values

partb <- na.omit(data.frame("Body_Wt" = body, "Brain_Wt" = brain))

kable(head(partb),caption="Brain and Body Weight for 62 Species")
```

Table 3: Brain and Body Weight for 62 Species

Body_Wt	Brain_Wt
3.385	44.5
0.480	15.5
1.350	8.1
465.000	423.0
36.330	119.5
27.660	115.0

```
kable(summary(partb), caption="Summary")
```

Table 4: Summary

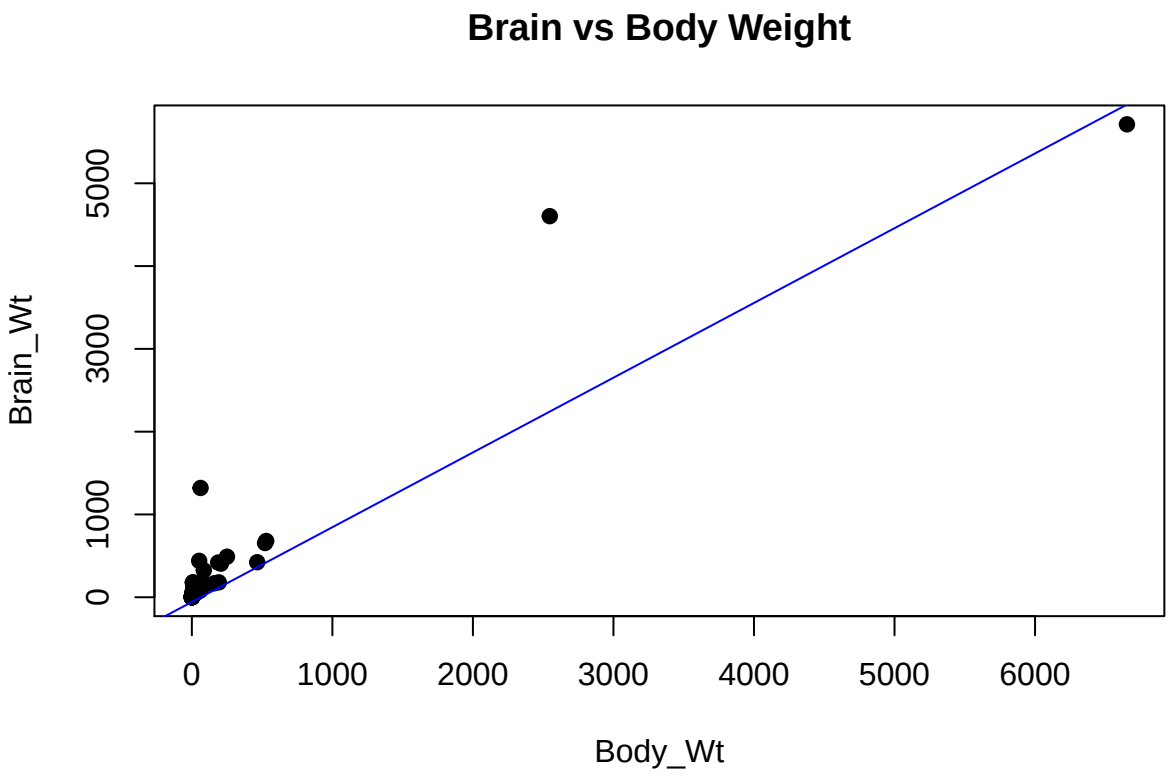
Body_Wt	Brain_Wt
Min. : 0.005	Min. : 0.10
1st Qu.: 0.600	1st Qu.: 4.25
Median : 3.342	Median : 17.25
Mean : 198.790	Mean : 283.13
3rd Qu.: 48.202	3rd Qu.: 166.00
Max. :6654.000	Max. :5712.00

```
model <- lm(Body_Wt ~ Brain_Wt, partb)

# Chose scatterplot with line of best fit to display the distribution of body weight

plot(partb, main="Brain vs Body Weight",pch=19)

abline(model, col="blue")
```



C

```
partc  <- fread("http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat",
                header=TRUE, sep="\t", sep2=" ", skip=-1)

# Separate data into individual columns

partc <- partc %>% separate(`Year Long Jump Year Long Jump Year Long Jump Year Long Jump`,
                            into = c("1","2","3","4","5","6"), convert=TRUE, sep = " ")

# Collapse columns

year <- as.vector(as.matrix(partc[,c(TRUE,FALSE)]))
ljump <- as.vector(as.matrix(partc[,c(FALSE, TRUE)]))

# Remove NA

partc <- na.omit(data.frame("Year" = year, "LongJump" = ljump))

kable(head(partc),caption="Long Jump Distance by Year")
```

Table 5: Long Jump Distance by Year

Year	LongJump
-4	249.75
0	282.88
4	289.00
8	294.50
12	299.25
20	281.50

```
kable(summary(partc), caption="Summary")
```

Table 6: Summary

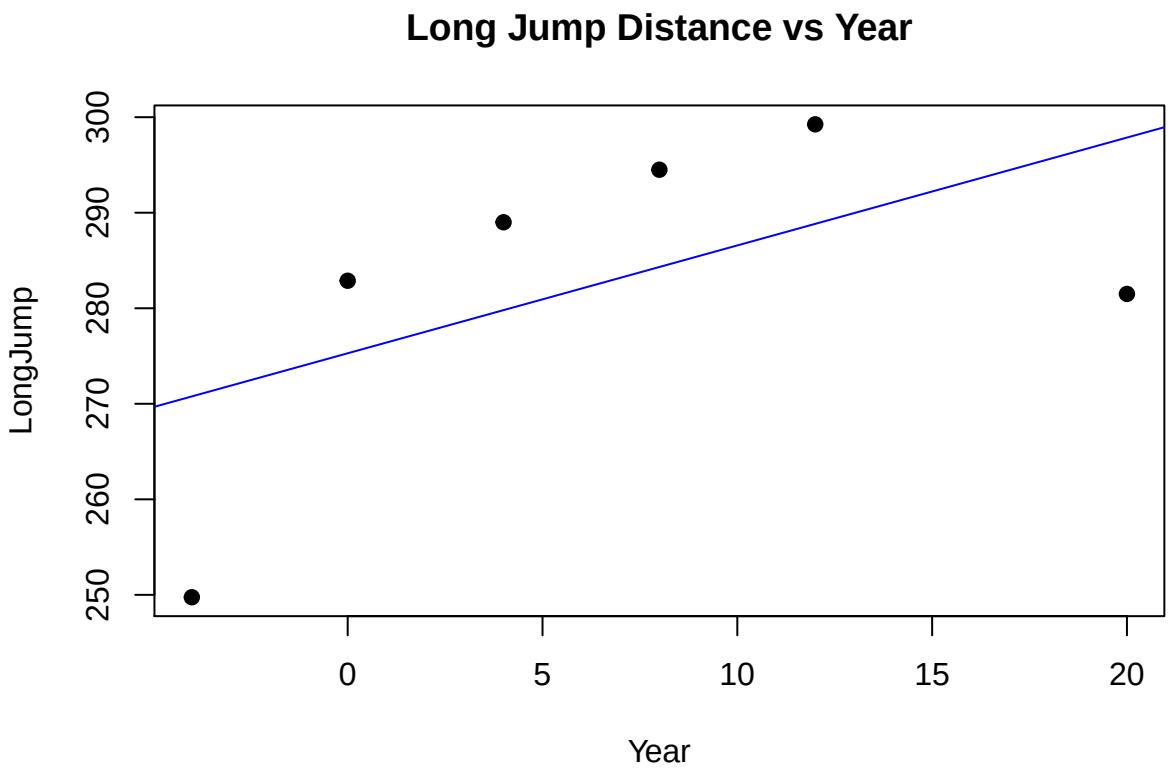
Year	LongJump
Min. :-4.000	Min. :249.8
1st Qu.: 1.000	1st Qu.:281.8
Median : 6.000	Median :285.9
Mean : 6.667	Mean :282.8
3rd Qu.:11.000	3rd Qu.:293.1
Max. :20.000	Max. :299.2

```
model <- lm(LongJump ~ Year, partc)

# Chose scatterplot with line of best fit to display the potential linear relationship between year and

plot(partc,main="Long Jump Distance vs Year", pch=19)

abline(model, col="blue")
```



D

```
partd  <- fread("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat",
                header=FALSE, sep=" ", sep2=",", skip=-1)

# Separate each of the yield columns into 3 columns based on planting density, measurement

for(i in 2:ncol(partd)){
  name <- paste("V", i, sep="")

  partd <- partd %>% tidyr::extract(name, into=paste((i-1)*10000, 1:3, sep="_"),
                                    regex='([^,]+),([^,]+),([^,]+)[,]*$', convert=TRUE)

}

# Melt the dataframe so that yield is distinguished by variety, density, and measurement

partd <-melt(partd, value.name="Yield",variable.name= 'Density_Measure') %>%
  rename(c("V1" =  "Variety"))
```

```
## Using V1 as id variables
```

```
# Separate the column generated by melting into two variables

partd <- partd %>% tidyr::extract(Density_Measure, into=c("Density", "Measurement_Number"),
```

```
                                regex='([^_]+)_([^_]+)', convert=TRUE)


kable(head(partd),caption="Tomato Yield Based on Variety, Planting Densities")
```

Table 7: Tomato Yield Based on Variety, Planting Densities

Variety	Density	Measurement_Number	Yield
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

```
kable(summary(partd), caption="Summary")
```

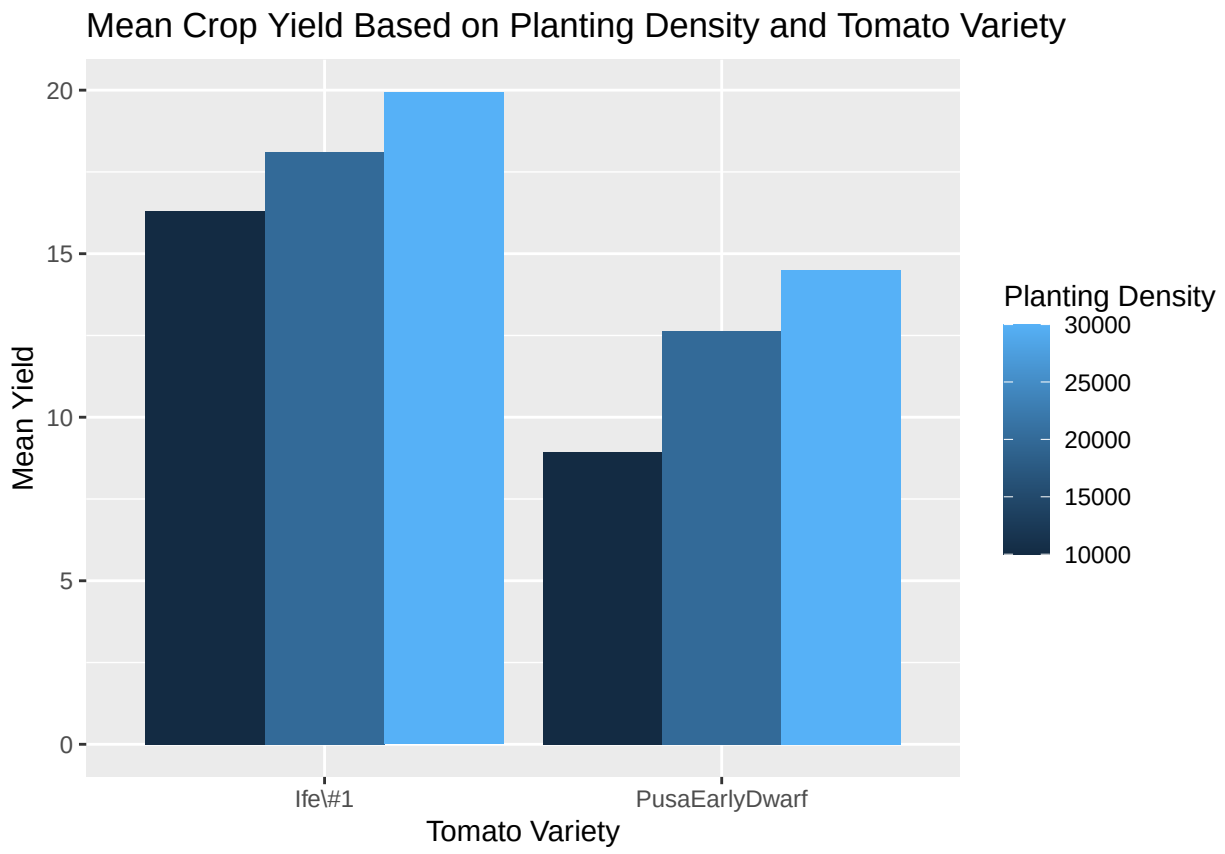
Table 8: Summary

Variety	Density	Measurement_Number	Yield
Length:18	Min. :10000	Min. :1	Min. : 8.10
Class :character	1st Qu.:10000	1st Qu.:1	1st Qu.:12.95
Mode :character	Median :20000	Median :2	Median :15.35
NA	Mean :20000	Mean :2	Mean :15.07
NA	3rd Qu.:30000	3rd Qu.:3	3rd Qu.:17.88
NA	Max. :30000	Max. :3	Max. :21.00

```
sum <- aggregate(partd$Yield, list(partd$Variety, partd$Density), FUN=mean)

# I chose to use a barplot to compare the mean observed crop yield between planting densities.

ggplot(sum,aes(x = Group.1,y = x,group=Group.2 ,fill = Group.2)) +
  geom_bar(stat = "identity",position = "dodge") +
  labs(x="Tomato Variety", y="Mean Yield", fill = "Planting Density",title="Mean Crop Yield Based on Pla
```



E

```
parte  <- fread("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LarvaeControl.dat",
                header=TRUE, sep=" ", sep2=" ", skip=1)

# Use renaming to note the age groups for convenience

parte <- parte %>% rename_with( ~ paste0("1_", .x), 2:6)

parte <- parte %>% rename_with( ~ paste0("2_", .x), 7:ncol(parte))

# Melt the data to distinguish count based on age, treatment

parte <- parte %>% melt(id.vars = c("Block"), variable.name = "Age_Treatment", value.name = "Count")

# Separate the melted column into 2 variables

parte <- parte %>% tidyr::extract(Age_Treatment, into=c("Age", "Treatment"), regex='([^_]+)_([^_]+)',
                                  convert=TRUE)

parte$Age <- as.factor(parte$Age)

kable(head(parte),caption="Larvae Counts Based on Age, Treatment")
```

Table 9: Larvae Counts Based on Age, Treatment

Block	Age	Treatment	Count
1	1	1	13
2	1	1	29
3	1	1	5
4	1	1	5
5	1	1	0
6	1	1	1

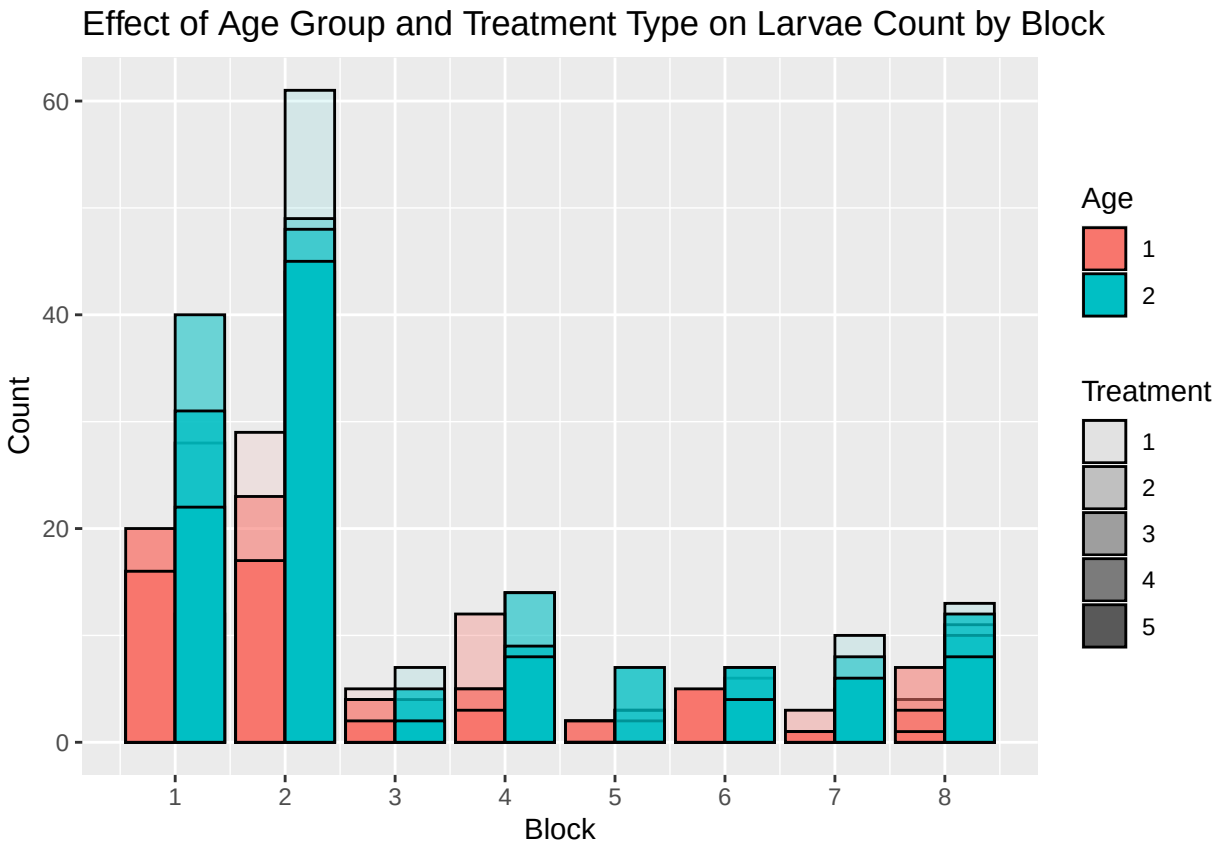
```
kable(summary(parte), caption="Summary")
```

Table 10: Summary

Block	Age	Treatment	Count
Min. :1.00	1:40	Min. :1	Min. : 0.00
1st Qu.:2.75	2:40	1st Qu.:2	1st Qu.: 2.75
Median :4.50	NA	Median :3	Median : 5.50
Mean :4.50	NA	Mean :3	Mean :10.50
3rd Qu.:6.25	NA	3rd Qu.:4	3rd Qu.:13.00
Max. :8.00	NA	Max. :5	Max. :61.00

```
# I chose to use a barplot that displays the count produced based on each age group across all blocks, 

ggplot(parte,aes(x = Block,y = Count,group=Age,fill = Age, alpha=Treatment)) +
  geom_bar(stat = "identity",position = "dodge", color="black") +
  labs(title="Effect of Age Group and Treatment Type on Larvae Count by Block") +
  scale_x_continuous(breaks = seq(0, 10, by = 1))
```


Effect of Age Group and Treatment Type on Larvae Count by Block
Count
60
40
20
0
1
2
3
4
5
6
7
8
Block
Age
1
2
Treatment
1
2
3
4
5