HW3_kexinx

Kexin Xie

Oct 15th 2021

Problem 3

Part A

```
url<-'https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/ThicknessGauge.dat'
thickness_raw<-read.table(url, header=F, skip=2, fill=T, stringsAsFactors = F)
colnames(thickness_raw)<-c("part","Operator1 Time1","Operator1 Time2","Operator2 Time1","Operator2 Time
```

We can notice that the column names in the original data set are placed in the wrong position, so I renamed each column and marked the measurement time. Tidying the dataset requires first melting, and then splitting the column into two variables: Operator and Times.

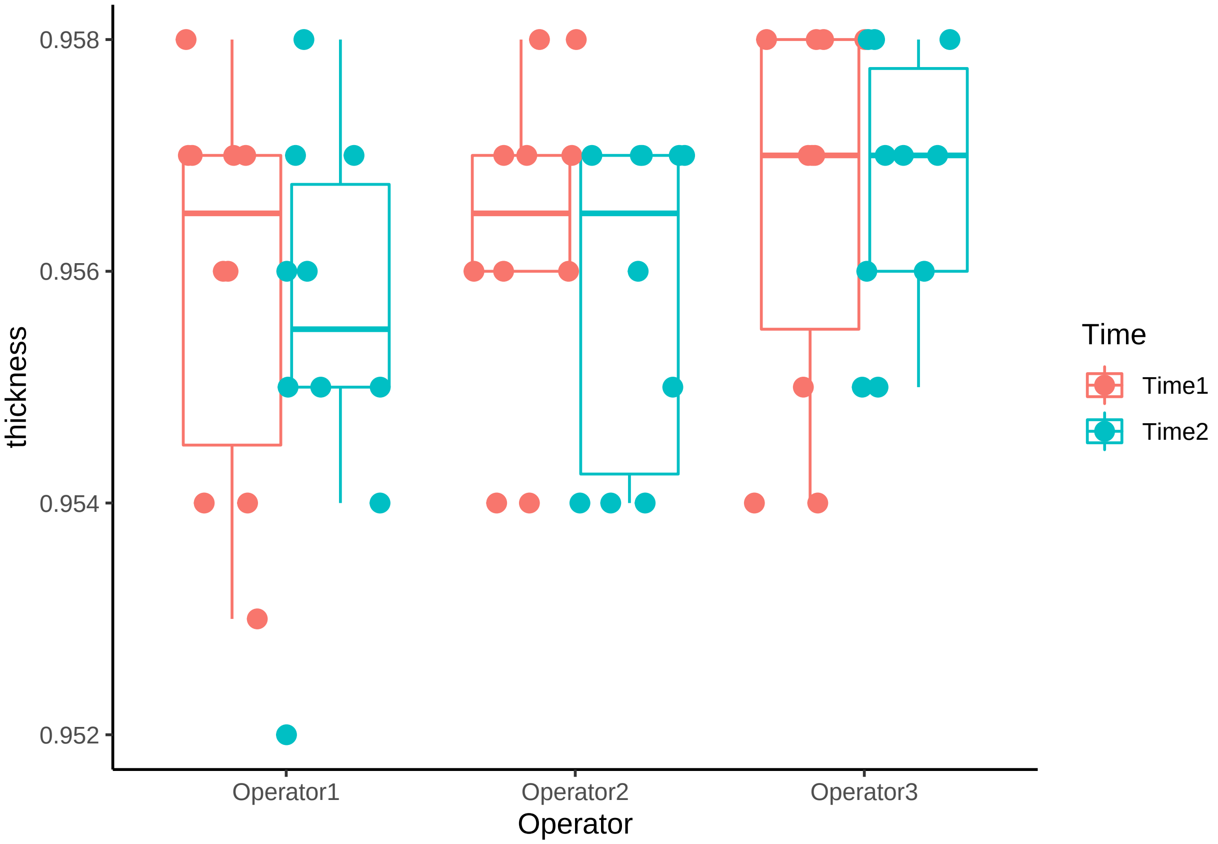
```
thickness_tidy<-thickness_raw %>%
                 gather(key = "Operator", value = "thickness",-part) %>%
                 separate(col = Operator, into = c("Operator","Times")) %>%
                 spread(key = Times, value = thickness)
knitr::kable(head(thickness_tidy),format = "markdown", caption="Tidy Thickness Data")
```

Table 1: Tidy Thickness Data

part	Operator	Time1	Time2
1	Operator1	0.953	0.952
1	Operator2	0.954	0.954
1	Operator3	0.954	0.956
2	Operator1	0.956	0.956
2	Operator2	0.956	0.957
2	Operator3	0.958	0.957

Thus, the dataset is in a tidy fashion. Then, we can use functions in ggplot package to plot our tidy data.

```
thickness_plot<-thickness_tidy %>%
  gather(`Time1`, `Time2`, key = "Time", value = "thickness") %>%
  ggplot(aes(x=Operator,y=thickness, color=Time))+geom_boxplot(outlier.shape = NA)+
    geom_point(aes(fill = Time), size = 3, shape = 21, position = position_jitterdodge())+
    theme_classic()
thickness_plot
```



The boxplot above indicates that operator 3 tends to measure higher values whereas operator 1 would like to measure lower values.

Part B

```
url<-'https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat'
BrainandBodyWeight_raw<-read.table(url, header=F,skip=2, fill=T, stringsAsFactors = F)
colnames(BrainandBodyWeight_raw)<-c("Body_Wt","Brain_Wt","Body_Wt","Brain_Wt","Body_Wt","Brain_Wt")
```

From the original dataset, two variables "Body weight" and "Brain weight" are listed in six columns. We should arrange the dataset into two columns to make each column contain each variable.

```
BrainandBodyWeight_tidy<-rbind(BrainandBodyWeight_raw[,1:2],
                               BrainandBodyWeight_raw[,3:4],BrainandBodyWeight_raw[,5:6])
BrainandBodyWeight_tidy<-BrainandBodyWeight_tidy %>% na.omit()
knitr::kable(head(BrainandBodyWeight_tidy),format = "markdown", caption="Tidy BrainandBodyWeight Data")
```

Table 2: Tidy BrainandBodyWeight Data

Body_Wt	Brain_Wt
0.48	15.5
1.35	8.1
465.00	423.0
36.33	119.5
27.66	115.0
14.83	98.2

```
knitr::kable(summary(BrainandBodyWeight_tidy),format = "markdown", caption="Summary BrainandBodyWeight
```

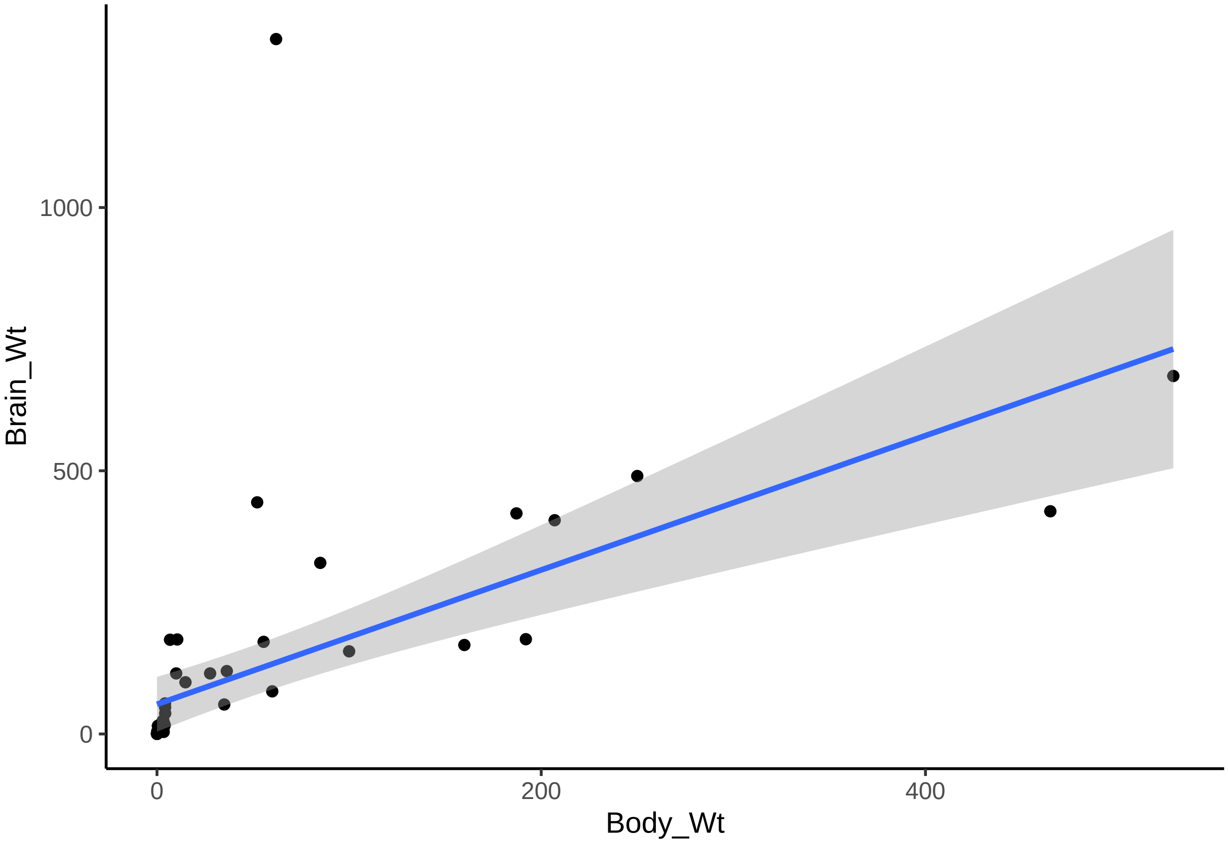
Table 3: Summary BrainandBodyWeight Data

Body_Wt	Brain_Wt
Min. : 0.005	Min. : 0.10
1st Qu.: 0.515	1st Qu.: 3.95
Median : 3.300	Median : 17.00
Mean : 199.968	Mean : 285.47
3rd Qu.: 44.245	3rd Qu.: 163.00
Max. :6654.000	Max. :5712.00

Then we use scatter plot and linear regression to see whether there is any relationship between Body weight and Brain weight.

```
BrainandBodyWeight_plot<-BrainandBodyWeight_tidy %>%
  filter(!Body_Wt>2000) %>%
  ggplot(aes(x=Body_Wt,y=Brain_Wt))+geom_point()+geom_smooth(method='lm')+theme_classic()
BrainandBodyWeight_plot
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We can notice that there are some outliers far away from the most data points. Looking at the original data, these outliers may be caused by the different units.

Part C

```
url<-'http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat'
LongJump_raw<-fread(url, header=F,skip=1,fill=T, stringsAsFactors = F)
```

Just like data in part b, we should rearrange the data from 8 columns to 2 columns and reformat the values of Year to make it readable for readers.

```
LongJump_raw<-as.data.frame(LongJump_raw)
colnames(LongJump_raw)<-c("Year","Long_jump","Year","Long_jump","Year","Long_jump","Year","Long_jump")
LongJump_tidy<-rbind(LongJump_raw[,1:2],LongJump_raw[,3:4],LongJump_raw[,5:6],LongJump_raw[,7:8])
LongJump_tidy<-LongJump_tidy%>%na.omit()%>%
     mutate(Year=Year+1900)
knitr::kable(head(LongJump_tidy),format = "markdown", caption="Tidy LongJump Data")
```

Table 4: Tidy LongJump Data

Year	Long_jump
1896	249.75
1900	282.88
1904	289.00
1908	294.50
1912	299.25
1920	281.50

```
knitr::kable(summary(LongJump_tidy),format = "markdown", caption="Summary LongJump Data")
```

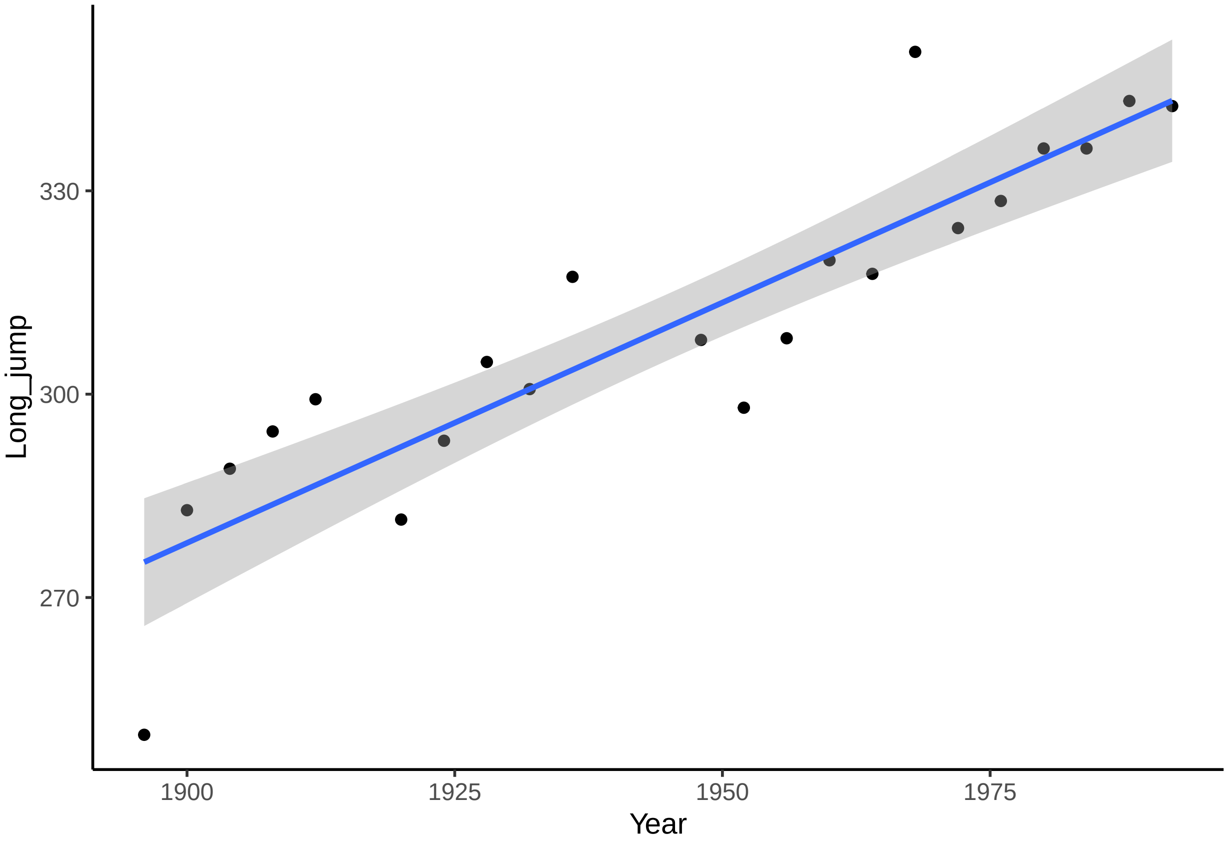
Table 5: Summary LongJump Data

	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

The relationship between year and long jump tends to linear.

```
LongJump_plot<-LongJump_tidy %>%
  ggplot(aes(x=Year,y=Long_jump))+geom_point()+geom_smooth(method='lm')+theme_classic()
LongJump_plot
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Part D

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

For this tomato data, we first separate the data in one columns to three columns and then combine the temporary tables together. To tidy the table, we also use gather to melt the data, and then splitting rename columns as Variety, Yield and Density.

```
tomato_tidy1<-tomato_raw[,1:2] %>% separate(col='V2',into=c("y1", "y2", "y3"),sep=',',remove=TRUE) %>%
              mutate(Density=rep(10000,2))
tomato_tidy2<-tomato_raw[,c(1,3)] %>% separate(col='V3',into=c("y1", "y2", "y3"),sep=',',remove=TRUE) %
              mutate(Density=rep(20000,2))
tomato_tidy3<-tomato_raw[,c(1,4)] %>% separate(col='V4',into=c("y1", "y2", "y3"),sep=',',remove=TRUE) %
              mutate(Density=rep(30000,2))
tomato_tidy<-rbind(tomato_tidy1,tomato_tidy2,tomato_tidy3)
tomato_tidy<-tomato_tidy %>% rename(c("Variety" =  "V1")) %>%
               gather(key=Times,value=Yield,c(y1,y2,y3))%>%
               arrange(Variety,Density) %>% select(-Times) %>%
               mutate(Yield=as.numeric(Yield), Variety=as.factor(Variety))
knitr::kable(head(tomato_tidy),format = "markdown", caption="Tidy Tomato Data")
```

Table 6: Tidy Tomato Data

Variety	Density	Yield
Ife#1	10000	16.1

Variety	Density	Yield
Ife#1	10000	15.3
Ife#1	10000	17.5
Ife#1	20000	16.6
Ife#1	20000	19.2
Ife#1	20000	18.5

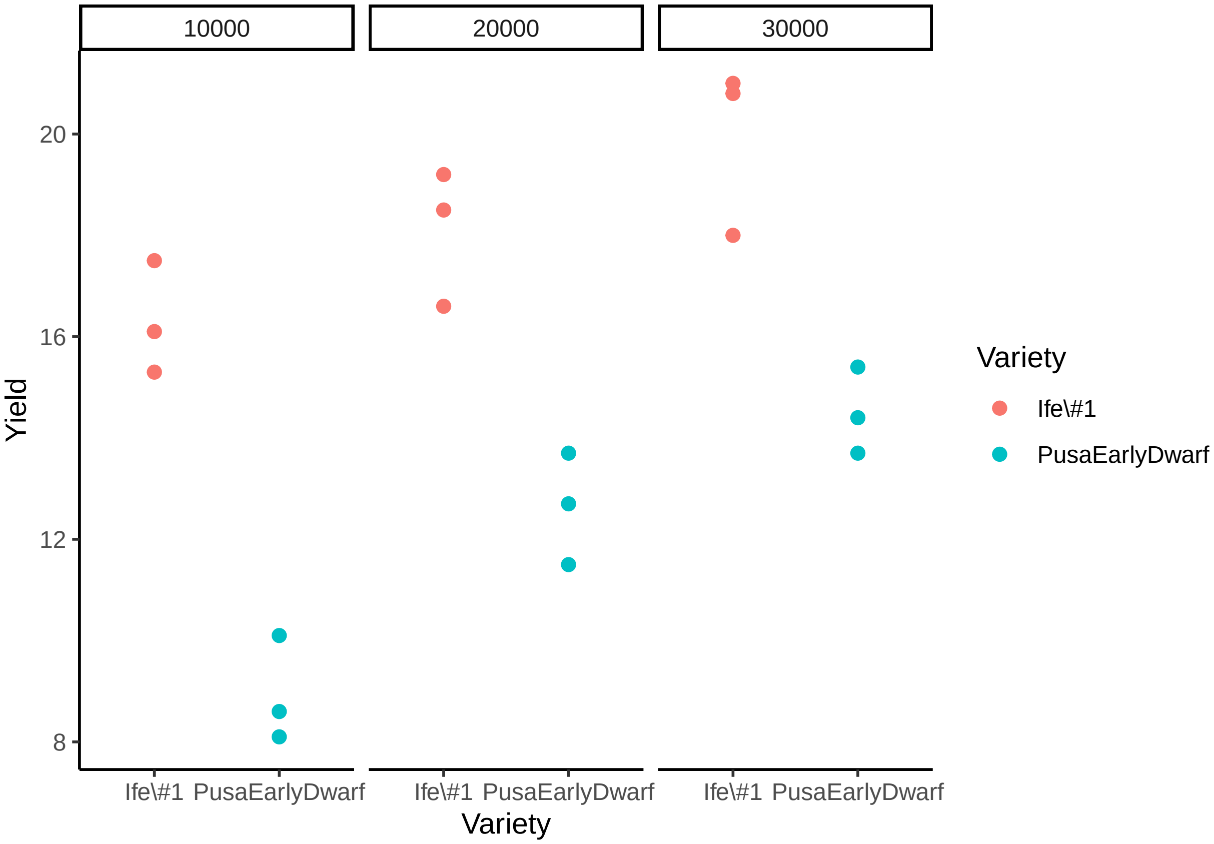
```r
knitr::kable(summary(tomato_tidy), caption="Summary Tomato Data")
```

Table 7: Summary Tomato Data

Variety	Density	Yield
Ife#1 :9	Min. :10000	Min. : 8.10
PusaEarlyDwarf:9	1st Qu.:10000	1st Qu.:12.95
NA	Median :20000	Median :15.35
NA	Mean :20000	Mean :15.07
NA	3rd Qu.:30000	3rd Qu.:17.88
NA	Max. :30000	Max. :21.00

From the scatter plot, it's easy to notice that the tomato yields increases as the planting densities increasing in both two varieties of tomato. Moreover, the first variety of tomato is always having higher yields than the other one.

```r
# plot
tomato_plot<-tomato_tidy %>%
  ggplot(aes(x=Variety,y=Yield,group=Variety, color=Variety))+
  geom_point(aes(fill = Variety), size = 2, shape = 21)+
  facet_grid(~factor(Density))+theme_classic()
tomato_plot
```



Part E

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

To tidy this data, we should rename the columns, combine the temporary tables togather and then melt the table.

```
LarvaeControl_tidy1<-LarvaeControl_raw[,2:6] %>%
  rename("Treatment1"=`1`,"Treatment2"=`2`,"Treatment3"=`3`,"Treatment4"=`4`,"Treatment5"=`5`)%>%
  mutate(Age=rep(1,8))
LarvaeControl_tidy2<-LarvaeControl_raw[,7:11] %>%
  rename("Treatment1"=`\t1`,"Treatment2"=`2`,"Treatment3"=`3`,"Treatment4"=`4`,"Treatment5"=`5`)%>%
  mutate(Age=rep(2,8))
LarvaeControl_tidy<-rbind(LarvaeControl_tidy1,LarvaeControl_tidy2)
LarvaeControl_tidynew<-LarvaeControl_tidy %>% gather(key=Treatment, value=Counts,-Age)
knitr::kable(head(LarvaeControl_tidynew),format = "markdown", caption="Tidy LarvaeControl Data")
```

Table 8: Tidy LarvaeControl Data

Age	Treatment	Counts
1	Treatment1	13
1	Treatment1	29
1	Treatment1	5
1	Treatment1	5
1	Treatment1	0

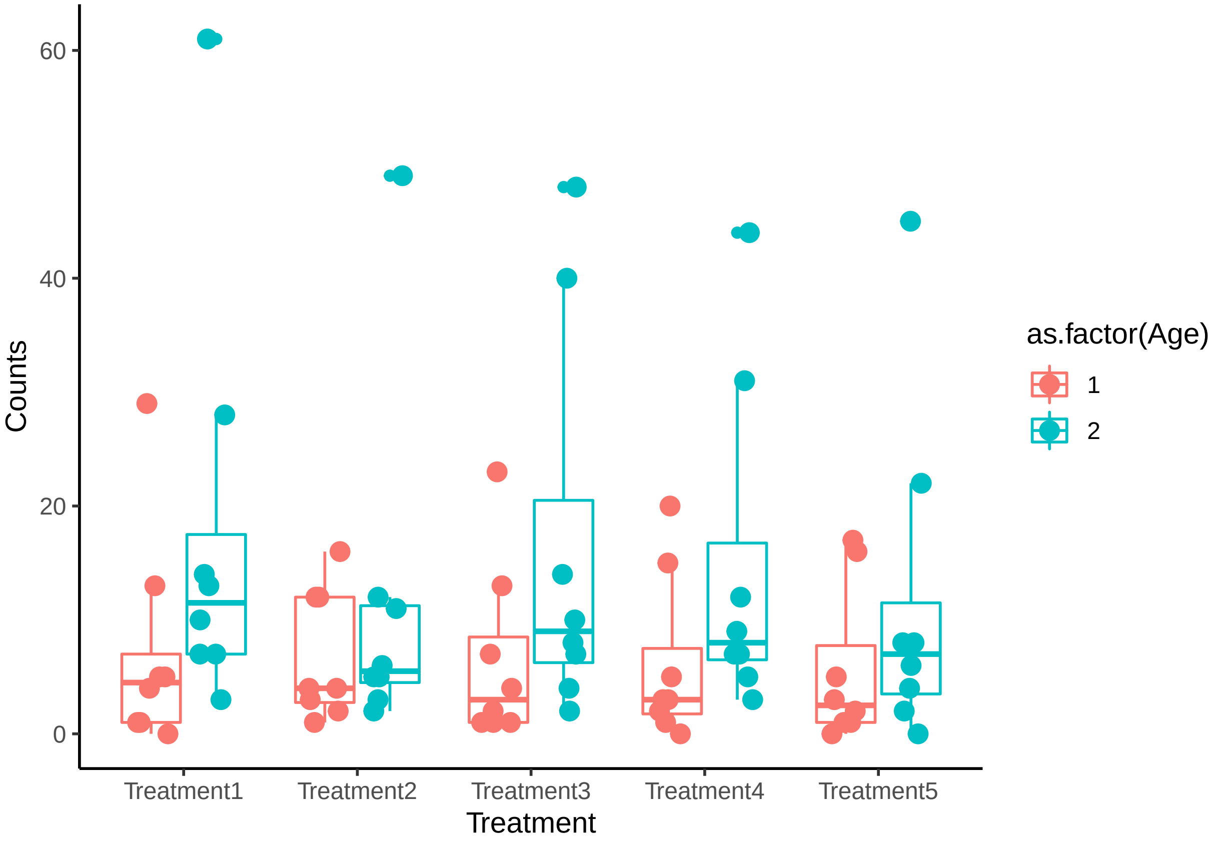
Age	Treatment	Counts
1	Treatment1	1

```
knitr::kable(summary(LarvaeControl_tidynew),format = "markdown", caption="Summary LarvaeControl Data")
```

Table 9: Summary LarvaeControl Data

Age	Treatment	Counts
Min. :1.0	Length:80	Min. : 0.00
1st Qu.:1.0	Class :character	1st Qu.: 2.75
Median :1.5	Mode :character	Median : 5.50
Mean :1.5	NA	Mean :10.50
3rd Qu.:2.0	NA	3rd Qu.:13.00
Max. :2.0	NA	Max. :61.00

```
# Plot
LarvaeControl_plot<-LarvaeControl_tidynew %>%
  ggplot(aes(x=Treatment,y=Counts,color=as.factor(Age)))+geom_boxplot()+
  geom_point(aes(fill = as.factor(Age)), size = 3, shape = 21, position = position_jitterdodge())+
  theme_classic()
LarvaeControl_plot
```



As shown in the boxplot, larvae in age group 2 have higher numbers.