STAT 5014 hw3

Owusu Noah

October 25, 2023

Part A

The issue associated with the thickness guage data is that each operator column appears trice and it is a duplication issue that requires tidying of the data.

```
#importing thickness guage data

dat <- read.csv(file = "Thickness.csv")
pivot_dat <- dat %>%
  pivot_longer(cols = c("opt1","opt2","opt3"), names_to = "Operator",
               values_to = "Thickness") #making the data more tidy
```

```
#creating table for few observations
kable(head(pivot_dat), align = "cr",
      caption = "First six rows of the Thickness dataset")
```

Table 1: First six rows of the Thickness dataset

Part	Operator	Thickness
1	opt1	0.953
1	opt2	0.954
1	opt3	0.954
2	opt1	0.956
2	opt2	0.956
2	opt3	0.958

```
#obtaining summary statistics of the data
s1 <- pivot_dat %>%
  group_by(Part) %>%
  get_summary_stats(Thickness, show = c("mean", "median", "min", "max"))

kable(head(s1), align = "cr",
      caption = "Summary statistics of the Thickness dataset across Operator")
```

Table 2: Summary statistics of the Thickness dataset across Operator

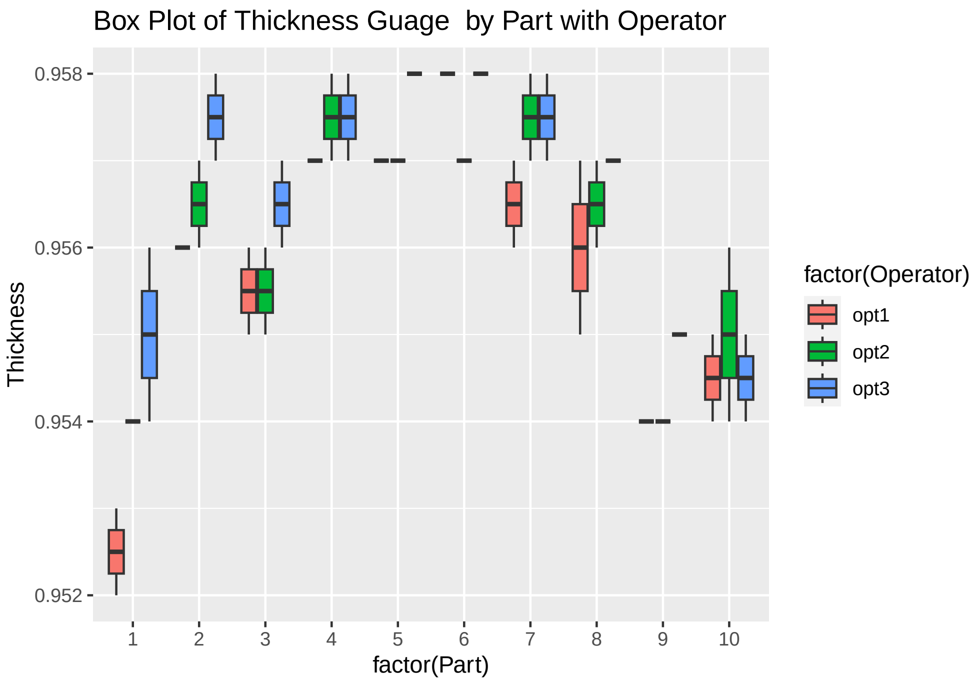
Part	variable	n	mean	median	min	max
1	Thickness	6	0.954	0.954	0.952	0.956
2	Thickness	6	0.957	0.956	0.956	0.958
3	Thickness	6	0.956	0.956	0.955	0.957
4	Thickness	6	0.957	0.957	0.957	0.958
5	Thickness	6	0.957	0.957	0.957	0.958
6	Thickness	6	0.958	0.958	0.957	0.958

```r
s2 <- pivot_dat %>%
  group_by(Operator) %>%
  get_summary_stats(Thickness, show = c("mean", "median", "min", "max"))
kable(head(s2), align = "cr",
      caption = "Summary statistics of the Thickness dataset across Part")
```

Table 3: Summary statistics of the Thickness dataset across Part

Operator	variable	n	mean	median	min	max
opt1	Thickness	20	0.956	0.956	0.952	0.958
opt2	Thickness	20	0.956	0.956	0.954	0.958
opt3	Thickness	20	0.957	0.957	0.954	0.958

```r
# Creating Box Plot for the data
ggplot(pivot_dat, aes(x = factor(Part), y = Thickness, fill = factor(Operator))) +
  geom_boxplot() +
  labs(title = "Box Plot of Thickness Guage  by Part with Operator")
```



PART B

Similar to part (A), the Body and Brain weight data has duplication issues. There are also missing values in this data.

```
# Importing the Body and Brain weight data
bbw <- read.csv("BodyBrain_wt.csv")

#changing colum names
colnames(bbw) = c("Body", "Brain", "Body", "Brain", "Body", "Brain")

bd <- as.vector(rbind(bbw$Body,bbw$Body.1,bbw$Body.2))
bn <- as.vector(rbind(bbw$Brain,bbw$Brain.1,bbw$Brain.2))

df <- data.frame("Body wt" = bd, "Brain wt" = bn)

p_df <- df %>%
  pivot_longer(cols = c("Body.wt","Brain.wt"), names_to = "variable",
               values_to = "weight")
```

```r
kable(head(p_df), align = "cr",
      caption = "First six rows of the Body and Brain weight dataset")
```

Table 4: First six rows of the Body and Brain weight dataset

variable	weight
Body.wt	3.385
Brain.wt	44.500
Body.wt	0.480
Brain.wt	15.500
Body.wt	1.350
Brain.wt	8.100

```r
#obtaining summary statistics
df1 <- p_df %>%
  group_by(factor(variable)) %>%
  get_summary_stats(weight, show = c("mean", "median", "min", "max"))
kable(head(df1), align = "cr",
      caption = "Summary statistics of the Body and Brain weight dataset")
```

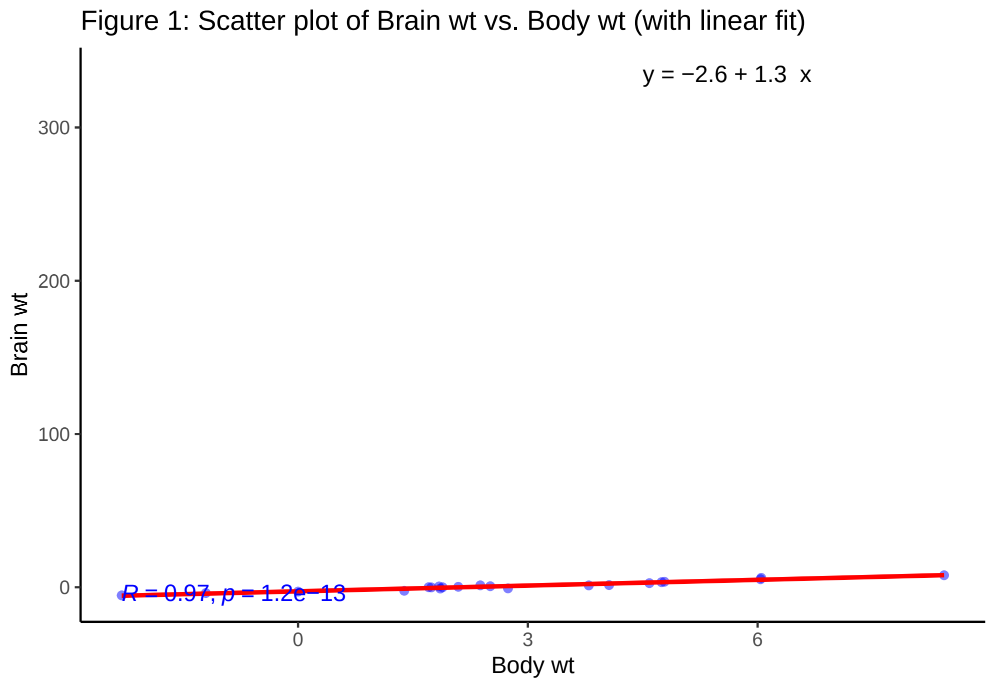
Table 5: Summary statistics of the Body and Brain weight dataset

factor(variable)	variable	n	mean	median	min	max
Body.wt	weight	21	157.052	1.7	0.005	2547
Brain.wt	weight	21	283.943	10.8	0.100	4603

```r
# Create scatter plot for body and brain weight data
new_df <- df %>%
  mutate(across(c(Body.wt, Brain.wt), function(x) log(x)))

new_df <- na.omit(new_df)

ggplot(new_df, aes(x = Brain.wt, y = Body.wt))+
geom_point(alpha = 0.5, col = "blue")+
geom_smooth(method = "lm", se = F, formula = y~x, col = "red")+
ggtitle("Figure 1: Scatter plot of Brain wt vs. Body wt (with linear fit)")+
stat_cor(method = ('pearson'),col='blue')+
stat_regline_equation(label.x=4.5, label.y=335, output.type = "latex")+
labs(x = "Body wt", y = "Brain wt")+
theme_classic()
```



PART C

The Long Jump data have duplication issue. Each variable appears four times on the column. There are also missing values in this data.

```
#importing long jump dataset
long <- read.csv("LongJump.csv")

l1 <- as.vector(rbind(long$Year,long$Year.1,long$Year.2))
l2 <- as.vector(rbind(long$Jump,long$Jump.1,long$Jump.2))

df <- data.frame("Year" = l1, "Jump" = l2)

ndf <- df %>%
  pivot_longer(cols = c("Year","Jump"), names_to = "variable",
               values_to = "value")
kable(head(ndf), align = "cr",
      caption = "First six rows of the Long Jump dataset")
```

Table 6: First six rows of the Long Jump dataset

variable	value
Year	-4.00
Jump	249.75
Year	24.00
Jump	293.13
Year	56.00

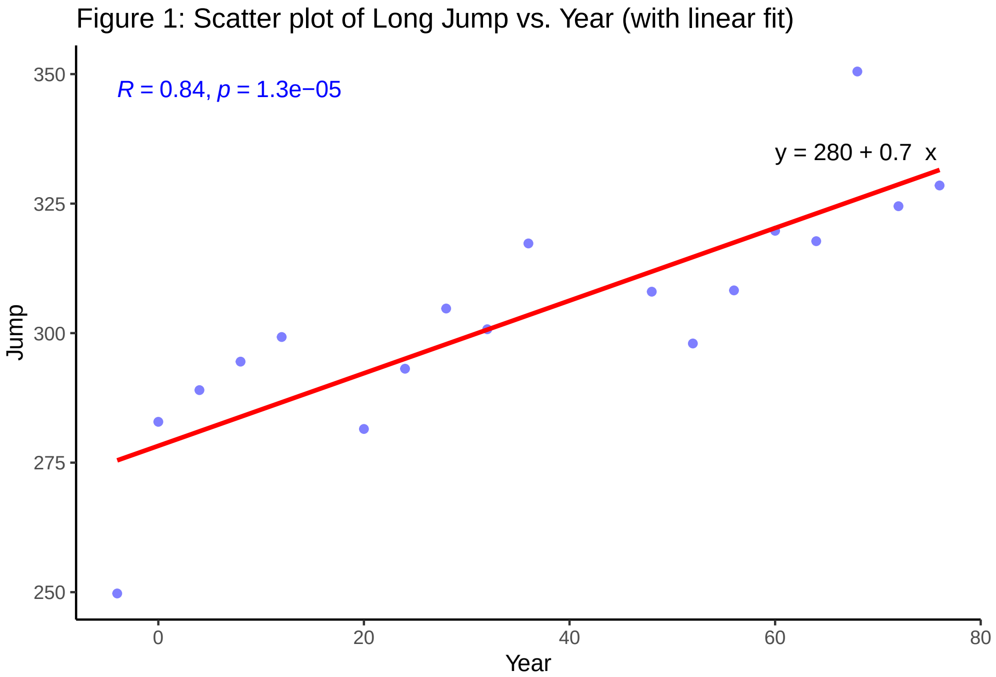
variable	value
Jump	308.25

```
ndf1 <- ndf %>%
  group_by(factor(variable)) %>%
  get_summary_stats(value, show = c("mean", "median", "min", "max"))
kable(head(ndf1), align = "cr", caption = "Summary statistics of the LongJump dataset")
```

Table 7: Summary statistics of the LongJump dataset

factor(variable)	variable	n	mean	median	min	max
Jump	value	18	303.782	302.75	249.75	350.5
Year	value	18	36.444	34.00	-4.00	76.0

```
#Creating a scatter plot for the long jump dataset
ggplot(df, aes(x = Year, y = Jump))+
geom_point(alpha = 0.5, col = "blue")+
geom_smooth(method = "lm", se = F, formula = y~x, col = "red")+
ggtitle("Figure 1: Scatter plot of Long Jump vs. Year (with linear fit)")+
stat_cor(method = ('pearson'),col='blue')+
stat_regline_equation(label.x=60, label.y=335, output.type = "latex")+
labs(x = "Year", y = "Jump")+
theme_classic()
```



PART D

The issue associated with the tomato data is that the values are embedded between each row and each column.

```r
#importing the tomatoe dataset
tom <- read.csv(file = "tomato.csv")
kable(head(tom), align = "cr",
      caption = "First six rows of the tomatoe dataset")
```

Table 8: First six rows of the tomatoe dataset

Variety	Unit	PlantDensity
Ife	10000	16.1
Ife	10000	15.3
Ife	10000	17.5
Ife	20000	16.6
Ife	20000	19.2
Ife	20000	18.5

```r
t1 <- tom %>%
  group_by(Variety) %>%
  get_summary_stats(PlantDensity, show = c("mean", "median", "min", "max"))
kable(head(t1), align = "cr", caption = "Summary statistics of the LongJump
      dataset across variety")
```

Table 9: Summary statistics of the LongJump dataset across variety

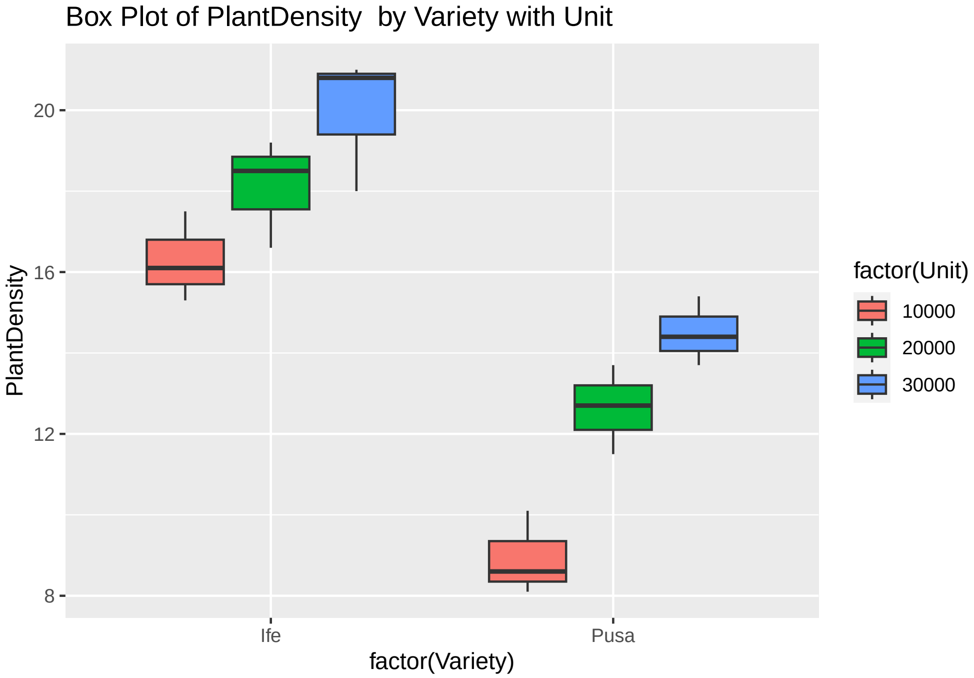
Variety	variable	n	mean	median	min	max
Ife	PlantDensity	9	18.111	18.0	15.3	21.0
Pusa	PlantDensity	9	12.022	12.7	8.1	15.4

```r
t2 <- tom %>%
  group_by(Unit) %>%
  get_summary_stats(PlantDensity, show = c("mean", "median", "min", "max"))
kable(head(t2), align = "cr",
      caption = "Summary statistics of the LongJump dataset across unit")
```

Table 10: Summary statistics of the LongJump dataset across unit

Unit	variable	n	mean	median	min	max
10000	PlantDensity	6	12.617	12.70	8.1	17.5
20000	PlantDensity	6	15.367	15.15	11.5	19.2
30000	PlantDensity	6	17.217	16.70	13.7	21.0

```
#Creating Box Plot for tomatoe dataset
ggplot(tom, aes(x = factor(Variety), y = PlantDensity, fill = factor(Unit))) +
  geom_boxplot() +
  labs(title = "Box Plot of PlantDensity  by Variety with Unit")
```



Part E

In the Larvae counts data, the Age and the Treatment variables seem to be interacted and have embedded values.

```
#importing Larve counts dataset
LV <- "LarvaeCounts.csv" %>%
read.csv()

lv_dat <- LV %>%
  arrange(desc(Counts))

kable(head(lv_dat), align = "cr", caption = "First six rows of Larvae Count
      dataset arranged in a descending order of counts")
```

Table 11: First six rows of Larvae Count dataset arranged in a descending order of counts

Block	Age	Treatment	Counts
2	Age 2	1	61
2	Age 2	2	49
2	Age 2	3	48
2	Age 2	5	45
2	Age 2	4	44
1	Age 2	3	40

```r
dt <- lv_dat %>%
  group_by(Treatment,Age) %>%
  get_summary_stats(Counts, show = c("mean", "median", "min", "max"))
kable(head(dt), align = "cr", caption = "Summary of the Larvae count dataset
      based on Age and Treatment")
```

Table 12: Summary of the Larvae count dataset based on Age and Treatment

Age	Treatment	variable	n	mean	median	min	max
Age 1	1	Counts	8	7.250	4.5	0	29
Age 2	1	Counts	8	17.875	11.5	3	61
Age 1	2	Counts	8	6.750	4.0	1	16
Age 2	2	Counts	8	11.625	5.5	2	49
Age 1	3	Counts	8	6.500	3.0	1	23
Age 2	3	Counts	8	16.625	9.0	2	48

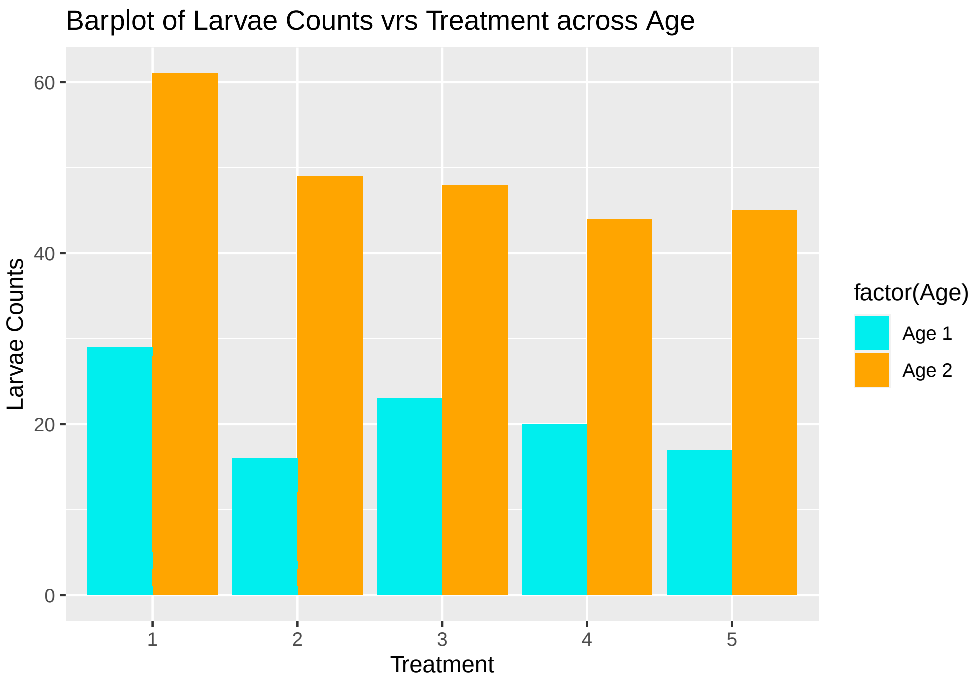
```r
dt <- lv_dat %>%
  group_by(Block) %>%
  get_summary_stats(Counts, show = c("mean", "median", "min", "max"))
kable(head(dt), align = "cr", caption = "Summary of the Larvae count dataset
      based on Block")
```

Table 13: Summary of the Larvae count dataset based on Block

Block	variable	n	mean	median	min	max
1	Counts	10	21.1	18.0	12	40
2	Counts	10	34.3	36.5	12	61
3	Counts	10	3.6	4.0	1	7
4	Counts	10	7.6	6.5	1	14
5	Counts	10	2.1	2.0	0	7

Block	variable	n	mean	median	min	max
6	Counts	10	4.2	4.5	1	7

```
#Creating barplot for Larvae counts dataset
ggplot(lv_dat, aes(x=factor(Treatment),y=Counts,fill=factor(Age))) +
geom_bar(position="dodge",stat='identity')+
ggtitle("Barplot of Larvae Counts vrs Treatment across Age")+
xlab("Treatment")+ ylab("Larvae Counts")+
scale_fill_manual(values = c("cyan2","orange"))
```


Barplot of Larvae Counts vrs Treatment across Age
Larvae Counts
Treatment
factor(Age)
Age 1
Age 2