HW4_kmg0122

Mingang Kim

2021 10 16

1.

2.

3.

```
library(tidyverse)
```

```
## Warning:   'tidyverse'  R   4.1.1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning:   'ggplot2'  R   4.1.1
```

```
## Warning:   'tibble'  R   4.1.1
```

```
## Warning:   'tidyr'  R   4.1.1
```

```
## Warning:   'readr'  R   4.1.1
```

```
## Warning:   'purrr'  R   4.1.1
```

```
## Warning:   'dplyr'  R   4.1.1
```

```
## Warning:   'forcats'  R   4.1.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(knitr)
```

```
## Warning:    'knitr'  R   4.1.1
```

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

```
## Warning:    'data.table'  R   4.1.1
```

```
##
##          : 'data.table'
```

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

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

(a).

```
data.3.a<-read.csv("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/ThicknessGauge.dat",
                   header=F, skip=2, sep = " ")


colnames(data.3.a)<-c("index", c("a1","a2","b1","b2","c1","c2"))


data.temp.1 <- data.3.a %>% select(index,contains("1")) %>% rename(a=a1, b=b1, c=c1)
data.temp.2 <- data.3.a %>% select(index,contains("2")) %>%
  rename(a=a2, b=b2, c=c2) %>% mutate(index=index+10)

dat.3.a <- bind_rows(data.temp.1, data.temp.2) %>%
  rename("Operator 1"=a,"Operator 2"= b, "Operator 3"=c)

kable(head(dat.3.a), caption="Wall Thickness")
```

Table 1: Wall Thickness

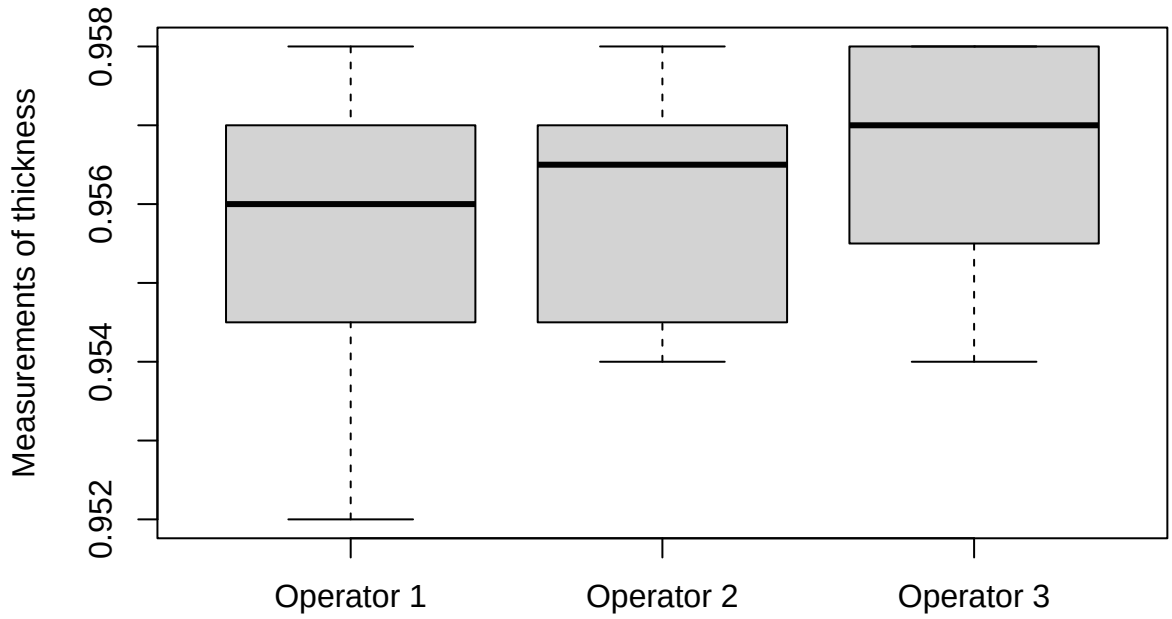
index	Operator 1	Operator 2	Operator 3
1	0.953	0.954	0.954
2	0.956	0.956	0.958
3	0.956	0.956	0.957
4	0.957	0.958	0.957
5	0.957	0.957	0.958
6	0.958	0.957	0.958

```
kable(summary(dat.3.a), caption="Wall Thickness Summary")
```

Table 2: Wall Thickness Summary

index	Operator 1	Operator 2	Operator 3
Min. : 1.00	Min. :0.9520	Min. :0.9540	Min. :0.9540
1st Qu.: 5.75	1st Qu.:0.9547	1st Qu.:0.9547	1st Qu.:0.9557
Median :10.50	Median :0.9560	Median :0.9565	Median :0.9570
Mean :10.50	Mean :0.9557	Mean :0.9560	Mean :0.9566
3rd Qu.:15.25	3rd Qu.:0.9570	3rd Qu.:0.9570	3rd Qu.:0.9580
Max. :20.00	Max. :0.9580	Max. :0.9580	Max. :0.9580

```
#Draw box plot to compare each operator's distribution
boxplot(dat.3.a[,2:4], ylab = "Measurements of thickness")
```



(b).

```
data.3.b<-read.csv("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat",
                   header=F, sep=" ", skip=1)

#rename each column because dplyr cannot road data when column name is duplicated.
```

```
colnames(data.3.b)<-paste(rep(c("Body_Wt","Brain_Wt"),3),c(1,1,2,2,3,3),sep="_")

#made temporary data to merge by row.
data.temp.1<-data.3.b %>% select(contains("1")) %>%

  rename( Body_Wt=Body_Wt_1, Brain_Wt=Brain_Wt_1)
data.temp.2<-data.3.b %>% select(contains("2")) %>%

  rename( Body_Wt=Body_Wt_2, Brain_Wt=Brain_Wt_2)
data.temp.3<-data.3.b %>% select(contains("3")) %>%

  rename( Body_Wt=Body_Wt_3, Brain_Wt=Brain_Wt_3)

#bind data by row. bind_row merge data by row which have the
# same column name. And then omitted NA data.
dat.3.b<-bind_rows(data.temp.1,data.temp.2,data.temp.3) %>% na.omit()

kable(head(dat.3.b), caption="Body and Brain weight Data")
```

Table 3: Body and Brain weight Data

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(dat.3.b), 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

```
#Draw scatter plot to see if there is any relationship between Body weight and Brain weight

plot(dat.3.b$Body_Wt, dat.3.b$Brain_Wt, xlab="Body Weight",
     ylab="Brain Weight", main="Scatter plot between Body weight and Brain weight")
```

(c).

```
data.3.c<-fread("http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat",
                sep=" ", fill=T, skip=1, col.names = c(paste0(rep(c("year", "Long_Jump"),4),c(1,1,2,2,3

#made temporary data to merge by row.
data.temp.1<-data.3.c %>% select(contains("1")) %>% rename(year=year1, Long_Jump=Long_Jump1) %>% mutate
data.temp.2<-data.3.c %>% select(contains("2")) %>% rename(year=year2, Long_Jump=Long_Jump2) %>% mutate
data.temp.3<-data.3.c %>% select(contains("3")) %>% rename(year=year3, Long_Jump=Long_Jump3) %>% mutate
data.temp.4<-data.3.c %>% select(contains("4")) %>% rename(year=year4, Long_Jump=Long_Jump4) %>% mutate

# data merge
dat.3.c<-bind_rows(data.temp.1,data.temp.2,data.temp.3, data.temp.4)

# data summary
kable(summary(dat.3.c), caption="Summary")
```

Table 5: Summary

year	Long_Jump
Min. :1896	Min. :249.8
1st Qu.:1921	1st Qu.:295.4
Median :1950	Median :308.1

year	Long_Jump
Mean :1945	Mean :310.3
3rd Qu.:1971	3rd Qu.:327.5
Max. :1992	Max. :350.5
NA's :2	NA's :2

```
#drop NA values
dat.3.c <- dat.3.c %>% na.omit()

kable(head(dat.3.c), caption="Body and Brain weight Data")
```

Table 6: Body and Brain weight Data

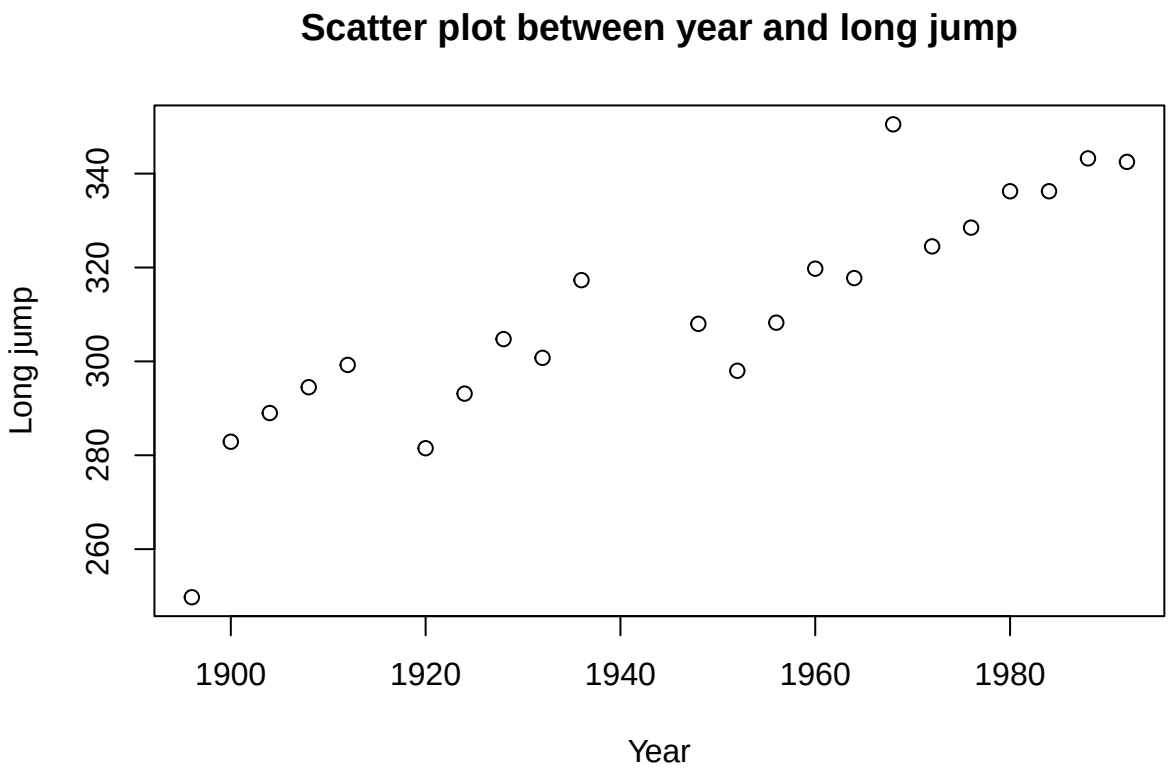
year	Long_Jump
1896	249.75
1900	282.88
1904	289.00
1908	294.50
1912	299.25
1920	281.50

```
#check that NA values disappear
kable(summary(dat.3.c), caption="Summary")
```

Table 7: Summary

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
plot(dat.3.c$year, dat.3.c$Long_Jump, xlab="Year",
     ylab="Long jump", main="Scatter plot between year and long jump")
```



(d).

```
data.3.d<-fread("http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat")
```

```
## Warning in fread("http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/
## tomato.dat"): Detected 3 column names but the data has 4 columns (i.e. invalid
## file). Added 1 extra default column name for the first column which is guessed
## to be row names or an index. Use setnames() afterwards if this guess is not
## correct, or fix the file write command that created the file to create a valid
## file.
```

```
data.temp.1 <- data.3.d[,c(1,2)] %>% separate(`10000`, c("y1", "y2", "y3"), sep=",") %>%
  mutate(Density=rep(10000,2))
```

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

```
data.temp.2 <- data.3.d[,c(1,3)] %>% separate(`20000`, c("y1", "y2", "y3"), sep=",") %>%
  mutate(Density=rep(20000,2))
data.temp.3 <- data.3.d[,c(1,4)] %>% separate(`30000`, c("y1", "y2", "y3"), sep=",") %>%
  mutate(Density=rep(30000,2))

dat.3.d <-bind_rows(data.temp.1, data.temp.2, data.temp.3 )
```

```
dat.3.d <-dat.3.d %>% rename(variety=V1) %>%
  gather(key="Try", value = "yields", y1, y2, y3) %>% select(-Try)

#change yields to numeric variable
dat.3.d<-dat.3.d %>% mutate(yields=as.numeric(yields), variety=as.factor(variety))

#print data
kable(head(dat.3.d), caption="Tomato yield")
```

Table 8: Tomato yield

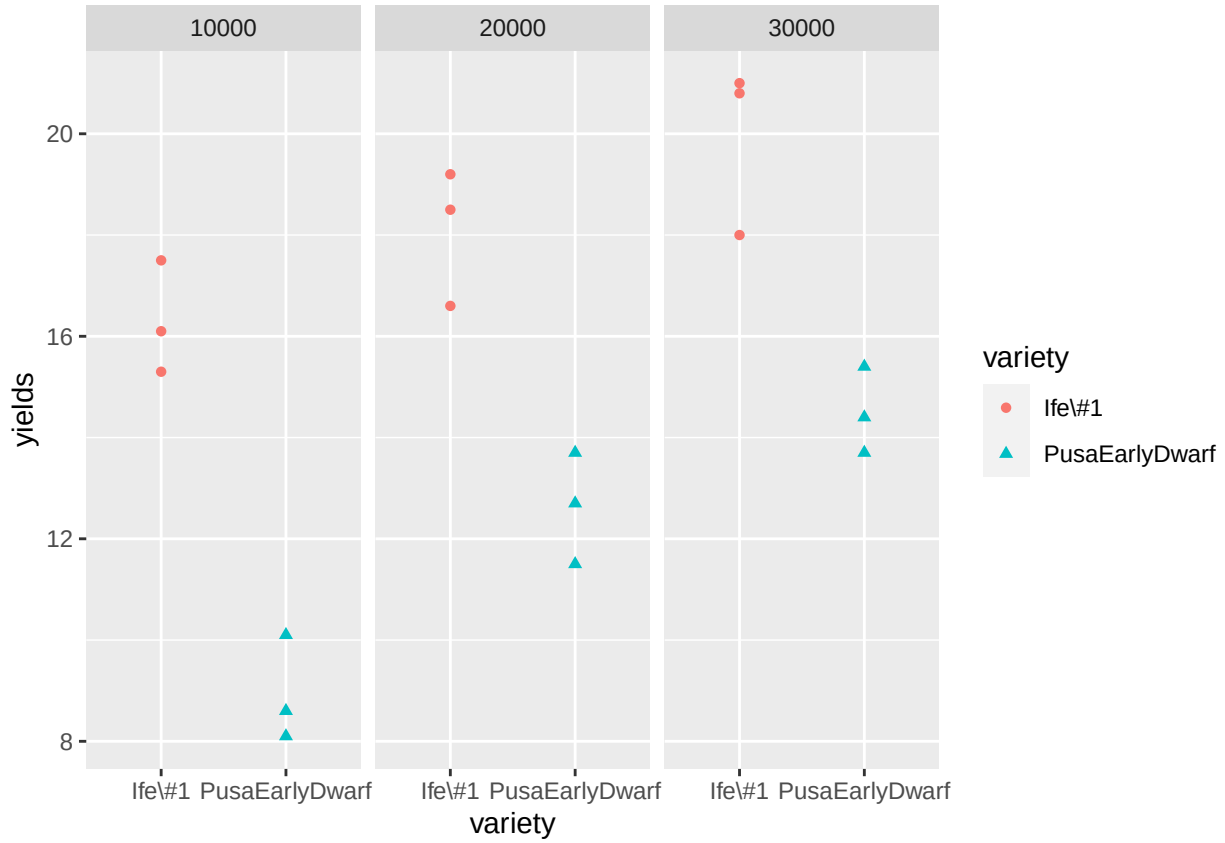
variety	Density	yields
Ife#1	10000	16.1
PusaEarlyDwarf	10000	8.1
Ife#1	20000	16.6
PusaEarlyDwarf	20000	12.7
Ife#1	30000	20.8
PusaEarlyDwarf	30000	14.4

```
#check summary
kable(summary(dat.3.d), caption="Summary")
```

Table 9: Summary

variety	Density	yields
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

```
#plot
dat.3.d %>% ggplot(aes(x=variety, y=yields,
           group=variety,
           shape=variety,
           color=variety)) +
           geom_point() +
           facet_grid(~factor(Density))
```



(e).

```
data.3.e<-fread("https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LarvaeControl.dat")

#add age colunm and rename number with treatment
data.temp.age1<-data.3.e[,c(1:6)] %>% rename(Trt1=`1`, Trt2=`2`, Trt3=`3`, Trt4=`4`, Trt5=`5`) %>%
  mutate(age=rep(1,8))
data.temp.age2 <- data.3.e[,c(1,7:11)] %>%
  rename(Trt1=` 1`, Trt2=`2`, Trt3=`3`, Trt4=`4`, Trt5=`5`) %>%
  mutate(age=rep(2,8))

dat.3.e <- bind_rows(data.temp.age1, data.temp.age2) %>%

  gather(key="Treatment", value="Counts", Trt1, Trt2, Trt3, Trt4, Trt5) %>%
  mutate(Treatment=substring(Treatment,4), age=as.factor(age), Block=as.factor(Block))

#print data
kable(head(dat.3.e), caption="Larvae Counts")
```

Table 10: Larvae Counts

Block	age	Treatment	Counts
1	1	1	13
2	1	1	29
3	1	1	5

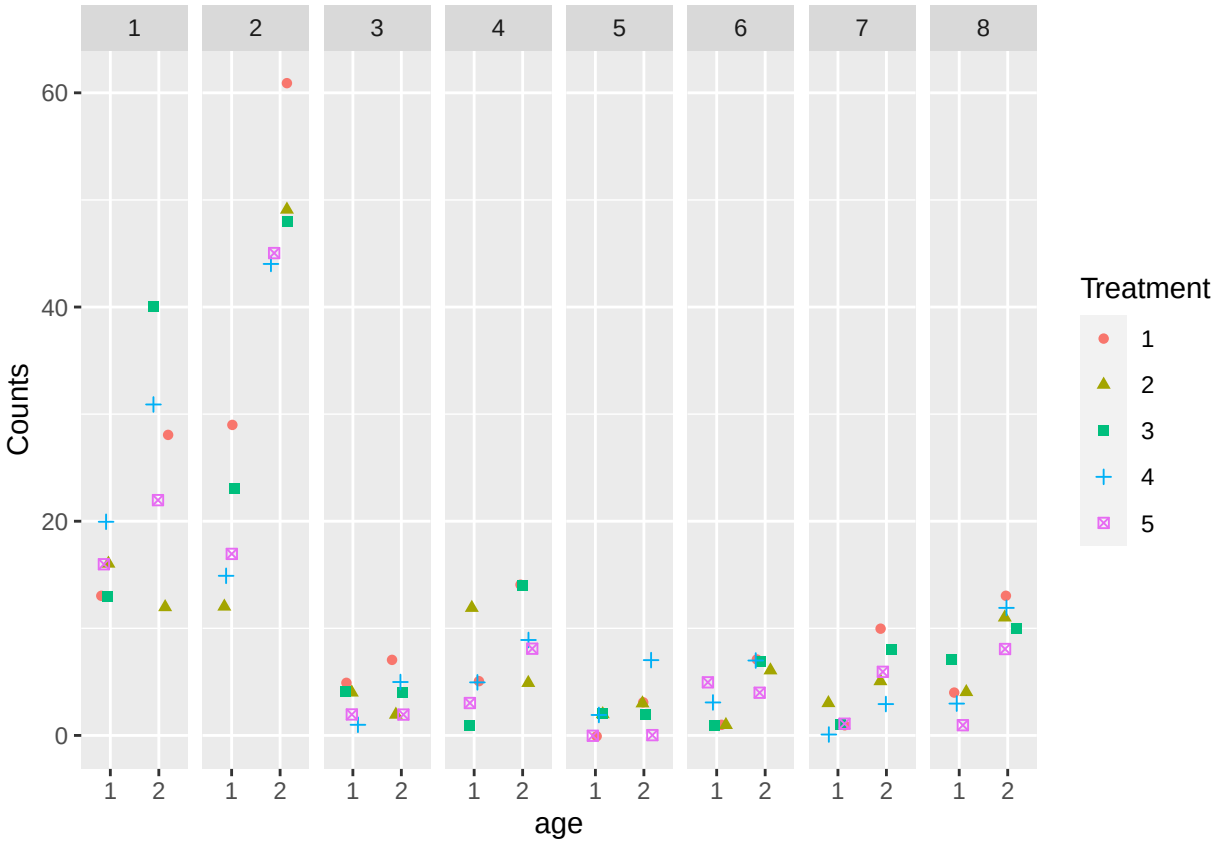
Block	age	Treatment	Counts
4	1	1	5
5	1	1	0
6	1	1	1

```
#check summary
kable(summary(dat.3.e), caption="Summary")
```

Table 11: Summary

Block	age	Treatment	Counts
1 :10	1:40	Length:80	Min. : 0.00
2 :10	2:40	Class :character	1st Qu.: 2.75
3 :10	NA	Mode :character	Median : 5.50
4 :10	NA	NA	Mean :10.50
5 :10	NA	NA	3rd Qu.:13.00
6 :10	NA	NA	Max. :61.00
(Other):20	NA	NA	NA

```
#plot
dat.3.e %>% ggplot(aes(x=age, y=Counts,
           group=Treatment,
           shape=Treatment,
           color=Treatment)) +
           geom_point(position=position_jitter(h=0.1, w=0.2)) +
           facet_grid(~Block)
```


1
2
3
4
5
6
7
8
60
40
20
0
Counts
1
2
age
Treatment
1
2
3
4
5