Homework2_zhiyuan

Zhiyuan Du

9/8/2020

Problem 3

Vision control is thing we cannot avoid in our onw lives even if we do not know it.

1. I would sue Github as my vision control and use branches to work on longer-term tasks and then merge the branch back into the main line when it's done.
2. I will compare whole sets of files to other branches or to past revisions to see what's different.
3. In the real work, I will track work over time through the vision control.

Problem 4

a. First, let's tidy the sensory data from five operators.

```
# getting the data: "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
url= "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
operator_data_raw=read.table(url,header = FALSE,skip=1,fill=TRUE,stringsAsFactors = FALSE)
saveRDS(operator_data_raw,"operator_data_raw.RDS")
operator_data_raw=readRDS("operator_data_raw.RDS")
```

First, let's tidy the data by the baseR, there are NA values in the raw data, where should be the values under the number of the item, so let's fill those NA values. Also, "item" and "operator" are categorical variables, we shuold finally make the data into a dataframe.

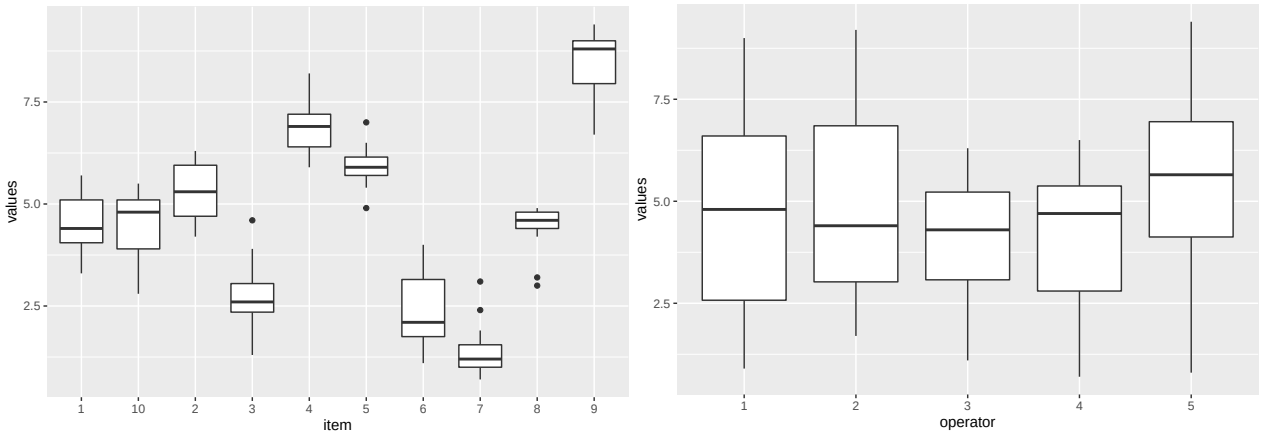
```
# use the for circle to fill the NA values and correct the item number.
for(i in 2:length(operator_data_raw$V6))
  {
  if(is.na(operator_data_raw$V6[i])){operator_data_raw$V6[i]=operator_data_raw$V1[i]
                                     operator_data_raw$V1[i]=operator_data_raw$V1[i-1]}
}
operator_data_tidy_bR=operator_data_raw[-1,]
operator_data_tidy_bR=data.frame(operator_data_tidy_bR$V1,
                                 operator=as.character(rep(c(1,2,3,4,5),2)),
                                 sapply(stack(operator_data_tidy_bR[,-1]),as.numeric))
operator_data_tidy_bR=operator_data_tidy_bR[,-4]
colnames(operator_data_tidy_bR)=c("item","operator","values")
```

The operator data has been tidyed by the baseR method. Let's do a summary of the data.

item	operator	values
Length:150	Length:150	Min. :0.700

item	operator	values
Class :character	Class :character	1st Qu.:3.025
Mode :character	Mode :character	Median :4.700
NA	NA	Mean :4.657
NA	NA	3rd Qu.:6.000
NA	NA	Max. :9.400

There are two categorical variables in the data, let's try to make two boxplots for the data:



From the plots, we can apparently tell the difference between them. The values disperse more when catgorized by item then operator.

Now, let's tidy the data by the tidyverse:

```
operator_data_tidy_tv=operator_data_raw[-1,]%>%
   mutate(V7=c(4.4,4.3,4.1,4.7,4.9,6.0,2.4,3.9,1.9,6.0,7.1,6.4,5.9
               ,5.8,5.8,1.7,3.0,2.1,0.7,1.3,0.9
               ,3.2,3.0,4.8,8.8,9.0,8.9,3.8,5.4,2.8))%>%
   mutate(operator=rep(c(1:5),6))%>%
   mutate(item=rep(c(1:10),each=3))%>%
   select(item,operator,V2:V5,V7)%>%
   gather(key="V9",value = "values",V2:V5,V7)%>%
   select(-V9)
```

---

b. Next, let's tidy the data of gold Medal performance for Olympic Men's Long Jump, year.

```
# getting the data: "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
url2= "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
longjump_data_raw=fread(url2)
saveRDS(longjump_data_raw,"longjump_data_raw.RDS")
longjump_data_raw=readRDS("longjump_data_raw.RDS")
```

First, let's tidy the data by the baseR, we need to put change the dimension and put the values in the right place.

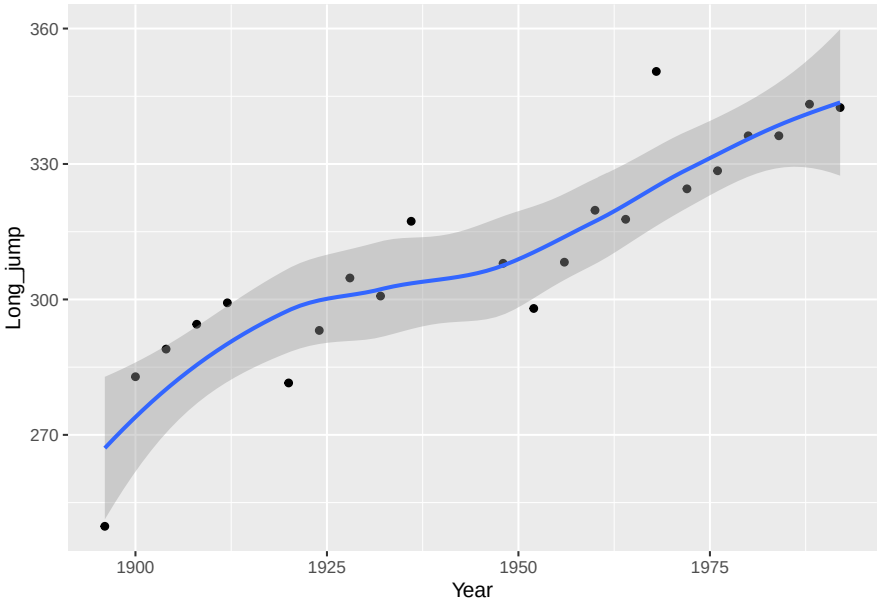
```
c1=data.frame(longjump_data_raw[,1]+1900,longjump_data_raw[,2])
c2=data.frame(longjump_data_raw[,3]+1900,longjump_data_raw[,4])
c3=data.frame(longjump_data_raw[,5]+1900,longjump_data_raw[,6])
c4=data.frame(longjump_data_raw[,7]+1900,longjump_data_raw[,8])
colnames(c2)=c("Year","Long")
```

```
colnames(c3)=c("Year","Long")
colnames(c4)=c("Year","Long")
longjump_data_tidy_bR=rbind(c1,c2,c3,c4)
longjump_data_tidy_bR=na.omit(longjump_data_tidy_bR)
colnames(longjump_data_tidy_bR)=c("Year","Long_jump")
```

The longjump data has been tidyed by the baseR method. Let's do a summary table of the 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

We assume that the long jump values would increase by years, let's make a scatter plot to have a check.



From the plot, we can apparently tell the line relationship between year and longjump, which means the long jump records are increasing by years.

Now, let's tidy the data by the tidyverse:

```
longjump_data_tidy_tv1=colnames(longjump_data_raw)=
c("col1","col2","col3","col4","col5","col6","col7",
"col8","col9","col10","col11","col12")
longjump_data_tidy_tv1=longjump_data_raw
longjump_data_tidy_tv1=longjump_data_raw%>%
        gather(key="col14",value="Long_jump",2,4,6,8)%>%
        select(Long_jump)
longjump_data_tidy_tv2=longjump_data_raw%>%
        gather(key="col13",value = "Year1",1,3,5,7)%>%
        mutate(Year=Year1+1900)%>%
        select(Year)
```

```
longjump_data_tidy_tv=longjump_data_tidy_tv2%>%
        bind_cols(longjump_data_tidy_tv1)%>%
        drop_na()
```

---

c. let's tidy the data of brain weight (g) and body weight (kg) for 62 species.

```
# getting the data: "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
url= "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
body_data_raw=read.table(url,header = FALSE,fill=TRUE,stringsAsFactors = FALSE)
saveRDS(body_data_raw,"body_data_raw.RDS")
body_data_raw=readRDS("body_data_raw.RDS")
```

First, let's tidy the data by the baseR, we need to put change the dimension and put the values in the right place.

```
d1=data.frame(body_data_raw[-1,1],body_data_raw[-1,2])
d2=data.frame(body_data_raw[-1,3],body_data_raw[-1,4])
d3=data.frame(body_data_raw[-1,5],body_data_raw[-1,6])

colnames(d1)=c("Body_weight","Brain_weight")
colnames(d2)=c("Body_weight","Brain_weight")
colnames(d3)=c("Body_weight","Brain_weight")
body_data_tidy_bR=rbind(d1,d2,d3)
body_data_tidy_bR=body_data_tidy_bR[-63,]
body_data_tidy_bR <- as.data.frame(sapply(body_data_tidy_bR, as.numeric))
```

Let's do a summary table for the data:

Body_weight	Brain_weight
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

After reading the data, we assume that there is positive relationship between the body weight and brain weight. Let's have a check by a scatter plot:

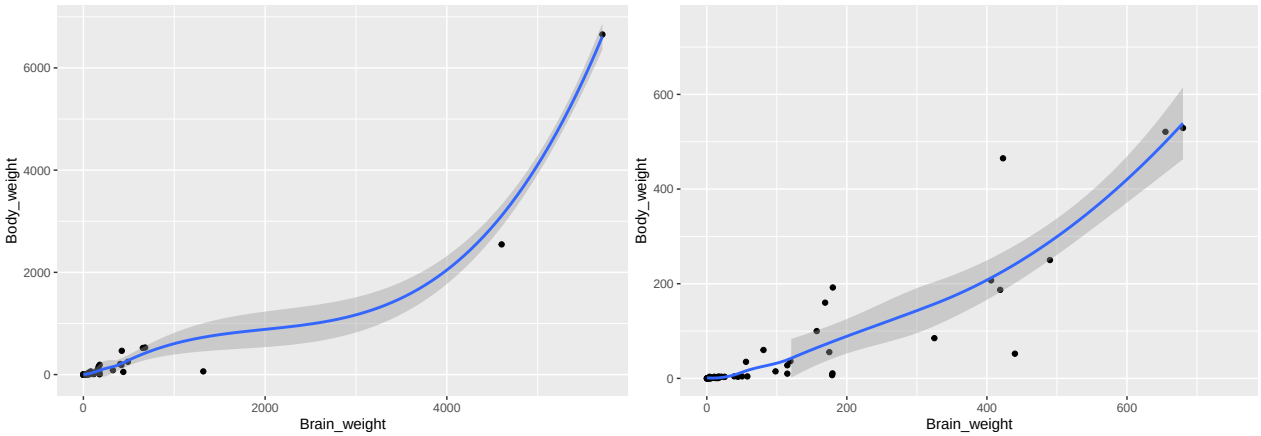


Figure 1: the plots for all species and removing the outliers

From the plot, there is a positive relationship between the body weight and brain weight, which means the species have more brain weight with more brain weight.

Now, let's tidy the data by the tidyverse:

```
body_data_tidy_tv=body_data_raw[-1,]%>%
        select(1,2,3,4,5,6)%>%
        mutate_all(as.numeric)%>%
        gather(key="d4",value = "Body_weight",1,3,5)%>%
        gather(key="d5",value = "Brain_weight",1,2,3)%>%
        select(Body_weight,Brain_weight)%>%
        slice(1:62)
```

---

d. Let's tidy the data of triplicate measurements of tomato yield for two varieties of tomatos at three planting densities.

```
# getting the data: "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"
url3= "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"
tomato_data_raw=fread(url3)
saveRDS(tomato_data_raw,"tomato_data_raw.RDS")
tomato_data_raw=readRDS("tomato_data_raw.RDS")
```

First, let's tidy the data by the baseR. The raw data is a table, we need to reshape it into a data that we can use to analyze, which means the each variable for one column.

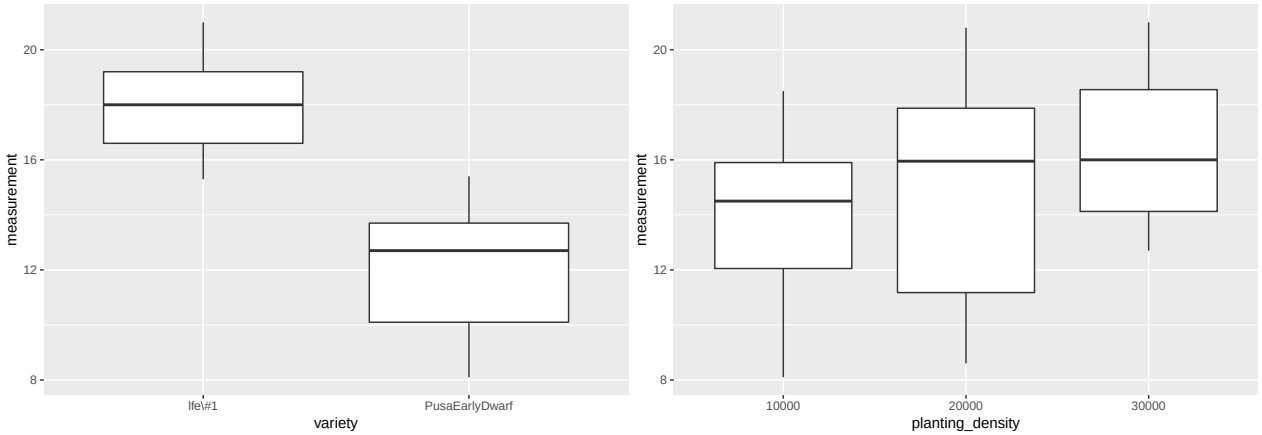
```
tomato_tidy_base_bR=separate(tomato_data_raw,"10000",into = c("100001","100002","100003")
                             ,sep = ",",convert = TRUE)
tomato_tidy_base_bR=separate(tomato_tidy_base_bR,"20000",into = c("200001","200002","200003")
                             ,sep = ",",convert = TRUE)
tomato_tidy_base_bR=separate(tomato_tidy_base_bR,"30000",into = c("300001","300002","300003")
                             ,sep = ",",convert = TRUE)

variety=rep(c("lfe\\#1","PusaEarlyDwarf"),9)
planting_density=rep(c("10000","10000","10000","20000","20000","20000","30000","30000","30000"),2)
measurement=stack(tomato_tidy_base_bR)[c(-1,-2),1]
tomato_tidy_base_bR=data.frame(variety,planting_density,measurement=sapply(measurement, as.numeric))
```

Let's make a summary table for the data:

variety	planting_density	measurement
Length:18	Length:18	Min. : 8.10
Class :character	Class :character	1st Qu.:12.95
Mode :character	Mode :character	Median :15.35
NA	NA	Mean :15.07
NA	NA	3rd Qu.:17.88
NA	NA	Max. :21.00

There are two categorical variables in the data, so let's make two boxplots:



Now, let's tidy the data by the tidyverse:

```
tomato_tidy_tv= tomato_data_raw%>%
separate("10000", into = c("100001","100002","100003"),sep = ",",convert = TRUE)%>%
separate("20000",into = c("200001","200002","200003"),sep = ",",convert = TRUE)%>%
separate("30000",into = c("300001","300002","300003"),sep = ",",convert = TRUE)%>%
gather(key="planting_density",value = "measurement",-1)%>%
rename("variety"="V1")%>%
separate(planting_density,into = c("planting_density","n"),sep=-1,remove=TRUE)%>%
select(variety,planting_density,measurement)
```