02_data_munging_summarizing_R

Assignment

Shreya Chandrasekharan

Problem 4

Version Control

1. Version control proves to be highly useful backing work up, forking and rewinding, and for collaborating.

2. Even for the solo analyst, it comes in handy while backing up in different physical locations, while also accurately showing changes in subsequent backups. Experimenting with data becomes an easier task since we can also revert to the last version of our work (branching).

3. Version control immensely helps in perfecting reproducible research.

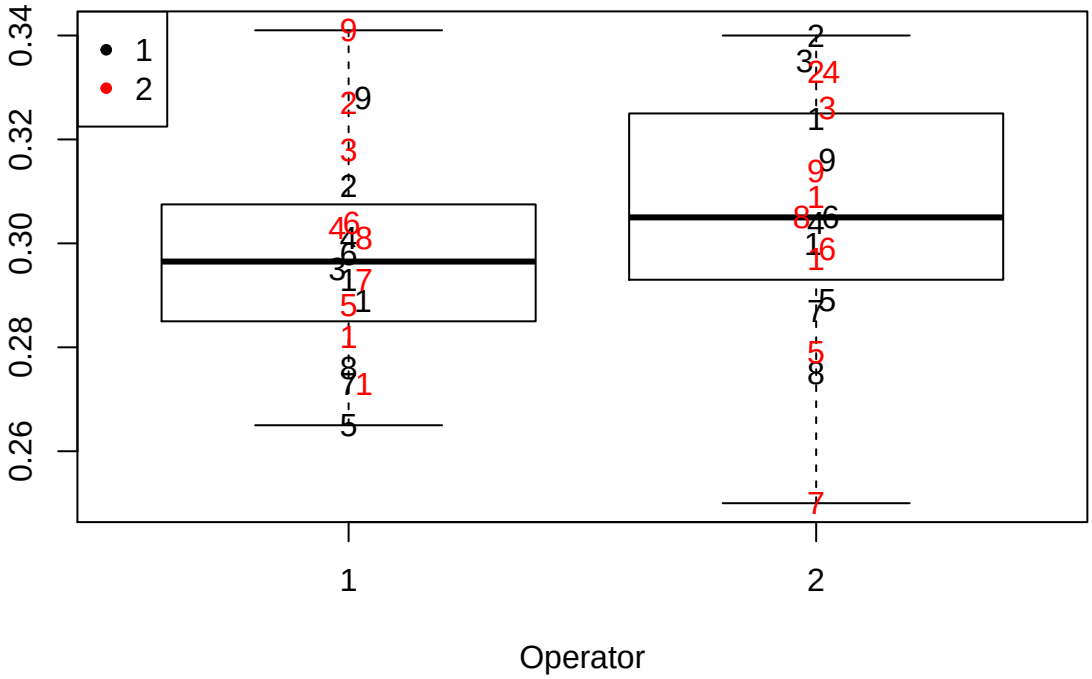
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:magrittr':
##
##     extract

##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2. http://CRAN.R-project.org/package=stargazer

##
## Attaching package: 'data.table'

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

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
## The following object is masked from 'package:base':
##
##     date
```

Problem 5

Table 1: CMM data summary

part	operator	replicate	value
Min. : 1.0	Length:40	Min. :1.0	Min. :0.2500
1st Qu.: 3.0	Class :character	1st Qu.:1.0	1st Qu.:0.2888
Median : 5.5	Mode :character	Median :1.5	Median :0.3010
Mean : 5.5	NA	Mean :1.5	Mean :0.3020
3rd Qu.: 8.0	NA	3rd Qu.:2.0	3rd Qu.:0.3165
Max. :10.0	NA	Max. :2.0	Max. :0.3410



Part (a): Sensory data

```
############################
#Problem5_Sensory_analysis
#get data
############################
url<-"http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
Sensory_raw<-read.table(url, header=F, skip=1, fill=T, stringsAsFactors = F)
Sensory_tidy<-Sensory_raw[-1,]
Sensory_tidy_a<-filter(.data = Sensory_tidy,V1 %in% 1:10) %>%
                rename(Item=V1,V1=V2,V2=V3,V3=V4,V4=V5,V5=V6)
Sensory_tidy_b<-filter(.data = Sensory_tidy,!(V1 %in% 1:10)) %>%
                mutate(Item=rep(as.character(1:10),each=2)) %>%
                mutate(V1=as.numeric(V1)) %>%
                select(c(Item,V1:V5))
Sensory_tidy<-bind_rows(Sensory_tidy_a,Sensory_tidy_b)
colnames(Sensory_tidy)<-c("Item",paste("Person",1:5,sep="_"))
Sensory_tidy<-Sensory_tidy %>%
    gather(Person,value,Person_1:Person_5) %>%
    mutate(Person = gsub("Person_","",Person)) %>%
    arrange(Item)

############################


knitr::kable(summary(Sensory_tidy), caption="Sensory data summary")
```

Table 2: Sensory data summary

Item	Person	value
Length:150	Length:150	Min. :0.700
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

Part (b): Long Jump data

```
############################
#Problem5_LongJump_analysis
#get data
############################
url<-"http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
LongJump_raw<-read.table(url, header=F, skip=1, fill=T, stringsAsFactors = F)
colnames(LongJump_raw)<-rep(c("V1","V2"),4)
LongJump_tidy<-rbind(LongJump_raw[,1:2],LongJump_raw[,3:4],
                     LongJump_raw[,5:6],LongJump_raw[,7:8])
LongJump_tidy<-LongJump_tidy %>%
    filter(!(is.na(V1))) %>%
    mutate(YearCode=V1, Year=V1+1900, dist=V2) %>%
    select(-V1,-V2)
```

```
############################


knitr::kable(summary(LongJump_tidy), caption="Long Jump data summary")
```

Table 3: Long Jump data summary

YearCode	Year	dist
Min. :-4.00	Min. :1896	Min. :249.8
1st Qu.:21.00	1st Qu.:1921	1st Qu.:295.4
Median :50.00	Median :1950	Median :308.1
Mean :45.45	Mean :1945	Mean :310.3
3rd Qu.:71.00	3rd Qu.:1971	3rd Qu.:327.5
Max. :92.00	Max. :1992	Max. :350.5

Part (c): Brain vs Body data

```
############################
#Problem5_BrainBody_analysis
#get data
############################
url<-"http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
BrainBody_raw<-read.table(url, header=F, skip=1, fill=T, stringsAsFactors = F)
colnames(BrainBody_raw)<-rep(c("Brain","Body"),3)
BrainBody_tidy<-rbind(BrainBody_raw[,1:2],BrainBody_raw[,3:4],
                      BrainBody_raw[,5:6])
BrainBody_tidy<-BrainBody_tidy %>%
    filter(!(is.na(Brain)))

############################


knitr::kable(summary(BrainBody_tidy), caption="Brain/Body weight data summary")
```

Table 4: Brain/Body weight data summary

Brain	Body
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.: 48.203	3rd Qu.: 166.00
Max. :6654.000	Max. :5712.00

Part (d): Tomato data

```
############################
#Problem5_Tomato_analysis
#get data
############################
url<-"http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"
```

```
    Tomato_raw<-read.table(url, header=F, skip=2, fill=T, stringsAsFactors = F, comment.char = "")
    Tomato_tidy<-Tomato_raw %>%
        separate(V2,into=paste("C10000",1:3,sep="_"),sep=",",remove=T, extra="merge") %>%
        separate(V3,into=paste("C20000",1:3,sep="_"),sep=",",remove=T, extra="merge") %>%
        separate(V4,into=paste("C30000",1:3,sep="_"),sep=",",remove=T, extra="merge") %>%
        mutate(C10000_3=gsub(",","",C10000_3)) %>%
        gather(Clone,value,C10000_1:C30000_3) %>%
        mutate(Variety=V1, Clone=gsub("C","",Clone)) %>%
        mutate(Variety=gsub("\\\\#"," ",Variety)) %>%
        separate(Clone,into = c("Clone","Replicate")) %>%
        select(-V1,Variety,Clone,value) %>%
        arrange(Variety)

    ############################


    knitr::kable(summary(Tomato_tidy), caption="Tomato data summary")
```

Table 5: Tomato data summary

Clone	Replicate	value	Variety
Length:18	Length:18	Length:18	Length:18
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

Problem 6

```
# Path to data
.datapath <- file.path(path.package('swirl'), 'Courses',
                       'R_Programming_E', 'Looking_at_Data',
                       'plant-data.txt')
# Read in data
plants <- read.csv(.datapath, strip.white=TRUE, na.strings="")

# Remove annoying columns
.cols2rm <- c('Accepted.Symbol', 'Synonym.Symbol')
plants <- plants[, !(names(plants) %in% .cols2rm)]

# Make names pretty
names(plants) <- c('Scientific_Name', 'Duration', 'Active_Growth_Period',
                   'Foliage_Color', 'pH_Min', 'pH_Max',
                   'Precip_Min', 'Precip_Max',
                   'Shade_Tolerance', 'Temp_Min_F')
```

Linear Model

```
my_lm <- lm(formula = Foliage_Color~pH_Min + pH_Max, data = plants)
my_lm
```

```
my_lm_anova <- anova(my_lm)
my_lm_anova
```

Comment: There is an error in the above functions. 'Foliage_Color' is a Qualitative Attribute, whereas 'pH_Min' and 'pH_Max' are Numeric Variables. I was unable to test for a linear relationship between 'Foliage_Color' and a statistic that combines information in 'pH_Min' and 'pH_Max' or get ANOVA results using my current statistical knowledge.

Problem 7

```
Car_Gebreken_raw <- read.csv("Open_Data_RDW__Gebreken.csv",stringsAsFactors = F,
                             nrows=200, header=T,quote = '"')
Car_Geconstat_raw <- read.csv("Open_Data_RDW__Geconstateerde_Gebreken.csv",
                              stringsAsFactors = F, nrows=200, header=T)
Car_Person_raw <- read.csv("Personenauto_basisdata.csv",stringsAsFactors = F,
                           nrows=200, header=T)

Car_Gebreken_raw.colclass <- sapply(Car_Gebreken_raw,class)
Car_Geconstat_raw.colclass <- sapply(Car_Geconstat_raw,class)
Car_Person_raw.colclass <- sapply(Car_Person_raw,class)

print("Gebreken")
print(Car_Gebreken_raw.colclass)

print("Geconstat")
print(Car_Geconstat_raw.colclass)

print("Personen")
print(Car_Person_raw.colclass)

#this had the defect code and description
 Car_Gebreken_select <- fread(input = "Open_Data_RDW__Gebreken.csv",
                              header = T, select=c(1,6), showProgress=F)
#this has the license plate, inspection date and defect code
Car_Geconstat_select <- fread(input = "Open_Data_RDW__Geconstateerde_Gebreken.csv",

                              header=T, select=c(1,3,5),showProgress=F)
#this has the license plate, make and model of vehicle
Car_Person_select <- fread(input = "Personenauto_basisdata.csv",
                           header=T, showProgress = F, select = c(1,3,4))

Car_Geconstat_select_2017 <-
  Car_Geconstat_select[grep("2017",
                            Car_Geconstat_select$"Meld datum door keuringsinstantie")]
############################
```

```
merge_License_plate <- merge(Car_Gebreken_select, Car_Geconstat_select,
                             Car_Person_select, by="Kenteken")
merge_Defect_code <- merge(Car_Gebreken_select,
                           Car_Geconstat_select, Car_Person_select,
                           by="Gebrek identificatie")
```

Comment: I have been unable to complete the tasks required under this problem. This goes on to show that

I require more practice of R codes and also better online search skills for when I get stuck in problems.
I hope to achieve computational efficiency with time, and be able to satisfactorily complete homework assignments at the very least. However, I aspire to be able to use the programming skills I am learning through this course for productive research work in the long run.