HW2_Do_Quyen

Quyen Do

August 29, 2018

Problem 1

Work through the "R Programming E" lesson parts 4-7, 14 (optional 12 - only takes 5 min) and "Getting and Cleaning Data" *swirl* lessons parts 1-4.

From the R command prompt:
install.packages("swirl")
library(swirl)
install_course("Getting_and_Cleaning_Data")
swirl()

Problem 2

Read through the Git help Chapters 1 and 2. https://git-scm.com/book/en/v2

Part A: setup Github

In Github, you will want to "fork" my class repo. Search for STAT_5014. Towards the right top of the page, you will see a little icon labeled "Fork". Click on this to create a linked copy of my repo in your GitHub repo set. You should now be in your Git repo set. Look at the repo name towards the top left, it should be /STAT_5014. IF so, click on the clone or download button to the middle right. Copy the https address which should look like https://github.com/_your username_/STAT_5014.git . MAKE sure the link has YOUR user name.

Part B: ssh key

Before continuing, if you didn't set up the SSH key in the last homework, do so now.

Part C: setup Rproject

In Rstudio, create a new Rproject using version control.

1. File –> New Project –> Version Control –> Git
2. In the Repository URL box, past the https address from Part A.
3. In the Project directory name box, type STAT_5015_homework
4. In the Create project as subdirectory, browse to where you want your homework files to live.

Part D: !!*Do this to get updates*!!

In the directory you put the project file, do:

1. git remote add upstream https://github.com/rsettlage/STAT_5014.git

2. git checkout master
3. git pull upstream master

Problem 3

In the last problem, you forked my class repo and then "cloned" it to your local hard drive. If you explore the files that were pulled down, you will notice that there is a subdirectory called "02_data_munging_summarizing_R".

Create a new R Markdown file within the project folder within the "02_data_munging_summarizing_R" subfolder (file–>new–>R Markdown–>save as.

The filename should be: HW2_lastname_firstname, i.e. for me it would be HW2_Settlage_Bob

You will use this new R Markdown file to solve problems 4-7.

Problem 4

Version control helps me manage and control my coding assignments. If I treat each assignment as a project. Version control helps me back up and save the progress of the assignment as I'm completing it. It is a way for me to keep the current codes while enabling backtracking to older versions if my new codes break. In a long run, version control helps me retain my codes, i.e. my work throught the course, so that I can always look back at what I did and got reminded of what I have learned.

Problem 5

In this exercise, you will import, munge, clean and summarize datasets from Wu and Hamada's *Experiments: Planning, Design and Analysis* book you will use in the Spring. For each one, please weave your code and text to describe both your process and observations. Make sure you create a tidy dataset describing the variables, create a summary table of the data, note issues with the data.

a. Sensory data from five operators.
   http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat

```
#Import raw data from url
url <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
sensory_raw <- read_csv(url,skip=1)
```

```
## Parsed with column specification:
## cols(
##   `Item 1 2 3 4 5` = col_character()
## )
```

```
#Set up partial_cleaned table (as recorded by the researcher)
columns_names <- c("Item","Operator 1","Operator 2", "Operator 3", "Operator 4","Operator 5")
sensory_partial_cleaned <- data.frame(matrix(nrow=nrow(sensory_raw),ncol=length(columns_names)))
colnames(sensory_partial_cleaned) <- columns_names

##To keep track of the current item the row is referred to
current_item <- 0

for (i in 1:nrow(sensory_raw))
{
  row <- sensory_raw[i,][[1]]
  #Split the row by space
```

```
  row_data <- strsplit(row,split=" ")[[1]]
  row_data <- as.numeric(row_data)

  # If row_data contain 6 figures, the first figure must be the item number of that row and the next fe
  if (length(row_data) == 6){
    current_item <- row_data[1]

    #Get rid of item number from the row_data
    row_data <- row_data[-1]
  }
  sensory_partial_cleaned[i,] <- c(current_item,row_data)
}

#Turn sensory_partial_cleaned to a cleaned table
sensory_cleaned <- gather(sensory_partial_cleaned,key="Operator",value="Value",
                          "Operator 1","Operator 2","Operator 3","Operator 4","Operator 5")
#Create summary table
kable(summary(sensory_cleaned),caption="Sensory Data Summary")
```

Table 1: Sensory Data Summary

Item	Operator	Value
Min. : 1.0	Length:150	Min. :0.700
1st Qu.: 3.0	Class :character	1st Qu.:3.025
Median : 5.5	Mode :character	Median :4.700
Mean : 5.5	NA	Mean :4.657
3rd Qu.: 8.0	NA	3rd Qu.:6.000
Max. :10.0	NA	Max. :9.400

b. Gold Medal performance for Olympic Men's Long Jump, year is coded as 1900=0.
http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat

```
url <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
medal_raw <- read.csv(url,header=FALSE, sep=" ",skip=1)
names(medal_raw) <- c("Year 1","LongJump 1", "Year 2","LongJump 2","Year 3", "LongJump 3","Year 4","Long

#medal_raw table is messy for the fact there are 8 columns depicting 2 variables. The 4 Year columns ne

#Create a clean dataframe with 2 variables
medal_cleaned <- data.frame(matrix(nrow=0,ncol=2))

#Subsequently extract Year and corresponding LongJump columns from raw table and bind them onto the cle
medal_cleaned <- rbind(medal_cleaned,medal_raw[,c("Year 1","LongJump 1")])

names(medal_cleaned) <- c("Year 2","LongJump 2")
medal_cleaned <- rbind(medal_cleaned,medal_raw[,c("Year 2","LongJump 2")])

names(medal_cleaned) <- c("Year 3","LongJump 3")
medal_cleaned <- rbind(medal_cleaned,medal_raw[,c("Year 3","LongJump 3")])

names(medal_cleaned) <- c("Year 4","LongJump 4")
medal_cleaned <- rbind(medal_cleaned,medal_raw[,c("Year 4","LongJump 4")])
```

```
#Format clean table
names(medal_cleaned) <- c("Year","Long Jump")

#Remove NA row
medal_cleaned<- medal_cleaned[-c(23,24),]

#Create a summary table
kable(summary(medal_cleaned),caption="Long Jump Gold Medal Data Summary")
```

Table 2: Long Jump Gold Medal Data Summary

Year	Long Jump
Min. :-4.00	Min. :249.8
1st Qu.:21.00	1st Qu.:295.4
Median :50.00	Median :308.1
Mean :45.45	Mean :310.3
3rd Qu.:71.00	3rd Qu.:327.5
Max. :92.00	Max. :350.5

c. Brain weight (g) and body weight (kg) for 62 species.
http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat

```
#Import data into a raw table
url <- "http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat"
bodyBrain_raw <- read.csv(url,sep=" ",header=FALSE,skip=1)
names(bodyBrain_raw) <- c("Body Wt 1","Brain Wt 1","Body Wt 2","Brain Wt 2","Body Wt 3","Brain Wt 3")

#bodyBrain_raw table is messy for the fact there are 6 columns depicting 2 variables. The 3 Body Wt col

#Create a clean dataframe with 2 variables
bodyBrain_cleaned <- data.frame(matrix(nrow=0,ncol=2))

#Subsequently extract Brain Wt and corresponding Body Wt columns from raw table and bind them onto the
bodyBrain_cleaned <- rbind(bodyBrain_cleaned,bodyBrain_raw[,c("Body Wt 1","Brain Wt 1")])

names(bodyBrain_cleaned) <- c("Body Wt 2","Brain Wt 2")
bodyBrain_cleaned <- rbind(bodyBrain_cleaned,bodyBrain_raw[,c("Body Wt 2","Brain Wt 2")])

names(bodyBrain_cleaned) <- c("Body Wt 3","Brain Wt 3")
bodyBrain_cleaned <- rbind(bodyBrain_cleaned,bodyBrain_raw[,c("Body Wt 3","Brain Wt 3")])

#Format clean table
names(bodyBrain_cleaned) <- c("Body Wt","Brain Wt")

#Remove NA row
bodyBrain_cleaned<- bodyBrain_cleaned[-63,]
```

d. Triplicate measurements of tomato yield for two varieties of tomatos at three planting densities.
http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat

```
url <- "https://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat"

tomato_raw <- read_lines(url)[-c(1,2)]
```

```
#Set up cleaned table
columns_names <- c("Variety","Density","Yield")
tomoto_cleaned <- data.frame(matrix(nrow=0,ncol=length(columns_names)))
added_row <- data.frame(matrix(nrow=1,ncol=length(columns_names)))
colnames(added_row) <- columns_names
colnames(tomoto_cleaned) <- columns_names

#Run through each line from the data to extract information
for (line in tomato_raw)
{
  #split each line by white space character
  line <- strsplit(line,split=" ")[[1]]

  #get rid of white space elements
  line <- line[line!=""]

  #extract the information
  variety <- line[1]
  density_10000<- as.numeric(strsplit(line[2],split=",")[[1]])
  density_20000 <- as.numeric(strsplit(line[3],split=",")[[1]])
  density_30000 <- as.numeric(strsplit(line[4],split=",")[[1]])

  for (i in density_10000)
  {
    added_row$Variety <- variety
    added_row$Density <- 10000
    added_row$Yield <- i
    tomoto_cleaned <- rbind(tomoto_cleaned,added_row)
  }

  for (i in density_20000)
  {
    added_row$Variety <- variety
    added_row$Density <- 20000
    added_row$Yield <- i
    tomoto_cleaned <- rbind(tomoto_cleaned,added_row)
  }

  for (i in density_30000)
  {
    added_row$Variety <- variety
    added_row$Density <- 30000
    added_row$Yield <- i
    tomoto_cleaned <- rbind(tomoto_cleaned,added_row)
  }
}
```

Problem 6

In the swirl lessons, you played with a dataset "plants". Our ultimate goal is to see if there is a relationship between pH and Foliage_Color. Consider a statistic that combines the information in pH_Min and pH_Max. Clean, summarize and transform the data as appropriate. Use function *lm* to test for a relationship. Report

both the coefficients and ANOVA results in table form.

Note that if you didn't just do the swirl lesson, it is now not available. Add the following code to your project to retrieve it.

```
# Path to data
library(swirl)
.datapath <- file.path(path.package('swirl'), 'Courses',
                       'R_Programming', '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')

# Consider the range of pH
plants$pH_range <- plants$pH_Max-plants$pH_Min
plants_working <- plants[,c('Foliage_Color','pH_range')]
plants_working$Foliage_Color <- as.factor(plants_working$Foliage_Color)
```

Problem 8

Finish this homework by pushing your changes to your repo and submitting me a pull request. In general, your workflow for this should be:

1. In R: pull (Git tab, down arrow) – to make sure you have the most recent repo
2. In R: do some work
3. In R: check files you want to commit
4. In R: commit, make message INFORMATIVE and USEFUL
5. In R: push – this pushes your local changes to the repo
6. In Github: submit a pull request – this tells me you are wanting me to pull in your changes to my master repo

If you have difficulty with steps 1-5, git is not correctly or completely setup. The above will pull from your repo, so does not include anything to get from MY repo, ie nothing new will show up.

To get stuff from my repo which you then can push to your repo, modify the above to be:

1. In R: do some work
2. At command prompt: git pull upstream master – to make sure you have the most recent repo
3. In R: check files you want to commit (this MAY include files I added/changed)

4. In R: commit, make message INFORMATIVE and USEFUL

5. In R: push – this pushes OUR local changes to YOUR repo

6. In Github: submit a pull request – this tells me you are wanting me to pull in your changes to my master repo

Only submit the .Rmd and .pdf solution files. Names should be formatted HW2_lastname_firstname.Rmd and HW2_lastname_firstname.pdf

Optional preperation for next class:

Next week we will talk about R logic and good programming practices. If you have time and are interested, please read:

Google's R Style Guide: https://google.github.io/styleguide/Rguide.xml
Hadley Wickam's R Style Guide: http://r-pkgs.had.co.nz/style.html