Homework 2

Due September 12, 2019

Eric Bae

2019-09-06

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about Reproducible Research, R and version control, getting, cleaning and munging data and finally, summarizing data. Again, we are focusing on Reproducible Analysis which, for us, is accomplished by mixing code, figures and text into a cohesive document that fully describes both the process we took to go from data to results and the rational behind our data driven conclusions. This week we begin creating tidy data sets. While others have proposed standards for sharing data with statiticians, as practicing data scientists, we realize the often onerous task of getting, cleaning and formatting data is usually in our hands. From here on out, we will use GitHub to retrieve and turn in the homework assignments.

Problem 1

Work through the "R Programming E" lesson parts 4-7, 14 (optional 12 - only takes 5 min).

From the R command prompt:

```
#install.packages("swirl")
library(swirl)
install_course("R_Programming_E")
swirl()
```

Problem 2

Create a new R Markdown file within your local GitHub repo folder (file–>new–>R Markdown–>save as).

The filename should be: HW2_lastname, i.e. for me it would be HW2_Settlage

You will use this new R Markdown file to solve problems 3-5.

Problem 3

In the lecture, there were two links to StackOverflow questions on why one should use version control. In your own words, summarize in 2-3 sentences how you think version control can help you in the classroom.

I think version control in this classroom is a good idea because I am prone to causing irreversible mistakes to my codes, and since we will continue to upload a large number of files to GitHub, which could lead to me losing track of some files. I have had situations where I accidentally overwrote a file against my intention. In fact, I already did with HW1, though that one was semi-intentional.

Problem 4

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
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

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

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

Problem 5

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
.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')
```

Problem 6

Finish this homework by pushing your changes to your repo. In general, your workflow for this should be:

1. git pull – to make sure you have the most recent repo
2. In R: do some work
3. git add – this tells git to track new files
4. git commit – make message INFORMATIVE and USEFUL
5. git push – this pushes your local changes to the repo

If you have difficulty with steps 1-5, git is not correctly or completely setup. See me for help.

Only submit the .Rmd and .pdf solution files. Names should be formatted HW2_lastname.Rmd and HW2_lastname.pdf

Optional preperation for next class:

TBD