Statistics 5014: Homework 2

Due Monday September 11, 10 am

2017-09-07

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) 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.

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

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.

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
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 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.

Problem 7

One common situation data scientists encounter is when data is spread across many data files. This can be that the data is simply split across data files OR different aspects of the data is in different data files. Here we will look at the second scenario: different aspects of a dataset are contained in different data files that need to be merged. In this case, we are going to munge some open data containing car records, reported defects, and defect descriptions. You should start this problem by looking at the help for variations on SQL like merge functions: merge, join, inner_join, left_join, right_join. As in the last problem, please create a tidy dataset, summarize and annotate the process, and report the indicated statistics as follows:

Personal car details: https://opendata.rdw.nl/api/views/qyrd-w56j/rows.csv?accessType=DOWNLOAD
Observed Defects: https://opendata.rdw.nl/api/views/a34c-vvps/rows.csv?accessType=DOWNLOAD
Defect Details: https://opendata.rdw.nl/api/views/hx2c-gt7k/rows.csv?accessType=DOWNLOAD

In this task, you should (suggested steps, not necessarily in order):

a. load all three datasets into R (consider saving, first two are ca. 1 GB)

b. merge/join the three datasets, by license plate, then by defect code

c. clean the data, remove NA, etc

d. report how many DIFFERENT makes and models of cars you end with (?unique ?distinct ?duplicated) considering only year 2017

e. report a table of the 5 most frequent defects (translated) and the top make/models having that defect (?count) again considering only year 2017

f. use function lm to test for a relationship between number of defects observed by make, report both the coefficient and anova tables (2017 only)

g. repeat (f) by model (2017 only)

h. comment on this workflow and how you might be more computationally efficient

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.

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