Statistics 5014: Homework 3

Due Monday September 11, 10 am

2017-09-12

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 R and version control, munging and 'tidying' data, good programming practice and finally some basic programming building blocs. To begin the homework, we will for the rest of the course, start by loading data and then creating tidy data sets.

Problem 1

Work through the "R Programming E" lesson parts 8 and 9.

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

Problem 2

In the last homework, you "forked" my repo, cloned it to your local computer, saved your homework in the "02_data_munging_summarizing_R" directory, pushed your homework to your remote repo to ultimately sent me a pull request. Here, we will follow the same workflow but put your homework in "03_good_programming_R_functions".

Create a new R Markdown file within the project folder within the "03_good_programming_R_functions" subfolder (file–>new–>R Markdown–>save as.

The filename should be: HW3_lastname_firstname, i.e. for me it would be HW3_Settlage_Bob

You will use this new R Markdown file to solve problems 4-10.

Problem 4

In the lecture, there were two links to programming style guides. What is your takeaway from this and what specifically are *you* going to do to improve your coding style?

Problem 5

Good programming practices start with this homework. In the last homework, you imported, munged, cleaned and summarized datasets from Wu and Hamada's *Experiments: Planning, Design and Analysis*. In this problem, please using *lintr* to lint your last homework (if you didn't do it, perhaps the time is now ;)). In my case, the command looks like this (takes a few moments to run):

```
    lint(filename = "./02_data_munging_summarizing_R_git/HW2_Settlage_Bob.Rmd")
```

Can you clean up your code to pass the major issues?? <— just a challenge, not part of the problem!!

Note that really all we have done is syntax checking and received a few stylistic suggestions. We COULD create a custom linter to check for indenting, etc. For now, I think it is enough to know there are a lot of opinions on what code should look like and working in teams requires you to code nicely!! So, clean up your code!!

From the messages, what are some things you need to change in your code?

Problem 6

A situation you may encounter is a data set where you need to create a summary statistic for each observation type. Sometimes, this type of redundancy is perfect for a function. Here, we need to create a single function to:

1. calculate the mean for dev1
2. calculate the mean for dev2
3. calculate the sd for dev1
4. calculate the sd for dev2
5. calculate the correlation between dev1 and dev2
6. return the above as a single data.frame

We will use this function to summarize a dataset which has multiple repeated measurements from two devices by thirteen Observers. In the current lecture directory, you will see a file "HW3_data.rds". Please load the file (?readRDS – really nice format for storing data objects), loop through the Observers collecting the summary statistics via your function for each Observer separately.

The output of this problem should be:

a. A single table of the means, sd, and correlation for each of the 13 Observers

b. A box plot of all the means to compare the spread of means from dev1 to dev2

c. A violin plot of all the sd to compare the spread of sd from dev1 to dev2

I will look at the code and comment on it, so make it NICE!!

Problem 7 – redo

Same as last time, please create and annotate the process to create a tidy dataset from http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BloodPressure.dat

Problem 8

Create a function to find solutions to (1) using Newton's method. The answer should include the solutions with tolerance used to terminate the loop and a plot showing the iterations on the path to the solution.

$$f(x) = 3^x - sin(x) + cos(5x) \tag{1}$$

Problem 9 – redo, make a good honest and professional attempt

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 10

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: git pull upstream – to make sure you have the most recent local 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 HW3_lastname_firstname.Rmd and HW3_lastname_firstname.pdf

Optional preperation for next class:

Next week we will talk about Exploratory Data Analysis and graphing. Swirl will be a bit part of this. Check out "Exploratory_Data_Analysis" Swirl lessons 1-10.