Statistics 5014: Homework 5

Due Tuesday October 3, 11am

2017-09-27

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 Exploratory Data Analysis (EDA) and plotting. In this homework, we will as usual, load, munge and create tidy data sets. In EDA, our goal is often to create informative (and perhaps pretty) plots showing features or perhaps deficiencies in the data. Perhaps more importantly, our goal is to find these deficiencies and explore relationships in the data. Efficiently.

Problem 1

Work through the Swirl "R_programming_E" lesson parts 10 and 11.

swirl()

Problem 2

As in the last homework, create a new R Markdown file within the project folder within the "05_R_apply_family" subfolder (file–>new–>R Markdown–>save as.

The filename should be: HW4_lastname_firstname, i.e. for me it would be HW5_Settlage_Bob

You will use this new R Markdown file to solve the following problems:

Problem 3

What are you thoughts for what makes a good figure?

Problem 4

a. Create a function that computes the proportion of successes in a vector. Use good programming practices.

b. Create a matrix to simulate 10 flips of a coin with varying degrees of "fairness" as follows:

```
set.seed(12345)
P4b_data <- matrix(rbinom(10, 1, prob = (30:40)/100), nrow = 10,
    ncol = 10)
```

c. Use your function in conjunction with apply to compute the proportion of success in P4b_data by column and then by row. What do you observe? What is going on?

d. You are to fix the above matrix by creating a function whose input is a probability and output is a vector whose elements are the outcomes of 10 flips of a coin. Now create a vector of the desired probabilities. Using the appropriate apply family function, create the matrix we really wanted above. Prove this has worked by using the function created in part a to compute and tabulate the appropriate marginal successes.

Problem 5

Load, munge, and explore the data given in Wu and Hamada from the starch experiment. Consider strength as the response. You do not need to form a model or otherwise analyze the dataset, you do need to explore the data, make any figures/tables necessary to make observations about the data, and generally annotate the process in text.

http://www2.isye.gatech.edu/~jeffwu/book/data/starch.dat

Problem 6

Our ultimate goal in this problem is to create an annotated map of the US. I am giving you the code to create said map, you will need to customize it to include the annotations.

Part a. Get and import a database of US cities and states. Here is some R code to help:

```
# we are grabbing a SQL set from here
# http://www.farinspace.com/wp-content/uploads/us_cities_and_states.zip

# download the files, looks like it is a .zip
library(downloader)
download("http://www.farinspace.com/wp-content/uploads/us_cities_and_states.zip",
    dest = "us_cities_states.zip")
unzip("us_cities_states.zip", exdir = "./")

# read in data, looks like sql dump, blah
library(data.table)
states <- fread(input = "./us_cities_and_states/states.sql",
    skip = 23, sep = "'", sep2 = ",", header = F, select = c(2,
        4))
### YOU do the CITIES I suggest the cities_extended.sql
### may have everything you need can you figure out how to
### limit this to the 50?
```

Part b. Create a summary table of the number of cities included by state.

Part c. Create a function that counts the number of occurances of a letter in a string. The input to the function should be "letter" and "state_name". The output should be a scalar with the count for that letter.

Create a for loop to loop through the state names imported in part a. Inside the for loop, use an apply family function to iterate across a vector of letters and collect the occurance count as a vector.

```
    ##pseudo code
    letter_count <- data.frame(matrix(NA,nrow=50, ncol=26))
    getCount <- function(what args){
        temp <- strsplit(state_name)
        # how to count??
        return(count)
    }
    for(i in 1:50){
        letter_count[i,] <- xx-apply(args)
    }
```

Part d.

Create 2 maps to finalize this. Map 1 should be colored by count of cities on our list within the state. Map 2 should highlight only those states that have more than 3 occurances of ANY letter in thier name.

Quick and not so dirty map:

```
# https://cran.r-project.org/web/packages/fiftystater/vignettes/fiftystater.html
library(ggplot2)
library(fiftystater)

data("fifty_states")  # this line is optional due to lazy data loading
crimes <- data.frame(state = tolower(rownames(USArrests)),
    USArrests)
# map_id creates the aesthetic mapping to the state name
# column in your data
p <- ggplot(crimes, aes(map_id = state)) + # map points to the fifty_states shape data
geom_map(aes(fill = Assault), map = fifty_states) + expand_limits(x = fifty_states$long,
    y = fifty_states$lat) + coord_map() + scale_x_continuous(breaks = NULL) +
    scale_y_continuous(breaks = NULL) + labs(x = "", y = "") +
    theme(legend.position = "bottom", panel.background = element_blank())

p
# ggsave(plot = p, file =
# 'HW5_Problem6_Plot_Settlage.pdf')
```

Problem 7

Push your homework and submit a pull request.

When it is time to submit, –ONLY– submit the .Rmd and .pdf solution files. Names should be formatted HW4_lastname_firstname.Rmd

Optional preperation for next class:

Next week we will talk about the dual handling of vectors and matrices in R. No swirl. :)