Statistics 5014: Homework 4

Due Wednesday September 19 in GitHub, 9 am

2018-09-12

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 Exploratory Data Analysis and plotting. To begin the homework, we will as usual, start by loading, munging and creating tidy data sets. In this homework, our goal is to create informative (and perhaps pretty) plots showing features or perhaps deficiencies in the data.

Problem 1

Work through the Swirl "Exploratory_Data_Analysis" lesson parts 1 - 10.

swirl()

Problem 2

As in the last homework, create a new R Markdown file within the project folder within the "04_projecting_knowledge_plots" subfolder (file–>new–>R Markdown–>save as.

The filename should be: HW4_lastname_firstname, i.e. for me it would be HW4_Settlage_Bob

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

Problem 3

In the lecture, there were a few links to Exploratory Data Analysis (EDA) materials. According to Roger Peng, what is the focus of the EDA stage of an analysis? Hint: this is summarized in the free sample portion of his online book.

Problem 4

In this weeks folder, there is an Excel file containing the dataset for this problem: HW4_data.xlsx. Read this into R (see below). Make sure you get (and combine) BOTH sheets of data. Work up a well annotated reproducible exploration of this dataset using the principles discussed in class or in the supplementary material.

```
library(xlsx)
prob4_data1 <- read.xlsx("HW4_data.xlsx", sheetIndex = 1)
```

Things you need to answer/show:

1. summary statistics (combined in a table?)

2. factor exploration, what factors are present?

3. create a (couple?) multipanel plot (lattice, ggplot2, base R)
   - Base R help: http://www.statmethods.net/advgraphs/layout.html
   - GGplot2: http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

4. correlation plots (?pairs, ?plotcorr)

5. This is the SAME dataset as in Problem 6 of the last homework with the factors relabeled. What is the lesson in this dataset when considering the summary statistics and the plots??

NOTE: DO use this version and NOT the previous version, i.e. load using the xlsx package as above so the factors are correctly labeled.

Problem 5

Using a dataset of your choice (internal R dataset, simulated, anything you wish to plot), create a function that takes as arguments an X-Y dataset and creates a multipanel plot with:

1. a scatter plot
2. marginal histogram in the X dimension
3. marginal histogram in the Y dimension

The scatter plot should be the majority of the plot space. The marginal histograms should be aligned with the appropriate axis.

Problem 6

Please push your completed homework to Github and submit a pull request.