MSDS 6306: Doing Data Science – Versioning

Live session Unit 04 assignment

Due: 1 hour before your 5th live session (September 27)

Submission

ALL (non-swirl) MATERIAL MUST BE KNITTED INTO A SINGLE, LEGIBLE, AND DOCUMENTED HTML DOCUMENT. Formatting can be basic, but it should be easily human-readable. Unless otherwise stated, please enable {r, echo=TRUE} so your code is visible.

Questions

1. **FiveThirtyEight Data (30 points):** Navigate on GitHub to https://github.com/rudeboybert/fivethirtyeight and **read README.md**. Seriously, it will include every command you need. Test out some commands on R.

   a. Install the fivethirtyeight package.

   b. In the *listing of Data sets in package 'fivethirtyeight,'* assign the eighteenth data set to an object 'df.'

   c. Use a *more detailed list of the data sets* to write out the URL in a comment to the related news story.

   d. Using R command(s), give the dimensions and column names of this data frame.

2. **Data Summary (30 points):** Use your newly assigned data frame for Question 2.

   a. Write an R command that gives you the column names of the data frame. Right after that, write one that counts the number of columns **but not** rows. **Hint:** The number should match one of your numbers in Question 1d for dimensions.

   b. Generate a count of each unique major_category in the data frame. I recommend using libraries to help. I have demonstrated one briefly in live-session. To be clear, this should look like a matrix or data frame containing the major_category and the frequency it occurs in the dataset. Assign it to major_count.

   c. To make things easier to read, enter par(las=2) before your plot to make the text perpendicular to the axis. Make a barplot of major_count. Make sure to label the title with something informative (check the vignette if you need), label the x and y axis, and make it any color other than grey. Assign the major_category labels to their respective bar. Flip the barplot horizontally so that bars extend to the right, not upward. All of these options can be done in a single pass of barplot(). **Note:** It's okay if it's wider than the preview pane.

   d. Write the fivethirtyeight data to a csv file. Make sure that it does not have row labels.

3. **Codebook (30 points)**:

    a. Start a new repository on GitHub for your SMU MSDS homework. On your local device, make sure there is a directory for Homework at the minimum; you are welcome to add whatever you would like to this repo in addition to your requirements here.

    b. Create a README.md file which explains the purpose of the repository, the topics included, the sources for the material you post, and contact information in case of questions. Remember, the one in the root directory should be general. You are welcome to make short READMEs for each assignment individually in other folders.

    c. In one (or more) of the nested directories, post your RMarkdown script, HTML file, and data from 'fivethirtyeight.' Make sure that in your README or elsewhere that you credit fivethirtyeight in some way.

    d. In your RMarkdown script, please provide the link to this GitHub so the grader can see it.

4. **Swirl (10 points)**: Complete Module 15 in the R Programming course of Swirl. *Copy your code/output to a separate .txt file. It does not need to be included in your RMarkdown file. The grader has requested at minimum to show the 90%-100% progress bar for the module and what output you had for it.*

    a. Complete "15: Graphics Basics"

<u>Reminder</u>

To complete this assignment, please submit **one** RMarkdown and matching HTML file that includes questions 1-2, and a .txt file containing solely your swirl output (Question 4) at least one hour before your live session on September 27, 2017. You do not need to submit a link to your GitHub: just note where it is in your RMarkdown file. Make sure it is public!! Please submit all files at the same time; only one submission is granted.

Good luck!