BIOS 6660, Spring 2019
Homework 5: Reproducible research
Due: Tuesday, February 26th at 10:30am

In this assignment, you will develop an end-to-end reproducible data analysis using a public dataset.

Instructions for turning in assignment: As always, submit a GitHub URL through Canvas for the repository version containing your final submission. We will look in your **`Homework_5`** directory for **`hw5.Rmd`**, **`hw5.html`**, and a **`data`** subdirectory containing the data .zip file, .csv file, and .txt file downloaded from the web. **We will run your R Markdown document** to make sure it runs from end to end with no manual intervention required. Before submitting your assignment, we recommend deleting the **`data`** directory and re-running your workflow to make sure everything is automated. Additionally, we will look back at your commit history to verify that you made multiple commits while developing the analysis.

Step 1: Setup

In your **`BIOS6660`** repository on your local computer, create a **`Homework_5`** subdirectory. Inside **`Homework_5`**, create a new R Markdown document named **`hw5.Rmd`**. Keep the provided line **`knitr::opts_chunk$set(echo = TRUE)`** so that all code will be included in the report.

Step 2: Data download

Create a new section of your R Markdown document titled “Data download”. Write a code chunk to create a **`data`** subdirectory if it doesn’t exist, download the zipped dataset from this URL into the **`data`** subdirectory, and unzip it within the **`data`** subdirectory. Run the code chunk, deleting any previously downloaded files, until the files download and unzip correctly. Then add code to check if the unzipped data file already exists so the download will not repeat the next time you run the code. Commit the current version of your code and the data files to GitHub.

Step 3: Overview

Now that you have downloaded the data and unzipped it, you can view the file **Facebook_metrics.txt** which contains a brief description of the dataset. Using this, create a section titled “Overview” at the very beginning of your report and write a brief description of the dataset. Include a mention of the **Facebook_metrics.txt** file for details on the data. Knit the report and commit all changes to GitHub.

Step 4: Dataset reproducibility

Add a section titled “Dataset reproducibility” after “Data download”. In the new section, add a code chunk to print the MD5 checksum of the .csv file as well as its **`file.info()`**. Knit the report and commit all changes to GitHub.

Step 5: Load data

Add a section titled “Load data” and a code chunk that uses **`read.table()`** to load the data from the .csv file into a data frame. Notice the field separator in the data file.

Step 6: Data processing

Add a section titled “Data processing” and a code chunk in which you add a column called **`Day`** to the dataset. **`Day`** should contain the day of the week spelled out (e.g. “Sunday”, “Monday”, etc.). You should use the **`Post.Weekday`** column which contains a numerical representation of the day on which the post appeared. As they do not specify the day encoding, assume 1 is Sunday. There are many ways to accomplish this step.

Step 7: Exploratory analysis

Add a section titled “Exploratory analysis”. Add a code chunk in which you use dplyr verbs to print the total number of posts of each type (“Link”, “Photo”, etc.) that the page posted. Add another code chunk in which you make a ggplot boxplot of **`Total.Interactions`** (*y*-axis) by **`Type`** (*x*-axis). For the boxplot, use a log10 scale on the *y*-axis. Then add a second boxplot of **`Total.Interactions`** by your new column **`Day`**. Optional: if you want to reorder the days in the second boxplot, you can follow the answer to this post.

“Exploratory analysis” should contain three code chunks. Each code chunk should be preceded by a brief explanation. As this is exploratory analysis, the plots don’t need to be pretty. Knit the report and commit all changes to GitHub.

Step 8: Data analysis

Add a section titled “Data analysis”. Add a code chunk in which you do some statistical test to check if there is a significant difference in **`Total.Interactions`** based on whether the post was **`Paid`** or not. Use text explanations to introduce the test and discuss the results. It doesn’t matter which test you choose or whether the test is actually appropriate for this situation.

Finally, add another code chunk with one more ggplot plot of your choice; the code chunk should be preceded by an explanation of the plot. Plot anything you think might be interesting about this dataset. Knit the report and commit all changes to GitHub.

Step 9: Software environment

Add a final section titled “Software environment” and a code chunk calling **`sessionInfo()`** to record your environment. Make sure you are satisfied with your entire workflow and knit a final version to commit to GitHub.