PHP 2550: Worksheet 2

Due: September 15th at 11:59pm

Reading Recap

1. For each of the listed readings below, recap the reading in three points. Each point could be summarizing a key takeaway, something that surprised you, or something that you want to remember.

    - Chapters 1-2 of the "The 9 Pitfalls of Data Science"
    - Ten Simple Rules for Reproducible Computational Research
    - Researcher Requests for Inappropriate Analysis and Reporting

2. How did the readings relate to each other? (1-2 paragraphs)

Reproducible Code and Documentation

1. What makes an analysis reproducible? Consider your answer in three settings: (a) when the data and code is available, (b) the data is available but the code can't be shared, and (c) when neither the data or code are available.
2. Read the code in the R script "gerrymander_sim.R" on Canvas in the Code Examples folder. The code uses the file "congressional_election_results_post1948.csv" and is implementing a simulation test for gerrymandering. Improve the code to be more reproducible by consider the following guidelines.

    - Use comments, white space, and variable names to improve the readability of your code.
    - When possible, wrap code into functions that include roxygen documentation.
    - Include all libraries and the random seed (if used) at the top of the document.
    - Avoid using magic numbers - numbers that are not explained or defined.

3. Data cleaning and documentation is a key part of the data science pipeline. How we preprocess the data can impact our analysis. Additionally, we should be checking our data for potential quality problems. To demonstrate this point, we have provided two data files from Providence's scooter share program in June 2019. The files are "prov_locations.csv" and "prov_events.csv". The first file contains information about scooter locations where each row is a period of time when a scooter was in a set location. The second file records every event in the system.

   Your goal is to do a quality check on the data. There are two types of checks you should think about: single-source checks or multiple-source checks. A single-source problem can relate to the attributes (columns), records (rows), or the data source. A multiple-source problem relates to how the two data sets relate to each other. To get started, look over the data codebook and then brainstorm some things you plan to check in the data. Your code and report should use the same reproducible guidelines above.