CSCE 587 – Big Data Analytics

R Homework Assignment 1

Due: Tuesday, 2/4, by 11:59pm

Deliverables: To the assignment in Blackboard, you will submit an R script (a .R file) with commands to answer the questions below, and comments clearly labelling each answer. You will also submit two PDFs with the plots. You may submit these as three separate files, or you may compress them to a zip file and upload it.

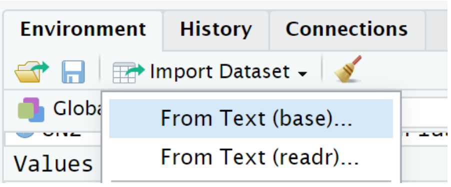
The Data: The data attached to the assignment in Blackboard provides population estimates for US states in each of the years from April 1, 2020 to July 1, 2024. The data was published by the US Census Bureau (at https://www2.census.gov/programs-surveys/popest/datasets/2020-2024/state/totals/NST-EST2024-ALLDATA.csv). A PDF with full descriptions of the fields in the dataset can be found at https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2020-2024/NST-EST2024-ALLDATA.pdf

Preparation:

You can download and save the csv file attached to the assignment in Blackboard OR you can use wget at the terminal to download the file from my CSE site to your virtual machine:

```
wget https://cse.sc.edu/~bhipp/587/popest2024.csv --no-check-certificate
```

Then click the "Import Dataset" button under Environment, select "From Text (base)", and select the popest2024.csv file to load it into your environment.



Additional Instructions of Note:

Your answer for each question must be clearly labelled in your R script and must include the command(s) necessary to generate the answer. Unlabeled answers and answers that don't contain the R command(s) to justify them will not receive credit.

The csv file contains nation, region, division, and state level data (DC and Puerto Rico are included with state level data). State level records have a value of 40 in the SUMLEV field. Each of the questions below are about the state level records. So, you will need to subset out the state level data with the appropriate R commands (those commands must be in your R script) to answer the questions.

For questions that start with "What state-level record…", be sure your answer includes the name of the state!

The questions / problems:

1. What state-level record had the most international migration from 7/1/2023 to 6/30/2024 (field INTERNATIONALMIG2024)?
2. What state-level record had the least international migration from 7/1/2023 to 6/30/2024 (field INTERNATIONALMIG2024)?
3. What state-level record had the highest death rate from 7/1/2023 to 6/30/2024 (field RDEATH2024)?
4. What state-level record had the lowest death rate from 7/1/2023 to 6/30/2024 (field RDEATH2024)?
5. What was the mean state-level population in 2024 (field POPESTIMATE2024)?
6. What was the median state-level population in 2024 (field POPESTIMATE2024)?
7. How many state-level records had a negative change in population from 7/1/2023 to 7/1/2024 (field NPOPCHG_2024)?
8. Use hist() to construct a histogram of state birth rates from 7/1/2023 to 6/30/2024 (field RBIRTH2024).
    - Set the breaks parameter to 16 (to set the number of bins used)
    - Update the horizontal axis label to have a viewer-friendly, meaningful label.
    - Update the plot title to be a viewer-friendly, meaningful title.
    - Save the plot to a PDF. This PDF must be submitted with your assignment.
9. Use plot() to create a scatter plot comparing domestic and international migration from 7/1/2023 to 6/30/2024
    - Use international migration (field INTERNATIONALMIG2024) as the horizontal axis
    - Use domestic migration (field DOMESTICMIG2024) as the vertical axis.
    - Ensure that the axes have meaningful labels.
    - Ensure that the plot is given a meaningful title.
    - Save the plot to a PDF. This PDF must be submitted with your assignment.
10. How many state-level records in the Midwest (Region == 2) had positive domestic migration from 7/1/2023 to 6/30/2024 (field DOMESTICMIG2024)?

Graduate Students Only:

11. Compute the mean international migration from 7/1/2023 to 6/30/2024 (field INTERNATIONALMIG2024) for states in the Northeast (REGION == 1)?
12. Compute the median international migration from 7/1/2023 to 6/30/2024 (field INTERNATIONALMIG2024) for states in the West (REGION == 4)?
13. How many southern states (REGION == 3) had more international migration than the maximum international migration for a midwestern state (REGION == 2) from 7/1/2023 to 6/30/2024? Output the names of those southern states.