MSDS 6306: Doing Data Science – Exploratory Data

Live session Unit 10 assignment

Due: 1 hour before your 11th live session

Submission

ALL MATERIAL MUST BE KNITTED INTO A SINGLE, LEGIBLE, AND DOCUMENTED HTML DOCUMENT. Use RMarkdown to create this file. 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

Background: Your organization is responsible for building new VA hospitals in the mainland of the United States. You are a little overwhelmed by the prospect, not sure which places need the most help. You decide to begin by exploring healthcare facility data recorded by the U.S. Government.

Disclaimer: While these are real data, the assignment is not an endorsement for any particular position on medical affairs or building hospitals. It is for instructional use only.

1. **Mental Health Clinics (40%)**

    a. This data set is a survey of every known healthcare facility that offers mental health services in the United States in 2015. Navigate to https://datafiles.samhsa.gov/study-dataset/national-mental-health-services-survey-2015-n-mhss-2015-ds0001-nid17098 and select the R download. Look through the codebook PDF for an explanation on certain variables. Upon opening the RDA file, the data set should be inserted into your global environment, which you can then reference.

    b. Please create code which lists the State abbreviations *without their counts*, one abbreviation per State value. It does not have to in data frame format. A vector is fine.

    c. Filter the data.frame from 1A. We are only interested in the Veterans Administration (VA) medical centers in the mainland United States—create a listing of counts of these centers by state, including only mainland locations. Alaska, Hawaii, and U.S. territories should be omitted. DC, while not a state, is in the mainland, so it should remain included. Convert this to data.frame()

    d. Create a ggplot barchart of this filtered data set. Vary the bar's colors by what State it has listed. Give it an appropriately professional title that is **centered**. Make sure you have informative axis labels. The State axis should be readable, not layered over each other. You're welcome to have a legend or not.

2. **Cleaning and Bringing in New Features (60%)**

a. This graph (1D) might be somewhat misleading, as bigger states may have more hospitals, but could be more sparsely located. Read statesize.csv into your R environment. This contains essentially a vector of square miles for each state. In trying to merge it with your data.frame() from 1C, you find that they don't match. Use **paste()** on your LST column in 1C to see what the matter is, and write what you observe in a comment.

b. Correct the problem with the LST column using any method in R that is programmatic and easily understandable. Once you have made these state abbreviations identical to statesize.csv's Abbrev column, merge the data.frame() from 1C and statesize.csv in order to add size information.

c. Calculate a new variable in your combined data.frame() which indicates the VA hospitals per **thousand** square miles.

d. Create another ggplot which considers the VAs per square thousand miles, rather than just frequency.

- Make sure the State axis is readable, like before. Change the title and axes as appropriate.
- Modify the ggplot syntax to make your bars in descending order (there are StackOverflow topics for this, and I have demonstrated how in Live Coding in prior classes).
- Color-code the bars based on Region (see the merged data.frame)—however, change the color scheme from the default. Any set of colors is fine, so long as it is readable.
- Keep the legend—you should have four regions and therefore four colors.

e. What patterns do you see? By this metric, is there any region that seems relatively high for VA medical centers per thousand square miles? How about low? Given these data, what advice might you give your boss before you start modeling (and why)?

Reminder

To complete this assignment, please submit **one** RMarkdown and matching HTML file at least one hour before your live session. Please submit all files at the same time; only one submission is granted.

Good luck!