Homework 4 - Dplyr and tidyR

Instructions

1 This homework is a little different. Because some of what we are asking you to do is inherently a little subjective, we aren't using Gradescope. Instead you will have to submit a link to a Github repository that contains all the code and output the assignment asks you to create.

2 Please remember to make this a public repository so we can view the work.

3 The detailed instructions for each question and the expected deliverables are given in the pages below. Please be sure to read these carefully.

4 The deadline for the homework is Friday September 27 at 11:59pm ET.

5 It is encouraged for you seek assistance from and collaborate with your peers if you are stuck. If you do take help from your peers, please be sure to credit them at the bottom of your pdf document. Note that even if you do collaborate, all the work you submit must still be your own.

6 As always if you have concerns/questions/need help, feel free to come talk to us.

7 A rough grading rubric. Your work will be assessed on: 1) Correctness of code where applicable. 2) The soundness of your reasoning and the thinking apparent in your choices. Your justifications and written responses will contribute significantly but not exhaustively to this. 3) Presentation. Your output should look professional. There are a number of aspects to this: The structure of your document; whether the visualizations have informative labels, are easy to parse and look good; whether your code is well commented; and the wording you use, which should be concise and cogent.

As we discussed in lecture on Wednesday, the National Oceanic and Atmospheric Administration(NOAA) keeps track of meteorological data from a number of buoys. For this exercise, we are interested in Buoy number 44013 located sixteen nautical miles east of Boston Harbour and the questions will deal with data from that buoy.

This is the link to the National Data Buoy Center: https://www.ndbc.noaa.gov/

This is the link to the buoy of interest: https://www.ndbc.noaa.gov/station_page.php?station=44013

a Your first exercise is to read in the data for all the years from 1985 to 2023. As discussed in class, you don't want to do this manually and will need to figure out a way to do it programmatically. We've given you a skeleton of how to do this for data for one year below. Your task is to adapt this to reading in multiple datasets from all the years in question. This example code is meant to be a guide and if you think of a better way to read the data in, go for it.

Keep in mind that initially, these datasets did not record units and then started to do so in the line below the column headers. So for some years you will have to skip 1 instead of 2.

In addition to reading in this data, use lubridate to create a proper date column.

```
file_root<-"https://www.ndbc.noaa.gov/view_text_file.php?filename=44013h"
year<-"2023"
tail<- ".txt.gz&dir=data/historical/stdmet/"
path<-paste0(file_root,year,tail)
header=scan(path,what= 'character',nlines=1)
buoy<-fread(path,header=FALSE,skip=2)
colnames(buoy)<-header
```

Save your code in an R Script with an appropriate name which you must include in your Github submission. Keep in mind that if we clone your repository, this script must run without errors for you to get credit. Remember to comment your code for readability.

For the following questions (b through d), you will need to put your code, its output, and your written answers into a pdf. One way to do this is to use R Markdown or Quarto and knit into a pdf. Include the rmd/qmd file as well the pdf in the github repository you submit.

b Your next exercise is to identify and deal with the null data in the dataset.Recall from class that for WDIR and some other variables these showed up as 999 in the dataset. Convert them to NA's. Is it always appropriate to convert missing/null data to NA's? When might it not be? Analyze the pattern of NA's. Do you spot any patterns in the way/dates that these are distributed?

Bonus: Can you think of other data sources you can add to this that might shed light on the NA's? Join those to your data and investigate. Look at government shutdowns and budget changes.

c Can you use the Buoy data to see the effects of climate change? Create visualizations to show this and justify your choices. Can you think of statistics you can use to bolster what your plots represent? Calculate these, justify your use of them. Add this code, its output, your answers and visualizations to your pdf.

d As part of this Homework, you have been given data for rainfall in Boston from 1985 to the end of 2013. Your job for this exercise is to see if you can spot any patterns between rainfall(whether it happens and how much of it there is) and the readings taken by the weather buoy in the same period. There are a number of ways you can do this but the broad steps are: 1) Acquaint yourself with the data. Look at distributions and summary statistics(dplyr is great for coaxing means, averages, counts out of data). 2) Create visualizations. Can you see patterns in the distributions and visualizations? Investigate these with more statistics and visualizations. 3) Try building a very simple model. Explain your choices at every step. Do you come away from this exercise with more sympathy for the weather people on TV who keep getting their forecasts wrong?

Structure your response to this question as an exploration of the data and as if it were a report. Show the graphs and outputs.

Expected Submission Items:

Your Github Repo should contain at least all of the following:

An R Script to read in all the buoy data

A pdf with your responses to questions (b) through (d)

The code you used to generate the pdf be it an rmd/qmd file or something else.

Any the data you need to include for your scripts and analysis to run