CPTR 420
Test 2
November 21, 2023 12:30 – 1:50 p.m.

You may use your notes and previous scripts to help you. You may use r cheat sheets files or official documentation. However, you must not use AI help or answer sites like Stack Overflow. You must confer with person within the class or online. You should neither share nor accept help from anyone.

The following exercise uses data from multiple sources:

The Hotel_Reviews.csv file contains reviews for several hotel in most states in the US (DC and some states are missing)

The State_Crimes_2019.csv file contains data about different types of crimes in states across the US (all states and the District of Columbia are included)

state.name is a dataset available in R. The dataset has the full name of all the US states in alphabetic order
state.abb is a dataset available in R. The dataset has the 2-letter abbreviation of all the US states in order by state name.

You will be manipulating these data to look for connections and to prepare visualizations.

Exercises:
Note that there are several ways to achieve a result. Your goal for each exercise is to get the result. Your code should be versatile enough to work if the data were changed e.g. avoid performing calculations in your head or with a calculator and using these results as arguments.

Your script should be well documented. Add a comment block at the beginning that has your name, the date and, the filename

Part 1: Prepare the environment and get the data (10 points)

1. Install and load the packages you will use, this may include, but are not be limited to:
   tidyverse (dplyr, tidyr, ggplot2, stringr), lubridate, corrplot, wordcloud2, tmap, maps, tm, tidytext

2. Get your datasets:
   a. yourInitials_hotel_df fill missing data with NAs
   b. yourInitials_crimes_df
   c. YourInitials_state_abb
   d. YourInitials_state_name

Part 2: Inspect and clean the data (30 points)

3. Create a dataframe called states_df that has a column for the name and another for the abbreviation of all states (i.e. the data in the state.name and the state.abb datasets). The column namesyou're your states_df should be state_name and state_abbr
4. The hotel file is your main data. Do the following that file:
    a. Find the number of rows and columns in the dataset
    b. List the first 5 rows of data
    c. Show the descriptive stats i.e. mean, median, percentiles, NAs, …
    d. The DateAdded and DateUpdated variables are character type that have both date and time. You want only the date to be stored in these variable as R dates.
    e. The following variables will not be needed, remove them:
        i. Keys, categories, reviews.date, sourceURLS, reviews.username, reviews.dateAdded, websites
    f. Change the name of the province variable to state_abbr

Part 3: Some Analysis (30 points)

5. How many hotel reviews are there for each state and what is the average rating for each state? Create a review_df that shows the state abbreviation, number of reviews, and average rating
6. Add the state_name from the state_df created in #3 above. Note that your review_df data does not include all states
7. We need to see the hotel review and crimes data together. Add the crimes data to the review_df
8. Some columns are not needed for our analysis. Remove the following:
    a. The Data.Totals columns e.g. Data.Totals.Property.All, Data.Totals.Property.Buglary, …
    b. Remove Data.Rates.Property.Burglary, Larveny, Motor
    c. Remove Data.Rates.Violent.Assualt, Murder, Rape, Robbery

Part 4: VIzualizations (20 points)

For each viz, add titles labels

9. Is there a correlation between Property crime rates and rating? Prepare a scatter plot with avg_rating and Data.Rates.Property.All. Your plot should be well labelled
10. Prepare a correlation matrix and corresponding plot for the numeric data in your review_df (do not include state name, state abbreviation, and population)

Do either 11 or 12 below, not both

11. Create an interactive map of the states that shows the state abbreviation, number of reviews, average rating, population
12. For any state of your choice, create a viz (wordcloud , bar chart, horizontal bar, …) of the 100 most popular words in the review title (not including common English stop words. **note that this data is in the original dataset)**

Save an image of each visualization to your folder

Upload your folder with the script and your visualizations to LearningHub. Do not upload the data