Assignment

Instructions: Your final document should be an 1) RMarkdown file and 2) HTML file knitted from R Markdown.

In answering each of the following questions **please include a) the question as a header in your Rmarkdown report, b) then include the raw code that you used to generate your results, and c) the top ten rows/values/or elements of the resulting dataframe, vector, or list created in your results (unless a lesser amount is requested).**

Feel free to refer to any R scripts provided throughout the course to answer the following questions:

1. Web-scrape all table data from the following web-page and build a data frame in R. Limit your final table to include columns for "Package", "Item", "Title", "Rows", and "Cols". Print the first five rows of the table.

http://vincentarelbundock.github.io/Rdatasets/datasets.html

2. Web-scrape the full links to every CSV file listed in the CSV column of the web-page. Add a new column to your data frame that includes these links. Name the column "CSV Links". Print the first five rows of the table.

Note: You may need to use string operations to recreate the full link to the csv files after they are scraped.

3. Use R code/functions to search the "Title" column to return the row of data with the title, "Violent Crime Rates by US State"

4. Import the csv file into R using the full link listed in the "CSV Links" column for this dataset. Create a new variable called "Violent_crime" that adds together data for all columns in the dataset that contain violent crime data (i.e.-add data from assault, murder, and rape columns together in new column called "Violent_crime").

5. Using one or more of the following example datasets that come preloaded in R (see ?data(state) for more information), add state region codes, state divisions, and all of the variables from the "state.x77" dataset to your Violent Crime Rates data frame. Print the first five lines of your new dataset.

- **state.abb, state.area, state.center, state.division, state.name, state.region**

Note 1: state names can be extracted to new columns for joins from these datasets by using row.names() if needed.

Note 2: Some of these state datasets may be matrix objects or lists. Be sure to convert them to data frames before joining them if needed.

6. Calculate the average for each numeric column in the dataset.

7. Group the data by region and then calculate the average for each numeric column in the dataset per region. Which region had the highest population (data is from the late 1970s)? Which region had the most violent crime?

8. Group the data by division and then calculate the average for each numeric column in the dataset per division. Which division had the highest population (data is from the late 1970s)? Which division had the most violent crime?

9. What SQL statement would you write to return two columns denoting income and Illiteracy in your state data?

10. What SQL statement would you write to return two columns denoting income and Illiteracy in your state data and sort the data from the highest to lowest income values?

11. What SQL statement would you write to return two columns denoting income and Illiteracy in your state data and sort the data from the highest to lowest income values and limit the data to incomes at or higher than 5000?

12. What SQL statement would you write to return two columns denoting income and Illiteracy in your state data and sort the data from the highest to lowest income values and limit the data to incomes at or higher than 5000 and return the top 10 rows only?

13. Create a new data frame that includes two columns from your state data denoting state names and violent crimes. Spread the state names to 50 unique columns with a single row that includes the violent crime data per state. Print the first five columns of the new dataset.

14. Take the dataset from question 13 and use a function from the apply family of functions to return a named list with each name denoting a state and each value per state indicating the square root of the value for violent crimes.

15. Subset the list you created in question 14 to extract values for Texas and New York.