DA5020 – Assignment 4

For the following questions, you will use the NYC Green Taxi Trip Records for December 2017; view the accompanying green trips data dictionary for additional information.

Load the NYC Green Taxi Trip Records data ***directly from the URL*** into a data frame and answer the questions below. Note: you should include any suitable data preparation steps from the previous assignment.

Question 1 — (5 points)

Inspect the data and identify at least three columns/fields that should be represented as factors and convert their respective data types to a factor. Hint: make use of the data dictionary to understand the expected values for each field.

Question 2 — (10 points)

Visualize the data to determine: 1) the most common way that New Yorkers request/hail a cab and 2) the three most popular pickup locations. Helpful fields are: **trip_type** and **PULocationID**. Explain your results.

Question 3 — (10 points)

Count the frequency of pickups for each day in December. Visualize the results using a bar chart; show the frequency on the y-axis and the date on the x-axis. Do you detect any patterns in the visualization? Note: do not include the time in your calculation or the visualization (only use the date).

Question 4 — (10 points)

Create a function called *HourOfDay()* that takes one argument which is a **textual representation of a timestamp in the format 'YYYY-MM-DD HH:MM:SS'** and uses a regular expression to extract the hour (or you can use the lubridate package to extract the hour). For example, the function should take a timestamp in the following format: '2017-12-01 11:10:25' and return '11'.

Question 5 — (5 points)

In a new R chunk, demonstrate that the *HourOfDay()* function works using the timestamps in the lpep_pickup_datetime column. You will need to: 1) provide each value in the lpep_pickup_datetime column as the argument to your function, and 2) save the result from the function in a new column called lpep_pickup_hour. Hint: you can use the **mutate** function in dplyr to do this.

Question 6 — (10 points)

Report the median trip distance grouped by the hour of the day (use the lpep_pickup_hour field that you created). Visualize the results and explain any patterns you observed.

Question 7 — (+5 bonus points)

This bonus question is optional and you can choose to respond to either question a OR b below. Write the appropriate R code to answer one of the following questions:

a) Filter the data to identify the date with the least trips. Do you detect anything interesting (or unusual)?

b) Demonstrate a technique that you would use to handle missing values in this dataset (or using a dataset of your choice).

Submission Details

- Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd File, DA5020.A4.FirstName.LastName.Rmd and your PDF/HTML DA5020.A4.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.
- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it.) Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.
- Not submitting a knitted PDF or HTML will result in reduction of 30 points.
- Not submitting the .Rmd file (or both) will result in a score of 0.

Useful Resources

- R Markdown Notebooks
- NYC Green Taxi Trip Records - December 2017 and Green Trips Data Dictionary
- Visit the following webpage to obtain more information on the dataset: NYC TLC Trip Record Data

** Note: Do not merely print the output for each question but always explain the results from your code. Points will be deducted for any missing explanation and any visualizations that are not clearly labeled.