DA5020 – Assignment 4

For the following questions, you will use the NYC Green Taxi Trip Records for February 2020 or 2018; view the accompanying green trips data dictionary for additional information.

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

You may want to complete the two data camps Introduction to Data Visualization with ggplot2 & Communicating with data in the Tidyverse before completing this assignment and will count for bonus points.

Bonus — (+5 bonus points, *All instructions must be followed)* (Both questions must be completed to obtain the bonus points)

1. Complete the data camp "Introduction to Data Visualization with ggplot2 "and a separate data camp "Communicating with data in the Tidyverse". A link to Data Camp will be shared within thecourse, so that you can access the resource for free using your Northeastern email. Please submit the statement of accomplishment from both data camps to earn credit (must also have your name on it). (No partial credit)

2. Filter the data and extract the date with the most trips. Do you detect anything interesting (or unusual) with the trips that occurred on that day?

Question 1 — (10 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 range of values for each field.

Question 2 — (10 points)

Analyze the data to determine what is the most common way that people: a) hail a cab and b) pay for the trip. Helpful fields are: **trip_type** and **payment_type**. Explain your results.

Question 3 — (10 points)

Count the frequency of pickups for each day in February (exclude any other months pickups). Visualize the results using a bar chart and show the frequency on the y-axis and the date on the x-axis. Do you detect any patterns? What are your observations? Note: do not include the time in your calculations 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: '2020-02-01 11:10:25' and return '11'. note: 2018 date format is a bit different

Question 5 — (5 points)

In a new R chunk, apply the function *HourOfDay()* using each value in the lpep_pickup_datetime column and save the results in a new column called lpep_pickup_hour. Hint: you can use the **mutate** function in dplyr to apply your function to each row in the dataframe.

Question 6 — (5 points)

Report the median trip distance grouped by the lpep_pickup_hour. Visualize the results and explain any patterns you observed based on the hour of day.

Submission Details

** 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.

- **Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd)**.

- Name your .R script using the following format: DA5020.A1.FirstName.LastName.R, where FirstName.LastName is your first and last name. If you submit an Rmd file, you should use DA5020.A1.FirstName.LastName.Rmd

- Ensure that you number each question correctly.

- We must be able to run your code. You will not receive credit for any code that does not run. 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.

- Describe answers for full marks