DA5020 – Assignment 3

For the following questions, you will use the NYC Green Taxi Trip Records for February 2020 or the 2018 data found in Canvas; view the accompanying green trips data dictionary for additional information. You may want to review the data camps Introduction to the Tidyverse & Reshaping Data with tidyr before doing this assignment. Completing both Data Camps will earn extra credit for this assignment

Question 1 — *(+5 Bonus Points, All instructions must be followed)*

1. Complete the data camp "Introduction to the Tidyverse and a seperate data camp "Reshaping Data with tidyr". A link to Data Camp will be shared within the course, 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)

Question 2 — (10 points)

Load the NYC Green Taxi Trip Records data *directly from the URL or data in Canvas* into a data frame called tripdata_df.

Inspect the data to identify its dimensions and the frequency of missing values. Helpful functions: dim(), glimpse() and summary(). Tip: it is also good practice to inspect the data type for each field/column to determine if the data was imported correctly.

Question 3 — (10 points)

Explore the data to determine if there are any inconsistent or invalid data; for example, examine the dates within the dataset to see if they align with your expectations (remember you downloaded a dataset for February 2020). Identify at least **Three (3)** things that stand out to you and remember that this is based on your observations about the data, so it's important to demonstrate what you found.

Question 4 — (10 points)

Create a histogram, showing the **trip_distance**. Is the data skewed? Explain what you observed using 1-2 sentences. Note: you may need to rescale the y-axis using a log scale to improve the visualization.

Question 5 — (10 points)

Analyze the **tip_amount** and **trip_distance** variables to identify any outliers. You can assume the outliers are 3 standard deviations from the mean. Comment on the outliers that were detected; after which, remove the outlier **tip_amount** from the data (building from Q4)

Question 6 — (5 points)

Filter the data from question 5 above (Trip_distance), and create a suitable visualization to show the frequency of trip_distance by **payment_type** (e.g. credit card, cash, etc). Ensure that your visualization(s) has a title and label both the x and y axis. (outliers should be removed)

Question 7 — (5 points)

State at least two methods/techniques that can be used to handle missing data. Which approach would you recommend, is suitable, to handle missing values in this dataset? How would you do this in R and provide examples on Taxi data.

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.A1.FirstName.LastName.Rmd and your PDF/HTML DA5020.A1.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