DA5020 – Assignment 3

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.

Question 1 — (5 points)

Load the NYC Green Taxi Trip Records data ***directly from the URL*** 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 2 — (10 points)

Explore the data to determine if there are any invalid data. For example, examine the dates to see if they align with your expectations (Hint: remember that you downloaded a dataset for December 2017). Identify at least three 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 3 — (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 x-axis or the y-axis using a log scale to improve the visualization. Remember that you did this in a previous assignment.

```
scales_x_log10()
```

Question 4 — (15 points)

Analyze the **tip_amount** 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 frame.

```
mean.sleep.total <- mean(msleep$sleep_total,
na.rm = TRUE) # calculate mean
sd.sleep.total <- sd(msleep$sleep_total, na.rm =
TRUE) # calculate sd
msleep$z.score <- (mean.sleep.total-
msleep$sleep_total)/sd.sleep.total # get z score
using definition
outliers.msleep <-
msleep[which(abs(msleep$z.score) >= 1.5),] #
subset dataset (outliers) according to z score
```

Question 5 — (10 points)

Create a suitable visualization that shows the *proportion* of trips based on the **payment_type** (e.g. credit card, cash, etc). Comment on the most common method of payment. Ensure that your visualization has a title and label both the x and y axis.

```
geom_bar(x=, y= ..prop.., group = 1)
```

Question 6 — (5 bonus points)

Research at least two methods/techniques that can be used to handle missing data. Which approach would you recommend to handle missing values in this dataset? Note: only your recommendation is required; you do not need to implement the code (for this question).

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.A3.FirstName.LastName.Rmd and your PDF/HTML DA5020.A3.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.