Hands On Workshop Questions

Below are some questions that you should be able to answer now that you've learned some important R functions for the data science process. Some of the phases have more questions than others. You don't have to answer every single question, but we strongly recommend that you attempt each one. You should implement your answers in a .R file by using hashtags (#) to denote each comment. We won't be checking your submissions, but you might want to revisit the code you rewrote in the future! I have also included some explanation on some questions, you don't need to answer it if you don't know but it doesn't hurt to guess!

Questions are rated on a 1-5 difficulty scale. If you aren't too experienced, don't worry about skipping questions with a 5. They aren't necessary to complete.

I. Data Acquisition

a. (1) Read the taxis file into your working environment. What file reading method did you choose, and why?

II. Data Cleaning & Transformation

a. (1) Examine the dataset, and check the type of each variable. Are there any variables that could be "cleaned"? If so, state which ones.

b. (1) Explain why these chosen variable(s) require cleaning. What kind of code could be used to make these variables work?

c. (1) Create a new variable, called pickup_char, which converts the pickup time into the character class.

d. (2) Examine your new variable by looking at the first few observations. Let's say we want to find taxi pick ups that occurred from 12 AM - 12:59 AM. What function can we use to do this?

e. (4) Create a string pattern that grabs these pickup times and execute it with the proper function. Make sure to save what output you get. After you have this output, try to put the values in the row section of your data frame. What happens? Save this result as a new data frame.

f. (3) Create a data frame that only has passenger trips that have multiple people. Call this variable taxis_group.

g. **(5)** Create a factor variable that categorizes the taxi trips by each day of the month. There are 31 days in the dataset, so you should have a factor should have 31 unique values. (Hint: You should consider looking at the seq(), paste() and factor())

h. (2) Create a new variable, called tip_percent. What two variables would you use to generate tip_percent? Attach this variable to the data frame.

i. (3) Create the time elapsed variable as mentioned at the beginning. Save it into taxis as time_elapsed. Afterwards, create a variable called time_elapsed_min, and save it into the same data frame. (Hint: Examine what some of the values of the time_elapsed variable looks like)

j. (3) Create a data frame that is grouped by the number of passengers and has one variable, the average total fare paid. Make sure to save and store this data frame.
k. (4) Now let's bring in multiple variables. Try creating a data frame with the average of the following variables: time_elapsed_min and tip percent. (Hint: Try looking at the cbind() function)
l. **(5)** Create a vector of group lengths for each day of the month. Then, create a data frame that also groups by day, and include the following variables by their mean: fare_amount, tip_percent, trip_distance, time_elapsed_min. Combine the vector with the data frame, into a single data frame and store it.

III. Exploratory Data Analysis

a. (1) Use the dim() function to report the dimensions of the dataset. What are they? What does each value represent?
b. (2) Generate the summary statistics for the total amount. Do you think there are outliers? Are there any values that seem misreported?
c. (3) Generate the summary statistics for passenger count, trip distance, tip amount, and fare amount and make sure they print out as a data frame. Consider applying summary to the proper set of columns.
d. (4) Generate count tables for vendor ID, passenger count, ratecode ID and payment type such that they all print out at once.
e. **(5)** Generate a table that shows the relationship between passenger count and payment type. You will have to modify what you have been previously doing to answer this question.
f. (3) Use the subset for trips with multiple people. Plot the relationship between trip distance and tip amount. Are there any outliers? What does this suggest about how we should present these variables?
g. **(5)** Find an optimal way to plot the relationship between passenger count and tip rate. Are there any alternative solutions towards using the plot function?

IV. Statistical Modeling

a. (4) Can the time elapsed in a taxi trip explain the tip percent? Use a linear model to examine this relationship. Extract the coefficients from the linear model. Report the linear model as Tip Percent=(Your Intercept)+(Coefficient for time elapsed)*Time Elapsed.

V. Reports & Presentation Graphics

a. (1) Install and load in the "ggplot2" package.
b. (3) Use the subset for trips with multiple people. Create a basic plot that visualizes the relationship between tip percent and time elapsed in minutes. If there are any outliers affecting your plot, consider using subsets if possible to improve the quality.
c. **(5)** Fit a line using the coefficients from the linear model between tip percent and time elapsed (with your subset!). Add this line to the plot from the previous step. (Hint: Consider reading the documentation of geom_smooth, and using model=lm). You may have to consider rewriting the plot with ggplot().

d. **(5)** Create a latitude and longitude plot for drop offs. Color the points such that they are blue if there is a low fare amount and orange if there is a high fare amount. If the plot looks off consider using subsets instead. You will need to use ggplot() and not qplot(). We recommend that even if you are new, you attempt this question. If we have time, we will lead you through the process!