AD699: Data Mining for Business Analytics
Individual Assignment #1

You will upload two files into Blackboard: a PDF with your answers and the .R file that contains your script.

Main Topics: Data Exploration & Data Visualization

Tasks:

- **Data Exploration & Visualization**:

1. Download the file 'parkingLA2017.csv' from our class Blackboard site.

2. Read this file into your R environment (if it takes a while for the file to load, don't worry -- this is normal. Be patient). Show the step that you used to accomplish this.
    a. What are the dimensions of this dataframe? Show the code that you used to determine this.

3. Filter the dataframe. Create a new object that only contains data for your assigned "Make" of car (the list of vehicle Make assignments can be found on our class Blackboard page, in the same folder that contains this assignment prompt). Show the code that you used to accomplish this. *For the next set of questions, use this new dataframe.*

    a. What are the dimensions of this new dataframe that only contains the rows for your assigned "Make" of car?

4. Dealing with missing values:

    a) Are there any missing values in your dataset? How do you know this? Show the R code that you used to answer this question, along with the results that appeared in your console.

    b) Does the variable RP.State.Plate contain any missing values? How do you know this? Again, show your steps and the results.

    c) Find and display the standard deviation for the variable Fine.amount. Does the variable Fine.amount contain any missing values? How do you know this? Again,

show your steps and the results. To deal with this issue, perform an imputation by replacing the NAs with a reasonable alternative. Now, find and display the standard deviation for this variable again. What happened? Why do you think this change occurred?

d) Replace any blank cells in the Location column with NA. In a sentence or two, what does this accomplish?

e) Now, remove all of the records that contain "NA" for AGENCY.SHORT.NAME from your dataset.

5. Dealing with the Date data type

   A. Right now, if you call the str() function on your dataset, you'll see that R does not recognize the Issue.Date variable as a date. Fix this by explicitly telling R to treat this variable as a date. Show the code that you used to accomplish this.
   B. Are the rows currently displayed in chronological (i.e. date) order? How do you know this? Use the arrange() function from dplyr to put the dates in order. Show the code that you used to accomplish this.
   C. What month were you born in? Using the subset() function, make a new object that only contains dates for your particular birth month. Show the code that you used to accomplish this. *We will not use this object again in any of the following steps.*

6. We won't need to use the variable 'ticket number' in our analysis. Remove this column from your dataframe. Show the code that you used to accomplish this.

7. Using the summary() function, find out even more about the distribution of fine amounts. Show a screenshot that displays the Minimum, 1st Quartile, Median, 3rd Quartile, Maximum, and Mean values for parking fine amounts.

8. Identify the five most common types of violation descriptions in your dataset. Show the code that you used to accomplish this, and a screenshot that shows the names of the five most common violation descriptions.

9. Create a new dataframe that only contains data for the five most common violations. Show the code that you used to accomplish this.

You will use this new dataframe for all the following steps in this assignment.

10. Using ggplot, create a barplot that displays the number of occurrences for the five most common violations. Be sure to label your axes, to give the graph a title, and to color each of your bars. In a sentence or two, describe what your barplot is showing you.

11. How did the average size of a fine vary from agency to agency? (use the AGENCY.SHORT.NAME variable to make this grouping). Find the average fine Size for each agency (show a screenshot of your code plus your results) and then display your results visually with a barplot built using ggplot. Give your barplot a title, and clearly label your x and y axes. Be sure to color your bars. In a sentence or two, describe what your barplot is showing you.

12. Using ggplot, create a violin plot that shows the agencies on the x-axis, and the fine amounts on the y-axis. Give your violin plot a title, and clearly label your x and y axes. In a sentence or two, describe what your violin plot is showing you.

13. Using ggplot, create a histogram that shows the frequency of ticket issuances per hour of the day. Use different colors to depict different types of violation descriptions.

    In a couple of sentences, describe what your histogram is showing you. What meaning could someone take from this? Why (or why not) does this histogram make sense, in terms of what it says about parking violations and times of day?