**GALE - Data Science Assessment**

This challenge is designed to test your technical and analytics skills on a real-world data set.

We'll use the New York taxis data - in particular the Green taxis - as our experimental data set.

Please download Green taxi data for February 2016 from this site:

http://www.nyc.gov/html/tlc/html/about/trip\_record\_data.shtml

Please zip all code and outputs and email them back to (Recruiter)

1. Download the data, load it into your favorite statistical programing software or database.

Report the number of rows and columns that you've loaded.

2. Visualize trip distance by time of day in any way you see fit, any observations?

3. What are the most popular pickup locations on weekdays vs weekend?

4. I want to know where I can most easily get a cab. Recommend a pick up spot where i can

find a cab given my lat long

-------------------------------------------------------------------------------------------------------------------------------

**My Answer to the Assessment Has Two Parts**:

For question 1-3, they are answered in *Green Taxi Case Study Part1.* I have attached the R markdown file in the submission and host the result via the link below.

<http://rpubs.com/couyang/379824>

For question 4, the question is answered in *Green Taxi Case Study Part2.* I have attached the R markdown (shiny) file in the submission. This part has the feature of inputting the longitude and latitude and it recommends 20 pick-up spots based on the input. In the end, it also provide the closest pick up spot among the top 20.

***Summary for Question 1 to 3:***

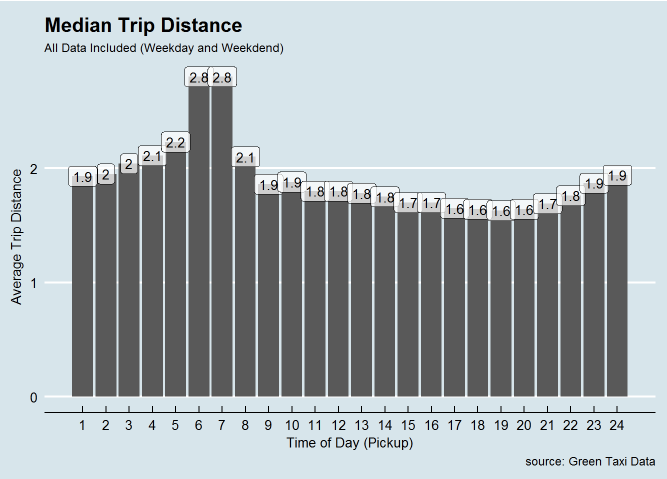
1. Download the data, load it into your favorite statistical programing software or database.

Report the number of rows and columns that you've loaded.

***Answer:*** *The first question can be answered by looking at the structure of the dataset. The dataset has 1510722 observations (rows) and 21 variables (columns).*

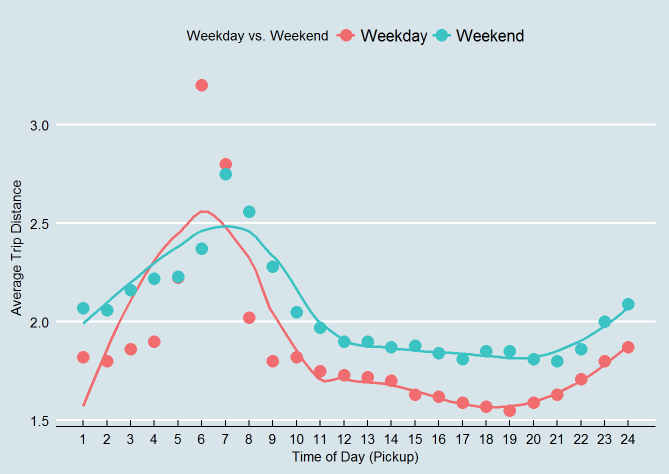
2. Visualize trip distance by time of day in any way you see fit, any observations?

***Answer:*** *From an initial look at the trip distance by time of day graph, the median trip distance is longer at 6 AM and 7 AM than any other time during the day. The difference between the two hours and the rest of the day is quite significant.*



*Same as the observation from the full dataset, 6 AM and 7 AM are still the two hour that people take taxi for the longest median trip distance. However, in this case, the median trip distance is slightly longer at 6 AM. At that day of the weekday, my assumption is most of the taxi usage is caused by the commute to go to work. Since people usually start work at 8 AM, one of the potential explanation is that people who live further from their work place tend to leave a bit earlier than the people who live close by. Moreover, it might also cause by that more people take taxi to work at 6 AM compared to 7 AM. Therefore, the median trip distance is slightly longer at 6 AM than 7 AM.*

*For the weekend, 7 AM and 8 AM turned out to be the two hours that have the longest trip distance. Additionally, the difference between the two hours and the rest of the day is not as significant as for the weekdays. My speculation is that people tend to sleep in, since they don’t have to go to work in the early morning. Therefore, the longest median trip distance is postponed one hour as well as the difference is not as significant as the weekdays.*



3. What are the most popular pickup locations on weekdays vs weekend?

***Answer:***

