**1. Business Context**

The New York City Taxi and Limousine Commission (TLC), created in 1971, is the agency responsible for licensing and regulating New York City's Medallion (Yellow) taxi cabs, for-hire vehicles (community-based liveries, black cars and luxury limousines), commuter vans, and paratransit vehicles. Over 200,000 TLC licensees complete approximately 1,000,000 trips each day. To operate for hire, drivers must first undergo a background check, have a safe driving record, and complete 24

hours of driver training. TLC-licensed vehicles are inspected for safety and emissions at TLC's Woodside Inspection Facility. It is a priority of the New York City Taxi and Limousine Commission to provide safe, reliable transportation options for all New Yorkers and to recognize and address the needs of our licensees. TLC supports and contributes to city-wide efforts of traffic safety, accessibility and technological improvements. Policy researchers at the TLC use data generated by licensees to observe changing trends in the industry and inform decisions made by agency and the City.

**2. Business Problem Understanding**

TLC licenses over 130,000 vehicles in New York City. Each vehicle receives comprehensive safety and emissions inspections by TLC and must be driven by TLC licensed drivers that have undergone background checks and passed TLC education requirements.

Green Taxis

* Green Taxis provide street hail service and prearranged service in northern Manhattan (above E 96th St and W 110th St) and in the outer boroughs.
* Green Taxis charge standard metered fares for all street hail trips. The price for prearranged trips are set by the base or smartphone app used to reserve the trip.
* Green Taxis are easily identified by their green color, taxi “T” markings, and license numbers on the roof and sides of the vehicle.

As TLC is license issuing authority, it needs to determine the number of new licenses to be released every year. For that purpose, its thinking of following a data driven analysis approach this time. It’s expecting your help to do a detailed analysis of the green taxi trip data for the year 2019

**3. Data Understanding**

For this analysis, the department is expecting you to explore the usage of Oracle Database, SQL Plus, Python for storage and querying the green taxi trip data accumulated for the year 2019. The data is available at the city portal [here](https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page). The relevant data dictionary can also be read from the same source at [here](https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf). Trip datasets contain the information about the each trip completed by green taxis during each day of each month of year 2019. Each trip record contains various details of a trip like vendor id, pickup and drop time, passenger count, fare details, pickup and drop points etc.

**4. Expected Outcomes**

You are expected to the following activities and find out the answers to following questions.

1. Use python to load the all the 2019 files to an Oracle Database table
2. Create the table structure with appropriate data types before loading data
3. Use Taxi Zone Lookup table from the city portal site to for drop and pickup location names
4. Answer the following questions
   1. Find the month wise trip count, average distance and average passenger count from the trips completed by green taxis in 2019.
   2. Find out the five busiest routes served by the green taxis during 2019. The name of start and drop points to be provided.
   3. What are the top 3 busiest hours of the day for the taxis?
   4. What is the most preferred way of payment used by the passengers?
   5. Write a PL/SQL block to read through each record and update ehail\_fee to 0.5 (capture the time taken for execution)
   6. Write a normal update statement to update ehail\_fee to 0.75 (capture the time taken for execution)
   7. Identify the time taken by e and f and provide your analysis on why each step took more/less time compared to other
5. **Submission**

A document file containing

* 1. A brief description of your understanding of data
  2. Any anomalies you identified in the provided dataset and a brief description of how you identified them and why do you think they are anomalies
  3. Queries you have written including the DDLs
  4. Results to the queries above

**Drop me a note at** [**Naveen.arigapudi@kpipartners.com**](mailto:Naveen.arigapudi@kpipartners.com) **if you have any questions**