Problem Statement

The New York City Taxi & Limousine Commission (TLC) has provided a dataset of trips made by the taxis in the New York City. The detailed trip-level data is more than just a vast list of taxi pickup and drop off coordinates.

The records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations (location coordinates of the starting and ending points), trip distances, itemized fares, rate types, payment types, driver-reported passenger counts etc. The data used was collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP).

This dataset was created by aggregating the aforementioned records. It provides precise location coordinates for where the trip started and ended, timestamps for when the trip started and ended, plus a few other variables including fare amount, payment method, and distance travelled.

The purpose of this dataset is to get a better understanding of the taxi system so that the city of New York can improve the efficiency of in-city commutes. Several exploratory questions can be asked about the travelling experience for passengers.

In this assignment, we ONLY consider the data of yellow taxis for November and December of the year 2017.

The dataset has been placed in the HDFS storage of the lab. The path to the data files is as follows:

'/common_folder/nyc_taxi_data/'

While performing this analysis, you will inevitably make several assumptions. As long as you state these assumptions, you will be awarded marks.

A few pointers before you start the assignment:

- While creating the tables, it is mandatory to define the integers as int and floating points as double. Certain results may be affected if this is not followed and in that case, marks will not be awarded.
- The solution file must contain all the necessary commands to set up the environment before you start querying. These are covered during the course - they involve adding a JAR file and setting parameters of Hive for partitioning. If these commands are not present in your solution file, marks will be deducted
- Lastly, your code should be syntactically correct, concise and commented. Marks are reserved for the comments present along with every question mentioned in the problem statement - make sure you write these comments as you go along writing your queries.

Here are the questions you need to answer.

Basic Data Quality Checks

1. How many records has each TPEP provider provided? Write a query that summarises the number of records of each provider.
2. The data provided is for months November and December only. Check whether the data is consistent, and if not, identify the data quality issues. Mention all data quality issues in comments.
3. You might have encountered unusual or erroneous rows in the dataset. Can you conclude which vendor is doing a bad job in providing the records using different columns of the dataset? Summarise your conclusions based on every column where these errors are present. For example, There are unusual passenger count, i.e. 0 which is unusual.

HINT: Use the Data Dictionary provided to validate the data present in the records provided.

Before answering the below questions, you need to create a clean, ORC partitioned table for analysis. Remove all the erroneous rows.

IMPORTANT: Before partitioning any table, make sure you run the below commands.

SEThive.exec.max.dynamic.partitions=100000;
SET hive.exec.max.dynamic.partitions.pernode=100000;

Analysis-I

1. Compare the overall average fare per trip for November and December.
2. Explore the 'number of passengers per trip' - how many trips are made by each level of 'Passenger_count'? Do most people travel solo or with other people?
3. Which is the most preferred mode of payment?
4. What is the average tip paid per trip? Compare the average tip with the 25th, 50th and 75th percentiles and comment whether the 'average tip' is a representative statistic (of the central tendency) of 'tip amount paid'. Hint: You may use percentile_approx(DOUBLE col, p): Returns an approximate pth percentile of a numeric column (including floating point types) in the group.
5. Explore the 'Extra' (charge) variable - what fraction of total trips have an extra charge is levied?

Analysis-II

1. What is the correlation between the number of passengers on any given trip, and the tip paid per trip? Do multiple travellers tip more compared to solo travellers? Hint: Use CORR(Col_1, Col_2)
2. Segregate the data into five segments of ‘tip paid’: [0-5), [5-10), [10-15) , [15-20) and >=20. Calculate the percentage share of each bucket (i.e. the fraction of trips falling in each bucket).
3. Which month has a greater average ‘speed’ - November or December? Note that the variable ‘speed’ will have to be derived from other metrics. Hint: You have columns for distance and time.
4. Analyse the average speed of the most happening days of the year, i.e. 31st December (New year’s eve) and 25th December (Christmas) and compare it with the overall average.