BITS Pilani

BIG DATA ANALYTICS

Submission Date: 28th JUNE 2021 11.50 PM

Weightage: 10%

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. Learn about the differences in the types of TLC-licensed vehicles as well as tips on how to identify a properly licensed vehicle.

Yellow Taxis

- Yellow Taxis are the only vehicles licensed to pick up street-hailing passengers anywhere in NYC.
- Yellow Taxis charge standard metered fares for all street hail trips.
- Yellow Taxi smartphone apps can offer set, upfront pricing for trips booked through an app.
- Yellow Taxis are easily identified by their yellow color, taxi "T" markings, and medallion license numbers on the roof and sides of the vehicle.

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 yellow and green taxi trip data for the year 2019 and help them determining the number of licenses to be released for the year 2021.

3. Data Understanding

For this analysis, the department is expecting you to explore the usage of Apache Hive for the storage and querying the yellow and green taxi trip data accumulated for the year 2019. The data is available at the city portal here. The relevant data dictionary can also be read from the same source at here and here.

Trip datasets contain the information about the each trip completed by yellow and 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. Data preparation and Exploratory Data Analysis

You are supposed to make utilizations of all the appropriate data pre-processing techniques on the given data set. If required, make appropriate assumptions and make it explicitly known while using them in the query. Make appropriate selection of the attributes with sound justification for the same. The data set allows for several new combinations of attributes and attribute exclusions, or the modification of the attribute type (categorical, integer, or real) depending on the purpose of the analysis.

5. Expected Outcomes

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

1) Which vendor provides the most useful data?

2) Find the month wise trip count, average distance and average passenger count from the trips completed by yellow and green taxis in 2019. Summary visualizations will be preferred for better analysis.

3) Find out the five busiest routes served by the yellow and green taxis during 2019. The name of start and drop points to be provided.

4) What are the top 3 busiest hours of the day for the taxis?

5) What is the most preferred way of payment used by the passengers? What are the weekly trends observed for the methods of payments?

The results should consists of

a) A document file containing answer to the five questions based on the analysis that you have carried out earlier along with the supporting Hive queries that you have written to extract the answers. Also the document should also describe the preprocessing steps carried out on the data before moving it into Hive for the analysis purpose.

- Refer the document used while registering the groups. In case of discrepancies, write to me separately (copying all your group members) with subject line as "NS BDA Group <your_group_number>". email to TA
- Using the Canvas, only the first member of group (as listed in the above mentioned doc) has to upload the file. No submission over email will be considered.
- Name the document file in format like "Grp_<your_group_number>.doc" only. Don't add anything into the file names.
- Make sure that you upload the file well ahead of deadline. At last moments, we have seen several groups have faced issues while doing the submissions.

6. References

- Data set link
- Hive documentation

- Hive Tutorial
- Groups information