450128

Homework 3

1-Preparing the file:

a.Download the dataset from http://3.236.139.48/dataset/NYCParkingViolation.csv into your home directory.



b.Can you tell us how many columns do you see here? Use Linux Commands.



Seeing that we have 12 ‘,’ we have 13 columns as there are n+1 columns for ‘,’

c.Copy the file into your HDFS folder.



2-Create a table in Hive called your ID\_Parking. For example, for a student with SID as 111111, the table name is going to be 111111\_Parking. This table is going to include NYCParkingViolation, then, make it corresponding to the data.

A picture containing text

Description automatically generated

3-Load the file into the created table.



4-Find the three most violated plate numbers (plate number of highest 3 violated cars).

Graphical user interface, text

Description automatically generated

Table

Description automatically generated

5-Find the average vehicle year per car maker. Please ignore rows with vehicle year as 0 and consider car maker having at least 10 violations. Which Car maker is the youngest one by average?

A picture containing text

Description automatically generated

Table

Description automatically generated

KI/MO is the youngest Vehicle Maker

6-Find the number of color per car maker. Ignore the rows with null values. Hint: Here null is defined as “empty”. So, when you say col!=”” you are dropping rows with null values in col.

A picture containing text

Description automatically generated

Table

Description automatically generated

7-For the most violated car model, find the top 10 violation code:

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

8-Find the total number of violations per daytime (morning or afternoon). Hint: Look at the Violation Time columns which indicates the time of the day.

Graphical user interface, text, application

Description automatically generated

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

Both resulted in 785 results for morning and afternoon.

9-Find the registration state of the violators with more than average number of violations. Find the top 3 violator states. Hint: Hive/Impala does not support Having with Subqueries. So we have to write two separate queries to first find the average number of violations and then find registration state of the violators with more than that number.?

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

A picture containing table

Description automatically generated

10-Find the number of afternoon violations for Hondas and BMWs.

A picture containing text

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated