Assignment: Analysis with Hive

Introduction:

In this Assignment, you will be working with the car.csv dataset that you can download from https://www.kaggle.com/mirosval/personal-cars-classifieds or use the wget command provided which can directly load the dataset into HDFS.

This dataset has the classified records for several Eastern European countries over several years. Beware that the data is not “clean” and investigating and cleaning the data is an important part of the Assignment.

Problem Background

You are the data analyst at a large investment firm that is contemplating to invest in a used car business. Your task is to provide data driven advice to the stakeholders, that will enable them to make a sound investment decision. Failure to make the best decision may result in large financial consequences and irreversible damage to the company reputation and brand.
Your manager has instructed you to use the cars.csv dataset, because the veracity of this data has been established.

Tasks:

Use the cars dataset, answer the following questions in ***Apache Hive.***

1. Write a Hive query to create a table called **used_cars** from data. Use a schema that is appropriate for the column headings

2. Look at the date column of the table used_cars. Why does the date column have all NULL values?

3. Create a table such that the date column is read correctly based on the format in the dataset (see posted text file)

4. Write Hive queries to see how many missing values you have in each attribute? Based on the results, document how many missing values in each column we have. Especially, mention those columns with more than 50% missing values.
5. Group the price column and count the number of unique prices. Do you notice if there is a single price that is repeating across the ads?
6. Write a Hive query to create a new table called **clean_used_cars** from **used_cars** with the following conditions:
    - Drop the columns with more than 50% missing values
    - The manufacture year between 2000 and 2017 including 2000 and 2017
    - Both maker and model exist in the row
    - The price range is from 3000 to 2000,000 (3000 ≤ price ≤ 2000,000)
    - Remove any price you singled out in Step 3 (ie a price that repeats too frequently for a random set of ads).

7. Write a Hive query to find how many records remained **clean_used_cars**
8. Write a Hive query to find the make and model for the cars with the top 10 **highest average price**
9. Write a Hive query to find the make and model for the cars with the top 10 **lowest average price**
10. Write a Hive query to recommend top five make and model for <u>**Economic**</u> segment customers (Top five manufacturers in the 3000 to 20,000 price range;3000≤price<20,000) - based on the top **average price**
11. Write a Hive query to recommend top five make and model for <u>**Intermediate**</u> segment customers (Top five manufacturers in the 20,000 to 300,000 price range; 3000≤price<20,000) - based on the top **average price**
12. Write a Hive query to recommend the top five make and model for the <u>**Luxury**</u> segment customers (Top five manufacturers in the 300,000 to 2000,000 price range; 300,000≤price<2000,000) - based on the top **average price**

Deliverables:

Based on the answers in the previous section, write a formal report to a group of investors outlining your findings and what cars you recommend. You can also expand on the above analysis and do your own queries if you prefer. Remember that your audience is not a technical audience and as such, the report should focus on analysis rather than technical details.

Any technical details can be provided in an appendix (example, loading of data into Hive etc). You can use external tools for visualization of any queries (example Excel, PowerBI, Tableau) but the actual analysis must be done in Hive.

Hints:

- Formal reports should be written in third person, do not use words such as “I”, “we”, etc.
- A formal report should be structured with an introduction, a plan and a body. Do not just “jump” into the report without any context.
- A formal report **is not** just listing the queries in the Tasks section. Such reports will receive poor marks. You can put the screenshot of your tasks in an Appendix along with any analysis you have done.
- You should be aware of your audience and write a report the is curtailed to their knowledge. Investors do not know about Hive, HDFS, SQL, they want the analysis. You can include any technical details in appendices.
- Reports should be free of grammar and spelling mistakes.
- Reports should have a good “flow”. In other words, your audience should be excited about reading the report.