# ***Assignment 1 – Data Analytics with Hive***

## *Introduction*

In this project I will be working with the car.csv dataset that is downloaded from https://www.kaggle.com/mirosval/personal-cars-classifieds

## *Setup the Database*

***Code***

```
hive> CREATE DATABASE cars_db;

hive> USE cars_db;
```

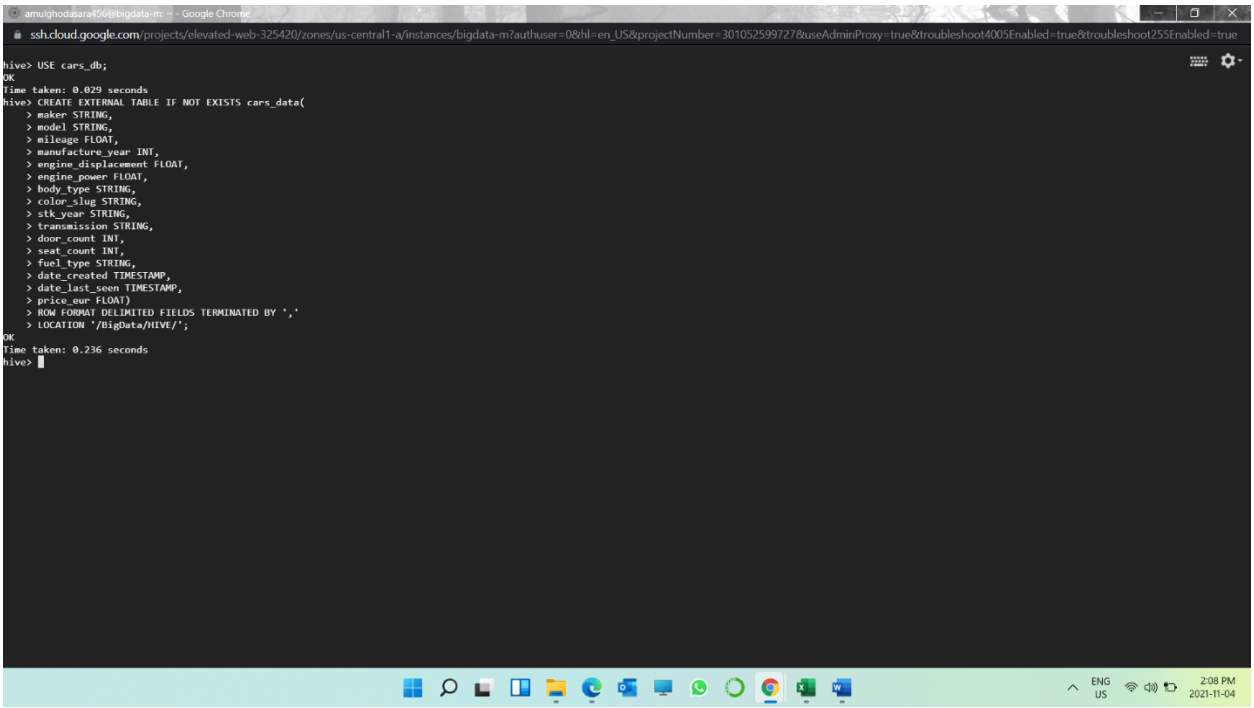
## *Create a table*

***Code***

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS cars_data(

    > maker STRING,

    > model STRING,

    > mileage FLOAT,

    > manufacture_year INT,

    > engine_displacement FLOAT,

    > engine_power FLOAT,

    > body_type STRING,

    > color_slug STRING,

    > stk_year STRING,

    > transmission STRING,

    > door_count INT,

    > seat_count INT,
```

```
> fuel_type STRING,
> date_created TIMESTAMP,
> date_last_seen TIMESTAMP,
>  price_eur FLOAT)
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
> LOCATION '/BigData/hive/';
```

### *Screenshot*

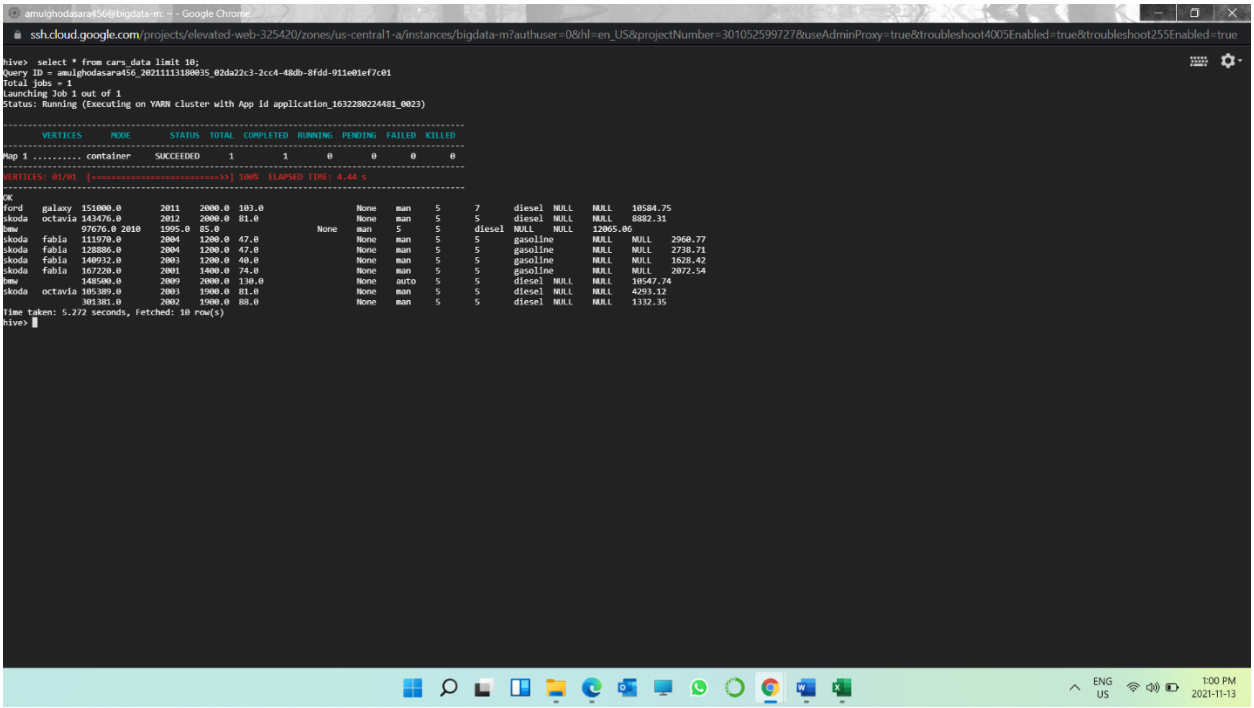



## *Load Data: -*

### *Code*

hive> select * from cars_data limit 10;

**Screenshot**




# Data Cleaning: -

## Cleaning 1

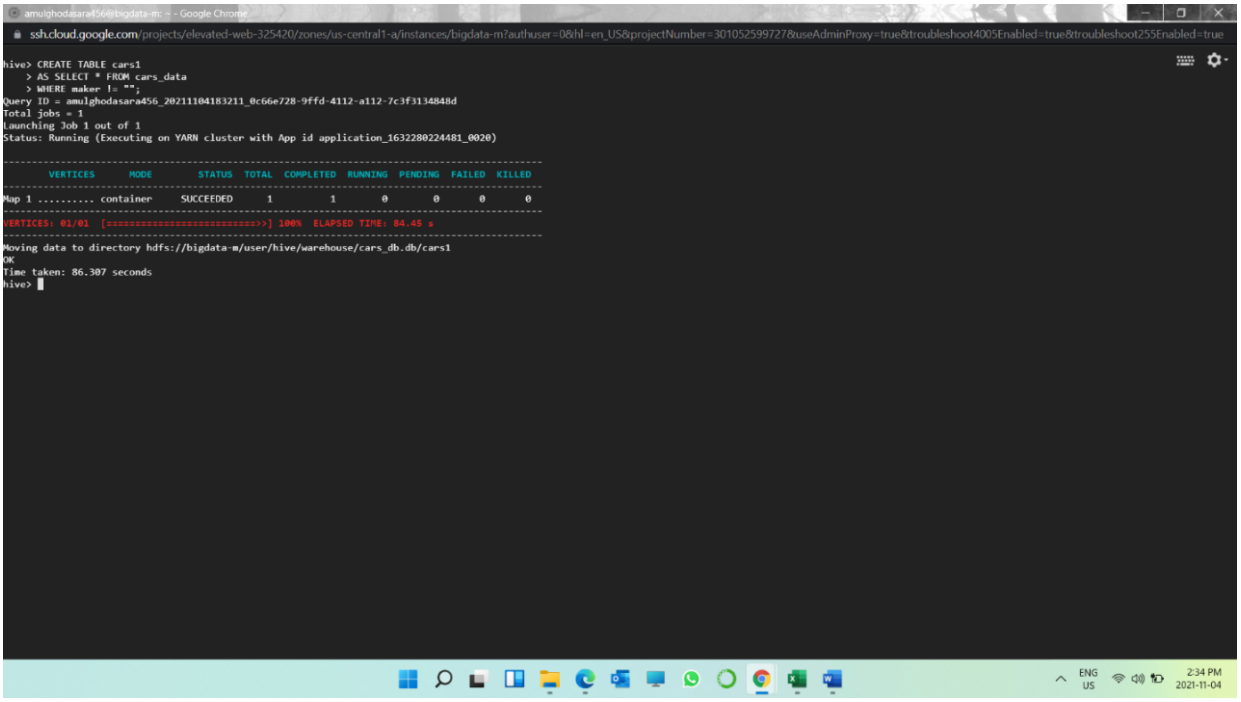
In cleaning 1, In the data set, there are many null values in maker column. Which will affect the result of analysis, So I decided to remove that null value for better result. We need maker column in the analysis questions that why I choose this step.

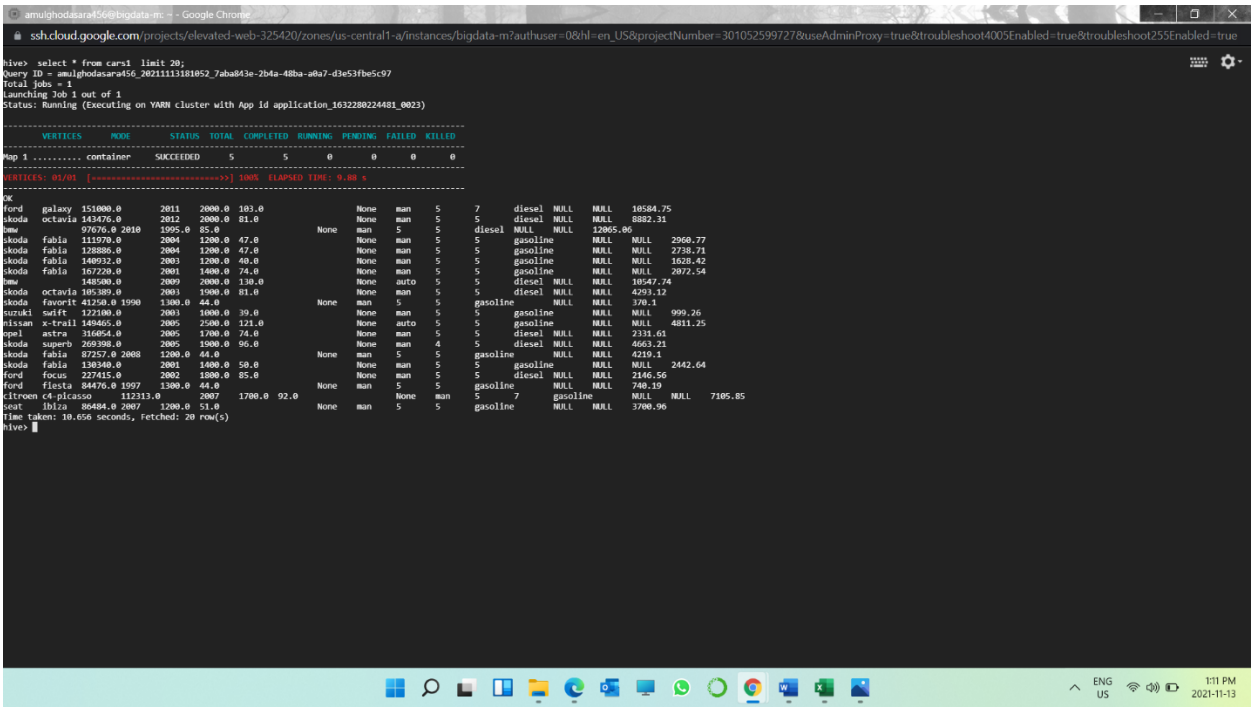
Here, I create new table which name in cars1. For cleaning, I use cars_data set for cleaning the maker column and transfer that data into new table which is called cars1. Finally, I displayed first 20 values.

**Code**

```
hive> CREATE TABLE cars1
> AS SELECT * FROM cars_data
> WHERE maker != "";
```

**Screenshot**





```
hive>  select * from cars1  limit 20;
Query ID = amulghodasara456_20211113181052_7aba843e-2b4a-48ba-a0a7-d3e53fbe5c97
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1632280224481_0023)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 9.00 s
--------------------------------------------------------------------------------
OK
ford    galaxy  151000.0        2011    2000.0  103.0                   None    man     5       7       diesel  NULL    NULL    10584.75
skoda   octavia 143476.0        2012    2000.0  81.0                    None    man     5       5       diesel  NULL    NULL    8882.31
bmw             97676.0 2010    1995.0  85.0                    None    man     5       5       diesel  NULL    NULL    12065.06
skoda   fabia   111970.0        2004    1200.0  47.0                    None    man     5       5       gasoline        NULL    NULL    2960.77
skoda   fabia   128886.0        2004    1200.0  47.0                    None    man     5       5       gasoline        NULL    NULL    2738.71
skoda   fabia   140932.0        2003    1200.0  40.0                    None    man     5       5       gasoline        NULL    NULL    1628.42
skoda   fabia   167220.0        2001    1400.0  74.0                    None    man     5       5       gasoline        NULL    NULL    2072.54
bmw             148500.0        2009    2000.0  130.0                   None    auto    5       5       diesel  NULL    NULL    10547.74
skoda   octavia 105389.0        2003    1900.0  81.0                    None    man     5       5       diesel  NULL    NULL    4293.12
skoda   favorit 41250.0 1990    1300.0  44.0                    None    man     5       5       gasoline        NULL    NULL    370.1
suzuki  swift   122100.0        2003    1000.0  39.0                    None    man     5       5       gasoline        NULL    NULL    999.26
nissan  x-trail 149465.0        2005    2500.0  121.0                   None    auto    5       5       gasoline        NULL    NULL    4811.25
opel    astra   316054.0        2005    1700.0  74.0                    None    man     5       5       diesel  NULL    NULL    2331.61
skoda   superb  260398.0        2005    1900.0  96.0                    None    man     4       5       diesel  NULL    NULL    4663.21
skoda   fabia   87257.0 2008    1200.0  44.0                    None    man     5       5       gasoline        NULL    NULL    4219.1
skoda   fabia   130340.0        2001    1400.0  50.0                    None    man     5       5       gasoline        NULL    NULL    2442.64
ford    focus   227415.0        2002    1800.0  85.0                    None    man     5       5       diesel  NULL    NULL    2146.56
ford    fiesta  84476.0 1997    1300.0  44.0                    None    man     5       5       gasoline        NULL    NULL    740.19
citroen c4-picasso      112313.0        2007    1700.0  92.0                    None    man     5       7       gasoline        NULL    NULL    7105.85
seat    ibiza   86484.0 2007    1200.0  51.0                    None    man     5       5       gasoline        NULL    NULL    3700.96
Time taken: 10.656 seconds, Fetched: 20 row(s)
hive>
```

### *Cleaning 2*

In cleaning 1, In the data set, there are many null values in model column. Which will affect the result of analysis, So I decided to remove that null value for better result. We need model column in the analysis questions that why I choose this step.

Here, I create new table which name in cars2. For cleaning, I used cars1 (In which, maker column is already cleaned) set for cleaning the model column and transfer that data into new table which is called cars2. Lastly, I displayed first 30 values.

### *Code*

```
hive> CREATE TABLE cars2
> AS SELECT * FROM cars1
> WHERE model != "";
```

## *Screenshot*

amulghodasara456@bigdata-m: ~ - Google Chrome

ssh.cloud.google.com/projects/elevated-web-325420/zones/us-central1-a/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=301052599727&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true

```
hive> CREATE TABLE cars2
    > AS SELECT * FROM cars1
    > WHERE model != "";
Query ID = amulghodasara456_20211104183447_920bf072-cbfe-488b-975b-3e53355be8e3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1632280224481_0020)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 19.04 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cars_db.db/cars2
OK
Time taken: 20.044 seconds
hive>
```

amulghodasara456@bigdata-m: ~ - Google Chrome

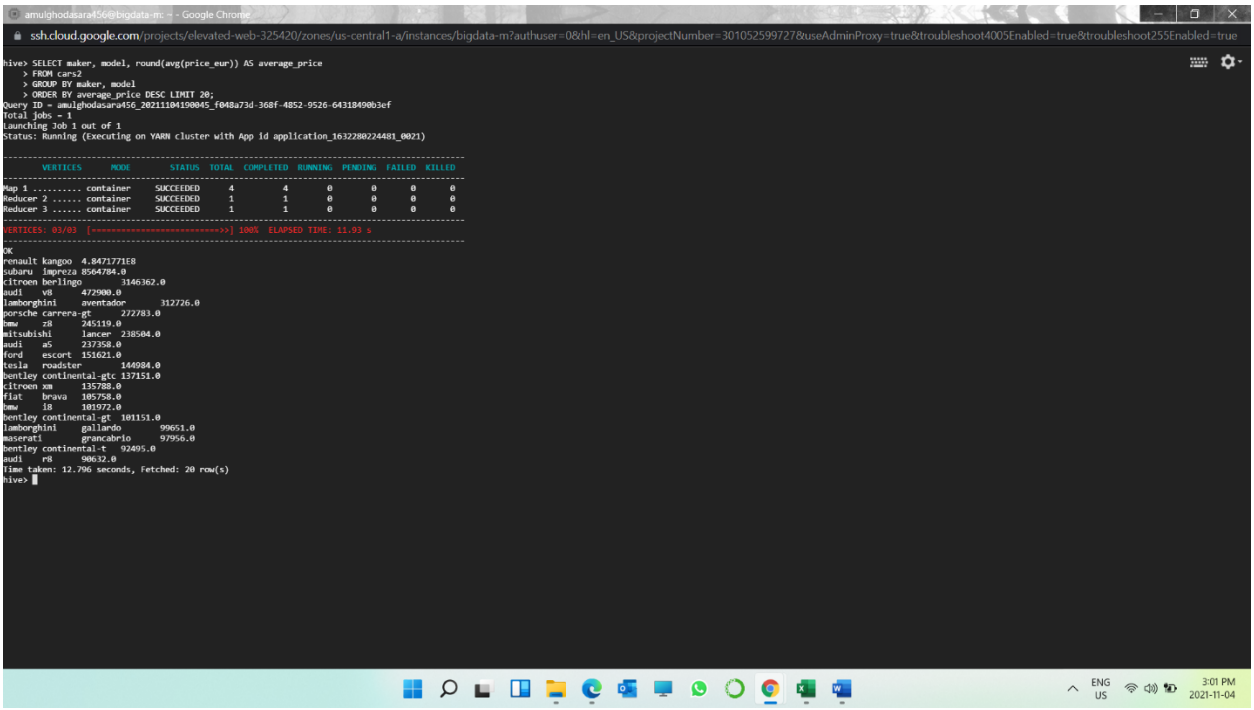
ssh.cloud.google.com/projects/elevated-web-325420/zones/us-central1-a/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=301052599727&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true

```
hive> select * from cars2 limit 30;
Query ID = amulghodasara456_20211113183123_9a6f8966-07cc-48a9-b235-f7a1abb8f28f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1632280224481_0026)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 8.64 s
--------------------------------------------------------------------------------
OK
volkswagen      caddy   152686.0        2012    1598.0  75.0    other                   man     NULL    2               NULL    NULL    8568.0
volkswagen      golf-sportsvan  0.0     NULL    1395.0  92.0    other                   auto    5       5               NULL    NULL    25610.0
volkswagen      caddy   85250.0 2012    1598.0  75.0    other           man     NULL    2       NULL    NULL    9758.0
volkswagen      caddy   152686.0        2012    1598.0  75.0    other                   man     NULL    2               NULL    NULL    8568.0
volkswagen      multivan        268000.0        2003    2460.0  128.0   other           2016    man     5       7               NULL    NULL    8999.0
toyota  celica  182000.0        2001    1795.0  141.0   other           man     3       4               NULL    NULL    4400.0
opel    corsa   208831.0        1994    1389.0  44.0    other           2016            2       NULL            NULL    NULL    500.0
volkswagen      golf    64000.0 2010    1397.0  77.0    other           2017    man     5       5               NULL    NULL    9900.0
volkswagen      caddy   141423.0        2012    1598.0  75.0    other                   man     NULL    2               NULL    NULL    8687.0
volkswagen      caddy   141423.0        2012    1598.0  75.0    other                   man     NULL    2               NULL    NULL    8687.0
peugeot 407     162000.0        2005    1997.0  100.0   other           man     4       5               NULL    NULL    3200.0
audi    a5      220000.0        1999    2496.0  110.0   other           man     5       5               NULL    NULL    17000.0
audi    a1      32000.0 2011    1390.0  90.0    other           2017    auto    3       4       NULL    NULL    9700.0
mini    cooper  121301.0        2007    1598.0  88.0    other           2018    man     3       4               NULL    NULL    7490.0
fiat    500     41500.0 2008    1242.0  51.0    other           man     3       4       NULL    NULL    6700.0
volkswagen      polo    1.0     NULL    1197.0  66.0    other           auto    5       NULL            NULL    NULL    16150.0
ford    fusion  125200.0        2004    1596.0  74.0    other           2017    man     5       5               NULL    NULL    2750.0
volkswagen      golf    38000.0 2012    1598.0  77.0    other           man     2       4               NULL    NULL    15500.0
mini    one     99000.0 2010    1560.0  66.0    other           2017    man     3       4       NULL    NULL    8200.0
volkswagen      golf    4900.0  2015    1984.0  221.0   other           auto    NULL    NULL            NULL    NULL    36590.0
citroen c5      161000.0        2010    1560.0  80.0    other           2017    man     4       5               NULL    NULL    5350.0
nissan  double-cab      130300.0        2008    2488.0  126.0   other           man     4       5               NULL    NULL    14000.0
volkswagen      sharan  200000.0        1999    1781.0  110.0   other           2016    man     5       7               NULL    NULL    2200.0
audi    a4      250000.0        1996    NULL    150.0   other           man     NULL    NULL            NULL    NULL    1100.0
renault clio    28000.0 2014    NULL    147.0   other           NULL    NULL            NULL    NULL    16500.0
audi    a3      180000.0        2001    1595.0  75.0    other           man     5       5               NULL    NULL    2950.0
opel    astra   121500.0        1999    1598.0  74.0    other           2018    man     5       5               NULL    NULL    1700.0
seat    leon    173000.0        2003    1597.0  77.0    other           man     5       5               NULL    NULL    1490.0
ford    fiesta  125000.0        2008    1399.0  50.0    other           man     5       5               NULL    NULL    5995.0
citroen c3      161000.0        2003    NULL    54.0    other           man     NULL    NULL            NULL    NULL    1800.0
Time taken: 9.443 seconds, Fetched: 30 row(s)
hive>
```

# *Analysis (Questions): -*

## Q1. What is the relationship between car makes, models and price?

### *Analysis*




According to my analysis, different maker and their model have different price even if their maker and model are same because of the condition of that perticular car. Here, I used average price of car with group by maker and model. In this dataset, there are many cars have same maker & model with different price and other variables. For instance, Lamborgini gallardo ( model ) is considering luxuries car but here its price in only 99651 EUR while ford escort ( model ) is not that much luxuries but its price is much higher than , Lamborgini gallardo ( model ).

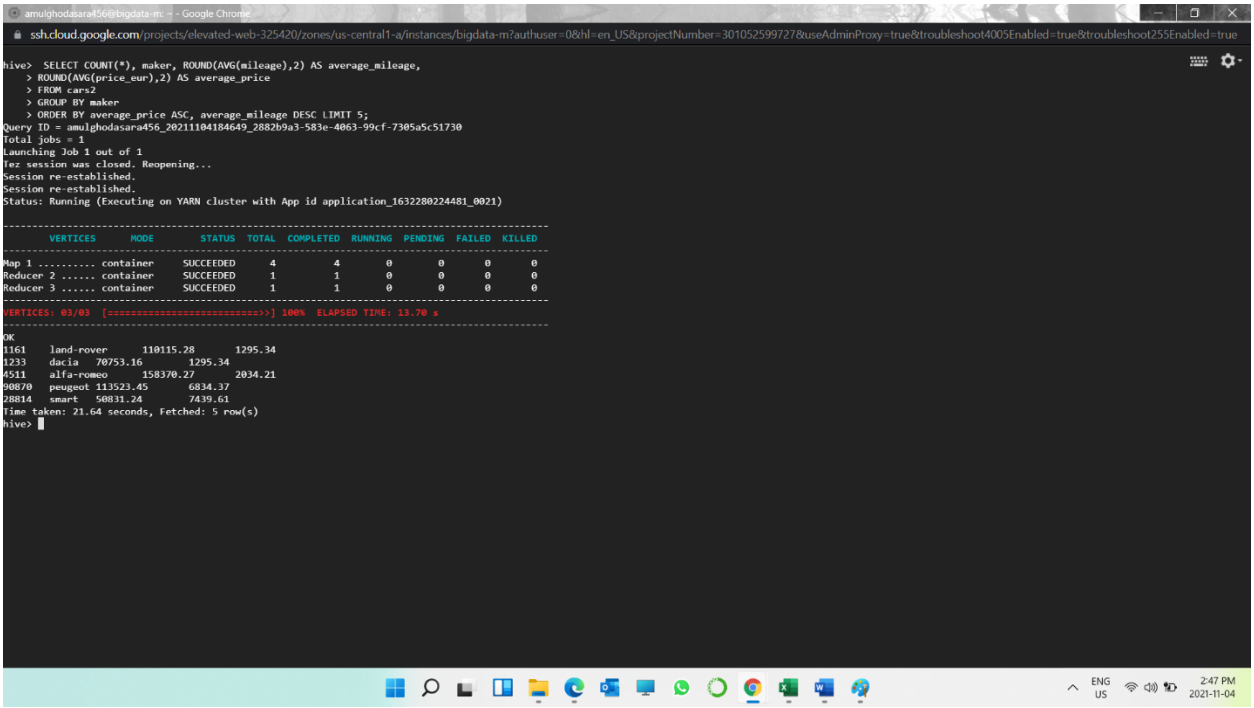
Overall, there are not strong relation between maker, model and price but other variables effects also.

### *Method*

For this research, I select maker, model and price_eur and find its average from cars2 table. After that, I group by them with maker and model and order by them with respect with average price. Finally, displey them with limit 20 so, it shows first 20 results.

## Q2. What are the top five vehicle manufacturers would you recommend? Why?

### *Analysis*




According to my research, if I want to recommend top five car manufactures, I will focus on its maker, milage & price. Here, I used average milage and price for my analysis.

Conclusion, I would recommend Land-Rover, Dacia, Alfa-Romeo, Peugeot, Smart manufactures. There is impotent reason behind this which is their higher milage and lower price than others maker. If any car has higher milage, after long time it will consider as cost effective and beneficial to the environment.

### *Method*

For this analysis, I select maker, price_eur and milage get its average to find top 5 car makers from cars2 table. Firstly, Group by them in respect with maker and order by them average milage and average price. Lastly, display the top five car makers.

**Q3. Does fuel type have any impact on the car price? Explain.**

**Analysis**

```
hive> SELECT ROUND(AVG(price_eur),2) AS average_price, fuel_type
    > FROM cars2
    >  WHERE fuel_type <> ""
    >  GROUP BY fuel_type
    >  ORDER BY average_price DESC LIMIT 10;
Query ID = amulghodasara456_20211104185420_a2ac7fa8-993a-483c-8cb8-2d9c42cc51ec
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1632280224481_0021)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 12.09 s
--------------------------------------------------------------------------------
OK
16050.38        gasoline
11217.19        diesel
1295.34 cng
1295.34 electric
1295.34 lpg
Time taken: 12.955 seconds, Fetched: 5 row(s)
hive>
```

According to me, yes fuel types have huge impact on the vehicle price. Here, I perform analysis and found that different fuel type car's price is different. In generally, car is running on 5 different fuel which are Gasoline, Diesel, CNG, LPG, Electric. For this research, I used average price of car and fuel type.

Overall, LPG, CNG and electric car price is same and as lower as Gasoline and Diesel. Moreover, Gasoline car is expensive among all others. Average price of Gasoline car is 16050.38 EUR while Diesel have 11217.19 and CNG, LPG & Electric.

**Method**

Here, for analysis, firstly I select price_eur and gets it average and fuel type from cars2 table. Then, group by it with fuel type and order by with average price for final output.