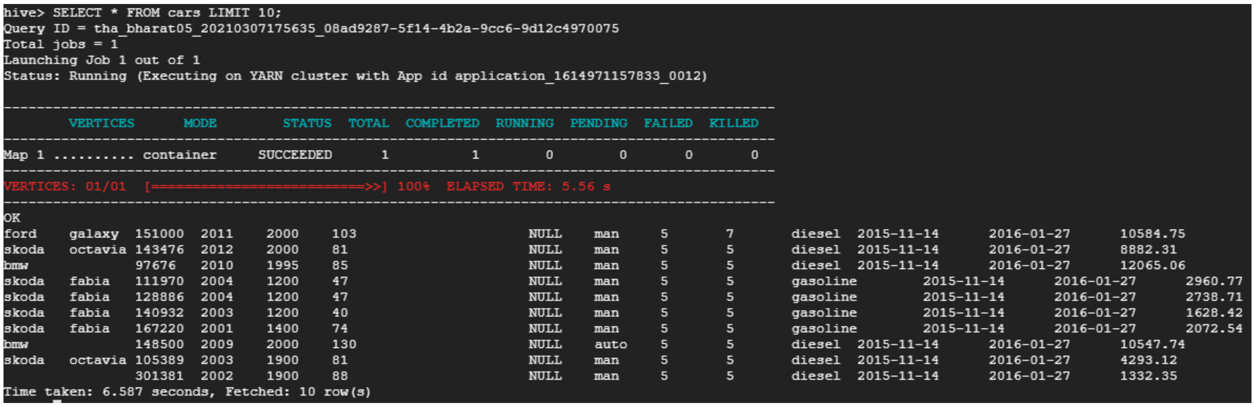
# Data Analytics with Hive

## Background

The dataset contains the car classified records for several Eastern European countries over several years.

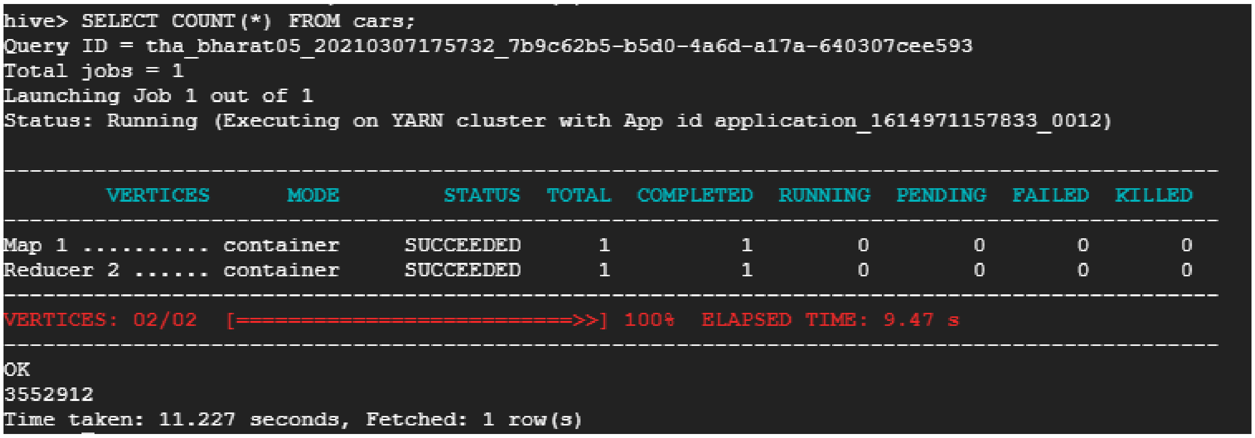
## Data Loading and Inspection

```
--check if data loaded correctly
SELECT * FROM cars LIMIT 10;
```



```
hive> SELECT * FROM cars LIMIT 10;
Query ID = tha_bharat05_20210307175635_08ad9287-5f14-4b2a-9cc6-9d12c4970075
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614971157833_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.56 s
----------------------------------------------------------------------------------------------
OK
ford    galaxy  151000  2011    2000    103                     NULL    man     5       7       diesel  2015-11-14      2016-01-27      10584.75
skoda   octavia 143476  2012    2000    81                      NULL    man     5       5       diesel  2015-11-14      2016-01-27      8882.31
bmw             97676   2010    1995    85                      NULL    man     5       5       diesel  2015-11-14      2016-01-27      12065.06
skoda   fabia   111970  2004    1200    47                      NULL    man     5       5       gasoline        2015-11-14      2016-01-27      2960.77
skoda   fabia   128886  2004    1200    47                      NULL    man     5       5       gasoline        2015-11-14      2016-01-27      2738.71
skoda   fabia   140932  2003    1200    40                      NULL    man     5       5       gasoline        2015-11-14      2016-01-27      1628.42
skoda   fabia   167220  2001    1400    74                      NULL    man     5       5       gasoline        2015-11-14      2016-01-27      2072.54
bmw             148500  2009    2000    130                     NULL    auto    5       5       diesel  2015-11-14      2016-01-27      10547.74
skoda   octavia 105389  2003    1900    81                      NULL    man     5       5       diesel  2015-11-14      2016-01-27      4293.12
                301381  2002    1900    88                      NULL    man     5       5       diesel  2015-11-14      2016-01-27      1332.35
Time taken: 6.587 seconds, Fetched: 10 row(s)
```

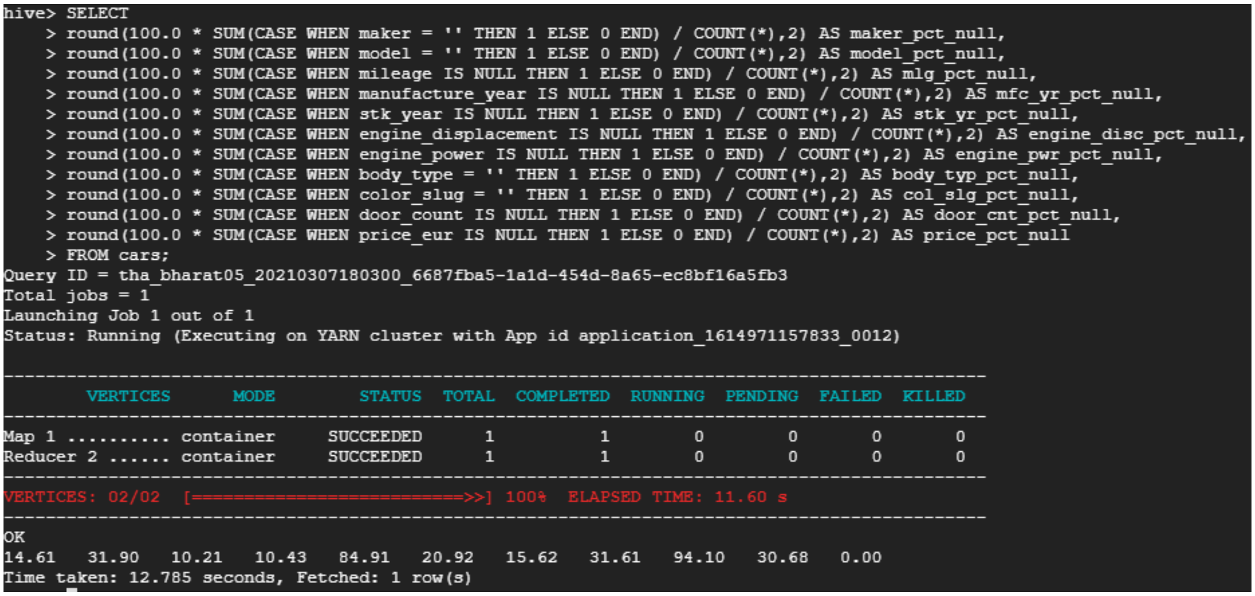
```
--Total row count
SELECT COUNT(*) FROM cars;
```



```
hive> SELECT COUNT(*) FROM cars;
Query ID = tha_bharat05_20210307175732_7b9c62b5-b5d0-4a6d-a17a-640307cee593
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614971157833_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.47 s
----------------------------------------------------------------------------------------------
OK
3552912
Time taken: 11.227 seconds, Fetched: 1 row(s)
```

1. There are 3552912 rows in the dataset.

```
-- Count Percent Null values in some columns

SELECT 100.0 * SUM(CASE WHEN maker = '' THEN 1 ELSE 0 END) / COUNT(*) AS
maker_pct_null,
100.0 * SUM(CASE WHEN model = '' THEN 1 ELSE 0 END) / COUNT(*) AS
model_pct_null,
100.0 * SUM(CASE WHEN mileage IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS
mlg_pct_null,
100.0 * SUM(CASE WHEN manufacture_year IS NULL THEN 1 ELSE 0 END) / COUNT(*)
AS mfc_yr_pct_null,
100.0 * SUM(CASE WHEN stk_year IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS
stk_yr_pct_null,
100.0 * SUM(CASE WHEN engine_displacement IS NULL THEN 1 ELSE 0 END) /
COUNT(*) AS engine_disc_pct_null,
100.0 * SUM(CASE WHEN engine_power IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS
engine_pwr_pct_null,
100.0 * SUM(CASE WHEN body_type = '' THEN 1 ELSE 0 END) / COUNT(*) AS
body_typ_pct_null,
100.0 * SUM(CASE WHEN color_slug = '' THEN 1 ELSE 0 END) / COUNT(*) AS
col_slg_pct_null,
100.0 * SUM(CASE WHEN door_count IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS
door_cnt_pct_null,
100.0 * SUM(CASE WHEN price_eur IS NULL THEN 1 ELSE 0 END) / COUNT(*) AS
price_pct_null
FROM cars;
```

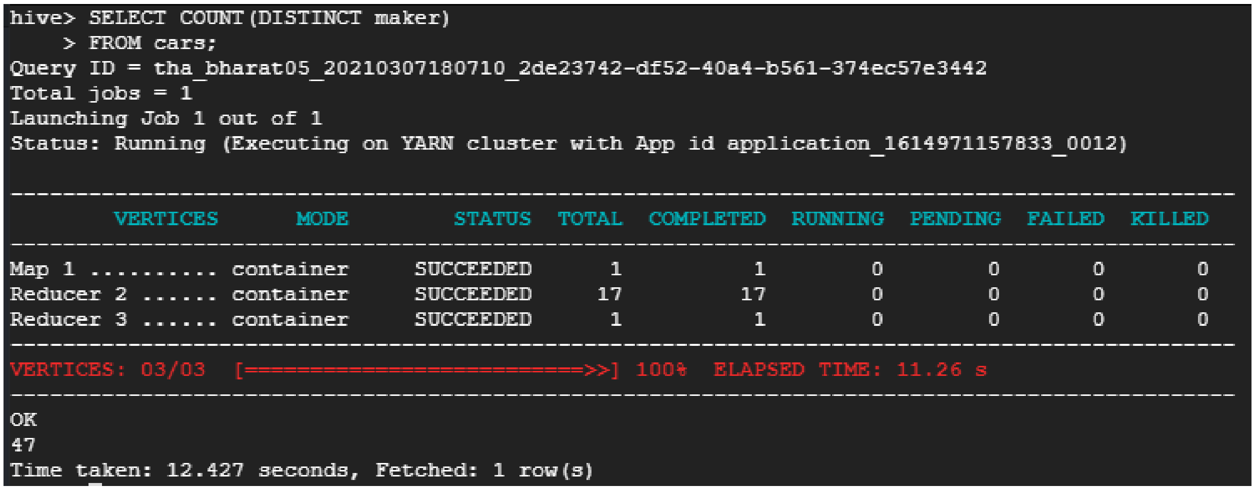


```
hive> SELECT
    > round(100.0 * SUM(CASE WHEN maker = '' THEN 1 ELSE 0 END) / COUNT(*),2) AS maker_pct_null,
    > round(100.0 * SUM(CASE WHEN model = '' THEN 1 ELSE 0 END) / COUNT(*),2) AS model_pct_null,
    > round(100.0 * SUM(CASE WHEN mileage IS NULL THEN 1 ELSE 0 END) / COUNT(*),2) AS mlg_pct_null,
    > round(100.0 * SUM(CASE WHEN manufacture_year IS NULL THEN 1 ELSE 0 END) / COUNT(*),2) AS mfc_yr_pct_null,
    > round(100.0 * SUM(CASE WHEN stk_year IS NULL THEN 1 ELSE 0 END) / COUNT(*),2) AS stk_yr_pct_null,
    > round(100.0 * SUM(CASE WHEN engine_displacement IS NULL THEN 1 ELSE 0 END) / COUNT(*),2) AS engine_disc_pct_null,
    > round(100.0 * SUM(CASE WHEN engine_power IS NULL THEN 1 ELSE 0 END) / COUNT(*),2) AS engine_pwr_pct_null,
    > round(100.0 * SUM(CASE WHEN body_type = '' THEN 1 ELSE 0 END) / COUNT(*),2) AS body_typ_pct_null,
    > round(100.0 * SUM(CASE WHEN color_slug = '' THEN 1 ELSE 0 END) / COUNT(*),2) AS col_slg_pct_null,
    > round(100.0 * SUM(CASE WHEN door_count IS NULL THEN 1 ELSE 0 END) / COUNT(*),2) AS door_cnt_pct_null,
    > round(100.0 * SUM(CASE WHEN price_eur IS NULL THEN 1 ELSE 0 END) / COUNT(*),2) AS price_pct_null
    > FROM cars;
Query ID = tha_bharat05_20210307180300_6687fba5-1a1d-454d-8a65-ec8bf16a5fb3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614971157833_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.60 s
----------------------------------------------------------------------------------------------
OK
14.61   31.90   10.21   10.43   84.91   20.92   15.62   31.61   94.10   30.68   0.00
Time taken: 12.785 seconds, Fetched: 1 row(s)
```

1. stk_year, color_slug have more than 80% null values.
2. model, body_type, and  door_count have over 30% blank values.
3. All the cars have price information.

## Exploratory Analysis Raw Dataset

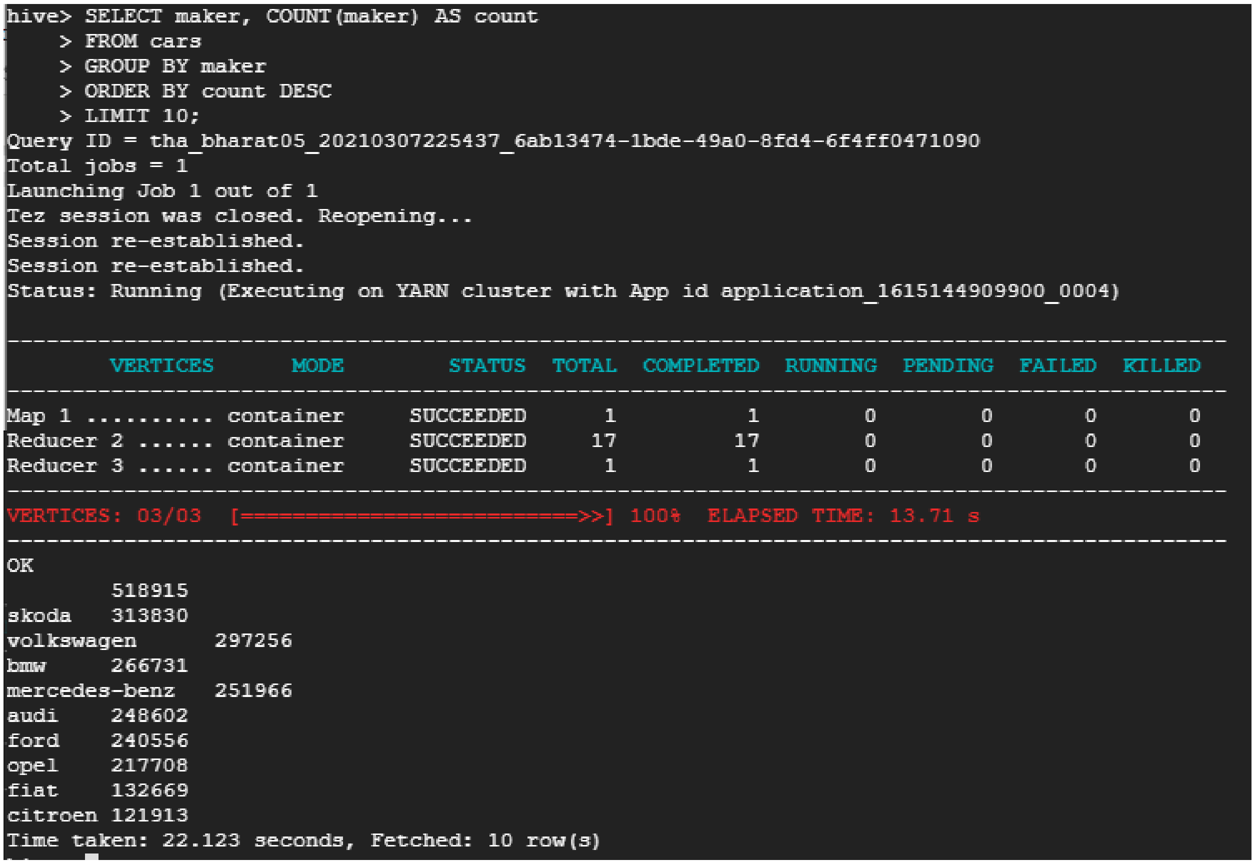
```
-- how many unique car makers
SELECT COUNT(DISTINCT maker)
FROM cars;
```



```
hive> SELECT COUNT(DISTINCT maker)
    > FROM cars;
Query ID = tha_bharat05_20210307180710_2de23742-df52-40a4-b561-374ec57e3442
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614971157833_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 11.26 s
----------------------------------------------------------------------------------------------
OK
47
Time taken: 12.427 seconds, Fetched: 1 row(s)
```

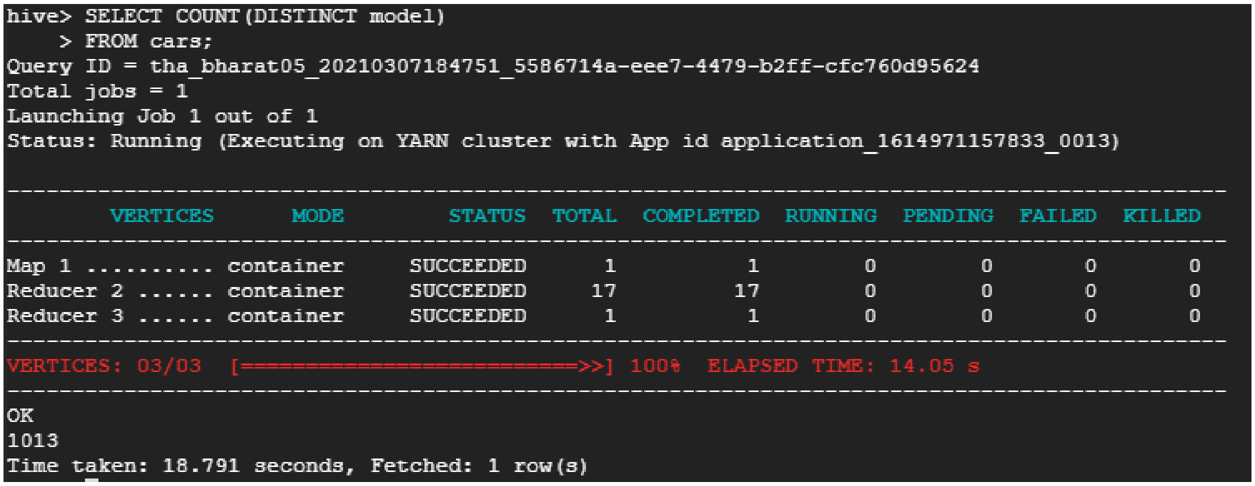
1. The dataset consists of car from total of 47 distinct car manufacturers.

```
-- Top 10 car makers
SELECT maker, COUNT(maker) AS count
FROM cars
GROUP BY maker
ORDER BY count DESC
LIMIT 10;
```



```
hive> SELECT maker, COUNT(maker) AS count
    > FROM cars
    > GROUP BY maker
    > ORDER BY count DESC
    > LIMIT 10;
Query ID = tha_bharat05_20210307225437_6ab13474-1bde-49a0-8fd4-6f4ff0471090
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1615144909900_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.71 s
----------------------------------------------------------------------------------------------
OK
        518915
skoda   313830
volkswagen      297256
bmw     266731
mercedes-benz   251966
audi    248602
ford    240556
opel    217708
fiat    132669
citroen 121913
Time taken: 22.123 seconds, Fetched: 10 row(s)
```

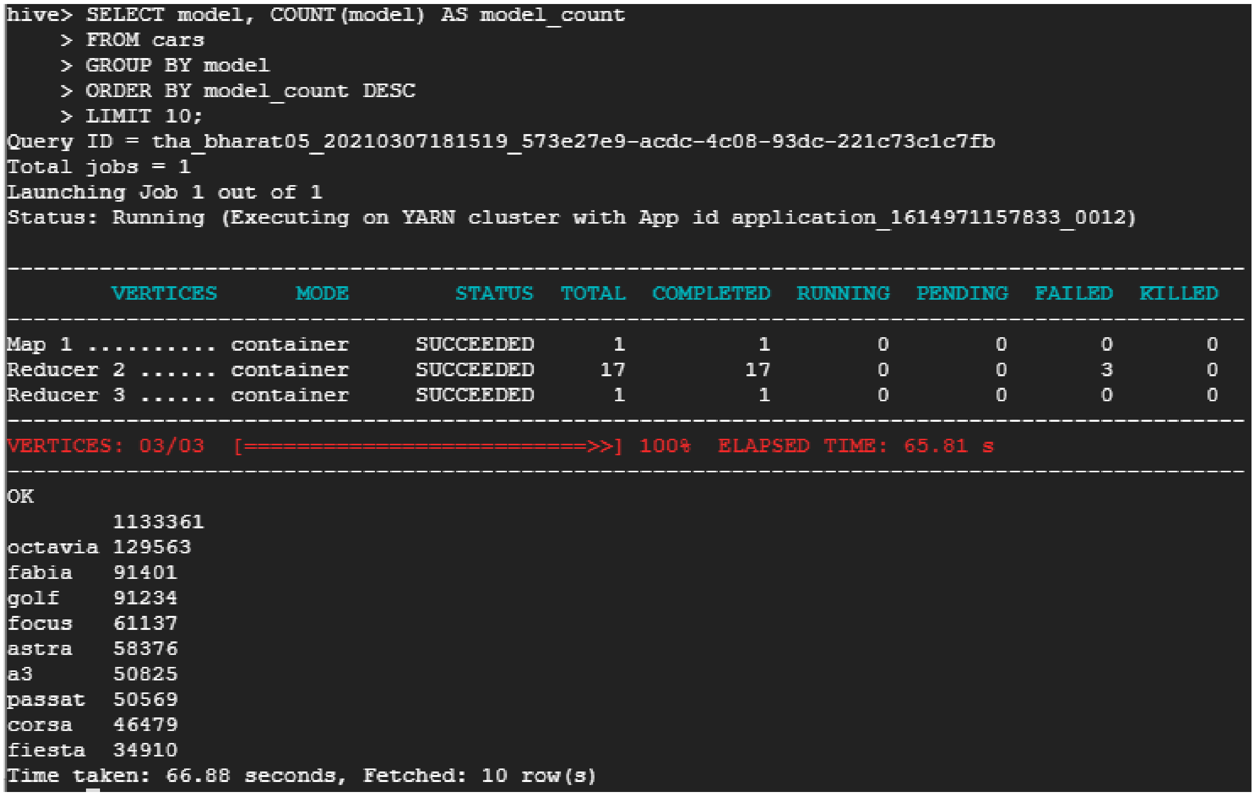
```
-- How many unique models
SELECT COUNT(DISTINCT model)
FROM cars;
```



```
hive> SELECT COUNT(DISTINCT model)
    > FROM cars;
Query ID = tha_bharat05_20210307184751_5586714a-eee7-4479-b2ff-cfc760d95624
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614971157833_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.05 s
----------------------------------------------------------------------------------------------
OK
1013
Time taken: 18.791 seconds, Fetched: 1 row(s)
```

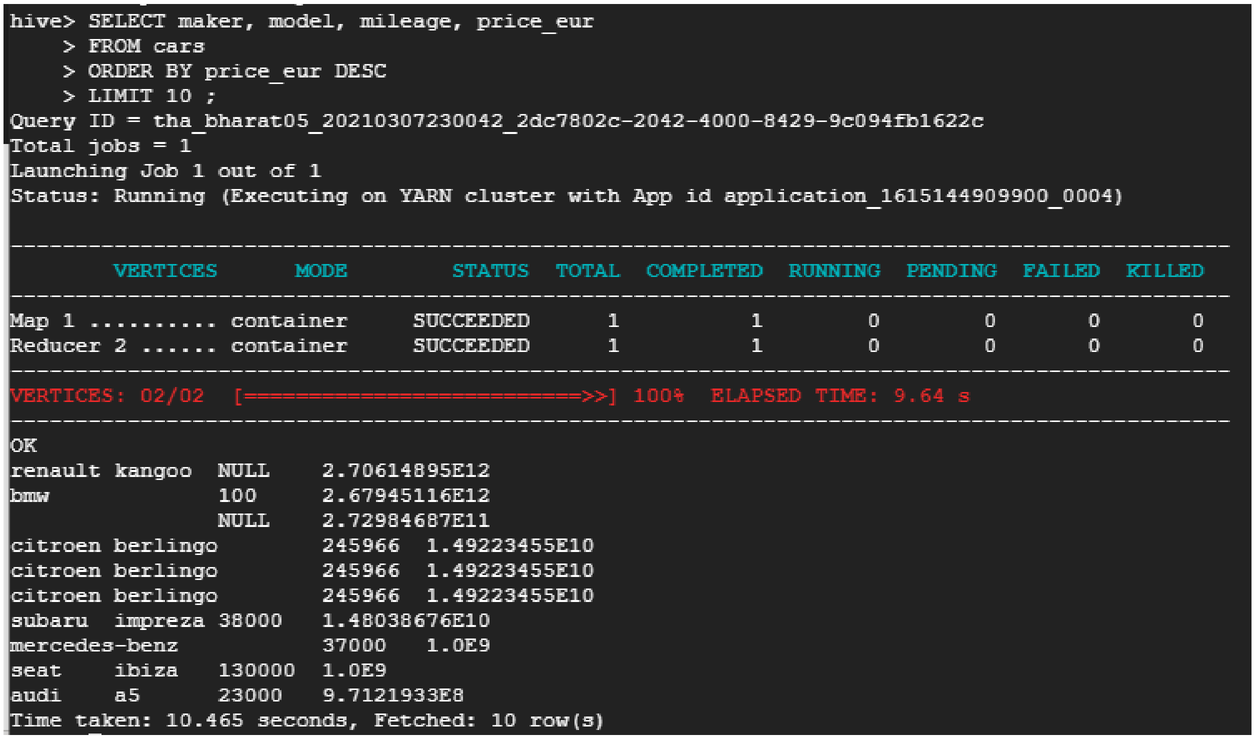
1. A total 1013 distinct car models.

```
-- Top 10 popular car models
SELECT model, COUNT(model) AS model_count
FROM cars
GROUP BY model
ORDER BY model_count DESC
LIMIT 10;
```



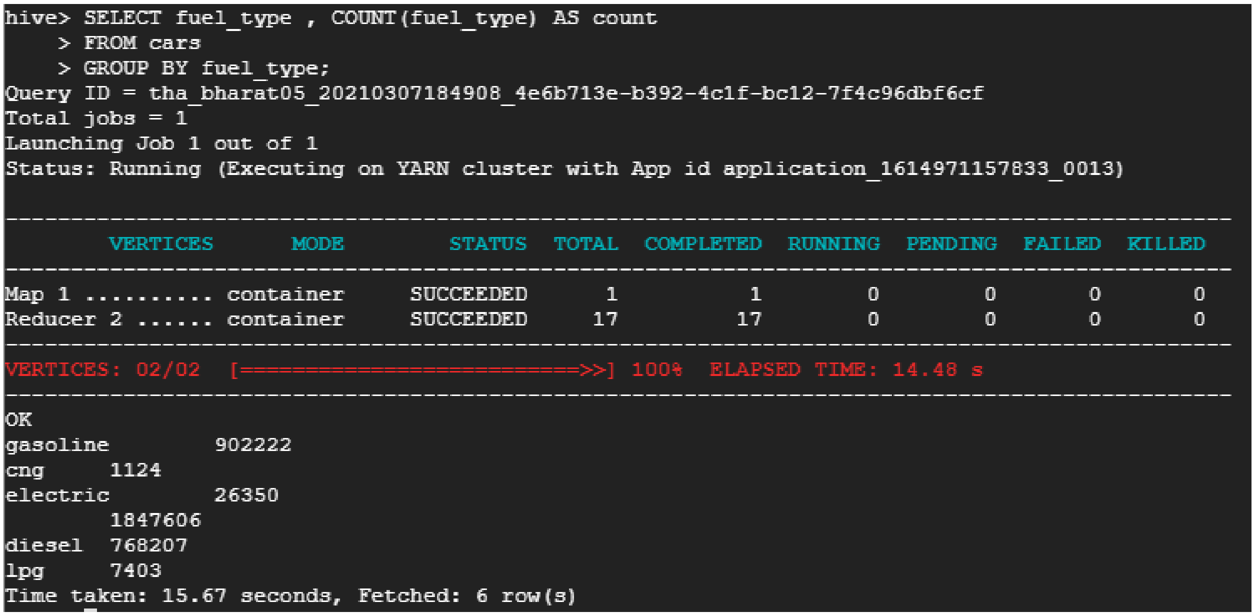
```
hive> SELECT model, COUNT(model) AS model_count
    > FROM cars
    > GROUP BY model
    > ORDER BY model_count DESC
    > LIMIT 10;
Query ID = tha_bharat05_20210307181519_573e27e9-acdc-4c08-93dc-221c73c1c7fb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614971157833_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       3       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 65.81 s
----------------------------------------------------------------------------------------------
OK
        1133361
octavia 129563
fabia   91401
golf    91234
focus   61137
astra   58376
a3      50825
passat  50569
corsa   46479
fiesta  34910
Time taken: 66.88 seconds, Fetched: 10 row(s)
```

```sql
--Top 10 most expensive cars
SELECT maker, model, mileage, price_eur
FROM cars
ORDER BY price_eur DESC
LIMIT 10;
```



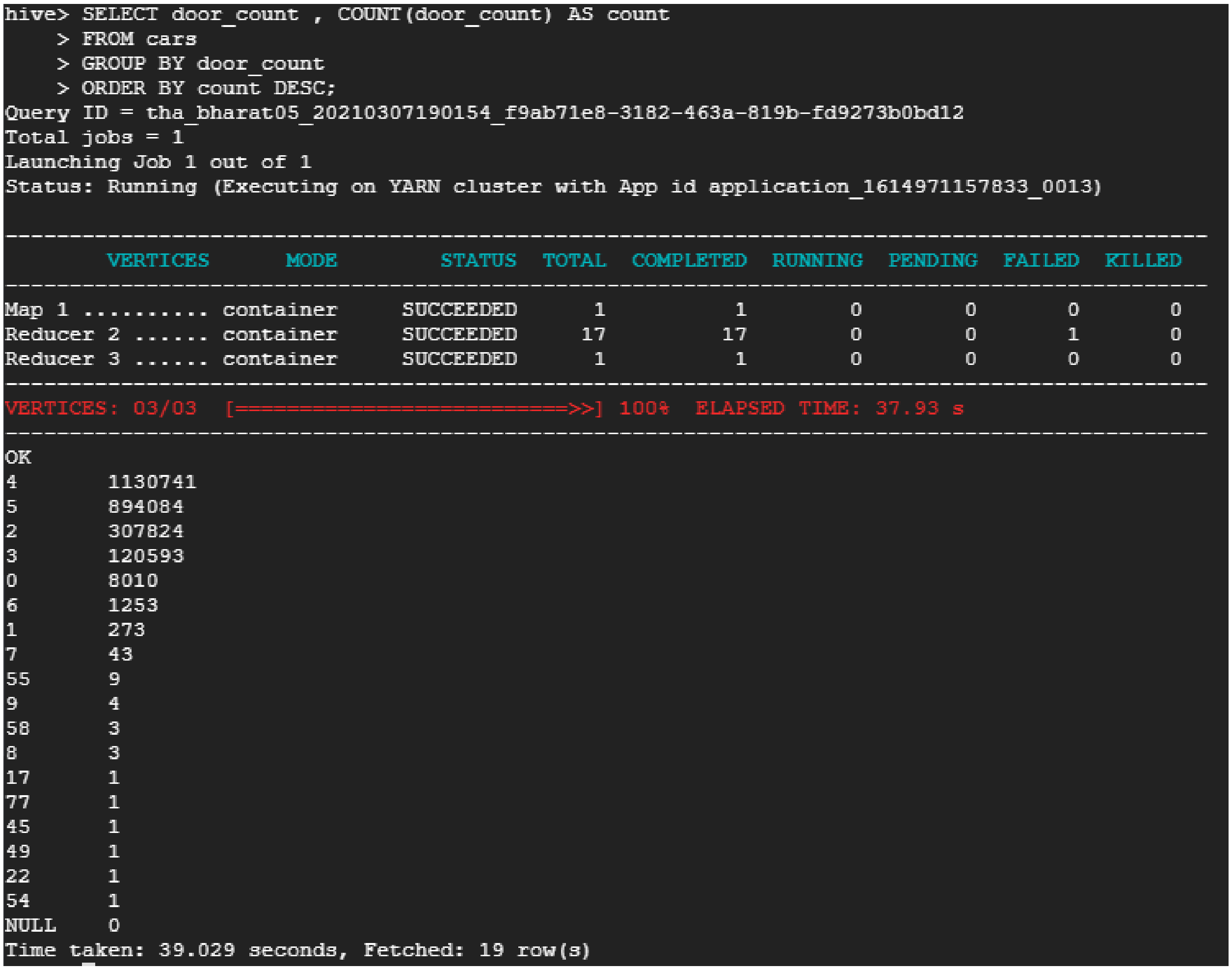
```
hive> SELECT maker, model, mileage, price_eur
    > FROM cars
    > ORDER BY price_eur DESC
    > LIMIT 10 ;
Query ID = tha_bharat05_20210307230042_2dc7802c-2042-4000-8429-9c094fb1622c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.64 s
----------------------------------------------------------------------------------------------
OK
renault kangoo  NULL    2.70614895E12
bmw             100     2.67945116E12
                NULL    2.72984687E11
citroen berlingo        245966  1.49223455E10
citroen berlingo        245966  1.49223455E10
citroen berlingo        245966  1.49223455E10
subaru  impreza 38000   1.48038676E10
mercedes-benz           37000   1.0E9
seat    ibiza   130000  1.0E9
audi    a5      23000   9.7121933E8
Time taken: 10.465 seconds, Fetched: 10 row(s)
```

```sql
-- Check different fuel types and their counts
SELECT fuel_type , COUNT(fuel_type) AS count
FROM cars
GROUP BY fuel_type;
```



```
hive> SELECT fuel_type , COUNT(fuel_type) AS count
    > FROM cars
    > GROUP BY fuel_type;
Query ID = tha_bharat05_20210307184908_4e6b713e-b392-4c1f-bc12-7f4c96dbf6cf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614971157833_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.48 s
----------------------------------------------------------------------------------------------
OK
gasoline        902222
cng     1124
electric        26350
        1847606
diesel  768207
lpg     7403
Time taken: 15.67 seconds, Fetched: 6 row(s)
```

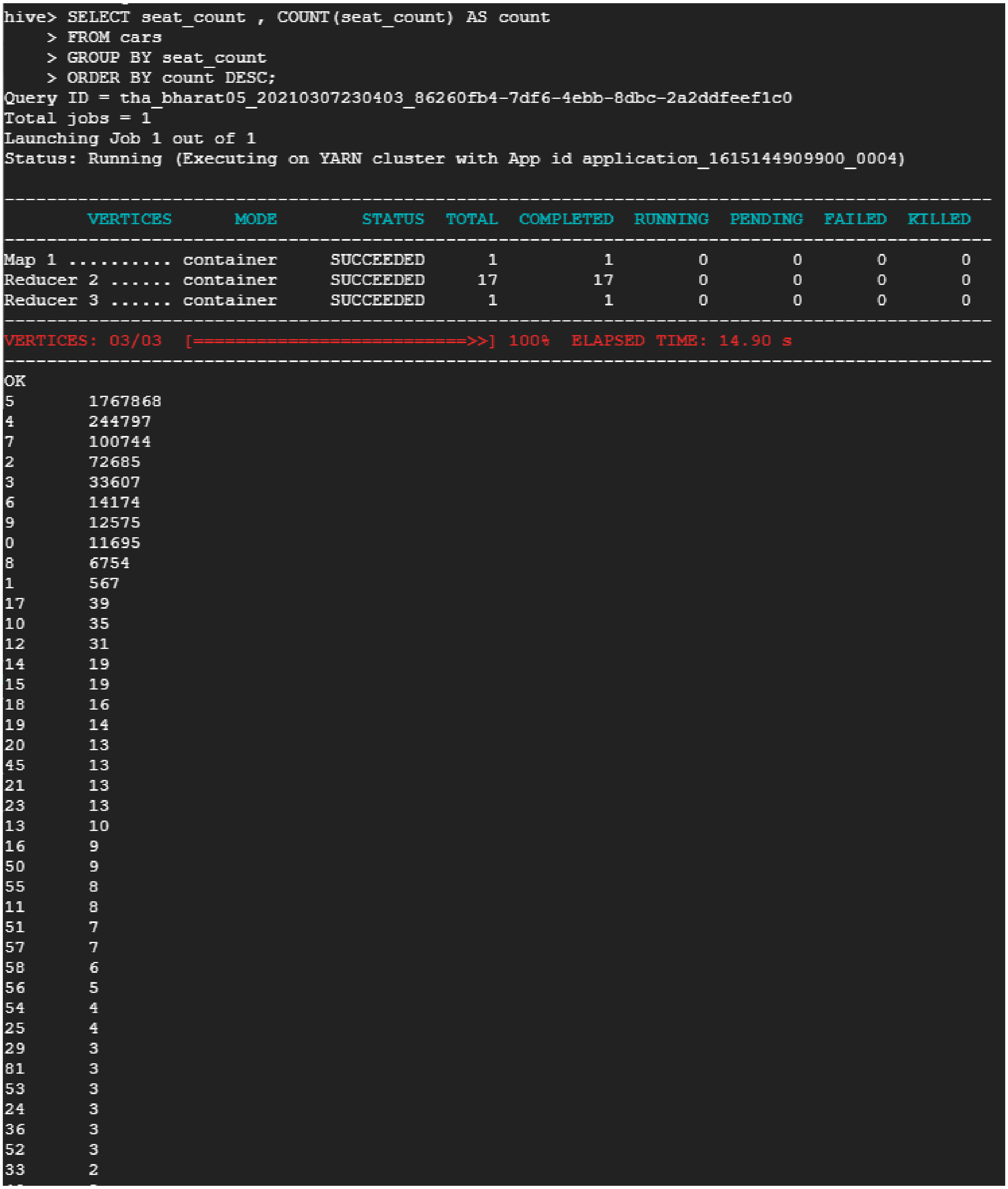
```
-- Check door_count and their counts
SELECT door_count , COUNT(door_count) AS count
FROM cars
GROUP BY door_count
ORDER BY count ASC;
```



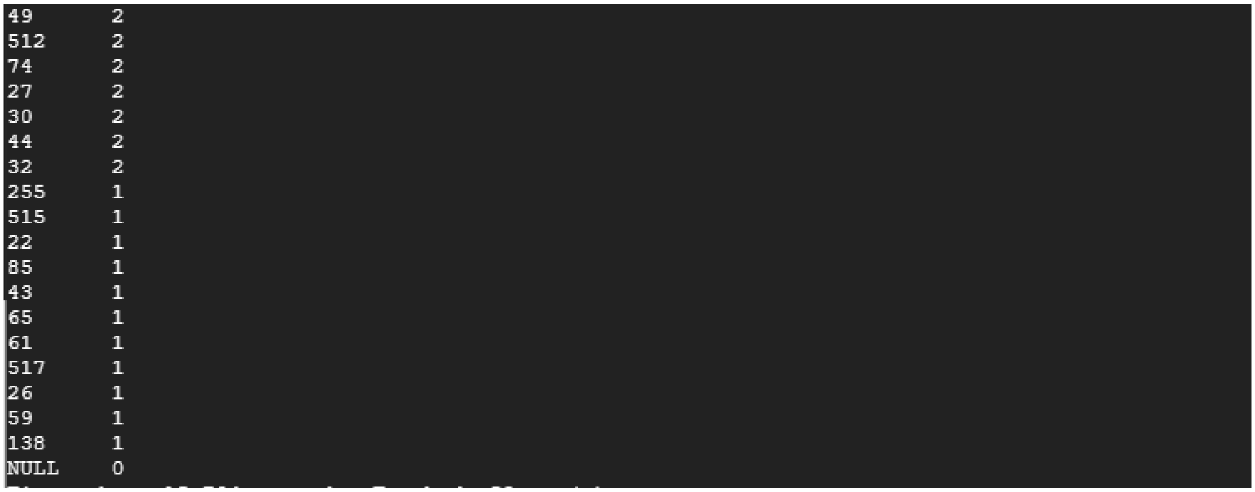
```
hive> SELECT door_count , COUNT(door_count) AS count
    > FROM cars
    > GROUP BY door_count
    > ORDER BY count DESC;
Query ID = tha_bharat05_20210307190154_f9ab71e8-3182-463a-819b-fd9273b0bd12
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1614971157833_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       1       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 37.93 s
----------------------------------------------------------------------------------------------
OK
4       1130741
5       894084
2       307824
3       120593
0       8010
6       1253
1       273
7       43
55      9
9       4
58      3
8       3
17      1
77      1
45      1
49      1
22      1
54      1
NULL    0
Time taken: 39.029 seconds, Fetched: 19 row(s)
```

1. Most of cars are 4 doored sedans as one would expect.

```sql
-- Check seat_count and their counts
SELECT seat_count , COUNT(seat_count) AS count
FROM cars
GROUP BY seat_count
ORDER BY count DESC;
```



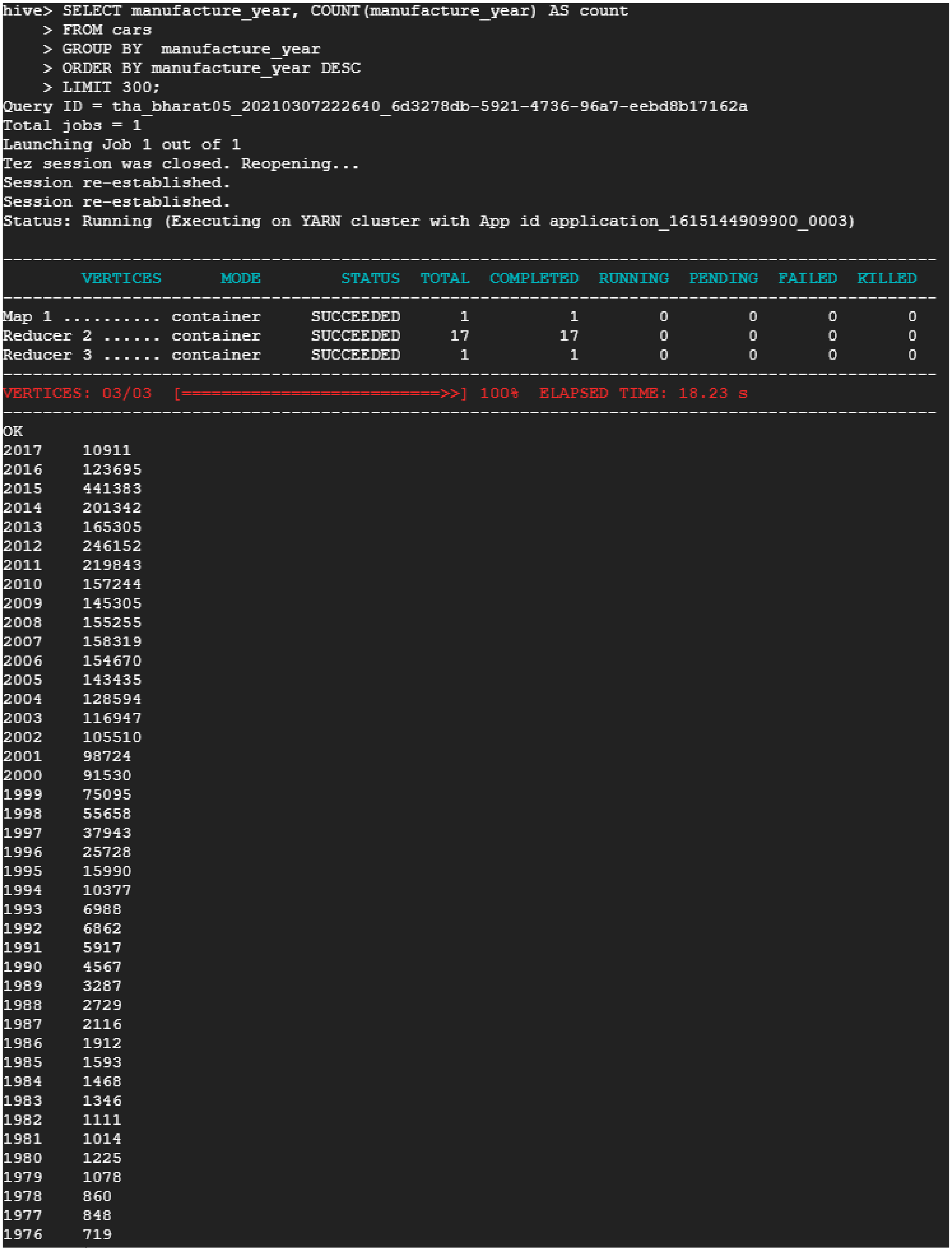
```
hive> SELECT seat_count , COUNT(seat_count) AS count
    > FROM cars
    > GROUP BY seat_count
    > ORDER BY count DESC;
Query ID = tha_bharat05_20210307230403_86260fb4-7df6-4ebb-8dbc-2a2ddfeef1c0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.90 s
----------------------------------------------------------------------------------------------
OK
5       1767868
4       244797
7       100744
2       72685
3       33607
6       14174
9       12575
0       11695
8       6754
1       567
17      39
10      35
12      31
14      19
15      19
18      16
19      14
20      13
45      13
21      13
23      13
13      10
16      9
50      9
55      8
11      8
51      7
57      7
58      6
56      5
54      4
25      4
29      3
81      3
53      3
24      3
36      3
52      3
33      2
```



```
49      2
512     2
74      2
27      2
30      2
44      2
32      2
255     1
515     1
22      1
85      1
43      1
65      1
61      1
517     1
26      1
59      1
138     1
NULL    0
```

1. Highest number of cars have 5 seat_counts, indicating they are 4 doored sedans.
2. Some values are unreasonable such as 517 etc.

```sql
--- manufacturing year and their count
SELECT manufacture_year, COUNT(manufacture_year) AS count
FROM cars
GROUP BY  manufacture_year
ORDER BY manufacture_year DESC
LIMIT 300;
```

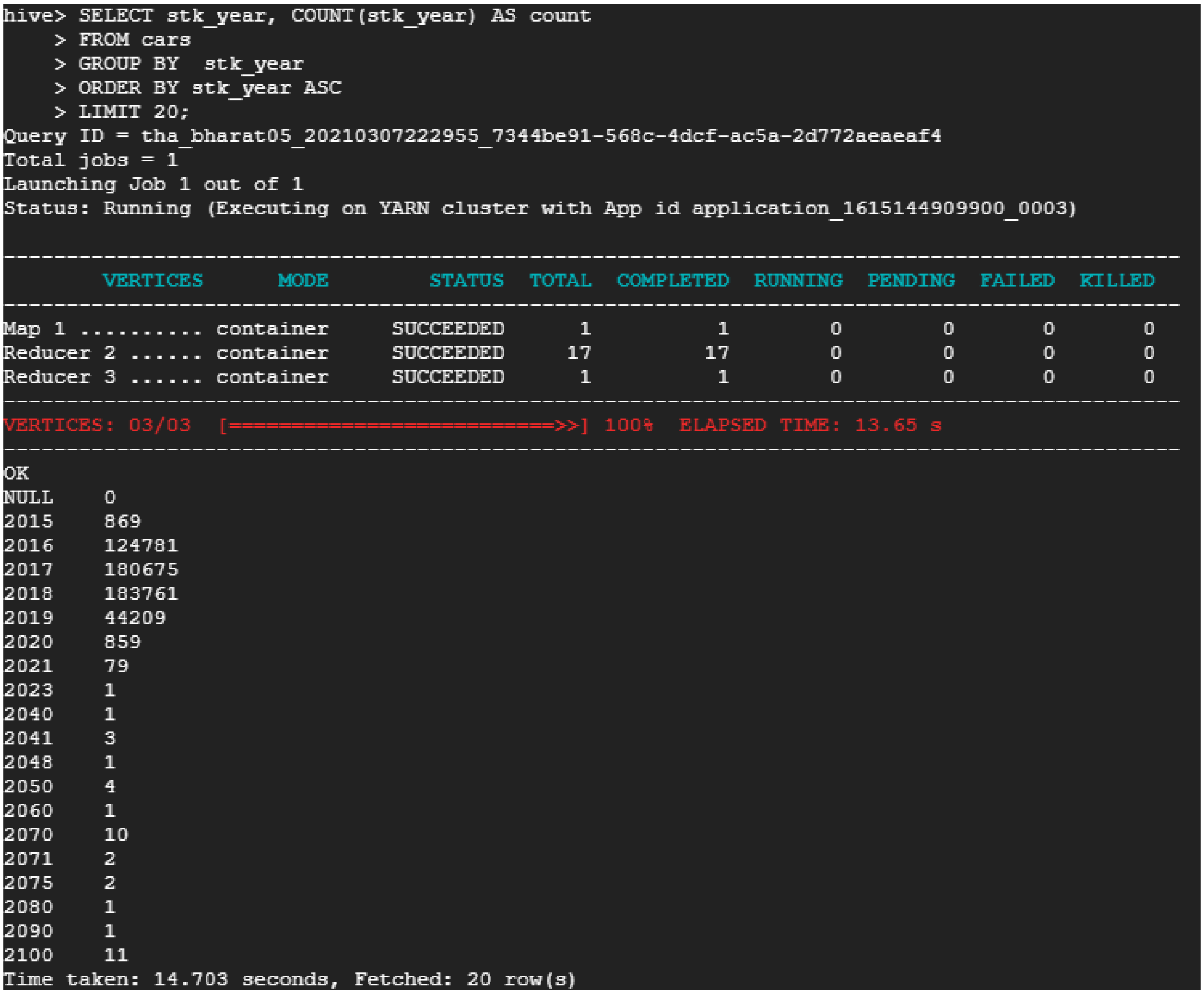


```
hive> SELECT manufacture_year, COUNT(manufacture_year) AS count
    > FROM cars
    > GROUP BY  manufacture_year
    > ORDER BY manufacture_year DESC
    > LIMIT 300;
Query ID = tha_bharat05_20210307222640_6d3278db-5921-4736-96a7-eebd8b17162a
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1615144909900_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 18.23 s
----------------------------------------------------------------------------------------------
OK
2017    10911
2016    123695
2015    441383
2014    201342
2013    165305
2012    246152
2011    219843
2010    157244
2009    145305
2008    155255
2007    158319
2006    154670
2005    143435
2004    128594
2003    116947
2002    105510
2001    98724
2000    91530
1999    75095
1998    55658
1997    37943
1996    25728
1995    15990
1994    10377
1993    6988
1992    6862
1991    5917
1990    4567
1989    3287
1988    2729
1987    2116
1986    1912
1985    1593
1984    1468
1983    1346
1982    1111
1981    1014
1980    1225
1979    1078
1978    860
1977    848
1976    719
```

(Not all the records are shown)

1. It is highly unlikely to have manufacturing years earlier than 1700, because cars were invented in 18th century. Earlier records are hard to explain.
2. Most of used cars are from last two decades.
3. 2017 shows fewer records than earlier years indicating incomplete records.

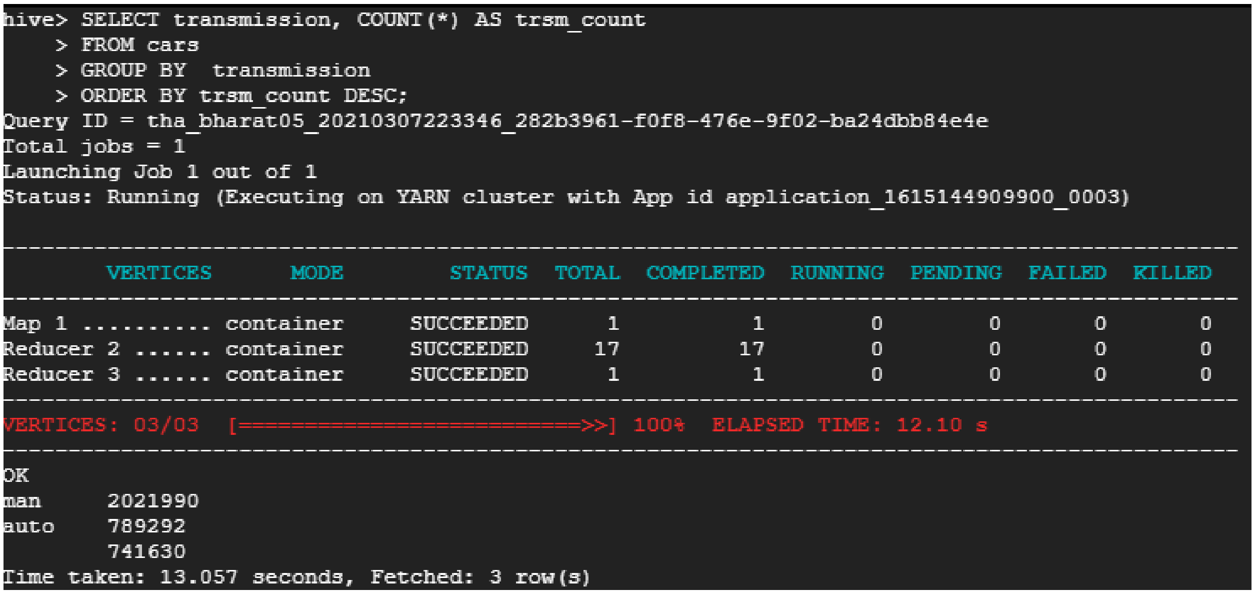
```
-- sticker years and their counts
SELECT stk_year, COUNT(stk_year) AS count
FROM cars
GROUP BY  stk_year
ORDER BY stk_year ASC
LIMIT 20;
```



```
hive> SELECT stk_year, COUNT(stk_year) AS count
    > FROM cars
    > GROUP BY  stk_year
    > ORDER BY stk_year ASC
    > LIMIT 20;
Query ID = tha_bharat05_20210307222955_7344be91-568c-4dcf-ac5a-2d772aeaeaf4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.65 s
----------------------------------------------------------------------------------------------
OK
NULL    0
2015    869
2016    124781
2017    180675
2018    183761
2019    44209
2020    859
2021    79
2023    1
2040    1
2041    3
2048    1
2050    4
2060    1
2070    10
2071    2
2075    2
2080    1
2090    1
2100    11
Time taken: 14.703 seconds, Fetched: 20 row(s)
```

1. Unreasonable values for sticker year for example 2100 etc.

```
--transmission types and their counts
SELECT transmission, COUNT(*) AS trsm_count
FROM cars
GROUP BY  transmission
ORDER BY trsm_count DESC;
```

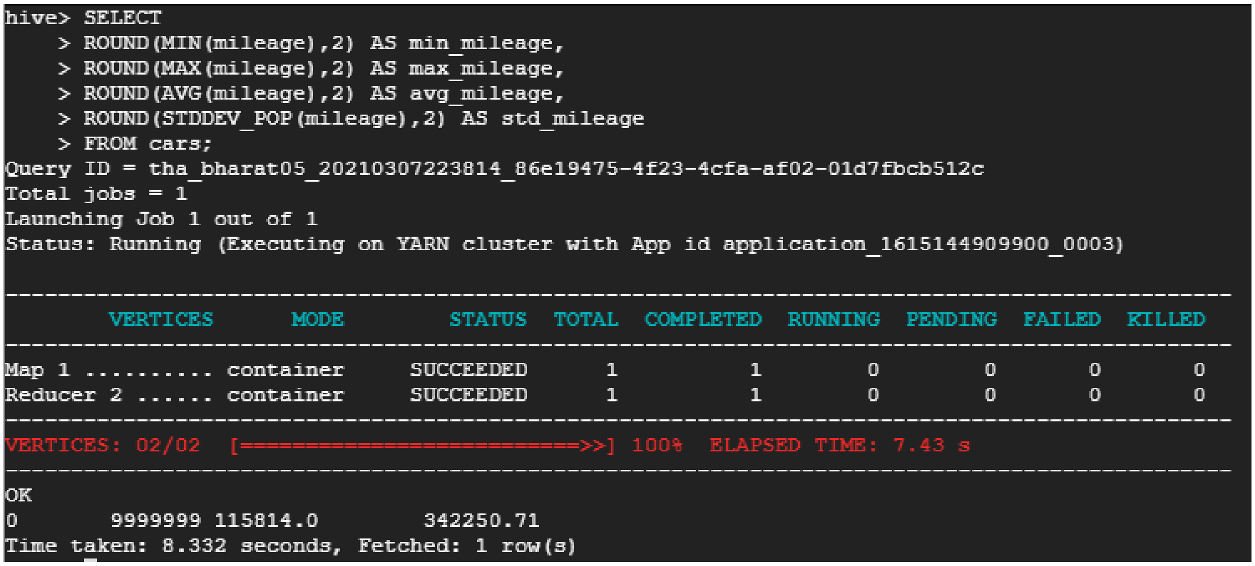


```
hive> SELECT transmission, COUNT(*) AS trsm_count
    > FROM cars
    > GROUP BY  transmission
    > ORDER BY trsm_count DESC;
Query ID = tha_bharat05_20210307223346_282b3961-f0f8-476e-9f02-ba24dbb84e4e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     17         17        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 12.10 s
----------------------------------------------------------------------------------------------
OK
man     2021990
auto    789292
        741630
Time taken: 13.057 seconds, Fetched: 3 row(s)
```

2. manual transmission is most common type of transmission.

## Descriptive Statistics Raw Dataset

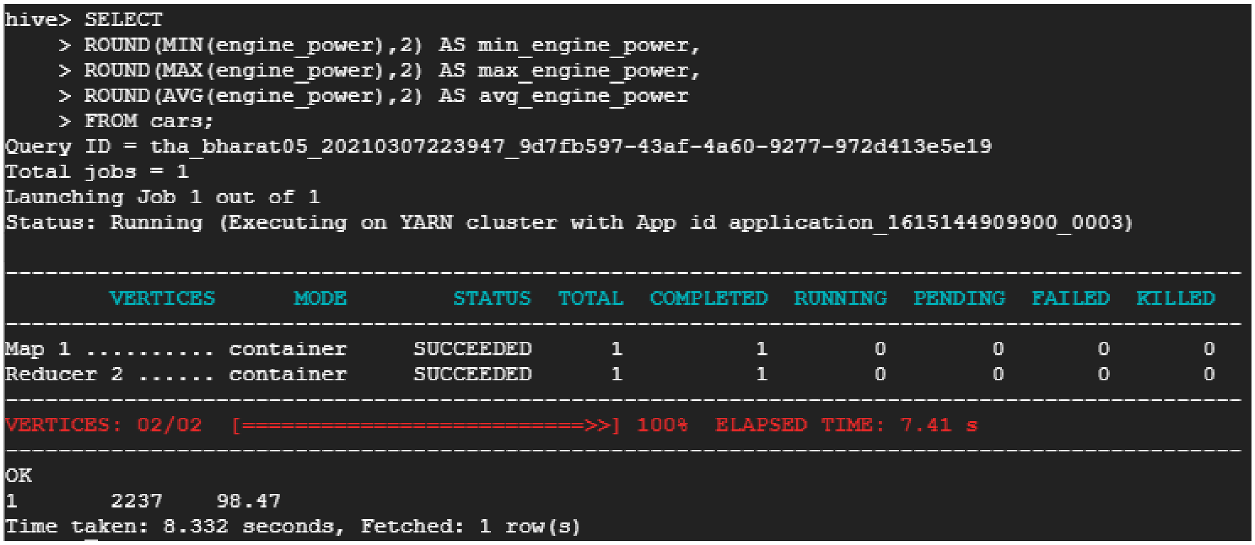
```
-- mileage
SELECT MIN(mileage) AS min_mileage,
MAX(mileage) AS max_mileage,
AVG(mileage) AS avg_mileage,
STDDEV_POP(mileage) AS std_mileage
FROM cars
```



```
hive> SELECT
    > ROUND(MIN(mileage),2) AS min_mileage,
    > ROUND(MAX(mileage),2) AS max_mileage,
    > ROUND(AVG(mileage),2) AS avg_mileage,
    > ROUND(STDDEV_POP(mileage),2) AS std_mileage
    > FROM cars;
Query ID = tha_bharat05_20210307223814_86e19475-4f23-4cfa-af02-01d7fbcb512c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.43 s
----------------------------------------------------------------------------------------------
OK
0       9999999 115814.0        342250.71
Time taken: 8.332 seconds, Fetched: 1 row(s)
```

1. Maximum mileage looks unreasonable, most likely some sort of default value.

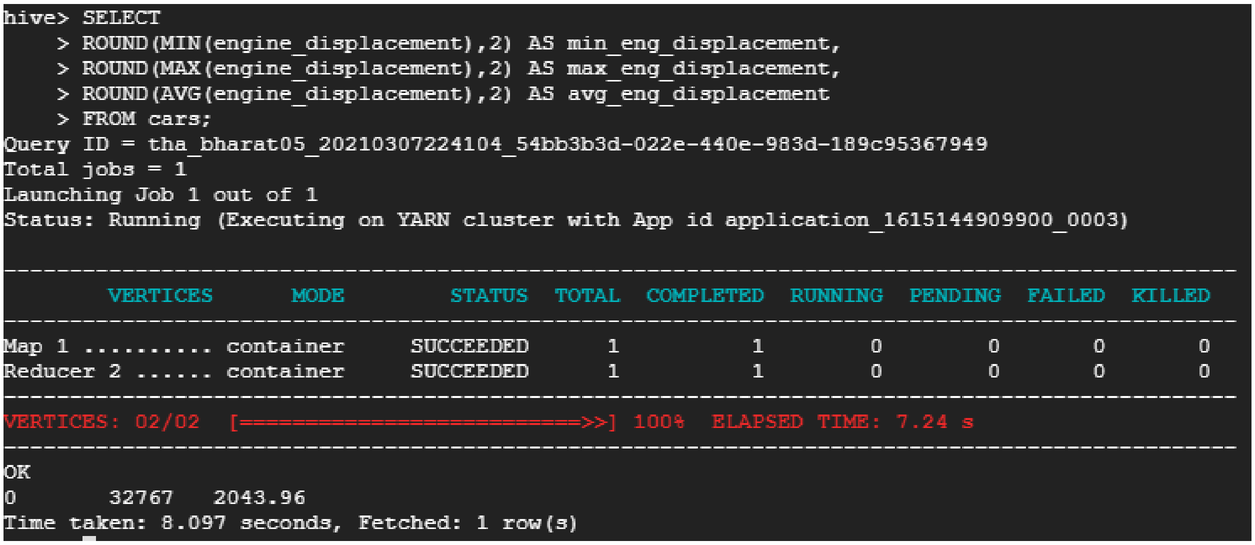
```
-- engine power
SELECT
ROUND(MIN(engine_power),2) AS min_engine_power,
ROUND(MAX(engine_power),2) AS max_engine_power,
ROUND(AVG(engine_power),2) AS avg_engine_power
FROM cars
```



```
hive> SELECT
    > ROUND(MIN(engine_power),2) AS min_engine_power,
    > ROUND(MAX(engine_power),2) AS max_engine_power,
    > ROUND(AVG(engine_power),2) AS avg_engine_power
    > FROM cars;
Query ID = tha_bharat05_20210307223947_9d7fb597-43af-4a60-9277-972d413e5e19
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.41 s
----------------------------------------------------------------------------------------------
OK
1       2237    98.47
Time taken: 8.332 seconds, Fetched: 1 row(s)
```

1. Min and max engine power need further investigation.

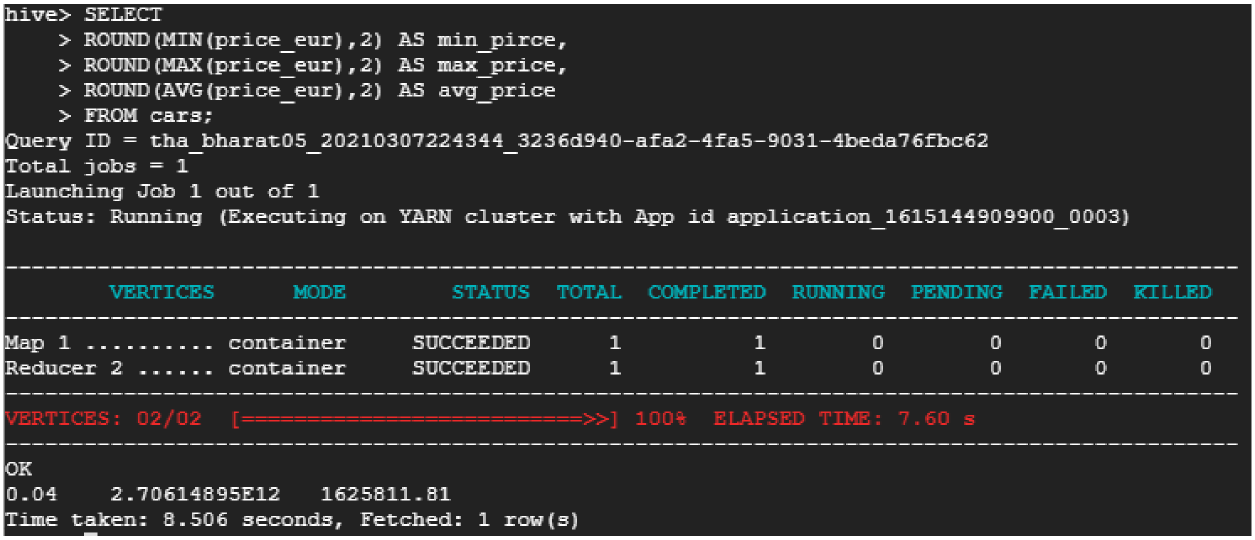
```sql
-- engine displacement
SELECT
ROUND(MIN(engine_displacement),2) AS min_eng_displacement,
ROUND(MAX(engine_displacement),2) AS max_eng_displacement,
ROUND(AVG(engine_displacement),2) AS avg_eng_displacement
FROM cars;
```



```
hive> SELECT
    > ROUND(MIN(engine_displacement),2) AS min_eng_displacement,
    > ROUND(MAX(engine_displacement),2) AS max_eng_displacement,
    > ROUND(AVG(engine_displacement),2) AS avg_eng_displacement
    > FROM cars;
Query ID = tha_bharat05_20210307224104_54bb3b3d-022e-440e-983d-189c95367949
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.24 s
----------------------------------------------------------------------------------------------
OK
0       32767   2043.96
Time taken: 8.097 seconds, Fetched: 1 row(s)
```

1. Just like power, minimum, and maximum engine displacement is hard to explain.

```sql
-- price
SELECT
ROUND(MIN(price_eur),2) AS min_pirce,
ROUND(MAX(price_eur),2) AS max_price,
ROUND(AVG(price_eur),2) AS avg_price
FROM cars;
```



```
hive> SELECT
    > ROUND(MIN(price_eur),2) AS min_pirce,
    > ROUND(MAX(price_eur),2) AS max_price,
    > ROUND(AVG(price_eur),2) AS avg_price
    > FROM cars;
Query ID = tha_bharat05_20210307224344_3236d940-afa2-4fa5-9031-4beda76fbc62
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.60 s
----------------------------------------------------------------------------------------------
OK
0.04    2.70614895E12   1625811.81
Time taken: 8.506 seconds, Fetched: 1 row(s)
```

1. Minimum price of 0.04 euros and 2.7 trillion euros looks very unreasonable.

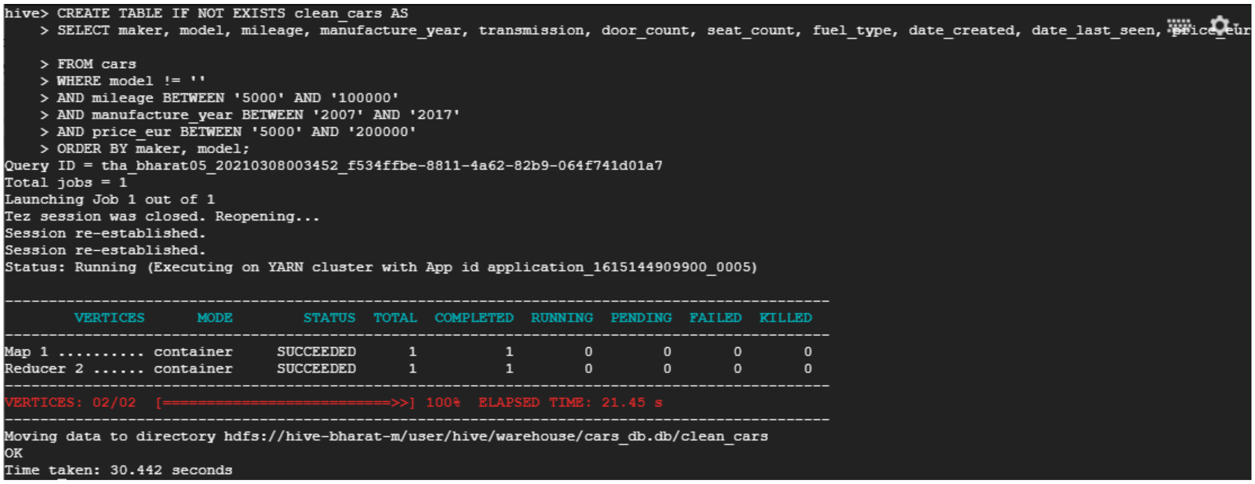
## Key Findings from Raw Data

- There are 3552912 cars in the raw dataset from 47 distinct car manufacturers. A total of 1013 distinct car models are present.
- stk_year, color_slug columns have more than 80% null values.
- model, body_type, and  door_count columns have over 30% blank values.
- All the cars have price information.
- Most of cars are 4 doored sedans as one would expect.
- Highest number of cars have 5 seat_counts, indicating they are 4 doored sedans. Some seat_count values are unreasonable such as 517 etc.
- It is highly unlikely to have manufacturing year earlier than 1700, because cars were invented in 18th century. Earlier records are hard to explain.
- Most of used cars are from last two decades.
- 2017 shows fewer records than earlier years indicating incomplete records.
- Unreasonable values for sticker year for example 2100 etc.
- Manual transmission is most common type of transmission.
- Maximum mileage looks unreasonable, most likely some sort of default value.
- Minimum and maximum engine power need further investigation.
- Just like power, minimum, and maximum engine displacement is hard to explain.
- Minimum price of 0.04 euros and maximum price 2.7 trillion euros looks very unreasonable.

## Data preparation

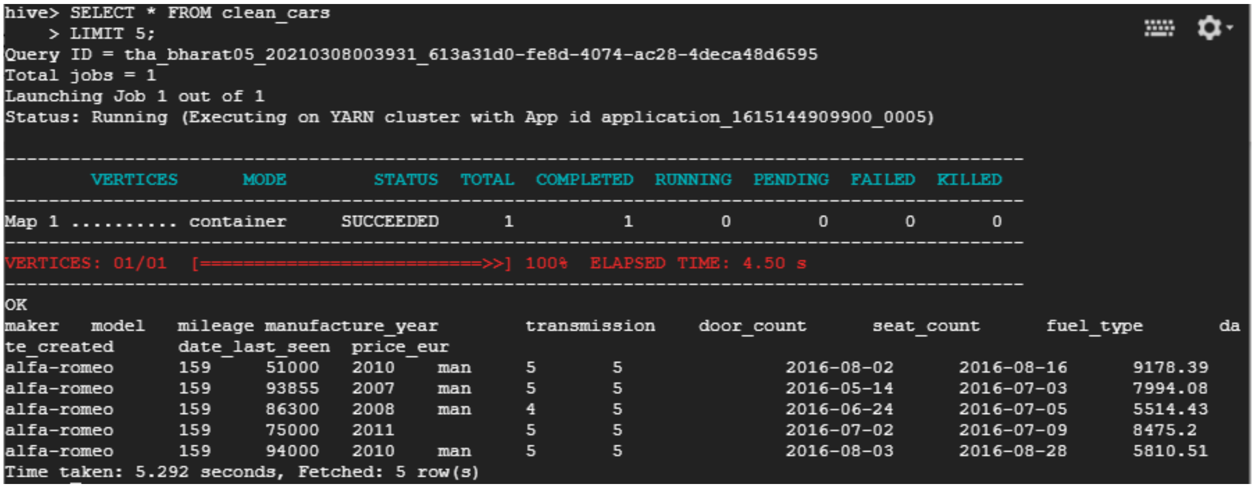
- To extract more reasonable analytical insights, only the data satisfying the below conditions was analyzed further:
- Sticker year and color slug columns will be ignored since they have than 80% null values.
- Rows which have model field as null will be ignored.
- Columns such as engine displacement, engine power, body type will also be ignored, since these columns have suspicious values, and this data can be easily verified from car manufacturer using manufacturing year and model.
- Door count and Seat count, although have some unreasonable values, are retained for trend analysis.
- Only considering cars with mileage greater than 5,000 and less than 100,000. Cars with mileage less than 5,000 will be priced at par with new car, so it would be better to buy new car instead. And Cars with mileage 100,000 will likely have high maintenance costs, so it is prudent to avoid those.
- Only cars which are less than 10 years old are selected. Cars outside this range will likely lead to high maintenance costs.
- Price range was selected between 5000 € and 200,000 €. This would eliminate the problematic values while retaining majority of the cars.

```sql
--Create new clean Table
CREATE TABLE IF NOT EXISTS clean_cars AS
SELECT maker, model, mileage, manufacture_year, transmission, door_count,
seat_count, fuel_type, date_created, date_last_seen, price_eur
FROM cars
WHERE model != ''
AND mileage BETWEEN '5000' AND '100000'
AND manufacture_year BETWEEN '2007' AND '2017'
AND price_eur BETWEEN '5000' AND '200000'
ORDER BY maker, model;
```



```
hive> CREATE TABLE IF NOT EXISTS clean_cars AS
    > SELECT maker, model, mileage, manufacture_year, transmission, door_count, seat_count, fuel_type, date_created, date_last_seen, price_eur

    > FROM cars
    > WHERE model != ''
    > AND mileage BETWEEN '5000' AND '100000'
    > AND manufacture_year BETWEEN '2007' AND '2017'
    > AND price_eur BETWEEN '5000' AND '200000'
    > ORDER BY maker, model;
Query ID = tha_bharat05_20210308003452_f534ffbe-8811-4a62-82b9-064f741d01a7
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1615144909900_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 21.45 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://hive-bharat-m/user/hive/warehouse/cars_db.db/clean_cars
OK
Time taken: 30.442 seconds
```

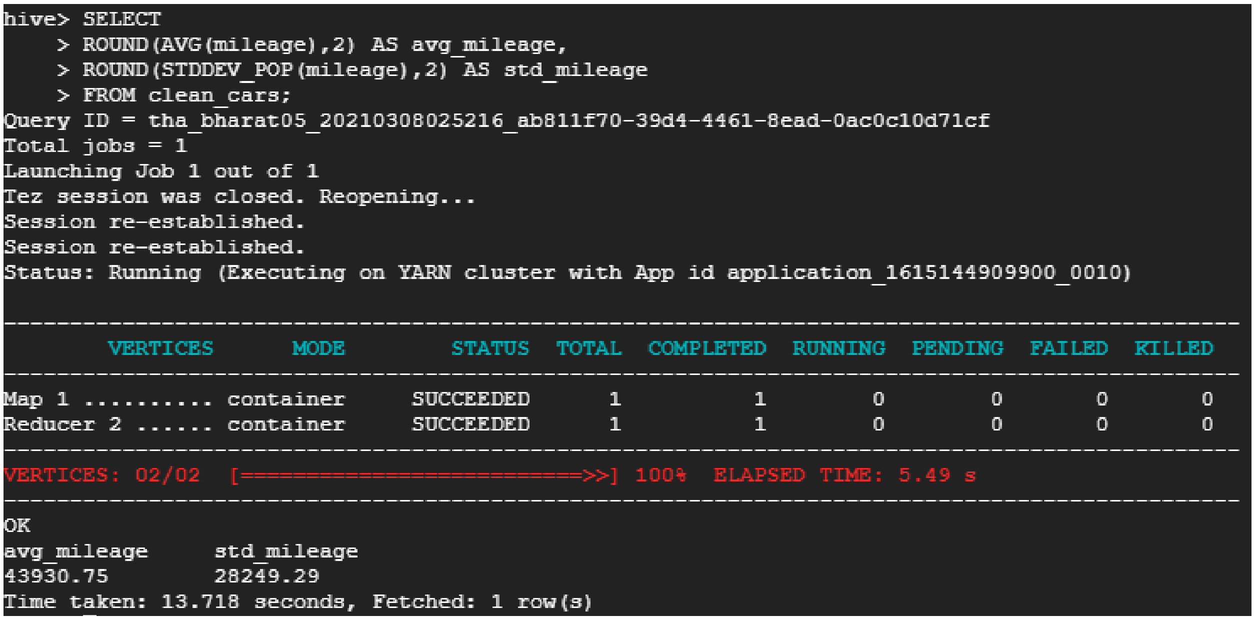
```sql
-- data preview
SELECT * FROM clean_cars
LIMIT 5;
```



```
hive> SELECT * FROM clean_cars
    > LIMIT 5;
Query ID = tha_bharat05_20210308003931_613a31d0-fe8d-4074-ac28-4deca48d6595
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.50 s
----------------------------------------------------------------------------------------------
OK
maker   model   mileage manufacture_year        transmission    door_count      seat_count      fuel_type       da
te_created      date_last_seen  price_eur
alfa-romeo      159     51000   2010    man     5       5               2016-08-02      2016-08-16      9178.39
alfa-romeo      159     93855   2007    man     5       5               2016-05-14      2016-07-03      7994.08
alfa-romeo      159     86300   2008    man     4       5               2016-06-24      2016-07-05      5514.43
alfa-romeo      159     75000   2011            5       5               2016-07-02      2016-07-09      8475.2
alfa-romeo      159     94000   2010    man     5       5               2016-08-03      2016-08-28      5810.51
Time taken: 5.292 seconds, Fetched: 5 row(s)
```

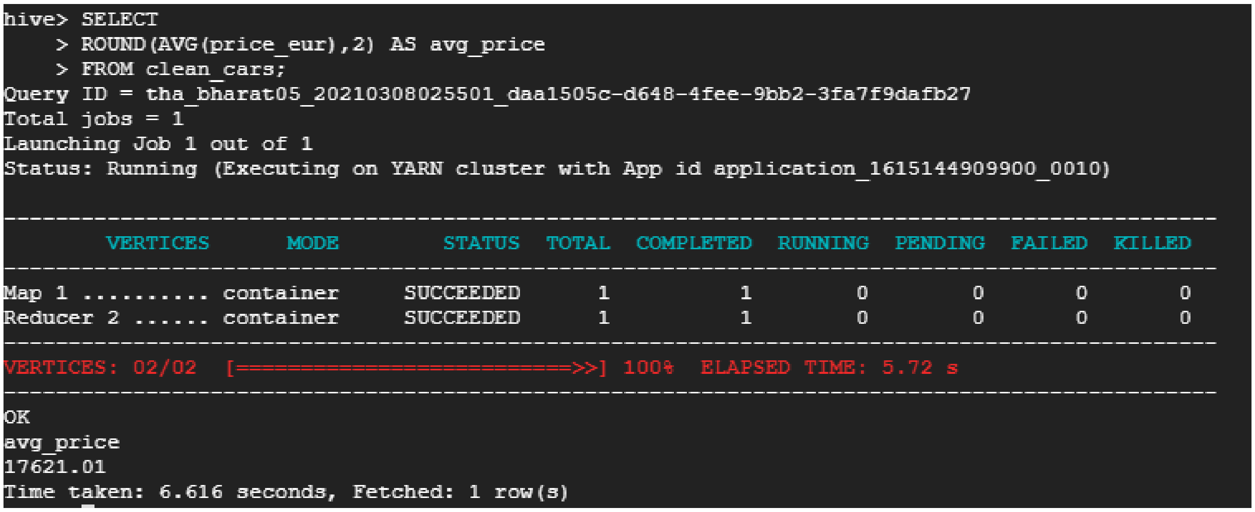
## Descriptive Statistics Clean Dataset

```
-- mileage
SELECT
ROUND(AVG(mileage),2) AS avg_mileage,
ROUND(STDDEV_POP(mileage),2) AS std_mileage
FROM clean_cars;
```



```
hive> SELECT
    > ROUND(AVG(mileage),2) AS avg_mileage,
    > ROUND(STDDEV_POP(mileage),2) AS std_mileage
    > FROM clean_cars;
Query ID = tha_bharat05_20210308025216_ab811f70-39d4-4461-8ead-0ac0c10d71cf
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1615144909900_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.49 s
----------------------------------------------------------------------------------------------
OK
avg_mileage     std_mileage
43930.75        28249.29
Time taken: 13.718 seconds, Fetched: 1 row(s)
```

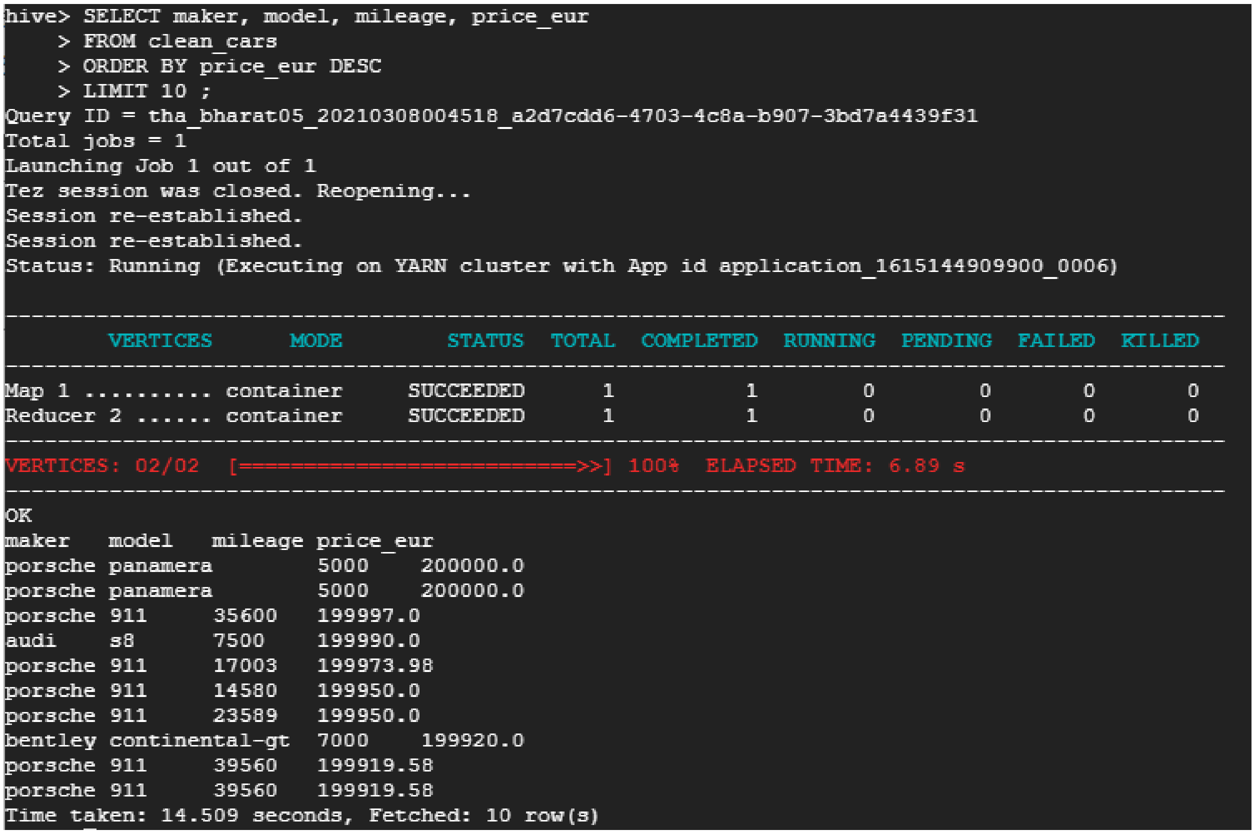
```
SELECT
ROUND(AVG(price_eur),2) AS avg_price
FROM clean_cars;
```



```
hive> SELECT
    > ROUND(AVG(price_eur),2) AS avg_price
    > FROM clean_cars;
Query ID = tha_bharat05_20210308025501_daa1505c-d648-4fee-9bb2-3fa7f9dafb27
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.72 s
----------------------------------------------------------------------------------------------
OK
avg_price
17621.01
Time taken: 6.616 seconds, Fetched: 1 row(s)
```

# Exploratory Analysis Clean Dataset

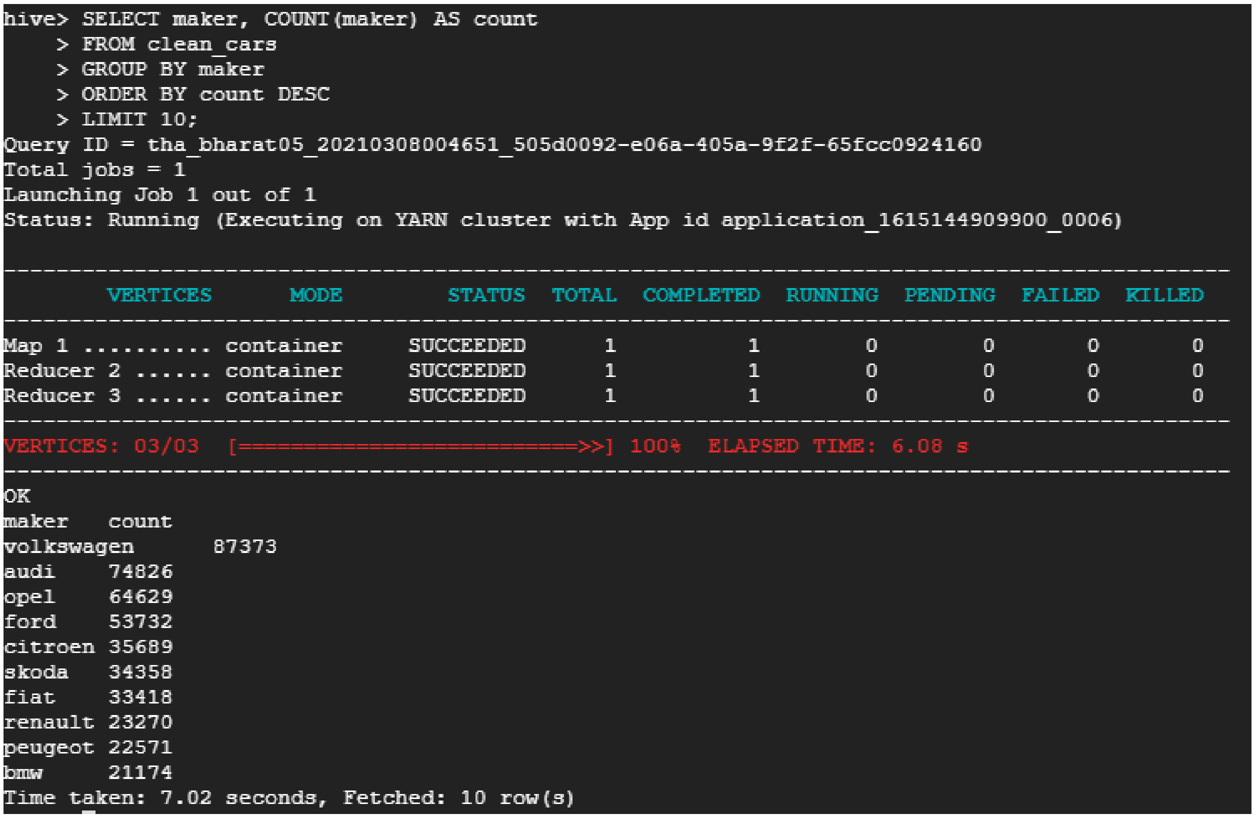
```
--Top 10 most expensive cars
SELECT maker, model, mileage, price_eur
FROM clean_cars
ORDER BY price_eur DESC
LIMIT 10 ;
```



```
hive> SELECT maker, model, mileage, price_eur
    > FROM clean_cars
    > ORDER BY price_eur DESC
    > LIMIT 10 ;
Query ID = tha_bharat05_20210308004518_a2d7cdd6-4703-4c8a-b907-3bd7a4439f31
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1615144909900_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.89 s
----------------------------------------------------------------------------------------------
OK
maker   model   mileage price_eur
porsche panamera        5000    200000.0
porsche panamera        5000    200000.0
porsche 911     35600   199997.0
audi    s8      7500    199990.0
porsche 911     17003   199973.98
porsche 911     14580   199950.0
porsche 911     23589   199950.0
bentley continental-gt  7000    199920.0
porsche 911     39560   199919.58
porsche 911     39560   199919.58
Time taken: 14.509 seconds, Fetched: 10 row(s)
```

1. In the clean dataset Porsche Panamera is the most expensive car.

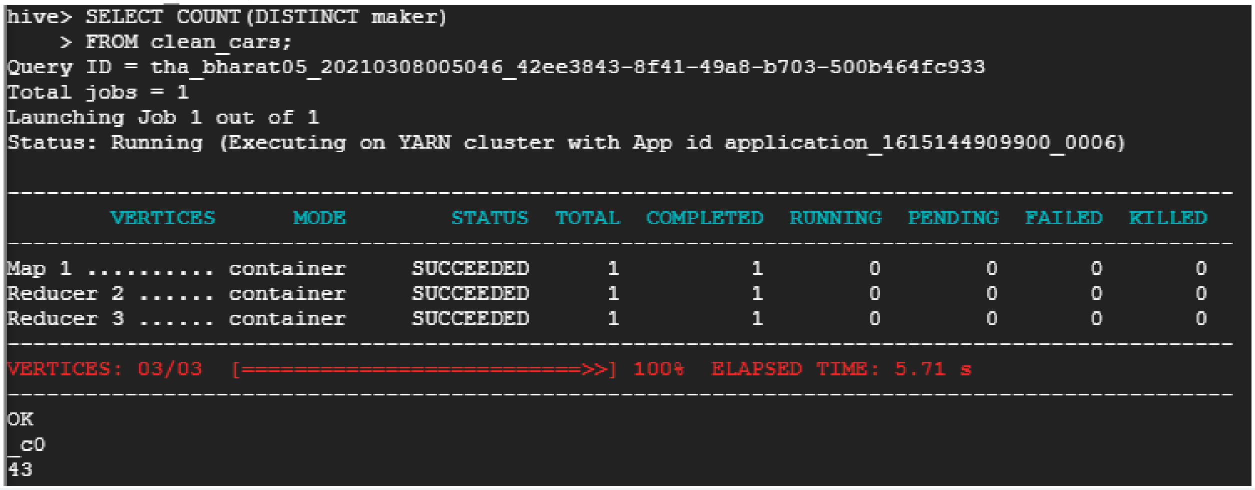
```
-- Top 10 popular car makers
SELECT maker, COUNT(maker) AS count
FROM clean_cars
GROUP BY maker
ORDER BY count DESC
LIMIT 10;
```



```
hive> SELECT maker, COUNT(maker) AS count
    > FROM clean_cars
    > GROUP BY maker
    > ORDER BY count DESC
    > LIMIT 10;
Query ID = tha_bharat05_20210308004651_505d0092-e06a-405a-9f2f-65fcc0924160
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 6.08 s
----------------------------------------------------------------------------------------------
OK
maker   count
volkswagen      87373
audi    74826
opel    64629
ford    53732
citroen 35689
skoda   34358
fiat    33418
renault 23270
peugeot 22571
bmw     21174
Time taken: 7.02 seconds, Fetched: 10 row(s)
```

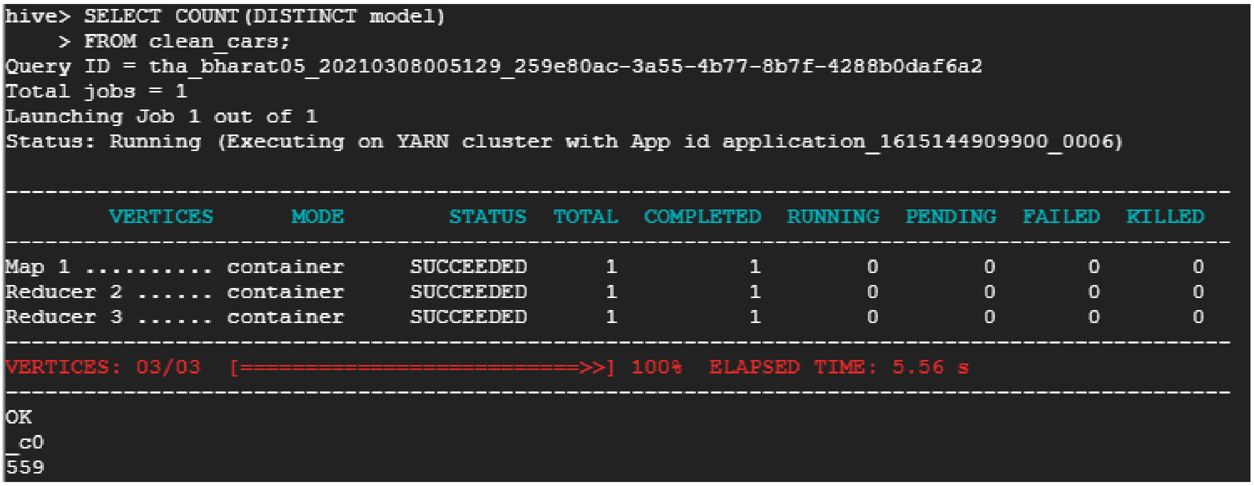
1. It turns out the in the clean dataset most cars are of Volkswagen make.

```
-- unique makers
SELECT COUNT(DISTINCT model)
FROM clean_cars;
```



```
hive> SELECT COUNT(DISTINCT maker)
    > FROM clean_cars;
Query ID = tha_bharat05_20210308005046_42ee3843-8f41-49a8-b703-500b464fc933
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 5.71 s
----------------------------------------------------------------------------------------------
OK
_c0
43
```

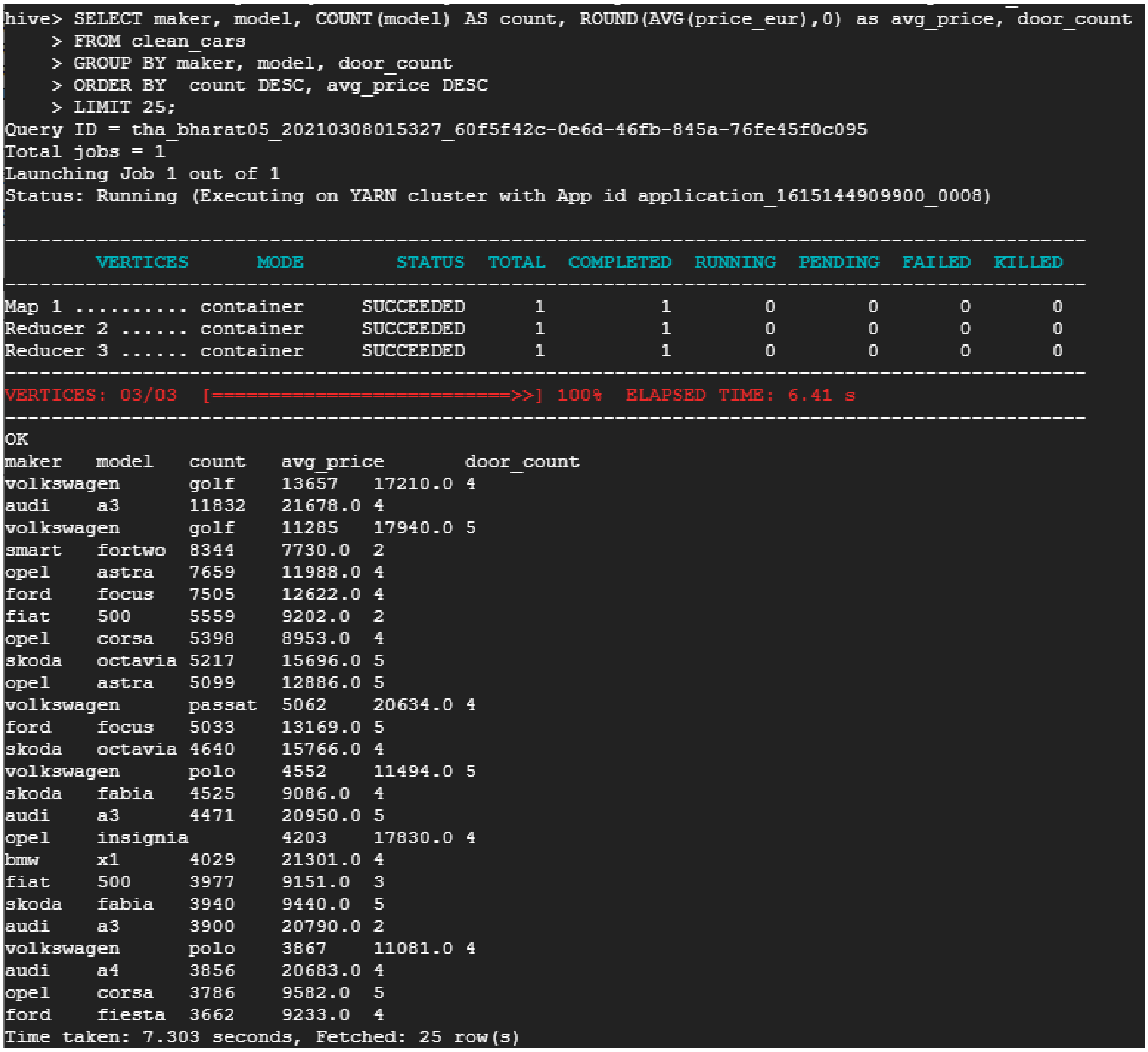
```
-- unique makers
SELECT COUNT(DISTINCT model)
FROM clean_cars;
```



```
hive> SELECT COUNT(DISTINCT model)
    > FROM clean_cars;
Query ID = tha_bharat05_20210308005129_259e80ac-3a55-4b77-8b7f-4288b0daf6a2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 5.56 s
----------------------------------------------------------------------------------------------
OK
_c0
559
```

1. Buyers can choose any of  559 model from 43 makers in the clean dataset.

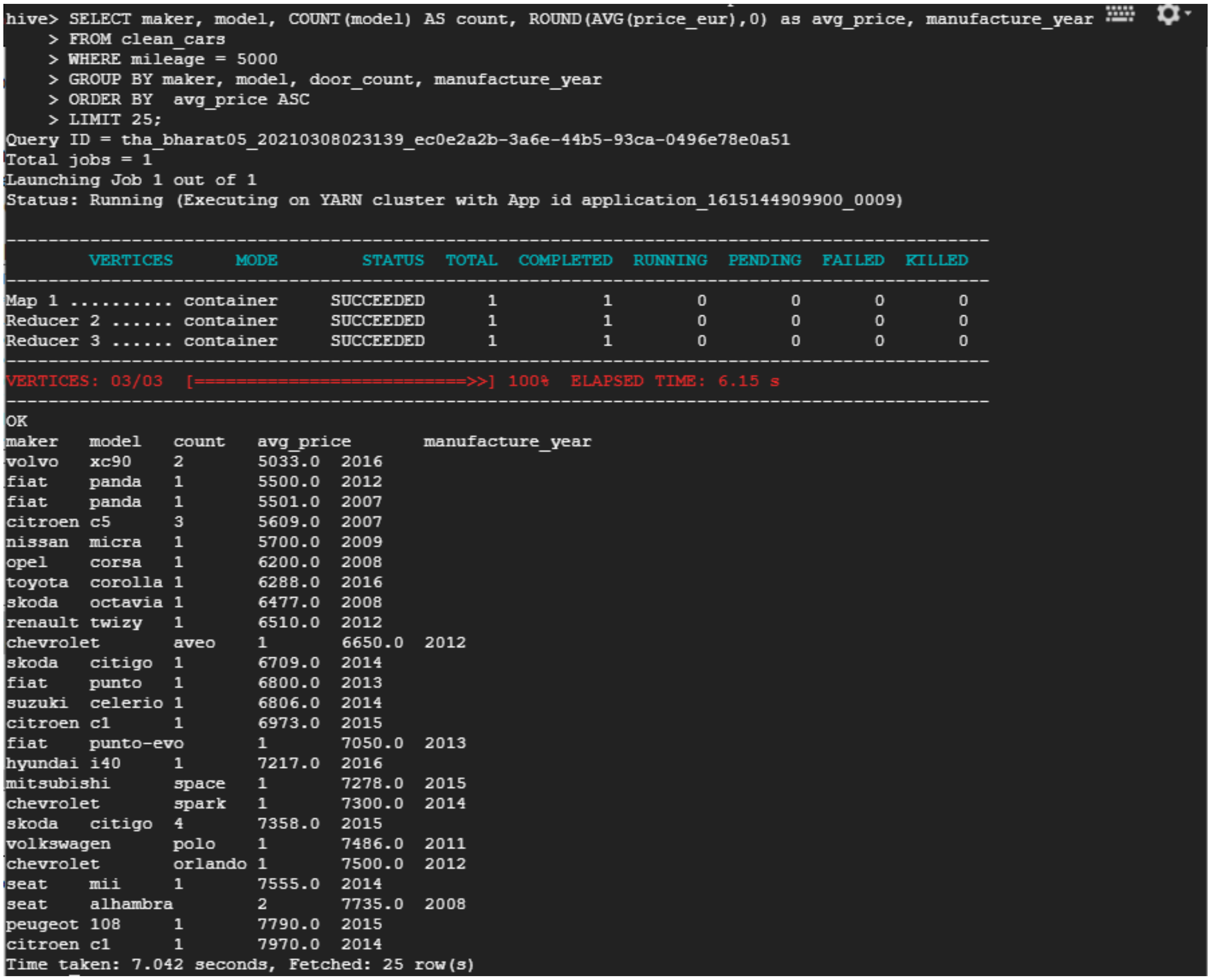
```
-- Top 25 available car models
SELECT maker, model, COUNT(model) AS count, ROUND(AVG(price_eur),0) as
avg_price
FROM clean_cars
GROUP BY maker, model
ORDER BY  count DESC, avg_price DESC
LIMIT 25;
```



```
hive> SELECT maker, model, COUNT(model) AS count, ROUND(AVG(price_eur),0) as avg_price, door_count
    > FROM clean_cars
    > GROUP BY maker, model, door_count
    > ORDER BY  count DESC, avg_price DESC
    > LIMIT 25;
Query ID = tha_bharat05_20210308015327_60f5f42c-0e6d-46fb-845a-76fe45f0c095
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0008)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 6.41 s
----------------------------------------------------------------------------------------------
OK
maker   model   count   avg_price       door_count
volkswagen      golf    13657   17210.0 4
audi    a3      11832   21678.0 4
volkswagen      golf    11285   17940.0 5
smart   fortwo  8344    7730.0  2
opel    astra   7659    11988.0 4
ford    focus   7505    12622.0 4
fiat    500     5559    9202.0  2
opel    corsa   5398    8953.0  4
skoda   octavia 5217    15696.0 5
opel    astra   5099    12886.0 5
volkswagen      passat  5062    20634.0 4
ford    focus   5033    13169.0 5
skoda   octavia 4640    15766.0 4
volkswagen      polo    4552    11494.0 5
skoda   fabia   4525    9086.0  4
audi    a3      4471    20950.0 5
opel    insignia        4203    17830.0 4
bmw     x1      4029    21301.0 4
fiat    500     3977    9151.0  3
skoda   fabia   3940    9440.0  5
audi    a3      3900    20790.0 2
volkswagen      polo    3867    11081.0 4
audi    a4      3856    20683.0 4
opel    corsa   3786    9582.0  5
ford    fiesta  3662    9233.0  4
Time taken: 7.303 seconds, Fetched: 25 row(s)
```

1. Now, Golf model has the highest availability.

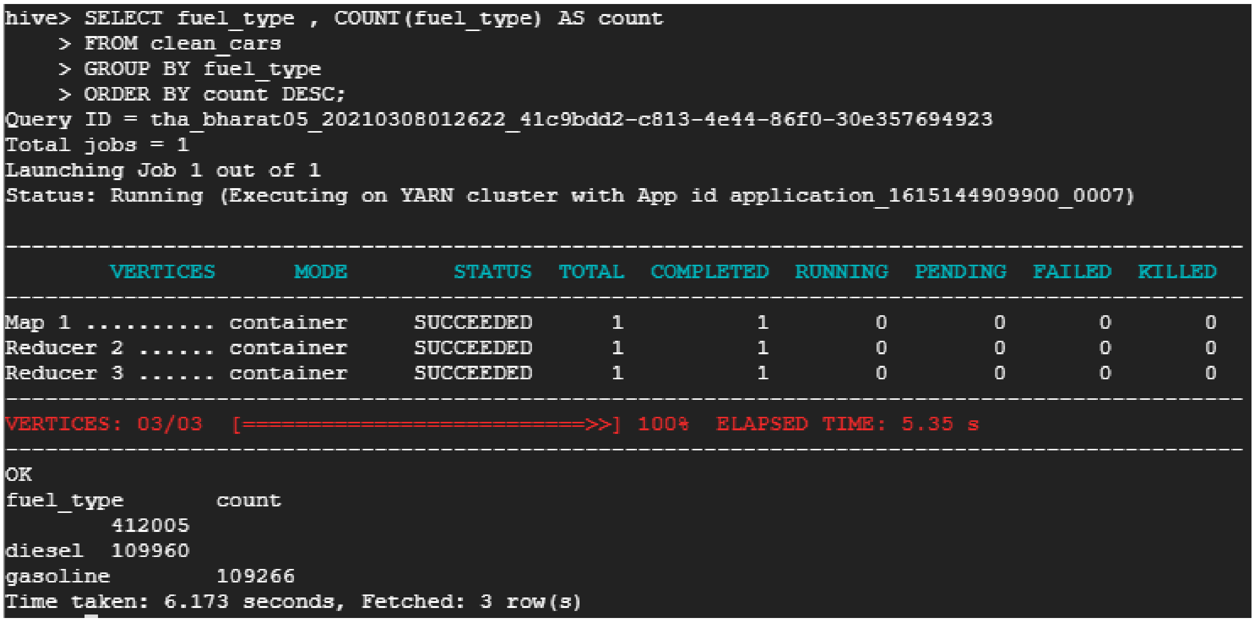
```
--- cars which have driven least i.e. 5000 km and are also cheaper
SELECT maker, model, COUNT(model) AS count, ROUND(AVG(price_eur),0) as
avg_price, manufacture_year
FROM clean_cars
WHERE mileage = 5000
GROUP BY maker, model, door_count, manufacture_year
ORDER BY  avg_price ASC
LIMIT 25;
```



```
hive> SELECT maker, model, COUNT(model) AS count, ROUND(AVG(price_eur),0) as avg_price, manufacture_year
    > FROM clean_cars
    > WHERE mileage = 5000
    > GROUP BY maker, model, door_count, manufacture_year
    > ORDER BY  avg_price ASC
    > LIMIT 25;
Query ID = tha_bharat05_20210308023139_ec0e2a2b-3a6e-44b5-93ca-0496e78e0a51
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 6.15 s
----------------------------------------------------------------------------------------------
OK
maker   model   count   avg_price       manufacture_year
volvo   xc90    2       5033.0  2016
fiat    panda   1       5500.0  2012
fiat    panda   1       5501.0  2007
citroen c5      3       5609.0  2007
nissan  micra   1       5700.0  2009
opel    corsa   1       6200.0  2008
toyota  corolla 1       6288.0  2016
skoda   octavia 1       6477.0  2008
renault twizy   1       6510.0  2012
chevrolet       aveo    1       6650.0  2012
skoda   citigo  1       6709.0  2014
fiat    punto   1       6800.0  2013
suzuki  celerio 1       6806.0  2014
citroen c1      1       6973.0  2015
fiat    punto-evo       1       7050.0  2013
hyundai i40     1       7217.0  2016
mitsubishi      space   1       7278.0  2015
chevrolet       spark   1       7300.0  2014
skoda   citigo  4       7358.0  2015
volkswagen      polo    1       7486.0  2011
chevrolet       orlando 1       7500.0  2012
seat    mii     1       7555.0  2014
seat    alhambra        2       7735.0  2008
peugeot 108     1       7790.0  2015
citroen c1      1       7970.0  2014
Time taken: 7.042 seconds, Fetched: 25 row(s)
```

1. These car prices appear too good to be true, one must exercise extra caution while buying lets say Volvo XC 90, manufactured in 2016 and driven only 5000, for about 5,000 euros.

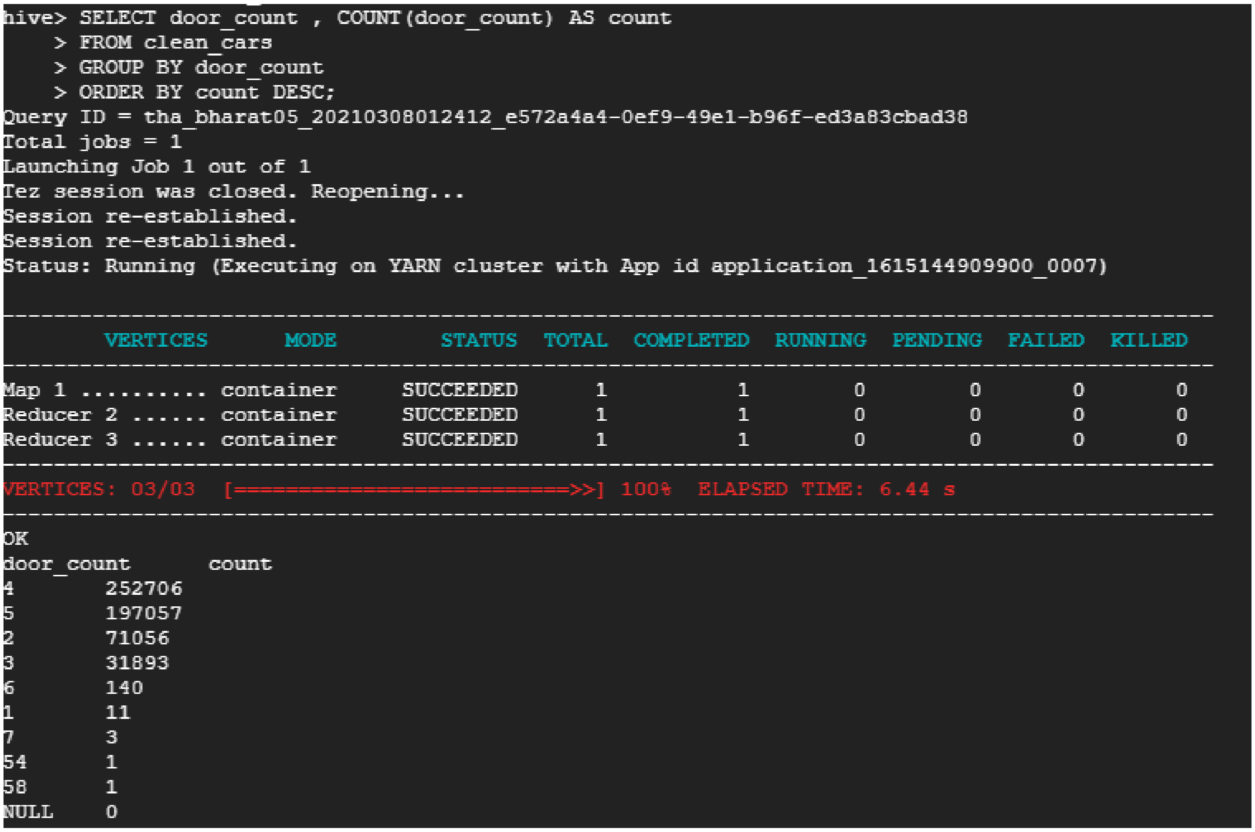
```
-- Check different fuel types and their counts
SELECT fuel_type , COUNT(fuel_type) AS count
FROM clean_cars
GROUP BY fuel_type
ORDER BY count DESC;
```



```
hive> SELECT fuel_type , COUNT(fuel_type) AS count
    > FROM clean_cars
    > GROUP BY fuel_type
    > ORDER BY count DESC;
Query ID = tha_bharat05_20210308012622_41c9bdd2-c813-4e44-86f0-30e357694923
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 5.35 s
----------------------------------------------------------------------------------------------
OK
fuel_type       count
        412005
diesel  109960
gasoline        109266
Time taken: 6.173 seconds, Fetched: 3 row(s)
```

1. It appears data cleaning led to removal of electric, cng and lpg cars. In the clean dataset there is almost equal number of gasoline powered and diesel powered cars.

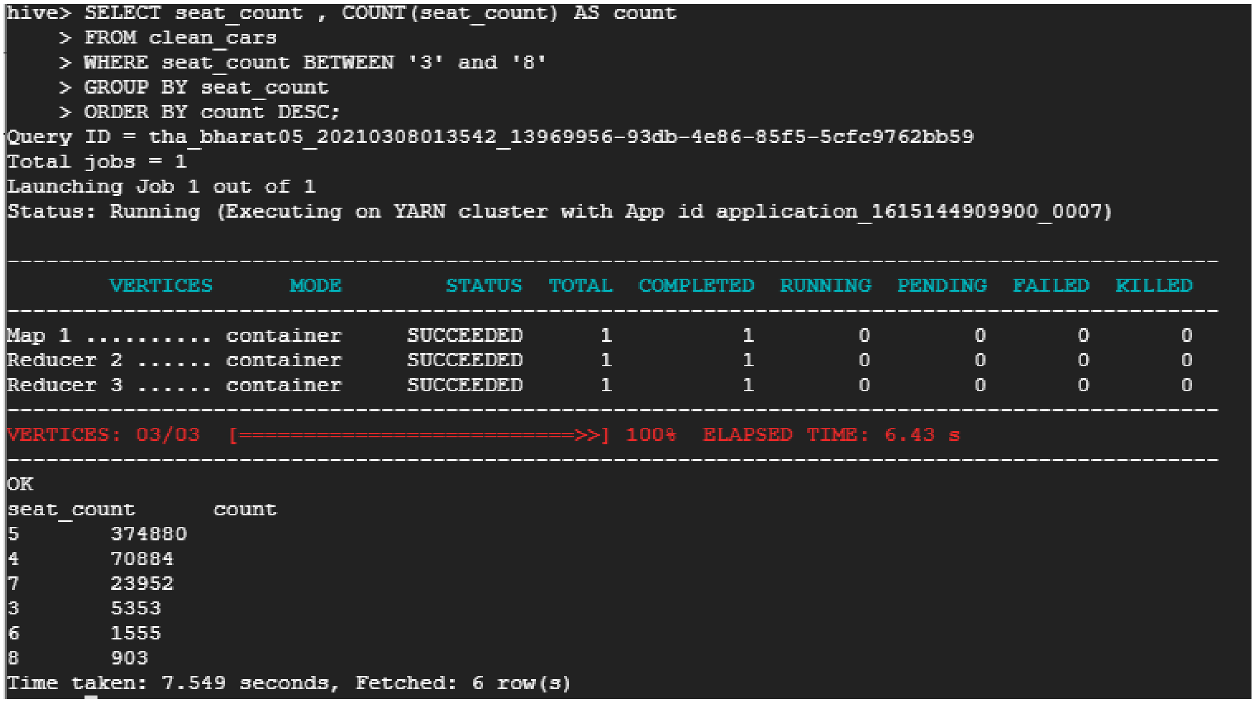
```
-- Check different fuel types and their counts
SELECT fuel_type , COUNT(fuel_type) AS count
FROM clean_cars
GROUP BY fuel_type
ORDER BY count DESC;
```



```
hive> SELECT door_count , COUNT(door_count) AS count
    > FROM clean_cars
    > GROUP BY door_count
    > ORDER BY count DESC;
Query ID = tha_bharat05_20210308012412_e572a4a4-0ef9-49e1-b96f-ed3a83cbad38
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1615144909900_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 6.44 s
----------------------------------------------------------------------------------------------
OK
door_count      count
4       252706
5       197057
2       71056
3       31893
6       140
1       11
7       3
54      1
58      1
NULL    0
```

1. 4 doored cars have the highest count. Still there are few vehicles with unreasonable door counts.

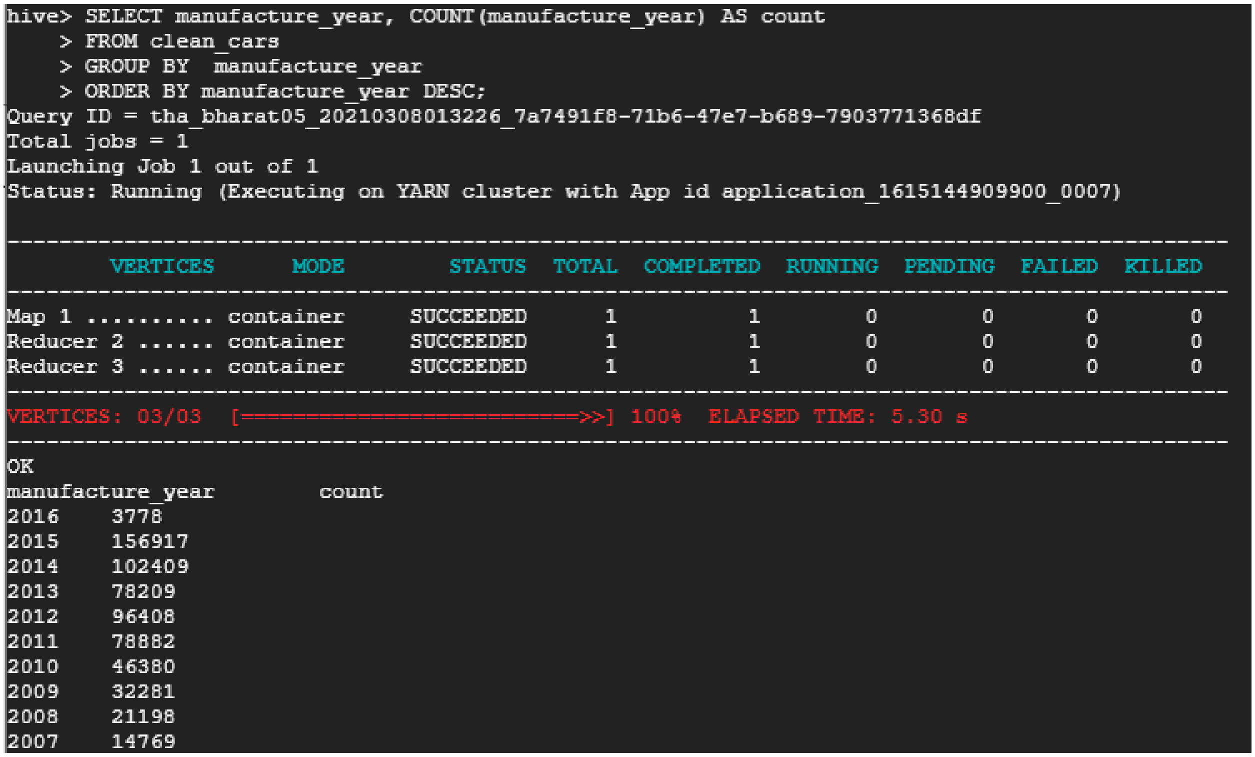
```sql
-- Check seat_count and their counts
SELECT seat_count , COUNT(seat_count) AS count
FROM clean_cars
WHERE seat_count BETWEEN '3' and '8'
GROUP BY seat_count
ORDER BY count DESC;
```



```
hive> SELECT seat_count , COUNT(seat_count) AS count
    > FROM clean_cars
    > WHERE seat_count BETWEEN '3' and '8'
    > GROUP BY seat_count
    > ORDER BY count DESC;
Query ID = tha_bharat05_20210308013542_13969956-93db-4e86-85f5-5cfc9762bb59
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 6.43 s
----------------------------------------------------------------------------------------------
OK
seat_count      count
5       374880
4       70884
7       23952
3       5353
6       1555
8       903
Time taken: 7.549 seconds, Fetched: 6 row(s)
```

1. It appears that the most of the cars in clean dataset have seat count of 5.

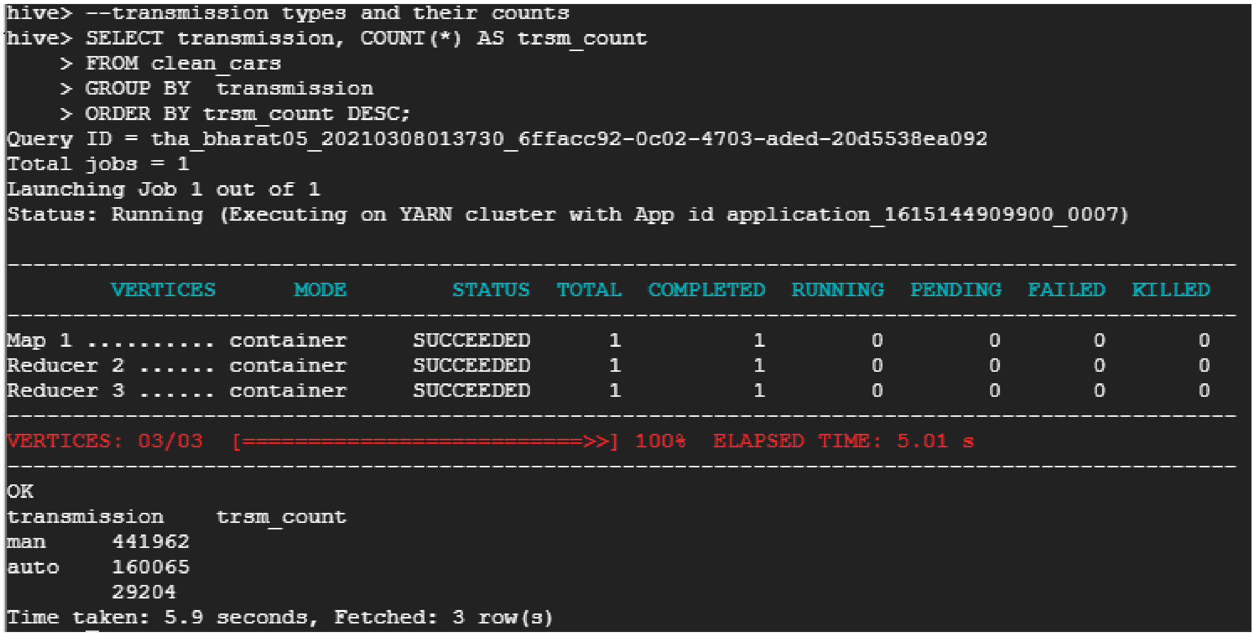
```sql
--- manufacturing year and their count
SELECT manufacture_year, COUNT(manufacture_year) AS count
FROM clean_cars
GROUP BY  manufacture_year
ORDER BY manufacture_year DESC;
```



```
hive> SELECT manufacture_year, COUNT(manufacture_year) AS count
    > FROM clean_cars
    > GROUP BY  manufacture_year
    > ORDER BY manufacture_year DESC;
Query ID = tha_bharat05_20210308013226_7a7491f8-71b6-47e7-b689-7903771368df
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 5.30 s
----------------------------------------------------------------------------------------------
OK
manufacture_year        count
2016    3778
2015    156917
2014    102409
2013    78209
2012    96408
2011    78882
2010    46380
2009    32281
2008    21198
2007    14769
```

1. Data cleaning led to removal of all of cars which were manufacture in 2017 and most of cars from 2016.

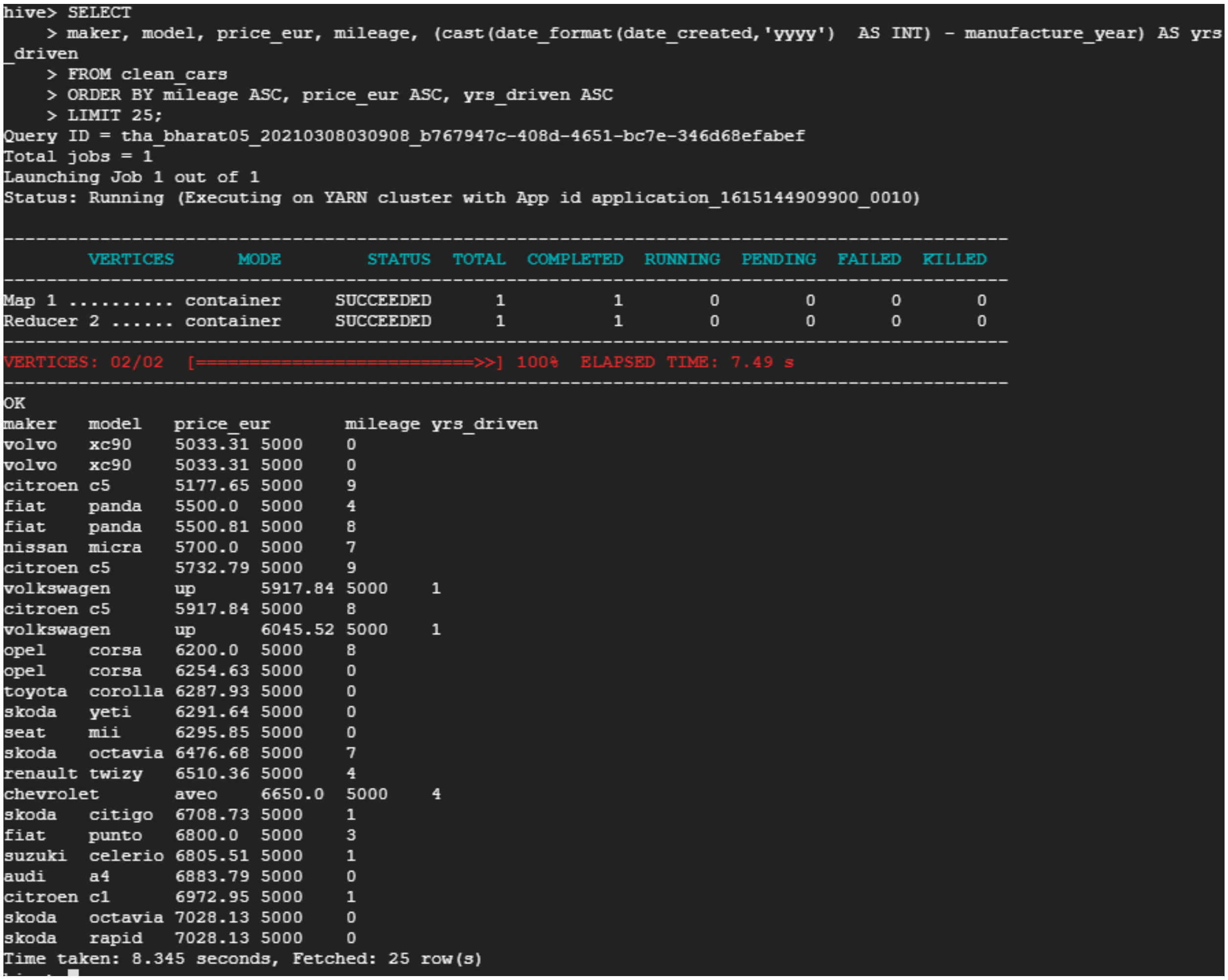
```
--transmission types and their counts
SELECT transmission, COUNT(*) AS trsm_count
FROM clean_cars
GROUP BY  transmission
ORDER BY trsm_count DESC;
```



```
hive> --transmission types and their counts
hive> SELECT transmission, COUNT(*) AS trsm_count
    > FROM clean_cars
    > GROUP BY  transmission
    > ORDER BY trsm_count DESC;
Query ID = tha_bharat05_20210308013730_6ffacc92-0c02-4703-aded-20d5538ea092
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        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: 5.01 s
----------------------------------------------------------------------------------------------
OK
transmission    trsm_count
man     441962
auto    160065
        29204
Time taken: 5.9 seconds, Fetched: 3 row(s)
```

1. There are 4 times as much manual cars as there are automatic cars.

```
-- Creating new column yrs_driven
SELECT
maker, model, price_eur, mileage, (cast(date_format(date_created,'yyyy')  AS
INT) - manufacture_year) AS yrs_driven
FROM clean_cars
ORDER BY mileage ASC, price_eur ASC, yrs_driven ASC
LIMIT 25;
```



```
hive> SELECT
    > maker, model, price_eur, mileage, (cast(date_format(date_created,'yyyy')  AS INT) - manufacture_year) AS yrs
_driven
    > FROM clean_cars
    > ORDER BY mileage ASC, price_eur ASC, yrs_driven ASC
    > LIMIT 25;
Query ID = tha_bharat05_20210308030908_b767947c-408d-4651-bc7e-346d68efabef
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1615144909900_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.49 s
----------------------------------------------------------------------------------------------
OK
maker   model   price_eur       mileage yrs_driven
volvo   xc90    5033.31 5000    0
volvo   xc90    5033.31 5000    0
citroen c5      5177.65 5000    9
fiat    panda   5500.0  5000    4
fiat    panda   5500.81 5000    8
nissan  micra   5700.0  5000    7
citroen c5      5732.79 5000    9
volkswagen      up      5917.84 5000    1
citroen c5      5917.84 5000    8
volkswagen      up      6045.52 5000    1
opel    corsa   6200.0  5000    8
opel    corsa   6254.63 5000    0
toyota  corolla 6287.93 5000    0
skoda   yeti    6291.64 5000    0
seat    mii     6295.85 5000    0
skoda   octavia 6476.68 5000    7
renault twizy   6510.36 5000    4
chevrolet       aveo    6650.0  5000    4
skoda   citigo  6708.73 5000    1
fiat    punto   6800.0  5000    3
suzuki  celerio 6805.51 5000    1
audi    a4      6883.79 5000    0
citroen c1      6972.95 5000    1
skoda   octavia 7028.13 5000    0
skoda   rapid   7028.13 5000    0
Time taken: 8.345 seconds, Fetched: 25 row(s)
```

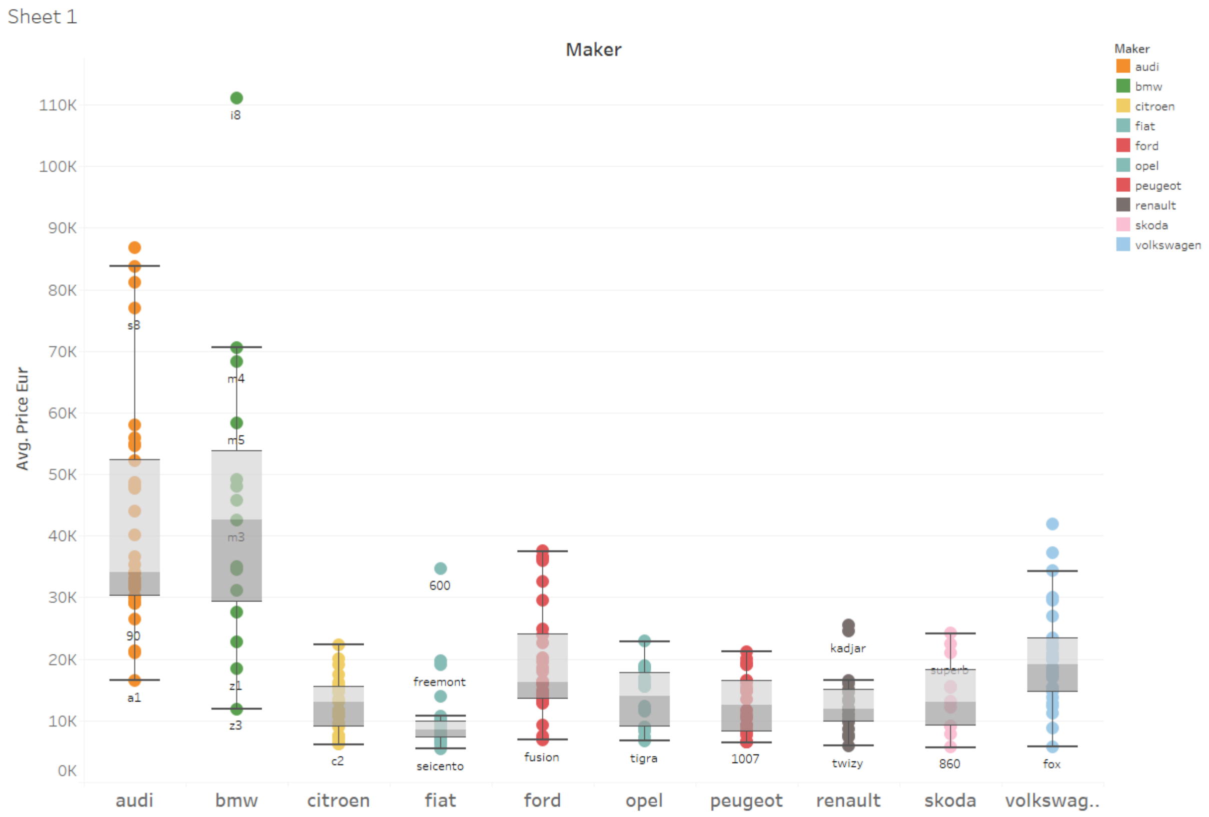
1. Years driven can be used to quickly sort relatively newer cars.

## Key Findings from Clean Dataset

- In the clean dataset there are 559 car models from 43 manufacturers.
- Now, Volkswagen Golf has the highest availability.
- In the filtered dataset Porsche Panamera is the most expensive car.
- There are 4 times as much manual cars as there are automatic cars.
- Average car mileage is around 40,000km and car price is 17,600 euros.
- Data cleaning led to removal of all of cars which were manufacture in 2017 and most of cars from 2016.
- This might lead to distortion of sales statistics.
- Most of the cars have 4 doors and 5 seats.
- It appears data cleaning led to removal of electric, cng and lpg cars.
- In the clean dataset there is almost equal number of gasoline powered and diesel powered cars.
- Some prices on the lowest end appear too good to be true, buyer must exercise great caution while making such a purchase.

## Analysis (Questions)

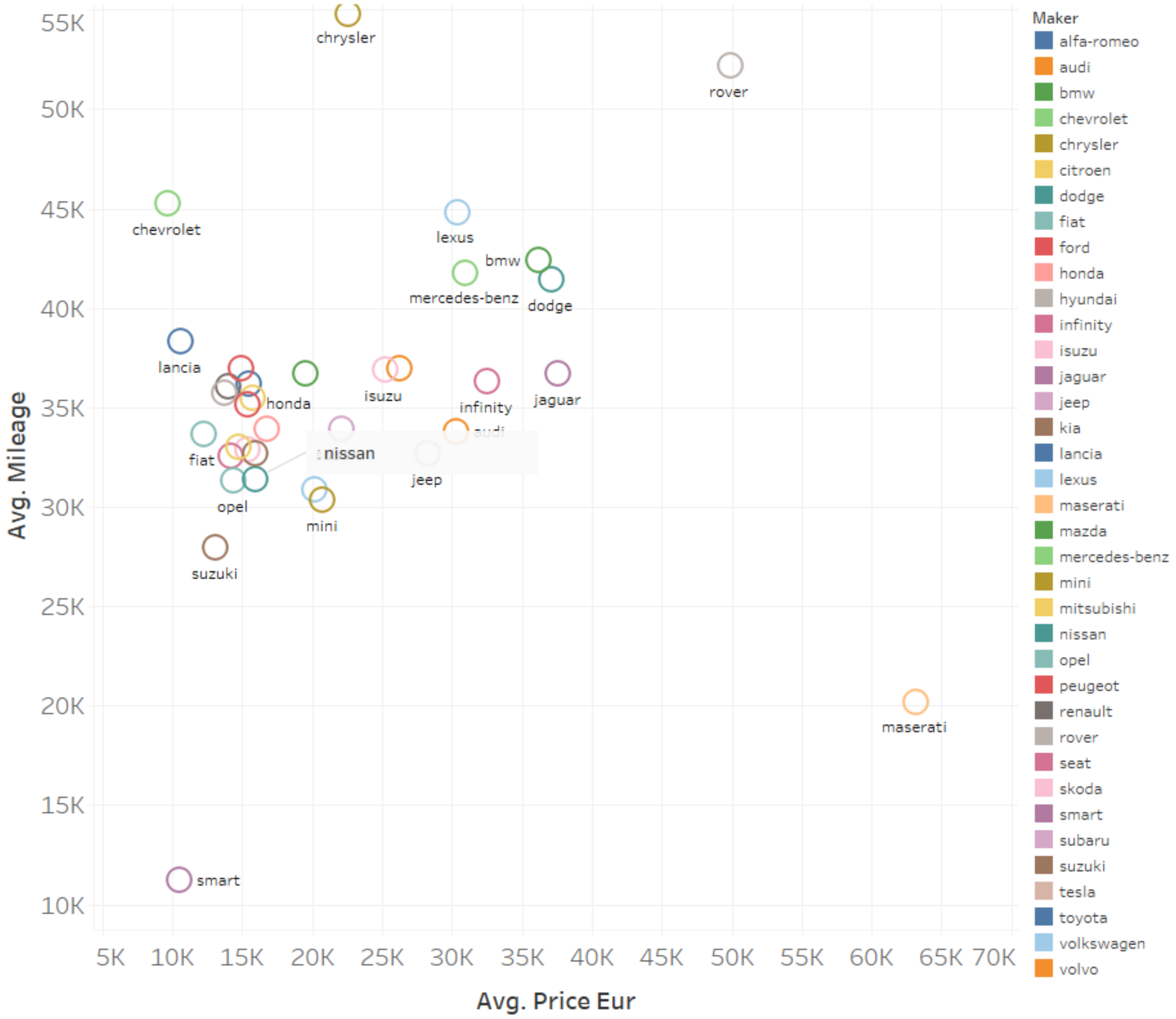
### 1. What is the relationship between car makes, models and price?



Average of Price Eur for each Maker. Color shows details about Maker. The marks are labeled by Model. The view is filtered on Maker, which keeps 10 of 43 members.

Car prices vary a great deal across different models for a particular make and also they are very different for different manufacturers.

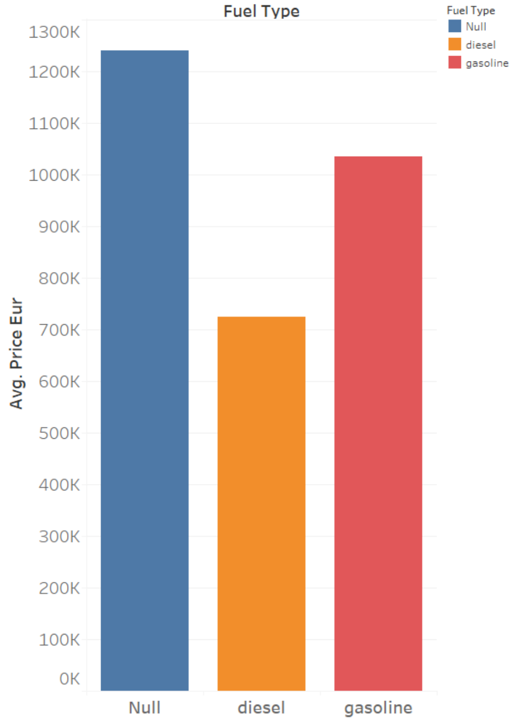
## 2. What are the top five vehicle manufacturers would you recommend? Why?



Average of Price Eur vs. average of Mileage. Color shows details about Maker. The marks are labeled by Maker. The data is filtered on Manufacture Year and Seats Count. The Manufacture Year filter keeps 2012, 2013, 2014, 2015 and 2016. The Seats Count filter keeps 5. The view is filtered on Maker, which excludes aston-martin, bentley, lamborghini and porsche.

Based on the scatter plot of Avg. Mileage and Avg. Price for vehicle manufactured from 2012-2016 (recent) and with 5 seats(most popular segment), we would want cars with lower mileage and lower price. So this would mean we should prefer car manufacturer on the bottom left corner of the plot i.e., Smart, Suzuki, Opel, Nissan, and Fiat if we are looking for biggest bang for the buck. Off course the answer will change depending upon budget, car segment etc.

### 3. Does fuel type have any impact on the car price? Explain



Average of Price Eur for each Fuel Type. Color shows details about Fuel Type. Details are shown for Maker.

If we exclude the Null values, it appears Diesel cars are cheaper than Gasoline cars. This could be attributed to two reasons. First higher price of Diesel fuel in Europe and secondly higher maintenance costs of Diesel engines as compared to their Gasoline counterparts.

# Appendix

**Data Dictionary of Extracted Dataset**

| | **Column** | **Data Type** | **Description** |
|---|---|---|---|
| 1 | maker | String | Name of car manufacturer |
| 2 | model | String | Name of car model |
| 3 | mileage | Float | Total distance travelled (km) |
| 4 | manufacture_year | Integer | Year in which car was manufactured. |
| 5 | transmission | String | The type of vehicle transmission – manual or automatic |
| 6 | door_count | Integer | The number of doors in the vehicle. |
| 7 | seat_count | Integer | The number of seats in the vehicle. |
| 8 | fuel_type | String | Type of fuel – gasoline/diesel/electric etc. |
| 9 | date_created | Date | The date on which ad was scraped |
| 10 | data_last_seen | Date | The date of the last time the ad was on the website |
| 11 | price_eur | Float | The vehicle price in Euro |