11/14/2021

ANALYSIS REPORT ON USED CARS

Amith Heiden

200497933

CONTENTS

INTRODUCTION: 2
Research Questions: 2
DATA CLEANING: 3
ANALYSIS AND VISUALIZATIONS: 4
CONCLUSION 12
APPENDIX 13

INTRODUCTION:

The report illustrates about the analysis on Used cars dataset which has certain variables.

There are roughly 3.5 Million rows and the following columns were present:

- maker – Make of the car
- model – Model of the car make
- mileage – Kilometres covered by the car
- manufacture_year – Year of car manufactured
- engine_displacement - Size of piston cylinder in CC
- engine_power - Power of Engine
- body_type – Type of body
- color_slug – Colour of car
- stk_year – Year of the last emission control
- transmission – Automatic or manual
- door_count – Number of doors
- seat_count – Number of seats
- fuel_type – Fuel type of cars
- date_created – When the ad was created
- date*last*seen – When the ad was last seen
- price_eur – Price in Euro

The purpose of the analysis is to determine the best car makers and models into which the firm would like to invest based on the outcomes of analysis.

RESEARCH QUESTIONS:

Below are the research questions that were taken into considerations for analysing the data and produce decision making results.

1. Which fuel type was most dominating?
   Among the fuel types such as diesel, gasoline, lpg, cng and electric, the analysis on the most preferred fuel type resulted to be Diesel and Gasoline.

2. Who were the Top 5 car manufacturers over a period of 8 years?
   Top 5 manufacturers will be determined based on the Total count of cars produced over the years

3. What was the seat count most commonly preferred?
   Seat counts play a major role in size of the car. Analysing on the count of seats over car will provide the most preferred seat count which would assist in focusing on the cars that would be sold at a maximum rate

4. What was the most popular door count for the top 5 manufacturers?
   Analysing door counts will result in focusing on the cars with door counts that were used and preferred the most.

5. Do the Top 5 manufacturers has the same level of productivity over the years?
   Analysing the Top 5 manufacturers productivity rate over the period of years will enunciate in which year each of the car manufacturers have their maximum productivity. The cars are more preferred to purchase if it is manufactured in the latter part of the period.

6. What is the impact of Price and mileage over models of each manufacturers?
   The analysis over the models based on the Price and mileage will assist in selecting the best cars to be invested further. Cars with low price and appropriate mileage coverage will be highly preferred.

DATA CLEANING:

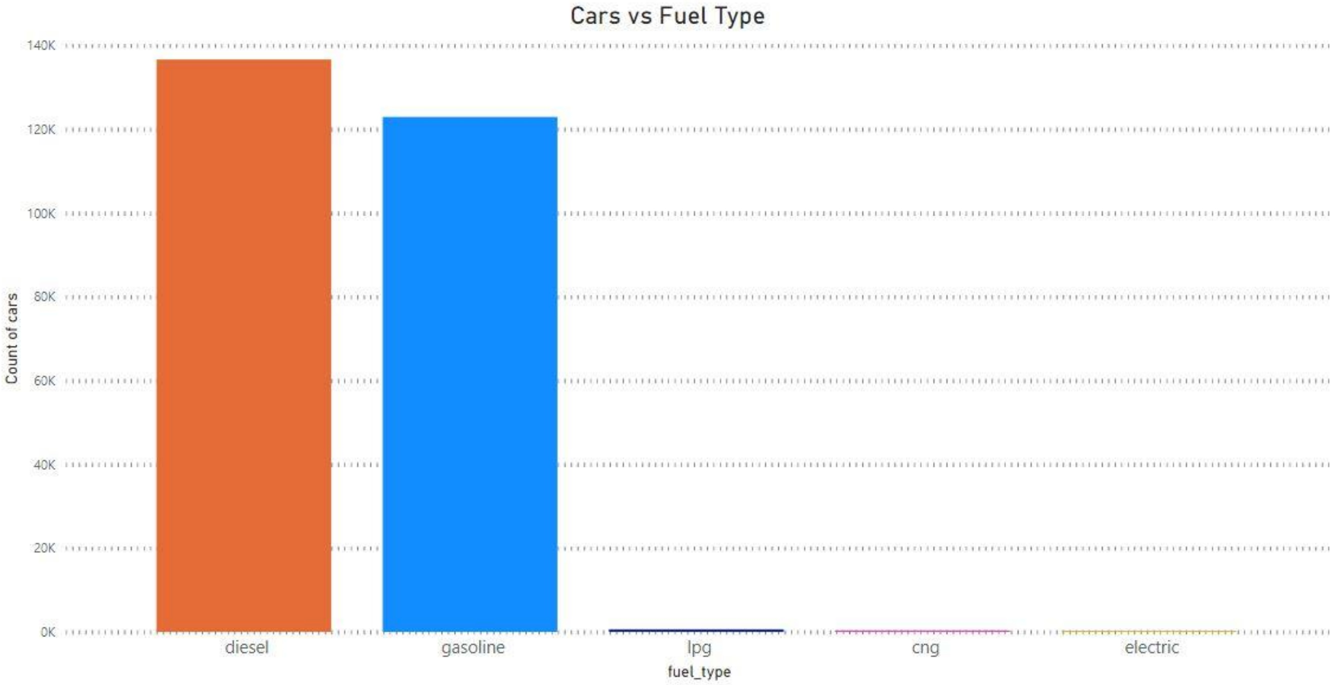
The Provided data is cleaned before Analysis and decision making.

Below are some of the Data Cleaning executed:

1. The columns body type, colour slug and stk year are removed due to insufficient data
2. Maker and Model columns have empty and null spaces which are cleaned
3. Mileage, Manufacture year, Engine Power and displacement, Transmission, Door count, Seat Count, Price Eur have Null values which were removed.
4. Manufacture year had Outliers such as 0,1,2 as years which is not possible. Hence they are removed
5. Door Count and Seat Counts had Outliers such as 1 and 0 which are out of context. These Outliers are removed for Analysis
6. The format of Date created and Date Last Seen are changed for calculating the Date gap for Analysis
7. Price_eur has outliers stating values in Billions which are cleaned for Analysis

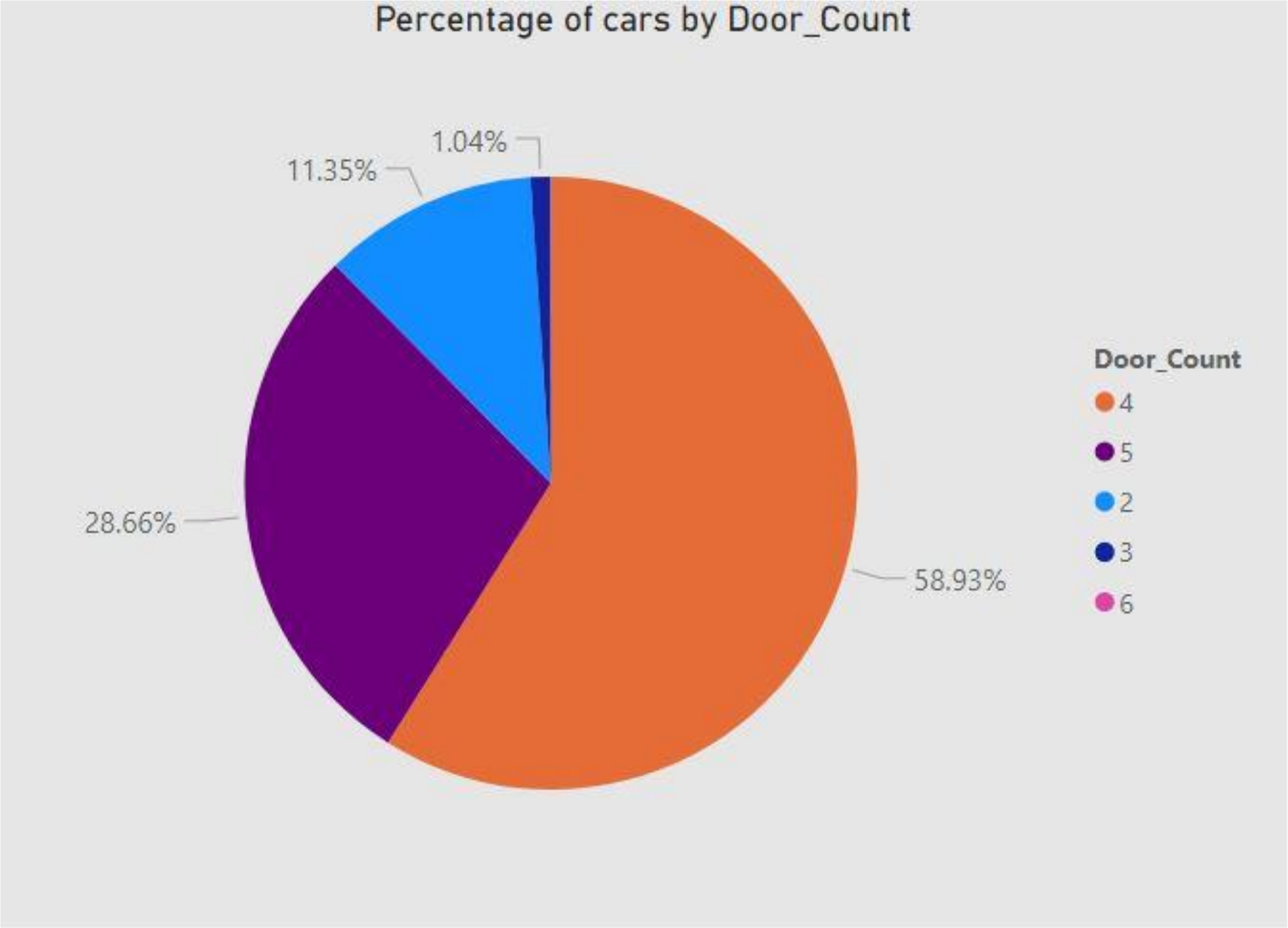
(Refer Appendix for cleaning pathway)

ANALYSIS AND VISUALIZATIONS:



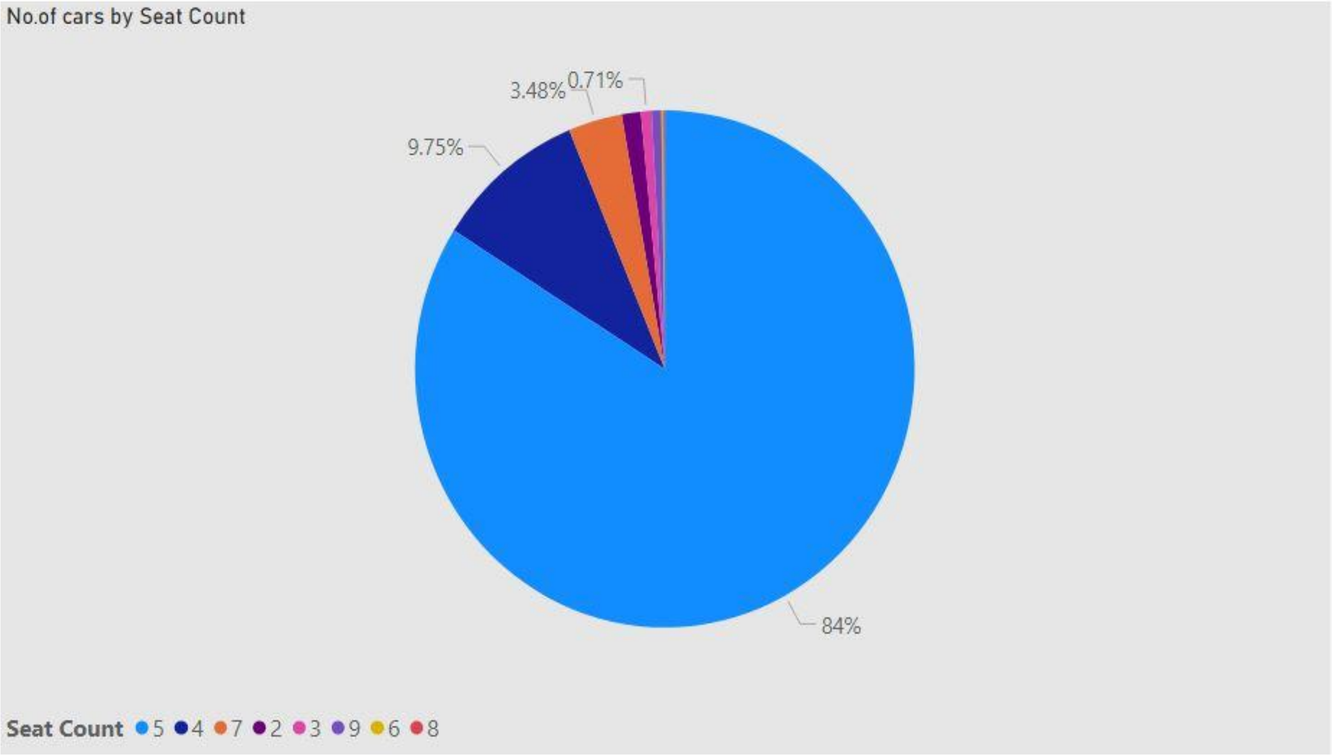
The above column chart shows the count for Fuel Type. As explained in the Research Question

- Diesel and Gasoline were used by most of the cars
- LPG CNG and Electric cars are few in number compared to Diesel and Gasoline cars
- Taking only Diesel and Gasoline cars for further analysis



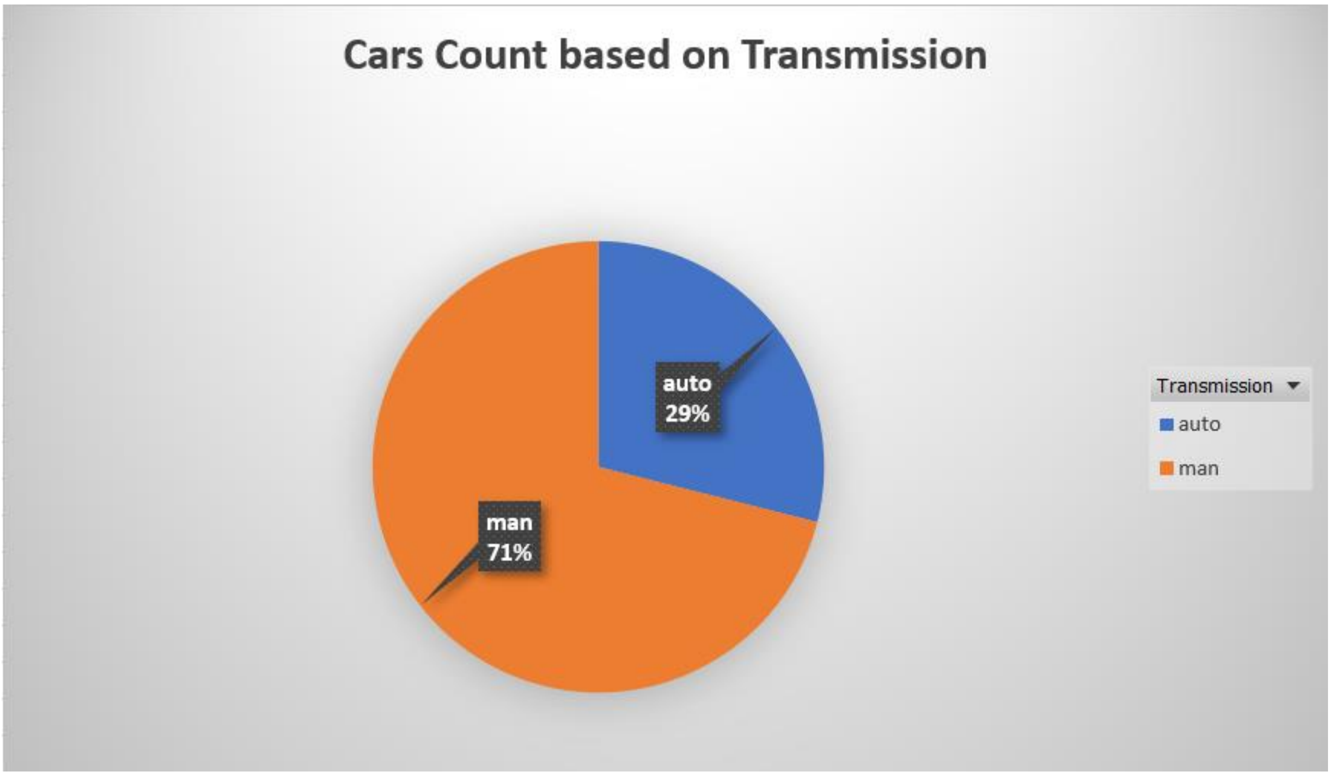
From the Pie chart It is clear that

- The Number of cars with 4 doors is above 50 percent of the total number of cars
- Cars with 4 doors are more popular, followed by cars with 5 doors

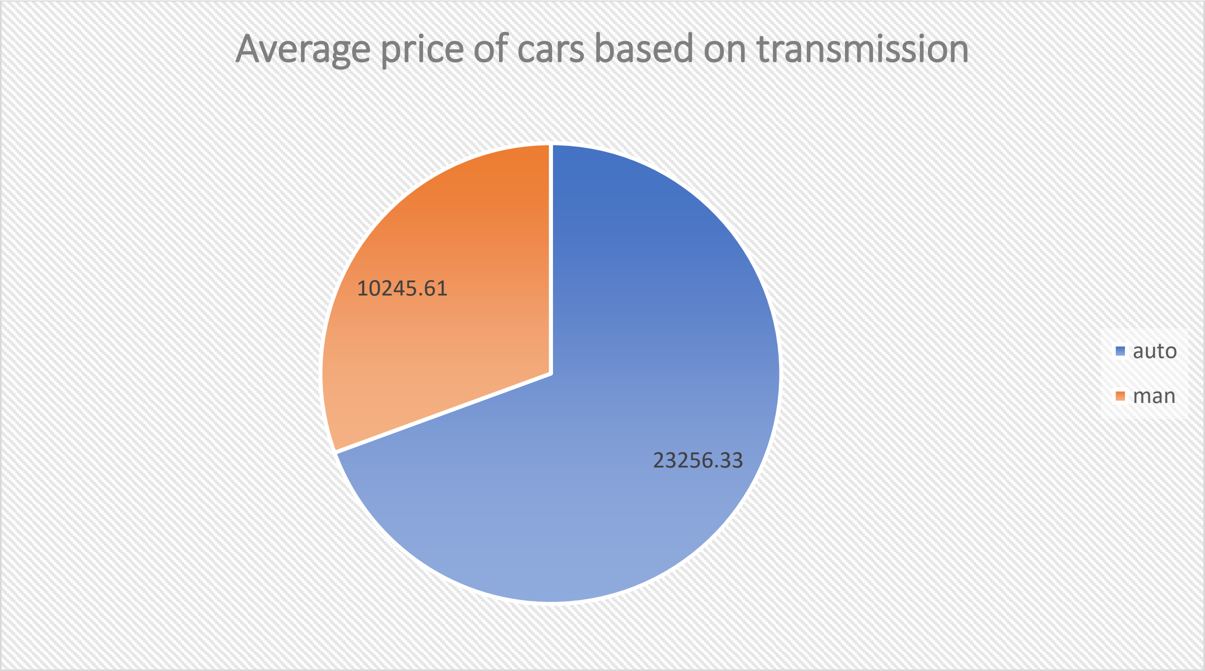


From the above graph It can be seen that:

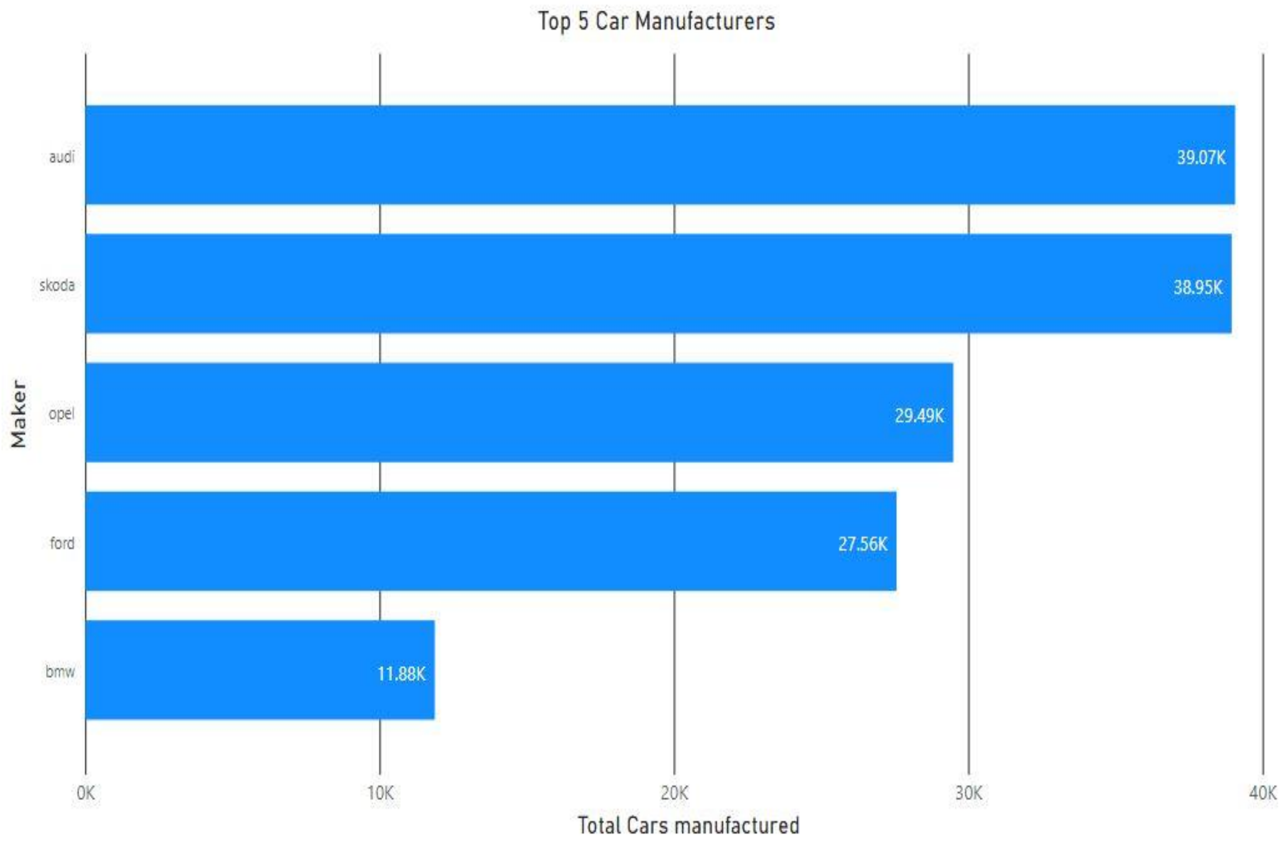
- Cars with seats cover up to 50 percent of the total car count and hence cars with 5 seats are more common.



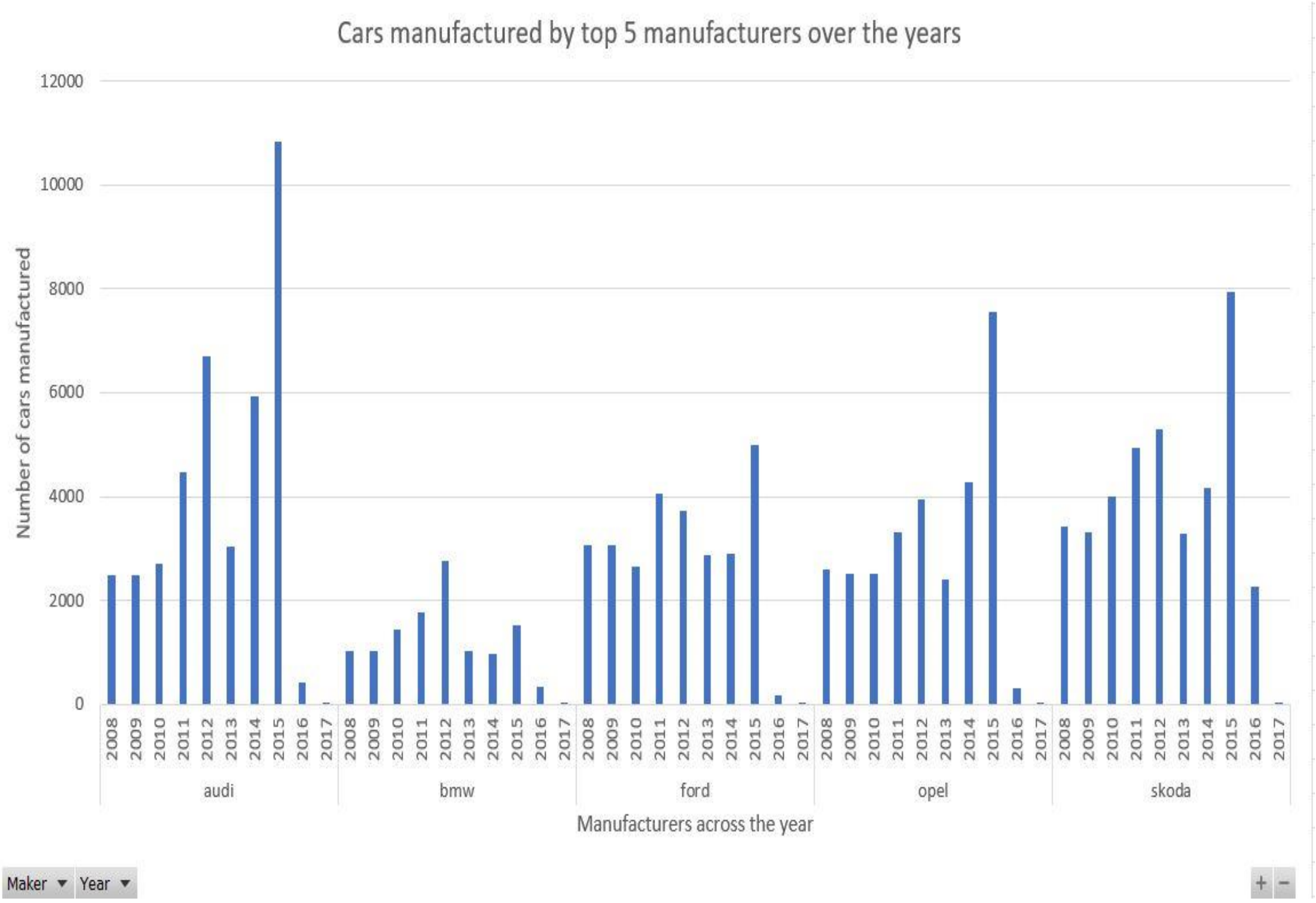
The graph shows that cars with manual transmission are more common than auto transmission.



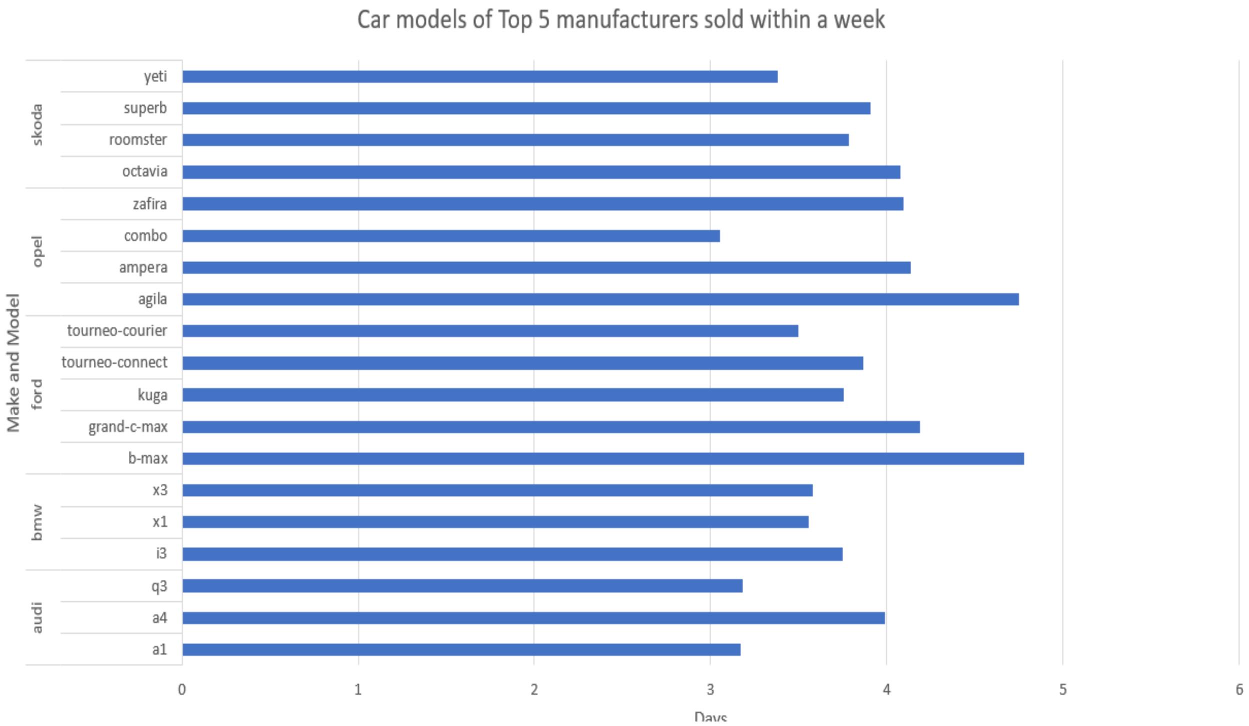
The above graph states that the average price of automatic transmission cars are higher than the average price of manual transmission cars despite the count of transmission



- The graph illustrates the top 5 makers of car based on the number of cars manufactured in a total
- Top 5 manufacturers include Audi, Skoda, Opel, Ford and BMW.
- These top 5 manufacturers are of Gasoline and Diesel Fuel type with seat and door counts equal to 4 and 5



- This subsidiary of the previous graph illustrates about the number of cars manufactured every year for the top 5 manufacturers
- Except for BMW all the other car makers have their maximum productivity during the period of 2015
- Audi has the highest productivity overall during 2015



Above graph illustrates the car models of the Top 5 manufacturers based on the ad display date and ad last seen date.

These cars were sold in a period of a week.

Based on the above Analysis, the below table suggests models of cars from the top 5 manufacturers considering their average Mileage and Average Price and the year of manufacture.

Maker	Model	Manufacturing Year	Avg Mileage	Avg Price
audi	a1	2015	10067.14	18,314.19 €
	a4	2016	9611.76	20,387.40 €
	a1	2016	2998.13	21,278.29 €
	q3	2015	11025.29	32,235.64 €
	q3	2016	3026.59	38,765.06 €
skoda	citigo	2010	11850.00	8,508.51 €
	roomster	2014	22897.87	9,491.54 €
	roomster	2015	14113.93	11,321.91 €
	octavia	2015	21386.24	12,111.94 €
	yeti	2015	15753.16	13,216.17 €
	superb	2015	22848.76	19,499.68 €
opel	adam	2015	8700.00	10,046.45 €
	agila	2015	120.00	12,395.93 €
	combo	2015	6442.86	15,575.02 €
	zafira	2016	2763.33	16,227.10 €
	zafira	2015	11046.86	18,444.69 €
	combo	2016	3000.00	20,900.00 €
	ampera	2014	10700.00	22,405.94 €
	ampera	2015	7900.00	26,909.16 €
	ampera	2016	3434.00	33,434.00 €
ford	kuga	2016	11237.33	11,652.80 €
	b-max	2015	11858.70	14,537.93 €
	grand-c-max	2015	10855.89	17,774.27 €
	b-max	2016	100.00	17,990.00 €
	tourneo-connect	2016	2004.00	17,995.00 €
	tourneo-connect	2015	10487.49	18,717.73 €
	grand-c-max	2016	5000.00	22,533.88 €
bmw	x1	2016	7907.11	10,825.01 €
	x3	2016	6165.26	27,317.92 €
	i3	2015	8881.17	35,695.64 €
	i3	2014	11150.29	37,036.96 €

Table 1

CONCLUSION

The Analysis emphasises that all the Research Questions are justified.

From the Analysis the following justifications were made:

1. Diesel and Gasoline are the most commonly used fuel types among all car brands
2. The most commonly preferred door and seat counts were calculated and thus the corresponding cars were determined
3. Manual transmission was predominantly used and found to be cheaper than Auto transmission
4. Top 5 car manufacturers based on the productivity rate were determined
5. Distribution of car productivity for the top 5 manufacturers for each year was compared and justified
6. Cars whose ad were posted and removed within a week were determined

7. The car models which have an appropriate mileage and considerable price is determined for the Top 5 car manufacturers as referred in **Table 2**

APPENDIX

Analysis Pathway:

Data Cleaning → Shortlisting the cars based on fuel type → Determining the most popular seat counts and door counts → Determining the Top 5 manufacturers based on productivity over years considering the previously analysed conditions→ Filtering the models of the Top 5 manufacturers based on Mileage, Price and Year of Manufacture

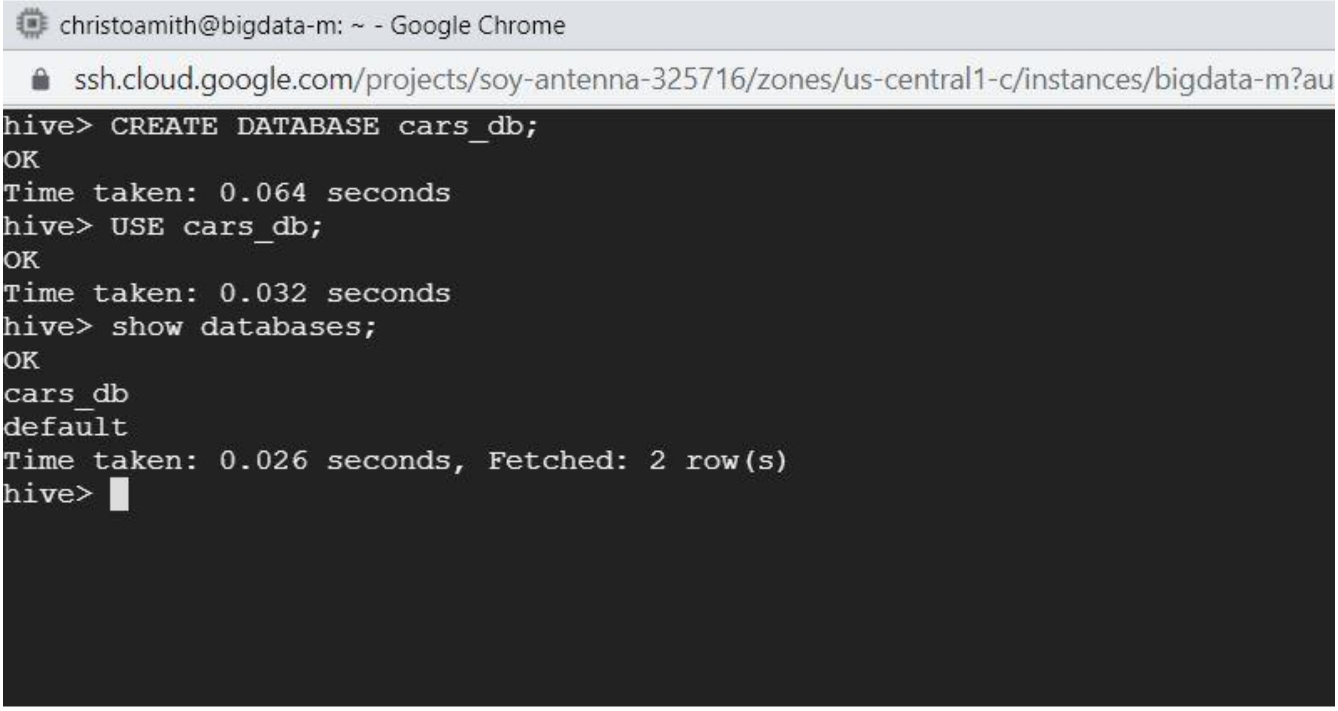
Tools Used for Analysis:

- Apache HIVE
- Excel
- Power BI

Data Cleaning Pathway:

STEP 1:

Creating database in HIVE

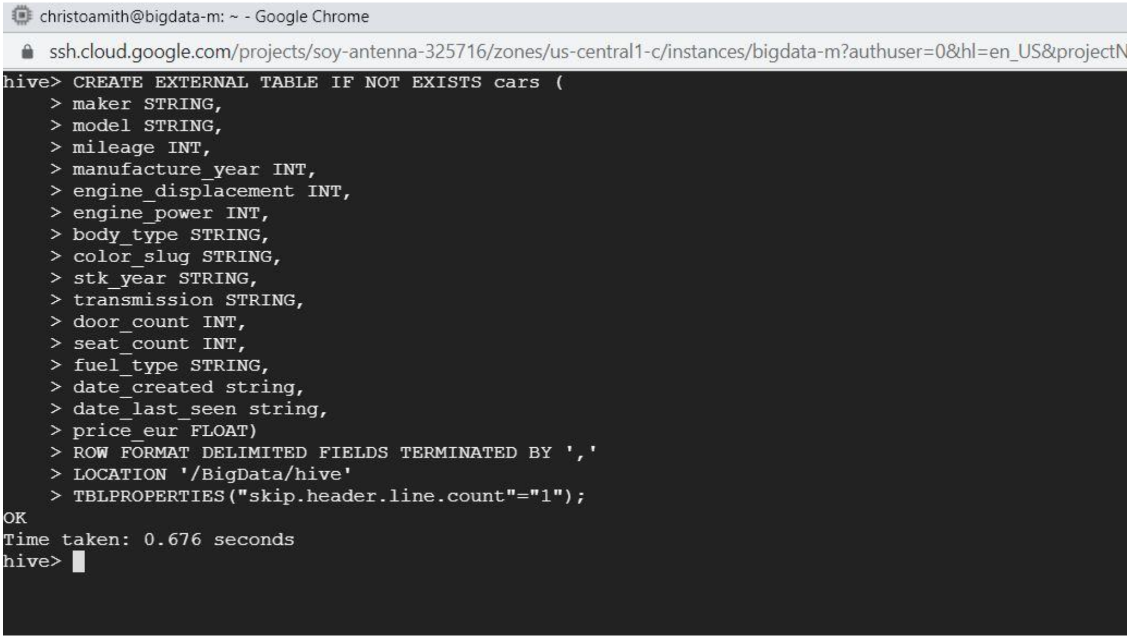


christoamith@bigdata-m: ~ - Google Chrome
ssh.cloud.google.com/projects/soy-antenna-325716/zones/us-central1-c/instances/bigdata-m?au

```
hive> CREATE DATABASE cars_db;
OK
Time taken: 0.064 seconds
hive> USE cars_db;
OK
Time taken: 0.032 seconds
hive> show databases;
OK
cars_db
default
Time taken: 0.026 seconds, Fetched: 2 row(s)
hive>
```

STEP 2a:

Creating raw data table in cars_db

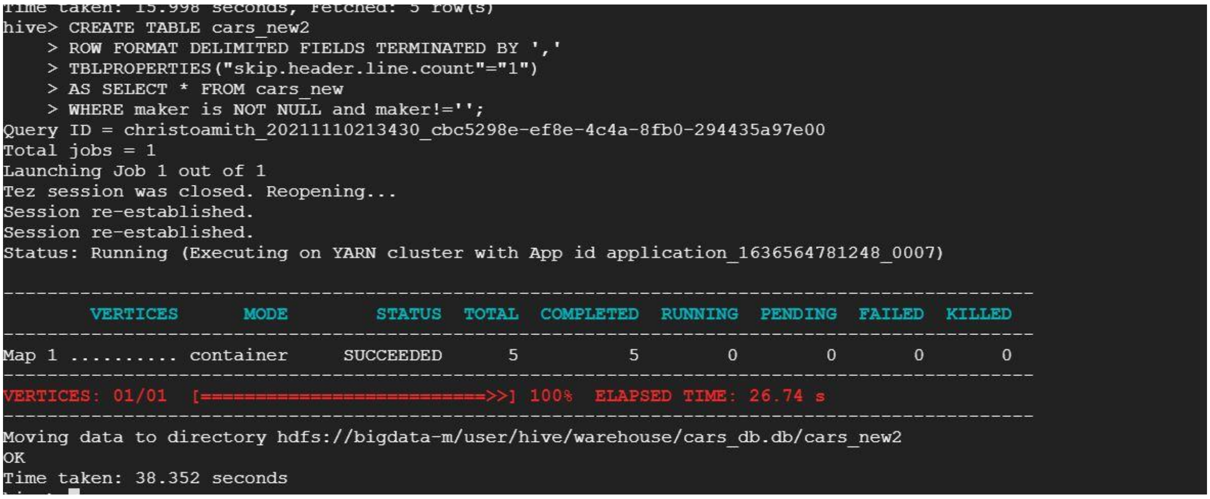


christoamith@bigdata-m: ~ - Google Chrome
ssh.cloud.google.com/projects/soy-antenna-325716/zones/us-central1-c/instances/bigdata-m?authuser=0&hl=en_US&projectN

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS cars (
    > maker STRING,
    > model STRING,
    > mileage INT,
    > manufacture_year INT,
    > engine_displacement INT,
    > engine_power INT,
    > body_type STRING,
    > color_slug STRING,
    > stk_year STRING,
    > transmission STRING,
    > door_count INT,
    > seat_count INT,
    > fuel_type STRING,
    > date_created string,
    > date_last_seen string,
    > price_eur FLOAT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/BigData/hive'
    > TBLPROPERTIES("skip.header.line.count"="1");
OK
Time taken: 0.676 seconds
hive> 
```

STEP 2b:

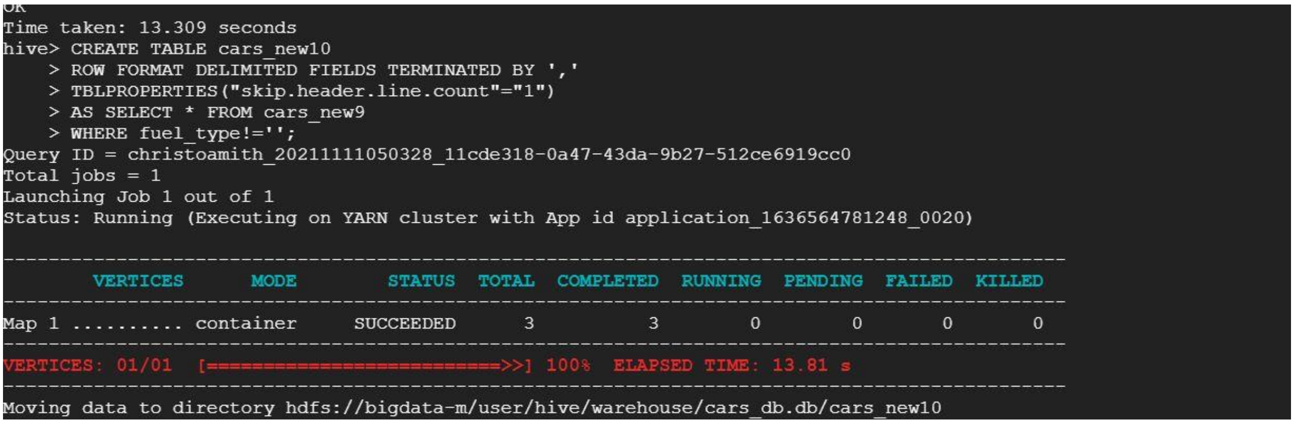
Cleaning makers column



```
Time taken: 15.998 seconds, Fetched: 5 row(s)
hive> CREATE TABLE cars_new2
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > TBLPROPERTIES("skip.header.line.count"="1")
    > AS SELECT * FROM cars_new
    > WHERE maker is NOT NULL and maker!='';
Query ID = christoamith_20211110213430_cbc5298e-ef8e-4c4a-8fb0-294435a97e00
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_1636564781248_0007)

----------------------------------------------------------------------------------------------
        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: 26.74 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cars_db.db/cars_new2
OK
Time taken: 38.352 seconds
```

STEP 3a:

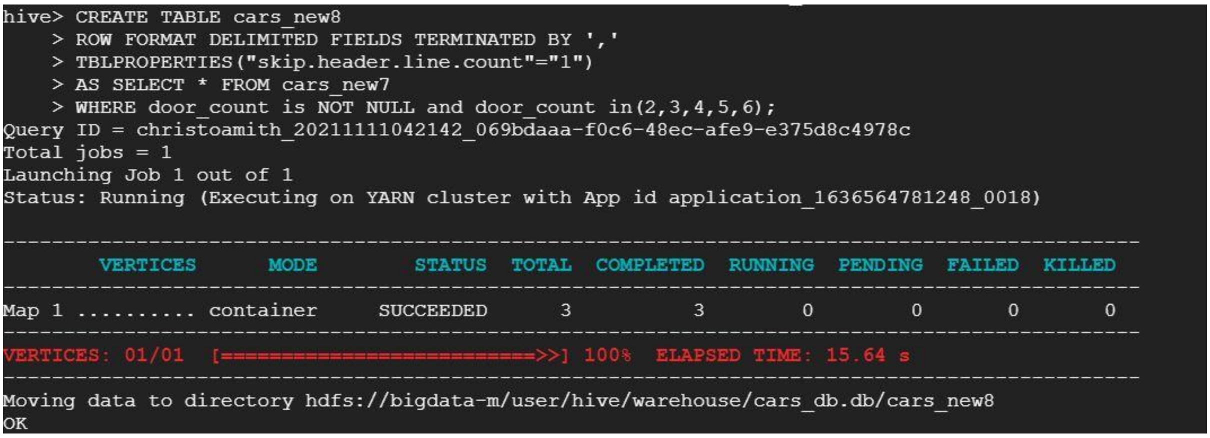
Cleaning Fuel_type column



```
OK
Time taken: 13.309 seconds
hive> CREATE TABLE cars_new10
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > TBLPROPERTIES("skip.header.line.count"="1")
    > AS SELECT * FROM cars_new9
    > WHERE fuel_type!='';
Query ID = christoamith_20211111050328_11cde318-0a47-43da-9b27-512ce6919cc0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1636564781248_0020)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 13.81 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cars_db.db/cars_new10
```

STEP 4:

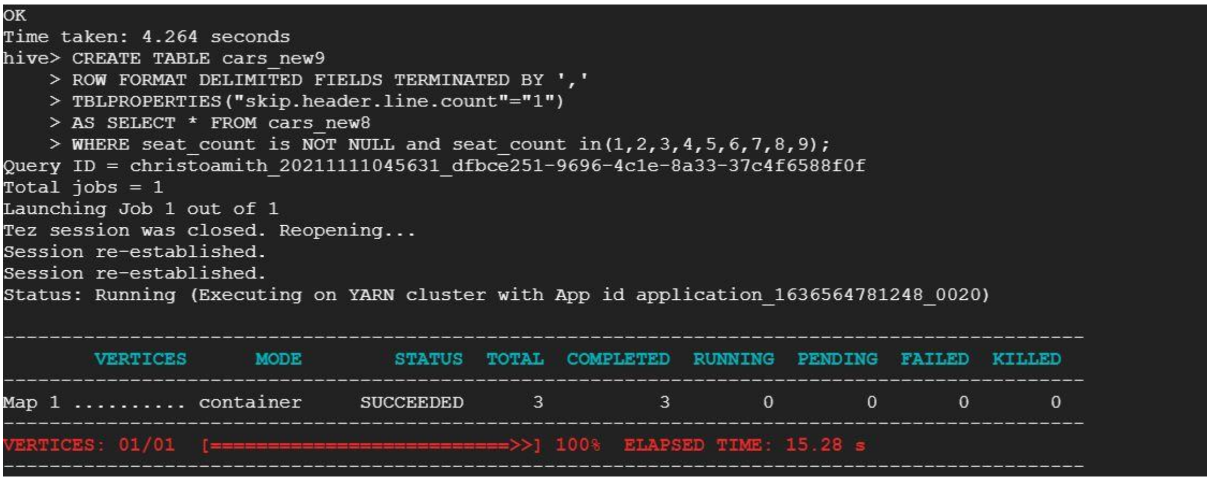
Cleaning door_count column



```
hive> CREATE TABLE cars_new8
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > TBLPROPERTIES("skip.header.line.count"="1")
    > AS SELECT * FROM cars_new7
    > WHERE door_count is NOT NULL and door_count in(2,3,4,5,6);
Query ID = christoamith_20211111042142_069bdaaa-f0c6-48ec-afe9-e375d8c4978c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1636564781248_0018)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 15.64 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cars_db.db/cars_new8
OK
```

STEP 5:

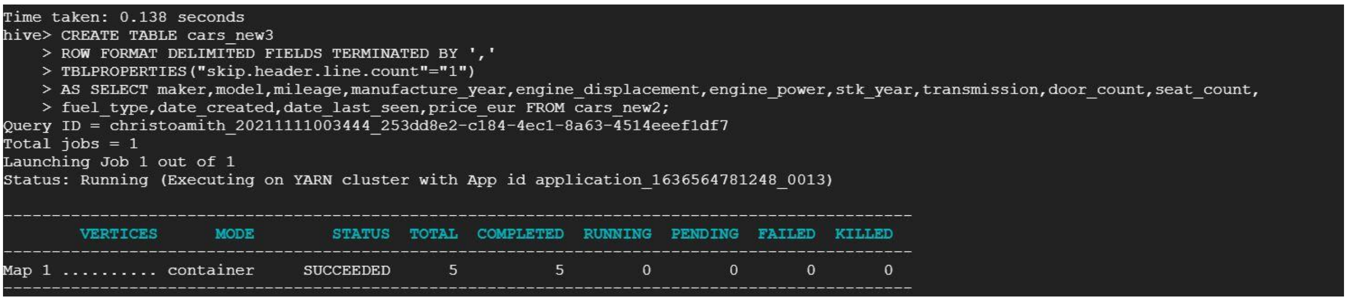
Cleaning the Seat_Count column



```
OK
Time taken: 4.264 seconds
hive> CREATE TABLE cars_new9
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > TBLPROPERTIES("skip.header.line.count"="1")
    > AS SELECT * FROM cars_new8
    > WHERE seat_count is NOT NULL and seat_count in(1,2,3,4,5,6,7,8,9);
Query ID = christoamith_20211111045631_dfbce251-9696-4c1e-8a33-37c4f6588f0f
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_1636564781248_0020)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 15.28 s
----------------------------------------------------------------------------------------------
```

STEP 6:

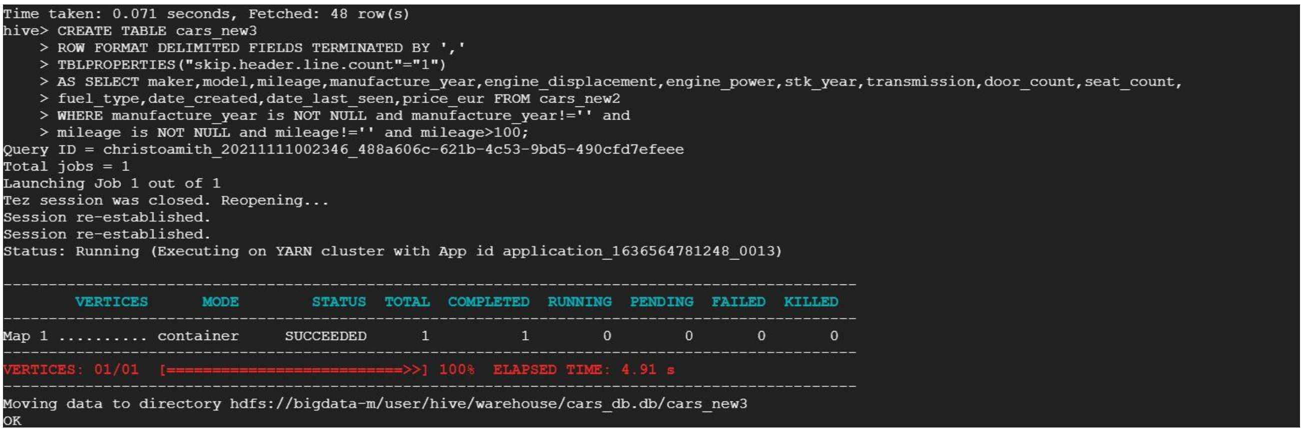
Eliminating the columns with insufficient data



```
Time taken: 0.138 seconds
hive> CREATE TABLE cars_new3
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > TBLPROPERTIES("skip.header.line.count"="1")
    > AS SELECT maker,model,mileage,manufacture_year,engine_displacement,engine_power,stk_year,transmission,door_count,seat_count,
    > fuel_type,date_created,date_last_seen,price_eur FROM cars_new2;
Query ID = christoamith_20211111003444_253dd8e2-c184-4ec1-8a63-4514eeef1df7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1636564781248_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
----------------------------------------------------------------------------------------------
```

STEP 7:

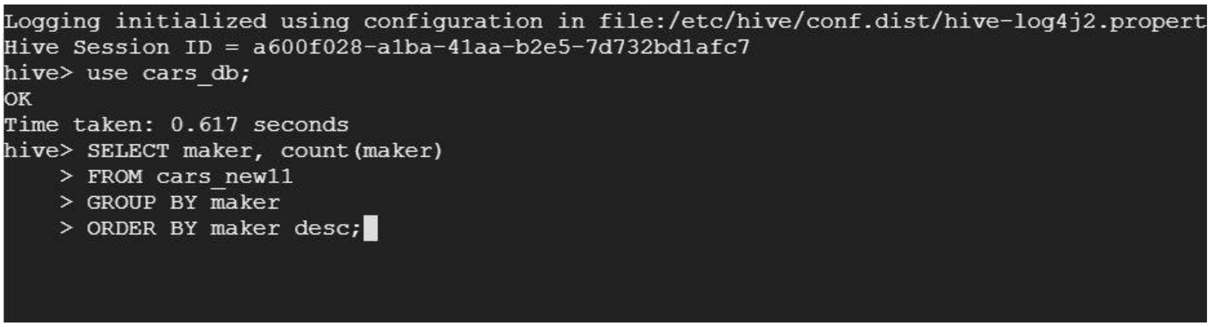
Cleaning data from Mileage, Manufacture year



```
Time taken: 0.071 seconds, Fetched: 48 row(s)
hive> CREATE TABLE cars_new3
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > TBLPROPERTIES("skip.header.line.count"="1")
    > AS SELECT maker,model,mileage,manufacture_year,engine_displacement,engine_power,stk_year,transmission,door_count,seat_count,
    > fuel_type,date_created,date_last_seen,price_eur FROM cars_new2
    > WHERE manufacture_year is NOT NULL and manufacture_year!='' and
    > mileage is NOT NULL and mileage!='' and mileage>100;
Query ID = christoamith_20211111002346_488a606c-621b-4c53-9bd5-490cfd7efeee
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_1636564781248_0013)

----------------------------------------------------------------------------------------------
        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: 4.91 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cars_db.db/cars_new3
OK
```

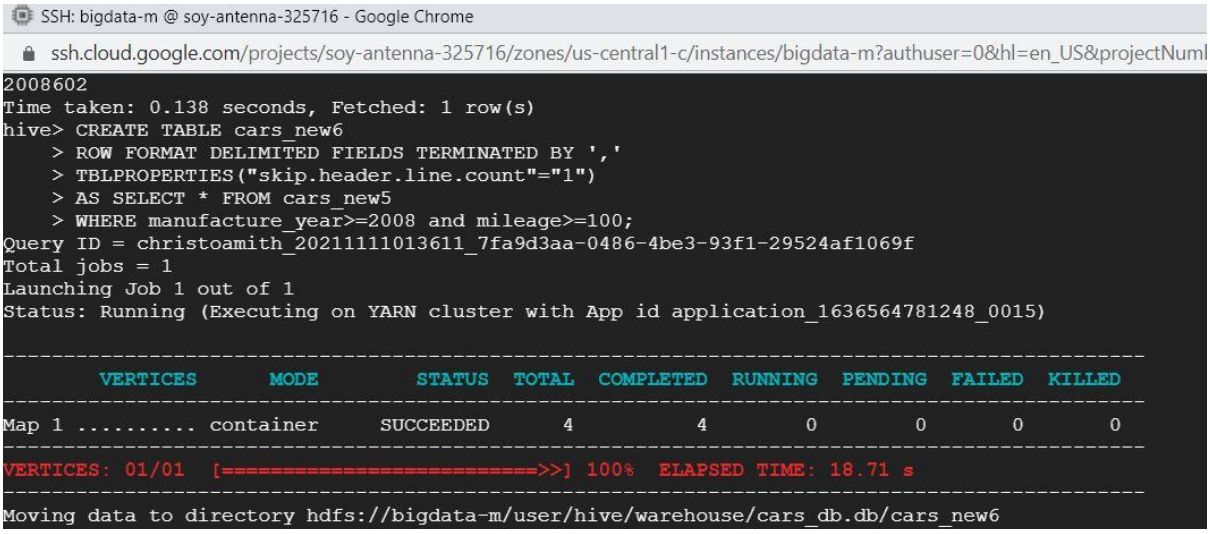
ANALYSIS:

Analysing the top 5 makers:



```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.propert
Hive Session ID = a600f028-a1ba-41aa-b2e5-7d732bd1afc7
hive> use cars_db;
OK
Time taken: 0.617 seconds
hive> SELECT maker, count(maker)
    > FROM cars_new11
    > GROUP BY maker
    > ORDER BY maker desc;
```

Filtering data less than 2008 and neglecting mileage with inappropriate values

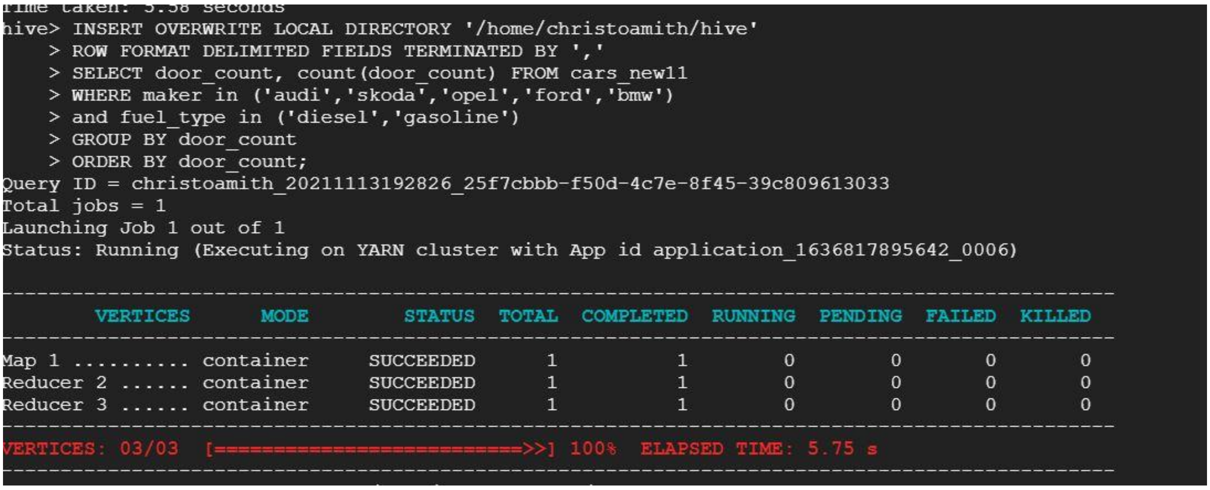


SSH: bigdata-m @ soy-antenna-325716 - Google Chrome

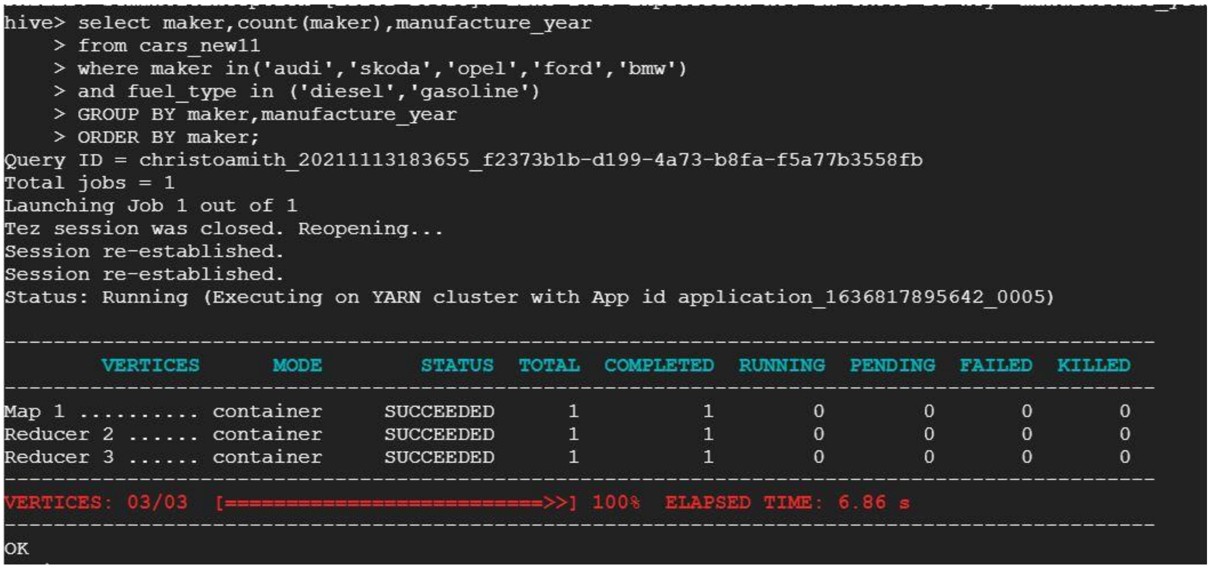
ssh.cloud.google.com/projects/soy-antenna-325716/zones/us-central1-c/instances/bigdata-m?authuser=0&hl=en_US&projectNuml

```
2008602
Time taken: 0.138 seconds, Fetched: 1 row(s)
hive> CREATE TABLE cars_new6
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > TBLPROPERTIES("skip.header.line.count"="1")
    > AS SELECT * FROM cars_new5
    > WHERE manufacture_year>=2008 and mileage>=100;
Query ID = christoamith_20211111013611_7fa9d3aa-0486-4be3-93f1-29524af1069f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1636564781248_0015)

----------------------------------------------------------------------------------------------
        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: 18.71 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cars_db.db/cars_new6
```

Selecting the Top 5 manufacturers and exporting them as a csv



```
Time taken: 5.58 seconds
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > SELECT door_count, count(door_count) FROM cars_new11
    > WHERE maker in ('audi','skoda','opel','ford','bmw')
    > and fuel_type in ('diesel','gasoline')
    > GROUP BY door_count
    > ORDER BY door_count;
Query ID = christoamith_20211113192826_25f7cbbb-f50d-4c7e-8f45-39c809613033
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1636817895642_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: 5.75 s
----------------------------------------------------------------------------------------------
```



```
hive> select maker,count(maker),manufacture_year
    > from cars_new11
    > where maker in('audi','skoda','opel','ford','bmw')
    > and fuel_type in ('diesel','gasoline')
    > GROUP BY maker,manufacture_year
    > ORDER BY maker;
Query ID = christoamith_20211113183655_f2373b1b-d199-4a73-b8fa-f5a77b3558fb
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_1636817895642_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.86 s
----------------------------------------------------------------------------------------------
OK
```

SQL queries used on HIVE for analysis:

----------**CREATE TABLE from RAW TABLE**----------

CREATE TABLE cars_new2

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT * FROM cars_new2

WHERE maker in ('audi','skoda','bmw','ford');

-----------**Verifying the result**-------------

SELECT * FROM cars_new2

```
ORDER BY price_eur desc

LIMIT 50;
```

-------------**Exporting the final result as a file**-----------------

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT fuel_type,count(fuel_type) FROM cars_new11

GROUP BY fuel_type;
```

--------------**Eliminating insufficient data columns**-------------------

```
CREATE TABLE cars_new3

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT maker,model,mileage,manufacture_year,engine_displacement,engine_power,stk_year,transmission,door_count,seat_count,

fuel_type,date_created,date_last_seen,price_eur FROM cars_new2;
```

---------------**Cleaning Mileage columns**------------

```
CREATE TABLE cars_new4

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT * FROM cars_new3

WHERE mileage!=" or mileage is NOT NULL;
```

---------------**Cleaning manufacturing year column**-----------

```
CREATE TABLE cars_new5

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT * FROM cars_new4

WHERE manufacture_year!=" or manufacture_year is NOT NULL;
```

----------------- **Filtering Mileage and manufacture year**-------------

```
CREATE TABLE cars_new6

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT * FROM cars_new5

WHERE manufacture_year>=2008 and mileage>=100;
```

------------------**Eliminating a column which is not used for analysis**-------------

```
CREATE TABLE cars_new7

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT maker,model,mileage,manufacture_year,engine_displacement,engine_power,transmission,door_count,seat_count,

fuel_type,date_created,date_last_seen,price_eur FROM cars_new6;
```

-------------------**Cleaning door count column**---------------

```
CREATE TABLE cars_new8

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT * FROM cars_new7

WHERE door_count is NOT NULL and door_count in(2,3,4,5,6);
```

---------------------**Cleaning seat count column**---------------

```
CREATE TABLE cars_new9

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT * FROM cars_new8

WHERE seat_count is NOT NULL and seat_count in(1,2,3,4,5,6,7,8,9);
```

--------------------- Cleaning fuel type---------------------

```
CREATE TABLE cars_new10

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT * FROM cars_new9

WHERE fuel_type!=";
```

-----------------------Cleaning transmission column-------------

```
CREATE TABLE cars_new11

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

TBLPROPERTIES("skip.header.line.count"="1")

AS SELECT * FROM cars_new10

WHERE transmission!=";
```

------------------------Exporting maker to detect top 5 manufacturers-------------

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT maker,count(maker) FROM cars_new11

GROUP BY maker

ORDER BY count(maker) desc;
```

-------------------Exporting data based on top 5 manufacturers, fuel type-------------

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT maker,count(maker),manufacture_year FROM cars_new11

WHERE maker in ('audi','skoda','opel','ford','bmw')

and fuel_type in ('diesel','gasoline')

GROUP BY maker,manufacture_year
```

```
ORDER BY maker;
```

--------------------Exporting data considering door count------------------

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT door_count, count(door_count) FROM cars_new11

WHERE maker in ('audi','skoda','opel','ford','bmw')

and fuel_type in ('diesel','gasoline')

GROUP BY door_count

ORDER BY door_count;
```

---------------------Exporting data considering seat count---------------------

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT seat_count, count(seat_count) FROM cars_new11

WHERE maker in ('audi','skoda','opel','ford','bmw')

and fuel_type in ('diesel','gasoline')

GROUP BY seat_count

ORDER BY seat_count;
```

--------------------- **Checking for average price and mileage**--------------

```
SELECT maker,model,manufacture_year,avg(price_eur),avg(mileage)

FROM cars_new11

WHERE maker in ('audi','skoda','opel','ford','bmw')

and fuel_type in ('diesel','gasoline') and door_count in (4,5) and seat_count in (4,5)

GROUP BY maker,model,manufacture_year
```

------------------------Exporting data based on average price and mileage------------

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT maker,model,manufacture_year,avg(price_eur),avg(mileage)
```

```
FROM cars_new11

WHERE maker in ('audi','skoda','opel','ford','bmw')

and fuel_type in ('diesel','gasoline') and door_count in (4,5) and seat_count in (4,5)

GROUP BY maker,model,manufacture_year;
```

-----------------Exporting data considering ad postings date------------------

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT maker,model,manufacture_year,avg(price_eur),avg(mileage)

FROM cars_new11

WHERE maker in ('audi','skoda','opel','ford','bmw')

and fuel_type in ('diesel','gasoline') and door_count in (4,5) and seat_count in (4,5) and DATEDIFF(date_last_seen,date_created)<=7

GROUP BY maker,model,manufacture_year;
```

------------------- Queries for Analysing transmission-------------------------

```
INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT transmission,count(transmission)

FROM cars_new11

GROUP BY transmission;


INSERT OVERWRITE LOCAL DIRECTORY '/home/christoamith/hive'

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

SELECT maker,model,count(model),date_last_seen,date_created

FROM cars_new11

WHERE maker in ('audi','skoda','opel','ford','bmw')

and fuel_type in ('diesel','gasoline') and door_count in (4,5) and seat_count in (4,5) and DATEDIFF(date_last_seen,date_created)<=7

GROUP BY maker,model,manufacture_year,date_last_seen,date_created;
```

```
SELECT maker,model,count(model),date_last_seen,date_created

FROM cars_new11

WHERE maker in ('audi','skoda','opel','ford','bmw')

and fuel_type in ('diesel','gasoline') and door_count in (4,5) and seat_count in (4,5) and DATEDIFF(date_last_seen,date_created)<=7

GROUP BY maker,model,manufacture_year,date_last_seen,date_created;
```