

CARS DATASET

DESCRIPTION:

The given dataset contains details of given car(model, price, cc, etc.) The cars dataset was collected in hope that it has certain influence on the factors determining price of a car. Can also be used to train models to accurately predict prices of models with only a few samples.

CONTENT:

- maker - normalized all lowercase
- model - normalized all lowercase
- mileage - in KM
- Manufacture year
- Engine displacement - in cc
- Engine power - in kW
- Body type - almost never present, but I scraped only personal cars, no motorcycles or utility vehicles

- Color slug - also almost never present
- stk year - year of the last emission control
- transmission - automatic or manual
- Door count
- Seat count
- Fuel type - gasoline, Diesel, electric
- Date created - when the ad was scraped
- Date *last* seen - when the ad was last seen. Our policy was to remove all ads older than 60 days
- Price Eur - list price converted to EUR

REFERENCE:

This dataset was taken from Kaggle :

<https://www.kaggle.com/mirosval/personal-cars-classifieds>

DATA CLEANING

1. MAKER:

Maker are independent and individual value. So, incase of blank values it is hard to replace.

Removed all the Blank cells using filter in Microsoft Excel.

2. MODEL:

Maker are independent and individual value. So, incase of blank values it is hard to replace.

Removed all the Blank cells using filter in Microsoft Excel.

3. MILEAGE:

We set the minimum value to be 10000 km covered per year and so car with mileage less are removed as erroneous value.

Removed all the Values (< 10000) using filter in Microsoft Excel because mileage should be at least 10000 kms per year.

4. MANUFACTURE YEAR:

This column contained values which are realistically impossible and so it was rectified.

Removed all the Values (< 1980) using filter in Microsoft Excel because car manufacture started at 1980s.

5. ENGINE DISPLACEMENT:

Removed all the Values (< 1000) and (> 10000) using filter in Microsoft Excel because Engine Displacement should be between 1000 and 10000cc.

6. ENGINE POWER:

Removed all the Values (< 30) using filter in Microsoft Excel because Engine Power should not be less than 30kw.

7. BODY TYPE:

This column was almost empty and so I filled it with two categories, namely, SUV and COMPACT with the help of Seat Count using Microsoft Excel.

Formula: =IF(J2>6,"SUV","COMPACT")

Also, renamed the column to Car Type.

8. COLOR SLUG:

Deleted this Column using Microsoft Excel because its of no use.

9. DOOR COUNT:

Deleted this Column using Microsoft Excel because its of no use.

10. STK CONTROL:

Renamed this into **Pollution Test**.

Categorized the value into **TRUE** OR **FALSE** by using IF and Look Up Table formula in Microsoft Excel.

Formula: =IF(all_anonymized_2015_11_2017_03!I2="None", FALSE, TRUE)

11. TRANSMISSION:

Renamed this into **Auto Facility**.

Categorized the value into **TRUE** OR **FALSE** by using IF and Look Up Table formula in Microsoft Excel.

Formula: =IF(all_anonymized_2015_11_2017_03!J2="auto", TRUE, FALSE)

12. SEAT COUNT:

Deleted this Column using Microsoft Excel because its of no use.

13. FUEL TYPE:

Didn't modify any Values in this column.

14. DATE CREATED:

Deleted this Column using Microsoft Excel because its of no use.

15. DATE LAST SEEN:

Deleted this Column using Microsoft Excel because its of no use.

16. PRICE EUR:

Renamed this column into **Price**.

This value had many up and down values and so, I took average of every individual car model and replaced by using IF and Look Like Table.

Formula: `=IF(O2="X",VLOOKUP(F2,Sheet5!A2:B27,2,TRUE),N2)`

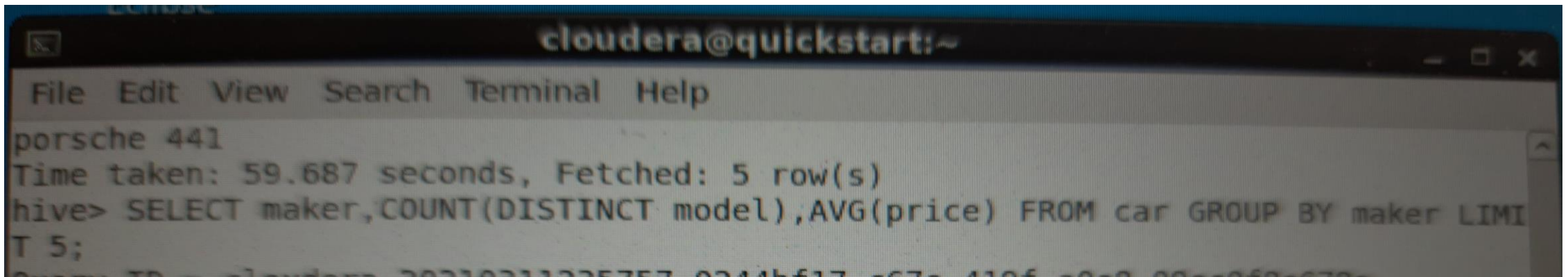
DATA ANALYSIS

QUESTIONS:

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

QUERY-1

```
SELECT maker, COUNT(DISTINCT model), AVG(price) FROM car GROUP BY maker  
LIMIT 5;
```

A screenshot of a terminal window titled "cloudera@quickstart:~". The window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The terminal output shows the query result for "porsche 441", the execution time "Time taken: 59.687 seconds, Fetched: 5 row(s)", and the Hive command "hive> SELECT maker,COUNT(DISTINCT model),AVG(price) FROM car GROUP BY maker LIMIT 5;".

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
porsche 441  
Time taken: 59.687 seconds, Fetched: 5 row(s)  
hive> SELECT maker,COUNT(DISTINCT model),AVG(price) FROM car GROUP BY maker LIMIT 5;
```


OUTPUT:

audi 34 23102.944724905043

bentley 8 74645.45802186376

bmw 16 26663.300917344222

chevrolet 33 16038.810314431224

chrysler 15 18126.342562759783

Total MapReduce CPU Time Spent: 6 seconds 540 msec

OK

audi 34 23102.944724905043

bentley 8 74645.45802186376

bmw 16 26663.300917344222

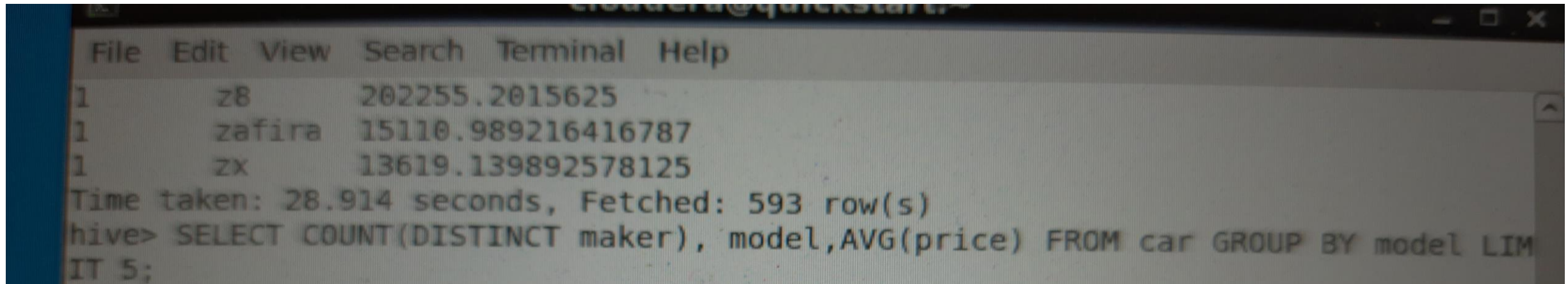
chevrolet 33 16038.810314431224

chrysler 15 18126.342562759783

Time taken: 36.239 seconds, Fetched: 5 row(s)

QUERY-2:

```
SELECT COUNT(DISTINCT maker), model, AVG(price) FROM car GROUP BY model  
LIMIT 5;
```



```
File Edit View Search Terminal Help  
1      z8      202255.2015625  
1      zafira  15110.989216416787  
1      zx      13619.139892578125  
Time taken: 28.914 seconds, Fetched: 593 row(s)  
hive> SELECT COUNT(DISTINCT maker), model,AVG(price) FROM car GROUP BY model LIM  
IT 5;
```

OUTPUT:

2	100	18467.826056419333
1	100-nx	12143.382269965277
1	105	14635.300903320312
1	115	17100.680114746094
1	116	19124.689127604168

Total MapReduce CPU Time Spent: 6 seconds 240 msec

OK

2	100	18467.826056419333
1	100-nx	12143.382269965277
1	105	14635.300903320312
1	115	17100.680114746094
1	116	19124.689127604168

Time taken: 28.474 seconds, Fetched: 5 row(s)

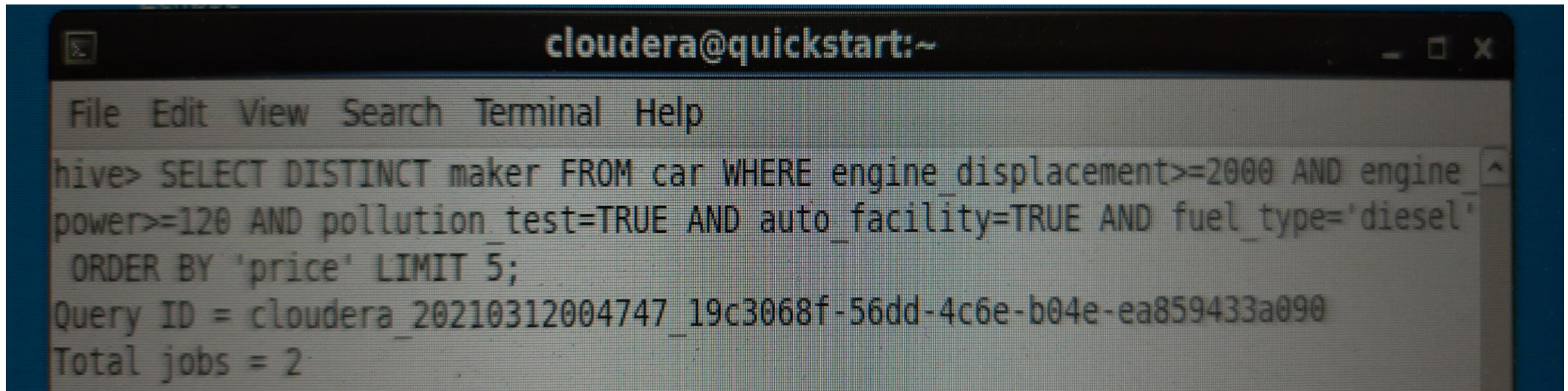
JUSTIFICATION FOR QUESTION- 1

ALL THREE COLUMNS,NAMELY, MAKER,MODEL AND PRICE ARE INTERDEPENDABLE BECAUSE PRICE IS DEPENDENT ON MODEL AND MODEL IS DEPENDENT ON MAKER.

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

QUERY:

```
SELECT DISTINCT maker FROM car WHERE engine_displacement >= 2000 AND  
engine_power >= 120 AND pollution_test=TRUE AND auto_facility = TRUE AND  
fuel_type = 'diesel' ORDER BY 'price' LIMIT 5;
```

A screenshot of a terminal window titled 'cloudera@quickstart:~'. The terminal shows a Hive query being executed. The query is: 'SELECT DISTINCT maker FROM car WHERE engine_displacement>=2000 AND engine_power>=120 AND pollution_test=TRUE AND auto_facility=TRUE AND fuel_type='diesel' ORDER BY 'price' LIMIT 5;'. Below the query, the terminal displays 'Query ID = cloudera_20210312004747_19c3068f-56dd-4c6e-b04e-ea859433a090' and 'Total jobs = 2'. The terminal has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
hive> SELECT DISTINCT maker FROM car WHERE engine_displacement>=2000 AND engine  
power>=120 AND pollution_test=TRUE AND auto_facility=TRUE AND fuel_type='diesel'  
ORDER BY 'price' LIMIT 5;  
Query ID = cloudera_20210312004747_19c3068f-56dd-4c6e-b04e-ea859433a090  
Total jobs = 2
```

OUTPUT:

volvo

toyota

skoda

porsche

opel

Total MapReduce CPU Time Spent: 10 seconds 400 msec

OK

volvo

toyota

skoda

porsche

opel

Time taken: 55.619 seconds, Fetched: 5 row(s)

JUSTIFICATION FOR QUESTION- 2

THE TOP 5 MANUFACTURERS ARE LISTED ABOVE . AND IT IS CALCULATED ON THE AVERAGE ENGINE DISPLACEMENT, AVERAGE ENGINE POWER, POLLUTION TEST CERTIFIED, AUTO FACILITY AND FUEL TYPE. FINALLY, PRICE IS CALCULATED ON THE CHEAPEST FIRST BASIS.

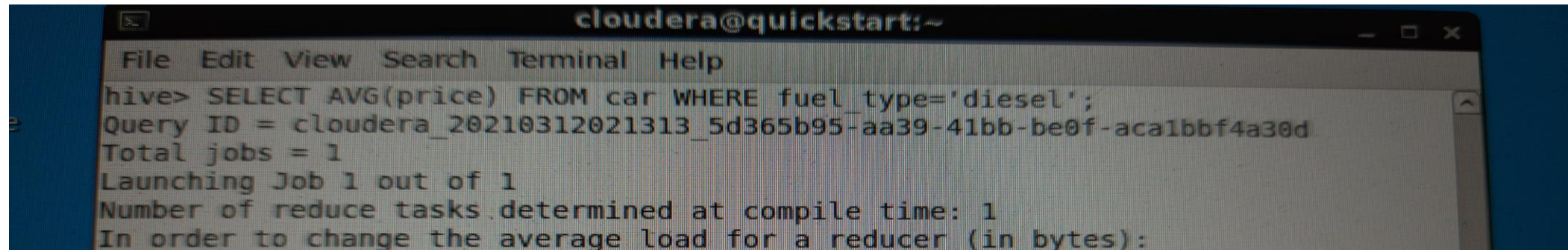
3.Does fuel type have any impact on the car price? Explain

QUERY - 1:

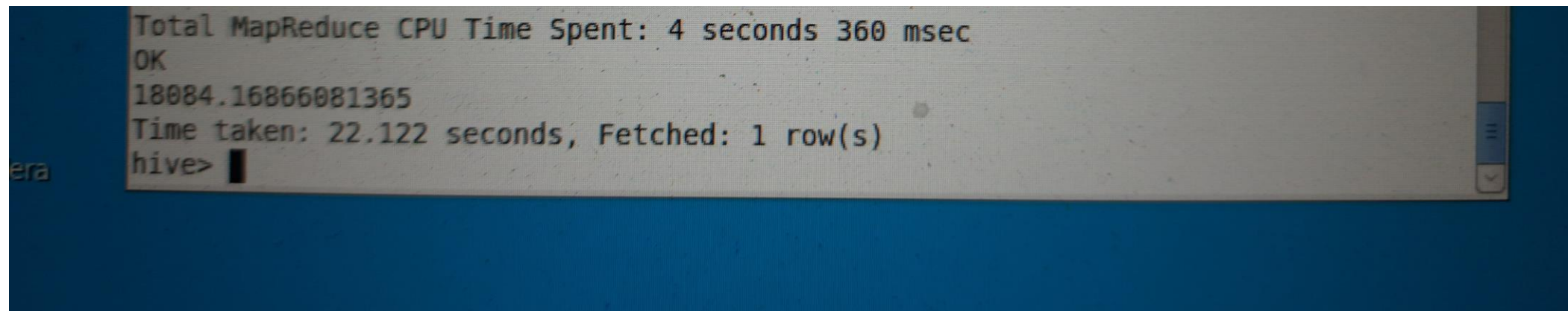
```
SELECT AVG(price) FROM car WHERE fuel_type='diesel';
```

OUTPUT:

18084.16866081365



```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
hive> SELECT AVG(price) FROM car WHERE fuel_type='diesel';  
Query ID = cloudera_20210312021313_5d365b95-aa39-41bb-be0f-aca1bbf4a30d  
Total jobs = 1  
Launching Job 1 out of 1  
Number of reduce tasks determined at compile time: 1  
In order to change the average load for a reducer (in bytes):
```



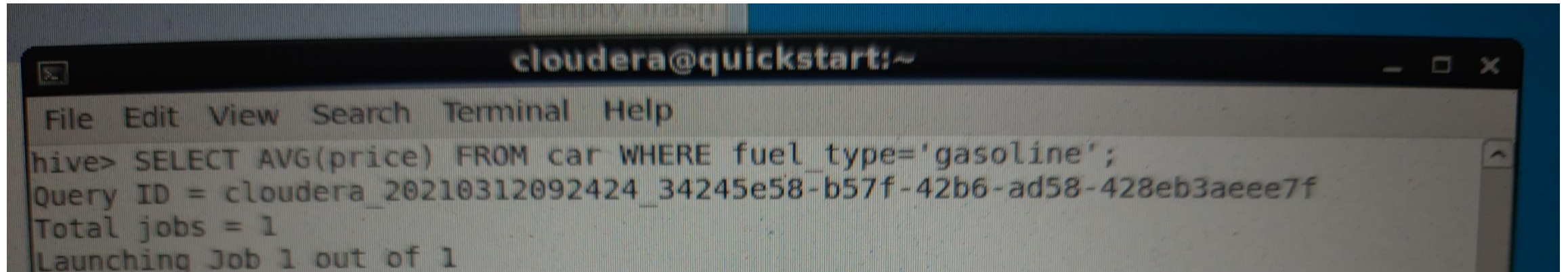
```
Total MapReduce CPU Time Spent: 4 seconds 360 msec  
OK  
18084.16866081365  
Time taken: 22.122 seconds, Fetched: 1 row(s)  
hive>
```


QUERY-2:

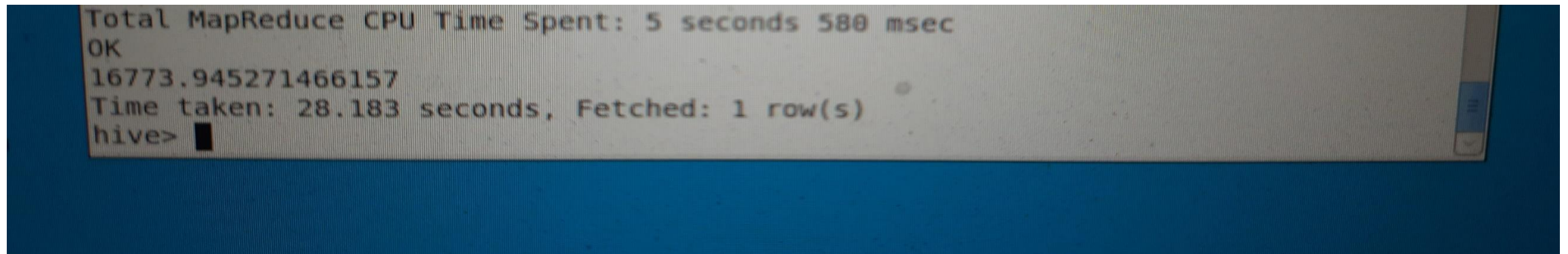
SELECT AVG(price) FROM car WHERE fuel_type='gasoline';

OUTPUT:

16773.945271466157

A screenshot of a terminal window titled "cloudera@quickstart:~". The terminal shows the execution of a Hive query: "hive> SELECT AVG(price) FROM car WHERE fuel_type='gasoline';". Below the query, it displays "Query ID = cloudera_20210312092424_34245e58-b57f-42b6-ad58-428eb3ae7f", "Total jobs = 1", and "Launching Job 1 out of 1".

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
hive> SELECT AVG(price) FROM car WHERE fuel_type='gasoline';  
Query ID = cloudera_20210312092424_34245e58-b57f-42b6-ad58-428eb3ae7f  
Total jobs = 1  
Launching Job 1 out of 1
```

A continuation of the terminal output from the previous screenshot. It shows the completion of the job: "Total MapReduce CPU Time Spent: 5 seconds 580 msec", "OK", the result value "16773.945271466157", "Time taken: 28.183 seconds, Fetched: 1 row(s)", and the prompt "hive>".

```
Total MapReduce CPU Time Spent: 5 seconds 580 msec  
OK  
16773.945271466157  
Time taken: 28.183 seconds, Fetched: 1 row(s)  
hive>
```

JUSTIFICATION FOR QUESTION- 3

THE AVERAGE PRICE OF DIESEL ENGINE IS 18084.16866081365. AND THE AVERAGE PRICE OF GASOLINE ENGINE IS 16773.945271466157. THE DIFFERENCE IS ABOUT 1311 AND SO THE FUEL TYPE HAVE IMPACT ON THE CAR PRICE.

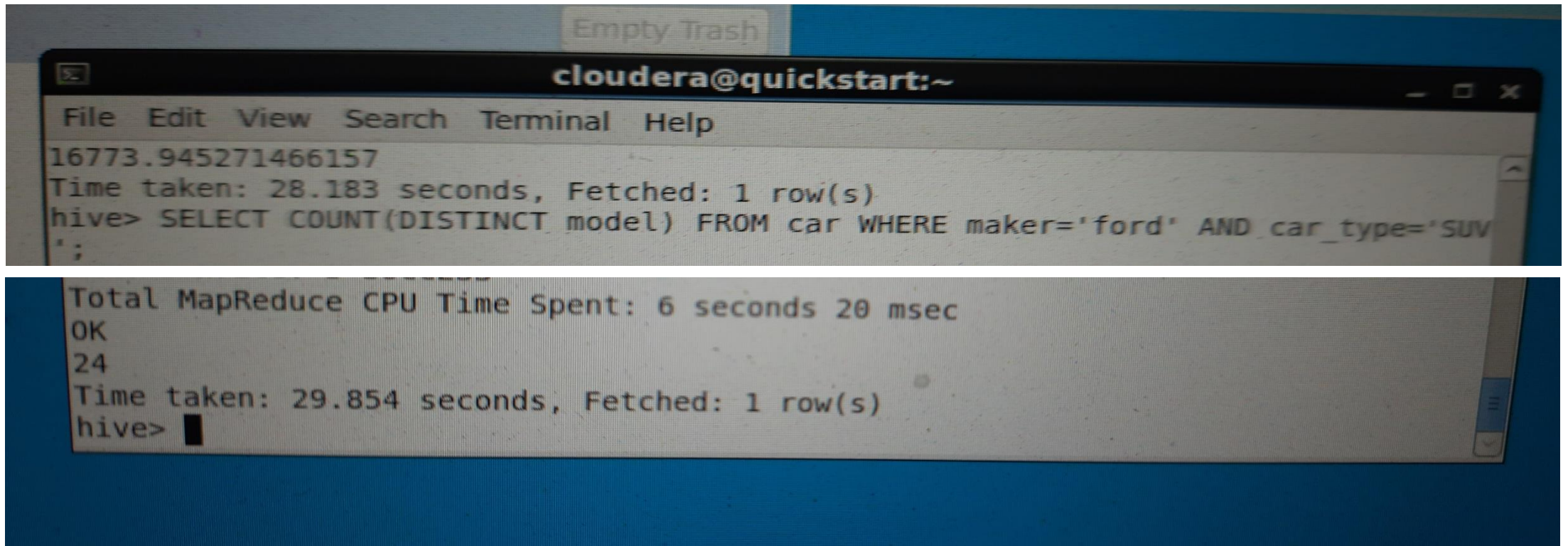
4.How many SUV models did FORD maker launched till now ?

QUERY:

```
SELECT COUNT(DISTINCT model) FROM car WHERE maker='ford' AND car_type = 'SUV';
```

OUTPUT:

24

A screenshot of a terminal window titled 'cloudera@quickstart:~'. The terminal shows the execution of a Hive query. The query is 'SELECT COUNT(DISTINCT model) FROM car WHERE maker='ford' AND car_type='SUV';'. The output shows '16773.945271466157' and 'Time taken: 28.183 seconds, Fetched: 1 row(s)'. Below this, the query is repeated, and the output shows 'Total MapReduce CPU Time Spent: 6 seconds 20 msec', 'OK', and '24'. The terminal also shows 'Time taken: 29.854 seconds, Fetched: 1 row(s)' and 'hive>' at the end.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
16773.945271466157  
Time taken: 28.183 seconds, Fetched: 1 row(s)  
hive> SELECT COUNT(DISTINCT model) FROM car WHERE maker='ford' AND car_type='SUV'  
;  
  
Total MapReduce CPU Time Spent: 6 seconds 20 msec  
OK  
24  
Time taken: 29.854 seconds, Fetched: 1 row(s)  
hive>
```

JUSTIFICATION FOR QUESTION- 4

THE FORD MAKER LAUNCHED 24 SUV CARS TILL NOW.

5. Which 5 car makers have top engine power and passed pollution test ?

QUERY:

```
SELECT DISTINCT maker,engine_power FROM car WHERE pollution_test=TRUE ORDER  
BY engine_power DESC LIMIT 5;
```

OUTPUT:

hyundai	812
audi	516
bentley	460
toyata	449
porsche	441

cloudera@quickstart:~

File Edit View Search Terminal Help

Time taken: 29.854 seconds, Fetched: 1 row(s)

```
hive> SELECT DISTINCT maker,engine_power FROM car WHERE pollution_test=TRUE ORDER BY engine_power DESC LIMIT 5;
```

Query ID = cloudera_20210312094747_47b09093-2c91-4d53-8152-dd0a9abc18bb

Total jobs = 2

Total MapReduce CPU Time Spent: 11 seconds 560 msec

OK

hyundai 812

audi 516

bentley 460

bentley 449

porsche 441

Time taken: 57.185 seconds, Fetched: 5 row(s)

JUSTIFICATION FOR QUESTION- 5

THE BEST 5 CAR MAKERS WHO HAVE TOP ENGINE POWER AND PASSED POLLUTION TEST ARE, NAMELY, HYUNDAI, AUDI, BENTLEY, PORSCHE.

REPORT

From this report, I conclude, the top 5 CAR MANUFACTURERS are

- 1.Volvo
- 2.Toyota
- 3.Skoda
- 4.Porsche
- 5.Opel

Above Car Manufacturers are arranged on the below given Rank Basis,

- 1.Lowest Price
- 2.Engine Displacement greater than or equal to 2000 cc.
- 3.Engine Power greater than or equal to 120 kw.
- 4.Should have passed Pollution emission control test.
- 5.Should have Automatic Gear facility.
6. Should run in Diesel type engine.

DATA VISUALIZATION

