Data analysis tools Analytics (DATA1202)

**Assignment-2 Analysis with hive**

**student number: 100818850**

**Student name: Jayarani Rajesh**

**Professor: Saber amini**

2022

**Introduction**

Since the first production of cars, they are bought and used without slowing down.

In addition, with the development of technology, cars gain new features. As people buy new cars, old cars are transferred to another market. Maybe one buys his/her first car from the used car market. In addition to this, while the internet surrounded our lives with developing technology, people started to buy their cars from the internet. Thus, such as the data used here, it provided an opportunity to conduct research that could analyze people's needs and understand the reasons for their preferences. With this report, some reasons for preference will be tried to be understood.

This report addresses the branch manager’s concerns with an investment in the used cars business. It has been suggested by the manager that we should use the cars.csv dataset to complete the analysis report. This report will clearly explain the used cars dataset and will make recommendations on which car manufacturer to invest in our company’s new business. Analysis is done with **hive** using **Google Cloud Platform**.

**About - Apache Hive**

Apache Hive - an open-source framework, is a data warehouse software project built on top of Apache Hadoop for providing data queries, analysis and it is also used to efficiently store, process large datasets (petabytes of data using SQL). Hive differs from a relational database in a way, that it stores the schema in a database and processed data into HDFS.

**Dataset Background**

The Classified ads for cars dataset from Kaggle has data about wide range of used cars from the countries of Germany and Czech Republic since 2015. The whole data is all about Car manufactures, Pricing, Engine capacity, Total door and seat counts, Fuel category and some dates about the car advertisement. The purpose of this assignment is Data Cleaning and Analyzing the dataset to provide valuable information to Stakeholders in order to make business decisions.

There are so many benefits in purchasing a used car. Some of them are, used cars prices and insurance rates are almost 45% lower than new cars with wider variety to choose from our own budget. Also buying a used car reduces the carbon dioxide output into the environment.

Data present in the Dataset ‘all\_anonymized\_2015\_11\_2017\_03.csv’



The dataset contains about ~3.5 million rows and 16 columns. As per author, the data is taken about one year ago and also sourced from unstructured sources. Therefore, It is anticipated that there would be fields having missing values or invalid values. Therefore, Data cleaning must be performed on the data before considering it in any kind of decision process.

**Methodology**

**Data Cleaning**

The raw data downloaded from the Kaggle website consists of 16 lines and 3.5

million rows. Hive will be used to clear and analyze the data that has thousands of

empty and wrong lines.

- In the initial state of the data, there are 3552912 rows.

- All rows with null values in maker or model are removed. As a result of this

operation, there are 1322853 rows remained in clean\_used\_cars table.

Removal of Null values and Variables:

* Variables with 50% of missing values are removed. The removed variables are, color\_slug and fuel\_type.
* Null values in the fields maker and model are removed as major analysis depends on these two variables.
* Cars with the price value of ‘1295.34’ is removed as it is too frequent.

Missing Value Analysis:



Changing data types of the variables:

The datatypes of the following variables has been changed from String to its respective datatypes as per the data present in each variable.



**Analysis**

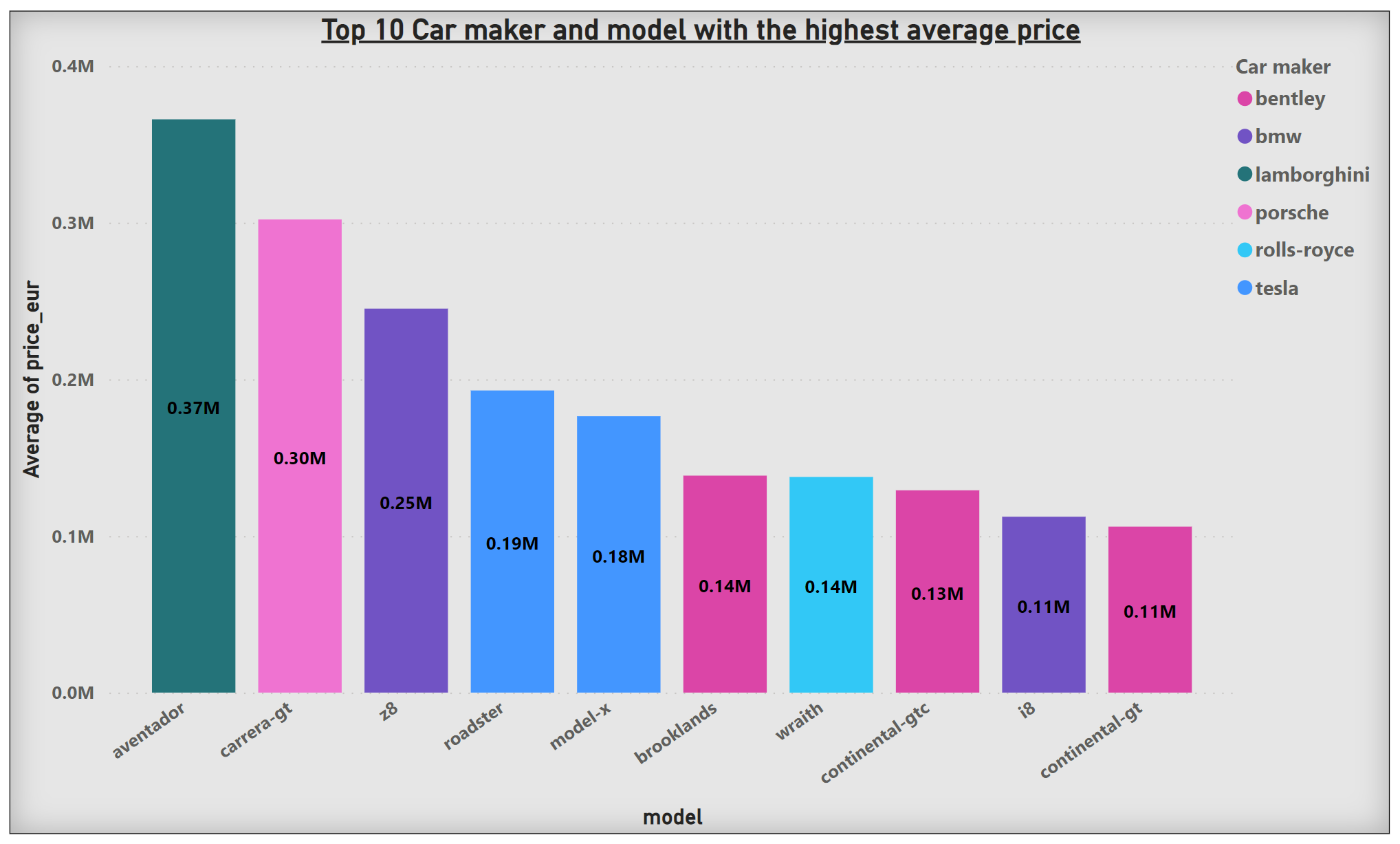
The price range of cars are set to provide options for the customers and investors based on three main segments which are economic, intermediate, and luxury. The economic segments has used cars with a price of 3000 – 20000, intermediate segment with a price range between 20,000 - 300,000, and luxury segment has a price range more 300,000 - 2000,000.

Top 10 and least 10 average car prices are found with respect to the maker and model of the cars.

The manufacture year range of the vehicles was determined as 2000-2017 to eliminate very old cars.

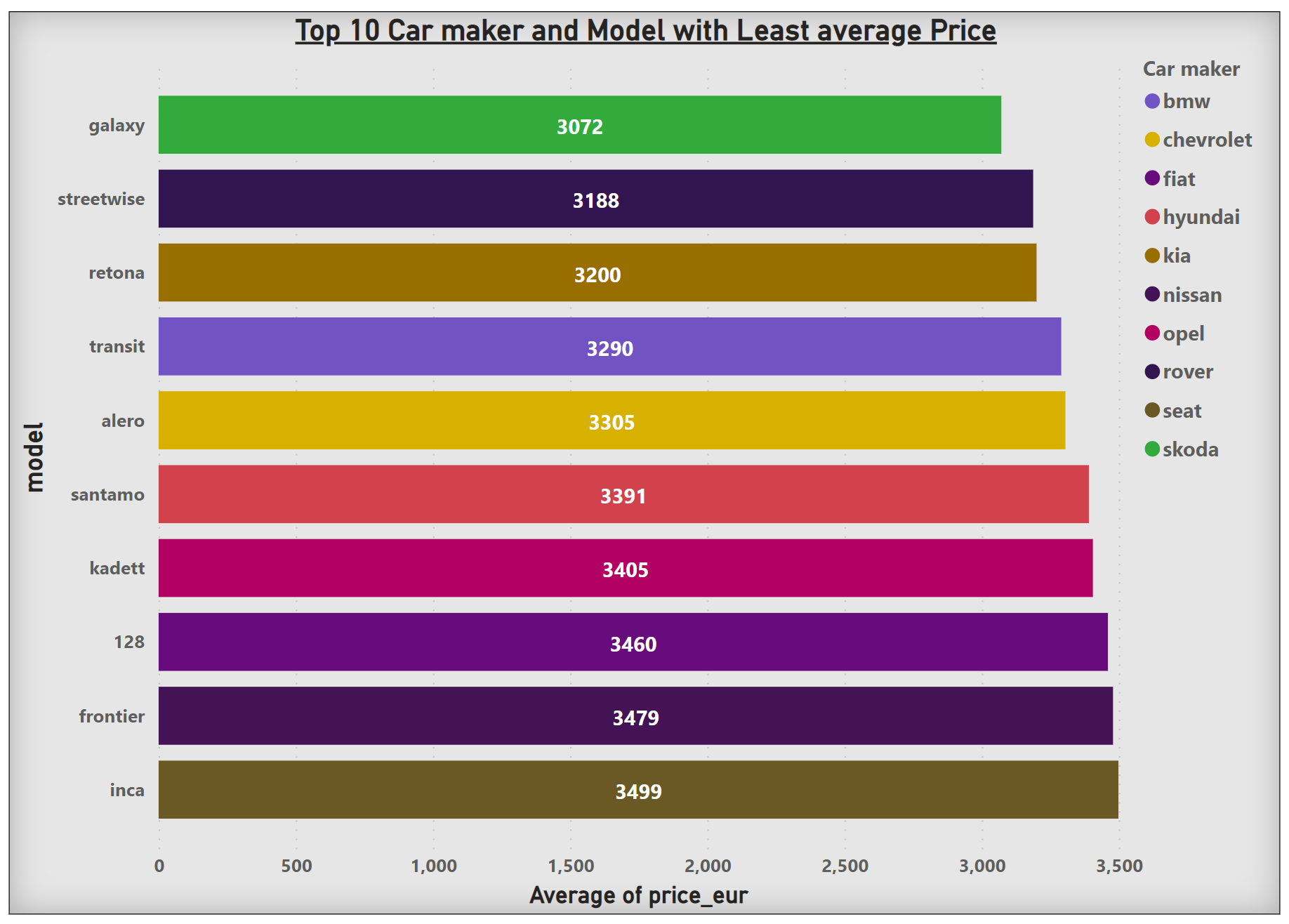
To best present the analysis mentioned. A series of visualizations are created using **PowerBI.**

**Top 10 highest average priced makers and models:**



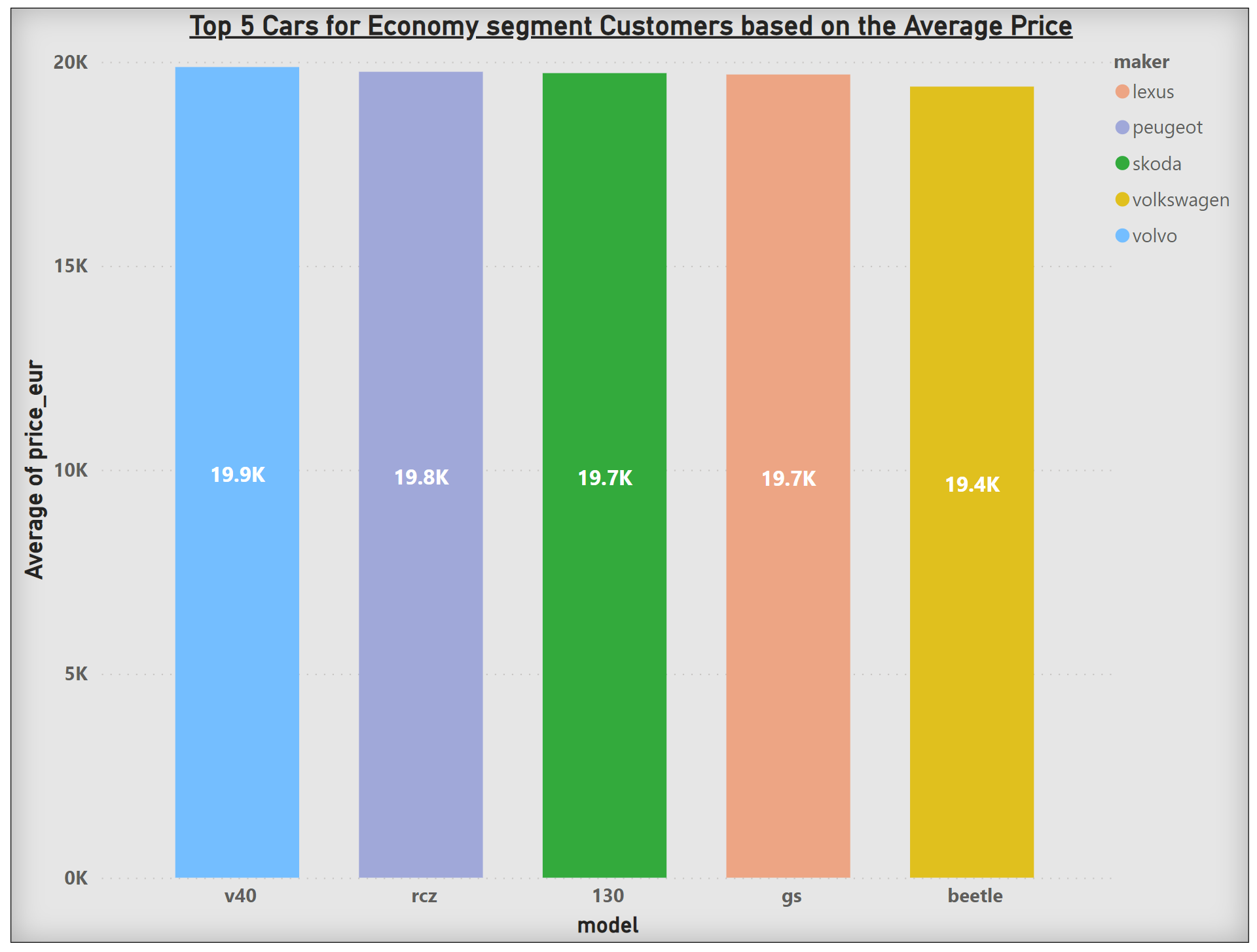
The bar graph depicts top 10 high priced cars by its makers and model names. The first car and its model name is **Lamborghini aventador** with an average price of **365960€. porsche carrera-gt** has the second highest average price of **302045€.** Next on the list is **bmw z8** with an average price around **245118€. Tesla roadster** is at four with an average price of **192880€. Tesla model-x** comes with an average price of **176418€. Bentley brooklands** has an average price of **138501€. Rolls-royce wraith** has an average price of **137663€. Bentley continental-gtc** comes at top 7 with an average price of **129138€. Bmw i8** has an average price of  **112273€** and last in the list is **bentley continental-gt** with an average price **105946€.**

**Top 10 makers and models with least average price:**



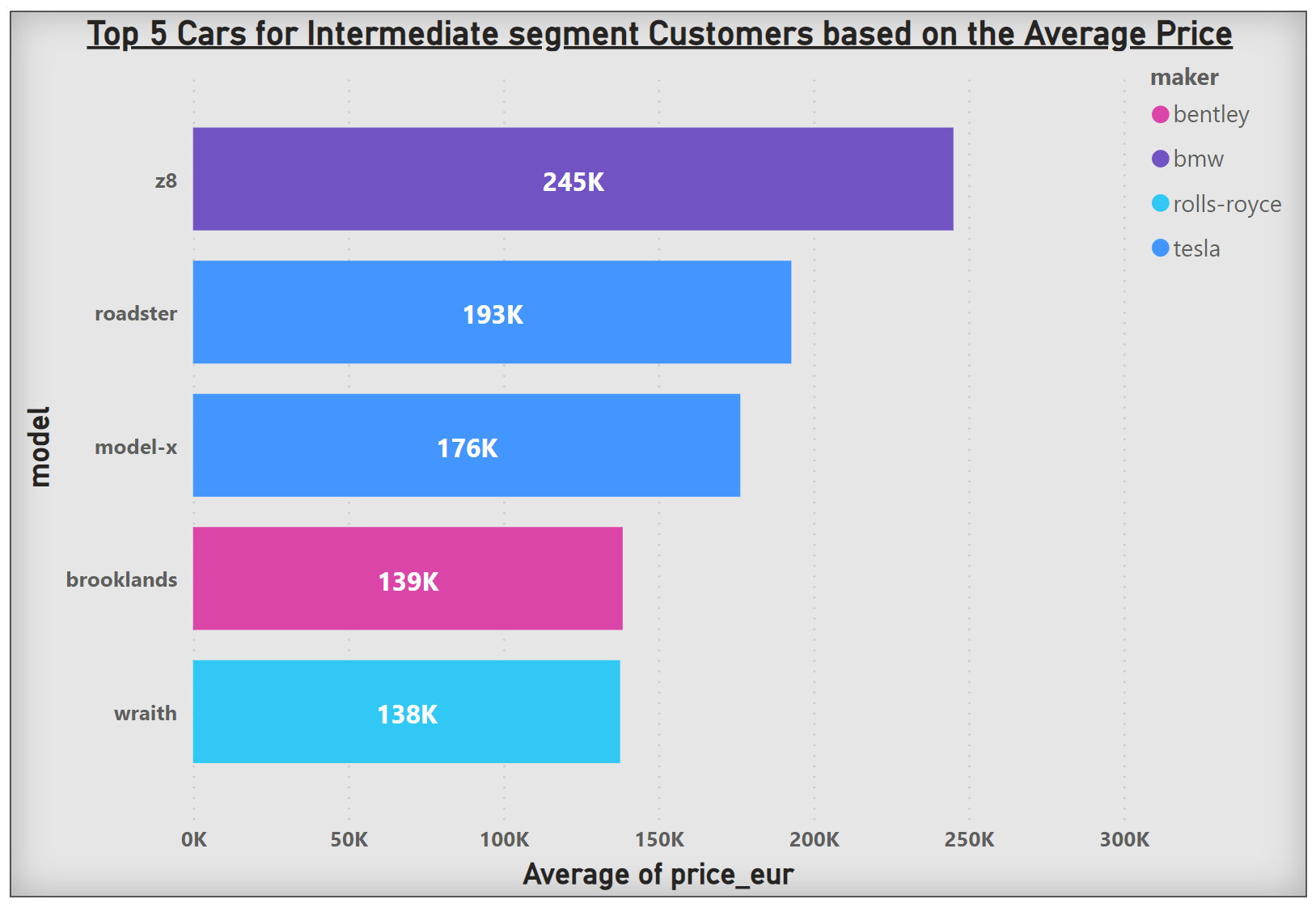
The horizontal bar graph lists top 10 least makers and models with respect to average price. The listed brands and models have almost equal average prices. **skoda galaxy** comes first in the list with a least average price of **3072€.** The second comes **rover streetwise** with an average price of **3188€.** And **kia retona** has an average price of **3200€. bmw transit** has an average price of **3290€. chevrolet alero** comes in place five with an average price of **3305€. hyundai santamo** with an average price of **3391€. opel kadett** is at place seven with an average price around **3405€. fiat 128** has an average price of **3460€. Nissan frontier** has an average price **3479€.** **Seat inca** comes last in the list of least average priced cars **3498€.**

**Top five make and model for Economic segment customers price range- 3000 ≤ price < 20,000**



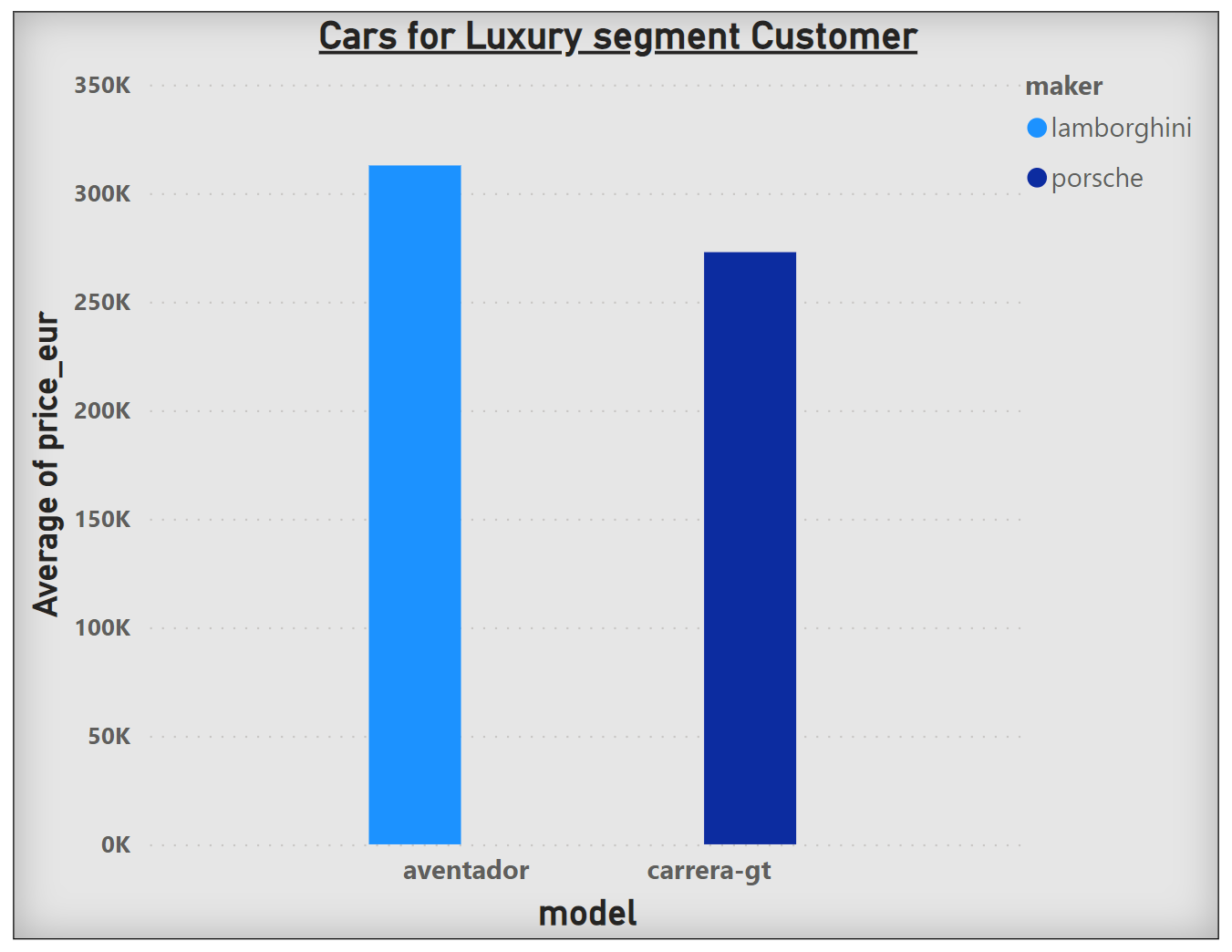
This bar graph shows top 5 manufacturers with their car models for average price between 3000 and 20000. The first car and its model is **volvo v40** is around **19.9k.** The second car manufacturer and model they make which is economic is **Peugeot rcz** is around **19.8k.** The maker and model in third place is **Skoda 130** is around **19.7k. Lexus gs** has an average price of **19.7k.** The car with least average price in this economic segment is **Volkwagen beatle** with an average price of **19.4k.**

**Top five manufacturers in the 20,000 to 300,000 price range; 3000≤price<20,000) - based on the top average price**



This horizontal bar graph shows intermediate segment prices for manufacturers and their model. The manufacturer with lowest intermediate price is **Rolls Royce** the model name is wraith with an average price of **138k.** The second lowest in the intermediate segment is **Bentley brooklands** with an average price of **139k.** The third manufacturer is **Tesla model-x** with an average price of **176k** and also **Tesla roadster** comes next with an average price of **193k.** The last in the list is **BMW z8** with an average price around **245k Euros.**

**The make and model for the Luxury segment customers price range- 300,000 ≤ price < 2000,000**



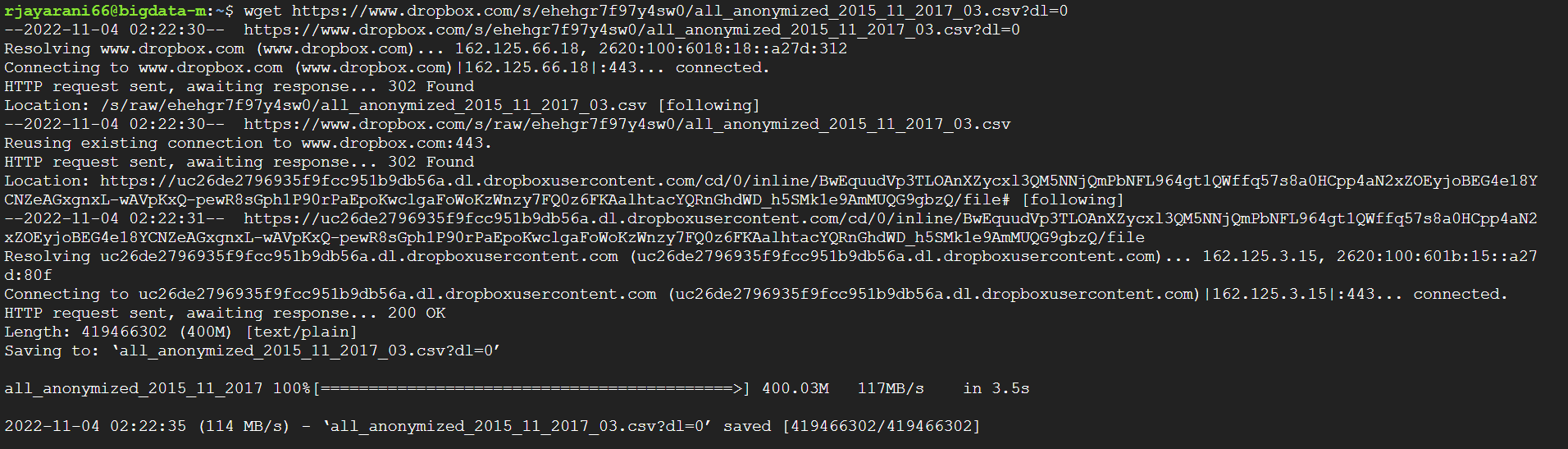
There are two luxury manufacturers for luxury customers the average price range is set between 300000 – 2000000. The first luxury car and its model name is **Lamborghini aventador** with an average price of **365960€.** The next one on the list is **Porsche Carrera-gt** with an average price of **302405€.**

**Conclusion and recommendation**

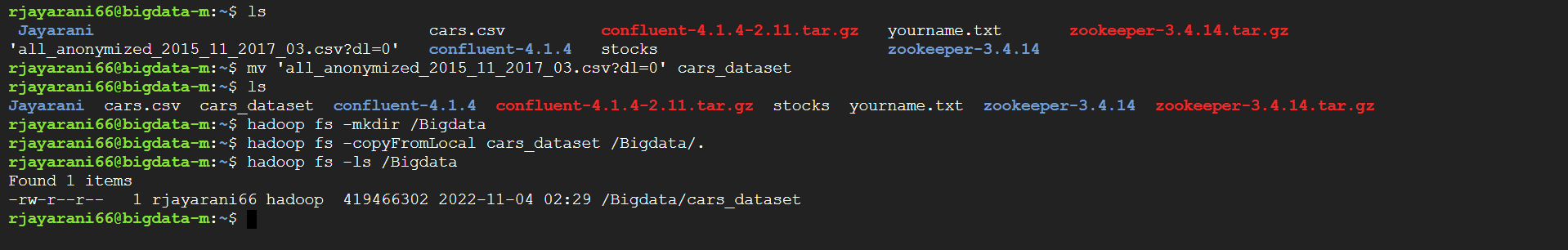
Analyses have been performed on Hive. Based on the analysis, a decision was made on what car brand and model to invest in. Based on their economic status, customers can make their own choices based on this analysis. As a result of this analysis, various options have been presented for each segment of people to help them decide which brand and model to buy. Based on the analysis results, it is possible to make an investment based on segments.

**Appendix**

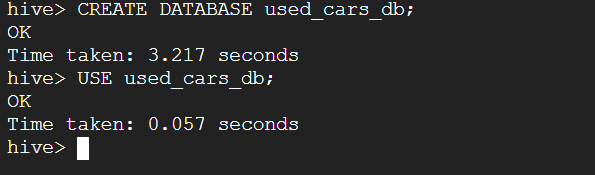
The dataset is downloaded



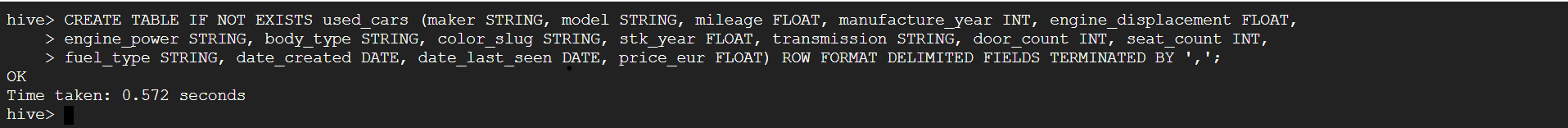
The dataset has been moved and renamed



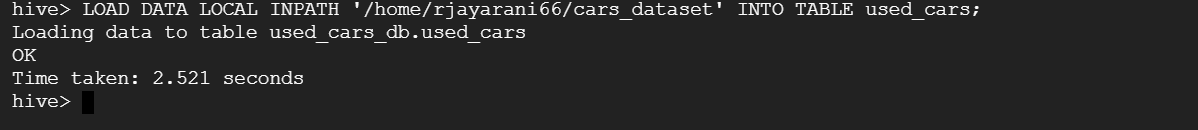
Started hive and created a database named it used\_cars\_db



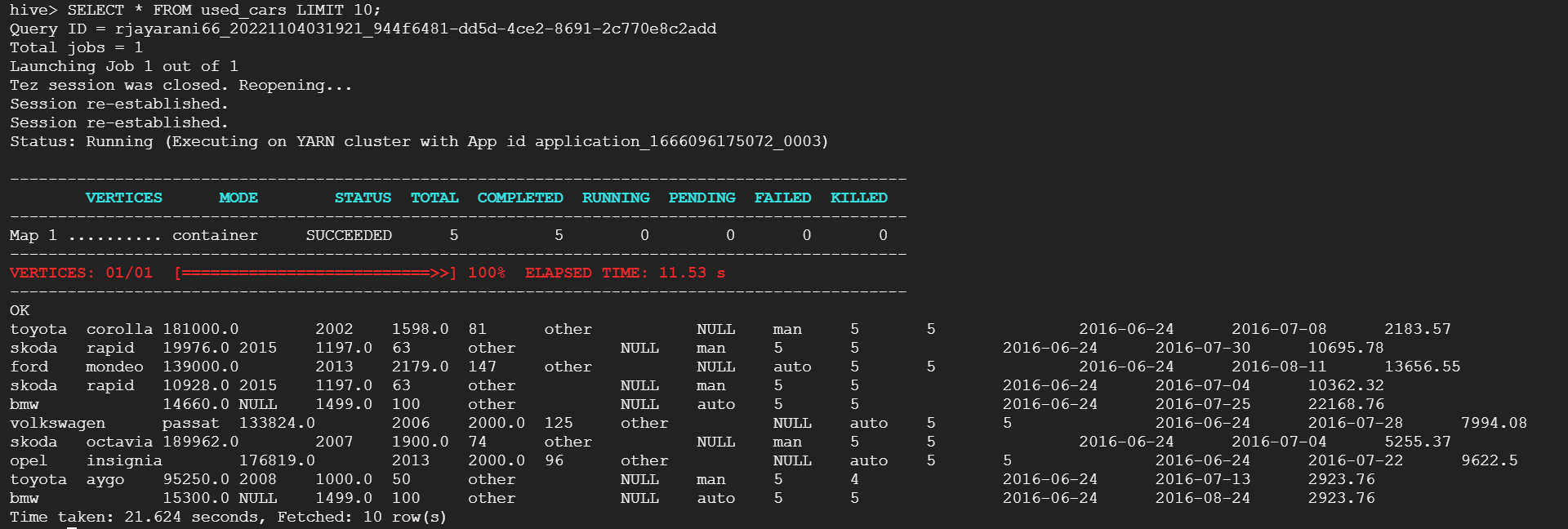
Created a schema for the table and named the table used\_cars



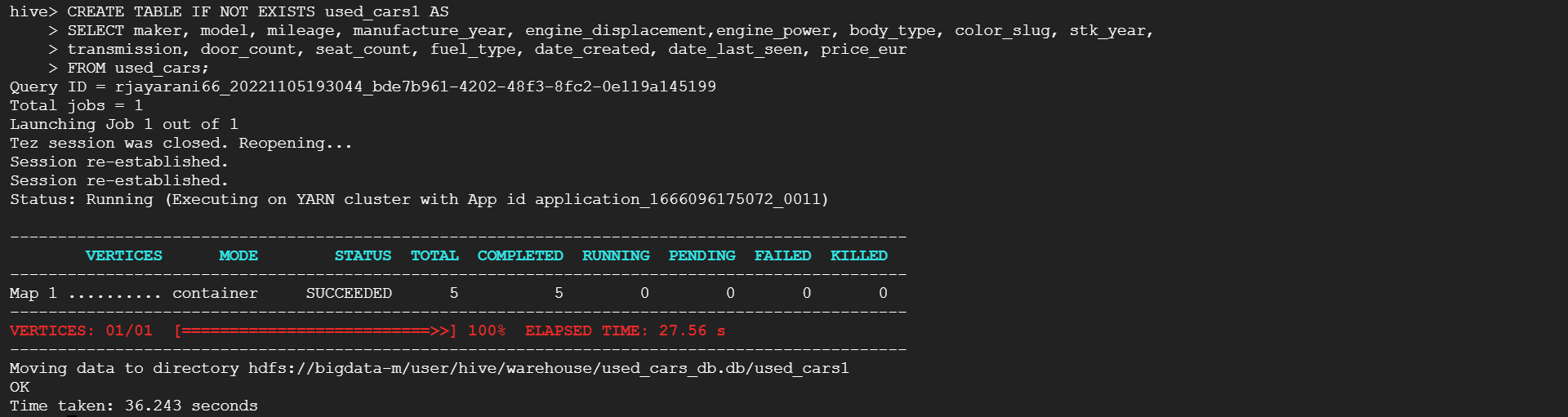
The dataset cars is loaded to the table skeleton created on the previous step.

****

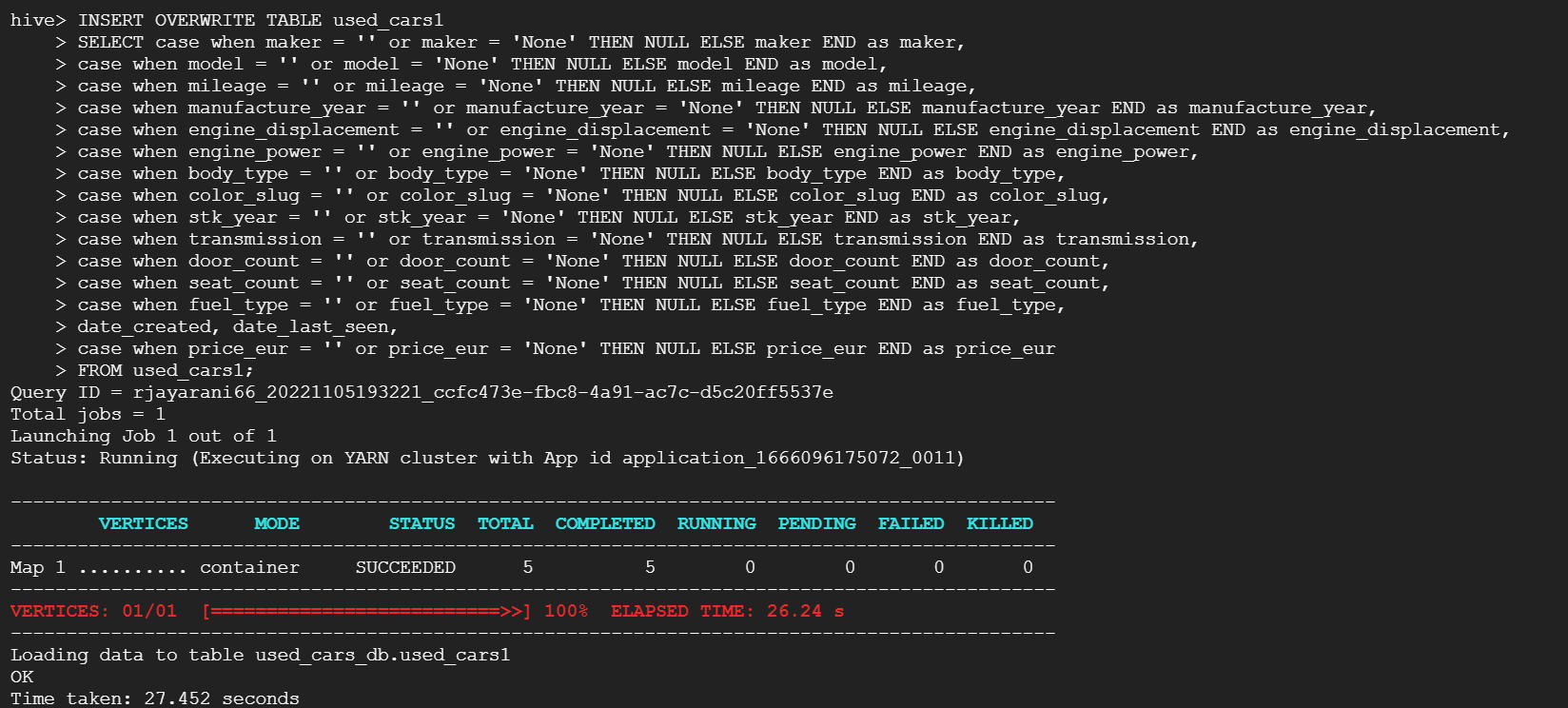
Checking if the dataset is loaded properly.

****

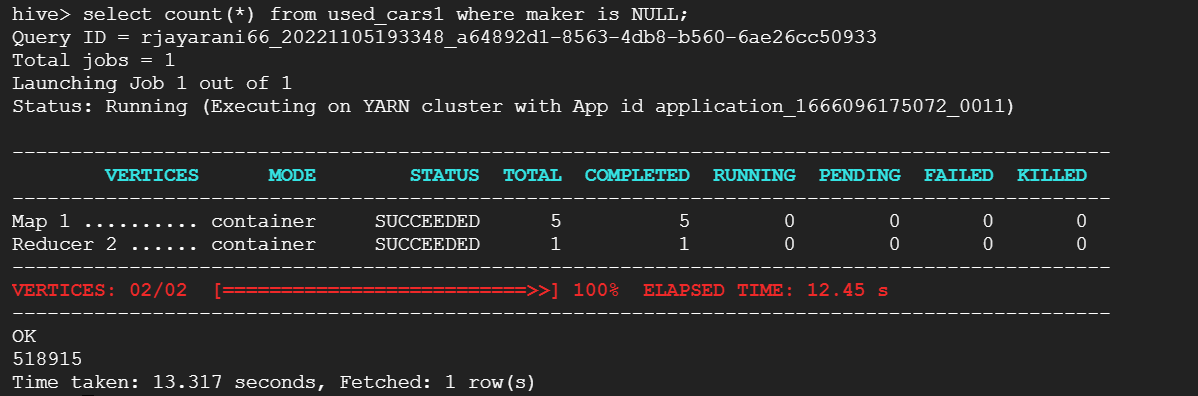
Creating a different table to mark missing values values to NULL.

****

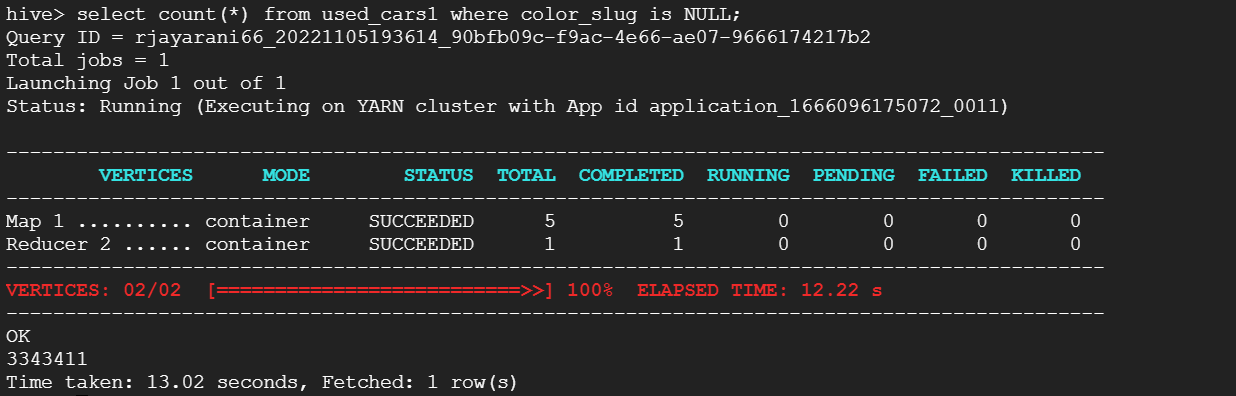
Created the overwrited table for further analysis

****

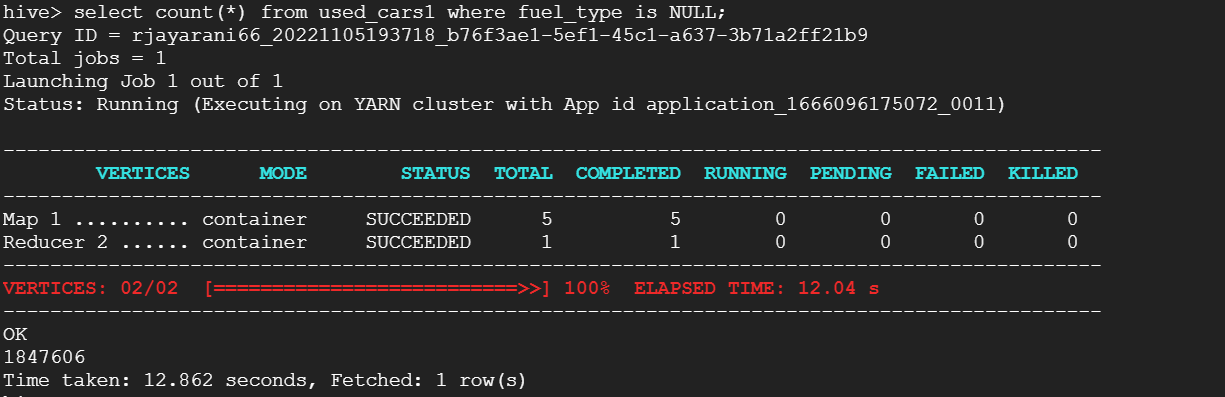
Counting null values from the fields of clean\_used\_cars.

****

Counting color\_slug null values.

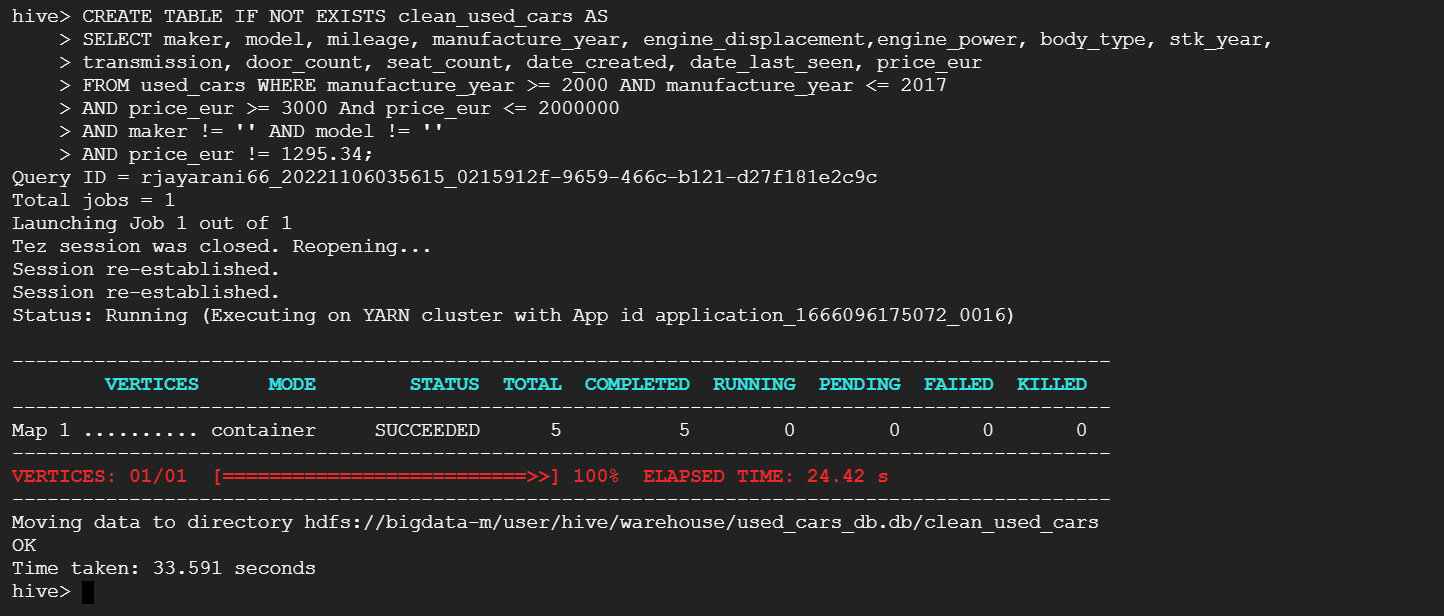
****

Counting fuel-type null values.

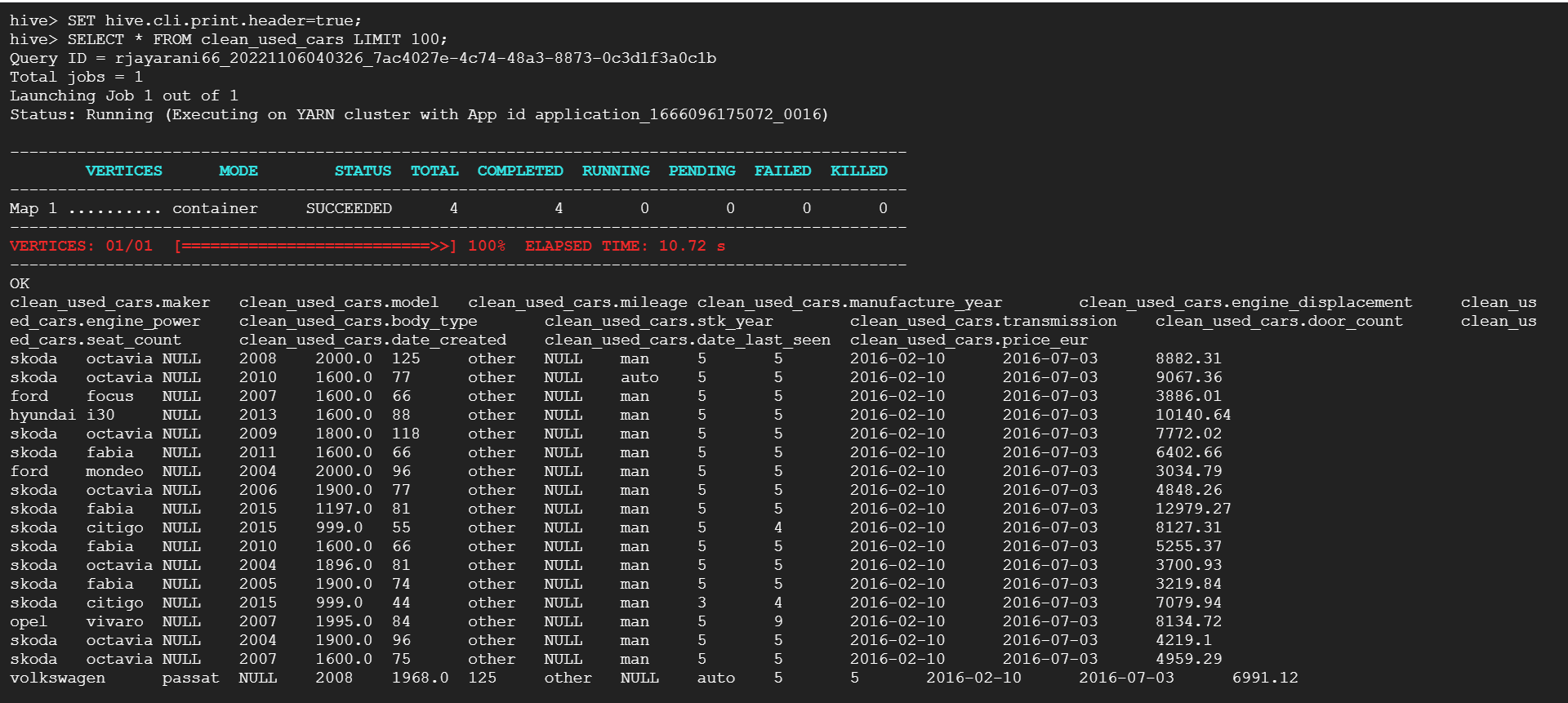
****

Creating a new table naming it clean\_used\_cars

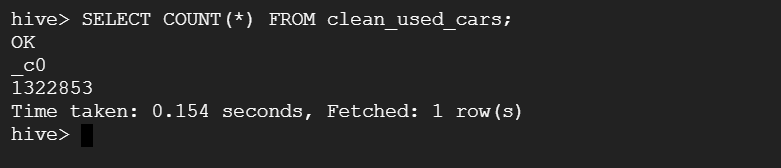
* Drop the columns with more than 50% missing values
* The manufacture year between 2000 and 2017 including 2000 and 2017
* Both maker and model exist in the row
* The price range is from 3000 to 2000,000 (3000 ≤ price ≤ 2000,000)
* Remove any price you singled out in Step 3 (i.e., a price that repeats too frequently for a random set of ads).

****

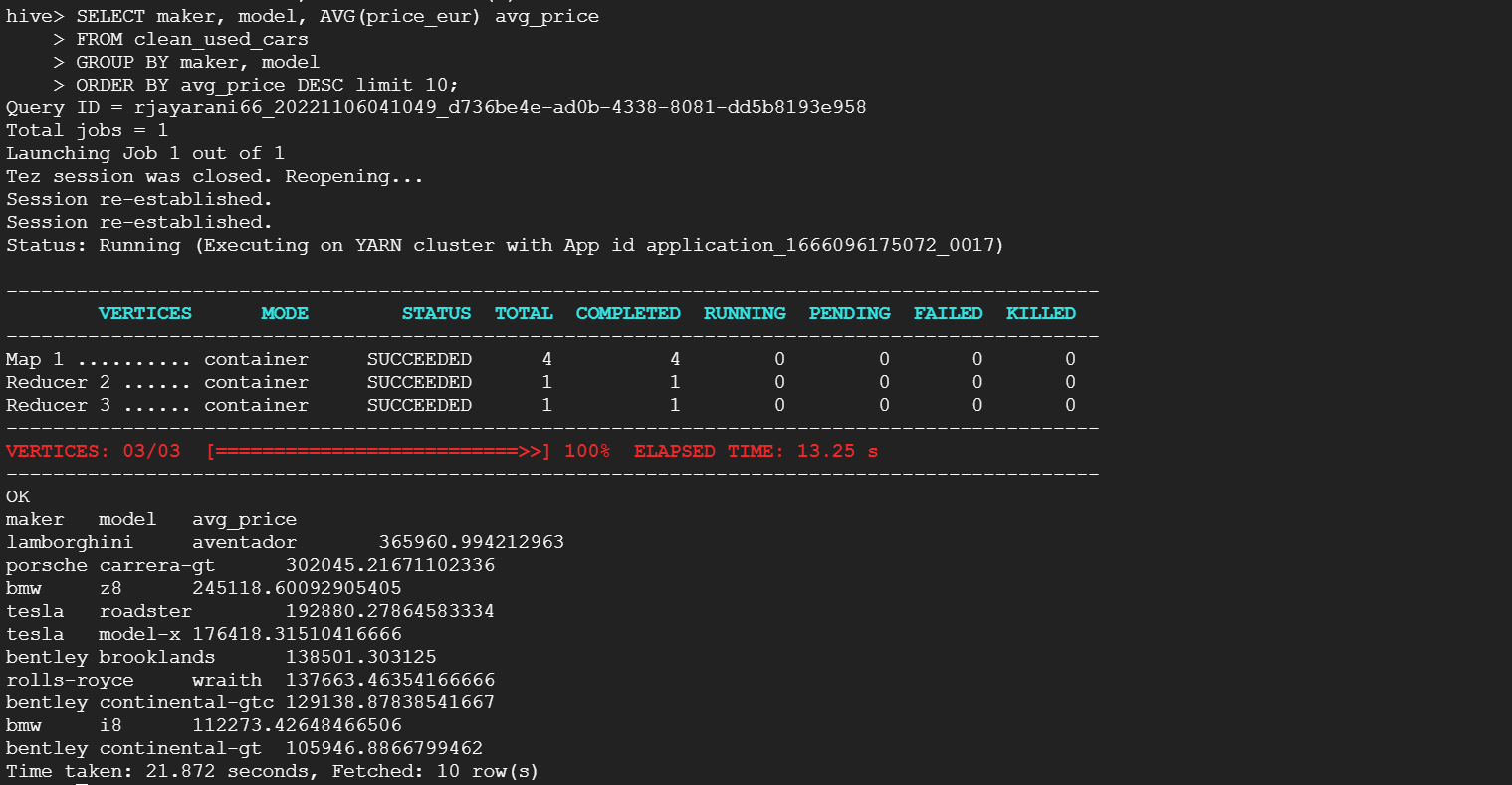
Checking if the table is loaded properly with header.

****

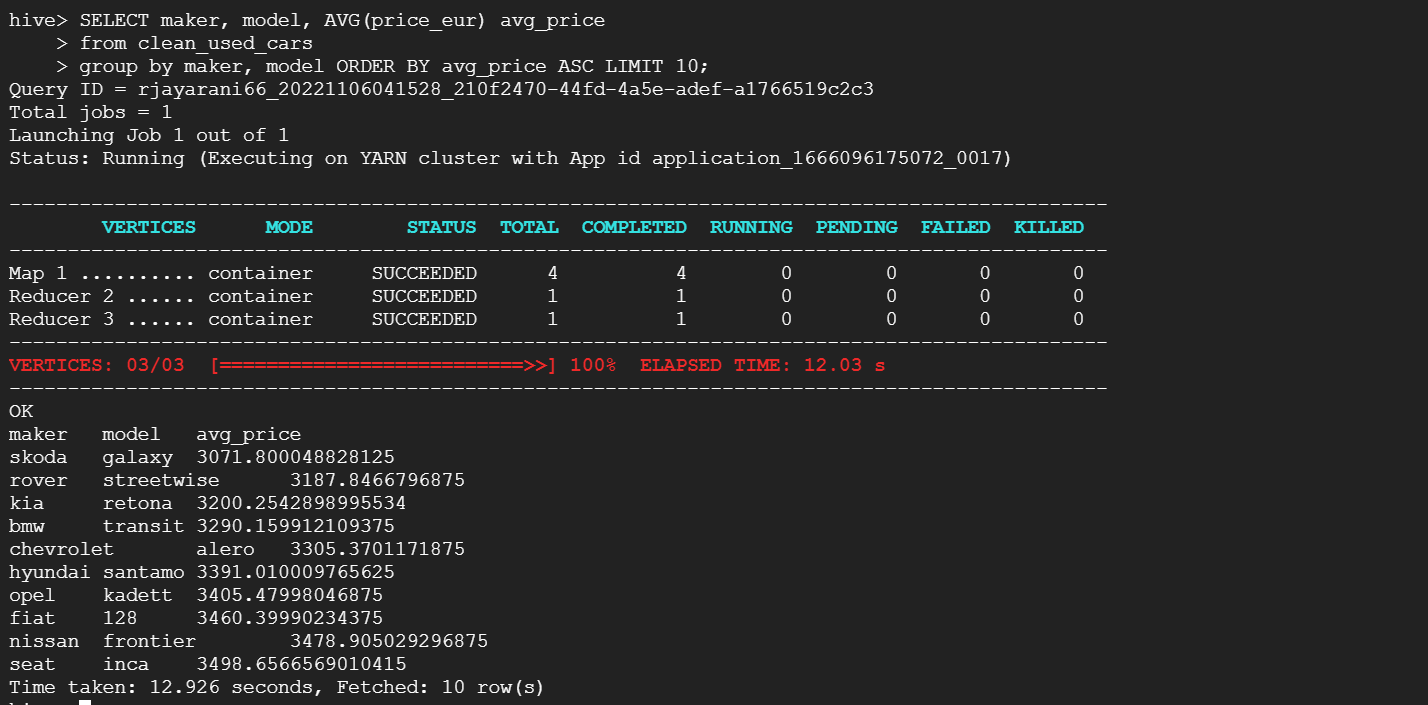
Count of number of rows in clean\_used\_cars table.

****

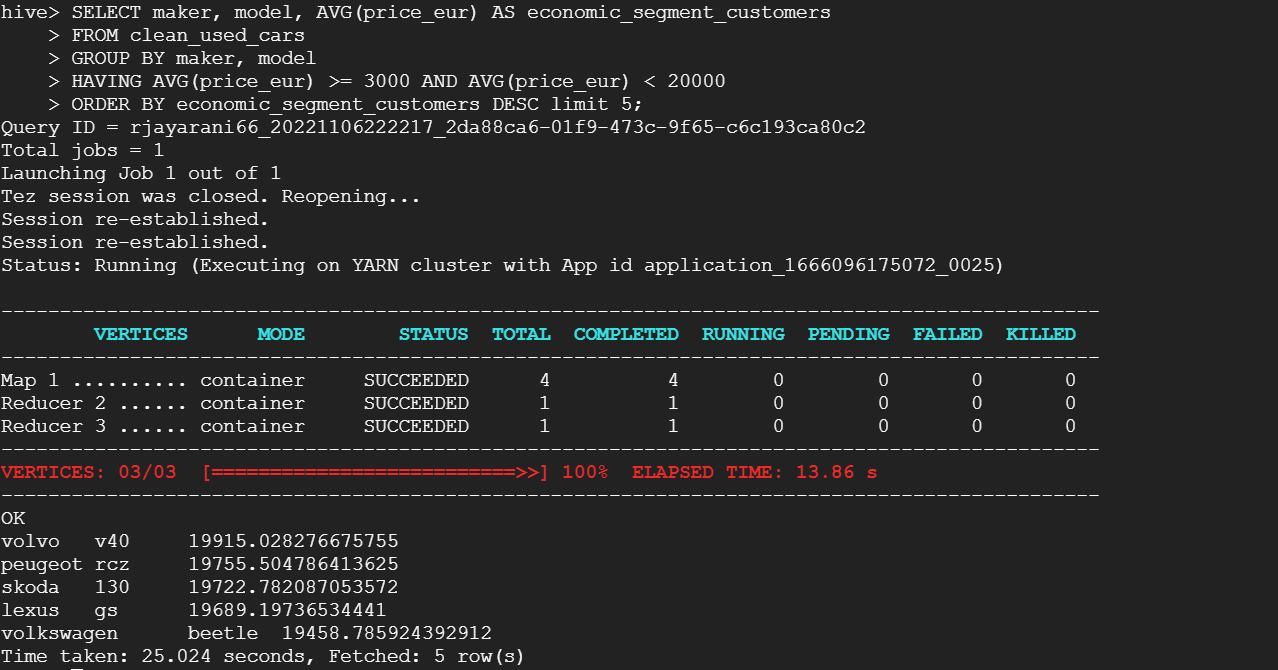
8. Write a Hive query to find the make and model for the cars with the top 10 **highest average price**

****

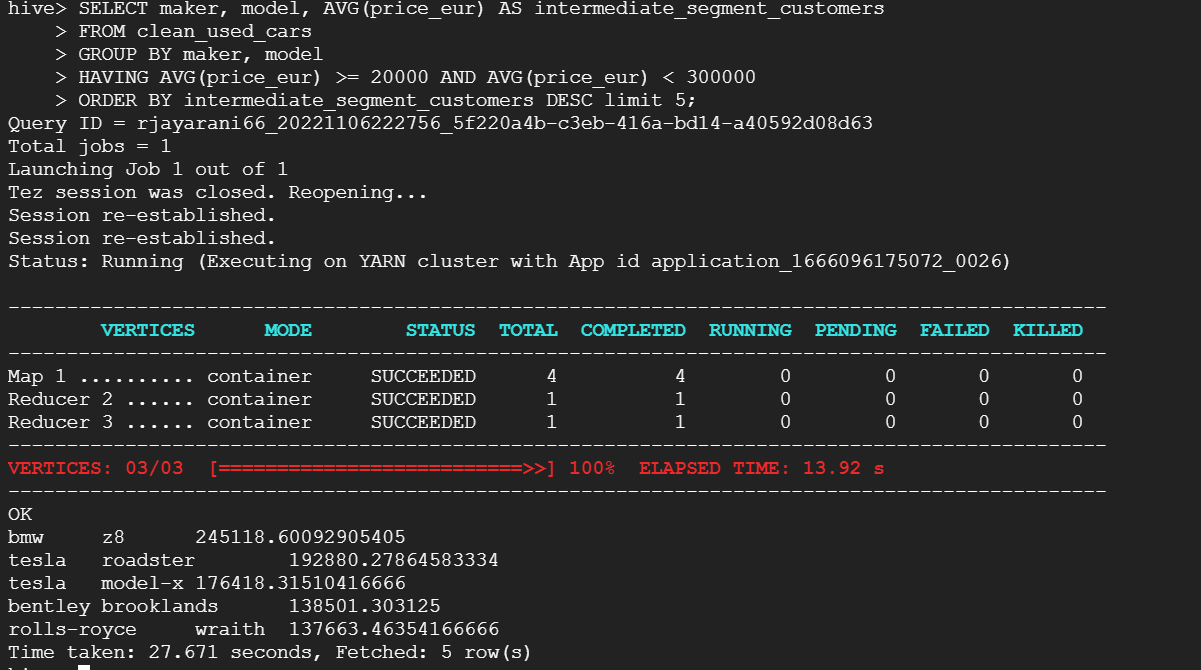
9. Write a Hive query to find the make and model for the cars with the top 10 **lowest average price**

****

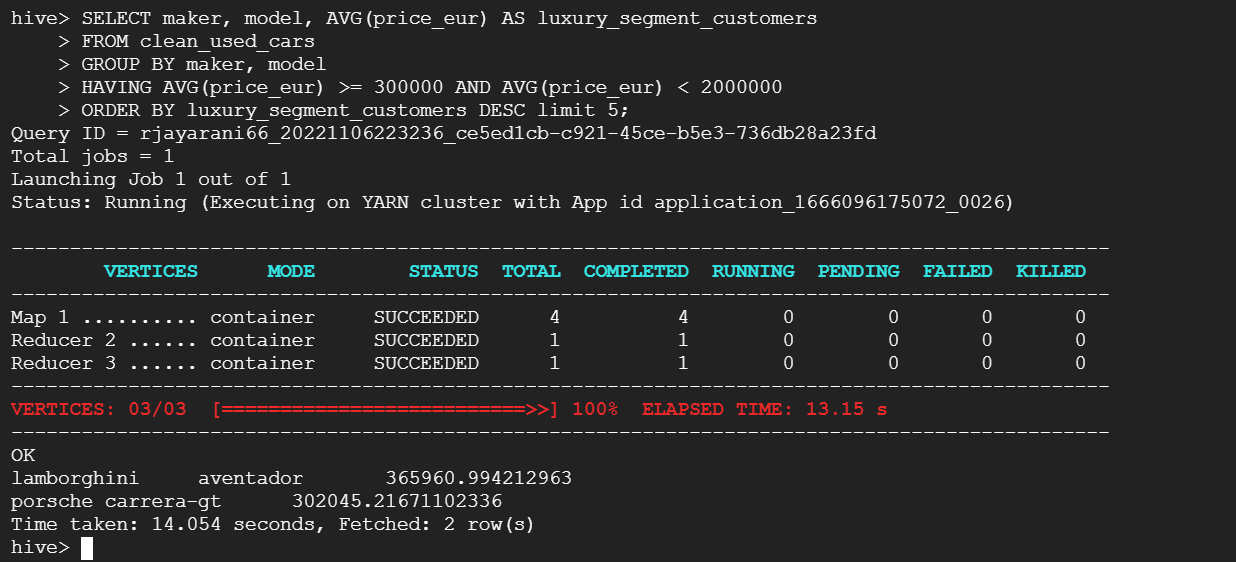
10. Write a Hive query to recommend top five make and model for **Economic** segment customers (Top five manufacturers in the 3000 to 20,000 price range;3000≤price<20,000) - based on the top **average price**



11. Write a Hive query to recommend top five make and model for **Intermediate** segment customers (Top five manufacturers in the 20,000 to 300,000 price range; 3000≤price<20,000) - based on the top **average price**



12. Write a Hive query to recommend the top five make and model for the **Luxury** segment customers (Top five manufacturers in the 300,000 to 2000,000 price range; 300,000≤price<2000,000) - based on the top **average price**

****

**Hive codes**

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

wget <https://www.dropbox.com/s/ehehgr7f97y4sw0/all_anonymized_2015_11_2017_03.csv?dl=0>

hadoop fs -mkdir /Bigdata

hadoop fs -copyFromLocal cars\_dataset /Bigdata/.

hadoop fs -ls /Bigdata/cars\_dataset

CREATE DATABASE used\_cars\_db;

USE used\_cars\_db;

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

**3.** Create a table such that the date column is read correctly based on the format in the dataset

CREATE TABLE IF NOT EXISTS used\_cars (maker STRING, model STRING, mileage FLOAT, manufacture\_year INT, engine\_displacement FLOAT,

engine\_power STRING, body\_type STRING, color\_slug STRING, stk\_year FLOAT, transmission STRING, door\_count INT, seat\_count INT,

fuel\_type STRING, date\_created DATE, date\_last\_seen DATE, price\_eur FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

DESCRIBE used\_cars;

!hadoop fs -ls /Bigdata/cars\_dataset

LOAD DATA LOCAL INPATH '/home/rjayarani66/cars\_dataset' INTO TABLE used\_cars;

SELECT \* FROM used\_cars LIMIT 10;

SELECT COUNT(\*) FROM used\_cars;

SET hive.cli.print.header=true;

SELECT maker, model FROM used\_cars LIMIT 10;

CREATE TABLE IF NOT EXISTS used\_cars1 AS

SELECT maker, model, mileage, manufacture\_year, engine\_displacement,engine\_power, body\_type, color\_slug, stk\_year,

transmission, door\_count, seat\_count, fuel\_type, date\_created, date\_last\_seen, price\_eur

FROM used\_cars;

INSERT OVERWRITE TABLE used\_cars1

SELECT case when maker = '' or maker = 'None' THEN NULL ELSE maker END as maker,

case when model = '' or model = 'None' THEN NULL ELSE model END as model,

case when mileage = '' or mileage = 'None' THEN NULL ELSE mileage END as mileage,

case when manufacture\_year = '' or manufacture\_year = 'None' THEN NULL ELSE manufacture\_year END as manufacture\_year,

case when engine\_displacement = '' or engine\_displacement = 'None' THEN NULL ELSE engine\_displacement END as engine\_displacement,

case when engine\_power = '' or engine\_power = 'None' THEN NULL ELSE engine\_power END as engine\_power,

case when body\_type = '' or body\_type = 'None' THEN NULL ELSE body\_type END as body\_type,

case when color\_slug = '' or color\_slug = 'None' THEN NULL ELSE color\_slug END as color\_slug,

case when stk\_year = '' or stk\_year = 'None' THEN NULL ELSE stk\_year END as stk\_year,

case when transmission = '' or transmission = 'None' THEN NULL ELSE transmission END as transmission,

case when door\_count = '' or door\_count = 'None' THEN NULL ELSE door\_count END as door\_count,

case when seat\_count = '' or seat\_count = 'None' THEN NULL ELSE seat\_count END as seat\_count,

case when fuel\_type = '' or fuel\_type = 'None' THEN NULL ELSE fuel\_type END as fuel\_type,

date\_created, date\_last\_seen,

case when price\_eur = '' or price\_eur = 'None' THEN NULL ELSE price\_eur END as price\_eur

FROM used\_cars1;

4. Write Hive queries to see how many missing values you have in each attribute? Based on the results, document how many missing values in each column we have. Especially, mention those columns with more than 50% missing values.

select count(\*) from used\_cars1 where maker is NULL

select count(\*) from used\_cars1 where model is NULL

select count(\*) from used\_cars1 where mileage is NULL

select count(\*) from used\_cars1 where manufacture\_year is NULL

select count(\*) from used\_cars1 where engine\_displacement is NULL

select count(\*) from used\_cars1 where engine\_power is NULL

select count(\*) from used\_cars1 where body\_type is NULL

select count(\*) from used\_cars1 where color\_slug is NULL

select count(\*) from used\_cars1 where stk\_year is NULL

select count(\*) from used\_cars1 where transmission is NULL

select count(\*) from used\_cars1 where door\_count is NULL

select count(\*) from used\_cars1 where seat\_count is NULL

select count(\*) from used\_cars1 where fuel\_type is NULL

select count(\*) from used\_cars1 where date\_created is NULL

select count(\*) from used\_cars1 where date\_last\_seen is NULL

select count(\*) from used\_cars1 where price\_eur is NULL;

5. Group the price column and count the number of unique prices. Do you notice if there is a single price that is repeating across the ads?

SELECT price\_eur, count(\*) agg(price\_eur)

FROM used\_cars1

GROUP BY price\_eur

ORDER BY price\_eur DESC;

* 1. 6. Write a Hive query to create a new table called **clean\_used\_cars** from **used\_cars** with the following conditions: o Drop the columns with more than 50% missing values
  2. o The manufacture year between 2000 and 2017 including 2000 and 2017
  3. o Both maker and model exist in the row
  4. o The price range is from 3000 to 2000,000 (3000 ≤ price ≤ 2000,000)
  5. o Remove any price you singled out in Step 3 (ie a price that repeats too frequently for a random set of ads).

CREATE TABLE IF NOT EXISTS clean\_used\_cars AS

SELECT maker, model, mileage, manufacture\_year, engine\_displacement,engine\_power, body\_type, stk\_year,

transmission, door\_count, seat\_count, date\_created, date\_last\_seen, price\_eur

FROM used\_cars WHERE manufacture\_year >= 2000 AND manufacture\_year <= 2017

AND price\_eur >= 3000 And price\_eur <= 2000000

AND maker != '' AND model != ''

AND price\_eur != 1295.34;

SELECT \* FROM clean\_used\_cars LIMIT 10;

7. Write a Hive query to find how many records remained **clean\_used\_cars**

SELECT COUNT(\*) FROM clean\_used\_cars;

8. Write a Hive query to find the make and model for the cars with the top 10 **highest average price**

SELECT maker, model, AVG(price\_eur) avg\_price

FROM clean\_used\_cars

GROUP BY maker, model

ORDER BY avg\_price DESC limit 10;

9. Write a Hive query to find the make and model for the cars with the top 10 **lowest average price**

SELECT maker, model, AVG(price\_eur) avg\_price

from clean\_used\_cars

group by maker, model

ORDER BY avg\_price ASC LIMIT 10;

10. Write a Hive query to recommend top five make and model for **Economic** segment customers (Top five manufacturers in the 3000 to 20,000 price range;3000≤price<20,000) - based on the top **average price**

SELECT maker, model, AVG(price\_eur) AS economic\_segment\_customers

FROM clean\_used\_cars

GROUP BY maker, model

HAVING AVG(price\_eur) >= 3000 AND AVG(price\_eur) < 20000

ORDER BY economic\_segment\_customers DESC limit 5;

11. Write a Hive query to recommend top five make and model for **Intermediate** segment customers (Top five manufacturers in the 20,000 to 300,000 price range; 3000≤price<20,000) - based on the top **average price**

SELECT maker, model, AVG(price\_eur) AS intermediate\_segment\_customers

FROM clean\_used\_cars

GROUP BY maker, model

HAVING AVG(price\_eur) >= 20000 AND AVG(price\_eur) < 300000

ORDER BY intermediate\_segment\_customers DESC limit 5;

12. Write a Hive query to recommend the top five make and model for the **Luxury** segment customers (Top five manufacturers in the 300,000 to 2000,000 price range; 300,000≤price<2000,000) - based on the top **average price**

SELECT maker, model, AVG(price\_eur) AS luxury\_segment\_customers

FROM clean\_used\_cars

GROUP BY maker, model

HAVING AVG(price\_eur) >= 300000 AND AVG(price\_eur) < 2000000

ORDER BY luxury\_segment\_customers DESC limit 5**;**

**Reference**

Lecture notes and exercises.