Data Systems Architecture: BDAT 1002
Mid-Term Project.

Prepared By:

Prateek Singh
Student ID: 200506901

Table of Contents

Overview:

This analysis is a deep dive into the sales of a Used Cars Businesses. The data obtained from all the dealerships of these businesses contains 3.5 million records of cars sold over two years, 2015 and 2016. Due to the sheet size of the data recorded, large computing tools such as Google Cloud Platform was used to analyze the data, with the subsequent reports and pivot tables being generated in MS Excel

Scope:

The Report outlines high level business trends that were obtained from this data and provides analyses that can be used to implement business decisions.

Methodology and Glossary:

The data was scraped from ads for used cars from several websites in German and Czech Republic. The original Dataset was messy and required significant cleaning and ordering, several values were dropped and the final set of usable values of the cars is as follows.

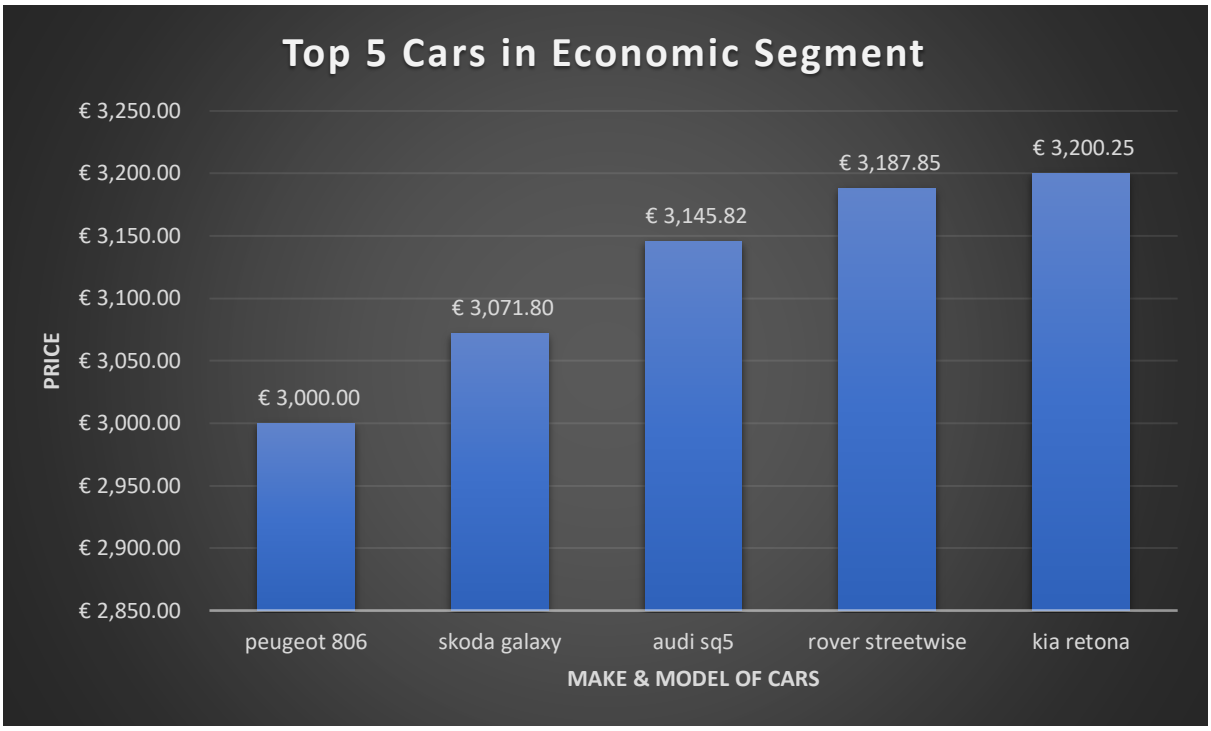
maker	Car Maker Company
model	Car Name
manufacture_year	Year of manufacture for the car
Phone_number (previously mileage, it was incorrectly scraped)	Number Listed on Ad
engine_displacement	the measure of the cylinder volume swept by all of the pistons of a piston engine, excluding the combustion chambers (in CCM)
engine_power	Engine power is the power that an engine can put out. It can be expressed in power units, in Kilowatts. (KW)
body_type	Only personal cars
color_slug	Body Color
stk_year	Year of the last emission control
transmission	Automatic or Manual
door_count	Number of doors on the car
seat_count	Number of seats in the car
fuel_type	Gasoline, diesel, cng, lpg, 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	Listed price converted to EUR

Market Analysis

Based on the data received and analyzed from the dataset the following observations can be made to improve profits.

Economic Segment Analysis:

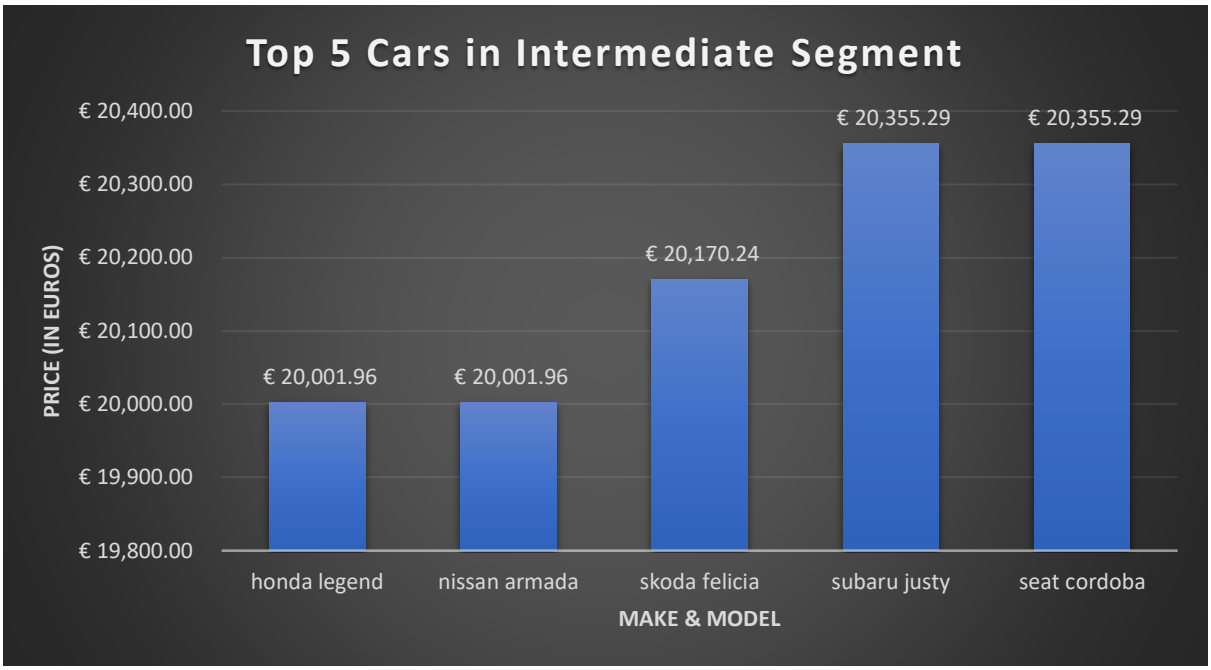
The cars of the economic segment continue to be the cars with the highest volume. This is not a surprise as apart from the cars that are priced at economic prices, the values of cars from the intermediate segment deprecate to prices that can be considered economic. The cars in this segment are all the cars that are priced between € 3,000.00 and € 20,000



These are the top 5 cars in the Economic segment. These are the cars that are affordable by even the lowest earning buyers in the Economic Segment. A company planning to capitalize on this market would need to stock more cars of these makes and models.

Intermediate Segment Analysis:

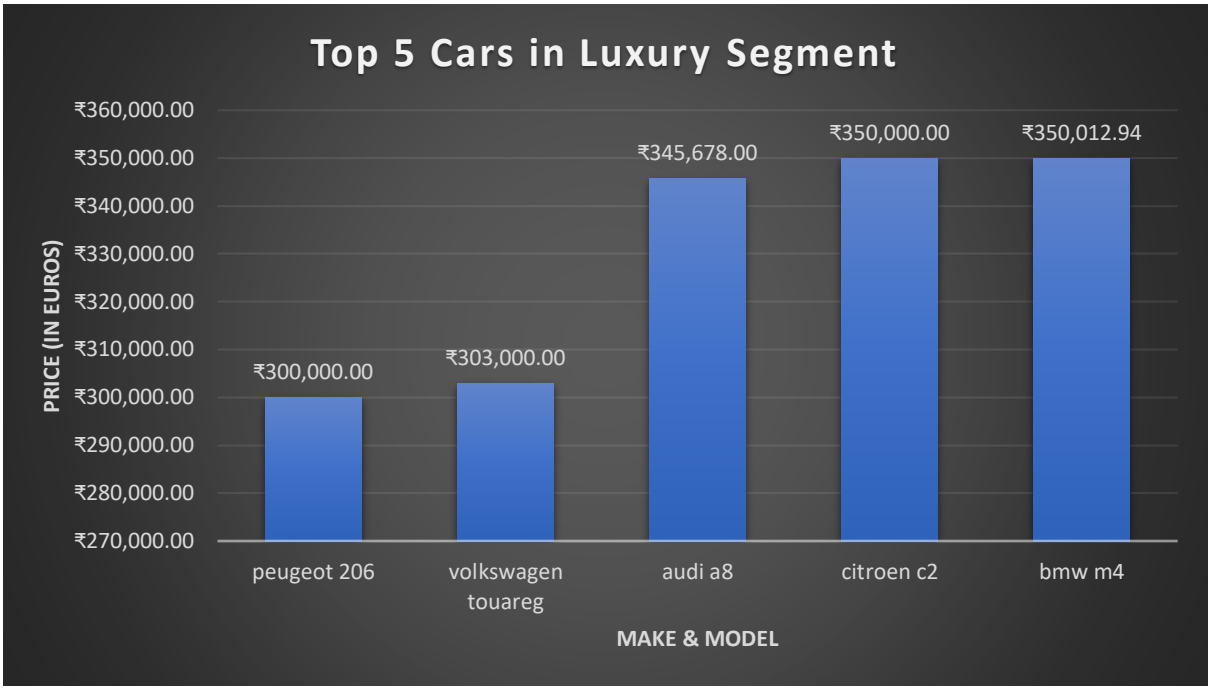
These are the cars that have an average price range from € 20,000 to € 300,000. The target market audience for cars like these would be mid to high-income customers.



These are the top 5 cars in the Economic segment. These are the cars that are affordable by even the lowest earning buyers in the Intermediate Segment. A company planning to capitalize on this market would need to stock more cars of these makes and models.

Luxury Segment Analysis:

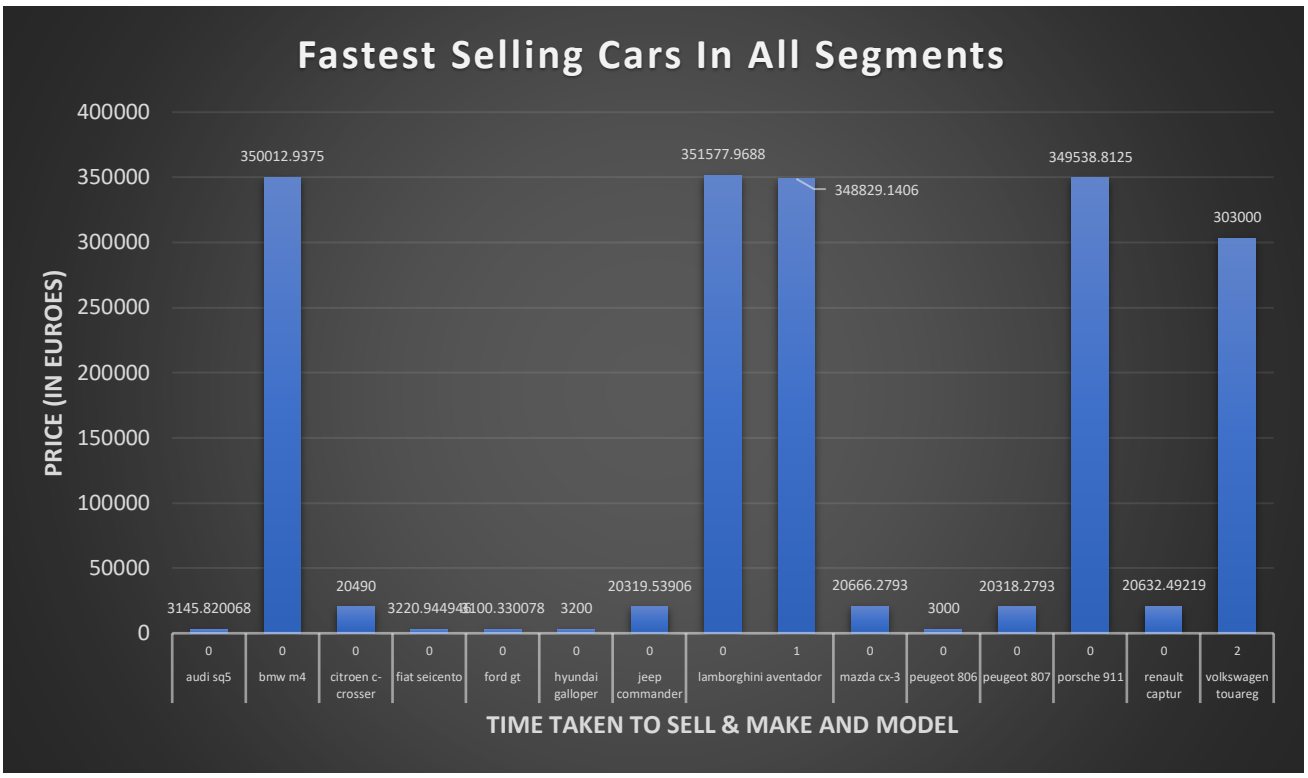
In this segment, car functionality takes the back burner with car prices in the range of €300,000 to €2000,000. This segment is reserved for Extremely high-income clients with a focus on the choice of the car rather than aspects such as mileage, engine power and price as is observed by clients in other segments.



A company whose model focuses on the uber wealthy must invest in a stock of makers such as these with an emphasis on these car manufacturing companies

Analysis of fastest selling cars:

An analysis of the fastest selling cars gives car dealership owners the chance to create a catalogue which ensures a steady income. Here, the time spent by an advertisement for the car was taken from the date it was listed and the date it was removed. The analysis shows the following for each segment:



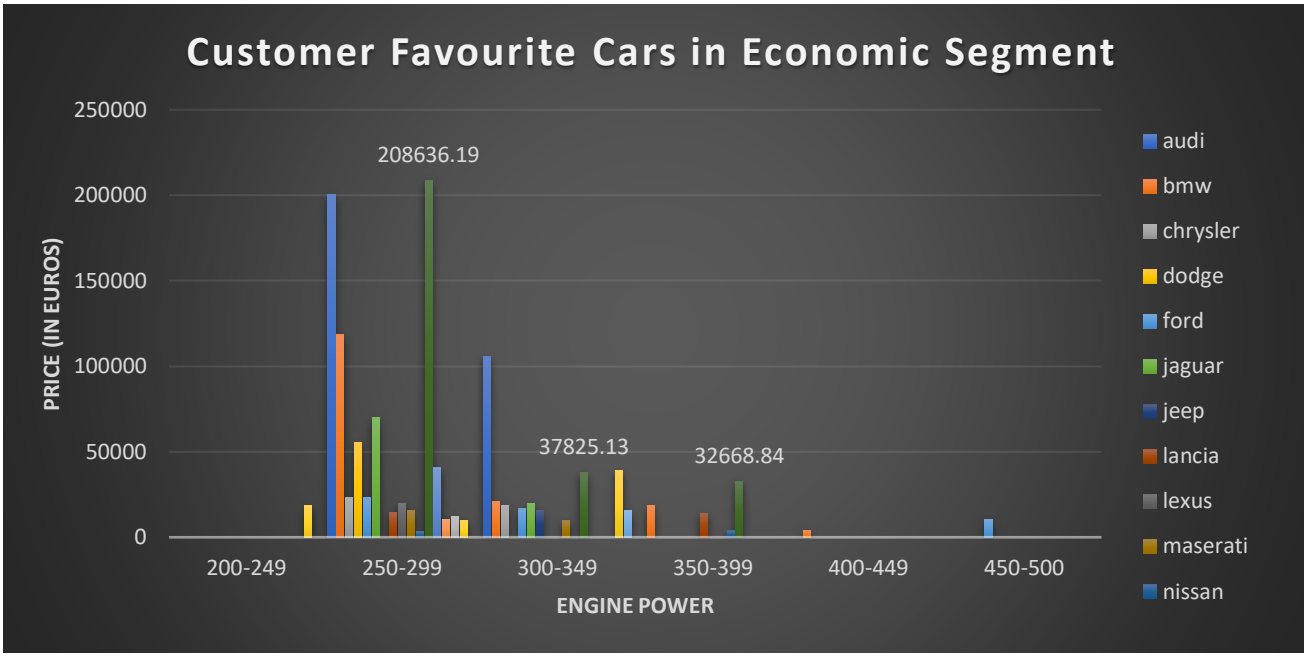
Here, the time taken to sell each car is given in the x-axis with the price at which it was sold on the y-axis. Each segment can be identified using the following table

Time Taken to sell (in Days)	Maker and model	Price (in Euros)
Economic Segment		
0	peugeot 806	3000
0	ford gt	3100.330078
0	audi sq5	3145.820068
0	hyundai galloper	3200
0	fiat seicento	3220.944946
Intermediate Segment		
0	peugeot 807	20318.2793
0	jeep commander	20319.53906
0	citroen c-crosser	20490
0	renault captur	20632.49219
0	mazda cx-3	20666.2793
Luxury Segment		
0	porsche 911	349538.8125
0	bmw m4	350012.9375
0	lamborghini aventador	351577.9688
1	lamborghini aventador	348829.1406
2	volkswagen touareg	303000

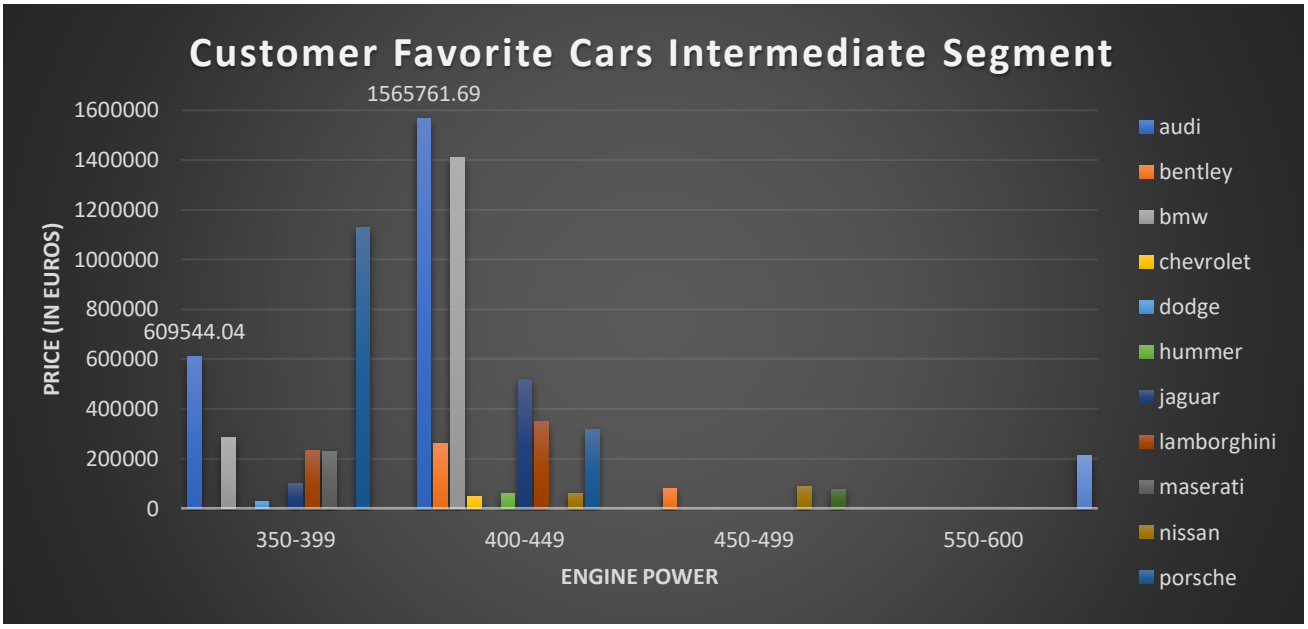
Even though the cars in the economic segment sell fast, they don't have a very high-profit margin. Used-cars businesses should target easily available cars in the intermediate and luxury segments to boost profit in a crunch.

Customer Preference Analysis:

An analysis of the fastest-selling cars also shows that even though there are cars with better performance in terms of engine power with negligible price differences, customers prefer to buy a particular make of car more than the other.



In the Economic Segment, it's Jaguar. It's observed in the graph above that even though there are better cars with lower price values, customers choose to buy Jaguars.



In the intermediate segment, the mantle goes to Audi. As it can be observed, Audi has one of the highest prices with comparatively lower engine power in its segment. Yet it was sold in less than 1 day.

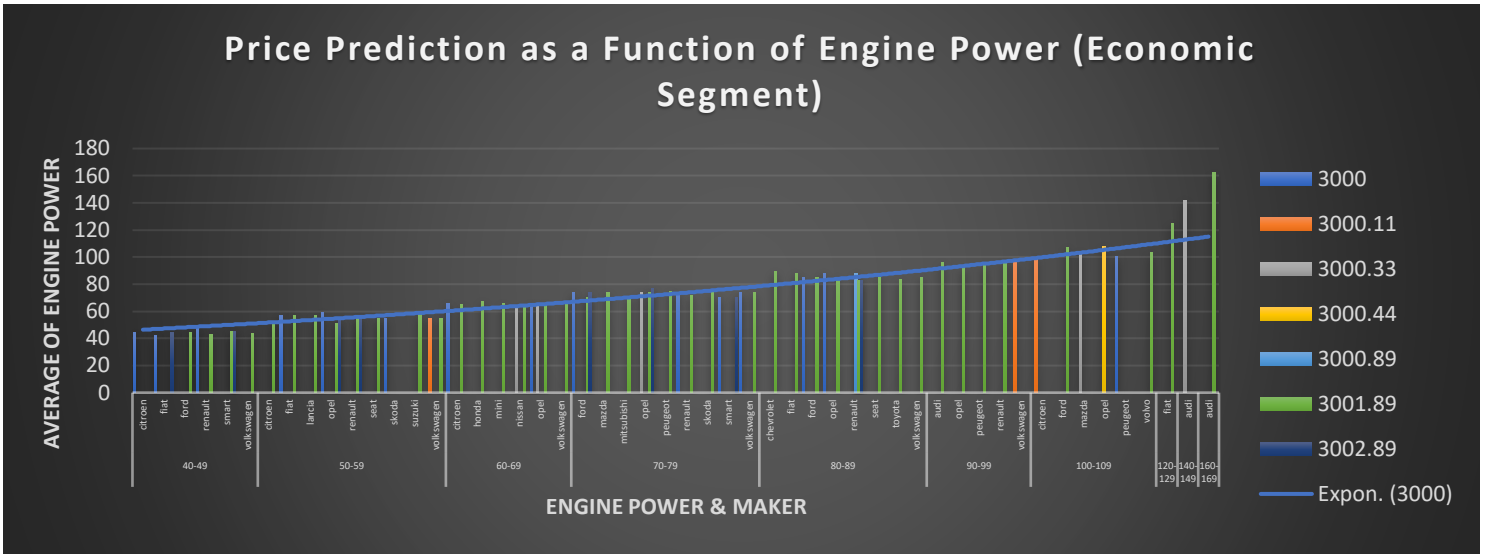
These analyses suggest dealerships should invest in acquiring these makes over others.

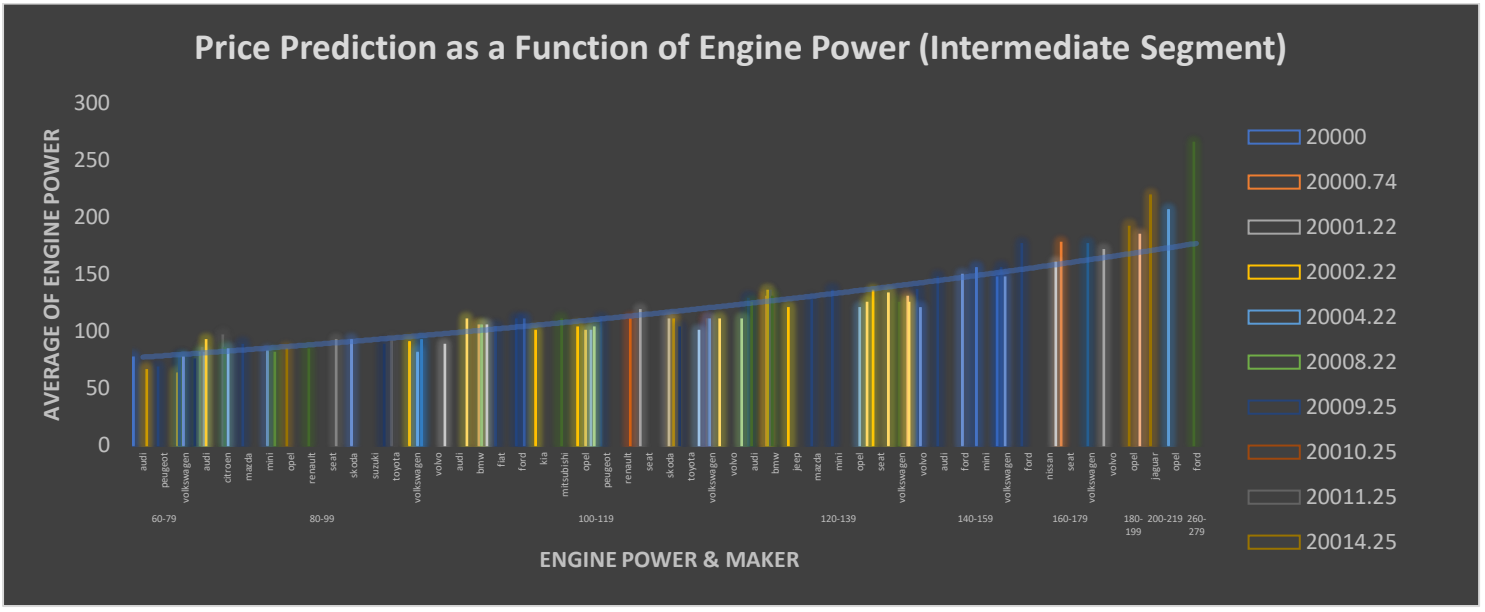
Unfortunately, this analysis only applies to these segments as people who shop in the Luxury segment do not show any interest in the performance of a vehicle.

Price Prediction Analysis

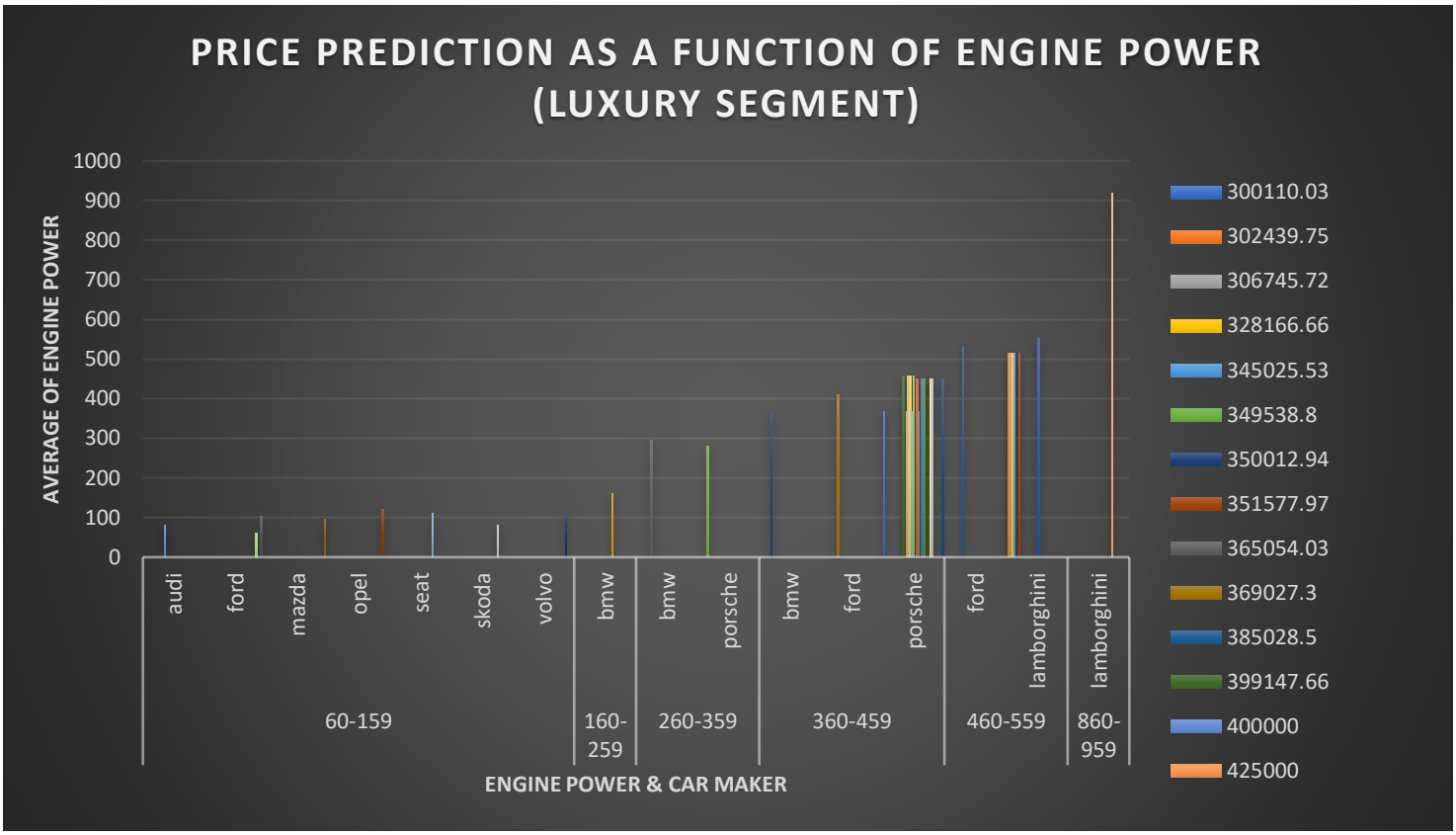
The analysis of the top 100 cars in the Economic and Intermediate Segments shows a direct relationship between the cost of the car and the engine power. This shows that engine power is a predominant factor when it comes to price determination.

The trend lines in the below-given graphs show an upward bell curve, proving that the company should also focus on investing capital in maintaining the engine performance of its vehicles. This will result in a justifiable higher price allocation, one which the client will not be able to deny.





The luxury segment shows no trend, this is a result of the lesser number of cars in this segment. With a higher number of cars, this trend would no doubt be visible here as well.



Codes and Screenshots:

Uploading Dataset and Creating Folder structure:

Codes:

```
wget https://www.dropbox.com/s/rsrxro7r1c5a4i2/cars.csv

hadoop fs -mkdir /Bigdata/

hadoop fs -mkdir /Bigdata/hive/

hadoop fs -copyFromLocal cars.csv /Bigdata/hive/

CREATE DATABASE midtermproject_db;

USE midtermproject_db;
```

```
ps21031994@cluster-08b2-m:~$ wget https://www.dropbox.com/s/rsrxro7r1c5a4i2/cars.csv
--2022-07-22 14:25:54--  https://www.dropbox.com/s/rsrxro7r1c5a4i2/cars.csv
Resolving www.dropbox.com (www.dropbox.com)... 162.125.3.18, 2620:100:6018:18::a27d:312
Connecting to www.dropbox.com (www.dropbox.com)|162.125.3.18|:443... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: /s/raw/rsrxro7r1c5a4i2/cars.csv [following]
--2022-07-22 14:25:55--  https://www.dropbox.com/s/raw/rsrxro7r1c5a4i2/cars.csv
Reusing existing connection to www.dropbox.com:443.
HTTP request sent, awaiting response... 302 Found
Location: https://uc13f12c716fc5cf3aa8fac58be6.dl.dropboxusercontent.com/cd/0/inline/BpkqWWG0isBNI1Ulu-qzS6ipKnI4a1En
9C7ITl3QgvrfO3HDOpfql2hqKZUzu9-wMfSq6SoU0NM_9UWa0cTNedSNIGe0k_BQe4iOR4-Y8dzZwjOpfK9p3E4VCUNdiaSObAouRNxgViDPcU3GDjfEm
UfgzFph5QQz1kh-unimNCXfug/file# [following]
--2022-07-22 14:25:55--  https://uc13f12c716fc5cf3aa8fac58be6.dl.dropboxusercontent.com/cd/0/inline/BpkqWWG0isBNI1Ulu
-qzS6ipKnI4a1En9C7ITl3QgvrfO3HDOpfql2hqKZUzu9-wMfSq6SoU0NM_9UWa0cTNedSNIGe0k_BQe4iOR4-Y8dzZwjOpfK9p3E4VCUNdiaSObAouRN
xgViDPcU3GDjfEmUfgzFph5QQz1kh-unimNCXfug/file
Resolving uc13f12c716fc5cf3aa8fac58be6.dl.dropboxusercontent.com (uc13f12c716fc5cf3aa8fac58be6.dl.dropboxusercontent.
com)... 162.125.3.15, 2620:100:6018:15::a27d:30f
Connecting to uc13f12c716fc5cf3aa8fac58be6.dl.dropboxusercontent.com (uc13f12c716fc5cf3aa8fac58be6.dl.dropboxusercont
ent.com)|162.125.3.15|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 419466302 (400M) [text/plain]
Saving to: 'cars.csv.1'

cars.csv.1                  100%[=============================================>] 400.03M  67.4MB/s    in 6.3s

2022-07-22 14:26:02 (63.1 MB/s) - 'cars.csv.1' saved [419466302/419466302]
```

```
hive> CREATE DATABASE midtermproject_db;
OK
Time taken: 0.068 seconds
```

```
hive>
    > USE midtermproject_db;
OK
Time taken: 0.039 seconds
```

Cleaning:

Creating tables:

Codes:

```
CREATE TABLE IF NOT EXISTS cars_rawdata (
maker STRING,
model STRING,
phone_number INT,
manufacture_year STRING,
engine_displacement FLOAT,
engine_power FLOAT,
body_type STRING,
color_slug STRING,
stk_year STRING,
transmission STRING,
door_count INT,
seat_count INT,
fuel_type STRING,
date_created STRING,
datelastseen STRING,
price_eur FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

LOAD DATA INPATH '/Bigdata/hive/cars.csv' INTO TABLE cars_rawdata;

CREATE TABLE IF NOT EXISTS used_cars (
maker STRING,
model STRING,
phone_number INT,
manufacture_year INT,
engine_displacement FLOAT,
engine_power FLOAT,
body_type STRING,
color_slug STRING,
stk_year INT,
transmission STRING,
door_count INT,
seat_count INT,
fuel_type STRING,
date_created TIMESTAMP,
datelastseen TIMESTAMP,
price_eur FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

INSERT INTO TABLE used_cars
SELECT maker, model, phone_number, manufacture_year, engine_displacement, engine_power,
body_type, color_slug, stk_year,
transmission ,door_count, seat_count, fuel_type,
from_unixtime(unix_timestamp(date_created, 'yyyy-mm-dd')),
from_unixtime(unix_timestamp(datelastseen, 'yyyy-mm-dd')), price_eur
FROM cars_rawdata;
```

```
hive> CREATE TABLE IF NOT EXISTS cars_rawdata (
    > maker STRING,
    > model STRING,
    > phone_number INT,
    > manufacture_year STRING,
    > engine_displacement FLOAT,
    > engine_power FLOAT,
    > body_type STRING,
    > color_slug STRING,
    > stk_year STRING,
    > transmission STRING,
    > door_count INT,
    > seat_count INT,
    > fuel_type STRING,
    > date_created STRING,
    > datelastseen STRING,
    > price_eur FLOAT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.11 seconds
hive> LOAD DATA INPATH '/Bigdata/hive/cars.csv' INTO TABLE cars_rawdata;
Loading data to table midtermproject_db.cars_rawdata
OK
Time taken: 0.187 seconds
```

```
hive> CREATE TABLE IF NOT EXISTS used_cars (
    > maker STRING,
    > model STRING,
    > phone_number INT,
    > manufacture_year INT,
    > engine_displacement FLOAT,
    > engine_power FLOAT,
    > body_type STRING,
    > color_slug STRING,
    > stk_year INT,
    > transmission STRING,
    > door_count INT,
    > seat_count INT,
    > fuel_type STRING,
    > date_created TIMESTAMP,
    > datelastseen TIMESTAMP,
    > price_eur FLOAT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.107 seconds
```

```
hive> INSERT INTO TABLE used_cars
    > SELECT maker, model, phone_number, manufacture_year, engine_displacement, engine_power, body_type, color_slug,
stk_year,
    > transmission ,door_count, seat_count, fuel_type, from_unixtime(unix_timestamp(date_created, 'yyyy-mm-dd')), fro
m_unixtime(unix_timestamp(datelastseen, 'yyyy-mm-dd')), price_eur
    > FROM cars_rawdata;
Query ID = ps21031994_20220722143904_23e2f840-311f-4fdd-a299-5bc67fc1059c
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_1658444852694_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 51.22 s
--------------------------------------------------------------------------------
Loading data to table midtermproject_db.used_cars
OK
Time taken: 59.691 seconds
```

Value Analysis:

Null Value Analysis:

Codes:

```
SELECT COUNT(maker) FROM used_cars WHERE maker = "";
SELECT COUNT(model) FROM used_cars WHERE model = "";
SELECT COUNT(ISNULL(phone_number)) FROM used_cars WHERE phone_number IS NULL;
SELECT COUNT(ISNULL(manufacture_year)) FROM used_cars WHERE manufacture_year IS NULL;
SELECT COUNT(ISNULL(engine_displacement)) FROM used_cars WHERE engine_displacement IS
NULL;
SELECT COUNT(ISNULL(engine_power)) FROM used_cars WHERE engine_power IS NULL;
SELECT COUNT(body_type) FROM used_cars WHERE body_type = "";
SELECT COUNT(color_slug) FROM used_cars WHERE color_slug = "";
SELECT COUNT(ISNULL(stk_year)) FROM used_cars WHERE stk_year IS NULL;
SELECT COUNT(transmission) FROM used_cars WHERE transmission = "";
SELECT COUNT(ISNULL(door_count)) FROM used_cars WHERE door_count IS NULL;
SELECT COUNT(ISNULL(seat_count)) FROM used_cars WHERE seat_count IS NULL;
SELECT COUNT(fuel_type) FROM used_cars WHERE fuel_type = "";
SELECT COUNT(ISNULL(date_created)) FROM used_cars WHERE date_created IS NULL;
SELECT COUNT(ISNULL(datelastseen)) FROM used_cars WHERE datelastseen IS NULL;
SELECT COUNT(ISNULL(price_eur)) FROM used_cars WHERE price_eur IS NULL;
```

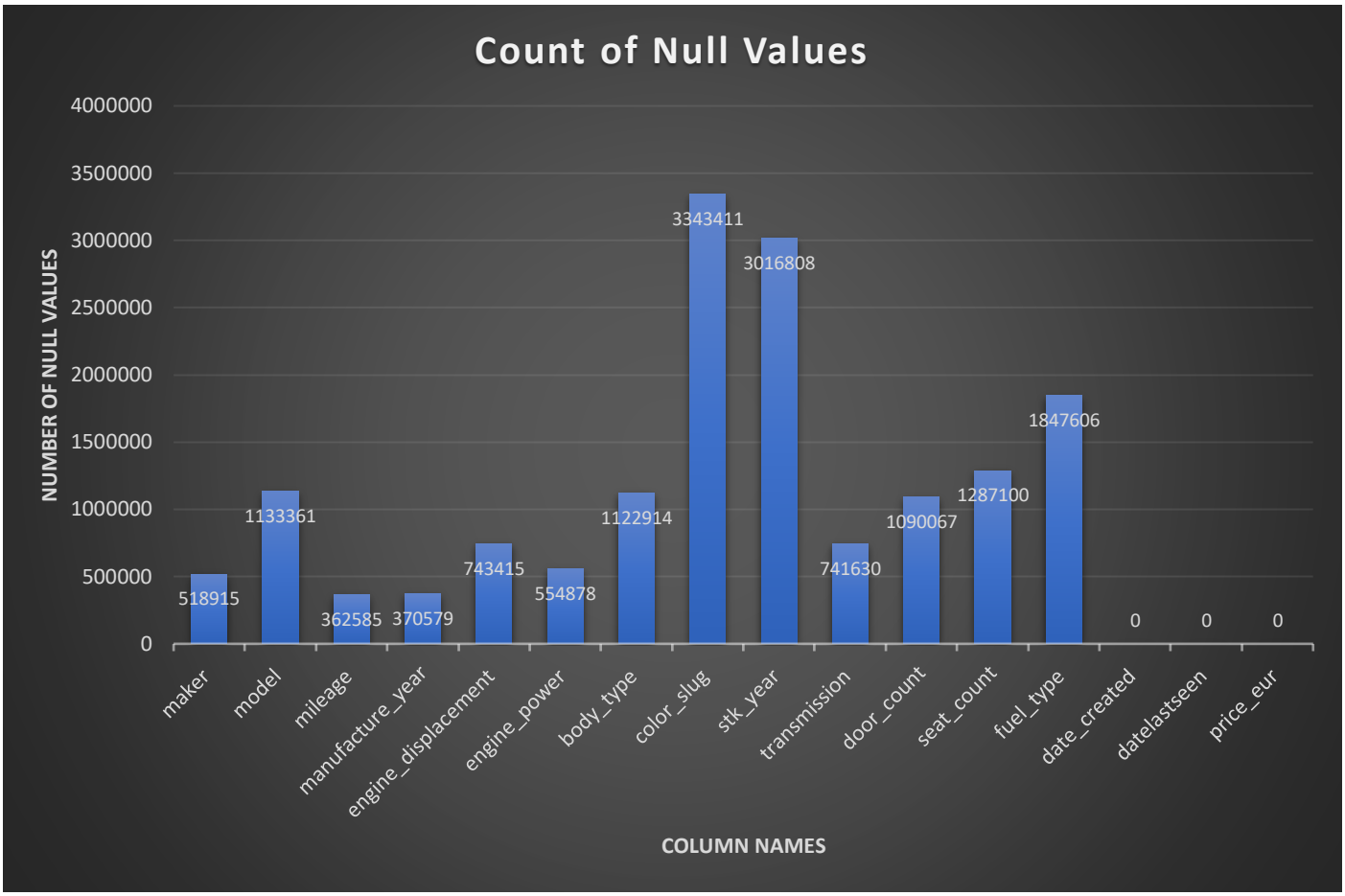
```
hive> SELECT COUNT(body_type) FROM used_cars WHERE body_type = "";
Query ID = ps21031994_20220722145152_e0a902d3-ebac-4425-b449-2a3e67421928
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.69 s
----------------------------------------------------------------------------------------------
OK
1122914
Time taken: 1.53 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(color_slug) FROM used_cars WHERE color_slug = "";
Query ID = ps21031994_20220722145153_c18fbcc6-3661-4158-bc30-b07b79053556
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.74 s
----------------------------------------------------------------------------------------------
OK
3343411
Time taken: 1.568 seconds, Fetched: 1 row(s)
```

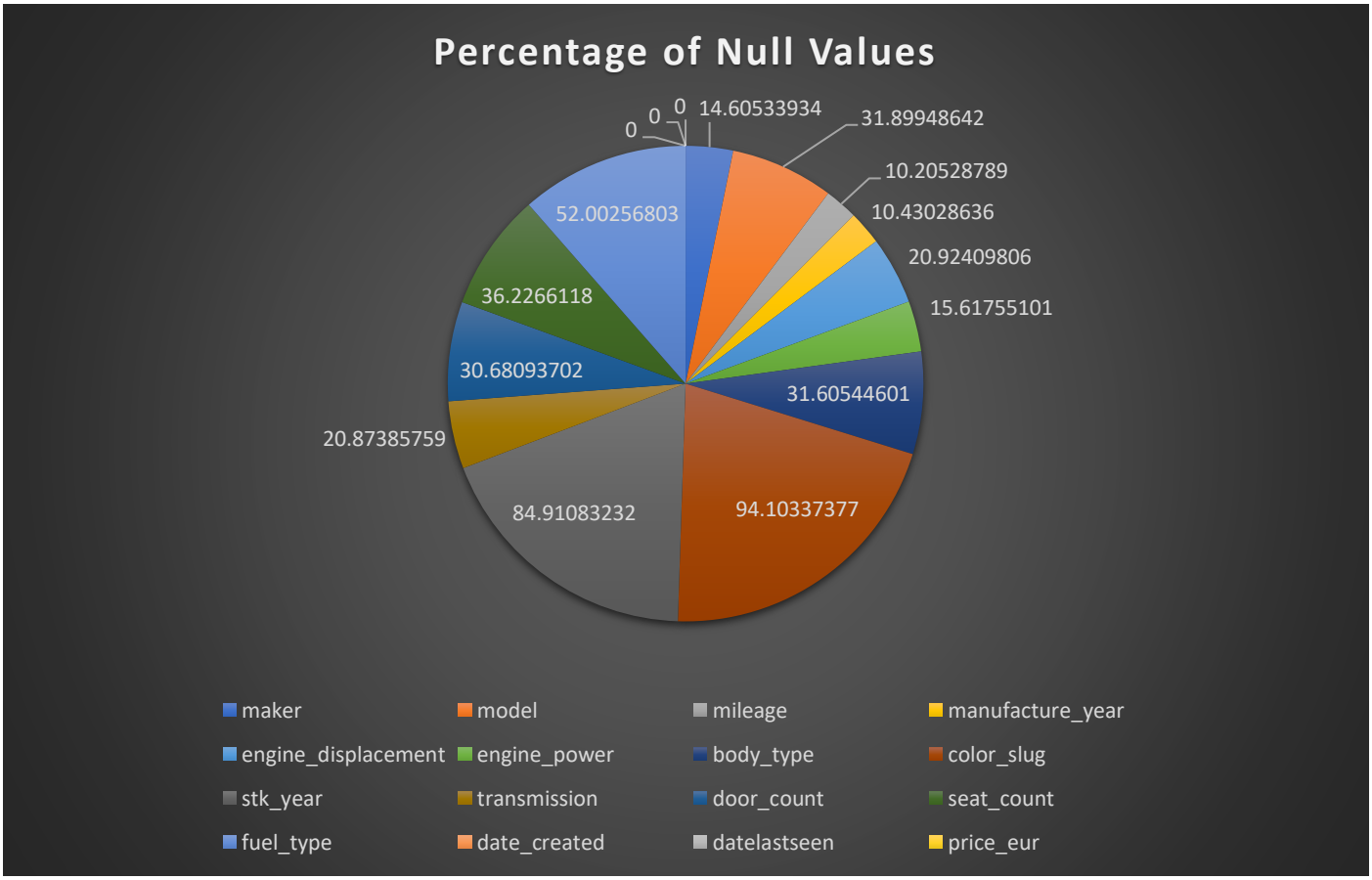
```
hive> SELECT COUNT(ISNULL(date_created)) FROM used_cars WHERE date_created IS NULL;
Query ID = ps21031994_20220722145205_ad22d07e-a0d1-4659-b23f-0c72d2fd0e78
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.82 s
--------------------------------------------------------------------------------
OK
1
Time taken: 6.707 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(ISNULL(datelastseen)) FROM used_cars WHERE datelastseen IS NULL;
Query ID = ps21031994_20220722145211_69777a35-2cd8-49ab-a152-cd56da1d6985
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.62 s
--------------------------------------------------------------------------------
OK
1
Time taken: 5.473 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(ISNULL(door_count)) FROM used_cars WHERE door_count IS NULL;
Query ID = ps21031994_20220722145159_05b4955a-8bd1-40f6-b0ad-f555b4d1527f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.82 s
--------------------------------------------------------------------------------
OK
1090067
Time taken: 1.781 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(ISNULL(engine_displacement)) FROM used_cars WHERE engine_displacement IS NULL;
Query ID = ps21031994_20220722145147_42dab374-0ba2-4180-9b27-139cd6fd4c69
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 1.72 s
--------------------------------------------------------------------------------
OK
743415
Time taken: 2.657 seconds, Fetched: 1 row(s)
```

```
hive> SELECT COUNT(ISNULL(engine_power)) FROM used_cars WHERE engine_power IS NULL;
Query ID = ps21031994_20220722145149_48c478cb-2ee5-4e84-8f89-e5f390718e03
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 1.59 s
--------------------------------------------------------------------------------
OK
554878
Time taken: 2.418 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(fuel_type) FROM used_cars WHERE fuel_type = "";
Query ID = ps21031994_20220722145203_b4ad782b-86f3-4ad5-b238-3800d321b051
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.77 s
--------------------------------------------------------------------------------
OK
1847606
Time taken: 1.682 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(maker) FROM used_cars WHERE maker = "";
Query ID = ps21031994_20220722145118_e55aaf0c-7f4c-4736-929e-d674bf4d25c1
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_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 13.71 s
--------------------------------------------------------------------------------
OK
518915
Time taken: 22.113 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(ISNULL(manufacture_year)) FROM used_cars WHERE manufacture_year IS NULL;
Query ID = ps21031994_20220722145145_bf5fa599-16d3-4911-ba7a-e4365360dfeb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.83 s
--------------------------------------------------------------------------------
OK
370579
Time taken: 1.735 seconds, Fetched: 1 row(s)
```

```
hive> SELECT COUNT(model) FROM used_cars WHERE model = "";
Query ID = ps21031994_20220722145140_44d0cb92-39a3-41c8-b8e7-abc0d816e02f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 1.57 s
--------------------------------------------------------------------------------
OK
1133361
Time taken: 2.451 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(ISNULL(phone_number)) FROM used_cars WHERE phone_number IS NULL;
Query ID = ps21031994_20220722145143_acbf9a1f-dd5f-42fb-9ed0-9fc7f99ff4ab
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 1.31 s
--------------------------------------------------------------------------------
OK
362585
Time taken: 2.307 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(ISNULL(price_eur)) FROM used_cars WHERE price_eur IS NULL;
Query ID = ps21031994_20220722150527_0274f2a9-ae38-47fb-bff0-5873e8af39bd
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_1658444852694_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 13.61 s
--------------------------------------------------------------------------------
OK
1
Time taken: 22.103 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(ISNULL(seat_count)) FROM used_cars WHERE seat_count IS NULL;
Query ID = ps21031994_20220722145201_d9f7ae2e-4510-47e8-95bb-e85588b8a21d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 1.43 s
--------------------------------------------------------------------------------
OK
1287100
Time taken: 2.351 seconds, Fetched: 1 row(s)
```

```
hive> SELECT COUNT(ISNULL(stk_year)) FROM used_cars WHERE stk_year IS NULL;
Query ID = ps21031994_20220722145155_6071acf2-dc86-4fce-bf4e-d5c60a81d0c3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.99 s
--------------------------------------------------------------------------------
OK
3016808
Time taken: 1.867 seconds, Fetched: 1 row(s)
hive> SELECT COUNT(transmission) FROM used_cars WHERE transmission = "";
Query ID = ps21031994_20220722145157_c6f5f0df-dda1-429e-82fb-0e7ab8860403
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.80 s
--------------------------------------------------------------------------------
OK
741630
Time taken: 1.781 seconds, Fetched: 1 row(s)
```


Count of Null Values
NUMBER OF NULL VALUES
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
518915
1133361
362585
370579
743415
554878
1122914
3343411
3016808
741630
1090067
1287100
1847606
0
0
0
maker
model
mileage
manufacture_year
engine_displacement
engine_power
body_type
color_slug
stk_year
transmission
door_count
seat_count
fuel_type
date_created
datelastseen
price_eur
COLUMN NAMES


Percentage of Null Values
0
0
0
14.60533934
31.89948642
10.20528789
10.43028636
20.92409806
15.61755101
31.60544601
94.10337377
84.91083232
20.87385759
30.68093702
36.2266118
52.00256803
maker
model
mileage
manufacture_year
engine_displacement
engine_power
body_type
color_slug
stk_year
transmission
door_count
seat_count
fuel_type
date_created
datelastseen
price_eur

Unique Price Analysis:

Code:

```
SELECT price_eur, COUNT(price_eur) FROM used_cars GROUP BY price_eur
ORDER BY COUNT(price_eur) DESC LIMIT 10;
```

```
hive> SELECT price_eur, COUNT(price_eur) FROM used_cars GROUP BY price_eur ORDER BY COUNT(price_eur) DESC LIMIT 10;
Query ID = ps21031994_20220722150904_e8fe4ee2-ac2c-4c14-af69-88c3477c3f2e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 16.89 s
--------------------------------------------------------------------------------
OK
1295.34 673623
9900.0  6609
10900.0 6497
12900.0 6274
11900.0 6169
8900.0  5935
6900.0  5657
13900.0 5597
4900.0  5557
14900.0 5556
```

Number of Distinct Prices: 227302
The price that occurs the most: 1295.34

Creating Clean Table:

Code:

```
CREATE TABLE IF NOT EXISTS clean_used_cars(
maker STRING,
model STRING,
phone_number INT,
manufacture_year INT,
engine_displacement FLOAT,
engine_power FLOAT,
body_type STRING,
transmission STRING,
door_count INT,
seat_count INT,
date_created TIMESTAMP,
datelastseen TIMESTAMP,
price_eur FLOAT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

```
hive> CREATE TABLE IF NOT EXISTS clean_used_cars(
    > maker STRING,
    > model STRING,
    > phone_number INT,
    > manufacture_year INT,
    > engine_displacement FLOAT,
    > engine_power FLOAT,
    > body_type STRING,
    > transmission STRING,
    > door_count INT,
    > seat_count INT,
    > date_created TIMESTAMP,
    > datelastseen TIMESTAMP,
    > price_eur FLOAT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.04 seconds
```

fuel_type, color_slug & stk_year columns have been dropped due to having greater than 50% missing values

Loading Data into clean_used_cars:

Code:

```
INSERT INTO TABLE clean_used_cars
SELECT maker, model, phone_number, manufacture_year, engine_displacement, engine_power,
body_type, transmission, door_count, seat_count, date_created, datelastseen, price_eur
FROM used_cars
WHERE manufacture_year >= '2000' AND manufacture_year <= '2017' AND maker != "" AND
model != "" AND price_eur != 1295.34;
```

```
hive> INSERT INTO TABLE clean_used_cars
    > SELECT maker, model, phone_number, manufacture_year, engine_displacement, engine_power, body_type, transmission
, door_count, seat_count, date_created, datelastseen, price_eur
    > FROM used_cars
    > WHERE manufacture_year >= '2000' AND manufacture_year <= '2017' AND maker != "" AND model != "" AND price_eur !
= 1295.34;
Query ID = ps21031994_20220722151243_4970ab40-01d6-476d-9f42-6a57043c7f05
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0005)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 27.07 s
----------------------------------------------------------------------------------------------
Loading data to table midtermproject_db.clean_used_cars
OK
Time taken: 28.479 seconds
```

```
SELECT COUNT(*) FROM clean_used_cars;
```

```
hive> SELECT COUNT(*) FROM clean_used_cars;
OK
1534714
Time taken: 2.084 seconds, Fetched: 1 row(s)
```

Market Analysis:

Cars with Highest Average Price:

Code:

```
SELECT maker, model, AVG(price_eur) AS average_price FROM clean_used_cars GROUP BY maker,model SORT BY average_price DESC LIMIT 10;
```

```
hive> SELECT maker, model, AVG(price_eur) AS average_price FROM clean_used_cars GROUP BY maker,model SORT BY average_
price DESC LIMIT 10;
Query ID = ps21031994_20220722151902_7b33f921-d77d-42ff-8c1e-5bb55dc4ceeb
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_1658444852694_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 13.82 s
--------------------------------------------------------------------------------
OK
subaru  impreza 1.5335643785554325E7
citroen berlingo        4902520.086321111
mitsubishi      lancer  397571.8683985986
lamborghini     aventador       361512.9448926972
porsche carrera-gt      302045.21671102336
audi    a5      264466.8896514627
bmw     z8      245118.60092905405
tesla   roadster        192880.27864583334
tesla   model-x 176418.31510416666
bentley brooklands      138501.303125
```

Cars with Lowest Average Price:

Code:

```
SELECT maker, model, AVG(price_eur) AS average_price FROM clean_used_cars GROUP BY maker,model SORT BY average_price ASC LIMIT 10;
```

```
hive> SELECT maker, model, AVG(price_eur) AS average_price FROM clean_used_cars GROUP BY maker,model SORT BY average_
price ASC LIMIT 10;
Query ID = ps21031994_20220722151925_4ec4dc6f-5db5-4be3-84b3-e190d8c9bd58
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 2.38 s
--------------------------------------------------------------------------------
OK
opel    calibra 360.8399963378906
peugeot 405     370.1000061035156
hyundai scoupe  444.1199951171875
toyota  starlet 462.0099983215332
skoda   favorit 536.575286767062
citroen zx      544.6583251953125
fiat    uno     600.3300170898438
ford    scorpio 758.6966756184896
opel    sintra  835.8866746690538
kia     sephia  886.13240234375
Time taken: 3.138 seconds, Fetched: 10 row(s)
hive>
```

Cars in Economic Segment:

Code:

```
SELECT maker, model, AVG(price_eur) FROM clean_used_cars WHERE price_eur >= 3000 AND price_eur < 20000 GROUP BY maker,model ORDER BY AVG(price_eur) LIMIT 5;
```

```
hive> SELECT maker, model, AVG(price_eur) FROM clean_used_cars WHERE price_eur >= 3000 AND price_eur < 20000 GROUP BY
 maker,model ORDER BY AVG(price_eur) LIMIT 5;
Query ID = ps21031994_20220724054424_bf7c51d3-5f7b-4f17-a5a5-42050c30e009
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0020)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.88 s
--------------------------------------------------------------------------------
OK
peugeot 806     3000.0
skoda   galaxy  3071.800048828125
audi    sq5     3145.820068359375
rover   streetwise      3187.8466796875
kia     retona  3200.2542898995534
Time taken: 19.692 seconds, Fetched: 5 row(s)
hive>
    >
```

Cars in Intermediate Segment:

Code:

```
SELECT maker, model, AVG(price_eur) FROM clean_used_cars WHERE price_eur >= 20000 AND price_eur < 300000 GROUP BY maker,model ORDER BY AVG(price_eur) LIMIT 5;
```

```
    > SELECT maker, model, AVG(price_eur) FROM clean_used_cars WHERE price_eur >= 20000 AND price_eur < 300000 GROUP
BY maker,model ORDER BY AVG(price_eur) LIMIT 5;
Query ID = ps21031994_20220724054514_25b500d6-431c-4604-994b-918a3106cca8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0020)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.76 s
--------------------------------------------------------------------------------
OK
honda   legend  20001.9609375
nissan  armada  20001.9609375
skoda   felicia 20170.240234375
subaru  justy   20355.2890625
seat    cordoba 20355.2890625
Time taken: 14.743 seconds, Fetched: 5 row(s)
hive>
```

Cars in Luxury Segment:

Code:

```
SELECT maker, model, AVG(price_eur) FROM clean_used_cars WHERE price_eur >= 300000 AND price_eur <= 2000000 GROUP BY maker,model ORDER BY AVG(price_eur) LIMIT 5;
```

```
hive>
    > SELECT maker, model, AVG(price_eur) FROM clean_used_cars WHERE price_eur >= 300000 AND price_eur <= 2000000 GRO
UP BY maker,model ORDER BY AVG(price_eur) LIMIT 5;
Query ID = ps21031994_20220724195555_2f5b2a9c-58bc-4d54-b378-2f34e4fa876d
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_1658444852694_0023)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.85 s
--------------------------------------------------------------------------------
OK
peugeot 206     300000.0
volkswagen      touareg 303000.0
audi    a8      345678.0
citroen c2      350000.0
bmw     m4      350012.9375
Time taken: 22.975 seconds, Fetched: 5 row(s)
hive>
```

Fastest Selling Cars in Each Segment:

Code:

<u>Economy Segment:</u>

```
SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), maker, model, AVG(price_eur) FROM clean_used_cars WHERE DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) >= 0 AND price_eur >= 3000 AND price_eur < 20000 GROUP BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)),maker,model ORDER BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), AVG(price_eur) LIMIT 10;
```

```
    > SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), maker, model, AVG(price_eur) FROM clean_used_car
s WHERE DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) >= 0 AND price_eur >= 3000 AND price_eur < 20000 GROUP
 BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)),maker,model ORDER BY DATEDIFF(TO_DATE(datelastseen), TO_DA
TE(date_created)), AVG(price_eur) LIMIT 10;
Query ID = ps21031994_20220724195949_2cadfd5f-6fdb-44df-ae5f-039d0c5f1dc1
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0023)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 17.86 s
--------------------------------------------------------------------------------
OK
0       peugeot 806     3000.0
0       ford    gt      3100.330078125
0       audi    sq5     3145.820068359375
0       hyundai galloper        3200.0
0       fiat    seicento        3220.9449462890625
0       mazda   cx-9    3290.159912109375
0       infinity        ex37    3293.860107421875
0       nissan  370-z   3330.8701171875
0       suzuki  wagon-r 3385.123291015625
0       hyundai trajet  3650.092529296875
Time taken: 18.943 seconds, Fetched: 10 row(s)
```

Intermediate Segment:

```
SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), maker, model,
AVG(price_eur) FROM clean_used_cars WHERE DATEDIFF(TO_DATE(datelastseen),
TO_DATE(date_created)) >= 0 AND price_eur >= 20000 AND price_eur < 300000 GROUP BY
DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)),maker,model ORDER BY
DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), AVG(price_eur) LIMIT 10;
```

```
hive> SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), maker, model, AVG(price_eur) FROM clean_used_car
s WHERE DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) >= 0 AND price_eur >= 20000 AND price_eur < 300000 GRO
UP BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)),maker,model ORDER BY DATEDIFF(TO_DATE(datelastseen), TO_
DATE(date_created)), AVG(price_eur) LIMIT 10;
Query ID = ps21031994_20220724200008_c5d4a8a2-6f84-4d83-b082-17cfcea6d3c4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0023)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.88 s
--------------------------------------------------------------------------------
OK
0       peugeot 807     20318.279296875
0       jeep    commander       20319.5390625
0       citroen c-crosser       20490.0
0       renault captur  20632.4921875
0       mazda   cx-3    20666.279296875
0       citroen c1      20721.240234375
0       dodge   durango 20734.310546875
0       peugeot 2008    20830.5359375
0       citroen jumper  21003.76953125
0       subaru  xv      21016.666666666668
Time taken: 7.833 seconds, Fetched: 10 row(s)
```

Luxury Segment:

```
SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), maker, model,
AVG(price_eur) FROM clean_used_cars WHERE DATEDIFF(TO_DATE(datelastseen),
TO_DATE(date_created)) >= 0 AND price_eur >= 300000 AND price_eur < 2000000 GROUP BY
DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)),maker,model ORDER BY
DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), AVG(price_eur) LIMIT 10;
```

```
hive> SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), maker, model, AVG(price_eur) FROM clean_used_car
s WHERE DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) >= 0 AND price_eur >= 300000 AND price_eur < 2000000 G
ROUP BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)),maker,model ORDER BY DATEDIFF(TO_DATE(datelastseen), T
O_DATE(date_created)), AVG(price_eur) LIMIT 10;
Query ID = ps21031994_20220724200030_a914a168-ec84-4ba4-901d-234f59ae2468
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0023)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 17.40 s
--------------------------------------------------------------------------------
OK
0       porsche 911     349538.8125
0       bmw     m4      350012.9375
0       lamborghini     aventador       351577.96875
1       lamborghini     aventador       348829.140625
2       volkswagen      touareg 303000.0
2       lamborghini     aventador       369505.28125
2       skoda   fabia   740266.75
3       audi    a3      400000.0
4       porsche 911     300110.03125
5       ford    gt      385028.5
Time taken: 18.304 seconds, Fetched: 10 row(s)
```

Customer Favourite Cars:

Code:

Economic Segment

```
SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, maker, model, price_eur FROM CLEAN_USED_CARS WHERE DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 3000 AND price_eur < 20000 GROUP BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, maker, model, price_eur ORDER BY engine_power DESC LIMIT 100;
```

```
hive> SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, maker, model, price_eur FROM CLEAN
_USED_CARS WHERE DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 3000 AND price_eur < 200
00 GROUP BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, maker, model, price_eur ORDER BY en
gine_power DESC LIMIT 100;
Query ID = ps21031994_20220724200825_95174981-e3d4-4d9a-9f46-c0d845e551c0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0024)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 18.75 s
--------------------------------------------------------------------------------
OK
0       493.0   ford    mustang 10140.64
0       408.0   bmw     x6      4034.05
0       397.0   lancia  delta   13900.52
0       373.0   bmw     m5      18507.81
0       368.0   porsche cayenne 17902.0
0       368.0   porsche 911     14766.84
0       357.0   nissan  gt-r    3700.93
0       338.0   volvo   460     15500.0
0       331.0   porsche cayenne 11435.97
0       331.0   audi    a8      9800.37
0       331.0   audi    a8      12951.44
```

Intermediate Segment:

SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, maker, model, price_eur FROM CLEAN_USED_CARS WHERE DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 20000 AND price_eur < 300000 GROUP BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, maker, model, price_eur ORDER BY engine_power DESC LIMIT 100;

```
hive> SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, maker, model, price_eur FROM CLEAN
_USED_CARS WHERE DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 20000 AND price_eur < 30
0000 GROUP BY DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, maker, model, price_eur ORDER BY
engine_power DESC LIMIT 100;
Query ID = ps21031994_20220724201049_c5c517bb-15d0-4978-b2f0-f35cf9d72fd0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0024)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 17.55 s
--------------------------------------------------------------------------------
OK
0       568.0   tesla   model-x 210202.06
0       465.0   rolls-royce     wraith  76980.01
0       464.0   nissan  gt-r    89009.88
0       456.0   bentley continental-gtc 80026.94
0       449.0   audi    r8      131202.11
0       449.0   bentley continental-gt  62107.88
0       445.0   audi    s8      125970.65
0       445.0   audi    rs6     100012.14
0       441.0   hummer  h2      59950.0
0       432.0   chevrolet       camaro  47946.08
```

Price Determination:

Economic Segment:

Code:

```
SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, engine_displacement, manufacture_year, maker, price_eur FROM clean_used_cars WHERE engine_power IS NOT NULL AND engine_displacement IS NOT NULL AND DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 3000 AND price_eur <= 20000 ORDER BY price_eur ASC LIMIT 100;
```

```
hive> SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, engine_displacement, manufacture_y
ear, maker, price_eur FROM clean_used_cars WHERE engine_power IS NOT NULL AND engine_displacement IS NOT NULL AND DAT
EDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 3000 AND price_eur <= 20000 ORDER BY price_e
ur ASC LIMIT 100;
Query ID = ps21031994_20220724201318_7c9fe1f3-04c3-4284-a1c7-db6a55036570
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0024)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 17.39 s
--------------------------------------------------------------------------------
OK
0       101.0   1997.0  2000    peugeot 3000.0
0       59.0    1229.0  2007    opel    3000.0
0       48.0    1461.0  2005    renault 3000.0
0       74.0    1753.0  2003    ford    3000.0
0       74.0    1896.0  2003    volkswagen      3000.0
0       100.0   1997.0  2002    peugeot 3000.0
0       57.0    1368.0  2008    fiat    3000.0
0       44.0    1124.0  2003    citroen 3000.0
0       70.0    1493.0  2006    smart   3000.0
0       66.0    1398.0  2005    citroen 3000.0
0       40.0    1108.0  2001    fiat    3000.0
0       85.0    1896.0  2004    ford    3000.0
0       44.0    1242.0  2003    fiat    3000.0
```

Intermediate Segment:

Code:

```
SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, engine_displacement, manufacture_year, maker, price_eur FROM clean_used_cars WHERE engine_power IS NOT NULL AND engine_displacement IS NOT NULL AND DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 20000 AND price_eur <= 300000 ORDER BY price_eur ASC LIMIT 100;
```

```
hive> SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, engine_displacement, manufacture_y
ear, maker, price_eur FROM clean_used_cars WHERE engine_power IS NOT NULL AND engine_displacement IS NOT NULL AND DAT
EDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 20000 AND price_eur <= 300000 ORDER BY price
_eur ASC LIMIT 100;
Query ID = ps21031994_20220724201337_c9a48c97-b9e8-454a-9009-443cc4f4b40c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0024)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.91 s
--------------------------------------------------------------------------------
OK
0       77.0    1598.0  2013    audi    20000.0
0       147.0   1984.0  2012    volkswagen      20000.0
0       155.0   1598.0  2012    mini    20000.0
0       177.0   1984.0  2007    seat    20000.74
0       130.0   1995.0  2011    bmw     20000.74
0       171.0   2400.0  2010    volvo   20001.22
0       118.0   1798.0  2010    seat    20001.22
0       100.0   1995.0  2011    kia     20002.22
0       135.0   1995.0  2012    bmw     20002.22
0       103.0   1364.0  2016    opel    20002.22
0       120.0   2143.0  2014    jeep    20002.22
0       63.0    1197.0  2016    volkswagen      20002.22
0       135.0   1968.0  2015    seat    20002.22
```

Luxury Segment:

Code:

```
SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power,
engine_displacement, manufacture_year, maker, price_eur FROM clean_used_cars WHERE
engine_power IS NOT NULL AND engine_displacement IS NOT NULL AND
DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) = 0 AND price_eur >= 300000 AND
price_eur <= 2000000 ORDER BY price_eur ASC LIMIT 100;
```

```
hive> SELECT DATEDIFF(TO_DATE(datelastseen), TO_DATE(date_created)), engine_power, engine_displacement, manufacture_y
ear, maker, price_eur FROM clean_used_cars WHERE engine_power IS NOT NULL AND engine_displacement IS NOT NULL AND DAT
EDIFF(TO_DATE(datelastseen), TO_DATE(date_created)) >= 0 AND price_eur >= 300000 AND price_eur <= 2000000 ORDER BY pr
ice_eur ASC LIMIT 100;
Query ID = ps21031994_20220724201805_0a248ba1-474d-4203-8574-1fe7fb454bc3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1658444852694_0024)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.04 s
--------------------------------------------------------------------------------
OK
4       368.0   3996.0  2015    porsche 300110.03
364     515.0   6498.0  2014    lamborghini     302439.75
366     515.0   6498.0  2013    lamborghini     306745.72
1       515.0   6498.0  2016    lamborghini     328166.66
8       515.0   6498.0  2015    lamborghini     345025.53
0       280.0   3600.0  2004    porsche 349538.8
0       368.0   2979.0  2016    bmw     350012.94
0       515.0   6498.0  2015    lamborghini     351577.97
372     294.0   4941.0  2002    bmw     365054.03
381     410.0   5400.0  2005    ford    369027.3
5       529.0   5409.0  2005    ford    385028.5
363     456.0   3600.0  2010    porsche 399147.66
3       81.0    1598.0  2014    audi    400000.0
```