Execution Screenshots

1. Data Loading

Create database and table (table named 'used_cars_yichen')

```
hive> CREATE DATABASE cardata;
OK
Time taken: 0.81 seconds
hive> USE cardata;
OK
Time taken: 0.058 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS used_cars_yichen (
    > maker STRING,
    > model STRING,
    > mileage INT,
    > manufacture_year STRING,
    > engine_displacement INT,
    > engine_power INT,
    > body_type STRING,
    > color_slug STRING,
    > stk_year STRING,
    > transmission STRING,
    > door_count INT,
    > seat_count INT,
    > fuel_type STRING,
    > date_created STRING,
    > datelastseen STRING,
    > price_eur FLOAT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/BigData/'
    > TBLPROPERTIES ('skip.header.line.count'='1','serialization.null.format' = '');
OK
Time taken: 0.473 seconds
```

```
hive> DESCRIBE FORMATTED used_cars_yichen;
OK
# col_name              data_type               comment
maker                   string
model                   string
mileage                 int
manufacture_year        string
engine_displacement     int
engine_power            int
body_type               string
color_slug              string
stk_year                string
transmission            string
door_count              int
seat_count              int
fuel_type               string
date_created            string
datelastseen            string
price_eur               float

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sun Nov 06 03:16:26 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/BigData
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        bucketing_version       2
        numFiles                2
        totalSize               847689511
        transient_lastDdlTime   1667704586

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             ,
        serialization.format    ,
Time taken: 0.135 seconds, Fetched: 45 row(s)
```

2. Examine missing values

Cleaning 2-1. Count missing value for each attribute

```
hive> SELECT count(*) from used_cars_yichen where maker is NULL;
Query ID = ychsiaoca_20221106164408_1a9e6325-7f2f-4fe9-b596-82e14dac9b60
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0001)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.32 s
--------------------------------------------------------------------------------
OK
518915
Time taken: 8.457 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where model is NULL;
Query ID = ychsiaoca_20221106165006_934ee7c3-dfb5-4a7f-a1fb-14443ad47172
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_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.46 s
--------------------------------------------------------------------------------
OK
1133361
Time taken: 16.7 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where mileage is NULL;
Query ID = ychsiaoca_20221106165049_49866448-1074-4336-9fe4-57fa571204b2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.25 s
--------------------------------------------------------------------------------
OK
362584
Time taken: 9.247 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where manufacture_year is NULL;
Query ID = ychsiaoca_20221106165140_51de1125-6455-4cff-baec-359c3f0039a5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.43 s
--------------------------------------------------------------------------------
OK
370578
Time taken: 8.393 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where engine_displacement is NULL;
Query ID = ychsiaoca_20221106165235_26b6ebfe-10a2-4505-9f12-743a48da3b04
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.55 s
--------------------------------------------------------------------------------
OK
743414
Time taken: 9.398 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where engine_power is NULL;
Query ID = ychsiaoca_20221106165334_25556506-f422-4669-948d-96724a66bebc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.91 s
--------------------------------------------------------------------------------
OK
554877
Time taken: 7.754 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where body_type is NULL;
Query ID = ychsiaoca_20221106165442_6728b369-fd2f-40fc-ba0d-50e7d386a1b8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.85 s
--------------------------------------------------------------------------------
OK
1122914
Time taken: 7.731 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where color_slug is NULL;
Query ID = ychsiaoca_20221106165541_e53f75db-d52e-4f7d-ae21-1627964ebdb0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.32 s
--------------------------------------------------------------------------------
OK
3343411
Time taken: 8.167 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where stk_year is NULL;
Query ID = ychsiaoca_20221106165635_541ab097-caea-4a65-929c-e9342a0134a3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.57 s
--------------------------------------------------------------------------------
OK
1708156
Time taken: 8.472 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where transmission is NULL;
Query ID = ychsiaoca_20221106165827_7087de65-cd54-4429-a002-223ca9a0e88b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.39 s
--------------------------------------------------------------------------------
OK
741630
Time taken: 9.231 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where door_count is NULL;
Query ID = ychsiaoca_20221106165851_31f61e34-952c-4b02-9e24-af2dbc9e0ce8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.61 s
--------------------------------------------------------------------------------
OK
1090066
Time taken: 8.48 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where seat_count is NULL;
Query ID = ychsiaoca_20221106165952_4a7b3388-576c-4251-b10a-8045d20ad842
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.16 s
--------------------------------------------------------------------------------
OK
1287099
Time taken: 9.196 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where fuel type is NULL;
Query ID = ychsiaoca_20221106170048_96a6294d-b008-45fc-99d0-b6241bdce3b0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.22 s
--------------------------------------------------------------------------------
OK
1847606
Time taken: 9.017 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where date_created is NULL;
Query ID = ychsiaoca_20221106170237_de4021ea-a0fe-4105-8c1c-6d18ff9979c4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.45 s
--------------------------------------------------------------------------------
OK
0
Time taken: 9.292 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where datelastseen is NULL;
Query ID = ychsiaoca_20221106170402_814bc3a7-6c7f-472f-9bd8-173d487cfcad
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.57 s
--------------------------------------------------------------------------------
OK
0
Time taken: 8.4 seconds, Fetched: 1 row(s)
```

```
hive> SELECT count(*) from used_cars_yichen where price_eur is NULL;
Query ID = ychsiaoca_20221106170428_e750cb24-d184-4eb0-ac50-66643e79cfb6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667752617512_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.77 s
--------------------------------------------------------------------------------
OK
0
Time taken: 9.609 seconds, Fetched: 1 row(s)
```

2-2. Remove any records that do not have a price

```
hive> CREATE TABLE used_cars_yichen_clean_2_2
    > AS SELECT *FROM used_cars_yichen
    > WHERE price_eur is not NULL;
Query ID = ychsiaoca_20221115003411_97f384a1-11a3-4a50-9af5-85b729524ebd
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_1668463771091_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 22.03 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_2_2
OK
Time taken: 30.211 seconds
```

- Note:
  - There is no row without a price.
  - Number of rows: 3552912 (the same with the number before cleaning)

2-3. Remove any records that do not have a model listed

```
hive> CREATE TABLE used_cars_yichen_clean_2_3
    > AS SELECT *FROM used_cars_yichen_clean_2_2
    > WHERE model is not NULL;
Query ID = ychsiaoca_20221115010454_342c99a8-ad71-43fe-8f4f-e16ddf87ee2d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668463771091_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 20.90 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_2_3
OK
Time taken: 25.259 seconds
```

- Note:
  - Number of rows without model: 1133361
  - Number of rows: 2419551 (=3552912 - 1133361)

3. Remove records with abnormally repetitive prices

3-1. Group the price column and count the number of unique prices.

```
hive> CREATE TABLE used_cars_yichen_clean_3_1
    > AS SELECT price_eur FROM  used_cars_yichen_clean_2_3
    > GROUP BY price_eur;
Query ID = ychsiaoca_20221115010937_53e0d281-071d-448b-93bc-65699bcaa871
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668463771091_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.03 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_3_1
OK
Time taken: 16.47 seconds
```

```
hive> DESCRIBE FORMATTED used_cars_yichen_clean_3_1;
OK
# col_name              data_type               comment
price_eur               float

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Tue Nov 15 01:09:53 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_3_1
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                2
        numRows                 167243
        rawDataSize             1261211
        totalSize               1428454
        transient_lastDdlTime   1668474593

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.071 seconds, Fetched: 31 row(s)
```

- Note: the number of unique prices: 167243

3-2. Remove records with abnormally repetitive prices

```
hive> CREATE TABLE used_cars_yichen_clean_3_2
    > AS SELECT price_eur, COUNT (price_eur) count_of_price  FROM  used_cars_yichen_clean_2_3
    > GROUP BY price_eur;
Query ID = ychsiaoca_20221115011249_f082dced-6acc-4040-822f-b73711f68a00
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668463771091_0009)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.66 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_3_2
OK
Time taken: 15.877 seconds
```

```
hive> SELECT price_eur, count_of_price FROM used_cars_yichen_clean_3_2
    > ORDER BY count_of_price DESC
    > LIMIT 10;
Query ID = ychsiaoca_20221115011413_b73ac2ba-9f8c-40dc-99b6-84f98c19521b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668463771091_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.65 s
--------------------------------------------------------------------------------
OK
1295.34 574753
9900.0  5298
12900.0 5002
10900.0 4954
11900.0 4807
4900.0  4590
8900.0  4481
5900.0  4254
4500.0  4146
3500.0  4146
Time taken: 7.526 seconds, Fetched: 10 row(s)
```

- Note: price '1295.34' repeat 574753 times

Remove rows with a price '1295.34'

```
hive> CREATE TABLE used_cars_yichen_clean_3_2_a
    > AS SELECT * FROM  used_cars_yichen_clean_2_3
    > WHERE price_eur != 1295.34;
Query ID = ychsiaoca_20221115011633_e739520b-f13b-46a4-a462-bde391704738
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668463771091_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 17.55 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_3_2_a
OK
Time taken: 18.907 seconds
```

- Note: Number of rows: 1844798 (=2419551-574753)

4. Fill in the maker value to complete the record

Check NULL values of 'maker'

```
hive> SELECT count(*) from used_cars_yichen_clean_3_2_a where maker is NULL;
Query ID = ychsiaoca_20221115012511_2d5d6059-6f76-4694-8646-80f84c6e6cd9
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668463771091_0010)

--------------------------------------------------------------------------------
        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: 11.66 s
--------------------------------------------------------------------------------
OK
0
Time taken: 12.536 seconds, Fetched: 1 row(s)
```

- Note: There is no NULL values of 'maker'

6. Remove records which are abnormal or cannot be trusted

6-1. Exclude the unreasonable prices according to market research and exclude the outliers according to the distribution.

Calculate the average and standard deviation of price by model_maker

```
hive> CREATE TABLE average_price_by_model_maker
    > AS SELECT model, maker, AVG(price_eur) avg_price ,STDDEV(price_eur) sd_price FROM
  used_cars_yichen_clean_3_2_a
    > GROUP BY model, maker;
Query ID = ychsiaoca_20221118041935_78a1bee2-ff21-4812-9f98-06ef21eda6ab
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668463771091_0024)

--------------------------------------------------------------------------------
-------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  K
ILLED
--------------------------------------------------------------------------------
-------
Map 1            container        INITED      5          0        0        5       0
--------------------------------------------------------------------------------
------- 2        container        INITED      1          0        0        1       0
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  K
ILLED  --------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------S: 00/02  [>>-------------------------] 0%    ELAPSED TIME: 0.00 s
Map 1            container        INITED      5          0        0        5       0
--------------------------------------------------------------------------------
------- 2        container        INITED      1          0        0        1       0
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  K
ILLED  --------------------------------------------------------------------------------
--------------------------------------------------------------------------------
-------S: 00/02  [>>-------------------------] 0%    ELAPSED TIME: 0.51 s
-------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/average_price_
by_model_maker
OK
Time taken: 16.009 seconds
```

```
hive> DESCRIBE FORMATTED average_price_by_model_maker;
OK
# col_name              data_type               comment
model                   string
maker                   string
avg_price               double
sd_price                double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Fri Nov 18 04:19:51 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/user/hive/warehouse/cardata.db/average_price_b
y_model_maker
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                1
        numRows                 826
        rawDataSize             40059
        totalSize               40885
        transient_lastDdlTime   1668745191

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.065 seconds, Fetched: 34 row(s)
```

Show the top 10 rows with highest averages

```
hive> SELECT * FROM average_price_by_model_maker
    > ORDER BY avg_price DESC
    > LIMIT 10;
Query ID = ychsiaoca_20221118050946_1105897e-98a3-4f7f-8ab2-9dad6945ea69
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_1668463771091_0027)

Map 1 .......... container     SUCCEEDED      1          1        0        0       0
    0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0
    0
--------------------------------------------------------------------------------
-------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.14 s
--------------------------------------------------------------------------------
-------
OK
kangoo  renault 7.009004778337162E8     4.3545727108604965E10
impreza subaru  1.0991529486164475E7    4.030591851533493E8
berlingo        citroen 4248744.7203723835      2.5152784926156342E8
v8      audi    588181.6242049324       3839765.68180293
aventador       lamborghini     368228.04374951654      140258.30912585667
lancer  mitsubishi      305817.8635682135       7774780.251419145
carrera-gt      porsche 272783.0702237216       347487.460756318
a5      audi    255972.56859279692      1.3655389152282294E7
z8      bmw     245118.60092905405      61622.247922119735
xm      citroen 200067.5044852933       2304741.5361049357
Time taken: 14.411 seconds, Fetched: 10 row(s)
```

- Note: According to the average and standard deviation list above, the average and standard deviation are biased by serious outliers.

Therefore, this analysis will first exclude the unreasonable prices according to market research and exclude the outliers according to the distribution.

Reasonable price range: **$1331.28 euro ~ $35766.07 euro**

-- MAX: high end car, 2021, very low mileage (<1,000)

For example: 2021 Mercedes-Benz A-Class, mileage=999, $37,075 = **$35766.07 euro**

--MIN: Most affordable car, 2000, very high mileage (400,000)

For example: 2000 Toyota 4Runner, mileage=400,000, $1380 = **$1331.28 euro**

(Reference: https://www.consumerreports.org/cars/car-value-estimator/)

```
hive> CREATE TABLE used_cars_yichen_clean_6_1
    > AS SELECT * FROM used_cars_yichen_clean_3_2_a
    > WHERE 1331.28 < price_eur AND price_eur < 35766.07;
Query ID = ychsiaoca_20221119042642_bda77b85-a527-4003-acd4-0698666203b2
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_1668829286622_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 18.22 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_6
_1
OK
Time taken: 26.593 seconds
```

```
hive> DESCRIBE FORMATTED used_cars_yichen_clean_6_1;
OK
# col_name              data_type               comment
maker                   string
model                   string
mileage                 int
manufacture_year        string
engine_displacement     int
engine_power            int
body_type               string
color_slug              string
stk_year                string
transmission            string
door_count              int
seat_count              int
fuel_type               string
date_created            string
datelastseen            string
price_eur               float

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 19 04:27:09 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_6_
1
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                5
        numRows                 1629858
        rawDataSize             201294278
        totalSize               202924136
        transient_lastDdlTime   1668832029

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.072 seconds, Fetched: 46 row(s)
```

- Note:
  - There are 1629858 rows in 'used_cars_yichen_clean_6_1
  - There are 214,940 rows with unreasonable prices are excluded (1844798 - 1629858)

Examine the average and standard deviation again

```
Time taken: 0.081 seconds, Fetched: 34 row(s)
hive> SELECT * FROM average_price_by_model_maker_2
    > ORDER BY avg_price DESC
    > LIMIT 10;
Query ID = ychsiaoca_20221118060115_ed641626-0e5c-46cc-9952-6627868c4cab
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668463771091_0029)

--------------------------------------------------------------------------------
-------S: 01/02  [==============>>-------------] 50%   ELAPSED TIME: 5.05 s
Map 1 .......... container     SUCCEEDED      1          1        0        0       0
    0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0
    0
--------------------------------------------------------------------------------
-------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.27 s
--------------------------------------------------------------------------------
-------
OK
q70     infinity        33841.746744791664      1704.098946630235
continental-flying-spur bentley 33350.037760416664      1647.619011733293
qx70    infinity        32700.252734375 966.3578728525526
continental-gt  bentley 32396.80234375  1583.2688725579544
h1      hummer  31669.272194602272      4242.638410159205
i3      bmw     29990.86916613223       5404.980436169807
qx56    infinity        29925.0 3075.0
q45     infinity        29900.0 0.0
rs3     audi    29855.20176003196       5809.4125171969035
lx-570  lexus   29454.5498046875        45.4501953125
Time taken: 6.164 seconds, Fetched: 10 row(s)
```

- The average and standard deviation of prices look reasonable now.

6-2. Exclude rows with unreasonable manufacture year

Examine manufacture year

```
hive> SELECT manufacture_year FROM used_cars_yichen_clean_6_1
    > GROUP BY manufacture_year
    > ORDER BY manufacture_year DESC
    > LIMIT 10;
Query ID = ychsiaoca_20221119043610_c129f5e2-7a6c-4059-8a8f-392505f6a50c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.06 s
--------------------------------------------------------------------------------
OK
992
991
990
99
988
974
965
960
959
958
Time taken: 14.006 seconds, Fetched: 10 row(s)
```

```
hive> SELECT manufacture_year FROM used_cars_yichen_clean_6_1
    > GROUP BY manufacture_year
    > ORDER BY manufacture_year
    > LIMIT 10;
Query ID = ychsiaoca_20221119043641_31337cd1-4cb9-4db8-bd7a-c6672db31ba5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.27 s
--------------------------------------------------------------------------------
OK
NULL
0
10
1000
1001
1003
1006
1007
1009
1010
Time taken: 14.15 seconds, Fetched: 10 row(s)
```

Exclude the unreasonable manufacture year

- No early than 1908 - the year first car came out; no later than the year when the ad posted)
  (Reference: https://www.carsguide.com.au/car-advice/who-invented-the-first-car-and-when-was-it-made-76976)

```
hive> CREATE TABLE used_cars_yichen_clean_6_2
    > AS SELECT * FROM  used_cars_yichen_clean_6_1
    > WHERE manufacture_year >= 1908 AND manufacture_year <= year(date_created);
Query ID = ychsiaoca_20221119045111_1cb8c5d8-bf31-475b-bd84-f6965d5a6ea8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 18.58 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_6
_2
OK
Time taken: 19.695 seconds
```

```
hive> DESCRIBE FORMATTED used_cars_yichen_clean_6_2;
OK
# col_name              data_type               comment
maker                   string
model                   string
mileage                 int
manufacture_year        string
engine_displacement     int
engine_power            int
body_type               string
color_slug              string
stk_year                string
transmission            string
door_count              int
seat_count              int
fuel_type               string
date_created            string
datelastseen            string
price_eur               float
```

```
# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 19 04:51:31 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_6_
2
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                5
        numRows                 1483586
        rawDataSize             183815219
        totalSize               185298805
        transient_lastDdlTime   1668833491

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.07 seconds, Fetched: 46 row(s)
```

- Note:
  - There are 1483586 rows in 'used_cars_yichen_clean_6_2'
  - 146,272 rows with a manufature_year < 1908 were excluded (1629858 - 1483586)

6-3. Exclude rows with unreasonable milage

Examine mileage

```
hive> SELECT mileage FROM used_cars_yichen_clean_6_2
    > GROUP BY mileage
    > ORDER BY mileage
    > LIMIT 10;
Query ID = ychsiaoca_20221119045812_44c05f62-fb41-4a97-a5c6-c29e939faf75
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_1668829286622_0006)

--------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 14.60 s
--------------------------------------------------------------------------------
OK
NULL
0
1
2
3
4
5
6
7
8
Time taken: 23.192 seconds, Fetched: 10 row(s)
```

```
hive> SELECT mileage FROM used_cars_yichen_clean_6_2
    > GROUP BY mileage
    > ORDER BY mileage DESC
    > LIMIT 10;
Query ID = ychsiaoca_20221119045951_77e07152-0858-44ee-a71b-4ccbbd1ebb9d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0006)

--------------------------------------------------------------------------------
        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
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 13.65 s
--------------------------------------------------------------------------------
OK
9999999
9899800
9754500
9635898
9370350
9352000
9309100
9294000
9288000
9284000
Time taken: 14.541 seconds, Fetched: 10 row(s)
```

Exclude rows with unreasonable milage

The mileage increases 10,000 miles (16,093 KM) to 12,000 (19,312 KM) per year for regular-use cars

(Reference: https://www.progressive.com/answers/used-car-mileage/).

Note: The unit of mileage in this car dataset is KM., 1 mile = 1.60934 KM.

Min: Assume the used cars in this dataset have been used at least one year before putting onto ads, this analysis deemed 'mileage < 16,093*0.5 KM' as abnormal mileage.

Max: this analysis deemed 'mileage > (year of ad – year of manufacture)* 19,312 *2 KM' as abnormal mileage.

```
hive> CREATE TABLE used_cars_yichen_clean_6_3
    > AS SELECT * FROM  used_cars_yichen_clean_6_2
    > WHERE mileage >= 8047 AND mileage < (year(date_created)-manufacture_year)*19312*2;
Query ID = ychsiaoca_20221119061018_d21d3c28-664e-434a-850d-404a7dbbb8e7
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_1668829286622_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      5          5        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 17.94 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_6
_3
OK
Time taken: 26.701 seconds
```

```
hive> DESCRIBE FORMATTED used_cars_yichen_clean_6_3;
OK
# col_name              data_type               comment
maker                   string
model                   string
mileage                 int
manufacture_year        string
engine_displacement     int
engine_power            int
body_type               string
color_slug              string
stk_year                string
transmission            string
door_count              int
seat_count              int
fuel_type               string
date_created            string
datelastseen            string
price_eur               float
# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 19 06:10:45 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_6_
3
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                5
        numRows                 1129441
        rawDataSize             140324326
        totalSize               141453767
        transient_lastDdlTime   1668838245

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.078 seconds, Fetched: 46 row(s)
```

- Note:
  - There are 1129441 rows in 'used_cars_yichen_clean_6_3'
  - 354145 rows with abnormal mileages were excluded (1483586 - 1129441)

Examine mileage again

```
hive> SELECT mileage FROM used_cars_yichen_clean_6_3
    > GROUP BY mileage
    > ORDER BY mileage
    > LIMIT 10;
Query ID = ychsiaoca_20221119061453_be223423-1e16-4e60-9cf1-f65b0c91643d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0010)

--------------------------------------------------------------------------------
        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: 12.46 s
--------------------------------------------------------------------------------
OK
8047
8048
8049
8050
8051
8052
8053
8054
8055
8056
Time taken: 13.409 seconds, Fetched: 10 row(s)
```

```
hive> SELECT mileage FROM used_cars_yichen_clean_6_3
    > GROUP BY mileage
    > ORDER BY mileage DESC
    > LIMIT 10;
Query ID = ychsiaoca_20221119061547_34efa1fe-afa9-494d-ba04-5098655b80cc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0010)

--------------------------------------------------------------------------------
        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: 13.05 s
--------------------------------------------------------------------------------
OK
1940000
1234567
1129000
1111111
1000000
999999
997652
990000
970000
933534
Time taken: 13.973 seconds, Fetched: 10 row(s)
```

5. Find the average price for cars of different models and makers. Then write queries that will remove any records where prices are multiple factors above this price.

The main factors that influence used car prices include model, maker, manufacture year, mileage.

5-1. Calculates the average and standard deviation of price by 4 factors (model*maker*manufacture year*mileage_level)

First, this analysis created a new column named 'mileage_level'.

100,000 miles (160,934 KM) is considered a cut-off point for used cars

- WHEN mileage >= 160,934 KM THEN 'High_mile'
- ELSE 'Low_mile'

(Reference: https://www.progressive.com/answers/used-car-mileage/)

```
hive> CREATE TABLE used_cars_yichen_clean_5_1
    > AS SELECT *,
    > CASE
    > WHEN mileage >= 160934  THEN 'High_mile'
    > ELSE 'Low_mile'
    > END AS mileage_level
    > FROM used_cars_yichen_clean_6_3;
Query ID = ychsiaoca_20221119061854_10c34117-f3e4-4842-b32e-74257cc33ff4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 15.67 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_5
_1
OK
Time taken: 16.665 seconds
```

- Note: There are 222527 'High_mile'; 906914 'Low_mile'.

Second, this analysis calculates the average and standard deviation of price by model*maker*manufacture year*mileage_level

```
hive> CREATE TABLE average_price_by_4_factors
    > AS SELECT model, maker, manufacture_year, mileage_level, AVG(price_eur) avg_price ,STDDEV(p
rice_eur) sd_price
    > FROM  used_cars_yichen_clean_5_1
    > GROUP BY model, maker, manufacture_year, mileage_level;
Query ID = ychsiaoca_20221119062257_a6bdc578-bd20-423e-a646-787016d01272
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0010)

----------------------------------------------------------------------------------------------
        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: 14.13 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/average_price_by_4_facto
rs
OK
Time taken: 15.256 seconds
```

```
hive> DESCRIBE FORMATTED average_price_by_4_factors;
OK
# col_name              data_type               comment
model                   string
maker                   string
manufacture_year        string
mileage_level           string
avg_price               double
sd_price                double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 19 06:23:12 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/user/hive/warehouse/cardata.db/average_price_by_4_factor
s
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                1
        numRows                 14084
        rawDataSize             822965
        totalSize               837049
        transient_lastDdlTime   1668838992

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.06 seconds, Fetched: 36 row(s)
```

```
hive> SELECT * FROM average_price_by_4_factors
    > LIMIT 20;
Query ID = ychsiaoca_20221119062531_0ea7d779-d5c9-490c-97fa-9550cdf1f35e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.38 s
--------------------------------------------------------------------------------
OK
100     audi    1961    Low_mile        12504.6298828125        0.0
100     audi    1963    Low_mile        14977.2001953125        0.0
100     audi    1967    Low_mile        3902.43994140625        0.0
100     audi    1971    Low_mile        6878.1900634765625      2815.9608848854577
100     audi    1973    Low_mile        18002.2640625   5687.37998857897
100     audi    1974    Low_mile        13250.47998046875       249.52001953125
100     audi    1975    Low_mile        5351.925048828125       651.405029296875
100     audi    1977    Low_mile        3214.1966959635415      1261.40292031476
100     audi    1980    High_mile       2250.070068359375       0.0
100     audi    1980    Low_mile        8251.869873046875       250.649658203125
100     audi    1982    High_mile       2967.126668294271       1376.836844693956
100     audi    1982    Low_mile        8004.836751302083       4.574924748962229
100     audi    1983    High_mile       3818.4500732421875      2116.1900634765625
100     audi    1983    Low_mile        2925.106689453125       676.8114490864003
100     audi    1984    Low_mile        2394.5550537109375      943.7449951171875
100     audi    1985    High_mile       3550.1224975585938      1231.296097613271
100     audi    1985    Low_mile        3812.4312438964844      505.9469133417928
100     audi    1986    High_mile       2321.745675223214       1274.3088322927408
100     audi    1986    Low_mile        3295.1875       1609.2149811223949
100     audi    1987    High_mile       4023.933807373047       2902.389939250223
Time taken: 6.16 seconds, Fetched: 20 row(s)
```

5-2. Join average and std to the table which will be used in next step

- Left join
  --left table: used_cars_yichen_clean_5_1
  --Right table: average_price_by_4_factors

```
hive> CREATE TABLE used_cars_yichen_clean_5_2
    > AS SELECT L.*, R.avg_price, R.sd_price
    > FROM used_cars_yichen_clean_5_1 L LEFT OUTER JOIN average_price_by_4_factors R
    > ON L.model = R.model AND L.maker = R.maker AND L.manufacture_year = R.manufacture_year AND
L.mileage_level = R.mileage_level;
Query ID = ychsiaoca_20221119062648_834d6aa2-2a5c-4b16-bd02-88dcc3335951
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0010)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 20.76 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_5
_2
OK
Time taken: 22.099 seconds
```

```
hive> DESCRIBE FORMATTED used_cars_yichen_clean_5_2;
OK
# col_name              data_type               comment
maker                   string
model                   string
mileage                 int
manufacture_year        string
engine_displacement     int
engine_power            int
body_type               string
color_slug              string
stk_year                string
transmission            string
door_count              int
seat_count              int
fuel_type               string
date_created            string
datelastseen            string
price_eur               float
mileage_level           string
avg_price               double
sd_price                double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 19 06:27:10 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_5_
2
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                4
        numRows                 1129441
        rawDataSize             192258303
        totalSize               193387744
        transient_lastDdlTime   1668839230

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.077 seconds, Fetched: 49 row(s)
```

5-3. Uses the calculated averages and standard deviations of prices to set up the rules to exclude outliers from the dataset.

Select rows that the ‘price_eur’ are between (avg + 3*std) and (avg - 3*std)

```
hive> CREATE TABLE used_cars_yichen_clean_5_3
    > AS SELECT * FROM used_cars_yichen_clean_5_2
    > WHERE price_eur > (avg_price - 3*sd_price ) AND price_eur < (avg_price + 3*sd_price );
Query ID = ychsiaoca_20221119063047_c27c219c-bf20-4b0d-b5ba-546ab16d336b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1668829286622_0010)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 19.26 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_5
_3
OK
Time taken: 20.274 seconds
```

```
hive>
    > DESCRIBE FORMATTED used_cars_yichen_clean_5_3;
OK
# col_name              data_type               comment
maker                   string
model                   string
mileage                 int
manufacture_year        string
engine_displacement     int
engine_power            int
body_type               string
color_slug              string
stk_year                string
transmission            string
door_count              int
seat_count              int
fuel_type               string
date_created            string
datelastseen            string
price_eur               float
mileage_level           string
avg_price               double
sd_price                double
```

```
# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 19 06:31:07 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://bigdata-m/user/hive/warehouse/cardata.db/used_cars_yichen_clean_5_
3
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                4
        numRows                 1114112
        rawDataSize             189712717
        totalSize               190826829
        transient_lastDdlTime   1668839467

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.08 seconds, Fetched: 49 row(s)
```

Data Analysis Execution Records

Part I. A snapshot of Luxury used car ads

- Car details group by model_maker

Group rows by model_maker, and calculate the quantity, average price, average mileage, average engine displacement.

```
hive> CREATE TABLE model_maker_info
    > AS SELECT model, maker, COUNT(CONCAT(model, maker)) model_maker_count,
    > AVG(price_eur), AVG(mileage), AVG(engine_displacement)
    > FROM  used_cars_yichen_clean_5_3
    > GROUP BY model, maker;
Query ID = ychsiaoca_20221126060604_290494d0-1b27-4f27-a019-a5d5b0fd1d89
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_1669439444647_0003)

--------------------------------------------------------------------------------
        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: 12.75 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info
OK
Time taken: 21.245 seconds
```

```
hive> DESCRIBE FORMATTED model_maker_info;
OK
# col_name              data_type               comment
model                   string
maker                   string
model_maker_count       bigint
_c3                     double
_c4                     double
_c5                     double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 26 06:06:25 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                1
        numRows                 711
        rawDataSize             48075
        totalSize               48786
        transient_lastDdlTime   1669442785

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.262 seconds, Fetched: 36 row(s)
```

- Select only luxury car brands to a new table, and order by quantity

(Luxury car brands in this dataset include Audi, BMW, Lexus, Mercedes-Benz, Porsche, Volvo, Bentley.)

```
hive> CREATE TABLE model_maker_info_luxury
    > AS SELECT * FROM model_maker_info
    > WHERE maker IN ('audi', 'bmw', 'lexus', 'mercedes-benz', 'porsche', 'volvo', 'bentley')
    > ORDER BY model_maker_count DESC;
Query ID = ychsiaoca_20221126061730_6cd2ad09-9b2d-4f51-8896-54987ba99bd4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669439444647_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.05 s
----------------------------------------------------------------------------------------------
Moving data to directory hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info_luxury
OK
Time taken: 5.959 seconds
```

```
hive> DESCRIBE FORMATTED model_maker_info_luxury;
OK
# col_name              data_type               comment
model                   string
maker                   string
model_maker_count       bigint
_c3                     double
_c4                     double
_c5                     double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 26 06:17:36 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info_luxury
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                1
        numRows                 123
        rawDataSize             8024
        totalSize               8147
        transient_lastDdlTime   1669443456

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.068 seconds, Fetched: 36 row(s)
```

Calculate quantity of luxury used car ads by maker

(Makers with most luxury used car ads)

```
hive> SELECT maker, SUM(model_maker_count) quantity FROM model_maker_info_luxury
    > GROUP BY maker
    > ORDER BY quantity DESC;
Query ID = ychsiaoca_20221126063915_18e1c9ea-d675-4ff0-9689-155432a9814d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669439444647_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.24 s
--------------------------------------------------------------------------------
OK
audi    105509
bmw     27865
volvo   22740
mercedes-benz   12601
porsche 3452
lexus   2111
bentley 42
Time taken: 6.147 seconds, Fetched: 7 row(s)
```

Show the top 20 cars (model_maker) with the highest quantity

```
hive> SELECT * FROM model_maker_info_luxury
    > LIMIT 20;
Query ID = ychsiaoca_20221126062004_13721cf4-1b4d-4504-9360-80d54f0536be
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669439444647_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.14 s
--------------------------------------------------------------------------------
OK
a3      audi    31798   14181.152440333653      95869.19051512674       1917.8924604869715
a4      audi    16932   10658.528921565743      145319.95298842428      2083.9165947656024
a6      audi    10855   13323.519051998917      162481.2965453708       2624.5430320898063
x1      bmw     8410    20773.71283674569       64690.60642092747       2026.8277721113216
x3      bmw     8330    19772.380408169127      116649.1581032413       2395.5252320185614
a5      audi    7427    23374.242081257573      88679.23858893228       2293.7235566164404
a1      audi    6895    15529.039311137487      41646.00145032632       1433.8169744088111
vito    mercedes-benz   5637    12759.511709481578      132807.44367571402      2195.380608974359
x5      bmw     5505    18416.04657739679       158566.21035422344      3142.2373137551285
q5      audi    5435    25912.869251802193      94971.51904323827       2239.73167675924
q3      audi    4037    26577.5532866396        49583.89373297003       1969.7459154929577
coupe   audi    3557    17690.168518838567      108968.63396120326      2226.323680241327
v70     volvo   3171    11290.892335678192      184351.7013560391       2280.8364845938377
v50     volvo   3159    7507.441990634089       155790.53402975627      1864.187238493724
xc60    volvo   2979    23420.73143203597       87217.08694192683       2243.8986280487807
v40     volvo   2912    13933.227938054682      92624.50309065935       1813.4715978749489
tt      audi    2650    15727.859958173643      109756.82301886793      2074.6162790697676
viano   mercedes-benz   2451    20586.234646782006      128801.46389228886      2463.499771167048
q7      audi    2356    22327.32588379735       150986.81536502545      3167.906421105027
a8      audi    2269    13122.185193898658      198972.8915821948       3729.4173640167364
Time taken: 4.849 seconds, Fetched: 20 row(s)
```

- LUXURY used cars manufactured within 10 years

car details group by model_maker_year

```
hive> CREATE TABLE model_maker_info_yearly
    > AS SELECT model, maker, COUNT(CONCAT(model, maker)) model_maker_count, manufacture_year,
    > AVG(price_eur), AVG(mileage), AVG(engine_displacement)
    > FROM  used_cars_yichen_clean_5_3
    > GROUP BY model, maker, manufacture_year;
Query ID = ychsiaoca_20221126205141_28ef2285-2597-4a9b-978a-01050cb1ad21
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669439444647_0011)

--------------------------------------------------------------------------------
        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: 14.23 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info_yearly
OK
Time taken: 18.548 seconds
```

```
hive> DESCRIBE FORMATTED model_maker_info_yearly;
OK
# col_name              data_type               comment
model                   string
maker                   string
model_maker_count       bigint
manufacture_year        string
_c4                     double
_c5                     double
_c6                     double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 26 20:51:59 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info_yearly
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                1
        numRows                 7287
        rawDataSize             487573
        totalSize               494860
        transient_lastDdlTime   1669495920

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.103 seconds, Fetched: 37 row(s)
```

LUXURY car details group by model_maker_year

```
hive> CREATE TABLE model_maker_info_luxury_yearly
    > AS SELECT * FROM model_maker_info_yearly
    > WHERE maker IN ('audi', 'bmw', 'lexus', 'mercedes-benz', 'porsche', 'volvo', 'bentley')
    > ORDER BY model_maker_count DESC;
Query ID = ychsiaoca_20221126205359_1b596208-d152-486d-bdea-cabaf938210f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669439444647_0011)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.37 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info_luxury_yearly
OK
Time taken: 6.468 seconds
```

```
hive> DESCRIBE FORMATTED model_maker_info_luxury_yearly;
OK
# col_name              data_type               comment
model                   string
maker                   string
model_maker_count       bigint
manufacture_year        string
_c4                     double
_c5                     double
_c6                     double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 26 20:54:05 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info_luxury_yearly
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                1
        numRows                 1307
        rawDataSize             83142
        totalSize               84449
        transient_lastDdlTime   1669496045

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.066 seconds, Fetched: 37 row(s)
```

- Top LUXURY 20 used cars less than 10 years by ads quantity

```
hive> CREATE TABLE model_maker_info_luxury_yearly_less10
    > AS SELECT * FROM model_maker_info_luxury_yearly
    > WHERE maker IN ('audi', 'bmw', 'lexus', 'mercedes-benz', 'porsche', 'volvo', 'bentley') AND manufacture_yea
r >= 2012
    > ORDER BY model_maker_count DESC;
Query ID = ychsiaoca_20221126205933_1072be4e-7350-4076-a95f-afe97d9ae4e0
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_1669439444647_0012)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.15 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info_luxury_yearly_less
10
OK
Time taken: 14.41 seconds
```

```
hive> DESCRIBE FORMATTED model_maker_info_luxury_yearly_less10;
OK
# col_name              data_type               comment
model                   string
maker                   string
model_maker_count       bigint
manufacture_year        string
_c4                     double
_c5                     double
_c6                     double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 26 20:59:47 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_info_luxury_yearly_less1
0
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                1
        numRows                 194
        rawDataSize             12448
        totalSize               12642
        transient_lastDdlTime   1669496387

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.064 seconds, Fetched: 37 row(s)
```

```
hive> SELECT * FROM model_maker_info_luxury_yearly_less10
    > LIMIT 20;
Query ID = ychsiaoca_20221126210107_35f80fdd-bba1-44b0-ac28-9de2f7e54946
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669439444647_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.46 s
----------------------------------------------------------------------------------------------
OK
a3      audi    5283    2015    24426.46548739767       18395.934317622563      1667.4238488783944
a3      audi    3634    2014    23403.72614650695       22472.53302146395       1675.35630407911
a3      audi    2343    2012    15896.069451807913      71060.41570635937       1702.698669201521
x1      bmw     2228    2012    20592.34527007125       63686.58482944345       1999.2152739346984
a3      audi    1981    2013    20450.453796910493      48795.18727915194       1733.9104739613808
x3      bmw     1833    2012    27781.431198428123      78899.48990725586       2176.2073669849933
a5      audi    1803    2012    24887.627958935627      80357.89129229063       2185.7235188509876
q3      audi    1720    2012    24161.95776196857       69675.72325581395       1971.3145478374836
a4      audi    1565    2012    18612.143877296327      86351.32779552716       2057.729769858946
a1      audi    1561    2015    18577.364297850938      17483.534913516974      1439.8053035589671
a1      audi    1480    2012    14255.322715635557      53620.13445945946       1427.0172932330827
q5      audi    1478    2012    28673.652510254567      76988.90054127199       2219.4128035320086
a6      audi    1431    2012    25445.171365904524      92598.01187980433       2564.298953662182
a1      audi    1229    2014    17038.321741984084      22364.19853539463       1408.2757973733583
x1      bmw     1200    2013    23051.95216715495       51841.31416666666       2002.8216106014272
a1      audi    950     2013    15856.804971217105      37217.98842105263       1407.2383720930231
x1      bmw     848     2014    25842.017229188164      24779.787735849055      1994.5512422360248
q3      audi    834     2013    26783.366850269784      49836.460431654676      2011.8703208556149
q3      audi    824     2014    29241.608057114685      23426.86650485437       1948.0561151079137
x1      bmw     673     2015    27091.90455226133       19706.017830609213      1988.9128919860627
Time taken: 5.268 seconds, Fetched: 20 row(s)
```

Part II. Recommended Top 15 Luxury Used Cars for Driving Sales

Select cars which meet primary criteria

```
hive> CREATE TABLE model_maker_selection_1
    > AS SELECT * FROM used_cars_yichen_clean_5_3
    > Where maker IN ('audi', 'bmw', 'lexus', 'mercedes-benz', 'porsche', 'volvo', 'bentley')
    > AND manufacture_year >= 2012
    > AND mileage < 16093
    > AND engine_displacement <= 2000;
Query ID = ychsiaoca_20221126222325_ea609415-3d83-49f0-986b-bb9ec18ee095
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_1669439444647_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 14.12 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_selection_1
OK
Time taken: 22.239 seconds
```

The cleaned dataset (analysis sample)

```
hive> DESCRIBE FORMATTED model_maker_selection_1;
OK
# col_name              data_type               comment
maker                   string
model                   string
mileage                 int
manufacture_year        string
engine_displacement     int
engine_power            int
body_type               string
color_slug              string
stk_year                string
transmission            string
door_count              int
seat_count              int
fuel_type               string
date_created            string
datelastseen            string
price_eur               float
mileage_level           string
avg_price               double
sd_price                double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 26 22:23:47 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_selection_1
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                4
        numRows                 7241
        rawDataSize             1198875
        totalSize               1206116
        transient_lastDdlTime   1669501427

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.066 seconds, Fetched: 49 row(s)
```

Group by model_maker and ordered by quantity

```
hive> CREATE TABLE model_maker_selection_2
    > AS SELECT model, maker, COUNT(CONCAT(model, maker)) model_maker_count,
    > AVG(price_eur), AVG(mileage), AVG(engine_displacement)
    > FROM  model_maker_selection_1
    > GROUP BY model, maker
    > ORDER BY model_maker_count DESC;
Query ID = ychsiaoca_20221126222838_6910d4c8-ae95-4a2d-ab6d-913a9c24a338
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1669439444647_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.98 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_selection_2
OK
Time taken: 6.047 seconds
```

```
hive> DESCRIBE FORMATTED model_maker_selection_2;
OK
# col_name              data_type               comment
model                   string
maker                   string
model_maker_count       bigint
_c3                     double
_c4                     double
_c5                     double

# Detailed Table Information
Database:               cardata
OwnerType:              USER
Owner:                  ychsiaoca
CreateTime:             Sat Nov 26 22:28:44 UTC 2022
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://hive-midterm-m/user/hive/warehouse/cardata.db/model_maker_selection_2
Table Type:             MANAGED_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   {\"BASIC_STATS\":\"true\"}
        bucketing_version       2
        numFiles                1
        numRows                 29
        rawDataSize             1734
        totalSize               1763
        transient_lastDdlTime   1669501724

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
Time taken: 0.067 seconds, Fetched: 36 row(s)
```

```
hive> SELECT * FROM model_maker_selection_2;
Query ID = ychsiaoca_20221126224139_ac21aa2a-6b28-4e2f-8817-937471985808
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_1669439444647_0015)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.12 s
--------------------------------------------------------------------------------
OK
```

```
a3      audi    3452    24605.478196516513      11678.950463499421      1659.3146002317496
a1      audi    1196    17879.725585120977      11303.17474916388       1403.2934782608695
q3      audi    575     29566.50589673913       11434.114782608696      1912.5808695652174
x1      bmw     403     26999.446223635237      12409.022332506203      1986.0942928039701
v40     volvo   319     20764.500355113636      12477.639498432602      1749.5924764890283
90      audi    192     22975.276789347332      11749.5 1632.546875
a4      audi    180     27015.44130859375       11794.116666666667      1928.7777777777778
a5      audi    172     31408.796579760176      11754.127906976744      1929.9883720930231
coupe   audi    132     30609.116225733902      11134.295454545454      1938.878787878788
v60     volvo   111     24595.723500844593      12792.972972972973      1809.054054054054
tt      audi    107     29510.00773948598       11548.29906542056       1962.392523364486
xc60    volvo   81      31248.816502700618      12618.851851851852      1976.4074074074074
z4      bmw     72      29360.65847439236       11545.222222222223      1997.0972222222222
q5      audi    39      31542.029954176684      12859.153846153846      1973.3333333333333
a6      audi    36      32327.101725260418      12699.638888888889      1965.5
v70     volvo   25      29117.87203125  12704.44        1961.48
80      audi    24      29416.806559244793      11881.291666666666      1834.75
ct-200h lexus   23      23671.23046875  11802.608695652174      1798.0869565217392
s60     volvo   20      24061.14296875  13460.55        1875.35
s3      audi    19      33110.44140625  11477.842105263158      1984.0
a4-allroad      audi    17      32849.97713694853       12099.588235294117      1968.9411764705883
x3      bmw     17      31541.713120404413      12706.941176470587      1995.4705882352941
vito    mercedes-benz   9       23231.747612847223      11367.111111111111      1598.0
i3      bmw     7       33416.76422991072       11780.57142857143       647.2857142857143
xc90    volvo   5       21366.14404296875       12580.2 1969.0
100     audi    3       28165.772786458332      10666.666666666666      1777.0
s80     volvo   2       24195.0 15750.0 1976.5
xc70    volvo   2       32802.923828125 14400.0 1969.0
c30     volvo   1       16682.859375    13500.0 1560.0
Time taken: 12.217 seconds, Fetched: 29 row(s)
```

Export primary selection result to csv file

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/ychsiaoca/model_maker_selection_2'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > select * from model_maker_selection_2;
Query ID = ychsiaoca_20221126230539_82b6a466-fbbc-4527-b81f-6b8d6b729d89
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_1669439444647_0016)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.29 s
--------------------------------------------------------------------------------
Moving data to local directory /home/ychsiaoca/model_maker_selection_2
OK
Time taken: 12.137 seconds
hive>
    > --download path: '/home/ychsiaoca/model_maker_selection_2/000000_0'
```