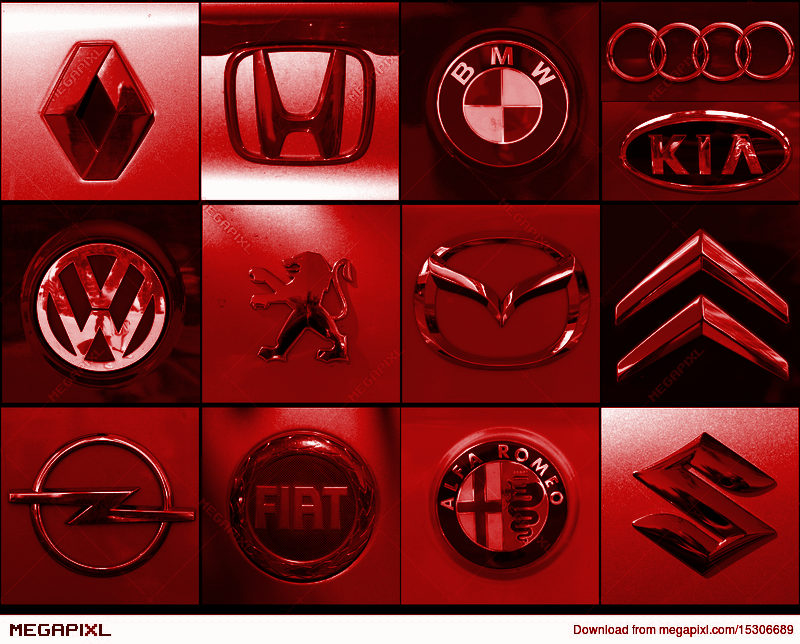
## horizontal line



Big Data Investment

Used Car Report

12/02/2020

**─**

Filipe Soares

CSDA 1020

**Overview**

Big Data Investment is one of the largest investment firms in the world and is looking to expand into the used car business. Since it has opened a new corporate office in Prague, the decision was made to look into the used car advertisements in the surrounding areas of Eastern Europe. Apache Hive was used to identify the best cars to purchase in order to take Big Data Investment to the next level. It has been decided that this company should buy cars that fit into the following criteria:

* Audi, Chevrolet, Honda, Mazda, and Mercedes-Benz
* Mainly diesel with some gasoline-powered vehicles
* From the years 2010-2013
* Almost exclusively manual transmission, very few automatic
* Hatchbacks (5-door cars), with a mix of 4-door and 2-door to a lesser extent
* Mainly cars with 5 seats, some with 4 seats.

The report will discuss how these conclusions were made and screenshots the relevant code for data cleaning and visualizations can be found in the appendix.

# Goal

To clean and analyze the data to provide insight that will determine the best cars to purchase based on the following questions:

1. What brands provide the best value based on availability, average price and average mileage?
2. Is there a market for greener fuel types?
3. Are there specific manufacturer years that should be avoided?
4. What body types, door count, seat count and transmission are ideal?

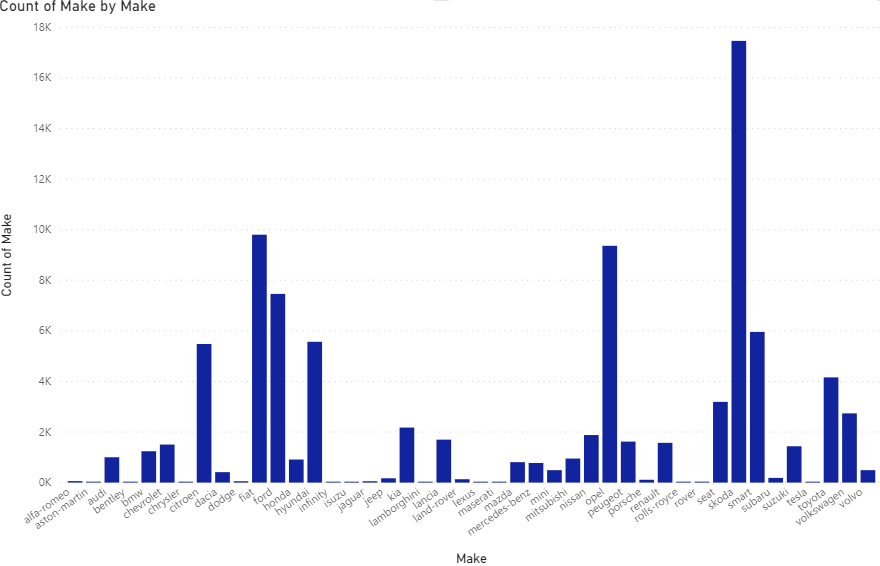
**Analysis**

Data Cleaning

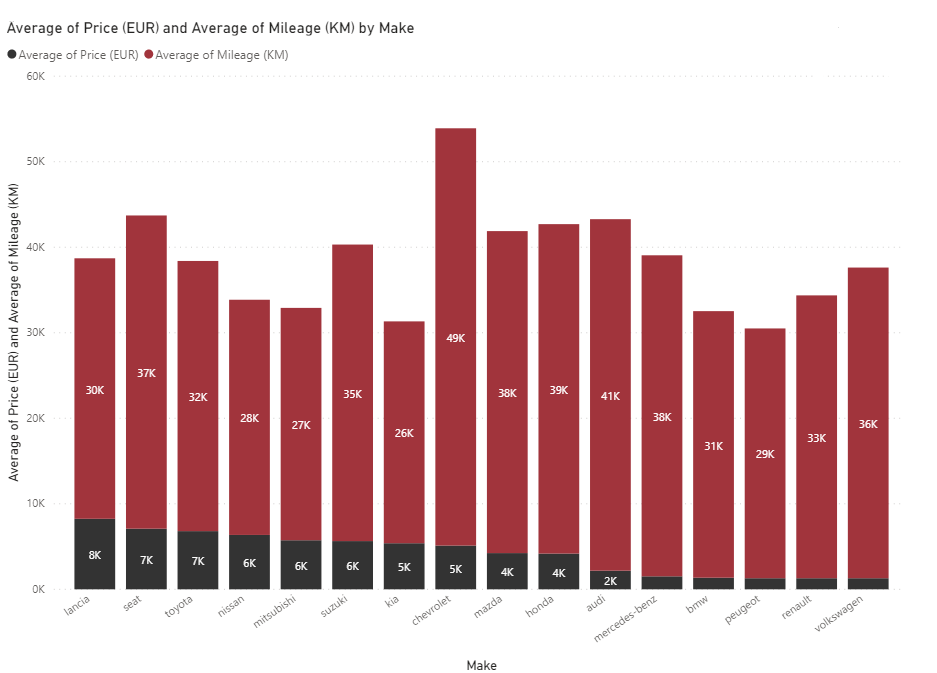
This section will provide a brief summary of how the data was cleaned and prepared for this report. The full details and screenshots will be found in the appendix. The original dataset contained about 3.5 million cars. Through filtering for; make (remove null), model (remove null), mileage (under 100,000 km), manufactured year (2010 and later), transmission (remove null), fuel type (remove null), and price (10,000 euros and below) the final table had just over 91,000 entries. Door and and seat count were included in the final table but did not get filtered as it was determined removing the null values would unnecessarily lower the pool of cars available. Other columns were deemed to be of less importance or extremely incomplete (date last seen had a large percentage of null values) so it was decided to drop these for the final table as they provide little value.

Brand

Identifying the ideal brands to invest in was the first piece of analysis completed. This was done by first narrowing down the selection to those that had between 500 and 5000 listings, mainly to eliminate the outliers. Brands with more than 5000 listings are likely not attracting enough buyers, and those below 500 are likely too niche of a market. Figure 1 demonstrates the distribution of cars by make. Skoda, Fiat, Opel and Ford had many more listings than the others, while Aston Martin, Lamborghini and Tesla were among those with the fewest.

**Figure 1**

After identifying which makes to eliminate based on availability, there were 16 car brands that remained: Audi, BMW, Chevrolet, Honda, Kia, Lancia, Mazda, Mercedes-Benz, Mitsubishi, Nissan, Peugeot, Renault, Seat, Suzuki, Toyota and Volkswagen. Looking at finding a balance between kilometers driven and price would illustrate how certain makes maintained their value in relation to how much the car was used. Figure 2 breaks down this relationship among the remaining 16 makes.

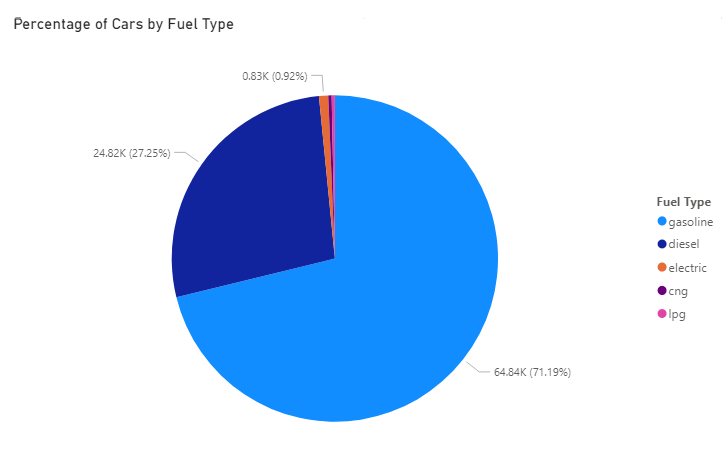


**Figure 2**

Analysis of Figures 1 and 2 determined that the top 5 makes were Mazda, Audi, Mercedes-Benz, Chevrolet, and Honda. At first the extremely low average price for Audi (2179 EUR) and Mercedes-Benz (1519.11 EUR) are alarming, indicating that perhaps there may be mechanical issues with the vehicles. However, most of the European-made vehicles (aside from Lancia and Seat) are significantly cheaper than the imports (North American & Asian) so those prices are not unusual. When you compare these makes to those of similar price ranges, these 2 brands have more mileage. This displays that there may be higher resale value than Peugeot, BMW and Renault. Volkswagen appeared at first glance to have great value as well but upon further examination the minimum and maximum price were the same, indicating there may be an error in the data (can be found in the Appendix). Chevrolet, Honda and Mazda are the 3 best imports. Honda has an average price of 4179.41 EUR and average mileage of 38,522.04 km. Mazda had an average price of 4235.57 EUR and an average of 37651.36 kilometres driven. Chevrolet’s average price is a bit higher at 5114.30 EUR but the average mileage is 48,794.21 km, which is by far the highest average mileage. The prices are affordable in the Honda, Mazda and Chevrolet while still having many kilometers driven, indicating the potential for higher resale value.

Fuel Type

Another aspect to the analysis was establishing whether or not there should be heavy investment in cleaner fuels. The percentage of each fuel type can be found in Figure 3. About 71% of the cars for sale use gasoline, while 27.25% take diesel. The remainder of the cars were split between electric, compressed natural gas (CNG) or liquefied petroleum gas (LPG).

**Figure 3**

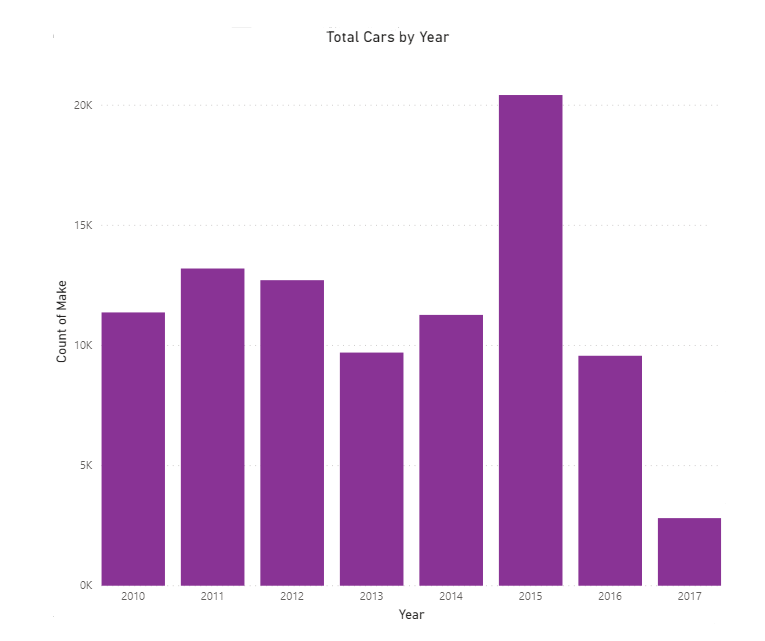
Upon first glance it would appear that there should be much greater investment in gasoline powered cars but further analysis found including average price and average mileage (found in Figure 4) indicates otherwise. Diesel cars have a lower average price to gasoline (4101.25 EUR and 5977.65 EUR respectively) and have an average mileage of 47,830.25 compared to gasoline which has an average of 31,287.58. It is not unusual for the price of the vehicle to decrease as it is used more often, however the low number of kilometres driven on the gasoline powered cars as well as the large amount of units available could indicate that drivers are less satisfied with this fuel type and are more eager to sell them. The cleaner fuel types are available at lower prices but they generally have much less kilometres driven than the diesel cars. The LPG vehicles actually have the highest amount of kilometres on them and have a similar price point to their other “green” counterparts, however the extremely low number available (288 cars) indicates the chances of customers buying this car are minimal.

****

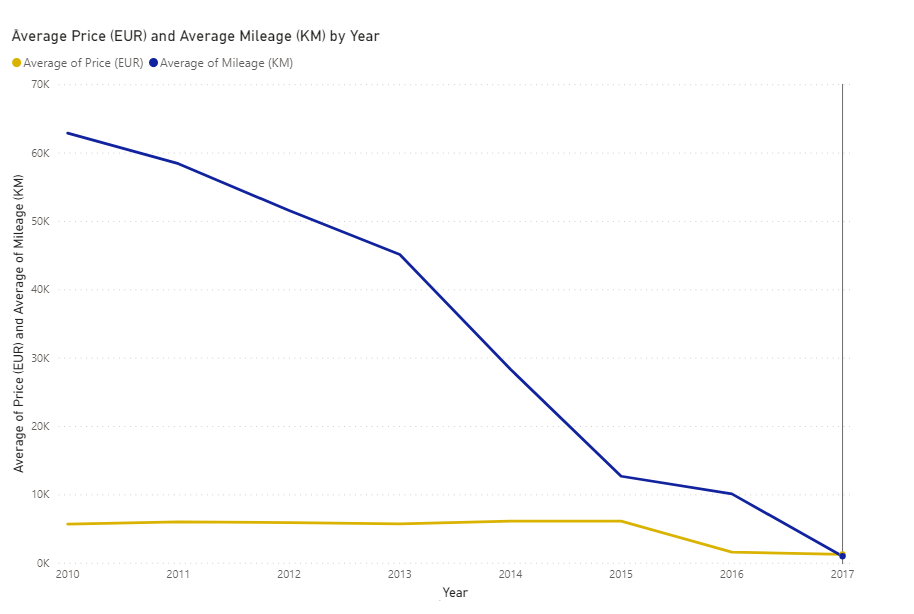
**Figure 4**

Year

The year cars were manufactured is also a factor in purchasing a vehicle. Cars that are too old are likely to be viewed as less reliable (unless the buyer is looking for certain classic vehicles) and cars that are new are usually more expensive or defective. Using the data available, the total cars on sale were divided by manufactured year in Figure 5. Here it is seen that there is a large number of vehicles on sale that were made in 2015, indicating that this year may be less desirable than the rest. There is also a very low number of vehicles from the year 2017 but it is difficult to determine if this is significant or if it is a matter of drivers who are currently satisfied with their newer vehicles. Figure 6 provides a clearer picture of what cars provide the best value. There is a steady increase in average price from 2010-2015, and then the price decreases sharply for 2016 and 2017. There is a steady decrease in average kilometers driven until a steep decline in 2014 that continues until 2017. Based on the available information, it is recommended to purchase cars between the years 2010-2013.



**Figure 5**

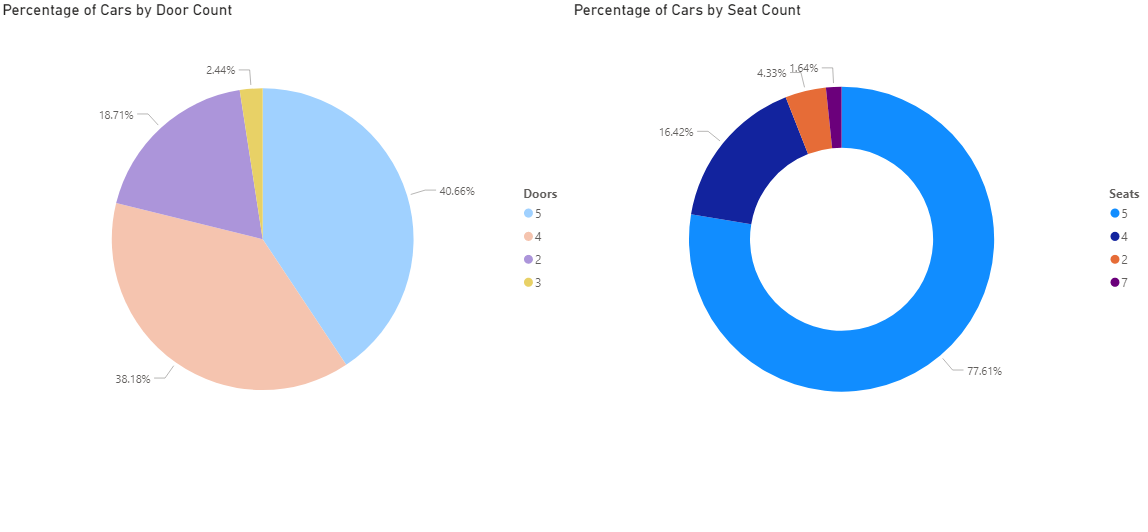


**Figure 6**

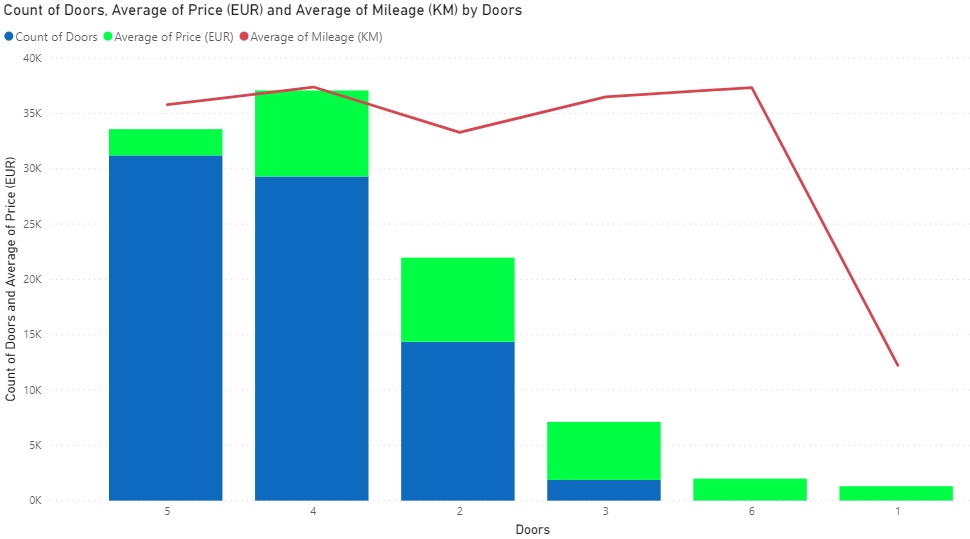
The increase in average price from 2013 to 2014 (5,767.26 EUR to 6,149.38 EUR) and the sharp decrease in average kilometres between the same years (about 17,000 km difference) indicates 2013 may be last year containing affordable prices and high resale value.

Doors & Seats

A brief breakdown of cars by door count and seat count was conducted to identify the most common body types. “Body Type” was one of the columns excluded in the data cleaning process because of the high percentage of missing data. Removing null values would have reduced the number of cars available unnecessarily. Figure 7 illustrates that 5-door vehicles, commonly known as hatchbacks, make up 40.66% of the available cars. Sports cars, pick-up trucks and coupes are most likely to have 2 doors, and they are 18.71% of the pool. All other vehicle body types generally fall into the 4-door category and amount to 38.18% of the cars for sale. 5-door vehicles have similar kilometres driven and lower prices on average than 4-door cars so this indicates excellent value. These findings can be found in Figure 8. This differs with the abundance of 2015 cars available because the price of 2015 cars was higher than all other years while the average kilometres driven was much lower than the previous years. Most vehicles have a back seat and have a capacity of 5 people, so it is not surprising to see that the vast majority of the cars for sale have 5 seats. From the data available it is recommended to buy a larger percentage of hatchbacks and mainly cars with a 5 seat capacity. There should also be high investment in cars with 4 doors with the remaining consisting of 2-door cars, and very few should have 2 or 7 seats.



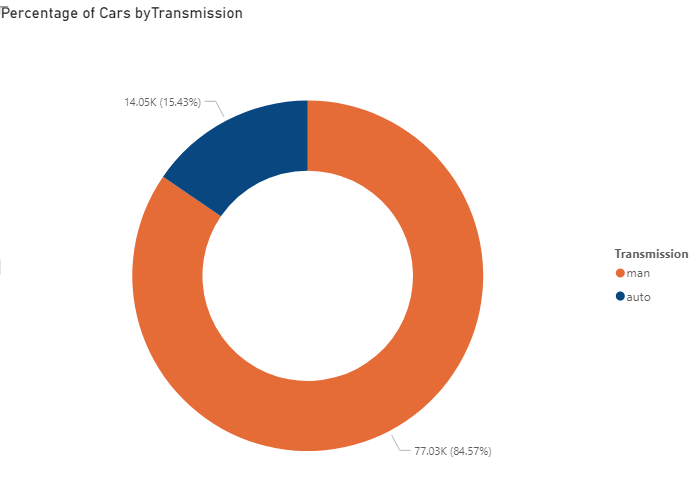
**Figure 7**



**Figure 8**

Transmission

Finally, the cars available were broken down by transmission type. Figure 9 reveals the overwhelming amount of cars with manual transmission. The high percentage likely means that manual is much more common than automatic in general, not only the used car ads. Due to the high imbalance of transmission type skewing heavily towards manual, this factor was not investigated any further. It is recommended that the cars purchased are almost exclusively manual.



**Figure 9**

**Recommendations**

The recommendations for Big Data Investment are as follows:

* The top 5 makes to purchase are Audi, Chevrolet, Honda, Mazda, and Mercedes-Benz. If the values of the Volkswagen can be verified then it would replace Mazda, however every vehicle having the same listed price is suspicious.
* About 45-50% of the vehicles should use diesel as they provide great value, with the remaining cars being powered by gasoline. At this moment it appears there is little to no value in buying cars of cleaner fuel types.
* Cars from the years 2010-2013 appear to hold their value the most, so investment in these vehicles is recommended. 2014 appears to be the beginning of a worrisome downward trend in average price per average kilometers driven.
* Buying cars with 4 or 5 doors is preferred (roughly 40% each), as well as a smaller percentage of 2-door vehicles (15-20%).
* About 75-80% of the cars should have 5 seats. A much lesser percentage of cars should have 4 seats.
* 85-90% of the cars purchased should have a manual transmission as that appears to be the most popular transmission type by a large margin.

# Appendix

## Data Cleaning

Creating a table removing the null values in the “make” column. This process was repeated to create tables removing null from model, transmission and fuel types. The same process was used to create a table with mileage below 100,000 km. No tables were made for seats and doors as they would have removed many rows unnecessarily.

## 

Creating the “final” data table with desired columns from the “fuel” table. The data was filtered to include only cars that had a price equal to or less than 10,000 EUR. SELECT COUNT (\*) was used to identify how many rows remained, which is 91087.

## 

Describing the “final” table using DESCRIBE FORMATTED to find the path and describe the file in detail.

## 

## Inserting the final table to a new directory.

## 

Downloading the table from local directory, using “copy to local”, ls (to confirm the file exists in the local directory) and using the “download file” option in the SSH

## 

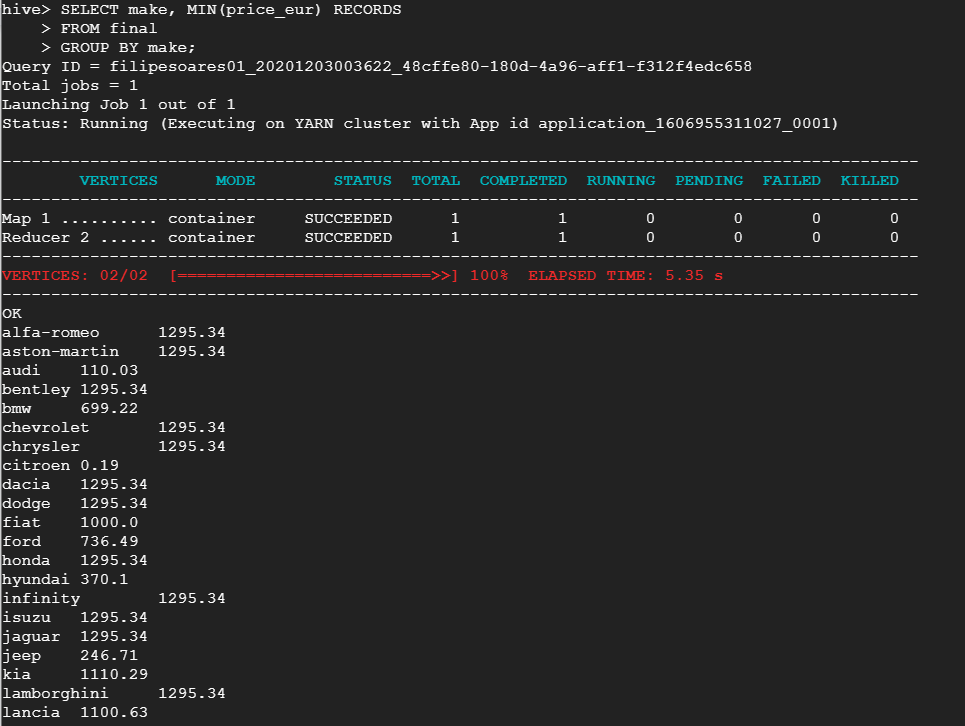
## Filtering For Visualizations (all visualizations created with Power BI)

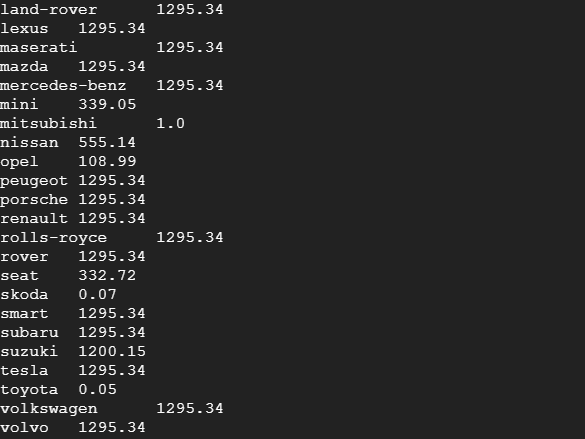
Selecting the data used for creating Figures 1 and 2

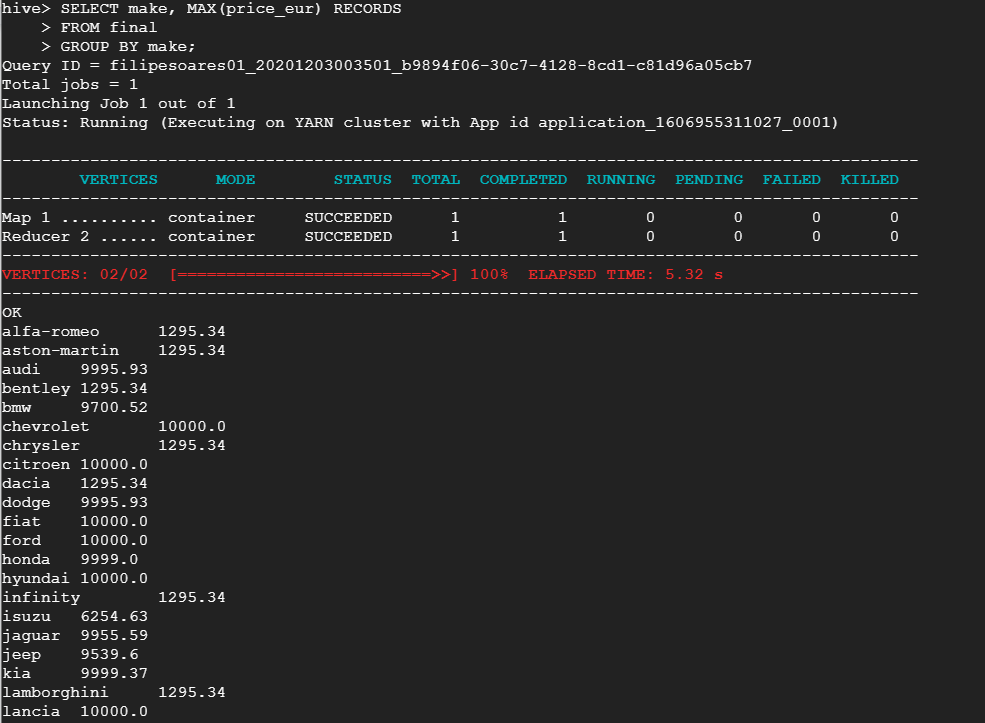
## 

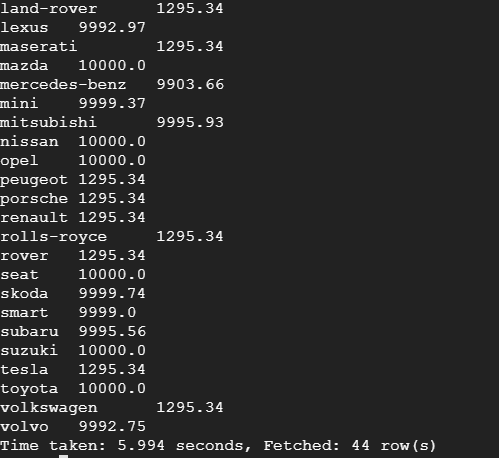
## 

## Minimum prices by “make” (2 images) followed by Maximum prices. This shows Volkswagen has the same minimum and maximum, which means all entries for Volkswagen in this table are of the same value.









## 

## Selecting data used for creating Figure 3

## 

## Selecting data used for creating Figure 4

## 

Selecting data used for creating Figure 5

## 

Selecting data used for creating Figure 6

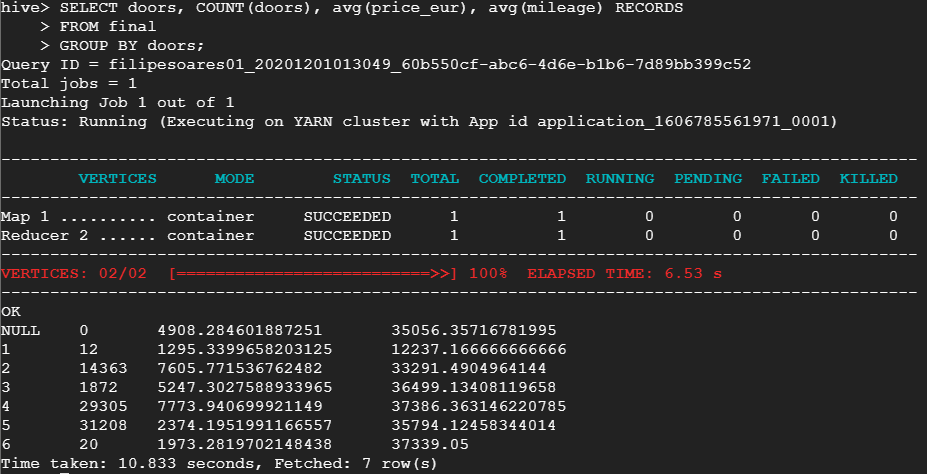
## 

Selecting data used for creating Figure 7

## 

## 

Selecting data used for creating Figure 8



Selecting data used for creating Figure 9

## 

## 

**Assumptions**

* This report was requested in 2017 or 2018, as the last available manufacture year is 2017
* No other data exists about the listed cars (reliability and safety ratings, yearly consumer awards/sales, etc.)
* The company is located in or looking to expand into Europe, eliminating the variables of differing car value (many of the cars made in Europe are cheaper, the American and Asian cars are more expensive in this dataset), car availability in North America (many of these brands are uncommon/not available to purchase in North America) and removing the cost of transporting the vehicles to North America.