AIT 580 - Final Project Report

SriUdaya Damuluri
G01111092

This report is based on the observations of over 370K used cars web scraped from Ebay Kleinanzeigen a German eBay organization, with the information of about 10 attributes related to the cars. The dataset is an important source for the decision making of the organization on the progress of business.

The data speaks about which used car was sold on eBay, for what price along with its attributes like kilometers travelled, model etc. All the attributes can be classified into two types of data String and Integer. The dataset provides valuable information about various brands, models etc to give an overall broad perspective of the insights into the business workflow in used cars.

For every used car advertisement that has been posted on Ebay, there is an attribute that speaks about no kilometers travelled by the car before it was posted. That is a valuable information for determining which cars models are being sold earlier than others and the prices associated with them. The fuelType attribute speaks about which fuel type is preferable. The dataset also projects information about the number of posts of cars Ebay displays from each postal code i.e. locality. All of this data can be used to predict to car values for a particular brand of car and a particular model.

Who collected the data?

Ebay Kleinanzeigen, is a German Ebay organization which is the Germany version of eBay Inc. - a multinational e-commerce corporation headquartered in San Jose, California that facilitates consumer-to-consumer and business-to-consumer sales through its website. It was established in 1995 by Pierre Omidyar on a Labor Day coding at home, he came up with eBay.

Ebay is one of the dominant partners in the e-commerce domain with many million users trying to buy and sell all categories of products using the eBay website. As Germany specializes in cars the sales of cars are highly predominant in the eBay website of Germany.

Need

The dataset would be highly important to make decisions on business improvements by speaking which brands are doing good and which brands of cars need improvement, what is the pricing of cars that is more likely to be accepted by the buyers. It can answer highly critical questions and predict the trends in the sales of second hand cars from eBay that will help fix the issues if any and focus more on the sectors doing poor in the sales and by investing more in those sectors. This collection and analysis of data would help figure out which brand cars are selling more, what average kilometers travelled will make the car more preferable to be sold.

It speaks about the pricing of cars whether it is too high or too low for any model of car. Also, projects stats about which model of cars have more sales.

There can be potentially many answers that can be found from the analysis on the dataset. Some of them are:

1. How many kilometers on an average did cars from each brand travel before people decided to sell it?
2. Which fuel type of car is sold more than others?
3. Average kilometers for each type of Vehicle being put on sale?
4. Average price of used cars for each brand, which is highest and lowest?

Requirements, resources needed:

There are many analytical tools like SQL, SQL Workbench, Tableau, etc. for analyzing the given dataset, I have used SQL and SQL Workbench as the analyzing tool. I have used a Macbook Pro with 16 GB RAM and running MacOSX. The response to graph plots in SQL WorkBench was very slow even though a small part of the dataset was analyzed using a 16GB memory hardware. It can be implied that a higher configuration of hardware and software must be used to for the complete dataset.

DataSet:

The metadata of the dataset is as given below:

- Name (String) - Name of the car
- Price (Integer) - Price of the used car
- VehicleType (String) - Type of the vehicle
- YearofRegistration (Integer) - Year when the car got registered
- PowerPS (Integer) - Power in PS
- Model (String) - Model of the car
- Kilometer (Integer) - Kilometers travelled
- Fueltype (String) – Type of the fuel used
- Brand (String) – Brand of the car
- PostalCode (Integer) – Postal Code where the car got registered

Fig.1 describes the metadata using SQL queries which contains the name of each column and the respective datatype.

```
4 DESCRIBE CARS;
5
```

COLUMN_NAME	DATA_TYPE	PK	NULLABLE	DEFAULT	AUTOINCREMENT
NAME	VARCHAR2(100 Byte)	NO	NO		NO
PRICE	NUMBER(30)	NO	NO		NO
VEHICLETYPE	VARCHAR2(100 Byte)	NO	NO		NO
YEAROFREGISTRATION	VARCHAR2(100 Byte)	NO	NO		NO
POWERPS	NUMBER(10)	NO	NO		NO
MODEL	VARCHAR2(20 Byte)	NO	NO		NO
KILOMETER	NUMBER(30)	NO	NO		NO
FUELTYPE	VARCHAR2(20 Byte)	NO	NO		NO
BRAND	VARCHAR2(20 Byte)	NO	NO		NO
POSTALCODE	NUMBER(10)	NO	NO		NO

Fig. 1

Results / Findings:

Potential questions that can be answered by looking at the data:

1) How many kilometers on an average did cars from each brand travel before people decided to sell it?

```
4 SELECT AVG(kilometer) as avg, brand FROM cars group by brand order by avg desc ;
5
```

Result 1 | Messages

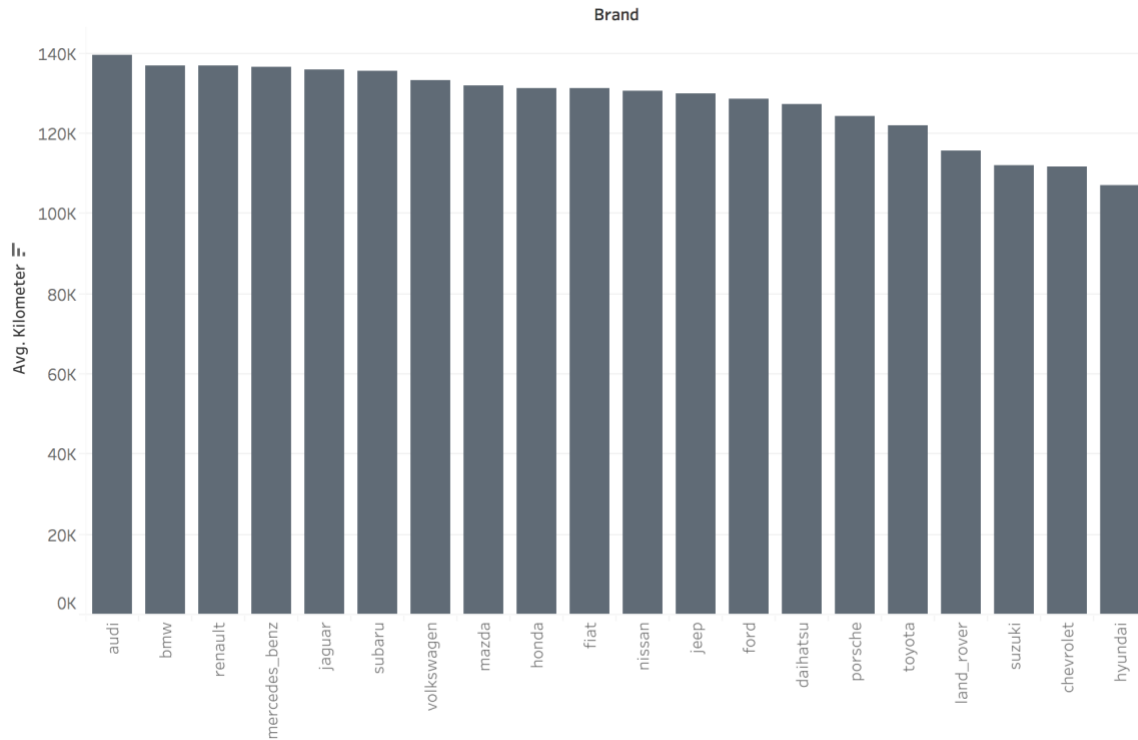| AVG	BRAND
139517.89	audi
137021.07	bmw
136765.68	renault
136412.92	mercedes_benz
136000	jaguar
135677.08	subaru
133179.40	volkswagen
131763.67	mazda
131361.87	honda
131351.12	fiat
130728.07	nissan
130078.13	jeep
128671.92	ford
127346.15	daihatsu
124210.53	porsche
122104.48	toyota
115795.45	land_rover
111930.50	suzuki
111569.34	chevrolet
107128.71	hyundai

Fig. 2

Fig. 3

We can see that Audi, BMW, and Renault are the most used brands which travelled the maximum kilometers.

2) What is the most common fuel type used?

```sql
SELECT count(*) as count, fuelType from cars group by fueltype order by count desc ;
```

Result 1   Messages

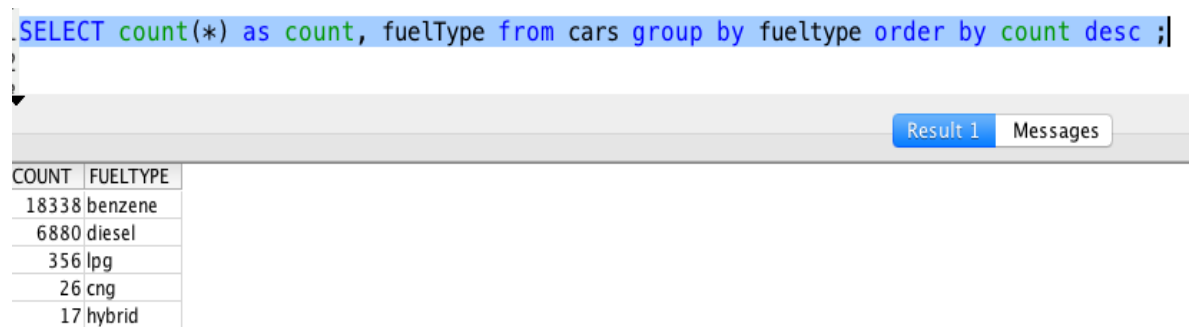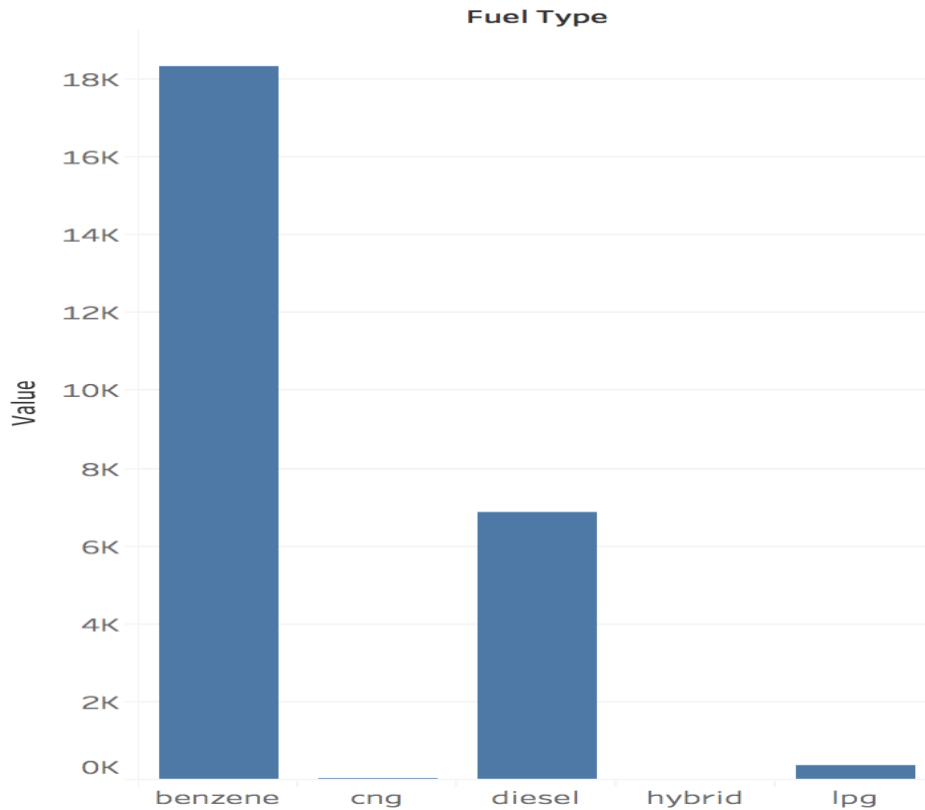| COUNT	FUELTYPE
18338	benzene
6880	diesel
356	lpg
26	cng
17	hybrid

Fig. 4

Fig. 5

This above representation gives an explanation about the different types of fuel used by various types of cars. From the graph, it can be figured out that benzene is being used by most of the vehicles.

3) Average kilometers travelled for each type of vehicle being put on sale?


```sql
SELECT AVG(kilometer) as avg, vehicleType FROM cars group by vehicleType order by avg desc ;
```

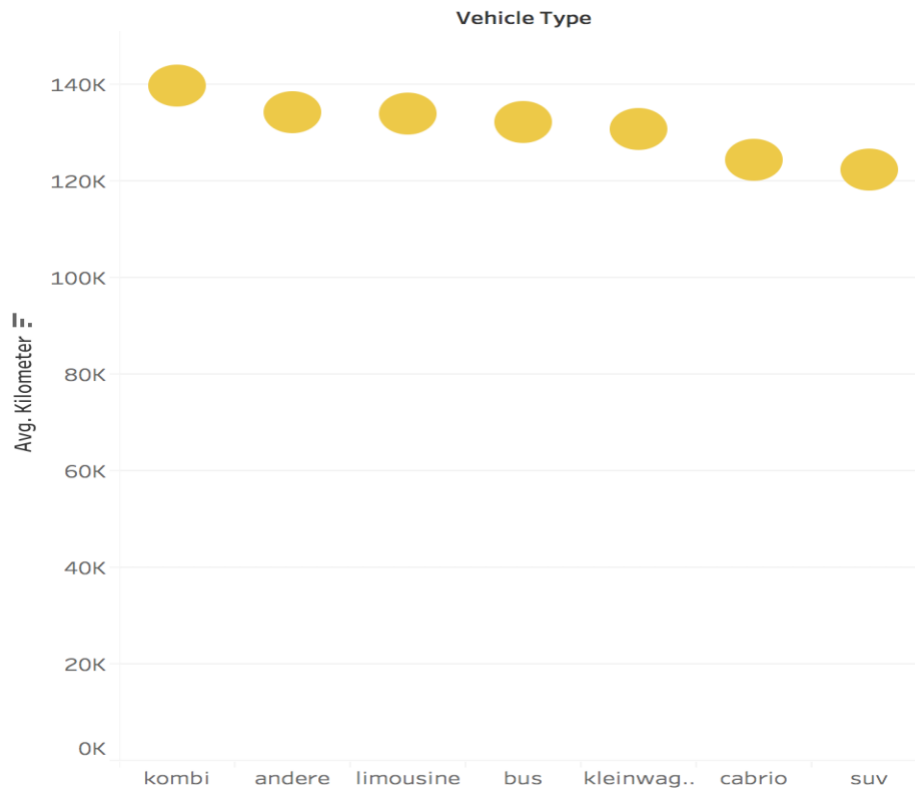| AVG	VEHICLETYPE
139782.33	kombi
134389.83	andere
133982.82	limousine
132373.93	bus
130708.23	kleinwagen
124464.44	cabrio
122465.36	suv

Fig. 6

Fig. 7

The illustration provided previously gives a brief description about the types of vehicle and their average amount of distance travelled in terms of kilometers.

4) Average price of used cars for each brand, which is highest and lowest?

```sql
24 SELECT AVG(price) as price, brand FROM cars group by brand order by price ;
25
```

Result 1   Messages

PRICE	BRAND
1064.08	daihatsu
1325.16	fiat
1382.28	renault
2013.29	ford
2538.28	nissan
2720.17	subaru
2889.30	mazda
2965.90	honda
3313.05	suzuki
3548.78	volkswagen
4660.62	hyundai
4829.17	toyota
5579.93	mercedes_benz
6186.88	chevrolet
6489.28	audi
6596.10	bmw
8016.72	jaguar
10459	jeep
15428.55	land_rover
16584.63	porsche

Fig. 8

price and brand

Fig. 9

The representation provided above gives an idea about the average price of the cars for each brand upon the time of resale. We can conclude that Porsche company has the highest resale value as compared to any other brand.

References

1. **Details of eBay History and Company Information found from the organization's website.**
    https://www.ebayinc.com/our-company/our-history/

2.   **The dataset for the eBay Germany version was scraped by kaggle website and converted from Germany to English for common people to understand.** © 2017 Kaggle Inc

    https://www.kaggle.com/orgesleka/used-cars-database

3.   **eBay Germany version website - (e-commerce website)**

    https://www.ebay-kleinanzeigen.de/