**Data exploration and Cleaning using HIVE**

**Dataset : Used cars for sale in Germany and Czech Republic since 2015**

**Source :** [**https://www.kaggle.com/mirosval/personal-cars-classifieds**](https://www.kaggle.com/mirosval/personal-cars-classifieds)

The data was scraped from several websites in Czech Republic and Germany over a period of more than a year.

**Data Cleaning**

Apache HIVE has been used to load and clean data.

The data is loaded into a table named cars.

Code used –

***hive> CREATE EXTERNAL TABLE IF NOT EXISTS cars (***

***> maker STRING,***

***> model STRING,***

***> mileage INT,***

***> manufacture\_year INT,***

***> 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,***

***> date\_last\_seen string,***

***> price\_eur FLOAT)***

***> ROW FORMAT DELIMITED FIELDS TERMINATED BY ','***

***> LOCATION '/BigData/hive'***

***> TBLPROPERTIES("skip.header.line.count"="1");***

Further, datatype of ***date\_created, date\_last\_seen*** changed to **‘date’** datatype as currently it is ‘**string**’ datatype. Following code was used to convert the datatypes and a new table **cars\_new** was created :

***CREATE TABLE IF NOT EXISTS cars\_new AS***

***SELECT maker, model, mileage, manufacture\_year, engine\_displacement, engine\_power ,***

***body\_type, color\_slug, stk\_year, transmission , door\_count, seat\_count, fuel\_type,***

***CAST(to\_date(from\_unixtime(unix\_timestamp(date\_created,'yyyy-MM-dd'))) AS date) as date\_created,***

***CAST(to\_date(from\_unixtime(unix\_timestamp(date\_last\_seen,'yyyy-MM-dd'))) AS date) as date\_last\_seen,***

***price\_eur FROM cars;***

The ‘**maker**’ column contained 518915 rows with blank values. These empty rows were removed and a new table was created named **cars\_new1**. Code used:

***hive> CREATE TABLE cars\_new1 AS SELECT \* FROM cars\_new WHERE maker <> '';***

Value of ‘**body\_type**’ column was blank for more than 32% records. Also, more than 94% records had blank value for ‘**color\_slug**’ column. For column **‘stk\_year’**, the values were either blank or ‘None’. All these columns have been removed for further analysis.

Values under **‘door\_count’** and **‘seat\_count’** were same for the records. Hence, **‘door\_count’** column was removed to avoid redundancy.

Also, columns **‘engine\_displacement’** and **‘date\_last\_seen’** were removed as they do not contain any major data which could be used in our analysis.

The values for **‘manufacture\_year’** has been kept older than 1996 and also, all the records which has price less than $500 and more than $1,000,000 has been removed for our data analysis as there very many values did not seem correct or logical.

Below code was used to perform all the operations and a new table – **‘cars\_cleaned’** was created :

**CREATE TABLE IF NOT EXISTS cars\_cleaned AS**

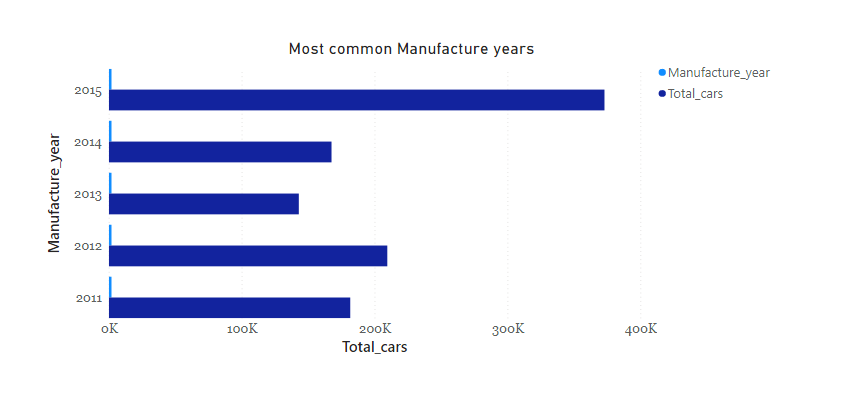
**SELECT maker, model, mileage, manufacture\_year, engine\_power, transmission, seat\_count, fuel\_type, date\_created,price\_eur**

**FROM cars\_final**

**WHERE price\_eur BETWEEN 500 AND 1000000 AND manufacture\_year > 1996 AND mileage <= 1000000;**

**Data Analysis**

1. Which are the most common manufacture years for all the listed cars?



2015 is the most common manufacture year with 372943 cars with year 2015. Then, it is followed by 2012, 2011, 2014 and 2013. These data were scrapped in from 2015 to 2017. Majority of cars are not older than 5 years.

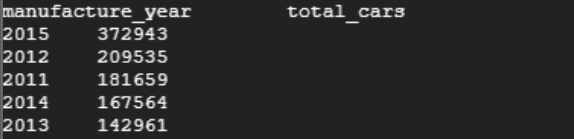
Code used to derive the results:

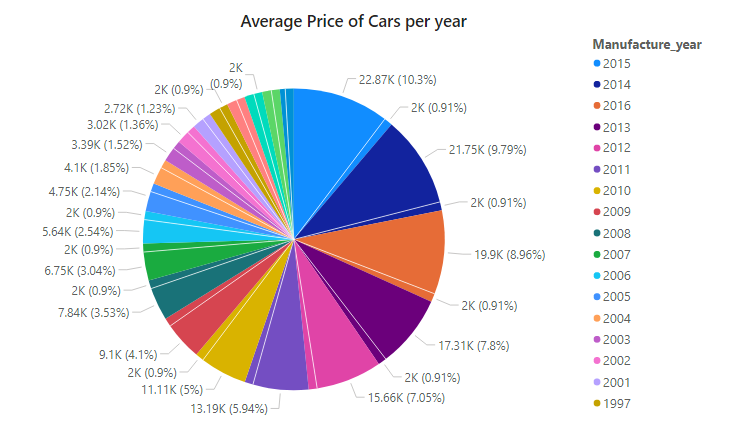
***SELECT manufacture\_year, total\_cars***

***FROM (SELECT COUNT(maker) as total\_cars, manufacture\_year FROM cars\_cleaned GROUP BY manufacture\_year)as temp***

***ORDER BY total\_cars DESC LIMIT 5;***

Output:



2) What is average price of the cars for every manufacture year? Or Does manufacture year influence car pricing?

The average price- 22872.25 is highest for cars with manufacture year 2015. The older the manufacturing year, lesser the price of car. The average price is almost ten times cheaper for cars with manufacture year 1997 than 2015.

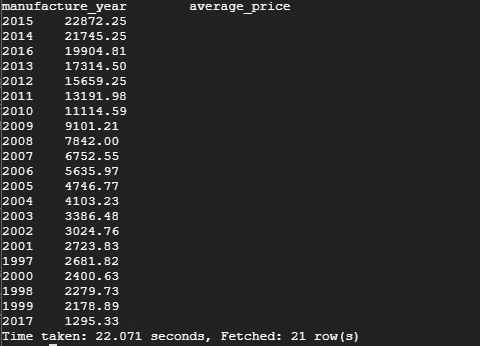
Code used to derive the results:

***SELECT manufacture\_year, Average\_price***

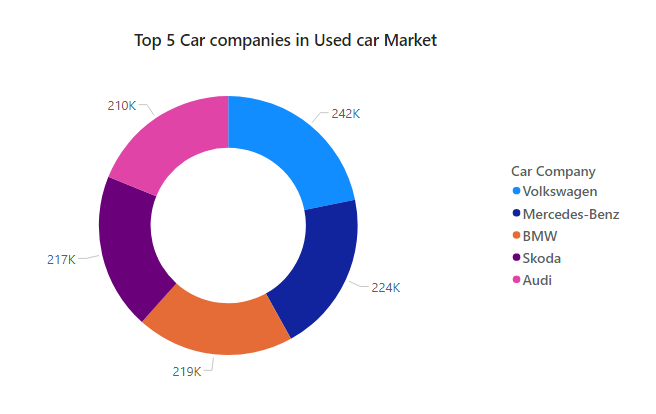
***FROM (SELECT manufacture\_year, ROUND(AVG(price\_eur),2) AS Average\_price FROM cars\_cleaned GROUP BY manufacture\_year) AS temp***

***ORDER BY Average\_price DESC;***

Output:



1. Which is the most common car manufacturer or company in used cars market?

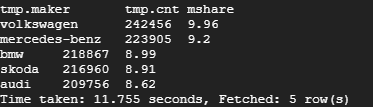


Volkaswagen, Mercedes-Benz, BMW, Skoda and Audi are the top five car makers in used car market as per the data.

Code used to derive the results:

***SELECT tmp.\*,ROUND(tmp.cnt/(SELECT SUM(cnt) AS snt FROM (SELECT maker,COUNT(\*) AS cnt FROM cars\_cleaned GROUP BY maker) AS tmp2)\*100,2) AS mshare FROM (SELECT maker,COUNT(\*) AS cnt FROM cars\_cleaned GROUP BY maker) AS tmp ORDER BY mshare DESC LIMIT 5;***

Output:



1. Does engine power impact car pricing?

To analyze the relationship of engine power with car pricing, different averages were calculated for three samples. First sample contains cars with least engine power. Second sample is cars with maximum engine power and last on is random selection of cars with engine power more than 400.

Code used to derive results:

***SELECT AVG( price\_eur) AS average\_price FROM cars\_cleaned WHERE engine\_power in (SELECT engine\_power FROM cars\_cleaned ORDER BY engine\_power DESC LIMIT 100);***

***Output:***



***SELECT AVG( price\_eur) AS average\_price FROM cars\_cleaned WHERE engine\_power in (SELECT engine\_power FROM cars\_cleaned WHERE engine\_power IS NOT NULL ORDER BY engine\_power ASC LIMIT 100);***

***Output:***



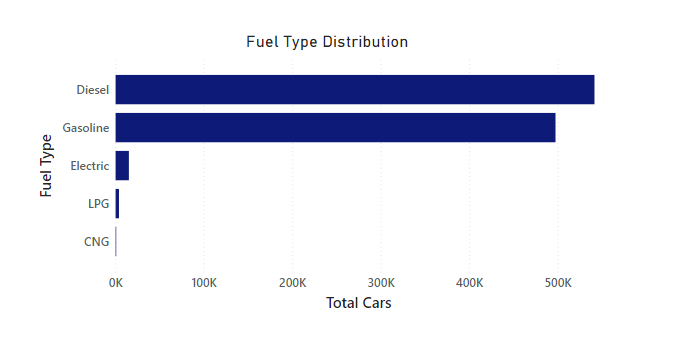
***SELECT AVG( price\_eur) AS average\_price FROM cars\_cleaned WHERE engine\_power in (SELECT engine\_power FROM cars\_cleaned ORDER BY engine\_power BETWEEN 400 AND 800 DESC LIMIT 100);***

***Output:***



Based on above results, there is no apparent impact of engine power on car prices as for maximum power engine, average car price is 25609.29 euros and for minimum power engine, it is 26162.09 which is almost equal to the first sample.

1. What are the types of fuel used by the cars?

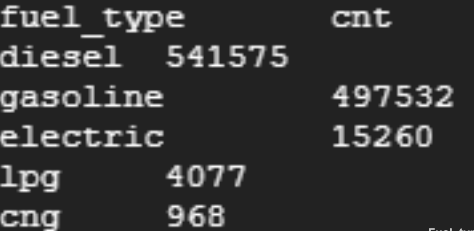


As per the data, diesel and gasoline are the two most common fuel types. CNG is the least used fuel.

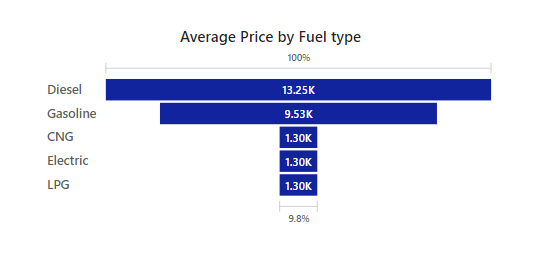
Code used to derive the results:

***SELECT fuel\_type, COUNT(fuel\_type) as cnt FROM cars\_cleaned WHERE fuel\_type <> '' GROUP BY fuel\_type ORDER BY cnt DESC;***

Output:



1. Which type of cars fuel-wise is comparatively expensive?

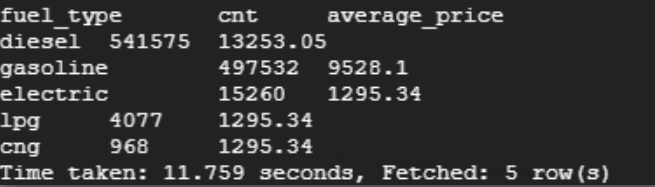


Diesel cars are most expensive among the fuels used. In the data present, diesel has an average of 13K euros. The second most expensive are gasoline cars.

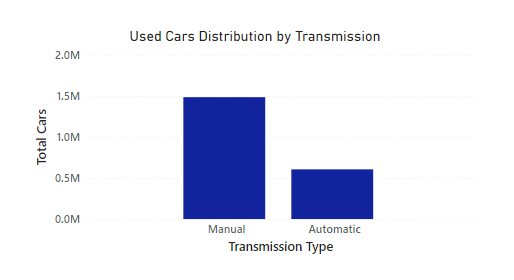
Code used to derive the results:

***SELECT fuel\_type, COUNT(fuel\_type) AS cnt, ROUND(AVG(price\_eur),2) AS average\_price FROM cars\_cleaned WHERE fuel\_type <> '' GROUP BY fuel\_type ORDER BY cnt DESC;***

OutPut:



1. Which is the most common transmission in used car market?

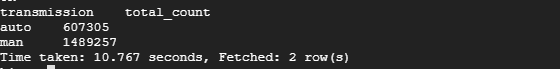


Manual transmission is more common than automatic.

Code used to derive the result:

***SELECT transmission, COUNT(transmission) AS total\_Count FROM cars\_cleaned WHERE transmission <> '' GROUP BY transmission;***

Output:



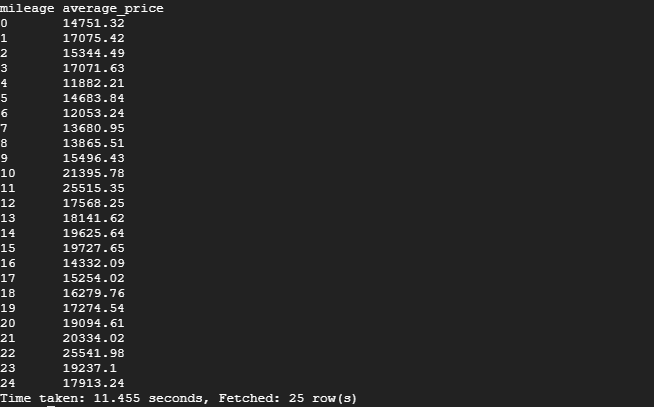
1. Does car mileage impact price of cars?

To analyze, two samples were considered. First, some data with maximum mileage were taken and average price was calculated. Second, data with minimum mileage and average price for these cars was calculated.

Code used to derive the results:

***SELECT mileage, ROUND(AVG(price\_eur),2) AS average\_price FROM cars\_cleaned WHERE mileage is not null GROUP BY mileage ORDER BY mileage ASC LIMIT 25;***

***Output:***



***SELECT mileage, ROUND(AVG(price\_eur),2) AS average\_price FROM cars\_cleaned GROUP BY mileage ORDER BY mileage DESC LIMIT 25;***

***OutPut:***



As per the results, the price is higher for cars where the mileage is lower. The price reduces with higher mileage.