AKASH PATEL

Data Analytics with Hive

# IMPORTING DATA INTO HADOOP

Kagle: <https://www.kaggle.com/mirosval/personal-cars-classifieds>

To download dataset :  
wget <https://www.dropbox.com/s/ffkwf3ixq1mjk7q/cars.csv>

hadoop fs -mkdir /BigData

hadoop fs -mkdir /BigData/hive

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

# CREATE DATABASE AND TABLE IN HIVE & LOAD DATA INTO TABLE

**CRETAE DATABASE**

CREATE DATABASE cars;

USE cars;

**Table creation named carsdata**

create table if not exists cardata (maker string, model string, mileage bigint, man\_year int, eng\_dis int, eng\_pow int, body\_type string, color\_slug string, skt\_year INT, 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 ',' tblproperties("skip.header.line.count"="1");

**Load dataset into table “carsdata”**

load data inpath '/BigData/hive/cars' into table cars.cardata;

Text

Description automatically generated

**Check loaded data**

select \* from cardata limit 5;

A computer screen capture

Description automatically generated with medium confidence

**Finding size of dataset**

Select count(\*) from cardata;

**Total 3552912 data found in table named cardata.**

Text

Description automatically generated

**Finding different manufactures from cardata**

select distinct maker from cardata where maker != ‘’;

select count(distinct maker) from cardata where maker != ‘’;

Total 46 different manufactures found in data

Text

Description automatically generated Text

Description automatically generatedText

Description automatically generated

**Finding average price of each manufacture**select maker, avg(price\_eur) from cardata where maker != ‘’ group by maker;

Text

Description automatically generated

**Finding total number of data having model and maker manufactures from cardata**

select count(maker) from cardata where maker != ‘’ and model != ‘’;

Total 2419551 having maker and model, means 31% rows missing those two values. That can be removed from dataset.

Text

Description automatically generated

# Data Cleaning

* **Null values**: Maker and Model are two key component of this dataset and without this two value rest data is not that much helpful, so here 31% of rows will be ignored which contains null values.
* **Filtering Columns**: color\_slug and stk\_year are two column which has around 94% and 50% data null. Also, these two features are not that useful when it comes to used cars. So that can be ignore and will be removed from final table.
* **Removing Outliers:** Assuming that most cars prices lie between 3K to 2000K, so rest will be removed considering outliers.
* **Filtering Data**: Most of the used cars buyers won’t prefer to buy cars which are older than 15 years, so here cars which are manufacture before 2005 year will be ignore for final dataset. Same way mileage will be filter out for values from 20000 to 160000.
* Also, car doors having more than 6 and less than 2 will be removed for door\_count as per cars specification. Same way seats count less than 2 and more than 10 will be removed for seats counts.

**Considering all above points and conversion of two fields to timestamp will be effective from below table creation.**

create table if not exists carsdata\_clean as select maker, model, mileage, man\_year ,eng\_dis , eng\_pow , body\_type , transmission , door\_count , seat\_count , fuel\_type , CAST(to\_date(from\_unixtime(unix\_timestamp(date\_created, 'yyyy-MM-dd'))) AS date) date\_created, CAST(to\_date(from\_unixtime(unix\_timestamp(date\_last\_seen, 'yyyy-MM-dd'))) AS date)date\_last\_seen,price\_eur from cardata where maker||model != '' and maker != 'maker' and door\_count between 2 and 6 and seat\_count between 2 and 10 and man\_year between 2005 and 2017 and price\_eur between 3000 and 2000000;

Text

Description automatically generated

Sample data of clean table ‘cardata\_clean’.

Graphical user interface

Description automatically generated

Average cars price by maker

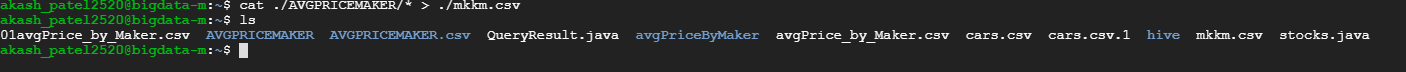
Text

Description automatically generated

Exporting csv file for visualization of avg car price by maker.

Text

Description automatically generated



Chart, bar chart

Description automatically generated

# Data Analysis

Relationship between car maker, model and price.

Text

Description automatically generated

From the above screenshot, its clearly visible that price of cars directly connected to maker of cars, for example Audi, Lexus, BMW, jaguar, Lamborghini all these brands car value is higher compared to other car makes, also these brands are known for luxuries cars. While Kia, Toyota, Nissan falls under economic car brands which are comparatively affordable to common people.

The reason behind high price for luxuries car brands is safely and features which makes these cars more comfortable and costly.

Cars which are expensive are luxuries, supercar, or larger and high-end cars regardless brand of car. Doesn’t matter which manufacture produced that car. This can be seen in above screenshot and below chart.

Highest price is for Lamborghini Aventador.

Text

Description automatically generated

Average car price for small cars is low as around 3K to 4K. Cars such as lupo, streetwise, accent, lybra are low-cost cars and more affordable.

Top car manufacture from dataset.

To fetch details about top car manufactures from dataset, first need is to decide base for it. Such as what factors should be considered to find out top manufactures.

Every person who buys a car or used car is more conscious about price or car, mileage of car. In data cleaning stage dataset is filter out for car price between 3000 to 2000000.

While buying a used car, most people look for how much car has been used and what mileage it gives and then what price is offered for this price. And what kind of damage has been done to the car.

To find out top manufacture all these factors should be considered and here we don’t have car damage data so here it will be decided based on car mileage and car price.

Text

Description automatically generated

From the above data we can see that top manufacture for cars under 50000 are lotus, tesla, infinity, isuzu, aston-martin, Maserati, jeep, Porsche, jaguar, Bentley.

Text

Description automatically generated

From the above data we can see that top manufacture for cars under 50K-100K are Rolls-Royce, Mitsubishi, Infinity, Seat, Chrysler, Subaru, Lotus, Jeep, Dodge, Volvo.

Text

Description automatically generated

From the above data we can see that top manufacture for cars under 100K-200K are Kia, Nissan, Dodge, Jaguar, Lancia, Rover, Lexus, Maserati, Peugeot, Mini.

Relationship between fuel type and car price

To find out relationship between these two, need to check for average price of each fuel type cars.

Text

Description automatically generated

We can see the difference between Average price of diesel and gasoline, almost 3000 difference.

To find out further, here all cars average price by fuel\_type and manufacture year.

Text

Description automatically generated

From above chart, it shows that by years price of cars increase regardless fuel type. But till 2012 car price of diesel and gasoline was almost same but then after 2012 one can see difference in prices till now. Also in the year 2016 gasoline car were cheaper than diesel cars and same pattern found for year 2015.

# Conclusion

To analyse data of used car and for this report Hive and Hadoop is used, all queries are performed with the use of HIVE and result of queries exported from Hadoop and Visualization done in Microsoft Excel.

To perform analysis on such a huge dataset contains of almost 4 million records, Hadoop is best tool and for optimization of time and get rid of programming Apache Hive is best suitable tool. First of all need to load this dataset into Hive by creating database and table. After loading data, find out about outliers and missing data and based on its cleaning has been done on this dataset.

Based on all analysis and queries and market research, used cars business is growing for all kind of cars luxurious, economy, large, small and to invest in this area at this time could be a best decision.