Analysis on Used car dataset using Big Data Queries and Visualization

CIS 5200: System Analysis and Design by Prof. Jongwook Woo
Contributed by Heta Parekh, Priya Ramdas, Savita Yadav, Sonam Suryawanshi, Vijay Muthupillai
Email Ids: hparekh2@calstatela.edu pramdas2@calstatela.edu syadav5@calstatela.edu ssuryaw@calstatela.edu
vmuthup@calstatela.edu

Abstract: In today's era, owning a car is very important-since it provides the opportunity to travel long distances due to a lack of public transport specially in densely populated cities. But not everyone can afford a brand-new car. In most of the cities, people have started buying vehicles which were used prior. To give a more preference-based car as per the needs and demands of consumers; an analysis has been prepared. This project uses Kaggle's used car dataset and performs analysis and visualization with the use of Hadoop and Hive.

1. Introduction

Used car dataset holds information related to the car design based on model type and brand with the seller rating *(Figure 5.1) a*nd description for the same. This paper contributes the analysis by inventory of cars in terms of make and model, geological location and price-range varying with brands *(Fig 5.3*). N-gram analysis has been performed based on the car description *(Figure 5.2);* paper contributes analysis done for car takes a longer time to selling as well as *c*lassification of car makes and model on the terms of accidents *(Figure 5.5 and 5.7)*. To get accurate results-big data tools, methods and various technologies are implemented in this term project. The dataset was collected as the used car dataset from Kaggle, and it consists of 3 million car records from the year 2005 to September 2020. The data set was massive to store and thus it needs to process with the framework of Hadoop to access the data. The dataset was first loaded to amazon s3. Using *wget,* the data is downloaded and uploaded to Hadoop where Hadoop Distributed file system (HDFS) is used to store the data and MapReduce to process the data. The data was further processed in Hadoop's ecosystem –Apache Hive by starting the Beeline and connecting it to the Hive server in to process and analyze to data draw insights.

2. Related Work

A related study and analysis "*Statistical Analysis on Second hand vehicle sales using Big data Technologies and Linear Regression"*[1] was performed in 2019 where they used Hadoop for processing the data and Python Programming language for performing regression analysis and building a model that predicts the price of used cars. Their dataset consists of 1.7 million records of car sales data from the year 1900 to 2019. In their study, they mainly focus on trend of used car sales over the years, State wise distribution of car sales and which manufacturer is the most trusted in the second hand vehicle sales market. Whereas, we have used hive to manipulate the data and built queries which are then used to analyze data and get useful insights and visualizations.

Related analysis on "Used car Price Analysis" [2] was a part of the Capstone project was performed in Jupyter Notebook. The major goal of their project was to build a model that decides if the asking price for a certain car is reasonable given the information provided in the listing. And they have provided code that will be useful in finding recommendations for related vehicles with a lower price, lower miles, and one that is slightly more expensive.

Similar analysis has been performed by Luc Frachon [3] who worked in automotive industry for 12 years. The dataset he used for analysis was found on Kaggle, the well-known Machine Learning Competition website. His analysis was focused on Used Cars Dataset to predict the price for the used cars by comparing features like date of manufacture, milage on meter. Analysis has the brand category variable that turned out to be quite useful as well as it exhibits a significant association to price. The brand and model variables tell us a lot about the German market, one of the most high-end market in Europe. Using the exploratory analysis, analysis built a linear model that explains 76.5% of the variance in the data. We did similar analysis by comparing card makers by models and the number of units sold in a particular time frame. This analysis compared the car makers, how much time it took to sell in consolidated days range to 0-91 ,91-182,182-365,365-above.

3. Specifications

Table 1 Data Specification

File Size	9.29 GB
No. of files	1
File format	CSV
Years analyzed	2006-2020
Country	USA
No. of records	3.0 Million

Table 2 H/W Specification

Cluster Version	Oracle Big Data Compute Edition
Number of Nodes	3
Memory size	180 GB
CPU	12 OCPUs
CPU Speed	2.20 GHz
HDFS capacity	147 GB
Storage	957 GB

4. Background

Apache Hadoop is an open source and consists of Linux based structure of tools. It is widely used to run applications for big data. Big data contains a massive amount of data in terms of high volume of unstructured data, Velocity at which data is performed and Variety of different data types. It has two vital components – MapReduce to process the data and Hadoop Distributed File System (HDFS) to store the data. There are two types of nodes in HDFS – data nodes where data is stored in replicated file blocks and the name node which forms a relation between the client and the data node. JBDC client such as beeline is the primary way to access Hive. Hive is a SQL-based data warehouse system for Hadoop that opens the door for data summarization, ad hoc queries, and the analysis of massive data sets stored in HDFS. (*as shown in Fig 3.1)*

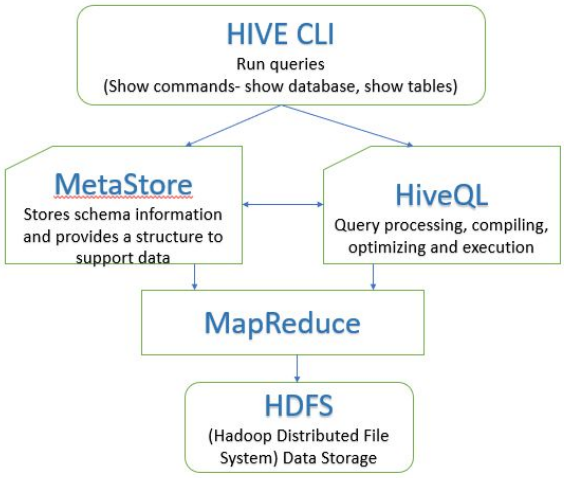

Figure 3.1 Hive Architecture

5. Architecture Workflow

For the project, we extracted the data source from Kaggle and staged in the Amazon AWS site. Then, we uploaded the dataset and used Hive to trigger queries and create tables based on the use-cases (Section 5). For the visualization, we used Tableau, SAP, PowerBI , Excel and to end we found the insights which can be useful for consumers to purchase used cars.

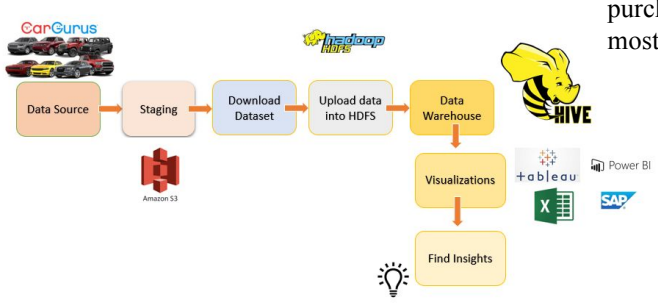

Figure 4.1 Architectural Workflow

6. Visualizations and Insights

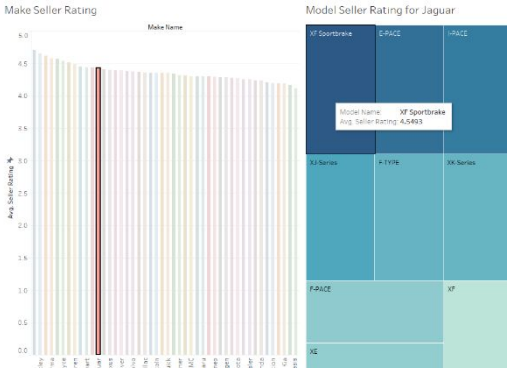

Figure 5.1 Seller rating of car brand and its models.

The above visualization signifies the seller rating of all car brands and its' models. Top 15 car makers having highest seller rating above 4.0 in last 5 years are 'Lotus', 'Bentley', 'Ferrari', 'Karma', 'Lamborghini', 'Rolls-Royce', 'Porsche', 'McLaren', 'Aston Martin', 'smart', 'MINI', 'Jaguar', 'Tesla', 'Lexus', 'Audi'.

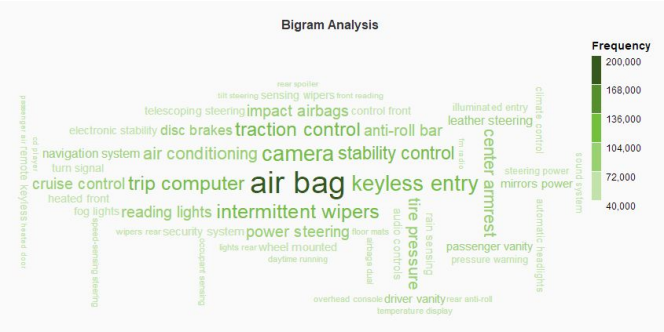

Figure 5.2 N-gram analysis of the top selling cars

This tag cloud represents bi-gram analysis of the description for the cars with higher seller ratings *(shown in Fig 5.2)*. Based on the analysis, it can be inferred that features, and technology are vital factors to purchase a car. Among the features and latest technology – most are likely to be safety and comfort related.

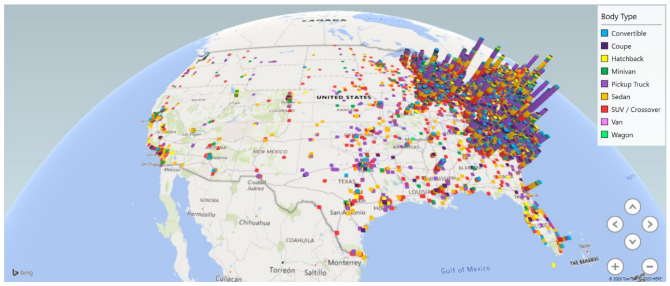

Figure 5.3 Inventory of cars by geospatial mapping

The geospatial map gives a realistic view of the data. It represents the most popular body styles across the US (SUVs, sedans, and pickup trucks). In general, the east coast of the US contains more listings than the west coast. SUVs dominate in the northeast region. Sedans and Pickups are popular on the east coast and parts of the Midwest regions like Michigan, Ohio, and Indiana. Overall, it will give the consumers a better idea of where to find reliable car dealers, and the dealers will have a view of where to bring their next car dealership in a particular region.

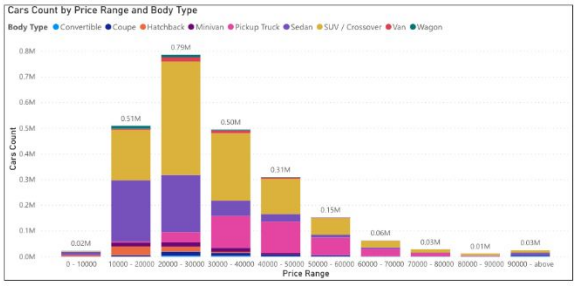

Figure 5.4 Inventory of cars based on body type and price range

The above visualization represents the inventory of cars based on body type and price range for the past 5 years. The price range is distributed by every $10,000 and the most used car body types are Convertible, Coupe, Hatchback, Minivan, Pickup Truck, Sedan, SUV / Crossover, Van, Wagon.

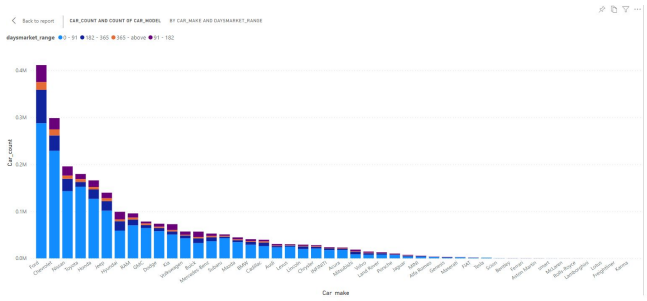

Figure 5.5 Car make & model takes a longer time to selling

Ford is the largest market player in the US Car Market and same belongs to resale car market sale also. From the Graph it shows, car maker able to successfully resale their models in similar fashion as their new car sales. There is no major gap which shows like any one of the makers is lagging in resale. Still there are few points which we can raise like Toyota which is a Japanese maker doing good in their resale market comparer to Nissan and Ford. If we compare German car makers Audi, BMW, and Mercedes we can see from data.

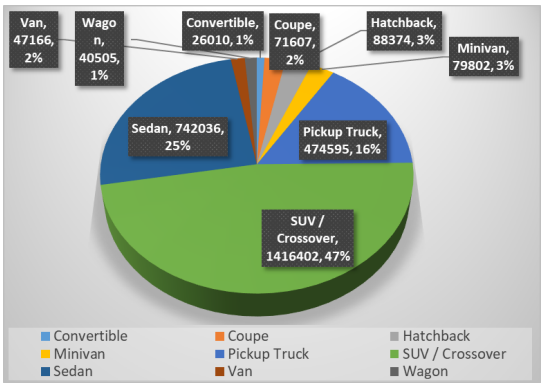

Figure 5.6 Inventory of cars by body type

This analysis is intended to find out the composition mix of Used Cars by body type in the CarGurus inventory of 3 million used cars. SUV / Crossovers dominate the inventory with close to 50% share which was expected as most Americans prefer bigger cars. However, the surprising data point to me is the higher Pickup Truck count than the Minivans. Offline analysis revealed that most American families in towns, farmlands and suburbs prefer Pickup Trucks for their day to day activities than the Minivan. That's an Interesting fact...

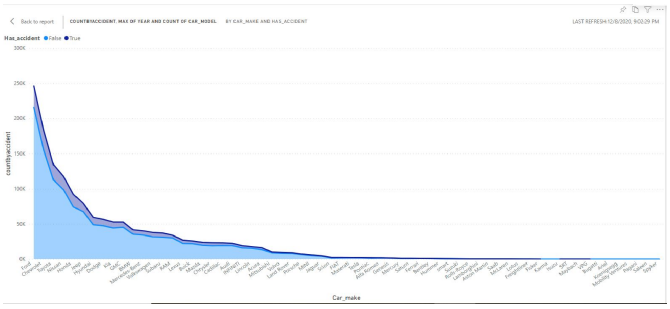

Figure 5.7 Inventory of cars having more accidents

Following Graph (*shown in Fig 5.7)* represents which car has more accident. We have compared between cars make such as 'Lotus', 'Bentley', 'Ferrari', 'Karma', 'Lamborghini',

'Rolls-Royce', 'Porsche', 'McLaren', 'Aston Martin', 'smart', 'MINI', 'Jaguar', 'Tesla', 'Lexus', 'Audi', how much time it took are to sell, we have also calculated how many models each car makes have. When we compare the accidents data on the roads, then we can easily see number of cars has accident is more from Ford and other players which are selling more cars. Accidents depends on driver's mindset mostly and the conditions of infrastructure, weather. We can see in US all these three factors are mostly constant and there is no difference in people if they are driving Ford or Ferrari. Has Accidents is in same proportion as the number of cars on the roads from the manufacture.

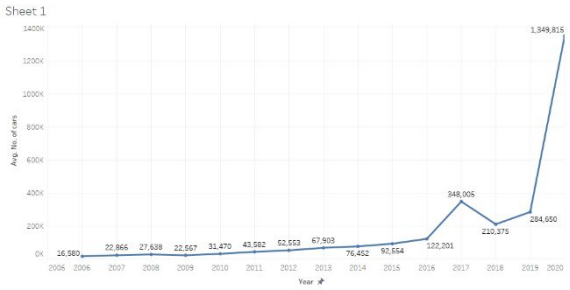

Figure 5.8 Inventory of cars by year

In the above visualization, it can be inferred that highest number of cars have been listed in the cargurus inventory for the year 2020. We can make an assumption here stating because of the Pandemic, many users wouldn't feel the need for using the car and hence they sold it off to third party. I.e cargurus in this case.

7. Conclusion

This project uses Kaggle's 'used car' dataset to perform analysis and visualization with the use of Hadoop, Hive and other technologies to help dealers, manufactures and consumers to understand the real value for the 'used cars' and growing market on the same. From the above analysis, we can conclude that safety is the key feature when it comes to 'used cars' and user can also look into other analysis to find out which car has high rating, number of accidents based on car models and accordingly purchase cars based on their price range.

References

[1] https://www.slideshare.net/YashIyengar/big-data-analysis-of-second-hand-car-sales
[2] https://github.com/clrife/CarPriceAnalysis
[3]https://rstudio-pubs-static.s3.amazonaws.com/248952_706edc85cfa84a369dfe401a763d32fc.html
https://github.com/MRPriya/UsedCarSet (GitHub)
http://usedcardataset.s3.amazonaws.com/archive.zip (Amazon s3)
https://www.kaggle.com/ananaymital/us-used-cars-dataset (Dataset)