CIS 4560 – Used Car Dataset Term Paper

Team 4 – Jesus CortezBonilla, Joshua Rowill Koa, Kevin De La Torre,
Bryan Mendoza, Samuel Mendoza
Department of Information Systems, California State University
Los Angeles

Abstract: The research of this documents helps explain the usefulness of the US Used Car dataset in analyzing and comparing various makes, models, and listed prices. Other qualifying information such as location gives us insight on where these cars are located and price differences. Various tools such as Hadoop, Tableau, SAP Analytic Cloud are used to analyze the given data. All together this helped in Identifying overpriced cars, classifying cars, and finding dealers who will generate the most savings.

1. Introduction

Our team based our project on queries, charts, and Geo Mapping because we wanted to get a better understanding of popular used cars and areas of concentrations of these used cars. For someone looking to buy or sell a used car, this data provides answers where there is a large assortment of used cars and places that have a quick turnaround by the listed data and days on the lot. Though word of mouth or recommendation can be useful in finding dealerships, this data, turned into information can give us statistics popular used cars and where one is able to purchase it. We were able to identify deals in the market and guide buying decisions, identify overpriced cars, generate information to evaluate dealers who will generate the most savings, and classify cars based on mileage, brand, and days on the market.

2. Related Work

The dataset that we analyzed needed to be filtered much like many other databases. Often, there is more information than is needed, which can cause problems when generating reports based on data. One study suggested that Machine learning could calculate fair prices for used cars. Contrary to our methods, it proved to be a question of what is being sold. If you're selling or buying one car, our data would help in located places where you will be able to save money or the better region to sell your car. In their report, "The answer is definitely yes if you have hundreds or thousands of vehicles to sell" [1]. Machine learning will can help in answering much larger questions than our method.

In recent years, the forecast of used car prices has assumed high significance. Consider an analysis done by Bryan Whiting [2] on his search to find the best deal on a new used car and how data can teach you a lot about an industry in a very limited amount of time. His procedure involved web scraping thousands of used cars from Truecars.com. The idea was to find a search query on Truecars that he liked and use html nodes to parse out the data he wanted. Afterwards he integrated linear regression to further clarify his data and then observed the intercepts. Through this framework he was able to compare make-model differences and scale mileage in order for easier interpretation. The researcher suggests that with such a model, you can identify deals in the market and help guide buyer decision. Analysis of a large volume of data by a linear regression can predict value of cars and evaluate great deals.

Luc Frachon has suggested studies on car price forecasting, which is one of the popular research topics in the field. It notes that it is a challenging task to construct a highly accurate and consistent prediction model where the number of distinct values must be taken into account. As the dataset was obtained from similar sources, Kaggle, it was interesting to compare the results against each other. Frachon established that engine power had the highest connection to price among continuous variables. "It was important to look at mileage because it revealed two distinct price similarities, one for new cars and one for vintage" [3]. Unfortunately for us, mileage, a large factor was not really a present contributor in this dataset due to the limited values. It would be useful to equate this dataset to used car data from other sources in order to see whether the model built here would give us the same degree of precision.

3. Background/Existing Work

Kaggle offered many different data sets regarding used cars. However, many of those data set were small in compared to the one that we used. The data set in question not only tower in size in comparison to the rest, had a clear purpose to be used by the public. The used card data set was analyzed, filtered, and aggregated for use. It consisted of 66 columns and over 3 million entries. This gave us various ways to approach our research. The data comprised of detailed parts of a vehicle including leg room, bed height, bed type, body type, brand name, fuel economy, and much more. Each column served a purpose in order to categorize the cars and their characteristics.

The data set also included some key attributes such as longitude and latitude along with names of dealership where used vehicles are sold. This helps in locating the exact coordinates of where these used cars are located around the United States. This data allowed us to translate the data into information that can better suit the public in buying and selling used cars.

4. Implementation (Your Work)

Specifications 4.1

For the calculations portion, most were done using a provided Hadoop cluster. This consists of an EMR cluster of 3 Nodes with version 20.3, 12OCPUs of Intel® Xeon® CPU E5-2699C V4 with a speed of 2.20GHZ, 180GB of

memory and a Storage capacity of 957Giga Bytes. The spark version which was used is 2.1 on hive.

The Usedcars dataset comprised on one 2 Giga Byte compressed zip file. The uncompressed file contained data of 9.29 GB on one CSV excel workbook. This workbook included 66 columns with over 3 million entries that had been gather by an online crawler that searched the web for data on used cars, vin numbers of the cars, locations, car attributes, and used car dealerships.

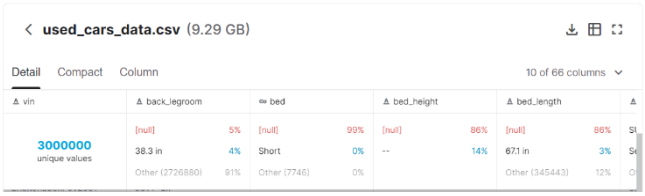

Figure 1. UsedCar Dataset from Kaggle.com

In uploading the csv file, we utilized the window command prompt as follows:
Pscs kdelat15@129.150.64.74 being the username and server address. :/home/kdelat15/UsedCarsData being the destination pathway to where the file was upload.

```
ws [Version 10.0.19041.685]
oft Corporation. All rights reserved.

pscp kdelat15@129.150.64.74:/home/kdelat15/UsedCarsData
```

Figure 2. Using pscp to upload a file to Hadoop server.

After the files have been uploaded, we transitioned them to hive in order to create tables. By creating tables In hive, it allowed use to filter unecessary data that was either null or allowing us to create queries. Some of the fields were litter with commas which created many problems along the way in analysing the data.

4.2 Creating Tables

By filtering out and creating tables, we essentially removed any columns that would slow down the processing and unassociated to the analysis. This also let us extract the main information that was necessary in order to better clarify our data. On a further note, there was a column whose fields contained many commas which kept giving us errors in establishing tables. Figure 3 illustrates the code that was used in hive in order to create one big table where we would be able to call on queries to make use of the data in questioned. We combed through the 66 columns to see what we would be able to use.

Figure 3. Tables created and used in Hive.

4.3 Data Flow

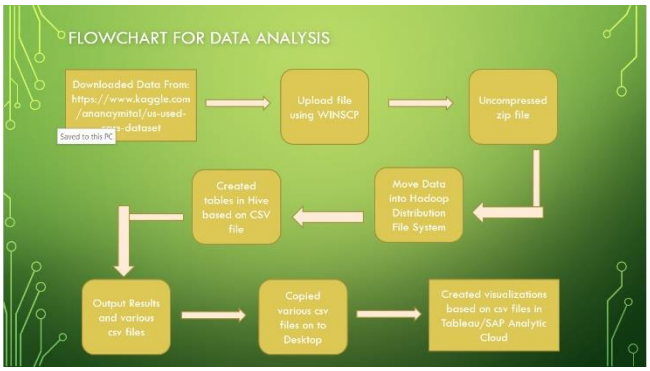

Figure 4. Data Flow Chart

Figure 4 illustrates the entire process of our project. It began by downloading the used cars data set from the Kaggle website. From there we uploaded the zip file through pscp script or through WINSCP. Either method works, however, WINSCP allows for a visual of the upload content.

Once the data had been uploaded, it needed to be uncompressed for usability. We were able to uncompressed the 2-gigabyte zip file into a 9-gigabyte csv file using Linux commands. Once uncompressed, the file was moved away from the local and into the Hadoop server. Of course, if a copy remained in the local server, it was deleted.

Next, the file was transition into Hive for the purpose of creating tables in order to be usable in our data analysis. There were unnecessary fields that were with left blank or unusable for our data analysis and unnecessary columns that did not fit in our analysis. By the end of the filtering we had enough clean data to use for the project. From here we were able to get enough queries and output various csv files to generate charts, maps, and geo spatial visualizations in Tableau, Powerbi, or SAP Analytic Cloud. The visualizations were generated in order for us to show the differences and representation of our results.

4.4 Analysis & Visualization

To begin the analysis, we opted to find out which were the ten most popular used cars for sale throughout the United States. After determining the most popular cars we could then gain insight on which cars were being sold the most. Alternatively, we also analyzed which were the most popular car types that were being sold along with the average mileage and price point each vehicle was being sold at based on its year of manufacture.

Once our queries were analyzed and finalized we went ahead and stored them in our HDFS by using INSERT OVERWRITE DIRECTORY and storing it under the ‘usedcars’ folder. We then used the –GET command to move the file to our local directories while also saving them as a .CSV, afterwards we used psftp to download the .CSV files to our local machines to be able to input them into Tableau and PowerBi so that we could visualize our data.

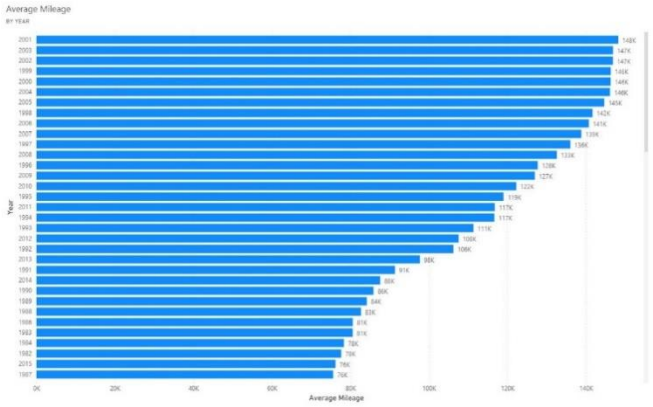

Figure 5. Bar Chart Average Mileage by Year

In this bar chart (Figure 5) we can see the average mileage of used cars for sale by year. The year with the greatest average mileage is 2001(148K). The second year with most mileage is 2003 (147K) followed by years 2002 (147K) , 1999 (147k), and 2000 (146K). As we can see these years are from the same period of time. Cars from this era are generally considered to be more reliable so it makes that these years have the highest average.

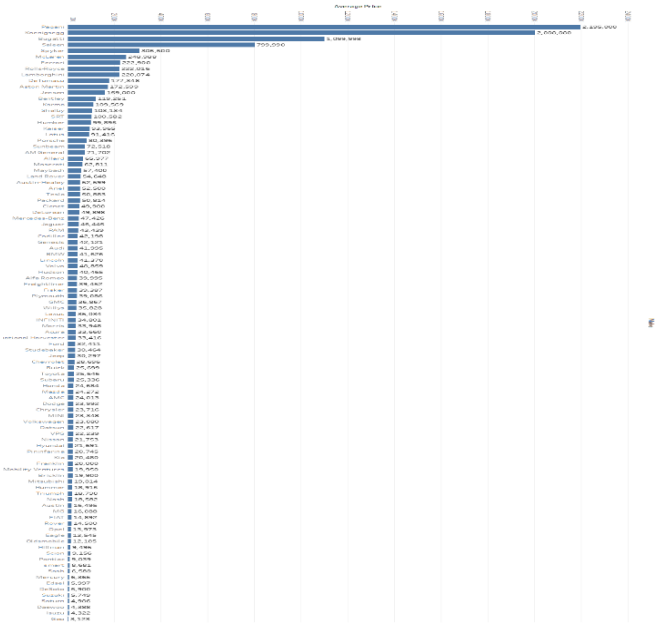

Figure 6. Bar Chart Average Price by Make

This column chart (Figure 6) shows the average price of used cars for sale by make. The most expensive make for sale is Pagani with an average price of 2,195,000 dollars. Pagani is followed by Koenigsegg at an average price of 2 million dollars. Third is Bugatti with an average price of 1,099,998 dollars. These three makes are known for making very expensive and high-end sport cars. The three car makes with the lowest average selling price are Daewoo ($4,388), Izuzu ($4,322), and Geo ($3,123). These three brands are known for making cheap economy cars which would explain their low average prices.

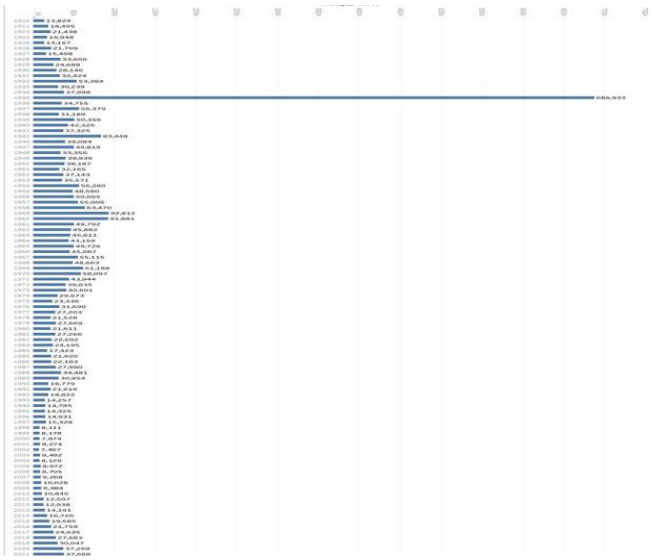

Figure 7. Bar Chart Average Price per Year

In this column chart (Figure 7), we can see the Average Price by Year. The year with the greatest price is 1935 ($686,933). This year can be considered an outlier as it the average price for this year is significantly greater compared to the rest of the data. It might be that this car was a very expensive antique car and the only car for sale with the model year 1935. The year second highest average was 1959 ($92,812) followed by the year 1962 ($91,881). These two years may have cars that are old enough to be expensive antiques but may also have less expensive cars bringing the average down. The three years with the lowest average price are 2002 ($7,467), 2000 ($7,874), 1998 ($8,111). We can see some overlap with Figure 1 cars from this time period on average have the highest amount of mileage. This would explain why cars from this year would have the lowest average prices.

As the idea is to identify great deals in the market, we also wanted to evluate dealers who would offer us the most savings. Therefore we calculated the average savings per seller. The chart below illustrates just a portion of the results as we limited the query to 1,000 rows and generated an output with a size of 40 kilobytes.

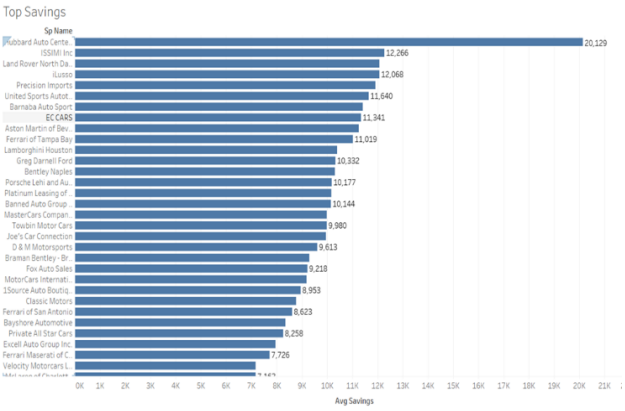

Figure 8. Bar Chart Average Savings per Seller (Top)

The results displayed by (Figure 8) shows us that Hubbard Auto Center of Scottsdale is at the very peak of Average Savings amounting to about an average of $20,129 dollars in savings. Followed by ISSIMI Inc and Land Rover North Dakota with a solid $8,000 gap. We also concluded that the sweet spot for what dealers were willing to lose out on to save you money was aroudn 3,300 dollars. Based on this graph you can certainly observe a lot of disparities between sellers as not many share similar savings for the customer.

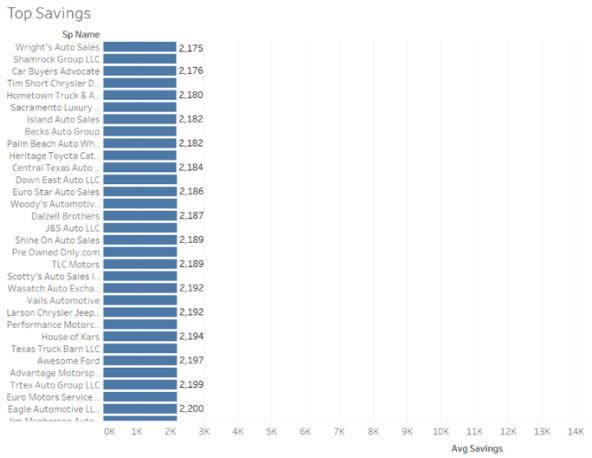

Figure 9. Bar Chart Average Savings per Seller (Low)

We also decided to look into the bottom half of the output generated by our query. As you may see above (Figure 9) this chart shows that Wright's Auto Sales generated the least amount of average savings starting at $2,175 dollars followed by other delearships showing minimal disparities in savings. This tells us that many of the dealers did not show a wide interest in standing out in regards to saving the customer some money.

5. Conclusions

The Hadoop and its Hive features have been used in this analysis to demonstrate the simplicity of operating large datasets in a certain environment. In this paper, a study was carried out on the data collection of used cars that would help us gain a deeper understanding of the different variables and their importance on the metric of price. With the assistance of Hive we were able to create queries and analyze results to help identify deals in the market and guide buyer decision. Although looking towards the future we may try to find ways to further improve our analysis, many of which include doing diagnostics to check for errors that are normally distributed. We may also check for outliers to prevent their affects on the accuracy of our results. Checking for interactions such as the change in price by cars on average may also be a focus for the next project. Ultimately, a car is purchased by irrational people, and sold by more experienced ones as well. So when you decide to make a deal, its best to review all information available to you.

References

[1] Lepchenkov, Kirill. Figuring out a Price of a Used Car in a Data Car in a Data Science Way, 12, Dec 2019. https://towardsdatascience.com/figuring-out-a-fair-price-of-a-used-car-in-a-data-science-way-11450b3b252b

[2]Whiting, Bryan. Buying a Used Car the Data Science Way (and Hacking Truecar). 10 Feb. 2018. https://www.bryanwhiting.com/2018/02/buying-a-used-car-the-data-science-way/

[3]Frachon, Luc. Used Cars Dataset – Exploratory Analysis. 19 Jan. 2017. https://rstudio-pubs-static.s3.amazonaws.com/248952_706edc85cfa84a369dfe401a763d32fc.html

[4]Kaggle.com Ananay Mital Sep, 2020 https://www.kaggle.com/ananaymital/us-used-cars-dataset

[5]Github.com, 2020 https://github.com/Smendo105/CIS4560