CIS 4560 Term Paper Used Car Data from CarsGuru

**Authors:** Andrew Rosalino; Andrew Tran; Edgar Manukian; Thomas Lew; Brian Sangalang

**Abstract:** We will use this dataset to determine the most popular used car. We can separate the inventory of each geo graphic location to determine the highest density of used cars around the United States. We will also generate insights as to whether the drive type, brand, mileage, or even fuel efficiency has a large impact on purchasing decision. The dataset also includes longitude and latitudes so we can generate 3d maps along with other visualizations via Power BI.

**1. Introduction**

The paper covers the process of analyzing the Used Car Data from CarsGuru. The process involves obtaining the dataset from Kaggle[[1]](#footnote-0), utilizing Hadoop, and analyzing the data with PowerBI and Tableau. The dataset consists of the inventory of used cars from CarsGuru in September 2020.

The team has chosen this dataset as the market of used cars is interesting due to the COVID-19 and how people are buying cars. Some reasons may be factories shutting down or running in limited function, people are not buying new cars, or using other means of traveling such as public transit or ride sharing. The data of used cars is interesting to determine where the affordable used cars are, determine the best used car brands, and see where the most inventory for used cars are located.

**2. Related Work**

Used cars are bought and sold everywhere, by many different people and many different businesses. Reports from many different sources and companies are readily available regarding performance in the business of used cars.

In a report from the National Independent Automobile Dealers Association (NIADA)[[2]](#footnote-1) they report all of the members' (used car dealers) inventory and statistics. They use surveys of each dealer to compile useful graphs showing information such as how long vehicles are in the inventory before they are sold, how many vehicles are in the inventory, and the average price of each vehicle in their inventory. This information is intended to be used by used car dealerships in order to compare their own performance against the bulk of other businesses in the same industry.

Cox Automotive, a company that owns several brands of used car dealerships, put out a report on November 18 regarding the prices and supply of used vehicles. In this report[[3]](#footnote-2), they are able to visualize the effect of a surge in inventory in the year 2020, and how it leads to a correlation between average price and number of unsold vehicles. Also in this report they were able to show other useful information regarding their dealerships, such as days in supply and the average mileage across the vehicles in their inventory.

The reports listed above use surveys of their dealerships in order to provide data from which their analysis is made from, which is then used to make custom graphs to put into their reports. The difference between the above reports and our analysis of data from CarsGuru is that we took a dataset from Kaggle, a community focussed on data science and machine learning, which we then manipulated in Hadoop and Pig. We then used our newly created .CSV in PowerBI and Tableau to make professional level charts and graphs from the data.

**3. Specifications**

The dataset contains around 3 million used car details obtained from the Cargurus inventory. The dataset is one .csv file, eight gigabytes uncompressed and two gigabytes compressed. There are about 66 columns in the dataset with 32 of them being strings, 10 of the columns being boolean, and eight being decimals. Other columns values range up to 16. In the dataset there are about 3,000,000 unique values for *vin* or vehicle identification number.

The table below shows the hardware specifications the team used for the project. The cluster version the team has worked with is version 2.7.1. The number of nodes the team utilized is three. The memory size is 180 gigabytes.

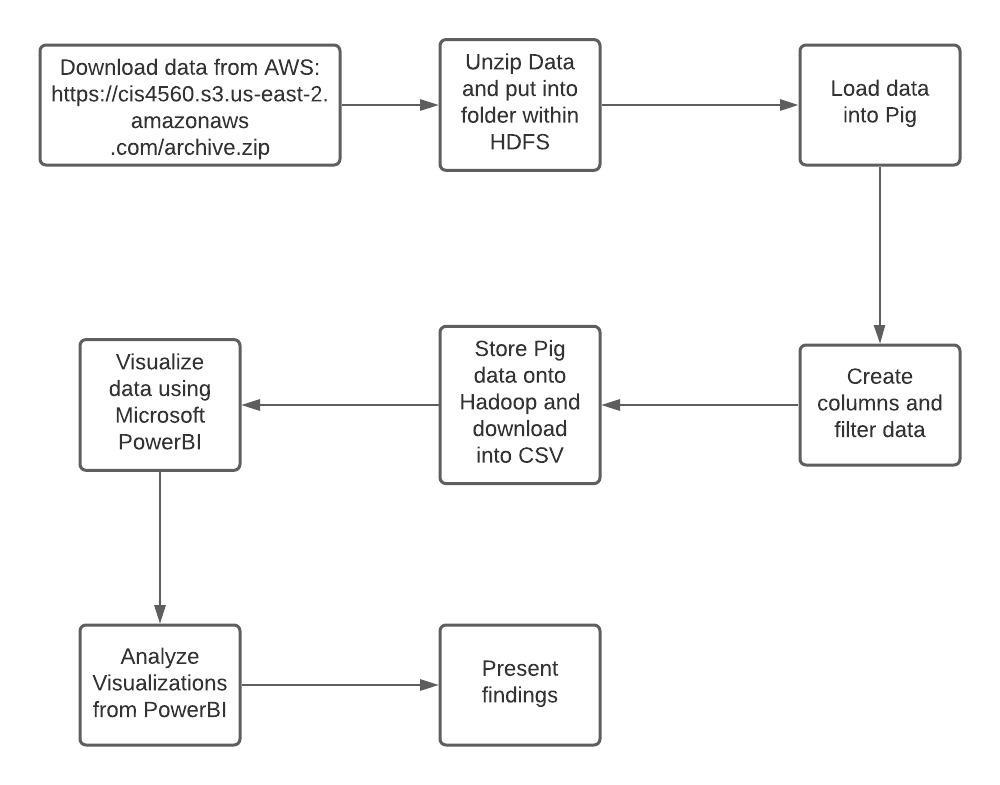
***Table 1 - Hardware Specifications***

|  |  |
| --- | --- |
| Cluster Version # | 2.7.1 |
| Cluster # of nodes | 3 |
| Memory Size | 180 GB |
| OCPU | 12 |

**4. Flowchart**

The initial raw data included details that were not necessary for analysis of the CarGurus inventory, such as the height and length of the vehicle, and the amount of legroom the back seats have. After the data was downloaded and unzipped, it was loaded into Pig and the unnecessary information was filtered out. After a workable CSV file was created, it was downloaded and opened in PowerBI and Tableau to be analyzed. Several visualizations including a geo-map and pie charts regarding the data were created so that the data could be presented in a clean manner. The flowchart can be seen below.

***Figure 1 - Flowchart***



**5. Data Cleaning**

The files are uploaded and stored in HDFS and the data is loaded into tables utilizing beeline. The team has started the shell terminal and downloaded the archive. After unzipping the archive we created a folder to store the dataset. The team ran beeline and connected the team called upon the database in Hive. The team created a car data table and tested the data to see if the process was successful. After running SQL queries testing the dataset the team has used pig to store and clean the data.

The team entered pig in the terminal and loaded the data and defined the columns. The information was cleaned and filtered to show relevant columns. The columns in pig that were shown are *vin* which is the vehicle identification number which serves to identify the specific automobile. The *make\_nam*e to track the brand and the *price* of the vehicle with the *city* for location. Some of these values will help provide the longitude and latitude for our analysis knowing the location of the city and tracking how many used cars are in the inventory. After filtering the data, it was tested to see if the process was successful.

Columns the team has removed are not relevant to our analysis. Values such as *back\_legroom* identify how much legroom is in the rear seat. The *bed* measure is a category of bed in a pickup truck with other strings following it which are *bed\_height* and *bed\_length* which are the height and length of a bed in a pickup truck. The *body\_type* involves what type of car it presents. The type of car meaning it can be a convertible or other nine body types in the dataset.

After testing the filtered data in pig we stored the newly created filtered\_car\_data.csv file and uploaded the dataset into PowerBI and Tableau. The dataset was present in Hadoop and downloaded using SCP.

**6. Analysis**

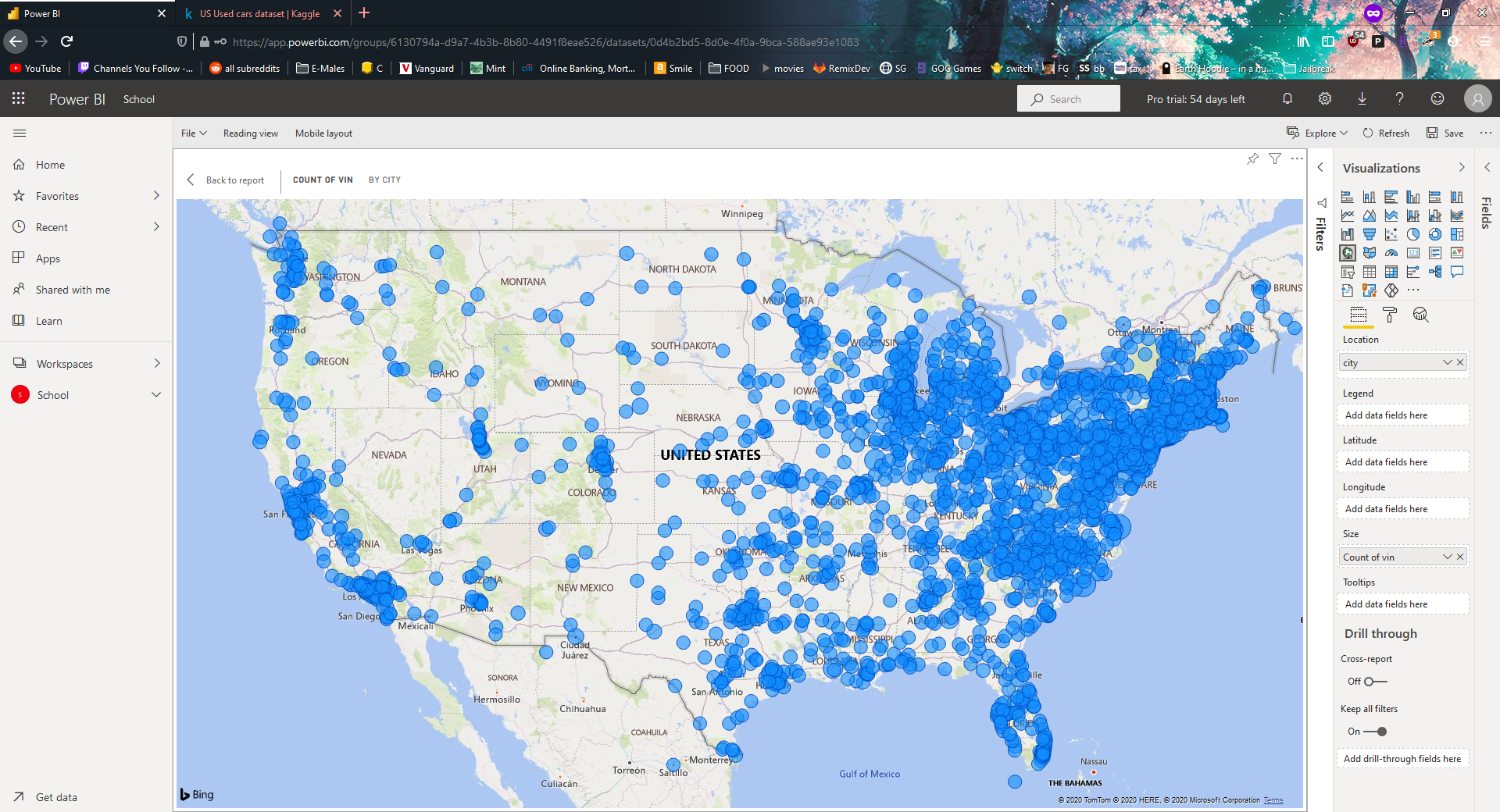
After cleaning the data for analysis the file was extracted in a .csv file. The team then made several charts and interactive maps of the dataset. The graphs were created utilizing PowerBI and Tableau. The analysis measured the various car brands by mileage, price, and location.

For PowerBI there are two visualizations the team has utilized. There is a geographic map presenting the areas of used cars from CarsGuru. The next figure is a bar chart by descending order from the amount of used cars from a certain brand.

The Tableau charts consist of bar charts comparing the various columns which are price, mileage, and the brand of the car.

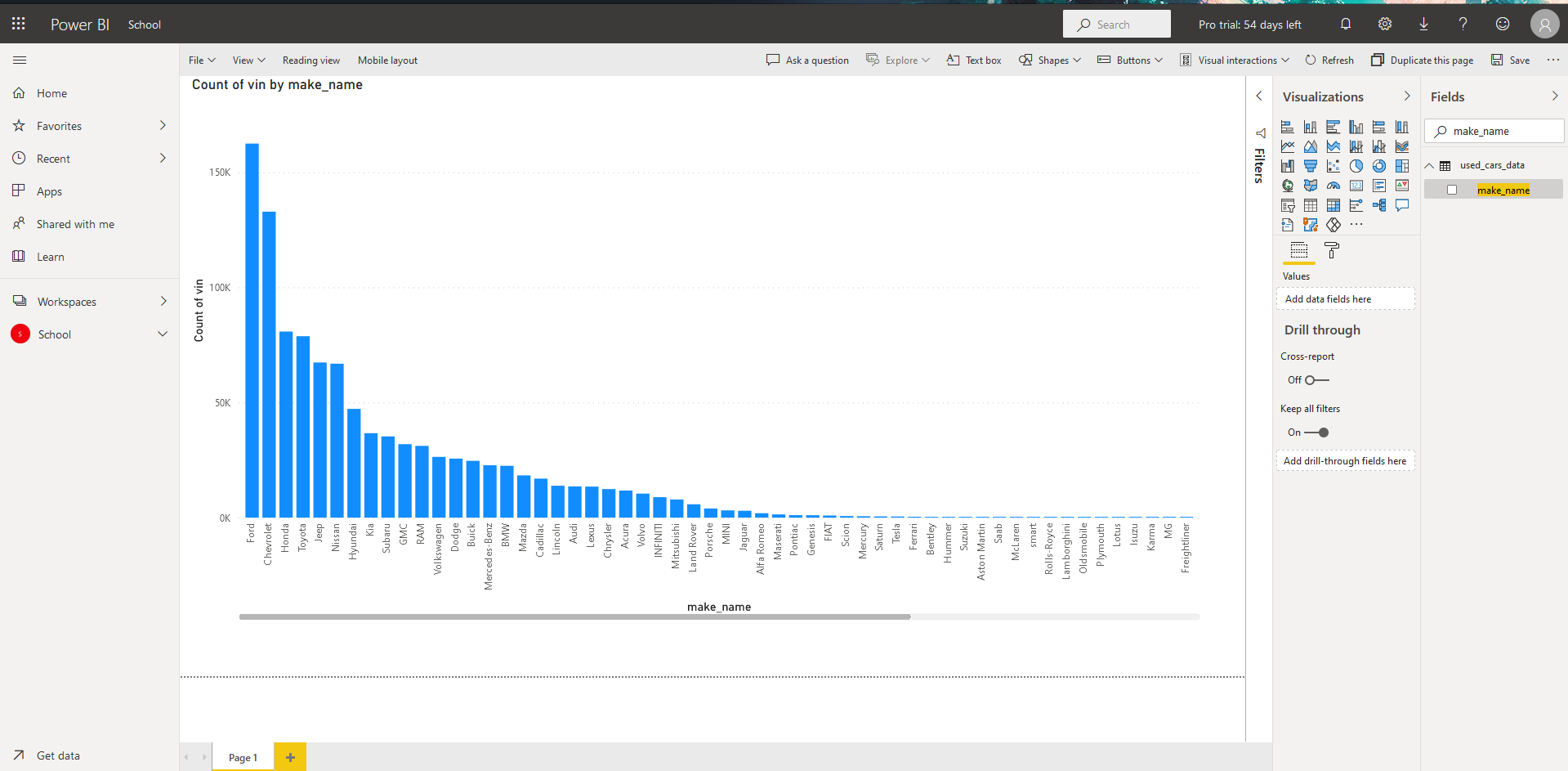
**6.1 PowerBI**

***Figure 2 - Used cars in the United State*s**

****

In Figure 2, the data presents the locations of used cars in the inventory of CarsGuru. From the visual, the team infers most used cars are found in populated cities. The eastern coast of the United States is filled with used cars, however CarsGuru is located in the east coast which may include more data. The midwest United States is fairly sparse with several locations of used cars.

***Figure 3 - Top used cars***



In figure 3 shows the top used cars in descending order using PowerBI. The *count of vin* is the count of the vehicle identification number in the Cars Guru inventory. The top brand for used cars is ford with a count of 150 thousand and more in the inventory. Followed by Ford is Chevrolet above 100 thousand count of *vin*. In third place is Honda above 50 thousand in count of *vin*.

**6.2 Tableau**

***Figure 4 - Mileage by Make***

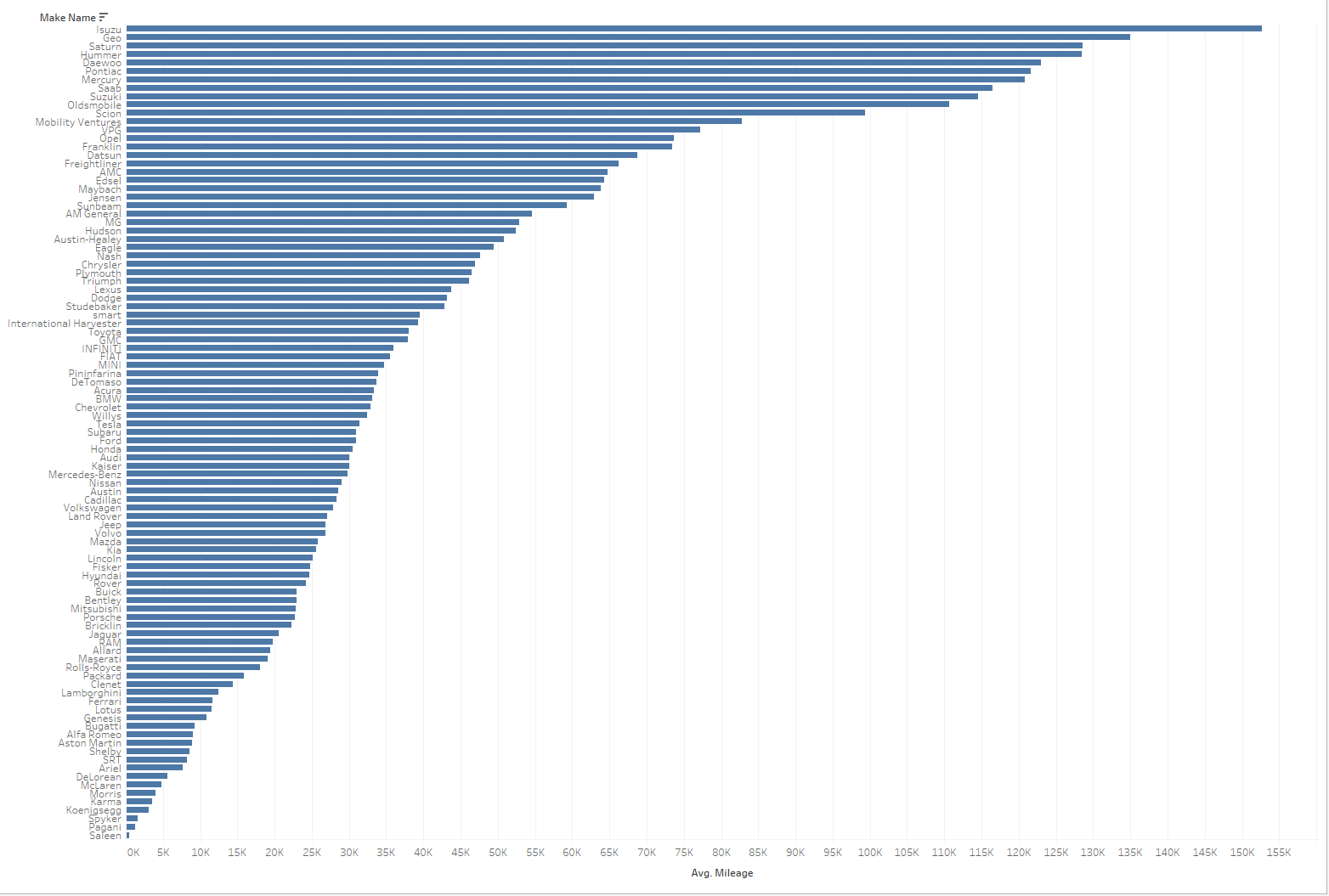


Figure 4 illustrates the average mileage of a particular make of car, sorted in descending order. The data shows that Isuzu cars generally have the most mileage. To note is that this result may be influenced by the fact that Isuzu primarily produces diesel work trucks.

***Figure 5 - Average Price by Make Filtered by Average Price Greater > $500,000***

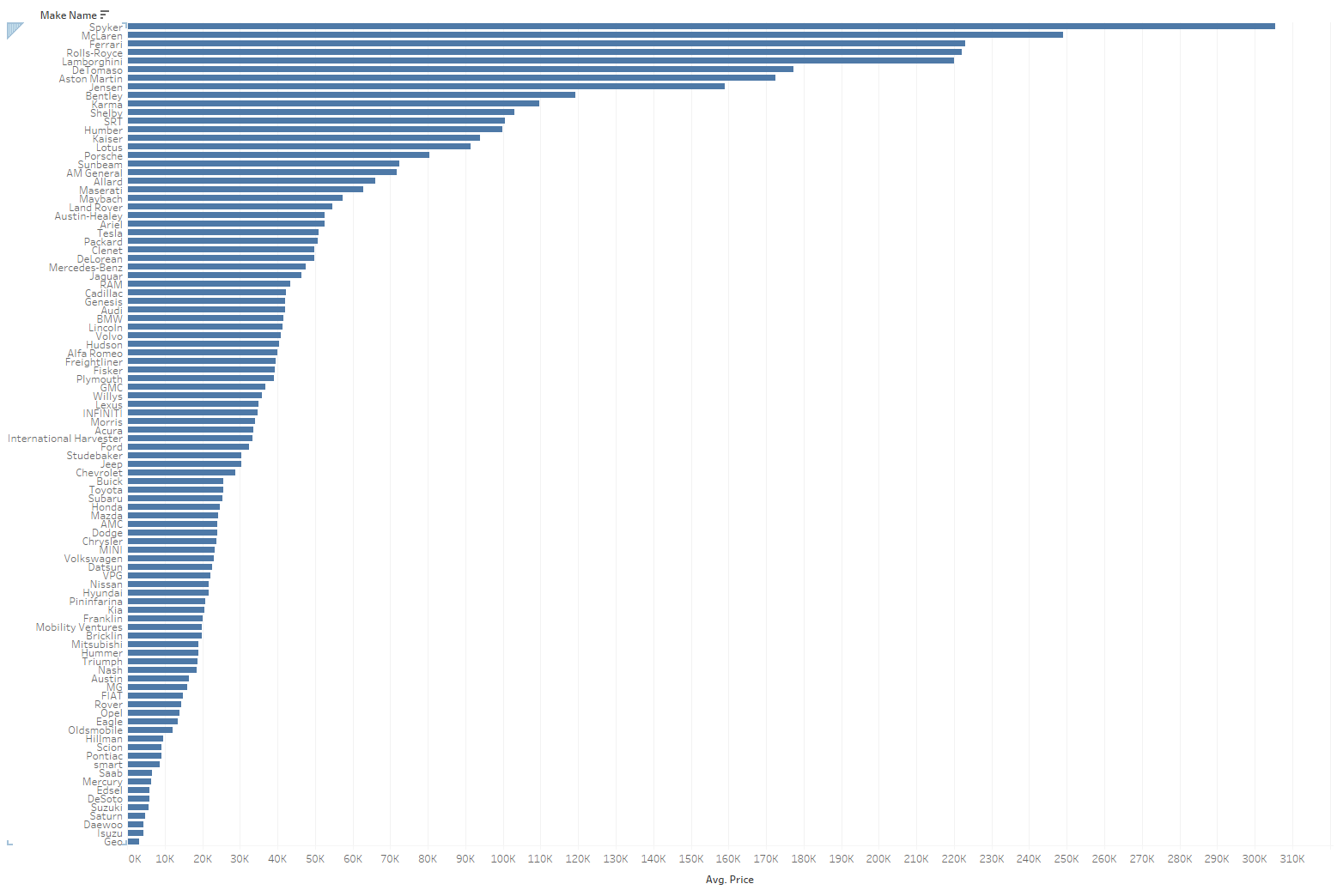


Figure 5 illustrates the average price of used cars by make, filtering out those greater than $500,000 to maintain a more readable chart. At the bottom are Japanese car makes, which means that these are generally the lowest priced used cars listed.

**7. Conclusion**

The team’s analysis by Tableau and PowerBI gave several insights with data provided by CarsGuru in September 2020. The dataset was analyzed loading it into hadoop, filtering and storing using Hive, and analyzing it by Tableau and PowerBI. The hardware that was used consisted of the Cluster Version 2.7.1. There were three nodes used and the memory size was 180 gigabytes. The analysis will determine the most popular used car. The analysis can separate the popularity between the eastern and western United States and see if either coast has a preference or more inventory of used cars. We can generate insights of the drive type, brand, mileage, or the fuel efficiency and determine what proves more useful in a purchasing decision of a used car.

The workflow was uploaded and stored in HDFS. The data was unzipped into a folder in HDFS. The data was loaded into Pig. In Pig the team has created columns and the data was filtered for the necessary data. Then the team stored the pig data into hadoop and downloaded it into a .csv file. The data was visualized using Microsoft PowerBI and was analyzed.

Our analysis showed that most used cars are found in the eastern coast of the United States. CarsGuru however is located on the east coast as shown in Figure 2. The top used cars are Ford, Chevrolet, and Honda in the bar chart on figure 3. In figure 4 it presents the average mileage of the make of car. The visualization is sorted in descending order showing that Isuzu cars generally have the most mileage. Isuzu works with diesel work trucks which may influence the make of the cars. The average price of used cars by the make of the car is filtered so any car greater than $500,000 are filtered as seen in figure 5. In the bottom of figure 5 it shows mainly Japanese cars which are the lowest priced as used cars listed.

2020 had the most used cars located in populated areas and found in the eastern coast of the United States. Japanese cars such as Suzuki, Honda, and Toyota with their sub brands are the lowest in the average price listing. The best mileage would be Isuzu cars. The top used cars in the inventory of CarsGuru is Ford, Cheverloet, and Honda.

### References

[1]Github: <https://github.com/aylmaokai/CIS4560>

[2]Direct Dataset Link: <https://www.kaggle.com/ananaymital/us-used-cars-dataset>

[3] AWS Link: <https://cis4560.s3.us-east-2.amazonaws.com/archive.zip>

[4] NIADA Report: <http://www.niada.com/uploads/dynamic_areas/5Nd7sQuZuYjQ5FCDUBXL/34/UCIR__RUNNING.pdf>

[5] CoxAuto Report: <https://www.coxautoinc.com/market-insights/used-vehicle-supply-and-prices-on-the-rise/>

1. Direct Dataset Link: <https://www.kaggle.com/ananaymital/us-used-cars-dataset> [↑](#footnote-ref-0)
2. NIADA Report: http://www.niada.com/uploads/dynamic\_areas/5Nd7sQuZuYjQ5FCDUBXL/34/UCIR\_\_RUNNING.pdf [↑](#footnote-ref-1)
3. CoxAuto Report: https://www.coxautoinc.com/market-insights/used-vehicle-supply-and-prices-on-the-rise/ [↑](#footnote-ref-2)