George Mason University

US Cars Dataset – Online Auction in North America

AIT 664-003 (Fall 2020)

Project Deliverable - 3

Group - 5

Sindhuja Pulluri

Doddanaik Basavaraj Vakkund

Vineel Vishwanth Busi

Introduction:

The Report "The US Vehicle Auction Market: Size, Trends and Forecast" includes the analysis of US vehicle Auction in terms of Value, Volume, and Segment. Our research project is related to Automobile industry where we are interested in analyzing the behavior of customers buying the used cars using Auction services. The competition within the various players is studied in the auction space. This report can assess these players to pick up the market capabilities and grow their profits by taking a decision from the data analysis report presented. (AuctionExport n.d.)

In the present situations, an automobile is the most important part in our daily routine. So, for buying a car or any vehicle we research on many factors like how much mileage it gives or how standard it is. Some people are fond of a particular brand for its features. So, through this data we are interested to find what all factors contribute in buying a car and price variation.

To explore the data, first it needs to be pre-processed and cleaned to remove any missing and unnecessary values. Then based on the information to be gained, visualization can be performed and for prediction machine learning algorithms can be applied.

Problem Statement:

Our goal in analyzing the data is to Predict the Price of Used cars based on different characteristics like Mileage, Year, color and brand. For this we want to first visualize the data to observe the trends and relation between variables.

Tools Used:

- Tableau: Tableau Software is an Interactive data visualization software. It queries relational databases, online analytical processing cubes, cloud databases and spreadsheets to generate graph-type data visualizations. (tableau n.d.)
- R Programming: R is a programming language and free software environment for statistical computing and graphics supported by R foundation for Statistical Computing. It's widely used among statisticians and data miners for developing statistical software and data analysis. (rstudio n.d.)

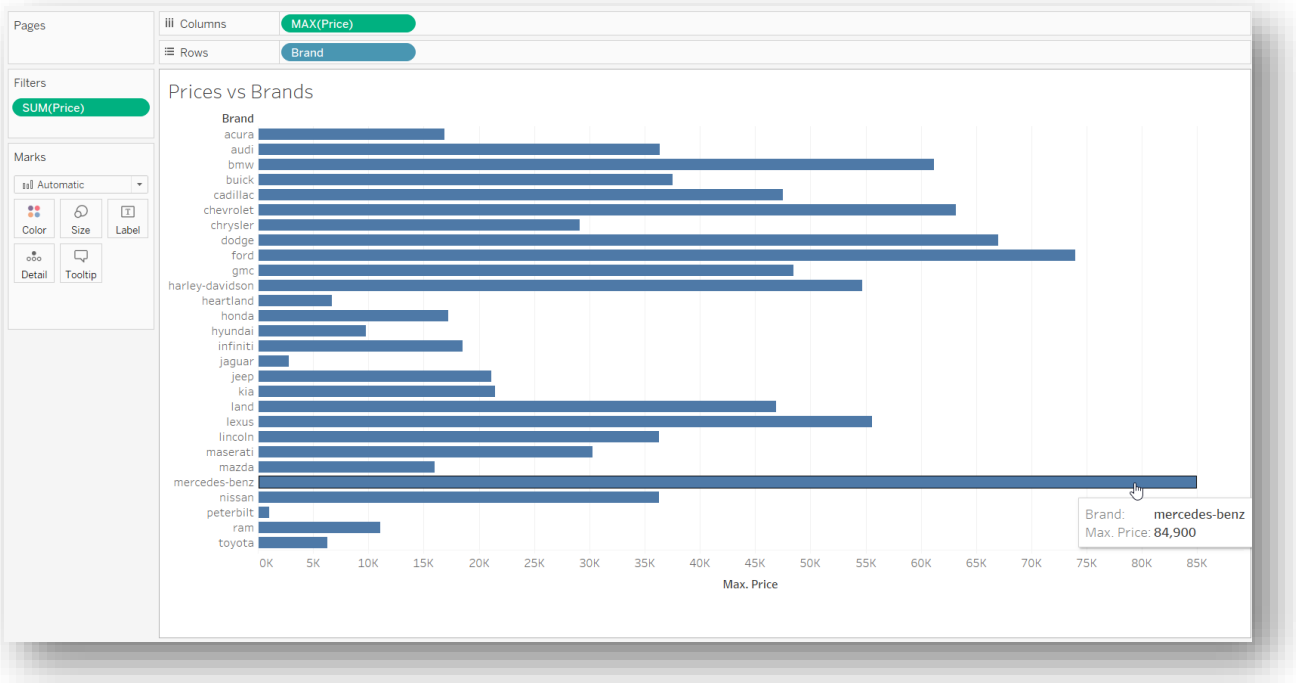
Exploratory Data Analysis:

Initially we have cleaned the dataset to set it into a proper format without any spelling errors and with proper case of words. Later performed exploratory data analysis for which we have used Tableau and R.

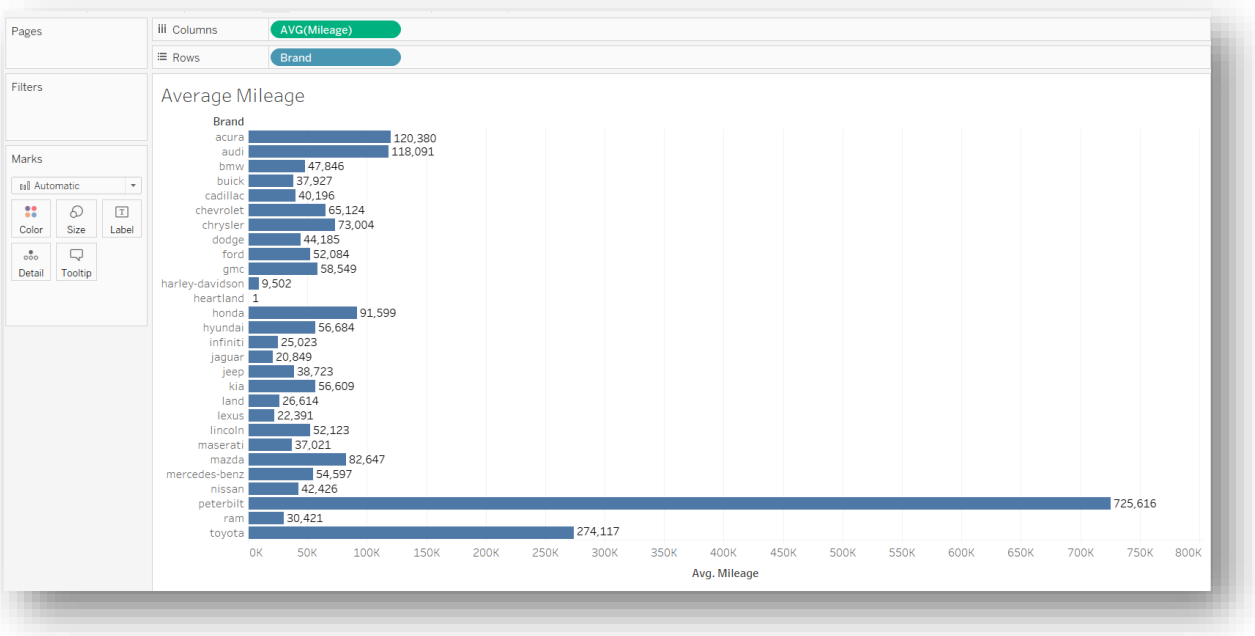
From the below screenshot it becomes easier to understand data rather than seeing it in tabular form. So, from the below summary statistics we can observe that the minimum Price is 0 and maximum is 84,900 and details like 1st quartile consists of data up to price 10,330 and so forth. Similarly, in "Brand" we can see how many cars of brand are present in the dataset for example Ford have 1,228 and dodge have 432 number of cars available for sale. The data is collected from 1973 till 2020.

```
> summary(cars)
     price              brand          model           year                title_status
 Min.   :    0   ford      :1228   door    : 645   Min.   :1973   clean vehicle    :2324
 1st Qu.:10330   dodge     : 432   f-150   : 218   1st Qu.:2016   salvage insurance: 157
 Median :16900   Nissan    : 309   doors   : 148   Median :2018
 Mean   :18859   Chevrolet: 296   caravan: 102   Mean   :2017
 3rd Qu.:25700   gmc       :  41   mpv     :  87   3rd Qu.:2019
 Max.   :84900   jeep      :  30   fusion  :  65   Max.   :2020
                 (Other)   : 145   (Other):1216
    mileage            color                    vin            lot
 Min.   :     71   white  :704     1g1al58f787159241:   2   Min.   :159348797
 1st Qu.:  21976   black  :511     1gndt13s632267445:   2   1st Qu.:167625635
 Median :  35595   gray   :393     1gnevhkw8jj148388:   2   Median :167745143
 Mean   :  52678   silver :298     3gcrkse37ag234620:   2   Mean   :167692903
 3rd Qu.:  64111   red    :191     19uua96529a004646:   1   3rd Qu.:167779865
 Max.   :1017936   blue   :150     19xfb2f81fe252000:   1   Max.   :167805500
                   (Other):234   (Other)            :2471
          state         country           condition
 Pennsylvania  : 293   Canada:   7   2 days left  :825
 Florida       : 245   USA   :2474   21 hours left:491
 Texas         : 214                 3 days left  :137
 California    : 187                 14 hours left:108
 Michigan      : 169                 1 days left  : 90
 north Carolina: 145                 8 days left  : 82
 (Other)       :1228                 (Other)      :748
```

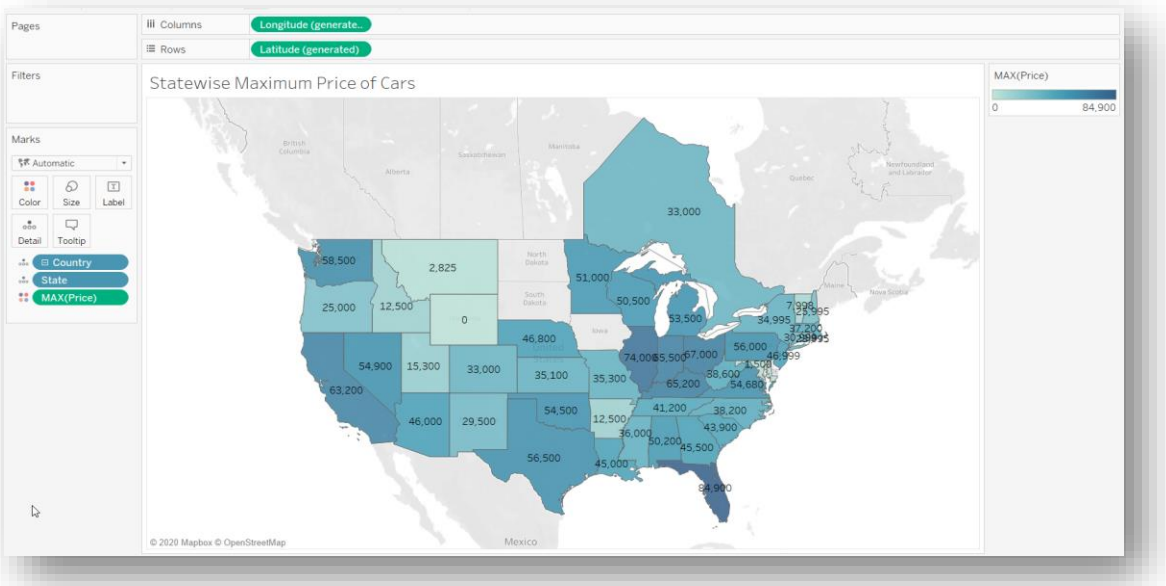
Every car has some manufacturing price. Here in our dataset, the price ranges with a minimum of $0 to a maximum of $84,900. The below Visualization shows the Maximum Price of the vehicles as per the brand. Through this visualization we can find which brand is costly.



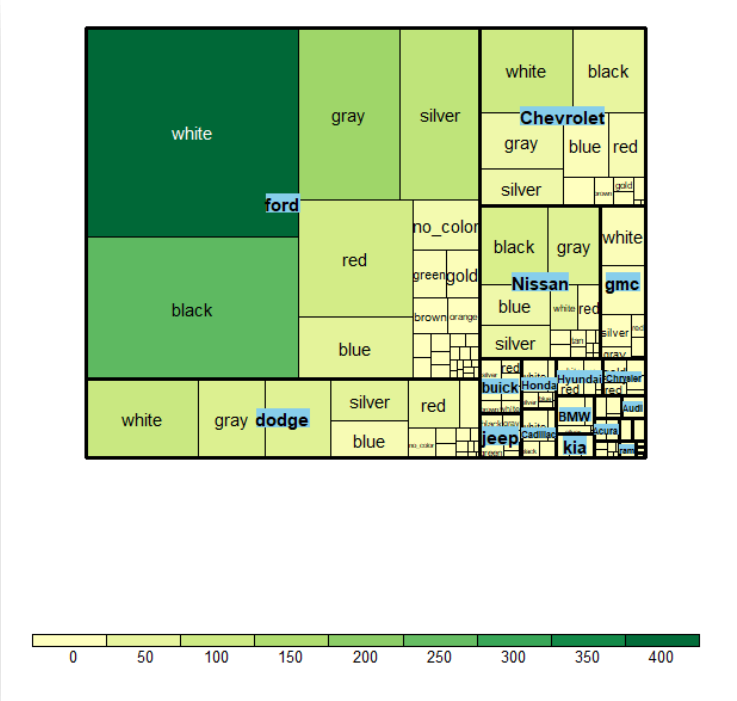
The below visualization is giving an information of average mileage for different brands of the cars. Here mileage is distance travelled by the car. Different averages of different brands are shown. The Mileage ranges from double digit to a value more than 6 digits which in turn says the cars travelled more.



Using Tableau, we can even visualize using the world map if the details about longitude and latitude of a place are known. The below Visualization is an example for that, which shows the maximum price of cars in each state. The intensity of color compares the value of Price. This is a good visualization which describes the sale of cars in all over the country and displays the maximum price in a particular state.

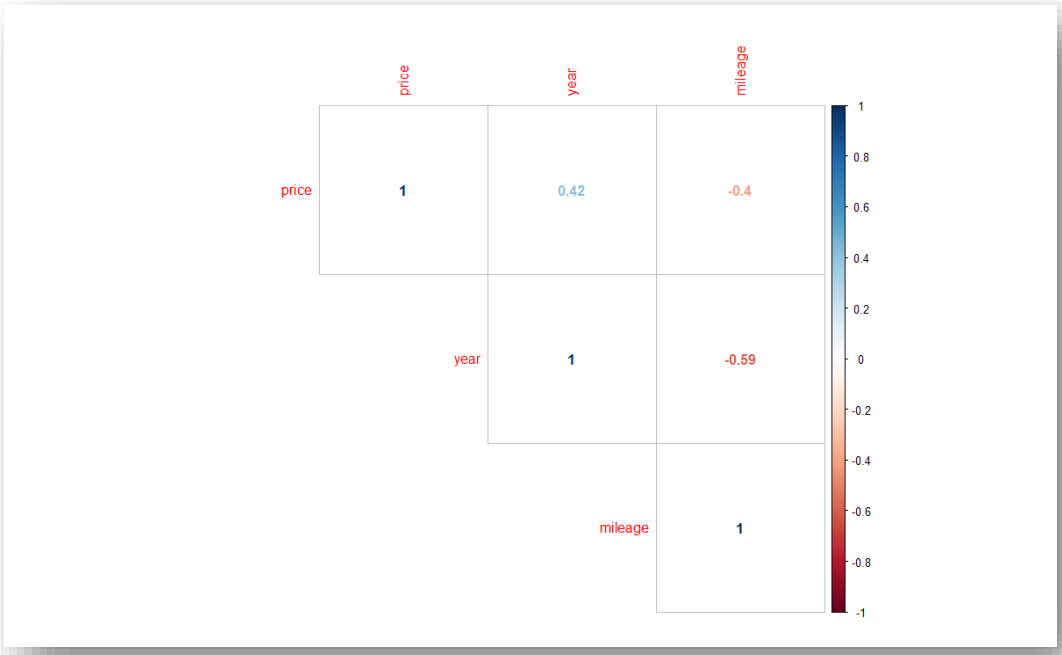


The below Visualization is obtained using R and is a treemap which describes the Colors present in each Brand and the size of each block and intensity of green color represents the count of that particular Cars. This is an interesting visualization which is easy to analyze the cars and their brands available.



Prior to running any sort of investigation, it is better to consider the correlation between the factors/features present in the dataset, and correlation plot is most popular for this reason as it shows the relation between all the variables present in the dataset. We have used the 'corrplot' package which is present in R Studio, to plot the Correlation between variables and to find whether the variables are linearly dependent or independent on each other.

From the below plot we can observe that the coefficient for Price and year is positive which means as new the vehicle is, the Price of the vehicle will be higher but the coefficient value is not nearer to 1 which means that is less correlated. In the same way the Price and Mileage are negatively correlated which means, as the number of miles the car has travelled increases, the price value decreases.

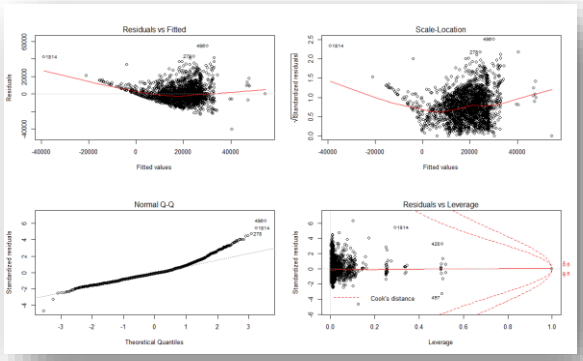


We have applied the model “Multiple Linear regression” on the data to check the Price which is dependent variable considering other factors as independent variables. Using the Backward Elimination Process, we removed the variables which are not significant and finally end up with only important variables to predict the Price value as shown below.

Most of the linear model assumptions are satisfied as seen from the below plots. And from the below plots we can say data is distributed properly and from Residual Vs Leverage plot we can say that there are no outliers in the dataset as nothing fell under cook’s distance.

```
> anova(fit)
Analysis of Variance Table

Response: price
              Df     Sum Sq    Mean Sq  F value    Pr(>F)
brand         26 4.2032e+10 1.6166e+09  19.1445 < 2.2e-16 ***
year           1 6.8863e+10 6.8863e+10 815.4944 < 2.2e-16 ***
mileage        1 1.8853e+10 1.8853e+10 223.2623 < 2.2e-16 ***
title_status   1 3.9555e+09 3.9555e+09  46.8422 9.726e-12 ***
state         42 2.5039e+10 5.9616e+08   7.0599 < 2.2e-16 ***
Residuals   2409 2.0342e+11 8.4443e+07
```



References

Alsenani, Doaa. n.d. *Kaggle.* https://www.kaggle.com/doaaalsenani/usa-cers-dataset.

n.d. *AuctionExport.* https://www.auctionexport.com/.

n.d. *rstudio.* https://rstudio.com/.

n.d. *tableau.* https://www.tableau.com/.