Predicting U.S. Car Preferences

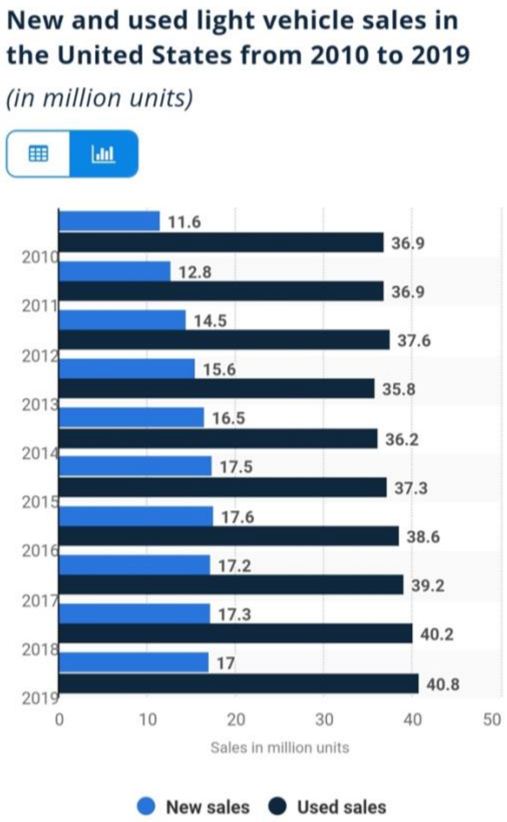
Martin S. Rivera Ritchie

August 2021

1. Introduction

1.1 Background

Used cars sales comprise an important chunk of sales every year. According to Statista in 2019 an estimated 40.8% of sales made in the US were used cars, in comparison to the 17% comprised by new car sales (graph 1).



(Img. 1, Graph from Statista depicting number of used car sales vs new car sales)

1.2 Business case

Having this background in mind, the business opportunity that I will investigate is how should a car dealership balance their inventory in order to improve their sales based on year, make, model, millage, price and state.

1.3 Interest

The main party of interest that this study is directed to are the owners/managers of car dealerships that wish to better understand which factors have an impact for a car to be sold, and what cars being sold the most in their state. There are two parties that could benefit from the study, one is the group of people who are interested in selling their used car. This group of people could benefit from the study to compare their car against the market and approximate a fair price for their vehicle. The third party is car manufacturers. This group could benefit from knowing which models are the most popular when making decisions for future product lines. However, in this study the focus will be geared towards car dealerships management teams that sell used and new cars, and wish to better balance their stocks between the two.

2. Data acquisition & Cleaning

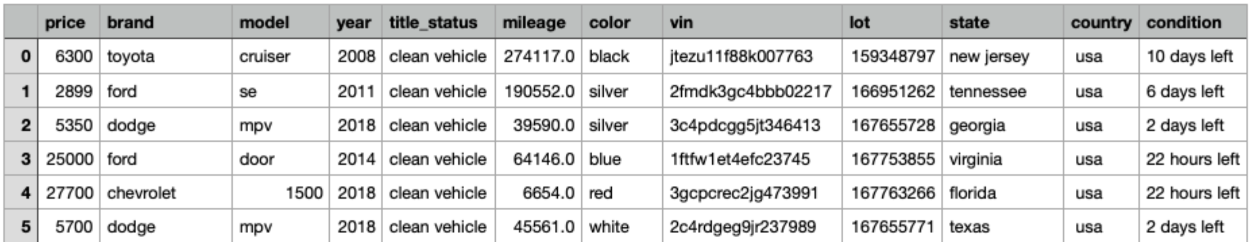
2.1 Data Sources

The data that I will evaluate is one that I acquired from *Kaggle* called *"US Cas Dataset"* which in turn was scrapped from *auctionexport.com*. This dataset contains 2,498 entries of which their classification is as follows:

Variable Name	Type	Description
Price	Float	Price at which vehicle was sold.
Brand	String	Make of the vehicle sold.
Model	String	Model of the vehicle sold.
Year	Float	Year of model pertaining to the vehicle.
Title_status	String	Whether, or not, the vehicle is identified as salvaged by the insurance; binary.
Millage	Float	Millage as read in the odometer of the vehicle.
Color	String	Paint of the vehicle.
VIN	String	Vehicle Identification Number.
Lot	Integer	Lot pertaining to the manufactured batch that the vehicle belongs to.
State	String	State in which transaction took place.
Country	String	Country in which the transaction took place.
Condition	String	Time that it took to sell the vehicle at auction.

(Table 1, Variable description)

The dataset as it was downloaded is in a .csv file, let’s look at the first five rows to get an idea of what it looks like in its raw form:



	price	brand	model	year	title_status	mileage	color	vin	lot	state	country	condition
0	6300	toyota	cruiser	2008	clean vehicle	274117.0	black	jtezu11f88k007763	159348797	new jersey	usa	10 days left
1	2899	ford	se	2011	clean vehicle	190552.0	silver	2fmdk3gc4bbb02217	166951262	tennessee	usa	6 days left
2	5350	dodge	mpv	2018	clean vehicle	39590.0	silver	3c4pdcgg5jt346413	167655728	georgia	usa	2 days left
3	25000	ford	door	2014	clean vehicle	64146.0	blue	1ftfw1et4efc23745	167753855	virginia	usa	22 hours left
4	27700	chevrolet	1500	2018	clean vehicle	6654.0	red	3gcpcrec2jg473991	167763266	florida	usa	22 hours left
5	5700	dodge	mpv	2018	clean vehicle	45561.0	white	2c4rdgeg9jr237989	167655771	texas	usa	2 days left

(Image 2, Demonstrating a portion of the dataset to be used from Kaggle)

2.2 Variable Selection

Looking at the dataset in its raw from there are a few steps that must be taken to clean the data before it can begin to be studied. These steps include eliminating some of the variables, standardizing the data in an existing column, and must importantly the addition of a column which will need to be randomized for the purposes of this exercise which is the characteristic “New vs. Used car”. The changes to the dataset are described in the following table:

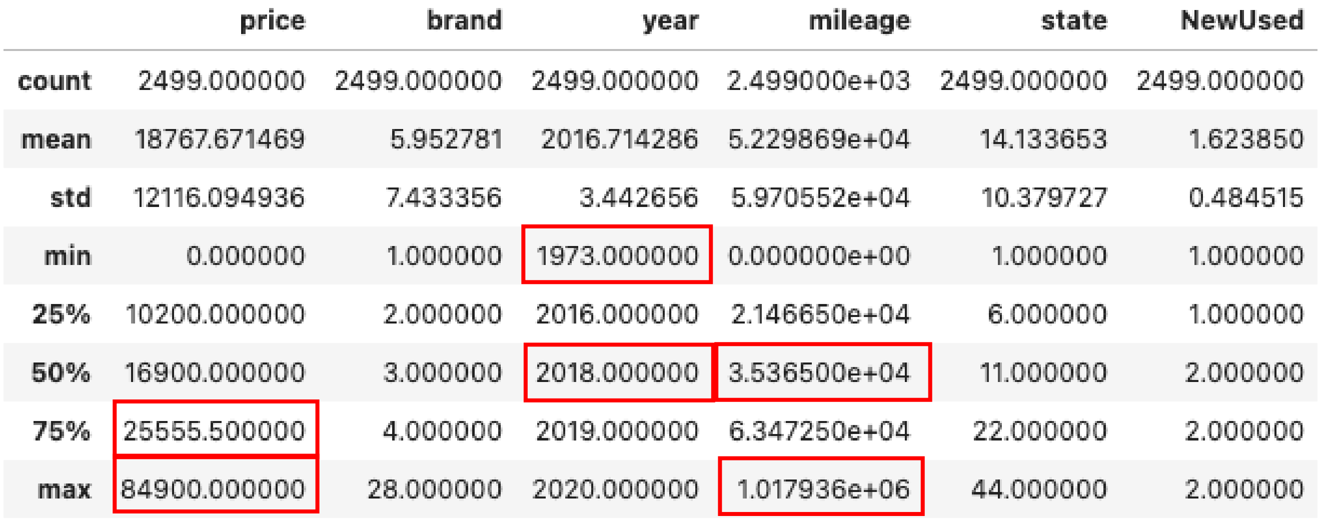
Variable	Action	Description of Change
Model	Standardize	The variable contains both numbers interpreted as integers and words interpreted as string values. All values in the variable must be transformed into string values. The variable also contains Null values, these values will be substituted with the median of model of the cars sold per brand.
Title_status	Eliminate	Variable is removed from the dataset.
VIN	Eliminate	Variable is removed from the dataset.
Lot	Eliminate	Variable is removed from the dataset.
Condition	Eliminate	Variable is removed from the dataset.
Country	Eliminate	Variable is removed from the dataset.
New/Used	Append	For the purpose of this exercise, where the range in the *Year* of the cars is from 1973 – 2020, all cars from 2019 – 2020 will be classified as *new*, which comprises 37% of the cars in the dataset.
Target	Append	Create Target variable based on a profile identified during the Exploratory Data Analysis.

(Table 2, Variables changes)

In order to produce a model that can identify the car profile with the better chance to get sold a target variable needs to be identified or created. The first step is to gain a full understanding of all the variables distributions, behaviors, and correlations. In this section these insights will be obtained with the use of statistical analysis, univariate exploration with histograms, and a bivariate approach to histograms.

2.3 Data Cleaning

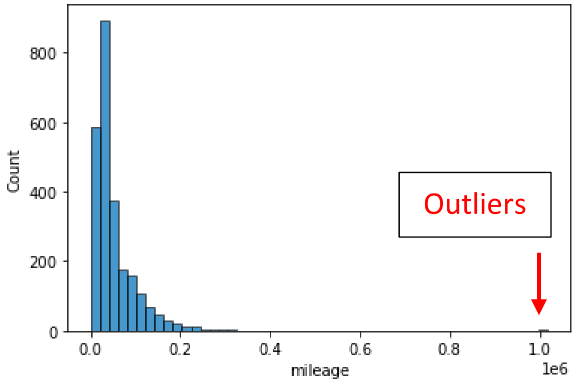
First, from the statistical summary I identify possible outliers.



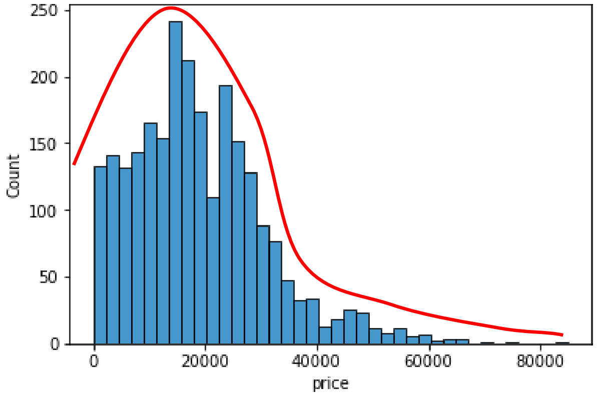
	price	brand	year	mileage	state	NewUsed
count	2499.000000	2499.000000	2499.000000	2.499000e+03	2499.000000	2499.000000
mean	18767.671469	5.952781	2016.714286	5.229869e+04	14.133653	1.623850
std	12116.094936	7.433356	3.442656	5.970552e+04	10.379727	0.484515
min	0.000000	1.000000	1973.000000	0.000000e+00	1.000000	1.000000
25%	10200.000000	2.000000	2016.000000	2.146650e+04	6.000000	1.000000
50%	16900.000000	3.000000	2018.000000	3.536500e+04	11.000000	2.000000
75%	25555.500000	4.000000	2019.000000	6.347250e+04	22.000000	2.000000
max	84900.000000	28.000000	2020.000000	1.017936e+06	44.000000	2.000000

(Img. 3, Statistical Summary)

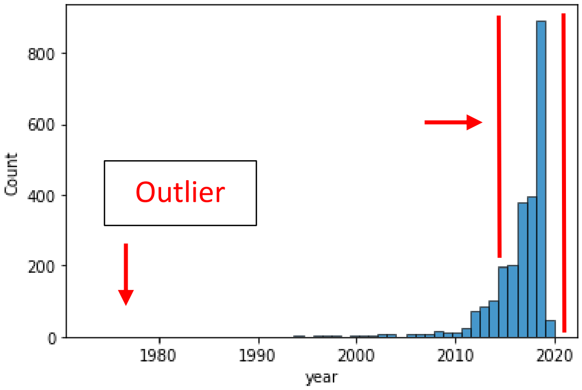
From the statistical analysis there seem to be distributions in the variables *Price, Year, Mileage* that stand out. The variable *Year* appears to have an outlier in tis minimum value of 1973 when compared to the median which is 2018. Within the variable *Price* the difference between the max value of $84,900 USD and the median of $16,900 USD indicates that the max value could be an outlier; similar with the variable mileage and its max value of 1,017,936 miles. In order to confirm that these are outliers which need to be removed, the histogram of each variable is plotted and analyzed.



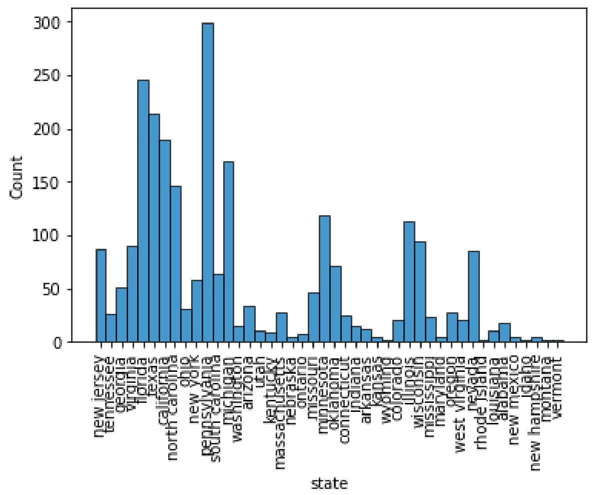
Mileage: The histogram shows that the majority of the vehicles sold are of low mileage with a clear outlier with 1,017,900 miles. Most of the data lies within the lower quartile.



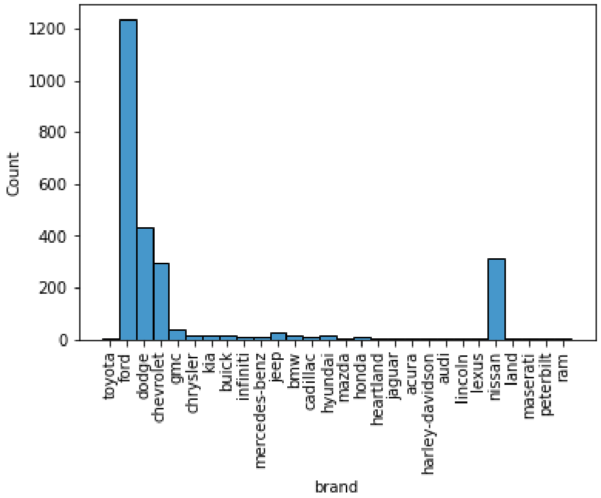
Price: The histogram exhibits a rough bell shape where there's a high concentration of sales being made around the middle quartile of the data which is at $16,900.



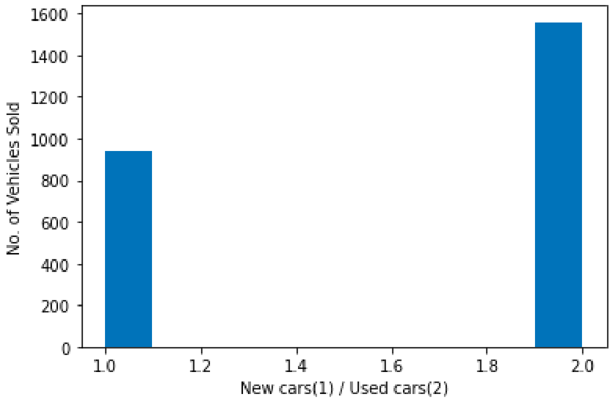
Year: The histogram indicates that the majority of the cars being sold are within the range of 2016-2020.



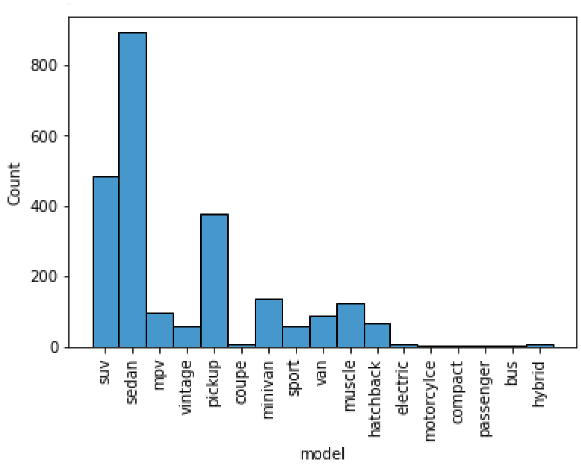
State: From the histogram above, the states with the most sales are South Carolina, Florida, Texas, California, North Carolina.



Brand Histogram: From the histogram above the four most popular brands are Ford, Dodge, Chevrolet, and Nissan.



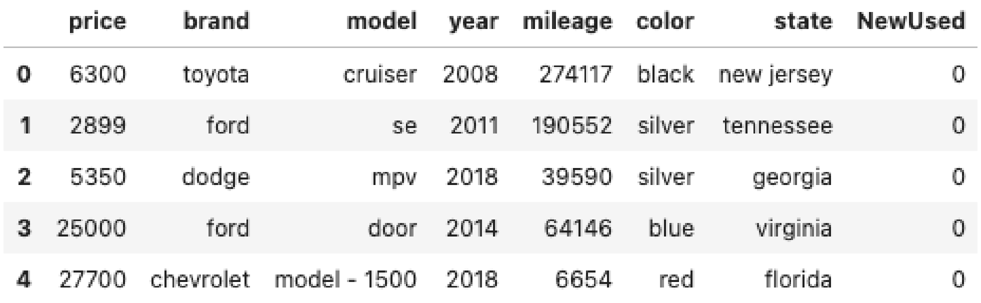
New vs Used Cars Sold: The histogram confirms that more used cars are sold, on average 1.6 more used cars vs. new.



Model Histogram: The histogram shows that the most popular models are sedans, SUV's, and Pickups.

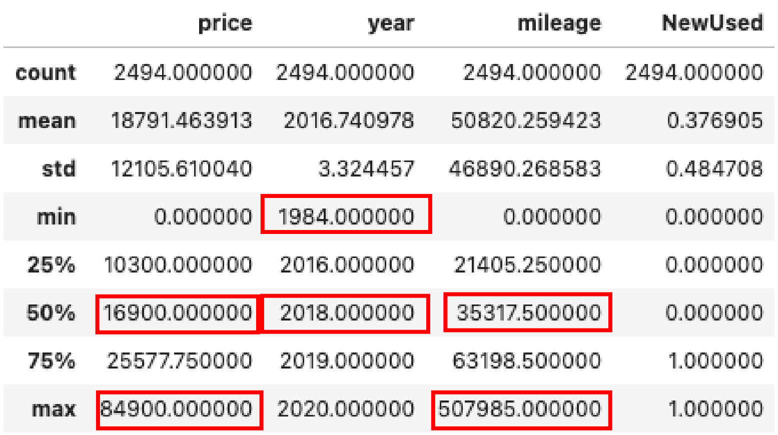
From the univariate analysis the max values in the column *Mileage* are identified as outliers and are erased, which are rows 490, 516, 528, 1827 with values over 902,041 miles. The same is done with the outlier identified in the variable *Year* which has a value of 1973. These changes are done in order to reduce the error due to variance when the variable target is created.

The resulting dataset looks as follows:



	price	brand	model	year	mileage	color	state	NewUsed
0	6300	toyota	cruiser	2008	274117	black	new jersey	0
1	2899	ford	se	2011	190552	silver	tennessee	0
2	5350	dodge	mpv	2018	39590	silver	georgia	0
3	25000	ford	door	2014	64146	blue	virginia	0
4	27700	chevrolet	model - 1500	2018	6654	red	florida	0

(Img. 4, First five rows of clean Dataset)



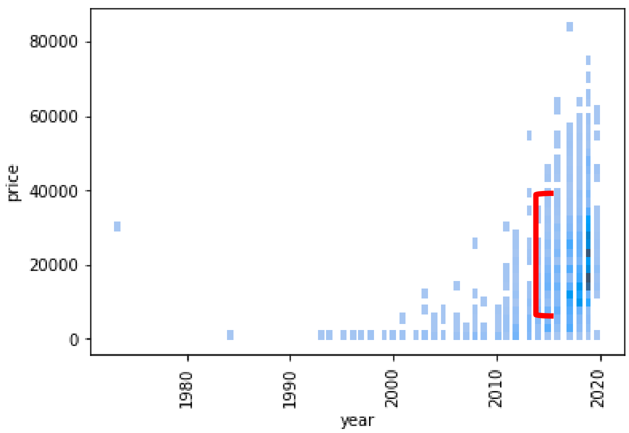
	price	year	mileage	NewUsed
count	2494.000000	2494.000000	2494.000000	2494.000000
mean	18791.463913	2016.740978	50820.259423	0.376905
std	12105.610040	3.324457	46890.268583	0.484708
min	0.000000	1984.000000	0.000000	0.000000
25%	10300.000000	2016.000000	21405.250000	0.000000
50%	16900.000000	2018.000000	35317.500000	0.000000
75%	25577.750000	2019.000000	63198.500000	1.000000
max	84900.000000	2020.000000	507985.000000	1.000000

(Img. 5, Table shows the change in statistics with the outliers removed.)

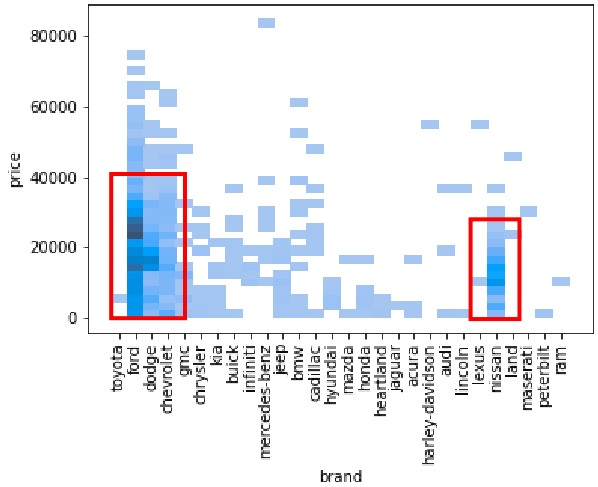
3. Analysis for Target variable profile

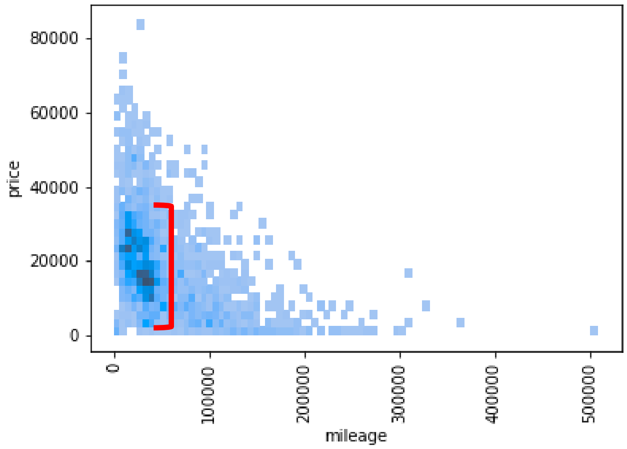
3.1 Bivariate Analysis

In the next step for the creation of the Target variable the variable *Price* is selected because it was the variable that required the least amount of cleaning, has the closes shape to bell curve in its distribution, and because in the U.S. the price at which a vehicle is sold can be adjusted by the sale people from purchase to purchase. For these reasons the variable *Price* is chosen as the most flexible and with the best range to compare it to the other variables; this will help to identify ranges for the construction of the Target variable.

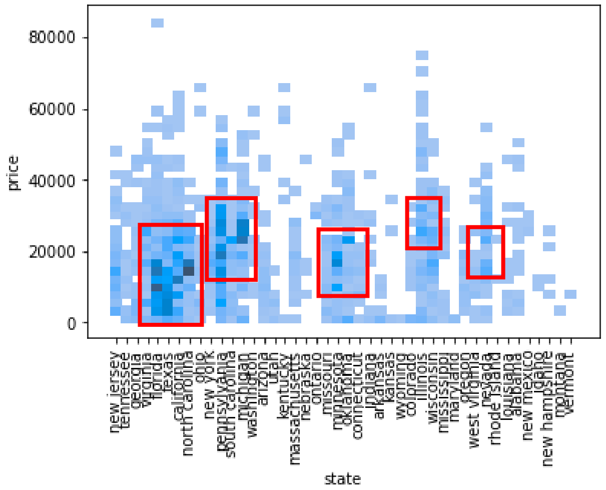


Price v. Year graph: shows a soft concentration of cars sales between the years 2018 and 2020, and a price range between $10,000 - $40,000 USD.



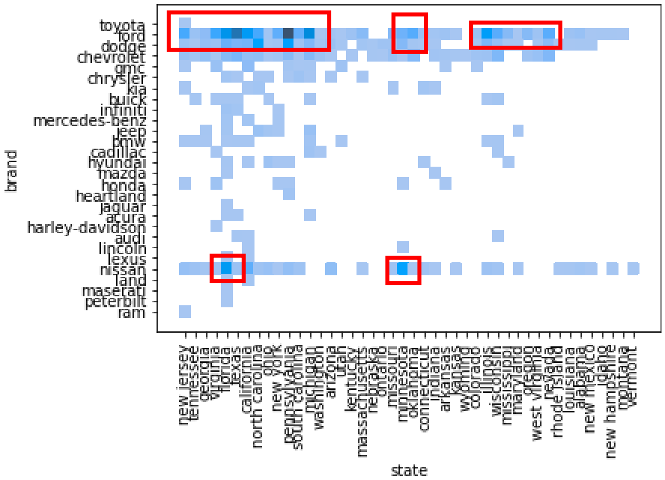


Price v. Mileage graph: shows a concentration of car sales with a mileage lower than 10,000 miles, and a Price range between $10,000 - $40,000 USD.

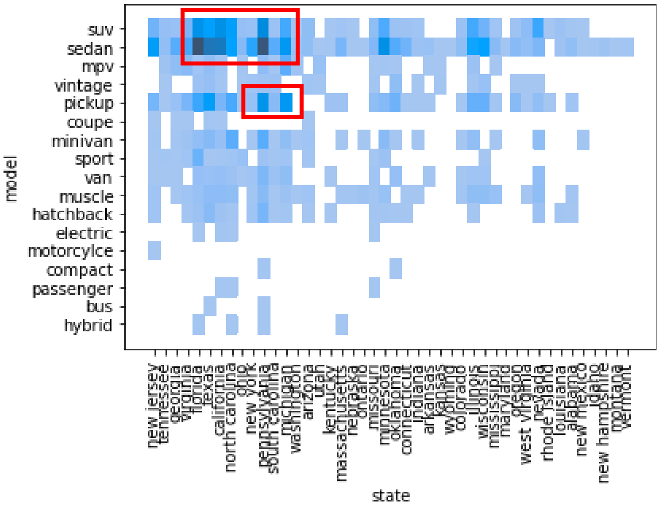


Price v. Brand graph: The graph shows the same price range for the 4 most popular brands which are Ford, Dodge, Chevrolet, and Nissan.

Price v. State graph: This graph better identifies additional states with a higher amount of car sales, which lie between the price range of $40,000-$10,000 USD.



Brand v. state graph: This additional histogram is made to better visualizes additional states that also contribute to the total sales of certain brands but were not easily identifiable in the previous graphs. These additional states are: New Jersey, Pennsylvania, Michigan, Minnesota, and Oklahoma.



Model v. state: The graph shows two concentrations of data, in the first one SUV's and Sedans are the most popular models in Florida, Texas, and California. The second concentration shows pickups being the most popular models in Pennsylvania and Michigan.

3.2 Target Variable Construction

The patterns identified in the previous areas will now be used to establish the parameters for the *Target* variable. The variable will identify clients that fall within the following profile:

I. Price range: $16,900 USD - $40,000
II. Brands: Ford, Dodge, Chevrolet, and Nissan.
III. Years: 2016 - 2019
IV. Mileage: 70 – 10,000 miles

The profiles resulting from this identification will be used to train the predicting models in order to identify what are the optimal conditions that better guarantee that a client will purchase a car. *Model* is being excluded from the *Target* variable because *Model* preference changes greatly between states; as seen from the *Model v. State* graph. Also, the variable *NewUsed* will not be factored into the Target profile so as to not introduce a bias regarding the condition of the car, I will let the models in the next section prove, or disprove, whether Americans are purchasing used cars over new ones.

4. Models

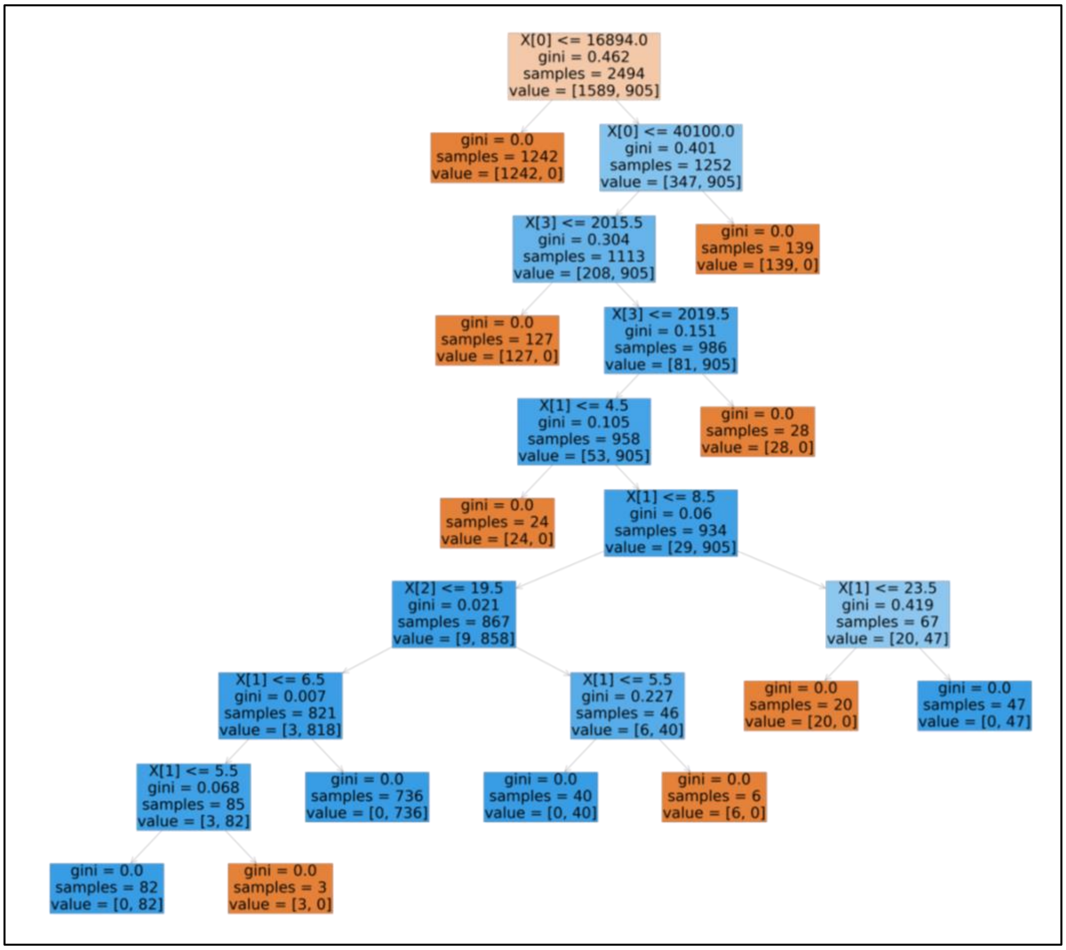
For the predicting models that follow the dataset will be split into training and testing sets, the distributions of these will be 75% of the data for training and 25% for testing. There are two ways to determine which are the most important characteristics to sell a car, one is through a categorization analysis, and the second one is by finding an accurate range with regressions. Making a profile is a lot easier using a classification approach, which is why I will be using a Decision Tree. Decision Trees are very visual tools that will help better map the way consumers pick their cars which can be very helpful when talking with dealership managers. The regression model will similarly help solidify the profile but in a much more numeric way, which can be better when talking to the finance department.

4.1 Decision Tree

First, the I will be using the Decision Tree to determine the most important factors for people to buy cars nationwide, across all 50 states. When I began programming the model, the first tree that I programmed ended up having an issue processing the variable labels with the string format, in order to correct that I used *Label Encoder* which transforms the string labels into integers but keeps them as labels without assigning a hierarchical structure to the data. The next problem I had was that the tree that I rendered was too large and could not be easily read; to solve this I had to prune the tree. For the pruning I limited the number of branches to a maximum of 2 per split and the maximum depth of 5 levels. This produced a much more legible tree, without compromising the classification of the data. Because of the Label Encoder function used the labels appear as factors of X, for this reason to read the Decision Tree the following key must be used.

Decision Tree Legend	
Representation in Tree	**Variable in Dataframe**
X[0]	Price
X[1]	Brand
X[2]	Model
X[3]	Year
X[4]	Mileage
X[5]	Color
X[6]	State
X[7]	NewUsed

(Table 3, Variable description)



(Img. 5, Variable description)

The tree above shows the decision patterns of the general population at a federal level. By the way the tree classifies the data, we can see that the most determining factors, nationwide, for clients to purchase a car in which are as follows in order of relevance:

1. Price: $16,894 USD
2. Year: 2015
3. Brand: Chevrolet
4. Model: SUV

From this data we can tell that the consumers in the data base are not buying expensive vehicles- they prefer to buy on a budget, they prefer used cars, the most popular brand is Chevrolet, and the most popular model across all brands are SUV.

4.2 Regression Models

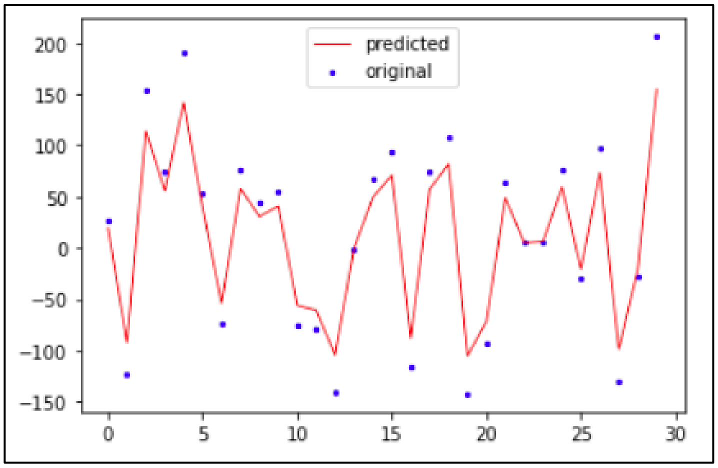
In order to find the model that most accurately identifies which clients are purchasing cars in accordance to the target profile identified in the Bivariate Analysis section, I first run 4 of the most popular models (in their default settings). I use the R-squared values which measure the amount of variance explained by the model to do this comparison, the following table shows the scores of each model.

Regression Evaluation Table	
Regression Model	**R-Squared Score**
Simple Linear Model	0.26
Logistic Model	0.78
Polynomial Model	0.73
Elastic Net Model	0.89

(Table 4, Variable description)

From the table it is clear that the model that best explains the data is the Elastic Net Regression Model. Since I used the tree to interpret the data at the national level, I will use the Elastic Net Regression to interpret the data by state.

After training, testing, and tunning the model I end up with a Predictive Model with a R-squared value of 0.93. The following graph shows that the model, though it has a high R-squared value, it is not over fitted.

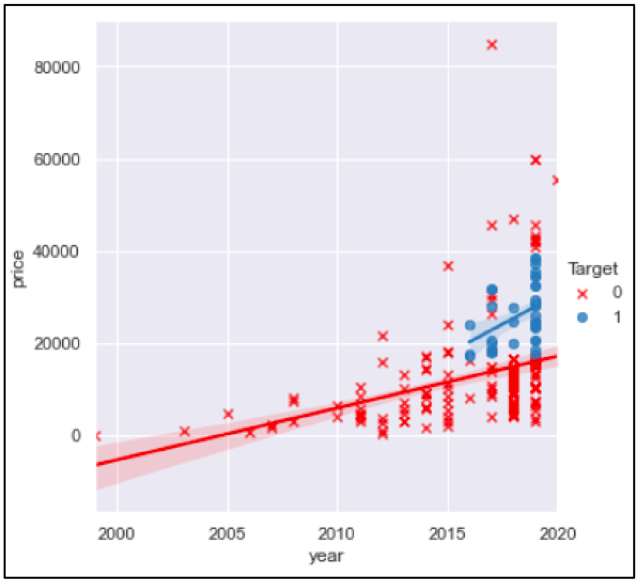


(Img. 6, Variable description)

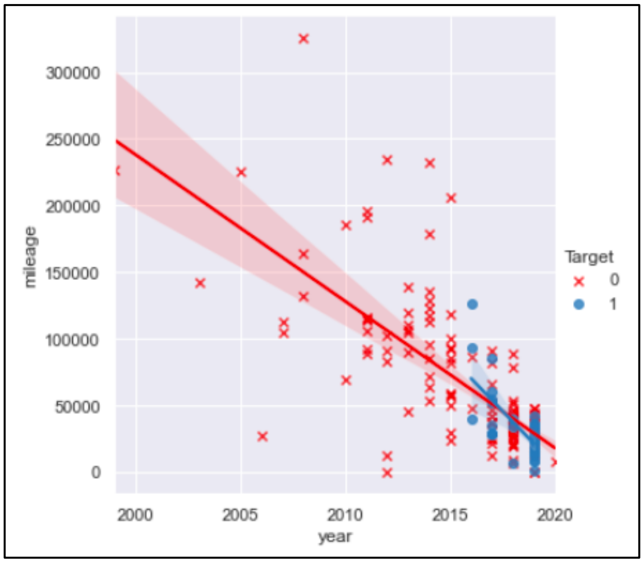
I now proceed to use this predictive model to profile 2 of the states with the most sales of vehicles in the country: Florida and California. To make the profile I use the predictive model in a similar way to the one used in the bivariate analysis section, by graphing two variables against each other I can use the Elastic Regression Predictive Model to find out what are the preferences in vehicles in each state.

4.3 Results of the top 2 states

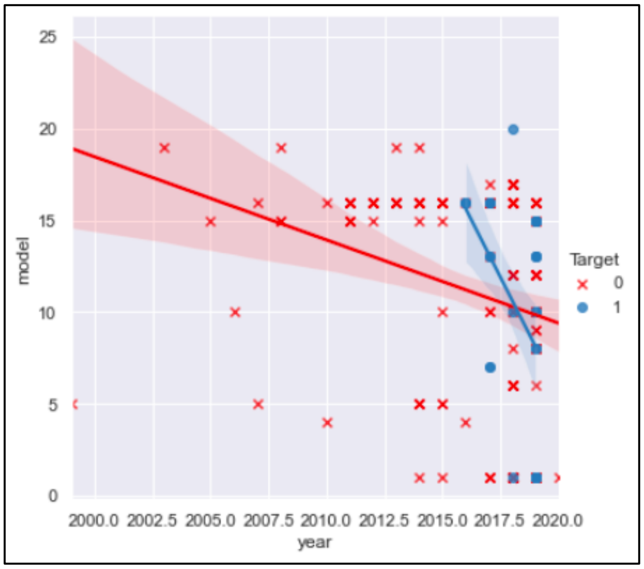
Florida



<u>Price vs. Year:</u> From this graph we can observe that the elastic regression predicts that the best price range to sell a car in Florida is between $20,000 - $30,000 USD.

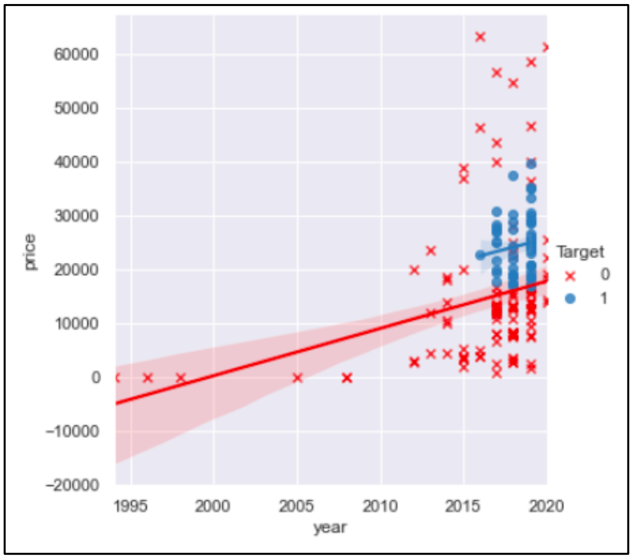


<u>Mileage vs. Ye</u>ar: From this graph we can observe that clients in Florida prefer new cars, or cars with low mileages; between 0 – 50,000 miles.

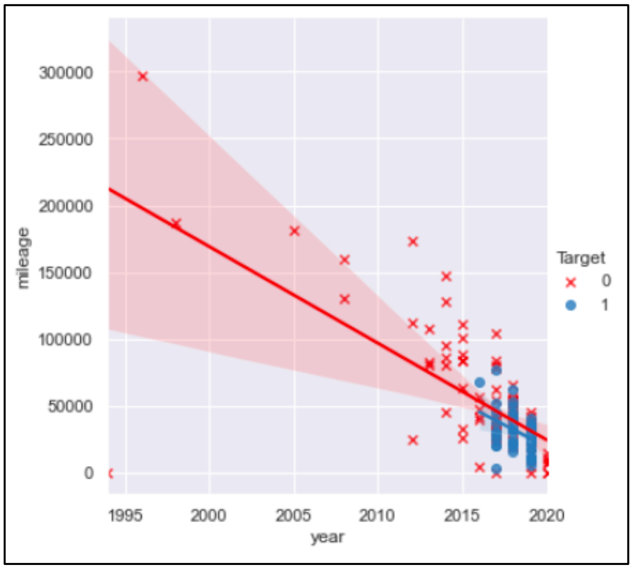


Model vs. Year: From this graph we observe that the 4 most popular models and the year they were made are 2019 four door pickups, 2018 Minivans, and 2017 4 door sedans, and 2016 single cab pickups.

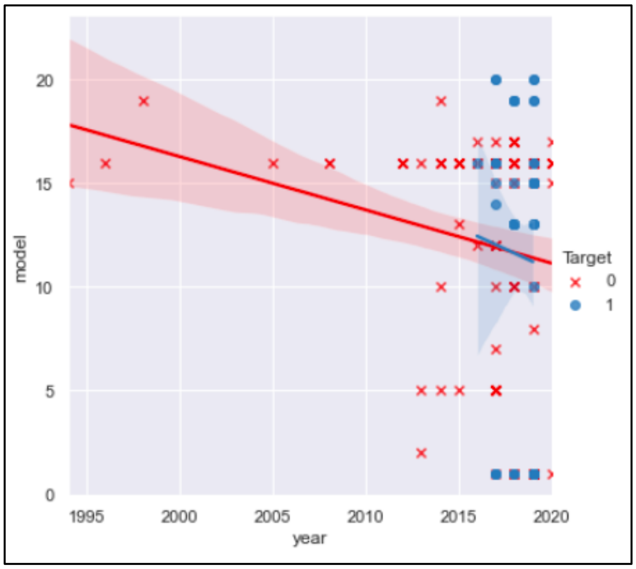
California



Price vs. Year: From the graph we see that the most popular price range for the purchasing of vehicles in California is between $21,000 - $25,000 USD.



Mileage vs. Year: From the graph we can see that the most popular cars are used cars with mileages between 30,000 – 50,000 miles.



Model vs. Year: From the graph we observe that the most popular years and models in California are 2019 Minivans, 2018 four door Sedans, 2017 Electric cars, 2016 Motorcycles.

5. Conclusions

Ultimately, the study successfully proved that the rising trend in used car sales is accurate. Using the decision tree I was able to reliably identify the trends nationwide amongst the five most relevant variables Price, Year, Model, and Brand. This information is mostly useful for car manufacturers that are making decisions regarding future car lines and are considering dropping certain models in order to focus on the cars that are most popular nationwide, which for the time being are SUVs.
With the regression model I was able to go a step further and find the preference per state with a high degree of accuracy. In Florida we see a preference towards minivans and pickups, while in California there's a propensity for electric cars and motorcycles.

6. Future directions

For future directions it would be useful to have a wider range of consumer data. As I established in the data acquisition portion, the data used belongs to a single auction website, and as such it is a small representation of the overall market. Having access to a bigger and more diverse data base could improve the profiling of the consumers nationwide and per state. With a broader data base new trends might be identified, and with a higher number of transactions, the information also becomes more indicative of purchasing patterns; after all the data base used only represents 2,498 clients of the over 200 million adults in the US.