Predicting Used Car Price with Cars.com Data

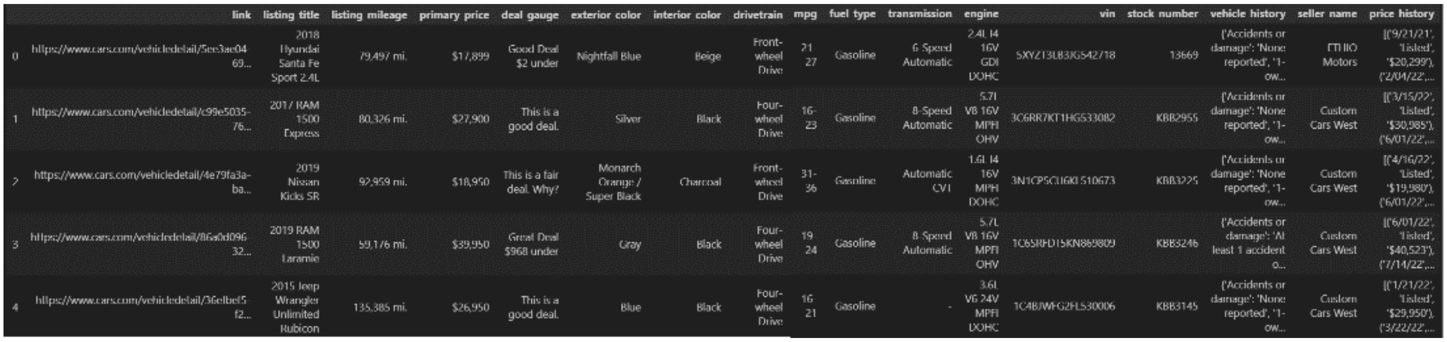
Evan Hollier, Paul Homuth, Han Lee

Introduction

The car market within the United States has recently experienced a large shakeup due to the COVID-19 pandemic. This shakeup is not only affecting prices, with reports of an 11% percent increase in the average cost between 2020 and 2021, but also with the means in which the cars are being sold as many large car companies are preparing for a 40% increase in online sales.[1] Other outlets are reporting a near 25% increase in price of a vehicle between the pre- and post-COVID eras.[2] There are many reasons for the boom in prices, but a main reason is the shutdown during COVID caused a gap in the supply chain of vehicles which is still being seen today.[2] With the uncertainty in the used car market, an exploratory data analysis and data modeling study was conducted in order to build a tool to help navigate the used car market to help the user better understand what a good price for a used vehicle is today.

Data Collection and Feature Engineering

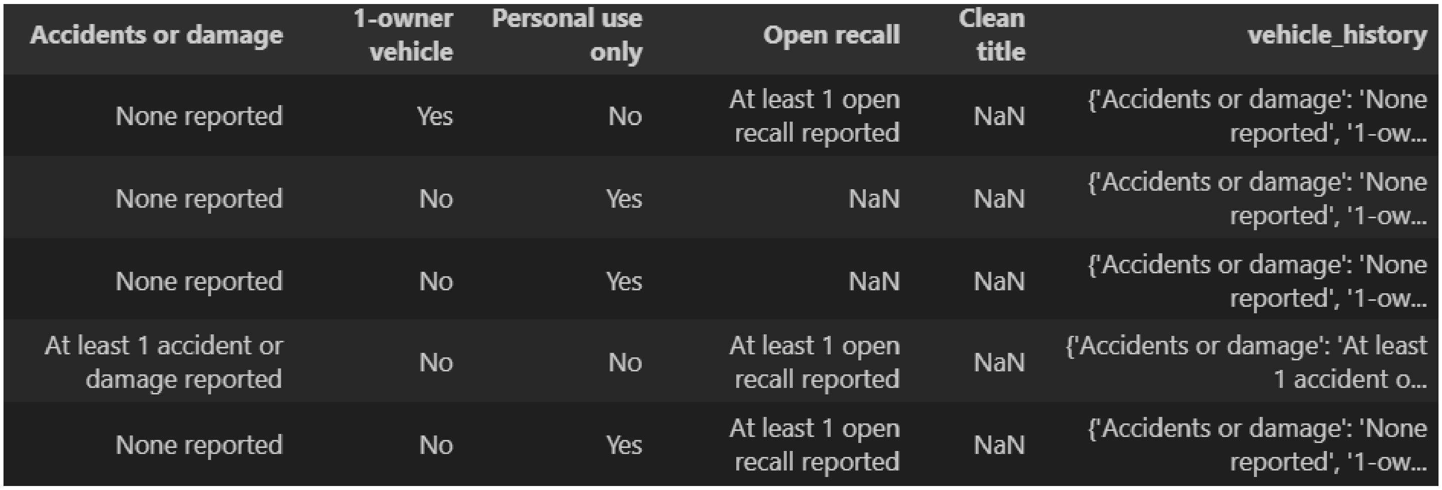
The data was collected through web scraping of cars.com with the use of Selenium. The program was set up to search for used vehicles within a ten-mile radius of the University of Denver. The program was designed to pull all possible listing information from the vehicles that ran on gasoline to ensure that each feature pulled would be a standardized format. After the program ran for about 15 hours, a data set of 6,002 vehicles had been collected with 17 features for each record. A sample of this dataset is shown below.



	link	listing title	listing mileage	primary price	deal gauge	exterior color	interior color	drivetrain	mpg	fuel type	transmission	engine	vin	stock number	vehicle history	seller name	price history
0	https://www.cars.com/vehicledetail/5ee3ae04-69...	2018 Hyundai Santa Fe Sport 2.4L	79,497 mi.	$17,899	Good Deal $2 under	Nightfall Blue	Beige	Front-wheel Drive	21 27	Gasoline	6-Speed Automatic	2.4L I4 16V GDI DOHC	5XYZT3LB3JG542718	13669	{'Accidents or damage': 'None reported', '1-ow...	ETHIO Motors	[('9/21/21', 'Listed', '$20,299'), ('2/04/22',...
1	https://www.cars.com/vehicledetail/c99e5035-76...	2017 RAM 1500 Express	80,326 mi.	$27,900	This is a good deal.	Silver	Black	Four-wheel Drive	16-23	Gasoline	8-Speed Automatic	5.7L V8 16V MPFI OHV	3C6RR7KT1HG533082	KBB2955	{'Accidents or damage': 'None reported', '1-ow...	Custom Cars West	[('3/15/22', 'Listed', '$30,985'), ('6/01/22',...
2	https://www.cars.com/vehicledetail/4e79fa3a-ba...	2019 Nissan Kicks SR	92,959 mi.	$18,950	This is a fair deal. Why?	Monarch Orange / Super Black	Charcoal	Front-wheel Drive	31-36	Gasoline	Automatic CVT	1.6L I4 16V MPFI DOHC	3N1CP5CU6KL510673	KBB3225	{'Accidents or damage': 'None reported', '1-ow...	Custom Cars West	[('4/16/22', 'Listed', '$19,980'), ('6/01/22',...
3	https://www.cars.com/vehicledetail/86a0d096-32...	2019 RAM 1500 Laramie	59,176 mi.	$39,950	Great Deal $968 under	Gray	Black	Four-wheel Drive	19 24	Gasoline	8-Speed Automatic	5.7L V8 16V MPFI OHV	1C6SRFDT5KN869809	KBB3246	{'Accidents or damage': 'At least 1 accident o...	Custom Cars West	[('6/01/22', 'Listed', '$40,523'), ('7/14/22',...
4	https://www.cars.com/vehicledetail/36e1be15-f2...	2015 Jeep Wrangler Unlimited Rubicon	135,385 mi.	$26,950	This is a good deal.	Blue	Black	Four-wheel Drive	16 21	Gasoline	-	3.6L V6 24V MPFI DOHC	1C4BJWFG2FL530006	KBB3145	{'Accidents or damage': 'None reported', '1-ow...	Custom Cars West	[('1/21/22', 'Listed', '$29,950'), ('3/22/22',...

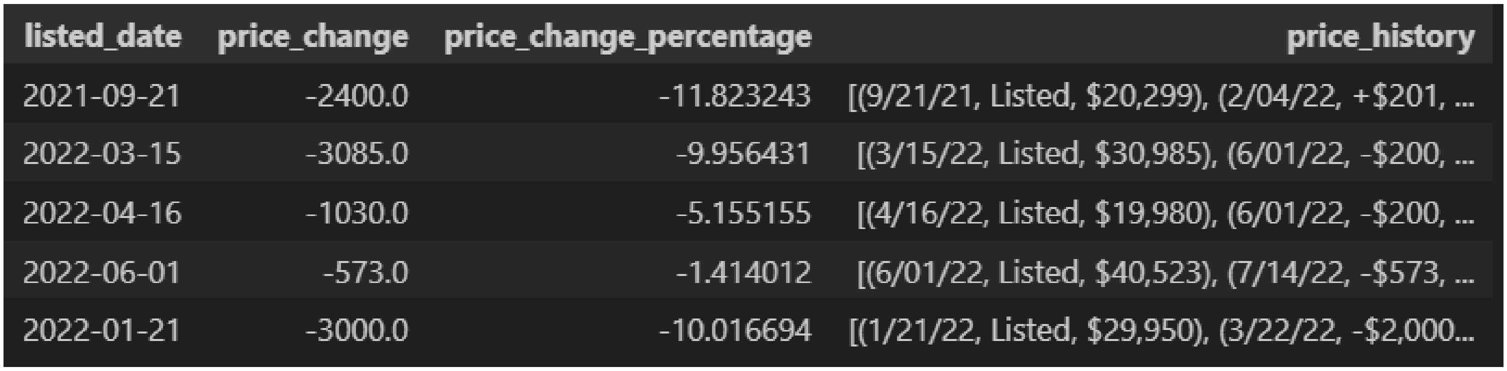
Once an adequate amount of data had been collected, the data set was to be processed and cleaned before data analysis could take place. The first step to cleaning was addressing number formatting from the web scraping. There were a number of characters or formats of data that existed to assist human readability, but they hurt a model's ability to understand the field. Some examples of this were the listing mileage and the primary price fields. These contained symbols and abbreviations, such as a dollar sign or the abbreviation for mile, to help a reader on the website. Thus, these were removed, and the fields were converted to integers. The MPG field was also designed for human readability by giving a range of MPG values for the vehicle. This field was separated into a high and low MPG so the information given could be better modeled.

The next step taken to clean the data set was to normalize the accident history field. Within the accident history, there are five pieces of information that are useful to the history of the car stored as a dictionary. This dictionary had to be removed, normalized, and then attached back to the data set as five new fields called Open recall, 1-owner vehicle, accidents reported, personal use only, and clean title. A few of these fields required more attention as they needed to change to the Boolean datatype to reflect the field more accurately, such as 1-owner history being either “yes” or “no”.



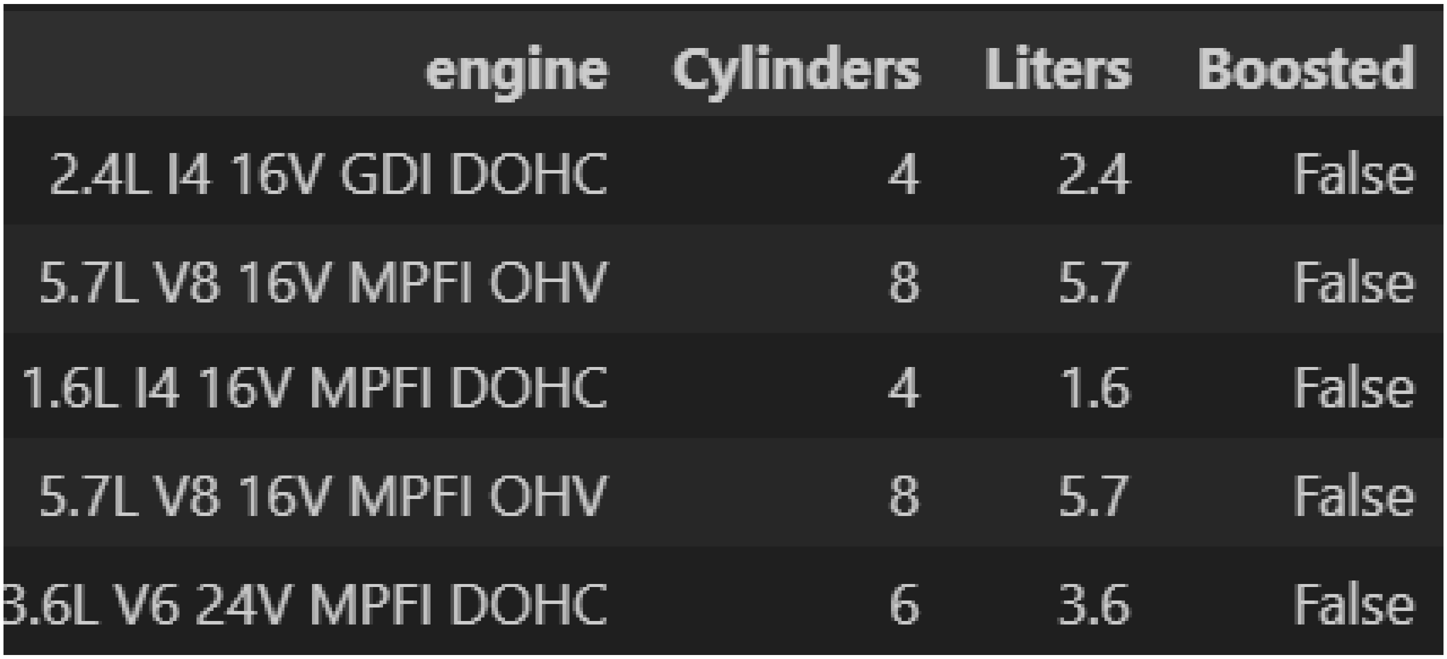
Accidents or damage	1-owner vehicle	Personal use only	Open recall	Clean title	vehicle_history
None reported	Yes	No	At least 1 open recall reported	NaN	{'Accidents or damage': 'None reported', '1-ow...
None reported	No	Yes	NaN	NaN	{'Accidents or damage': 'None reported', '1-ow...
None reported	No	Yes	NaN	NaN	{'Accidents or damage': 'None reported', '1-ow...
At least 1 accident or damage reported	No	No	At least 1 open recall reported	NaN	{'Accidents or damage': 'At least 1 accident o...
None reported	No	Yes	At least 1 open recall reported	NaN	{'Accidents or damage': 'None reported', '1-ow...

Another field that required cleaning was price history. This field contains a list of dates and price changes that describe how the listed car changed in price over time. This list was broken into three new fields describing the original listing date, the total change in price, and the percentage change in price. This provided a more intuitive look at the history of the listing by giving a clearer view of how many days the car has been available for purchase and how much the seller has adjusted that price during the same time. Cars.com only lists a price history for cars that have had their price adjusted, meaning that a large portion (about 25%) of the cars did not have this information. For this reason, these fields were elected to be used for exploratory data analysis and not for an independent variable within the machine learning and regression models.



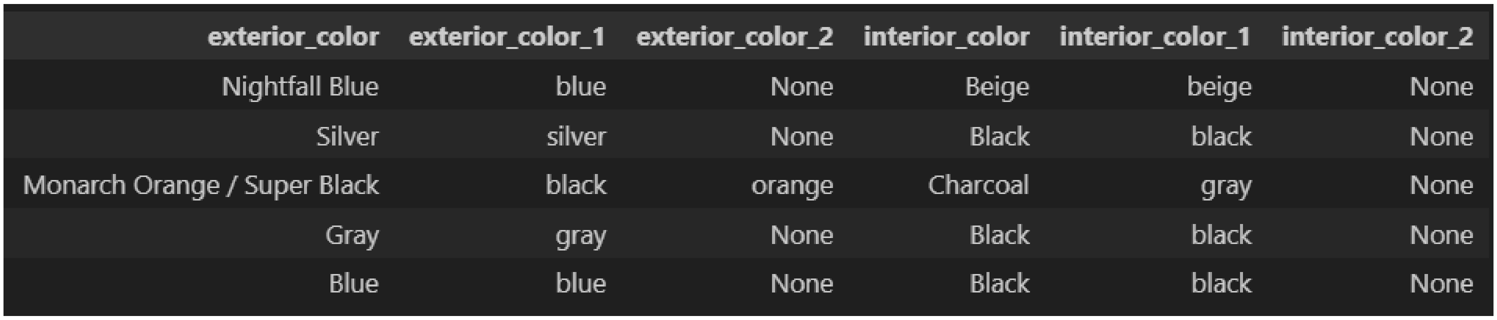
listed_date	price_change	price_change_percentage	price_history
2021-09-21	-2400.0	-11.823243	[(9/21/21, Listed, $20,299), (2/04/22, +$201, ...
2022-03-15	-3085.0	-9.956431	[(3/15/22, Listed, $30,985), (6/01/22, -$200, ...
2022-04-16	-1030.0	-5.155155	[(4/16/22, Listed, $19,980), (6/01/22, -$200, ...
2022-06-01	-573.0	-1.414012	[(6/01/22, Listed, $40,523), (7/14/22, -$573, ...
2022-01-21	-3000.0	-10.016694	[(1/21/22, Listed, $29,950), (3/22/22, -$2,000...

Next, the engine field was broken up into the most important features. The features that were deemed most important were based on research and discussion with other people who had extensive knowledge about engines. They ended up being the number of cylinders, the volume of each cylinder, and whether the engine is boosted. Extracting these elements required an understanding of how engines are named, examining discrepancies between how engines were listed in the dataset, and careful regular expressions to handle every way an engine was listed. Cylinders were typically listed as the arrangement pattern represented as a letter followed by the number of cylinders – for example V8 means the cylinders are arranged in a V shape and there are 8 cylinders. Liters were typically listed as a number followed by an L, though a few had the word "Liter" written out, so capturing that was much simpler. Lastly, the Boosted feature was simply the existence of the words "turbo" or "supercharge" in the engine.



engine	Cylinders	Liters	Boosted
2.4L I4 16V GDI DOHC	4	2.4	False
5.7L V8 16V MPFI OHV	8	5.7	False
1.6L I4 16V MPFI DOHC	4	1.6	False
5.7L V8 16V MPFI OHV	8	5.7	False
3.6L V6 24V MPFI DOHC	6	3.6	False

Similarly, the two fields of exterior color and interior color needed to be altered into a more useable state. These two fields contained a description of the color of the vehicle as a categorical variable, but the total amount of unique color descriptions added too much granularity to the field. To fix this problem, a color map was developed to map the descriptions in the field to a more standard description. This turned more unusual colors, such as cement or café latte, to more standard categories, like gray and brown. Some mappings had more than one color listed, but this proved to be a small enough subset to be ignored. This adjusted the total number of unique colors for exterior and interior color from 6,020 and 6,266 respectively to a much more manageable 13 and 12.

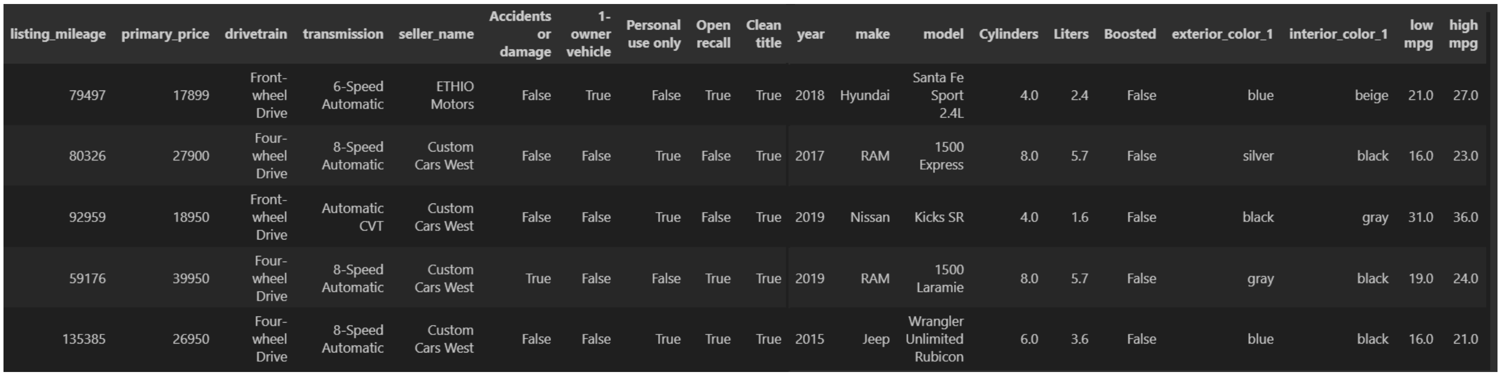


exterior_color	exterior_color_1	exterior_color_2	interior_color	interior_color_1	interior_color_2
Nightfall Blue	blue	None	Beige	beige	None
Silver	silver	None	Black	black	None
Monarch Orange / Super Black	black	orange	Charcoal	gray	None
Gray	gray	None	Black	black	None
Blue	blue	None	Black	black	None

The final few cleaning steps involved breaking up listing title into smaller fields and removing unnecessary features. Listing title was split into vehicle make, model, and year. The fields that were unique to each vehicle, such as VIN or listing link, were removed to create the dataset that would be used to train and test the models. Fuel type was verified to be gasoline for all listings and was thus also removed.

After the columns were cleaned, missing values were checked. As previously mentioned, price history data was missing for 25% of the data. The linear regression and K-means regression both performed better with the columns related to price history dropped. This is probably due to there being over 1,000 more rows with the price history columns dropped because of how much was missing. Thus, the price history features from the models were removed. Other than that, there weren't any serious issues with missing data, and 80% of the original 6,000 data points was able to used.

Below is a sample of the data after being cleaned.

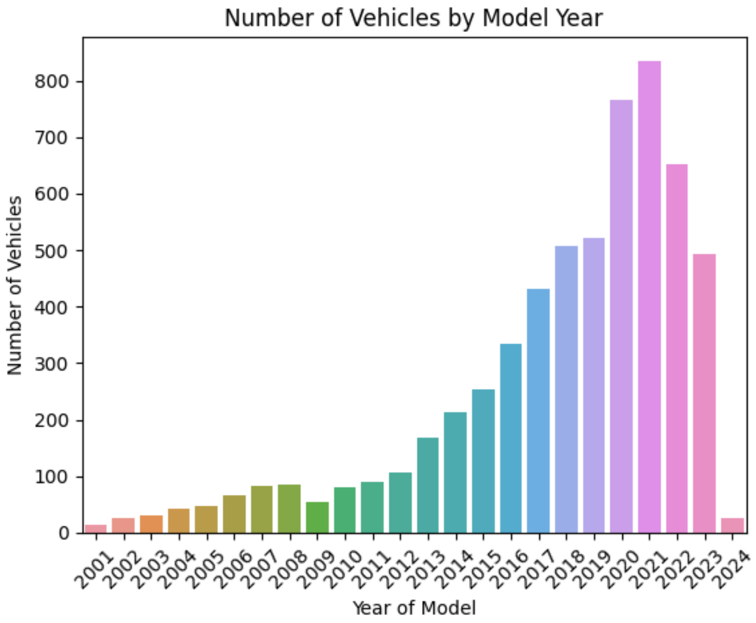


listing_mileage	primary_price	drivetrain	transmission	seller_name	Accidents or damage	1-owner vehicle	Personal use only	Open recall	Clean title	year	make	model	Cylinders	Liters	Boosted	exterior_color_1	interior_color_1	low mpg	high mpg
79497	17899	Front-wheel Drive	6-Speed Automatic	ETHIO Motors	False	True	False	True	True	2018	Hyundai	Santa Fe Sport 2.4L	4.0	2.4	False	blue	beige	21.0	27.0
80326	27900	Four-wheel Drive	8-Speed Automatic	Custom Cars West	False	False	True	False	True	2017	RAM	1500 Express	8.0	5.7	False	silver	black	16.0	23.0
92959	18950	Front-wheel Drive	Automatic CVT	Custom Cars West	False	False	True	False	True	2019	Nissan	Kicks SR	4.0	1.6	False	black	gray	31.0	36.0
59176	39950	Four-wheel Drive	8-Speed Automatic	Custom Cars West	True	False	False	True	True	2019	RAM	1500 Laramie	8.0	5.7	False	gray	black	19.0	24.0
135385	26950	Four-wheel Drive	8-Speed Automatic	Custom Cars West	False	False	True	True	True	2015	Jeep	Wrangler Unlimited Rubicon	6.0	3.6	False	blue	black	16.0	21.0

After the data was cleaned, categorical variables were one-hot encoded, and a train-test split was performed.

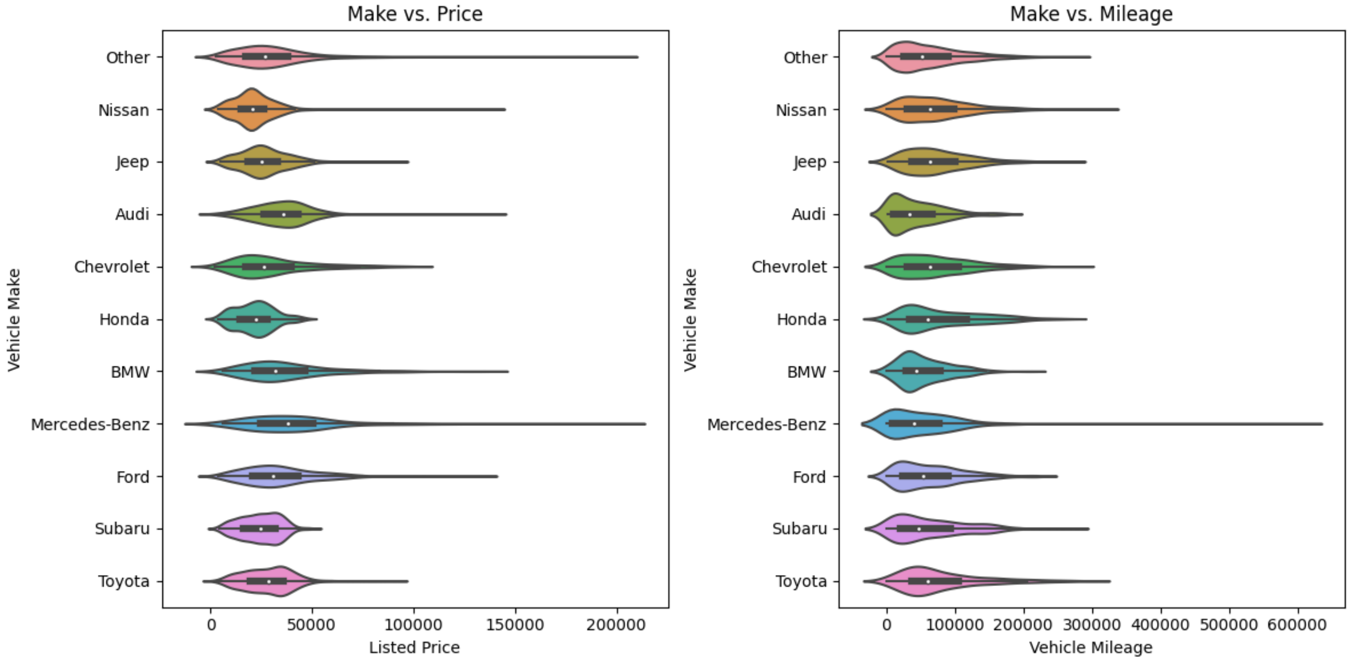
Exploratory Data Analysis

With the data in the correct format, the data could start to be explored. The first thing analyzed was the distribution of vehicles available by model year. This distribution shows the total number of vehicles found available on the website that were made since the year 2000. The most common car found in the data set was

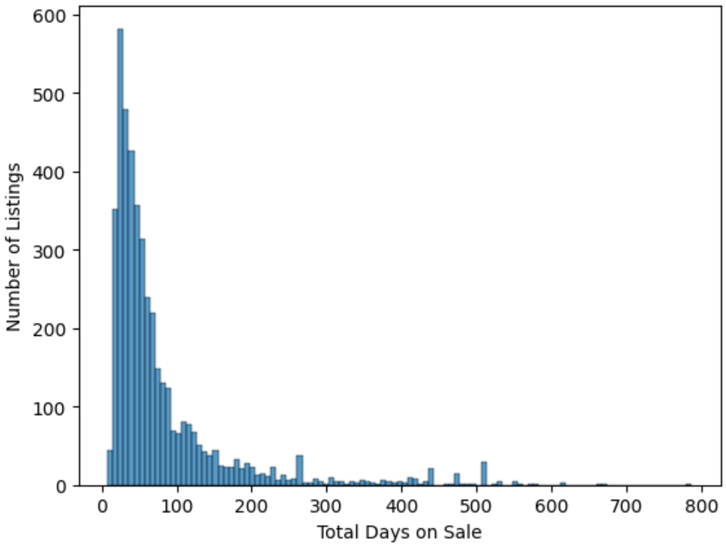


produced in the year 2021, but the most interesting feature in the distribution is the shoulder that appears in the years 2018 and 2019. With new car production reportedly dropping during the COVID-19 pandemic, it could be possible that a higher focus was placed on buying used cars made before or at the start of the pandemic to compensate for the lack of new vehicles available due to low production.

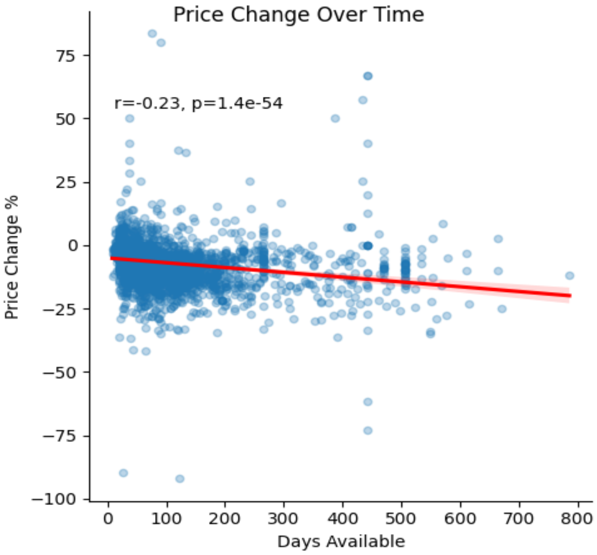
The next most prominent feature to be explored was the make of the vehicles. In total, there were 49 unique different makes found in the data set with the top contributors being Ford, Toyota, Jeep, Chevrolet, Subaru, Nissan, BMW, Honda, Audi, and Mercedes-Benz accounting for 3,847 vehicles of the dataset. Data for these vehicle makes was aggregated to get a better picture of the distribution of price and miles for these groups. The distributions for both price and mileage both had similar shapes with high density toward the lower end of the range of density, but every group seemed to have a long tail that imply that each group has a small handful of vehicles with unusually large values.



After gaining a better understanding of the prices within the data set, the next goal was to find out how the prices change over time. Each vehicle was assigned a number corresponding to the total number of days the vehicle had been for sale, which was calculated by the original date the vehicle was listed on the website.



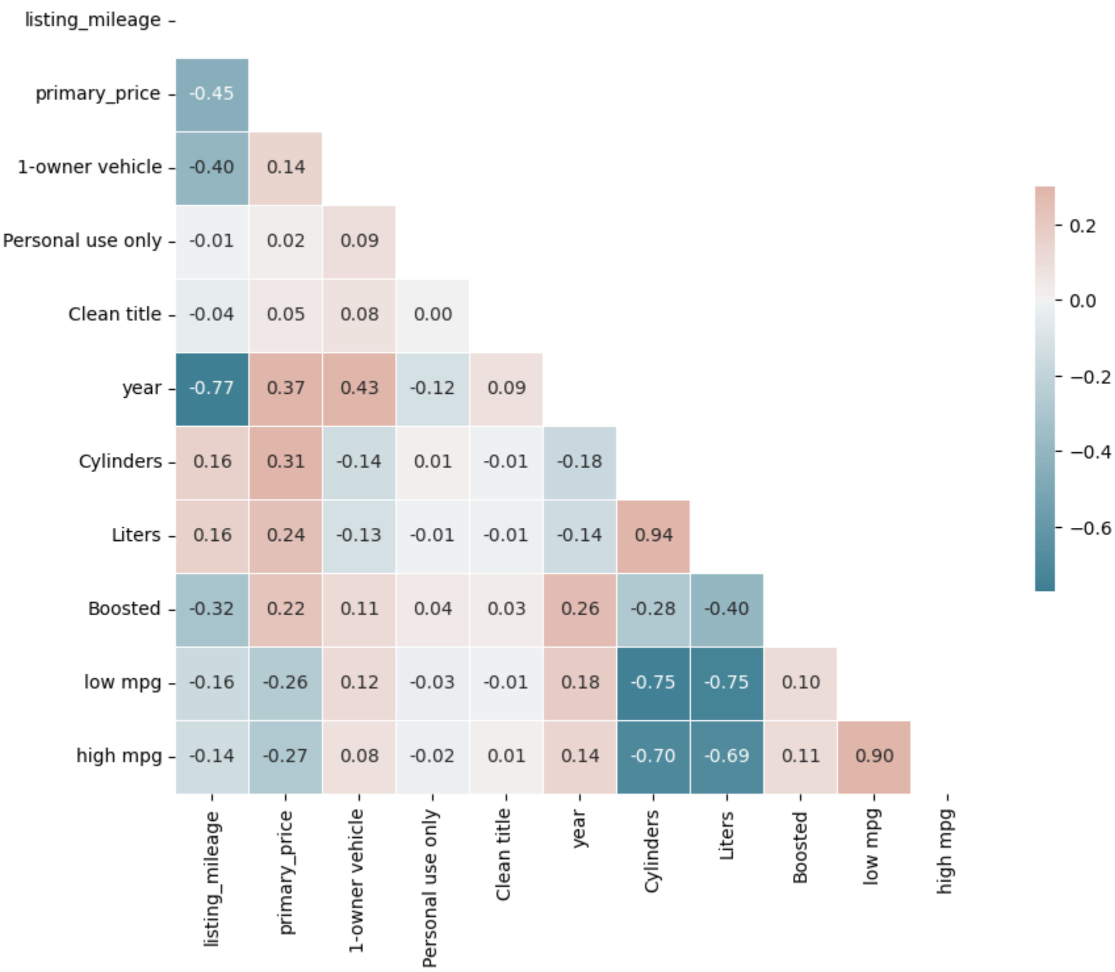
Plotting this field to a histogram showed that most vehicles had only been for sale for 150 days or less. It was also seen that often the cars which had been available for longer periods of time experienced a higher drop in price compared to those which had not been on the market for as long. This relationship was explored in a linear regression plot. The plot implies that the two fields are most likely related, with a p-value 1.4×10^{-54}, but that the relationship is most likely not a linear relationship, with an R^2 of 0.3.



Model Selection

To develop a tool that could assist a buyer in navigating the used car market, three models were developed and tested to try to predict the price of the vehicle based on the features found in the listing. These models were Ordinary Least Squares (OLS) linear regression, a Deep Neural Network (DNN) and a K-means clustering regression. The correlation matrix for the variables in the dataset that were passed into the model is shown below.

Correlation Matrix for Model Variables



	listing_mileage	primary_price	1-owner vehicle	Personal use only	Clean title	year	Cylinders	Liters	Boosted	low mpg	high mpg
listing_mileage											
primary_price	-0.45										
1-owner vehicle	-0.40	0.14									
Personal use only	-0.01	0.02	0.09								
Clean title	-0.04	0.05	0.08	0.00							
year	-0.77	0.37	0.43	-0.12	0.09						
Cylinders	0.16	0.31	-0.14	0.01	-0.01	-0.18					
Liters	0.16	0.24	-0.13	-0.01	-0.01	-0.14	0.94				
Boosted	-0.32	0.22	0.11	0.04	0.03	0.26	-0.28	-0.40			
low mpg	-0.16	-0.26	0.12	-0.03	-0.01	0.18	-0.75	-0.75	0.10		
high mpg	-0.14	-0.27	0.08	-0.02	0.01	0.14	-0.70	-0.69	0.11	0.90	

The first model was chosen to be OLS regression to test the viability of less complex methods within this setting, and to also act as a baseline for the more complex method. This model was fed data on a random 80:20 train-test split with the categorical variables one-hot encoded to produce the proper vectors for OLS.

This model attempts to predict price by fitting a straight line through the data that minimizes the distance between all points in the data set to find the best fitting relationship.

The second model created was also designed to mimic a field within the data from the website. Deal gauge is an assigned value of fair, good, or great from cars.com that assess the quality of the deal of the listed vehicle based on over 100 factors.[3] Thus, the second model was chosen to be a K-means to attempt to group the fields into similar groups. K-means clustering is the process of grouping data into different groups by assessing the similarity of the records within the data set. The clustering model was trained on a data set with all categorical variables one-hot encoded to predict the deal gauge outcome.

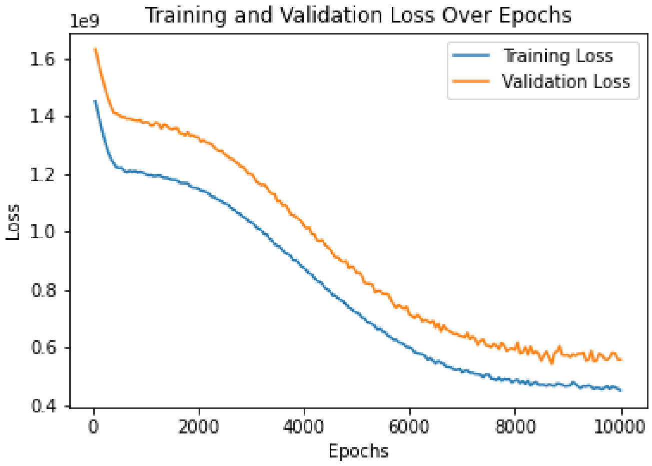
Lastly, the final model was chosen to be a deep neural network. A deep neural network is a machine learning model that simulates multiple layers of neurons to learn and predict outcomes from various fields. This model was chosen for its flexibility to handle different kinds of data as well as giving the data set a chance to be put through a more advanced model. This model was trained on the same data set as the regression models. Hyperparameters were tuned by training the model over 1,000 epochs with 81 different combinations of hyperparameter values. The best performing hyperparameters were found to be a learning rate of 0.0001, 1 hidden layer, 64 neurons per layer, and a dropout rate of 0.5.

Results and Discussion

The OLS regression model's performance on the testing data returned a root mean square error of 16,546 with an R^2 of 0.605. A high root mean square error with a questionable R^2 implies that the linear relationship produced by the model did not perform very well when attempting to predict a vehicles price based on the fields it was given. The model did find multiple fields within the regression that do have a statistically significant relationship with the listing price, such as listing mileage, 1-owner vehicle, and accidents or damage reported. One interesting note is that the R^2 for the training data set showed a much greater agreement at a value of 0.983. This large difference in the R^2 values shows that the model could be overfit on the training set. Any future edition of this model should explore removing features to limit the overfitting of the training set.

The K-means clustering set was tested for its accuracy when compared to the deal gauge found on cars.com. The model returned an accuracy score of 0.3207 for the tested data set. This means that only 32% of the time the clustering algorithm was able to put the vehicle in the correct deal grouping. Given that there are only three possible groups to be put in of fair, good, and great, the clustering model showed similar performance to random assignment of the groups.

Finally, the Deep Neural Network was trained and tested on the same dataset as the two regressions. After training, the DNN stabilized at a loss value of 465,916,192 on the training data while the validation set stabilized at a 561,570,880 value after 10,000 epochs.



Once the model had stabilized, its ability to predict a vehicles price was tested for accuracy. This returned a R^2 of 0.190 with a root mean square error of 23,697. Comparing this to the OLS Regression's of 0.605 and 16,546, it was seen that the increase in complexity in the model did not return better predictions for this data set. One advantage the DNN did provide was it was not as prone to over fitting on the training data. The training statistics for R^2 and root mean square error returned values of 0.1960 and 21,585 which performed much closer to the testing set compared to the OLS model counterpart.

One issue encountered in all models was the presence of categorical variables. None of these models are specifically designed to handle categorical variables, but the issue is compounded by the data set. Specifically, the field that describes the vehicle model has a total of 1,966 unique car models. This number of unique categories is too much for these models to handle appropriately. To improve this for future investigations, the vehicle model field would have to be mapped to more standardized options similarly to how the color fields were handled.

Conclusion

An exploratory data analysis and data modeling study was performed on a data set of 6,002 vehicles collected from web scraping used vehicles within a ten-mile radius of the University of Denver from cars.com, with the goal of building a tool to assist customers with navigating the used car market online. Three models were built to predict price from the features found within the listings on cars.com. The Ordinary Least Squares regression model was unable to accurately predict the price, but it did indicate that the fields collected did have statistical significance to price. The K-means clustering model was built to mimic the deal gauge found within the website, but the model was unable to correctly classify the vehicles into the same groups as cars.com. Finally, the last model applied was a Deep Neural Network. The DNN model showed little agreement to the testing data, but it was able to show resistance to overfitting with our data set.

References

1. Cable News Network. (2021, March 9). How COVID has changed the way used cars are sold | CNN business. CNN. https://www.cnn.com/2021/03/09/success/online-used-car-shopping-feseries/index.html

2. Domonoske, C. (2023, March 18). *Why car prices are still so high - and why they are unlikely to fall anytime soon*. NPR. https://www.npr.org/2023/03/18/1163278082/car-prices-used-cars-electric-vehicles-pandemic
3. Cars.com. (n.d.). *Cars.com leverages algorithm to redefine best-in-class car search and shopping experience: Press release*. Cars.com. https://www.cars.com/articles/cars-com-leverages-algorithm-to-redefine-best-in-class-car-search-and-shopping-experience-1420697515716/
4. evanhollier/Cars_Project (github.com)

Supporting Information

Original Listing Date for New Incoming Listings

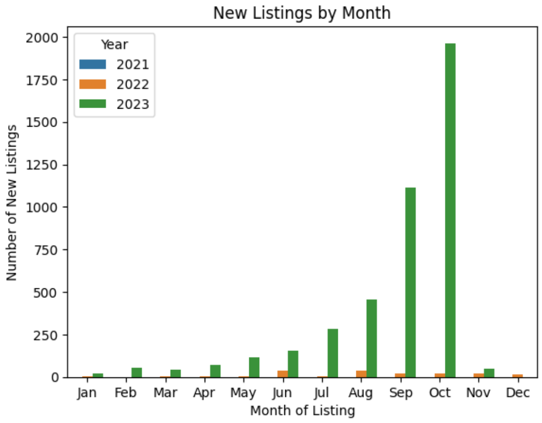


Figure 1: Bar Chart showing when each vehicle in the data set first appears on cars.com. Most listings have occurred in recent months of 2022.

Facet Grid of Listing Price vs Vehicle Miles

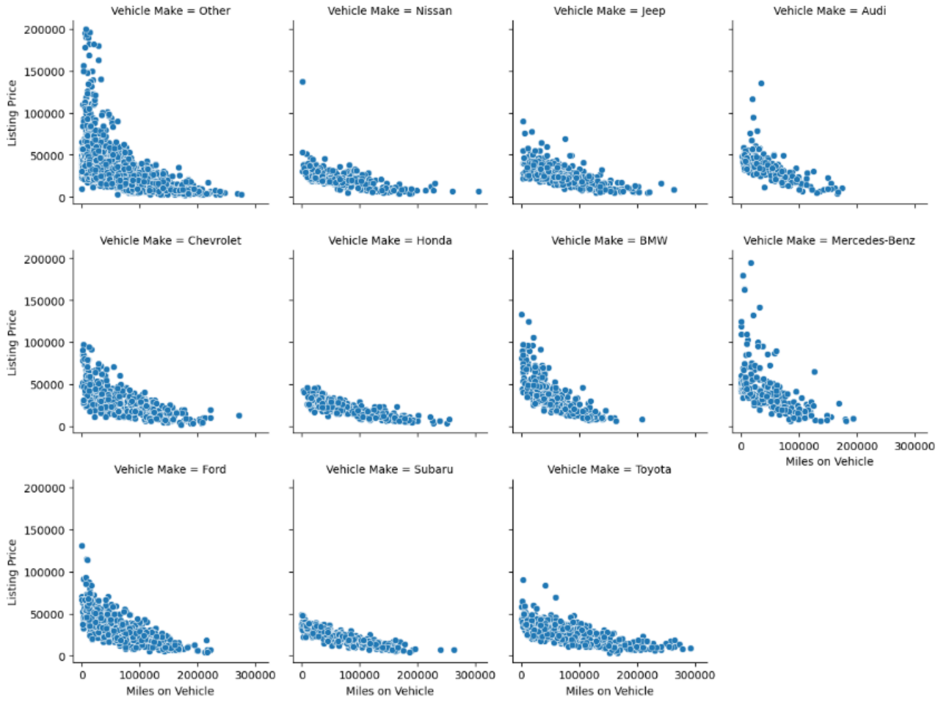
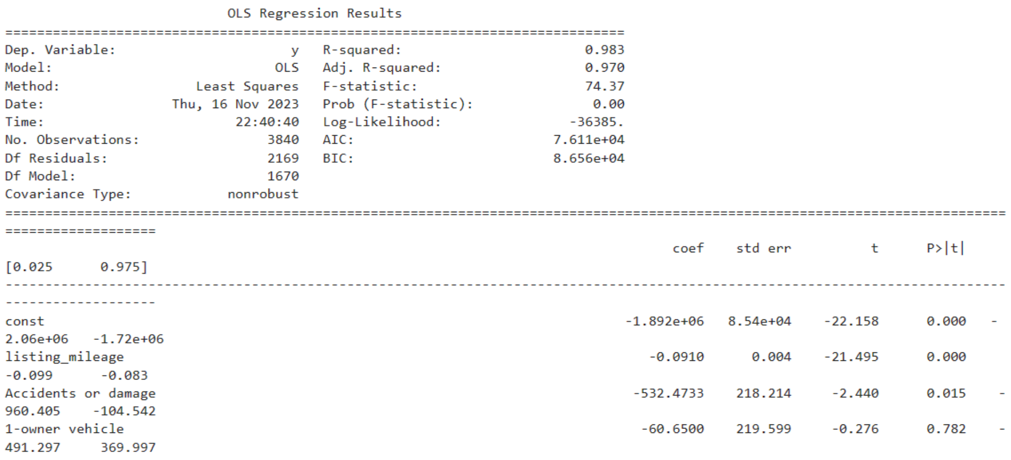


Figure 2: A facet grid of the scatter plot of the main numeric fields found in cars.com, price and mileage. The plots are separated into the top ten vehicle makes plus an "other" category. The scatter plots imply that having a mileage and price are inversely related.

Linear Regression



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.983
Model:                            OLS   Adj. R-squared:                  0.970
Method:                 Least Squares   F-statistic:                     74.37
Date:                Thu, 16 Nov 2023   Prob (F-statistic):               0.00
Time:                        22:40:40   Log-Likelihood:                -36385.
No. Observations:                3840   AIC:                         7.611e+04
Df Residuals:                    2169   BIC:                         8.656e+04
Df Model:                        1670
Covariance Type:            nonrobust
==============================================================================================================================
===================
                                                                                    coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------------------------------------------------------
-------------------
const                                                                         -1.892e+06   8.54e+04    -22.158      0.000   -
2.06e+06   -1.72e+06
listing_mileage                                                                  -0.0910      0.004    -21.495      0.000
-0.099      -0.083
Accidents or damage                                                            -532.4733    218.214     -2.440      0.015    -
960.405    -104.542
1-owner vehicle                                                                 -60.6500    219.599     -0.276      0.782    -
491.297     369.997
```

Figure 3: A sample of the output of the OLS regression run in python showing coefficient, error, and significance for input variables. See modeling.ipynb[5] for full output.

Deal Gauge Breakdown

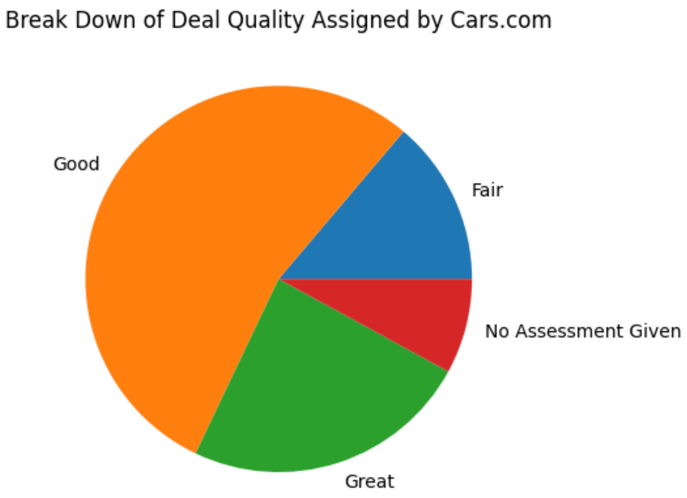
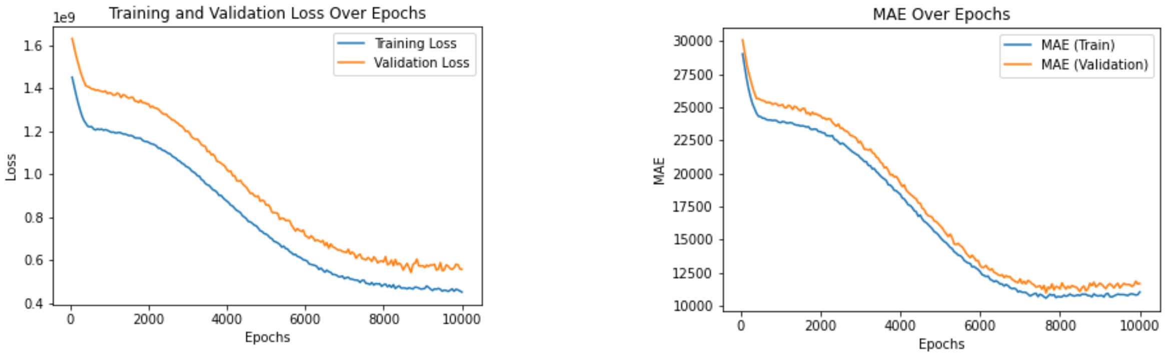


Figure 4: A pie chart showing the distribution of the assigned deal gauge from cars.com. This was attempted to be reproduced by the K-means clustering algorithm.

Deep Neural Network Epoch Graphs



Figures 5,6: showing how the performance and learning of the deep neural network performed across the training epochs. The graphs show that the statistics begin to stabilize around 8000 epochs.