Predictive Analytics Using SAS

Pricing Cars For Sale on CraigsList

Suhas Patil

Introduction

The used vehicle market accounts for about 70% of the total number of vehicles sold in the United States. Since 2009, used vehicle sales have increased from 35.5 to 40.42 million a total of 13.24% increase in the number of sold used vehicles.[1] This report explores the many variables taken into account when purchasing a used vehicle, such as make and model, and looks at how they affect the vehicle's sale price. The data for the analysis found in this report is based off a continually updated dataset which contains information about cars being sold on CraigsList. CraigsList is an open, web-based platform for C2C suppliers and buyers to trade goods that contains the largest database when it comes to used vehicles in the United States. The basis for the analysis is to see if we can predict the price of a vehicle accurately with parameters given in the dataset.

Data Description

The dataset around CraigsList used vehicles for sale is available on Kaggle.com. The data is scraped from the CraigsList site every few months and contains relevant information around car sales such as price, condition, manufacturer, and latitude/longitude.[2] The data is entered by an everyday, average user which means there is significant human error present and many missing or erroneous values (i.e. $1 as the price of an entire vehicle, wrong VIN etc.). How we accounted for these kinds of issues will be addressed in the preprocessing data section below.

Preprocessing Data

The dataset contains many missing values, as well as erroneous values that did not make common sense. How we handled these issues is listed below per feature of the dataset which we deemed potentially significant in the modeling of the price of posted vehicle:

Price: If the price was greater than $200,000 or less than $50, then we deleted that row from the dataset since this is the variable around a posting we're attempting to predict with our model. More on this is below in the "Pre-Processing Our To-Be-Predicted Variable: Price" section.

Transmission: If the transmission was listed as a blank or as other, we changed the value to automatic. We did this for two reasons, because 1) the data showed that automatic transmission was 88% of the values and 2) the dataset is US vehicles only and the most popular transmission type in the US is automatic.

Title_Status: If the title status was blank, we changed it to clean because clean was the most popular value (mode) in the data (95%) and we figured that if someone was selling a car not on the black market, they'd probably have a good title!

Manufacturer: With this feature, we attempted to parse the manufacturer from the make field by using the list of non-blank manufacturers (i.e. the make may read "Ford F-150" instead of Ford being the manufacturer and make being F-150). We also cleaned this column by changing entries like "Harley" to "Harley-Davidson" or "Infinity" to "Infiniti." Any that were still left blank after these steps were changed to "unknown_manufacturer."

Make: In this field, we cleaned this column by changing entries like "1500" to "Silverado 1500" or "f150" to "F-150;" any that were still left blank after these steps were changed to "unknown_make".

Due to the limitation in processing power and time constraints, we are considering only the cars with popular make (at least 100 cars of the same make are present), which makes up the bulk of data we have at hand.

Size: We tried to predict the size of the vehicle from Type, but the logistic regression was inaccurate due to conflicting data in the column "type" which refers to the vehicle's class. We also tried this prediction using the make of the car, but due to the large number of various make and limitations of the memory and time on the server, we had to drop it.

Type, Cylinders, Condition, and Paint_Color: If these features were left blank, we changed them to "unknown_type," "unknown_cylinders," etc.

State_Fips: This value is specific to each state in the US. If this feature was left blank, we entered 99 as the value instead to create a new state_fips category for "unknown_state" (highest in the dataset was 56).

VIN: With the VIN, we created a new column called VIN_Flag; if the VIN was valid (non-blank and 17 characters long), the VIN_Flag column was 1 and otherwise it was 0.

Drive: We changed any blanks in this field to "4WD," "FWD," or "RWD" based off the existing values in the data, finding the most repetitive drive for each model of car.

Odometer: From this feature, we created a new column called "Odometer_Flag" which was 1 if the odometer was valid and 0 if invalid. A valid odometer entry means that the odometer is non-blank, greater than 1 (even a brand-new car purchased from a car lot has a couple of miles on it), and less than 750,000; to get this number, we did a Google search on "what's the highest number on an odometer" and found an article from AutoTrader.[3] Additionally, any invalid values as described above in the odometer column were changed to a predicted odometer reading from price, city, and year since these are features which we thought would have an effect on the average total distance the car is driven. Our group argued over the methodology around entering in modeled values for this column, but found that sellers with high odometer readings are less likely to enter their reading, making this a Missing not at Random issue and thus could potentially skew the data if the blank values were just deleted.[8] Though the accuracy of the prediction model was not high (44%), it seemed to be significantly better than the mean values. The *city* seemed to not be efficient for the model in reducing errors and we removed it to get an accuracy of 43%. The remaining were replaced with the mean reading. In the next section, you will see in a visual representation that there are 32.32% of invalid/missing data in the odometer feature that we replaced.

Year: Regarding this data, we researched online what year the first cars that are most similar to what we drive today were invented/released and found that 1884 was the first year.[4] Therefore, we deleted rows with blank years and changed any non-blanks that were before 1884 to 1884.

Fuel: If the fuel was blank, we changed it to gas since it's both the most common in the dataset (87%) and the most common fuel type for cars in the US (and this dataset is ONLY cars posted on Craigslist in the US). We confirmed this via the EPA.gov website which stated that in 2018, gasoline accounted for 92% of transportation fuel consumption.[5]

State Name and State code: These fields provide information on which state of US the car is being sold in. The same information is being provided by the state_fips and hence, we chose to exclude these features.

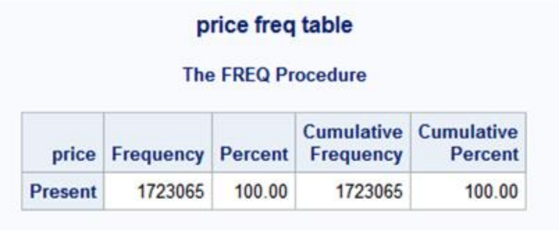
State name had a value called 'FAILED', which had no values in corresponding county or state related columns. After examining some of these values using their corresponding latitude and longitude coordinates in google maps, we found these locations to be outside of the United States and considered them as missing values which were handled in the state fips feature.

County Name and County Code: These fields provide information on the county in which the car is being sold. From our analysis, we found that county_fips and state_fips provide almost the same information regarding the price of the car as they were highly correlated. Since, we were already using the state_fips as a feature, we chose to exclude these.

Weather, Latitude, Longitude, URL, and Image URL: After researching and discussing these features, we chose to exclude it as it would not have an effect on the car price.

Pre-Processing Our To-Be-Predicted Variable: Price

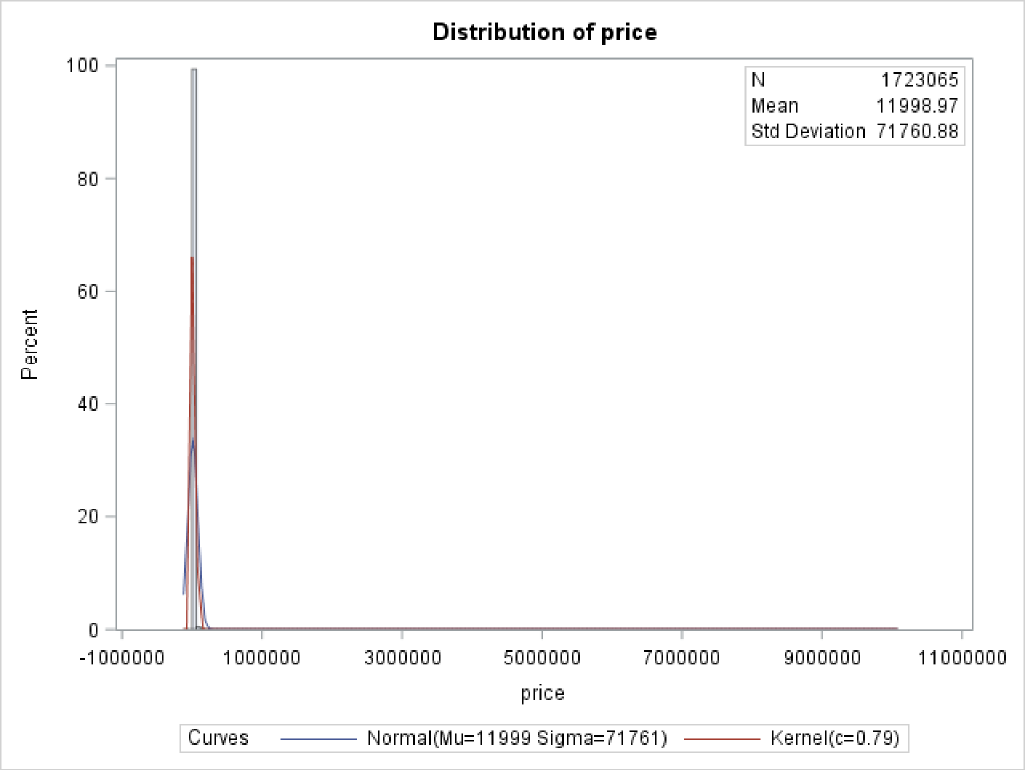
Our first observations of the price variable in this data set was that the raw data of price didn't have any missing values. We found this to be a promising sign; however, after digging deeper, we determined that the standard deviation was too large, meaning the mean and distribution are highly impacted by extreme values/outliers. There were significant amounts of $1 or $2 for the price, which after looking through postings on Craigslist, we realized these are not true values for the price, but ploys from sellers to get customers' attention. There were also high amounts that were 10 digits long, which we assumed were phone numbers instead of a price point, again a marketing ploy from sellers. Our findings are graphically represented below:



price freq table

The FREQ Procedure

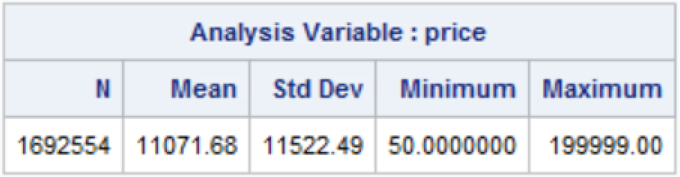
price	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Present	1723065	100.00	1723065	100.00



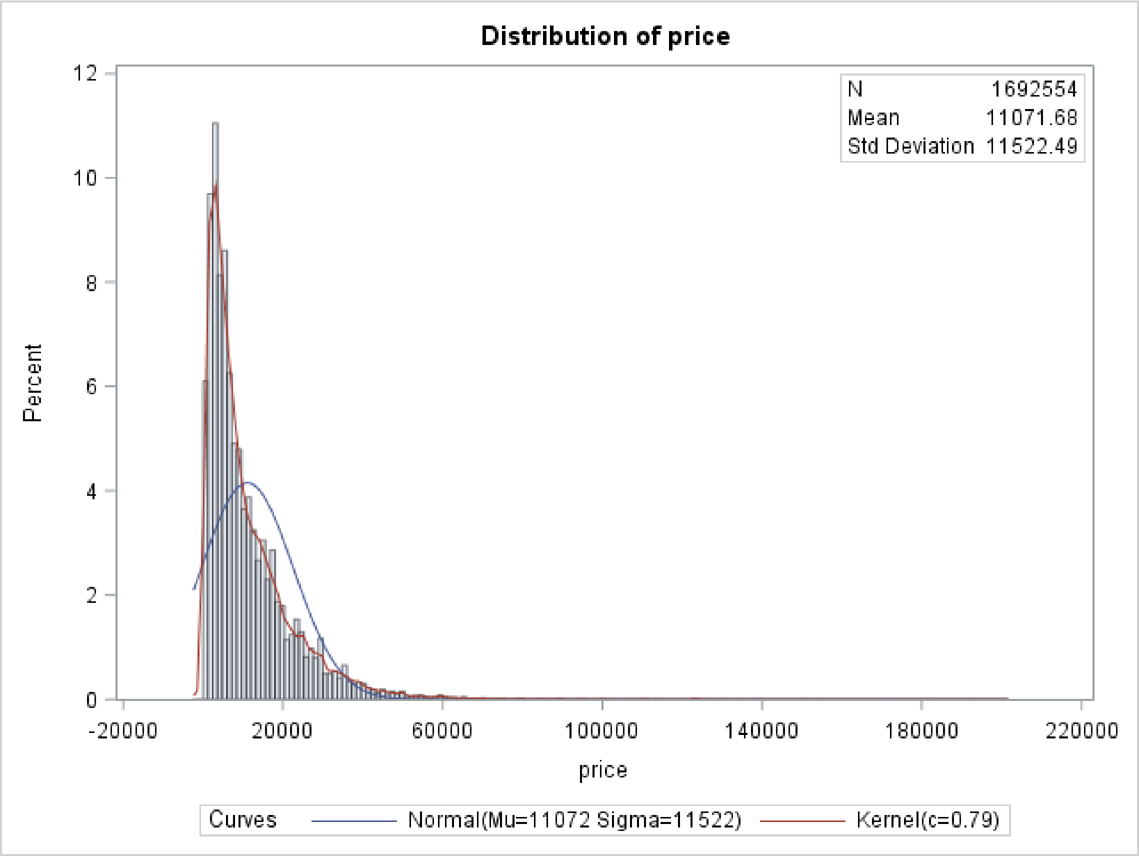
To overcome these issues with the raw data, we deleted rows from the data set which satisfied either of the below conditions:

1. Price greater than or equal to $200,000 or price less than $50 (cuts out the extreme/outlying values)
2. Price is non-numeric

After cleaning, the number of observations or ad postings decreased by 30,511 from 1,723,065 down to 1,692,554. The mean of the price did not show significant change, but the standard deviation dropped significantly, making the data much more reasonable and ready for further processing. Graphical representations of the new stats are below:



Analysis Variable : price				
N	Mean	Std Dev	Minimum	Maximum
1692554	11071.68	11522.49	50.0000000	199999.00



Exploratory Data Analysis

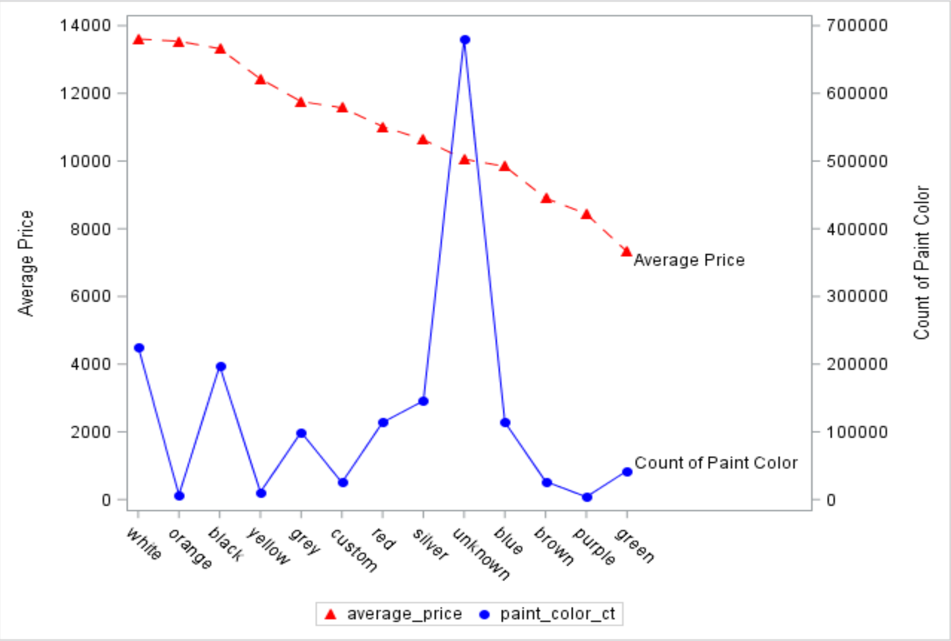
Upon digging into the dataset, we decided to attempt to predict the price at which a vehicle would be sold. We pre-determined that the variables which would most affect the price of a vehicle for sale (common-sense-wise) would be the paint color, the odometer reading, the year of manufacture, the manufacturer, the condition, the number of cylinders, the fuel type, the title status, the transmission type, whether or not the posting had an image, whether or not the posting had a VIN listed, the size, the make of the car, the vehicle class, the US state in which the car is sold, and the drive type.

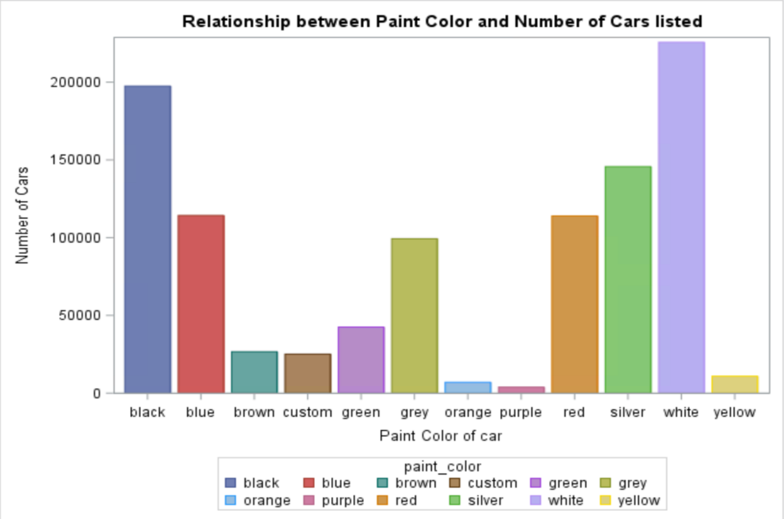
Special note: we ran ANOVA and logistic regression to find the relationship between the price and the categorical features for successful proof of significant variation between atleast one set of category groups of each feature on the price.

Feature: Paint Color

Our initial thoughts with paint color were that the more popular the color, the higher the price would be; however, some colors like orange and yellow are higher priced, yet not popular and blue is a popular color, but lower in price. However, there is a relationship seen between the color of the vehicle and its average price, so the paint_color instead of a flag as to whether it is a popular color should be tested in a model for significance.

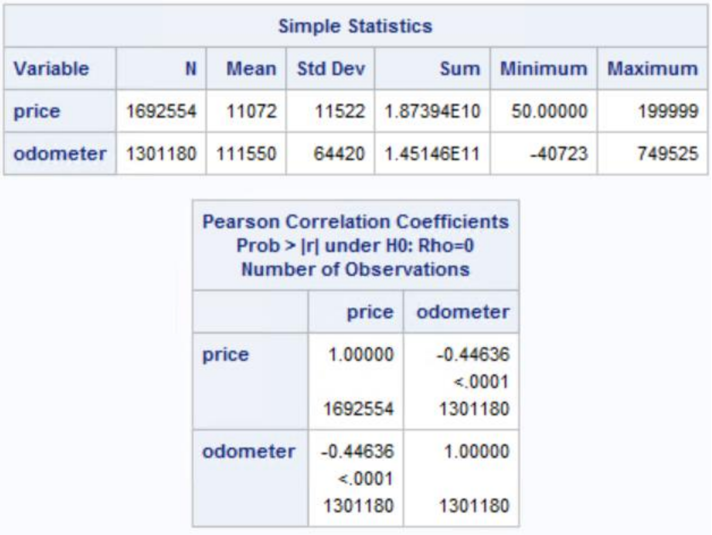
There are large numbers of black and white color cars in the list for sale.





Feature: Odometer Reading

As a rule of thumb, the higher the odometer, the lower the price. We can find this relationship in our dataset as well, from the correlation characteristics (-0.44) and also, as seen in the scatter plot below; the graphic represents the negative correlation between price and odometer.

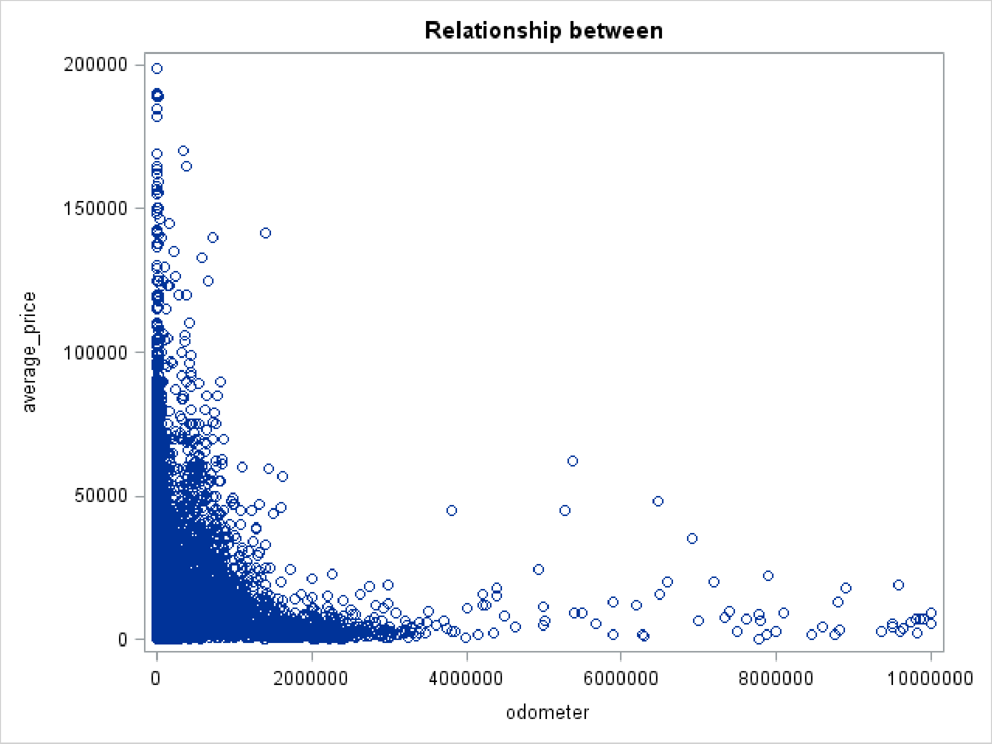


Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
price	1692554	11072	11522	1.87394E10	50.00000	199999
odometer	1301180	111550	64420	1.45146E11	-40723	749525

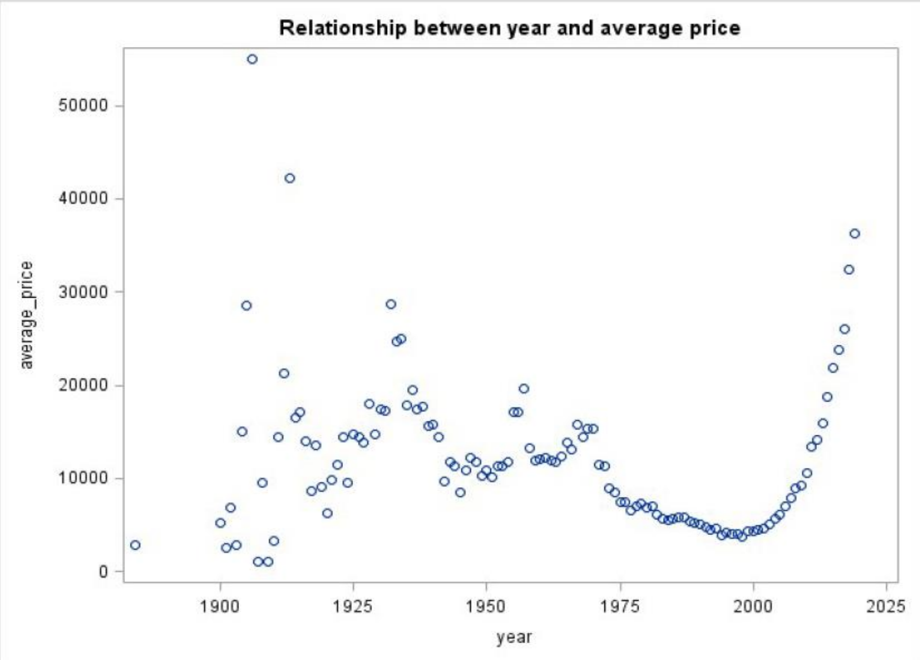
Pearson Correlation Coefficients
Prob > \|r\| under H0: Rho=0
Number of Observations

	price	odometer
price	1.00000 1692554	-0.44636 <.0001 1301180
odometer	-0.44636 <.0001 1301180	1.00000 1301180



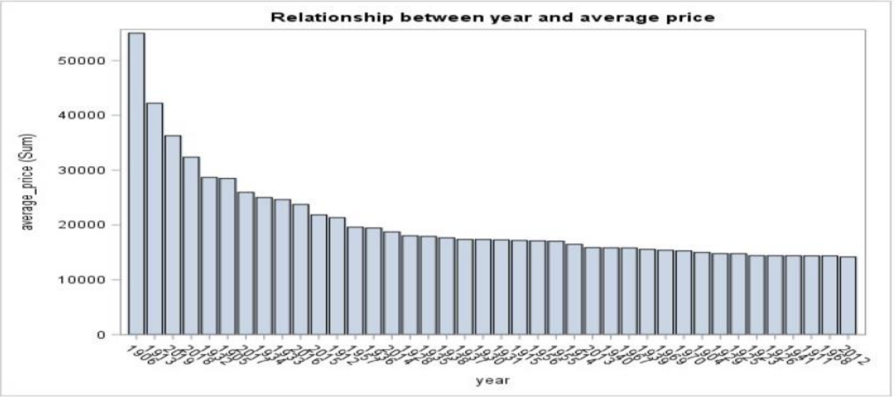
Feature: Year

During our initial discussion, we all agreed that the newer the year of the vehicle sold on the resale market, the higher it will be priced. And as seen in a graphical representation of the data set, a positive relationship seems to exist between year and the average price.

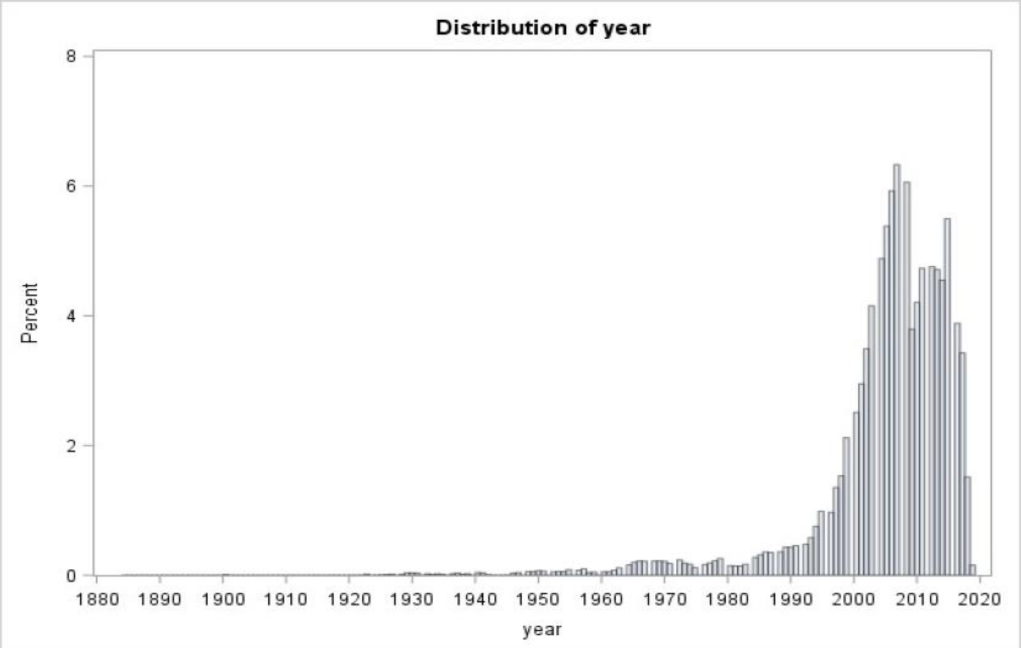


From the plot, we can see the price of cars drop with year but then start to increase as the year value decreases with some peaks going above the recent car prices. This might be due to the high price for vintage and classic cars in the market. Such cars are usually auctioned and bought by enthusiastic collectors for high prices. Since there are no gradual pricing changes or any measurable relationship of such cars with the price, we will not be able to predict the value of such cars without a data entry of the year or make feature in the dataset. Hence, year is a significant feature affecting the resale price of the car.

From the below plot, we can see most of years the top 40 average price of cars in the year list, to be from the 1900s.

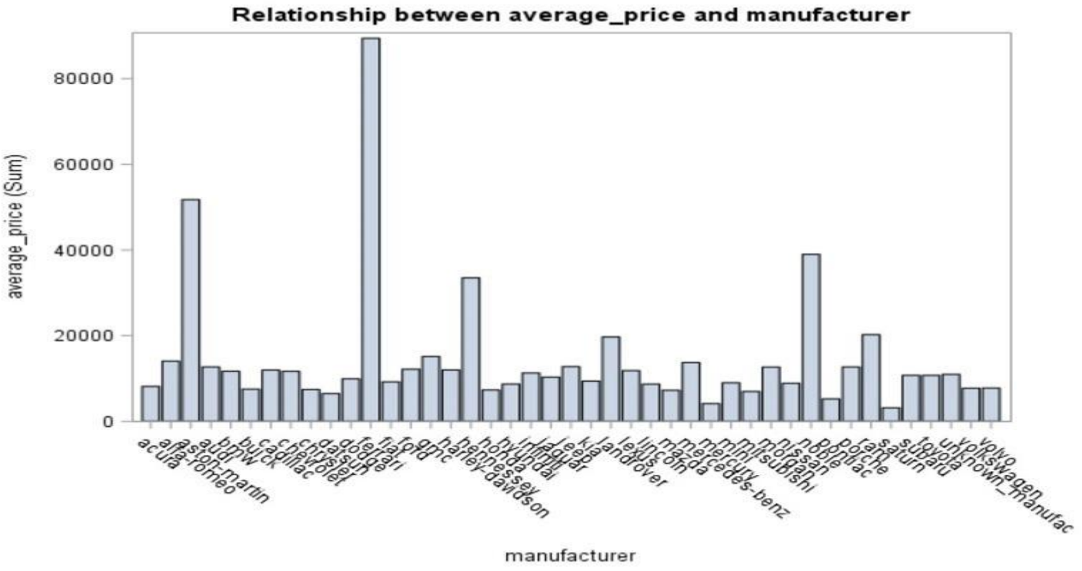


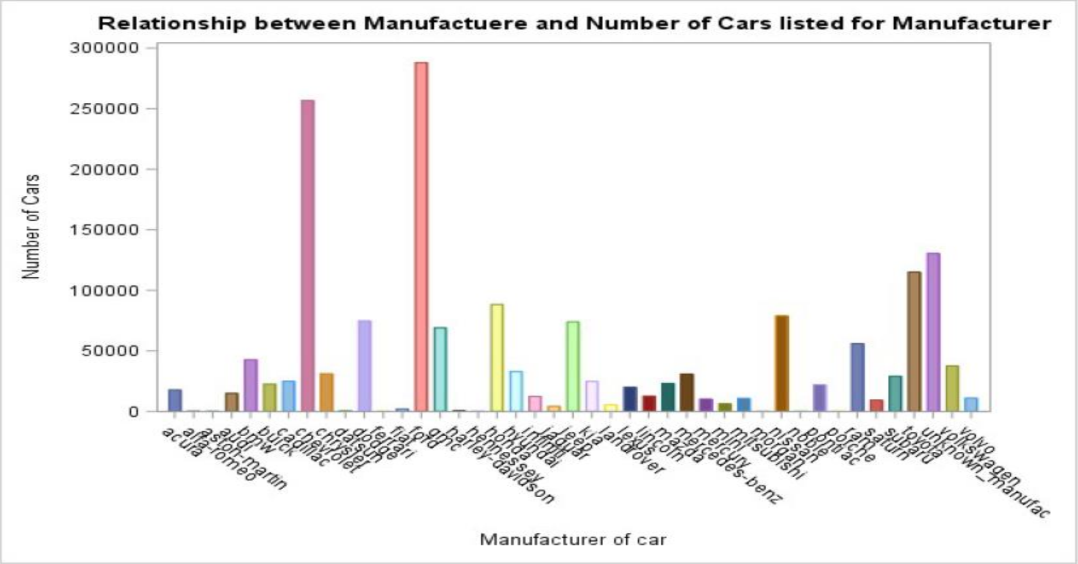
Though the average price of cars for these years are high, from the distribution plot we can see that the number of cars manufactured in these years to be low. Major bulk of the data are in the 2000s but the few cars in the 1900s are priced very high to make the average price stay high in those years, thus we determined that including the year is crucial to our predictive model.



Feature: Manufacturer

Based from the graph below, the manufacturers Aston-Martin, Ferrari, Hennessey, and Noble have the highest average price on Craigslist and we would expect cars from these manufacturers would cost more since they are high-end brands. Whereas Saturn and Mercury have the lowest average prices and we would expect that from these kinds of economy brands. Therefore, the manufacturer of the car seems to correlate with the quality of the car and the pricing and is significant to use as a predictor for pricing model.

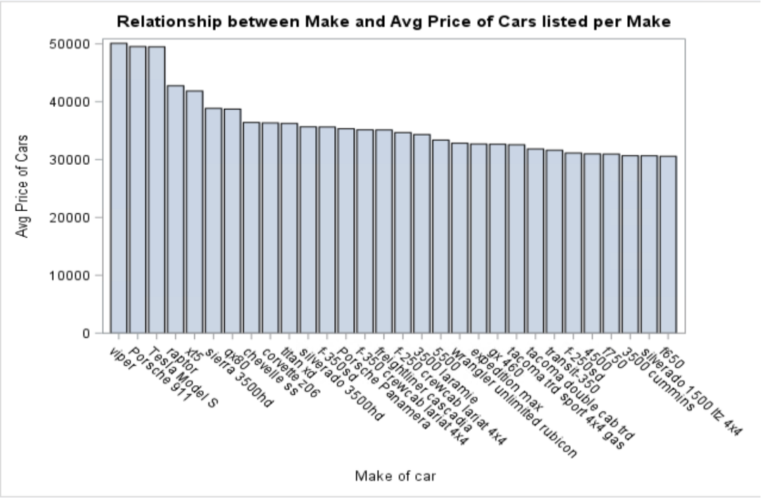


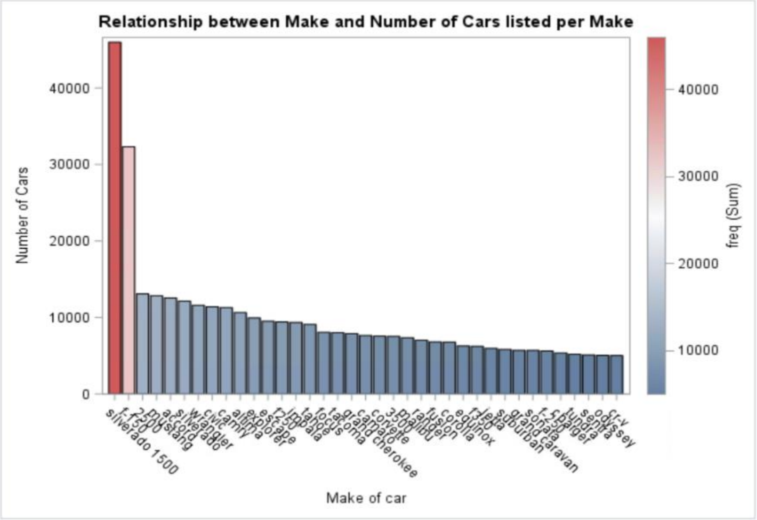


Feature: Make of Car (Model, Variant)

The make of the car is one of the important feature in the feature list which affects the price of the car. Each make of the car is priced differently based on the segment to which it belongs and the variants ranging from an entry level to a fully loaded version of the make. Since there is no specific validation or format of entry check in Craigslist, we have data redundancy and anomalies in this feature. Due to the huge number of different make of cars produced through the years, we will not be to check and correct every make in the list.

Having separate fields for model and variant fields in the data collection process, we will be able to analyze the data with better clarity and accuracy.





The sports cars like Viper, Porche 911 and electric car Tesla seem to have high average prices compared to other models.

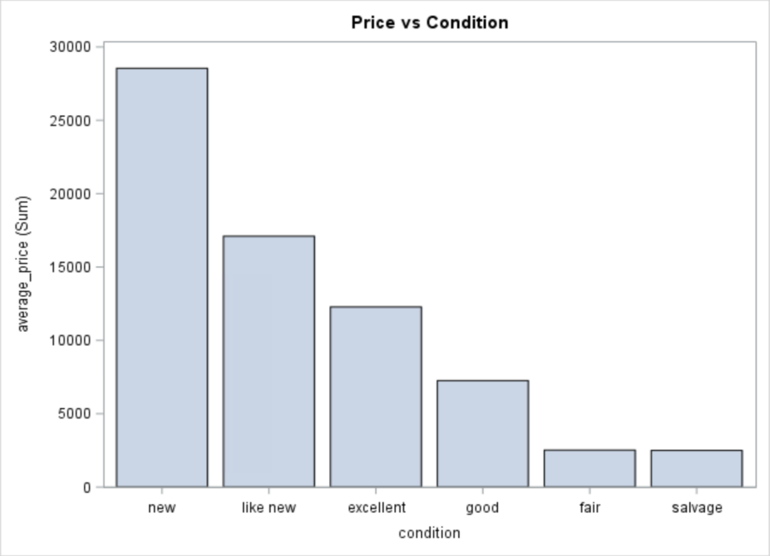
Pickup trucks like Silverado 1500, 2500 and f-150 have the highest numbers in the list.

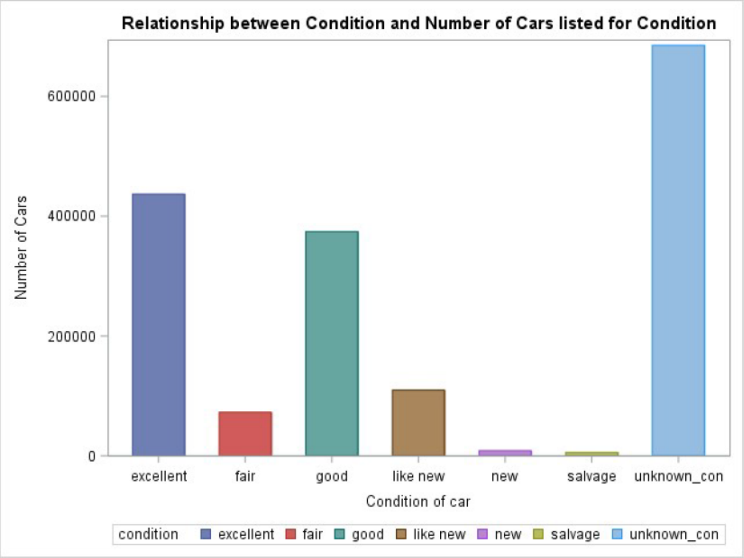
Do note that, due to improper make values in the list, the average prices for less frequent cars cannot be relied upon.

We did run a prediction model without the *make* feature to avoid the unreliable data which was significant and had about 63% accuracy. Although some of the features can be directly derived from the *make* feature, we do not want to remove this feature from our model.

Feature: Condition

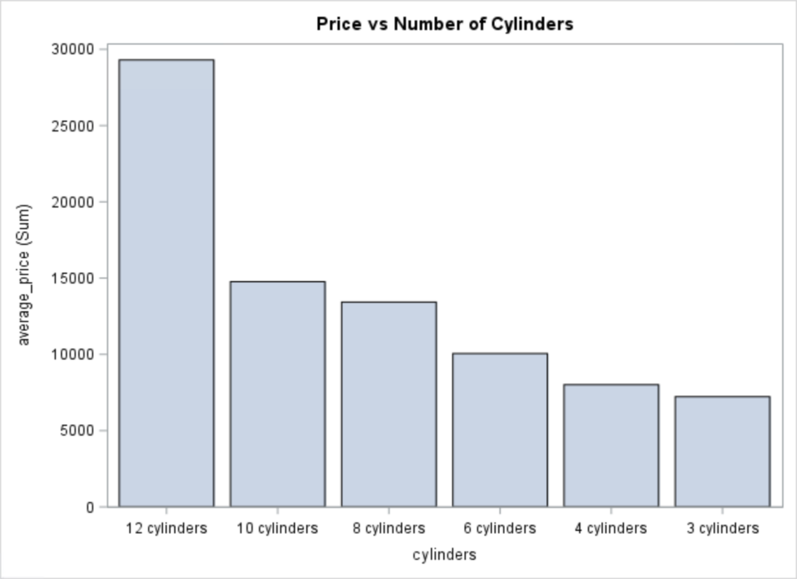
We expected to see a direct relationship between the condition of the vehicle and its corresponding price: the better the condition of the vehicle, the higher the selling price. So when evaluating a car with an identical make, model, and year but with a different condition status, we expected the car with the better condition of the two to have the higher selling price. From the graphs below, we can see that our hypothesis proves true and condition does have a direct relationship with price.

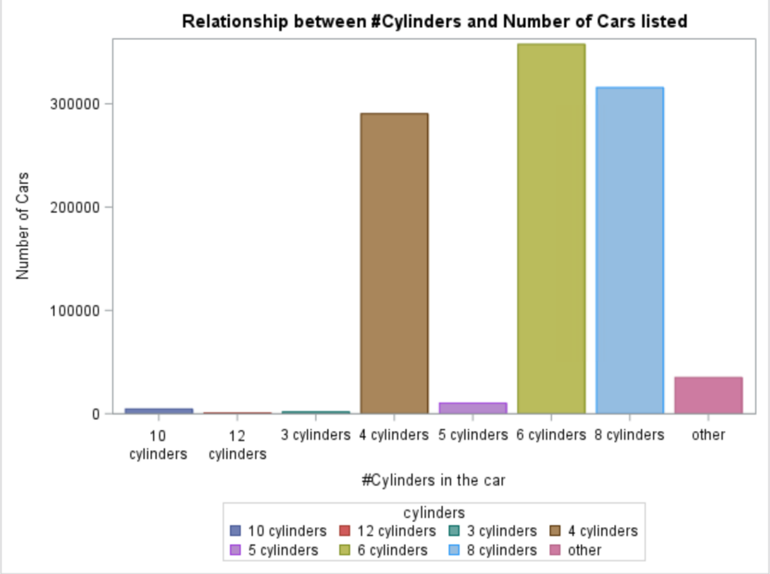




Feature: Number of Cylinders

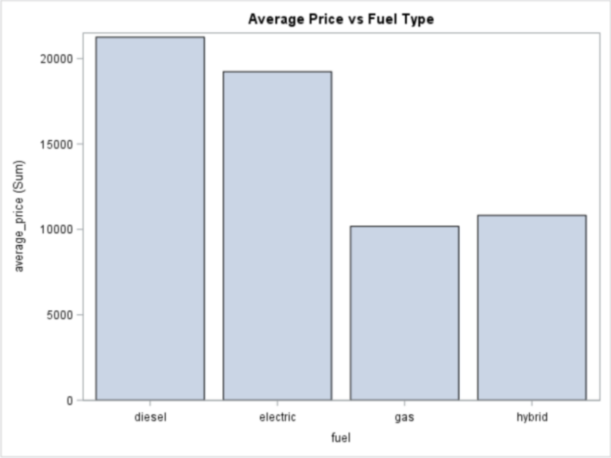
Looking into the feature of number of cylinders in a vehicle's engine, we expected that a higher count of cylinders in the vehicle's engine would result in a higher selling price. From the graph below, we can see this relationship between price and number of cylinders. We can also see that the average price of vehicles with 12 cylinders is significantly higher than the rest of the vehicles. This is most likely since cars with 12 cylinders are supercars that have extremely higher than average price tags.



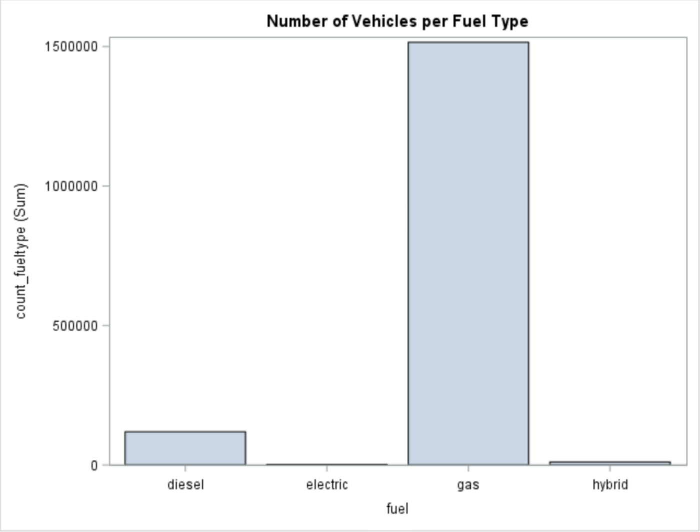


Feature: Fuel Type

For fuel type, we expected to see a relationship with the posted price of the vehicle where more common fuel types like gas were less expensive and less common fuel types like diesel were more expensive. From the graph below we can see that the price varies depending on the fuel type of the vehicle.

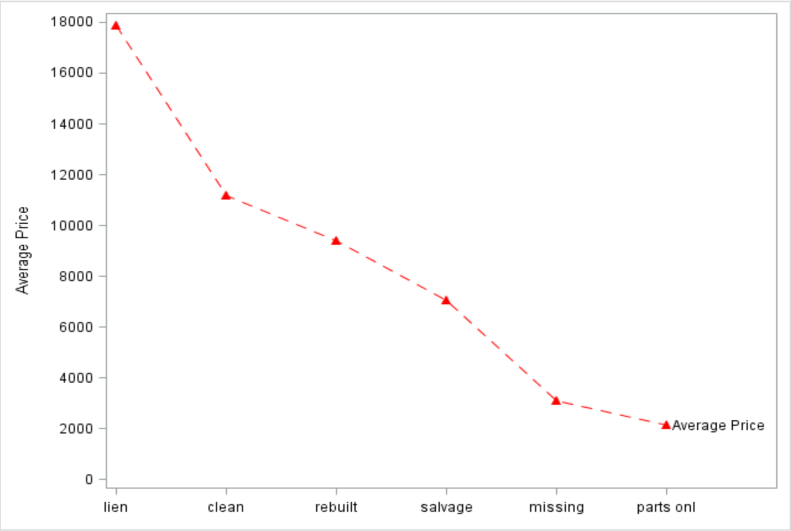


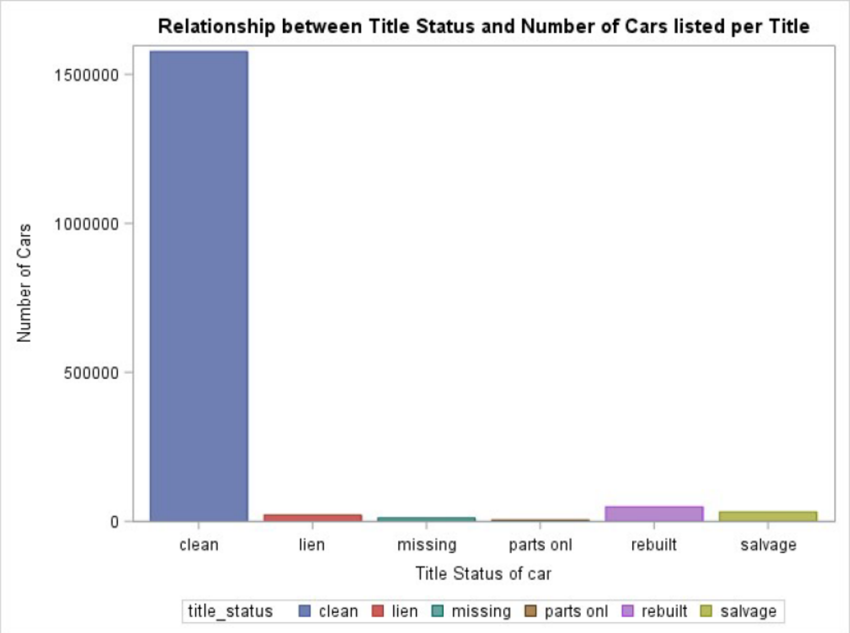
However, we must also consider the number of vehicles per fuel type. From the graph below we can see that most vehicles on CraigsList have a fuel type of gasoline. Therefore, we can expect a larger variance in price between vehicles of gasoline fuel type as opposed to those vehicles in the other three categories where we can expect a price closer to the mean we have calculated in the chart above.



Feature: Title Status

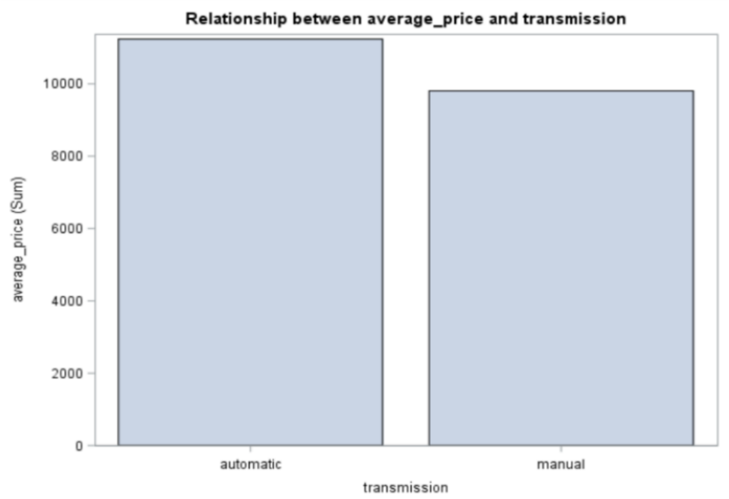
With the title status, we expected to see a relationship with the posted price of the vehicle. As per the graph below, vehicles with a title status of lien are highest priced, which we would expect as lien vehicles are being sold by a bank and tend to be deemed more reliable by buyers, thus worth more to a buyer. And on the other end, rebuilt, salvage, missing, and parts only titles we would expect to be worth less than a lien or clean title, so we will try to use this feature in our model.

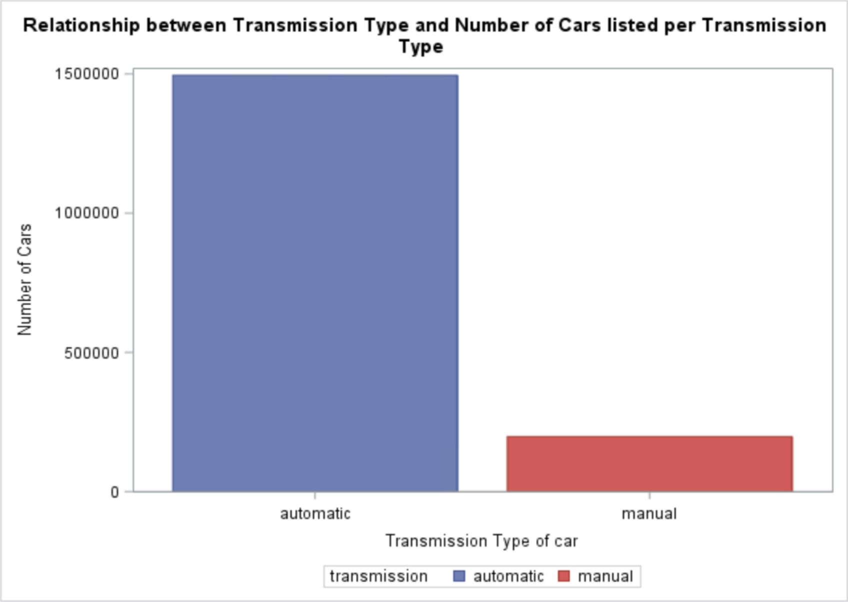




Feature: Transmission Type

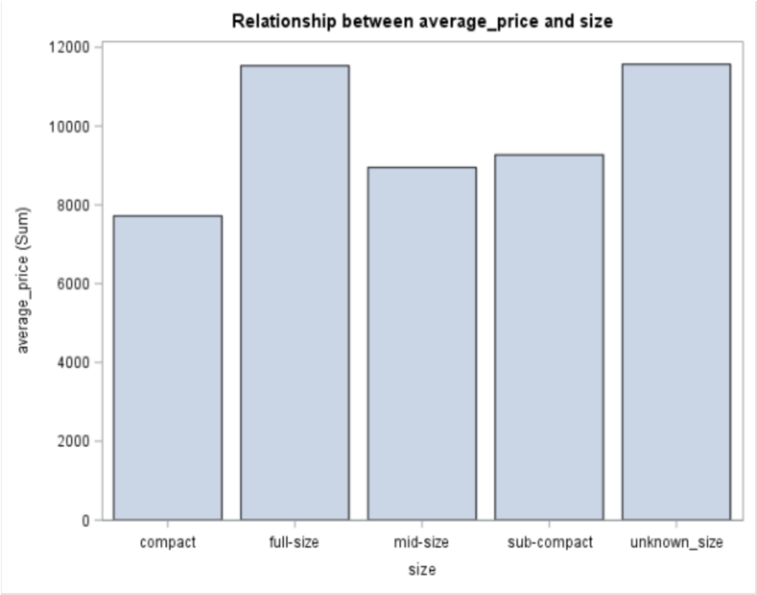
Regarding the transmission type feature, we expected to see manual cars to be slightly cheaper on average than automatic transmissions. However, we also discussed that sportier cars like a Corvette tend to be manual and are also expensive, so we hypothesized that the transmission type would have an effect on the price, but that there is an intertwined relationship between transmission type and the vehicle's class on the posted price of the car.

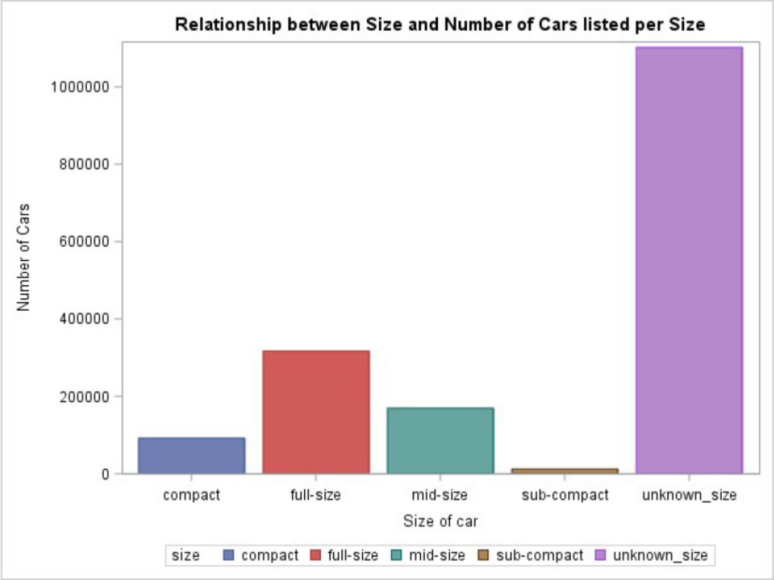




Feature: Size

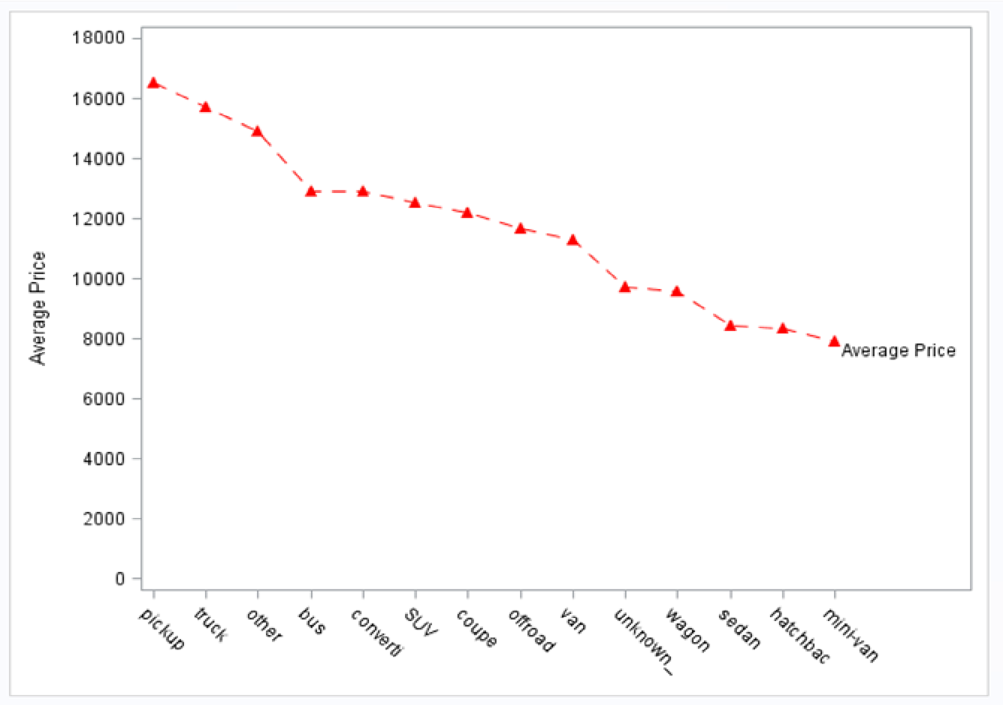
Our group discussed that due to the stability, comfort, power, and luggage space, we assumed that a full-size vehicle would be more expensive than a compact; and based from the graph below, our assumptions seem correct. Digging further into the data, there does seem to be a relationship between the size of the vehicle and its average price and this should feature should be added to our model to test for significance.

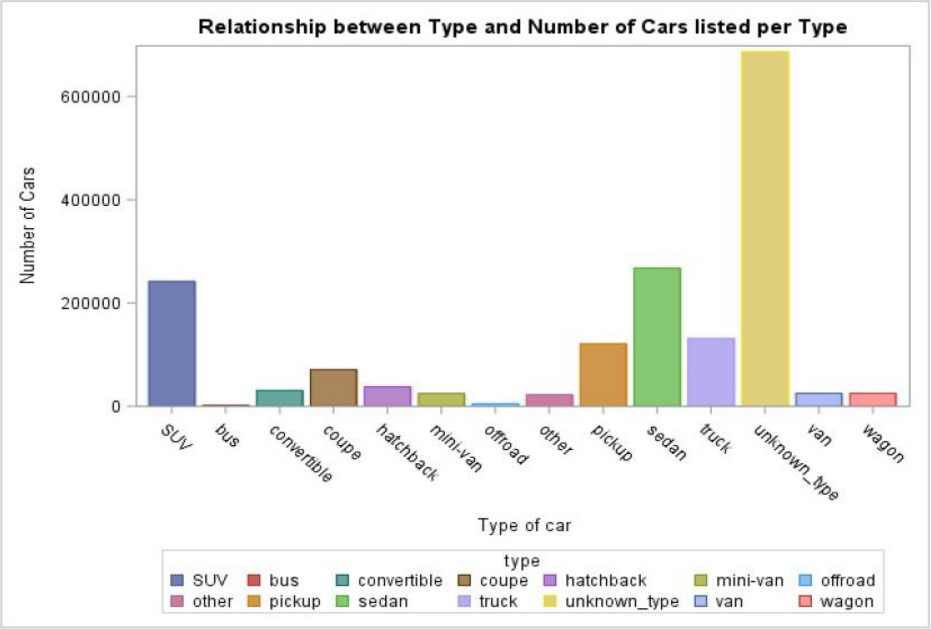




Feature: Vehicle Class (Type)

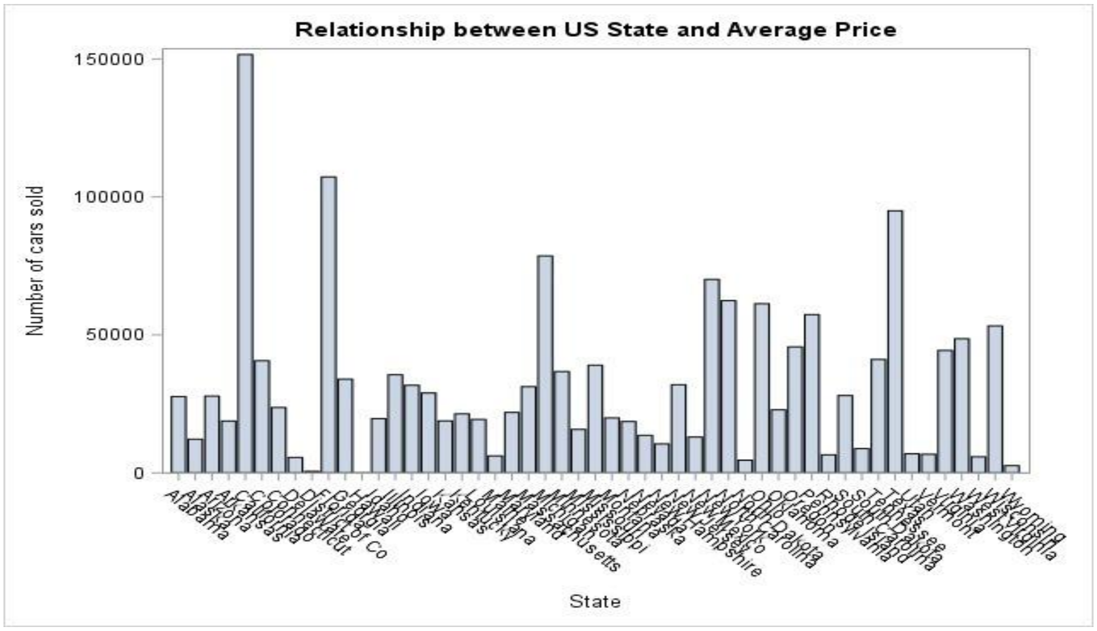
The "type" column of the data set referred to the vehicle's class, i.e. SUV or minivan, and so we hypothesized that big trucks like a Dodge Ram or a Ford F-150 tend to be pricier than smaller car types like a sedan. Based off the graph below, our hypothesis seems to hold true since there is variation between the average price of a car and the class of the car.

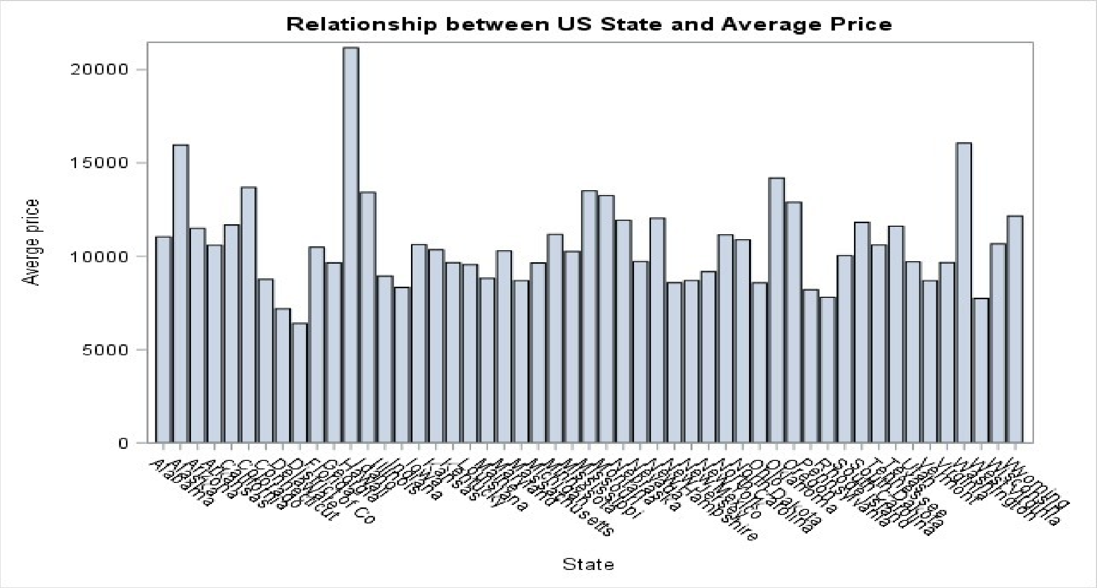




Feature: US State in Which the Car is Sold

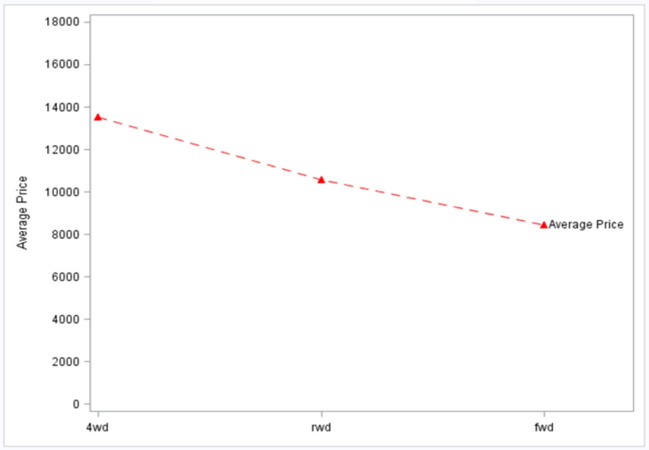
Our group discussed that where the car was being sold would have an effect on the price; i.e. if the car is being sold in Wyoming or Montana, it's probably going to sell for a higher price due to shipping costs to more remote regions and in states with higher populations per capita like Texas or California, cars would see for cheaper prices on average. As shown in the graphs below, there is indeed variation on the average price of vehicles sold in certain states and so we will put this variable in our model.

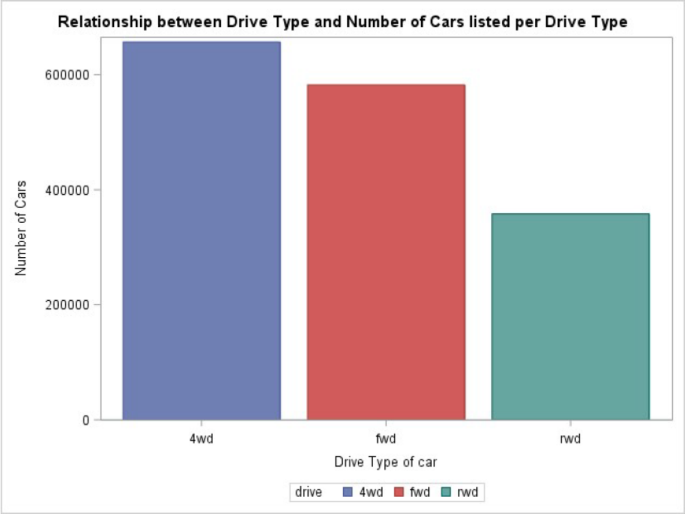




Feature: Drive Type

Our group discussed that 4-wheel-drivetrains would cost more than a vehicle with a rear-wheel-drivetrain or a front-wheel-drivetrain. A 4-wheel drive will usually be provided in full size vehicles like SUV or Trucks, which are usually priced higher than the lower sized cars. Based off the graph below, we hypothesized correctly, and this feature should be one that is tested in our model for significance.



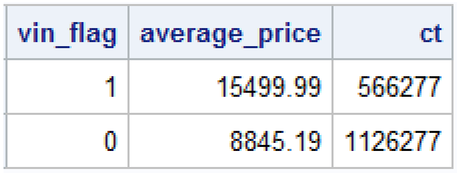


Feature: Whether or Not the Posting Had an Image

With this feature, we expected for the vehicle to be sold at a higher price if the posting on CraigList had a picture versus if it did not have a picture. However after digging into this feature, this column in the dataset had an entry for every single row, so it did not seem to be a valid indicator of whether or not the posting had an image and thus not a good feature to include in our model.

Feature: Whether or Not the Posting had a VIN Listed

One of the features we considered was if the posting had a VIN listed, did that increase the price or value of the car? The table below would seem to show that having a VIN listed in your CraigsList posting would increase the price. However, we decided that as a buyer, we wouldn't care if a VIN was posted or not and as an ad poster, we may not want to list the VIN with the concern of data theft. Additionally, the rows of the VIN column with values listed were not all valid, i.e. a lister would use the VIN field to list their contact email address or contact number instead of the VIN. Therefore, we hypothesized that this feature might not have a valuable benefit to our model.



vin_flag	average_price	ct
1	15499.99	566277
0	8845.19	1126277

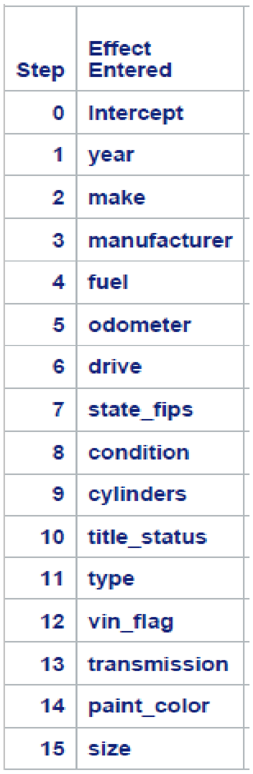
Empirical Analysis

As described above in the exploratory data analysis section, we can see clear linear relationships of price with several features, so we decided to use a generalized linear model to attempt to predict the sale price of a car based on its features.

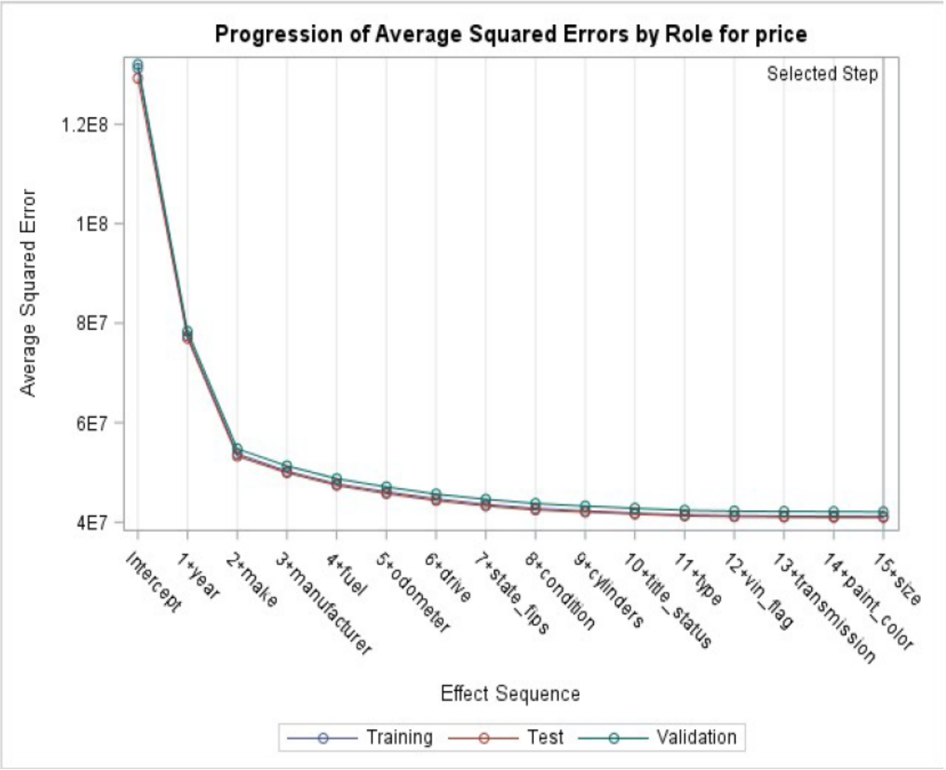
Since our model will contain both categorical (i.e. paint color, manufacturer, etc.) and continuous (i.e. odometer) independent variables as predictor features, the prediction algorithm will be based on Analysis of Covariance (ANCOVA). ANCOVA is a statistics tool used to test the mean and interaction effects of categorical variables on a continuous dependent variable, controlling for the effects of selected other continuous variables, which co-vary with the dependent variable; the control variables (categories) in an ANCOVA are called the "covariates."[6] In layman's terms, ANCOVA is a smart enough tool to be able to predict something like the price of a car from both text and numeric type details about a car.

List of features that we considered: odometer, vin_flag, make, year, manufacturer, condition, cylinders, fuel, drive, type, state_fips, title_status, transmission, size, paint_color, and state_fips.

The features cylinders, size and condition have a specific order in which each value has a different level corresponding to another. Considering and implementing the information provided by such orders over the price values will help better our model.When we ran a prediction model on all the features selected, we get an accuracy of about 68.5% in the prediction of price and from the results of the model, we can see that all the selected features are playing a significant role towards predicting the price of the car. The below table gives us the list of features in descending order of their magnitude of effect in accurately predicting the price of a vehicle. From the graph we can also see the amount of reduction in the prediction error per feature and after the first 6 features, the reduction is minimal.

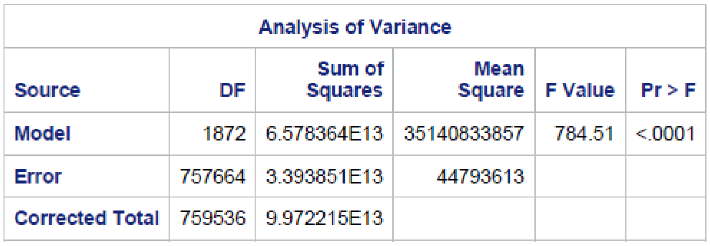


Step	Effect Entered
0	Intercept
1	year
2	make
3	manufacturer
4	fuel
5	odometer
6	drive
7	state_fips
8	condition
9	cylinders
10	title_status
11	type
12	vin_flag
13	transmission
14	paint_color
15	size

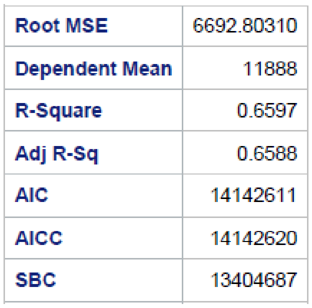


Additionally, a user or dealer who is browsing for a quick online estimate of the price at which they should list their car for might not be to keen on having to enter 15 separate, tedious metrics or the user might not know or have access to the values at that point in time.

Due to these reasons, we want to build a parsimonious and efficient model for our price prediction, so we tried to reduce the number of features by looking at their magnitude of effect in predicting the price. After removing all but the first 6 features (year, make, manufacturer, drive, odometer reading and fuel), our new model gives us an accuracy of about 66% which seems reasonable considering the significant gain in performance.



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1872	6.578364E13	35140833857	784.51	<.0001
Error	757664	3.393851E13	44793613		
Corrected Total	759536	9.972215E13			



Root MSE	6692.80310
Dependent Mean	11888
R-Square	0.6597
Adj R-Sq	0.6588
AIC	14142611
AICC	14142620
SBC	13404687

SAS Code

UTD BOX Link:

https://utdallas.box.com/s/m5ympt39ycbwkqq0a82repbh8d15sudl

Conclusions

Through the process of data analysis and reasoning, we can confidently confirm that we have generated a model which can be used to estimate the resale value of a car by having only 6 feature values entered by the user. From our model's estimates, while having any and all non-referenced features as constant, we can note the below:

- Having a 4-wheel-drivetrain will increase the average price of the car's resale price by $1,525, whereas having a front-wheel-drivetrain will decrease it by $3,061 compared to rear wheel drive cars.
- For every 100 additional miles driven by the car, the price value decreases by $2.77.
- A diesel-fueled vehicle has a higher resale price by about $8,591, a hybrid reduces the price by $279, and an electric car fetches about $2,850 additional as compared to a gasoline variant of the same car.
- A car from manufacturers like Ferrari, Aston-Martin, Jaguar, Porsche, and Lexus will yield higher resale values compared to a car with same features from Hyundai.
- Apart from the cars from most recent years, cars from 1930's with same features yield higher resale values compared to same car from year 2000.

From our experience with this dataset, we believe that with significant additional processing power, all the values in the make feature could be used instead of our limit to 100 occurrences or more to build a higher performing model which could provide better accuracy without losing accessibility to users wanting a quick quote. Additionally, with more time, we would want to use natural language processing practices to better classify the make and manufacturer features.

References

1. https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/

2. https://www.kaggle.com/austinreese/craigslist-carstrucks-data#craigslistVehicles.csv

3. https://www.autotrader.com/car-news/these-are-the-7-highest-mileage-cars-listed-on-autotrader-256616

4. https://www.clunkers4charity.org/facts-about/11-oldest-cars-world/

5. https://www.eia.gov/energyexplained/?page=us_energy_transportation

6. https://www.lehigh.edu/~wh02/ancova.html

7. https://documentation.sas.com

8. https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4