Final Project: Craigslist Used Cars

Group 7:
Hayden Reid
Sumit Paul
Lidya Lulseged
Annah Lee
Sruthi Theddu

Table of Contents

Executive Summary

Overview

Many people are turning to Craigslist instead of traditional car dealerships to sell their used cars. As a result, prices of cars on Craigslist are determined differently than prices of cars at traditional dealerships. In this report, we analyze how different variables affect price, specifically how the manufacturer of the car affects the price, and which manufacturers vehicles might hold value better in the industry. For this analysis, we used the "Used Cars Dataset" from Kaggle, which is a cross sectional dataset that is web-scraped from Craigslist. The dataset consists of over one million observations and 26 descriptive variables. After preprocessing the data, we used multiple analytical techniques including ordinary least squares and regression trees to model our data, as well as creating a predictive model for our analysis. From our analysis, we concluded that the mileage on the odometer, condition, and the type of the car are the three most important variables that predict the price of the car. Using our analysis, we can recommend types of cars and manufacturers depending on the consumer's needs.

Key Take-Aways

If the consumer's main requirement is to purchase a vehicle that has been utilized the least, they should stay away from Hyundai, Kia, and Dodge Rams. Instead they should look into purchasing a convertible or coupe since those vehicle types had the lowest average mileage. GMC, Ford, Chevy, Acura, Honda, Toyota, Lexus, Volvo, or Infiniti are the manufacturers that have the least average mileage, consumers should pay the most attention to them if their main requirement is to purchase a vehicle that will last them a long time. If looking for high value retention, then consumers should pay close attention to 6-cylinder vehicles manufactured in the Asian region. If looking for the best overall value, the buyer should focus on Asian buses and trucks, American coupes, and European SUVs. Non-luxury and non-performance-oriented vehicles that have the best value are Buick, Chevrolet, GMC, Ram, and Volkswagen.

The biggest factors in predicting the price of a vehicle was found to be the odometer reading and the condition of the vehicle, while the size of the vehicle was the least important. Our linear regression model was able to accurately describe ~75% of the variation in price, and we can confidently say that our model accurately describes the relationship between the variables chosen in the model and the price of vehicles listed on Craigslist.

From a business perspective, we have some recommendations as well as applications for our model for both the buyers and sellers in the used car market on Craigslist.
Sellers should provide the most information about their vehicle as possible. A buyer could use our predictive model as a guide to decide whether the price of a listed vehicle is reasonable or not. If the listed price is widely different from the predicted price, the buyer could use this information to negotiate on the price.

As for Craigslist, they could provide a more intuitive and uniform template about vehicle information. Craigslist could incentivize the seller to include more information about their vehicle by promoting their listing in the advertisement list as to be seen by more potential buyers. They could also implement a predictive model similar to ours and present it in their listings to provide a better understanding to potential buyers about the listing they are looking at in an effort to remove the ambiguity associated with the car buying process.

Introduction

Founded in 1995 in the San Francisco Bay area, Craigslist is known for being one of the largest consumer to consumer marketplaces, taking advantage of people's desire to get rid of unwanted items or realize value from rare or niche items. According to Forbes, Craigslist is conservatively valued at $3 billion, indicating its nationwide rapport as an advertisement platform. Consequently, it has become a top choice as a platform for the exchange of vehicles in the C2C space. Often times, buying a vehicle from a car dealership comes with much uncertainty and strings attached, as well as additional costs. On Craigslist, merchants and consumers alike are free to haggle for better prices and make sure they get the most for their dollar.

As a result of Craigslist's growing popularity in the third-party vehicle space, it has also become an interesting platform for acquiring data on new and used vehicles. Our team utilizes a dataset which contains vehicle listings from Craigslist across the United States along with vehicle information such as color, type, manufacturer, odometer value, etc. The dataset presents powerful potential in terms of analytics, allowing us to predict the price of vehicles based on the other information collected on them, understand which features are important in pricing and value, and derive insights as to which manufactures' vehicles retain their value over time.

Data Description

The data was collected via a web scraper every few months and was updated with the most current vehicle listings across the United States. The most recent version of the Craigslist vehicles dataset includes 1.7 million observations and 26 descriptive variables. Vehicles in the dataset are not limited to just personal vehicles, as it also contains commercial vehicles (work trucks, buses, delivery vans, etc.) and motorcycles as well. Below is a description of the variables included in the dataset:

Url - Link to listing
City - Craigslist region
Price - Price of vehicle
Year - Year of manufacturing
Manufacturer - Manufacturer of vehicle
Make - Model of vehicle
Condition - Vehicle condition
Cylinders - Number of cylinders
Fuel - Type of fuel required
Odometer - Miles traveled
Title_status - Title status (e.g. clean, missing, etc.)
Transmission - Type of transmission
Vin - Vehicle Identification Number
Drive - Drive of vehicle
Size - Size of vehicle
Type - Type of vehicle
Paint_color - Color of vehicle
Image_url - Link to image
Lat - Latitude of listing
Long - Longitude of listing
County_fips - Federal Information Processing Standards code
County_name - County of listing
State_fips - Federal Information Processing Standards code
State_code - 2 letter state code
State_name - State name
Weather - Historical average temperature for location in October/November

Objective

For our analysis, we plan to examine the differences between different car manufacturers and their inherent value in the third-party resale space. More specifically, we plan to determine which manufacturer's vehicles hold value the longest as compared to other manufacturers. Using our analysis, we plan to make recommendations for different prospective vehicle buyers looking for certain features in a vehicle. Our hypothesis is that American made vehicles do not hold their value quite as well as their foreign counterparts. We also plan to determine if any manufacturers have any "top performers" in terms of longevity and/or resale value, and if these vehicles are outliers to that manufacturer (meaning they last much longer and hold their value), or if they are aligned with the manufacturer's overall performance.

Additionally, we plan to examine the influences that the utilization of the vehicle has on its value. In this case we are considering two variables, the condition of the vehicle at the time of the sale and the average miles driven per year, and their relationship to the price. To create robust models, we need to first understand the relationships among our variables in the dataset. This will indicate which variables are influential to each other and can help us further determine which variables explain the most variation within our dataset. Next, we can use these relationships to reduce the number of unnecessary variables in the models that we construct and increase predictive accuracy. Using our predictive models, we can predict the prices of vehicles with certain attributes.

Preprocessing Data

Indexing/Variable Removal

In this stage we plan to frame the data around our analysis goals and make the dataset run as efficiently as possible in SAS. Our first step was to index the dataset to create a unique identifier for each observation in our data. This way, despite null values being present in almost every attribute, we can accurately say there is some data present there. Secondly, we need to determine which variables have no inherent value to our analysis. The first variables we determined to have no value were the "URL" and the "image URL" variables. While there may be some textual information in the hyperlink to the listing, we've decided to remove these variables as they provide no statistical power and should be removed. Next, we determined the redundancy between the "state_name" and "state_code" variables were unnecessary, so we decided to keep only the "state_name" variable in the dataset. The "fips" and "state_fips" variables represent the federal and state information processing codes, and we also decided to remove these variables from the data as they provide no meaningful information for our analysis. The VIN numbers for the vehicles were also included in the dataset, however they are arbitrary and have no meaningful interpretation, so we removed that variable as well. Finally, we decided to remove the weather variable as well. This variable, while perhaps useful, may not be a good indicator of the condition of the car. In this case, we would rather use the "condition" variable as an indicator to the vehicle's physical quality.

Data Sub-setting/Null Values

We should now consider values at the extreme ranges of our variables. Most notably the year variable has a minimum value of 302 which is clearly an inaccurate observation. We decided to subset the data to only represent vehicles with a year of 1940 or above, as anything below that might be a niche or collectors' vehicle rather than a consumer vehicle. This removed just over 3% of our original 1.04 million observations. Now that the unnecessary variables are removed and the year variable is in a reasonable range, we must decide how to handle null values in our dataset. First, we can consider the observations that provide no details on the vehicle at all. These observations would have no values for "manufacturer", "make", or "type". These observations would not be considered important because they only provide us with a price and a year, with no additional understanding of how that price was determined or the quality of that manufacturer's vehicles. In this case, these rows can be removed. Now our dataset consists 1.01 million observations (-3.5% of our original dataset). The remaining observations that contain null values still contain meaningful data, as they have a price and either "type", "manufacturer", and/or "make", which can still provide value for our analysis and therefore should not be deleted.

Variable Manipulation

Looking at our data there were many observations which listed the make of the vehicle but did not list the manufacturer. However, our dataset contained so many observations that duplicates were plentiful, and in many cases these duplicate observations did contain the information on the manufacturer. Using an index of all unique manufacturer/make combinations, we were able to populate many of the missing manufacturer values in our dataset. This process enriched our dataset with more information about the manufacturers by providing us with more observations per manufacturer. Additionally, we wanted to modify our manufacturers to be consistent when representing the same value. For example, the dataset had instances of manufacturer being listed as "chevy" and "chevrolet", which represent the same manufacturer. In this case we changed the manufacturer to be consistent for all Chevrolet models. We then further divided manufacturer to different regions of the world by grouping. For example, we group "ford" and "chevrolet" to be of region "american", in an effort to see the differences between cars manufactured around the world. As a convenience, we also wanted to add an "age" variable to our dataset by taking 2019 and subtracting the year the vehicle was made. Using this variable, we can determine how age affects the price of a vehicle as well as the longevity of vehicles belonging to certain manufacturers. We also wanted to calculate some additional metrics for our observations. We chose to calculate a price per mile as price per mile using the price of the vehicle divided by the odometer value, as well as odometer/age as mile_per_year by dividing the odometer value by the age of the vehicle. These metrics allow us to compare any vehicles in the dataset to each other, making our analysis more consistent.

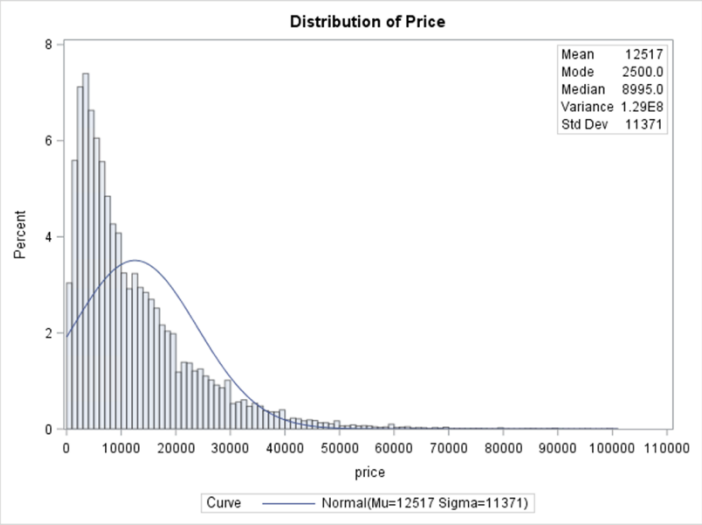
Exploratory Data Analysis

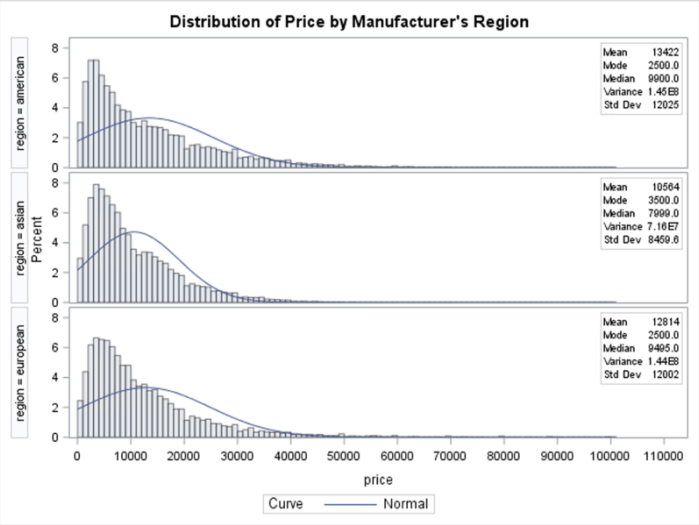
Price

To determine the distribution of price across our entire dataset, we were able to run a histogram fitted to a normal distribution. When analyzing the results, it was clear that there were outliers that severely skewed the dataset. To combat this, we set our endpoints to $0 - $100,000 (where the bulk of our data was present).

Once these parameters were defined, we were able to estimate the mean of the overall prices to be $12,517 +/- 11,371. This high standard deviation, and in turn high variance, could be pointing to inconsistencies in the dataset that were not combated by specifying our endpoints. The raw data shows us that these inconsistencies were due to irregularities in inputting of data. For example, some sellers listed their price as $0 possibly indicating that they are willing to negotiate.

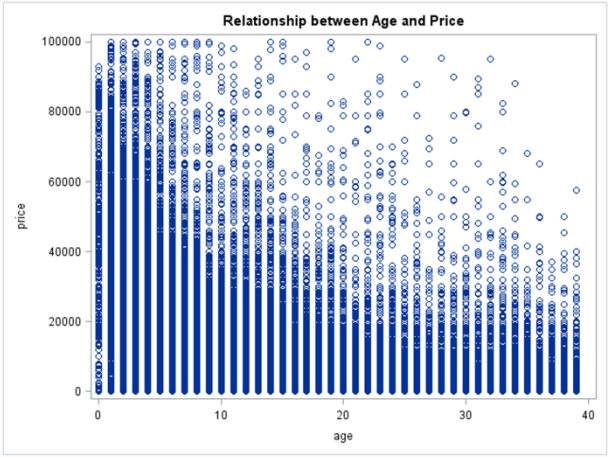
In an effort to further dissect the price variance among the dataset we constructed a stacked histogram that compared the prices of different manufacturers as dictated by their respective regions. These distinctions were broken up by American, European, and Asian vehicle manufacturers. These comparative histograms showed us that American manufacturers had an estimated mean of $13,422 +/- 12,025, Asian manufacturers had an estimated mean of $10,564 +/- 8,459.60, and European manufacturers had an estimated mean of $12,814 +/- 12,002. The results of the histograms indicate that all regions have a high standard deviation along with variance, but Asian manufacturers has the lowest price and deviation.





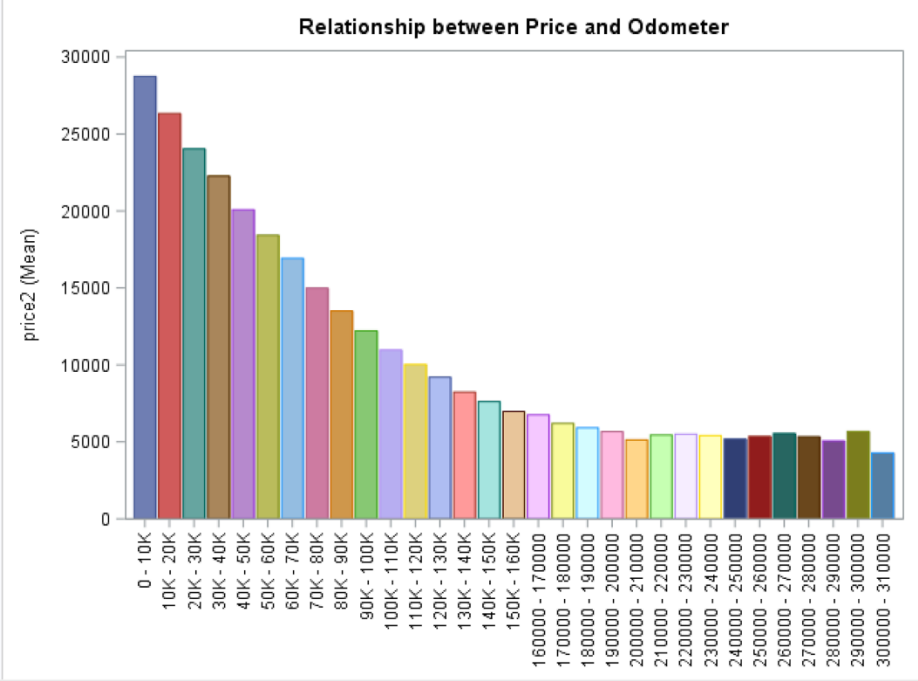
Price and Age

The age variable was calculated by subtracting the year of the car by the current year. In order to compare the monetary value of the vehicles in our entire dataset as it relates to its age, we constructed the scatterplot below that visualized this relationship. The plot shows that there is an inverse relationship, as the age of the vehicle increases the value (price) of the vehicle decreases.



Price and Mileage

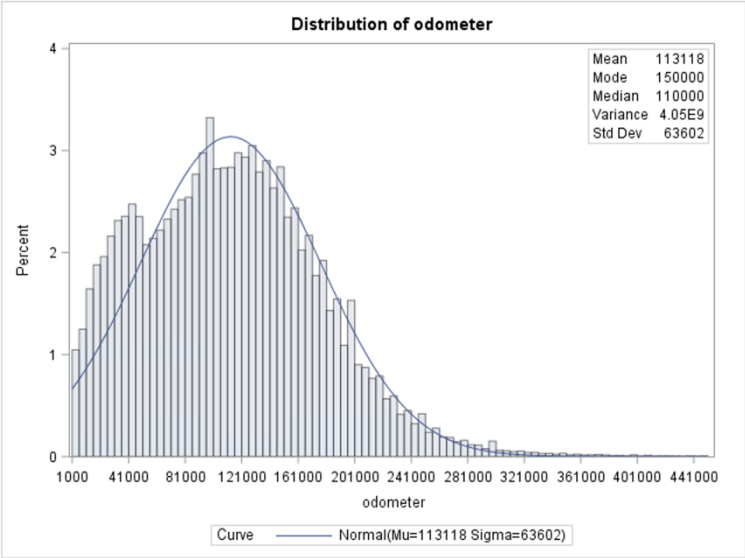
In order to determine a clear relationship between the price against the number of miles on the vehicle, we set the parameters for the price to be between $0-$100,000 and the odometer to be between 1000-300,000 miles. With these endpoints in place, we constructed buckets to group the miles in 10,000-mile increments. The graph below shows the resulting relationship, which indicates that, in this dataset, as the average mileage on the car increases the value (price) decrease.

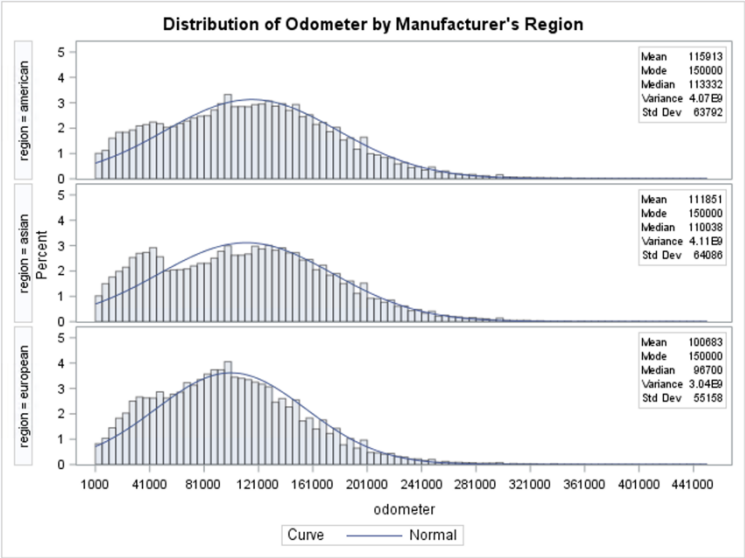


Mileage

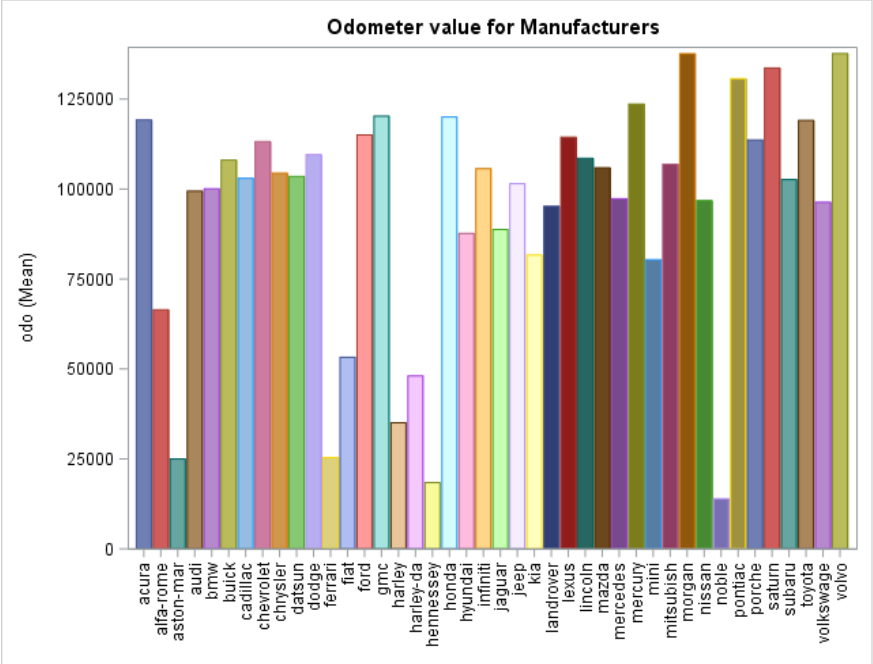
In order to further analyze the distribution of the odometer readings throughout the dataset we constructed another histogram that was fitted to normal distribution. Like the price, there were outliers that heavily skewed our results, and in an effort to elevate this we set our end-points to be between 1,000-450,000 miles. With these new parameters it was determined that the estimated mean of mileage of the entire dataset is 113,118miles +/- 63,602.

Staying consistent with the dissection of the price, we compared the odometer readings by the region the vehicles' manufacturers were from. The American manufactured vehicles had an estimated odometer reading mean of 115,913miles +/- 63,792, the Asian manufactures estimated mean was 111,851miles +/- 64,086, and the European manufacturer was 100,683miles +/- 55,158. This comparison showed us that the European manufactured vehicles had a lower average mileage with less deviation as compared to the other manufacturer regions.





In order to take a deeper look into which specific manufacturers had the highest mileage, we plotted the average number of miles on the odometer for the individual manufactures. As seen below, Volvo, Morgan, Chevrolet, Saturn, and Acura are among the manufacturers that had the highest mileage, on average, in this dataset.

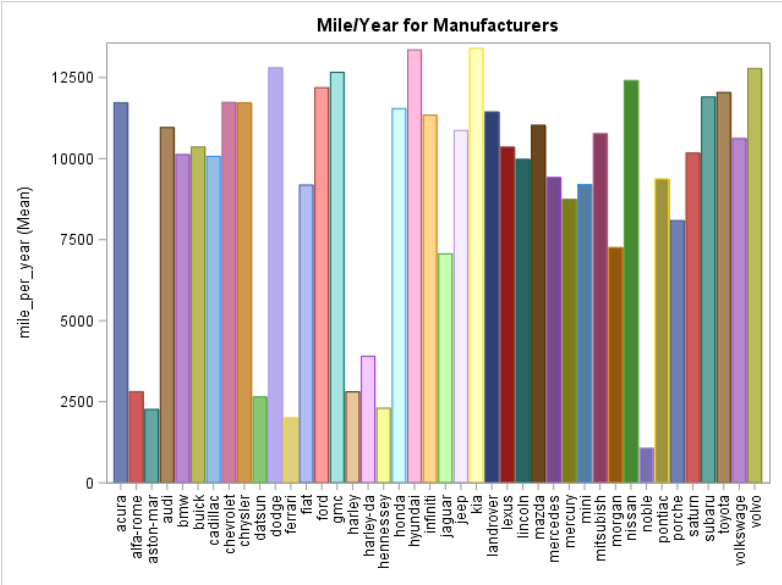


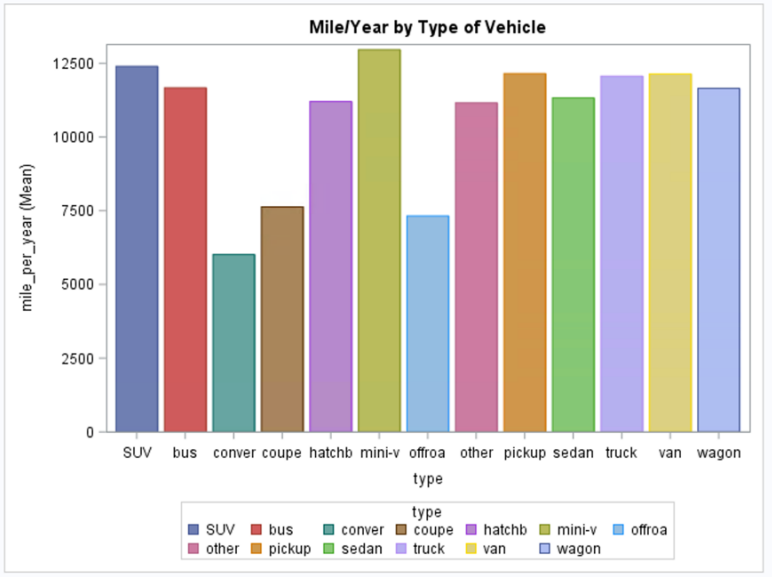
Mile/Year

To determine the wear and tear of the vehicles in the dataset we established the variable, mile per year. Mile per year was calculated by dividing the miles on the vehicle by the age of the vehicle. After establishing this variable, we plotted it by the manufactures present in the dataset. This lets the customer

know which manufactures have vehicles with the highest average number of miles driven per year. As seen below, Dodge, Kia, Hyundai, Volvo, GMC, and Nissan are amongst the manufacturers that had the highest average miles driven per year. This indicates that these manufacturers have the vehicles with the most wear and tear in this dataset.

In the graph below, we've also plotted mile per year against the type of vehicle to determine which types of vehicles had the most wear and tear throughout the years. From this plot we can see that, vehicles typed as "mini-van", "SUV", "pickup"/" truck", and "van" have the highest wear and tear in this dataset.

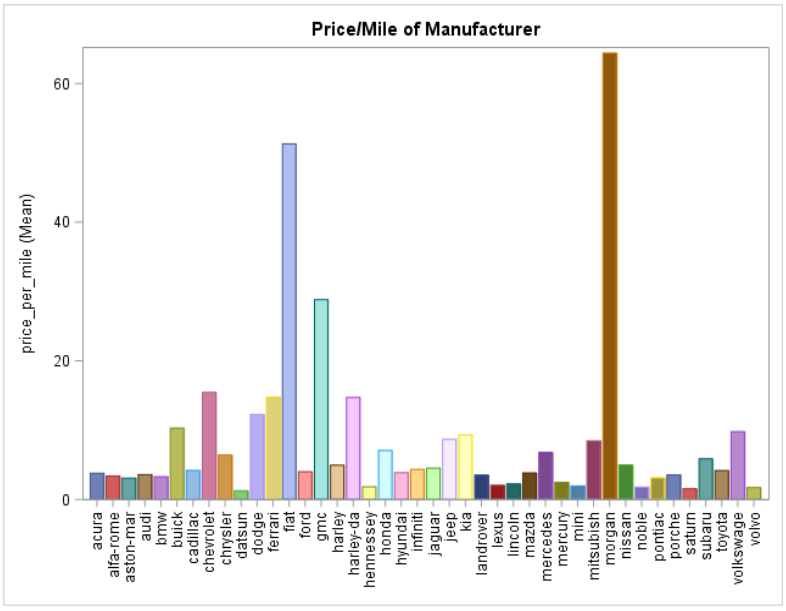


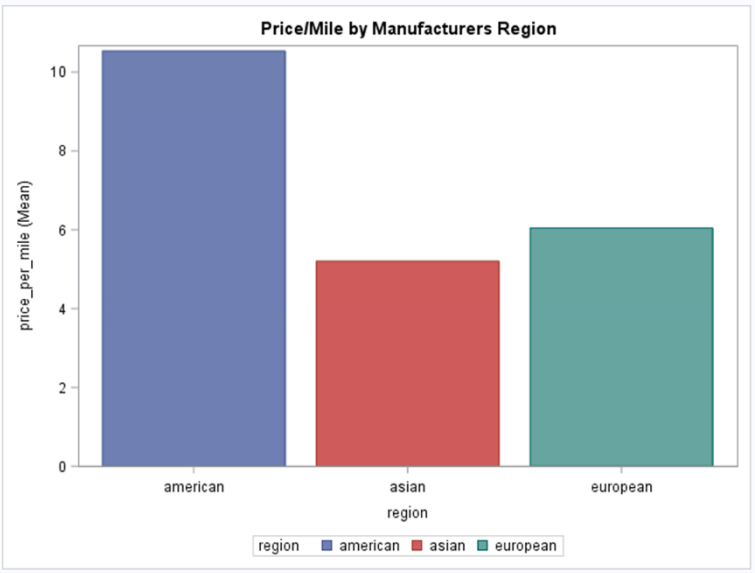


Price/Mile

Price per mile is a variable which was created by taking the price of the vehicle and dividing it by the odometer value the vehicle had in the listing. By plotting the price per mile variable against the manufacturer, we can see that certain vehicles retain value better given the number of miles that were driven, indicated by the height of the bars above each manufacturer. We can see that Fiat, GMC, Morgan, and Chevrolet vehicles really stand out with respect to price per mile within this dataset. Morgan is a British car manufacturer which builds high-end performance-oriented vehicles by hand. They are listed at very high prices and are wanted more as collectible vehicles rather than as daily drivers, so they generally will have very low miles and very high prices which will cause their average price per mile value to be extremely high as compared to other manufacturers within the dataset(*Morgan Motor)*. For the rest of the manufacturers, this plot indicates that their vehicles retain value better given their price and mileage.

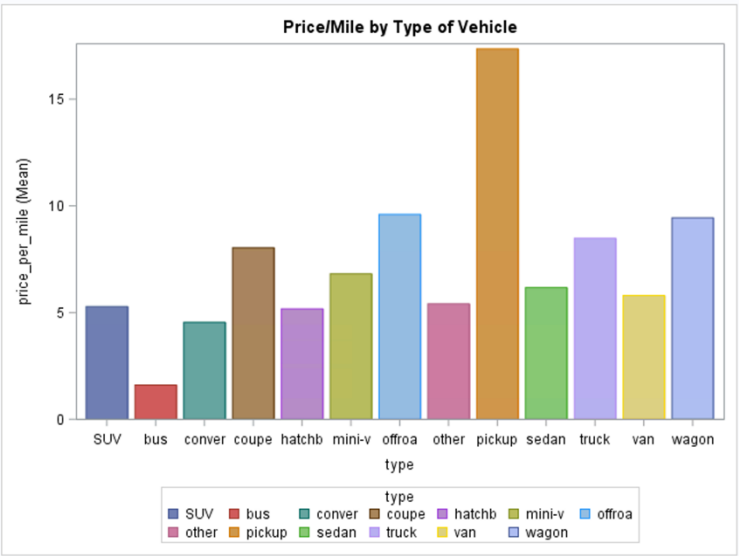
Taking it a step further, in the graph below we can see that the average price per mile for vehicles from the American region is higher than the average price per mile for vehicles from other regions. This can be due to the fact that the frequency of American vehicles which consists of many muscle cars and other highly valued cars in the dataset out number vehicles from other regions, and is skewing the overall average price per mile for the vehicles from the American region.

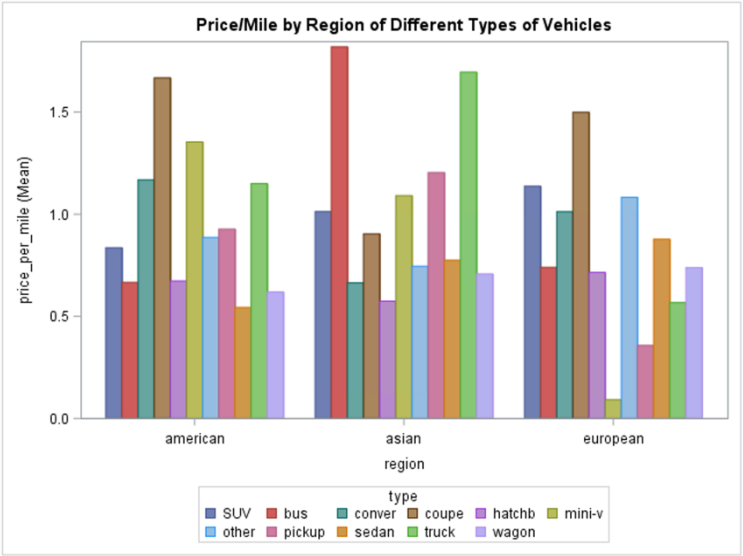




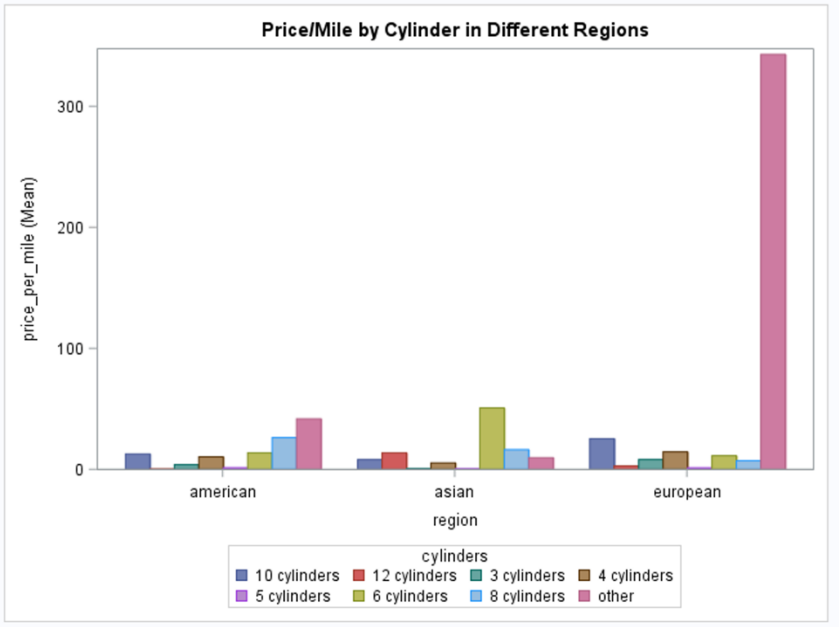
In the graph below, we have plotted the price per mile against the type of the vehicle to see which types of vehicles retain their value the most across the mileage and price values. From the plot, we can see that the vehicle type "pickup" has a high price per mile as compared to other vehicle types within the dataset. This indicates that pickup trucks are probably more highly valued according to this dataset, most likely due to their durability and capabilities.

Furthermore, the adjacent graph shows the average price per mile for each region and the type of vehicles within that region. We can see that the coupes from the American region have a very high price per mile compared to coupes from other regions. We can also see that Asian buses and trucks have a higher average price per mile as compared to other regions. European SUVs have a higher average price per mile as compared to SUVs from other regions. This indicates that a person shopping for a car on craigslist can potentially get better value for their vehicle purchase by either buying an American coupe, Asian buses and trucks, or European SUVs according to the type of vehicle they want to purchase.





In the chart below, the average price per mile of vehicles in different regions and different number of engine cylinders. We can see that the Asian region has the highest average price per mile for 6-cylinder engines, which indicates that 6 cylinder engine cars from Asia retain a higher price given their mileage compared to other 6 cylinder cars from the other regions. We can also see that American made vehicles that have 8 cylinders have a higher price per mile average as compared to other 8-cylinder cars from other regions. This could be because a majority of muscle cars have 8-cylinder engines and they are also American made and since they are performance-oriented vehicles, they also retain their value better because of their build quality and demand.

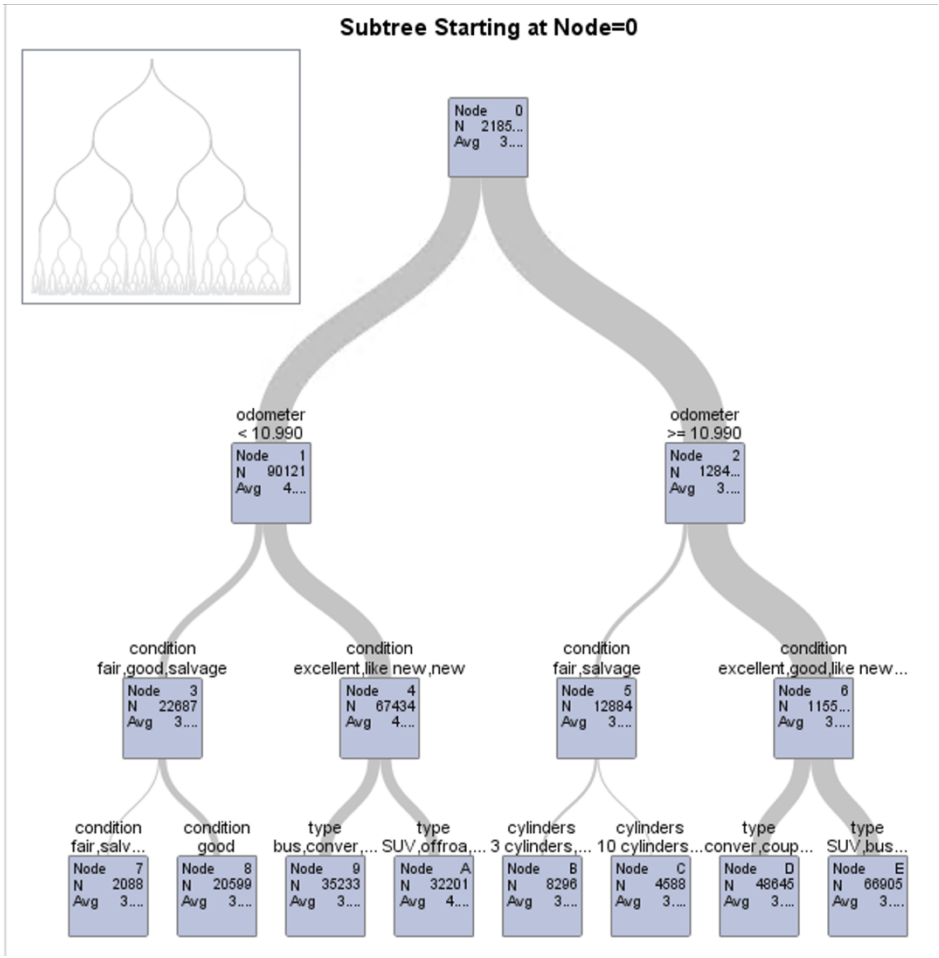


Empirical Analysis

Regression Tree

In order to determine which variables are most valuable in the dataset, we constructed a regression tree shown below. As seen below, we have 218,555 observations in the top node. The first split is whether odometer is less than 10.990 or more than/equal to 10.990. Among the cars with odometer less than 10.990 the next best split is whether condition is fair good and salvageable, or condition is excellent, like new, and new. The next best split is condition for the left node and type of car for the right node. For the cars with odometer greater than 10.990, the next best split is between condition. The left node's best split is the number cylinders of the car. The right node's best split is the types of car. The full regression tree in the top left corner shows the general shape of the tree with all its decision nodes and branches. The regression tree helps show the ranking of the important variables with the most important variable odometer being the first best split.

We chose to include the variables age, type, cylinder, odometer, region, size, and title status in the decision trees model without including price_per_mile, mile_per_year, paint_color, make, fuel, and manufacturer. The reason we did not include those variables was because there could be correlation among those variables and the ones that we did include in our decision trees model. For example, price per mile and miles per year would be correlated with age, odometer, and price which would incorrectly change the importance of our main variables since they would redundantly be explaining the same variation within the data as the variables that we did include. Make and manufacturer have too many different values, and the decision tree would incorrectly skew importance to those variables; therefore, we decided to include region instead to capture that information without skewing the importance. We can see from the decision tree output for the variable importance we can see that the odometer value of the vehicle is most important when deciding the price of a vehicle. The least important variable in the model for deciding the price of a vehicle is title status.



The HPSPLIT Procedure

Model-Based Fit Statistics for Selected Tree

N Leaves	ASE	RSS
372	0.1478	32296.2

Variable Importance

Variable	Training Relative	Training Importance	Count
odometer	1.0000	89.5870	19
condition	0.9048	81.0558	27
type	0.7803	69.9058	81
age	0.4572	40.9585	34
cylinders	0.3830	34.3092	57
size	0.2782	24.9273	40
region	0.2402	21.5171	56
title_status	0.2174	19.4728	57

Predictive Modeling Using OLS

In this model, we are using ordinary least squares to predict the price of the vehicles using the other variables available to us in the dataset. First, we began by modeling the price using the odometer values as we hypothesized this to be the most important variable in predicting the price. The objective for this preliminary model was simply to get an idea for how to properly model these two variables together. It should be noted that the data used for the subsequent modeling contains only present values in the odometer column and no nulls or blanks. Below are the results:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.485262E12	2.485262E12	0.08	0.7708
Error	701737	2.053938E19	2.926934E13		
Corrected Total	701738	2.053938E19			

Root MSE	5410115	R-Square	0.0000
Dependent Mean	42029	Adj R-Sq	-0.0000
Coeff Var	12872		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	43656	8537.69165	5.11	<.0001
odo	1	-0.01429	0.04902	-0.29	0.7708

As we can see, the overall model is not significant, and the odometer variable is deemed insignificant as well. From our exploratory analysis and understanding of used cars, we know this to be untrue. Our remedy is to next subset the data, limiting to vehicles that are most reasonably being sold to everyday car buyers. Our subset includes cars that are no older than 40 years, odometer values less than 200,000 and prices less than $100,000. (**Appendix Figure 4,5, & 6).** Using this new subset, we can now check the model for significance:

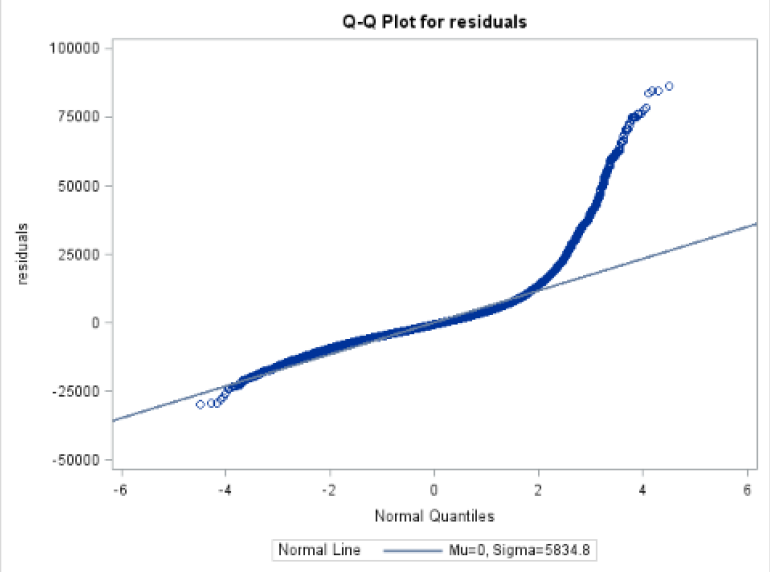
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.333499E13	2.333499E13	276881	<.0001
Error	590340	4.975275E13	84278133		
Corrected Total	590341	7.308774E13			

Root MSE	9180.31222	R-Square	0.3193
Dependent Mean	13440	Adj R-Sq	0.3193
Coeff Var	68.30619		

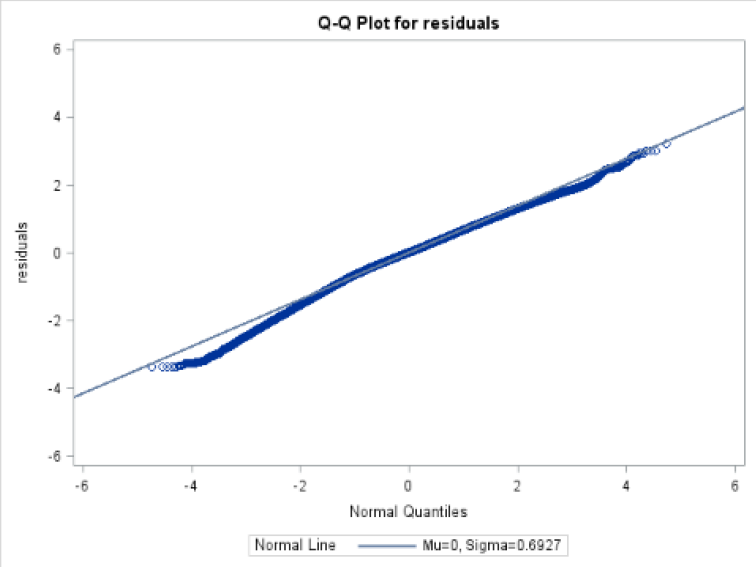
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	25881	26.49074	976.98	<.0001
odo	1	-0.12251	0.00023282	-526.19	<.0001

Now we can see a significant model along with the odometer variable being statistically significant, which verifies our assumption and understanding of the data. This model also explains ~32% of the variation in the price, which is quite high as a single variable model. Next, we can introduce more variables to the model to increase predictive power. We will introduce “age”, “type”, “condition”, “size”, “cyl”, “drive”, “fuel”, “region”, and “transmission”**.** Using the new parameters, our model now explains 75% of the variation in price. Before we interpret the results, we should first verify the assumptions on OLS, i.e. our residuals are normally distributed with a mean of zero and an equal variance. First, we examine the Q-Q plot of the residuals to check for normality:

Initial Q-Q Plot:



Final Q-Q Plot:



In our initial Q-Q plot, we saw fairly normal distribution up until the second quantile, and then the distribution trails off. We suspect the odometer variable is to blame for this strange effect because it is the only continuous variable (other than age, which is likely accurate after our subset) that might generate a different distribution. To account for this, we can then check for a log-linear relationship between price and the other predictor variables **(Appendix Figure 8).** Here we see a slightly more reasonable distribution, however it is still clearly non-linear. Next, we can check for a polynomial relationship between price and odometer by squaring the odometer term. Here we finally see a relatively normal distribution of our residuals. We next need to check for the mean and variance of our residual term. To check for the mean, we can simply output some statistics on the residuals saved from our model (see also **Appendix Figure 9** for residual vs predicted plot):

The UNIVARIATE Procedure
Variable: residuals

Moments			
N	178991	Sum Weights	178991
Mean	0	Sum Observations	0
Std Deviation	0.42999771	Variance	0.18489803
Skewness	-0.263761	Kurtosis	3.30362452
Uncorrected SS	33094.8987	Corrected SS	33094.8987
Coeff Variation	.	Std Error Mean	0.00101637

The REG Procedure
Model: MODEL1
Dependent Variable: lprice

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
4	8452.71	<.0001

Here we can see that the residuals do in fact have a mean of zero, however performing a test of equal variance yields that the variances are unequal. To correct this, our model should specify that the variances are unequal and adjust our standard error estimates (white correction). Lastly, before our model interpretations, we check for multicollinearity using the variance inflation factor. Below is our output:

Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	B	7.50102	0.10424	71.96	<.0001	0
odo	1	-0.00000216	2.326537E-8	-92.97	<.0001	2.77419
odo*odo	1	2.52811E-13	3.57342E-15	70.75	<.0001	2.60889
age	1	-0.03748	0.00027327	-137.17	<.0001	1.44410
type SUV	B	0.18142	0.01480	12.26	<.0001	9.91203
type bus	B	0.44914	0.06931	6.48	<.0001	1.05151
type conver	B	0.56249	0.01920	29.30	<.0001	2.32333
type coupe	B	0.35819	0.01647	21.74	<.0001	4.12895
type hatchb	B	0.15377	0.01769	8.69	<.0001	2.91728
type mini-v	B	0.13705	0.01966	6.97	<.0001	2.28990
type offroa	B	0.52456	0.03101	16.91	<.0001	1.27937
type other	B	0.37911	0.02824	13.42	<.0001	1.33869
type pickup	B	0.29599	0.01599	18.51	<.0001	5.74342
type sedan	B	0.00270	0.01465	0.18	0.8536	10.44271
type truck	B	0.38645	0.01566	24.70	<.0001	7.74034
type van	B	0.28379	0.01911	14.85	<.0001	2.39554
type wagon	0	0	.	.	.	.
condition excellent	B	1.82489	0.02941	62.05	<.0001	51.67218
condition fair	B	0.62018	0.03029	20.47	<.0001	13.23662
condition good	B	1.39321	0.02939	47.40	<.0001	48.00972
condition like new	B	2.03325	0.02990	68.00	<.0001	23.57527
condition new	B	1.97879	0.03849	51.41	<.0001	2.38880
condition salvage	0	0	.	.	.	.
size compact	B	0.01398	0.01774	0.79	0.4307	9.27085
size full-size	B	0.21334	0.01780	11.99	<.0001	19.16891
size mid-size	B	0.15145	0.01757	8.62	<.0001	15.97541
size sub-compact	0	0	.	.	.	.
cyl 3	B	-0.57633	0.09468	-6.09	<.0001	2.97119
cyl 4	B	-0.44107	0.07711	-5.72	<.0001	300.11079
cyl 5	B	-0.52613	0.07934	-6.63	<.0001	15.93021
cyl 6	B	-0.41222	0.07694	-5.36	<.0001	331.11752
cyl 8	B	-0.16104	0.07693	-2.09	0.0363	318.95404
cyl 10	B	-0.58347	0.08155	-7.15	<.0001	9.06443
cyl 12	0	0	.	.	.	.
drive 4wd	B	0.08545	0.00622	13.74	<.0001	2.28808
drive fwd	B	-0.27307	0.00695	-39.29	<.0001	2.70907
drive rwd	0	0	.	.	.	.
fuel diesel	B	0.46170	0.05339	8.65	<.0001	40.25005
fuel electr	B	0.05186	0.16504	0.31	0.7533	1.11480
fuel gas	B	-0.11384	0.05275	-2.16	0.0309	45.25845
fuel hybrid	B	0.16973	0.05727	2.96	0.0030	6.75558
fuel other	0	0	.	.	.	.
region american	B	-0.13749	0.00764	-17.99	<.0001	3.40044
region asian	B	-0.02617	0.00791	-3.18	0.0015	3.14159
region european	0	0	.	.	.	.
transmission automatic	1	0.63013	0.02956	21.32	<.0001	19.69603
transmission manual	1	0.75229	0.03025	24.87	<.0001	19.78282

NOTE: this test was performed solely for the purpose of determining multicollinearity, and the coefficient estimates, and other statistics should be ignored.

Typically, we use a threshold value of 10 on our VIF to determine if there is multicollinearity in the model. Here we see VIFs over 10 for "SUV" and "Sedan" car types, "excellent", "good", "fair", and "like new" conditions, "full-size" and "mid-size" sizes, 4,5,6, and 8 cylinders, "diesel" and "gas" fuel, and "automatic" and "manual" transmissions. While these values indicate the problem of multicollinearity in our model, we can justify the presence because these categorical variables contain three or more categories; therefore, they represent multicollinearity among themselves rather than with other predictor variables. **(***Statistical Horizons***).** Under this final model, we have verified the assumptions on OLS and achieved an r-squared value of .75. We can now interpret the results of our predictive model **(Appendix Figure 7).** It should first be noted that by selecting the variables that we used in our model, the model will only use the observations for which there is data in those variables, so our final observation count for the model is 178990. Some statistically significant interpretations from our model are as follows:

- Holding all else equal, a vehicle that drives 10,000 more miles will decrease its price by ~6.67%.
- For every additional year that a vehicle is driven, its price decreases by ~6.3%.
- SUVs have an average price of ~20.2% higher as compared to sedans.
- Trucks have an average price that is ~39.6% higher as compared to sedans.
- Vehicles that have a condition of "excellent" have a ~6.4% higher average price than vehicles with a "new" condition. Likewise, vehicles with a condition of "like new" have a ~9% higher average price as compared to vehicles with a new condition.
- As expected, vehicles with more than 8, 10 and 12 cylinders have higher average prices as compared to vehicles with 6 cylinders.
- Vehicles with four-wheel drive are, on average, ~11% higher in price as compared to vehicles with rear wheel drive.

Statistically insignificant interpretations:

- According to our output, the only statistically insignificant variable observed is the "other" class of the "fuel" variable. All other variables in our analysis, including each class value, was deemed significant by the model. This is indicative of a model which accurately describes the relationship between our variables and the price.

Conclusion

Our exploratory data analysis has revealed some insight as to what recommendations we can make to people who are looking to buy vehicles on craigslist. Based on the new measures we created using the data set (miles driven per year and price per mile) we were able to provide some insight as to which type of vehicle and manufacturer might retain more value or be more reliable than others.

- We saw that minivans have the highest miles driven per year, while coupes, convertibles, and off-road vehicles had the lowest miles driven per year. For someone who is looking for vehicles that have been utilized the least, this insight might guide them towards buying a convertible or a coupe. These vehicles are generally well maintained because they're often collectible vehicles and owners of the vehicles usually don't drive them as daily drivers.

- On the other hand, minivans and SUVs have the highest miles driven per year most likely because they are family-oriented vehicles and used for road trips and long-distance travel. Through our EDA, we

also saw that Volvo, Morgan, GMC, Ford, Land Rover, Acura, and Chevy manufacturers had the highest average mileage across their vehicles within the dataset. Aston Martin and Ferrari had the lowest average mileage across their vehicles. This is probably because they are high end performance-oriented vehicle manufacturers and not meant to be used as daily drivers. For someone looking for a vehicle that can last a long time, they might look into buying a GMC, Ford, Chevy, Acura, Honda, Toyota, Lexus, Volvo, or Infiniti with low mileage because we see from the dataset that these manufacturers' average vehicle mileage is very high, indicating that their vehicles have the potential to last well over the 100,000 miles range.

- Looking at specific vehicle brands and their average miles driven per year, we can make the recommendation to avoid buying a Hyundai, Kia, and Ram if the consumer is looking for specific manufacturers which are used less. Hyundai, Kia, and Ram all have average miles per year over 12,500, which is higher than most manufacturers within the dataset indicating that these vehicles are used more per year than other vehicles.

- By looking at the price per mile, we are able to see which vehicles have more value given their price and odometer value. We saw that Asian manufacturers that have 6-cylinder cars have the highest price per mile compared to 6-cylinder cars from other regions. In the case of 4-cylinder cars, European manufacturers have the highest price per mile as compared to 4-cylinder cars from other regions. Using this information, we can clearly recommend that 6-cylinder cars whose manufacturer is from the Asian region would be a better buy for a buyer who is looking for value in their vehicle. We even know from the real world that 6-cylinder cars from Acura and Lexus, which are Asian manufacturers, have some of the highest resale value in the industry because of their build quality and reliability, so it's not too surprising that the data shows us this information. Asian buses and trucks, American coupes, and European SUVs also have a high price per mile, so buyers might focus on those types of vehicles from those regions if they are looking for value.

- And finally, we saw that Fiat, Morgan, Ram, and GMC had high price per mile. Morgan is a luxury performance car manufacturer so it's not surprising that their vehicles would have very high value. Many Fiats are collectible cars now, so they would also have a very high price as compared to how many miles they have. If looking for a manufacturer whose vehicles retain value, yet they are not luxury or performance-oriented vehicles, buyers may want to focus on Buick, Chevrolet, GMC, Ram, and Volkswagen according to our analysis with this dataset.

- Our hypothesis was that American made vehicles would not hold value as well as their foreign counterparts, but what we found in our data made us conclude otherwise. Most of the vehicles that retained their value overall were actually American vehicles such as GMC and Ram and do hold their value quite well as compared to other car manufacturers. These American manufacturers solely focus on building large trucks and SUVs which is why they might hold their value better because of their specialization in that sector of the vehicle industry.

Sources

Der, Geoff, and Brian Everitt. *A Handbook of Statistical Analyses Using SAS*. CRC Press, 2009.

"MORGAN: A BRITISH MOTORING ICON." *Morgan Motor*, www.morgan-motor.com/about-morgan/.

Reese, Austin. "Used Cars Dataset." *Kaggle*, 15 July 2019, www.kaggle.com/austinreese/craigslist-carstrucks-data#craigslistVehicles.csv.

"SAS Documentation." *SAS Support*, support.sas.com/en/documentation.html.

"When Can You Safely Ignore Multicollinearity?" *Statistical Horizons*, statisticalhorizons.com/multicollinearity.

Appendix

Appendix A: Preprocessing Data

We mostly used Microsoft Excel as our tool to preprocess the data. It should be noted that Excel has a maximum capacity of just over one million rows that can be opened at one time. During our preprocessing, we were working with the maximum number of rows, which is likely not the entire dataset. For the purposes of our analysis, the rows we ultimately used contained the most amount of relevant information we deemed necessary, and despite not containing all of the available observations, should still represent the Craigslist vehicle population well.

To add the index, we simply created values ranging from 1 to the end of our observations counting by 1. To filter the data, we again used Excel's filtering functions to find observations meeting certain criteria. After filtering to find rows that were unnecessary for our data, we deleted these rows from our dataset.

To repopulate the "manufacturer" column with values, we used VLOOKUP to create an index using our existing manufacturer/make data. This index included all of the unique combinations of manufacturer and make in a separate table. This way, we could use our existing data to derive missing data and have more information available when it came time to perform analysis.

Appendix B: Exploratory Data Analysis

Figure 1:
This box plot shows the average odometer reading for each manufacturer as well as the 1st and 3rd quartile odometer readings for each manufacturer. We can see that the Morgan manufacturer has very low average mileage reading which indicates why the price per mile value for that manufacturer was so high. There were so many outliers for each of these manufacturers that our team had to zoom into a specific range of odometer values to see the actual box plots. We can see that most of the vehicles have a similar average odometer reading with the manufacturer "Saturn" having the lowest odometer reading.

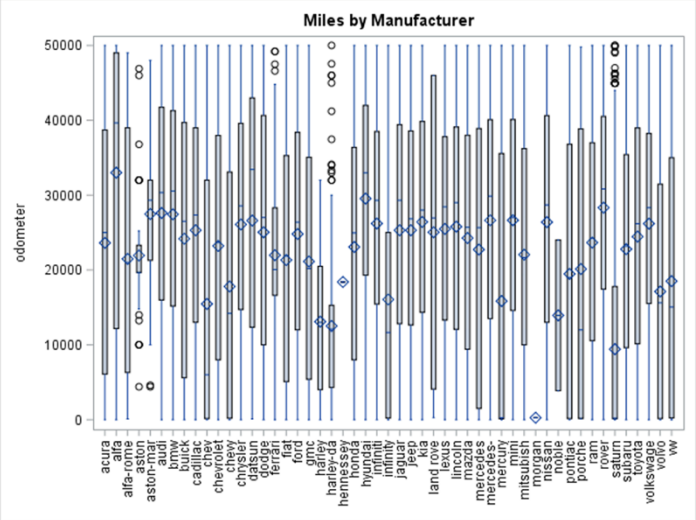


Figure 2:

In this graph we were able to determine the condition of the top 7 manufacturers in the dataset as determined by the seller. As seen in the graph below the top 7 manufacturers averaged out the same number of vehicles that were described as “like new”, “good”, “fair”, and “excellent”. Nissan slightly stands out from the rest of the pack by having more cars being described as “excellent” and “like new”.

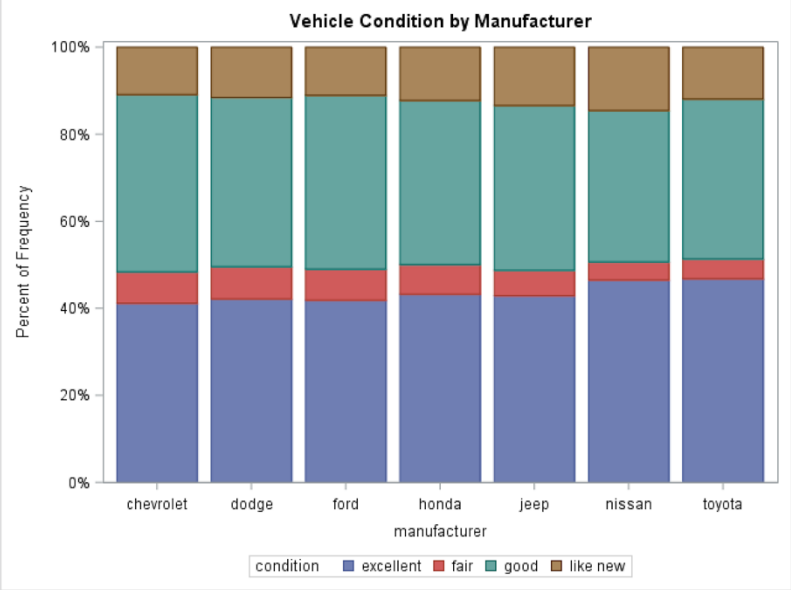


Figure 3:

This stacked bar chart shows the distribution of the cars sold on craigslist by the type of car. Then each bar shows the distribution of each type of car by the region the car manufacturer is from. We can see that American manufacturers make a large portion of SUVs, pickups, trucks, coupes, convertibles, mini vans, vans, off roads and others. Cars such as SUVs, trucks, pickups, mini vans, and vans are larger and more expensive than smaller cars such as sedans. American manufacturers have a higher frequency of these types of cars. Asian manufacturers have the largest share for sedan, hatchback, and wagon. European have the lowest frequency and have lower shares than American and Asian manufacturers in each car type except for convertible.

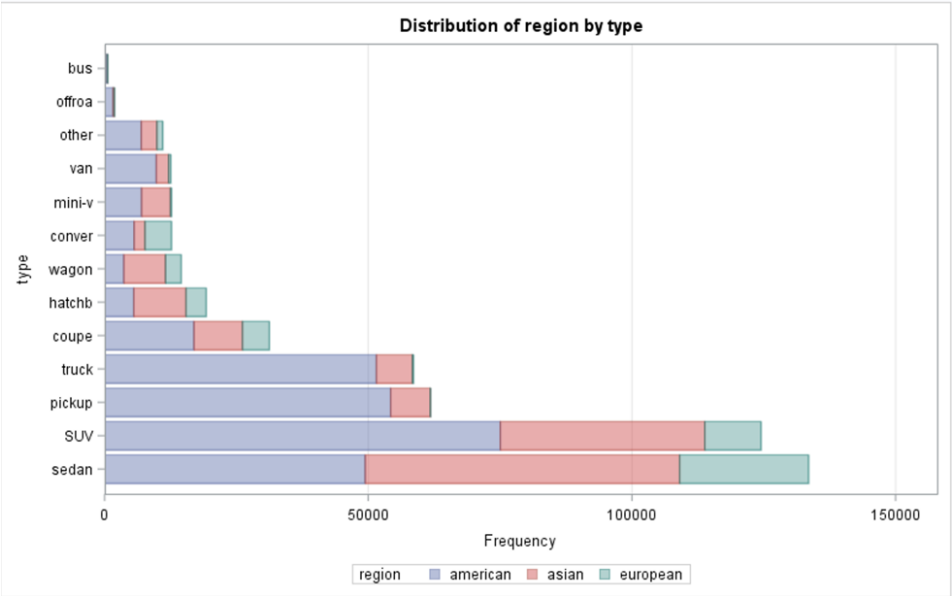


Figure 4:

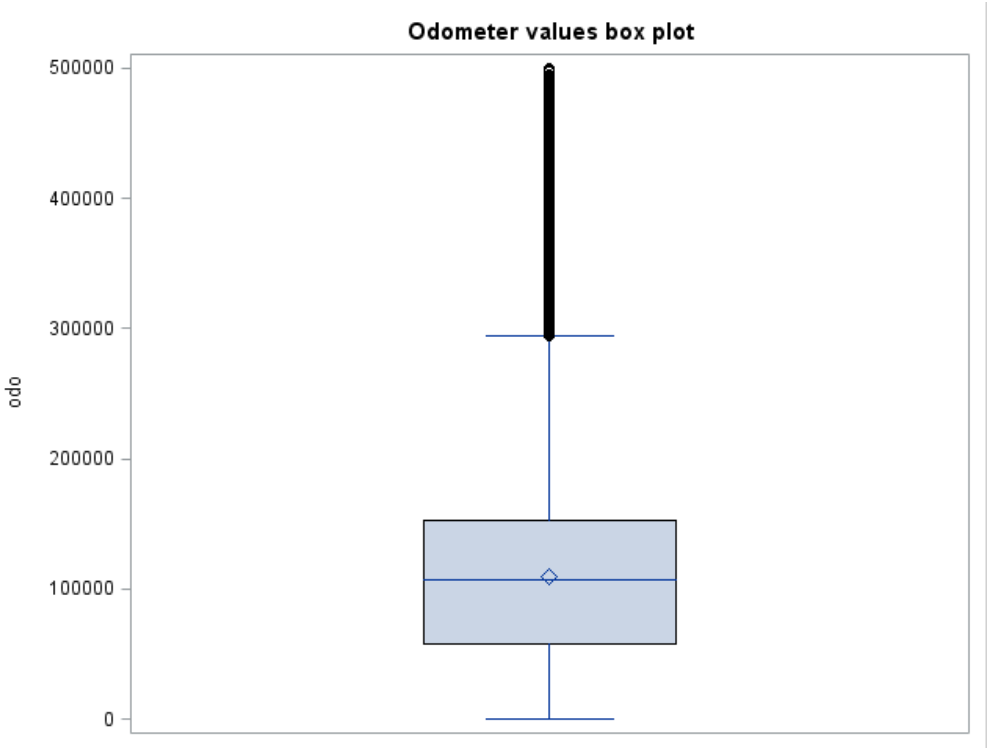
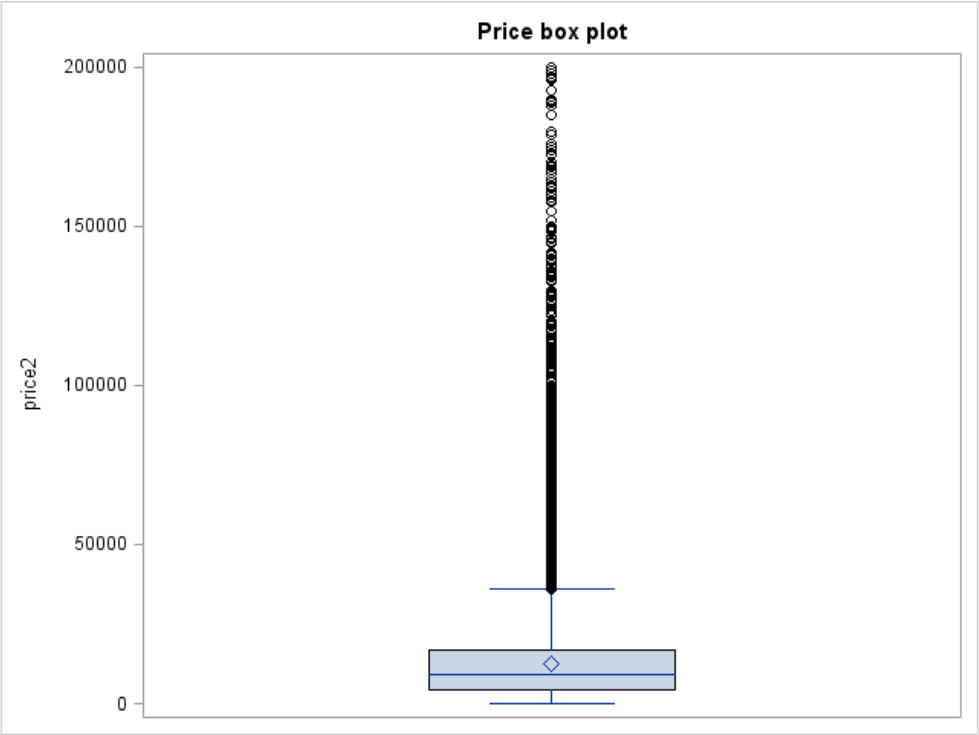


Figure 5:

Figure 6:



Appendix C: Empirical Analysis

Figure 7:

Intercept	1	10.103129	0.015304	660.15	<.0001
odo	1	-0.000006674	8.8823138E-8	-75.13	<.0001
odo*odo	1	5.216193E-12	3.988296E-13	13.08	<.0001
age	1	-0.063454	0.000201	-315.10	<.0001
type SUV	1	0.201914	0.003392	59.52	<.0001
type bus	1	0.364670	0.036838	9.90	<.0001
type conver	1	0.392461	0.006885	57.01	<.0001
type coupe	1	0.230361	0.004844	47.56	<.0001
type hatchb	1	0.058638	0.005603	10.47	<.0001
type mini-v	1	0.139306	0.007017	19.85	<.0001
type offroa	1	0.604467	0.014881	40.62	<.0001
type other	1	0.324944	0.013750	23.63	<.0001
type pickup	1	0.343446	0.004700	73.07	<.0001
type truck	1	0.396427	0.004301	92.17	<.0001
type van	1	0.264317	0.006655	39.72	<.0001
type wagon	1	0.035650	0.007369	4.84	<.0001
type sedan	0	0	.	.	.
condition excellent	1	0.063692	0.014613	4.36	<.0001
condition fair	1	-0.675350	0.015588	-43.33	<.0001
condition good	1	-0.167566	0.014691	-11.41	<.0001
condition like new	1	0.089787	0.014735	6.09	<.0001
condition salvage	1	-0.795577	0.025794	-30.84	<.0001
condition new	0	0	.	.	.
region american	0	0	.	.	.
transmission manual	1	0.150247	0.003922	38.31	<.0001
transmission other	1	0.078042	0.015481	5.04	<.0001
transmission automatic	0	0	.	.	.

size compact	1	-0.044193	0.003559	-12.42	<.0001
size full-size	1	0.032298	0.002612	12.37	<.0001
size sub-compact	1	-0.085740	0.008662	-9.90	<.0001
size mid-size	0	0	.	.	.
cyl 3	1	-0.366568	0.028065	-13.06	<.0001
cyl 4	1	-0.154393	0.003073	-50.24	<.0001
cyl 5	1	-0.152042	0.010437	-14.57	<.0001
cyl 8	1	0.261833	0.002998	87.34	<.0001
cyl 10	1	0.503581	0.014709	34.24	<.0001
cyl 12	1	0.602249	0.039141	15.39	<.0001
cyl 6	0	0	.	.	.
drive 4wd	1	0.113929	0.003177	35.86	<.0001
drive fwd	1	-0.212034	0.003502	-60.54	<.0001
drive rwd	0	0	.	.	.
fuel diesel	1	0.601202	0.004899	122.72	<.0001
fuel electr	1	0.147882	0.072759	2.03	0.0421
fuel hybrid	1	0.265339	0.010780	24.61	<.0001
fuel other	1	0.034423	0.025751	1.34	0.1813
fuel gas	0	0	.	.	.
region asian	1	0.167975	0.002617	64.19	<.0001
region european	1	0.170420	0.003729	45.70	<.0001

Figure 8:

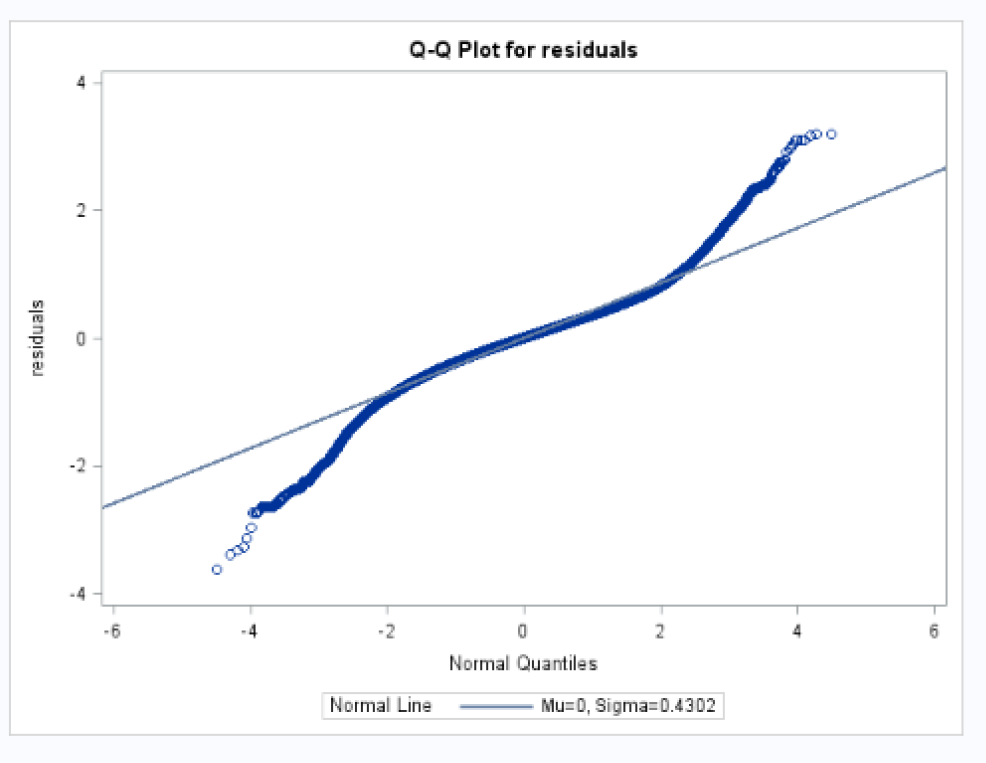
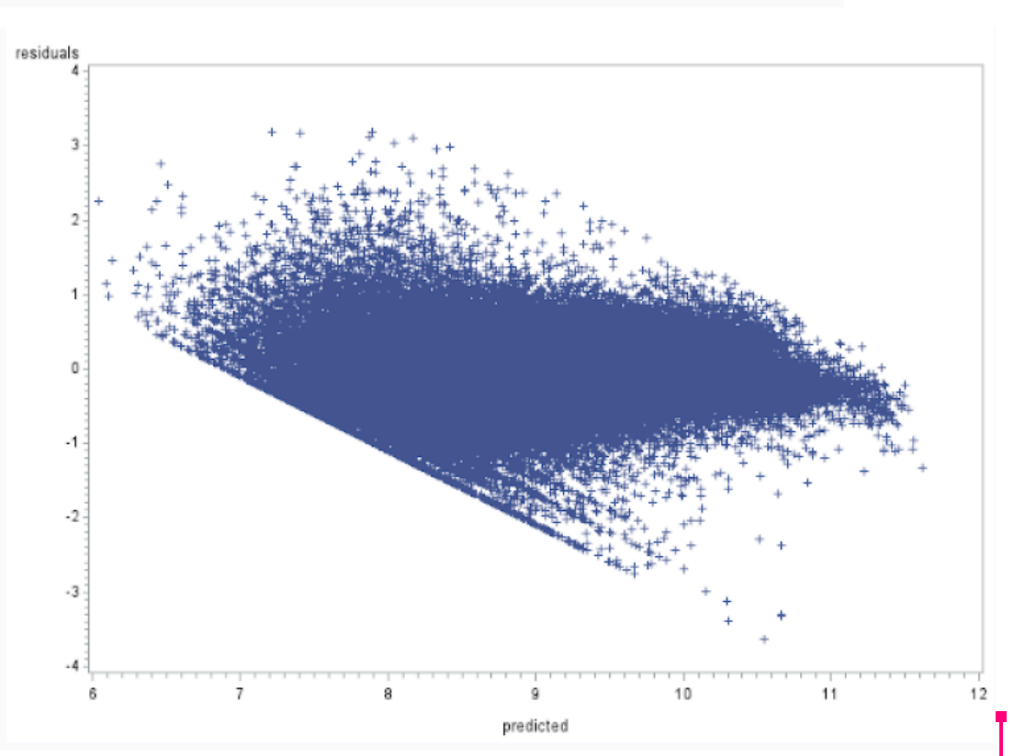


Figure 9:



Appendix D: SAS CODE

```
/*************************************************************
FINAL PROJECT: Group 7

Last updated: July 29, 2019
*************************************************************/


proc contents data = Craigslist;
run;
/*Variables age, price per mile, mile per year, lprice added*/
data work.Craigslist;
set work.Craigslist;
        age = 2019 - year;
        price_per_mile = price/odometer;
        mile_per_year = odometer/age;
        lprice = log(price);
run;
/*Variable for region where manufacturer is found added*/
data work.Craigslist;
set work.Craigslist;
if manufacturer = 'audi' or manufacturer = 'bmw' or manufacturer = 'alfa'or
manufacturer = 'alfa-rome' or manufacturer = 'aston' or manufacturer = 'aston-
mar' or manufacturer = 'ferrari' or manufacturer = 'fiat' or manufacturer = 'jaguar'
or manufacturer = 'land rove' or manufacturer = 'landrover' or manufacturer =
'mercedes' or manufacturer = 'mercedes-' or manufacturer = 'mercedesb' or
manufacturer = 'mini' or manufacturer = 'morgan' or manufacturer = 'noble' or
manufacturer = 'porsche' or manufacturer = 'rover' or manufacturer = 'volkswage'
or manufacturer = 'volvo' or manufacturer = 'vw'
then region = "european";

if manufacturer = 'buick' or manufacturer = 'cadillac' or manufacturer = 'chev' or
manufacturer = 'chevrolet' or manufacturer = 'chevy' or manufacturer = 'chrysler'
or manufacturer = 'dodge' or manufacturer = 'ford' or manufacturer = 'gmc' or
manufacturer = 'jeep' or manufacturer = 'lincoln' or manufacturer = 'mercury' or
manufacturer = 'pontiac' or manufacturer = 'ram' or manufacturer = 'saturn'
then region= "american";
```

```
if manufacturer = 'acura' or manufacturer = 'datsun' or manufacturer = 'honda' or
manufacturer = 'hyundai' or manufacturer = 'infiniti' or manufacturer = 'infinity' or
manufacturer = 'kia' or manufacturer = 'lexus' or manufacturer = 'mazda' or
manufacturer = 'mitsubish' or manufacturer = 'nissan' or manufacturer = 'subaru'
or manufacturer = 'toyota'
then region= "asian";
run;
/*Distribution of Price by Region Histogram*/
proc univariate data= Craigslist;
        where price between 0 and 100000;
        class region;
        var price;
        histogram price /endpoints = 0 to 100000 by 1000
                                nrows=3 odstitle="Distribution of Price by
Manufacturer's Region"
                                normal(noprint);
        inset mean mode median var="Variance" std="Std Dev" / pos = ne format =
6.3;
        title;
        run;
/*Distribution of Price of entire dataset histogram*/
proc univariate data= Craigslist;
        where price between 0 and 100000;
        var price;
        histogram price /endpoints = 0 to 100000 by 1000
                        odstitle="Distribution of Price"
                                normal(noprint);
        inset mean mode median var="Variance" std="Std Dev" / pos = ne format =
6.3;
        title;
        run;
/*Distribution of Odometer histogram*/
proc univariate data= Craigslist;
        where odometer between 1000 and 450000;
        var odometer;
        histogram /endpoints = 1000 to 450000 by 5000
                        normal(noprint);
```

```
        inset mean mode median var="Variance" std="Std Dev" / pos = ne format =
6.3;
        run;
/*Distribution of Odometer by region histogram*/
proc univariate data= Craigslist;
        where odometer between 1000 and 450000;
        class region;
        var odometer;
        histogram odometer/endpoints = 1000 to 450000 by 5000
                        nrows=3 odstitle="Distribution of Odometer by
Manufacturer's Region"
                        normal(noprint);
        inset mean mode median var="Variance" std="Std Dev" / pos = ne format =
6.3;
        run;
/*Relationship between Odometer and Price scatterplot*/
proc sgplot data = Craigslist;
        where price between 1 and 100000;
        where also odometer between 1000 and 450000;
        scatter x = odometer y = price;
        title 'Relationship between Odometer and Price';
        run;
        quit;
/*Relationship between Age and Price scatterplat*/
proc sgplot data = Craigslist;
        where price between 1 and 100000;
        where also age < 40;
        scatter x = age y = price;
        title 'Relationship between Age and Price';
        run;
quit;

/*Distribution of Price per Mile by Manufacturer's Region*/
proc univariate data= Craigslist;
        where price_per_mile between 0 and .5;
        class region;
        var price_per_mile;
        histogram price_per_mile/endpoints = 0 to .5 by .01
```

```
                        nrows=3 odstitle="Distribution of Price per Mile by
Manufacturer's Region"
                        normal(noprint);
        inset mean mode median var="Variance" std="Std Dev" / pos = ne format =
6.3;
        run;
/*Scatterplot of log(price) by odometer*/
proc sgplot data = Craigslist;
        where price between 1000 and 100000;
        where also odometer between 1000 and 300000;
        scatter y = lprice x = odometer;
        title "Relationship between Odometer and Log(price)";
        run;
quit;
/*Miles by Manufacturers*/
proc sgplot data= Craigslist;
where odometer between 0 and 1000000;
where also manufacturer NE '#N/A';
where also manufacturer NE '#NAME?';
vbar manufacturer/response = odometer group = manufacturer
groupdisplay = cluster stat= mean;
xaxis display=(nolabel) valuesrotate=vertical;
title 'Miles by Manufacturers';
run;
/*Price/Mile by Manufacturers Region*/
proc sgplot data= Craigslist;
where price_per_mile between 0 and 4000;
vbar region /response = price_per_mile
group = region groupdisplay = cluster stat= mean;
title 'Price/Mile by Manufacturers Region';
run;
/*Price/Mile of Manufacturers*/
proc sgplot data= Craigslist;
where price_per_mile between 0 and 4000;
where also manufacturer NE "#N/A";
where also manufacturer NE "#NAME?";
vbar manufacturer / response = price_per_mile
group = manufacturer groupdisplay = cluster stat= mean;
xaxis display=(nolabel) valuesrotate=vertical;
```

```
title 'Price/Mile of Manufacturers';
run;
/*Miles by Manufacturer*/
proc sgplot data =Craigslist;
where odometer between 0 and 50000;
where also manufacturer NE '#N/A';
where also manufacturer NE '#NAME?';
vbox odometer/category = manufacturer;
xaxis display=(nolabel) valuesrotate=vertical;
title 'Miles by Manufacturer';
run;
/*Price/Mile by Type of Vehicle*/
proc sgplot data= Craigslist;
where price_per_mile between 0 and 4000;
vbar type /response= price_per_mile group = type
groupdisplay = cluster stat= mean;
title 'Price/Mile by Type of Vehicle';
run;
/*Mile/Year by Type of Vehicle*/
proc sgplot data= Craigslist;
where mile_per_year between 0 and 70000;
vbar type /response = mile_per_year group = type
groupdisplay = cluster stat= mean;
title 'Mile/Year by Type of Vehicle';
run;
/*Price/Mile by Region of Different Types of Vehicles*/
proc sgplot data = Craigslist;
where price <=100000;
where also odometer >100;
where also age >=1;
/*where also cyl in (4,6,8,10);*/
where also type NE "offroa"; where also type NE "van";
where also condition NE 'new';
vbar region /response = price_per_mile group = type
groupdisplay = cluster stat= mean;
title 'Price/Mile by Region of Different Types of Vehicles';
run;
/*Mile/Year by Manufacturer*/
proc sgplot data= Craigslist;
```

```
where mile_per_year;
vbar manufacturer /response = mile_per_year group = manufacturer
groupdisplay = cluster stat= mean;
xaxis display=(nolabel) valuesrotate=vertical;
title 'Mile/Year by Manufacturer';
run;
/*Relationship between price and odometer*/
proc format;
value odo_bins
0 = "Exactly 0"
0 -< 10000 = "0 - 10K"
10000 -< 20000 = "10K - 20K"
20000 -< 30000 = "20K - 30K"
30000 -< 40000 = "30K - 40K"
40000 -< 50000 = "40K - 50K"
50000 -< 60000 = "50K - 60K"
60000 -< 70000 = "60K - 70K"
70000 -< 80000 = "70K - 80K"
80000 -< 90000 = "80K - 90K"
90000 -< 100000 = "90K - 100K"
100000 -< 110000 = "100K - 110K"
110000 -< 120000 = "110K - 120K"
120000 -< 130000 = "120K - 130K"
130000 -< 140000 = "130K - 140K"
140000 -< 150000 = "140K - 150K"
150000 -< 160000 = "150K - 160K"
160000 -< 170000 = "160000 - 170000"
170000 -< 180000 = "170000 - 180000"
180000 -< 190000 = "180000 - 190000"
190000 -< 200000 = "190000 - 200000"
200000 -< 210000 = "200000 - 210000"
210000 -< 220000 = "210000 - 220000"

ods graphics on;
proc freq data= Craigslist;
tables odometer/ missing plots=freqplot;
format odometer odo_bins.;
run;
```

```
proc sgplot data = Craigslist;
scatter y = price x = odo;
format odometer odo_bins.;
run;
quit;

proc sgplot data = Craigslist;
where price <= 100000
and odometer between 1000 and 300000;
vbar odometer / response = price group = odo
groupdisplay = cluster stat = mean;
format odometer odo_bins.;
xaxis display = (nolabels) valuesrotate= vertical;
title "Relationship between Price and Odometer";
run;
/*Regression Tree*/
proc hpsplit data=work.cars2 seed=123;
class condition region cylinders type size title_status;
model log_price = age condition cylinders odometer region type size title_status;
run;
/*modeling*/
proc reg data = proj2;
model price2 = odo;
run;

*/ initial model had no significance between odometer value and price, which
logically makes no sense. Some data
points are likely skewing our values. Reference figure (3). Next we decided on
ranges for price, odometer, and age
based on boxplots and understanding of vehicle markets;
proc reg data = proj2;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
model price2 = odo;
run;
*/ Now our model is significant and the odometer variable explains quite a bit of
variation in the price. Using this
```

```
model as a base, we should now include more variables to better predict the
price.;
ods graphics on;
proc glmselect data = proj2;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
class type condition size cyl drive fuel region transmission;
glmodel: model price2 = type condition odo size cyl age drive fuel region
transmission /selection = none;
title "MODEL";
output out=out r=residuals;
run;
quit;
*/ After including more variables in our model, we reach an r-squared value of .59,
meaning that about 59% of the
variation in the price is explained by our model. Before we interpret the results, we
should verify that the
assumptions on OLS regression are satisfied.;
proc univariate data=out;
var residuals;
histogram residuals / normal kernel;
qqplot residuals / normal(mu=est sigma=est);
title "";
run;
*/ Checking for the residuals being normally distributed, we see they are certainly
not. We should check for a log-linear
relationship between price and our predictor variables. ;

proc glmselect data = proj2;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
class type condition size cyl drive fuel region transmission;
glmodel: model lprice = odo age type condition size cyl drive fuel region
transmission / selection = none;
title "MODEL";
output out=out r=residuals;
run;
```

```
quit;

proc univariate data=out;
var residuals;
histogram residuals / normal kernel;
qqplot residuals / normal(mu=est sigma=est);
run;
*/ Fitting a log-linear model, we yield the same results as before: the residuals are
not normally distributed. Next,
we should check for a polynomial relationship in our model. Logically, of our two
continuous variables odometer
and age, odometer is most likely the variable causing the non-normality of the
residuals. We will try squaring
the term to see if the model fits better;

proc glm data = proj2;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
/* class type condition size cyl drive fuel region transmission;*/
glmodel: model price2 = odo odo*odo;
title "MODEL";
output out=out r=residuals;
run;
quit;

proc univariate data=out;
var residuals;
histogram residuals / normal kernel;
qqplot residuals / normal(mu=est sigma=est);
run;
/* Fitting the polynomial model, we still see non-normality in the error term. Next
we can check again for a log-linear
relationship in the polynomial model. ;
/* Best Model So Far MODEL*/
proc glm data = proj2;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
```

```
glmodel: model lprice = odo odo*odo;
title "MODEL";
output out=out r=residuals p= predicted;
run;
quit;

proc univariate data=out;
var residuals;
histogram residuals / normal kernel;
qqplot residuals / normal(mu=est sigma=est);
run;
ods graphics on;
proc glm data = proj2 plot = diagnostics;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
glmodel: model lprice = odo odo*odo;
title "MODEL";
output out=out r=residuals p= predicted;
run;
quit;
/* now we finally see a fairly normal error term. We should next evaluate the mean
and variance to verify the other
OLS assumptions. */
proc gplot data = out;
plot residuals*predicted;
run;
ods graphics on;
proc model data = proj2 plot = resplots;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
lprice = odo + odo**2;
fit "lprice"
title "MODEL";
run;
quit;

proc glm data = proj2;
```

```
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
class type condition size cyl drive fuel region transmission;
glmodel: model lprice = odo odo*odo age type condition size cyl drive fuel region
transmission;
title "MODEL";
output out=out r=residuals p= predicted;
run;
quit;
proc gplot data = out;
plot residuals*predicted;
run;
ods graphics on;
proc univariate;
var residuals;
run;

proc reg data = proj2;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
model lprice = odo odosq / white;
title "MODEL";
output out=out r=residuals p= predicted;
run;
quit;
proc univariate;
var residuals;
run;
proc glmselect data = proj2;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
class type(ref="sedan") condition(ref="new") size(ref="mid-size") cyl(ref="6")
drive(ref="rwd")
fuel(ref="gas") region(ref="american") transmission(ref="automatic");
glmodel: model lprice = odo odo*odo age type condition size cyl drive fuel region
transmission / selection = none ;
```

```
title "MODEL";
run;
quit;
proc model data = proj2;
where price2 between 1000 and 100000
and odo between 1000 and 200000
and age < 40;
lprice = odo + odosq + age + type + condition + size + cyl + drive + fuel + region +
transmission ;
fit lprice / ols HCCME = 3;
title "MODEL";
run;
quit;

proc glmmod noprint outdesign=ds2 data=proj2;
  class type condition size cyl drive fuel region transmission;
  model lprice = odo odo*odo age type condition size cyl drive fuel region
transmission;
run;

proc reg data=ds2;
  model lprice = col2 col3 col4 col5 col6 col7 col8 col9 col10 col11 col12 col13 col14
col15 col16
col17 col18 col19 col20 col21 col22 col23 col24 col25 col26 col27 col28 col29
col30 col31 col32 col33 col34 col35 col36
col37 col38 col39 col40 col41 col42 col43 col44 col45 col46 col47 / vif;
```