PREDICTIVE ANALYTICS – PROJECT UPDATE

Ali Abbas – axa180088

Neeladri Mohapatra – nnm170003

Bhumika Panchal – bsp190005

Juan Antonio Sillero – jxs190031

Bohan Zhang – bxz190005

Executive Summary

The technological advancements that have occurred since the turn of the century have made a tremendous impact on many different industries throughout the world. One of the more important advancements relates to analytics, and the role they currently play in business. The rise of analytics has made data ubiquitous in business, making it one of the most valuable assets. In our assessment, we will use predictive analytics to analyze the price of different automobiles based on different characteristics, such as location, make, model, and other variables that will be further analyzed in this project. Scrapped in January 2020, this dataset contains relevant information on used car listings on Craigslist. Craigslist contains the world's largest collection of used vehicles for sale, making it a relevant source of data. Our analysis will consist of various techniques of data pre-processing, along with many different multivariate statistical techniques.

A. Introduction

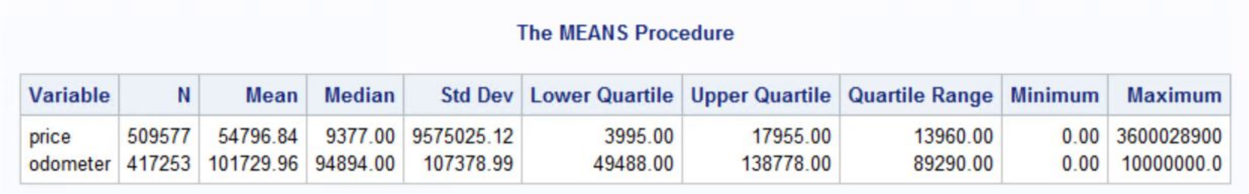
Craigslist is a classified advertisements website that contains different sections, which include listings pertaining to job openings, housing, and used cars. The website supports many different languages and is accessible in many different countries. The dataset selected for analysis is a compilation of used car listings, which include 25 variables. This process can help both the dealers and the customers of used cars in several ways: (i) identify the correlation between price and odometer; (ii) identify the probability of a used car being listed for sale, given the miles on the odometer; (iii) identify the possible discrepancies in price amongst states; These analyses can help customers by offering insights on whether a deal is better or worse than expected, or if a deal is a rare find or a bad offer. This can also help dealers understand the true value of their car, possibly increasing the amount they could receive in the car. The price of used cars is also dependent on the condition of the car, as well as what type of car it is.

B. Data Description

The data that is included in this data set contains every used vehicle entry in all the North-American regions on Craigslist. The columns that are included in the data set are the following:

- Year
- VIN
- URL
- Type
- Transmission
- Title_Status
- State
- Size
- Region
- Region_URL
- Price
- Paint_color
- Odometer
- Transmission
- Model
- Manufacturer
- Longitude
- Latitude
- Image_URL
- ID
- Fuel
- Drive
- Description
- Cylinder
- County
- Condition

The below initial data analysis is for the numeric columns odometer and price in the dataset. The maximum values show the outliers for the price. Compared to the rest of the data, it is an outlier for used cars listing to be at 3600 million. While processing the data, we eliminated extreme outliers.



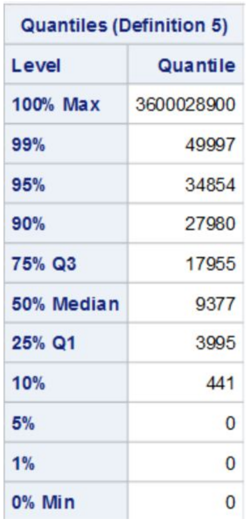
The MEANS Procedure

Variable	N	Mean	Median	Std Dev	Lower Quartile	Upper Quartile	Quartile Range	Minimum	Maximum
price	509577	54796.84	9377.00	9575025.12	3995.00	17955.00	13960.00	0.00	3600028900
odometer	417253	101729.96	94894.00	107378.99	49488.00	138778.00	89290.00	0.00	10000000.0

C. Preprocessing Data

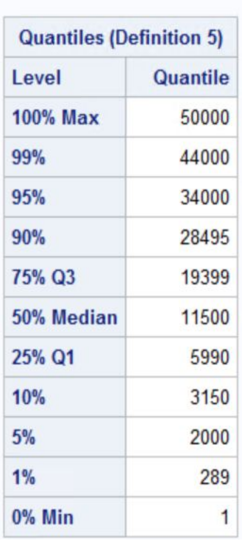
This dataset contained a large number of records with either a missing variable or variables that served little to no benefit to our analysis. The following variables were dropped from the data set:
URL, VIN, Region_URL, Longitude, Latitude, Image_URL, ID, and Description. These variables were dropped, either because they contained unique values and served no purpose in the predictive analysis, or because they contained insignificant values that serve no purpose in the predictive analysis.

The results attached below provide statistics on the price of the used cars included in the dataset. These statistics identified that the maximum value of the dataset was over 3.6B. Considering that 99% of the records were within 49,997, proving the outlier to be an irrelevant record, a condition was placed on price so that the values over 50,000 were not included in the results. The minimum value of the dataset was 0 -- since this accounted for over 5% of the data, a condition was added to ensure that only vehicles with a price of 0 were excluded in the results. As a result, only the records with a price between 1 and 50,000 made it to the result set. Please see below for the difference in the quartiles once the change was made. While performing research, it was discovered that craigslist does provide used cars for free under a charity program for low-income families. Additionally, cars are also placed without a price, as negotiation is done independently with the buyer and seller. Since the circumstances of those listings are unknown, removing these records better served the goal of the analysis.



Quantiles (Definition 5)	
Level	Quantile
100% Max	3600028900
99%	49997
95%	34854
90%	27980
75% Q3	17955
50% Median	9377
25% Q1	3995
10%	441
5%	0
1%	0
0% Min	0

Before



Quantiles (Definition 5)	
Level	Quantile
100% Max	50000
99%	44000
95%	34000
90%	28495
75% Q3	19399
50% Median	11500
25% Q1	5990
10%	3150
5%	2000
1%	289
0% Min	1

After

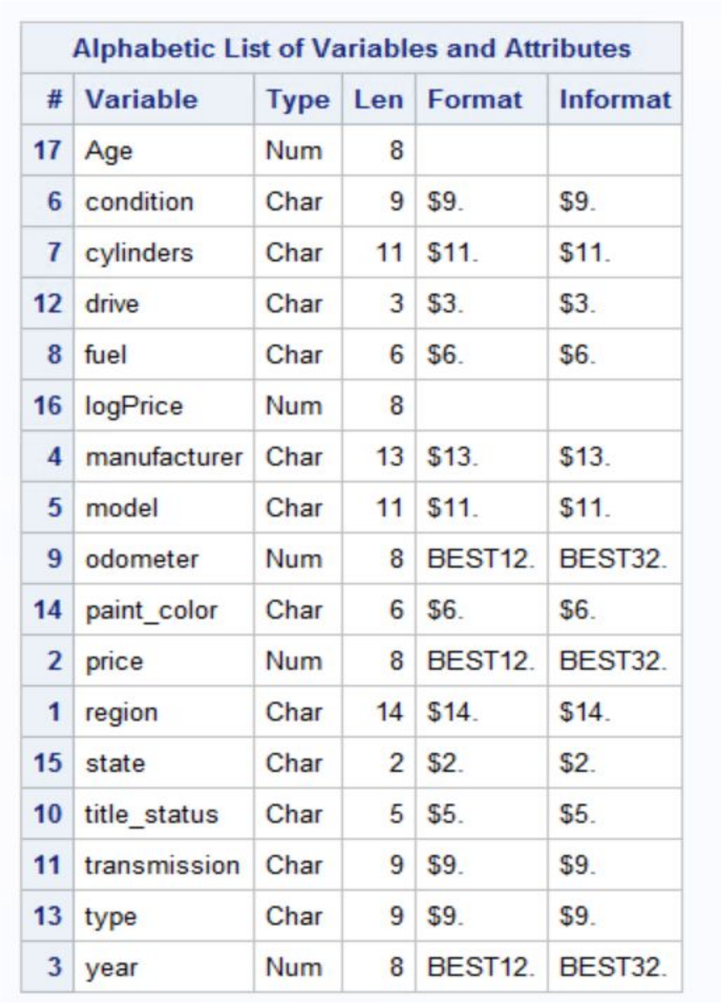
The “County” variable was empty in the dataset, making it an obvious drop candidate. The variable “Region” better served our analysis -- however, this was not an ideal variable either. The records that were in the region of “Santa Barbara” had a null value for variable “State”. Since this was the only exception, the null value was replaced with “CA” to adequately update the results.

The variable “Size” is dropped because the data for around 67% of those records were missing. After applying the above preprocessing steps, the content description of the modified dataset is as follows in Fig 1:

The missing value count for all the variables was calculated on the exported data. Below is the screenshot for the missing value report in fig 2.

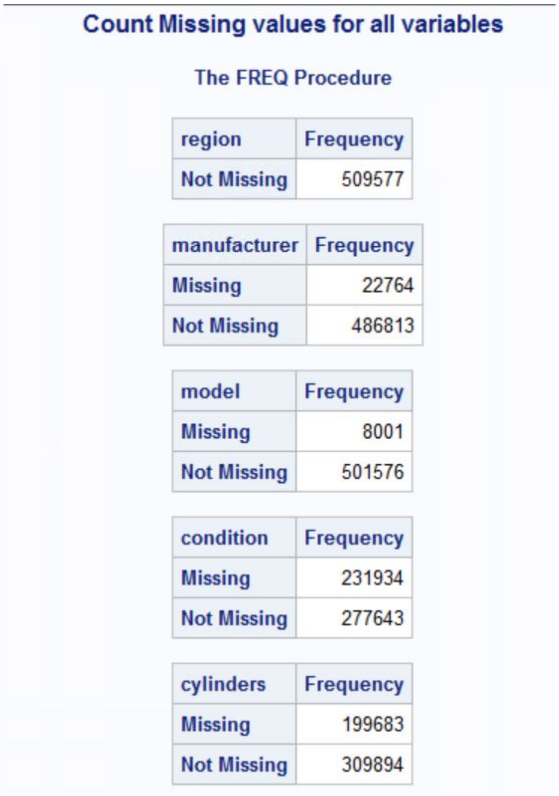
If the missing values are to be replaced and imputed in the dataset, the following assumptions can be made.

1. The missing count shows that there is more than 70% of data with a null value for variable “Size” in the dataset, making the variable droppable.
2. Drop the variable “County”, as over 99% of data is missing.
3. For the variable “Cylinder”, grouping the results by model, replace the missing values with the mode of the group.
4. Replace the missing values in type with the mode of the group by model and year.



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
17	Age	Num	8		
6	condition	Char	9	$9.	$9.
7	cylinders	Char	11	$11.	$11.
12	drive	Char	3	$3.	$3.
8	fuel	Char	6	$6.	$6.
16	logPrice	Num	8		
4	manufacturer	Char	13	$13.	$13.
5	model	Char	11	$11.	$11.
9	odometer	Num	8	BEST12.	BEST32.
14	paint_color	Char	6	$6.	$6.
2	price	Num	8	BEST12.	BEST32.
1	region	Char	14	$14.	$14.
15	state	Char	2	$2.	$2.
10	title_status	Char	5	$5.	$5.
11	transmission	Char	9	$9.	$9.
13	type	Char	9	$9.	$9.
3	year	Num	8	BEST12.	BEST32.



Count Missing values for all variables

The FREQ Procedure

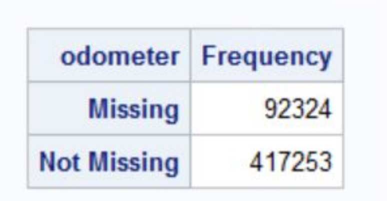
region	Frequency
Not Missing	509577

manufacturer	Frequency
Missing	22764
Not Missing	486813

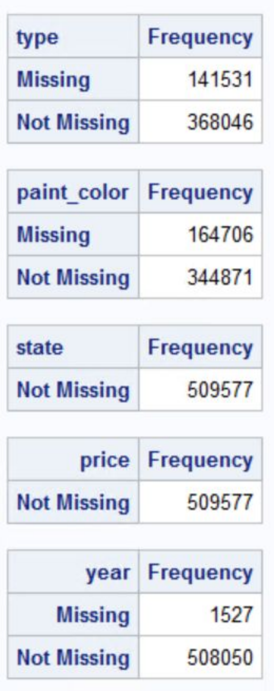
model	Frequency
Missing	8001
Not Missing	501576

condition	Frequency
Missing	231934
Not Missing	277643

cylinders	Frequency
Missing	199683
Not Missing	309894



odometer	Frequency
Missing	92324
Not Missing	417253



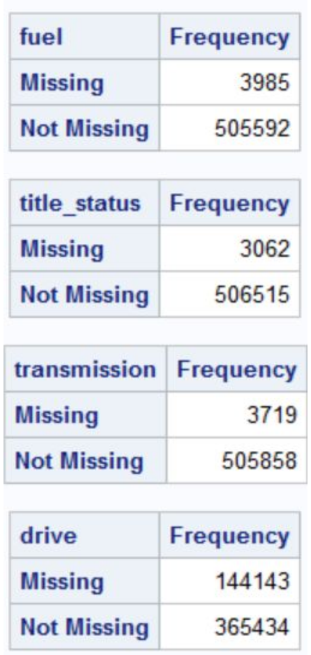
type	Frequency
Missing	141531
Not Missing	368046

paint_color	Frequency
Missing	164706
Not Missing	344871

state	Frequency
Not Missing	509577

price	Frequency
Not Missing	509577

year	Frequency
Missing	1527
Not Missing	508050



fuel	Frequency
Missing	3985
Not Missing	505592

title_status	Frequency
Missing	3062
Not Missing	506515

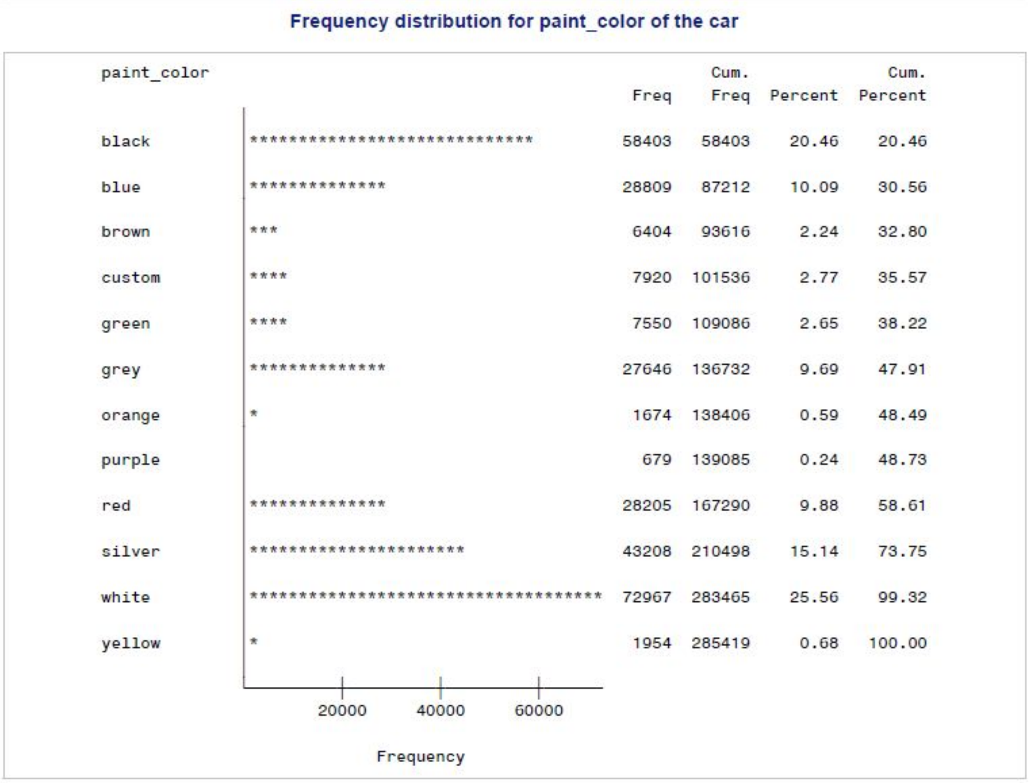
transmission	Frequency
Missing	3719
Not Missing	505858

drive	Frequency
Missing	144143
Not Missing	365434

Fig 2

D. Exploratory Data Analysis

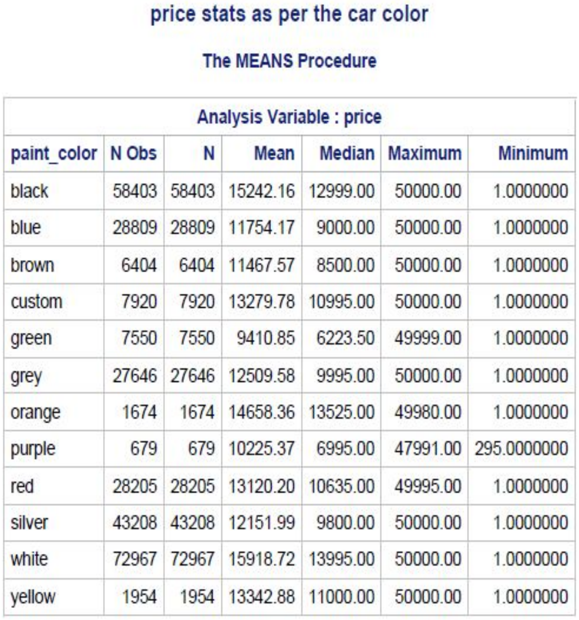
1. The paint color of the car impacts the price. About 45% of the listing was either black or white, while another 35% were grey, silver, or blue. The below figures show the color impacting the car price. The black cars show the highest mean price of the used cars.



Frequency distribution for paint_color of the car

```
paint_color                                                      Cum.              Cum.
                                                         Freq     Freq   Percent   Percent
                   |
black              | *****************************         58403    58403     20.46     20.46
                   |
blue               | **************                        28809    87212     10.09     30.56
                   |
brown              | ***                                    6404    93616      2.24     32.80
                   |
custom             | ****                                   7920   101536      2.77     35.57
                   |
green              | ****                                   7550   109086      2.65     38.22
                   |
grey               | **************                        27646   136732      9.69     47.91
                   |
orange             | *                                      1674   138406      0.59     48.49
                   |
purple             |                                         679   139085      0.24     48.73
                   |
red                | **************                        28205   167290      9.88     58.61
                   |
silver             | **********************                43208   210498     15.14     73.75
                   |
white              | ************************************  72967   283465     25.56     99.32
                   |
yellow             | *                                      1954   285419      0.68    100.00
                   |
                   ----------+---------+---------+---
                          20000     40000     60000

                               Frequency
```



price stats as per the car color

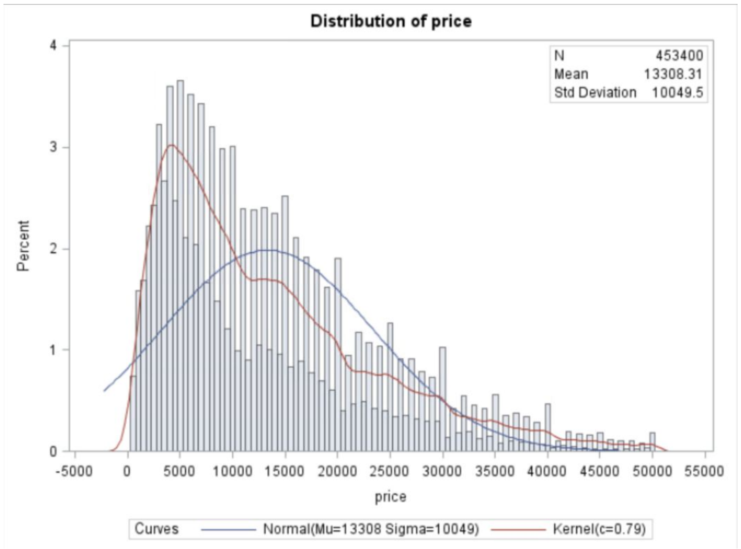
The MEANS Procedure

Analysis Variable : price

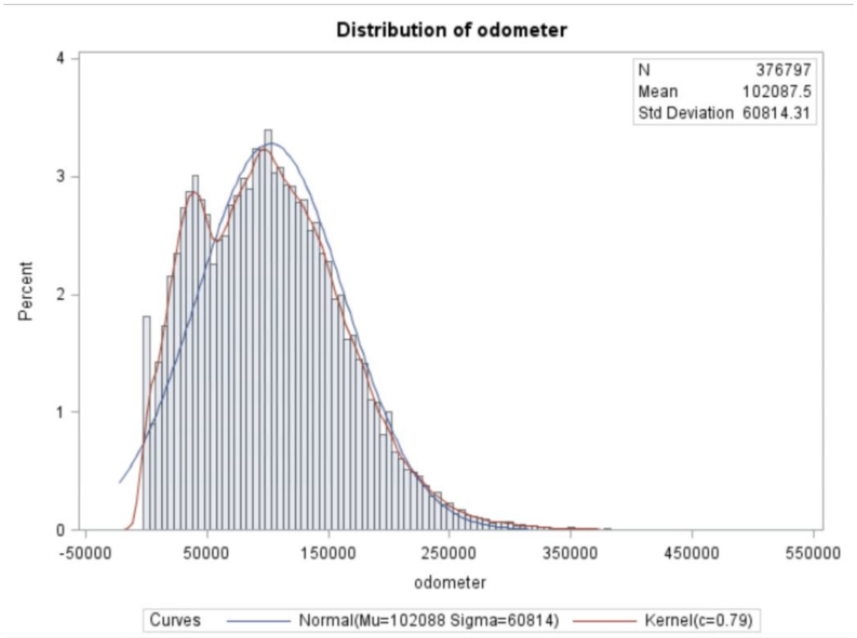
paint_color	N Obs	N	Mean	Median	Maximum	Minimum
black	58403	58403	15242.16	12999.00	50000.00	1.0000000
blue	28809	28809	11754.17	9000.00	50000.00	1.0000000
brown	6404	6404	11467.57	8500.00	50000.00	1.0000000
custom	7920	7920	13279.78	10995.00	50000.00	1.0000000
green	7550	7550	9410.85	6223.50	49999.00	1.0000000
grey	27646	27646	12509.58	9995.00	50000.00	1.0000000
orange	1674	1674	14658.36	13525.00	49980.00	1.0000000
purple	679	679	10225.37	6995.00	47991.00	295.0000000
red	28205	28205	13120.20	10635.00	49995.00	1.0000000
silver	43208	43208	12151.99	9800.00	50000.00	1.0000000
white	72967	72967	15918.72	13995.00	50000.00	1.0000000
yellow	1954	1954	13342.88	11000.00	50000.00	1.0000000

2. Histograms for price and odometer.

The histogram below illustrates the distribution of price. As seen below, the more expensive a car becomes, the less percent of the market share it has in the Craigslist used car market. Additionally, there is a drop off for the number of used cars listed for sale at over $20,000.



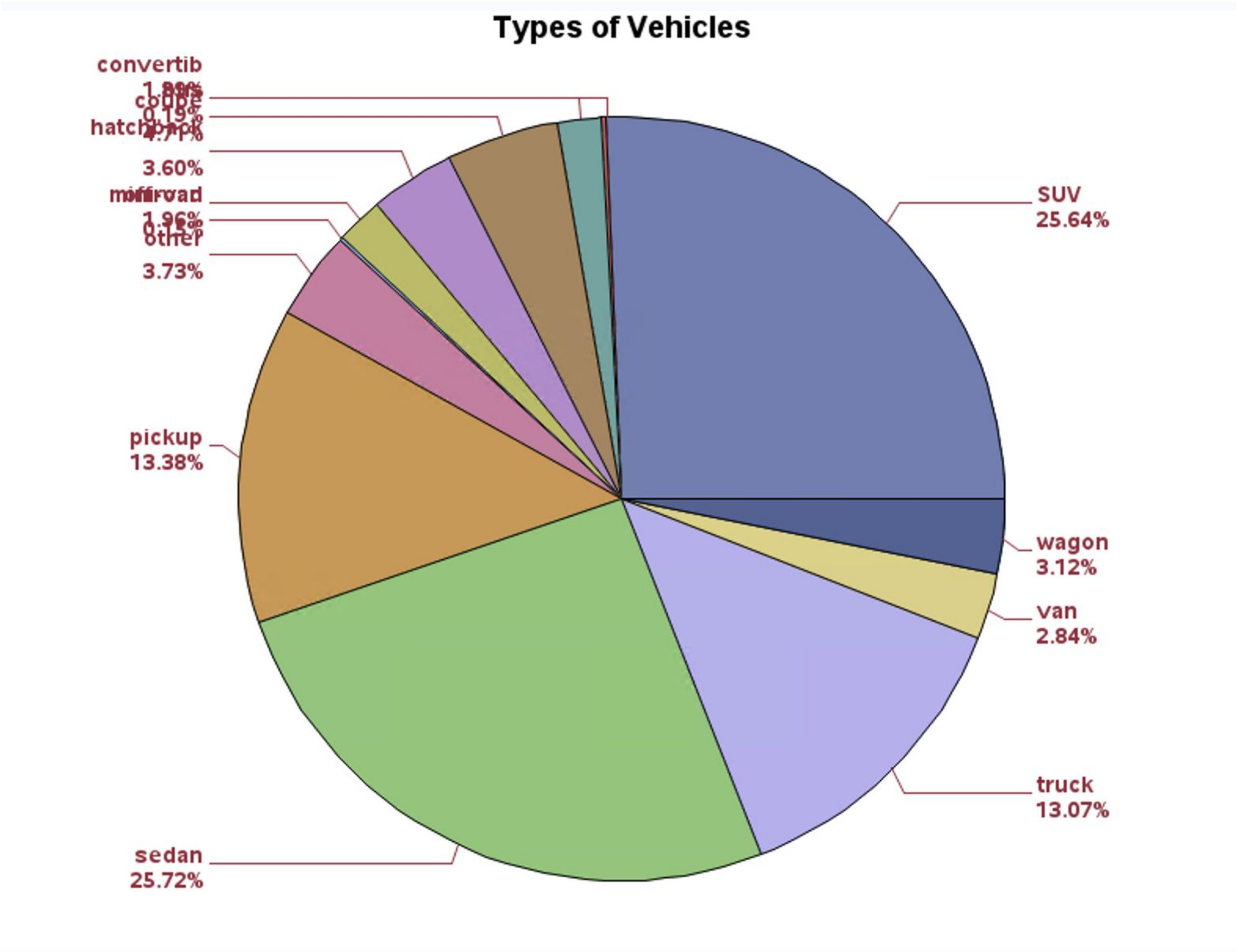
The histogram below illustrates the distribution of the odometer. The average miles on the odometer of a car is a little over 100,000 miles.



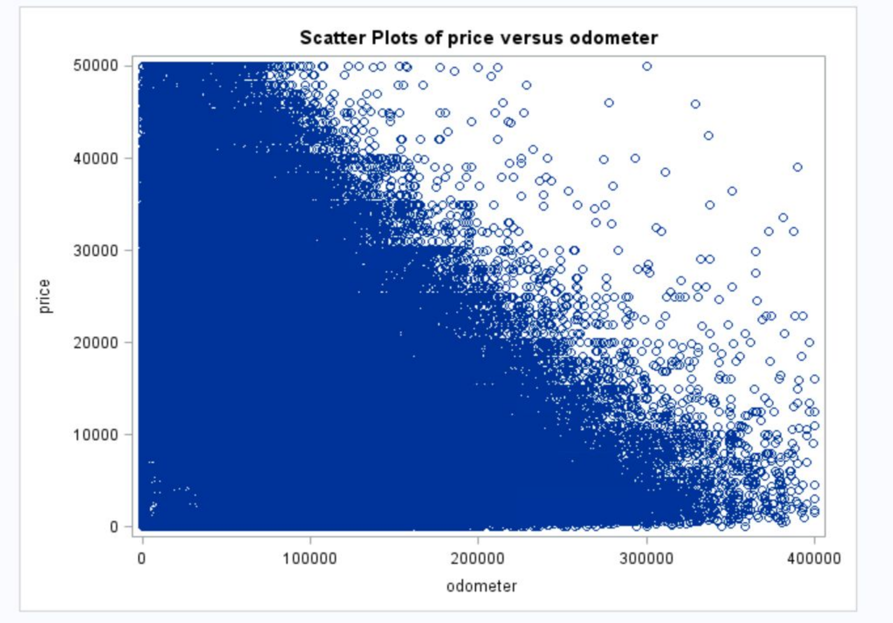
Frequency plot results for various variables show the following results.

a. Fuel: 88% of the car listings use **gas**.
b. Title_status: 96% of the cars are titled as **clean**.
c. Transmission: 91% of the cars have an **automatic** transmission as a type.
d. Type: About 50% of the listings have car types as **sedan and SUV**.
e. Condition: About 50% of the car listings say that the condition of the car is **excellent** followed by **good**.
f. Cylinders: Most of the car listings have cars with **4,6 or 8 cylinders** coming up with a total of 96% of the data.
g. States: Most of the car listings are from the registered states of CA, FL, and TX contributing to the 26% of the data.

The below pie chart shows the distribution of the data as per the type of car.

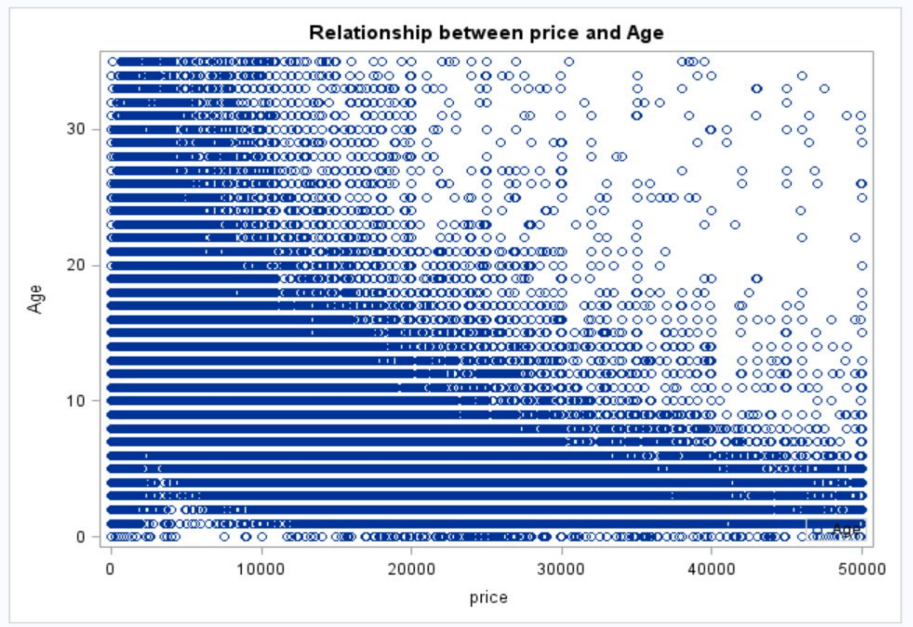


E. Empirical Anagtbd odometer



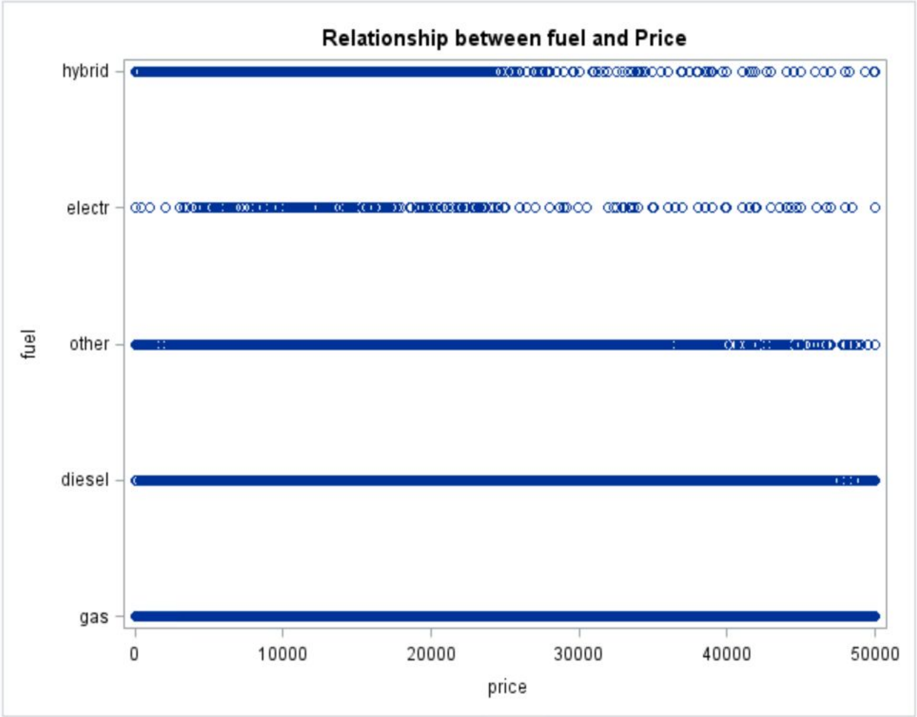
a. Relationship between Price and age:

The scatter plot illustrates that cars with a lesser age (indicated less used) tend to demand a higher sales price than cars with greater age. That being said, there are instances where cars with a higher age demand a higher price, suggesting there can be other factors adding to the price of the car. Cars aged 30+ years generate an average listing price from $10k to $15k and a maximum of 40k.



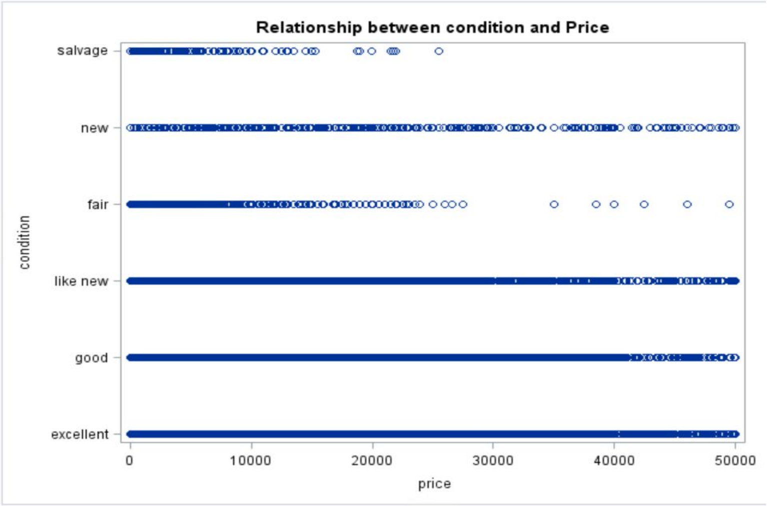
b. Relationship between fuel and price of the car.

The price of the car is affected by fuel but to a lesser extent. The average range of price for electric cars is between $5000 to $30,000, meanwhile, cars that run on gas have a price range from $0 to $50,000.



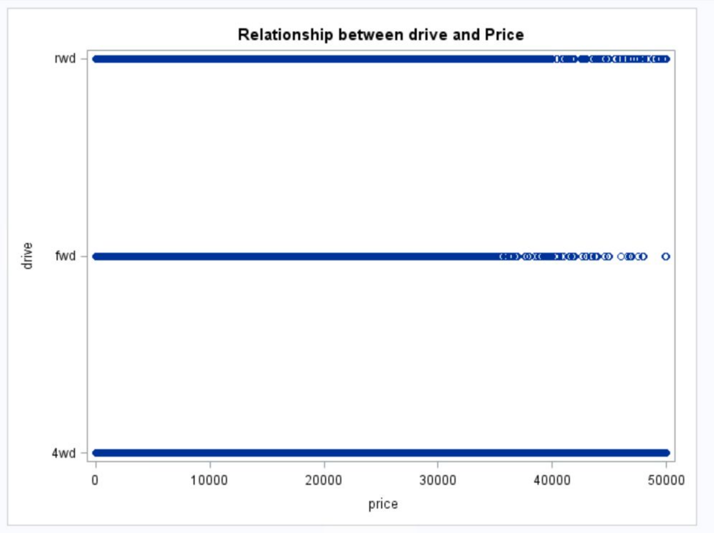
c. Relationship between price and condition of the car

As the price gets closer to the maximum limit of $50,000 as the condition improves. As the condition of the car decreases from excellent to fair, the average price of cars decreases to around $25,000. Thus the better the condition of the car, the greater the price of the car is.



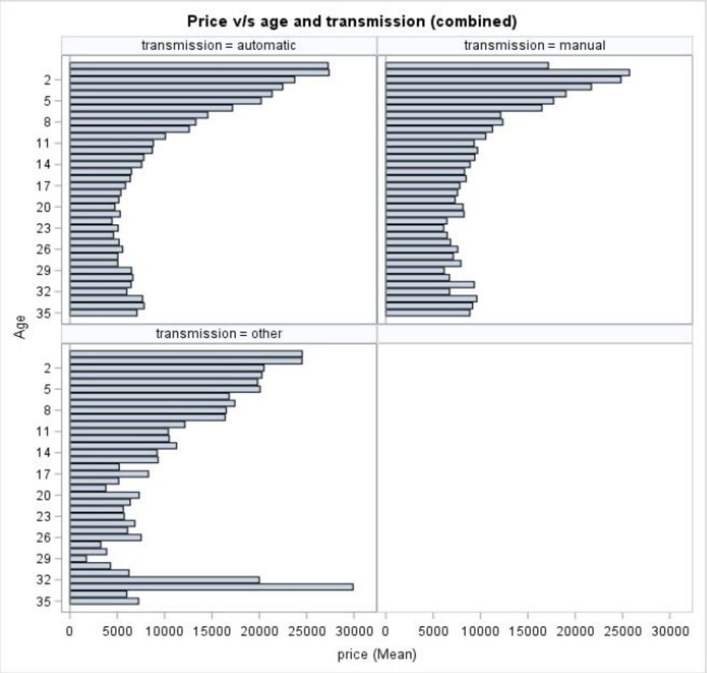
d. Relationship between drive and price

The drive data does not help much in predicting the price of the car.



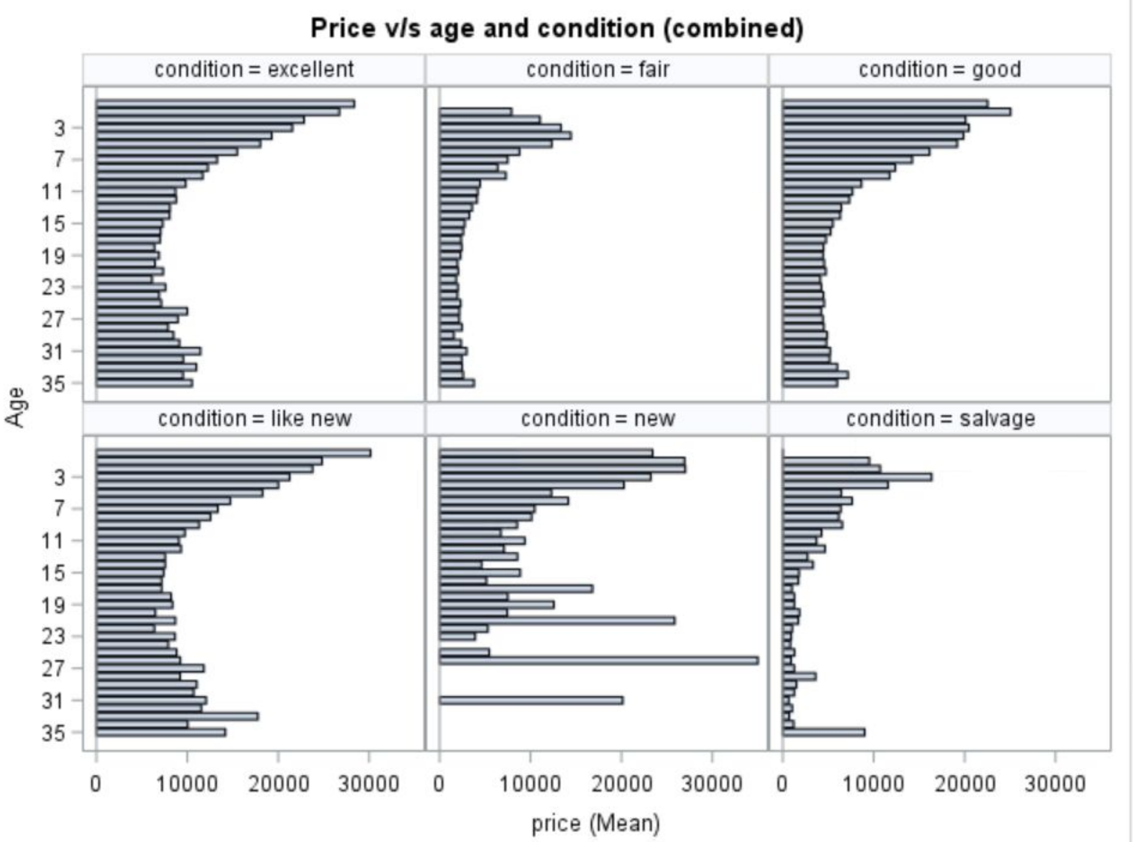
e. Price v/s transmission and age combined

For transmission type automatic and manual, the price of the car is maximum for age less than 5yrs. The graph for both types of transmission is exponential as the price drastically decreases as the age increases. While cars with other transmission types do not show the same relationship between age and price. Some older cars too have higher prices for different types of transmission in the car.



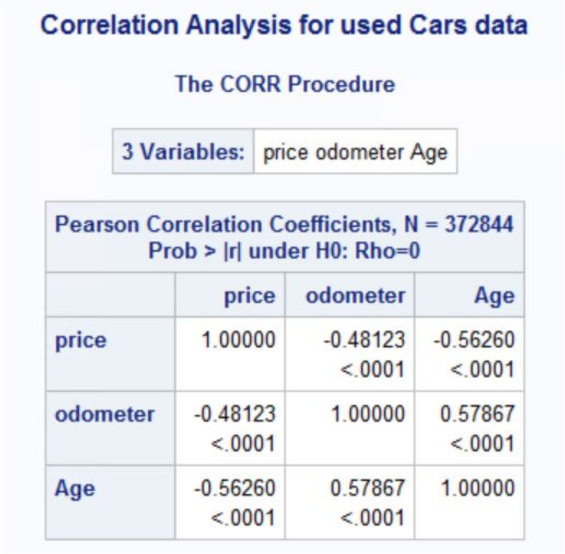
f. Price v/s condition and age combined

The graph behaves differently than the rest in case of the condition of cars as new. The prices somewhat hike with the age of new-like cars between the age of 23 to 35 yrs.



F. Correlation Analysis

Below is the correlation matrix for numeric data in the dataset. The correlation matrix shows odometer and price has negative weak correlation (If the threshold value considered is 0.5). The age and price have a slightly stronger correlation of -0.56. Thus as the age increases, the price of the car decreases. The age and odometer are positively correlated showing that if the age of the car is more, it is more used which is indicated by its odometer readings.



Correlation Analysis for used Cars data

The CORR Procedure

3 Variables:	price odometer Age

Pearson Correlation Coefficients, N = 372844 Prob > |r| under H0: Rho=0			
	price	odometer	Age
price	1.00000	-0.48123 <.0001	-0.56260 <.0001
odometer	-0.48123 <.0001	1.00000	0.57867 <.0001
Age	-0.56260 <.0001	0.57867 <.0001	1.00000

G. Regression Models

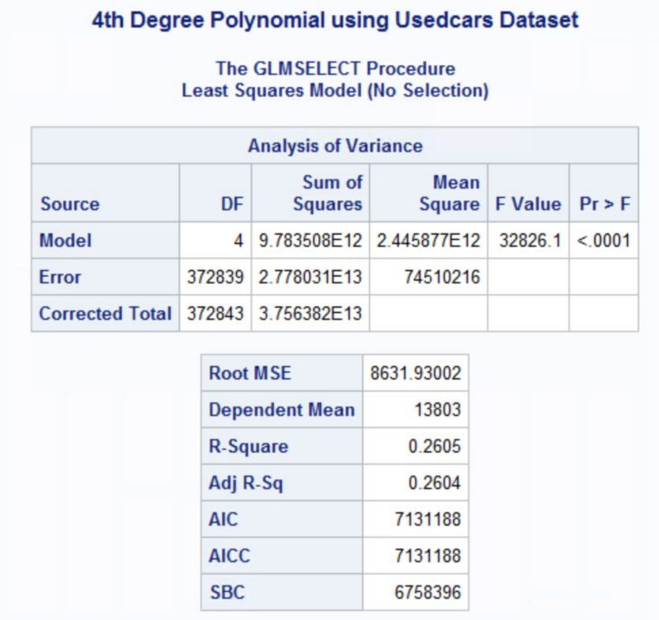
1. Linear Regression:

While running the degree one linear regression, the below output for odometer v/s price is obtained. The regression output shows the p-value of the odometer as less than 0.05. thus the data related to odometers is significant in analyzing and predicting the price of the car.

However, linear regression of degree one has a very low R Squared value, at 0.2316.

The model is the linear model for odometer v/s price using GLMSelect Procedure.

Applying the fourth-degree polynomial regression improved the R-square value of the model to 0.2605. The model for price v/s odometer is not a good-fit model.



4th Degree Polynomial using Usedcars Dataset

The GLMSELECT Procedure
Least Squares Model (No Selection)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	9.783508E12	2.445877E12	32826.1	<.0001
Error	372839	2.778031E13	74510216		
Corrected Total	372843	3.756382E13			

Root MSE	8631.93002
Dependent Mean	13803
R-Square	0.2605
Adj R-Sq	0.2604
AIC	7131188
AICC	7131188
SBC	6758396

The diagnostic plot shows that the model is a significant model. The collinearity diagnostic shows the adjusted r-square of 0.35.

H. Conclusions

After performing all the data exploration and polynomial regression, the following conclusions are made about the predictors and the target variables of the used car.

1. The price of the car on the craigslist car listings highly depends on the odometer and the age which is calculated by the year column in the dataset.
2. Price and odometer have a negative linear relationship.
3. Price and age have a negative linear relationship.
4. The condition of the used car is one of the strong price predictors.
5. The linear regression model has very low R-square value of 0.2316
6. However, the polynomial regression model of degree four improves the R-square value to 0.26
7. Concluding, the model is not a good-fit model. It has a higher RMSE value which makes it difficult to predict the price of the used cars in the future.