4/23/2020

Predictive Analytics

Analysis of Craigslist Car Sales

Executive Summary

As the world increases its reliance on analytics to predict the world around us, used cars are not to be left behind. In this document, we have presented the data from its origins to its new self after cleaning. Then we have detailed what we saw from the data and what values are important for creating a good price for a car. We also used these values to predict the cost of a car if we are given certain values that are important so we can help people to better understand a good price for buying or selling. Predominantly prices of cars are affected by variables such as the year the car was built in, the type of the car, the size of the car, number of miles it has been driven, the transmission of the car and other variables.

We also got to see that for cleaning the dataset several aspects need to be considered such as the importance of variables, percentage of missing value, what values need to be imputed and all of these combined to cross verify if our model fits the data well eventually.

We have hence tried our hands on various different modelling algorithms with data cleaning techniques to eventually come down to selection of better models after cleaning the data thoroughly.

1. Introduction

Craigslist is an online advertising website where users can list and view jobs, housing, items for sale, items wanted, services, etc. This allows anyone to see opportunities in their area whether they are looking for a new couch or needing to fill a job position. It was founded in 1995 and generates over a billion dollars in revenue.

The largest draw that users must use Craigslist is the fact that it’s fairly open in what a person can list and view listings for their needs. Unlike other sites that force it into a category, Craigslist allows for an open notice so discussions between people can happen with ease.

Our project focuses on predicting the price of a car by using the Craigslist used car sales which range across the USA.

2. Data Description

This dataset is scraped from the Craigslist website every month, it contains mostly all relevant information that Craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories.

The dataset that we are working on has about half a million records (0.56 million) with the following 25 variables:

1. Id - Entry id
2. Url - Listing url
3. Region - Craigslist region
4. Region_url - Region url

5. Price - Entry price
6. Year - Entry year
7. Manufacturer - Manufacturer of vehicle
8. Model - Model of vehicle
9. Condition - Condition of vehicle
10. Cylinders - Number of cylinders
11. Fuel - Fuel type
12. Odometer - Miles travelled by vehicle
13. Title_status - Title status of the vehicle
14. Transmission - Transmission of the vehicle
15. Vin - Vehicle identification number
16. Drive - Type of drive
17. Size - Size of vehicle
18. Type - Generic type of vehicle
19. Paint_color - Color of vehicle
20. Image_url - Image url
21. Description - Listed description of vehicle
22. County - County of listing
23. State - State of listing
24. Lat - Latitude of listing
25. Long - Longitude of listing

3. Exploratory Data Analysis and Data Preprocessing

To clean the data, we first began by looking at the data columns and consider which ones could be relevant to our analysis. In order to predict the price of the used cars we found that the following variables would not have helped us in predicting the price of the cars.

a. Id - An identification number and not required in price prediction
b. Url - A URL and not required in price prediction as these are mere links
c. Region - Region name not required in price prediction
d. Region_url - URL again has no dependency on price
e. Vin - Vehicle identification number has no impact on price
f. State - We could not see any relationship between price and state so it was omitted.
g. County - The names of the counties did not offer any valuable explanation of price. We believe this might be caused by the US having over three thousand counties.
h. Description - Description of the car is not required for price prediction. There is another variable which we have considered used to describe the condition of the car
i. Image_url - URL again has no dependency on price

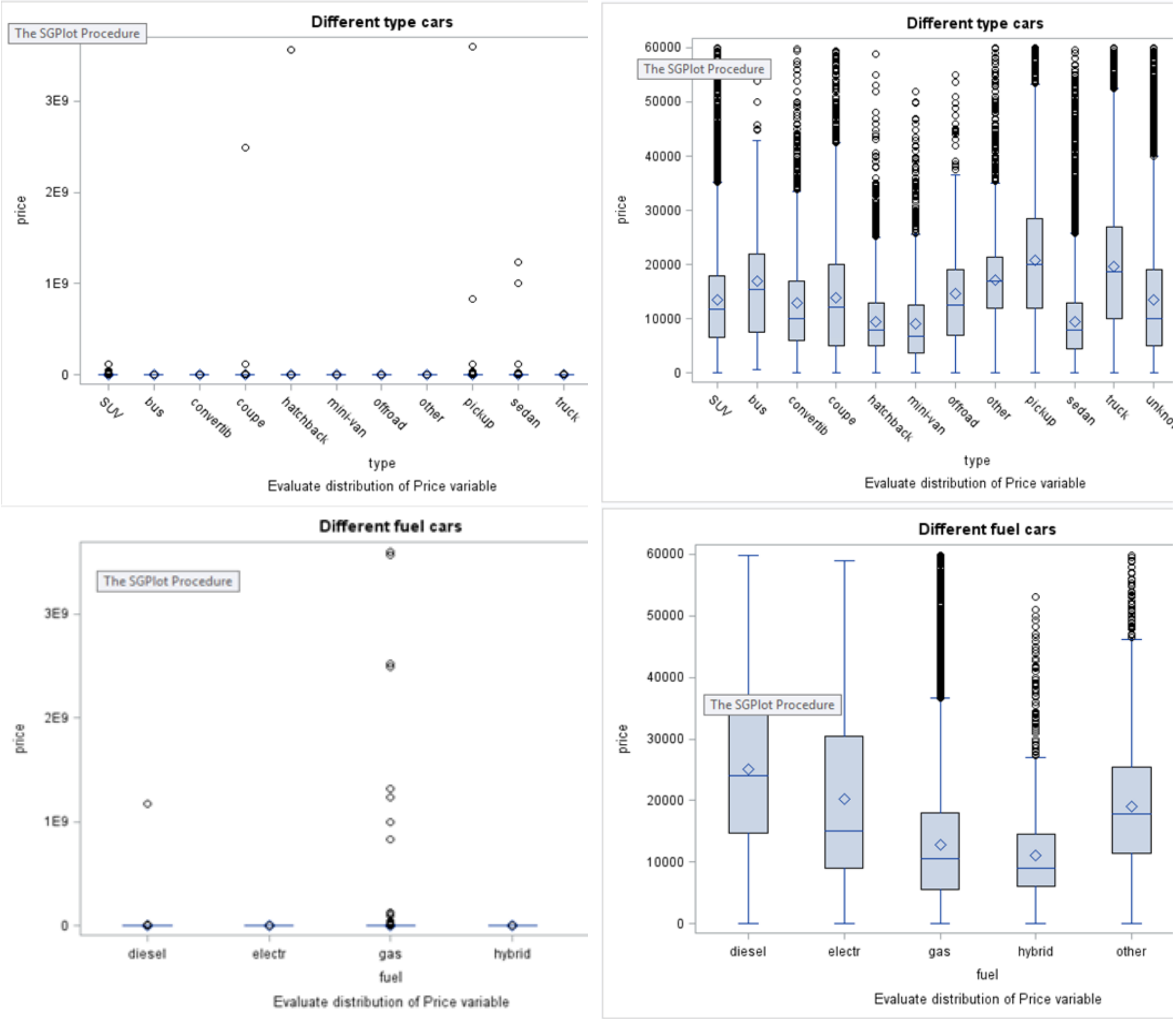
j. Lat - Latitude values not required for price prediction which shows the exact location of the listing. Another reason to remove this variable was there were many garbage values.
k. Long - Longitude values not required for price prediction which shows the exact location of the listing. Another reason to remove this variable was there were many garbage values.
l. Model - Due to there being several different types of models and their names changing throughout the years it made model values become insignificant for predicting price.

Next, we analyzed the number of missing values in each variable. Any variables containing null values more than 80% or garbage value and not making a significant contribution to the dependent variable were either removed or not used for running the regression model.

Below table shows the number of missing values we found in each variable that we considered after initial processing:

Variable name	**Number of missing values**	**Number of non-empty values**
manufacturer	23,584	516,175
condition	236,052	303,707
cylinders	218,945	321,264
fuel	3,393	536,366
title_status	2,940	536,819
transmission	3,973	535,786
drive	155,772	383,987
size	371,209	168,550
type	147,469	392,290
price	0	539,759
year	987	538,772
odometer	98,976	440,783

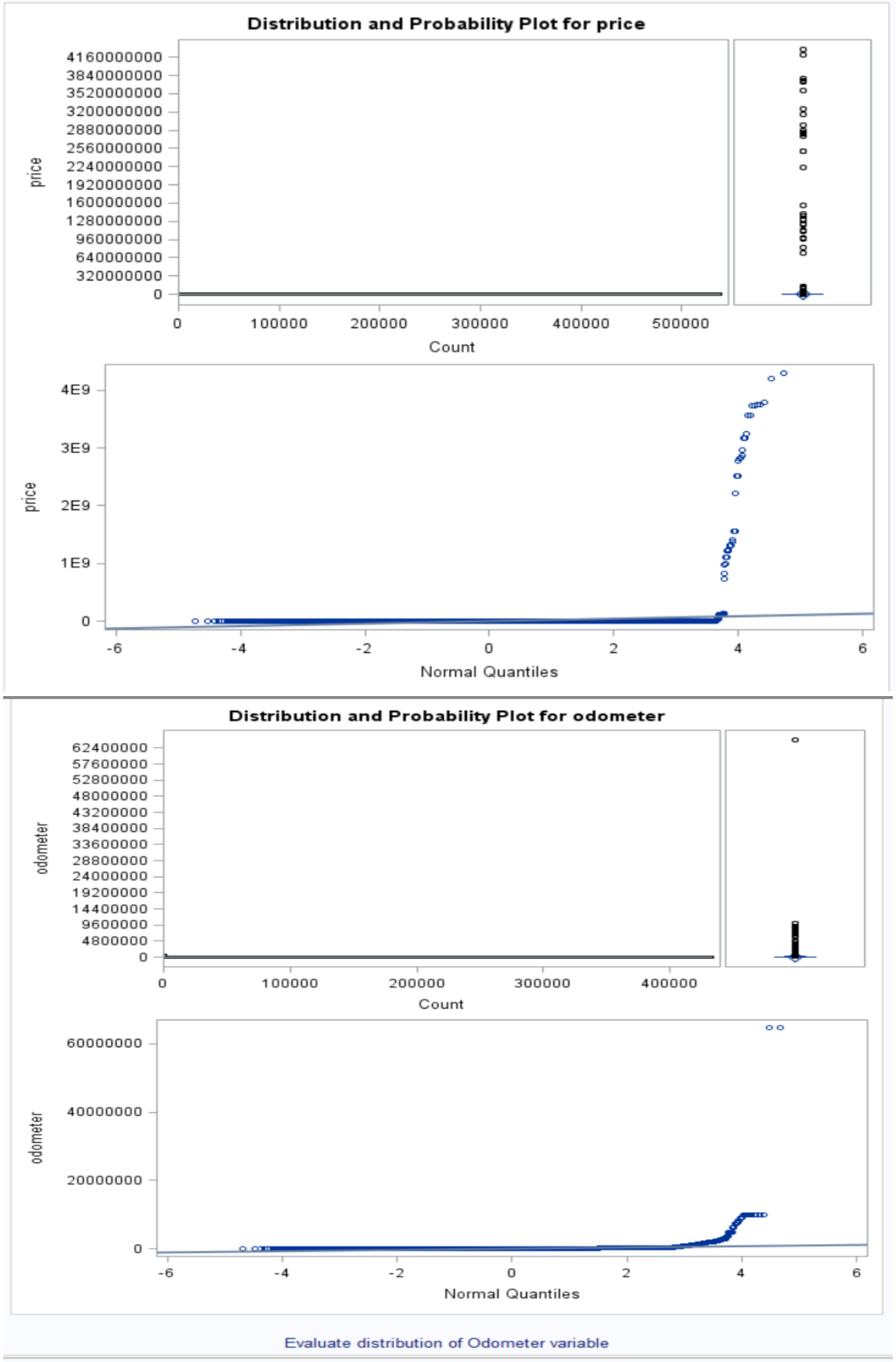
Price Distribution



In the above 4 plots you can see the before (left side) and after (right side) distribution of price based on the type of car (top plots) and distribution of price based on fuel type (bottom plots). As we can see there are a lot of unreasonably high prices(above 100k) from the box plot above. Hence, we removed the outlier during the data preprocessing step resulting in the even price distribution as shown in the box plots (Right side).

Of these variables there were many irrelevant values that we found with respect to price and year.

Plot to check the distribution of price variable

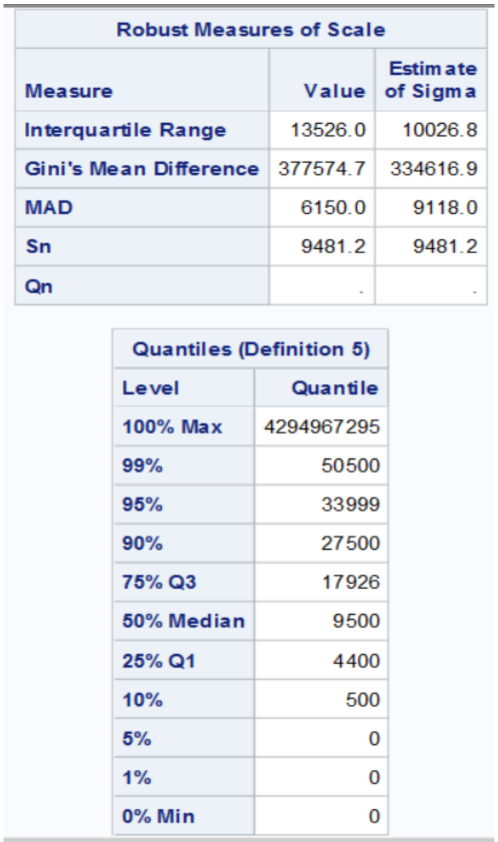


Plot to check distribution of odometer variable

As we can see, the price and the odometer variable has a lot many outliers thus barely being able to see the quartiles.

We can also see that the price and odometer had many values as 0 and so did the year with 2021. It made no sense that used cars have their price and odometer as 0 and year as 2021 in the future. Hence we removed these values.

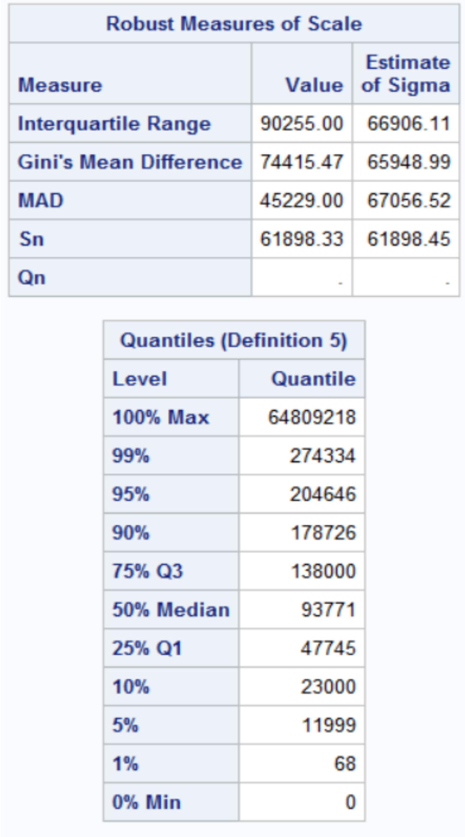
To add to that we also calculated what values can be considered as outliers based on the Interquartile Range and the First and the third quantiles for the variables price and odometer.



Robust Measures of Scale		
Measure	Value	Estimate of Sigma
Interquartile Range	13526.0	10026.8
Gini's Mean Difference	377574.7	334616.9
MAD	6150.0	9118.0
Sn	9481.2	9481.2
Qn	.	.

Quantiles (Definition 5)	
Level	Quantile
100% Max	4294967295
99%	50500
95%	33999
90%	27500
75% Q3	17926
50% Median	9500
25% Q1	4400
10%	500
5%	0
1%	0
0% Min	0

IQR for variable price



Robust Measures of Scale		
Measure	Value	Estimate of Sigma
Interquartile Range	90255.00	66906.11
Gini's Mean Difference	74415.47	65948.99
MAD	45229.00	67056.52
Sn	61898.33	61898.45
Qn	.	.

Quantiles (Definition 5)	
Level	Quantile
100% Max	64809218
99%	274334
95%	204646
90%	178726
75% Q3	138000
50% Median	93771
25% Q1	47745
10%	23000
5%	11999
1%	68
0% Min	0

IQR for variable odometer

We also calculated the outliers based on the values mentioned in the image above and decided the threshold to remove the outlier values would be the upper and lower outer fences. Based on that we considered choosing values as follows:

Price variable:

Upper outer fence = Q3 + 3 * IQR = 58,504

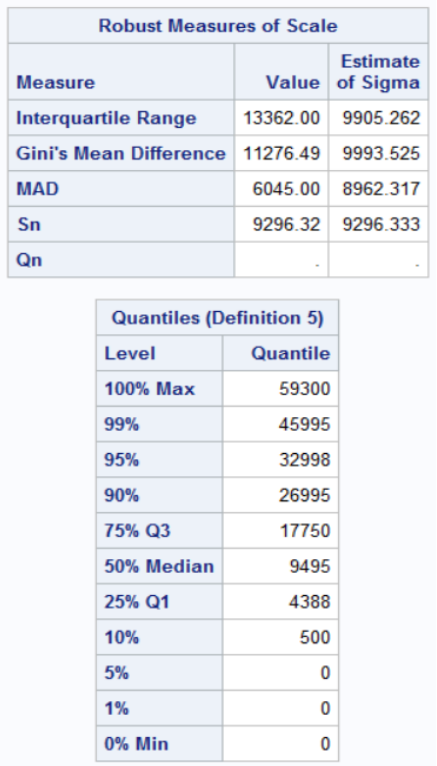
Lower outer Fence = Q1 - 3 * IQR = -36,178

As we can see the upper threshold that we decided was based on these values. This removed most of the values for price such as 4294967295 which can be seen in the 100% Max value in the plot above. These cannot be prices for a car and hence the calculation statistically helped us in removing such outliers.

Similarly although it does not make sense for a car to be sold at a price as less as $1 we have still kept it in the dataset since statistically this does not fall outside the lower outer fence which is negative.

Odometer variable:
Upper outer fence = Q3 + 3 * IQR = 408,765
Lower outer Fence = Q1 - 3 * IQR = -361,020

A single odometer was written at 60 million miles which is unheard of so seems like an inconsequential outlier. But after verifying statistically we can say indeed its an outlier (also anything beyond 408,765).



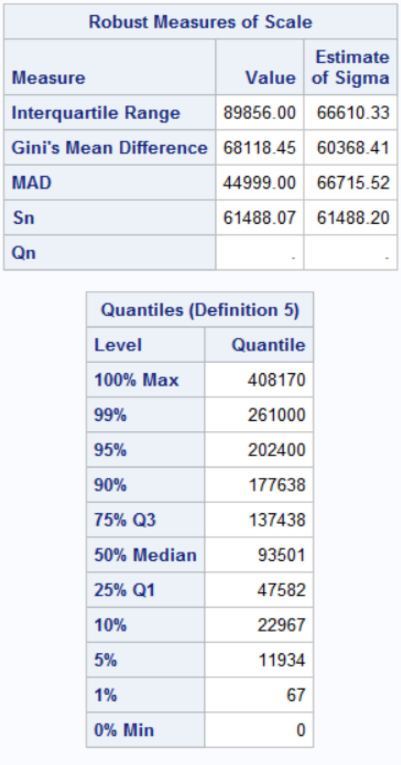
Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	13362.00	9905.262
Gini's Mean Difference	11276.49	9993.525
MAD	6045.00	8962.317
Sn	9296.32	9296.333
Qn	.	.

Quantiles (Definition 5)

Level	Quantile
100% Max	59300
99%	45995
95%	32998
90%	26995
75% Q3	17750
50% Median	9495
25% Q1	4388
10%	500
5%	0
1%	0
0% Min	0

IQR for price after cleaning



Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	89856.00	66610.33
Gini's Mean Difference	68118.45	60368.41
MAD	44999.00	66715.52
Sn	61488.07	61488.20
Qn	.	.

Quantiles (Definition 5)

Level	Quantile
100% Max	408170
99%	261000
95%	202400
90%	177638
75% Q3	137438
50% Median	93501
25% Q1	47582
10%	22967
5%	11934
1%	67
0% Min	0

IQR for odometer after cleaning

For some other variables we also removed garbage values such as “missi” and “parts only” for *title_status*. We also removed the missing records for the variables *year* and *manufacturer* as the number of missing records were less than 5% of the overall number of records and also did not contribute much to the model.

Imputation:-

There were many missing values in most of the variables that were needed in the regression model.

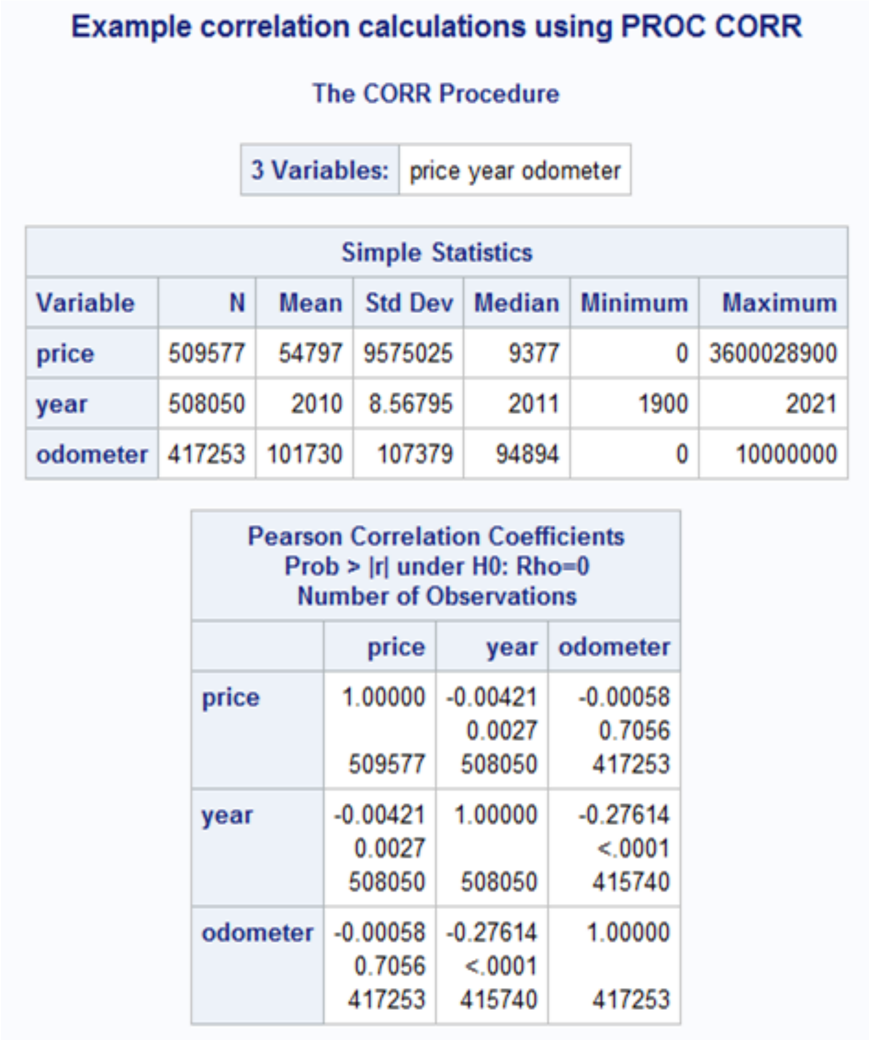
For example the variable *drive* almost had 30% missing values. Since this is a categorical variable we assigned weights to each category depending on the price of the vehicle. So higher the cost higher the weights were assigned to *4wd, rwd and fwd* cars respectively. Also since the mode in this case was *rwd* we imputed the missing values with the weight of *rwd*.

For the variables *cylinders, fuel, title_status* and *type* we imputed the missing values with *unknown* keyword.

Like *drive* variable we imputed the mode for *transmission* variable.

Next step was going back through the columns and rechecking to make sure the columns were still valuable to our analysis and not containing too many missing data points. This is when we had to remove the cylinder since there were only a few data points with over four hundred thousand rows left.

After cleaning the data extensively we are left with 17 variables of which 5 were added in by us for calculations. In total we were left with 389,326 observations, so 150,433 observations were dropped.



Example correlation calculations using PROC CORR

The CORR Procedure

3 Variables:	price year odometer

Simple Statistics						
Variable	N	Mean	Std Dev	Median	Minimum	Maximum
price	509577	54797	9575025	9377	0	3600028900
year	508050	2010	8.56795	2011	1900	2021
odometer	417253	101730	107379	94894	0	10000000

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

	price	year	odometer
price	1.00000 509577	-0.00421 0.0027 508050	-0.00058 0.7056 417253
year	-0.00421 0.0027 508050	1.00000 508050	-0.27614 <.0001 415740
odometer	-0.00058 0.7056 417253	-0.27614 <.0001 415740	1.00000 417253

We begin with a correlation table (above) so we can view what values have significant meaning to another. This is the first step in evaluating how to predict price for cars. We can see that price and year aren't well correlated since their correlation coefficient is below .005. We can also see that odometer is correlated with price since theirs were .7 which means they have correlation and should be evaluated further.

5. Empirical Analysis

There are quite a few predictive modeling algorithms to choose from but based on our requirement we choose Linear Regression. Using Regression, we were able to identify the relationship between Price of the car with other variables like year, condition, manufacturer, etc.

We predicted the price of used cars in the linear regression model. We used the 'glmselect' procedure as the dataset contains both categorical and numerical features which are used in predicting price of used cars.

There are a lot of missing values in the data that needs to be imputed. After analyzing the missing values, there is no pattern in the missing data. Hence, the missing values are considered as another category and imputed with "unknown" for variables – cylinder, fuel, title_status and type. For variables – transmission and drive, the missing values are imputed with their respective modes.

Linear Regression:

Since the objective here is predicting the price of a car, we are working on the Regression problem.
We performed four types of regression:

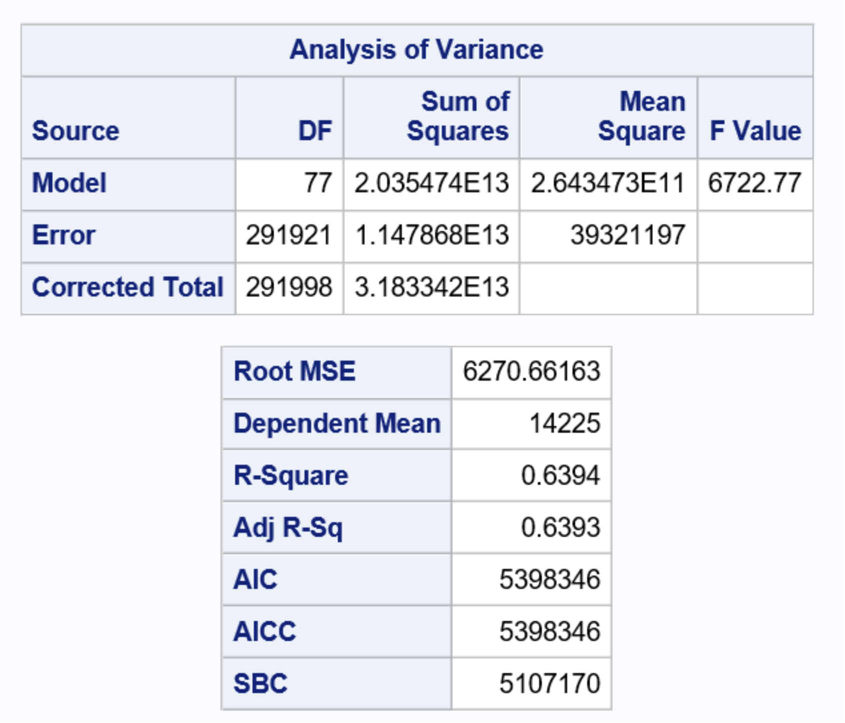
1. Base Model – includes all the features at once
2. Forward Selection – starts with null model and adds features that are significant
3. Backward Selection – starts with base model and drops the features that are not significant
4. Stepwise Regression- modified version of forward selection but checks the significance of features already in the model at every step.

Model Selection:

The best model is selected based on the lowest AIC value after the regression is performed.

Model	AIC	R-squared	Adjusted R-squared
Base Model	5398346	0.6394	0.6393
Forward Selection	5284316	0. 7584	0.7571
Backward Selection	5284316	0.7584	0.7571
Stepwise Regression	5284316	0.7584	0.7571

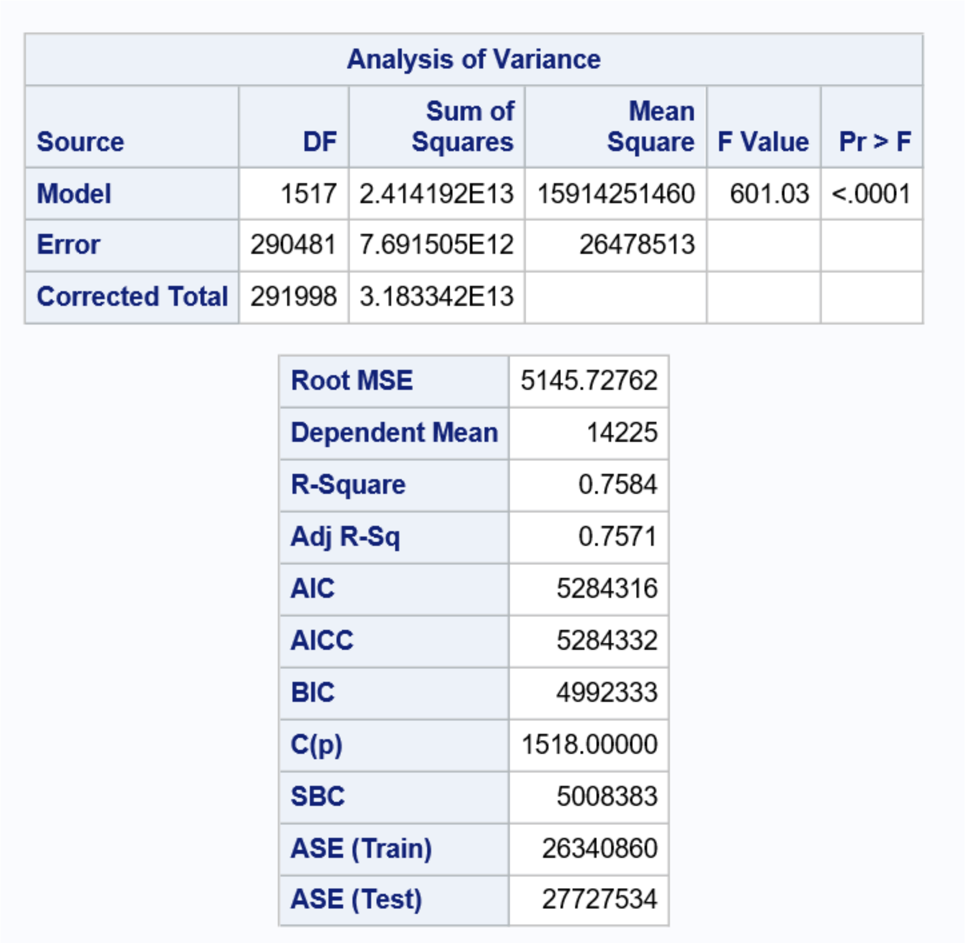
From the results above, we can observe that feature selection models – forward, backward or stepwise regression performed better than base models.



Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	77	2.035474E13	2.643473E11	6722.77
Error	291921	1.147868E13	39321197	
Corrected Total	291998	3.183342E13		

Root MSE	6270.66163
Dependent Mean	14225
R-Square	0.6394
Adj R-Sq	0.6393
AIC	5398346
AICC	5398346
SBC	5107170

Base model



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1517	2.414192E13	15914251460	601.03	<.0001
Error	290481	7.691505E12	26478513		
Corrected Total	291998	3.183342E13			

Root MSE	5145.72762
Dependent Mean	14225
R-Square	0.7584
Adj R-Sq	0.7571
AIC	5284316
AICC	5284332
BIC	4992333
C(p)	1518.00000
SBC	5008383
ASE (Train)	26340860
ASE (Test)	27727534

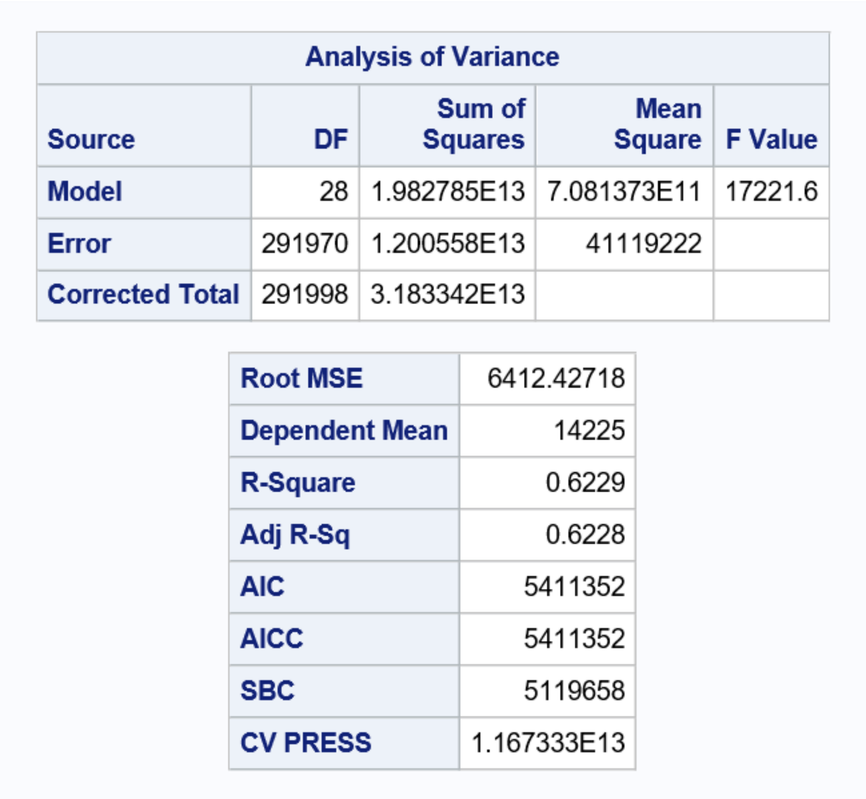
Forward, Backward and Stepwise Model

Additionally, we also implemented regularization models – Ridge, Lasso and Elastic Net. In most cases, regularization doesn't necessarily improve the performance on the data set that the algorithm used to learn the model parameters. Regularization is typically used for ill-posed problems; wherein additional information is used to constrain the solution set. However, it can improve the generalization performance thereby avoiding overfitting since we have a lot of features.

Below are results for regularization models:

Regularization	AIC	R-squared	Adjusted R-square
Ridge	5411352	0.6229	0.6228
Lasso	5398756	0.6389	0.6388
Elastic Net	5398585	0.6391	0.6390

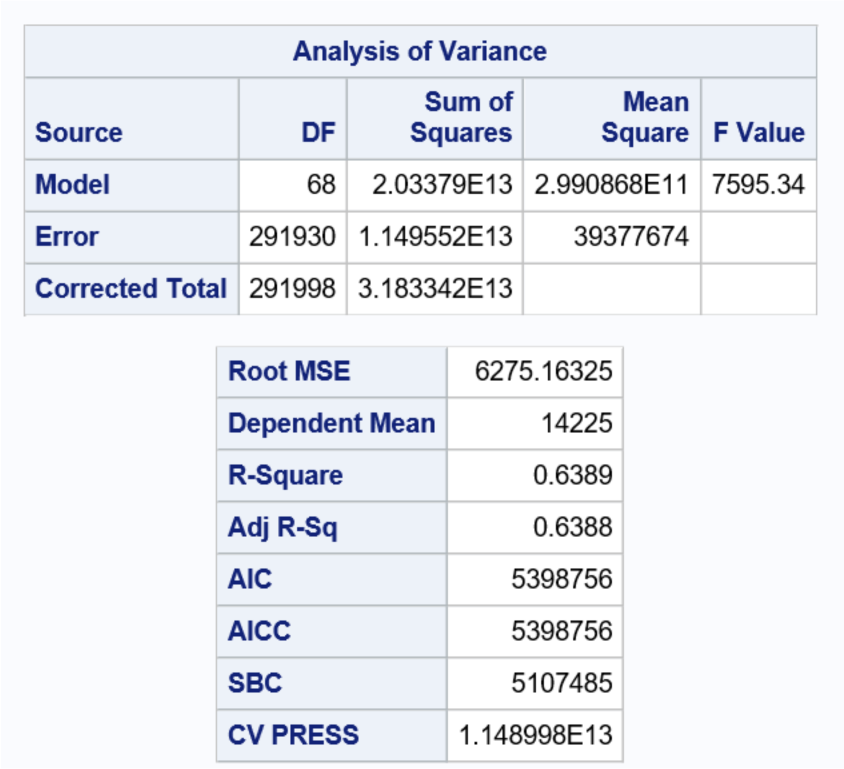
We've examined several possible models of linear regression. In our model and given the number of parameters exhibited in our observations, it's impossible to consider other regression models.



Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	28	1.982785E13	7.081373E11	17221.6
Error	291970	1.200558E13	41119222	
Corrected Total	291998	3.183342E13		

Root MSE	6412.42718
Dependent Mean	14225
R-Square	0.6229
Adj R-Sq	0.6228
AIC	5411352
AICC	5411352
SBC	5119658
CV PRESS	1.167333E13

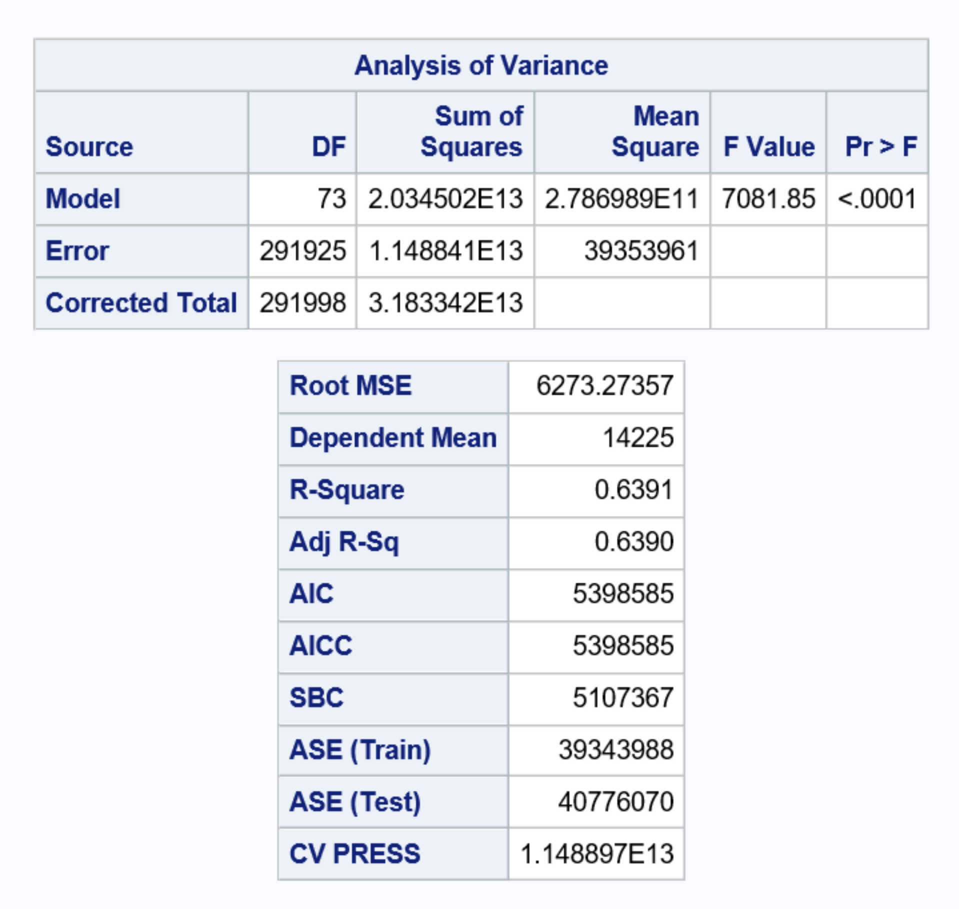
Ridge Model



Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	68	2.03379E13	2.990868E11	7595.34
Error	291930	1.149552E13	39377674	
Corrected Total	291998	3.183342E13		

Root MSE	6275.16325
Dependent Mean	14225
R-Square	0.6389
Adj R-Sq	0.6388
AIC	5398756
AICC	5398756
SBC	5107485
CV PRESS	1.148998E13

Lasso Model



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	73	2.034502E13	2.786989E11	7081.85	<.0001
Error	291925	1.148841E13	39353961		
Corrected Total	291998	3.183342E13			

Root MSE	6273.27357
Dependent Mean	14225
R-Square	0.6391
Adj R-Sq	0.6390
AIC	5398585
AICC	5398585
SBC	5107367
ASE (Train)	39343988
ASE (Test)	40776070
CV PRESS	1.148897E13

Elastic Net Model

However, in several models, we do include interactions of variables, e.g., age*odometer, as our regressors. This greatly increases the flexibility of the model and provides sufficient explanatory power in all the variability in the overall observations (or more specifically the “cleaned” data).

The model which only includes the linear effect of all the variables gives explanatory power, with an adjusted R squared value of 0.6394. This is already a well-established result. The benefits of this model are twofold. It gives us a simple enough model to operate (providing us a benchmark with good enough performance) and also delivers clear-cut messages on how the variables influence the price of a car. For example, if we want to know what’s the marginal contribution of age on the price (apparently negative) without knowing any more detailed information about the car, this model gives us -458.24, and it is statistically significant. The model, in some sense, gives us a list of “rule of thumbs”, where we can evaluate the influence of one single variable’s variation on the predicted price of the car.

Another important variable is “manufacturer”, where the variable takes on value of 1 when the car is manufactured by that specific manufacturer and zero otherwise. So it’s basically shifting upwards and downwards when concerning different manufacturer's cars' predicted prices in a hyperdimensional plane. So roughly speaking, those with negative slope parameters are, in some sense, below (weighted) average and those with positive slope parameters are above average. For example, Hyundai, Harley-dav and mazda are below average and lexus, Toyota and Cadillac are above average. This, in some sense, gives us a sense on which car brand is more premium and which one is cheaper.

There are two things to be noted as well:

1) Because we have limited data, it's possible that the observations are biased and not representative enough. For example, Lincoln and rand rover are also below average, which is a little bit counterintuitive. But if we are restricting our attention at the second-hand market, it's possible that people sell those cars only when they are in bad shape whereas people sell other cars when they don't feel like having them anymore (thus, more arbitrary and random). So it's possible that we find the average age of Lincoln is higher than other car brands and thus, explains our findings. But this still validates our prediction, in the sense that if a car is delivered to the second-hand market, we know that some car brands will have cheaper predicted prices whereas others don't. All we have to do is to restrict our attention to the case where our datasets are representative enough to make credible inference.

2) Some parameters are not statistically significant. These parameters should be treated with care when we are interested in those parameters' changes' influences on the predicted price.

We also have several other models, which give us as high an R squared as almost 0.75. These models are somewhat more difficult to explain as in their slope parameters, because they are including interaction terms. However, because of their high reliability as in prediction (for example, in one model, we use 80% of the data for training the model and the other 20% to test the model, and the resulted prediction is more reliable than the model we have talked about above)

For example, we own a dealer shop and we want to deliver reliable (predicted) car prices to the potential seller. With the model, we can give both point estimation and interval estimation with high accuracy. This is a good business practice. Although we have all the parameters' estimations, we are not interested in their marginal effects, but rather focus on the prediction part of the model. Put it otherwise, it's somewhat like a "black box", but still provides great tractability (unlike machine learning or deep learning). Such model is suitable for implementation in the small dealership business of cars.

CONCLUSION

The used cars database from Craigslist was analyzed by our group to determine the factors that played the largest role in the price of a car. Our conclusion was that 12 variables played a major role in deciding what a car’s price would be; manufacturer, condition, cylinders, fuel, title_status, transmission, driver, size, type, price, year, and odometer.

Appendix

Code:

```
dm 'clear log'; dm 'clear output';  /* clear log and output */

libname vehicles "E:\Users\sck160030\Documents\My SAS
Files\9.4\SAS_Data_Project1";
title;

proc import datafile="E:\Users\sck160030\Documents\My SAS Files\9.4\vehicles.csv"
        dbms=csv
        out=work.vehicles3
        replace;
        getnames=yes;
        datarow=2;
        run;
proc contents data=work.vehicles3;
        run;
proc sql;
        create table vehicles as select * from work.vehicles3;
        run;
        quit;
DATA data3(DROP = url region_url vin county description id image_url lat long
description region model size);
        SET work.vehicles3;
        run;
/*Frequency*/
proc freq data = data3;
        tables fuel title_status transmission manufacturer condition cylinders drive type
paint_color /missing;
        run;
/*Corellation metrix*/
proc corr data=data3;
        run;
proc corr data=data3 pearson spearman;
title 'Example correlation calculations using PROC CORR';
        run;
/* categorical variables frequency */
proc freq data = data3;
        tables fuel title_status transmission manufacturer condition cylinders drive type;
        run;
/*Data preprocessing*/
data data3;
set data3;
        if price = 0 then delete;
        run;
data data3;
set data3;
        if price > 59900 then delete;
        run;
data data3;
set data3;
        if year = '.' then delete;
        run;
data data3;
set data3;
        if year = 1900 then delete;
        run;
data data3;
set data3;
```

```
        if year = " " then delete;
        Run;

data data3;
set data3;
        if odometer > 408765 then delete;
        run;
data data3;
set data3;
        if odometer = '.' then delete;
        run;
data data3;
set data3;
        if odometer ne . then do; odometer = 1000000-odometer; end;
        Run;
data data3;
set data3;
        if title_status = "parts" then delete;
        Run;
data data3;
set data3;
        if title_status = "missi" then delete;
        Run;
data data3;
set data3;
        if manufacturer = " " then delete;
        Run;

/* find missing data */
proc format;
 value $missfmt ' '='Missing' other='Not Missing';
 value  missfmt  . ='Missing' other='Not Missing';
run;
proc freq data=work.vehicles;
format _CHAR_ $missfmt.; /* apply format for the duration of this PROC */
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;
```

```
/* Imputing condition of the car with odometer rating */
proc summary data=data3 QNTLDEF=3 qmethod=OS;
        var odometer;
        output out=summary MIN=p1= p25=p50= p75=MAX= / autoname;
        run;
proc print data=summary;
        run;
data data3;
set data3;
        if odometer = 0 then delete;
        run;
data data3;
set data3;
        if odometer < 50000 then condition2='excellent';
        else if odometer > 50000 and odometer < 95515 then condition2='good';
        else if odometer > 95514 and odometer < 139584 then condition2='fair';
        else condition2 ='salvage';
        run;
data data3;
        set data3;
        age = 2021-year;
        drop year;
        run;
data data3;
set data3;
        if age = 0 then delete;
        run;

/* Implemented Mode for transmission */
data data3;
set data3;
        if transmission = 'automatic' then transmission=1;
        else if transmission = 'manual' then transmission=2;
        else transmission= 1;
        run;

proc print data= data3(obs=10);
        run;
/*Histogram Plot for normal and log values*/
proc univariate data = data3 plots;
```

```
          var price age odometer;
          histogram; inset n mean std min max;
          run;
data data3;
          set data3;
          log_price = log(price);
          log_age = log(age);
          log_odometer = log(odometer);
          run;
proc univariate data = data3;
          var log_price log_age log_odometer;
          histogram; inset n mean std min max;
          run;
/*Importing Unknown for Blanks */
data data3;
set data3;
          if cylinders = '' then cylinders='unknown';
          if fuel = '' then fuel='unknown';
          if title_status = '' then title_status='unknown';
          if type = '' then type='unknown';
          run;

proc print data=data3 (obs=20);run;

proc contents data=data3; run;

/*Assigning weights and imputing mode as rwd*/
data data3;
set data3;
          if drive = '4wd' then drive=30;
          else if drive = 'rwd' then drive=20;
          else if drive = 'fwd' then drive=10;
          else drive= 20;
          run;
proc print data= data3(obs=100);
          run;

proc print data= data(obs=10);
          run;
proc print data= data3(obs=10);
```

```
        run;
/*Regression*/
proc glmselect data=data3 outdesign=data4_li;
        class manufacturer fuel transmission type cylinders condition2 title_status drive ;
        linear: model price = manufacturer fuel transmission cylinders type condition2
title_status drive odometer age;
        title "Linear Regression";
        run;
/*Random Sample*/
proc surveyselect data=data3 out=sample_data method=srs sampsize=10000
seed=987654321;
        run;
/* Create training and test datasets 80% of sample in train and 20 in test  */
proc surveyselect data=sample_data out=hwdata_sampled outall samprate=0.8
seed=2;
        run;
data train_data test_data;
 set hwdata_sampled;
        if selected then output train_data;
        else output test_data;
        run;

/* Forward------------ */
proc glmselect data=train_data testdata=test_data  plots=all;
        class manufacturer fuel transmission type cylinders condition2 title_status drive;
        model price = age|odometer| manufacturer |fuel| transmission| drive | type|
title_status| cylinders| condition2 @2
         /selection=forward(select=cp) hierarchy=single showpvalues ;
        performance buildsscp=incremental;
        run;

/* Backward------------ */
proc glmselect data=train_data testdata=test_data  plots=all;
        class manufacturer fuel transmission type cylinders condition2 title_status drive;
        model price = age|odometer| manufacturer |fuel| transmission| drive | type|
title_status| cylinders| condition2 @2
        /selection=backward(select=cp) hierarchy=single showpvalues ;
        performance buildsscp=incremental;
        run;
```

```
/* Stepwise------------ */
proc glmselect data=train_data testdata=test_data  plots=all;
        class manufacturer fuel transmission type cylinders condition2 title_status drive;
        model price = age|odometer| manufacturer |fuel| transmission| drive | type|
title_status| cylinders| condition2 @2
        /selection=stepwise(select=cp) hierarchy=single showpvalues;
        performance buildsscp=incremental;
        run;

/* Ridge ---------- */
proc glmselect data=data3;
        class manufacturer fuel transmission type cylinders condition2 title_status drive;
        model price = manufacturer fuel transmission cylinders type condition2
title_status drive odometer age / selection=ridge(choose=cv l1=0 stop=cv);
        score data=data3 p out=ridge;
        title "Ridge";
        run;
/* Lasso ---------- */
proc glmselect data=data3;
        class manufacturer fuel transmission type cylinders condition2 title_status drive;
        model price = manufacturer fuel transmission cylinders type condition2
title_status drive odometer age / selection=lasso(choose=cv stop=cv);
        score data=data3 p out=ridge;
        title "Lasso";
        run;
/* Elastic Net -------- */
  proc glmselect data=data3 testdata=test_data seed=2 plots=all;
 class manufacturer fuel transmission type cylinders condition2 title_status drive;
 model price = manufacturer fuel transmission cylinders type condition2 title_status
drive odometer age @2 /selection=elasticnet(choose=cv stop=none)
 hierarchy=single cvmethod=random(10) showpvalues ; performance
buildsscp=incremental; run;
proc univariate data = data3;
        var log_price log_age;
        histogram / normal kernel;
        qqplot / normal(mu=est sigma=est);
        inset n mean std;
        run;

/* price distribution by categories */
```

```
proc univariate data = data3; class age; var log_price; histogram; run;
proc univariate data = data3; class condition; var log_price; histogram; run;
proc univariate data = data3; class cylinders; var log_price; histogram; run;
proc univariate data = data3; class drive; var log_price; histogram; run;
proc univariate data = data3; class fuel; var log_price; histogram; run;
proc univariate data = data3; class manufacturer; var log_price; histogram; run;
proc univariate data = data3; class paint_color; var log_price; histogram; run;
proc univariate data = data3; class size; var log_price; histogram; run;
proc univariate data = data3; class title_status; var log_price; histogram; run;
proc univariate data = data3; class transmission; var log_price; histogram; run;
proc univariate data = data3; class type; var log_price; histogram; run;
```