USED CARS AND TRUCKS SALES ON CRAIG'S LIST

PREDICTIVE ANALYTICS

Ajinkya Ponkshe, Mohammed Kuzhimbadath, Siqi Tang, Ziyang Feng

PROJECT – GROUP 3

7/21/2021

Table of Contents

Executive Summary

Craigslist is one of the world's largest collection of used vehicles for sale. Used cars market has been one of the top priorities in the United States. This market has helped a lot of students and Middle-class people to afford cars of their choice at the same time opening a lot of options to select from. The demand for used cars are increasing at a high pace and its necessary to understand various factors and variables that will affect the buyer power for a customer. Through our analysis we want to find interesting trends and correlations between the various features of the sale of the used cars. This project would be relevant for used car business especially in the United States for the seller as well as the buyer. Figuring out the various relationship between the prices and other variables help the manufacturer to better understand their standing in the current market. They would be able to analyze the results on what kind of cars are selling more for them in the market and what features are important for the customers regarding the quality and preferences while selecting a car. Accordingly, they can make changes in their production lines or features which will help them to increase the sales of their specific brands and thus increase their brand market.

Data Description

The data before cleaning has 426880 entries and 26 variables. Data set provides us with the information on cars available in the used car market in respect to its various features such as Price of the car, the miles covered , manufacturer ,model, transmission type, the state where it is located etc. This data set also covers other basic information on the car such as the condition or title of the car , whether it is an excellent condition or good condition or if there are any salvages. The year and odometer values are also available to inform the customer how old or new the car has been into the market. This data set has all the basic and detailed information of the entire car so that it makes it easy to understand the car well and decide between various choices before purchasing the car. The variables and description of them are as below:

Id – the unique number for which each cars information are listed.

URL : the web link at which information about the vehicle is available on craigslist.

Region : Specific region of the state where the car is available

Price : Selling price of the car listed.

Year : The year at which vehicle is manufactured

Manufacturer : The company that has produced the car.

Model : The model of the car brand

Condition : Is the car in good or bad condition

Cylinders: No. of cylinders in the car

Fuel: Kind of fuel the car is using

Odometer : The miles car has travelled.

Transmission : The gear system in the car

VIN : The VIN number of the car

Drive : Front or rear or four-wheel drive

Size : Type of the car size – compact, sub-compact, Mid-size etc.

Type : The type of car – hatch back, sedan, SUV etc.

Paint color : The color of the car.

image URL : The web link at car's image is available.

Description : Describes about the car details.

County : The county at which car is available.

State: The state of the country at which car is available.

Lat : Latitude

Long : Longitude

Posting date : The date at which car is posted for selling.

Objective

Our main objective of this project is to identify the major trends in the sales of cars and truck in the American used car industry. We will be working on identifying relations between price and various variables to get a more detailed idea on how the prices are dependent on various variables. The process will start with looking at the data carefully regarding the cars manufacturer, year , odometer reading etc. Then we will make initial analysis to figure out outliers which are not significant to our data models. We are using to use ANOVA test, Linear and multiple regression models, diagnostic tests etc. to test the null hypothesis and to find out the trend between price and other independent variable. During this project we would working towards rejecting the null hypothesis we have come up from our analysis based on various models.

Following are the null hypotheses we will test using various statistical methods.

- Higher odometer value cars are not less costly.
- Used cars manufactured in latest years are less costly.
- Title of cars does not have an impact on the price of the cars.
- Condition of used cars does not have impact on price of the cars.

Data Pre-Processing / Data Cleaning

The data set contains a lot of missing values and variables that have no significance or logically dependence on our analysis. It is important that we do necessary data cleaning steps to transform the data set into features and variables that are really going to attract our attention in the analysis. We decided to follow the following steps to clean the data before we start with the actual analysis:

- *Dropping the unnecessary variables from the data set*

After the initial review of the data set, we concluded that variables such as id, url, size, region_url, VIN, paint_color, image_url, description, county, lat, long, posting date are not useful for our price analysis and thus we dropped them using the PROC SQL AND DROP command. Especially, the county variable has no values in any of its entries making it of no use to our analysis.

- *Removing the rows where price=0*

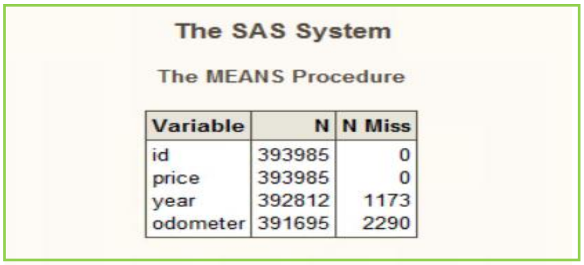
We know that the main purpose of using craigslist is to buy or sell a car that has been used for a particular amount of time. When prices are zero , it means those rows does not fit in the business scenario as we know a seller would not sell a car for $0.

- *Removing the records where all the variables are blank*

We figured out from the data set that there are number of rows where the values are entirely missing for all the variables. These rows with missing values do not add any value to analysis so it would be appropriate to drop them from the analysis.

- *Finding the missing values for numeric data*

We used the proc means and nmiss to find out the missing values in the numeric data variables.

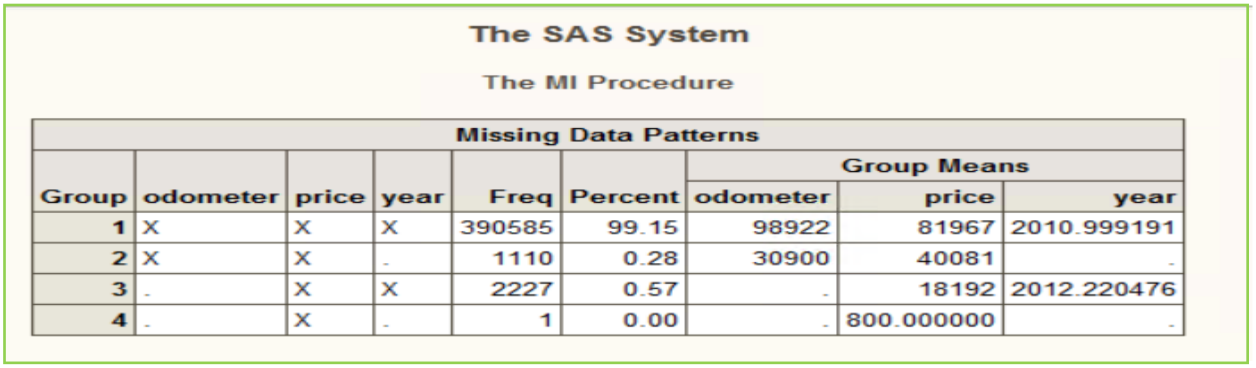


The SAS System

The MEANS Procedure

Variable	N	N Miss
id	393985	0
price	393985	0
year	392812	1173
odometer	391695	2290

- *Showing the trend of missing numeric data between the variable we found out before.*



The SAS System

The MI Procedure

Missing Data Patterns

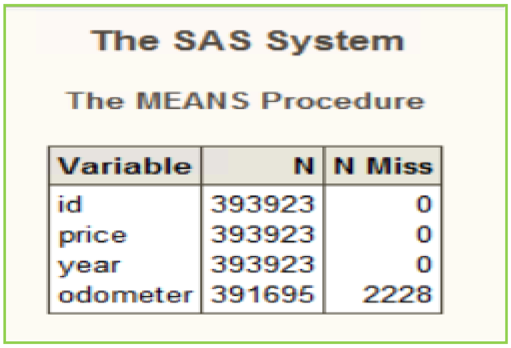
						Group Means		
Group	odometer	price	year	Freq	Percent	odometer	price	year
1	X	X	X	390585	99.15	98922	81967	2010.999191
2	X	X	.	1110	0.28	30900	40081	.
3	.	X	X	2227	0.57	.	18192	2012.220476
4	.	X	.	1	0.00	.	800.000000	.

The above table gives combinations of numerical variables and missing records. As we can see we have three major numerical variables namely odometer, price and year. There are 390585 records where the values for all the three numerical variables are present but there are 3338 records where at least one of the values is missing.

- *Imputing the year variable with median of year*

We used the impute with median method to fill the missing values of year. The years ranging in the dataset are from early 1900 to 2020. Thus, considering categorical nature of the variable we decided to use median to impute the missing values.

- *Checking again if the number of missing values of year is zero after imputing.*



The SAS System

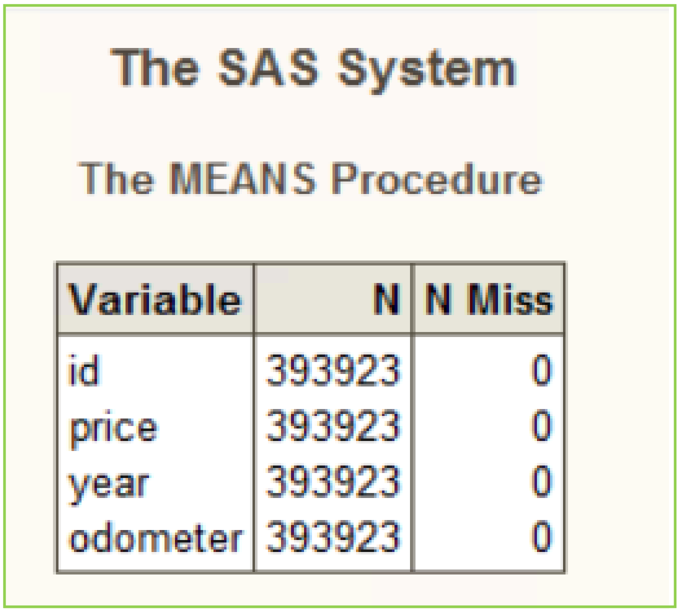
The MEANS Procedure

Variable	N	N Miss
id	393923	0
price	393923	0
year	393923	0
odometer	391695	2228

- *Imputing the values of odometer with mean of odometer grouped by year.*

We know the variable odometer and year are related. When the year increases, the odometer, value is expected to increase as the car might have run for a long time. So, we decided to impute the missing values of Odometer with mean of odometer grouped by year.

- *Checking if all the numeric values are imputed*



The SAS System

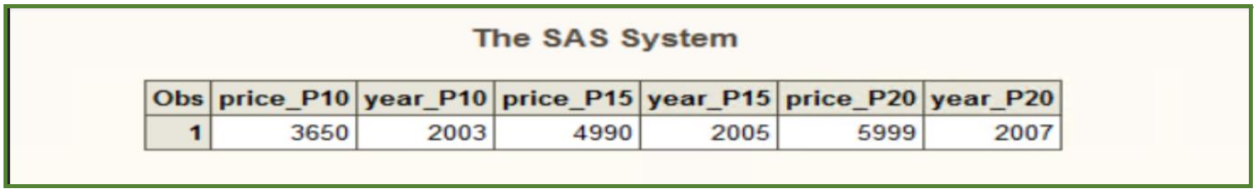
The MEANS Procedure

Variable	N	N Miss
id	393923	0
price	393923	0
year	393923	0
odometer	393923	0

- *Removing the outliers*

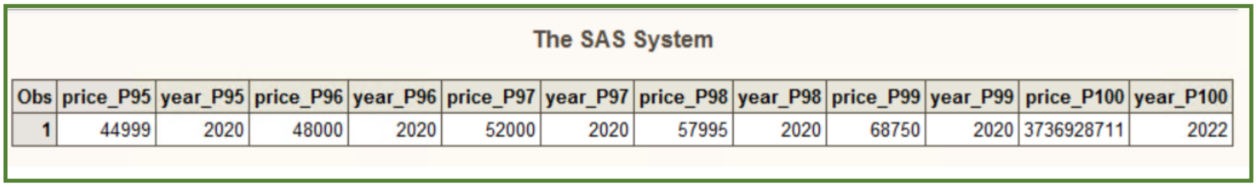
We found out the lower and higher quantile percentages for year and price to figure out the outliers and reduced our data set to range of price from $3000 to $70000 and year from 2003 to 2021.

After calculating the quantile ranges, to remove the outliers, we had decided to confine the range of year to 2003 to 2021 because we believe most of the customers are not likely to look at a car that is too old as before 2000



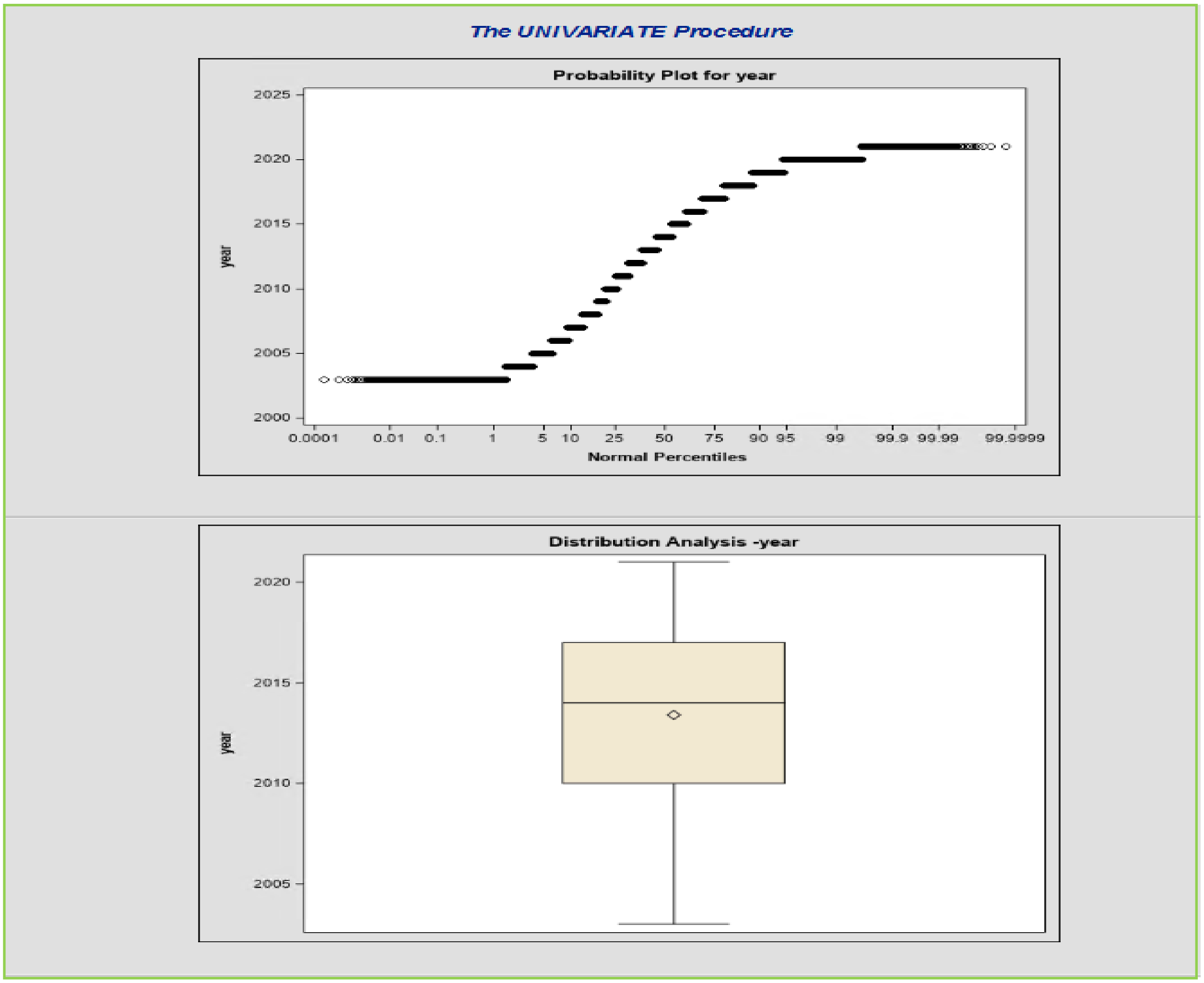
The SAS System

Obs	price_P10	year_P10	price_P15	year_P15	price_P20	year_P20
1	3650	2003	4990	2005	5999	2007



The SAS System

Obs	price_P95	year_P95	price_P96	year_P96	price_P97	year_P97	price_P98	year_P98	price_P99	year_P99	price_P100	year_P100
1	44999	2020	48000	2020	52000	2020	57995	2020	68750	2020	3736928711	2022



The UNIVARIATE Procedure

- *Imputing categorical variables*

We had some missing values in the manufacturer variable and could not find a relationship with any other variables. So, we decided to replace the missing values with "missing" as there is not information available for customers on the manufacturer of the car.

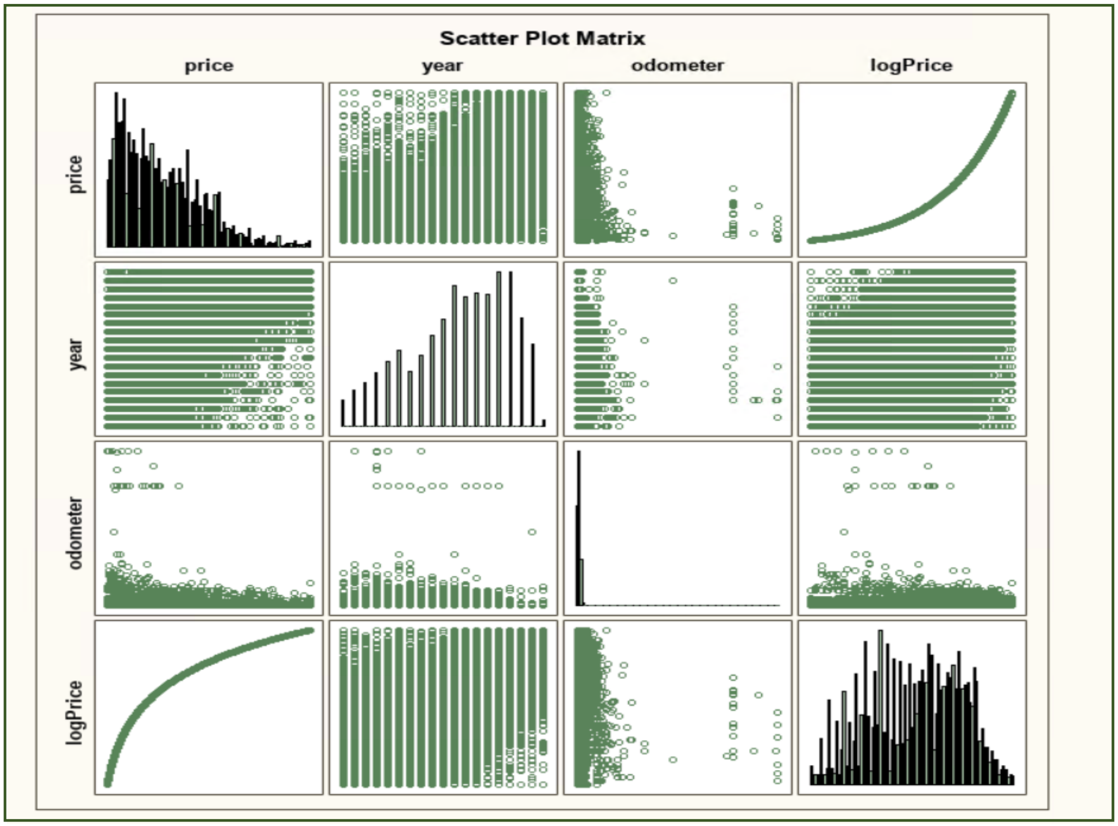
For variables such as condition, cylinders, fuel, transmission, drive, type we decided to replace the missing values with the most occurring values in each variable or the dominating values in each variable.

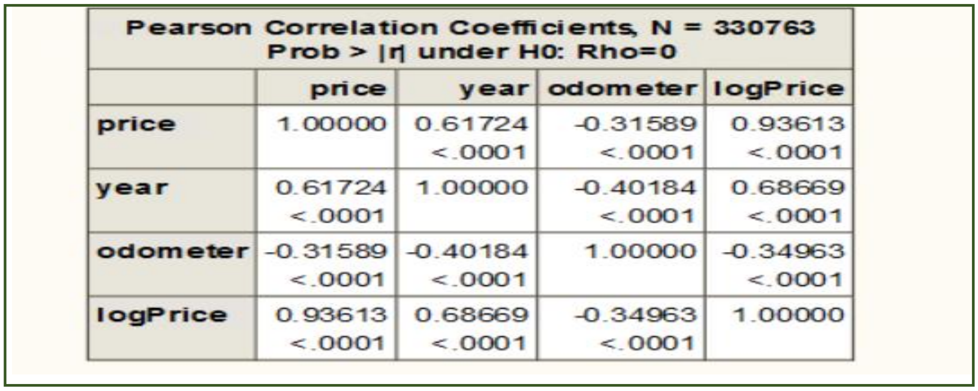
We used mode imputation (using proc SQL) to impute missing values and impute data for the categorical variables.

Exploratory Data Analysis

We used the exploratory data analysis to visualize the relationship of different variables with each other and their relationship with price of a used car.

Exploratory Analysis: Used Car Data



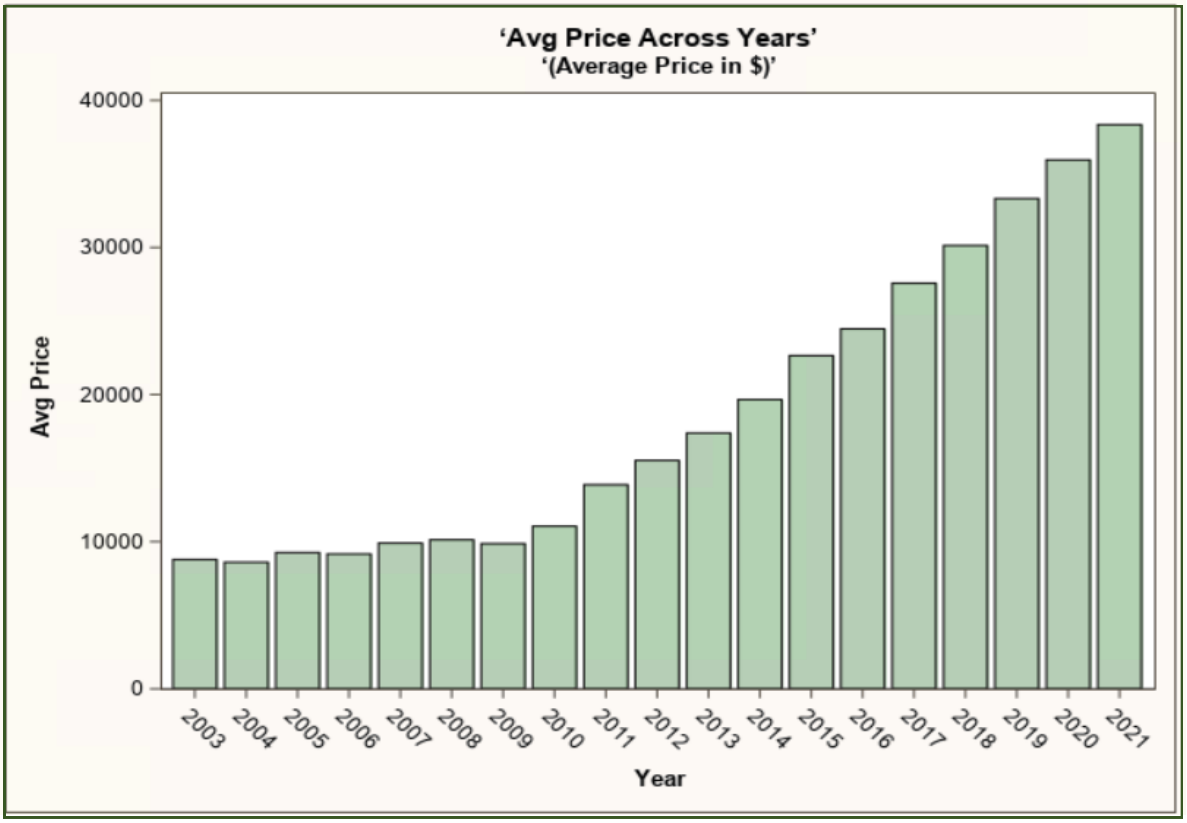


Pearson Correlation Coefficients, N = 330763 Prob > |r| under H0: Rho=0				
	price	year	odometer	logPrice
price	1.00000	0.61724 <.0001	-0.31589 <.0001	0.93613 <.0001
year	0.61724 <.0001	1.00000	-0.40184 <.0001	0.68669 <.0001
odometer	-0.31589 <.0001	-0.40184 <.0001	1.00000	-0.34963 <.0001
logPrice	0.93613 <.0001	0.68669 <.0001	-0.34963 <.0001	1.00000

We started with plotting a correlation matrix and scatter plot to study any first look observational relationship between the variables. The price versus odometer reflects a negative correlation. As we know when the odometer reading increases means car have been run for a longer mile and it would reduce the cost of the car as the odometer value increases.

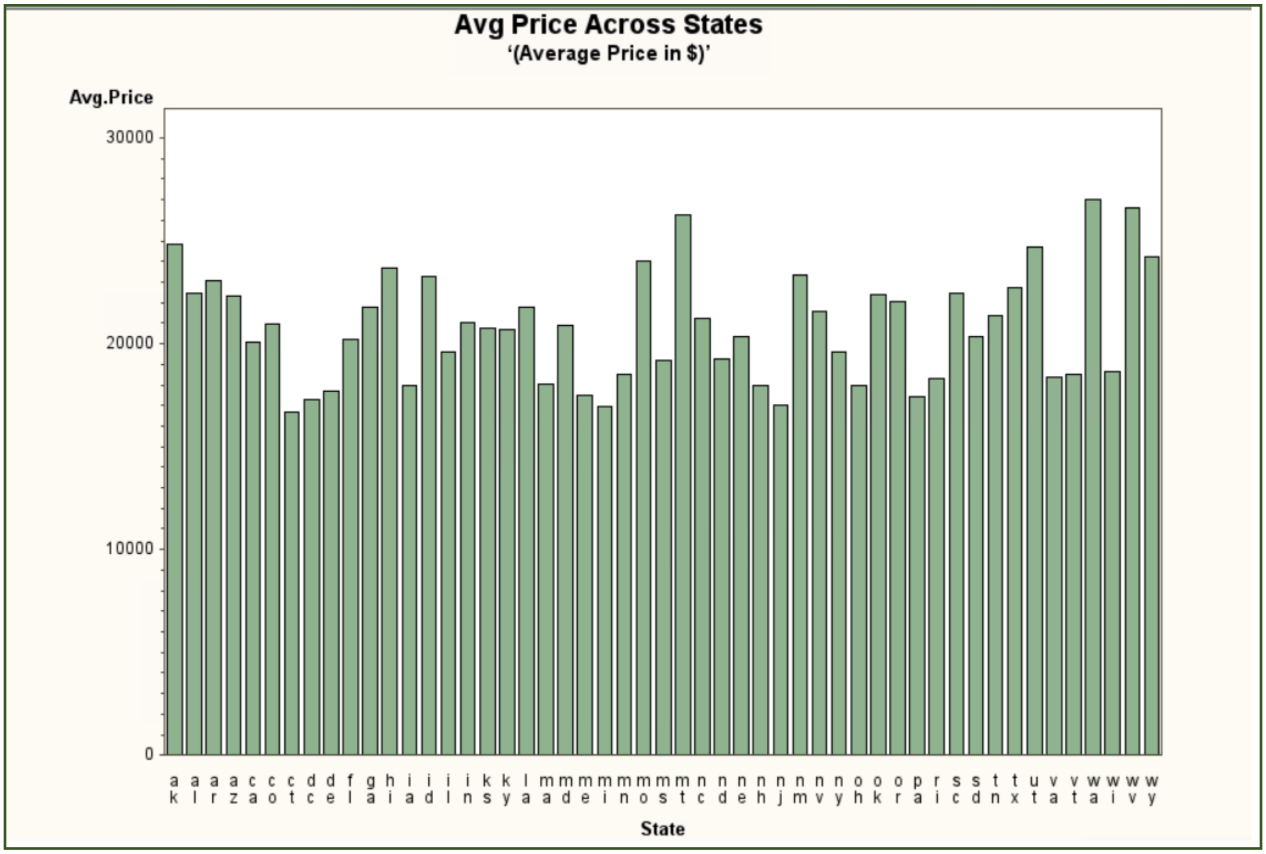
Further we take forward our analysis to get to a conclusion.

Average price of cars across years



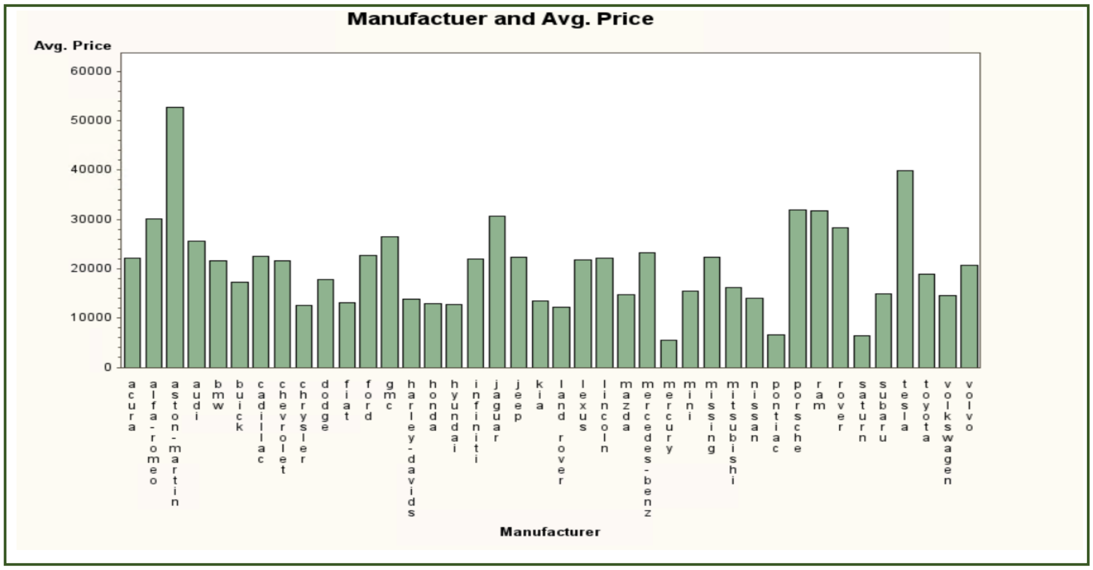
From the graph we can that cars towards the latest years are more costly compared to the cars that are in older years. The market of car prices has taken a big boom ever since new technology has got into the effect. The cars that are made in later years are more equipped with features and maintains a good condition compared to the cars in earlier 2000's. In addition to this new car will have high value as compared to old cars in the resale business as older cards expected to fetch lower price as compared to newer cars

Average Price of the cars across the states in USA



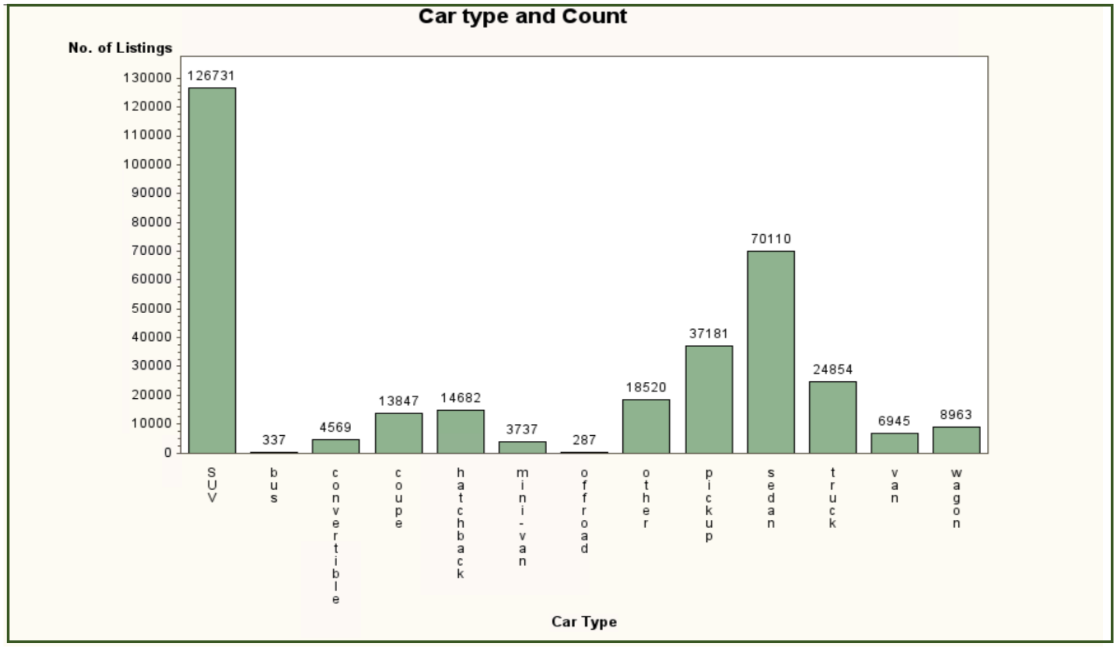
We can see from the graph that cars available in states like Washington , West Virginia, Montana are costlier than other states. This was expected since the prices of cars are more in richer states due to the geographical location across the country.

Average price of cars between different Manufactures



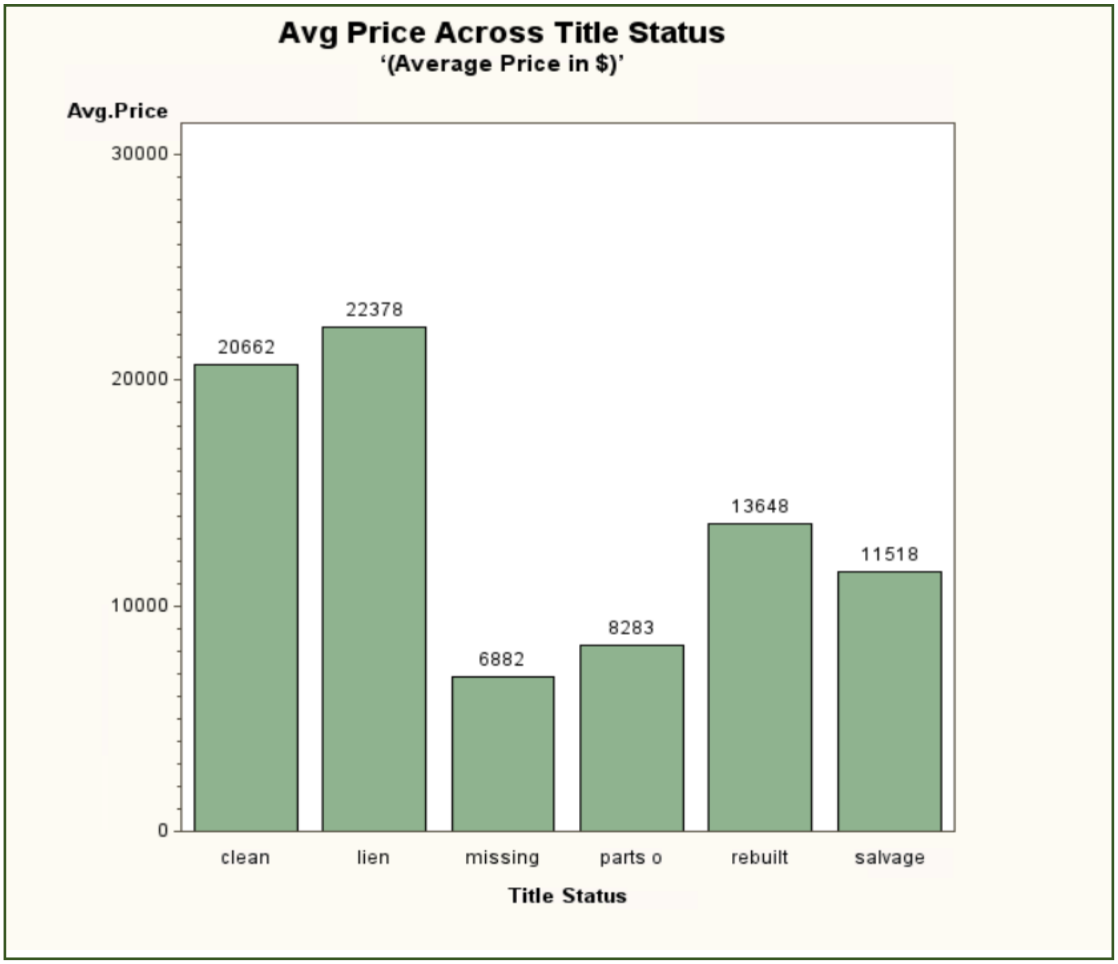
We noticed that the price of cars is higher for cars such as Aston martin, Tesla, Porsche etc. These cars are generally labelled as luxurious brands and the higher prices of them are justified for its luxury branding. Meanwhile prices of cars such as Saturn, Kia, mercury etc. are at a lower end since these are car brands that are known for budget friendly smaller cars available to the common people.

No of cars by Type



SUV and sedan are the greatest number of cars in the listing. This is because, people in United States prefer to have big cars like SUV so that they can have enough space to travel around with their family and it also provides good amount of boot space for the languages. Also, Sedan is one of the commonly used cars by small family people especially office working people , doctors and youth because of the style and comfort it gives them in using the car. We can see that a good number of cars in truck category. Coupe and hatch back cars are lower in number because these are not for every common as these cars are costly and considered under luxury car segment.

Average price of cars across the Title Status



We know that people are more interested to purchase cars that have a clean and good title. Also, we can see that lien title is when car is still in a mortgage or in EMI which will have to be transferred to the new owner who is willing to buy the car. This is the reason why the price of the cars is higher for clean and Lien titles. There are some title statuses that are marked as “missing”. This does not give much idea about the status of the car and the customer may have to consider some other features to make a decision.

Hypothesis Testing:

Following are the null hypotheses we will test using various statistical methods.

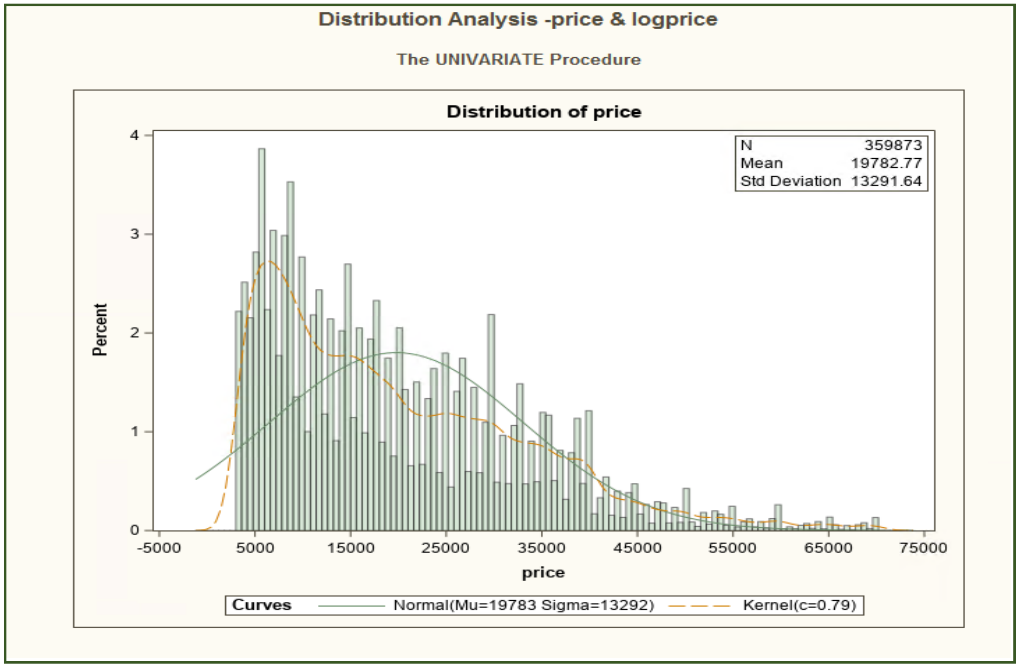
- Higher odometer value cars are not less costly.
- Used cars manufactured in latest years are less costly.
- Title of cars does not have an impact on the price of the cars.
- Condition of used cars does not have impact on price of the cars.

Linear Regression

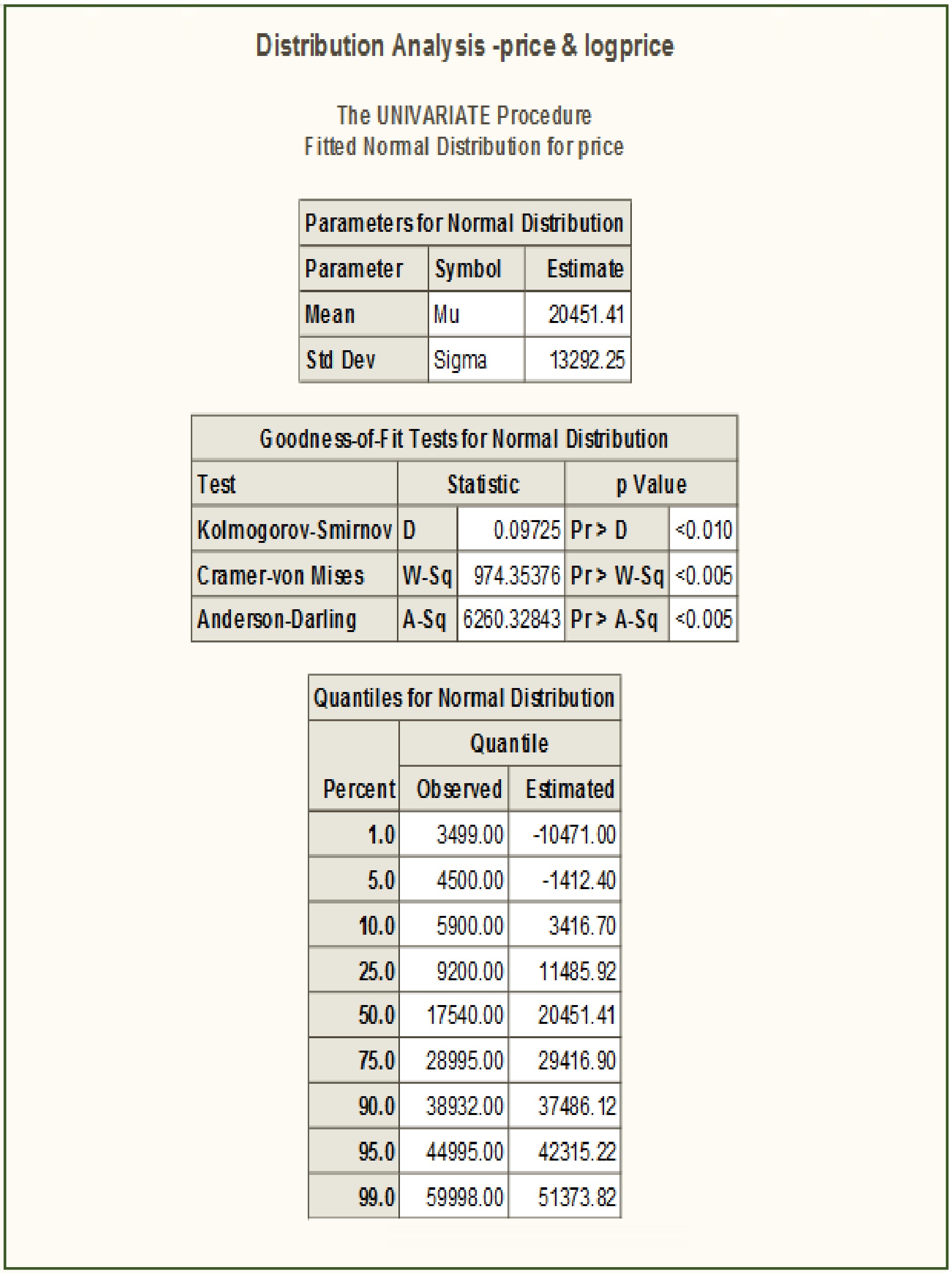
we build linear regression models using different variables to find the best fit and compare it with the base model which we have using Anova. After model selection, we also perform analysis of residuals.

1.Response variable- Price

We first decided to look at the distribution of price:



From the chart, we can see that the price is positively skewed. We may need to make some adjustment towards the price variable.



Distribution Analysis -price & logprice

The UNIVARIATE Procedure
Fitted Normal Distribution for price

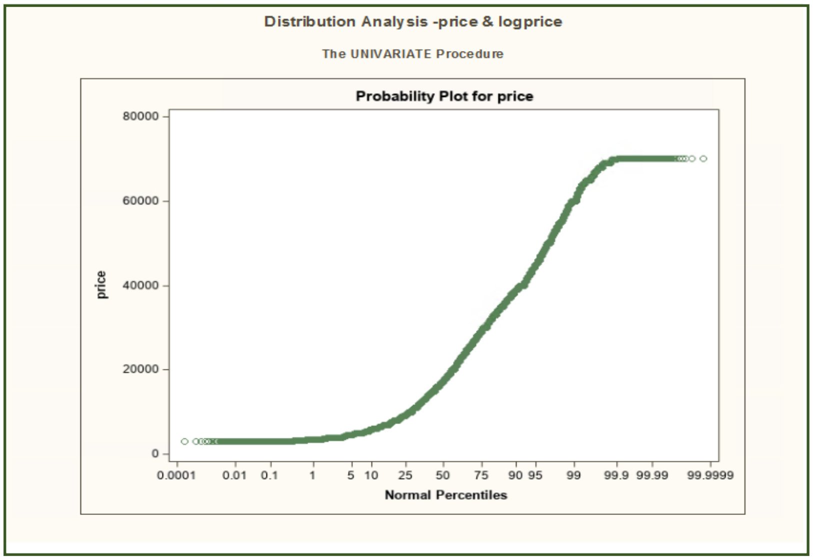
Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	20451.41
Std Dev	Sigma	13292.25

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.09725	Pr > D	<0.010
Cramer-von Mises	W-Sq	974.35376	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	6260.32843	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	3499.00	-10471.00
5.0	4500.00	-1412.40
10.0	5900.00	3416.70
25.0	9200.00	11485.92
50.0	17540.00	20451.41
75.0	28995.00	29416.90
90.0	38932.00	37486.12
95.0	44995.00	42315.22
99.0	59998.00	51373.82

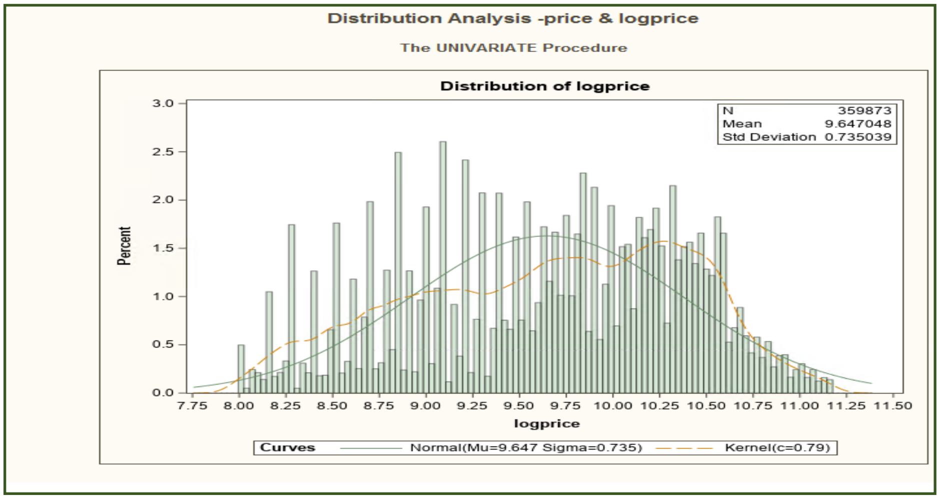
From the Goodness-of-Fit table, we see that the p-value is less than 0.01. We reject the null hypothesis and conclude that the price variable is normally distributed.

Q-Q plot for price variable

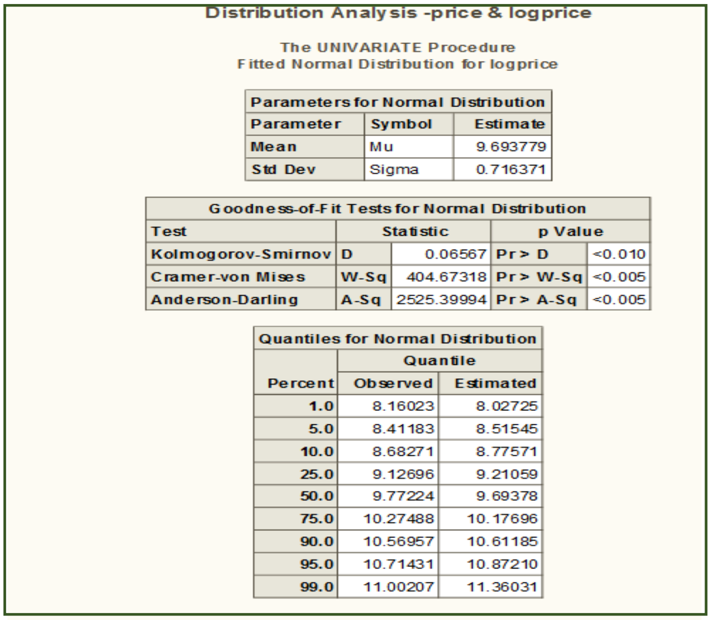


From Q-Q plot, the price variable looks normally distributed and it is consistent with the previous analysis we made.

So, we decided to log the price variable to have a better interpretation for the rest of our Analysis:



As we can see from the above chart, after we logged the price ,the data does not look skewed anymore.



Distribution Analysis -price & logprice

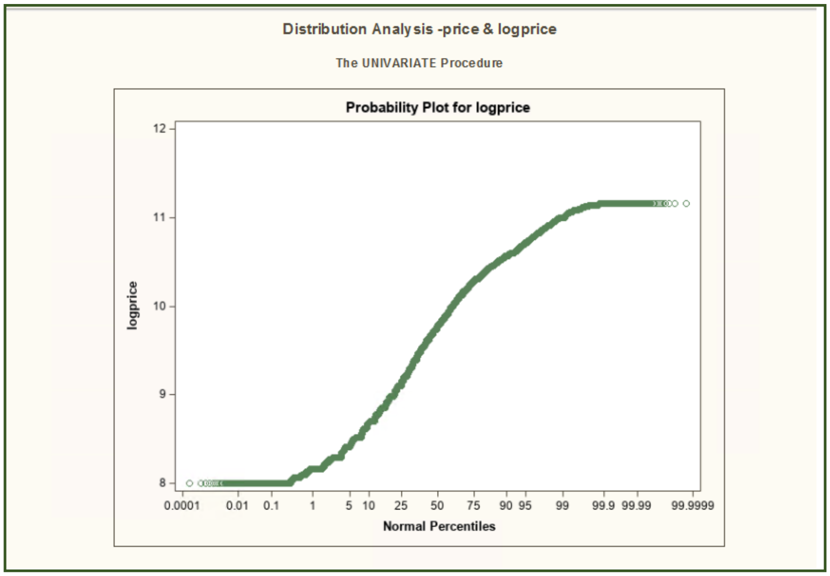
The UNIVARIATE Procedure
Fitted Normal Distribution for logprice

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	9.693779
Std Dev	Sigma	0.716371

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.06567	Pr > D	<0.010
Cramer-von Mises	W-Sq	404.67318	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	2525.39994	Pr > A-Sq	<0.005

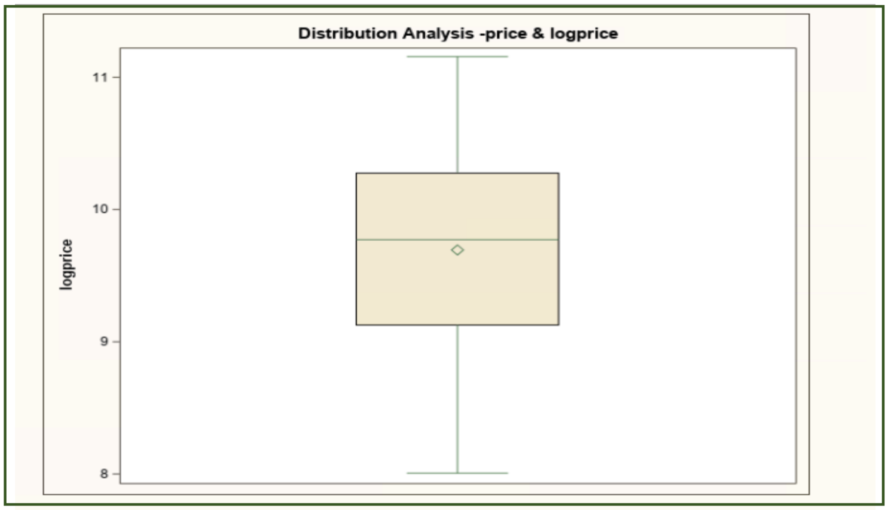
Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	8.16023	8.02725
5.0	8.41183	8.51545
10.0	8.68271	8.77571
25.0	9.12696	9.21059
50.0	9.77224	9.69378
75.0	10.27488	10.17696
90.0	10.56957	10.61185
95.0	10.71431	10.87210
99.0	11.00207	11.36031

From Goodness-of-Fit test, the p-value is less than 0.01. We reject the null hypothesis and conclude that log of price variable is normally distributed.



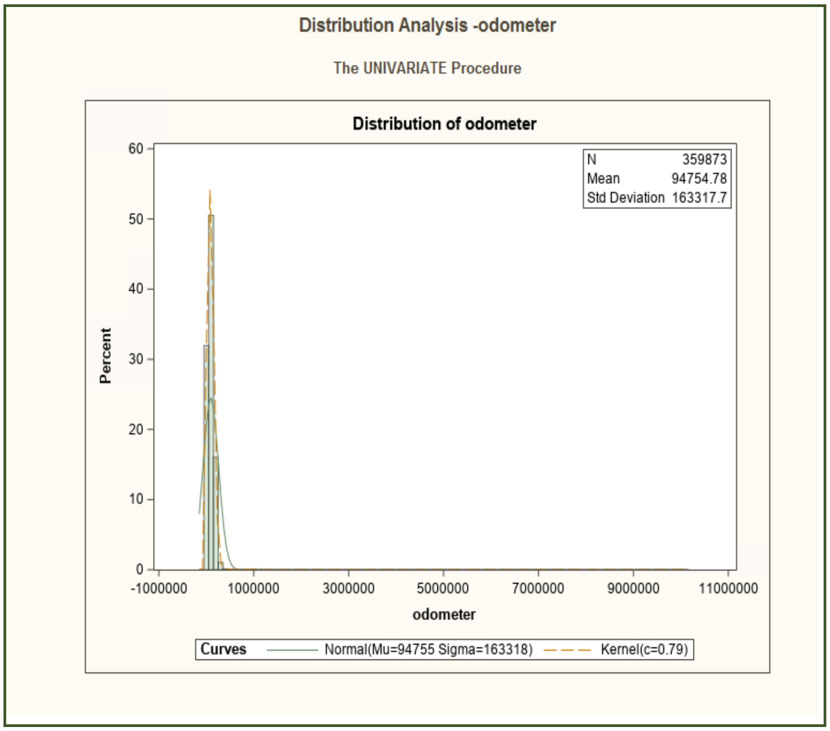
The Q-Q plot also looks slightly better this time after we logged the price.

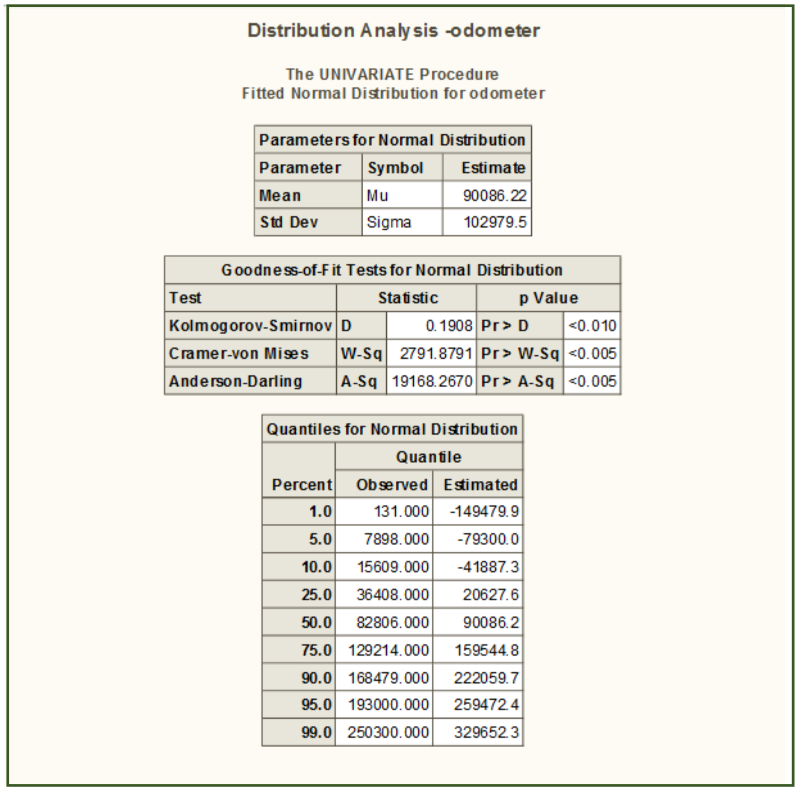
Just to make sure, **we created a box plot as well and checked for the outliers.** The variable looks normally distributed and there are no obvious outliers at this moment.



2. **Predictor Variable – Odometer, Year**

We generated the univariate plot to observe the distribution of odometer variable.





Distribution Analysis -odometer

The UNIVARIATE Procedure
Fitted Normal Distribution for odometer

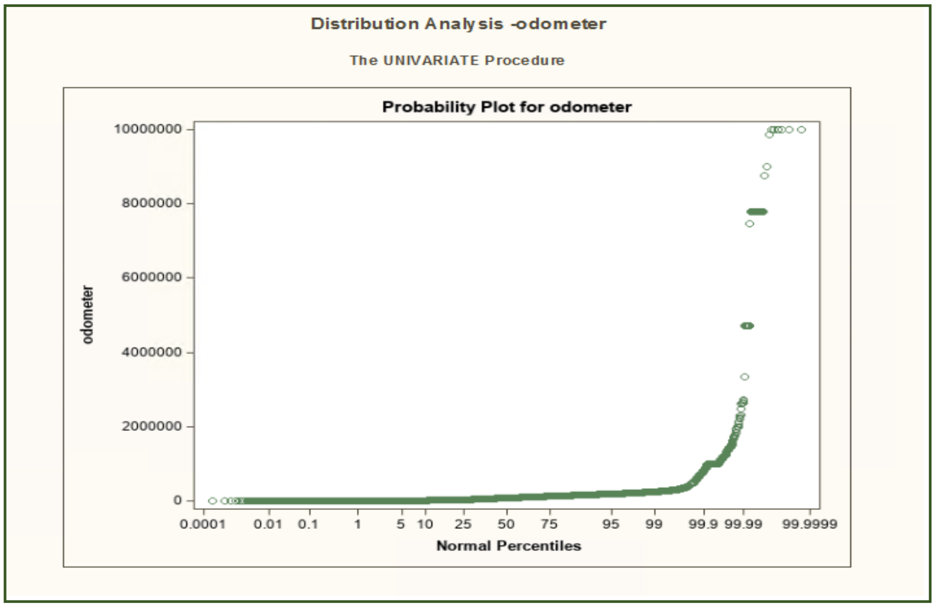
Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	90086.22
Std Dev	Sigma	102979.5

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.1908	Pr > D	<0.010
Cramer-von Mises	W-Sq	2791.8791	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	19168.2670	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	131.000	-149479.9
5.0	7898.000	-79300.0
10.0	15609.000	-41887.3
25.0	36408.000	20627.6
50.0	82806.000	90086.2
75.0	129214.000	159544.8
90.0	168479.000	222059.7
95.0	193000.000	259472.4
99.0	250300.000	329652.3

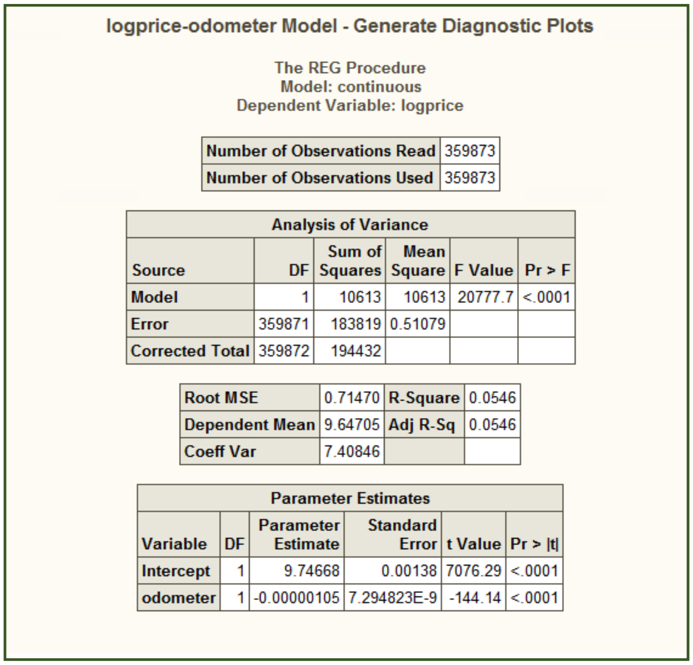
According to the p-value of Goodness-of -Fit tests, the p-value is less than 0.01. Thus, the odometer variable is normally distributed.

Distribution analysis of Odometer:



We can see from the the Q-Q plot that the percentile does not look quite linear. We believe there may need some modifications later for odometer variable.

We use log of price as the response variable and odometer as the predictor variable to run a simple regression model.



logprice-odometer Model - Generate Diagnostic Plots

The REG Procedure
Model: continuous
Dependent Variable: logprice

Number of Observations Read	359873
Number of Observations Used	359873

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10613	10613	20777.7	<.0001
Error	359871	183819	0.51079		
Corrected Total	359872	194432			

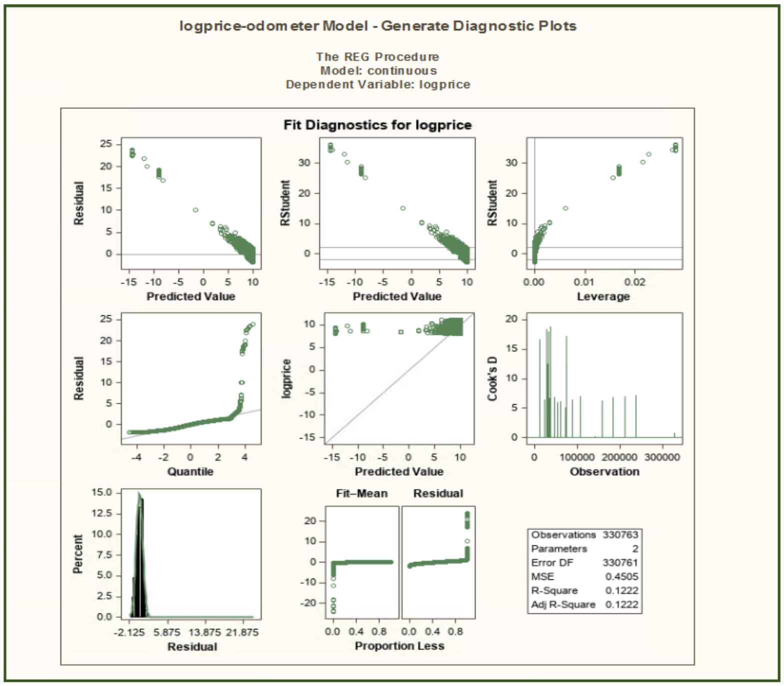
Root MSE	0.71470	R-Square	0.0546
Dependent Mean	9.64705	Adj R-Sq	0.0546
Coeff Var	7.40846		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.74668	0.00138	7076.29	<.0001
odometer	1	-0.00000105	7.294823E-9	-144.14	<.0001

The F-value is <0.0001. Overall, this model is statistically significant, which means there is a linear relationship between price and odometer.

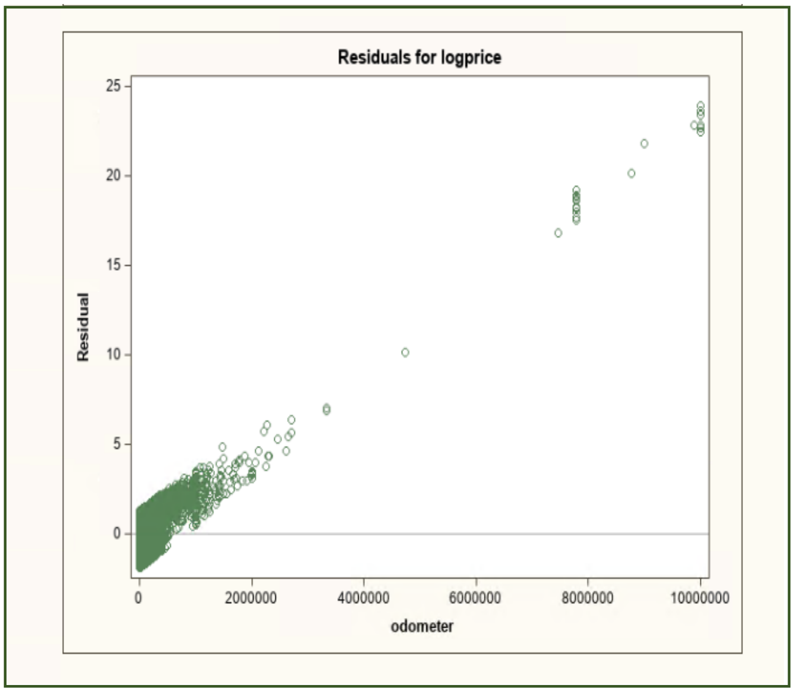
The estimate of odometer is -0.000000105. The t-value is -144.14. The p-value is <0.001. This coefficient is statistically significant. However, the estimate is a very small number, and it is hard to be interpreted straight-forwardly.



logprice-odom eter Model - Generate Diagnostic Plots

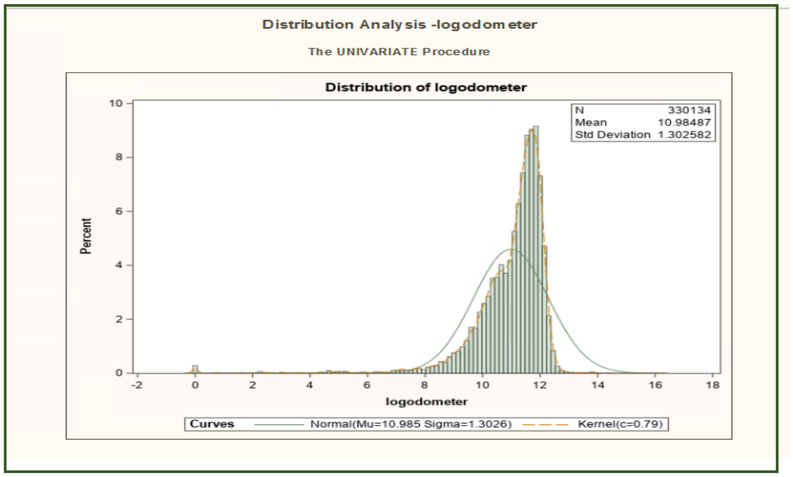
The REG Procedure
Model: continuous
Dependent Variable: logprice

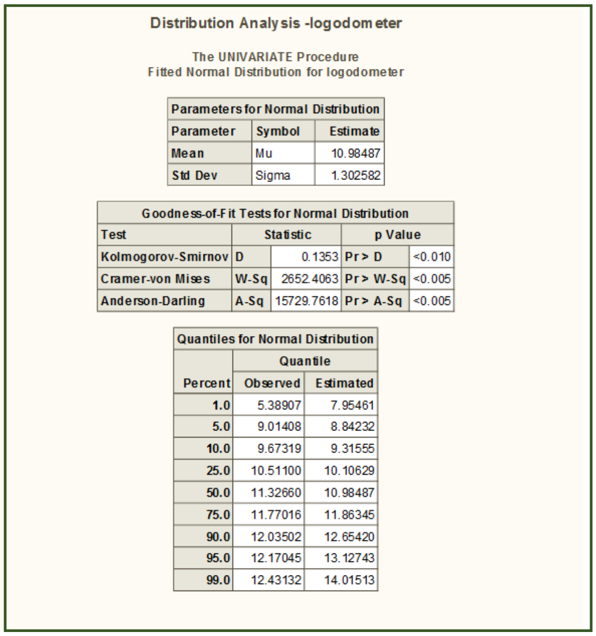
Observations	330763
Parameters	2
Error DF	330761
MSE	0.4505
R-Square	0.1222
Adj R-Square	0.1222



In addition, from the diagnostic plots and residual plots, we can see the pattern as above.

So, we decided to **try to log the odometer variable.**





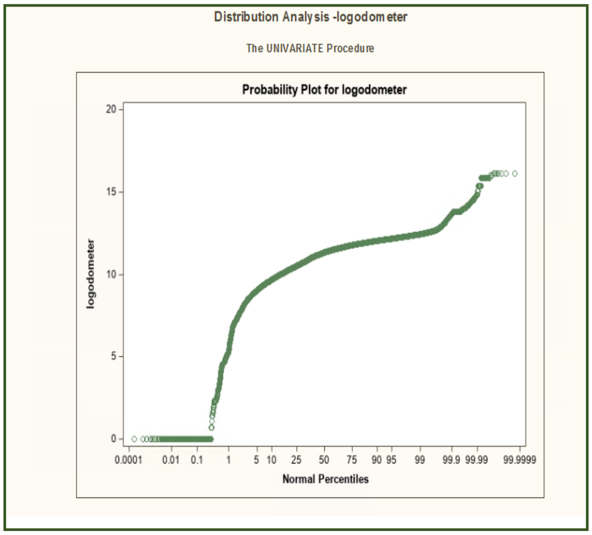
Distribution Analysis -logodometer

The UNIVARIATE Procedure
Fitted Normal Distribution for logodometer

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	10.98487
Std Dev	Sigma	1.302582

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.1353	Pr > D	<0.010
Cramer-von Mises	W-Sq	2652.4063	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	15729.7618	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	5.38907	7.95461
5.0	9.01408	8.84232
10.0	9.67319	9.31555
25.0	10.51100	10.10629
50.0	11.32660	10.98487
75.0	11.77016	11.86345
90.0	12.03502	12.65420
95.0	12.17045	13.12743
99.0	12.43132	14.01513

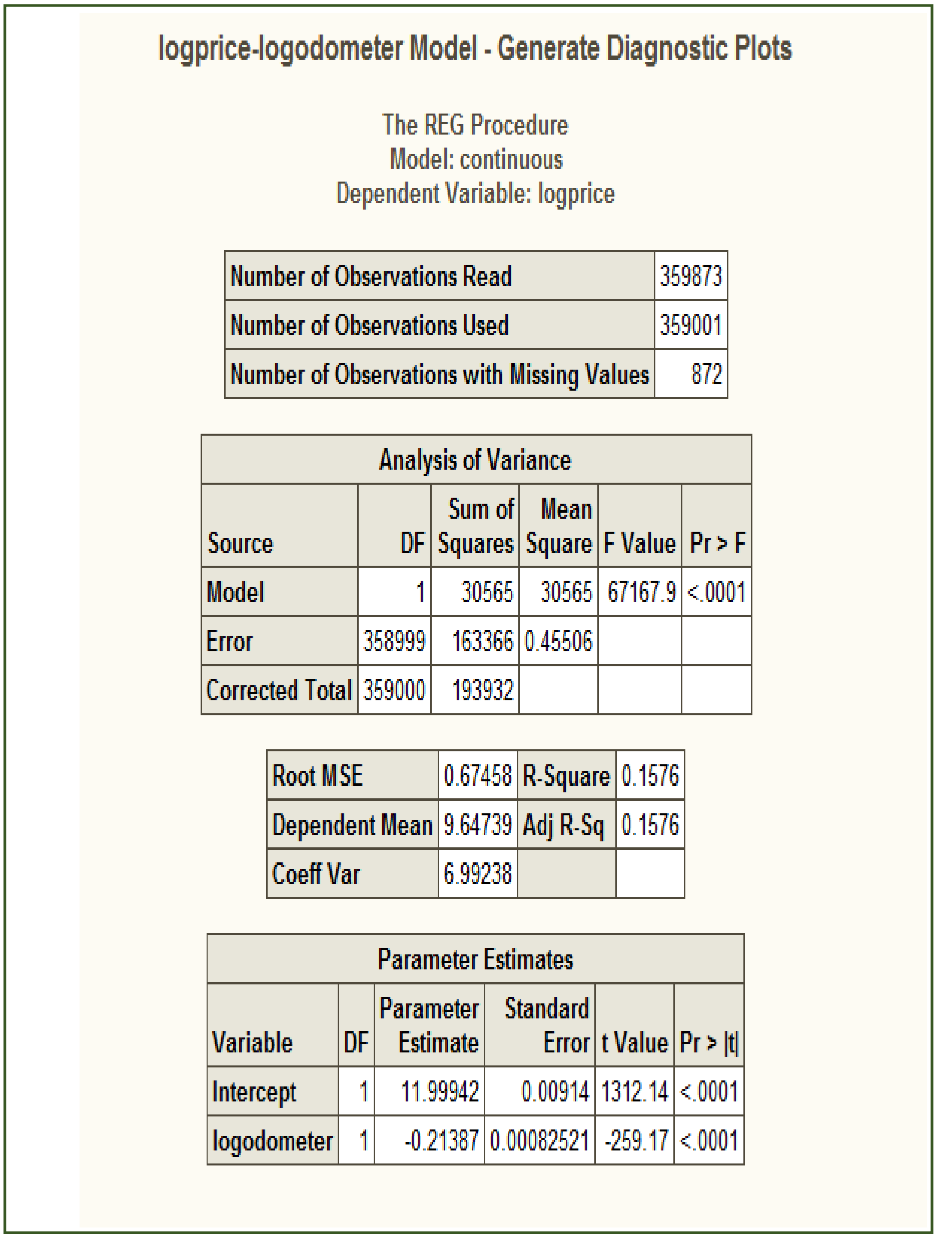


Distribution Analysis -logodometer

The UNIVARIATE Procedure

This time, the log of odometer variable looks slightly negatively skewed, but according to the p-value of Goodness-of-Fit tests, it is still normally distributed. The Q-Q plot looks better than previous one.

We use log of price as the response variable and log of odometer as the predictor variable to run a simple regression model again.



logprice-logodometer Model - Generate Diagnostic Plots

The REG Procedure
Model: continuous
Dependent Variable: logprice

Number of Observations Read	359873
Number of Observations Used	359001
Number of Observations with Missing Values	872

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	30565	30565	67167.9	<.0001
Error	358999	163366	0.45506		
Corrected Total	359000	193932			

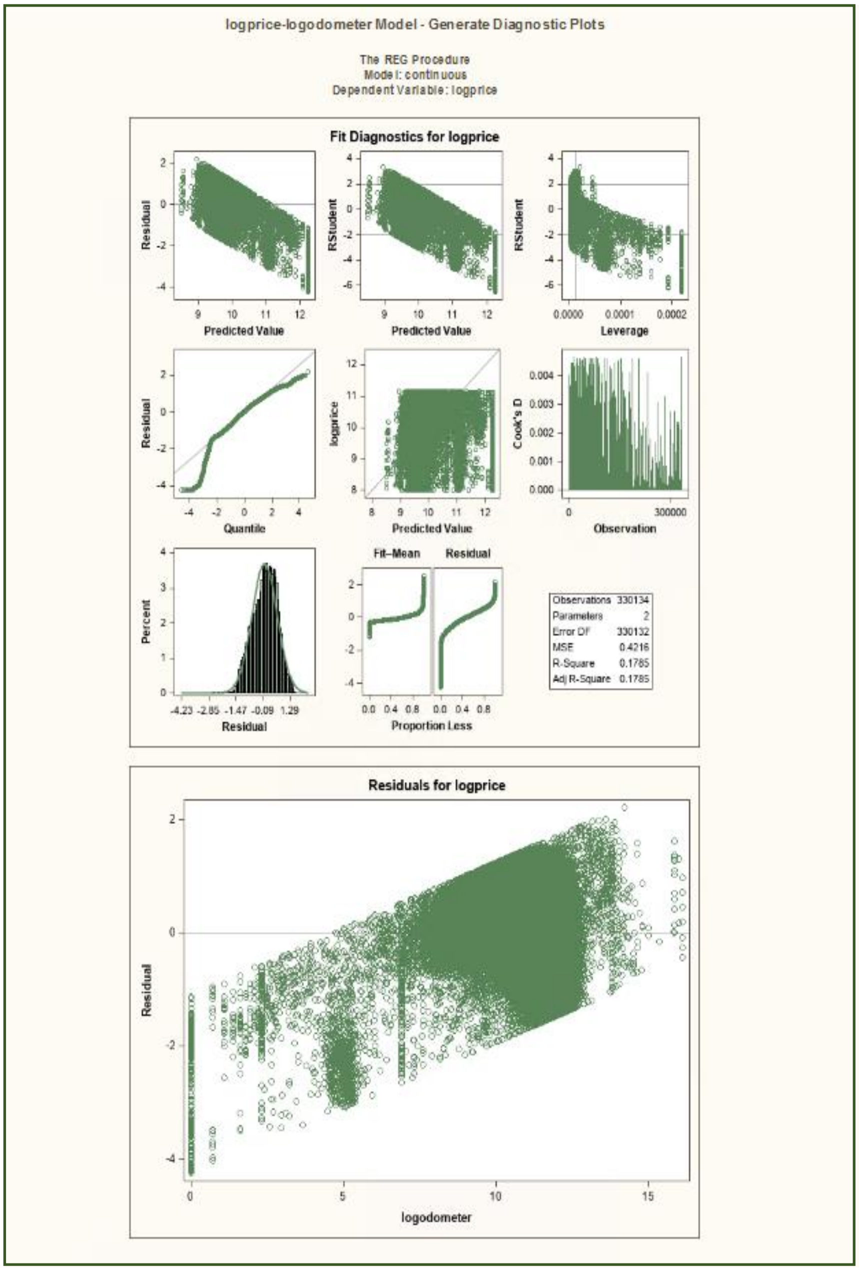
Root MSE	0.67458	R-Square	0.1576
Dependent Mean	9.64739	Adj R-Sq	0.1576
Coeff Var	6.99238		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.99942	0.00914	1312.14	<.0001
logodometer	1	-0.21387	0.00082521	-259.17	<.0001

The p-value of F test is less than 0.0001 indicating the model is statistically significant. The estimate of log of odometer is -0.21. The t-value is -259.17. The p-value is less than 0.0001.

We can reject the null hypothesis. The price and odometer have a **negative relationship**. For every percent increase in odometer, the price of used cars will be decreased 0.21%.



logprice-logodometer Model - Generate Diagnostic Plots

The REG Procedure
Model: continuous
Dependent Variable: logprice

Fit Diagnostics for logprice

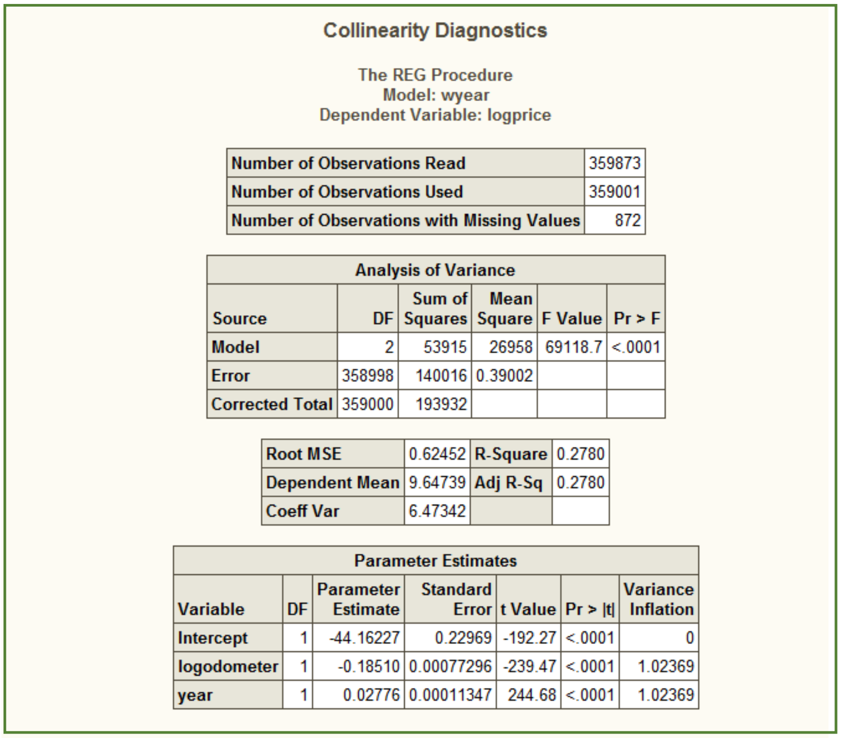
Observations	330134
Parameters	2
Error DF	330132
MSE	0.4216
R-Square	0.1785
Adj R-Square	0.1785

Residuals for logprice

From the diagnostic plots. The quantile and percent of residual looks good. However, there are plenty of influential observations according to RStudent plot and Cook's D chart. Since the price of luxury cars are very expensive regardless of used cars, it depends on the perspective if the researchers will take them as outliers.

Next, we included year as another predictor variable and run the multiple linear regressions.

We ran the **VIF test** and check if there is multicollinearity between the two predictor variables.



Collinearity Diagnostics

The REG Procedure
Model: wyear
Dependent Variable: logprice

Number of Observations Read	359873
Number of Observations Used	359001
Number of Observations with Missing Values	872

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	53915	26958	69118.7	<.0001
Error	358998	140016	0.39002		
Corrected Total	359000	193932			

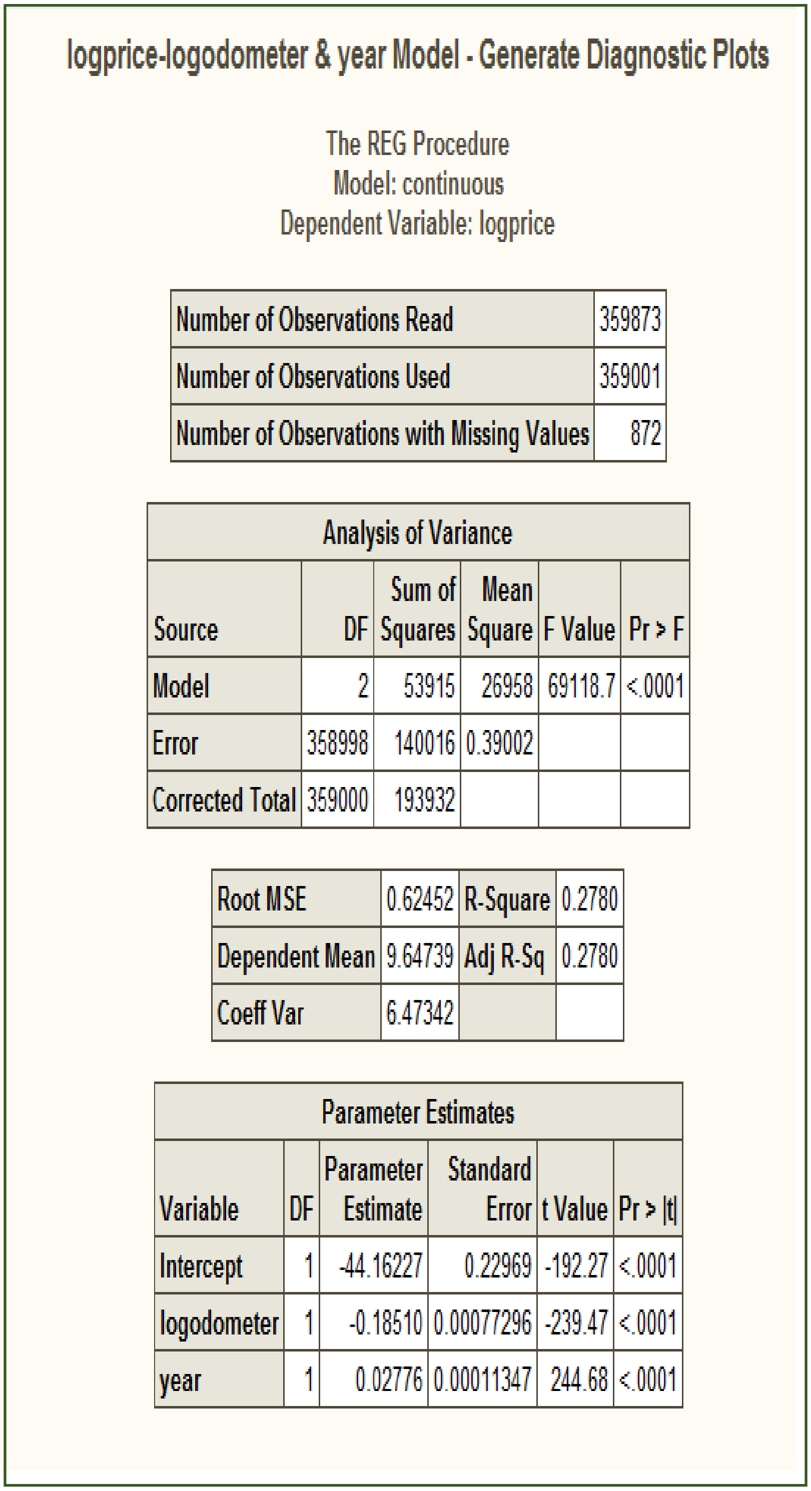
Root MSE	0.62452	R-Square	0.2780
Dependent Mean	9.64739	Adj R-Sq	0.2780
Coeff Var	6.47342		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-44.16227	0.22969	-192.27	<.0001	0
logodometer	1	-0.18510	0.00077296	-239.47	<.0001	1.02369
year	1	0.02776	0.00011347	244.68	<.0001	1.02369

The variance inflation is 1.02369. This number is less than 10 so we do not need to drop any one of these two variables.

SO , we run the multiple linear regressions model:



logprice-logodometer & year Model - Generate Diagnostic Plots

The REG Procedure
Model: continuous
Dependent Variable: logprice

Number of Observations Read	359873
Number of Observations Used	359001
Number of Observations with Missing Values	872

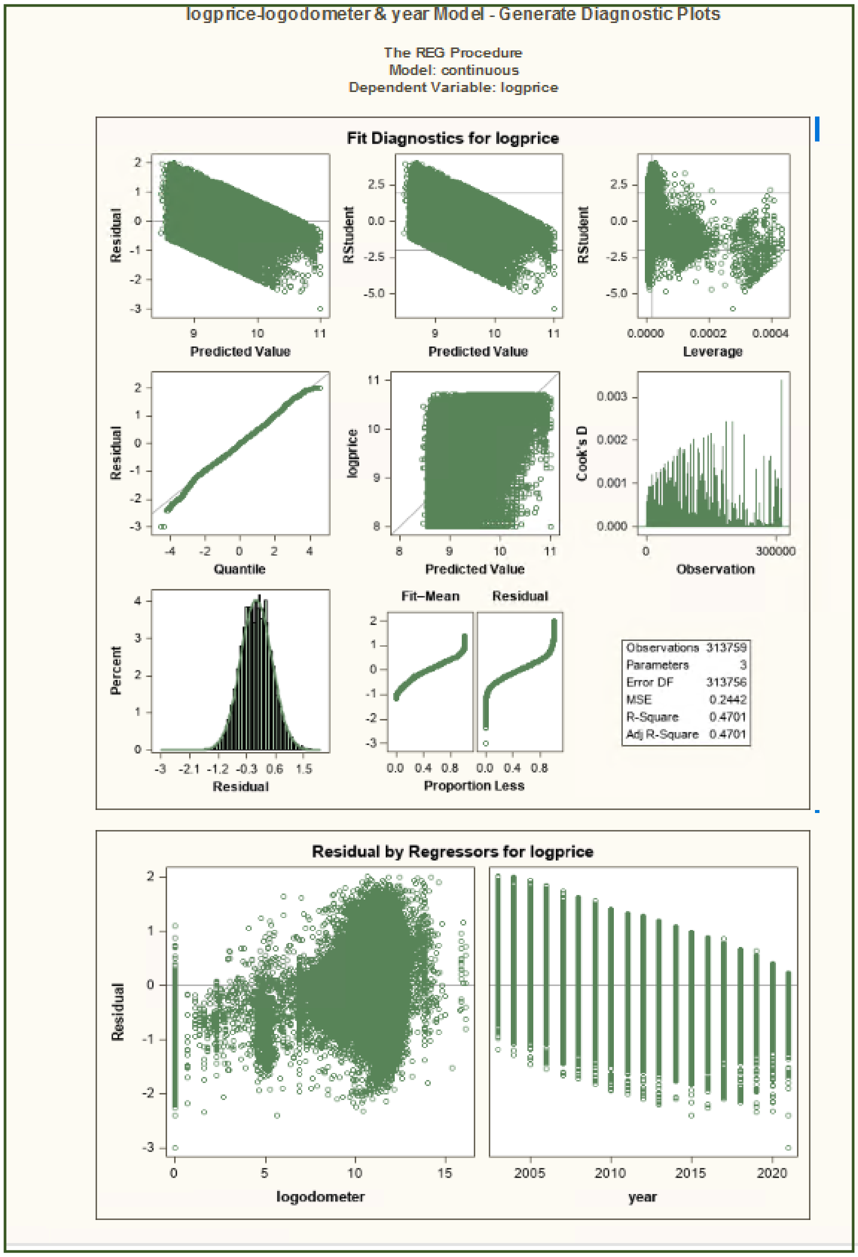
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	53915	26958	69118.7	<.0001
Error	358998	140016	0.39002		
Corrected Total	359000	193932			

Root MSE	0.62452	R-Square	0.2780
Dependent Mean	9.64739	Adj R-Sq	0.2780
Coeff Var	6.47342		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-44.16227	0.22969	-192.27	<.0001
logodometer	1	-0.18510	0.00077296	-239.47	<.0001
year	1	0.02776	0.00011347	244.68	<.0001



logprice-logodometer & year Model - Generate Diagnostic Plots

The REG Procedure
Model: continuous
Dependent Variable: logprice

Fit Diagnostics for logprice

Observations	313759
Parameters	3
Error DF	313756
MSE	0.2442
R-Square	0.4701
Adj R-Square	0.4701

Residual by Regressors for logprice

The p-value of F test is less than 0.0001 indicating the model is statistically significant. The estimate of log of odometer is -0.185. The t-value is -239.47. The p-value is less than 0.0001.

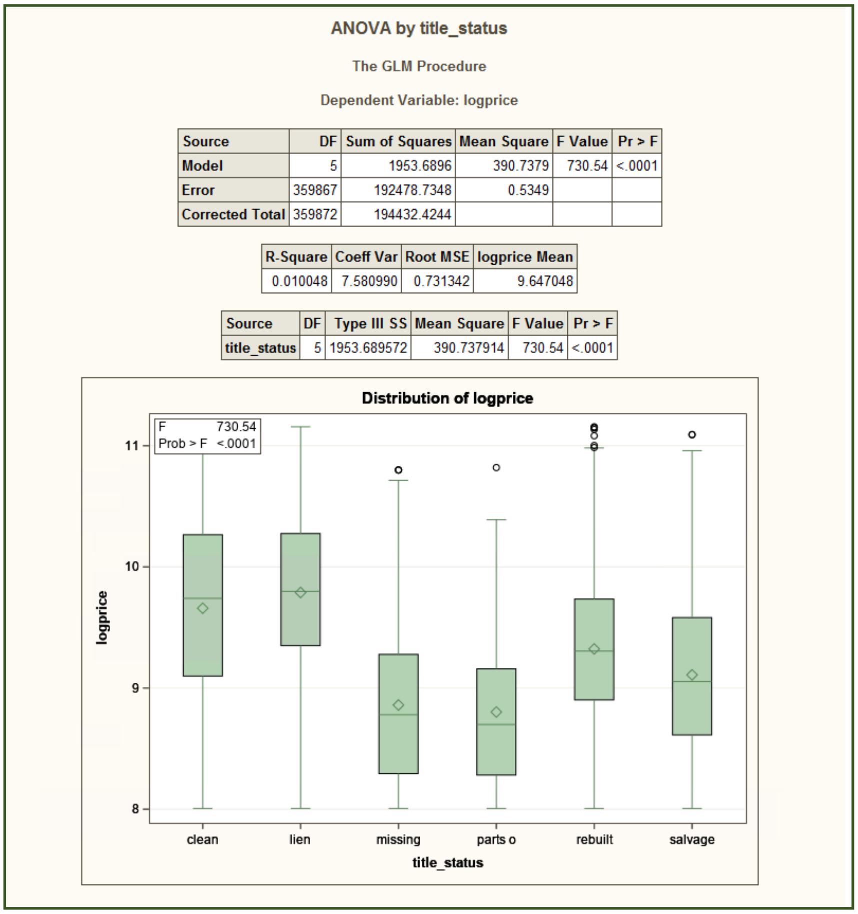
We can reject the null hypothesis. The price and odometer have a negative relationship. For every percent increase in odometer, the price of the used car will be decreased by 0.18%.

The estimate of year variable is 0.02. The t-value is 244.68. The p-value is less than 0.0001.We can reject the null hypothesis. The price and year have a positive relationship. For each year newer of the used car, the price of it will be increased by about 2%.

So, we can conclude from the above results that, we reject the null hypothesis that Higher odometer value cars are not less costly and that used cars manufactured in latest years are less costly.

ANOVA

We use log of price as the response variable group by title of status of used cars and ran the ANOVA.



ANOVA by title_status

The GLM Procedure

Dependent Variable: logprice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1953.6896	390.7379	730.54	<.0001
Error	359867	192478.7348	0.5349		
Corrected Total	359872	194432.4244			

R-Square	Coeff Var	Root MSE	logprice Mean
0.010048	7.580990	0.731342	9.647048

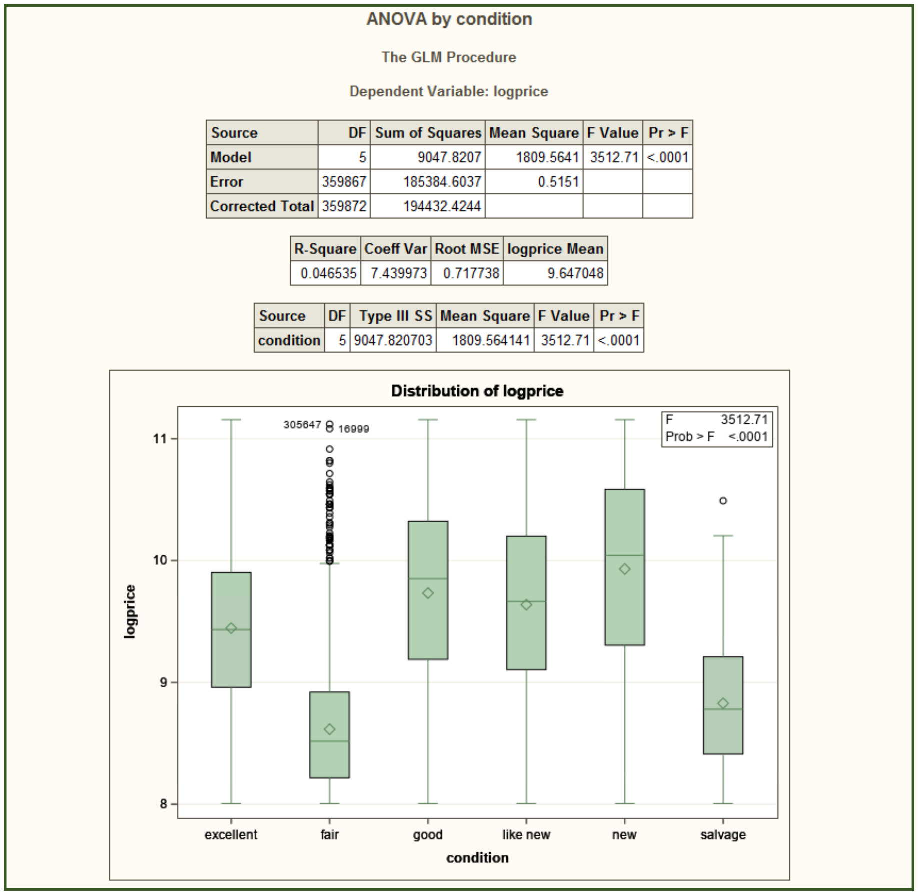
Source	DF	Type III SS	Mean Square	F Value	Pr > F
title_status	5	1953.689572	390.737914	730.54	<.0001

Distribution of logprice

From the plot, we can see that title status can be separated by 6 groups. Our null hypothesis is that the title of status does not have impact on the price of used cars. The F-value 730.54 and p-value is less than 0.0001, which indicates the model is overall significant. This means we Can reject the null hypothesis. **Title status of used cars does have impact on the price of used cars**.

From the box plot, we can infer that if the title is clean or lien, the price will be higher. If the title is missing or parts, the price will comparatively lower.

We used log of price as the response variable group by condition of used cars and ran the ANOVA.



ANOVA by condition

The GLM Procedure

Dependent Variable: logprice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9047.8207	1809.5641	3512.71	<.0001
Error	359867	185384.6037	0.5151		
Corrected Total	359872	194432.4244			

R-Square	Coeff Var	Root MSE	logprice Mean
0.046535	7.439973	0.717738	9.647048

Source	DF	Type III SS	Mean Square	F Value	Pr > F
condition	5	9047.820703	1809.564141	3512.71	<.0001

Distribution of logprice

From the plot, we can see that condition of used cars is divided into 6 groups. Our null hypothesis is that the condition of used cars does not have impact on the price of used cars.

The F-value is 3512.71 and p value is less than 0.0001, which indicates the model is overall significant. This means we reject the null hypothesis. **Condition of used cars does have impact on the price of used cars.**

From the box plot, we can infer that if the condition is new, excellent, like new or good, the price will be comparatively higher. If the condition is fair or salvage, the price will comparatively lower. However, we do notice there are several outliers in the fair group.

Conclusion

The data set we used for our analysis was the used cars and trucks sales data from craigslist. The business problem we were trying to identify was the relationship between price of used cars and odometer, year, title status and the condition of the used car. We started our analysis by performing Exploratory Data analysis on different variables and were able to produce useful insights and results. The EDA graphs helped us in identifying a positive relationship between price of used car and the manufacturing year of the car. Also, we noticed that cars with clean status are priced higher as compared to the other status. We also plotted the correlation matrix to find the relationship between price and odometer which showed a negative relationship between each other. This was logically expected as cars that have been run for a long time are expected to have higher odometer values and will be less priced in the market.

One of our main objectives in this project was to reject our null hypothesis and conclude on the significance of various variables and price. We used linear regression models, diagnostic tests, and ANOVA to reach this objective. From our linear regression results we were able to reject the null hypothesis that Higher odometer value cars are not less costly and that used cars manufactured in latest years are less costly. The price and odometer have a negative relationship. For every percent increase in odometer, the price of the used car will be decreased by 0.18%. Also, the price and year have a positive relationship. For each year newer of the used car, the price of it will be increased by about 2%.

We also performed the ANOVA to test our null hypothesis on title status and condition of the used cars. We rejected the null hypothesis – both title status and condition of used cars – and concluded that cars of cleaner title status are going to be more attracted to the buyers and are expected to have higher prices. The same holds for the condition of cars as well. From the three linear regression models – two simple linear and one multiple regressions, we compared the R2 and the fit diagnostics , we decided to choose multiple regression model as the best one which can explain the relationships between price and other variables better.

This project would be relevant for used car business especially in the United States for the seller as well as the buyer. Figuring out the various relationship between the prices and other variables help the manufacturer to better understand their standing in the current market. They would be able to analyze the results on what kind of cars are selling more for them in the market and what features are important for the customers in regard to the quality and preferences while selecting a car. Accordingly, they can make changes in their production lines or features which will help them to increase the sales of their specific brands and thus increase their brand market. This analysis will also help the purchaser in terms of understanding the current market trends of used cars and how well these cars are been sold and bought by different customers. This will help the customers in this used car market to make better decisions on what to expect and what to decide while selecting a used car from market.

References:

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