Summer 2019

BUAN 6337: Predictive Analytics using SAS

A Report On

"Analysis of Craigslist Cars Data"

THE UNIVERSITY
OF TEXAS AT DALLAS

Submitted by

Vikram Arikath
Nikhil Kumar Kasham
Laila Khan
Sriharshith Reddy Nimmala
Zeeshaan Ramani
Gowtham Sivashanmugam

Under the Guidance of

Dr. Sourav Chatterjee

TABLE OF CONTENTS

PROBLEM STATEMENT

We have decided to build a model based on 2 perspectives - the perspective of a buyer and the perspective of a seller. For example, A seller is looking forward to price his year-old full-size car that runs on gasoline. He has driven the car for 95,000 miles and the car is in good condition. The car has a 4-cylinder engine with front wheel drive and transmission is automatic. Secondly, a student wants to buy a mid-sized sedan to commute to the university. She wants a car that is not older than 15 years. She doesn't really know much about the technical specification, but she finds out from the internet that most of the cars have 4-cylinder front wheel drive. She wants her drive to be comfortable, so she wants a car with automatic transmission. She is not willing to afford more than $4000 and now she wants to know the mileage of the car that she can expect. We will be building a model to find a solution to address these two problems. We want the sellers and buyers to use our model as a base for making a decision to buy and sell a vehicle in America. We will also be creating a logit model to support our findings and conclusions.

INTRODUCTION

The dataset that is the used in this project is from Craigslist, a popular American website used for advertising products and services. Craigslist has separate categories within itself for jobs, housing, products etc. We are exploring the dataset related to the Craigslist Cars section where there are over 1.7 million records related to cars that are advertised in the website. The data is of .csv (Comma Separated Values) format and has 1723065 rows and 26 variables.

The details of cars in the dataset includes the following.

Observations	1723065
Variables	26
Indexes	0
Observation Length	386
Deleted Observations	0
Compressed	NO
Sorted	NO

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
2	city	Char	10	$10.	$10.
7	condition	Char	11	$11.	$11.
21	county_fips	Char	7	$7.	$7.
22	county_name	Char	11	$11.	$11.
8	cylinders	Char	14	$14.	$14.
14	drive	Char	5	$5.	$5.
9	fuel	Char	10	$10.	$10.
18	image_url	Char	61	$61.	$61.
19	lat	Char	11	$11.	$11.
20	long	Char	13	$13.	$13.
6	make	Char	23	$23.	$23.
5	manufacturer	Char	11	$11.	$11.
10	odometer	Char	8	$8.	$8.
17	paint_color	Char	7	$7.	$7.
3	price	Char	7	$7.	$7.
15	size	Char	13	$13.	$13.
24	state_code	Char	4	$4.	$4.
23	state_fips	Char	4	$4.	$4.
25	state_name	Char	14	$14.	$14.
11	title_status	Char	9	$9.	$9.
12	transmission	Char	11	$11.	$11.
16	type	Char	8	$8.	$8.
1	url	Char	85	$85.	$85.
13	vin	Char	19	$19.	$19.
26	weather	Char	4	$4.	$4.
4	year	Char	6	$6.	$6.

S. No	Column name	Description
1	url	URL that leads to the link of classified advertisement
2	City	City where the car is located
3	Price	Price of the car in USD ($)
4	Year	Year of manufacturing
5	Manufacturer	Name of the car manufacturer
6	Make	Name of the car
7	Condition	If the car is in excellent condition, like new etc
8	Cylinders	Number of engine cylinders of the car
9	Fuel	If the car runs on Gasoline, diesel or hybrid
10	Odometer	No of miles the car has been driven
11	Title_status	If the car is clean, Salvage
12	Transmission	If the car is Manual transmission or Automatic transmission
13	Vin	Vehicle Identification Number of the car
14	Drive	If the car is 4-wheel drive, Front wheel drive, Rear wheel drive
15	Size	Size of the car, if it is a compact, mid-size car etc.
16	Type	Type of car, if it is a sedan, SUV, Hatchback etc.
17	Paint_Color	Color of the car
18	Image_url	URL that links to the image of the car
19	Lat	Latitude coordinates of the location of car
20	Long	Longitude coordinates of the location of car
21	County_fips	ID number of the county where the car is located
22	County_name	Name of the county where the car is located
23	State_fips	ID number of the state where the car is located
24	State_code	State code of the state where the car is located
25	State_name	Name of the state where the car is located
26	Weather	Temperature in Farenheit

This dataset requires lot of preprocessing before any analysis could be performed on it. Then the EDA will be reported that will give us the initial analysis of the data. The objective of the project is well defined and is to develop a model to predict the price of the car based on all the other relevant variables. Hence the price is the dependent variable with the rest of the variables being independent. The goal is to create a model with the best use of given data to predict the price of car and to study how the dependent variables affect the price of the car.

DATA PREPROCESSING

As expected from a real world data-set there is a lot of missing and irregular values in all columns in the dataset. Other than that, it is also to be noted from running the PROC CONTENTS code that all the columns are in character format, including columns such as price, year, odometer, latitude, longitude, state_code and weather. So the primary step for data processing is to convert the required columns to continuous variables using the below code:

```
/* Convert character to numeric type */
data craig1;
 set craig;
 odometer_new=input(odometer,best12.);
 price_new=input(price,best12.);
 weather_new=input(weather,best12.);
 year_new=input(year,best12.);
run;
```

This results in the creation of 4 new columns such as odometer_new, price_new, weather_new and year_new all in the format of a continuous variable.

Now, among these continuous variables it is noted that there are missing and irregular values among the price variable. To ensure that there is no disturbance while exploring the data or building the model we first clear the outliers, the extreme values on the upper and lower bound and then impute the mean values of the column into the missing fields. The ranges for the upper and lower bound values are set in such a way that the removing them will not cause too much data to be lost. Hence, for cleaning the price variable, we use the following codes:

```
 /* Cleaning up price variable */
data craig_p1;
     set craig_m1;
     If price_new < 50 then delete;
 run;

data craig_p2;
     set craig_p1;
     If price_new > 100000 then delete;
 run;

/* Imputing Mean Values for Price */
proc stdize data=craig_p2 reponly method=mean out=craig_p1;
var price_new;
run;
```

The above codes help in eliminating the outliers and fillinf up the mean values for the missing data. After this data cleaning step we have 1697087 rows left in our dataset.

Next comes the odometer column. Over 33% of the data is missing in the dataset. It is not advisable to remove the empty records in this column since there is a huge loss in data and the predictive model's accuracy will be affected. Hence, only the extreme values of the column are removed using the following code:

```
 /* Cleaning up odometer variable */
data craig_o1;
     set craig_p2;
     if odometer_new > 500000 then delete;
 run;
```

Here, the outliers are deleted and we get a remaining of 1693093 rows.

For the year variable we do the same as odometer and delete extreme values such as dates before 1900, since on a logical perspective as well it is not possible to have a car older than 1900 as cars were introduced only around the 1900s. We use the following code :

```
/* Cleaning up year variable */
data craig_year;
    set craig_o1;
    If year_new <= 1900 then delete;
run;
```

Here, the irregular records are erased and we also get a remaining of 1686662 columns.

It is also to be noted that the limits for outliers were set on each variable after an initial exploratory analysis was done on each variable. The exploratory analysis will be discussed in detail in the next section of the report.

Now, there are several categorical variables in the data which have multiple empty values but are crucial for our predictive and logisitc model. It is not advisable to drop the empty records since a large chunk of data will be lost, which will affect the model. Hence the categorical variables are imputed with the unkown or other category in different columns to ensure they are filled up as well. The following code is used for the same :

```
/* Filling up missing values of categorical variables */
data craig_cat;
    set craig_year;
    If state_name = "FAILED" then state_name="unkown";
    If state_name = " " then state_name="unkown";
    If manufacturer = " " then manufacturer="other";
    If transmission = " " then transmission="other";
    If cylinders = " " then cylinders="unkown";
    If condition = " " then condition="unkown";
    If size = " " then size="unkown";
    If type = " " then type="unkown";
    If paint_color = " " then paint_color="unkown";
    If drive = " " then drive="NA";
run;
```

The above code ensures that the empty missing values are filled up with another category.

As a last step to data pre-preprocessing, it is known categorical variables as such cannot be used for model building for a regression model and hence it is essential to convert them into dummy variables to process them for a multi-linear model. We do it by using the following code :

```
DATA craig_dummy;
      SET craig_cat ;
IF fuel = "gas" THEN fuel_gas = 1;
        ELSE fuel_gas = 0;
IF fuel = "hybrid" THEN fuel_hybrid = 1;
        ELSE fuel_hybrid = 0;
IF fuel = "diesel" THEN fuel_diesel = 1;
        ELSE fuel_diesel = 0;
IF transmission = "automatic" THEN transmission_automatic = 1;
        ELSE transmission_automatic = 0;
IF transmission = "manual" THEN transmission_manual = 1;
        ELSE transmission_manual = 0;
IF size = "full-size" THEN size_fullsize = 1;
        ELSE size_fullsize = 0;
IF size = "mid-size" THEN size_midsize = 1;
        ELSE size_midsize = 0;
IF size = "compact" THEN size_compact = 1;
        ELSE size_compact = 0;
IF size = "sub-compact" THEN size_subcompact = 1;
        ELSE size_subcompact = 0;
IF size = "unkown" THEN size_unkown = 1;
        ELSE size_unkown = 0;
IF cylinders = "6 cylinders" THEN cylinders_6 = 1;
        ELSE cylinders_6 = 0;
IF cylinders = "4 cylinders" THEN cylinders_4 = 1;
        ELSE cylinders_4 = 0;
IF cylinders = "8 cylinders" THEN cylinders_8 = 1;
        ELSE cylinders_8 = 0;
IF cylinders = "other" THEN cylinders_other = 1;
        ELSE cylinders_other = 0;
IF cylinders = "unkown" THEN cylinders_other = 1;
        ELSE cylinders_other = 0;
IF condition = "excellent" THEN condition_excellent = 1;
        ELSE condition_excellent = 0;
IF condition = "good" THEN condition_good = 1;
        ELSE condition_good = 0;
IF condition = "like new" THEN condition_likenew = 1;
        ELSE condition_likenew = 0;
IF condition = "fair" THEN condition_fair = 1;
        ELSE condition_fair = 0;
IF condition = "unkown" THEN condition_unkown = 1;
        ELSE condition_unkown = 0;
IF drive = "4wd" THEN drive_4wd = 1;
        ELSE drive_4wd = 0;
IF drive = "rwd" THEN drive_rwd = 1;
        ELSE drive_rwd = 0;
IF drive = "fwd" THEN drive_fwd = 1;
        ELSE drive_fwd = 0;
IF drive = "NA" THEN drive_na = 1;
        ELSE drive_na = 0;
RUN;
```

This code helps generate several new columns with binary values (0,1) as records and this is will further help while building the predictive and logistic model for the dataset. This concludes the data pre-processing section of the report.

EXPLORATORY DATA ANALYSIS

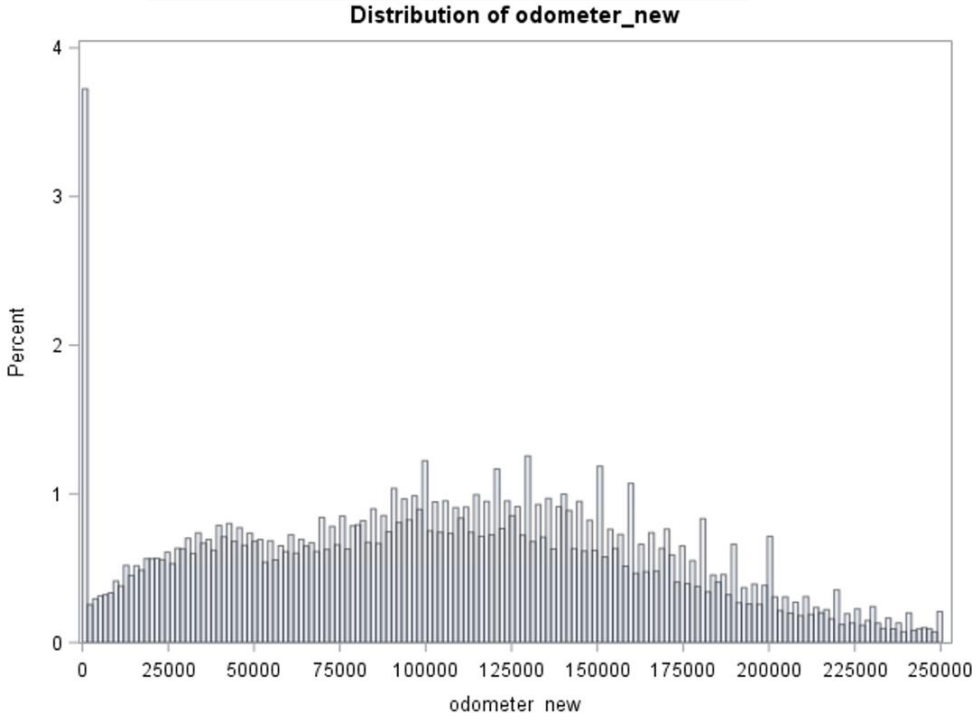
Analysis of the variable: Odometer

The UNIVARIATE Procedure
Variable: odometer_new

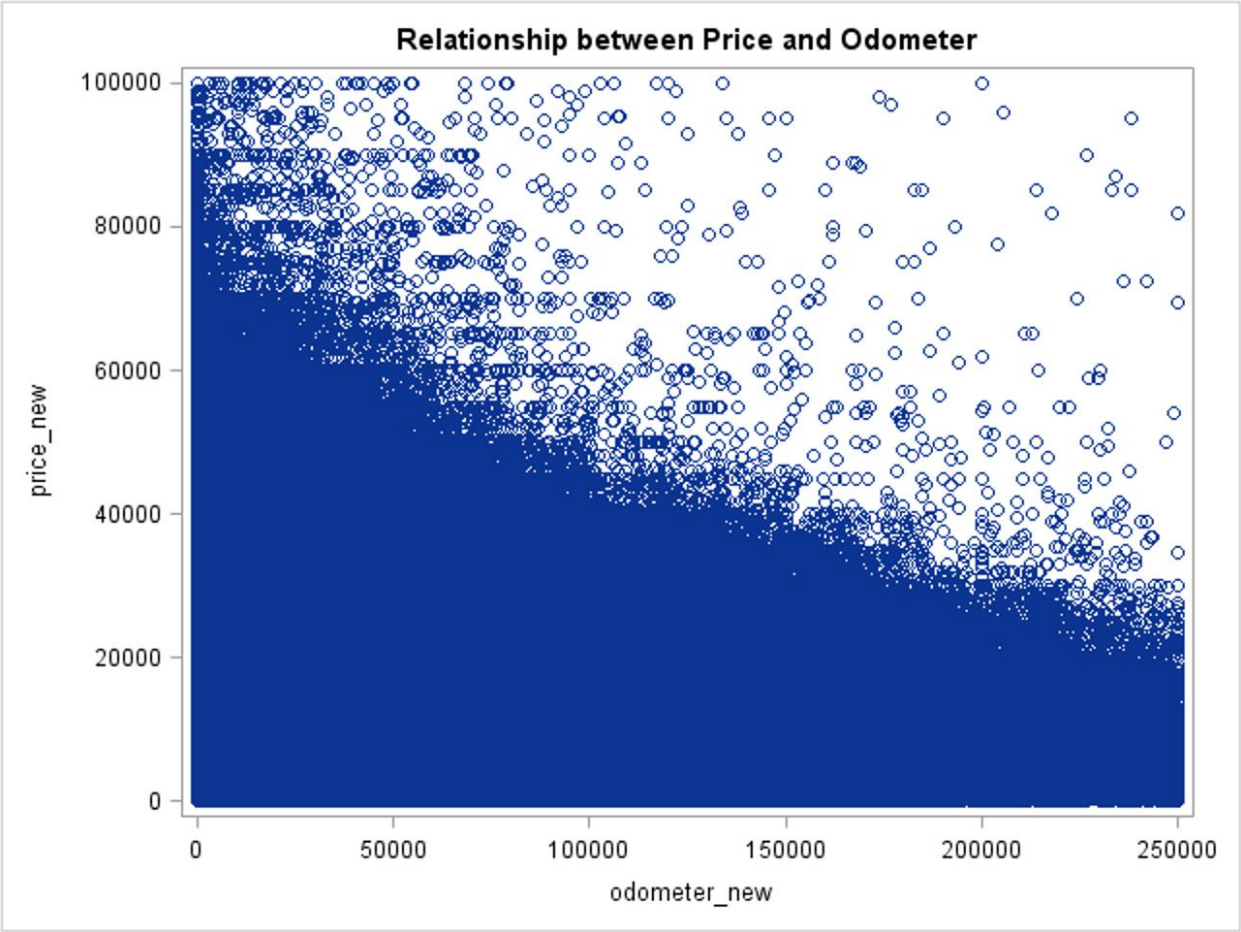
Moments			
N	1117249	Sum Weights	1117249
Mean	105019.007	Sum Observations	1.17332E11
Std Deviation	59946.5003	Variance	3593582893
Skewness	0.1201888	Kurtosis	-0.7581369
Uncorrected SS	1.63371E16	Corrected SS	4.01492E15
Coeff Variation	57.0815723	Std Error Mean	56.7138129

Basic Statistical Measures			
Location		Variability	
Mean	105019.0	Std Deviation	59947
Median	105000.0	Variance	3593582893
Mode	150000.0	Range	250000
		Interquartile Range	92000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	1851.736	Pr > |t|	<.0001
Sign	M	556094.5	Pr >= |M|	<.0001
Signed Rank	S	3.092E11	Pr >= |S|	<.0001



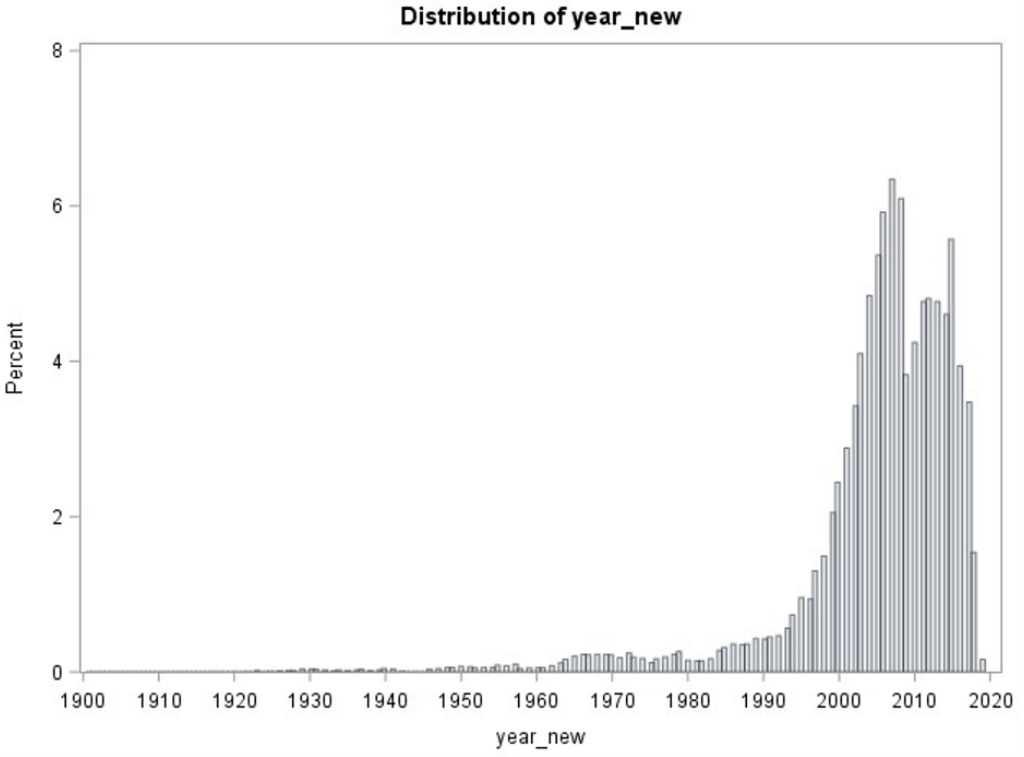
The univariate procedure displays the following distribution for the "Odometer" variable. The distribution is heavily concentrated between the values 25,000 miles and 175,000 miles before tailing off. The mean value for odometer is 105,019 miles with a median of 105,000. The kurtosis value for odometer is -0.75 which indicates that the distribution has lighter tails and a flatter peak compared to the normal distribution.



Using PROC SGPLOT function, we were able to construct a scatter plot to determine the correlation and relationship between odometer and price. From the scatter plot above, we can conclude that odometer and price have a negative relationship. As price of a vehicle increases, the odometer value decreases. The closer the odometer is to 0 or the less miles there are driven for a vehicle, the value of the vehicle is high compared to a vehicle with a high odometer, or a lot of miles driven. This relationship is extremely relevant in the real world. For example, if safety and sustainability was a priority for a family that is car shopping, they would choose an option that has a low odometer reading to avoid old/used up engines, bad condition, and constant hassle with maintenance.

Analysis of the variable: Year

Basic Statistical Measures			
Location		**Variability**	
Mean	2004.972	**Std Deviation**	11.87492
Median	2007.000	**Variance**	141.01363
Mode	2007.000	**Range**	118.00000
		Interquartile Range	10.00000

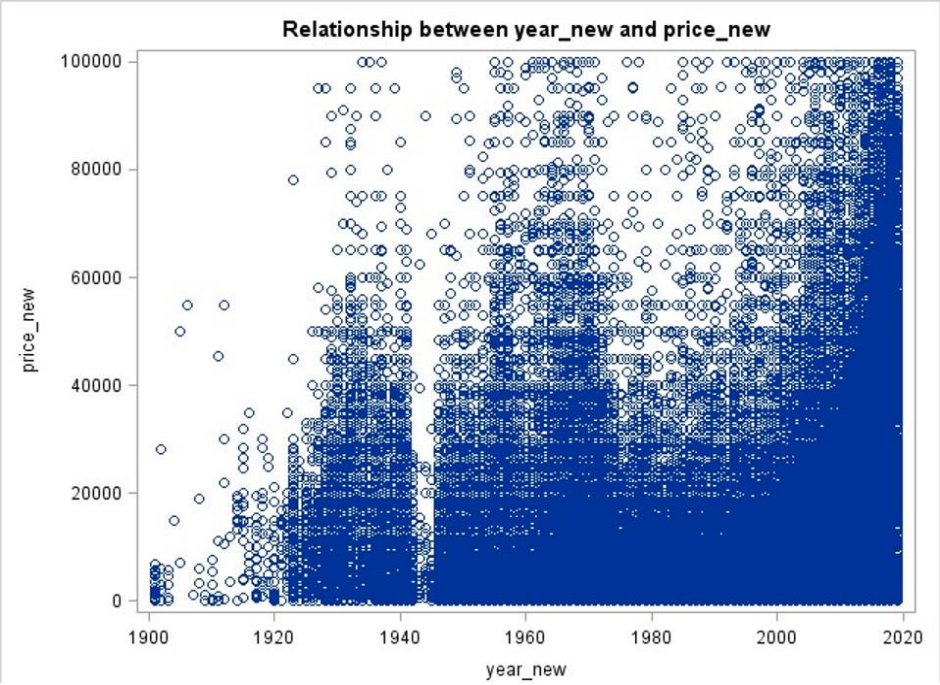


Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
1901	1.62E6	2019	1.66E6
1901	1.61E6	2019	1.66E6
1901	1.38E6	2019	1.66E6
1901	1.18E6	2019	1.67E6
1901	931913	2019	1.67E6

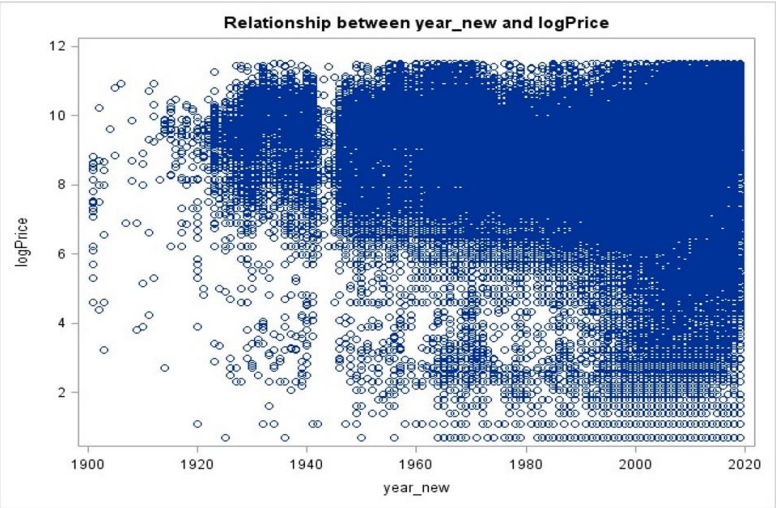
It can be seen from the univariate procedure of the year that the median of the continuous variables is 2007 which means that if we split the number of cars in our dataset into two equal parts, then by 2007 half of the cars in our dataset were manufactured. Since we have observations starting from the year 1900, this means that half of the cars in our data set were made in 107 years and the remaining half were made from

2007 to 2019 i-e in 12 years, so it can be concluded that people started buying more cars and the companies started making more cars after 2007. The mode of the data is also 2007 which shows that 2007 was the year having the largest number of manufactured cars that were to be sold.

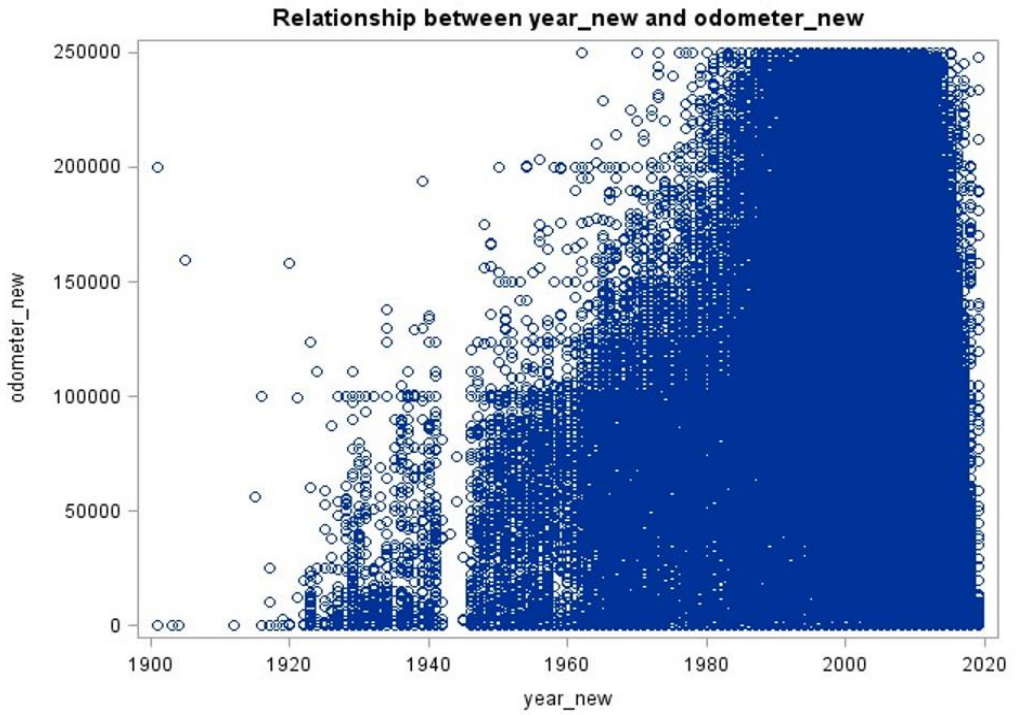
It is obvious from the distribution and probability plot of year _new that the distribution of the data is not normal. It seems that the data is negatively skewed.



The SG Plot between price_new and year_new shows a greater variance which is not desirable. Also, we can see that the prices are decreasing with the increase in years and more cars are available at a cheaper price which isn't the case evident from our dataset. For that reason, we take the log of the price and create an SG plot of logprice and the year_new.

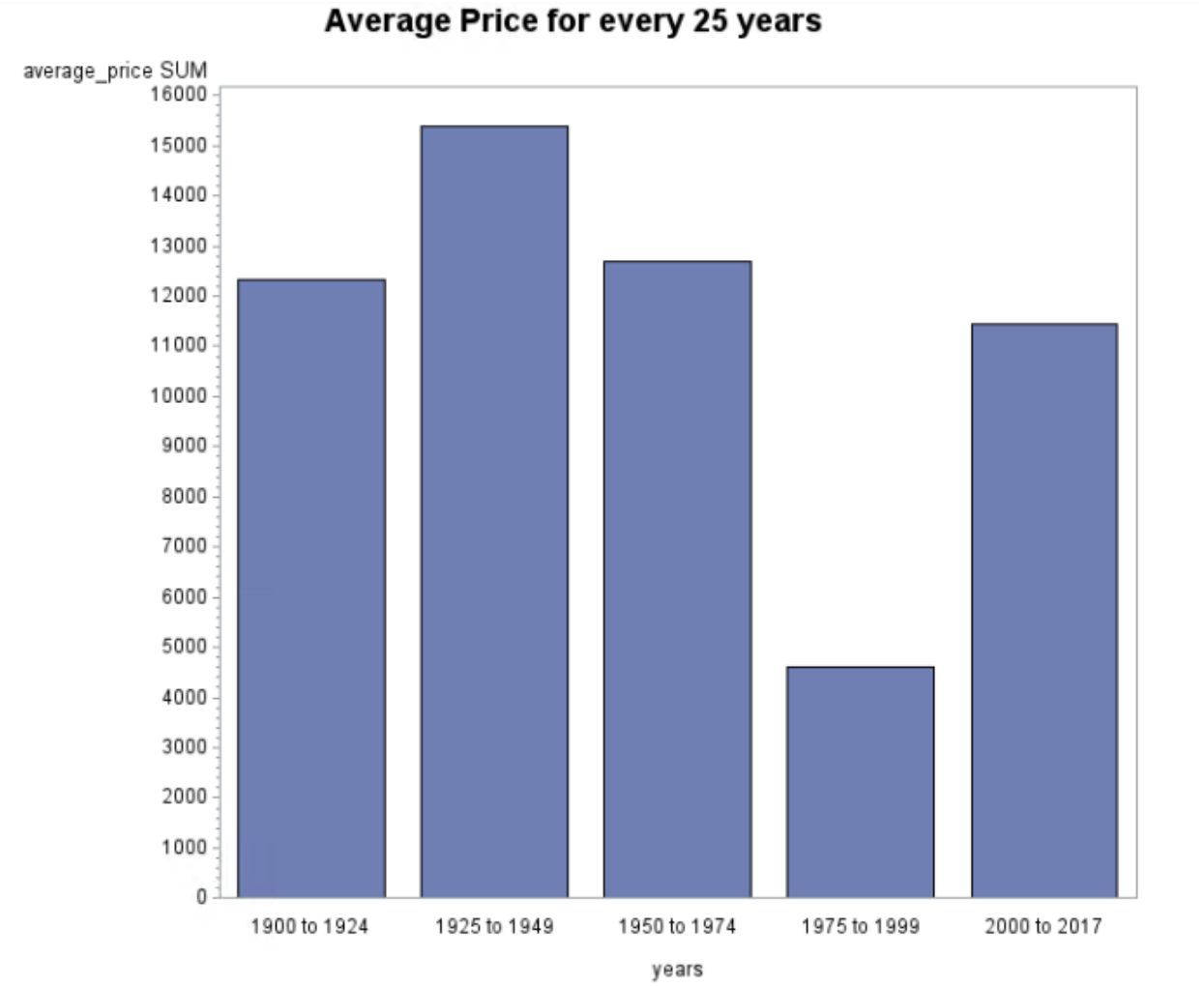


Taking the log of the price reduces the above-mentioned issue of greater variance. If we look at the SG Plot to analyze the relationship between log of price and year, fewer data points can be seen from 1900 to 1980 but as the time progresses more data points can be seen, and the value of the log price can be seen increasing with time especially after 1990. It could be said that people were willing to pay more for the cars or newer technology was used to manufacture cars for which the prices were higher.



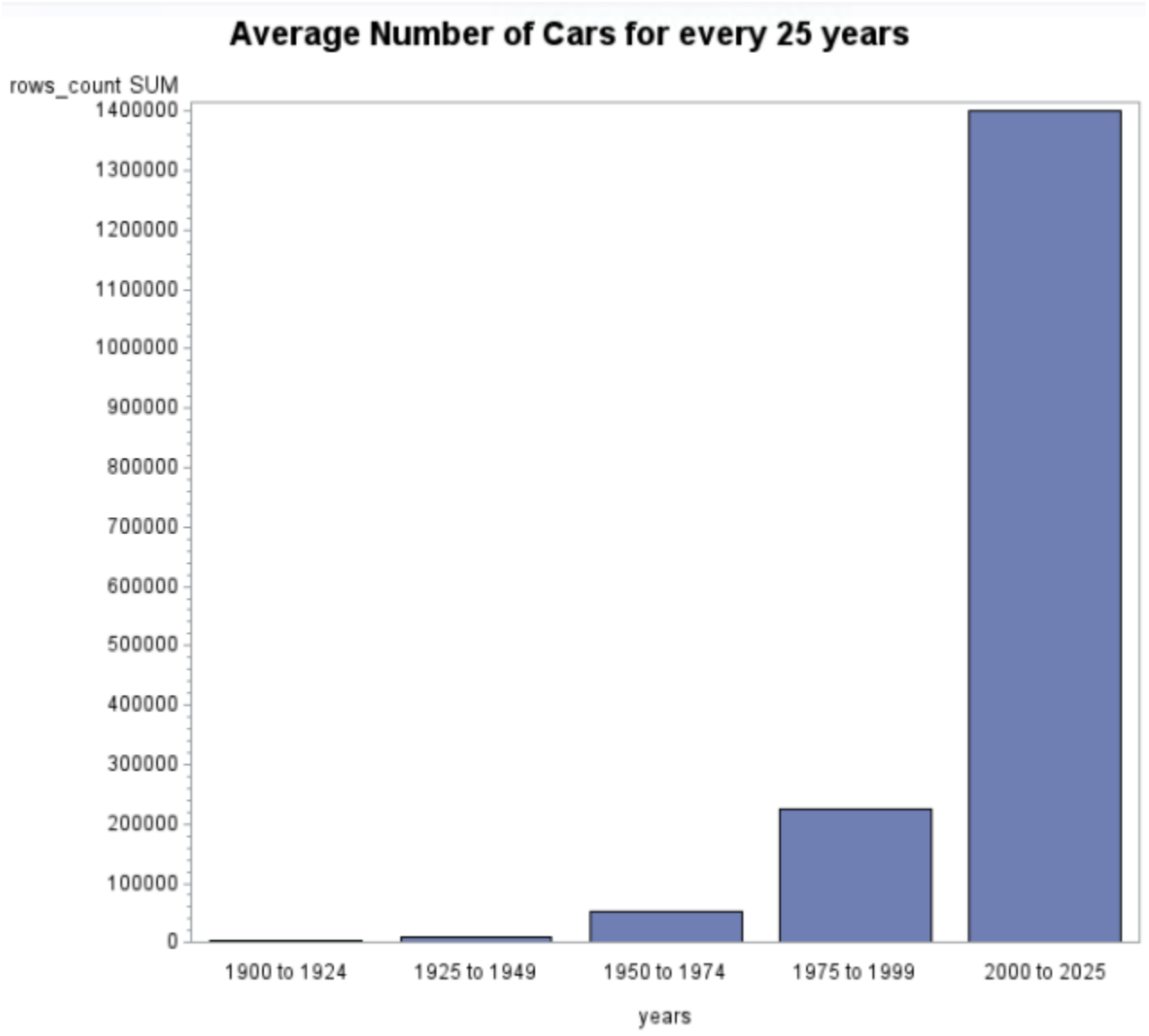
If we look at the sg plot to analyze the relationship between odometer and year, fewer data points can be seen up until 1980 but after that we can see many data points and larger distances driven. It can be said that the use of cars was greater, and people started travelling larger distances after the year 1980 based on our dataset. It could also be said that more efficient cars were manufactured after 1980 which resulted in larger distance driven.

	years	average_price
1	1900 to 1924	12331.988215
2	1925 to 1949	15369.404131
3	1950 to 1974	12674.185568
4	1975 to 1999	4615.9037585
5	2000 to 2017	11456.268092



We have also looked at the average prices of the cars by grouping the number of observations on years into an interval of 25 years, as a result we got 5 groups. In the group 1900 to 1925, the average price of the cars manufactured then was 12331 and between 2000 to 2017, the highest average price of 11456 was observed for the cars.

	years	rows_count
1	1900 to 1924	594
2	1925 to 1949	10071
3	1950 to 1974	52881
4	1975 to 1999	223677
5	2000 to 2025	1399439



While grouping the number of cars manufactured every 25 years we see shows that only 594 cars were manufactured between 1900 and 1924. The largest number of cars manufactured, 1399439 cars, were made between 2000 to 2025 where the actual automobile boom occurred.

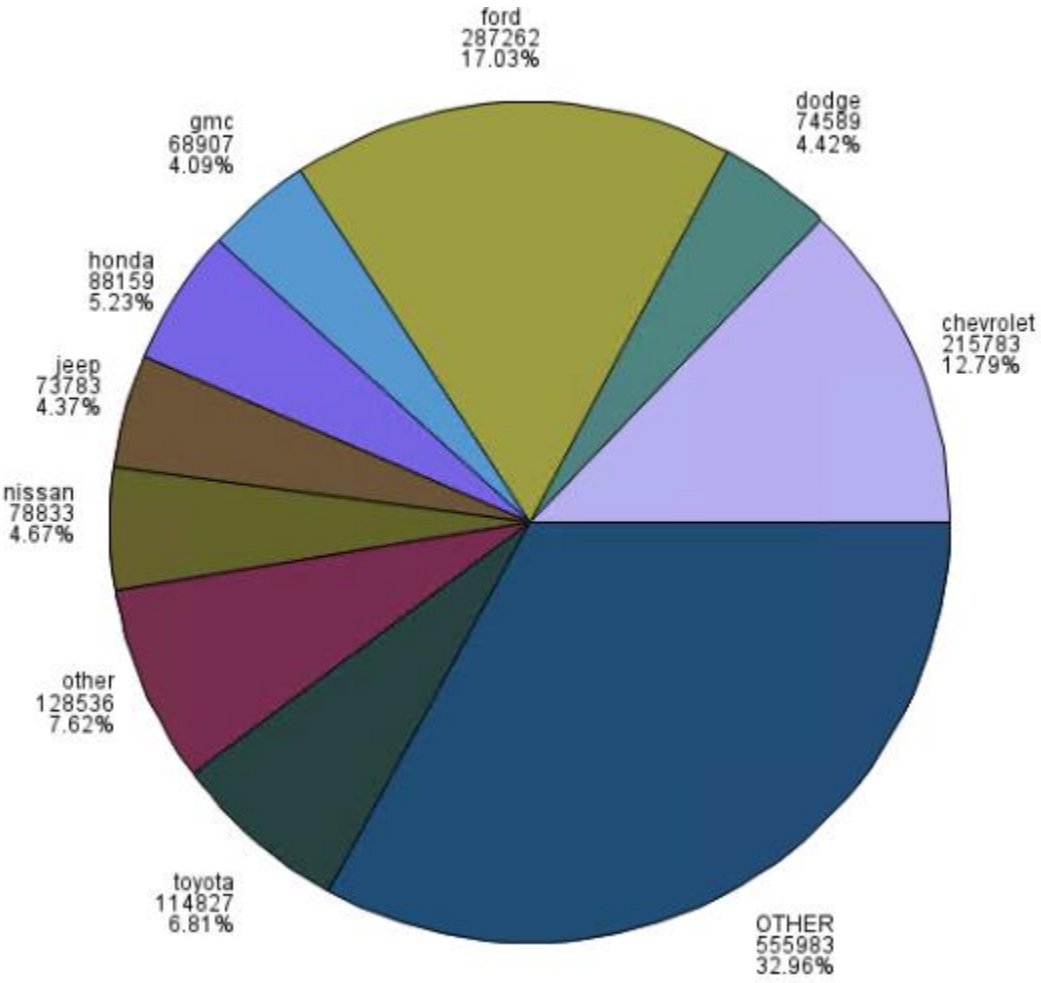
Analysis of the variable: City

city	Frequency	Percent	Cumulative Frequency	Cumulative Percent
springfiel	10124	0.61	10124	0.61
charleston	8146	0.49	18270	1.10
cosprings	8126	0.49	26396	1.58
sfbay	8089	0.49	34485	2.07
anchorage	8077	0.48	42562	2.55
miami	8045	0.48	50607	3.04
sacramento	8027	0.48	58634	3.52
grandrapid	7978	0.48	66612	4.00
orangecoun	7924	0.48	74536	4.47
losangeles	7908	0.47	82444	4.94
omaha	7900	0.47	90344	5.42
boise	7880	0.47	98224	5.89
nashville	7880	0.47	106104	6.36
hartford	7825	0.47	113929	6.83
boston	7824	0.47	121753	7.30

The first column in Table illustrates Manufactured City of a vehicle. The second column in the table denotes the frequency distribution of the model and we used this value to determine how the city of the vehicle is classified. The third column in this table includes the percentage of the distribution. The fourth column in the table provides the Cumulative frequency i.e the analysis of the frequency of occurrence of values of a phenomenon less than a reference value. The phenomenon may be time or space-dependent. And lastly the fifth column represents the cumulative percent which adds a percentage from one period to the percentage of another period. This calculation is important in statistics because it shows how the percentages add together over a period of time. We can find the cumulative percentage by dividing the number of times an event occurs by the total sample size. There is also a row which tells us about the missing frequency.

Exploring the City variable we see that the highest number of cars were recorded in city Springfield with 0.61 % of the total with 10124 frequency, followed by Charleston, Co springs, SF bay with a 0.49% having 8146, 8126, 8089 frequencies respectively and few of lowest recorded cities were Pei, Sherbrooke, Sudbury, Owen Sound, Newfoundla, Yellowknif, Territorie all put together gives almost a negligible percentage contribution to total with frequencies 25, 22, 20, 13, 11, 3, 1 respectively.

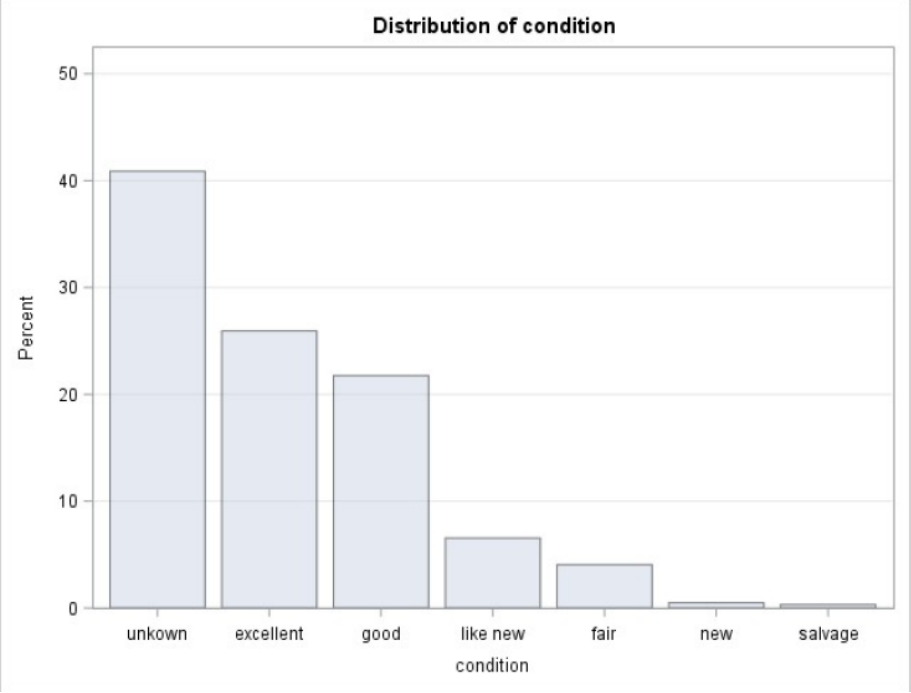
Analysis of the variable: Manufacturer



The above plot shows the distribution of manufacturer for given vehicle. As can be seen above, approximately 50% of cars are manufactured by Ford, Chevrolet, other, Toyota, Honda with Ford being the highest with 16.96% and 282690 frequency, followed by Chevrolet with 12.77% having 21296 frequency and remaining three of top 5 contribute around 20% with 7.64%,6.73%,5.20% having frequencies 127433, 112277, 86710 respectively. Manufacturers such as Morgan, Mercedes Benz, Noble, Hennessey having frequencies 10,4,2,1 all put together does not even contribute around 0.01% of the total manufacturers and Hennessey Being The Lowest recorded with Frequency of 1 with negligible percent of given dataset.

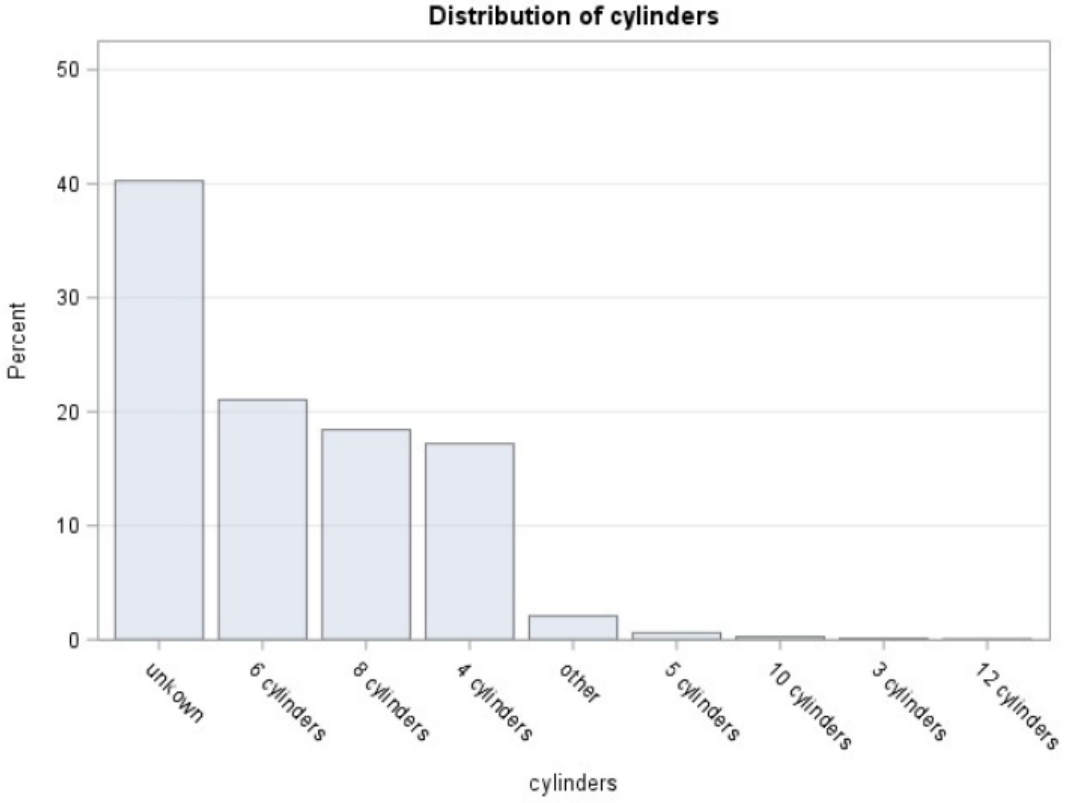
Analysis of the variable: Condition

condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
unknown	681599	40.88	681599	40.88
excellent	432373	25.93	1113972	66.81
good	362749	21.76	1476721	88.57
like new	109193	6.55	1585914	95.12
fair	67658	4.06	1653572	99.18
new	8312	0.50	1661884	99.68
salvage	5371	0.32	1667255	100.00



The below plot shows distribution of condition of a car. As can be seen above, there is a lot of empty records, showing that the condition of the car is unknown and with the top 3 Excellent, Good and Like New it totally contributes the most with cumulative percent of 88.57, Most number of cars were in unknown condition with 40.88% having 681599 frequency sits on top among all conditions we are given with, followed by cars with excellent condition with 25.93% having 432373 frequency then good condition with 21.76% having 362749 frequency. Remaining all types Like New, Fair, New, Salvage all put together consists of 11% of the total distribution of conditions having 109193, 67658, 8312, 5371 frequencies with percentages 6.55,4.06,0.50,0.32 respectively

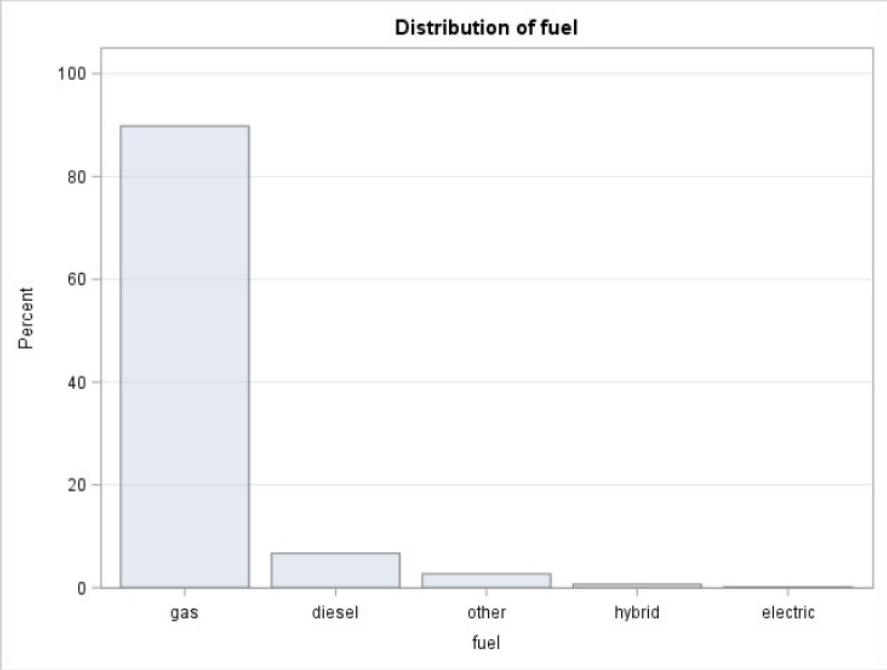
Analysis of the variable: Cylinders



The above plot shows the distribution of cylinders for given vehicles. For this variable as well we can see that there are a lot of unknown records. Of all 9 types of cylinders mentioned, top 3 types of cylinders which has the major contribution among the different types of cylinders are 6 Cylinders,8 Cylinders,4 Cylinders contribute with 96.91% and unknown stands top among all types with 40.26% having 671192 frequency of total and the rest of top 4 having frequencies 350872, 306983, 286669 with 21.04%, 18.41%, 17.19% respectively. The Remaining types namely Other, 5-Cylinders, 10-Cylinders, 3-Cylinders, 12-Cylinders collectively contribute about 3% of the total types of cylinders with respective frequencies 34624, 10063, 4345, 1745, 753.

Analysis of the variable: Fuel

fuel	Frequency	Percent	Cumulative Frequency	Cumulative Percent
gas	1488693	89.83	1488693	89.83
diesel	110543	6.67	1599236	96.51
other	44857	2.71	1644093	99.21
hybrid	10679	0.64	1654772	99.86
electric	2379	0.14	1657151	100.00
Frequency Missing = 10104				

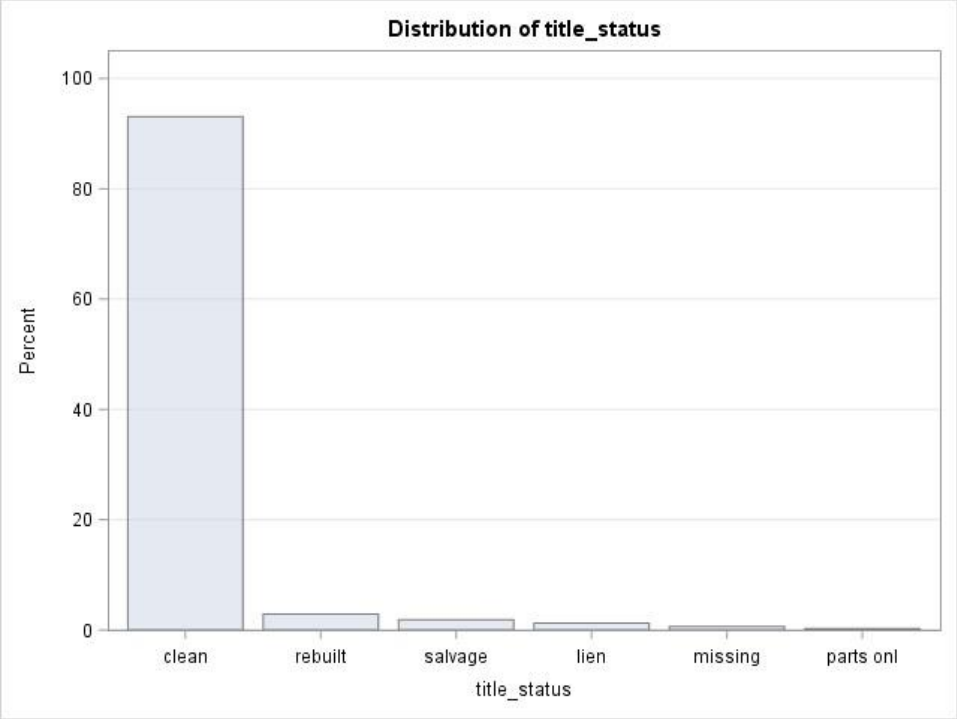


The Above graph depicts the distribution of fuel for the given vehicles. There are 5 types under attribute fuel namely Gas, Diesel, Other, Hybrid, Electric. Out of all types mentioned there is only one type of fuel that most cars have, Gas with 89.83% having 1488693 Frequency. Followed By Diesel With 6.67% having 110543 frequency and the of the fuel types with 2.71% ,0.64% and 0.14% having frequencies 44857,10679,2379 respectively. This trend may change in the future since many hybrid and electrical cars are being manufactured that will also help the environment.

Also, there are a total of 10104 frequencies which are missing from the dataset which are negligible.

Analysis of the variable: Title_status

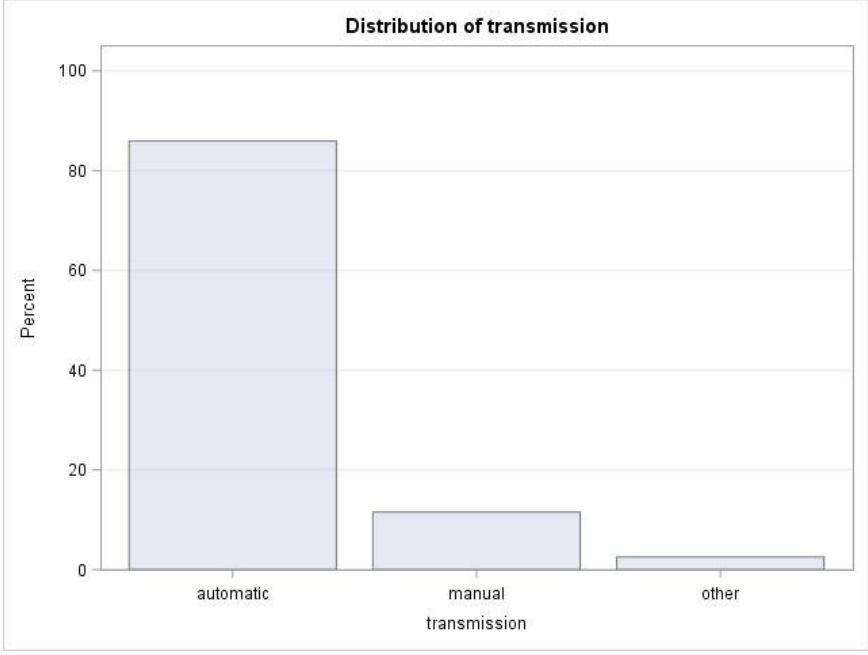
title_status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
clean	1549468	93.08	1549468	93.08
rebuilt	48261	2.90	1597729	95.98
salvage	31143	1.87	1628872	97.85
lien	21017	1.26	1649889	99.11
missing	10554	0.63	1660443	99.74
parts onl	4291	0.26	1664734	100.00
Frequency Missing = 2521				



As we can depict from the figure, maximum of the cars are clean, 93.08 % of the total titles were categorized as a clean title and having the highest number of frequency at 1549468 followed by rebuilt title with a 2.90% with a major margin difference in frequency by 48261. Salvage title, lien title, missing title and parts only title all put together consists of 4% of the titles with 1.87% ,1.26% ,0.63% and 0.26% respectively. While the cumulative frequency of the title status is 1664734 and there are 2521 frequencies which are missing from the dataset. This gives us an idea that people tend to keep their car clean to attract more buyers.

Analysis of the variable: Transmission

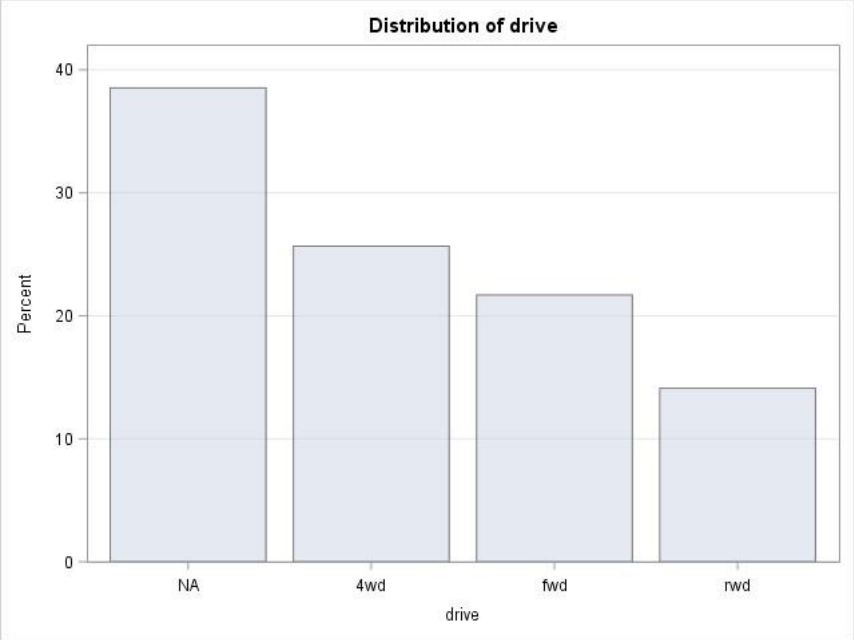
transmission	Frequency	Percent	Cumulative Frequency	Cumulative Percent
automatic	1432601	85.93	1432601	85.93
manual	192394	11.54	1624995	97.47
other	42260	2.53	1667255	100.00



We can observe from the figure that automatic transmission stands at 85.93% whereas manual transmission is at 11.54% holding a difference of 74.39% between automatic and manual transmission. Lastly other types are at 2.53% being the lowest amongst all and a less popular option when compared to others. This gives us an insight that American people like to but cars with automatic transmission more as compared to the rest of the world. In India, for example people prefer manual transmission cars than automatic transmission cars, but there are several factors such as the traffic and road types that determine the preference of transmission type.

Analysis of the variable: Drive

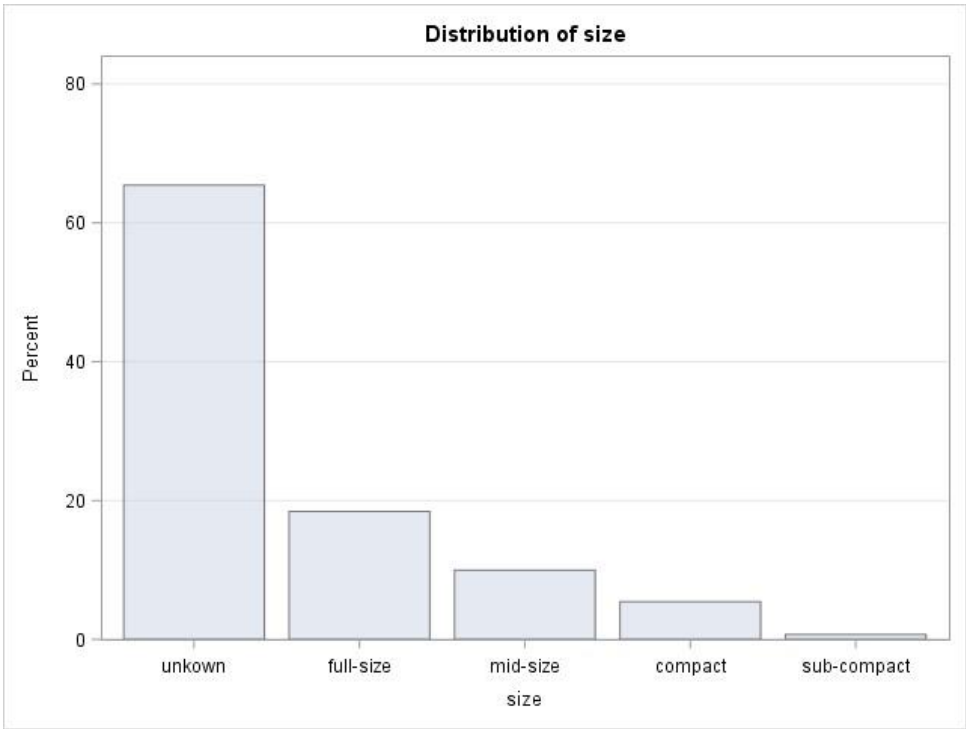
drive	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NA	642417	38.53	642417	38.53
4wd	427907	25.67	1070324	64.20
fwd	361727	21.70	1432051	85.89
rwd	235204	14.11	1667255	100.00



Drive type of a car is also a variable with high number of empty records labelled as 'NA'. In America, a vast majority of cars were rear-wheel drive through the 1970s. The configuration is becoming increasingly rare in 2019. We can interpret from the figure that 642417 number of vehicles drive mechanism is not available which consists of 38.53 % of the total data. The most used driving mechanism is the four-wheel drive which consists of a quarter of all the vehicles i.e 25.67% with an exact figure of 427907. The second most preferred is the front wheel drive which is a little less than the four wheel drive at 21.70 % by 361727.The least preferred one is the rear wheel drive mechanism with a 14.11 % of the total at 235204.

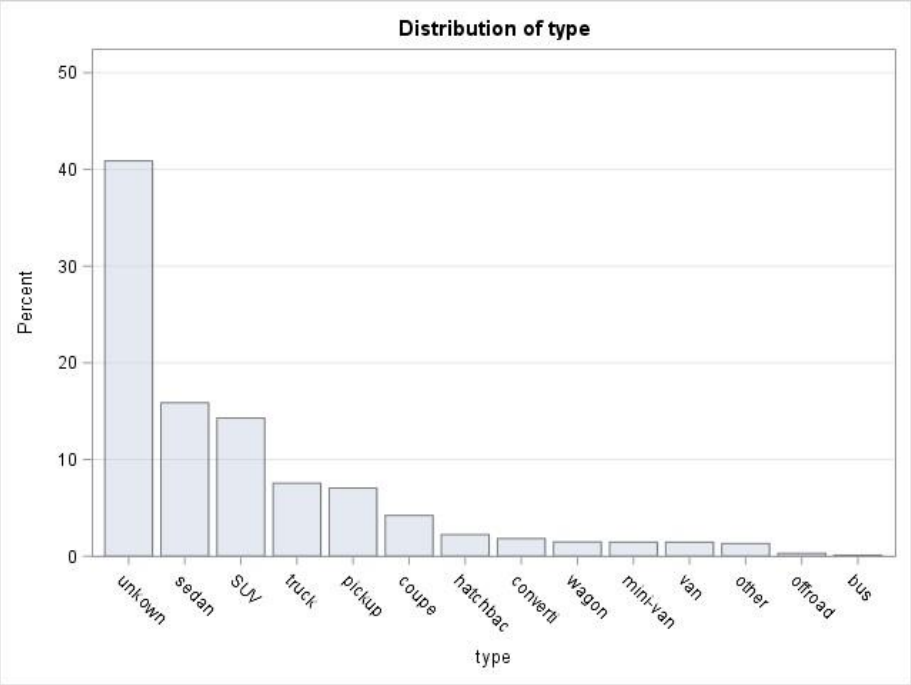
Analysis of the variable: Size

size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
unkown	1090932	65.43	1090932	65.43
full-size	307432	18.44	1398364	83.87
mid-size	166436	9.98	1564800	93.85
compact	90585	5.43	1655385	99.29
sub-compact	11870	0.71	1667255	100.00



We can interpret from the figure that the majority of the cars belong to the unknown category with 1090932 of the vehicles. This category consists of empty values 65.43% of the total data available. The most used car size is the full size which consists 18.44% with an exact figure of 307432.The second most preferred is the mid-size vehicles which is almost half of the full-size vehicles at 9.98% by 166436. The next preferred size is the compact size vehicle with a 5.43 % of the total at 90585. Lastly comes the sub-compact vehicle which is less than a percent of the total vehicles at 11870 vehicles by 0.71%. This makes us understand that the majority of American people tend to prefer full-size vehicles than compact or sub-compact sized vehicle for road travel.

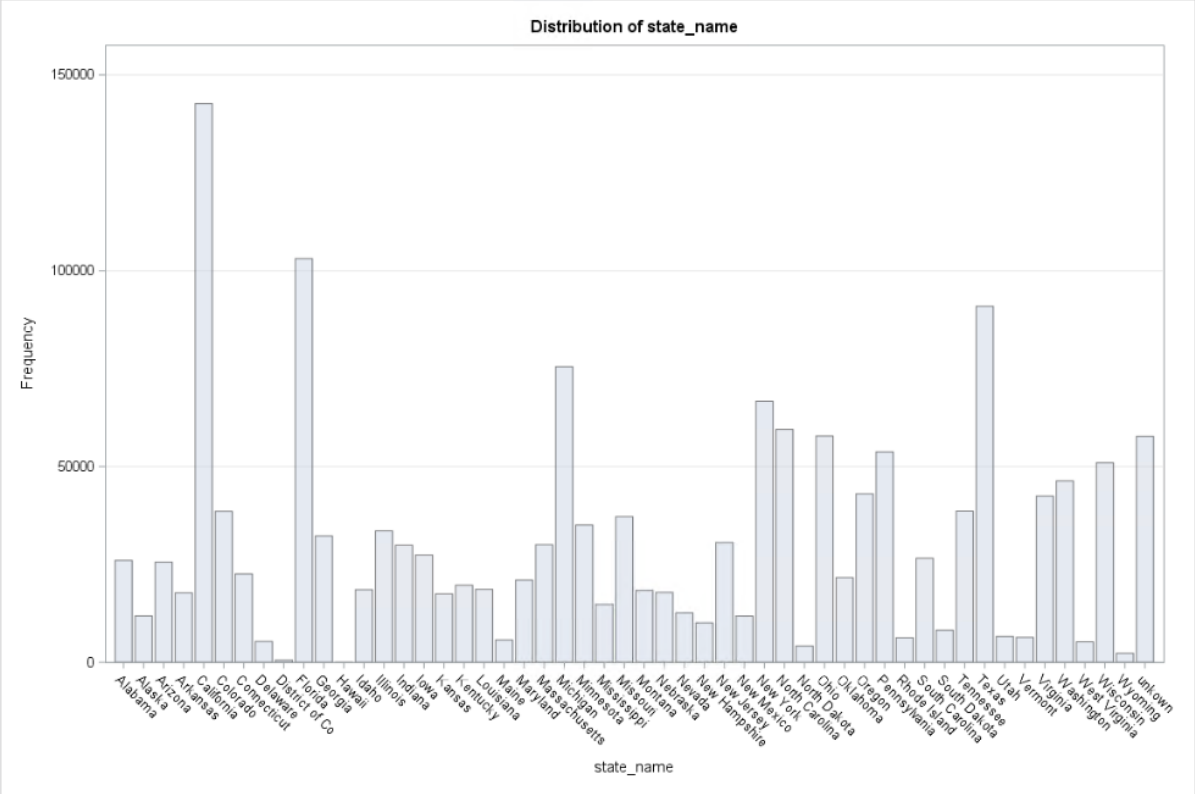
Analysis of the variable: Type



From the figure we understand that most of the car type records have unknown values. This unknown category consists of 40.88 % of the total data available. The most used car size is the sedan which consists 15.87 % with an exact figure of 264610.The second most preferred is the SUV vehicles which is almost similar to the sedan by a difference of 1.59% at 14.28 % by 238104. The next preferred vehicle type are the trucks and then the pickup vehicle with a 7.55 % and 7.04% of the total at 125882 and 117329.A few of the vehicle types like the convertible, wagon, mini-van, van and off-road vehicles are around 1% of all the vehicles each ranging from 30421 to 4923 vehicles respectively. The bus being the least owned vehicle type with a mere 0.11% of the vehicles resulting in 1815 busses in total. This also helps us understand that Sedans are highly preferred among the American population.

Analysis of the variable: State_name

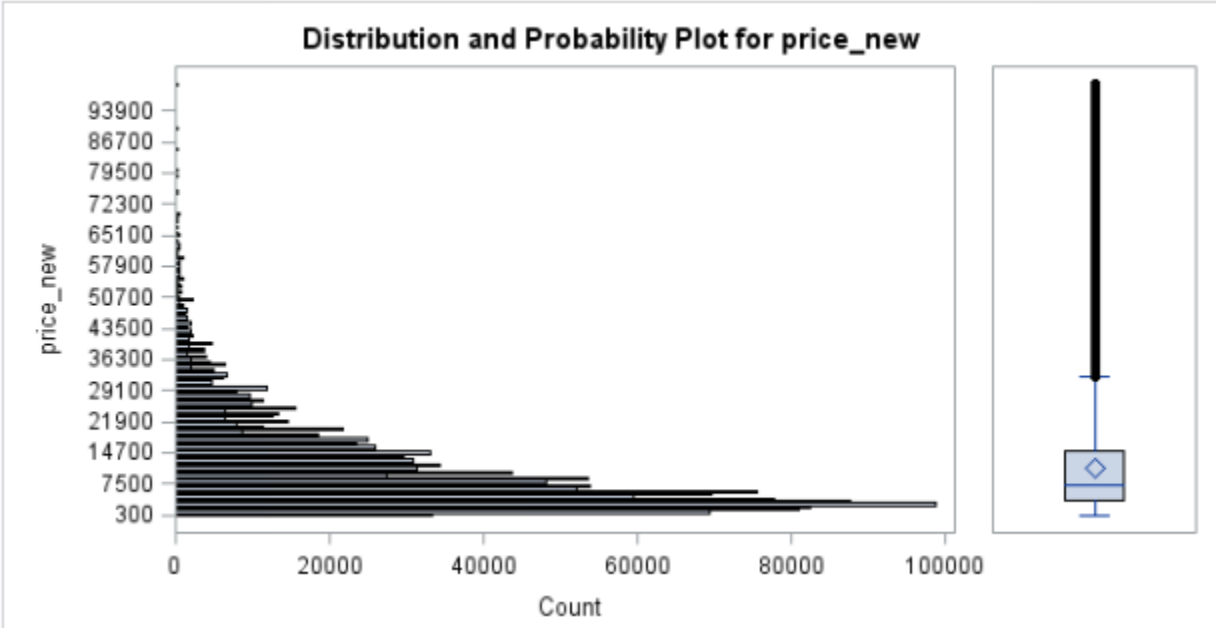
state_name	Frequency	Percent	Cumulative Frequency	Cumulative Percent
California	142670	8.88	142670	8.88
Florida	103087	6.41	245757	15.29
Texas	90933	5.66	336690	20.95
Michigan	75490	4.70	412180	25.65
New York	66663	4.15	478843	29.80
North Carolina	59480	3.70	538323	33.50
Ohio	57794	3.60	596117	37.09
unkown	57716	3.59	653833	40.68
Pennsylvania	53745	3.34	707578	44.03
Wisconsin	50958	3.17	758536	47.20
Washington	46335	2.88	804871	50.08
Oregon	43026	2.68	847897	52.76
Virginia	42430	2.64	890327	55.40



There are a total of 51 states including an unknown value. California being on top has the highest number of vehicles at 149901 by a percentage of 8.99%. It is followed by Florida by 105801 at 6.35%.The District of Columbia and Hawaii are the least of them, 567 and 22 vehicles each. The presence of Hawaii hardly makes any difference as it is 0.00% of the total data.

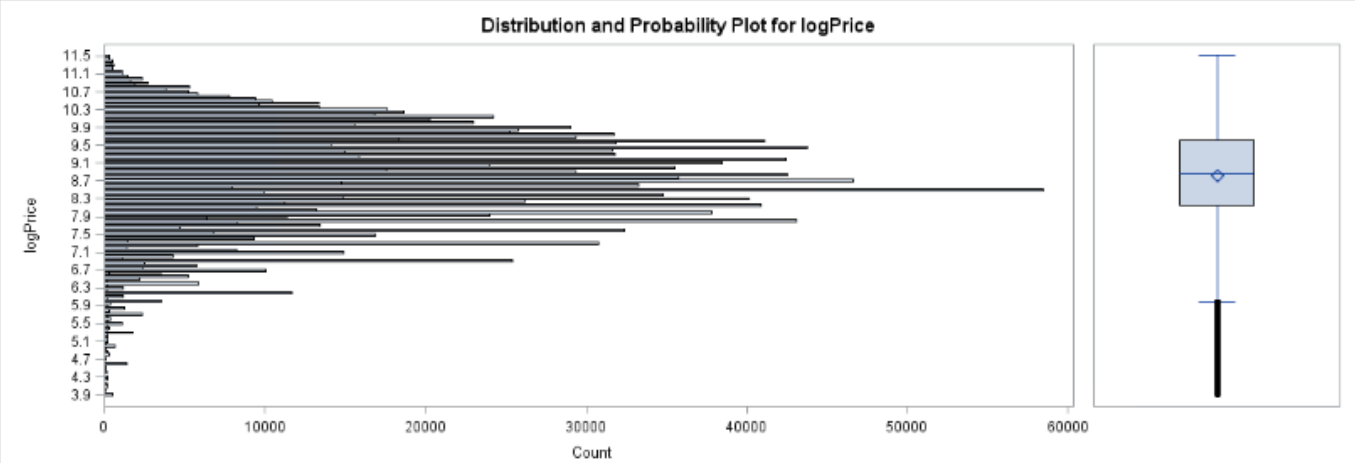
PREDICTION MODEL FOR PRICE

Basic Statistical Measures			
Location		Variability	
Mean	10948.51	Std Deviation	10852
Median	7200.00	Variance	117757584
Mode	2500.00	Range	99950
		Interquartile Range	11501



By running the basic PROC UNIVARITE function on the Price variable we get the following data as seen above. The average price for a car is close to eleven thousand dollars, with many cars having the price-tag of 2500$. We also notice a lot of cars within the 300$ to 20,000$ range.

Basic Statistical Measures			
Location		Variability	
Mean	8.816837	Std Deviation	1.07516
Median	8.881836	Variance	1.15598
Mode	7.824046	Range	7.60090
		Interquartile Range	1.45557



We convert the Price variable to the its log values and here we can immediately observe the values are more normally distributed with much less variance and range. This in turn will help to get a develop a better and accurate model.

A multiple linear regression model was created using PROC REG with price of the car as independent variable and the independent variables as odometer, year, fuel, transmission, size, cylinders, condition and drive. The adjuster r-squared of the model is 0.4981. The regression output of the model using SAS is,

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	6.946663E 13	3.020288E 12	49217.8	<.0001
Error	1.14E 6	6.999185E 13	61365734		
Corrected Total	1.14E 6	1.394585E 14			

Root MSE	7833.62843	R-Square	0.4981
Dependent Mean	12312	Adj R-Sq	0.4981
Coeff Var	63.62349		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-663043	1978.45170	-335.13	<.0001
odometer_new	1	-0.06867	0.00012403	-553.61	<.0001
year_new	1	342.94502	0.98019	349.88	<.0001
fuel_gas	1	-1533.02388	38.74632	-39.57	<.0001
fuel_hybrid	1	-137.44250	92.86154	-1.48	0.1389
fuel_diesel	1	9620.56389	47.32180	203.30	<.0001

transmission_automatic	1	-1612.94528	48.93238	-32.96	<.0001
transmission_manual	1	162.53196	54.78412	2.97	0.0030
size_fullsize	1	-344.82220	21.73954	-15.86	<.0001
size_midsize	1	-437.72071	25.30161	-17.30	<.0001
size_compact	1	-877.50031	32.29872	-27.17	<.0001
size_subcompact	1	-1213.30399	78.84127	-15.39	<.0001
cylinders_other	1	-634.63271	41.97864	-15.12	<.0001
cylinders_6	1	-1709.73161	40.33407	-42.39	<.0001
cylinders_4	1	-3865.31247	41.44373	-93.27	<.0001
cylinders_8	1	1057.46228	42.11168	25.11	<.0001
condition_excellent	1	-3429.49003	90.81906	-37.76	<.0001
condition_likenew	1	-1182.60052	93.80259	-12.61	<.0001
condition_fair	1	-6224.77500	97.75900	-63.67	<.0001
condition_good	1	-5329.71786	91.21635	-58.43	<.0001
condition_unkown	1	-2223.58413	91.52600	-24.29	<.0001
drive_rwd	1	2478.08776	27.36963	90.54	<.0001
drive_fwd	1	-1445.00139	23.55415	-61.35	<.0001
drive_4wd	1	4428.24928	22.51625	196.67	<.0001

The variance of the variable price was not uniform when it was plotted against year. We took a log of price and ran a separate model and we have interpreted it below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	607455	26411	59088.7	<.0001
Error	1.14E6	509804	0.44697		
Corrected Total	1.14E6	1117260			

Root MSE	0.66856	R-Square	0.5437
Dependent Mean	9.00729	Adj R-Sq	0.5437
Coeff Var	7.42244		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-59.23019	0.16885	-350.78	<.0001
odometer_new	1	-0.00000555	1.058564E-8	-523.94	<.0001
year_new	1	0.03421	0.00008365	408.95	<.0001
fuel_gas	1	-0.16022	0.00331	-48.45	<.0001
fuel_hybrid	1	0.08455	0.00793	10.67	<.0001
fuel_diesel	1	0.64084	0.00404	158.68	<.0001
transmission_automatic	1	-0.22295	0.00418	-53.39	<.0001
transmission_manual	1	-0.07523	0.00468	-16.09	<.0001

size_fullsize	1	-0.01140	0.00186	-6.15	<.0001
size_midsize	1	-0.05088	0.00216	-23.56	<.0001
size_compact	1	-0.11590	0.00276	-42.05	<.0001
size_subcompact	1	-0.10776	0.00673	-16.02	<.0001
cylinders_other	1	-0.06744	0.00358	-18.82	<.0001
cylinders_6	1	-0.19223	0.00344	-55.84	<.0001
cylinders_4	1	-0.31439	0.00354	-88.89	<.0001
cylinders_8	1	0.10409	0.00359	28.96	<.0001
condition_excellent	1	0.54335	0.00775	70.10	<.0001
condition_likenew	1	0.62248	0.00801	77.76	<.0001
condition_fair	1	-0.47524	0.00834	-56.96	<.0001
condition_good	1	0.21889	0.00778	28.12	<.0001
condition_unkown	1	0.57257	0.00781	73.30	<.0001
drive_rwd	1	0.30993	0.00234	132.68	<.0001
drive_fwd	1	-0.06078	0.00201	-30.23	<.0001
drive_4wd	1	0.42943	0.00192	223.47	<.0001

We have observed that the accuracy of the model has improved, but to make sure that the model we have achieved is optimal we have further run stepwise regression on the model and the following output is obtained.

Stepwise Model Selection for Price with logPrice as response variable

The GLMSELECT Procedure

Stepwise Selection Summary

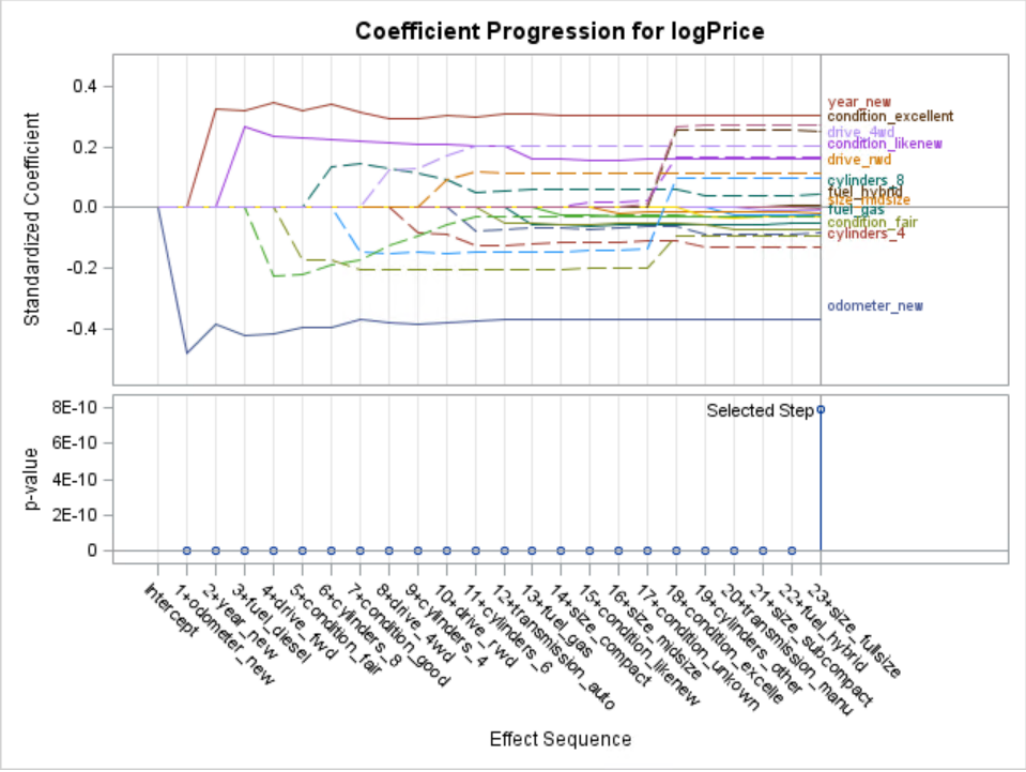
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	odometer_new		2	341128	<.0001
2	year_new		3	166335	<.0001
3	fuel_diesel		4	130463	<.0001
4	drive_fwd		5	101176	<.0001
5	condition_fair		6	61288.3	<.0001
6	cylinders_8		7	36913.0	<.0001
7	condition_good		8	44379.0	<.0001
8	drive_4wd		9	27978.6	<.0001
9	cylinders_4		10	12858.8	<.0001
10	drive_rwd		11	12818.0	<.0001
11	cylinders_6		12	9053.32	<.0001
12	transmission_automat		13	6238.19	<.0001
13	fuel_gas		14	3413.31	<.0001
14	size_compact		15	1565.61	<.0001
15	condition_likenew		16	825.42	<.0001
16	size_midsize		17	611.73	<.0001
17	condition_unkown		18	322.91	<.0001
18	condition_excellent		19	5005.61	<.0001
19	cylinders_other		20	362.03	<.0001

20	transmission_manual		21	268.94	<.0001
21	size_subcompact		22	250.18	<.0001
22	fuel_hybrid		23	114.28	<.0001
23	size_fullsize		24	37.78	<.0001

Selection stopped because all effects are in the final model.

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	23	607455	26411	59088.7
Error	1.14E6	509804	0.44697	
Corrected Total	1.14E6	1117260		

Root MSE	0.66856
Dependent Mean	9.00729
R-Square	0.5437
Adj R-Sq	0.5437
AIC	222150
AICC	222150
SBC	-918158



As we can see the model retains the same accuracy of 54% and all the coefficients relevant to the final model. The Coeffecient Progression for logPrice graph gives us an idea of how the stepwise regression has run.

From the output of the regression model we can interpret that the price of the car decreases by 0.55% for every 1000 miles increase in the odometer reading. The variable is statistically significant with p-value <0.0001

As the car is newer by 1 year the price of the car increases by 3.4%

The price of a hybrid car is 24.44% higher than the price of a gasoline car and the price of the diesel car is 80.1% higher than the gasoline cars. This make sense because most of the diesel vehicles will be commercial vehicles like buses. So, the size might have a bigger impact than the type of fuel in cars.

The price of the manual transmission car is 14.772% greater than the price of an automatic transmission car. Automatic cars are more common in America than the manual transmission vehicles.

The price of a subcompact car is 0.814% higher than the price of a compact car. The price of a mid-size car is 6.5% higher than the price of the compact car. The price of a full-size car is 10.47% higher than the price of a compact car.

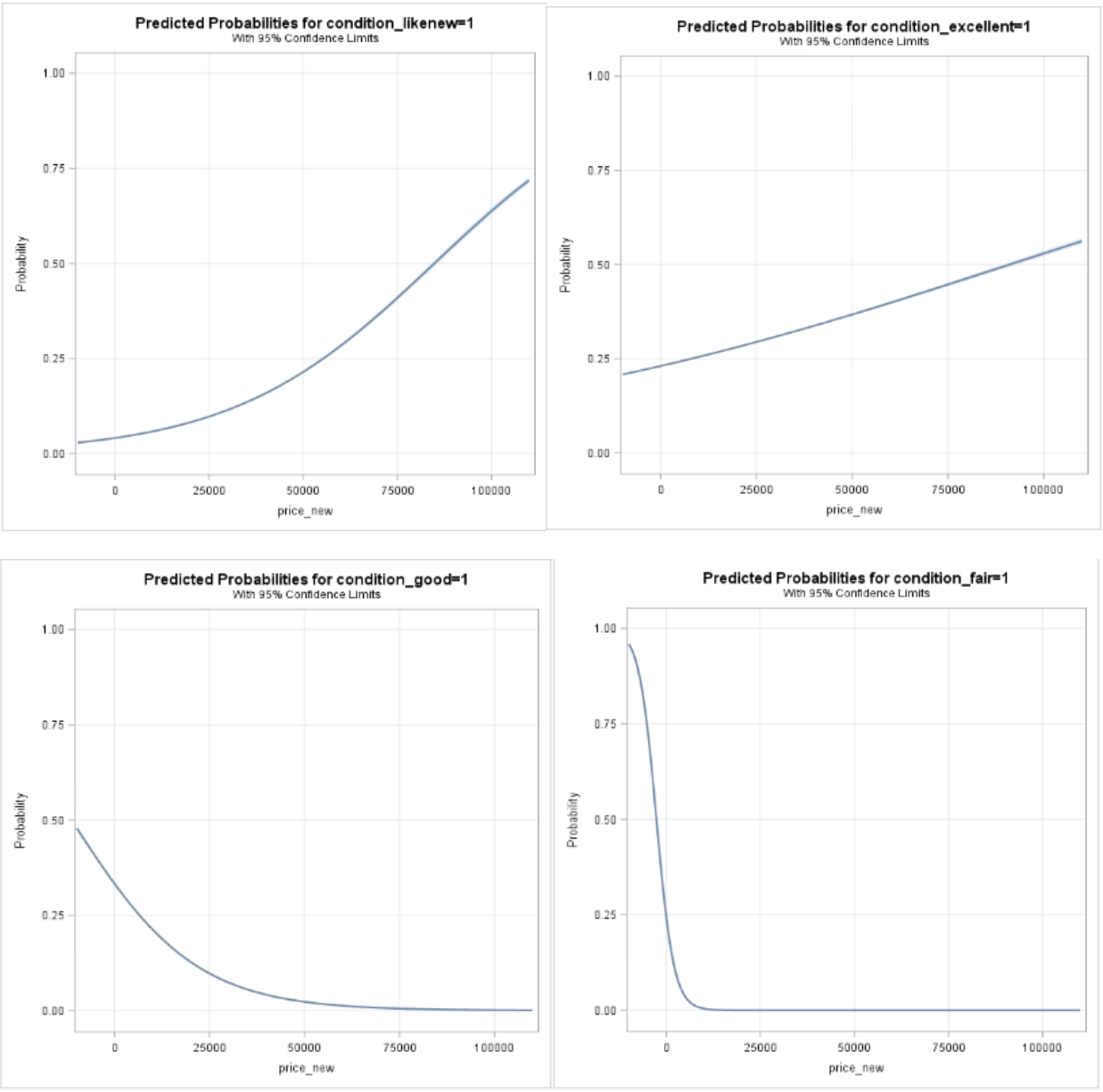
The price of a 6-cylinder vehicle is 12.21% higher than the price of a 4-cylinder vehicle and the price of an 8-cylinder vehicle is 41.84% higher than the price of a 4-cylinder vehicle. This is because bigger vehicles which costs more has a greater number of cylinders as they require higher horse power.

The cars which are like new are expensive than the cars that are fare by 109.7%. The cars which are excellent are 7.88% cheaper than the cars which are like new. The cars which are in good condition are % cheaper than the cars that are like new 32.45%.

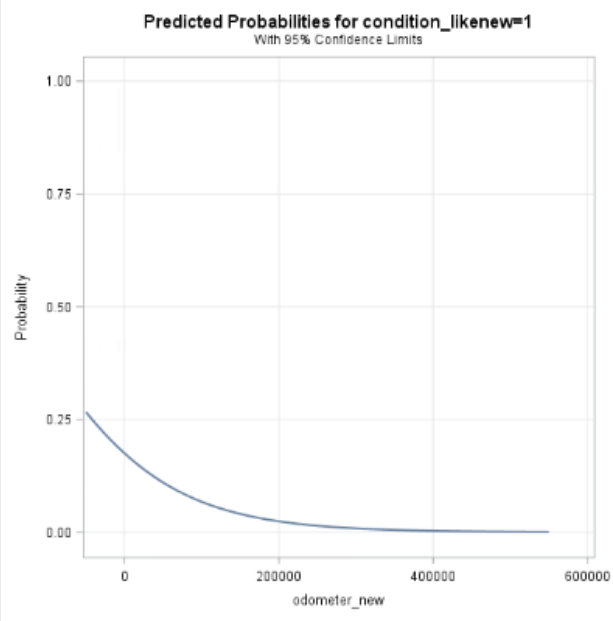
The price of a rear wheel drive is 36.99% higher than the price of a front wheel drive car. The reason may be most of the rear wheel drive vehicles might be an SUV or larger. The price of a 4-wheel drive car is 49% higher than the price of a front wheel drive car. Most of the rear wheel and 4-wheel drive vehicles might be large SUVs and commercial vehicles.

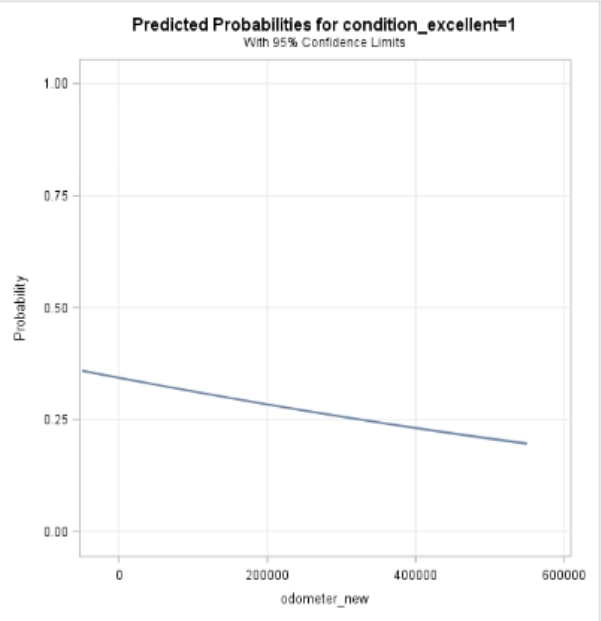
All the variables are statistically significant in the model with p-values of all the variables being <0.0001, even being significant at 1% significance level.

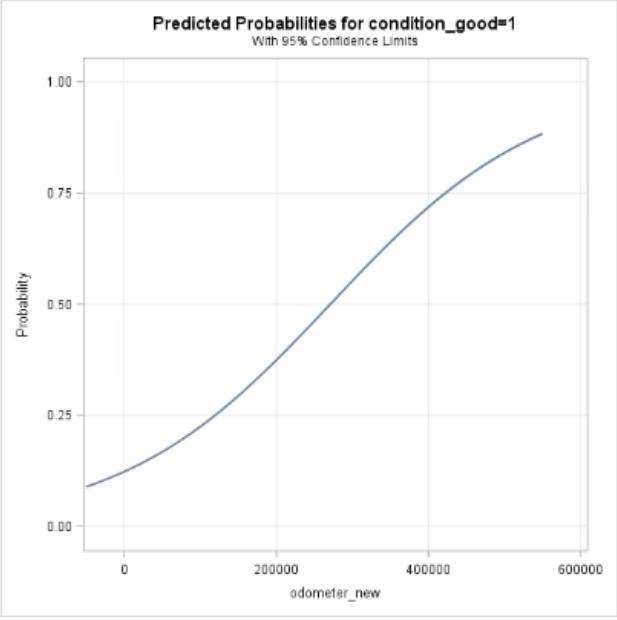
LOGISITIC MODELING FOR CONDITION

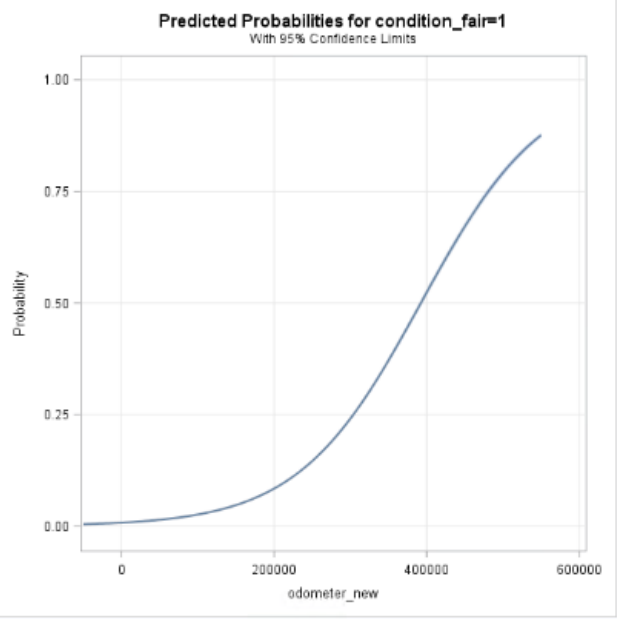


Running 4 logistic models we get the graphs above with predicted probabilities. The logistic function was such that the response variable is the specific condition of the car and the independent variable is the price. The 4 models correspond to the 4 conditions of the car, namely like new, excellent, good and fair. As seen from the graphs we see that the probability of a car being expensive increases if the condition is like new or excellent and the probability of a car being cheaper is more when the condition is good and fair. From the nature of the slopes of the graphs we understand how the probability of the condition of the car depends on the pricing of the car.









Similar to the previous 4 graphs we have run 4 more regression models where the response variable is condition of the car and the independent variable is the odometer values of each car. From the 4 graphs we can see that the probability of a car having a good or fair condition is much more if the odometer reading is higher and the probability of a car being in a like new or excellent condition is much lower and almost close to 0 when the odometer readings are high. From the nature of the slopes of the graphs we understand how the probability of the condition of the car depends on the odometer reading of a car. This helps to back the results we have obtained from the regression model earlier as well.

CONCLUSION

Based on the craigslist dataset, using SAS we have explored and gathered insights on various variables and how they behave in the real world. Since the data was for more than 100 years, we could also observe the pattern of cars manufactured and used from 1900 to 2019. We also had challenges while processing the data as many of the categorical and numerical values were missing. This might have affected the accuracy of the model to some extent but since we had large data, we were able to develop in a decent model that best reflects the actual scenario. The multilinear model that we created could be used for the 2 possible scenarios; the buyer and the seller perspective, that have been described in our problem statement. The price of a car depends on not just the variables specified in the model but it is also affected by various other variables that are not easy to capture. Our logit model has also helped in deriving conclusions and supports the multilinear model developed. Hence, making these models a base for making decisions for a given factors or conditions.

SOURCES

Books :

- Lora D. Delwich and Susan J. Slaughter. The Little SAS Book: A Primer, Fifth Edition. SAS Institute.
- Jeffrey M. Wooldridge. Introductory Econometrics: A Modern Approach, 6th edition. South-Western College Pub.

Websites :

- https://support.sas.com/documentation/
- https://www.kaggle.com/austinreese/craigslist-carstrucks-data#craigslistVehicles.csv
- https://www.wikipedia.org/

APPENDIX

```
/* Importing the required dataset */
proc import out= WORK.craig
     datafile= "H:\craig1.csv"
     dbms=CSV REPLACE;
     getnames=YES;
     datarow=2;
run;

proc contents data=craig;
run;

/* Convert character to numeric type */
data craig1;
 set craig;
 odometer_new=input(odometer,best12.);
 price_new=input(price,best12.);
 weather_new=input(weather,best12.);
 year_new=input(year,best12.);
run;

proc contents data=craig1;
run;

/* Cleaning up price variable */
data craig_p1;
      set craig1;
      If price_new < 50 then delete;
run;

data craig_p2;
    set craig_p1;
      If price_new > 100000 then delete;
run;

/* Imputing Mean Values for Price */
proc stdize data=craig_p2 reponly method=mean out=craig_p1;
var price_new;
run;

/* Cleaning up odometer variable */
data craig_o1;
      set craig_p2;
      if odometer_new > 500000 then delete;
run;

/* Cleaning up year variable */
data craig_year;
      set craig_o1;
      If year_new <= 1900 then delete;
run;

data craig_year1;
      set craig_year;
      If year_new => 2020 then delete;
run;
```

```
/* Filling up missing values of categorical variables */
data craig_cat;
      set craig_year;
      If state_name = "FAILED" then state_name="unkown";
      If state_name = " " then state_name="unkown";
      If manufacturer = " " then manufacturer="other";
      If transmission = " " then transmission="other";
      If cylinders = " " then cylinders="unkown";
      If condition = " " then condition="unkown";
      If size = " " then size="unkown";
      If type = " " then type="unkown";
      If paint_color = " " then paint_color="unkown";
      If drive = " " then drive="NA";
run;

/* Generating Dummy variables for required columns */
DATA craig_dummy;
      SET craig_cat ;
IF fuel = "gas" THEN fuel_gas = 1;
        ELSE fuel_gas = 0;
IF fuel = "hybrid" THEN fuel_hybrid = 1;
        ELSE fuel_hybrid = 0;
IF fuel = "diesel" THEN fuel_diesel = 1;
        ELSE fuel_diesel = 0;
IF transmission = "automatic" THEN transmission_automatic = 1;
        ELSE transmission_automatic = 0;
IF transmission = "manual" THEN transmission_manual = 1;
        ELSE transmission_manual = 0;
IF size = "full-size" THEN size_fullsize = 1;
        ELSE size_fullsize = 0;
IF size = "mid-size" THEN size_midsize = 1;
        ELSE size_midsize = 0;
IF size = "compact" THEN size_compact = 1;
        ELSE size_compact = 0;
IF size = "sub-compact" THEN size_subcompact = 1;
        ELSE size_subcompact = 0;
IF size = "unkown" THEN size_unkown = 1;
        ELSE size_unkown = 0;
IF cylinders = "6 cylinders" THEN cylinders_6 = 1;
        ELSE cylinders_6 = 0;
IF cylinders = "4 cylinders" THEN cylinders_4 = 1;
        ELSE cylinders_4 = 0;
IF cylinders = "8 cylinders" THEN cylinders_8 = 1;
        ELSE cylinders_8 = 0;
IF cylinders = "other" THEN cylinders_other = 1;
        ELSE cylinders_other = 0;
IF cylinders = "unkown" THEN cylinders_other = 1;
        ELSE cylinders_other = 0;
IF condition = "excellent" THEN condition_excellent = 1;
        ELSE condition_excellent = 0;
IF condition = "good" THEN condition_good = 1;
        ELSE condition_good = 0;
IF condition = "like new" THEN condition_likenew = 1;
        ELSE condition_likenew = 0;
IF condition = "fair" THEN condition_fair = 1;
        ELSE condition_fair = 0;
```

```
IF condition = "unkown" THEN condition_unkown = 1;
        ELSE condition_unkown = 0;
IF drive = "4wd" THEN drive_4wd = 1;
        ELSE drive_4wd = 0;
IF drive = "rwd" THEN drive_rwd = 1;
        ELSE drive_rwd = 0;
IF drive = "fwd" THEN drive_fwd = 1;
        ELSE drive_fwd = 0;
IF drive = "NA" THEN drive_na = 1;
        ELSE drive_na = 0;
RUN;

/* Generating logPrice variable */
data craig_lprice;
    set craig_dummy;
      logPrice = log(price_new);
      run;

/* Exploratory Data Analysis */
/* EDA for Price */
proc sgplot data=craig_lprice;
  title "Relationship between logPrice and Odometer";
  scatter y=logPrice x=odometer_new;
run;

proc univariate data=craig_lprice plots;
      histogram price_new logPrice / normal kernel;
      inset n mean std / position = ne;
      probplot logPrice;
      var logPrice price_new;
run;

/*EDA for Year*/
proc univariate data=craig_lprice plots;
    title " Exploring Year variable ";
    histogram year_new;
      var year_new;
run;

*Distribution of average price every 25 years;
proc sql;
  create table avg_price as
  select
    case
      when input(year,?4.) between 1900 and 1924 then '1900 to 1924'
      when input(year,?4.) between 1925 and 1949 then '1925 to 1949'
      when input(year,?4.) between 1950 and 1974 then '1950 to 1974'
      when input(year,?4.) between 1975 and 1999 then '1975 to 1999'
      when input(year,?4.) between 2000 and 2017 then '2000 to 2017'
      else 'out of range'
    end
    as years
  , avg(price_new) as average_price
  from work.craig_lprice
  group by years
  having years not in ('out of range');
```

```
quit;

proc gchart data=avg_price;
   vbar years / sumvar=average_price maxis=axis1;
   title "Average Price for every 25 years";
run;

*Distribution of number of cars manufactured every 25 years;
proc sql;
  create table count_cars as
  select
    case
      when input(year,?4.) between 1900 and 1924 then '1900 to 1924'
      when input(year,?4.) between 1925 and 1949 then '1925 to 1949'
      when input(year,?4.) between 1950 and 1974 then '1950 to 1974'
      when input(year,?4.) between 1975 and 1999 then '1975 to 1999'
      when input(year,?4.) between 2000 and 2025 then '2000 to 2025'
      else 'out of range'
    end
    as years,
      count(*) as rows_count
  from work.craig_lprice
  group by years
;
quit;

proc gchart data=count_cars;
   vbar years / sumvar=rows_count maxis=axis1;
   title "Average Number of Cars for every 25 years";
run;

*EDA for top 40 years;
ods graphics on / width=12in height=8in;
proc freq data=craig_lprice;
      tables year_new/
      plots(only)=freqplot(scale=freq);
      where year_new in (1980:2020);
run;


/* EDA for Odometer */
proc univariate data=craig_lprice plots;
    title " Exploring Odometer Values ";
    histogram odometer_new;
      var odometer_new;
run;


/* EDA for categorical variables */
proc freq data=craig_lprice order=freq;
      tables city manufacturer condition cylinders fuel title_status transmission
drive size type paint_color state_name/
      plots(only)=freqplot(scale=percent);
run;

*Pie chart for manufacturers;
proc gchart data=craig_lprice;
```

```
   pie manufacturer/ value=outside percent=outside;
   title "Exploring more of manufacturers";
run;

*Pie chart for condition;
proc gchart data=craig_lprice;
   pie condition/ value=outside percent=outside;
   title "Exploring more of manufacturers";
run;

*Graphs for state_name;
ods graphics on / width=12in height=8in;
proc freq data=craig_lprice order=freq;
      tables state_name/
      plots(only)=freqplot(scale=percent);
run;


/* Running the MLR model */
proc reg data=craig_lprice;
model price_new = odometer_new year_new fuel_gas fuel_hybrid fuel_diesel
transmission_automatic transmission_manual size_fullsize size_midsize size_compact
size_subcompact cylinders_other  cylinders_6 cylinders_4 cylinders_8
condition_excellent condition_likenew condition_fair condition_good
condition_unkown drive_rwd drive_fwd drive_4wd ;
        title "MLR model for Price Prediction with Price as response variable";
run;
quit;


proc reg data=craig_lprice;
model logPrice = odometer_new year_new fuel_gas fuel_hybrid fuel_diesel
transmission_automatic transmission_manual size_fullsize size_midsize size_compact
size_subcompact cylinders_other  cylinders_6 cylinders_4 cylinders_8
condition_excellent condition_likenew condition_fair condition_good
condition_unkown drive_rwd drive_fwd drive_4wd ;
        title "MLR model for Price Prediction with logPrice as response variable";
run;
quit;

proc glmselect data=craig_lprice plots=all;
   STEPWISE: model logPrice = odometer_new year_new fuel_gas fuel_hybrid
fuel_diesel transmission_automatic transmission_manual size_fullsize size_midsize
size_compact size_subcompact cylinders_other cylinders_6 cylinders_4 cylinders_8
condition_excellent condition_likenew condition_fair condition_good
condition_unkown drive_rwd drive_fwd drive_4wd / selection=stepwise
                   details=steps select=SL slentry=0.05 slstay=0.05;
   title "Stepwise Model Selection for Price with logPrice as response variable";
      run;
      quit;

/*Logistic Regression*/
ods graphics on / width=6in height=6in;
/* Condition compared to Price */
proc logistic data=craig_lprice
    plots(only)=(effect (clband showobs));
```

```
    MLogit2: model condition_likenew(event='1') = price_new;
    title 'Prob of like_new condition wrt price';
run;

proc logistic data=craig_lprice
    plots(only)=(effect (clband showobs));
    MLogit2: model condition_excellent(event='1') = price_new;
    title 'Prob of excellent condition wrt price';
run;

proc logistic data=craig_lprice
    plots(only)=(effect (clband showobs));
    MLogit2: model condition_good(event='1') = price_new;
    title 'Prob of good condition wrt price';
run;

proc logistic data=craig_lprice
    plots(only)=(effect (clband showobs));
    MLogit2: model condition_fair(event='1') = price_new;
    title ' Prob of fair condition wrt price';
run;

/* Condition compared to Odometer */
proc logistic data=craig_lprice
    plots(only)=(effect (clband showobs));
    MLogit2: model condition_likenew(event='1') = odometer_new;
    title 'Prob of like_new condition wrt odometer';
run;

proc logistic data=craig_lprice
    plots(only)=(effect (clband showobs));
    MLogit2: model condition_excellent(event='1') = odometer_new;
    title 'Prob of excellent condition wrt odometer';
run;

proc logistic data=craig_lprice
    plots(only)=(effect (clband showobs));
    MLogit2: model condition_good(event='1') = odometer_new;
    title 'Prob of good condition wrt odometer';
run;

proc logistic data=craig_lprice
    plots(only)=(effect (clband showobs));
    MLogit2: model condition_fair(event='1') = odometer_new;
    title 'Prob of fair condition wrt odometer';
run;
```