BUAN 6337.5U1

Predictive Analytics Using SAS

Summer 2019

Research Question

How do vehicle attributes affect its price?

Mentored By

Dr. Sourav Chatterjee

Prepared By

Pavan Sai Krishna Gorantla

Himaja Barla

Nirlep Shah

Ashley Creek

Siril Sai Alpuri

Contents

1. Abstract

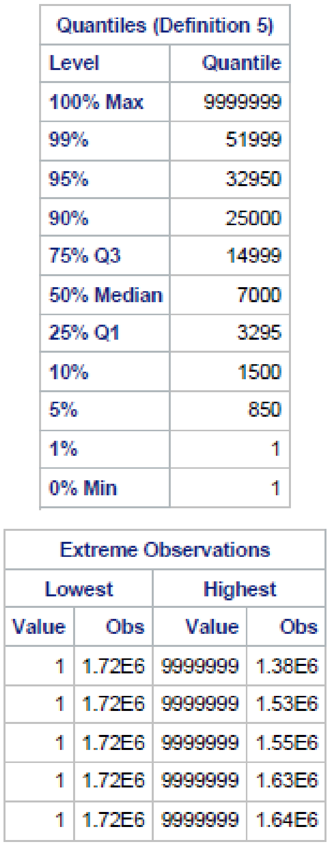
Craigslist is a popular platform used for buying and selling new and used goods by either private individuals or businesses. A dataset has been scraped from this website containing information about vehicles that are posted for sale on craigslist. The data has information on the sale price and other attributes of the vehicle. In this study, we focus on how the price depends on the attributes of vehicle and which attribute has significant effects on its sale price. Since most of the variables are categorical in nature, we run generalized linear regression models and interpret the effect shown on price. We aim to find best combination of attributes that can fetch highest price.

2. Data Description

We have cross-sectional dataset for 1,723,065 individual vehicles posted over several years span. Following below are variable descriptions and the green highlighted ones are vehicle attributes. We will be using only the highlighted variables and will be ignoring the rest of all other variables.

Variable	**Definition**
url	Website URL containing the vehicle details (unique identifier)
city	City in which the vehicle is available
price	Sale price of the vehicle
year	Year in which the vehicle was manufactured
manufacturer	Manufacturer name
make	Detailed description of make and model
condition	Describes vehicle condition (new, excellent, good, fair, like new, salvage)
cylinders	Number of cylinders in the vehicle (3, 4, 5, 6, 8, 10, 12, other)
fuel	Fuel type of the vehicle (diesel, electric, gas, hybrid, other)
odometer	Current odometer reading of the vehicle
title_status	Title about past history (clean, lien, missing, parts only, rebuilt, salvage)
transmission	Transmission Type of the vehicle (automatic, manual, other)
vin	Vehicle Identification Number
drive	Drive type (front wheel drive, rear wheel drive, four wheel drive)
size	Vehicle size (compact, full-size, mid-size, sub-compact)
type	Vehicle type (suv, bus, convertible, coupe, hatchback, mini-van, offroad, pickup, sedan, truck, van, wagon, other)
paint_color	Vehicle color (black, blue, brown, green, grey, orange, purple, red, silver, white, yellow, custom)
image_url	URL containing the vehicle's image
lat	Latitude of the location where the vehicle is available
long	Longitude of location where the vehicle is available
county_fips	Federal Information Processing Standard codes for the county
county_name	Name of the county
state_fips	Federal Information Processing Standard codes for the state
state_code	Two letter state code
state_name	Name of the state
weather	Weather code of the state

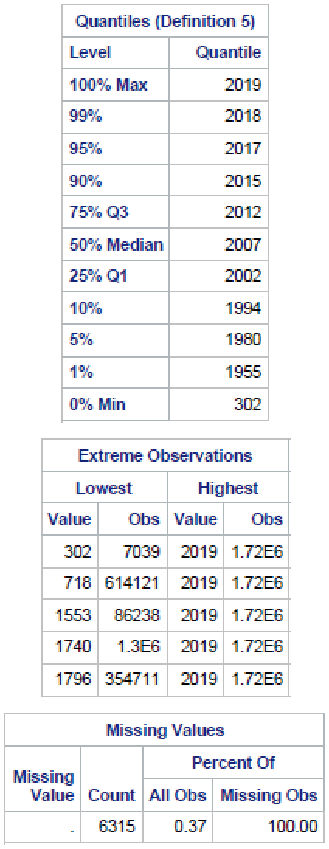
2.1. Outliers and Missing Values



Quantiles (Definition 5)	
Level	Quantile
100% Max	9999999
99%	51999
95%	32950
90%	25000
75% Q3	14999
50% Median	7000
25% Q1	3295
10%	1500
5%	850
1%	1
0% Min	1

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	1.72E6	9999999	1.38E6
1	1.72E6	9999999	1.53E6
1	1.72E6	9999999	1.55E6
1	1.72E6	9999999	1.63E6
1	1.72E6	9999999	1.64E6

price

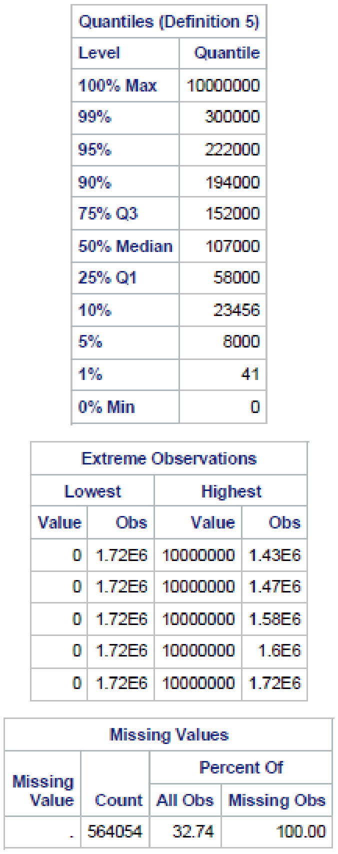


Quantiles (Definition 5)	
Level	Quantile
100% Max	2019
99%	2018
95%	2017
90%	2015
75% Q3	2012
50% Median	2007
25% Q1	2002
10%	1994
5%	1980
1%	1955
0% Min	302

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
302	7039	2019	1.72E6
718	614121	2019	1.72E6
1553	86238	2019	1.72E6
1740	1.3E6	2019	1.72E6
1796	354711	2019	1.72E6

Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	6315	0.37	100.00

year



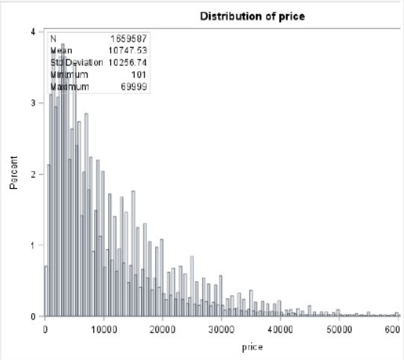
Quantiles (Definition 5)	
Level	Quantile
100% Max	10000000
99%	300000
95%	222000
90%	194000
75% Q3	152000
50% Median	107000
25% Q1	58000
10%	23456
5%	8000
1%	41
0% Min	0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1.72E6	10000000	1.43E6
0	1.72E6	10000000	1.47E6
0	1.72E6	10000000	1.58E6
0	1.72E6	10000000	1.6E6
0	1.72E6	10000000	1.72E6

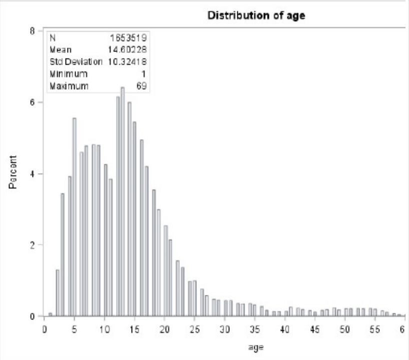
Missing Values			
		Percent Of	
Missing Value	Count	All Obs	Missing Obs
.	564054	32.74	100.00

odometer

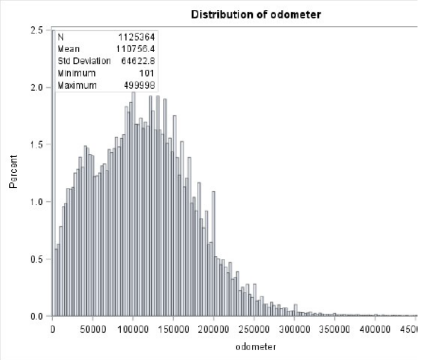
To remove illogical values, we filtered rows with *price* between 100 and 70,000 and *odometer* between 500 and 500,000 and *year* more than 1950. The *odometer* has 32.74% missing values. And we converted year into ***age*** by subtracting it from 2020 and new distribution is as follows:



price



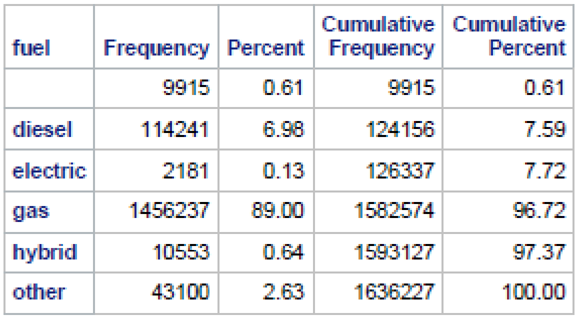
age



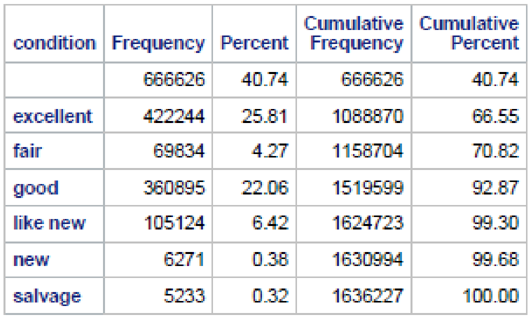
odometer

Since, all graphs are right-skewed, its better to normalize the data by taking the natural logarithm.

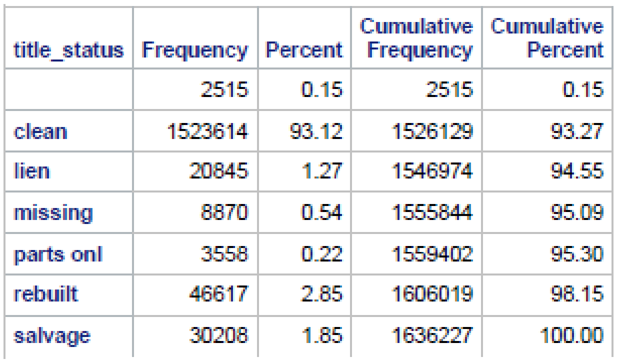
Among non-numerical variables, *make* is very long descriptive text and we will be ignoring that. Below are other variables frequency distributions using which we can find missing values share.



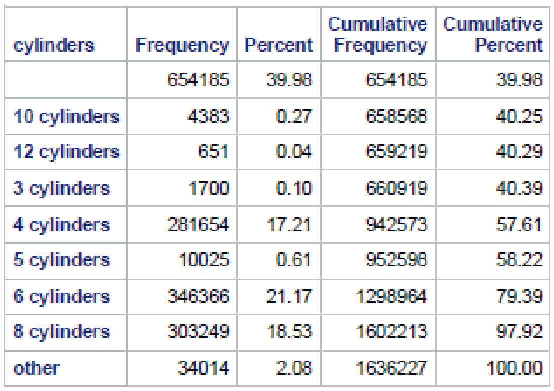
fuel	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	9915	0.61	9915	0.61
diesel	114241	6.98	124156	7.59
electric	2181	0.13	126337	7.72
gas	1456237	89.00	1582574	96.72
hybrid	10553	0.64	1593127	97.37
other	43100	2.63	1636227	100.00



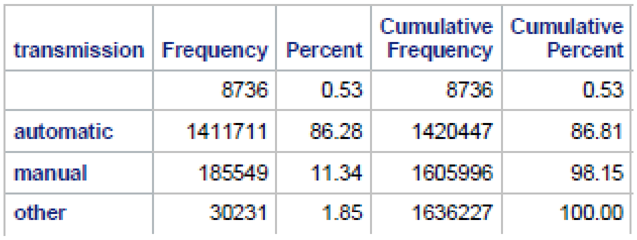
condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	666626	40.74	666626	40.74
excellent	422244	25.81	1088870	66.55
fair	69834	4.27	1158704	70.82
good	360895	22.06	1519599	92.87
like new	105124	6.42	1624723	99.30
new	6271	0.38	1630994	99.68
salvage	5233	0.32	1636227	100.00



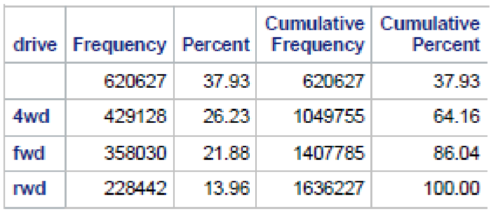
title_status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	2515	0.15	2515	0.15
clean	1523614	93.12	1526129	93.27
lien	20845	1.27	1546974	94.55
missing	8870	0.54	1555844	95.09
parts onl	3558	0.22	1559402	95.30
rebuilt	46617	2.85	1606019	98.15
salvage	30208	1.85	1636227	100.00



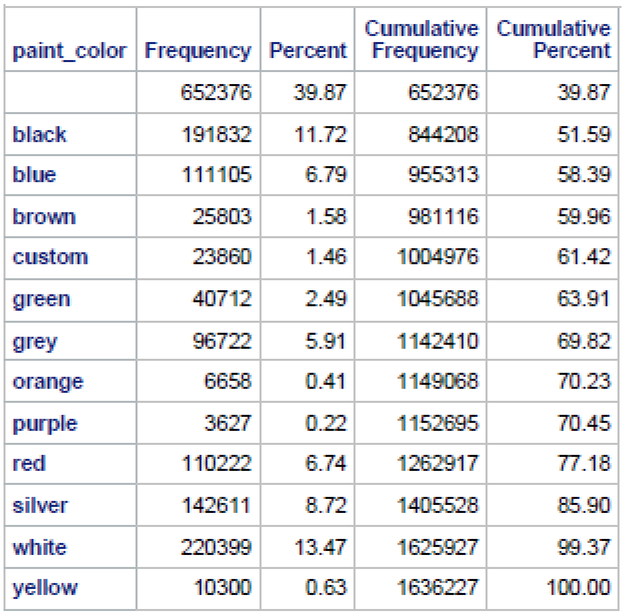
cylinders	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	654185	39.98	654185	39.98
10 cylinders	4383	0.27	658568	40.25
12 cylinders	651	0.04	659219	40.29
3 cylinders	1700	0.10	660919	40.39
4 cylinders	281654	17.21	942573	57.61
5 cylinders	10025	0.61	952598	58.22
6 cylinders	346366	21.17	1298964	79.39
8 cylinders	303249	18.53	1602213	97.92
other	34014	2.08	1636227	100.00



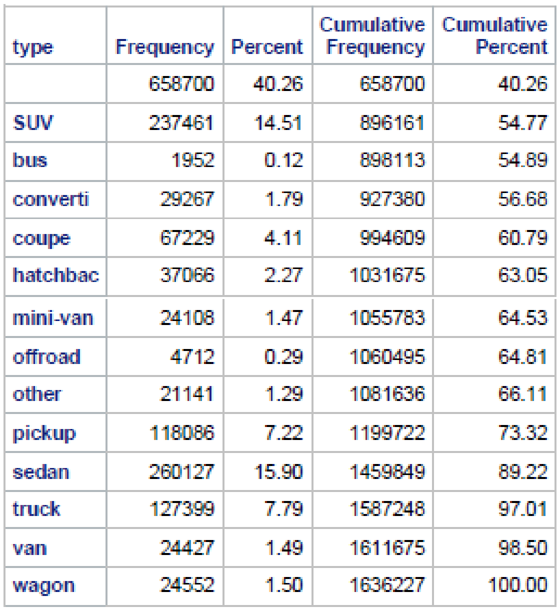
transmission	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	8736	0.53	8736	0.53
automatic	1411711	86.28	1420447	86.81
manual	185549	11.34	1605996	98.15
other	30231	1.85	1636227	100.00



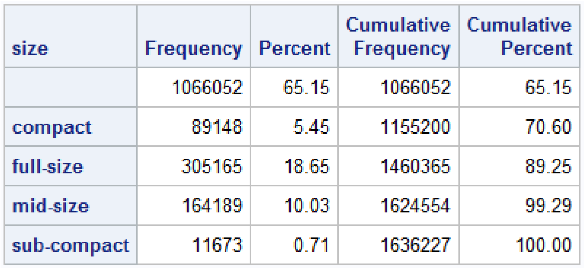
drive	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	620627	37.93	620627	37.93
4wd	429128	26.23	1049755	64.16
fwd	358030	21.88	1407785	86.04
rwd	228442	13.96	1636227	100.00



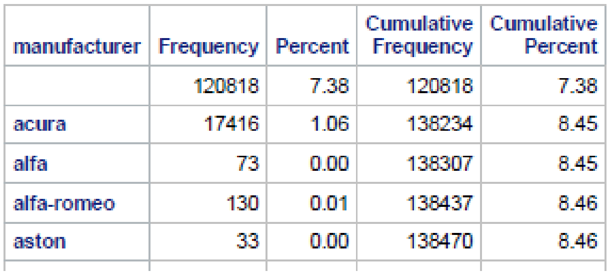
paint_color	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	652376	39.87	652376	39.87
black	191832	11.72	844208	51.59
blue	111105	6.79	955313	58.39
brown	25803	1.58	981116	59.96
custom	23860	1.46	1004976	61.42
green	40712	2.49	1045688	63.91
grey	96722	5.91	1142410	69.82
orange	6658	0.41	1149068	70.23
purple	3627	0.22	1152695	70.45
red	110222	6.74	1262917	77.18
silver	142611	8.72	1405528	85.90
white	220399	13.47	1625927	99.37
yellow	10300	0.63	1636227	100.00



type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	658700	40.26	658700	40.26
SUV	237461	14.51	896161	54.77
bus	1952	0.12	898113	54.89
converti	29267	1.79	927380	56.68
coupe	67229	4.11	994609	60.79
hatchbac	37066	2.27	1031675	63.05
mini-van	24108	1.47	1055783	64.53
offroad	4712	0.29	1060495	64.81
other	21141	1.29	1081636	66.11
pickup	118086	7.22	1199722	73.32
sedan	260127	15.90	1459849	89.22
truck	127399	7.79	1587248	97.01
van	24427	1.49	1611675	98.50
wagon	24552	1.50	1636227	100.00



size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	1066052	65.15	1066052	65.15
compact	89148	5.45	1155200	70.60
full-size	305165	18.65	1460365	89.25
mid-size	164189	10.03	1624554	99.29
sub-compact	11673	0.71	1636227	100.00



manufacturer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	120818	7.38	120818	7.38
acura	17416	1.06	138234	8.45
alfa	73	0.00	138307	8.45
alfa-romeo	130	0.01	138437	8.46
aston	33	0.00	138470	8.46

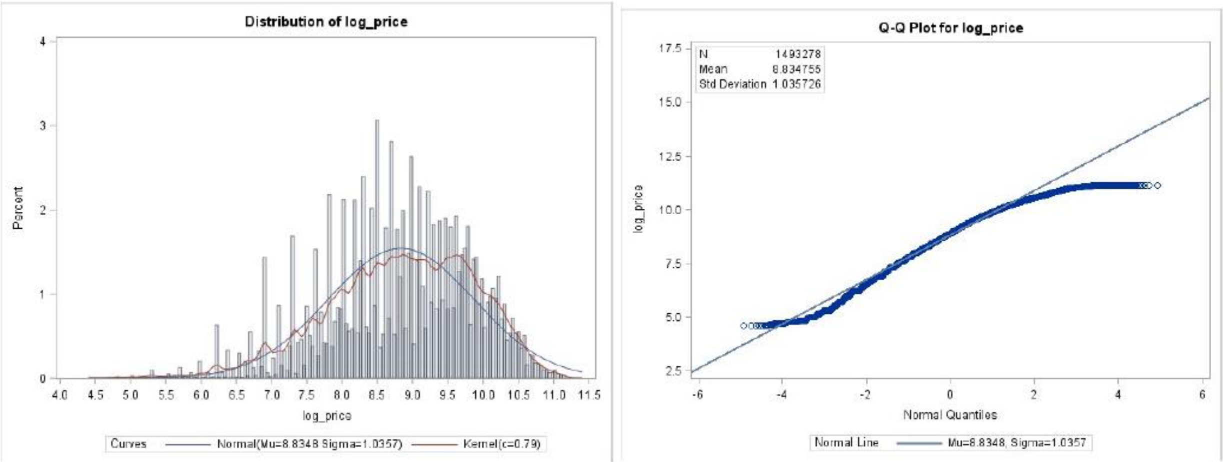
The *manufacturer* has 7% missing values but *size* has 65% missing values. As two-thirds of this variable is empty, we are ignoring the *size* variable. Also, *odometer*, *condition*, *cylinders*, *drive*, *type*, *paint_color* have 30-40% missing data. Hence, they are separated from other variables into a new dataset with smaller number of rows, upon deleting the missing rows mentioned as below:

Dataset	Variables	Rows	Rows %
With 4 category variables (*4cat*)	*price, age, log_price, log_age,* *fuel, title_status, transmission, manufacturer*	1493278	86.66 %
With 9 category variables (*9cat*)	*price, age, odometer, log_price, log_age, log_odometer,* *fuel, title_status, transmission, manufacturer,* *condition, cylinders, drive, type, paint_color*	441328	25.61 %

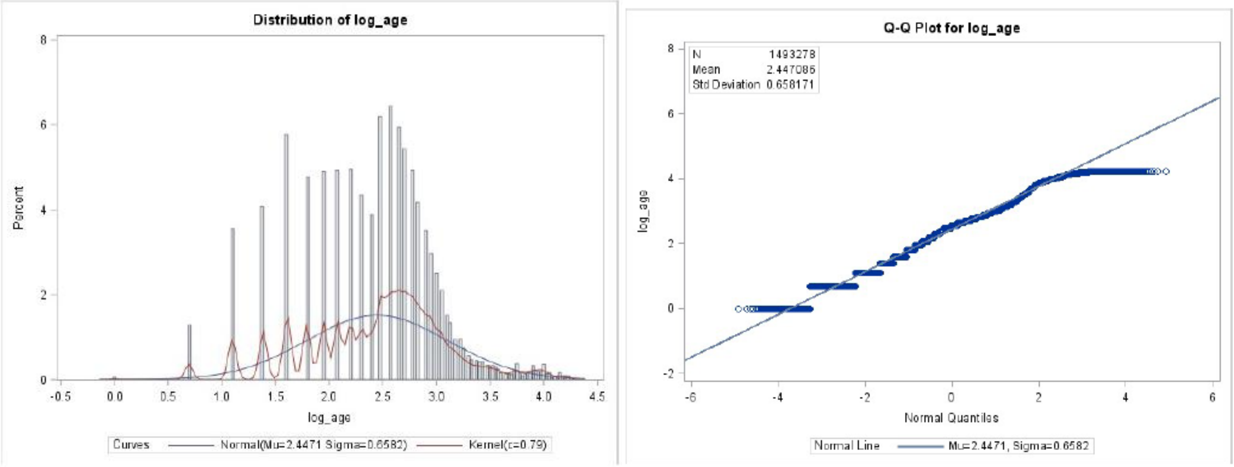
We will be using first table while using *fuel*, *title_status*, *transmission*, *manufacturer* as predictors and the second table while using *condition*, *cylinders*, *drive*, *type*, *paint_color* as the predictors.

2.2. Exploratory Data Analysis

After removing the missing value rows, *log_price* and *log_age* (in *4cat*) are normally distributed.

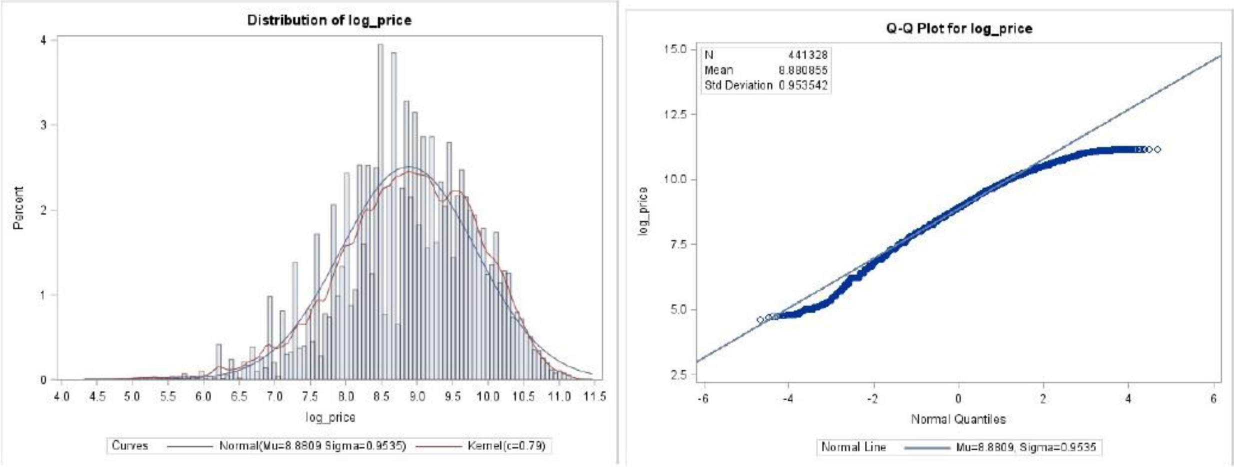


4cat* Dataset - Distribution of *log_price

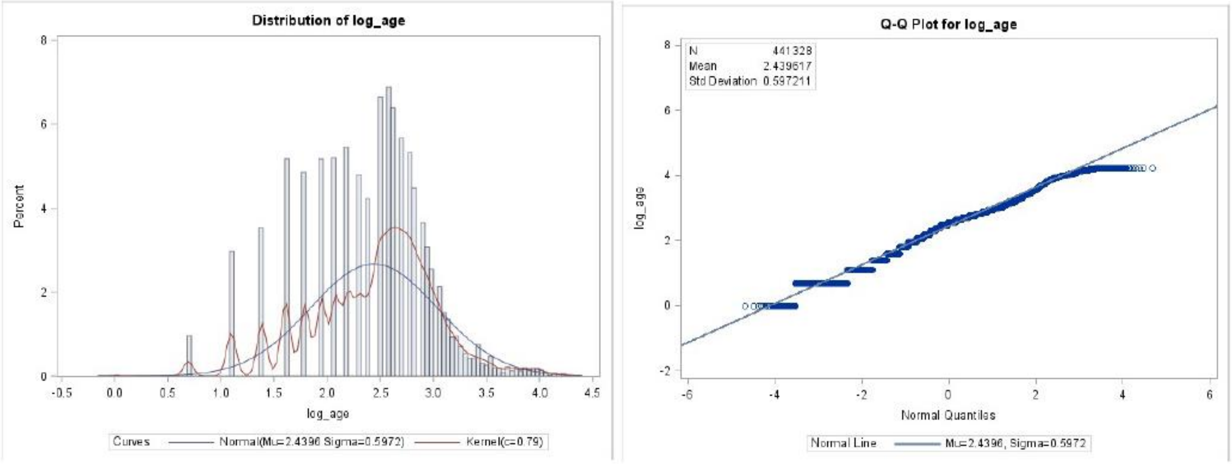


4cat* Dataset - Distribution of *log_age

After removing the missing value rows, *log_price* and *log_age* (in *9cat*) are normally distributed.

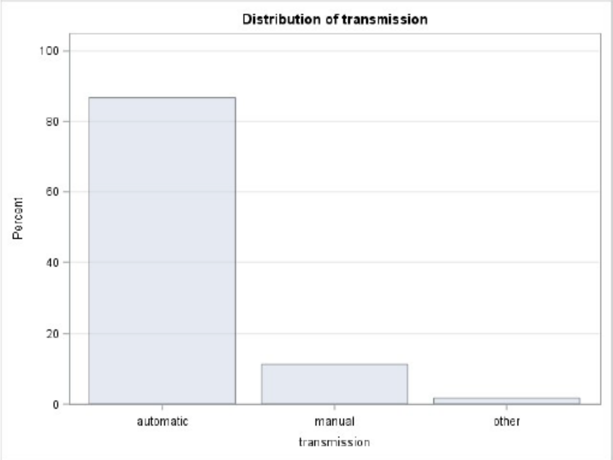


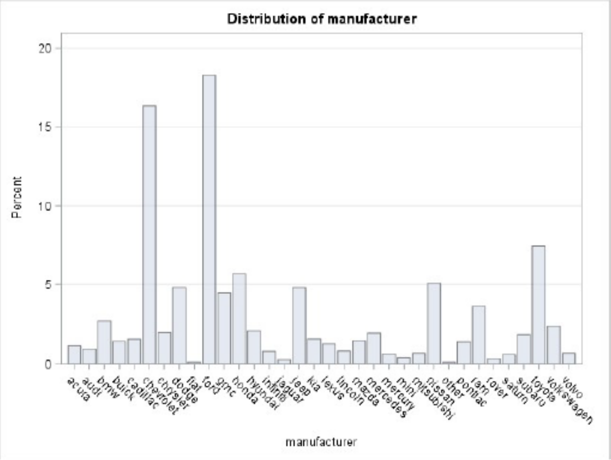
9cat* Dataset - Distribution of *log_price



9cat* Dataset - Distribution of *log_age

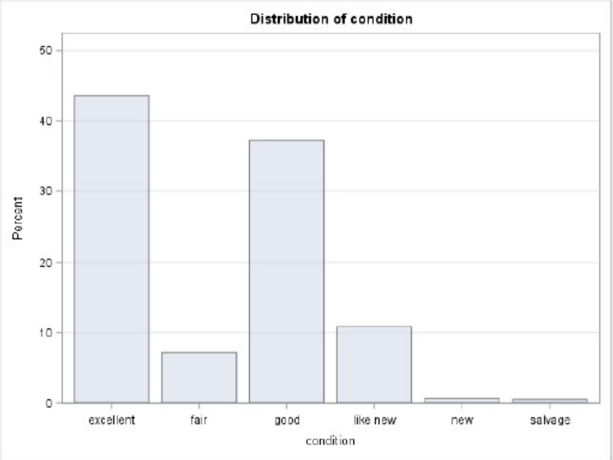
Below are the vehicles frequency distributions plots of all major categorical variables in the data

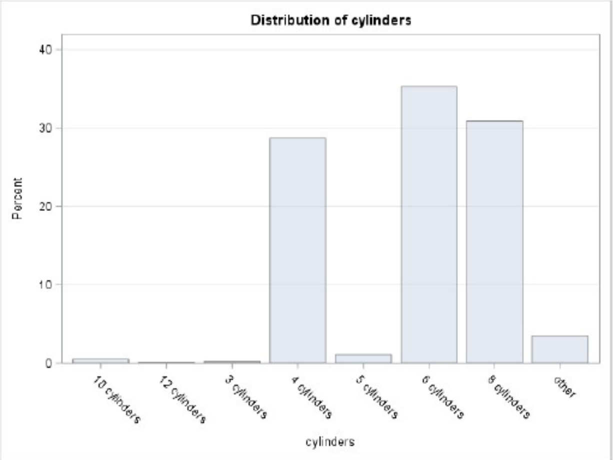




Most vehicles have automatic transmission

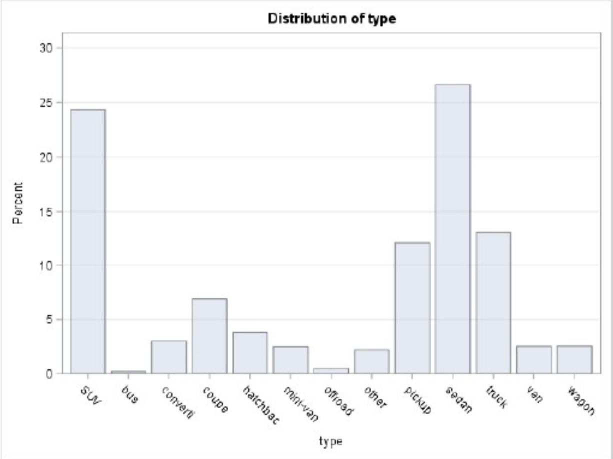
Top vehicles brands are ford and chevrolet

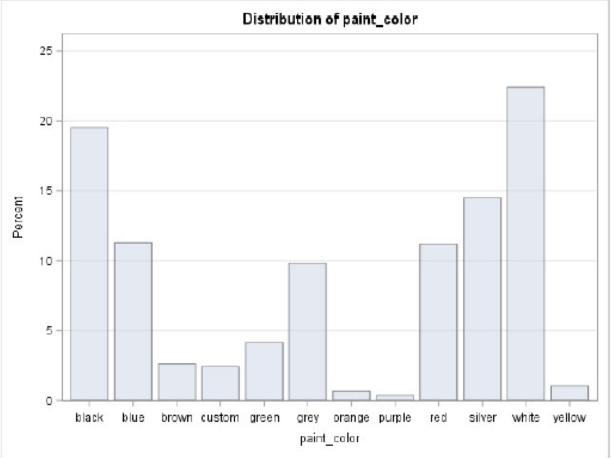




Most vehicles are in excellent or good condition

Most vehicles have either 4, 6 or 8 cylinders

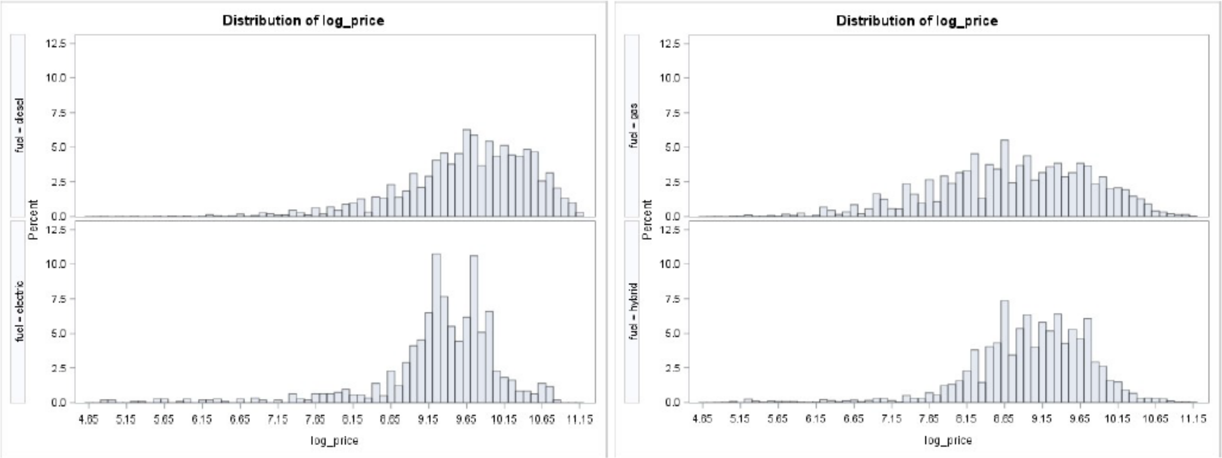




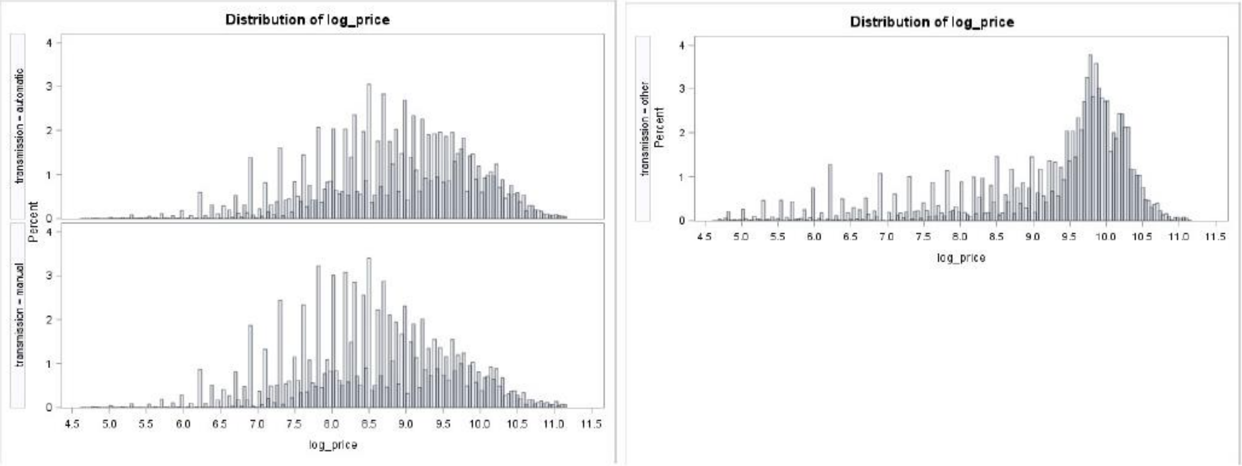
Top vehicles types are sedan, suv, truck, prickup

Most used vehicle colors are black, white, silver

Below are distribution plots of *log_price* split by some category variables like *fuel*, *transmission*

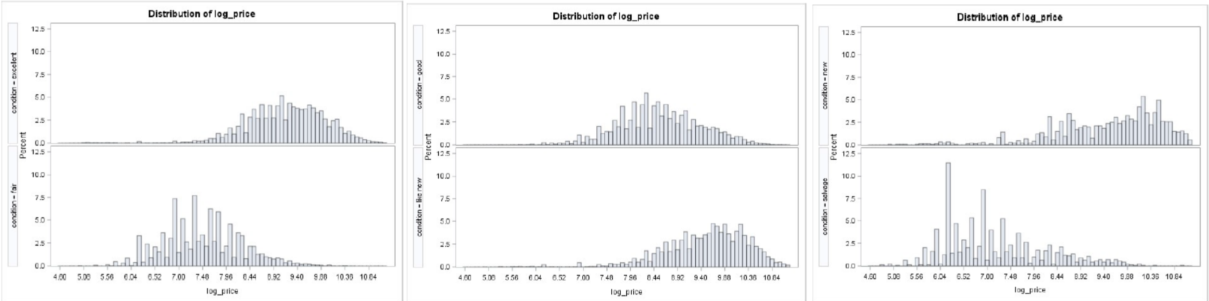


***4cat* Dataset - Variance of *log_price* is more when *fuel* is gas**



***4cat* Dataset - *log_price* is more left skewed when transmission type is neither automatic nor manual**

Below are the histogram distribution plots of *log_price* split by the category variable *condition*

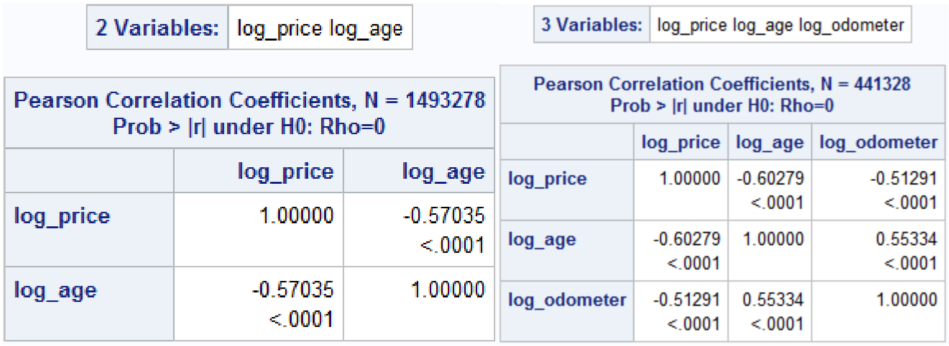


left-top-excellent, left-down-fair, middle-top-good, middle-down-like new, right-top-new, right-down-salvage

As we expect, mean of *price* is higher when *condition* is excellent, like new or new. For salvage, the *price* is scattered all over with high variance as people are skeptical about salvaged vehicles.

2.3. Correlations among predictors

As we have continuous and categorical variables, we need to use Pearson linear correlation for continuous, chi-square test and Cramer for categorical, anova test for continuous and categorical. Pearson correlation coefficients between *price* and *age* in *4cat*, *price*, *age* and *odometer* in *9cat*



2 Variables:	log_price log_age

Pearson Correlation Coefficients, N = 1493278
Prob > |r| under H0: Rho=0

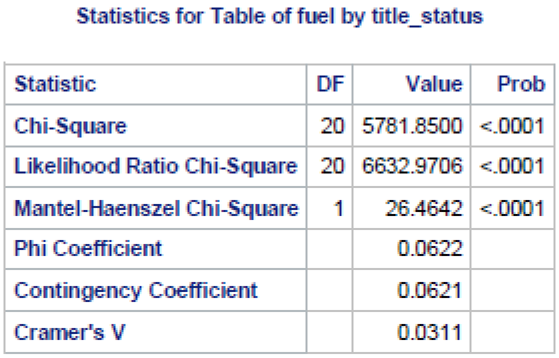
	log_price	log_age
log_price	1.00000	-0.57035 <.0001
log_age	-0.57035 <.0001	1.00000

3 Variables:	log_price log_age log_odometer

Pearson Correlation Coefficients, N = 441328
Prob > |r| under H0: Rho=0

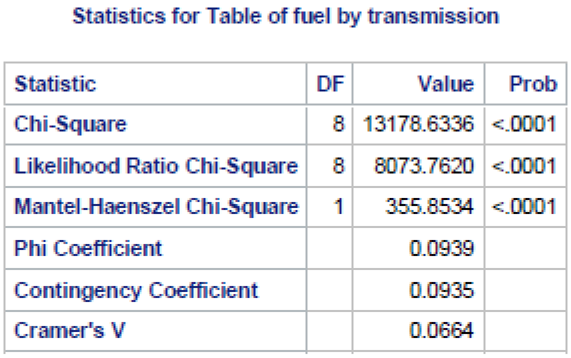
	log_price	log_age	log_odometer
log_price	1.00000	-0.60279 <.0001	-0.51291 <.0001
log_age	-0.60279 <.0001	1.00000	0.55334 <.0001
log_odometer	-0.51291 <.0001	0.55334 <.0001	1.00000

Mantel-Haenszel chi-square test and cramer's v score between categorical variables in *4cat* table



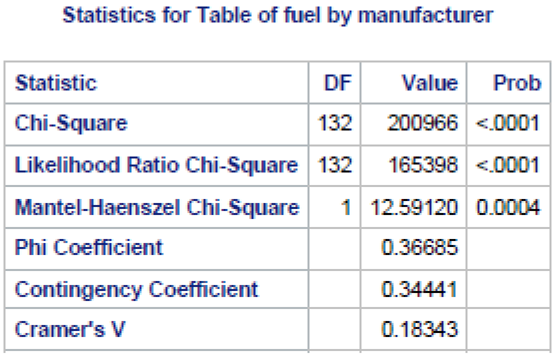
Statistics for Table of fuel by title_status

Statistic	DF	Value	Prob
Chi-Square	20	5781.8500	<.0001
Likelihood Ratio Chi-Square	20	6632.9706	<.0001
Mantel-Haenszel Chi-Square	1	26.4642	<.0001
Phi Coefficient		0.0622	
Contingency Coefficient		0.0621	
Cramer's V		0.0311	



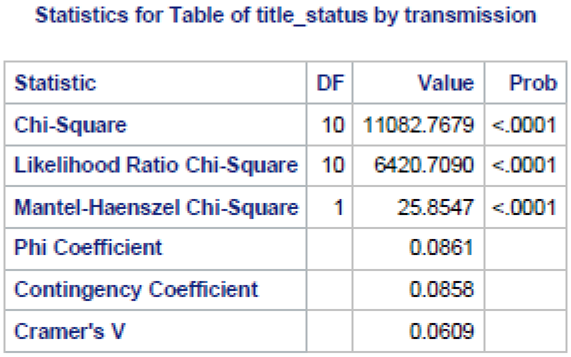
Statistics for Table of fuel by transmission

Statistic	DF	Value	Prob
Chi-Square	8	13178.6336	<.0001
Likelihood Ratio Chi-Square	8	8073.7620	<.0001
Mantel-Haenszel Chi-Square	1	355.8534	<.0001
Phi Coefficient		0.0939	
Contingency Coefficient		0.0935	
Cramer's V		0.0664	



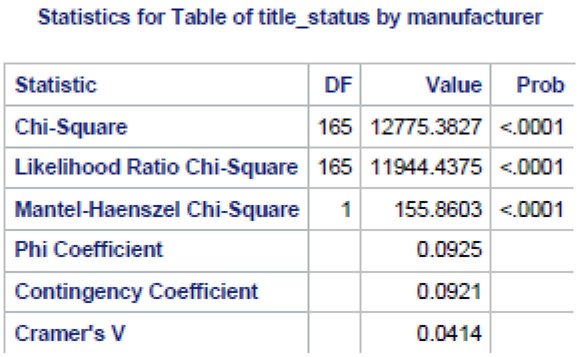
Statistics for Table of fuel by manufacturer

Statistic	DF	Value	Prob
Chi-Square	132	200966	<.0001
Likelihood Ratio Chi-Square	132	165398	<.0001
Mantel-Haenszel Chi-Square	1	12.59120	0.0004
Phi Coefficient		0.36685	
Contingency Coefficient		0.34441	
Cramer's V		0.18343	



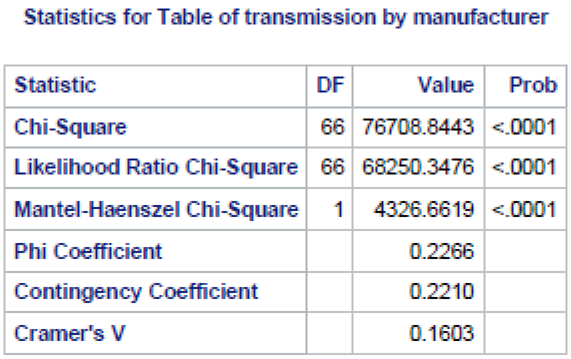
Statistics for Table of title_status by transmission

Statistic	DF	Value	Prob
Chi-Square	10	11082.7679	<.0001
Likelihood Ratio Chi-Square	10	6420.7090	<.0001
Mantel-Haenszel Chi-Square	1	25.8547	<.0001
Phi Coefficient		0.0861	
Contingency Coefficient		0.0858	
Cramer's V		0.0609	



Statistics for Table of title_status by manufacturer

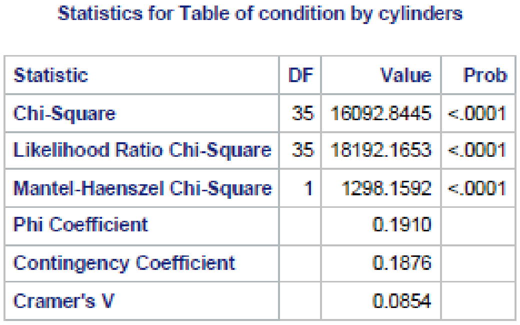
Statistic	DF	Value	Prob
Chi-Square	165	12775.3827	<.0001
Likelihood Ratio Chi-Square	165	11944.4375	<.0001
Mantel-Haenszel Chi-Square	1	155.8603	<.0001
Phi Coefficient		0.0925	
Contingency Coefficient		0.0921	
Cramer's V		0.0414	



Statistics for Table of transmission by manufacturer

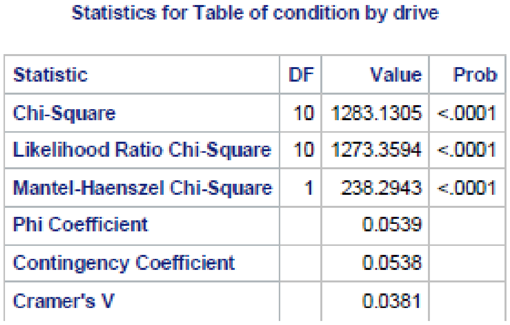
Statistic	DF	Value	Prob
Chi-Square	66	76708.8443	<.0001
Likelihood Ratio Chi-Square	66	68250.3476	<.0001
Mantel-Haenszel Chi-Square	1	4326.6619	<.0001
Phi Coefficient		0.2266	
Contingency Coefficient		0.2210	
Cramer's V		0.1603	

Mantel-Haenszel chi-square test and Cramer's V score between categorical variables in *9cat* table



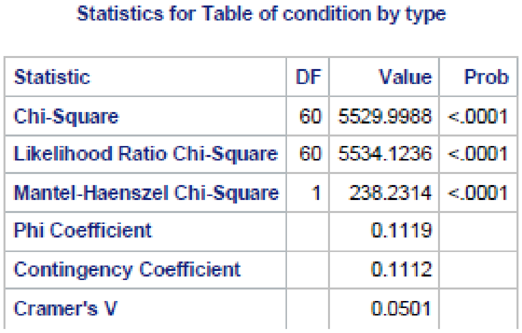
Statistics for Table of condition by cylinders

Statistic	DF	Value	Prob
Chi-Square	35	16092.8445	<.0001
Likelihood Ratio Chi-Square	35	18192.1653	<.0001
Mantel-Haenszel Chi-Square	1	1298.1592	<.0001
Phi Coefficient		0.1910	
Contingency Coefficient		0.1876	
Cramer's V		0.0854	



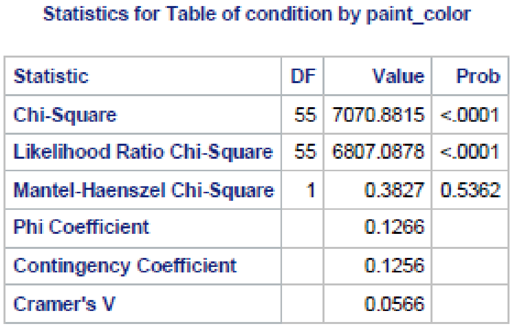
Statistics for Table of condition by drive

Statistic	DF	Value	Prob
Chi-Square	10	1283.1305	<.0001
Likelihood Ratio Chi-Square	10	1273.3594	<.0001
Mantel-Haenszel Chi-Square	1	238.2943	<.0001
Phi Coefficient		0.0539	
Contingency Coefficient		0.0538	
Cramer's V		0.0381	



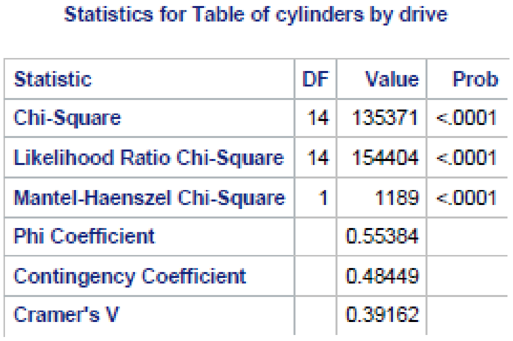
Statistics for Table of condition by type

Statistic	DF	Value	Prob
Chi-Square	60	5529.9988	<.0001
Likelihood Ratio Chi-Square	60	5534.1236	<.0001
Mantel-Haenszel Chi-Square	1	238.2314	<.0001
Phi Coefficient		0.1119	
Contingency Coefficient		0.1112	
Cramer's V		0.0501	



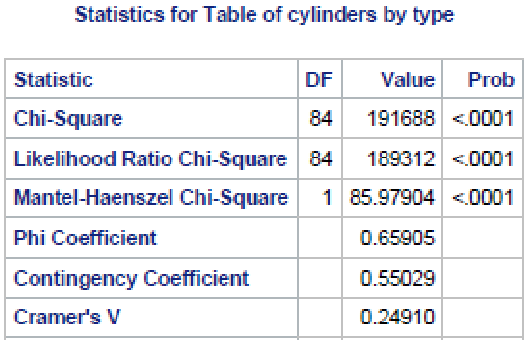
Statistics for Table of condition by paint_color

Statistic	DF	Value	Prob
Chi-Square	55	7070.8815	<.0001
Likelihood Ratio Chi-Square	55	6807.0878	<.0001
Mantel-Haenszel Chi-Square	1	0.3827	0.5362
Phi Coefficient		0.1266	
Contingency Coefficient		0.1256	
Cramer's V		0.0566	



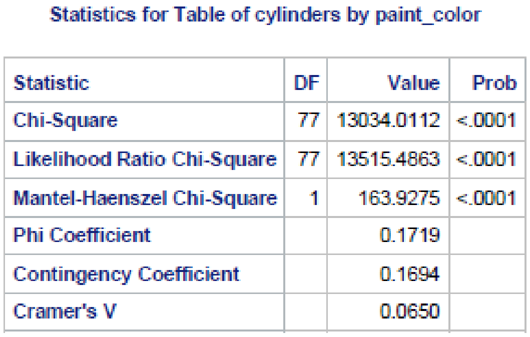
Statistics for Table of cylinders by drive

Statistic	DF	Value	Prob
Chi-Square	14	135371	<.0001
Likelihood Ratio Chi-Square	14	154404	<.0001
Mantel-Haenszel Chi-Square	1	1189	<.0001
Phi Coefficient		0.55384	
Contingency Coefficient		0.48449	
Cramer's V		0.39162	



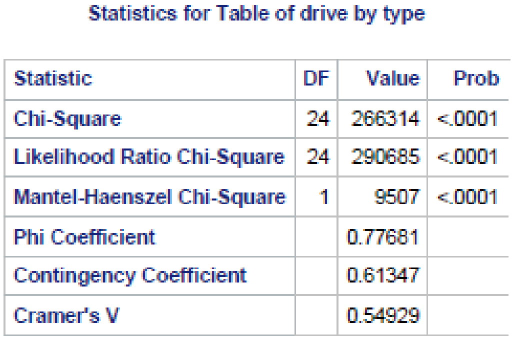
Statistics for Table of cylinders by type

Statistic	DF	Value	Prob
Chi-Square	84	191688	<.0001
Likelihood Ratio Chi-Square	84	189312	<.0001
Mantel-Haenszel Chi-Square	1	85.97904	<.0001
Phi Coefficient		0.65905	
Contingency Coefficient		0.55029	
Cramer's V		0.24910	



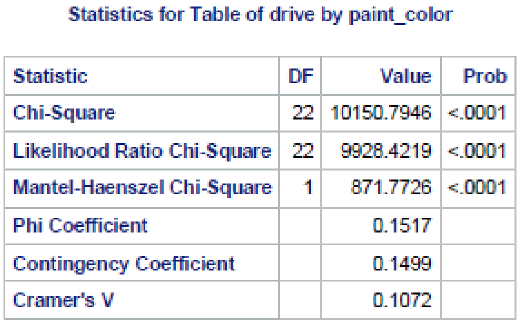
Statistics for Table of cylinders by paint_color

Statistic	DF	Value	Prob
Chi-Square	77	13034.0112	<.0001
Likelihood Ratio Chi-Square	77	13515.4863	<.0001
Mantel-Haenszel Chi-Square	1	163.9275	<.0001
Phi Coefficient		0.1719	
Contingency Coefficient		0.1694	
Cramer's V		0.0650	



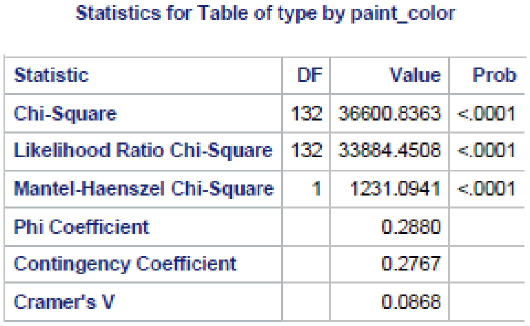
Statistics for Table of drive by type

Statistic	DF	Value	Prob
Chi-Square	24	266314	<.0001
Likelihood Ratio Chi-Square	24	290685	<.0001
Mantel-Haenszel Chi-Square	1	9507	<.0001
Phi Coefficient		0.77681	
Contingency Coefficient		0.61347	
Cramer's V		0.54929	



Statistics for Table of drive by paint_color

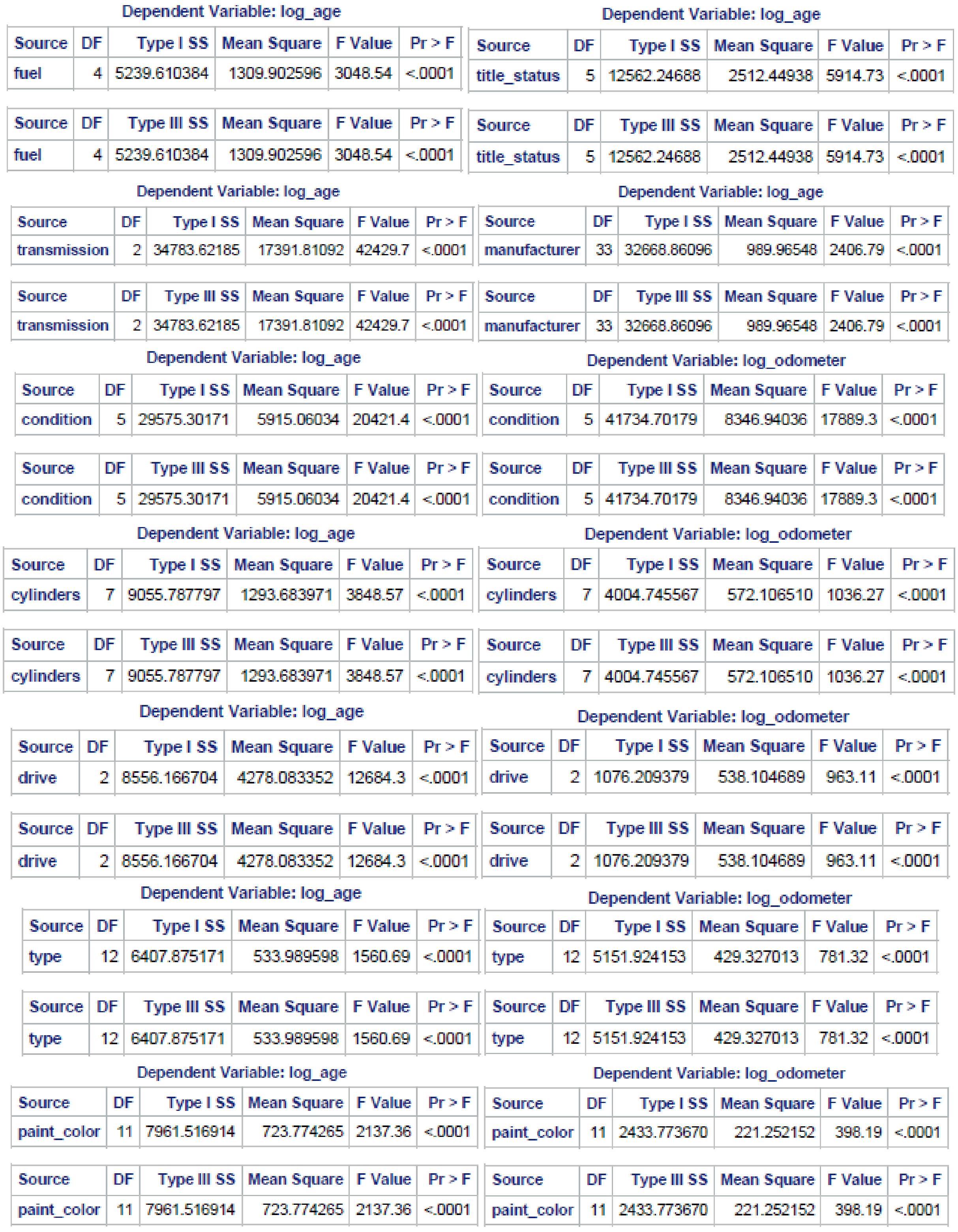
Statistic	DF	Value	Prob
Chi-Square	22	10150.7946	<.0001
Likelihood Ratio Chi-Square	22	9928.4219	<.0001
Mantel-Haenszel Chi-Square	1	871.7726	<.0001
Phi Coefficient		0.1517	
Contingency Coefficient		0.1499	
Cramer's V		0.1072	



Statistics for Table of type by paint_color

Statistic	DF	Value	Prob
Chi-Square	132	36600.8363	<.0001
Likelihood Ratio Chi-Square	132	33884.4508	<.0001
Mantel-Haenszel Chi-Square	1	1231.0941	<.0001
Phi Coefficient		0.2880	
Contingency Coefficient		0.2767	
Cramer's V		0.0868	

Anova between *log_age*, *log_odometer* and other categorical variables in *4cat* and *9cat* datasets



Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
fuel	4	5239.610384	1309.902596	3048.54	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
fuel	4	5239.610384	1309.902596	3048.54	<.0001

Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
title_status	5	12562.24688	2512.44938	5914.73	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
title_status	5	12562.24688	2512.44938	5914.73	<.0001

Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
transmission	2	34783.62185	17391.81092	42429.7	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
transmission	2	34783.62185	17391.81092	42429.7	<.0001

Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
manufacturer	33	32668.86096	989.96548	2406.79	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
manufacturer	33	32668.86096	989.96548	2406.79	<.0001

Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
condition	5	29575.30171	5915.06034	20421.4	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
condition	5	29575.30171	5915.06034	20421.4	<.0001

Dependent Variable: log_odometer

Source	DF	Type I SS	Mean Square	F Value	Pr > F
condition	5	41734.70179	8346.94036	17889.3	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
condition	5	41734.70179	8346.94036	17889.3	<.0001

Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
cylinders	7	9055.787797	1293.683971	3848.57	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
cylinders	7	9055.787797	1293.683971	3848.57	<.0001

Dependent Variable: log_odometer

Source	DF	Type I SS	Mean Square	F Value	Pr > F
cylinders	7	4004.745567	572.106510	1036.27	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
cylinders	7	4004.745567	572.106510	1036.27	<.0001

Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
drive	2	8556.166704	4278.083352	12684.3	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
drive	2	8556.166704	4278.083352	12684.3	<.0001

Dependent Variable: log_odometer

Source	DF	Type I SS	Mean Square	F Value	Pr > F
drive	2	1076.209379	538.104689	963.11	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
drive	2	1076.209379	538.104689	963.11	<.0001

Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
type	12	6407.875171	533.989598	1560.69	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
type	12	6407.875171	533.989598	1560.69	<.0001

Dependent Variable: log_odometer

Source	DF	Type I SS	Mean Square	F Value	Pr > F
type	12	5151.924153	429.327013	781.32	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
type	12	5151.924153	429.327013	781.32	<.0001

Dependent Variable: log_age

Source	DF	Type I SS	Mean Square	F Value	Pr > F
paint_color	11	7961.516914	723.774265	2137.36	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
paint_color	11	7961.516914	723.774265	2137.36	<.0001

Dependent Variable: log_odometer

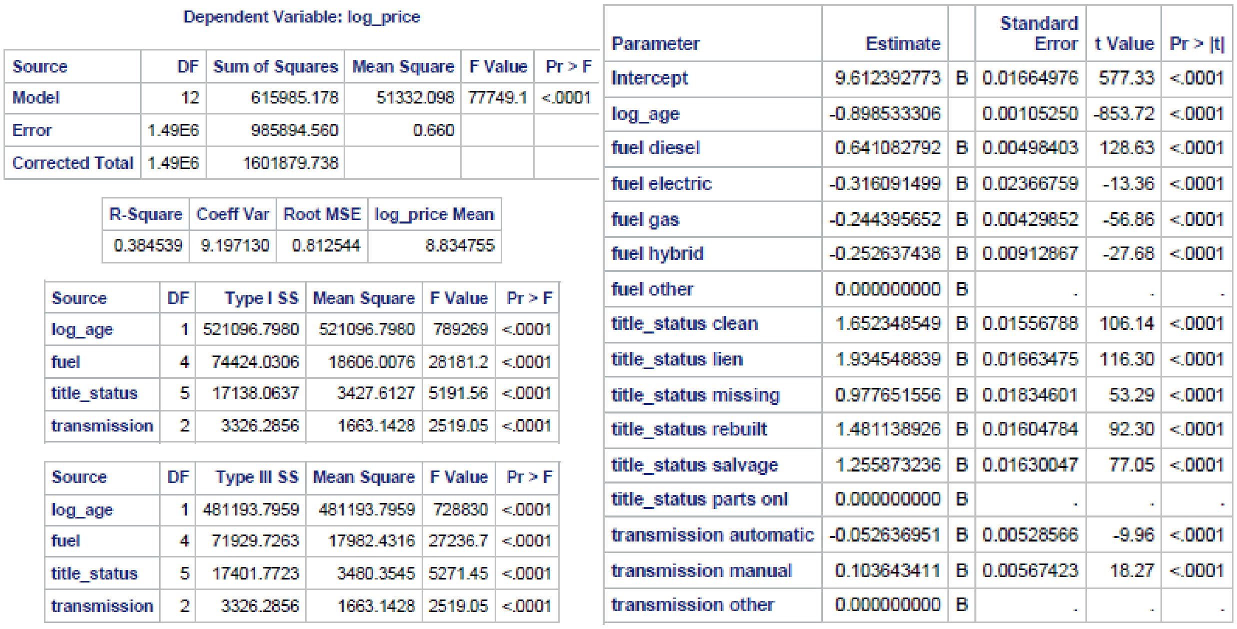
Source	DF	Type I SS	Mean Square	F Value	Pr > F
paint_color	11	2433.773670	221.252152	398.19	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
paint_color	11	2433.773670	221.252152	398.19	<.0001

3. Regression Models

We are building the following 4 regression models listed below using the GLM procedure in SAS.
Model 1 - *log_price* as the dependent and *fuel*, *transmission*, *title_status* as predictors on *4cat*
Model 2 - *log_price* as the dependent and *manufacturer* as predictors on *4cat*
Model 3 - *log_price* as the dependent and *condition*, *cylinders*, *drive* as predictors on *9cat*
Model 4 - *log_price* as the dependent and *type*, *paint_color* as predictors on *9cat*

3.1. Pricing Model with Fuel, Transmission, Title



Dependent Variable: log_price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	615985.178	51332.098	77749.1	<.0001
Error	1.49E6	985894.560	0.660		
Corrected Total	1.49E6	1601879.738			

R-Square	Coeff Var	Root MSE	log_price Mean
0.384539	9.197130	0.812544	8.834755

Source	DF	Type I SS	Mean Square	F Value	Pr > F
log_age	1	521096.7980	521096.7980	789269	<.0001
fuel	4	74424.0306	18606.0076	28181.2	<.0001
title_status	5	17138.0637	3427.6127	5191.56	<.0001
transmission	2	3326.2856	1663.1428	2519.05	<.0001

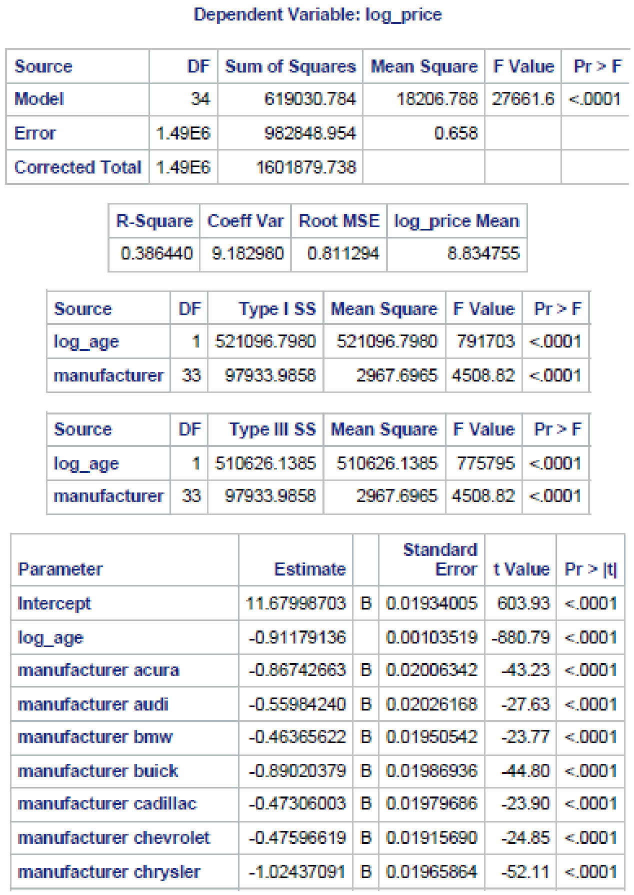
Source	DF	Type III SS	Mean Square	F Value	Pr > F
log_age	1	481193.7959	481193.7959	728830	<.0001
fuel	4	71929.7263	17982.4316	27236.7	<.0001
title_status	5	17401.7723	3480.3545	5271.45	<.0001
transmission	2	3326.2856	1663.1428	2519.05	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	9.612392773	B	0.01664976	577.33	<.0001
log_age	-0.898533306		0.00105250	-853.72	<.0001
fuel diesel	0.641082792	B	0.00498403	128.63	<.0001
fuel electric	-0.316091499	B	0.02366759	-13.36	<.0001
fuel gas	-0.244395652	B	0.00429852	-56.86	<.0001
fuel hybrid	-0.252637438	B	0.00912867	-27.68	<.0001
fuel other	0.000000000	B	.	.	.
title_status clean	1.652348549	B	0.01556788	106.14	<.0001
title_status lien	1.934548839	B	0.01663475	116.30	<.0001
title_status missing	0.977651556	B	0.01834601	53.29	<.0001
title_status rebuilt	1.481138926	B	0.01604784	92.30	<.0001
title_status salvage	1.255873236	B	0.01630047	77.05	<.0001
title_status parts onl	0.000000000	B	.	.	.
transmission automatic	-0.052636951	B	0.00528566	-9.96	<.0001
transmission manual	0.103643411	B	0.00567423	18.27	<.0001
transmission other	0.000000000	B	.	.	.

All estimates are found to be significant at 1% and R^2 is 38%. From the above, we can interpret:

When the age of the vehicle increases by 10% then the sale price of the vehicle decreases by 9%.
If the vehicle fuel type is diesel, then price is more by 64% when compared to “other” fuel type.
If the vehicle fuel type is electric, then price is less by 31% when compared to “other” fuel type.
If the vehicle fuel type is gas, then price is less by 24% when compared to the “other” fuel type.
If the vehicle fuel type is hybrid, then price is less by 25% when compared with “other” fuel type.
If the vehicle title status is clean, then price is more by 165% when compared to “parts only” title.
If the vehicle title status is lien, then price is less by 193% when compared to a “parts only” title.
If the vehicle title status is missing, then price is less by 98% when compared to “parts only” title.
If the vehicle title status is rebuilt, then price is less by 148% when compared to “parts only” title.
If the vehicle title status is salvage then price is less by 125% when compared to “parts only” title.
If vehicle transmission is automatic, price is less by 5% when compared to “other” transmission.
If vehicle transmission is manual, price is more by 10% when compared to “other” transmission.

3.2. Pricing Model with Manufacturer



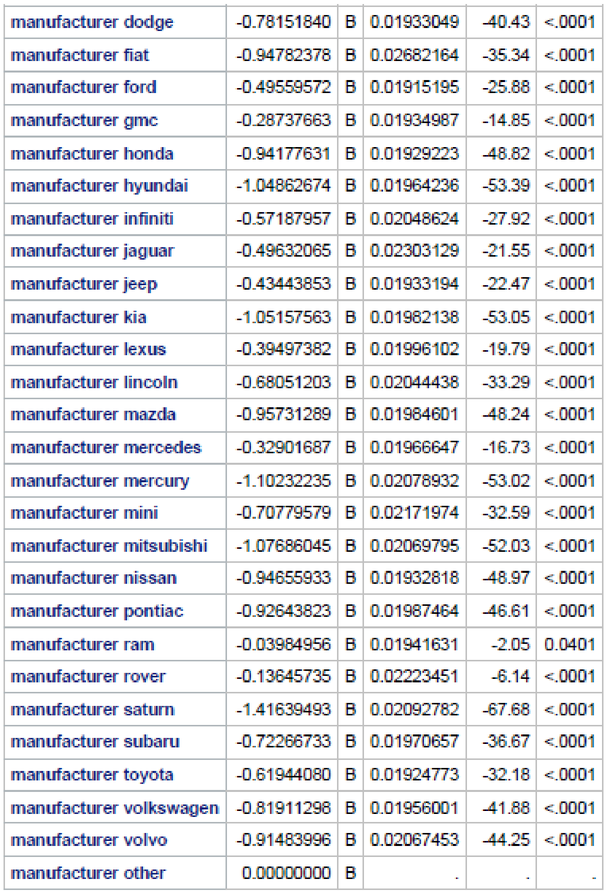
Dependent Variable: log_price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	34	619030.784	18206.788	27661.6	<.0001
Error	1.49E6	982848.954	0.658		
Corrected Total	1.49E6	1601879.738			

R-Square	Coeff Var	Root MSE	log_price Mean
0.386440	9.182980	0.811294	8.834755

Source	DF	Type I SS	Mean Square	F Value	Pr > F
log_age	1	521096.7980	521096.7980	791703	<.0001
manufacturer	33	97933.9858	2967.6965	4508.82	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
log_age	1	510626.1385	510626.1385	775795	<.0001
manufacturer	33	97933.9858	2967.6965	4508.82	<.0001

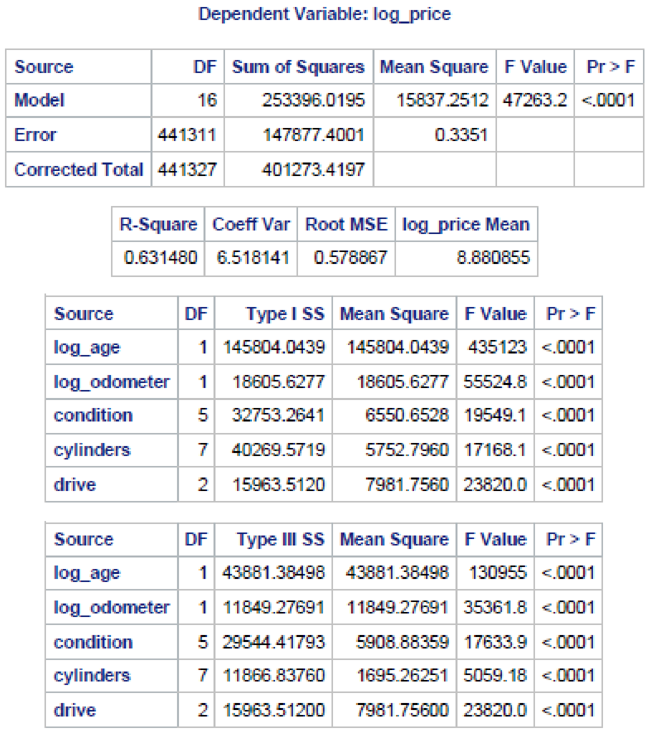


Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	11.67998703	B	0.01934005	603.93	<.0001
log_age	-0.91179136		0.00103519	-880.79	<.0001
manufacturer acura	-0.86742663	B	0.02006342	-43.23	<.0001
manufacturer audi	-0.55984240	B	0.02026168	-27.63	<.0001
manufacturer bmw	-0.46365622	B	0.01950542	-23.77	<.0001
manufacturer buick	-0.89020379	B	0.01986936	-44.80	<.0001
manufacturer cadillac	-0.47306003	B	0.01979686	-23.90	<.0001
manufacturer chevrolet	-0.47596619	B	0.01915690	-24.85	<.0001
manufacturer chrysler	-1.02437091	B	0.01965864	-52.11	<.0001
manufacturer dodge	-0.78151840	B	0.01933049	-40.43	<.0001
manufacturer fiat	-0.94782378	B	0.02682164	-35.34	<.0001
manufacturer ford	-0.49559572	B	0.01915195	-25.88	<.0001
manufacturer gmc	-0.28737663	B	0.01934987	-14.85	<.0001
manufacturer honda	-0.94177631	B	0.01929223	-48.82	<.0001
manufacturer hyundai	-1.04862674	B	0.01964236	-53.39	<.0001
manufacturer infiniti	-0.57187957	B	0.02048624	-27.92	<.0001
manufacturer jaguar	-0.49632065	B	0.02303129	-21.55	<.0001
manufacturer jeep	-0.43443853	B	0.01933194	-22.47	<.0001
manufacturer kia	-1.05157563	B	0.01982138	-53.05	<.0001
manufacturer lexus	-0.39497382	B	0.01996102	-19.79	<.0001
manufacturer lincoln	-0.68051203	B	0.02044438	-33.29	<.0001
manufacturer mazda	-0.95731289	B	0.01984601	-48.24	<.0001
manufacturer mercedes	-0.32901687	B	0.01966647	-16.73	<.0001
manufacturer mercury	-1.10232235	B	0.02078932	-53.02	<.0001
manufacturer mini	-0.70779579	B	0.02171974	-32.59	<.0001
manufacturer mitsubishi	-1.07686045	B	0.02069795	-52.03	<.0001
manufacturer nissan	-0.94655933	B	0.01932818	-48.97	<.0001
manufacturer pontiac	-0.92643823	B	0.01987464	-46.61	<.0001
manufacturer ram	-0.03984956	B	0.01941631	-2.05	0.0401
manufacturer rover	-0.13645735	B	0.02223451	-6.14	<.0001
manufacturer saturn	-1.41639493	B	0.02092782	-67.68	<.0001
manufacturer subaru	-0.72266733	B	0.01970657	-36.67	<.0001
manufacturer toyota	-0.61944080	B	0.01924773	-32.18	<.0001
manufacturer volkswagen	-0.81911298	B	0.01956001	-41.88	<.0001
manufacturer volvo	-0.91483996	B	0.02067453	-44.25	<.0001
manufacturer other	0.00000000	B	.	.	.

The conclusions about relation between price and age still hold valid in this regression model too. All estimates are found to be significant at 5% and R^2 is 38%. From the above, we can interpret:

Except Ram, all estimates are significant at 1%. Cheapest one is Saturn and costliest one is Ram. Other cheap vehicles include manufacturers like Mercury, Mitsubishi, Kia, Hyundai, and Chrysler. Other costly vehicles include manufacturers like Rover, GMC, Mercedes, Lexus, Jeep and BMW.

3.3. Pricing Model with Condition, Cylinders and Drive



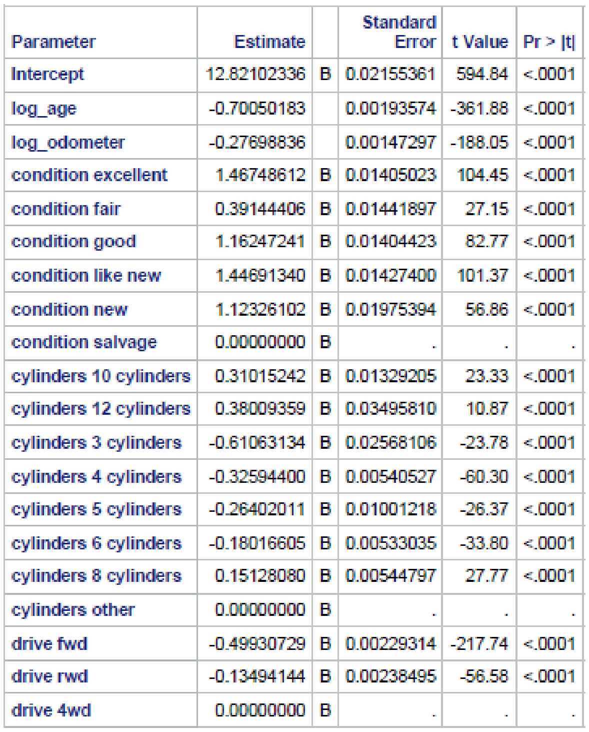
Dependent Variable: log_price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	253396.0195	15837.2512	47263.2	<.0001
Error	441311	147877.4001	0.3351		
Corrected Total	441327	401273.4197			

R-Square	Coeff Var	Root MSE	log_price Mean
0.631480	6.518141	0.578867	8.880855

Source	DF	Type I SS	Mean Square	F Value	Pr > F
log_age	1	145804.0439	145804.0439	435123	<.0001
log_odometer	1	18605.6277	18605.6277	55524.8	<.0001
condition	5	32753.2641	6550.6528	19549.1	<.0001
cylinders	7	40269.5719	5752.7960	17168.1	<.0001
drive	2	15963.5120	7981.7560	23820.0	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
log_age	1	43881.38498	43881.38498	130955	<.0001
log_odometer	1	11849.27691	11849.27691	35361.8	<.0001
condition	5	29544.41793	5908.88359	17633.9	<.0001
cylinders	7	11866.83760	1695.26251	5059.18	<.0001
drive	2	15963.51200	7981.75600	23820.0	<.0001

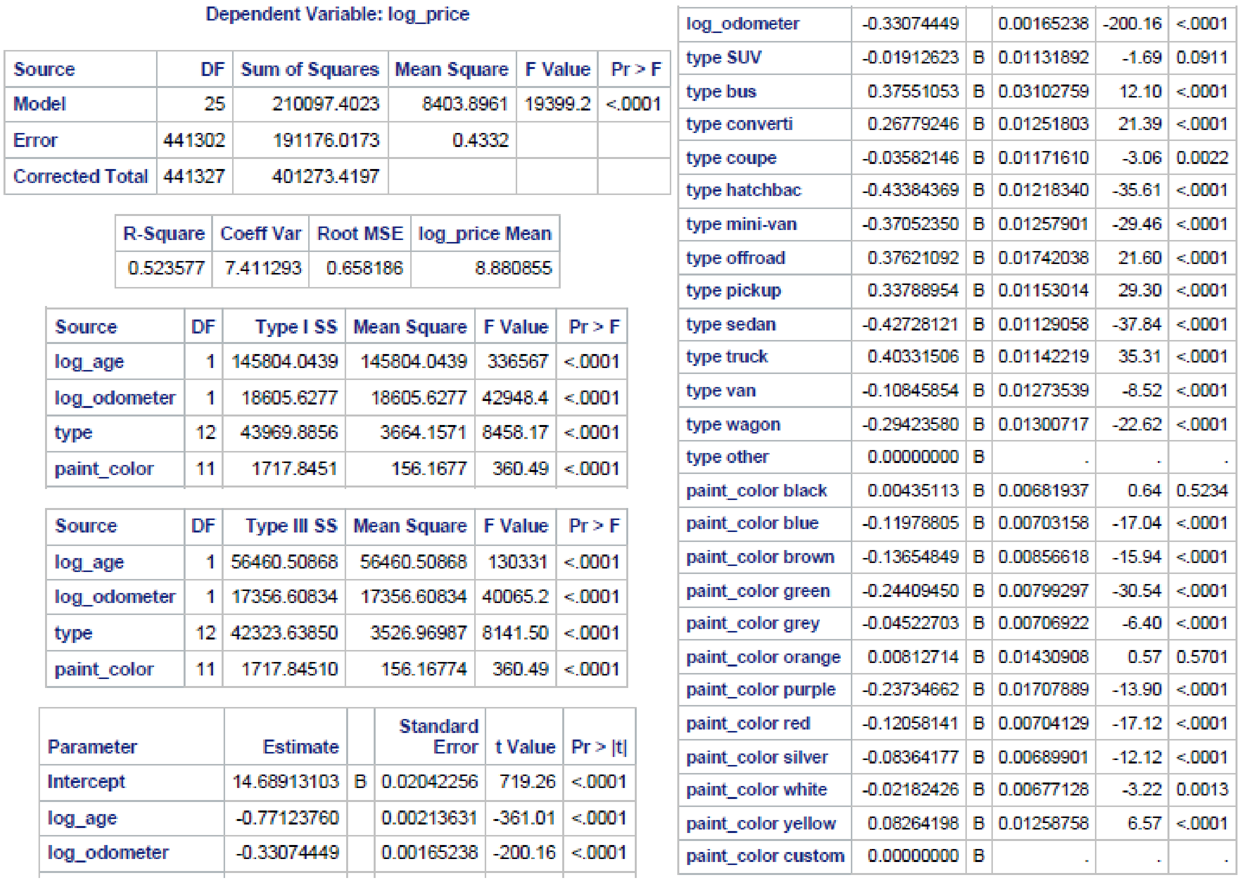


Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	12.82102336	B	0.02155361	594.84	<.0001
log_age	-0.70050183		0.00193574	-361.88	<.0001
log_odometer	-0.27698836		0.00147297	-188.05	<.0001
condition excellent	1.46748612	B	0.01405023	104.45	<.0001
condition fair	0.39144406	B	0.01441897	27.15	<.0001
condition good	1.16247241	B	0.01404423	82.77	<.0001
condition like new	1.44691340	B	0.01427400	101.37	<.0001
condition new	1.12326102	B	0.01975394	56.86	<.0001
condition salvage	0.00000000	B	.	.	.
cylinders 10 cylinders	0.31015242	B	0.01329205	23.33	<.0001
cylinders 12 cylinders	0.38009359	B	0.03495810	10.87	<.0001
cylinders 3 cylinders	-0.61063134	B	0.02568106	-23.78	<.0001
cylinders 4 cylinders	-0.32594400	B	0.00540527	-60.30	<.0001
cylinders 5 cylinders	-0.26402011	B	0.01001218	-26.37	<.0001
cylinders 6 cylinders	-0.18016605	B	0.00533035	-33.80	<.0001
cylinders 8 cylinders	0.15128080	B	0.00544797	27.77	<.0001
cylinders other	0.00000000	B	.	.	.
drive fwd	-0.49930729	B	0.00229314	-217.74	<.0001
drive rwd	-0.13494144	B	0.00238495	-56.58	<.0001
drive 4wd	0.00000000	B	.	.	.

All estimates are found to be significant at 1% and R^2 is 63%. From the above, we can interpret:

When the age of the vehicle increases by 10% then the sale price of the vehicle decreases by 7%.
When odometer reading increases by 10% then the sale price of the vehicle decreases by 2.77%.
If the vehicle condition is excellent, price is more by 146% when compared to a salvaged vehicle.
If the vehicle condition is fair, then price is more by 39% when compare with a salvaged vehicle.
If the vehicle condition is good, then price is more by 116% when compared to salvaged vehicle.
If the vehicle condition is like new, price is more by 144% when compared with a salvaged vehicle.
If the vehicle condition is new, then price is more by 112% when compared with salvaged vehicle.
If the vehicle has 3 cylinders, then price is less by 61% when compared to “other” cylinders count.
If the vehicle has 4 cylinders, then price is less by 32% when compared to “other” cylinders count.
If the vehicle has 5 cylinders, then price is less by 26% when compared to “other” cylinders count.
If the vehicle has 6 cylinders, then price is less by 18% when compared to “other” cylinders count.
If vehicle has 8 cylinders, then price is more by 15% when compared to “other” cylinders count.
If vehicle has 10 cylinders, then price is more by 31% when compared to “other” cylinders count.
If vehicle has 12 cylinders, then price is more by 38% when compared to “other” cylinders count.
If vehicle has front wheel drive, price is less by 50% when compared to four wheel drive vehicle.
If vehicle has rear wheel drive, price is less by 13% when compared to a four wheel drive vehicle.
As number of cylinders goes up, vehicle gets costly. Front wheel drive vehicles are more cheaper.

3.4. Price Model with Type and Color



Dependent Variable: log_price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	210097.4023	8403.8961	19399.2	<.0001
Error	441302	191176.0173	0.4332		
Corrected Total	441327	401273.4197			

R-Square	Coeff Var	Root MSE	log_price Mean
0.523577	7.411293	0.658186	8.880855

Source	DF	Type I SS	Mean Square	F Value	Pr > F
log_age	1	145804.0439	145804.0439	336567	<.0001
log_odometer	1	18605.6277	18605.6277	42948.4	<.0001
type	12	43969.8856	3664.1571	8458.17	<.0001
paint_color	11	1717.8451	156.1677	360.49	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
log_age	1	56460.50868	56460.50868	130331	<.0001
log_odometer	1	17356.60834	17356.60834	40065.2	<.0001
type	12	42323.63850	3526.96987	8141.50	<.0001
paint_color	11	1717.84510	156.16774	360.49	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	14.68913103	B	0.02042256	719.26	<.0001
log_age	-0.77123760		0.00213631	-361.01	<.0001
log_odometer	-0.33074449		0.00165238	-200.16	<.0001
type SUV	-0.01912623	B	0.01131892	-1.69	0.0911
type bus	0.37551053	B	0.03102759	12.10	<.0001
type converti	0.26779246	B	0.01251803	21.39	<.0001
type coupe	-0.03582146	B	0.01171610	-3.06	0.0022
type hatchbac	-0.43384369	B	0.01218340	-35.61	<.0001
type mini-van	-0.37052350	B	0.01257901	-29.46	<.0001
type offroad	0.37621092	B	0.01742038	21.60	<.0001
type pickup	0.33788954	B	0.01153014	29.30	<.0001
type sedan	-0.42728121	B	0.01129058	-37.84	<.0001
type truck	0.40331506	B	0.01142219	35.31	<.0001
type van	-0.10845854	B	0.01273539	-8.52	<.0001
type wagon	-0.29423580	B	0.01300717	-22.62	<.0001
type other	0.00000000	B	.	.	.
paint_color black	0.00435113	B	0.00681937	0.64	0.5234
paint_color blue	-0.11978805	B	0.00703158	-17.04	<.0001
paint_color brown	-0.13654849	B	0.00856618	-15.94	<.0001
paint_color green	-0.24409450	B	0.00799297	-30.54	<.0001
paint_color grey	-0.04522703	B	0.00706922	-6.40	<.0001
paint_color orange	0.00812714	B	0.01430908	0.57	0.5701
paint_color purple	-0.23734662	B	0.01707889	-13.90	<.0001
paint_color red	-0.12058141	B	0.00704129	-17.12	<.0001
paint_color silver	-0.08364177	B	0.00689901	-12.12	<.0001
paint_color white	-0.02182426	B	0.00677128	-3.22	0.0013
paint_color yellow	0.08264198	B	0.01258758	6.57	<.0001
paint_color custom	0.00000000	B	.	.	.

All the estimates except SUV type, Black and Orange colors, are significant at 1% and R^2 is 52%.

When the age of the vehicle increases by 10% then the sale price of the vehicle decreases by 8%. When odometer reading increases by 10% then the sale price of the vehicle decreases by 3.31%. If vehicle type is Bus, Covertible, Offroad, Pickup, Truck then sale price is more than “other” type. If vehicle is SUV, Coupe, Hatchback, Mini-Van, Sedan, Van, Wagon, price is less than “other” type. All the above vehicle colors except for Yellow color have prices less than Custom colored vehicles. Purple and Green color vehicles have least sale price when compared to Custom colored vehicles.

4. Conclusion

According to our regression results, a vehicle with below attributes fetches best possible price:

A Pickup, Truck or Offroad type vehicle that is in New or Excellent condition with Diesel fuel type and Clean title status, having Manual transmission, more than 6 cylinders and Four wheel drive, manufactured by Ram, Rover, GMC, Mercedes, Lexus, Jeep or BMW will get the highest sale price.