PREDECTIVE ANALYTICS PROJECT

123837205

DWIJ DUA

EXPLORATORY DATA ANALYTICS

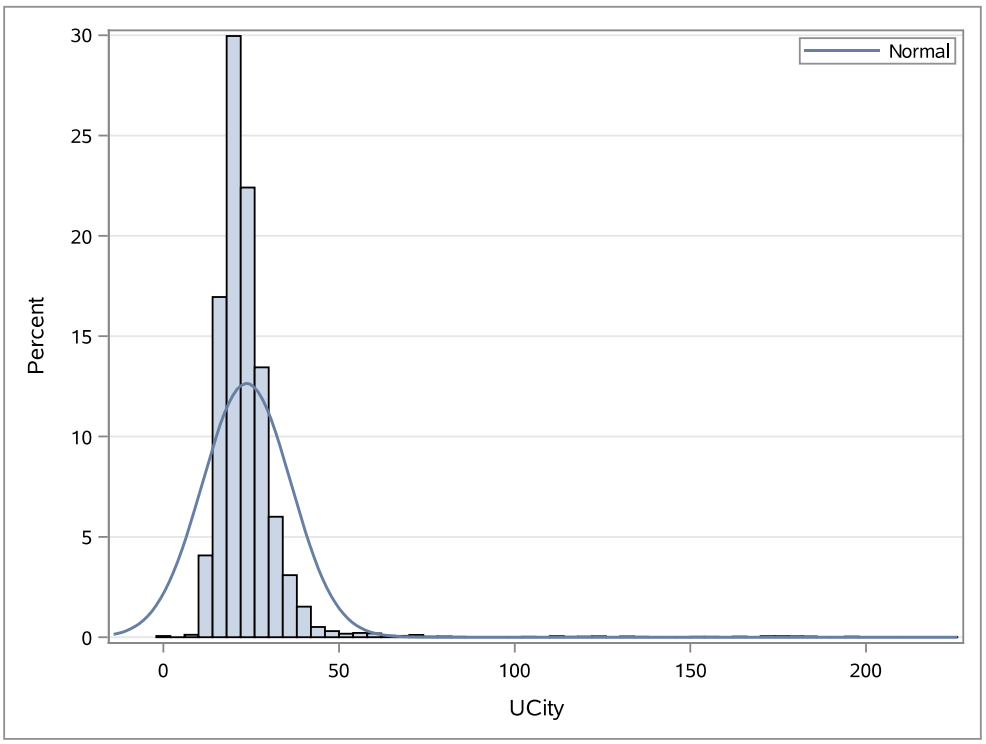
Our dataset has 43921 rows and 83 variables.

UCity - unadjusted city MPG for fuel type1 is the dependent variable and other 82 are independent variables.

We have numeric variables (for example charge, displ, cylinders, etc.), Categorical variables (for example trany, vClass, etc.) and others, such as model, id, etc.

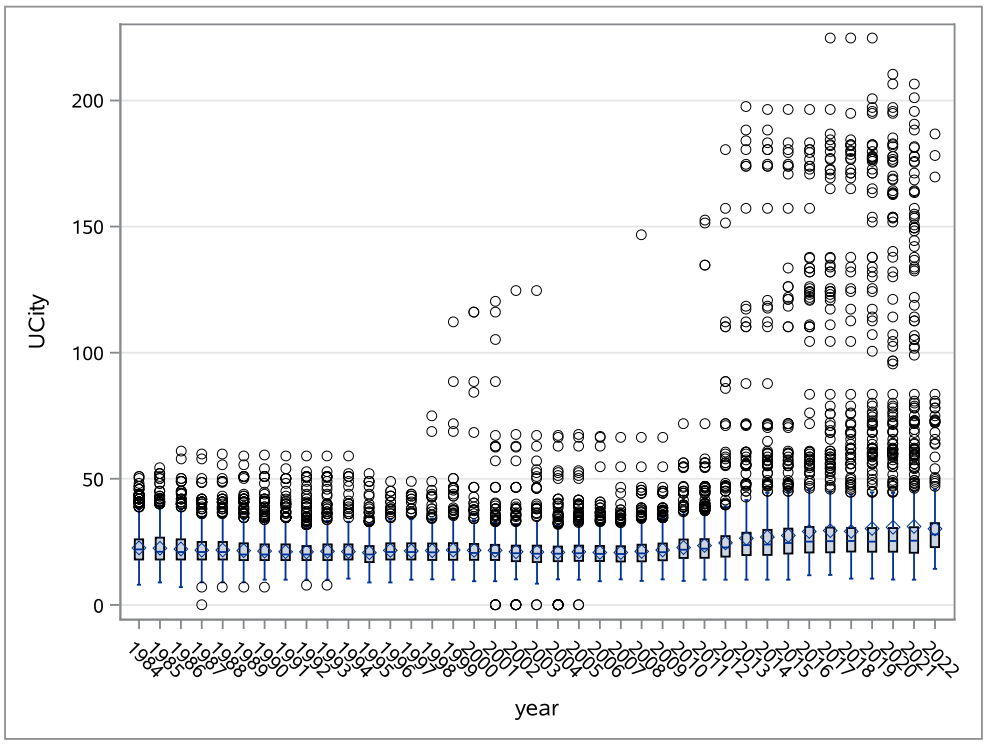
I imported the data in SAS Studio carry our EDA and Data Cleaning.

VARIABLE UCity (Dependent)

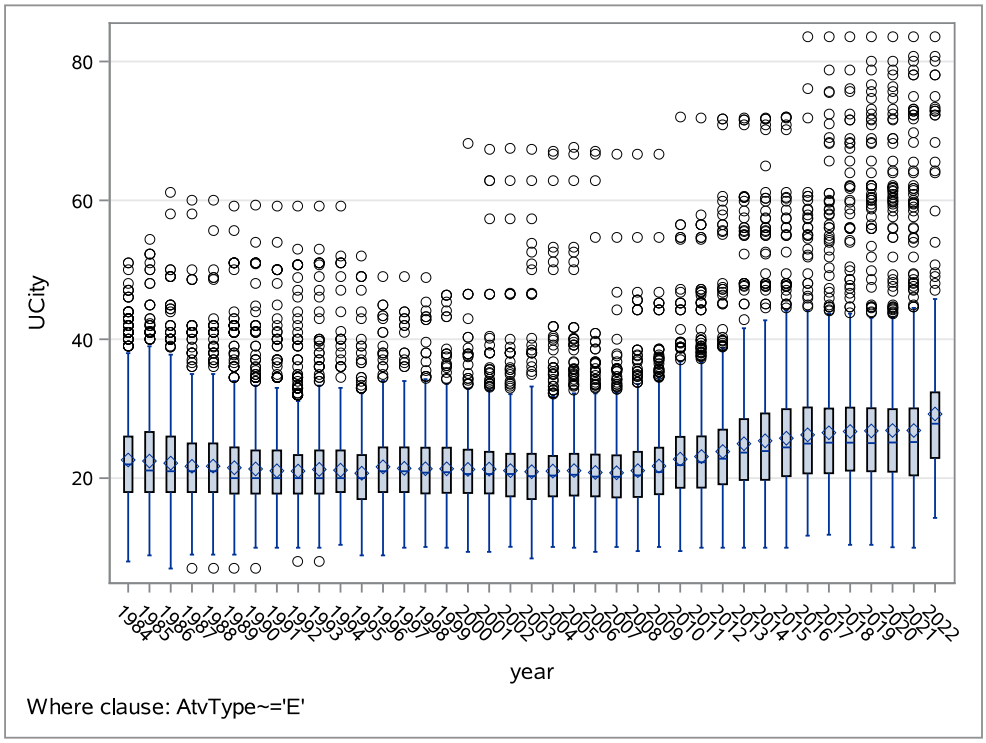


The variable is right skewed in nature.

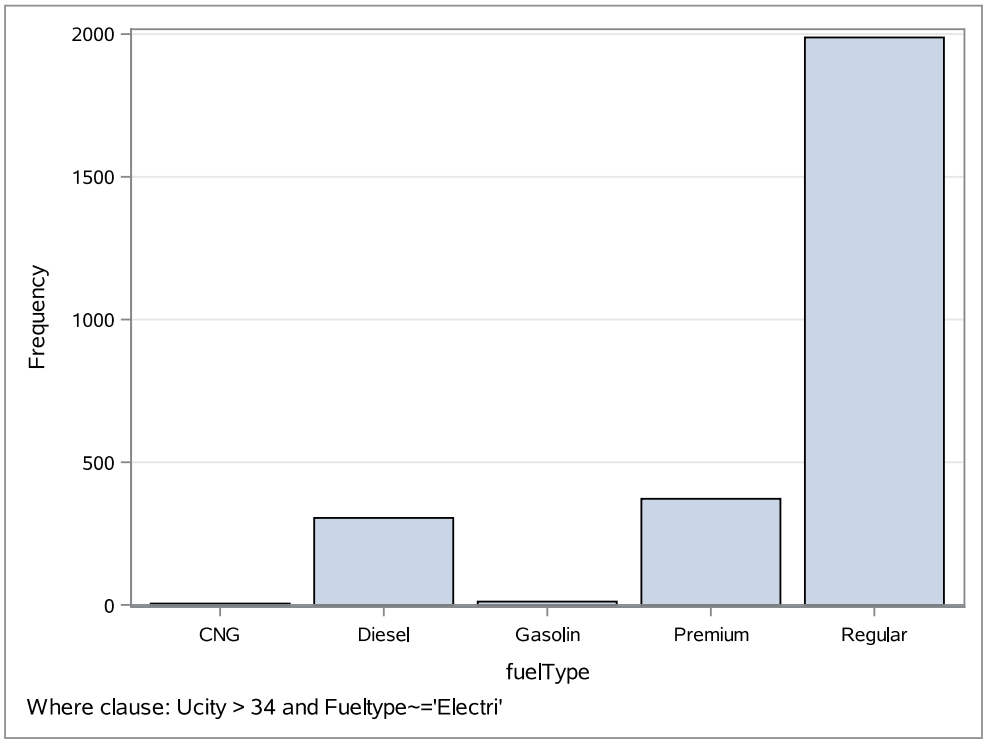
The minimum and maximum values of UCity are 0 and 224.8 respectively. Values beyond 34 are outliers in our data, which are mostly coming from Electric cars. Thus, those records and the records where the value of UCity is 0 have not been considered for model building.



The above boxplot shows high number of outliers, which are increasing in value as we move down the timeline.



Most of the outliers were because of Electric cars. Another boxplot was created to look at outliers not coming from electric cars. Hybrid cars too were giving many outliers. Over the years, hybrid cars were gaining popularity, giving higher mileages.



A frequency graph for the outliers in our data not coming from Electric cars is shown above.

Basic Statistical Measures			
Location		**Variability**	
Mean	22.90002	**Std Deviation**	7.19209
Median	21.62420	**Variance**	51.72615
Mode	20.00000	**Range**	76.55980
		Interquartile Range	7.60000

Moments			
N	43609	**Sum Weights**	43609
Mean	22.90002	**Sum Observations**	998646.97
Std Deviation	7.1920893	**Variance**	51.7261485
Skewness	2.05257861	**Kurtosis**	8.98410579
Uncorrected SS	25124709.4	**Corrected SS**	2255673.88
Coeff Variation	31.4064761	**Std Error Mean**	0.03444031

We have summary statistics for UCity including mean, median, standard deviation, skewness, kurtosis, etc.

INDEPENDENT VARIABLES

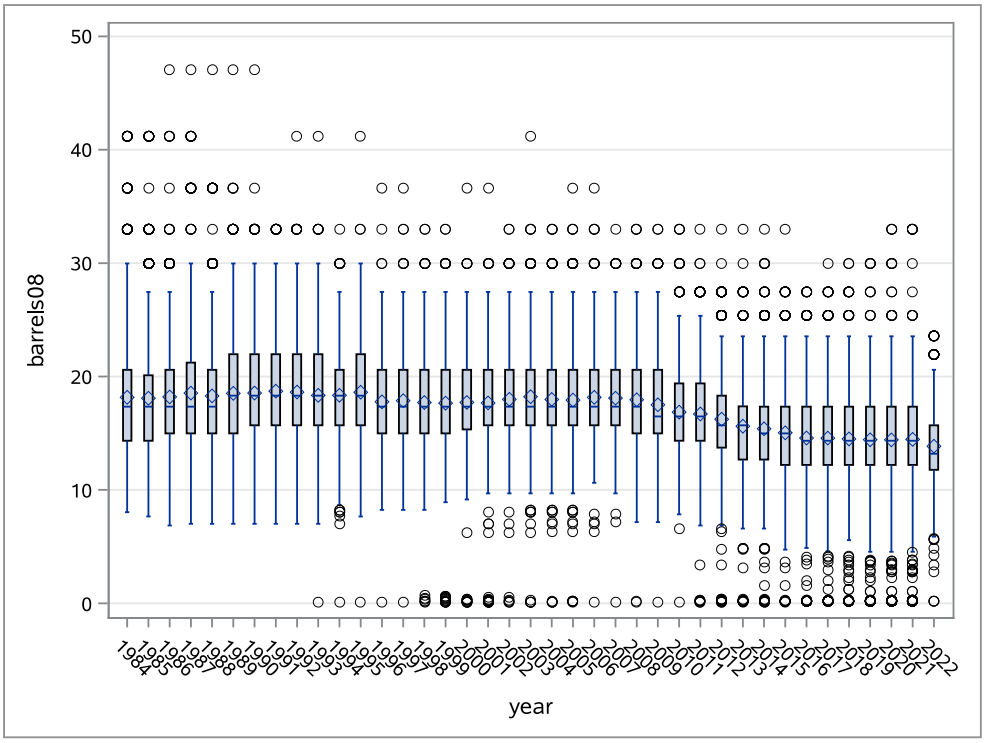
Numeric Variables:

1. **Barrels –** annual petroleum consumption in barrels for fuel type 1

Basic Statistical Measures			
Location		**Variability**	
Mean	17.09953	**Std Deviation**	4.68205
Median	16.48050	**Variance**	21.92162
Mode	18.31167	**Range**	47.02714
		Interquartile Range	5.05795

Moments			
N	43896	**Sum Weights**	43896
Mean	17.0995256	**Sum Observations**	750600.775
Std Deviation	4.68205301	**Variance**	21.9216204
Skewness	0.33978267	**Kurtosis**	2.12072942
Uncorrected SS	13797166.7	**Corrected SS**	962249.529
Coeff Variation	27.3811866	**Std Error Mean**	0.02234724

The above tables give summary statistics for barrels.



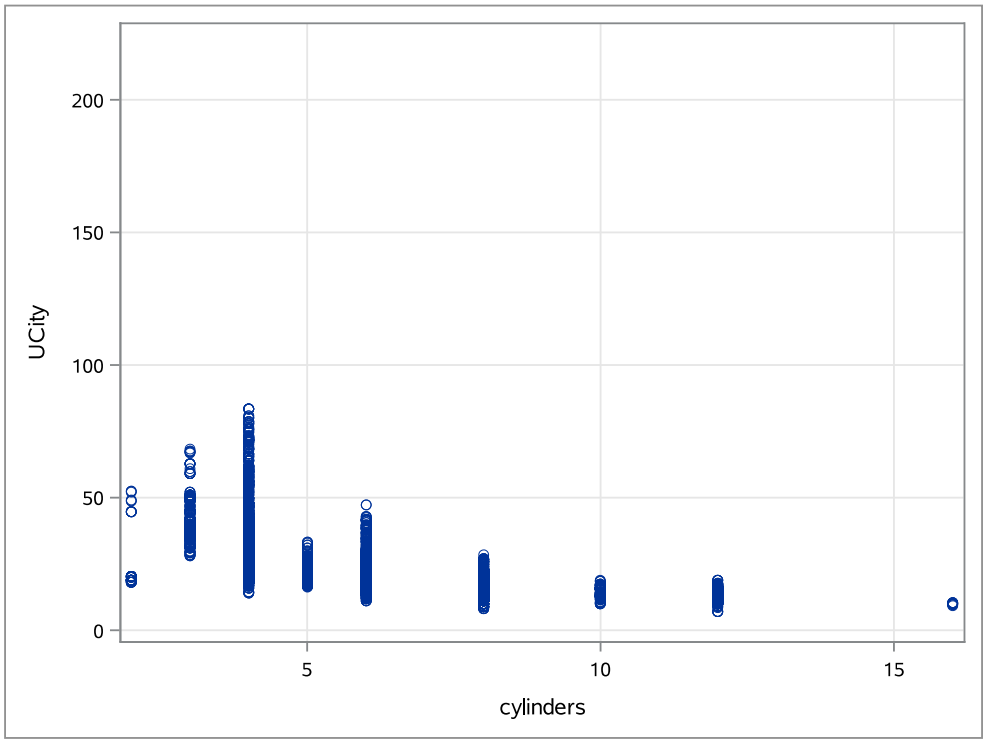
We have the annual fuel consumption of petroleum in barrels for fuel type 1 over the years. There is a declining trend in the number of barrels over the years.

2. Cylinders – The number of engine cylinders in the vehicle.

Basic Statistical Measures			
Location		**Variability**	
Mean	5.707517	**Std Deviation**	1.76656
Median	6.000000	**Variance**	3.12074
Mode	4.000000	**Range**	14.00000
		Interquartile Range	2.00000

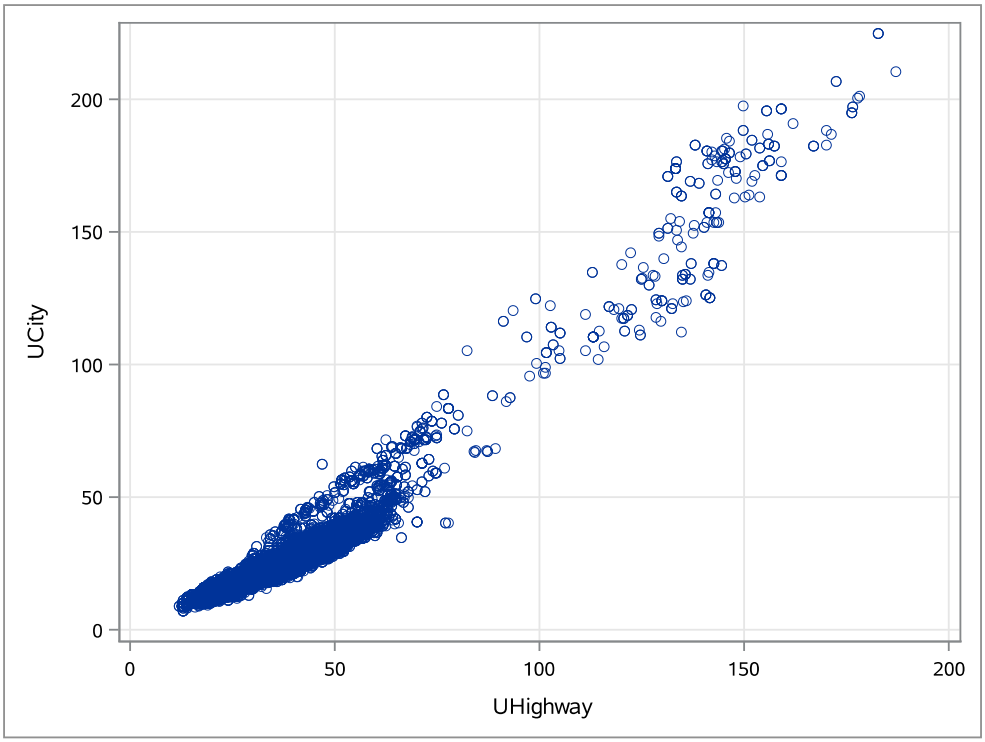
cylinders	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	63	0.14	63	0.14
3	348	0.80	411	0.94
4	17160	39.35	17571	40.29
5	774	1.77	18345	42.07
6	15050	34.51	33395	76.58
8	9328	21.39	42723	97.98
10	187	0.43	42910	98.40
12	682	1.56	43592	99.97
16	14	0.03	43606	100.00
Frequency Missing = 290				

Summary statistics display a total of 290 missing values.



From the above plot, it is clear that the lesser the number of cylinders, the higher the value of city MPG.

3. Highway – Gives information about unadjusted highway MPG.



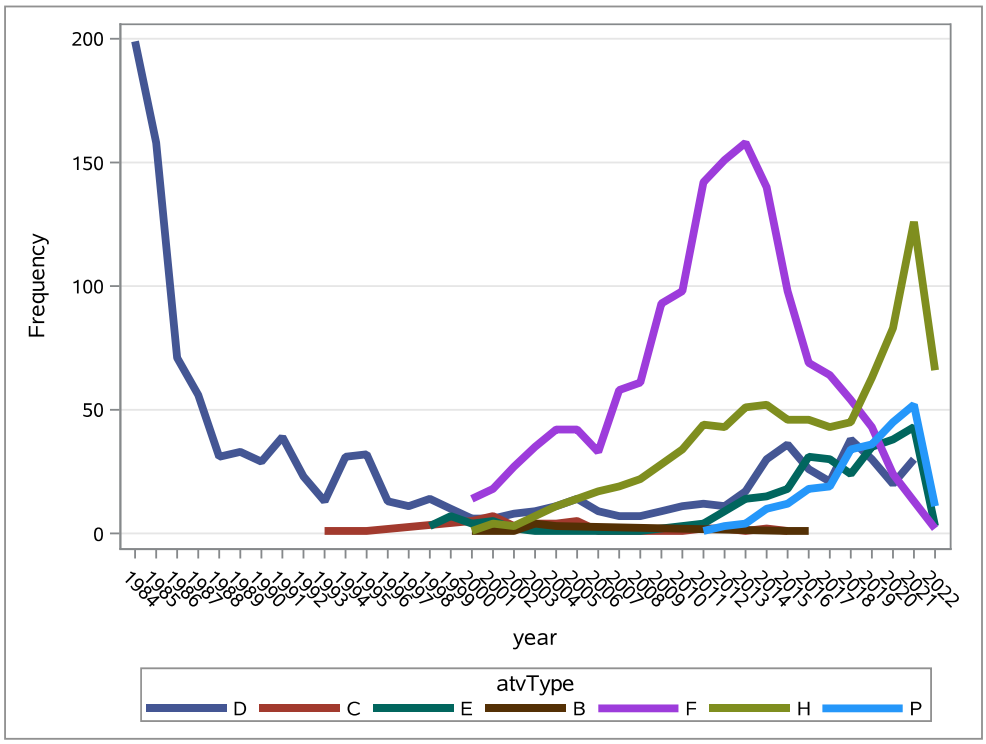
We see that the variables are positively related to each other.

4. ATV Type – It describes the type of alternative fuel or advanced technology vehicle

atvType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
B	12	0.30	12	0.30
C	44	1.08	56	1.38
D	1131	27.82	1187	29.19
E	287	7.06	1474	36.25
F	1479	36.37	2953	72.63
H	868	21.35	3821	93.97
P	245	6.03	4066	100.00
Frequency Missing = 39830				

The ATV types are as follows :
B : Bifuel, C : CNG, D : Diesel, E : EV, F : FVV, H : Hybrid and P : Plug-in hybrid

91.33% of observations are missing.



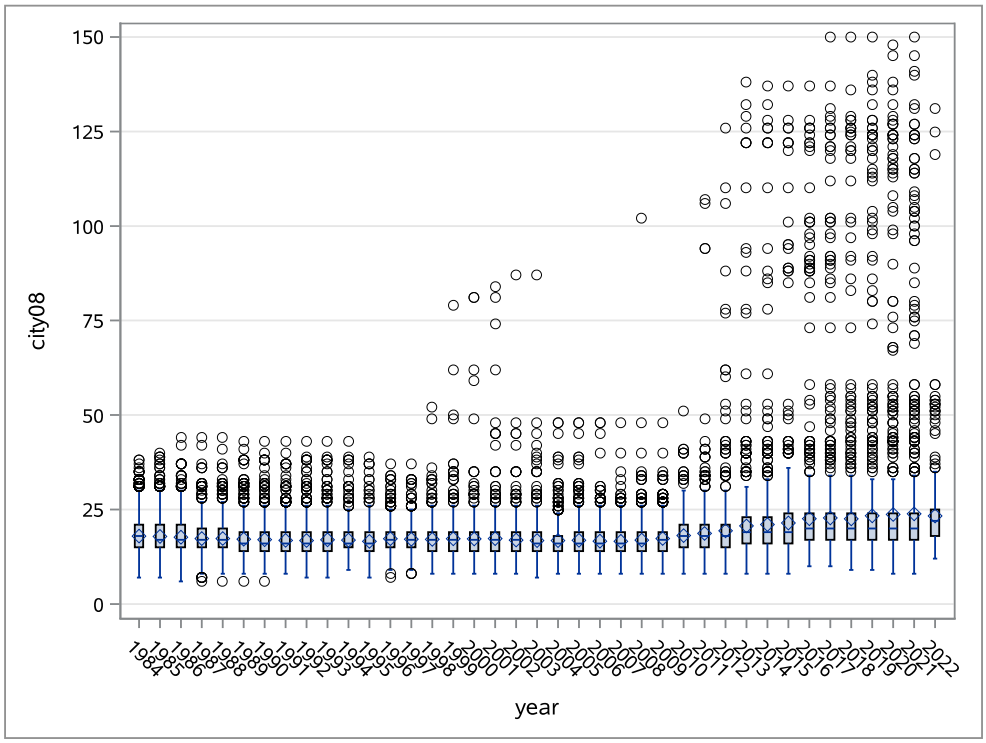
The graph depicts the decline in Diesel vehicles and a rise in hybrid vehicles.

5. City - city MPG for fuelType1

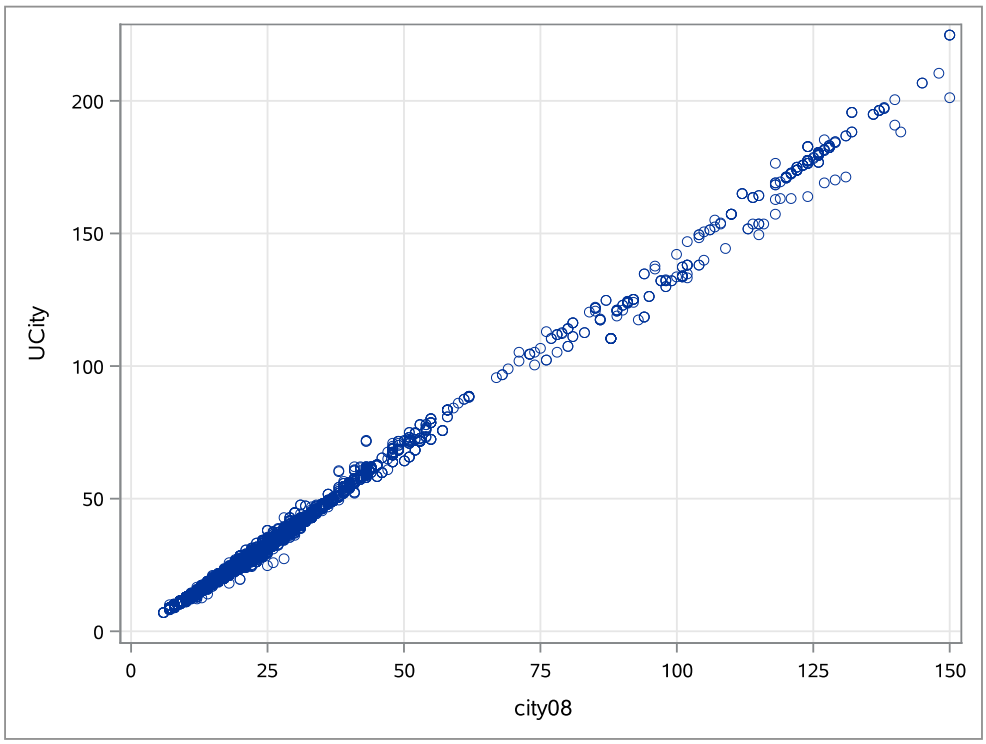
Basic Statistical Measures			
Location		**Variability**	
Mean	18.71521	**Std Deviation**	8.87409
Median	17.00000	**Variance**	78.74944
Mode	15.00000	**Range**	144.00000
		Interquartile Range	6.00000

Moments			
N	43896	**Sum Weights**	43896
Mean	18.7152132	**Sum Observations**	821523
Std Deviation	8.87408839	**Variance**	78.7494448
Skewness	7.50596322	**Kurtosis**	80.3659779
Uncorrected SS	18831685	**Corrected SS**	3456706.88
Coeff Variation	47.4164429	**Std Error Mean**	0.04235564

Above are the summary statistics.

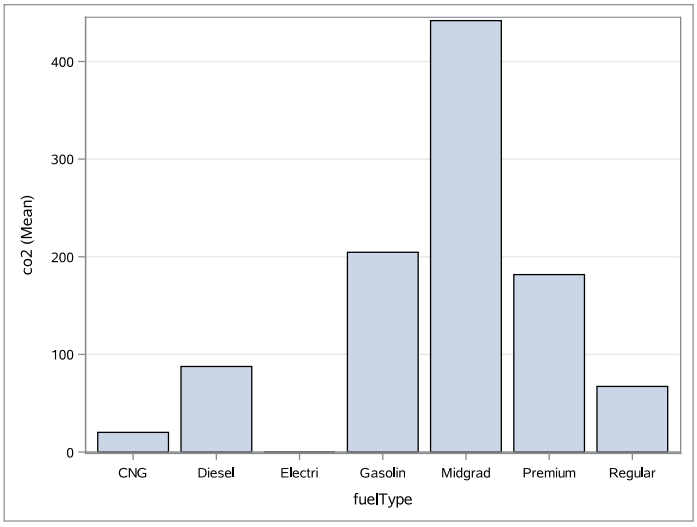


We can see that over the years, both the frequency and value of outliers have increased. This might be because of the use of Electric and Hybrid cars, with give a greater value of mpg.



Further, a strong positive relationship can be seen between UCity and city08.

6. **Fuel Type – We have the following types:**
   - fuelType : fuel type with fuelType1 and fuelType2 (if applicable)
   - fuelType1 : fuel type 1. For single fuel vehicles, this will be the only fuel. For dual fuel vehicles, this will be the conventional fuel.
   - fuelType2 : fuel type 2. For dual fuel vehicles, this will be the alternative fuel (e.g. E85, Electricity, CNG, LPG). For single fuel vehicles, this field is not used



co2 representstailpipe CO2 in grams/mile for fuelType.

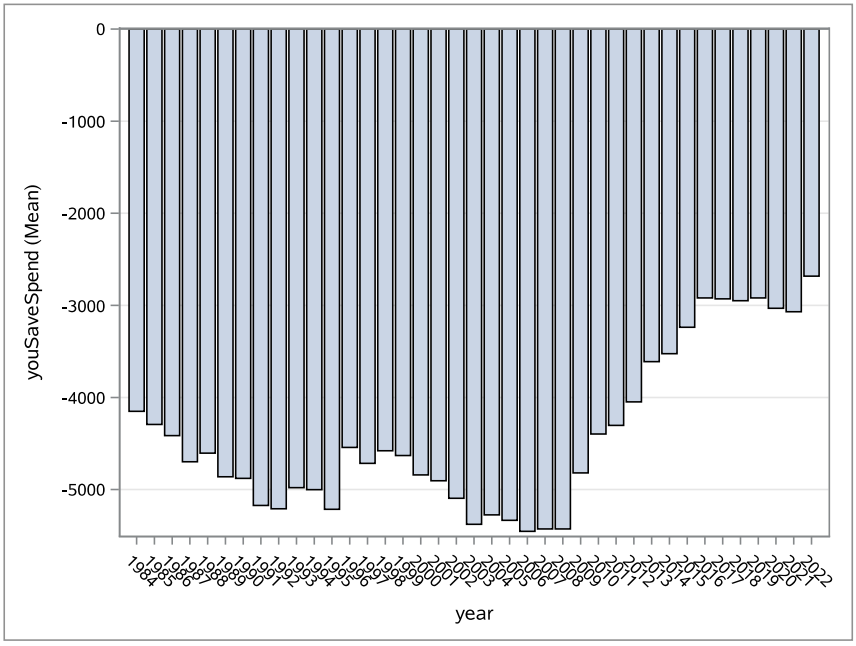
The bar graph shows mean consumption of co2 (in grams per mile) for various fuel types, where ‘Midgrad’ has the highest value and Electric vehicles show the lowest value.

7. youSaveSpend – Money in $ an individual saves or spends over 5 years compared to an average car

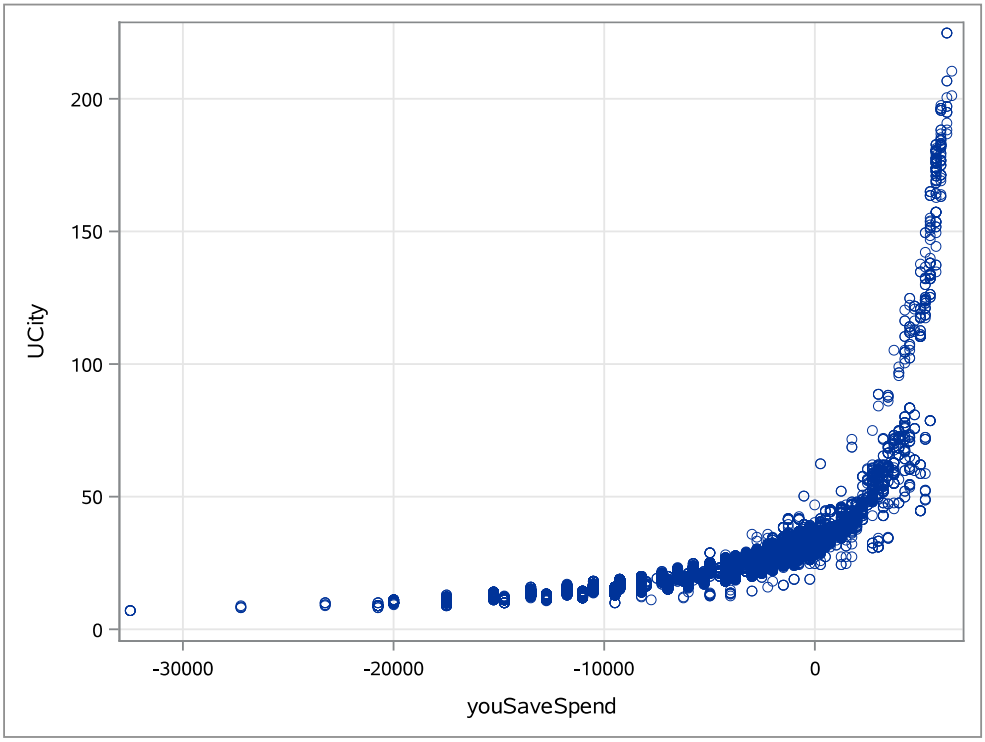
Basic Statistical Measures			
Location		**Variability**	
Mean	-4358.31	**Std Deviation**	3701
Median	-4250.00	**Variance**	13698825
Mode	-4250.00	**Range**	39000
		Interquartile Range	4500

Moments			
N	43896	**Sum Weights**	43896
Mean	-4358.3128	**Sum Observations**	-191312500
Std Deviation	3701.19231	**Variance**	13698824.5
Skewness	-0.6817707	**Kurtosis**	1.69512653
Uncorrected SS	1.43511E12	**Corrected SS**	6.0131E11
Coeff Variation	-84.922594	**Std Error Mean**	17.6656319

Summary statistics for the variable are given.

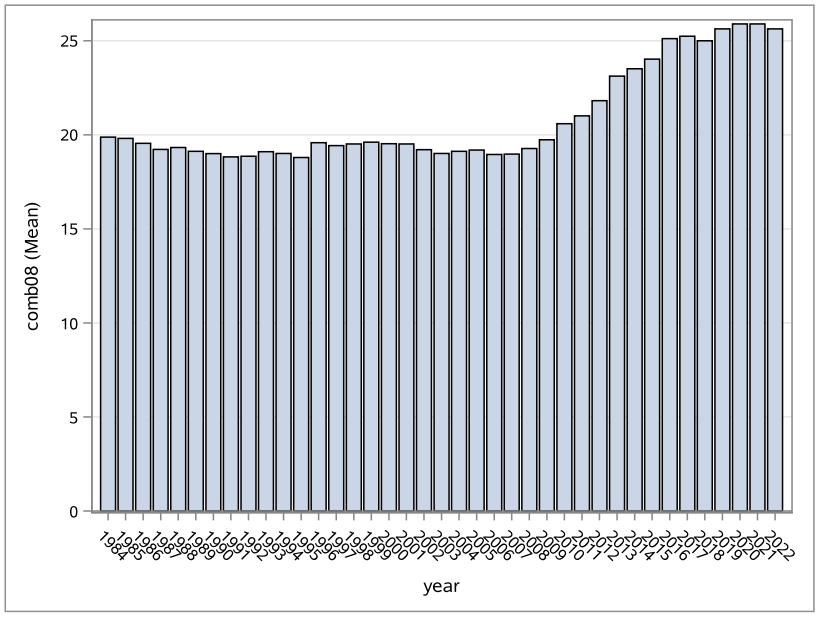


Maximum amount of money was spent by people in the years 2003-2009.

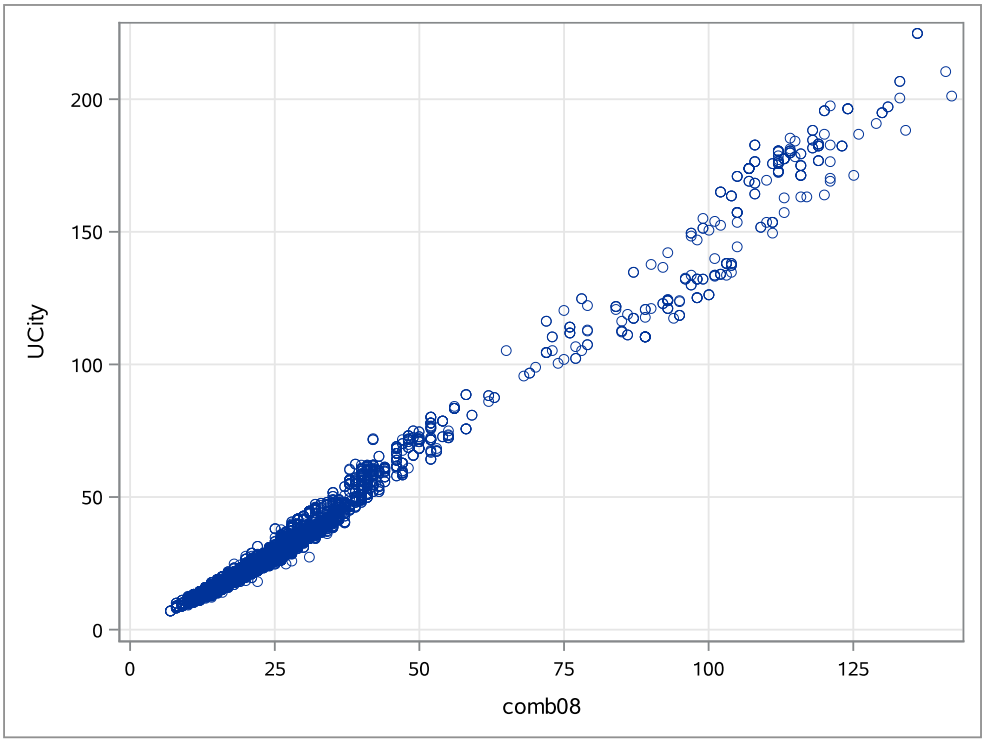


As the value of mpg reaches 50, that's when a person starts saving money. As the mileage value keeps increasing after 50, we see an exponential growth in savings.

8. Comb – combined MPG for fuelType1 (comb08).



The values for combined MPG for fuelType1 gradually increase over the years.



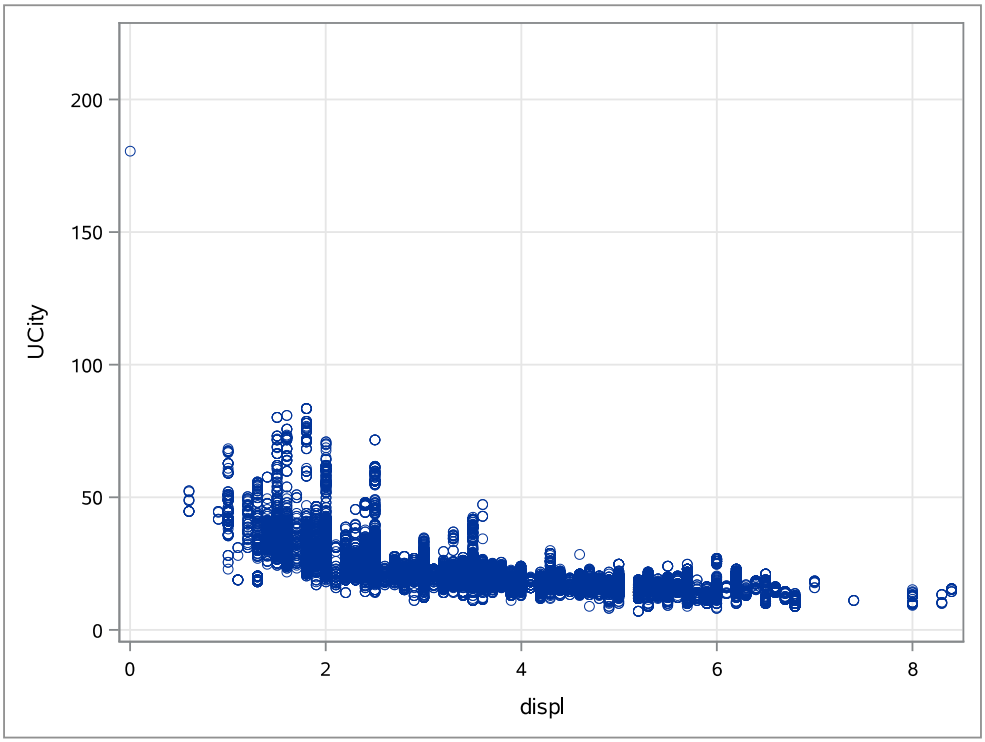
A strong positive relation exists between comb08 and UCity.

9. Displ – Represents combined volume of pistons inside the cylinders.

Basic Statistical Measures			
Location		**Variability**	
Mean	3.280616	**Std Deviation**	1.35679
Median	3.000000	**Variance**	1.84087
Mode	2.000000	**Range**	8.40000
		Interquartile Range	2.00000

Moments			
N	43608	**Sum Weights**	43608
Mean	3.28061594	**Sum Observations**	143061.1
Std Deviation	1.35678721	**Variance**	1.84087153
Skewness	0.66461585	**Kurtosis**	-0.4808918
Uncorrected SS	549603.41	**Corrected SS**	80274.8847
Coeff Variation	41.3576972	**Std Error Mean**	0.00649724

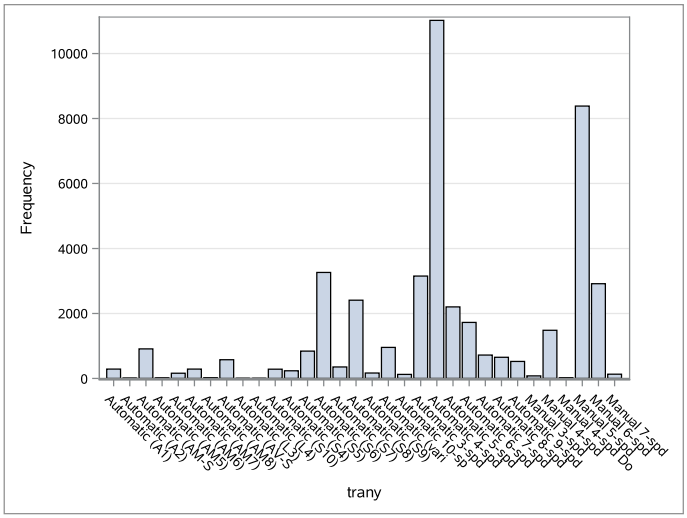
The summary statistics are given above.



The scatter plot shows an inverse relationship between the number of cylinders and city MPG.

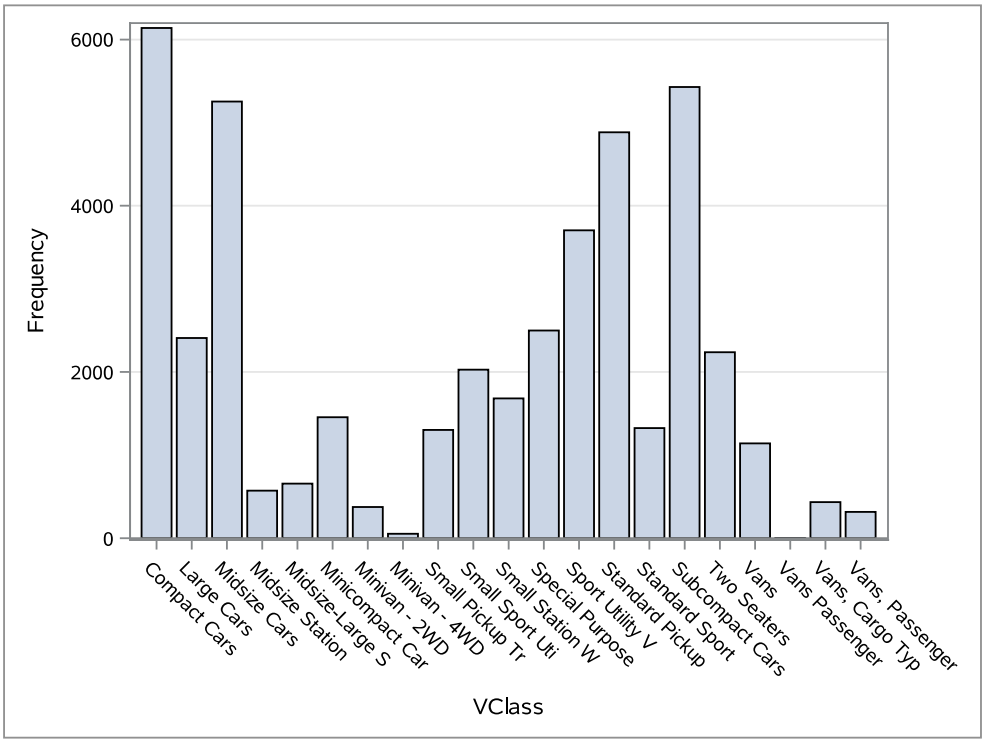
Categorical Variables

1. Trany – transmission type



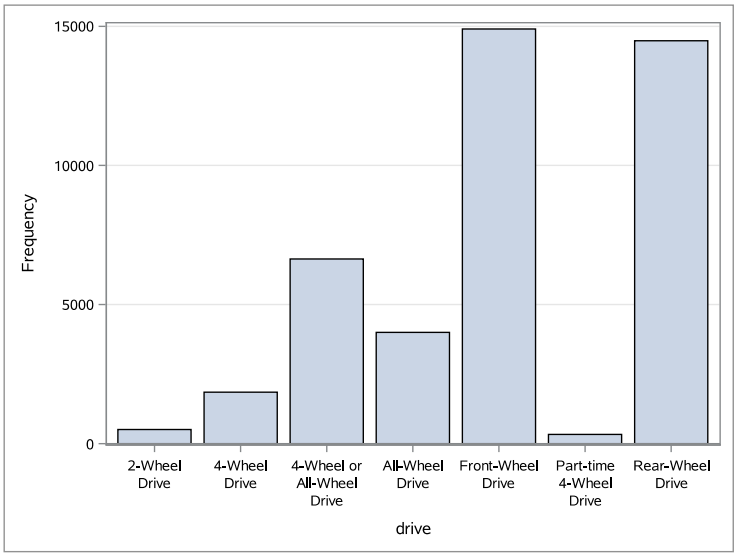
The most common transmission type in our dataset is Automatic 4 and 4 Speed.

2. Vclass – Gives information about EPA vehicle class sizes.



Compact cars are the most frequent in our data and Vans passenger vehicles have the least number of records.

3. Drive – Represents drive axle type



Thefrequency distribution of the variable is shown. There are maximum number of Front-Wheel driven vehicles and minimum number of Part-time 4-Wheel driven vehicles.

Correlation Analysis

11 Variables:	barrels08 city08 co2 comb08 cylinders displ fuelCost08 rangeHwy year youSaveSpend UCity

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

	barrels08	city08	co2	comb08	cylinders	displ	fuelCost08	rangeHwy	year	youSaveSpend	UCity
barrels08	1.00000 43896	-0.71265 <.0001 43896	-0.13242 <.0001 43896	-0.76766 <.0001 43896	0.73497 <.0001 43606	0.78584 <.0001 43608	0.91974 <.0001 43896	-0.24961 <.0001 43896	-0.28459 <.0001 43896	-0.92469 <.0001 43896	-0.69551 <.0001 43896
city08	-0.71265 <.0001 43896	1.00000 43896	0.03506 <.0001 43896	0.99150 <.0001 43896	-0.68024 <.0001 43606	-0.71384 <.0001 43608	-0.65031 <.0001 43896	0.67630 <.0001 43896	0.20503 <.0001 43896	0.65128 <.0001 43896	0.99808 <.0001 43896
co2	-0.13242 <.0001 43896	0.03506 <.0001 43896	1.00000 43896	0.05699 <.0001 43896	0.09919 <.0001 43606	0.06527 <.0001 43608	-0.03854 <.0001 43896	-0.03957 <.0001 43896	0.70656 <.0001 43896	0.03961 <.0001 43896	0.03385 <.0001 43896
comb08	-0.76766 <.0001 43896	0.99150 <.0001 43896	0.05699 <.0001 43896	1.00000 43896	-0.68393 <.0001 43606	-0.72850 <.0001 43608	-0.70052 <.0001 43896	0.66084 <.0001 43896	0.23338 <.0001 43896	0.70087 <.0001 43896	0.98749 <.0001 43896
cylinders	0.73497 <.0001 43606	-0.68024 <.0001 43606	0.09919 <.0001 43606	-0.68393 <.0001 43606	1.00000 43606	0.90507 <.0001 43606	0.77628 <.0001 43606	. . 43606	0.05232 <.0001 43606	-0.77578 <.0001 43606	-0.66154 <.0001 43606
displ	0.78584 <.0001 43608	-0.71384 <.0001 43608	0.06527 <.0001 43608	-0.72850 <.0001 43608	0.90507 <.0001 43606	1.00000 43608	0.76800 <.0001 43608	-0.01158 0.0156 43608	0.00476 0.3205 43608	-0.76801 <.0001 43608	-0.69443 <.0001 43608
fuelCost08	0.91974 <.0001 43896	-0.65031 <.0001 43896	-0.03854 <.0001 43896	-0.70052 <.0001 43896	0.77628 <.0001 43606	0.76800 <.0001 43608	1.00000 43896	-0.18420 <.0001 43896	-0.13482 <.0001 43896	-0.99853 <.0001 43896	-0.63205 <.0001 43896
rangeHwy	-0.24961 <.0001 43896	0.67630 <.0001 43896	-0.03957 <.0001 43896	0.66084 <.0001 43896	. . 43606	-0.01158 0.0156 43608	-0.18420 <.0001 43896	1.00000 43896	0.09121 <.0001 43896	0.18324 <.0001 43896	0.67081 <.0001 43896
year	-0.28459 <.0001 43896	0.20503 <.0001 43896	0.70656 <.0001 43896	0.23338 <.0001 43896	0.05232 <.0001 43606	0.00476 0.3205 43608	-0.13482 <.0001 43896	0.09121 <.0001 43896	1.00000 43896	0.13935 <.0001 43896	0.20480 <.0001 43896
youSaveSpend	-0.92469 <.0001 43896	0.65128 <.0001 43896	0.03961 <.0001 43896	0.70087 <.0001 43896	-0.77578 <.0001 43606	-0.76801 <.0001 43608	-0.99853 <.0001 43896	0.18324 <.0001 43896	0.13935 <.0001 43896	1.00000 43896	0.63330 <.0001 43896
UCity	-0.69551 <.0001 43896	0.99808 <.0001 43896	0.03385 <.0001 43896	0.98749 <.0001 43896	-0.66154 <.0001 43606	-0.69443 <.0001 43608	-0.63205 <.0001 43896	0.67081 <.0001 43896	0.20480 <.0001 43896	0.63330 <.0001 43896	1.00000 43896

A correlation matrix was created for above mentioned numeric variables. The ones marked in red have a strong negative correlation and the ones marked in green display a strong positive correlation.

Pearson Correlation Coefficients
Number of Observations

	cylinders	barrels08	city08	cityA08	co2	comb08	displ	highway08	fuelCost08	youSaveSpend
UCity	-0.66154 43606	-0.69551 43896	0.99808 43896	0.09261 43896	0.03385 43896	0.98749 43896	-0.69443 43608	0.93859 43896	-0.63205 43896	0.63330 43896

The variables city08, comb08, highway08, year have a strong positive correlation with UCity.

Data Cleaning and Preparation

The next step was to clean and prepare our data to build a model upon.

The procedure is as follows :

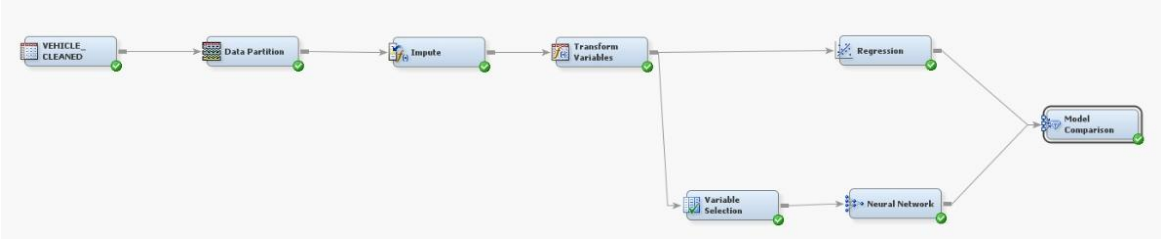
- Dropping records for Electric and Hybrid vehicles.
- Dropping all variables except Barrels08, Ucity, Uhighway, Vclass, city08, comb08, cylinders, displ, drive, fueltype, highway08, mpgData, fuelcost08, Trany, year and yousavespend.
- Rounding UCity, Barrels and UHighway to the nearest quarter.

Converting Categorical to Numeric variables using the following transformations:

- mpgData – 1 : 'Y', 2 : otherwise.
- Fuel type – 1 : CNG, 2 : Diesel, 3 : Midgrad, 4 : Premium and 5 : otherwise.
- Drive – 1 : Rear-wheel drive, 2 : Two-wheel drive (Front wheel, 2-wheel and part time 4-wheel), 3 : 4-wheel drive.
- Transmission – 1 : Manual, 2 : otherwise.
- Vclass – 1 : Mini, 2 : Vans, 3 : Mid sized, 4 : Small sized, 5 : Standard sized, 6 : Sports vehicles, 7 : Special vehicles and 8 : Large vehicles.

- Imputing outliers in UCity with mean.
- Imputing missing values in UHighway, barrels08, cylinders and displacement with their respective means.
- Exporting the dataset to be used for model building in SAS Miner

Models



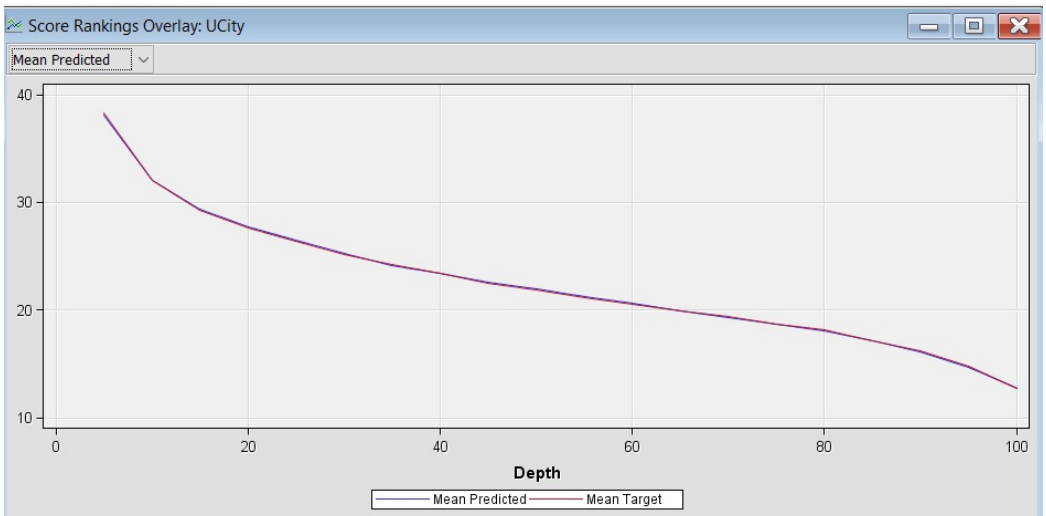
After loading the cleaned dataset, I portioned the dataset into training and testing using a 80:20 split.

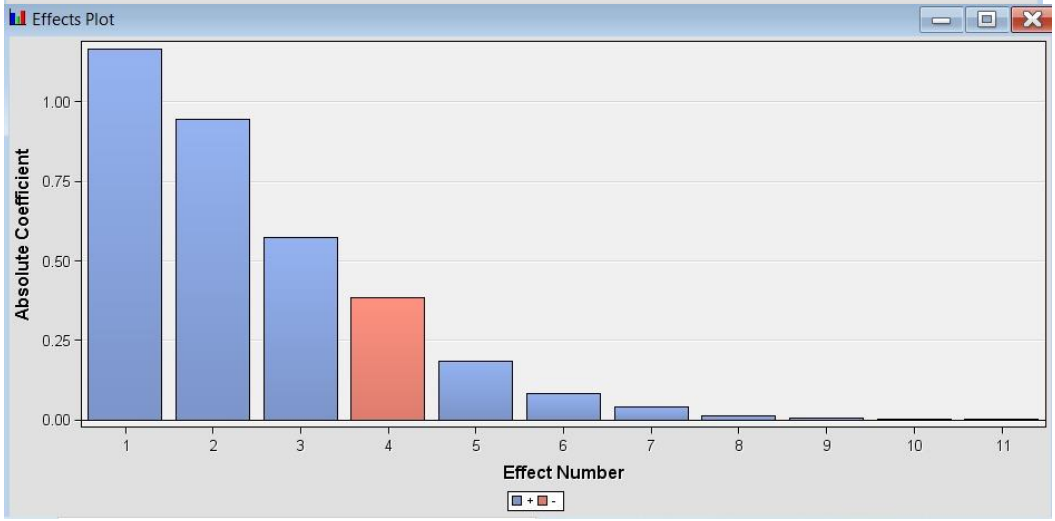
Then, I made sure that the variables were not skewed and followed a normal distribution using the Transform variables step.

Model 1 – Regression

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F
1	SQRT_comb08	1	1	488041	<.0001
2	SQRT_highway08	1	2	14752.6	<.0001
3	UHighway	1	3	25283.0	<.0001
4	PWR_city08	1	4	1707.89	<.0001
5	PWR_barrels08	1	5	1242.97	<.0001
6	IMP_cylinders	1	6	541.15	<.0001
7	PWR_youSaveSpend	1	7	683.90	<.0001
8	FuelNum	1	8	1518.76	<.0001
9	Transmission	1	9	352.28	<.0001
10	IMP_displ	1	10	243.91	<.0001
11	Vtype	1	11	144.80	<.0001
12	DriveNum	1	12	15.67	<.0001
13	PWR_fuelCost08	1	13	7.09	0.0078
14	MPGdNum	1	14	1.30	0.2544



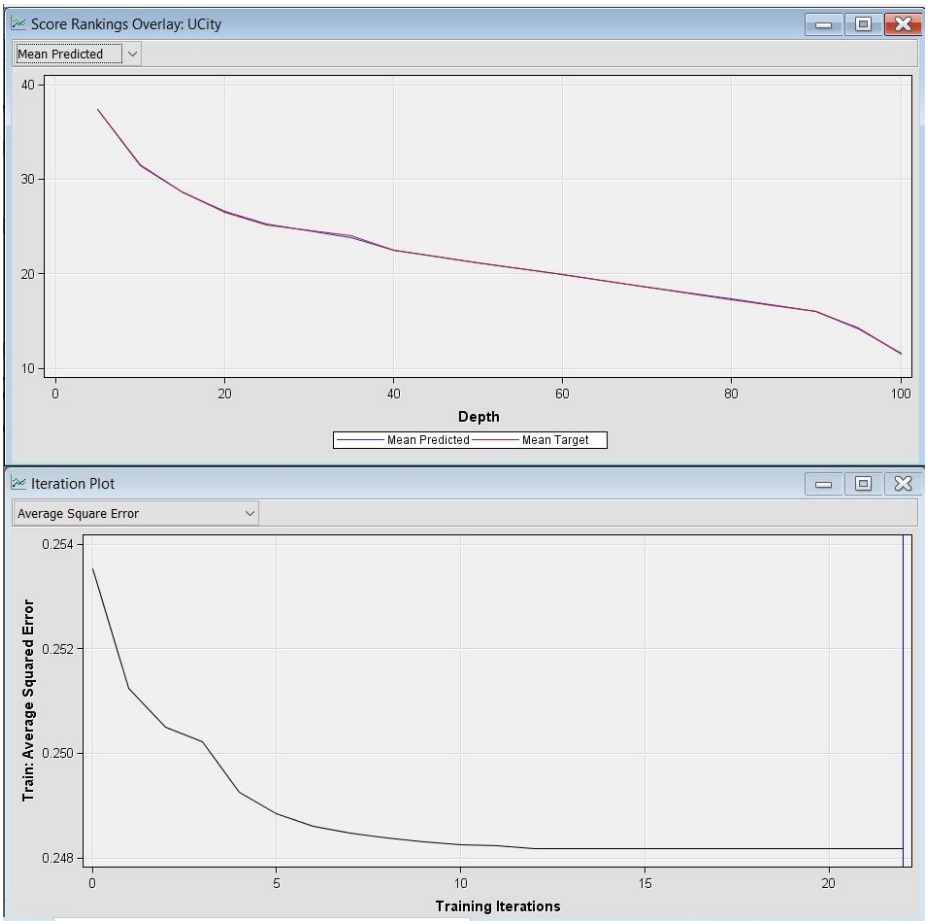


Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.1679	0.5400	2.16	0.0306
DriveNum	1	0.0132	0.00325	4.06	<.0001
FuelNum	1	0.0392	0.00458	8.57	<.0001
UHighway	1	0.1844	0.00318	57.97	<.0001
Vtype	1	0.00381	0.00116	3.28	0.0010
barrels08	1	0.0823	0.00231	35.56	<.0001
city08	1	0.9444	0.00499	189.42	<.0001
comb08	1	0.5740	0.00715	80.25	<.0001
fuelCost08	1	0.000061	0.000012	5.15	<.0001
highway08	1	-0.3852	0.00480	-80.21	<.0001
year	1	-0.00296	0.000267	-11.07	<.0001

Model 2 – Neural Network

Depth	Number of Observations	Mean Target	Mean Predicted
5	1590	38.2808	38.2949
10	1590	32.0425	31.9968
15	1603	29.2583	29.2348
20	1576	27.6218	27.6150
25	1593	26.3845	26.4171
30	1587	25.1348	25.1976
35	1662	24.1727	24.0970
40	1698	23.2842	23.2538
45	1432	22.5189	22.5381
50	1734	21.8408	21.8781
55	1459	21.0888	21.1850
60	1616	20.5476	20.5629
65	1588	19.9137	19.9093
70	1734	19.2823	19.2044
75	1385	18.6708	18.6489
80	1615	18.1297	18.1010
85	1574	17.1541	17.1966
90	1725	16.1919	16.1233
95	1452	14.6581	14.6634
100	1577	12.6623	12.6823



Neural Network gave the above model.

Model Comparison

	Neural Network	Regression
AIC	-51529.62	-4428282.6
RMSE	0.428	0.841

We can see that for Neural Network, the value for both AIC and RMSE is lower, meaning that Neural Network is a better performing model on our dataset.

I, Dwij Dua declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. I have not distributed my work to other students.