Sardar Patel Institute of Technology, Mumbai
Department of Electronics and Telecommunication Engineering
B.E. Sem-VII

Experiment Exploratory Data Analysis SAS

Name: Sruthi Shivaramakrishnan Batch:A UID:2019110059 Branch: ETRX

Objective: To perform Exploratory Data Analysis using SAS
Platform : SAS
Data Set : SAS help Cars Dataset
Code and output:
Importing the dataset and assigning to a new Cars:

```
data Cars;
set sashelp.cars;
run;
```

Printing the data:

```
proc print data=Cars;
```

Output:

Obs	Make	Model	Type	Origin	DriveTrain	MSRP	Invoice	EngineSize	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight	Wheelbase	Length
1	Acura	MDX	SUV	Asia	All	$36,945	$33,337	3.5	6	265	17	23	4451	106	189
2	Acura	RSX Type S 2dr	Sedan	Asia	Front	$23,820	$21,761	2.0	4	200	24	31	2778	101	172
3	Acura	TSX 4dr	Sedan	Asia	Front	$26,990	$24,647	2.4	4	200	22	29	3230	105	183
4	Acura	TL 4dr	Sedan	Asia	Front	$33,195	$30,299	3.2	6	270	20	28	3575	108	186
5	Acura	3.5 RL 4dr	Sedan	Asia	Front	$43,755	$39,014	3.5	6	225	18	24	3880	115	197
6	Acura	3.5 RL w/Navigation 4dr	Sedan	Asia	Front	$46,100	$41,100	3.5	6	225	18	24	3893	115	197
7	Acura	NSX coupe 2dr manual S	Sports	Asia	Rear	$89,765	$79,978	3.2	6	290	17	24	3153	100	174
8	Audi	A4 1.8T 4dr	Sedan	Europe	Front	$25,940	$23,508	1.8	4	170	22	31	3252	104	179
9	Audi	A41.8T convertible 2dr	Sedan	Europe	Front	$35,940	$32,506	1.8	4	170	23	30	3638	105	180
10	Audi	A4 3.0 4dr	Sedan	Europe	Front	$31,840	$28,846	3.0	6	220	20	28	3462	104	179
11	Audi	A4 3.0 Quattro 4dr manual	Sedan	Europe	All	$33,430	$30,366	3.0	6	220	17	26	3583	104	179
12	Audi	A4 3.0 Quattro 4dr auto	Sedan	Europe	All	$34,480	$31,388	3.0	6	220	18	25	3627	104	179
13	Audi	A6 3.0 4dr	Sedan	Europe	Front	$36,640	$33,129	3.0	6	220	20	27	3561	109	192
14	Audi	A6 3.0 Quattro 4dr	Sedan	Europe	All	$39,640	$35,992	3.0	6	220	18	25	3880	109	192
15	Audi	A4 3.0 convertible 2dr	Sedan	Europe	Front	$42,490	$38,325	3.0	6	220	20	27	3814	105	180
16	Audi	A4 3.0 Quattro convertible 2dr	Sedan	Europe	All	$44,240	$40,075	3.0	6	220	18	25	4013	105	180
17	Audi	A6 2.7 Turbo Quattro 4dr	Sedan	Europe	All	$42,840	$38,840	2.7	6	250	18	25	3836	109	192
18	Audi	A6 4.2 Quattro 4dr	Sedan	Europe	All	$49,690	$44,936	4.2	8	300	17	24	4024	109	193
19	Audi	A8 L Quattro 4dr	Sedan	Europe	All	$69,190	$64,740	4.2	8	330	17	24	4399	121	204
20	Audi	S4 Quattro 4dr	Sedan	Europe	All	$48,040	$43,556	4.2	8	340	14	20	3825	104	179
21	Audi	RS 6 4dr	Sports	Europe	Front	$84,600	$76,417	4.2	8	450	15	22	4024	109	191
22	Audi	TT 1.8 convertible 2dr (coupe)	Sports	Europe	Front	$35,940	$32,512	1.8	4	180	20	28	3131	95	159
23	Audi	TT 1.8 Quattro 2dr (convertible)	Sports	Europe	All	$37,390	$33,891	1.8	4	225	20	28	2921	96	159
24	Audi	TT 3.2 coupe 2dr (convertible)	Sports	Europe	All	$40,590	$36,739	3.2	6	250	21	29	3351	96	159

Statistical analysis of the data:

```
proc means data=Cars mean median mode std var min max;
```

Output:

The MEANS Procedure

Variable	Label	Mean	Median	Mode	Std Dev	Variance	Minimum	Maximum
MSRP		32774.86	27635.00	13270.00	19431.72	377591613	10280.00	192465.00
Invoice		30014.70	25294.50	14207.00	17642.12	311244319	9875.00	173560.00
EngineSize	Engine Size (L)	3.1967290	3.0000000	3.0000000	1.1085947	1.2289822	1.3000000	8.3000000
Cylinders		5.8075117	6.0000000	6.0000000	1.5584426	2.4287434	3.0000000	12.0000000
Horsepower		215.8855140	210.0000000	200.0000000	71.8360316	5160.42	73.0000000	500.0000000
MPG_City	MPG (City)	20.0607477	19.0000000	18.0000000	5.2382176	27.4389240	10.0000000	60.0000000
MPG_Highway	MPG (Highway)	26.8434579	26.0000000	26.0000000	5.7412007	32.9613857	12.0000000	66.0000000
Weight	Weight (LBS)	3577.95	3474.50	3175.00	758.9632146	576055.52	1850.00	7190.00
Wheelbase	Wheelbase (IN)	108.1542056	107.0000000	107.0000000	8.3118130	69.0862352	89.0000000	144.0000000
Length	Length (IN)	186.3621495	187.0000000	178.0000000	14.3579913	206.1519129	143.0000000	238.0000000

The variables Invoice, MSRP, Weight show high deviation from mean value indicating less clustering and more unique values.

Cleaning the data:
proc means data=Cars nmiss;

Output:

The MEANS Procedure

Variable	Label	N Miss
MSRP		0
Invoice		0
EngineSize	Engine Size (L)	0
Cylinders		2
Horsepower		0
MPG_City	MPG (City)	0
MPG_Highway	MPG (Highway)	0
Weight	Weight (LBS)	0
Wheelbase	Wheelbase (IN)	0
Length	Length (IN)	0

The variable cylinder has two null values.

Removing the null values:

```
data Cars_clean;
   SET Cars;
   IF cmiss(of _character_)
   OR nmiss(of _numeric_) > 0
    THEN
     DELETE;
run;
```

proc means data=Cars_clean nmiss;

Output:

The MEANS Procedure

Variable	Label	N Miss
MSRP		0
Invoice		0
EngineSize	Engine Size (L)	0
Cylinders		0
Horsepower		0
MPG_City	MPG (City)	0
MPG_Highway	MPG (Highway)	0
Weight	Weight (LBS)	0
Wheelbase	Wheelbase (IN)	0
Length	Length (IN)	0

The null values have been dropped and the dataset is cleaned with 0 null values.

Unique values:

```
proc sql;


select count(distinct 'EngineSize'n) as 'EngineSize'n,
      count(distinct Horsepower) as Horsepower,
      count(distinct MPG_City) as MPG_City,
      count(distinct 'Weight'n) as 'Weight'n,
      count(distinct 'Wheelbase'n) as 'Wheelbase'n
  from Cars_clean;
```

Output:

EngineSize	Horsepower	MPG_City	Weight	Wheelbase
42	109	28	346	40

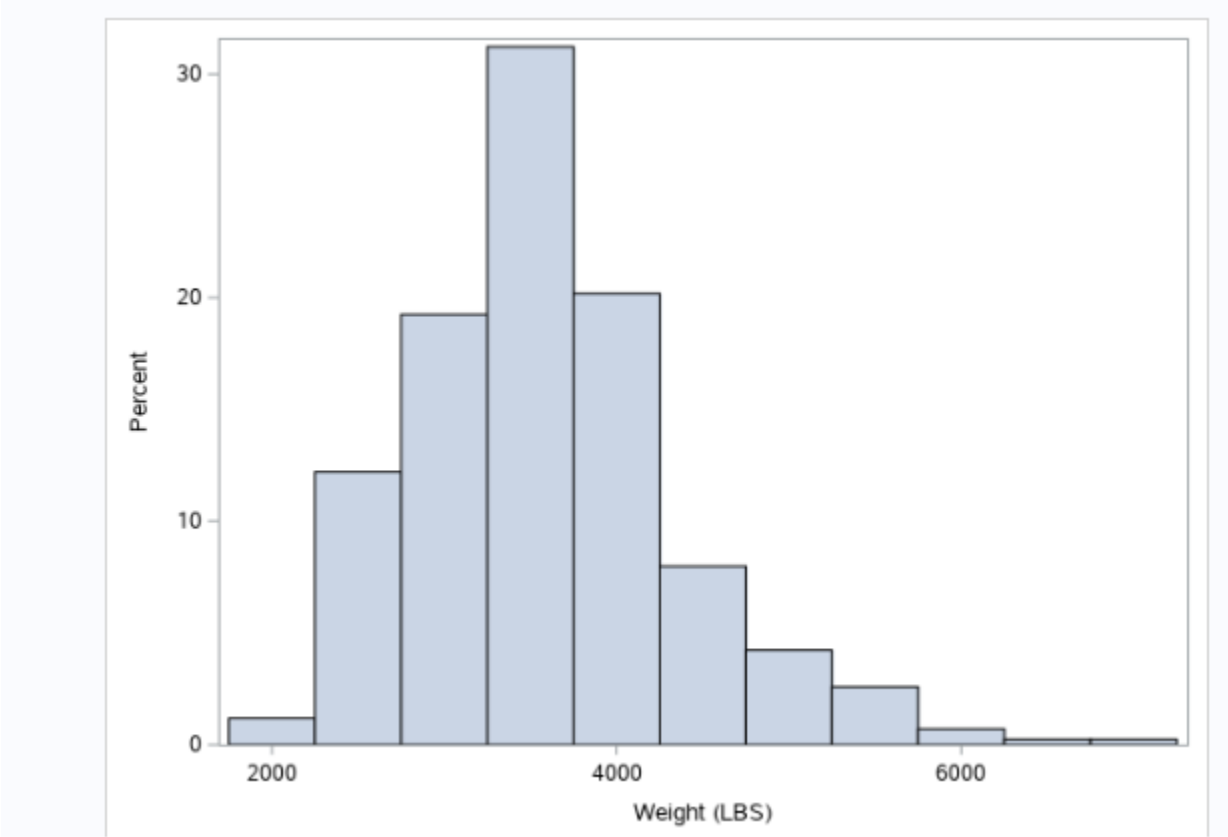
Weight shows high standard deviation because of more unique values.

Normalizing the data:

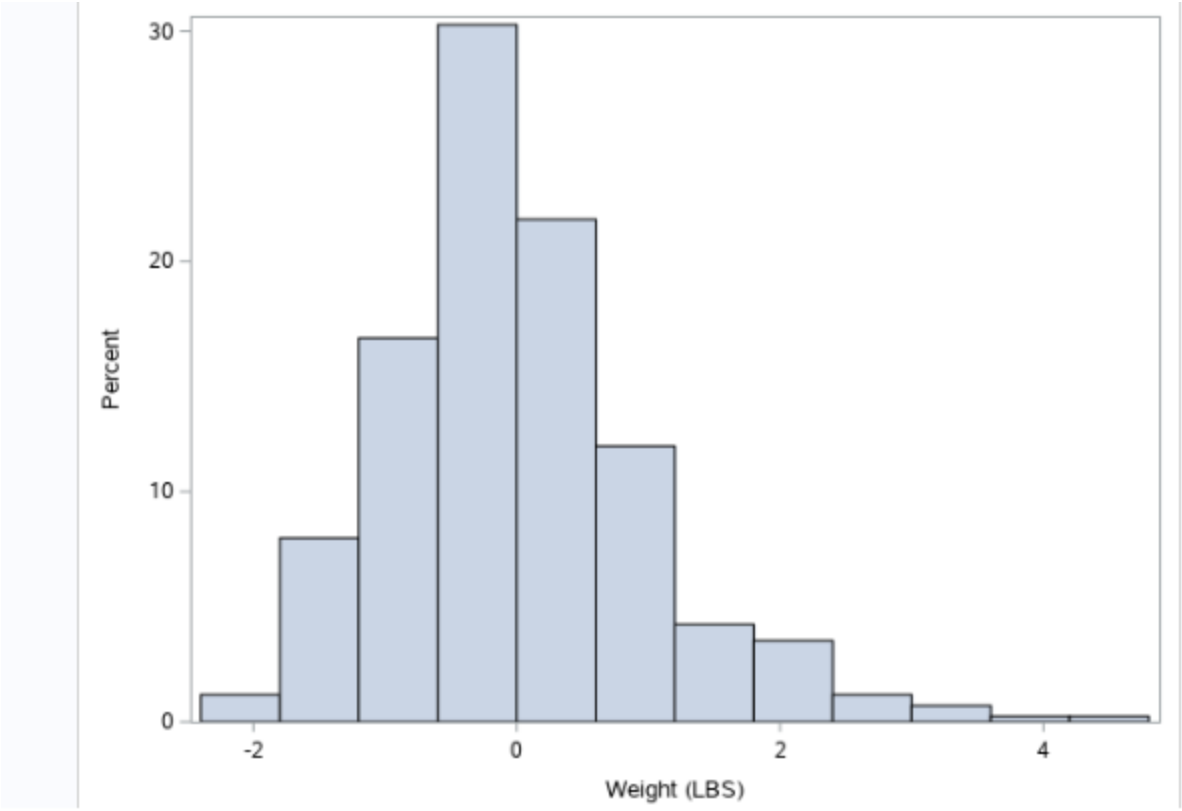
```
proc stdize data=Cars_clean out=normalized_data;
   var Weight;
run;
proc stdize data=Cars_clean out=normalized_data;
   var Invoice;
run;



ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=Cars_clean;
```

```
        histogram 'Weight'n /;

ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=normalized_data;
        histogram 'Weight'n /;
```

Output:
Before Normalisation:



After Normalisation:



\

The range of weights have been normalized.

Univariate analysis:

```
ods graphics on;
proc Univariate data=Cars_clean;
var Invoice;
run;
quit;
```

Variable: Invoice

Moments			
N	426	Sum Weights	426
Mean	30040.6549	Sum Observations	12797319
Std Deviation	17679.4301	Variance	312562249
Skewness	2.82587466	Kurtosis	13.8667565
Uncorrected SS	5.17279E11	Corrected SS	1.32839E11
Coeff Variation	58.85168	Std Error Mean	856.571188

Basic Statistical Measures			
Location		Variability	
Mean	30040.65	Std Deviation	17679
Median	25521.50	Variance	312562249
Mode	14207.00	Range	163685
		Interquartile Range	16956

Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	35.07082	Pr > |t|	<.0001
Sign	M	213	Pr >= |M|	<.0001
Signed Rank	S	45475.5	Pr >= |S|	<.0001

The above output shows the statistical parameters for the variable Invoice. Mean value is 30040 dollars with a huge variation in data indicated by standard deviation. The p value is greater than 0.001 indicating that we reject the null hypothesis suggesting mean value is not zero.

Quantiles (Definition 5)	
Level	Quantile
100% Max	173560.0
99%	88324.0
95%	66830.0
90%	48377.0
75% Q3	35777.0
50% Median	25521.5
25% Q1	18821.0
10%	14375.0
5%	12830.0
1%	10642.0
0% Min	9875.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
9875	207	88324	260
10107	169	113388	269
10144	381	117854	270
10319	344	119600	261
10642	383	173560	333

The above snippet shows the extreme values of the cost ranging from 9875 dollars to 173560 dollars for the cars.

Bivariate analysis:
proc freq data=Cars_clean;
table origin*type;
run;
Output:

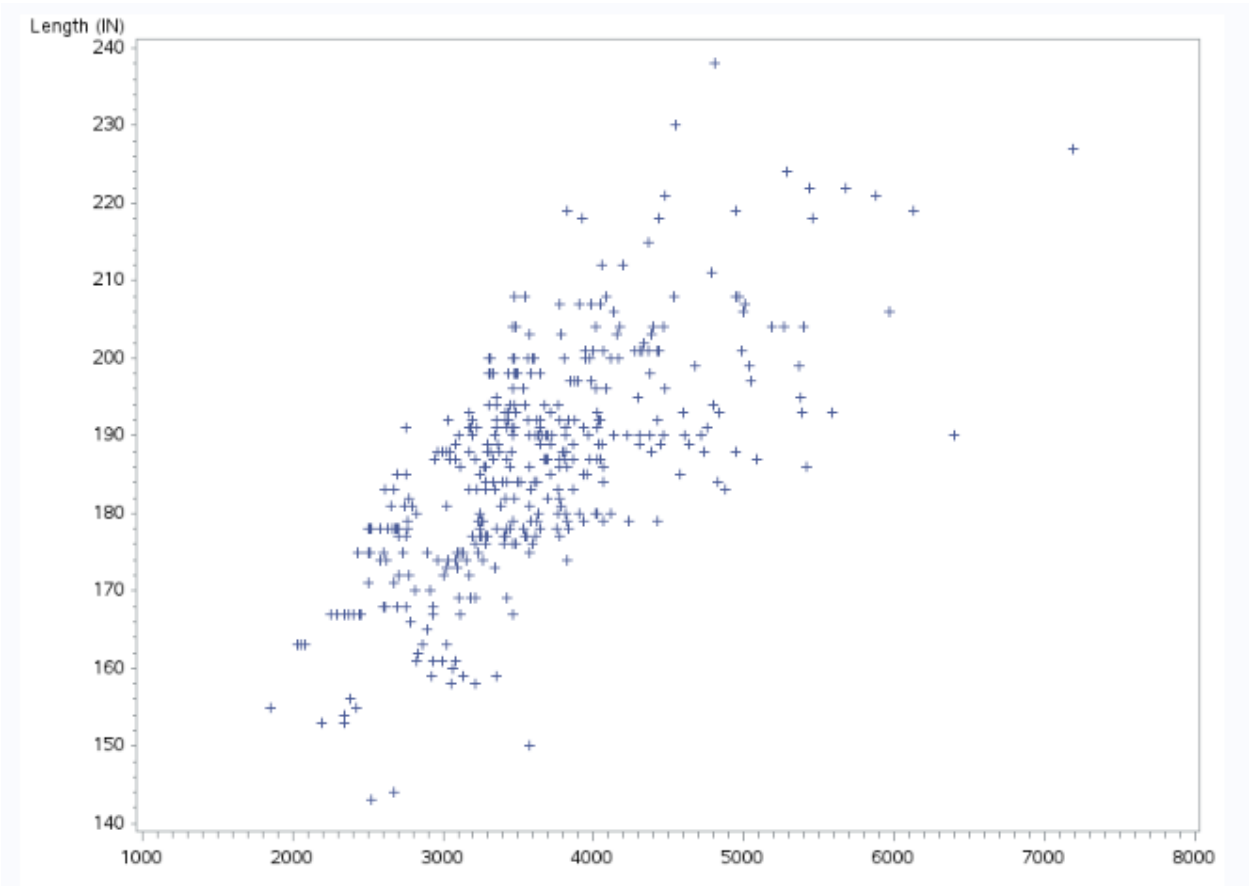
Frequency
Percent
Row Pct
Col Pct

Table of Origin by Type							
	Type						
Origin	Hybrid	SUV	Sedan	Sports	Truck	Wagon	Total
Asia	3 0.70 1.92 100.00	25 5.87 16.03 41.67	94 22.07 60.26 35.88	15 3.52 9.62 31.91	8 1.88 5.13 33.33	11 2.58 7.05 36.67	156 36.62
Europe	0 0.00 0.00 0.00	10 2.35 8.13 16.67	78 18.31 63.41 29.77	23 5.40 18.70 48.94	0 0.00 0.00 0.00	12 2.82 9.76 40.00	123 28.87
USA	0 0.00 0.00 0.00	25 5.87 17.01 41.67	90 21.13 61.22 34.35	9 2.11 6.12 19.15	16 3.76 10.88 66.67	7 1.64 4.76 23.33	147 34.51
Total	3 0.70	60 14.08	262 61.50	47 11.03	24 5.63	30 7.04	426 100.00

The Type of Cars from each region is shown in the table. Asia, Europe and the USA have the highest amount of Sedan cars manufacturing ,overall. Europe shows a high amount of Sports car manufacturing also.

```
ods graphics on;
proc gplot data=Cars_clean;
plot length*weight;
run;
quit;
```

Output:



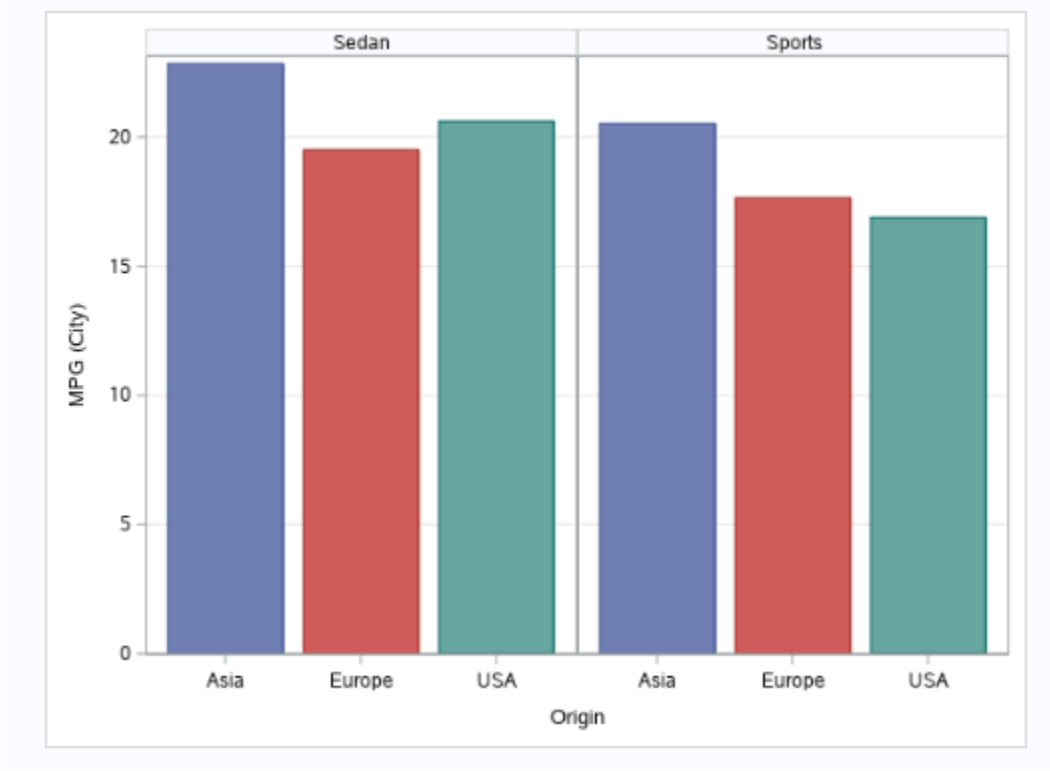
The above plot is a scatter plot between length and the weight of the car
Length and weight attribute plotting show a high positive correlation with each other.
Hence we can drop the variable

```
data Cars_clean;
  set Cars_clean(drop= length);

run;
```

Obs	Make	Model	Type	Origin	DriveTrain	MSRP	Invoice	Engine Size	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight	Wheelbase
1	Acura	MDX	SUV	Asia	All	$36,945	$33,337	3.5	6	265	17	23	4451	106
2	Acura	RSX Type S 2dr	Sedan	Asia	Front	$23,820	$21,761	2.0	4	200	24	31	2778	101
3	Acura	TSX 4dr	Sedan	Asia	Front	$26,990	$24,647	2.4	4	200	22	29	3230	105
4	Acura	TL 4dr	Sedan	Asia	Front	$33,195	$30,299	3.2	6	270	20	28	3575	108
5	Acura	3.5 RL 4dr	Sedan	Asia	Front	$43,755	$39,014	3.5	6	225	18	24	3880	115

The above snippet shows the database after dropping the variable.

```
proc sgpanel data=Cars_clean(where=(type in ('Sedan' 'Sports'))) noautolegend;
  panelby Type / novarname columns=2 onepanel;
  vbar origin / response=mpg_city stat=mean group=origin;
  rowaxis grid;
  run;
```

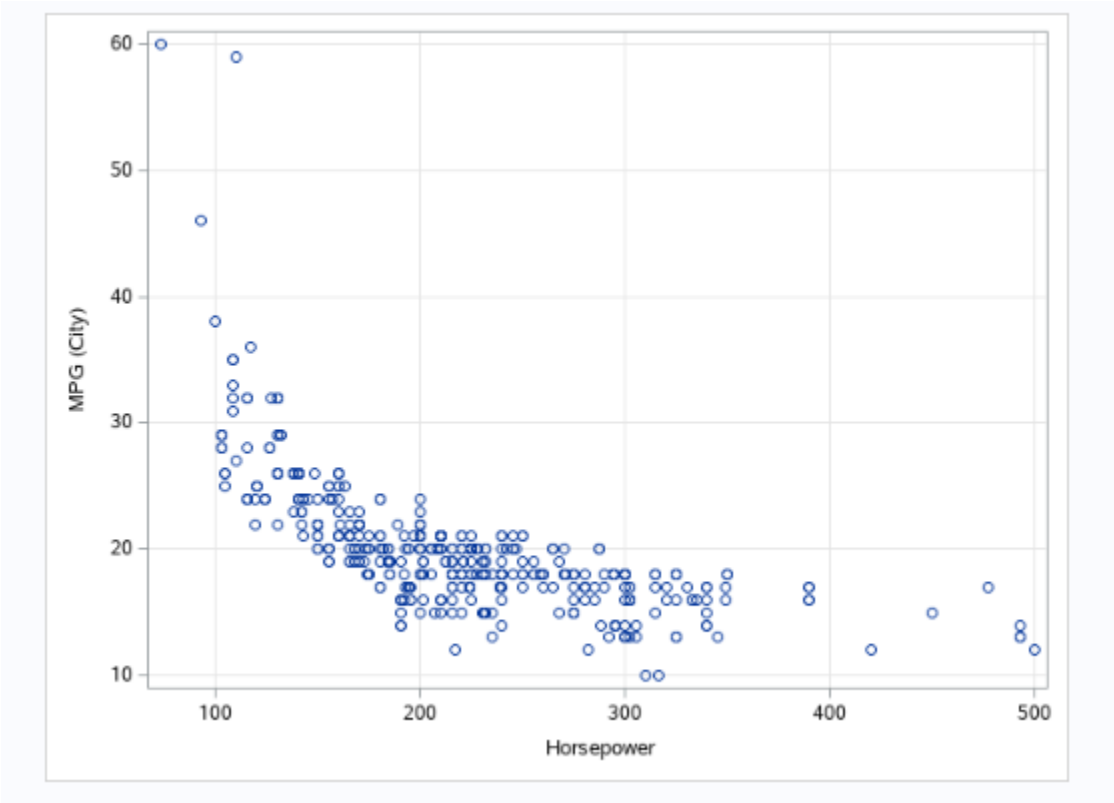


\

The above bar plot shows the MPG value for different cars from each place. Sedan cars and Sports cars have the highest fuel consumption in Asia followed by other countries like Europe and USA

```
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=Cars_clean;
        scatter x='Horsepower'n y='MPG_City'n /;
        xaxis grid;
        yaxis grid;
```

ods graphics / reset;
Output:



The above plots show a scatter plot between MPG_City and Horsepower. According to the plot MPG_City and Horsepower are negatively correlated to each other. The MPG_City plot shows a very high value for a few 100 Horsepower cars.

ods noproctitle;
ods graphics / imagemap=on;
proc corr data=Cars_clean pearson nosimple noprob plots=none;
var 'MSRP'n 'EngineSize'n 'Cylinders'n 'Horsepower'n 'MPG_City'n 'MPG_Highway'n 'Weight'n;
run;
Output:

7 Variables:	MSRP EngineSize Cylinders Horsepower MPG_City MPG_Highway Weight

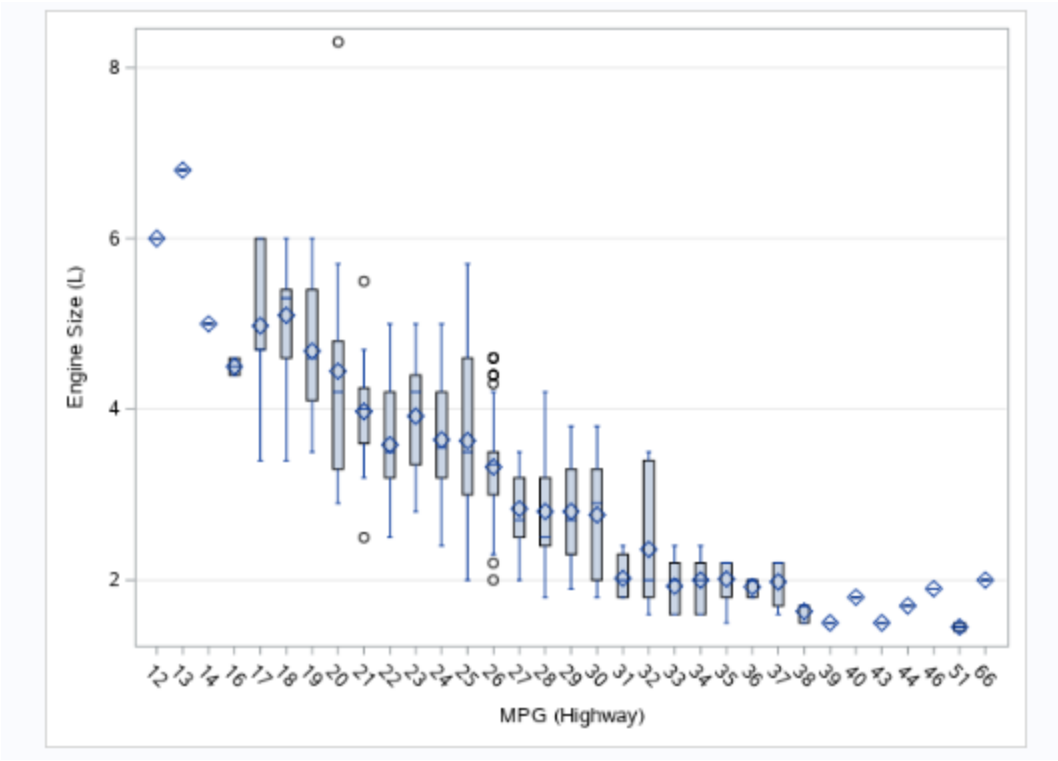
Pearson Correlation Coefficients, N = 426

	MSRP	Engine Size	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight
MSRP	1.00000	0.57324	0.64974	0.82730	-0.47592	-0.44052	0.44799
Engine Size Engine Size (L)	0.57324	1.00000	0.90800	0.79325	-0.71786	-0.72590	0.80871
Cylinders	0.64974	0.90800	1.00000	0.81034	-0.68440	-0.67610	0.74221
Horsepower	0.82730	0.79325	0.81034	1.00000	-0.67703	-0.64743	0.63176
MPG_City MPG (City)	-0.47592	-0.71786	-0.68440	-0.67703	1.00000	0.94099	-0.74042
MPG_Highway MPG (Highway)	-0.44052	-0.72590	-0.67610	-0.64743	0.94099	1.00000	-0.79362
Weight Weight (LBS)	0.44799	0.80871	0.74221	0.63176	-0.74042	-0.79362	1.00000

The above output shows correlation between the different variables.Engine Size and cylinders show a strong positive correlation. All the variables show a moderately strong correlation with Engine size. Increasing Engine size can increase the MSRP, Cylinders, Horsepower,Weight but can decrease the Miles per gallon.

```
ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=Cars_clean;
        vbox 'EngineSize'n  / category='MPG_Highway'n;
        yaxis grid;
run;
ods graphics / reset;
```

Output:



The above plot shows the trend in Engine size as MPG increases. For high miles per gallon engine size can be 2 or 3. Whereas for low miles per gallon engine size is 5 to 6.

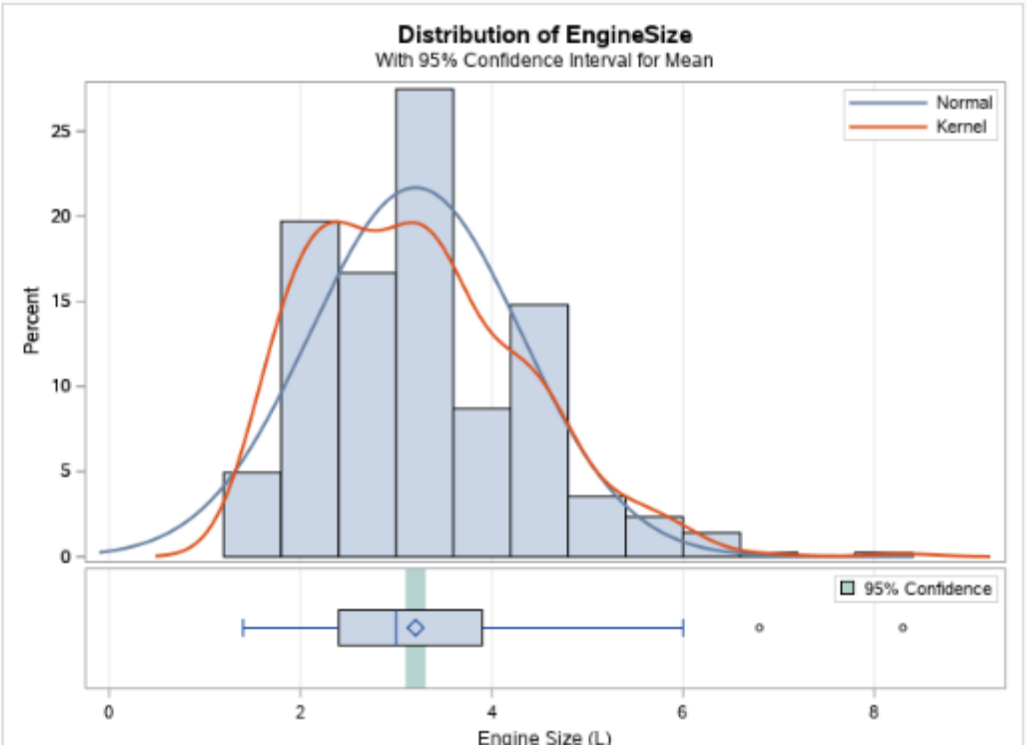
```
proc ttest data=Cars_clean ALPHA=0.05 H0=3;
   var EngineSize;
run;
```

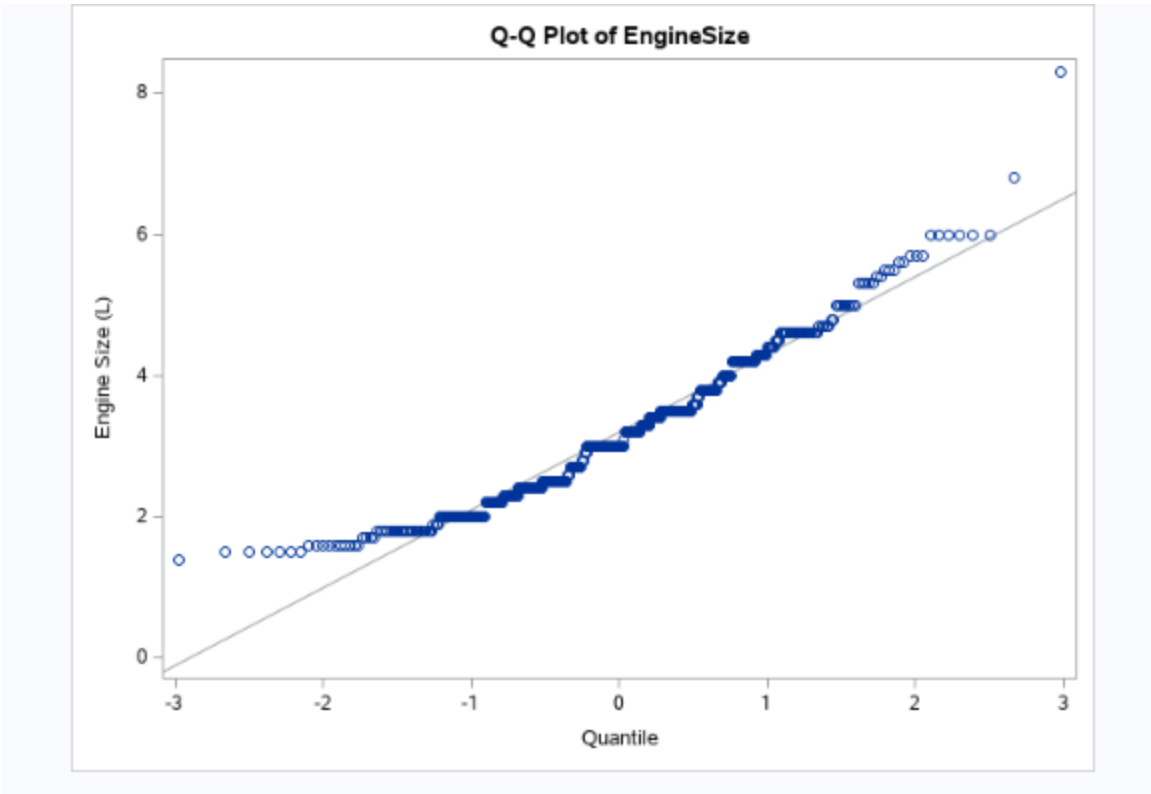
Output:

N	Mean	Std Dev	Std Err	Minimum	Maximum
426	3.2056	1.1035	0.0535	1.4000	8.3000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3.2056	3.1005	3.3107	1.1035	1.0341	1.1831

DF	t Value	Pr > |t|
425	3.85	0.0001





The above plot shows the t test results for mean of Engine Size. The p value is greater than 0.005. Hence we can reject the null hypothesis stating that the mean value of Engine is greater than 3 for the data.

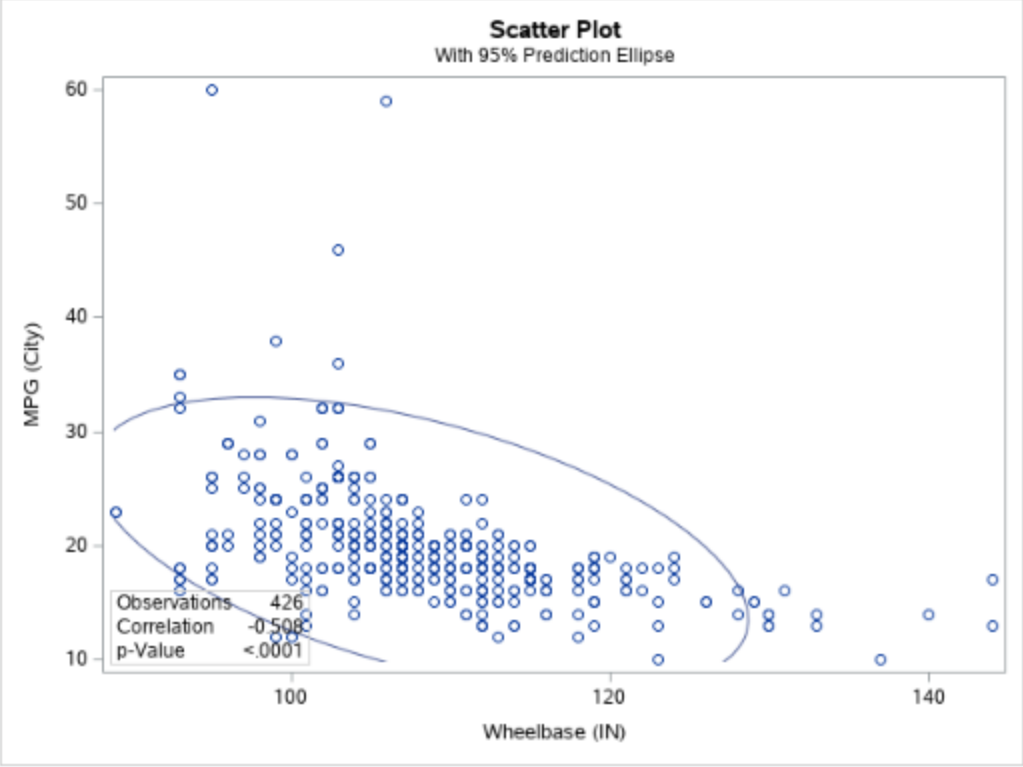
```
PROC CORR DATA= Cars_clean PLOTS=SCATTER(NVAR=all);
   VAR Wheelbase;
   WITH MPG_City;

RUN;
```

```
proc corr data= Cars_clean nosimple spearman;
var  Horsepower;
with MPG_Highway;
run;
```

Output:

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
MPG_City	426	20.07042	5.24862	8550	10.00000	60.00000	MPG (City)
Wheelbase	426	108.16432	8.33003	46078	89.00000	144.00000	Wheelbase (IN)

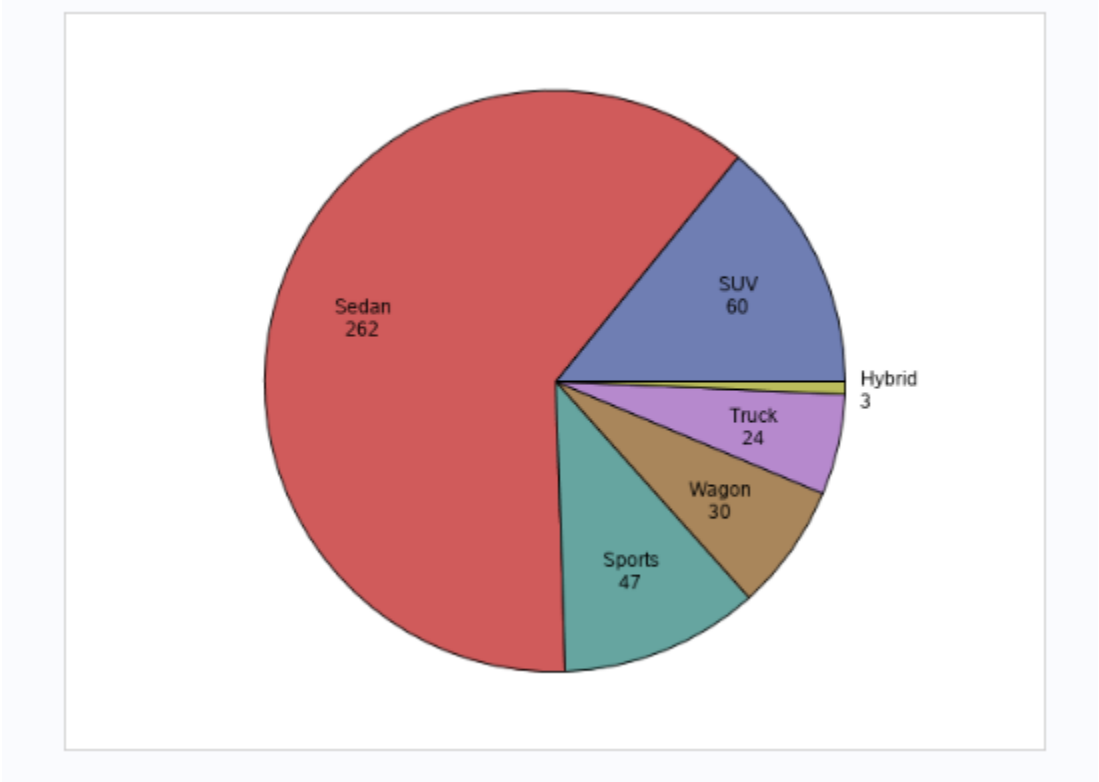
Pearson Correlation Coefficients, N = 426 Prob > |r| under H0: Rho=0	
	Wheelbase
MPG_City MPG (City)	-0.50803 <.0001



The above plot shows a scatter plot for Wheelbase and MPG_City.The plot shows a negative correlation of Wheelbase with MPG_City. The p value is less than 0.001 hence the variable MPG_City is significant.

```
proc template;
        define statgraph SASStudio.Pie;
                begingraph;
                layout region;
                piechart category=Type /;
                endlayout;
                endgraph;
        end;
run;
```

```
ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgrender template=SASStudio.Pie data=WORK.CARS_CLEAN;
run;

ods graphics / reset;
```
Output:

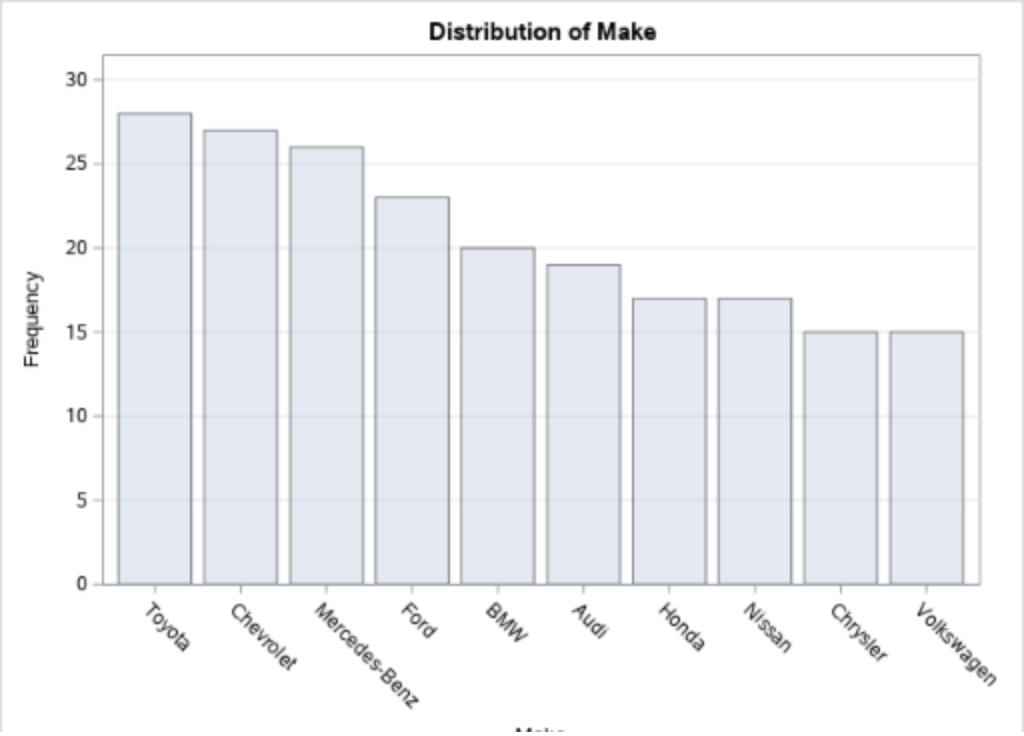


The above plot shows the market share of each type of car. Sedan cars have the highest share followed by SUV and Sports cars.

```
%let TopN = 10;
proc freq data=Cars_clean ORDER=FREQ;
  tables make / maxlevels=&TopN Plots=FreqPlot;
run;
```

Output:

Make	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Toyota	28	6.57	28	6.57
Chevrolet	27	6.34	55	12.91
Mercedes-Benz	26	6.10	81	19.01
Ford	23	5.40	104	24.41
BMW	20	4.69	124	29.11
Audi	19	4.46	143	33.57
Honda	17	3.99	160	37.56
Nissan	17	3.99	177	41.55
Chrysler	15	3.52	192	45.07
Volkswagen	15	3.52	207	48.59
The first 10 levels are displayed.				



The above chart shows the Top 10 frequency of make for each car. The highest make is for Toyota followed by Chevrolet and Merceded Benz.

Inference:

1. The analysis shows Engine and miles per gallon to be one of the most important parameters for a vehicle efficiency.
2. The mean engine size for vehicles is 3.5
3. Although for a few vehicles Engine size increases MPG, there are some vehicles which show a lower MPG with higher Engine Size.This can be because of other factors influencing the efficiency of the car.
4. The Efficiency of a car can be influenced by Engine Size to a large extent but other factors also influence the efficiency as the correlation plot shows strong correlation of Engine with other variables.

5. Price increases with Engine Size in cars.
6. The highest type of cars used is Sedan manufactured in Asia, Europe and USA.
7. The highest make is for Toyota cars in the respective countries.
8. This can be due to lower engine, size higher MPG and lower price for the cars.