Sardar Patel Institute of Technology,Mumbai
Department of Electronics and Telecommunication Engineering
B.E. Sem-VII (2021-2022)
Data Analytics

Experiment: Exploratory Data Analysis (EDA)

Name: Sneha Ghuge **UID:** 2019110015 **BE ETRX** **DA LAB 2**

Aim: Using the SAS software to analyze statistical data

Problem Statement : Study and understand the workings of SAS studio by referring to the online materials and documentation, etc., and then implement a small problem.

CODE & OUTPUT:

For this experiment I used the already present database in the SAS studio (My Libraries -> SASHELP -> SASHELP.CARS). I just made some modifications in the dataset and only kept the required column for the experiment.
Here I have directly used the SQL procedures to process the SQL statements.

```
Proc Sql;
create table cars as
SELECT Make, MSRP, Horsepower, MPG_City, MPG_Highway
    From SASHELP.CARS
;
RUN;
```

The output table looks like this.

	Make	MSRP	Horsepower	MPG_City	MPG_Highway
1	Acura	$36,945	265	17	23
2	Acura	$23,820	200	24	31
3	Acura	$26,990	200	22	29
4	Acura	$33,195	270	20	28
5	Acura	$43,755	225	18	24
6	Acura	$46,100	225	18	24
7	Acura	$89,765	290	17	24

The output of the **proc print** is shown below.

```
PROC PRINT DATA=cars(obs=10);
RUN;
```

Obs	Make	MSRP	Horsepower	MPG_City	MPG_Highway
1	Acura	$36,945	265	17	23
2	Acura	$23,820	200	24	31
3	Acura	$26,990	200	22	29
4	Acura	$33,195	270	20	28
5	Acura	$43,755	225	18	24
6	Acura	$46,100	225	18	24
7	Acura	$89,765	290	17	24
8	Audi	$25,940	170	22	31
9	Audi	$35,940	170	23	30
10	Audi	$31,840	220	20	28

We can use **proc freq** to produce frequency tables.

Below, we use it to make frequency tables for Make, Horsepower, MPG_City and MPG_Highway.

```
15 PROC FREQ DATA=cars;
16   TABLES make ;
17 RUN;
18
19 PROC FREQ DATA=cars;
20   TABLES Horsepower ;
21 RUN;
22
23 PROC FREQ DATA=cars;
24   TABLES MPG_City ;
25 RUN;
26
27 PROC FREQ DATA=cars;
28   TABLES MPG_Highway ;
29 RUN;
```

Make	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Acura	7	1.64	7	1.64
Audi	19	4.44	26	6.07
BMW	20	4.67	46	10.75
Buick	9	2.10	55	12.85
Cadillac	8	1.87	63	14.72
Chevrolet	27	6.31	90	21.03
Chrysler	15	3.50	105	24.53
Dodge	13	3.04	118	27.57
Ford	23	5.37	141	32.94

Horsepower	Frequency	Percent	Cumulative Frequency	Cumulative Percent
73	1	0.23	1	0.23
93	1	0.23	2	0.47
100	1	0.23	3	0.70
103	5	1.17	8	1.87
104	3	0.70	11	2.57
108	5	1.17	16	3.74
110	2	0.47	18	4.21
115	6	1.40	24	5.61
117	1	0.23	25	5.84
119	2	0.47	27	6.31
120	3	0.70	30	7.01

MPG (City)				
MPG_City	Frequency	Percent	Cumulative Frequency	Cumulative Percent
10	2	0.47	2	0.47
12	4	0.93	6	1.40
13	12	2.80	18	4.21
14	13	3.04	31	7.24
15	17	3.97	48	11.21
16	31	7.24	79	18.46
17	41	9.58	120	28.04
18	69	16.12	189	44.16
19	37	8.64	226	52.80
20	57	13.32	283	66.12

MPG (Highway)				
MPG_Highway	Frequency	Percent	Cumulative Frequency	Cumulative Percent
12	1	0.23	1	0.23
13	1	0.23	2	0.47
14	1	0.23	3	0.70
16	2	0.47	5	1.17
17	9	2.10	14	3.27
18	11	2.57	25	5.84
19	16	3.74	41	9.58
20	13	3.04	54	12.62

Proc means can be used to produce summary statistics.
Below, **proc means** is used to get descriptive statistics for the variable MSRP, Horsepower, **MPG_City, MPG_Highway**

```
34 PROC MEANS DATA=cars;
35   VAR MSRP;
36 RUN;
37
38
39 PROC MEANS DATA=cars;
40   VAR Horsepower;
41 RUN;
42
43
44 PROC MEANS DATA=cars;
45   VAR MSRP;
46 RUN;
47
48
49 PROC MEANS DATA=cars;
50   VAR MPG_City;
51 RUN;
52
53 PROC MEANS DATA=cars;
54   VAR MPG_Highway;
55 RUN;
```

Analysis Variable : MSRP				
N	Mean	Std Dev	Minimum	Maximum
428	32774.86	19431.72	10280.00	192465.00

Analysis Variable : Horsepower				
N	Mean	Std Dev	Minimum	Maximum
428	215.8855140	71.8360316	73.0000000	500.0000000

Analysis Variable : MPG_City MPG (City)				
N	Mean	Std Dev	Minimum	Maximum
428	20.0607477	5.2382176	10.0000000	60.0000000

Analysis Variable : MPG_Highway MPG (Highway)				
N	Mean	Std Dev	Minimum	Maximum
428	26.8434579	5.7412007	12.0000000	66.0000000

Using proc univariate for detailed summary statistics
You can use proc univariate to get more detailed summary statistics, as shown below.

```
54 PROC UNIVARIATE DATA=cars;
55   VAR MSRP;
56 RUN;
57
58 PROC UNIVARIATE DATA=cars;
59   VAR MPG_City;
60 RUN;
```

The UNIVARIATE Procedure
Variable: MSRP

Moments			
N	428	Sum Weights	428
Mean	32774.8551	Sum Observations	14027638
Std Deviation	19431.7167	Variance	377591613
Skewness	2.79809927	Kurtosis	13.8792055
Uncorrected SS	6.20985E11	Corrected SS	1.61232E11
Coeff Variation	59.2884899	Std Error Mean	939.267478

Basic Statistical Measures			
Location		Variability	
Mean	32774.86	Std Deviation	19432
Median	27635.00	Variance	377591613
Mode	13270.00	Range	182185
		Interquartile Range	18886

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	34.89406	Pr > |t|	<.0001
Sign	M	214	Pr >= |M|	<.0001
Signed Rank	S	45903	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	192465.0
99%	94820.0
95%	73195.0
90%	52795.0
75% Q3	39215.0
50% Median	27635.0
25% Q1	20329.5
10%	15460.0
5%	13670.0
1%	11155.0
0% Min	10280.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
10280	207	94820	262
10539	169	121770	271
10760	383	126670	272
10995	346	128420	263
11155	208	192465	335

The UNIVARIATE Procedure
Variable: MPG_City (MPG (City))

Moments			
N	428	Sum Weights	428
Mean	20.0607477	Sum Observations	8586
Std Deviation	5.23821764	Variance	27.438924
Skewness	2.7820718	Kurtosis	15.7911473
Uncorrected SS	183958	Corrected SS	11716.4206
Coeff Variation	26.1117767	Std Error Mean	0.25319881

Basic Statistical Measures			
Location		Variability	
Mean	20.06075	Std Deviation	5.23822
Median	19.00000	Variance	27.43892
Mode	18.00000	Range	50.00000
		Interquartile Range	4.50000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	79.22923	Pr > |t|	<.0001
Sign	M	214	Pr >= |M|	<.0001
Signed Rank	S	45903	Pr >= |S|	<.0001

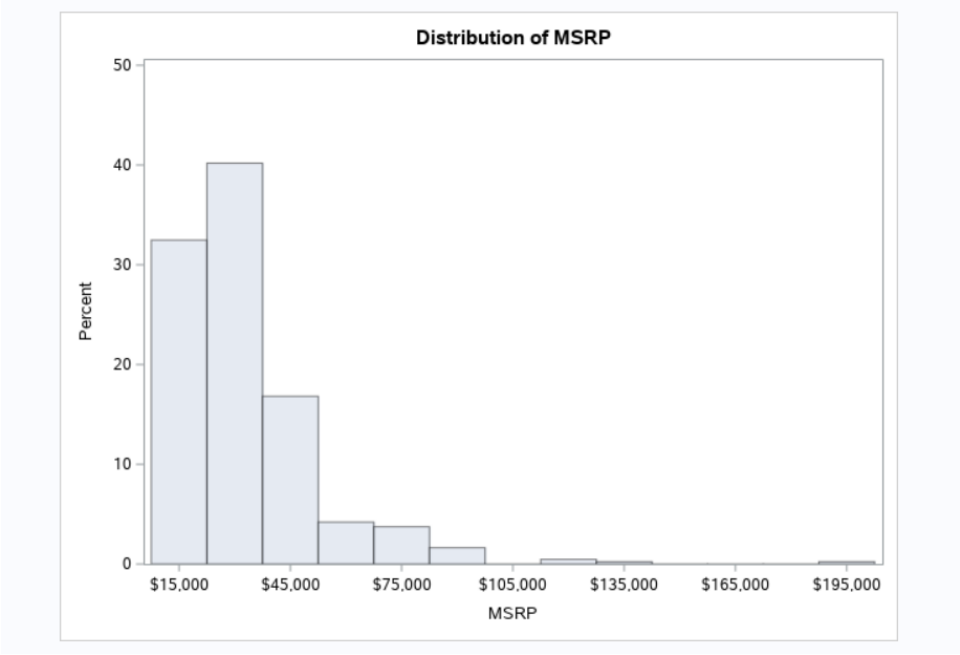
Quantiles (Definition 5)	
Level	Quantile
100% Max	60.0
99%	36.0
95%	29.0
90%	26.0
75% Q3	21.5
50% Median	19.0
25% Q1	17.0
10%	15.0
5%	14.0
1%	12.0
0% Min	10.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
10	167	36	156
10	119	38	405
12	413	46	150
12	217	59	374
12	216	60	151

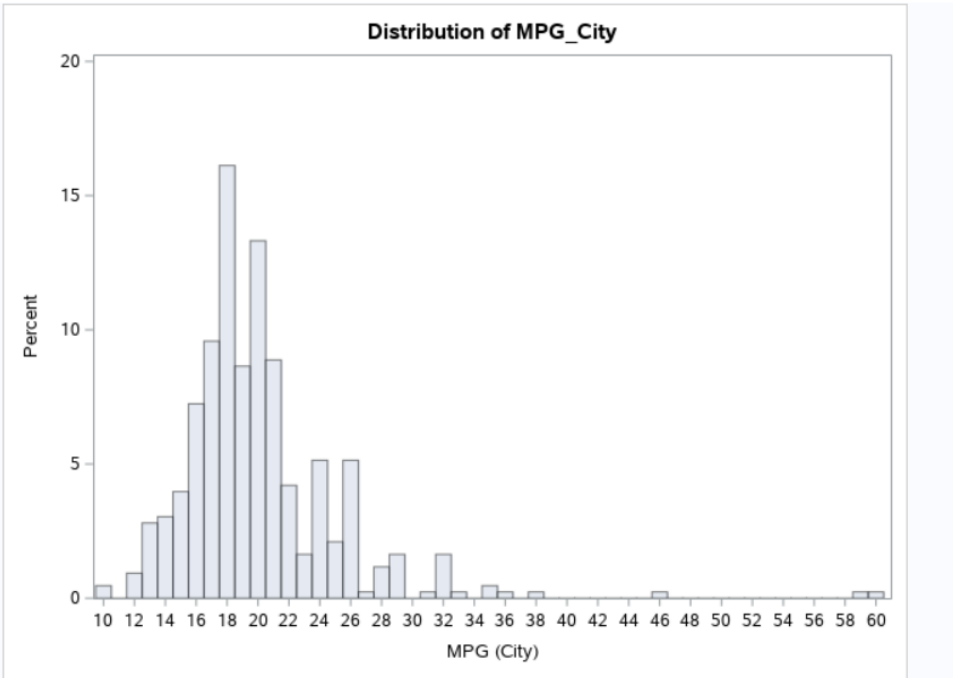
Now lets do Data representation to view and analyze the data in a better manner. The first kind was plotting histograms which is a graphical display of data using bars of different heights. It groups the various numbers in the data set into many ranges. It also represents the estimation of the probability of distribution of a continuous variable.

In SAS the PROC UNIVARIATE is used to create histograms.

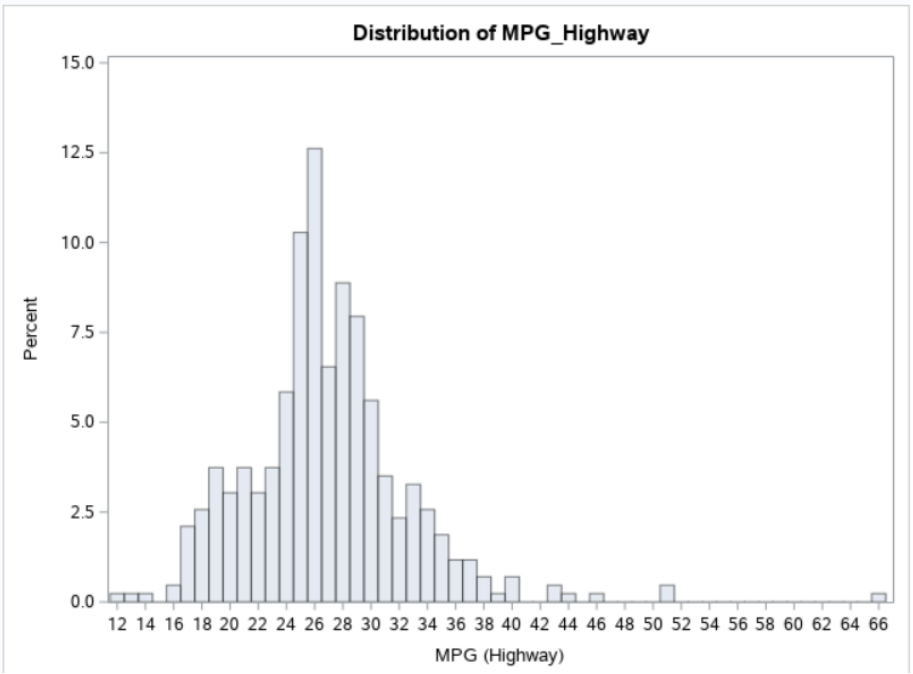
```
Proc univariate data=cars;
    histogram MSRP
    / midpoints= 25000 to 50000;
run;
```




```
Proc univariate data=cars;
    histogram MPG_City
    / midpoints= 15 to 20;
run;
```




```
Proc univariate data=cars;
    histogram MPG_Highway
    / midpoints= 25 to 50;
run;
```



CONCLUSION :

1. SAS provides us various functionalities like Data Management, Statistical Analysis, Report formation with perfect graphics etc that I have used in the above experiment.

2. In the above experiment I loaded the already existing dataset by just making some changes in the original dataset to create my own data and then after cleaning the values, filling missing values I performed various kinds of data analysis by plotting the various utility functions for plotting that SAS provides.