Name	Jayesh Vikas Bane
UID	2019120006
Class	BE EXTC
Batch	A

Aim: Exploratory Data Analysis in SAS.

Dataset Description:
I have used the built-in cars dataset from the SASHELP library. It contains the following columns:

- Make
- Model
- MSRP
- Invoice
- Engine Size
- Cylinders
- Horsepower
- MPG_City
- MPG_Highway
- Weight
- Wheelbase
- Length

EDA:

Summary Statistics for each numeric column

Variable	Label	Mean	Std Dev	Minimum	Maximum	N
MSRP		32774.86	19431.72	10280.00	192465.00	428
Invoice		30014.70	17642.12	9875.00	173560.00	428
Horsepower		215.8855140	71.8360316	73.0000000	500.0000000	428
Weight	Weight (LBS)	3577.95	758.9832146	1850.00	7190.00	428
EngineSize	Engine Size (L)	3.1967290	1.1085947	1.3000000	8.3000000	428
Cylinders		5.8075117	1.5584426	3.0000000	12.0000000	426
MPG_City	MPG (City)	20.0607477	5.2382176	10.0000000	60.0000000	428
MPG_Highway	MPG (Highway)	26.8434579	5.7412007	12.0000000	66.0000000	428
Wheelbase	Wheelbase (IN)	108.1542056	8.3118130	89.0000000	144.0000000	428
Length	Length (IN)	186.3621495	14.3579913	143.0000000	238.0000000	428

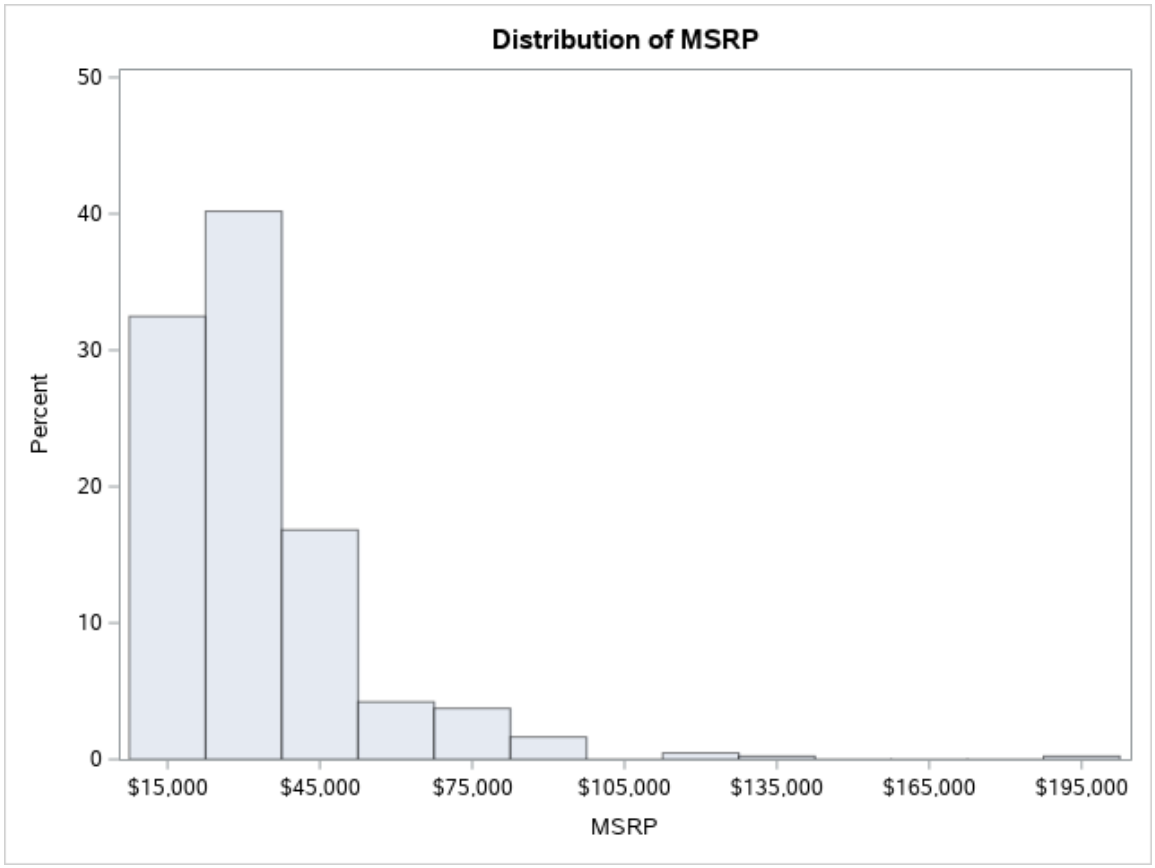
Correlation Analysis for every variable:

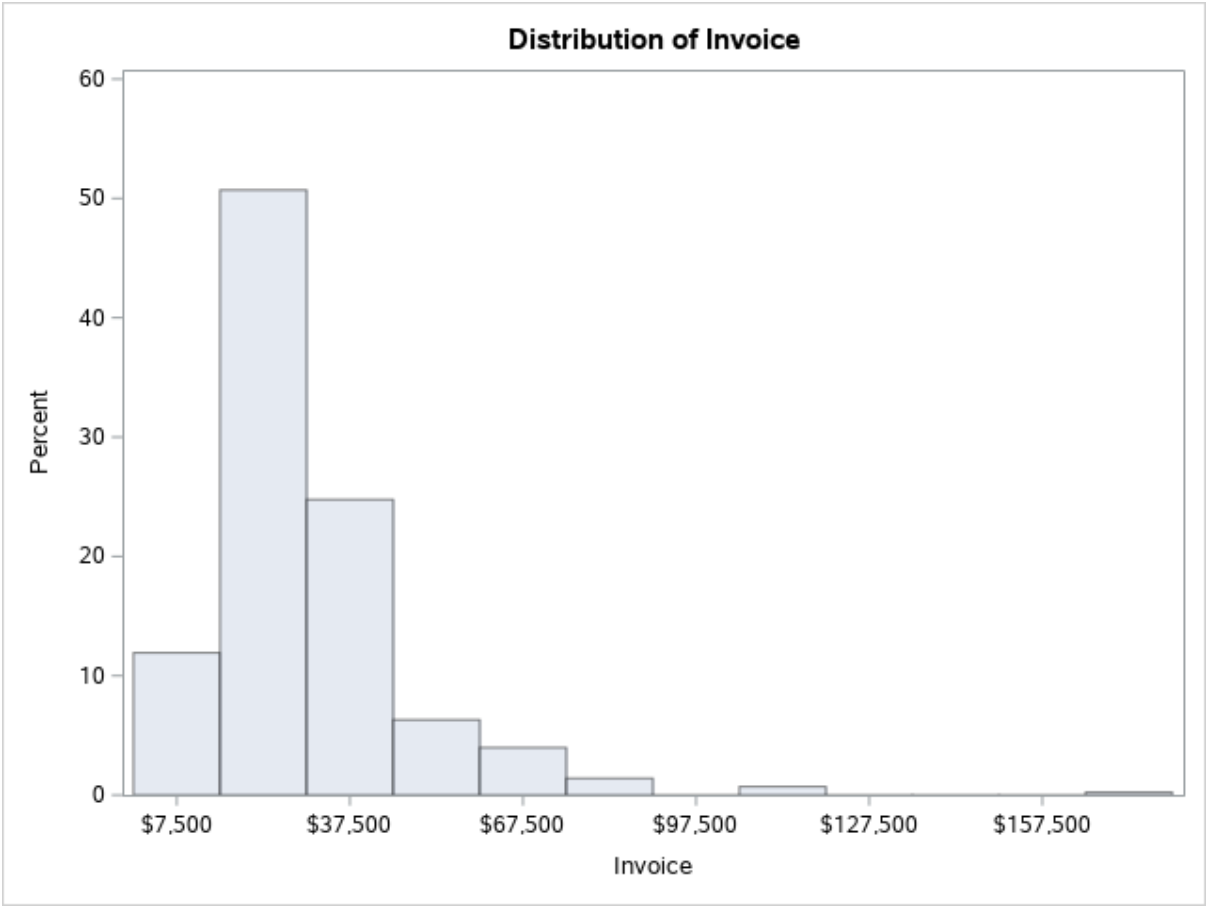
10 Variables:	MSRP Invoice EngineSize Cylinders Horsepower MPG_City MPG_Highway Weight Wheelbase Length

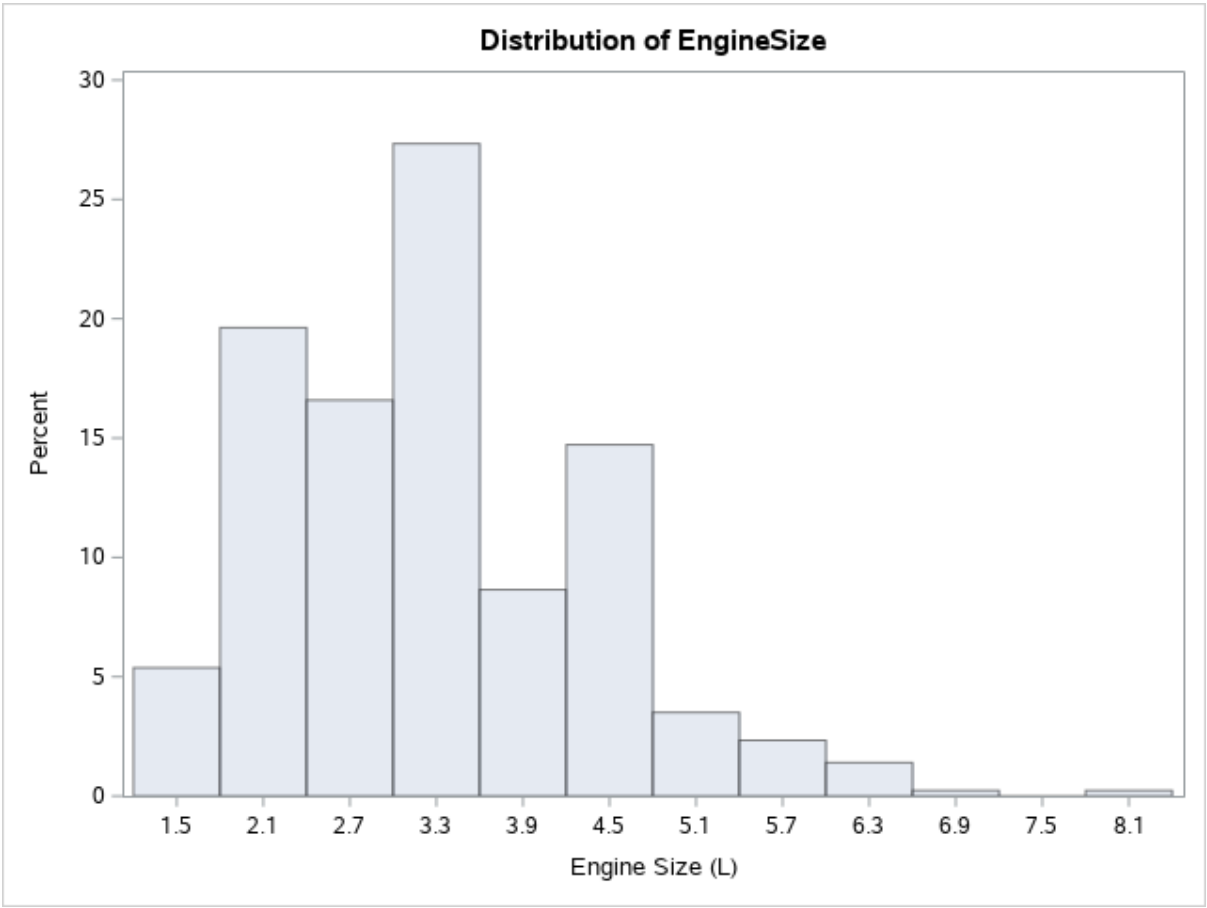
Pearson Correlation Coefficients
Number of Observations

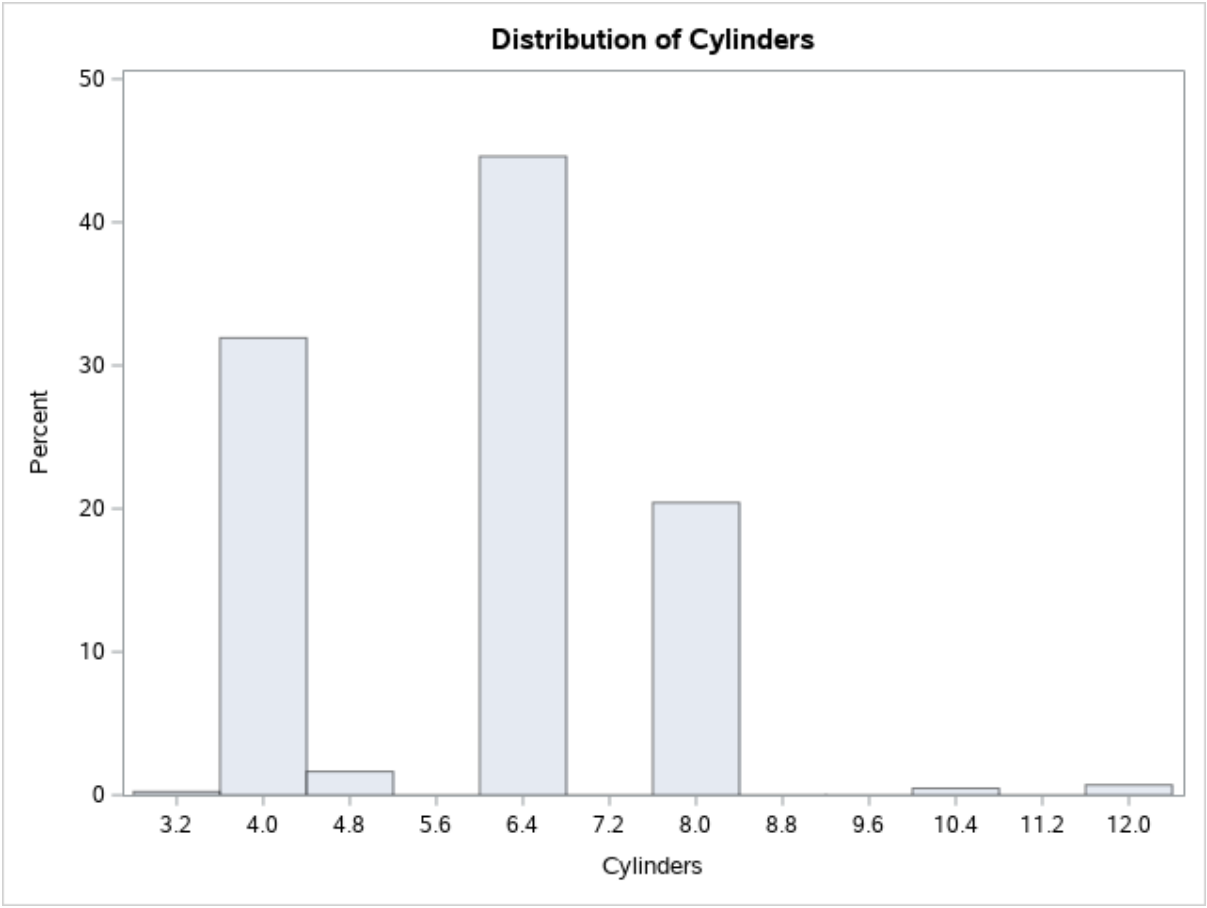
	MSRP	Invoice	EngineSize	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight	Wheelbase	Length
MSRP	1.00000 428	0.99913 428	0.57175 428	0.64974 426	0.82695 428	-0.47502 428	-0.43962 428	0.44843 428	0.15200 428	0.17204 428
Invoice	0.99913 428	1.00000 428	0.56450 428	0.64523 426	0.82375 428	-0.47044 428	-0.43459 428	0.44233 428	0.14833 428	0.16659 428
EngineSize Engine Size (L)	0.57175 428	0.56450 428	1.00000 428	0.90800 426	0.78743 428	-0.70947 428	-0.71730 428	0.80787 428	0.63652 428	0.63745 428
Cylinders	0.64974 426	0.64523 426	0.90800 426	1.00000 426	0.81034 426	-0.68440 426	-0.67610 426	0.74221 426	0.54673 426	0.54778 426
Horsepower	0.82695 428	0.82375 428	0.78743 428	0.81034 426	1.00000 428	-0.67670 428	-0.64720 428	0.63080 428	0.38740 428	0.38155 428
MPG_City MPG (City)	-0.47502 428	-0.47044 428	-0.70947 428	-0.68440 426	-0.67670 428	1.00000 428	0.94102 428	-0.73797 428	-0.50728 428	-0.50153 428
MPG_Highway MPG (Highway)	-0.43962 428	-0.43459 428	-0.71730 428	-0.67610 426	-0.64720 428	0.94102 428	1.00000 428	-0.79099 428	-0.52466 428	-0.46609 428
Weight Weight (LBS)	0.44843 428	0.44233 428	0.80787 428	0.74221 426	0.63080 428	-0.73797 428	-0.79099 428	1.00000 428	0.76070 428	0.69002 428
Wheelbase Wheelbase (IN)	0.15200 428	0.14833 428	0.63652 428	0.54673 426	0.38740 428	-0.50728 428	-0.52466 428	0.76070 428	1.00000 428	0.88919 428
Length Length (IN)	0.17204 428	0.16659 428	0.63745 428	0.54778 426	0.38155 428	-0.50153 428	-0.46609 428	0.69002 428	0.88919 428	1.00000 428

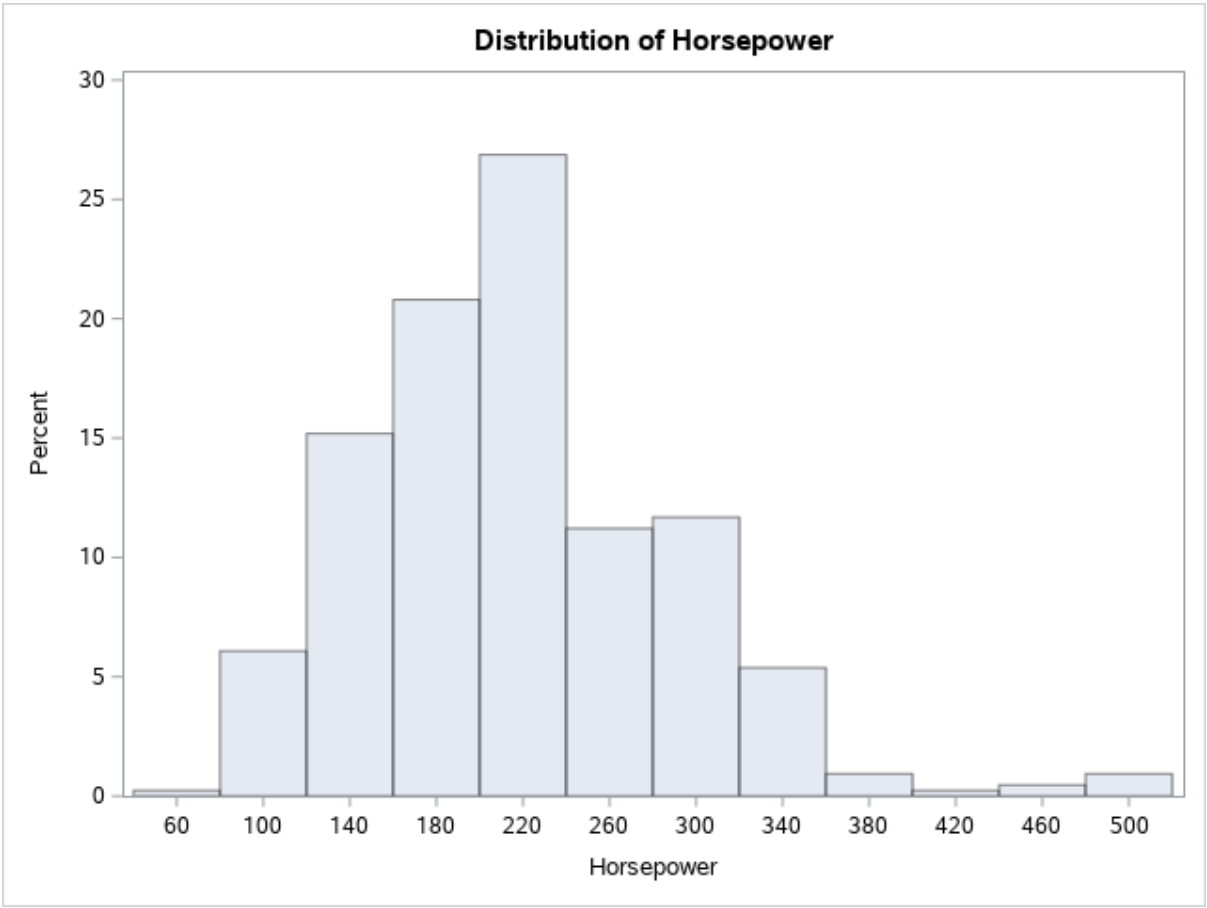
Distribution plots for each column:

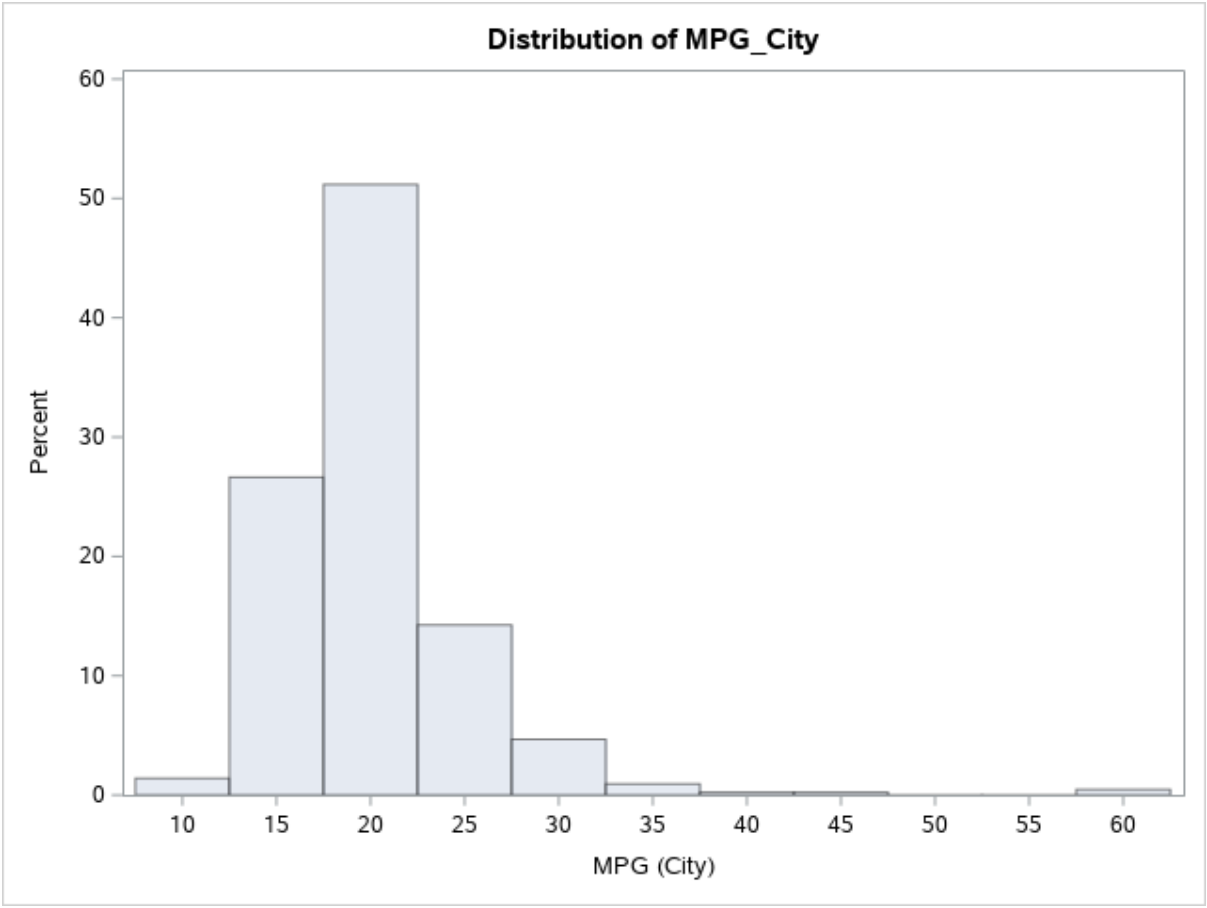


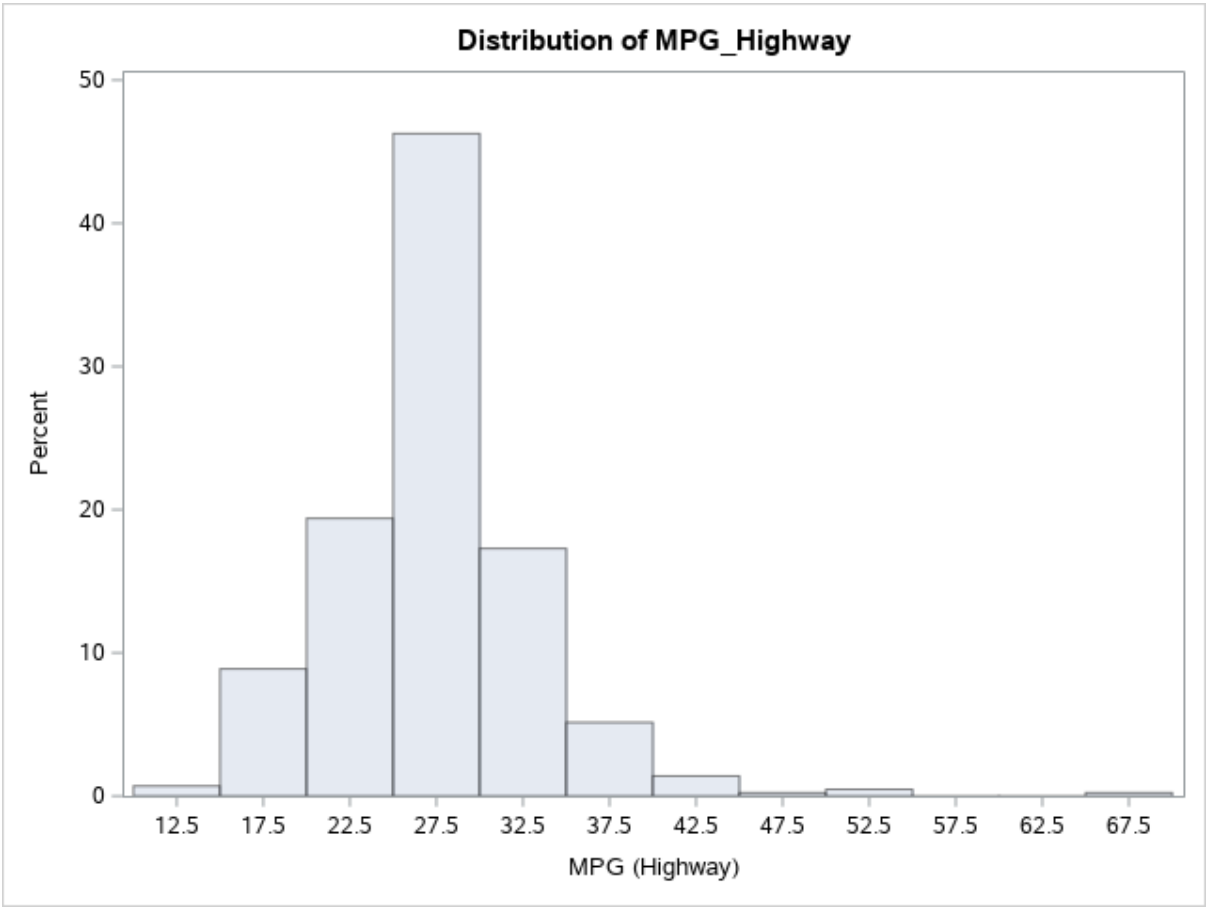

Distribution of Invoice
Percent
60
50
40
30
20
10
0
$7,500
$37,500
$67,500
$97,500
$127,500
$157,500
Invoice

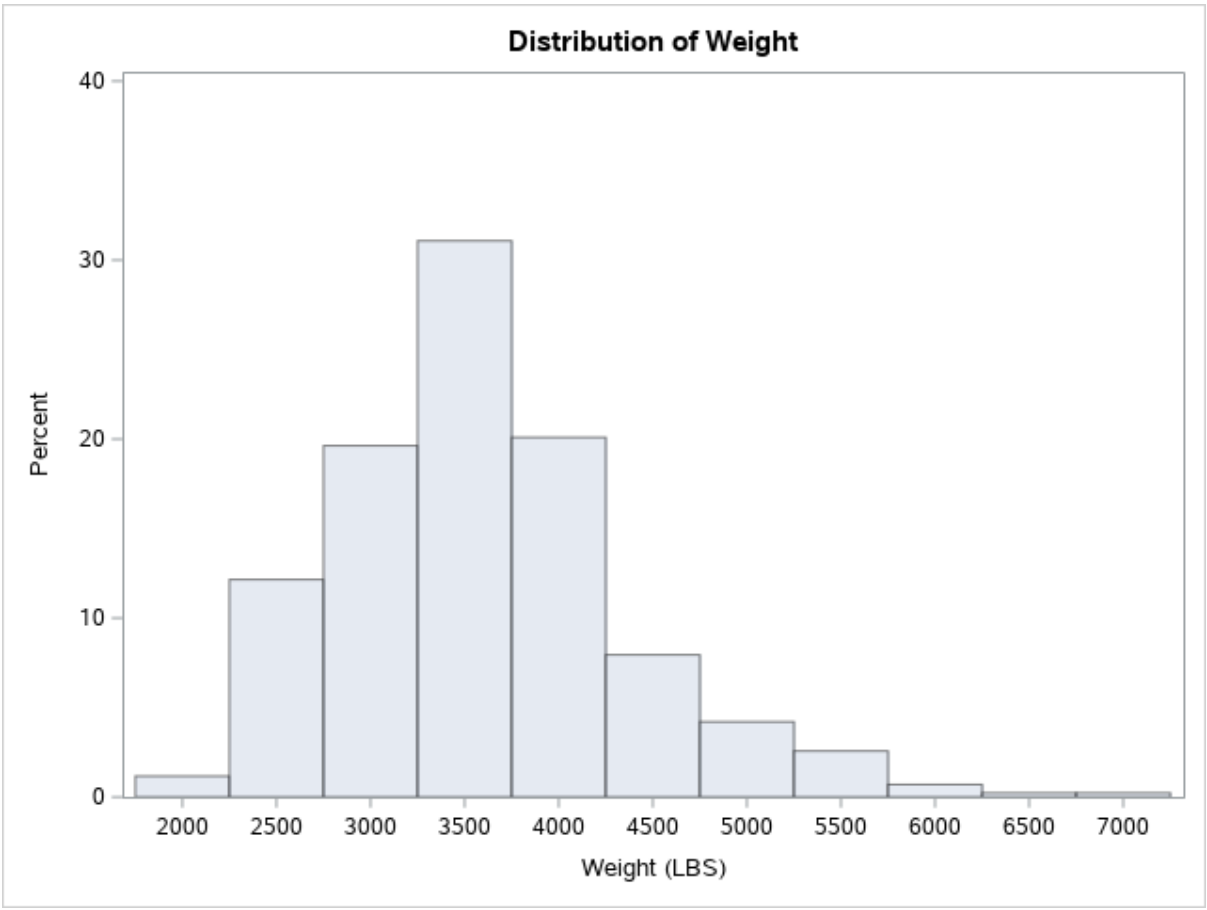

Distribution of EngineSize
Percent
30
25
20
15
10
5
0
1.5
2.1
2.7
3.3
3.9
4.5
5.1
5.7
6.3
6.9
7.5
8.1
Engine Size (L)

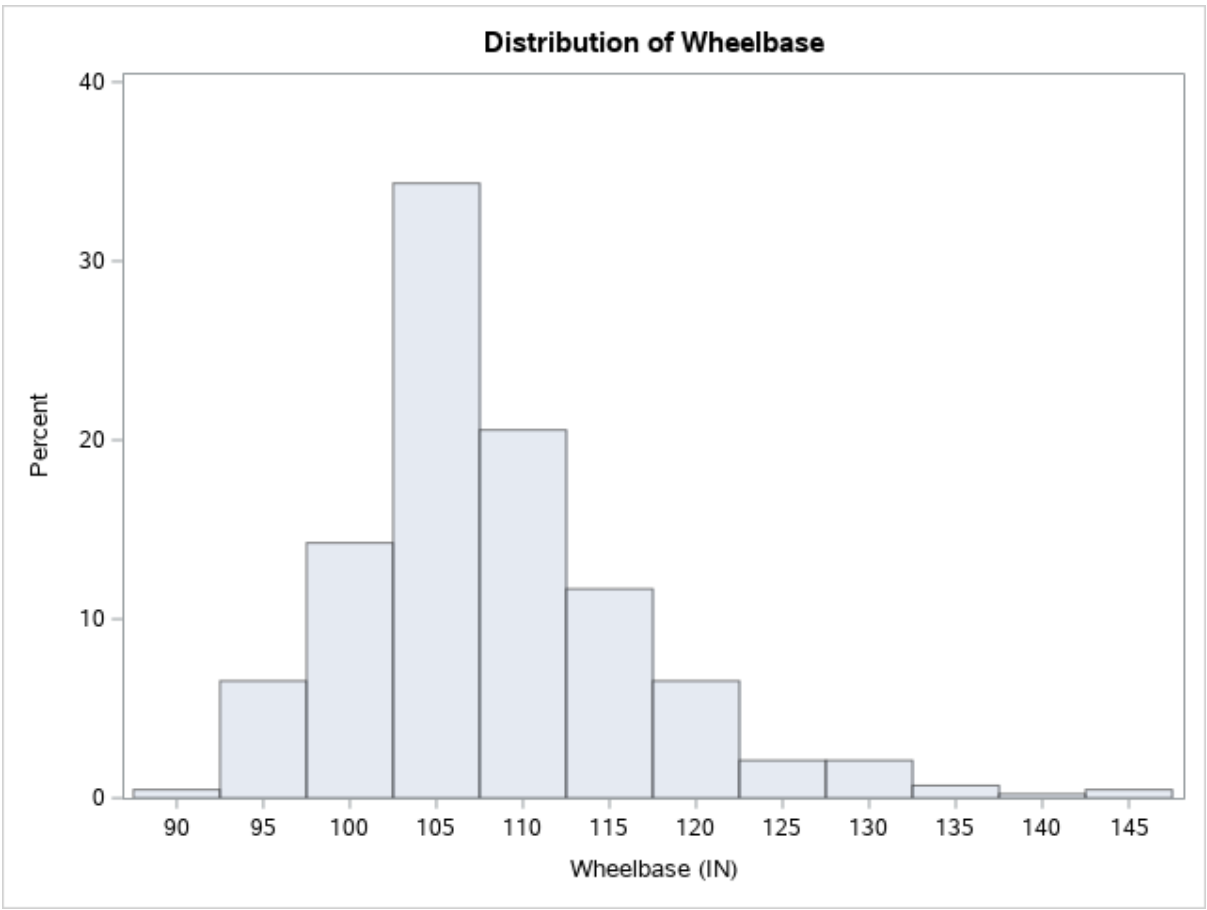

Distribution of Cylinders
Percent
50
40
30
20
10
0
3.2
4.0
4.8
5.6
6.4
7.2
8.0
8.8
9.6
10.4
11.2
12.0
Cylinders

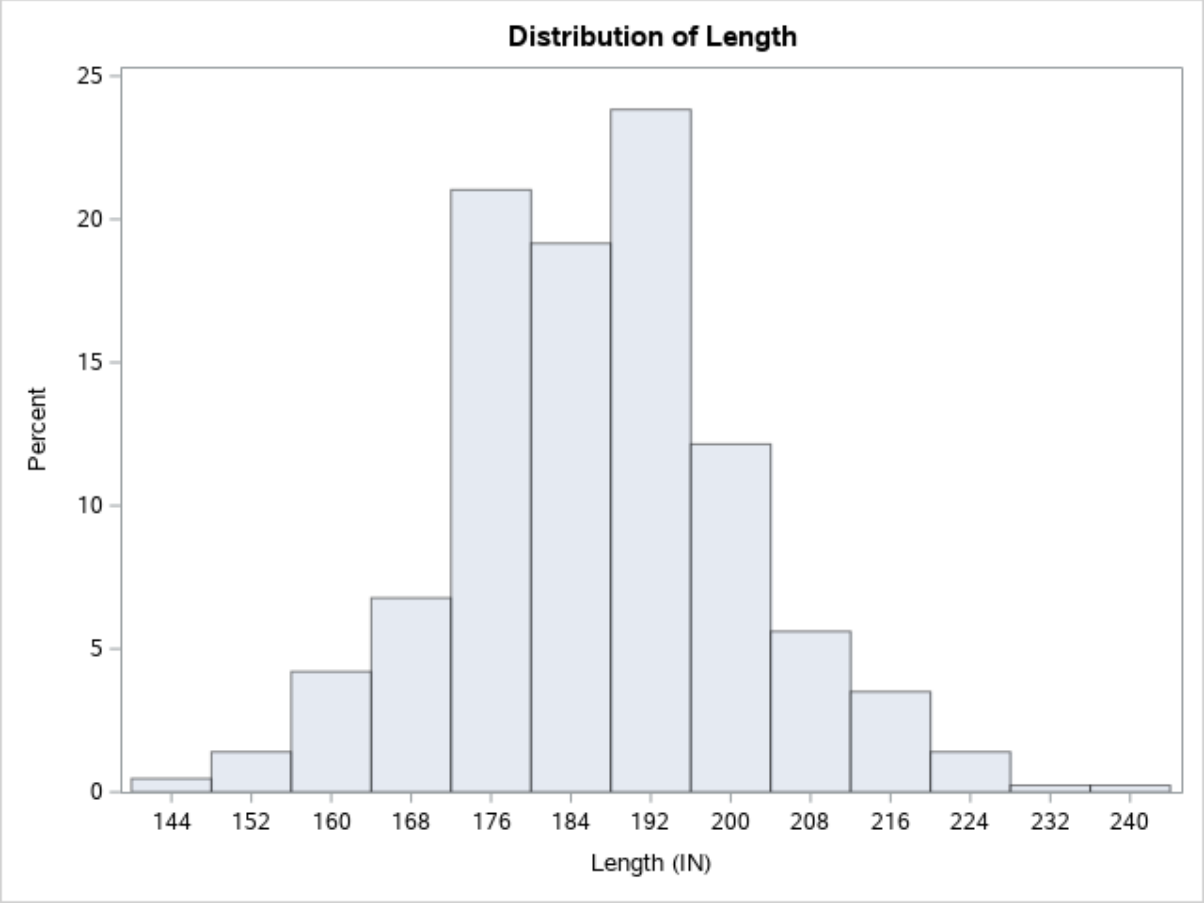

Distribution of Horsepower
Percent
30
25
20
15
10
5
0
60
100
140
180
220
260
300
340
380
420
460
500
Horsepower


Distribution of MPG_City
Percent
60
50
40
30
20
10
0
10
15
20
25
30
35
40
45
50
55
60
MPG (City)

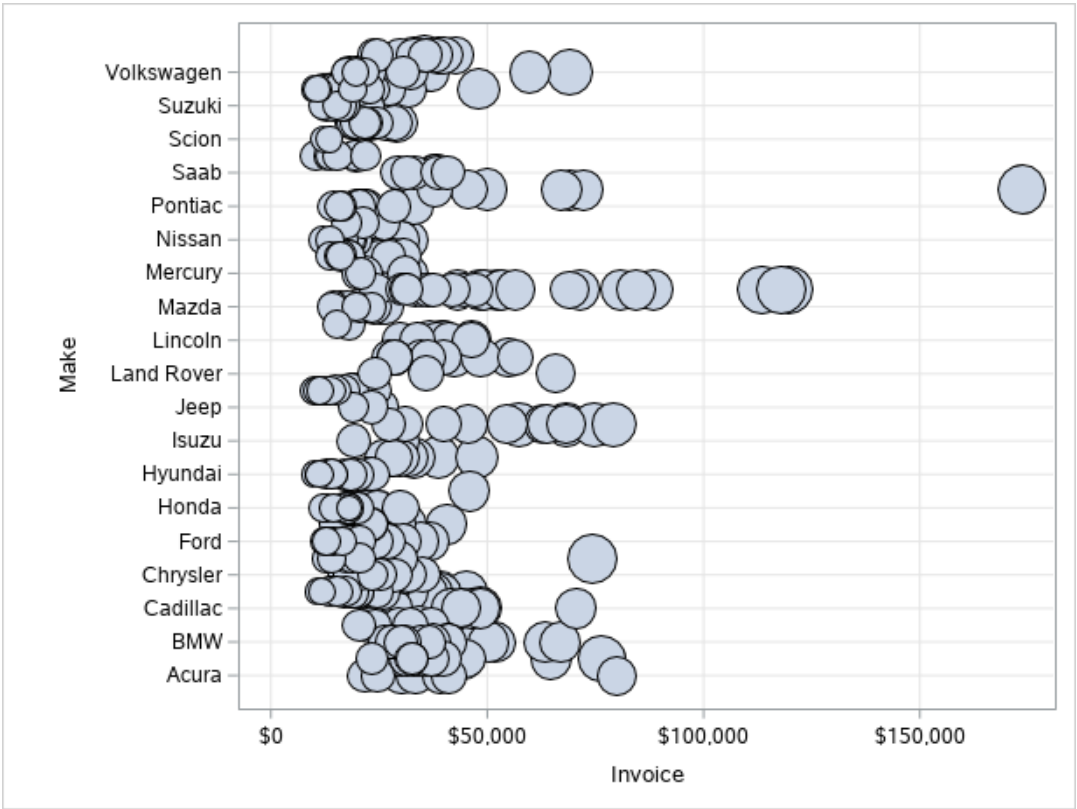

Distribution of MPG_Highway
Percent
50
40
30
20
10
0
12.5
17.5
22.5
27.5
32.5
37.5
42.5
47.5
52.5
57.5
62.5
67.5
MPG (Highway)


Distribution of Weight
Percent
40
30
20
10
0
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
Weight (LBS)

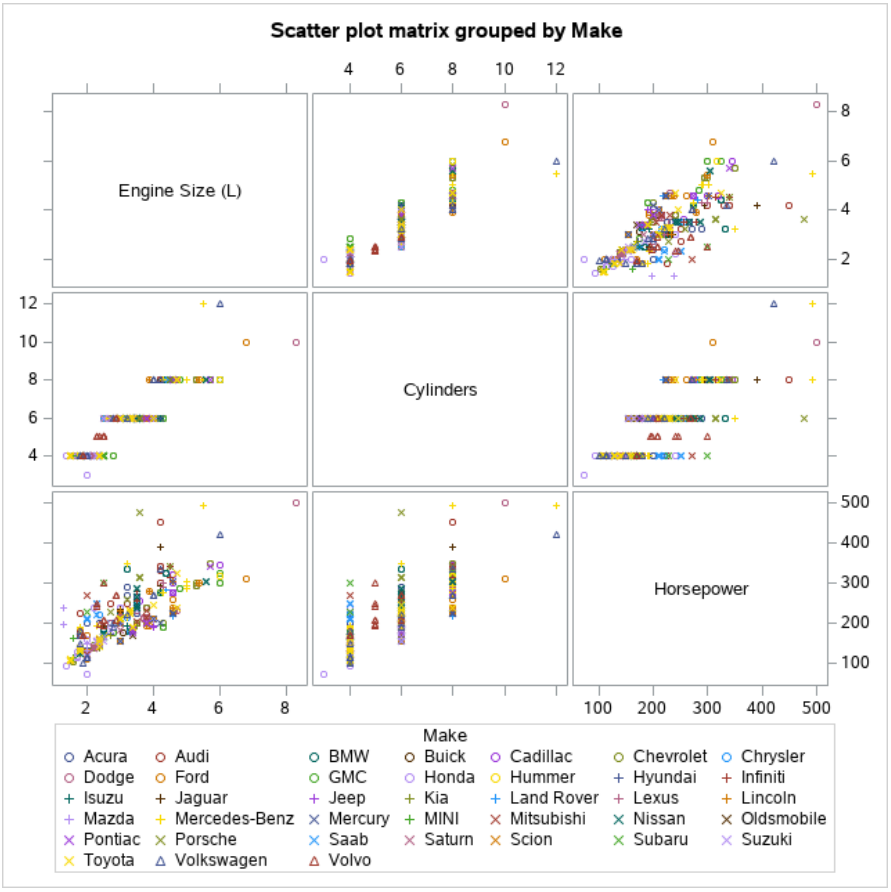

Distribution of Wheelbase
Percent
40
30
20
10
0
90
95
100
105
110
115
120
125
130
135
140
145
Wheelbase (IN)

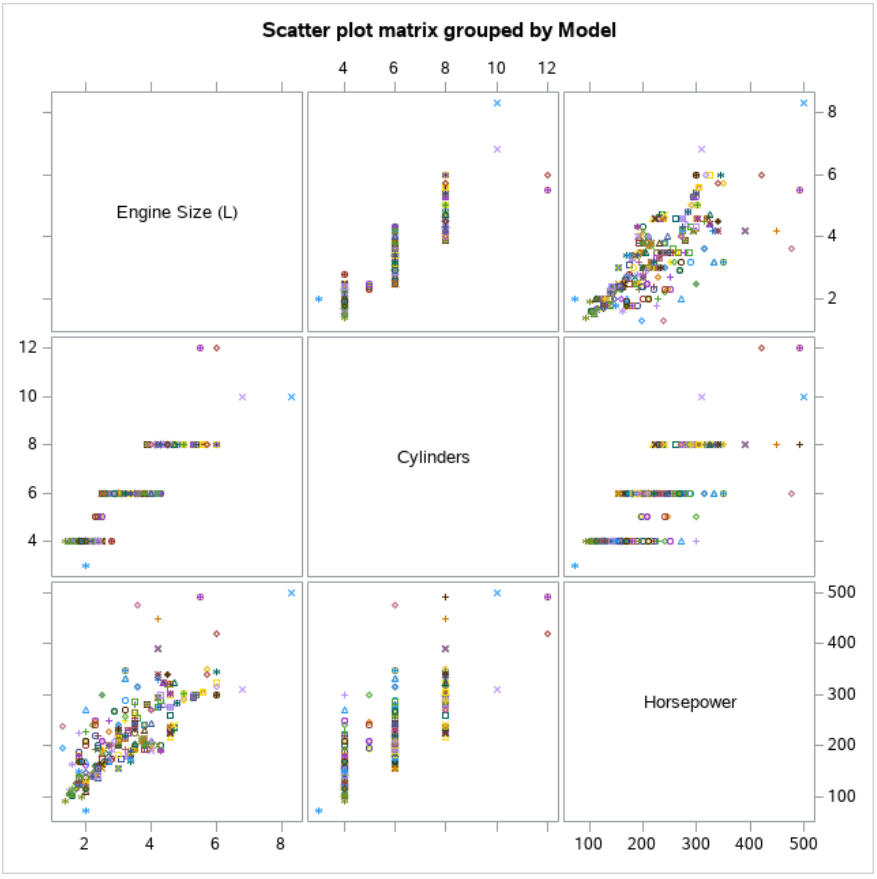


Bubble Plot between Invoice and Make depending on horsepower.

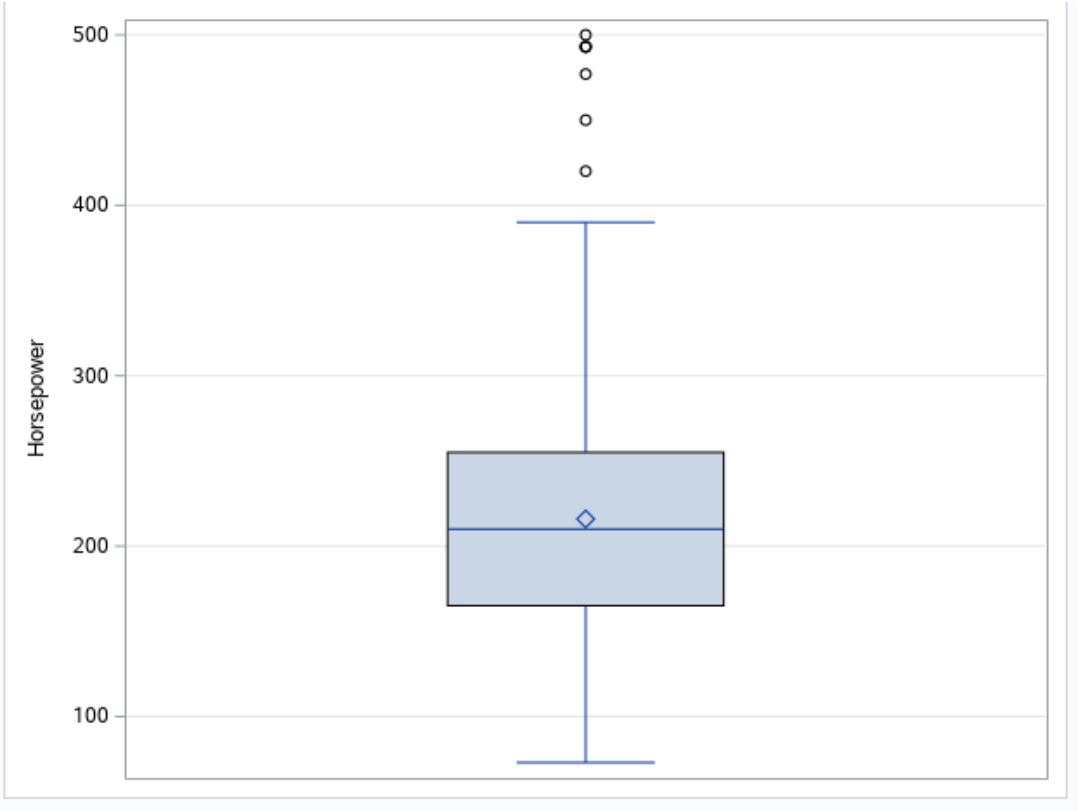


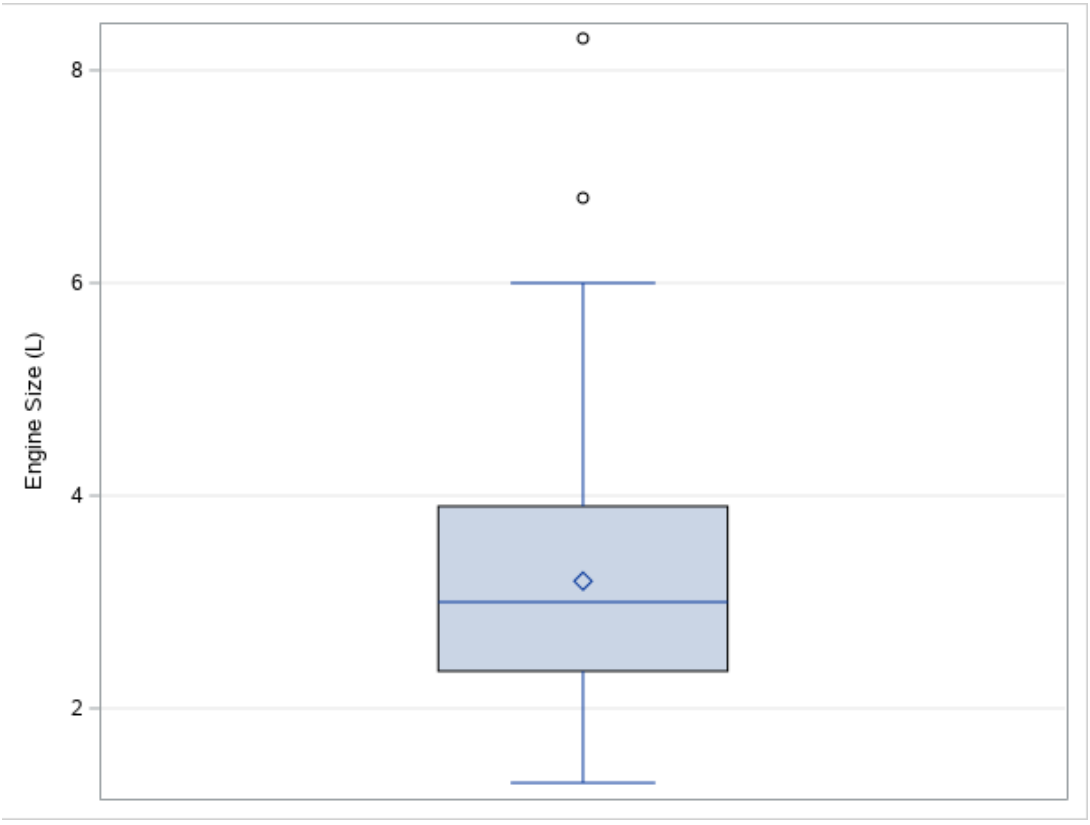
Scatter Plots of Engine Size, Cylinders and Horsepower, grouped by Make and Model

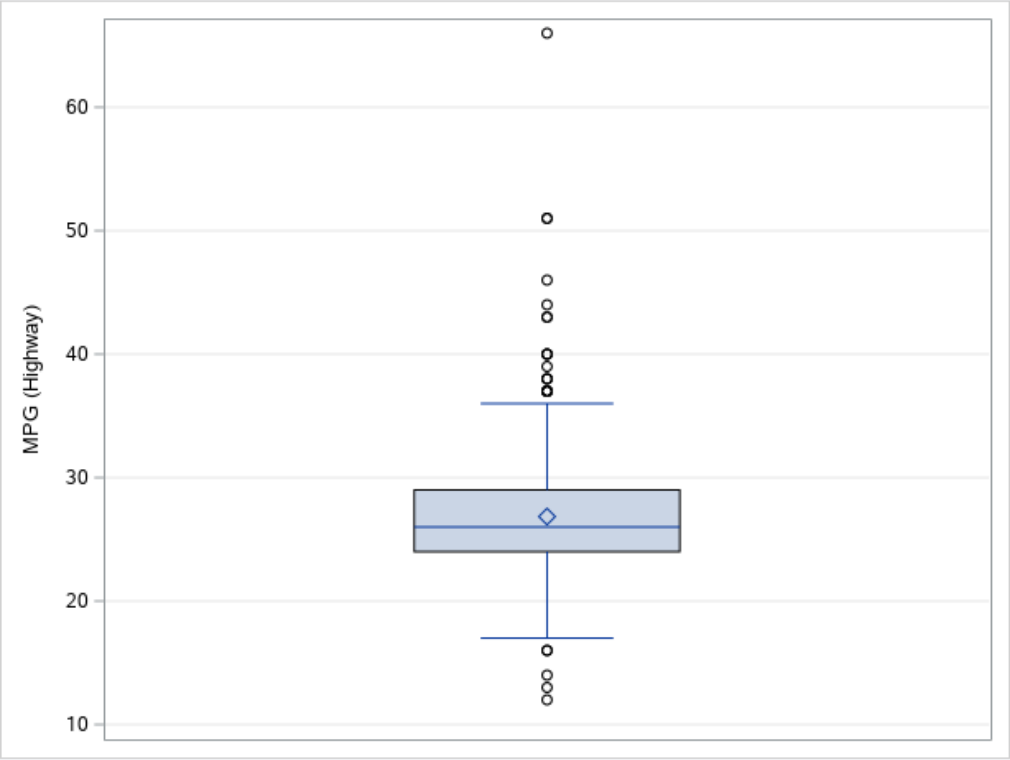




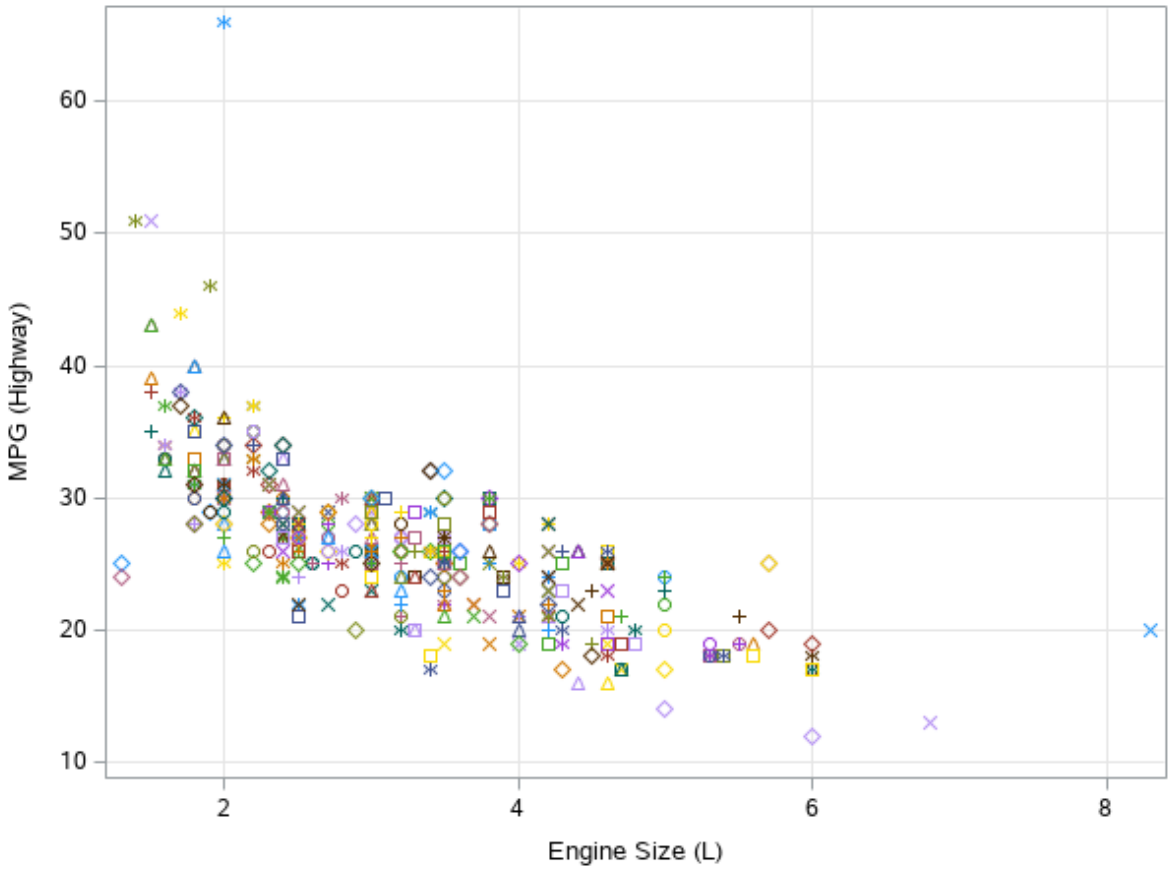
Box Plots for some columns:

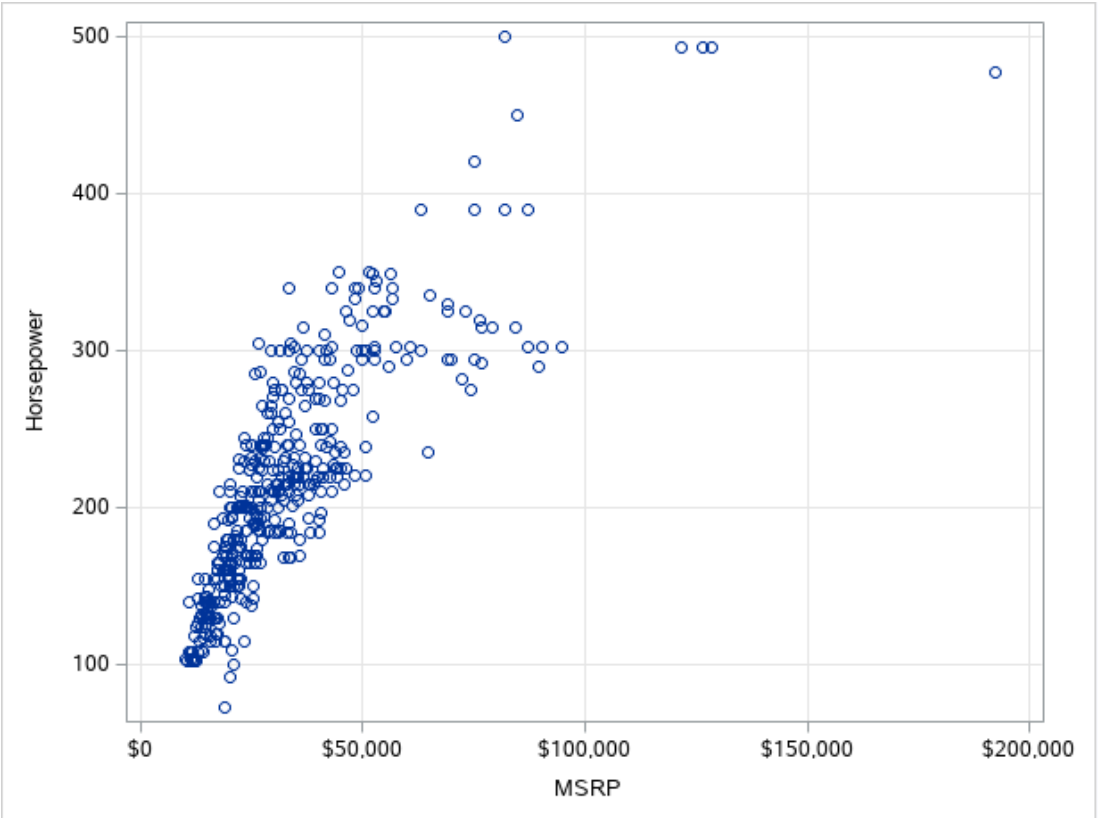






Scatter Plots between some strongly correlated variables





Conclusion:

- Computed the summary statistics of each numeric column, which is the SAS equivalent of Python's describe method.
- We clearly observe a near-perfect correlation between MSRP and Invoice, however, that is to be expected.
- The list of variables with a high correlation coefficient (i.e., > |0.7|):
    - Horsepower and Invoice
    - Engine Size and Cylinders
    - MPG City/ Highway and Engine Size
    - Weight and Engine Size.
    - Weight and MPG City/Highway
    - Wheelbase and Length
    - Engine Size and Weight
- Most of the columns in our data have a negative value for skewness, indicated by the inclination towards the left of the mean on the distribution plots.
- Plotted box plots to display the outliers for Horsepower, Engine Size and MPG Highway columns.
- Plotted the relations between Engine Size, Cylinders and Horsepower through scatter plots grouped by Make and Model of the car.
- Scatter plot between MPG and Engine Size verifies our negative correlation coefficient.
- Scatter plot between MSRP and Horsepower verifies our positive correlation coefficient.