BAN 100 Assignment: 1993 Cars Analysis

In this data audit report, we will be analyzing a data set from a client that contains information about 1993 vehicles to understand it more clearly as well as explore its statistical characteristics.

This data set has **93 rows** and **26 columns** as the table below describes the meaning of each column.

Variable	Description	Type		Unit
Manufacturer	The brand name of the vehicle manufacturer	Character	Nominal	
Model	The name of the vehicle product line	Character	Nominal	
Category	Vehicle category classification (Midsize, Compact, large, Sporty, Van)	Character	Nominal	
Min_Price	The lowest selling price of the vehicle	Numeric	Continuous	Thousand Dollars
Mid_Price	The average selling price of the vehicle	Numeric	Continuous	Thousand Dollars
Max_Price	The highest selling price of the vehicle	Numeric	Continuous	Thousand Dollars
City_Fuel	The distance in miles that a vehicle can travel per gallon in city conditions	Numeric	Discrete	Miles per Gallon
Hwy_Fuel	The distance in miles that a vehicle can travel per gallon in on highway conditions	Numeric	Discrete	Miles per Gallon
Air_Bags	Number of Airbags installed in the vehicle	Numeric	Discrete	
Drive_Train	Number of Drivetrain installed in the vehicle	Numeric	Discrete	
Cylinders	Number of Cylinders the vehicle engine has	Numeric	Discrete	
Engine_Size	The measurement of the total volume of the cylinders in the engine	Numeric	Continuous	Liters
Max_HP	The maximum amount of horsepower the engine can produce	Numeric	Discrete	HP
Max_HP_RPM	The engine RPM (rotation per minute) at the highest Horsepower	Numeric	Discrete	RPM
RPM_high*		Numeric	Discrete	RPM
Manual	If the vehicle has a manual transmission (1 is True, 0 is False)	Character	Nominal	1 = True, 2 = False
Fuel_Tank	The volume capacity of a vehicle's fuel tank in gallons	Numeric	Continuous	Gallons
Passenger	The maximum number of passengers	Numeric	Discrete	
Length	The length of a vehicle in inches	Numeric	Discrete	Inches
Wheel_Base	The horizontal distance between the centers of the front and rear wheels	Numeric	Discrete	Inches
Width	The width of a vehicle in inches	Numeric	Discrete	Inches
U_Turn_Diam	The minimum diameter of available space required for a vehicle to make a U turn	Numeric	Discrete	Feet
Rear_Room	The volume capacity of the rear seats area in cubic feet	Numeric	Continuous	Cubic Feet
Luggage	The volume capacity of the luggage area in cubic feet	Numeric	Discrete	Cubic Feet
Weight	The weight of the vehicle in pounds	Numeric	Discrete	Pounds
Domestic	If the vehicle was manufactured in the United States (1 is True, 0 is False)	Character	Nominal	1 = True, 2 = False

- More information needed for the RPM_high variable as it is not clear.

First, I have uploaded the data to SAS and use the **PROC MEANS** command to look at the data and find any missing or wrong values. The program returned the table below.

Variable	N	Mean	Std Dev	Minimum	Maximum
Min_Price	93	17.1258065	8.7460290	6.7000000	45.4000000
Mid_Price	93	20.9817204	16.7041842	7.4000000	151.0000000
Max_Price	93	21.8989247	11.0304568	7.9000000	80.0000000
City_Fuel	93	22.3655914	5.6198115	15.0000000	46.0000000
Hwy_Fuel	93	29.0860215	5.3317260	20.0000000	50.0000000
Air_Bags	93	0.8064516	0.7110627	0	2.0000000
Drive_Train	93	0.9354839	0.5276373	0	2.0000000
Cylinders	92	4.9673913	1.3046922	3.0000000	8.0000000
Engine_Size	93	2.6677419	1.0373630	1.0000000	5.7000000
Max_HP	93	143.8279570	52.3744095	55.0000000	300.0000000
Max_HP_RPM	93	5280.65	596.7316899	3800.00	6500.00
RPM_high	93	2332.20	496.5065252	1320.00	3755.00
Manual	93	0.6559140	0.4776442	0	1.0000000
Fuel_Tank	93	16.6645161	3.2793705	9.2000000	27.0000000
Passenger	93	5.0860215	1.0389785	2.0000000	8.0000000
Length	93	183.2043011	14.6023815	141.0000000	219.0000000
Wheel_Base	93	103.9462366	6.8196736	90.0000000	119.0000000
Width	93	69.3763441	3.7789865	60.0000000	78.0000000
U_Turn_Diam	93	38.9569892	3.2232645	32.0000000	45.0000000
Rear_Room	91	27.8296703	2.9890725	19.0000000	36.0000000
Luggage	82	13.8902439	2.9979668	6.0000000	22.0000000
Weight	93	3072.90	589.8965102	1695.00	4105.00
Domestic	93	0.5161290	0.5024484	0	1.0000000

Three columns (Cylinders, Rear_Room, Luggage) have missing values (1, 2, and 9) and those missing values means that the vehicle doesn't have the mentioned feature. For example, the Mazda RX-7 Sporty has a rotary engine that does not have any cylinders as well as no rear room or luggage space.

Mid_Price has a maximum value of 151 and after looking at the data it seems like it is a typographical error and the value was corrected.

There are 16 data points with zero value in the **Drive_train** variable that should be looked into as the drive train can not be zero.

I also added a column **Average_Fuel** that calculates the mean of City_Fuel and Hwy_Fuel to have a simple number to represent fuel consumption.

After fixing the data, I have decided to highlight the price and performance variables and further analyze their numbers by calculating measures of central tendencies and spread.

Variable	Mean	Median	Mode	Minimum	Maximum	Range	Lower Quartile	Upper Quartile	Quartile Range	Variance	Std Dev
Min_Price	17.1	14.7	8.4	6.7	45.4	38.7	10.8	20.3	9.5	76.5	8.7
Mid_Price	19.5	17.7	15.9	7.4	61.9	54.5	12.2	23.3	11.1	93.5	9.7
Max_Price	21.9	19.6	18.4	7.9	80.0	72.1	14.7	25.3	10.6	121.7	11.0
City_Fuel	22.4	21.0	18.0	15.0	46.0	31.0	18.0	25.0	7.0	31.6	5.6
Hwy_Fuel	29.1	28.0	26.0	20.0	50.0	30.0	26.0	31.0	5.0	28.4	5.3
Average_Fuel	25.7	24.5	23.5	17.5	48.0	30.5	22.0	27.5	5.5	29.1	5.4
Max_HP	143.8	140.0	110.0	55.0	300.0	245.0	103.0	170.0	67.0	2743.1	52.4

The table shows some interesting findings, for example even though Max_Price has a maximum value of $80,000 still 50% of the vehicles were sold between $14,700 and $25,300.

Frequency Distribution By Manufacturer's Vehicles Model

This frequency distribution illustrates the number of models produced by manufacturer, Chevrolet and Ford had the highest number of eight models made followed by Dodge, Mazda, and Pontiac. These five manufacturers contributed to 34.41% of all vehicles in the data.

Manufacturer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Chevrolet	8	8.60	8	8.60
Ford	8	8.60	16	17.20
Dodge	6	6.45	22	23.66
Mazda	5	5.38	27	29.03
Pontiac	5	5.38	32	34.41
Buick	4	4.30	36	38.71
Hyundai	4	4.30	40	43.01
Nissan	4	4.30	44	47.31
Oldsmobil	4	4.30	48	51.61
Toyota	4	4.30	52	55.91
Volkswage	4	4.30	56	60.22
Chrysler	3	3.23	59	63.44
Honda	3	3.23	62	66.67
Subaru	3	3.23	65	69.89
Acura	2	2.15	67	72.04
Audi	2	2.15	69	74.19
Cadillac	2	2.15	71	76.34
Eagle	2	2.15	73	78.49
Geo	2	2.15	75	80.65
Lexus	2	2.15	77	82.80
Lincoln	2	2.15	79	84.95
Mercedes-	2	2.15	81	87.10
Mercury	2	2.15	83	89.25
Mitsubish	2	2.15	85	91.40
Volvo	2	2.15	87	93.55
BMW	1	1.08	88	94.62
Infiniti	1	1.08	89	95.70
Plymouth	1	1.08	90	96.77
Saab	1	1.08	91	97.85
Saturn	1	1.08	92	98.92
Suzuki	1	1.08	93	100.00

Other Frequency Distributions

Category	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Midsize	22	23.66	22	23.66
Small	21	22.58	43	46.24
Compact	16	17.20	59	63.44
Sporty	14	15.05	73	78.49
Large	11	11.83	84	90.32
Van	9	9.68	93	100.00

Air_Bags	Frequency	Percent
0	34	36.56
1	43	46.24
2	16	17.20

Cylinders	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	49	53.26	49	53.26
6	31	33.70	80	86.96
8	7	7.61	87	94.57
3	3	3.26	90	97.83
5	2	2.17	92	100.00
Frequency Missing = 1				

Drive_Train	Frequency	Percent
0	16	17.20
1	67	72.04
2	10	10.75

Manual	Frequency	Percent
0	32	34.41
1	61	65.59

Passenger	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5	41	44.09	41	44.09
4	23	24.73	64	68.82
6	18	19.35	82	88.17
7	8	8.60	90	96.77
2	2	2.15	92	98.92
8	1	1.08	93	100.00

Domestic	Frequency	Percent
0	45	48.39
1	48	51.61

These frequency distributions show details such as vehicles by category where small and mid size vehicles made up 45.24%.

Most of the vehicles 53.26% had 5 passenger seats

63.44% had at least 1 Airbag

65.59% were manual transmission

51.61% of the vehicles were domestic

Question about the data

Does price affect Engine size, Horsepower, and Fuel consumption?
Do we need a bigger engine size in order to reach higher horsepower?
Is there a relation between weight and fuel consumption?

Variables Correlating with Mid_Price

Variables Corralating with Mid_Price

The CORR Procedure

3 With Variables:	Max_HP Engine_Size Average_Fuel
1 Variables:	Mid_Price

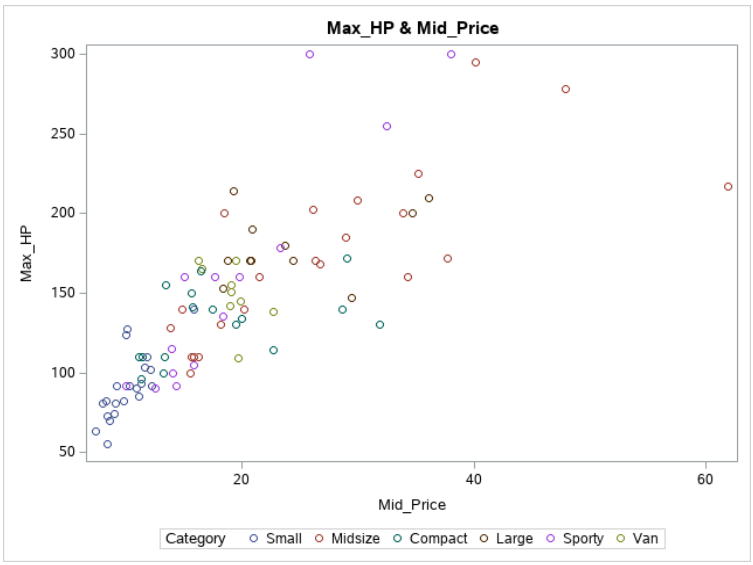
Pearson Correlation Coefficients, N = 93	
	Mid_Price
Max_HP	0.78828
Engine_Size	0.59716
Average_Fuel	-0.58671

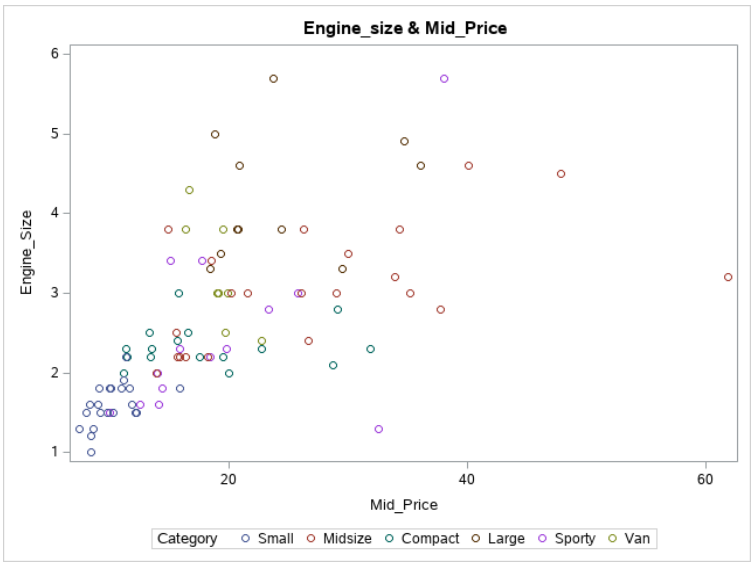
There is a strong positive correlation **0.788** between Average price and vehicle Horsepower produced by the engine.

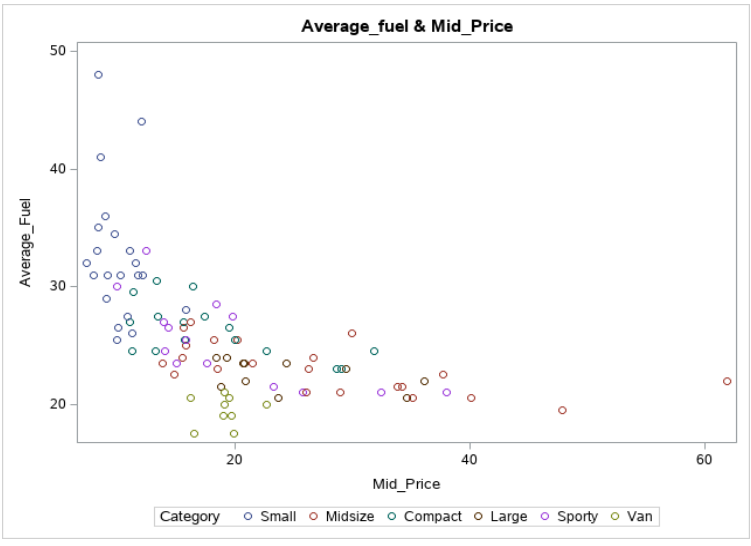
Engine Size follows the same trend but with weaker correlation **0.597** which means the more powerful the vehicle it is most likely to be more expensive

In terms of Average fuel, lower price vehicles travel a higher number of miles per gallon. There is a moderate negative correlation of **-0.586**

The graphs below illustrate the mentioned correlations and also groups categories by color.







Variables Correlating with Mid_Price

Variables Correlating with Engine_Size

The CORR Procedure

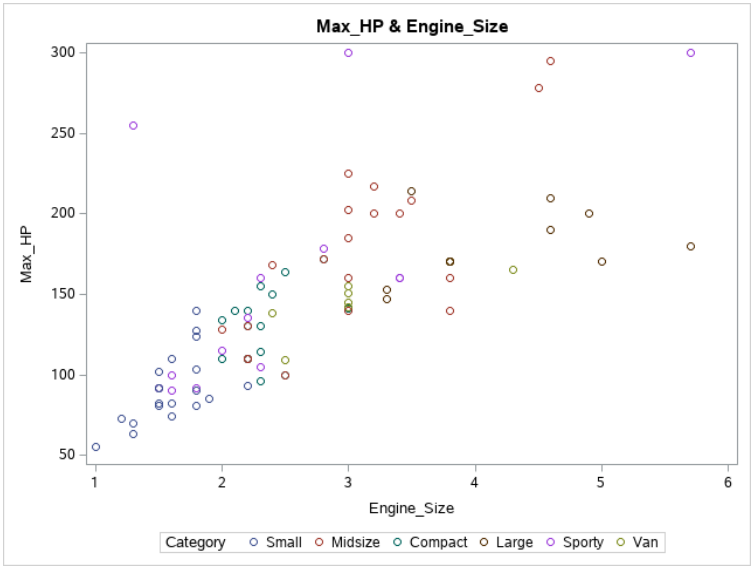
2 With Variables:	Max_HP Average_Fuel
1 Variables:	Engine_Size

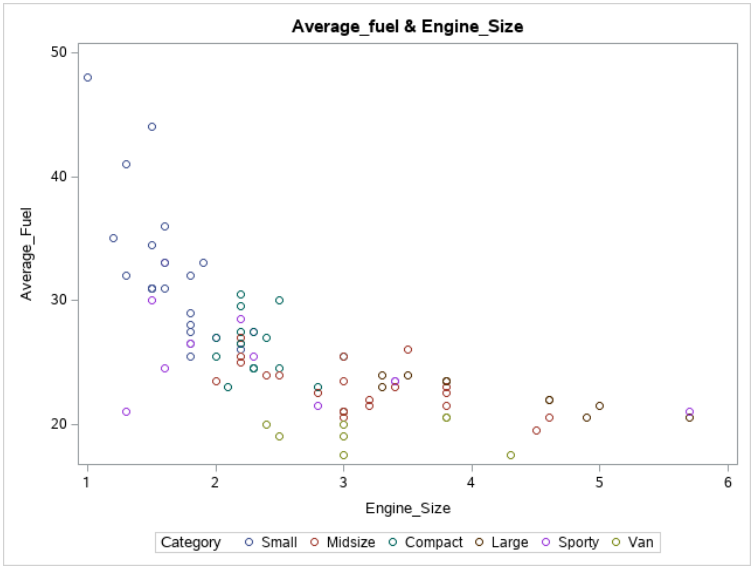
Pearson Correlation Coefficients, N = 93	
	Engine_Size
Max_HP	0.73212
Average_Fuel	-0.67907

Max_HP have a strong positive correlation **0.732** with engine size meaning the bigger the Engine_size the more likely it has higher Horsepower.

Average fuel has a moderate negative correlation **-0.679** where bigger engines travel less miles per gallon.

One vehicle (Mazda RX-7 Sporty) uses different engine technology and that is why it is able to produce 255 Horsepower with a smaller size engine of 1.3L





Variables correlating with Weight

Variables Correlating with Engine_Size

The CORR Procedure

1 With Variables:	Average_Fuel
1 Variables:	Weight

Pearson Correlation Coefficients, N = 93	
	Weight
Average_Fuel	-0.83916

There is a strong Negative correlation between Weight and Average_fuel **-0.839** which means heavier vehicles travel less miles per gallon.