Assignment 2 Data Surveying

For answering this assignment's questions, python is used for all calculations and visualizing the data. This report will present the results and a brief explanation of codes. Statistical computation and plotting different types of charts are implemented by numpy, pandas, and matplotlib libraries.

As a preprocess, this dataset is read by pandas method (pd.read_csv), then attributes, tuples, and their types should be identified. Furthermore, missing numeric values are picked out and substituted for mean values and nominal missing data replaced with the mode as well.

- **Question 1**

The Auto Imports Database file includes 26 attributes and 205 instances which 15 attributes are continuous numeric data, 10 nominal data, and just 1 attribute is integer. Here is the Type of each of them:

No.	Attribute	Type	Range
1	symbolling	Integer	From -3 to 3
2	normalized-losses	Continuous	From 65 to 256
3	make	Nominal	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4	fuel-type	Nominal	diesel, gas
5	aspiration	Nominal	std, turbo
6	num-of-doors	Nominal	four, two
7	body-style	Nominal	hardtop, wagon, sedan, hatchback, convertible

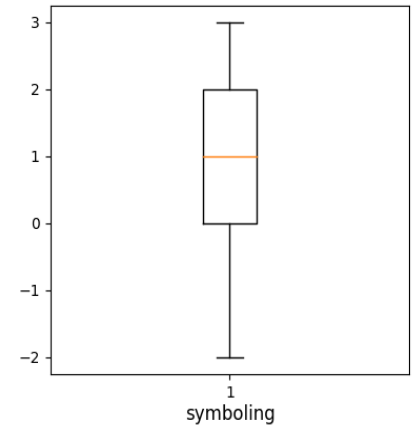
8	drive-wheels	Nominal	4wd, fwd, rwd
9	engine-location	Nominal	front, rear
10	wheel-base	Continuous	86.6 to 120.9
11	length	Continuous	141.1 to 208.1
12	width	Continuous	60.3 to 72.3
13	height	Continuous	47.8 to 59.8
14	curb-weight	Continuous	1488 to 4066
15	engine-type	Nominal	dohc, dohcv, L, ohc, ohcf, ohcv, rotor
16	num-of-cylinders	Nominal	eight, five, four, six, three, twelve, two
17	engine-size	Continuous	61 to 326
18	fuel-system	Nominal	1bbl, 2bbl,4bbl, idi, mfi, mpfi, spdi, spfi
19	bore	Continuous	2.54 to 3.94
20	stroke	Continuous	2.07 to 4.17
21	compression-ratio	Continuous	7 to 23
22	horsepower	Continuous	48 to 288
23	peak-rpm	Continuous	4150 to 6600
24	city-mpg	Continuous	13 to 49
25	highway-mpg	Continuous	16 to 54
26	price	Continuous	5118 to 45400

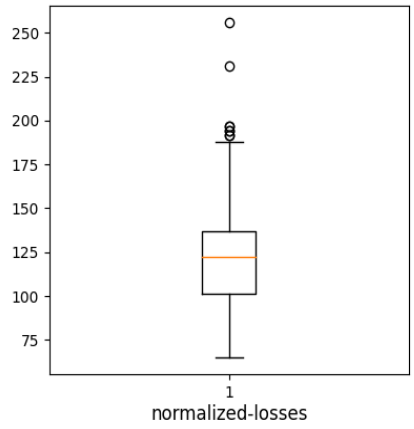
- **Question 2**

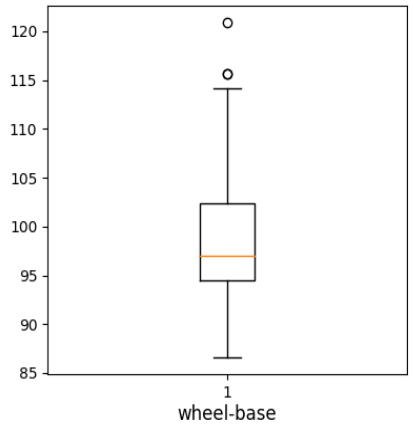
All the computational process to get mean, standard deviation, mode, min, 25%, median, 75%, and max of numerical data are done by pandas df.describe() function and df.mode(). All their boxplots are also displayed by matplotlib function (axs.boxplot)

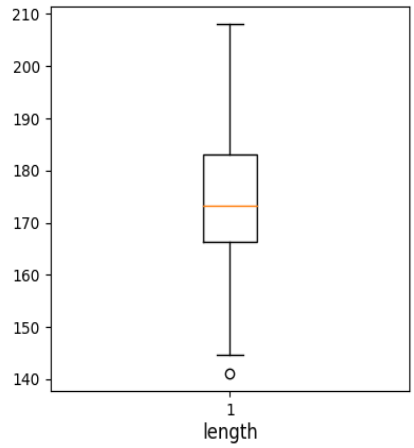
No	attribute	Mean	Standard deviation	mode	min	25%	median	75%	max
1	symboling	0.83	1.24	0	-2	0	1	2	3
2	normalized-losses	122	31.68	122	65	101	122	137	256
3	wheel-base	98.76	6.02	94.5	86.6	94.5	97	102.4	120.9
4	length	174.05	12.34	157.3	141.1	166.3	173.2	183.1	208.1
5	width	65.91	2.14	63.8	60.3	64.1	65.5	66.9	72.3
6	height	53.72	2.44	50.8	47.8	52	54.1	55.5	59.8
7	curb-weight	2555.56	520.68	2385	1488	2145	2414	2935	4066
8	engine-size	126.91	41.64	92	61	97	120	141	326
9	bore	3.33	0.27	3.62	2.54	3.15	3.31	3.58	3.94
10	stroke	3.25	0.31	3.4	2.07	3.11	3.29	3.41	4.17
11	compression-ratio	10.14	3.97	9.0	7.0	8.6	9.0	9.4	23.0
12	horsepower	104.26	39.52	68.0	48.0	70.0	95.0	116.0	288.0
13	peak-rpm	5125.37	476.98	5500.0	4150.0	4800.0	5200.0	5500.0	6600.0
14	city-mpg	25.22	6.54	31	13.0	19.0	24.0	30.0	49.0
15	highway-mpg	30.75	6.89	25	16.0	25.0	30.0	34.0	54.0
16	price	13207.13	7868.77	13207.13	5118.0	7788.0	10595.0	16500.0	45400.0

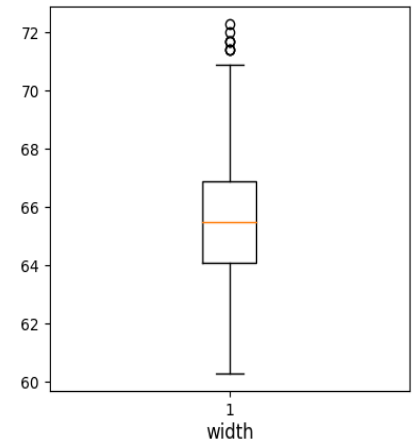
Boxplot of numeric data:

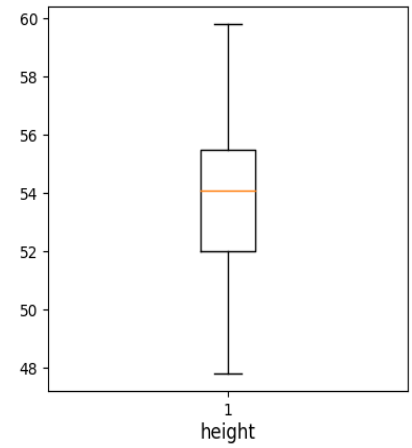


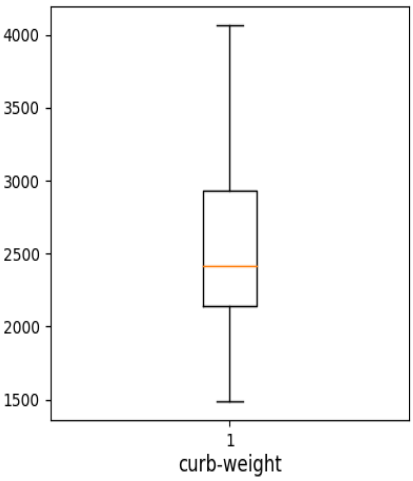


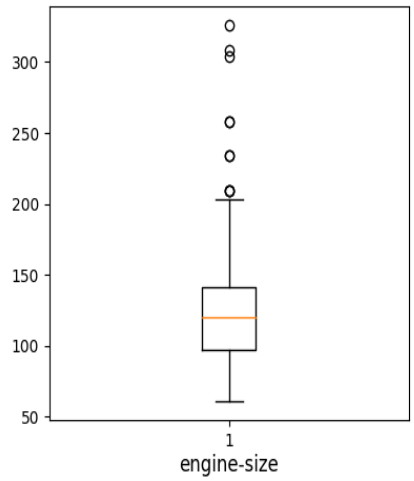


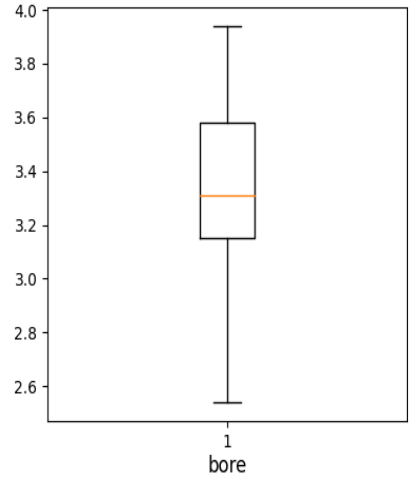


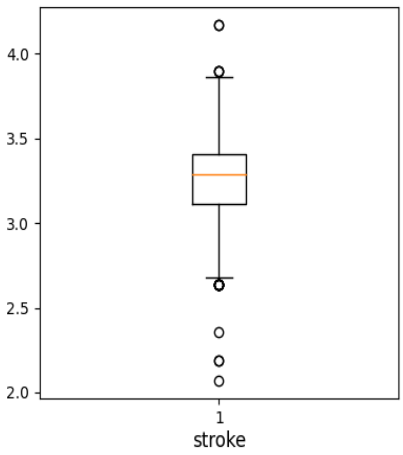


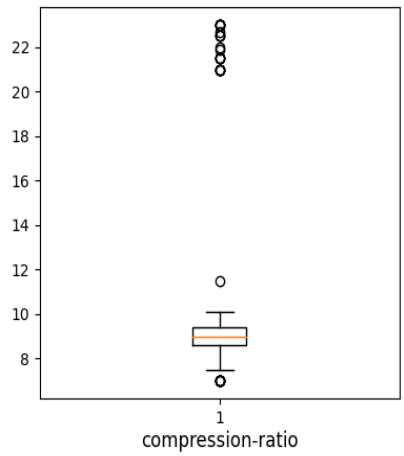


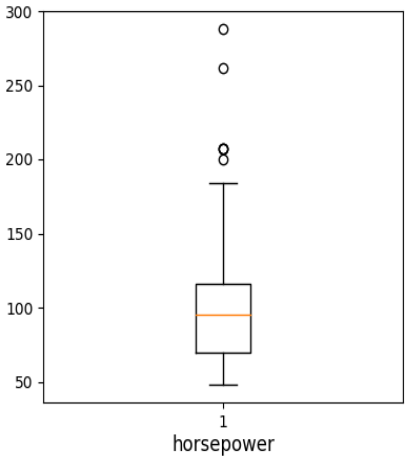


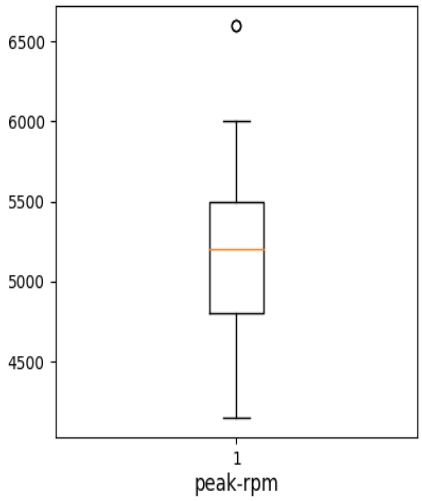


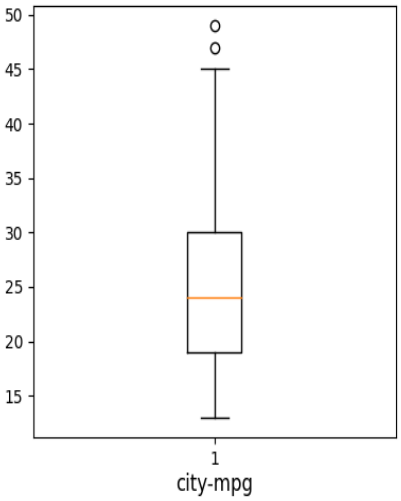


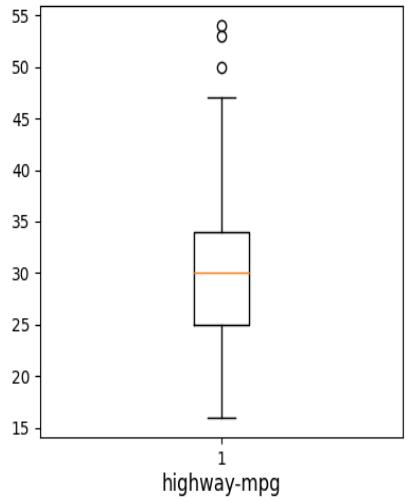

4.0
3.5
3.0
2.5
2.0
1
stroke

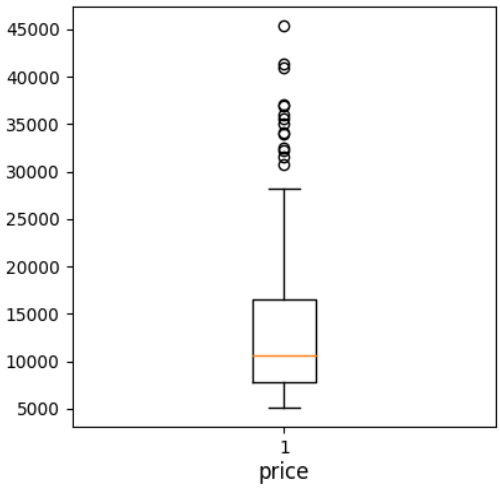

22
20
18
16
14
12
10
8
1
compression-ratio


300
250
200
150
100
50
1
horsepower


6500
6000
5500
5000
4500
1
peak-rpm


50
45
40
35
30
25
20
15
1
city-mpg

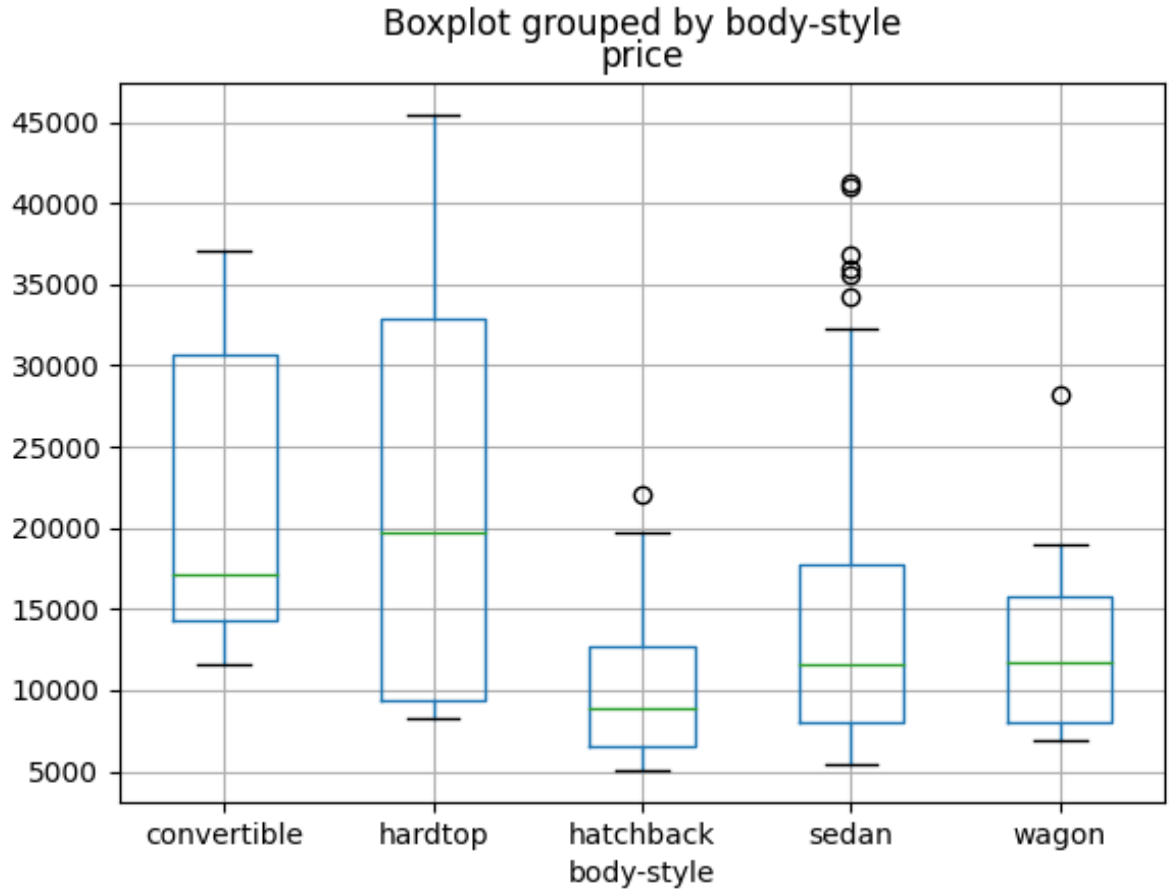

55
50
45
40
35
30
25
20
15
1
highway-mpg


45000
40000
35000
30000
25000
20000
15000
10000
5000
1
price

- **Question 3**

These boxplots show the distributions of price between the different body styles category.

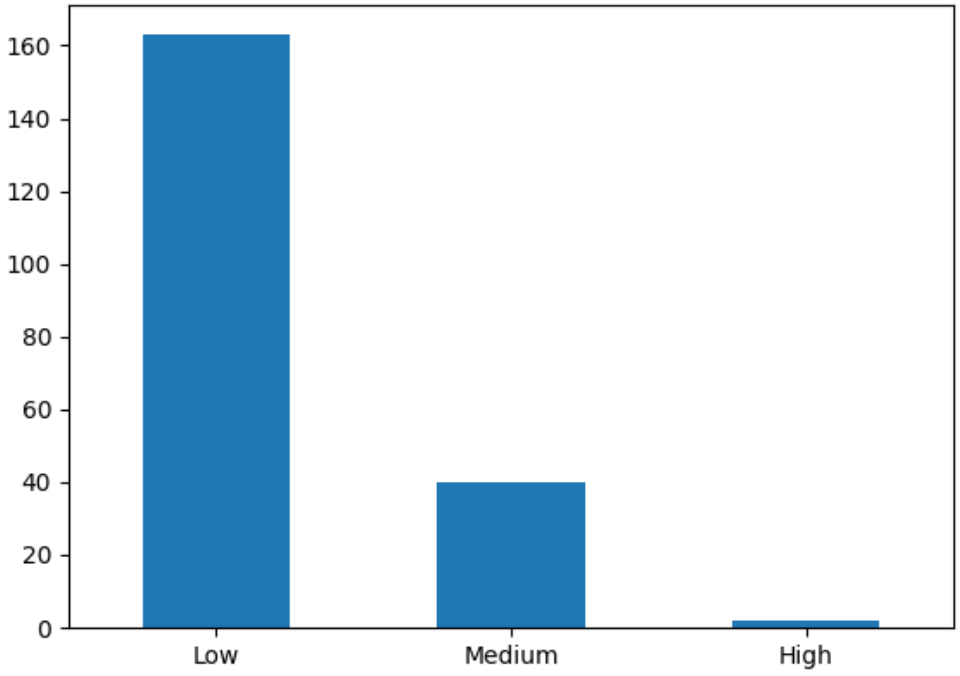
In pandas library, boxplots of variables distributions grouped by the values of a third variable can be created using the argument *by*.



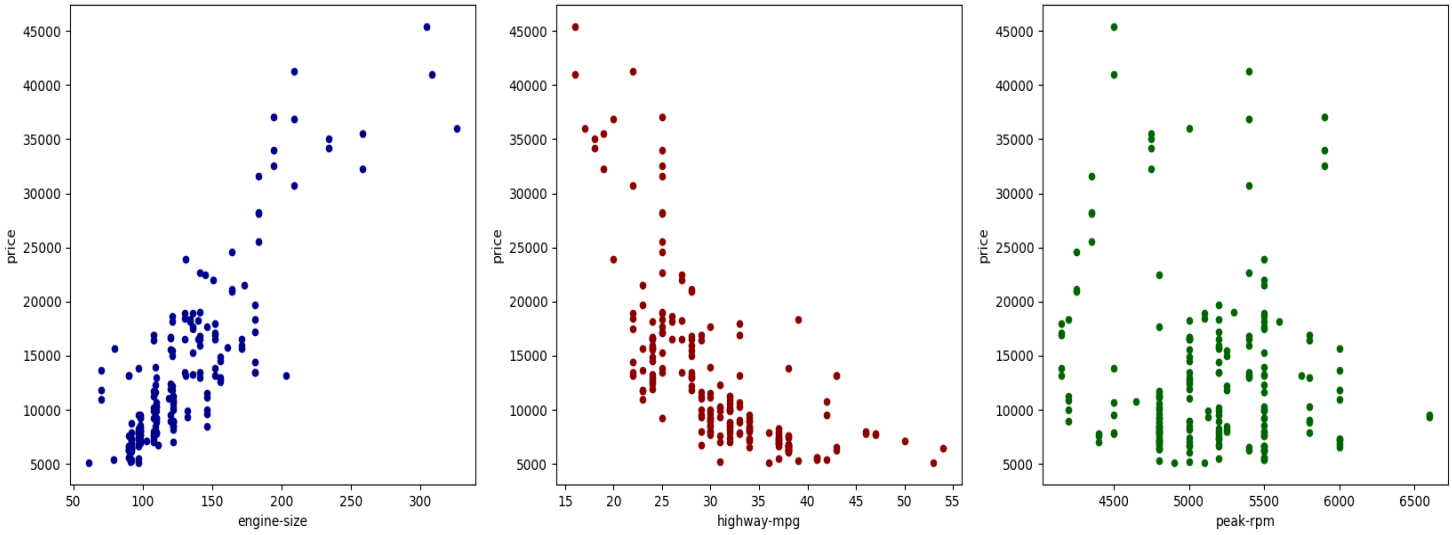
- **Question 4**

This histogram shows the distribution of horsepower. The values are between 48 and 288 and divided into three equal lengths.

For displaying this histogram, one attribute (hp_category) is added to the data with nominal objects (Low, Medium, and High) in order to define the counts of each new distinct values of this attribute and display its histogram by using df.plot.bar() function.



- **Question 5**



The first scatterplot shows the relatively positive correlation between engine size and price. Thus the larger engine size of the car, we could expect a higher price. The scatterplot between highway-mpg and price shows the different correlation. It approximately has a negative correlation which means the price of the car gets lower when the value of highway-mpg is bigger. The data in the third plot has no relationship which is seen in the two other plots. It seems that there is no relation between peak-rpm and the car price.

Using the df.plot.scatter function is sufficient for these data in order to show mentioned scatter plots in python.

- **Question 6**

To visualizing three different attributes (body style, drive wheel, and price) that two of them are categorical (nominal) and the other one is numeric and continuous data, the chart below could describe the relationships among them.

On the axes of the chart, there are two categorical attributes. So there are 15 combinations, for instance: cars with hard top body-style and RWD drive-wheel. But as the chart shows, neither hardtop body-style with 4wd wheel-style nor convertible with 4wd wheel-style doesn't exist in the data. For each of these combinations, the mean price is calculated and shown as a purple dot. The larger dot for one case points out the higher price for it. Hence, this chart indicates that any types of body-style with RWD wheel-style are more expensive than other wheel-styles. On the other hand, all types of body-style with FWD wheel-style are the cheapest. Although 4WD style doesn't manufacture for hardtop and convertible body styles, the prices of this kind of wheel for other body-style are a little bit more expensive than cars with FWD style.

In python, the dictionary data structure is used to store all combinations of car's body-style and wheel style mapped to each groups' count and sum price. Then the mean price is calculated for each key. Then we create a new pandas data frame with three attributes (drive-wheel, body-style, and mean price). The last step in the code is using pandas scatter plot so as to visualize this new data frame and analyze it.