In [1]:

```
import numpy as np
import pandas as pd
```

In [2]:

```
df = pd.read_csv('Automobile price data _Raw_.csv')
```

In [3]:

```
df.head()
```

Out[3]:

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	whee bas
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.
3	2	164	audi	gas	std	four	sedan	fwd	front	99.
4	2	164	audi	gas	std	four	sedan	4wd	front	99.

5 rows × 26 columns

In [4]:

```
df.shape
```

Out[4]:

```
(205, 26)
```

In [5]:

```
df.describe()
```

Out[5]:

	symboling	wheel-base	length	width	height	curb-weight	engine-size	c
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	
mean	0.834146	98.756585	174.049268	65.907805	53.724878	2555.565854	126.907317	
std	1.245307	6.021776	12.337289	2.145204	2.443522	520.680204	41.642693	
min	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	
25%	0.000000	94.500000	166.300000	64.100000	52.000000	2145.000000	97.000000	
50%	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	
75%	2.000000	102.400000	183.100000	66.900000	55.500000	2935.000000	141.000000	
max	3.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000	

In [6]:

```
df.isnull().sum()
```

Out[6]:

```
symboling            0
normalized-losses    0
make                 0
fuel-type            0
aspiration           0
num-of-doors         0
body-style           0
drive-wheels         0
engine-location      0
wheel-base           0
length               0
width                0
height               0
curb-weight          0
engine-type          0
num-of-cylinders     0
engine-size          0
fuel-system          0
bore                 0
stroke               0
compression-ratio    0
horsepower           0
peak-rpm             0
city-mpg             0
highway-mpg          0
price                0
dtype: int64
```

In [7]:

```
df['symboling'].unique()
```

Out[7]:

```
array([ 3,  1,  2,  0, -1, -2], dtype=int64)
```

In [8]:

```
df['normalized-losses'].unique()
```

Out[8]:

```
array(['?', '164', '158', '192', '188', '121', '98', '81', '118', '148',
       '110', '145', '137', '101', '78', '106', '85', '107', '104', '113',
       '150', '129', '115', '93', '142', '161', '153', '125', '128',
       '122', '103', '168', '108', '194', '231', '119', '154', '74',
       '186', '83', '102', '89', '87', '77', '91', '134', '65', '197',
       '90', '94', '256', '95'], dtype=object)
```

In [9]:

```
df['normalized-losses'].loc[df['normalized-losses'] == '?'].count()
```

Out[9]:

```
41
```

In [10]:

```
df['normalized-losses'].str.isnumeric().value_counts()
```

Out[10]:

```
True     164
False     41
Name: normalized-losses, dtype: int64
```

In [11]:

```
df['normalized-losses'].loc[df['normalized-losses'].str.isnumeric() == False]
```

Out[11]:

```
0      ?
1      ?
2      ?
5      ?
7      ?
9      ?
14     ?
15     ?
16     ?
17     ?
43     ?
44     ?
45     ?
46     ?
48     ?
49     ?
63     ?
66     ?
71     ?
73     ?
74     ?
75     ?
82     ?
83     ?
84     ?
109    ?
110    ?
113    ?
114    ?
124    ?
126    ?
127    ?
128    ?
129    ?
130    ?
131    ?
181    ?
189    ?
191    ?
192    ?
193    ?
Name: normalized-losses, dtype: object
```

In [12]:

```
nl = df['normalized-losses'] .loc[df['normalized-losses']  != '?']
nl_m = nl.astype(str).astype(int).mean()
df['normalized-losses']  = df['normalized-losses'] .replace('?',nl_m).astype(int)
df['normalized-losses'].head()
```

Out[12]:

```
0    122
1    122
2    122
3    164
4    164
Name: normalized-losses, dtype: int32
```

In [13]:

```
df['make'].unique()
```

Out[13]:

```
array(['alfa-romero', 'audi', 'bmw', 'chevrolet', 'dodge', 'honda',
       'isuzu', 'jaguar', 'mazda', 'mercedes-benz', 'mercury',
       'mitsubishi', 'nissan', 'peugot', 'plymouth', 'porsche', 'renault',
       'saab', 'subaru', 'toyota', 'volkswagen', 'volvo'], dtype=object)
```

In [14]:

```
df['num-of-doors'].unique()
```

Out[14]:

```
array(['two', 'four', '?'], dtype=object)
```

In [15]:

```
df['fuel-type'].unique()
```

Out[15]:

```
array(['gas', 'diesel'], dtype=object)
```

In [16]:

```
df['aspiration'].unique()
```

Out[16]:

```
array(['std', 'turbo'], dtype=object)
```

In [17]:

```
df['num-of-doors'].unique()
```

Out[17]:

```
array(['two', 'four', '?'], dtype=object)
```

In [18]:

```
df['num-of-doors'].isnull().sum()
```

Out[18]:

```
0
```

In [19]:

```
df['num-of-doors'].loc[df['num-of-doors'] == '?'].count()
```

Out[19]:

```
2
```

In [20]:

```
df['num-of-doors'].str.isnumeric().value_counts()
```

Out[20]:

```
False    205
Name: num-of-doors, dtype: int64
```

In [21]:

```
df['num-of-doors'].loc[df['num-of-doors'].str.isnumeric() == False]
```

Out[21]:

```
0       two
1       two
2       two
3      four
4      four
       ... 
200    four
201    four
202    four
203    four
204    four
Name: num-of-doors, Length: 205, dtype: object
```

In [22]:

```
# remove the records which are having the value '?'
df['num-of-doors'].loc[df['num-of-doors'] == '?']
df = df[df['num-of-doors'] != '?']
df['num-of-doors'].loc[df['num-of-doors'] == '?']
```

Out[22]:

```
Series([], Name: num-of-doors, dtype: object)
```

In [23]:

```
df['drive-wheels'].unique()
```

Out[23]:

```
array(['rwd', 'fwd', '4wd'], dtype=object)
```

In [24]:

```
df['drive-wheels']  = df['drive-wheels'] .replace('4wd','rwd')
df['drive-wheels'].head()
```

Out[24]:

```
0    rwd
1    rwd
2    rwd
3    fwd
4    rwd
Name: drive-wheels, dtype: object
```

In [25]:

```
df['drive-wheels'].unique()
```

Out[25]:

```
array(['rwd', 'fwd'], dtype=object)
```

In [26]:

```
df['engine-location'].unique()
```

Out[26]:

```
array(['front', 'rear'], dtype=object)
```

In [27]:

```
df['length'].unique()
```

Out[27]:

```
array([168.8, 171.2, 176.6, 177.3, 192.7, 178.2, 176.8, 189. , 193.8,
       197. , 141.1, 155.9, 158.8, 157.3, 174.6, 173.2, 144.6, 150. ,
       163.4, 157.1, 167.5, 175.4, 169.1, 170.7, 172.6, 199.6, 191.7,
       159.1, 166.8, 169. , 177.8, 175. , 190.9, 187.5, 202.6, 180.3,
       208.1, 199.2, 178.4, 173. , 172.4, 165.3, 170.2, 165.6, 162.4,
       173.4, 181.7, 184.6, 178.5, 186.7, 198.9, 167.3, 168.9, 175.7,
       181.5, 186.6, 156.9, 157.9, 172. , 173.5, 173.6, 158.7, 169.7,
       166.3, 168.7, 176.2, 175.6, 183.5, 187.8, 171.7, 159.3, 165.7,
       180.2, 183.1, 188.8])
```

In [28]:

```
df['width'].unique()
```

Out[28]:

```
array([64.1, 65.5, 66.2, 66.4, 66.3, 71.4, 67.9, 64.8, 66.9, 70.9, 60.3,
       63.6, 63.8, 64.6, 63.9, 64. , 65.2, 62.5, 66. , 61.8, 69.6, 70.6,
       64.2, 65.7, 66.5, 66.1, 70.3, 71.7, 70.5, 72. , 68. , 64.4, 65.4,
       68.4, 68.3, 65. , 72.3, 66.6, 63.4, 65.6, 67.7, 67.2, 68.9, 68.8])
```

In [29]:

```
df['height'].unique()
```

Out[29]:

```
array([48.8, 52.4, 54.3, 53.1, 55.7, 55.9, 52. , 53.7, 56.3, 53.2, 50.8,
       50.6, 59.8, 50.2, 52.6, 54.5, 58.3, 53.3, 54.1, 51. , 53.5, 51.4,
       52.8, 47.8, 49.6, 55.5, 54.4, 56.5, 58.7, 54.9, 56.7, 55.4, 54.8,
       49.4, 51.6, 54.7, 55.1, 56.1, 49.7, 56. , 50.5, 55.2, 52.5, 53. ,
       59.1, 53.9, 55.6, 56.2, 57.5])
```

In [30]:

```
df['curb-weight'].unique()
```

Out[30]:

```
array([2548, 2823, 2337, 2824, 2507, 2844, 2954, 3086, 3053, 2395, 2710,
       2765, 3055, 3230, 3380, 3505, 1488, 1874, 1909, 1876, 2128, 1967,
       1989, 2535, 2811, 1713, 1819, 1837, 1940, 1956, 2010, 2024, 2236,
       2289, 2304, 2372, 2465, 2293, 2734, 4066, 3950, 1890, 1900, 1905,
       1945, 1950, 2380, 2385, 2500, 2410, 2425, 2670, 2700, 3515, 3750,
       3495, 3770, 3740, 3685, 3900, 3715, 2910, 1918, 1944, 2004, 2145,
       2370, 2328, 2833, 2921, 2926, 2365, 2405, 2403, 1889, 2017, 1938,
       1951, 2028, 1971, 2037, 2008, 2324, 2302, 3095, 3296, 3060, 3071,
       3139, 3020, 3197, 3430, 3075, 3252, 3285, 3485, 3130, 2191, 2818,
       2778, 2756, 2800, 3366, 2579, 2460, 2658, 2695, 2707, 2758, 2808,
       2847, 2050, 2120, 2240, 2190, 2340, 2510, 2290, 2455, 2420, 2650,
       1985, 2040, 2015, 2280, 3110, 2081, 2109, 2275, 2094, 2122, 2140,
       2169, 2204, 2265, 2300, 2540, 2536, 2551, 2679, 2714, 2975, 2326,
       2480, 2414, 2458, 2976, 3016, 3131, 3151, 2261, 2209, 2264, 2212,
       2319, 2254, 2221, 2661, 2563, 2912, 3034, 2935, 3042, 3045, 3157,
       2952, 3049, 3012, 3217, 3062], dtype=int64)
```

In [31]:

```
df['engine-type'].unique()
```

Out[31]:

```
array(['dohc', 'ohcv', 'ohc', 'l', 'rotor', 'ohcf', 'dohcv'], dtype=object)
```

In [32]:

```
df['engine-type'].replace({'ohcv':'ohc','dohcv':'dohc','ohcf':'ohc'}, inplace=True)
df['engine-type'].head()
```

Out[32]:

```
0    dohc
1    dohc
2     ohc
3     ohc
4     ohc
Name: engine-type, dtype: object
```

In [33]:

```
df['engine-type'].unique()
```

Out[33]:

```
array(['dohc', 'ohc', 'l', 'rotor'], dtype=object)
```

In [34]:

```
df['engine-type'].loc[df['engine-type'] == 'l']
df = df[df['engine-type'] != 'l']
df['engine-type'].loc[df['engine-type'] == 'l']
```

Out[34]:

```
Series([], Name: engine-type, dtype: object)
```

In [35]:

```
df['num-of-cylinders'].unique()
```

Out[35]:

```
array(['four', 'six', 'five', 'twelve', 'two', 'eight'], dtype=object)
```

In [36]:

```
df['fuel-system'].unique()
```

Out[36]:

```
array(['mpfi', '2bbl', 'mfi', '1bbl', 'spfi', '4bbl', 'idi', 'spdi'],
      dtype=object)
```

In [37]:

```
df['fuel-system'].replace({'spdi':'spfi'}, inplace=True)
```

In [38]:

```
df['fuel-system'].unique()
```

Out[38]:

```
array(['mpfi', '2bbl', 'mfi', '1bbl', 'spfi', '4bbl', 'idi'], dtype=object)
```

In [39]:

```
df['bore'].unique()
```

Out[39]:

```
array(['3.47', '2.68', '3.19', '3.13', '3.50', '3.31', '3.62', '3.03',
       '2.97', '3.34', '3.60', '2.91', '2.92', '3.15', '3.43', '3.63',
       '3.54', '3.08', '?', '3.39', '3.76', '3.58', '3.46', '3.80',
       '3.78', '3.17', '3.35', '3.59', '2.99', '3.33', '3.94', '3.74',
       '2.54', '3.05', '3.27', '3.24', '3.01'], dtype=object)
```

In [40]:

```
b = df['bore'] .loc[df['bore']  != '?']
bm=pd.to_numeric(b).mean()
bm
```

Out[40]:

```
3.318395721925134
```

In [41]:

```
df['bore'] = df['bore'] .replace('?',bm)
df['bore'] = pd.to_numeric(df['bore'])
```

In [42]:

```
df['stroke'].unique()
```

Out[42]:

```
array(['2.68', '3.47', '3.40', '2.80', '3.19', '3.39', '3.11', '3.23',
       '3.46', '3.90', '3.41', '3.07', '3.58', '4.17', '2.76', '3.15',
       '?', '3.16', '3.64', '3.10', '3.35', '3.12', '3.86', '3.29',
       '3.27', '2.90', '2.07', '2.36', '2.64', '3.03', '3.08', '3.50',
       '3.54', '2.87'], dtype=object)
```

In [43]:

```
df['horsepower'].loc[df['horsepower'] == '?'].count()
```

Out[43]:

```
2
```

In [44]:

```
df['horsepower'].str.isnumeric().value_counts()
```

Out[44]:

```
True     189
False      2
Name: horsepower, dtype: int64
```

In [45]:

```
df['horsepower'].loc[df['horsepower'].str.isnumeric() == False]
```

Out[45]:

```
130    ?
131    ?
Name: horsepower, dtype: object
```

In [46]:

```
hp = df['horsepower'] .loc[df['horsepower']  != '?']
hpm = hp.astype(str).astype(int).mean()
df['horsepower']  = df['horsepower'] .replace('?',hpm).astype(int)
df['horsepower'].head()
```

Out[46]:

```
0    111
1    111
2    154
3    102
4    115
Name: horsepower, dtype: int32
```

In [47]:

```
df['peak-rpm'].unique()
```

Out[47]:

```
array(['5000', '5500', '5800', '4250', '5400', '4800', '6000', '4750',
       '4200', '4350', '4500', '5200', '5900', '5750', '?', '5250',
       '4900', '4400', '6600', '5100', '5300'], dtype=object)
```

In [48]:

```
df['peak-rpm'].str.isnumeric().value_counts()
```

Out[48]:

```
True     189
False      2
Name: peak-rpm, dtype: int64
```

In [49]:

```
df['peak-rpm'].loc[df['peak-rpm'].str.isnumeric() == False]
```

Out[49]:

```
130    ?
131    ?
Name: peak-rpm, dtype: object
```

In [50]:

```
pr = df['peak-rpm'] .loc[df['peak-rpm']  != '?']
prm = pr.astype(str).astype(int).mean()
df['peak-rpm']  = df['peak-rpm'] .replace('?',prm).astype(int)
df['peak-rpm'].head()
```

Out[50]:

```
0    5000
1    5000
2    5000
3    5500
4    5500
Name: peak-rpm, dtype: int32
```

In [51]:

```
df['city-mpg'].unique()
```

Out[51]:

```
array([21, 19, 24, 18, 17, 16, 23, 20, 15, 38, 37, 31, 49, 30, 27, 25, 13,
       26, 22, 14, 45, 32, 28, 35, 34, 29, 33], dtype=int64)
```

In [52]:

```
df['highway-mpg'].unique()
```

Out[52]:

```
array([27, 26, 30, 22, 25, 20, 29, 28, 43, 41, 38, 24, 54, 42, 34, 33, 31,
       19, 17, 23, 32, 39, 18, 16, 37, 50, 36, 47, 46], dtype=int64)
```

In [53]:

```
df['price'].loc[df['price'] == '?'].count()
```

Out[53]:

```
4
```

In [54]:

```
df['price'].str.isnumeric().value_counts()
```

Out[54]:

```
True     187
False      4
Name: price, dtype: int64
```

In [55]:

```
df['price'].loc[df['price'].str.isnumeric() == False]
```

Out[55]:

```
9      ?
44     ?
45     ?
129    ?
Name: price, dtype: object
```

In [56]:

```
p = df['price'] .loc[df['price']  != '?']
pm = p.astype(str).astype(int).mean()
df['price']  = df['price'] .replace('?',pm).astype(int)
df['price'] .head()
```

Out[56]:

```
0    13495
1    16500
2    16500
3    13950
4    17450
Name: price, dtype: int32
```

In [57]:

```
df.to_csv('automobile_cleaned.csv', index=False)
```

In [ ]:

```

```

In [ ]:

```

```

In [ ]:

```

```