Assignment 1

Data Exploration and Analysis

Semester 2 2023

Student Name: Thuan Nguyen

Student ID: 23194073

PAPER NAME: Data Analysis

PAPER CODE: COMP517

Due Date: Friday 1st Sep 2023 (midnight)

TOTAL MARKS: 100

Contents

Table of figure

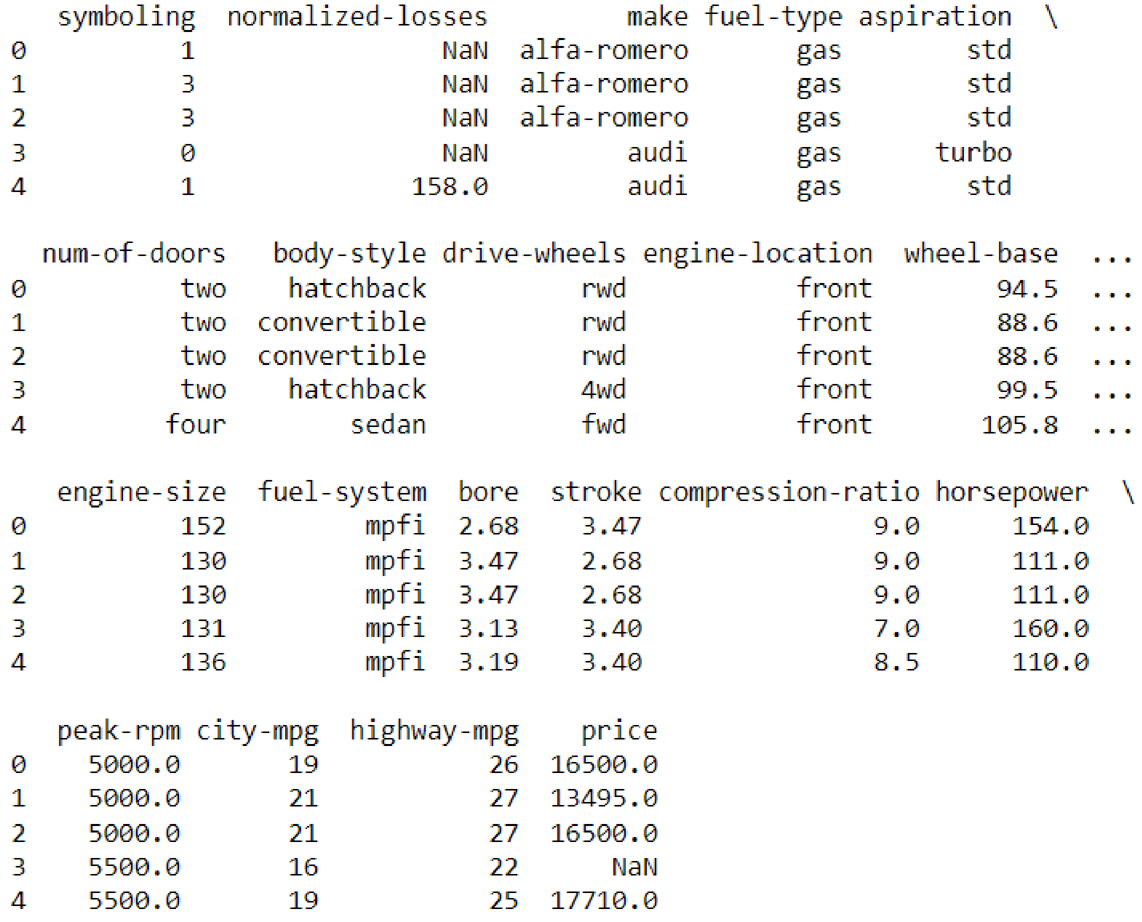
Figure 11. pie chart of 'make.' 14
Figure 12.Heatmap for numerical columns 15
Figure 13.Distribution of price across fuel type and body styles....................16
Figure 14.Histogram for mean and median value of drive-wheel....................17
Figure 15.Boxplot of body styles and price....................18

Task 1: Dataset

Part 1: The purpose and short summary about the dataset.

The purpose of this report is to provide deep insight of the dataset called “Car Dataset.csv”, which could be used for further analysis.

Part 2: Display heading rows of the dataset.



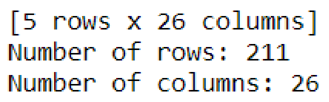
```
   symboling  normalized-losses         make fuel-type aspiration  \
0          1                NaN  alfa-romero       gas        std
1          3                NaN  alfa-romero       gas        std
2          3                NaN  alfa-romero       gas        std
3          0                NaN         audi       gas      turbo
4          1              158.0         audi       gas        std

  num-of-doors   body-style drive-wheels engine-location  wheel-base  ...
0          two    hatchback          rwd           front        94.5  ...
1          two  convertible          rwd           front        88.6  ...
2          two  convertible          rwd           front        88.6  ...
3          two    hatchback          4wd           front        99.5  ...
4         four        sedan          fwd           front       105.8  ...

   engine-size  fuel-system  bore  stroke compression-ratio horsepower  \
0          152         mpfi  2.68    3.47               9.0      154.0
1          130         mpfi  3.47    2.68               9.0      111.0
2          130         mpfi  3.47    2.68               9.0      111.0
3          131         mpfi  3.13    3.40               7.0      160.0
4          136         mpfi  3.19    3.40               8.5      110.0

   peak-rpm city-mpg  highway-mpg    price
0    5000.0       19           26  16500.0
1    5000.0       21           27  13495.0
2    5000.0       21           27  16500.0
3    5500.0       16           22      NaN
4    5500.0       19           25  17710.0
```

Figure 1. first few rows of dataset

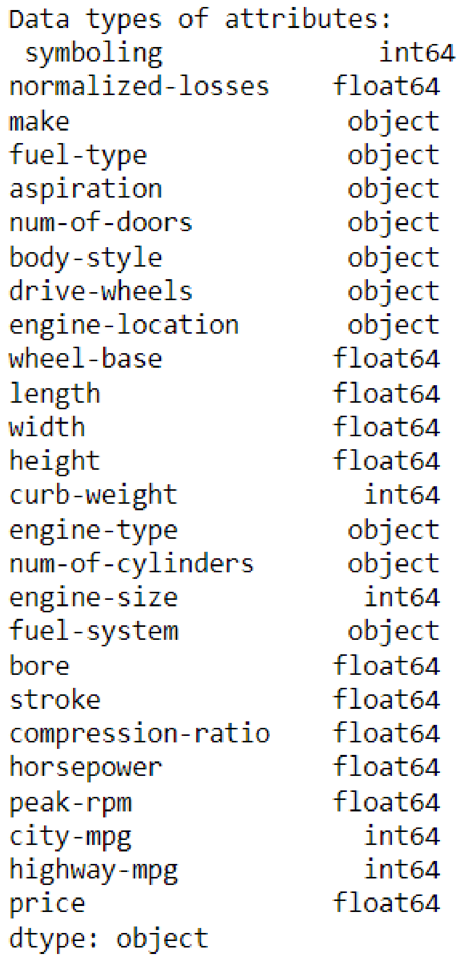
As seen above, the dataset contains car information such as brand, fuel types, body types, number of doors, horsepower, etc. Moreover, the dataset also contains insurance risk ratings, and normalized loss values.

Part 3: Observation:

The dataset contains 211 rows (index) and 26 columns (attributes).



```
[5 rows x 26 columns]
Number of rows: 211
Number of columns: 26
```



```
Data types of attributes:
 symboling                int64
normalized-losses    float64
make                  object
fuel-type             object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight             int64
engine-type           object
num-of-cylinders      object
engine-size             int64
fuel-system           object
bore                 float64
stroke               float64
compression-ratio    float64
horsepower           float64
peak-rpm             float64
city-mpg                int64
highway-mpg             int64
price                float64
dtype: object
```

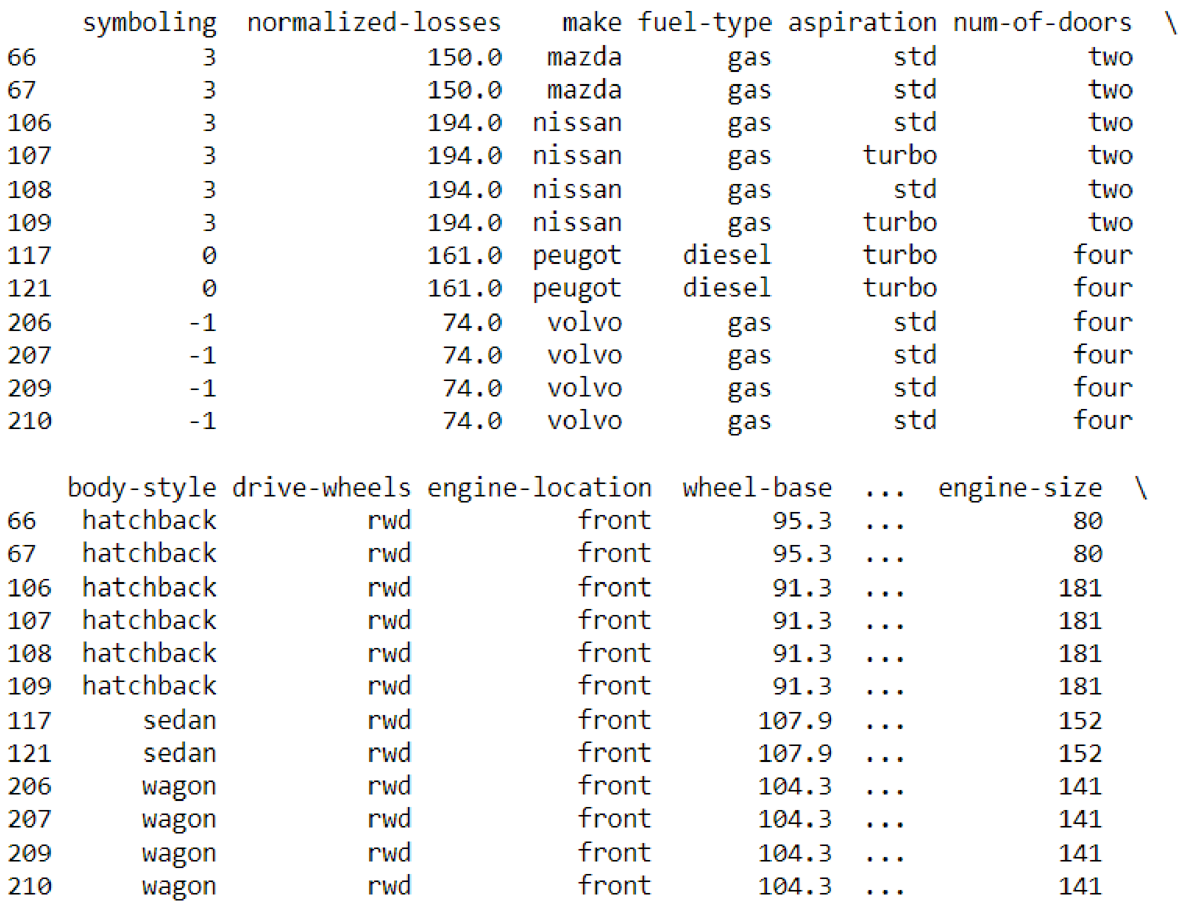
Figure 2.Data type

From figure 2, dataset contains numerical attributes, whole numbers and decimal numbers (int64, float64) and categorical attributes (object).

Task 2: Data Pre-processing

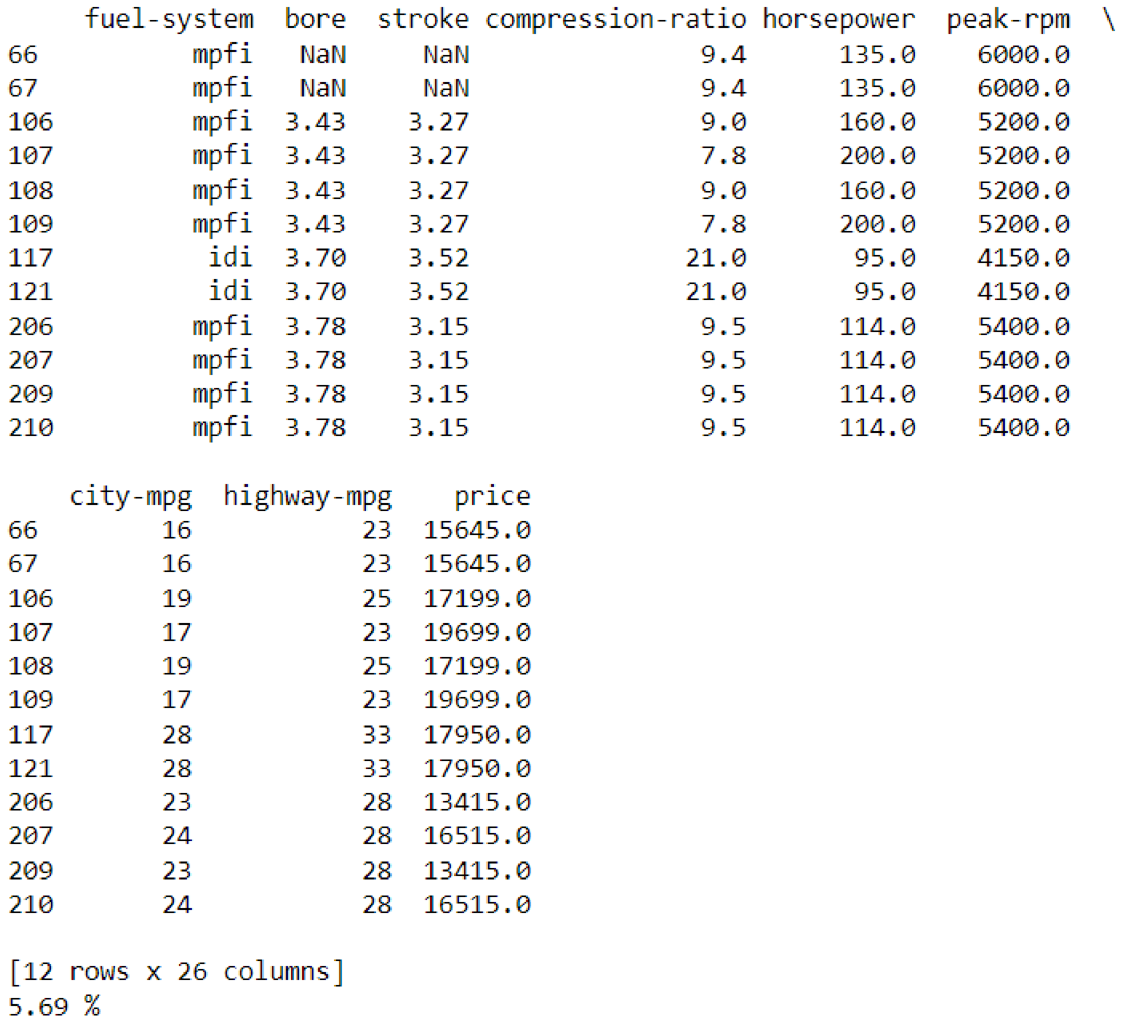
Part 1: Handling duplicates

After processing, the sample of duplicated rows is displayed below:



```
     symboling  normalized-losses    make fuel-type aspiration num-of-doors  \
66           3              150.0   mazda       gas        std          two
67           3              150.0   mazda       gas        std          two
106          3              194.0  nissan       gas        std          two
107          3              194.0  nissan       gas      turbo          two
108          3              194.0  nissan       gas        std          two
109          3              194.0  nissan       gas      turbo          two
117          0              161.0  peugot    diesel      turbo         four
121          0              161.0  peugot    diesel      turbo         four
206         -1               74.0   volvo       gas        std         four
207         -1               74.0   volvo       gas        std         four
209         -1               74.0   volvo       gas        std         four
210         -1               74.0   volvo       gas        std         four

    body-style drive-wheels engine-location  wheel-base  ...  engine-size  \
66   hatchback          rwd           front        95.3  ...           80
67   hatchback          rwd           front        95.3  ...           80
106  hatchback          rwd           front        91.3  ...          181
107  hatchback          rwd           front        91.3  ...          181
108  hatchback          rwd           front        91.3  ...          181
109  hatchback          rwd           front        91.3  ...          181
117      sedan          rwd           front       107.9  ...          152
121      sedan          rwd           front       107.9  ...          152
206      wagon          rwd           front       104.3  ...          141
207      wagon          rwd           front       104.3  ...          141
209      wagon          rwd           front       104.3  ...          141
210      wagon          rwd           front       104.3  ...          141
```

Figure 3. duplicate rows



```
     fuel-system  bore  stroke compression-ratio horsepower  peak-rpm  \
66          mpfi   NaN     NaN               9.4      135.0    6000.0
67          mpfi   NaN     NaN               9.4      135.0    6000.0
106         mpfi  3.43    3.27               9.0      160.0    5200.0
107         mpfi  3.43    3.27               7.8      200.0    5200.0
108         mpfi  3.43    3.27               9.0      160.0    5200.0
109         mpfi  3.43    3.27               7.8      200.0    5200.0
117          idi  3.70    3.52              21.0       95.0    4150.0
121          idi  3.70    3.52              21.0       95.0    4150.0
206         mpfi  3.78    3.15               9.5      114.0    5400.0
207         mpfi  3.78    3.15               9.5      114.0    5400.0
209         mpfi  3.78    3.15               9.5      114.0    5400.0
210         mpfi  3.78    3.15               9.5      114.0    5400.0

    city-mpg  highway-mpg   price
66        16           23  15645.0
67        16           23  15645.0
106       19           25  17199.0
107       17           23  19699.0
108       19           25  17199.0
109       17           23  19699.0
117       28           33  17950.0
121       28           33  17950.0
206       23           28  13415.0
207       24           28  16515.0
209       23           28  13415.0
210       24           28  16515.0

[12 rows x 26 columns]
5.69 %
```

Figure 4. duplicated rows

As can see from figure 3 and figure 4, first value is original value, so there are 6 rows repeated, which is considered as a duplicate.

In general, duplicated rows can skew analysis results, especially for visualizing distributions. In another hand, for this circumstance, the number of repeated rows is not too much. It might be caused by computing mistake, noticed one value many times. Since there is no indication that the duplicates are intentional or represent different observations, removing those duplicates rows is necessary for further analysis.

After removing, the dataset now contains 205 rows and 26 columns.

Part 2: Handling outliers:

Visualization method:

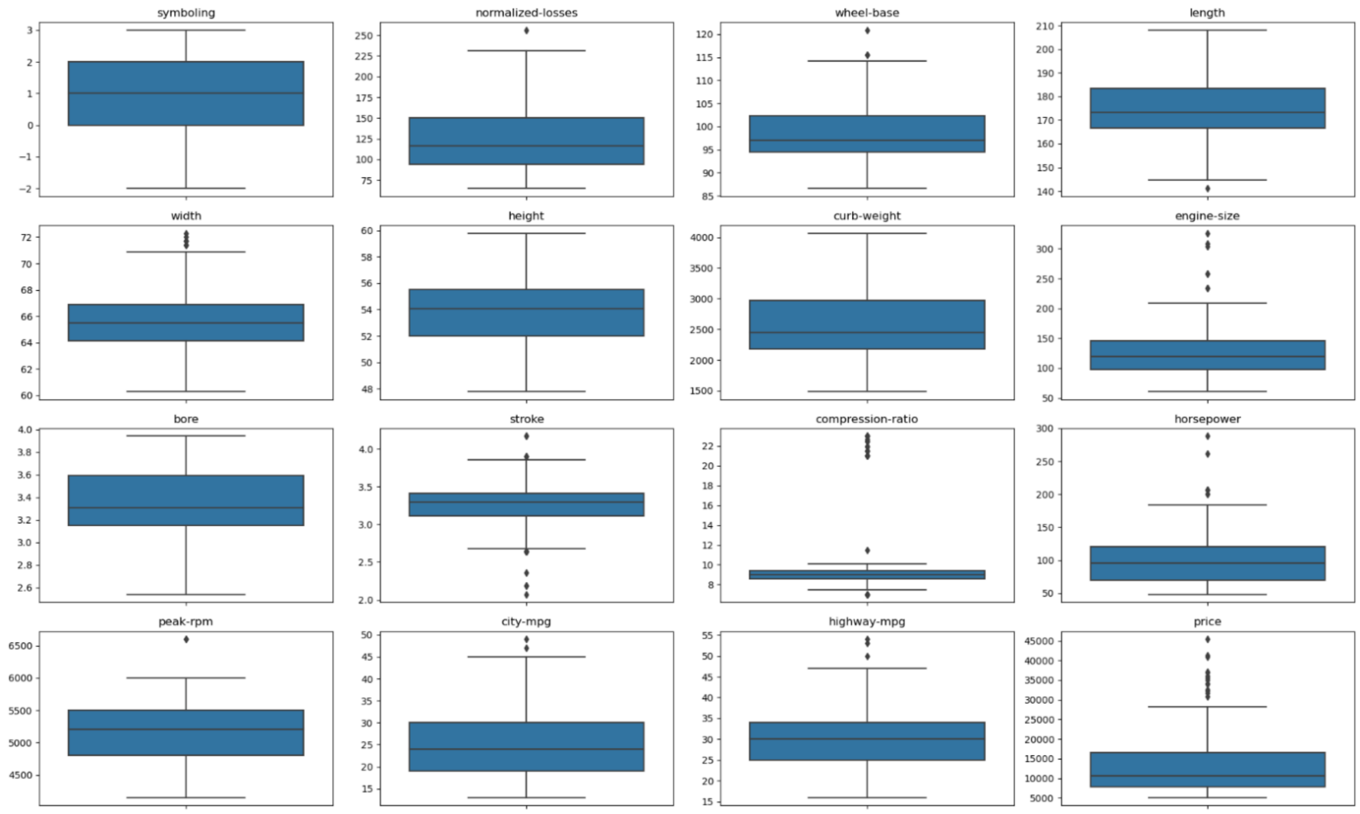
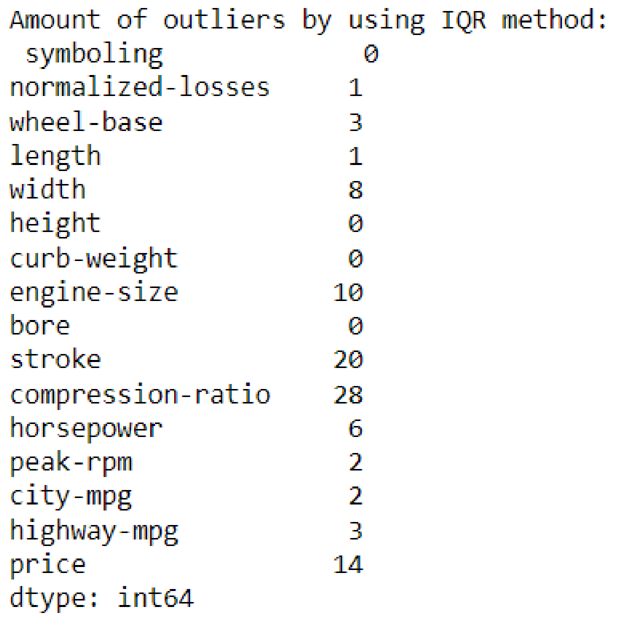


Figure 5. boxplot for numerical data

From the plots (figure 5), easily observe that attributes like normalized losses, wheelbase, length, width, engine size, stroke, compression ratio, horsepower, peak-rpm, city-mpg, highway-mpg, price, appear to have potential outliers. Additionally, compression ratio, engine size has more noticeable potential outliers (the outliers of those attributes are far away than other data point).

Furthermore, most of the attributes of the dataset appear to be skewed and non-normally distributed. Thus, using the IQR method is more prudent to calculate and reinforce affirmation.

After using the IQR method to calculate, here is the result:



```
Amount of outliers by using IQR method:
 symboling             0
normalized-losses     1
wheel-base            3
length                1
width                 8
height                0
curb-weight           0
engine-size          10
bore                  0
stroke               20
compression-ratio    28
horsepower            6
peak-rpm              2
city-mpg              2
highway-mpg           3
price                14
dtype: int64
```

Similar results arise after the comparison of the two methods. There are 12 attributes which contain outliers: normalize-losses, wheelbase, length, width, engine-size, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg, and price.

When visualizing the data for the attributes with outliers by using scatterplot, it is easier to understand the context and relationship of those attributes. It also provides a better view to identify the potential impact of these outliers as well as some valuable information it might contain, which helps to find suitable way of handling outlier of the data.

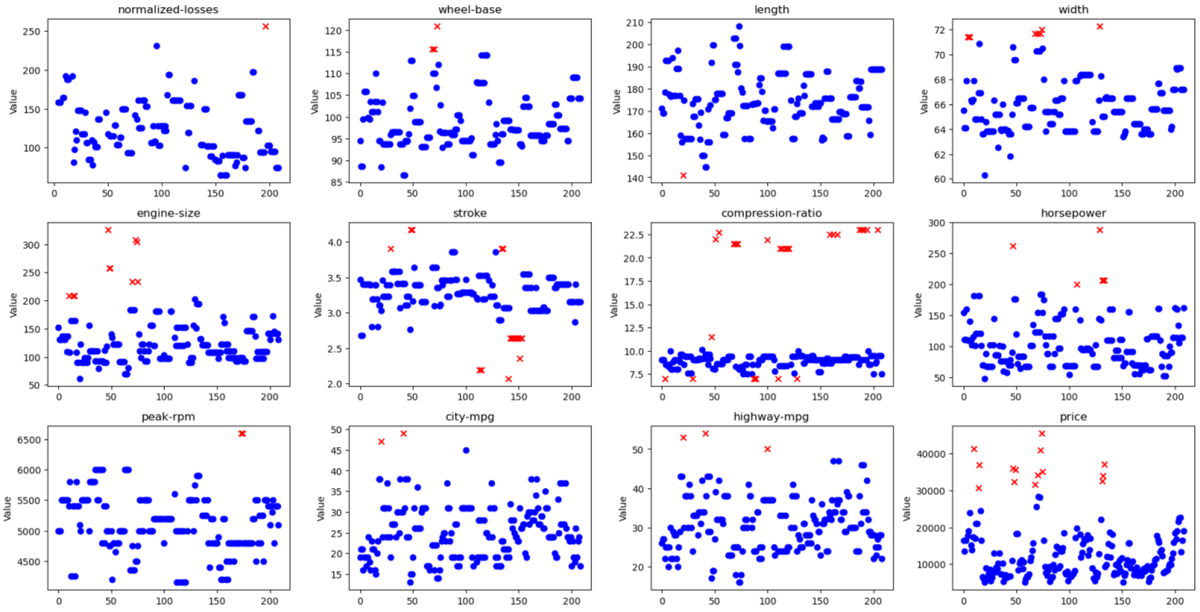


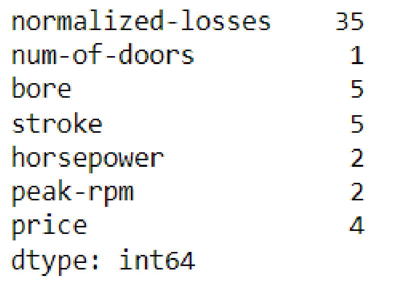
Figure 6. Scatterplot for attributes contain outliers.

Generally, outliers can make a big impact on statistical analyses and skew the distribution, and some extreme values can impact statistical power, which could make the insight become biased. After considering based on scatterplot that show above, it is noticed that:

- Those attributes such as ‘normalized-losses’, ‘wheelbase’, ‘length’, ‘width’, ’stroke’, ‘peak-rpm’, ‘city-mpg’ and ‘highway-mpg’ have outliers that do not seem to be extreme values (not far apart from other data points)., which mean it would not affect distribution seriously. So the solution is keeping those outliers.
- For attributes such as ‘engine-size’, ‘compression-ratio’, ‘horsepower’, and ‘price’ have outliers can be noticed as extreme values, but those attributes present actual data. Furthermore, higher compression ratios are often found in diesel engines, which represent for efficiency and power of diesel engines. Additionally,cars with higher price and larger engines size or horsepower are generally luxuries or sport cars, so removing them could affect badly to reality insight. Therefore, for these attributes, suitable handling method is tranforming to ‘log10’, which means that decreasing the value of outliers to make it less skew the distribution, but still keeping the price as it is because price reflect market segmentation.

.

Part 3: Handling missing values:



```
normalized-losses    35
num-of-doors          1
bore                  5
stroke                5
horsepower            2
peak-rpm              2
price                 4
dtype: int64
```

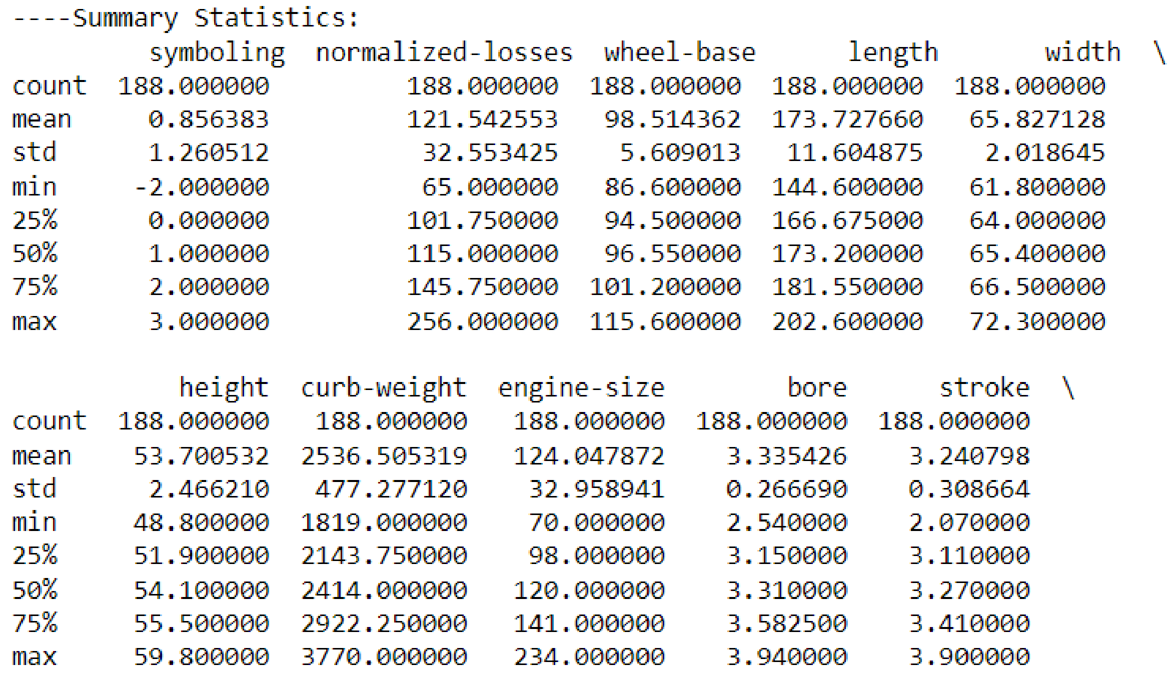
Figure 7. Attributes with missing value

From figure 7, there are two types of data that above attributes contain missing values, object, and number. Therefore, the solution for handling these missing values is:

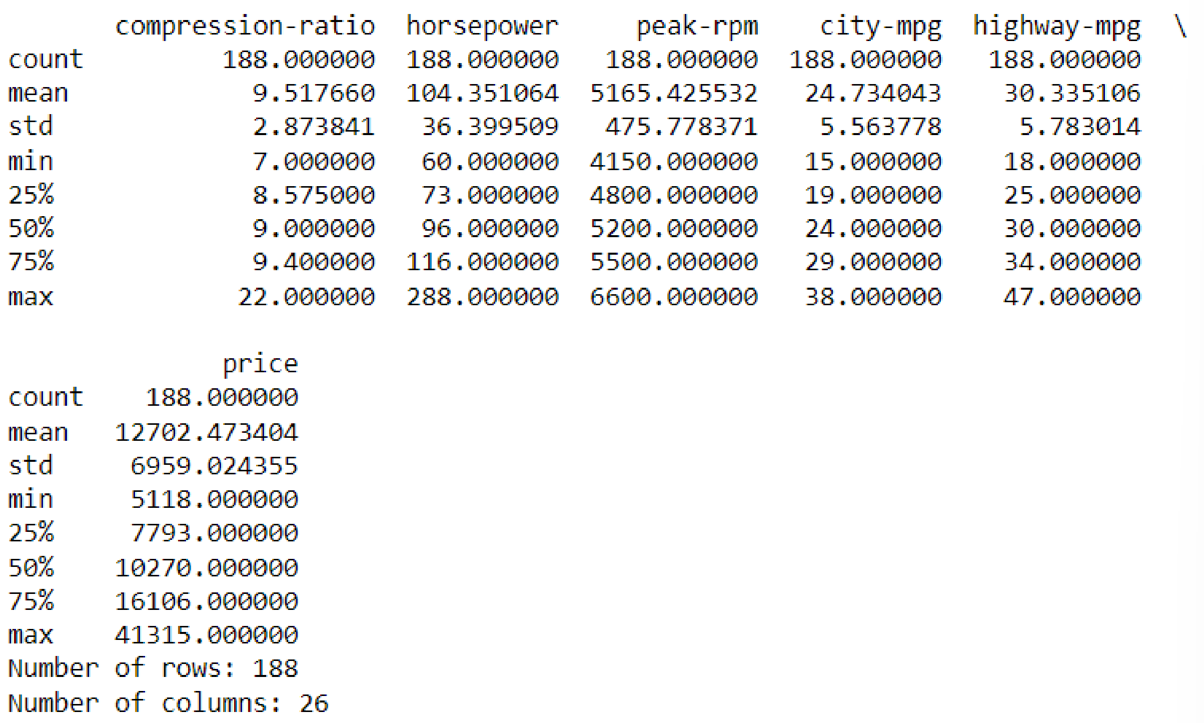
- For numerical attributes, impute them with median, because median is less sensitive to the outliers.
- For categorical attributes, impute them with 'Unknown'.

Task 3: Explore and Visualize the Clean Dataset

Part 1: Calculate basic summary statistics for numerical columns and explain the finding.



```
----Summary Statistics:
         symboling  normalized-losses  wheel-base      length       width  \
count   188.000000         188.000000  188.000000  188.000000  188.000000
mean      0.856383         121.542553   98.514362  173.727660   65.827128
std       1.260512          32.553425    5.609013   11.604875    2.018645
min      -2.000000          65.000000   86.600000  144.600000   61.800000
25%       0.000000         101.750000   94.500000  166.675000   64.000000
50%       1.000000         115.000000   96.550000  173.200000   65.400000
75%       2.000000         145.750000  101.200000  181.550000   66.500000
max       3.000000         256.000000  115.600000  202.600000   72.300000

            height  curb-weight  engine-size        bore      stroke  \
count   188.000000   188.000000   188.000000  188.000000  188.000000
mean     53.700532  2536.505319   124.047872    3.335426    3.240798
std       2.466210   477.277120    32.958941    0.266690    0.308664
min      48.800000  1819.000000    70.000000    2.540000    2.070000
25%      51.900000  2143.750000    98.000000    3.150000    3.110000
50%      54.100000  2414.000000   120.000000    3.310000    3.270000
75%      55.500000  2922.250000   141.000000    3.582500    3.410000
max      59.800000  3770.000000   234.000000    3.940000    3.900000
```



```
       compression-ratio  horsepower     peak-rpm    city-mpg  highway-mpg  \
count         188.000000  188.000000   188.000000  188.000000   188.000000
mean            9.517660  104.351064  5165.425532   24.734043    30.335106
std             2.873841   36.399509   475.778371    5.563778     5.783014
min             7.000000   60.000000  4150.000000   15.000000    18.000000
25%             8.575000   73.000000  4800.000000   19.000000    25.000000
50%             9.000000   96.000000  5200.000000   24.000000    30.000000
75%             9.400000  116.000000  5500.000000   29.000000    34.000000
max            22.000000  288.000000  6600.000000   38.000000    47.000000

              price
count    188.000000
mean   12702.473404
std     6959.024355
min     5118.000000
25%     7793.000000
50%    10270.000000
75%    16106.000000
max    41315.000000
Number of rows: 188
Number of columns: 26
```

Figure 8. Summary statistic of numerical data

Figure 8 provides information about basic summary statistics, including the highest and the lowest value, the range where the value located, average value as well as the median value.

Part 2: Visualization data

Visualization method to understand the distribution of numerical data.

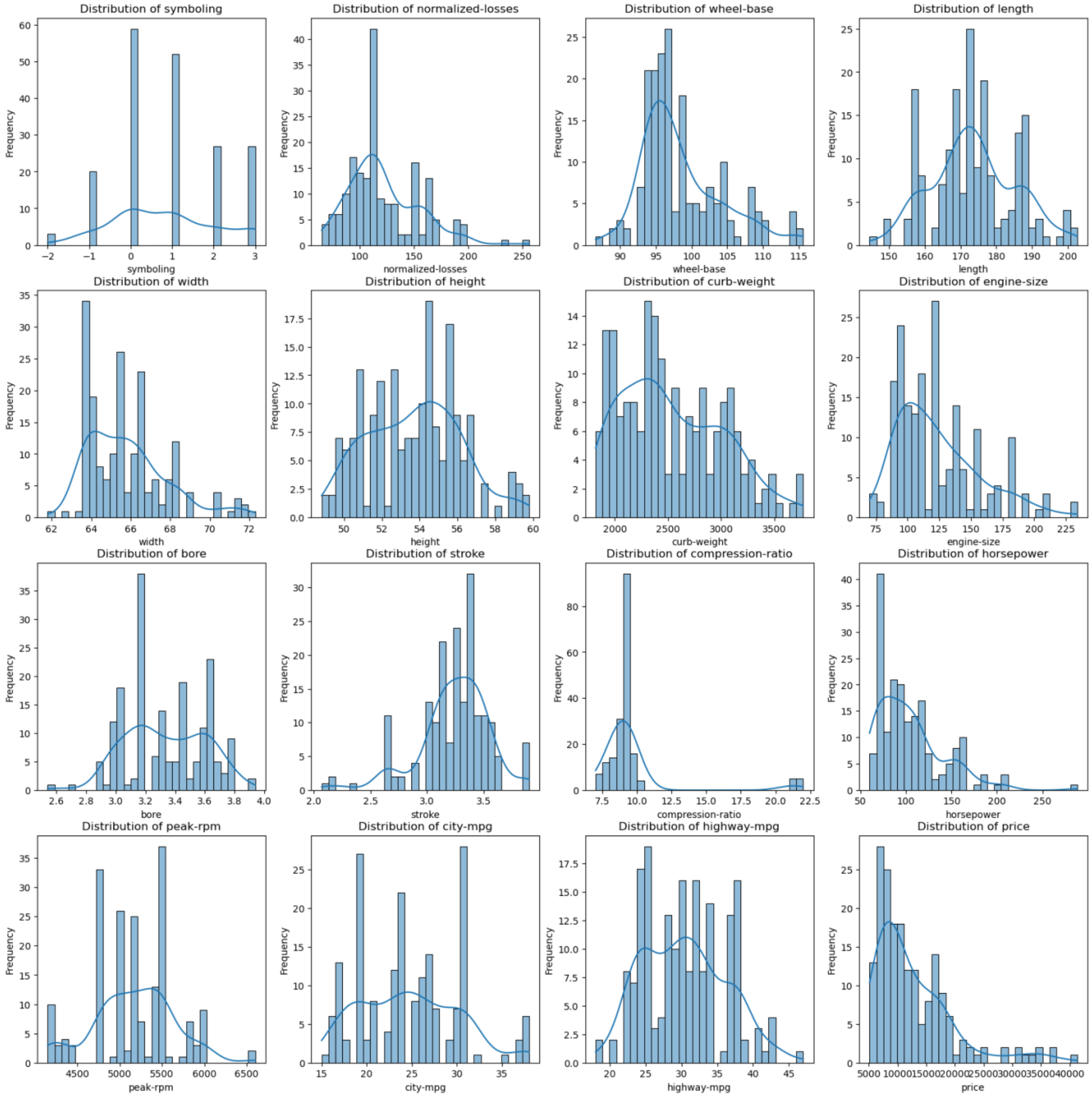


Figure 9. Histogram numerical data

As being seen above, the histograms generally provide insight into the distribution of the numerical attributes:

- ‘symbolling’ is discrete data, most of the values around 0 to 2, indicating that they are moderately risky. The distribution is likely to spread equally, there are some values occurring more than others but not so distinctive.
- Attributes that giving point of view about market segmentation, such as price, which distribution is clearly right skewed, indicate that most cars in the dataset are low to moderated and car dimensions, such as wheelbase, length, width, and height, show varied distribution, which means that there is no fixed shape for cars. In more details, ‘length’ has normal distribution, ‘height’ has “uniform” shape while distribution of wheelbase and width slightly skewed to the right.

- Attributes such as ‘engine-size’, ‘horsepower’, ‘compression-ratio’, indicate the performance of the car. In more details, Engine size of most cars in the dataset are between 80 and 125, with the distribution being right-skewed. Most cars have compression ratio around 8 and 10 and horsepower index around 60 to 120.
- Attributes for understanding fuel efficiency, City-mpg and Highway-mpg: both these attributes are slightly left-skewed, indicating that a majority of cars in the dataset have moderate to high fuel consumption. The distribution of ‘Curb-weight’ is slightly right-skewed, indicating that there are more cars with lower curb weight.

Visualization method to understand the distribution of categorical data.

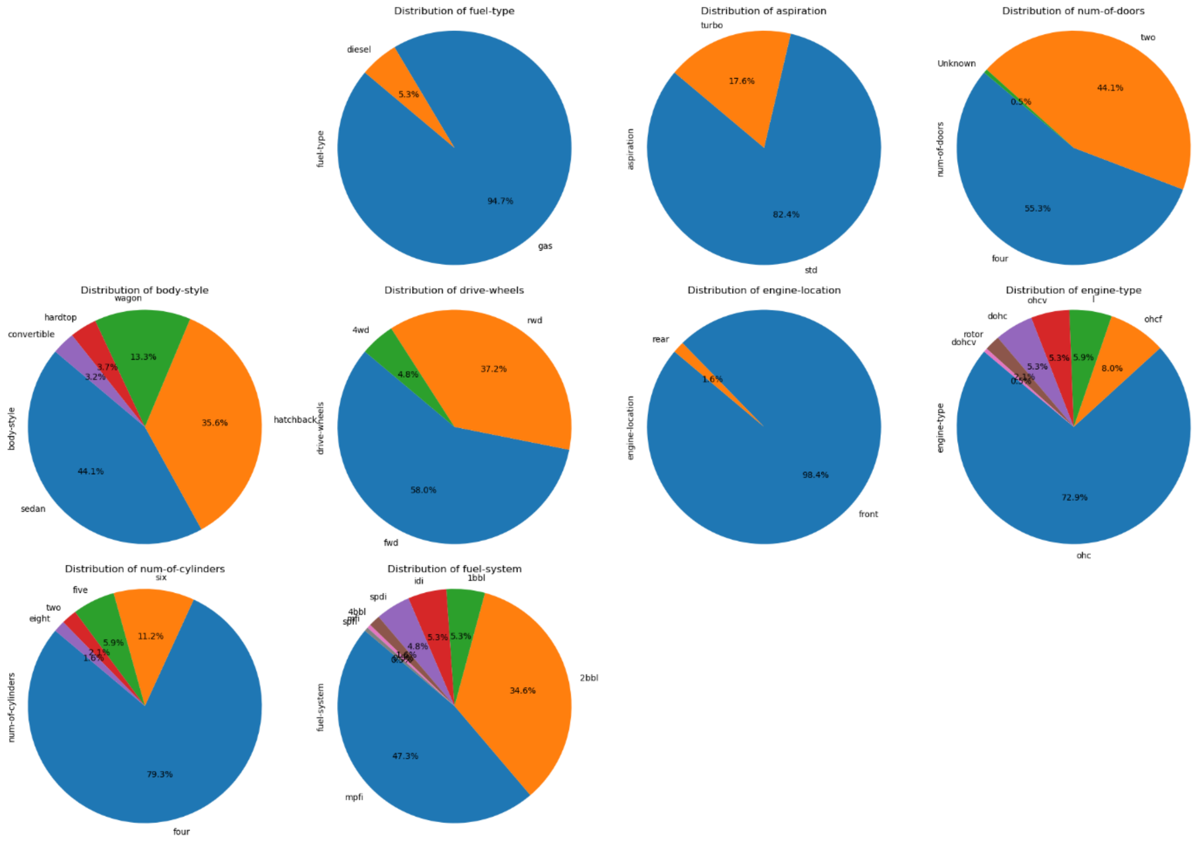


Figure 10. pie charts of categorical data

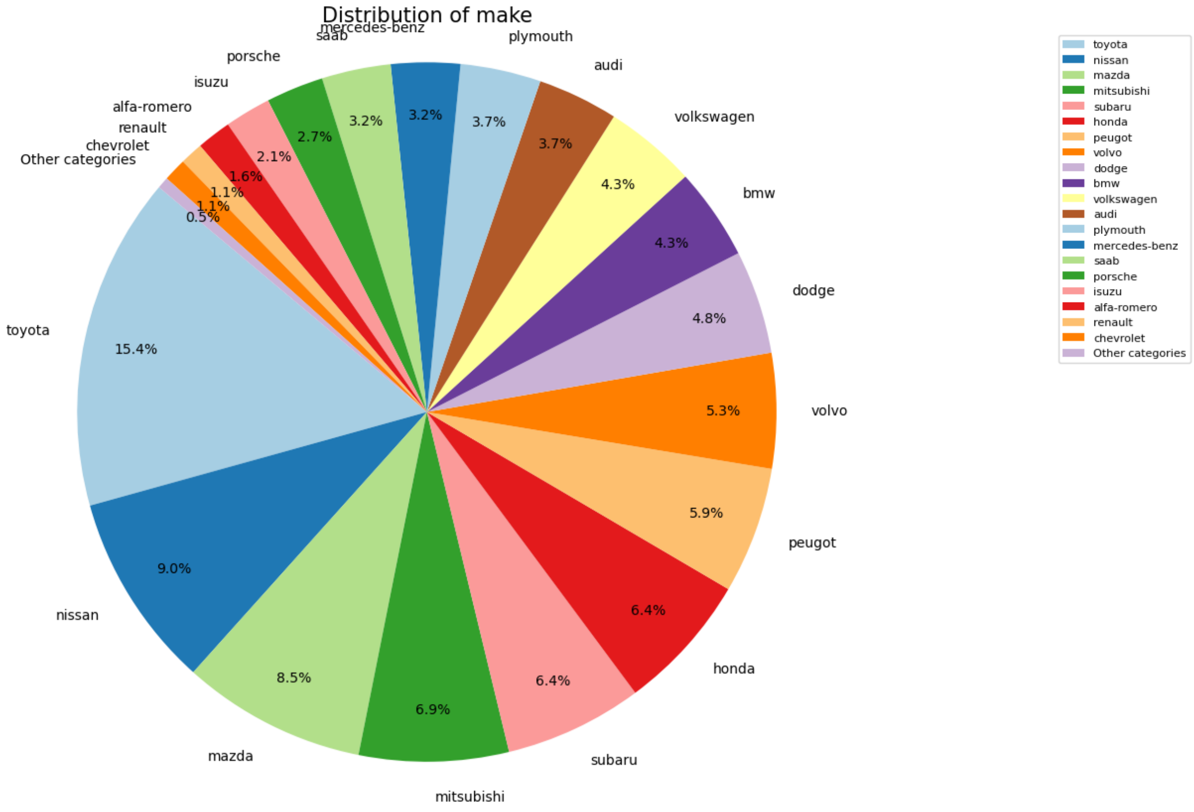


Figure 11. pie chart of 'make.'

Generally, when comparing different attributes from figure 10, there is only one element(blue) that is superior to the rest in terms of proportionality.

- A vast majority of cars use gasoline, with only a small percentage using diesel.
- Most cats have a standard aspiration, with a small portion having turbo.
- Four-door cars are more common than two-door cars in the dataset.
- The most common body styles are sedan and hatchback.
- Front-wheel drive (fwd) is the most common followed by rear-wheel drive (rwd).
- Almost all cats have their engines located in the front, the percentage of cars that have their engines located in the back is trivial.
- ‘ohc’ makes up the majority of engine styles, the rest area is equally divided by left elements.
- An attribute called ‘num- of cylinder’, cars which have four cylinders take the main part of the dataset.
- For fuel systems, ‘mpfi’ and ‘2bbl’ are two of the dominant parts of the dataset.

From figure 11, ‘make’ attributes: the distinction between elements is not excessive. The dataset contains cars form from a variety of manufacturers, with Toyota being the most prominent.

Task 4: Multivariate Analysis

Part 1: Correlation Analysis

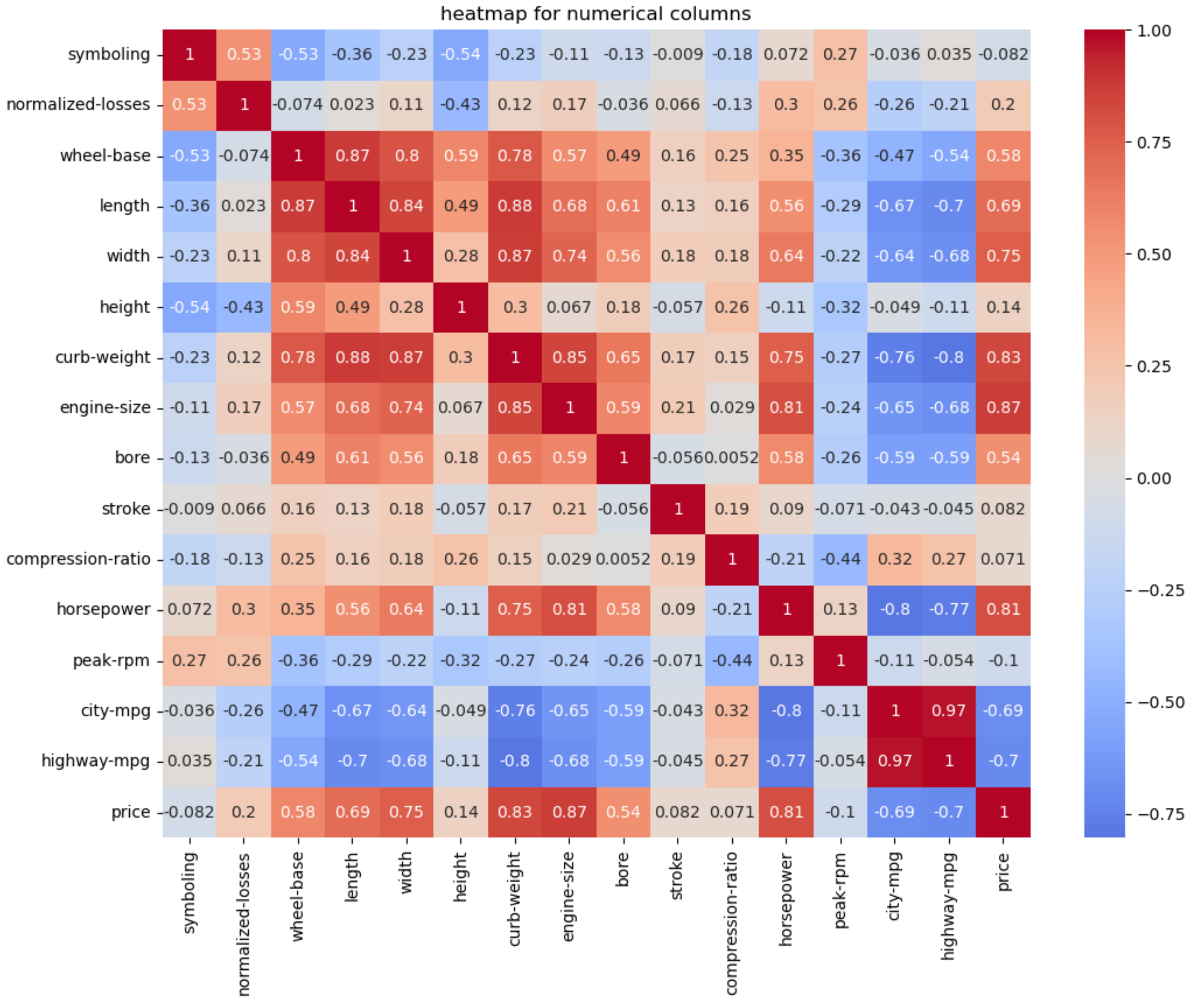


Figure 12.Heatmap for numerical columns

The correlation heatmap (figure 12) provides insights into the relationship between numerical attributes, there are some correlations need to be focused on:

Positive Correlations:

- ‘Engine-size’ has a strong positive correlation with curb-weight (0.87) and horsepower (0.81), implying that cars with larger engines tend to be heavier and stronger.
- City-mpg and highway-mpg are strongly correlated (0.98), which is expected since cars with good city mileage generally have good highway mileage.
- Length, width, and curb-weight are positively correlated, suggesting that larger cars (in terms of length and width) tend to be heavier.

Negatively correlations:

- ‘Engine-size’ is negatively correlated with ‘city-mpg’ (-0.68) and highway-mpg (-0.71). This means that cars with larger engines tend to have lower fuel efficiency.
- ‘Curb-weight’ also has a negative correlation with city-mpg (-0.79) and highway-mpg (-0.79), implying that heavier cars are generally less fuel-efficient.

Weak or no correlations:

- ‘Compression-ratio’ has weak or no significant correlations with other attributes in the dataset. It has strongest negative relationship with ‘peak-rpm’ and no correlations with others (from -0.21 to 0.32).
- ‘symbolling’ has moderate correlations (0.53) with attributes like normalize-losses,
- ‘Peak-rpm’ has slightly equal correlation with other attributes related to car’s performance characteristics. (Around 0.26, 0.27).

Part 2: Perform multivariate analysis of data.

Categorical variables chosen for this analysis:

- ‘Fuel-type’: natural resources, could become rare due to objective reasons regarding to exploit, import, export, government policies about emission lead to the affection of market demand which affect directly on the average price varies between gasoline and diesel cars.
- ‘Body-styles’: Different body styles might be priced differently, and certain body styles might be priced differently more commonly associated with specific types of fuel.

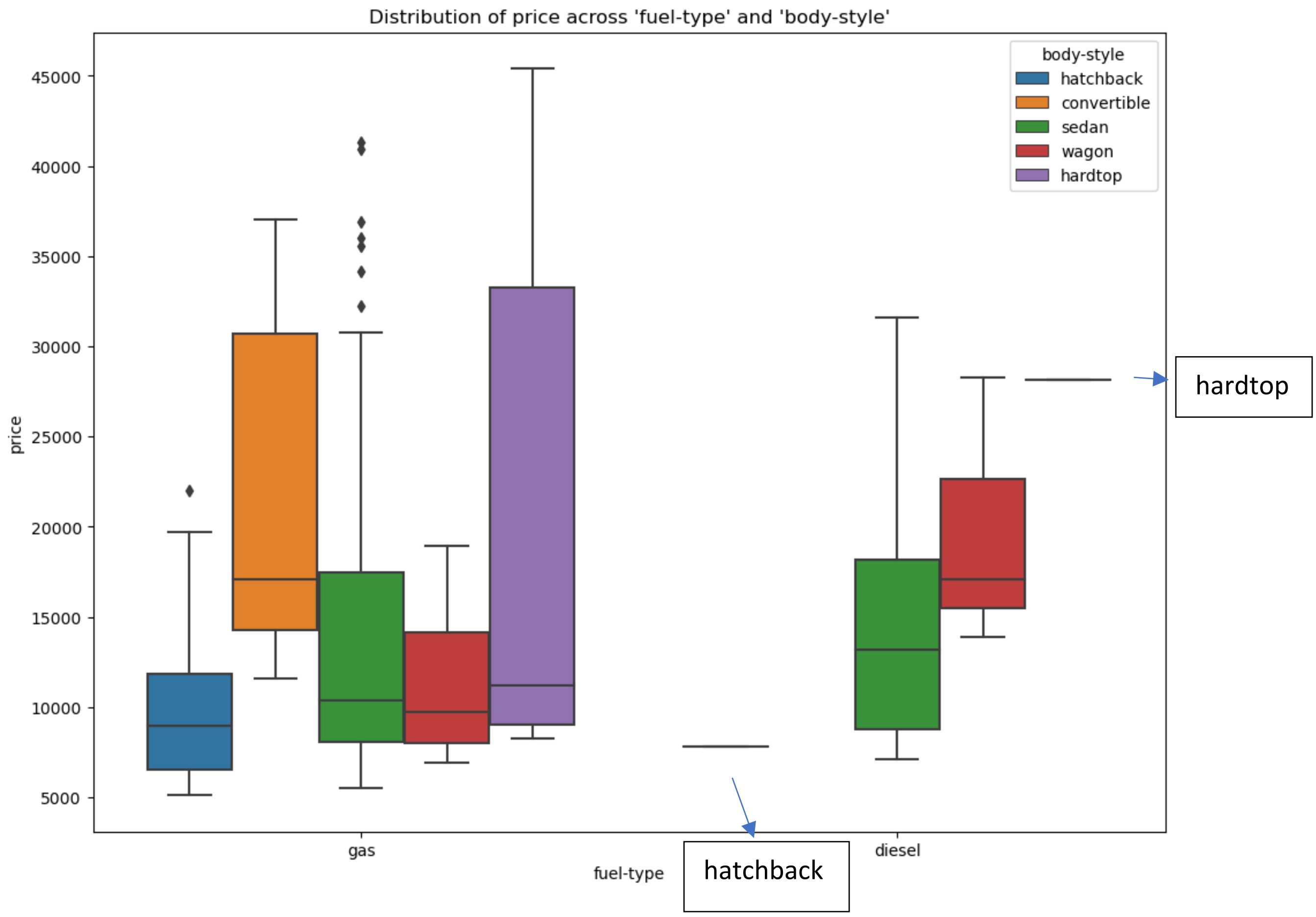


Figure 13.Distribution of price across fuel type and body styles

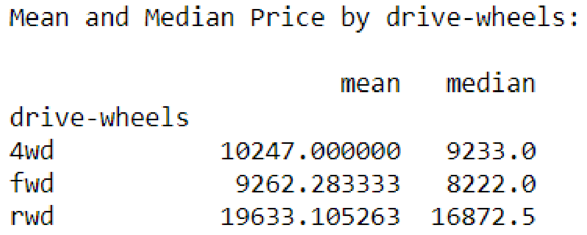
According to figure 13:

- Gasoline cars which have its body style is sedan, have the most outliers compared to others while the median price just in middle among the styles and lower than the median price of those using diesel.

- There is no convertible car that uses diesel fuel, but it has the second widest range of convertibles car using gasoline.
- The wagon has the smallest car range for cars using gasoline as well as diesel fuel. Furthermore, its median price for cars using diesel fuel is the highest even just medium for those using gasoline.
- The hardtop has the widest car range for cars using gasoline with the moderate median price among the dataset, which indicates that it has different segment from lower-price car to luxury car, but the majority is lower to moderate car price because the distribution skew to the right. In contrast, there is only one car with hardtop style using diesel, as present by only one ‘line’ from the figure above.
- The hatchback which uses gasoline has the lowest median price but seems to be priced higher due to the outlier, but like hardtop, only one hatchback car using diesel.

Overall, the price of a car is not only affected by the body style and fuel type but also the combination of those two.

Part 3: Perform aggregate analysis.



```
Mean and Median Price by drive-wheels:

                      mean   median
drive-wheels
4wd           10247.000000   9233.0
fwd            9262.283333   8222.0
rwd           19633.105263  16872.5
```

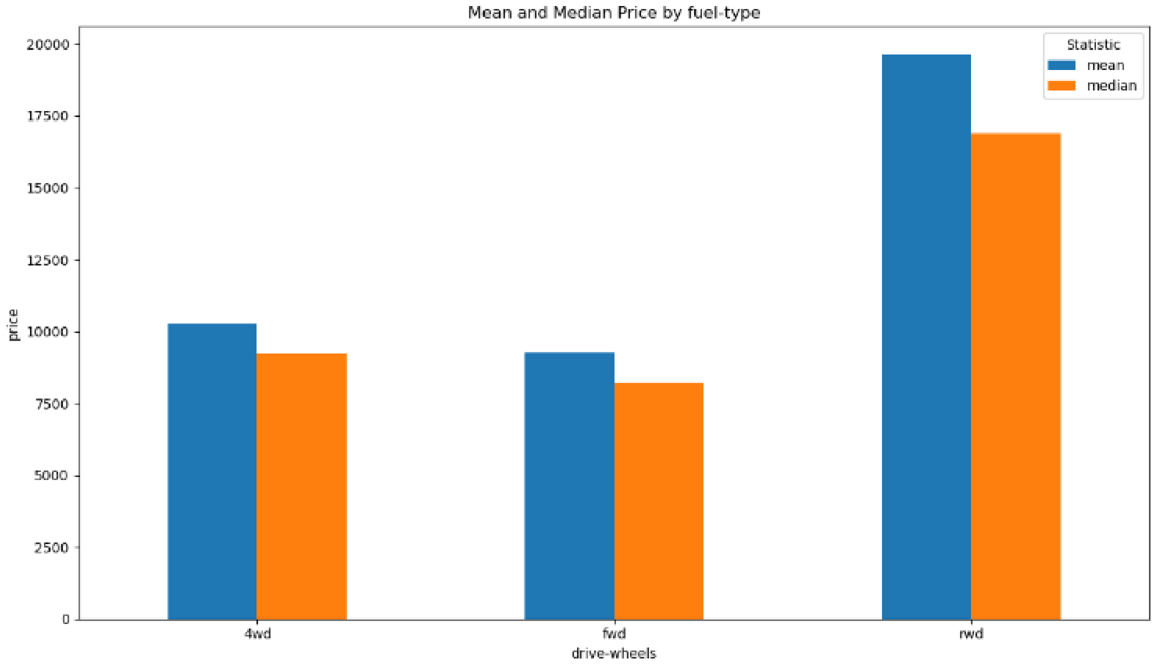


Figure 14.Histogram for mean and median value of drive-wheel

From figure 14 above, it is observable that:

- Four-wheel drive (4wd): has a mean (approximately 10000) and median price (approximately 9000) that is in the mid-range among the drive types.
- Front-wheel drive (fwd): has the lowest mean (approximately 9000) and median price (approximately 8000)

- Rear-wheel drive (rwd): has the highest mean (almost 20000) and median price (around 17000), nearly as twice as four-wheel and front-wheel drive.

In conclusion, front-wheel drive cars in the dataset are easier to afford while rear-wheel drive cars are priced higher than their counterparts.

Part 4: Perform analysis between ‘body-styles’ and ‘price.’

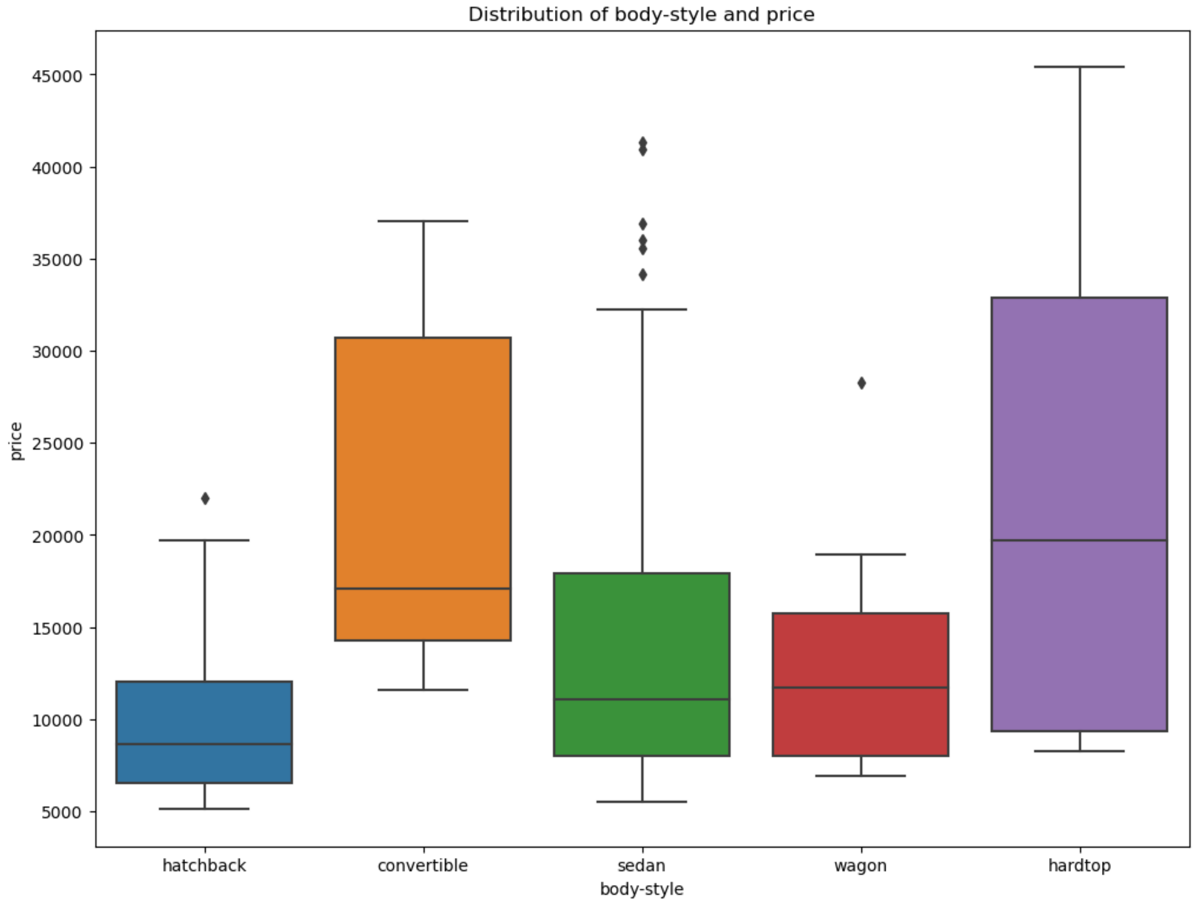


Figure 15.Boxplot of body styles and price

As can be seen from the boxplot above, it is observable that:

- Hatchback: primarily clustered in the lower price range, it has the lowest median price among the body styles. This body style contains outlier but not priced much higher.
- Sedan: has the second highest price range, with the most outliers compared to others.
- Convertibles: have a higher median price compared to others except hardtops, indicating that they are priced higher.
- Wagon: has the smallest price range, but it has an outlier which are priced much higher.
- Hardtop: has the highest median price and the largest price range.

Significances of performing such analysis:

1) Predict price and budget for the buyers:

If the consumer has insight into median price and price range of each body styles, they could have clearer visualization to make better decision. Additionally, if they are interested in certain body styles, they could have prediction about how much they should prepare for in advance. Furthermore, the wider the price range is, the more prices for consumer to choose. For instance, if the consumer is interested in hardtops, which might be more expensive than other body types because they have the highest median price, they could know how much they should budget, and how many choices they have based on the diversity of price.

2) Insights and market segmentation for producers.

Producers can efficiently divide the market and place their products depending on price points. Moreover, it can be the reference engineering to customize the product based on buyers' reference. For instance, if the companies want to focus on low for moderate buyers, they should produce hatchback and wagon, and for premium buyers they should produce hardtops.

Task 5: Conclusion

Part 1: Summary of the key findings and insights

Dataset Overview:

- The dataset provided insights into various attributes related to cars, including their brand, specifications, and pricing. Moreover, dataset provides information on risky level as well as value loss.
- The dataset is raw data, so it required pre-processing to handle missing values, outliers, and duplicate to obtain important information.

Price:

- The price of cars varies based on multiple factors. For instance, convertibles and hardtops tend to be priced higher, while hatchbacks are more affordable.
- Rear-wheel drive cars are typically more expensive compared to front-wheel drive vehicles.

Correlation insights

- Cars with larger engines or stronger tend to be less fuel efficient, as observed from the negative correlation between engine size, horsepower and miles-per-gallon attributes.
- Some attributes, such as engine size and curb weight, show strong correlation, indicating they depend on each other, and this can be useful for engineering and design considerations.
- The body styles and fuel type of cars significantly influence their pricing. This information is vital for manufacturers to position their products in the market effectively.

Part 2: Challenge faced during the analysis.

- Firstly, the dataset contains missing values represented as '?' instead of 'Nan', which required careful handling. Those problems required a deep understanding about the function with its parameters used in the code.
- Secondly, about handling outliers, there are different solutions for each outlier in certain attributes, which is required to have domain knowledge about cars and how certain attributes affect the whole dataset, needed consideration to ensure they didn't skew the analysis.
- Thirdly, needed to know how to use graphs for analyzing and visualizing specific attributes of the dataset appropriately.

Part 3: Further analysis suggestion.

- Deeper analysis about ‘symbolling’ as known as risk level and other attributes to identify which characteristics contribute to higher or lower risk. So that they could adjust the insurance price for specific car types.
- Add more data related to the market, such as customer reviews, brand reputation, resell price after N years. This additional data could provide more comprehensive insights into factors influencing car pricing. Information about features and technologies can directly affect the price.
- create new attributes based on existing attributes. For example, create attributes called ‘car-size’ based on existing attributes such as ‘height’, ‘length’, ‘width’.