Automobile Dataset Analysis

By

Chethan K

Data overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   symboling          205 non-null    int64  
 1   normalized-losses  205 non-null    object 
 2   make               205 non-null    object 
 3   fuel-type          205 non-null    object 
 4   aspiration         205 non-null    object 
 5   num-of-doors       205 non-null    object 
 6   body-style         205 non-null    object 
 7   drive-wheels       205 non-null    object 
 8   engine-location    205 non-null    object 
 9   wheel-base         205 non-null    float64
 10  length             205 non-null    float64
 11  width              205 non-null    float64
 12  height             205 non-null    float64
 13  curb-weight        205 non-null    int64  
 14  engine-type        205 non-null    object 
 15  num-of-cylinders   205 non-null    object 
 16  engine-size        205 non-null    int64  
 17  fuel-system        205 non-null    object 
 18  bore               205 non-null    object 
 19  stroke             205 non-null    object 
 20  compression-ratio  205 non-null    float64
 21  horsepower         205 non-null    object 
 22  peak-rpm           205 non-null    object 
 23  city-mpg           205 non-null    int64  
 24  highway-mpg        205 non-null    int64  
 25  price              205 non-null    object 
dtypes: float64(5), int64(5), object(16)
memory usage: 41.8+ KB
```

This dataset consist of data From 1985 Ward's Automotive Yearbook from bellow sources:

- 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.
- Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038
- Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

There are 26 columns and 205 rows, with some missing values.

Target : Price

Categorical features : 16

Numerical features : 10

Initial plan for data exploration

- Check for missing value and treatment that must be done
- Perform feature engineering on data (if needed)
- Perform EDA with visualization
- Perform hypothesis analysis on the dataset

Missing value

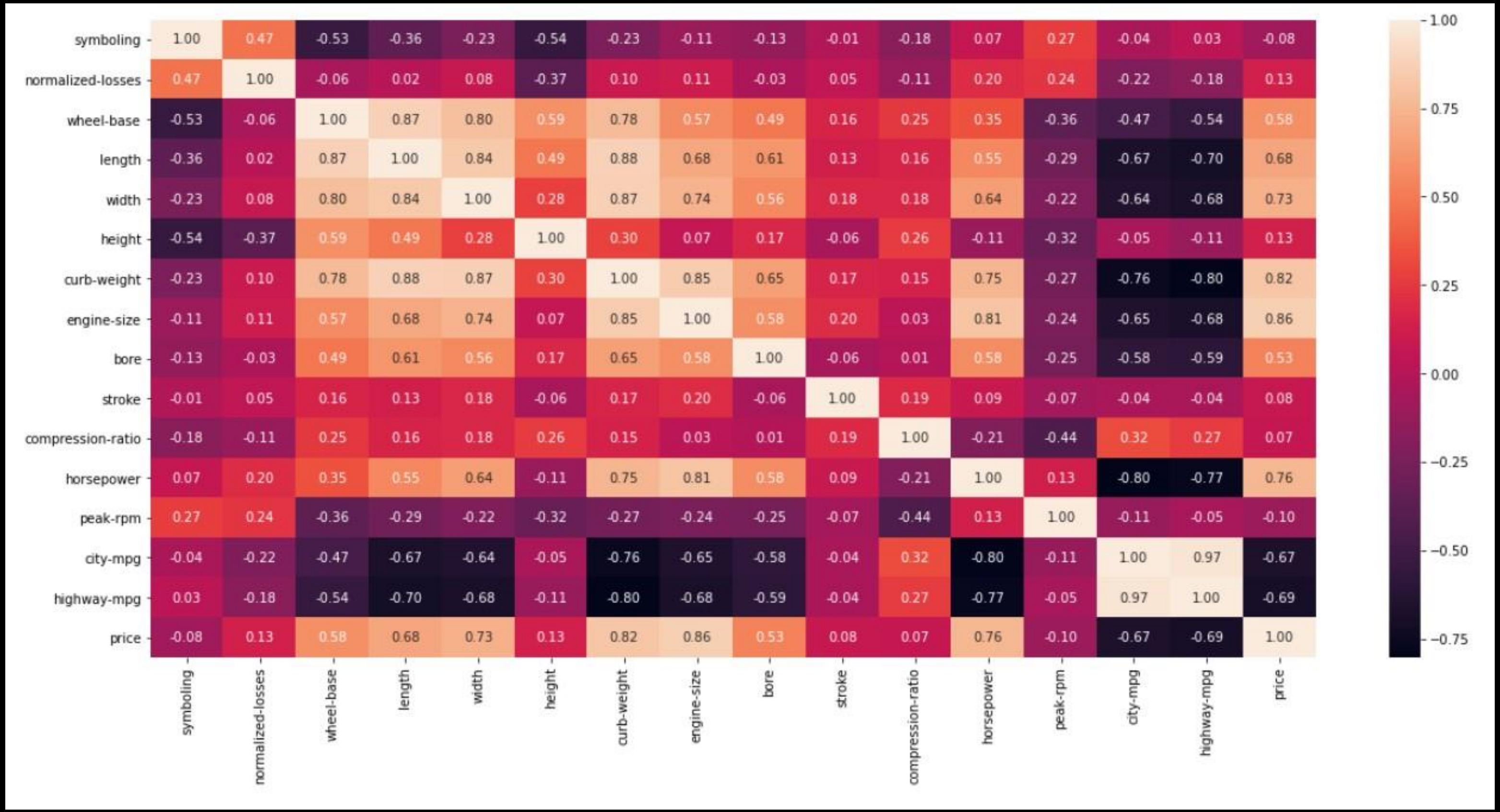
```
symboling             0
normalized-losses    41
make                  0
fuel-type             0
aspiration            0
num-of-doors          2
body-style            0
drive-wheels          0
engine-location       0
wheel-base            0
length                0
width                 0
height                0
curb-weight           0
engine-type           0
num-of-cylinders      0
engine-size           0
fuel-system           0
bore                  4
stroke                4
compression-ratio     0
horsepower            2
peak-rpm              2
city-mpg              0
highway-mpg           0
price                 4
dtype: int64
```

In our data set, there are several missing values with the highest number on normalized-losses feature.

Treatment for missing value:

- Remove rows with missing value in the price variable.
- Discard the normalized-losses column because the amount of missing value is too high.
- Performs imputation on stroke, bore, peak-rpm, num-of-doors, and horsepower features with their median, mean, or mode values (depending on the type and distribution of data)

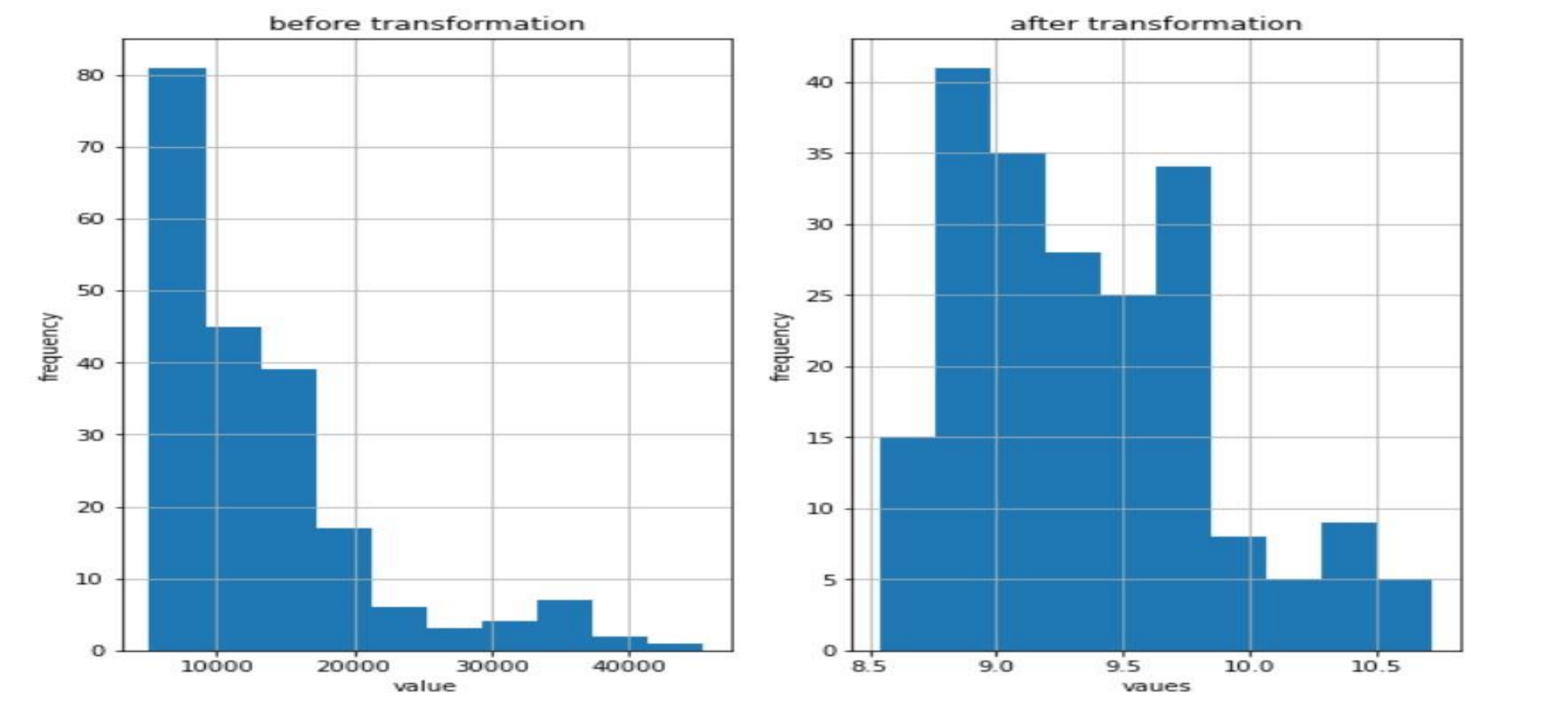
Exploratory Data Analysis



target variable has a strong correlation with several variables such as:

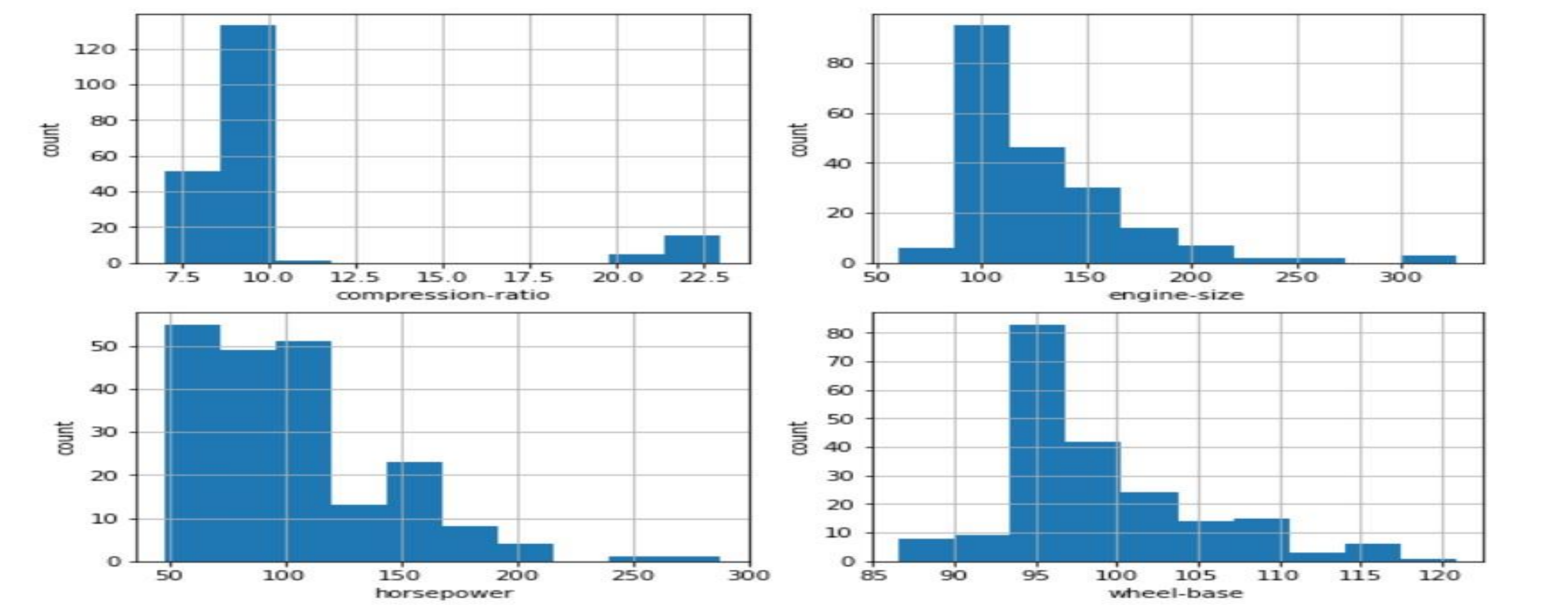
- make
- engine-size
- curb-weight
- horsepower
- num-of-cylinders
- width
- etc

Price Skew Transformation



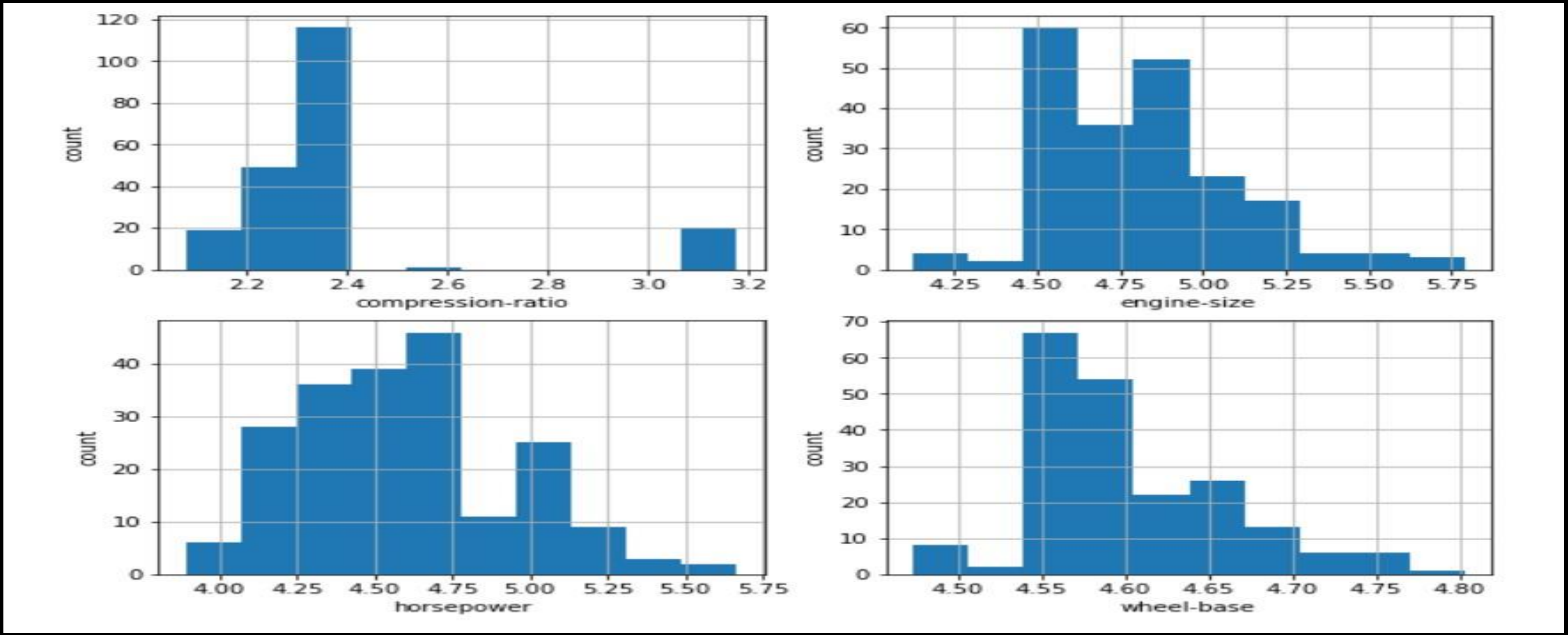
Price has high skew value ,Which has transformed .

High skew variables



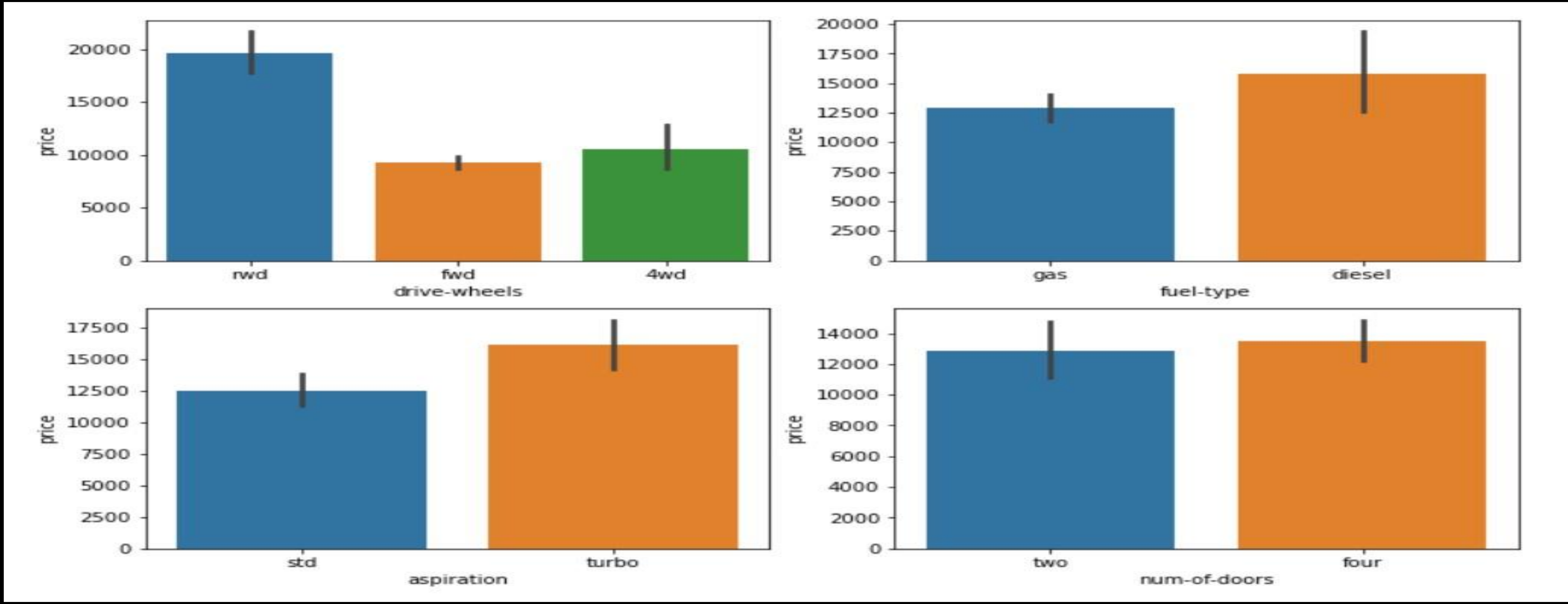
Compression-ratio, horsepower, engine-size, wheel-base have highly skewed distribution. We will do log transformation to these variables to get a more normal distribution.

After transformation



After transformation, we have slightly a more normal distribution.

Multivariate Analysis



- Diesel car have higher average car price compared to the other category.
- Car with four doors have slightly higher price compared to car with two doors.
- Car with rwd type have more higher average car price.

Hypothesis

From the previous slide, we can formulate below 3 hypothesis:

Hypothesis # 1:

H0 = Car with a fuel-type diesel has the same average price as gas car.

H1 = Diesel fuel-type car has an average price that is different from gas car.

Hypothesis # 2:

H0 = Car with std aspiration has the same average price as a turbo aspiration car.

H1 = Car with std aspiration has an average price that is different from a turbo aspiration car.

Hypothesis # 3:

H0 = A two-door car has the same average price as a four-door car.

H1 = A two-door car has an average price different from a four-door car.

Hypothesis Testing

We will do hypothesis testing on first (#1) hypothesis using T-test with 5% significance.

Hypothesis # 1:

H0 = Car with a fuel-type diesel has the same average price as a gas car. H1 = Car with a fuel-type diesel has an average price that is different from gas car.

We get a P-Value of more than 0.05.

Conclusion: Accept H0 (we don't have enough evidence to reject H0)

```
import scipy.stats as st

ttest = st.ttest_ind(a = diesel_car['price'], b = gas_car['price'])
p_value = ttest.pvalue
print('P-Value :',p_value)
if p_value >= 0.05:
    print('Car with a fuel-type diesel has the same average price as a gas car.')
else:
    print('Car with a fuel-type diesel has an average price that is different from gas car.')
```

```
P-Value : 0.1189625443809135
Car with a fuel-type diesel has the same average price as a gas car.
```

Recommendation

Suggestions for next steps in analyzing this data:

- Do a deeper analysis of other variables because the dataset has quite a number of variables
- Perform hypothesis testing on other variables
- Perform regression modeling to predict car prices

Data quality:

Good

The quality of the data is quite good because there are only a few missing values.

- The data format is also clean.
- The quality of the data is quite good because there are only a few missing values.
- There are quite a number of variables.
- There are quite a number of variables.

Bad

- The number of observations is quite small, so that if possible additional observations are necessary to make better model.