Car price prediction model

Mayank Prashar

Steps to develop the model

1. Understanding the data
2. Data cleaning
3. Data visualization
4. Model building

Understanding the data

- Import various libraries required
- Read the data
- Check the various attributes of data:-
  - Name of columns
  - How the data is distributed
  - What kind of data types we have

Observations made:-

- The data had 26 columns and 205 entries.
- Most of the data was in numerical form and there was less of the categorical data.

Steps to develop the model

1. Understanding the data
2. Data cleaning
3. Data visualization
4. Model building

Data cleaning and preparation

- splitted the ‘carname’ to the ‘company name’ and the ‘car model’ and created a separate column for them.
- Many of the company names were written wrong, replaced those names with the right ones.
- There were no duplicate values found,so didn't had to do anything with that.
- Lastly converted all the column names to lower case,to avoid any case difference error.

Visualising the data

Categorical Data

- Company
- Symboling
- Fuel type
- Engine type
- Carbody
- Door number
- Engine location
- Fuel system
- Cylinder number
- Aspiration
- Drive wheel

Numerical Variables

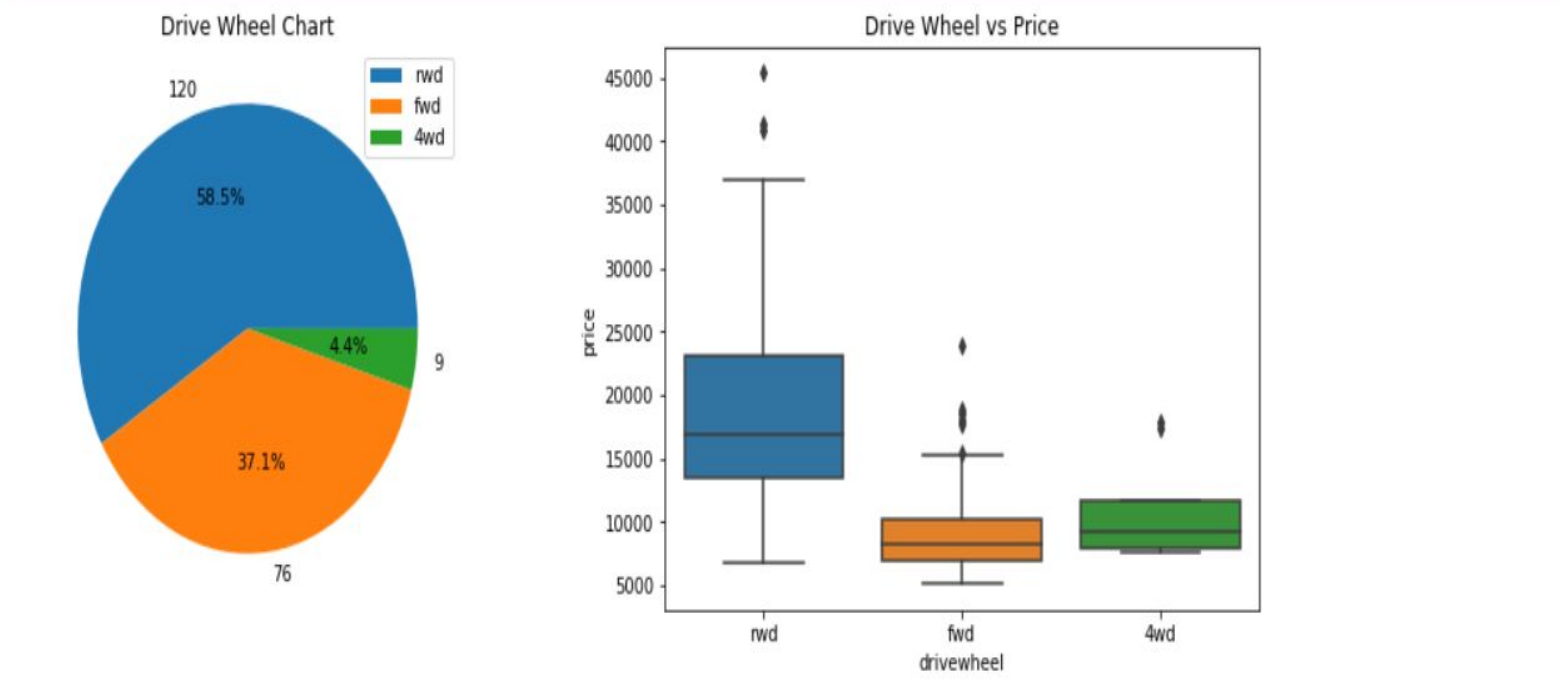
- Car Length
- Car Width
- Car Height
- Curb Weight
- Horsepower
- Bore Ratio
- Compression Ratio
- Highway miles per gallon (mpg)
- Engine Size
- Stroke
- City Miles per gallon (mpg)
- Peak Revolutions per Minute (rpm)
- Wheel Base

Categorical data

- Plotted a set of various graphs representing each categorical data and its relation with the price to understand there relation with price.
- Graphs used here were **pie charts ,bar graphs and histograms** for plotting the categorical data distributions and for plotting their relations with price ,the graphs used were mostly **box plots**.

Fig1.1 The data distribution in the drive wheel

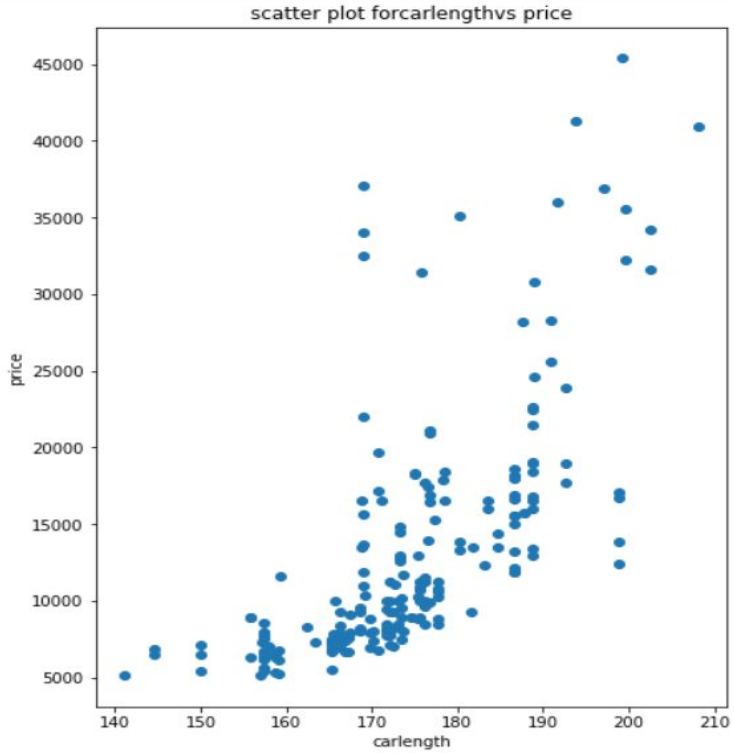
Fig1.2 The relationship between drive wheel and the price.

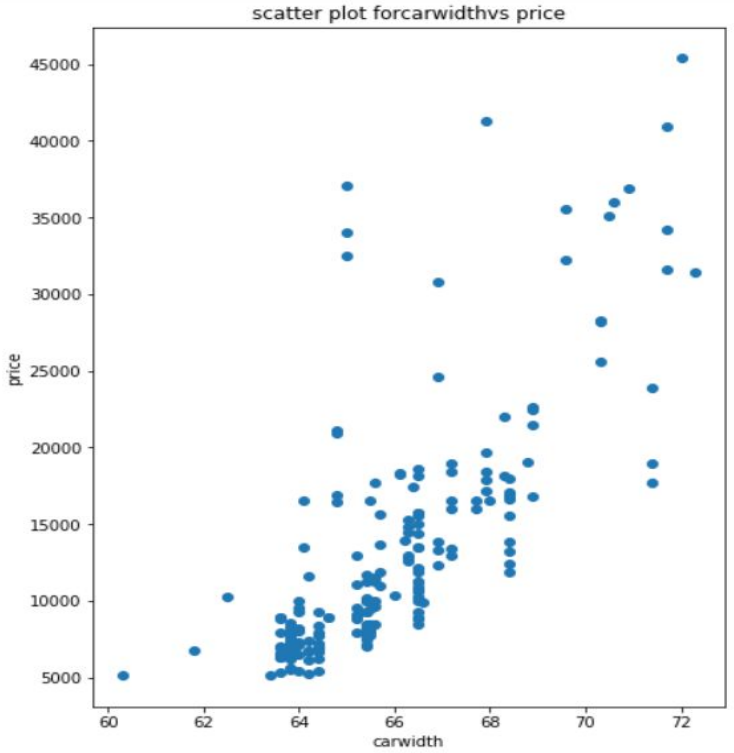


Numerical variables

- Plotted a set of various graphs representing each numerical variable and its relation with the price to understand there relation with price.
- Graphs used here were **scatter plot** for plotting the numerical variables distributions and also for their relations with price .

Fig2.1 the scatterplot showing the relationship of car length with price
Fig2.2 the scatterplot showing the relationship of car width with price





Steps to develop the model

1. Understanding the data
2. Data cleaning
3. Data visualization
4. Model building

Choosing the variables for the regression

- Finding the correlation of all the variables with respect to price and sort them as per their relations.

	price	correlation
wheelbase	0.578	strongly positive
carlength	0.683	strongly positive
carwidth	0.759	strongly positive
curbweight	0.835	strongly positive
enginesize	0.874	strongly positive
boreratio	0.553	strongly positive
horsepower	0.808	strongly positive
citympg	-0.686	strongly negative
highwaympg	-0.698	strongly negative
price	1.000	strongly positive

- Now,took only those variables with strongly negative or strongly positive correlation with the price.
- Converting all the categorical data into numerical data.

Model building

import the train_test_split from sklearn.

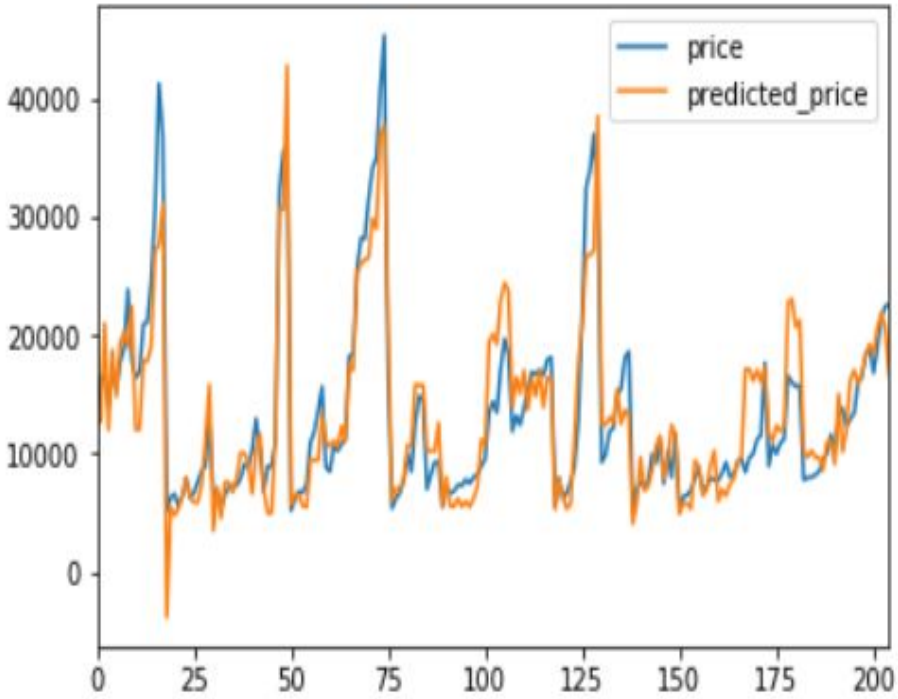
Split the data into train and test taking 20% test and 80% train data.

Now apply the linear regression(as the data is continuous) and predict the value based on the given data.

Check for the accuracy of the data.comes out to be 78%.

VISUALIZE THE RESULT

<Figure size 1800x1080 with 0 Axes>




thank
you