03_correlation

August 8, 2025

1 Import Required Libraries

This cell imports the necessary Python libraries for data analysis and visualization.

[86]:
```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

2 Load Data

Read the automobile dataset and split into features and target variable.

[87]:
```python
df = pd.read_csv("F:\\projects\\car_pricing\\data\\automobile.csv")
print(df.head)
```

```
<bound method NDFrame.head of      symboling  normalized-losses         make
fuel-type aspiration  \
0            3                NaN  alfa-romero       gas        std
1            3                NaN  alfa-romero       gas        std
2            1                NaN  alfa-romero       gas        std
3            2              164.0         audi       gas        std
4            2              164.0         audi       gas        std
..         …                …          …       …        …
196         -1               95.0        volvo       gas        std
197         -1               95.0        volvo       gas      turbo
198         -1               95.0        volvo       gas        std
199         -1               95.0        volvo    diesel      turbo
200         -1               95.0        volvo       gas      turbo

    num-of-doors   body-style drive-wheels engine-location  wheel-base  …  \
0            two  convertible          rwd           front        88.6  …
1            two  convertible          rwd           front        88.6  …
2            two    hatchback          rwd           front        94.5  …
3           four        sedan          fwd           front        99.8  …
4           four        sedan          4wd           front        99.4  …
..           …          …          …             …         …  …
196         four        sedan          rwd           front       109.1  …
```

```
197          four        sedan          rwd          front       109.1  …
198          four        sedan          rwd          front       109.1  …
199          four        sedan          rwd          front       109.1  …
200          four        sedan          rwd          front       109.1  …

     engine-size  fuel-system  bore  stroke compression-ratio horsepower  \
0            130         mpfi  3.47    2.68               9.0      111.0
1            130         mpfi  3.47    2.68               9.0      111.0
2            152         mpfi  2.68    3.47               9.0      154.0
3            109         mpfi  3.19    3.40              10.0      102.0
4            136         mpfi  3.19    3.40               8.0      115.0
..           …          …   …     …              …         …
196          141         mpfi  3.78    3.15               9.5      114.0
197          141         mpfi  3.78    3.15               8.7      160.0
198          173         mpfi  3.58    2.87               8.8      134.0
199          145          idi  3.01    3.40              23.0      106.0
200          141         mpfi  3.78    3.15               9.5      114.0

     peak-rpm city-mpg  highway-mpg  price
0      5000.0       21           27  13495
1      5000.0       21           27  16500
2      5000.0       19           26  16500
3      5500.0       24           30  13950
4      5500.0       18           22  17450
..        …      …          …    …
196    5400.0       23           28  16845
197    5300.0       19           25  19045
198    5500.0       18           23  21485
199    4800.0       26           27  22470
200    5400.0       19           25  22625

[201 rows x 26 columns]>
```

3 Display the DataFrame

Show the loaded automobile data for inspection.

[88]:
```python
from IPython.display import display

# Display the dataframe
display(df)
```

```
     symboling  normalized-losses         make fuel-type aspiration  \
0            3                NaN  alfa-romero       gas        std
1            3                NaN  alfa-romero       gas        std
2            1                NaN  alfa-romero       gas        std
3            2              164.0         audi       gas        std
4            2              164.0         audi       gas        std
```

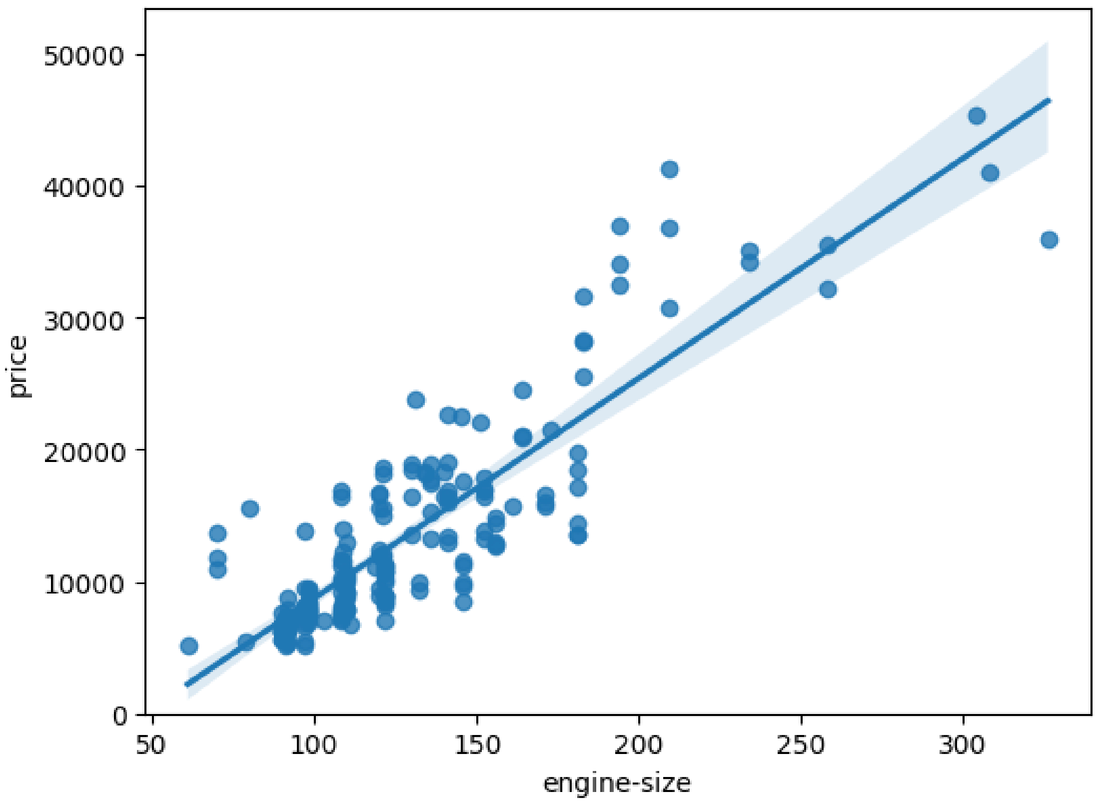
```
..          …                …          …       …         …
196        -1             95.0      volvo      gas       std
197        -1             95.0      volvo      gas     turbo
198        -1             95.0      volvo      gas       std
199        -1             95.0      volvo   diesel     turbo
200        -1             95.0      volvo      gas     turbo

    num-of-doors   body-style drive-wheels engine-location  wheel-base  …  \
0            two  convertible          rwd           front        88.6  …
1            two  convertible          rwd           front        88.6  …
2            two    hatchback          rwd           front        94.5  …
3           four        sedan          fwd           front        99.8  …
4           four        sedan          4wd           front        99.4  …
..           …            …          …              …          …   …
196         four        sedan          rwd           front       109.1  …
197         four        sedan          rwd           front       109.1  …
198         four        sedan          rwd           front       109.1  …
199         four        sedan          rwd           front       109.1  …
200         four        sedan          rwd           front       109.1  …

     engine-size  fuel-system  bore  stroke compression-ratio horsepower  \
0            130         mpfi  3.47    2.68               9.0      111.0
1            130         mpfi  3.47    2.68               9.0      111.0
2            152         mpfi  2.68    3.47               9.0      154.0
3            109         mpfi  3.19    3.40              10.0      102.0
4            136         mpfi  3.19    3.40               8.0      115.0
..           …            …   …     …                …         …
196          141         mpfi  3.78    3.15               9.5      114.0
197          141         mpfi  3.78    3.15               8.7      160.0
198          173         mpfi  3.58    2.87               8.8      134.0
199          145          idi  3.01    3.40              23.0      106.0
200          141         mpfi  3.78    3.15               9.5      114.0

     peak-rpm city-mpg  highway-mpg  price
0      5000.0       21           27  13495
1      5000.0       21           27  16500
2      5000.0       19           26  16500
3      5500.0       24           30  13950
4      5500.0       18           22  17450
..        …        …          …     …
196    5400.0       23           28  16845
197    5300.0       19           25  19045
198    5500.0       18           23  21485
199    4800.0       26           27  22470
200    5400.0       19           25  22625

[201 rows x 26 columns]
```

4 Visualize Engine Size vs Price

Plot a regression line to examine the relationship between engine size and price.

[89]:
```
# Engine size as potential predictor variable of price
import seaborn as sns
sns.regplot(x="engine-size", y="price", data=df)
plt.ylim(0,)
```

[89]:
```
(0.0, 53445.84773783944)
```



5 Correlation: Engine Size and Price

Calculate the correlation coefficient between engine size and price.

[90]:
```
df[["engine-size", "price"]].corr()
```

[90]:

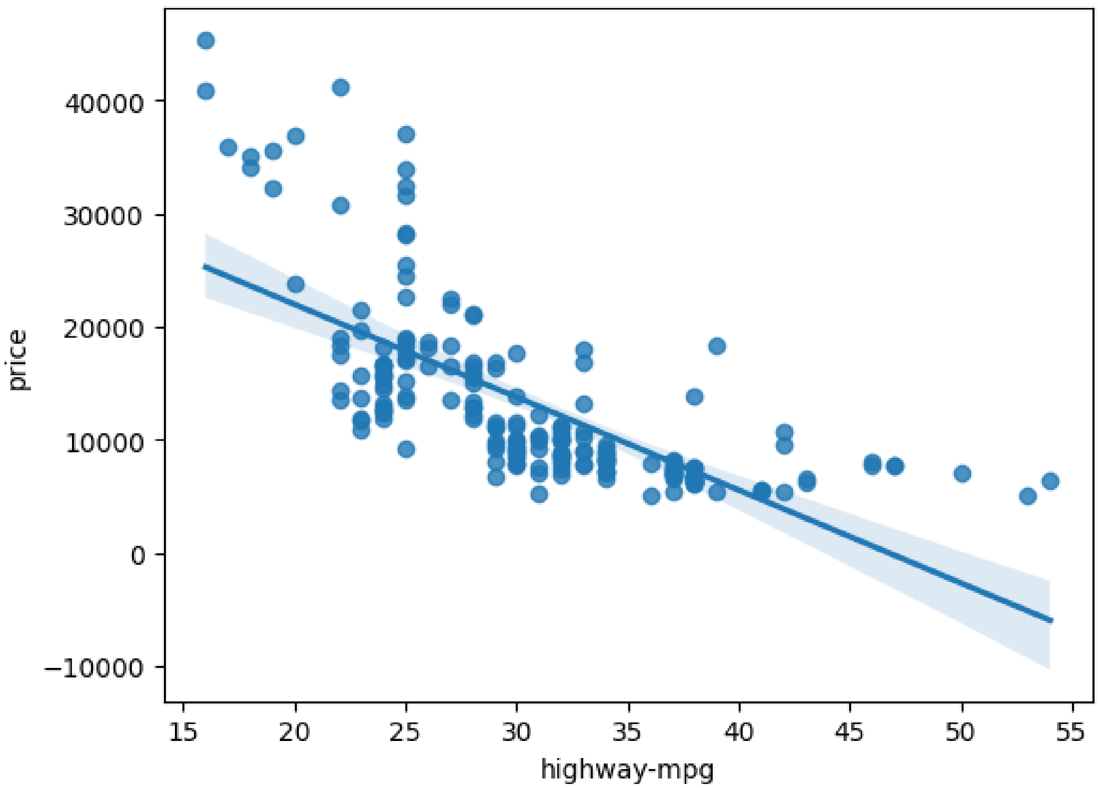
	engine-size	price
engine-size	1.000000	0.872335
price	0.872335	1.000000

6 Visualize Highway MPG vs Price

Plot a regression line to examine the relationship between highway-mpg and price.

[91]:
```
sns.regplot(x="highway-mpg", y="price", data=df)
```

[91]: <Axes: xlabel='highway-mpg', ylabel='price'>



7 Correlation: Highway MPG and Price

Calculate the correlation coefficient between highway-mpg and price.

[92]:
```
df[['highway-mpg', 'price']].corr()
```

[92]:

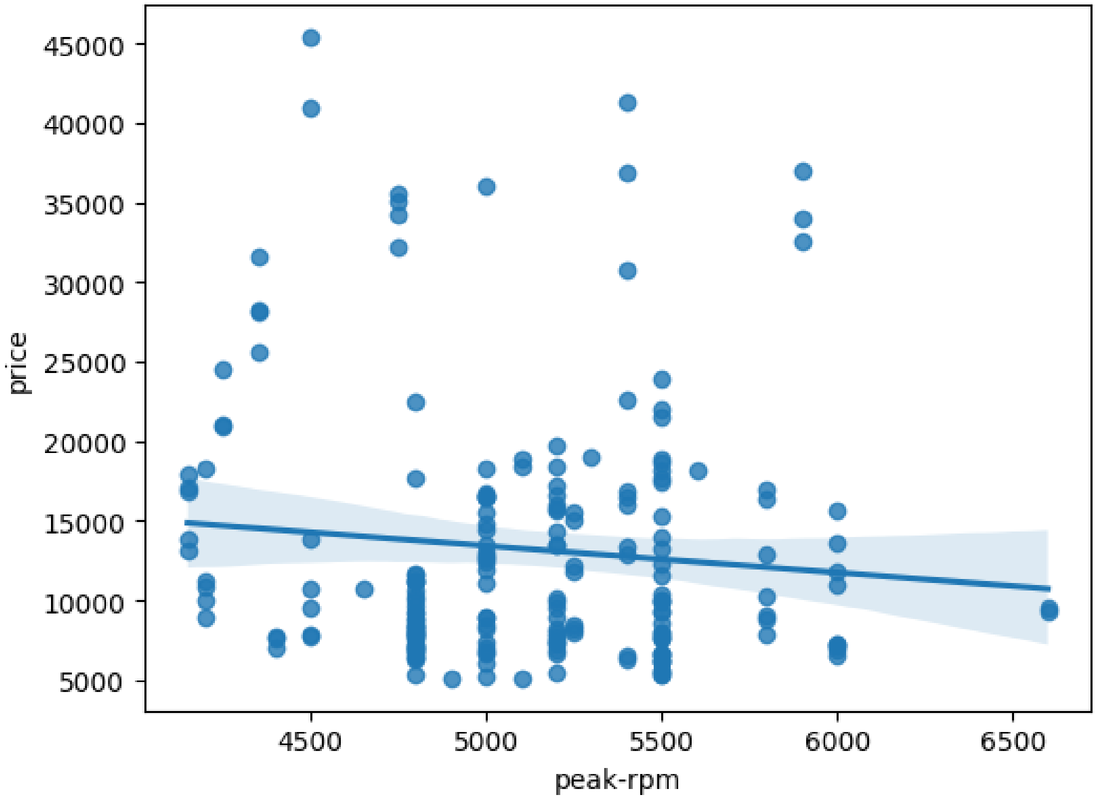
	highway-mpg	price
highway-mpg	1.000000	-0.704692
price	-0.704692	1.000000

8 Visualize Peak RPM vs Price

Plot a regression line to examine the relationship between peak-rpm and price.

[93]:
```
sns.regplot(x="peak-rpm", y="price", data=df)
```

[93]:
```
<Axes: xlabel='peak-rpm', ylabel='price'>
```



9 Correlation: Peak RPM and Price

Calculate the correlation coefficient between peak-rpm and price.

[94]:
```
df[['peak-rpm','price']].corr()
```

[94]:

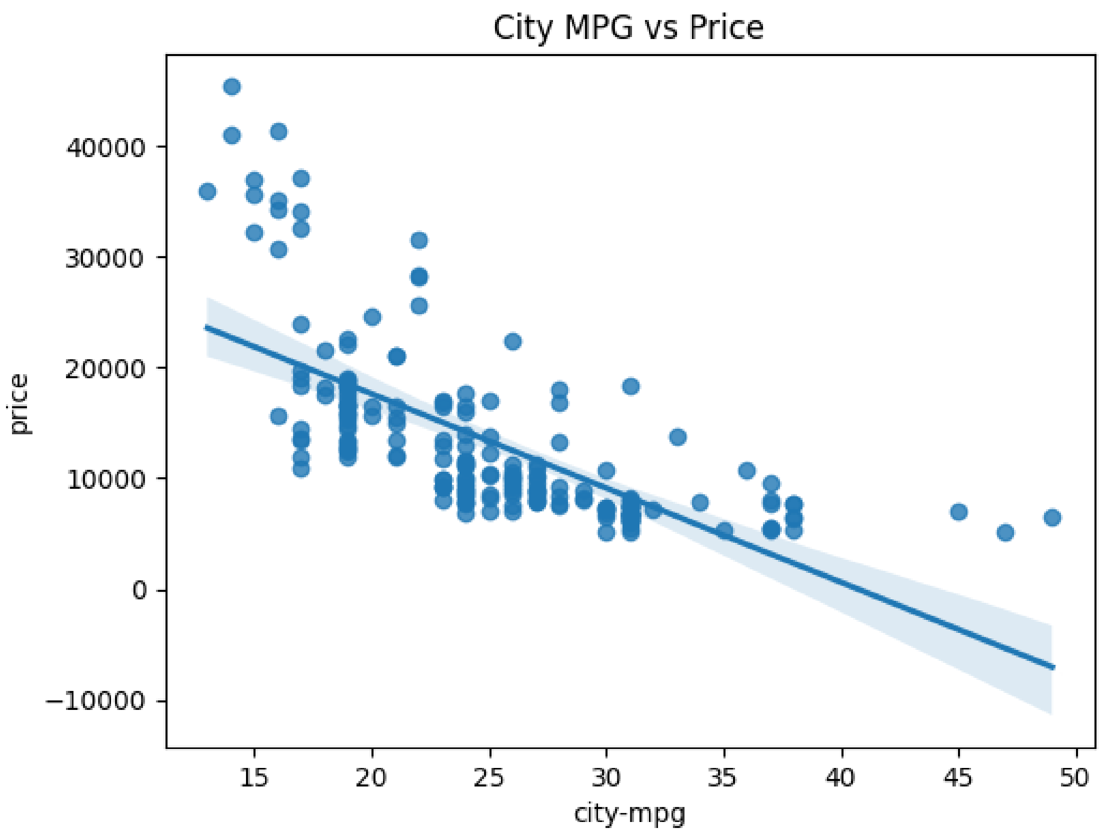
	peak-rpm	price
peak-rpm	1.000000	-0.101649
price	-0.101649	1.000000

10 Visualize City MPG vs Price

Plot a regression line to examine the relationship between city-mpg and price.

[95]:
```
# Plotting a negative relationship between two variables
sns.regplot(x="city-mpg", y="price", data=df)
plt.title("City MPG vs Price")
```

```
plt.show()
```



11 Correlation: City MPG and Price

Calculate the correlation coefficient between city-mpg and price.

[96]:
```
df[['city-mpg','price']].corr()
```

[96]:

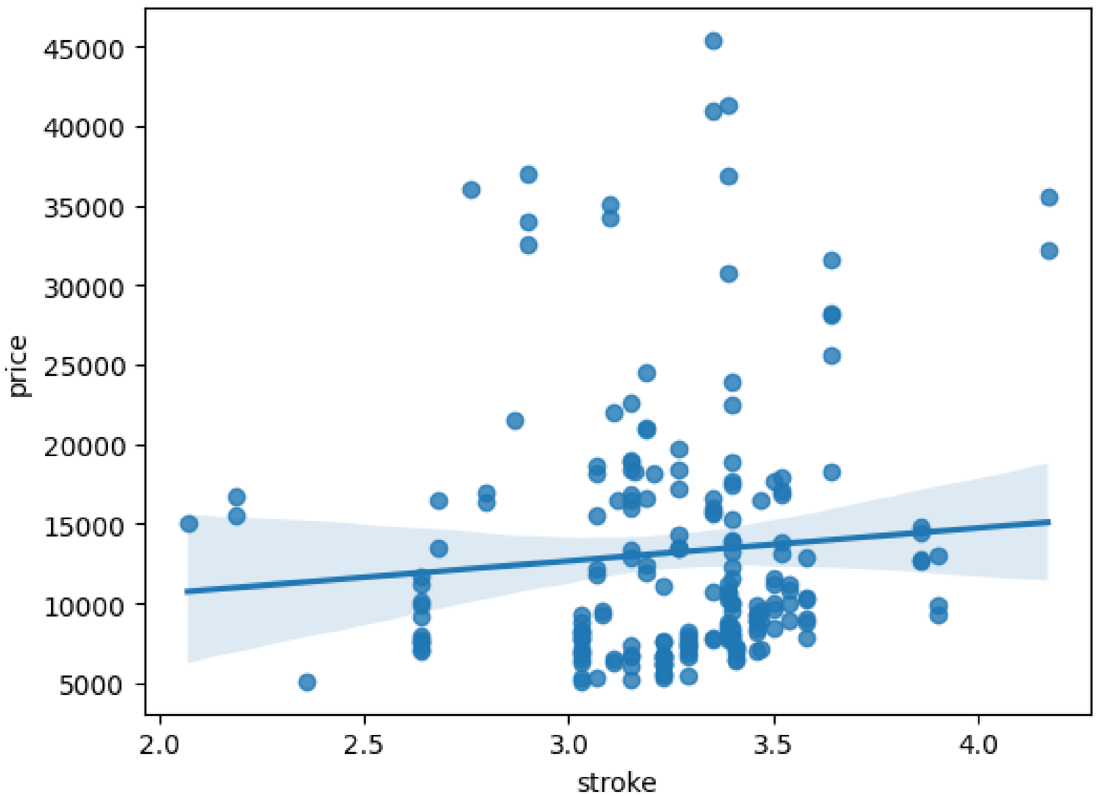
	city-mpg	price
city-mpg	1.000000	-0.686571
price	-0.686571	1.000000

12 Visualize Stroke vs Price

Plot a regression line to examine the relationship between stroke and price.

[97]:
```
sns.regplot(x="stroke", y="price", data=df)
```

[97]:
```
<Axes: xlabel='stroke', ylabel='price'>
```



13 Correlation: Stroke and Price

Calculate the correlation coefficient between stroke and price.

[98]:
```
df[["stroke","price"]].corr()
```

[98]:

	stroke	price
stroke	1.00000	0.08231
price	0.08231	1.00000

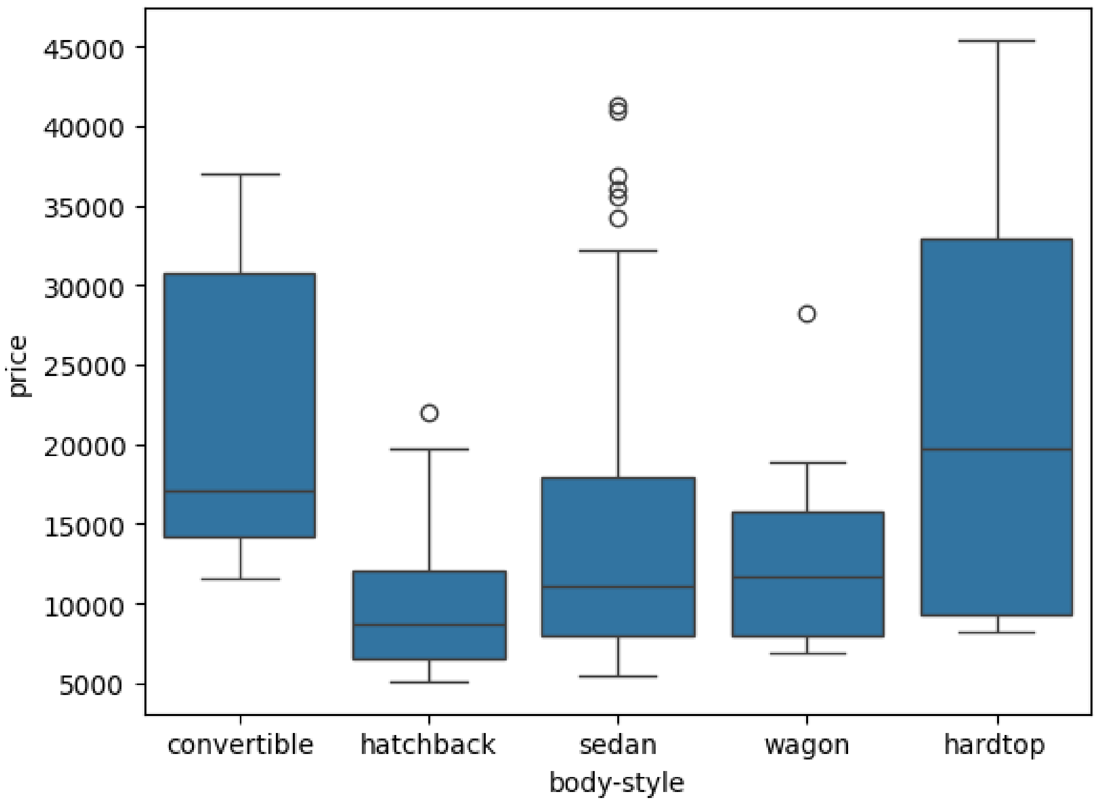
Categorical Variables

These are variables that describe a 'characteristic' of a data unit, and are selected from a small group of categories. The categorical variables can have the type "object" or "int64". A good way to visualize categorical variables is by using boxplots.

Let's look at the relationship between "body-style" and "price".

[99]:
```
sns.boxplot(x="body-style", y="price", data=df)
```

[99]: <Axes: xlabel='body-style', ylabel='price'>



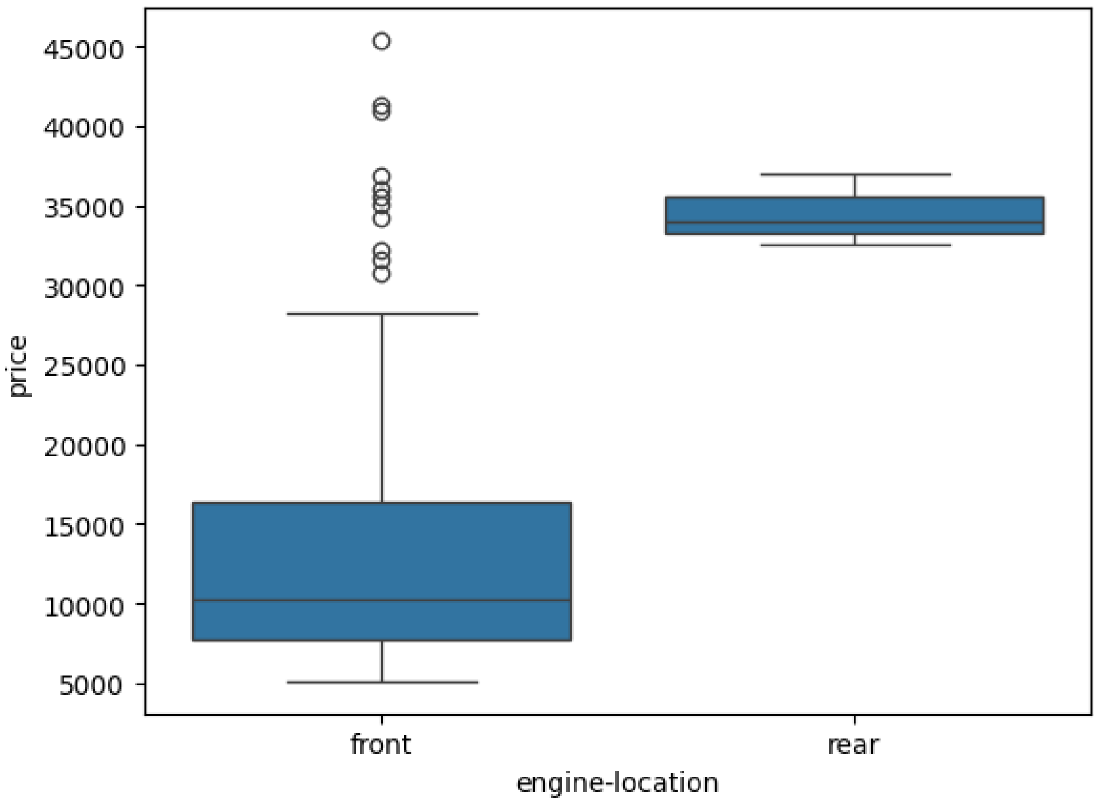
14 Visualize Body Style vs Price

Use a boxplot to examine the relationship between body-style and price.

We see that the distributions of price between the different body-style categories have a significant overlap, so body-style would not be a good predictor of price. Let's examine engine "engine-location" and "price":

[100]:
```
sns.boxplot(x="engine-location", y="price", data=df)
```

[100]: <Axes: xlabel='engine-location', ylabel='price'>



15 Visualize Engine Location vs Price

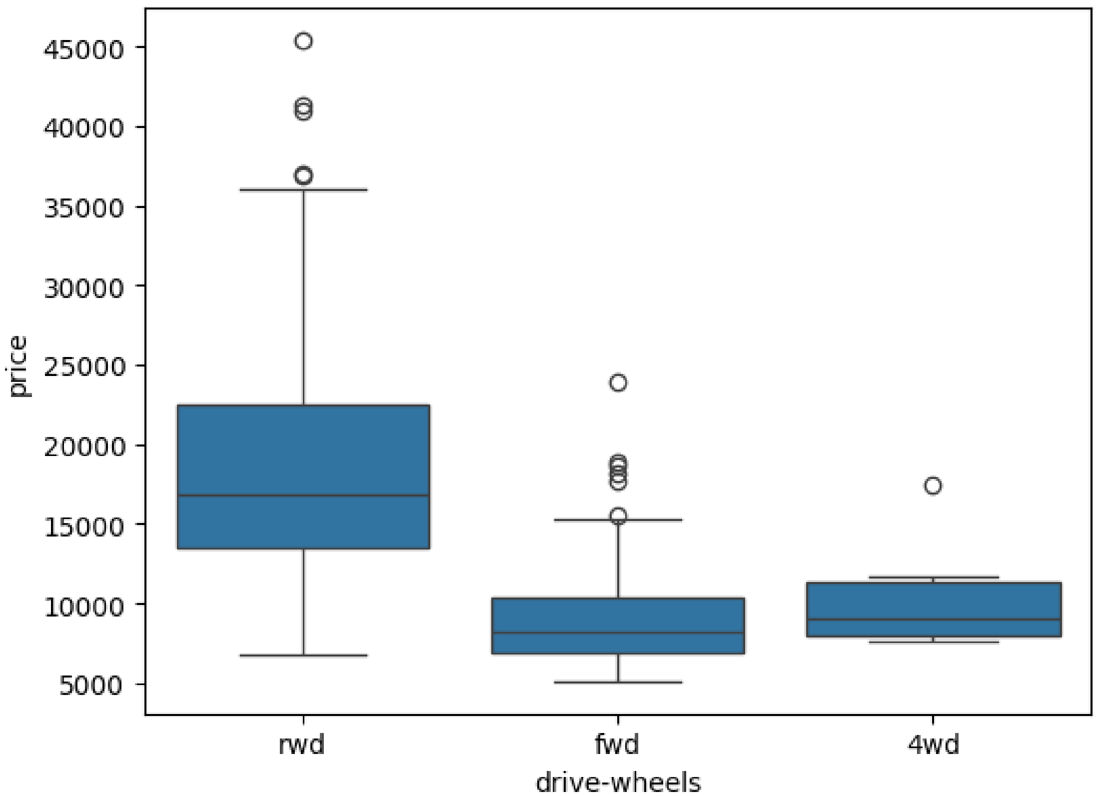
Use a boxplot to examine the relationship between engine-location and price.

Here we see that the distribution of price between these two engine-location categories, front and rear, are distinct enough to take engine-location as a potential good predictor of price.

Let's examine "drive-wheels" and "price".

[101]:
```
# drive-wheels
sns.boxplot(x="drive-wheels", y="price", data=df)
```

[101]: <Axes: xlabel='drive-wheels', ylabel='price'>



Here we see that the distribution of price between the different drive-wheels categories differs. As such, drive-wheels could potentially be a predictor of price.

16 Conclusion

In this analysis, we explored the relationships between various numerical and categorical features and car price.

- **Positive Correlation:** Features like engine size showed a strong positive correlation with price, meaning as engine size increases, the price tends to increase as well.
- **Negative Correlation:** Features such as highway-mpg and city-mpg had a strong negative correlation with price, indicating that higher mileage is associated with lower car prices.
- **Weak Correlation:** Some features, like stroke and peak-rpm, showed weak or no significant correlation with price, so they are less useful for predicting price.
- **Categorical Variables:** Among categorical variables, engine location and drive-wheels showed clear differences in price distributions, making them potentially useful predictors.

Understanding the direction and strength of these correlations helps in selecting the best features for building predictive models for car pricing.

[102]:
```
!jupyter nbconvert --to pdf --output "03_correlation.pdf" "03_correlation.ipynb"
```

```
[NbConvertApp] Converting notebook 03_correlation.ipynb to pdf
[NbConvertApp] Support files will be in 03_correlation_files\
[NbConvertApp] Making directory .\03_correlation_files
[NbConvertApp] Writing 51484 bytes to notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', 'notebook']
[NbConvertApp] WARNING | b had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 281089 bytes to 03_correlation.pdf
```