Car Price Prediction and Key Determinants in the U.S. Market

Content

1. Data Information

2. Explore Data Analysis

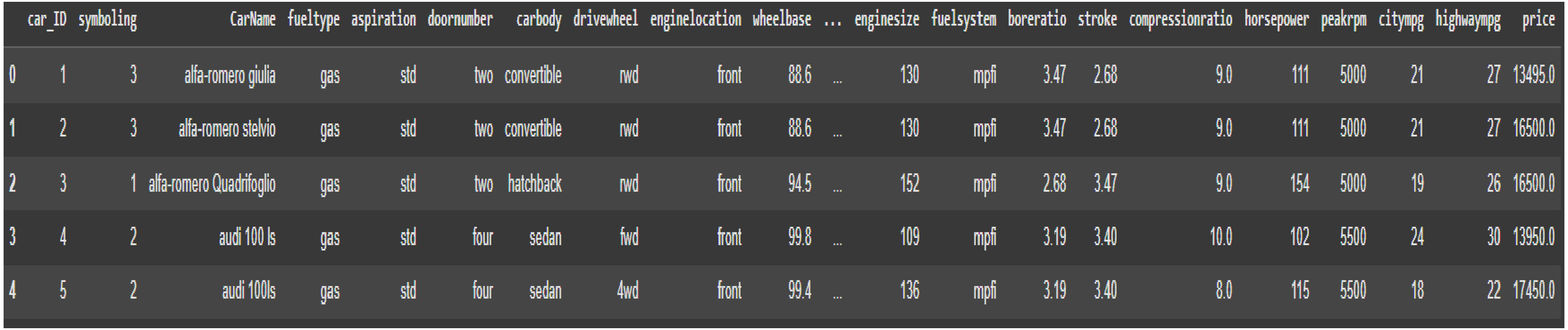
3. Model Selection

4. Evaluation and Result

DATA INFORMATION

A Chinese automobile company, Geely Auto, plans to enter the U.S. market by establishing a local manufacturing unit to produce cars that can compete with American and European brands. Specifically, the company wants to understand:

➔ What are the key factors that determine car pricing?
➔ Which technical specifications of a car have the most significant impact on its price?



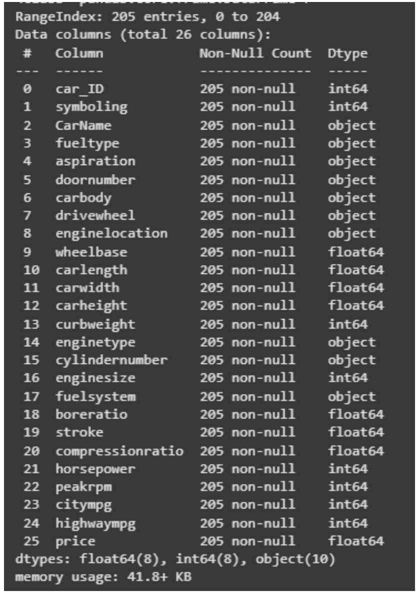
	car_ID	symboling	CarName	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	...	enginesize	fuelsystem	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price
0	1	3	alfa-romero giulia	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111	5000	21	27	13495.0
1	2	3	alfa-romero stelvio	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111	5000	21	27	16500.0
2	3	1	alfa-romero Quadrifoglio	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	154	5000	19	26	16500.0
3	4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0	102	5500	24	30	13950.0
4	5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0	115	5500	18	22	17450.0

DATA INFORMATION

1. Car price data dictionary

Column name	Description
Car_ID	Unique id of each observation (Integer)
Symboling	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
carCompany	Name of car company (Categorical)
fueltype	Car fuel type i.e gas or diesel (Categorical)
aspiration	Aspiration used in a car (Categorical)
doornumber	Number of doors in a car (Categorical)
carbody	body of car (Categorical)
drivewheel	type of drive wheel (Categorical)
enginelocation	Location of car engine (Categorical)
wheelbase	Wheelbase of car (Numeric)
carlength	Length of car (Numeric)
carwidth	Width of car (Numeric)
curbweight	The weight of a car without occupants or baggage. (Numeric)
enginetype	Type of engine. (Categorical)
cylindernumber	cylinder placed in the car (Categorical)
enginesize	Size of car (Numeric)
fuelsystem	Fuel system of car (Categorical)
boreratio	Bore Ratio of car (Numeric)
stroke	Stroke or volume inside the engine (Numeric)
compressionratio	compression ratio of car (Numeric)
horsepower	Horsepower (Numeric)
peakrpm	car peak rpm (Numeric)
citympg	Mileage in city (Numeric)
highwaympg	Mileage on highway (Numeric)
price(Dependent variable)	Price of car (Numeric)

EXPLORE DATA ANALYSIS



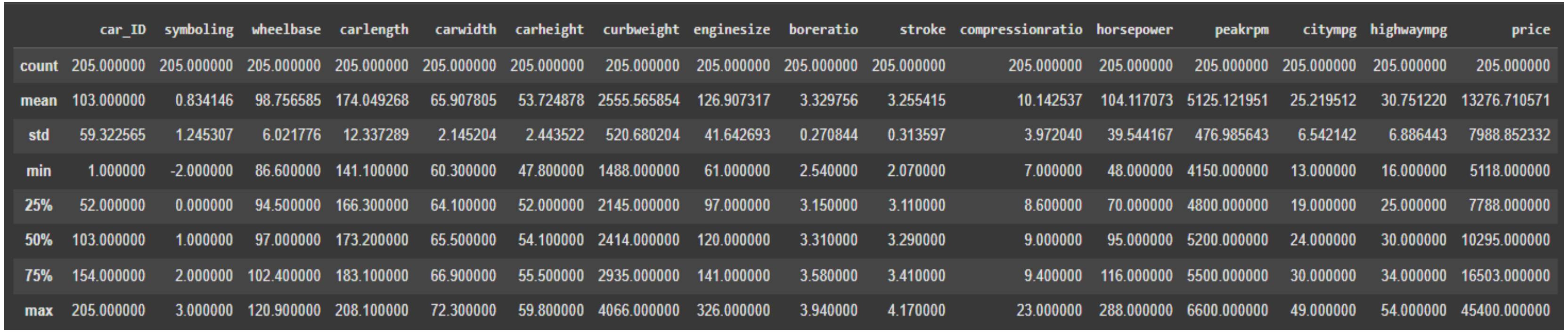
```
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   car_ID            205 non-null    int64  
 1   symboling         205 non-null    int64  
 2   CarName           205 non-null    object 
 3   fueltype          205 non-null    object 
 4   aspiration        205 non-null    object 
 5   doornumber        205 non-null    object 
 6   carbody           205 non-null    object 
 7   drivewheel        205 non-null    object 
 8   enginelocation    205 non-null    object 
 9   wheelbase         205 non-null    float64
 10  carlength         205 non-null    float64
 11  carwidth          205 non-null    float64
 12  carheight         205 non-null    float64
 13  curbweight        205 non-null    int64  
 14  enginetype        205 non-null    object 
 15  cylindernumber    205 non-null    object 
 16  enginesize        205 non-null    int64  
 17  fuelsystem        205 non-null    object 
 18  boreratio         205 non-null    float64
 19  stroke            205 non-null    float64
 20  compressionratio  205 non-null    float64
 21  horsepower        205 non-null    int64  
 22  peakrpm           205 non-null    int64  
 23  citympg           205 non-null    int64  
 24  highwaympg        205 non-null    int64  
 25  price             205 non-null    float64
dtypes: float64(8), int64(8), object(10)
memory usage: 41.8+ KB
```

1. Data Inspection

- Identified and handled missing values
- Checked for duplicated records
- Detected outliers using the IQR
- Fixed misspelled categorical values
- Removed irrelevant or inappropriate

2. Findings

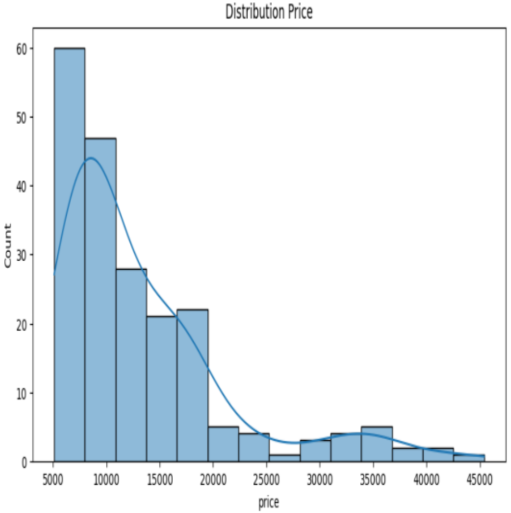
- Dataset contains **205 records** and **15 numerical columns**.
- Numerical features like **carlength**, **curbweight**, and **enginesize** show wide but reasonable ranges.
- Average **engine size** is around **127 cc** and **horsepower** around **104 HP**, indicating diverse car types.
- Average price is **$13,276**, with values ranging from **$5,118** to **$45,400**, suggesting high variability.
- Fuel efficiency averages **25 city MPG** and **30 highway MPG**, consistent with typical car performance.



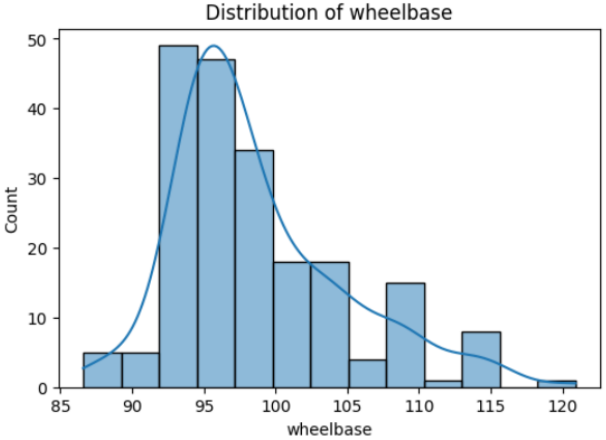
	car_ID	symboling	wheelbase	carlength	carwidth	carheight	curbweight	enginesize	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	price
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	103.000000	0.834146	98.756585	174.049268	65.907805	53.724878	2555.565854	126.907317	3.329756	3.255415	10.142537	104.117073	5125.121951	25.219512	30.751220	13276.710571
std	59.322565	1.245307	6.021776	12.337289	2.145204	2.443522	520.680204	41.642693	0.270844	0.313597	3.972040	39.544167	476.985643	6.542142	6.886443	7988.852332
min	1.000000	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	2.540000	2.070000	7.000000	48.000000	4150.000000	13.000000	16.000000	5118.000000
25%	52.000000	0.000000	94.500000	166.300000	64.100000	52.000000	2145.000000	97.000000	3.150000	3.110000	8.600000	70.000000	4800.000000	19.000000	25.000000	7788.000000
50%	103.000000	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	3.310000	3.290000	9.000000	95.000000	5200.000000	24.000000	30.000000	10295.000000
75%	154.000000	2.000000	102.400000	183.100000	66.900000	55.500000	2935.000000	141.000000	3.580000	3.410000	9.400000	116.000000	5500.000000	30.000000	34.000000	16503.000000
max	205.000000	3.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000	3.940000	4.170000	23.000000	288.000000	6600.000000	49.000000	54.000000	45400.000000

EXPLORE DATA ANALYSIS

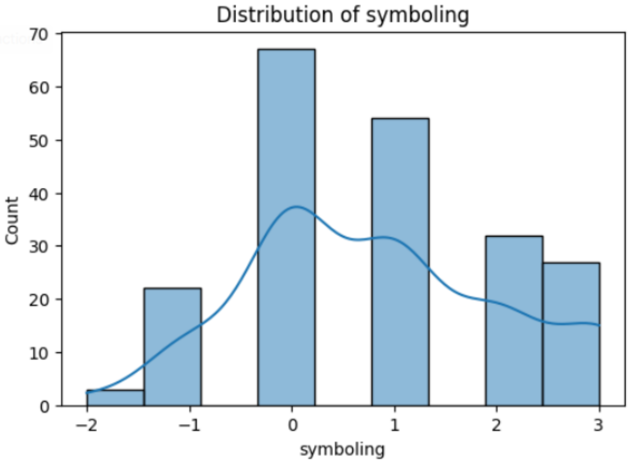
Univariate Analysis
Numerical Features:

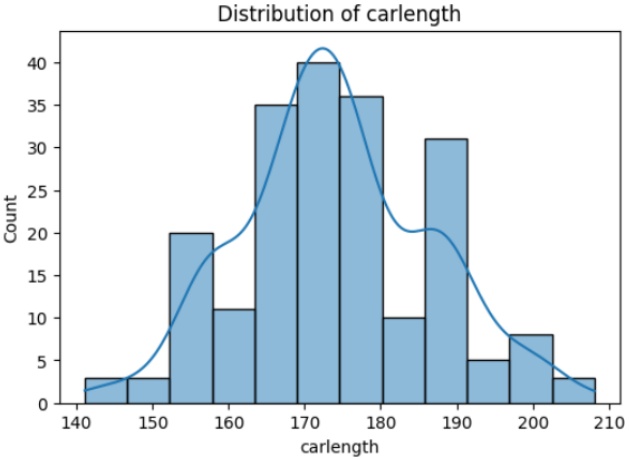


1. Price
Distribution: Right-skewed.
Insight: Most cars are priced below $15,000, while a few luxury cars exceed $30,000, indicating possible outliers.



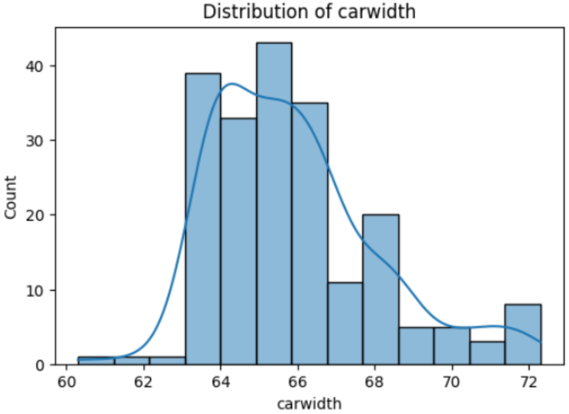
2. Wheelbase
Distribution: Slightly right-skewed.
Insight: Majority of vehicles have a wheelbase between 95–100 inches, with a few larger models beyond 110 inches.
3. Symboling
Distribution: Multi-modal and discrete.
Insight: Represents categorical insurance risk levels rather than a continuous variable.
4. Carlength
Distribution: Approximately normal.
Insight: Most car lengths range from 170–180 inches, with few long vehicles above 190 inches.



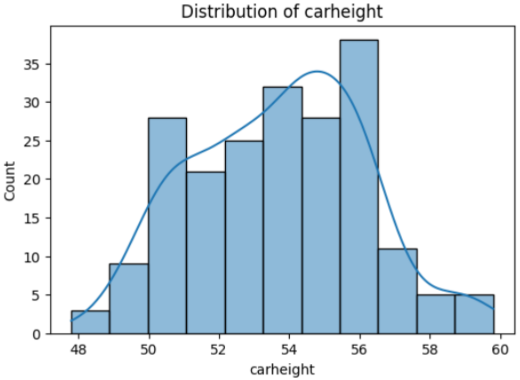


EXPLORE DATA ANALYSIS

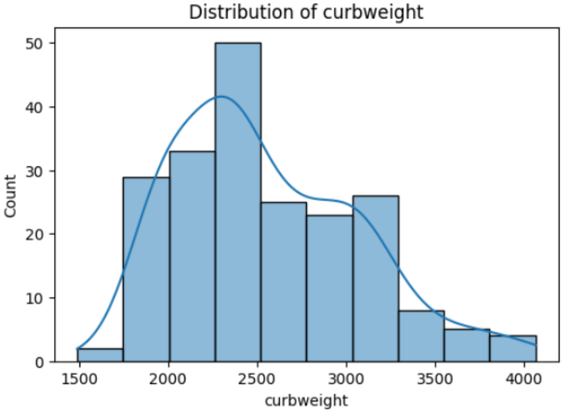
Univariate Analysis Numerical Features:

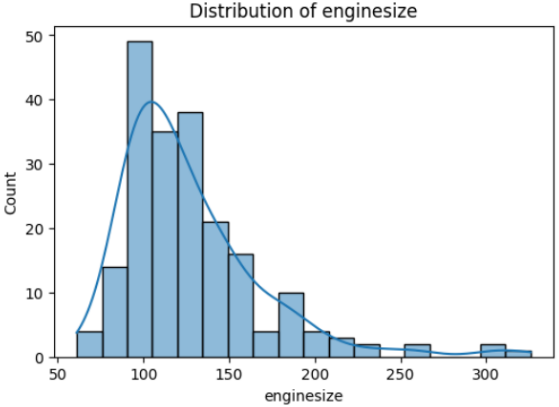


1. Carwidth
Distribution: Nearly normal with a slight right skew.
Insight: Most vehicles have widths between 64–67 inches, with a few wider models above 70 inches.



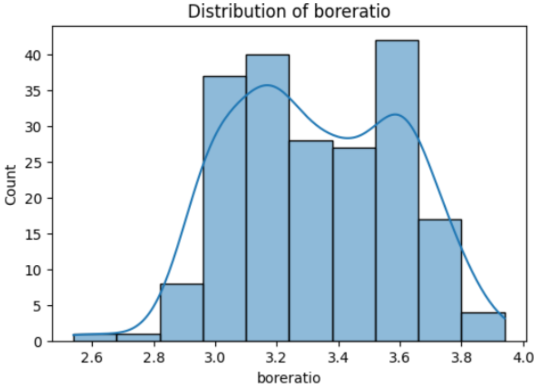
2. Carheight
Distribution: Approximately normal, slightly left-skewed.
Insight: Most vehicles have heights between **52–56 inches**, with very few shorter or taller cars.
3. Curbweight
Distribution: Right-skewed.
Insight: Majority of vehicles weigh between 2,000–3,000 lbs, with heavier luxury cars reaching above 3,500 lbs.
4. Enginesize
Distribution: Strongly right-skewed.
Insight: Most cars have small to mid-size engines (under 150), while a few high-performance cars have engines over 250.





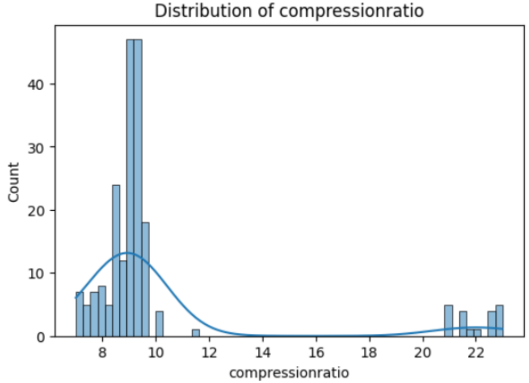
EXPLORE DATA ANALYSIS

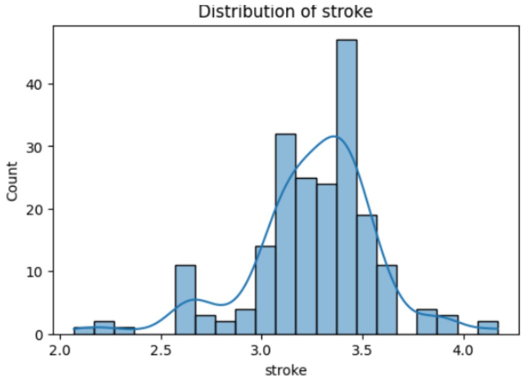
Univariate Analysis Numerical Features:

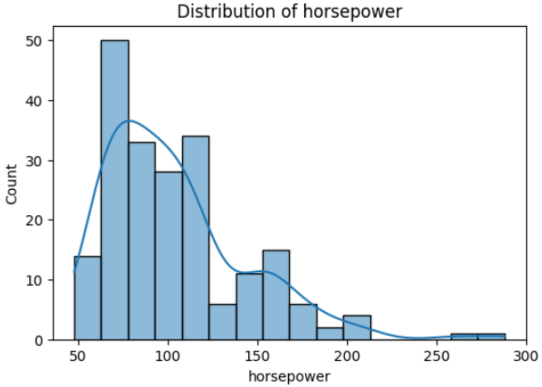


1. Boreratio
Distribution: Approximately normal.
Insight: Most values range between **3.0 and 3.6**, suggesting consistent engine bore proportions across models.

2. Compressionratio
Distribution: Highly right-skewed.
Insight: Most cars have a compression ratio around **8–10**, but a few models have very high ratios (above 20).
3. Stroke
Distribution: Roughly normal with mild variation.
Insight: Majority of vehicles have stroke values between **3.0 and 3.5**, indicating similar piston stroke design.
4. Horsepower
Distribution: Strongly right-skewed.
Insight: Most cars produce **under 120 HP**, while high-performance vehicles reach over **200 HP**.

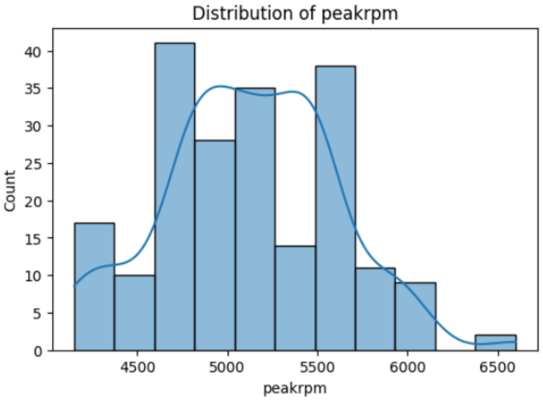


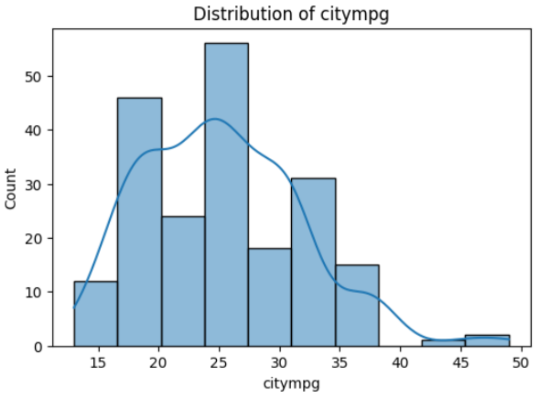


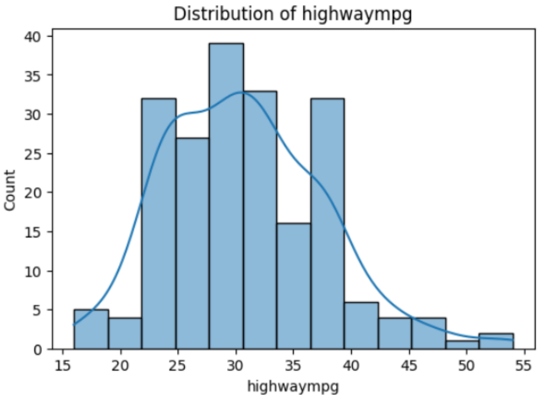


EXPLORE DATA ANALYSIS

Univariate Analysis
Numerical Features:







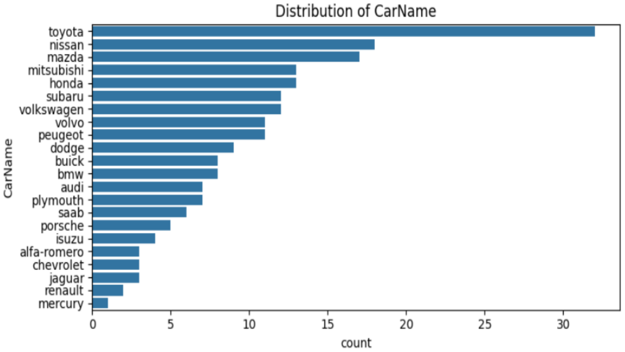
1. Peakrpm
Distribution: Approximately normal with slight right skew.
Insight: Most vehicles reach peak revolutions between **4,800 and 5,500 rpm**, with only a few exceeding 6,000.

2. Citympg
Distribution: Right-skewed.
Insight: The majority of cars have city mileage between **20–30 mpg**, while a few highly efficient models reach up to **45–50 mpg**.

3. Highwaympg
Distribution: Nearly normal.
Insight: Most highway mileage values fall between **25–40 mpg**, showing a balanced distribution across vehicle types.

EXPLORE DATA ANALYSIS

Univariate Analysis
Categorical Features:



1. CarName

Insight:

Toyota is the most frequent brand in the dataset, followed by Nissan, Mazda, and Mitsubishi.

Brands such as Mercury, Renault, and Jaguar appear least frequently.

2. Fueltype

Insight:

The majority of cars use gasoline (gas), while only a small portion use diesel.

3. Aspiration

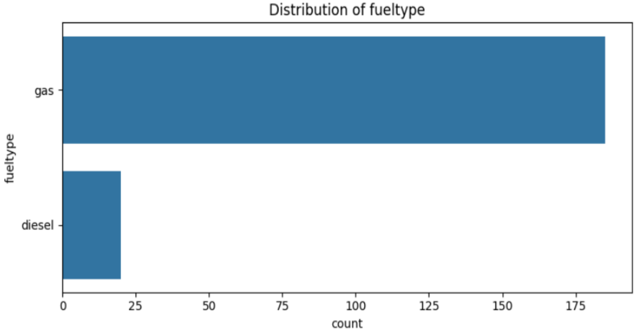
Insight:

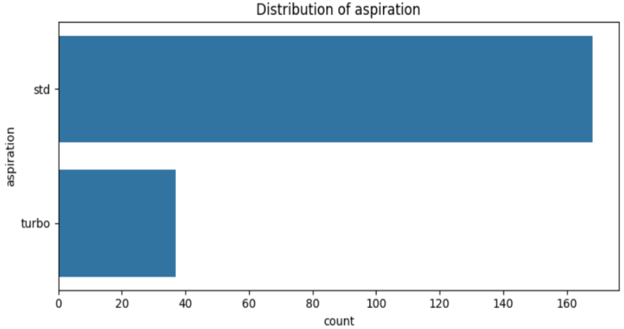
Most vehicles have standard (std) aspiration systems, while a smaller proportion are turbocharged (turbo).

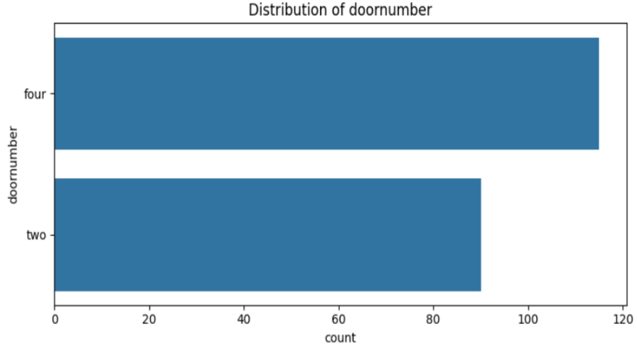
4. Doornumber

Insight:

Cars with four doors slightly outnumber those with two doors.

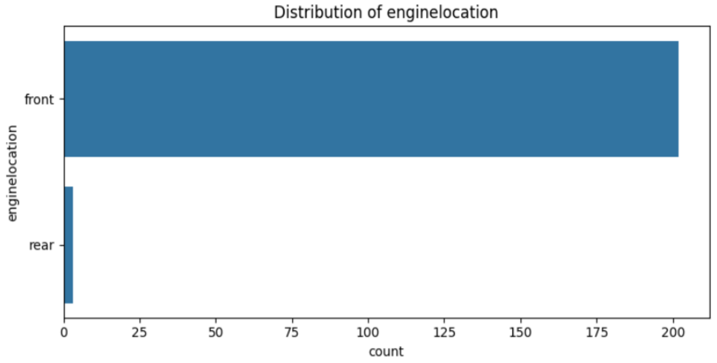


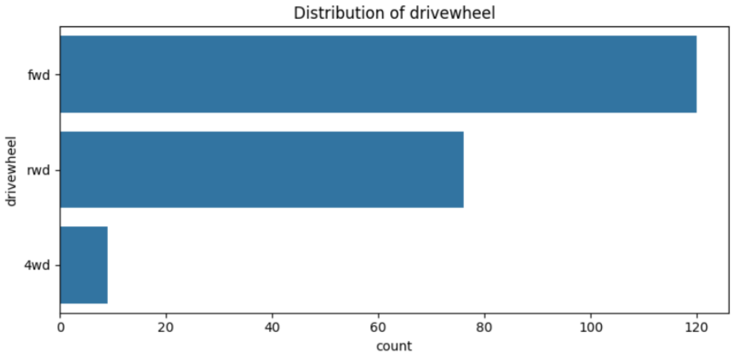


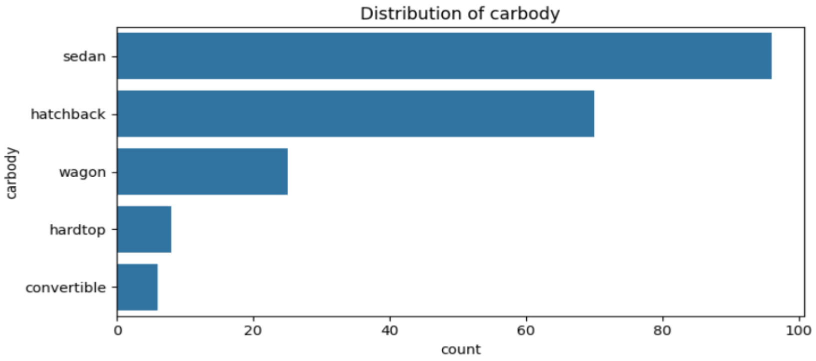


EXPLORE DATA ANALYSIS

Univariate Analysis Categorical Features:







1. Carbody

Insight: The sedan body type is the most common in the dataset, followed by hatchback and wagon.

Hardtop and convertible models are relatively rare.

2. Enginelocation

Insight: Nearly all vehicles have their engine located in the front, with only a very small number having rear engines.

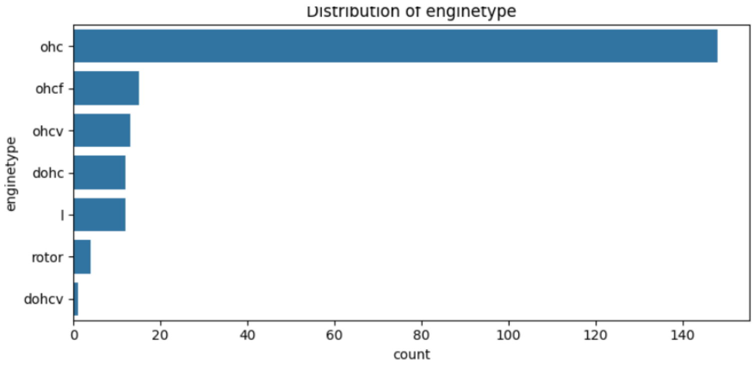
3. Drivewheel

Insight: Front-wheel drive (fwd) cars dominate, followed by rear-wheel drive (rwd).

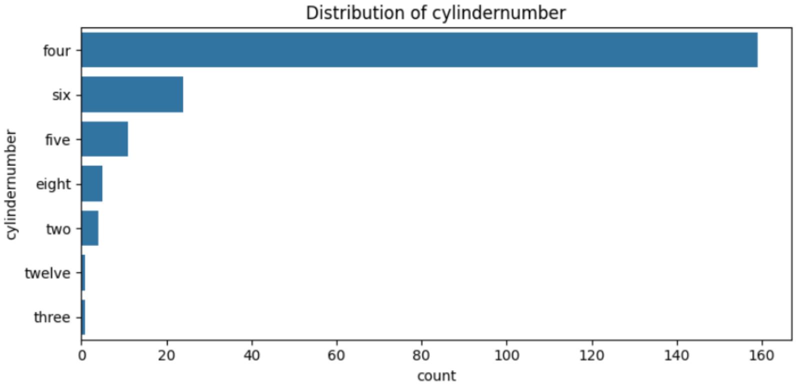
Four-wheel drive (4wd) cars are very few in number.

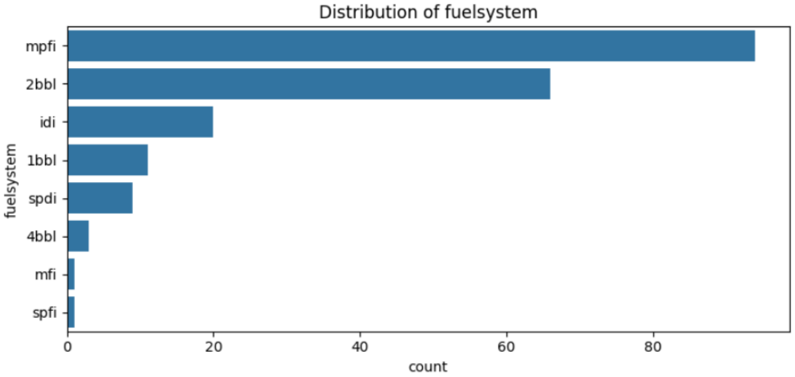
EXPLORE DATA ANALYSIS

Univariate Analysis
Categorical Features:



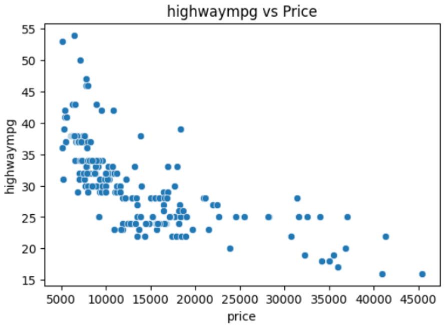
1. Enginetype
Insight: The OHC (Overhead Camshaft) engine type dominates the dataset, followed by OHCF, OHCV, and DOHC types.
Other types such as rotor, L, and DOHCV are very rare.
2. Cylindernumber
Insight: Cars with four cylinders are the most common, making up the majority of the dataset.
Other configurations like six, five, or eight cylinders appear much less frequently.
3. Fuelsystem
Insight: The MPFI (Multi-Point Fuel Injection) system is the most prevalent, followed by 2BBL and IDI.
Other systems like SPFI and MFI occur rarely.

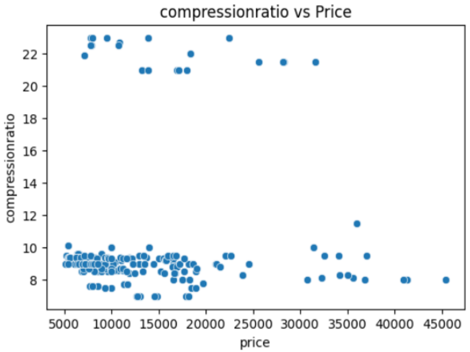


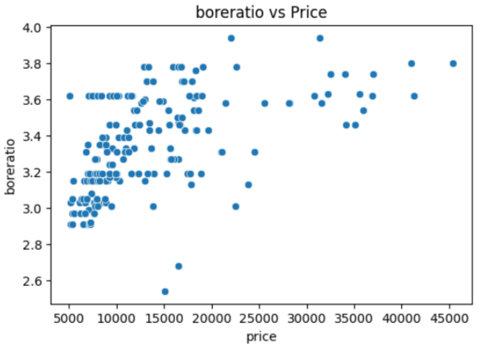


EXPLORE DATA ANALYSIS

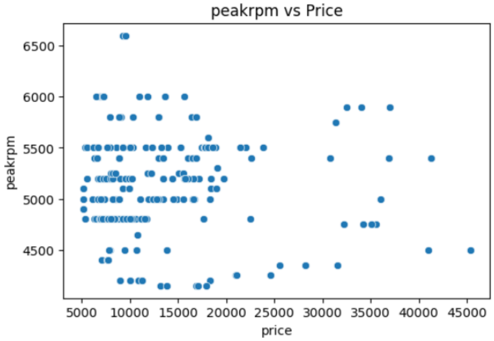
Bivariate Analysis Numerical Features:

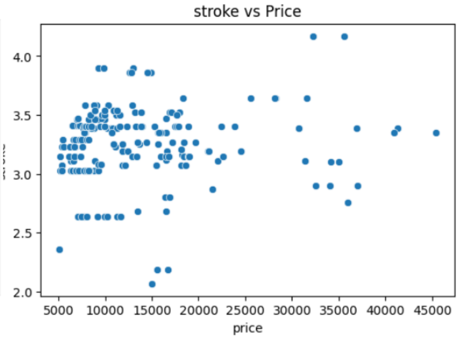

highwaympg vs Price

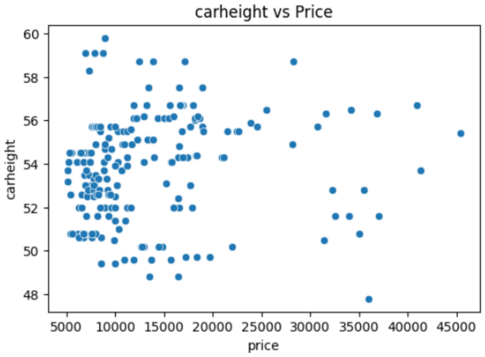

compressionratio vs Price

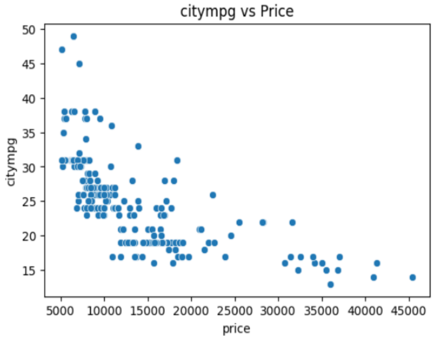

boreratio vs Price

Overall Conclusion:
There is no strong linear relationship between most numerical features and price.
However, a **slight negative correlation** can be observed between **price** and **fuel efficiency variables** such as *citympg* and *highwaympg* — indicating that **more expensive cars tend to have lower fuel efficiency**.
Other features like *compressionratio*, *boreratio*, *peakrpm*, *stroke*, and *carheight* show **weak or no visible correlation** with price.


peakrpm vs Price

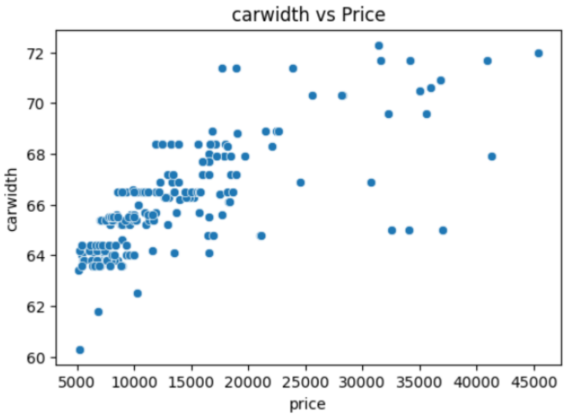

stroke vs Price

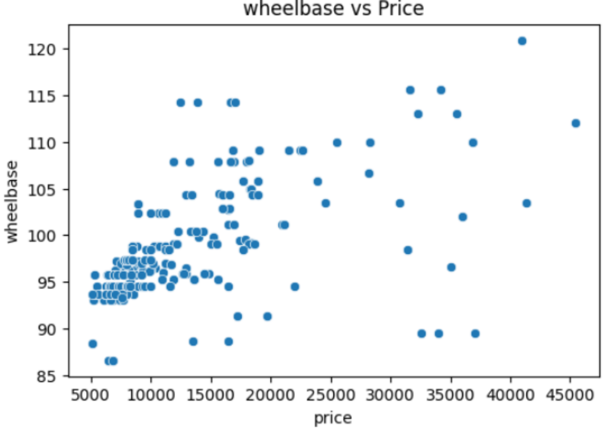

carheight vs Price


citympg vs Price

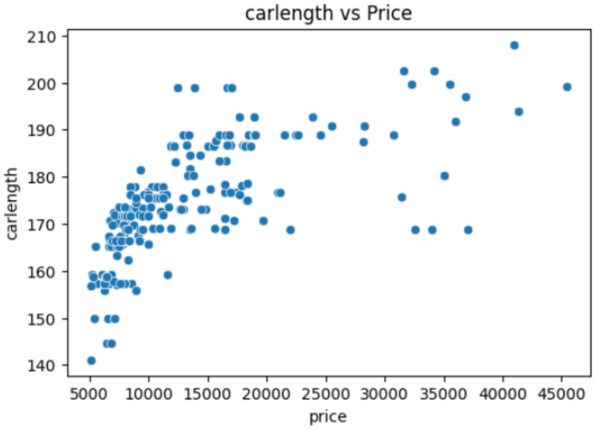
EXPLORE DATA ANALYSIS

Bivariate Analysis
Numerical Features:



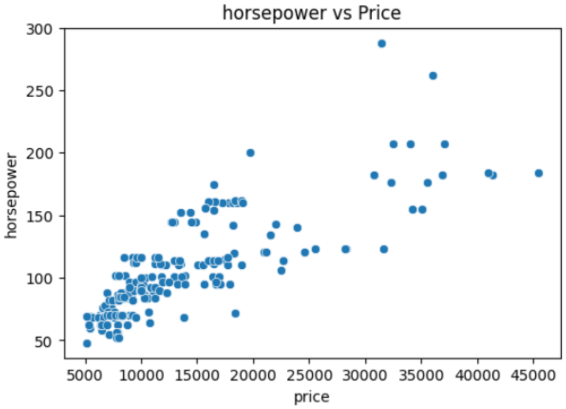


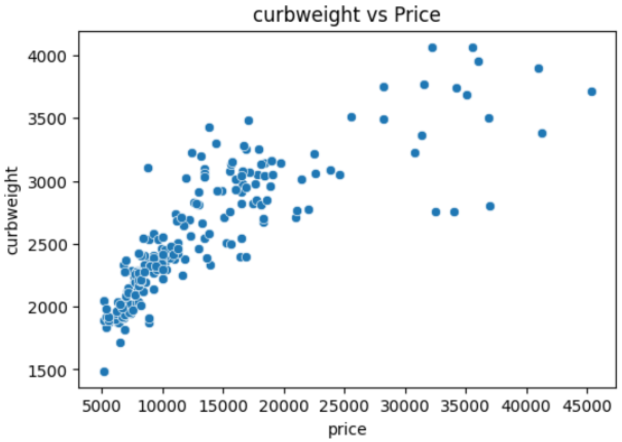
1. Carwidth
Insight: There is a strong positive correlation between *carwidth* and *price*.
Wider cars tend to be more expensive, indicating that *carwidth* is a strong indicator of luxury and performance level.
2. Wheelbase
Insight: *Wheelbase* shows a positive relationship with *price.*
Cars with longer wheelbases generally have higher prices.
3. Carlength
Insight: *Carlength* demonstrates a moderate positive correlation with *price.*
Longer vehicles tend to cost more, though the correlation is not as strong as for *carwidth* or *wheelbase.*



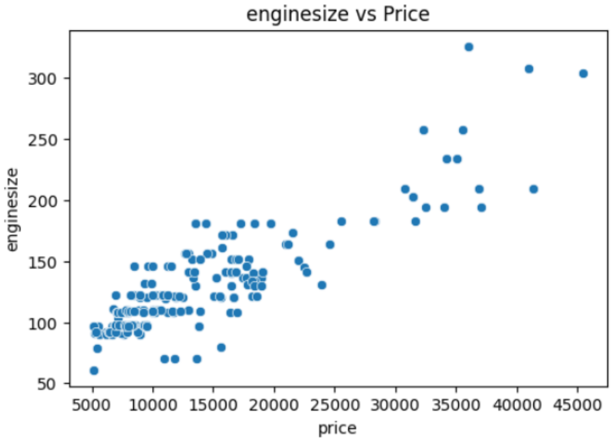
EXPLORE DATA ANALYSIS

Bivariate Analysis
Numerical Features:





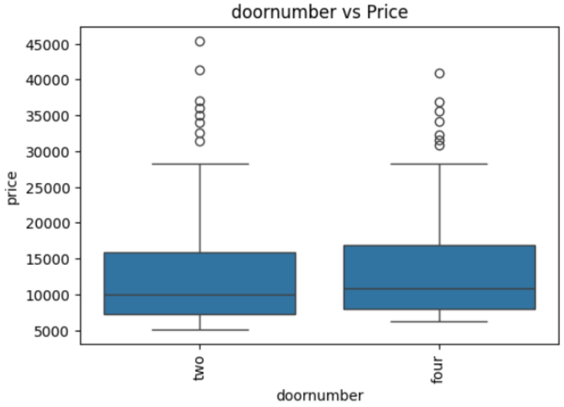
4. Horsepower
Insight: There is a clear positive correlation between *horsepower* and *price.* Cars with higher horsepower tend to have significantly higher prices.
5. Curbweight
Insight: *Curbweight* shows a strong positive relationship with *price.* Heavier cars generally have higher prices.
6. Enginesize
Insight: *Enginesize* has one of the strongest correlations with *price.* Cars with larger engine sizes are consistently more expensive.

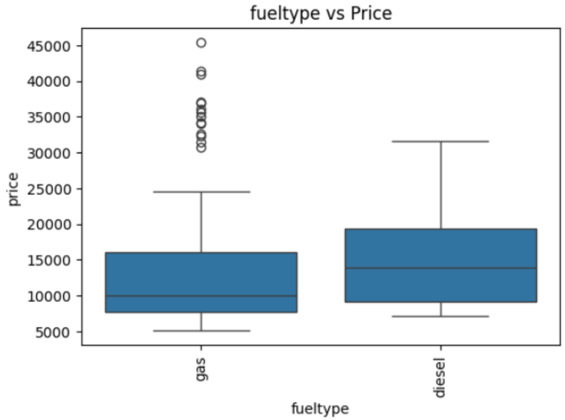


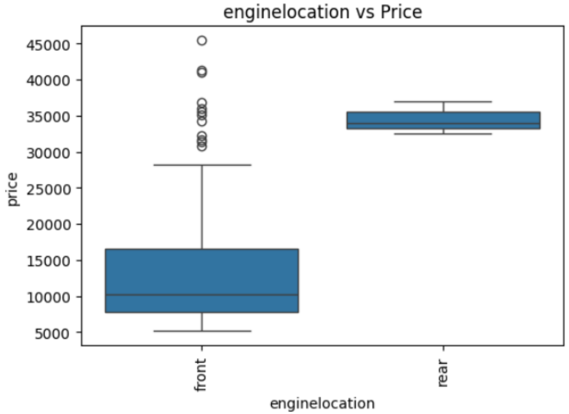
EXPLORE DATA ANALYSIS

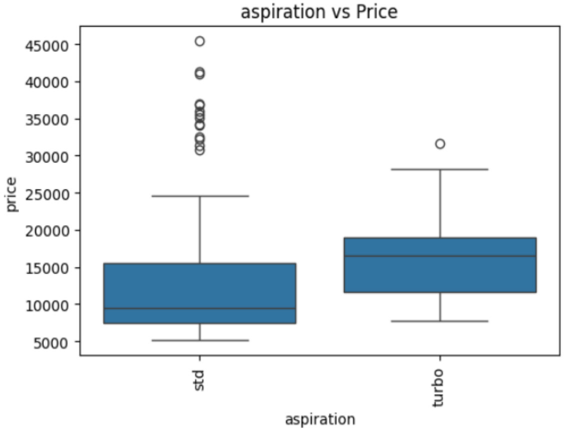
Bivariate Analysis Categorical Features:

1. Doornumber
Insight: There is little difference in price between cars with *two* and *four* doors.
2. Fueltype
Insight: *Diesel* cars generally have slightly higher prices than *gas* cars.
3. Enginelocation
Insight: Cars with *rear* engine placement are significantly more expensive than those with *front* engines.





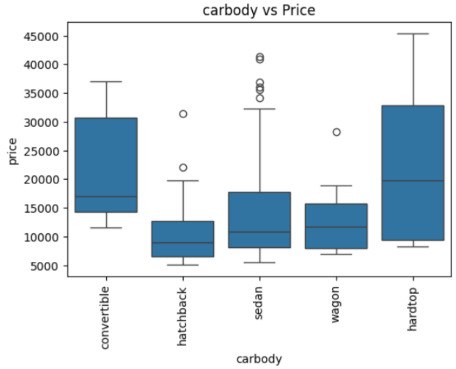


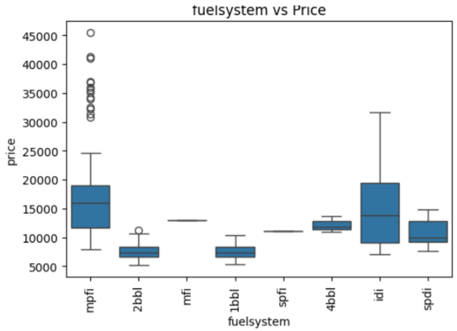


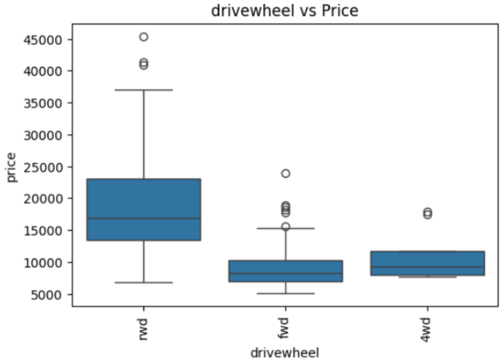
EXPLORE DATA ANALYSIS

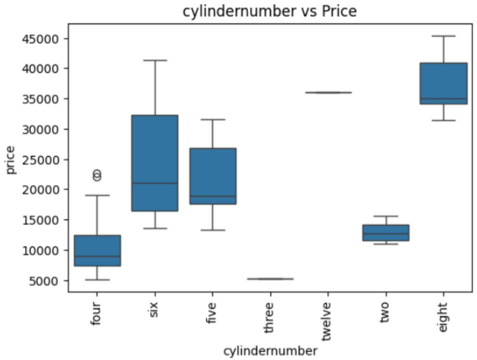
Bivariate Analysis
Categorical Features:

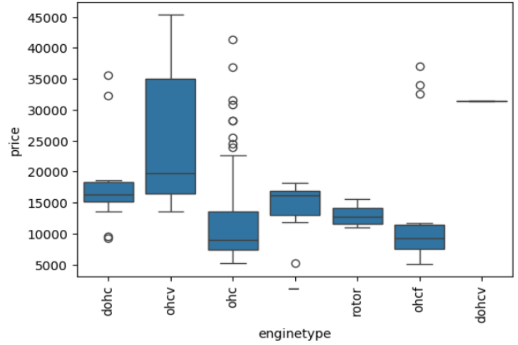
1. Fuelsystem
Insight: The *MPFI* (Multi-Point Fuel Injection) system is associated with higher car prices, while systems like *2bbl*, *1bbl*, and *mfi* correspond to lower-priced models.
2. Carbody
Insight: *Convertible* and *hardtop* cars tend to have the highest prices, while *hatchback* and *sedan* models are more affordable.
3. Drivewheel
Insight: Cars with *rwd* (rear-wheel drive) generally show higher prices than *fwd* or *4wd* vehicles.
4. Cylindernumber
Insight: Cars with *eight* or *twelve* cylinders are significantly more expensive, while *four-cylinder* engines dominate lower price ranges.
5. Enginetype
Insight: *DOHC* (Double Overhead Camshaft) and *DOHCV* engines correspond to higher-priced cars, while *OHCF* and *rotor* types are generally in the lower range.







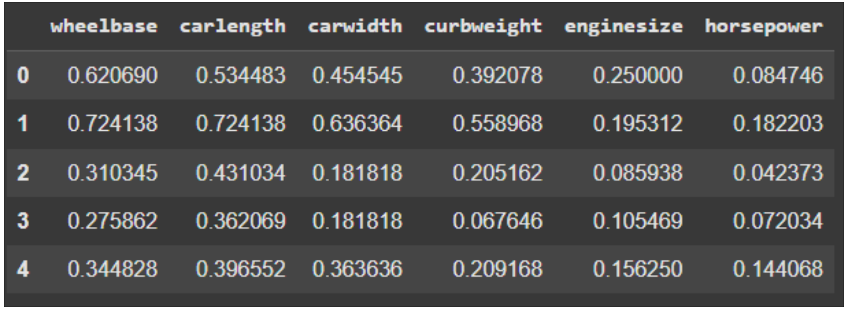




MODEL SELECTION

Data Preprocessing

- Before feeding the data into the model, all features were **encoded** and **scaled** to ensure consistent value ranges.
- The features selected for model training include: **wheelbase**, **carlength**, **carwidth**, **curbweight**, **enginesize**, and **horsepower** — all showing strong positive correlation with **price**, as seen in the heatmap.

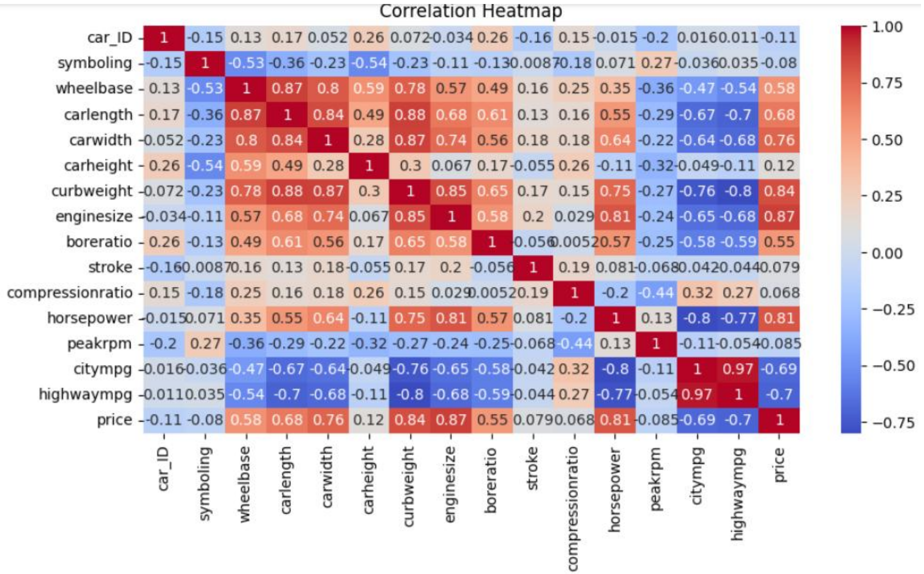


	wheelbase	carlength	carwidth	curbweight	enginesize	horsepower
0	0.620690	0.534483	0.454545	0.392078	0.250000	0.084746
1	0.724138	0.724138	0.636364	0.558968	0.195312	0.182203
2	0.310345	0.431034	0.181818	0.205162	0.085938	0.042373
3	0.275862	0.362069	0.181818	0.067646	0.105469	0.072034
4	0.344828	0.396552	0.363636	0.209168	0.156250	0.144068

Training Process

The dataset was divided into **training (80%)** and **testing (20%)** subsets.

All models were trained using their **default hyperparameters** as the initial setup.



Models Used

- Random Forest Regressor
- CatBoost Regressor
- XGBoost Regressor
- Gradient Boosting Regressor
- Linear Regressor
- LightGBM Regressor

EVALUATION AND RESULT

Model	MAE	MSE	RMSE	R2
Linear Regression	2699	14248520.0	3774.0	0.820
Random Forest Regressor	1308	3807549.0	1951.0	0.952
Gradient Boosting Regressor	1620	7410204.0	2722.0	0.906
XGBRegressor	1577	5458101.0	2336.0	0.931
LGBMRegressor	2228	15193283.0	3897.0	0.808
CatBoostRegressor	1266	5231102.0	2287.0	0.934

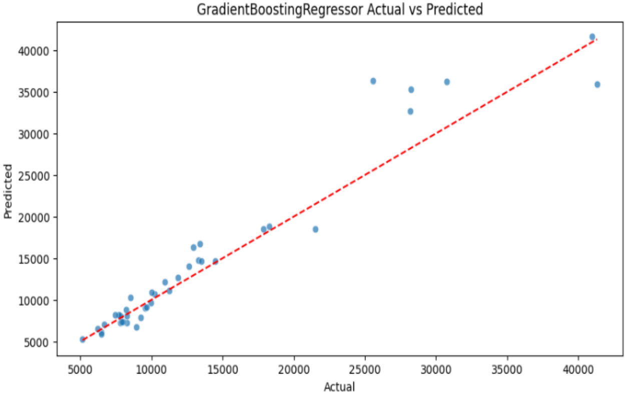
1. Model Comparison

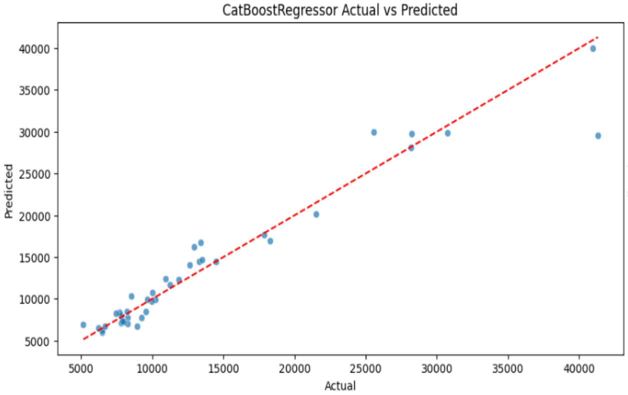
- From the table, several regression models were tested to predict car prices.
  The performance metrics (MAE, MSE, RMSE, R^2) indicate that:
- **Random Forest Regressor** delivers the **best performance** among all models.
- MAE = **1308**, RMSE = **1951.0**, R^2 = **0.952**
- It achieves the **lowest prediction error** and the **highest R^2 score**, meaning it fits the data very well.
- **CatBoost Regressor** also performs strongly, with an R^2 of **0.934**, but slightly below Random Forest.
- **Linear Regression** performs the weakest, showing the highest error and lowest accuracy, suggesting the relationship between features and price is **non-linear**.

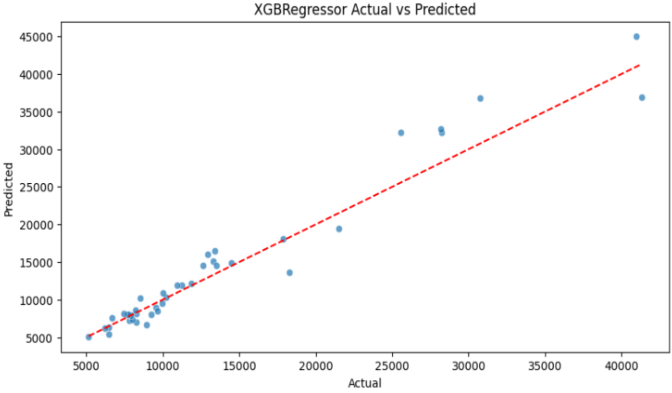
Conclusion

Random Forest Regressor is the **most accurate and reliable model** for predicting car prices in this dataset.

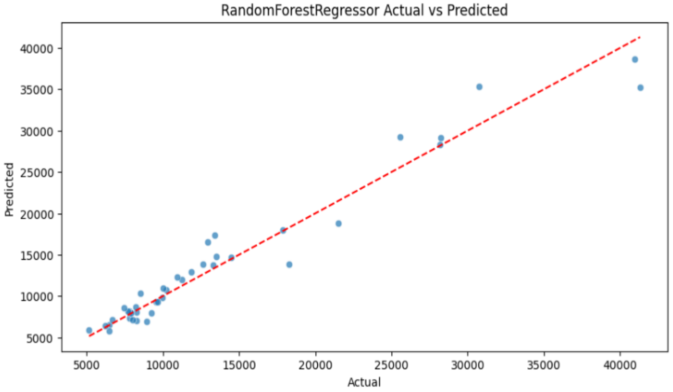
EVALUATION AND RESULT







- **Random Forest Regressor** shows the **closest alignment** between predicted and actual values.
Most of its data points lie **very near the diagonal line**, indicating **high prediction accuracy** and **strong model fit**.
- **CatBoost** and **XGBoost** also show good performance, with data points clustering close to the line, but with **slightly more variance**.
- **Gradient Boosting Regressor** performs well but exhibits **larger deviations** for higher price values.



Conclusion

Random Forest Regressor is the **best-performing model**, offering the **most reliable and consistent predictions** for car price estimation.

EVALUATION AND RESULT

Hyperparameter Tunning Results:

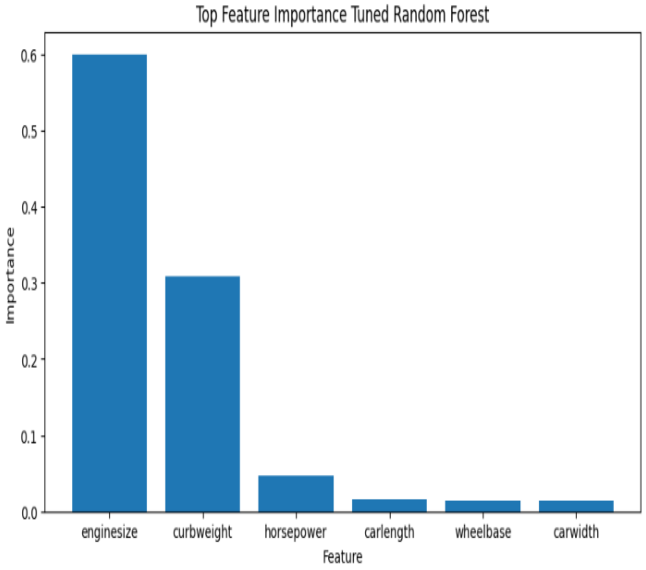
Model	MAE	MSE	RMSE	R2
Random Forest Regressor	1421	4314052	2077	0.945
Gradient Boosting Regressor	1526	4874845	2207	0.938
XGBRegressor	1705	5964858	2442	0.924
CatBoostRegressor	1260	4655855	2157	0.94

After hyperparameter tuning:
Most models showed **better balance between bias and variance**.
Random Forest still achieved one of the **best R² scores (0.945)** and **lowest RMSE**, confirming its **consistency and reliability**.
CatBoost emerged as a strong competitor with a slightly higher R² than Gradient Boosting and XGB.

Conclusion

- **Random Forest Regressor** remains the **most robust and accurate model overall**, even after tuning.
- **Gradient Boosting Regressor** benefited the most from tuning and now performs comparably.
- **CatBoost Regressor** also showed reliable improvement with strong generalization.

EVALUATION AND RESULT



Insight:

- **Enginesize** is by far the most influential feature, contributing around **60%** of the model's predictive power.
- **Curbweight** is the second most important feature, with about **30%** importance.
- Other features such as **horsepower**, **carlength**, **wheelbase**, and **carwidth** have minimal impact on predicting car prices.

Interpretation:

- The results indicate that **engine size** and **vehicle weight** are the key determinants of car price — larger engines and heavier cars tend to correlate with higher prices.
- Features like **horsepower** and **car dimensions** (length, width, wheelbase) have comparatively smaller effects, meaning price variations are less sensitive to these factors when other main attributes are known.
- The **Random Forest model** effectively identifies these dominant factors, confirming that performance and structural size are critical price drivers in this dataset.