FINAL TEST

ĐỀ 1:

Car Price Prediction

Học viên: Nguyễn Công Trúc

OVERVIEW

Bối cảnh

Một công ty ô tô Trung Quốc Geely Auto muốn thâm nhập thị trường Mỹ bằng cách thành lập đơn vị sản xuất của họ ở đó và sản xuất ô tô đểcạnh tranh với các đối tác Mỹ và châu Âu.

Do vậy, họ đã ký hợp đồng với một công ty tư vấn ô tô để hiểu các yếu tố ảnh hưởng đến việc định giá ô tô tại thị trường Mỹ, vì những yếu tố đó có thể rất khác so với thị trường Trung Quốc.

Vai trò

Tôi là một nhà phân tích dữ liệu cho công ty tư vấn ô tô, vai trò của tôi phải:

- Hiểu rõ vễ việc những yếu tố ảnh hưởng đến việc định giá ô tô tại thị trường Mỹ.

- Tìm hiểu xem có thể đưa ra mô hình dự đoán giá xe tại thị trường Mỹ.

Mục tiêu

Xây dựng mô hình giá xe hướng tới mục tiêu:

- Yếu tố quan trọng ảnh hưởng tới giá xe.

- Phát triển mô hình dự đoán giá.

- Phân tích giá theo phân khúc thị trường

TABLE OF CONTENTS

01

Discovery

Khám phá và hiểu dữ liệu

02

Analyze

Phân tích dữ liệu để trả lời các câu hỏi cụ thể

03

Model Prediction

sử dụng các mô hình học máy để dự đoán các giá trị trong tương lai.

04

Suggestion

cung cấp các gợi ý và khuyến nghị dựa trên dữ liệu

01

Discovery

Khám phá và hiểu dữ liệu giá xe.

INFORMATION ABOUT DATASET

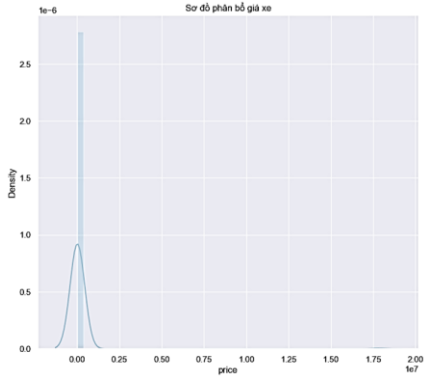
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):

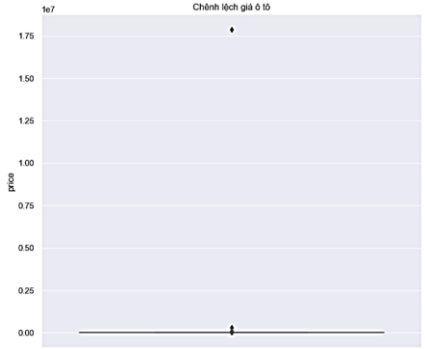
#	Column	Non-Null Count	Dtype
0	car_ID	205 non-null	int64
1	symboling	205 non-null	int64
2	CarName	205 non-null	object
3	fueltype	205 non-null	object
4	aspiration	205 non-null	object
5	doornumber	205 non-null	object
6	carbody	205 non-null	object
7	drivewheel	205 non-null	object
8	enginelocation	205 non-null	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)

Bộ dữ liệu gồm có: 26 cột, 205 dòng không có giá trị Null

COLUMNS DISTRIBUTION



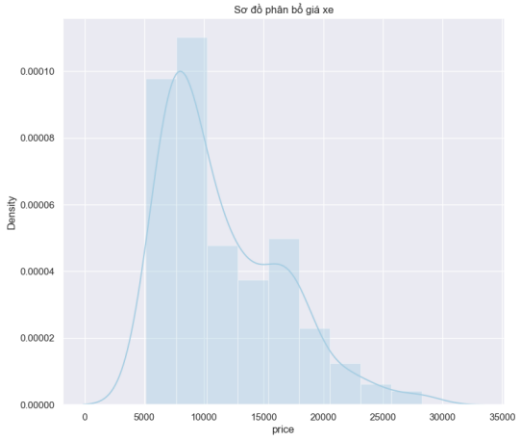


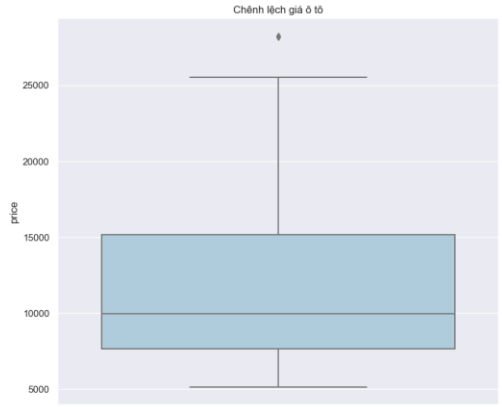
	price
count	2.050000e+02
mean	1.024690e+05
std	1.246484e+06
min	5.118000e+03
25%	7.788000e+03
50%	1.059500e+04
75%	1.655800e+04
max	1.785917e+07

- Dựa trên giá trung bình và độ lệch chuẩn cao, ta có thể thấy rằng giá cả sản phẩm phân bố không đều.
- Có một số ít sản phẩm có giá rất cao so với phần lớn các sản phẩm khác.
- Điều này khiến đồ thị phân bố không được rõ ràng .

=> Quyết định xử lý outlier để cải thiện phân bố trong cột price.

COLUMN DISTRIBUTION





Sau khi xử lý outliers:

Chúng ta có thể thấy rằng có (205 - 187)=28 bản ghi trong cột price là các giá trị outliers trong tập dữ liệu.

```
In [205]: df.shape
Out[205]: (187, 25)
```

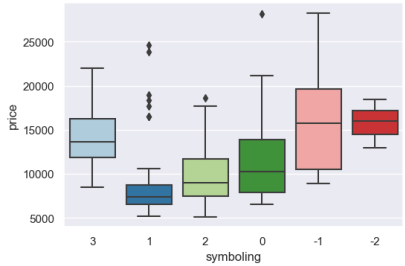
Nhận xét:
Các cột có sự phân bố lệch về trái .
=> Điều này cho thấy hầu hết giá xe hầu hết phân bố trong khoảng từ 5000$ đến 15000$ và khoảng trên 15000$ có sự tập trung ít hơn trên thị trường.

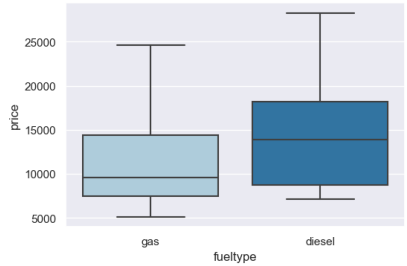
02

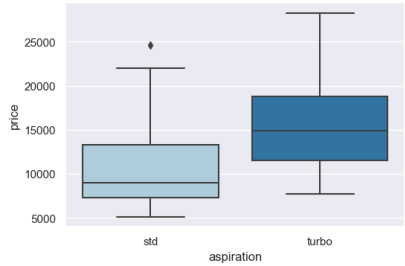
Analyze

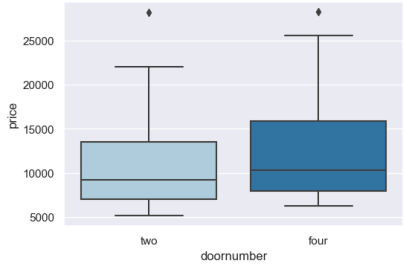
Phân tích dữ liệu để xác định các yếu tố ảnh hưởng đến giá xe.

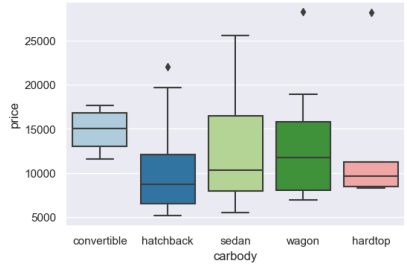
BUILDING HYPOTHESIS

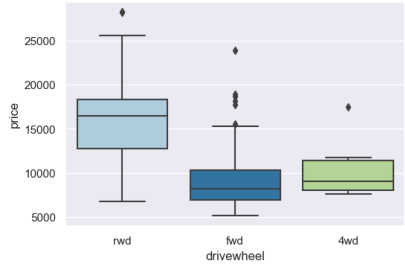


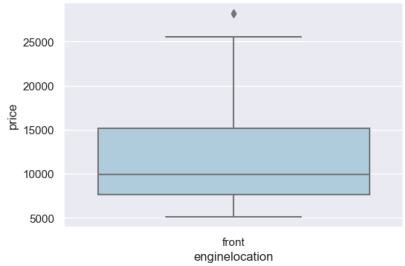


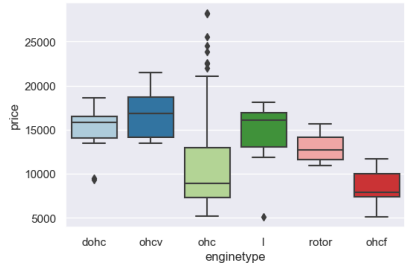


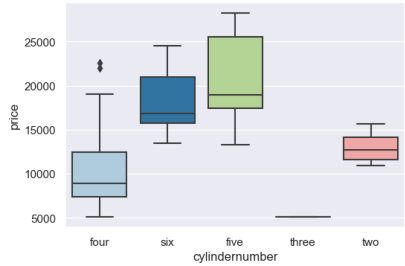


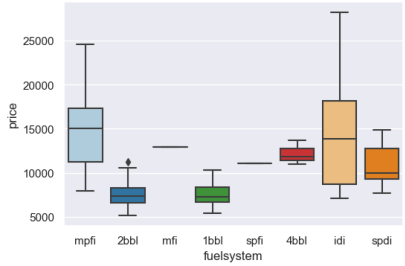






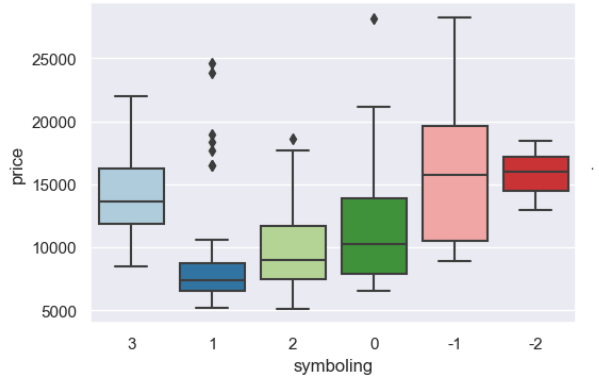


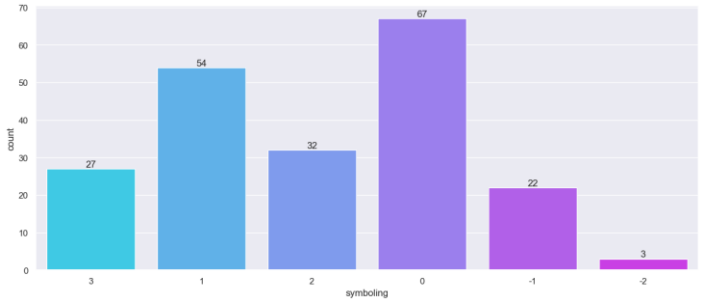




So với các thông số, giá xe nó được ảnh hưởng bởi những yếu tố liên quan như ...
Nhưng nhìn chung thì lại không có sự chênh lệch quá nhiều tại các labels.
=>
Nhận định: Không có sự ảnh hưởng đáng kể ở những cột thành phần ngoại trừ symboling,cylindernumber. enginetype và fuelsystem.
Giả thuyết:
Giá trị của ô tô càng cao thì sẽ có mức độ rủi ro thấp.
Giá xe tỷ lệ thuận với tính phổ biến,giá cả của những linh kiện trên xe.

VALIDATING HYPOTHESIS





Dẫn chứng:

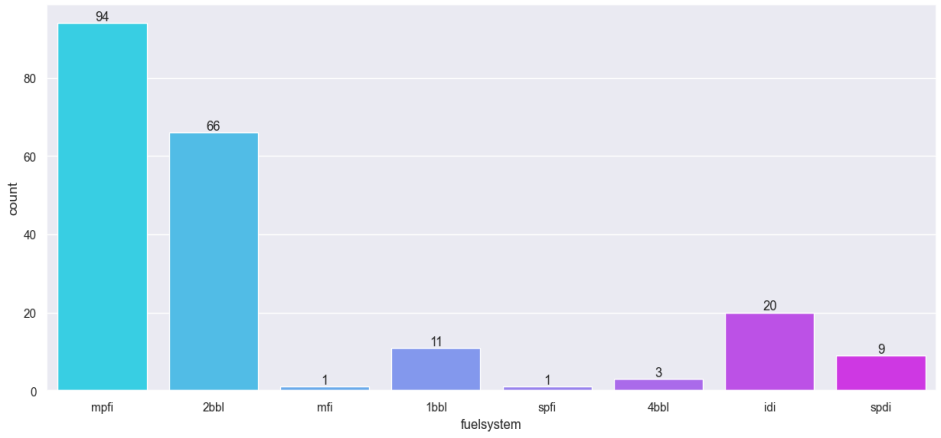
•Trong dữ liệu, mặc dù chỉ có 3 xe có mức độ an toàn ở mức -2 (an toàn) nhưng giá lên đến 15.000$. Điều này có thể giải thích do đây là những xe có Symboling cao, thiết kế được kiểm chứng và ít rủi ro hơn, dẫn đến mức giá cao hơn.

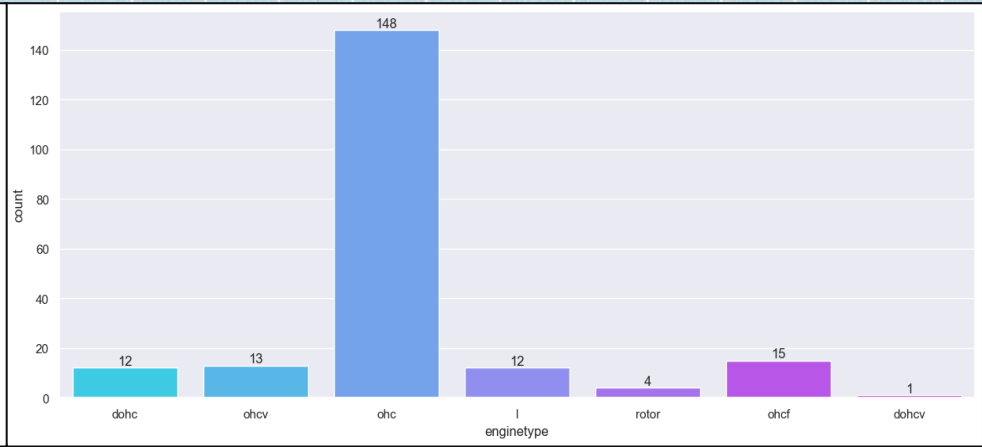
•Ngược lại, có nhiều xe ở mức Symboling 0 và giá cả ở mức trung bình. Điều này có thể do những xe này có thiết kế phổ biến hơn, tiềm ẩn nhiều rủi ro tiềm ẩn hơn, dẫn đến mức giá thấp hơn.

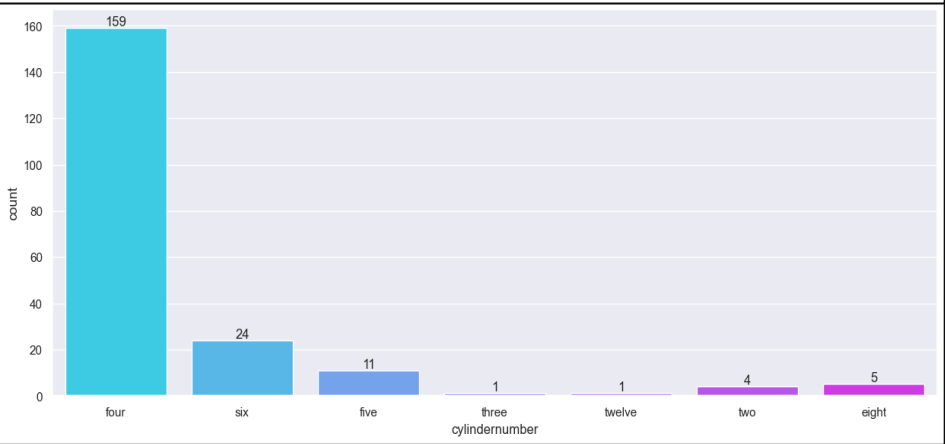
Xe có Symboling cao:Thường có thiết kế tương tự như các mẫu xe khác của cùng nhà sản xuất, đã được kiểm chứng về độ an toàn qua quá trình sản xuất và sử dụng.Ít có khả năng gặp sự cố do lỗi thiết kế hoặc sản xuất.Do đó, được các công ty bảo hiểm đánh giá rủi ro thấp hơn và xếp hạng an toàn hơn.

Xe có Symboling thấp:Thường có thiết kế độc đáo, khác biệt so với các mẫu xe khác.Ít có dữ liệu về độ an toàn trong thực tế do số lượng sản xuất ít hơn.Do đó, tiềm ẩn nhiều rủi ro tiềm ẩn do lỗi thiết kế hoặc sản xuất.Được các công ty bảo hiểm đánh giá rủi ro cao hơn và xếp hạng rủi ro hơn.

VALIDATING HYPOTHESIS







- **Số lượng xi-lanh (cylindernumber):** 4 và 5
- **Loại động cơ (enginetype):** ohc và ohcv
- **Hệ thống nhiên liệu (fuelsystem):** mpfi và idi

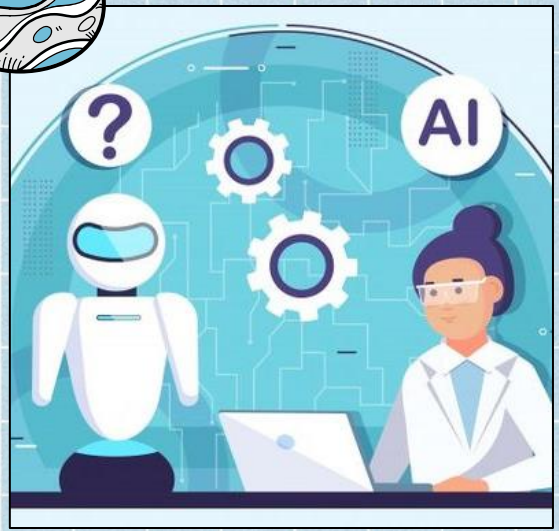
Đánh giá:

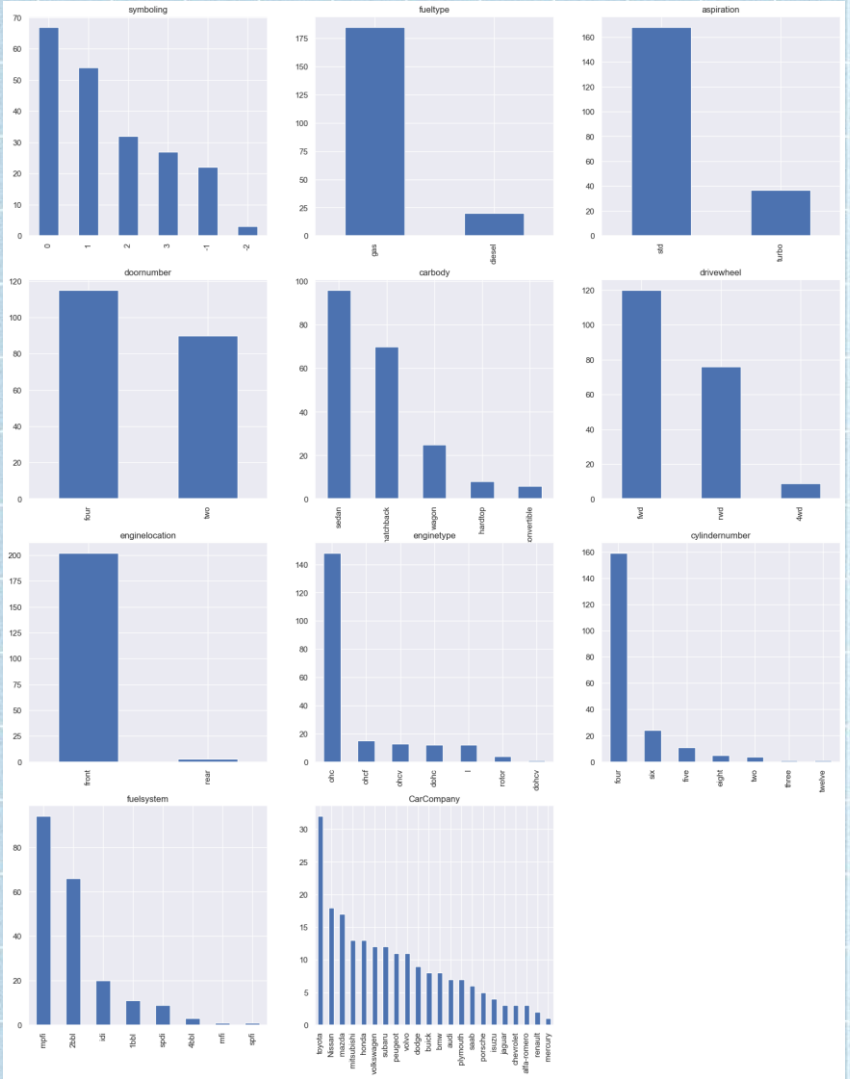
- **Số lượng xi-lanh:** Chi phí xe giá cao thường thể thấy xe đó có 4 xi-lanh,5 xi-lanh bởi vì ta nhìn vào 4 xi-lanh được sử dụng khá phổ biến và 5 xi-lanh được phát triển để tối ưu công xuất nên thường chi phí sẽ cao hơn.
- **Loại động cơ:** tương tự ohc là động cơ sử dụng nhiều lên đến 148 xe và ohcv là động cơ được phát triển tiên tiến về công xuất lên dù số lượng ít nhưng giá thành vẫn cao.
- **Hệ thống nhiên liệu:** tương tự ở trên mpfi sẽ có sự phổ biến và idi có sự phát triển nên có thì giá của những xe sử dụng hệ thống này cao.

03

Model Prediction

Sử dụng các mô hình học máy để dự đoán giá xe cho các mẫu xe trong thời gian tới.



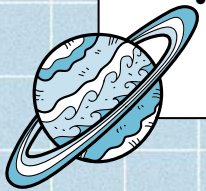


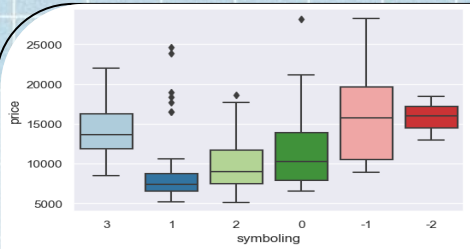
DATA VALIDATION AND PREPROCESSING

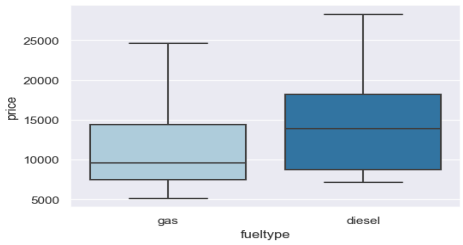
CHECK BALANCE

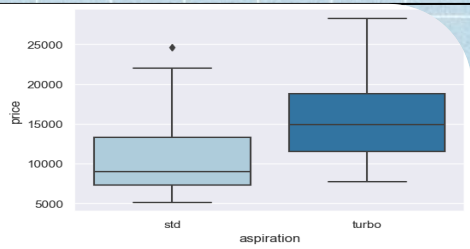
Dựa trên phân tích dữ liệu, chúng ta thấy có sự mất cân bằng dữ liệu trong các cột sau:

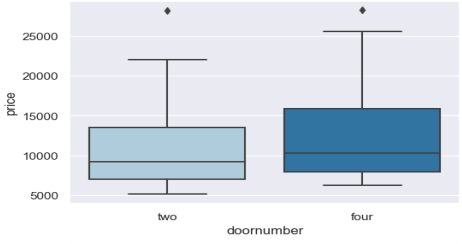
- **symboling:** Rất ít xe có xếp hạng -2.
- **fuletype:** Tất cả các xe đều có loại nhiên liệu là Xăng (Gas), do xe Diesel đã bị loại bỏ trong quá trình xử lý các điểm bất thường (outlier).
- **aspiration:** Số xe sử dụng công nghệ turbo ít hơn so với loại tiêu chuẩn (std).
- **engineloaction:** Vị trí của tất cả các động cơ đều ở phía trước, do tất cả các xe động cơ đặt sau đã bị loại bỏ trong quá trình xử lý outlier.
- **enginetype:** Số xe sử dụng kiểu động cơ ohc (trục cam đôi trên đỉnh) đáng kể hơn so với các loại khác.
- **cylindernumber:** Số xe 4 xi-lanh nhiều hơn đáng kể so với các loại khác.
- **fuelsystem:** Xe sử dụng hệ thống nhiên liệu mpfi và 2bbl phổ biến hơn so với các loại khác.
- **CarCompany:** Hầu hết các xe được khảo sát thuộc thương hiệu Toyota.

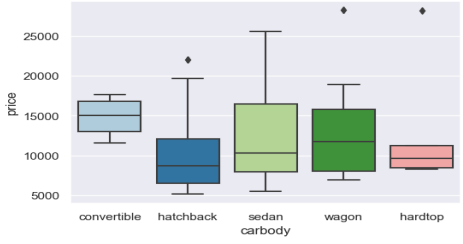


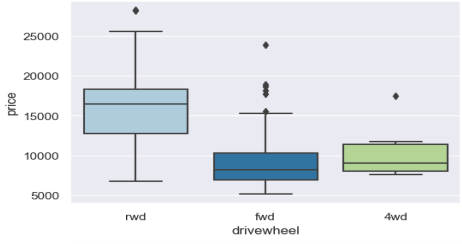


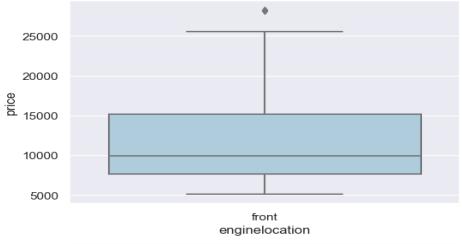


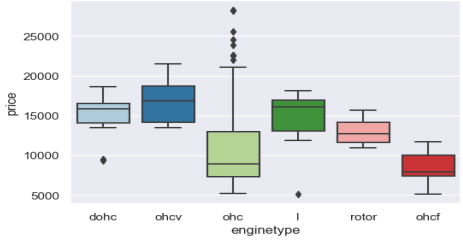


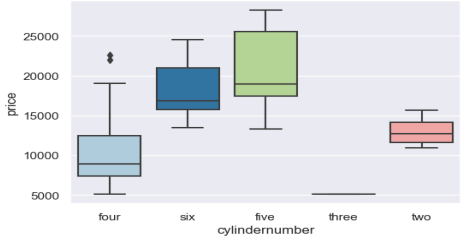


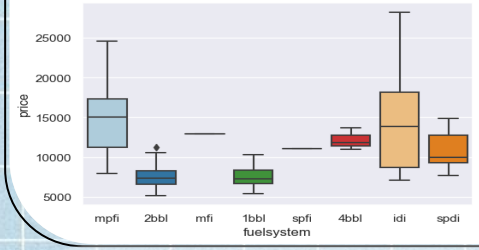












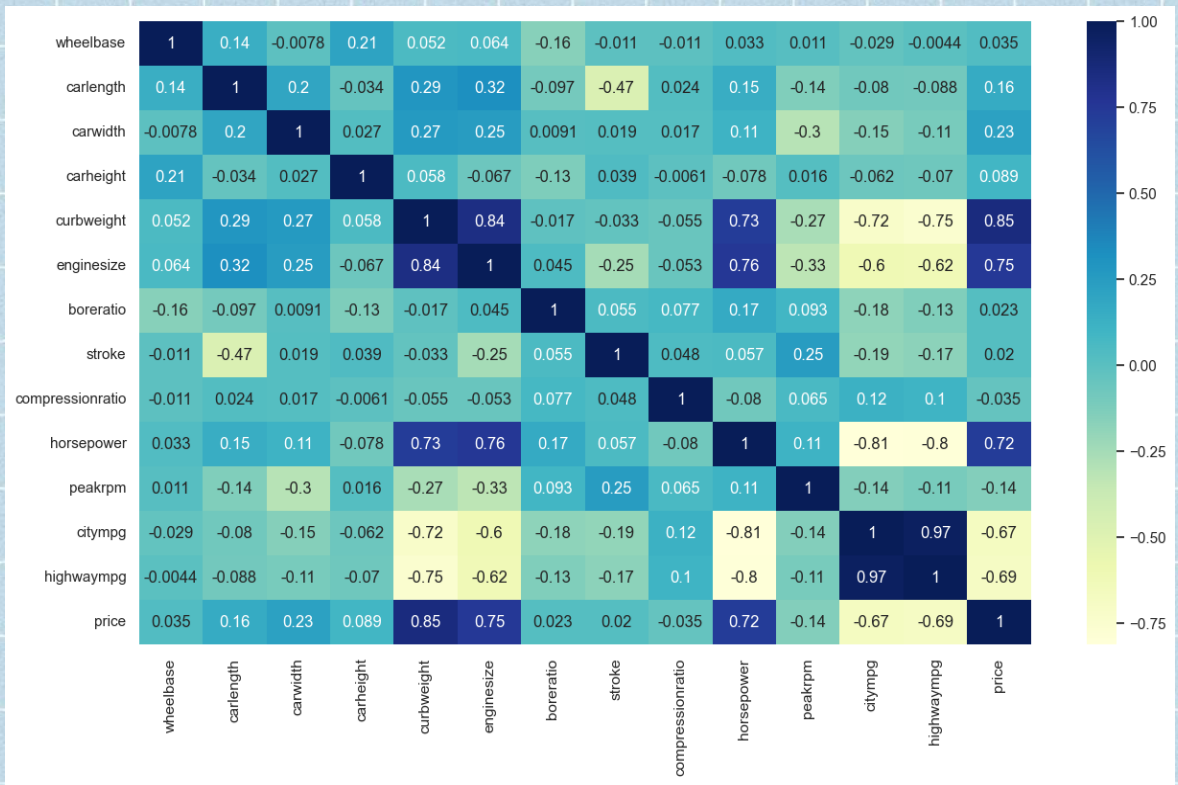
Data Validation and Preprocessing Check Outlier

Dữ liệu không có quá nhiều giá trị Outlier ở các cột fuelstype, aspiration, doornumber,…

Giá trị Outlier xuất hiện ở những cột symboling, enginetype, driverwheel nhưng không đáng kể bởi những diểm này không nằm quá xa vùng phân bố cho phép.

=> Giá trị của xe không bị ảnh hưởng nhiều đến chênh lệch quá cả.

DATA VALIDATION AND PREPROCESSING
CHECKCORRELATION AND FEATURESSELECTION

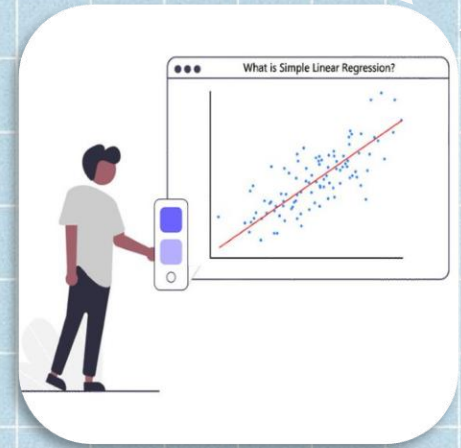


Price (Giá): Tương quan dương nhẹ với Wheelbase, Carlength , Curbweight, Enginesize và Horsepower . Điều này cho thấy rằng xe có giá cao hơn thường có chiều dài cơ sở dài hơn, chiều dài, trọng lượng, kích thước động cơ và công suất cao hơn.
Chỉ có một số labels như citympg, highwaympg có hệ số tương quan rất thấp.

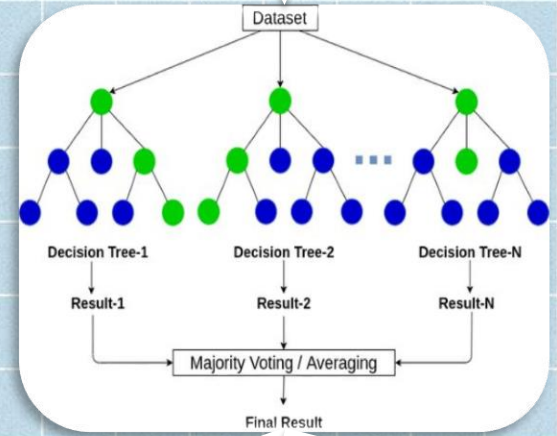
Chúng ta có thể thấy rằng hầu hết các biến rất yếu tương quan với nhãn. Chỉ có một số trường hợp tương quan thấp.
Nhưng có khả năng sử dụng được tất cả cho model.

Building model to predict car price

Models

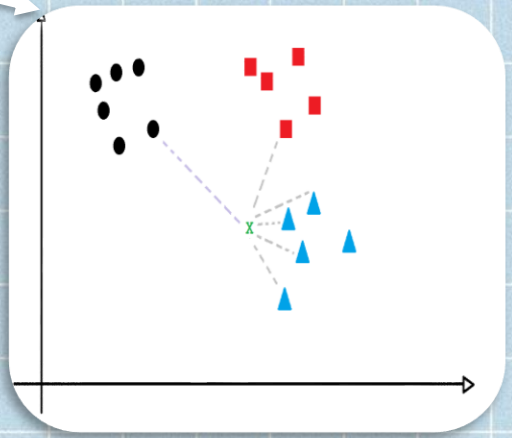


Linear Regression



Decision Tree

Random Forest



KNeighbors Regressor

Building model to predict car price

Kết quả	MSE	R2_Test
Linear Regression	0.14	0.85
Decision Tree	0.13	0.86
Random Forest	0.13	0.87
KNeighbors Regressor	0.20	0.69

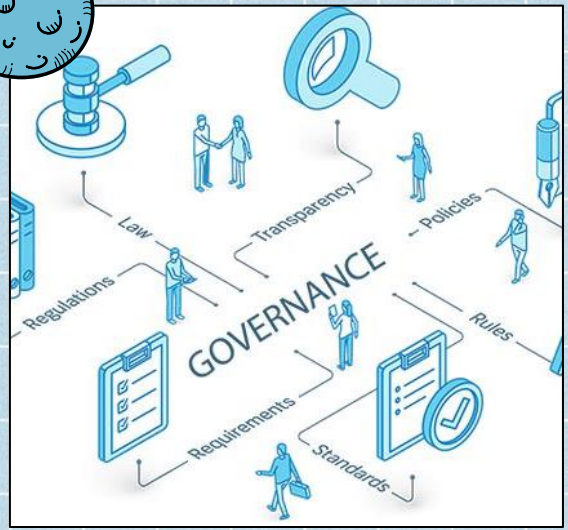
Xét về độ chính xác:

- **Random Forest** có hiệu suất tốt nhất với giá trị RMSE thấp nhất (0.13) và R-squared cao nhất (0.87).
- **Decision Tree** xếp ở vị trí thứ hai với RMSE là 0.13 và R-squared là 0.86.
- **Linear Regression** và **KNeighbors Regressor** có hiệu suất thấp hơn với RMSE lần lượt là 0.14 và 0.20, và R-squared lần lượt là 0.85 và 0.69.

04

Suggestion

Cung cấp các gợi ý và khuyến nghị cho người mua xe dựa trên nhu cầu và ngân sách của họ.



Suggestion

Dự đoán giá xe ô tô bằng mô hình học máy: Giải pháp cho thị trường minh bạch và hiệu quả

- Thị trường xe ô tô thường xuyên biến động với nhiều mức giá khác nhau khiến người mua gặp khó khăn trong việc lựa chọn và so sánh giá cả. Nhằm giải quyết vấn đề này, việc ứng dụng mô hình học máy vào việc dự đoán giá xe ô tô mang lại giải pháp tiềm năng cho thị trường minh bạch và hiệu quả hơn.
- Thông qua việc thu thập dữ liệu chi tiết về thuộc tính xe từ nhiều nguồn uy tín, mô hình học máy được xây dựng và huấn luyện để dự đoán giá xe chính xác. Mô hình này mang lại nhiều lợi ích thiết thực cho cả người mua xe và đại lý xe:

Đối với người mua xe:

- Dễ dàng so sánh giá xe từ các nguồn khác nhau, tránh mua xe giá cao.
- Lựa chọn mua xe phù hợp với nhu cầu và ngân sách.
- Tiết kiệm thời gian và công sức trong quá trình tìm kiếm và mua xe.

Đối với đại lý xe:

- Định giá xe cạnh tranh và hợp lý hơn, thu hút khách hàng tiềm năng.
- Tăng doanh số bán hàng và lợi nhuận.
- Giảm chi phí hoạt động và quản lý.

Việc triển khai mô hình học máy dự đoán giá xe ô tô sẽ góp phần tạo dựng thị trường xe minh bạch, hiệu quả và lành mạnh hơn, mang lại lợi ích cho cả người mua và người bán.

Thank you