BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH

TRƯỜNG CÔNG NGHỆ VÀ THIẾT KẾ

ĐỒ ÁN MÔN HỌC

ĐỀ TÀI:

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN ĐỊNH GIÁ XE HƠI Ở BA LAN

Học phần: Lập trình phân tích dữ liệu

Nhóm 9:

1. Đào Thị Phương Quỳnh
2. Văn Ngọc Như Quỳnh
3. Nguyễn Thị Phương Thảo
4. Lý Gia Thuận

Chuyên Ngành: KHOA HỌC DỮ LIỆU
Khóa: K47
Giảng Viên: TS. Nguyễn An Tế

TP. Hồ Chí Minh, Ngày 03 tháng 12 năm 2023

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1 Giới thiệu bài toán

Đề tài này phân tích dữ liệu về các đặc điểm xe ô tô tại Ba Lan được tổng hợp trên một trang quảng cáo. Trong quá trình mua bán xe ô tô, vấn đề quan trọng nhất chính là cân nhắc kỹ lưỡng để không đưa ra một mức giá cao hơn (hoặc thấp hơn) giá trị thực của chiếc xe. Nếu đưa ra một mức giá cao hơn so với giá thị trường, khả năng bán được chiếc xe đó sẽ giảm đi hoặc làm kéo dài thời gian bán được xe. Bên cạnh đó, người tiêu dùng sẽ không chọn mua một chiếc xe đã qua sử dụng mà lại có giá cao hơn giá trị thực. Dựa vào phân tích các dữ liệu về đặc điểm của xe như mẫu xe, năm sản xuất, tổng quãng đường đã đi, màu sắc,... các doanh nghiệp sẽ có cái nhìn tổng quát về công việc kinh doanh và từ đó đưa ra những chiến lược phù hợp.

1.2 Mục tiêu nghiên cứu

Sử dụng để dự đoán giá xe và phân tích ảnh hưởng của các yếu tố khác nhau trong những đặc điểm cụ thể của từng loại xe. Từ đó có thể giúp các doanh nghiệp trong ngành ô tô tìm hiểu về xu hướng thị trường, tìm ra thông tin quan trọng về sản phẩm và người tiêu dùng, cũng như đưa ra quyết định kinh doanh phù hợp. Đồng thời, có thể giúp cho người tiêu dùng đưa ra quyết định để lựa chọn loại xe phù hợp với bản thân.

1.3 Phương pháp nghiên cứu

- Phân tích các biến để thấy được sự tương quan với nhau.
- Sử dụng các mô hình hồi quy để dự đoán giá xe.

1.4 Tài nguyên sử dụng

- Bộ dữ liệu Poland Cars For Sale lấy từ Kaggle.

1.5 Ngôn ngữ sử dụng:

- Ngôn ngữ lập trình phân tích dữ liệu Python.

CHƯƠNG 2: TỔNG QUAN BỘ DỮ LIỆU

2.1 Sơ lược về bộ dữ liệu:

Bộ dữ liệu chứa các mẫu quảng cáo xe ở Ba Lan, với số thuộc tính là 25 và số dòng là 208.304 dòng.

2.2 Mô tả thuộc tính của bộ dữ liệu:

STT	Tên thuộc tính	Mô tả	Loại biến	Ghi chú
1	Index	ID của mẫu quảng cáo	Categorical	
2	Price	Giá của từng mẫu xe	Numerical	
3	Currency	Đơn vị tiền tệ	Categorical	2 giá trị: EUR và PLN
4	Condition	Tình trạng xe	Categorical	2 giá trị: New và Used
5	Vehicle_brand	Thương hiệu xe	Categorical	
6	Vehicle_model	Mẫu xe	Categorical	
7	Vehicle_version	Phiên bản xe	Categorical	
8	Vehicle_generation	Thế hệ xe	Categorical	
9	Production_year	Năm sản xuất	Categorical	
10	Mileage_km	Số km mà xe đã đi	Numerical	
11	Power_HP	Công suất động cơ của xe, với đơn vị đo là mã lực	Numerical	
12	Displacement_cm3	Dung tích xi lanh của động cơ xe, với đơn vị đo là cm3	Numerical	
13	Fuel_type	Loại nhiên liệu mà xe sử dụng	Categorical	8 giá trị unique

14	CO2_emissions	Lượng khí thải CO2 của xe, với đơn vị đo là g/km	Numerical	
15	Drive	Hệ thống dẫn động của xe	Categorical	5 giá trị unique
16	Transmission	Loại hộp số của xe	Categorical	2 giá trị: Manual và Automatic
17	Type	Kiểu dáng của xe	Categorical	9 giá trị unique
18	Doors_number	Số cửa mà xe có	Numerical	
19	Colour	Màu sắc của xe	Categorical	
20	Origin_country	Nguồn gốc của xe	Categorical	
21	First_owner	Liệu chủ xe có phải là người chủ đầu tiên không?	Categorical	1 giá trị: Yes
22	First_registration_date	Ngày đăng ký xe đầu tiên	Categorical	
23	Offer_publication_date	Ngày đăng tin quảng cáo	Categorical	
24	Offer_location	Địa chỉ được cung cấp bởi nhà phát hành	Categorical	
25	Features	Danh sách các đặc điểm khác (ABS, airbag, parking sensors…)	Categorical	

CHƯƠNG 3: TIỀN XỬ LÝ DỮ LIỆU

3.1 Mô tả tổng quan về dữ liệu

Để có cái nhìn tổng quan về dữ liệu, cần thu thập thông tin về các yếu tố sau: số lượng dòng và cột, xác định sự hiện diện của dữ liệu bị thiếu, nếu có, cần biết dữ liệu thiếu ở những dòng cụ thể và trong những cột cụ thể nào, cùng với tỷ lệ phần trăm mà dữ liệu thiếu chiếm trong toàn bộ tập dữ liệu.

Đầu tiên, kiểm tra sự toàn vẹn của bộ dữ liệu:

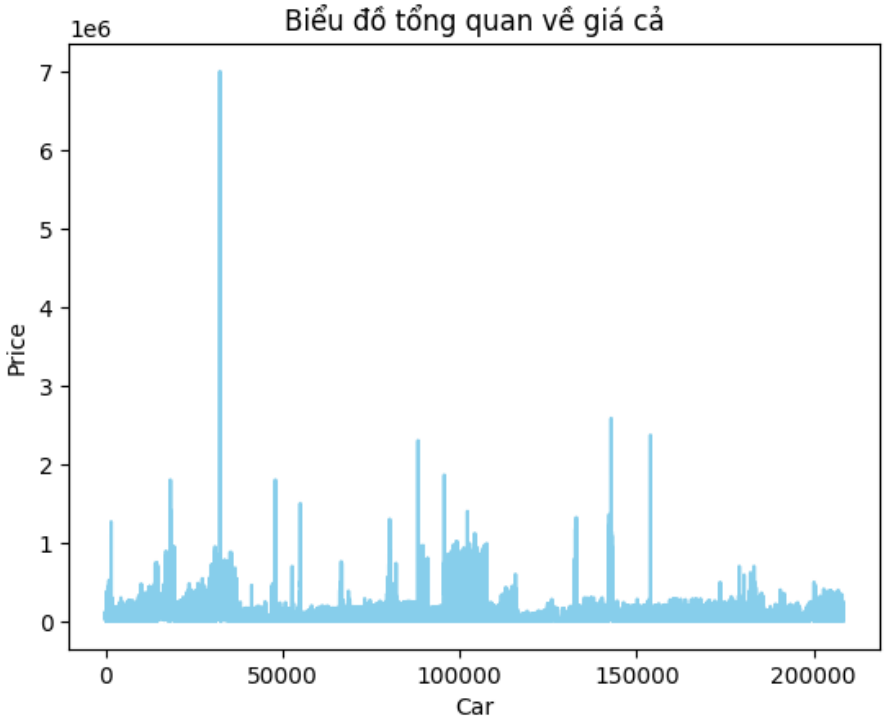
```
# Phần trăm các cột các dữ liệu bị thiếu

missing_val_over40 = []

for col in df:

    missing_value = df[col].isnull().sum()

    if missing_value > 0:

        missing_percentage = round(missing_value*100/df.shape[0],5)

        print(f"Phần trăm dữ liệu bị thiếu {col}: {missing_percentage}
%")

        print("")

        if missing_percentage > 40:

            missing_val_over40.append(col)

print(f"""Các cột có phần trăm dữ liệu mất lớn (trên 40%):

{missing_val_over40}""")
```

Output:

```
Phần trăm dữ liệu bị thiếu Vehicle_version: 33.71131 %

Phần trăm dữ liệu bị thiếu Vehicle_generation: 29.01721 %

Phần trăm dữ liệu bị thiếu Mileage_km: 0.47191 %

Phần trăm dữ liệu bị thiếu Power_HP: 0.30868 %

Phần trăm dữ liệu bị thiếu Displacement_cm3: 0.94381 %

Phần trăm dữ liệu bị thiếu CO2_emissions: 54.85108 %

Phần trăm dữ liệu bị thiếu Drive: 7.2375 %

Phần trăm dữ liệu bị thiếu Transmission: 0.22995 %

Phần trăm dữ liệu bị thiếu Doors_number: 0.71386 %

Phần trăm dữ liệu bị thiếu Origin_country: 43.20224 %

Phần trăm dữ liệu bị thiếu First_owner: 68.75048 %

Phần trăm dữ liệu bị thiếu First_registration_date: 58.50056 %

Các cột có phần trăm dữ liệu mất lớn (trên 40%):
['CO2_emissions', 'Origin_country', 'First_owner', 'First_registration_date']
```

Nhận xét: Sơ lược tổng quan của bộ dữ liệu có 25 cột và 208304 dòng, trong đó có 12 cột tồn tại missing values, có 3 cột có tỷ lệ missing values trên 50% bao gồm cột "First_owner" với 68.75%, cột "First_registration_date" với 58.5% và cột "CO2_emissions" với 54.85%. Các dữ liệu missing values sẽ được xử lý ở trong phần tiền xử lý dữ liệu.

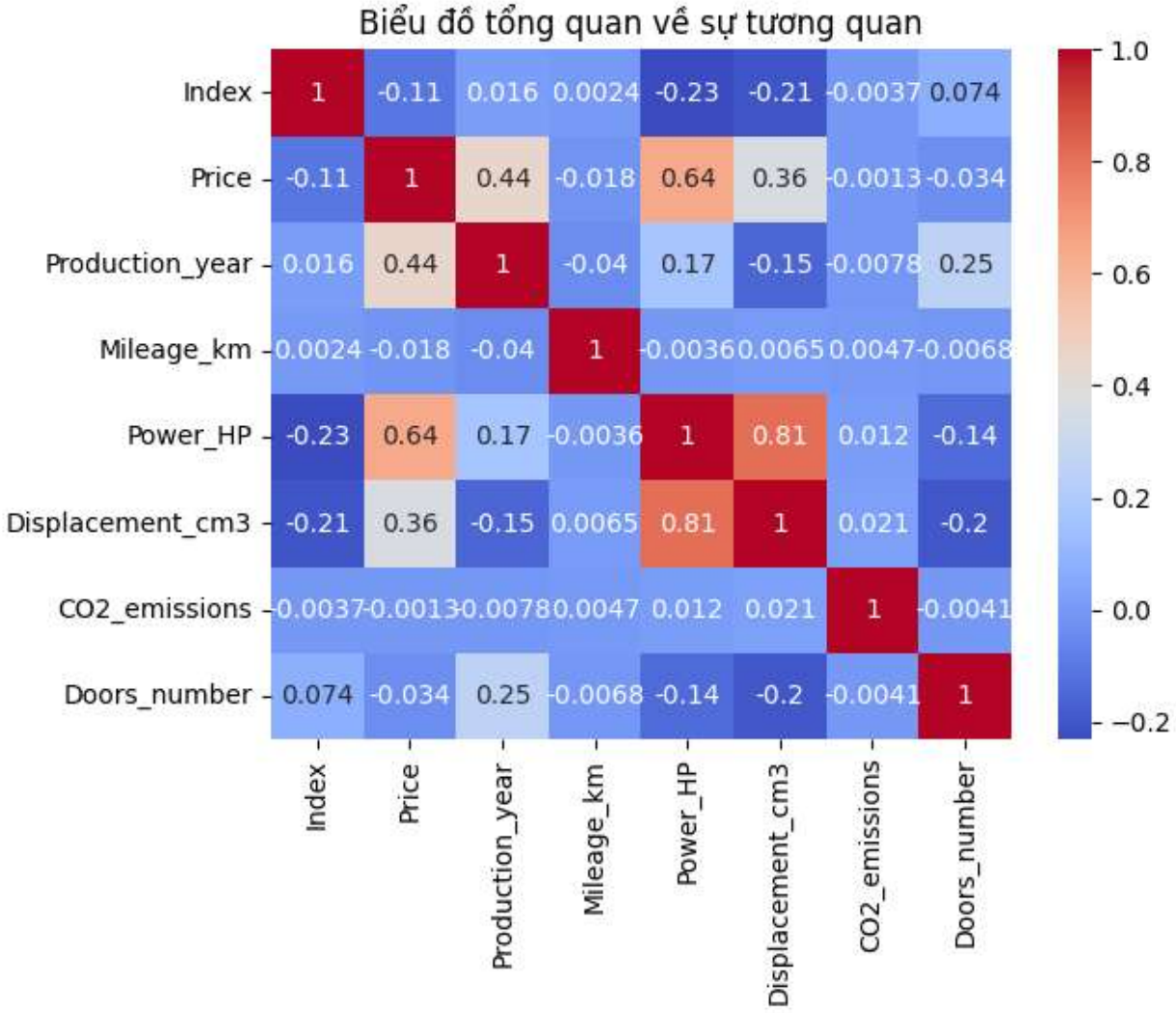
Bước tiếp theo trong quá trình phân tích dữ liệu là tiến hành khám phá chi tiết về các biến có trong bộ dữ liệu. Trong trường hợp này, biến mục tiêu mà chúng ta quan tâm là biến "Price". Chúng ta thực hiện một loạt các phân tích và tính toán để hiểu rõ hơn về tính chất của biến này. Điều này bao gồm việc xem xét các thông số thống kê như giá trị trung bình, phương sai, phân phối dữ liệu và biểu đồ tương ứng; và các thước đo khác để xác định sự biến động, xu hướng và tính chất của biến "Price".



	Stats	Values (In Thousand)
0	count	208.304000
1	mean	63.053834
2	std	86.659673
3	min	0.500000
4	25%	17.800000
5	50%	35.700000
6	75%	75.990000
7	max	6999.000000

Nhận xét: Giá của các loại xe được quảng cáo nằm trong khoảng từ 0.500 đến 6999.000, với giá trị trung bình và trung vị khá gần nhau. Phần lớn (75%) của các giá trị nằm dưới 75.990, trong khi một số mẫu dữ liệu có giá trị cao hơn nhiều, giá lớn nhất lên đến 6.999000. Độ lệch chuẩn khoảng 86.660 cho thấy sự phân tán rộng của dữ liệu. Có thể thấy được rằng có sự biến động đáng kể giữa mẫu dữ liệu trong biến "Price".

Tiếp theo, chúng ta cùng tìm hiểu tổng quan sự tương quan giữa biến Price và các biến Numerical thông qua biểu đồ Heatmap:



Nhận xét: Có thể quan sát một sự tương quan mạnh giữa biến "Power_HP" (công suất động cơ) và "Displacement_cm3" (dung tích động cơ) với biến mục tiêu "Price" (giá cả). Sự tương quan này là một yếu tố quan trọng trong việc hiểu sự biến đổi của giá xe trên thị trường.

Biến "Power_HP" thường đóng vai trò quan trọng trong xác định giá trị của một chiếc xe, vì công suất động cơ thường liên quan đến hiệu suất và khả năng vận hành. Một động cơ mạnh có thể tương ứng với một chiếc xe có giá cao hơn.

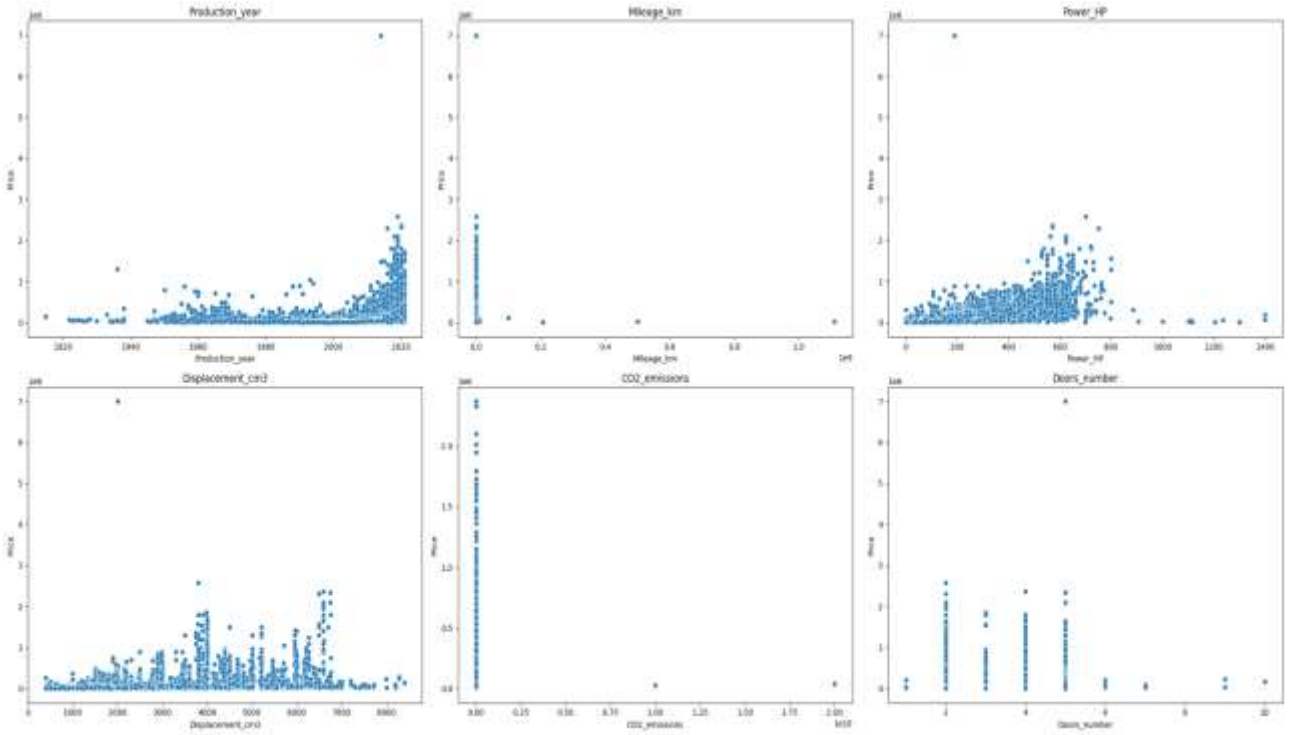
Tương tự, biến "Displacement_cm3" cũng là một yếu tố quan trọng. Dung tích động cơ có thể liên quan đến khả năng vận hành, tiết kiệm nhiên liệu và hiệu suất. Một dung tích động cơ lớn có thể dẫn đến một chiếc xe có giá trị cao hơn.

Chỉ số tương quan Pearson của 0.64 giữa "Power_HP" và "Price" cho thấy một mối tương quan cao, và chỉ số 0.36 giữa "Displacement_cm3" và "Price" cũng cho thấy một mối tương quan trung bình. Điều này có nghĩa rằng, trong mô hình dự đoán giá xe, công suất động cơ và dung tích động cơ là những yếu tố quan trọng mà chúng ta cần xem xét, và chúng có khả năng ảnh hưởng lớn đến giá cả của xe hơi trên thị trường.

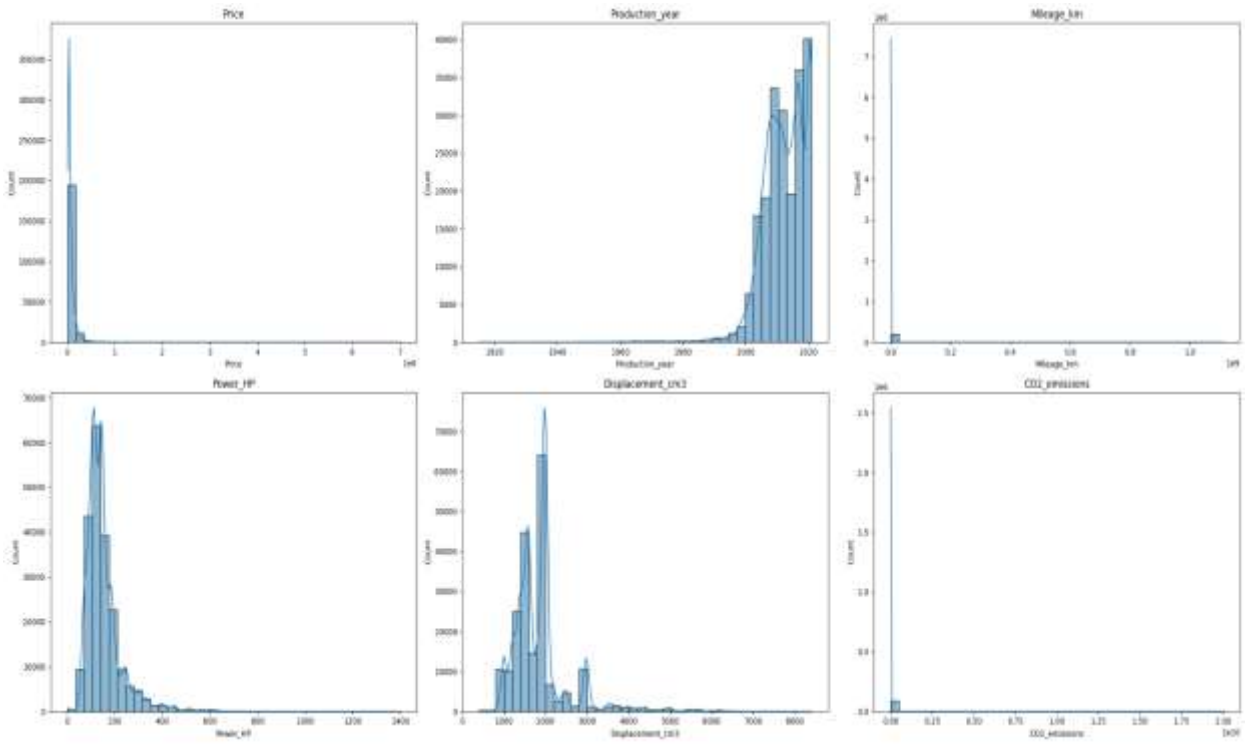
Bên cạnh đó, chúng ta cũng có thể quan sát một mối tương quan mạnh giữa biến "Production_year" và "Price", với chỉ số tương quan là 0.44. Điều này ngụ ý rằng năm sản xuất của xe đóng một vai trò quan trọng trong ảnh hưởng đến giá trị của chúng, và sự thay đổi của năm sản xuất có thể tác động đáng kể đến giá cả.

Các biến như "Mileage_km", "CO2_emissions", và "Doors_number" mang giá trị âm thể hiện mức độ tương quan ngược chiều với biến "Price", với những chỉ số tương quan nhỏ như vậy cho thấy rằng các yếu tố như quãng đường đã đi, lượng khí CO2 phát thải, và số cửa của xe không có ảnh hưởng lớn đến giá trị của xe, ít nhất là từ góc độ tương quan trong phạm vi dữ liệu hiện có.

Để kiểm chứng lại các nhận xét ở trên, chúng ta có thể sử dụng biểu đồ subplots để trực quan hóa sự tương qua của biến Price và các biến Numerical:



Tiếp theo, nhóm quyết định xem sự phân bổ của các biến thông qua biểu đồ:




```
for col in numeric_x:
  skew = df[col].skew()
  kurt = df[col].kurtosis()
  print(f"{col}: Độ lệch: {skew.round(2)}, đỉnh: {kurt.round(2)}")
  if skew < 0:
    print(f"=> phân phối lệch trái: {col}\n")
  elif skew > 0:
    print(f"=> phân phối lệch phải: {col}\n")
```

Output:

```
Price: Độ lệch: 7.52, đỉnh: 250.77
=> phân phối lệch phải: Price
Production_year: Độ lệch: -1.95, đỉnh: 11.84
=> phân phối lệch trái: Production_year
Mileage_km: Độ lệch: 310.34, đỉnh: 106902.72
```

```
=> phân phối lệch phải: Mileage_km

Power_HP: Độ lệch: 2.53, đỉnh: 10.55

=> phân phối lệch phải: Power_HP

Displacement_cm3: Độ lệch: 2.5, đỉnh: 9.46

=> phân phối lệch phải: Displacement_cm3

CO2_emissions: Độ lệch: 246.86, đỉnh: 63950.49

=> phân phối lệch phải: CO2_emissions
```

Nhận xét: Ta thấy đa số các phân phối đều lệch phải, chỉ có Production_year là lệch trái với giá trị -1.95. Có các thuộc tính có đỉnh rất nhọn (CO2_emissions: 63950 và Mileage_km: 106902) cho thấy các giá trị cực trị là rất nhiều.

Sau khi đã xem xét các biến numeric, tiếp theo chúng ta cùng tìm hiểu về các biến categorical có trong bộ dữ liệu:

Nhận xét: Thông qua các biểu đồ, có thể thấy được sự phân phối rõ giữa các giá trị trong các biến:

Ở biến Condition, có sự khác biệt rõ rệt giữa số lượng xe đã qua sử dụng và số lượng xe mới được đăng bán. Số lượng xe đã qua sử dụng cao hơn rất nhiều so với số lượng xe mới, ước tính khoảng gấp 7 lần.

Hay sự phân hóa rõ rệt giữa các loại nhiên liệu được các xe sử dụng trong biến Fuel_type, 2 loại nhiên liệu được sử dụng chính là Gasoline và Diesel, mỗi loại chiếm tỷ trọng cao hơn gấp khoảng 10 lần so với loại nhiên liệu đứng thứ 3 là Gasoline + LPG. Điều này có thể phản ánh thị trường và sự ưa thích của người tiêu dùng đối với các loại nhiên liệu này, có thể là do hiệu suất, giá cả hoặc yếu tố môi trường.

Ngoài ra, việc trực quan hóa trên biểu đồ còn giúp chúng ta có thể nhận thấy rằng hầu hết các loại xe sử dụng hộp số thủ công (manual), và tỷ lệ này gần như gấp đôi so với hộp số tự động (automatic). Điều này cho thấy sự ưa thích của một phần lớn người dùng về hộp số thủ công trong việc lái xe. Sự thịnh hành của hộp số thủ công có thể được hiểu như một biểu hiện của sự ưa thích của một phần lớn người lái xe. Mặc dù hộp số tự động thường mang lại sự thuận tiện và dễ sử dụng hơn, người lái có thể ưa chuộng hộp số thủ công vì nó mang lại trải nghiệm lái xe và sự tương tác nhiều hơn với chiếc xe.

Dường như hệ thống dẫn động Front Wheels chiếm đa số trong số các xe, với khoảng 140,000 chiếc sử dụng loại này, có thể phản ánh sự ưa chuộng về hiệu suất trong điều kiện đường bình thường và tiết kiệm nhiên liệu.

3.2 Làm sạch dữ liệu

Làm sạch dữ liệu là một bước quan trọng trong quá trình phân tích, xử lý dữ liệu. Để đảm bảo chất lượng của kết quả phân tích, dữ liệu cần được làm sạch trước khi tiến hành kiểm tra và xử lý. Làm sạch dữ liệu là quá trình loại bỏ, xử lý các dữ liệu không mong muốn hoặc thiếu sót ra khỏi tập dữ liệu. Các dữ liệu không mong muốn có thể bao gồm các giá trị sai lệch, trùng lặp, thiếu thông tin, hoặc không phù hợp với mục tiêu phân tích. Tất cả những vấn đề này có thể ảnh hưởng đến chất lượng của việc phân tích và dẫn đến những kết luận sai lệch. Mục tiêu của việc làm sạch dữ liệu (data cleaning) là tạo ra một tập dữ liệu sạch, đồng nhất và phù hợp với mục tiêu phân tích. Do đó, trước khi bắt đầu quá trình phân tích dữ liệu, chúng ta cần tiến hành làm sạch dữ liệu để đảm bảo chất lượng của kết quả.

3.2.1 Xử lý giá trị trùng lặp

Trước tiên, ta tiến hành kiểm tra số lượng dòng bị trùng lặp của mỗi cột trong DataFrame.

```
# Kiểm tra các hàng trùng lặp

duplicate = df[df.duplicated()]

print('Các hàng bị trùng lặp:', len(duplicate))
```

Output:

```
Các hàng bị trùng lặp: 0
```

Nhận xét: Tập dữ liệu không chứa giá trị trùng lặp nào, do đó nhóm không thực hiện các bước để xử lý giá trị trùng lặp.

3.2.2 Hợp lý hóa đơn vị tiền

Chuyển các dòng có giá trị đơn vị tiền tệ từ EUR sang PLN trong cột Currency. Việc chuyển đổi này giúp đơn giản hóa việc phân tích và so sánh dữ liệu.

```
# Convert EUR to PLN. 1 EUR = 4.45 PLN

df_EUR = df.loc[df['Currency'] == 'EUR']

for index, row in df_EUR.iterrows():

    df.at[index, 'Price'] = row['Price'] * 4.45
```

```
df.drop(columns=['Currency'],axis=1,inplace=True)
```

Sau khi chuyển đơn vị tiền tệ, ta tiến hành loại bỏ cột Currency bởi vì lúc này tất cả các giá trị trong cột Price đều được biểu diễn bằng cùng một đơn vị tiền tệ, do đó thông tin về đơn vị tiền tệ ban đầu không còn mang lại giá trị gì cho việc phân tích dữ liệu.

3.2.3 Xử lý giá trị bị thiếu

Giá trị thiếu trong dữ liệu có thể ảnh hưởng nghiêm trọng đến kết quả phân tích. Do đó, việc kiểm tra giá trị thiếu là một bước vô cùng cần thiết để đảm bảo tính chính xác của kết quả phân tích. Việc này giúp chúng ta xác định được những thông tin nào đang bị thiếu trong bộ dữ liệu, từ đó có thể đưa ra các phương án xử lý phù hợp.

Trước tiên, nhóm tiến hành đánh giá tỷ lệ phần trăm giá trị thiếu trong từng cột dữ liệu:

```
missing_val_over40 = []

for col in df:

    missing_value = df[col].isnull().sum()

    if missing_value > 0:

        missing_percentage = round(missing_value*100/df.shape[0],5)

        print(f"Phần trăm dữ liệu bị thiếu {col}: {missing_percentage}
%")

        print("")
```

```
        if missing_percentage > 40:
            missing_val_over40.append(col)
print(f"""Các cột có phần trăm dữ liệu mất lớn (trên 40%):
{missing_val_over40}""")
```

Output :

```
Phần trăm dữ liệu bị thiếu Vehicle_version: 33.71131 %
Phần trăm dữ liệu bị thiếu Vehicle_generation: 29.01721 %
Phần trăm dữ liệu bị thiếu Mileage_km: 0.47191 %
Phần trăm dữ liệu bị thiếu Power_HP: 0.30868 %
Phần trăm dữ liệu bị thiếu Displacement_cm3: 0.94381 %
Phần trăm dữ liệu bị thiếu CO2_emissions: 54.85108 %
Phần trăm dữ liệu bị thiếu Drive: 7.2375 %
Phần trăm dữ liệu bị thiếu Transmission: 0.22995 %
Phần trăm dữ liệu bị thiếu Doors_number: 0.71386 %
Phần trăm dữ liệu bị thiếu Origin_country: 43.20224 %
Phần trăm dữ liệu bị thiếu First_owner: 68.75048 %
Phần trăm dữ liệu bị thiếu First_registration_date: 58.50056 %

Các cột có phần trăm dữ liệu mất lớn (trên 40%):
['CO2_emissions', 'Origin_country', 'First_owner', 'First_registration_date']
```

a. Xóa các cột có dữ liệu bị thiếu ra khỏi bộ dữ liệu:

Sau khi đánh giá tỷ lệ phần trăm giá trị thiếu trong từng cột dữ liệu, ta có thể quan sát thấy rằng có 4 cột chứa lượng thông tin bị thiếu đáng kể, vượt mức 40%, bao gồm: CO2_emissions, First_owner, First_registration_date và Origin_country. Vậy nên, nhóm quyết định loại bỏ những cột này ra khỏi tập dữ liệu. Việc loại bỏ những cột này sẽ không gây ảnh hưởng đáng kể đến chất lượng của việc phân tích dữ liệu, bởi vì chúng không quá quan trọng đối với mục tiêu phân tích.

Ngoài ra, xóa Index ra khỏi Dataframe vì trùng với Index của DataFrame. Hai thuộc tính Vehicle_version và Vehicle_generation là những thuộc tính chứa mixed value và khó để xử lý nên loại bỏ khỏi bộ dữ liệu.

```
# Xóa các thuộc tính có trên 40% tổng dữ liệu bị mất

df.drop(columns = missing_val_over40 + ['Index', 'Vehicle_version',
'Vehicle_generation'], axis=1,inplace=True)
```

b. Thay thế bằng giá trị median

Đối với thuộc tính Mileage_km nhóm sử dụng phương pháp điền giá trị trung vị để cân bằng các phân phối lệch phải và tránh ảnh hưởng các cực trị. Phương pháp này giúp đảm bảo tính chính xác của dữ liệu, đồng thời giúp duy trì số lượng quan sát ban đầu.

Xử lý giá trị bị thiếu của cột Mileage_km:

```
median_mileage = df['Mileage_km'].median()

# Thay thế giá trị bị thiếu bằng giá trị trung vị

df['Mileage_km'].fillna(median_mileage, inplace=True)

df['Mileage_km']
```

Output:

```
0              1.0
1          59000.0
2          52000.0
3          29000.0
4            600.0
             ...
208298      6000.0
208300     63518.0
208301     11880.0
208302    100000.0
208303     20056.0
Name: Mileage_km, Length: 208289, dtype: float64
```

c. Thay thế bằng giá trị Mode:

Để xử lý các giá trị bị thiếu cho các cột Power_HP, Displacement_cm3, Transmission và Drive trong bộ dữ liệu, ta xử lý giá trị bị thiếu bằng cách sử dụng phương pháp thay thế bằng giá trị phổ biến nhất trong cột đó. Điều này giúp giảm thiểu sự biến đổi trong dữ liệu ban đầu và duy trì tính chất phân loại của biến.

Xử lý giá trị bị thiếu của cột Power_HP:

```
# Thay thế giá trị bị thiếu bằng giá trị xuất hiện nhiều nhất

df['Power_HP'].fillna(df['Power_HP'].mode()[0], inplace=True)

df['Power_HP']
```

Output:

```
0         145.0
1          75.0
2         180.0
3         160.0
4         165.0
          ...
208298    150.0
208300     70.0
208301     60.0
208302     36.0
208303     70.0
Name: Power_HP, Length: 208289, dtype: float64
```

Xử lý giá trị bị thiếu của cột Displacement_cm3:

```
# Thay thế giá trị bị thiếu bằng giá trị trung bình

df['Displacement_cm3'].fillna(df['Displacement_cm3'].mode()[0],
```

```
inplace=True)

df['Displacement_cm3']
```

Output:

```
0          1400.0
1          1100.0
2          1368.0
3          1368.0
4          1368.0
            ...
208298      750.0
208300     2120.0
208301     2120.0
208302     2200.0
208303     2120.0
Name: Displacement_cm3, Length: 208289, dtype: float64
```

- Việc điền mode cho các giá trị bị thiếu ở cột Power_HP và Displacement là để đảm bảo các giá trị được thêm vào nằm ở miền phân phối phổ biến và đồng thời tránh ảnh hưởng của các cực trị.

Xử lý giá trị bị thiếu của cột Transmission:

```
# Thay thế giá trị bị thiếu bằng giá trị phổ biến nhất

df['Transmission'].fillna(df['Transmission'].mode()[0], inplace=True)

df['Transmission']
```

Output:

```
0             Manual
1             Manual
2          Automatic
3             Manual
4             Manual
              ...
208298        Manual
208300        Manual
208301        Manual
208302        Manual
208303        Manual
Name: Transmission, Length: 208289, dtype: object
```

Xử lý giá trị bị thiếu của cột Drive:

```
# # Xử lý missing value của cột Drive

df['Drive'].fillna(df['Drive'].value_counts().idxmax(),inplace=True)
```

Output:

```
0          Front wheels
1          Front wheels
2          Front wheels
3          Front wheels
4          Front wheels
               ...
208299     Front wheels
208300      Rear wheels
208301     Front wheels
208302     Front wheels
208303     Front wheels
```

```
Name: Drive, Length: 208289, dtype: object
```

d. Thay thế dựa trên dữ liệu liên quan

Cuối cùng, để xử lý các giá trị bị thiếu trong cột Doors_number, ta sử dụng phương pháp imputation dựa trên dữ liệu liên quan, cụ thể ở đây là dựa vào thông tin ở cột Type.

Đầu tiên, nhóm tiến hành loại bỏ các dòng trong cột Doors_number có giá trị là 1, 7, 9, 10 vì chúng không hợp lệ.

Loại bỏ những giá trị không hợp lệ:

```
# Xóa các dòng có giá trị là 9, 10, 7 hoặc 1 trong cột 'Doors_number'
vì không hợp lệ

df = df[~df['Doors_number'].isin([9, 10, 7, 1])]
```

Như đã thấy trong Crosstab ở bên dưới thì các giá trị 1, 9, 7, 10 chỉ xuất hiện 1-2 lần và không phù hợp với loại xe. Ví dụ, loại xe nhỏ (small_cars) không thể có đến 9 cửa.

```
doors_type_ct = pd.crosstab(df['Doors_number'],df['Type'])

doors_type_ct
```

Output:

Type	SUV	city_cars	compact	convertible	coupe	minivan	sedan	small_cars	station_wagon
Doors_number									
1.0	0	0	1	0	1	0	0	2	0
2.0	137	259	189	2037	3744	120	207	451	71
3.0	495	5204	3122	321	1089	150	100	2288	76
4.0	1163	347	1051	52	520	2898	19791	66	1987
5.0	39461	18132	27605	56	954	18760	12924	3078	37846
6.0	3	0	0	0	0	31	1	0	16
7.0	3	0	0	0	0	5	0	0	0
9.0	0	0	0	0	0	1	0	1	0
10.0	0	0	0	0	0	1	0	0	0

```
df['Doors_number'] =
df['Doors_number'].fillna(df['Type'].map(doors_type_mode))
```

Sau khi loại bỏ các giá trị không hợp lệ, nhóm tiến hành chia nhỏ vấn đề và xử lý từng loại xe. Cụ thể, dựa vào Crosstab để xác định từng loại cửa được gia công nhiều nhất cho từng loại xe như kết quả bên dưới. Tiếp đến, tiến hành điền các ô dữ liệu thiếu bằng doors mode của từng loại xe.

```
# Xử lý missing value của cột Doors_number

doors_type_mode = dict()

for car in doors_type_ct:

    doors_mode = doors_type_ct[car].idxmax()

    doors_type_mode[car] = doors_mode

doors_type_mode
```

Output:

```
{'SUV': 5.0,
 'city_cars': 5.0,
 'compact': 5.0,
```

```
'convertible': 2.0,
'coupe': 2.0,
'minivan': 5.0,
'sedan': 4.0,
'small_cars': 5.0,
'station_wagon': 5.0}
```

Sau quá trình loại bỏ các dòng chứa giá trị bị thiếu từ các cột: Vehicle_version, Vehicle_generation, Mileage_km, Power_HP, Displacement_cm3, CO2_emissions, Drive, Transmission, Doors_number, Origin_country, First_owner và First_registration_date, bộ dữ liệu không còn giá trị bị thiếu ở bất kỳ cột nào.

3.3 Chuyển đổi và phân loại kiểu dữ liệu

Đầu tiên, để cải thiện tính nhất quán và độ chính xác của dữ liệu, nhóm chuyển cột Offer_publication_date và cột Production_year sang dạng Datetime. Điều này có thể giúp cho quá trình phân tích dữ liệu trở nên thuận tiện và chính xác hơn.

Tiếp đến, chuyển cột Doors_number sang dạng Categorical bởi vì mặc dù cột Doors_number trong tập dữ liệu được biểu diễn dưới dạng số nhưng thực chất nó là một biến phân loại. Hơn nữa, cột này có ít hơn 10 giá trị duy nhất, do đó, việc xem xét nó như một biến phân loại thường mang lại nhiều ý nghĩa hơn so với việc xem nó như một biến số. Thế nên, việc chuyển đổi Doors_number sang dạng Categorical sẽ giúp ta phản ánh đúng hơn bản chất của dữ liệu.

Cuối cùng, ta tiến hành phân loại các cột trong tập dữ liệu và in ra danh sách các cột theo từng loại để có cái nhìn tổng quan về cấu trúc và tính chất của dữ liệu. Việc này giúp chúng ta hiểu rõ hơn về dữ liệu và từ đó có thể đưa ra các phương pháp tiền xử lý phù hợp.

Các thuộc tính Doors_number, Production_year dù có dtype là object và datetime nhưng format vẫn là số nên cần loại ra khỏi list numerical_features.

Đây là các thuộc tính sau khi xử lý:

Các cột có kiểu dữ liệu numerical là:

['Price', 'Mileage_km', 'Power_HP', 'Displacement_cm3']

Các cột có kiểu dữ liệu categorical là:

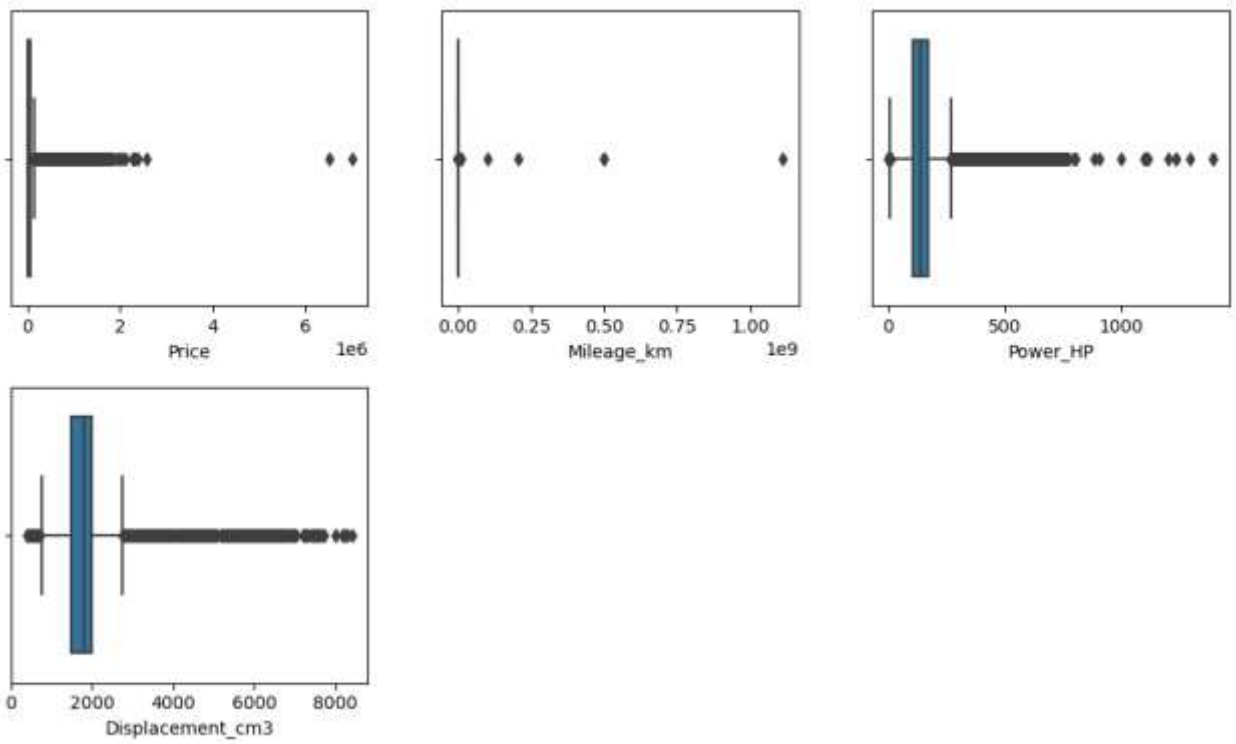
['Condition', 'Vehicle_brand', 'Vehicle_model', 'Fuel_type', 'Drive', 'Transmission', 'Type', 'Doors_number', 'Colour', 'Offer_location', 'Features']

Các cột có kiểu dữ liệu datetime là:

['Offer_publication_date', 'Production_year']

3.4 Xử lý outliers của các biến định lượng

Kiểm tra outlier của các biến:

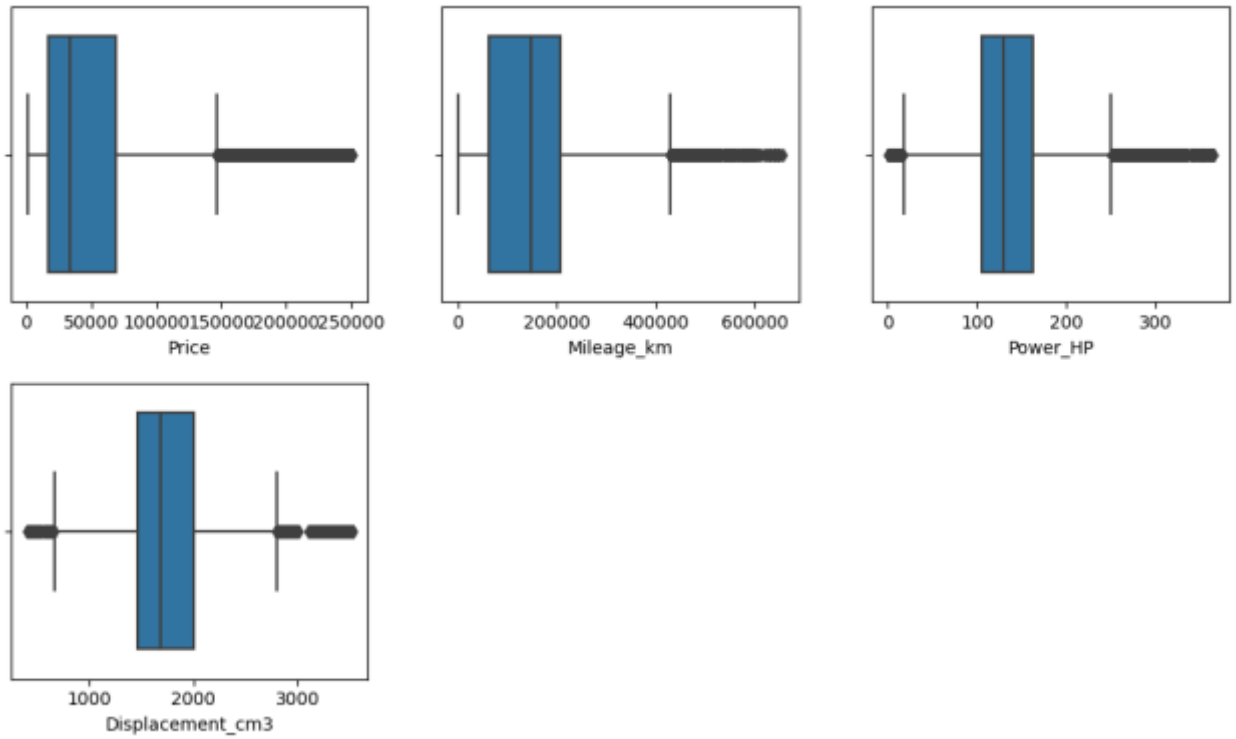


Dùng hàm iqr tính toán khoảng tứ phân vị cho một cột trong DataFrame để tiến hành kiểm tra Outliers. Ở đây, ta tính tứ phân vị thứ nhất (Q1) và thứ ba (Q3).

Sau khi đã có index của các outliers, tiến hành loại bỏ Outliers ra khỏi bộ dữ liệu. Để xử lý Outliers, ta loại ra khỏi Dataframe.

Với multiplier = 3, sẽ chỉ loại những siêu ngoại lai để đảm bảo miền giá trị của thuộc tính định lượng là nhiều nhất.

Kiểm tra lại bằng boxplot:



Lưu ý: vẫn còn một số Boxplot còn thể hiện giá trị ngoài khoảng 3 sigma nhưng không nằm trong khoảng Lower/Upper bound của IQR nên căn bản ngoại lai đã được xóa bỏ.

Bộ dữ liệu ban đầu gồm 208304 dòng và 25 cột. Sau khi thực hiện các bước kiểm tra và xử lý giá trị trùng lặp, giá trị thiếu và giá trị outliers thì bộ dữ liệu còn 195234 và 18 cột.

CHƯƠNG 4: EDA - PHÂN TÍCH KHÁM PHÁ DỮ LIỆU

4.1 Thao tác thuộc tính

4.1.1 Thêm thuộc tính

Tạo cột Vehicle mới từ 2 cột Vehicle_brand và Vehicle_model:

```
df['Vehicle'] = df['Vehicle_brand'] + ' ' + df['Vehicle_model']

categorical_features.append('Vehicle')
```

Nhận xét: Việc tạo một cột mới từ hai cột Vehicle_brand và Vehicle_model có thể giúp thực hiện việc phân tích dễ dàng hơn. Cột Vehicle_brand chứa tên các hãng xe và cột Vehicle_model chứa tên các dòng xe, việc tạo cột Vehicle sẽ giúp chúng ta hiển thị thông tin xe cụ thể, ví dụ: Vehicle_brand: Toyota, Vehicle_model: Camry sẽ được kết hợp thành một cột Vehicle: Toyota Camry. Như vậy, ta có thể thấy dòng xe cụ thể là Toyota Camry mà không cần phải nhìn vào cả hai cột. Điều này có thể giúp ích trong việc phân tích dữ liệu và làm cho dữ liệu dễ hiểu hơn.

Thêm cột vehicle vào danh sách các thuộc tính định danh

```
df.drop(columns=['Vehicle_model'],inplace=True)

categorical_features = [element for element in
categorical_features if element != 'Vehicle_model']
```

Nhận xét: Việc tạo ra cột Vehicle, là tên xe bao gồm cả Brand - thương hiệu và Model - mẫu xe thì xóa cột Model vì không còn mang ý nghĩa như Vehicle. Sau đó xóa bỏ giá trị đó trong danh sách Categorical_features.

4.1.2 Xử lý chuỗi

```
df['Features'] = df['Features'].replace('[]', 'No Features')

df['Features'] = df['Features'].str.strip('[]')
```

Nhận xét: Thay thế những giá trị [ ] trong features bằng No Features, những giá trị trong này có nghĩa là xe không sở hữu tính năng đặc biệt nào. Xóa đi hai dấu ngoặc vuông ở cột Features để thuận tiện cho việc xử lý chuỗi về sau:

```
df[['Features']]
```

	Features
0	No Features
1	No Features
2	'ABS', 'Electric front windows', 'Drivers airb...
3	'ABS', 'Electric front windows', 'Drivers airb...
4	'ABS', 'Electrically adjustable mirrors', 'Pas...
...	...
208298	No Features
208300	No Features
208301	No Features
208302	No Features
208303	No Features

195226 rows × 1 columns

Trích xuất ra tên thành phố:

```
from unidecode import unidecode

df['Offer_location'] = df['Offer_location'].apply(lambda
x:unidecode(x))
```

Liệt kê ra những địa danh:

```
# List of Polish cities
polish_location = [
    "Warsawa", "Krakow", "Lodz", "Wroclaw", "Poznan", "Gdansk", "Szczecin",
    "Bydgoszcz", "Lublin", "Bialystok", "Katowice", "Gdynia", "Czestochowa",
    "Radom", "Sosnowiec", "Torun", "Kielce", "Gliwice", "Zabrze", "Bytom",
    "Bielsko-Biala", "Olsztyn", "Rzeszow", "Ruda Slaska", "Rybnik", "Nowe Tychy",
    "Dabrowa Gornicza", "Plock", "Elblag", "Opole", "Walbrzych", "Wloclawek",
    "Tarnow", "Chorzow", "Kalisz", "Koszalin", "Legnica", "Grudziadz", "Jaworzno",
    "Slupsk", "Jastrzebie Zdroj", "Nowy Sacz", "Jelenia Gora", "Konin",
    "Piotrkow Trybunalski", "Inowroclaw", "Lubin", "Siedlce", "Piekary Slaskie",
    "Myslowice", "Ostrowiec Swietokrzyski", "Siemianowice Slaskie", "Ostrow Wielkopolski",
    "Suwalki", "Gniezno", "Stargard", "Glogow", "Wejherowo", "Przemysl", "Zamosc",
    "Leszno", "Lomza", "Chelm", "Tomaszow Mazowiecki", "Pruszkow", "Stalowa Wola",
    "Zgierz", "Starachowice", "Skarzysko-Kamienna", "Tarnowskie Gory", "Kedzierzyn-Kozle",
    "Leczyca", "Mielec", "Tczew", "Swidnica", "Pabianice", "Sochaczew", "Otwock",
    "Swinoujscie", "Belchatow", "Swarzedz", "Bedzin", "Zory", "Krosno", "Jaworzno",
    "Biala Podlaska", "Kutno", "Nowa Sol", "Pila", "Sieradz", "Zory", "Swietochlowice",
    "Ostroleka", "Siemiatycze", "Zdunska Wola", "Legionowo", "Turek", "Myszkow",
    "Nysa", "Kolobrzeg", "Gostyn", "Boleslawiec", "Olawa", "Bierun", "Zlotow",
    "Grodzisk Mazowiecki", "Wadowice"
]
```

Tạo hàm để thay thế, đồng thời bao gồm thêm những giá trị tỉnh thành khác theo biểu thức chính quy và thay thế những giá trị thành phố /tỉnh lụy chứa trong location thành tên nơi đó.

```
def replace_city(location):
    for city in polish_location:
        if city in location:
            return city
    return location


def extract_province(address):
    match = re.search(r',\s*([^,]+?)\s*(?:\(|$)', address)
    return match.group(1).strip() if match else ''


extracted_phrases = [extract_province(addr) for addr in
df['Offer_location']]
polish_location.extend(extracted_phrases)
df['Offer_location'] = df['Offer_location'].apply(replace_city)
```

Thay thế những giá trị trích xuất trả về chuỗi trống bằng mode:

```
# Replace những giá trị không xử lí được bằng mode

df['Offer_location'] = df['Offer_location'].replace(' ',
df['Offer_location'].mode()[0])
```

Nhận xét: Việc chuẩn hóa các ký tự Ba Lan về UTF-8 sẽ giúp chúng ta dễ tiếp cận với địa điểm buôn bán xe và tránh các lỗi về mã hóa. Sau đó thay thế các nơi rao bán cụ thể bằng tên thành phố/tỉnh lụy, có nghĩa là chỉ quan tâm về thông tin rao bán xe ở các thành phố/ tỉnh lụy được trích xuất.

Việc trích xuất này sẽ không thể chính xác tuyệt đối vì độ phức tạp của tổ chức chuỗi và sắp xếp ký tự không đồng nhất trong 'Offer_location', do đó các giá trị không trích xuất được sẽ được điền bằng giá trị xuất hiện nhiều nhất ở cột thành phố/tỉnh lụy đã qua xử lý.

4.1.3 Rời rạc hóa dữ liệu

Tạo hai khoảng thời gian (trước và sau năm 2000) để có cái nhìn bao quát về mối liên hệ của các thuộc tính khác dựa trên hai khoảng thời gian này. Sau đó tái gán lại vào list categorical features.

```
df['Production_period'] = pd.cut(df['Production_year'],

                                 bins =
[df['Production_year'].min(),2000,df['Production_year'].max()],

                                 labels = ['Before 2000','After
2000'])

categorical_features.append('Production_period')
```

```
regions = {
    "North": [
        "Pomorskie", "Gdansk", "Gdynia", "Sopot", "Trzew", "Slupsk", "Wejherowo",
        "Kujawsko-pomorskie", "Bydgoszcz", "Torun", "Wloclawek",
        "Warminsko-mazurskie", "Olsztyn", "Elblag", "Sopot"
        "Zachodniopomorskie", "Szczecin", "Koszalin", "Swinoujscie", "Stargard", "Kolobrzeg"
    ],
    "West": [
        "Wielkopolskie", "Poznan", "Kalisz", "Konin", "Gniezno", "Pila", "Ostrów Wielkopolski",
        "Ostrow Wielkopolsk",
        "Lubuskie", "Zielona Góra", "Gorzów Wielkopolski",
        "Dolnoslaskie", "Wroclaw", "Legnica", "Walbrzych", "Jelenia Gora", "Lubin", "Swidnica",
        "Gorzów Wielkopolski"
    ],
    "South": [
        "Malopolskie", "Krakow", "Nowy Sacz", "Tarnow", "Wadowice",
        "Slaskie", "Katowice", "Czestochowa", "Sosnowiec", "Olwice", "Zabrze", "Bytom", "Ruda Slaska",
        "Chorzow", "Rybnik", "Tychy", "Dabrowa Gornicza", "Jaworzno", "Piekary Slaskie", "Bierun",
        "Hyslowice", "Zory", "Siemianowice Slaskie",
        "Podkarpackie", "Rzeszow", "Przemysl", "Mielec", "Sanok", "Krosno",
        "Swietokrzyskie", "Kielce", "Ostrowiec Swietokrzyski",
        "Opolskie", "Opole", "Tychy", "Bielsko-Biala", "Bedzin"
    ],
    "East": [
        "Lubelskie", "Lublin", "Zamosc", "Chelm", "Biala Podlaska",
        "Podlaskie", "Bialystok", "Lomza", "Suwalki", "Siemiatycze", "Sanok",
        "Rembertow"
    ],
    "Central": [
        "Mazowieckie", "Warsaw", "Mokotow", "Bialoleka", "Wlochy", "Wola", "Ursynow", "Bemowo",
        "Targowek", "Bielany", "Ursus", "Praga-Poludnie", "Srodmiescie", "Otwock", "Piaseczno", "Radom", "Siedlce", "Plock", "Sochaczew",
        "Grodzisk Mazowiecki", "Legionowo","Lodzkie", "Lodz", "Piotrków Trybunalski",
        "Piotrków Trybunalski","Skierniewice", "Turek"
    ]
}
```

Tạo hàm để map các giá trị trong list polish_location vào sau đó thực hiện rời rạc hóa.

Tạo biến Offer_region để rời rạc hóa các giá trị thành phố tỉnh lụy độc nhất quá lớn và dễ trực quan hơn theo vùng.

Tái gán vào danh sách thuộc tính định danh

```
def map_to_region(index):

    for region, indices in regions.items():

        if index in indices:

            return region

    return 'Unknown'


df['Offer_region'] = df['Offer_location'].apply(map_to_region)

categorical_features.append('Offer_region')
```

Những giá trị trả về Unknown sẽ được gán bằng giá trị mode của thuộc tính (vùng) đó.

```
df['Offer_region'].replace('Unknown',df['Offer_region'].mode()[0],inplace=True)
```

4.1.4 Tái thiết lập thứ tự

```
new_order = ['Price',
 'Condition',
 'Vehicle_brand',
 'Vehicle',
 'Mileage_km',
 'Power_HP',
 'Displacement_cm3',
 'Fuel_type',
 'Drive',
 'Transmission',
 'Type',
 'Doors_number',
 'Colour',
 'Production_year',
 'Production_period',
 'Offer_publication_date',
 'Offer_location',
 'Offer_region',
 'Features'
 ]
```

Tạo một danh sách chứa các thuộc tính mới và cũ theo thứ tự hợp lý và gán vào dataframe:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 195226 entries, 0 to 208303
Data columns (total 19 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   Price                  195226 non-null  float64
 1   Condition              195226 non-null  object
 2   Vehicle_brand          195226 non-null  object
 3   Vehicle                195226 non-null  object
 4   Mileage_km             195226 non-null  float64
 5   Power_HP               195226 non-null  float64
 6   Displacement_cm3       195226 non-null  float64
 7   Fuel_type              195226 non-null  object
 8   Drive                  195226 non-null  object
 9   Transmission           195226 non-null  object
 10  Type                   195226 non-null  object
 11  Doors_number           195226 non-null  object
 12  Colour                 195226 non-null  object
 13  Production_year        195226 non-null  int64
 14  Production_period      195224 non-null  category
 15  Offer_publication_date 195226 non-null  datetime64[ns]
 16  Offer_location         195226 non-null  object
 17  Offer_region           195226 non-null  object
 18  Features               195226 non-null  object
dtypes: category(1), datetime64[ns](1), float64(4), int64(1), object(12)
memory usage: 28.5+ MB
```

Kiểm tra lại các danh sách phân loại dữ liệu sau khi chỉnh sửa thuộc tính:

- Các cột có kiểu dữ liệu numerical là: ['Price', 'Mileage_km', 'Power_HP', 'Displacement_cm3']
- Các cột có kiểu dữ liệu categorical là: ['Condition', 'Vehicle_brand', 'Fuel_type', 'Drive', 'Transmission', 'Type', 'Colour', 'Offer_location', 'Features', 'Doors_number', 'Vehicle', 'Production_period', 'Offer_region']

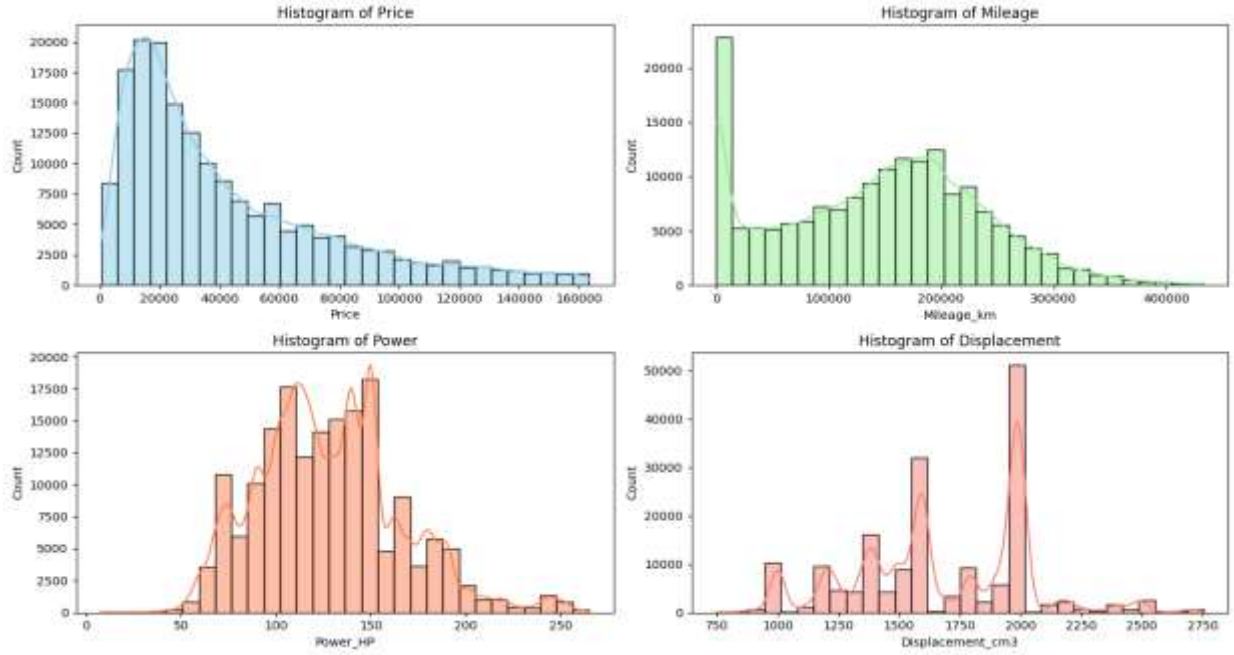
- Các cột có kiểu dữ liệu datetime là: ['Offer_publication_date', 'Production_year']

4.2. Phân tích đơn biến

4.2.1 Biến định lượng

	Price	Mileage_km	Power_HP	Displacement_cm3
count	195226.000000	195226.000000	195226.000000	195226.000000
mean	50777.184098	141785.731683	138.259202	1757.378182
std	48284.398447	95616.596719	50.472770	468.084447
min	585.000000	1.000000	1.000000	400.000000
25%	16900.000000	61200.000000	105.000000	1461.000000
50%	32999.000000	149000.000000	130.000000	1686.000000
75%	68900.000000	208239.000000	163.000000	1995.000000
max	250600.000000	657000.000000	365.000000	3518.000000

Phân phối từng biến định lượng:



Nhận xét:

- **Biến Price:**
  + Dựa vào số liệu thống kê của biến Price, có thể nhận thấy rằng giá của các loại xe sau khi được xử lý nằm trong khoảng từ 585 PLN đến 250.600 PLN, vì đây là bộ dữ liệu về các loại xe cũ nên có thể thấy rằng một số xe cũ có giá rất rẻ, có thể là vì xe đã qua sử dụng và có số km đã di chuyển cao, còn một số xe khác lại có giá rất cao trên thị trường thì có thể đó là những chiếc xe cổ, được sản xuất giới hạn và có giá trị thương mại cao. Giá trị trung bình là 50.777 PLN cao hơn trung vị là 32.999

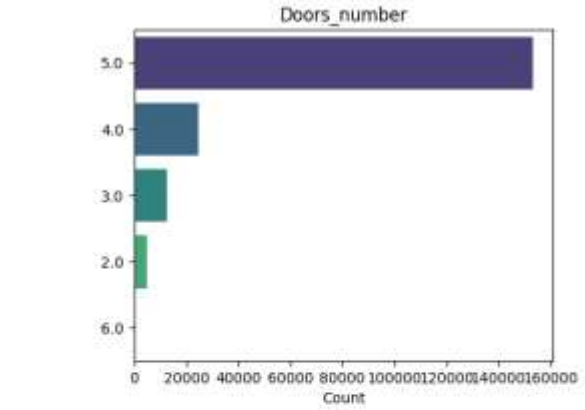
PLN, cho thấy sự ảnh hưởng của các mẫu xe có giá cao đặc biệt. Độ lệch chuẩn lớn, đạt 48.284 USD, cho thấy sự biến động đáng kể trong miền giá trị. Phần lớn (75%) các mẫu ô tô trong tập dữ liệu có giá dưới 68.900 PLN, trong khi 25% còn lại có giá trên mức này, chỉ ra sự phân bố không đồng đều của giá trị.

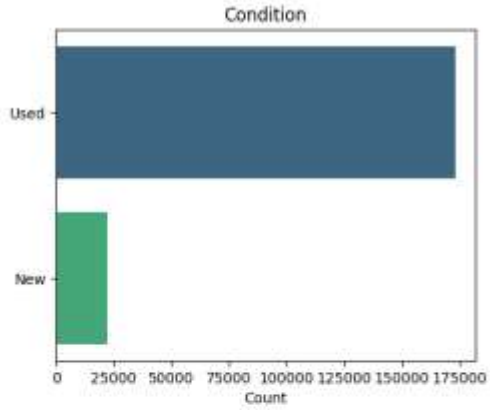
+ Dựa vào biểu đồ trực quan có thể thấy được tính chất lệch phải của dữ liệu, cho thấy một số lượng lớn các ô tô có giá thấp hơn so với số lượng các ô tô có giá cao. Điều này làm nổi bật sự không đồng đều trong phân phối giá trị của biểu Price trong tập dữ liệu.

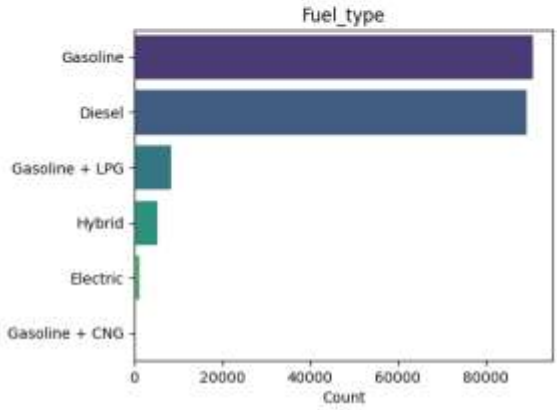
- **Biến Mileage_km:**
  + Dựa vào phân tích số liệu của biến Mileage_km, có thể nhận thấy một sự đa dạng lớn trong dữ liệu, với giá trị thấp nhất là 1 km và giá trị cao nhất là 657.000 km. Điều này cho thấy sự đối lập giữa các xe mới và những chiếc xe đã chạy rất nhiều. Trung bình Mileage_km đạt 141.785 km, là một con số đáng chú ý, đồng thời thể hiện rằng hầu hết các xe trong tập dữ liệu đã trải qua một quãng đường di chuyển đáng kể.
  + Độ lệch chuẩn là 95.616 km, cho thấy sự đồng đều trong dữ liệu và mức độ biến động không lớn, đặc biệt là phần lớn các xe có mileage không chênh lệch quá nhiều so với trung bình. Chỉ số phân vị (25%, 50%, 75%) thể hiện rõ xu hướng phổ biến với một số lượng đáng kể các xe nằm trong khoảng từ 61.200 km đến 208.239 km. Điều này làm nổi bật thị trường ô tô đã qua sử dụng, với nhiều xe có mức độ di chuyển ở mức trung bình và cao.
- **Biến Power_HP:**
  + Dựa vào phân tích số liệu của biến Power_HP, ta có thể nhận thấy một độ đa dạng đáng kể trong công suất động cơ của các mẫu xe. Với giá trị thấp nhất là 8 HP và giá trị cao nhất là 365 HP, cho thấy sự khác biệt giữa các mô hình xe.
  + Với giá trị trung bình lên đến 138,25 HP, tập dữ liệu này thể hiện một trung bình công suất động cơ khá lớn, có sự biến động đáng kể. Độ lệch chuẩn là 50,43 HP, cho thấy sự đa dạng lớn giữa các mẫu xe, từ những chiếc xe có công suất động cơ yếu cho đến những xe có công suất động cơ mạnh hơn.
  + Các giá trị phân vị từ 105 đến 163 mã lực cho thấy hầu hết xe có công suất tập trung trong khoảng đó.
- **Biến Displacement_cm3:**
  + Dựa vào phân tích số liệu của biến Displacement_cm3, ta có thể nhận thấy sự đa dạng rõ ràng trong dung tích xi lanh của các loại xe. Từ giá trị thấp nhất là 400 cm3 đến giá trị cao nhất là 3.518 cm3, có sự biến động đáng kể, cho thấy sự linh hoạt và đa dạng trong các thông số động cơ của thị trường ô tô.
  + Với dung tích trung bình là 1.757 cm3, có thể nhận thấy rằng đa số các xe trong tập dữ liệu có dung tích xi lanh khá lớn. Độ lệch chuẩn lên đến 468 cm3, thể hiện sự biến động đáng kể trong dung tích xi lanh của các xe, làm nổi bật sự đa dạng và sự lựa chọn rộng lớn từ phía người tiêu dùng.
  + Các giá trị phân vị từ 1.461 cm3 đến 1.995 cm3 cho thấy đa số xe trong tập dữ liệu có dung tích xi-lanh ở mức trung bình đến cao.

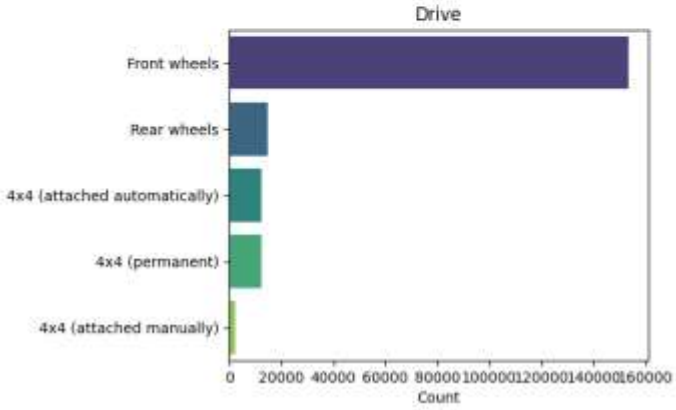
4.2.2 Biến định danh

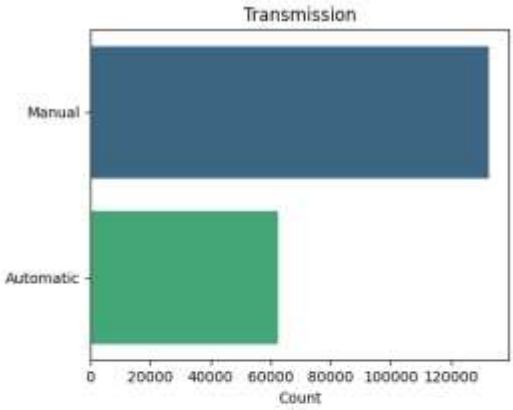
a) Biến định danh có dưới 10 giá trị unique

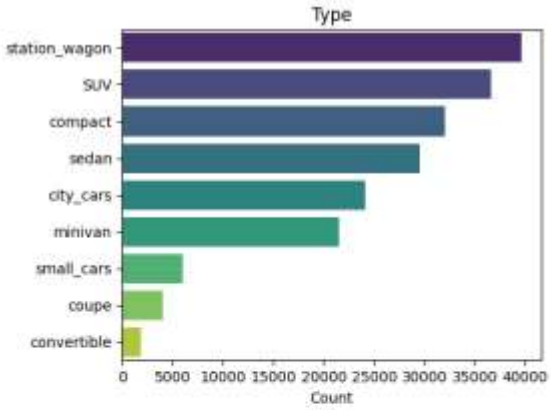


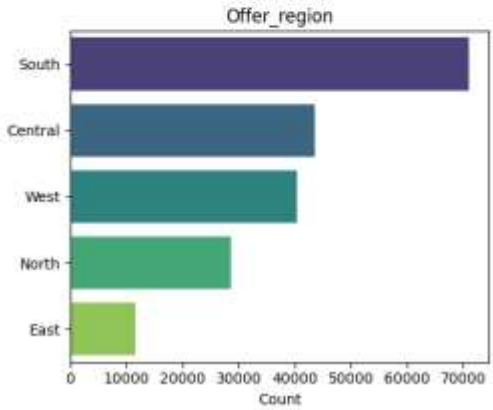












Nhận xét: Thông qua các biểu đồ, có thể thấy rõ sự phân bố giữa các giá trị trong các biến:

- **Door_numbers:**

+ Trong biến Door_numbers, việc xe 5 cửa chiếm đa số với hơn 140.000 chiếc là một biểu hiện rõ ràng của sự thịnh hành và sự lựa chọn ưu tiên của người tiêu dùng. Sự ưa chuộng của mô hình này có thể phản ánh nhu cầu ngày càng tăng cao về tính tiện ích và linh hoạt trong việc sử dụng xe hơi. Mô hình 5 cửa thường mang lại sự thuận tiện khi cần sử dụng không gian nội thất, đồng thời cũng tạo ra một trải nghiệm lái xe thoải mái và dễ dàng.
+ Sự chênh lệch đáng kể giữa số lượng xe 5 cửa và 4 cửa, với tỷ lệ lên đến 5 lần, thậm chí có thể đánh dấu một xu hướng thị trường rõ ràng, trong đó người tiêu dùng đặt ưu tiên cao vào thiết kế có sự tiện ích và tính linh hoạt hơn, thay vì sự hạn chế của mô hình 4 cửa. Các nhà sản xuất ôtô có thể thấy được nhu cầu này và đã tích hợp chúng vào nhiều dòng xe khác nhau để đáp ứng đa dạng và thay đổi của người mua.

- **Condition:**

+ Với hơn 175.000 chiếc xe cũ so với chỉ gần 25.000 chiếc xe mới được quảng cáo, đặt ra câu hỏi về sự ưa chuộng và ưu tiên của người tiêu dùng. Số lượng lớn của xe đã qua sử dụng có thể phản ánh nhu cầu chủ yếu của thị trường trong việc mua và bán xe cũ, có thể do những ưu điểm như giá trị hấp dẫn, sự đa dạng về mô hình và lựa chọn.
+ Có lẽ cơ sở kinh doanh này muốn hướng đến tệp khách hàng ưu tiên quan tâm đến các xe đã qua sử dụng để tận dụng giá trị cao hơn và giảm chi phí sở hữu xe, đồng thời có sự đa dạng trong việc chọn lựa dựa trên sự phong phú của thị trường xe đã qua sử dụng.

- **Fuel_type:**

+ Sự phân hóa đáng kể giữa các loại nhiên liệu trong bộ dữ liệu là một chi tiết quan trọng, với Gasoline và Diesel là hai loại nhiên liệu nổi bật nhất. Sự ưa chuộng của Gasoline và Diesel có thể phản ánh sự lựa chọn của người tiêu dùng dựa trên nhu cầu và ưu tiên cá nhân.
+ Gasoline thường được ưa chuộng với khả năng cung cấp hiệu suất động cơ tốt và là lựa chọn phổ biến cho các dòng xe hơi cá nhân. Điều này có thể liên quan đến sự dễ dàng trong việc sử dụng, giá thành thấp và sự phổ biến của các trạm xăng.
+ Trái lại, Diesel thường được lựa chọn cho các loại xe có nhu cầu vận chuyển nặng, như xe tải và xe du lịch chuyên dụng. Điều này thường do Diesel mang lại hiệu suất năng lượng tốt, đặc biệt là ở tốc độ cao và trong điều kiện đường hầm. Đối với một số người tiêu dùng, sự tiết kiệm nhiên liệu và khả năng vận hành trong điều kiện nặng có thể là lý do chọn Diesel.
+ Sự phân hóa rõ rệt giữa các loại nhiên liệu thể hiện sự đa dạng trong thị trường và sự phản ánh của người tiêu dùng về lựa chọn xe hơi phù hợp với nhu cầu cá nhân và môi trường lái xe.

- **Drive:**

+ Hệ thống dẫn động Front wheels chiếm phần lớn trong bộ dữ liệu, có lẽ là do sự ưa chuộng về hiệu suất và chi phí sử dụng của loại hệ thống này. Hệ thống dẫn động Front wheels thường mang lại sự ổn định và hiệu suất tốt, đặc biệt là trên các đường bề mặt khó khăn.

Điều này có thể đáp ứng nhu cầu của những người muốn sở hữu xe với chi phí vận hành và bảo dưỡng thấp, mà vẫn giữ được sự linh hoạt trong lái xe hàng ngày.

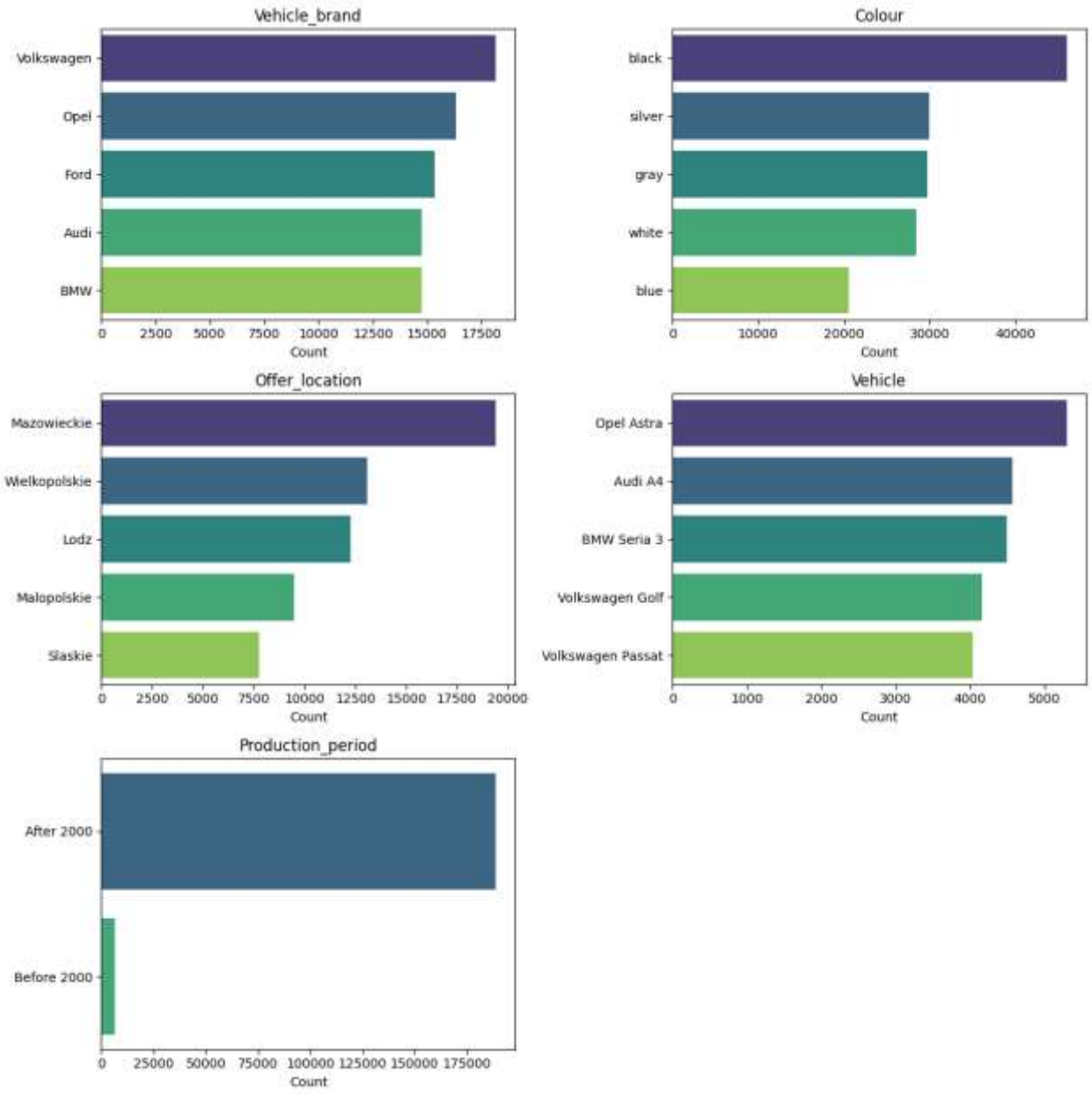
- **Transmission:**

+ Sự ưu tiên cho hộp số Manual với hơn 130.000 chiếc có thể phản ánh sự ưa thích của một phần đáng kể người tiêu dùng đối với trải nghiệm lái xe thủ công. Hộp số Manual thường mang lại cảm giác lái xe tích cực và tương tác chặt chẽ giữa người lái và chiếc xe. Điều này tạo ra một trải nghiệm lái xe độc đáo, nơi người lái có sự kiểm soát cao hơn, cảm nhận được tốc độ và công suất của xe thông qua việc thay đổi số và quản lý đồng thời cảm nhận của mình với đường đua.
+ Mặc dù hộp số Manual chiếm ưu thế với hơn 130.000 chiếc trong bộ dữ liệu, nhưng cũng không thể phủ nhận sự phổ biến và thuận tiện của hộp số Automatic, đặc biệt là trong môi trường lái xe hàng ngày. Với hơn 60.000 chiếc, hộp số Automatic đóng vai trò quan trọng trong sự đa dạng của thị trường ôtô và là lựa chọn ưa chuộng của một phần lớn người tiêu dùng. Hộp số Automatic thường được đánh giá cao về tính tiện lợi, đặc biệt là trong các điều kiện giao thông đô thị, đôi khi làm giảm mệt mỏi cho người lái. Chúng giúp giảm độ phức tạp trong quá trình lái xe, đặc biệt là khi phải di chuyển trong các đoạn đường đông đúc hoặc khi lái xe trong các thành phố lớn. Sự dễ sử dụng của hộp số Automatic cũng là một yếu tố quan trọng thuận lợi cho những người mới học lái xe.
+ Tóm lại, sự đa dạng giữa Hộp số Manual và hộp số Automatic trong bộ dữ liệu là biểu hiện rõ ràng của sự đáp ứng của thị trường ô tô đối với sự đa dạng và đối tượng người mua có nhu cầu khác nhau về trải nghiệm lái xe.

- **Offer_region:**

+ Sự chênh lệch đáng kể giữa khu vực phía Nam và phía Đông trong phân phối xe hơi có thể là một phản ánh rõ ràng của sự không đồng đều về cơ sở hạ tầng, nhu cầu sử dụng, và sự phát triển kinh tế giữa các vùng trong quốc gia. Các yếu tố này đóng vai trò quan trọng trong việc xác định mức độ tiếp cận và sử dụng xe hơi, tạo nên một bức tranh phong phú về thị trường ôtô ở nước này.
+ Các vùng Central và West, xấp xỉ nhau, có thể thể hiện một sự cân bằng hơn trong việc phân phối xe hơi. Các yếu tố như sự đồng đều về phát triển kinh tế, cơ sở hạ tầng và nhu cầu sử dụng có thể tạo ra một thị trường ô tô ổn định và đồng đều hơn ở các vùng này.
+ Tổng cộng, sự chênh lệch lớn trong phân phối xe hơi giữa các vùng phản ánh sự đa dạng và thách thức của thị trường ô tô trong việc đáp ứng nhu cầu và điều kiện đặc biệt của từng khu vực trong quốc gia.

b) Biến định tính có trên 10 giá trị unique:



Nhận xét:

- **Vehicle_brand:**

+ Trong thị trường ôtô, sự độc đáo và đa dạng của các thương hiệu hàng đầu thể hiện sự cạnh tranh khốc liệt và nhu cầu đa dạng của người tiêu dùng. Volkswagen, với khoảng 17.500 chiếc xe được quảng cáo, không chỉ là một thương hiệu lớn mà còn nổi tiếng với sự đa dạng của các mô hình, từ sedan đến SUV. Sự linh hoạt này không chỉ giúp thương hiệu này chiếm lĩnh một vị trí đầu bảng mà còn đáp ứng linh hoạt với sự biến động của thị trường và sở thích của khách hàng.
+ Opel và Ford cũng chiếm một tỷ trọng quan trọng trong danh sách, với số lượng xe quảng cáo không kém cạnh. Cùng với đó, Audi và BMW cũng giữ vị trí đáng chú ý, thể hiện sức hút của các mô hình hạng sang trong phân khúc này.

+ Sự đa dạng này trong danh sách các thương hiệu hàng đầu không chỉ là kết quả của sự cạnh tranh mà còn là dấu hiệu của sự phát triển và tiến bộ trong ngành công nghiệp ôtô, nơi sự sáng tạo và đáp ứng nhanh chóng là chìa khóa để giữ vững và mở rộng thị trường.

- **Colour:**

+ Màu sắc không chỉ là một yếu tố trang trí, mà còn đóng một vai trò quan trọng trong quá trình quyết định mua sắm xe hơi. Dữ liệu quảng cáo cho thấy rằng người tiêu dùng có sự ưa thích đặc biệt đối với màu đen, với hơn 40.000 chiếc xe.
+ Ngoài ra, màu bạc, xám, trắng và xanh dương cũng đứng trong danh sách phổ biến, cho thấy sự đa dạng trong sở thích cá nhân của khách hàng.
+ Sự đa dạng trong màu sắc có thấy đây không chỉ là một xu hướng thị trường mà còn là một chiến lược của ngành công nghiệp ô tô để tiên đoán và đáp ứng linh hoạt với nhu cầu đa dạng của người tiêu dùng. Việc tạo ra nhiều lựa chọn cho người mua không chỉ tăng cường trải nghiệm mua sắm mà còn thể hiện cam kết của ngành công nghiệp đối với sự đổi mới và sự linh hoạt trong việc đáp ứng mong muốn ngày càng đa dạng của khách hàng.

- **Offer_location:**

+ Về địa chỉ xe được nhà phát hành cung cấp, gần 20.000 xe có vị trí ở Mazowieckie, việc tập trung lớn nhất tại Mazowieckie không chỉ là một dấu hiệu về sự đô thị hóa mà còn phản ánh sự quan trọng của khu vực này trong bối cảnh kinh tế và thị trường ôtô. Vị trí này thường là trung tâm kinh tế và văn hóa, nơi mà nhu cầu về giao thông cá nhân thường xuyên tăng cao. Sự tập trung lớn ở đây có thể là kết quả của cả yếu tố dân số đông đúc và nhu cầu vận chuyển cao.
+ Ngoài ra, còn có sự đóng góp quan trọng từ các khu vực như Wielkopolskie, Lodz, Malopolskie và Slaskie, đồng đều phân bố trên khắp quốc gia. Sự phân bố rộng rãi của xe ô tô ở các vùng này có thể là kết quả của sự phát triển kinh tế, cũng như nhu cầu vận chuyển đa dạng từ cả thành thị đến vùng nông thôn.
+ Từ cái nhìn chiến lược, sự phân bố rộng rãi này có thể là dấu hiệu của việc ngành công nghiệp ô tô đang tích cực đáp ứng và phát triển không chỉ ở các trung tâm đô thị lớn mà còn ở các khu vực lân cận, thúc đẩy sự tiện lợi và tiếp cận đối với các sản phẩm ôtô trên toàn quốc.

- **Vehicle:**

+ Danh sách các phiên bản xe tiếp tục thể hiện sự ưa chuộng đặc biệt đối với một số mô hình cụ thể. Opel Astra, với hơn 5.000 chiếc, nổi bật là một trong những lựa chọn phổ biến nhất trong dữ liệu. Sự phổ biến của Opel Astra có thể là kết quả của sự đa dạng và linh hoạt trong các tính năng, giá trị, và hiệu suất của mô hình này, thu hút sự quan tâm của đa dạng đối tượng người mua.
+ Tiếp theo, Audi A4 và BMW Seria 3 là những tên tuổi đáng chú ý, đặc biệt là trong phân khúc xe hạng sang. Sự ưa chuộng của chúng có thể phản ánh sự hấp dẫn của các mô hình hạng sang với khả năng kết hợp giữa thiết kế tinh tế, hiệu suất đáng kể và công nghệ tiên tiến. Người mua thường chọn những mô hình này để trải nghiệm sự sang trọng và đẳng cấp khi lái xe.

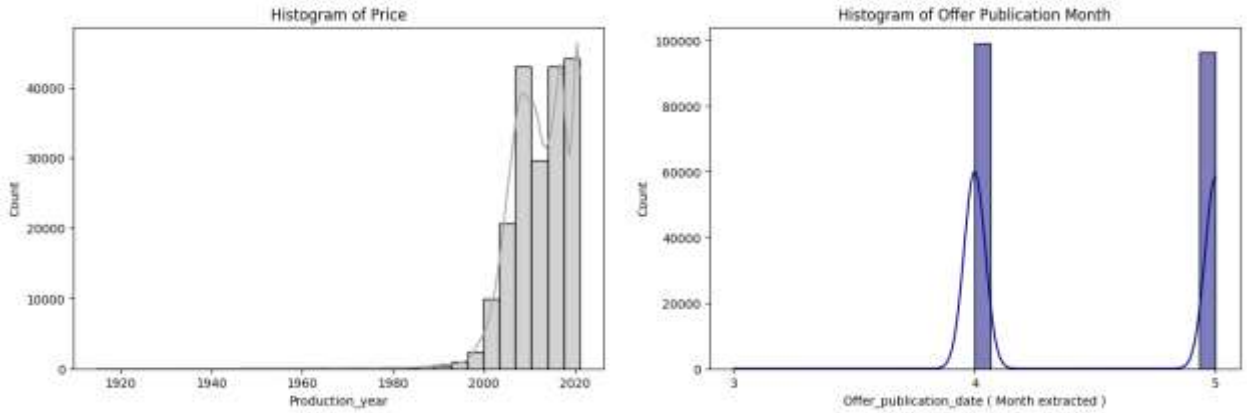
- **Production_period:**

+ Dựa vào dữ liệu năm sản xuất trong tập dữ liệu, có thể thấy rõ xu hướng chung của thị trường ôtô, nơi phần lớn các xe đều được sản xuất sau năm 2000, chiếm tỷ lệ lớn với hơn

175.000 chiếc. Điều này thể hiện sự tiên tiến và phát triển không ngừng của ngành công nghiệp ôtô, với việc áp dụng công nghệ mới, tính năng an toàn và hiệu suất cao vào sản phẩm. Ngoài ra còn cho thấy được xu hướng sử dụng xe ô tô ngày càng tăng.

+ Số lượng xe sản xuất trước năm 2000, mặc dù khá nhỏ với khoảng 10.000 chiếc, thường đánh dấu sự đa dạng và độc đáo trong thị trường. Những chiếc xe cổ điển này có thể đại diện cho sự đam mê của một nhóm nhỏ người mua có sở thích sưu tầm đồ cổ.

4.2.3 Biến thời gian



Nhận xét:

- **Production_year:**

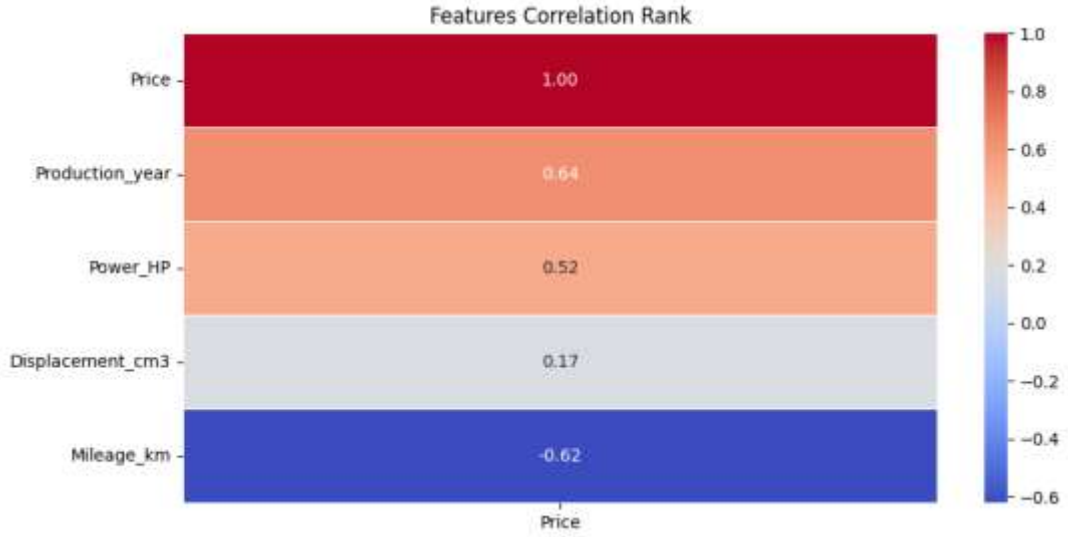
+ Năm sản xuất sau khi tiền xử lý không thay đổi nhiều so với trước đó, ta thấy có phân phối lệch trái, khác với các số liệu định lượng khác, điều này có thể hiểu là các giá trị phân phối nằm ở miền phải cho thấy các số lượng xe được rao bán trong bộ dữ liệu này tăng dần theo năm sản xuất, các xe được sản xuất càng gần với thực tế sẽ được chú trọng nhiều hơn, số xe được rao bán nhiều nhất ở năm 2021.

- **Offer_publication_date (Chia theo tháng):**

+ Số lượng xe được đăng tin quảng cáo chủ yếu vào những tháng 4, 5 và một số ở trong tháng 3. Có thể là do bộ dữ liệu chủ yếu được thu thập vào các tháng này.

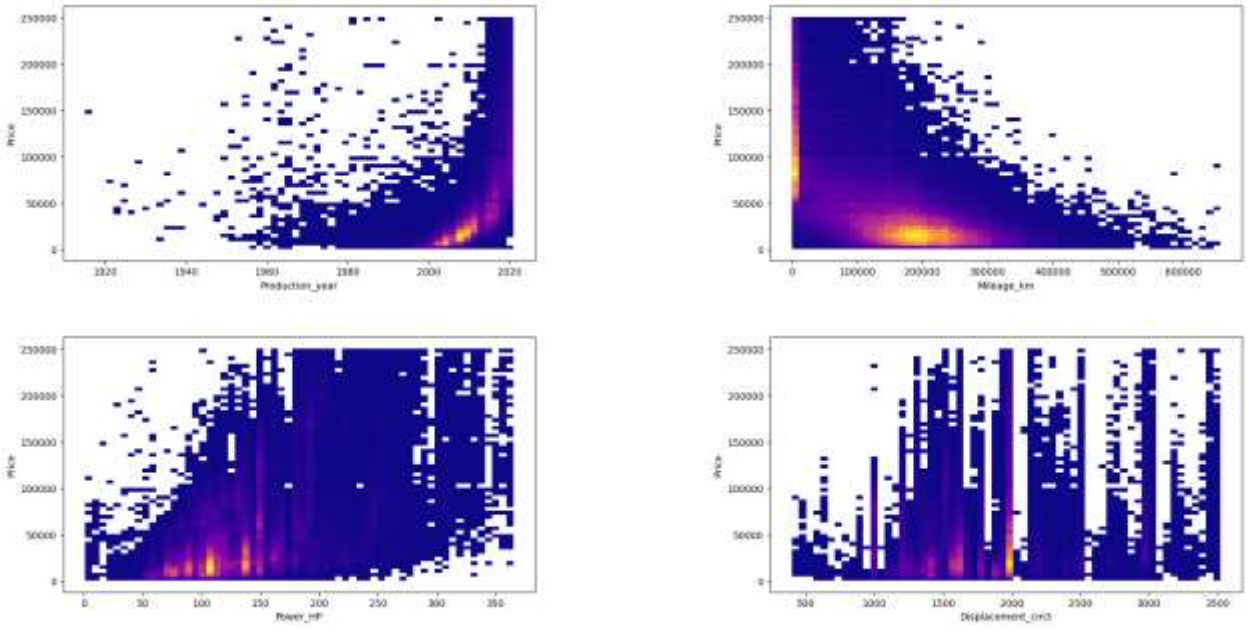
4.3 Phân tích đa biến

4.3.1 Phân tích tổng quan về sự tương quan giữa các biến numeric



Nhận xét: Mức giá bán của các loại có độ tương quan tuyến tính dương tương đối mạnh với Production_year, correlation = 0.64, điều này có nghĩa khi năm sản xuất tăng dần thì các loại xe có giá bán tăng theo.

Ngược lại, quãng đường đi được, hay hiểu cách khác là số odo, correlation = -0.62. Sự tương quan âm thể hiện rõ rệt khi những xe có số quãng đường đã đi được càng cao thì sẽ xu hướng giá sẽ giảm.



Nhận xét:

- **Production_year**: Số lượng xe được rao bán chủ yếu có năm sản xuất từ 1948 về sau. Các quan sát về xe có mức giá phân bổ vào khoảng năm từ 2000 đến 2020 và được bán dưới 50.000 PLN chiếm tỉ lệ rất lớn. Mức giá cũng tăng mạnh theo mốc thời gian sản xuất trong khoảng này.

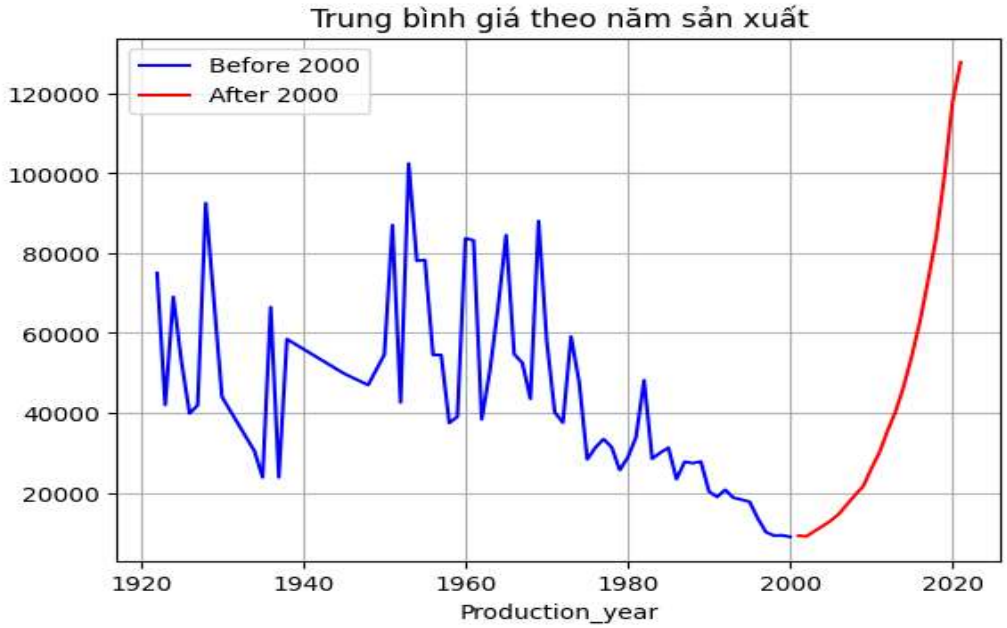
- **Mileage_km**: Các xe có số odo ở khoảng 100.000 - 300.000 km và có định giá dưới 50.000 PLN xuất hiện nhiều nhất. Mức giá tăng dần thì giá xe cũng giảm dần. Tuy nhiên, ở số odo 1 km, mức giá bán biến động rất mạnh và số lượng xe nằm trong khoảng này là tương đối nhiều.

- **Power_HP**: Công suất động cơ (Mã lực) có quan hệ tuyến tính dương tương đối nên khi công suất tăng thì xu hướng định giá cũng tăng, các xe có giá bán dưới 50,000 PLN và công suất trong khoảng 50 - 150 mã lực là tập giá trị phổ biến. Cho thấy các xe chủ yếu được bán có công suất từ 50 - 150 mã lực và có mức giá dưới 50.000 PLN.

- **Displacement_cm3**: Dung tích xilanh không có mối tương quan rõ ràng với định giá, nên các khi giá tăng, dung tích xilanh sẽ không bị ảnh hưởng hoặc ngược lại. Các xe có dung tích ở 1500 cm3 và 2000 cm3 được bán và chủ yếu ở dưới 50.000 PLN, quan sát ở mốc này là khá giống nhau.

4.3.2 Phân tích sự tương quan giữa giá xe trung bình và năm sản xuất

Năm sản xuất và giá xe được xem xét như hai yếu tố quan trọng, chịu ảnh hưởng lớn từ các yếu tố ngoại vi, tạo ra sự biến động trong sản lượng và giá cả của xe hơi hàng năm. Nhìn chung, số lượng xe sản xuất và mức giá có thể biến động tùy thuộc vào các điều kiện thị trường và kinh tế, tạo ra sự chênh lệch đáng kể về giá trung bình giữa các năm. Bằng cách phân tích trực quan và sử dụng các thống kê mô tả, nhóm hy vọng sẽ đưa ra những nhận định chính xác hơn về mối quan hệ giữa giá xe và năm sản xuất.



	Tương quan	Độ lệch chuẩn	Độ biến động
Trước 2000	-0.569	22965.820	52.300
Sau 2000	0.930	36905.472	83.823

Nhận xét: Các xe được sản xuất trước năm 2000 có mức định giá biến động không đồng đều, có nhiều mức giá trị lớn bất thường dẫn đến đường phân bố bị chia cắt mạnh và xu hướng chung giảm dần và đạt đáy ở 2000. Sau 2000, ta thấy quan hệ tuyến tính rõ rệt giữa năm sản xuất và mức định giá của xe được bán và đạt đỉnh tại mốc thời gian thu thập dữ liệu.

Trước năm 2000:

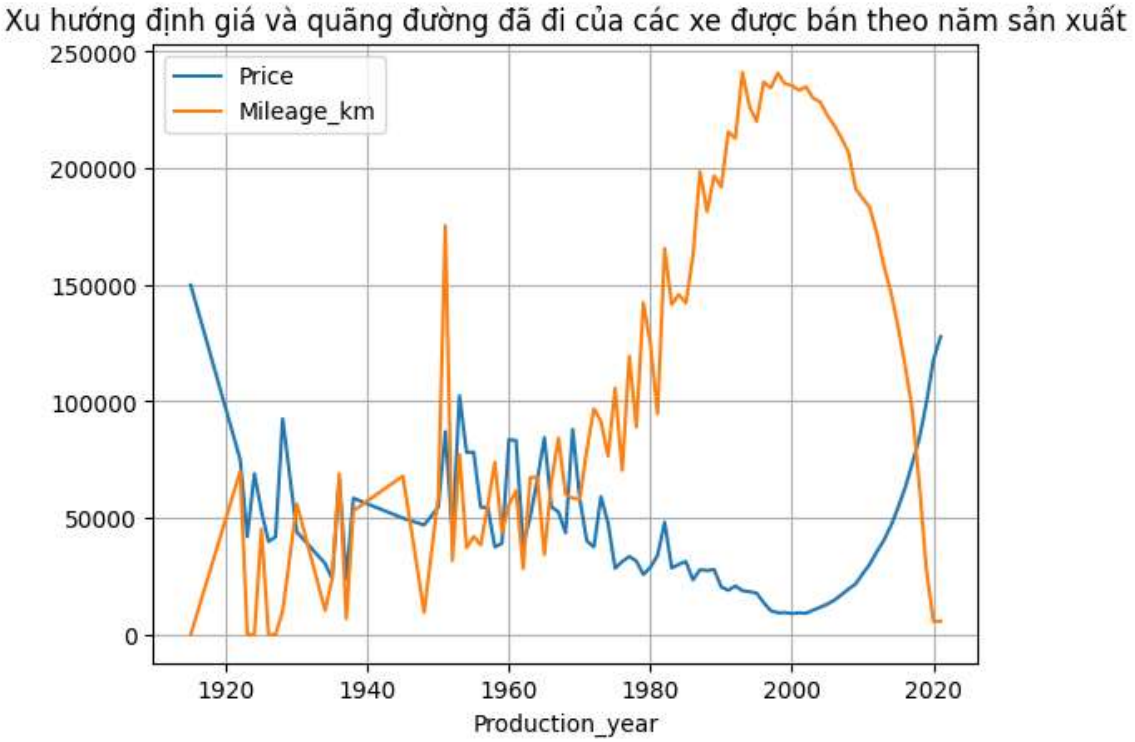
- Ta thấy tương quan của những xe trước năm 2000 có giá trị ước lượng là -0.569, điều này cho thấy các xe có năm sản xuất từ trước năm 2000 đều có xu hướng định giá giảm dần theo năm sản xuất.
- Độ lệch chuẩn khá cao (22,965.820) cho thấy có sự phân tán lớn trong giá xe so với trung bình trước năm 2000.
- Hệ số biến động là 52.3%, cho thấy có một mức độ biến động đáng kể trong giá xe trước năm 2000 với khoảng giá trị nhỏ nhất là 9.011 đến 102.388 PLN. Mức chênh lệch lên đến hơn 100 lần.

Sau năm 2000:

- Tương quan của những xe sau năm 2000 có giá trị ước lượng là 0.93, điều này có ý nghĩa các xe có năm sản xuất từ trước năm 2000 đều có xu hướng định giá giảm dần theo năm sản xuất.
- Độ lệch chuẩn còn cao hơn (36,905.384), điều này cho thấy mức độ phân tán của giá xe so với trung bình là còn lớn hơn sau năm 2000
- Hệ số biến động tăng lên đáng kể (83.823%), phản ánh một mức độ không ổn định cao hơn nhiều trong giá xe so với trước năm 2000 với khoảng giá trị nhỏ nhất là 9.121 đến 127.726 PLN. Mức chênh lệch là cao hơn so với trước năm 2000.

4.3.3 Phân tích xu hướng định giá xe trung bình theo mileage_km và năm sản xuất

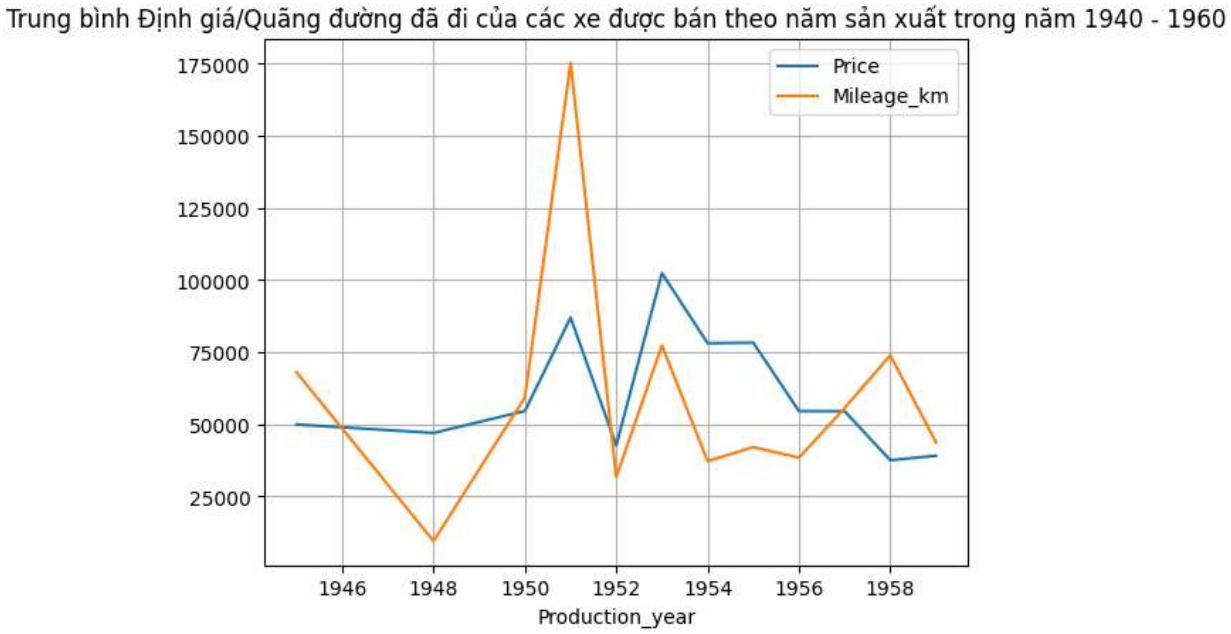
Sau khi đã có cái nhìn tổng quan về biến động giá trung bình theo năm sản xuất, nhóm nghiên cứu quyết định tiến xa hơn để xác định những yếu tố chi tiết ảnh hưởng đến giá xe. Thông qua phân tích tổng quan về dữ liệu đã chỉ ra rằng phần lớn các mẫu xe trong tập dữ liệu là xe đã qua sử dụng. Do đó, quãng đường di chuyển có thể là một yếu tố quan trọng ảnh hưởng đến giá bán. Nhóm sẽ thực hiện một phân tích trực quan thông qua biểu đồ để xem xét mối quan hệ giữa quãng đường di chuyển trung bình, giá xe trung bình và năm sản xuất.



Nhận xét:

- Thông qua phân tích giá ở trên và biểu đồ thể hiện xu hướng định giá và quãng đường đã đi trung bình của các xe theo năm sản xuất, có thể thấy được sự biến động mạnh của trung bình quãng được đi theo năm sản xuất. Có nhiều giá trị khá lớn được ghi nhận ở khoảng thời gian từ 1940 - 1960. Sau đó đường xu hướng trung bình số odo sau đó tăng mạnh và đạt đỉnh đến năm 2000, sau đó quãng đường đi được giảm mạnh về năm 2020.

- Khi trung bình quãng đường các xe được ghi nhận ở các xe được bán càng cao thì giá của nó càng giảm (như ở khoảng 2000). Rút ra kết luận rằng khi các xe này có chỉ số quãng đường đi được càng nhiều sẽ làm ảnh hưởng đến các yếu tố động cơ lẫn các yếu tố bên ngoài (vỏ xe, kính, ...) và làm giá rao bán bị giảm đáng kể.



Quan hệ giữa định giá và quãng đường đã đi thể hiện mối tương quan tuyến tính âm ở hầu hết các giai đoạn sản xuất, tuy nhiên ở một vài giai đoạn sản xuất của xe, mối quan hệ tuyến tính này thể hiện không rõ ràng (năm sản xuất từ năm 1940 - 1960). Bằng tính toán, hệ số tương quan giữa các xe có năm sản xuất trong giai đoạn này có hệ số tương quan dương = 0.21.

```
corr_4060 = df[(df['Production_year'] > 1940) & (df['Production_year'] < 1960)][['Price','Mileage_km']].corr().iloc[1,0]

print('Tương quan giá / quãng đường ( trung bình ) đi được từ 1940-1960:', corr_4060)
```

Output:

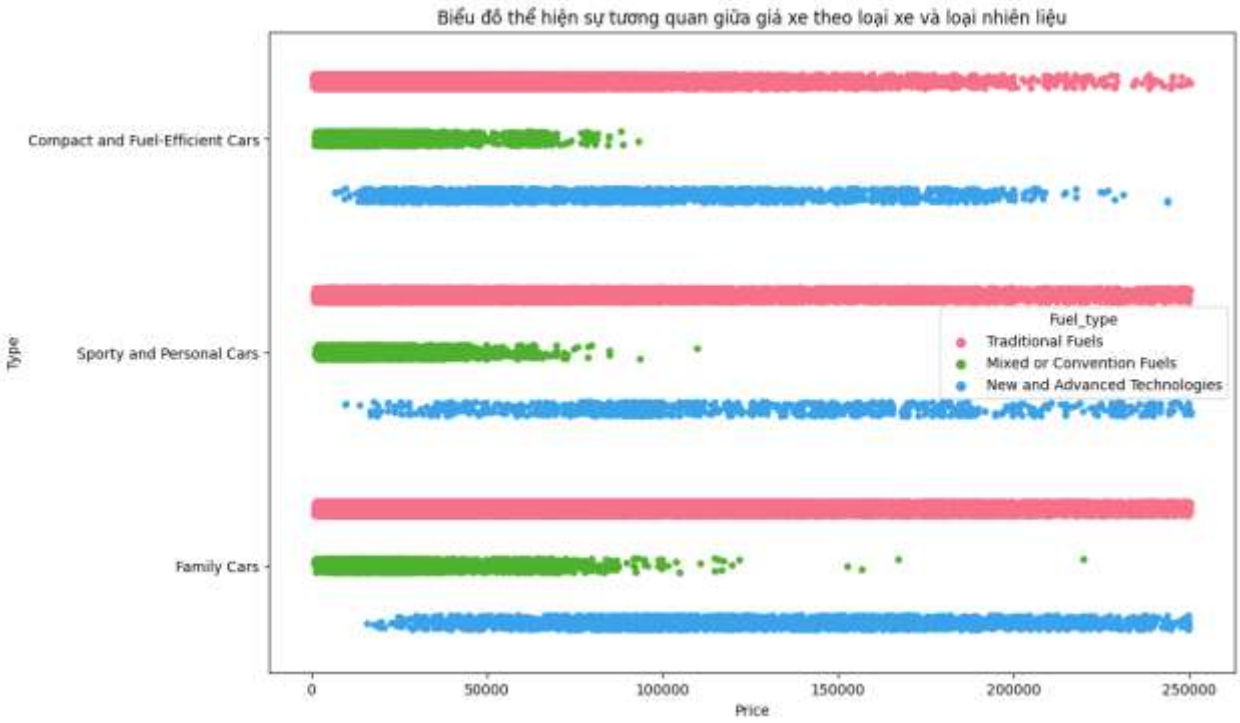
```
Tương quan giá / quãng đường (trung bình) đi được từ 1940-1960:
0.21107280371033416
```

Nhận xét: Mặc dù hệ số tương quan Pearson không quá cao nhưng cũng nói lên được rằng có một số nhất định những loại xe được sản xuất ở khoảng thời gian này có thể được định giá không thông qua mối tương quan của sự gia tăng giữa quãng đường đã đi. Có thể ở các xe được sản xuất trong giai đoạn này, có nhiều chiếc được định giá là xe cổ, số lượng xe giới hạn dẫn đến giá thành cao hơn dù đã là xe qua sử dụng.

4.3.4 Phân tích sự tương quan giữa giá xe trung bình, loại xe và loại nhiên liệu mà xe sử dụng

Có nhiều yếu tố khác có thể có ảnh hưởng đáng kể đến giá xe, chẳng hạn như loại xe và loại nhiên liệu xe sử dụng. Để hiểu rõ hơn về tác động của những yếu tố này, nhóm tiếp tục thực hiện phân tích chi tiết về sự tương quan giữa các yếu tố như loại xe, loại nhiên liệu và giá xe.

Vì các loại xe trong dữ liệu có các đặc điểm tương đồng với nhau về mục đích sử dụng, do đó chúng ta sẽ gom chúng lại thành 3 nhóm là: *Nhóm xe có kích thước nhỏ gọn và tiết kiệm nhiên liệu, xe cá nhân & thể thao*, và cuối cùng là *xe gia đình* để dễ dàng thấy được sự khác nhau giữa các nhóm. Tương tự, ta cũng sẽ gộp các loại nhiên liệu thành 3 nhóm là: *nhiên liệu truyền thống*, *nhiên liệu bán truyền thống* và *nhiên liệu mới*.



Nhận xét:

Thông qua việc phân tích biểu đồ, chúng ta có thể rõ ràng nhận thấy xu hướng chung trong ba nhóm xe khác nhau. Tất cả đều có sự đa dạng trong việc sử dụng cả nhiên liệu truyền thống và nhiên liệu mới. Mức giá bán của các loại xe này phản ánh sự đa dạng này, với một phổ giá trải dài. Điều đáng chú ý là số lượng xe được bán ở mức giá cao hơn trong cả ba nhóm so với nhóm sử dụng nhiên liệu bán truyền thống.

Nguyên nhân dẫn đến vấn đề này có thể là do:

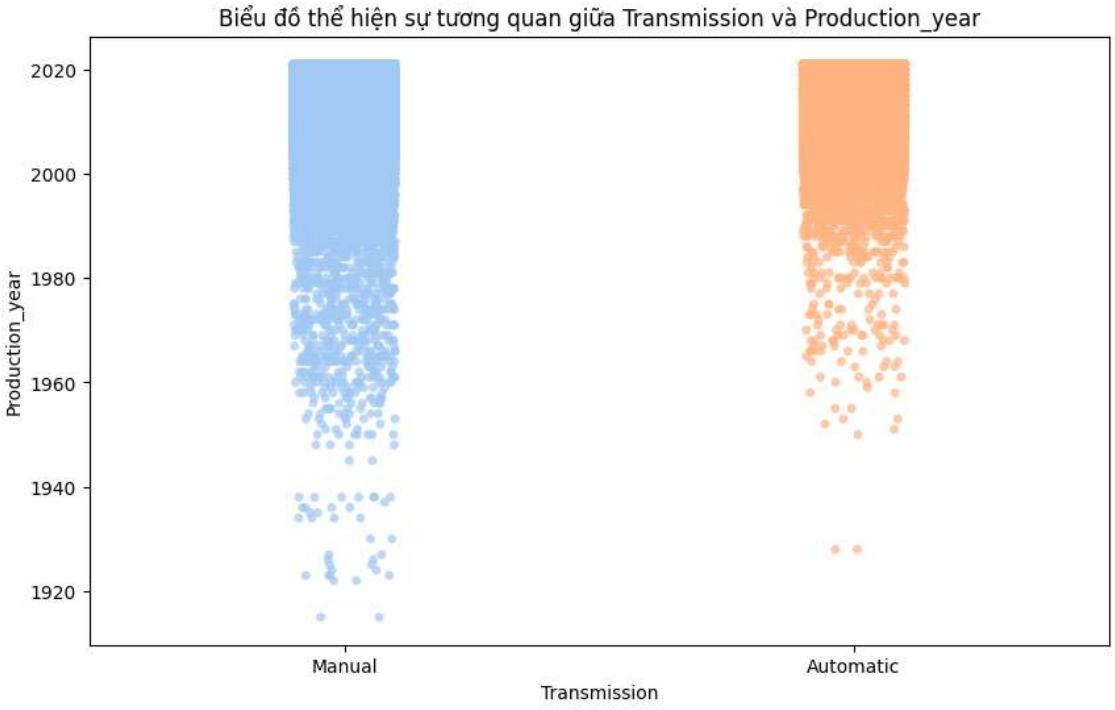
+ Xe sử dụng nhiên liệu truyền thống có độ phổ biến cao trong giai đoạn trước năm 2000 và kể cả sau năm 2000 do các loại nhiên liệu như Gasoline và Diesel dễ tiếp cận.

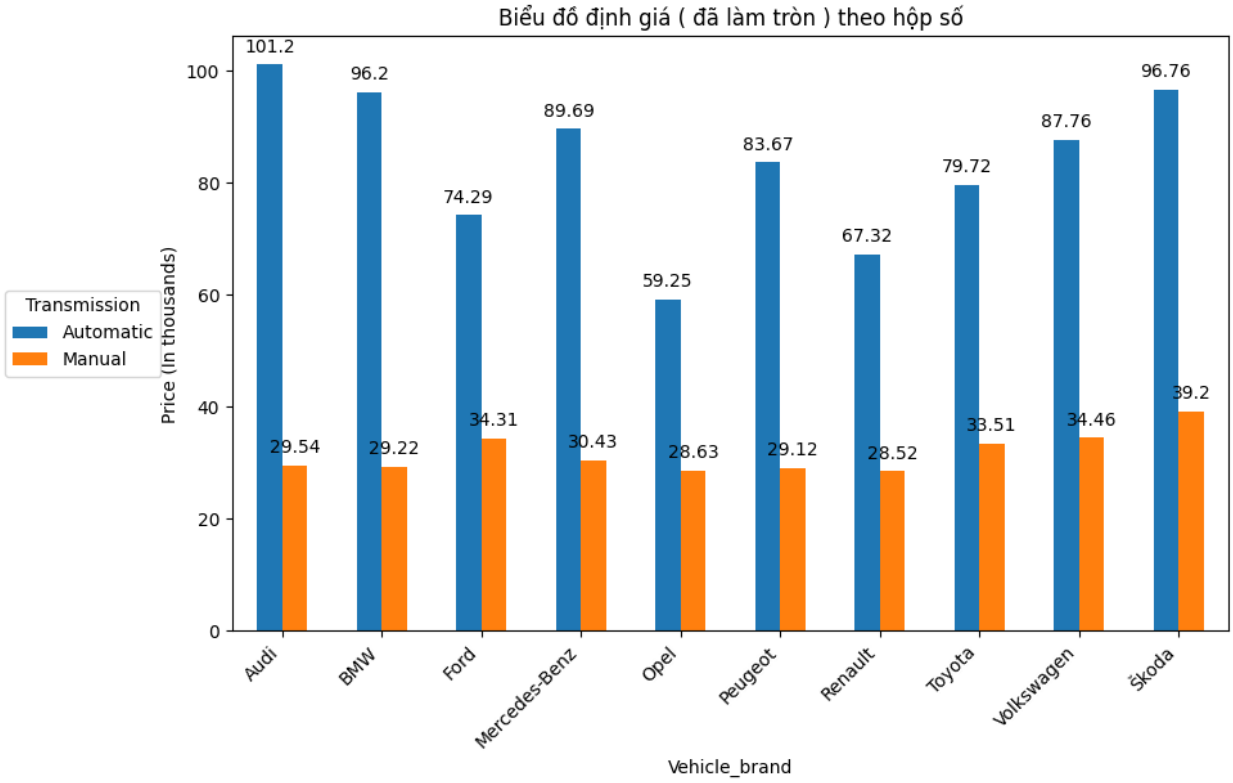
+ Xe sử dụng nhiên liệu hiện đại như Hybrid và Electric dù chỉ mới được sử dụng nhiều trong khoảng thời gian gần đây nhưng do được trang bị công nghệ hiện đại đi kèm với độ thân thiện với môi trường nên làm chi phí đắt hơn.
+ Xe sử dụng nhiên liệu bán truyền thống như Gasoline + LPG và Gasoline + CNG có giá thấp hơn 2 loại trên có thể là do đây là sự kết hợp động cơ điện và động cơ xăng, chi phí sản xuất thường thấp hơn xe điện do sử dụng công nghệ đơn giản hơn nên nó sẽ rẻ hơn xe điện. Cũng như xe sử dụng nhiên liệu bán truyền thống thường được hưởng ưu đãi thuế so với xe xăng, do đó có thể giúp giảm giá thành.

Thấy được sự ưu tiên đối với sử dụng nhiên liệu truyền thống, nhóm xe này sử dụng nhiên liệu này vẫn giữ vững vị thế trong thị trường. Đồng thời, biểu đồ cũng phản ánh sự chuyển đổi và chuyển giao từ xe sử dụng nhiên liệu truyền thống sang các loại xe sử dụng nhiên liệu mới như Electric hay Hybrid. Điều này là một dấu hiệu của sự nhạy bén và thích nghi của thị trường ô tô đối với các xu hướng và công nghệ mới. Điều này có thể thúc đẩy ý thức bảo vệ môi trường trong cộng đồng xe hơi và khuyến khích sự chuyển đổi đối với các phương tiện sử dụng nguồn năng lượng sạch, góp phần tích cực vào mục tiêu bảo vệ môi trường toàn cầu và tiết kiệm nhiên liệu.

4.3.5 Phân tích sự về xu hướng chuyển dịch trong loại hộp số của xe, xem xét sự khác biệt của giá xe theo loại hộp số của top 10 hãng xe được quảng cáo nhiều nhất.

Sau khi tìm hiểu sự tác động của các yếu tố có thể ảnh hưởng đến giá thành của xe, nhóm quyết định mở rộng nghiên cứu để tìm hiểu sâu hơn về những hãng xe được mọi người đặc biệt quan tâm. Ngoài ra nhóm nhận thấy được rằng loại hộp số xe là một yếu tố quan trọng và được khách hàng cân nhắc rất kỹ lưỡng trước khi quyết định mua xe. Nên trong phân tích này nhóm mong muốn biết được giá cả của những hãng xe được quan tâm nhất, song song với đó, sự thay đổi theo thời gian về thị phần của các loại hộp số, có sự khác biệt về giá cả của các xe sử dụng loại hộp số khác nhau hay không.





Nhận xét:

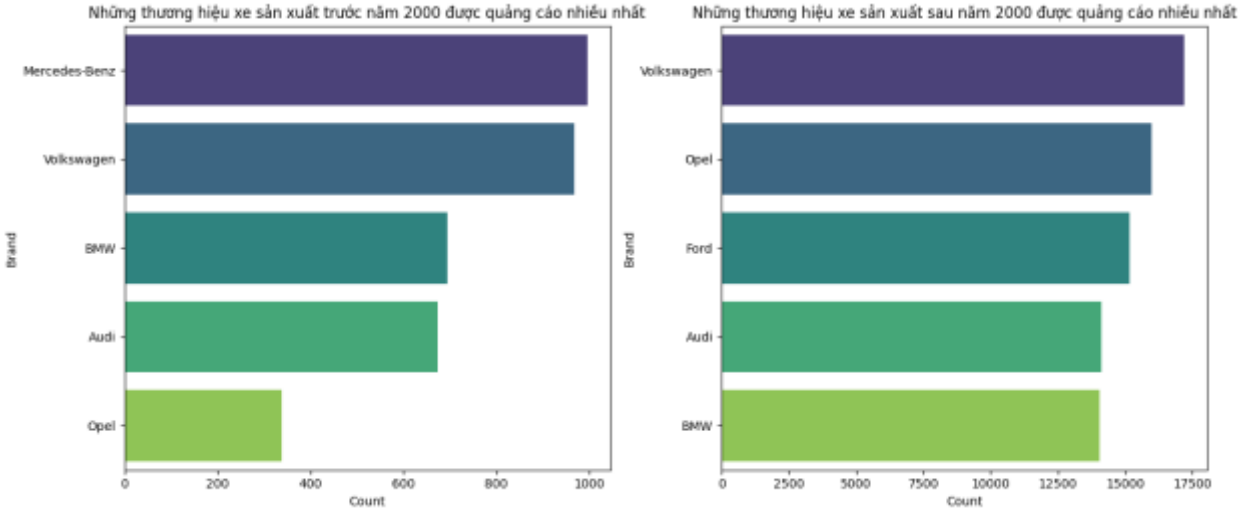
Phân tích biểu đồ đầu tiên, nhóm quan sát thấy được sự tiến triển của công nghệ hộp số trong ngành ô tô. Xe có hộp số sàn được xuất hiện trước các xe có hộp số tự động, dù ra đời và chiếm thị phần trước nhưng kể từ sau năm 1980, các xe sử dụng hộp số tự động đã bắt đầu được sử dụng rộng rãi hơn, phản ánh sự chuyển đổi và đa dạng hóa nhu cầu của người tiêu dùng đối với

loại hộp số. Sự thay đổi này có thể được hiểu như một phản ánh của sự tiện lợi và thoải mái mà hộp số tự động mang lại, phù hợp với các nhóm khách hàng có nhu cầu sử dụng xe đa dạng.

Tuy nhiên, khi quan sát biểu đồ thứ hai, sự khác biệt giữa giá trung bình của xe sử dụng hộp số tự động và hộp số sàn trên cùng một hãng xe đưa ra một góc nhìn mới. Các xe sử dụng hộp số tự động thường có giá trung bình cao hơn đáng kể so với phiên bản sử dụng hộp số sàn. Điều này có thể phản ánh chi phí sản xuất và tính năng tiện ích cao của hộp số tự động. Trong một số trường hợp như Audi và BMW, sự chênh lệch giá giữa hai loại hộp số có vẻ lớn, đặc biệt là khi khách hàng chọn các phiên bản cao cấp với tính năng và hiệu suất tối ưu. Điều này có thể làm nổi bật sự ưa chuộng và sẵn lòng chi trả cao của một số đối tượng khách hàng đối với xe sử dụng hộp số tự động, có thể là do mong muốn trải nghiệm lái xe tiện lợi và sang trọng hơn.

4.3.6 Phân tích yếu tố ảnh hưởng đến sự quảng bá của top 5 hãng xe ở mốc thời gian trước và sau năm 2000

Tiếp theo trong bài phân tích này, nhóm mong muốn phân tích chi tiết hơn về các hãng xe được quan tâm nhiều nhất theo thời gian. Nhóm hi vọng thông qua sự phân tích này có thể tìm ra được những lý do nào khiến cho một dòng xe trở nên phổ biến và được quảng cáo nhiều như vậy. Ngoài ra nhóm còn mong muốn xác định xem có sự khác biệt nào đáng kể về top những xe được quảng cáo nhiều nhất ở 2 giai đoạn trước và sau năm 2000 không. Bằng cách này nhóm mong muốn có cái nhìn trực quan hơn về xu hướng chuyển động của thị trường ô tô từ thế kỷ 20 đến thế kỷ 21.

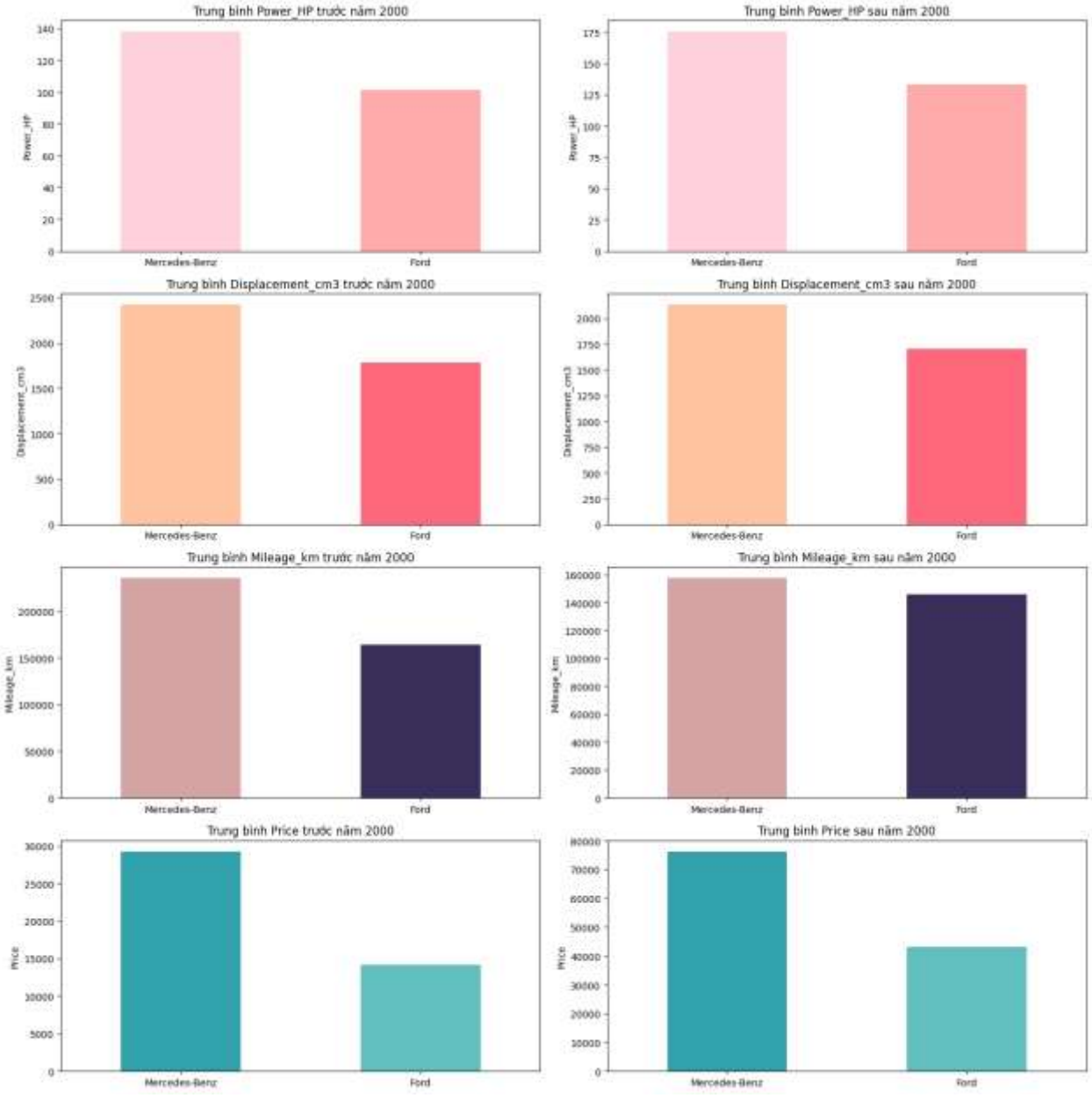


Nhận xét:

- Tổng quan, có sự tăng lên về số lượng xe được sản xuất sau năm 2000 và được quảng cáo nhiều hơn so với các xe sản xuất trước năm đó.
- Trong danh sách các thương hiệu xe sản xuất trước năm 2000, Mercedes-Benz dẫn đầu với hơn 1000 xe được quảng cáo, theo sau là Volkswagen, BMW, Audi và Opel.
- Trong những thương hiệu được sản xuất sau năm 2000, Volkswagen đã từ vị trí thứ 2 vươn lên thành hãng xe được quảng cáo nhiều nhất với hơn 17500 xe, Opel từ vị trí thứ 5 thành

vị trí thứ 2. Trái lại, BMW từ vị trí thứ 3 thành vị trí thứ 5. Mercedes-Benz từ hãng xe được quảng cáo nhiều nhất trong số những hãng sản xuất trước năm 2000 thì bị rơi khỏi top 5 những thương hiệu được sản xuất sau năm 2000.

Để tìm hiểu tại sao lại có sự quảng bá khác biệt đối với các hãng xe sản xuất trước và sau năm 2000, ta sẽ tìm hiểu các yếu tố ảnh hưởng đến sự thay đổi này:



Nhóm quyết định trực quan những yếu tố được đánh giá sẽ có sự ảnh hưởng đến sự thay đổi trong các hãng xe được quảng cáo nhiều thông qua biểu đồ trực quan. Từ đó có thể đi đến những kết luận rằng:

- Dựa vào biểu đồ về công suất động cơ trung bình, có thể thấy được sự tăng lên về công suất động cơ ở cả 2 hãng xe này theo thời gian. Công suất trung bình của hãng Ford sau năm 2000 xấp xỉ bằng với công suất động cơ của xe Mercedes-Benz ở trước năm 2000, ở mức động cơ khoảng 140 HP, có vẽ khách hàng vẫn ưu tiên quan tâm nhiều đến các loại xe có mức công suất động cơ này.

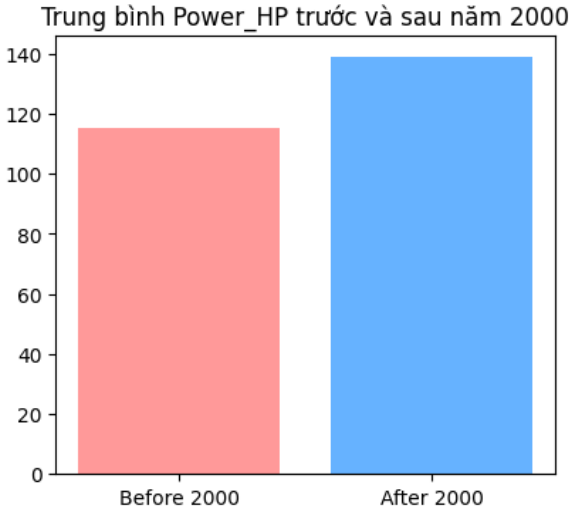
- Còn ở biểu đồ về dung tích xi lanh trung bình, có thể thấy được xe của hãng Mercedes-Benz sản xuất trước năm 2000 có dung tích xi lanh giảm nhẹ so với những xe sản xuất sau năm 2000. Xe của hãng Ford cũng có dung tích xi lanh không quá chênh lệch giữa 2 giai đoạn, do vậy có thể nói rằng các xe ở thuộc hãng Mercedes-Benz đã có sự điều chỉnh để phù hợp và tối ưu hơn trong dung tích xi lanh.

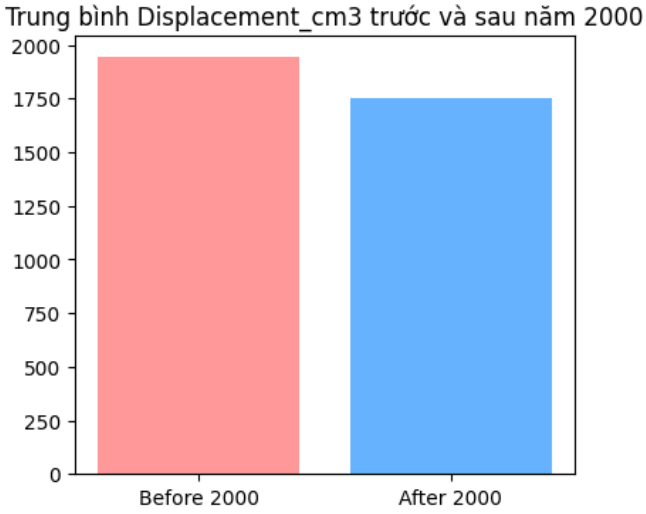
- Về số quãng đường di chuyển trung bình và giá trung bình của các hãng xe, các xe thuộc hãng Mercedes-Benz sản xuất trước năm 2000 có quãng đường di chuyển trung bình lớn hơn đồng thời giá thấp, nên có thể thu hút sự quan tâm của khách hàng nhiều hơn, còn sau năm 2000, trung bình quãng đường di chuyển của các xe thuộc hãng này thấp hơn, đồng nghĩa với việc giá cả của xe sẽ cao hơn, có thể vượt quá khả năng chi trả của khách hàng nên nó không còn được ưu tiên quảng cáo nữa. Còn về hãng xe Ford, mặc dù không có sự thay đổi nhiều về quãng đường di chuyển trung bình của các xe, những giá bán các xe sau năm 2000 vẫn nhỉnh hơn trước đó, tuy nhiên vẫn nằm trong tầm giá 40000, xấp xỉ bằng một nửa giá trung bình của các xe thuộc hãng Mercedes-Benz. Từ đây có thể kết luận rằng càng về sau, có hãng xe đều đã có những điều chỉnh về công suất động cơ để tối ưu hơn về công suất, tiết kiệm nhiên liệu. Tuy nhiên có thể thấy giá cả vẫn là một yếu tố quan trọng trong việc quyết định chọn mua của khách hàng. Đây có thể là một yếu tố quan trọng để các hãng xe xác định được tệp khách mình mong muốn hướng đến để tối ưu trong cả công suất động cơ lẫn chi phí nhằm thu hút được nhiều khách hàng tiềm năng hơn.

4.3.7 Phân tích sự thay đổi của động cơ của xe theo thời gian

Sau quá trình phân tích để tìm ra sự thay đổi dẫn đến sự ưa chuộng về các mẫu xe theo thời gian, nhóm nhận thấy rằng có những thay đổi liên quan đến điều chỉnh động cơ xe. Điều này đặc biệt này được nhóm làm rõ ở phân tích trước, có thể thấy được sự đáp ứng linh hoạt của hai đại diện hàng đầu trong ngành công nghiệp ô tô - Ford và Mercedes-Benz - đối với nhu cầu ngày càng đa dạng của khách hàng.

Qua đó, nhóm quyết định phân tích trên tổng thể dữ liệu, bằng cách phân tích toàn diện và sâu sắc, nhóm hy vọng rằng sẽ có thể đưa ra những hiểu biết chi tiết về cách mà điều chỉnh động cơ đã làm thay đổi bức tranh tổng thể của ngành công nghiệp ô tô. Điều này không chỉ giúp các nhà nghiên cứu hiểu rõ hơn về chiến lược cụ thể của các hãng xe mà còn hỗ trợ người tiêu dùng trong quá trình lựa chọn xe, đồng thời thách thức và kích thích sự đổi mới trong ngành.





Nhận xét:

Thông qua biểu đồ có thể thấy được đã có sự thay đổi theo thời gian ở các chỉ số động cơ của xe trước và sau năm 2000.

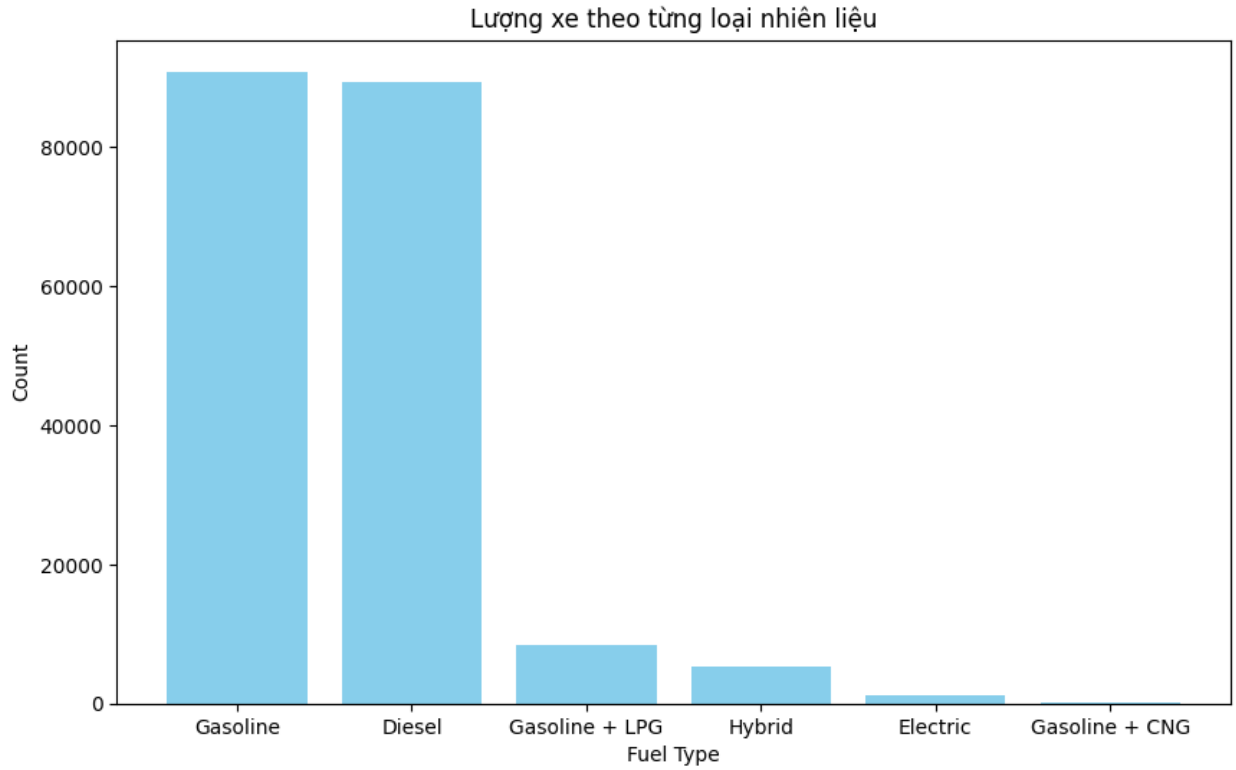
Đầu tiên, khi xem xét về công suất động cơ, điều đáng chú ý là xu hướng tăng dần của công suất, đặc biệt là sau năm 2000. Xe được sản xuất trong thời kỳ này thường có mức công suất trung bình cao hơn khoảng 20 mã lực so với các mô hình trước đó, có thể thấy được sự tiến bộ và đổi mới trong công nghệ động cơ.

Ngoài ra, trong việc đánh giá dung tích xi lanh, chúng ta thấy một hướng đi ngược lại. Cụ thể, có sự giảm tỷ lệ dung tích xi lanh trên các mô hình xe được sản xuất sau năm 2000. Điều này có thể cho thấy sự chú trọng vào hiệu suất năng lượng và khả năng tối ưu hóa trong quản lý nhiên liệu, thậm chí có thể kết hợp với xu hướng tăng công suất để tạo ra các động cơ mạnh mẽ, nhưng vẫn hiệu quả và thân thiện với môi trường.

4.4 Kiểm định

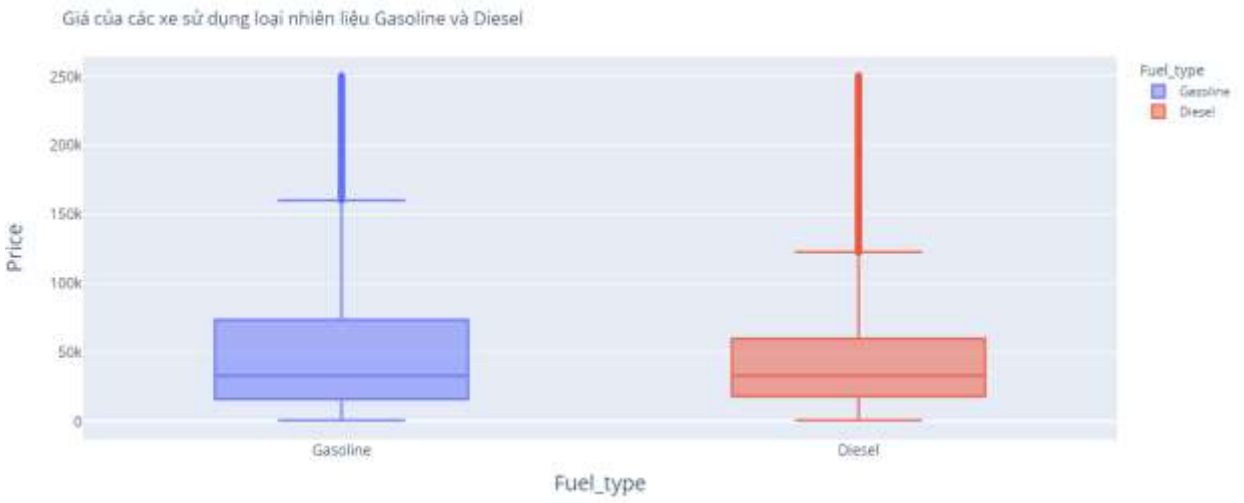
4.4.1 Giá cả của xe

Giả thuyết 1: Không có sự khác biệt giữa giá của những xe sử dụng loại nhiên liệu Gasoline và Diesel.



Qua phân tích biểu đồ, dễ thấy rằng Gasoline và Diesel là hai loại nhiên liệu nổi bật nhất. Sự phổ biến của cả hai loại này thể hiện sự ưa chuộng đặc biệt từ phía người tiêu dùng đối với các xe sử dụng 2 loại nhiên liệu Gasoline và Diesel. Nhóm quyết định đi phân tích sâu hơn về giá cả của những xe ở dụng 2 loại nhiên liệu này.

Mối tương quan giữa loại nhiên liệu Gasoline và Diesel đối với Giá xe:



Phân tích chi tiết hơn về các xe sử dụng Gasoline và Diesel tiết lộ một xu hướng đáng chú ý: giá của các xe sử dụng nhiên liệu Gasoline thường cao hơn so với xe sử dụng Diesel. Tuy nhiên, để chắc chắn rằng sự chênh lệch giá này có ý nghĩa thống kê, chúng ta cần kiểm định giả thuyết với mức độ tin cậy là 95%.

Kiểm định giả thuyết 1: Không có sự khác biệt giữa giá của những xe sử dụng loại nhiên liệu Gasoline và Diesel

H0: μ{Price}[Gasoline] = μ{Price}[Diesel]

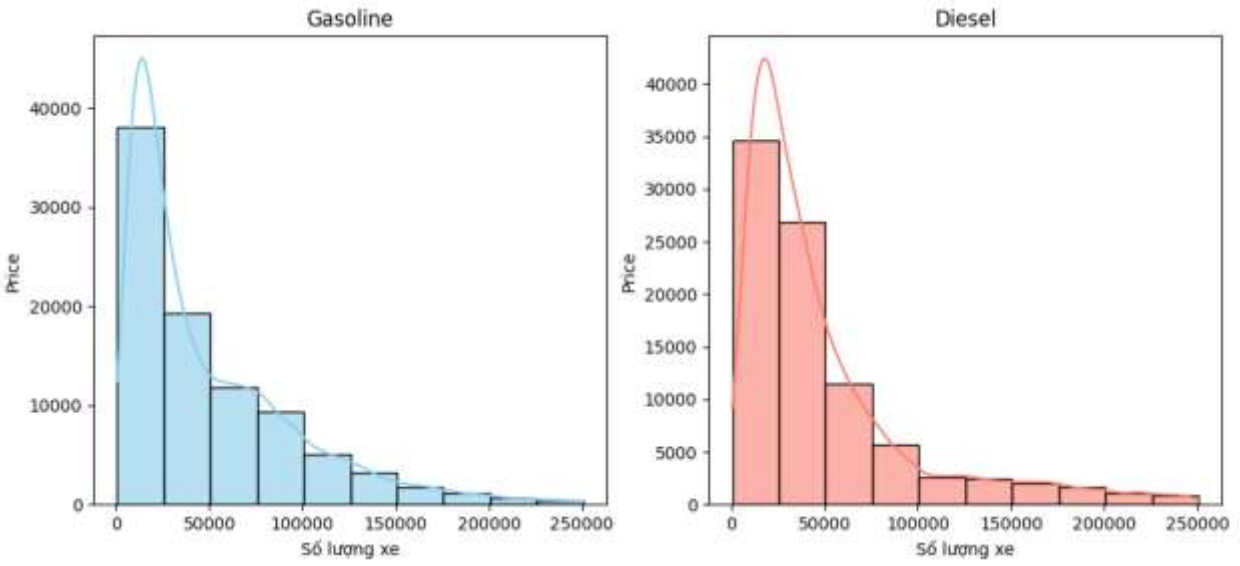
H1: μ{Price}[Gasoline] ≠ μ{Price}[Diesel]

```
price_gasoline = df[df['Fuel_type'] == 'Gasoline']['Price']

price_diesel = df[df['Fuel_type'] == 'Diesel']['Price']

# Tính toán phương sai

variance_gasoline = price_gasoline.var()

variance_diesel = price_diesel.var()

n_gasoline = len(price_gasoline)

n_diesel = len(price_diesel)

print('Phương sai của giá xe sử dụng gasoline:', variance_gasoline)

print('Phương sai của giá xe sử dụng diesel:', variance_diesel)

# Tính toán giá trị Z và p-value

z_statistic = (price_gasoline.mean() - price_diesel.mean()) /
((variance_gasoline/n_gasoline + variance_diesel/n_diesel)**0.5)

p_value = 2 * (1 - stats.norm.cdf(abs(z_statistic)))

print('Z_statistic:', z_statistic)

print('Giá trị p:', p_value)

alpha = 0.05

confidence_level = 1 - alpha

if (p_value < alpha):

    print(f'Trị số p = {p_value:.4f} < {alpha}',

          'nên bác bỏ H0.\n=> μ[Gasoline] != μ[Diesel].\n=> Có sự khác
biệt về giá giữa những xe sử dụng loại nhiên liệu Gasoline và
Diesel.')
```

```
else:
    print(f'Trị số p = {p_value:.4f} >= {alpha}',
          'nên chấp nhận H0.\n=> μ[Gasoline] == μ[Diesel].\n=> Không có sự khác biệt về giá giữa những xe sử dụng loại nhiên liệu Gasoline và Diesel.')
```

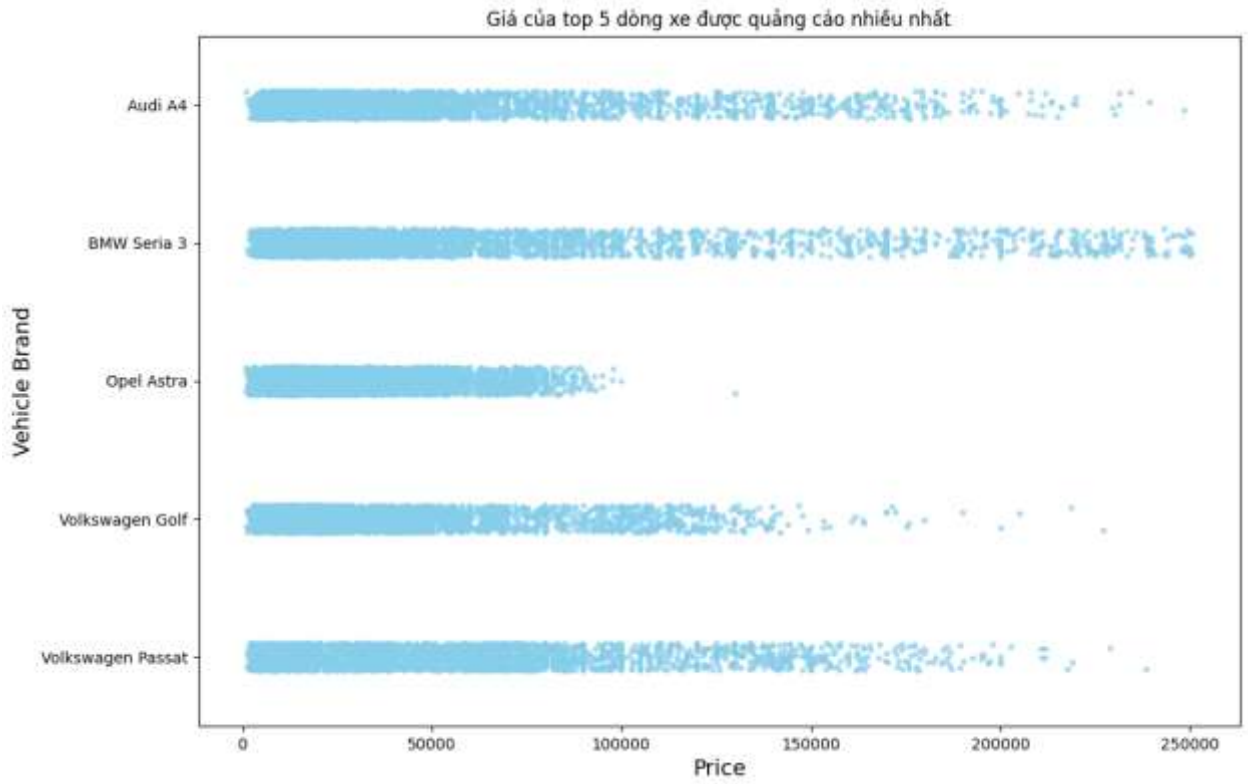
Output:

```
Phương sai của giá xe sử dụng gasoline: 2134528795.9964607
Phương sai của giá xe sử dụng diesel: 2320340850.4154143
Z_statistic: 2.7359048632142753
Giá trị p: 0.006220900837485832
Trị số p = 0.0062 < 0.05 nên bác bỏ H0.
=> μ[Gasoline] != μ[Diesel].
=> Có sự khác biệt về giá giữa những xe sử dụng loại nhiên liệu Gasoline và Diesel.
```



Nhận xét: Thông qua Kiểm định về giả thuyết, chúng ta đi đến kết luận rằng có sự khác biệt về giá của những xe sử dụng nhiên liệu là Gasoline và những xe sử dụng nhiên liệu là Diesel. Thông qua biểu đồ có thể thấy có đến 25000 xe có giá trung bình xấp xỉ 40000, trong khi đó giá xe trung bình của những xe sử dụng nhiên liệu là Diesel xấp xỉ từ 35000 trở xuống. Sự chênh lệch này là minh chứng cho việc giá cả thị trường của các loại xe có sự biến động phụ thuộc vào loại nhiên liệu mà xe sử dụng.

Giả thuyết 2: Không có sự khác biệt về giá xe trung bình của top 5 dòng xe được quảng cáo nhiều nhất.



Qua quan sát biểu đồ, chúng ta nhận thấy top 5 dòng xe được quảng cáo nhiều nhất thuộc các hãng Audi, BMW, Opel và Volkswagen. Trong số này, giá trung bình của các dòng xe như Audi A4, BMW Seria 3, và Volkswagen Passat cao hơn so với hai dòng xe còn lại.

Tuy đã có nhận định về sự chênh lệch giá, nhóm quyết định kiểm định giả thuyết với mức độ tin cậy là 95%, nhằm xác định tính chắc chắn của những khác biệt này. Điều này giúp đảm bảo rằng các quyết định chiến lược có thể dựa trên cơ sở số liệu thống kê đầy đủ và đáng tin cậy.

Trước tiên, nhóm quyết định tiến hành kiểm định Levene để kiểm tra sự đồng nhất phương sai .

Kiểm định giả thuyết : Không có sự khác biệt đáng kể về phương sai giữa các nhóm.

H0: Không có sự khác biệt đáng kể về phương sai giữa các nhóm.

H1: Có sự khác biệt đáng kể về phương sai giữa các nhóm.

```
from scipy.stats import levene

top_5_brands = df['Vehicle'].value_counts()[:5].index

df_top_5 = df[df['Vehicle'].isin(top_5_brands)]

# Tạo danh sách chứa dữ liệu giá xe cho từng dòng xe

price_data = [df_top_5[df_top_5['Vehicle'] == brand]['Price'] for
brand in top_5_brands]
```

```
# Tiến hành kiểm định Levene

statistic, p_value = levene(*price_data)

print(f"Levene Statistic: {statistic}")

print(f"P-Value (Levene): {p_value}")

alpha = 0.05

if p_value < alpha:

    print("Có đủ bằng chứng để bác bỏ giả thuyết H0.\n=> Có sự khác biệt về phương sai giữa các nhóm.")

else:

    print("Không bác bỏ giả thuyết H0.\n=> Không có sự khác biệt đáng kể về phương sai giữa các nhóm.")
```

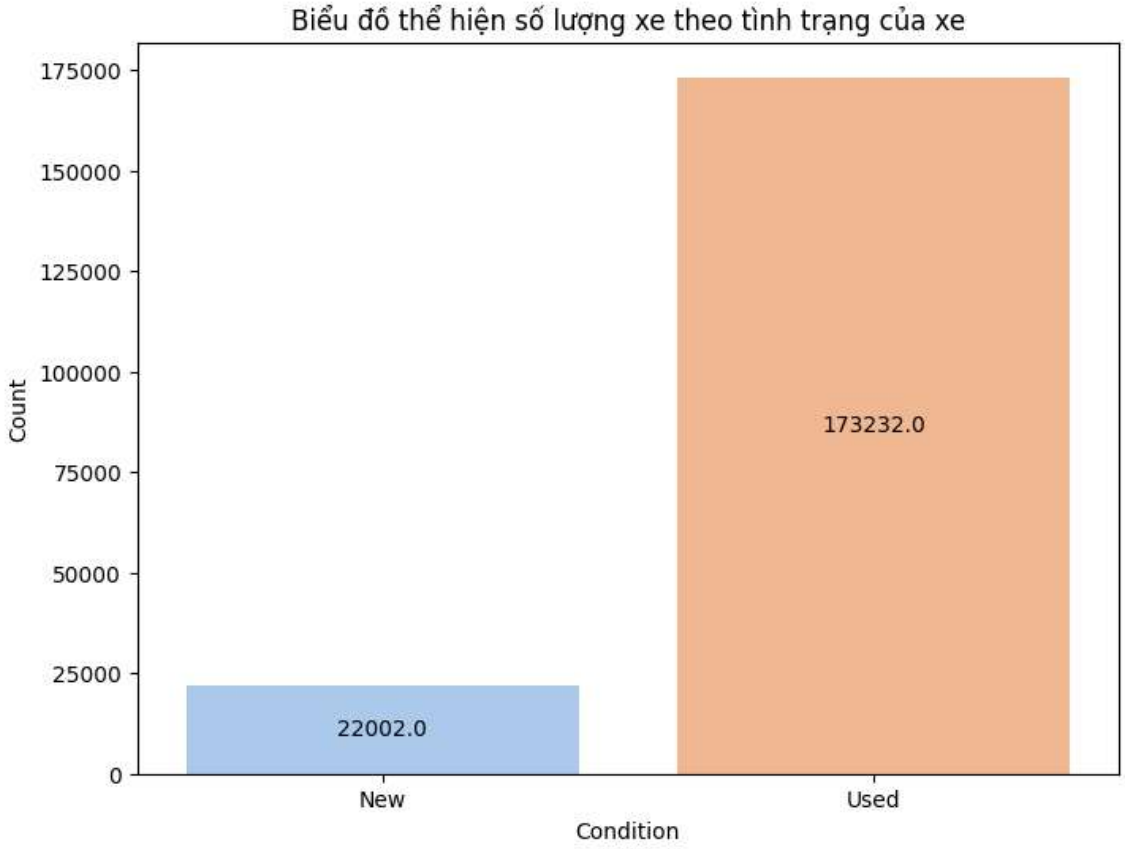
Output:

```
Levene Statistic: 257.88862884744304

P-Value (Levene): 4.6163416573339865e-217

Có đủ bằng chứng để bác bỏ giả thuyết H0.

=> Có sự khác biệt về phương sai giữa các nhóm.
```

Nhận xét: Kiểm định Levene cho thấy có bằng chứng để bác bỏ giả thuyết có sự đồng nhất phương sai giữa các nhóm. Vậy nên, chúng ta không thể dùng kiểm định Anova ở đây vì nó sẽ không mang lại kết quả chính xác.

4.4.2 Động cơ và nhiên liệu xe

Giả thuyết 3: Xe có chỉ số Mileage_km trên 4884 thì có khả năng là xe cũ hơn so với các xe còn lại.



Thông qua các bước phân tích ở trước và biểu đồ cột ở trên, nhóm nhận thấy được rằng dữ liệu đề cập phần lớn đến các xe ô tô đã qua sử dụng. Tuy nhiên vẫn có một số lượng xe được xem là xe mới, vậy có sự khác biệt như thế nào trong quãng đường các xe đã đi để xác định xem tình trạng của xe là xe cũ hay là mới. Để đi đến được nhận định, nhóm quyết định thực hiện kiểm định giả thuyết để xem xét rằng có sự khác biệt về số quãng đường mà của các xe ở 2 tình trạng này hay không.

Nhóm thực hiện tính quãng đường di chuyển trung bình của tất cả các xe mới và thu được kết quả rằng các xe mới có số quãng đường di chuyển trung bình là 4884 km và quyết định chọn đây là ngưỡng để phân biệt 2 tình trạng này. Tiếp theo nhóm thực hiện kiểm định liệu các xe có chỉ số Mileage_km trên 4884 thì có khả năng là xe cũ hơn so với các xe còn lại không.

Kiểm định giả thuyết 3: Xe có chỉ số Mileage_km trên 4884 thì có khả năng là xe cũ hơn so với các xe còn lại.

H0: μ{Mileage_km > 4884}[Condition] = μ{Mileage_km <= 4884}[Condition]

H1: μ{Mileage_km > 4884}[Condition] ≠ μ{Mileage_km <= 4884}[Condition]

```
# Lọc ra những chiếc xe có tình trạng là cũ

used_cars = df[df['Condition'] == 'New']

# Tính số km trung bình của những chiếc xe có tình trạng là cũ

average_mileage_used_cars = used_cars['Mileage_km'].mean()

# In số km trung bình

print(f"Số km trung bình của những xe cũ là:
{average_mileage_used_cars} km")
```

Output:

```
Số km trung bình của những xe cũ là: 4884.459640032725 km
```

```
df_condition = df.copy()

df_condition['Condition'] = df_condition['Condition'].replace({

    'New' : 1,

    'Used' : 0

})

condition_counts = df_condition['Condition'].value_counts()

print(condition_counts)
```

Output:

```
Used    173224

New      22002
```

```
import scipy.stats as stats

from statsmodels.stats.weightstats import ztest

# Tạo 2 series chứa thông tin về tình trạng các xe dựa trên ngưỡng
quãng đường đã đi là 4884 km

group_over = df_condition[df_condition['Mileage_km'] >
```

```
4884]['Condition']

group_under = df_condition[df_condition['Mileage_km'] <=
4884]['Condition']

# Kiểm định z test

z_statistic, p_value = ztest(group_over, group_under,
alternative='two-sided')

print(f"Z-Statistic: {z_statistic}")

print(f"Giá trị p: {p_value}")

# Kiểm tra giả thuyết

alpha = 0.05

if p_value < alpha:

    print(f"Trị số p = {p_value} < {alpha}",

          'có đủ bằng chứng để bác bỏ H0\n=> Có sự chênh lệch giữa
quãng đường di chuyển của 2 nhóm xe này')

else:

    print(f"Trị số p = {p_value} >= {alpha}",

          'không đủ bằng chứng để bác bỏ H0\n=> Có sự chênh lệch giữa
quãng đường di chuyển của 2 nhóm xe này')
```

Output:

```
Z-Statistic: -943.9764916752024

Giá trị p: 0.0

Trị số p = 0.0 < 0.05 có đủ bằng chứng để bác bỏ H0

=> Có sự chênh lệch giữa quãng đường di chuyển của 2 nhóm xe này
```

Nhận xét: Dựa vào các thông số từ kết quả sau khi kiểm định với p = 0.0 nhỏ hơn mức ý nghĩa alpha = 0.05, ta có đủ bằng chứng để bác bỏ giả thuyết H0. Điều này có nghĩa là có sự chênh lệch giữa quãng đường di chuyển của 2 nhóm xe này. Ngoài ra, với giá trị thống kê z là -943.9988, đây là một con số rất lớn và khẳng định được sự chênh lệch đáng kể giữa các nhóm.

Giả thuyết 4: Không có sự liên quan giữa loại nhiên liệu và loại hộp số của xe

Sau khi tiến hành phân tích kiểm định những yếu tố ảnh hưởng đến giá xe, trong phần tiếp theo, nhóm muốn phân tích kỹ hơn về các yếu tố ảnh hưởng lẫn nhau động hệ thống động cơ liệu, liệu các thành phần này có sự tác động, ảnh hưởng đến nhau.

Đầu tiên nhóm muốn tiến hành kiểm định về sự ảnh hưởng giữa nhiên liệu và loại hộp số của xe. Tương tự như ở phần trước nhóm chọn ra 2 loại nhiên liệu phổ biến nhất là Gasoline và Diesel để tiến hành kiểm định. Bằng cách này, nhóm mong muốn xác định được sự tương quan và tác động giữa loại nhiên liệu là Gasoline và Diesel và loại hộp số nhằm hiểu rõ hơn về sự tương tác giữa các yếu tố này và ảnh hưởng đồng thời của chúng đến giá cả xe. Điều này sẽ cung cấp thông tin quan trọng cho quá trình ra quyết định phát triển chiến lược và phát triển sản phẩm

Kiểm định giả thuyết 4: Không có sự liên quan giữa loại nhiên liệu và loại hộp số của xe

H0: Loại nhiên liệu (Gasoline, Diesel) và loại hộp số của xe là Độc Lập.

H1: Loại nhiên liệu (Gasoline, Diesel) và loại hộp số của xe là Phụ Thuộc lẫn nhau.

```
filtered_df = df[df['Fuel_type'].isin(['Gasoline', 'Diesel'])]

fuel_type_counts = filtered_df['Fuel_type'].value_counts()

fuel_type_counts
```

Output:

```
Gasoline    90878
Diesel      89343
Name: Fuel_type, dtype: int64
```

```
df3 = filtered_df[['Fuel_type', 'Transmission']]

crosstab = pd.crosstab(df3['Fuel_type'], df3['Transmission'])

crosstab
```

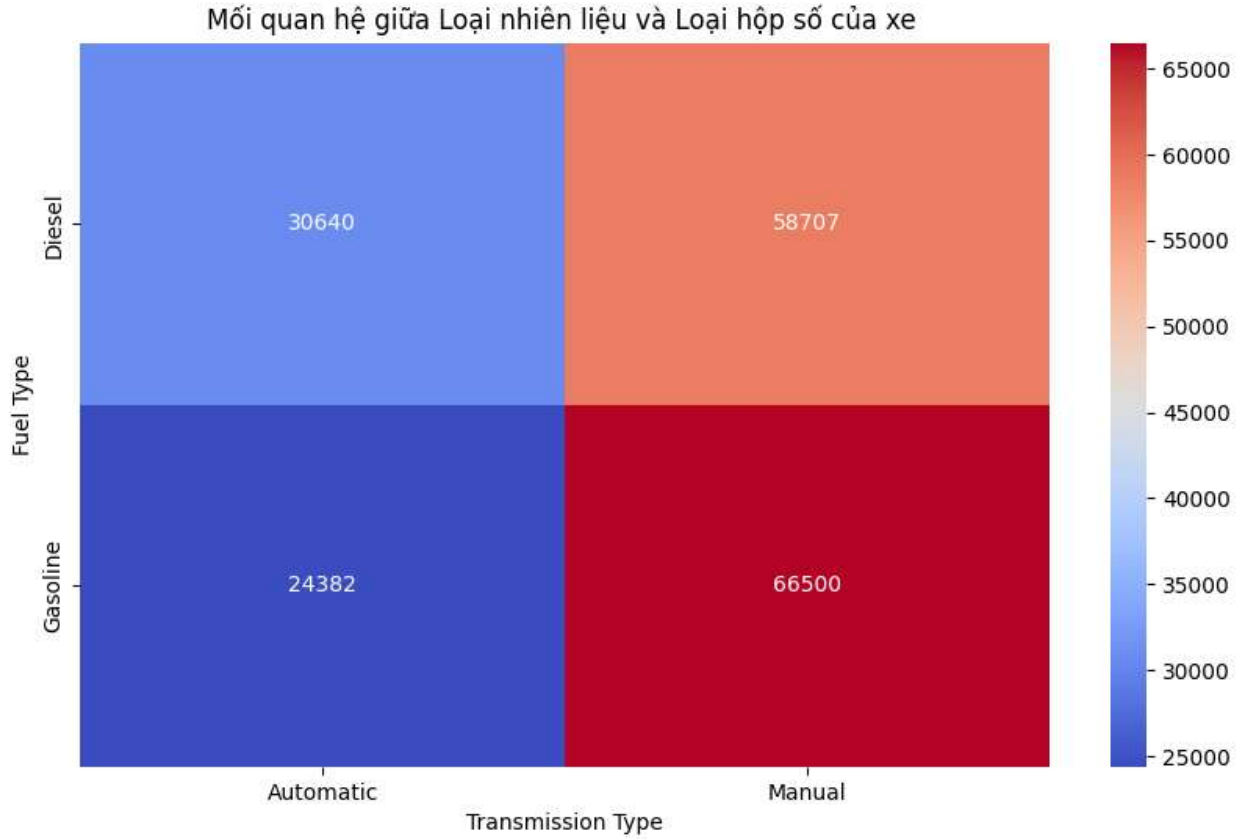
Output:

Transmission	Automatic	Manual
Fuel_type		
Diesel	30639	58704
Gasoline	24381	66497

```
contingency_table = pd.crosstab(df3['Fuel_type'], df3['Transmission'])

# Kiểm định chi-squared

chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

print(f'Chi-squared statistic: {chi2_stat}')

print(f'Giá trị p: {p_value}')

# Kiểm tra giả thuyết

alpha = 0.05

if p_value < alpha:

    print(f'Trị số p = {p_value} < {alpha}',

     'nên bác bỏ H0\n => Loại nhiên liệu (Gasoline, Diesel) và loại
hộp số của xe là Phụ Thuộc lẫn nhau')

else:

    print(f'Trị số p = {p_value} >= {alpha}',

      'nên không bác bỏ H0\n => Loại nhiên liệu (Gasoline, Diesel) và
loại hộp số của xe là Độc Lập')
```

Output:

```
Chi-squared statistic: 1183.514419198795

Giá trị p: 2.333975795723498e-259

Trị số p = 2.333975795723498e-259 < 0.05 nên bác bỏ H0

 => Loại nhiên liệu (Gasoline, Diesel) và loại hộp số của xe là Phụ
Thuộc lẫn nhau
```



Nhận xét:

Thông qua kiểm định giả thuyết, chứng minh được rằng các loại nhiên liệu (Gasoline, Diesel) và loại hộp số của xe có sự phụ thuộc lẫn nhau. Kết quả này là một bước quan trọng để hiểu rõ hơn về mối quan hệ giữa các thành phần trong động cơ của xe.

Ngoài ra, nhận định này không chỉ giúp chúng ta hiểu rõ hơn về các yếu tố kỹ thuật trong thị trường ô tô, mà còn là cơ hội để tối ưu hóa chiến lược và phát triển sản phẩm trong ngành công nghiệp này. Đồng thời, nó cũng làm nổi bật sự phức tạp và tương tác đa chiều của các yếu tố ảnh hưởng đến quyết định mua xe của người tiêu dùng.

Giả thuyết 5: Không có sự liên quan giữa kiểu dáng của xe và loại nhiên liệu mà xe sử dụng.

Vì mục đích sử dụng của mỗi loại xe có những nét tương đồng, nên nhóm quyết định gom chúng lại thành 3 nhóm xe chính bao gồm: Nhóm xe có kích thước nhỏ gọn và tiết kiệm nhiên liệu, xe cá nhân & thể thao, và cuối cùng là xe gia đình. Tương tự với các loại nhiên liệu được đề cập đến trong bộ dữ liệu, nhóm cũng quyết định phân loại chúng vào các nhóm bao gồm: Nhiên liệu truyền thống, Nhiên liệu bán truyền thống và Nhiên liệu mới.

Sau khi tiến hành phân tích trong phần phân tích đa biến, nhóm thấy được có sự khác biệt nhẹ về việc sử dụng các loại nhiên liệu theo từng nhóm xe, nên nhóm quyết định thực hiện kiểm định để xét liệu có quan hệ giữa kiểu dáng thân và loại nhiên liệu mà xe sử dụng.

Kiểm định giả thuyết 5: Không có sự liên quan giữa kiểu dáng của xe và loại nhiên liệu mà xe sử dụng.

H0: Kiểu dáng của xe và loại nhiên liệu mà xe sử dụng là Độc Lập.

H1: Kiểu dáng của xe và loại nhiên liệu mà xe sử dụng là Phụ Thuộc lẫn nhau.

```
df_type = df.copy()

df_type['Type'] = df_type['Type'].replace({

  'station_wagon' : 'Family Cars',

  'SUV' : 'Family Cars',

  'minivan' : 'Family Cars',

  'compact' : 'Compact and Fuel-Efficient Cars',

  'city_cars' : 'Compact and Fuel-Efficient Cars',

  'small_cars' : 'Compact and Fuel-Efficient Cars',

  'sedan' : 'Sporty and Personal Cars',

  'coupe' : 'Sporty and Personal Cars',

  'convertible' : 'Sporty and Personal Cars'

})
```

```
df_fuel = df.copy()

df_fuel['Fuel_type'] = df_fuel['Fuel_type'].replace({

    'Gasoline' : 'Traditional Fuels',

    'Diesel' : 'Traditional Fuels',

    'Gasoline + LPG' : 'Mixed or Convention Fuels',

    'Gasoline + CNG' : 'Mixed or Convention Fuels',

    'Hybrid' : 'New and Advanced Technologies',

    'Electric' : 'New and Advanced Technologies'

})
```

```
crosstab_table = pd.crosstab(df_type['Type'], df_fuel['Fuel_type'])

crosstab_table
```

Output:

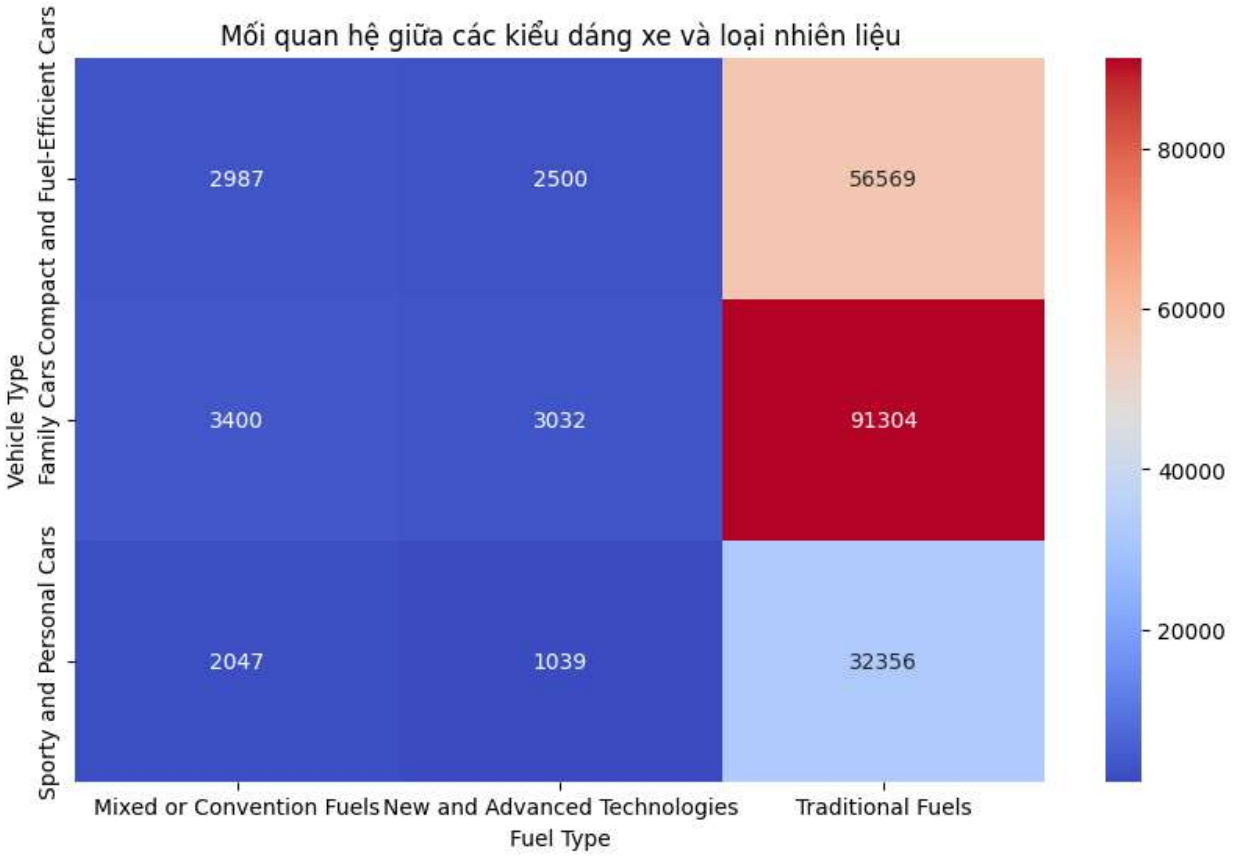
Fuel_type Type	Mixed or Convention Fuels	New and Advanced Technologies	Traditional Fuels
Compact and Fuel-Efficient Cars	2987	2500	56569
Family Cars	3400	3032	91296
Sporty and Personal Cars	2047	1039	32356

```
contingency_table = pd.crosstab(df_type['Type'], df_fuel['Fuel_type'])

# Kiểm định chi bình phương

chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

print(f'Chi-squared statistic: {chi2_stat}')

print(f'Giá trị p: {p_value}')

alpha = 0.05

if p_value < alpha:

  print(f'Trị số p = {p_value} < {alpha}',

     'nên bác bỏ H0\n => Kiểu dáng của xe và loại nhiên liệu mà xe sử dụng là Phụ Thuộc lẫn nhau')

else:

  print(f'Trị số p = {p_value} >= {alpha}',

    'nên không bác bỏ H0\n => Kiểu dáng của xe và loại nhiên liệu mà xe sử dụng là Độc Lập')
```

Output:

```
Chi-squared statistic: 515.6681805272433

Giá trị p: 2.735891230490413e-110

Trị số p = 2.735891230490413e-110 < 0.05 nên bác bỏ H0
```

```
=> Kiểu dáng của xe và loại nhiên liệu mà xe sử dụng là Phụ Thuộc lẫn
nhau.
```



Nhận xét:

Qua quá trình kiểm định và phân tích dữ liệu, nhóm đã thu được kết quả p-value đặc biệt thấp, thấp hơn cả mức alpha quy định (0.05). Điều này ngụ ý rằng kiểu dáng thân xe và loại nhiên liệu mà xe sử dụng có sự phụ thuộc lẫn nhau. Qua đây, ta có cơ sở để kết luận rằng sự lựa chọn về kiểu dáng thân xe của một chiếc xe cụ thể có ảnh hưởng đáng kể đến quyết định về loại nhiên liệu mà nó sử dụng.

Thông qua việc trực quan bằng biểu đồ đã mô tả rõ sự phân bố của các nhóm xe và loại nhiên liệu tương ứng. Đa số các nhóm xe được quan sát sử dụng loại nhiên liệu truyền thống như xăng và dầu, điều này là điều dễ hiểu trong bối cảnh hiện nay. Tuy nhiên, điểm đáng chú ý là có một số ít xe trong các nhóm, đặc biệt là nhóm xe nhỏ gọn và nhóm xe gia đình, đã chọn lựa sử dụng các loại nhiên liệu mới và thân thiện với môi trường.

Nhìn chung, xu hướng này có thể được coi là một tín hiệu tích cực về sự chuyển đổi từ các nhiên liệu truyền thống sang những lựa chọn có tác động tích cực đối với môi trường. Đặc biệt, sự lựa chọn của nhóm xe gia đình, thường xuyên có nhu cầu sử dụng xe lâu dài và di chuyển xa, là một dấu hiệu quan trọng về hướng đi mới trong ngành công nghiệp xe hơi.

Thông qua nhận định này, có thể thấy được cái nhìn sâu sắc về xu hướng sử dụng nhiên liệu của các loại xe khác nhau và đặt ra những câu hỏi hứa hẹn về tương lai của ngành công nghiệp xe hơi, đặc biệt là trong bối cảnh nhu cầu ngày càng cao về sự bền vững và bảo vệ môi trường.

Giả thuyết 6: Không có sự liên quan giữa loại hộp số và hệ thống dẫn động của xe.

Nhóm đã quan sát rằng các giá trị trong biến "Drive" (Hệ thống dẫn động) có sự khác biệt đáng kể trong mục đích sử dụng của xe nên nhóm đã quyết định phân loại các giá trị này thành 3 nhóm chính: "Front wheels" (Dẫn động trước), "Rear wheels" (Dẫn động sau), và "4x4" (Dẫn động 4 bánh). Mỗi nhóm được xác định để phục vụ cho các mục đích sử dụng khác nhau của người lái xe.

Thông qua tìm hiểu, nhóm nhận thấy được rằng hệ thống dẫn động và loại hộp số của một xe ô tô có sự liên quan mật thiết, và để đi đến kết luận chính xác cho nhận định này, nhóm đã quyết định thực hiện kiểm định với 2 biến Drive và Transmission được đề cập đến trong bộ dữ liệu. Bằng cách này, nhóm hy vọng sẽ có cái nhìn rõ ràng hơn về cách các yếu tố này tương tác và ảnh hưởng đến trải nghiệm lái xe.

Kiểm định giả thuyết 6: Loại hộp số và hệ thống dẫn động của xe là Độc Lập

H0: Loại hộp số và hệ thống dẫn động của xe là Độc Lập

H1: Loại hộp số và hệ thống dẫn động của xe là Phụ Thuộc lẫn nhau

```
df_drive = df.copy()

df_drive['Drive'] = df_drive['Drive'].replace({

    'Front wheels': 'Front wheels',

    'Rear wheels': 'Rear wheels',

    '4x4 (attached automatically)': '4x4',

    '4x4 (permanent)': '4x4',

    '4x4 (attached manually)': '4x4'

})
```

```
crosstab_table = pd.crosstab(df_drive['Drive'],
df_drive['Transmission'])

crosstab_table
```

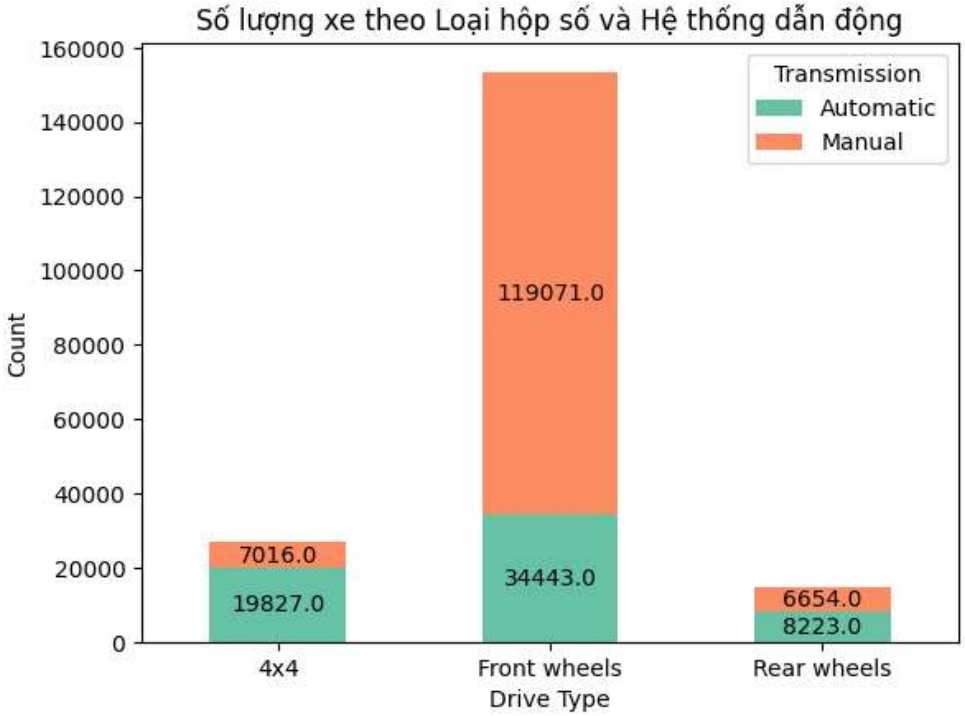
Output:

Transmission / Drive	Automatic	Manual
4x4	19826	7015
Front wheels	34442	119066
Rear wheels	8223	6654

```
contingency_table = pd.crosstab(df_drive['Drive'], df['Transmission'])

# Kiểm định chi-squared

chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

print(f'Chi-squared statistic: {chi2_stat}')

print(f'Giá trị p: {p_value}')

alpha = 0.05

if p_value < alpha:

    print(f'Trị số p = {p_value} < {alpha}',

     'nên bác bỏ H0\n => Loại hộp số và hệ thống dẫn động của xe là Phụ Thuộc lẫn nhau')

else:

    print(f'Trị số p = {p_value} >= {alpha}',

      'nên không bác bỏ H0\n => Loại hộp số và hệ thống dẫn động của xe là Độc Lập')
```

Output:

```
Chi-squared statistic: 31769.034585851827

Giá trị p: 0.0

Trị số p = 0.0 < 0.05 nên bác bỏ H0

 => Loại hộp số và hệ thống dẫn động của xe là Phụ Thuộc lẫn nhau
```



Nhận xét: Dựa vào kết quả kiểm định thống kê, nhóm thu được giá trị p-value là 0.0, mức ý nghĩa thấp hơn mức alpha quy định (0.05). Điều này chứng minh một cách rõ ràng rằng loại hộp số của xe và hệ thống dẫn động của xe có sự phụ thuộc lẫn nhau.

Tình hình này được minh họa rõ trong biểu đồ, nơi mà chúng ta có thể quan sát sự khác biệt đáng kể trong việc sử dụng các loại hộp số tùy thuộc vào nhóm hệ thống dẫn động khác nhau. Trong nhóm sử dụng hệ thống dẫn động trước (Front wheels), loại hộp số thủ công được ưa chuộng hơn, với số lượng xe sử dụng gần 3.5 lần so với loại hộp số tự động. Điều này có thể do người lái xe ở nhóm này đánh giá cao sự thuận tiện trong giao thông, nơi mà việc sử dụng hộp số thủ công có thể linh hoạt và dễ quản lý hơn.

Ngược lại, ở nhóm sử dụng hệ thống dẫn động 4x4, chúng ta thấy số lượng xe sử dụng hộp số tự động cao hơn đáng kể so với loại hộp số thủ công. Điều này có thể liên quan đến mục đích sử dụng xe, nơi mà người lái xe ưa chuộng hiệu suất và đa năng trên nhiều điều kiện đường đi khác nhau. Trong khi đó, ở nhóm sử dụng hệ thống dẫn động sau (Rear wheels), sự phân bố giữa hai loại hộp số này khá cân đối, với số lượng xe sử dụng tự động và thủ công xấp xỉ nhau.

Sự phụ thuộc giữa loại hộp số và hệ thống dẫn động có thể được hiểu dựa trên những yếu tố như sự thuận tiện, hiệu suất và mục đích sử dụng của người lái xe. Điều này có thể hỗ trợ các nhà sản xuất trong việc phát triển sản phẩm và chiến lược tiếp thị phù hợp với nhu cầu thị trường.

Giả thuyết 7: Không có sự liên quan giữa kiểu dáng của xe và các khu vực.

Sau khi kiểm tra sự phụ thuộc giữa kiểu dáng của xe và loại nhiên liệu mà xe sử dụng, nhóm đặt ra giả thuyết về việc có sự khác biệt trong các kiểu dáng xe ở các vùng khác nhau hay không. Để đánh giá và chứng minh hoặc phủ định giả thuyết này, nhóm quyết định thực hiện một kiểm định để xem liệu có sự phụ thuộc giữa kiểu dáng của xe và khu vực địa lý hay không.

Kiểm định giả thuyết 7 : Không có sự liên quan giữa kiểu dáng của xe và các khu vực.

H0: Kiểu dáng của xe và các khu vực là Độc Lập

H1: Kiểu dáng của xe và các khu vực là Phụ Thuộc lẫn nhau.

```
crosstab_table = pd.crosstab(df_type['Type'], df['Offer_region'])

crosstab_table
```

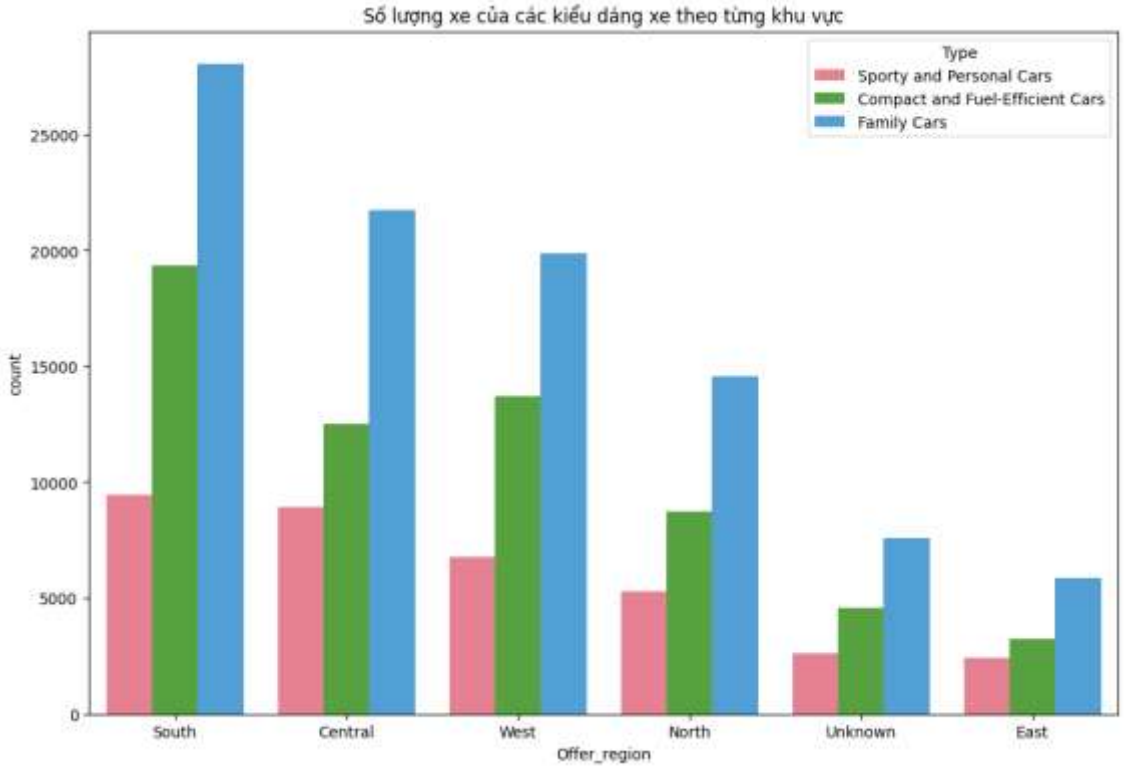
Output:

Offer_region Type	Central	East	North	South	West
Compact and Fuel-Efficient Cars	12676	3212	8731	23731	13706
Family Cars	21911	5899	14581	35456	19881
Sporty and Personal Cars	9017	2424	5287	11925	6789

```
contingency_table = pd.crosstab(df_type['Type'], df['Offer_region'])

# Kiểm định chi bình phương

chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

print(f'Chi-squared statistic: {chi2_stat}')

print(f'Giá trị p: {p_value}')

alpha = 0.05

if p_value < alpha:

  print(f'Trị số p = {p_value} < {alpha}',

     'nên bác bỏ H0\n => Kiểu dáng của xe và các khu vực là Phụ Thuộc
lẫn nhau')

else:

  print(f'Trị số p = {p_value} >= {alpha}',

    'nên không bác bỏ H0\n => Kiểu dáng của xe và các khu vực là Độc
Lập')
```

Output:

```
Chi-squared statistic: 622.9650645158116
Giá trị p: 2.6989471975185646e-129
Trị số p = 2.6989471975185646e-129 < 0.05 nên bác bỏ H0
 => Kiểu dáng của xe và các khu vực là Phụ Thuộc lẫn nhau
```



Nhận xét: Thông qua kiểm định, có thể thấy được chỉ số p-value rất nhỏ so với mức alpha quy định (0.05), từ đây có thể rút ra nhận định rằng kiểu dáng của xe và các khu vực có sự phụ thuộc lẫn nhau.

Từ góc nhìn trực quan, biểu đồ thể hiện rằng South và Central là hai khu vực tập trung nhiều quảng cáo xe, đặc biệt là trong phân khúc xe gia đình. Các dòng xe nhỏ gọn và tiết kiệm nhiên liệu cũng thu hút sự quan tâm đáng kể. Trái ngược, nhóm xe thể thao và cá nhân chỉ chiếm một phần nhỏ thị phần, chủ yếu là đối tượng khách hàng trẻ và chưa có gia đình.

Có thể nhận định rằng người tiêu dùng vẫn đánh giá cao các mẫu xe gia đình, có kích thước lớn và sự thoải mái, phù hợp cho cả đi lại hàng ngày và những chuyến đi xa. Ngoài ra, có một lượng lớn khách hàng quan tâm đến xe tiết kiệm nhiên liệu thể hiện xu hướng tăng cường ý thức về tiết kiệm nhiên liệu và bảo vệ môi trường đồng thời tối ưu được chi phí. Còn đối với các mô hình xe thể thao và cá nhân, chúng chỉ thu hút một tầng khách hàng hẹp.

CHƯƠNG 5: XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH

5.1 Mục đích xây dựng mô hình

Dựa trên các thông tin về các mẫu xe, việc xây dựng mô hình hồi quy để dự đoán giá xe có thể cung cấp các thông tin hữu ích cho người mua và người bán khi đưa ra quyết định. Người mua có thể xem xét để đảm bảo rằng giá cả đề xuất là hợp lý, trong khi người bán có thể sử dụng mô hình hồi quy để xác định giá bán phù hợp với thị trường. Ngoài ra, mô hình hồi quy tốt sẽ có thể được sử dụng để dự đoán giá của các xe mới xuất hiện trên thị trường, giúp cập nhật giá cả và đáp ứng nhanh chóng với biến động của thị trường ô tô.

5.2 Huấn luyện mô hình

- Tạo một bản copy của dataframe sau lúc tiền xử lý, việc sử dụng bản nháp này để tránh các ảnh hưởng của thuộc tính trong quá trình thao tác thuộc tính (thêm bớt cột,...).
- Chọn ra các thuộc tính định tính có số lượng giá trị duy nhất bé hơn 10, việc chọn các thuộc tính định danh có quá nhiều giá trị unique sẽ gây ảnh hưởng đến bộ nhớ và tài nguyên của máy:
  Các thuộc tính định lượng được chọn lọc:

  ['Price', 'Production_year', 'Mileage_km', 'Power_HP', 'Displacement_cm3']

  Các thuộc tính định tính được chọn lọc:

  ['Condition', 'Fuel_type', 'Drive', 'Transmission', 'Type', 'Doors_number']

- Tái gán các cột đã chọn vào DataFrame và áp dụng dummy encoding
- Loại bỏ biến target ra khỏi danh sách các biến được chọn
- Đối với mô hình Linear Regression và Support Vector Regression: Là những mô hình rất nhạy cảm với các giá trị có miền phân bố lớn nên việc scale theo MinMaxScaler (trả về khoảng giá trị 0-1) để thu hẹp khoảng cách của dữ liệu là cần thiết. Riêng với mô hình Random Forest thì việc scale dữ liệu là không cần thiết.
- Chia ra ba tập train, test và validation, tập train/temp được chia theo tỉ lệ 70/30 và tiếp tục chia tập test và validation theo tỉ lệ 50/50 và đưa vào huấn luyện.

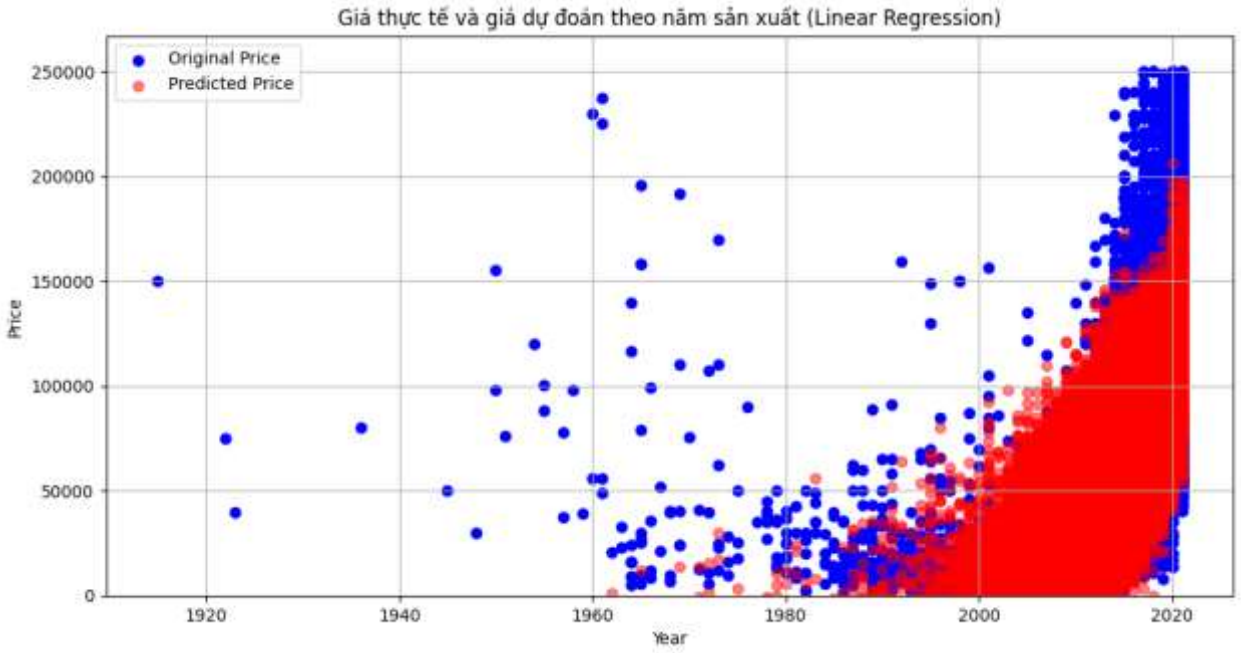
5.3 Đánh giá mô hình

5.3.1 Các chỉ số đánh giá

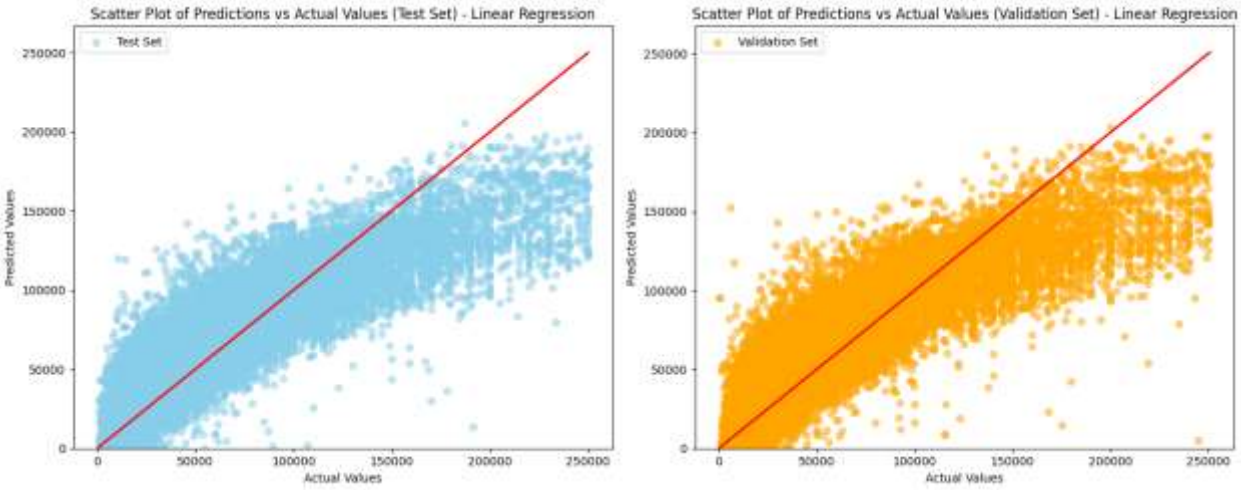
Chỉ số đánh giá	Hồi quy tuyến tính	Support Vector Regressor	Random Forest
Mean Absolute Error (MAE)	**15.958,72** Đây là giá trị trung bình của sai số tuyệt đối, cho thấy mô hình này dự đoán lệch gần 16 nghìn đơn vị PLN so với giá trị thực tế.	**17.771,93** SVR có giá trị trung bình của sai số tuyệt đối là lớn nhất trong cả ba mô hình, điều này có nghĩa các dự đoán lệch đến khoảng 17 nghìn PLN so với giá trị thực tế.	**7.283,217** Đây là MAE thấp nhất trong ba mô hình, cho thấy Random Forest có xu hướng dự đoán chính xác hơn, giá trị dự đoán có sai số tuyệt đối chỉ khoảng 7.000 PLN
Root Mean Squared Error (RMSE)	**23.180,41** RMSE lớn cũng chỉ ra rằng mô hình có những dự đoán sai lệch khá xa so với giá trị thực.	**31.663,07** RMSE cao nhất trong số ba mô hình, báo hiệu rằng các dự đoán của SVR có các mức sai số lớn nhất trong các mô hình được chọn	**13.173,666** RMSE thấp nhất, phản ánh sự chính xác cao nhất trong việc dự đoán của mô hình Random Forest so với hai mô hình còn lại.
R2	**0,7706** Hệ số này cho biết khoảng 77,06% sự biến thiên của định giá có thể được giải thích bởi mô hình hồi quy.	**0,572** Từ những sai số lớn, mô hình SVR chỉ dự đoán đúng khoảng 57,2 %, hoặc chỉ có 57,2% sự biến thiên của định giá có thể được giải thích bởi mô hình SVR	**0,926** R2 cao nhất, cho thấy mô hình Random Forest giải thích được 92,58% sự biến thiên, là dấu hiệu của một mô hình dự đoán tốt.

5.3.2 Biểu diễn trực quan các giá trị dự đoán và thực tế

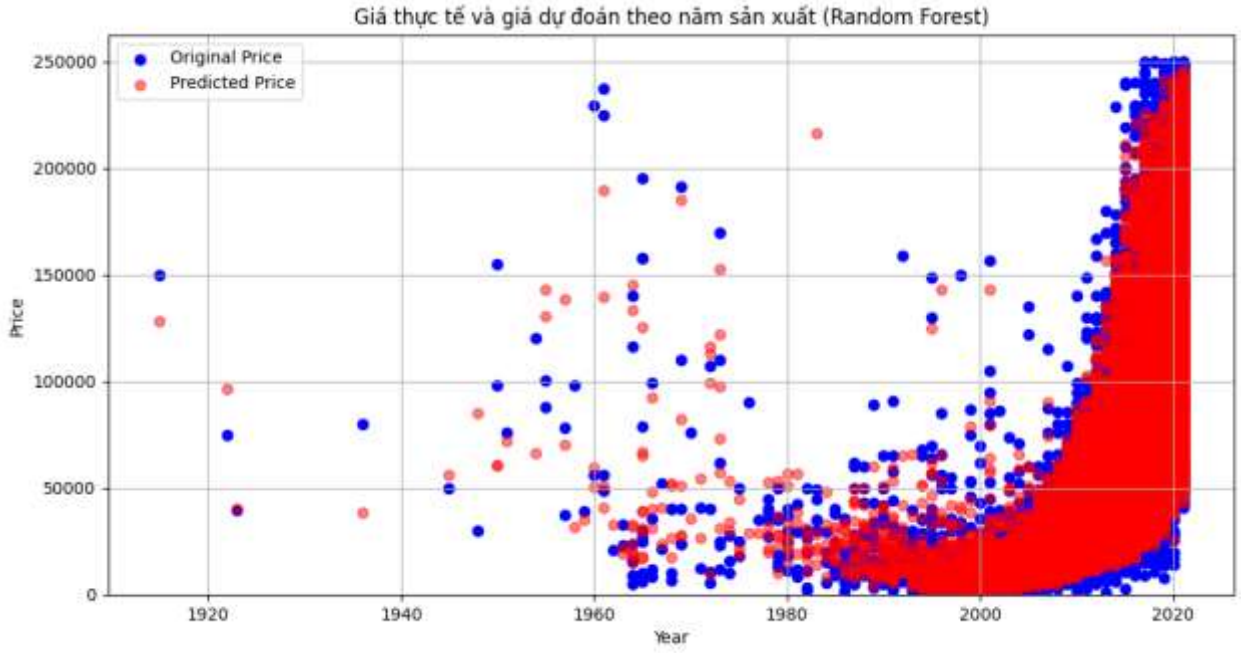
a) **Hồi quy tuyến tính (Linear Regression):** Mô hình Hồi quy tuyến tính dự đoán các giá trị định giá theo năm sản xuất sản bị sai lệch tương đối, Tuy nhiên, các giá trị dự đoán trên mức 0 cho thấy đây là một mô hình tương đối tốt khi các mức giá từ 0 - 200.000 PLN được dự đoán tương đối chính xác và có sai lệch không quá lớn so với thực tế.



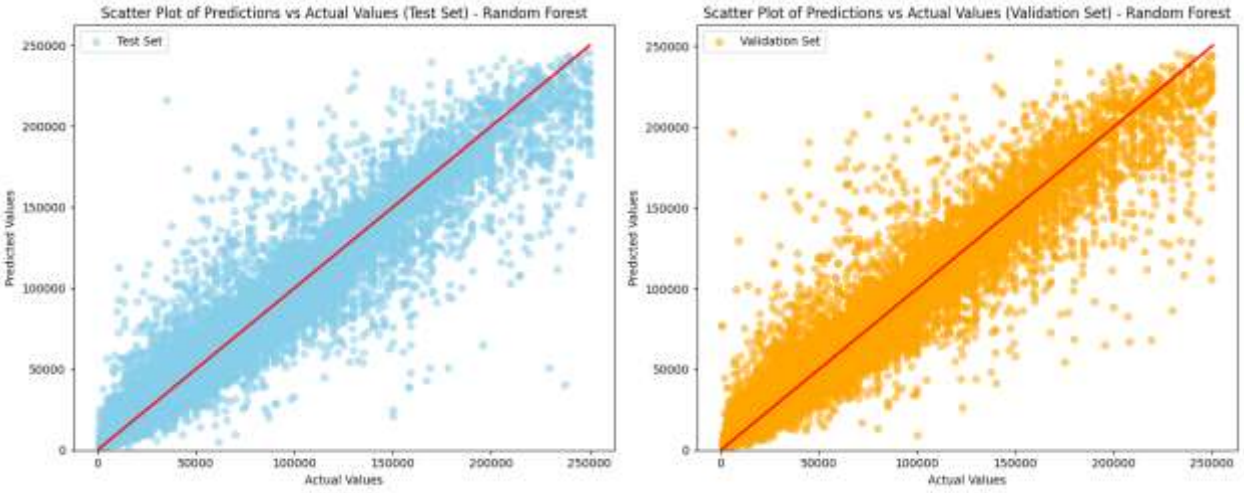
Đối chiếu trên tập đánh giá:



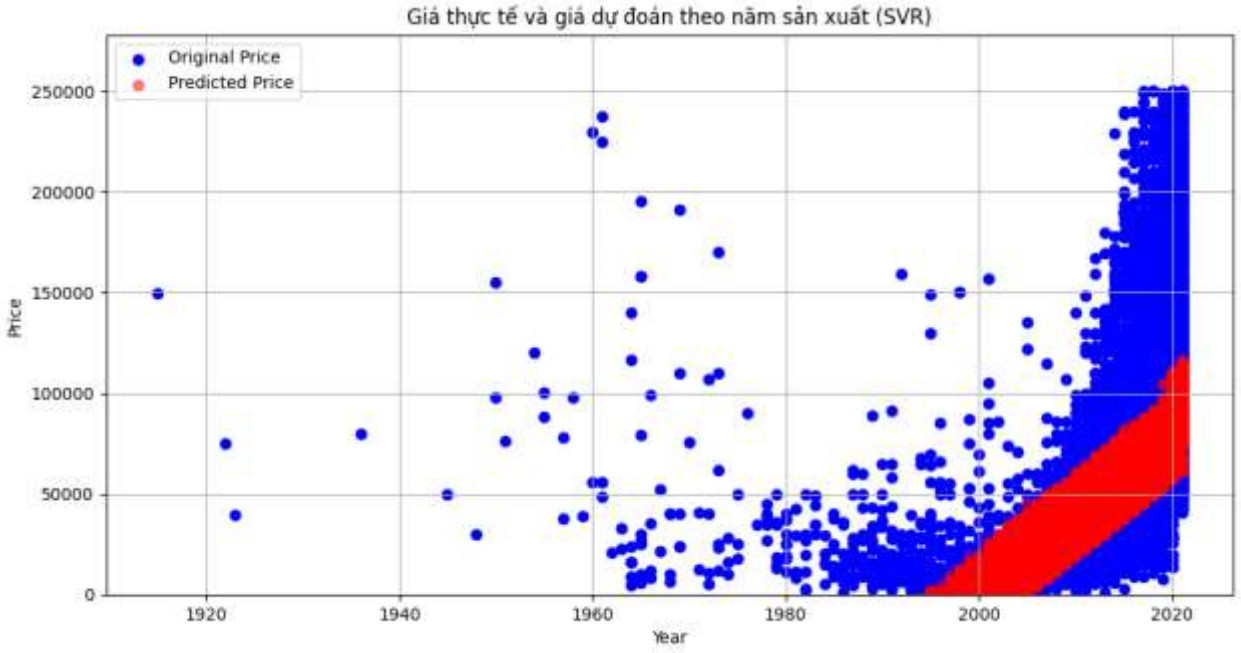
b) **Random Forest:** Mô hình Random Forest dự đoán các giá trị định giá theo năm sản xuất đạt độ chính xác rất cao, đặc biệt với những xe có năm sản xuất từ năm 1980 trở đi. Không có giá trị âm được quan sát từ tập dự đoán trên Test set. Các điểm dữ liệu phân phối rất theo sát đường tuyến tính.



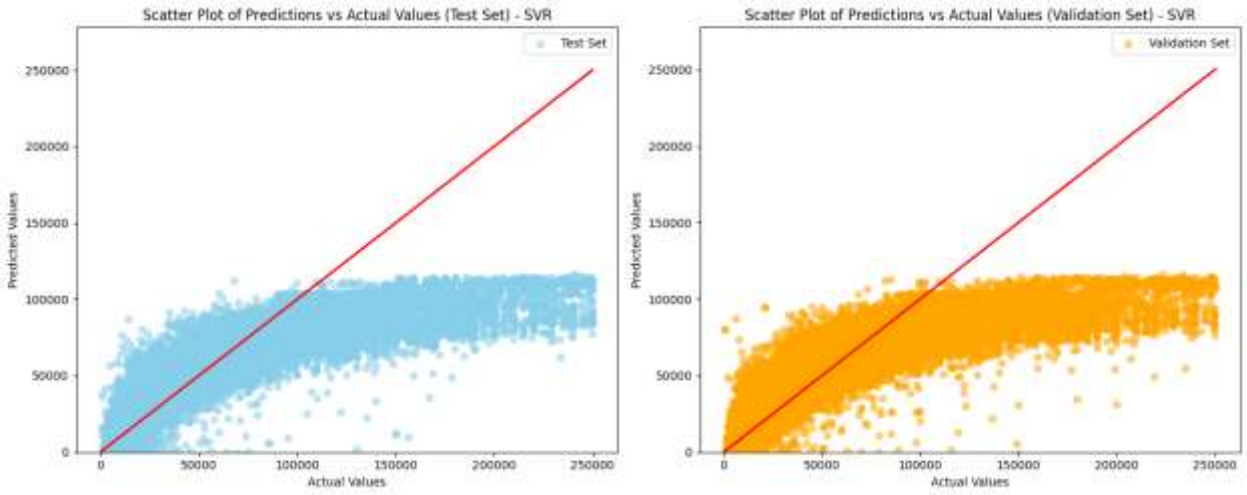
Đối chiếu trên tập đánh giá:



c) **Support Vector Regressor:** Mô hình SVR dự đoán các giá trị định giá theo năm sản xuất bị sai lệch tương đối lớn và có sai số rất nhiều, không quan sát được các định giá trước khoảng 1990 được dự đoán.



Đối chiếu trên tập đánh giá:



5.3.3 Kết luận

Dựa trên các mô hình được chọn thì ta thấy mô hình Random Forest có khả năng dự đoán đúng rất cao trên cả tập test và validation với sai số tuyệt đối chỉ lệch khoảng 7.000 đơn vị PLN, và tổng các sai số dự đoán là 173 nghìn PLN với tất cả các giá trị thực tế, nhỏ hơn gấp nhiều lần so với hai mô hình còn lại (MSE nhỏ hơn xấp xỉ 3 lần với Linear Regression và nhỏ hơn tới 5,8 lần khi so với SVR).

Thông qua mô hình dự đoán giá xe ô tô và kết quả dự báo, quá trình này không chỉ cung cấp một cái nhìn chính xác về giá trị dự kiến của chiếc xe dựa trên các đặc điểm quan trọng như mẫu xe, năm sản xuất, tổng quãng đường đi, động cơ, mà còn mang lại nhiều lợi ích quan trọng cho cả người mua và người bán xe.

Đối với người mua, việc có thông tin chính xác về giá trị dự kiến của chiếc xe giúp họ đưa ra quyết định mua sắm thông minh. Họ có thể so sánh giá trị dự kiến với giá bán thực tế, từ đó xác định xem chiếc xe có được định giá công bằng hay không. Thông tin này làm tăng tính minh bạch và tin cậy trong quá trình mua bán, giảm thiểu rủi ro của người mua khi đối mặt với giá cả không chính xác hoặc không hợp lý.

Ngược lại, đối với doanh nghiệp, việc có khả năng dự đoán giá chính xác giúp họ đưa ra chiến lược kinh doanh hiệu quả. Những thông tin này có thể hỗ trợ quyết định về việc giữ hay bán xe, đặt giá bán mục tiêu, và điều chỉnh chiến lược quảng cáo. Ngoài ra, việc nghiên cứu thị trường và dự đoán xu hướng giúp họ nắm bắt được nhu cầu thị trường, đưa ra quyết định thông minh về quảng cáo và khuyến mãi để thu hút khách hàng.

Việc định hình chiến lược kinh doanh thông qua thông tin về giá giúp doanh nghiệp cạnh tranh được thị phần. Chiến lược quảng cáo và giá cả cạnh tranh có thể được xây dựng dựa trên thông tin chính xác về giá trị xe ô tô, làm tăng khả năng thu hút và giữ chân khách hàng.

Tổng cộng, việc dự đoán giá xe ô tô không chỉ là một công cụ hữu ích cho quá trình mua bán, mà còn là một yếu tố quan trọng hỗ trợ quyết định kinh doanh và chiến lược phát triển trên thị trường ô tô cạnh tranh ngày nay.

Tài liệu tham khảo

[0] Giáo trình Biểu diễn Trực quan Dữ liệu, TS. Nguyễn An Tế, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.

[1] Giáo trình Máy học, TS. Nguyễn An Tế, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.

[2] Giáo trình Lập trình Phân tích Dữ liệu, TS. Nguyễn An Tế, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.

[3] Giáo trình Khai phá Dữ liệu, TS. Nguyễn An Tế, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.

[4] Poland Car Dealership EDA.

[5] Data Visualization with Python.