

nhom_5

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KÌ

MÔN HỌC:

NHẬP MÔN KHOA HỌC DỮ LIỆU

Đề tài:

PHÂN TÍCH GIÁ XE CŨ ĐỂ XÁC ĐỊNH XU HƯỚNG THỊ TRƯỜNG

GV: ThS. Lê Minh Tân

SVTH: Nhóm 5

- 21110202 Bùi Quốc Khang
- 21110206 Lê Việt Khanh
- 21110298 Đặng Kim Thành

CHƯƠNG 1: MỞ ĐẦU

"Phân Tích Giá Xe Cũ để xác định xu hướng thị trường" là một giải pháp toàn diện nhằm phân tích những yếu tố phức tạp trong thị trường ô tô, tập trung vào việc xác định xu hướng thị trường và của các xe ô tô đã qua sử dụng. Với sự biến động liên tục của sự kiện toàn cầu và ảnh hưởng không ngừng đổi mới ngành công nghiệp ô tô, việc hiểu rõ nhịp đập của thị trường là quan trọng đối với người tiêu dùng, nhà sản xuất, nhà đầu tư và nhà phân phối.

Mục đích thực hiện

- Khám phá những yếu tố ảnh hưởng đến giá xe ô tô.
- Đưa ra các kết luận về xu hướng thị trường.

CHƯƠNG 2: DATASETS

2.1 Dataset "Used Cars Price Prediction" (DatasetMixed)

Tổng Quan Về Tập Dữ Liệu

- Số lượng trường : 13 fields
- Số lượng bản ghi : 6020 records
- Tác giả : Co-learning Lounge | Expert | Kaggle
- Mô tả : Tập dữ liệu bao gồm thông tin về các xe đã qua sử dụng, với các chi tiết như tên xe, vị trí, năm sản xuất, số km đã di, loại nhiên liệu, hộp số, loại chủ sở hữu, mức tiêu thụ nhiên liệu, dung tích động cơ, công suất, số ghế, giá mới của xe (nếu có), và giá bán.
- Link to Dataset : [Used Cars Price Prediction \(kaggle.com\)](#)

Khám Phá Tổng Quát và Định Hướng Phân Tích

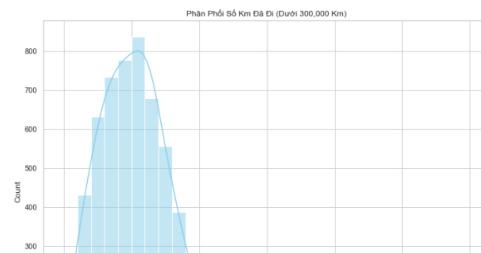
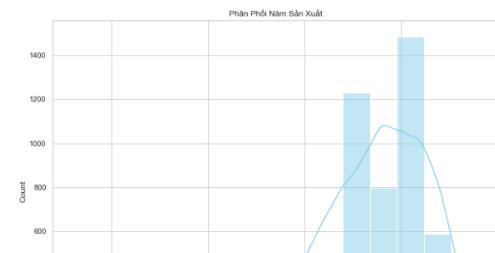
Cấu trúc và loại thông tin

- Name: Tên và mẫu xe.
- Location: Thành phố nơi xe được bán.
- Year: Năm sản xuất.
- Kilometers_Driven: Tổng số kilômét xe đã di.
- Fuel_Type: Loại nhiên liệu mà xe sử dụng (Xăng, Dầu diesel, ...).
- Transmission: Loại hộp số (Số sàn hoặc Tự động).
- Owner_Type: Số người từng sở hữu.
- Mileage: Mức tiêu thụ nhiên liệu của xe theo km/l hoặc km/kg.
- Engine: Dung tích động cơ của xe tính bằng CC.
- Power: Công suất của xe tính bằng mã lực.
- Seats: Số chỗ ngồi trong xe.
- New_Price: Giá của xe khi mới (hầu như không có)
- Price: Giá của xe đã qua sử dụng tính bằng Lakh (1 Lakh = 100000 INR = 29.801.000 VNĐ)

Year	Kilometers_Driven	Seats	Price
count	6019.000000	6.019000e+03	5977.000000
mean	2013.358199	5.873838e+04	5.278735
std	3.269742	9.126844e+04	0.888840
min	1998.000000	1.710800e+02	0.000000
25%	2011.000000	3.400000e+04	5.000000
50%	2014.000000	5.300000e+04	5.000000
75%	2016.000000	7.300000e+04	5.000000
max	2019.000000	6.500000e+06	10.000000
			160.000000

Kiểm tra giá trị thiếu

- Mileage: 2 giá trị bị thiếu.
- Engine: 36 giá trị bị thiếu.
- Power: 36 giá trị bị thiếu.
- Seats: 42 giá trị bị thiếu.
- New_Price: 5195 giá trị bị thiếu, đây là cột có số lượng giá trị bị thiếu lớn nhất.



Price (Giá xe):

- Giá xe có sự biến động lớn từ 0.44 đến 160 Lakh với giá trung bình là khoảng 9.48 Lakh. Điều này phản ánh sự đa dạng về giá của xe trên thị trường, từ xe giá rẻ đến xe sang trọng.
- Độ skewness cao (Skewness = 3.34) và độ nhọn cao (Kurtosis = 17.09) cho thấy phân phối giá xe lệch nhiều về phía giá thấp với một số ít xe có giá rất cao, tạo ra đuôi phải dài trong phân phối.

Kilometers_Driven (Số km đã đi):

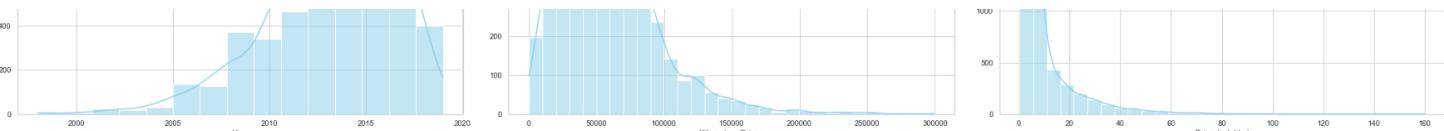
- Phạm vi số km đã đi rất rộng, từ 171 km đến 6,500,000 km, cho thấy sự đa dạng lớn về tình trạng sử dụng của xe.

- Độ skewness cao (Skewness = 58.72 và Kurtosis = 4125.09), phản ánh sự tồn tại của các giá trị ngoại lệ cực lớn, có thể là do nhập liệu sai lầm hoặc xe được sử dụng rất nhiều.

Year (Năm sản xuất):

- Xe trong dữ liệu chủ yếu sản xuất từ năm 1998 đến 2019, với số lượng xe tăng dần về phía những năm gần đây, thể hiện qua giá trị trung bình năm là khoảng 2013.36 và trung vị là 2014.

- Độ skewness âm (Skewness = -0.85) chỉ ra rằng phân phối có xu hướng hơi lệch về phía các năm gần đây, trong khi độ nhọn (Kurtosis = 0.89) gần với phân phối chuẩn cho thấy phân bố tương đối đều.

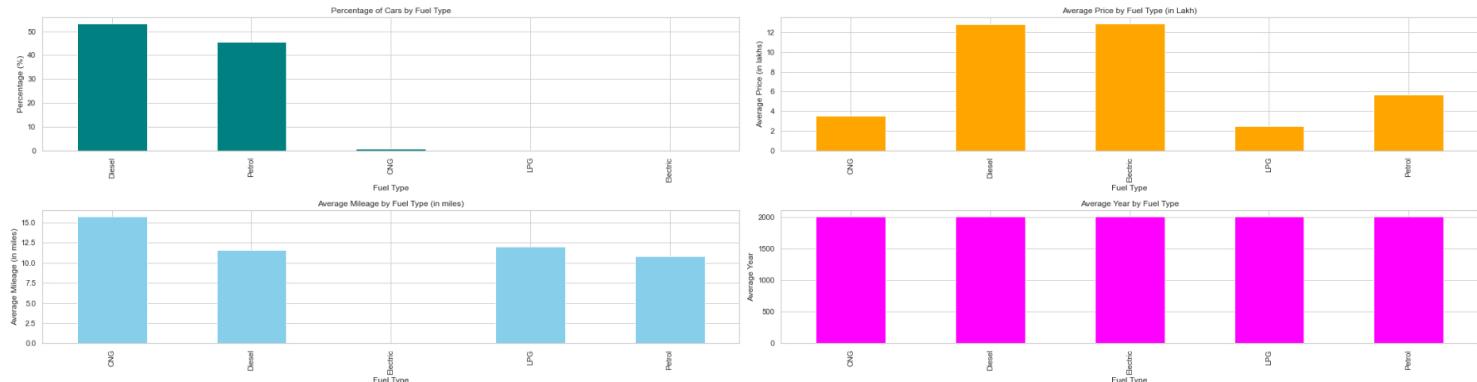


Biểu đồ phân phối cho các biến đã cho thấy:

- Năm Sản Xuất: Phần lớn xe thuộc về giai đoạn sau năm 2010, với sự tăng dần về số lượng xe mới được bán ra.
- Số Km Đã Đì: Số km đã đi dưới 300,000 km, với dịnh phân phối khoảng 50,000 km, cho thấy hầu hết xe đã qua sử dụng được bán lại có số km đi vừa phải.
- Giá Xe: Phân phối giá có sự tập trung cao ở phía dưới, với một số xe có giá rất cao, điều này phản ánh sự chênh lệch lớn về giá giữa các loại xe.

Thống kê mô tả

- Xe trong dữ liệu có từ năm 1998 đến 2019, với trung bình khoảng năm 2013. Điều này cho thấy dữ liệu bao gồm cả xe cũ và xe khá mới.
- Số km đã đi có phạm vi rất lớn, từ chỉ 171 km đến 6,500,000 km, nhưng trung bình là khoảng 58,738 km. Có vẻ như có một số bản ghi với giá trị cực lớn có thể là ngoại lệ.
- Số ghế dao động từ 0 (có thể là dữ liệu bị thiếu hoặc nhập sai) đến 10 ghế, với số ghế phổ biến nhất là 5 ghế.
- Giá xe có sự biến động lớn, từ 0.44 đến 160 Lakh (tiền Ấn Độ), cho thấy sự đa dạng về loại xe từ giá rẻ đến xe sang trọng.



- Tỷ Lệ Phân Trâm Xe Theo Loại Nhiên Liệu:** Diesel chiếm tỷ lệ lớn nhất trong tập dữ liệu, tiếp theo là Petrol, điều này phản ánh sự ưu tiên nhiên liệu trong thị trường xe đã qua sử dụng.
- Giá Trung Bình Theo Loại Nhiên Liệu:** Xe chạy điện có giá trung bình cao nhất, điều này có thể do công nghệ xe điện còn khá mới và giá xe ban đầu cao. Xe Diesel vẫn có một mức giá cao hơn xe Petrol mà thực tế là điều này có thể do hiệu suất và độ tin cậy của Diesel đối với các chuyến đi dài và khả năng tiêu thụ nhiên liệu hiệu quả. Các loại nhiên liệu còn lại có giá trung bình thấp hơn có thể do chi phí nhiên liệu thấp và không phải là lựa chọn phổ biến nhất.
- Quãng Đường Trung Bình (Mileage) Theo Loại Nhiên Liệu:** Xe Diesel có quãng đường trung bình cao hơn so với các loại xe khác, có thể do Diesel thường được sử dụng cho các chuyến đi dài và vận hành bền bỉ.
- Năm Sản Xuất Trung Bình Theo Loại Nhiên Liệu:** Các loại xe khác nhau không có sự chênh lệch đáng kể về năm sản xuất, cho thấy rằng không có sự ưu tiên rõ ràng về năm sản xuất giữa các loại nhiên liệu.

Diesel và Petrol là hai loại nhiên liệu phổ biến, nhưng sự chênh lệch về giá có thể phản ánh tầm nhìn dài hạn của người mua xe về chi phí nhiên liệu và bảo dưỡng. Xe điện có giá cao có thể do người tiêu dùng sẵn lòng trả giá cao hơn cho công nghệ mới và thân thiện với môi trường. Xe Diesel thường được sử dụng nhiều hơn, điều này phản ánh trong quãng đường trung bình cao của chúng.

2.2 Dataset “Used Cars data form websites” (DatasetIndiaWebsite)

Tổng Quan Về Tập Dữ Liệu

- Số lượng trường: 13 fields
- Số lượng bản ghi: 8128 records
- Tác giả: Nikhil Kushwaha & Nishant Verma
- Mô tả: Tập dữ liệu bao gồm thông tin về các xe đã qua sử dụng, với các chi tiết như tên xe, vị trí, năm sản xuất, số km đã đi, loại nhiên liệu, hộp số, loại chủ sở hữu, mức tiêu thụ nhiên liệu, dung tích động cơ, công suất, số ghế, giá mới của xe (nếu có), và giá bán.
- Link to Dataset : [Used Cars data form websites](#)

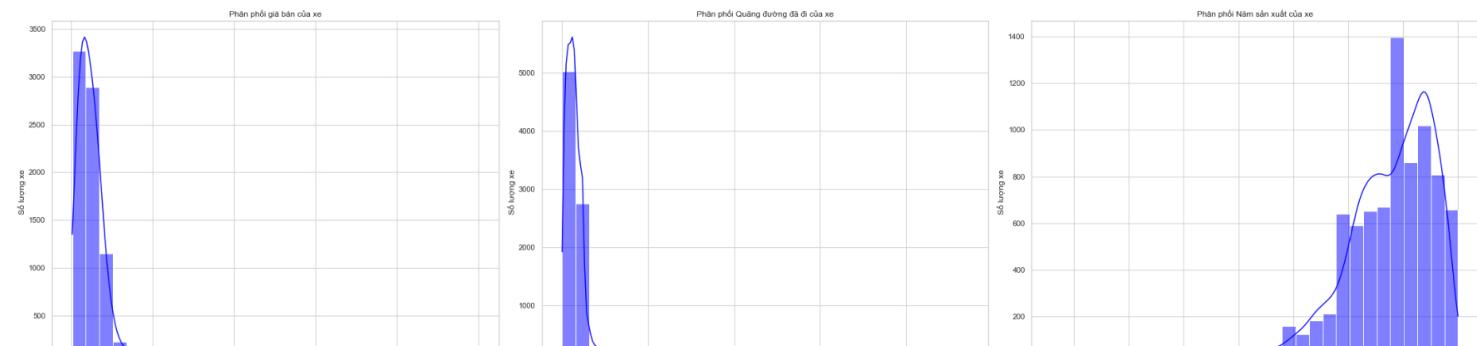
Khám Phá Tổng Quát và Định Hướng Phân Tích

Cấu trúc và loại thông tin

- name: Tên của xe (ví dụ: Maruti Swift Dzire VDI).
- year: Năm sản xuất của xe.
- selling_price: Giá bán của xe (đơn vị tiền tệ không được chỉ định, giả sử là INR, vì tập dataset được lấy ở Ấn Độ).
- km_driven: Số kilômét mà xe đã đi.
- fuel: Loại nhiên liệu sử dụng (Diesel, Petrol, v.v.).
- seller_type: Loại người bán (Cá nhân, Đại lý, v.v.).
- transmission: Loại hộp số (Manual hoặc Automatic).
- owner: Thông tin về chủ sở hữu hiện tại (First Owner, Second Owner, v.v.).
- mileage: Mức tiêu thụ nhiên liệu của xe (kmpl hoặc km/kg).
- engine: Dung tích động cơ của xe (ví dụ: 1248 CC).
- max_power: Công suất tối đa của xe (ví dụ: 74 bhp).
- torque: Mô-men xoắn của động cơ.
- seats: Số chỗ ngồi trong xe.

Kiểm tra giá trị bị thiếu

- mileage: 221 giá trị bị thiếu.
- engine: 221 giá trị bị thiếu.
- max_power: 215 giá trị bị thiếu.
- torque: 222 giá trị bị thiếu.
- seats: 221 giá trị bị thiếu.





Phân bố giá: Biểu đồ trên thể hiện phân phối của giá bán xe trong tập dữ liệu. Ta có thể thấy rằng phần lớn các xe có giá bán tập trung ở mức thấp hơn, với số lượng giảm dần khi giá bán tăng lên. Điều này cho thấy rằng xe có giá rẻ là phổ biến hơn trên thị trường xe cũ.

Phân bố quãng đường đã đi: Biểu đồ phân phối quãng đường đã đi của xe cho thấy số các xe được bán lại khi chưa di quá nhiều, với phần lớn tập trung dưới 100.000 km. Phân phối lệch phải này phản ánh thực tế là xe càng di nhiều, số lượng xe còn lại trên thị trường càng ít, điều này có thể liên quan đến lo ngại về độ bền và chi phí bảo trì tăng theo thời gian sử dụng xe. Sự hiện diện của các xe có quãng đường cực cao là khá hiếm và có thể chỉ ra rằng đây là những xe được bảo dưỡng kỹ lưỡng hoặc sử dụng trong các ngành công nghiệp cụ thể như taxi hoặc vận chuyển.

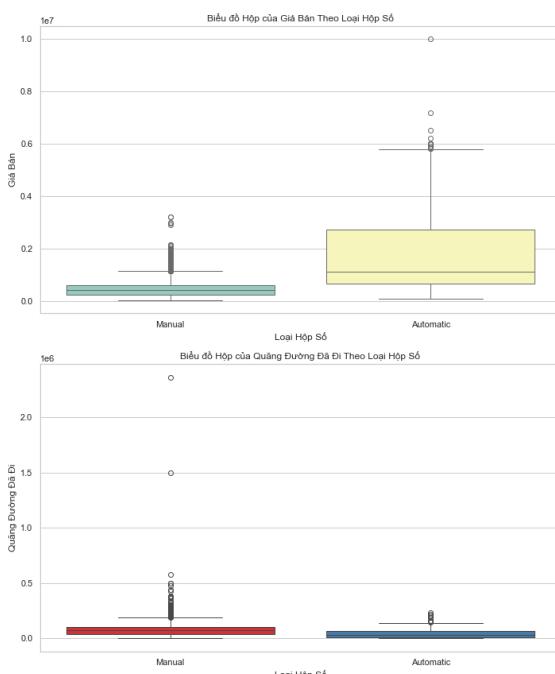
Phân bố năm đăng ký: Biểu đồ phân phối năm sản xuất của xe cho thấy sự tăng trưởng rõ rệt về số lượng xe được sản xuất và bán lại trong những năm gần đây. Điều này có thể phản ánh nhu cầu mua sắm xe mới liên tục tăng và xu hướng của người dùng là thay đổi xe thường xuyên hơn. Ngoài ra, sự tập trung dày đặc các xe sản xuất từ năm 2010 trở đi cũng cho thấy thị trường xe cũ có xu hướng ưa chuộng các mẫu xe mới hơn lý do kinh tế và an toàn.

	selling_price	km_driven	year
count	8,128,000+03	8,128,000+03	8128,000000
mean	6,382718e+05	6,091951e+04	2013,894011
std	8,062534e+05	5,655055e+04	4,442429
min	2,099900e+04	1,000000e+00	1983,000000
25%	2,549990e+05	3,500000e+04	2011,000000
50%	4,500000e+05	6,000000e+04	2015,000000
75%	6,750000e+05	9,000000e+04	2017,000000
max	1,000000e+07	2,360457e+06	2020,000000
variance	6,590446e+11	3,197965e+08	16,355948
kurtosis	2,108129e+01	3,840974e+02	1,707013
skewness	4,193533e+00	1,117091e+01	-1,072293

• **Giá bán (selling_price):** Giá bán có phạm vi rộng từ 29,999 đến 10,000,000, với độ lệch và độ nhọn cao, cho thấy sự phân bố không đồng đều và có một số giá trị ngoại lệ cao rất đáng kể. Điều này có thể ảnh hưởng đến quyết định giá xe cũ trên thị trường.

• **Quãng đường đã đi (km_driven):** Dữ liệu này cũng cho thấy độ lệch và độ nhọn rất cao, với một số xe có số km đã đi đặc biệt cao so với mức trung bình, phản ánh sự khác biệt lớn về mức độ sử dụng của xe.

• **Năm sản xuất (year):** Phân bố sản xuất cho thấy hầu hết các xe trong dữ liệu được sản xuất gần đây, với sự tập trung chính từ năm 2011 đến 2020. Độ lệch âm cho thấy một xu hướng nhẹ về phía các mẫu xe mới hơn trong dữ liệu.



Biểu đồ Hộp của Giá Bán Theo Loại Hộp Số

Phân tích:

- **Hộp số tự động:** Biểu đồ cho thấy giá bán của xe với hộp số tự động có sự phân tán rất rộng, với một số giá trị cực cao (như thể hiện bởi các điểm dữ liệu ngoại lệ). Điều này có thể chỉ ra rằng xe tự động có thể bao gồm cả mô hình cơ bản và các mô hình cao cấp với nhiều tính năng hơn. Đây có thể là một dấu hiệu của sự đa dạng trong nhóm xe tự động về cấu hình và giá bán.

- **Hộp số sàn:** Phạm vi và phân tán của giá bán xe với hộp số sàn thì hẹp hơn nhiều. Điều này có thể cho thấy mức độ giá bán của xe hộp số sàn ít biến động hơn và chúng thường nằm trong phân khúc giá thấp hơn.

Biểu đồ Hộp của Quãng Đường Đã Di Theo Loại Hộp Số

Phân tích:

- **Hộp số sàn:** Median (đường trong biểu đồ hộp) nằm khá thấp, gần với cạnh dưới của hộp, cho thấy rằng một nửa số xe có quãng đường đi rất thấp. Hộp có phạm vi (khoảng cách giữa cạnh trên và cạnh dưới) khá hẹp, điều này cho thấy đa số xe hộp số sàn có quãng đường đi tương đối đồng đều và không phân tán rộng. Có một số ngoại lệ (các điểm dữ liệu đơn lẻ nằm xa hộp), điều này chỉ ra rằng mặc dù số xe hộp số sàn có quãng đường thấp nhưng vẫn có một số xe đi được quãng đường rất xa.

- **Hộp số tự động:** Median của hộp số tự động cũng nằm khá thấp và thậm chí còn thấp hơn so với hộp số sàn, cho thấy một nửa số xe tự động có quãng đường đi thấp. Sự phân bố của các điểm dữ liệu ngoại lệ không đáng kể so với hộp số sàn, điều này có thể cho thấy xe tự động có xu hướng được sử dụng trong phạm vi hạn chế hơn.

2.3 Dataset “Used Cars Prices in UK” (DatasetUK)

Tổng Quan Về Tập Dữ Liệu

- Số lượng trường : 13 fields
- Số lượng bài ghi : 4727 records
- Tác giả : Muhammad Awais Tayyab
- Mô tả : Tập dữ liệu này tập trung vào thị trường xe cũ ở Vương quốc Anh, bao gồm thông tin về giá xe, năm sản xuất, số kilomet đã đi, loại nhiên liệu, và nhiều hơn nữa. Đây là nguồn thông tin hữu ích để phân tích và hiểu về xu hướng giá xe cũ ở Vương quốc Anh.
- Link to Dataset : [Used Cars Prices in UK](#)

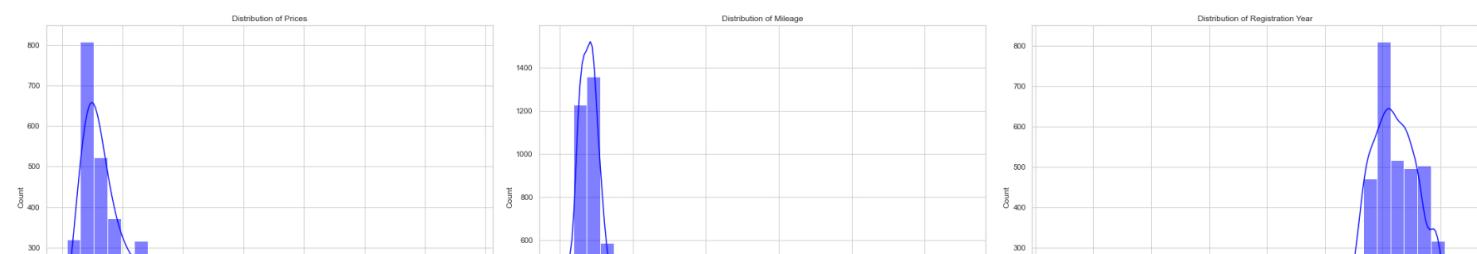
Khám Phá Tổng Quát và Định Hướng Phân Tích

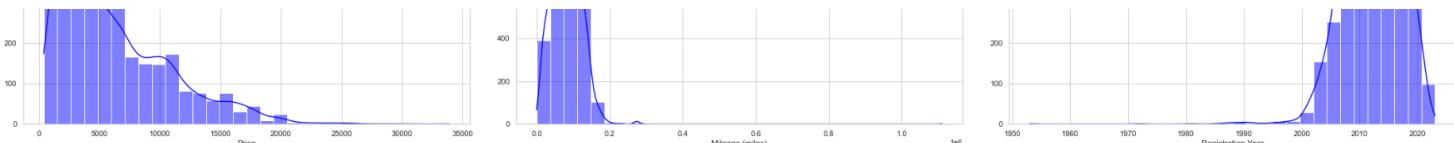
Cấu trúc và loại thông tin

- Tiêu đề xe (Title): Tên của xe, có thể bao gồm thương hiệu và mẫu xe.
- Giá (Price): Giá của xe ô tô đã qua sử dụng.
- Quãng đường đã đi (Mileage(miles)): Số dặm xe đã đi.
- Năm đăng ký (Registration_Year): Năm mà xe được đăng ký lần đầu.
- Số chủ sở hữu trước (Previous Owners): Số lượng chủ sở hữu trước của xe.
- Loại nhiên liệu (Fuel type): Loại nhiên liệu mà xe sử dụng.
- Kiểu thân xe (Body type): Kiểu dáng của xe, như hatchback, sedan, v.v.
- Động cơ (Engine): Thông tin về dung tích động cơ của xe.
- Hộp số (Gearbox): Loại hộp số, tự động hoặc số mi.
- Số cửa (Doors): Số lượng cửa của xe.
- Số chỗ ngồi (Seats): Số chỗ ngồi trong xe.
- Lớp phát thải (Emission Class): Tiêu chuẩn khí thải của xe.
- Lịch sử bảo dưỡng (Service history): Thông tin về lịch sử bảo dưỡng của xe.

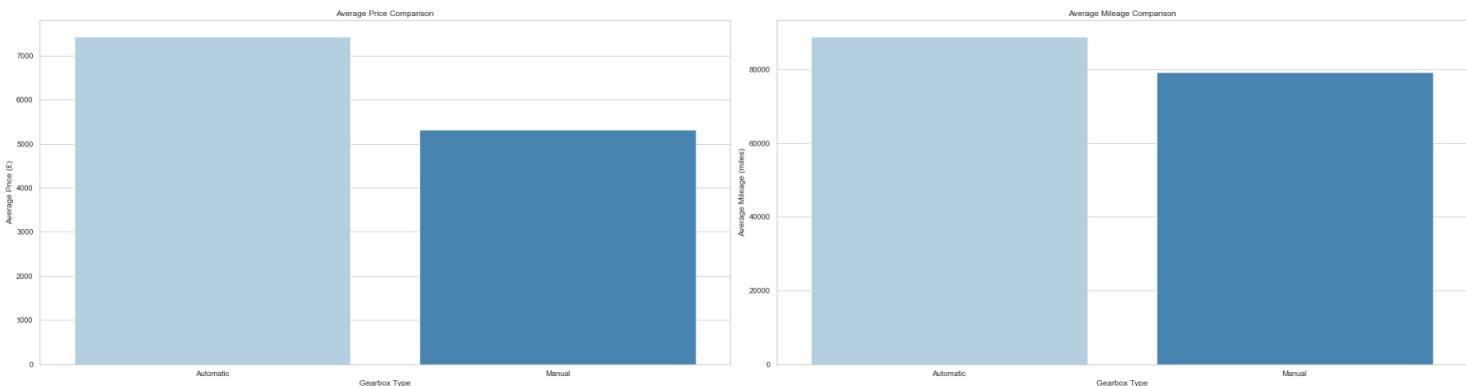
Kiểm tra giá trị bị thiếu

- Previous Owners: 1,409 giá trị thiếu
- Engine: 45 giá trị thiếu
- Doors: 25 giá trị thiếu
- Seats: 35 giá trị thiếu
- Emission Class: 87 giá trị thiếu
- Service history: 3,145 giá trị thiếu



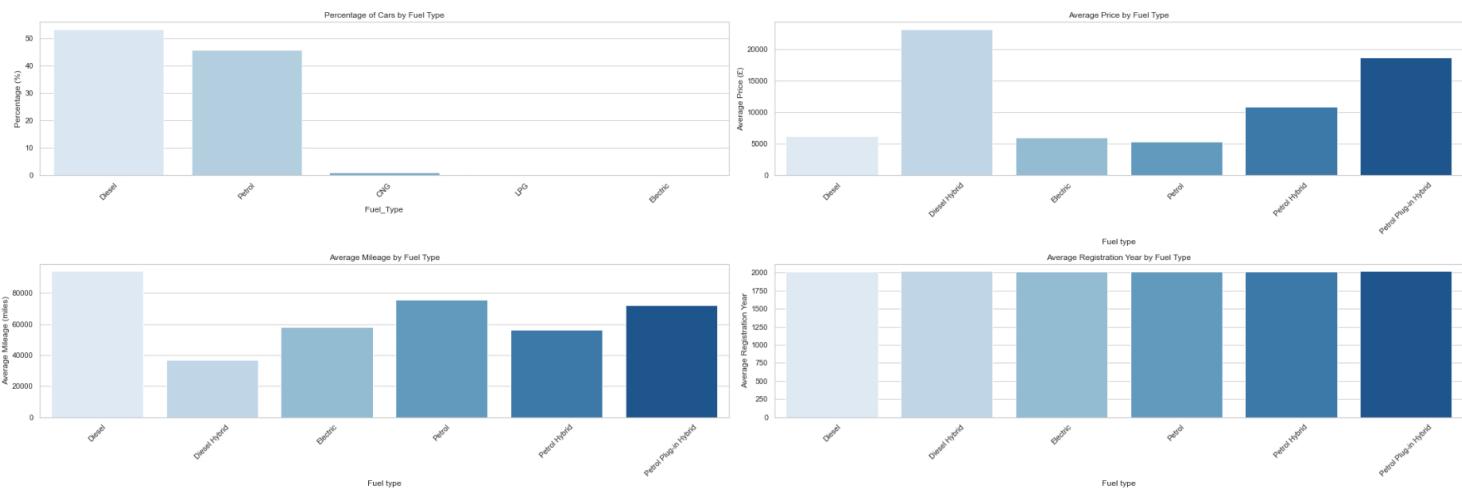


- Phân bố giá:** Biểu đồ phân bố giá cho thấy phần lớn xe nằm trong tầm giá thấp và trung bình, với một số ít xe có giá rất cao.
- Phân bố quãng đường đã đi:** Phần lớn các xe có quãng đường từ 50,000 đến 100,000 dặm, cho thấy đây là khoảng quãng đường phổ biến cho xe đã qua sử dụng trên thị trường.
- Phân bố năm đăng ký:** Có sự tập trung cao của xe đăng ký trong khoảng 10 năm gần đây, phản ánh xu hướng mua bán xe có độ mới cao hơn.



- Giá Trung Bình:** Có sự chênh lệch rõ ràng giữa xe tự động và xe sàn, với xe tự động có giá trung bình cao hơn. Điều này phản ánh sự chênh lệch về giá trị và ưu tiên của người mua đối với sự tiện lợi và thoải mái mà hộp số tự động mang lại.
- Quãng Đường Đã Di Trung Bình:** Xe tự động cũng cho thấy quãng đường đã di trung bình cao hơn so với xe sàn, điều này có thể liên quan đến loại người dùng và môi trường sử dụng xe.

Xe tự động thường được xem là có giá trị giá tăng do tính năng tiện lợi, dễ lái, đặc biệt trong điều kiện giao thông đô thị hay các tình huống đòi hỏi ít sự tương tác từ phía người lái, điều này có thể giải thích sự chênh lệch về giá và quãng đường đã di giữa hai nhóm xe.



- Tỷ Lệ Phân Trăm của Xe theo Loại Nhiên Liệu:** Xe sử dụng xăng chiếm tỷ lệ cao nhất trong tất cả các loại nhiên liệu, điều này cho thấy sự phổ biến của xe xăng trên thị trường. Xe dùng dầu (Diesel) cũng chiếm một tỷ lệ đáng kể, nhưng ít hơn xe xăng. Các loại xe hybrid và điện chiếm tỷ lệ thấp hơn nhiều, phản ánh xu hướng chuyển dịch sang xe sạch vẫn còn khá mới và chưa được ưa chuộng rộng rãi.
- Giá Trung Bình:** Xe hybrid cắm sạc (Petrol Plug-in Hybrid) có giá trung bình cao nhất, có thể do công nghệ tiên tiến và chi phí sản xuất cao hơn. Xe hybrid xăng (Petrol Hybrid) và điện cũng có mức giá khá cao so với các loại xe truyền thống. Xe dùng xăng và dầu có giá thấp hơn, phản ánh sự phổ biến và số lượng lớn trên thị trường.
- Quãng Đường Đã Di Trung Bình:** Không có sự khác biệt đáng kể về số km trung bình giữa các loại xe, chỉ ra rằng dù loại nhiên liệu có khác nhau nhưng mức độ sử dụng xe có thể tương đương. Điều này cũng có thể phản ánh rằng các loại xe mới hơn như xe hybrid và điện chưa được sử dụng trong thời gian dài như xe xăng hay dầu.
- Năm Đăng Ký Trung Bình:** Xe hybrid cắm sạc và xe hybrid xăng có năm đăng ký cao nhất, cho thấy chúng là những mẫu xe tương đối mới trên thị trường. Xe điện cũng có năm đăng ký khá gần đây, phù hợp với xu hướng chuyển đổi sang xe không khí. Xe xăng và dầu có năm đăng ký trung bình thấp hơn, có nghĩa là chúng đã có mặt trên thị trường lâu hơn.

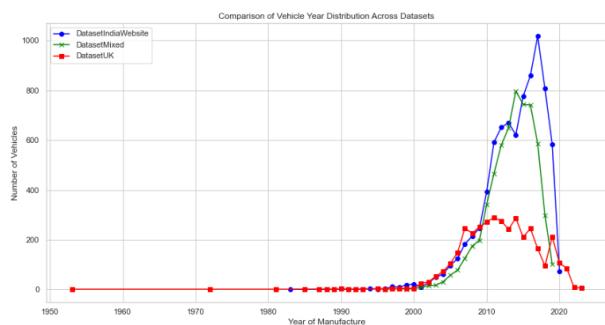
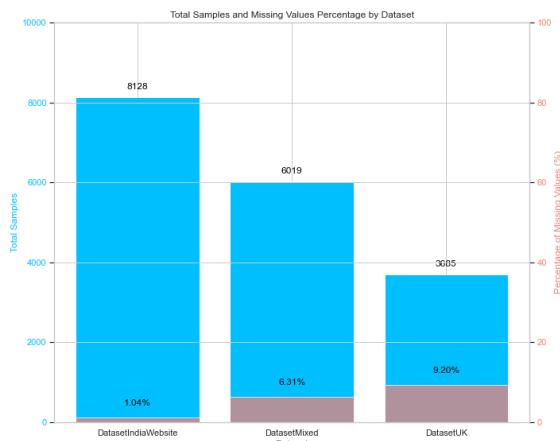
Xe sử dụng nhiên liệu xăng vẫn chiếm ưu thế do giá thành thấp và sự phổ biến, trong khi xe diesel được ưa chuộng vì khả năng tiết kiệm nhiên liệu. Xe hybrid và plug-in hybrid tuy chiếm tỷ lệ nhỏ nhưng có giá cao và năm đăng ký gần đây, cho thấy sự quan tâm ngày càng tăng đối với xe tiết kiệm năng lượng và thân thiện với môi trường.

CHƯƠNG 3. SO SÁNH DỮ LIỆU VÀ CHỌN TẬP DATASET CHÍNH

3.1 So sánh tổng quát

- Sự Tương Đồng:** Cả ba bộ dữ liệu đều chứa thông tin cơ bản về xe cũ như tên, năm sản xuất, loại nhiên liệu, số chỗ ngồi, thông tin động cơ, và giá xe. Điều này giúp chúng ta có thể so sánh và phân tích giá xe trên các thị trường và các điều kiện sử dụng khác nhau.
- Sự Khác Biệt:**

Cột	DatasetMixed	DatasetIndiaWebsite	DatasetUK
Cấu trúc dữ liệu	Có 13 cột nhưng bao gồm thông tin vị trí xe được bán và giá xe mới dự kiến (chỉ có cho một số mẫu). Tổng tin này hữu ích cho việc phân tích sự chênh lệch giữa giá mới và giá bán lại, cũng như ảnh hưởng của địa điểm đến giá xe.	Bao gồm 13 cột thông tin liên quan đến các chi tiết xe như tên xe, năm sản xuất, giá bán, loại nhiên liệu, hộp số, và các thông số kỹ thuật động cơ. Tập dữ liệu này rất chi tiết về thông tin kỹ thuật xe, nhưng thiếu thông tin về vị trí bán.	Gồm 14 cột, bao gồm các thông tin về sức chở lớn nhất, số lượng thắc xe, sức vận hành, chi phí, lực cản, độ động. Sự bao lưu này cung cấp cái nhìn sâu hơn về trạng thái và các tiêu chí chất lượng.
Định dạng dữ liệu	Tương tự như DatasetIndiaWebsite, nhưng có thêm cột New Price với dữ liệu bị thiếu đáng kể, chỉ có giá trị cho một phần nhỏ của tập dữ liệu.	Chủ yếu là dữ liệu dạng văn bản (object) với số (int, float). Các trường như mileage, engine, và max_power được lưu trữ dưới dạng chuỗi, có thể cần chuyển đổi để phản ánh dữ liệu.	Có sự khác biệt về định dạng, nhưng cả các cột Emission Class và Service history, cung cấp thông tin chi tiết hơn về trạng thái và môi trường của xe.
Quy mô dữ liệu	Có 6019 mẫu, đa dạng về địa điểm và cơ sở dữ liệu về giá xe mới.	Là tập dữ liệu lớn nhất với 8128 mẫu, cung cấp độ phủ rộng về số lượng xe được bán.	Nhỏ nhất với 3685 mẫu, nhưng chi tiết về các yếu tố ảnh hưởng đến giá bán lại và môi trường.



- DatasetMixed với 6.019 mẫu, đứng ở mức trung bình về số lượng mẫu nhưng có tỷ lệ thiếu dữ liệu cao hơn ở mức 6.79%. Tập dữ liệu này có thể chứa đủ thông tin cho nhiều loại phân tích nhưng yêu cầu công sức nhiều hơn để xử lý dữ liệu thiếu.
- DatasetIndiaWebsite nổi bật với số lượng mẫu lớn nhất là 8.128 và tỷ lệ dữ liệu bị thiếu thấp nhất chỉ 1.04%. Điều này cho thấy tập dữ liệu này không chỉ có quy mô lớn mà còn rất đầy đủ, giảm thiểu khả năng phát sinh lỗi do thiếu dữ liệu trong các phân tích.
- DatasetUK có số lượng mẫu thấp nhất là 3.685 và tỷ lệ dữ liệu bị thiếu cao nhất là 9.20%. Sự thiếu hụt này có thể hạn chế khả năng phân tích tách rời và chính xác các xu hướng hoặc mẫu từ tập dữ liệu.

Từ những phân tích trên, DatasetIndiaWebsite dường như là lựa chọn tốt nhất để sử dụng cho các nhu cầu phân tích dữ liệu tiếp theo. Tập dữ liệu này không chỉ cung cấp số lượng mẫu lớn nhất, mà còn đảm bảo tính toàn vẹn và độ chính xác cao nhờ vào tỷ lệ dữ liệu bị thiếu rất thấp. Điều này sẽ giúp phân tích trở nên đáng tin cậy hơn, giảm thiểu rủi ro và thời gian cần thiết cho việc xử lý sạch dữ liệu.

CHƯƠNG 4. TIỀN XỬ LÝ VÀ CHUẨN HÓA DỮ LIỆU

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8128 entries, 0 to 8127
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   name        8128 non-null   obj    
 1   year        8128 non-null   int64  
 2   selling_price 8128 non-null  int64  
 3   km_driven    8128 non-null   int64  
 4   fuel         8128 non-null   obj    
 5   seats        8128 non-null   int64  
 6   engine        8128 non-null   int64  
 7   seats_boxplot 8128 non-null   float64
 8   engine_boxplot 8128 non-null   float64
 9   used_car_age 8128 non-null   int64  
 10  mileage       8128 non-null   int64  
 11  km_driven_boxplot 8128 non-null   float64
 12  selling_price_boxplot 8128 non-null   float64
 13  target_variable 8128 non-null   float64
dtypes: int64(6), float64(3), object(4)
memory usage: 78.4 KB

```

	name	year	selling_price	km_driven	fuel	seats	engine	used_car_age	target_variable
0	Maruti Swift Dzire VXi	2014	...	5	Diesel	5.0	6254	10	0.0
1	Skoda Rapid 1.5 TDI Ambition	2014	...	5	Petrol	7.8	1128	10	0.0
2	Honda City 2017-2020 EXLi	2006	...	5	CNG	8.8	236	18	0.0
3	Hyundai i20 Sportz Diesel	2010	...	5	LPG	4.0	133	14	0.0
4	Maruti Swift VXI BSIII	2007	...	5	Name: count, dtype: int64	9.0	89	17	0.0
						6.0	62		0.0
						10.0	19		0.0
						2.0	2		0.0
						14.0	1		0.0

Bước 1: Kiểm tra các giá trị null

Bước 2: Chỉ sử dụng 2 loại fuel là Diesel và Petrol

Bước 3: Loại bỏ các hàng có giá trị null ở cột [seats]

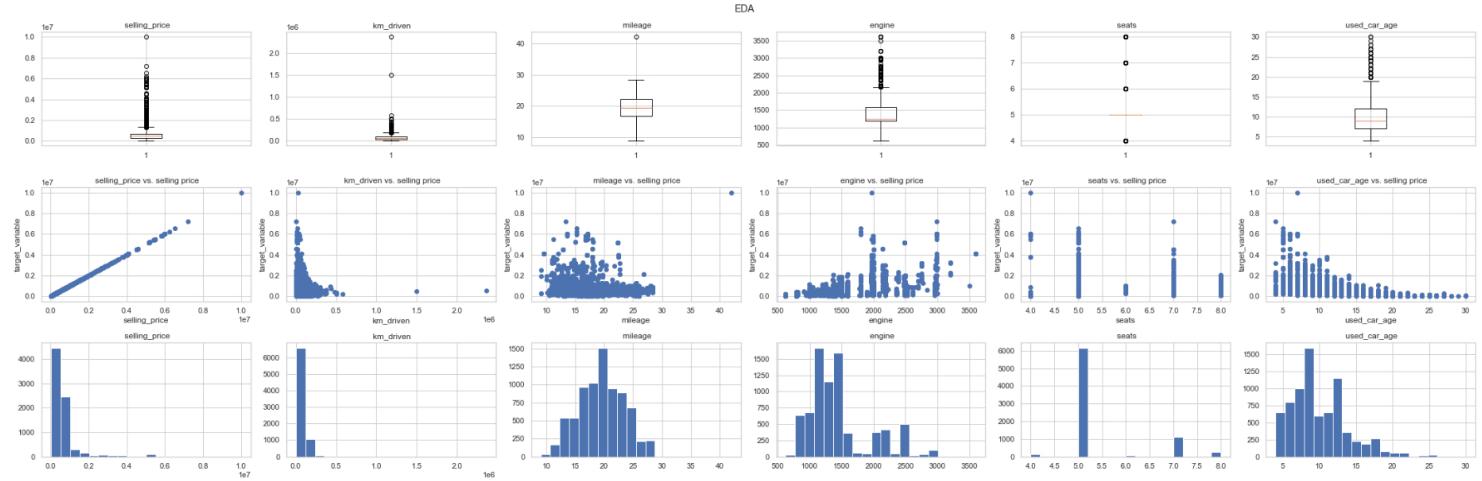
Bước 4: Chỉ sử dụng những xe có số ghế ngồi từ 8 trở xuống

Bước 5: Thay thế chuỗi "kmpl" trong cột [mileage]

Bước 6: Đổi kiểu dữ liệu của các cột

Bước 7: Loại bỏ các hàng có giá trị [mileage] là 0.0

Bước 8: Tính tuổi của xe đã qua sử dụng bằng cách trừ năm hiện tại với năm mua xe

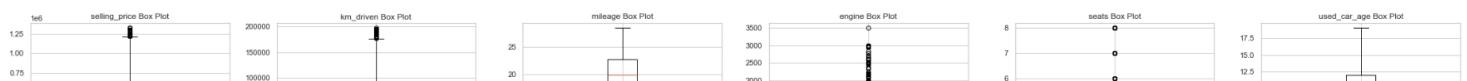


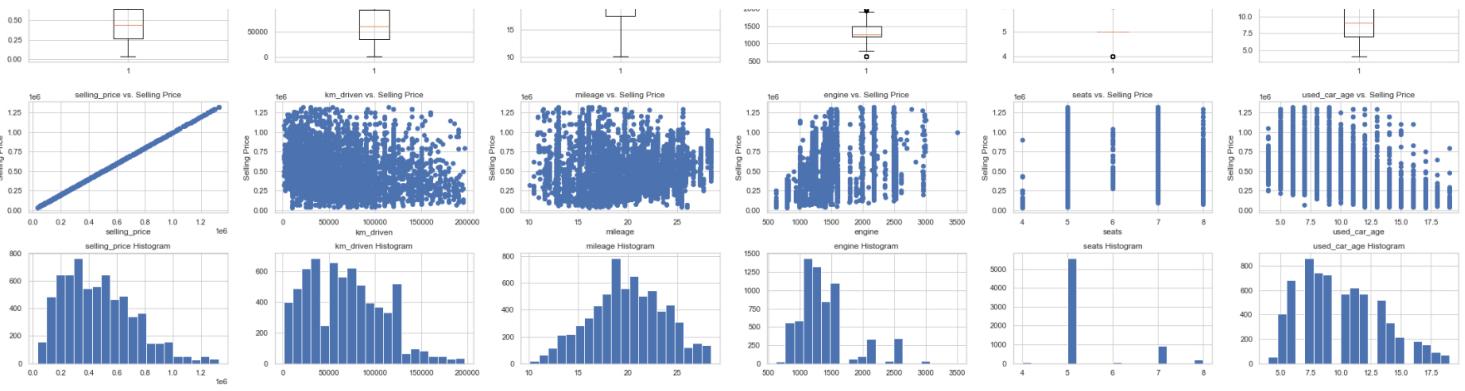
- Ở đây số lượng các giá trị ngoại lệ trong các cột của DataFrame được tìm thấy như sau: selling_price: 590 giá trị ngoại lệ, km_driven: 192 giá trị ngoại lệ, mileage: 1 giá trị ngoại lệ, used_car_age: 163 giá trị ngoại lệ

Tiến hành kiểm tra, loại bỏ ngoại lai và trực quan hóa dữ liệu:

```
Checking outliers:
Number of outliers in column 'selling_price': 590
Number of outliers in column 'km_driven': 192
Number of outliers in column 'mileage': 1
Number of outliers in column 'used_car_age': 163
```

EDA After Removing Outliers

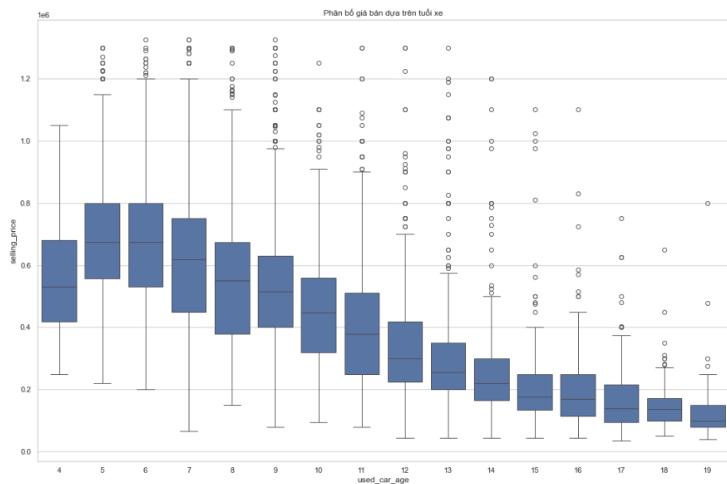




- Kết luận:** sau khi kết thúc phần tiền xử lí, nhìn chung ta thấy phạm vi dữ liệu đã được thu hẹp, với ít hoặc không có giá trị ngoại lệ còn sót lại, các biến và giá bán mà không bị ảnh hưởng bởi ngoại lệ, phân phối đồng đều hơn của các biến sau khi loại bỏ các giá trị bất thường, giúp cung cấp cái nhìn sâu sắc hơn về hình dạng phân phối thực tế của dữ liệu.
- Tổng kết:** dữ liệu của tập dataset đã được làm sạch và có 6,811 hàng và 10 cột, tất cả các cột đều không có giá trị null. Điều này có nghĩa là sau quá trình loại bỏ các giá trị ngoại lệ và dữ liệu không phù hợp, tổng số hàng đã giảm đi khoảng 16.20%, giúp cho dữ liệu trở nên sạch sẽ hơn và có khả năng cải thiện chất lượng của các phân tích và mô hình hóa sẽ được thực hiện sau này.

CHƯƠNG 5: KHAI PHÁ DỮ LIỆU

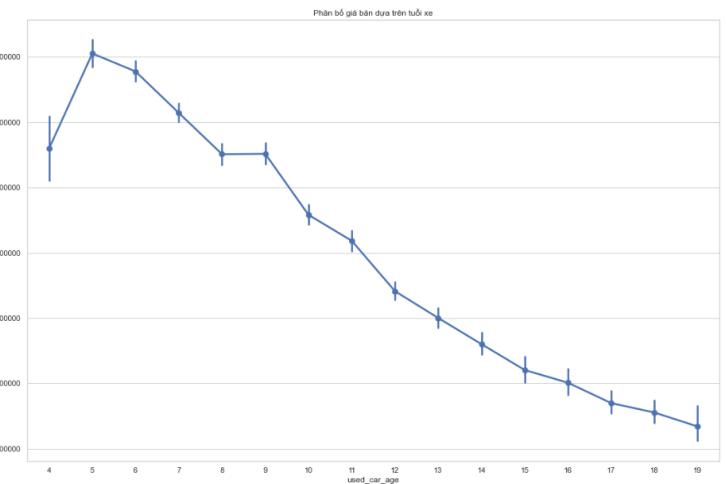
5.1 Mối quan hệ giữa tuổi xe và giá bán



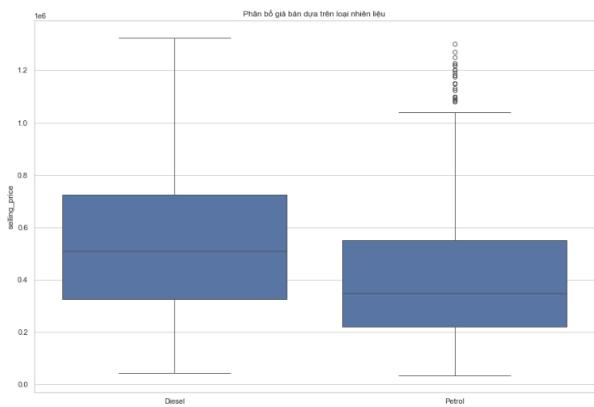
- Giá giảm theo tuổi xe: Giá xe có xu hướng giảm khi tuổi xe tăng, điều này phản ánh sự mất giá theo thời gian của xe. Xe càng mới thì thường có giá bán cao hơn.
- Biến động giá: Các xe mới hơn (tuổi từ 0 đến khoảng 5 năm) có biến động giá lớn hơn so với xe cũ hơn. Điều này có thể được thấy qua các "whisker" và "outliers" rộng hơn trên biểu đồ cho các nhóm tuổi này.
- Outliers: Có một số giá bán ngoại lệ ở các xe mới hơn, cho thấy một số mới có giá bán rất cao so với phần lớn xe cùng độ tuổi. Điều này có thể phản ánh sự khác biệt về chất lượng, thương hiệu hoặc các yếu tố đặc biệt khác của xe.
- Tập trung giá ở các xe cũ: Các xe cũ từ 10 năm trở lên có giá bán tập trung hơn và thấp hơn nhiều, điều này cho thấy sự thống nhất hơn về giá trị xe sau một khoảng thời gian sử dụng dài.

Vậy chúng ta có thể thấy được tuổi xe là yếu tố quan trọng quyết định đến giá xe. Xe càng mới thường có giá cao hơn do giá trị hao mòn thấp hơn, và giá giảm dần theo thời gian sử dụng do hao mòn và lỗi thời.

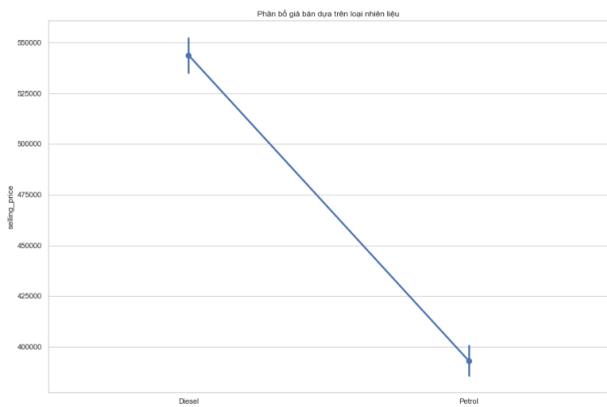
5.2 Mối quan hệ giữa loại nhiên liệu và giá bán



- Xu hướng giảm giá rõ rệt: Có thể thấy rõ ràng giá bán trung bình của xe giảm đáng kể theo thời gian. Điều này phù hợp với kỳ vọng là giá trị xe hao mòn theo thời gian.
- Sự biến động của giá bán: Biểu đồ cho thấy một số biến động trong giá trung bình ở các khoảng tuổi xe nhất định. Điều này có thể phản ánh sự khác biệt trong cung cầu, đặc điểm của các mẫu xe trong mỗi nhóm tuổi, hoặc sự chênh lệch giá giữa các loại xe cao cấp và bình dân.
- Giá bán của những xe cục mới và cục cũ: Xe mới (dưới 3 năm) có giá trung bình cao hơn hẳn so với các nhóm tuổi khác. Tương tự, giá bán của xe cũ từ 15 năm trở lên có xu hướng ổn định hơn và thấp, phản ánh giá trị thực của chúng sau một thời gian dài sử dụng.



- Biến động giá: Có sự khác biệt rõ rệt về phạm vi giá giữa các loại nhiên liệu. Xe chạy dầu diesel có phạm vi giá cao hơn và trung bình cũng cao hơn so với xe chạy xăng. Điều này có thể do xe diesel thường được ưu chuộng vì khả năng tiết kiệm nhiên liệu tốt hơn và thường được sử dụng cho các mục đích thương mại hoặc xe lớn hơn, có giá bán cao hơn.
- Outliers: Các loại nhiên liệu đều có những ngoại lệ về giá, nhưng xe chạy diesel có xu hướng có nhiều giá cao hơn, cho



- Giá trung bình theo loại nhiên liệu: Xe chạy bằng diesel có giá trung bình cao hơn so với xe chạy bằng xăng. Điều này phù hợp với kỳ vọng rằng xe diesel thường có giá bán đầu cao hơn do độ bền và hiệu suất nhiên liệu tốt hơn.
- Biến động giá: Biểu đồ không hiển thị khoảng tin cậy, nhưng sự khác biệt rõ ràng về giá trung bình giữa hai loại nhiên liệu cho thấy rằng loại nhiên liệu là một yếu tố quan trọng ảnh hưởng đến giá xe trên thị trường xe đã qua sử dụng.

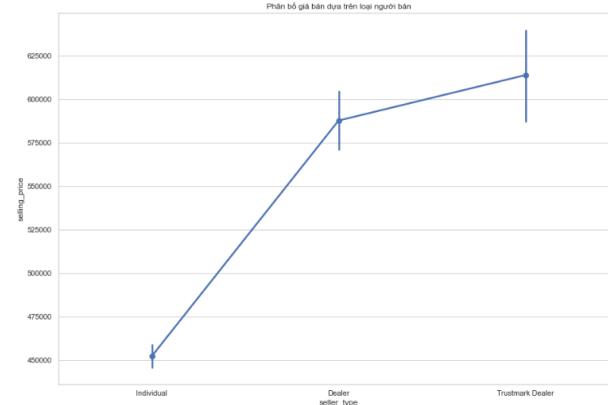
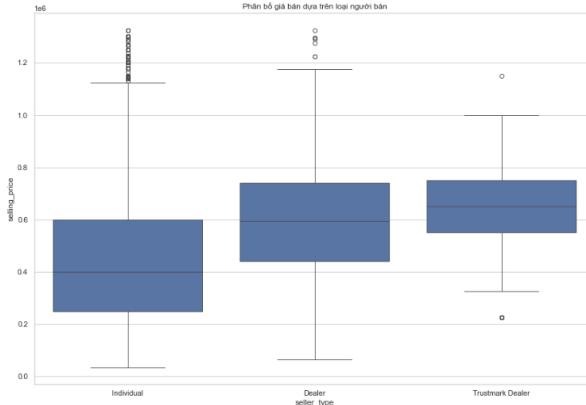
thấy sự phổ biến của một số mẫu xe diesel cao cấp hoặc lớn hơn trên thị trường.

- Giá bán trung bình: Xe diesel có giá bán trung bình cao hơn xe chạy xăng, phản ánh sự ưu chuộng và giá trị bền vững của chúng trên thị trường.

- Xe diesel thường có giá cao hơn xe xăng do tính bền và tuổi thọ dài hơn, phù hợp với việc lái xe thường xuyên và trong điều kiện khắc nghiệt.
- Động cơ diesel hiệu quả về mặt nhiên liệu, đặc biệt khi lái ở tốc độ cao hoặc vận chuyển hàng nặng, giúp tiết kiệm chi phí nhiên liệu đáng kể.
- Mô-men xoắn lớn hơn ở tốc độ thấp làm cho xe diesel lý tưởng cho việc kéo hoặc vận chuyển, đặc biệt phù hợp với xe tải và SUV.
- Xe diesel giữ giá tốt hơn khi bán lại do độ bền và hiệu quả nhiên liệu cao, làm tăng giá trị ban đầu và giá bán lại của chúng.

Những yếu tố này biến xe diesel thành lựa chọn kinh tế hơn trong dài hạn trên thị trường xe đã qua sử dụng.

5.3 Mối quan hệ giữa loại người bán và giá bán



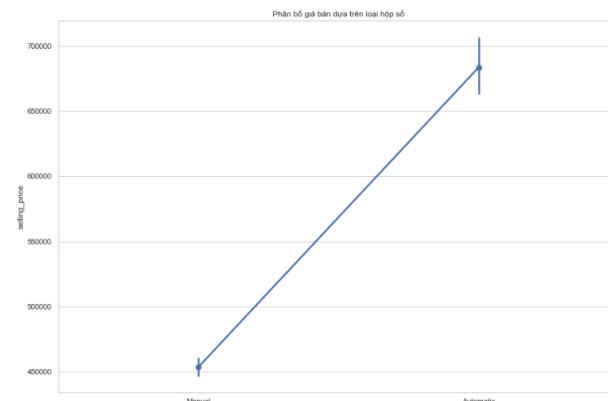
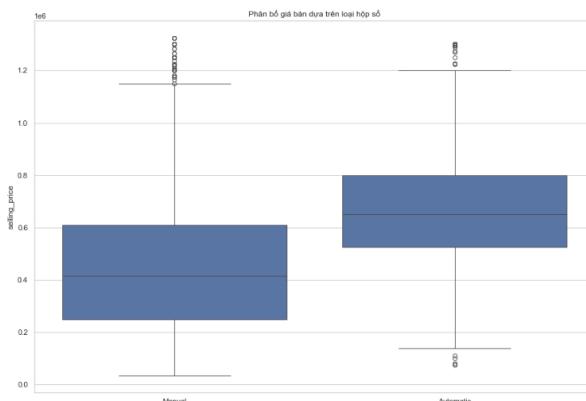
- Biến động giá:** Có sự khác biệt rõ rệt về phạm vi giá giữa các loại người bán. Người bán cá nhân thường có giá bán thấp hơn so với các loại người bán khác như đại lý hoặc người bán chuyên nghiệp. Điều này có thể do người bán cá nhân thường cung cấp giá thấp hơn để thu hút người mua trong một thị trường cạnh tranh.
- Giá cao nhất và thấp nhất:** Các người bán chuyên nghiệp và đại lý có thể có các mẫu xe có giá cao hơn do chất lượng xe tốt hơn hoặc đã qua tân trang, cũng như việc cung cấp bảo hành và các dịch vụ bổ sung.
- Outliers:** Sự tồn tại của outliers trong mỗi loại người bán cho thấy có những xe có giá rất cao hoặc rất thấp so với mức trung bình của loại hình bán hàng đó.

- Giá bán trung bình:** Biểu đồ cho thấy người bán chuyên nghiệp thường có giá bán trung bình cao hơn so với người bán cá nhân. Điều này có thể phản ánh việc các người bán chuyên nghiệp cung cấp xe có chất lượng tốt hơn, đã qua kiểm định nghiêm ngặt hơn, hoặc cung cấp thêm các dịch vụ hậu mãi và bảo hành.

- Sự chênh lệch giá:** Sự chênh lệch giá giữa các loại người bán cũng phản ánh chiến lược giá và đối tượng mục tiêu khác nhau. Người bán cá nhân có thể cung cấp giá thấp hơn để thu hút người mua trong khi các đại lý và người bán chuyên nghiệp có thể nhắm đến phân khúc cao cấp hơn.
- Độ tin cậy và bảo hành:** Người mua có thể sẵn sàng trả giá cao hơn cho các xe được bán bởi người bán chuyên nghiệp do sự đảm bảo về độ tin cậy và các lợi ích bảo hành đi kèm.

Sự chênh lệch giá giữa các loại người bán xe đã qua sử dụng chủ yếu xuất phát từ khác biệt về chất lượng xe, dịch vụ hỗ trợ khách hàng, và mức độ bảo hành mà họ cung cấp.

5.4 Mối quan hệ giữa hộp số và giá bán



- Giá bán trung bình:** Xe có hộp số tự động (Automatic) có giá bán trung bình cao hơn đáng kể so với xe có hộp số sàn (Manual). Điều này phản ánh xu hướng thị trường hiện nay, nơi mà xe hộp số tự động được đánh giá cao vì tính tiện lợi và thoải mái khi lái, đặc biệt trong điều kiện giao thông đông đúc.
- Phạm vi giá và Outliers:** Cả hai loại hộp số đều có các outliers, nhưng xe hộp số tự động có phạm vi giá cao hơn và nhiều outliers ở mức giá cao, cho thấy sự tồn tại của các mẫu xe tự động cao cấp trên thị trường.
- Phản bội giá:** Sự chênh lệch giá giữa hai loại hộp số cũng phản ánh sự chênh lệch về công nghệ và chi phí sản xuất. Xe tự động thường phức tạp hơn và đắt tiền hơn để sản xuất, dẫn đến mức giá bán cao hơn.

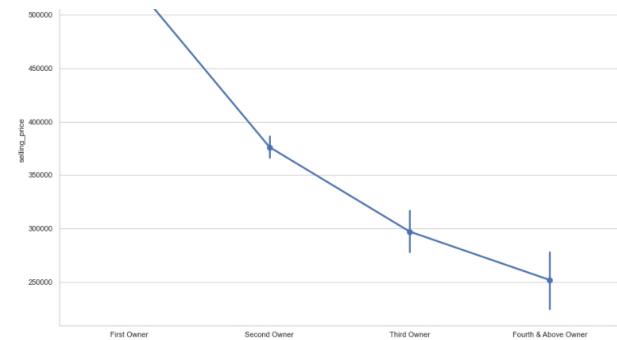
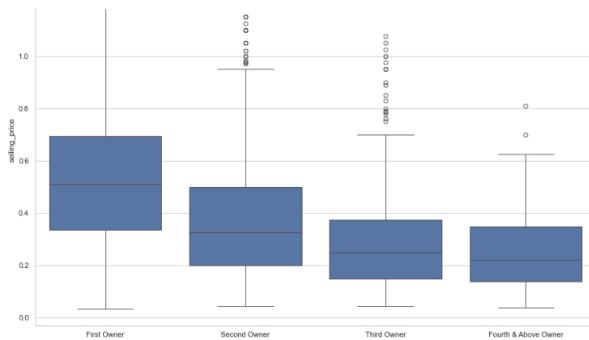
- Giá bán trung bình:** Có sự chênh lệch rõ rệt về giá bán trung bình giữa xe có hộp số tự động và hộp số sàn. Xe hộp số tự động có giá cao hơn nhiều so với xe hộp số sàn, phản ánh xu hướng thị trường và sự ưu chuộng của người tiêu dùng đối với sự tiện lợi và thoải mái mà hộp số tự động mang lại.

- Tính thị trường:** Sự phổ biến của hộp số tự động trong các mẫu xe mới và cao cấp hơn cũng góp phần làm tăng giá trung bình của dòng xe này. Xe tự động thường kết hợp với các tính năng an toàn và thoải mái cao cấp hơn, cũng như được định vị trong phân khúc thị trường cao hơn.
- Tác động đến quyết định mua:** Thông tin này rất hữu ích cho người mua khi cân nhắc giữa chi phí ban đầu cao hơn của hộp số tự động so với lợi ích lâu dài về mặt sự tiện nghi và giá trị bán lại, đặc biệt trong bối cảnh xe tự động ngày càng được ưa chuộng.

Xe có hộp số tự động thường có giá cao hơn so với xe hộp số sàn chủ yếu do chi phí sản xuất cao hơn và các công nghệ tiên tiến hơn được tích hợp trong hệ thống truyền động. Hộp số tự động cung cấp sự tiện lợi và thoải mái cho người lái, đặc biệt trong điều kiện giao thông đông đúc, vì nó loại bỏ nhu cầu phải thay đổi số thường xuyên, làm giảm sự mệt mỏi khi lái xe.

5.5 Mối quan hệ giữa số lượng chủ sở hữu và giá bán

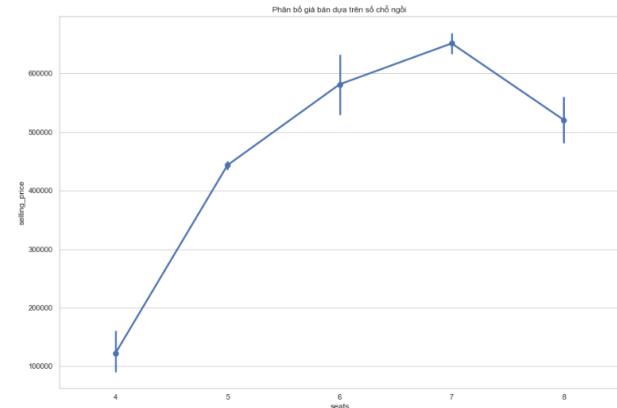
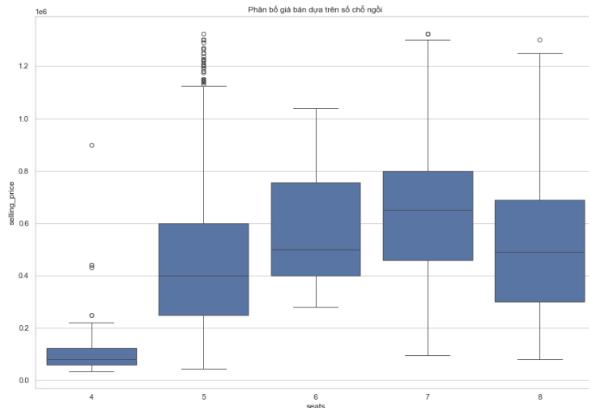




- Giá bán trung bình: Xe thuộc sở hữu của "First Owner" (chủ sở hữu đầu tiên) có giá trung bình cao hơn so với các loại chủ sở hữu khác. Điều này phản ánh sự tin tưởng cao hơn đối với xe có ít chủ sở hữu trước, vì nó thường được giữ gìn tốt hơn và có ít vấn đề hơn.
- Phạm vi giá và Outliers: Xe thuộc sở hữu của "Second Owner" (chủ sở hữu thứ hai) và các loại sở hữu sau đó do thấy phạm vi giá thấp hơn và nhiều outliers giảm dần. Điều này cho thấy giá trị xe giảm theo số lượng chủ sở hữu tăng lên, và các xe này có thể đã trải qua nhiều sửa chữa hoặc bảo trì hơn.
- Giá bán giảm theo sở hữu: Biểu đồ cũng cho thấy rõ ràng rằng giá bán giảm khi xe chuyển từ tay người này sang tay người khác. Xe thuộc sở hữu "Third Owner" và "Fourth & Above Owner" thường có giá thấp hơn đáng kể, phản ánh sự mất giá do sử dụng lâu dài và khả năng đã được bán lại nhiều lần.

Giá bán của xe đã qua sử dụng có mối quan hệ nghịch với số lượng chủ sở hữu mà xe đó đã trải qua.

5.6 Mối quan hệ giữa số chỗ ngồi và giá bán



- Sự khác biệt giá theo số chỗ ngồi: Có sự khác biệt rõ rệt về giá bán giữa các xe có số chỗ ngồi khác nhau. Xe có 5 chỗ ngồi có phạm vi giá rộng và nhiều outliers, phản ánh sự đa dạng của phân khúc này trên thị trường.
- Xe 7 chỗ: Xe có 7 chỗ ngồi thường có giá cao hơn so với xe 5 chỗ, có thể do chúng thường là SUV hoặc minivan, những loại xe thường được trang bị nhiều tính năng hơn và có khả năng vận chuyển tốt hơn.
- Xe 2 và 4 chỗ: Xe có 2 và 4 chỗ ngồi thường có giá thấp hơn, điều này có thể liên quan đến kích thước nhỏ hơn và ít tính năng hơn. Xe 2 chỗ ngồi thường là xe thể thao hoặc xe chuyên dụng, có thể có giá cao hoặc thấp tùy theo mẫu và trang bị.
- Phân khúc và tính năng: Số chỗ ngồi cũng phản ánh phân khúc của xe; xe có nhiều chỗ ngồi thường phục vụ cho nhu cầu gia đình hoặc thương mại, yêu cầu không gian rộng rãi và tính năng an toàn cao, từ đó đẩy giá bán lên cao.

Số lượng chỗ ngồi trong một chiếc xe có mối quan hệ trực tiếp với giá bán của nó, phản ánh nhu cầu và mục đích sử dụng đa dạng của các loại xe. Xe có 7 chỗ ngồi thường có giá cao hơn do chúng phục vụ tốt cho nhu cầu của các gia đình lớn hoặc sử dụng trong kinh doanh, như vận chuyển khách hoặc dịch vụ shuttle, và thường được trang bị nhiều tính năng tiện nghi và an toàn.

CHƯƠNG 6. KIỂM ĐỊNH THỐNG KÊ VÀ TUYÊN BỐ GIÁ THUYẾT (HYPOTHESIS STATEMENT)

6.1 Giả thuyết kiểm định giá bán theo loại nhiên liệu

- Mục đích: Kiểm định sự khác biệt về giá bán trung bình giữa xe sử dụng nhiên liệu diesel và petrol, giúp hiểu rõ hơn về ảnh hưởng của loại nhiên liệu đến giá bán xe.
- Nhận thức vấn đề: Nhận định ban đầu về giả thuyết là xe chạy diesel có giá bán cao hơn xe chạy xăng do các yếu tố như hiệu quả nhiên liệu tốt hơn, độ bền cao hơn. T-test (kiểm định t) được áp dụng để kiểm tra giả thuyết này, có hay không sự khác biệt về giá bán trung bình đáng kể giữa hai nhóm nhiên liệu.
 - Giả thuyết Không (H0): Phản ánh giả định rằng không có sự khác biệt về giá bán giữa hai nhóm, hoặc giá của nhóm diesel không cao hơn nhóm xăng (có khoảng tin cậy là 95% và mức ý nghĩa là 5%).
 - Giả thuyết Nghịch (H1): Có sự khác biệt đáng kể về giá xe giữa các loại nhiên liệu khác nhau, cụ thể là phản ánh rằng giá bán của nhóm diesel cao hơn nhóm xăng.

```
Hypothesis test result
sample size for diesel 3571
sample size for petrol 3248
-----
Hypothesis:
H0 = Used car price with fuel type diesel <= Used car price with fuel type petrol
H1 = Used car price with fuel type diesel > Used car price with fuel type petrol
-----
t-statistic: 25.62565462232831
p-value: 2.752725671450283e-138
alpha : 0.05
-----
Based on t-test, it can be concluded:
Null hypothesis rejected
```

- Kết quả từ kiểm định t cho thấy một chênh lệch đáng kể, với giá trị p-value rất nhỏ ($p < \alpha$), cho thấy sự khác biệt về giá bán giữa hai nhóm là có ý nghĩa thống kê.
- Kết luận: Việc kiểm định thống kê đã cung cấp bằng chứng rõ ràng về mối quan hệ giữa loại nhiên liệu và giá bán xe đã qua sử dụng. Bằng cách bác bỏ giả thuyết null, chúng ta có thể khẳng định rằng xe diesel được bán với giá cao hơn so với xe petrol, điều này có ý nghĩa quan trọng trong việc định hình các quyết định liên quan đến mua bán và sở hữu xe.

6.2 Giả thuyết kiểm định giá bán theo loại hộp số

- Mục đích: Kiểm tra liệu có sự khác biệt về giá bán trung bình giữa xe hộp số tự động và hộp số sàn. Kiểm định này giúp xác định ảnh hưởng của loại hộp số đến giá bán xe, vốn là thông tin quan trọng đối với người mua và người bán trong việc định giá xe.
- Nhận thức vấn đề: Xe hộp số tự động thường có giá cao hơn do tiện ích và sự thoải mái cho người lái. Nhưng để kết luận chính xác cần phải kiểm định bằng thống kê, T-test hỗ trợ kiểm tra giả định này bằng cách so sánh hai nhóm hộp số về giá bán.
- Giả thuyết Không (H0): Giá bán của xe sử dụng hộp số sàn không cao hơn xe sử dụng hộp số tự động (có khoảng tin cậy là 95% và mức ý nghĩa là 5%).
- Giả thuyết Nghịch (H1): Giá bán của xe sử dụng hộp số tự động cao hơn xe sử dụng hộp số sàn.

```
Hypothesis test result
sample size for manual 6258
sample size for automatic 553
-----
Hypothesis:
H0 = Used car price with transmission type manual <= Used car price with type automatic
H1 = Used car price with transmission type automatic > Used car price with transmission type manual
-----
t-statistic: -21.018078899494657
p-value: 2.423223621083974e-75
alpha : 0.05
-----
Based on t-test, it can be concluded:
Reject H0
```

- Kết quả kiểm định này cho thấy có sự khác biệt đáng kể giữa hai loại hộp số, với giá trị p-value thấp, cho thấy sự khác biệt về mặt thống kê.
- Kết luận: Cân cứ vào kết quả của kiểm định t, chúng ta có thể khẳng định rằng xe có hộp số tự động có giá cao hơn đáng kể so với xe có hộp số thủ công. Bằng cách bác bỏ giả thuyết null, chúng ta có thể khẳng định rằng loại hộp số là một yếu tố quan trọng ảnh hưởng đến giá xe cũ.

6.3 Giả thuyết kiểm định giá bán theo loại người bán

- Mục đích: Kiểm định sự khác biệt về giá bán giữa các loại người bán khác nhau (cá nhân, đại lý, đại lý được chứng nhận). Đây cũng là một yếu tố quan trọng ảnh hưởng đến giá xe.
- Nhận thức vấn đề: Giả sử rằng xe bán bởi đại lý được chứng nhận có giá cao hơn do uy tín và dịch vụ hậu mãi. ANOVA (Phân tích phương sai) kiểm tra điều này bằng cách so sánh ba nhóm với nhau về giá bán, phù hợp khi có nhiều hơn hai nhóm cần so sánh.
- Giả thuyết Không (H0): Không có sự khác biệt đáng kể về giá bán trung bình giữa các loại người bán (có khoảng tin cậy là 95% và mức ý nghĩa là 5%).
- Giả thuyết Nghịch (H1): Có sự khác biệt đáng kể về giá bán trung bình giữa ít nhất hai loại người bán.

```
Hypothesis:
H0 = There is no significant difference in the average selling_price between seller_type
H1 = There is a significant difference in the average selling_price between at least two seller_type
-----
p-value: 4.864194801591785e-57
alpha : 0.05
-----
Based on ANOVA test, it can be concluded:
Reject H0
```

- Kết luận: Cân cứ vào kết quả của kiểm định ANOVA, giá trị p-value thấp hơn mức ý nghĩa đã định (alpha = 0.05), điều này cho thấy có sự khác biệt đáng kể về giá bán trung bình giữa các loại người bán.

6.4 Giả thuyết kiểm định giá bán theo số lượng người đã từng sở hữu

- Mục đích: So sánh giá bán giữa các xe với số lượng chủ sở hữu trước khác nhau.
- Nhận thức vấn đề: Các xe có ít chủ sở hữu trước thường giá tốt hơn. ANOVA giúp kiểm định sự khác biệt về giá bán tổng cộng giữa các nhóm xe dựa vào số lượng chủ sở hữu trước đó, thích hợp khi so sánh nhiều hơn hai nhóm.
- Giả thuyết Không (H0): Không có sự khác biệt đáng kể về giá bán trung bình giữa các loại chủ sở hữu. (có khoảng tin cậy là 95% và mức ý nghĩa là 5%).
- Giả thuyết Nghịch (H1): Có sự khác biệt đáng kể về giá bán trung bình giữa ít nhất hai loại chủ sở hữu.

```
Hypothesis:
H0 = There is no significant difference in the average selling_price between owner_type
H1 = There is a significant difference in the average selling_price between at least two owner_type
-----
p-value: 4.009711536756785e-185
alpha : 0.05
-----
Based on ANOVA test, it can be concluded:
Reject H0
```

- Kết luận: Cân cứ vào kết quả của kiểm định, giá trị p-value thấp hơn mức ý nghĩa đã định (alpha = 0.05), điều này cho thấy có sự khác biệt đáng kể về giá bán trung bình giữa các nhóm chủ sở hữu khác nhau.

CHƯƠNG 7: XÂY DỰNG MÔ HÌNH HỒI QUY VÀ DỰ BÁO GIÁ

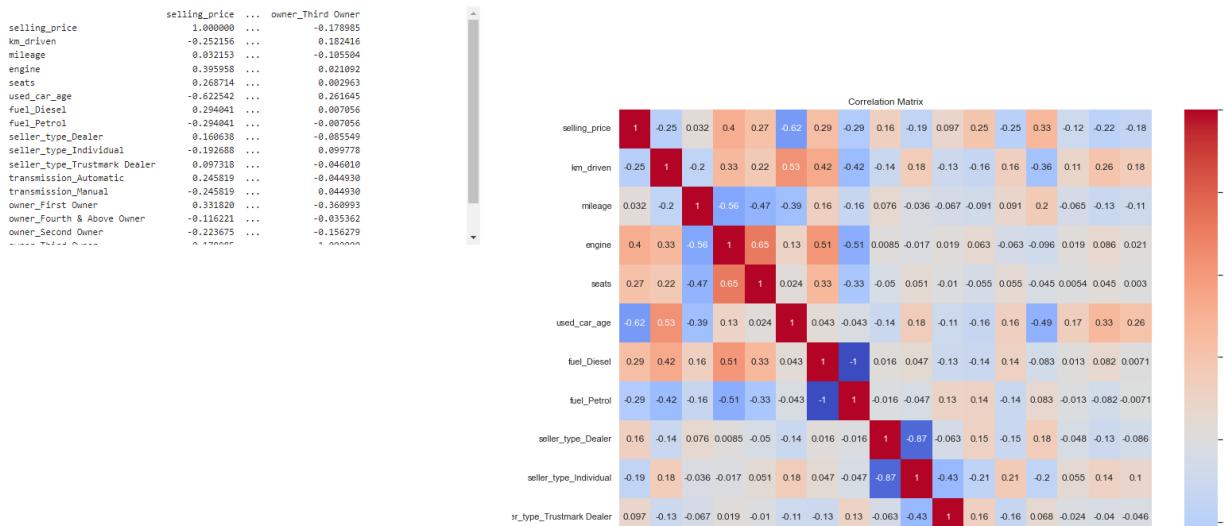
7.1 Phân tích mối tương quan giữa các biến

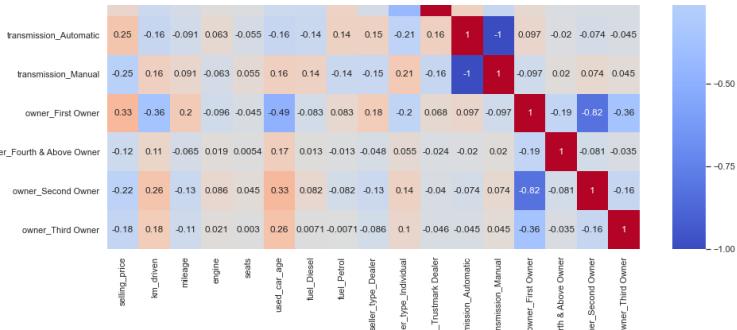
Mã trận tương quan đã được phân tích để xác định mối liên hệ giữa các biến số. Các biến có tương quan mạnh mẽ với giá xe như engine, used_car_age, fuel_Diesel, và transmission_Automatic được lựa chọn cho mô hình hồi quy vì chúng có hệ số tương quan rõ rệt với giá xe:

- engine với giá xe (0.4): Xác định rằng xe có dung tích động cơ lớn hơn thường có giá cao hơn.
- used_car_age với giá xe (-0.62): Giá xe giảm đáng kể theo thời gian sử dụng.
- fuel_Diesel (0.29): Xe chạy dầu diesel có giá bán cao hơn các loại nhiên liệu khác.

	selling_price	km_driven	...	owner_Second	owner_Third
Owner	0	450000	145500	...	0
	0	370000	120000	...	1
	1	158000	140000	...	0
	2	225000	127000	...	0
	3	130000	120000	...	0
	4	130000	120000	...	0

[5 rows x 17 columns]





7.2 Loại bỏ biến không cần thiết

- Dựa trên ma trận tương quan, các biến seats, km_driven, và mileage được loại bỏ khỏi phân tích do tương quan thấp với giá xe (dưới 0.2)

7.3 Xây dựng mô hình hồi quy tuyến tính

	selling_price	engine	transmission_Manual	owner_First Owner
0	4500000	1248	...	1
1	3700000	1498	...	1
2	1580000	1497	...	1
3	2250000	1396	...	1
4	1300000	1298	...	1

[5 rows x 8 columns]

7.4 Kết quả từ mô hình hồi quy OLS

	used_car_age	fuel_Diesel	fuel_Petrol	transmission_Automatic	transmission_Manual	owner_First Owner
used_car_age	-4.725e-04	634.006	-74.519	-4.85e-04	-4.6e-04	
fuel_Diesel	1.949e+00	402.164	46.381	0.000	1.87e+05	2.03e+05
fuel_Petrol	1.239e-05	3883.300	472	0.000	1.12e+05	1.3e+05
transmission_Automatic	2.223e-05	5957.237	45.849	0.000	2.12e+05	2.32e+05
transmission_Manual	9.649e-04	3735.168	25.832	0.000	8.92e-04	1.04e+05
owner_First Owner	3.477e-04	4465.364	7.786	0.000	3.6e-04	4.35e-04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 7.64e-30. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

7.5 Đánh giá và xác nhận mô hình

Mô hình được đánh giá thông qua:

- R-squared: 0.644, chỉ ra rằng mô hình có thể giải thích 64.4% sự biến thiên trong giá bán xe dựa trên các biến đã chọn. Đây là một chỉ số tốt, cho thấy mô hình có khả năng dự đoán giá xe một cách hiệu quả.
- F-statistic: 2465, với một giá trị p gần như bằng 0, chứng tỏ mô hình có sự khác biệt thống kê đáng kể so với một mô hình không có các biến giải thích này.
- Coefficients và P-values: Tất cả các biến độc lập đều có P-value nhỏ hơn 0.05, cho thấy mỗi biến đều có ảnh hưởng đáng kể và ý nghĩa thống kê đến giá xe.

Kết Luận

- Mô hình hồi quy đã phát triển thành công và cho phép dự đoán giá xe một cách chính xác, phục vụ cho các quyết định mua bán và đầu tư xe cũ.

CHƯƠNG 8: SO SÁNH TỔNG QUÁT VÀ KẾT LUẬN CHO TOÀN BỘ BÀI TOÁN

8.1 Kết luận và khuyến nghị

Những kết luận chính

- Các yếu tố ảnh hưởng đến giá: Phân tích xác nhận rằng công suất động cơ, loại hộp số, loại nhiên liệu, tuổi xe, số kilomet đã đi, uy tín thương hiệu và loại xe là những yếu tố quan trọng trong việc xác định giá xe cũ.
- Xu hướng thị trường: Có sự chuyển dịch rõ rệt về hộp số tự động và các xe có tính năng an toàn tiên tiến, thường có giá trị bán lại cao hơn. Ngoài ra, sự quan tâm ngày càng tăng của người tiêu dùng đối với xe thân thiện với môi trường cũng ảnh hưởng đến nhu cầu thị trường và sự ổn định giá.
- Biến động khu vực: Giá cả cũng có sự khác biệt đáng kể giữa các khu vực khác nhau, phản ánh điều kiện kinh tế địa phương, sự phổ biến của một số loại xe và sự khác biệt trong sở thích của người tiêu dùng.

Khuyến nghị

- Đối với người mua: Người mua nên xem xét chi phí sở hữu tổng thể, bao gồm khâu hao, bảo hiểm và bảo trì, ngoài giá mua ban đầu. Mua các mẫu xe cũ hơn một chút hoặc xe hết hạn cho thuê có thể cung cấp giá trị tốt hơn do giá khâu hao ban đầu cao.
- Đối với người bán: Người bán nên tập trung vào những chiếc xe được bảo dưỡng tốt và có số kilomet thấp để tối đa hóa giá trị bán lại. Cung cấp lịch sử dịch vụ chi tiết và tiến hành kiểm tra trước khi bán cũng có thể tăng lòng tin của người mua và có khả năng tăng giá bán.
- Đối với nhà phân tích thị trường: Việc tiếp tục theo dõi các xu hướng mới nổi như sự chuyển dịch về xe điện và thay đổi hành vi tiêu dùng sau đại dịch là rất quan trọng. Những yếu tố này có khả năng ảnh hưởng đáng kể đến giá trị xe trong tương lai và động thái của thị trường.

8.2 Các vấn đề tiềm tàng và hướng đi trong tương lai

Các vấn đề tiềm tàng

- Hạn chế về dữ liệu: Các mô hình hiện tại phụ thuộc nhiều vào dữ liệu lịch sử, có thể không đầy đủ để nắm bắt những thay đổi nhanh chóng về công nghệ hoặc sở thích của người tiêu dùng. Tốc độ chấp nhận công nghệ mới như các tính năng lái xe tự động có thể thay đổi mô hình già một cách đáng kể.
- Biến động thị trường: Sự dao động kinh tế và các yếu tố bên ngoài như thay đổi chính sách của chính phủ liên quan đến khí thải xe có thể dẫn đến những thay đổi đột ngột trên thị trường, ảnh hưởng đến độ chính xác dự đoán của các mô hình hiện tại.

Hướng đi trong tương lai

- Tích hợp dữ liệu thời gian thực: Các nghiên cứu tương lai nên bao gồm nguồn dữ liệu động, thời gian thực để nắm bắt ngay lập tức các tác động của thay đổi thị trường và tiến bộ công nghệ đến giá xe cũ.
- Công cụ phân tích nâng cao: Việc sử dụng các mô hình học máy phức tạp hơn có thể cung cấp cái nhìn sâu sắc hơn và dự đoán chính xác hơn bằng cách xử lý các mối quan hệ phi tuyến tính phức tạp và tương tác giữa một tập hợp lớn các biến.

- Yếu tố bền vững: Khi các yếu tố về môi trường ngày càng trở nên quan trọng, việc nghiên cứu thêm về cách thức bền vững ảnh hưởng đến sự lựa chọn của người tiêu dùng và giá trong thị trường xe cũ sẽ rất cần thiết.
- Phân tích thị trường toàn cầu: Mở rộng phân tích để bao gồm xu hướng thị trường toàn cầu và so sánh các khu vực địa lý khác nhau có thể tiết lộ các cơ hội và thách thức độc đáo trong thị trường xe cũ quốc tế.

Tài liệu tham khảo

1	CS 229 Project Report: Predicting Used Car Prices n.d., Stanford https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf
2	Ferizqa, D. (2023, July 12) Statistical analysis: Used car price. Medium https://medium.com/@dsyafz/statistical-analysis-used-car-price-132b073439d5
3	Lau, S., Gonzalez, J., & Nolan, D. (2023) Learning data science. O'Reilly Media
4	Marcus Collard. (n.d.) Price Prediction for Used Cars https://www.diva-portal.org/smash/get/diva2:1674070/FULLTEXT01.pdf
5	Sekseria, A. (2020, January 14) Exploratory data analysis on used cars dataset. Medium https://medium.com/@sekseriaankit0657/exploratory-data-analysis-on-used-cars-dataset-4ddacd58cdb6
6	Fischer, J. (2024, May 2) Used car price trends for 2024 (Updated weekly). CarEdge https://caredge.com/guides/used-car-price-trends-for-2024

Boost was developed by Le Minh Tan