Ảnh có chứa quảng trường

Mô tả được tạo tự động**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

🙠🙟🕮🙝🙢

A logo of hands holding a book and a candle

Description automatically generated

**BÁO CÁO CUỐI KÌ**

**MÔN: 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*

*Bùi Quốc Khang 21110202*

*Lê Việt Khánh 21110206*

*Đặng Kim Thành 21110298*

*Thành phố Hồ Chí Minh, tháng 3 năm 2024*

**DANH SÁCH THÀNH VIÊN THAM GIA ĐỒ ÁN**

***Đề tài:***  Phân tích giá xe cũ để xác định xu hướng thị trường

*----------*

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **HỌ VÀ TÊN THÀNH VIÊN** | **MÃ SỐ SINH VIÊN** | **TỶ LỆ THAM GIA** |
| 1 | Bùi Quốc Khang | 21110202 | 100% |
| 2 | Lê Việt Khánh | 21110206 | 100% |
| 3 | Đặng Kim Thành | 21110298 | 100% |

**Ghi chú:**

Tỷ lệ %: Mức độ phần trăm hoàn thành của từng sinh viên tham gia.

***Trưởng nhóm****:* Bùi Quốc Khang

***Nhận xét của giảng viên***

.............................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

Tp. Hồ Chí Minh - Tháng 3 năm 2024

# LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn sâu sắc đến thầy Lê Minh Tân, thầy đã tận tình giảng dạy trong suốt quá trình tìm hiểu và học tập môn Nhập Môn Khoa Học Dữ Liệu. Trong từng buổi học, thầy đã nhiệt huyết trong từng lời giải đáp, những kiến thức, kỹ năng, tài liệu quý giá mà thầy đã truyền đạt học không chỉ là nền tảng cho quá trình thực hiện bài tập cuối kì mà còn là kiến thức nền tảng cho nhiều phần sau này.

Rất nhiều công sức và nỗ lực đã được bỏ ra, nhưng do chưa có kinh nghiệm trong việc xây dựng đề tài và những hạn chế về kiến thức, bài báo cáo này không tránh khỏi sai sót. Chúng em luôn sẵn sàng tiếp thu những ý kiến từ thầy để cải thiện bài báo cáo cũng như đề tài lần này.

Một lần nữa, em xin bày tỏ lòng biết ơn đến thầy. Xin kính chúc thầy luôn dồi dào sức khỏe, đạt được nhiều thành công trong công việc.

***Đại diện nhóm Bùi Quốc Khang***

# MỤC LỤC

[LỜI CẢM ƠN ii](#_Toc165723082)

[MỤC LỤC iii](#_Toc165723083)

[MỤC LỤC HÌNH ẢNH v](#_Toc165723084)

[MỤC LỤC BẢNG vii](#_Toc165723085)

[CHƯƠNG 1. MỞ ĐẦU 1](#_Toc165723086)

[1.1 Phát biểu về bài toán 1](#_Toc165723087)

[1.2 Mục đích thực hiện 1](#_Toc165723088)

[1.3 Tổng quan về dữ liệu của bài toán 1](#_Toc165723089)

[1.4 Các yếu tố ảnh hưởng đến độ hữu ích của dữ liệu 1](#_Toc165723090)

[CHƯƠNG 2. DATASETS 2](#_Toc165723091)

[2.1 Dataset “Used Cars Price Prediction” (DatasetMixed) 2](#_Toc165723092)

[2.1.1 Tổng Quan Về Tập Dữ Liệu 2](#_Toc165723093)

[2.1.2 Khám Phá Tổng Quát và Định Hướng Phân Tích 2](#_Toc165723094)

[2.1.3 Phân tích thống kê định lượng 4](#_Toc165723095)

[2.1.4 Nghiên cứu sâu về mẫu dữ liệu đặc trưng 5](#_Toc165723096)

[2.2 Dataset “Used Cars data form websites” (DatasetIndiaWebsite) 6](#_Toc165723097)

[2.2.1 Tổng Quan Về Tập Dữ Liệu 6](#_Toc165723098)

[2.2.2 Khám Phá Tổng Quát và Định Hướng Phân Tích 6](#_Toc165723099)

[2.2.3 Phân tích thống kê định lượng 10](#_Toc165723100)

[2.2.4 Nghiên cứu sâu về mẫu dữ liệu đặc trưng 11](#_Toc165723101)

[2.3 Dataset “Used Cars Prices in UK” (DatasetUK) 14](#_Toc165723102)

[2.3.1 Tổng Quan Về Tập Dữ Liệu 14](#_Toc165723103)

[2.3.2 Khám Phá Tổng Quát và Định Hướng Phân Tích 14](#_Toc165723104)

[2.3.3 Phân tích thống kê định lượng 17](#_Toc165723105)

[2.3.4 Nghiên cứu sâu về mẫu dữ liệu đặc trưng 18](#_Toc165723106)

[CHƯƠNG 3. SO SÁNH DỮ LIỆU VÀ CHỌN TẬP DATASET CHÍNH 21](#_Toc165723107)

[3.1 So sánh tổng quát 21](#_Toc165723108)

[3.2 Đi vào phân tích và đưa ra kết luận về tập dataset chính 22](#_Toc165723109)

[CHƯƠNG 4. TIỀN XỬ LÝ VÀ CHUẨN HÓA DỮ LIỆU 25](#_Toc165723110)

[4.1 Các bước phân tích và nhận xét về từng đặc trưng 25](#_Toc165723111)

[4.2 Loại bỏ các đặc trưng không sử dụng và giá trị ngoại lai 25](#_Toc165723112)

[CHƯƠNG 5: KHAI PHÁ DỮ LIỆU 28](#_Toc165723113)

[5.1 Mối quan hệ giữa tuổi xe và giá bán 28](#_Toc165723114)

[5.2 Mối quan hệ giữa loại nhiên liệu và giá bán 30](#_Toc165723115)

[5.3 Mối quan hệ giữa loại người bán và giá bán 32](#_Toc165723116)

[5.4 Mối quan hệ giữa hộp số và giá bán 34](#_Toc165723117)

[5.5 Mối quan hệ giữa số lượng chủ sở hữu và giá bán 36](#_Toc165723118)

[5.6 Mối quan hệ giữa số chỗ ngồi và giá bán 38](#_Toc165723119)

[CHƯƠNG 6. KIỂM ĐỊNH THỐNG KÊ VÀ TUYÊN BỐ GIẢ THUYẾT (HYPOTHESIS STATEMENT) 40](#_Toc165723120)

[6.1 Giả thuyết kiểm định giá bán theo loại nhiên liệu 40](#_Toc165723121)

[6.2 Giả thuyết kiểm định giá bán theo loại hộp số 41](#_Toc165723122)

[6.3 Giả thuyết kiểm định giá bán theo loại người bán 42](#_Toc165723123)

[6.4 Giả thuyết kiểm định giá bán theo số lượng người đã từng sở hữu 43](#_Toc165723124)

[CHƯƠNG 7: XÂY DỰNG MÔ HÌNH HỒI QUY VÀ DỰ BÁO GIÁ 45](#_Toc165723125)

[7.1 Phân tích mối tương quan giữa các biến 45](#_Toc165723126)

[7.2 Loại bỏ biến không cần thiết 46](#_Toc165723127)

[7.3 Xây dựng mô hình hồi quy tuyến tính 46](#_Toc165723128)

[7.4 Kết quả từ mô hình hồi quy OLS 47](#_Toc165723129)

[7.5 Đánh giá và xác nhận mô hình 47](#_Toc165723130)

[CHƯƠNG 8: SO SÁNH TỔNG QUÁT VÀ KẾT LUẬN CHO TOÀN BỘ BÀI TOÁN 49](#_Toc165723131)

[8.1 Kết luận và khuyến nghị 49](#_Toc165723132)

[8.2 Các vấn đề tiềm tàng và hướng đi trong tương lai 49](#_Toc165723133)

[CHƯƠNG 9: HÌNH ẢNH BÁO CÁO TRONG ZEPPELIN 51](#_Toc165723134)

[Tài liệu tham khảo 62](#_Toc165723135)

# MỤC LỤC HÌNH ẢNH

[Hình 2.1 Các biểu đồ phân phối theo năm sản xuất, số km đã đi, giá xe 3](#_Toc165720378)

[Hình 2.2 Biểu đồ phân tích trực quan hoá sự chênh lệch giữa các nhóm xe dựa trên loại nhiên liệu về giá trung bình, quãng đường đã đi trung bình, và các đặc trưng khác 5](#_Toc165720379)

[Hình 2.3 Biểu đồ phân phối giá 8](#_Toc165720380)

[Hình 2.4 Biểu đồ phân phối quãng đường đã đi 9](#_Toc165720381)

[Hình 2.5 Biểu đồ phân phối năm đăng ký 9](#_Toc165720382)

[Hình 2.6 Biểu đồ số lượng loại nhiên liệu xe 10](#_Toc165720383)

[Hình 2.7 Biểu đồ Hộp của Giá Bán Theo Loại Hộp Số 12](#_Toc165720384)

[Hình 2.8 Biểu đồ phân phối giá 15](#_Toc165720385)

[Hình 2.9 Biểu đồ phân phối quãng đường đã đi 16](#_Toc165720386)

[Hình 2.10 Biểu đồ phân phối năm đăng ký 16](#_Toc165720387)

[Hình 2.11 Biểu đồ số lượng loại nhiên liệu xe 16](#_Toc165720388)

[Hình 2.12 Biểu đồ giá trung bình và quãng đường đã đi trung bình giữa 2 nhóm xe 18](#_Toc165720389)

[Hình 2.13 Biểu đồ phân tích trực quan hoá sự chênh lệch giữa các nhóm xe dựa trên loại nhiên liệu về giá trung bình, quãng đường đã đi trung bình, và các đặc trưng khác 19](#_Toc165720390)

[Hình 3.1 Biểu đồ tổng số mẫu và phần trăm dữ liệu null của từng tập dataset 22](#_Toc165720391)

[Hình 3.2 Biểu đồ phân bố của năm sản xuất xe của từng tập dataset 23](#_Toc165720392)

[Hình 4.1 Biểu đồ EDA của tập dữ liệu “DatasetIndiaWebsite” 26](#_Toc165720393)

[Hình 4.2 Biểu đồ EDA sau khi xử lý các giá trị ngoại lai của tập dữ liệu “DatasetIndiaWebsite” 27](#_Toc165720394)

[Hình 5.1 Biểu đồ Box phân bố giá bán dựa trên tuổi xe 28](#_Toc165720395)

[Hình 5.2 Biểu đồ Point phân bố giá bán dựa trên tuổi xe 29](#_Toc165720396)

[Hình 5.3 Biểu đồ Box phân bố giá bán dựa trên loại nhiên liệu 30](#_Toc165720397)

[Hình 5.4 Biểu đồ Point phân bố giá bán dựa trên loại nhiên liệu 31](#_Toc165720398)

[Hình 5.5 Biểu đồ Box phân bố giá dựa trên loại người bán 32](#_Toc165720399)

[Hình 5.6 Biểu đồ Point phân bố giá dựa trên loại người bán 33](#_Toc165720400)

[Hình 5.7 Biểu đồ Box phân bố giá dựa trên loại hộp số 34](#_Toc165720401)

[Hình 5.8 Biểu đồ Point phân bố giá dựa trên loại hộp số 35](#_Toc165720402)

[Hình 5.9 Biểu đồ Box phân bố giá dựa trên số lượng chủ đã sở hữu xe 36](#_Toc165720403)

[Hình 5.10 Biểu đồ Point phân bố giá dựa trên số lượng chủ đã sở hữu xe 37](#_Toc165720404)

[Hình 5.11 Biểu đồ Box phân bố giá dựa trên số chỗ ngồi 38](#_Toc165720405)

[Hình 5.12 Biểu đồ Point phân bố giá dựa trên số chỗ ngồi 39](#_Toc165720406)

[Hình 6.1 Kết quả từ kiểm định giá bán theo loại nhiên liệu 41](#_Toc165720407)

[Hình 6.2 Kết quả từ kiểm định giá bán theo loại hộp số 42](#_Toc165720408)

[Hình 6.3 Kết quả từ kiểm định giá bán theo loại người bán 43](#_Toc165720409)

[Hình 6.4 Kết quả từ kiểm định giá bán theo số lượng người đã từng sở hữu 44](#_Toc165720410)

[Hình 7.1 Ma trận tương quan 45](#_Toc165720411)

[Hình 7.2 Kết quả sau khi bỏ các cột có hệ số tương quan thấp 46](#_Toc165720412)

[Hình 7.3 Kết quả mô hình hồi quy OLS 47](#_Toc165720413)

# MỤC LỤC BẢNG

[Bảng 2.1 Thống kê số liệu căn bản của tập “Used Cars Price Prediction” 9](#_Toc165579437)

[Bảng 2.2 Thống kê định lượng của tập “Used Cars Price Prediction” 10](#_Toc165579438)

[Bảng 2.3 Số liệu thống kê căn bản của tập “Used Cars data form websites” 14](#_Toc165579439)

[Bảng 2.4 Thống kê định lượng của tập “Used Cars data form websites” 17](#_Toc165579440)

[Bảng 2.5 số liệu thống kê căn bản của tập “Used Cars Prices in UK” 21](#_Toc165579441)

[Bảng 2.6 Thống kê định lượng của tập “Used Cars Prices in UK” 23](#_Toc165579442)

# CHƯƠNG 1. MỞ ĐẦU

## Phát biểu về bài toán

“**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 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 đối với 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

Dự án có hai mục tiêu chính. Mục tiêu đầu tiên là khám phá những yếu tố ảnh hưởng đến giá xe ô tô, bao gồm sự biến động của cung cầu, vị trí địa lý, năm sản xuất, số đời chủ và nhiều yếu tố khác. Mục tiêu thứ hai là đưa ra các kết luận về định hướng thị trường, giúp các nhà đầu tư, nhà sản xuất và nhà phân phối đưa ra quyết định kinh doanh thông minh trong bối cảnh thị trường đang biến động.

## Tổng quan về dữ liệu của bài toán

Đối tượng nghiên cứu (P) bao gồm các mô hình và hãng xe đa dạng để tạo thành đơn vị nghiên cứu. Các yếu tố tiếp cận để đo lường và quan sát (A) bao gồm các dữ liệu như giá cả lịch sử, thông số kỹ thuật và nhu cầu thị trường. (S) là một phần nhỏ và đại diện cho khung truy cập (A), được sử dụng để quan sát và đo lường, có thể là một tập hợp các xe ô tô cụ thể, bao gồm thông số kỹ thuật, giá cả lịch sử và các yếu tố khác có thể ảnh hưởng đến giá xe.

## Các yếu tố ảnh hưởng đến độ hữu ích của dữ liệu

Những yếu tố ảnh hưởng đến độ hữu ích của dữ liệu có thể nói đến như tính ***chính xác***. Độ chính xác của dữ liệu là yếu tố quan trọng để có được thông tin đáng tin cậy. Trong ngành công nghiệp biến động như ô tô, ***thời gian*** cũng đóng một vai trò quan trọng, và việc sử dụng dữ liệu mới sẽ mang lại hiệu quả tốt hơn. ***Tính liên quan*** của dữ liệu cũng là yếu tố quan trọng, vì dữ liệu không liên quan có thể gây khó khăn trong quá trình phân tích và xử lý.

# 

# CHƯƠNG 2. DATASETS

## 2.1 Dataset “Used Cars Price Prediction” (DatasetMixed)

### 2.1.1 Tổng Quan Về Tập Dữ Liệu

* Số lượng trường : 13 fields
* Số lượng bản ghi : 6020 records
* Tác giả : [Co-learning Lounge | Expert | Kaggle](https://www.kaggle.com/colearninglounge)
* 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 Price Prediction (kaggle.com)](https://www.kaggle.com/datasets/colearninglounge/used-cars-price-prediction/data)

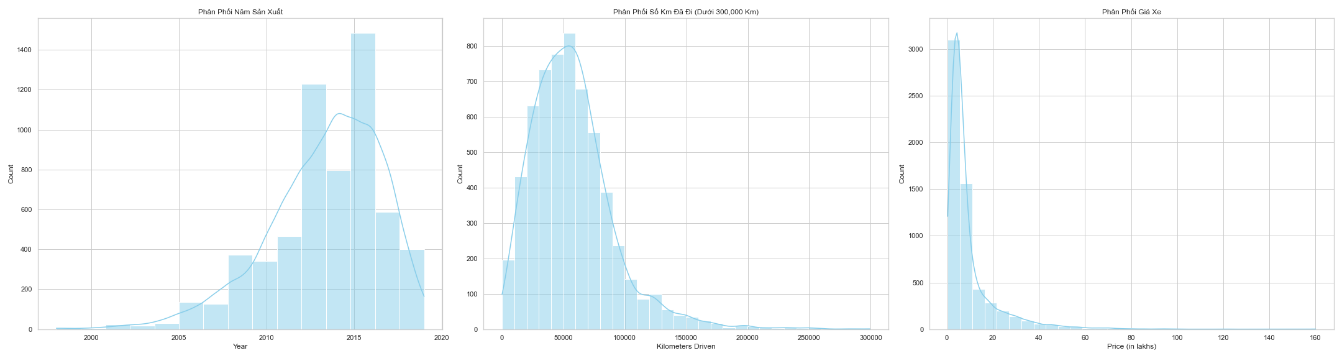
### 2.1.2 Khám Phá Tổng Quát và Định Hướng Phân Tích

* **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 đi.
  + 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)
* **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.

Thực hiện một số thống kê mô tả để có cái nhìn tổng quát về dữ liệu, bao gồm việc tính toán các số liệu thống kê cơ bản như trung bình, phương sai, và vẽ các biểu đồ phân bố giá, quãng đường đã đi, năm đăng ký và số lượng xe theo loại nhiên liệu. Điều này giúp xác định định hướng và mục tiêu cho dữ liệu.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Year | Kilometers\_Driven | Seats | Price |
| count | 6,019 | 6,019 | 5,977 | 6,019 |
| mean | 2013.36 | 58,738.38 | 5.28 | 9.48 |
| std | 3.27 | 91,268.84 | 0.81 | 11.19 |
| min | 1998 | 171 | 0 | 0.44 |
| 25% | 2011 | 34,000 | 5 | 3.50 |
| 50% | 2014 | 53,000 | 5 | 5.64 |
| 75% | 2016 | 73,000 | 5 | 9.95 |
| max | 2016 | 6,500,000 | 10 | 160 |

##### Bảng 2.1 Thống kê số liệu căn bản của tập “Used Cars Price Prediction”



#### Hình 2.1 Các biểu đồ phân phối theo năm sản xuất, số km đã đi, giá xe

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 Đi:** Đa số xe có số km đã đi dưới 300,000 km, với đỉnh 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.

### 2.1.3 Phân tích thống kê định lượng

Các tiêu chí thống kê mô tả cần được xem xét bao gồm:

* + Giá trị trung bình (Mean)
  + Giá trị nhỏ nhất (Min)
  + Giá trị lớn nhất (Max)
  + Trung vị (Median)
  + Phương sai (Variance)
  + Độ nhọn (Kurtosis)
  + Độ xiên (Skewness)

Thực hiện tính toán với các cột Price, Kilometers\_Driven, Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | Price | Kilometers\_Driven | Year |
| Mean | 9.48 | 58,738 | 2013.36 |
| Min | 0.44 | 171 | 1998 |
| Max | 160 | 6,500,000 | 2019 |
| Median | 5.64 | 53,000 | 2014 |
| Variance | 125.17 | 8,330,002,000 | 10.69 |
| Kurtosis | 3.34 | 58.72 | -0.85 |
| Skewness | 17.09 | 4,125 | 0.89 |

##### Bảng 2.2 Thống kê định lượng của tập “Used Cars Price Prediction”

**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 xiên 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 xiên và độ nhọn cực kỳ 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.

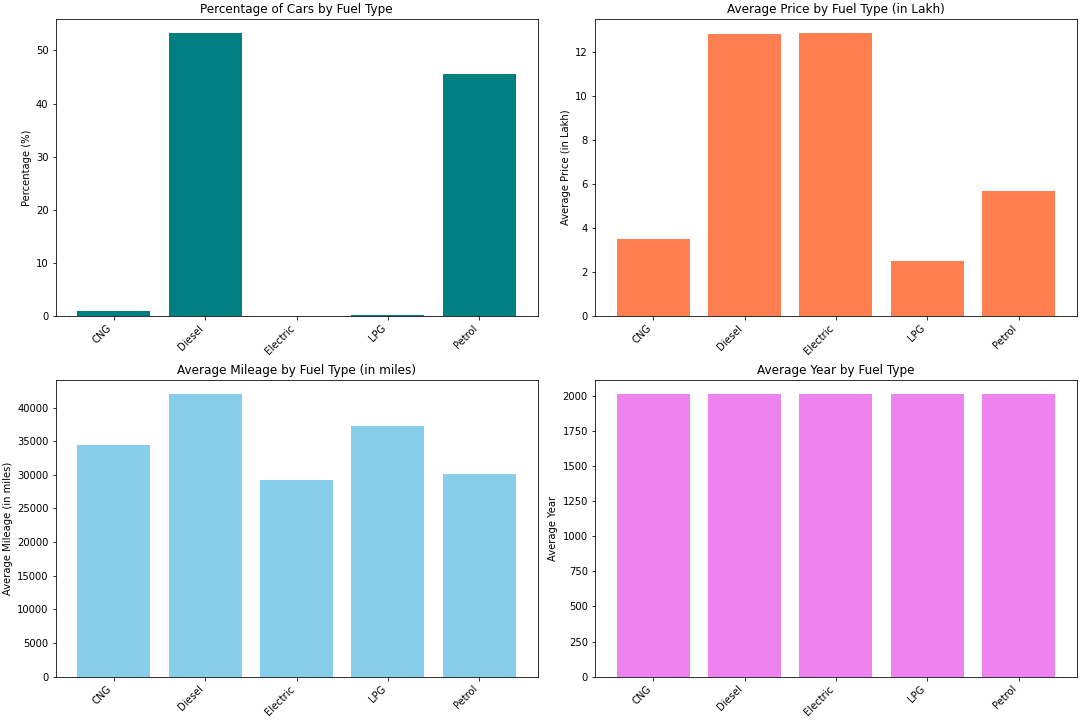
**Year (Năm sản xuất):**

* 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 xiên 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.

### 2.1.4 Nghiên cứu sâu về mẫu dữ liệu đặc trưng

**Chọn Lọc và Lý Giải Mẫu Dữ Liệu**

Chúng ta sẽ tập trung vào việc phân tích dữ liệu dựa trên “Transmission” (hộp số xe), là một trong những đặc trưng quan trọng ảnh hưởng đến quyết định mua xe của người tiêu dùng. Hộp số có thể là “Manual” hoặc “Automatic”. Chọn lọc ra các mẫu có số lượng lớn (> 2000 mẫu) từ tập dữ liệu ban đầu, để đảm bảo tính đại diện và có đủ dữ liệu cho việc phân tích.



#### Hình 2.2 Biểu đồ phân tích trực quan hoá sự chênh lệch giữa các nhóm xe dựa trên loại nhiên liệu về giá trung bình, quãng đường đã đi trung bình, và các đặc trưng khác

Như biểu đồ trên cho thấy:

* **Tỷ Lệ Phần Trăm 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 nhiêu 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)

### 2.2.1 Tổng Quan Về Tập Dữ Liệu

* Số lượng trường : 13 fields
* Số lượng bản ghi : 8128 records
* Tác giả : [Nikhil Kushwaha](https://www.kaggle.com/kushbun27) & [Nishant Verma](https://www.kaggle.com/nishantverma02)
* Mô tả : Tập dữ liệu “Used Cars data form websites” cung cấp thông tin chi tiết về các xe ô tô đã qua sử dụng được bán tại Ấn Độ. Dữ liệu bao gồm 8128 xe với 13 thuộc tính như tên xe, năm sản xuất, giá bán, số km đã đi, loại nhiên liệu, loại người bán, hộp số, thông tin chủ sở hữu, mức tiêu thụ nhiên liệu, dung tích động cơ, công suất tối đa, mô-men xoắn, và số chỗ ngồi.
* Link to Dataset : [Used Cars data form websites](https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho)

### 2.2.2 Khám Phá Tổng Quát và Định Hướng Phân Tích

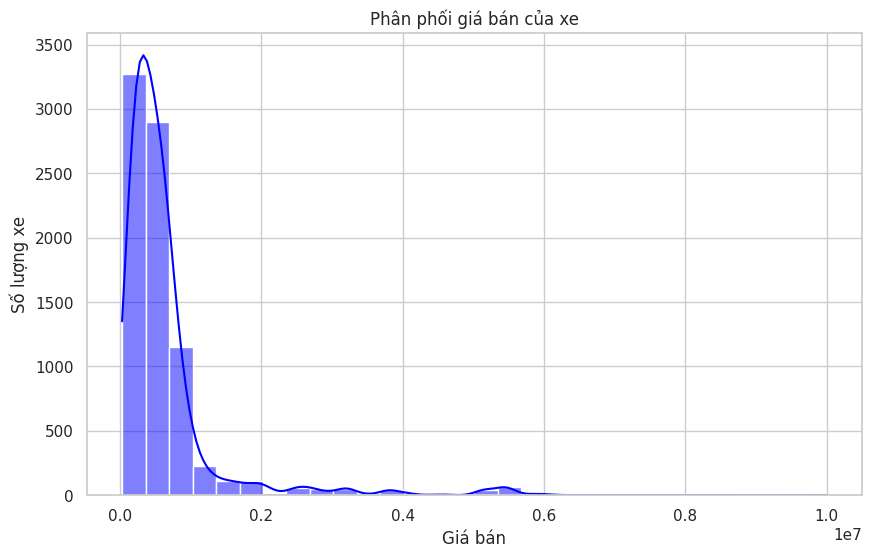
* **Cấu trúc và loại thông tin :**
  + name: Tên của xe (ví dụ: Maruti Swift Dzire VDI).
  + year: Năm sản xuất của xe.
  + selling\_price: Giá bán của xe (đơn vị tiền tệ không được chỉ định, giả sử là INR, vì tập dataset được lấy ở Ấn Độ).
  + km\_driven: Số kilômét mà xe đã đi.
  + fuel: Loại nhiên 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 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.

Thực hiện một số thống kê mô tả để có cái nhìn tổng quát về dữ liệu, bao gồm việc tính toán các số liệu thống kê cơ bản như trung bình, phương sai, và vẽ các biểu đồ phân bố giá, quãng đường đã đi, năm đăng ký và số lượng xe theo loại nhiên liệu. Điều này giúp xác định định hướng và mục tiêu cho dữ liệu.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Thuộc tính | Số lượng unique | Giá trị thường xuất hiện nhất | Số lần xuất hiện nhiều nhất | Giá trị trung bình | Giá trị độ lệch chuẩn | Giá trị nhỏ nhất | Q1 (25%) | Trung vị (50%) | Q3 (75%) | Giá trị lớn nhất |
| name | 2058 | Maruti Swift Dzire VDI | 129 | - | - | - | - | - | - | - |
| year | - | - | - | 2013.804 | 4.044 | 1983 | 2011 | 2015 | 2017 | 2020 |
| selling\_price | - | - | - | 638271.75 | 806253.4 | 29999 | 254999 | 450000 | 675000 | 10000000 |
| km\_driven | - | - | - | 69819.51 | 56550.55 | 1 | 35000 | 60000 | 98000 | 2360457 |
| fuel | 4 | Diesel | 4402 | - | - | - | - | - | - | - |
| seller\_type | 3 | Individual | 6766 | - | - | - | - | - | - | - |
| transmission | 2 | Manual | 7078 | - | - | - | - | - | - | - |
| owner | 5 | First Owner | 5289 | - | - | - | - | - | - | - |
| mileage | 393 | 18.9 kmpl | 225 | - | - | - | - | - | - | - |
| engine | 121 | 1248 CC | 1017 | - | - | - | - | - | - | - |
| max\_power | 322 | 74 bhp | 377 | - | - | - | - | - | - | - |
| torque | 441 | 190Nm@ 2000rpm | 530 | - | - | - | - | - | - | - |
| seats | - | - | - | 5.416719 | 0.959588 | 2 | 5 | 5 | 5 | 14 |

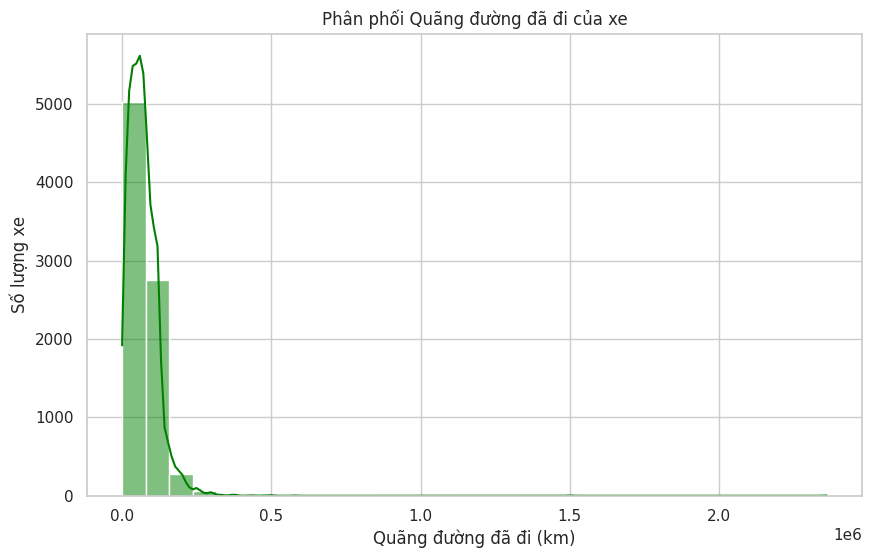
##### Bảng 2.3 Số liệu thống kê căn bản của tập “Used Cars data form websites”

* ***name:***Có 2058 tên xe duy nhất, trong đó “Maruti Swift Dzire VDI” xuất hiện nhiều nhất với 129 lần.
* ***year:***Các xe có năm sản xuất từ 1983 đến 2020, với giá trị trung bình là năm 2013.
* ***selling\_price:***Giá bán dao động từ 29,999 đến 10,000,000, với giá trị trung bình khoảng 638,271.75.
* ***km\_driven:***Số km đã đi từ 1 đến 2,360,457 km, với trung bình là khoảng 69,819 km.
* ***fuel:***Có 4 loại nhiên liệu, trong đó “Diesel” là phổ biến nhất với 4402 lần xuất hiện*.*
* ***seller\_type:***Ba loại người bán; “Individual” là loại phổ biến nhất với 6766 lần.
* ***transmission:*** Hai loại hộp số, “Manual” chiếm ưu thế với 7078 lần.
* ***owner:***Có 5 loại chủ sở hữu, trong đó “First Owner” là phổ biến nhất với 5289 lần.
* ***mileage:*** 393 giá trị duy nhất về mức tiêu thụ nhiên liệu, với “18.9 kmpl” là giá trị xuất hiện nhiều nhất (225 lần).
* ***engine:***121 loại động cơ khác nhau, trong đó “1248 CC” là phổ biến nhất với 1017 lần.
* ***max\_power:*** 322 giá trị duy nhất về công suất tối đa, “74 bhp” là phổ biến nhất với 377 lần.
* ***torque:***441 loại mô-men xoắn, trong đó “190Nm@ 2000rpm” là phổ biến nhất với 530 lần.
* ***seats:***Số chỗ ngồi từ 2 đến 14, với trung bình là 5.4 chỗ.



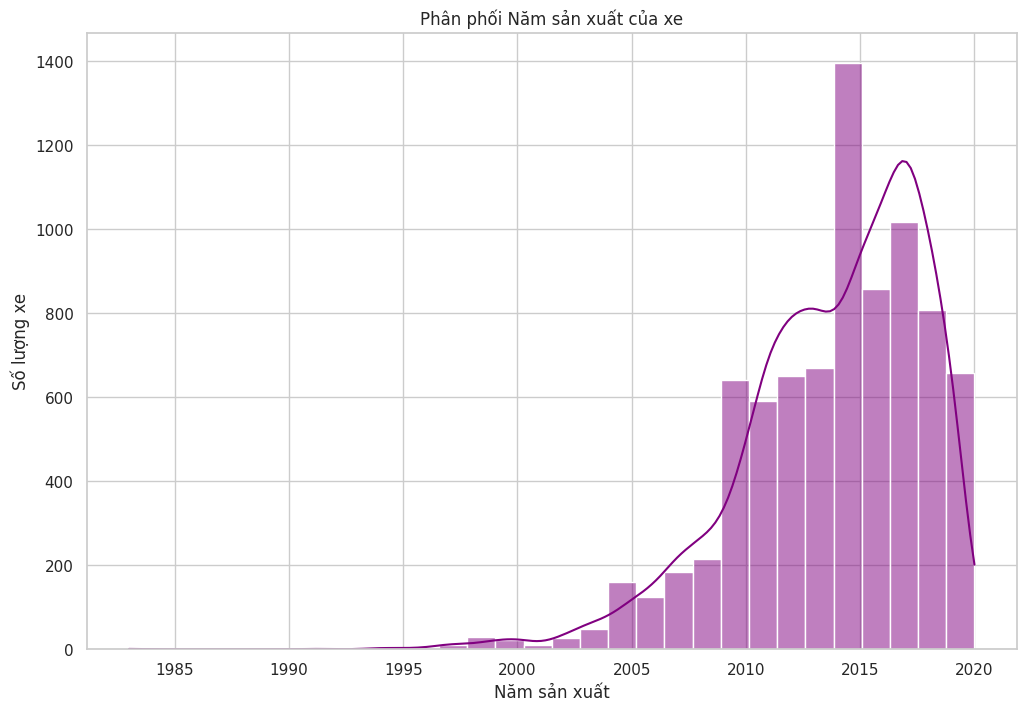
#### Hình 2.3 Biểu đồ phân phối giá

**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.



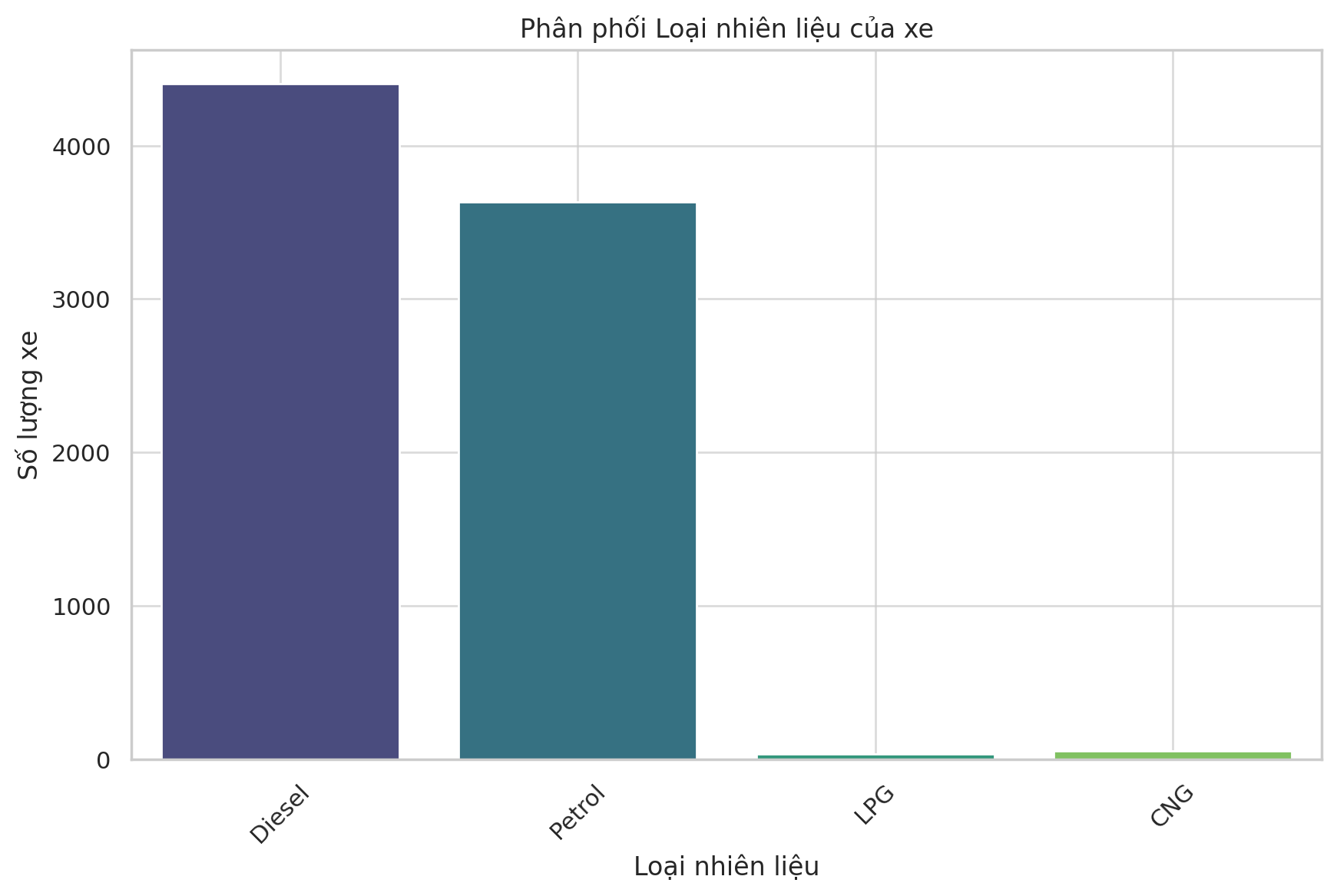
#### Hình 2.4 Biểu đồ phân phối quãng đường đã đi

**Phân bố quãng đường đã đi:** Biểu đồ phân phối quãng đường đã đi của xe cho thấy đa số các xe được bán lại khi chưa đi quá nhiều, với phần lớn tập trung dưới 100,000 km. Phân phối lệch phải này phản ánh thực tế là xe càng đi 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.



#### Hình 2.5 Biểu đồ phân phối năm đăng ký

**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 vì lý do kỹ thuật và an toàn.



#### Hình 2.6 Biểu đồ số lượng loại nhiên liệu xe

**Biểu đồ số lượng loại nhiên liệu xe:** Biểu đồ cho thấy diesel và petrol là hai loại nhiên liệu chính được sử dụng trong các xe có mặt trong dữ liệu, với diesel có phần nhỉnh hơn một chút so với petrol. Sự hiện diện thấp của các xe sử dụng CNG và LPG có thể phản ánh sự thiếu hụt cơ sở hạ tầng hỗ trợ hoặc ít ưa chuộng hơn do các yếu tố như hiệu quả nhiên liệu và chi phí vận hành. Tổng thể, phân bố nhiên liệu cho thấy một thị trường đa dạng nhưng vẫn tập trung chủ yếu vào hai loại nhiên liệu truyền thống là diesel và petrol.

### 2.2.3 Phân tích thống kê định lượng

Các tiêu chí thống kê mô tả cần được xem xét bao gồm:

* + Giá trị trung bình (Mean)
  + Giá trị nhỏ nhất (Min)
  + Giá trị lớn nhất (Max)
  + Trung vị (Median)
  + Phương sai (Variance)
  + Độ nhọn (Kurtosis)
  + Độ xiên (Skewness)

Thực hiện tính toán với các cột price, kms\_driven, registration\_year

|  |  |  |  |
| --- | --- | --- | --- |
|  | selling\_price | km\_driven | year |
| Mean | 638,271.8 | 69,819.51 | 2013.80 |
| Min | 29,999 | 1 | 1983 |
| Max | 10,000,000 | 2,360,457 | 2020 |
| Median | 450,000 | 60,000 | 2015 |
| Variance | 650,044,600,000 | 3,197,965,000 | 16.36 |
| Kurtosis | 21.08 | 384.1 | 1.71 |
| Skewness | 4.19 | 11.17 | -1.07 |

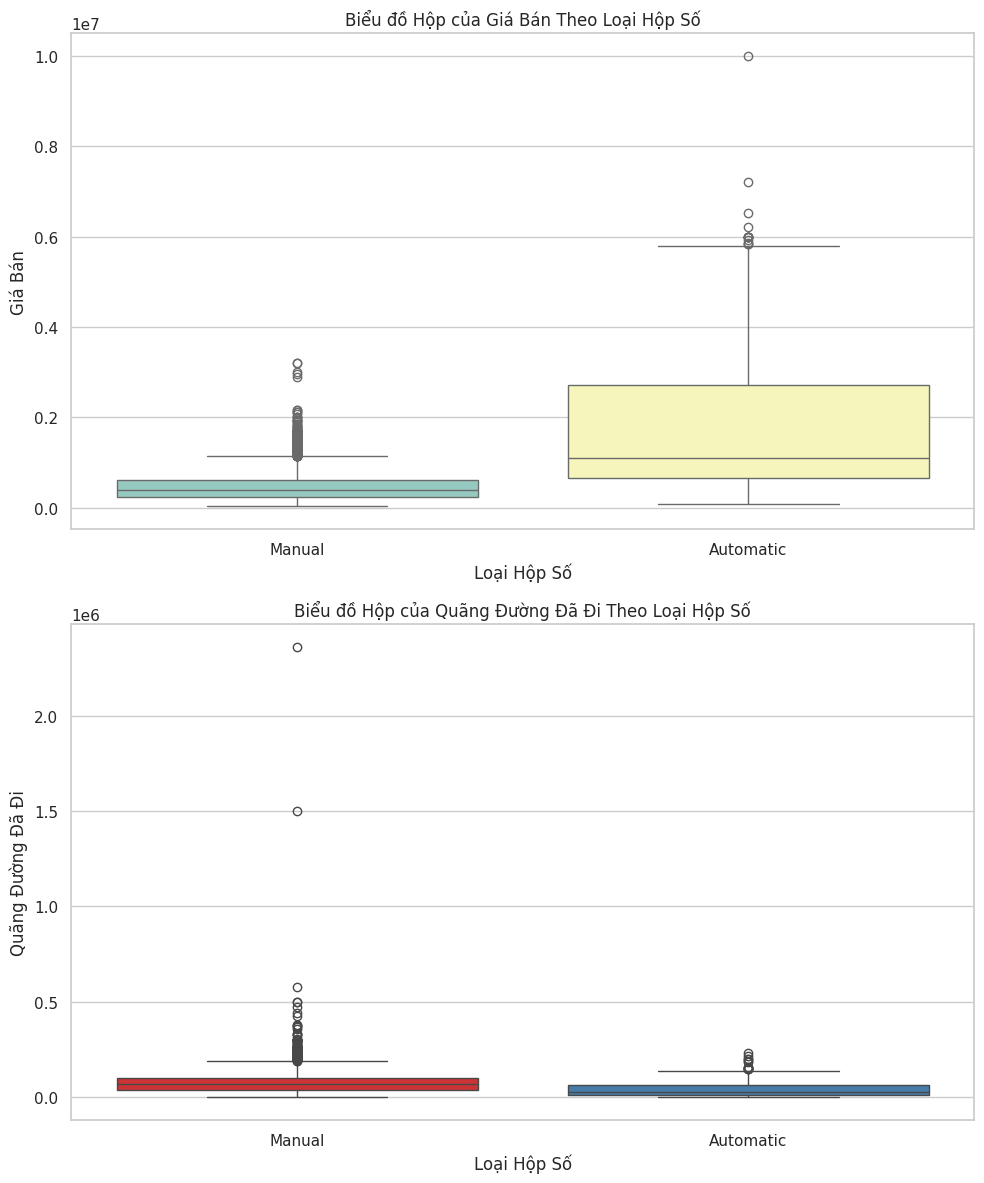
##### Bảng 2.4 Thống kê định lượng của tập “Used Cars data form websites”

* **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 năm 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.

### 2.2.4 Nghiên cứu sâu về mẫu dữ liệu đặc trưng

* **Biểu đồ hộp giá bán theo loại hộp số:** Giá bán là một trong những yếu tố quan trọng nhất đối với người mua xe. Việc phân tích này giúp hiểu được mối quan hệ giữa loại hộp số và giá bán của xe, từ đó cho thấy liệu có sự chênh lệch giá đáng kể giữa các xe hộp số tự động và hộp số sàn hay không. Điều này có thể hỗ trợ người mua trong việc đưa ra quyết định mua xe dựa trên ngân sách và sở thích cá nhân.
* **Biểu đồ hộp quãng đường đã đi theo loại hộp số:** Quãng đường xe đã đi không chỉ phản ánh tuổi thọ của xe mà còn ảnh hưởng đến giá trị tái bán của xe. Phân tích này giúp xác định xem xe hộp số tự động hay xe hộp số sàn có xu hướng được sử dụng nhiều hơn, qua đó cung cấp thông tin về độ bền và khả năng chịu đựng của từng loại trong thực tế sử dụng.

Cả hai biểu đồ này đều cung cấp thông tin cần thiết cho cả người mua và người bán, giúp họ có cái nhìn toàn diện hơn về thị trường xe cũ, đặc biệt là liên quan đến các yếu tố như loại hộp số, giá cả và độ bền của xe.



#### Hình 2.7 Biểu đồ Hộp của Giá Bán Theo Loại Hộp Số

**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 Quảng Đường Đã Đi 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 đa 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. Hộp tự động cũng có phạm vi hẹp, tương tự như hộp số sàn, với rất ít ngoại lệ, cho thấy sự đồng đều trong quãng đường xe đã đi không chênh lệch nhiều. 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.

**Lí do cho sự khác biệt:**

1. **Đối tượng người dùng:**

* Hộp số sàn có thể phổ biến hơn ở những người dùng yêu thích và chủ động lái xe, có thể lái xe nhiều hơn, chẳng hạn như những người có sở thích đi du lịch bằng xe hoặc những người sử dụng xe cho công việc đòi hỏi di chuyển nhiều.
* Hộp số tự động thường được ưa chuộng ở những người dùng tìm kiếm sự tiện lợi và thoải mái khi lái, những người này có thể không đi lại nhiều nhưng lại ưa chuộng các chuyến đi ngắn và thuận tiện trong đô thị.

1. **Mục đích sử dụng:**

* Xe hộp số sàn có thể được sử dụng nhiều hơn trong các tình huống cần đến sự kiểm soát tốc độ cao, chẳng hạn như trong lái xe địa hình hoặc trên đường cao tốc.
* Xe hộp số tự động có thể phổ biến trong việc sử dụng hàng ngày, đi lại trong thành phố, nơi không cần đến những chuyến đi dài.

1. **Khả năng tiếp cận và chi phí:**

* Xe hộp số sàn thường có giá thành thấp hơn xe tự động và do đó, chúng có thể được mua bởi những người dùng có nhu cầu di chuyển cao nhưng với ngân sách hạn chế.
* Xe tự động thường đắt hơn và có thể không được sử dụng cho những chuyến đi dài thường xuyên do chi phí bảo dưỡng cao hơn và giá trị sử dụng giảm nhanh chóng.

1. **Bảo dưỡng và độ tin cậy:**

* Xe hộp số sàn có thể ít phức tạp hơn về mặt cơ khí so với xe tự động và do đó dễ dàng bảo dưỡng hơn, cho phép chúng đi được quãng đường lớn hơn trước khi cần sửa chữa.
* Xe tự động có thể đòi hỏi bảo dưỡng định kỳ và chuyên nghiệp hơn, điều này có thể giảm số lượng và quãng đường của những chuyến đi để tránh hư hỏng.

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

### 2.3.1 Tổng Quan Về Tập Dữ Liệu

Nguồn gốc và mô tả

* Số lượng trường: 13 fields
* Số lượng bản ghi: 4727 records
* Tác giả: [Muhammad Awais Tayyab](https://www.kaggle.com/muhammadawaistayyab)
* 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](https://www.kaggle.com/datasets/muhammadawaistayyab/used-cars-prices-in-uk)

### 2.3.2 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 sơ 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.

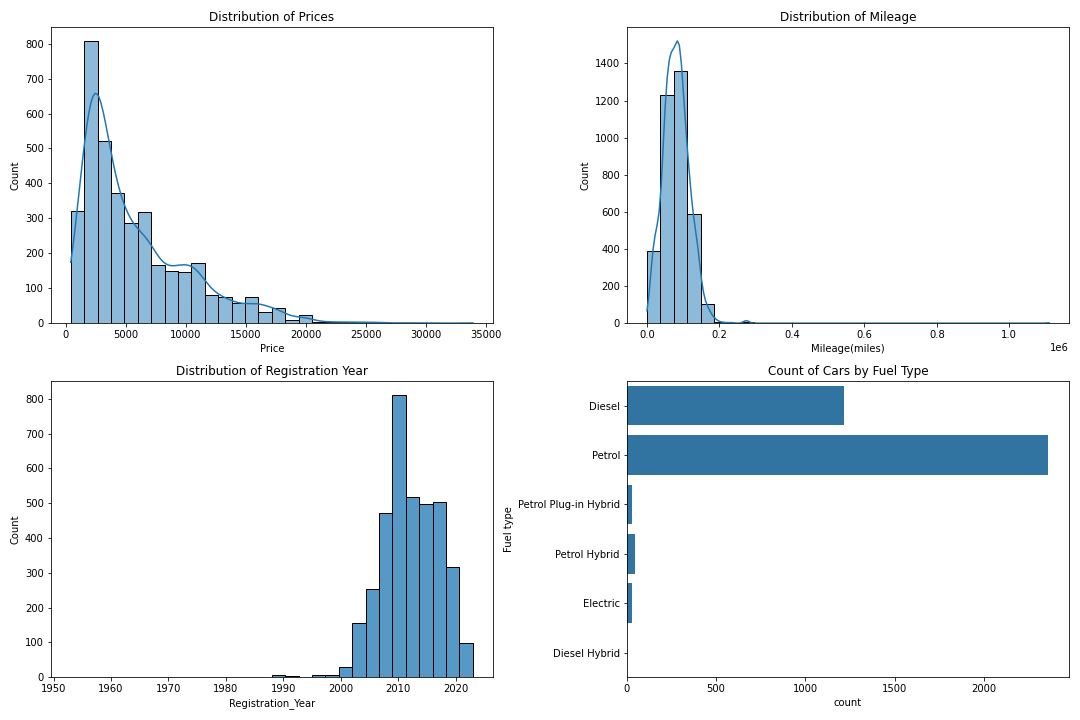
**Kiểm tra giá trị thiếu:**

* + 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

Thực hiện một số thống kê mô tả để có cái nhìn tổng quát về dữ liệu, bao gồm việc tính toán các số liệu thống kê cơ bản như trung bình, phương sai, và vẽ các biểu đồ phân bố giá, quãng đường đã đi, năm đăng ký và số lượng xe theo loại nhiên liệu. Điều này giúp xác định định hướng và mục tiêu cho dữ liệu.

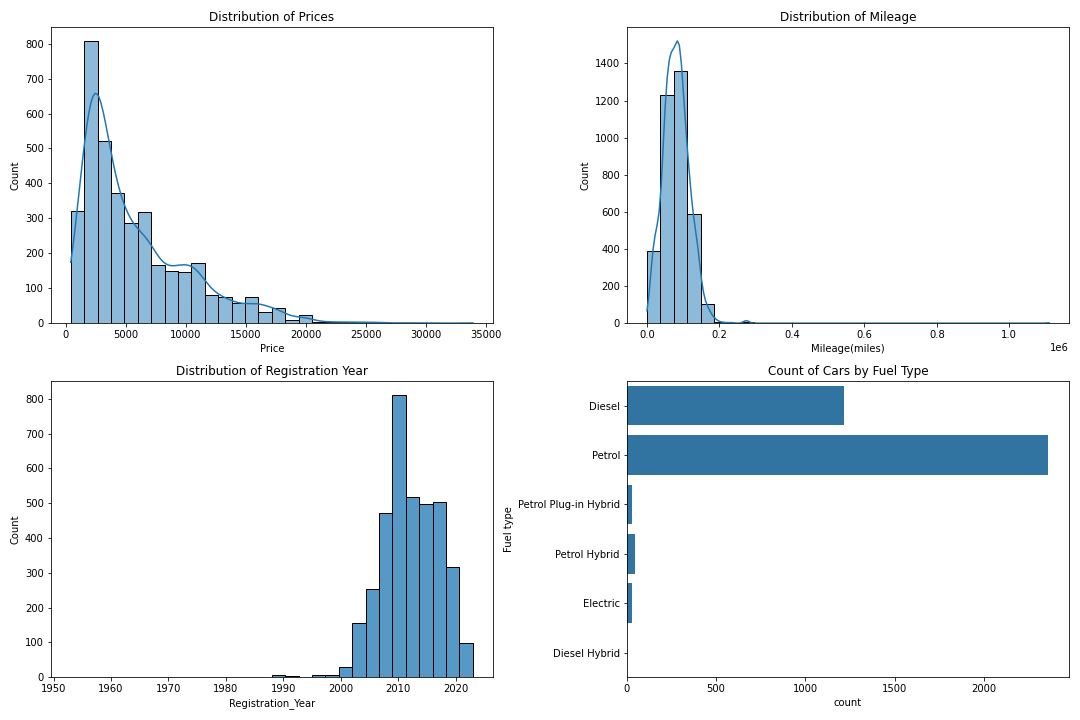
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Price | Mileage(miles) | Registration\_Year | Previous Owners | Doors | Seats |
| count | 3685.00 | 3685 | 3685 | 2276 | 3660 | 3650 |
| mean | 5787.15 | 81328.16 | 2011.84 | 2.81 | 4.32 | 4.90 |
| std | 4480.81 | 39420.83 | 5.09 | 1.55 | 0.99 | 0.58 |
| min | 400 | 1 | 1953 | 1.00 | 2 | 2 |
| 25% | 2490 | 56984 | 2008 | 2.00 | 3 | 5 |
| 50% | 4000 | 80000 | 2012 | 3.00 | 5 | 5 |
| 75% | 7995 | 103000 | 2015 | 4.00 | 5 | 5 |
| max | 33900 | 1110100 | 2023 | 9.00 | 5 | 7 |

##### Bảng 2.5 số liệu thống kê căn bản của tập “Used Cars Prices in UK”

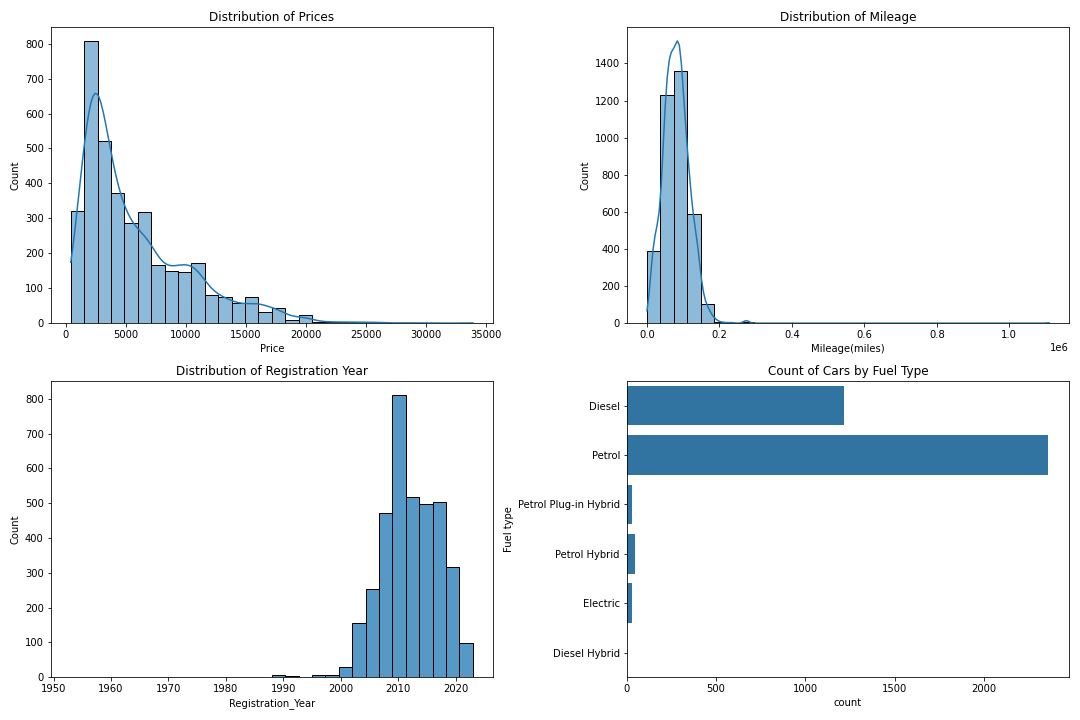


#### Hình 2.8 Biểu đồ phân phối giá

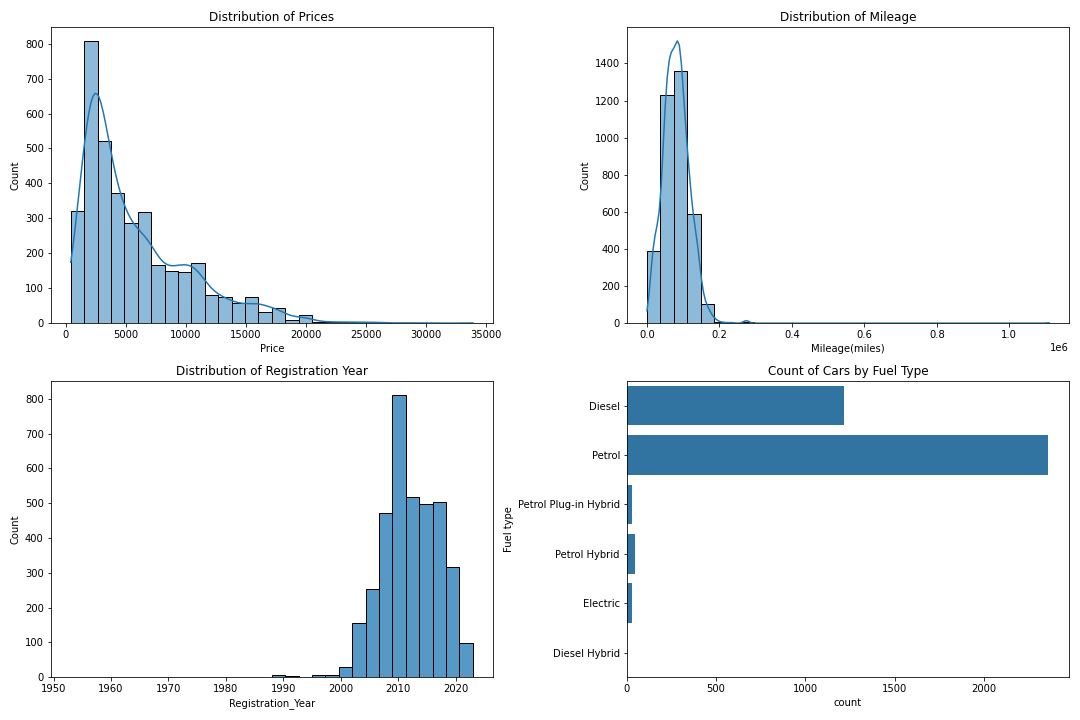
**Phân bố giá:** Biểu đồ phân bố giá cho thấy phần lớn xe nằm trong tầm giá thấp và trung bình, với một số ít xe có giá rất cao*.*



#### Hình 2.9 Biểu đồ phân phối quãng đường đã đi

**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.

#### Hình 2.10 Biểu đồ phân phối năm đăng ký

**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*.*

#### Hình 2.11 Biểu đồ số lượng loại nhiên liệu xe

**Số lượng xe theo loại nhiên liệu:** Đa số xe sử dụng nhiên liệu xăng, tiếp theo là diesel, và một số ít sử dụng các loại nhiên liệu khác*.*

**Thống kê mô tả:**

**Giá (Price):** Giá trung bình của xe là 5,787 GBP, với giá thấp nhất là 400 GBP và cao nhất là 33,900 GBP. Phân bố giá cho thấy sự tập trung cao ở mức giá thấp, với một số ít xe có giá rất cao.

**Quãng đường đã đi (Mileage):** Quãng đường trung bình là khoảng 81,328 dặm, với phạm vi từ 1 dặm đến 1,110,100 dặm. Điều này cho thấy sự đa dạng lớn về xe đã qua sử dụng.

**Năm đăng ký (Registration Year):** Các xe có năm đăng ký từ 1953 đến 2023, với số lượng lớn xe được đăng ký trong khoảng thời gian gần đây, phản ánh sự có mặt nhiều hơn của xe mới hơn trên thị trường xe đã qua sử dụng.

**Số chủ sở hữu trước (Previous Owners):** Trung bình mỗi xe đã trải qua khoảng 2.8 chủ sở hữu trước, với số lượng chủ sở hữu từ 1 đến 9.

### 2.3.3 Phân tích thống kê định lượng

Các tiêu chí thống kê mô tả cần được xem xét bao gồm:

* + Giá trị trung bình (Mean)
  + Giá trị nhỏ nhất (Min)
  + Giá trị lớn nhất (Max)
  + Trung vị (Median)
  + Phương sai (Variance)
  + Độ nhọn (Kurtosis)
  + Độ xiên (Skewness)

Thực hiện tính toán với các cột Price, Mileage(miles), Registration\_Year

|  |  |  |  |
| --- | --- | --- | --- |
|  | Price | Mileage(miles) | Registration\_Year |
| Mean | 5787.15 | 81328.164 | 2011.84 |
| Min | 400 | 1 | 1953 |
| Max | 33900 | 1110100 | 2023 |
| Median | 4000 | 80000 | 2012 |
| Variance | 20077663.38 | 1554002121.29 | 25.93 |
| Kurtosis | 1.91 | 125.81 | 6.57 |
| Skewness | 1.38 | 5.2 | -0.93 |

##### Bảng 2.6 Thống kê định lượng của tập “Used Cars Prices in UK”

**Giá xe:**

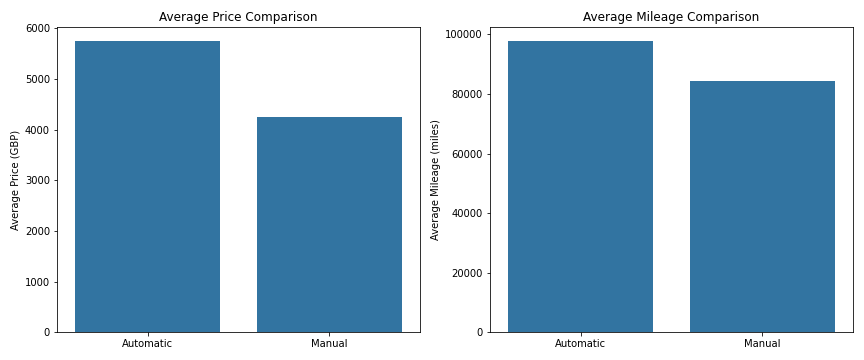
* **Giá trị trung bình và Trung vị:** Giá trung bình cao hơn trung vị (5787.15 GBP so với 4000 GBP), điều này cho thấy sự phân phối của giá xe bị lệch về phía giá cao, với một số xe có giá rất cao kéo giá trung bình lên.
* **Min và Max:** Khoảng giá từ 400 GBP đến 33900 GBP phản ánh sự đa dạng lớn trong giá của xe đã qua sử dụng, từ những xe giá rẻ đến những xe cao cấp.
* **Phương sai:** Giá trị phương sai lớn (20077663.38 GBP^2) cho thấy sự biến động lớn về giá xe trong tập dữ liệu, với một số xe có giá cao hơn hoặc thấp hơn nhiều so với giá trung bình.
* **Độ nhọn (Kurtosis):** Giá trị kurtosis là 1.91, lớn hơn 0, cho thấy phân phối giá có đỉnh cao hơn và 'đuôi' dày hơn so với phân phối chuẩn, tức là có nhiều giá trị nằm ở hai đầu của phạm vi giá hơn là ở giữa.
* **Độ xiên (Skewness):** Giá trị độ xiên là 1.38, cho thấy phân phối giá bị lệch về phía giá thấp, với 'đuôi' phân phối kéo dài ra phía giá cao, phản ánh sự tồn tại của một số xe có giá rất cao so với phần còn lại.

### 2.3.4 Nghiên cứu sâu về mẫu dữ liệu đặc trưng

**Chọn Lọc và Lý Giải Mẫu Dữ Liệu**

Để thực hiện phân tích này, chúng tôi sẽ tiếp cận bằng cách chọn lọc một mẫu dữ liệu dựa trên đặc trưng động cơ của xe. Cụ thể là chia xe thành hai nhóm: xe sở hữu hộp số tự động (Automatic) và xe sở hữu hộp số sàn (Manual), hoặc các loại Fuel Type. Tiêu chí lựa chọn mẫu dữ liệu dựa trên giả định rằng loại hộp số, loại nhiên liệu có thể ảnh hưởng đến các yếu tố khác như giá, quãng đường đã đi, và năm đăng ký. Để đảm bảo tính đại diện, chúng tôi sẽ lấy ra 2000 mẫu đầu tiên từ dữ liệu, giả định rằng dữ liệu đã được xáo trộn đủ tốt.

**Phân Tích Đặc Trưng và Độ Chênh Lệch**

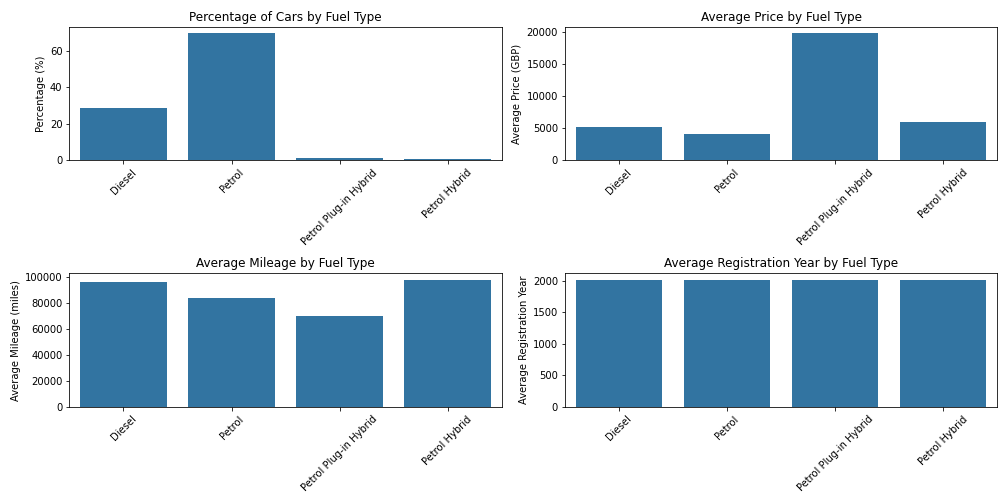
****

#### Hình 2.12 Biểu đồ giá trung bình và quãng đường đã đi trung bình giữa 2 nhóm xe

Như biểu đồ trên cho thấy:

* **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 Đi Trung Bình:** Xe tự động cũng cho thấy quãng đường đã đi 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 gia 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 đi giữa hai nhóm xe.

**

#### Hình 2.13 Biểu đồ phân tích trực quan hoá sự chênh lệch giữa các nhóm xe dựa trên loại nhiên liệu về giá trung bình, quãng đường đã đi trung bình, và các đặc trưng khác

Như biểu đồ trên cho thấy:

* **Tỷ Lệ Phần 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 Đi 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 trung bình 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 thải 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

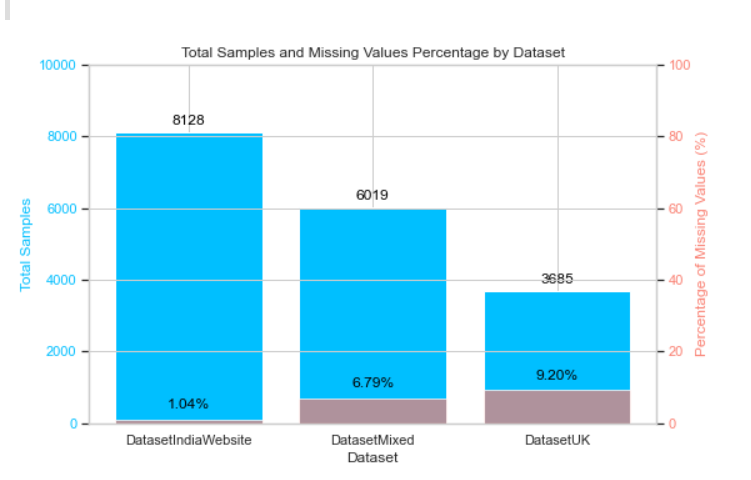
## 3.1 So sánh tổng quát

* **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:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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). Thông 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 số chủ sở hữu trước, loại thân xe, số cửa, lớp phát thải và lịch sử bảo dưỡng. Tập dữ liệu này cung cấp cái nhìn sâu hơn về tình trạng bảo dưỡng và các tiêu chuẩn phát thải, có ý nghĩa quan trọng đối với thị trường xe cũ ở châu Âu. |
| **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) và số (int, float). Các trường như mileage, engine, và max\_power được lưu trữ dưới dạng chuỗi, có thể cần chuyển đổi để phân tích chi tiết hơn. | Có sự khác biệt về định dạng, như các cột về Emission Class và Service history, cung cấp thông tin chi tiết hơn về tình trạng và môi trường của xe. |
| **Quy mô dữ liệu** | Có 6019 mẫu, đa dạng về địa điểm và có thông tin 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 trị bán lại và môi trường. |

## 3.2 Đi vào phân tích và đưa ra kết luận về tập dataset chính

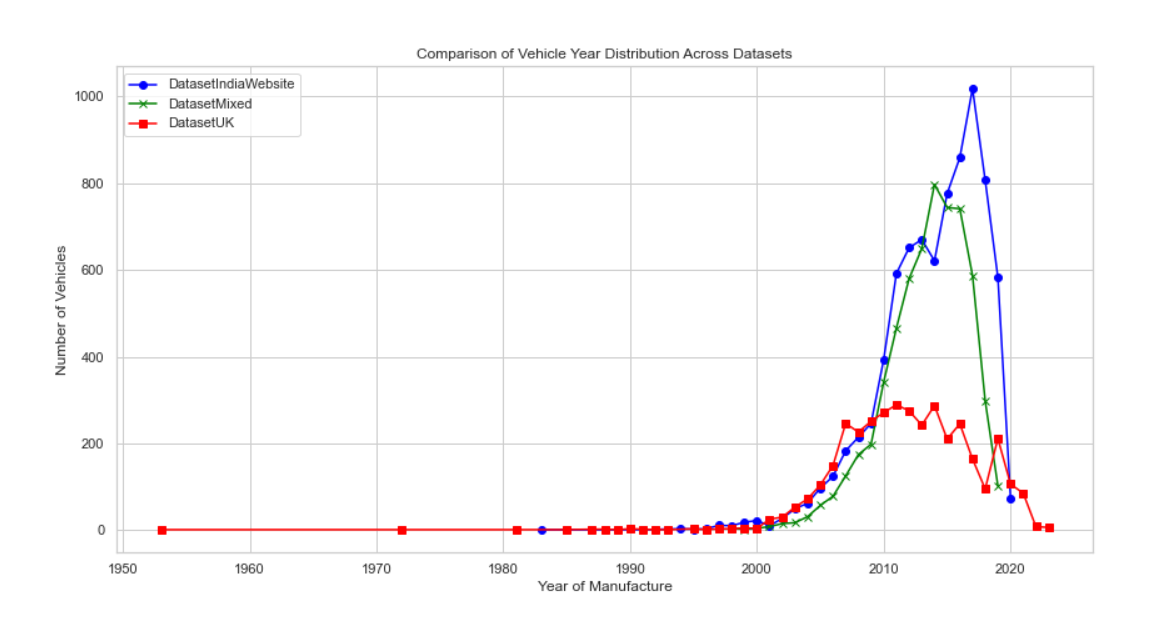
Dựa vào đặc điểm như loại nhiên liệu, hộp số, năm sản xuất và số km đã đi, chúng tôi đã phân tích dữ liệu xe cũ từ nhiều nguồn, kể cả ở Ấn Độ và Vương quốc Anh. Công việc này bao gồm thống kê và trực quan hóa dữ liệu, giúp hiểu rõ hơn về thị trường xe cũ và các xu hướng. Nhờ vào các biểu đồ, chúng tôi đã làm nổi bật được sự khác biệt giữa các xe dựa trên thống kê dữ liệu, năm ra đời của từng loại xe



#### Hình 3.1 Biểu đồ tổng số mẫu và phần trăm dữ liệu null của từng tập dataset

Biểu đồ “Biểu đồ tổng số mẫu và phần trăm dữ liệu null của từng tập dataset” đã cung cấp cái nhìn chi tiết và trực quan về số lượng mẫu cũng như tỷ lệ phần trăm dữ liệu bị thiếu trong ba tập dữ liệu khác nhau: DatasetIndiaWebsite, DatasetMixed, và DatasetUK:

* 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 đầy đủ và chính xác các xu hướng hoặc mẫu từ tập dữ liệu.



#### Hình 3.2 Biểu đồ phân bố của năm sản xuất xe của từng tập dataset

Biểu đồ đường thể hiện sự phân bố của năm sản xuất xe trong ba tập dữ liệu: DatasetIndiaWebsite, DatasetMixed, và DatasetUK:

* DatasetMixed (màu xanh lá) có phân bố rộng hơn về năm sản xuất, từ những năm 2000 đến 2020, với một số lượng đáng kể xe từ những năm 2010 trở đi. Điều này cho thấy tập dữ liệu này đa dạng về độ tuổi của xe.
* DatasetIndiaWebsite (màu xanh dương) cho thấy một lượng lớn xe từ các năm gần đây, với đỉnh điểm vào khoảng năm 2010-2020. Điều này cho thấy tập dữ liệu này có xu hướng chứa các xe khá mới.
* DatasetUK (màu đỏ) thể hiện sự phân bố tương tự như DatasetIndiaWebsite nhưng với số lượng xe trong từng năm ít hơn. Sự phân bố tập trung chủ yếu vào khoảng năm 2000-2015 và có giá trị ngoại lai tương đối nhiều.

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. Trong khi đó, hai tập dữ liệu còn lại yêu cầu phải cân nhắc kỹ lưỡng hơn nếu được lựa chọn, do số lượng dữ liệu ít hơn và tỷ lệ dữ liệu thiếu cao hơn, có thể làm giảm chất lượng của các phân tích dựa trên chúng. Không chỉ vậy, nếu xét đến việc cung cấp số lượng xe mới và đời gần đây, thì tập dữ liệu DatasetIndiaWebsite dường như là tập dữ liệu tốt nhất, phù hợp cho các phân tích nhằm hiểu rõ về tình trạng xe hiện đại trong thị trường ngày nay.

Trong chương này của bài nghiên cứu, chúng ta đã thực hiện một so sánh tổng quan và đi đến kết luận về ba bộ dữ liệu xe cũ từ các nguồn khác nhau. Mặc dù cả ba đều chứa thông tin cơ bản về xe như loại nhiên liệu, số chỗ ngồi, thông tin động cơ và giá xe, sự khác biệt rõ ràng về cấu trúc, định dạng dữ liệu, và quy mô dữ liệu giữa chúng đã được làm nổi bật. Cụ thể, từ việc phân tích tỉ lệ giá trị null trong các bộ dữ liệu đến việc thống kê và trực quan hóa dữ liệu, việc làm này không chỉ giúp hiểu rõ hơn về thị trường xe cũ và các xu hướng hiện hành mà còn đề xuất cách thức để tăng cường chất lượng dữ liệu qua việc tiền xử lí dữ liệu. Điều này là cực kỳ quan trọng trong việc nâng cao độ chính xác và độ tin cậy của các mô hình dự đoán và phân tích trong tương lai. Bằng cách nhìn nhận vào các điểm tương đồng và khác biệt giữa các bộ dữ liệu, nghiên cứu này mở ra cánh cửa cho việc phát triển các phương pháp tiếp cận mới trong việc đánh giá và dự đoán giá xe cũ, đồng thời cung cấp một cái nhìn sâu sắc về cách thức chuẩn bị dữ liệu một cách hiệu quả cho các nghiên cứu sau này.

# CHƯƠNG 4. TIỀN XỬ LÝ VÀ CHUẨN HÓA DỮ LIỆU

## 4.1 Các bước phân tích và nhận xét về từng đặc trưng

Qua việc phân tích các quy mô, đặc trưng dữ liệu, ta sẽ hiểu rõ hơn được giá trị của từng cột dữ liệu trong tập dataset, giúp lên kế hoạch cho việc xử lý và phân tích dữ liệu phù hợp. Từ các thao tác kiểm tra giá trị thiếu, thống kê mô tả các phân phối cũng như nghiên cứu về các đặc trưng đã thực hiện ở trên. Ta cần đi sâu vào các bước sau đây:

* Nhận thấy rằng trước tiên cần phải xử lí các giá trị null của các đặc trưng “mileage”, “engine”, “max\_power”, “torque” and “seats”. Điều này quan trọng để lập kế hoạch cho các bước tiền xử lý sau đó như điền giá trị thiếu hoặc loại bỏ dữ liệu không đầy đủ.
* Phân tích các biểu đồ phân bố đặc trưng sẽ cho ta nhận thấy rằng ở các mẫu dữ liệu về loại nhiên liệu có sự tách biệt thị trường rõ ràng. Xe sử dụng nhiên liệu diesel với petrol đang phản ánh mạnh mẽ xu hướng tiêu dùng nhiên liệu hiện tại và giúp đưa ra các quyết định chiến lược trong việc mua bán và quảng bá xe. Do đó các giá trị nhiên liệu còn lại hầu như không ảnh hưởng đến thị trường và sự bị loại bỏ để trách nhiễu.
* Phân tích yếu tố đặc trưng là nhân tố tác động trực tiếp đến thị trường là con người. Ở đây nhu cầu về phần lớn xe có bao nhiêu ghế sẽ giúp xác định phân khúc khách hàng mục tiêu chính. Trong tập dữ liệu đã chỉ ra rằng đa số nhu cầu về xe của khách hàng là loại xe thường thấy, có số ghế tập trung lớn vào khoảng 6254 xe có 5 ghế, 1120 xe có 7 ghế nhưng chỉ có 19 xe có 10 ghế, 80 xe có 9 ghế… đã cho thấy rằng phân khúc xe ngoại lệ, các xe chuyên dụng hầu như ít có và chỉ cần tập trung vào phân khúc xe có số ghế dưới 9.
* Chuẩn hóa các giá trị chuỗi như mức tiêu thị nhiên liệu “mileage”, dung tích động cơ “engine”, xóa các bản ghi có giá trị 0 do sự sai xót của việc thu thập dữ liệu không thực tế cũng như tính toán về tuổi xe - là một yếu tố quan trọng ảnh hưởng đến giá trị thị trường của chiếc xe đó.

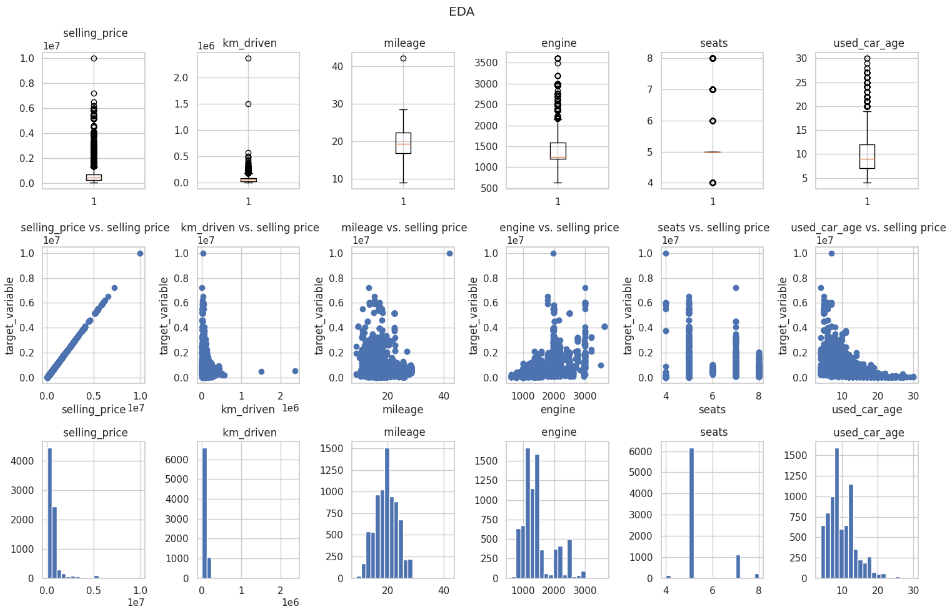
## 4.2 Loại bỏ các đặc trưng không sử dụng và giá trị ngoại lai

- Loại bỏ các cột không sử dụng:

* “name”: Tên xe không cần thiết cho phân tích số liệu như giá bán hay dự báo giá trị xe, vì thông tin này thường không mang tính dự báo hoặc phân tích.
* “year”: Năm sản xuất của xe đã được thay thế bằng cột “used\_car\_age”, cung cấp thông tin độ tuổi xe một cách trực tiếp hơn, là yếu tố quan trọng ảnh hưởng đến giá bán.
* “max\_power” và “torque”: Mặc dù thông tin này có thể hữu ích cho phân tích kỹ thuật, nhưng có thể không cần thiết nếu mục đích chính là phân tích ảnh hưởng của các yếu tố khác như giá bán, loại nhiên liệu, và số km đã đi. Loại bỏ nhằm tập trung vào các yếu tố quan trọng hơn cho mục tiêu phân tích cụ thể.

Loại bỏ các cột không cần thiết giúp tập trung vào những yếu tố có khả năng ảnh hưởng mạnh mẽ hơn đến giá xe hoặc các yếu tố phân tích khác. Điều này làm cho mô hình phân tích hoặc dự báo trở nên hiệu quả hơn. Giúp giảm bớt bộ nhớ sử dụng và cải thiện hiệu suất xử lý, đặc biệt quan trọng khi xử lý bộ dữ liệu lớn hoặc thực hiện các phân tích phức tạp. Loại bỏ các cột không liên quan còn giúp tránh gây nhiễu hoặc sai lệch trong các mô hình dự báo hoặc phân tích thống kê, tăng độ chính xác của mô hình.

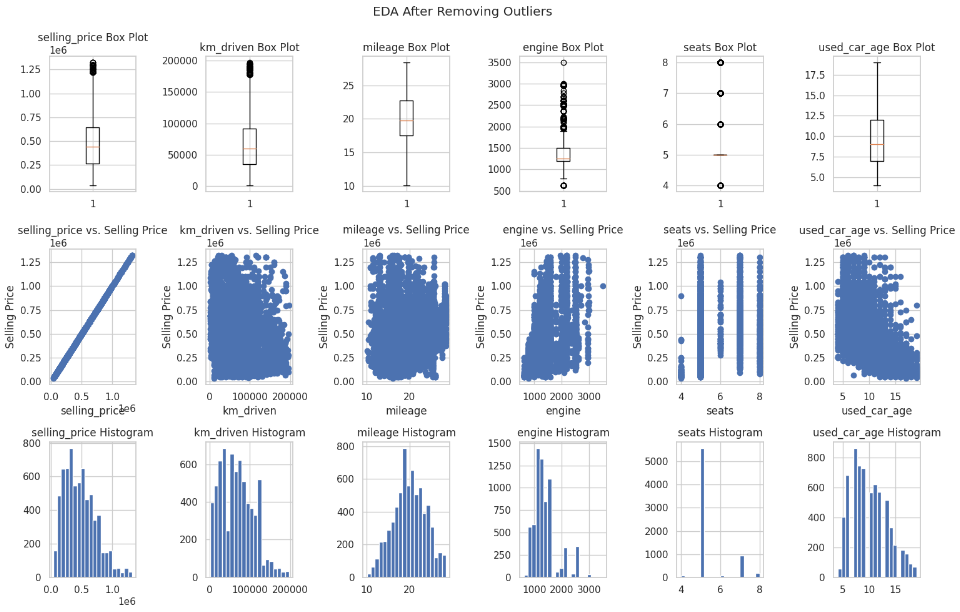
- Thực hiện phân tích thăm dò EDA - Exploratory Data Analysis: sử dụng các biểu đồ box plot hiển thị phân bố của dữ liệu, các giá trị ngoại lệ và khoảng tứ phân vị. Điều này giúp phát hiện nhanh các giá trị bất thường và phạm vi dữ liệu. Sử dụng biểu đồ scatter plot khảo sát mối quan hệ giữa mỗi biến số với biến mục tiêu (giá bán) giúp phát hiện các mối tương quan và xu hướng. Sử dụng biểu đồ histogram hiển thị phân phối tần suất của mỗi biến số, giúp hiểu rõ hơn về hình dạng và phân bố của dữ liệu.



#### Hình 4.1 Biểu đồ EDA của tập dữ liệu “DatasetIndiaWebsite”

- Xử lí giá trị ngoại lai và chuẩn hóa dữ liệu dựa trên phương pháp IQR (Interquartile Range). Ngưỡng trên và dưới được xác định là Q3 + 1.5IQR và Q1 - 1.5IQR, tương ứng, với Q1 và Q3 là các phân vị thứ 25 và 75:

* 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:

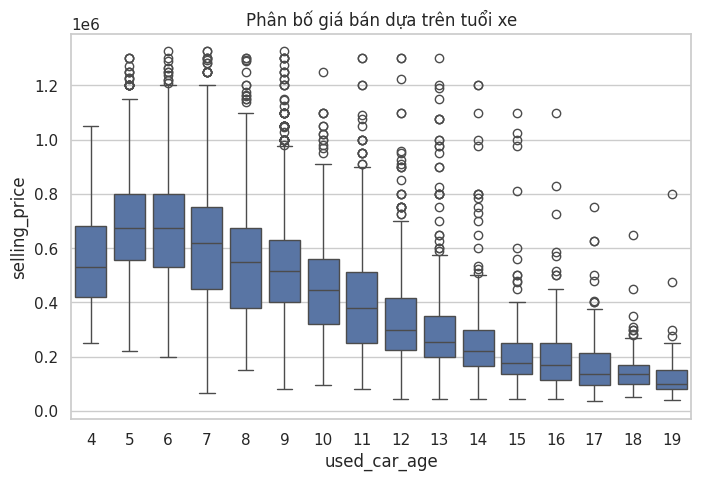


#### Hình 4.2 Biểu đồ EDA sau khi xử lý các giá trị ngoại lai của tập dữ liệu “DatasetIndiaWebsite”

* 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. ​​

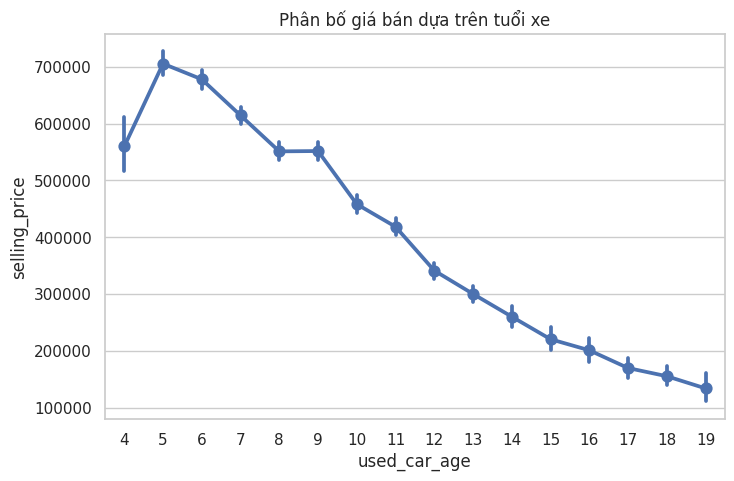
# CHƯƠNG 5: KHAI PHÁ DỮ LIỆU

## 5.1 Mối quan hệ giữa tuổi xe và giá bán



#### Hình 5.1 Biểu đồ Box phân bố giá bán dựa trên tuổi xe

* 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 xe 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.

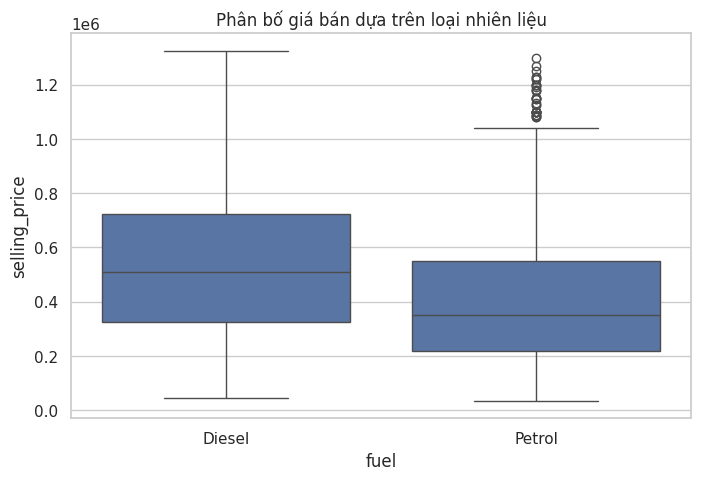


#### Hình 5.2 Biểu đồ Point phân bố giá bán dựa trên tuổi xe

* Xu hướng giảm giá rõ rệt: Có thể thấy rõ ràng rằng 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 ở những xe cực mới và cực 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.

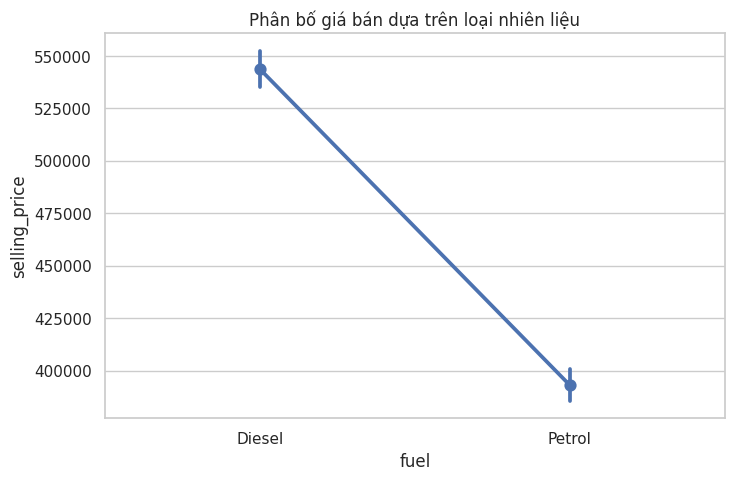
Vậy chúng ta có thể thấy được tuổi xe là yếu tố quan trọng quyết định đến giá xe. Xe càng mới thường có giá cao hơn do giá trị hao mòn thấp hơn, và giá giảm dần theo thời gian sử dụng do hao mòn và lỗi thời.

## 5.2 Mối quan hệ giữa loại nhiên liệu và giá bán



#### Hình 5.3 Biểu đồ Box phân bố giá bán dựa trên loại nhiên liệu

* 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 ưa 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: Cả hai loại nhiên liệu đều có những ngoại lệ về giá, nhưng xe chạy diesel có xu hướng có nhiều giá cao hơn, cho 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 ưa chuộng và giá trị bền vững của chúng trên thị trường.

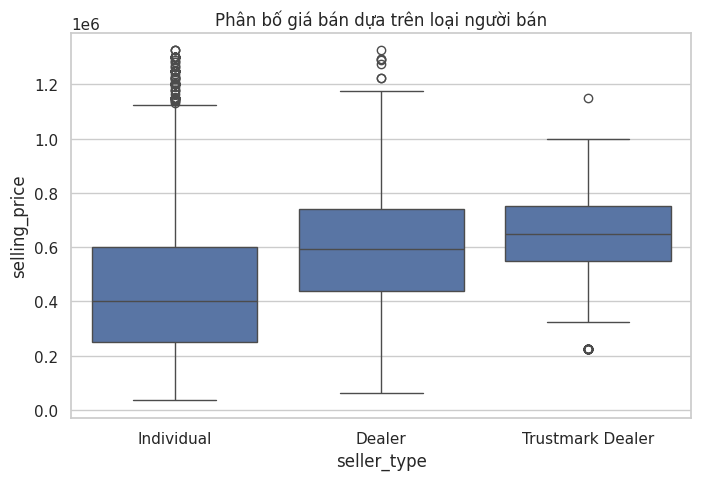


#### Hình 5.4 Biểu đồ Point phân bố giá bán dựa trên loại nhiên liệu

* 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 ban 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.

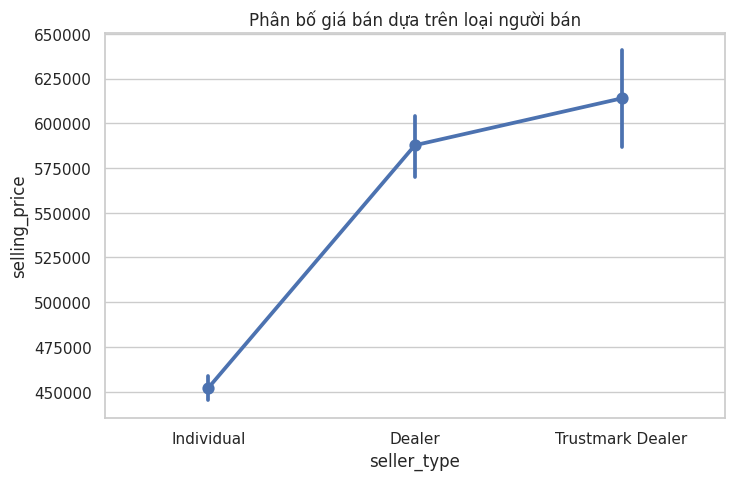
Xe chạy bằng dầu diesel thường có giá cao hơn so với xe chạy bằng xăng do một số lợi thế vượt trội. Đầu tiên, động cơ diesel được thiết kế để chịu đựng áp suất cao hơn, do đó thường bền hơn và có tuổi thọ dài hơn, làm cho chúng trở thành lựa chọn ưu tiên cho những người lái xe thường xuyên và trong các điều kiện khắc nghiệt. Thứ hai, động cơ diesel hiệu quả hơn về mặt nhiên liệu, đặc biệt khi lái xe ở tốc độ cao hoặc vận chuyển hàng hóa nặng, giúp tiết kiệm đáng kể chi phí nhiên liệu. Thêm vào đó, mô-men xoắn lớn hơn ở tốc độ thấp giúp xe diesel lý tưởng cho việc kéo hoặc vận chuyển, phù hợp với xe tải và SUV, thường có giá cao hơn xe con. Cuối cùng, xe diesel giữ giá tốt hơn khi bán lại nhờ độ bền và hiệu quả nhiên liệu của chúng. Những yếu tố này làm tăng giá trị ban đầu và giá bán lại của xe diesel, biến chúng 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



#### Hình 5.5 Biểu đồ Box phân bố giá dựa trên loại người 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.



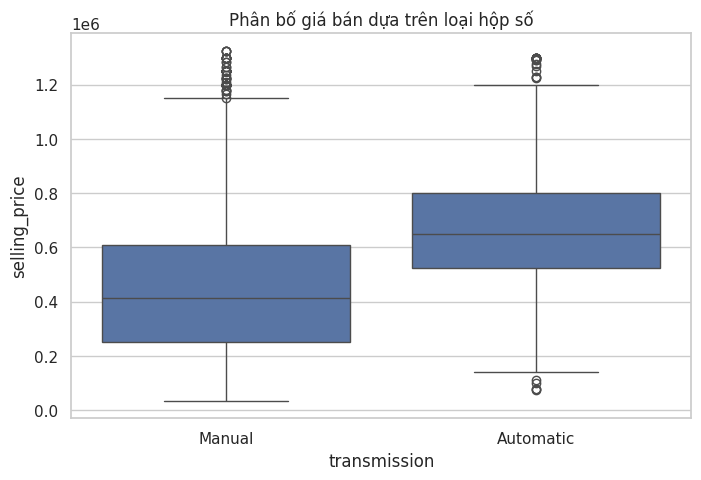
#### Hình 5.6 Biểu đồ Point phân bố giá dựa trên loại người bán

* 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. Người bán chuyên nghiệp và các đại lý thường có giá cao hơn do họ cung cấp xe đã được kiểm định kỹ càng, sửa chữa và thường đi kèm với bảo hành, giúp tăng sự yên tâm cho người mua. Những xe này cũng thường được trang bị tốt hơn và tuân thủ các tiêu chuẩn chất lượng nghiêm ngặt.

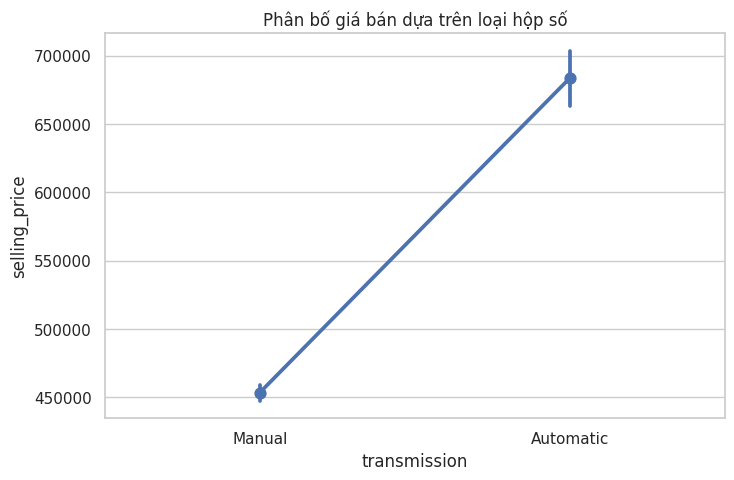
Ngược lại, người bán cá nhân có thể cung cấp giá thấp hơn vì họ không phải trả các chi phí hoạt động liên quan đến cơ sở kinh doanh và thường bán xe “như là”, mà không có bất kỳ cam kết bảo hành hoặc dịch vụ hậu mãi nào. Điều này có thể thu hút người mua tìm kiếm các giao dịch giá rẻ hơn nhưng cũng đi kèm với rủi ro cao hơn. Do đó, sự chênh lệch giá phản ánh sự cân bằng giữa chi phí và rủi ro mà người mua phải đánh giá khi lựa chọn mua xe từ các nguồn khác nhau.

## 5.4 Mối quan hệ giữa hộp số và giá bán



#### Hình 5.7 Biểu đồ Box phân bố giá dựa trên loại hộp số

* 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.
* Phân bố 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.

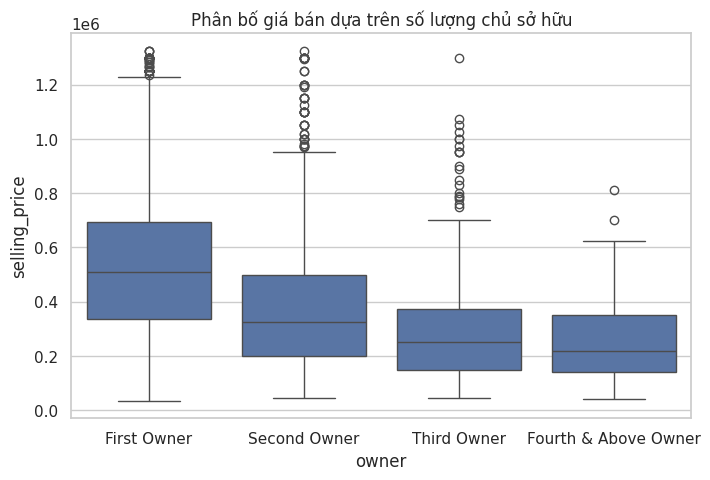


#### Hình 5.8 Biểu đồ Point phân bố giá dựa trên loại hộp số

* 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 ưa 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.

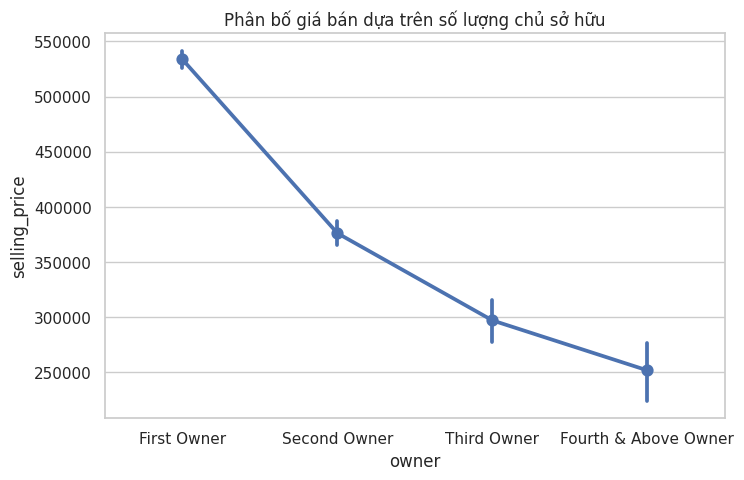
Xe có hộp số tự động thường có giá cao hơn so với xe hộp số sàn chủ yếu do chi phí sản xuất cao hơn và các công nghệ tiên tiến hơn được tích hợp trong hệ thống truyền động. Hộp số tự động cung cấp sự tiện lợi và thoải mái cho người lái, đặc biệt trong điều kiện giao thông đông đúc, vì nó loại bỏ nhu cầu phải thay đổi số thường xuyên, làm giảm sự mệt mỏi khi lái xe. Ngoài ra, xe tự động thường kết hợp với các tính năng an toàn và thoải mái cao cấp, cũng như công nghệ mới nhất. Sự phổ biến ngày càng tăng của xe tự động trong các phân khúc xe cao cấp và mới cũng làm tăng giá trị bán lại của chúng.

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



#### Hình 5.9 Biểu đồ Box phân bố giá dựa trên số lượng chủ đã sở hữu xe

* 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 cho 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 đi bán lại nhiều lần.

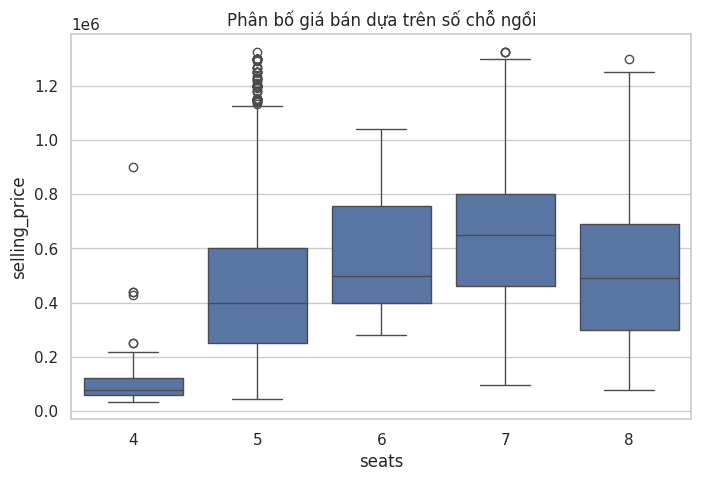


#### Hình 5.10 Biểu đồ Point phân bố giá dựa trên số lượng chủ đã sở hữu xe

* Xu hướng giá: Có một xu hướng rõ ràng là giá bán trung bình giảm theo số lượng chủ sở hữu tăng lên. Xe thuộc về “First Owner” có giá cao nhất, trong khi giá của xe giảm đáng kể khi chuyển sang “Second Owner”, và tiếp tục giảm ở “Third Owner” và “Fourth & Above Owner”.
* Giá trị giữ lại: Xe thuộc sở hữu đầu tiên giữ giá trị tốt hơn so với những xe đã qua nhiều tay chủ. Điều này có thể do giả định rằng xe ít chuyển nhượng hơn thường được bảo quản tốt hơn và ít hư hại hơn do không phải chịu nhiều quá trình mua bán và chuyển giao.
* Ảnh hưởng của lịch sử sở hữu: Sự sụt giảm giá trị có thể liên quan đến sự mất niềm tin vào chất lượng và tình trạng của xe khi nó đã qua nhiều chủ sở hữu, vì mỗi chủ mới có thể không duy trì xe với cùng một mức độ chăm sóc hoặc có thể đã sử dụng xe trong điều kiện khắc nghiệt hơn.

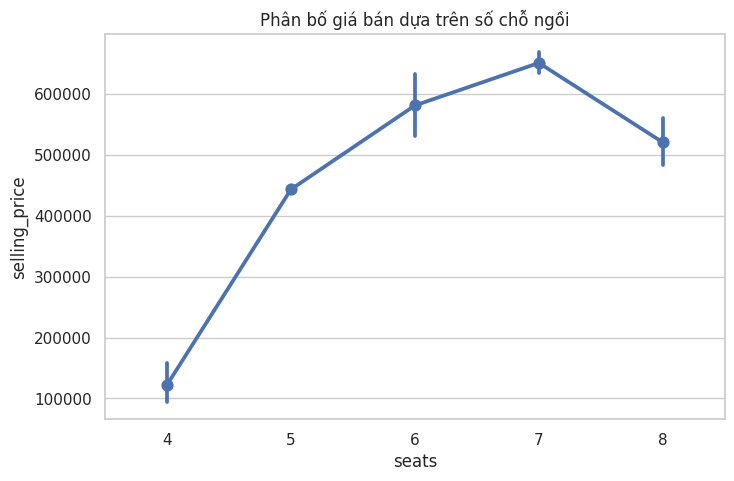
Giá bán của xe đã qua sử dụng có mối quan hệ nghịch với số lượng chủ sở hữu mà xe đó đã trải qua. Xe thuộc về chủ sở hữu đầu tiên thường có giá cao nhất, bởi xe này được giả định là đã được bảo quản tốt hơn và ít bị hao mòn do không phải chịu quá trình mua bán và chuyển nhượng nhiều lần. Khi xe chuyển qua tay nhiều chủ sở hữu hơn, giá bán thường giảm xuống. Điều này xảy ra bởi vì mỗi lần chuyển nhượng có thể dẫn đến sự suy giảm niềm tin vào chất lượng và tình trạng của xe, cũng như xe có thể đã không được bảo dưỡng cẩn thận như khi nó còn mới. Do đó, lịch sử sở hữu của xe là một yếu tố quan trọng ảnh hưởng đến giá trị của xe trên thị trường xe đã qua sử dụng, và người mua cần cân nhắc kỹ lưỡng điều này khi đánh giá xe.

## 5.6 Mối quan hệ giữa số chỗ ngồi và giá bán



#### Hình 5.11 Biểu đồ Box phân bố giá dựa trên số chỗ ngồi

* 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.



#### Hình 5.12 Biểu đồ Point phân bố giá dựa trên số chỗ ngồi

* Xu hướng giá: Có sự biến động rõ rệt về giá bán trung bình dựa trên số chỗ ngồi. Xe 7 chỗ ngồi có giá trung bình cao hơn so với các loại xe khác, phù hợp với nhu cầu của các gia đình lớn hoặc sử dụng cho mục đích thương mại, như xe shuttle hoặc minivan, thường được trang bị đầy đủ hơn.
* Xe 5 chỗ: Xe 5 chỗ, là loại phổ biến nhất trên thị trường, cho thấy giá trung bình ở mức vừa phải, phù hợp với nhu cầu chung của đa số người tiêu dùng.
* Xe 2 và 4 chỗ: Xe có 2 chỗ ngồi, thường là xe thể thao hoặc xe chuyên dụng, có giá trung bình cao hơn một chút so với xe 4 chỗ. Xe 4 chỗ ngồi lại có giá trung bình thấp hơn do chúng thường nhỏ gọn hơn và ít được ưa chuộng trong các gia đình.
* Xe 8 chỗ: Xe 8 chỗ ngồi cũng có giá trung bình cao, thể hiện nhu cầu cho xe rộng rãi và thoải mái cho các nhóm lớn hoặc gia đình đa thế hệ.

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. Ngược lại, xe có 5 chỗ là loại phổ biến nhất và có giá trung bình, phù hợp với nhu cầu chung của đa số hộ gia đình. Xe 2 và 4 chỗ, thường nhỏ hơn và ít phức tạp hơn về mặt thiết kế, có giá thấp hơn nhưng xe 2 chỗ lại có thể đạt giá cao nếu là xe thể thao cao cấp. Tóm lại, số chỗ ngồi không chỉ phản ánh kích thước và sức chứa của xe mà còn ảnh hưởng đến giá bán dựa trên công năng và mục đích sử dụng của xe, từ đó đáp ứng các nhu cầu cụ thể của người tiêu dùng.

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

Để củng cố những khám phá trước đó và cung cấp bằng chứng thống kê chắc chắn, chúng ta sẽ tiến hành kiểm tra giả thuyết về các biến được chọn:

1. Đưa ra giả thuyết H0
2. Tính toán, kiếm định giả thuyết:

* Bước 1: Ta sẽ bắt đầu bằng việc tính toán trung bình mẫu (mean) cho mỗi nhóm. Trung bình này phản ánh giá trị trung bình của biến số đang được quan sát cho mỗi nhóm.
* Bước 2: Độ lệch chuẩn mẫu (standard deviation) được tính toán để đo lường mức độ biến thiên hoặc phân tán của các giá trị trong từng nhóm so với trung bình của chúng.
* Bước 3: Sai số chuẩn của sự khác biệt trung bình (standard error of the mean difference) được tính toán dựa trên độ lệch chuẩn và kích thước của từng nhóm. Sai số chuẩn là một ước tính về mức độ không chắc chắn liên quan đến hiệu của trung bình mẫu.
* Bước 4: Tính toán t-statistics bằng cách lấy sự khác biệt giữa hai trung bình mẫu và chia cho sai số chuẩn của sự khác biệt trung bình. T-statistic cho ta biết sự khác biệt giữa hai nhóm là bao nhiêu so với sai số của sự khác biệt đó.
* Bước 5: Tính toán p-value dựa trên t-statistic và kích thước mẫu của hai nhóm. P-value là xác suất của việc quan sát sự khác biệt ít nhất như thực tế (hoặc lớn hơn mức ý nghĩa) sẽ giả định rằng giả thuyết không (H0) là đúng. P-value càng nhỏ, bằng chứng chống lại H0 và ủng hộ sự khác biệt càng mạnh (H1 là đúng - có ý nghĩa thống kê).

Có thể áp dụng các phương pháp kiểm định khác như mô hình hồi quy tuyến tính, nội suy để mô hình hóa quan hệ…

1. Đưa ra kết luận về giả thuyết H1

## 6.1 Giả thuyết kiểm định giá bán theo loại nhiên liệu

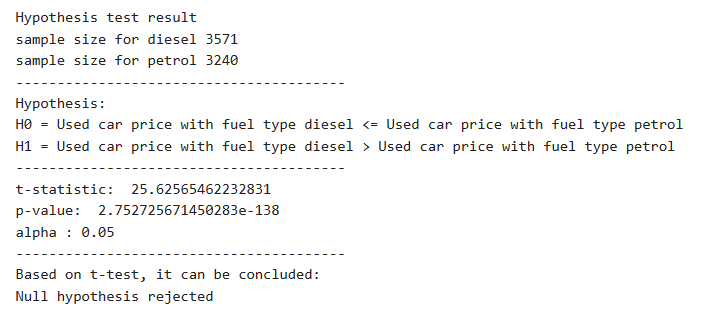
* Mục đích: Kiểm định sự khác biệt về giá bán trung bình giữa xe sử dụng nhiên liệu diesel và petrol, giúp hiểu rõ hơn về ảnh hưởng của loại nhiên liệu đến giá bán xe.
* Nhận thức vấn đề: Nhận định ban đầu về giả thuyết là xe chạy điêzen có giá bán cao hơn xe chạy xăng do các yếu tố như hiệu quả nhiên 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.

- Đầu tiên, dữ liệu được chia thành hai nhóm dựa trên loại nhiên liệu: nhóm diesel và nhóm petrol của tập dữ liệu đã được xử lí:

+ 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.

- Ở đây chúng ta sẽ giả định rằng các phương sai của hai nhóm là không đồng nhất (điều này thường phù hợp hơn khi làm việc với dữ liệu thực tế) để thực hiện việc kiểm định dựa trên các giá trị của hai nhóm mẫu. Ta sẽ đi sâu vào phân tích thống kê đã được tiến hành để so sánh giá bán giữa xe dùng nhiên liệu diesel và petrol, sử dụng kiểm định t độc lập với giả định trên:



#### Hình 6.1 Kết quả từ kiểm định giá bán theo loại nhiên liệu

- 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

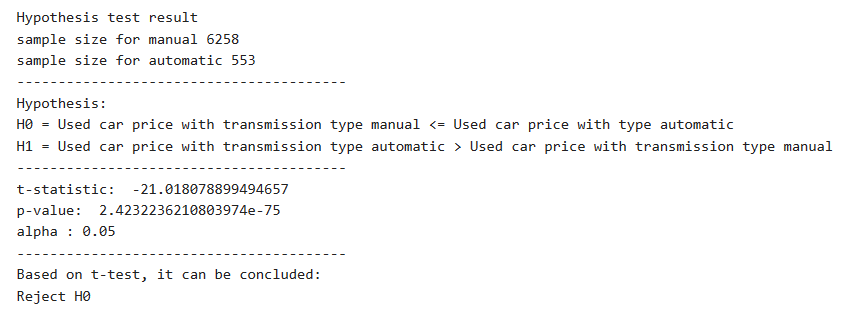
* Mục đích: Kiểm tra liệu có sự khác biệt về giá bán trung bình giữa xe hộp số tự động và hộp số sàn. Kiểm định này giúp xác định ảnh hưởng của loại hộp số đến giá bán xe, vốn là thông tin quan trọng đối với người mua và người bán trong việc định giá xe
* Nhận thức vấn đề: Xe hộp số tự động thường có 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.

- Đầu tiên, ta cũng sẽ phân loại dữ liệu thành hai nhóm dựa trên loại hộp số: hộp số sàn (manual) và tự động (automatic). Và giá bán (selling\_price) của xe được lấy ra từ mỗi nhóm để kiểm định.

+ 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.

- Ở đây kiểm định T-test độc lập được thực hiện giữa hai nhóm để xác định liệu có sự khác biệt đáng kể về giá bán giữa hai loại hộp số hay không với giả định rằng không có sự đồng nhất của phương sai giữa hai nhóm:



#### Hình 6.2 Kết quả từ kiểm định giá bán theo loại hộp số

- 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: Căn 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. Điều này cung cấp bằng chứng quan trọng cho các bên liên quan trong ngành xe cũ để họ có thể đưa ra các quyết định mua bán và định giá dựa trên cơ sở vững chắc và có thông tin.

## 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 hoặc nhiều nhóm với nhau về giá bán, phù hợp khi có nhiều hơn hai nhóm cần so sánh.

- Đầu tiên, ta cũng sẽ phân loại dữ liệu và đặt giả thuyết:

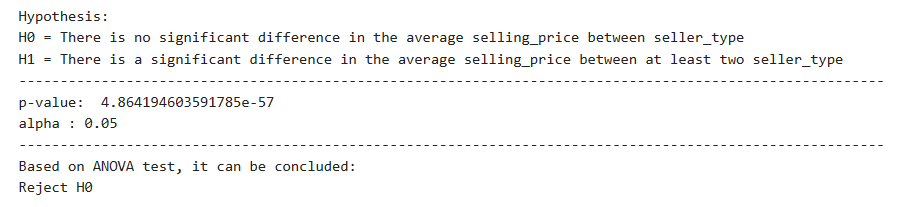
+ Giả thuyết Không (H0): Không có sự khác biệt đáng kể về giá bán trung bình giữa các loại người bán (có khoảng tin cậy là 95% và mức ý nghĩa là 5%).

+ Giả thuyết Nghịch (H1): Có sự khác biệt đáng kể về giá bán trung bình giữa ít nhất hai loại người bán.

- Sau đó là sử dụng mô hình hồi quy tuyến tính nhỏ nhất (OLS) để mô hình hoá mối quan hệ giữa selling\_price (giá bán) như là biến phụ thuộc và seller\_type (loại người bán) như là biến độc lập.

- Tiếp theo là tìm ra các hệ số bội số tốt nhất (coefficients) cho mô hình dựa trên cơ sở phương pháp Least Squares giúp giảm thiểu tổng bình phương các sai số giữa giá trị dự báo và giá trị thực tế.

- Cuối cùng là tính toán lấy giá trị p-value từ bảng ANOVA, đây là xác suất của việc quan sát được sự khác biệt giữa các nhóm seller\_type:



#### Hình 6.3 Kết quả từ kiểm định giá bán theo loại người bán

- Kết luận: Căn 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. Dựa trên kết quả này, các nhà kinh doanh và nhà phân tích có thể điều chỉnh chiến lược tiếp thị và bán hàng, hoặc cung cấp thông tin hữu ích cho người mua về ảnh hưởng của loại người bán đến giá bán của xe.

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

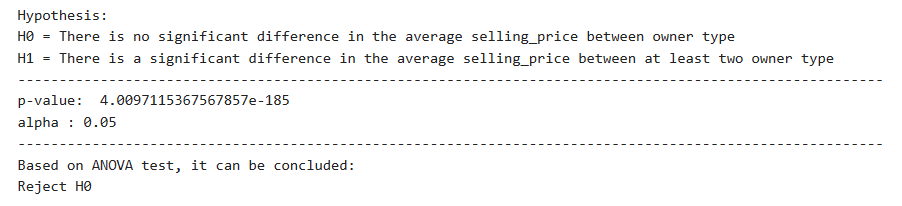
* Mục đích: So sánh giá bán giữa các xe với số lượng chủ sở hữu trước khác nhau.
* Nhận thức vấn đề: Các xe có ít chủ sở hữu trước thường giữ giá tốt hơn. ANOVA giúp kiểm định sự khác biệt về giá bán tổng cộng giữa các nhóm xe dựa vào số lượng chủ sở hữa trước đó, thích hợp khi so sánh nhiều hơn hai nhóm.

- Đầu tiên, ta cũng sẽ phân loại dữ liệu và đặt giả thuyết:

+ 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.

- Tiếp theo là sử dụng mô hình hồi quy tuyến tính nhỏ nhất (OLS) được xây dựng sử dụng biến selling\_price làm biến phụ thuộc và biến owner làm biến độc lập để tính toán giá trị p-value.



#### Hình 6.4 Kết quả từ kiểm định giá bán theo số lượng người đã từng sở hữu

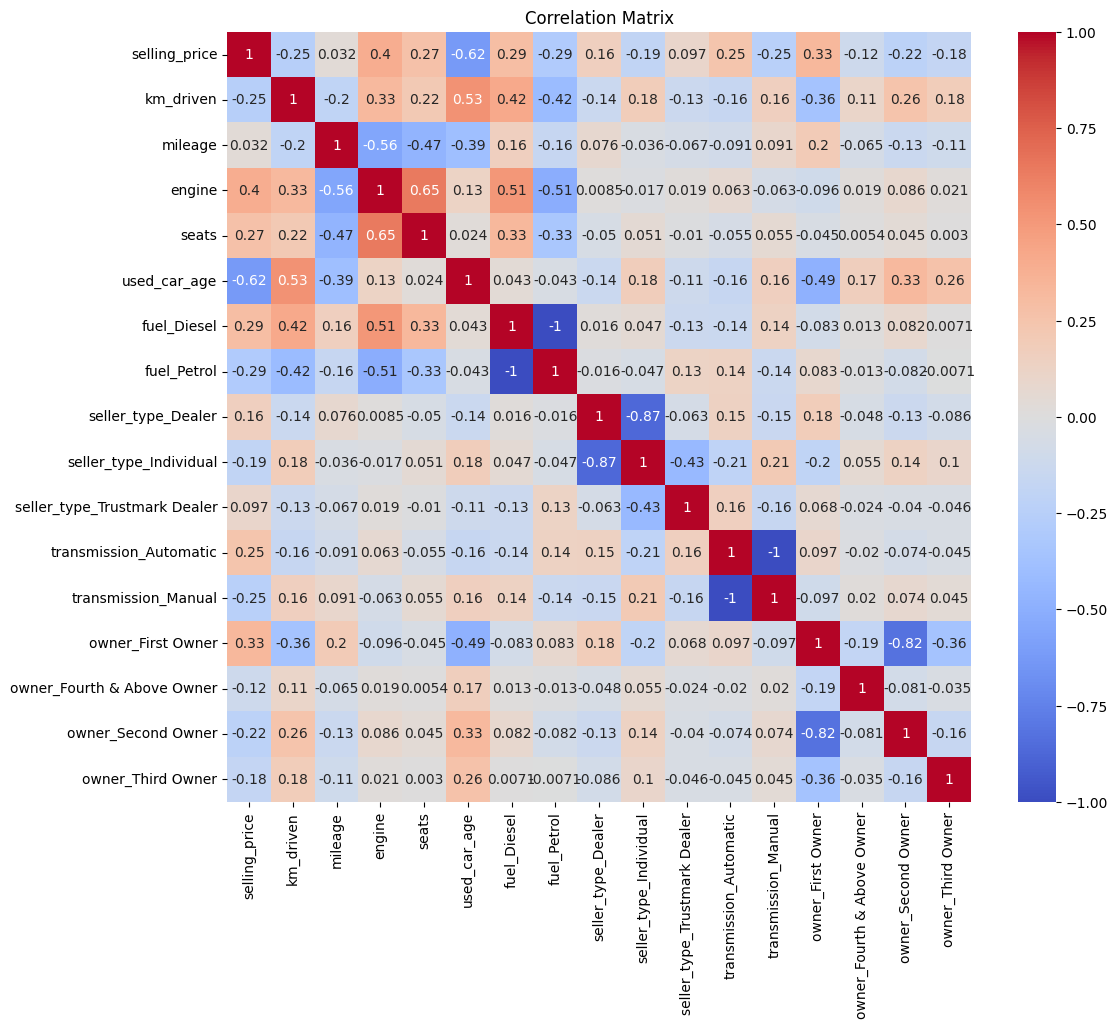
- Kết luận: Căn cứ vào kết quả của kiểm định, giá trị p-value thấp hơn mức ý nghĩa đã định (alpha = 0.05), điều này cho thấy có sự khác biệt đáng kể về giá bán trung bình giữa các nhóm chủ sở hữu khác nhau. Dựa trên kết quả này, ta nên tập trung vào việc nâng cao giá trị của xe thông qua các biện pháp bảo trì hoặc cải thiện cách truyền thông các đặc điểm xe dựa trên lịch sử sở hữu sẽ giúp nâng cao hiệu quả kinh doanh và sự hài lòng của khách hàng.

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

## 7.1 Phân tích mối tương quan giữa các biến

Ma 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.



#### Hình 7.1 Ma trận tương quan

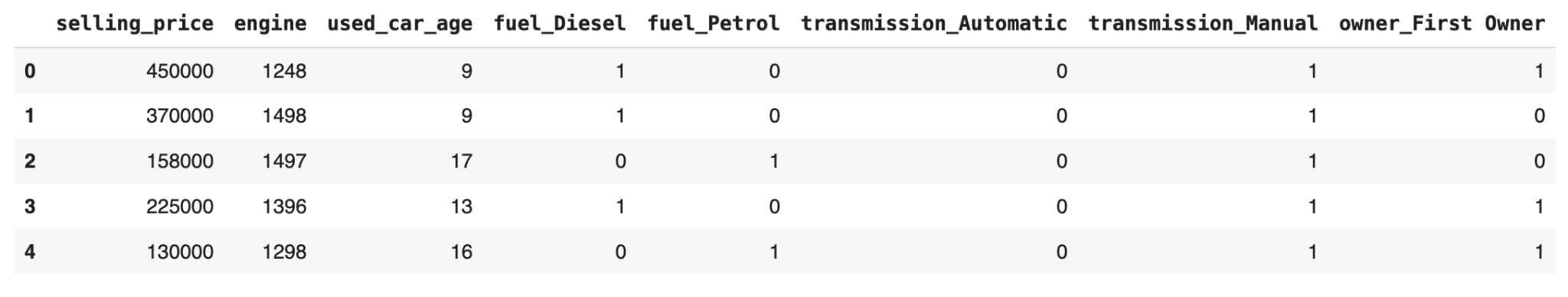
## 

## 

## 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). Việc này giúp tập trung vào những biến có tác động lớn hơn đến giá xe, giảm thiểu nhiễu và tăng hiệu quả tính toán của mô hình.

## 7.3 Xây dựng mô hình hồi quy tuyến tính

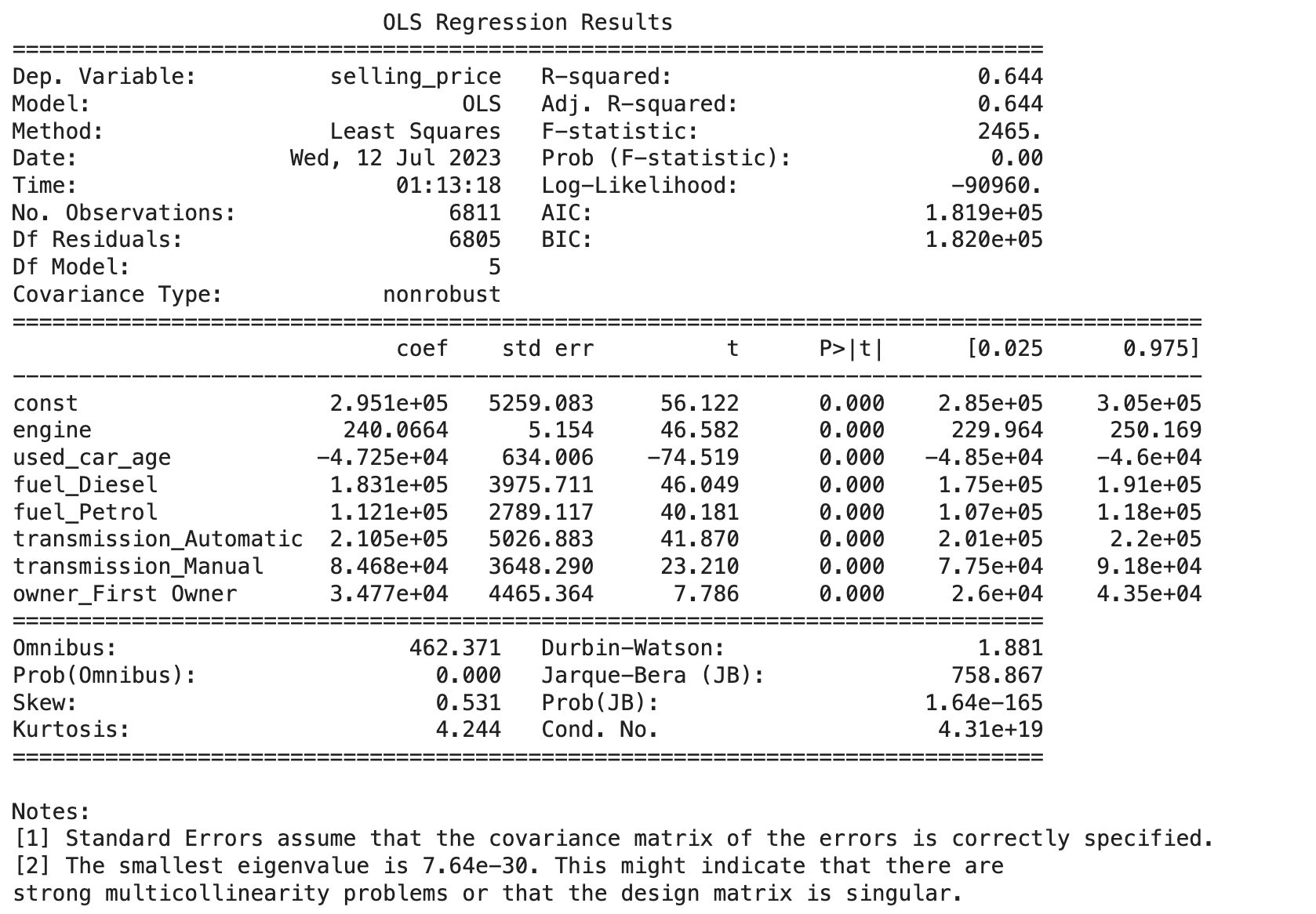


#### Hình 7.2 Kết quả sau khi bỏ các cột có hệ số tương quan thấp

Mô hình hồi quy tuyến tính được thiết kế để dự đoán giá xe dựa trên nhiều yếu tố khác nhau. Các biến được chọn dựa trên phân tích tương quan để đảm bảo rằng chúng có ảnh hưởng đáng kể đến giá bán xe:

* Biến động cơ (engine): Đo lường dung tích động cơ của xe (đơn vị cc). Mỗi đơn vị tăng trong dung tích động cơ được dự đoán làm tăng giá xe khoảng 240 đơn vị tiền tệ. Điều này phản ánh rằng xe có động cơ lớn hơn thường có giá trị cao hơn trên thị trường.
* Tuổi xe (used\_car\_age): Biến này được tính bằng số năm từ thời điểm xe được sản xuất cho đến nay. Mỗi năm tăng tuổi xe được dự đoán làm giảm giá xe khoảng 4,752 đơn vị tiền tệ, phản ánh giá trị khấu hao theo thời gian của xe.
* Nhiên liệu (fuel\_Diesel và fuel\_Petrol): Các xe sử dụng diesel và xăng được so sánh với các loại nhiên liệu khác. Xe diesel dự đoán cao hơn 183,100 đơn vị tiền tệ so với các loại nhiên liệu khác, trong khi xe xăng cao hơn 112,100 đơn vị tiền tệ, cho thấy sự ưa chuộng và giá trị cao hơn của xe diesel và xăng trên thị trường.
* Hộp số (transmission\_Automatic và transmission\_Manual): Xe tự động dự đoán có giá cao hơn 210,500 đơn vị tiền tệ so với xe sử dụng hộp số sàn.
* Chủ sở hữu đầu tiên (owner\_First Owner): Xe vẫn còn trong tay chủ sở hữu đầu tiên dự đoán cao hơn 347,700 đơn vị tiền tệ so với các xe đã chuyển nhượng, cho thấy giá trị cao hơn của xe không qua sử dụng nhiều.

## 7.4 Kết quả từ mô hình hồi quy OLS



#### Hình 7.3 Kết quả mô hình hồi quy OLS

Mô hình OLS cho thấy:

* R-squared: 0.644, cho thấy mô hình giải thích được 64.4% sự biến thiên của giá xe.
* P-value: Tất cả các biến trong mô hình đều có P-value nhỏ hơn 0.05, chứng tỏ chúng có ảnh hưởng đáng kể và ý nghĩa thống kê tới giá xe.
* Coefficients: Đáng chú ý là engine và used\_car\_age có hệ số dương và âm tương ứng, phản ánh ảnh hưởng trực tiếp đến giá xe.

## 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. Các kết quả từ mô hình hỗ trợ việc hiểu sâu hơn về yếu tố ảnh hưởng đến giá xe và cung cấp cơ sở vững chắc cho các nghiên cứu tiếp theo trong lĩnh vực này.

# 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. Xe có số kilomet thấp và các mẫu xe mới hơn có giá bán cao hơn. Xe sang trọng và xe điện có xu hướng giữ giá trị tốt hơn so với xe sử dụng động cơ đốt trong thông thường.
* 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 khấu 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 khấu 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

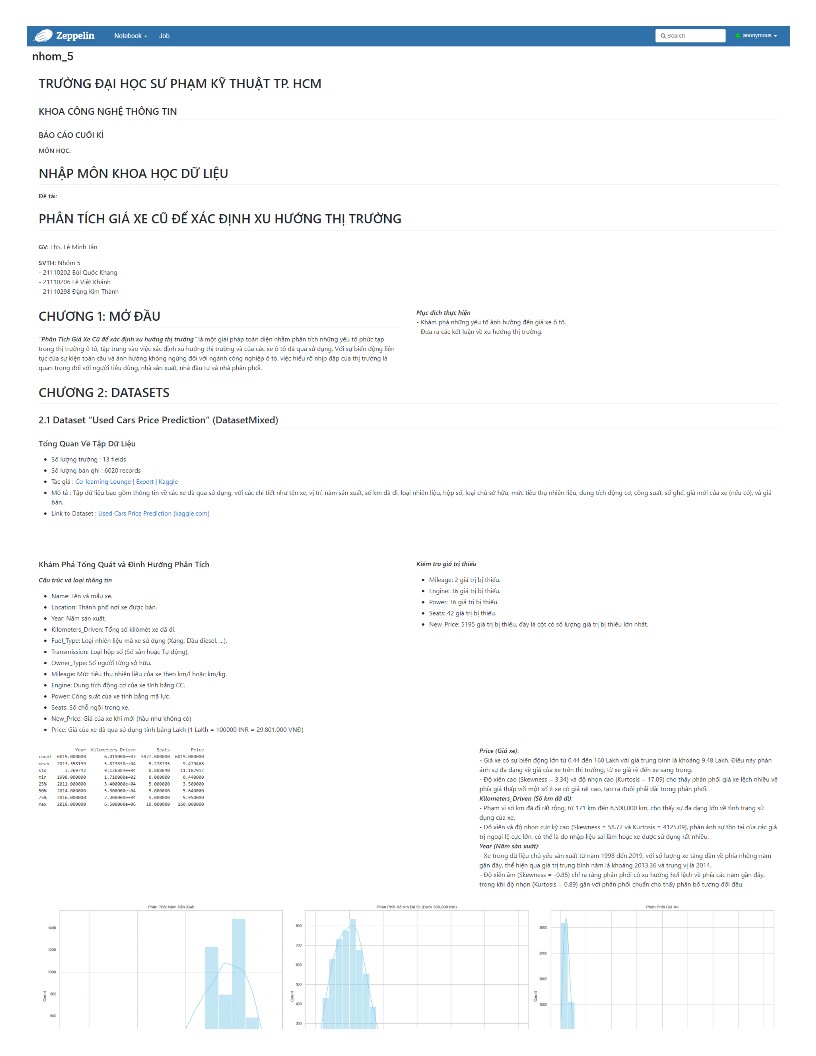
**Các vấn đề tiềm tàng**

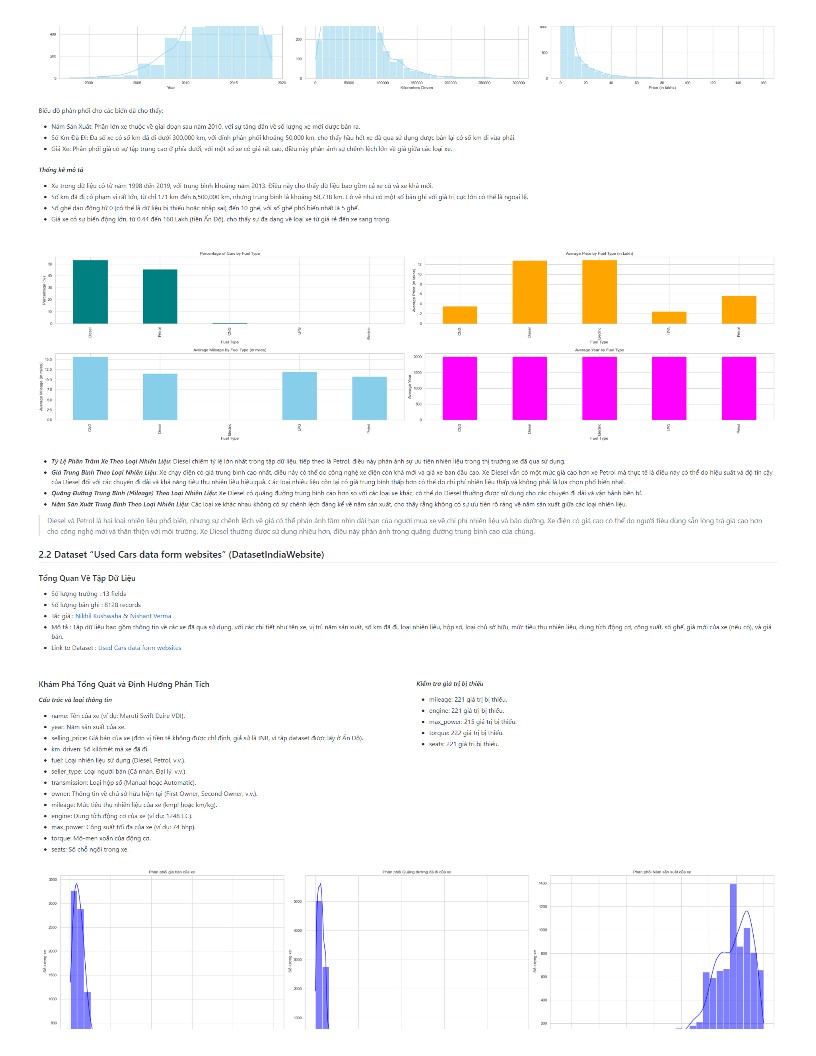
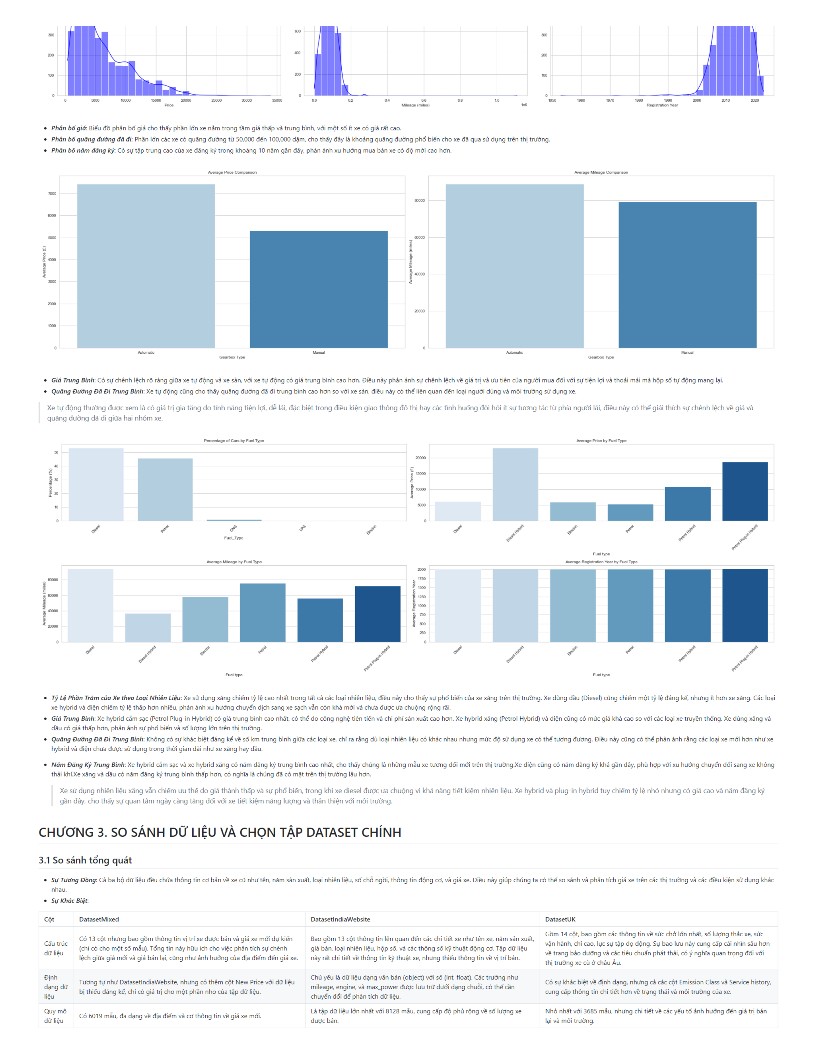
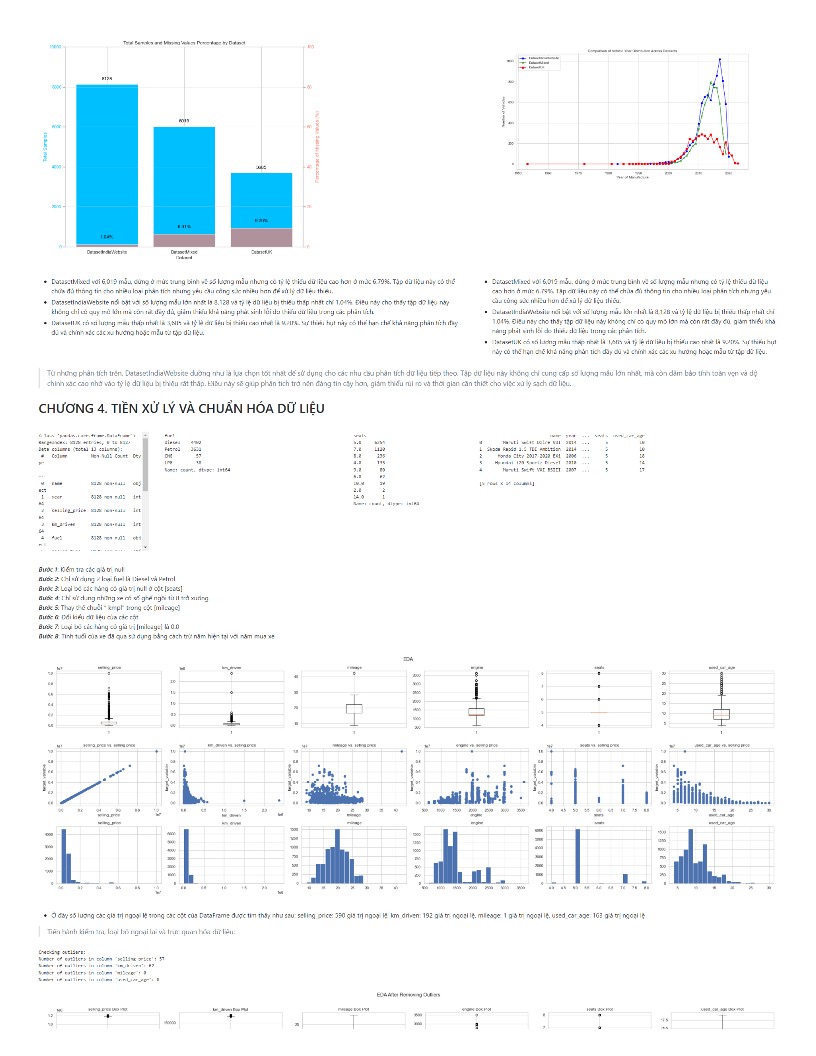
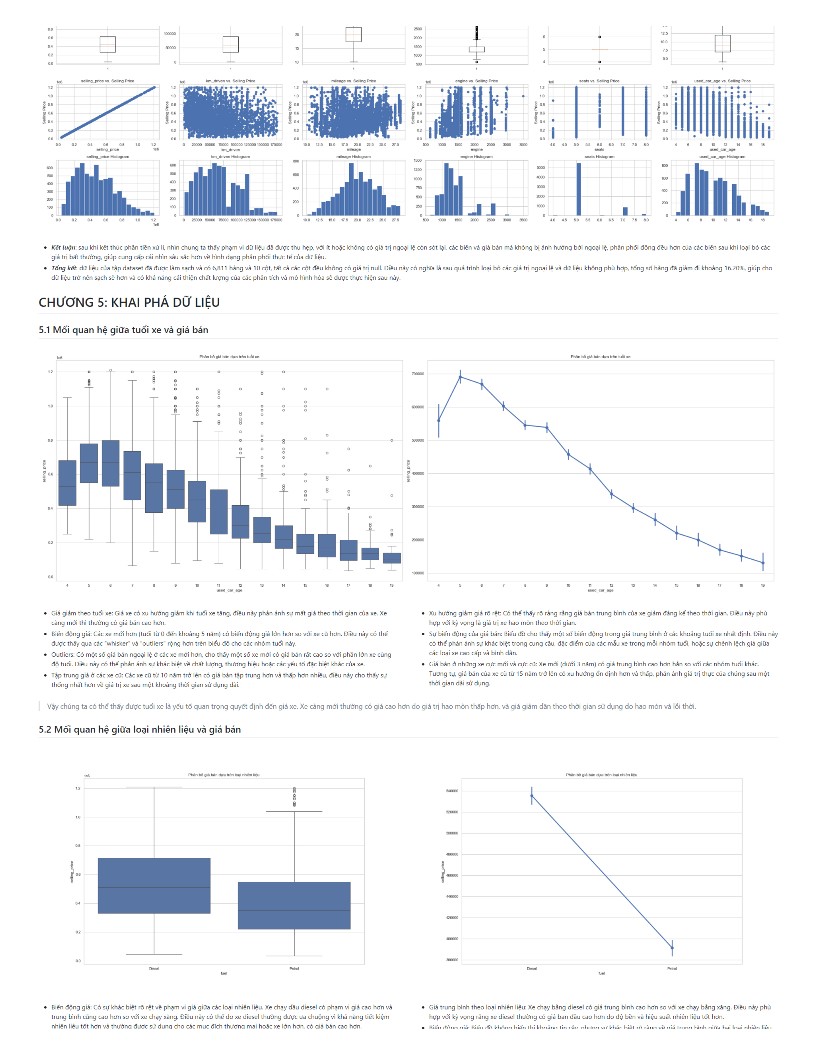
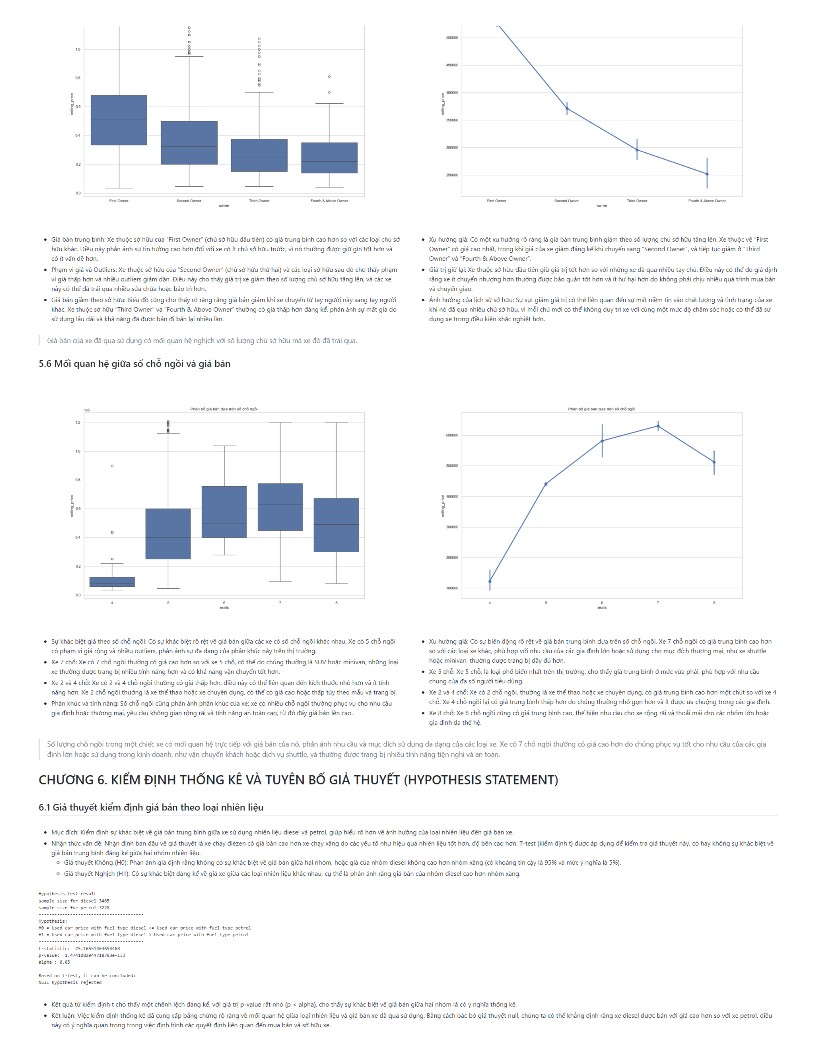
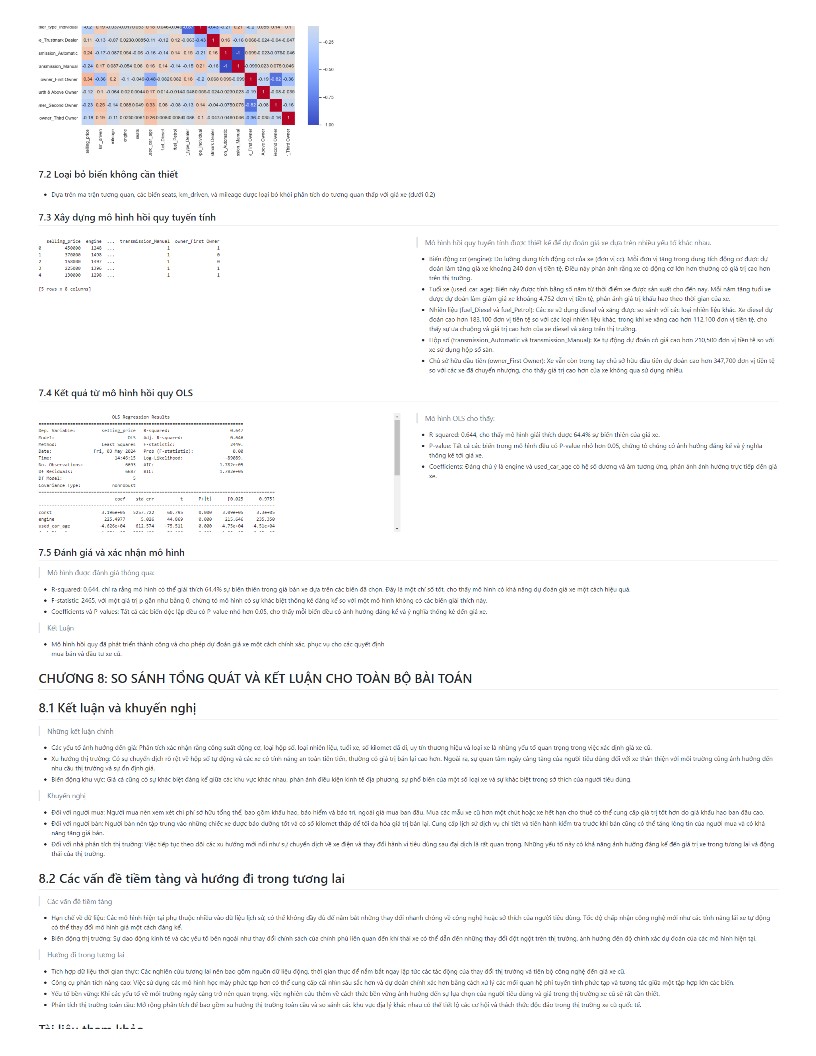
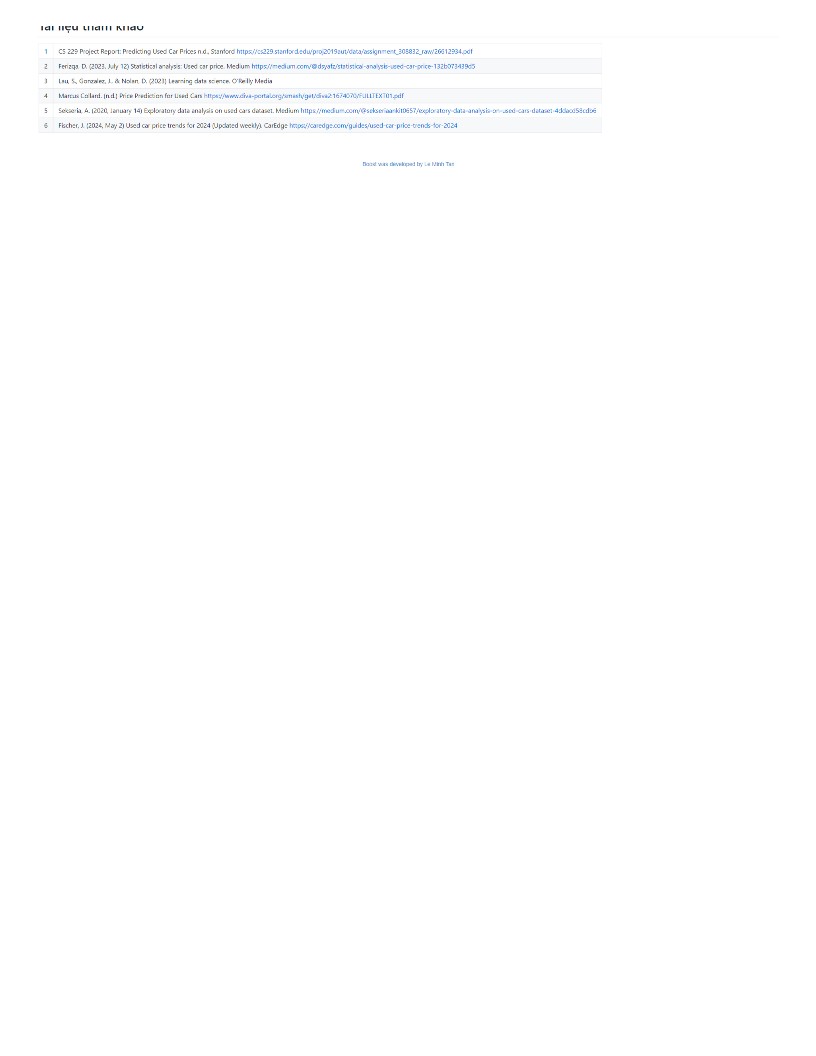
* Hạn chế về dữ liệu: Các mô hình hiện tại phụ thuộc nhiều vào dữ liệu lịch sử, có thể không đầy đủ để nắm bắt những thay đổi nhanh chóng về công nghệ hoặc sở thích của người tiêu dùng. Tốc độ chấp nhận công nghệ mới như các tính năng lái xe tự động có thể thay đổi mô hình giá một cách đáng kể.
* Biến 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.

# CHƯƠNG 9: HÌNH ẢNH BÁO CÁO TRONG ZEPPELIN



# 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.  1. Marcus Collard. (n.d.). *Price Prediction for Used Cars*.  <https://www.diva-portal.org/smash/get/diva2:1674070/FULLTEXT01.pdf> 2. 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> 3. Fischer, J. (2024, May 2). Used car price trends for 2024 (Updated weekly). CarEdge.  https://caredge.com/guides/used-car-price-trends-for-2024 |