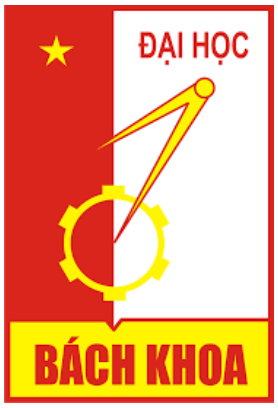
ĐẠI HỌC BÁCH KHOA HÀ NỘI

TRƯỜNG CNTT&TT

-----□□📖□□-----



BÁO CÁO NHẬP MÔN KHOA HỌC DỮ LIỆU

Đề tài: Dự đoán giá xe ô tô cũ

Mã học phần IT4930 Mã lớp 144938

Giảng viên hướng dẫn: TS. Trần Việt Trung

TS. Bùi Thị Mai Anh

TS. Nguyễn Thị Oanh

PGS. TS. Thân Quang Khoát

Nhóm sinh viên thực hiện: Nhóm 1

Họ tên sinh viên	MSSV
Trần Đức Hân	20204649
Nguyễn Trung Hiếu	20204552
Lê Quang Vũ	20204624
Bùi Đức Đăng	20200147
Cao Thành Huy	20204656

Hà Nội 12/2023

Mục lục

Mục lục hình ảnh

Phần 2: Thu thập dữ liệu

Phần 3: Phân tích và trực quan hóa dữ liệu

Phần 4: Mô hình

Phần 1: Phân tích tổng quan

1.1 Giới thiệu chung

Thị trường ô tô cũ ở Việt Nam đang phát triển mạnh mẽ, nhưng việc định giá xe cũ là một vấn đề khó khăn, đòi hỏi kinh nghiệm và kiến thức chuyên môn. Điều này dẫn đến tình trạng mua bán xe cũ thường bị chênh lệch giá, gây thiệt thòi cho cả người bán và người mua. Đề tài "Dự đoán giá xe ô tô cũ" nhằm giải quyết vấn đề này bằng cách xây dựng một mô hình dự đoán giá xe ô tô cũ dựa trên các thông số của xe, chẳng hạn như hãng xe, năm sản xuất, số km đã đi, số chỗ ngồi, màu xe, ... Mô hình dự đoán sẽ được xây dựng dựa trên các phương pháp học máy. Đề tài có ý nghĩa thực tiễn cao, giúp người bán và người mua xe cũ có thể định giá xe một cách chính xác hơn, giảm thiểu rủi ro và tránh bị bán lỗ mua hớ đồng thời cũng giúp các công ty dự đoán giá xe cải thiện chất lượng phục vụ tốt hơn.

1.2 Phân tích bài toán

Tên bài toán: Bài toán dự đoán giá xe cũ

Mục tiêu: Dựa vào các thông số của xe như hãng xe, năm sản xuất, số km đã đi , số chỗ ngồi, màu xe , … để đưa ra dự đoán về giá của xe.

Phương pháp: đưa về bài toán hồi quy tuyến tính

Input: vector chứa thông tin các trường dữ liệu thông số của xe X[x1, x2, x3,…, xn]

Output: Nhãn cho input (Y) cụ thể là giá xe

Các bước tiến hành:

- Thu thập dữ liệu từ các trang web rao bán ô tô
- Tiến hành phân tích, trực quan hóa dữ liệu.
- Tiền xử lý dữ liệu
- Modeling: Sử dụng các mô hình khác nhau để huấn luyện
- Đánh giá chất lượng mô hình.
- Triển khai hệ thống đầy đủ.

Phần 2: Thu thập dữ liệu

2.1 Công nghệ sử dụng

2.1.1 Selenium

Selenium là một bộ công cụ tự động hóa ứng dụng web nguồn mở. Nó được sử dụng để kiểm tra tự động, phát triển web và các mục đích khác. Selenium cung cấp một API cho nhiều ngôn ngữ lập trình, bao gồm Java, Python, C#, JavaScript và Ruby.

Các tính năng chính của Selenium:

- Hỗ trợ nhiều ngôn ngữ lập trình: Selenium cung cấp một API cho nhiều ngôn ngữ lập trình phổ biến, giúp các nhà phát triển dễ dàng sử dụng nó trong các dự án của mình.
- Hỗ trợ nhiều trình duyệt web: Selenium hỗ trợ nhiều trình duyệt web phổ biến, bao gồm Chrome, Firefox, Edge và Safari. Điều này giúp các nhà phát triển có thể kiểm tra ứng dụng của mình trên nhiều trình duyệt khác nhau.
- Hỗ trợ nhiều loại kiểm tra: Selenium hỗ trợ nhiều loại kiểm tra, bao gồm kiểm tra hộp đen, kiểm tra hộp trắng và kiểm tra chức năng. Điều này giúp các nhà phát triển có thể kiểm tra ứng dụng của mình theo nhiều cách khác nhau.
- Hỗ trợ nhiều loại báo cáo: Selenium hỗ trợ nhiều loại báo cáo, bao gồm báo cáo văn bản, báo cáo HTML và báo cáo XML. Điều này giúp các nhà phát triển dễ dàng theo dõi kết quả của các bài kiểm tra.

Ứng dụng của Selenium:

- Kiểm tra tự động ứng dụng web: Selenium là một công cụ kiểm tra tự động mạnh mẽ, được sử dụng để kiểm tra ứng dụng web. Nó có thể tự động thực hiện các tác vụ như nhập văn bản, nhấp chuột và cuộn trang.
- Phát triển web: Selenium cũng có thể được sử dụng để phát triển web. Nó có thể được sử dụng để tự động hóa các tác vụ như tạo và chỉnh sửa trang web.
- Tự động hóa các tác vụ web: Selenium có thể được sử dụng để tự động hóa các tác vụ web khác nhau, chẳng hạn như đăng nhập vào trang web, thanh toán trực tuyến và đặt lịch hẹn.
- Tự động hóa các tác vụ dữ liệu: Selenium có thể được sử dụng để tự động hóa các tác vụ dữ liệu khác nhau, chẳng hạn như nhập dữ liệu vào cơ sở dữ liệu và tải xuống dữ liệu từ trang web.

2.1.2 MinIO

MinIO là một dịch vụ lưu trữ đối tượng đám mây mã nguồn mở, được thiết kế để cung cấp hiệu suất cao, độ tin cậy và khả năng mở rộng. MinIO được xây dựng dựa trên mã nguồn của Amazon S3, và tương thích API với Amazon S3.

Các tính năng chính của MinIO

- Hiệu suất cao: MinIO được thiết kế để cung cấp hiệu suất cao, với khả năng xử lý hàng triệu yêu cầu mỗi giây.
- Độ tin cậy: MinIO được xây dựng trên các nguyên tắc thiết kế chịu lỗi, giúp đảm bảo tính sẵn sàng cao.
- Khả năng mở rộng: MinIO có thể được mở rộng theo chiều ngang để đáp ứng nhu cầu lưu trữ ngày càng tăng.

Các ứng dụng của MinIO

MinIO có thể được sử dụng cho nhiều ứng dụng khác nhau, bao gồm:

- Lưu trữ đám mây: MinIO có thể được sử dụng để lưu trữ dữ liệu đám mây, chẳng hạn như dữ liệu ứng dụng, dữ liệu phân tích, và dữ liệu sao lưu.
- Trực tuyến: MinIO có thể được sử dụng để lưu trữ nội dung trực tuyến, chẳng hạn như hình ảnh, video, và tài liệu.
- Sự kiện: MinIO có thể được sử dụng để lưu trữ dữ liệu sự kiện, chẳng hạn như dữ liệu từ các cảm biến và thiết bị IoT.

2.1.3 MongoDB

MongoDB là một cơ sở dữ liệu NoSQL nguồn mở. Nó sử dụng mô hình dữ liệu tài liệu để lưu trữ dữ liệu. MongoDB được sử dụng rộng rãi trong các ứng dụng web, ứng dụng di động và ứng dụng dữ liệu lớn.

Các tính năng chính của MongoDB:

- Sử dụng mô hình dữ liệu tài liệu: MongoDB sử dụng mô hình dữ liệu tài liệu, trong đó dữ liệu được lưu trữ dưới dạng các tài liệu JSON. Điều này giúp MongoDB dễ sử dụng và quản lý hơn các cơ sở dữ liệu quan hệ truyền thống.
- Hỗ trợ lưu trữ dữ liệu lớn: MongoDB hỗ trợ lưu trữ dữ liệu lớn, lên đến hàng petabyte.

- Hỗ trợ truy cập dữ liệu nhanh chóng: MongoDB sử dụng một số kỹ thuật để tăng tốc độ truy cập dữ liệu, bao gồm phân cấp dữ liệu và cache.
- Hỗ trợ tính sẵn sàng cao: MongoDB được thiết kế để có tính sẵn sàng cao, ngay cả khi một số máy chủ bị lỗi.

Ứng dụng của MongoDB:

- Lưu trữ dữ liệu web: MongoDB thường được sử dụng để lưu trữ dữ liệu web, chẳng hạn như dữ liệu người dùng, dữ liệu sản phẩm và dữ liệu giao dịch.
- Lưu trữ dữ liệu di động: MongoDB cũng có thể được sử dụng để lưu trữ dữ liệu di động, chẳng hạn như dữ liệu vị trí, dữ liệu lịch sử và dữ liệu giao dịch.
- Lưu trữ dữ liệu dữ liệu lớn: MongoDB có thể được sử dụng để lưu trữ dữ liệu dữ liệu lớn, chẳng hạn như dữ liệu cảm biến, dữ liệu hình ảnh và dữ liệu video.

2.1.4 Airflow

Airflow là một công cụ tự động hóa dòng công việc. Nó được sử dụng để lên lịch và quản lý các tác vụ tự động. Airflow được sử dụng rộng rãi trong các ứng dụng phân tích dữ liệu, xử lý dữ liệu lớn và máy học.

Các tính năng chính của Airflow:

- Lên lịch và quản lý các tác vụ tự động: Airflow cho phép người dùng lên lịch và quản lý các tác vụ tự động, chẳng hạn như chạy các công việc phân tích dữ liệu, xử lý dữ liệu lớn và đào tạo mô hình máy học.
- Hỗ trợ nhiều loại tác vụ: Airflow hỗ trợ nhiều loại tác vụ, bao gồm tác vụ Python, tác vụ shell và tác vụ HTTP.
- Hỗ trợ nhiều loại nguồn dữ liệu: Airflow hỗ trợ nhiều loại nguồn dữ liệu, bao gồm cơ sở dữ liệu, file và API.
- Hỗ trợ tích hợp với các công cụ khác: Airflow có thể được tích hợp với các công cụ khác, chẳng hạn như Hadoop, Spark và Kubernetes.

Ứng dụng của Airflow:

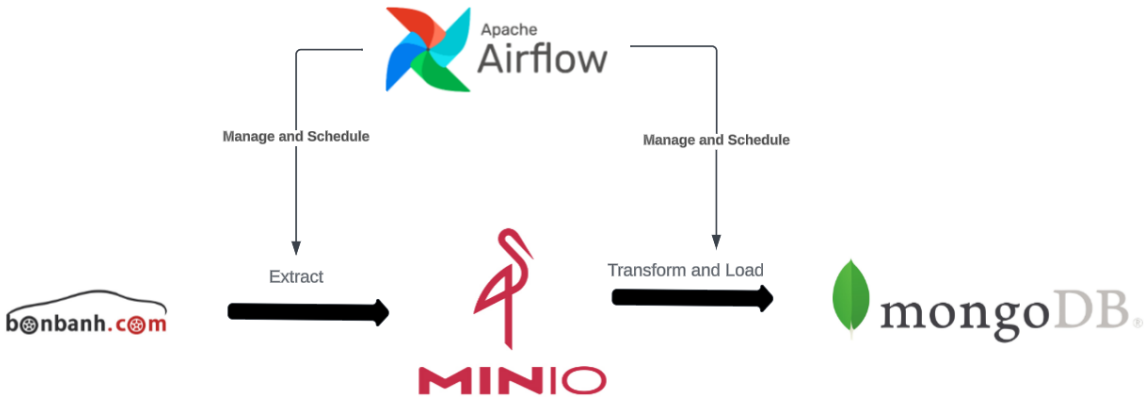
- Lên lịch các tác vụ phân tích dữ liệu: Airflow có thể được sử dụng để lên lịch các tác vụ phân tích dữ liệu, chẳng hạn như chạy các công việc trích xuất, chuyển đổi và tải (ETL), phân tích dữ liệu và báo cáo dữ liệu.

- Lên lịch các tác vụ xử lý dữ liệu lớn: Airflow có thể được sử dụng để lên lịch các tác vụ xử lý dữ liệu lớn, chẳng hạn như chạy các công việc phân tích dữ liệu lớn, xử lý dữ liệu thời gian thực và đào tạo mô hình máy học.
- Lên lịch các tác vụ máy học: Airflow có thể được sử dụng để lên lịch các tác vụ máy học, chẳng hạn như chạy các công việc đào tạo mô hình, triển khai mô hình và đánh giá mô hình.

2.2 Thu thập dữ liệu

2.2.1 Tiến trình thu thập và lưu trữ

Luồng thu thập và lưu trữ dữ liệu thể hiện như sau



Hình 2. 1 Luồng thu thập dữ liệu

Dữ liệu được thu thập trong đồ án môn học chủ yếu được lấy từ trang web Bonbanh.com. Là một trang web chuyên về mua bán ô tô và xe máy tại Việt Nam, Bonbanh.com cung cấp một nền tảng trực tuyến cho người dùng đăng tin mua bán ô tô kể cả xe mới và xe cũ với nhiều hãng xe, kiểu dáng xe, giá cả đa dạng. Người dùng có thể đăng thông tin về xe của mình để bán hoặc tìm kiếm xe cần mua trên trang web. Quá trình thu thập dữ liệu nhóm sử dụng hai công cụ chính là Selenium và Beautifulsoup4, được chia làm hai bước chính.

Bước thứ nhất, truy cập từng page trong phần mua xe (có khoảng 1800 pages), thu thập đường dẫn URL của các xe được giao bán. Các đường dẫn này sẽ được lưu vào một file, phục vụ cho bước thứ hai.

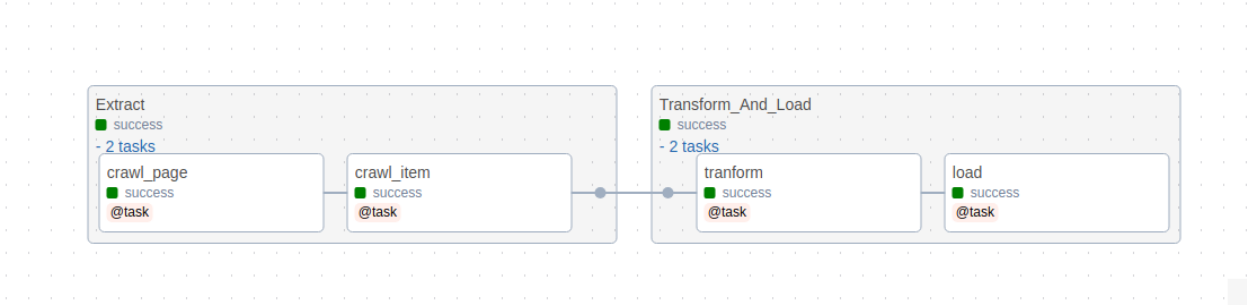
Bước hai, duyệt qua các đường dẫn có trong file đã được thu thập ở bước trước, tiến hành thu thập dữ liệu của từng xe. Sau khi dữ liệu được thu thập, sẽ được lưu ở dưới các file định dạng JSON và được lưu trữ trên MiniO.

Sau quá trình thu thập và lưu trữ, dữ liệu sẽ được tiền xử lý để lưu vào kho dữ liệu cuối cùng. Dữ liệu thu thập của nhóm chỉ được thu thập ở một trang web, có các trường chủ yếu dưới dạng văn bản nên việc tiền xử lý chủ yếu ở việc làm sạch dữ liệu, chuyển đổi dữ liệu và loại bỏ các trùng lặp, ngoại lệ.

Ở quy trình làm sạch dữ liệu, nhóm chủ yếu hướng tới việc điền hoặc loại bỏ các giá trị thiếu. Ví dụ, nếu trường số ki lô mét đã đi nếu thiếu có thể tiến hành điền hoặc loại bỏ dữ trên thông tin trình trạng của xe là mới hay cũ. Tiếp đến là việc chuyển đổi dữ liệu, dữ liệu sau khi thu thập được có các trường như số Ki lô mét đã đi, giá xe lẽ ra phải là dạng số nhưng lại là dạng văn bản, ví dụ 1000 km, 2000 km, 3 tỉ 200 triệu, 899 triệu, ... Những trường này cần được chuyển đổi sang dạng số, để tiện cho việc truy vấn, huấn luyện sau này. Ngoài ra, việc loại bỏ trùng lặp, ngoại lệ cũng được thực hiện ở bước tiền xử lý.

Dữ liệu được lưu trữ ở MiniO sau khi tiền xử lý sẽ được lưu trữ ở MongoDB.

Tất cả các công việc trên sẽ được quản lý và lập lịch bằng Airflow, hiện nay nhóm đang cài đặt toàn bộ các công việc phí trên định kì được thực hiện 1 ngày một lần theo đồ thị sau:



Hình 2. 2 Lập lịch và quản lý công việc bằng Airflow

2.2.2 Khó khăn

Dữ liệu về xe ô tô được rao bán rất quan trọng với một trang web mua bán xe, nó là cơ sở vận hành cho toàn bộ trang web,. Các lập trình viên lập trình nên trang web luôn cố gắng bảo vệ dữ liệu của trang web bằng nhiều cách khác nhau. Trong quá trình thu thập dữ liệu, nhóm gặp tình trạng chặn thu thập bằng cách chặn địa chỉ IP khi gửi quá nhiều request.

Như đã trình bày ở trên, trong quá trình thực hiện thu thập dữ liệu trang web, em sử dụng công cụ chính là Selenium. Việc sử dụng Selenium giúp giảm đáng kể việc chặn địa chỉ IP khi thu thập dữ liệu do Selenium sẽ thực hiện tự động một tab trình duyệt, sau đó truy cập địa chỉ của trang web và mô phỏng thao tác người dùng để thu thập dữ liệu. Ngoài ra, nhóm cũng sử dụng kết hợp Selenium và Beautiful Soup 4 trong quá trình thu thập dữ liệu có thể tận dụng những ưu điểm của cả hai công cụ. Selenium có thể được sử dụng để tương tác với các trang web động và thu thập dữ liệu theo yêu cầu, trong khi Beautiful Soup 4 có thể được sử dụng để phân tích và trích xuất dữ liệu từ các thành phần cụ thể trên trang web.

Với cách tiếp cận này, nhóm có thể tự động hóa quy trình thu thập dữ liệu và trích xuất thông tin từ các trang web phức tạp một cách hiệu quả. Việc kết hợp cả hai công cụ có thể tạo ra một quy trình thu thập dữ liệu mạnh mẽ, linh hoạt và hiệu quả. Sự kết hợp giữa khả năng tự động hóa và tương tác với trình duyệt của Selenium và khả năng phân tích và trích xuất dữ liệu của Beautiful Soup 4 giúp nhóm thu thập dữ liệu theo yêu cầu và trích xuất thông tin một cách chính xác và linh hoạt từ các trang web.

Khó khăn tiếp theo là việc triển khai các bộ cung cụ được đã được trình bày trên một hạ tầng để tất cả các thành viên trong nhóm có thể tiến hành truy cập và truy vấn dữ liệu. Ở đây, nhóm sử dụng giải pháp là triển khai các công cụ trên hạ tầng đám mây, được cung cấp bởi Microsoft Azure. Đối với MongoDB sẽ được chạy trực tiếp trên server, công cụ còn lại là Minio sẽ được triển khai trên Docker. Do kinh phí có hạn, server chỉ có 8GB RAM, vì thế việc tự động hóa, thu thập dữ liệu, tiền xử lý dữ liệu sẽ chỉ được thực hiện ở trên máy của thành viên nhóm. Server Azure chỉ là nơi lưu trữ, cung cấp điểm truy cập trực tuyến cho toàn bộ các thành viên trong nhóm.

Phần 3: Data Analysis

3.1 Cơ sở lý thuyết, công nghệ sử dụng

Giới thiệu về Google colab

Google Colab là một nền tảng máy tính đám mây cung cấp môi trường tính toán Jupyter Notebook trực tuyến miễn phí. Google Colab cho phép người dùng tạo, chỉnh sửa và chạy mã Python, R, Julia, và nhiều ngôn ngữ lập trình khác.

Google Colab có một số ưu điểm so với các nền tảng máy tính đám mây khác, bao gồm:

- Miễn phí: Google Colab cung cấp môi trường tính toán miễn phí, không giới hạn thời gian sử dụng.
- Dễ sử dụng: Google Colab sử dụng giao diện Jupyter Notebook, một giao diện thân thiện với người dùng.
- Có sẵn nhiều thư viện: Google Colab có sẵn nhiều thư viện Python, R, và Julia, giúp người dùng có thể thực hiện các tác vụ phân tích dữ liệu một cách dễ dàng.

Các thư viện sử dụng:

- Pandas là một thư viện Python dành cho phân tích và xử lý dữ liệu. Pandas cung cấp các hàm và phương thức để làm việc với dữ liệu bảng, bao gồm đọc, ghi, xử lý, phân tích và trực quan hóa dữ liệu.
- NumPy là một thư viện Python dành cho tính toán khoa học. NumPy cung cấp các hàm và phương thức để làm việc với các mảng, ma trận và vector.
- JSON là một định dạng trao đổi dữ liệu phổ biến. JSON sử dụng văn bản thuần túy để biểu diễn dữ liệu, bao gồm các số, chuỗi, danh sách và đối tượng.
- Matplotlib là một thư viện Python dành cho trực quan hóa dữ liệu. Matplotlib cung cấp các hàm và phương thức để tạo các biểu đồ và đồ thị dữ liệu, bao gồm biểu đồ đường, biểu đồ cột, biểu đồ tròn, biểu đồ phân tán, và nhiều loại khác.
- Seaborn là một thư viện Python dành cho trực quan hóa dữ liệu. Seaborn dựa trên Matplotlib, cung cấp các hàm và phương thức để tạo các biểu đồ và đồ thị dữ liệu trực quan và chuyên nghiệp.

Các thư viện này được sử dụng rộng rãi trong phân tích dữ liệu, xử lý dữ liệu, và trực quan hóa dữ liệu.

3.2 Data understanding

Thông tin về dữ liệu được mô tả như sau:

- Thông tin chung:
  - Số mẫu dữ liệu: 15,435
  - Số lượng trường: 22

- Các trường dữ liệu chính:
  - title: Tiêu đề liên quan đến xe.
  - author: Tác giả liên quan đến thông tin về xe.
  - engine: Loại động cơ của xe.
  - type: Loại hình xe (được giả định).
  - number_Km: Số km đã đi.
  - gear: Loại hộp số của xe.
  - status: Tình trạng của xe.
  - origin: Nguồn gốc xuất xứ của xe.
  - number_seats: Số lượng ghế ngồi trong xe.
  - number_door: Số cửa của xe.
  - price: Giá của xe.
  - exterior_color: Màu ngoại thất của xe.
  - interior_color: Màu nội thất của xe.
  - fuel_consumption: Sự tiêu thụ nhiên liệu (có vẻ có dữ liệu trống).
  - description: Mô tả chi tiết về xe.
  - car_name: Tên mẫu xe.
  - car_company: Tên hãng xe sản xuất.
  - year_man: Năm sản xuất của xe.
  - img_link: Liên kết đến hình ảnh của xe.
  - date: Ngày đăng thông tin.
  - phone_number: Số điện thoại liên hệ.
  - link: Liên kết đến thông tin chi tiết về xe.

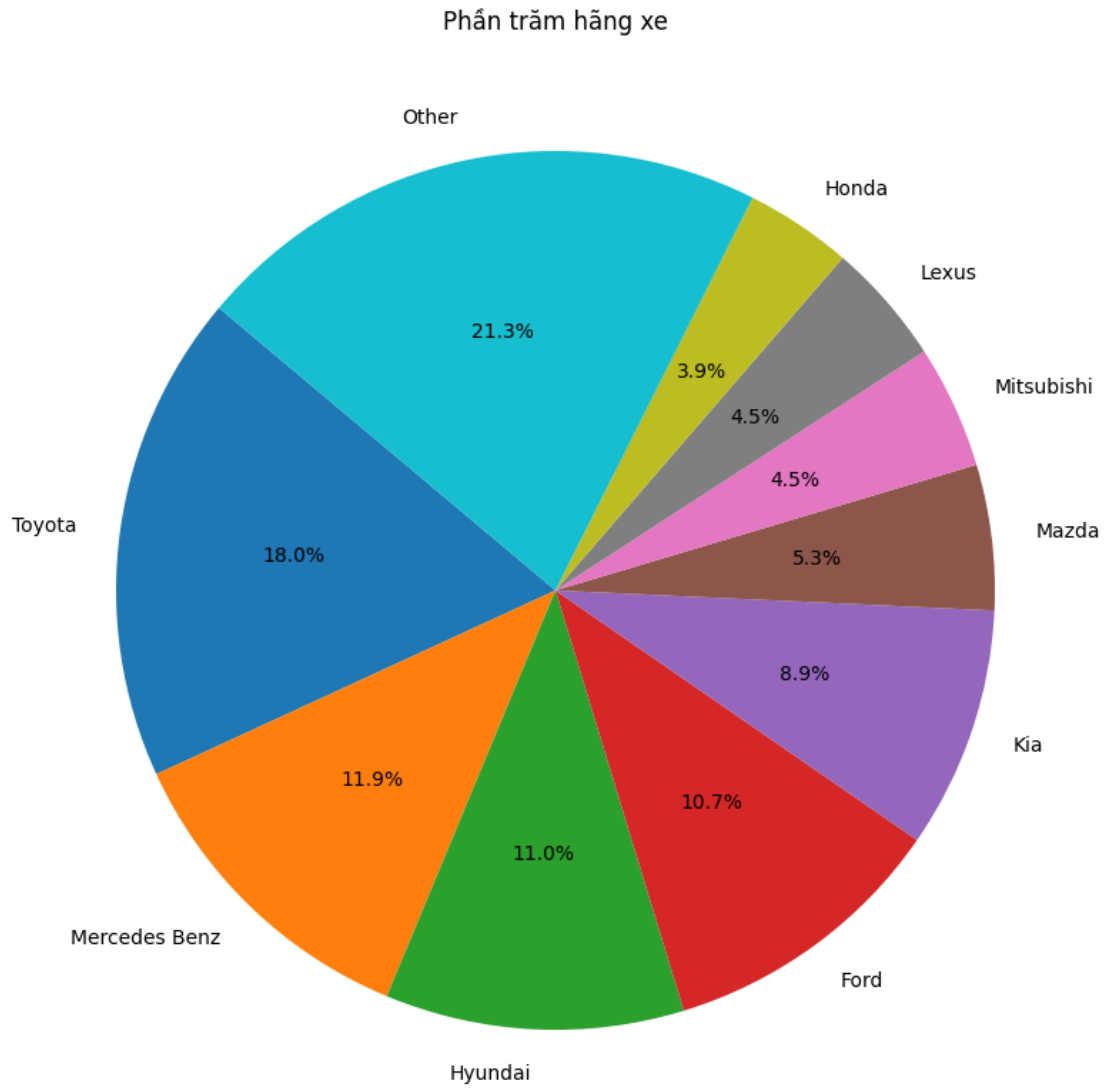
- Dữ liệu kiểu dữ liệu chính:
  - object: Thường là chuỗi ký tự hoặc các giá trị không liên tục.
  - int64: Dữ liệu số nguyên.
  - float64: Dữ liệu số thực.
  - datetime64[ns]: Dữ liệu thời gian.

- Vấn đề cần xử lý:
    - Dữ liệu trống (NULL) trong cột fuel_consumption cần được xử lý.

Bộ dữ liệu này cung cấp một loạt thông tin đa dạng về các chi tiết của xe ô tô, từ thông số kỹ thuật đến thông tin về giá và màu sắc. Sau đây sẽ là các phân tích và trực quan hóa để hiểu rõ hơn về đặc điểm và mối quan hệ giữa các trường trong dữ liệu.

3.3 Trực quan hóa dữ liệu

- **Biểu đồ tỷ trọng các hãng xe**



Hình 3. 1 Biểu đồ tỉ trọng hãng xe

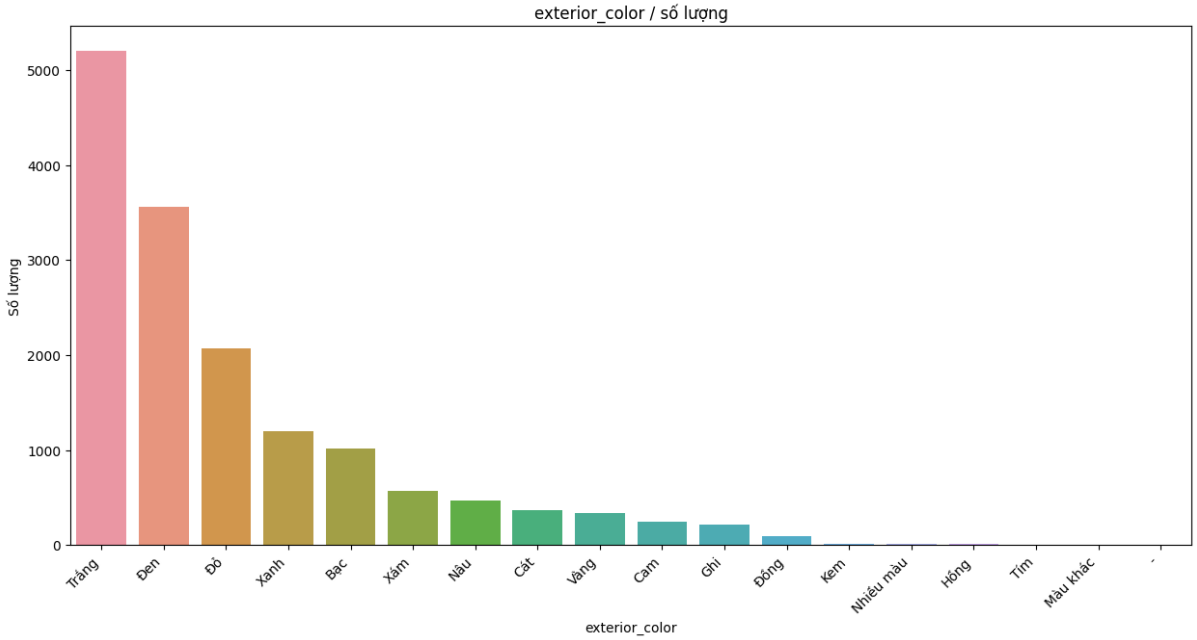
Các hãng xe Hàn Quốc chiếm thị phần lớn nhất với 46,2%, trong đó Hyundai chiếm 21,3%, Kia chiếm 11,9% và VinFast chiếm 11,0%. Đây là một thị trường tiềm năng đối với các hãng xe Hàn Quốc, đặc biệt là Hyundai và Kia. Các hãng này đã có nhiều năm kinh nghiệm hoạt động tại Việt Nam và đã xây dựng được thương hiệu uy tín với người tiêu dùng.

Các hãng xe Nhật Bản cũng có thị phần đáng kể với 34,5%, trong đó Toyota chiếm 18,0%, Mitsubishi chiếm 7,8% và Honda chiếm 6,0%. Các hãng xe Nhật Bản được đánh

giá cao về chất lượng, độ bền và khả năng tiết kiệm nhiên liệu. Đây là những yếu tố quan trọng thu hút khách hàng Việt Nam.

Các hãng xe khác chiếm thị phần còn lại với 9,3%, trong đó Mazda chiếm 8,9%, Ford chiếm 5,3%, Mercedes-Benz chiếm 4,5% và Lexus chiếm 3,9%. Các hãng xe này đang nỗ lực tăng thị phần tại Việt Nam bằng cách tung ra các mẫu xe mới với giá cả cạnh tranh và các chương trình khuyến mãi hấp dẫn.

- **Biểu đồ thể hiện số lượng màu ngoại thất của xe**



Hình 3. 2 Biểu đồ số lượng màu ngoại thất của xe

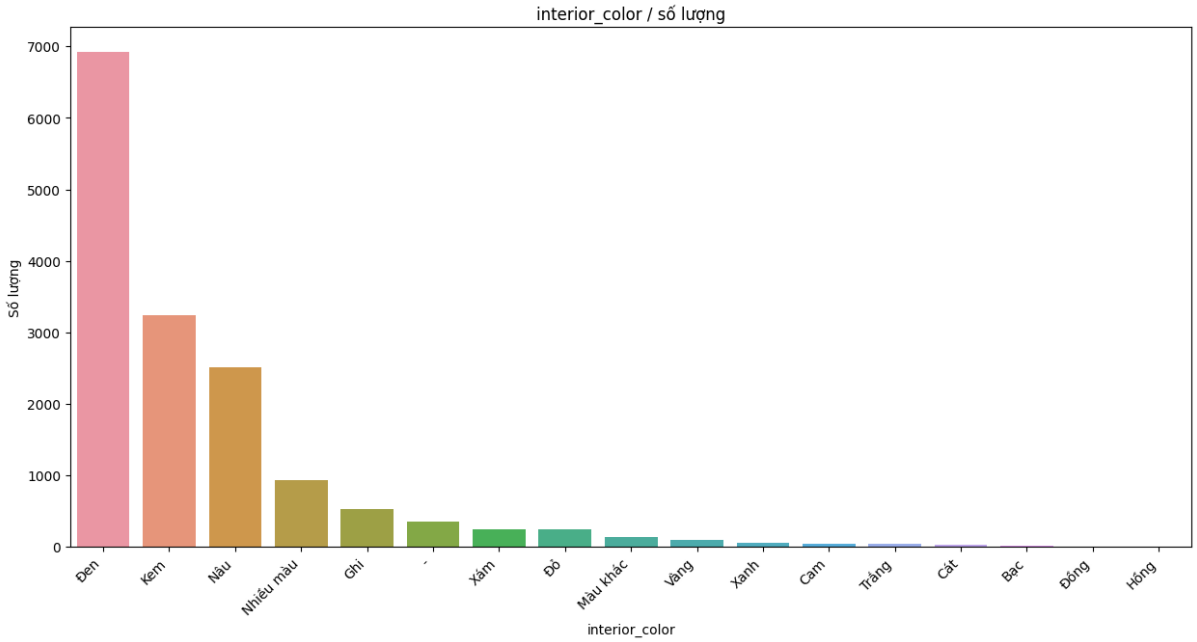
Màu trắng là màu phổ biến nhất, chiếm hơn 10% tổng số ô tô được bán ra. Điều này có thể là do màu trắng được coi là một màu trung tính và dễ phối hợp với các màu khác. Nó cũng là một màu phổ biến cho các doanh nghiệp và tổ chức, vì nó được coi là một màu chuyên nghiệp và đáng tin cậy.

Màu đen là màu phổ biến thứ hai, chiếm hơn 9% tổng số ô tô được bán ra. Màu đen thường được coi là một màu sang trọng và tinh tế. Nó cũng là một màu phổ biến cho các chiếc xe thể thao và xe cao cấp.

Màu đỏ là màu phổ biến thứ ba, chiếm hơn 8% tổng số ô tô được bán ra. Màu đỏ thường được coi là một màu trẻ trung và năng động. Nó cũng là một màu phổ biến cho các chiếc xe thể thao và xe hiệu suất cao.

Các màu khác được bán ra với số lượng ít hơn, với màu xám, xanh lam, vàng và cam là những màu phổ biến nhất tiếp theo.

- **Biểu đồ thể hiện số lượng xe có màu nội thất**



Hình 3. 3 Biểu đồ số lượng xe có màu nội thất

Màu đen là màu phổ biến nhất, chiếm hơn 6% tổng số ô tô được bán ra. Điều này có thể là do màu đen được coi là một màu sang trọng và tinh tế. Nó cũng là một màu phổ biến cho các chiếc xe cao cấp và xe sang trọng.

Màu kem là màu phổ biến thứ hai, chiếm hơn 5% tổng số ô tô được bán ra. Màu kem thường được coi là một màu trung tính và dễ phối hợp với các màu khác. Nó cũng là một màu phổ biến cho các chiếc xe gia đình và xe cỡ nhỏ.

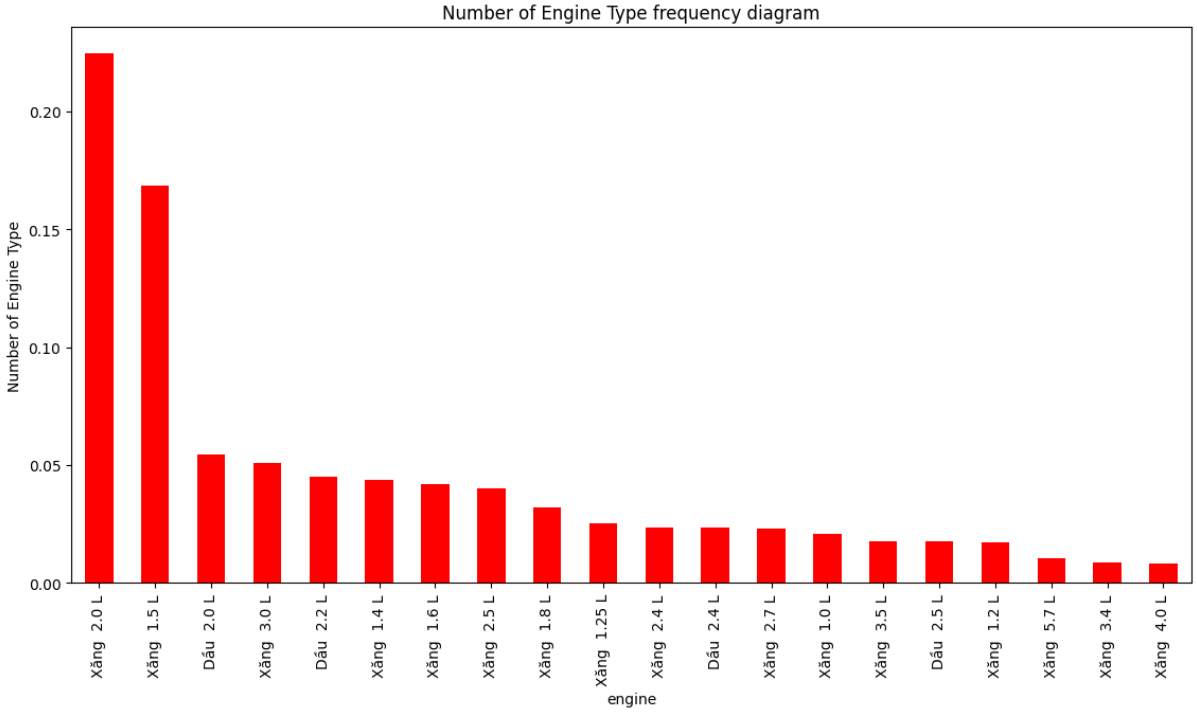
Màu nâu là màu phổ biến thứ ba, chiếm hơn 4% tổng số ô tô được bán ra. Màu nâu thường được coi là một màu sang trọng và cổ điển. Nó cũng là một màu phổ biến cho các chiếc xe SUV và xe bán tải.

- **Biểu đồ thể hiện số lượng loại động cơ**

Động cơ xăng chiếm hơn 80% tổng số ô tô được bán ra. Điều này cho thấy rằng động cơ xăng vẫn là loại động cơ phổ biến nhất trên thị trường ô tô hiện nay.

Động cơ diesel chiếm hơn 10% tổng số ô tô được bán ra. Điều này cho thấy rằng động cơ diesel vẫn là một lựa chọn phổ biến, đặc biệt là trong các chiếc xe tải và SUV.

Các loại động cơ khác, chẳng hạn như động cơ điện, động cơ hybrid và động cơ khí tự nhiên, chiếm phần còn lại của thị trường. Điều này cho thấy rằng các loại động cơ này đang trở nên phổ biến hơn, nhưng chúng vẫn chưa chiếm được thị phần đáng kể.



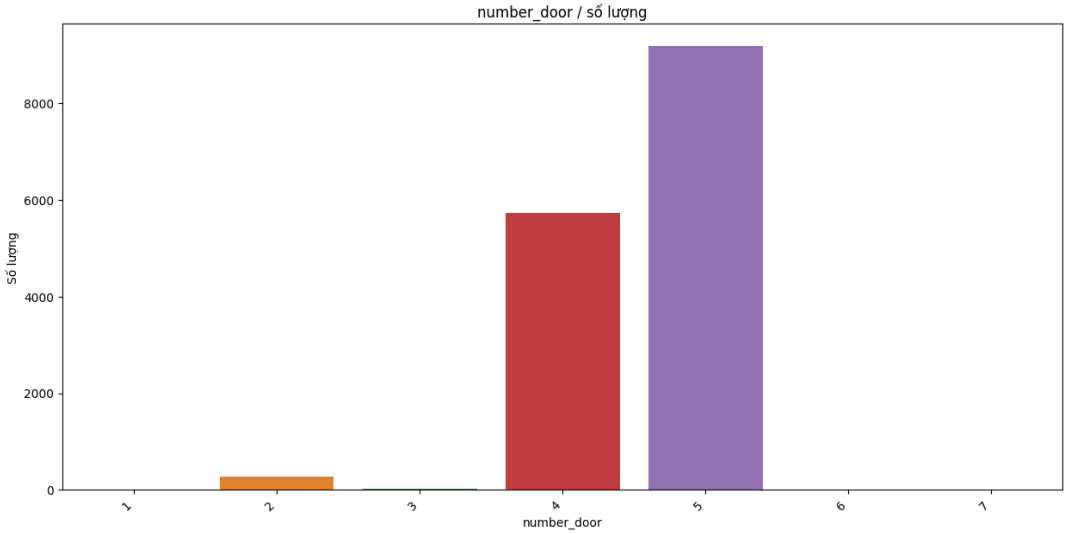
Hình 3. 4 Biểu đồ số lượng động cơ

Động cơ xăng chiếm hơn 80% tổng số ô tô được bán ra. Điều này cho thấy rằng động cơ xăng vẫn là loại động cơ phổ biến nhất trên thị trường ô tô hiện nay.

Động cơ diesel chiếm hơn 10% tổng số ô tô được bán ra. Điều này cho thấy rằng động cơ diesel vẫn là một lựa chọn phổ biến, đặc biệt là trong các chiếc xe tải và SUV.

Các loại động cơ khác, chẳng hạn như động cơ điện, động cơ hybrid và động cơ khí tự nhiên, chiếm phần còn lại của thị trường. Điều này cho thấy rằng các loại động cơ này đang trở nên phổ biến hơn, nhưng chúng vẫn chưa chiếm được thị phần đáng kể.

- **Biểu đồ số cửa ô tô**



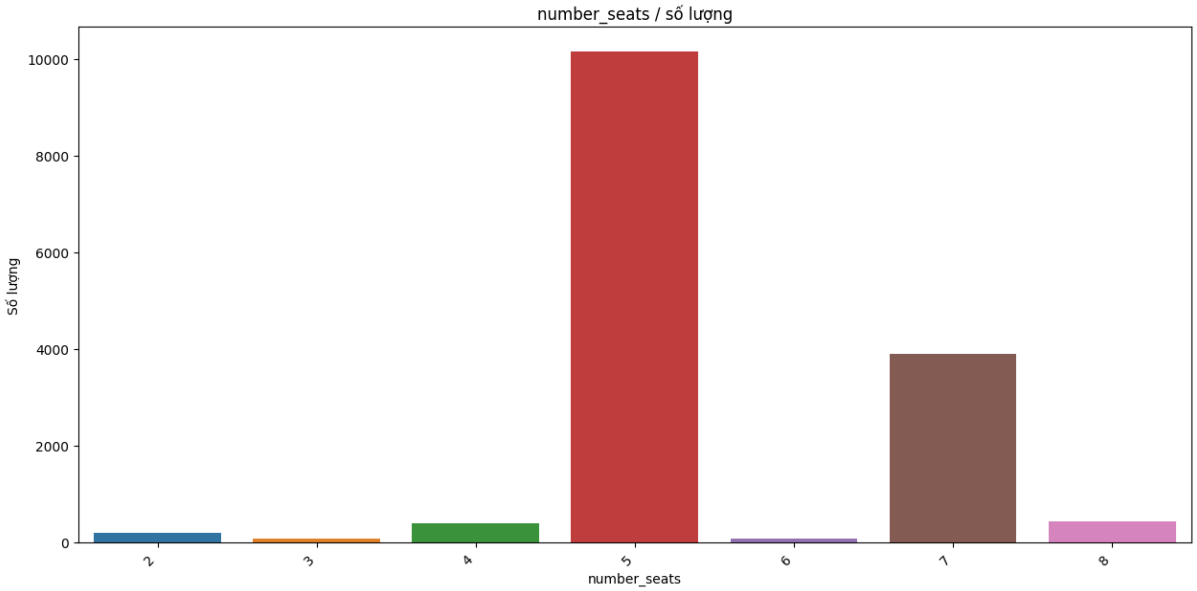
Hình 3. 5 Biểu đồ số cửa ô tô

Biểu đồ này thể hiện số lượng ô tô có số cửa khác nhau. Theo đó, số ô tô có 4 cửa chiếm đa số, với hơn 8.000 chiếc. Tiếp theo là số ô tô có 5 cửa, với hơn 6.000 chiếc. Số ô tô có 3 cửa và 2 cửa tương đương nhau, với khoảng 4.000 chiếc mỗi loại. Số ô tô có 1 cửa và 6 cửa rất ít, chỉ khoảng 2.000 chiếc mỗi loại.

Từ biểu đồ này, có thể thấy số cửa ô tô phổ biến nhất là 4 cửa, chiếm khoảng 60% tổng số ô tô. Số ô tô có 5 cửa chiếm khoảng 40% tổng số ô tô. Số ô tô có 3 cửa và 2 cửa tương đương nhau, chiếm khoảng 20% tổng số ô tô. Số ô tô có 1 cửa và 6 cửa rất ít, chỉ chiếm khoảng 10% tổng số ô tô.

Số lượng ô tô có 4 cửa nhiều hơn do đây là kiểu dáng phổ biến nhất, phù hợp với nhiều loại xe khác nhau, từ xe sedan, hatchback, crossover đến SUV. Số ô tô có 5 cửa cũng khá phổ biến, phù hợp với các gia đình có nhiều thành viên hoặc những người thường xuyên chở nhiều đồ đạc. Số ô tô có 3 cửa và 2 cửa thường được sử dụng cho các dòng xe thể thao hoặc xe mui trần. Số ô tô có 1 cửa và 6 cửa là những loại xe khá đặc biệt, thường được sử dụng cho các mục đích đặc biệt, chẳng hạn như xe cứu thương hoặc xe chở chất nguy hiểm.

- **Biểu đồ số chỗ ngồi ô tô**



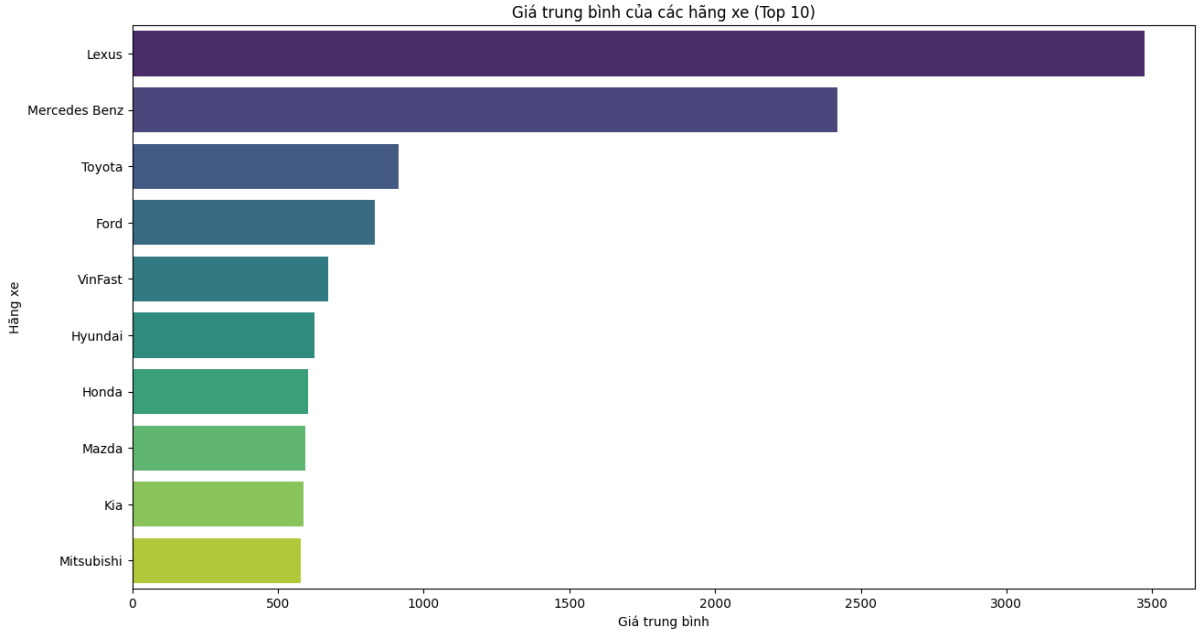
Hình 3. 6 Biểu đồ số chỗ ngồi của ô tô

Xe ô tô 4 chỗ ngồi là loại xe phổ biến nhất ở Việt Nam. Điều này có thể được giải thích bởi nhu cầu sử dụng ô tô của người dân Việt Nam chủ yếu là cho gia đình, với số lượng thành viên từ 2 đến 4 người.

Xe ô tô 5 chỗ ngồi cũng là một lựa chọn phổ biến, phù hợp với các gia đình có từ 4 đến 5 thành viên.

Xe ô tô 2 chỗ ngồi và ô tô 7 chỗ ngồi có tỷ lệ thấp hơn, do nhu cầu sử dụng không phổ biến. Xe ô tô 2 chỗ ngồi thường được sử dụng cho mục đích cá nhân, như đi lại, giải trí,... Xe ô tô 7 chỗ ngồi thường được sử dụng cho mục đích gia đình, với số lượng thành viên đông hoặc cho mục đích kinh doanh, như vận chuyển hành khách, hàng hóa,...

- **Biểu đồ thể hiện tác động của hãng xe lên giá**



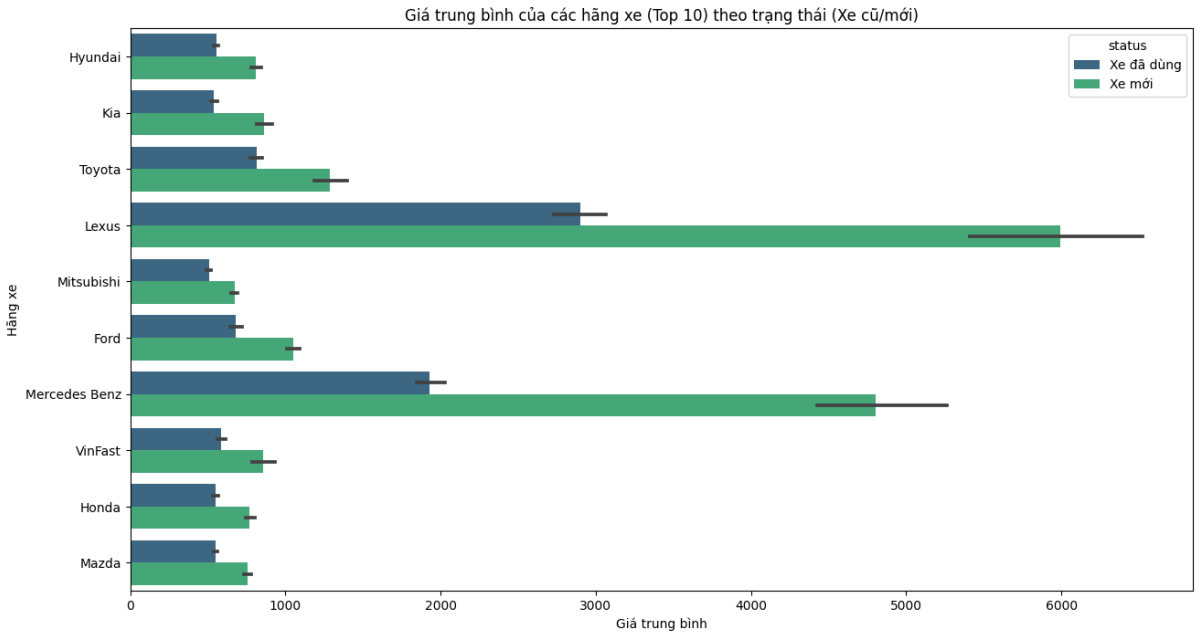
Hình 3. 7 Biểu đồ thể hiện tác động của hãng xe lên giá

Biểu đồ thể hiện tác động của hãng xe lên giá cho thấy rằng, các hãng xe sang và cao cấp có giá bán trung bình cao hơn đáng kể so với các hãng xe phổ thông. Cụ thể, Lexus là hãng xe có giá bán trung bình cao nhất, với giá trung bình là 3.500 triệu đồng. Tiếp theo là Mercedes-Benz, với giá trung bình là 3.000 triệu đồng. Toyota, Ford, VinFast, Hyundai, Honda, Mazda và Mitsubishi là các hãng xe có giá bán trung bình thấp hơn, từ 500 triệu đồng đến 2.500 triệu đồng.

Có thể thấy rằng, giá bán của một chiếc ô tô phụ thuộc vào nhiều yếu tố, trong đó có thương hiệu là một yếu tố quan trọng. Các hãng xe sang và cao cấp thường có giá bán cao hơn do nhiều lý do, chẳng hạn như:

- Sử dụng các vật liệu cao cấp và công nghệ tiên tiến hơn
- Có các tính năng và tiện nghi cao cấp hơn
- Có thương hiệu uy tín và danh tiếng

- **Biểu đồ thể hiện tác động của hãng xe lên giá trong từng trạng thái: xe đã dùng / xe mới**



Hình 3. 8 Biểu đồ thể hiện tình trạng xe lên giá

Biểu đồ thể hiện tác động của hãng xe lên giá trong từng trạng thái: xe đã dùng / xe mới cho thấy rằng, các hãng xe sang và cao cấp thường có giá bán cao hơn đáng kể so với các hãng xe phổ thông, ở cả hai trạng thái xe đã dùng và xe mới.

Xe mới

Ở trạng thái xe mới, Lexus là hãng xe có giá bán trung bình cao nhất, với giá trung bình là 3.500 triệu đồng. Tiếp theo là Mercedes-Benz, với giá trung bình là 3.000 triệu đồng. Toyota, Ford, VinFast, Hyundai, Honda, Mazda và Mitsubishi là các hãng xe có giá bán trung bình thấp hơn, từ 500 triệu đồng đến 2.500 triệu đồng.

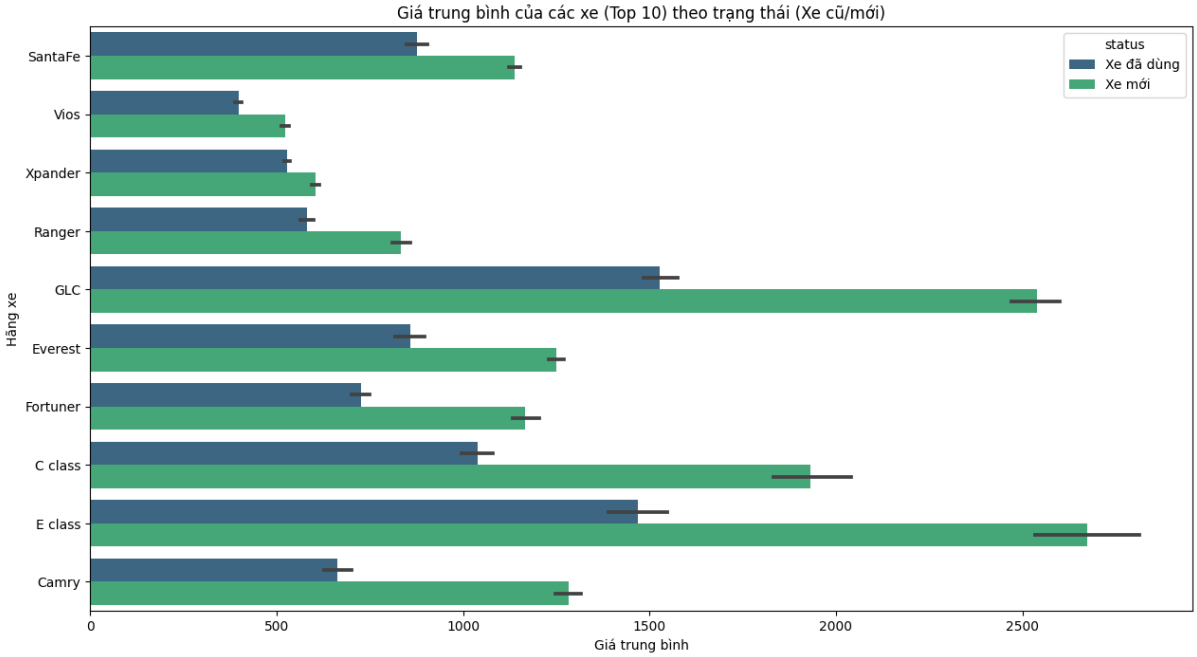
Xe đã dùng

Ở trạng thái xe đã dùng, Lexus vẫn là hãng xe có giá bán trung bình cao nhất, với giá trung bình là 2.500 triệu đồng. Tiếp theo là Mercedes-Benz, với giá trung bình là 2.000 triệu đồng. Toyota, Ford, VinFast, Hyundai, Honda, Mazda và Mitsubishi là các hãng xe có giá bán trung bình thấp hơn, từ 500 triệu đồng đến 1.500 triệu đồng.

Có thể thấy rằng, giá bán của một chiếc ô tô đã dùng thường thấp hơn đáng kể so với giá bán của một chiếc ô tô mới. Tuy nhiên, trong trường hợp của các hãng xe sang và cao

cấp, giá bán của một chiếc ô tô đã dùng vẫn có thể cao hơn đáng kể so với giá bán của một chiếc ô tô mới của các hãng xe phổ thông.

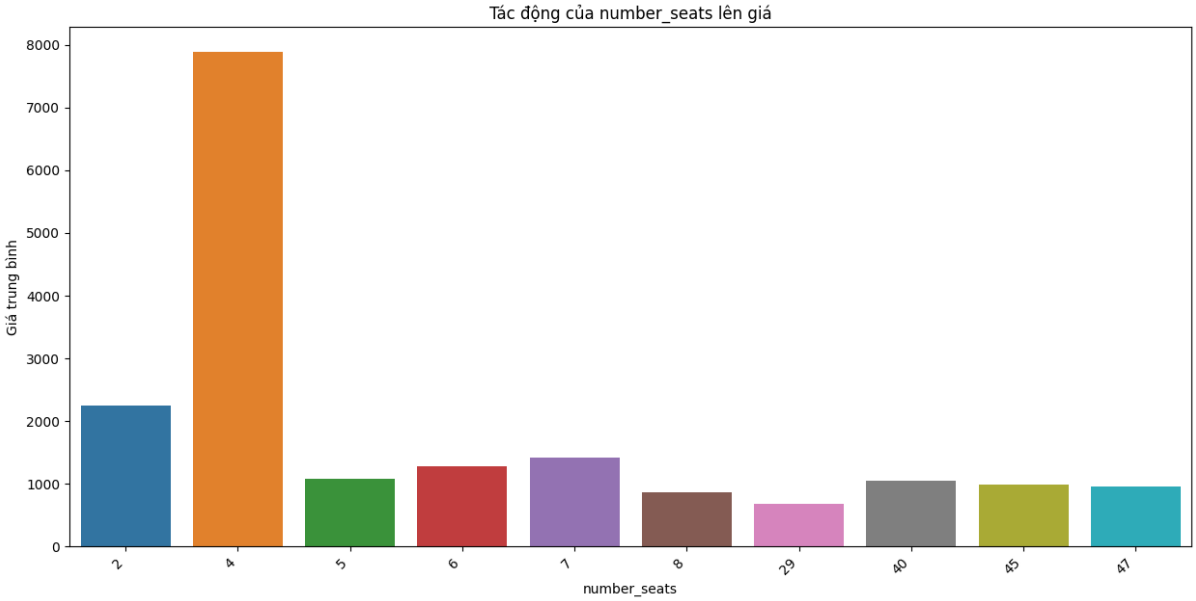
- **Biểu đồ thể hiện tác động của tên xe lên giá**



Hình 3. 9 Biểu đồ thể hiện tác động của tên xe lên giá

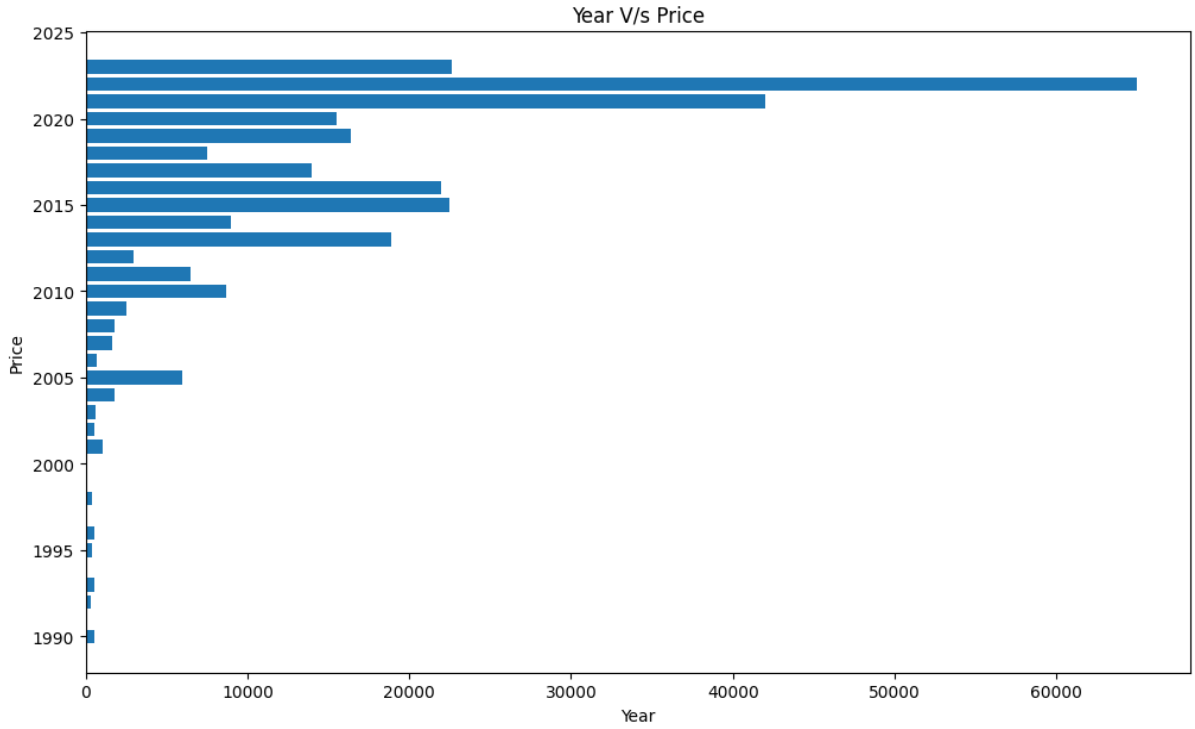
Cũng tương tự như biểu đồ trong 2.2, ta có thể thấy trong từng hãng xe mỗi dòng xe khác nhau sẽ có mức giá khác nhau tùy thuộc vào từng chi tiết kỹ thuật của các dòng xe.

- **Biểu đồ thể hiện sự tác động của số chỗ ngồi lên giá**



Hình 3. 10 Biểu đồ thể hiện tác động của số chỗ ngồi lên giá

- **Biểu đồ thể hiện tác động của năm sản xuất lên giá**

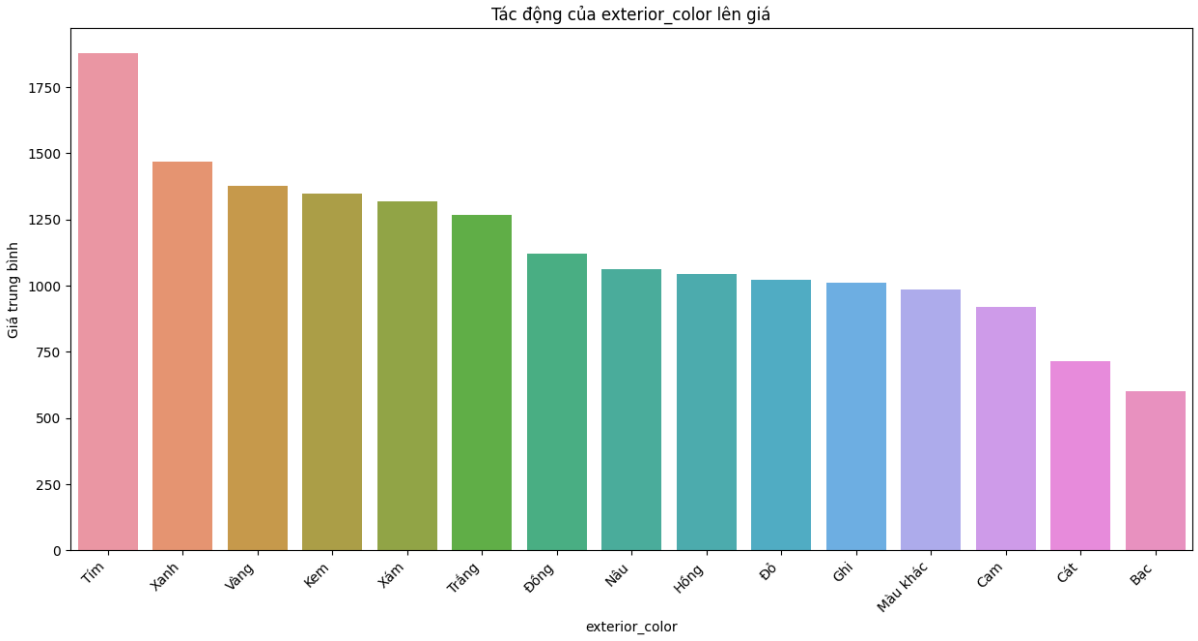


Hình 3. 11 Biểu đồ thể hiện tác động của năm sản xuất lên giá

Biểu đồ thể hiện tác động của năm sản xuất lên giá bán của ô tô trong bộ dữ liệu thu thập được cho thấy rằng, giá ô tô thường giảm dần theo năm sản xuất. Cụ thể, giá ô tô mới thường cao hơn đáng kể so với giá ô tô đã dùng. Giá ô tô đã dùng thường giảm dần theo thời gian, do nhiều yếu tố, bao gồm:

- Giá trị khấu hao của xe
- Chi phí bảo dưỡng và sửa chữa
- Tình trạng thị trường

- **Biểu đồ thể hiện màu ngoại thất ảnh hưởng lên giá xe**



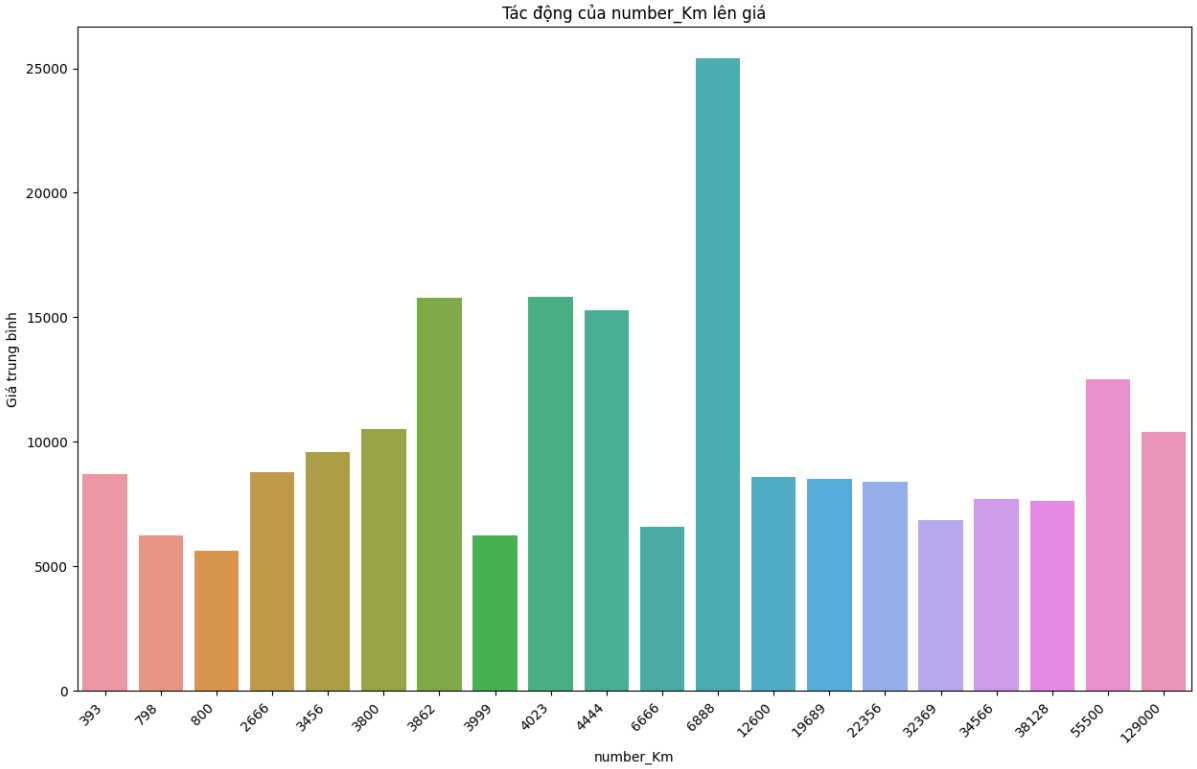
Hình 3. 12 Biểu đồ thể hiện tác động của màu ngoại thất lên giá

Biểu đồ tác động của màu ngoại thất lên giá bán ô tô cho thấy rằng, màu sắc ngoại thất của ô tô có thể ảnh hưởng đáng kể đến giá bán của xe. Cụ thể, màu trắng thường có giá cao nhất, tiếp theo là màu đen, bạc, và các màu khác.

Có thể thấy rằng, màu trắng là màu phổ biến nhất và được ưa chuộng nhất trên thị trường ô tô. Điều này có thể được giải thích bởi một số lý do, bao gồm:

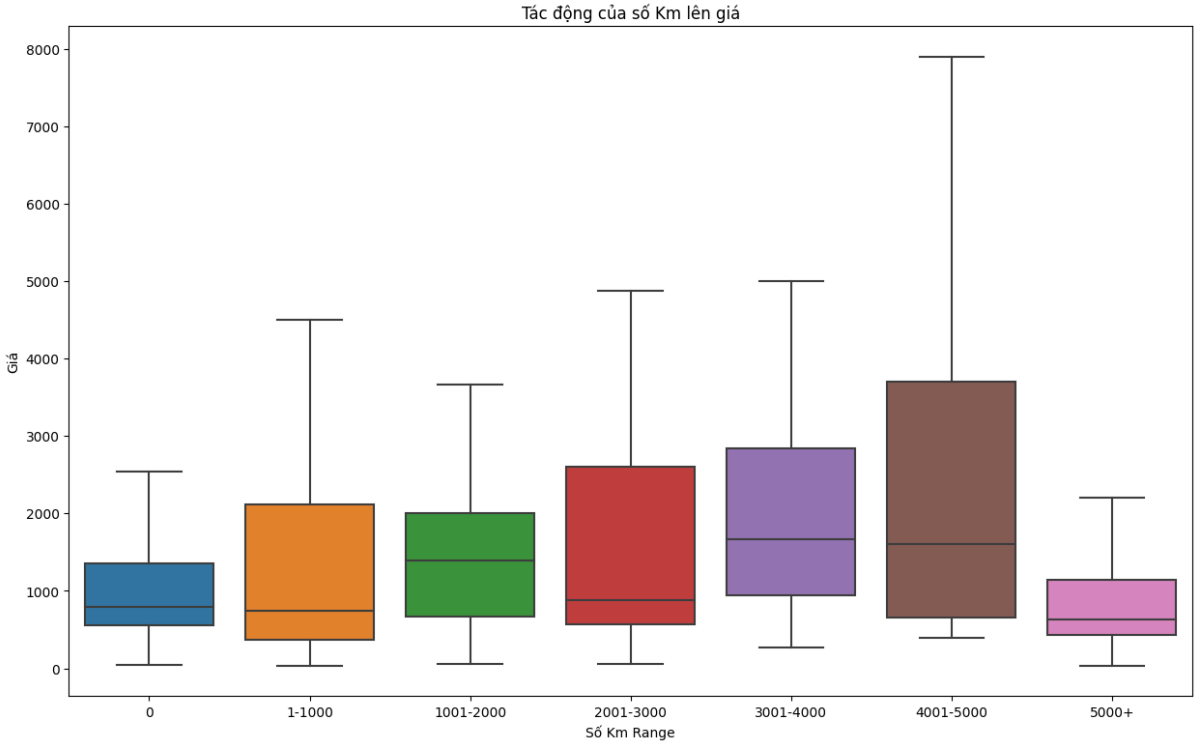
- Màu trắng là màu trung tính, phù hợp với nhiều phong cách và sở thích khác nhau.
- Màu trắng tạo cảm giác sang trọng và tinh tế.
- Màu trắng dễ dàng giữ sạch và bảo dưỡng.

- **Biểu đồ thể hiện tác động của số Kilomet xe đã đi lên giá của chúng**



Hình 3. 13 Biểu đồ thể hiện tác động của số Km đã đi lên giá

- **Biểu đồ thể hiện sự phân bố và tác động của số Km đã đi lên giá**

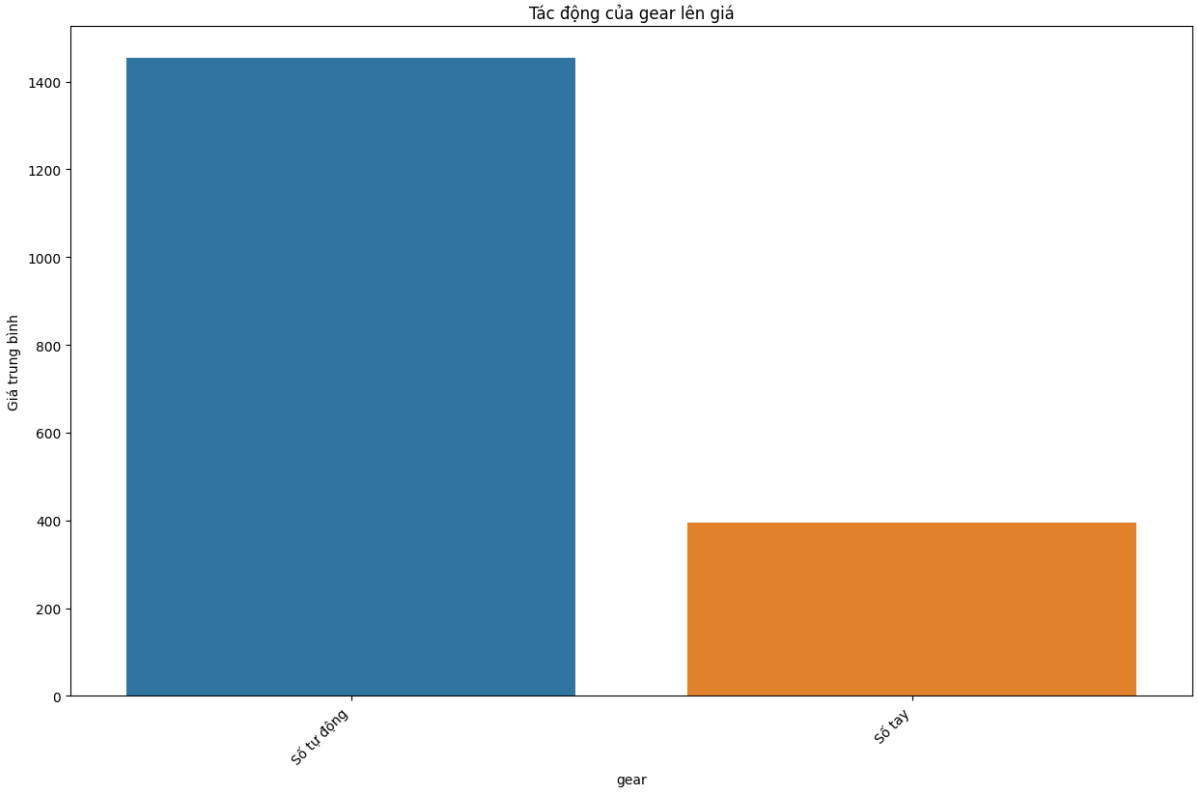


Hình 3. 14 Biểu đồ thể hiện phân bố và tác động của số Km đã đi lên giá

Biểu đồ thể hiện tác động của số Kilomet xe đã đi lên giá của chúng cho thấy rằng, số Kilomet xe đã đi càng nhiều thì giá xe càng thấp. Cụ thể, giá xe giảm dần theo số Kilomet xe đã đi, với mức độ giảm dần theo thời gian.

Có thể thấy rằng, số Kilomet xe đã đi là một yếu tố quan trọng ảnh hưởng đến giá bán của xe. Điều này là do các xe đã đi nhiều Kilomet thường có khả năng bị hao mòn và hư hỏng cao hơn.

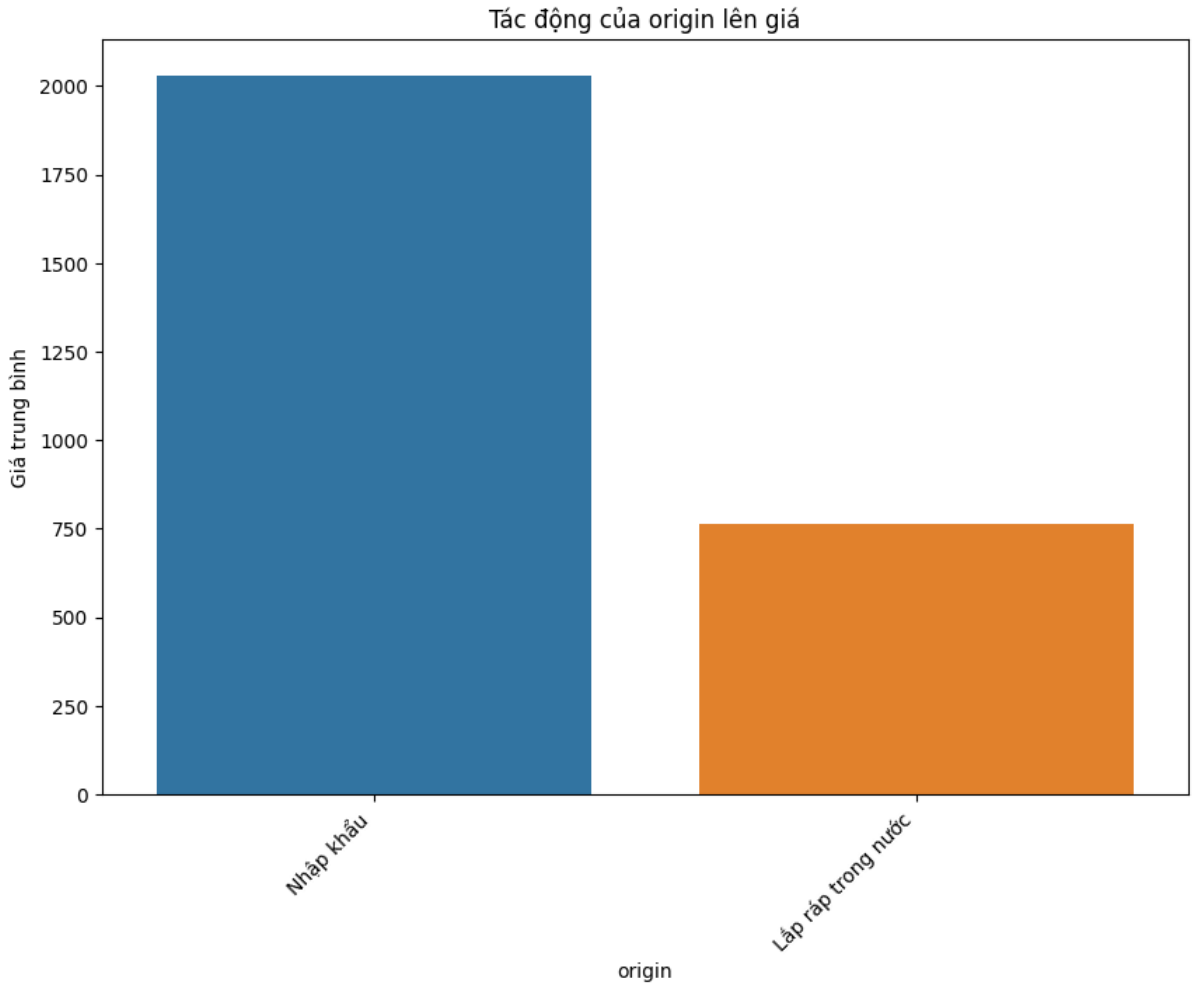
- **Biểu đồ thể hiện tác động của hộp số lên giá**



Hình 3. 15 Biểu đồ thể hiện tác động của hộp số lên giá

Ta có thể thấy loại hộp số ảnh hưởng khá nhiều đến giá, hộp số tự động sẽ có mức giá cao hơn số sàn vì thực tế hộp số tự động có cấu tạo và chi tiết kỹ thuật phức tạp hơn các hộp số sàn truyền thống, điều này khiến giá thành có sự chênh lệch.

- **Biểu đồ thể hiện tác động của nguồn gốc lên giá**



Hình 3. 16 Biểu đồ thể hiện tác động của nguồn gốc lên giá

Ta có thể thấy xe nhập khẩu có mức giá cao hơn rất nhiều xe lắp ráp ở trong nước. Điều này có thể liên hệ với thực tế vấn đề liên quan đến thuế nhập khẩu, các xe nhập khẩu sẽ bị đánh thuế khá cao, điều này ảnh hưởng rất nhiều đến giá.

Phần 4: Modeling

4.1 Cơ sở lý thuyết

4.1.1. Mô hình hồi quy tuyến tính (Linear Regression)

Giới thiệu chung

Hồi quy tuyến tính là một mô hình dự đoán giá trị đầu ra dựa trên các biến đầu vào.

Giả định rằng mối quan hệ giữa biến đầu vào và đầu ra là tuyến tính.

Mục tiêu là tìm ra đường thẳng (hoặc siêu mặt phẳng trong không gian đa chiều) sao cho tổng bình phương sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất (phương pháp bình phương tối thiểu).

Ưu điểm:

- Đơn giản và dễ hiểu.
- Tính tuyến tính giúp giải thích dễ dàng.

Hạn chế:

- Giả định về tuyến tính có thể không phù hợp cho dữ liệu có mối quan hệ phức tạp.
- Nhạy cảm với nhiễu và outliers.

4.1.2. Mô hình KNeighborsRegressor (KNN)

Giới thiệu chung:

KNN là một mô hình dựa trên việc xác định giá trị đầu ra dựa trên giá trị của k láng giềng gần nhất trong không gian đầu vào.

KNN không tạo ra một hàm dự đoán toàn cục, mà chỉ lưu trữ dữ liệu đào tạo để trực tiếp tìm kiếm và dự đoán.

Ưu điểm:

- Dễ hiểu và triển khai.
- Khả năng làm việc tốt cho dữ liệu có cấu trúc không tuyến tính.

Hạn chế:

- Nhạy cảm với nhiễu và outliers.
- Yêu cầu lưu trữ toàn bộ tập dữ liệu đào tạo.

4.1.3. Mô hình Gradient Boosting Regressor

Giới thiệu chung:

Gradient Boosting là một phương pháp học máy ensemble (tập hợp nhiều mô hình yếu thành một mô hình mạnh).

Xây dựng mô hình theo cách tuần tự, mỗi mô hình cố gắng cải thiện và sửa sai của mô hình trước đó.

Tổng các dự đoán của các cây con tạo ra dự đoán cuối cùng.

Ưu điểm:

- Hiệu suất cao và có khả năng xử lý cả dữ liệu có tính chất phức tạp.
- Khả năng làm việc tốt với dữ liệu có nhiễu và outliers.

Hạn chế:

- Đòi hỏi thời gian và tài nguyên tính toán lớn hơn so với một số mô hình khác.

4.1.4. Mô hình XGBoost Regressor

Giới thiệu chung:

XGBoost (eXtreme Gradient Boosting) đại diện cho một biến thể mạnh mẽ của Gradient Boosting, tích hợp nhiều tối ưu hóa và cải tiến hiệu suất.

Mô hình này sử dụng kỹ thuật "regularization" để kiểm soát overfitting, đồng thời áp dụng "shrinkage" để giảm tác động của từng cây trong quá trình huấn luyện. Điều này giúp tăng tính ổn định và hiệu suất của mô hình, đặc biệt là khi xử lý dữ liệu có độ phức tạp cao và tránh tình trạng quá mức học.

Ưu điểm:

- Hiệu suất cao và khả năng xử lý dữ liệu lớn.
- Hiệu quả tính toán và có khả năng tùy chỉnh nhiều tham số.

Hạn chế:

- Đòi hỏi hiểu biết sâu sắc về các tham số và cách tinh chỉnh chúng.

4.2 Huấn luyện model

4.2.1 Chuẩn bị dữ liệu

Từ những phân tích đánh giá và hiểu dữ liệu ở bước **Data Analysis** thực hiện một số biến đổi dữ liệu để phù hợp đưa vào model như sau:

- Xóa các dòng có giá trị null trong cột 'number_seats' và chỉ giữ lại những dòng có giá trị 'number_seats' từ 2 đến 16
- Chỉ giữ lại các dòng có **color** = ['Đen', 'Đỏ', 'Bạc', 'Xanh', 'Trắng', 'Nâu', 'Cát', 'Vàng', 'Cam', 'Ghi', 'Xám', 'Đồng'] do các màu còn lại có số lượng ít hơn rất nhiều so với các màu khác.
- Xóa đi các dòng có năm sản xuất **year_man** từ 2006 trở về trước .
- Xóa các hàng có giá trị **gear** không phải "Số tự động" hoặc "Số tay":
- Xóa các dòng có giá trị **price** không nằm trong khoảng từ 0.2 đến 10
- Xóa các dòng có **number_Km** là null và lớn hơn 600,000,000.
- Sử dụng thư viện LabelEncoder để chuyển đổi các giá trị của các cột **'engine', 'status', 'exterior_color','car_name','car_company','type','gear'** thành số.

Sau khi đã có một số biến đổi ta đã thu được dữ liệu sạch có thể đưa vào để train model tiếp theo đưa đầu vào **X** thành dạng vector và nhãn là **y** như sau:

X = ['engine', 'number_Km', 'status', 'number_seats', 'exterior_color', 'car_name', 'car_company', 'year_man']

y = ['price']

Tiếp theo chia tập train và test với tỉ lệ 7:3.

4.2.2 Tiến hành train, tinh chỉnh

Mô hình hồi quy tuyến tính (Linear Regression)

Kết quả thu được :
Mean Absolute Error: 0.8372027436815721

RMSE: 1.2141077476212396

R-squared: 0.14959680828488375

Từ những thông số trên cho thấy model khá tệ và model này không phù hợp với bộ dữ liệu thu thập được.

Mô hình KNeighborsRegressor (KNN)

Với số lân cận (n_neighbors) là 5 và đánh giá trên tập kiểm tra, kết quả là:

Mean Absolute Error: 0.7041207272315361

RMSE: 1.1496242267353105

R-squared: 0.2375309224603379

Từ những thông số trên cho thấy model này khá tệ và có vẻ không phù hợp với dữ liệu.

Mô hình Gradient Boosting Regressor

Kết quả thu được :
Mean Absolute Error: 0.26352304404505705

RMSE: 0.4587590101651996

R-squared: 0.8785829618301579

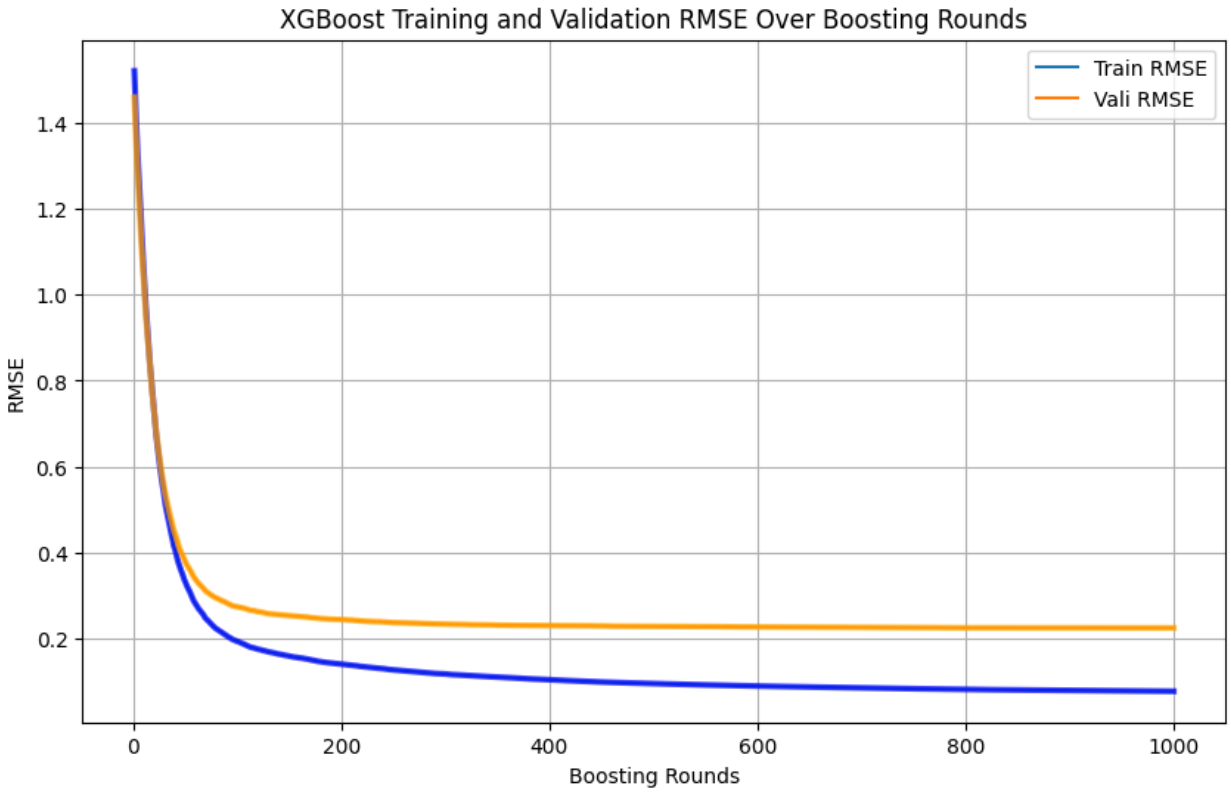
Từ những thông số trên cho thấy model này có độ chính xác khá tốt và phù hợp với bộ dữ liệu.

Mô hình XGBoost Regressor

Chia lại dữ liệu thành tập train , test , vali như sau :
X_train, X_temp, y_train, y_temp = train_test_split(X, y, test_size=0.4, random_state=42)

X_vali, X_test, y_vali, y_test = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)

Tiến hành tinh chỉnh các siêu tham số đạt kết quả như sau:



Hình 4. 1 Biểu đồ giá trị Train RMSE và Vali RMSE

Từ biểu đồ trên có thể thấy RMSE giảm và khá nhỏ trên cả tập train và vali.

MAE: 0.09859067118

RMSE: 0.2212945152

R-squared: 0.9731187777

Qua so sánh cho thấy XGBoost Regressor làm việc tốt nhất trên bộ dữ liệu này.

4.2.3 Phân tích và cải tiến

Thực hiện thêm bớt một số trường để tìm ra được số trường phù hợp và tốt nhất cho model XGBoost Regressor. Ưu tiên những trường dữ liệu quan trọng ảnh hưởng lớn đến giá trị giá. Thực hiện với 6 trường và 9 trường

Với 6 trường được cho là có mức ảnh hưởng lớn nhất

Các trường sử dụng : **'number_Km', 'status', 'number_seats', 'car_name', 'car_company', 'year_man'** (bớt đi hai trường là **engine** và **exterior _color** so với thử nghiệm ở mục 2.2)

Kết quả thu được :
MAE: 0.1170774036

RMSE: 0.2839504259

R-squared: 0.9557419005

Từ kết quả thu được ở trên cho thấy khi bớt đi hai trường **engine** và **exterior _color** model thu được có độ chính xác thấp hơn.

Sử dụng cả 9 trường tiềm năng

Các trường sử dụng : **'engine', 'number_Km', 'status', 'number_seats', 'exterior_color', 'car_name', 'car_company', 'year_man' , 'gear'** (thêm trường **gear** so với thử nghiệm ở mục 2.2).

Kết quả thu được:
MAE: 0.09384314842

RMSE: 0.2156524101

R-squared: 0.9744720263

Từ kết quả trên cho thấy model thu được có độ chính xác lớn hơn việc thử nghiệm với 8 trường trong mục 4.2 và đây là model tốt nhất hiện tại thu được.

Phần 5: Kết luận

5.1 Đánh giá kết quả

Kết quả đạt được

- Quá trình thu thập dữ liệu có hiệu quả và phù hợp với quy mô bài toán đặt ra
- Dữ liệu thu thập được phản ánh được giá cả mặt bằng chung thị trường xe ô tô ở Việt Nam. Thị trường phổ biến là loại xe hạng vừa và trung và các hãng xe nổi tiếng
- Qua mô hình huấn luyện có thể thấy các mô hình dự đoán đơn giản như hồi quy tuyến tính hay KNN cho kết quả dự đoán khá tồi trong khi các mô hình phức tạp hơn như Gradient Boosting hay XGBoot cho kết quả với độ chính xác tốt

Hạn chế

- Nguồn thu thập dữ liệu chưa được đa dạng
- Lượng dữ liệu còn hạn chế
- Quy trình thực hiện thủ công phức tạp

5.2 Hướng phát triển

Theo những hạn chế ở trên, trong tương lai hệ thống cần phát triển thêm ở một số khía cạnh để hoàn thiện hơn như:

- Mở rộng thêm các nguồn dữ liệu khác nhau

- Huấn luyện thêm bằng các phương pháp, dữ liệu khác nhau

- Tích hợp thêm việc deploy model