

TRƯỜNG ĐẠI HỌC SÀI GÒN

-----🙞🙜🕮🙞🙜-----



BÁO CÁO

Phân Tích Dữ Liệu

ĐỀ TÀI: Phân Tích Và Dự Đoán Giá Nhà

*Giảng Viên:* TS. Đỗ Như Tài

*Nhóm Thực Hiện:*

Trần Minh Phú - 3122410318

Đinh Nguyễn Duy Phong - 3122410306

Nguyễn Hoàng Phú - 3122410314

Phan Hoàng Vũ - 3123410436

LỜI CẢM ƠN

Trong quá trình thực hiện đồ án này, nhóm chúng em đã may mắn nhận được sự quan tâm và chỉ dẫn tận tình từ quý thầy cô khoa Công nghệ thông tin. Đặc biệt, chúng em xin được bày tỏ lòng biết ơn sâu sắc và chân thành nhất tới thầy Đỗ Như Tài – giảng viên trực tiếp hướng dẫn học phần. Thầy đã dành nhiều thời gian và tâm huyết để truyền giảng những kiến thức chuyên môn nền tảng cũng như chuyên sâu về lĩnh vực phân tích dữ liệu và các mô hình học máy hiện đại. Không dừng lại ở việc giảng dạy lý thuyết, thầy còn luôn theo sát, định hướng và đưa ra những góp ý vô cùng quý báu, giúp nhóm chúng em tháo gỡ những vướng mắc kỹ thuật, từ đó hoàn thiện phương pháp nghiên cứu, biết cách tiếp cận bài toán một cách đa chiều và trình bày kết quả đạt độ khoa học, logic và tiễn hơn. Chính sự tận tâm, tinh thần trách nhiệm cao cả và những lời động viên kịp thời của thầy đã trở thành nguồn động lực to lớn, tiếp thêm sức mạnh cho chúng em vượt qua những giai đoạn khó khăn, áp lực trong suốt quá trình triển khai dự án. Một lần nữa, chúng em xin chân thành cảm ơn thầy và kính chúc thầy luôn dồi dào sức khỏe, hạnh phúc, tiếp tục gặt hái được nhiều thành công rực rỡ hơn nữa trong sự nghiệp giảng dạy cũng như trên con đường nghiên cứu khoa học của mình.

MỤC LỤC

[CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI 1](#_Toc217073241)

[1.1. Lý do chọn đề tài 1](#_Toc217073242)

[1.2. Mục tiêu nghiên cứu 2](#_Toc217073243)

[1.2.1. Mục tiêu chung 2](#_Toc217073244)

[1.3. Câu hỏi nghiên cứu 2](#_Toc217073245)

[1.4. Đối tượng và phạm vi nghiên cứu 3](#_Toc217073246)

[1.4.1. Đối tượng nghiên cứu 3](#_Toc217073247)

[1.4.2. Phạm vi nghiên cứu 3](#_Toc217073248)

[1.5. Đóng góp nghiên cứu 4](#_Toc217073249)

[1.7.1. Cấu trúc dữ liệu 4](#_Toc217073250)

[1.7.2. Tính uy tín của dữ liệu 5](#_Toc217073251)

[1.7.3. Các biến dữ liệu đầu vào và đầu ra 6](#_Toc217073252)

[CHƯƠNG 2. CƠ SỞ LÝ LUẬN 7](#_Toc217073253)

[2.1. Tổng quan nghiên cứu 7](#_Toc217073254)

[2.2. Khoảng trống nghiên cứu 7](#_Toc217073255)

[2.3. Cơ sở lý thuyết về định giá bất động sản 8](#_Toc217073256)

[2.3.1. Lý thuyết Hedonic Pricing 8](#_Toc217073257)

[2.3.2. Lý thuyết vị trí (Location Theory) 8](#_Toc217073258)

[2.3.3. Đặc điểm vật lý của bất động sản 9](#_Toc217073259)

[2.3.4. Lý thuyết về pháp lý và quyền sở hữu 9](#_Toc217073260)

[2.3.5. Lý thuyết học máy trong dự báo giá 10](#_Toc217073261)

[CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU 11](#_Toc217073262)

[3.1. Quy trình nghiên cứu 11](#_Toc217073263)

[3.2. Thu thập và tiền xử lý dữ liệu 11](#_Toc217073264)

[3.2.1. Đánh giá chất lượng dữ liệu ban đầu 11](#_Toc217073265)

[3.2.2. Tiền xử lý dữ liệu 13](#_Toc217073266)

[3.3. Phương pháp phân tích khám phá và trực quan hóa dữ liệu 19](#_Toc217073267)

[3.3.1 Phân tích thống kê mô tả 19](#_Toc217073268)

[3.3.2 Trực quan hóa phân bố dữ liệu 19](#_Toc217073269)

[3.3.3 Phân tích mối quan hệ giữa các biến Biểu đồ phân tán 20](#_Toc217073270)

[3.4. Phân cụm bất động sản bằng K-means 20](#_Toc217073271)

[3.4.1. Giới thiệu phương pháp K-means 20](#_Toc217073272)

[3.4.2. Lựa chọn biến cho phân cụm 21](#_Toc217073273)

[3.4.3. Xác định số cụm tối ưu (Elbow & Silhouette) 21](#_Toc217073274)

[3.5. Phương pháp luận học máy 22](#_Toc217073275)

[3.5.1. Bài toán hồi quy 22](#_Toc217073276)

[3.5.2. Các thuật toán học máy được áp dụng 22](#_Toc217073277)

[3.5.3. Phương pháp đánh giá mô hình 23](#_Toc217073278)

[3.5.4. Kiểm định giả thuyết thống kê 24](#_Toc217073279)

[3.6. Triển khai xây dựng và đánh giá mô hình 24](#_Toc217073280)

[3.6.1. Tiền xử lý dữ liệu cho modeling 24](#_Toc217073281)

[3.6.2. Tinh chỉnh siêu tham số 25](#_Toc217073282)

[3.6.3. Đánh giá và lựa chọn mô hình 25](#_Toc217073283)

[3.6.4. Diễn giải mô hình và phân tích yếu tố ảnh hưởng 26](#_Toc217073284)

[CHƯƠNG 4. THẢO LUẬN KẾT QUẢ NGHIÊN CỨU 27](#_Toc217073285)

[4.1. Phân tích khám phá dữ liệu 27](#_Toc217073286)

[4.1.1 Phân tích đơn biến 27](#_Toc217073287)

[4.1.2 Phân tích đa biến 31](#_Toc217073288)

[4.2. Xây dựng mô hình phân cụm và dự đoán giá nhà 47](#_Toc217073289)

[4.2.1 Phương pháp phân cụm (K-means) 47](#_Toc217073290)

[4.2.2 Dự đoán giá nhà bằng mô hình máy học 54](#_Toc217073291)

[4.3. Diễn giải ý nghĩa kinh tế 62](#_Toc217073292)

[CHƯƠNG 5. KẾT LUẬN VÀ KHUYẾN NGHỊ 64](#_Toc217073293)

[5.1. Kết luận 64](#_Toc217073294)

[5.2. Hàm ý thực tiễn 65](#_Toc217073295)

[5.3. Hạn chế của nghiên cứu 65](#_Toc217073296)

[5.4. Hướng phát triển trong tương lai 66](#_Toc217073297)

DANH MỤC HÌNH ẢNH

[Hình 3.1: Phần trăm dữ liệu bị thiếu ở các cột 12](#_Toc217072132)

[Hình 3.2: Thống kê mô tả các cột số 13](#_Toc217072133)

[Hình 3.3: Kiểm tra số lượng dữ liệu bị thiếu 14](#_Toc217072134)

[Hình 3.4: Kiểm tra lại dữ liệu bị thiếu 15](#_Toc217072135)

[Hình 3.5: Kết quả sau khi xử lý 19](#_Toc217072136)

[Hình 3.6: Kết quả sau khi xử lý dữ liệu 19](#_Toc217072137)

[Hình 4.1 : Biểu đồ Histogram và Boxplot của biến Area 27](#_Toc217072138)

[Hình 4.2: Biểu đồ Histogram và Boxplot của biến Frontage 28](#_Toc217072139)

[Hình 4.3: Biểu đồ Histogram và Boxplot của biến Access Road. 28](#_Toc217072140)

[Hình 4.4: Biểu đồ Histogram và Boxplot của biến Floors. 29](#_Toc217072141)

[Hình 4.5: Biểu đồ Histogram và Boxplot của biến Bedrooms. 30](#_Toc217072142)

[Hình 4.6: Biểu đồ Histogram và Boxplot của biến Bathrooms. 30](#_Toc217072143)

[Hình 4.7: Biểu đồ Histogram và Boxplot của biến Price. 31](#_Toc217072144)

[Hình 4.8: Biểu đồ Boxplot phân bố giá nhà tại các thành phố trọng điểm (Top 10) 32](#_Toc217072145)

[Hình 4.9: Biểu đồ cột thể hiện số lượng tin đăng (quy mô giao dịch) theo từng Thành phố 33](#_Toc217072146)

[Hình 4.10: Giá nhà trung vị theo nhóm diện tích (Area\_bin) 34](#_Toc217072147)

[Hình 4.11: Giá nhà trung vị theo nhóm mặt tiền (Frontage\_bin) 35](#_Toc217072148)

[Hình 4.12: Giá nhà trung vị theo nhóm bề rộng đường trước nhà (Access\_Road\_bin) 36](#_Toc217072149)

[Hình 4.13: Giá nhà trung vị theo số tầng (Floors) 37](#_Toc217072150)

[Hình 4.14: Giá nhà trung vị theo số phòng tắm (Bathrooms) 38](#_Toc217072151)

[Hình 4.15: Giá nhà trung vị theo số phòng ngủ (Bedrooms) 39](#_Toc217072152)

[Hình 4.16: Heatmap tương quan giữa đặc điểm vật lý và giá nhà. 40](#_Toc217072153)

[Hình 4.17: Giá nhà trung vị theo City và nhóm diện tích (Area\_bin). 41](#_Toc217072154)

[Hình 4.18: Giá nhà trung vị theo City và nhóm mặt tiền (Frontage\_bin) 42](#_Toc217072155)

[Hình 4.19: Giá nhà trung vị theo City và nhóm bề rộng đường trước nhà (Access\_Road\_bin) 43](#_Toc217072156)

[Hình 4.20: Biểu đồ Boxplot phân bổ giá nhà theo Tình trạng Pháp lý và Mức độ Nội thất 45](#_Toc217072157)

[Hình 4.21: Biểu đồ so sánh Giá trung vị theo Thành phố dựa trên Pháp lý và Nội thất 46](#_Toc217072158)

[Hình 4.22: Biểu đồ Giá trung vị theo Thành phố và Mức độ nội thất 47](#_Toc217072159)

[Hình 4.23: Biểu đồ Elbow xác định số cụm tối ưu cho thuật toán K-Means 49](#_Toc217072160)

[Hình 4.24: Biểu đồ hệ số Silhouette theo số cụm K 49](#_Toc217072161)

[Hình 4.25: Biểu đồ PCA hai chiều thể hiện kết quả phân cụm K-Means với K = 3. 51](#_Toc217072162)

[Hình 4.26: Phân bố giá theo cụm 53](#_Toc217072163)

[Hình 4.27: Hình biểu diễn giá thực và giá dự đoán 58](#_Toc217072164)

[Hình 4.28: Tầm quan trọng của các biến 59](#_Toc217072165)

DANH MỤC BẢNG BIỂU

[Bảng 4.1: Giá trị trung bình các đặc trưng vật lý và giá nhà theo từng cụm. 51](#_Toc217073239)

[Bảng 4.2: Kết quả của 2 mô hình máy học: 56](#_Toc217073240)

# CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

## 1.1. Lý do chọn đề tài

Trong những năm gần đây, thị trường bất động sản Việt Nam có nhiều biến động mạnh mẽ, đặc biệt trong giai đoạn 2022–2024 khi nền kinh tế chịu tác động từ lạm phát, chính sách siết tín dụng và các thay đổi về pháp lý. Giá nhà ở liên tục gia tăng tại nhiều khu vực đô thị lớn, tạo ra khoảng cách ngày càng lớn giữa khả năng chi trả của người dân và mặt bằng giá thị trường. Tại các thành phố như Hà Nội và TP. Hồ Chí Minh, vấn đề nhà ở không chỉ mang tính kinh tế mà còn trở thành một vấn đề xã hội được quan tâm sâu sắc.

Bất động sản là lĩnh vực chứa đựng khối lượng dữ liệu rất lớn liên quan đến vị trí, diện tích, đặc điểm vật lý, tình trạng pháp lý, mức độ hoàn thiện nội thất và nhiều yếu tố khác. Tuy nhiên, phần lớn dữ liệu hiện nay mới chỉ được khai thác ở mức mô tả rời rạc, chưa được phân tích một cách hệ thống để làm rõ cơ chế hình thành giá. Trong bối cảnh khoa học dữ liệu và trí tuệ nhân tạo phát triển mạnh mẽ, việc ứng dụng các phương pháp phân tích dữ liệu và học máy để khám phá các yếu tố ảnh hưởng đến giá nhà, cũng như xây dựng mô hình dự đoán giá, trở thành một hướng nghiên cứu có ý nghĩa thực tiễn cao.

Xuất phát từ thực tế đó, nhóm lựa chọn đề tài “Phân tích và Dự đoán Giá Nhà ở Việt Nam từ dữ liệu Batdongsan.com.vn” với mong muốn cung cấp một cái nhìn toàn diện về thị trường nhà ở Việt Nam thông qua phân tích dữ liệu, xác định các yếu tố quan trọng nhất tác động đến giá bất động sản và ứng dụng các mô hình học máy để hỗ trợ dự đoán giá. Bộ dữ liệu được thu thập từ Kaggle năm 2024, có nguồn gốc rõ ràng, quy mô lớn và mang tính cập nhật, phù hợp cho nghiên cứu phân tích dữ liệu và mô hình hóa. Do đó, đề tài không chỉ mang giá trị học thuật mà còn có ý nghĩa ứng dụng đối với người mua nhà, nhà đầu tư và doanh nghiệp bất động sản.

## 1.2. Mục tiêu nghiên cứu

### 1.2.1. Mục tiêu chung

Nghiên cứu hướng đến việc phân tích các yếu tố ảnh hưởng đến giá nhà ở Việt Nam và xây dựng mô hình dự đoán giá nhà dựa trên dữ liệu thực tế từ batdongsan.com.vn (2024).

1.2.2. Mục tiêu cụ thể

* Phân tích thực trạng giá nhà theo khu vực, diện tích, pháp lý, số phòng…
* Tìm ra các yếu tố ảnh hưởng mạnh nhất đến giá bất động sản.
* Tìm ra các nhóm nhà (phân khúc) dựa vào đặc trưng vật lý của bất động sản.
* Xây dựng mô hình dự đoán giá nhà sử dụng các thuật toán học máy ( Random Forest, XGBoos).
* So sánh hiệu suất mô hình và chọn ra mô hình dự đoán tối ưu.
* Đề xuất hàm ý thực tiễn dựa trên kết quả phân tích.

## 1.3. Câu hỏi nghiên cứu

Câu 1: Phân tích theo vị trí Giá nhà ở các thành phố Việt Nam hiện nay phân bố như thế nào giữa các khu vực, và sự khác biệt này phản ánh điều gì về mức độ đô thị hóa, nhu cầu nhà ở và chênh lệch phát triển giữa các thành phố?

Câu 2: Phân tích theo đặc điểm vật lý Giá nhà thay đổi như thế nào theo các đặc điểm vật lý của bất động sản (diện tích, mặt tiền, số tầng, số phòng ngủ, số phòng tắm), và các đặc điểm nào phản ánh rõ nhất sự khác biệt giữa các phân khúc nhà ở (nhỏ – trung bình – lớn)?

Câu 3: Tình trạng pháp lý (Legal status) và mức độ hoàn thiện nội thất (Furniture) ảnh hưởng như thế nào đến giá nhà, và mức độ ảnh hưởng này có khác nhau giữa các khu vực và phân khúc nhà ở hay không?

Câu 4: Thị trường nhà ở có thể được phân chia thành những nhóm (phân khúc) nào dựa trên các đặc điểm vật lý của bất động sản, và mỗi nhóm có đặc điểm giá như thế nào?

Câu 5: Mô hình học máy dự đoán giá nhà dựa trên các đặc điểm của bất động sản và vị trí hoạt động hiệu quả đến mức nào (so sánh và tối ưu hiệu suất mô hình), và những yếu tố nào đóng vai trò quan trọng nhất trong việc quyết định giá nhà (feature importance)?

Câu 6: Từ kết quả phân tích, mô hình dự đoán và phân cụm, có thể rút ra những hàm ý gì cho người mua, nhà đầu tư và nhà quản lý trong việc định giá, đầu tư và phát triển thị trường nhà ở theo hướng minh bạch và bền vững?

## 1.4. Đối tượng và phạm vi nghiên cứu

### 1.4.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là các bất động sản nhà ở tại Việt Nam, với các thuộc tính đặc trưng như vị trí địa lý, diện tích, mặt tiền, đường vào, số tầng, số phòng ngủ, số phòng tắm, tình trạng pháp lý và mức độ hoàn thiện nội thất, cùng với giá bán trên thị trường trong năm 2024. Tất cả các yếu tố này được trích xuất từ dữ liệu đăng tin trên nền tảng batdongsan.com.vn.

### 1.4.2. Phạm vi nghiên cứu

Về không gian, nghiên cứu tập trung vào dữ liệu nhà ở tại các tỉnh và thành phố lớn của Việt Nam, tiêu biểu như TP. Hồ Chí Minh, Hà Nội, Bình Dương và các khu vực lân cận. Về thời gian, dữ liệu được thu thập trong năm 2024, phản ánh tương đối sát thực trạng thị trường hiện tại. Về nội dung, nghiên cứu chỉ tập trung vào phân tích các yếu tố ảnh hưởng và dự đoán giá nhà ở, không xem xét các loại hình bất động sản khác như đất nông nghiệp, đất nền dự án hoặc bất động sản thương mại.

## 1.5. Đóng góp nghiên cứu

Về mặt học thuật, nghiên cứu cung cấp một phân tích thực nghiệm dựa trên bộ dữ liệu bất động sản cập nhật tại Việt Nam, đồng thời so sánh hiệu quả của các thuật toán học máy trong bài toán dự đoán giá nhà. Kết quả nghiên cứu góp phần làm rõ vai trò của các yếu tố vật lý, vị trí và pháp lý trong cơ chế hình thành giá bất động sản, cũng như mở rộng hướng tiếp cận định lượng trong nghiên cứu thị trường nhà ở Việt Nam.

Về mặt thực tiễn, nghiên cứu hỗ trợ người mua nhà, môi giới và nhà đầu tư trong việc ước lượng giá trị bất động sản dựa trên dữ liệu thực tế. Ngoài ra, kết quả phân tích và mô hình dự đoán có thể được sử dụng như một công cụ tham khảo cho doanh nghiệp bất động sản trong quá trình định giá sơ bộ và đánh giá thị trường.

1.6. Cấu trúc nghiên cứu

Nghiên cứu gồm 5 chương:

Chương 1: Giới thiệu đề tài

Chương 2: Cơ sở lý luận và tổng quan nghiên cứu

Chương 3: Phương pháp nghiên cứu

Chương 4: Phân tích dữ liệu & mô hình dự đoán

Chương 5: Kết luận và khuyến nghị

1.7. Giới thiệu tổng quan về Dataset

Bộ dữ liệu House Price Prediction Dataset Vietnam – 2024 được thu thập từ batdongsan.com.vn và công bố trên Kaggle (license CC0 – Public Domain).

### 1.7.1. Cấu trúc dữ liệu

Các cột dữ liệu chính:

* Address: Địa chỉ đầy đủ của bất động sản, bao gồm tên đường, phường/xã, quận/huyện và thành phố. Đây là thông tin quan trọng giúp xác định vị trí địa lý và giá trị khu vực.
* Area (m²): Tổng diện tích sử dụng của căn nhà hoặc lô đất, được tính theo mét vuông. Diện tích là yếu tố ảnh hưởng trực tiếp đến giá trị bất động sản.
* Frontage (m): Chiều rộng mặt tiền của ngôi nhà hoặc lô đất, đo bằng mét. Mặt tiền lớn thường làm tăng giá trị vì phù hợp kinh doanh hoặc buôn bán.
* Access Road (m): Chiều rộng của con đường dẫn vào bất động sản, tính bằng mét. Đường rộng thể hiện khả năng tiếp cận tốt và thường tương quan với giá cao hơn.
* House Direction / Balcony Direction: Hướng của mặt tiền ngôi nhà và hướng của ban công (ví dụ: Đông, Tây, Nam, Bắc). Các thông tin này phản ánh yếu tố phong thủy và ánh sáng, nhưng trong nghiên cứu này có tỷ lệ thiếu cao nên không được sử dụng trong mô hình.
* Floors: Số tầng của căn nhà. Số tầng nhiều hơn thường cho thấy diện tích sử dụng lớn hơn.
* Bedrooms: Số lượng phòng ngủ của bất động sản.
* Bathrooms: Số lượng phòng tắm/toilet đi kèm.
* Legal Status: Tình trạng pháp lý của bất động sản, như có Sổ đỏ/Sổ hồng (Have certificate) đã có đầy đủ giấy chứng nhận quyền sử dụng đất và quyền sở hữu nhà, hợp đồng mua bán (Sale contract) chưa hoàn tất cấp giấy chứng nhận quyền sở hữu. Pháp lý minh bạch làm tăng giá trị và tính thanh khoản.
* Furniture State: Mức độ hoàn thiện nội thất, gồm ba nhóm: đầy đủ (full), một phần (basic), hoặc không nội thất (N/A).
* Price (tỷ VND): Mức giá chào bán của bất động sản, được biểu diễn bằng tỷ Việt Nam đồng. Đây là biến mục tiêu trong mô hình dự báo.

### 1.7.2. Tính uy tín của dữ liệu

Bộ dữ liệu House Price Prediction Dataset Vietnam – 2024 có độ uy tín cao:

* Khả dụng (Usability): 10/10, dữ liệu đầy đủ, dễ sử dụng.
* Độ đầy đủ (Completeness): 100%, không thiếu cột quan trọng.
* Độ tin cậy (Credibility): 100%, nguồn rõ ràng từ batdongsan.com.vn, được Kaggle đánh giá.
* Khả năng tương thích (Compatibility): 100%, định dạng chuẩn, giấy phép CC0 – Public Domain, có thể sử dụng cho nghiên cứu, học tập và thương mại.
* Mô tả cột dữ liệu (Column Description): Mỗi cột có mô tả rõ ràng, hỗ trợ tiền xử lý và phân tích.

### 1.7.3. Các biến dữ liệu đầu vào và đầu ra

Trong nghiên cứu này, các biến dữ liệu được chia thành hai nhóm chính gồm biến đầu vào và biến đầu ra nhằm phục vụ cho quá trình phân tích và xây dựng mô hình dự đoán giá nhà.

Nhóm biến đầu vào bao gồm cả các biến định lượng và biến danh mục, phản ánh toàn diện đặc điểm vật lý, vị trí và trạng thái của bất động sản. Cụ thể, các biến định lượng gồm diện tích sử dụng (Area), chiều rộng mặt tiền (Frontage), độ rộng đường vào (Access Road), số tầng (Floors), số phòng ngủ (Bedrooms) và số phòng tắm (Bathrooms), là những yếu tố thể hiện quy mô, mức độ tiện nghi và khả năng khai thác công năng của ngôi nhà. Bên cạnh đó, các biến danh mục như thành phố và quận/huyện (City, District), tình trạng pháp lý (Legal Status) và mức độ hoàn thiện nội thất (Furniture State) phản ánh đặc điểm vị trí, tính minh bạch về pháp lý và chất lượng hoàn thiện của bất động sản, qua đó có tác động đáng kể đến giá trị thị trường.

Biến đầu ra của nghiên cứu là giá bán bất động sản (Price), được đo lường bằng tỷ đồng Việt Nam, đóng vai trò là biến mục tiêu trong các mô hình hồi quy và học máy nhằm dự đoán giá nhà dựa trên tập hợp các đặc trưng đầu vào nêu trên.

# CHƯƠNG 2. CƠ SỞ LÝ LUẬN

## 2.1. Tổng quan nghiên cứu

Giá bất động sản là một trong những chủ đề được nghiên cứu rộng rãi trong lĩnh vực kinh tế và tài chính đô thị trên thế giới cũng như tại Việt Nam. Các nghiên cứu trước đây chủ yếu tập trung vào việc xác định những yếu tố ảnh hưởng đến giá nhà như diện tích, số phòng, vị trí địa lý, tình trạng pháp lý và điều kiện thị trường. Trong bối cảnh quốc tế, nhiều công trình đã chỉ ra rằng giá trị bất động sản được hình thành từ sự kết hợp của các yếu tố vật lý, vị trí và môi trường xung quanh, đồng thời chịu tác động mạnh từ mức độ phát triển hạ tầng và khả năng kết nối giao thông.

Tại Việt Nam, các nghiên cứu về giá nhà vẫn còn tương đối hạn chế, đặc biệt là các nghiên cứu sử dụng dữ liệu lớn và dữ liệu trực tuyến có tính cập nhật cao. Phần lớn các nghiên cứu trước đây chủ yếu dựa trên phương pháp thống kê truyền thống hoặc tập trung vào một số khu vực cụ thể như Hà Nội và TP. Hồ Chí Minh. Việc ứng dụng các mô hình học máy hiện đại trong dự đoán giá bất động sản vẫn chưa được khai thác đầy đủ, trong khi thị trường nhà ở Việt Nam có đặc điểm phi tuyến mạnh và chịu ảnh hưởng lớn bởi yếu tố vị trí và pháp lý. Do đó, việc mở rộng nghiên cứu theo hướng kết hợp dữ liệu thực tế và các phương pháp phân tích hiện đại là cần thiết.

## 2.2. Khoảng trống nghiên cứu

Mặc dù đã có nhiều nghiên cứu liên quan đến định giá bất động sản, vẫn tồn tại một số khoảng trống đáng chú ý. Thứ nhất, các nghiên cứu sử dụng dữ liệu bất động sản trực tuyến có tính cập nhật cao từ các nền tảng như batdongsan.com.vn còn khá hạn chế tại Việt Nam. Thứ hai, việc áp dụng và so sánh các thuật toán học máy hiện đại như Random Forest và XGBoost trong bài toán dự đoán giá nhà chưa được thực hiện một cách hệ thống. Thứ ba, nhiều nghiên cứu chỉ tập trung vào một số thành phố lớn mà chưa phân tích toàn diện sự khác biệt theo khu vực, đặc điểm vật lý và tình trạng pháp lý của bất động sản.

Nghiên cứu này được thực hiện nhằm lấp đầy những khoảng trống trên bằng cách sử dụng bộ dữ liệu bất động sản cập nhật năm 2024, kết hợp phân tích khám phá dữ liệu, phân cụm và mô hình học máy để đánh giá các yếu tố ảnh hưởng đến giá nhà ở Việt Nam một cách toàn diện và thực tiễn hơn.

## 2.3. Cơ sở lý thuyết về định giá bất động sản

### 2.3.1. Lý thuyết Hedonic Pricing

Lý thuyết Hedonic Pricing cho rằng giá trị của một bất động sản không được quyết định bởi một yếu tố riêng lẻ mà là kết quả tổng hợp của nhiều đặc điểm cấu thành khác nhau, bao gồm đặc điểm vật lý, vị trí địa lý và các yếu tố môi trường xung quanh. Theo Rosen (1974), mỗi đặc tính của bất động sản đều mang một mức giá ngầm định và tổng giá trị thị trường được hình thành từ sự kết hợp của các mức giá này. Trong nghiên cứu giá nhà, mô hình Hedonic thường được triển khai thông qua các phương pháp hồi quy nhằm ước lượng mức độ đóng góp của từng yếu tố như diện tích, số phòng, mặt tiền hay tình trạng pháp lý. Tuy nhiên, trong bối cảnh thị trường bất động sản có tính phi tuyến cao và sự khác biệt lớn giữa các khu vực như tại Việt Nam, việc mở rộng lý thuyết Hedonic bằng các phương pháp phân tích dữ liệu hiện đại là cần thiết để phản ánh đầy đủ hơn cơ chế hình thành giá.

### 2.3.2. Lý thuyết vị trí (Location Theory)

Lý thuyết vị trí trong kinh tế đô thị nhấn mạnh vai trò then chốt của yếu tố không gian địa lý đối với giá trị bất động sản. Theo các nghiên cứu kinh điển của Alonso (1964) và Fujita (1989), giá nhà có xu hướng cao hơn tại các khu vực gần trung tâm đô thị, nơi tập trung nhiều hoạt động kinh tế, dịch vụ và hạ tầng giao thông, và giảm dần theo khoảng cách ra khu vực ngoại thành. Yếu tố vị trí không chỉ phản ánh khoảng cách địa lý mà còn bao gồm khả năng tiếp cận các tiện ích công cộng như trường học, bệnh viện, trung tâm thương mại và hệ thống giao thông. Thực tiễn thị trường bất động sản Việt Nam cho thấy sự chênh lệch giá rõ rệt giữa các quận trung tâm và vùng ven, khẳng định vai trò quan trọng của vị trí địa lý trong việc hình thành giá nhà.

### 2.3.3. Đặc điểm vật lý của bất động sản

Các đặc điểm vật lý của bất động sản đóng vai trò quan trọng trong việc quyết định giá trị thị trường do chúng phản ánh trực tiếp quy mô, công năng và mức độ tiện nghi của ngôi nhà. Những yếu tố như diện tích sử dụng, chiều rộng mặt tiền, độ rộng đường vào, số tầng, số phòng ngủ và số phòng tắm thường có mối quan hệ chặt chẽ với giá bán. Bất động sản có diện tích lớn, mặt tiền rộng và khả năng tiếp cận giao thông thuận lợi thường được định giá cao hơn do đáp ứng tốt hơn nhu cầu sinh hoạt và kinh doanh. Đồng thời, số tầng và số phòng thể hiện khả năng khai thác không gian và mức độ phù hợp với các hộ gia đình có quy mô khác nhau. Việc phân tích các đặc điểm vật lý này giúp làm rõ sự khác biệt giữa các phân khúc nhà ở trên thị trường.

### 2.3.4. Lý thuyết về pháp lý và quyền sở hữu

Trong lĩnh vực bất động sản, tình trạng pháp lý và quyền sở hữu được xem là một trong những yếu tố nền tảng quyết định giá trị và tính thanh khoản của tài sản. Các yếu tố pháp lý như giấy chứng nhận quyền sử dụng đất và quyền sở hữu nhà ở (thường được gọi là sổ đỏ hoặc sổ hồng), hợp đồng mua bán hợp lệ, cũng như sự rõ ràng trong quyền sở hữu và quyền sử dụng đóng vai trò quan trọng trong việc đảm bảo tính an toàn pháp lý cho các bên tham gia giao dịch. Những bất động sản có hồ sơ pháp lý đầy đủ và minh bạch thường được thị trường đánh giá cao hơn do giúp giảm thiểu rủi ro tranh chấp, rút ngắn thời gian giao dịch và tạo niềm tin cho người mua.

Ngược lại, các bất động sản có pháp lý chưa hoàn chỉnh hoặc tồn tại các vấn đề liên quan đến quyền sở hữu thường gặp khó khăn trong việc chuyển nhượng, vay vốn ngân hàng hoặc khai thác giá trị kinh tế, từ đó làm giảm mức độ hấp dẫn và giá bán trên thị trường. Trong bối cảnh thị trường bất động sản Việt Nam, nơi khung pháp lý còn đang hoàn thiện và nhận thức của người mua ngày càng cao, yếu tố pháp lý không chỉ ảnh hưởng đến quyết định mua bán mà còn tác động trực tiếp đến khả năng thanh khoản của bất động sản. Do đó, những căn nhà có quyền sở hữu rõ ràng và pháp lý minh bạch thường được định giá cao hơn so với các bất động sản tương đồng về đặc điểm vật lý nhưng có mức độ rủi ro pháp lý lớn hơn.

### 2.3.5. Lý thuyết học máy trong dự báo giá

Khái niệm về một số thuật toán học máy phổ biến dùng trong phân tích dự báo giá nhà:

Random Forest: Tổ hợp nhiều cây quyết định, giảm overfitting và nâng cao độ chính xác dự đoán. Thích hợp với dữ liệu có nhiều thuộc tính và tương tác phức tạp.

XGBoost (Extreme Gradient Boosting): XGBoost là thuật toán boosting mạnh mẽ dựa trên việc xây dựng tuần tự nhiều cây quyết định, trong đó mỗi cây mới được huấn luyện để khắc phục sai số còn lại của các cây trước đó. XGBoost được tối ưu hóa về tốc độ, khả năng xử lý quan hệ phi tuyến và có cơ chế regularization (L1, L2) giúp cải thiện khả năng tổng quát hóa. Nhờ đó, XGBoost thường cho độ chính xác cao trong các bài toán dự báo giá nhà và được sử dụng rộng rãi trong các cuộc thi dữ liệu thực tế.

Việc áp dụng các thuật toán học máy này giúp mô hình dự báo giá dựa trên nhiều yếu tố cùng lúc, nhận diện các mối quan hệ phức tạp trong dữ liệu và tăng cường độ chính xác so với các mô hình hồi quy tuyến tính truyền thống.

# CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU

## 3.1. Quy trình nghiên cứu

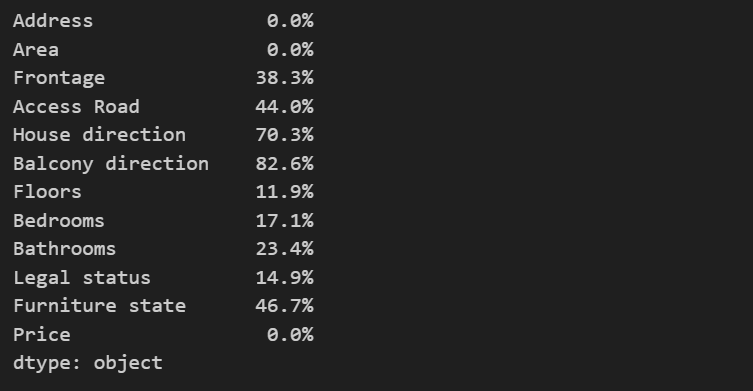
Nghiên cứu được thực hiện theo các bước chính:

1. Thu thập dữ liệu từ batdongsan.com.vn / Kaggle 2024.
2. Đánh giá và tiền xử lý dữ liệu, loại bỏ giá trị thiếu hoặc bất thường.
3. Phân tích khám phá (EDA) và trực quan hóa dữ liệu để hiểu phân bố và mối quan hệ giữa các thuộc tính.
4. Dùng thuật toán K-means để phân cụm các đặc trưng của bất động sản.
5. Xây dựng và huấn luyện mô hình học máy (Random Forest, XGBoost).
6. Đánh giá mô hình, lựa chọn mô hình tối ưu và diễn giải kết quả.
7. Rút ra kết luận và đề xuất ứng dụng thực tiễn.

## 3.2. Thu thập và tiền xử lý dữ liệu

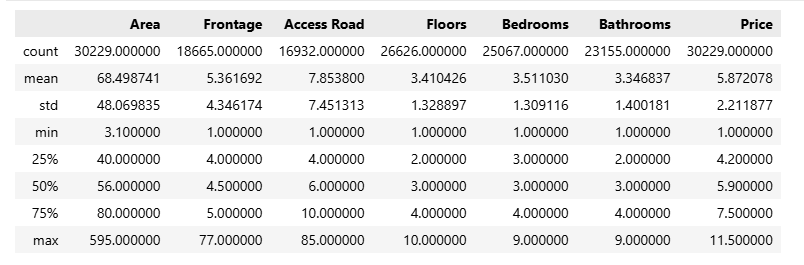
### 3.2.1. Đánh giá chất lượng dữ liệu ban đầu

Bộ dữ liệu ban đầu gồm hơn 30.229 mẫu thông tin bất động sản, đảm bảo quy mô đủ lớn để thực hiện phân tích thống kê và xây dựng các mô hình dự đoán có độ tin cậy cao. Với hơn 10.265 địa chỉ duy nhất trên toàn bộ 30.229 bản ghi, dữ liệu thể hiện mức độ phủ rộng theo khu vực rất tốt, phản ánh đa dạng nhiều loại hình nhà ở tại các thành phố và tỉnh thành khác nhau. Sự phân tán này giúp mô hình học máy thu được thông tin phong phú về đặc trưng vị trí, đồng thời giảm nguy cơ bị lệch do tập trung vào một khu vực quá hẹp. Tuy nhiên, số lượng địa chỉ duy nhất lớn cũng làm tăng tính biến động của dữ liệu thô, do đó cần trích xuất thêm các biến tổng quát hơn như City và District để mô hình khai thác thông tin vị trí hiệu quả và ổn định hơn.



#### Hình 3.1: Phần trăm dữ liệu bị thiếu ở các cột

Tỷ lệ dữ liệu bị thiếu giữa các biến trong bộ dữ liệu có sự chênh lệch đáng kể. Các biến cốt lõi gồm Address, Area và Price đều không có giá trị thiếu, đảm bảo nền tảng dữ liệu đủ vững cho các bước phân tích và xây dựng mô hình. Tuy nhiên, nhiều biến quan trọng khác lại tồn tại tỷ lệ thiếu tương đối cao. Đặc biệt, các biến House direction (70.3%) và Balcony direction (82.6%) có mức thiếu vượt quá 70%, cho thấy phần lớn tin đăng không cung cấp thông tin về hướng nhà hoặc hướng ban công; do đó hai biến này khó có khả năng sử dụng trong phân tích định lượng. Các biến liên quan đến cấu trúc và đặc điểm vật lý như Frontage (38.3%), Access Road (44.0%) và Furniture state (46.7%) cũng bị thiếu từ 40–50%, phản ánh sự không đồng nhất trong cách người dùng cung cấp thông tin trên nền tảng bất động sản. Các biến Floors, Bedrooms và Bathrooms có tỷ lệ thiếu từ 11% đến 23%, ở mức có thể xử lý thông qua phương pháp điền giá trị theo nhóm địa lý hoặc nhóm diện tích. Nhìn chung, dù một số biến bị thiếu nhiều cần loại bỏ hoặc xử lý đặc biệt, bộ dữ liệu vẫn đảm bảo chất lượng tốt ở các trường quan trọng, đủ điều kiện để tiếp tục các bước tiền xử lý và mô hình hóa.



#### Hình 3.2: Thống kê mô tả các cột số

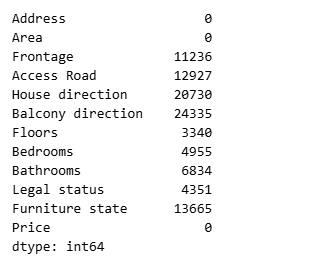
Kiểm tra thống kê mô tả cũng cho thấy sự xuất hiện giá trị ngoại lai (outliers) ở các biến số như diện tích, mặt tiền, độ rộng đường và số phòng. Một số giá trị có độ lệch lớn so với phân bố chung (ví dụ: Area đến 595 m², Frontage 77 m, Access Road 85 m). Những ngoại lai này có thể đến từ lỗi nhập liệu, tài sản đặc biệt hiếm hoặc phân khúc không phù hợp với phạm vi nghiên cứu. Riêng biến Price hầu như không tồn tại ngoại lai rõ rệt, phản ánh phân bố giá khá hợp lý và phù hợp với thị trường thực tế.

Nhìn chung, bộ dữ liệu có chất lượng tốt ở các biến cốt lõi nhưng vẫn cần thực hiện các bước làm sạch như xử lý giá trị thiếu và loại bỏ hoặc điều chỉnh ngoại lai để đảm bảo độ tin cậy cho các phân tích và mô hình dự báo ở các phần tiếp theo.

### 3.2.2. Tiền xử lý dữ liệu

A. Loại bỏ ngoại lai giai đoạn 1 (diện tích bất thường)

Loại các căn nhà có diện tích < 20 m² hoặc > 200 m² vì đây chủ yếu là kios, nhà kho hoặc biệt thự siêu sang – không đại diện cho phân khúc nhà ở phổ biến tại Việt Nam. Kết quả còn 29.406 bản ghi. Kiểm tra lại số lượng dữ liệu bị thiếu.



#### Hình 3.3: Kiểm tra số lượng dữ liệu bị thiếu

House direction & Balcony direction: thiếu >70–80% → quá nhiều.Frontage, Access Road, Bedrooms, Bathrooms… cũng thiếu khá nhiều

B. Xóa 2 cột thiếu quá nhiều dữ liệu Balcony direction và House direction

Vì thiếu >70% → điền vào cũng không có ý nghĩa, dễ gây nhiễu cho dữ liệu khi phân tích.

C. Điền giá trị thiếu theo nhóm địa lý + nhóm diện tích

Thay vì điền median/mode toàn bộ (dễ gây bias), thực hiện điền giá trị thiếu theo các bước :

Tạo nhóm diện tích:Nhóm 1: Diện tích dưới 40 m² (thường là nhà nhỏ, nhà phố hẻm, căn hộ mini), Nhóm 2: Từ 40 đến dưới 70 m² (nhà phố phổ thông, chiếm tỷ lệ lớn), Nhóm 3: Từ 70 đến dưới 150 m² (nhà phố trung-cao cấp, biệt thự nhỏ), Nhóm 4: Từ 150 m² trở lên (biệt thự, nhà mặt tiền lớn).Việc chia nhóm này giúp khi điền giá trị thiếu của các cột số tầng, số phòng ngủ, phòng tắm, mặt tiền… thì hệ thống sẽ ưu tiên lấy giá trị phổ biến của chính nhóm diện tích đó tại cùng khu vực địa chỉ (Address), thay vì lấy trung bình của cả nước

Điền median theo nhóm (Address + nhóm diện tích) cho các cột:Floors (số tầng), Bedrooms (số phòng ngủ),Bathrooms (số phòng tắm),Frontage (mặt tiền).

Nếu vẫn còn thiếu → điền median theo Address.Nếu vẫn còn thiếu → điền median toàn bộ dữ liệu

Riêng Access Road (đường vào nhà) điền mode(giá trị xuất hiện nhiều nhất)

theo Address (vì đường cùng khu vực thường giống nhau).

Áp dụng tương tự cho Bedrooms, Bathrooms, Frontage.

A screenshot of a computer

AI-generated content may be incorrect.

#### Hình 3.4: Kiểm tra lại dữ liệu bị thiếu

D. Điền các giá trị còn thiếu cuối cùng

Sau bước trên vẫn còn một lượng nhỏ dữ liệu ở vài cột trong dữ liệu còn tình trạng NaN:

- Floors, Bedrooms, Bathrooms → điền = 1 (nhà nhỏ nhất vẫn phải có ít nhất 1 tầng/phòng).

- Frontage, Access Road → điền median toàn bộ.

- Legal status → điền “Sale contract” (giả định chưa có sổ đỏ).

-Furniture state → điền “No furnishings” (giả định không có nội thất).

E. Chuyển kiểu dữ liệu Bedrooms và Bathrooms từ float → int (làm tròn xuống)

Sau khi điền giá trị thiếu bằng median theo nhóm, các cột Bedrooms và Bathrooms có kiểu float (ví dụ 3.5). Do số lượng phòng ngủ và phòng tắm trong thực tế luôn là số nguyên.

F. Xử lý bản ghi trùng lặp

Phát hiện 305 dòng hoàn toàn trùng lặp → xóa bỏ: Còn lại 29.101 bản ghi sau khi xóa.

G. Loại bỏ ngoại lai giai đoạn 2 (các cột còn lại)

Sau khi đã điền thiếu, tiếp tục loại bỏ các giá trị bất thường không hợp lý còn lại:

- Mặt tiền < 2 m hoặc > 30 m hầu như không tồn tại ở nhà phố thông thường (thường là nhà tạm, kios hoặc biệt thự siêu lớn).

- Đường vào nhà < 2 m gần như không thể có xe hơi ra vào, > 12 m thường chỉ xuất hiện ở các dự án biệt thự cao cấp hoặc dữ liệu sai.

- Nhà ở Việt Nam hiếm khi có trên 7 tầng nếu không phải chung cư (đối tượng nghiên cứu là nhà phố, biệt thự, nhà riêng lẻ).

-Số phòng ngủ và phòng tắm trên 8 phòng thường thuộc phân khúc siêu sang hoặc dữ liệu nhập sai.

Kết quả: loại bỏ thêm 3.525 bản ghi bất thường, còn lại 25.576 bản ghi.

H. Xử lý dữ liệu bị lỗi

Kiểm tra tính hợp lý vật lý và logic của dữ liệu bất động sản dựa trên kinh nghiệm thực tế thị trường nhà ở Việt Nam. Mặt tiền (Frontage) không hợp lý so với diện tích (Area)

* Nếu mặt tiền ≥ 12 m thì diện tích phải ≥ 80 m²
* Nếu mặt tiền 8–12 m thì diện tích phải ≥ 60 m²
* Nếu mặt tiền 5–8 m thì diện tích phải ≥ 40 m²
* Nếu mặt tiền < 5 m thì diện tích phải ≥ 20 m² (Một căn nhà không thể có mặt tiền quá rộng trong khi tổng diện tích rất nhỏ.)
* Số phòng ngủ không hợp lý so với diện tích:Mỗi phòng ngủ cần ít nhất khoảng 9 m² → Bedrooms > Area / 9 là bất thường.
* Số phòng tắm không hợp lý so với số phòng ngủ:Thông thường số phòng tắm không vượt quá 2 lần số phòng ngủ (ví dụ: nhà 4 phòng ngủ hiếm khi có hơn 8 phòng tắm).

Kết quả kiểm tra:

* Số dòng lỗi về mặt tiền: 1073 bản ghi.
* Số dòng lỗi về số phòng ngủ: 2479 bản ghi.
* Số dòng lỗi về số phòng tắm: 37 bản ghi.
* Tổng số dòng có ít nhất một lỗi logic: 3369 bản ghi.
* Số bản ghi sau khi loại bỏ dòng lỗi: 22207 bản ghi.

I. Thêm 2 cột City và District từ cột Address Chỉnh lại các Thành phố/Tỉnh cho thống nhất

Từ cột Address (dạng chuỗi dài) tách ra 2 cột mới City và District rất cần thiết để phân tích theo khu vực.

- Loại bỏ giá trị rỗng hoặc NaN.

- Loại bỏ hoàn toàn các dòng bị lỗi do người đăng nhập sai vị trí tỉnh/thành Ví dụ: cuối địa chỉ lại ghi "giá 3 tỷ", "giá 6ty", "giá 2tỷ380"…

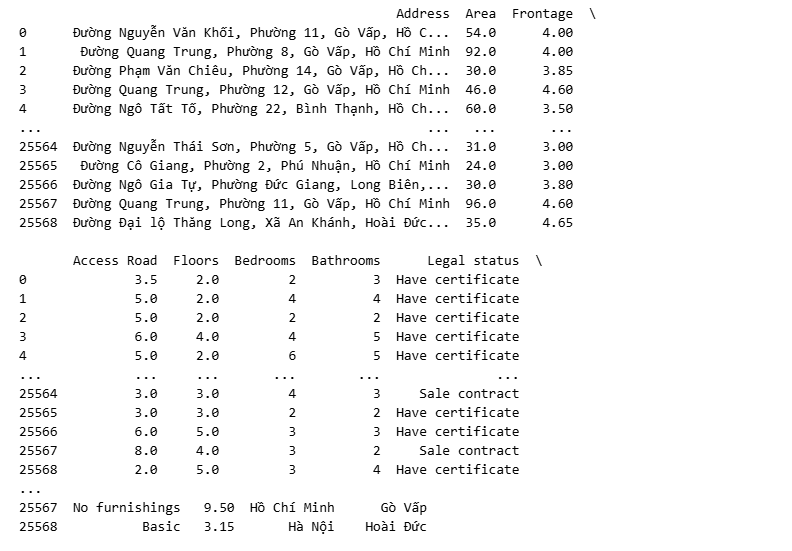
- Loại bỏ một số giá trị rõ ràng sai:"phòng công chứng nguyễn thị thành", "Quận Nam Từ Liêm", "TP Cam Ranh","Đường số 11".

- Chuẩn hóa tên tỉnh/thành phố về dạng chính thức: "TP Hồ Chí Minh": "Hồ Chí Minh", "TP HCM": "Hồ Chí Minh", "TP. HCM": "Hồ Chí Minh", "TPHCM": "Hồ Chí Minh", "HCM": "Hồ Chí Minh", "HN": "Hà Nội", "Quận 8": "Hồ Chí Minh", "Quận Bình Thạnh": "Hồ Chí Minh", "Bán nhà chính chủ Phó Đức Chính khu Bà Chiểu - trung tâm Bình Thạnh giá cực tốt":"Hồ Chí Minh", "Hồ Chí Minh giá 2tỷ380":"Hồ Chí Minh", "TpHCM": "Hồ Chí Minh", "Hồ Chí Mính": "Hồ Chí Minh", "Bình Dương (gần cafe Xóm Vắng 2)": "Bình Dương".

- Cuối cùng còn 55 tỉnh/thành chính thức .

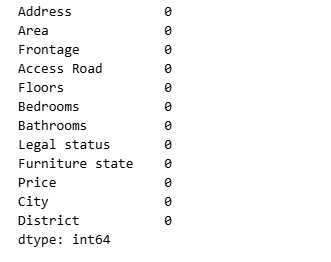
J. Làm sạch tên Quận/Huyện

* Loại bỏ các giá trị rỗng hoặc không hợp lệ (None, chuỗi trắng).
* Giữ nguyên dạng “Quận + số” (ví dụ: Quận 1, Quận 12, Quận 7) → chuẩn hóa viết hoa đầu: Quận 1, Quận 9, Quận 12.
* Với các quận có tên riêng (ví dụ: Quận Bình Thạnh, Quận Gò Vấp) → bỏ từ “Quận” ở đầu → chỉ giữ lại tên: Bình Thạnh, Gò Vấp, Tân Bình.
* Với huyện/thành phố trực thuộc tỉnh → loại bỏ tiền tố “Huyện”, “Thành phố”, “Thị xã”:Ví dụ: Huyện Long Thành → Long Thành, Thành phố Dĩ An → Dĩ An , Thị xã Sa Pa → Sa Pa , Huyện Nhà Bè → Nhà Bè.
* Tiếp tục loại bỏ hoàn toàn các bản ghi không hợp lệ còn sót lại:
* Loại bỏ các hàng có District = None (do lỗi tách chuỗi).
* Loại bỏ các hàng mà giá trị District thực chất là phường/xã/thị trấn (do lỗi tách sai cấp hành chính), ví dụ: “Phường Linh Trung”, “Xã An Khánh”, “Thị trấn Tân Túc”…



#### Hình 3.5: Kết quả sau khi xử lý

K. Kiểm tra cuối cùng và lưu dữ liệu sạch



#### Hình 3.6: Kết quả sau khi xử lý dữ liệu

Kết quả*:* Không còn giá trị NaN nào, không còn bản ghi trùng lặp, kiểu dữ liệu hợp lý.

## 3.3. Phương pháp phân tích khám phá và trực quan hóa dữ liệu

### 3.3.1 Phân tích thống kê mô tả

Tính các chỉ số trung bình, trung vị, độ lệch chuẩn, độ lệch (skewness) và độ nhọn (kurtosis) của các biến định lượng như diện tích, mặt tiền, số tầng, số phòng ngủ, phòng tắm và giá bán. Các chỉ số này giúp nhận diện sự bất đối xứng trong phân bố và phát hiện sớm các giá trị ngoại lai

### 3.3.2 Trực quan hóa phân bố dữ liệu

Sử dụng biểu đồ histogram kết hợp đường cong mật độ (KDE), biểu đồ hộp (boxplot) và biểu đồ barchart để quan sát hình dạng phân bố, giá trị tập trung và các điểm ngoại lai của từng biến. Đặc biệt, biểu đồ hộp được sử dụng rộng rãi để so sánh phân bố giá nhà giữa các nhóm (theo thành phố, quận/huyện, pháp lý, nội thất).

### 3.3.3 Phân tích mối quan hệ giữa các biến Biểu đồ phân tán

Trong quá trình phân tích mối quan hệ giữa các biến, nghiên cứu sử dụng biểu đồ phân tán kết hợp với đường hồi quy tuyến tính nhằm đánh giá mức độ tương quan và xu hướng biến động của giá bán bất động sản theo các đặc trưng vật lý. Các biến như diện tích, chiều rộng mặt tiền, độ rộng đường vào, số tầng và số phòng được xem xét để xác định mối quan hệ giữa từng đặc điểm và giá nhà. Thông qua biểu đồ phân tán, nghiên cứu có thể quan sát trực quan xu hướng tăng hoặc giảm của giá bán khi các đặc trưng vật lý thay đổi, đồng thời nhận diện các mô hình phi tuyến hoặc hiện tượng phân tán không đồng đều trong dữ liệu. Việc kết hợp đường hồi quy tuyến tính giúp làm rõ chiều hướng và mức độ ảnh hưởng trung bình của từng biến, qua đó hỗ trợ đánh giá vai trò của các đặc trưng vật lý trong việc hình thành giá bất động sản và cung cấp cơ sở cho việc lựa chọn biến trong các mô hình dự đoán tiếp theo.

## 3.4. Phân cụm bất động sản bằng K-means

Nhằm khám phá các phân khúc tiềm ẩn của thị trường bất động sản dựa trên các đặc điểm vật lý của căn nhà, nghiên cứu áp dụng phương pháp phân cụm không giám sát. Phân cụm cho phép nhóm các bất động sản có đặc điểm tương đồng vào cùng một nhóm, từ đó hỗ trợ phân tích cấu trúc thị trường và làm cơ sở cho các phân tích chuyên sâu ở các chương tiếp theo. Trong nghiên cứu này, thuật toán K-means được lựa chọn để thực hiện quá trình phân cụm.

### 3.4.1. Giới thiệu phương pháp K-means

K-means là một thuật toán phân cụm không giám sát (unsupervised learning), được sử dụng phổ biến trong phân tích dữ liệu nhằm chia tập dữ liệu thành K nhóm (cụm) sao cho các quan sát trong cùng một cụm có mức độ tương đồng cao, trong khi sự khác biệt giữa các cụm là lớn nhất. Mỗi cụm được đại diện bởi một tâm cụm (centroid), là giá trị trung bình của các điểm dữ liệu thuộc cụm đó.

Quy trình hoạt động của thuật toán K-means bao gồm các bước cơ bản:

1. Khởi tạo K tâm cụm ban đầu (thường ngẫu nhiên).
2. Gán mỗi điểm dữ liệu vào cụm có tâm gần nhất dựa trên khoảng cách Euclidean.
3. Cập nhật lại vị trí tâm cụm dựa trên các điểm dữ liệu đã được gán.
4. Lặp lại các bước trên cho đến khi các tâm cụm hội tụ hoặc không còn sự thay đổi đáng kể.

K-means có ưu điểm là dễ triển khai, hiệu quả về mặt tính toán và phù hợp với các tập dữ liệu có kích thước lớn. Tuy nhiên, thuật toán phụ thuộc vào việc lựa chọn số cụm K và nhạy cảm với các giá trị ngoại lệ, do đó cần kết hợp với các phương pháp đánh giá phù hợp trong quá trình xác định số cụm.

### 3.4.2. Lựa chọn biến cho phân cụm

Mục tiêu của phân cụm trong nghiên cứu là phân khúc bất động sản dựa trên đặc điểm vật lý và công năng sử dụng, do đó các biến được lựa chọn cho quá trình phân cụm bao gồm: diện tích sử dụng (Area), chiều ngang mặt tiền (Frontage), độ rộng đường vào (Access Road), số tầng (Floors), số phòng ngủ (Bedrooms), số phòng tắm (Bathrooms)

Các biến này phản ánh quy mô, mức độ tiện nghi và khả năng tiếp cận của bất động sản, là những yếu tố có ý nghĩa trong việc phân loại và hình thành các phân khúc nhà ở trên thị trường.

Biến giá bán (Price) không được đưa vào quá trình phân cụm nhằm đảm bảo việc phân nhóm được thực hiện một cách khách quan, tránh việc giá nhà chi phối trực tiếp cấu trúc cụm. Trước khi tiến hành phân cụm, các biến được chuẩn hóa bằng phương pháp chuẩn hóa Z-score (Standardization) để đưa về cùng thang đo, giúp cải thiện độ ổn định và độ chính xác của thuật toán K-means.

### 3.4.3. Xác định số cụm tối ưu (Elbow & Silhouette)

Việc xác định số cụm K phù hợp là một bước quan trọng trong quá trình phân cụm bằng K-means. Trong nghiên cứu này, hai phương pháp phổ biến được sử dụng để hỗ trợ lựa chọn số cụm tối ưu là phương pháp Elbow và chỉ số Silhouette.

Phương pháp Elbow dựa trên việc phân tích tổng bình phương khoảng cách trong cụm (Within-Cluster Sum of Squares – WCSS). Khi số cụm K tăng, giá trị WCSS có xu hướng giảm do các điểm dữ liệu được phân bổ vào các cụm nhỏ hơn. Việc quan sát đồ thị WCSS theo K giúp xác định điểm mà tại đó tốc độ giảm của WCSS bắt đầu chậm lại, thể hiện sự đánh đổi hợp lý giữa số lượng cụm và mức độ đồng nhất trong cụm.

Bên cạnh đó, chỉ số Silhouette được sử dụng để đánh giá chất lượng phân cụm dựa trên mức độ gắn kết của các điểm dữ liệu trong cùng cụm và mức độ tách biệt giữa các cụm khác nhau. Chỉ số này cung cấp một thước đo định lượng để so sánh hiệu quả phân cụm giữa các giá trị K khác nhau, hỗ trợ quá trình lựa chọn số cụm phù hợp cho dữ liệu.

Các phương pháp trên được sử dụng nhằm đảm bảo số cụm được lựa chọn có cơ sở khoa học, phục vụ tốt cho việc phân tích và diễn giải kết quả phân cụm ở các chương tiếp theo của báo cáo.

## 3.5. Phương pháp luận học máy

### 3.5.1. Bài toán hồi quy

Trong nghiên cứu này, mục tiêu chính là xây dựng mô hình dự đoán giá bán của bất động sản dựa trên các đặc trưng mô tả ngôi nhà và vị trí của nó. Đây là một bài toán hồi quy điển hình, trong đó biến đầu ra là giá trị liên tục (đơn vị tỷ đồng). Mô hình được huấn luyện nhằm tìm ra mối quan hệ giữa giá bán và các yếu tố đầu vào như diện tích, số phòng ngủ, số tầng, tình trạng pháp lý, nội thất hay khu vực địa lý. Dữ liệu sau khi làm sạch và mã hóa được chia thành tập huấn luyện và tập kiểm tra để đảm bảo mô hình học được xu hướng tổng quát thay vì ghi nhớ dữ liệu.

### 3.5.2. Các thuật toán học máy được áp dụng

Trong nghiên cứu này, hai thuật toán chính được sử dụng để dự đoán giá bán bất động sản, bao gồm Random Forest và XGBoost. Đây là hai phương pháp học máy mạnh mẽ, phù hợp cho các bài toán hồi quy với dữ liệu có nhiều đặc trưng và mối quan hệ phi tuyến.

Thuật toán Random Forest thuộc nhóm mô hình rừng cây quyết định, hoạt động bằng cách xây dựng một tập hợp nhiều cây quyết định độc lập và tổng hợp dự đoán từ tất cả các cây để đưa ra giá trị cuối cùng. Cách tiếp cận này giúp mô hình giảm sai số phương sai, ít bị ảnh hưởng bởi nhiễu trong dữ liệu và hạn chế hiện tượng quá khớp. Random Forest cũng cho phép tính toán mức độ quan trọng của từng đặc trưng, nhờ đó hỗ trợ đánh giá yếu tố nào ảnh hưởng mạnh nhất đến giá nhà đây là một lợi thế quan trọng cho bài toán phân tích kinh tế.

Thuật toán XGBoost (Extreme Gradient Boosting) là một phiên bản nâng cao của mô hình boosting, xây dựng tuần tự nhiều cây quyết định sao cho mỗi cây mới sẽ học từ phần sai số còn lại của các cây trước đó. XGBoost sử dụng cơ chế tối ưu hóa dựa trên đạo hàm bậc hai, regularization (L1, L2) và kỹ thuật giảm độ phức tạp mô hình, nhờ đó đạt hiệu suất dự đoán cao và khả năng tổng quát hóa tốt ngay cả khi dữ liệu có nhiễu. So với Random Forest, XGBoost có khả năng nắm bắt các quan hệ phi tuyến phức tạp tốt hơn và thường cho sai số dự đoán thấp hơn khi được tối ưu tham số đúng cách.

Việc lựa chọn hai thuật toán này giúp nghiên cứu cân bằng giữa tính ổn định và khả năng giải thích của mô hình (Random Forest) với sức mạnh dự đoán và hiệu quả trong việc mô hình hóa dữ liệu phức tạp (XGBoost). Sau khi huấn luyện, hai mô hình được đánh giá trên cùng một bộ dữ liệu nhằm so sánh hiệu quả và xác định mô hình phù hợp nhất cho bài toán dự đoán giá nhà ở thị trường Việt Nam.

### 3.5.3. Phương pháp đánh giá mô hình

Để đánh giá mức độ chính xác của các mô hình dự đoán giá, nghiên cứu sử dụng kết hợp nhiều chỉ số phổ biến trong bài toán hồi quy, bao gồm sai số tuyệt đối trung bình (MAE), căn sai số bình phương trung bình (RMSE) và hệ số xác định R² . Các chỉ số này phản ánh không chỉ sai số trung bình giữa giá dự đoán và giá thực tế mà còn thể hiện mức độ mô hình giải thích được phương sai của dữ liệu. Dữ liệu được chia thành hai tập độc lập gồm tập huấn luyện để xây dựng mô hình và tập kiểm tra để đánh giá hiệu quả tổng quát, giúp giảm nguy cơ overfitting và đảm bảo mô hình có thể ứng dụng trong thực tế.

### 3.5.4. Kiểm định giả thuyết thống kê

Trong nghiên cứu, các kiểm định giả thuyết thống kê được sử dụng nhằm đánh giá sự khác biệt và mối quan hệ giữa các nhóm dữ liệu trước khi đưa vào xây dựng mô hình học máy. Cụ thể, nhóm tiến hành kiểm định đối với một số biến danh mục để xác định xem các biến này có ảnh hưởng đáng kể đến giá bán bất động sản hay không. Kiểm định ANOVA được áp dụng để kiểm tra sự khác biệt về giá nhà giữa nhiều nhóm phân loại, chẳng hạn như tình trạng pháp lý của bất động sản (có giấy chứng nhận hoặc chỉ có hợp đồng mua bán) và mức độ hoàn thiện nội thất (hoàn thiện đầy đủ, cơ bản hoặc chưa hoàn thiện). Trong trường hợp so sánh giữa hai nhóm dữ liệu, kiểm định t-test được sử dụng, ví dụ để đánh giá sự khác biệt về giá bán giữa các bất động sản có sổ đỏ và không có sổ đỏ. Bên cạnh đó, kiểm định Chi-square được thực hiện nhằm xem xét mức độ phụ thuộc giữa các biến danh mục như thành phố, quận/huyện và các nhóm giá bất động sản. Kết quả của các kiểm định thống kê giúp xác định những biến có ý nghĩa thống kê thực sự, từ đó hỗ trợ quá trình lựa chọn đặc trưng phù hợp, loại bỏ các biến ít đóng góp và góp phần nâng cao hiệu quả dự đoán cũng như giảm độ phức tạp của mô hình học máy.

## 3.6. Triển khai xây dựng và đánh giá mô hình

## 3.6.1. Tiền xử lý dữ liệu cho modeling

Trước khi tiến hành xây dựng mô hình học máy, dữ liệu được tiền xử lý nhằm đảm bảo tính nhất quán và phù hợp cho quá trình huấn luyện mô hình. Cụ thể, tập dữ liệu được tách thành hai phần gồm tập biến đầu vào (X) chứa các đặc trưng của bất động sản và tập biến đầu ra (y) là giá bán (Price), đóng vai trò là biến mục tiêu trong bài toán dự đoán. Các biến danh mục như thành phố, quận/huyện, tình trạng pháp lý và mức độ hoàn thiện nội thất được mã hóa bằng phương pháp Label Encoding để chuyển đổi dữ liệu dạng phân loại sang dạng số, giúp mô hình có thể xử lý hiệu quả. Bên cạnh đó, các biến định lượng bao gồm diện tích, mặt tiền, độ rộng đường vào, số tầng, số phòng ngủ và số phòng tắm được chuẩn hóa bằng phương pháp StandardScaler nhằm đưa các đặc trưng về cùng một thang đo và tránh hiện tượng một số biến có giá trị lớn chi phối kết quả học của mô hình. Cuối cùng, dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80% và 20% nhằm đánh giá khả năng tổng quát hóa của mô hình trên dữ liệu chưa từng xuất hiện trong quá trình huấn luyện.

## 3.6.2. Tinh chỉnh siêu tham số

Random Forest : Các siêu tham số quan trọng như n\_estimators, max\_depth, min\_samples\_split và min\_samples\_leaf được tối ưu bằng phương pháp GridSearchCV kết hợp với cross-validation. Quy trình này giúp lựa chọn cấu hình mô hình tối ưu, giảm thiểu hiện tượng quá khớp và đảm bảo khả năng tổng quát hóa tốt trên dữ liệu mới.

XGBoost : Đối với mô hình XGBoost, các siêu tham số như n\_estimators, learning\_rate, max\_depth, subsample, colsample\_bytree, gamma, reg\_lambda và reg\_alpha được tinh chỉnh thông qua GridSearchCV. XGBoost sử dụng cơ chế boosting để mô hình hóa quan hệ phi tuyến giữa các đặc trưng đầu vào và giá nhà, đồng thời áp dụng regularization nhằm cải thiện hiệu suất dự đoán và hạn chế quá khớp.

## 3.6.3. Đánh giá và lựa chọn mô hình

Việc đánh giá hiệu quả dự đoán của các mô hình học máy được thực hiện trên tập dữ liệu kiểm tra nhằm đảm bảo tính khách quan và khả năng tổng quát hóa của mô hình. Trong nghiên cứu này, ba chỉ số đánh giá phổ biến gồm sai số tuyệt đối trung bình (MAE), căn bậc hai của sai số bình phương trung bình (RMSE) và hệ số xác định (R²) được sử dụng để so sánh hiệu suất giữa các mô hình. Chỉ số MAE phản ánh mức độ sai lệch trung bình giữa giá trị dự đoán và giá trị thực tế, qua đó thể hiện mức sai số thực tiễn của mô hình. Trong khi đó, RMSE đo lường mức độ sai lệch thông qua bình phương sai số, do đó nhạy cảm hơn với các dự đoán sai lệch lớn và phản ánh mức độ phạt cao đối với các điểm ngoại lai. Hệ số xác định R² cho biết tỷ lệ biến thiên của giá nhà được mô hình giải thích bởi các biến đầu vào, với giá trị càng gần 1 thể hiện mức độ phù hợp càng cao của mô hình. Dựa trên các chỉ số này, mô hình được lựa chọn là mô hình có giá trị MAE và RMSE thấp, đồng thời đạt hệ số R² cao, đảm bảo sai số dự đoán nhỏ và khả năng giải thích dữ liệu tốt. Mô hình có hiệu suất nổi trội theo các tiêu chí trên sẽ được sử dụng cho bài toán dự đoán giá nhà trong nghiên cứu.

## 3.6.4. Diễn giải mô hình và phân tích yếu tố ảnh hưởng

Để hiểu rõ hơn cơ chế dự đoán và xác định các yếu tố tác động mạnh nhất đến giá bất động sản, nghiên cứu tiến hành phân tích mô hình sau khi huấn luyện.

Để diễn giải mô hình XGBoost sau khi tối ưu, nghiên cứu sử dụng thước đo feature\_importances\_ do mô hình cung cấp. Đây là chỉ số phản ánh mức độ đóng góp của từng đặc trưng vào quá trình dự đoán, dựa trên mức giảm sai số khi đặc trưng đó được sử dụng trong các cây quyết định của mô hình.

Các giá trị này được tổng hợp thành bảng và trực quan hóa bằng biểu đồ bar chart, trong đó trục hoành thể hiện mức độ quan trọng và trục tung liệt kê các đặc trưng đầu vào. Biểu đồ được sắp xếp theo thứ tự giảm dần, giúp dễ dàng nhận diện những yếu tố có tác động mạnh nhất đến giá bất động sản.

Phân tích biểu đồ cho phép xác định các biến ảnh hưởng nổi bật như diện tích (Area), vị trí (City, District), tình trạng pháp lý (Legal status), hoặc các yếu tố cấu trúc như số phòng và số tầng. Nhờ vậy, mô hình không chỉ cho phép dự đoán giá mà còn hỗ trợ hiểu rõ cơ chế tác động của từng đặc trưng, từ đó cung cấp các hàm ý hữu ích cho người mua nhà, nhà đầu tư và doanh nghiệp bất động sản.

# CHƯƠNG 4. THẢO LUẬN KẾT QUẢ NGHIÊN CỨU

## 4.1. Phân tích khám phá dữ liệu

### 4.1.1 Phân tích đơn biến

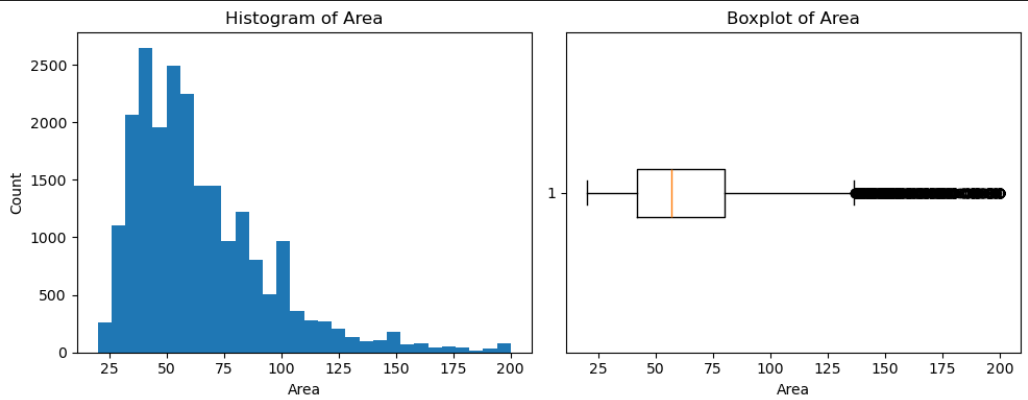
Mục tiêu và phương pháp (phân tích đơn biến)

Phân tích đơn biến nhằm mô tả phân bố của từng biến, phát hiện độ lệch (skewness), tính rời rạc, và ngoại lệ (outliers) để có quyết định tiền xử lý phù hợp trước khi phân tích đa biến/mô hình hóa.​

Nhóm sử dụng Histogram để quan sát dạng phân bố và Boxplot để nhận diện trung vị, IQR và các điểm ngoại lệ; các biểu đồ được tạo tự động cho từng biến trong danh sách

A. Area (Diện tích)

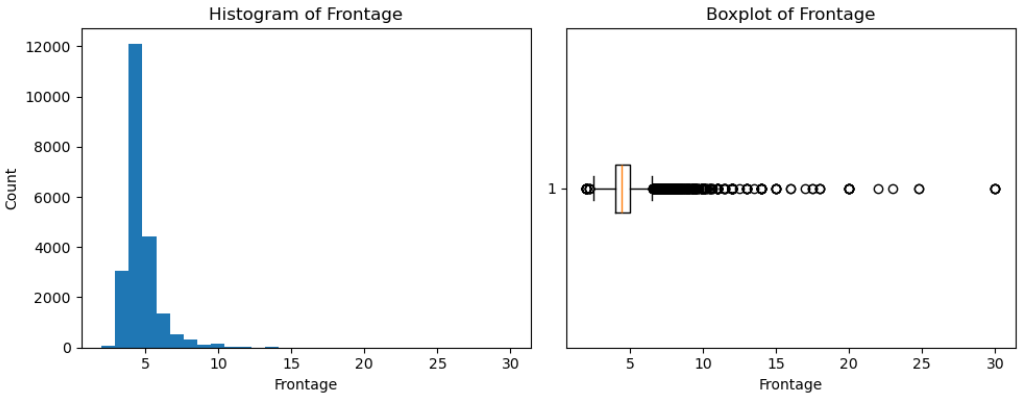
Biến Area thể hiện một phân bố lệch phải rất mạnh được thể hiện rõ qua biểu đồ Histogram với đỉnh tần suất tập trung cực đại trong khoảng 25 đến 50 đơn vị diện tích. Tần suất giảm rất nhanh sau mốc 100 đơn vị, chỉ ra rằng phần lớn thị trường là các căn nhà có diện tích nhỏ và vừa. Đồng thời, biểu đồ Boxplot xác nhận rằng mặc dù 50\% dữ liệu tập trung chặt chẽ (hộp tứ phân vị hẹp), sự hiện diện dày đặc của giá trị ngoại lai (outliers) kéo dài liên tục về phía giá trị lớn đã làm tăng phương sai tổng thể của biến một cách nghiêm trọng. Do đó, việc áp dụng Log Transformation là bắt buộc để giảm độ lệch và ảnh hưởng của các điểm dữ liệu cực đoan này.



#### Hình 4.1 : Biểu đồ Histogram và Boxplot của biến Area

B. Frontage (Mặt tiền)

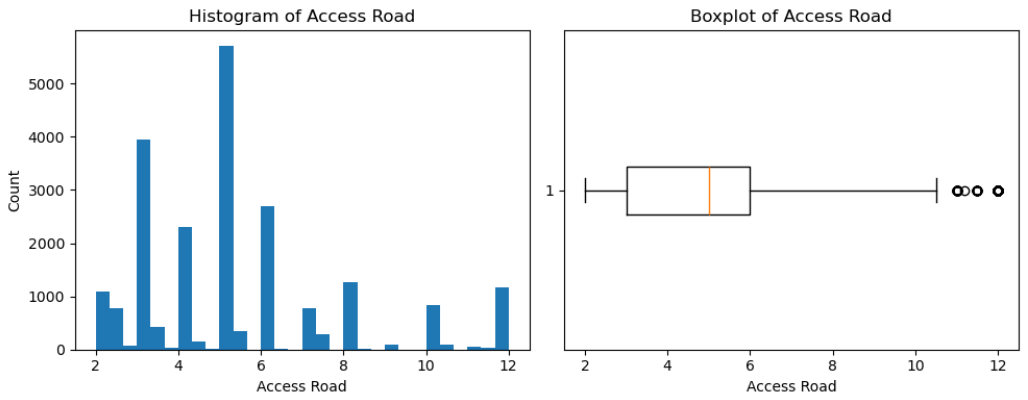
Phân tích biến Frontage cho thấy sự lệch phải cực đoan thậm chí còn nghiêm trọng hơn Area. Biểu đồ Histogram có đỉnh tần suất tập trung gần như toàn bộ tại các giá trị mặt tiền rất nhỏ, chủ yếu dưới 5 đơn vị. Biểu đồ Boxplot củng cố nhận định này khi hộp tứ phân vị (IQR) nằm ở mức thấp, nhưng chỉ ra sự tồn tại của các giá trị ngoại lai trải dài liên tục đến 30 đơn vị. Sự mất cân bằng dữ liệu nghiêm trọng này yêu cầu phải thực hiện Log Transformation để chuẩn hóa và giúp biến này đóng góp một cách tuyến tính và ổn định vào mô hình.



#### Hình 4.2: Biểu đồ Histogram và Boxplot của biến Frontage

C. Access Road

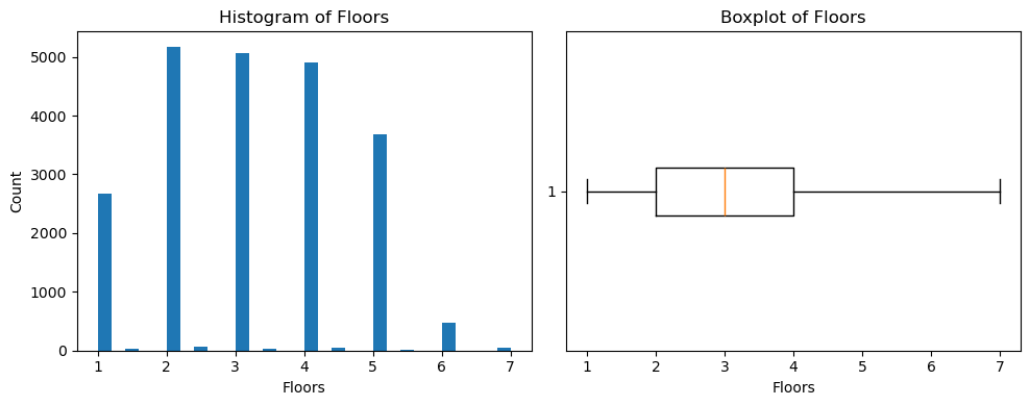
Biến Access Road có tính chất rời rạc và đa đỉnh rõ ràng. Biểu đồ Histogram không thể hiện phân bố liên tục mà tập trung tại các giá trị rời rạc cụ thể như 5, 6, 3 và 12. Điều này khẳng định đây là một biến định danh (Categorical) đã được mã hóa. Biểu đồ Boxplot cho thấy sự tập trung của dữ liệu, nhưng cũng chỉ ra các giá trị ngoại lai ở phía trên (khoảng 11 đến 12). Do đó, biến này phải được xử lý bằng One-Hot Encoding để tránh giả định một mối quan hệ tuyến tính không tồn tại giữa các loại đường.



#### Hình 4.3: Biểu đồ Histogram và Boxplot của biến Access Road.

D. Floors (số tầng)

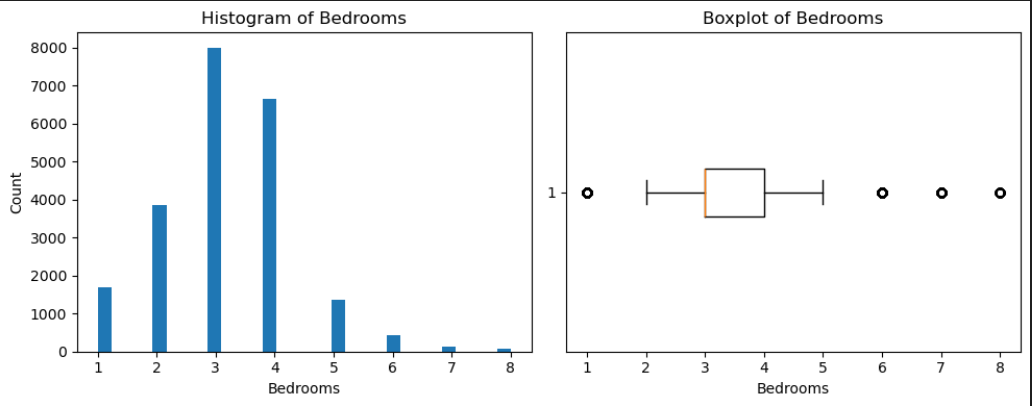
Histogram cho thấy số tầng là biến rời rạc, tập trung chủ yếu ở nhóm 2–5 tầng, trong đó các cột ở mức 2, 3, 4, 5 cao hơn rõ rệt so với các mức còn lại, còn 6–7 tầng xuất hiện ít. Boxplot cho thấy median khoảng 3 tầng, IQR nằm quanh vùng 2–4 tầng, và râu phải kéo dài tới khoảng 7 tầng → phân bố lệch phải nhẹ (một số nhà cao tầng hơn mức phổ biến), nhưng không thấy outlier tách biệt quá mạnh vì biến nhỏ và rời rạc.



#### Hình 4.4: Biểu đồ Histogram và Boxplot của biến Floors.

E. Bedrooms (số phòng ngủ)

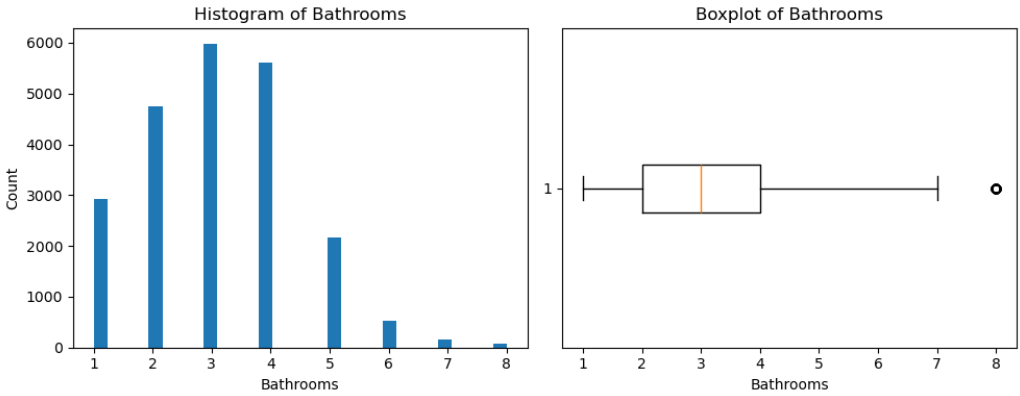
Histogram tập trung rõ ở 3–4 phòng ngủ (phổ biến nhất), tiếp theo là 2 và 5; các mức 6–8 phòng ngủ rất hiếm nên tạo đuôi phải. Boxplot thể hiện median khoảng 3–4, IQR hẹp (đa số nhà nằm trong vùng 3–4), đồng thời có các điểm rời phía phải (6–8) và một vài điểm thấp bất thường → cho thấy tồn tại một nhóm nhỏ nhà rất nhiều phòng ngủ (ngoại lệ) so với mặt bằng chung.



#### Hình 4.5: Biểu đồ Histogram và Boxplot của biến Bedrooms.

F. Bathrooms (số phòng tắm)

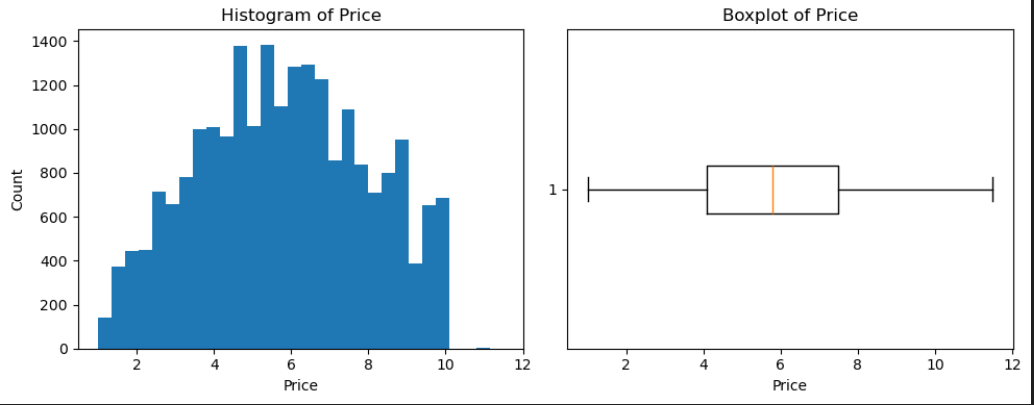
Histogram tập trung nhiều nhất ở 2–4 phòng tắm, đặc biệt các mức 3–4 nổi bật, trong khi 6–8 rất ít nên tạo đuôi phải. Boxplot cho thấy median khoảng 3, IQR nằm khoảng 2–4, râu phải kéo dài đến ~7 và có điểm ngoại lệ xa hơn → cho thấy một số nhà có số phòng tắm cao bất thường so với đa số, phù hợp với nhóm nhà lớn/nhà cao cấp



#### Hình 4.6: Biểu đồ Histogram và Boxplot của biến Bathrooms.

G. Price (giá)

Histogram của Price có dạng phân bố tương đối “phình” ở vùng giữa và thưa dần về phía phải, nghĩa là phần lớn bất động sản nằm trong khoảng giá trung bình, còn nhóm giá rất cao ít hơn nên tạo đuôi phải. Boxplot cho thấy median nằm gần vùng trung tâm của hộp, IQR tương đối rộng, râu phải dài hơn râu trái → giá có xu hướng lệch phải và có độ phân tán đáng kể; nhìn nhanh cũng thấy giá biến động mạnh hơn các biến rời rạc như số tầng/phòng.



#### Hình 4.7: Biểu đồ Histogram và Boxplot của biến Price.

KẾT LUẬN :

* Các biến cấu trúc (Floors/Bedrooms/Bathrooms/Access Road) phần lớn là rời rạc và có “cụm mức phổ biến”, trong khi Area và Price thể hiện đặc trưng lệch phải + đuôi dài; điều này phù hợp hành vi thị trường bất động sản (nhiều sản phẩm tầm trung, ít sản phẩm cao cấp).​
* Việc dùng histogram + boxplot giúp nhóm phát hiện ngoại lệ và ra quyết định xử lý dữ liệu theo hướng robust (clip/log/giữ ngoại lệ hợp lý) thay vì loại bỏ tùy tiện, từ đó làm nền cho phân tích đa biến và mô hình dự báo giá đáng tin cậy hơn.

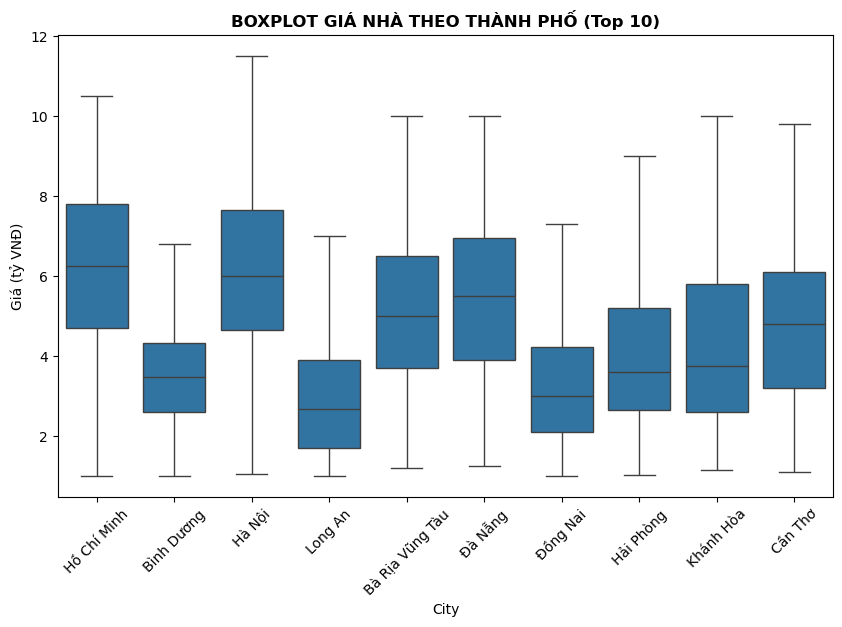
### 4.1.2 Phân tích đa biến

4.1.2.1. Phân tích theo vị trí.

Câu hỏi nghiên cứu :

Phân tích theo vị trí Giá nhà ở các thành phố Việt Nam hiện nay phân bố như thế nào giữa các khu vực, và sự khác biệt này phản ánh điều gì về mức độ đô thị hóa, nhu cầu nhà ở và chênh lệch phát triển giữa các thành phố?

A. Phân tích Phân bố Giá theo Khu vực



#### Hình 4.8: Biểu đồ Boxplot phân bố giá nhà tại các thành phố trọng điểm (Top 10)

* Sự thống trị của các trung tâm kinh tế: Biểu đồ cho thấy Hồ Chí Minh và Hà Nội có mức giá trung vị (median) cao nhất (khoảng 6 tỷ VNĐ) và dải giá rộng nhất. Điều này minh chứng cho mức độ đô thị hóa sâu, nơi thị trường tồn tại đầy đủ các phân khúc từ nhà giá rẻ đến bất động sản cao cấp.
* Đặc thù vùng miền và khoảng cách phát triển: Miền Nam có sự phân hóa rõ rệt nhất với TP.HCM ở đỉnh cao và Long An ở mức thấp hơn (trung vị khoảng 2.8 tỷ VNĐ). Sự chênh lệch vị trí hộp giữa các vùng phản ánh sự phát triển không đồng đều về hạ tầng và dịch vụ giữa lõi đô thị và vùng vệ tinh.
* Sự trỗi dậy của các đô thị vệ tinh: Các thành phố như Bình Dương và Đồng Nai có dải giá tập trung ở phần thấp hơn, phản ánh quá trình đô thị hóa đang mở rộng để giải quyết nhu cầu nhà ở cho lực lượng lao động tại các khu công nghiệp.

B. Phân tích Quy mô và Nhu cầu Thị trường

A graph with blue bars

AI-generated content may be incorrect.

#### Hình 4.9: Biểu đồ cột thể hiện số lượng tin đăng (quy mô giao dịch) theo từng Thành phố

* Sức hút của "Siêu đô thị": Số lượng mẫu tại Hồ Chí Minh áp đảo hoàn toàn các khu vực khác, theo sau là Hà Nội. Nhu cầu nhà ở thực tại Việt Nam hiện nay vẫn đang xoay quanh các trung tâm tài chính lớn, nơi cơ hội việc làm và hạ tầng phát triển nhất.
* Bất cân xứng nguồn cung và dịch chuyển nhu cầu: Việc nhu cầu tập trung quá mức vào hai đầu cầu gây áp lực tăng giá liên tục, buộc người dân phải dịch chuyển sang các vùng giá "dễ thở" hơn (thường tập trung ở mức 2.8 - 4 tỷ VNĐ) như Long An hay Bình Dương để tìm kiếm cơ hội an cư.

Kết luận:

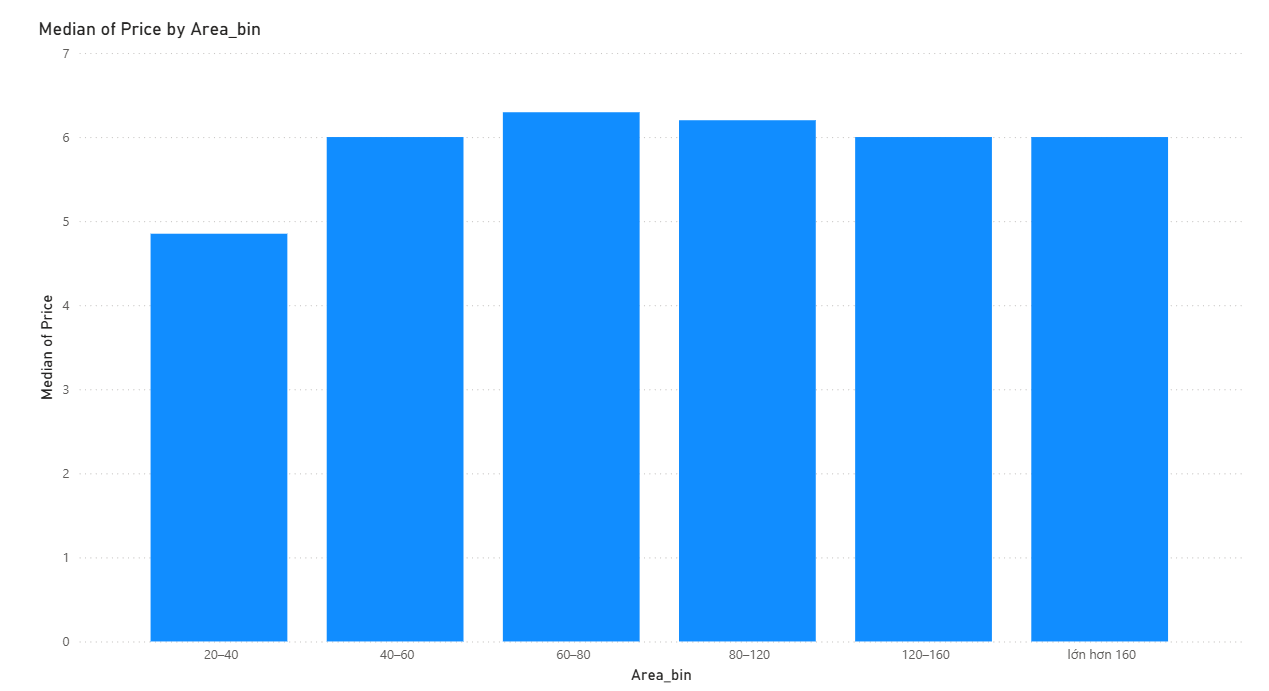
* "Từ số liệu thực tế, chúng ta thấy thị trường nhà đất Việt Nam hiện nay có sự phân hóa rất rõ rệt: Hà Nội và TP.HCM vẫn là 'đầu tàu' giữ vị trí số 1 cả về mức giá đắt đỏ lẫn số lượng người quan tâm mua bán. Tại hai thành phố này, giá nhà không chỉ cao mà còn có đủ mọi phân khúc, cho thấy khi đất chật người đông và hạ tầng hiện đại, giá trị nhà đất sẽ luôn ở mức rất cao.
* Tuy nhiên, một xu hướng rất thực tế đang diễn ra là người dân bắt đầu chuyển hướng sang các khu vực lân cận như Bình Dương, Đồng Nai hay Long An. Đây là giải pháp tất yếu để cân bằng giữa ngân sách và nhu cầu sở hữu nhà trong bối cảnh giá cả tại các đô thị lõi liên tục thiết lập mặt bằng mới".

4.1.2.2. Phân tích theo đặc trưng vật lý

Câu hỏi nghiên cứu

Phân tích theo đặc điểm vật lý Giá nhà thay đổi như thế nào theo các đặc điểm vật lý của bất động sản (diện tích, mặt tiền, số tầng, số phòng ngủ, số phòng tắm), và các đặc điểm nào phản ánh rõ nhất sự khác biệt giữa các phân khúc nhà ở?

A. Ảnh hưởng của diện tích (Area) đến giá



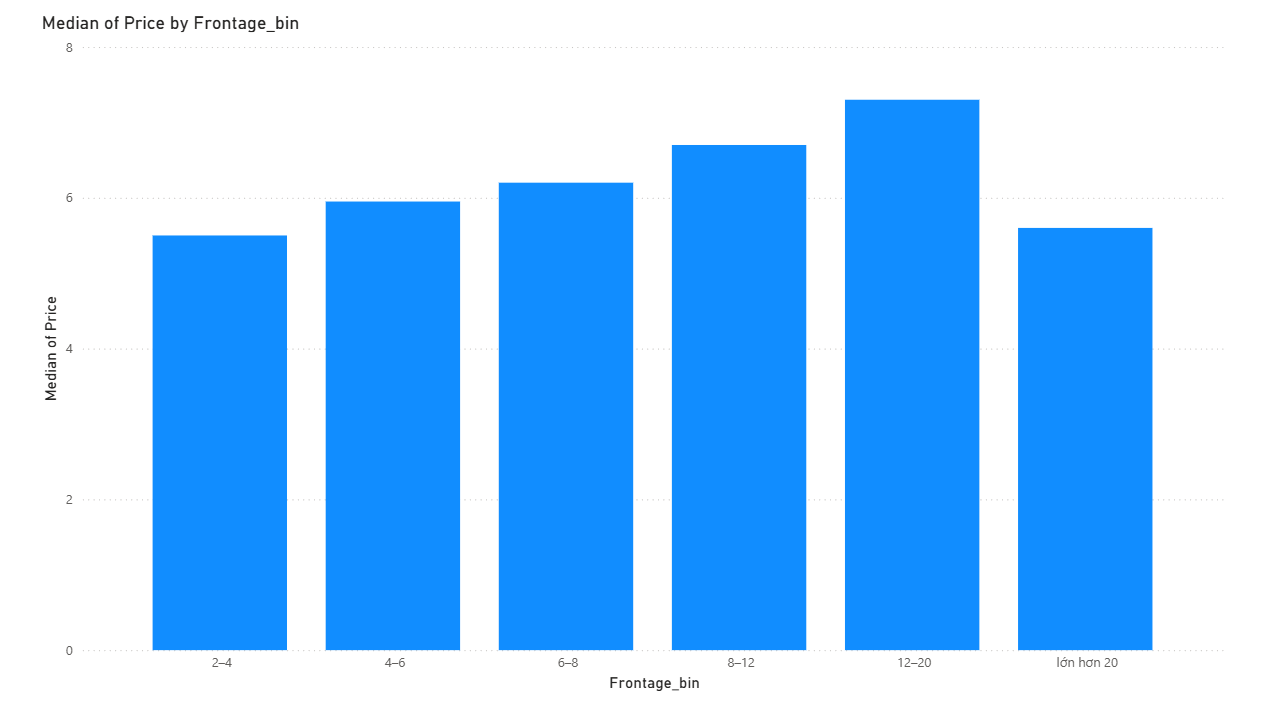
#### Hình 4.10: Giá nhà trung vị theo nhóm diện tích (Area\_bin)

Nhóm 20–40m² có median giá thấp nhất; lên 40–60m² thì median tăng rõ.​ Median đạt mức cao nhất quanh 60–80m², sau đó các nhóm 80–120, 120 – 160 và lớn hơn160 không tăng thêm nhiều (thậm chí giảm nhẹ so với đỉnh).​

Cách hiểu hợp lý: khi diện tích vượt ngưỡng “đủ dùng”, giá phụ thuộc nhiều hơn vào vị trí (City/District), mặt tiền, đường trước nhà và mức hoàn thiện/công năng (tầng, phòng ngủ, phòng tắm) thay vì chỉ tăng thêm diện tích.

Giá tăng mạnh khi chuyển từ phân khúc nhỏ sang trung bình, và bão hòa ở nhóm diện tích lớn. Diện tích một mình không giải thích tốt chênh lệch giá trên toàn bộ dữ liệu vì còn bị chi phối bởi vị trí và các yếu tố công năng khác.

B. Ảnh hưởng của mặt tiền (Frontage) đến giá

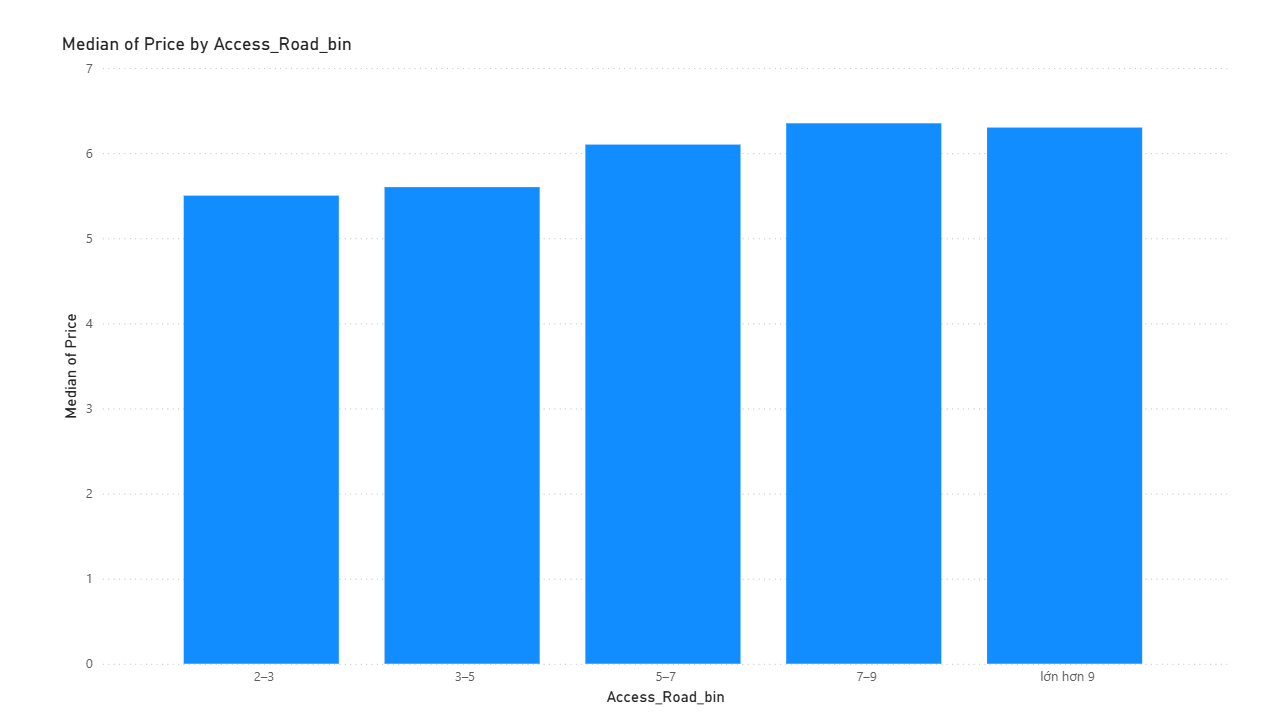


#### Hình 4.11: Giá nhà trung vị theo nhóm mặt tiền (Frontage\_bin)

Từ 2–4m → 4–6m → 6–8m → 8–12m → 12–20m: median giá tăng đều qua từng bậc. Điều này hợp lý vì mặt tiền rộng thường tăng giá trị sử dụng (thoáng, dễ bố trí công năng, phù hợp kinh doanh/để xe, tính “đẹp” của nhà phố).​

Nhóm “lớn hơn 20m” giảm xuống so với 12–20m. Mặt tiền có tác động dương đến giá trong phần lớn các nhóm (đặc biệt từ 2–4m lên 12–20m), nhưng hiệu ứng bão hòa/đảo chiều ở nhóm >20m cho thấy mặt tiền không phải yếu tố quyết định.

C. Ảnh hưởng của đường trước nhà (Access Road) đến giá



#### Hình 4.12: Giá nhà trung vị theo nhóm bề rộng đường trước nhà (Access\_Road\_bin)

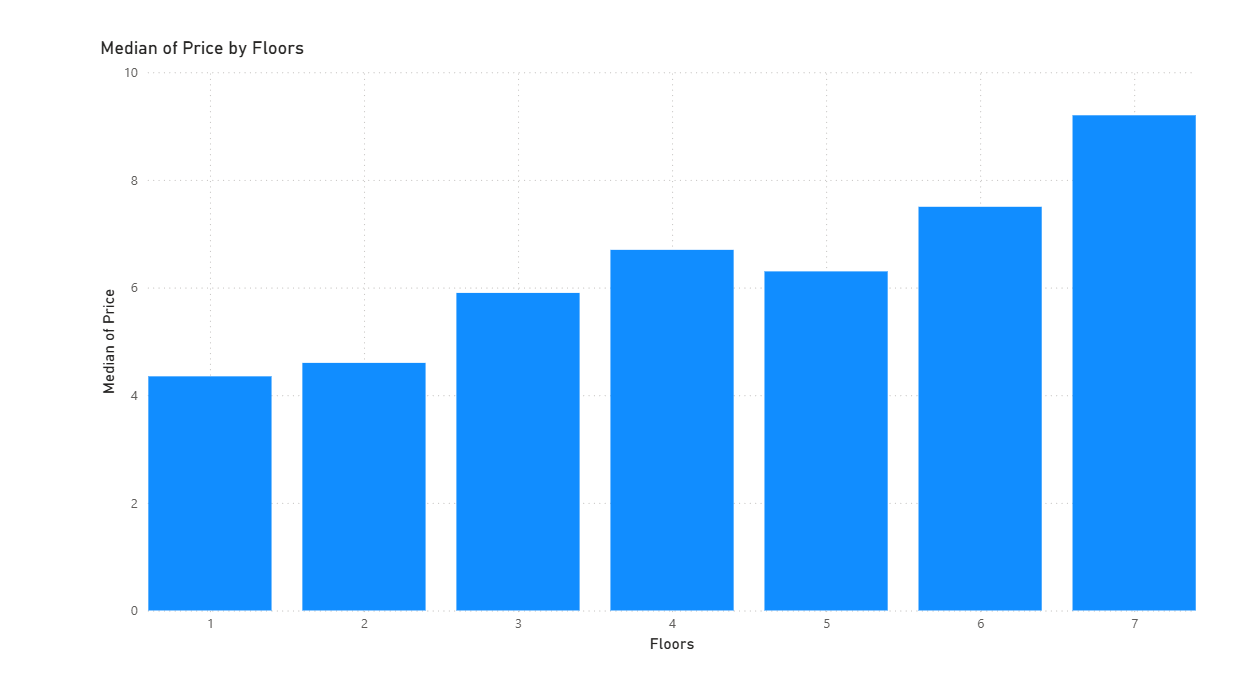
Nhóm 2–3m và 3–5m có median giá xấp xỉ nhau (tăng nhẹ), cho thấy khi đường còn hẹp, chênh lệch 1–2m chưa tạo khác biệt lớn về giá trị.​

Bước tăng rõ nằm ở nhóm 5–7m và 7–9m: median giá tăng lên mức cao hơn, phản ánh lợi ích thực tế của đường rộng (xe ô tô ra vào thuận tiện, khả năng kinh doanh/mặt bằng tốt hơn, tính thanh khoản cao hơn).​

Nhóm “lớn hơn 9m” không tăng thêm nhiều so với 7–9m, gợi ý hiệu ứng “bão hòa”: vượt một ngưỡng tiện ích, giá không còn tăng mạnh

Bề rộng đường trước nhà có quan hệ dương với giá: đường càng rộng thì median giá càng cao, đặc biệt khi vượt ngưỡng khoảng 5m (từ hẻm nhỏ sang đường ô tô).

D. Ảnh hưởng của số tầng (Floor) đến giá



#### Hình 4.13: Giá nhà trung vị theo số tầng (Floors)

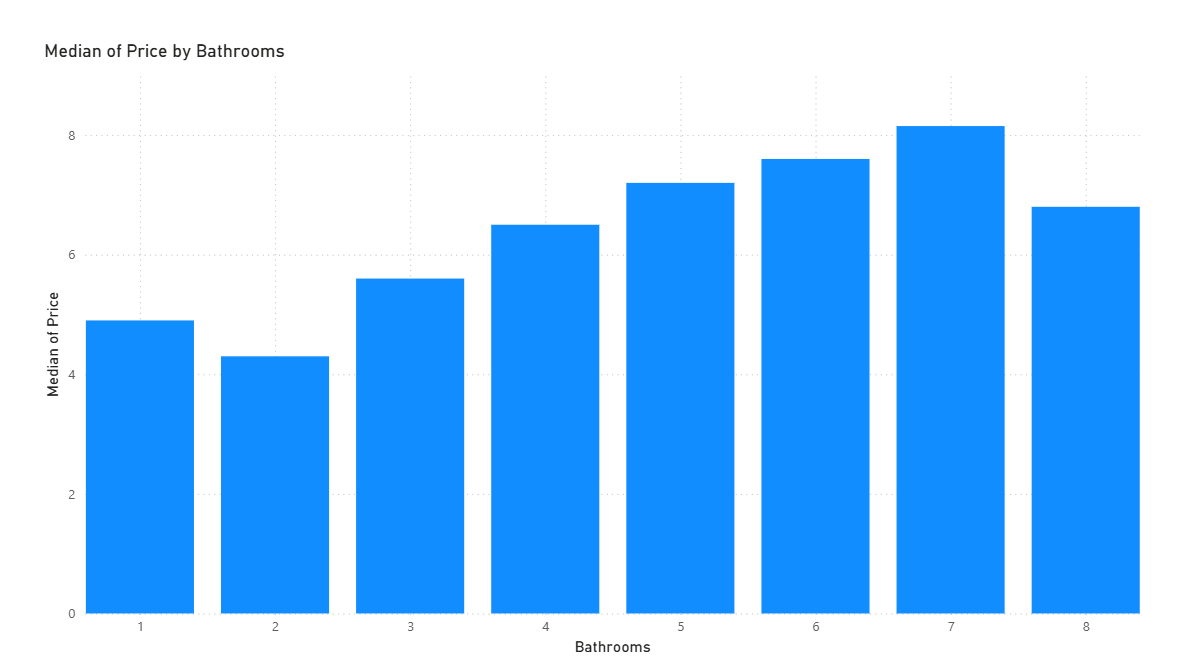
Nhóm 1 tầng có median giá khoảng 4.35, lên 2 tầng khoảng 4.60 (tăng nhẹ).

Ở 3 tầng: median lên khoảng 5.90 (tăng ~1.125 so với 2 tầng), và 4 tầng lên khoảng 6.70 (tăng thêm ~0.10 so với 3.5–4 tuỳ nhóm).

Ở 5 tầng, median giảm nhẹ so với 4 tầng (khoảng 6.30 so với 6.70), rồi tăng lên 6 tầng khoảng 7.50 và 7 tầng khoảng 9.20.

Số tầng phản ánh quy mô xây dựng và khả năng khai thác sử dụng (nhiều không gian, có thể chia phòng/cho thuê/kinh doanh), vì vậy giá có xu hướng tăng theo số tầng.

E. Ảnh hưởng của số phòng tắm (Bathrooms) đến giá



#### Hình 4.14: Giá nhà trung vị theo số phòng tắm (Bathrooms)

Nhóm 1–2 phòng tắm: median giá lần lượt ~4.90 và ~4.30 ; mức giảm ở 2 phòng tắm có thể do nhóm này tập trung ở phân khúc nhà nhỏ/ít tiện nghi hoặc cấu trúc mẫu.

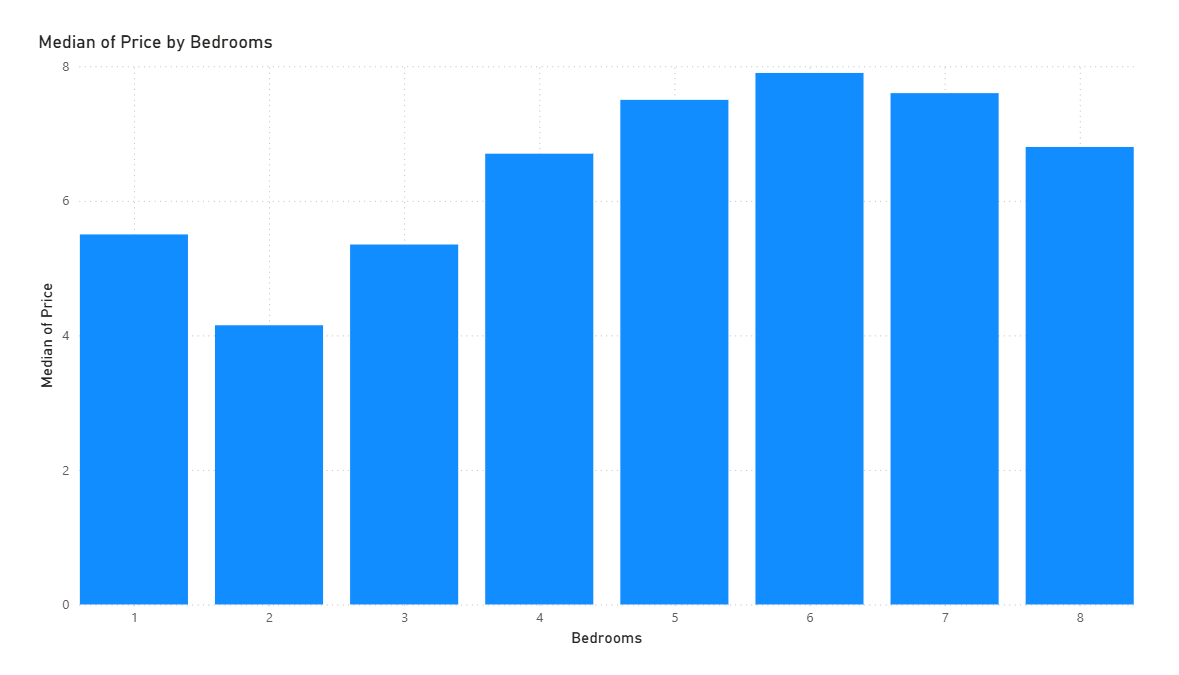
Từ 3 lên 4 phòng tắm: median tăng mạnh từ ~5.60 lên ~6.50 , cho thấy khi nhà đạt mức “đủ tiện nghi cho gia đình lớn/nhà nhiều tầng”, giá nhảy đáng kể.

Từ 5 lên 7 phòng tắm: median tiếp tục tăng từ ~7.20 → ~7.60 → ~8.15 , phản ánh nhóm nhà nhiều phòng tắm thường là nhà quy mô lớn hơn, khai thác tốt hơn (nhiều phòng ngủ/tầng, phù hợp ở kết hợp cho thuê/kinh doanh).

8 phòng tắm: median giảm còn ~6.80 vì cỡ mẫu nhỏ, đây dễ là hiệu ứng nhiễu (ít quan sát) hoặc “mix” loại hình đặc thù, xu hướng tổng thể tăng đến 7 phòng tắm và cảnh báo ngoại lệ ở mức rất cao.

Bathrooms có xu hướng làm tăng giá nhà vì phản ánh mức độ tiện nghi và quy mô sử dụng; trong dữ liệu, median giá tăng rõ nhất từ nhóm 3–4 phòng tắm trở lên và đạt cao nhất quanh 7 phòng tắm.

F. Ảnh hưởng của số phòng ngủ (Bedrooms) đến giá



#### Hình 4.15: Giá nhà trung vị theo số phòng ngủ (Bedrooms)

Nhóm 1–2 phòng ngủ: median lần lượt ~5.50 và ~4.15; việc 2 phòng ngủ thấp hơn 1 phòng ngủ cho thấy nhóm “1 phòng ngủ” trong dữ liệu có thể không phải nhà nhỏ (có thể là nhà phố/nhà đặc thù ghi 1 phòng ngủ nhưng giá cao do vị trí), hoặc do sai khác cấu trúc mẫu.

Từ 3 lên 4 phòng ngủ: median tăng mạnh từ ~5.35 lên ~6.70, đây là bước nhảy lớn nhất (tăng ~1.35) và phản ánh nhu cầu thị trường cho nhà gia đình nhiều phòng.

5–6 phòng ngủ: median tiếp tục tăng lên ~7.50 và ~7.90 , thể hiện giá trị công năng tăng thêm (phù hợp gia đình lớn/khai thác cho thuê).

7–8 phòng ngủ: median giảm nhẹ ~7.60 với ~6.80 ; do cỡ mẫu nhỏ và có thể chuyển sang phân khúc/địa bàn khác, nên nên mô tả là “xu hướng bão hòa/dao động” thay vì kết luận giảm chắc chắn.

Cho thấy số phòng ngủ là biến phản ánh quy mô sử dụng và mức tiện nghi, góp phần phân biệt các phân khúc nhà ở. Trong dữ liệu, median giá tăng mạnh nhất khi chuyển từ 3 → 4 phòng ngủ và tiếp tục tăng đến khoảng 6 phòng ngủ; ở mức 7–8 phòng ngủ, giá dao động do ít quan sát và khả năng nhiễu theo loại hình/vị trí.

G. Phân tích tương quan

A screenshot of a graph

AI-generated content may be incorrect.

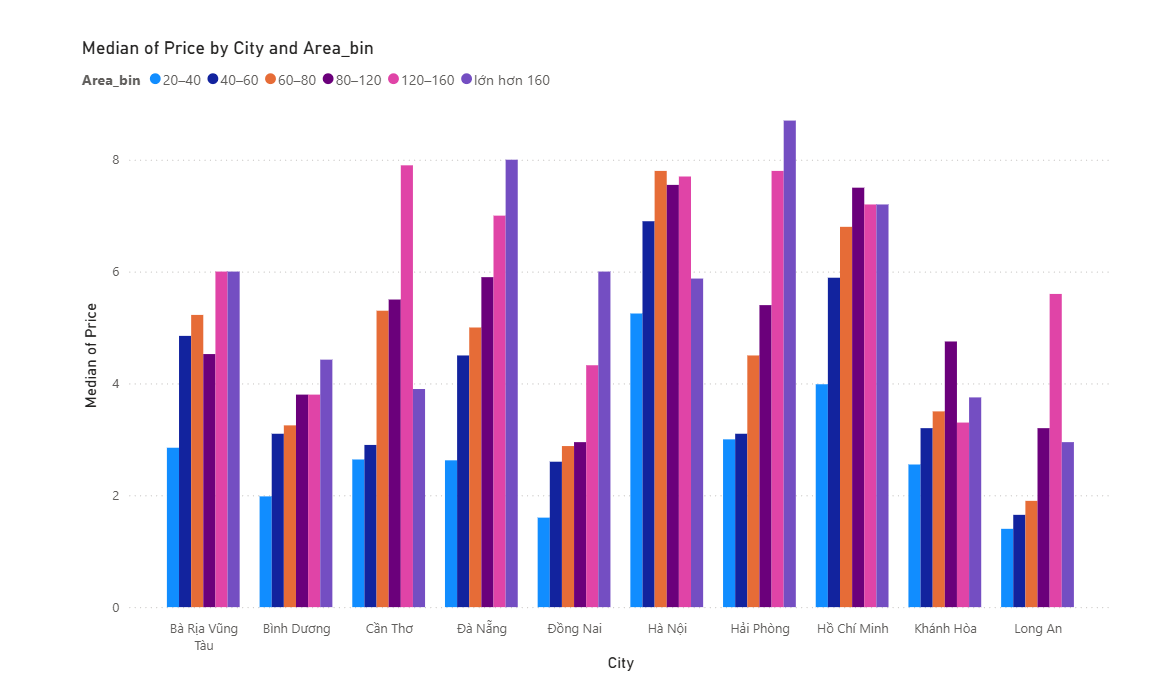
#### Hình 4.16: Heatmap tương quan giữa đặc điểm vật lý và giá nhà.

Heatmap cho thấy giá nhà (Price) tương quan trung bình với nhóm biến công năng: Bathrooms (~0.36), Floors (~0.34), Bedrooms (~0.33). Ba biến này tăng thì giá có xu hướng tăng tương đối rõ, phù hợp với câu chuyện “nhà nhiều tầng/nhiều phòng → tiện nghi hơn, khai thác được tốt hơn → giá cao hơn”.​

Ngược lại, các biến “kích thước/tiếp cận” có tương quan yếu với giá: Area (~0.12), Access Road (~0.11), Frontage (~0.07). Điều này không có nghĩa chúng không quan trọng, mà là quan hệ không tuyến tính và bị “nhiễu” bởi vị trí (City/District) và loại hình BĐS; ví dụ diện tích/mặt tiền tăng đến ngưỡng thì giá bão hòa hoặc đảo chiều ở nhóm rất lớn.​

Heatmap cũng cho thấy các biến công năng đồng biến mạnh với nhau: Bedrooms–Bathrooms rất cao (~0.75), Bathrooms–Floors(~0.56), Bedrooms–Floors (~0.43). Vì chúng đi theo “gói” (nhà nhiều tầng thường nhiều phòng ngủ và phòng tắm), khi diễn giải cần nhấn mạnh đây là mối liên hệ tổng hợp về quy mô/tiện nghi chứ không phải từng biến tác động độc lập hoàn toàn.

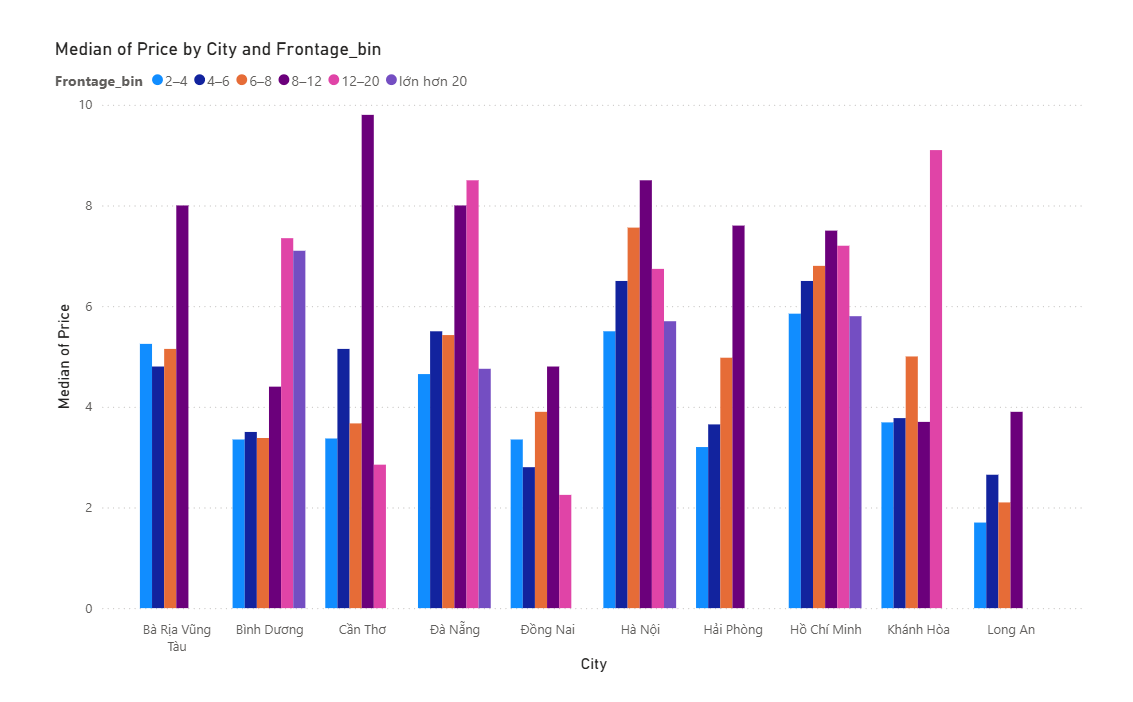
H. Phân tích theo City: Area/Frontage/Access Road × City × Price



#### Hình 4.17: Giá nhà trung vị theo City và nhóm diện tích (Area\_bin).

Hà Nội và TP.HCM có mặt bằng median giá cao ở hầu hết các nhóm diện tích, và thường tăng rõ khi lên các bin diện tích lớn (đặc biệt ở nhóm 60–160m²), cho thấy tại đô thị lớn, diện tích vẫn “có giá”  
 Bình Dương/Đồng Nai/Long An có median giá thấp hơn rõ rệt so với Hà Nội/TP.HCM ở cùng bin diện tích: cùng 60–80m² nhưng giá giữa các tỉnh có thể chênh lớn

Phân tích theo City cho thấy diện tích có xu hướng làm giá tăng, nhưng tác động phụ thuộc mạnh vào City: cùng một mức diện tích, median giá giữa các City khác nhau đáng kể.



#### Hình 4.18: Giá nhà trung vị theo City và nhóm mặt tiền (Frontage\_bin)

Ở nhiều City, median giá thường tăng khi đi từ nhóm 2–4m lên các nhóm 6–8m và 8–12m/12–20m, phản ánh lợi thế mặt tiền rộng (thoáng, dễ bố trí công năng, phù hợp kinh doanh/để xe).​

Tuy nhiên, các cột ở nhóm 12–20m và “>20m” dao động mạnh giữa các City (có nơi tăng vọt, có nơi giảm), cho thấy nhóm mặt tiền rất rộng dễ “lẫn” loại hình BĐS khác nhau và chịu ảnh hưởng lớn của vị trí/khu dân cư

Mặt tiền có xu hướng tác động dương đến giá, nhưng tác động này không ổn định giữa các thành phố và dễ bão hòa/dao động ở nhóm mặt tiền rất lớn; do đó tương quan tuyến tính tổng thể thấp. Khi tách theo City, mối liên hệ mặt tiền–giá rõ hơn ở một số thị trường,



#### Hình 4.19: Giá nhà trung vị theo City và nhóm bề rộng đường trước nhà (Access\_Road\_bin)

Với các đô thị lớn như Hà Nội và TP.HCM, các nhóm đường rộng (5–7m, 7–9m, >9m) thường có median giá cao hơn nhóm đường hẹp, phản ánh lợi thế ô tô ra vào, kinh doanh, và tính thanh khoản.​

Ở các tỉnh như Bình Dương/Đồng Nai/Long An, đường rộng cũng làm giá tăng nhưng mặt bằng giá chung thấp hơn; vì vậy cùng một bin đường (ví dụ 5–7m) vẫn có giá khác nhau lớn giữa City.

Bề rộng đường trước nhà có xu hướng làm tăng giá do cải thiện khả năng tiếp cận và khai thác (ô tô ra vào, kinh doanh), thể hiện rõ khi so sánh theo từng City và theo các bậc 2–3m → 5–7m → 7–9m/>9m. Tuy nhiên, vì khác biệt mặt bằng giá giữa các City rất lớn và phân phối mẫu theo từng bin không đồng đều nê tương quan tuyến tính tổng thể chỉ ở mức thấp

KẾT LUẬN:

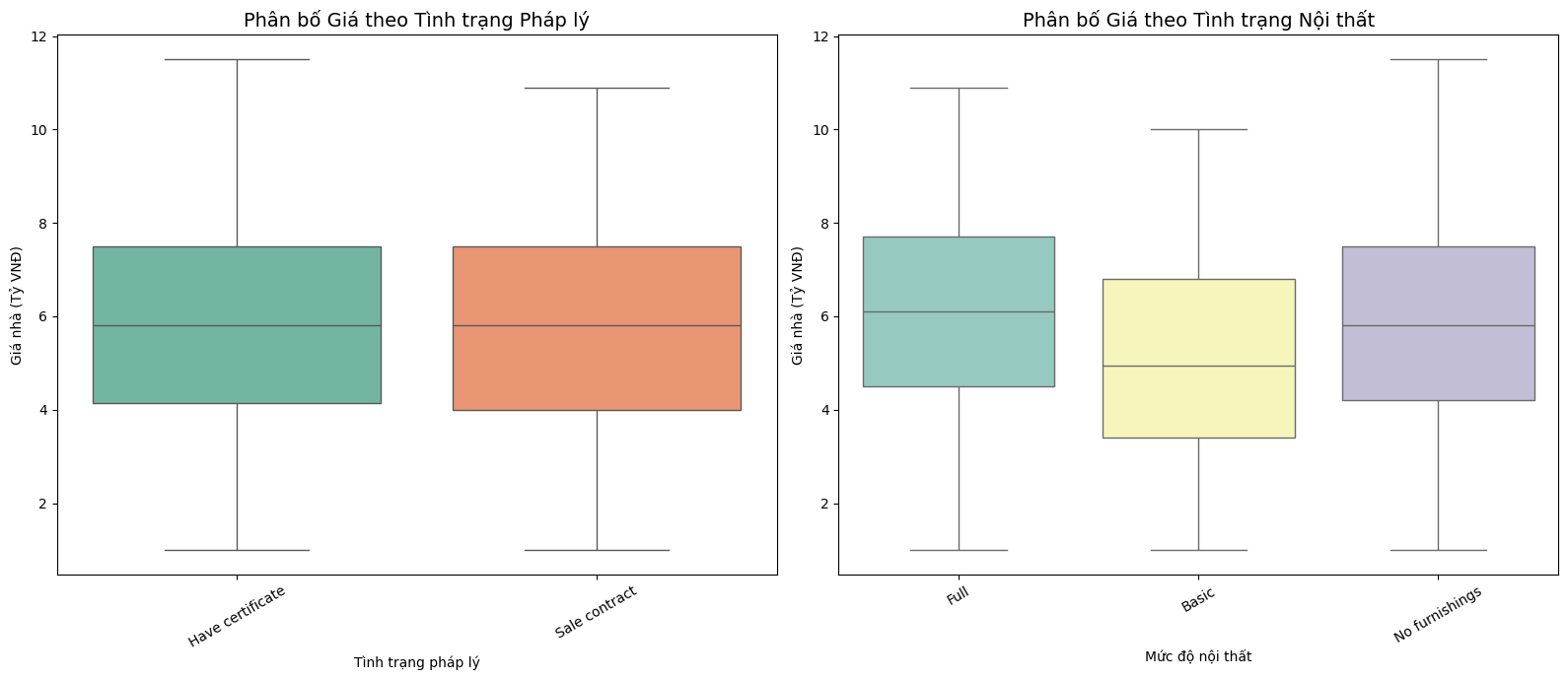
* Giá nhà trong bộ dữ liệu chịu ảnh hưởng rõ nhất bởi các đặc điểm “công năng” của bất động sản, cụ thể số phòng tắm, số tầng và số phòng ngủ; các biến này có xu hướng đồng biến với giá và phản ánh trực tiếp mức tiện nghi cũng như khả năng khai thác sử dụng của căn nhà. Ngược lại, diện tích, mặt tiền và bề rộng đường trước nhà thể hiện mối liên hệ dương nhưng yếu hơn khi xét toàn bộ mẫu; tác động của nhóm biến này thường không tuyến tính (tăng đến ngưỡng rồi bão hòa/dao động) và dễ bị chi phối bởi khác biệt thị trường theo khu vực.
* Khi phân chia nhà theo quy mô (nhỏ–trung bình–lớn) dựa trên diện tích, sự khác biệt giữa các phân khúc thể hiện rõ nhất ở diện tích; đồng thời các đặc điểm đi kèm như mặt tiền và đường trước nhà có xu hướng tăng theo phân khúc, còn số tầng giảm dần (nhà nhỏ thường xây cao tầng hơn), phản ánh khác biệt về loại hình và cách khai thác sử dụng giữa các nhóm. Vì vậy, có thể kết luận rằng nhóm biến công năng là nhóm giải thích sự biến động giá tốt nhất, còn diện tích và các đặc điểm tiếp cận/cấu trúc là những dấu hiệu hữu ích để nhận diện và mô tả sự khác nhau giữa các phân khúc nhà ở.

4.1.2.3. Phân tích theo pháp lý và nội thất

Câu hỏi nghiên cứu :

Tình trạng pháp lý (Legal status) và mức độ hoàn thiện nội thất (Furniture) ảnh hưởng như thế nào đến giá nhà, và mức độ ảnh hưởng này có khác nhau giữa các khu vực và phân khúc nhà ở hay không?

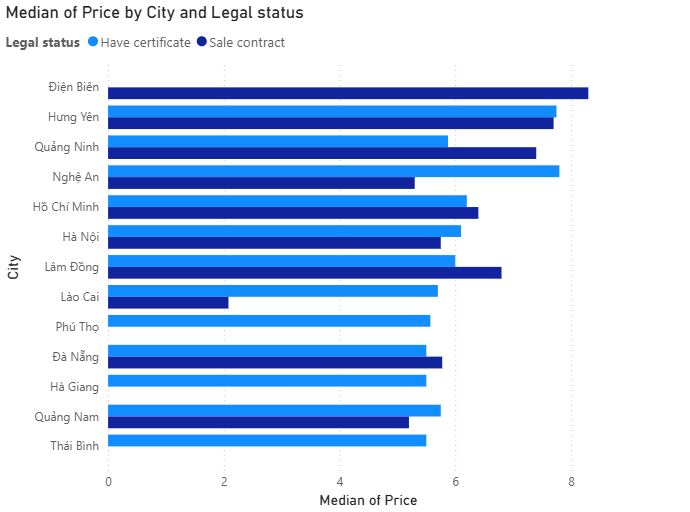
A. Ảnh hưởng của pháp lý và nội thất đến giá nhà.



#### Hình 4.20: Biểu đồ Boxplot phân bổ giá nhà theo Tình trạng Pháp lý và Mức độ Nội thất

* Tình trạng pháp lý – “Bảo chứng” cho phân khúc giá cao: Biểu đồ Boxplot bên trái cho thấy nhóm “Have certificate" (Đã có sổ) xuất hiện ở vùng giá rất cao, đạt ngưỡng trên 11 tỷ VNĐ. Điều này hàm ý các tài sản giá trị lớn thường đi kèm pháp lý rõ ràng để đảm bảo an toàn tuyệt đối cho nhà đầu tư. Tuy nhiên, giá trung vị của nhóm có sổ và “Sale contract" khá gần nhau (quanh mức 6 tỷ VNĐ), cho thấy ở phân khúc phổ thông người mua vẫn có thể chấp nhận pháp lý chưa hoàn thiện nếu giá phù hợp.
* Mức độ nội thất – "Động lực" tăng trưởng giá trị: Quan sát biểu đồ bên phải, nhóm "Full" (Đầy đủ nội thất) có mức giá trung vị cao nhất (vượt mốc 6 tỷ VNĐ), đóng vai trò là đòn bẩy giúp căn nhà thoát khỏi phân khúc bình dân để tiến vào dải giá cao hơn. Trong khi đó, nhóm "Basic" có hộp biểu đồ hẹp và thấp nhất, phản ánh phân khúc dành cho người mua muốn tối ưu chi phí để tự hoàn thiện.

B. Phân tích sự khác biệt của pháp lý theo Khu vực (City)



#### Hình 4.21: Biểu đồ so sánh Giá trung vị theo Thành phố dựa trên Pháp lý và Nội thất

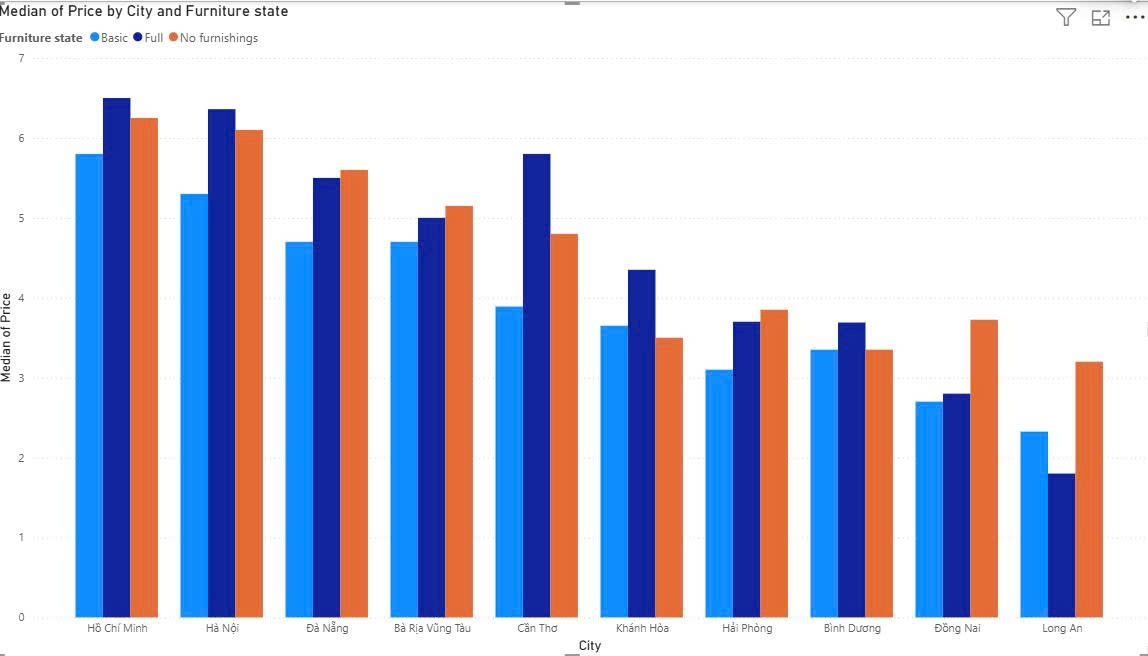
Kết quả cho thấy, tại hầu hết các thành phố, nhà có pháp lý đầy đủ có mức giá trung vị cao hơn so với nhà chỉ có hợp đồng mua bán. Điều này phản ánh vai trò quan trọng của yếu tố pháp lý trong việc hình thành giá nhà, do pháp lý đầy đủ giúp giảm thiểu rủi ro giao dịch và nâng cao khả năng chuyển nhượng trong tương lai.

Tuy nhiên, mức độ chênh lệch giá giữa hai trạng thái pháp lý không đồng đều giữa các khu vực. Tại các thành phố lớn như Hồ Chí Minh, Hà Nội và Đà Nẵng, sự khác biệt về giá giữa hai nhóm pháp lý tương đối nhỏ. Điều này cho thấy tại các thị trường phát triển, người mua có xu hướng chấp nhận rủi ro pháp lý cao hơn, do nhu cầu nhà ở lớn và kỳ vọng khả năng sinh lời trong dài hạn.

Ngược lại, tại các tỉnh/thành có quy mô thị trường nhỏ hơn như Lào Cai, Phú Thọ, Nghệ An hoặc Quảng Nam, giá nhà có hợp đồng mua bán thấp hơn rõ rệt so với nhà có giấy chứng nhận. Kết quả này cho thấy người mua tại các khu vực này nhạy cảm hơn với rủi ro pháp lý, và pháp lý trở thành yếu tố mang tính quyết định trong quá trình định giá bất động sản.

Tổng hợp lại, có thể nhận định rằng ảnh hưởng của pháp lý đến giá nhà phụ thuộc mạnh vào đặc thù từng khu vực. Pháp lý đóng vai trò quan trọng hơn tại các thị trường nhỏ và kém phát triển, trong khi tại các đô thị lớn, yếu tố này có xu hướng tương tác với các đặc điểm khác như vị trí và đặc trưng vật lý của bất động sản.

C. Phân tích sự khác biệt của nội thất theo Khu vực (City)



#### Hình 4.22: Biểu đồ Giá trung vị theo Thành phố và Mức độ nội thất

* Đòn bẩy giá tại trung tâm: Biểu đồ cột nhóm minh chứng tại các thị trường đô thị hóa cao như Hà Nội, TP.HCM và Cần Thơ, nhóm nhà "Full nội thất" (cột màu xanh đậm) luôn thiết lập mức giá trung vị cao vượt trội so với các nhóm còn lại.
* Xu hướng tiêu dùng: Điều này phản- ánh nhu cầu về sự tiện nghi "chìa khóa trao tay" và tiêu chuẩn sống cao của người dân tại các thành phố lớn.
* Sự khác biệt tại vùng vệ tinh: Ở các khu vực như Bình Dương hay Long An, khoảng cách giữa các loại hình nội thất không quá lớn, thậm chí nhóm nhà trống (No furnishings) vẫn giữ mức giá ổn định, cho thấy người mua tại đây ưu tiên giá trị thực của đất và diện tích hơn là phần hoàn thiện nội thất.

KẾT LUẬN:

“Số liệu thực tế cho thấy: Pháp lý (Legal status – tình trạng pháp lý) là điều kiện cần giúp tăng độ tin cậy và thường gắn với các tài sản ở vùng giá rất cao, còn Nội thất (Furniture state – mức hoàn thiện nội thất) là ‘đòn bẩy’ giúp nâng mặt bằng giá và tăng sức cạnh tranh. Mức ảnh hưởng không cố định mà thay đổi theo khu vực: có nơi pháp lý tạo chênh lệch rất rõ, có nơi chênh lệch nhỏ hơn do nhiều yếu tố khác như vị trí cùng lúc chi phối giá.”

## 4.2. Xây dựng mô hình phân cụm và dự đoán giá nhà

### 4.2.1 Phương pháp phân cụm (K-means)

A. Mục tiêu phân cụm

Mục tiêu của phân tích phân cụm là nhằm nhận diện các phân khúc nhà ở khác nhau dựa trên các đặc điểm vật lý của bất động sản, từ đó làm rõ cấu trúc thị trường nhà ở và hỗ trợ việc diễn giải kết quả dự đoán giá nhà.

Cụ thể, phân cụm giúp trả lời câu hỏi :

Thị trường nhà ở có thể được chia thành những nhóm (phân khúc) nào dựa trên đặc điểm vật lý, và mỗi nhóm có đặc điểm giá như thế nào?

B. Dữ liệu và biến sử dụng

Phân tích phân cụm được thực hiện không sử dụng biến giá nhà (Price) để đảm bảo tính khách quan. Các biến đầu vào bao gồm các đặc điểm vật lý của bất động sản: Area (Diện tích), Frontage (Mặt tiền), Access Road (Độ rộng đường vào), Floors (Số tầng), Bedrooms (Số phòng ngủ), Bathrooms (Số phòng tắm)

Các biến này phản ánh quy mô và công năng sử dụng của bất động sản, phù hợp cho việc phân khúc thị trường nhà ở.

C. Chuẩn hóa dữ liệu

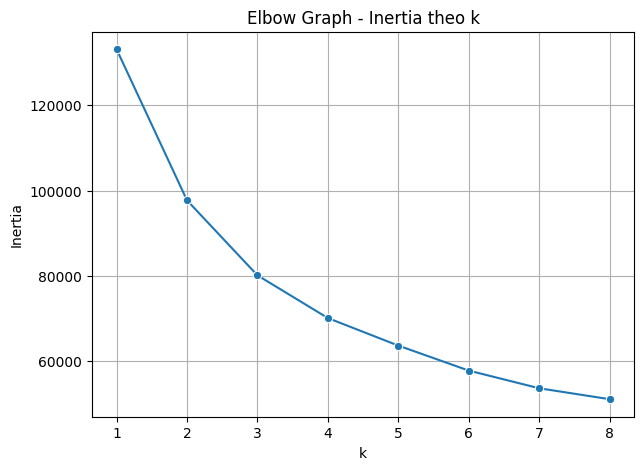
Trước khi thực hiện phân cụm, ta cần chuẩn hóa dữ liệu bằng phương pháp StandardScaler để đưa các biến về cùng thang đo, tránh việc các biến có giá trị lớn (như diện tích) chi phối kết quả phân cụm.

D. Xác định số cụm tối ưu

Số cụm tối ưu được xác định thông qua phương pháp Elbow, dựa trên tổng bình phương khoảng cách trong cụm (Within-Cluster Sum of Squares – WCSS).

Khi số cụm K tăng, giá trị WCSS giảm dần do các điểm dữ liệu được chia vào các cụm nhỏ hơn. Tuy nhiên, sau một ngưỡng nhất định, tốc độ giảm của WCSS chậm lại rõ rệt, tạo thành một “điểm khuỷu tay” (elbow point).

Giá trị K tại điểm này được xem là số cụm hợp lý, vì việc tăng thêm cụm không mang lại cải thiện đáng kể về mức độ gắn kết trong cụm.



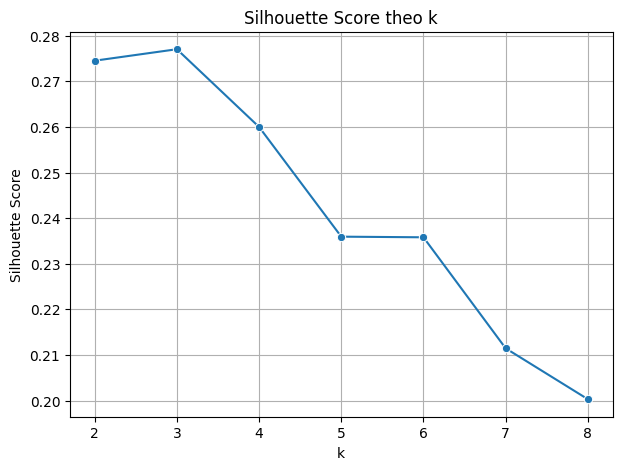
#### Hình 4.23: Biểu đồ Elbow xác định số cụm tối ưu cho thuật toán K-Means

Kết quả từ biểu đồ Elbow cho thấy điểm khuỷu tay xuất hiện tại K = 3

Bên cạnh phương pháp Elbow, nghiên cứu sử dụng hệ số Silhouette để đánh giá chất lượng phân cụm.

Hệ số Silhouette đo lường mức độ: gắn kết giữa các quan sát trong cùng một cụm (cohesion) và phân tách giữa các cụm khác nhau (separation)

Giá trị Silhouette nằm trong khoảng từ –1 đến 1, trong đó giá trị càng cao thể hiện chất lượng phân cụm càng tốt.



#### Hình 4.24: Biểu đồ hệ số Silhouette theo số cụm K

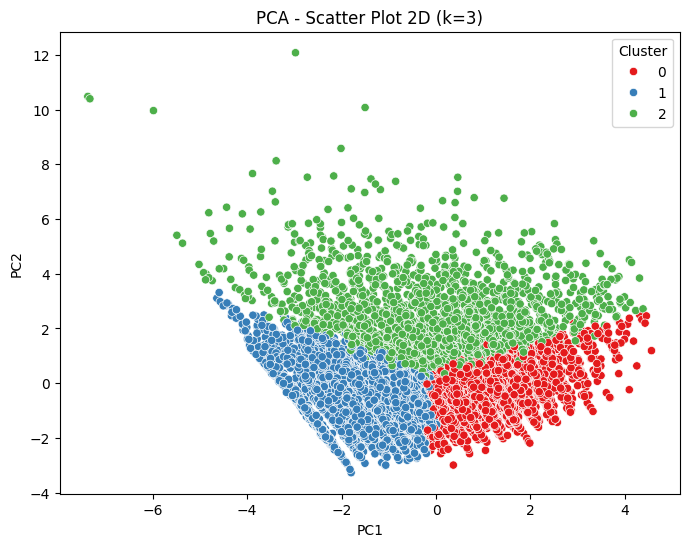
Kết quả thực nghiệm cho thấy hệ số Silhouette đạt giá trị cao nhất (hoặc ổn định nhất) tại K = 3, phù hợp với kết quả thu được từ phương pháp Elbow.

Xác định cụm tối ưu:

* Kết quả từ phương pháp Elbow cho thấy tổng bình phương khoảng cách trong cụm (WCSS) giảm mạnh khi số cụm tăng từ 1 lên 3, và tốc độ giảm chậm lại rõ rệt kể từ k = 3, cho thấy điểm khuỷu tay xuất hiện tại k = 3.
* Đồng thời, phân tích hệ số Silhouette cho thấy giá trị Silhouette đạt mức cao nhất tại k = 3 (Silhouette = 0.2771), phản ánh các cụm có mức độ gắn kết nội bộ tốt và sự tách biệt rõ ràng giữa các cụm.
* Dựa trên sự đồng thuận giữa hai phương pháp Elbow và Silhouette, nghiên cứu lựa chọn k = 3 làm số cụm tối ưu cho thuật toán K-Means nhằm phân khúc thị trường nhà ở.

E. Kết quả phân cụm

Sau khi áp dụng thuật toán K-Means với số cụm K = 3, mỗi bất động sản được gán vào một cụm tương ứng. Để đánh giá và minh họa trực quan kết quả phân cụm, nghiên cứu sử dụng phương pháp phân tích thành phần chính (Principal Component Analysis – PCA) nhằm giảm không gian đặc trưng nhiều chiều xuống còn hai thành phần chính (PC1 và PC2).



#### Hình 4.25: Biểu đồ PCA hai chiều thể hiện kết quả phân cụm K-Means với K = 3.

Biểu đồ phân tán trong không gian PC1–PC2 cho thấy ba cụm được hình thành có sự phân tách tương đối rõ ràng. Mỗi điểm dữ liệu đại diện cho một bất động sản và được tô màu theo cụm tương ứng. Mặc dù vẫn tồn tại một số vùng chồng lấn giữa các cụm – phản ánh đặc trưng liên tục của dữ liệu nhà ở thực tế – cấu trúc phân cụm tổng thể vẫn thể hiện rõ ràng.

Kết quả này cho thấy thuật toán K-Means đã nhận diện được các nhóm bất động sản có đặc điểm vật lý khác biệt, đồng thời xác nhận tính hợp lý của việc lựa chọn số cụm K = 3.

F. Đặc điểm các cụm nhà ở

Để làm rõ sự khác biệt giữa các cụm, nghiên cứu tiến hành tổng hợp giá trị trung bình của các biến đặc trưng vật lý trong từng cụm, bao gồm diện tích, mặt tiền, độ rộng đường vào, số tầng, số phòng ngủ và số phòng tắm.

##### Bảng 4.1: Giá trị trung bình các đặc trưng vật lý và giá nhà theo từng cụm.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Area | Frontage | Access Road | Floors | Bedrooms | Bathrooms | Price |
| 0 | 48.1 | 4.2 | 4.0 | 4.2 | 3.7 | 3.7 | 6.2 |
| 1 | 68.2 | 4.5 | 5.3 | 2.0 | 2.2 | 1.8 | 4.8 |
| 2 | 93.7 | 5.6 | 8.2 | 3.0 | 4.1 | 4.0 | 6.6 |

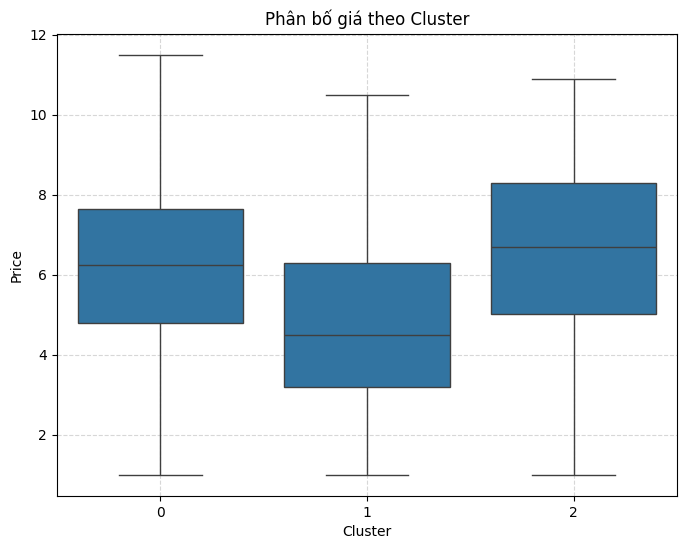
Cụm1– Nhà quy mô nhỏ, công năng thấp:  
 Cụm 1 có đặc trưng là các bất động sản có quy mô và công năng sử dụng thấp nhất trong ba cụm. Diện tích trung bình tương đối nhỏ (khoảng 68 m²), số tầng thấp, số phòng ngủ và phòng tắm hạn chế. Bên cạnh đó, độ rộng đường vào và mặt tiền ở mức trung bình thấp, cho thấy nhóm nhà này chủ yếu nằm trong các khu dân cư hiện hữu, hẻm nhỏ hoặc khu vực có khả năng tiếp cận giao thông hạn chế. Với các đặc điểm trên, cụm 1 phản ánh phân khúc nhà ở phục vụ nhu cầu ở cơ bản, thường gắn với mức giá thấp và tính ổn định cao.

Cụm 0 – Nhà ở gia đình quy mô trung bình:  
 Cụm 0 đại diện cho nhóm nhà ở có quy mô và công năng ở mức trung bình, với diện tích trung bình khoảng 48 m² nhưng được bù đắp bởi số tầng và số phòng chức năng tương đối cao. Các chỉ tiêu về mặt tiền và đường vào ở mức vừa phải, cho thấy khả năng tiếp cận giao thông thuận lợi hơn so với cụm 1. Nhìn chung, các bất động sản trong cụm này có cấu trúc không gian hợp lý, đáp ứng tốt nhu cầu sinh hoạt của các hộ gia đình. Đây là phân khúc phổ biến trên thị trường, đóng vai trò trung tâm trong cấu trúc thị trường nhà ở đô thị.

Cụm 2 – Nhà quy mô lớn, công năng cao:  
 Cụm 2 bao gồm các bất động sản có quy mô lớn và mức độ hoàn thiện cao nhất trong ba cụm. Diện tích trung bình đạt gần 94 m², cao hơn rõ rệt so với hai cụm còn lại. Đồng thời, các chỉ tiêu về mặt tiền và độ rộng đường vào cũng cao hơn, phản ánh vị trí thuận lợi và khả năng tiếp cận giao thông tốt. Số phòng ngủ, phòng tắm và số tầng ở mức cao cho thấy khả năng khai thác linh hoạt cho nhiều mục đích sử dụng. Nhóm nhà này thường gắn với phân khúc giá cao, phù hợp cho mục đích đầu tư, kinh doanh hoặc khai thác thương mại.

G. Phân tích giá nhà theo cụm

Mặc dù biến giá nhà (Price) không được sử dụng trong quá trình phân cụm, nghiên cứu tiến hành phân tích và so sánh giá trung vị (median) giữa các cụm nhằm đánh giá mối liên hệ giữa các đặc điểm vật lý của bất động sản và giá trị thị trường.



#### Hình 4.26: Phân bố giá theo cụm

Kết quả biểu đồ boxplot cho thấy sự phân hóa rõ rệt về mức giá giữa các cụm. Việc sử dụng giá trung vị thay vì giá trị trung bình giúp hạn chế ảnh hưởng của các giá trị ngoại lai và phản ánh chính xác hơn mức giá điển hình của từng phân khúc nhà ở.

Cụ thể, cụm 1 có giá trung vị thấp nhất và khoảng phân vị (IQR) tương đối hẹp, cho thấy mức giá thấp và độ biến động giá nhỏ. Điều này phù hợp với đặc điểm của nhóm nhà có quy mô nhỏ và công năng sử dụng thấp, chủ yếu phục vụ nhu cầu ở cơ bản.

Cụm 0 có giá trung vị cao hơn cụm 1, đồng thời độ phân tán giá ở mức trung bình, phản ánh phân khúc nhà ở gia đình quy mô trung bình với mức giá tương đối ổn định.

Trong khi đó, cụm 2 có giá trung vị cao nhất và độ phân tán giá lớn hơn rõ rệt so với hai cụm còn lại. Điều này cho thấy nhóm nhà có quy mô lớn và công năng cao thường gắn với mức giá cao, đồng thời thể hiện tiềm năng đầu tư đi kèm với mức độ rủi ro lớn hơn.

Nhìn chung, mặc dù giá nhà không được sử dụng trong quá trình hình thành cụm, kết quả phân tích boxplot cho thấy các cụm nhà được phân loại theo đặc điểm vật lý có sự khác biệt rõ ràng về giá trị thị trường. Kết quả này củng cố giả định rằng các đặc điểm vật lý đóng vai trò quan trọng trong việc hình thành giá nhà và là cơ sở hợp lý cho các mô hình dự đoán giá ở các phần tiếp theo của nghiên cứu.

H. Ý nghĩa và thảo luận kết quả phân cụm

Mặc dù biến giá nhà (Price) không được sử dụng trong quá trình phân cụm, kết quả phân tích cho thấy các cụm bất động sản hình thành dựa trên đặc điểm vật lý vẫn thể hiện sự khác biệt rõ rệt về giá trị thị trường. Việc so sánh phân bố giá nhà giữa các cụm, thông qua biểu đồ boxplot, cho phép đánh giá mối liên hệ giữa đặc điểm vật lý và giá nhà một cách khách quan.

Kết quả trực quan hóa cho thấy các cụm có mức giá trung vị và độ biến động giá khác nhau, phản ánh sự phân hóa rõ ràng giữa các phân khúc nhà ở. Cụ thể, các cụm nhà có quy mô và công năng sử dụng lớn hơn thường đi kèm với mức giá trung vị cao hơn. Ngược lại, các cụm nhà có quy mô nhỏ và công năng thấp có mức giá thấp hơn và độ phân tán giá nhỏ hơn, phản ánh tính ổn định tương đối của phân khúc này. Một số cụm có mức giá cao đồng thời đi kèm với độ biến động lớn, cho thấy tiềm năng đầu tư song hành với mức độ rủi ro cao hơn.

Những kết quả trên cho thấy, mặc dù giá nhà không tham gia vào quá trình hình thành cụm, các đặc điểm vật lý của bất động sản có mối quan hệ chặt chẽ với giá trị thị trường. Điều này khẳng định tính hợp lý của việc sử dụng các biến vật lý để phân khúc thị trường nhà ở và củng cố ý nghĩa kinh tế của các cụm thu được từ thuật toán K-Means.

Từ góc độ thực tiễn, kết quả phân cụm mang lại nhiều hàm ý quan trọng. Đối với người mua nhà, việc phân chia thị trường thành các phân khúc có đặc điểm vật lý tương đồng giúp so sánh giá nhà một cách hợp lý, tránh so sánh giữa các loại hình bất động sản khác biệt về quy mô và công năng. Đối với nhà đầu tư, kết quả phân cụm hỗ trợ nhận diện các phân khúc phù hợp với khẩu vị rủi ro, từ các phân khúc ổn định đến các phân khúc có tiềm năng sinh lời cao nhưng biến động lớn. Đối với nhà quản lý và hoạch định chính sách, phân cụm giúp làm rõ cấu trúc thị trường nhà ở, qua đó hỗ trợ công tác quy hoạch đô thị và quản lý thị trường bất động sản hiệu quả hơn.

Trả lời cho câu hỏi nghiên cứu:

* Dựa trên các đặc điểm vật lý của bất động sản, thị trường nhà ở có thể được chia thành ba phân khúc chính. Phân khúc thứ nhất là nhà quy mô nhỏ, công năng thấp, đặc trưng bởi diện tích nhỏ, ít tầng và số phòng hạn chế; nhóm này thường có mức giá trung vị thấp và biến động giá nhỏ. Phân khúc thứ hai là nhà ở gia đình quy mô trung bình, với các đặc điểm vật lý và công năng sử dụng ở mức vừa phải; phân khúc này có mức giá trung vị trung bình và tương đối ổn định. Phân khúc thứ ba là nhà quy mô lớn, công năng cao, có diện tích lớn, mặt tiền và đường vào rộng, nhiều phòng chức năng; nhóm này gắn với mức giá trung vị cao nhất và độ biến động giá lớn hơn, phản ánh tiềm năng đầu tư đi kèm với rủi ro cao hơn.

### 4.2.2 Dự đoán giá nhà bằng mô hình máy học

A. Mục tiêu và câu hỏi nghiên cứu

Mục tiêu của chương này là xây dựng và đánh giá các mô hình học máy nhằm dự đoán giá nhà dựa trên các đặc điểm của bất động sản và yếu tố vị trí. Thông qua việc so sánh và tối ưu hiệu suất của các mô hình, nghiên cứu hướng đến việc xác định mô hình dự đoán phù hợp nhất, đồng thời phân tích mức độ quan trọng của các đặc trưng đầu vào trong việc quyết định giá nhà. Ngoài ra, nghiên cứu cũng xem xét sự khác biệt về mức độ ảnh hưởng của các yếu tố này giữa các thành phố khác nhau.

Câu hỏi nghiên cứu đặt ra là:  
 Mô hình học máy dự đoán giá nhà dựa trên các đặc điểm của bất động sản và vị trí hoạt động hiệu quả đến mức nào, những yếu tố nào đóng vai trò quan trọng nhất trong việc quyết định giá nhà?

B. Dữ liệu và đặc trưng đầu vào

Mô hình được huấn luyện trên bộ dữ liệu giá nhà đã được tiền xử lý ở các chương trước. Dữ liệu bao gồm thông tin về đặc điểm vật lý của bất động sản, đặc điểm vị trí và giá nhà làm biến mục tiêu.

Các đặc trưng đầu vào được chia thành hai nhóm chính:

* Về đặc trưng vật lý: Area, Frontage, Access Road, Floors, Bedrooms, Bathrooms.
* Về đặc trưng vị trí: Thành phố (City), quận/huyện (District), Legal status, Furniture state.

Lý do không dùng phân cụm để làm đặc trưng (biến đầu vào của mô hình):

* Lý do là các cụm được hình thành chủ yếu dựa trên đặc điểm vật lý, trong khi mức độ chênh lệch giá nhà giữa các cụm không đủ lớn và có sự chồng lấn đáng kể về phân bố giá. Do đó, việc đưa biến Cluster vào mô hình có thể không mang lại thêm nhiều thông tin dự đoán mới, thậm chí có nguy cơ gây nhiễu và làm giảm khả năng tổng quát hóa của mô hình.

C. Các mô hình máy học được sử dụng.

Random Forest Regression – mô hình phi tuyến, nắm bắt quan hệ phức tạp.

XGBoost – mô hình tăng cường, tối ưu hiệu suất dự đoán

D. Quy trình huấn luyện và đánh giá mô hình.

Chia tập dữ liệu: Dữ liệu được chia thành tập huấn luyện (80%) và tập kiểm tra (20%) nhằm đánh giá khả năng tổng quát hóa của mô hình.

Chuẩn hóa dữ liệu:

* Trước khi xây dựng mô hình học máy, dữ liệu được tiền xử lý nhằm đảm bảo tính nhất quán và hiệu quả trong quá trình huấn luyện. Các biến phân loại được mã hóa nhãn (Label Encoding) để chuyển đổi từ dạng định tính sang dạng số, giúp mô hình có thể xử lý được thông tin phân loại.\
* Đối với các biến số, nghiên cứu sử dụng phương pháp StandardScaler để chuẩn hóa dữ liệu, đưa các đặc trưng về cùng thang đo với giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Việc chuẩn hóa bằng StandardScaler giúp tránh hiện tượng các biến có giá trị lớn (như diện tích) chi phối mô hình, đồng thời cải thiện độ ổn định và hiệu suất của các thuật toán học máy.

E. So sánh kết quả và tối ưu tham số

Nghiên cứu tiến hành xây dựng và so sánh hiệu suất của hai mô hình học máy gồm Random Forest và XGBoost trong bài toán dự đoán giá nhà. Hiệu suất mô hình được đánh giá thông qua các chỉ số RMSE, MAE và R².

##### Bảng 4.2: Kết quả của 2 mô hình máy học:

|  |  |  |  |
| --- | --- | --- | --- |
| Mô hình | RMSE | MAE | R² |
| XGBoost | 1.32 | 0.99 | 0.6488 |
| Random Forest | 1.42 | 1.06 | 0.5926 |

Kết quả cho thấy mô hình XGBoost đạt hiệu suất tốt hơn so với Random Forest trên cả ba chỉ số đánh giá. Cụ thể, XGBoost có RMSE = 1.32 và MAE = 0.99, thấp hơn so với Random Forest (RMSE = 1.42, MAE = 1.06), cho thấy sai số dự đoán trung bình của XGBoost nhỏ hơn. Đồng thời, hệ số xác định R² của XGBoost đạt 0.6488, cao hơn đáng kể so với Random Forest (R² = 0.5926), phản ánh khả năng giải thích biến thiên của giá nhà tốt hơn.

Nhìn chung, kết quả này cho thấy XGBoost là mô hình phù hợp hơn cho bài toán dự đoán giá nhà trong nghiên cứu, nhờ khả năng mô hình hóa các mối quan hệ phi tuyến và tương tác phức tạp giữa các đặc trưng bất động sản.

Tối ưu hóa siêu tham số cho mô hình XGBoost:

Sau khi xác định XGBoost là mô hình có hiệu suất tốt nhất trong giai đoạn so sánh ban đầu, nghiên cứu tiếp tục thực hiện tối ưu hóa siêu tham số nhằm cải thiện khả năng dự đoán và mức độ khái quát của mô hình. Quá trình tối ưu được thực hiện bằng phương pháp Grid Search kết hợp với Cross-Validation (CV).

Cụ thể, các siêu tham số quan trọng của mô hình XGBoost liên quan đến độ phức tạp của cây, tốc độ học, khả năng lấy mẫu ngẫu nhiên và regularization được xem xét đồng thời. Việc đánh giá các tổ hợp tham số được thực hiện thông qua cross-validation 3 lần, sử dụng hệ số R² làm thước đo chính để lựa chọn mô hình tối ưu.

Kết quả Grid Search cho thấy bộ siêu tham số tối ưu của mô hình XGBoost bao gồm: số lượng cây lớn kết hợp với tốc độ học nhỏ, giúp mô hình học ổn định và giảm nguy cơ overfitting, độ sâu cây ở mức trung bình, cân bằng giữa khả năng học quan hệ phi tuyến và tính khái quát và các tham số regularization được điều chỉnh nhằm kiểm soát độ phức tạp của mô hình

Việc lựa chọn các tham số này cho thấy mô hình XGBoost đạt hiệu quả cao nhất khi được thiết kế theo hướng học chậm, nhiều cây và có cơ chế kiểm soát overfitting rõ ràng. Trên tập kiểm tra, mô hình XGBoost sau tối ưu cho kết quả dự đoán tốt hơn so với phiên bản ban đầu, qua đó được lựa chọn làm mô hình cuối cùng cho các phân tích tiếp theo về tầm quan trọng của đặc trưng và diễn giải kết quả dự đoán giá nhà.

Kết quả tối ưu hóa mô hình XGBoost:

Optimized XGBoost (GridSearch): RMSE: 1.31, MAE: 0.97, R²: 0.6528

Kết quả sau tối ưu cho thấy hiệu suất của mô hình XGBoost được cải thiện nhẹ nhưng ổn định. Cụ thể, RMSE giảm từ 1.32 xuống 1.31, MAE giảm từ 0.99 xuống 0.97, trong khi hệ số xác định R² tăng từ 0.6488 lên 0.6528. Những cải thiện này cho thấy việc tinh chỉnh siêu tham số giúp mô hình dự đoán giá nhà chính xác hơn và giải thích tốt hơn mức độ biến thiên của biến mục tiêu.

Mặc dù mức cải thiện không quá lớn, kết quả này khẳng định XGBoost sau tối ưu là mô hình có hiệu suất cao nhất trong nghiên cứu, và được lựa chọn làm mô hình cuối cùng để phân tích tầm quan trọng của các đặc trưng và phục vụ diễn giải kết quả dự đoán giá nhà.

So sánh giá thực tế và giá dự đoán

Mối quan hệ giữa giá nhà thực tế và giá nhà dự đoán bởi mô hình XGBoost sau khi tối ưu thông qua biểu đồ scatter. Đường chéo y=x được sử dụng làm đường tham chiếu, đại diện cho trường hợp mô hình dự đoán hoàn toàn chính xác.



#### Hình 4.27: Hình biểu diễn giá thực và giá dự đoán

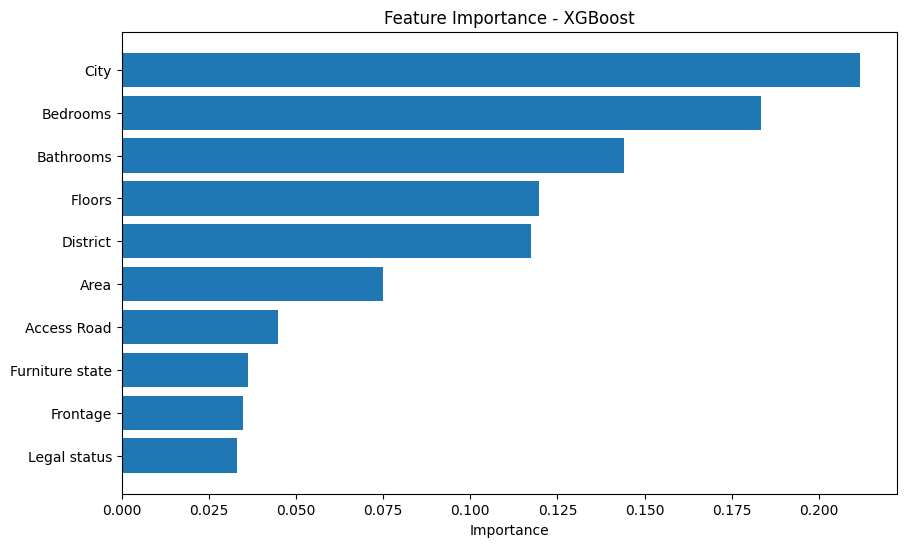
Quan sát hình cho thấy phần lớn các điểm dữ liệu phân bố tương đối sát đường y=x, đặc biệt trong khoảng giá trung bình, cho thấy mô hình có khả năng dự đoán tốt đối với đa số bất động sản trên thị trường. Điều này phản ánh mức độ phù hợp khá cao giữa giá dự đoán và giá thực tế, phù hợp với các chỉ số đánh giá định lượng đã trình bày trước đó.

Tuy nhiên, tại vùng giá cao, các điểm dữ liệu có xu hướng phân tán rộng hơn quanh đường chuẩn, cho thấy sai số dự đoán tăng lên đối với các bất động sản có giá trị lớn. Hiện tượng này cho thấy mô hình gặp nhiều khó khăn hơn khi dự đoán các trường hợp nhà ở cao cấp, vốn có đặc điểm đa dạng và chịu ảnh hưởng mạnh từ các yếu tố vị trí hoặc yếu tố đặc thù chưa được phản ánh đầy đủ trong dữ liệu.

Nhìn chung, biểu đồ Actual vs Predicted cho thấy mô hình XGBoost tối ưu hoạt động ổn định và đáng tin cậy ở phân khúc giá phổ biến, đồng thời gợi ý hướng cải thiện trong tương lai đối với phân khúc nhà giá cao.

F. Phân tích tầm quan trọng của đặc trưng

Dựa trên mô hình XGBoost đã được tối ưu, nghiên cứu tiến hành phân tích tầm quan trọng của các đặc trưng đầu vào nhằm xác định những yếu tố có ảnh hưởng lớn nhất đến giá nhà. Phân tích này phản ánh mức độ đóng góp tương đối của từng biến trong quá trình ra quyết định của mô hình.



#### Hình 4.28: Tầm quan trọng của các biến

Kết quả cho thấy yếu tố vị trí đóng vai trò quan trọng nhất trong việc quyết định giá nhà. Cụ thể, biến City có mức độ quan trọng cao nhất trong mô hình, cho thấy sự khác biệt đáng kể về mặt bằng giá giữa các thành phố. Bên cạnh đó, biến District cũng nằm trong nhóm đặc trưng quan trọng, phản ánh tác động rõ rệt của vị trí chi tiết hơn trong cùng một thành phố đến giá trị bất động sản.

Nhóm các đặc trưng phản ánh quy mô và công năng sử dụng của bất động sản cũng có ảnh hưởng lớn đến giá nhà. Các biến như Bedrooms, Bathrooms và Floors có mức độ quan trọng cao, cho thấy số lượng không gian sử dụng và khả năng khai thác công năng là những yếu tố then chốt trong việc hình thành giá. Đáng chú ý, trong nhóm này, số phòng ngủ và phòng tắm có mức ảnh hưởng lớn hơn diện tích, cho thấy giá trị sử dụng thực tế đôi khi được mô hình đánh giá cao hơn quy mô thuần túy.

Biến Area (diện tích) tuy vẫn giữ vai trò quan trọng nhưng có mức độ ảnh hưởng thấp hơn so với các biến về công năng và vị trí. Điều này cho thấy diện tích không phải là yếu tố duy nhất quyết định giá nhà, mà cần được xem xét đồng thời với các yếu tố khác như vị trí và thiết kế công năng.

Các yếu tố liên quan đến khả năng tiếp cận và điều kiện pháp lý – tiện nghi, bao gồm Access Road, Frontage, Furniture state và Legal status, có mức độ ảnh hưởng thấp hơn nhưng vẫn góp phần giải thích sự khác biệt về giá nhà. Những yếu tố này thường đóng vai trò bổ trợ, giúp điều chỉnh giá trị bất động sản trong các trường hợp cụ thể, đặc biệt đối với nhà mặt tiền hoặc nhà có tiềm năng khai thác thương mại.

Kết quả feature importance

* Nhìn chung, kết quả phân tích feature importance cho thấy giá nhà chịu tác động mạnh mẽ bởi sự kết hợp giữa vị trí địa lý và đặc điểm công năng của bất động sản. Đồng thời, sự khác biệt về mức độ quan trọng của các đặc trưng cũng lý giải vì sao mô hình XGBoost có thể dự đoán tốt ở phân khúc giá phổ biến nhưng gặp nhiều thách thức hơn ở phân khúc giá cao, nơi các yếu tố đặc thù và khó định lượng đóng vai trò lớn hơn. Kết quả này không chỉ giúp nâng cao khả năng diễn giải mô hình học máy, mà còn cung cấp các hàm ý thực tiễn cho người mua, nhà đầu tư và nhà quản lý trong việc đánh giá và so sánh giá trị bất động sản.

G. Ý nghĩa và thảo luận kết quả mô hình máy học

Trong nghiên cứu này, mô hình học máy được xây dựng nhằm dự đoán giá nhà dựa trên các đặc điểm vật lý, hạ tầng và yếu tố vị trí của bất động sản. Kết quả đánh giá mô hình tốt nhất (XGBoost đã tối ưu tham số) cho thấy khả năng dự đoán tương đối chính xác, với các chỉ số: RMSE = 1.31, MAE = 0.97 và R² = 0.6528.

Điều này chứng tỏ mô hình có thể giải thích được khoảng 65% sự biến thiên của giá nhà trong tập dữ liệu, phản ánh hiệu quả dự đoán ở mức khá và có thể ứng dụng trong thực tiễn định giá tự động hoặc hỗ trợ ra quyết định đầu tư.

Kết quả phân tích mức độ quan trọng của các đặc trưng đầu vào cho thấy các yếu tố ảnh hưởng lớn nhất đến giá nhà gồm:

Vị trí địa lý: City và District là hai đặc trưng có mức độ quan trọng cao nhất, khẳng định vị trí là yếu tố quyết định hàng đầu trong định giá bất động sản. Điều này phù hợp với thực tiễn thị trường khi giá nhà biến động mạnh theo khu vực đô thị, địa hình và khả năng tiếp cận dịch vụ tiện ích.

Đặc điểm công năng sử dụng của căn nhà: Bedrooms, Bathrooms và Floors có mức đóng góp đáng kể. Điều này cho thấy người mua chú trọng đến giá trị sử dụng thực tế, tiện ích sinh hoạt và tiêu chuẩn không gian hơn là chỉ xem xét diện tích thuần túy.

Diện tích và hạ tầng tiếp cận: Area, Access Road, Frontage và Legal status tuy có mức độ quan trọng thấp hơn nhưng vẫn ảnh hưởng rõ rệt trong quá trình hình thành giá. Các yếu tố này đóng vai trò điều chỉnh và tạo sự chênh lệch giá giữa các căn nhà cùng khu vực.

Kết quả trên giúp trả lời câu hỏi nghiên cứu đặt ra:

* Mô hình học máy (XGBoost) hoạt động hiệu quả với mức độ dự đoán tương đối chính xác, thể hiện qua chỉ số R² đạt khoảng 0.65. Các yếu tố quan trọng nhất quyết định giá nhà là vị trí địa lý (thành phố, quận / huyện ), tiếp theo là đặc điểm công năng sử dụng (số phòng ngủ, phòng tắm, số tầng) và diện tích. Các yếu tố về hạ tầng tiếp cận, mặt tiền và pháp lý ảnh hưởng ở mức điều chỉnh.

## 4.3. Diễn giải ý nghĩa kinh tế

Kết quả phân tích mức độ quan trọng của các đặc trưng mang lại nhiều ý nghĩa kinh tế trong việc lý giải cơ chế hình thành giá bất động sản tại Việt Nam, đồng thời phản ánh rõ hành vi định giá của thị trường nhà ở đô thị.

Trước hết, sự nổi trội của biến City cho thấy vị trí địa lý vĩ mô đóng vai trò quyết định trong việc hình thành mặt bằng giá bất động sản. Các đô thị lớn như Hà Nội và TP. Hồ Chí Minh, với mức độ đô thị hóa cao, nhu cầu nhà ở lớn và nguồn cung đất hạn chế, thường có giá nhà cao hơn đáng kể so với các tỉnh thành khác. Mô hình học máy đã phản ánh chính xác thực tế này thông qua mức độ quan trọng vượt trội của biến City. Tương tự, biến District có importance cao cho thấy sự phân hóa giá rõ rệt ngay trong cùng một thành phố, trong đó các quận trung tâm như Quận 1, Quận 3, Ba Đình hay Hoàn Kiếm có giá nhà cao hơn đáng kể so với các khu vực ngoại thành. Điều này khẳng định vai trò chi phối của vị trí trong cấu trúc giá nhà tại thị trường Việt Nam.

Thứ hai, nhóm các biến phản ánh công năng và quy mô sử dụng của ngôi nhà bao gồm Bathrooms, Floors và Bedrooms có mức độ ảnh hưởng cao ngay sau yếu tố vị trí. Kết quả này cho thấy giá nhà không chỉ phụ thuộc vào diện tích đất, mà còn phụ thuộc mạnh vào khả năng đáp ứng nhu cầu sinh hoạt và khai thác sử dụng. Trong đó, biến Bathrooms có mức độ quan trọng cao nhất trong nhóm này, phản ánh xu hướng thị trường hiện nay khi người mua nhà, đặc biệt tại các đô thị lớn, ngày càng coi trọng tiện nghi và chất lượng không gian sống hơn là diện tích thuần.

Thứ ba, biến Area mặc dù vẫn có ảnh hưởng đáng kể nhưng không phải là yếu tố quyết định hàng đầu. Điều này phản ánh đúng thực trạng thị trường bất động sản Việt Nam, nơi mà diện tích lớn chưa chắc đi kèm với giá trị cao nếu bất động sản nằm ở vị trí kém thuận lợi. Ngược lại, các căn nhà có diện tích nhỏ hoặc trung bình nhưng nằm tại khu vực trung tâm, hạ tầng tốt vẫn có thể đạt mức giá rất cao. Mô hình đã học được mối quan hệ phi tuyến này thông qua dữ liệu thực tế.

Bên cạnh đó, các yếu tố như Frontage, Access Road và Legal Status đóng vai trò bổ trợ trong quá trình định giá. Nhà có mặt tiền rộng hoặc đường vào thuận tiện thường phù hợp cho mục đích kinh doanh, từ đó có giá trị cao hơn so với nhà trong hẻm nhỏ. Trong khi đó, yếu tố pháp lý (Legal Status) tuy có mức độ ảnh hưởng thấp hơn so với các đặc trưng khác nhưng vẫn mang ý nghĩa quan trọng trong việc phân biệt mức độ an toàn và tính thanh khoản của bất động sản, đặc biệt giữa các tài sản có đầy đủ giấy tờ pháp lý và các tài sản chưa hoàn chỉnh về mặt pháp lý.

Tổng hợp các kết quả cho thấy thị trường bất động sản Việt Nam vận hành dựa trên hai nhóm yếu tố chính:  
 Vị trí địa lý, quyết định mặt bằng giá và mức độ chênh lệch giữa các khu vực;  
 Công năng sử dụng và tiện ích của ngôi nhà, điều chỉnh mức giá cụ thể trong từng khu vực và từng phân khúc.

Những phát hiện này không chỉ giúp cải thiện khả năng dự đoán của mô hình học máy mà còn cung cấp cơ sở định lượng quan trọng cho việc định giá sơ bộ, tư vấn đầu tư và phân tích thị trường bất động sản theo hướng thực tiễn và phù hợp với bối cảnh Việt Nam.

# CHƯƠNG 5. KẾT LUẬN VÀ KHUYẾN NGHỊ

## 5.1. Kết luận

Nghiên cứu đã thực hiện phân tích toàn diện thị trường nhà ở Việt Nam dựa trên bộ dữ liệu bất động sản năm 2024, kết hợp giữa phân tích khám phá dữ liệu (EDA), mô hình học máy dự đoán giá và phương pháp phân cụm nhằm làm rõ cấu trúc và động lực hình thành giá nhà.

Kết quả phân tích theo vị trí cho thấy thị trường nhà ở Việt Nam có sự phân hóa rõ rệt giữa các khu vực. Các đô thị lớn như TP. Hồ Chí Minh và Hà Nội có mặt bằng giá cao hơn đáng kể so với các tỉnh vệ tinh như Bình Dương, Đồng Nai hay Long An. Sự khác biệt này phản ánh mức độ đô thị hóa, áp lực dân số, hạ tầng và nhu cầu nhà ở tập trung tại các trung tâm kinh tế lớn.

Phân tích theo đặc điểm vật lý cho thấy các biến công năng như số tầng, số phòng ngủ và số phòng tắm có mối quan hệ rõ ràng và tương đối mạnh với giá nhà. Ngược lại, các biến kích thước như diện tích, mặt tiền và độ rộng đường trước nhà tuy có tác động tích cực nhưng thể hiện xu hướng bão hòa, tức là giá chỉ tăng mạnh đến một ngưỡng nhất định rồi không còn tăng tương ứng. Điều này cho thấy giá nhà không chỉ phụ thuộc vào kích thước vật lý mà còn chịu ảnh hưởng lớn từ vị trí và khả năng khai thác sử dụng.

Đối với các yếu tố danh mục, tình trạng pháp lý (Legal status) và mức độ hoàn thiện nội thất (Furniture state) cho thấy ảnh hưởng đáng kể đến giá nhà. Các bất động sản có pháp lý đầy đủ (có sổ đỏ/sổ hồng) và nội thất hoàn chỉnh thường có mức giá cao hơn và ổn định hơn so với các bất động sản chỉ có hợp đồng mua bán hoặc chưa hoàn thiện nội thất. Mức độ ảnh hưởng của các yếu tố này cũng khác nhau giữa các khu vực, phản ánh đặc thù thị trường địa phương.

Trong phần mô hình hóa, các mô hình học máy được xây dựng và đánh giá bằng các chỉ số RMSE, MAE và R². Kết quả cho thấy mô hình XGBoost sau khi tối ưu đạt hiệu suất dự đoán tốt nhất, giải thích được phần lớn biến động của giá nhà. Khi huấn luyện mô hình riêng cho từng thành phố lớn, hiệu suất dự đoán có sự khác biệt, cho thấy cấu trúc hình thành giá nhà không hoàn toàn giống nhau giữa các khu vực.

Ngoài ra, phương pháp K-means clustering đã giúp phân chia thị trường nhà ở thành các phân khúc rõ ràng dựa trên đặc điểm vật lý. Các cụm nhà đại diện cho những nhóm phổ biến như nhà hẻm nhỏ, nhà mặt tiền, nhà nhiều tầng hoặc nhà diện tích lớn, mỗi nhóm có đặc điểm giá và công năng khác nhau. Điều này góp phần làm rõ cấu trúc phân khúc của thị trường bất động sản Việt Nam.

Tổng hợp các kết quả trên cho thấy việc kết hợp phân tích dữ liệu, học máy và phân cụm là một hướng tiếp cận hiệu quả trong nghiên cứu giá nhà, vừa cung cấp khả năng dự đoán, vừa hỗ trợ diễn giải kinh tế và thực tiễn.

## 5.2. Hàm ý thực tiễn

Đối với người mua nhà, kết quả nghiên cứu cho thấy việc đánh giá giá trị bất động sản không nên chỉ dựa vào diện tích mà cần xem xét tổng hợp các yếu tố công năng, pháp lý và vị trí. Người mua có thể tận dụng thông tin về phân khúc và giá trung vị để lựa chọn bất động sản phù hợp với ngân sách và nhu cầu sử dụng.

Đối với nhà đầu tư, mô hình dự đoán và kết quả phân cụm cung cấp cơ sở định lượng để nhận diện các phân khúc tiềm năng, đặc biệt là các khu vực vệ tinh có mức giá thấp hơn nhưng đang trong quá trình đô thị hóa. Việc chú trọng vào các yếu tố như số tầng, khả năng khai thác công năng và pháp lý đầy đủ có thể giúp tối ưu hóa lợi nhuận đầu tư.

Đối với doanh nghiệp và nhà quản lý thị trường, nghiên cứu cho thấy dữ liệu bất động sản trực tuyến có thể được khai thác hiệu quả để xây dựng các công cụ định giá minh bạch. Các mô hình học máy có thể hỗ trợ định giá sơ bộ, giám sát biến động thị trường và góp phần giảm bất cân xứng thông tin trong giao dịch bất động sản.

## 5.3. Hạn chế của nghiên cứu

Mặc dù nghiên cứu đã đạt được nhiều kết quả tích cực trong việc phân tích và dự đoán giá nhà ở Việt Nam, vẫn tồn tại một số hạn chế nhất định cần được nhìn nhận một cách khách quan.

Thứ nhất, nguồn dữ liệu được thu thập chủ yếu từ các tin đăng bất động sản trực tuyến. Do đó, mức giá được sử dụng trong nghiên cứu là giá chào bán, không hoàn toàn phản ánh giá giao dịch thực tế trên thị trường. Trên thực tế, giá giao dịch có thể thấp hơn hoặc cao hơn giá chào bán tùy thuộc vào khả năng thương lượng, tình trạng thị trường và các yếu tố cá nhân của người mua – người bán. Điều này có thể ảnh hưởng đến độ chính xác tuyệt đối của mô hình dự đoán.

Thứ hai, một số đặc trưng quan trọng có khả năng tác động mạnh đến giá nhà chưa được đưa vào mô hình do hạn chế về dữ liệu. Các yếu tố như hướng nhà, khoảng cách đến trung tâm thành phố, chất lượng hạ tầng giao thông, tiện ích xung quanh (trường học, bệnh viện, trung tâm thương mại) hoặc môi trường sống (không gian xanh, mức độ ô nhiễm) chưa được khai thác đầy đủ. Việc thiếu các biến này có thể làm giảm khả năng giải thích toàn diện sự biến động giá nhà giữa các khu vực.

Thứ ba, nghiên cứu chưa xét đến yếu tố thời gian trong phân tích. Giá bất động sản thường chịu ảnh hưởng mạnh từ chu kỳ kinh tế, chính sách tín dụng, lãi suất ngân hàng và các yếu tố vĩ mô khác. Việc sử dụng dữ liệu tại một thời điểm hoặc trong một giai đoạn ngắn có thể chưa phản ánh đầy đủ xu hướng biến động dài hạn của thị trường nhà ở.

Cuối cùng, phương pháp phân cụm và mô hình học máy được sử dụng trong nghiên cứu chủ yếu tập trung vào các đặc trưng vật lý và vị trí cơ bản của bất động sản. Do đó, kết quả phân khúc thị trường mang tính tổng quát, chưa đi sâu vào các phân khúc đặc thù như nhà đầu tư cho thuê, nhà ở xã hội hoặc bất động sản nghỉ dưỡng.

## 5.4. Hướng phát triển trong tương lai

Dựa trên những hạn chế đã nêu, nghiên cứu trong tương lai có thể được mở rộng và hoàn thiện theo nhiều hướng khác nhau nhằm nâng cao giá trị học thuật và tính ứng dụng thực tiễn.

Thứ nhất, việc kết hợp dữ liệu giao dịch thực tế từ các sàn bất động sản hoặc cơ quan quản lý nhà nước sẽ giúp cải thiện đáng kể độ chính xác của mô hình dự đoán. Ngoài ra, việc bổ sung dữ liệu về hạ tầng, tiện ích đô thị và yếu tố không gian địa lý sẽ cho phép mô hình phản ánh sát hơn giá trị thực của bất động sản.

Thứ hai, các phương pháp diễn giải mô hình nâng cao như SHAP (SHapley Additive exPlanations) hoặc Permutation Importance có thể được áp dụng để phân tích sâu hơn mức độ ảnh hưởng của từng đặc trưng đến giá nhà. Điều này không chỉ giúp cải thiện tính minh bạch của mô hình học máy mà còn hỗ trợ người dùng hiểu rõ hơn cơ chế hình thành giá.

Thứ ba, nghiên cứu có thể phát triển theo hướng phân tích không gian (spatial analysis), kết hợp dữ liệu tọa độ địa lý để đánh giá sự lan tỏa giá nhà giữa các khu vực lân cận. Cách tiếp cận này đặc biệt phù hợp với các đô thị lớn, nơi mà vị trí và kết nối hạ tầng đóng vai trò then chốt trong việc định giá bất động sản.

Cuối cùng, việc xây dựng mô hình dự báo theo chuỗi thời gian (time-series forecasting) sẽ giúp phân tích và dự đoán xu hướng giá nhà trong tương lai, từ đó hỗ trợ tốt hơn cho các quyết định đầu tư, hoạch định chính sách và quản lý thị trường bất động sản theo hướng bền vững.

TÀI LIỆU THAM KHẢO

* Rosen, S. (1974). *Hedonic prices and implicit markets: Product differentiation in pure competition.* Journal
* Fujita, M. (1989). *Urban Economic Theory: Land Use and City Size.* Cambridge University Press.
* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021).
* Breiman, L. (2001). *Random Forests.* Machine Learning, 45(1), 5–32.
* Chen, T., & Guestrin, C. (2016).
* Hastie, T., Tibshirani, R., & Friedman, J. (2017).
* MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.*
* *Tan, P.-N., Steinbach, M., & Kumar, V. (2019).*