ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH DỮ LIỆU GIÁ CHUNG CƯ TẠI TP.HCM

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Phòng Lai Bảo Minh	20522217
2	Nguyễn Nhật Hoàng	20520516

TP. HỒ CHÍ MINH – 12/2024

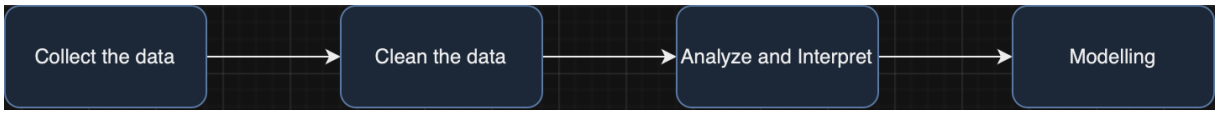
1. GIỚI THIỆU

Thông tin Chung cư TP.HCM là bộ dữ liệu được thu thập và xử lý từ nền tảng Batdongsan.com.vn, cụ thể trong mục **Chung cư bán tại TP.HCM** [1], một trong những trang web bất động sản lớn nhất Việt Nam. Nội dung bộ dữ liệu bao gồm các thông tin chi tiết liên quan đến các căn hộ chung cư đang được bán tại TP. Hồ Chí Minh, với nhiều yếu tố như tiêu đề, địa chỉ, mức giá, diện tích, số phòng ngủ, và vị trí địa lý (latitude, longitude).

Đề tài sẽ sử dụng công cụ như Pandas, Numpy để xử lý dữ liệu và Matplotlib, Seaborn để trực quan hóa nhằm phân tích các yếu tố tác động đến thị trường chung cư. Qua đó, nhóm hướng đến việc nhận định xu hướng, hỗ trợ người mua và nhà đầu tư đưa ra quyết định hợp lý, đồng thời phát triển tư duy phân tích và giải quyết vấn đề từ dữ liệu thực tế.

2. NỘI DUNG

Phương pháp phân tích cho bài toán:



Hình 1. Quy trình phân tích dữ liệu.

2.1. Thu thập và tiền xử lý dữ liệu

Quá trình crawl dữ liệu bắt đầu bằng việc thu thập liên kết tin đăng căn hộ từ trang web bất động sản TP.HCM. Mã truy cập vào từng URL để lấy thông tin chi tiết như địa chỉ, giá, diện tích, số phòng, hướng nhà và tọa độ, sau đó lưu vào file CSV với tên "Thông tin Chung cư TP.HCM". Kế đến nhóm tiến hành kiểm tra và làm sạch dữ liệu để tạo ra bộ dữ liệu hoàn chỉnh.

Quy trình xử lý dữ liệu:

1. Loại bỏ các dòng dữ liệu trùng lặp.
2. Bổ sung cột tên "Đường" nhằm cung cấp thông tin hữu ích cho quá trình phân tích sau này.
3. Chuẩn hóa “Mức giá” và “Giá/m^2” về cùng đơn vị tiền (triệu).
4. Thống kê tỷ lệ phần trăm giá trị thiếu (missing) cho từng cột.
5. Quyết định loại bỏ các cột có tỷ lệ phần trăm giá trị thiếu lớn hơn 50%.
6. Sử dụng XGBoost [2] để điền các giá trị còn thiếu sau khi đã xử lý những giá trị thiếu lớn.
7. Loại bỏ các giá trị nhiễu bằng DBSCAN [3] để đảm bảo tính chính xác của dữ liệu.

2.2. Mô tả dữ liệu

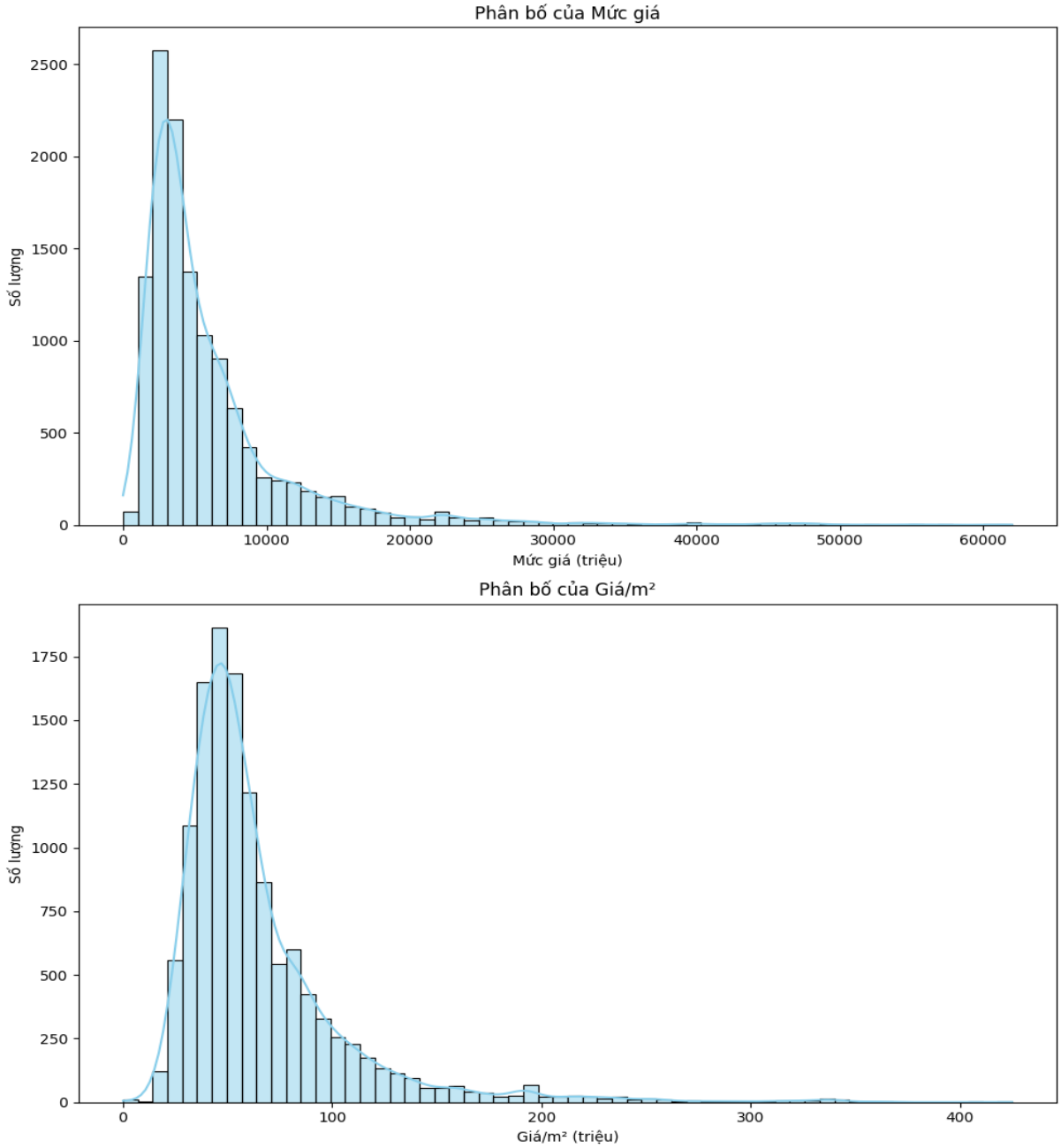
Kích thước dữ liệu ban đầu gồm 20068 dòng và 15 cột. Sau khi hoàn thành tiền xử lý thu được một bộ dữ liệu mới với kích thước 12550 dòng và 11 cột, với thông tin các cột như sau:

STT	Thuộc Tính	Kiểu Dữ Liệu	Miền Dữ Liệu
1	Địa chỉ	Object	Dự án Celadon City, Đường N1, Phường Sơn Kỳ, Tân Phú, Hồ Chí Minh, …
2	Mức giá (triệu)	Float64	3.4 - 62000
3	Giá/m^2 (triệu/m^2)	Float64	0.03867 - 424.849518
4	Số Phòng Ngủ	Int64	1 - 5
5	Huyện	Object	Nhà Bè, Thủ Đức, ...
6	Diện tích	Float64	20 - 392
7	Số toilet	Int64	1 - 5
8	Nội thất	Object	Đầy đủ, full nội thất, …
9	Longtitude	Float64	105.841194-106.851036
10	Lattitude	Float64	10.643736 - 21.029581
11	Đường	Object	Đào Trí, Phú Thuận, …

Bảng 1. Thông tin các thuộc tính bộ dữ liệu.

Sau khi giải quyết xong nhóm đã hiểu hơn về bộ dữ liệu và phát hiện ra những vấn đề sâu sắc hơn mà bộ dữ liệu mang lại, nhóm tiếp tục tiến hành phân tích thăm dò để giải quyết những vấn đề mới. Cuối cùng nhóm sẽ tổng hợp những kết quả đã phân tích được.

2.3. Phân phối giá và đơn giá của chung cư tại TP.HCM



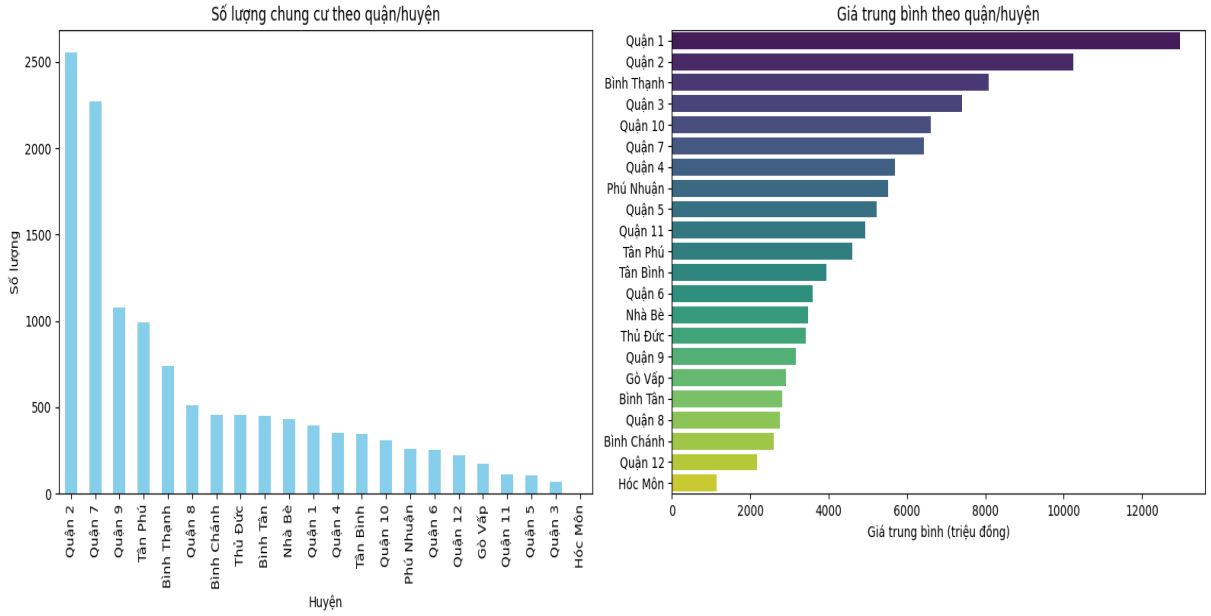
Hình 2. Phân phối giá và đơn giá chung cư tại TP.HCM.

Giá bất động sản tập trung chủ yếu dưới 10 tỷ đồng, với đỉnh phân bố khoảng 3 tỷ đồng. Khi mức giá vượt 10 tỷ, số lượng giảm mạnh, và chỉ một tỷ lệ rất nhỏ các bất động sản cao cấp có giá trên 20 tỷ đồng, thậm chí trên 60 tỷ đồng. Phân bố giá thể hiện sự lệch phải rõ rệt, với phần lớn các căn hộ thuộc mức giá trung bình thấp.

Đơn giá bất động sản chủ yếu nằm trong khoảng 20 - 100 triệu đồng/m², tập trung quanh mức 50 triệu đồng/m². Các bất động sản trên 100 triệu đồng/m² chiếm tỷ lệ rất nhỏ, phản ánh sự khan hiếm của các căn hộ cao cấp ở vị trí đắc địa. Biểu đồ cũng cho thấy sự lệch phải, với một số giá trị cực cao kéo dài phân phối về phía bên phải

2.4. Tác động của vị trí địa lý đối với giá chung cư

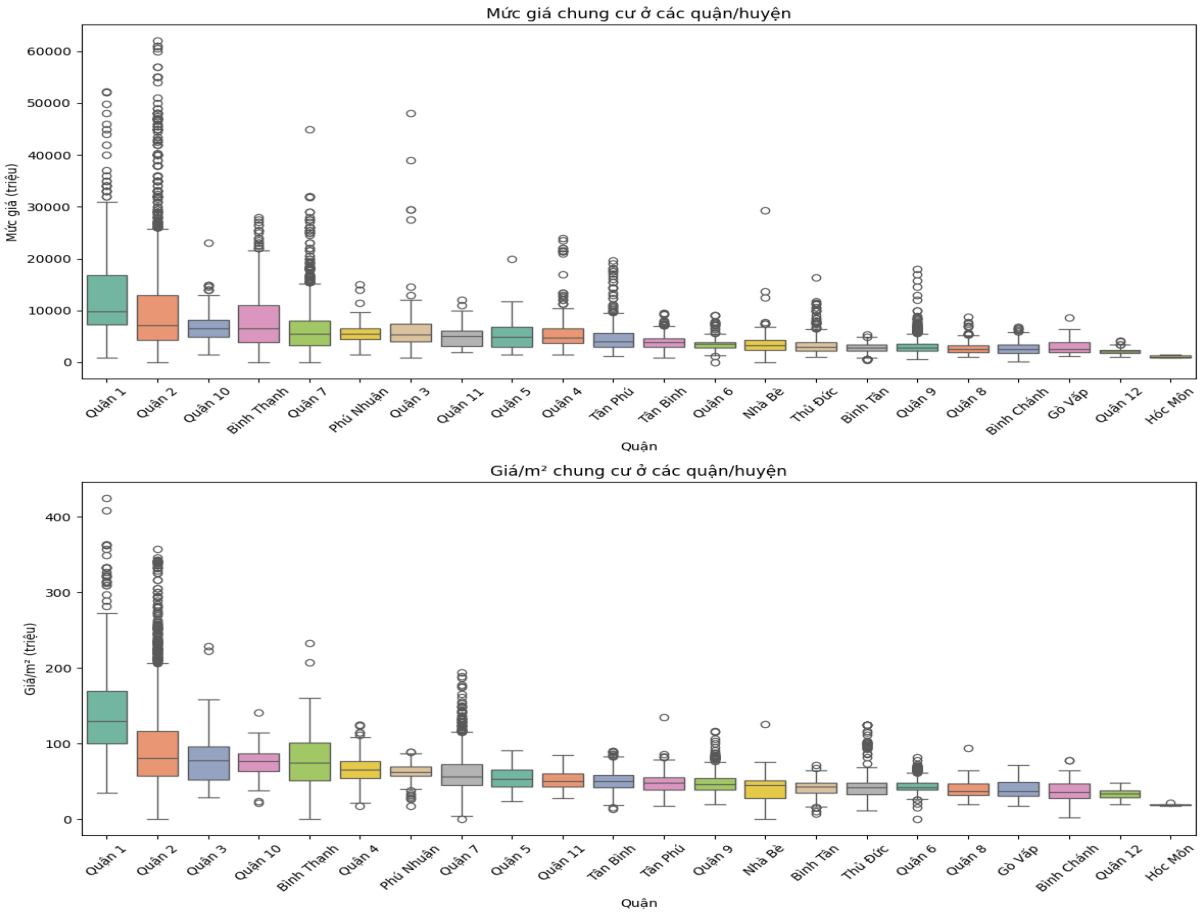
2.4.1. Quận/huyện tác động thế nào?



Hình 3. Số lượng chung cư và giá bán trung bình tại từng quận.

Quận 2 dẫn đầu với hơn 2.500 chung cư, nhờ quy hoạch tốt và hạ tầng hiện đại. Quận 7, Quận 9, và Tân Phú cũng phát triển mạnh với số lượng chung cư đáng kể. Các quận như Bình Chánh, Quận 8, và Thủ Đức có số lượng chung cư trung bình (500-1.000), trong khi Quận 3, Quận 5, và Hóc Môn ghi nhận rất ít chung cư, dưới 100, do nhu cầu thấp hoặc phát triển chậm.

Về giá trung bình, Quận 1 dẫn đầu với hơn 12 tỷ đồng nhờ vị trí đắc địa và tiện ích phát triển. Quận 2 và Bình Thạnh theo sau với mức trên 8 tỷ đồng, còn các quận nội thành như Quận 3, 10 và 7 dao động từ 5-7 tỷ đồng. Trong khi đó, các quận ngoại thành như Bình Chánh và Hóc Môn có giá trung bình dưới 2 tỷ đồng, phản ánh sự chênh lệch lớn giữa trung tâm và vùng ven.

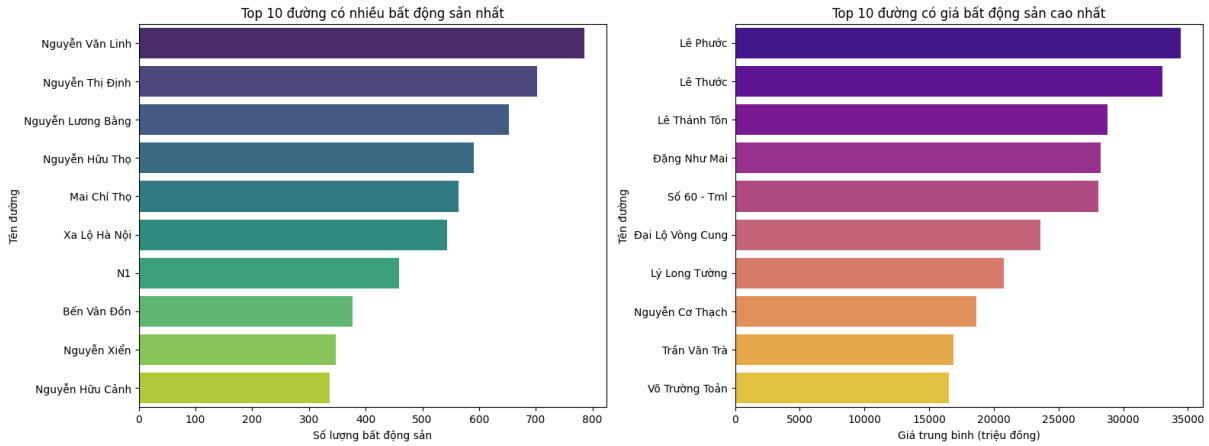


Hình 4. Giá và đơn giá ở từng quận phân bố như thế nào.

Biểu đồ boxplot cho thấy sự chênh lệch rõ rệt về giá chung cư giữa trung tâm và vùng ven TP.HCM. Quận 1 và Quận 2 có mức giá cao nhất, với nhiều outliers vượt 60 tỷ đồng, phản ánh sự phát triển mạnh và sự xuất hiện của các dự án cao cấp. Trong khi đó, các quận như Bình Thạnh, Tân Phú, Thủ Đức và Quận 7 có giá trung bình từ 5-10 tỷ, cân đối giữa tiện ích và giá cả. Các quận ngoại thành như Hóc Môn, Bình Chánh và Gò Vấp ghi nhận median dưới 5 tỷ đồng, với mức giá ổn định và ít biến động.

Giá/m² cũng thể hiện sự chênh lệch lớn, với Quận 1 và Quận 2 ghi nhận mức giá vượt 100 triệu/m², trong khi Hóc Môn và Bình Chánh có mức giá thấp hơn đáng kể. Sự phân hóa giá trong cùng quận, như ở Quận 1 và Quận 7, cho thấy sự tồn tại của các dự án cao cấp và trung cấp. Nhìn chung, giá chung cư tại TP.HCM phản ánh sự phát triển đa dạng giữa trung tâm và vùng ven, đáp ứng nhu cầu của nhiều đối tượng khách hàng.

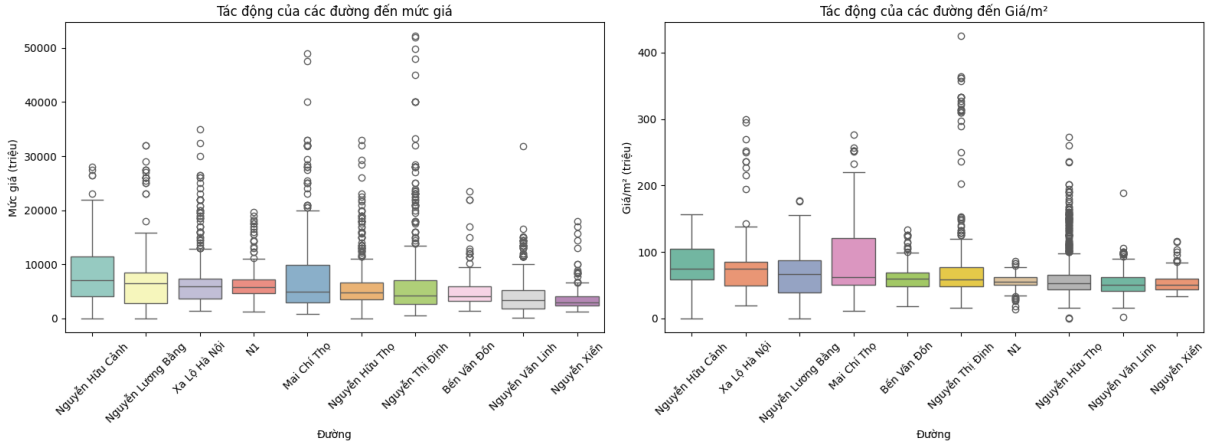
2.4.2. Đường tác động thế nào?



Hình 5. Những con đường có số lượng chung cư nhiều nhất và giá chung cư cao nhất.

Biểu đồ cho thấy Nguyễn Văn Linh là con đường có số lượng bất động sản cao nhất trong top 10, với hơn 700 bất động sản, theo sau là Nguyễn Thị Định, Nguyễn Lương Bằng và Nguyễn Hữu Thọ, mỗi con đường có trên 500 bất động sản. Các tuyến đường chính, có vị trí giao thông thuận lợi và gần các tiện ích, chiếm phần lớn bất động sản, phản ánh sự phát triển không đồng đều trong khu vực.

Về giá trị bất động sản, có sự phân hóa rõ rệt với các con đường như Lê Phước và Lê Thước có giá chung cư trung bình vượt 30 tỷ đồng, trong khi các con đường còn lại có mức giá trên 15 tỷ đồng. Những tuyến đường này đều có giá bất động sản cao, phản ánh vị trí đắc địa và sự hiện diện của nhiều bất động sản cao cấp.



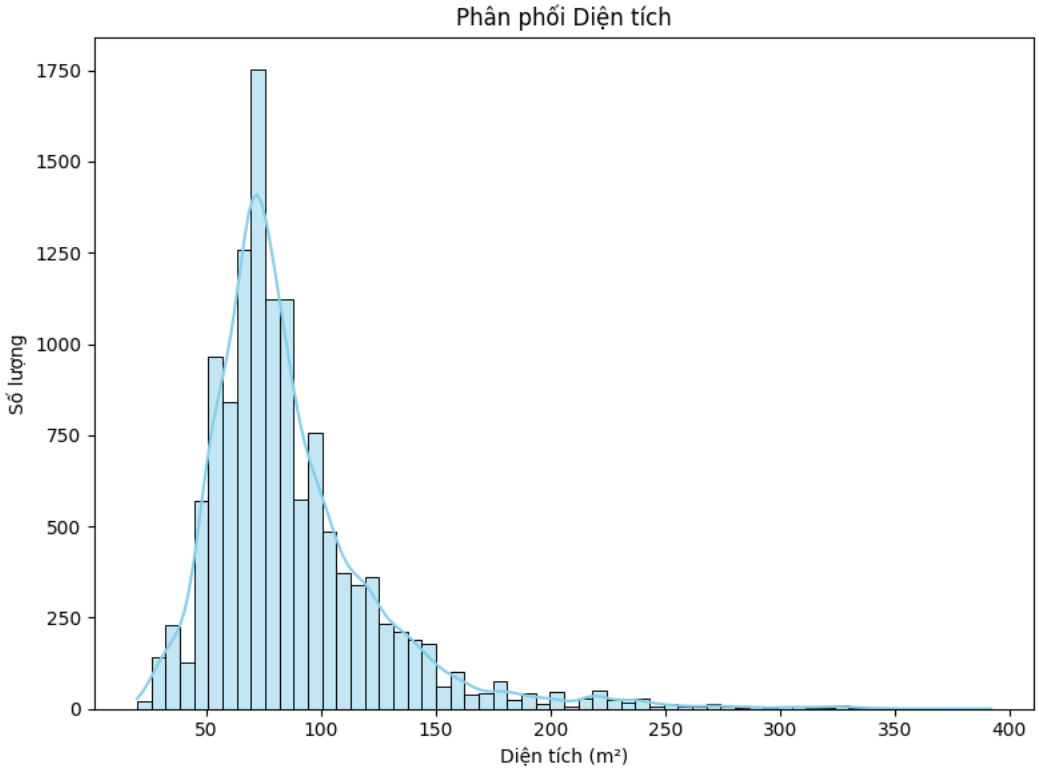
Hình 6. Phân bố giá trên những con đường có đông chung cư nhất.

Biểu đồ boxplot cho thấy sự khác biệt rõ rệt về mức giá chung cư tại TP.HCM. Các tuyến đường như Nguyễn Hữu Cảnh và Nguyễn Lương Bằng có mức giá trung bình cao, với nhiều điểm dữ liệu vượt 20 tỷ, phản ánh sự hiện diện của các dự án cao cấp ở

vị trí đắc địa. Trong khi đó, Nguyễn Văn Linh và Nguyễn Xiển có giá thấp hơn, chủ yếu dưới 10 tỷ, cho thấy các khu vực này đang phát triển với nhiều dự án tầm trung.

Về đơn giá, các tuyến đường như Nguyễn Hữu Cảnh, Xa Lộ Hà Nội và Nguyễn Lương Bằng có đơn giá cao, vượt 100 triệu/m², cho thấy sự tập trung của các dự án cao cấp. Ngược lại, Nguyễn Xiển và Nguyễn Văn Linh có mức giá thấp hơn, dưới 50 triệu/m², phản ánh các dự án trung bình. Biểu đồ cũng cho thấy sự phân tán giá lớn ở một số tuyến đường, với nhiều outliers, phản ánh sự đa dạng trong các loại hình bất động sản.

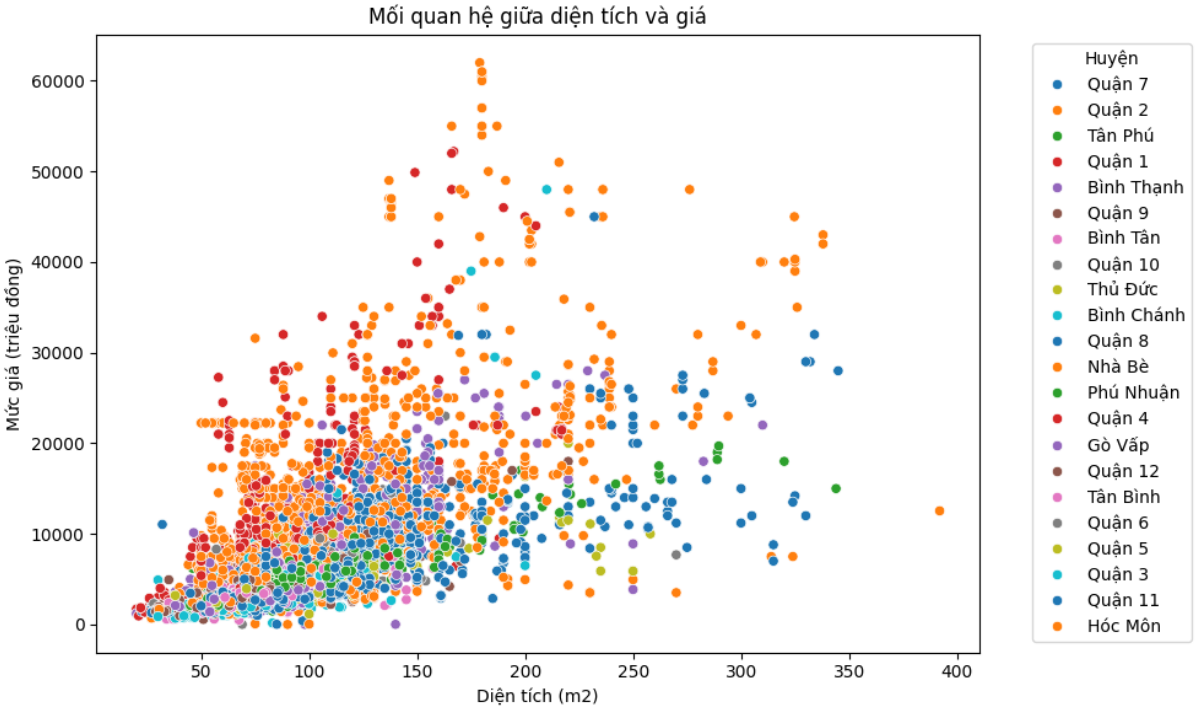
2.5. Tác động của yếu tố diện tích đối với giá và đơn giá của chung cư tại TP.HCM



Hình 7. Phân phối diện tích.

Thị trường bất động sản chủ yếu tập trung vào các căn hộ nhỏ, dưới 100 m², với xu hướng cao nhất ở diện tích 70-80 m², phù hợp với nhu cầu của hộ gia đình trẻ hoặc người độc thân. Mặc dù có một số căn hộ lớn, thường là cao cấp hoặc văn phòng cho thuê, nhưng chúng ít phổ biến hơn.

Điều này phản ánh xu hướng hiện nay là người mua nhà ngày càng quan tâm đến các căn hộ có thiết kế thông minh, tận dụng tối đa không gian, giúp tiết kiệm chi phí sinh hoạt.

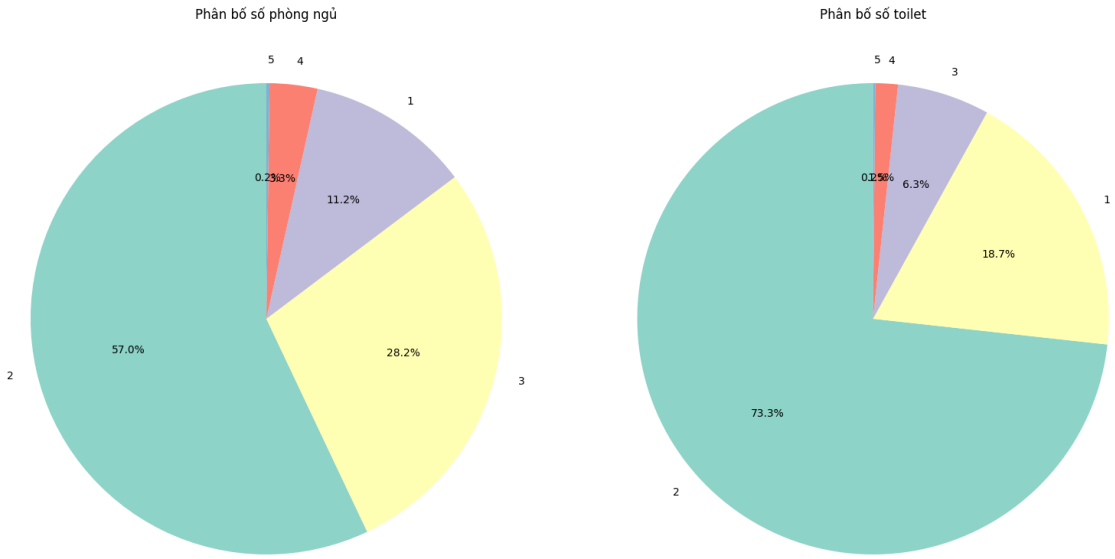


Hình 8. Mối quan hệ giữa diện tích và giá.

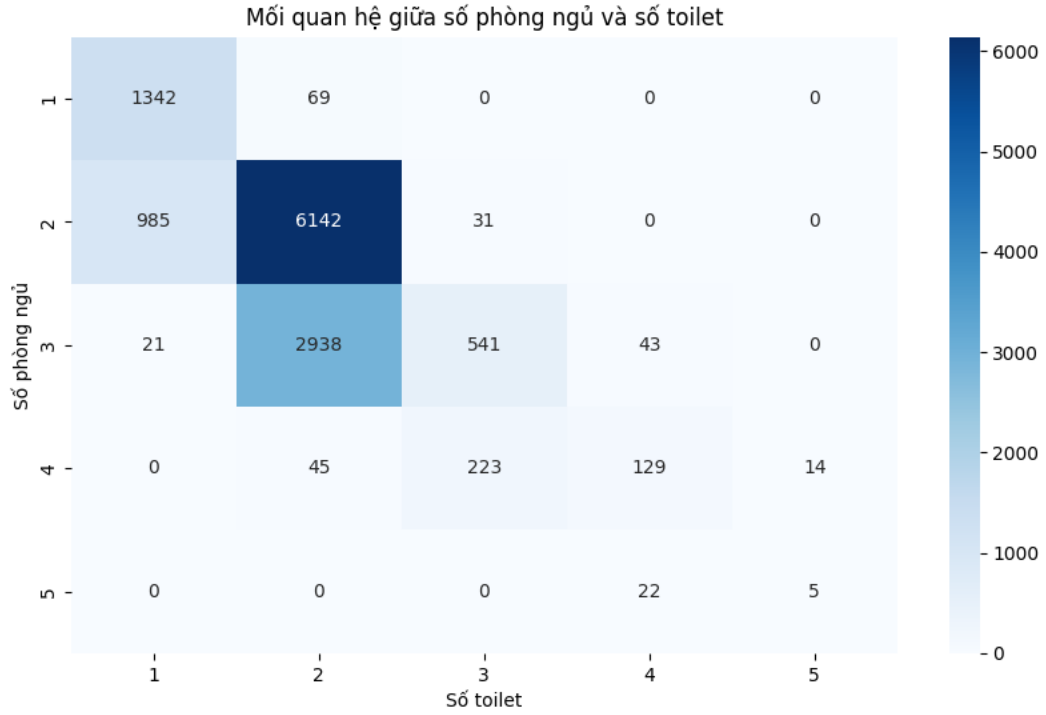
Biểu đồ cho thấy một xu hướng chung là giá nhà tăng khi diện tích tăng, nhưng mối quan hệ này không hoàn toàn tuyến tính, và có sự phân tán lớn giữa các điểm dữ liệu. Điều này cho thấy rằng, mặc dù diện tích có ảnh hưởng trực tiếp đến giá, nhưng nó không phải yếu tố duy nhất quyết định giá trị bất động sản. Các yếu tố khác như vị trí quận, tiện ích xung quanh, chất lượng xây dựng và thời điểm giao dịch cũng đóng vai trò quan trọng trong việc hình thành giá trị căn nhà. Cụ thể, các căn nhà ở các quận trung tâm TP.HCM thường có giá cao hơn so với các quận ngoại thành, ngay cả khi diện tích của chúng tương đương.

Một số điểm dữ liệu nằm ngoài xu hướng chung, cho thấy có những căn nhà có diện tích lớn nhưng giá lại thấp hoặc ngược lại. Điều này có thể được giải thích bởi một số yếu tố khác ngoài diện tích, như tình trạng của căn nhà (cần sửa chữa), vị trí không thuận tiện, hoặc các vấn đề pháp lý liên quan đến bất động sản. Các yếu tố này góp phần làm cho giá trị của bất động sản có sự biến động lớn dù diện tích có vẻ tương đồng.

2.6. Tác động của yếu tố số phòng ngủ và số toilet đối với giá và đơn giá của chung cư tại TP.HCM



Hình 9. Tỷ lệ số phòng ngủ và số toilet.

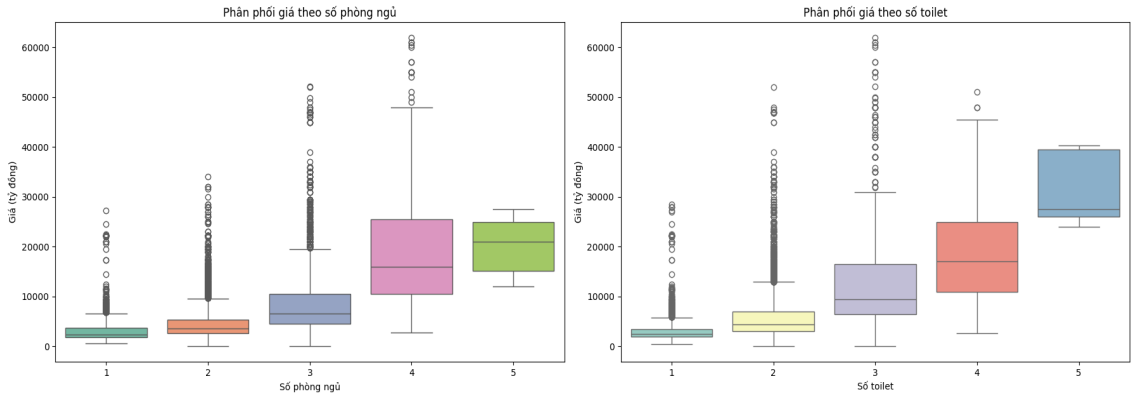


Hình 10. Quan hệ giữa số phòng ngủ và số toilet.

Phần lớn các căn hộ có 2 hoặc 3 phòng ngủ, chiếm 85.2% tổng số căn hộ, phản ánh nhu cầu về không gian sống rộng rãi, đặc biệt là đối với các gia đình trẻ hoặc cặp vợ chồng. Căn hộ 1 phòng ngủ chiếm tỷ lệ nhỏ, cho thấy thị trường đang chú trọng vào phân khúc với diện tích vừa phải. Về phòng vệ sinh, hầu hết các căn hộ có 2 phòng vệ

sinh (73.3%), với căn hộ 1 phòng vệ sinh chiếm tỷ lệ thấp. Nhu cầu về tiện nghi và sự thoải mái cao hơn, khiến khách hàng chú trọng vào số lượng phòng vệ sinh.

Thị trường bất động sản đang tập trung vào căn hộ 2-3 phòng ngủ và 2 phòng vệ sinh, đáp ứng nhu cầu sinh hoạt cơ bản. Mặc dù có sự tương quan giữa số phòng ngủ và số toilet, một số căn hộ có số phòng ngủ lớn nhưng lại chỉ có 1 toilet, điều này có thể do thiết kế đặc biệt hoặc hạn chế diện tích.

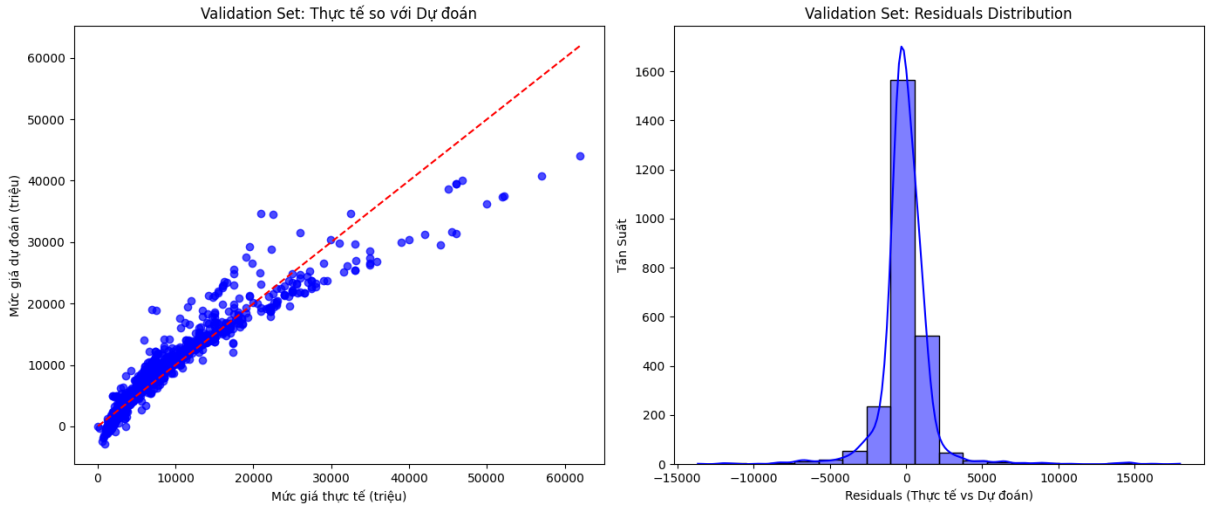


Hình 11. Phân phối giá theo số phòng ngủ và số toilet.

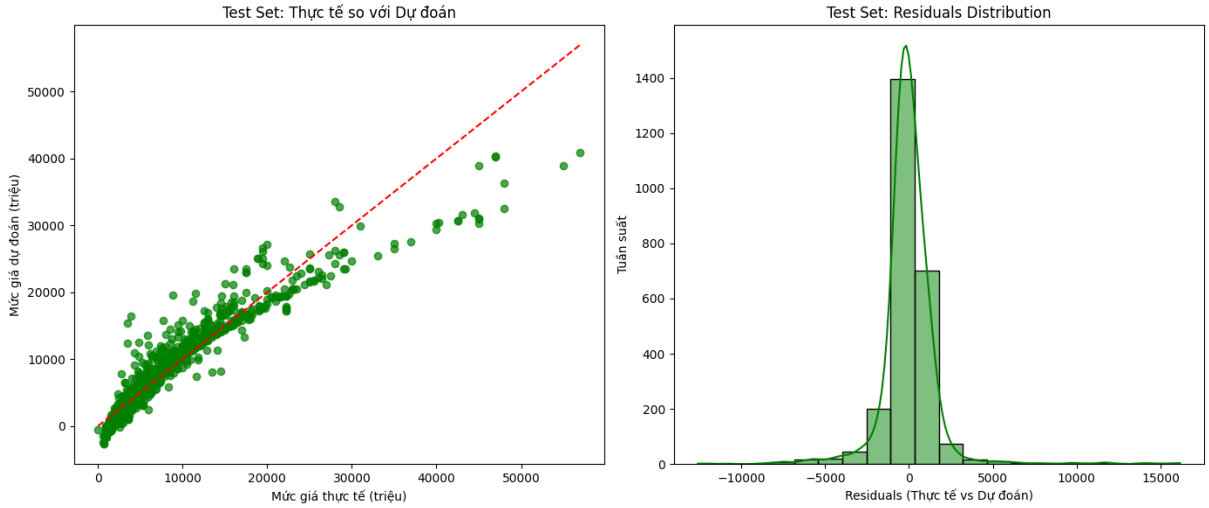
Giá nhà thường tăng theo số lượng phòng ngủ và số toilet, vì căn nhà có nhiều phòng ngủ và toilet thường có diện tích lớn hơn và tiện nghi đầy đủ hơn. Tuy nhiên, sự phân hóa giá vẫn lớn giữa các căn nhà có cùng số lượng phòng ngủ hoặc toilet, do ảnh hưởng của các yếu tố như vị trí, diện tích, hướng nhà, và tiện ích. Một số căn nhà có giá cao hoặc thấp bất thường so với những căn cùng loại, nhờ vào những yếu tố đặc biệt như vị trí đắc địa hoặc thiết kế độc đáo.

2.7. Mô hình dự đoán và đánh giá

Mô hình Linear Regression đạt hiệu suất tốt trên cả tập validation và test. Cụ thể, trên tập validation, MSE là 3,135,178.55, RMSE là 1,770.64, và R^2 là 0.91. Kết quả trên tập test tương tự với MSE 3,024,200.33, RMSE 1,739.02, và R^2 0.91. Điều này cho thấy mô hình ổn định, không bị overfitting hoặc underfitting.



Hình 12. Scatter Plot và Residual Plot trên Validation Set.



Hình 13. Scatter Plot và Residual Plot trên Test Set.

3. KẾT LUẬN

Bộ dữ liệu cung cấp thông tin chi tiết về thị trường chung cư tại TP.HCM, bao gồm giá trị, diện tích, và các đặc điểm liên quan đến từng quận và đường phố. Điều này giúp phân tích sự phân bố và xu hướng giá cả trên từng khu vực, từ đó cung cấp cái nhìn sâu sắc về sự biến động của thị trường bất động sản.

Với bộ dữ liệu này, các nhà đầu tư, môi giới, và người mua có thể đưa ra quyết định sáng suốt hơn khi lựa chọn bất động sản. Nó cũng là nền tảng quan trọng để phát triển các mô hình dự đoán giá trị và sự tăng trưởng trong tương lai, giúp tối ưu hóa chiến lược đầu tư và ra quyết định thị trường.

TÀI LIỆU THAM KHẢO

[1] "Batdongsan.com.vn," [Trực tuyến]. Available: https://batdongsan.com.vn/ban-can-ho-chung-cu-tp-hcm.

[2] "machinelearningmastery," [Trực tuyến]. Available: https://machinelearningmastery.com/navigating-missing-data-challenges-with-xgboost/.

[3] "medium," [Trực tuyến]. Available: https://medium.com/biased-algorithms/dbscan-for-outlier-detection-in-python-d24a9c949a50#:~:text=to%20identify%20them.-,Identifying%20Outliers,t%20belong%20to%20any%20cluster..

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Phòng Lai Bảo Minh	- Viết source code. - Viết báo cáo - Phân tích tổng quan trên dữ liệu. - Tổng hợp các kết quả vào báo cáo.
2	Nguyễn Nhật Hoàng	- Làm slide. - Làm report