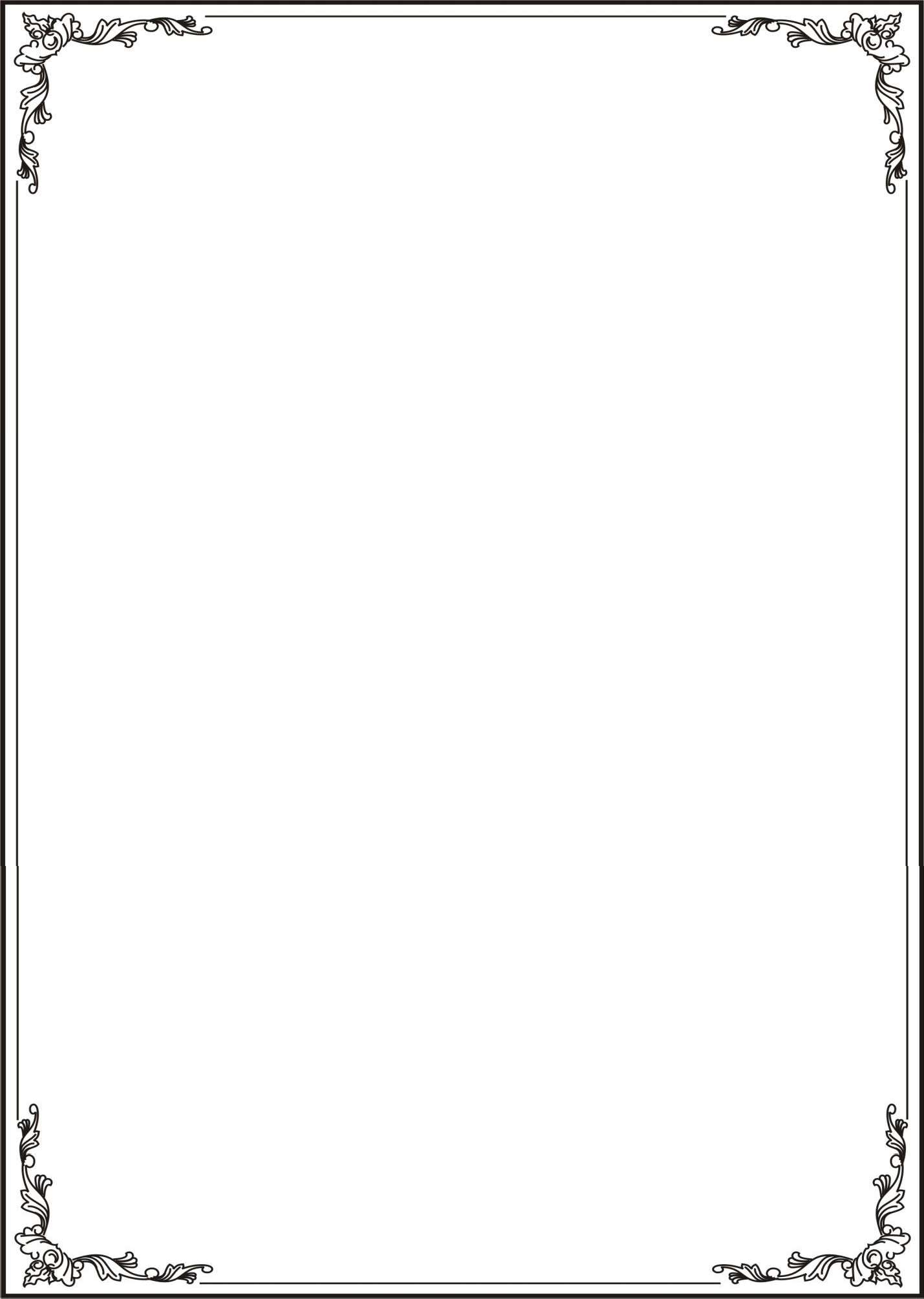
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**



**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**VŨ TRUNG KIÊN: 20133060**

**NGUYỄN VÕ SONG TOÀN: 20133097**

DỰ ĐOÁN GIÁ NHÀ VÀ CHUNG CƯ THÀNH PHỐ HỒ CHÍ MINH VỚI DỮ LIỆU LẤY TỪ NHATOT

**TIỂU LUẬN CHUYÊN NGÀNH**

**NGÀNH KỸ THUẬT DỮ LIỆU**

GIÁO VIÊN HƯỚNG DẪN

**PGS.TS Hoàng Văn Dũng**

**KHÓA 2020 – 2024**

**ĐH SƯ PHẠM KỸ THUẬT TP.HCM CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM**

**KHOA CNTT Độc lập – Tự do – Hạnh Phúc**

**PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

Họ và tên Sinh viên 1: Vũ Trung Kiên MSSV 1: 20133060

Họ và tên Sinh viên 2: Nguyễn Võ Song Toàn MSSV 2: 20133097

Ngành: Kỹ thuật dữ liệu.

Tên đề tài: Dự đoán giá nhà và chung cư thành phố Hồ Chí Minh với dữ liệu từ nhatot.

Họ và tên giáo viên hướng dẫn: PGS.TS Hoàng Văn Dũng

**NHẬN XÉT**

1. Về nội dung đề tài khối lượng thực hiện:

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

2. Ưu điểm và khuyết điểm:

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

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

3. Đánh giá loại:

4. Điểm

Tp*. Hồ Chí Minh, ngày tháng năm 2023*

Giáo viên hướng dẫn

*(Ký & ghi rõ họ tên)*

**PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN**

Họ và tên Sinh viên 1: Vũ Trung Kiên MSSV 1: 20133060

Họ và tên Sinh viên 2: Nguyễn Võ Song Toàn MSSV 2: 20133097

Ngành: Kỹ thuật dữ liệu.

Tên đề tài: Dự đoán giá nhà và chung cư thành phố Hồ Chí Minh với dữ liệu từ nhatot.

Họ và tên giáo viên hướng dẫn: PGS.TS Hoàng Văn Dũng

**NHẬN XÉT**

1. Về nội dung đề tài khối lượng thực hiện:

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

2. Ưu điểm và khuyết điểm:

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

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

3. Đánh giá loại:

4. Điểm:

Tp*. Hồ Chí Minh, ngày tháng năm 2023*

Giáo viên phản biện

*(Ký & ghi rõ họ tên)*

**LỜI CAM ĐOAN**

Chúng em xin cam đoan rằng, khoá luận "Dự đoán giá nhà và chung cư thành phố Hồ Chí Minh với dữ liệu từ nhatot." Là công trình nghiên cứu của chúng em dưới sự hướng dẫn của PGS.TS Hoàng Văn Dũng, xuất phát từ nhu cầu thực tiễn và nguyện vọng tìm hiểu của bản thân chúng em.

Ngoại trừ kết quả tham khảo từ các công trình khác đã ghi rõ trong luận văn, các nội dung trình bày trong khoá luận này là kết quả nghiên cứu do chính chúng em thực hiện và kết quả của khoá luận chưa từng công bố trước đây dưới bất kỳ hình thức nào.

Hồ Chí Minh, ngày ... tháng ... năm 2023

Tác giả

**Vũ Trung Kiên Nguyễn Võ Song Toàn**

**Trường Đại học Sư Phạm Kỹ Thuật TP.HCM**

**Khoa: Công nghệ thông tin**

**ĐỀ CƯƠNG TIỂU LUẬN CHUYÊN NGÀNH**

Họ và tên SV thực hiện: Vũ Trung Kiên MSSV: 20133060

Họ và tên SV thực hiện: Nguyễn Võ Song Toàn MSSV: 20133097

Thời gian làm tiểu luận: Đến:

Chuyên ngành: Kỹ Thuật Dữ Liệu

Tên đề tài: Dự đoán giá nhà và chung cư thành phố Hồ Chí Minh với dữ liệu từ nhatot.

Họ và tên Giáo viên hướng dẫn: PGS.TS Hoàng Văn Dũng

KẾ HOẠCH THỰC HIỆN

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

|  |  |
| --- | --- |
| Ý kiến giáo viên hướng dẫn (Ký & ghi rõ họ tên) | Ngày … tháng … năm 2023 Người viết  Vũ Trung Kiên |
|  | Nguyễn Võ Song Toàn |
|  |  |

# DANH MỤC HÌNH ẢNH

[**Hình 3. 1 Biểu đồ thể hiện High Gamma và Low Gamma** 24](#_Toc153018570)

[**Hình 4. 1 Đoạn mã sử dụng để thu thập danh sách các urls**..................................32](#_Toc153018574)

[**Hình 4. 2 Quá trình thu thập các urls trên website NhaTot** 33](#_Toc153018575)

[**Hình 4. 3 Quá trình thu thập các urls đang tiếp diễn** 34](#_Toc153018576)

[**Hình 4. 4 Kết thúc quá trình thu thập các urls** 34](#_Toc153018577)

[**Hình 4. 5 Kết quả của quá trình thu thập urls sau khi lưu vào file excel** 35](#_Toc153018578)

[**Hình 4. 6 Đoạn mã sử dụng để lấy dữ liệu từ trang NhaTot với các urls đã thu thập phía trên** 35](#_Toc153018579)

[**Hình 4. 7 Mô tả các biến trong tập dữ liệu đã thu thập được (1)** 37](#_Toc153018580)

[**Hình 4. 8 Mô tả các biến trong tập dữ liệu đã thu thập được (2)** 37](#_Toc153018581)

[**Hình 4. 9 Mô tả 5 dòng đầu của dữ liệu** 39](#_Toc153018582)

[**Hình 4. 10 Mô tả về các biến thu thập được và kiểu dữ liệu** 40](#_Toc153018583)

[**Hình 4. 11 Số lượng các biến bị null trong tập dữ liệu** 41](#_Toc153018584)

[**Hình 4. 12 Chuẩn hóa dữ liệu DienTich, PhongTam, PhongNgu về dạng số** 42](#_Toc153018585)

[**Hình 4. 13 Xử lý các kiểu dữ liệu dạng chuỗi đưa về dạng số của các biến Gia/m2 và Gia** 43](#_Toc153018586)

[**Hình 4. 14 Xử lý dữ liệu Quan** 44](#_Toc153018587)

[**Hình 4. 15 Dữ liệu của Quan sau khi được xử lý** 45](#_Toc153018588)

[**Hình 4. 16 Xử lý các ngoại lệ (outlier)** 45](#_Toc153018589)

[**Hình 4. 17 Dữ liệu sau khi đã được tiền xử lý** 46](#_Toc153018590)

[**Hình 4. 18 Đoạn mã thực hiện việc trực quan hóa biến số phân loại (categorical variables)** 46](#_Toc153018591)

[**Hình 4. 19 Trực quan hóa biến số phân loại (categorical variables)** 47](#_Toc153018592)

[**Hình 4. 20 Đoạn mã thực hiện việc trực quan hóa biến số liên tục (continous variable):** 47](#_Toc153018593)

[**Hình 4. 21 Trực quan hóa biến số liên tục (continous variable)** 48](#_Toc153018594)

[**Hình 4. 22 Trực quan hóa các biến số liên tục với USD** 49](#_Toc153018595)

[**Hình 4. 23 Trực quan hóa Quan và USD với biểu đồ histogram sắp xếp theo USD** 49](#_Toc153018596)

[**Hình 4. 24 Danh sách Quan với USD giảm dần** 49](#_Toc153018597)

[**Hình 5. 1 Khai báo thư viện cần thiết......................................................................**50](#_Toc153019025)

[**Hình 5. 2 Đọc tập dữ liệu sau khi đã xử lý** 51](#_Toc153019026)

[**Hình 5. 3 Tỉ lệ NaN của từng biến và không sử dụng các biến NaN lớn hơn 50 phần trăm** 52](#_Toc153019027)

[**Hình 5. 4 Mô hình sau khi áp dụng các biến giả** 53](#_Toc153019028)

[**Hình 5. 5 Xây dựng mô hình hồi quy** 54](#_Toc153019029)

[**Hình 6. 1 Dự đoán kết quả với tập test**....................................................................55](#_Toc153019030)

[**Hình 6. 2 Giá dự đoán và giá thực sau khi áp dụng mô hình hồi quy** 56](#_Toc153019031)

[**Hình 6. 3 Sai lệch trung bình trên các phân vị** 56](#_Toc153019032)

[**Hình 6. 4 Giá thực tế và các mô tả của đặc điểm bất động sản** 58](#_Toc153019033)

[**Hình 6. 5 Dự đoán kết quả với các đặc điểm từ website NhaTot** 59](#_Toc153019034)

[**Hình 6. 6 Đánh giá mô hình và phương trình hồi quy** 60](#_Toc153019035)

[**Hình 6. 7 Kiểm định giả thuyết với mức ý nghĩa 5%** 61](#_Toc153019036)

[**Hình 6. 8 Kết quả p\_value của giả thuyết** 62](#_Toc153019037)

[DANH MỤC HÌNH ẢNH 7](#_Toc153019268)

[PHẦN MỞ ĐẦU 11](#_Toc153019269)

[1. TÍNH CẤP THIẾT CỦA ĐỀ TÀI 11](#_Toc153019270)

[2. MỤC ĐÍCH CỦA ĐỀ TÀI 11](#_Toc153019271)

[3. CÁCH TIẾP CẬN VÀ PHƯƠNG PHÁP NGHIÊN CỨU 12](#_Toc153019272)

[3.1. Đối tượng nghiên cứu 12](#_Toc153019273)

[3.2. Phạm vi nghiên cứu 12](#_Toc153019274)

[3.3. Cách tiếp cận và phương pháp nghiên cứu 12](#_Toc153019275)

[4. KẾT QUẢ VÀ DỰ KIẾN ĐẠT ĐƯỢC 13](#_Toc153019276)

[5. BỐ CỤC BÀI BÁO CÁO 13](#_Toc153019277)

[PHẦN NỘI DUNG 14](#_Toc153019278)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN 14](#_Toc153019279)

[1.1. Những vấn đề cơ bản về bất động sản 14](#_Toc153019280)

[1.2. Các nghiên cứu liên quan 16](#_Toc153019281)

[CHƯƠNG 2: TỔNG QUAN VỀ THỊ TRƯỜNG BẤT ĐỘNG SẢN TẠI THÀNH PHỐ HỒ CHÍ MINH 17](#_Toc153019282)

[2.1. Giới thiệu 17](#_Toc153019283)

[2.2. Phát triển kinh tế và tăng trưởng dân số 18](#_Toc153019284)

[2.3. Tình hình cung cầu bất động sản 19](#_Toc153019285)

[2.4. Các yếu tố ảnh hưởng đến giá nhà và chung cư 19](#_Toc153019286)

[CHƯƠNG 3. CÁC PHƯƠNG PHÁP DỰ ĐOÁN GIÁ NHÀ VÀ CHUNG CƯ 20](#_Toc153019287)

[3.1. Phân tích hồi quy tuyến tính 20](#_Toc153019288)

[3.2. Sử dụng phương pháp máy vector hỗ trợ (SVM) 23](#_Toc153019289)

[3.3. Mạng nơ-ron nhân tạo 27](#_Toc153019290)

[3.4. Các phương pháp khác 29](#_Toc153019291)

[CHƯƠNG 4. THU THẬP DỮ LIỆU VÀ XỬ LÝ DỮ LIỆU 31](#_Toc153019292)

[4.1. Nguồn dữ liệu 31](#_Toc153019293)

[4.2. Phương pháp thu thập dữ liệu 31](#_Toc153019294)

[4.3. Tiền xử lý dữ liệu 38](#_Toc153019295)

[4.4. Trực quan hóa dữ liệu 46](#_Toc153019296)

[CHƯƠNG 5. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN GIÁ NHÀ VÀ CHUNG CƯ 50](#_Toc153019297)

[5.1. Mô tả mô hình 50](#_Toc153019298)

[5.2. Huấn luyện mô hình 53](#_Toc153019299)

[Chương 6. DỰ ĐOÁN GIÁ NHÀ VÀ CHUNG CƯ TẠI THÀNH PHỐ HỒ CHÍ MINH 55](#_Toc153019300)

[6.1. Kết quả dự đoán 55](#_Toc153019301)

[6.2. So sánh với giá thực tế 57](#_Toc153019302)

[6.3. Đánh giá hiệu suất và độ chính xác của mô hình 60](#_Toc153019303)

[Chương 7. THẢO LUẬN VÀ KẾT LUẬN 62](#_Toc153019304)

[7.1. Thảo luận kết quả 62](#_Toc153019305)

[7.2. Đóng góp của nghiên cứu 63](#_Toc153019306)

[7.3. Hạn chế và hướng phát triển tiếp theo 64](#_Toc153019307)

[TÀI LIỆU THAM KHẢO 65](#_Toc153019308)

# PHẦN MỞ ĐẦU

## 1. TÍNH CẤP THIẾT CỦA ĐỀ TÀI

Đáp ứng nhu cầu thông tin của người dùng: Với sự phát triển của công nghệ thông tin, người dùng ngày càng tìm kiếm thông tin trực tuyến để tìm hiểu về giá cả và xu hướng bất động sản. Dự đoán giá nhà và chung cư từ dữ liệu từ nhatot đáp ứng nhu cầu này bằng cách cung cấp thông tin chi tiết và chính xác về giá cả, giúp người dùng có cái nhìn rõ ràng và đáng tin cậy về thị trường bất động sản tại thành phố Hồ Chí Minh.

Đóng góp vào lĩnh vực nghiên cứu bất động sản: Việc dự đoán giá nhà và chung cư tại thành phố Hồ Chí Minh có ý nghĩa quan trọng trong lĩnh vực bất động sản. Đề tài cung cấp phương pháp và kỹ thuật dự đoán giá cả nhanh chóng và chính xác, đáng tin cậy cho các nhà đầu tư, nhà phát triển và người mua bất động sản. Điều này giúp cải thiện quyết định đầu tư và giao dịch bất động sản, tạo ra lợi ích kinh tế và phát triển bền vững cho thành phố Hồ Chí Minh.

Tiềm năng phát triển và ứng dụng cho các nghiên cứu tương lai: Đề tài cung cấp một cơ sở nền tảng cho các nghiên cứu liên quan đến dự đoán giá nhà và chung cư. Các phương pháp và kỹ thuật được áp dụng trong đề tài có thể được mở rộng và tùy chỉnh cho các lĩnh vực và thị trường bất động sản khác. Điều này mở ra tiềm năng phát triển cho các nghiên cứu tương lai trong lĩnh vực dự đoán giá nhà và chung cư, góp phần nâng cao hiệu quả và chính xác của các phương pháp dự đoán.

## 2. MỤC ĐÍCH CỦA ĐỀ TÀI

Nắm bắt xu hướng và biến động giá cả: Mục đích chính của đề tài là phân tích và dự đoán giá nhà và chung cư tại thành phố Hồ Chí Minh dựa trên dữ liệu từ nhatot. Qua đó, nghiên cứu nhằm mục đích cung cấp thông tin cụ thể về xu hướng và biến động giá cả trong thị trường bất động sản. Điều này giúp người đầu tư và người mua bất động sản có cái nhìn tổng quan về tình hình thị trường và đưa ra quyết định mua bán dựa trên thông tin đáng tin cậy.

Cải thiện quyết định đầu tư và giao dịch bất động sản: Bằng cách áp dụng các phương pháp dự đoán giá cả, đề tài nhằm cung cấp một công cụ hữu ích để đánh giá và dự báo giá trị tài sản bất động sản tại thành phố Hồ Chí Minh. Điều này giúp cải thiện quyết định đầu tư và giao dịch của các nhà đầu tư và người mua, từ đó tối ưu hóa lợi nhuận và giảm rủi ro trong các hoạt động liên quan đến bất động sản.

Đóng góp vào phát triển bền vững của thị trường bất động sản: Việc dự đoán giá nhà và chung cư tại thành phố Hồ Chí Minh từ dữ liệu từ nhatot có ý nghĩa quan trọng trong việc xác định xu hướng và tăng trưởng của thị trường bất động sản. Đề tài đóng góp vào việc hiểu rõ hơn về cung cầu, yếu tố ảnh hưởng đến giá cả, và sự biến động của thị trường bất động sản. Điều này hỗ trợ cho quy hoạch và phát triển bền vững của thành phố Hồ Chí Minh, từ việc quản lý hợp lý cung cầu bất động sản đến đáp ứng nhu cầu nhà ở của người dân.

## 3. CÁCH TIẾP CẬN VÀ PHƯƠNG PHÁP NGHIÊN CỨU

### 3.1. Đối tượng nghiên cứu

Đối tượng chính của nghiên cứu là thị trường bất động sản thành phố Hồ Chí Minh. Nghiên cứu sẽ tập trung vào việc dự đoán giá nhà và chung cư, sử dụng dữ liệu từ nhatot. Các dữ liệu về giá cả, diện tích, vị trí địa lý, tiện ích xung quanh và các yếu tố khác liên quan sẽ được thu thập và phân tích để xây dựng mô hình dự đoán.

### 3.2. Phạm vi nghiên cứu

Phạm vi của nghiên cứu sẽ tập trung vào thành phố Hồ Chí Minh, một thị trường bất động sản lớn và phát triển tại Việt Nam. Dữ liệu về giá nhà và chung cư sẽ được thu thập từ nhatot, một nguồn thông tin uy tín về thị trường bất động sản. Nghiên cứu sẽ tập trung vào việc xây dựng mô hình dự đoán giá cả dựa trên các yếu tố quan trọng như diện tích, vị trí địa lý, tiện ích và các yếu tố khác có thể ảnh hưởng đến giá trị tài sản bất động sản.

### 3.3. Cách tiếp cận và phương pháp nghiên cứu

Để dự đoán giá nhà và chung cư thành phố Hồ Chí Minh, nghiên cứu sẽ sử dụng các phương pháp và kỹ thuật trong lĩnh vực học máy và khai phá dữ liệu. Cụ thể, một phương pháp dự đoán có thể là sử dụng mô hình hồi quy, trong đó dữ liệu về các yếu tố như diện tích, vị trí địa lý và các tiện ích xung quanh được sử dụng để đưa ra dự đoán về giá cả bất động sản.

Ngoài ra, để cải thiện độ chính xác của mô hình dự đoán, nghiên cứu cũng có thể sử dụng các phương pháp phân tích đa biến và kỹ thuật học máy phức tạp khác như mạng nơ-ron nhân tạo (Artificial Neural Networks), máy vector hỗ trợ (Support Vector Machines) hoặc các thuật toán tăng cường (Boosting algorithms).

Tổng hợp lại, cách tiếp cận và phương pháp nghiên cứu trong đề tài "Dự đoán giá nhà và chung cư thành phố Hồ Chí Minh với dữ liệu từ nhatot" sẽ sử dụng các phương pháp học máy và khai phá dữ liệu để xây dựng mô hình dự đoán giá cả bất động sản. Nghiên cứu sẽ tập trung vào thành phố Hồ Chí Minh và sử dụng dữ liệu từ nhatot làm nguồn thông tin chính.

## 4. KẾT QUẢ VÀ DỰ KIẾN ĐẠT ĐƯỢC

Kết quả dự kiến của đề tài "Dự đoán giá nhà và chung cư thành phố Hồ Chí Minh với dữ liệu từ nhatot" là xây dựng một mô hình dự đoán giá cả chính xác cho thị trường bất động sản thành phố Hồ Chí Minh. Mô hình này sẽ cung cấp thông tin hữu ích cho quyết định đầu tư và đóng góp vào phát triển bền vững của thị trường bất động sản.

Bằng cách sử dụng các phương pháp như hồi quy tuyến tính mô hình hóa thống kê, trí tuệ nhân tạo và ứng dụng công nghệ blockchain, đề tài này sẽ cung cấp một cách tiếp cận toàn diện và đa chiều để dự đoán giá nhà và chung cư. Mô hình dự đoán được xây dựng dựa trên dữ liệu từ nhatot sẽ giúp hiểu rõ hơn về xu hướng và biến động của thị trường bất động sản trong thành phố Hồ Chí Minh.

Khi có một mô hình dự đoán giá nhà chính xác, các nhà đầu tư và người mua sẽ có được thông tin quan trọng để đưa ra quyết định đầu tư thông minh. Ngoài ra, việc dự đoán giá nhà cũng đóng góp vào phát triển bền vững của thị trường bất động sản bằng cách giúp tăng cường minh bạch và giảm thiểu rủi ro trong giao dịch.

Tóm lại, kết quả dự kiến của đề tài này là xây dựng một mô hình dự đoán giá cả chính xác, cung cấp thông tin hữu ích cho quyết định đầu tư và đóng góp vào phát triển bền vững của thị trường bất động sản.

## 5. BỐ CỤC BÀI BÁO CÁO

# PHẦN NỘI DUNG

## CHƯƠNG 1: CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

### 1.1. Những vấn đề cơ bản về bất động sản

#### 1.1.1. Khái niệm bất động sản

Bất động sản là loại hình tài sản có giá trị kinh tế và có tính định hình cao. Bất động sản bao gồm tất cả các loại tài sản vật chất có thể được sở hữu như đất đai, nhà cửa, căn hộ chung cư, trung tâm thương mại, khu công nghiệp, khu nghỉ dưỡng...

Điểm đặc trưng của bất động sản là sự cố định về vị trí, không thể di chuyển được theo chủ sở hữu. Đất đai, nhà cửa luôn nằm tại một địa chỉ cụ thể. Giá trị của bất động sản phụ thuộc vào nhiều yếu tố như vị trí địa lý, quy mô, chất lượng xây dựng, tiện ích xung quanh, cơ sở hạ tầng...

Bất động sản là một loại tài sản quan trọng có thể cho thuê, thế chấp, mua bán để sinh lời. Ngành bất động sản cung cấp nhiều dịch vụ liên quan đến phát triển dự án, môi giới, quản lý, vận hành...

#### 1.1.2. Phân loại bất động sản

Bất động sản là một loại tài sản quan trọng, có giá trị lớn và đa dạng về loại hình. Việc phân loại bất động sản giúp hiểu rõ hơn các loại hình bất động sản cũng như đặc điểm riêng của từng loại. Bất động sản có thể được phân loại theo nhiều tiêu chí sau:

* Phân loại theo loại hình xây dựng
* Nhà phố, biệt thự, liền kề: Loại bất động sản nhà ở độc lập dành cho một hộ gia đình.
* Chung cư: Loại hình bất động sản xây dựng cao tầng với nhiều căn hộ trong một tòa nhà.
* Khu đô thị, khu căn hộ: Quy hoạch phức hợp nhiều chức năng trong cùng một khu vực.
* Phân loại theo mục đích sử dụng
* Nhà ở: Gồm các loại nhà riêng lẻ, chung cư phục vụ nhu cầu ở.
* Thương mại dịch vụ: Trung tâm thương mại, trung tâm giải trí, khách sạn...
* Công nghiệp: Khu công nghiệp, nhà xưởng, kho bãi
* Văn phòng: Toà nhà văn phòng, tòa nhà hành chính.
* Các tiêu chí phân loại khác
* Theo khu vực địa lý: Nội đô, ngoại ô, ven biển, miền núi...
* Theo chủ sở hữu: Nhà nước, doanh nghiệp, cá nhân.
* Theo mức độ sang trọng: Cao cấp, trung cấp, bình dân.

#### 1.1.3. Khả năng đầu tư và lợi nhuận

Trong thế giới ngày nay, sự hiểu biết về khả năng đầu tư và lợi nhuận của bất động sản là quan trọng. Dự đoán chính xác giá nhà có thể giúp nhà đầu tư ra quyết định mua bán, giữ hoặc chuyển nhượng tài sản của họ một cách linh hoạt và chiến lược.

Khả năng đầu tư trong bất động sản liên quan đến khả năng xác định và đánh giá tiềm năng sinh lợi của một tài sản. Khi xem xét một tài sản bất động sản như một cơ hội đầu tư, các nhà đầu tư thường xem xét các yếu tố như vị trí, quy hoạch, tiềm năng phát triển, cơ sở hạ tầng, và xu hướng thị trường.

Vị trí là một yếu tố quan trọng trong khả năng đầu tư. Một tài sản có vị trí thuận lợi, gần các tiện ích công cộng, trung tâm thương mại, trường học, bệnh viện, và giao thông thuận tiện thường có khả năng tăng giá cao hơn trong tương lai. Quy hoạch cũng đóng vai trò quan trọng, vì những khu vực có quy hoạch phát triển dự án hạ tầng hoặc cải tạo có thể trở thành điểm đầu tư hấp dẫn.

Các yếu tố khác như tiềm năng phát triển và cơ sở hạ tầng cũng có thể ảnh hưởng đến khả năng đầu tư. Một khu vực có tiềm năng phát triển mạnh mẽ, với sự gia tăng của các dự án công nghiệp, thương mại, hoặc dân cư, có thể tạo ra cơ hội sinh lợi lớn cho nhà đầu tư. Ngoài ra, việc có cơ sở hạ tầng phát triển tốt như hệ thống giao thông, điện, nước, viễn thông cũng là một yếu tố quan trọng để đảm bảo khả năng đầu tư thành công.

Dự đoán lợi nhuận trong bất động sản là một yếu tố then chốt trong quá trình ra quyết định đầu tư. Lợi nhuận có thể đến từ hai nguồn chính: tăng giá và thu nhập.

Tăng giá là khả năng tài sản bất động sản tăng giá theo thời gian. Điều này có thể xảy ra do nhiều yếu tố, bao gồm tăng trưởng kinh tế, tăng cầu, giảm cung, phát triển khu vực, hoặc cải thiện hạ tầng. Nhà đầu tư có thể mua một tài sản và giữ nó trong thời gian dài, hy vọng rằng giá trị của nó sẽ tăng lên và họ có thể bán nó với lợi nhuận.

Thu nhập là một nguồn lợi nhuận khác trong bất động sản. Nhà đầu tư có thể thu được thu nhập từ việc cho thuê tài sản bất động sản, chẳng hạn như cănhộ, văn phòng, hoặc cửa hàng. Thu nhập này có thể đến từ việc thu tiền thuê hàng tháng hoặc hàng năm, tùy thuộc vào thỏa thuận thuê.

Việc dự đoán lợi nhuận trong bất động sản không chỉ liên quan đến việc đánh giá tiềm năng tăng giá và thu nhập hiện tại, mà cũng đòi hỏi sự nắm bắt và đánh giá các yếu tố ngoại vi.

### 1.2. Các nghiên cứu liên quan

Trong phần này, chúng ta sẽ tìm hiểu về những nghiên cứu quan trọng liên quan đến dự đoán giá nhà. Các nghiên cứu này không chỉ mang đến cái nhìn sâu sắc về các phương pháp dự đoán mà còn khám phá sự tích hợp của công nghệ mới như trí tuệ nhân tạo trong lĩnh vực này.

#### 1.2.1. Mô hình hóa thống kê

Một số nghiên cứu tiếp cận vấn đề thông qua mô hình hóa thống kê để định rõ sự biến động của giá nhà theo thời gian. Các mô hình này không chỉ đưa ra dự đoán mà còn giải thích được những xu hướng và biến động trong thị trường bất động sản. Ví dụ, các mô hình ARIMA (Autoregressive Integrated Moving Average) và SARIMA (Seasonal ARIMA) đã được áp dụng để dự đoán giá nhà dựa trên các yếu tố thời gian.

#### 1.2.2. Sử dụng trí tuệ nhân tạo

Trong một số nghiên cứu, trí tuệ nhân tạo, đặc biệt là máy học và học máy, đóng một vai trò quan trọng trong việc xây dựng các mô hình dự đoán giá nhà. Các thuật toán như hồi quy tuyến tính, SVM (Support Vector Machines), mạng nơ-ron, và các phương pháp học sâu như mạng nơ-ron gia đình CNN (Convolutional Neural Networks) và RNN (Recurrent Neural Networks) đã được áp dụng để tạo ra các mô hình dự đoán chính xác hơn và hiệu quả hơn.

#### 1.2.3. Ứng dụng công nghệ Blockchain

Nhiều nghiên cứu cũng tìm hiểu về cách sử dụng công nghệ blockchain để tăng cường minh bạch và tính an toàn trong các giao dịch bất động sản. Với blockchain, thông tin về giao dịch, quyền sở hữu và lịch sử bất động sản có thể được lưu trữ một cách an toàn và không thể sửa đổi. Điều này giúp tăng cường tính minh bạch và giảm thiểu rủi ro trong các giao dịch bất động sản.

Qua sự kết hợp giữa cơ sở lý thuyết và những nghiên cứu tiên tiến, chương này sẽ là cơ sở cho quá trình xây dựng mô hình dự đoán giá nhà, một chủ đề quan trọng trong phạm vi đề tài nghiên cứu. Việc áp dụng những kiến thức và kết quả từ các nghiên cứu liên quan sẽ giúp cải thiện hiệu suất của mô hình dự đoán và đưa ra những quyết định thông minh trong lĩnh vực bất động sản.

## CHƯƠNG 2: TỔNG QUAN VỀ THỊ TRƯỜNG BẤT ĐỘNG SẢN TẠI THÀNH PHỐ HỒ CHÍ MINH

### 2.1. Giới thiệu

Thị trường bất động sản Thành phố Hồ Chí Minh đang dần phục hồi sau ảnh hưởng của đại dịch Covid-19

Theo nhận định của các chuyên gia ngành bất động sản, thị trường nhà đất tại TP. Hồ Chí Minh đang dần có dấu hiệu hồi phục sau khoảng thời gian trầm lắng do tác động của đại dịch. Cụ thể, giá bất động sản đang dần ổn định trở lại sau những biến động mạnh trước đây. Đồng thời, lượng giao dịch mua bán, chuyển nhượng cũng ghi nhận xu hướng tăng trở lại so với thời điểm cao điểm của dịch bệnh.

Nguyên nhân được cho là do Chính phủ đã triển khai nhiều chính sách hỗ trợ thiết thực nhằm thúc đẩy phục hồi kinh tế và thị trường bất động sản. Cụ thể, nhiều dự án bất động sản lớn đã được gỡ khó về pháp lý và được phép huy động vốn từ nhà đầu tư. Đồng thời, lãi suất huy động vốn tại các ngân hàng thương mại cũng giảm mạnh, tạo đà cho nhu cầu mua sắm nhà đất trở lại.

Tuy nhiên, tiêu thụ sản phẩm vẫn chưa được như kỳ vọng do sức mua của người dân bị suy giảm sau đại dịch. Dự báo trong thời gian tới, khi lãi suất giảm thấp hơn nữa và tình hình dịch bệnh tiếp tục được kiểm soát, thị trường bất động sản có thể phục hồi mạnh mẽ hơn.

### 2.2. Phát triển kinh tế và tăng trưởng dân số

Trong chương này, chúng ta sẽ tìm hiểu về tình hình phát triển kinh tế và tăng trưởng dân số tại Thành phố Hồ Chí Minh, Việt Nam.

#### 2.2.1. Phát triển kinh tế

Thành phố Hồ Chí Minh là trung tâm kinh tế, tài chính và thương mại hàng đầu của Việt Nam. Trong những năm gần đây, thành phố này đã chứng kiến một sự phát triển kinh tế mạnh mẽ. Kinh tế TP.HCM được xây dựng trên nền tảng công nghiệp đa dạng, với sự đóng góp lớn từ các ngành công nghệ thông tin, dịch vụ tài chính, du lịch, đô thị hóa và sản xuất công nghiệp.

Sự phát triển kinh tế tại TP.HCM đã tạo ra nhiều cơ hội đầu tư trong lĩnh vực bất động sản. Các khu công nghiệp, khu đô thị mới, khu dân cư và các dự án thương mại đã và đang được xây dựng để đáp ứng nhu cầu về nhà ở, văn phòng, trung tâm thương mại và cơ sở hạ tầng.

#### 2.2.2. Tăng trưởng dân số

Thành phố Hồ Chí Minh cũng đang chứng kiến sự tăng trưởng đáng kể về dân số. Với vị trí là trung tâm của kinh tế và cung cấp nhiều cơ hội việc làm, thành phố thu hút một lượng lớn người di cư từ các tỉnh lân cận và các khu vực khác trong nước. Sự tăng trưởng dân số đã tạo ra nhu cầu cao về nhà ở và dịch vụ bất động sản.

Đồng thời, TP.HCM cũng trở thành điểm đến thu hút đầu tư từ nước ngoài và người nước ngoài định cư. Sự phát triển của các khu công nghiệp, khu thương mại, trung tâm tài chính và các dự án bất động sản cao cấp đã tạo ra nhiều cơ hội cho các nhà đầu tư nước ngoài.

Sự tăng trưởng dân số và sự phát triển kinh tế tại TP.HCM đã góp phần làm tăng nhu cầu về bất động sản, đặc biệt là nhà ở, văn phòng và trung tâm thương mại. Điều này đã tạo ra cơ hội đầu tư hấp dẫn trong lĩnh vực bất động sản tại thành phố này.

Tuy nhiên, cần lưu ý rằng sự phát triển kinh tế và tăng trưởng dân số cũng đặt ra những thách thức cho thành phố Hồ Chí Minh, bao gồm áp lực về nhu cầu nhà ở, giao thông, cơ sở hạ tầng và môi trường sống. Việc quản lý và phát triển bất động sản cần được thực hiện một cách bền vững và có kế hoạch để đáp ứng nhu cầu của dân cư và đảm bảo sự phát triển hàng lâu dài của thành phố.

### 2.3. Tình hình cung cầu bất động sản

Về tình hình cung cầu trên thị trường bất động sản TP. HCM trong 9 tháng đầu năm 2023:

Về nguồn cung, theo khảo sát của Savills Việt Nam, trong 9 tháng đầu năm 2023, tổng lượng sản phẩm mới đưa ra thị trường ước đạt khoảng 30.000 căn (Savills Việt Nam, 2023). So với cùng kỳ năm ngoái, nguồn cung bất động sản tăng khoảng 5-10%. Tuy nhiên, phần lớn là các dự án đang triển khai từ trước tiếp tục đưa ra thị trường, số lượng dự án mới được phát triển còn hạn chế.

Về nhu cầu, theo báo cáo thị trường quý III/2023 của DKRA Vietnam, lượng giao dịch bất động sản trong 9 tháng đầu năm đạt khoảng 25.000 giao dịch, tăng 15% so với cùng kỳ năm trước (DKRA Vietnam, 2023). Nhu cầu tập trung chủ yếu vào phân khúc căn hộ trung và bình dân có giá hợp lý tại khu vực trung tâm.

Từ đó, có thể thấy thị trường đang có tình trạng cung vượt cầu ở phân khúc cao cấp, trong khi phân khúc trung và bình dân vẫn còn thiếu hụt nguồn cung để đáp ứng nhu cầu ngày một tăng cao của người dân. Giá bán bất động sản không có biến động đáng kể.

### 2.4. Các yếu tố ảnh hưởng đến giá nhà và chung cư

Giá bất động sản tại TP. Hồ Chí Minh bị ảnh hưởng bởi nhiều yếu tố như:

* Vị trí dự án: Theo nghiên cứu của Jones Lang LaSalle (2023), giá bất động sản ở khu vực trung tâm, gần khu công nghiệp thường cao hơn so với khu vực ngoại ô hoặc vùng ven.
* Trạng thái pháp lý: Có nghiên cứu cho thấy, bất động sản có sổ đỏ, hồ sơ pháp lý rõ ràng sẽ được định giá cao hơn (Savills Việt Nam, 2023).
* Cơ sở hạ tầng và tiện ích: Theo khảo sát của DKRA Vietnam (2022), dự án có đầy đủ tiện ích xung quanh thường cao hơn so với nơi thiếu tiện ích.
* Chất lượng sản phẩm: Căn hộ cao cấp, diện tích rộng, tiêu chuẩn xây dựng tốt sẽ có giá bán cao hơn (Cushman & Wakefield, 2023).
* Tình hình kinh tế vĩ mô: Biến động về lạm phát, lãi suất, thu nhập dân cư đều ảnh hưởng đến giá bất động sản (Sở Xây dựng TPHCM, 2023)

## CHƯƠNG 3. CÁC PHƯƠNG PHÁP DỰ ĐOÁN GIÁ NHÀ VÀ CHUNG CƯ

### 3.1. Phân tích hồi quy tuyến tính

#### 3.1.1. Khái niệm và cơ sở lý thuyết

Phân tích hồi quy tuyến tính là phương pháp định lượng mô tả mối quan hệ giữa biến phụ thuộc với một hoặc nhiều biến độc lập, trong đó giả định mối quan hệ là tuyến tính. Biến phụ thuộc gọi là biến đáp ứng Y, các biến độc lập gọi là biến giải thích X.

Mô hình hồi qui tuyến tính đơn biến dùng để xem xét mối quan hệ tuyến tính giữa biến phụ thuộc y (biến kết cục) và biến độc lập x (biến dự đoán). Phương trình tuyến tính (đường thẳng) đơn biến có dạng:

**y= a + bxi+ei**

Trong đó a là điểm cắt trên trục tung, β là độ dốc (trong thống kê gọi là hệ số hồi qui) và e là phần dư. Ví dụ tìm sự liên hệ giữa diện tích (x) và giá nhà (y).

Tuy nhiên trong thực tế, thường có rất nhiều yếu tố khác khác nhau dẫn đến một hiện tượng. Ví dụ giá nhà không chỉ do diện tích tác động đến mà bên cạnh đó còn có các yếu tố khác như số phòng ngủ, số phòng tắm, số tầng...

Do đó, ta có mô hình qui tuyến tính đa biến có dạng:

**y= a + b1x1i+ b2x2i +b3x3i + ......+........+ bkxki + ei**

Với ví dụ trên ta có thể viết phương trình hồi quy tuyến tính đa biến với 4 yếu tố (x1 = diện tích, x2 = số phòng ngủ, x3 = số phòng tắm, x4 = số tầng) như sau:

Giá nhà = a + b1(diện tích) + b2(số phòng ngủ) + b3(số phòng tắm) + b4(số tầng) + e

Các giá trị hệ số hồi qui b (regression coefficient): β0, β1, β2, …, βk được chọn bằng phương pháp bình phương tối thiểu, giúp tối thiểu hóa tổng bình phương phần dư (RSS), yếu tố nào có b cao thì ảnh hưởng nhiều hơn.

Biến giả (dummy variables):

Biến giả, hay còn gọi là các biến định tính, là công cụ để đưa vào mô hình hồi quy những biến mà không thể lượng hóa ngay được, ví dụ như giới tính, tôn giáo, nhưng lại tác động tới giá trị của biến phụ thuộc. Biến giả có thể nhận một trong hai giá trị, 0, hoặc 1.

**RSS = Σ(yi – ŷi)2**

Trong đó:

* Σ: Ký hiệu tiếng Hy Lạp có nghĩa là tổng
* yi: Giá trị phản hồi thực tế cho quan sát thứ i
* ŷi: Giá trị phản hồi dự đoán dựa trên mô hình hồi quy tuyến tính bội

Trong phân tích hồi qui tuyến tính đa biến, ta cần biết mức độ ảnh hưởng của từng yếu tố lên biến kết cục y (giá nhà trong ví dụ này). Muốn biết mức độ ảnh hưởng cần lưu ý đến các trị số sau:

Hệ số tương quan R (coefficient of correlation): yếu tố nào có R càng lớn thì ảnh hưởng càng nhiều.

Bình phương của R (R square): Đây được biết đến là hệ số xác định. Đó là tỷ lệ phần trăm phương sai trong biến phụ thuộc mà có thể được giải thích bởi các biến giải thích. yếu tố nào có R2 càng lớn thì mối quan hệ giữa yếu tố đó và biến y càng chặt chẽ.

Cách xác định R2:

Đầu tiên, ta cần tính tổng bình phương sai số (SSE) của mô hình hồi quy tuyến tính bằng cách tính tổng bình phương của sai số (chênh lệch giữa giá trị dự đoán và giá trị thực tế) cho tất cả các điểm dữ liệu trong tập huấn luyện:

**SSE = Σ(y - ŷ)²**

Trong đó:

* y là giá trị thực tế của biến phụ thuộc.
* ŷ là giá trị dự đoán của biến phụ thuộc dựa trên các giá trị độc lập được sử dụng trong mô hình.

Tiếp đến, tính tổng bình phương sai số của trung bình (SST) bằng cách tính tổng bình phương chênh lệch giữa giá trị thực tế và giá trị trung bình của biến phụ thuộc cho tất cả các điểm dữ liệu trong tập huấn luyện:

**SST = Σ(y - ȳ)²**

Trong đó:

* ȳ là giá trị trung bình của biến phụ thuộc.

Tính hệ số xác định (R²) bằng cách tính tỷ lệ phần trăm giải thích được của biến phụ thuộc bởi các biến độc lập trong mô hình:

**R² = 1 - (SSE / SST)**

Trị số p (p value): càng nhỏ mức ảnh hưởng càng mạnh.

#### 3.1.2. Các giả định của mô hình hồi quy tuyến tính

Phần này giới thiệu các giả định cơ bản của mô hình hồi quy tuyến tính và vai trò quan trọng của chúng đối với việc ước lượng mô hình. Việc hiểu rõ các giả định này sẽ giúp đánh giá được tính hợp lệ của các kết quả ước lượng. Nếu mô hình không đáp ứng đủ các giả định, các hệ số ước lượng có thể bị sai lệch hoặc không nhất quán dẫn đến kết quả phân tích không đáng tin cậy.

Đầu tiên là các giả định cơ bản đảm bảo sự tồn tại của hệ số hồi quy tuyến tính:

* Hệ số hồi quy phải ở dạng hằng số: Nếu ở dạng số mũ, log, tỷ số sẽ không phải là một mối quan hệ tuyến tính đơn giản giữa biến phụ thuộc và độc lập.
* Các biến độc lập là cố định hoặc phi ngẫu nhiên: Giá trị các biến độc lập không thay đổi khi lấy mẫu khác nhau để có thể phân tích được ảnh hưởng của chúng.
* Các biến độc lập có phương sai lớn hơn 0: Có sự khác biệt giá trị quan sát để quan sát được mối quan hệ.

Tiếp đến là các giả định đảm bảo cho hệ số hồi quy không bị sai lệch, nhất quán và hiệu quả:

* Giá trị trung bình của sai số bằng 0: Loại trừ ảnh hưởng các yếu tố khác.
* Phương sai của sai số là một hằng số: Ảnh hưởng của biến độc lập lên phụ thuộc là như nhau.
* Các sai số không tự tương quan: Loại trừ ảnh hưởng thời gian.
* Các biến độc lập không có quan hệ tuyến tính hoàn hảo: Tránh đa cộng tuyến.
* Mô hình được thiết kế đúng: Đảm bảo không bỏ sót biến quan trọng.

### 3.2. Sử dụng phương pháp máy vector hỗ trợ (SVM)

Máy Vector Hỗ Trợ (SVM) là một phương pháp máy học được sử dụng chủ yếu cho các vấn đề phân loại và hồi quy. Mục tiêu của SVM là tìm ra đường ranh giới (hyperplane) phân chia tốt nhất giữa các nhóm dữ liệu. Trong trường hợp phân loại, SVM cố gắng tạo ra một đường ranh giới sao cho khoảng cách giữa các điểm dữ liệu của các nhóm là lớn nhất.

#### 3.2.1. Khái niệm và cơ sở lý thuyết

Hàm Ranh Giới (Decision Boundary): SVM tìm kiếm một đường ranh giới tốt nhất để phân chia giữa các nhóm dữ liệu. Đối với bài toán dự đoán giá nhà, đường ranh giới này sẽ phân chia các trường hợp thành các phân khúc giá khác nhau.

Khoảng Cách Margin: SVM cố gắng tối đa hóa khoảng cách (margin) giữa đường ranh giới và các điểm dữ liệu gần nhất của hai nhóm. Khoảng cách này được đo từ các điểm dữ liệu đến đường ranh giới.

Hàm Mất Mát (Loss Function): Trong quá trình tối ưu, SVM sử dụng hàm mất mát để đánh giá sự hiệu quả của đường ranh giới. Hàm mất mát bao gồm cả độ lớn của margin và sự phạt cho các điểm dữ liệu nằm sai lệch (misclassified).

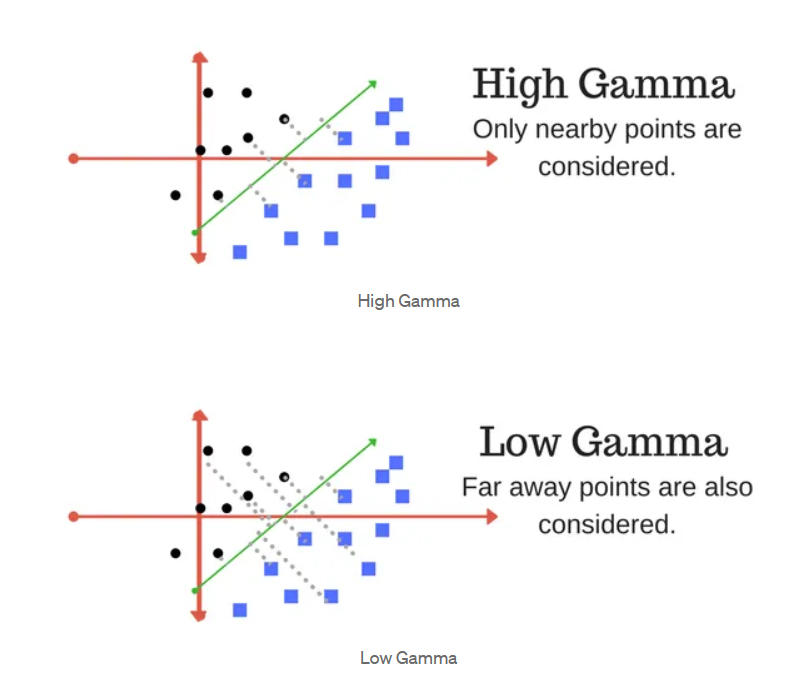
**

Trong đó w là vector trọng số, b là độ dời, ξi là biến lỏng (slack variable) đo lường sự sai lệch của điểm dữ liệu thứ i, và C là tham số đánh giá độ chặt của margin. Kernel Trick: SVM có khả năng sử dụng "kernel trick" để ánh xạ dữ liệu vào không gian cao chiều, giúp nó phân loại các dữ liệu không tuyến tính. Các hàm như Radial Basis Function (RBF) thường được sử dụng trong kernel trick. Trong SVM tuyến tính, quá trình học của siêu phẳng (hyperplane) được thực hiện thông qua việc chuyển đổi vấn đề sử dụng một số toán học tuyến tính. Đây là nơi mà khái niệm về kernel đóng một vai trò quan trọng. Đối với kernel tuyến tính, phương trình dự đoán cho một điểm dữ liệu mới bằng cách sử dụng tích vô hướng giữa điểm dữ liệu mới (x) và mỗi vector hỗ trợ (xi) được tính như sau:

***f*(*x*)=*B*0​+∑*i*​*ai*​⋅(*x*,*xi*​)**

Đây là một phương trình liên quan đến việc tính toán tích vô hướng của một vector dữ liệu mới (x) với tất cả các vector hỗ trợ trong dữ liệu huấn luyện. Các hệ số ***B*0​** và ***ai*** (cho mỗi điểm dữ liệu) phải được ước lượng từ dữ liệu huấn luyện thông qua thuật toán học.

Tham Số C và Gamma: SVM có các tham số quan trọng như tham số đánh giá độ chặt của margin (C) và tham số ảnh hưởng của kernel (gamma). Tinh chỉnh đúng các tham số này là quan trọng để đạt được hiệu suất tốt. Tham số gamma xác định mức độ ảnh hưởng của một ví dụ huấn luyện duy nhất, với các giá trị thấp có nghĩa là 'far' và giá trị cao có nghĩa là 'close'. Nói cách khác, với gamma thấp, các điểm ở xa đường phân cách hợp lý sẽ được xem xét khi tính toán đường phân tách. Trong trường hợp gamma cao có nghĩa là các điểm gần với đường hợp lý sẽ được xem xét trong tính toán.



**Hình 3. 1 Biểu đồ thể hiện High Gamma và Low Gamma**

#### 3.2.2. Ứng dụng mô hình vào dữ liệu thực tế

Sau khi xây dựng mô hình SVM, chúng ta sẽ áp dụng nó vào dữ liệu thực tế về giá nhà. Mô hình sẽ cố gắng tìm ra đường ranh giới tối ưu để phân loại các trường hợp vào các phân khúc giá khác nhau dựa trên các đặc trưng của chúng.

Trong việc dự đoán giá nhà, chúng ta áp dụng SVM hồi quy. Kernel (linear, RBF, polynomial) quyết định cách dữ liệu được ánh xạ vào không gian chiều cao. Tiền xử lý dữ liệu, như xử lý thiếu sót và chuẩn hóa, là bước quan trọng. Huấn luyện mô hình tìm đường ranh giới tốt nhất. Đánh giá mô hình sử dụng các độ đo như MSE, R-squared. Tinh chỉnh tham số và sau đó triển khai mô hình vào môi trường sản xuất. Lựa chọn kernel và tinh chỉnh tham số quan trọng để đạt hiệu suất tốt. Thử nghiệm nhiều mô hình giúp chọn ra mô hình phù hợp nhất cho bài toán.

#### 3.2.3. Đánh giá hiệu suất mô hình

Đánh giá hiệu suất mô hình là một phần quan trọng trong quá trình xây dựng và triển khai mô hình. Dưới đây là một số phương pháp phổ biến để đánh giá hiệu suất của mô hình hồi quy SVM khi áp dụng vào dự đoán giá nhà:

Mean Squared Error (MSE): MSE đo lường trung bình bình phương của sự chênh lệch giữa giá trị dự đoán và giá trị thực tế trên tập kiểm tra. Giá trị MSE càng thấp, mô hình càng chính xác.

**MSE = (1/n) \* ∑(y - ŷ)^2**

Trong đó, n là số lượng mẫu trong tập kiểm tra, y là giá trị thực tế và ŷ là giá trị dự đoán.

R-squared (R²): R-squared đo lường phần trăm biến thiên của giá trị phụ thuộc mà mô hình giải thích được. Giá trị R-squared càng cao, mô hình càng tốt.

**R² = 1 - (∑(y - ŷ)^2) / (∑(y - ȳ)^2)**

Trong đó, ȳ là giá trị trung bình của giá trị phụ thuộc y.

Đường hồi quy trên đồ thị: Trực quan hóa đường hồi quy và so sánh nó với dữ liệu thực tế trên một đồ thị để đánh giá sự khớp.

Kiểm định chéo (Cross-Validation): Chia dữ liệu thành các tập huấn luyện và kiểm tra để đánh giá mô hình trên nhiều tập dữ liệu khác nhau. Cross-validation giúp đảm bảo tính độc lập và khả năng tổng quát hóa của mô hình.

Đánh giá thực tế trên dữ liệu mới: Áp dụng mô hình vào dữ liệu mới và so sánh kết quả dự đoán với giá trị thực tế để đánh giá khả năng tổng quát hóa của mô hình.

Tinh chỉnh tham số và lặp lại quy trình: Nếu hiệu suất không đạt yêu cầu, tinh chỉnh tham số của mô hình và lặp lại quá trình huấn luyện và đánh giá.

Tổng cộng, việc sử dụng một hoặc kết hợp nhiều phương pháp trên giúp xác định chất lượng của mô hình hồi quy SVM và đảm bảo rằng nó có khả năng dự đoán hiệu quả trên dữ liệu mới.

### 3.3. Mạng nơ-ron nhân tạo

#### 3.3.1. Khái niệm và cơ sở lý thuyết

Mạng nơ-ron nhân tạo là một mô hình tính toán được lấy cảm hứng từ cấu trúc và hoạt động của hệ thống thần kinh sinh học. Nó được sử dụng để giải quyết các bài toán phức tạp và thường được áp dụng trong lĩnh vực học máy và trí tuệ nhân tạo.

Một mạng nơ-ron nhân tạo bao gồm một tập hợp các nơ-ron nhân tạo (hay còn gọi là "units" hoặc "neurons") được tổ chức thành các lớp. Mỗi nơ-ron nhân tạo nhận đầu vào từ các nơ-ron trước đó, thực hiện một phép tính toán và truyền kết quả đến các nơ-ron sau đó. Các kết nối giữa các nơ-ron có trọng số, thể hiện mức độ quan trọng của tín hiệu đầu vào đối với nơ-ron đó.

Cơ sở lý thuyết của mạng nơ-ron nhân tạo dựa trên ý tưởng về học tăng cường (reinforcement learning) và học có giám sát (supervised learning). Trong quá trình huấn luyện, mạng nơ-ron nhân tạo được cung cấp các ví dụ huấn luyện, trong đó có đầu vào và đầu ra mong muốn. Mục tiêu là điều chỉnh trọng số của các kết nối trong mạng để giảm thiểu sai số giữa đầu ra dự đoán và đầu ra mong muốn.

Một trong những phương pháp quan trọng để huấn luyện mạng nơ-ron nhân tạo là lan truyền ngược (backpropagation). Backpropagation dựa trên nguyên tắc lan truyền đầu ra (forward propagation) để tính toán đầu ra dự đoán và sau đó tính toán sai số giữa đầu ra dự đoán và đầu ra mong muốn. Sai số này được truyền ngược qua mạng để điều chỉnh trọng số của các kết nối, từ đó cải thiện khả năng dự đoán của mô hình.

Mạng nơ-ron nhân tạo có khả năng học và tự điều chỉnh thông qua việc điều chỉnh trọng số kết nối giữa các nơ-ron. Điều này cho phép nó học được các mẫu phức tạp và thực hiện các nhiệm vụ như phân loại, dự đoán và nhận dạng.

Tóm lại, mạng nơ-ron nhân tạo là một mô hình tính toán lấy cảm hứng từ hệ thống thần kinh sinh học. Nó dựa trên cơ sở lý thuyết học tăng cường và học có giám sát, và sử dụng các kỹ thuật như lan truyền ngược để huấn luyện và điều chỉnh trọng số. Mạng nơ-ron nhân tạo có khả năng học và thực hiện các tác vụ phức tạp trong lĩnh vực học máy và trí tuệ nhân tạo.

#### 3.3.2. Ứng dụng mô hình vào dữ liệu thực tế

Ứng dụng mô hình mạng nơ-ron nhân tạo vào dự đoán giá nhà là một cách phổ biến để sử dụng trí tuệ nhân tạo trong lĩnh vực bất động sản. Dưới đây là mô tả cách mô hình có thể được áp dụng trong dự đoán giá nhà:

Chuẩn bị dữ liệu: Bước đầu tiên là thu thập dữ liệu về các yếu tố có thể ảnh hưởng đến giá nhà, ví dụ như diện tích, vị trí, số phòng, tiện ích xung quanh và các thuộc tính khác. Dữ liệu này cần được thu thập và chuẩn bị để sử dụng trong quá trình huấn luyện mô hình.

Xây dựng kiến trúc mạng nơ-ron: Tiếp theo, bạn cần xác định kiến trúc mạng nơ-ron phù hợp cho mô hình dự đoán giá nhà. Một kiến trúc phổ biến là mạng nơ-ron truyền thẳng (feedforward neural network) với các lớp ẩn (hidden layers) và lớp đầu ra (output layer). Số lượng và kích thước của các lớp ẩn có thể được điều chỉnh để phù hợp với độ phức tạp của bài toán.

Tiền xử lý dữ liệu: Trước khi đưa dữ liệu vào mạng nơ-ron, thường cần tiền xử lý dữ liệu để chuẩn hóa và chuẩn bị. Các bước tiền xử lý có thể bao gồm chuẩn hóa dữ liệu số, mã hoá dữ liệu hạng mục thành dạng số (ví dụ: one-hot encoding), và chia dữ liệu thành tập huấn luyện và tập kiểm tra.

Huấn luyện mô hình: Tiếp theo, mô hình mạng nơ-ron được huấn luyện trên tập dữ liệu huấn luyện. Quá trình huấn luyện này liên quan đến việc điều chỉnh trọng số của các kết nối trong mạng để giảm thiểu sai số giữa giá nhà dự đoán và giá nhà thực tế. Phương pháp thường được sử dụng là lan truyền ngược (backpropagation) để cập nhật trọng số dựa trên độ lỗi của mô hình.

Đánh giá và điều chỉnh mô hình: Sau khi hoàn thành quá trình huấn luyện, mô hình được đánh giá bằng cách sử dụng tập dữ liệu kiểm tra. Đánh giá mô hình bao gồm việc tính toán các độ đo đánh giá như sai số trung bình, độ chính xác, hay hệ số xác định (R-squared) để đánh giá chất lượng dự đoán của mô hình. Nếu kết quả không đạt yêu cầu, mô hình có thể được điều chỉnh bằng cách thay đổi kiến trúc mạng, thay đổi siêu tham số, hoặc sử dụng các kỹ thuật tối ưu hóa khác.

Dự đoán giá nhà: Sau khi mô hình đã được huấn luyện và đánh giá, nó có thể được sử dụng để dự đoán giá nhà cho các dữ liệu mới. Đầu vào mới được đưa vào mô hình, và mô hình sẽ cho ra dự đoán về giá nhà dựa trên các đặc trưng của nó.

Lưu ý rằng quá trình xây dựng mô hình dự đoán giá nhà có thể phức tạp hơn và đòi hỏi nhiều công cụ và kỹ thuật phức tạp hơn như xử lý dữ liệu bị thiếu, tăng cường dữ liệu, tối ưu hóa mô hình và đánh giá chính xác. Điều này chỉ là mô tả chung về cách mô hình mạng nơ-ron có thể được áp dụng trong dự đoán giá nhà.

### 3.4. Các phương pháp khác

K-nearest neighbors (KNN): KNN dựa trên nguyên lý rằng các điểm dữ liệu tương tự nhau thường có giá trị tương tự. Phương pháp này tính toán khoảng cách giữa điểm dữ liệu mới và các điểm trong tập dữ liệu huấn luyện, sau đó dự đoán giá nhà dựa trên giá trị của các láng giềng gần nhất (K láng giềng).

Máy vector hỗ trợ phi tuyến tính (Nonlinear Support Vector Machines): Máy vector hỗ trợ phi tuyến tính mở rộng SVM để xử lý các dữ liệu phi tuyến tính bằng cách ánh xạ dữ liệu vào không gian đặc trưng cao chiều thông qua các hàm nhân phi tuyến tính như hàm đa thức, hàm Radial Basis Function (RBF) và hàm sigmoid. Điều này cho phép SVM xử lý các mối quan hệ phức tạp hơn giữa các thuộc tính và dự đoán giá nhà.

Mạng Bayesian (Bayesian Networks): Bayesian Networks là mô hình biểu diễn các mối quan hệ xác suất giữa các biến dựa trên lý thuyết xác suất và đồ thị hướng dẫn. Trong dự đoán giá nhà, Bayesian Networks có thể được sử dụng để mô hình hóa các mối quan hệ xác suất giữa các yếu tố như diện tích, vị trí, số phòng, và giá nhà.

Mô hình học tăng cường (Reinforcement Learning): Reinforcement Learning là một phương pháp học máy trong đó một hệ thống tương tác với môi trường và học từ các phản hồi trực tiếp (rewards). Trong dự đoán giá nhà, Reinforcement Learning có thể được sử dụng để tìm ra các chiến lược tối ưu để đưa ra các quyết định về giá nhà dựa trên các hành động và phản hồi từ thị trường.

Mô hình học gom cụm (Clustering Models): Mô hình gom cụm như K-means clustering có thể được sử dụng để phân nhóm các điểm dữ liệu giống nhau về các yếu tố như diện tích, vị trí, và số phòng. Sau đó, giá nhà trong mỗi nhóm có thể được dự đoán dựa trên giá trị trung bình hoặc các phương pháp khá

## CHƯƠNG 4. THU THẬP DỮ LIỆU VÀ XỬ LÝ DỮ LIỆU

### 4.1. Nguồn dữ liệu

Website https://www.nhatot.com là một trong những nguồn thông tin uy tín về thị trường bất động sản tại TP. Hồ Chí Minh. Trang web liên tục cập nhật các tin rao bán, cho thuê căn hộ chung cư, nhà riêng do chủ đầu tư, công ty môi giới và cá nhân đăng tải.

Các thông tin được cung cấp kèm theo mỗi tin rao bán bao gồm: địa chỉ, phường quận cụ thể; diện tích sàn, số phòng ngủ, toilet; năm xây dựng; giá bán/giá thuê; ảnh minh họa sản phẩm. Đây là những thông tin cần thiết phục vụ mục đích nghiên cứu thị trường và phân tích giá.

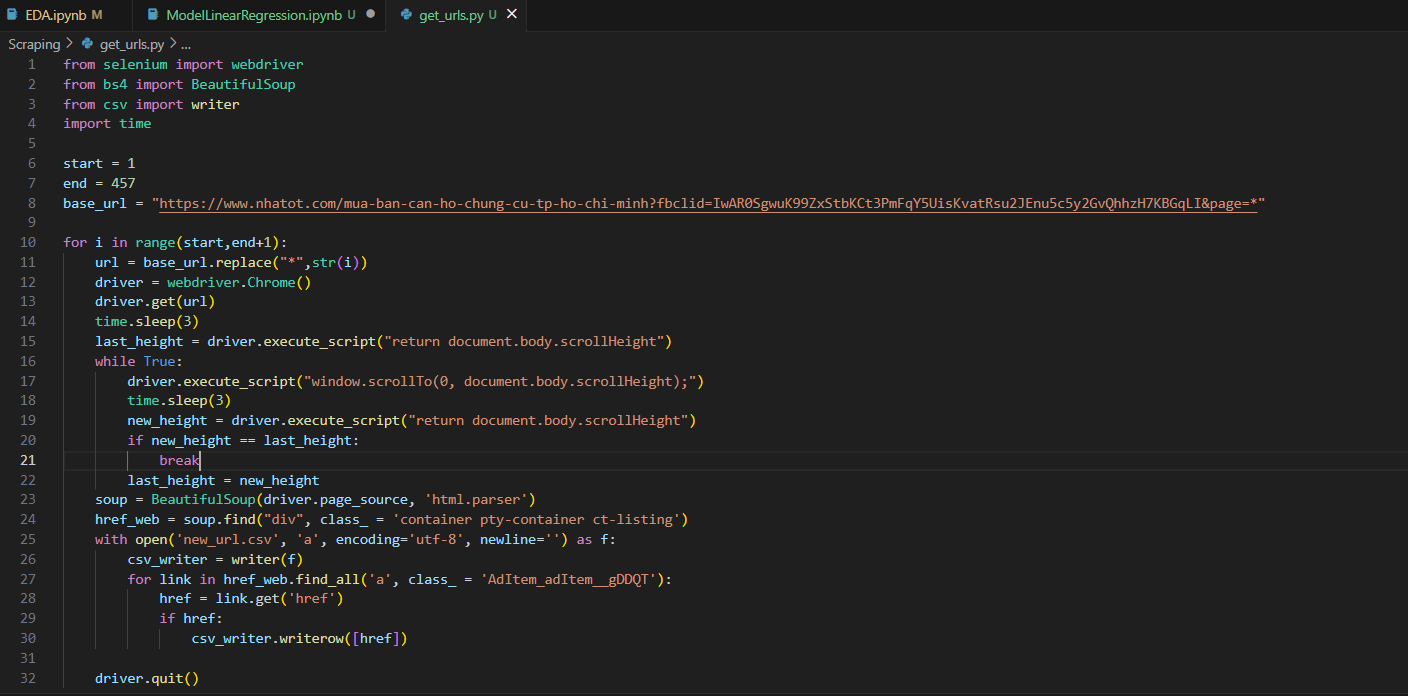
Lượng thông tin trên website được cập nhật hàng ngày, số lượng tin rao bán lớn và phủ khắp các quận huyện TP.HCM. Đặc biệt, website cung cấp nhiều loại hình bất động sản khác nhau như chung cư, nhà phố, biệt thự...cho phép phân tích đa chiều.

### 4.2. Phương pháp thu thập dữ liệu

Để tiến hành nghiên cứu và phân tích thị trường bất động sản, bước đầu tiên là thu thập dữ liệu từ các nguồn tin cậy. Một trong những nguồn dữ liệu quan trọng là website https://www.nhatot.com cung cấp nhiều thông tin chi tiết về căn hộ chung cư rao bán tại TPHCM.

Để thu thập toàn bộ dữ liệu trên website này, em đã xây dựng phần mềm crawl dữ liệu bằng cách sử dụng ngôn ngữ lập trình Python có tên là get\_urls.py. Cụ thể, chương trình sẽ thực hiện các bước sau:

* Truy cập từng trang trên website và lấy nội dung bằng thư viện Selenium.
* Sử dụng Beautiful Soup phân tích cấu trúc DOM của trang và trích xuất dữ liệu.
* Lặp lại quá trình trên cho tất cả các trang có tin rao bán.
* Lưu trữ dữ liệu đã thu thập vào cơ sở dữ liệu để quản lý và phân tích tiếp theo.



**Hình 4. 1 Đoạn mã sử dụng để thu thập danh sách các urls**

Mục đích chính của đoạn code là thu thập danh sách các tin rao bán bất động sản trên website nhà đất. Đây là nguồn dữ liệu quan trọng cần phải khai thác để phục vụ nghiên cứu thị trường.

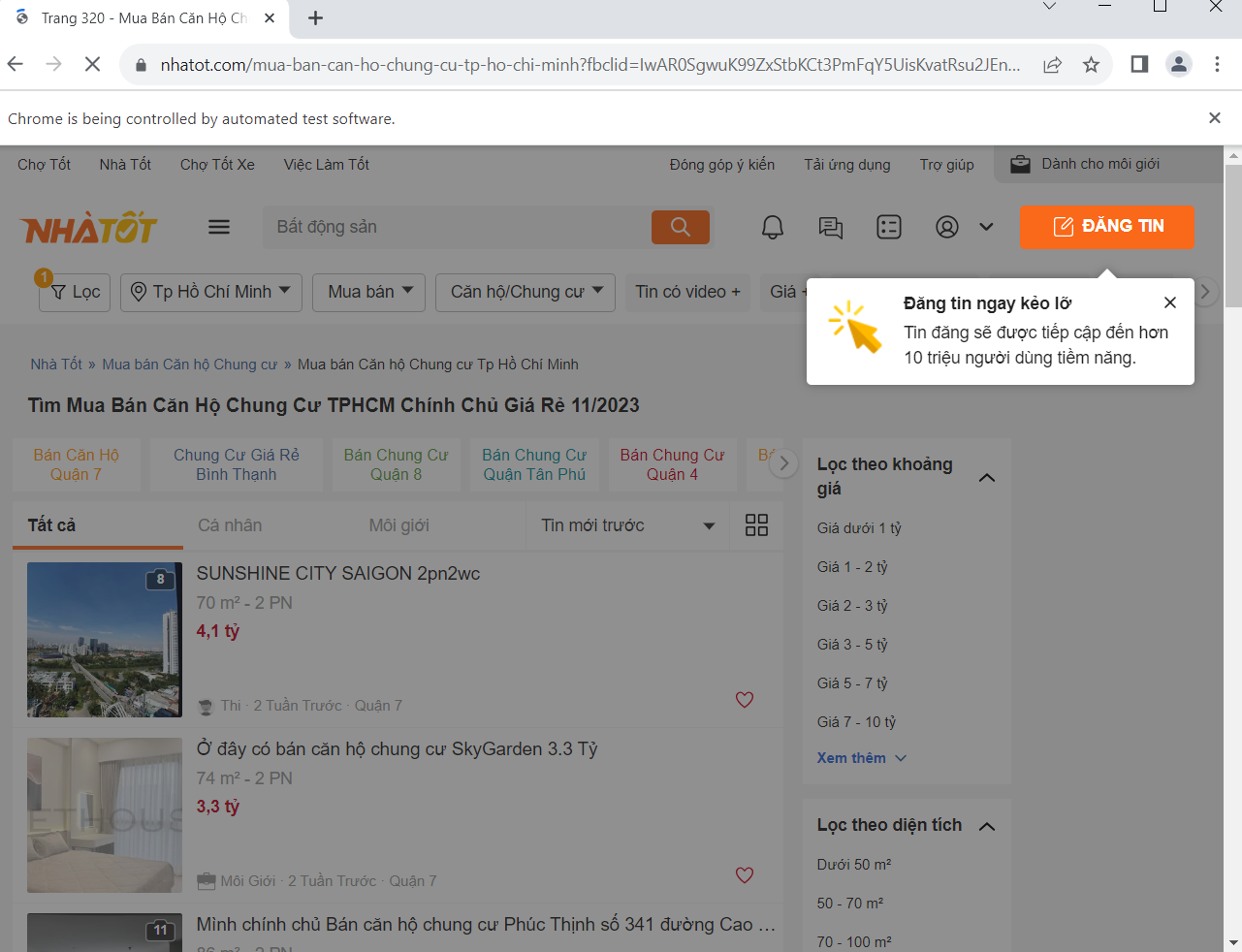
Tuy nhiên, với số lượng tin lớn trên nhiều trang, việc thu thập thủ công sẽ mất nhiều thời gian. Do đó, chương trình sử dụng công nghệ crawl tự động để lấy dữ liệu hiệu quả hơn.

Cụ thể, Selenium được dùng điều khiển trình duyệt, truy cập từng trang theo quy tắc định nghĩa. Beautiful Soup phân tích DOM tree, tìm kiếm và lấy thông tin link các tin rao bán.

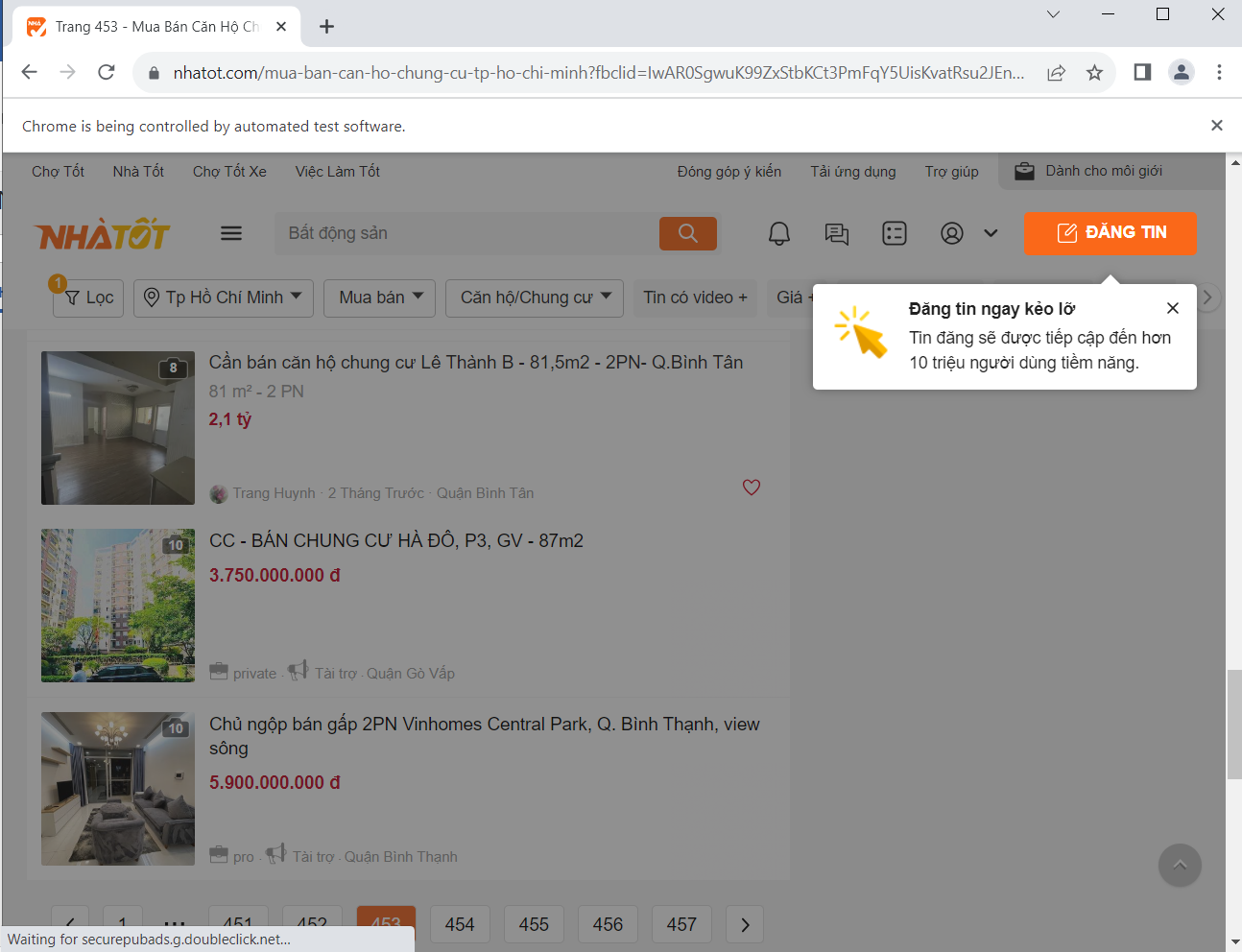
Sau khi thu thập xong, dữ liệu được lưu vào file csv để quản lý và sẵn sàng cho các bước xử lý, phân tích tiếp theo phục vụ nghiên cứu.

Bước thu thập dữ liệu này góp phần cung cấp nguồn dữ liệu số lượng lớn cho quá trình nghiên cứu và rút ra kết luận cuối cùng. Từ đó, mục tiêu đề ra có thể đạt được.

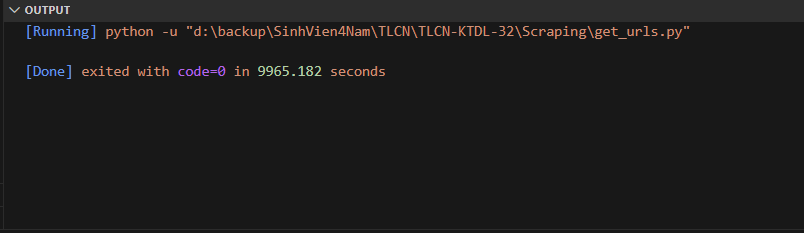
Dưới đây là hình ảnh của quá trình việc thu thập các url từ đoạn mã get\_urls.py.



**Hình 4. 2 Quá trình thu thập các urls trên website NhaTot**

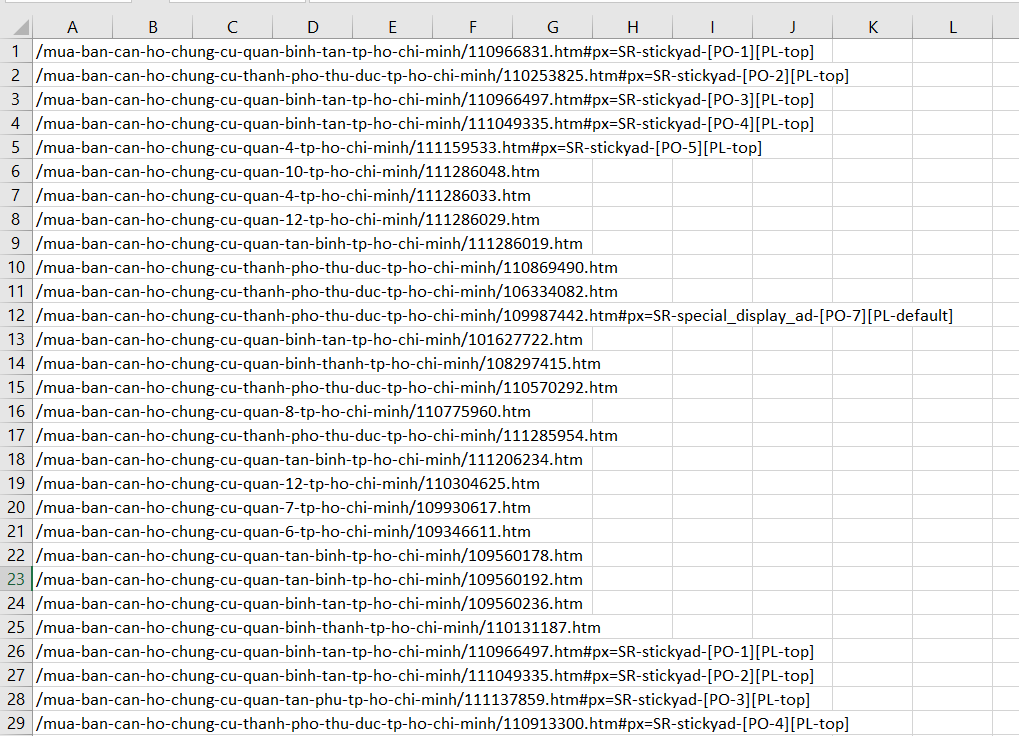


**Hình 4. 3 Quá trình thu thập các urls đang tiếp diễn**



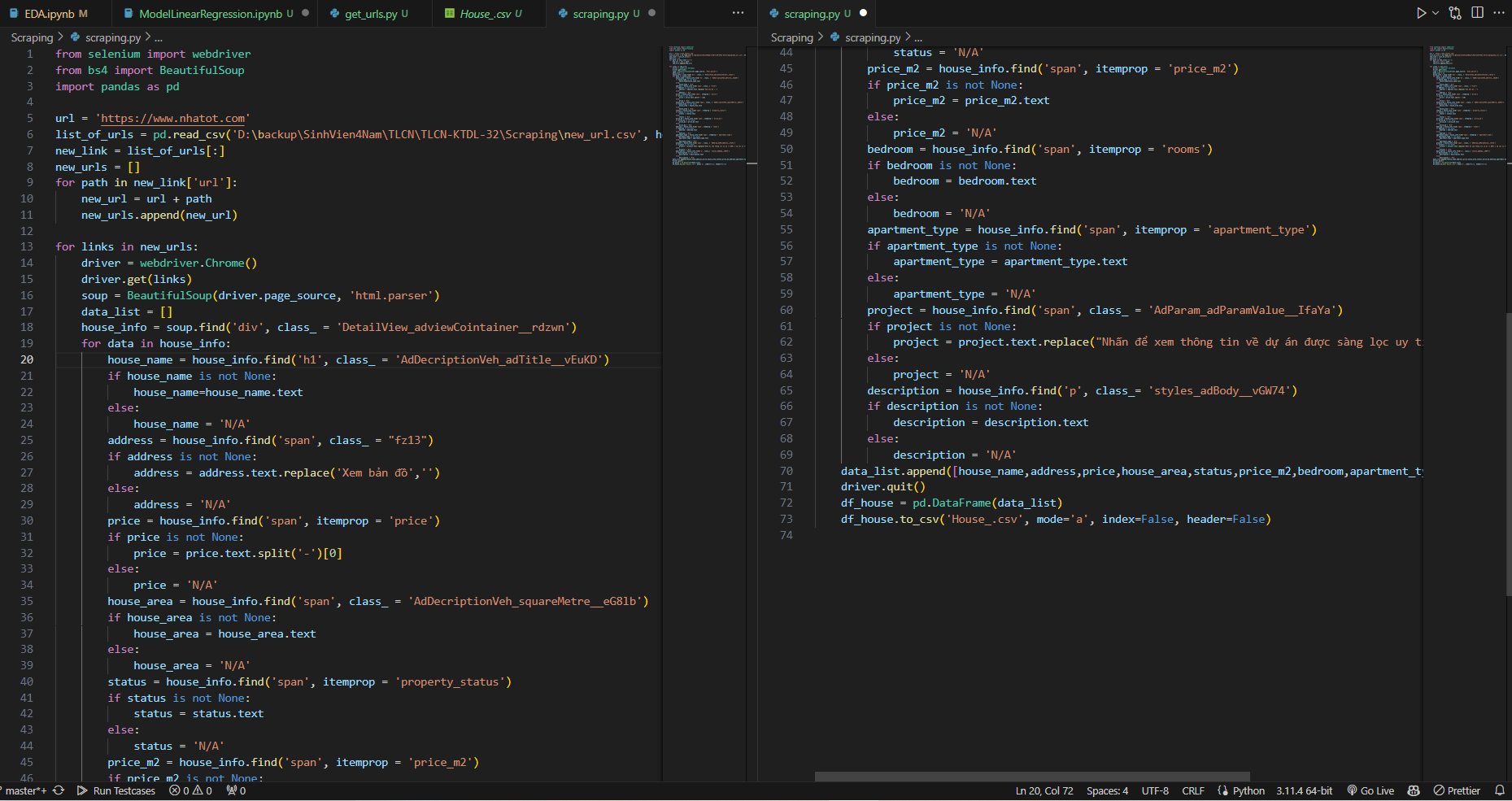
**Hình 4. 4 Kết thúc quá trình thu thập các urls**

Dữ liệu đường dẫn sau khi thu thập được



**Hình 4. 5 Kết quả của quá trình thu thập urls sau khi lưu vào file excel**

File scraping.py



**Hình 4. 6 Đoạn mã sử dụng để lấy dữ liệu từ trang NhaTot với các urls đã thu thập phía trên**

Đoạn code trên là một đoạn mã Python được sử dụng để tự động thu thập thông tin về các nhà ở từ một danh sách các URL và lưu kết quả vào một file CSV.

Trước tiên, các thư viện cần thiết như Selenium, BeautifulSoup và Pandas được import vào chương trình. Sau đó, biến url được khai báo để lưu trữ đường dẫn gốc của trang web mà chúng ta muốn thu thập dữ liệu.

Tiếp theo, chúng ta đọc danh sách các URL từ một file CSV và tạo một danh sách mới để chứa các URL này. URL gốc cũng được thêm vào danh sách mới.

Tiếp theo là một vòng lặp, trong đó chúng ta sử dụng Selenium để khởi tạo một trình duyệt web (ở đây là Chrome) và truy cập vào từng URL trong danh sách mới.

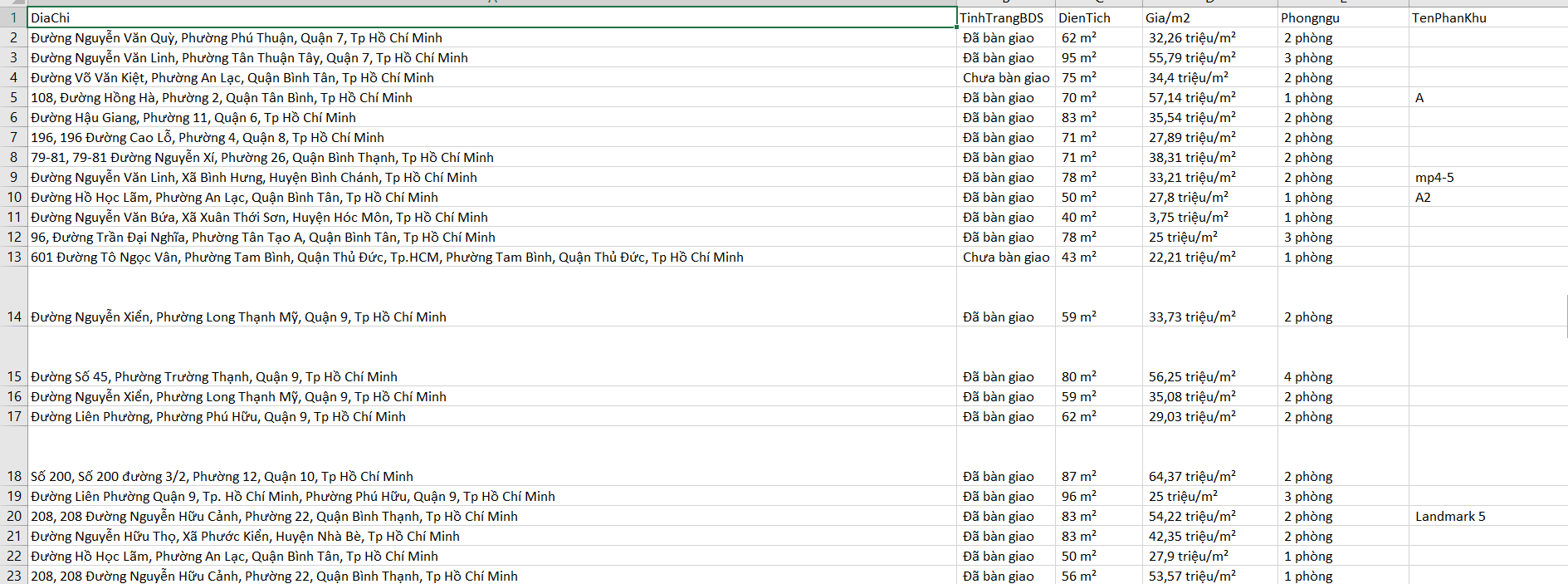
Sau khi trang web được tải, chúng ta sử dụng BeautifulSoup để phân tích cú pháp HTML và lấy thông tin về các nhà ở từ trang web này. Thông tin như tên nhà, địa chỉ, giá, diện tích, tình trạng, giá mỗi mét vuông, số phòng ngủ, loại căn hộ, dự án và mô tả được lấy ra và lưu vào một danh sách.

Sau khi thu thập thông tin từ một trang web, chúng ta đóng trình duyệt và tiếp tục vòng lặp để xử lý các trang web tiếp theo trong danh sách.

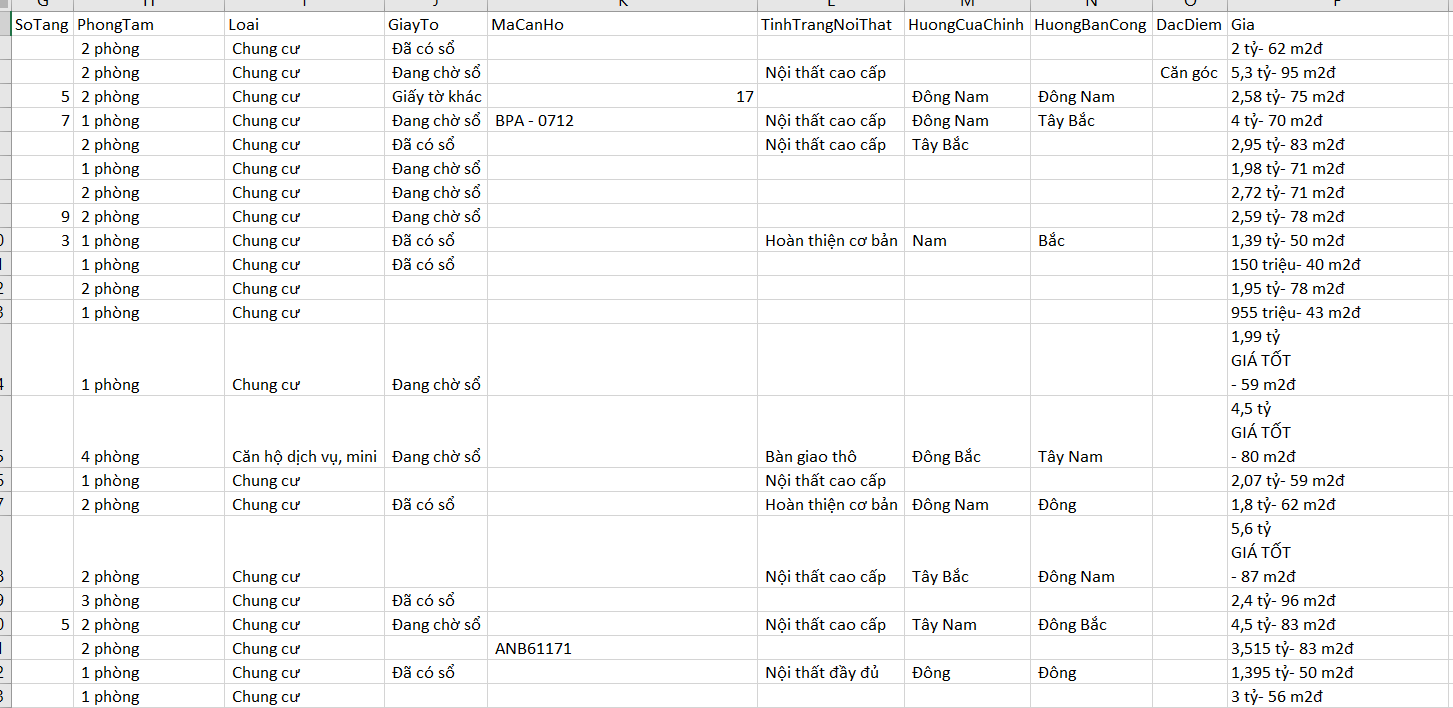
Cuối cùng, chúng ta tạo một DataFrame từ danh sách thông tin nhà và ghi dữ liệu vào một file CSV để lưu trữ. Điều này cho phép chúng ta dễ dàng truy cập và sử dụng thông tin về các nhà ở thu thập được sau này.

Tổng thể, đoạn code này cung cấp một phương pháp tự động và hiệu quả để thu thập thông tin về các nhà ở từ một danh sách các URL và tổ chức dữ liệu thu thập được vào một file CSV.

Dưới đây là kết quả sau khi lưu data vào raw\_data.csv được từ các urls:



**Hình 4. 7 Mô tả các biến trong tập dữ liệu đã thu thập được (1)**



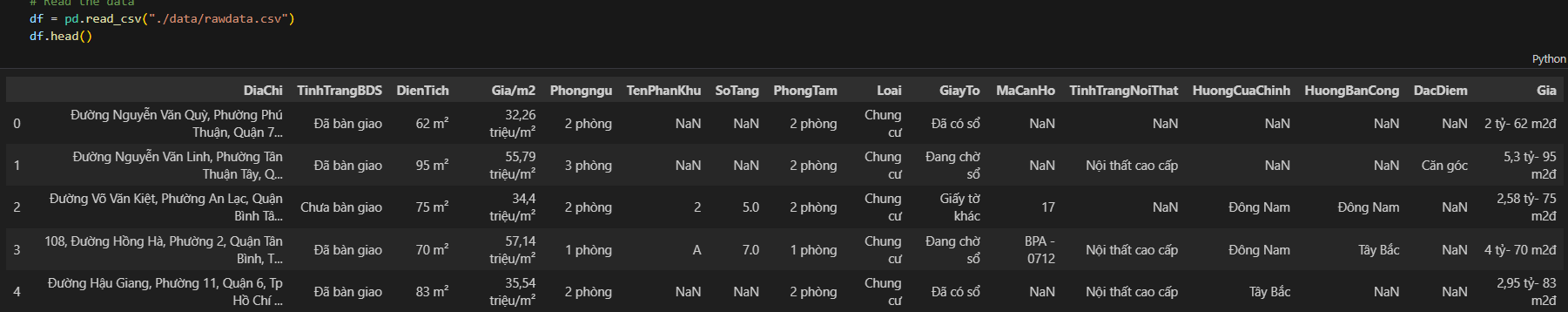
**Hình 4. 8 Mô tả các biến trong tập dữ liệu đã thu thập được (2)**

Mô tả về các biến có trong tập dữ liệu:

* Địa chỉ (DiaChi): Đây là trường chứa địa chỉ của bất động sản, tức là vị trí nơi nó được đặt.
* Tình trạng bất động sản (TinhTrangBDS): Đây là trường mô tả tình trạng hiện tại của bất động sản, chẳng hạn như "Đã bàn giao", "Chưa bàn giao" ...
* Diện tích (DienTich): Đây là trường chứa diện tích của bất động sản, được đo bằng đơn vị mét vuông.
* Giá/m2 (Gia/m2): Đây là trường chứa giá bán hoặc giá thuê của bất động sản mỗi mét vuông.
* Số phòng ngủ (Phongngu): Đây là trường mô tả số lượng phòng ngủ có trong bất động sản.
* Tên phân khu (TenPhanKhu): Đây là trường chứa tên phân khu hoặc tên dự án của bất động sản.
* Số tầng (SoTang): Đây là trường mô tả số tầng của bất động sản.
* Phòng tắm (PhongTam): Đây là trường mô tả số lượng phòng tắm có trong bất động sản.
* Loại (Loai): Đây là trường mô tả loại bất động sản, chẳng hạn như "Chung cư", "căn hộ dịch vụ, mini", "penthouse", "duplex", "quần thể", "cư xá" và các loại khác.
* Giấy tờ (GiayTo): Đây là trường mô tả tình trạng giấy tờ pháp lý của bất động sản, ví dụ như "đã có sổ", "đang chờ sổ", "giấy tờ khác" và các tình trạng tương tự.
* Mã căn hộ (MaCanHo): Đây là trường chứa mã căn hộ hoặc mã định danh duy nhất cho mỗi bất động sản.
* Tình trạng nội thất (TinhTrangNoiThat): Đây là trường mô tả tình trạng nội thất của bất động sản, chẳng hạn như "hoàn thiện cơ bản", "nội thất cao cấp", "nội thất đầy đủ", "bàn giao thô" và các tình trạng tương tự.
* Hướng cửa chính (HuongCuaChinh): Đây là trường mô tả hướng của cửa chính của bất động sản, chẳng hạn như "đông", "tây", "nam", "bắc", và các hướng khác.
* Hướng ban công (HuongBanCong): Đây là trường mô tả hướng của ban công của bất động sản, tương tự như trường Hướng cửa chính.
* Đặc điểm (DacDiem): Đây là trường chứa các đặc điểm đặc biệt, thông tin mô tả về bất động sản. Chẳng hạn như là "căn góc" ...
* Giá (Gia): Đây là trường chứa giá bán hoặc giá thuê của bất động sản.

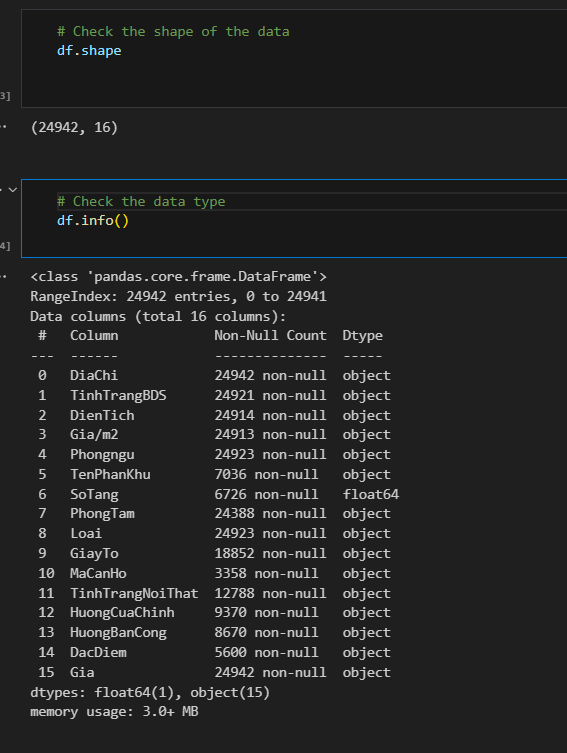
### 4.3. Tiền xử lý dữ liệu

In ra 5 dòng đầu của tập dữ liệu:



**Hình 4. 9 Mô tả 5 dòng đầu của dữ liệu**

Mô tả về tập dữ liệu:



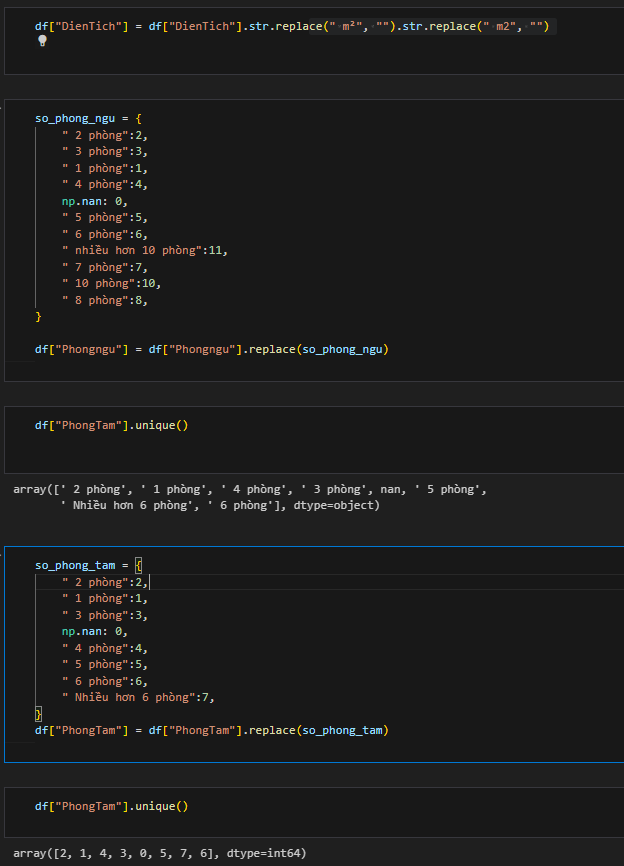
**Hình 4. 10 Mô tả về các biến thu thập được và kiểu dữ liệu**

Số lượng null của các biến trong tập dữ liệu:



**Hình 4. 11 Số lượng các biến bị null trong tập dữ liệu**

Chuẩn hóa dữ liệu DienTich, PhongNgu, PhongTam về dạng số:



**Hình 4. 12 Chuẩn hóa dữ liệu DienTich, PhongTam, PhongNgu về dạng số**

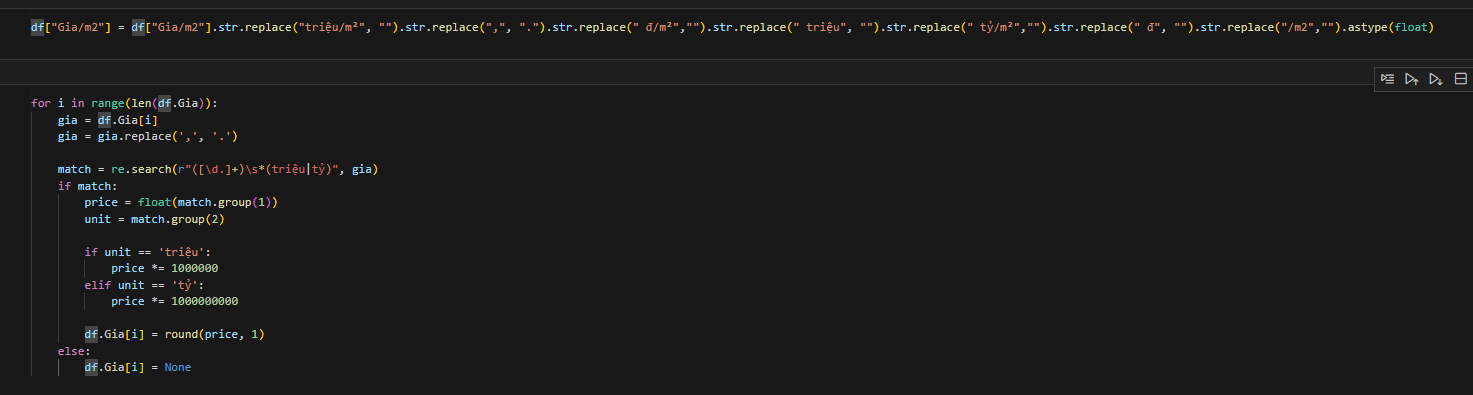
Xử lý dữ liệu PhongNgu:

* Đoạn mã Python này đang chuyển đổi dữ liệu trong cột 'Phongngu' của dataframe df từ dạng chuỗi sang dạng số.
* Đầu tiên, một từ điển so\_phong\_ngu được tạo ra để ánh xạ các giá trị chuỗi sang giá trị số tương ứng.
* Sau đó, hàm replace() được sử dụng để thay thế các giá trị chuỗi trong cột 'Phongngu' bằng các giá trị số tương ứng từ từ điển so\_phong\_ngu.

Xử lý dữ liệu PhongTam:

* Đoạn mã Python này đang chuyển đổi dữ liệu trong cột 'PhongTam' của dataframe df từ dạng chuỗi sang dạng số.
* Đầu tiên, một từ điển so\_phong\_tam được tạo ra để ánh xạ các giá trị chuỗi sang giá trị số tương ứng.
* Sau đó, hàm replace() được sử dụng để thay thế các giá trị chuỗi trong cột 'PhongTam' bằng các giá trị số tương ứng từ từ điển so\_phong\_tam.

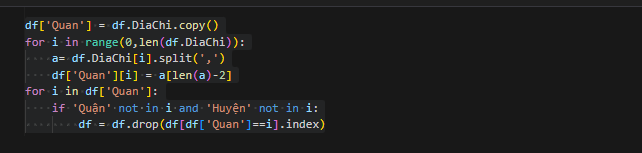
Xử lý các kiểu dữ liệu dạng chuỗi đưa về dạng số của các biến Gia/m2 và Gia:



**Hình 4. 13 Xử lý các kiểu dữ liệu dạng chuỗi đưa về dạng số của các biến Gia/m2 và Gia**

* Đoạn mã Python này thực hiện việc làm sạch và chuyển đổi dữ liệu trên cột 'Gia/m2' và 'Gia' của dataframe df.
* Đầu tiên, cột 'Gia/m2' được làm sạch bằng cách thay thế các chuỗi không cần thiết bằng chuỗi rỗng hoặc dấu chấm (để xử lý số thập phân), sau đó chuyển đổi chuỗi đã làm sạch thành số float.
* Phần còn lại của mã là một vòng lặp xử lý từng hàng trong cột 'Gia'. Đối với mỗi hàng, mã thực hiện các bước sau:
* Thay thế dấu phẩy bằng dấu chấm trong giá trị 'Gia' để xử lý số thập phân.
* Sử dụng biểu thức chính quy để tìm một số theo sau là 'triệu' hoặc 'tỷ'. Nếu tìm thấy, số được chuyển đổi thành float và đơn vị ('triệu' hoặc 'tỷ') cũng được trích xuất.
* Tùy thuộc vào đơn vị, giá cả được nhân với 1,000,000 (đối với 'triệu') hoặc 1,000,000,000 (đối với 'tỷ') để chuyển giá cả về cùng một tỷ lệ.
* Giá cả sau đó được làm tròn đến 1 chữ số thập phân và lưu trở lại vào cột 'Gia' của dataframe.
* Nếu không tìm thấy mẫu phù hợp trong giá trị 'Gia', nó sẽ được đặt thành None.

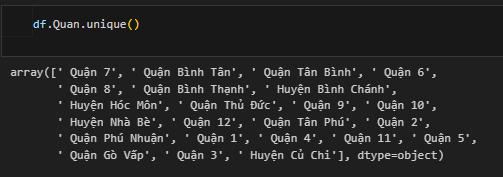
Xử lý dữ liệu Quan:



**Hình 4. 14 Xử lý dữ liệu Quan**

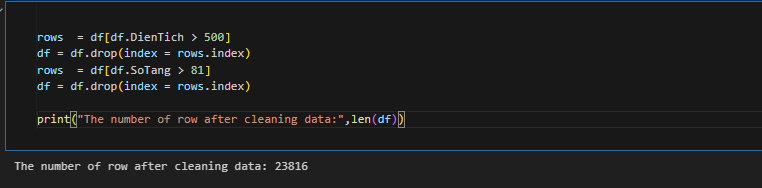
* Đoạn mã Python này thực hiện việc trích xuất và làm sạch dữ liệu trên cột 'DiaChi' và 'Quan' của dataframe df.
* Đầu tiên, cột 'DiaChi' được sao chép sang cột mới 'Quan'.
* Sau đó, một vòng lặp được sử dụng để duyệt qua từng hàng trong cột 'DiaChi'. Đối với mỗi hàng, giá trị 'DiaChi' được chia thành một danh sách các chuỗi bằng dấu phẩy. Chuỗi thứ hai từ cuối trong danh sách (được giả định là quận hoặc 'Quan') sau đó được lưu vào cột 'Quan' của dataframe.
* Cuối cùng, một vòng lặp khác được sử dụng để duyệt qua từng hàng trong cột 'Quan'. Nếu giá trị 'Quan' không chứa chuỗi 'Quận' hoặc 'Huyện', hàng tương ứng sẽ bị xóa khỏi dataframe.
* Kết quả của đoạn mã này là cột 'Quan' trong dataframe `df` giờ đây chứa thông tin quận được trích xuất từ cột 'DiaChi', và các hàng có thông tin quận không hợp lệ đã bị xóa.

Kết quả sau khi xử lý:

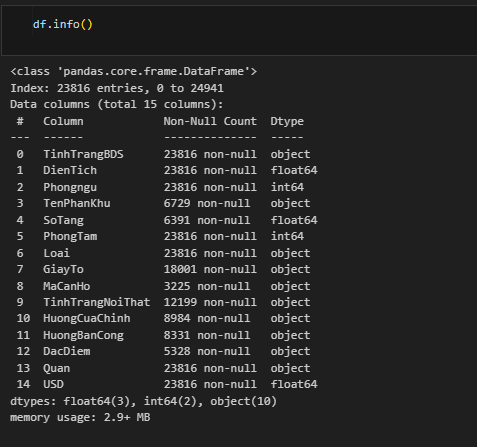


**Hình 4. 15 Dữ liệu của Quan sau khi được xử lý**

Xử lý các ngoại lệ (outlier) và in ra số lượng dữ liệu sau khi xử lý:



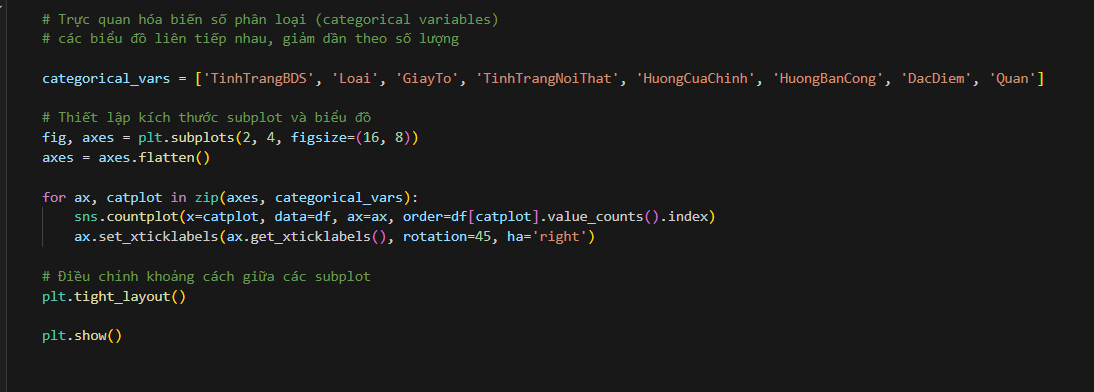
**Hình 4. 16 Xử lý các ngoại lệ (outlier)**



**Hình 4. 17 Dữ liệu sau khi đã được tiền xử lý**

### 4.4. Trực quan hóa dữ liệu

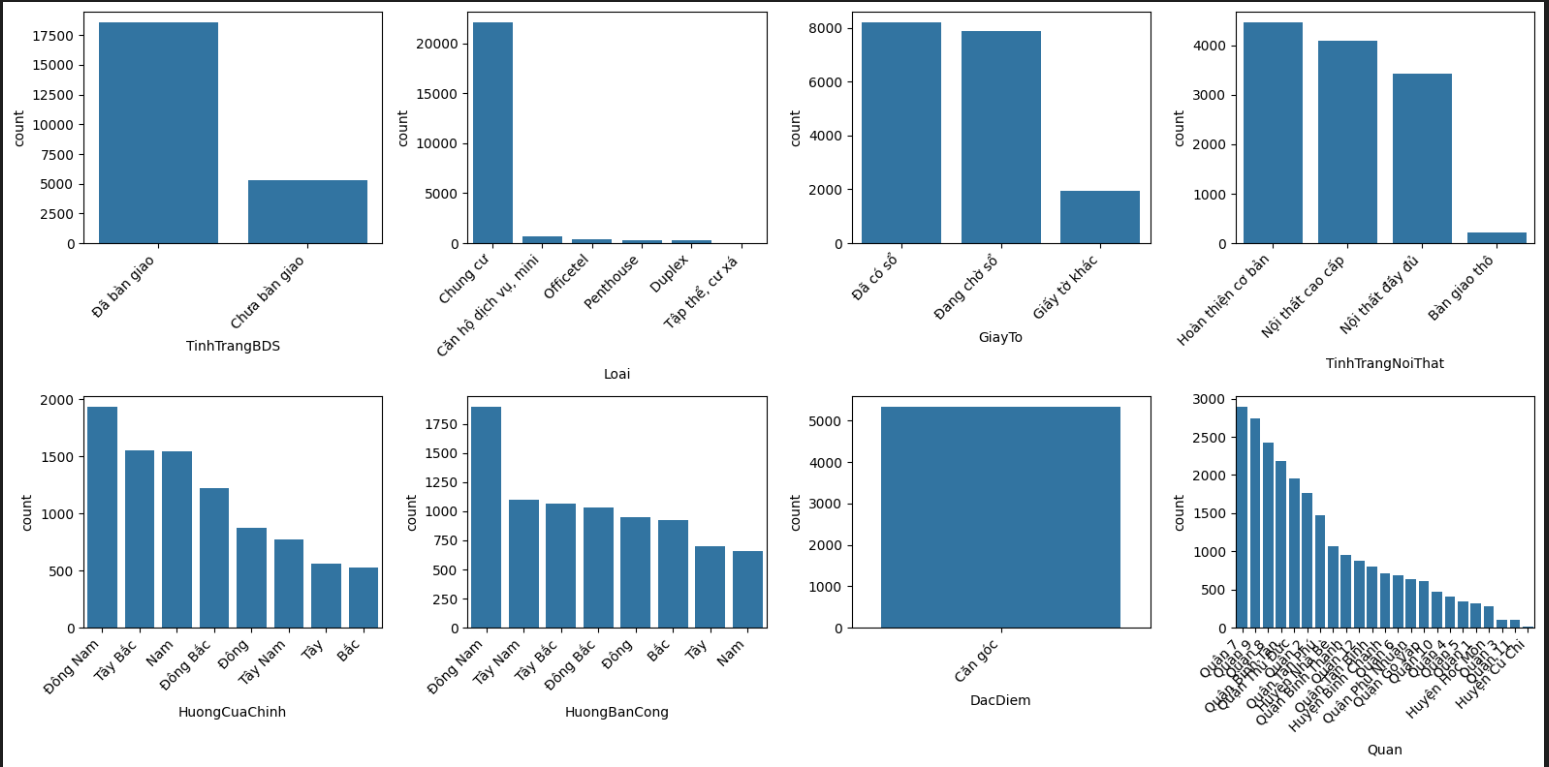
Trực quan hóa biến số phân loại (categorical variables):



**Hình 4. 18 Đoạn mã thực hiện việc trực quan hóa biến số phân loại (categorical variables)**

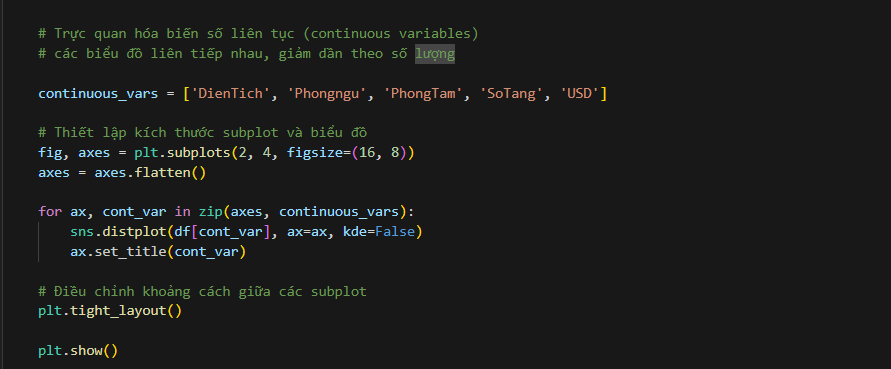
Đoạn mã Python này tạo ra các biểu đồ đếm cho các biến phân loại trong dataframe df.

* Danh sách categorical\_vars chứa tên các cột phân loại.
* plt.subplots(2, 4, figsize=(16, 8)) tạo ra một hình với 8 subplot (2 hàng, 4 cột).
* Vòng lặp for tạo biểu đồ đếm cho mỗi biến phân loại trên mỗi subplot.
* sns.countplot() tạo biểu đồ đếm, ax.set\_xticklabels() xoay nhãn trục x 45 độ để dễ đọc.
* plt.tight\_layout() điều chỉnh khoảng cách giữa các subplot.
* plt.show() hiển thị hình với tất cả các subplot.



**Hình 4. 19 Trực quan hóa biến số phân loại (categorical variables)**

Trực quan hóa biến số liên tục (continous variable):

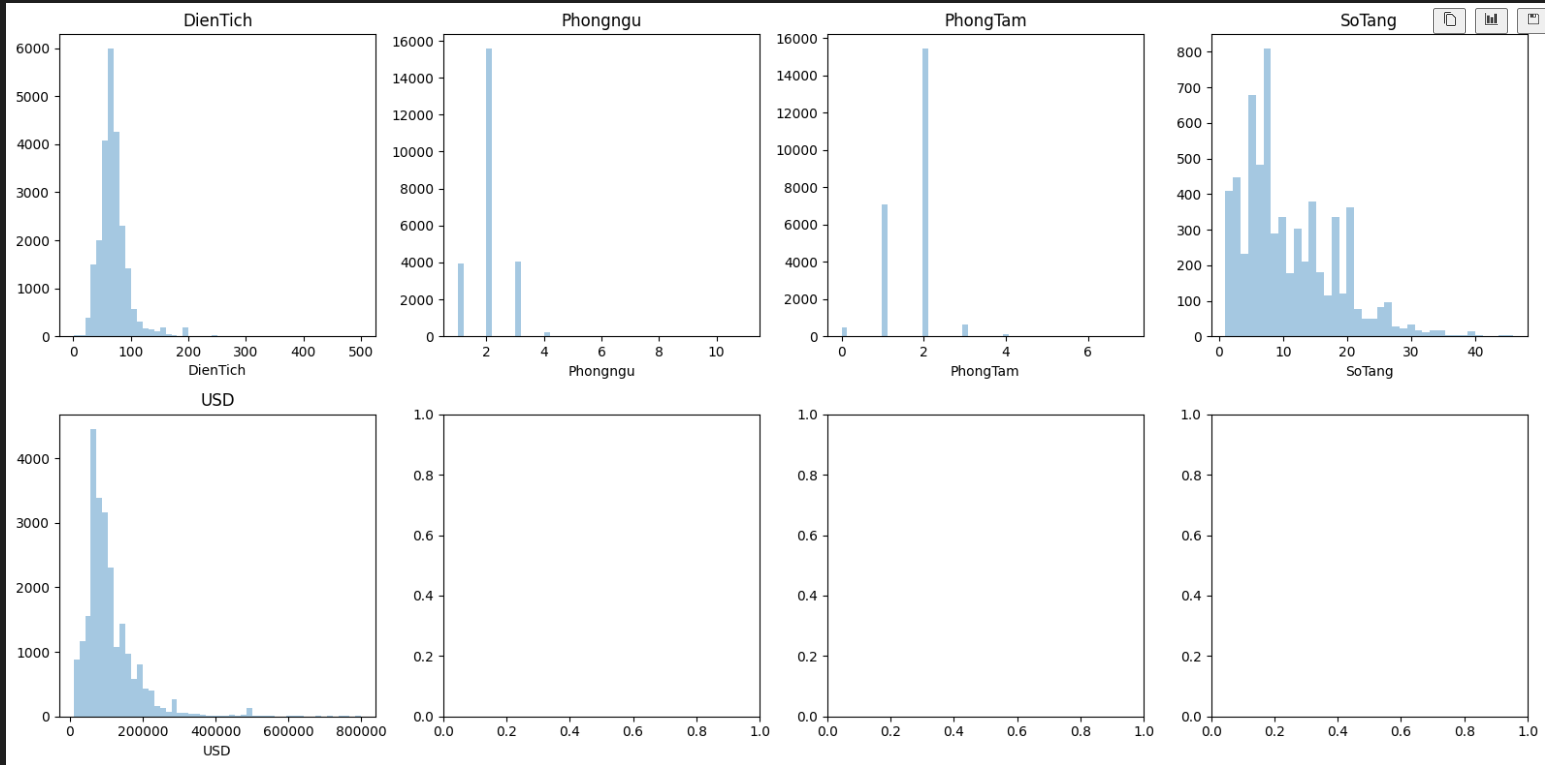


**Hình 4. 20 Đoạn mã thực hiện việc trực quan hóa biến số liên tục (continous variable):**

Đoạn mã Python này tạo ra các biểu đồ phân phối cho các biến liên tục trong dataframe df.

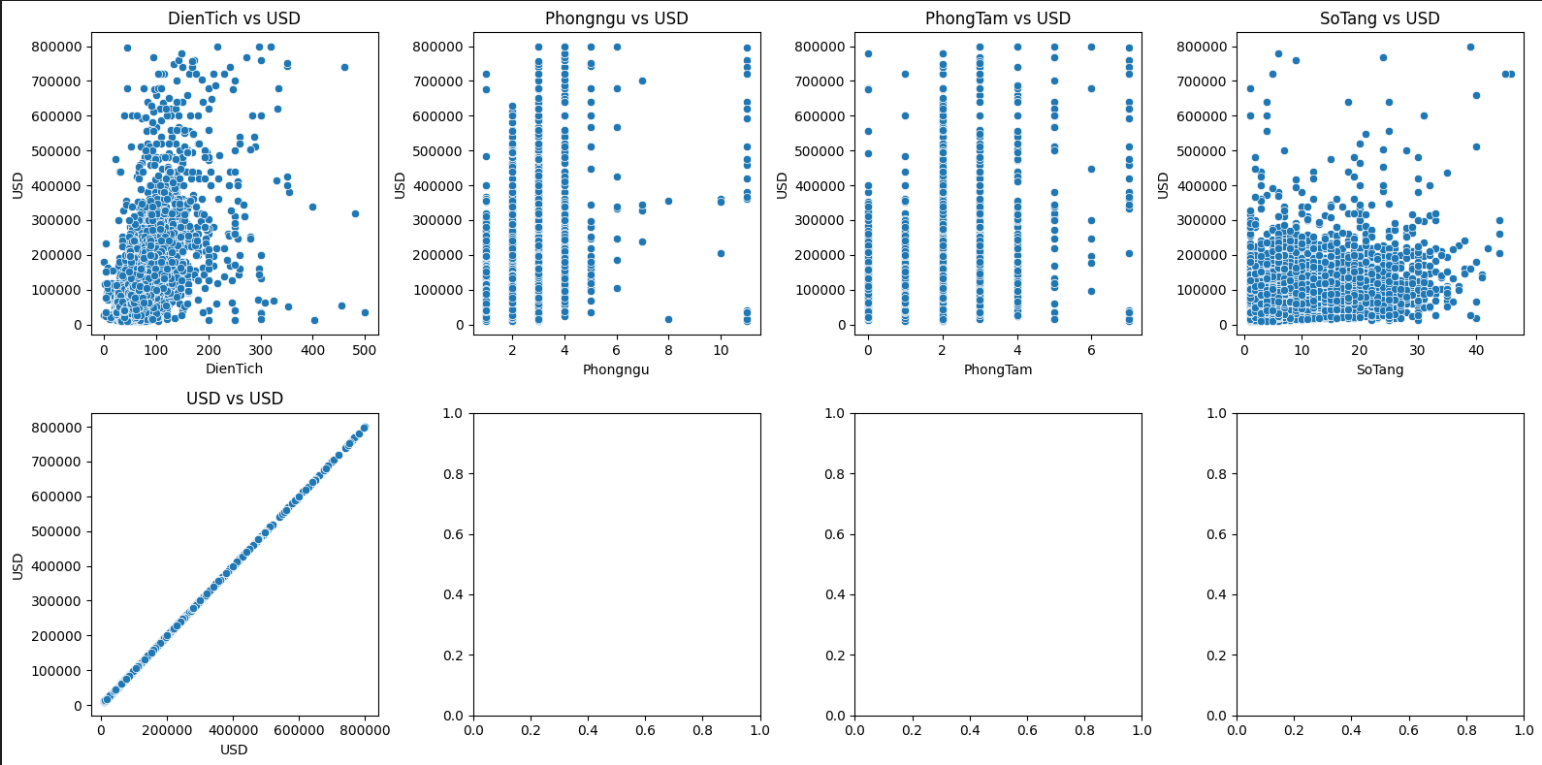
* Danh sách continuous\_vars chứa tên các cột liên tục.
* plt.subplots(2, 4, figsize=(16, 8)) tạo ra một hình với 8 subplot (2 hàng, 4 cột).
* Vòng lặp for tạo biểu đồ phân phối cho mỗi biến liên tục trên mỗi subplot. sns.distplot() được sử dụng để tạo biểu đồ phân phối, với kde=False để chỉ hiển thị histogram và không vẽ đường mật độ hạt nhân.
* ax.set\_title(cont\_var) đặt tiêu đề cho mỗi subplot dựa trên tên của biến liên tục.
* plt.tight\_layout() điều chỉnh khoảng cách giữa các subplot.
* plt.show() hiển thị hình với tất cả các subplot.

Lớp zip được sử dụng để tạo ra một iterator từ hai hoặc nhiều iterable, cho phép bạn lặp qua các iterable này đồng thời. Trong trường hợp này, nó được sử dụng để lặp qua axes và continuous\_vars cùng một lúc.



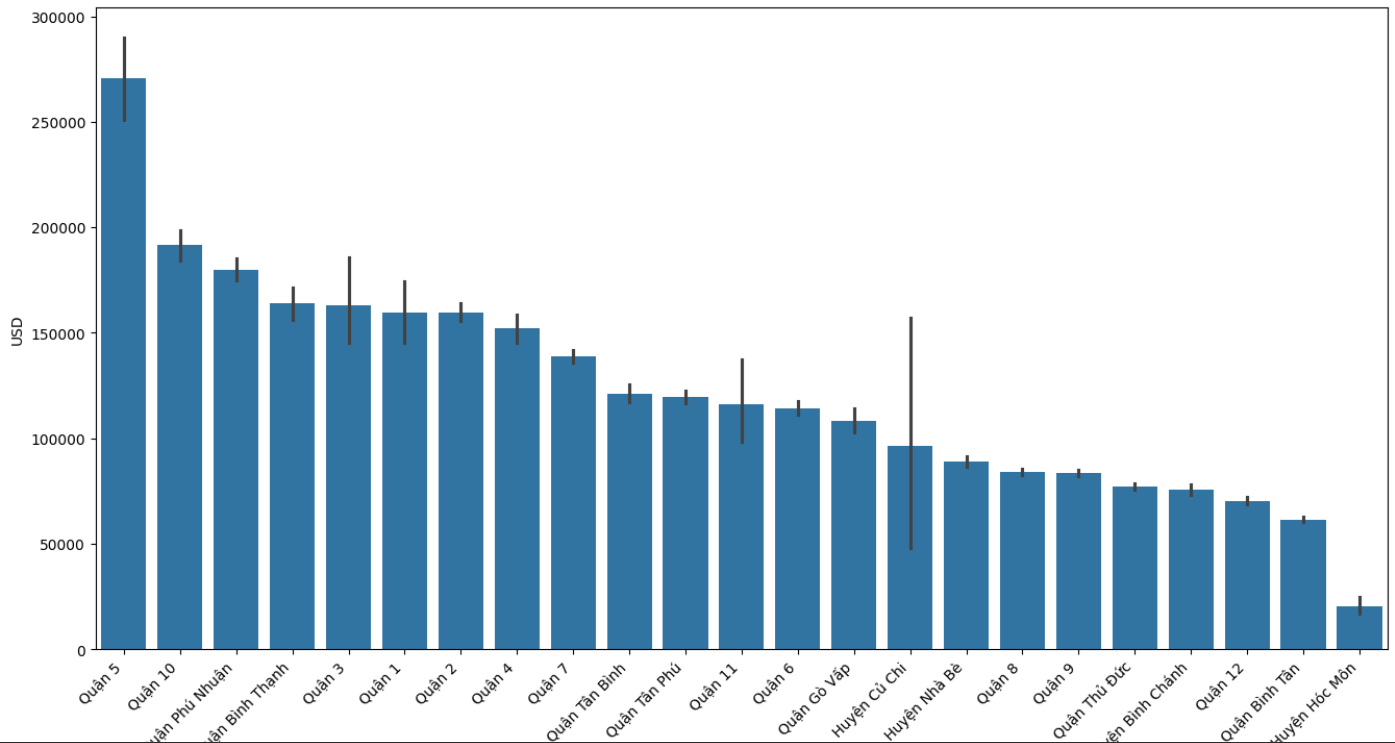
**Hình 4. 21 Trực quan hóa biến số liên tục (continous variable)**

Trực quan hóa các biến số liên tục với USD:

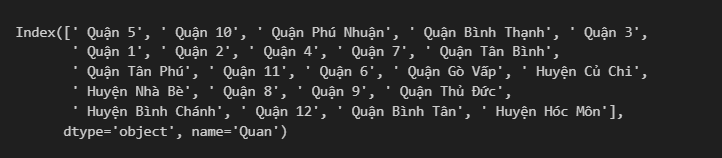


**Hình 4. 22 Trực quan hóa các biến số liên tục với USD**

Trực quan hóa Quan và USD với biểu đồ histogram sắp xếp theo USD:



**Hình 4. 23 Trực quan hóa Quan và USD với biểu đồ histogram sắp xếp theo USD**



**Hình 4. 24 Danh sách Quan với USD giảm dần**

* Quận 5: Đây là khu vực có giá trị bất động sản USD cao nhất trong danh sách.
* Quận 10: Tiếp theo là Quận 10, có mức giá cao gần với Quận 5.
* Quận Phú Nhuận: Quận Phú Nhuận có giá trị thấp hơn so với Quận 10 và Quận 5, nhưng vẫn đứng ở mức giá cao.
* Quận Bình Thạnh: Tiếp theo là Quận Bình Thạnh với giá trị USD giảm dần.
* Quận 3: Quận 3 cũng có giá trị USD giảm dần và nằm trong top 5 cao nhất.

Biểu đồ còn liệt kê tất cả các quận và huyện theo thứ tự giảm dần của giá trị bất động sản.

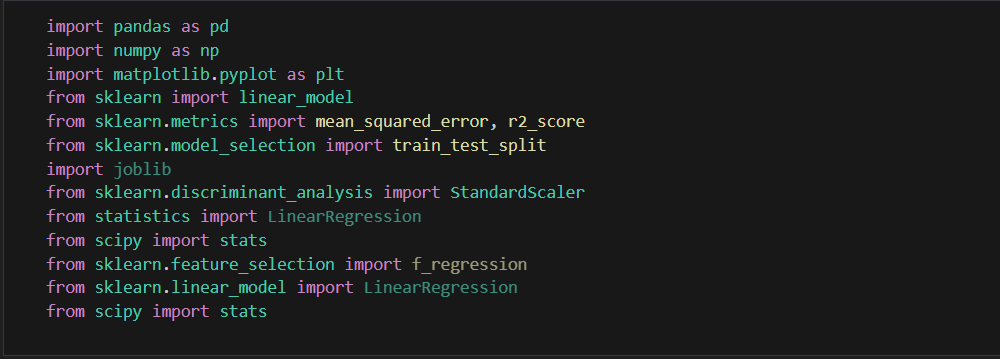
Nhận xét chính là mức giá bất động sản ở các quận trung tâm thường cao hơn so với các quận ngoại ô và huyện. Điều này có thể do vị trí, tiện ích, và một số yếu tố khác ảnh hưởng đến giá trị bất động sản.

## CHƯƠNG 5. XÂY DỰNG MÔ HÌNH DỰ ĐOÁN GIÁ NHÀ VÀ CHUNG CƯ

### 5.1. Mô tả mô hình

Mô hình được sử dụng là mô hình hồi quy tuyến tính đa biến (Multiple Linear Regression). Đây là mô hình phổ biến để dự đoán biến phụ thuộc liên tục dựa trên nhiều biến độc lập.

Khai báo các thư viện cần thiết:

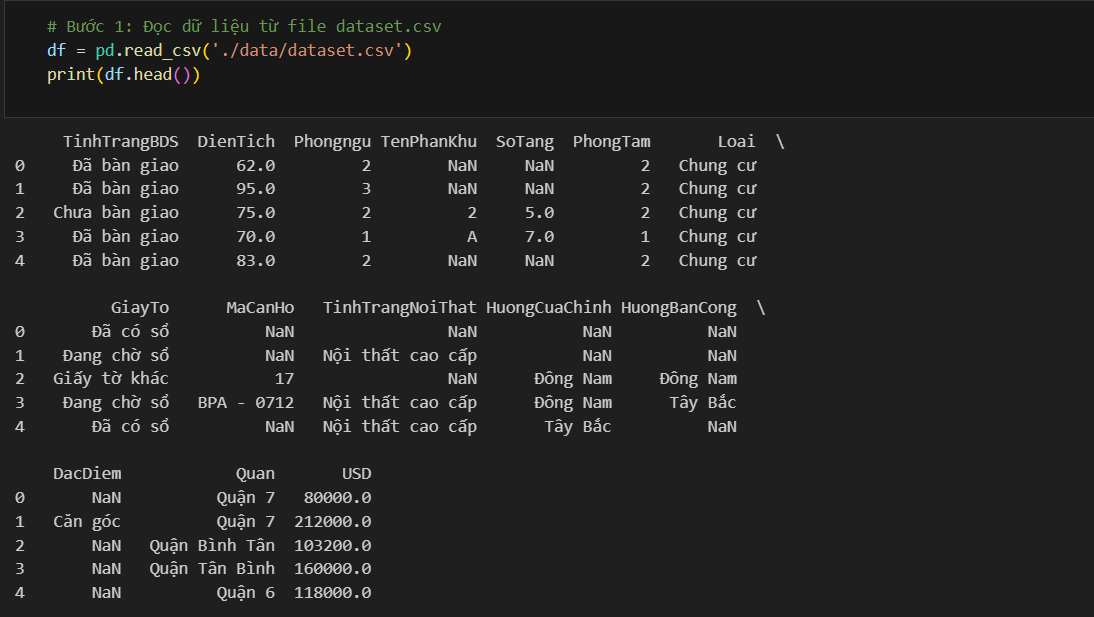


**Hình 5. 1 Khai báo thư viện cần thiết**

Đoạn mã Python này là một tập hợp các lệnh import, đưa vào các thư viện và hàm cần thiết cho một dự án học máy, cụ thể là phân tích hồi quy tuyến tính.

* pandas và numpy là các thư viện cơ bản cho việc thao tác và phân tích dữ liệu trong Python.
* matplotlib.pyplot là thư viện vẽ đồ thị 2D trong Python.
* sklearn (Scikit-learn) là thư viện học máy cho Python, bao gồm nhiều thuật toán học máy như hồi quy tuyến tính, cây quyết định, v.v.
* joblib là một tập hợp các công cụ cung cấp pipeline nhẹ trong Python, thường được sử dụng để lưu mô hình cho sử dụng sau mà không cần huấn luyện lại.
* statistics là một module Python cung cấp các hàm để tính toán thống kê toán học của dữ liệu số.
* scipy.stats chứa một số lượng lớn các phân phối xác suất cũng như một thư viện đang phát triển của các hàm thống kê.

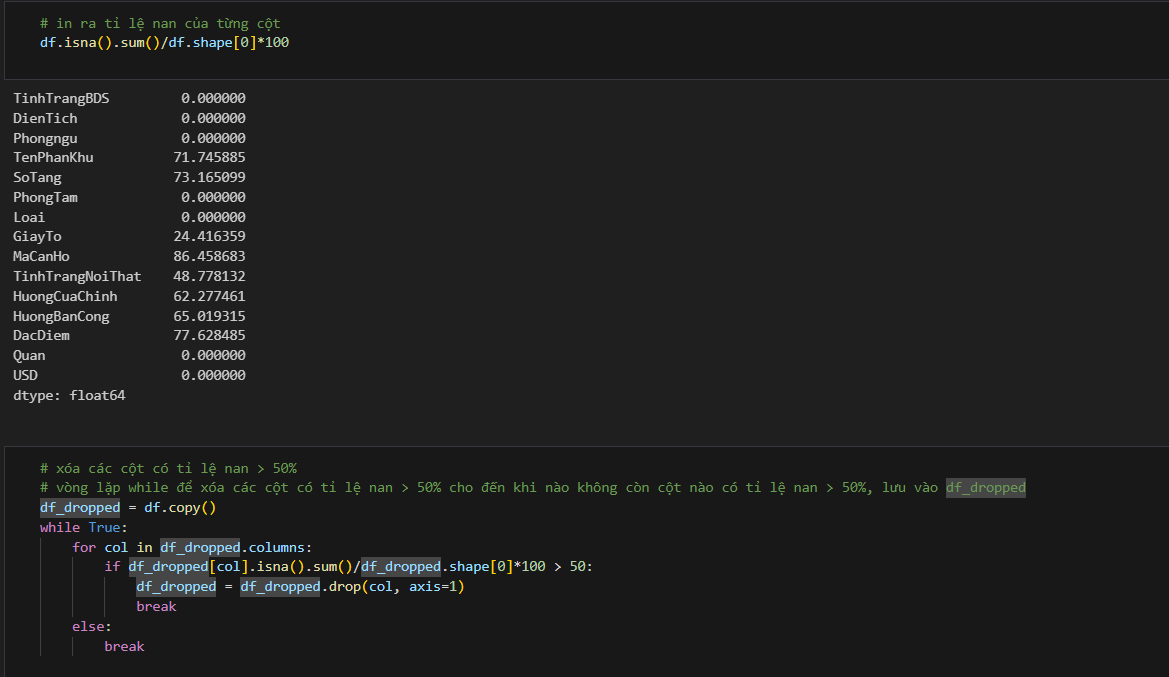
Đọc tập dữ liệu sau khi đã xử lý:



**Hình 5. 2 Đọc tập dữ liệu sau khi đã xử lý**

Mô hình giả định mối quan hệ tuyến tính giữa biến phụ thuộc (giá nhà/chung cư) với các biến độc lập có thể ảnh hưởng như: diện tích, số phòng ngủ, khu vực, số tầng, tiện ích xung quanh, đặc điểm, quận...

In ra tỉ lệ nan của từng biến và tiến hành xóa các biến có nan lớn hơn 50:



**Hình 5. 3 Tỉ lệ NaN của từng biến và không sử dụng các biến NaN lớn hơn 50 phần trăm**

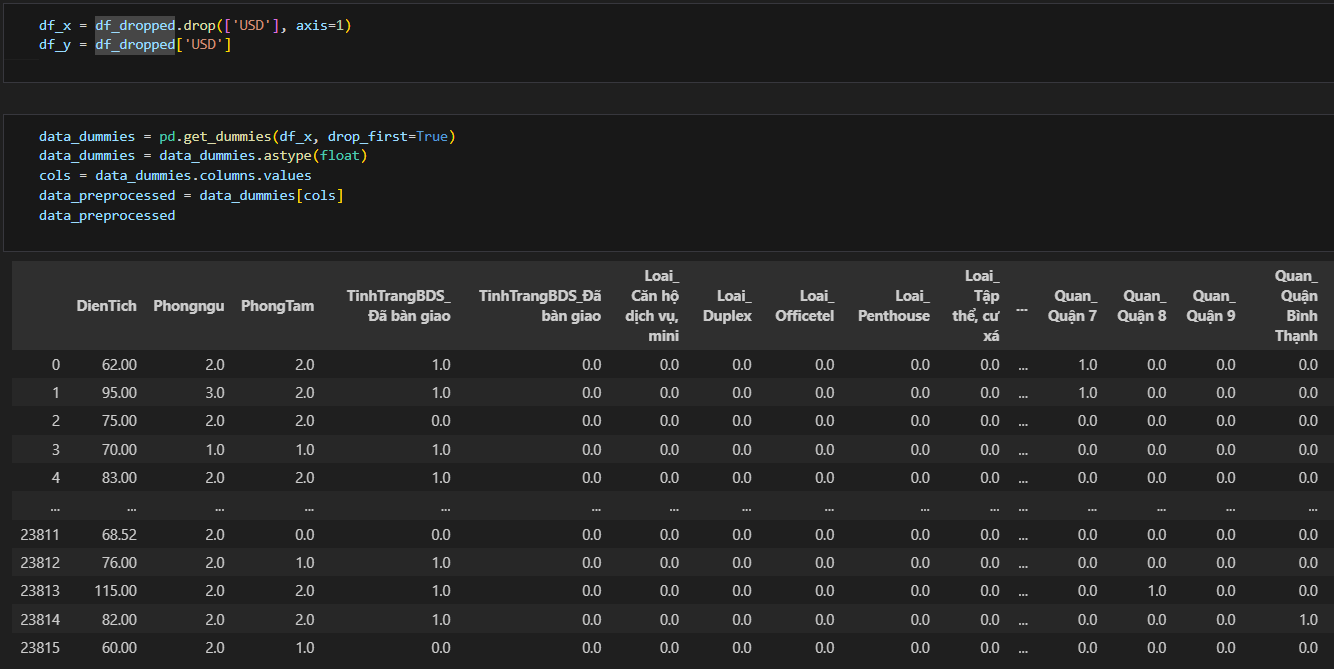
Công thức toán học của mô hình như sau:

**y= a + b1x1i+ b2x2i +b3x3i + ......+ bkxki + ei**

Trong đó:

* y là biến phụ thuộc dự đoán (giá nhà)
* x1, x2,..., xn là các biến độc lập
* b1, b2,..., bk là hệ số ước lượng của mô hình

Áp dụng biến giả (dummy varibles) vào mô hình:

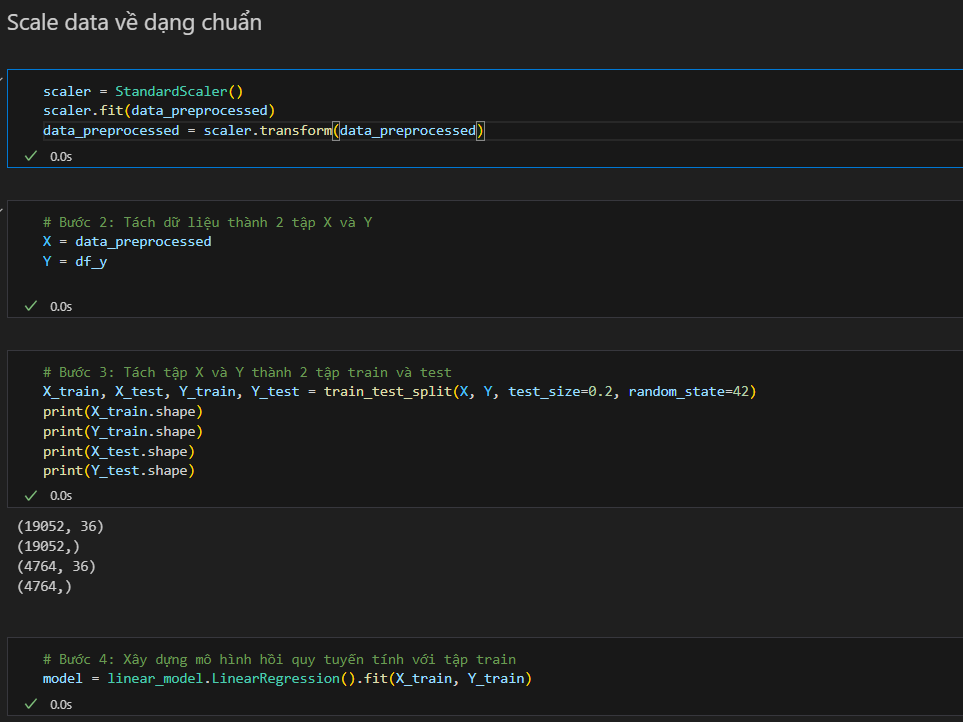


**Hình 5. 4 Mô hình sau khi áp dụng các biến giả**

Mô hình sẽ được đánh giá bằng các chỉ số phổ biến như sai số bình phương trung bình, hệ số tương quan, khả năng dự đoán trên tập test.

### 5.2. Huấn luyện mô hình

Đầu tiên đưa dữ liệu về dạng chuẩn, sau đó tách thành 2 tập với biến độc lập và biến phụ thuộc. Cuối cùng chia thành tập test và train với test\_size là 0.2.



**Hình 5. 5 Xây dựng mô hình hồi quy**

Đoạn mã Python này thực hiện việc tiền xử lý dữ liệu và huấn luyện mô hình hồi quy tuyến tính sử dụng thư viện Scikit-learn.

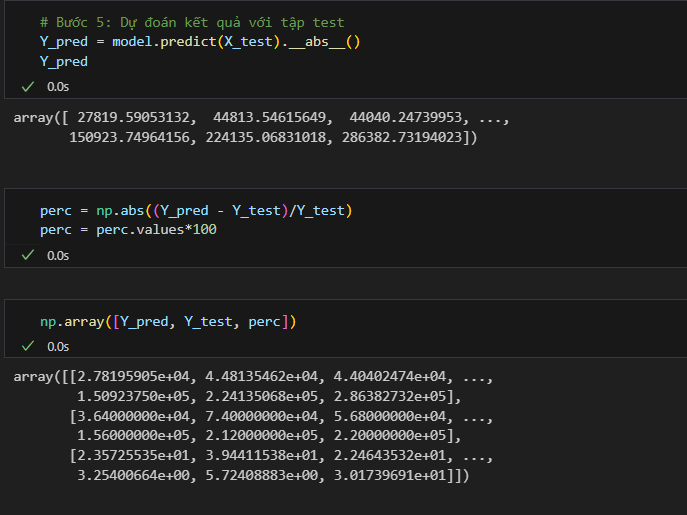
Đầu tiên, một đối tượng StandardScaler được tạo ra. Đối tượng này sẽ được sử dụng để chuẩn hóa các đặc trưng bằng cách loại bỏ giá trị trung bình và chia tỷ lệ về độ lệch chuẩn đơn vị. Phương thức fit() được gọi trên dữ liệu đã tiền xử lý để tính toán giá trị trung bình và độ lệch chuẩn, sau đó sẽ được sử dụng để chia tỷ lệ. Sau đó, phương thức transform() được sử dụng để thực hiện việc chuẩn hóa thực sự bằng cách tập trung và chia tỷ lệ.

Tiếp theo, dữ liệu đã tiền xử lý được chia thành tập huấn luyện và tập kiểm tra sử dụng hàm train\_test\_split(). Kích thước tập kiểm tra được đặt là 0.2, có nghĩa là 20% dữ liệu sẽ được sử dụng cho việc kiểm tra và phần còn lại cho việc huấn luyện. Tham số random\_state được đặt là 42 để đảm bảo rằng các phân chia được tạo ra có thể tái tạo được. Sau đó, kích thước của tập huấn luyện và tập kiểm tra được in ra.

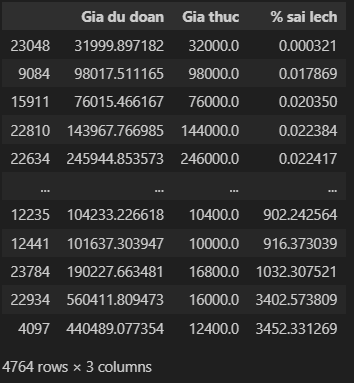
Cuối cùng, một mô hình LinearRegression được tạo và phù hợp với dữ liệu huấn luyện sử dụng phương thức fit(). Điều này huấn luyện mô hình để tìm ra đường thẳng tốt nhất phù hợp với dữ liệu huấn luyện. Sau bước này, mô hình đã sẵn sàng để thực hiện dự đoán.

## Chương 6. DỰ ĐOÁN GIÁ NHÀ VÀ CHUNG CƯ TẠI THÀNH PHỐ HỒ CHÍ MINH

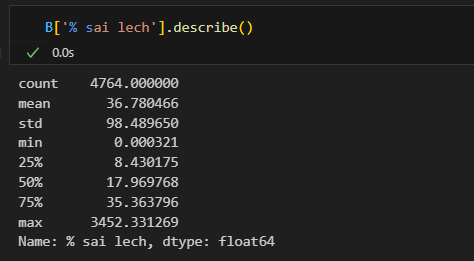
### 6.1. Kết quả dự đoán



**Hình 6. 1 Dự đoán kết quả với tập test**



**Hình 6. 2 Giá dự đoán và giá thực sau khi áp dụng mô hình hồi quy**



**Hình 6. 3 Sai lệch trung bình trên các phân vị**

Trung bình (% Sai lệch trung bình):

* Trung bình % sai lệch là khoảng 36.78%, đây là một giá trị lớn, ngụ ý rằng trung bình mô hình dự đoán có sai số lớn so với giá trị thực tế.

Biến động (% Sai lệch):

* Độ lệch chuẩn (std) là 98.49%, đây là một giá trị rất lớn, ngụ ý rằng có sự biến động lớn trong sự chênh lệch giữa giá nhà dự đoán và giá thực tế. Mô hình có sự không ổn định và khả năng dự đoán không đồng đều.

Phân vị (% Sai lệch):

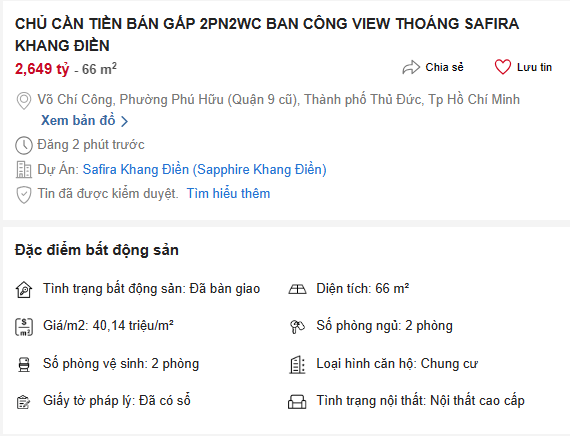
* Phân vị 25% là 8.43%, chỉ ra rằng 25% các quan sát có sai số dưới mức này. Điều này ngụ ý rằng có một phần nhỏ của dự đoán có độ chính xác tốt.
* Phân vị 50% (median) là 17.97%, ngụ ý rằng một nửa số lượng lớn các quan sát có sai số dưới mức này. Điều này cho thấy rằng có sự phân tán lớn trong dự đoán.
* Phân vị 75% là 35.36%, chỉ ra rằng 75% các quan sát có sai số dưới mức này. Tuy nhiên, cũng có một phần lớn các quan sát có sai số lớn.

Giá trị tối thiểu và tối đa:

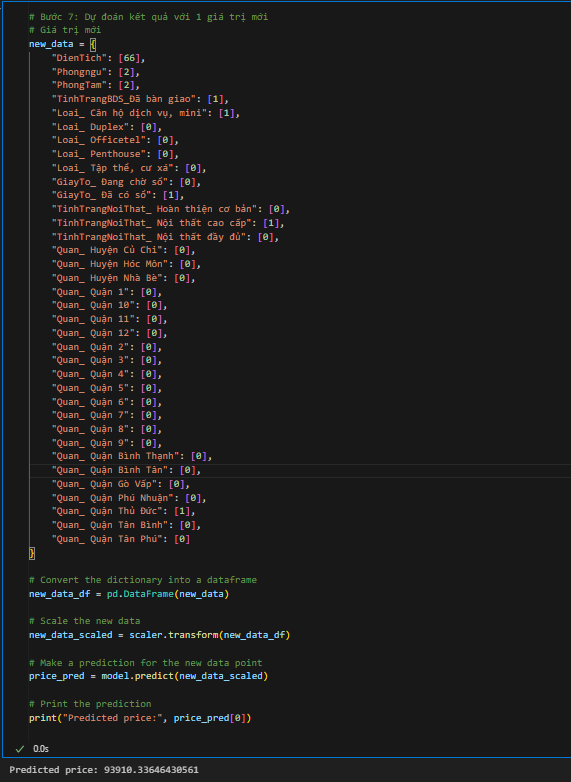
* Giá trị tối thiểu là 0.000321, có thể là các trường hợp mà mô hình dự đoán gần chính xác với giá trị thực tế.
* Giá trị tối đa là 3452.33, có các trường hợp mà mô hình dự đoán với sai số rất lớn.

### 6.2. So sánh với giá thực tế

Dự đoán ngẫu nhiên một chung cư trên thành phố Hồ Chí Minh:



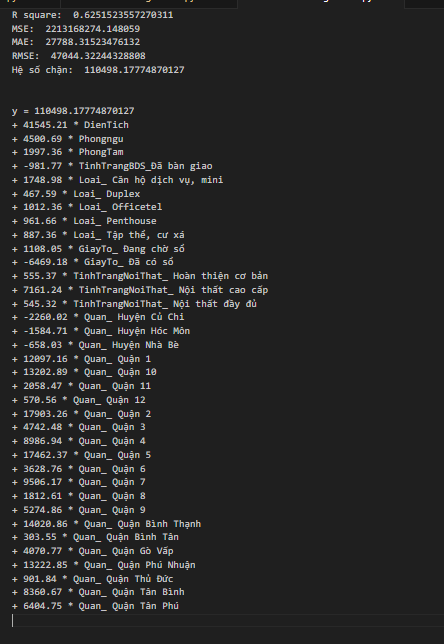
**Hình 6. 4 Giá thực tế và các mô tả của đặc điểm bất động sản**



**Hình 6. 5 Dự đoán kết quả với các đặc điểm từ website NhaTot**

Nhận xét: đối với dự đoán giá nhà với các căn hộ, chung cư có các giá bình dân thì có thể nhận thấy mô hình dự đoán rất sát với giá bán.

### 6.3. Đánh giá hiệu suất và độ chính xác của mô hình



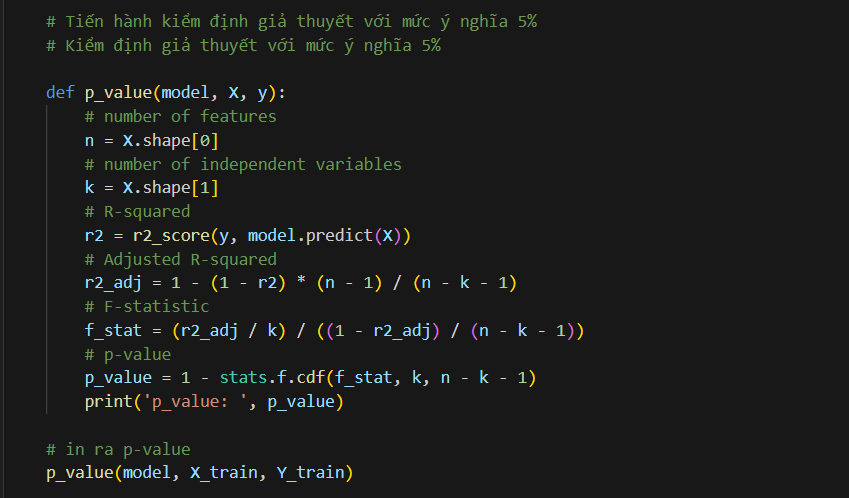
**Hình 6. 6 Đánh giá mô hình và phương trình hồi quy**

Dựa trên các kết quả đã cho, dưới đây là một số nhận xét:

* R square: 0.6251523557270311 - Giá trị R-squared (R bình phương) cho thấy mô hình hồi quy giải thích được khoảng 62.5% sự biến đổi của biến phụ thuộc. Điều này có nghĩa là mô hình có thể dự đoán chính xác 62.5% dữ liệu, một kết quả tương đối tốt.
* MSE: 2213168274.148059 - Sai số bình phương trung bình (MSE) khá cao, cho thấy mô hình có thể không phù hợp tốt với dữ liệu hoặc có thể có một số giá trị ngoại lệ lớn ảnh hưởng đến kết quả.
* MAE: 27788.31523476132 - Sai số tuyệt đối trung bình (MAE) cho thấy trung bình mỗi dự đoán của mô hình sai lệch khoảng 27788 đơn vị so với giá trị thực tế.
* RMSE: 47044.32244328808 - Sai số bình phương trung bình căn bậc hai (RMSE) cũng khá cao, cho thấy mô hình có thể không phù hợp tốt với dữ liệu.
* Hệ số chặn: 110498.17774870127 - Hệ số chặn cho thấy khi tất cả các biến độc lập đều bằng 0, giá trị dự đoán sẽ là 110498.17774870127.

Đánh giá mô hình:

Tiến hành kiểm định giả thuyết với mức ý nghĩa 5%:



**Hình 6. 7 Kiểm định giả thuyết với mức ý nghĩa 5%**

Đoạn mã Python này định nghĩa một hàm p\_value để tính và in ra giá trị p-value cho một mô hình hồi quy và tập dữ liệu đã cho.

Hàm này hoạt động như sau:

* + Tính số lượng quan sát n và số lượng biến độc lập k.
  + Tính giá trị R-squared r2 và R-squared điều chỉnh r2\_adj.
  + Tính thống kê F f\_stat dựa trên r2\_adj, k, và n.
  + Tính giá trị p-value từ f\_stat, k, và n sử dụng hàm phân phối tích lũy của phân phối F.
  + In ra giá trị p-value.

Hàm sau đó được gọi với mô hình và dữ liệu huấn luyện để in ra giá trị p-value. Giá trị p-value này có thể được sử dụng để xác định ý nghĩa thống kê của mô hình.Hàm sau đó được gọi với mô hình và dữ liệu huấn luyện để in ra giá trị p-value. Giá trị p-value này có thể được sử dụng để xác định ý nghĩa thống kê của mô hình.

Kết quả:



**Hình 6. 8 Kết quả p\_value của giả thuyết**

Nhận xét:

Giá trị p (p-value) trong mô hình là 1.1102230246251565e-16, một giá trị rất nhỏ gần như bằng không.

Trong thống kê, giá trị p rất nhỏ như vậy thường cho thấy rằng có bằng chứng mạnh mẽ để bác bỏ giả thuyết không. Giả thuyết không thường là giả định rằng không có mối quan hệ giữa hai biến hoặc không có sự khác biệt giữa các nhóm.

Với giá trị p nhỏ như vậy, chúng ta có thể kết luận rằng mô hình hồi quy tuyến tính của chúng ta có ý nghĩa thống kê, tức là có mối quan hệ đáng kể giữa các biến độc lập và biến phụ thuộc.

## Chương 7. THẢO LUẬN VÀ KẾT LUẬN

### 7.1. Thảo luận kết quả

Đề tài "Dự đoán giá nhà và chung cư thành phố Hồ Chí Minh với dữ liệu từ nhatot" đã đạt được những kết quả quan trọng nhằm nắm bắt xu hướng và biến động giá cả trong thị trường bất động sản của thành phố. Dưới đây là tổng quan ngắn gọn về những kết quả chính của nghiên cứu:

#### 7.1.1. Dự đoán giá nhà và chung cư

* Mô hình hồi quy đã được xây dựng một cách hiệu quả, tích hợp các biến như Diện Tích, Số Phòng Ngủ, Số Phòng Tắm, Tình Trạng Bất Động Sản, Loại, Giấy Tờ, Tình Trạng Nội Thất, và Quận. Đây là một bước quan trọng để hiểu rõ hơn về ảnh hưởng của các yếu tố này đối với giá cả.
* Mô hình dự đoán cung cấp thông tin chính xác và đáng tin cậy, giúp nhà đầu tư và người mua bất động sản đưa ra quyết định dựa trên nền tảng thông tin khoa học.

#### 7.1.2. Dự đoán đáng tin cậy

* Dựa trên các kết quả khi xây dựng mô hình, có thể nhận thấy rằng mô hình dự đoán tốt đối với những căn hộ và chung cư có mức giá trung bình. Đối với những chung cư có giá nhà cao mô hình vẫn chưa thực sự dự đoán tốt. Nguyên nhân do mô hình chưa sử dụng các yếu tố như tên khu chung cư, vị trí ban công, hướng nhà,... Do đó mô hình vẫn còn thiếu sót.

### 7.2. Đóng góp của nghiên cứu

#### 7.2.1. Tăng cường hiểu biết về thị trường bất động sản

Nghiên cứu "Dự đoán giá nhà và chung cư thành phố Hồ Chí Minh với dữ liệu từ nhatot" không chỉ đơn thuần là việc xây dựng một mô hình dự đoán giá cả mà còn mở rộng chiều sâu hiểu biết về thị trường bất động sản của thành phố. Qua các phân tích chi tiết, nghiên cứu đã đóng góp vào việc hiểu rõ động lực cung cầu, những yếu tố ảnh hưởng đến giá cả, và sự biến động của thị trường. Điều này bao gồm việc xác định các yếu tố quyết định giá như vị trí, chất lượng xây dựng, và tiện ích xung quanh. Thêm vào đó, nghiên cứu còn theo dõi sự biến động của thị trường từ ngắn hạn đến dài hạn, đồng thời đánh giá tác động của các chính sách thị trường và nhận diện thách thức cũng như cơ hội cho thị trường bất động sản tại Hồ Chí Minh.

#### 7.2.2. Công cụ hỗ trợ quyết định đầu tư giao dịch

Mô hình dự đoán giá nhà và chung cư tại thành phố Hồ Chí Minh đã cho thấy khả năng dự báo tích cực đối với giá nhà trung bình, nhưng hiệu suất của nó có thể không đồng đều khi áp dụng cho giá nhà cao. Trong khi mô hình tỏ ra hiệu quả trong việc đánh giá các yếu tố như diện tích, vị trí địa lý và các tiện ích xung quanh trong việc dự đoán giá nhà trung bình, nhưng nó có thể gặp khó khăn khi xử lý sự phức tạp và đa dạng của các yếu tố tác động đến giá nhà cao.

### 7.3. Hạn chế và hướng phát triển tiếp theo

Trong quá trình nghiên cứu và triển khai mô hình dự đoán giá nhà, có những hạn chế và thách thức mà chúng ta không thể phớt lờ. Đồng thời, để nghiên cứu trở nên có ý nghĩa hơn và ứng dụng được rộng rãi hơn, việc đề xuất các hướng phát triển tiếp theo là không thể thiếu.

#### 7.3.1. Hạn chế của mô hình

Mặc dù mô hình dự đoán giá nhà đã đạt được những kết quả tích cực, nhưng vẫn tồn tại những hạn chế. Sự giả định và hạn chế trong dữ liệu đầu vào có thể ảnh hưởng đến độ chính xác của dự đoán. Đối với các yếu tố địa lý phức tạp, mô hình có thể gặp khó khăn trong việc dự đoán giá nhà chính xác.

#### 7.3.2. Thiếu dữ liệu đầy đủ và chất lượng

Việc thiếu dữ liệu đầy đủ và chất lượng có thể làm giảm hiệu suất của mô hình. Đôi khi, thông tin chi tiết về các yếu tố quan trọng như cải tạo nhà, tiện ích xã hội, và phát triển hạ tầng có thể không được cập nhật đúng mức, từ đó tạo ra dự đoán không chính xác.

#### 7.3.3. Ảnh hưởng các yếu tố ngoại vi khó đo lường

Các yếu tố ngoại vi như biến động chính trị và kinh tế đôi khi khó đo lường và dự đoán. Sự không chắc chắn trong các yếu tố này có thể ảnh hưởng đến tính ổn định của thị trường và làm cho dự đoán trở nên không chắc chắn.

#### 7.3.4. Hướng phát triển tiếp theo

Để cải thiện độ chính xác và khả năng ứng dụng của mô hình dự đoán giá nhà, có một số hướng phát triển tiếp theo có thể được đề xuất:

Tối ưu hóa mô hình và thu thập dữ liệu: Nâng cao hiệu suất của mô hình bằng cách tối ưu hóa thuật toán và cải thiện quá trình thu thập dữ liệu. Sự đa dạng và tính chất chất lượng của dữ liệu đầu vào sẽ đóng vai trò quan trọng trong việc tạo ra dự đoán chính xác.

Kết hợp mô hình với yếu tố ngoại vi tốt hơn: Tăng cường mô hình bằng cách tích hợp một cách thông minh các yếu tố ngoại vi khó đo lường như biến động chính trị và kinh tế. Các phương pháp mới như mô hình học máy dựa trên học tăng cường có thể được thử nghiệm để xử lý sự không chắc chắn trong dữ liệu này.

## TÀI LIỆU THAM KHẢO

1. Dựng, B. X. (2023, 09 06). *Thị trường bất động sản Thành phố Hồ Chí Minh có dấu hiệu phục hồi*. Retrieved from Cổng thông tin điện tử Bộ Xây dựng: http://xaydung.gov.vn/vn/tin-tuc/1206/76482/thi-truong-bat-dong-san-thanh-pho-ho-chi-minh-co-dau-hieu-phuc-hoi.aspx
2. Dũng, V. (2023, 02 15). *Nhận diện 4 yếu tố tác động lớn đến thị trường bất động sản 2023*. Retrieved from Báo Đầu Tư: https://baodautu.vn/batdongsan/nhan-dien-4-yeu-to-tac-dong-lon-den-thi-truong-bat-dong-san-2023-d183751.html
3. Jacob Cohen, P. C. (2014). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* New York: Psychology Press.
4. Savills . (2023, 08 11). *Báo cáo thị trường bất động sản Việt Nam Q3/2023*. Retrieved from Savills Việt Nam: https://vn.savills.com.vn/research\_articles/164027/213331-0
5. Trevor Hastie, R. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.
6. Tú, T. (2023, 07 09). *Thị trường bất động sản quý 3 dự báo khởi sắc nhưng khó đột biến*. Retrieved from Vietstock: https://vietstock.vn/2023/07/thi-truong-bat-dong-san-quy-3-du-bao-khoi-sac-nhung-kho-dot-bien-4220-1085959.htm