**TRƯỜNG ĐẠI HỌC KINH TẾ**

**ĐẠI HỌC ĐÀ NẴNG**

….🖎…🕮…✍….



**BÁO CÁO**

**ĐỀ TÀI: REGRESSION**

**HỌC PHẦN: PHÂN TÍCH DỮ LIỆU BẰNG PYTHON**

GVHD : Lê Diên Tuấn

Nhóm : 2

Sinh viên thực hiện : Nguyễn Thị Mùi

Nguyễn Thị Thu Hoài

Nguyễn Thị Tiên

Nguyễn Minh Phượng

Lê Âu Anh Thư

*Đà Nẵng, 2024*

**MỤC LỤC**

[I. Đánh giá 3](#_Toc163314085)

[II. Lời cảm ơn 2](#_Toc163314086)

[III. GIỚI THIỆU TỔNG QUAN 2](#_Toc163314087)

[1. Giới thiệu về bài báo cáo 2](#_Toc163314088)

[2. Mục tiêu của bài báo cáo 1](#_Toc163314089)

[3. Đối tượng và phạm vi bài báo cáo 1](#_Toc163314090)

[4. Phương pháp thực hiện bài báo cáo 1](#_Toc163314091)

[IV. Cơ sở lý thuyết 2](#_Toc163314092)

[1. Tổng quan về hồi quy 2](#_Toc163314093)

[1.1 Hồi quy tuyến tính đơn bội 2](#_Toc163314094)

[1.2 Hồi quy tuyến tính đa bội 3](#_Toc163314095)

[1.3 Hồi quy logistics 4](#_Toc163314096)

[V. Dữ liệu 6](#_Toc163314097)

[VI. Framework 7](#_Toc163314098)

[VII. Chi tiết quy trình triển khai bằng python 7](#_Toc163314099)

[1. Tiền xử lý dữ liệu 7](#_Toc163314100)

[2. Khám phá và phân tích 12](#_Toc163314101)

[2.1 Thống kê 12](#_Toc163314102)

[2.2 . Trực quan hóa dữ liệu 14](#_Toc163314103)

[3. REG 19](#_Toc163314104)

[3.1 Hồi quy tuyến tính đơn bội 19](#_Toc163314105)

[3.2 Hồi quy tuyến tính đa bội 26](#_Toc163314106)

[4. Kết quả nghiên cứu. 32](#_Toc163314107)

[VIII. Video 33](#_Toc163314108)

[TÀI LIỆU THAM KHẢO 33](#_Toc163314109)

**DANH MỤC HÌNH ẢNH**

[Hình 1.Các mối quan hệ của hồi quy tuyến tính đơn bội 3](#_Toc163314180)

[Hình 2.Mô hình hồi quy logistic 4](#_Toc163314181)

[Hình 3.Dữ liệu 6](#_Toc163314182)

[Hình 4.Framework 7](#_Toc163314183)

[Hình 5. Code đọc dữ liệu 8](#_Toc163314184)

[Hình 6. Kết quả đọc dữ liệu 8](#_Toc163314185)

[Hình 7. Kiểm tra các giá trị null 9](#_Toc163314186)

[Hình 8. Thông tin các cột dữ liệu 10](#_Toc163314187)

[Hình 9. Xem outliers 10](#_Toc163314188)

[Hình 10. Loại bỏ outlier bằng IQR 11](#_Toc163314189)

[Hình 11.Đồ thị boxplot sau loại bỏ outliers 11](#_Toc163314190)

[Hình 12. Mô tả dữ liệu sau khi loại bỏ outliers 12](#_Toc163314191)

[Hình 13. Thống kê dữ liệu 12](#_Toc163314192)

[Hình 14. Biểu đồ giá nhà trung bình theo diện tích 14](#_Toc163314193)

[Hình 15. Biểu đồ giá nhà trung bình theo số lượng phòng ngủ 15](#_Toc163314194)

[Hình 16. Biểu đồ giá nhà trung bình theo nội thất 16](#_Toc163314195)

[Hình 17. Biểu đồ giá nhà lớn nhất và giá nhà nhỏ nhất theo số tầng 17](#_Toc163314196)

[Hình 18. Diện tích trung bình của các tầng 17](#_Toc163314197)

[Hình 19. Biểu đồ giá cả trung bình theo tầng 18](#_Toc163314198)

[Hình 20. Tạo bản sao của dữ liệu 19](#_Toc163314199)

[Hình 21. Lấy 2 cột dữ liệu area và price 20](#_Toc163314200)

[Hình 22. In ra cột price 20](#_Toc163314201)

[Hình 23. Thêm hệ số chặn 21](#_Toc163314202)

[Hình 24. Kiểm định giả thuyết 21](#_Toc163314203)

[Hình 25. Biểu đồ phân tán trực quan 2 biến area và price 22](#_Toc163314204)

[Hình 26. Biểu đồ boxplot 23](#_Toc163314205)

[Hình 27. Code đặt biến giả đổi kiểu dữ liệu 23](#_Toc163314206)

[Hình 28. Tính ma trận tương quan 23](#_Toc163314207)

[Hình 29. Đưa 2 biến vào data Frame 24](#_Toc163314208)

[Hình 30. Xây dựng dự đoán 24](#_Toc163314209)

[Hình 31. Dự đoán giá nhà theo diện tích 25](#_Toc163314210)

[Hình 32. Vẽ biểu đồ scatter plot thể hiện mối quan hệ tuyến tính giữa 2 biến 25](#_Toc163314211)

[Hình 33. Tạo bản sao và xử lý biến giả 26](#_Toc163314212)

[Hình 34. In ra dữ liệu sau khi xử lý biến giả 26](#_Toc163314213)

[Hình 35. In ra hệ số tương quan 27](#_Toc163314214)

[Hình 36. Ma trận hệ số tương quan 27](#_Toc163314215)

[Hình 37. In ra biến độc lập 28](#_Toc163314216)

[Hình 38. In ra biến phụ thuộc 28](#_Toc163314217)

[Hình 39. Tạo intercept B0 29](#_Toc163314218)

[Hình 40. Code kiểm định giả thuyết 29](#_Toc163314219)

[Hình 41. In ra kết quả kiểm định giả thuyết 30](#_Toc163314220)

[Hình 42. Hiển thị thông tin tổng quan về dữ liệu 31](#_Toc163314221)

[Hình 43. Code so sánh giá trị dự đoán và giá trị thực tế 31](#_Toc163314222)

[Hình 44. In ra kết quả so sánh giá trị dự đoán và giá trị thực tế 31](#_Toc163314223)

[Hình 45. Dự đoán giá trị mới 32](#_Toc163314224)

# Đánh giá

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và Tên** | **Phần trăm đóng góp** |
| 1 | Lê Âu Anh Thư (Nhóm trưởng)  Công việc: Làm tổng quan, cơ sở lý thuyết, làm báo cáo, làm và sửa slide, xem nội dung bài làm và tổng kết | 100% |
| 2 | Nguyễn Minh Phượng  Công việc: Làm code hồi quy tuyến tính đơn, làm slide, tiền xử lý dữ liệu | 100% |
| 3 | Nguyễn Thị Mùi  Công việc: Quay video, làm code hồi quy tuyến tính bội, đưa ra kết quả nghiên cứu | 100% |
| 4 | Nguyễn Thị Tiên  Công việc: Vẽ framework, mô tả dữ liệu, đưa ra nhận xét, khám phá và phân tích | 100% |
| 5 | Nguyễn Thị Thu Hoài  Công việc: Làm tiền xử lý dữ liệu và trực quan hóa, tìm hiểu code hồi quy | 100% |

# Lời cảm ơn

**LỜI CAM ĐOAN**

Nhóm nghiên cứu xin cam đoan nội dung nghiên cứu **“Dự báo giá nhà ở Mỹ theo phương pháp hồi quy”** là sản phẩm mà nhóm 2 đã nỗ lực nghiên cứu dưới sự hướng dẫn của giảng viên Lê Diên Tuấn. Các kết quả trong báo cáo nghiên cứu này là nhóm 2 đã thu thập được từ trong quá trình học và tham khảo các tài liệu hướng dẫn của thầy Lê Diên Tuấn. Nhóm 2 xin chịu hoàn toàn trách nhiệm về tính chính xác của thông tin dữ liệu của báo cáo nghiên cứu này.

# GIỚI THIỆU TỔNG QUAN

## Giới thiệu về bài báo cáo

Trong thời đại hiện nay, thị trường bất động sản đang chứng kiến một cuộc tăng giá nhà vô cùng đáng kể. Sự tăng giá này đặt ra nhiều thách thức đối với các nhà đầu tư, chính phủ và các bên liên quan.Từ đó mà việc hiểu rõ và tận dụng thông tin của giá nhà trong quy mô giá nhà đa dạng như hiện nay là một khía cạnh đóng vai trò vô cùng quan trọng để nâng cao hiệu quả kinh doanh và tạo ra sự cạnh tranh bền vững cho thị trường bất động sản.Sự biến động của giá nhà phụ thuộc vào nhiều yếu tố, bao gồm kinh tế, chính trị, xã hội, cơ sở hạ tầng và nhu cầu của thị trường. Để hiểu và dự đoán xu hướng giá nhà, phân tích hồi quy đã trở thành một công cụ quan trọng.Thông qua việc áp dụng phương pháp phân tích hồi quy, nhóm hy vọng sẽ có cái nhìn sâu sắc về xu hướng giá nhà và các yếu tố ảnh hưởng đến nó trong thời đại tăng giá nhà vượt trội. Việc hiểu rõ và tận dụng thông tin về giá nhà trong quy mô giá nhà đa dạng như hiện nay đóng vai trò quan trọng để nâng cao hiệu quả kinh doanh và tạo ra sự cạnh tranh bền vững cho thị trường bất động sản. Trong đó, phương pháp hồi quy .đóng vai trò quan trọng để xác định mối quan hệ giữa giá nhà và các yếu tố ảnh hưởng, từ đó dự đoán và đánh giá giá trị tương lai của bất động sản. Phân tích hồi quy giúp chúng ta tìm ra những biến số quan trọng nhất và đo lường sự ảnh hưởng của chúng đối với giá nhà. Vì vậy, hiểu và áp dụng phương pháp phân tích hồi quy trong học phần phân tích dữ liệu giúp chúng ta có cái nhìn sâu sắc hơn về mối quan hệ giữa giá nhà và các yếu tố ảnh hưởng, cũng như khả năng dự đoán và đánh giá trong lĩnh vực bất động sản.

## Mục tiêu của bài báo cáo

* Tìm hiểu và phân tích xu hướng giá nhà trong thời đại tăng giá mạnh, cùng với các yếu tố ảnh hưởng đến giá nhà. Bằng cách áp dụng phương pháp phân tích hồi quy, bài báo cáo nhằm xác định mối quan hệ giữa giá nhà và các yếu tố quan trọng, từ đó dự đoán và đánh giá giá trị tương lai của bất động sản. Ngoài ra, bài báo cáo cũng nhấn mạnh vai trò quan trọng của việc hiểu rõ và tận dụng thông tin về giá nhà trong quy mô giá nhà đa dạng như hiện nay để nâng cao hiệu quả kinh doanh và tạo ra sự cạnh tranh bền vững cho thị trường bất động sản.

## Đối tượng và phạm vi bài báo cáo

* Đối tượng: Các nhà đầu tư, chính phủ và các bên liên quan trong lĩnh vực bất động sản
* Phạm vi:
* Đề tài được thực hiện dựa trên phạm vi tập dữ liệu giá nhà ở Mỹ
* Các mô hình hồi quy tuyến tính

## Phương pháp thực hiện bài báo cáo

* Xác định mô hình hồi quy: Lựa chọn phương pháp hồi quy phù hợp để phân tích mối quan hệ giữa giá nhà và các yếu tố ảnh hưởng. Sử dụng các phương pháp hồi quy tuyến tính, hồi quy đa biến phù hợp với dữ liệu và mục tiêu nghiên cứu.
* Thực hiện phân tích hồi quy: Áp dụng mô hình hồi quy vào dữ liệu để ước lượng các tham số và đánh giá mức độ ảnh hưởng của các yếu tố đối với giá nhà. Phân tích kết quả để hiểu rõ hơn về xu hướng giá nhà và đánh giá sự ảnh hưởng của các yếu tố

# Cơ sở lý thuyết

## Tổng quan về hồi quy

* Phân tích hồi quy là nghiên cứu sự phụ thuộc của một biến (biến phụ thuộc) vào một hay nhiều biến khác (biến độc lập), nhằm mục đích ước lượng (hay dự đoán) giá trị trung bình của biến phụ thuộc trên cơ sở các giá trị biết trước của các biến độc lập.
* Mục đích:
* Dự đoán: Hồi quy tuyến tính có thể được sử dụng để dự đoán giá trị của biến phụ thuộc dựa trên các giá trị của biến độc lập.
* Phân tích tương quan: Hồi quy tuyến tính có thể giúp xác định mức độ tương quan giữa hai biến số. Bằng cách xây dựng một mô hình hồi quy, bạn có thể kiểm tra xem liệu có một mối quan hệ tuyến tính giữa các biến hay không.
* Phân tích ảnh hưởng: Hồi quy tuyến tính cho phép bạn xác định độ ảnh hưởng của mỗi biến độc lập lên biến phụ thuộc. Bằng cách xem xét các hệ số hồi quy, bạn có thể đo lường và so sánh tác động của các biến và xác định xem biến nào có tác động lớn hơn hoặc nhỏ hơn.
* Dự báo và kế hoạch: Hồi quy tuyến tính có thể được sử dụng để xây dựng các mô hình dự báo và kế hoạch. Với dữ liệu lịch sử, bạn có thể sử dụng hồi quy tuyến tính để dự đoán xu hướng tương lai và đưa ra kế hoạch dựa trên dự báo đó.

=>Tóm lại, hồi quy tuyến tính là một công cụ mạnh mẽ cho phân tích dữ liệu và dự báo. Nó có thể được sử dụng trong nhiều lĩnh vực khác nhau như kinh tế, khoa học, y tế, xã hội học và nhiều lĩnh vực khác.

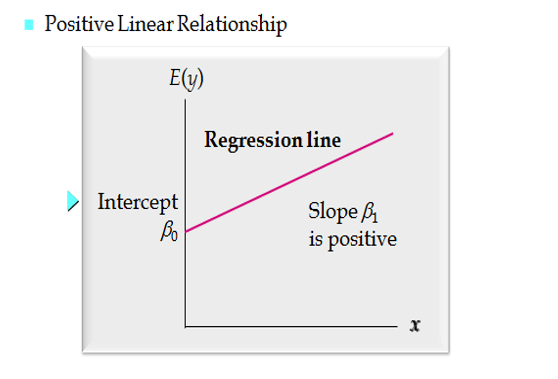
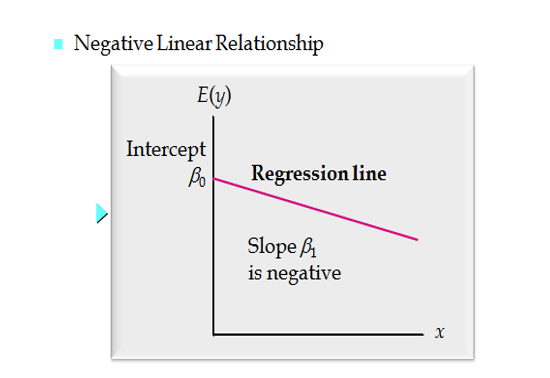
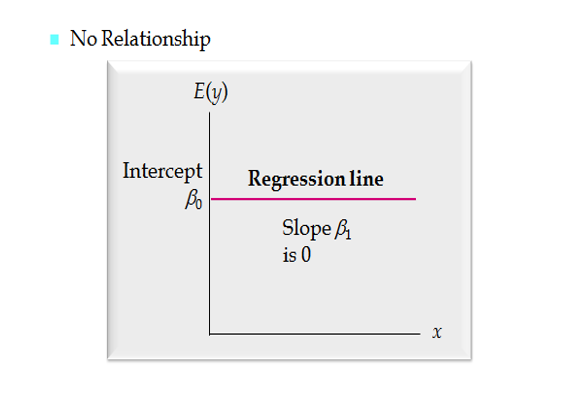
### . Hồi quy tuyến tính đơn bội

* Hồi quy tuyến tính đơn giản liên quan đến một biến độc lập và một biến phụ thuộc.
* Hồi quy tuyến tính đơn giản được định nghĩa bằng hàm tuyến tính:

Y= β0\*X + β1 + ε

* Y: Biến phụ thuộc.
* β0: Hệ số chặn (intercept).
* β1: Hệ số dốc (slope).
* X: Biến độc lập.
* ε: Sai số (error).
* Hệ số β0 β1:
* β0: Giá trị của Y khi X bằng 0.
* β1: Biểu thị mức độ thay đổi của Y khi X tăng lên 1 đơn vị.
* Các mối quan hệ giữa hai biến thể hiện bởi đường thẳng

+ Quan hệ thuận + Quan hệ nghịch + Không có quan hệ

Hình 1.Các mối quan hệ của hồi quy tuyến tính đơn bội

### . Hồi quy tuyến tính đa bội

* Định nghĩa:
* Phân tích hồi quy bội (Multiple Rsegression Analysis) là một phương pháp thống kê để xác định mối quan hệ giữa một biến phụ thuộc và hai hoặc nhiều biến độc lập.
* Trong phân tích hồi quy tuyến tính bội, tập dữ liệu chứa một biến phụ thuộc và nhiều biến độc lập. Hàm đường hồi quy tuyến tính thay đổi để bao gồm nhiều yếu tố như sau:

Y= β0\*X0 + β1\*X1 + β2\*X2+…… βn\*Xn+ ε

* Y: Biến phụ thuộc.
* β0: Hệ số chặn (intercept).
* β1, β2, ..., βk: Hệ số hồi quy cho các biến độc lập X1, X2, ..., Xk.
* X1, X2, ..., Xk: Biến độc lập.
* ε: Sai số (error).
* Hệ số β0 và βi:
* β0: Giá trị của Y khi tất cả các biến X bằng 0.
* βi: Biểu thị mức độ thay đổi của Y khi Xi tăng lên 1 đơn vị, giữ nguyên các biến X khác.
* Sai số ε:
* Biểu thị sự khác biệt giữa giá trị thực tế của Y và giá trị dự đoán bởi mô hình.
* Giả sử ε có phân phối chuẩn với trung bình bằng 0 và phương sai không đổi. Khi số lượng biến dự báo tăng lên, các hằng số β cũng tăng lên tương ứng.

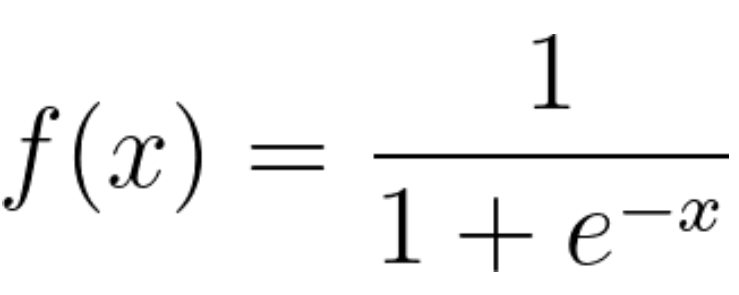
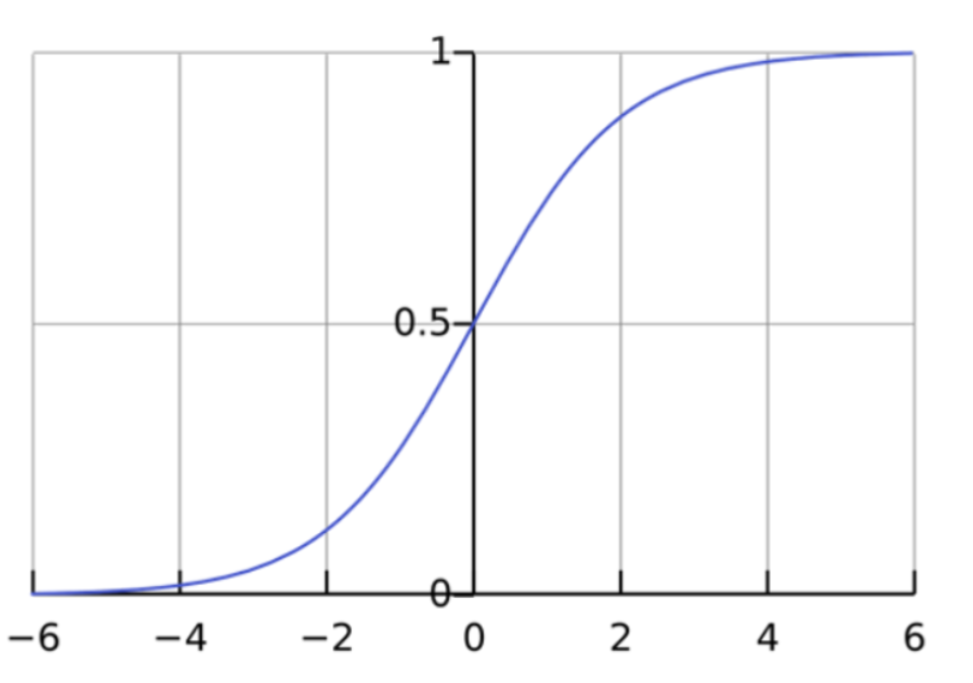
### . Hồi quy logistics

#### Khái niệm

* Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

#### Mô hình

* Hồi quy logistic là một mô hình thống kê sử dụng hàm logistic, hay hàm logit trong toán học làm phương trình giữa x và y. Hàm logit ánh xạ y làm hàm sigmoid của x.
* Phương trình hồi quy logistic

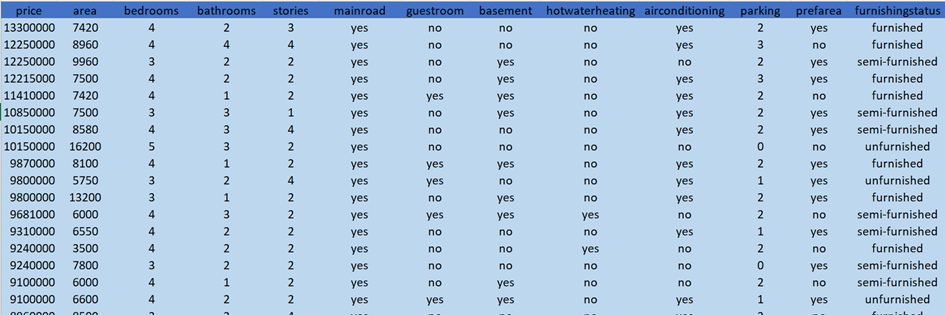
 

Hình 2.Mô hình hồi quy logistic

* Hàm logistic chỉ trả về các giá trị giữa 0 và 1 cho biến phụ thuộc, dù giá trị của biến độc lập là gì. Đây là cách hồi quy logistic ước tính giá trị của biến phụ thuộc. Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc.
* Có ba cách tiếp cận phân tích hồi quy logistic dựa trên kết quả của biến phụ thuộc.
* Hồi quy logistic nhị phân
* Hồi quy logistic nhị phân phù hợp với các vấn đề phân lớp nhị phân chỉ có hai kết quả có thể xảy ra. Biến phụ thuộc chỉ có thể có hai giá trị, chẳng hạn như có và không hoặc 0 và 1.
* Dù hàm logistic tính toán một phạm vi giá trị giữa 0 và 1, mô hình hồi quy nhị phân vẫn sẽ làm tròn kết quả đến các giá trị gần nhất. Nói chung, kết quả dưới 0,5 sẽ được làm tròn thành 0 và kết quả trên 0,5 sẽ được làm tròn thành 1, do đó hàm logistic trả về một kết quả nhị phân.
* Hồi quy logistic đa thức
* Hồi quy đa thức có thể phân tích các vấn đề có một số kết quả có thể xảy ra, miễn là số kết quả hữu hạn. Ví dụ: kỹ thuật này có thể dự đoán xem giá nhà sẽ tăng 25%, 50%, 75% hay 100% dựa trên dữ liệu dân số, nhưng sẽ không thể dự đoán được giá trị chính xác của một ngôi nhà.
* Hồi quy logistic đa thức hoạt động bằng cách ánh xạ các giá trị kết quả cho các giá trị khác nhau giữa 0 và 1. Hàm logistic có thể trả về một khoảng dữ liệu liên tục như 0,1, 0,11, 0,12, v.v., do đó hồi quy đa thức cũng nhóm đầu ra đến các giá trị gần nhất có thể có.
* Hồi quy logistic thứ tự
* Hồi quy logistic thứ tự, hay mô hình logit có thứ tự, là một loại hồi quy đa thức đặc biệt cho các vấn đề trong đó các số đại diện cho các bậc chứ không phải là giá trị thực tế. Ví dụ: bạn sẽ sử dụng hồi quy thứ tự để dự đoán đáp án cho câu hỏi khảo sát yêu cầu khách hàng đánh giá dịch vụ của bạn ở mức kém, ổn, tốt hoặc xuất sắc dựa trên một giá trị số, chẳng hạn như số lượng mặt hàng họ mua từ bạn trong năm.

# Dữ liệu

* Dữ liệu:



Hình 3.Dữ liệu

* Mô tả dữ liệu: Bộ dữ liệu này cung cấp thông tin toàn diện về dự đoán giá nhà ở Mỹ, với 13 cột dữ liệu, 545 bản ghi.Bộ dữ liệu bao gồm các thuộc tính khác nhau bao gồm Giá, Khu vực, Thành Phòng ngủ, Phòng tắm, Số tầng, Gần đường chính, Phòng khách, Tầng hầm, Nước nóng, Máy lạnh, Chỗ đậu xe, Khu vực ưa thích, Nội thất.
* Tóm tắt: Bộ dữ liệu cho phép phân tích và lập mô hình để hiểu các yếu tố ảnh hưởng đến giá nhà và đưa ra dự đoán chính xác về thị trường bất động sản.
* Chuẩn bị dữ liệu:Kiểm tra dữ liệu: Xem xét các giá trị bị thiếu, ngoại lệ hoặc không hợp lệ trong bộ dữ liệu và xử lý chúng nếu cần.

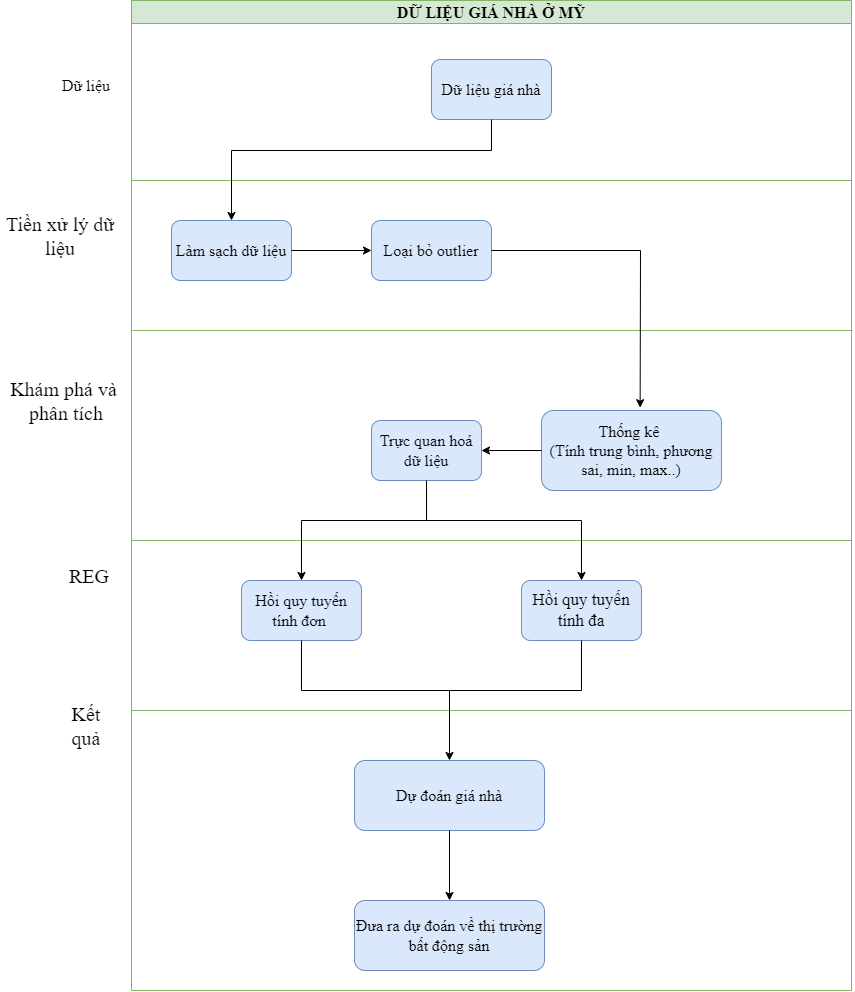
+ Tiền xử lý dữ liệu: Áp dụng các phương pháp tiền xử lý để chuẩn hóa dữ liệu.

+ Tách dữ liệu: Chia bộ dữ liệu thành hai phần, một phần để huấn luyện mô hình và một phần để kiểm tra mô hình.

+ Xây dựng mô hình: Sử dụng các phương pháp hồi quy như hồi quy tuyến tính,, hoặc các phương pháp hồi quy khác để xây dựng mô hình dự đoán giá nhà.

+ Đánh giá mô hình: Đánh giá hiệu suất của mô hình bằng các độ đo như sai số trung bình

# Framework

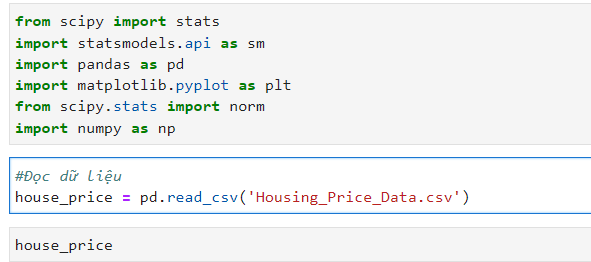


Hình 4.Framework

# Chi tiết quy trình triển khai bằng python

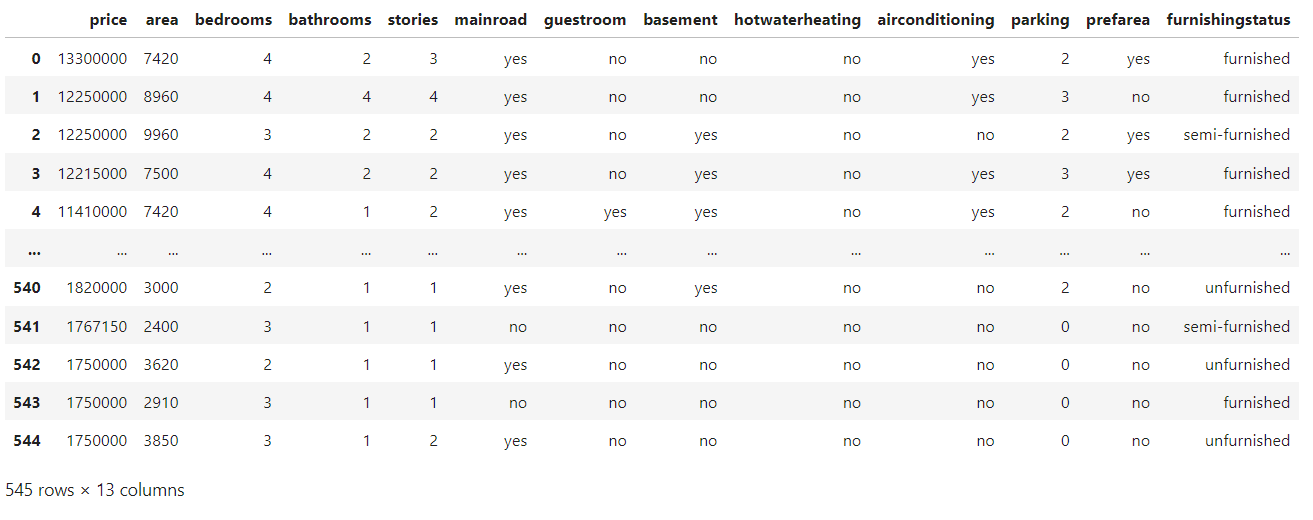
## Tiền xử lý dữ liệu

* Đọc dữ liệu



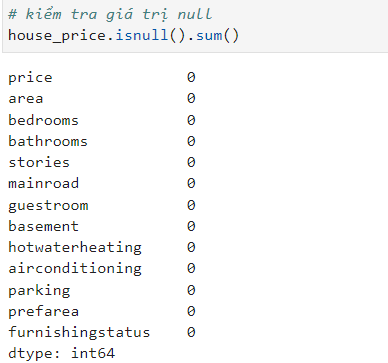
Hình 5. Code đọc dữ liệu

* Kết quả:



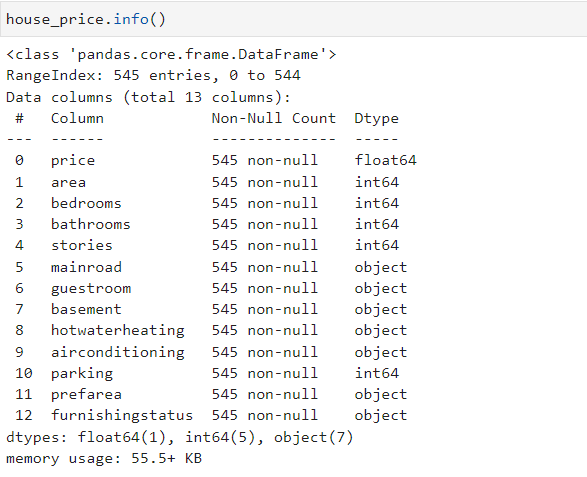
Hình 6. Kết quả đọc dữ liệu

* Kiểm tra các giá trị null và loại bỏ nếu có trong dữ liệu



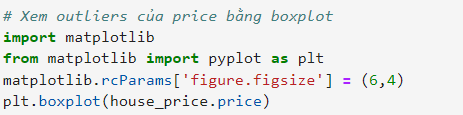
Hình 7. Kiểm tra các giá trị null

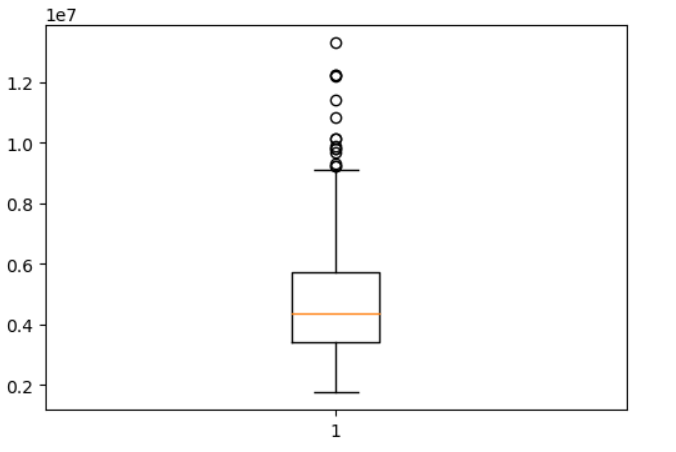
* Thông tin các cột dữ liệu



Hình 8. Thông tin các cột dữ liệu

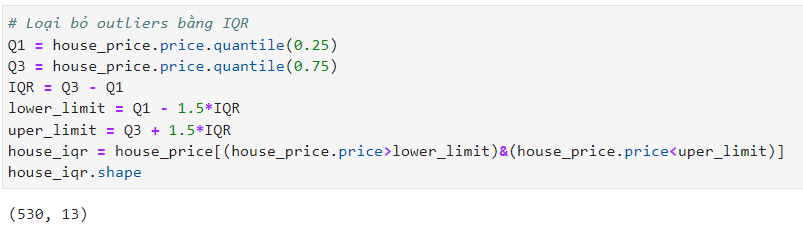
* Tiền xử lý dữ liệu là một kỹ thuật được sử dụng để chuyển đổi dữ liệu thô thành một tập dữ liệu sạch.
* Từ đó, quá trình xử liệu dữ liệu nhằm cải thiện chất lượng dữ liệu và do đó cải tiến chất lượng của kết quả phân tích và khai phá dữ liệu.
* Dữ liệu khi chưa được loại bỏ:



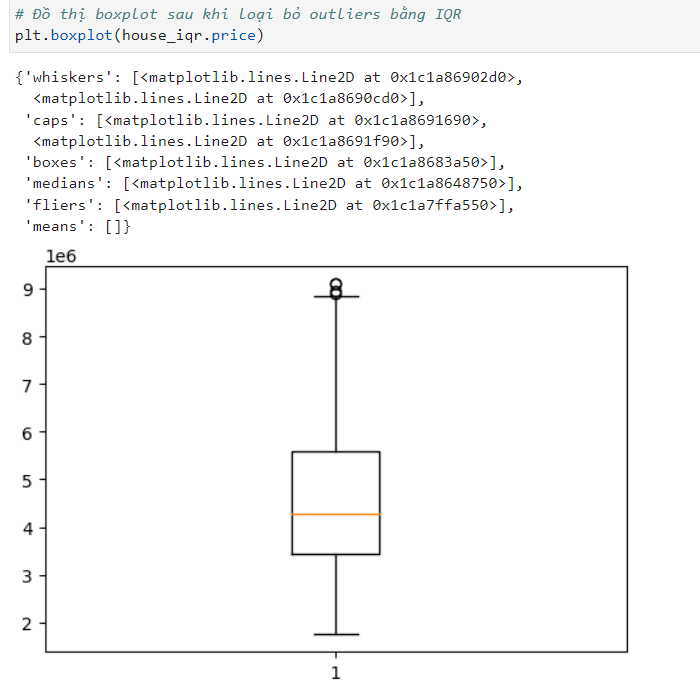


Hình 9. Xem outliers

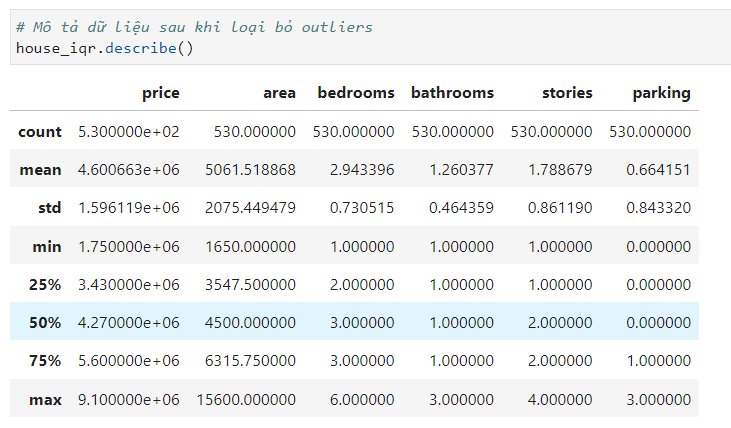
* Quá trình tiền xử lý dữ liệu nhóm sử dụng phương pháp IQR để loại bỏ các trị outlier để đem lại chất lượng cao cho quá trình phân tích và khám phá dữ liệu về giá nhà tại Mỹ.



Hình 10. Loại bỏ outlier bằng IQR



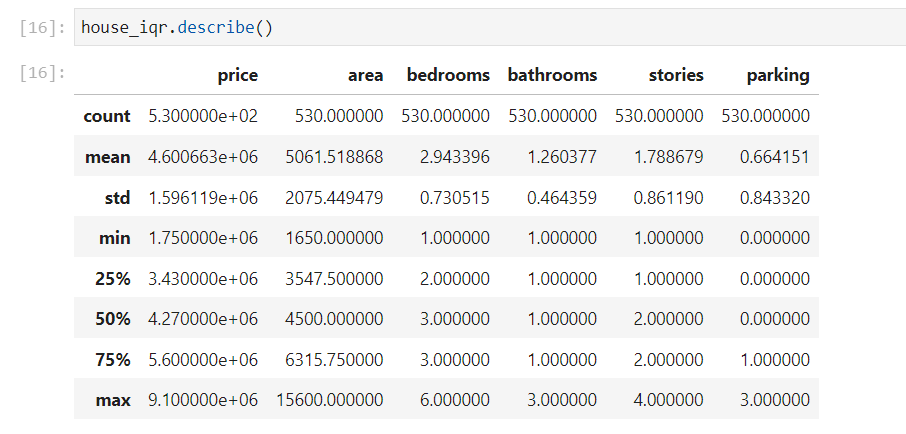
Hình 11.Đồ thị boxplot sau loại bỏ outliers



Hình 12. Mô tả dữ liệu sau khi loại bỏ outliers

## Khám phá và phân tích

### Thống kê

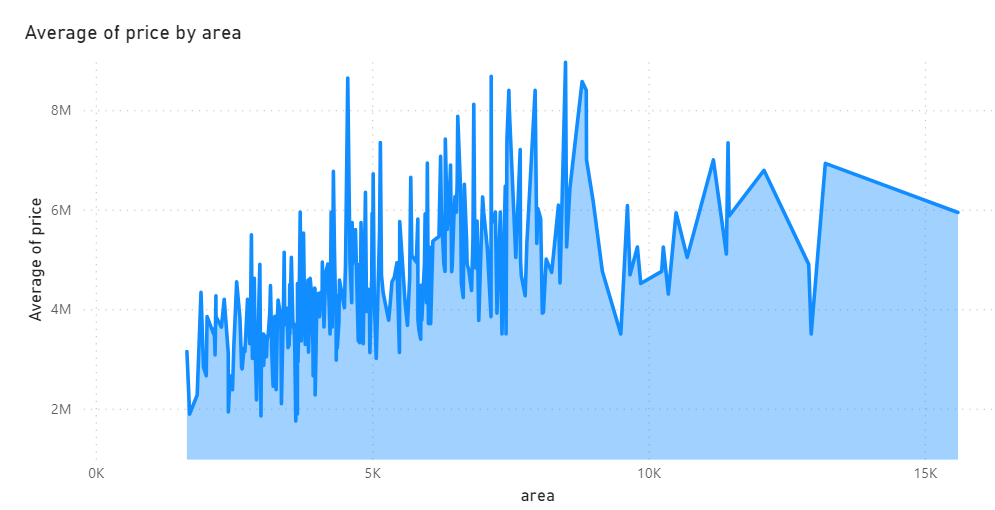


Hình 13. Thống kê dữ liệu

* Qua bảng thống kê ta thấy khám phá và phân tích được những thông tin:
* Giá: Giá nhà trung bình.
* Diện tích: Diện tích trung bình của ngôi nhà.
* Phòng ngủ: Số phòng ngủ trung bình trong nhà.
* Phòng tắm: Số phòng tắm trung bình trong nhà.
* Tầng: Số tầng trung bình của ngôi nhà.
* Bãi đậu xe: Số chỗ đậu xe trung bình trong nhà.
* Dữ liệu dựa trên 530 ngôi nhà được bán ở Hoa Kỳ vào năm 2023.
* Nhận xét:
* Giá nhà trung bình là 4,6 triệu đô la.
* Diện tích trung bình của một ngôi nhà là 5.061 feet vuông.
* Ngôi nhà trung bình có 2,94 phòng ngủ.
* Ngôi nhà trung bình có 1,26 phòng tắm.
* Ngôi nhà trung bình có 1,79 tầng.
* Ngôi nhà trung bình có 0,66 chỗ đậu xe.
* Xu hướng:
* Giá nhà đang tăng.
* Diện tích nhà đang tăng.
* Số phòng ngủ trong nhà đang tăng.
* Số phòng tắm trong nhà đang tăng.
* Số tầng của nhà đang tăng.
* Số chỗ đậu xe trong nhà đang tăng.
* **Tác động:**
* Giá nhà cao khiến việc mua nhà trở nên khó khăn hơn đối với nhiều người.
* Những ngôi nhà lớn hơn đang tiêu thụ nhiều tài nguyên hơn.
* Nhiều phòng ngủ và phòng tắm hơn đang dẫn đến việc sử dụng nước và năng lượng nhiều hơn.
* Nhiều tầng hơn đang dẫn đến việc sử dụng đất nhiều hơn.
* Nhiều chỗ đậu xe hơn đang dẫn đến việc sử dụng ô tô nhiều hơn.

### . Trực quan hóa dữ liệu

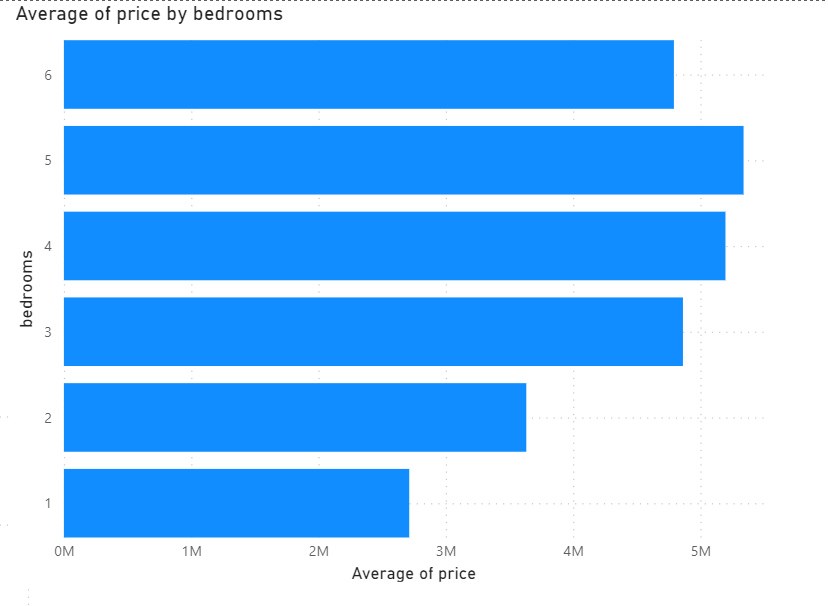
#### Giá trung bình theo diện tích



Hình 14. Biểu đồ giá nhà trung bình theo diện tích

* Giá trung bình bất động sản có xu hướng tăng khi diện tích tăng. Điều này cho thấy tồn tại một mối quan hệ tuyến tính dương giữa giá và diện tích.
* Biến động lớn ở các khu vực có diện tích nhỏ: Nhìn vào biểu đồ, chúng ta thấy rằng giá trung bình có xu hướng biến đổi mạnh mẽ ở các khu vực có diện tích nhỏ. Điều này có thể liên quan đến việc thị trường bất động sản ở những khu vực này thường không ổn định và dễ bị ảnh hưởng bởi các yếu tố như cung cầu, thay đổi kinh tế, và quy hoạch đô thị.
* Sự ổn định hơn khi diện tích tăng: Ngược lại, khi diện tích tăng, giá trung bình có xu hướng ổn định hơn. Các khu vực lớn hơn thường ít bị ảnh hưởng bởi các biến động ngắn hạn và có thể có sự ổn định hơn trong việc đầu tư và phát triển.
* Cân nhắc về quy mô và đầu tư: Dựa trên biểu đồ, việc đầu tư vào bất động sản cần xem xét cả quy mô và diện tích. Có thể có cơ hội lớn hơn ở các khu vực lớn hơn, nhưng cũng cần cân nhắc rủi ro và tiềm năng sinh lời.

#### Giá trung bình theo số lượng phòng ngủ

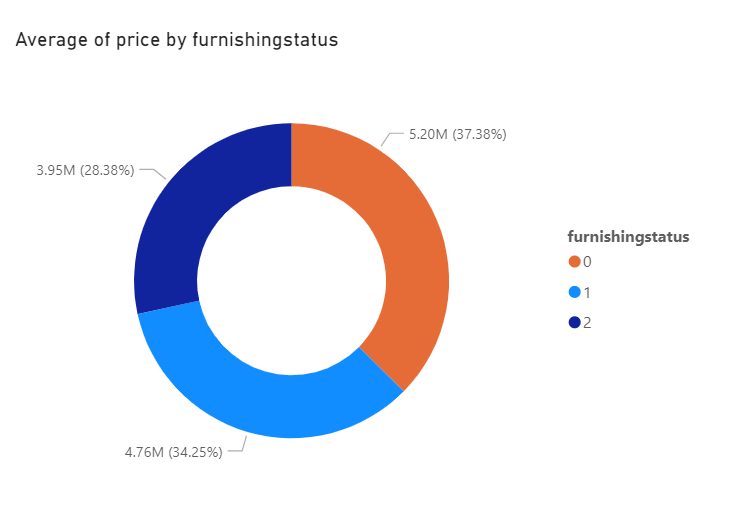


Hình 15. Biểu đồ giá nhà trung bình theo số lượng phòng ngủ

* Dựa trên sự phân tích biểu đồ, chúng ta có thể nhận thấy rằng giá trị của một căn nhà có xu hướng tăng theo số lượng phòng ngủ. Điều này cho thấy sự tồn tại một mối quan hệ tuyến tính giữa giá và diện tích của căn nhà. Biểu đồ này cung cấp một cái nhìn tổng quan về sự tương quan giữa các yếu tố quan trọng trong việc xác định giá trị căn nhà. Khi số lượng phòng ngủ tăng lên, giá trị của căn nhà cũng tăng theo. Điều này có thể được giải thích bằng việc rằng các căn nhà có số phòng ngủ lớn hơn thường có diện tích lớn hơn và do đó có giá trị cao hơn.

#### Giá trung bình theo nội thất

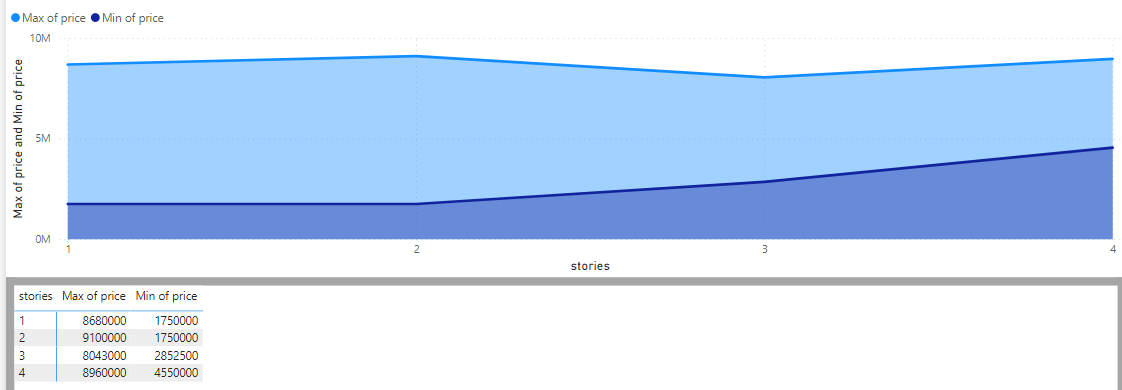
* Nội thất đầy đủ (0);
* Nội thất 1 nửa chưa đầy đủ (1);
* Không có nội thất (2)



Hình 16. Biểu đồ giá nhà trung bình theo nội thất

* Dựa vào phân tích biểu đồ, chúng ta có thể nhận thấy rằng giá cả của căn phòng thay đổi theo mức độ đầy đủ nội thất của nó. Điều này cho thấy sự tồn tại một mối quan hệ tương quan tuyến tính giữa giá cả và sự đầy đủ nội thất của căn phòng. Mối quan hệ này có ý nghĩa quan trọng trong việc xác định giá trị của một căn phòng. Khi căn phòng được trang bị đầy đủ nội thất, giá cả thường có xu hướng tăng lên. Ngược lại, nếu căn phòng thiếu nội thất hoặc không đầy đủ, giá cả có thể giảm đi.
* Tuy nhiên, cần lưu ý rằng mối quan hệ giữa giá cả và sự đầy đủ nội thất không phải lúc nào cũng là một mối quan hệ tuyệt đối. Ngoài sự đầy đủ nội thất, giá trị căn phòng cũng phụ thuộc vào nhiều yếu tố khác như vị trí, kích thước căn phòng, chất lượng xây dựng, tiện ích xung quanh và nhu cầu thị trường.

#### Giá lớn nhất và giá nhỏ nhất theo số tầng



Hình 17. Biểu đồ giá nhà lớn nhất và giá nhà nhỏ nhất theo số tầng

* Dựa vào biểu đồ ta thấy:
* Giá phòng thấp nhất của phòng tầng 1 là 1750000$ và giá trị lớn nhất 8680000$.
* Giá phòng thấp nhất của phòng tầng 2 là 1750000$ và giá trị lớn nhất 9100000$.
* Giá phòng thấp nhất của phòng tầng 3 là 2852500$ và giá trị lớn nhất 8043000$.
* Giá phòng thấp nhất của phòng tầng 4 là 4550000$ và giá trị lớn nhất 8960000$.

→ Giá phòng tầng 4 có giá phòng cao nhất và tầng 2 có giá từ rẻ đến đắt.

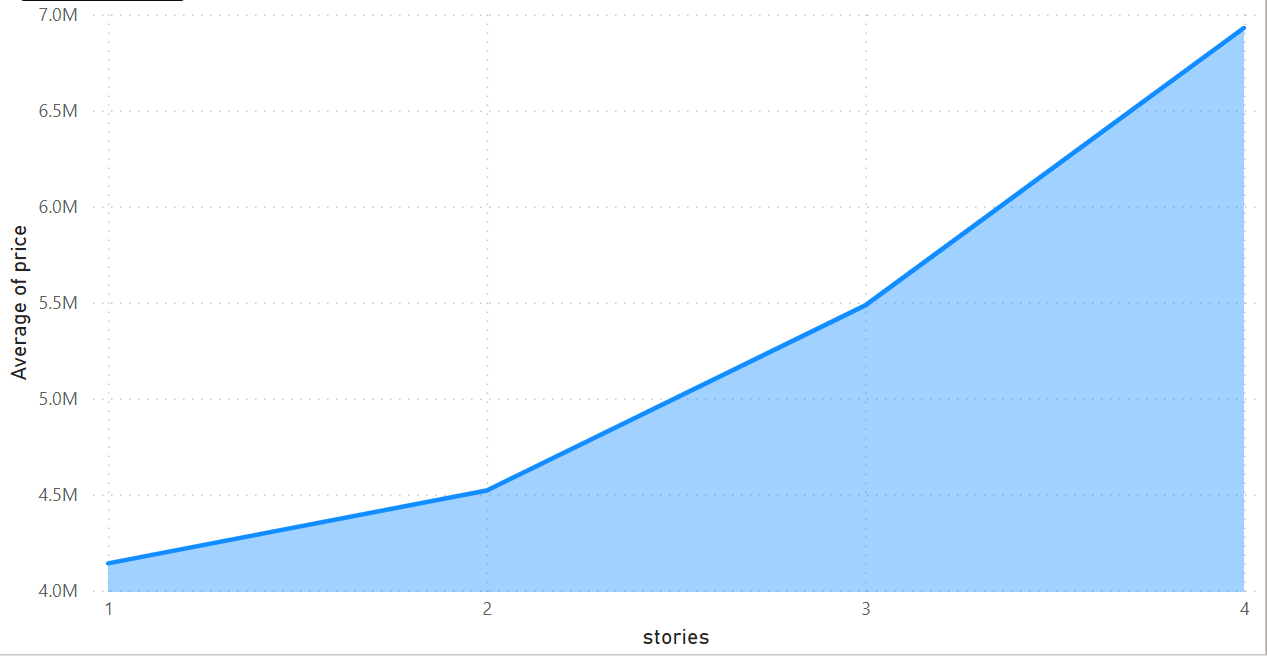
#### Diện tích trung bình của các tầng

#### 

Hình 18. Diện tích trung bình của các tầng

* Dựa vào biểu đồ ta thấy:
* Diện tích căn hộ trong tầng 2 có giá trị nhỏ nhất, chỉ đạt 4587,38m2. Điều này cho thấy rằng các căn hộ ở tầng 2 thường có diện tích nhỏ hơn so với các tầng khác. Có thể do diện tích hạn chế trong tầng này hoặc có các yếu tố khác như thiết kế, vị trí của căn hộ. Tầng 4 lại có diện tích căn hộ lớn nhất, với giá trị là 6406,92m2. Điều này cho thấy rằng các căn hộ ở tầng 4 thường có diện tích lớn hơn so với các tầng khác.

#### Giá cả trung bình theo tầng



Hình 19. Biểu đồ giá cả trung bình theo tầng

* Dựa vào biểu đồ ta thấy:
* Giá trung bình của tầng 1 thấp nhất với giá 4141103.98$. Có thể có nhiều yếu tố đã dẫn đến sự khác biệt này, bao gồm diện tích, thiết kế, vị trí, tiện ích và các yếu tố khác. Tầng 1 có thể là tầng mặt đất, nơi có thể có sự tiếp xúc trực tiếp với đường phố hoặc khu vực công cộng, điều này có thể ảnh hưởng đến giá trị bất động sản.
* Giá trung bình tầng 4 cao nhất với giá 6930169.47$. Điều này cho thấy tầng 4 có mức giá trung bình cao hơn so với các tầng khác trong tòa nhà hoặc khu vực. Một số yếu tố có thể giải thích sự khác biệt này bao gồm tầm nhìn, cách biển cận, tiện ích hoặc thiết kế độc đáo. Tầng 4 có thể có lợi thế về tầm nhìn hoặc tiếp xúc với ánh sáng tự nhiên tốt hơn so với các tầng thấp hơn. Ngoài ra, trong một số trường hợp, các tầng cao hơn có thể được trang bị các tiện ích cao cấp hoặc các không gian sống rộng rãi hơn, điều này có thể tạo ra sự khác biệt về giá trị.

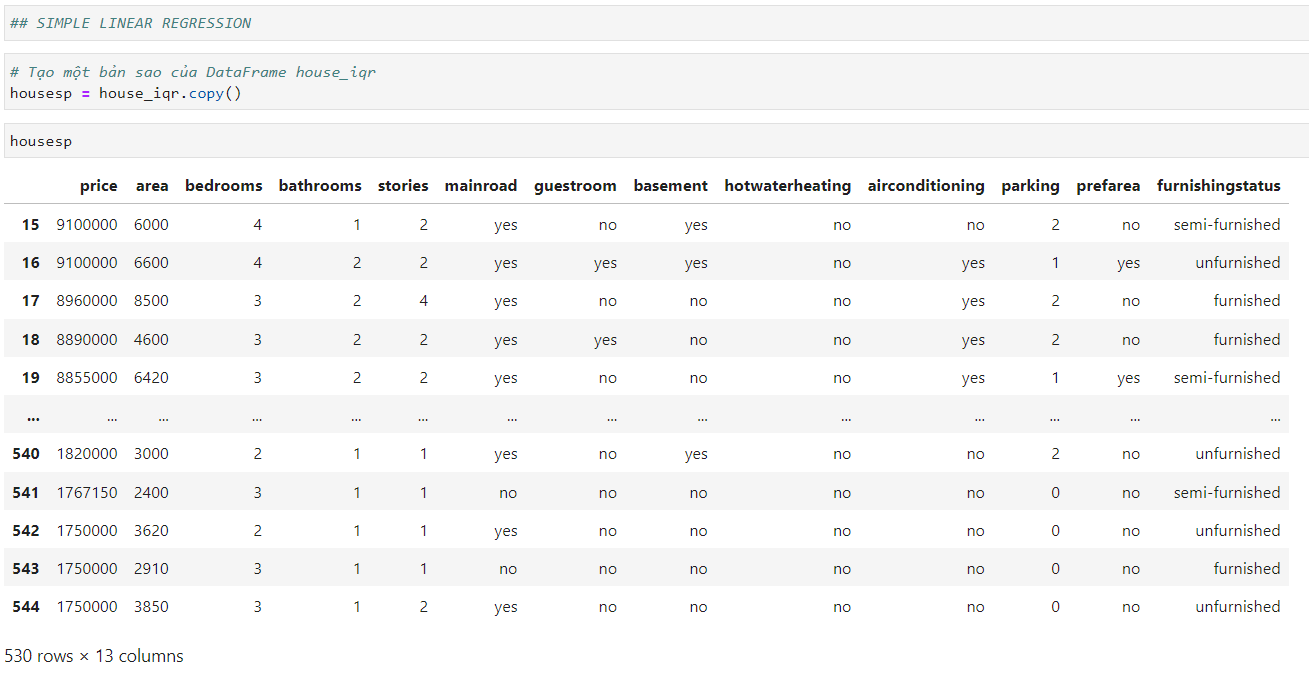
## REG

### Hồi quy tuyến tính đơn bội

* Giả thuyết: H0: β1 = 0 X(area) và Y(price) không có liên hệ tuyến tính

H1: β1 ≠ 0 X(area) và Y(price) có liên hệ tuyến tính

* Mức ý nghĩa: *a*  = 0.05
* Quy trình kiểm định giả thuyết bằng python:

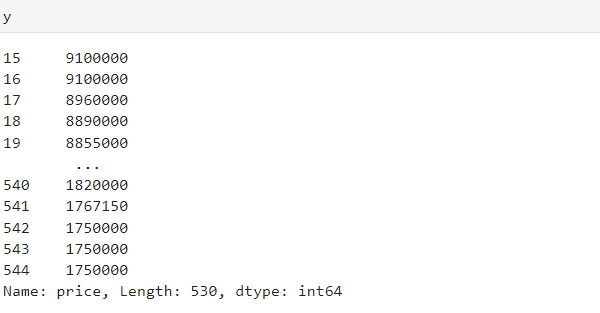


Hình 20. Tạo bản sao của dữ liệu



Hình 21. Lấy 2 cột dữ liệu area và price

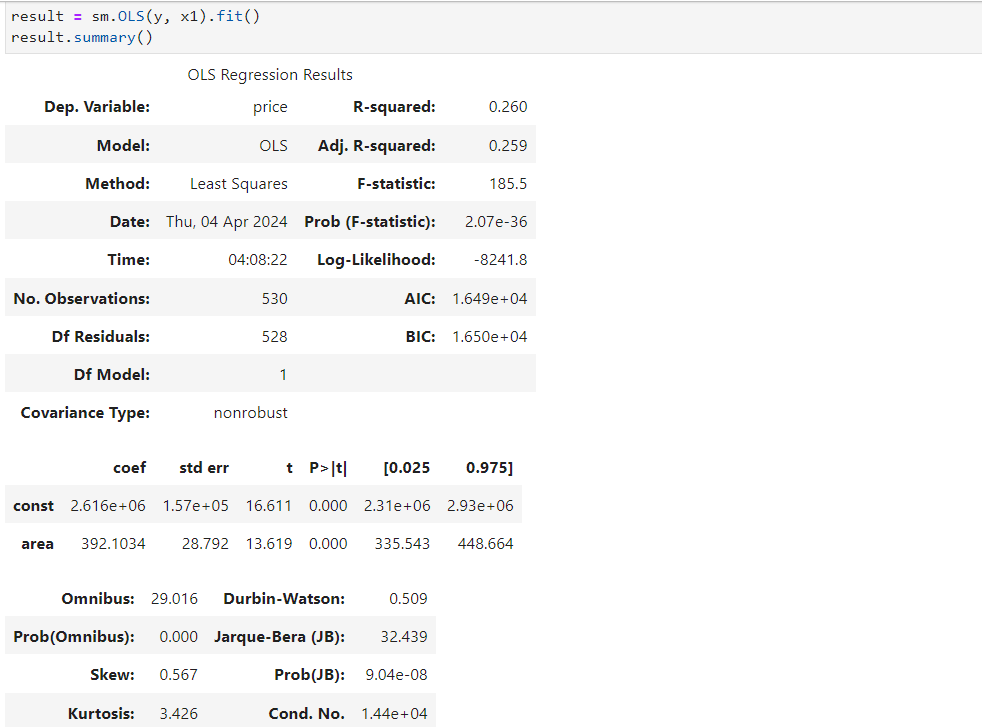
* Kết quả



Hình 22. In ra cột price



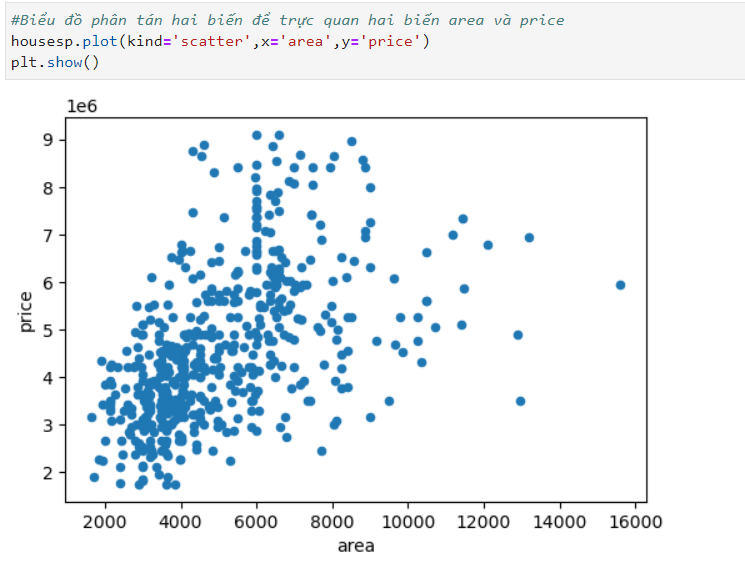
Hình 23. Thêm hệ số chặn



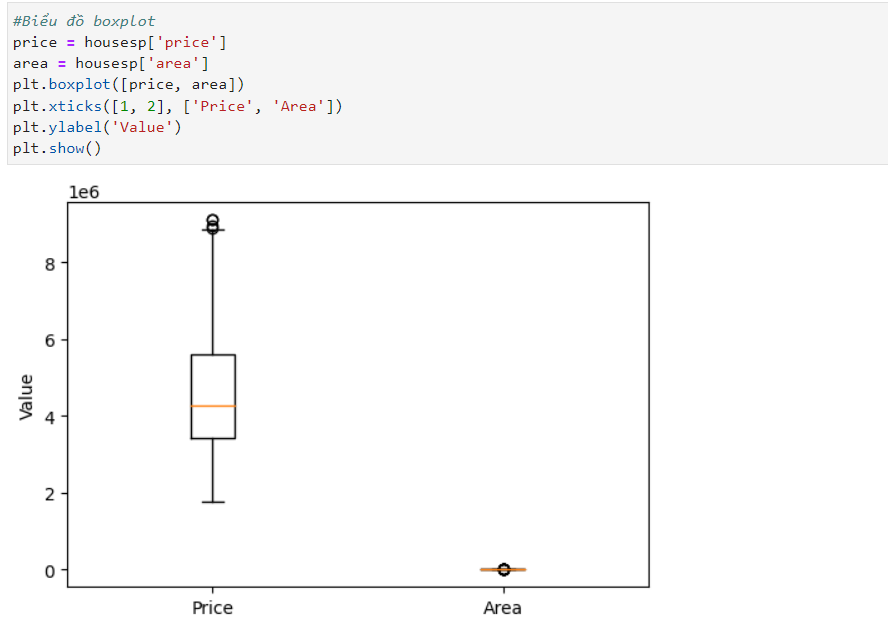
Hình 24. Kiểm định giả thuyết

=> Chúng ta sử dụng hai mảng price và area để tạo dữ liệu mẫu. Trước khi thực hiện kiểm định t-test, chúng ta thêm intercept vào dữ liệu bằng cách sử dụng sm.add\_constant(area). Sau đó, chúng ta khởi tạo mô hình hồi quy tuyến tính bằng sm.OLS(price, X) và fit mô hình vào dữ liệu. Cuối cùng, chúng ta lấy giá trị t-test và p-value cho hệ số hồi quy thông qua results.tvalues[1] và results.pvalues[1].

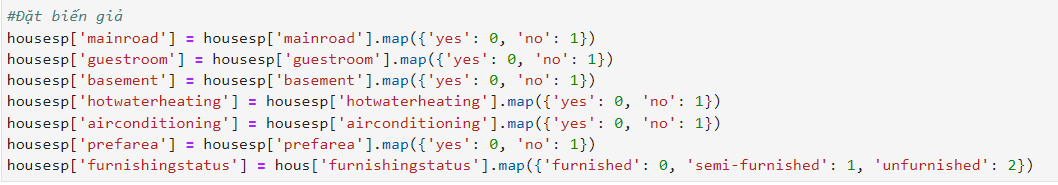
* Kết quả sẽ là giá trị t-test và p-value cho hệ số hồi quy của biến area. p-value nhỏ hơn ngưỡng xác định trước (thường là 0.05), ta bác bỏ giả thuyết rằng hệ số hồi quy bằng 0 (bác bỏ H0) và kết luận rằng biến area có tác động đáng kể đến biến phụ thuộc price.
* Khi kiểm định giả thiết area và price có mối quan hệ tuyến tính. Ta thực hiện dự báo price dựa trên area

****

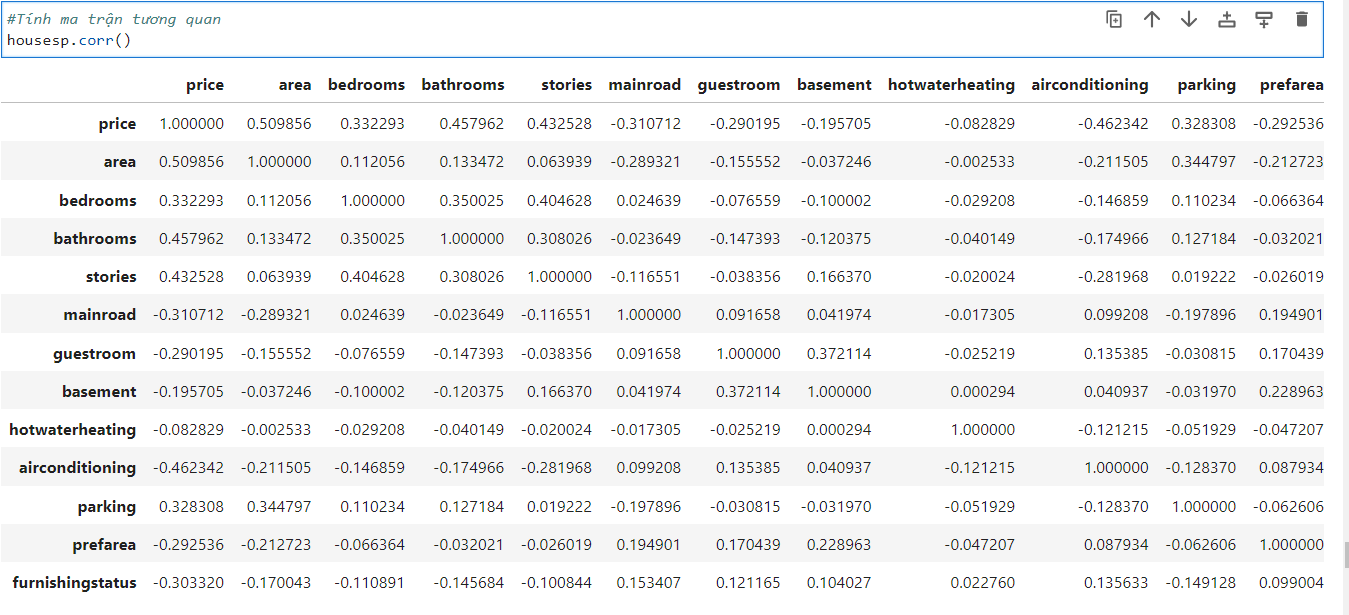
Hình 25. Biểu đồ phân tán trực quan 2 biến area và price

****

Hình 26. Biểu đồ boxplot

****

Hình 27. Code đặt biến giả đổi kiểu dữ liệu

****

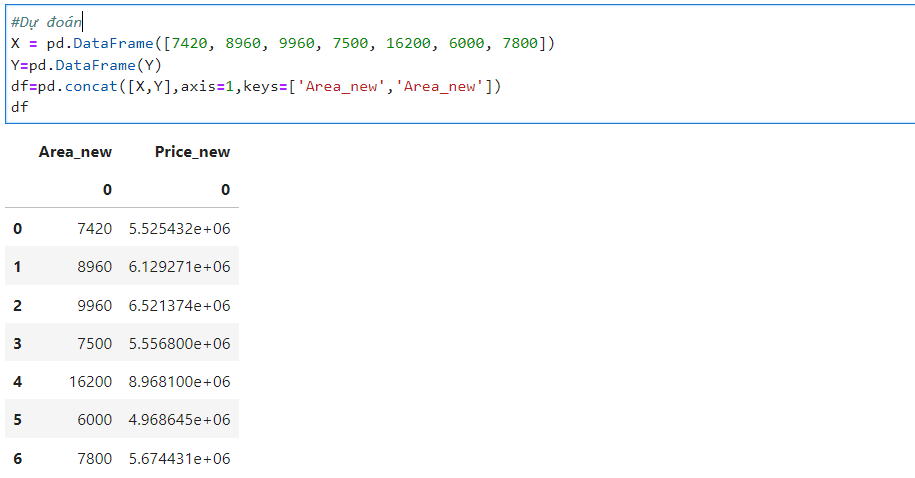
Hình 28. Tính ma trận tương quan

****

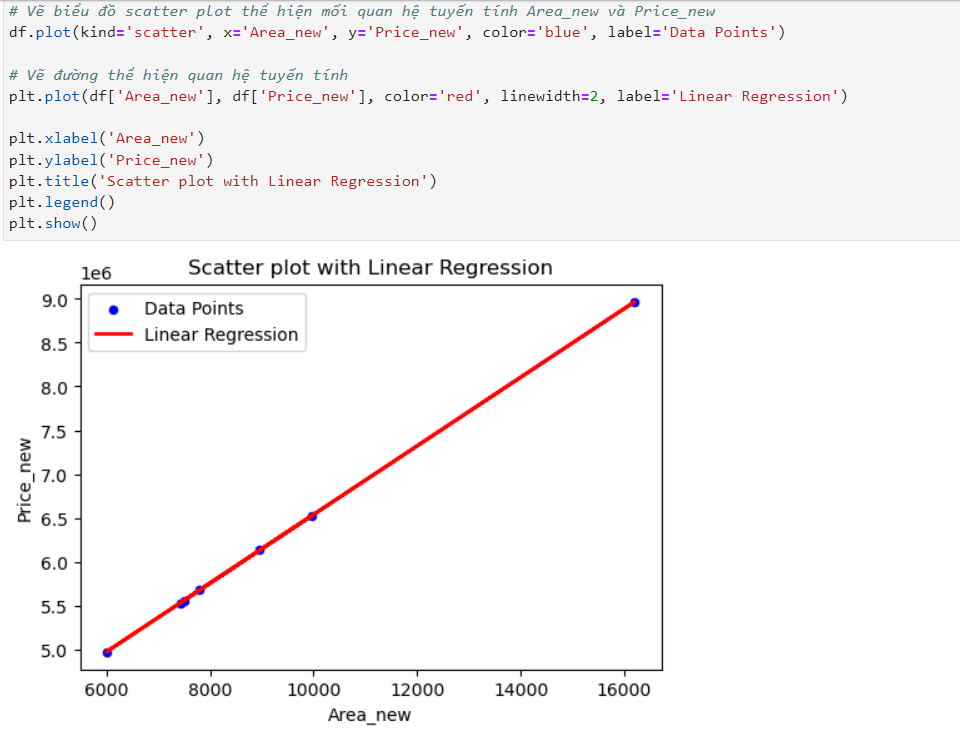
Hình 29. Đưa 2 biến vào data Frame

****

Hình 30. Xây dựng dự đoán

****

Hình 31. Dự đoán giá nhà theo diện tích

****

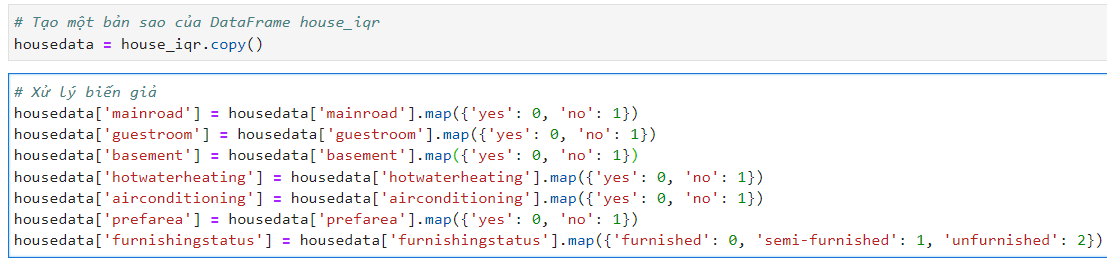
Hình 32. Vẽ biểu đồ scatter plot thể hiện mối quan hệ tuyến tính giữa 2 biến

### Hồi quy tuyến tính đa bội

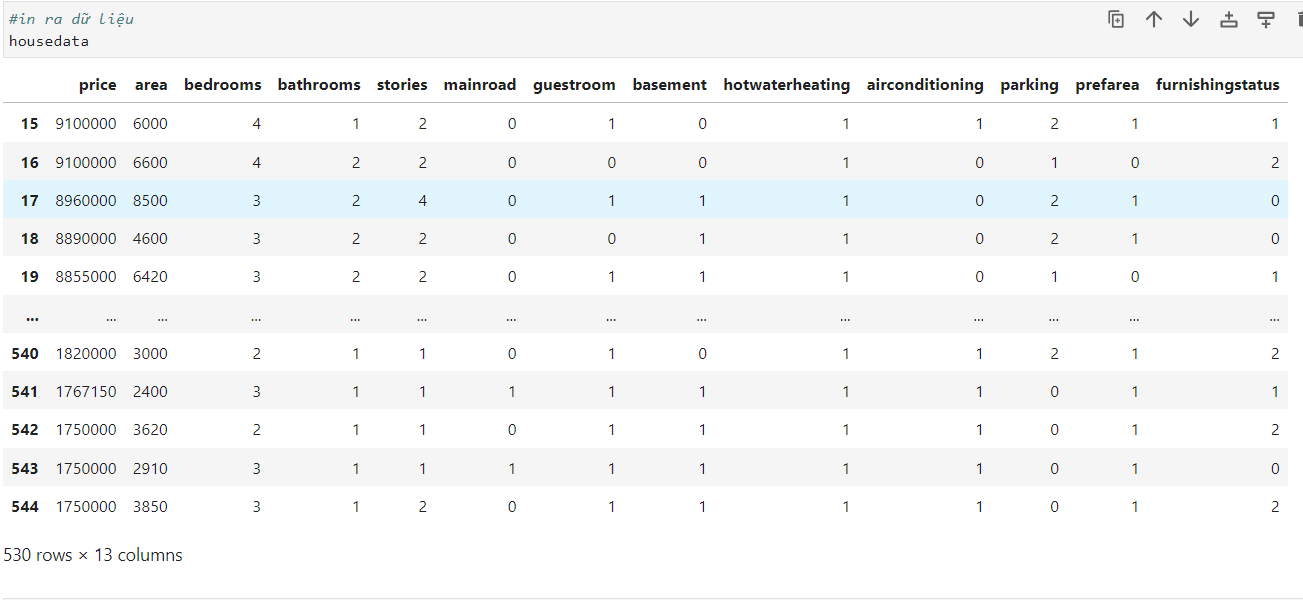
* Giả thuyết: H0: β1 =β2 =β3 =β4 =β5 =β6 = 0 (area, bedrooms, bathrooms, stories, mainroads, furnishingstatus và price không có mối quan hệ tuyến tính)

H1: β1 =β2 =β3 =β4 =β5 =β6 ≠ 0 (area, bedrooms, bathrooms, stories, mainroads, furnishingstatus và price có mối quan hệ tuyến tính)

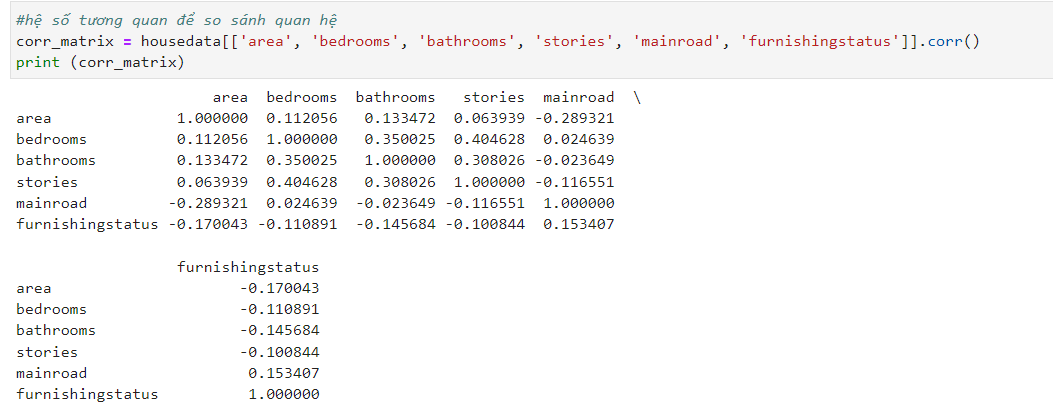
* Mức ý nghĩa: *a*  = 0.05
* Quy trình kiểm định giả thuyết bằng python:



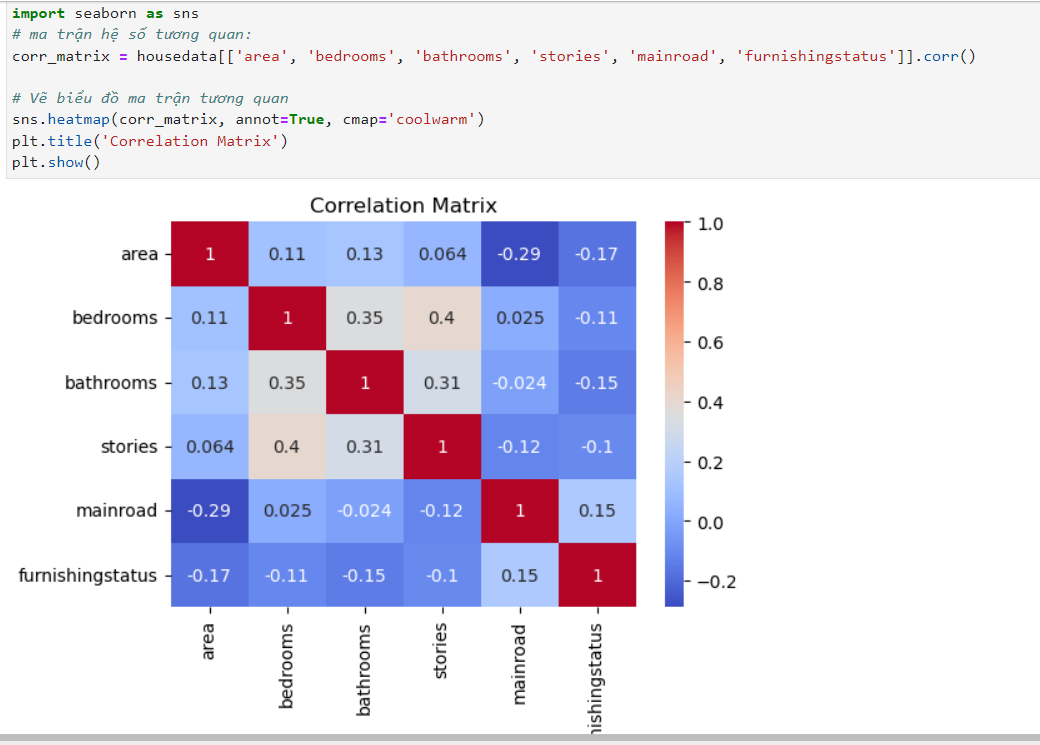
Hình 33. Tạo bản sao và xử lý biến giả



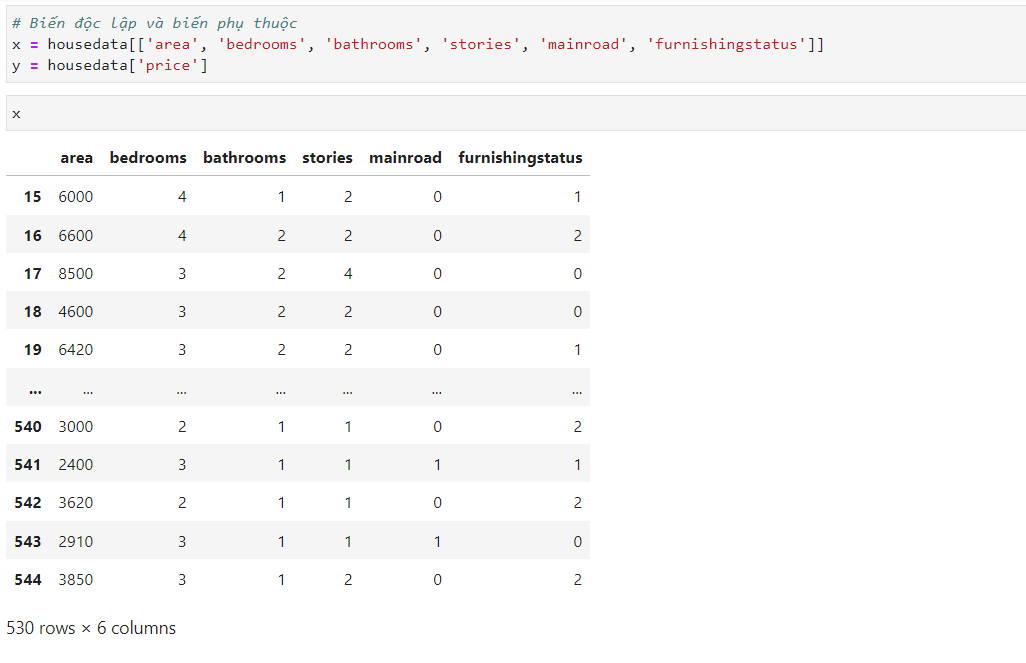
Hình 34. In ra dữ liệu sau khi xử lý biến giả



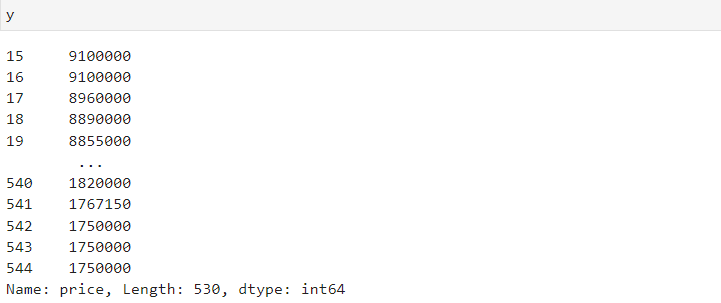
Hình 35. In ra hệ số tương quan



Hình 36. Ma trận hệ số tương quan

****

Hình 37. In ra biến độc lập



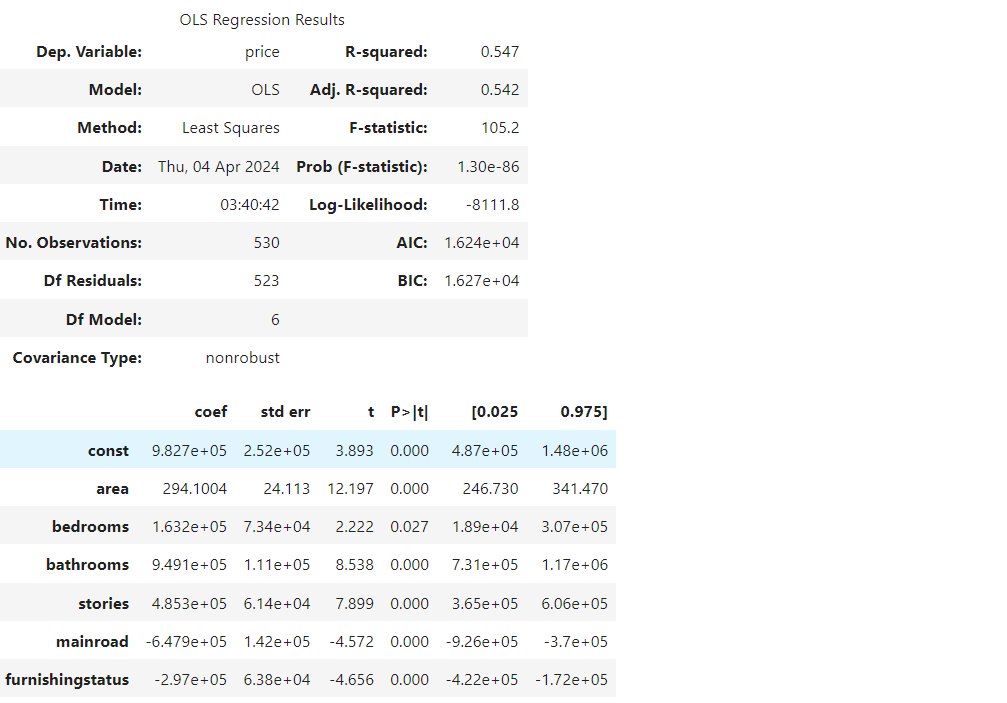
Hình 38. In ra biến phụ thuộc



Hình 39. Tạo intercept B0



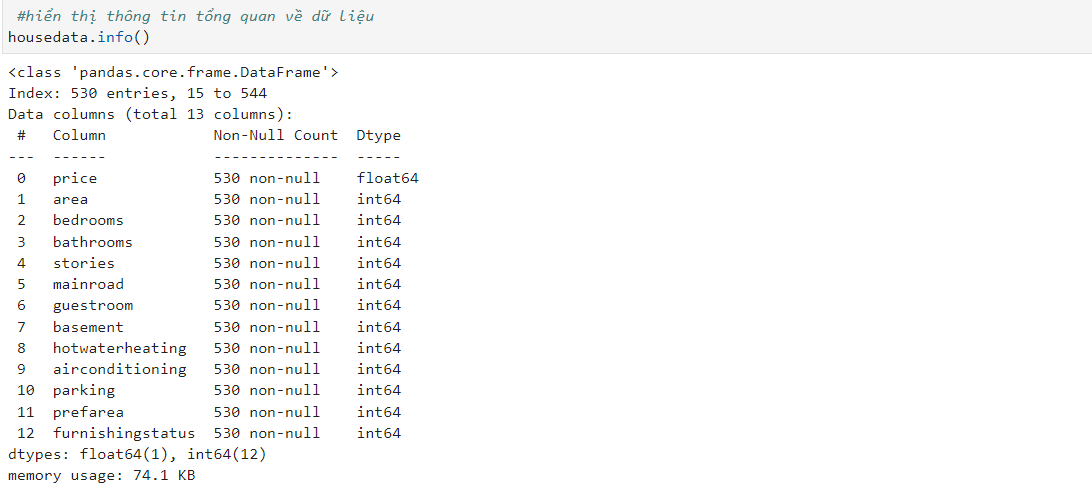
Hình 40. Code kiểm định giả thuyết



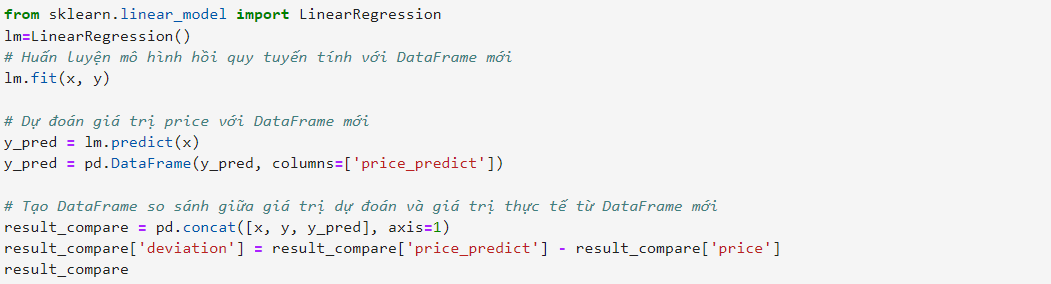
Hình 41. In ra kết quả kiểm định giả thuyết

=> Chúng ta sử dụng hai mảng price và (area, bedrooms, bathrooms, stories, mainroasds, furnishingstatus) để tạo dữ liệu mẫu. Trước khi thực hiện kiểm định t-test, chúng ta thêm intercept vào dữ liệu bằng cách sử dụng sm.add\_constant(area, bedrooms, bathrooms, stories, mainroasds, furnishingstatus). Sau đó, chúng ta khởi tạo mô hình hồi quy tuyến tính bằng sm.OLS(price, X) và fit mô hình vào dữ liệu. Cuối cùng, chúng ta lấy giá trị t-test và p-value cho hệ số hồi quy thông qua results.tvalues[1] và results.pvalues[1].

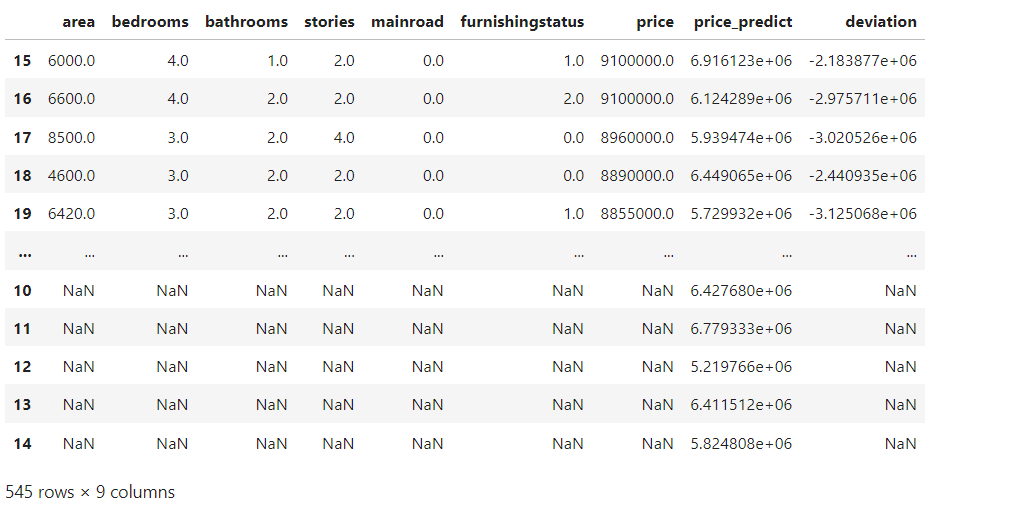
* Kết quả sẽ là giá trị t-test và p-value cho hệ số hồi quy của biến area. p-value nhỏ hơn ngưỡng xác định trước (thường là 0.05), ta bác bỏ giả thuyết rằng hệ số hồi quy bằng 0 (bác bỏ H0) và kết luận rằng biến area,bedrooms, bathrooms, stories, mainroasds, furnishingstatus có tác động đáng kể đến biến phụ thuộc price.
* Khi kiểm định giả thiết area, bedrooms, bathrooms, stories, mainroasds, furnishingstatus và price có mối quan hệ tuyến tính. Ta thực hiện dự báo price dựa trên area, bedrooms, bathrooms, stories, mainroasds, furnishingstatus.



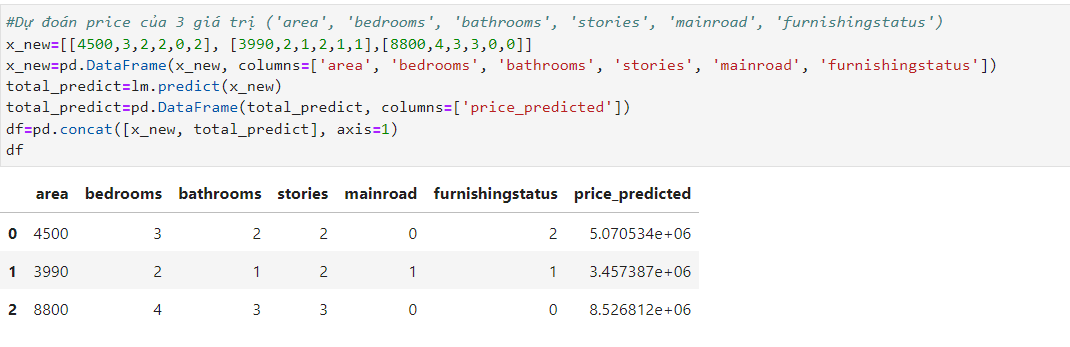
Hình 42. Hiển thị thông tin tổng quan về dữ liệu



Hình 43. Code so sánh giá trị dự đoán và giá trị thực tế



Hình 44. In ra kết quả so sánh giá trị dự đoán và giá trị thực tế



Hình 45. Dự đoán giá trị mới

## Kết quả nghiên cứu.

* Sau khi sử dụng hồi quy tuyến tính, nhóm đã có cái nhìn chi tiết hơn về dữ liệu cũng như có thể đưa ra các dự đoán về giá nhà khi có những yếu tố ảnh hưởng đến giá. Dựa vào mô hình hồi quy tuyến tính, nhóm đã xác định các yếu tố ảnh hưởng đến giá nhà. Các yếu tố ảnh hưởng đến giá nhà bao gồm diện tích, số phòng ngủ, phòng khách, phòng tắm, hướng ra mặt chính, tình trạng nội thất.Tuy nhiên diện tích nhà là yếu tố quan trọng ảnh hưởng đến giá nhà. Các yếu tố khác như số phòng ngủ, số phòng tắm, số tầng, vị trí nhà trên đường chính và tình trạng nội thất cũng có ảnh hưởng đến giá, nhưng mức ảnh hưởng ít hơn.
* Dự đoán:
* Area (diện tích) có thể có tác động lớn đến giá nhà. Thường thì diện tích càng lớn, giá nhà càng cao.
* Số phòng ngủ (bedrooms) và số phòng tắm (bathrooms) cũng có thể ảnh hưởng đáng kể đến giá nhà. Thông thường, những căn nhà có số phòng ngủ và phòng tắm nhiều hơn sẽ có giá cao hơn.
* Số phòng khách (guestrooms) có thể cũng có tác động nhưng có thể không mạnh như diện tích, số phòng ngủ và phòng tắm.
* Có hướng mặt tiền đường chính (mainroads) có thể là một yếu tố quan trọng. Những căn nhà có mặt tiền đường chính thường có giá trị cao hơn so với những căn nhà không có mặt tiền đường chính.
* Tình trạng nội thất (furnishingstatus) cũng có thể ảnh hưởng đến giá nhà. Những căn nhà được trang bị nội thất cao cấp có thể có giá trị cao hơn so với những căn nhà không được trang bị nội thất hoặc có nội thất đơn giản.

=> Dự đoán xu hướng bất động sản của một khu vực.

# Video

Link video báo cáo:

# TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | CafeDev, "Triển khai hồi quy tuyến tính bằng python," *Triển khai hồi quy tuyến tính bằng python,* vol. 50, no. 2, p. 10, 2024. |
| [2] | Amazon, "Tại sao hồi quy logistic lại quan trọng?," *Tại sao hồi quy logistic lại quan trọng?,* vol. 50, no. 2, p. 8, 2024. |
|  |
| [3] | V1Study, "Hồi quy tuyến tính trong python," *Python: Hồi quy tuyến tính trong python,* vol. 50, no. 2, p. 15, 2024. |