ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

HUỲNH KHẢI SIÊU – 18520348

TRẦN THỊ MỸ LINH – 18520999

DƯƠNG THỊ HỒNG HẠNH – 18520711

LÊ PHAN THÀNH ĐẠT – 18520570

MÔN: THIẾT KẾ VÀ PHÂN TÍCH THỰC NGHIỆM

LỚP: DS304.K21

PHÂN TÍCH ẢNH HƯỞNG CỦA MỘT SỐ YẾU TỐ ĐẾN GIÁ NHÀ

BỘ DỮ LIỆU: REAL ESTATE PRICE PREDICTION

KHDL2018

GIẢNG VIÊN HƯỚNG DẪN: TS. ĐỖ TRỌNG HỢP

TP. HỒ CHÍ MINH, 2020

1. Mở đầu:

Giá nhà đất là một chủ đề luôn nhận được sự quan tâm rất lớn của cộng đồng hiện nay, với quy mô và sự gia tăng dân số mỗi năm dẫn đến tình trạng "đất chật người đông", nhu cầu thu mua nhà vì thế cũng không ngừng tăng lên. Không những thế, nhà đất còn là lĩnh vực kinh doanh, là cơ hội đầu tư mang về nguồn lợi nhuận lớn cho nhiều người.
Vì vậy mà sự biến động của giá nhà đất trở thành vấn đề quan trọng đối với cả người mua lẫn người bán. Đối với người mua thì nên mua ở đâu, mua khi nào thì mới có được mức giá hợp lý hay đối với người bán thì nên bán ra thời điểm nào để có lời cao ? Đây luôn là câu hỏi khó, bởi diễn biến giá nhà đất là không hề đơn giản và phụ thuộc vào rất nhiều yếu tố tác động lên nó.

Từ đó cho thấy việc phân tích sự ảnh hưởng của các yếu tố đến giá nhà đất là rất cần thiết. Dự đoán giá nhà đất là một giải pháp để giải quyết bài toán cho các doanh nghiệp kinh doanh nhà đất và nhiều ngành nghề liên quan. Từ kết quả dự đoán, doanh nghiệp sẽ đưa ra những đề xuất phù hợp cho khách hàng, kịp thời đưa ra các giải pháp, đối phó với những nguy cơ trong tương lai bằng cách cắt giảm chi phí hoặc thay đổi chiến lược phù hợp. Đối với người có nhu cầu mua nhà, họ sẽ dựa vào kết quả dự đoán để xem xét chọn 1 ngôi nhà có giá cả phù hợp thu nhập tài chính cá nhân, đúng với địa điểm mong muốn. Ví dụ: Gần các cửa hàng tiện lợi, gần ga tàu điện, trung tâm thành phố,...

Để phục vụ những nhu cầu và yêu cầu trên, quá trình phân tích đánh giá và đưa ra kết quả dưới đây sẽ phần nào nói lên kết cấu của sự biến động giá nhà đất thông qua bộ dữ liệu sẵn có.

2. Giới thiệu bộ dữ liệu:

Bộ dữ liệu gồm thông tin các thuộc tính và giá cả của 414 ngôi nhà được bán trong ngày 01/01/1970 ở thành phố Tân Đài Bắc, Đài Loan. Cụ thể như sau:

- Tên bộ dữ liệu: **Real estate price prediction.**
- Nguồn: https://www.kaggle.com/quantbruce/real-estate-price-prediction
- Thông tin chi tiết:

Bảng 2.1: Codebook của bộ dữ liệu.

Tên bộ dữ liệu	Real estate price prediction

Chức năng	Bộ dữ liệu được tạo ra nhằm mục đích sử dụng cho phân tích hồi quy, nghiên cứu mô hình hồi quy tuyến tính đơn/ đa biến và xây dựng mô hình dự đoán kết quả giá nhà.
Số dòng	Gồm 414 điểm dữ liệu và 8 thuộc tính.
Số thuộc tính	8 thuộc tính.
Các thuộc tính	No: (int64) Số thứ tự X1 transaction date: (float64) Ngày giao dịch. X2 house age: (float64) Tuổi ngôi nhà. X3 distance to the nearest MRT station: (float64) Khoảng cách đến trạm ga tàu gần nhất. X4 number of convenience stores: (float64) Số cửa hàng tiện lợi ở gần đó. X5 latitude: (float64) Vĩ độ của ngôi nhà. X6 longitude: (float64) Kinh độ của ngôi nhà.
Tác giả	Bruce Thông tin chi tiết: https://www.kaggle.com/quantbruce

3. Triển khai thực hiện

3.1. Quan sát, thăm dò, tiền xử lý dữ liệu

3.1.1. Quan sát dữ liệu

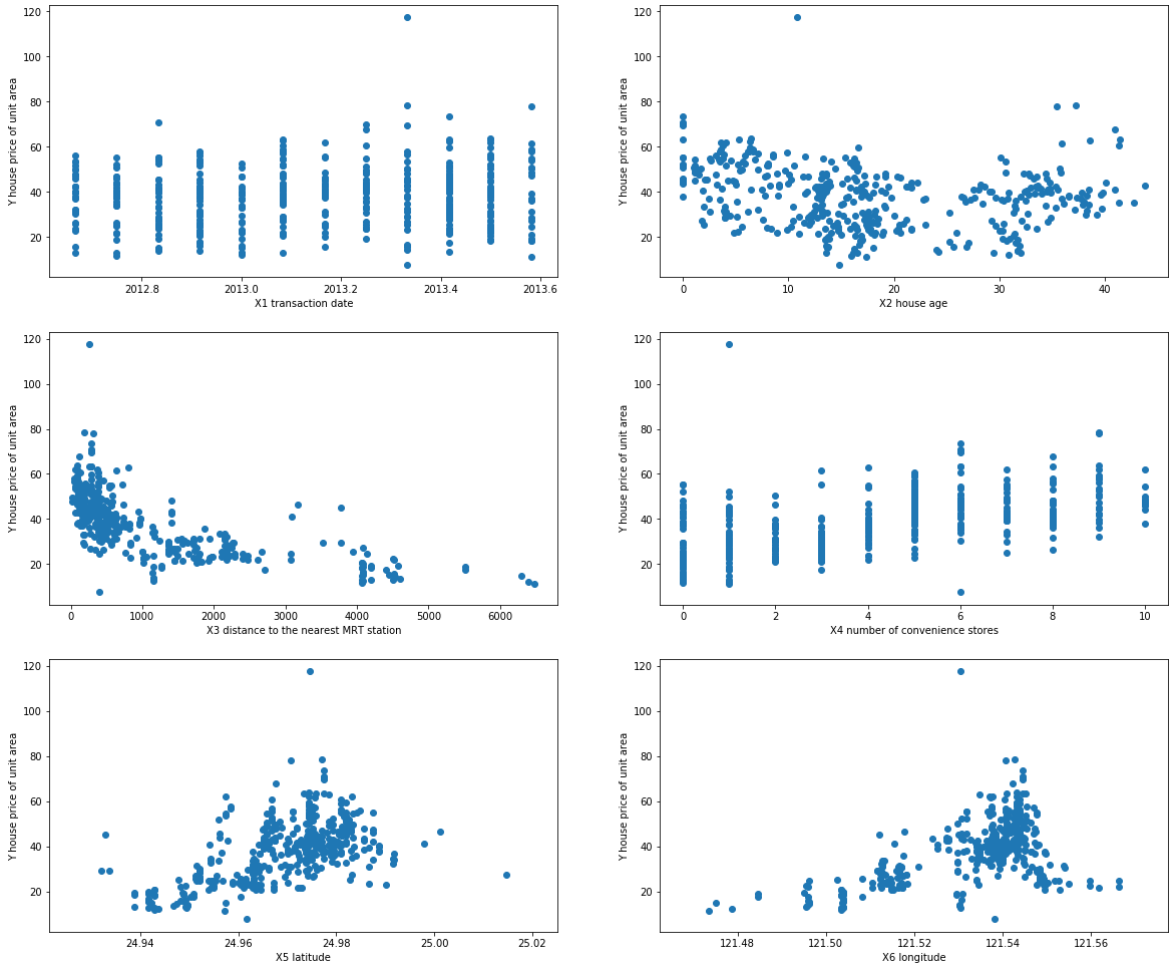
Bảng 3.1.1: Thông tin ban đầu của các thuộc tính

Tên thuộc tính	**Giá trị trung bình**	**Miền giá trị**
No	207.5	[1, 414]
X1 transaction date	2013.149	[2012.667, 2013.583]
X2 house age	17.7126	[0, 43.8]
X3 distance to the nearest MRT station	**1083.8857**	[23.38284, **6488.021**]
X4 number of convenience stores	4.0942	[0, 10]

X5 latitude	24.96903	[24.93207, 25.01459]
X6 longitude	121.53361	[121.4735, 121.5663]
Y house price ofunit area	**37.98**	[7.6, **117.5**]

3.1.2. Thăm dò dữ liệu

- **Trực quan dữ liệu của từng thuộc tính so với biến mục tiêu**

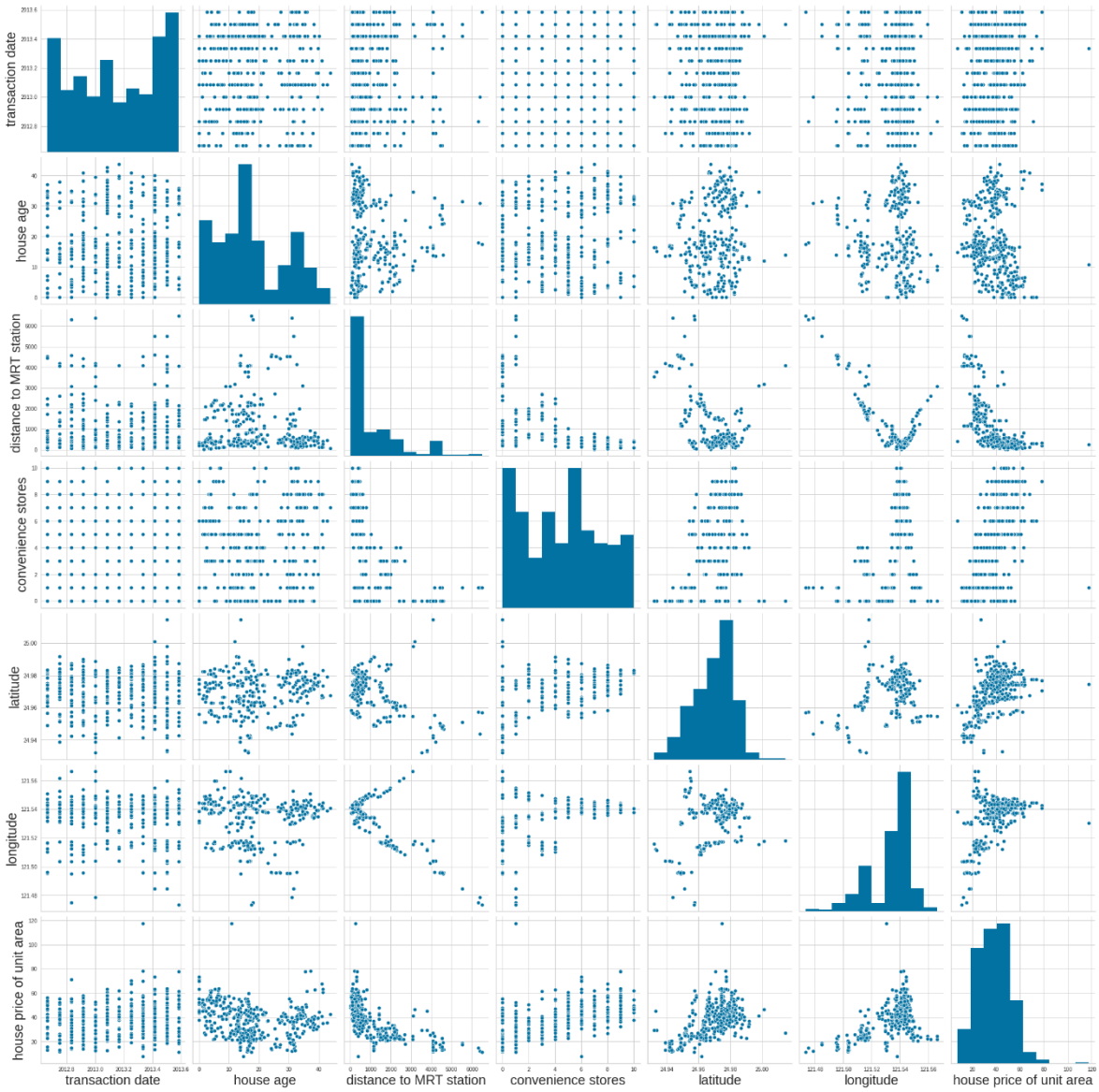


Hình 3.1.2a: Trực quan dữ liệu của từng thuộc tính X theo Y

Hầu như tất cả các thuộc tính từ X1(X1 transaction date) đến X6 (X6 longitude) đều có sự biến động rõ rệt tại mỗi giá trị được nhận so với Y (Y house price of unit area).

Ngoài ra tại một số vị trí còn có các điểm dữ liệu nằm tách biệt, khá xa so với những điểm dữ liệu còn lại.

- **Trực quan dữ liệu của từng thuộc tính so với các thuộc tính khác**



Hình 3.1.2b: Trực quan dữ liệu của từng thuộc tính so với những thuộc tính khác

Để xét sự tương tác giữa các cặp thuộc tính với nhau, ta có thể nhìn vào biểu đồ trên và đưa ra đánh giá ban đầu về sự tương tác đó. Ví dụ đối với biểu đồ của X1 so với X3 (hình 2.1.2), ta có thể thấy rằng các điểm dữ liệu phân bố hầu như rời rạc, cách xa nhau và không theo quy luật. Do đó ta có thể nhận xét rằng, nếu xét sự tương tác giữa cặp thuộc tính X1 và X3, chúng không ảnh hưởng nhiều đến sự thay đổi của nhau, thậm chí là không ảnh hưởng.

Biểu đồ nằm trên đường chéo chính(hình 2.1.2) nhận trục hoành làm thuộc tính đang xét (X1, X2, X3…Y) và nhận trục tung làm số lượng hay tần suất xuất hiện của từng giá trị. Biểu đồ trên đường chéo chính cho chúng ta biết mức độ phân bố của dữ liệu bên trong

từng thuộc tính đang xét. Từ đó có thể suy ra được dạng phân phối hay mức độ lệch (**skew**) của biểu đồ:

Bảng 3.1.2: Chỉ số mức độ lệch của từng thuộc tính dựa theo biểu đồ.

Tên thuộc tính	**Chỉ số lệch**
X1 transactrion date	-0.1500256905888924
X2 house age	0.38153741817729087
X3 distance to the nearest MRT station	1.8819063601148036
X4 number of convenience stores	0.1540458341286463
X5 latitude	-0.43700771816804596
X6 longitude	-1.2151682334072738
Y house price of unit area	0.5976770142537495

Ta nhận thấy hầu như tất cả chỉ số đều ở mức âm cao hoặc dương cao, trong khi một tập dữ liệu phân phối chuẩn phải có chỉ số lệch bằng 0. Do đó ta có thể nhận xét, dữ liệu bên trong toàn bộ thuộc tính phân bố một cách chênh lệch đầy biến động.

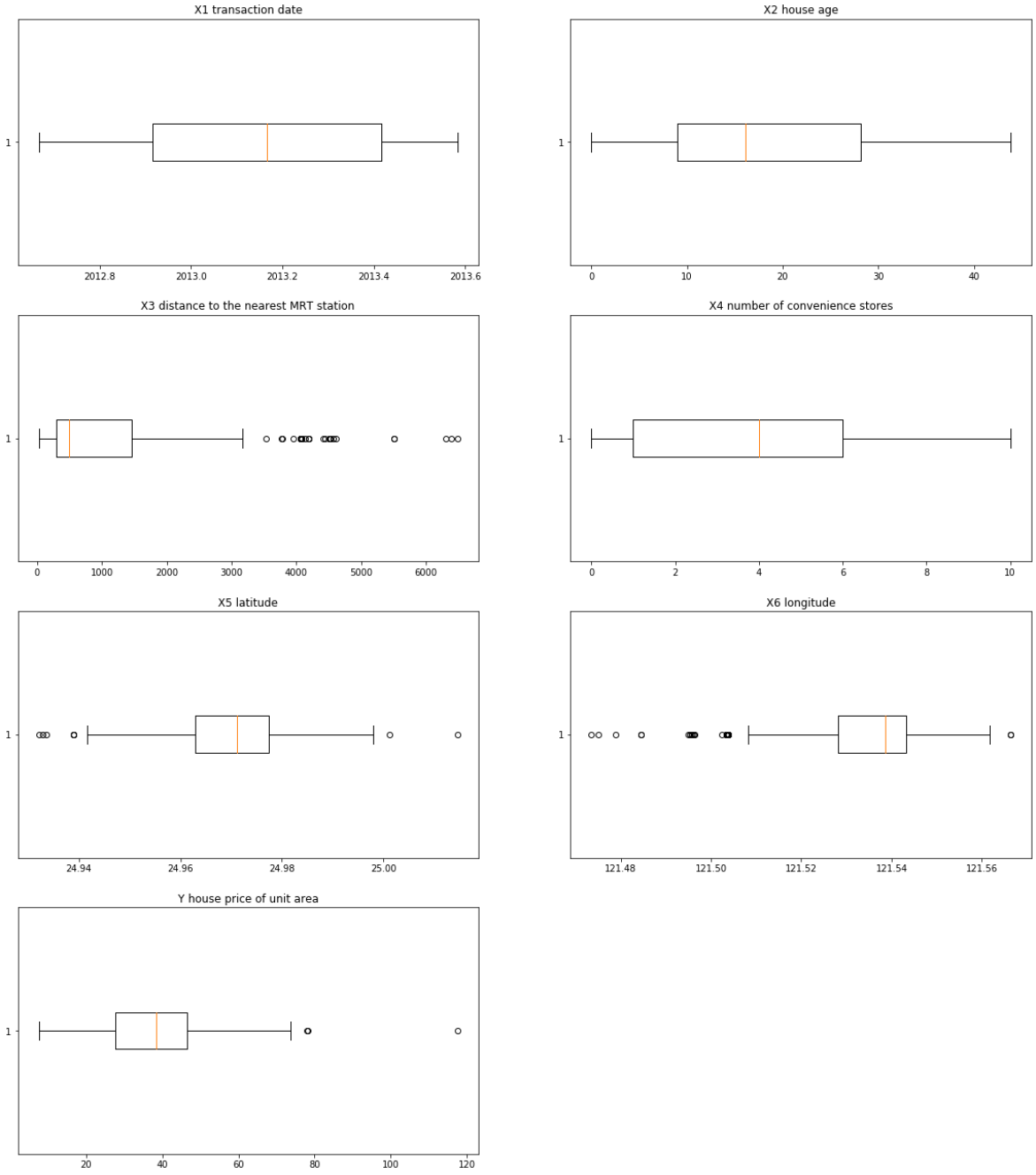
- **Ma trận tương quan**

Hình 3.1.2c: Ma trận tương quan của toàn bộ thuộc tính có trong tập dữ liệu

Để xét sự tương tác của từng cặp thuộc tính một cách chi tiết và rõ ràng hơn, ta có thể xét thông qua chỉ số tương quan của từng cặp thuộc tính ở ma trận tương quan như trên. Chỉ số tương quan biểu thị cho mức độ tương tác của thuộc tính này với thuộc tính kia, hay sự ảnh hưởng của thuộc tính này đến thuộc tính kia. Chỉ số càng lớn biểu thị mức độ tương quan càng cao. Số âm biểu thị tương quan nghịch biến và số dương biểu thị mức tương quan đồng biến.

Cụ thể, xét X1 và X3(hình 2.1.3) ta thấy chỉ số tương quan nằm ở mức rất nhỏ: 0.0096 và có màu tím tương ứng với mức tương quan **đồng biến** thấp. Điều này có nghĩa rằng mọi sự thay đổi của X1 sẽ làm thay đổi X3 ở một mức rất nhỏ, hầu như không có và ta có thể bỏ qua sự thay đổi nhỏ đó – đúng với sự phân tích biểu đồ hình 2.1.2. Nếu xét X3 và X6, ta thấy chỉ số tương quan ở mức âm cao: -0.81 và có màu đen tương ứng mức tương quan **nghịch biến** cao. Ta nói rằng, X3 và X6 có tác động đến sự thay đổi lẫn nhau trong quá trình vận hành của dữ liệu. Hiện tượng này được gọi là **Đa cộng tuyến**, có ảnh hưởng xấu nếu xây dựng mô hình và sự đánh giá sau này, cụ thể là mô hình hồi quy tuyến tính đa biến.

- **Biểu đồ hộp**



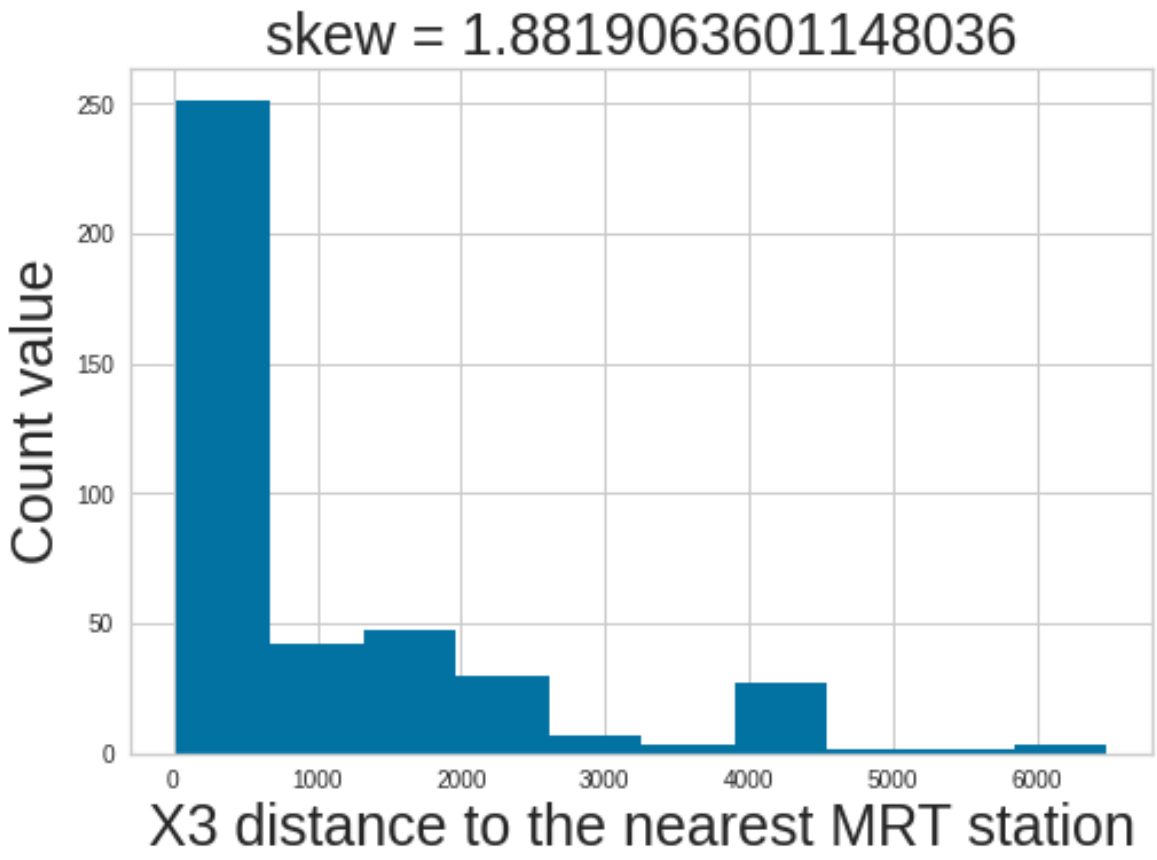
Hình 3.1.2d: Biểu đồ hộp của từng thuộc tính.

Biểu đồ hộp cho thấy sự phân bố của dữ liệu bên trong từng thuộc tính. Dữ liệu nằm giữa hai đầu của đường kẻ là dữ liệu tốt. Dữ liệu nằm ngoài khoảng này là dữ liệu phân bố

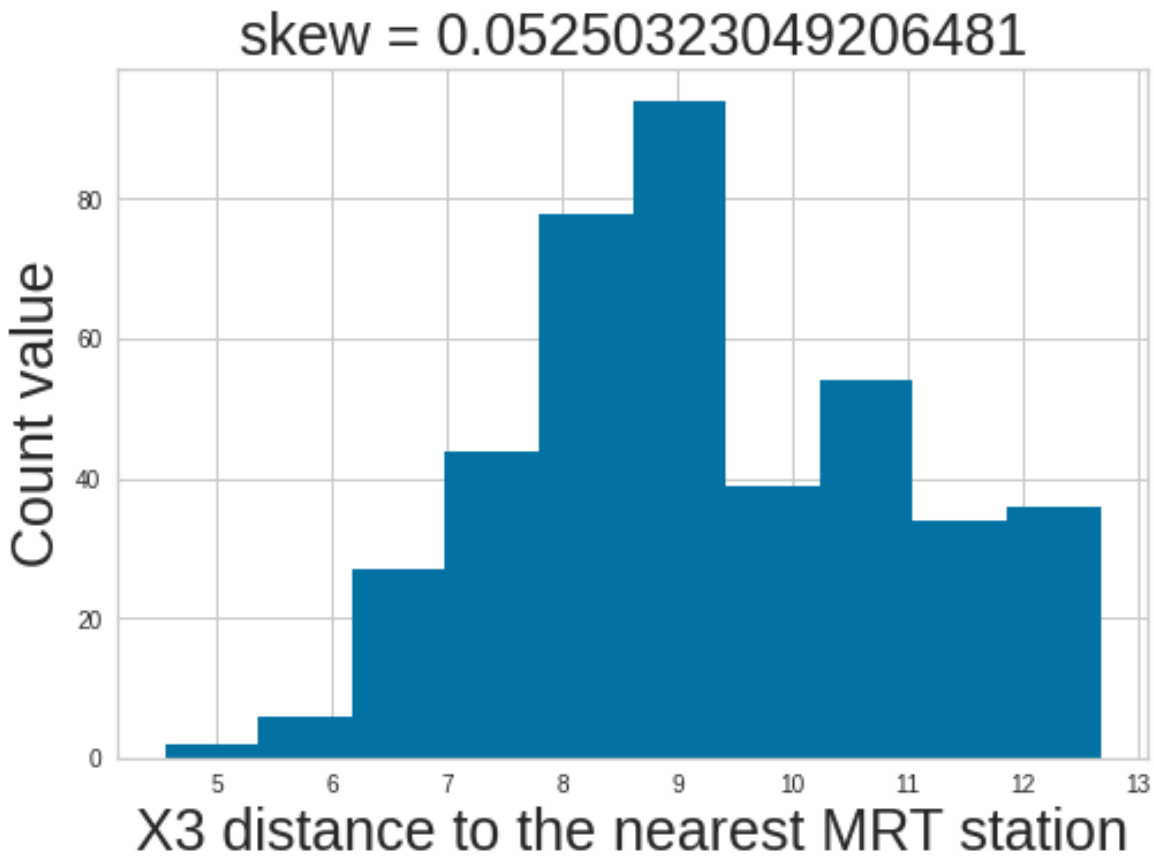
không tốt. Cụ thể: ở các biểu đồ hộp X3 có chứa các giá trị xấp xỉ lớn hơn 3000 trở đi nằm ngoài miền biểu diễn của hộp, do đó chúng có khả năng là các giá trị nhiễu, hoặc là các ngoại lệ. Tương tự như X5, X6, Y cũng có các giá trị nằm ngoài miền biểu diễn.

3.2. Xử lý ngoại lệ

Dựa vào quá trình phân tích ảnh hưởng của các yếu tố X đến giá nhà Y và sự phân bố dữ liệu bên trong từng thuộc tính, kết hợp chỉ số lệch và ma trận tương quan, ta có thể xem xét để xử lý, loại bỏ những ngoại lệ không phù hợp.



Hình 3.2a: Biểu đồ phân bố của dữ liệu bên trong X3.



Hình 3.2b: Biểu đồ phân bố của dữ liệu bên trong log(X3).

Đối với biến X3, ta thấy biểu đồ đang bị lệch về phía bên trái, chỉ số skew đang ở mức dương ≈ 1.88 chứng tỏ mức độ phân tán của dữ liệu khá cao. Sau khi dùng thuật toán **logarit cơ số 10** cho toàn bộ dữ liệu có trong X3, biểu đồ trở nên cân bằng hơn, chỉ số skew giảm mạnh về mức ≈ 0.05. Điều này cho thấy quá trình biến đổi dữ liệu thuộc tính X3 về dạng **logarit cơ số 10** sẽ làm giảm độ lệch của dữ liệu, thuận lợi hơn cho quá trình phân tích.

Để loại bỏ ngoại lệ, ta sử dụng công thức **IQR** cho dữ liệu mà ta xét.

Ví dụ, đối với thuộc tính X3, ta sẽ loại bỏ ngoại lệ như sau:

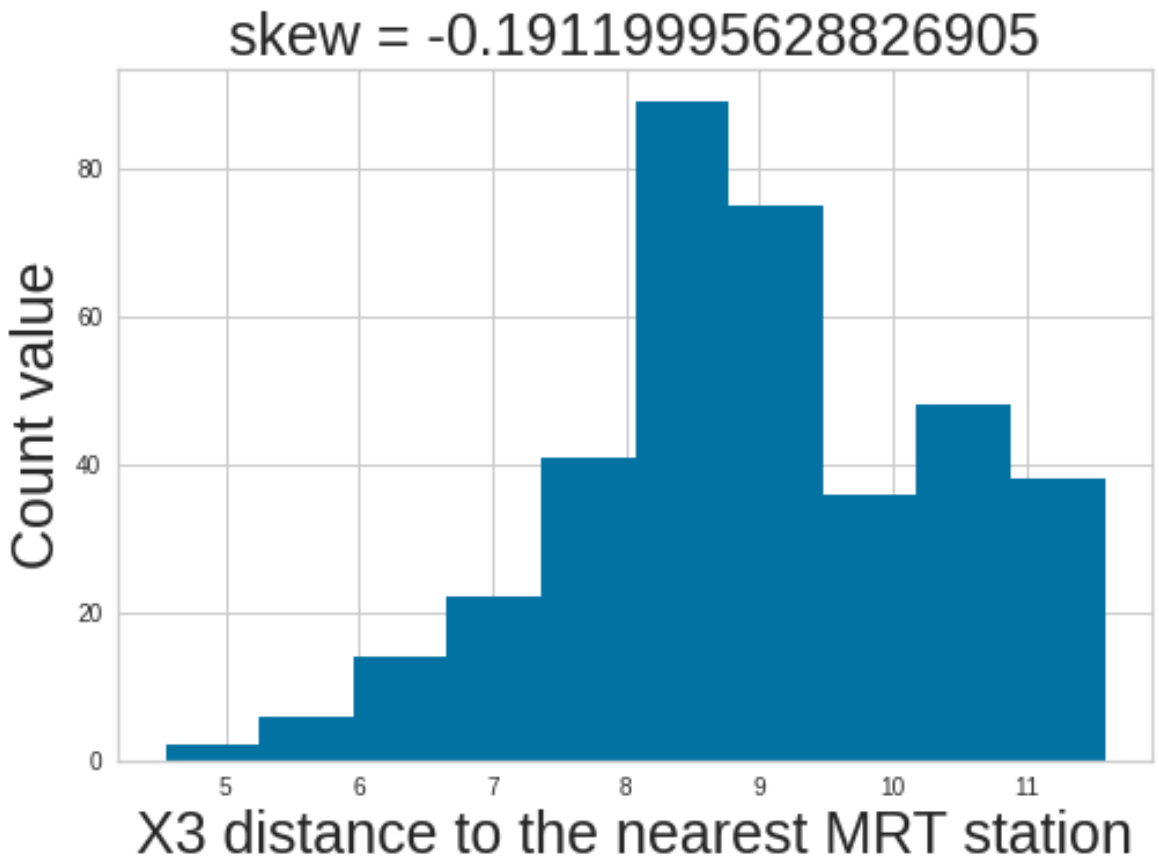
[Q1 + (IQR)*1.5 <= X3 <= Q3 + (IQR)*1.5]

Trong đó: Q1 là mức **tứ phân vị** thứ 1 (0.25) của X3.

Q3 là mức **tứ phân vị** thứ 3 (0.75) của X3.

IQR = Q3 – Q1.

1.5 là hệ số tùy chỉnh theo mức người dùng đặt.



Hình 3.2c: Biểu đồ phân bố của dữ liệu bên trong X3 sau khi loại bỏ ngoại lệ.

Sau khi loại bỏ ngoại lệ, biểu đồ biểu diễn dữ liệu bên trong X3 đã giảm lệch đáng kể, nhưng chỉ số lệch ở mức ≈ -0.191 là lớn hơn so với việc sử dụng thuật toán **logarit cơ số 10**. Để xét xem quá trình xây dựng mô hình hồi quy, phân tích có phù hợp với bộ dữ liệu được loại bỏ ngoại lệ này hay không, ta sẽ sử dụng dữ liệu này (X3 sau khi sử dụng thuật toán **logarit cơ số 10** sau đó loại bỏ ngoại lệ) để phân tích.

Thực hiện tương tự đối với các thuộc tính X5, X6, Y để xử lý ngoại lệ. Ta thu được kết quả như sau:

Bảng 3.2a : Thay đổi của dữ liệu trước và sau khi xử lý.

Tên thuộc tính cũ	**Miền giá trị cũ**	**Tên thuộc tính mới**	**Miền giá trị mới**
No	[1, 414]	-	-

X1 transactrion date	[2012.667, 2013.583]	X1 transactrion date	[2012.667, 2013.583]
X2 house age	[0, 43.8]	X2 house age	[0, 43.8]
X3 distance to the nearest MRT station	[**23.38284, 6488.021**]	X3 distance to the nearest MRT station	[**4.547378, 11.59113**]
X4 number of convenience stores	[0, 10]	X4 number of convenience stores	[0, 10]
X5 latitude	[**24.93207, 25.01459**]	X5 latitude	[**24.94883, 24.998**]
X6 longitude	[**121.4735, 121.5663**]	X6 longitude	[**121.5083, 121.5617**]
Y house price of unit area	[7.6, **117.5**]	Y house price of unit area	[7.6, **73.6**]

Bảng 3.2b: Thông tin các thuộc tính sau khi thăm dò và tiền xử lý.

Tên thuộc tính	**Miền dữ liệu**	**Giá trị trung bình**
X1 transactrion date	[2012.667, 2013.583]	2013.144003
X2 house age	[0, 43.8]	17.475202
X3 distance to the nearest MRT station	[4.547378, 11.59113]	8.925642
X4 number of convenience stores	[0, 10]	4.493261

X5 latitude	[24.94883, 24.998]	24.971224
X6 longitude	[121.5083, 121.5617]	121.536577
Y house price of unit area	[7.6, 73.6]	39.527763
Tổng:	371 điểm dữ liệu, 7 thuộc tính.	

3.2. Xử lý dữ liệu nâng cao: Phân cụm dữ liệu

Theo như quan sát, ta có thể thấy bộ dữ liệu có chứ hai thuộc tính X5 latitude và X6 longitude là vĩ độ và kinh độ của ngôi nhà. Thông thường, kinh độ và vĩ độ luôn tồn tại thành cặp tưng ứng, thể hiện vị trí địa lý nhưng ở bộ dữ liệu này chúng tồn tại rời rạc nhau. Chính vì vậy, chúng ta sẽ thực gom nhóm hai thuộc tính này thành khu vực địa lý bằng phương pháp phân cụm (Clustering), mà cụ thể là sử dụng phương pháp K-Means Clustering.

3.2.1. Giới thiệu về phân cụm dữ liệu

Phân cụm là một kỹ thuật học không giám sát (Unsupervised), được sử dụng trong khai phá dữ liệu nhằm phân chia tập dữ liệu ban đầu thành các cụm riêng biệt mà tại đó dữ liệu trong từng cụm có sự tương đồng với nhau, trái lại, dữ liệu nằm ngoài cụm có sự khác biệt phân biệt được với dữ liệu bên trong cụm.

Số cụm được xác định tùy theo kinh nghiệm làm việc hoặc dựa vào thuật toán được cài đặt để tự động xác định.

Mục tiêu của phân cụm dữ liệu là khai thác đặc điểm, tính năng, tiện ích, thông tin chung của từng nhóm dữ liệu được đưa vào để ứng dụng vào các lĩnh vực khác nhau như xử lý ảnh, phân tích kinh doanh, nghiên cứu thị trường,…

Quá trình phân cụm trải qua nhiều bước, nhưng chung quy vẫn phải trải qua các bước thăm dò, tiền xử lý dữ liệu để loại bỏ các yếu tố gây nhiễu, ảnh hưởng xấu đến chất lượng và kết quả nghiên cứu.

- **Một số ứng dụng của phân cụm dữ liệu**

– Xử lý ảnh: Phân cụm đa mô hình để phân đoạn ảnh viễn thám. Cung cấp chính xác thông tin dựa theo bước sóng được xác định.

- Phân tích kinh doanh: Nhóm đối tượng khách hàng thành từng cụm riêng biệt. Mỗi nhóm đối tượng lại có một tiềm năng riêng, khả năng sinh lời riêng. Do đó, có thể thay đổi chính sách và cơ cấu sản xuất phù hợp, tăng doan thu.
- Nghiên cứu thị trường: Sự thay đổi của cơ cấu sản xuất, chuyển biến của từng ngành cụ thể được phân cụm rõ ràng. Dựa vào kết quả thu được để thay đổi chiến lược, thay thế mô hình sản xuất, bình ổn thị trường.
- Phân nhóm bệnh nhân: Mỗi nhóm bệnh nhân có chung đặc điểm, triệu chứng, số liệu y tế sẽ được phân cụm để đưa ra những phương pháp chữa bệnh, các loại thuốc và phương hướng nghiên cứu y học sau này.
- Phân loại gian lận báo cáo tài chính: Nghiên cứu số liệu trên mỗi báo cáo tài chính của các công ty, phân cụm từng nhóm đối tượng. Từ đó xác định đặc điểm, tỷ suất, dấu hiệu phát hiện gian lận trong báo cáo tài chính.

• K-Means Clustering

❖ **Định nghĩa**

K-Means Clustering là một phương pháp phân cụm không giám sát (Unsupervised), sử dụng khoảng cách giữa các điểm dữ liệu với nhau để tính toán tìm ra được trung tâm cụm thích hợp(Centroid) và phân nhóm từng điểm dữ liệu vào từng trung tâm thích hợp nhất.

❖ **Tóm tắt thuật toán**

Đầu vào: Dữ liệu X và số lượng cluster cần tìm K.

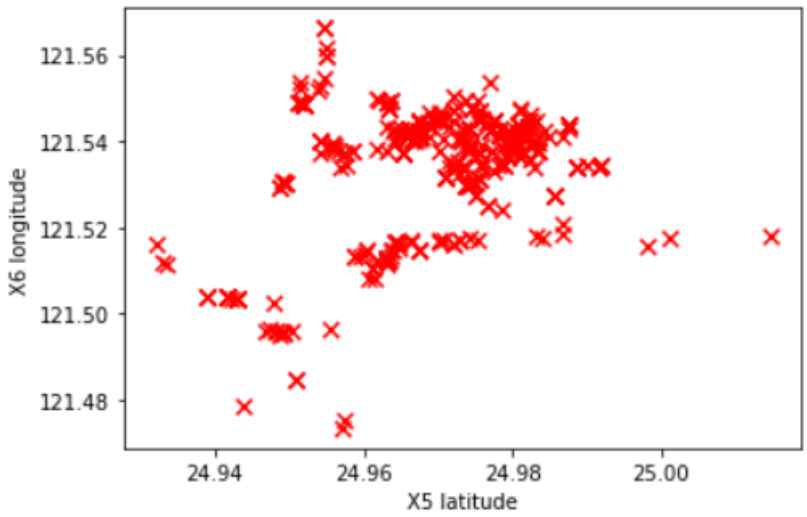
Đầu ra: Các center M và label vector cho từng điểm dữ liệu Y.

- Chọn K điểm bất kỳ làm các center ban đầu.
- Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
- Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
- Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
- Quay lại bước 2.
- Chúng ta có thể đảm bảo rằng thuật toán sẽ dừng lại sau một số hữu hạn vòng lặp.

3.2.2. Thực hiện phân cụm

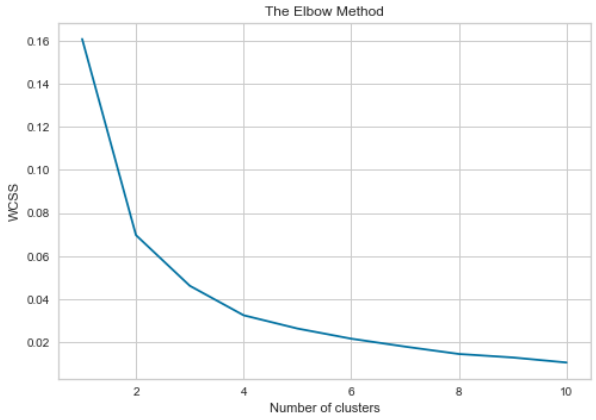
– Thực hiện trên hai bộ dữ liệu: đã qua xử lý và chưa xử lý.

• Phân cụm trên bộ dữ liệu chưa xử lý



Hình 3.2.2a: Ảnh Trực quan hóa vị trí địa lý của các ngôi nhà theo kinh độ và vĩ độ trên bộ dữ liệu chưa xử lý.

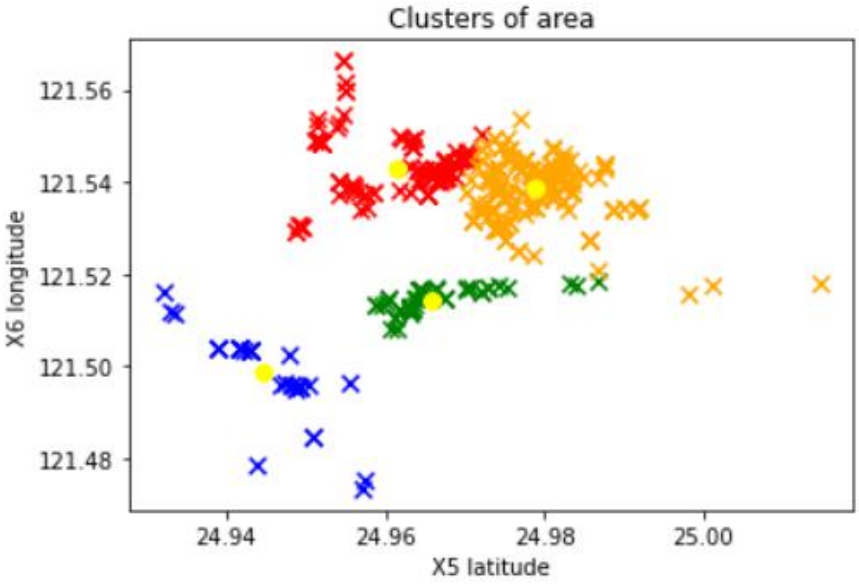
Đầu tiền, ta sẽ dựa vào ảnh trực quan hóa giá trị lỗi của K-Means để xác định số lượng cụm thích hợp:



Hình 3.2.2b: Ảnh trực quan hóa giá trị lỗi của K-Means trên bộ dữ liệu chưa xử lý.

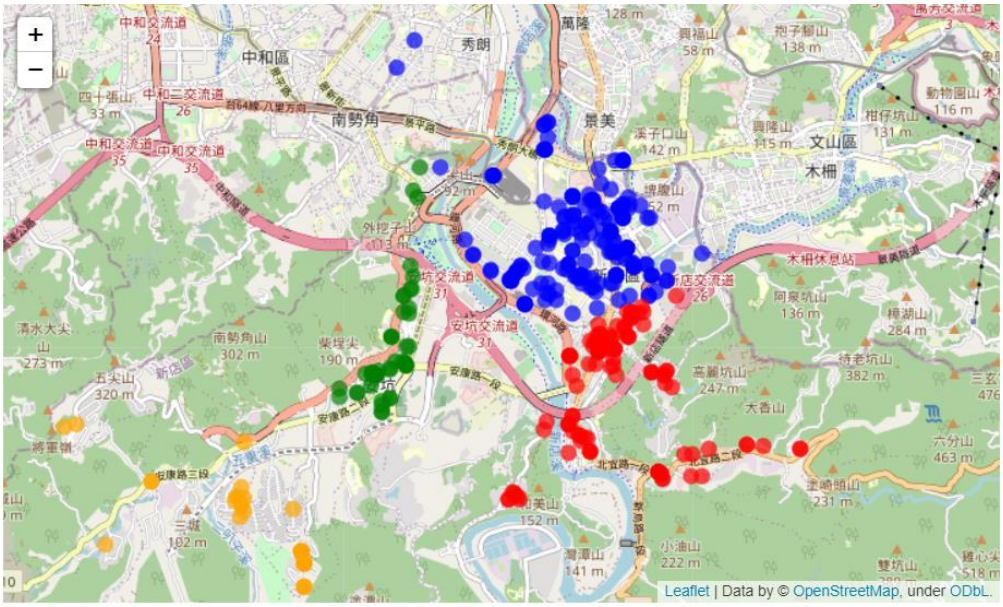
Ở đây ta có thể thấy giá trị hàm lỗi giảm mạnh k=1 đến k=4, sau đó giảm nhẹ dần về sau. Do đó, chọn k=4 là số lượng cụm hợp lý.

Tiếp đến, thực hiện huấn luyện mô hình với n_cluster=4 ta sẽ được kết quả như sau:



Hình 3.2.2c: Trực quan kết quả sau khi phân cụm trên bộ dữ liệu chưa xử lý

Và đây là ảnh được thể hiện chúng trên bản đồ:



Hình 3.2.2d: Ảnh phân cụm trên dữ liệu chưa xử lý được thể hiện trên bản đồ.

Cuối cùng, thay thế kinh độ và vĩ độ bằng thuộc tính khu vực địa lý X5 Area, ta thu được một dataset mới như sau:

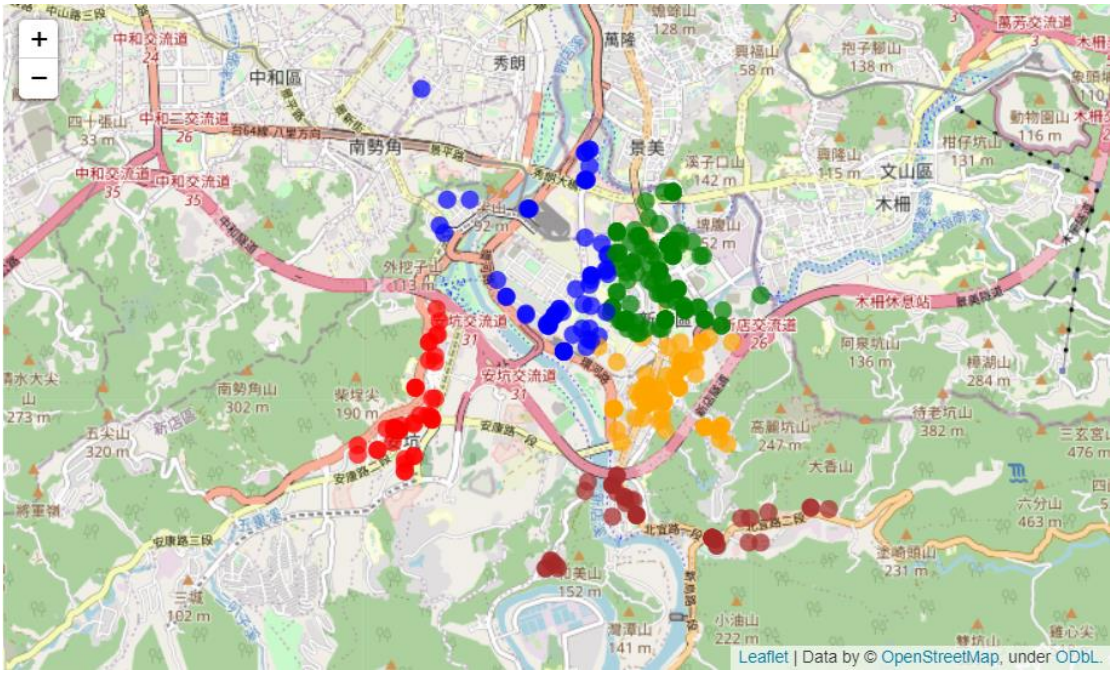
	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	Y house price of unit area	X5 Area
0	1	2012.917	32.0	84.87882	10	37.9	3
1	2	2012.917	19.5	306.59470	9	42.2	3
2	3	2013.583	13.3	561.98450	5	47.3	3
3	4	2013.500	13.3	561.98450	5	54.8	3
4	5	2012.833	5.0	390.56840	5	43.1	3
...	...	...	...	...	...	...	...
409	410	2013.000	13.7	4082.01500	0	15.4	2
410	411	2012.667	5.6	90.45606	9	50.0	3
411	412	2013.250	18.8	390.96960	7	40.6	3
412	413	2013.000	8.1	104.81010	5	52.5	0
413	414	2013.500	6.5	90.45606	9	63.9	3

414 rows × 7 columns

Hình 3.2.2e: Bộ dữ liệu thu được sau khi đã thực hiện phân cụm trên dữ liệu chưa xử lý.

• **Thực hiện trên dữ liệu đã qua xử lý**

Thực hiện tương tự như trên, dựa vào kinh độ, vĩ độ ta phân được thành 5 khu vực địa lý như ảnh bên dưới và thu được một bộ dữ liệu mới.



Hình 3.2.2f: Ảnh phân cụm các khu vực địa lý được thể hiện trên bảng đồ.

3.3. KIỂM ĐỊNH T-TEST MỘT YẾU TỐ

Để xét mức độ ảnh hưởng của một biến X đến kết quả Y, ta sử dụng **t-test** để kiểm tra mức ý nghĩa của nó có phù hợp hay không. Bảng dưới đây mô tả giá trị **p-value** của phép kiểm định **t-test** cho từng thuộc tính X so với Y trên bộ dữ liệu đã được xử lý ở mức ý nghĩa α = **0.05**.

Bảng 3.3: Giá trị p-value của từng biến độc lập X so với biến phụ thuộc Y

Tên thuộc tính	**Giá trị p-value (t-test)**
X1 transactrion date	0.0767 (.)
X2 house age	<2e-16 (***)
X3 distance to the nearest MRT station	<2e-16 (***)
X4 number of convenience stores	<2e-16 (***)
X5 latitude	<2e-16 (***)

X6 longitude	<2e-16 (***)

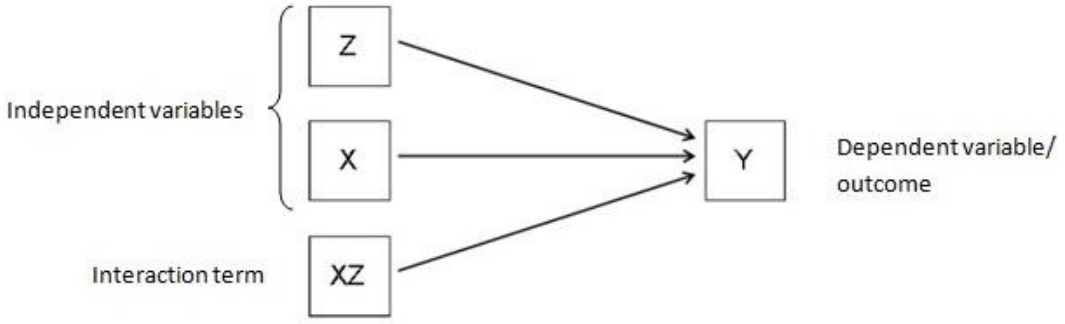
Chú thích: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Dựa vào giá trị p-value có trong bảng 4.1 ta thấy: nếu xét sự ảnh hưởng của từng thuộc tính X đến kết quả Y, với mức ý nghĩ $\alpha = 0.05$, thuộc tính X1 không có ý nghĩa thống kê (0.0767 > 0.05), còn lại các thuộc tính khác đều có ý nghĩa thống kê. Nói cách khác, nếu xét mô hình hồi quy tuyến tính đơn biến, ngoài mô hình X1 ~ Y không có ý nghĩa thì các mô hình còn lại đều có ý nghĩa để phân tích.

3.4. Mô hình hồi quy

Mô hình được xây dựng dựa trên những thuật toán cơ bản cũng như tìm kiếm các mối liên hệ giữa những yếu tố (interaction) ảnh hưởng đến kết quả để đưa vào mô hình.

3.4.1. Interaction



(Source: Barron & Kenny, 1986)

Các biến độc lập X và Z có ảnh hưởng đến kết quả Y. Ngoài ra, các biến kết hợp như XZ, cũng có tác động đến Y. Điều này có nghĩa là nếu loại bỏ Z sẽ thay đổi tác động của X tới Y và ngược lại.

3.4.2. Interaction in R

Để kiểm tra tương tác giữa các yếu tố có ý nghĩa thống kê hay không, hay nói cách khác là có ảnh hưởng đến kết quả hay không ta sử dụng hàm aov() trong R. Ví dụ muốn kiểm tra tương tác giữa ba biến x1,x2,x3 ảnh hưởng đến y ta thực hiện đoạn code sau:

```
av<-aov(y~(x1+x2+x3)^3, data=mydata)
summary(av)
```

Khi đó, tất cả các tương tác giữa các yếu tố sẽ được liệt kê với P(>F) tương ứng. Việc cần làm là chọn ra các tương tác có ý nghĩa để kiểm tra trong mô hình.

3.4.3. Tổng quan về Regression (Phân Tích Hồi Quy)

Regression là phương pháp nghiên cứu mối quan hệ giữa 2 biến mà cụ thể một biến sẽ là biến độc lập (ảnh hưởng đến biến mục tiêu), và biến còn lại sẽ là biến mục tiêu (bị ảnh hưởng bởi biến độc lập), mô hình hóa, định lượng hóa mối quan hệ này để qua đó có thể xác định được giá trị của biến mục tiêu nếu các biến độc lập thay đổi như thế nào. Và phân tích hồi quy sẽ cho ra kết quả dự báo của biến mục tiêu.

Trong bài này, chúng ta quan tâm đến các mô hình hồi quy sau:

- **Linear Regression:** đây được xem là mô hình đơn giản nhất nghiên cứu mối quan hệ tuyến tính giữa một biến độc lập và biến phụ thuộc, áp dụng cho biến định lượng, và đồ thị là dạng đường thẳng.
- **Multiple regression:** áp dụng cho nghiên cứu mối quan hệ của nhiều biến độc lập và một biến phụ thuộc, áp dụng cho biến định lượng.
- **Polynominal Regression:** áp dụng cho các trường hợp mà biến độc lập x có bậc mũ lớn hơn 1, và y là biến định lượng. Đồ thị là một đường cong.
- **Ridge Regression:** phân tích mối quan hệ giữa các biến độc lập và biến phụ thuộc sử dụng phương pháp điều chỉnh mô hình sao cho giảm thiểu các vấn đề Overfitting, tối ưu hay kiểm soát mức độ phức tạp của mô hình để cân đối giữa Biased và Variance qua đó giảm sai số của mô hình.

3.4.4. Regression in R

Trong R đã tích hợp sẵn các thuật toán Regression với các hàm tương ứng.

Với Linear Regression, Multiple Regression, Polynominal Regression, ta gọi hàm lm() để tiến hành cài đặt thuật toán, và hàm summary() để xem toàn bộ các yếu tố đánh giá mô hình (R-squared,…). Tuy nhiên, ở mỗi thuật toán thì phần công thức có sự khác biệt.

Ví dụ:

Linear Regression:

relation <- lm(formula = **y ~ x**)

summary(relation)

Multiple Regression :

relation <- lm(formula = **y ~ x1 + x2+ x3**)

summary(relation)

Polynominal Regression:

relation <- lm(formula = **y ~ x1 + I(x1^2)+I(x1^3)**)

summary(relation)

Ngoài ra chúng ta có thể sử dụng công thức mở rộng khác, như cộng thêm tích giữa các yếu tố, logarit của yếu tố, … (khi giữa chúng có interaction) vào mô hình.

3.4.5. R-squared và Adjust R-Squared

- **R-squared:**

Công thức:

$$R^2 = 1 - (ESS / TSS)$$

R^2: tổng các độ lệch bình phương giải thích từ hồi quy

ESS: tổng các độ lệch bình phương phần dư

TSS: tổng các độ lệch bình phương toàn bộ

Cho biết mức độ phù hợp của mô hình với tập dữ liệu, hay mức độ giải thích của các biến độc lập với biến phụ thuộc.

Càng đưa thêm nhiều biến vào mô hình thì giá trị **R2** sẽ tăng. Lý do là khi càng đưa thêm biến giải thích vào mô hình thì sẽ càng khiến phần dư giảm xuống, trong khi **TSS** không đổi, dẫn tới **R2** luôn luôn tăng.

Giá trị **R2** tăng khả năng giải thích của mô hình, nhưng bản chất thì lại không làm rõ được tầm quan trọng của biến đưa vào, do đó nếu dựa vào giá trị **R2** để đánh giá tính hiệu quả của mô hình sẽ dẫn đến tình huống không chính xác vì sẽ đưa quá nhiều biến không cần thiết, làm phức tạp mô hình.

Để ngăn chặn tình trạng như đã nêu trên, một phép đo khác về mức độ thích hợp được sử dụng thường xuyên hơn. Phép đo này gọi là **Adjust R-Squared.**

- **Adjust R-Squared:**

Công thức:

$$R^2 = 1 - \frac{ESS/(n-k)}{TSS/(n-1)} = 1 - \frac{ESS(n-1)}{TSS(n-k)}$$

n: số lượng mẫu quan sát.

k: số tham số của mô hình, bằng số lượng biến độc lập cộng 1.

R2(R2-Adjusted): hệ số R bình phương hiệu chỉnh.

Việc thêm vào một biến dẫn đến tăng **R2** nhưng cũng làm giảm đi một bậc tự do. **R2 hiệu chỉnh** là một phép đo độ thích hợp tốt hơn bởi vì nó cho phép đánh đổi giữa việc tăng R2 và giảm bậc tự do.

Trong quá trình xây dựng mô hình, chúng ta sẽ quan tâm tới giá trị **Adjust R-Squared** để đánh giá cũng như so sánh giữa các mô hình để lựa chon được mô hình phù hợp nhất.

3.4.6. Xây dựng mô hình hồi quy

Với mục đích:

- Tìm hiểu về sự ảnh hưởng của quá trình làm sạch và tiền xử lý dữ liệu đối kết quả xây dựng mô hình hồi quy dự đoán.

- Khảo sát việc phân cụm nhà theo khu vực địa lý dựa vào kinh độ và vĩ độ sẽ cho kết quả mô hình khác biệt như thế nào so với việc xây dựng mô hình trên bộ dữ liệu có kinh độ và vĩ độ rời rạc.

Thực hiện xây dưng mô hình hồi quy đa biến (sử dụng lm()) và mô hình hồi quy Ridge trên cả bốn bộ dữ liệu:

1. Dữ liệu chưa qua xử lý và chưa phân cụm.
2. Dữ liệu chưa qua xử lý và đã phân cụm.
3. Dữ liệu đã xử lý và chưa phân cụm.
4. Dữ liệu đã xử lý và đã phân cụm.

- **Dữ liệu thực hiện**

1. Dữ liệu chưa qua xử lý và chưa phân cụm.

- Số điểm dự liệu: 414 điểm
- Số thuộc tính: 7
  - X1 transaction date

- X2 house age
- X3 distance to the nearest MRT Station
- X4 number of convenience stores
- X5 latitude
- X6 longitude
- Y house price of unit area

2. Dữ liệu chưa qua xử lý và đã phân cụm.

- Số điểm dự liệu: 414 điểm
- Số thuộc tính: 6
  - X1 transaction date
  - X2 house age
  - X3 distance to the nearest MRT Station
  - X4 number of convenience stores
  - X5 Area
  - Y house price of unit area

3. Dữ liệu đã xử lý và chưa phân cụm.

- Số điểm dự liệu: 371 điểm
- Số thuộc tính: 7
  - X1 transaction date
  - X2 house age
  - X3 distance to the nearest MRT Station
  - X4 number of convenience stores
  - X5 latitude
  - X6 longitude
  - Y house price of unit area

4. Dữ liệu đã xử lý và đã phân cụm.

- Số điểm dự liệu: 371 điểm
- Số thuộc tính: 6
  - X1 transaction date
  - X2 house age
  - X3 distance to the nearest MRT Station

- X4 number of convenience stores
- X5 Area
- Y house price of unit area

• Chia dữ liệu

Nhằm xây dựng và đánh giá chất lượng mô hình hồi quy một cách chính xác, ta sẽ thực hiện chia dữ liệu thành 2 phần Train và Test theo tỷ lệ (8:2).

Bảng 3.4.6a: Chia dữ liệu Train-Test

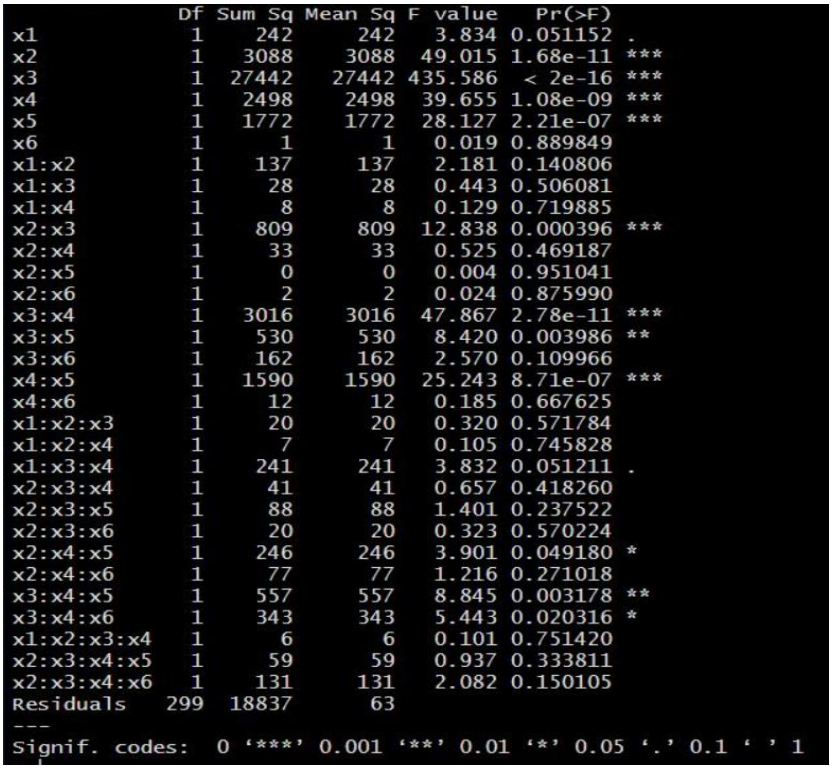
STT	Bộ dữ liệu	TRAIN	TEST
1	Dữ liệu chưa qua xử lý và chưa phân cụm.	331	83
2	Dữ liệu chưa qua xử lý và đã phân cụm.		
3	Dữ liệu đã xử lý và chưa phân cụm.	296	75
4	Dữ liệu đã xử lý và đã phân cụm.		

• Khảo sát ảnh hưởng của các tương tác

Sử dụng Two-way ANOVA Table để xác định ảnh hưởng của các tương tác giữa các yếu tố đối với giá nhà.

Kết quả thu được:

a. Dữ liệu chưa qua xử lý và chưa phân cụm.

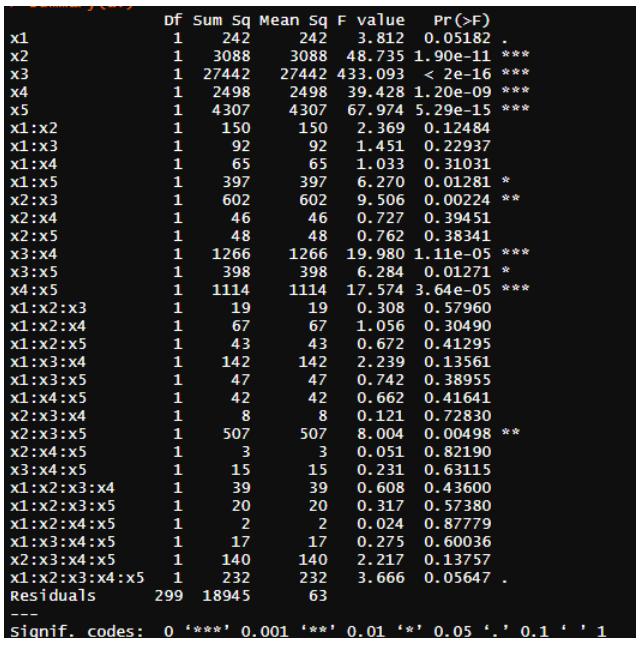


	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x1	1	242	242	3.834	0.051152	.
x2	1	3088	3088	49.015	1.68e-11	***
x3	1	27442	27442	435.586	< 2e-16	***
x4	1	2498	2498	39.655	1.08e-09	***
x5	1	1772	1772	28.127	2.21e-07	***
x6	1	1	1	0.019	0.889849	
x1:x2	1	137	137	2.181	0.140806	
x1:x3	1	28	28	0.443	0.506081	
x1:x4	1	8	8	0.129	0.719885	
x2:x3	1	809	809	12.838	0.000396	***
x2:x4	1	33	33	0.525	0.469187	
x2:x5	1	0	0	0.004	0.951041	
x2:x6	1	2	2	0.024	0.875990	
x3:x4	1	3016	3016	47.867	2.78e-11	***
x3:x5	1	530	530	8.420	0.003986	**
x3:x6	1	162	162	2.570	0.109966	
x4:x5	1	1590	1590	25.243	8.71e-07	***
x4:x6	1	12	12	0.185	0.667625	
x1:x2:x3	1	20	20	0.320	0.571784	
x1:x2:x4	1	7	7	0.105	0.745828	
x1:x3:x4	1	241	241	3.832	0.051211	.
x2:x3:x4	1	41	41	0.657	0.418260	
x2:x3:x5	1	88	88	1.401	0.237522	
x2:x3:x6	1	20	20	0.323	0.570224	
x2:x4:x5	1	246	246	3.901	0.049180	*
x2:x4:x6	1	77	77	1.216	0.271018	
x3:x4:x5	1	557	557	8.845	0.003178	**
x3:x4:x6	1	343	343	5.443	0.020316	*
x1:x2:x3:x4	1	6	6	0.101	0.751420	
x2:x3:x4:x5	1	59	59	0.937	0.333811	
x2:x3:x4:x6	1	131	131	2.082	0.150105	
Residuals	299	18837	63			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 3.4.6a: Two-way ANOVA table trên dữ liệu chưa xử lý-chưa phân cụm

b. Dữ liệu chưa qua xử lý và đã phân cụm.

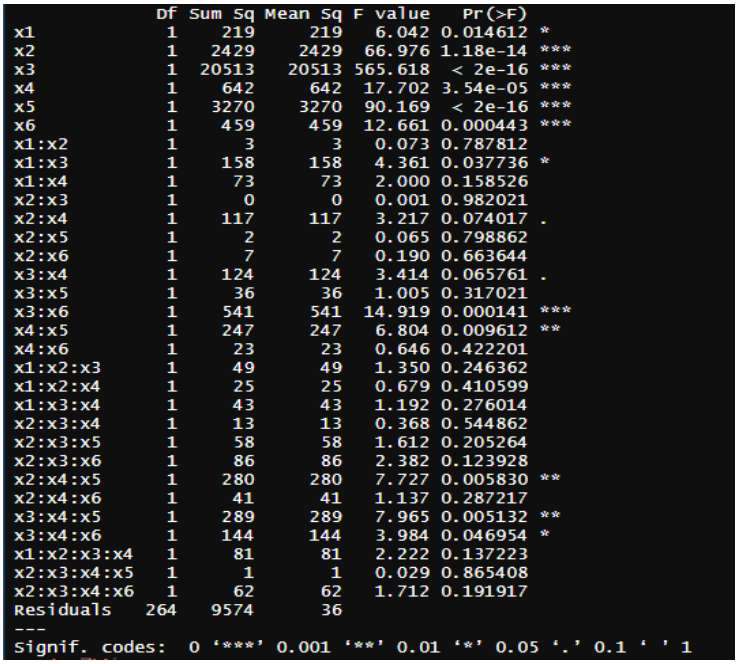


	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x1	1	242	242	3.812	0.05182	.
x2	1	3088	3088	48.735	1.90e-11	***
x3	1	27442	27442	433.093	< 2e-16	***
x4	1	2498	2498	39.428	1.20e-09	***
x5	1	4307	4307	67.974	5.29e-15	***
x1:x2	1	150	150	2.369	0.12484	
x1:x3	1	92	92	1.451	0.22937	
x1:x4	1	65	65	1.033	0.31031	
x1:x5	1	397	397	6.270	0.01281	*
x2:x3	1	602	602	9.506	0.00224	**
x2:x4	1	46	46	0.727	0.39451	
x2:x5	1	48	48	0.762	0.38341	
x3:x4	1	1266	1266	19.980	1.11e-05	***
x3:x5	1	398	398	6.284	0.01271	*
x4:x5	1	1114	1114	17.574	3.64e-05	***
x1:x2:x3	1	19	19	0.308	0.57960	
x1:x2:x4	1	67	67	1.056	0.30490	
x1:x2:x5	1	43	43	0.672	0.41295	
x1:x3:x4	1	142	142	2.239	0.13561	
x1:x3:x5	1	47	47	0.742	0.38955	
x1:x4:x5	1	42	42	0.662	0.41641	
x2:x3:x4	1	8	8	0.121	0.72830	
x2:x3:x5	1	507	507	8.004	0.00498	**
x2:x4:x5	1	3	3	0.051	0.82190	
x3:x4:x5	1	15	15	0.231	0.63115	
x1:x2:x3:x4	1	39	39	0.608	0.43600	
x1:x2:x3:x5	1	20	20	0.317	0.57380	
x1:x2:x4:x5	1	2	2	0.024	0.87779	
x1:x3:x4:x5	1	17	17	0.275	0.60036	
x2:x3:x4:x5	1	140	140	2.217	0.13757	
x1:x2:x3:x4:x5	1	232	232	3.666	0.05647	.
Residuals	299	18945	63			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 3.4.6b: Two-way ANOVA table trên dữ liệu chưa xử lý-đã phân cụm.

c. Dữ liệu đã xử lý và chưa phân cụm.

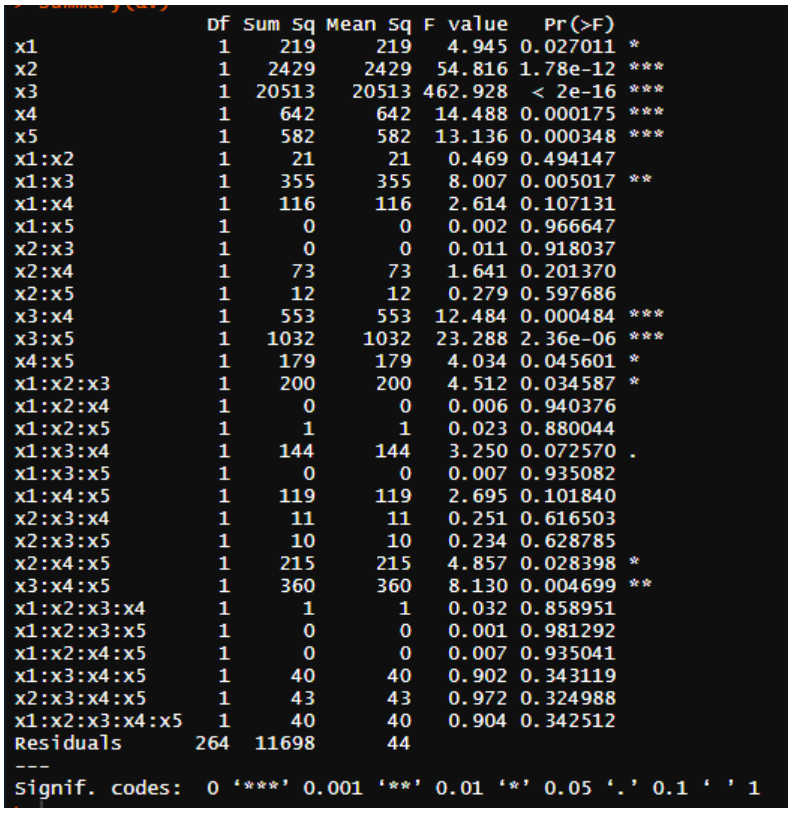


	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x1	1	219	219	6.042	0.014612	*
x2	1	2429	2429	66.976	1.18e-14	***
x3	1	20513	20513	565.618	< 2e-16	***
x4	1	642	642	17.702	3.54e-05	***
x5	1	3270	3270	90.169	< 2e-16	***
x6	1	459	459	12.661	0.000443	***
x1:x2	1	3	3	0.073	0.787812	
x1:x3	1	158	158	4.361	0.037736	*
x1:x4	1	73	73	2.000	0.158526	
x2:x3	1	0	0	0.001	0.982021	
x2:x4	1	117	117	3.217	0.074017	.
x2:x5	1	2	2	0.065	0.798862	
x2:x6	1	7	7	0.190	0.663644	
x3:x4	1	124	124	3.414	0.065761	.
x3:x5	1	36	36	1.005	0.317021	
x3:x6	1	541	541	14.919	0.000141	***
x4:x5	1	247	247	6.804	0.009612	**
x4:x6	1	23	23	0.646	0.422201	
x1:x2:x3	1	49	49	1.350	0.246362	
x1:x2:x4	1	25	25	0.679	0.410599	
x1:x3:x4	1	43	43	1.192	0.276014	
x2:x3:x4	1	13	13	0.368	0.544862	
x2:x3:x5	1	58	58	1.612	0.205264	
x2:x3:x6	1	86	86	2.382	0.123928	
x2:x4:x5	1	280	280	7.727	0.005830	**
x2:x4:x6	1	41	41	1.137	0.287217	
x3:x4:x5	1	289	289	7.965	0.005132	**
x3:x4:x6	1	144	144	3.984	0.046954	*
x1:x2:x3:x4	1	81	81	2.222	0.137223	
x2:x3:x4:x5	1	1	1	0.029	0.865408	
x2:x3:x4:x6	1	62	62	1.712	0.191917	
Residuals	264	9574	36			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 3.4.6c: Two-way ANOVA table trên dữ liệu đã xử lý-chưa phân cụm.

d. Dữ liệu đã xử lý và đã phân cụm.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x1	1	219	219	4.945	0.027011	*
x2	1	2429	2429	54.816	1.78e-12	***
x3	1	20513	20513	462.928	< 2e-16	***
x4	1	642	642	14.488	0.000175	***
x5	1	582	582	13.136	0.000348	***
x1:x2	1	21	21	0.469	0.494147	
x1:x3	1	355	355	8.007	0.005017	**
x1:x4	1	116	116	2.614	0.107131	
x1:x5	1	0	0	0.002	0.966647	
x2:x3	1	0	0	0.011	0.918037	
x2:x4	1	73	73	1.641	0.201370	
x2:x5	1	12	12	0.279	0.597686	
x3:x4	1	553	553	12.484	0.000484	***
x3:x5	1	1032	1032	23.288	2.36e-06	***
x4:x5	1	179	179	4.034	0.045601	*
x1:x2:x3	1	200	200	4.512	0.034587	*
x1:x2:x4	1	0	0	0.006	0.940376	
x1:x2:x5	1	1	1	0.023	0.880044	
x1:x3:x4	1	144	144	3.250	0.072570	.
x1:x3:x5	1	0	0	0.007	0.935082	
x1:x4:x5	1	119	119	2.695	0.101840	
x2:x3:x4	1	11	11	0.251	0.616503	
x2:x3:x5	1	10	10	0.234	0.628785	
x2:x4:x5	1	215	215	4.857	0.028398	*
x3:x4:x5	1	360	360	8.130	0.004699	**
x1:x2:x3:x4	1	1	1	0.032	0.858951	
x1:x2:x3:x5	1	0	0	0.001	0.981292	
x1:x2:x4:x5	1	0	0	0.007	0.935041	
x1:x3:x4:x5	1	40	40	0.902	0.343119	
x2:x3:x4:x5	1	43	43	0.972	0.324988	
x1:x2:x3:x4:x5	1	40	40	0.904	0.342512	
Residuals	264	11698	44			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Hình 3.4.6d: Two-way ANOVA table trên dữ liệu đã xử lý-đã phân cụm.

❖ Nhận xét:

– Có sự khác giao về số lượng tương tác giữa các yếu tố trên dữ liệu đã xử lý và chưa xử lý tương ứng.

– Có những yếu tố khi đứng riêng lẻ thì không ảnh hưởng đến kết quả nhưng khi kết hợp với yếu tố khác có thể ảnh hưởng đến kết quả. Cụ thể, ta có thể thấy ở bảng Two-way ANOVA trên dữ liệu chưa qua xử lý và chưa phân cụm: x6 không có ý nghĩa thống kê đối với kết quả nhưng sự tương tác giữa x3:x4:x6 thì có.

– Ở dữ liệu chưa xử lý, khi để X5 latitude và X6 longitude rời rạc thì X6 không có ý nghĩa thống đối với kết quả nhưng trên phân cụm theo khu vực địa lý thì yếu tố X5 Area lại có ý nghĩa thống kê.

• Định hướng thực hiện

– Sử dụng tất cả các yếu tố của dữ liệu Train để xây dựng mô hình hồi quy đa biến.

– Căn cứ vào Adjusted R-square để xác định chất lượng của mô hình.

a. Đối với các mô hình sử dụng hàm lm() để xây dựng mô hình (Linear, Multiple, Polynomial Regression):

– Dựa vào Two-way ANOVA table, lần lượt đưa các yếu tố và tương tác giữa các yếu tố có ảnh đến kết quả vào mô hình theo mức độ từ đơn giản đến phức tạp:

- Đầu tiên là các yếu tố đơn có ảnh hưởng đến kết quả:
  Ví dụ: x1, x2, x3, x4, ..
- Tiếp theo là lũy thừa của các yếu tố đơn có ảnh hưởng đến kết quả:
  Ví dụ: I(x2^2), I(x5^3), …
- Đưa tiếp các tương tác giữa 2 yếu tố, 3 yếu tố, … có ảnh hưởng đến kết quả vào mô hình:
  Ví dụ: I(x1*x2*x3*x4), I(x5*x6), …
- Và thêm tiếp những thứ phúc tạp hơn như log(), sqrt (), …

Thực hiện thêm đến khi mô hình cho Adjusted R-square cao nhất có thể.

– Đồng thời, trong suốt quá trình đó, ta lần lượt loại bỏ đi các yếu tố và tương tác không có ý nghĩa đến kết quả để tránh làm phức tạp mô hình.

b. Đối với mô hình sử dụng thư viện glmnet (Ridge Regression):

– Lựa chọn giá trị Lambda tối ưu thông qua hàm cv.glmnet()

– Tìm kiếm mô hình tốt nhất thông qua K-Cross Validation.

– Xây dựng mô hình cuối cùng.

– Đánh giá mô hình trên train và test.

- **Kết quả thu được**

a. Dữ liệu chưa qua xử lý và chưa phân cụm.

– **Mô hình hồi quy đa biến sử dụng hàm lm():**

- **Các biến tham gia vào mô hình:**

relation <- lm(formula = y ~ x1+x2+x3+x4+x5+I(x2^2)+I(x5^2) +I(x4*x3)+I(x5*x3)+I(x4*x5) +I(x1*x3*x4)+I(x5*x3*x4))

```
Residuals:
    Min      1Q  Median      3Q     Max
-33.997  -3.966  -0.600   2.930  70.345

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.751e+06  1.930e+06   1.944 0.052821 .
x1                1.062e+01  2.237e+00   4.745 3.15e-06 ***
x2               -9.546e-01  1.436e-01  -6.648 1.29e-10 ***
x3                6.545e+00  9.623e-01   6.802 5.14e-11 ***
x4                2.348e+03  5.538e+02   4.240 2.93e-05 ***
x5               -3.032e+05  1.546e+05  -1.962 0.050690 .
I(x2^2)           1.755e-02  3.480e-03   5.042 7.73e-07 ***
I(x5^2)           6.093e+03  3.096e+03   1.968 0.049931 *
I(x4 * x3)        5.236e+00  1.526e+00   3.432 0.000679 ***
I(x5 * x3)       -2.624e-01  3.856e-02  -6.804 5.07e-11 ***
I(x4 * x5)       -9.398e+01  2.218e+01  -4.238 2.96e-05 ***
I(x1 * x3 * x4) -1.708e-03  7.099e-04  -2.406 0.016679 *
I(x5 * x3 * x4) -7.201e-02  2.365e-02  -3.044 0.002525 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.636 on 318 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6899
F-statistic: 62.18 on 12 and 318 DF,  p-value: < 2.2e-16
```

- Adjusted R-square = 68,99%
- Tham số mô hình:

```
Coefficients:
    (Intercept)              x1              x2              x3
      3.751e+06       1.062e+01      -9.546e-01       6.545e+00
             x4              x5         I(x2^2)         I(x5^2)
      2.348e+03      -3.032e+05       1.755e-02       6.093e+03
      I(x4 * x3)      I(x5 * x3)      I(x4 * x5) I(x1 * x3 * x4)
       5.236e+00      -2.624e-01      -9.398e+01      -1.708e-03
I(x5 * x3 * x4)
      -7.201e-02
```

– **Mô hình Ridge Regression:**

- Giá trị Lambda tối ưu nhất: 0.5011872
- Tham số mô hình:

```
                                                  s0
(Intercept)                                -1.768350e+04
X3.distance.to.the.nearest.MRT.station     -4.024714e-03
X2.house.age                               -2.534493e-01
X6.longitude                                2.037883e+01
X1.transaction.date                         4.744628e+00
X4.number.of.convenience.stores             1.098555e+00
X5.latitude                                 2.281844e+02
```

- Adjusted R-square =56.42%

b. Dữ liệu chưa qua xử lý và đã phân cụm.

– **Mô hình hồi quy đa biến sử dụng hàm lm():**

- Các biến tham gia vào mô hình:

relation <- lm(formula = y ~ x1+ I(x2^2) +I(x2^3) +I(x1*x5) +I(x4*x5)
+sqrt(x4)+sqrt(I(x2*x3))+sqrt(I(x3*x4))
+I(x1*sqrt(x5)+I(x2*sqrt(x3))+I(x4*sqrt(x5)))

```
Residuals:
    Min      1Q  Median      3Q     Max
-33.086  -4.175  -0.319   3.130  68.570

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.267e+04  3.103e+03  -4.084 5.61e-05 ***
x1                6.312e+00  1.541e+00   4.095 5.36e-05 ***
I(x2^2)          -3.401e-02  7.232e-03  -4.703 3.82e-06 ***
I(x1 * x5)        7.269e-03  2.883e-03   2.522   0.0122 *
I(x4 * x5)       -3.874e+00  2.023e+00  -1.915   0.0563 .
I(x2^3)           6.805e-04  1.705e-04   3.991 8.16e-05 ***
sqrt(x4)          9.135e+00  1.340e+00   6.818 4.64e-11 ***
sqrt(I(x2 * x3)) -1.591e-01  3.524e-02  -4.515 8.91e-06 ***
sqrt(I(x3 * x4)) -1.881e-01  4.215e-02  -4.462 1.13e-05 ***
I(x1 * sqrt(x5)) -8.659e-03  4.867e-03  -1.779   0.0761 .
I(x2 * sqrt(x3))  2.536e-02  6.126e-03   4.140 4.45e-05 ***
I(x4 * sqrt(x5))  5.994e+00  3.468e+00   1.728   0.0849 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.699 on 319 degrees of freedom
Multiple R-squared:  0.6953,    Adjusted R-squared:  0.6848
F-statistic: 66.16 on 11 and 319 DF,  p-value: < 2.2e-16
```

- Adjusted R-square = 68,48%

- Tham số mô hình:

```
Coefficients:
     (Intercept)               x1            I(x2^2)        I(x1 * x5)        I(x4 * x5)           I(x2^3)
      -1.267e+04         6.312e+00        -3.402e-02         7.269e-03        -3.874e+00         6.805e-04
        sqrt(x4)  sqrt(I(x2 * x3))  sqrt(I(x3 * x4))  I(x1 * sqrt(x5))  I(x2 * sqrt(x3))  I(x4 * sqrt(x5))
        9.135e+00        -1.591e-01        -1.881e-01        -8.659e-03         2.536e-02         5.994e+00
```

– **Mô hình Ridge Regression:**

- Giá trị Lambda tối ưu nhất: 0.1995262

• Hệ số mô hình:

```
                                                  s0
(Intercept)                                -9.710766e+03
X3.distance.to.the.nearest.MRT.station     -5.225673e-03
X2.house.age                               -2.785913e-01
X1.transaction.date                         4.842947e+00
X4.number.of.convenience.stores             1.184405e+00
X5.Area                                     2.780170e+00
```

• Adjusted R-square = 62.79%

c. Dữ liệu đã xử lý và chưa phân cụm.

– **Mô hình hồi quy đa biến sử dụng hàm lm():**

• Các biến tham gia vào mô hình:

```
relation <- lm(formula = y ~ x1+x2+x4+x5+I(x2^2)+I(x4^2)
                            +I(x3*x6)+I(x3*x4*x5)+I(x3*x6*x4)
                            +I(log(x3))+I(log(x6)))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.3635  -4.0274  -0.1225   3.3582  22.0243

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6.356e+04  2.581e+04  -2.463 0.014384 *
x1                5.334e+00  1.310e+00   4.071 6.07e-05 ***
x2               -8.221e-01  1.191e-01  -6.901 3.36e-11 ***
x4                6.137e+00  1.800e+00   3.410 0.000744 ***
x5                4.813e+02  7.109e+01   6.770 7.37e-11 ***
I(x2^2)           1.382e-02  2.866e-03   4.822 2.32e-06 ***
I(x4^2)          -1.538e-01  5.711e-02  -2.693 0.007500 **
I(x3 * x6)       -1.284e-01  2.695e-02  -4.766 3.00e-06 ***
I(x3 * x4 * x5) -4.725e+00  1.948e+00  -2.425 0.015929 *
I(x3 * x6 * x4)  9.666e-01  4.004e-01   2.414 0.016405 *
I(log(x3))        1.169e+02  3.113e+01   3.755 0.000210 ***
I(log(x6))        8.485e+03  5.241e+03   1.619 0.106574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.98 on 284 degrees of freedom
Multiple R-squared:  0.7436,    Adjusted R-squared:  0.7337
F-statistic: 74.89 on 11 and 284 DF,  p-value: < 2.2e-16
```

• Adjusted R-square = 73,37%

• Tham số mô hình:

```
Coefficients:
    (Intercept)              x1              x2              x4
     -6.356e+04       5.334e+00      -8.221e-01       6.137e+00

             x5         I(x2^2)         I(x4^2)      I(x3 * x6)
      4.813e+02       1.382e-02      -1.538e-01      -1.284e-01

I(x3 * x4 * x5)  I(x3 * x6 * x4)       I(log(x3))       I(log(x6))
     -4.725e+00       9.666e-01        1.169e+02        8.485e+03
```

– **Mô hình Ridge Regression:**

• Giá trị Lambda tối ưu nhất: 0.2511886

• Hệ số mô hình:

```
(Intercept)                                -3.648258e+04
X3.distance.to.the.nearest.MRT.station     -4.323389e+00
X2.house.age                               -2.439160e-01
X6.longitude                                1.322139e+02
X1.transaction.date                         5.793242e+00
X4.number.of.convenience.stores             3.734196e-01
X5.latitude                                 3.536853e+02
```

• Adjusted R-square = 69.48%

d. Dữ liệu đã xử lý và đã phân cụm.

– **Mô hình hồi quy đa biến sử dụng hàm lm():**

• Các biến tham gia vào mô hình:

relation <- lm(formula = y ~ x1+x5 +I(x3^2)+I(x5^2) + I(x2^3) +I(x4^3)
+I(x5^3)+I(x4*x5)+sqrt(x2)+sqrt(x5)+I(x5*x4*sqrt(x3)))

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.3647  -3.9639   0.2365   3.5476  22.8663

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.090e+04  2.642e+03  -4.125 4.88e-05 ***
x1                       5.440e+00  1.313e+00   4.144 4.51e-05 ***
x5                      -1.353e+02  2.135e+01  -6.340 9.03e-10 ***
I(x3^2)                 -1.361e-01  3.495e-02  -3.893 0.000123 ***
I(x5^2)                  3.626e+01  6.426e+00   5.642 4.07e-08 ***
I(x4 * x5)               1.888e+00  7.049e-01   2.678 0.007837 **
sqrt(x2)                -2.844e+00  3.766e-01  -7.551 5.91e-13 ***
sqrt(x5)                 1.168e+02  1.640e+01   7.123 8.69e-12 ***
I(x2^3)                  7.563e-05  3.240e-05   2.334 0.020283 *
I(x4^3)                 -3.945e-03  2.076e-03  -1.900 0.058439 .
I(x5^3)                 -4.314e+00  7.869e-01  -5.483 9.24e-08 ***
I(x5 * x4 * sqrt(x3)) -5.131e-01  2.402e-01  -2.137 0.033485 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.971 on 284 degrees of freedom
Multiple R-squared:  0.7444,    Adjusted R-squared:  0.7345
F-statistic:  75.2 on 11 and 284 DF,  p-value: < 2.2e-16
```

• Adjusted R-square : 73,45%

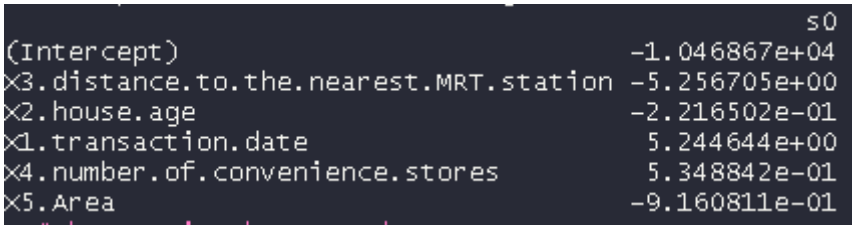
• Tham số mô hình:

```
Coefficients:
         (Intercept)                    x1                    x5              I(x3^2)              I(x5^2)
           -1.090e+04             5.440e+00             -1.353e+02            -1.361e-01            3.626e+01
            I(x4 * x5)               sqrt(x2)              sqrt(x5)              I(x2^3)              I(x4^3)
             1.888e+00             -2.844e+00             1.168e+02             7.563e-05            -3.945e-03
               I(x5^3)  I(x5 * x4 * sqrt(x3))
            -4.314e+00             -5.131e-01
```

– **Mô hình Ridge Regression:**

• Giá trị Lambda tối ưu nhất: 0.07943282

• Hệ số mô hình:



```
                                                  s0
(Intercept)                                -1.046867e+04
X3.distance.to.the.nearest.MRT.station     -5.256705e+00
X2.house.age                               -2.216502e-01
X1.transaction.date                         5.244644e+00
X4.number.of.convenience.stores             5.348842e-01
X5.Area                                    -9.160811e-01
```

• Adjusted R-square = 57.99%

Bảng 3.4.6b: Tổng kết Adjusted R2 của các mô hình trên dữ liệu train.

	Chưa xử lí- Chưa phân cụm	**Chưa xử lí – Đã phân cụm**	**Xử lí – Chưa phân cụm**	**Xử lí – Đã phân cụm**
LM	68,99%	68,48%	73,37%	**73,45%**
Ridge	56,42%	62,79%	69,48%	57,99%

❖ **Nhận xét:**

Trên dữ liệu huấn luyện:

– Giữa 2 thuật toán thì hồi quy đa thức cho kết quả cao ở tất cả các bộ dữ liệu so với thuật toán hồi quy Ridge.

– Mô hình huấn luyện trên dữ liệu đã xử lý cho kết quả tốt hơn chưa xử lý (Hồi quy đa thức)

– Mô hình trên dữ liệu đã phân cụm và chưa phân cụm cho kết quả gần bằng nhau (Hồi quy đa thức)

– Mô hình huấn luyện trên dữ liệu đã xử lý và đã phân cụm cho kết quả tốt nhất với adjusted R2 là 73,45%.

– Mô hình cho kết quả adjusted R2 thấp nhất là mô hình ở dữ liệu chưa xử lý và chưa phân cụm, áp dụng thuật toán hồi quy Ridge (56,42%).

– Thuật toán hồi quy đa biến cho ra mô hình có sự phân loại rõ ràng hơn Ridge.

– Nhìn chung, nếu chỉ dựa vào kết quả thu được ở tập huấn luyện thì hồi quy đa thức phù hợp với dữ liệu hơn là Ridge.

• Kết quả áp dụng mô hình lên dữ liệu Test:

Bảng 3.4.6c: Tổng kết Adjusted R2 của các mô hình trên dữ liệu test.

	Chưa xử lí- Chưa phân cụm	**Chưa xử lí – Đã phân cụm**	**Xử lí – Chưa phân cụm**	**Xử lí – Đã phân cụm**
LM	76,54%	**77,7%**	64,61%	66,11%
Rigde	65,50%	72,44%	55,40%	44,24%

❖ **Nhận xét:**

Trên dữ liệu Test:

– Bộ dữ liệu phù hợp với thuật toán Hồi quy đa thức hơn là Hồi quy Ridge.

– Mô hình trên dữ liệu chưa xử lý cho kết quả cao hơn đã xử lý.

– Mô hình trên dữ liệu chưa xử lý và đã phân cụm, áp dụng hồi quy đa thức cho kết quả dự đoán tốt nhất trên tập test với adjusted R2 là 77,7%.

– Mô hình cho kết quả adjusted R2 thấp nhất là mô hình ở dữ liệu đã xử lí- đã phân cụm Rigde.

4. Kết luận

4.1. Tổng kết mô hình

Mô hình cho kết quả tốt nhất hiện tại là mô hình thu được trên dữ liệu chưa xử lý và đã phân cụm, tức là chưa loại bỏ các outlier, chưa chuẩn hóa và đã áp dụng thực hiện thay thế các biến kinh độ và vĩ độ (X5 latitude, X6 longitude) thành khu vực địa lý (X5 Area).

Mô hình cuối cùng có dạng như sau:

$$Y = (-1{,}267e+04)+(6{,}312e+00)x_1+(-3{,}402e-02)x_2^2$$
$$+(6{,}805e-04)x_2^3+(7{,}269e-03)x_1x_5+(-3{,}874e+00)x_4x_5$$
$$+(9{,}135e+00)\sqrt{x_4}+(-1{,}591e-01)\sqrt{x_2x_3}+(-1{,}881e-01)\sqrt{x_3x_4}$$
$$+(-8{,}659e-02)x_1\sqrt{x_5}+(2{,}536e-02)x_2\sqrt{x_3}+(5{,}994e+00)x_4\sqrt{x_5}$$

4.2. Kết luận chung

– Việc phân bố của dữ liệu cũng ảnh hưởng lớn đến kết quả của mô hình.

– Không phải lúc nào việc loại bỏ outlier cũng mang lại kết quả tốt. Cần chú ý trong quá trình xử lí outlier, có thể tách riêng ra nhưng không được tùy tiện loại bỏ, bởi có những outlier phản ánh xu hướng, đặc trưng của dữ liệu.

– Có những yếu tố khi đứng riêng lẻ thì không ảnh hưởng đến kết quả nhưng khi kết hợp với yếu tố khác có thể ảnh hưởng đến kết quả.

– Giữa các mô hình, không phải mô hình nào cho kết quả tốt trên train đều sẽ cho kết quả tốt trên test, việc đánh giá mô hình phải thông qua thực nghiệm, mô hình chỉ tốt khi nó tốt trên thực nghiệm.

– Việc tìm kiếm các mối liên hệ giữa các yếu tố vô cùng quan trọng, không thể vì một yếu tố nào đó đứng một mình không ảnh hưởng đến kết quả mà loại bỏ nó.

5. Tài liệu tham khảo

1. https://bigdatauni.com/vi/tin-tuc/tong-quan-ve-regression-phan-tich-hoi-quy.html
2. https://rstatisticsblog.com/data-science-in-action/machine-learning/ridge-regression-in-r/?fbclid=IwAR3ANrfqFyYSE5Y2UCnNlRxDiufJwfLz0jw8em3H8x4OhHmwWHfyyj-ijRM
3. https://machinelearningcoban.com/2017/01/01/kmeans/