Phần chung

**Bài 1:** Tập tin "gia\_nha.csv" chứa thông tin về giá bán ra thị trường (đơn vị đô

la) của 21613 ngôi nhà ở quận King nước Mỹ trong khoảng thời gian từ tháng

5/2014 đến 5/2015. Bên cạnh giá nhà, dữ liệu còn bao gồm các thuộc tính mô tả

chất lượng ngôi nhà. Dữ liệu gốc được cung cấp tại:

https://www.kaggle.com/harlfoxem/housesalesprediction.

• price: Giá nhà được bán ra.

• sqft\_living15: Diện tích trung bình của 15 ngôi nhà gần nhất trong khu dân cư.

• floors: Số tầng của ngôi nhà được phân loại từ 1 − 3.5.

• condition: Điều kiện kiến trúc của ngôi nhà từ 1 − 5, 1: rất tệ và 5: rất tốt.

• sqft\_above: Diện tích ngôi nhà.

• sqft\_living: Diện tích khuôn viên nhà.

**Câu hỏi:**

1. **Đọc dữ liệu (Import data)**

Input:

Logo

Description automatically generated with medium confidence

Giải thích: Đọc tập tin gianha.csv và lưu dữ liệu với tên gia\_nha

Output:

Table

Description automatically generated

1. Làm sạch dữ liệu (Data cleaning)
2. Hãy trích ra một dữ liệu con đặt tên là new\_DF chỉ bao gồm các biến chính mà taquan tâm như đã trình bày trong phần giới thiệu dữ liệu. Từ câu hỏi này về sau, mọi yêu cầu xử lý đều dựa trên tập dữ liệu con new\_DF này.

Input:

Chart, bubble chart

Description automatically generated

Giải thích: Trích tất cả tập tin và đánh số thứ tự có trong gia\_nha

Output:

Graphical user interface, text

Description automatically generated

Input:

Graphical user interface

Description automatically generated with low confidence

Giải thích: Lập bảng dữ liệu con với tên được đặt là new\_DF bao gồm các biến chọn theo số thứ tự được đánh dấu như trên

Output

Table

Description automatically generated

1. Kiểm tra các dữ liệu bị khuyết trong tập tin. (Các câu lênh tham khảo: is.na(),which(), apply()). Nếu có dữ liệu bị khuyết, hãy đề xuất phương pháp thay thế cho những dữ liệu bị khuyết này.

Input:

A picture containing logo

Description automatically generated

OutputText

Description automatically generated with medium confidence

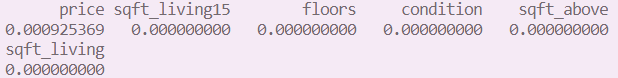
Input:

A picture containing logo

Description automatically generated

Giải thích: Tạo ra bảng thống kê số lượng giá trị khuyết trong New\_DF

Output:



Input:

Graphical user interface, text

Description automatically generated with medium confidence

Giải thích: Tính tỉ lệ giá trị khuyết tật có trong dữ liệu New\_DF

Output:

Nhận xét: Dựa trên giá trị tính được như trên và bảng thống kê số lượng giá trị khuyết ở trên ta thấy rằng các giá trị khuyết tại biến price. Do đó ta chỉ cần tập trung xử lí các giá trị khuyết đó bằng cách thay thế giá trị trung bình của cột vào giá trị bị khuyết

Input:

A picture containing graphical user interface

Description automatically generated

Output:

Table

Description automatically generated

Table

Description automatically generated

1. Làm rõ dữ liệu
2. Chuyển đổi các biến price, sqft\_living15, sqft\_above, sqft\_living lần lượt thành log(price), log(sqft\_living15), log(sqft\_above), và log(sqft\_living). Từ đây mọi sự tính toán với các biến trên được hiểu là đã qua đổi biến dạng log.

Input:

A picture containing graphical user interface

Description automatically generated

Giải thích: Chuyển đổi các biến price thành log(price), sqft\_living15 thành log(sqft\_living15), sqft\_above thành log(sqft\_above) và biến sqft\_living thành log(sqft\_living).

Table

Description automatically generatedTable

Description automatically generatedOutput:

1. Đối với các biến liên tục, hãy tính các giá trị thống kê mô tả bao gồm: trung bình,trung vị, độ lệch chuẩn, giá trị lớn nhất và giá trị nhỏ nhất. Xuất kết quả dưới dạngbảng. (Hàm gợi ý: mean(), median(), sd(), min(), max() , apply(), as.data.frame(), rownames()).

Input:

A picture containing logo

Description automatically generated

Giải thích: tính trung bình các biến price, sqft\_living15, sqft\_above, sqft\_living và lưu vào biến trungbinh

Input:

A picture containing text

Description automatically generated

Giải thích: tính trung vị các biến price, sqft\_living15, sqft\_above, sqft\_living và lưu vào biến trungvi

Input:

A picture containing logo

Description automatically generated

Giải thích: tính độ lệch chuẩn các biến price, sqft\_living15, sqft\_above, sqft\_living và lưu vào biến dolechchuan

Input:

A picture containing graphical user interface

Description automatically generated

Giải thích: tính giá trị nhỏ nhất các biến price, sqft\_living15, sqft\_above, sqft\_living và lưu vào biến giatrinhonhat

Input:

A picture containing graphical user interface

Description automatically generated

Giải thích: tính giá trị lớn nhất các biến price, sqft\_living15, sqft\_above, sqft\_living và lưu vào biến giatrilonnhat

Input:

A picture containing text

Description automatically generated

Giải thích: Lập bảng thống kê với các biến trungbinh, trungvi, dolechchuan, giatrilonnhat, giatrinhonhat

Output:

Table

Description automatically generated

1. Đối với các biến phân loại, hãy lập một bảng thống kê số lượng cho từng chủng loại (Hàm gợi ý: table()).

Input:

Text

Description automatically generated

Giải thích: Lập bảng thống kê số lượng cho biến floors

Output:

Table

Description automatically generated

Input:

Graphical user interface, text

Description automatically generated

Giải thích: Lập bảng thống kê số lượng cho từng biến condition

Output:

Table

Description automatically generated

1. Hãy dùng hàm hist() để vẽ đồ thị phân phối của biến price.

Input:

Timeline

Description automatically generated

Giải thích: Vẽ biểu đồ histogram biến price

Output:

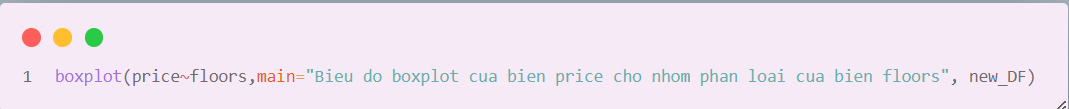
Chart, histogram

Description automatically generated

Nhận xét: Đồ thị phân phối của biến price có hình dạng phân phối chuẩn. Trong đó, già nhà tập trung phần lớn ở mức giá từ 12$ - 14$, phân bố tần suất cao nhất vào khoảng 12.5$ - 13$, phân bố tần suất thấp nhất vào khoảng 15.5% - 16$.

1. Hãy dùng hàm boxplot() vẽ phân phối của biến price cho từng nhóm phân loại của biến floors và biến condition.

Input:



Giải thích: Vẽ biểu đồ boxplot của biến price cho nhóm phân loại của biến floors.

Chart, box and whisker chart

Description automatically generatedOutput:

Nhận xét:

* **Với nhóm nhà 1 tầng**

Giá nhà bán ra thị trường cao nhất khoảng 15.5$

Giá nhà bán ra thị trường thấp nhất 11.2$

Có khoảng 25% nhà bán ra thị trường với giá dưới 12.5$

Có khoảng 50% nhà bán ra thị trường với giá dưới 12.9$

Có khoảng 75% nhà bán ra thị trường với giá dưới 13.25$

Phạm vi liên phân tử (IQR) khoảng 0.75$

* **Với nhóm nhà 1.5 tầng**

Giá nhà bán ra thị trường cao nhất khoảng 15.25$

Giá nhà bán ra thị trường thấp nhất khoảng 11.5$

Có khoảng 25% nhà bán ra thị trường với giá dưới 12.85$

Có khoảng 50% nhà bán ra thị trường với giá dưới 13.25$

Có khoảng 75% nhà bán ra thị trường với giá dưới 13.4$

Phạm vi liên phân tử (IQR) khoảng 0.55$

* **Với nhóm nhà 2 tầng**

Giá nhà bán ra thị trường cao nhất khoảng 15.75$

Giá nhà bán ra thị trường thấp nhất khoảng 11.5$

Có khoảng 25% nhà bán ra thị trường với giá dưới 12.9$

Có khoảng 50% nhà bán ra thị trường với giá dưới 13.3$

Có khoảng 75% nhà bán ra thị trường với giá dưới 13.5$

Phạm vi liên phần tử (IQR) khoảng 0.6$

* **Với nhóm nhà 2.5 tầng**

Giá nhà bán ra thị trường cao nhất khoảng 15.85$

Giá nhà bán ra thị trường thấp nhất khoảng 12.5$

Có khoảng 25% nhà bán ra thị trường với giá dưới 13.25$

Có khoảng 50% nhà bán ra thị trường với giá dưới 13.5$

Có khoảng 75% nhà bán ra thị trường với giá dưới 14.15$

Phạm vi liên phần tử (IQR) khoảng 0.9$

* **Với nhóm nhà 3 tầng**

Giá nhà bán ra thị trường cao nhất khoảng 14.8$

Giá nhà bán ra thị trường thấp nhất khoảng 12.4$

Có khoảng 25% nhà bán ra thị trường với giá dưới 12.9$

Có khoảng 50% nhà bán ra thị trường với giá dưới 13.1$

Có khoảng 75% nhà bán ra thị trường với giá dưới 13.3$

Phạm vi liên phần tử (IQR) khoảng 0.4$

* **Với nhóm nhà 3.5 tầng**

Giá nhà bán ra thị trường cao nhất khoảng 14.85$

Giá nhà bán ra thị trường thấp nhất khoảng 12.9$

Có khoảng 25% nhà bán ra thị trường với giá dưới 13.1$

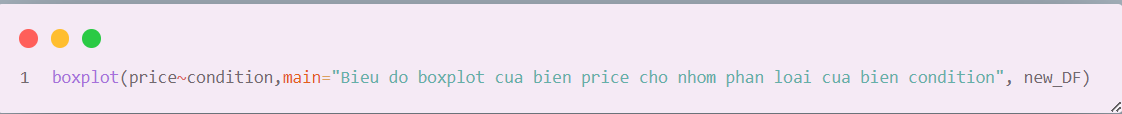
Có khoảng 50% nhà bán ra thị trường với giá dưới 13.2$

Có khoảng 75% nhà bán ra thị trường với giá dưới 13.5$

Phạm vi liên phần tử (IQR) khoảng 0.4$

Nhận xét: Nhóm nhà có số tầng từ 2 đến 2.5 có giá bán ra cao nhất và nhóm nhà có 1 tầng có giá bán ra thấp nhất. Cho thấy rằng nhóm nhà từ 2 đến 2.5 tầng được người dân sống tại quận King ưa chuộng và cho rằng đây là nhóm nhà phù hợp nhất. Trài với đó là nhóm nhà 1 tầng ít được chuộng nhất, và nhóm nhà 3.5 tầng được bán với giá đồng đều nhất so với các nhóm khác.

Input:



Chart, box and whisker chart

Description automatically generatedGiải thích: Vẽ biểu đồ boxplot của biến price cho nhóm phân loại của biến condition.

Nhận xét:

* **Với nhóm nhà có điều kiện kiến trúc rất tệ (1)**

Giá nhà bán ra thị trường cao nhất khoảng 14.25$

Giá nhà bán ra thị trường thấp nhất khoảng 11.25$

Có khoảng 25% nhà bán ra thị trường với giá dưới 11.9$

Có khoảng 50% nhà bán ra thị trường với giá dưới 12.5$

Có khoảng 75% nhà bán ra thị trường với giá dưới 13$

Phạm vi liên phần tử (IQR) khoảng 1.1

* **Với nhóm nhà có điều kiện kiến trúc tệ(2):**

Giá nhà bán ra thị trường cao nhất khoảng 14.8$

Giá nhà bán ra thị trường thấp nhất khoảng 11.25$

Có khoảng 25% nhà bán ra thị trường với giá dưới 12.25$

Có khoảng 50% nhà bán ra thị trường với giá dưới 12.6$

Có khoảng 75% nhà bán ra thị trường với giá dưới 12.9$

Phạm vi liên phần tử (IQR) khoảng 0.65$

* **Với nhóm nhà có điều kiện kiến trúc trung bình(3):**

Giá nhà bán ra thị trường cao nhất khoảng 15.8$

Giá nhà bán ra thị trường thấp nhất khoảng 11.15$

Có khoảng 25% nhà bán ra thị trường với giá dưới 12.7$

Có khoảng 50% nhà bán ra thị trường với giá dưới 13$

Có khoảng 75% nhà bán ra thị trường với giá dưới 13.3$

Phạm vi liên phần tử (IQR) khoảng 0.6$

* **Với nhóm nhà có điều kiện kiến** t**rúc tốt**(**4):**

Giá nhà bán ra thị trường cao nhất khoảng 15.9$

Giá nhà bán ra thị trường thấp nhất khoảng 11.4$

Có khoảng 25% nhà bán ra thị trường với giá dưới 12.6$

Có khoảng 50% nhà bán ra thị trường với giá dưới 13$

Có khoảng 75% nhà bán ra thị trường với giá dưới 13.25$

Phạm vi liên phần tử (IQR) khoảng 0.65$

* **Với nhóm nhà có điều kiện kiến trúc rất** **tốt(5):**

Giá nhà bán ra thị trường cao nhất khoảng 15.2$

Giá nhà bán ra thị trường thấp nhất khoảng 11.6$

Có khoảng 25% nhà bán ra thị trường với giá dưới 12.8$

Có khoảng 50% nhà bán ra thị trường với giá dưới 13.1$

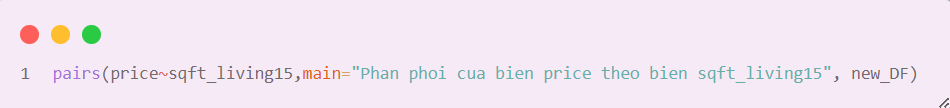
Có khoảng 75% nhà bán ra thị trường với giá dưới 13.4$

Phạm vi liên phần tử (IQR) khoảng 0.6$

Nhận xét: Ta thấy giá nhà tỉ lệ thuận với điều kiện kiến trúc ngôi nhà, vì nhu cầu của người dân ở các ngôi nhà có điều kiện kiến trúc cao là khá cao, rất phù hợp với thực tiễn đời sống.

1. Dùng lệnh pairs() vẽ các phân phối của biến price lần lượt theo các biến sqft\_living15 , sqft\_above, và sqft\_living.

Input:



Giải thích: Vẽ phân phối biến price theo biến sqft\_living15 và đặt tên cho biểu đồ là “Phan phoi cua bien price theo sqft\_living15”

Chart, scatter chart

Description automatically generated Output:

Input:

A picture containing graphical user interface

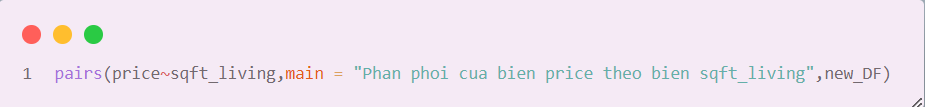
Description automatically generated

Giải thích: Vẽ phân phối biến price theo biến sqft\_living15 và đặt tên cho biểu đồ là “Phan phoi cua bien price theo sqft\_above”

Chart, scatter chart

Description automatically generatedOutput:

Input:



Giải thích: Vẽ phân phối biến price theo biến sqft\_living15 và đặt tên cho biểu đồ là “Phan phoi cua bien price theo sqft\_living”

Chart

Description automatically generated Output:

Nhận xét:

Biến price và sqft\_living có một mối quan hệ tuyến tính dương rõ ràng, tức là khi diện tích nhà càng lớn thì giá nhà càng cao. Điều đó cho thấy rõ ràng về nhu cầu và cung cấp của thị trường bất động sản.

Biến price và sqft\_living15 cũng có mối quan hệ tuyến tính dương nhưng không rõ ràng như biến sqft\_living. Điều đó nghĩa là ngoài diện tích nhà, một số yếu tố khác như thiết kế, nội thất,… cũng ảnh hưởng đến giá nhà.

Biến price và sqft\_above như sqft\_living có mối quan hệ tuyến tính dương khá rõ ràng.Điều đó cho thấy biến diện tích tầng trên càng lớn thì giá nhà cũng càng cao, và cho thấy rằng diện tích tầng trên tương đương tổng diện tích theo nhu cầu trên thị trường.

1. Xây dựng các mô hình hồi quy tuyến tính (Fitting linear regression models):
2. Xét mô hình hồi quy tuyến tính bao gồm biến price là một biến phụ thuộc, và tất cả các biến còn lại đều là biến độc lập. Hãy dùng lệnh lm() để thực thi mô hình hồi quy tuyến tính bội.

Input:

Graphical user interface, text

Description automatically generated

Giải thích: Xây dựng mô hình hồi quy tuyến tính m1 và lập bảng thống kê kết quả tính toán mô hình này.

Output:

Table

Description automatically generated

Nhận xét: Từ kết quả phân tích ta thu được =5.451345, = 0.429764, = 0.13687,

= 0.085407, = – 0.178314, = 0.685977.   
Như vậy ta thu được đường thẳng hồi quy ước lượng cho bởi phương trình sau:

Price =5.451345 + 0.429764 x sqft\_living15 + 0.13687 x floors + 0.085407 x

condition – 0.178314 x sqft\_above + 0.685977 x sqft\_living

Hệ số R-squared bằng 0,4984 nghĩa là 49,84% phương sai của biến price được giải thích bởi các biến floors, sqft\_living15, sqft\_above, condition, sqft\_living. Hệ số này cho thấy mô hình hồi quy tuyến tính có giá trị khớp trung bình với dữ liệu, do đó chỉ dự đoán được 1 phần giá trị biến phụ thuộc và vẫn có nhiều sai số.

1. Dựa vào kết quả của mô hình hồi quy tuyến tính trên, những biến nào bạn sẽ loại khỏi mô hình tương ứng với mức ý nghĩa 5%?

Dựa vào kết quả của mô hình trên, ta không bỏ biến nào khỏi mô hình viến Pr(>|t|) của các biến độc lập đều rất nhỏ, bé hơn nhiều so với 0.05. Do đó ta không cần loại bỏ biến độc lập nào vì biến nào cũng đều rất nhỏ và bé hơn 5% nên rất có ý nghĩa thống kê.

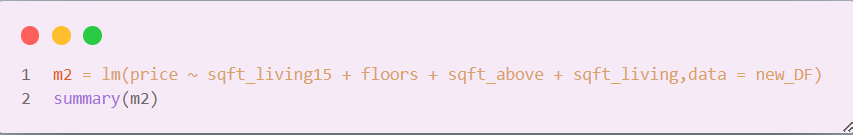
1. Xét 2 mô hình tuyến tính cùng bao gồm biến price là biến phụ thuộc nhưng:

+ Mô hình M1 chứa tất cả các biến còn lại là biến độc lập

+Mô hình M2 là loại bỏ biến condition từ mô hình M1.

Hãy dùng lệnhh anova() để đề xuất mô hình hồi quy hợp lý hơn.

Input:



Giải thích: ta lập mô hình tương tự mô hình nhưng bỏ biến độc lập condition

Table

Description automatically generated Output:

Nhận xét:

Từ kết quả phân tích ta thu được =5.81042, =0.414445, =0.112811,

= –0.203136, = 0.721355.  
Như vậy ta thu được đường thẳng hồi quy ước lượng cho bởi phương trình sau:

Price =5.81042 + 0.414445.sqft\_living15 +0.112811.floors – 0.203136.sqft\_above

+ 0.721355.sqft\_living

Hệ số R-squared bằng 0,4982 nghĩa là 49,82% phương sai của biến price được giải thích bởi các biến floors, sqft\_living15, sqft\_above, sqft\_living. Hệ số này cho thấy mô hình hồi quy tuyến tính có giá trị khớp trung bình với dữ liệu, do đó chỉ dự đoán được 1 phần giá trị biến phụ thuộc và vẫn có nhiều sai số.

Input:

Text, letter

Description automatically generated

Ta đặt giả thiết

H0: hai mô hình có hiệu quả như nhau

H1: hai mô hình có hiệu quả khác nhau

Nhận xét: Với kết quả thu được trong bảng Annova cho 2 mô hình M1 và M2 như trên ta thu được P value ( Pr(>F) ) =2.2e-16 rất là nhỏ so với mức ý nghĩa 5%, do đó xác suất quan sát được kết quả như vậy với H0 là rất khó. Do đó ta có thể bác bỏ H0 và chấp nhận H1. Do đó, ta cần cân nhắc chọn mô hình M1 hay mô hình M2. Tuy nhiên, ở mô hình M1 ta đã có kết luận rằng không nên bỏ các biến độc lập nào khỏi mô hình vì các biến có ý nghĩa thống kê, mà mô hình M2 lại bỏ đi biến condition so với mô hình M1. Do đó, ta có cơ sở để chọn mô hình M1. Thật vậy ta xét thấy R2 (hệ số xác định) ở mô hình M1 = 0,4984 > M2 = (0,4882) điều đó cho thấy các biến độc lập trong mô hình M1 giải thích biến phụ thuộc nhiều hơn so với mô hình M2. Do đó ta chọn mô hình M1 sẽ hợp lý hơn.

1. Chọn mô hình hợp lý hơn từ câu (c) hãy suy luận sự tác động của các biến lên giá nhà.

Ta nhận thấy biến price có liên quan tuyến tính đến các các biến độc lập còn lai. Ngoài ra như đã nói p value của các biến độc lập rất nhỏ và chúng nó rất có ý nghĩa thống kê nên những sự thay đổi của các biến độc lập này có ảnh hưởng nhiều đến giá nhà. Cụ thể khi các biến độc lập tăng lên 1 đơn vị thì giá nhà sẽ tăng 1 khoảng hệ số ước lượng của khoảng này.

Ví dụ như: hệ số hồi quy ứng với sqft\_living15 = 0.430556 thì ứng với sqft\_living15

tăng 1 đơn vị thì ta có thể kỳ vọng giá nhà có thể tăng 0,4305 đơn vị (giả sử rằng các biến dự báo còn lại không đổi). Tương tự cũng như hệ số hồi quy ứng với floors = 0.137069 thì ứng với floors tăng 1 đơn vị thì ta có thể kỳ vọng giá nhà tăng 0.137069 (giả sử rằng các biến dự báo còn lại không đổi)

1. Từ mô hình hồi quy mà bạn chọn ở câu (c) hãy dùng lệnh plot() để vẽ đồ thị biểu thị sai số hồi quy (residuals) và giá trị dự báo (fitted values). Nêu ý nghĩa và nhận xét đồ thị.

Input:

A picture containing text

Description automatically generated

Giải thích: Vẽ đồ thị biểu thị sai số hồi quy (residuals) và giá trị dự báo (fitted values)

Chart, scatter chart

Description automatically generatedOutput:

Nhận xét:

Dựa trên đồ thị ta có thể thấy các điểm phần dư phân bố ngẫu nhiên xung quanh y=0 (ngoại trừ một số giá trị ngoại lai) chứng tỏ phương sai của các sai số là hằng số hay phương sai đồng nhất.

* Các giả định trong mô hình hồi quy tuyến tính:

X và Y có mối quan hệ tuyến tính

Sai số hồi quy theo tuân theo phân phối chuẩn

Các sai số phải độc lập với nhau

Phương sai và các sai số phải là hằng số

* Vẽ thêm biểu đồ để kiểm tra các giả định của mô hình hồi quy tuyến tính

Input:

A picture containing logo

Description automatically generated

Giải thích: vẽ biểu đồ phần tram phân vị cho mô hình M1 (biểu đồ Q-Q). Đây là đồ thị dùng để kiểm tra giả định về phân phối chuẩn của các sai số. Nếu các điểm thặng dư nằm trên cùng 1 đường thẳng thì điều kiện về phân phối chuẩn được thỏa mãn.

Chart, line chart

Description automatically generated Output:

Nhận xét: Từ đồ thị ta thấy các biến tập trung đa số xung quanh đường thẳng 45o, ở khoảng đầu và cuối có vài giá trị ngoại lai nhưng không đáng kể. Vậy giả định về các sai số tuân theo phân phối chuẩn được thỏa mãn.

Input:

A picture containing logo

Description automatically generated

Giải thích: Vẽ biểu đồ thang đo vị trí của căn bậc hai của các sai số tuyệt đối so với các giá trị khớp trong mô hình M1.

Chart, scatter chart

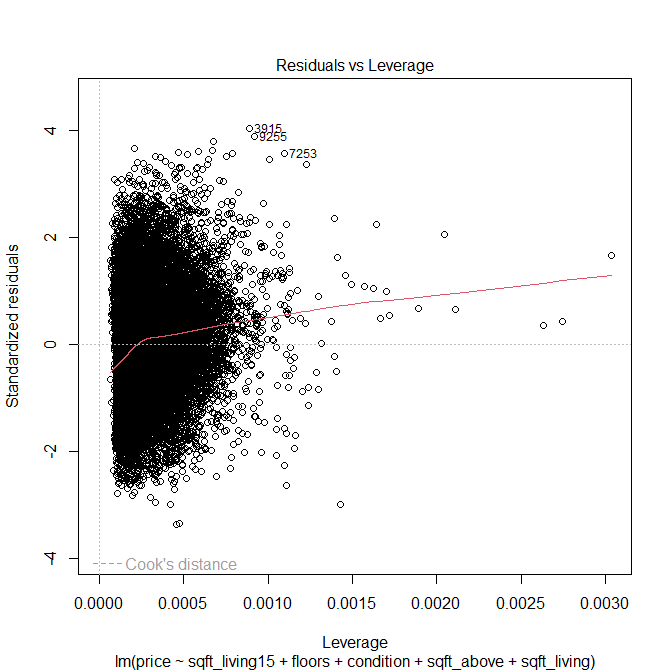
Description automatically generatedOutput:

Nhận xét: Từ biểu đồ ta thấy các điểm phân tán xung quanh đường thẳng màu đỏ nằm ngang nên giả định các phương sai các sai số là hằng số là đúng.

Input:

A picture containing logo

Description automatically generated

Giải thích: Vẽ biểu đồ Residual vs Leverage giúp ta xác định được các quan sát có ảnh hưởng đến mô hình hồi quy. Nếu các điểm nằm ngoài đường nét đứt màu đỏ thì đây là quan sát ảnh hưởng.

Output:

Nhận xét: Dựa vào đồ thị ta thấy các quan sát thứ 3915,9255,7253 có thể là các điểm

ảnh hưởng cao trong bộ dữ liệu, tuy nhiên các điểm này chưa vượt qua đường khoảng

cách cook’s distance. Vì thế các điểm này chưa thực sự ảnh hưởng cao, nên ta không cần

loại bỏ chúng khi phân tích.

1. Dự báo (Predictions)
2. Từ mô hình bạn chọn trong câu (c), hãy dùng lệnh predict() để dự báo giá nhà tại 2 thuộc tính như sau:

x1: sqft\_living15 = mean(sqft\_living15), sqft\_above = mean(sqft\_above),sqft\_living = mean(sqft\_living), floor = 2, condition = 3

x2: sqft\_living15 = max (sqft\_living15), sqft\_above = max(sqft\_above), sqft\_living = max(sqft\_living), floor = 2, condition = 3.

So sánh khoảng tin cậy cho 2 giá trị dự báo này.

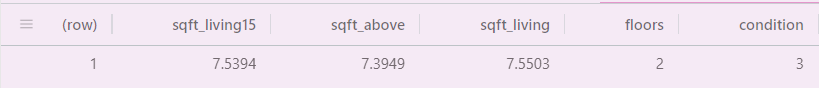
Input:

Text

Description automatically generated with medium confidence

Giải thích: Đoạn code này tạo một data frame có tên là X1 với các cột là sqft\_living15, sqft\_above, sqft\_living, floors và condition. Giá trị của các cột này được tính bằng cách lấy trung bình của các cột tương ứng trong data frame new\_DF, ngoại trừ floors và condition được gán cố định là 2 và 3.

Output:



Input:

Graphical user interface, text, application, Word

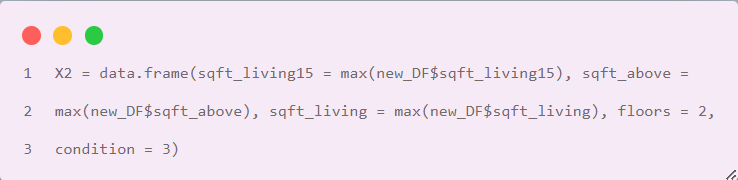
Description automatically generated

Giải thích: Dự đoán giá trị của biến phụ thuộc cho một data frame X1 dựa trên mô hình hồi quy M1

Output:

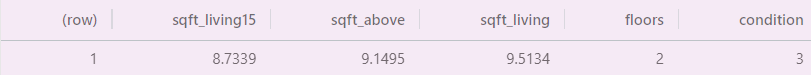
Graphical user interface

Description automatically generated with low confidence

Input:

Giải thích: Đoạn code này tạo một data frame có tên là X2 với các cột là sqft\_living15, sqft\_above, sqft\_living, floors và condition. Giá trị của các cột này được tính bằng cách lấy giá trị lớn nhất của các cột tương ứng trong data frame new\_DF, ngoại trừ floors và condition được gán cố định là 2 và 3.

Output:



Input:

Graphical user interface, text

Description automatically generated

Giải thích: Dự đoán giá trị của biến phụ thuộc cho một data frame X2 dựa trên mô hình hồi quy M1

Output:

Graphical user interface, application

Description automatically generated with medium confidence

1. So sánh khoảng tin cậy cho 2 giá trị dự báo này

Input:

A picture containing graphical user interface

Description automatically generated

Giải thích: Nối bảng predict\_X1 và predict\_x2 lại và tạo một bảng với tên pred

Output:

Table

Description automatically generated

Input:

A picture containing text

Description automatically generated

Giải thích: Đổi tên dòng thành X1 và X2

Table

Description automatically generatedTable

Description automatically generatedOutput:

Input:

Logo, company name

Description automatically generated

Giải thích: tạo thêm một cột range với giá trị bằng cột upr – cột lwr

Output:

Table

Description automatically generated