CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

1. Môi trường thực nghiệm

* Google Colab là một công cụ mạnh mẽ để phát triển, thử nghiệm, và triển khai các mô hình học máy, bao gồm cả việc huấn luyện và dự đoán với mô hình hồi quy tuyến tính.

****

Hình 3.1. Google Colab

* Ngôn ngữ lập trình: Python 3.7 và các thư viện hỗ trợ khác như Sklearn, Numpy, Pandas, Matplotlib và Seaborn.



Hình 3.2. Ngôn ngữ lập trình Python

* Lưu trữ: Google Drive, dịch vụ lưu trữ đám mây, cho phép người dùng lưu trữ và truy cập các tệp tin trực tuyến.



Hình 3.3. Google Drive

1. Dữ liệu thực nghiệm

Bộ dữ liệu sử dụng trong nghiên cứu này được thu thập từ cuộc điều tra dân số thực tế ở California năm 1990. Dữ liệu liên quan được thu thập trong một quận nhất định của California và một số số liệu thống kê tóm tắt về chúng dựa trên cuộc điều tra dân số.

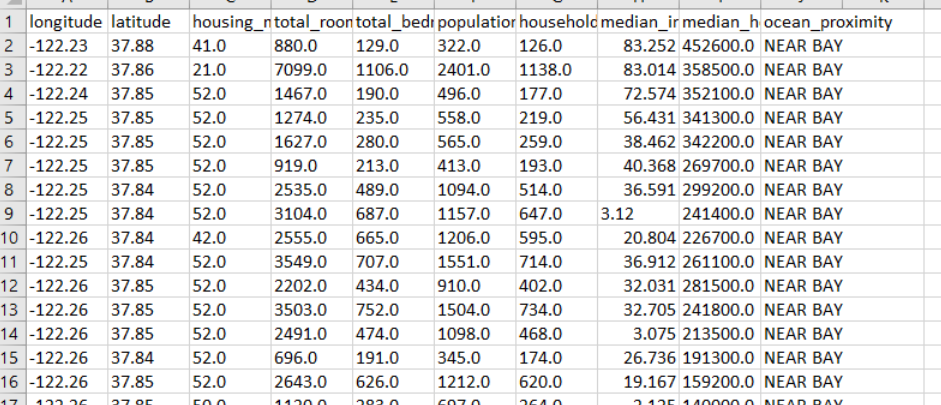
Bộ dữ liệu được phân tích ở đây là file dataset (.csv) có tên cụ thể là housing.csv, chứa 20640 bản ghi thông tin về giá nhà California.

Thông tin cụ thể như sau:

- Tên bộ dữ liệu: California Housing Prices

- Nguồn: <https://www.kaggle.com/datasets/camnugent/california-housing-prices/data>

- Dữ liệu 15 dòng đầu của dataset:

****

Hình 3.4. 15 dòng đầu của bộ dữ liệu gốc

**-** Thông tin cụ thể các cột của dataset như sau:

**longitude**: Kinh độ của vị trí địa lý.

**latitude**: Vĩ độ của vị trí địa lý.

**housing\_median\_age**: Tuổi trung bình của nhà ở trong khu vực.

**total\_rooms**: Tổng số phòng trong khu vực nhà ở.

**total\_bedrooms**: Tổng số phòng ngủ trong khu vực nhà ở.

**population**: Dân số trong khu vực nhà ở.

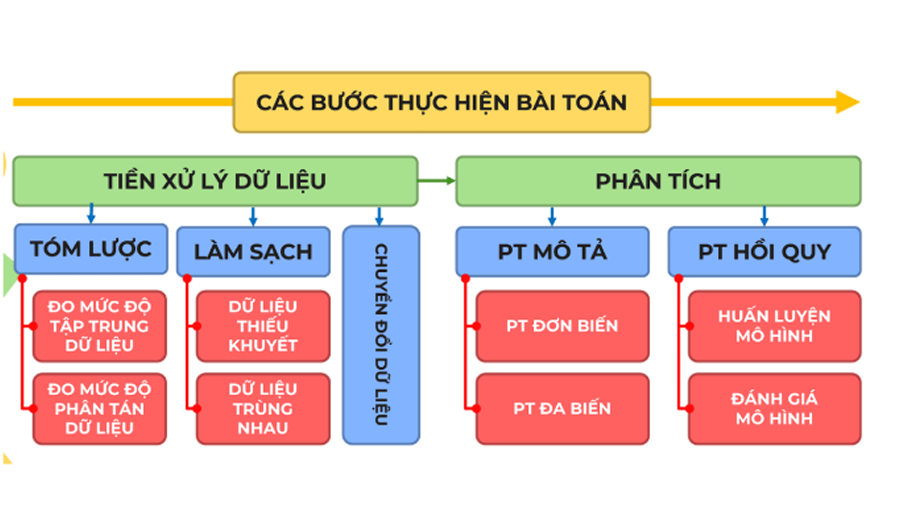
**households**: Số hộ gia đình trong khu vực nhà ở.

**median\_income**: Thu nhập trung bình của các hộ gia đình trong khu vực

**median\_house\_value**: Giá trị trung bình của các ngôi nhà trong khu vực

**ocean\_proximity**: Khoảng cách đến đại dương .

1. Quy Trình thực nghiệm

****

Hình 3.5. Quy trình thực nghiệm đề tài phân tích dữ liệu

1. Đặt mục tiêu

* Phân tích mô tả để thể hiện mối quan hệ giữa các giá trị của dữ liệu, từ đó đánh giá được tương quan của giá nhà .
* Phân tích hồi quy để dự báo giá nhà dựa theo mô hình hồi quy tuyến tính.

1. Tiền xử lý dữ liệu

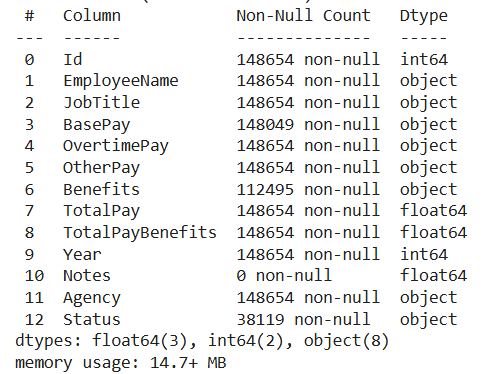
* Làm sạch dữ liệu:

Làm sạch dữ liệu là quá trình loại bỏ các sai sót, lỗi, nhiễu và thông tin không chính xác hoặc không cần thiết khỏi tập dữ liệu ban đầu để đảm bảo dữ liệu đáng tin cậy và phù hợp cho việc phân tích và xử lý tiếp theo. Quá trình làm sạch dữ liệu thường là một phần quan trọng trong tiền xử lý dữ liệu trước khi bắt đầu phân tích mô tả và cả phân tích hồi quy.

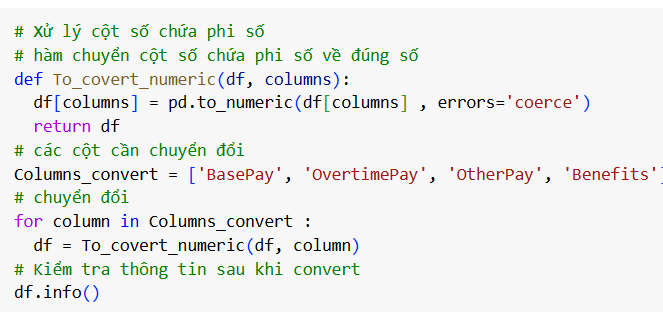
Một số tác vụ chính trong quá trình làm sạch dữ liệu bao gồm:

* Loại bỏ dữ liệu trùng lặp: Loại bỏ các bản ghi bị trùng lặp trong tập dữ liệu để tránh ảnh hưởng đến kết quả phân tích.
* Xử lý dữ liệu thiếu: Điền vào các giá trị thiếu hoặc quyết định loại bỏ chúng dựa trên ngữ cảnh và mục tiêu của phân tích.
* Xử lý giá trị ngoại lai : Xử lý giá trị ngoại lai bao gồm việc phát hiện và xử lý các điểm dữ liệu bất thường, nằm xa phân phối dữ liệu chính, bằng các phương pháp như loại bỏ, thay thế hoặc biến đổi
* Sửa lỗi và sai sót: Điều tra và sửa các lỗi cú pháp, sai sót chính tả hoặc sai sót logic trong dữ liệu.
* Chọn lọc đặc trưng: Xác định và lựa chọn các đặc trưng quan trọng nhất để sử dụng trong phân tích hoặc mô hình hóa.
* Xử lý cột số chứa giá trị phi số :

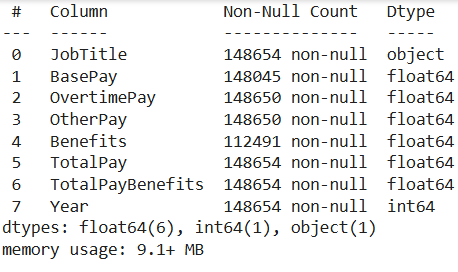
Thông tin dữ liệu trước khi xử lý :



* Xử lý cột số chứa phi số :



-Thông tin dữ liệu sau khi xử lý :

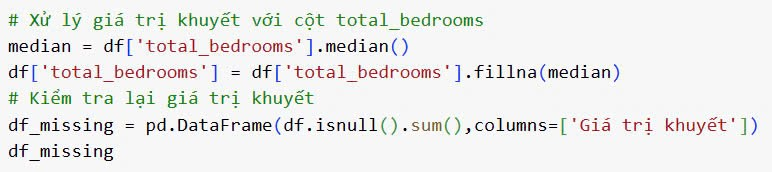


* Kiểm tra dữ liệu khuyết :



Hình 3.6. Kiểm tra dữ liệu bị khuyết

* Ta sẽ điền khuyết bằng median :

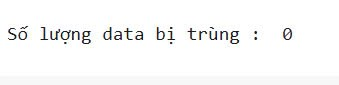
**

Hình 3.7. Xử lý dữ liệu bị khuyết

* Xử lý giá trị trùng lặp :



Hình 3.8. Xử lý dữ liệu bị trùng



Hình 3.9. Kết quả xử lý dữ liệu bị trùng

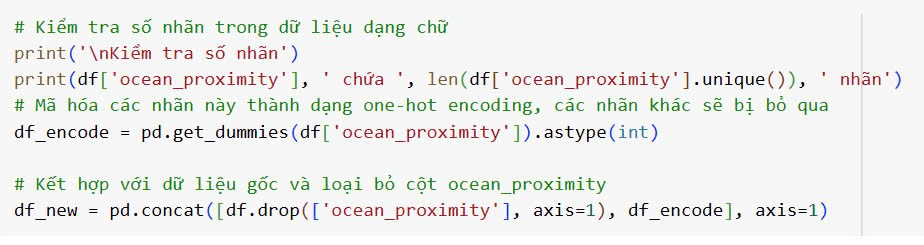
* Xử lý giá trị ngoại lai bằng phương pháp IQR



Hình 3.10. Xử lý giá trị ngoại lai

Chuyển đổi dữ liệu: Chuyển đổi dữ liệu trong phân tích dữ liệu là quá trình thay đổi cách thức biểu diễn, xử lý hoặc áp dụng các phép toán trên dữ liệu ban đầu để tạo ra dữ liệu mới có ý nghĩa hoặc thuận tiện hơn cho mục đích phân tích. Nó có vai trò quan trọng trong việc biểu diễn trực quan hơn dataset, thuận tiện hơn trong việc phân tích dữ liệu.

Trong project này, ta thấy có cột duy nhất đang ở dạng Object, tức phi số, vì vậy, ta sẽ chuyển hóa cột dữ liệu này bằng dummy.



Hình 3.11. chuyển hóa cột dữ liệu

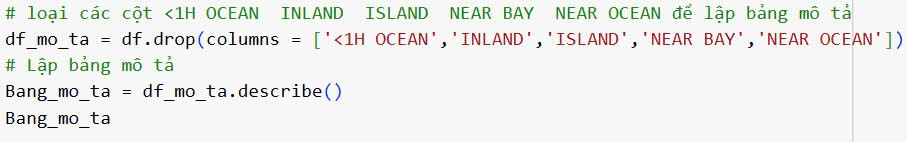
1. Phân tích mô tả

Phân tích mô tả trong phân tích dữ liệu là quá trình tóm tắt, mô tả và hiểu sâu về các đặc điểm, mẫu thái và thông tin quan trọng của tập dữ liệu. Với mục tiêu đó, ta sẽ tiến hành phân tích mô tả cho bộ dữ liệu của project theo cả 2 hướng phân tích đơn biến (trên từng biến) và phân tích đa biến (trên nhiều biến) bằng cách biểu diễn dưới các biểu đồ khác nhau.

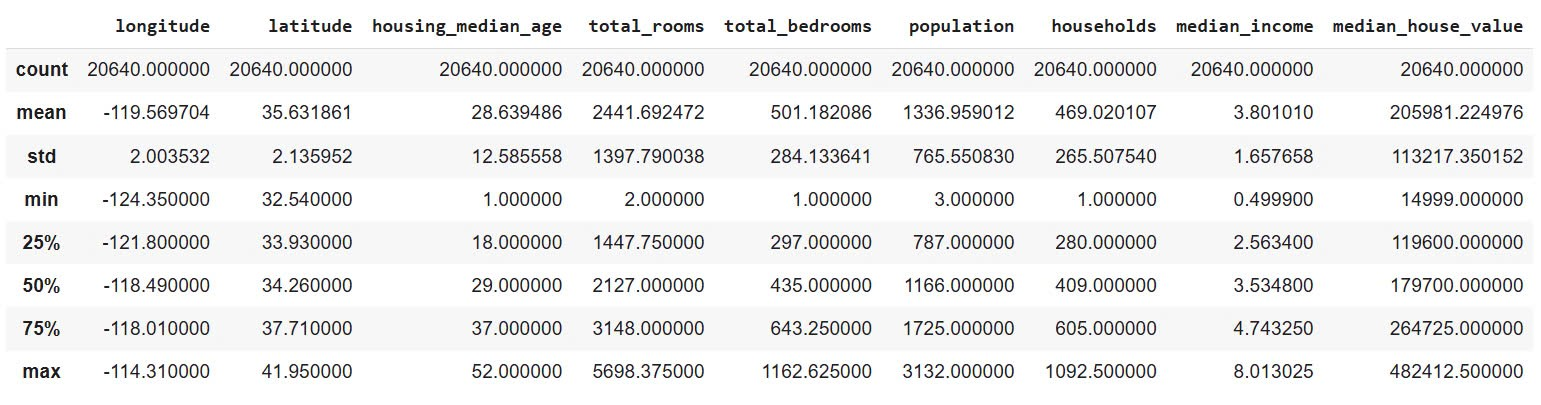
* Tóm lược dữ liệu:

+ Tóm lược dữ liệu trong phân tích dữ liệu là quá trình tổng hợp, trích xuất và trình bày các thông tin quan trọng và chính xác từ tập dữ liệu ban đầu. Mục tiêu của việc tóm lược dữ liệu là giúp người đọc hoặc người xem nắm bắt được những điểm quan trọng và khái quát của dữ liệu mà không cần phải đọc hoặc xem toàn bộ dữ liệu gốc. Tóm lược dữ liệu bao gồm 2 loại đo: Đo mức độ tập trung dữ liệu (mean, median, mode, ...) và Đo mức độ phân tán dữ liệu (quartile, interquartile, standard deviation, ...).

+ Ta sẽ tiến hành tổng hợp các thông tin về độ tập trung và phân tán của dữ liệu. Những thông số này chỉ tương thích với các cột dữ liệu dạng thông số, vậy nên sẽ chỉ có” longitude, latitude, housing\_median\_age, total\_rooms, total\_bedrooms, population, households, median\_income, median\_house\_value” là được phân tích. Dưới đây là kết quả tóm lược dữ liệu bao gồm các thuộc tính count, mean, std, min, 25%, 50%, 75%, max, mode, median của các dữ liệu trên:

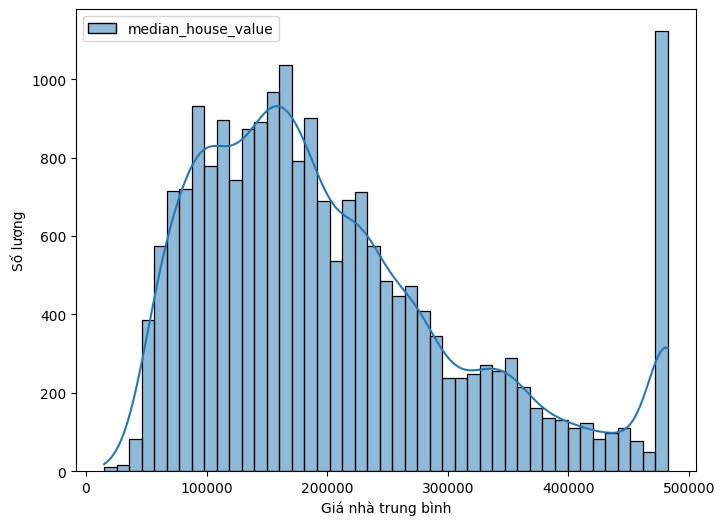


Hình 3.12. Tóm lược dữ liệu



Hình 3.13. Bảng tóm lược dữ liệu

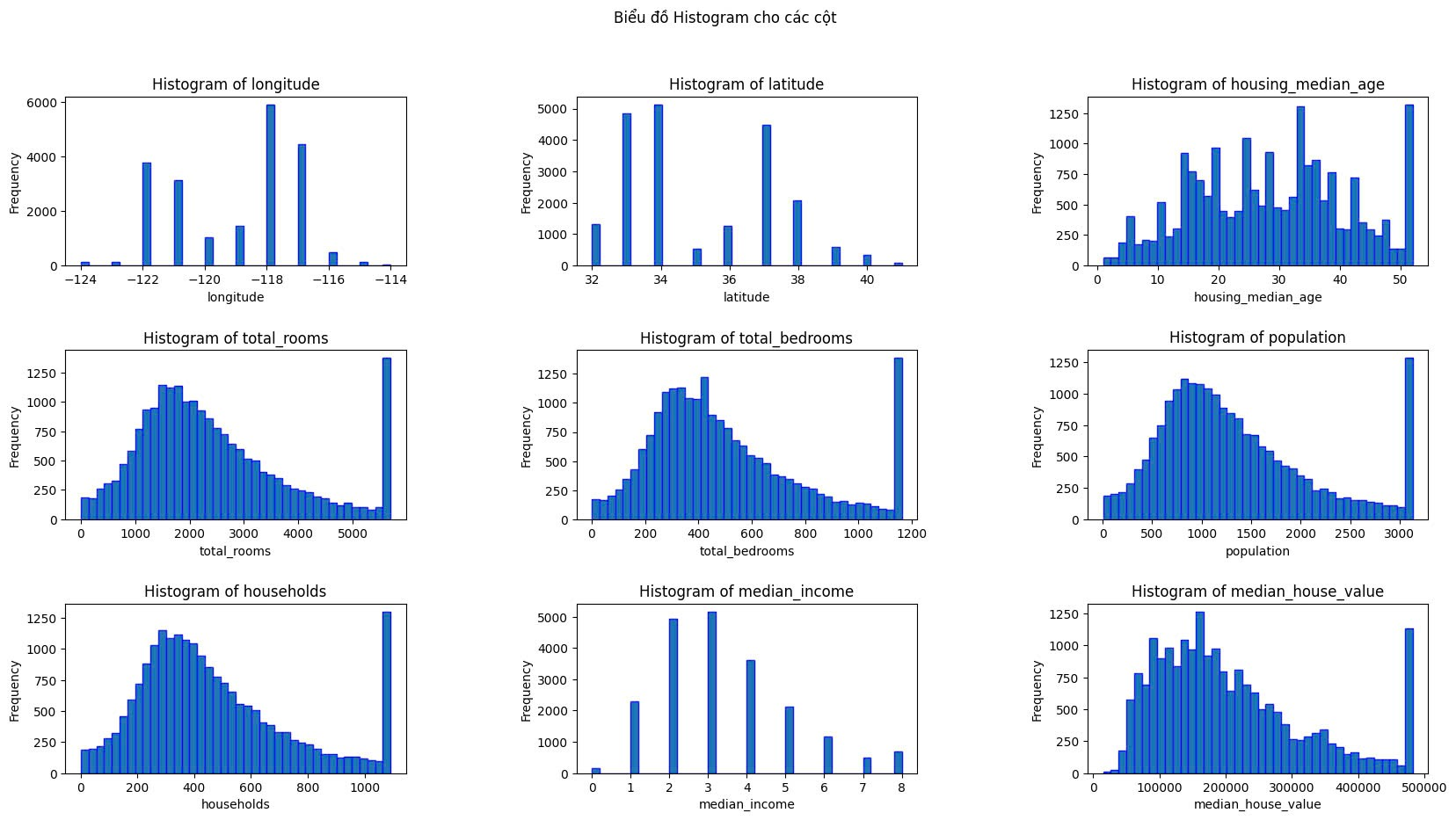
Biểu đồ Hist Plot biểu diễn giá nhà trung bình :



Hình 3.14. Biểu đồ Hist Plot

Biểu đồ hist plot thể hiện phân bố giá trị trung bình của nhà ở. Dữ liệu có xu hướng tập trung nhiều nhất ở khoảng giá trị từ 100,000 đến 200,000. Phân bố không hoàn toàn đối xứng, có vẻ hơi lệch phải, với một số lượng đáng kể các giá trị cao hơn ở phía trên 400,000 và một số ít giá trị rất thấp. Đường cong mật độ giúp trực quan hóa xu hướng này rõ hơn.

Biểu đồ Histogram cho các cột:

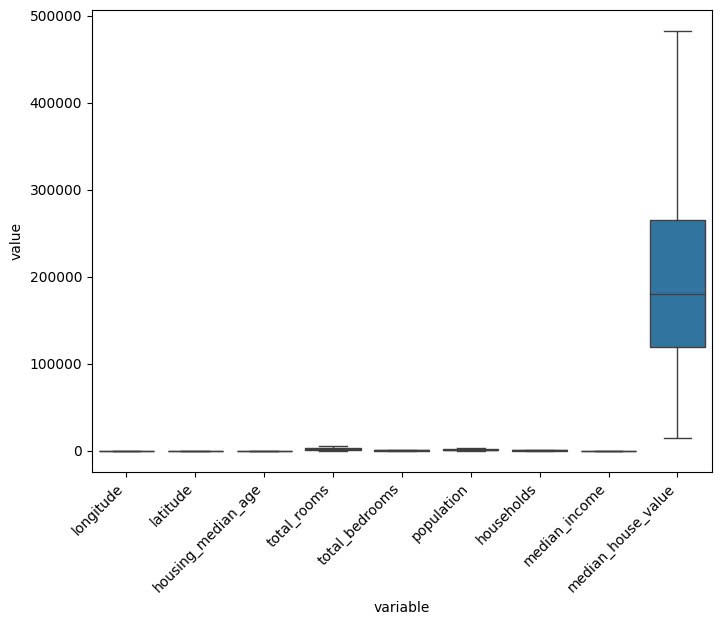


Hình 3.15. Biểu đồ Histogram cho các cột

Nhận xét:

* + Histogram of longitude**:** Phân bố không đồng đều, cho thấy sự tập trung của dữ liệu ở một vài khu vực kinh độ cụ thể, có thể là do mật độ dân số hoặc các yếu tố địa lý khác. Có một đỉnh chính ở khoảng -118, cho thấy một phần lớn dữ liệu tập trung tại khu vực kinh độ đó.
* Histogram of latitude: Tương tự như kinh độ, phân bố không đều, tập trung ở một số khu vực vĩ độ nhất định. Cho thấy một sự phân bố không đồng đều về mặt địa lý. Các đỉnh cao cho thấy khu vực có nhiều dữ liệu hơn.
* Histogram of housing\_median\_age: Phân bố gần như đối xứng, có một phần nhỏ lệch phải, tập trung nhiều nhất ở giữa (khoảng 25-35 tuổi). Điều này cho thấy nhiều nhà ở có tuổi đời trung bình.
* Histogram of total\_rooms: Phân bố lệch phải rõ rệt, nhiều nhà ở có ít phòng hơn, và có một số lượng nhỏ nhà ở có số lượng phòng rất lớn. Chứng tỏ sự bất bình đẳng về quy mô nhà ở.
* Histogram of total\_bedrooms: Tương tự như total\_rooms, phân bố lệch phải, cho thấy nhiều nhà ở có ít phòng ngủ hơn và ít nhà ở có số lượng phòng ngủ rất lớn.
* Histogram of population: Phân bố lệch phải. Hầu hết các khu vực có dân số tương đối nhỏ, với một số ít khu vực có dân số lớn hơn nhiều.
* Histogram of households: Cũng có phân bố lệch phải, tương tự total\_rooms và total\_bedrooms, cho thấy số hộ gia đình nhỏ chiếm đa số, và số hộ gia đình lớn hơn khá ít.
* Histogram of median\_income: Phân bố đa đỉnh, cho thấy một số mức thu nhập trung bình phổ biến trong bộ dữ liệu. Điều này cho thấy sự phân tầng rõ rệt về thu nhập.
* Histogram of median\_house\_value: Phân bố lệch phải, cho thấy nhiều nhà ở có giá trị trung bình tương đối thấp, và một số ít nhà ở có giá trị rất cao. Đây là một phân bố điển hình của giá nhà ở, thường lệch phải do sự tồn tại của một số lượng nhỏ nhà ở có giá trị rất cao.

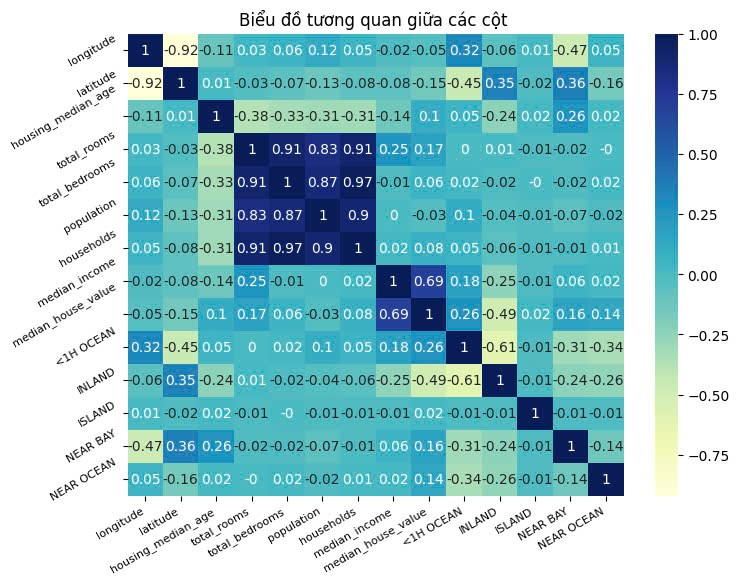
Biểu đồ Box Plot :



Hình 3.16. Biểu đồ Box Plot

Dựa theo sự phân phối trên, ta thấy các cột có mỗi quan hệ rất yếu so với cột median\_house\_value.

Biểu đồ tương quan các cột :

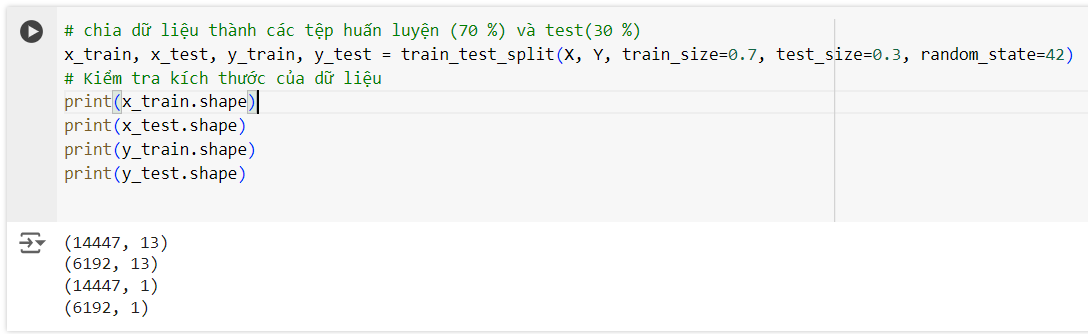


Hình 3.17. Biểu đồ Box Plot

Heatmap thể hiện tương quan mạnh giữa total\_rooms, total\_bedrooms, population, households; tương quan dương giữa median\_income và median\_house\_value; và ảnh hưởng của vị trí địa lý đến giá nhà. Một số biến có tương quan yếu hoặc không tương quan.

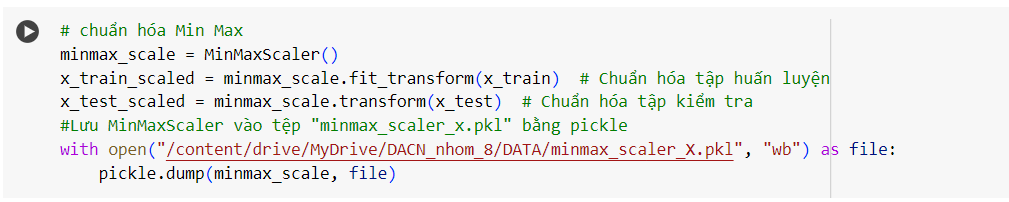
1. Phân tích hồi quy

Ta tiến hành chia tập dữ liệu thành 2 phần , 70% dùng để huấn luyện (train\_size=0.7) và 30% còn lại dùng để kiểm tra (test\_size=0.3).



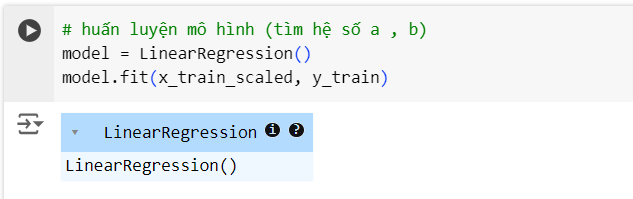
Hình 3.18: Chia tập dữ liệu huấn luyện và kiểm tra

Tiếp đó, tiến hành chuẩn hóa dữ liệu cho các biến độc lập



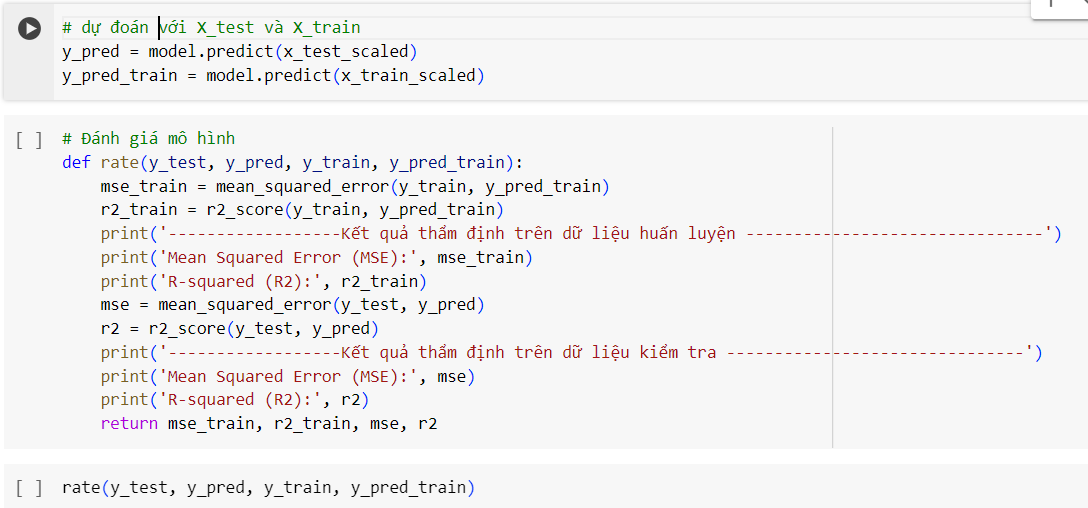
Hình 3.19: Chuẩn hóa dữ liệu

Ta tiến hành huấn luyện bằng mô hình “Hồi quy tuyến tính” được thể hiện thông qua LinearRegression().



Hình 3.20: Huấn luyện bằng mô hình Hồi quy tuyến tính

Cuối cùng, ta dự đoán kết quả trên tập X\_train và X\_test sau đó sử dụng hàm rate() để tiến hành đánh giá mô hình.



Hình 3.21: Dự đoán và đánh giá mô hình

Hàm rate sẽ cho chúng ta thấy được 2 độ đo là MSE (Mean Square Error) và R2\_score được thể hiện trên dữ liệu huấn luyện (train) và dữ liệu thẩm định (test). Sau đây là kết quả:

------------------Kết quả thẩm định trên dữ liệu huấn luyện -----------------------

Mean Squared Error (MSE): 4238473157.105766

R-squared (R2): 0.671074490160668

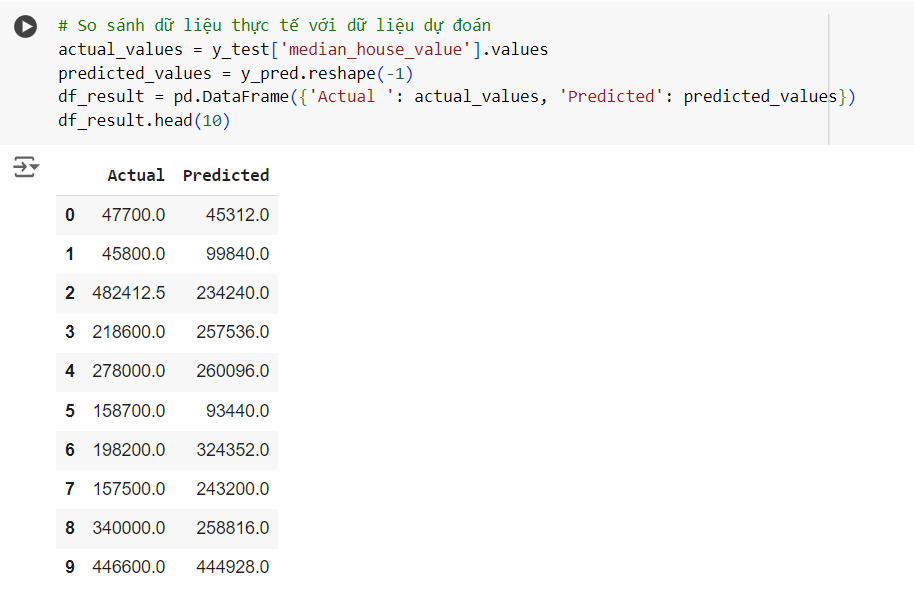
------------------Kết quả thẩm định trên dữ liệu kiểm tra --------------------------

Mean Squared Error (MSE): 4448074117.059552

R-squared (R2): 0.6485968222151116

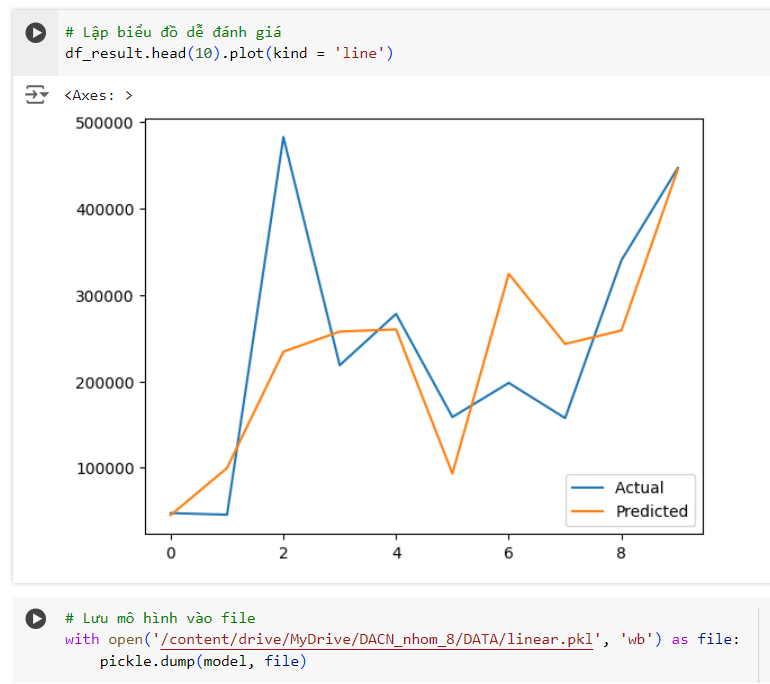
(4238473157.105766, 0.671074490160668, 4448074117.059552, 0.6485968222151116)

Ta sẽ so sánh dữ liệu sau khi dự đoán với dữ liệu thực tế:



Hình 3.22: Kết quả so sánh dữ liệu dự đoán với dữ liệu thực tế

Cuối cùng, ta thực hiện lập biểu đồ để trực quan hóa kết quả đánh giá mô hình để dễ dàng theo dõi mà cải thiện hiệu suất mô hình. Lưu mô hình cho lần sử dụng sau.



Hình 3.23: Biểu đồ so sánh giữa dữ liệu dự đoán với dữ liệu thực tế

1. Đánh giá

Phần phân tích mô tả đã phân tích được bộ dữ liệu ra các biểu đồ phù hợp và cho ta cái nhìn tổng quan về các yếu tố ảnh hưởng đến giá nhà tại California, chẳng hạn như mối liên hệ giữa vị trí địa lý, số phòng, và thu nhập trung bình. Tuy nhiên, đối với bài toán dự báo, mô hình phân tích hồi quy tuyến tính đang không đạt hiệu quả cao với điểm R2 chỉ ở mức trung bình. Điều này cho thấy mô hình chưa thể nắm bắt tốt mối quan hệ giữa các đặc điểm và biến mục tiêu (‘median\_house\_value’), có thể do mối quan hệ giữa các biến phức tạp hoặc dữ liệu chưa được xử lý tối ưu.

1. Kết luận

Chương 3 đã trình bày phần thực nghiệm và đánh giá của dự án thông qua đầy đủ các bước, từ tiền xử lý dữ liệu, phân tích mô tả, đến xây dựng và đánh giá mô hình dự báo. Qua đó, đã đưa ra được các nhận xét về hiệu suất mô hình hồi quy tuyến tính, chỉ ra những hạn chế và các yếu tố cần cải thiện. Đồng thời, chương cũng đề xuất các hướng phát triển tiềm năng để nâng cao độ chính xác và hiệu quả của dự án trong tương lai.