**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**NHẬP MÔN HỌC MÁY**

**ĐỀ TÀI:**

**SỬ DỤNG MÔ HÌNH HỌC MÁY HỒI QUY TUYẾN TÍNH**

**(LINEAR REGRESSION) CHO BÀI TOÁN DỰ ĐOÁN NHÀ ĐẤT**

|  |  |
| --- | --- |
|  |  |
| |  |  |  | | --- | --- | --- | | **Sinh viên thực hiện** | **: Đặng Quang Thành**  **Nguyễn Kiên Sơn** | | | **Giảng viên hướng dẫn** | | **: Vũ Văn Định** | | | **Ngành** | | **: Công nghệ thông tin** | | | **Chuyên ngành** | | **: Công nghệ phần mềm** | | | **Lớp** | | **: D16CNPM1** | | | **Khóa** | | **: 2021-2025** | | |  | |
| ***Hà Nội, tháng 4 năm 2024*** |  |

**PHIẾU CHẤM ĐIỂM**

**Sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| Họ và tên | Chữ ký | Công việc | Điểm |
| Nguyễn Kiên Sơn |  |  |  |
| Đặng Quang Thành |  |  |  |

**Giảng viên chấm:**

|  |  |  |
| --- | --- | --- |
| Giảng viên | Chữ ký | Ghi chú |
| Giảng viên chấm 1: |  |  |
| Giảng viên chấm 2: |  |  |

**CHƯƠNG 1:GIỚI THIỆU VỀ HỌC MÁY VÀ THUẬT TOÁN HỒI QUY TUYẾN TÍNH**

**1.1.Giới thiệu về học máy**

**1.1.1.Khái niệm về học máy**

Học máy là một lĩnh vực trong trí tuệ nhân tạo (AI) mà máy tính được lập trình để tự động học và cải thiện từ dữ liệu mà không cần phải được lập trình một cách cụ thể cho mỗi nhiệm vụ. Mục tiêu của học máy là phát triển các thuật toán và mô hình máy tính để hiểu và ánh xạ dữ liệu để có thể tự động học từ kinh nghiệm.

Học máy hiện nay vẫn cần sự đánh gía của con người trong việc tìm hiểu dữ liệu và lựa chọn kĩ thuật phù hợp để phân tích dữ liệu.Dữ liệu cần phải sạch,không có dữ liệu giả và không có sai lệch.Các mô hình học máy yêu cầu các dữ liệu phải đủ lớn để có thể học và đánh giá mô hình.

**1.1.2.Ứng dụng của học máy**

**Máy truy tìm dữ liệu:** Học máy được sử dụng để tìm kiếm thông tin trong một lượng lớn dữ liệu.

**Chẩn đoán y khoa:** Học máy được sử dụng để chẩn đoán bệnh và đưa ra phương án điều trị.

**Phát hiện gian lận:** Học máy được sử dụng để phát hiện các hoạt động gian lận trong tài chính, bảo hiểm,...

**Phân tích thị trường chứng khoán:** Học máy được sử dụng để dự đoán giá cổ phiếu và đưa ra quyết định đầu tư.

**Phân loại các chuỗi DNA:** Học máy được sử dụng để phân loại các chuỗi DNA và phát hiện các đột biến gen.

**Nhận dạng tiếng nói và chữ viết:** Học máy được sử dụng để nhận dạng tiếng nói và chữ viết trong các ứng dụng như trợ lý ảo, dịch thuật tự động,...

**Chơi trò chơi:** Học máy được sử dụng để tạo ra các chương trình chơi cờ vua, cờ vây,... có thể đánh bại con người.

**Cử động rô-bốt:** Học máy được sử dụng để điều khiển rô-bốt thực hiện các nhiệm vụ phức tạp.

Những ứng dụng này chỉ là một phần nhỏ trong số nhiều cách mà học máy đang thay đổi cách chúng ta làm việc và sống hàng ngày.

**1.2.Thuật toán hồi quy tuyến tính trong học máy**

**1.2.1.Giới thiệu về thuật toán hồi quy tuyến tính**

Hồi quy tuyến tính(Linear Aggression) là một thuật toán học máy **học có giám sát** được sử dụng để **dự đoán** một giá trị liên tục (biến phụ thuộc) dựa trên một hoặc nhiều giá trị liên tục hoặc phân loại (biến độc lập). Nó hoạt động bằng cách **tìm một đường thẳng** mô tả tốt nhất mối quan hệ giữa các biến độc lập và biến phụ thuộc trong tập dữ liệu huấn luyện.

Nói cách khác,”Hồi quy tuyến tính” là một phương pháp để dự đoán biến phụ thuộc(Y) dựa trên giá trị của biến độc lập(X).Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán số lượng liên tục.

**1.2.2.Dạng thuật toán**

Dạng của Linear Aggression (Hồi quy Tuyến tính) có thể được biểu diễn bằng phương trình sau:

**Hồi quy tuyến tính đơn giản:**

Hồi quy tuyến tính đơn giản được định nghĩa bằng hàm tuyến tính:

Y= β0\*X + β1 + ε

β0 và β1 là hai hằng số chưa biết đại diện cho độ dốc hồi quy, trong khi ε (epsilon) là thuật ngữ sai số.

Bạn có thể sử dụng hồi quy tuyến tính đơn giản để lập mô hình mối quan hệ giữa hai biến, chẳng hạn như sau:

* Lượng mưa và sản lượng cây trồng
* Tuổi và chiều cao ở trẻ em
* Nhiệt độ và sự giãn nở của thủy ngân kim loại trong nhiệt kế

**Hồi quy tuyến tính bội:**

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 + β1X1 + β2X2+…… βnXn+ ε

Khi số lượng biến dự báo tăng lên, các hằng số β cũng tăng lên tương ứng.

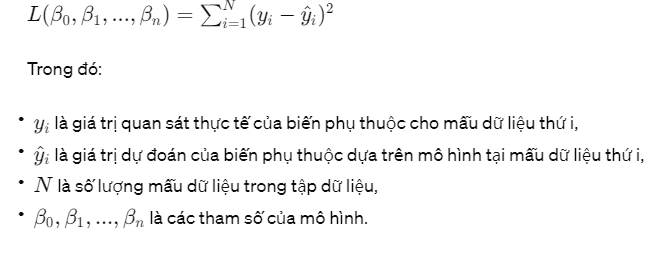
 Hồi quy tuyến tính bội lập mô hình nhiều biến và tác động của chúng lên một kết quả:

* Lượng mưa, nhiệt độ và mức độ sử dụng phân bón đối với sản lượng cây trồng
* Chế độ ăn uống và tập thể dục đối với bệnh tim
* Tăng trưởng tiền lương và lạm phát trên lãi suất cho vay gia đình

**1.2.3.Hàm mất mát(Loss Function)**

Trong ngữ cảnh của hồi quy tuyến tính, hàm mất mát thường được sử dụng để đo lường mức độ sai lệch giữa các dự đoán của mô hình và các giá trị thực tế. Mục tiêu của việc điều chỉnh các tham số của mô hình thông qua quá trình học máy là làm cho hàm mất mát đạt giá trị nhỏ nhất có thể.

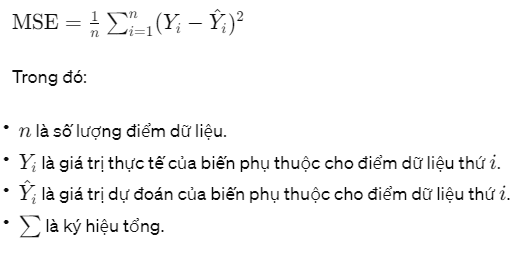
Trong hồi quy tuyến tính, hàm mất mát thường được chọn là tổng bình phương của sai số giữa dự đoán và giá trị quan sát:



Mục tiêu của việc tối ưu hóa hàm mất mát là tìm ra các giá trị của tham số , sao cho hàm mất mát đạt giá trị nhỏ nhất có thể

**1.2.4.Mức độ lỗi của mô hình hồi quy tuyến tính**

Sai số trung bình bình phương (Mean Squared Error - MSE): MSE là sự trung bình của bình phương của các sai số giữa các giá trị dự đoán từ mô hình và các giá trị thực tế.



Với MSE càng nhỏ,mô hình dự đoán càng chính xác

**CHƯƠNG 2 ỨNG DỤNG THUẬT TOÁN**

**2.1 Giới thiệu bài toán**

Lấy bối cảnh ở nước Mỹ, chúng ta sẽ đóng vai trò như một đại lý nhà nước để dự đoán giá nhà cho các khu vực. Với tập dữ liệu đã được chuẩn bị sẵn, nhiệm vụ bây giờ đó là sử dụng mô hình hồi quy tuyến tính để có thể ước tính ngôi nhà sẽ được bán với giá bao nhiêu.

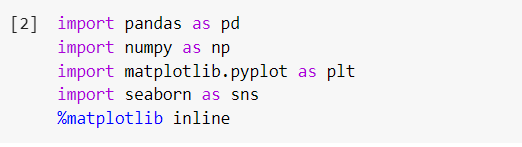
Tệp dữ liệu là file mở rộng CSV. Trong tệp dữ liệu này có 7 cột và 5000 hàng:

*  Avg. Area Income: Thu nhập trung bình tại khu vực ngôi nhà đã bán
*  Avg. Area House Age: Trung bình tuổi của một ngôi nhà đã bán
*  Avg. Area Number of Rooms: Trung bình diện tích các phòng
*  Avg. Area Number of Bedrooms: Trung bình diện tích phòng ngủ
*  Area Population: Dân số tại khu vực bán nhà
*  Price: Giá ngôi nhà đã bán
*  Address: Địa chỉ ngôi nhà bán

**2.2 Giải quyết bài toán**

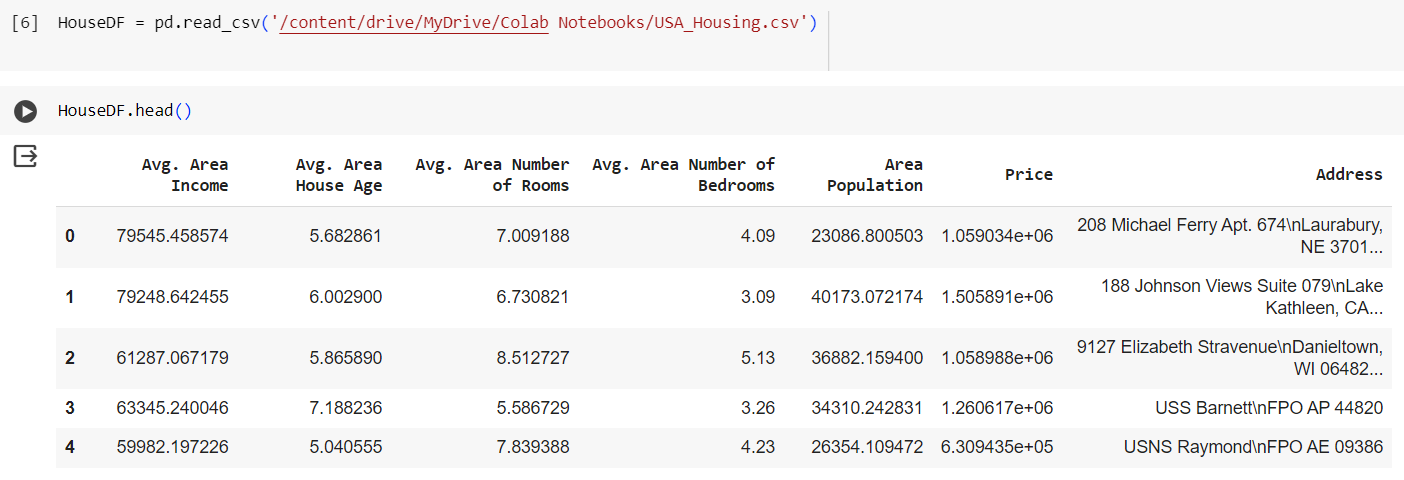
**2.2.1 Phân tích dữ liệu dataset**

Chúng ta sẽ sử dụng 1 số thư viện cần thiết để xử lí và phân tích trực quan dữ liệu:



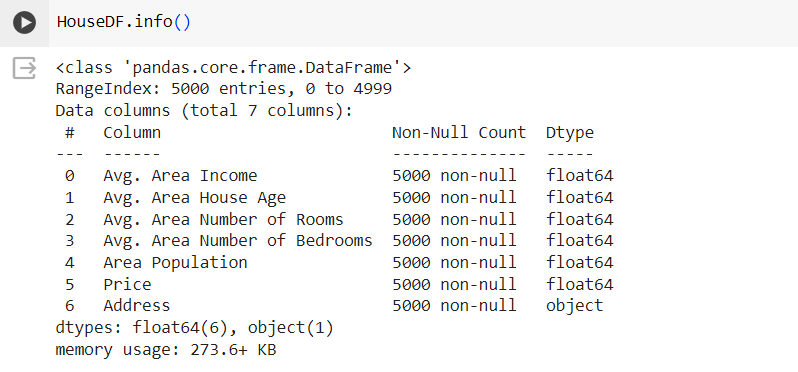
Hình 2.1 Import thư viện

Tiếp theo chúng ta sẽ load tệp dữ liệu có đuôi .csv:



Hình 2.2 Load dữ liệu từ file csv

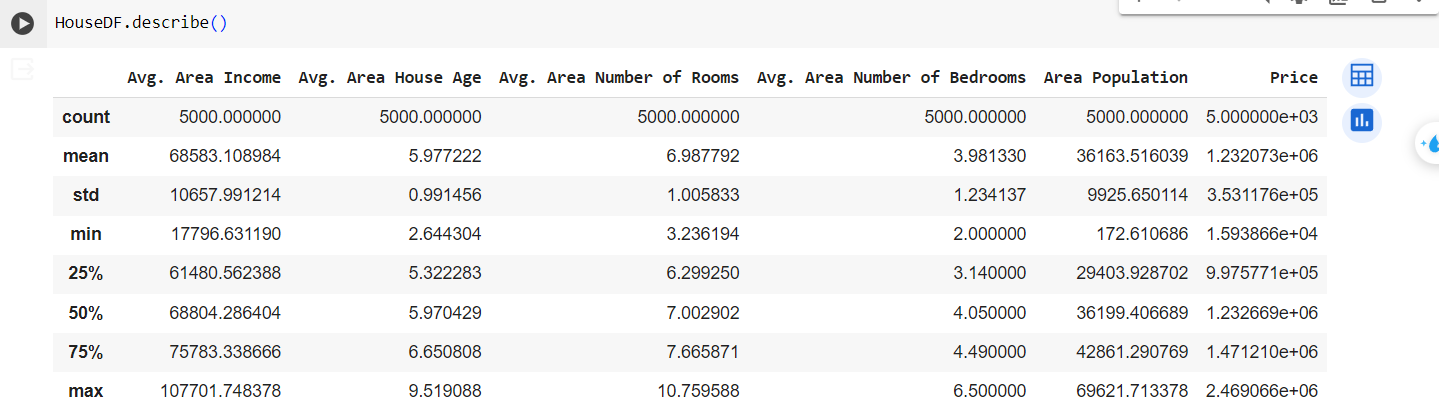
Để có thể xem tổng quát về dataset này ta có thể sử dụng:



Hình 2.3 Tổng quan về dataset

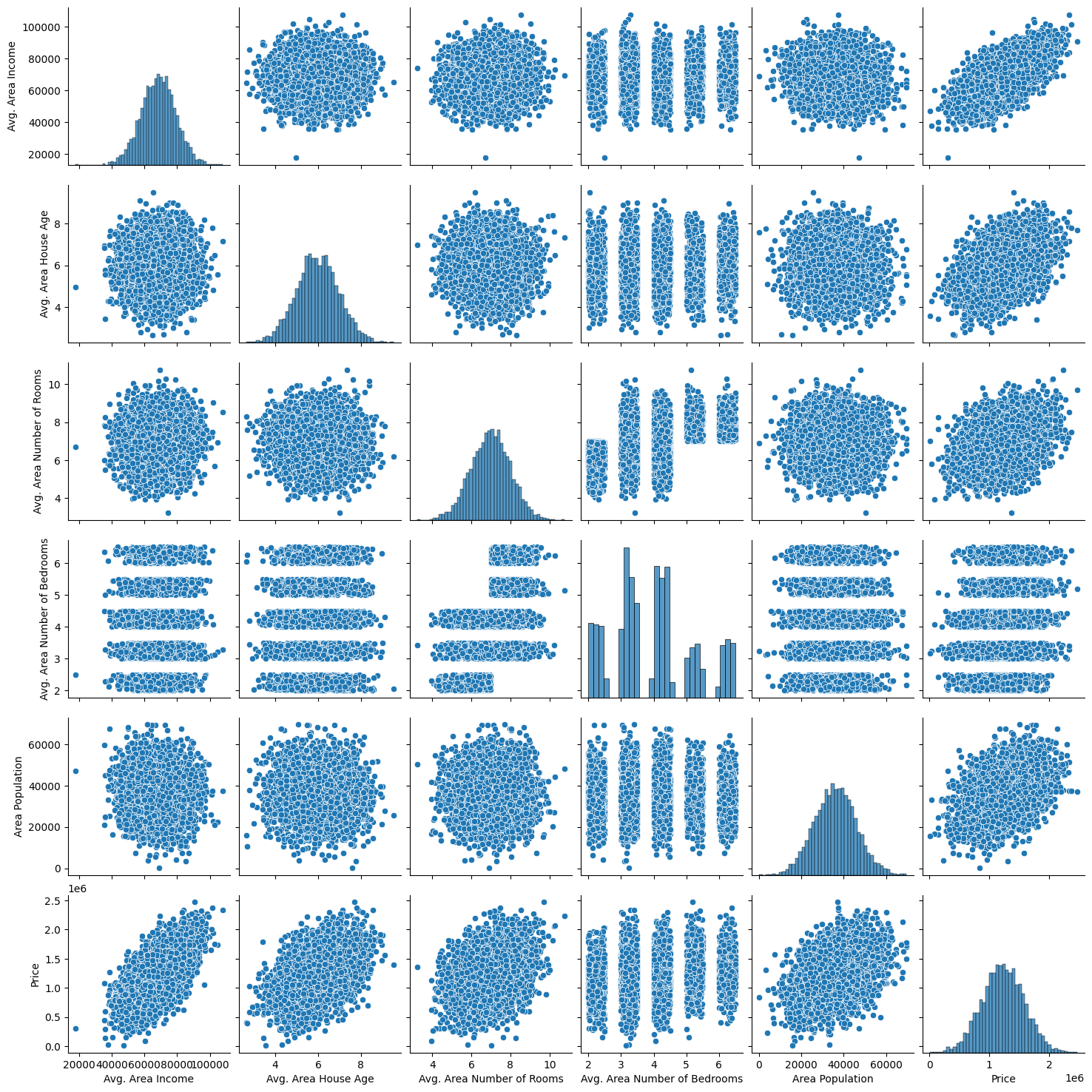
Dựa trên dữ liệu này ta có 5000 dòng tương ứng với 5000 ngôi nhà đã được bán. Lượng dữ liệu này đủ để có thể xây dựng một mô hình học máy

Ngoài ra ta còn có thể thống kê dữ liệu này bằng describe():



Nhờ hàm describe() giúp trả về dataframe mới với số hàng được hiển thị ra các thông số như số hàng, giá trị trung bình, độ lệch chuẩn, min, max, tỉ lệ phần trăm của các cột.

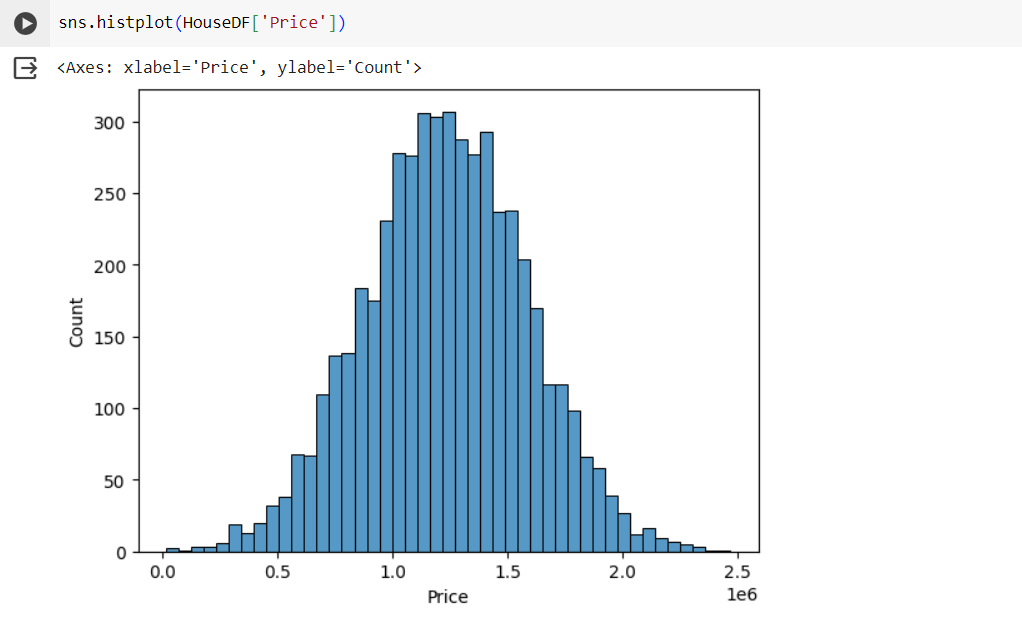
Phân bố bằng Seaborn ta có thể thấy được tương quan về dữ liệu



Hình 2.5 Phân bố bằng Seaborn

Về tương quan giữa các cột, ta thấy Cột Price có kiểu phân tán theo mô hình tuyến tính, dựa trên thông tin này, ta xây dựng mô hình máy học hồi quy tuyến tính để dự đoán nó dựa trên giá trị các cột khác, trừ địa chỉ (Address) ngôi nhà.

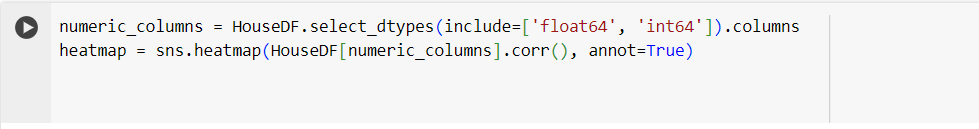
Bây giờ chúng ta sẽ sử dụng hisplot() để vẽ biểu đồ giá ngôi nhà:

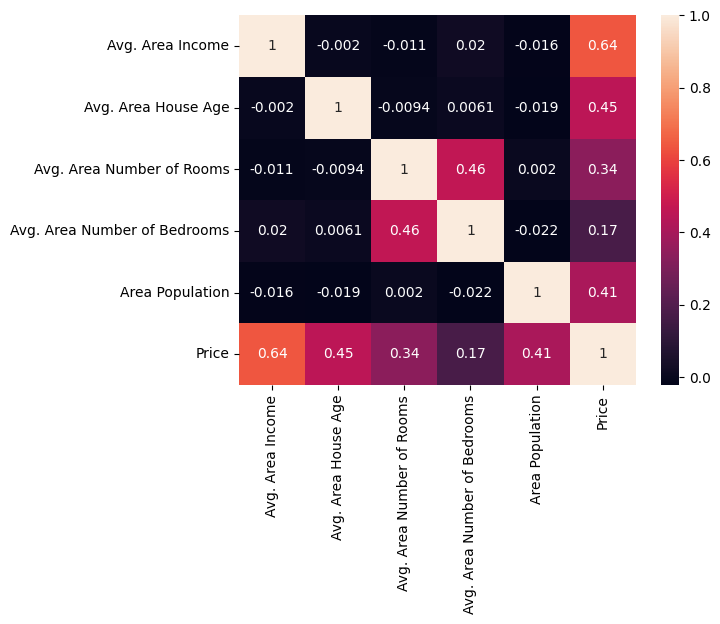


Hình 2.6 sẽ sử dụng histplot() để vẽ biểu đồ giá nhà

Ta thấy giá các ngôi nhà đã bán thường tập trung ở mức giá 0.5 đến 2.0, và nhiều nhất là 0.8 đến 1.7

Chúng ta sẽ sử dụng bản đồ nhiệt để kiểm tra độ tương quan giữa các cột:





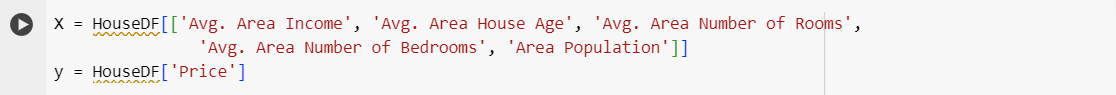
Hình 2.7 bản đồ nhiệt để kiểm tra độ tương quan giữa các cột

Qua đó, ta phân tích được các cột có giá trị tương quan như như thế nào với nhau. Về cơ bản, cột giá (price) có chút tương quan với các cột còn lại nhiều nhất, chứng tỏ các yếu tố đó có tác động ít nhiều lên giá nhà.

**2.2.2 Xây dựng mô hình dự đoán bằng thư viện Scikit – Learn**

**2.2.2.1 Phân tách dữ liệu thành train và test**

Bây giờ chúng ta hãy bắt đầu đào tạo mô hình hồi quy. Trước tiên, chúng ta sẽ cần tách dữ liệu của mình thành một mảng X chứa các tính năng cần đào tạo (các biến độc lập) và một mảng y với biến mục tiêu (biến phụ thuộc), trong trường hợp này là cột Giá. Chúng ta sẽ loại bỏ cột Địa chỉ vì nó chỉ có thông tin văn bản mà mô hình hồi quy tuyến tính không thể sử dụng



Hình 2.8 tách dữ liệu thành một mảng X chứa các tính năng cần đào tạo và một mảng y với biến mục tiêu

Giờ ta đã có hai biến x, y theo yêu cầu của mô hình, hai biến này dựa trên dữ liệu là dataset ta có được để đào tạo mô hình. Giờ là ta tách các biến trên thành giá trị train và test, hai giá trị này chúng ta sẽ luôn gặp và sử dụng trong qua trình xây dựng mô hình máy học.

Đầu tiên, ta từ thư viện Scikti - Learn model\_selection ta import train\_test\_split, phương thức này giúp ta tạo mô hình hồi quy