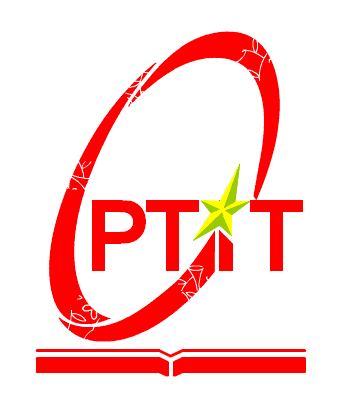
**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



BÁO CÁO THỰC TẬP CƠ SỞ

Chuyên ngành: **CÔNG NGHỆ THÔNG TIN**

Đề tài: Data Analyst sử dụng Python

Dự đoán giá nhà

**Giảng viên hướng dẫn:** TS. Đào Ngọc Phong

**Sinh viên thực hiện:** Lê Hoàng Long

Mã sinh viên: B20DCCN405

Lớp: D20CQCN405

SĐT:   0949404106

Email: lehoanglonght12@gmail.com

***Hà Nội, 4/2022***

**Lời cảm ơn**

Em muốn gửi lời cảm ơn tới Thầy Đào Ngọc Phong – và nhà trường đã tạo điều kiện để chúng em có môn thực hành cơ sở. Nhờ có môn học này mà em đã có thể tự thúc đẩy mình trong quá trình tiếp thu kiến thức và tự tìm hiểu các kiến thức mới. Trong quá trình tìm hiểu về đề tài Python và Data như làm báo cáo thực tập này, do trình độ và kinh nghiệm thực tiễn còn hạn chế nên không tránh khỏi những thiếu sót, em rất mong nhận được ý kiến đóng góp từ giảng viên hướng dẫn để em có thể học hỏi, rút kinh nghiệm và cải thiện trong tương lai.

I, Các thư viện Python sử dụng trong dự án này:  
+ NumPy: là viết tắt của "Numerical Python" và cung cấp một thư viện xử lý mảng mạnh mẽ cho Python. Các mảng NumPy được sử dụng để lưu trữ các bộ số lớn và thư viện bao gồm một loạt các hàm cho các phép toán toán học trên các mảng này. Đây là một số tính năng chính của NumPy:

* Mảng hiệu quả: các mảng NumPy hiệu quả hơn nhiều so với danh sách Python thông thường để lưu trữ và xử lý lượng lớn dữ liệu số.
* Hàm toán học: NumPy bao gồm một loạt các hàm toán học cho việc làm việc với các mảng, bao gồm các hàm cho phép tính toán cơ bản, lượng giác và đại số tuyến tính.
* Broadcasting: NumPy bao gồm một tính năng gọi là "broadcasting" cho phép bạn thực hiện các phép toán số trên các mảng với các hình dạng khác nhau.
* Tạo số ngẫu nhiên: NumPy bao gồm một module để tạo ra số ngẫu nhiên, rất hữu ích cho mô phỏng và phân tích thống kê.

+ Pandas: là một thư viện phân tích dữ liệu cung cấp các công cụ để làm việc với dữ liệu có cấu trúc trong Python. Nó được xây dựng trên nền tảng NumPy và cung cấp các chức năng nâng cao hơn cho việc xử lý dữ liệu. Đây là một số tính năng chính của Pandas:

* Dataframes: Pandas cung cấp một đối tượng "DataFrame" mạnh mẽ tương tự như một bảng tính hoặc bảng SQL. DataFrames có thể được sử dụng để lưu trữ và xử lý dữ liệu có cấu trúc.
* Xử lý dữ liệu: Pandas bao gồm một loạt các hàm để xử lý dữ liệu, bao gồm các hàm để lọc, sắp xếp và nhóm dữ liệu.
* Dữ liệu bị thiếu: Pandas bao gồm các hàm để xử lý dữ liệu bị thiếu trong tập dữ liệu của bạn, chẳng hạn như điền các giá trị bị thiếu hoặc xóa các hàng có giá trị bị thiếu.
* Ghép dữ liệu: Pandas bao gồm các hàm để hợp nhất nhiều bộ dữ liệu với nhau, điều này rất hữu ích để kết hợp dữ liệu từ nhiều nguồn.

+ Matplotlib: là một thư viện vẽ đồ thị cho Python cung cấp các công cụ để tạo hình ảnh của dữ liệu của bạn. Dưới đây là một số tính năng chính của Matplotlib:

* Các hàm vẽ đồ thị: Matplotlib bao gồm một loạt các hàm để tạo ra các loại đồ thị khác nhau, bao gồm các đồ thị đường, đồ thị phân tán và đồ thị histogram.
* Tùy chỉnh: Matplotlib cung cấp một loạt các tùy chọn để tùy chỉnh các biểu đồ của bạn, chẳng hạn như thay đổi màu sắc hoặc kiểu của các đường hoặc điểm đánh dấu.
* Các bản phân đoạn (subplots): Matplotlib cho phép bạn tạo nhiều biểu đồ trong một hình sử dụng các bản phân đoạn.

+ Seaborn : là một thư viện trực quan hóa dữ liệu dựa trên Matplotlib. Nó cung cấp một giao diện cao cấp để tạo ra đồ họa thống kê hấp dẫn và thông tin. Seaborn có thể được sử dụng để trực quan hóa phân phối đơn biến và đa biến, các mô hình hồi quy và các biến phân loại. Các tính năng chính của Seaborn bao gồm:

* Hỗ trợ cho các loại biểu đồ khác nhau như biểu đồ phân tán, biểu đồ đường, biểu đồ thanh, biểu đồ hộp, biểu đồ violin, bản đồ nhiệt và nhiều hơn nữa.
* Dễ dàng tùy chỉnh các yếu tố thẩm mỹ của biểu đồ bao gồm màu sắc, phông chữ và kích thước.
* Hỗ trợ tích hợp cho các biến phân loại và các lưới nhiều biểu đồ.
* Tích hợp với cấu trúc dữ liệu Pandas để dễ dàng xử lý và phân tích dữ liệu.
* Tích hợp với Matplotlib cho các tùy chỉnh và tùy chọn vẽ biểu đồ nâng cao hơn.

+ scikit-learn : là một thư viện học máy cho Python cung cấp một loạt các công cụ để xây dựng các mô hình dự đoán. Dưới đây là một số tính năng chính của scikit-learn:

* Các thuật toán học máy: scikit-learn bao gồm một loạt các thuật toán học máy, bao gồm các thuật toán hồi quy, phân loại và gom cụm.
* Các chức năng tiền xử lý: scikit-learn bao gồm các chức năng để tiền xử lý dữ liệu của bạn trước khi xây dựng một mô hình, chẳng hạn như tỷ lệ hoặc chuẩn hóa dữ liệu của bạn.
* Đánh giá mô hình: scikit-learn bao gồm các chức năng để đánh giá hiệu suất của mô hình của bạn, chẳng hạn như tính độ chính xác hoặc tính ma trận nhầm lẫn.
* Các đường ống (pipelines): scikit-learn bao gồm một tính năng gọi là "đường ống" cho phép bạn kết nối các bước tiền xử lý và một mô hình thành một đối tượng duy nhất.

+ Folium : là một thư viện Python để tạo bản đồ tương tác bằng cách sử dụng Leaflet.js. Nó cho phép bạn trực quan hóa dữ liệu địa lý trên bản đồ tương tác với đánh dấu, cửa sổ bật lên, bản đồ choropleth và nhiều hơn nữa. Folium được xây dựng trên các thư viện Python như Pandas, Numpy và Matplotlib và có thể được sử dụng để tạo bản đồ tương tác từ nhiều nguồn dữ liệu bao gồm các tệp CSV, tệp GeoJSON và Pandas DataFrames. Các tính năng chính của Folium bao gồm:

* API Python dễ sử dụng để tạo bản đồ tương tác.
* Hỗ trợ cho các loại bản đồ khác nhau bao gồm bản đồ tile, bản đồ choropleth và bản đồ nhiệt.
* Tích hợp với Pandas DataFrames để dễ dàng xử lý và phân tích dữ liệu.
* Hỗ trợ cho các biểu tượng và đánh dấu tùy chỉnh.
* Hỗ trợ tích hợp cho các cửa sổ bật lên và thông tin chú thích.
* Khả năng xuất bản đồ dưới dạng tệp HTML để dễ dàng

II, Giới thiệu tập dữ liệu phân tích:

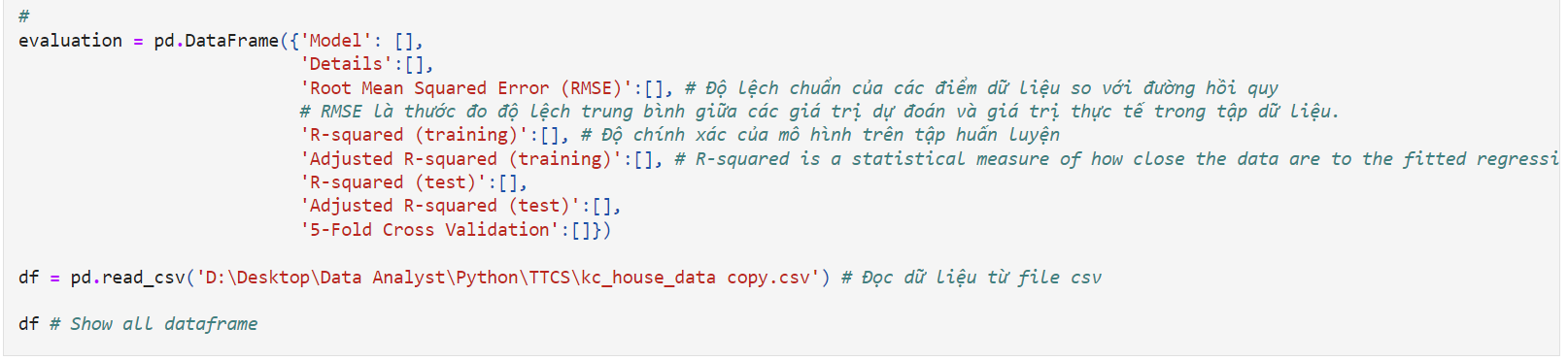
Đây là một tập dữ liệu về giá nhà trên Kaggle gồm 21613 dòng và 21 cột dữ liệu . Tập dữ liệu này chứa giá bán nhà cho Quận King, bao gồm thành phố Seattle. Nó bao gồm các ngôi nhà được bán từ tháng 5 năm 2014 đến tháng 5 năm 2015. Dưới đây là mô tả các column trong tập dữ liệu

|  |  |
| --- | --- |
| **Tên cột** | **Mô tả** |
| Id | Id dùng để phân biệt các ngôi nhà |
| Date | Ngày nhà được bán |
| price | GIá nhà - Đây là column ta muốn dự đoán |
| bedrooms | Số phòng ngủ |
| bathrooms | Số phòng tắm |
| sqft\_living | Diện tích sống của ngôi nhà |
| sqft\_lot | Diện tích của khu đất |
| floors | Tổng số tầng trong ngôi nhà |
| waterfront | Nhà có tầm nhìn ra sông hoặc biển không |
| view | Đã được xem |
| condition | Tình trạng chung của ngôi nhà |
| grade | Mức độ tổng thể của ngôi nhà, dựa trên hệ thống xếp loại của King County |
| sqft\_above | Diện tích của nhà ngoài tầng hầm |
| sqft\_basement | Diện tích của tầng hầm |
| yr\_built | Năm xây dựng |
| yr\_renovated | Năm nhà được cải tạo |
| zipcode | Mã Zip |
| lat | Vĩ độ |
| long | Kinh độ |
| sqft\_living15 | Diện tích phòng khách vào năm 2015 ( một số cải tạo) Điều này có thể đã ảnh hưởng hoặc không ảnh hưởng đến diện tích khu đất |
| sqft\_lot15 | Diện tích khu đất vào năm 2015 (một số cải tạo) |

III, Demo phần phân tích :

Lưu ý : Phần này trình hơi khó nhìn vì trình bày cả hình lẫn chữ, do các công thức toán học khá khó viết trên Word nên em chụp từ phần Markdown bên Jupiter Notebook sang luôn. Mong thầy thông cảm vì sự bất tiện này ạ .

# Để thực hiện một số phân tích, chúng ta cần thiết lập môi trường làm việc của mình. Để làm điều này, em đã nhập một số module và đọc dữ liệu. Đầu ra dưới đây là toàn bộ dữ liệu em đã xác định một bảng dữ liệu rỗng. Bảng dữ liệu này bao gồm **Root Mean Squared Error (RMSE là thước đo độ lệch trung bình giữa các giá trị dự đoán và giá trị thực tế trong tập dữ liệu), R-squared (R-squared là một đại lượng thống kê đo độ lệch của dữ liệu với đường hồi quy được sử dụng để khớp với dữ liệu), Adjusted R-squared** và trung bình các giá trị **R-squared** được đạt được bằng phương pháp **Kiểm định chéo k-Fold**, đây là các chỉ số quan trọng để so sánh các mô hình khác nhau. Một giá trị **R-squared** gần với một và RMSE càng nhỏ thì sai số càng ít



A picture containing text, screenshot, font, number

Description automatically generated

