BÁO CÁO ĐỒ ÁN

1712302- Bùi Lý Chung

1712575 -Hoàng Xuân Long

1/14/2021

I. THU THẬP DỮ LIỆU

- Hình thức thu thập dữ liệu: Crawler

- Nguồn dữ liệu : https://www.zillow.com/manchester-nh/sold/
- Mô tả: Dữ liệu về những căn nhà **đã được bán** ở Manchester, NH
- Lí do lựa chọn dữ liệu: nguồn cung cấp thực phẩm và sự hiểu biết khoa học về mối quan hệ giữa khẩu phần ăn và sức khỏe đã phát triển qua nhiều năm do nhu cầu về sức khỏe của con người ngày càng cao

HÌNHẢNHVỀDỮLIỆUTHUTHẬPĐƯỢC

2. Khám phá dữ liệu

In [102]:
```
# Đọc dữ liệu
House_df = pd.read_csv('courses.csv',index_col=0)
House_df.head(10)
```

Out[102]:

	Address	Bedrooms	Bathrooms	LivingArea	Heating	Basement	YearBuild	Fireplace	Garage	Price
0	216 Walek Farms Rd, Manchester, CT 06040	3 bd	3 ba	2,556 sqft	Baseboard, Gas	Finished	1995	Yes	2	$300,000
1	39 S Hawthorne St, Manchester, CT 06040	3 bd	1 ba	1,080 sqft	Forced air, Gas	Partially finished	1945	None	0	$170,000
2	16 Saddle Hill Rd, Manchester, CT 06040	4 bd	4 ba	3,832 sqft	Forced air	Finished	2007	Yes	2	$447,900
3	99 Pond Ln, Manchester, CT 06042	4 bd	3 ba	1,852 sqft	Forced air, Oil	NaN	1964	Yes	3	$299,000
4	430 E Center St #430, Manchester, CT 06040	1 bd	1.5 ba	1,074 sqft	Forced air, Gas	None	1976	None	0	$1,200
5	443 Main St, Manchester, CT 06040	-- bd	5 ba	6,079 sqft	Forced air, Gas, Oil	Full With Hatchway	1900	None	0	$378,500
6	40 S Alton St, Manchester, CT 06040	4 bd	1 ba	1,155 sqft	Forced air, Gas	Full With Hatchway	1945	None	0	$175,000
7	11 Durant St, Manchester, CT 06040	3 bd	2 ba	1,416 sqft	Baseboard, Other, Gas	Finished	1950	Yes	4	$195,000
8	180 Mountain Rd, Manchester, CT 06040	4 bd	2 ba	2,628 sqft	Other, Oil	Unfinished, Interior Entry, Full With Walk-Ou...	1959	Yes	2	$265,000
9	33 Hartland Rd, Manchester, CT 06042	3 bd	2 ba	2,040 sqft	Baseboard, Oil	Finished	1951	Yes	None	$215,000

Shape của dữ liệu

In [45]:
```
House_df.shape
```

Out[45]: (360, 10)

II. KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

- Số dòng, cột của dữ liệu: (360, 10)
- Số dòng cột output thiếu: 0
- Các cột dữ liệu khác bị thiếu:
  - Bedrooms: 10
  - Bathrooms: 33
  - LivingArea: 3
  - Heating: 0
  - YearBuild: 7
  - Garage: 5

II. KHÁM PHÁ VÀ TIỀN XỬ LÝ DỮ LIỆU

- **Ý nghĩa của mỗi dòng dữ liệu:**

  - Address: Địa chỉ nhà.
  - Price: Gía nhà (là giá của nhà đã bán).
  - Bedrooms: số phòng ngủ.
  - Bathrooms: số phòng tắm.
  - Living Area: Diện tích căn nhà (sqft).
  - Heating : Hệ thống sưởi.
  - Basement : Tầng hầm(Yes/No).
  - Yearbuild: Năm xây dựng.
  - Fireplace: Lò sưởi(Yes/No).
  - Garage: Nhà xe.

ĐẶT CÂU HỎI?

- Giá nhà được tính theo công thức nào từ input?
  - Input : Số phòng ngủ, nhà tắm, tầng hầm, năm xây dựng, Garage,Diện tích,lò sưởi, hệ thống nhiệt, mã vùng
  - Output : Giá nhà
- Việc tìm ra câu trả lời cho câu hỏi này có ý nghĩa:
  - Giúp người mua nhà biết được giá nhà mình định mua trong cùng khu vực đó là có bị hớ hay không?
  - Giúp người muốn bán nhà có thể đưa ra được mức giá phù hợp
- Nguồn cảm hứng của câu hỏi:
  - Có lẽ bắt nguồn từ chính gợi ý của thầy trong buổi học cũng như lúc nói về đồ án cuối kì

KHÁM PHÁ VÀ TIỀN XỬ LÝ TRƯỚC KHI TÁCH TẬP

- Xử lý thô với các cột chứa kí tự không mong muốn (bd, ba, sqft)
- Kiểm tra cột ouput, loại bỏ kí tự $

TÁCH CÁC TẬP DỮ LIỆU

- Tách thành 2 tập train, validation ,test:60%, 20%,20%
- Vector input X: tất cả các cột dữ liệu ngoại trừ cột “Price
- Output Y: cột dữ liệu “Price

KHÁM PHÁ VÀ TIỀN XỬ LÝ TẬP HUẤN LUYỆN (KHÁM PHÁ)

- Kiểu dữ liệu của các cột trong vector input X: float64, object

```
In [318]: train_X_df.dtypes

Out[318]: Address        object
          Bedrooms      float64
          Bathrooms     float64
          LivingArea    float64
          Basement       object
          YearBuild     float64
          Fireplace      object
          Garage        float64
          dtype: object
```

KHÁM PHÁ VÀ TIỀN XỬ LÝ TẬP HUẤN LUYỆN (TIỀN XỬ LÝ)

- Với các cột có mising:
  - Fireplace: tỉ lệ missing quá lớn, loại bỏ cột này
  - Basement: chuyển đổi về Yes/No, điền bằng giá trị most_freq
  - Garage: điền bằng giá trị most_freq
- Đối với cột Address: Nhận thấy mã khu vực có vẻ sẽ ảnh hưởng lớn tới kết quả dự đoán, ta xử lý bằng cách bỏ cột này và thay bằng cột chứa mã vùng được tách ra từ cột address
- Các cột Bedrooms, Bathrooms hay LivingArea ta chuyển về dạng số

KHÁM PHÁ VÀ TIỀN XỬ LÝ TẬP HUẤN LUYỆN (TIỀN XỬ LÝ)

- Xây dựng class ColDropper kế thừa class BaseEstimator và TransformerMixin để thực hiện nhiệm vụ tiền xử lý xóa các cột đã nêu trên
- Xây dựng process_pipeline gồm các giai đoạn:
  - ColDropper(num_top_titles=4): Lấy 4 giá trị mã vùng cao nhất
  - SimpleImputer(stragety='mean'): thực hiện điền dữ liệu bị thiếu bằng chiến lược mean
  - StandardScaler(): chuẩn hóa tỉ lệ dữ liệu

- Xây dựng process_pipeline_full:
  - Thêm vào với process_pipeline tạo thành process_pipeline_full

TIỀN XỬ LÝ VÀ MÔ HÌNH HÓA (VALIDATION)

- Thực hiện transform trên tập validation
- Lựa chọn mô hình:
  - MLPRegression

MÔ HÌNH MLPRegression

- Thử nghiệm mô hình neural network với:
  - hidden_layer_sizes=(50), activation='tanh', solver='lbfgs', random_state=0, max_iter=5000
  - Siêu tham số alpha của MLPRegressor với 5 giá trị khác nhau: 0.1, 1, 10, 100, 1000
  - Siêu tham số num_top_titles của MLPRegressor với 5 giá trị khác nhau: 1, 3, 5, 7, 9, 11
  - Sau khi xây dựng mô hình, tạo một full_pipeline chứa process_pipeline_full và mô hình

- Tìm được best_alpha=100 và best_val_err=56.2

NHÌN LẠI QUÁ TRÌNH LÀM ĐỒ ÁN

- Khó khăn:
  - Khó khăn trong việc đi tìm nguồn dữ liệu chính thống với thông tin chính xác, ít nhiễu, ít thô
  - Khó khăn trong việc craw data về, document về api của web quá ít thông tin
  - Khá gấp rút làm đồ án vì vướng trong thời gian thi cử
  - Khó khăn trong việc tìm hiểu knowledge domain của dữ liệu
  - Khó khăn trong việc tiền xử lý dữ liệu

- Những điều hữu ích học được:
  - Được làm việc hoàn chỉnh trên một mô hình khoa học dữ liệu
  - Kĩ năng khám phá và tiền xử lí dữ liệu
  - Kĩ năng đọc và tìm hiểu dữ liệu, cũng như nghiên cứu các mô hình
  - Kĩ năng debug lỗi