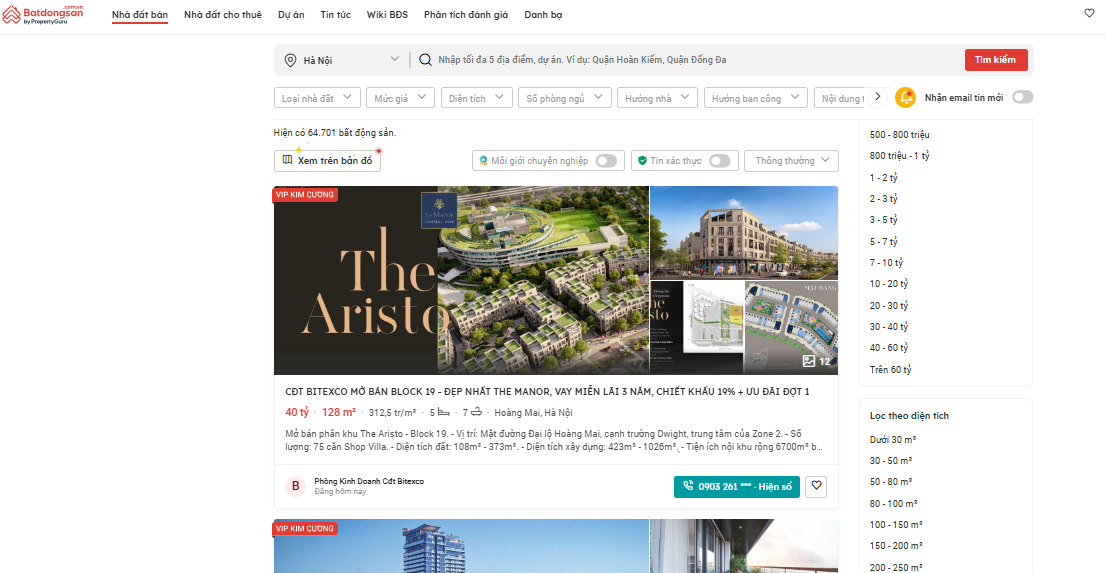
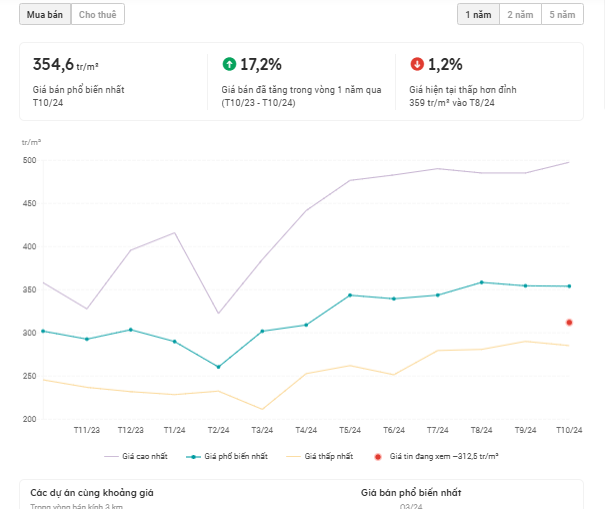
Phân tích dự liệu : Giá bất động sản trong thành phố Hà Nội

**Bước 1: Thu thập dữ liệu**- Sử dụng thư viện selenium để lấy dữ liệu trên trang web :

***https://batdongsan.com.vn/nha-dat-ban-ha-noi***



- Khó khăn:  
+Thời gian dành cho việc thu thập không được đảm bảo trong việc lấy dữ liệu từ thời gian trước.

+ Phải sử dụng chuột ảo để lấy thông tin tại các điểm trên biểu đồ (do không có cách khác để lấy thông tin giá thời gian trước)  


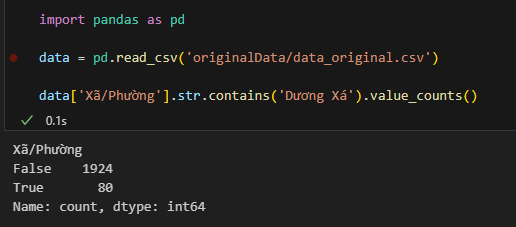
**Bước 2: Xử lý dữ liệu**

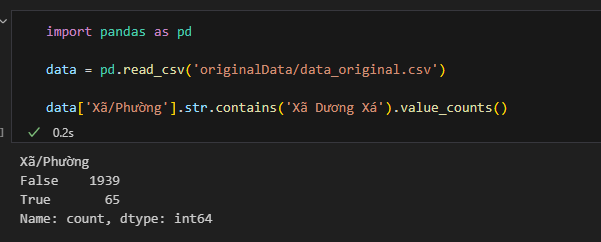
Lọc bỏ các bất động sản không phải ở Hà Nội  
Lọc bỏ các trường có thông tin ‘Đang cập nhật’

Lọc bỏ Mức giá ‘Thỏa thuận’

Lọc bỏ các trường ‘Không có thông tin’

**+ Xử lý trường ‘Xã/Phường’ và ‘Quận/Huyện’:**

Dữ liệu thô khi thu thập về sẽ có vấn đề : Có dữ liệu có chữ ‘Xã’ hoặc ‘Phường’ hoặc ‘Thị trấn’ có cái không  
Đây là ví dụ : ****

****

Vậy nên sẽ phải bỏ hết từ ‘Xã’ ‘Phường’ và ‘Thị trấn’ cho có tính thống nhất

**+ Xử lý trường Diện tích:**

**-** Bỏ chữ m^2 và chuyển về dạng số float **+ Xử lý trường Mức giá:**- Bỏ các kí tự đi và chuyển về số float   
- Dữ liệu thu thập sẽ có giá cả căn và giá trên đơn vị diện tích -> Thống nhất tất cả theo giá trên đơn vị diện tích (triệu/m^2)

**+ Xử lý trường Phòng ngủ và Phòng tắm:** Bỏ chữ phòng và chuyển về dạng số  
**+ Xử lý trường Pháp lý:   
-** Dữ liệu vào có Viết tắt Hợp đồng mua bán thành HĐMB -> Chuyển về Hợp đồng mua bán

- Xử lý các từ đồng nghĩa

**+ Xử lý trường Nội thất:**- Chúng tôi chia làm 3 loại : Cơ bản, Đầy đủ, Cao cấp- Xử lý các từ đồng nghĩa

**+ Xử lý trường Tên dự án hay chủ đầu tư:**

- Các dự án hay chủ đầu tư ‘Không có thông tin’ chính là các bất động sản Tư nhân

+ **Xử lý thông tin về giá**

Chia làm 2 loại : giá theo quý và giá theo tháng và không có thông tin và đặt các kí hiệu

Bước 3 : Trực quan và Phân tích (Dự đoán xu hướng)