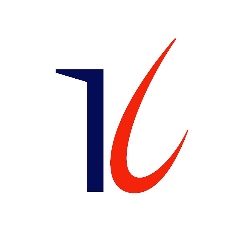
BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC THĂNG LONG**



**THU THẬP DỮ LIỆU**

**BẤT ĐỘNG SẢN TẠI HÀ NỘI**

**NHÓM SINH VIÊN THỰC HIỆN: A45839- NGUYỄN KHÁNH LINH**

**A44933- KHƯƠNG VĂN CƯƠNG**

**GIÁO VIÊN HƯỚNG DẪN: ĐOÀN TRUNG PHONG**

**HÀ NỘI – 2025**

**MỤC LỤC**

[CHƯƠNG 1. GIỚI THIỆU CHUNG 4](#_Toc188115025)

[1. Giới thiệu bài toán 4](#_Toc188115026)

[2. Giới hạn đối tượng khách thể, đối tượng nghiên cứu 4](#_Toc188115027)

[2.1. Đối tượng khách thế 4](#_Toc188115028)

[2.2. Đối tượng nghiên cứu 4](#_Toc188115029)

[3. Các nghiên cứu liên quan 5](#_Toc188115030)

[a. Lượng hóa các nhân tố ảnh hưởng đến giá đất ở 5](#_Toc188115031)

[b. Yếu tố ảnh hưởng đến quyết định mua bất động sản tại thành phố Hạ Long, tỉnh Quảng Ninh 5](#_Toc188115032)

[References 6](#_Toc188115033)

[CHƯƠNG 2. Cơ sở lý thuyết và công cụ sử dụng 7](#_Toc188115034)

[1. Cơ sở lý thuyết 7](#_Toc188115035)

[1.1. Thu thập dữ liệu 7](#_Toc188115036)

[1.2. Làm sạch dữ liệu 7](#_Toc188115037)

[1.3. Trực quan hoá dữ liệu 7](#_Toc188115038)

[2. Công nghệ, công cụ sử dụng 7](#_Toc188115039)

[2.1. Selenium 7](#_Toc188115040)

[2.2. BeautifulSoup 8](#_Toc188115041)

[2.3. Thư viện Pandas 8](#_Toc188115042)

[2.4. Thư viện Numpy 8](#_Toc188115043)

[2.5. Thư viện matplotlib 8](#_Toc188115044)

[2.6. Thư viện Seaborn 9](#_Toc188115045)

[2.7. Jupyter 9](#_Toc188115046)

[CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN 10](#_Toc188115047)

[1. Quy trình thực hiện 10](#_Toc188115048)

[CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM 11](#_Toc188115049)

[1. Kết quả thu thập dữ liệu 11](#_Toc188115050)

[2. Làm sạch dữ liệu 11](#_Toc188115051)

[2.1. Lỗi logic 11](#_Toc188115052)

[2.2. Đổi kiểu dữ liệu 11](#_Toc188115053)

[2.3. Xử lý dữ liệu NAN 12](#_Toc188115054)

[2.4. Xử lý dữ liệu trùng lặp 12](#_Toc188115055)

[2.5. Xử lí dữ liệu ngoại lai 13](#_Toc188115056)

[3. Phân tích khai thác dữ liệu 15](#_Toc188115057)

[3.1. Phân tích số lượng căn theo từng quận 15](#_Toc188115058)

[3.2. Phân tích giá trung bình theo từng quận 16](#_Toc188115059)

[3.3. Phân tích phân phối giá theo diện tích 18](#_Toc188115060)

[3.4. Phân tích mối quan hệ giữa diện tích(m2 ), số phòng ngủ và số phòng tắm 19](#_Toc188115061)

[3.5. Phân tích xu hướng giá cả theo thời gian 20](#_Toc188115062)

[3.6. Phân tích bất động sản theo quận 21](#_Toc188115063)

[4. Nhận xét kết quả 21](#_Toc188115064)

[4.1. Thu thập dữ liệu 21](#_Toc188115065)

[4.2. Làm sạch dữ liệu 21](#_Toc188115066)

[4.3. Trực quan hoá dữ liệu 22](#_Toc188115067)

[5. Định hướng phát triển trong tương lai 22](#_Toc188115068)

[6. Bảng phân công nhiệm vụ 22](#_Toc188115069)

# CHƯƠNG 1. GIỚI THIỆU CHUNG

## Giới thiệu bài toán

Trong bối cảnh thị trường bất động sản ngày càng phát triển và chịu ảnh hưởng bởi nhiều yếu tố kinh tế, xã hội, việc xác định và dự đoán giá trị bất động sản trở thành một yếu tố quan trọng đối với người mua, người bán, nhà đầu tư và các tổ chức nghiên cứu thị trường. Giá trị của bất động sản có thể bị tác động bởi nhiều yếu tố khác nhau như khu vực, diện tích, tiện ích, tình trạng của tài sản và các yếu tố vĩ mô.

Bài toán đặt ra là thu thập và phân tích dữ liệu về giá bất động sản tại Hà Nội để tìm hiểu các yếu tố ảnh hưởng đến giá trị của bất động sản. Mục tiêu là cung cấp các thông tin quan trọng hỗ trợ quyết định đầu tư, mua bán và nghiên cứu thị trường. Các yếu tố như diện tích, khu vực, số phòng, tiện ích và tình trạng của tài sản sẽ được phân tích để hiểu rõ hơn về tác động của chúng đối với giá trị bất động sản, giúp các nhà đầu tư và người mua bán đưa ra quyết định chính xác hơn.

## Giới hạn đối tượng khách thể, đối tượng nghiên cứu

### 2.1. Đối tượng khách thế

* Đối tượng khách thể của bài toán là những cá nhân hoặc tổ chức tham gia vào thị trường bất động sản tại Hà Nội:
* Người mua bất động sản: Cá nhân, gia đình hoặc nhà đầu tư có nhu cầu mua bất động sản tại Hà Nội.
* Người bán bất động sản: Chủ sở hữu tài sản, các nhà môi giới hoặc công ty kinh doanh bất động sản muốn bán tài sản của mình.
* Nhà đầu tư: Những cá nhân hoặc tổ chức muốn phân tích giá trị và tiềm năng sinh lời của bất động sản để quyết định đầu tư.
* Các tổ chức nghiên cứu thị trường: Các cơ quan, công ty nghiên muốn thu thập dữ liệu để đánh giá thị trường bất động sản tại Hà Nội.

### 2.2. Đối tượng nghiên cứu

* Đối tượng nghiên cứu của bài toán là dữ liệu và các yếu tố liên quan đến bất động sản tại Hà Nội:
* Dữ liệu giá bất động sản
* Đặc điểm của bất động sản: Diện tích, số lượng phòng, tiện ích, tình trạng, chính sách pháp lý
* Yếu tố địa lý và khu vực: Vị trí bất động sản( Quận/huyện )

## Các nghiên cứu liên quan

Một số nghiên cứu liên quan và các bài viết của các tác giả đã nghiên cứu về giá bất động sản hoặc thị trường bất động sản, với các chủ đề tương tự như việc phân tích giá bất động sản tại Hà Nội

### Lượng hóa các nhân tố ảnh hưởng đến giá đất ở

* Nguyễn Thị Vân Anh, Trần Thị Hồng Vân, Nguyễn Thị Thu Hằng :“Kết quả tổng quan các công trình nghiên cứu cho thấy có 5 nhóm nhân tố ảnh hưởng đến giá bất động sản, gồm (1) nhóm nhân tố phản ánh chất lượng sản phẩm; (2) nhóm nhân tố pháp lý; (3) nhóm nhân tố môi trường rủi ro; (4) nhóm nhân tố vĩ mô và (5) nhóm nhân tố phản ánh đặc điểm khách hàng." [1]

### Yếu tố ảnh hưởng đến quyết định mua bất động sản tại thành phố Hạ Long, tỉnh Quảng Ninh

* TS. Vũ Thị Thương, TS. Lê Khánh Cường, Lê Phương Thảo:"Nghiên cứu cho thấy, các yếu tố: Vị trí; Tài chính; Pháp lý; Tâm linh; Cơ sở hạ tầng và Môi trường xung quanh đều tác động đến Quyết định mua bất động sản của người dân trên địa bàn TP. Hạ Long, tỉnh Quảng Ninh." [2]

# References

|  |  |
| --- | --- |
| [1] | Nguyễn Thị Vân Anh, Trần Thị Hồng Vân, Nguyễn Thị Thu Hằng, Lượng hóa các nhân tố ảnh hưởng đến giá đất ở. |
| [2] | T. L. K. C. L. P. T. TS. Vũ Thị Thương, "Yếu tố ảnh hưởng đến quyết định mua bất động sản tại thành phố Hạ Long, tỉnh Quảng Ninh". |

# CHƯƠNG 2. Cơ sở lý thuyết và công cụ sử dụng

## Cơ sở lý thuyết

### Thu thập dữ liệu

* Thu thập dữ liệu là quá trình tìm kiếm, thu thập và tổ chức thông tin từ nhiều nguồn khác nhau, với mục đích phân tích và sử dụng thông tin đó để đưa ra quyết định, giải quyết vấn đề.
* Dự án này thu thập dữ liệu về thị trường bất động sản thông qua việc **khai thác dữ liệu từ trang web bất động sản https://batdongsan.com.vn/nha-dat-ban-ha-noi.** Chúng tôi sử dụng công cụ BeautifulSoup để phân tích nội dung HTML và Selenium để tự động hóa việc thu thập thông tin từ các trang web.

### Làm sạch dữ liệu

* Làm sạch dữ liệu là quá trình xử lý và sửa chữa các lỗi trong dữ liệu, bao gồm loại bỏ các giá trị thiếu, xử lý dữ liệu sai định dạng, loại bỏ dữ liệu ngoại lai hoặc không chính xác, và chuẩn hóa các giá trị.
* Dự liệu sau khi đã thu thập, chúng tôi tiến hành xử lý và sửa chữa các lỗi.Việc làm sạch dữ liệu là cần thiết vì dữ liệu không sạch có thể dẫn đến kết quả phân tích sai lệch và ảnh hưởng đến quá trình ra quyết định. Dữ liệu sạch giúp cải thiện độ tin cậy và hiệu quả của các phân tích.

### Trực quan hoá dữ liệu

* Trực quan hóa dữ liệu là quá trình sử dụng hình ảnh như đồ thị, biểu đồ hoặc bản đồ để trình bày dữ liệu, biến dữ liệu phức tạp hoặc số liệu lớn thành các hình ảnh dễ hiểu, hỗ trợ phân tích và ra quyết định.
* Sau khi làm sạch dữ liệu, chúng tôi sử dụng thư viện **Seaborn** và **Matplotlib** để tạo ra các biểu đồ phục vụ cho việc phân tích và hiểu rõ hơn về dữ liệu, giúp khám phá sâu hơn về các đặc điểm của dữ liệu và hiểu rõ mối quan hệ giữa các yếu tố ảnh hưởng.

## Công nghệ, công cụ sử dụng

### Selenium

* Selenium là một công cụ tự động hóa trình duyệt web phổ biến được sử dụng cho việc kiểm thử tự động trên ứng dụng web và tự động hóa các tác vụ trên trình duyệt web.
* Trong dự án này, chúng tôi sử dụng **Selenium** để tự động hóa trình duyệt (trong dự án này là Chrome).
* Tài liệu tham khảo: [Selenium](https://www.selenium.dev/documentation/)

### BeautifulSoup

* **Beautifulsoup** là một thư viện [Python](https://csc.edu.vn/data-science-machine-learning/Fundamentals-of-Python-Lap-trinh-Python-co-ban_189) được sử dụng để phân tích cú pháp HTML và XML, giúp cho việc trích xuất thông tin từ các trang web trở nên dễ dàng hơn.
* Trong dự án này,chúng tôi dùng **BeautifulSoup** để trích xuất dữ liệu từ mã nguồn HTML của trang web, phân tích và tìm kiếm các phần tử HTML như tiêu đề, mã sản phẩm, ngày tháng, giá cả, diện tích,... dựa trên các thẻ HTML và lớp CSS.
* Tài liệu tham khảo: [BeautifulSoup](https://www.crummy.com/software/BeautifulSoup/bs4/doc/)

### Thư viện Pandas

* **Pandas** là một thư viện mã nguồn mở trong Python, được thiết kế chủ yếu để làm việc với dữ liệu có cấu trúc dạng bảng. Pandas cung cấp các công cụ mạnh mẽ giúp xử lý, phân tích, làm sạch và trực quan hóa dữ liệu dễ dàng hơn.
* Trong dự án này, chúng tôi sử dụng pandas để làm sạch dữ liệu, bao gồm xử lý giá trị thiếu (NaN), chuẩn hóa định dạng, xử lý các giá trị ngoại lai (outliers), và chuyển đổi các dữ liệu không hợp lệ thành định dạng chuẩn.
* Tài liệu tham khảo: [Thư viện Pandas](https://pandas.pydata.org/docs/)

### Thư viện Numpy

* **Numpy** là thư viện toán học mạnh mẽ của Python, được thiết kế để làm việc hiệu quả với ma trận và mảng. Nó cung cấp các hàm toán học tối ưu, giúp xử lý dữ liệu số lớn nhanh chóng và tiết kiệm tài nguyên hơn. NumPy đặc biệt hữu ích khi làm việc với các phép toán ma trận phức tạp, giúp tăng hiệu suất xử lý dữ liệu đáng kể.
* **Trong dự án này**, chúng tôi sử dụng numpy để lọc dữ liệu ngoại lai cột Price thông qua việc áp dụng các kỹ thuật thống kê như tính trung bình, độ lệch chuẩn, hoặc các giá trị ngưỡng (threshold), giúp loại bỏ các giá trị bất thường, cải thiện chất lượng dữ liệu đầu vào và đảm bảo kết quả phân tích chính xác hơn.
* Tài liệu tham khảo: [Thư viện Numpy](https://numpy.org/doc/)

### Thư viện matplotlib

* Matplotlib là thư viện cho phép tạo ra các hình, hỗ trợ vẽ các biểu đồ đơn giản hoặc phức tạp và có khả năng tùy chỉnh cao. Thư viện này rất phổ biến trong khoa học dữ liệu và học máy, giúp chuyển các dữ liệu phức tạp thành đồ họa dễ hiểu, hữu ích cho việc phân tích và trình bày kết quả.
* Tài liệu tham khảo: [Thư viện matplotlib](https://matplotlib.org/stable/index.html)

### Thư viện Seaborn

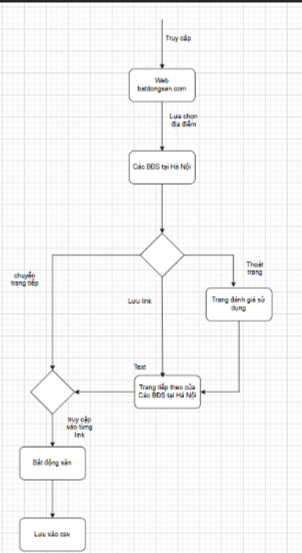
* Seaborn là thư viện Python được xây dựng trên matplotlib, giúp tạo ra các biểu đồ thống kê trực quan đẹp mắt và dễ hiểu. Nó cung cấp các công cụ mạnh mẽ để phân tích dữ liệu.
* Thư viện tham khảo: [Thư viện seaborn](https://seaborn.pydata.org/)

### Jupyter

* Trong dự án này, chúng tôi sử dụng công cụ Jupyter để viết code thu thập dữ liệu, làm sạch, trực quan hoá dữ liệu, phân tích dữ liệu từ dữ liệu đã thu thập được.
* Tài liệu tham khảo:[Jupyter](https://jupyter.org/)

# CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN

## Quy trình thực hiện



* Bước 1: Truy cập vào trang web batdongsan.com.vn.
* Bước 2: Chọn địa điểm bất động sản là Hà Nội
* Bước 3: L**ấy link thông tin bất động sản từ các trang (1 đến 700):** Thu thập các URL hợp lệ từ mỗi trang, loại bỏ quảng cáo và lưu vào mảng.
* Bước 4: Truy cập từng link trong mảng để lấy thông tin của từng bất động sản, sau đó lưu vào file.

# CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM

## Kết quả thu thập dữ liệu

* Sau khi chạy chương trình thu được **12948**  **dòng** và **11 cột** dữ liệu:
* Title: Tiêu đề
* Code: Mã tin
* Date: Ngày đăng
* Price per m2: Giá trị của bất động sản trên m2 tính theo đơn vị trăm triệu
* Total Price: Tổng giá trị bất động sản
* Area: Diện tích của bất động sản
* Bedroom Number: Số phòng ngủ của bất động sản.
* Bathroom Number: Số phòng tắm của bất động sản.
* Legal Status: Giấy tờ pháp lý
* Furniture: Tính trạng nội thất của bất động sản
* District: Quận/ huyện

## Làm sạch dữ liệu

### Lỗi logic

* Có 945 dòng bị nhầm lẫn thông tin giữa 2 cột Price per m2 và Total Price
* Giải phảp: Đổi chỗ 2 giá trị này trong các dòng bị lỗi

### Đổi kiểu dữ liệu

* Đổi kiểu dữ liệu cột ‘Date’ sang datetime.
* Đổi kiểu dữ liệu cột ‘Price(Tỷ)’ sang float
* Đổi kiểu dữ liệu cột ‘Total Price’ sang float
* Đổi kiểu dữ liệu cột ‘Area’ sang float.
* Đổi kiểu dữ liệu cột ‘Bedroom Number’ sang float.
* Đổi kiểu dữ liệu cột ‘Bathroom Number’ sang float.
* Đổi kiểu dữ liệu cột ‘Code’ sang float.

### Xử lý dữ liệu NAN

Nhận xét:

* Cột Code thiếu 4 giá trị
* Cột Date thiếu 1 giá trị
* Cột Price per m2 thiếu 1403 giá trị (có thể bài đăng không có Total Price nên cũng không có Price per m2)
* Cột Total Price (tỷ) thiếu 1402 giá trị (do là giá thoả thuận)
* Cột Area thiếu 11 giá trị
* Cột Bedroom Number thiếu 4326 giá trị
* Cột Bathroom Number thiếu 4995 giá trị
* (Do người đăng bài không cập nhật thông tin số phòng ngủ và phòng tắm)

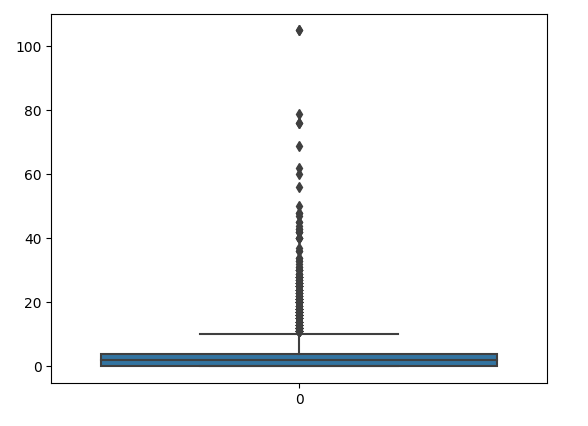
Giải pháp:

* Xoá các dòng bị thiếu giá trị Code và Date do mỗi bài đăng có mã Code riêng
* Thay thế các giá trị bị thiếu bằng giá trị 0

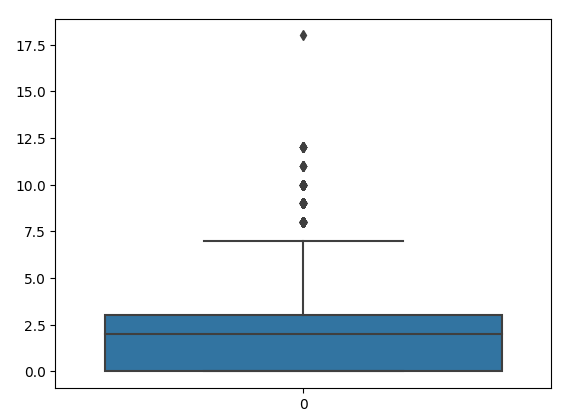
### Xử lý dữ liệu trùng lặp

* Quá trình thu thập dữ liệu có 530 dòng bị lặp lại do đó xóa.

### Xử lí dữ liệu ngoại lai



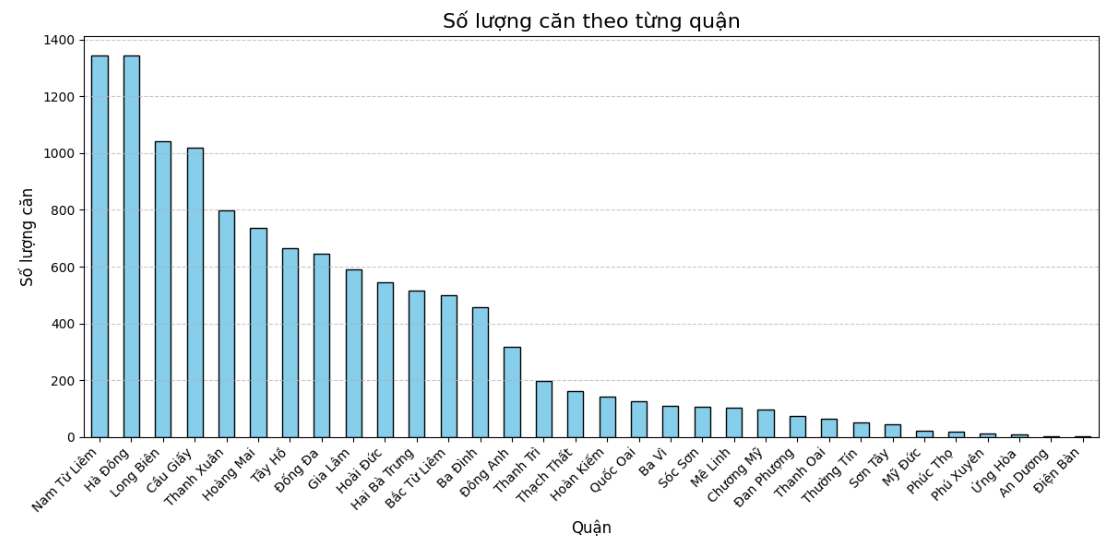
* Loại bỏ ngoại dựa trên cột “Bedroom”
* Sử dụng phương pháp IQR: khoảng cách giữa giá trị thứ 75% (Q3) và giá trị thứ 25% (Q1). Các giá trị nằm ngoài khoảng [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR] thường được coi là ngoại lai.



* Loại bỏ ngoại dựa trên cột “Bathroom”
* Sử dụng phương pháp IQR: khoảng cách giữa giá trị thứ 75% (Q3) và giá trị thứ 25% (Q1). Các giá trị nằm ngoài khoảng [Q1 - 1.5 \* IQR, Q3 + 1.5 \* IQR] thường được coi là ngoại lai.

## Phân tích khai thác dữ liệu

### Phân tích số lượng căn theo từng quận

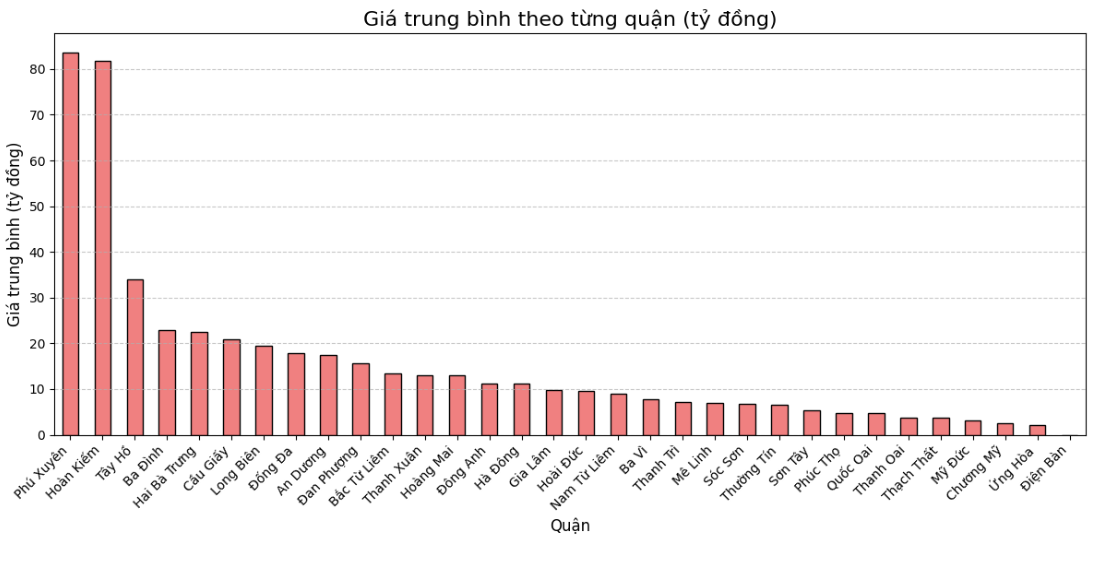


*Hình 1.1. Biểu đồ số lượng căn theo từng quận*

- Nhận xét:

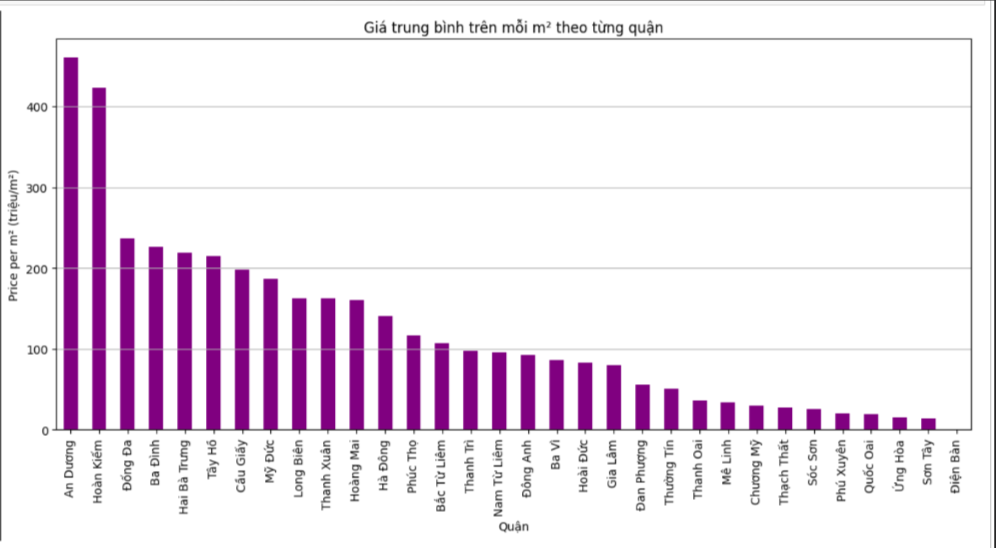
* Nam Từ Liêm, Hà Đông, và Long Biên là các quận dẫn đầu với số lượng căn cao nhất, mỗi quận vượt hơn 1000 căn.Điều này cho thấy các quận này có tốc độ phát triển đô thị mạnh mẽ, có thể là khu vực tập trung nhiều dự án bất động sản.
* Các quận như Cầu Giấy, Thanh Xuân, Hoàng Mai, Tây Hồ, và Đống Đa có số lượng căn nằm ở mức trung bình (từ 500 đến 1000 căn).Đây là những quận có quy hoạch lâu đời hoặc phát triển vừa phải so với nhóm quận dẫn đầu.
* Hoàn Kiếm, Sơn Tây, Chương Mỹ, Đan Phượng, và các quận khác thuộc nhóm cuối có số lượng căn rất ít.Nguyên nhân có thể do diện tích quận nhỏ (như Hoàn Kiếm) hoặc mức độ phát triển đô thị chưa cao (như các quận ngoại thành như Chương Mỹ).
* Biểu đồ thể hiện sự phân bổ không đồng đều về số lượng căn theo từng quận, với sự tập trung lớn ở một số quận trung tâm hoặc quận mới phát triển mạnh.Điều này phản ánh sự bất đối xứng trong phát triển đô thị giữa các khu vực nội thành và ngoại thành Hà Nội.

### Phân tích giá trung bình theo từng quận



*Hình 1.2. Biểu đồ thể hiện giá trung bình theo từng quận*

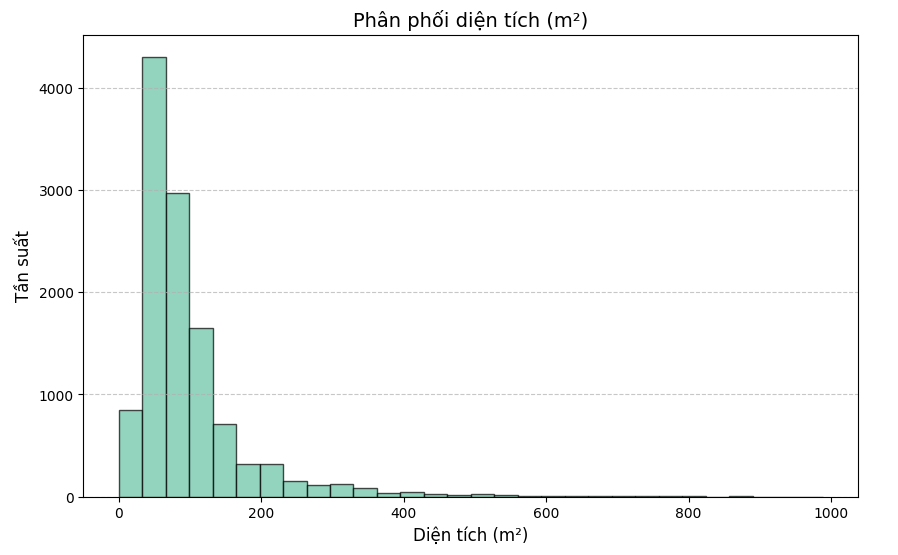
* Nhận xét:
* Phú Xuyên và Hoàn Kiếm dẫn đầu với giá trung bình cao vượt trội, trên 80 tỷ đồng.Điều này có thể phản ánh các giao dịch bất động sản giá trị cao tại các khu vực này, đặc biệt là Hoàn Kiếm – trung tâm của Hà Nội với bất động sản giá trị lịch sử và thương mại.
* Các quận như Tây Hồ, Ba Đình, Hai Bà Trưng, và Cầu Giấy có mức giá trung bình trong khoảng từ 20 tỷ đến 40 tỷ đồng.Đây là các khu vực có giá trị đất cao nhờ vị trí đắc địa và hạ tầng phát triển, thu hút nhiều nhà đầu tư bất động sản.
* Các quận như Thạch Thất, Mỹ Đức, Chương Mỹ, và các quận ngoại thành khác có mức giá trung bình dưới 5 tỷ đồng. Điều này phản ánh sự chênh lệch giữa giá trị bất động sản ở khu vực trung tâm so với vùng ven và ngoại thành.
* Biểu đồ cho thấy sự chênh lệch rất lớn giữa các quận trung tâm (Hoàn Kiếm, Tây Hồ) và các quận ngoại thành.Các quận trung tâm tập trung các bất động sản cao cấp, trong khi ngoại thành có xu hướng giao dịch các bất động sản giá thấp hơn.



*Hình 1.3. Biểu đồ thể hiện giá trung bình trên mỗi m2 theo quận*

* Nhận xét:
* An Dương và Hoàn Kiếm dẫn đầu với giá trung bình trên mỗi m² cao nhất, vượt mốc 400 triệu đồng/m². Điều này cho thấy đây là khu vực có giá trị bất động sản cao cấp hoặc nằm tại các vị trí trung tâm, đắc địa.
* Các quận trung tâm khác như Đống Đa, Ba Đình, Hai Bà Trưng, và Tây Hồ cũng có giá trung bình cao, dao động từ 200 - 300 triệu đồng/m². Những quận này thường là khu vực dân cư sầm uất hoặc có hạ tầng phát triển tốt.
* Các quận như Thạch Thất, Sơn Tây, Phú Xuyên, và Ứng Hòa có giá trung bình trên mỗi m² thấp nhất, dưới 50 triệu đồng/m². Điều này có thể liên quan đến vị trí xa trung tâm và hạ tầng chưa phát triển mạnh.
* Biểu đồ thể hiện rõ sự chênh lệch giá trị bất động sản giữa khu vực trung tâm (như Hoàn Kiếm, Ba Đình) và các vùng ngoại thành.
* Các quận có sự phát triển nhanh về hạ tầng hoặc gần trung tâm (như Cầu Giấy, Long Biên, Thanh Xuân) có giá trị ở mức trung bình khá, từ 100 - 200 triệu đồng/m².

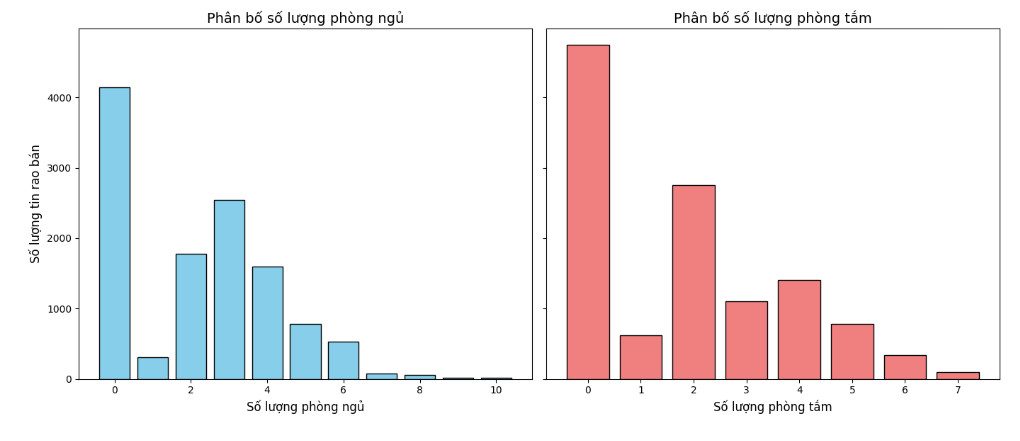
### Phân tích phân phối giá theo diện tích



*Hình 1.4. Biểu đồ thể hiện phân phối giá theo diện tích*

* Phần lớn các căn nhà/tin rao bán có diện tích dưới 100 m², tập trung chủ yếu trong khoảng từ 0 đến 50 m². Đây có thể là các căn hộ nhỏ hoặc nhà ở diện tích hẹp.
* Biểu đồ có dạng lệch phải, tức là phần lớn dữ liệu nằm ở mức diện tích nhỏ, trong khi một số ít tin rao bán có diện tích rất lớn, vượt trên 400 m², nhưng số lượng này không đáng kể.
* Có những căn/tin rao bán diện tích lớn hơn 600 m², nhưng tần suất cực kỳ thấp. Điều này cho thấy thị trường chủ yếu tập trung vào các bất động sản diện tích trung bình hoặc nhỏ

### Phân tích mối quan hệ giữa diện tích(m2 ), số phòng ngủ và số phòng tắm



*Hình 1.5. Biểu đồ thể hiện mối quan hệ giữa diện tích (m²), số phòng ngủ và số phòng tắm*

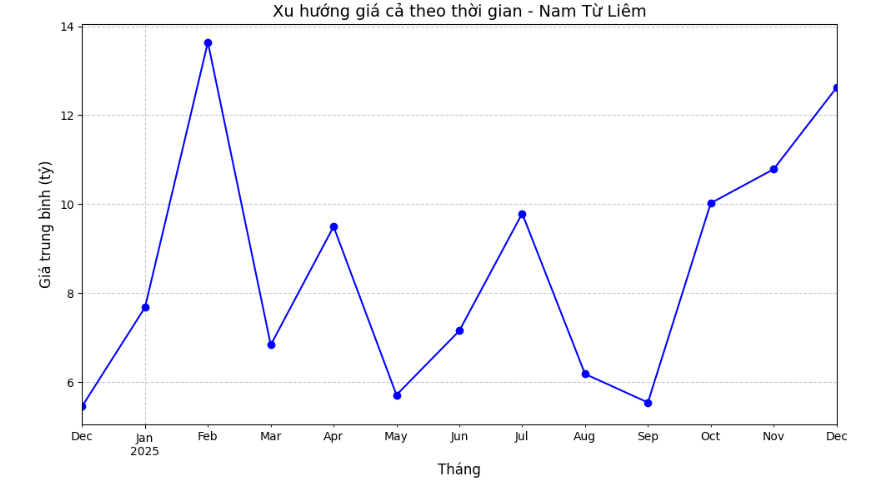
1. Phân bố số lượng phòng ngủ:

* Tập trung chính ở số lượng nhỏ
* Số lượng tin rao bán có 2 phòng ngủ chiếm đa số, tiếp theo là 3 phòng ngủ.
* Rất ít tin rao có số lượng phòng ngủ từ 6 trở lên.
* Ít tin rao với số lượng phòng ngủ lớn
* Các căn nhà/tin rao có từ 8-10 phòng ngủ gần như không phổ biến, có thể là bất động sản đặc thù (biệt thự hoặc nhà lớn).

1. Phân bố số lượng phòng tắm:

* Phòng tắm ít hơn phòng ngủ:
* Phổ biến nhất là các bất động sản có 1 hoặc 2 phòng tắm.
* Số lượng phòng tắm lớn hơn (5-7 phòng) rất hiếm gặp, thể hiện sự tương đồng với xu hướng phòng ngủ.
* Các căn có số lượng phòng ngủ hoặc phòng tắm lớn có thể đại diện cho các phân khúc nhà cao cấp hoặc biệt thự.
* Trong khi đó, các căn có ít phòng ngủ/phòng tắm phù hợp với các phân khúc nhà ở phổ thông.

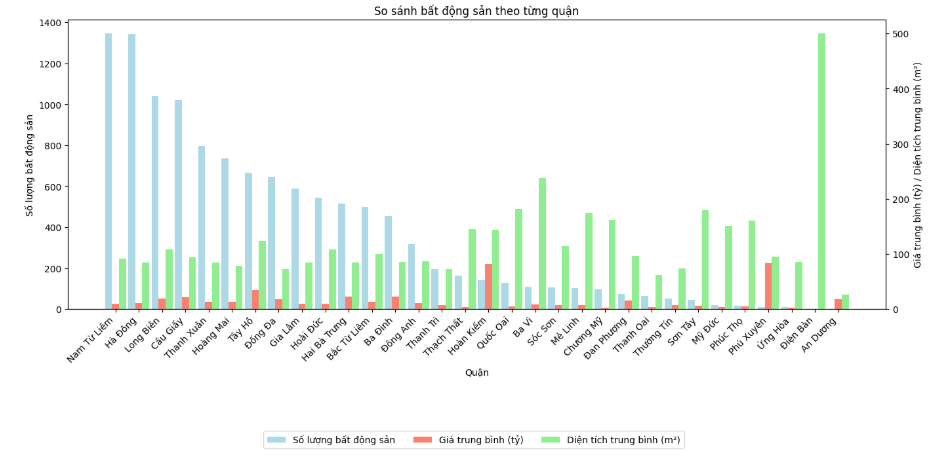
### Phân tích xu hướng giá cả theo thời gian



*Hình 1.6. Biểu đồ thể hiện xu hướng giá cả theo thời gian*

* Nhận xét:
* Giá trung bình có sự biến động đáng kể qua các tháng.
* Tháng 2 đạt đỉnh cao nhất với giá trung bình xấp xỉ 14 tỷ đồng.
* Sau tháng 2, giá giảm mạnh vào tháng 3 và tháng 5 xuống dưới 6 tỷ đồng.
* Từ tháng 9, giá trung bình tăng mạnh trở lại, đặc biệt từ tháng 10 đến tháng 12.
* Biểu đồ cho thấy khả năng giá cả chịu ảnh hưởng bởi mùa vụ hoặc các yếu tố đặc biệt trong từng thời kỳ.
* Giai đoạn cuối năm (tháng 10 đến tháng 12) giá tăng ổn định, có thể liên quan đến nhu cầu mua bán nhà đất cao hơn.

### Phân tích bất động sản theo quận



*Hình 1.7. Biểu đồ so sánh bất động sản theo từng quận*

* Nhận xét:
* Quận Nam Từ Liêm và Hà Đông có số lượng bất động sản lớn nhất, thể hiện qua chiều cao vượt trội của cột xanh nhạt.
* Giá trung bình tại các quận nội thành như Ba Đình, Tây Hồ, và Hoàn Kiếm cao hơn so với các quận ngoại thành, mặc dù số lượng bất động sản ít hơn.
* Diện tích trung bình tại các quận ngoại thành (như Ba Vì, Sóc Sơn, Quốc Oai) thường lớn hơn so với các quận nội thành. Điều này phản ánh đặc điểm phân bố bất động sản ở vùng nội và ngoại thành.

## Nhận xét kết quả

### Thu thập dữ liệu

* Ưu điểm: Dữ liệu thuộc tính các cột đa dạng, quá trình thu thập dữ liệu không xảy ra lỗi
* Nhược điểm: Giới hạn thu thập ở Hà Nội do hạn chế về thiết bị yếu

### Làm sạch dữ liệu

* Ưu điểm: xử lý tối ưu hết các lỗi logic trong dữ liệu như xử lý giá trị thiếu (NaN), chuẩn hóa định dạng, xử lý các giá trị ngoại lai (outliers), và chuyển đổi các dữ liệu không hợp lệ thành định dạng chuẩn.
* Nhược điểm: có phần điền giá trị NaN phòng ngủ và tắm chưa tối ưu.

### Trực quan hoá dữ liệu

* Ưu điểm: Sử dụng đồ thị, biểu để trình bày dữ liệu, biến dữ liệu phức tạp,số liệu lớn thành các biểu đồ dễ hiểu, hỗ trợ phân tích và ra quyết định. Tạo sự trực cho dữ liệu phức tạp.
* Nhược điểm: cần phân tích thêm

## Định hướng phát triển trong tương lai

* Lấy thêm được nhiều dữ liệu hơn.
* Mở rộng phạm vi
* Áp dụng mô hình học máy để có thể dự đoán được giá bất động sản.

## Bảng phân công nhiệm vụ

|  |  |  |
| --- | --- | --- |
| STT | Người thực hiện | Công việc |
| 1 | Nguyễn Khánh Linh | Thu thập dữ liệu, làm báo cáo |
| 2 | Khương Văn Cương | Làm sạch dữ liệu, phân tích dữ liệu, powerpoint |