ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Khoa Công nghệ Thông tin

TRỰC QUAN HÓA DỮ LIỆU

ĐỒ ÁN CUỐI KỲ

Hồ Chí Minh - 3/2024

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

1 Giới thiệu

1.1 Chủ đề

Hãy phân tích và trực quan hoá các dữ liệu liên quan đến đất nước Việt Nam.

Chủ đề lựa chọn: Phân tích và trực quan số liệu nhà ở tại thủ đô Hà Nội.

1.2 Tập dữ liệu

Dữ liệu được lấy từ kho dữ liệu trực tuyến mã nguồn mở **Kaggle**, truy cập bộ dữ liệu tại đây.

Dữ liệu gốc gồm có 12 thuộc tính bao gồm: 6 thuộc tính định tính và 6 thuộc tính định lượng như sau:

- Ngày: Ngày ghi nhận về thông tin căn nhà này.
- Địa chỉ: địa chỉ của căn nhà này tại thủ đô Hà Nội.
- Quận: Quận (hoặc ngang cấp quận) mà căn nhà này toạ lạc.
- Huyện: Huyện (hoặc ngang cấp Huyện) mà căn nhà này toạ lạc.
- Loại hình nhà ở: Thông tin về loại hình mà căn nhà này thuộc về.
- Giấy tờ pháp lý: Thông tin về giấy tờ hợp pháp của căn nhà này.
- Số tầng: Số tầng của căn nhà này.
- Số phòng ngủ: Số phòng ngủ của căn nhà này.
- Diện tích: thông tin về diện tích của căn nhà.
- Dài: thông tin về chiều dài của căn nhà.
- Rộng: thông tin về chiều rộng của căn nhà.
- Giá: Thông tin về giá tiền trên 1 m^2 của căn nhà.

1.3 Tiền xử lý dữ liệu

Loại bỏ chữ: Ở trong các thuộc tính định lượng như chiều dài rộng, giá, và diện tích. Việc loại bỏ các đơn vị đo lường (triệu, đồng, m, m^2) nhằm dễ dàng hơn trong việc trực quan hoá trên nền tàng tableu cũng như tinh toán về sau.

Xử lý trùng lắp: Trong tập dữ liệu gốc có tới hơn 80.000 dòng dữ liệu nên việc trùng lặp là không thể tránh khỏi, chính vì vậy, việc xử lý trùng lắp cũng khiến cho việc tính toán trở nên công bằng và thực tế hơn.

Loại bỏ cột không cần thiết: Có 1 cột "Unnamed: 0" không có giá trị trong dữ liệu cũng

như bài phân tích và trực quan này nên cần được loại bỏ để dữ liệu sạch hơn.

Loại bỏ giá trị rỗng: Các giá trị bị thiếu hay không được cung cấp sẽ bị loại bỏ để thuận tiện hơn trong quá trình trực quan dữ liệu trở nên mượt mà hơn.

Loại bỏ nhiễu: Dùng phương pháp 1.5 IQR để loại bỏ các giá trị nhiễu để tập trung vào các giá trị phổ biến làm các biểu đổ trực quan dễ quan sát cũng như cung cấp được nhiều thông tin hơn.

Loại bỏ dữ liệu sai quy định pháp luật: Các dữ liệu nhà ngõ, hẻm có số tầng lớn hơn 4; nhà phố liền kề có số tầng lớn hơn 6; nhà biệt thự có số tầng lớn hơn 3; nhà mặt phố mặt tiền có số tầng lớn hơn 8 sẽ bị loại bỏ khỏi bộ dữ liệu để phù hợp với quy định của nhà nước.

2 Các thông số thống kê

2.1 Thống kê các biến định lượng

2.1.1 Bảng biểu thống kê

	count	mean	std	min	25%	50%	75%	max
Số tầng	5751.0	3.867849	0.954401	1.0	4.00	4.00	4.00	8.00
Số phòng ngủ	5751.0	3.549470	1.203883	1.0	3.00	4.00	4.00	11.00
Diện tích	5751.0	41.890871	12.927284	4.0	33.00	40.00	50.00	86.00
Dài	5751.0	10.011707	2.682718	3.0	8.00	10.00	12.00	17.80
Rộng	5751.0	4.115239	0.888052	2.0	3.50	4.00	4.80	7.00
Giá/m2	5751.0	86.179467	32.550020	0.0	64.52	81.25	103.03	182.69

Hình 2.1 Bảng thống kê các thuộc tính định lượng

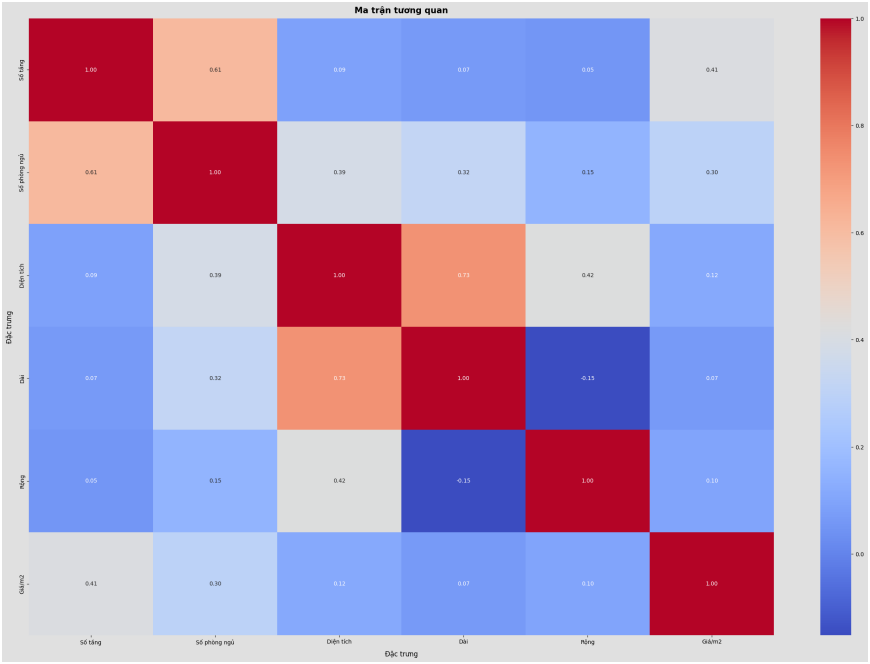
2.1.2 Các nhận xét

Nhận xét về thống kê chung:

- Thông số **Số tầng**: Biểu đồ này cho thấy số tầng của các tòa nhà được khảo sát. Phần lớn các tòa nhà có từ 4 đến 6 tầng, với số lượng tòa nhà 5 tầng là nhiều nhất. Điều này có thể cho thấy một xu hướng chung trong thiết kế xây dựng tại khu vực này, ưu tiên các tòa nhà vừa và nhỏ.

- Thông số **diện tích**:Biểu đồ này thể hiện sự phân bố diện tích của các căn nhà hoặc đất được khảo sát. Dữ liệu tập trung chủ yếu trong khoảng từ 20 đến 60m^2, với số lượng lớn nhất là khoảng 40m^2.Điều này có thể cho thấy rằng đa số các căn nhà hoặc mảnh đất ở khu vực khảo sát có diện tích vừa phải, phù hợp với nhu cầu sinh sống của hầu hết các hộ gia đình.

- Thông số **Giá/**m^2: Biểu đồ này hiển thị giá tiền trên mỗi mét vuông, một thông số quan trọng để đánh giá giá trị bất động sản. Sự phân bố tập trung chủ yếu ở khoảng 50 đến 125 triệu đồng/m^2, với điểm cao nhất là khoảng 75 triệu đồng/m^2. Điều này cho thấy giá cả bất động sản trong khu vực này khá đa dạng nhưng phần lớn vẫn ở mức vừa phải.



Hình 2.2 Biểu đồ mức độ tác động giữa thuộc tính định lượng

khoảng tin cậy 95% cho giá trị trung bình của mỗi cột:

	Giá trị chặn dưới	Giá trị chặn trên
Số tầng	3.069950	4.665748
Số phòng ngủ	2.542998	4.555941
Diện tích	31.083392	52.698351
Dài	7.768898	12.254515
Rộng	3.372809	4.857669
Giá/m2	58.966969	113.391965

Hình 2.3 khoảng tin cậy 95% cho giá trị trung bình của mỗi cột:

Nhận xét về tác đông giữa các thuộc tính:
Tương quan mạnh:

- Có tương quan cao giữa "Diện tích" và "Dài" với hệ số là 0.73. Điều này có nghĩa là chiều dài của bất động sản thường ảnh hưởng lớn đến diện tích tổng thể, điều này là dễ hiểu vì diện tích thường được tính bằng chiều dài nhân với chiều rộng.
- "Số tầng" và "Số phòng ngủ" cũng có mối tương quan khá cao (0.61), cho thấy rằng số lượng tầng trong một căn nhà thường có liên quan đến số lượng phòng ngủ.

Tương quan trung bình:

- "Số phòng ngủ" và "Diện tích" có hệ số tương quan là 0.39, cho thấy một mối quan hệ tích cực nhưng không quá mạnh. Điều này có thể do các căn nhà lớn hơn có nhiều phòng ngủ hơn nhưng không phải lúc nào cũng vậy.

- "Dài" và "Đặc trưng" có hệ số tương quan là 0.42, cho thấy một mối quan hệ vừa phải giữa đặc trưng này và kích thước chiều dài của bất động sản.

Tương quan thấp hoặc không đáng kể:

- Nhiều cặp biến như "Số tầng" với "Diện tích" hoặc "Dài" với "Số phòng ngủ" cho thấy hệ số tương quan thấp (dưới 0.15), cho thấy không có mối quan hệ rõ ràng hoặc mạnh mẽ giữa các biến này.
- "Giá/m2" có tương quan thấp với hầu hết các đặc trưng khác, cho thấy giá trên mỗi mét vuông không phụ thuộc nhiều vào kích thước hoặc cấu trúc của bất động sản.

2.2 Phân tích về các biến định tính

2.2.1 Bảng biểu thống kê

	num_values	value_ratios
Ngày	79	{'2020-08-04': 2.7, '2020-07-24': 2.6, '2020-0...
Địa chỉ	2512	{'Đường Khương Trung, Phường Khương Trung, Quậ...
Quận	22	{'Hà Đông': 19.5, 'Đống Đa': 12.5, 'Thanh Xuân...
Huyện	202	{'Phường Yên Nghĩa': 3.9, 'Phường Khương Trung...
Loại hình nhà ở	4	{'Nhà ngõ, hẻm': 66.5, 'Nhà mặt phố, mặt tiền'...
Giấy tờ pháp lý	4	{'Đã có sổ': 97.8, 'nan': 0.8, 'Đang chờ sổ': ...
Location	22	{'Hà Nội, Quận Hà Đông': 19.5, 'Hà Nội, Quận Đ...

Hình 2.4 Bảng thống kê các thuộc tính định tính

2.2.2 Các nhận xét

Nhận xét về thống kê chung:

- Thông số **Loại hình nhà ở**: Biểu đồ này thể hiện sự phân bố số lượng theo loại hình nhà ở. Phần lớn các tài sản là "Nhà riêng lẻ", tiếp theo là "Nhà liền kề" với số lượng đáng kể thấp hơn, và cuối cùng là "Nhà biệt thự", "Nhà tập thể" có số lượng ít nhất. Sự chênh lệch lớn giữa các loại hình nhà ở cho thấy phong cách và lựa chọn ở của người dân trong khu vực. Điều này có thể phản ánh nhu cầu và khả năng kinh tế của cư dân, cũng như xu hướng phát triển đô thị tại các khu vực đó. Những thông tin này rất quan trọng trong việc phân tích thị trường bất động sản và lập kế hoạch đô thị.
- Thông số **Giấy tờ pháp lý**: Biểu đồ này thể hiện số lượng giấy tờ pháp lý của các nhà, đất được phân theo loại. Sự chênh lệch lớn giữa những tài sản có giấy tờ đầy đủ và những tài sản không có hoặc chỉ có sổ tạm, cho thấy mức độ pháp lý của bất động sản tại khu vực này. Tình trạng này có thể ảnh hưởng lớn đến giá trị và khả năng giao dịch của bất động sản.

- Thông số **Quận**: Biểu đồ này cho thấy sự phân bố số lượng theo các quận trong thành phố. Quận 1 và Quận 7 nổi bật với số lượng cao nhất, cho thấy sự tập trung dân cư hoặc hoạt động thương mại đáng kể trong các khu vực này. Điều này có thể phản ánh mức độ phát triển và sự ưu tiên trong đầu tư cơ sở hạ tầng và dịch vụ.