**MÔ TẢ BỘ DỮ LIỆU**

1. **Bộ dữ liệu**

Vì dữ liệu của chúng tôi tự thu thập dựa trên các tin được đăng tại trang ***batdongsan.com.vn***, do đó chúng tôi sẽ cần phải đánh giá việc tin được đăng lên có phải là tin ảo hay không. Chúng tôi sẽ chia dữ liệu dựa trên 2 khía cạnh, thông tin bài đăng và thông tin bất động sản.

1. **Các cột dữ liệu:**

**Nhóm: Thông tin bài đăng**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thuộc tính** | **Định nghĩa** | **Kiểu dữ liệu** | **Ví dụ** |
| *post\_id* | Mã số bài đăng | int64 | 38205009 |
| *post\_title* | Nội dung bài đăng | object | Cho thuê căn 88m2 (3PN) giá 9.5tr - căn 2PN giá….. |
| *post\_verified* | Tin được xác thực bởi quản trị viên | int64 | 0 |
| *post\_type* | Loại bài đăng (Tin vip Kim cương, Vàng, Bạc; Tin thường) | object | vip-diamond |
| *upload\_date* | Thời gian đăng bài | object | 06/10/2023 |
| *owner\_id* | Mã số người đăng bài | int64 | 2524547 |
| *owner\_name* | Tên người đăng bài | object | Thiên Phúc |

**Nhóm: Thông tin căn hộ**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thuộc tính** | **Định nghĩa** | **Kiểu dữ liệu** | **Ví dụ** |
| *building\_type* | Loại bất động sản (Chung cư căn hộ, Nhà trọ Phòng trọ, Nhà riêng) | object | Căn hộ chung cư |
| *squares* | Diện tích | object | 88 m² |
| *nums\_bedroom* | Số phòng ngủ | object | 3 phòng |
| *nums\_wc* | Số nhà vệ sinh | object | 2 phòng |
| *nums\_floor* | Số tầng (Chỉ dành cho nhà riêng) | object | 2 tầng |
| *path\_size* | Độ rộng đường trước nhà (Chỉ dành cho nhà riêng) | object | 3 m |
| *front\_size* | Độ rộng mặt tiền (Chỉ dành cho nhà riêng) | object | 2 m |
| *building\_direction* | Hướng cửa chính | object | Tây - Nam |
| *window\_direction* | Hướng cửa số | object | Đông - Nam |
| *furniture* | Nội thất trong nhà | object | Đầy đủ |
| *legal* | Giấy tờ | object | Hợp đồng mua bán |
| *others* | Thông tin khác | float64 |  |
| *location* | Vị trí căn hộ | object | Dự án The Western Capital, Đường Lý Chiêu Hoàng, Phường 10, Quận 6, Hồ Chí Minh |
| ***price*** | Giá tiền (Mục tiêu cần dự đoán) | object | 9,5 triệu/tháng |

* Tổng số trường thông tin chúng tôi có được là 21 và 1740 dòng dữ liệu. Trong đó, có 4 biến số và 17 biến phân loại.

1. **Đánh giá tổng quan:**

Nhóm chúng tôi sẽ đánh giá trên 2 khía cạnh: thực tế và theo bộ dữ liệu.

* **Thực tế:**
* Thực tế, giá nhà sẽ phụ thuộc vào các thông tin liên quan các thông số cơ bản của 1 căn hộ/1 căn nhà như: *squares, nums\_bedroom, nums\_wc, nums\_floor, path\_size, front\_size, building\_direction, window\_direction, furniture.* Do đó, chúng tôi sẽ tập trung vào các dữ liệu này.
* Ngoài ra, có thể thấy giá nhà còn phụ thuộc vào vị trí địa lý: nhà ở Quận 1 sẽ có giá cao hơn nhà ở Quận 9, đối với đường cũng sẽ có ảnh hưởng tương tự.
* Đối với riêng chung cư, chúng tôi nhận thấy giá nhà còn phụ thuộc vào dự án, các mức dự án chia ra còn có: dự án cao cấp, trung bình và nhà ở xã hội.
* Đối với loại hình, chúng tôi tạm chia thành 3 nhóm: chung cư, nhà trọ/phòng trọ, nhà riêng. Thông thường, phòng trọ là dạng phòng dịch vụ, thường nằm trong 1 căn nhà 5-6 tầng, diện tích nhỏ và thường không chia phòng rõ ràng khác biệt so với chung cư. Tương tự, nhà trọ cũng thuộc loại nhà 1 phòng, có gác lửng mà không chia tầng rõ ràng, diện tích nhỏ, khác biệt với nhà riêng. Vì lý do đó, nên chúng tôi vẫn giữ 3 nhóm thay vì gộp lại làm 1.
* **Bộ dữ liệu**
* Giá nhà phụ thuộc khá nhiều vào môi giới/chủ nhà. Cùng 1 kiểu căn, ở địa chỉ gần nhau, nếu chủ nhà cho thuê, sẽ có giá rẻ hơn so với thuê qua môi giới. Qua đó có thể thấy giá nhà cũng ảnh hưởng bởi người đăng bài.
* Tuy nhiên, với các thông số *nums\_floor, path\_size, front\_size* chỉ có với nhà riêng, do đó, chúng tôi sẽ đánh giá mức độ phụ thuộc của giá đối với các thông số này để xác định xem liệu bộ dữ liệu có bị mất cân bằng dữ liệu hay khi các dự án chung cư không có hay không?
* Trường location bao gồm khá nhiều thông tin cần thiết và có khả năng ảnh hưởng đến giá nhà (dự đoán), chúng tôi sẽ xử lý bằng cách tách thành các trường thông tin mới như project, streets, ward, district.

1. **Tiền xử lý dữ liệu cơ bản**
2. Loại bỏ cột ‘other’: Đây là cột rỗng, chúng tôi sẽ tiến hành loại bỏ
3. Loại bỏ các dòng bị trùng lặp
4. Loại bỏ các dòng có cột ‘price’ có giá trị là thỏa thuận
5. Quy đổi giá trị price thành dạng số và có đơn vị là triệu/tháng
6. Tách cột ‘location’ thành các cột project, streets, ward, district

***Kết quả thu được sau khi tiền xử lý dữ liệu cơ bản:***

**Nhóm: Thông tin bài đăng**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thuộc tính** | **Định nghĩa** | **Kiểu dữ liệu** | **Ví dụ** |
| *post\_id* | Mã số bài đăng | int64 | 38205009 |
| *post\_title* | Nội dung bài đăng | object | Cho thuê căn 88m2 (3PN) giá 9.5tr - căn 2PN giá….. |
| *post\_verified* | Tin được xác thực bởi quản trị viên | int64 | 0 |
| *post\_type* | Loại bài đăng (Tin vip Kim cương, Vàng, Bạc; Tin thường) | object | vip-diamond |
| *upload\_date* | Thời gian đăng bài | object | 06/10/2023 |
| *owner\_id* | Mã số người đăng bài | int64 | 2524547 |
| *owner\_name* | Tên người đăng bài | object | Thiên Phúc |

**Nhóm: Thông tin căn hộ**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thuộc tính** | **Định nghĩa** | **Kiểu dữ liệu** | **Ví dụ** |
| *building\_type* | Loại bất động sản (Chung cư căn hộ, Nhà trọ Phòng trọ, Nhà riêng) | object | Căn hộ chung cư |
| *squares* | Diện tích | object | 88 m² |
| *nums\_bedroom* | Số phòng ngủ | object | 3 phòng |
| *nums\_wc* | Số nhà vệ sinh | object | 2 phòng |
| *nums\_floor* | Số tầng (Chỉ dành cho nhà riêng) | object | 2 tầng |
| *path\_size* | Độ rộng đường trước nhà (Chỉ dành cho nhà riêng) | object | 3 m |
| *front\_size* | Độ rộng mặt tiền (Chỉ dành cho nhà riêng) | object | 2 m |
| *building\_direction* | Hướng cửa chính | object | Tây - Nam |
| *window\_direction* | Hướng cửa số | object | Đông - Nam |
| *furniture* | Nội thất trong nhà | object | Đầy đủ |
| *legal* | Giấy tờ | object | Hợp đồng mua bán |
| *project\_names* | Tên dự án căn hộ | object | Dự án The Western Capital |
| *streets* | Tên đường | object | Đường Lý Chiêu Hoàng |
| *wards* | Tên phường | object | Phường 10 |
| *districts* | Tên quận | object | 6 |
| ***price*** | Giá tiền (Mục tiêu cần dự đoán) | float64 | 9,5 triệu/tháng |

* Tổng số trường thông tin chúng tôi có được là 22 và 1708 dòng dữ liệu. Trong đó, có 4 biến số và 18 biến phân loại.