Analyze real estate data

Report by: Le Thanh Dat

Contents

Chapter 1: Overview of the project

1.1. Problem

- The project with analyze the house price between three city: Ha Noi, Tp. Ho Chi Minh and Da Nang.
- Understand which features affect the house price
- Predict the house price base on linear regression model
- The data will crawl from website nhatot.com.

1.2. Overview of the house data in nhatot.com

Here is the overciew of the raw data:

	ad_id	list_id	list_time	date	account_id	account_oid	account_name	state	subject	body	category	category_name
0	153400852	112848721	1712249766000	15 phút trước	23093184	9757fd9ebb02faa9e6710e5b721bb42c	Lương Thị Đức Hậu	accepted	Bán nhà dân xây phố Kẻ Tạnh-Việt Hưng-38m2-4tầ...	Cần bán nhà 4 tầng, 38m2, mt 4m, tự xây kiên c...	1020	Nhà ở
1	156221470	115266058	1712249597259	17 phút trước	26816952	18e0aa380bfcfcc741e975d528ba24f7	Nguyễn Lam	accepted	CCMN kinh doanh 33 căn hộ khép kín gác lửng 14...	✅DÒNG TIỀN 2,4 TỶ/ NĂM - DT 146M XD 8 TẦNG TH...	1020	Nhà ở
2	156221612	115266086	1712249495000	19 phút trước	20980923	1c58aacaec12d253e135a4da158ec9cf	Đoàn Sang	accepted	Bán nhà Trần Quốc Vượng, mới tinh, ngõ oto nhỏ	BÁN NHÀ TRẦN QUỐC VƯỢNG – MỚI	1020	Nhà ở

Chapter 2: Data cleaning

2.1. Cleaning the house data

Step 1: Cleaning the unnecessary fields and rename fields

```
df_raw =
df_raw.drop(columns=['pty_characteristics','price_million_per_m2','shop_alias','date' ,'sho
p.alias', 'shop.status', 'shop.name', 'shop.address', 'shop.profileImageUrl',
'shop.createdDate', 'shop.modifiedDate', 'shop.urls', 'shop.shopsCategoriesRelationships',
'ad_id', 'list_id',  'account_id', 'account_oid', 'account_name', 'image', 'webp_image',
'videos', 'number_of_images', 'avatar', 'seller_info.full_name', 'seller_info.avatar',
'seller_info.sold_ads', 'seller_info.live_ads', 'has_video', 'company_logo',
'special_display_images', 'special_display', 'state', 'land_feature', 'street_id',
'size_unit', 'property_back_condition', 'property_road_condition', 'unitnumber_display',
'projectimages', 'project_oid', 'projectid', 'orig_list_time', 'streetnumber_display',
'ad_labels', 'label_campaigns', 'pty_jupiter', 'zero_deposit', 'escrow_can_deposit',
'protection_entitlement', 'owner', 'phone_hidden', 'contain_videos', 'type', 'params',
'address', 'location', 'longitude', 'latitude', 'company_ad', 'category', 'area',
'region','body', 'region_v2', 'area_v2', 'ward'], axis=1)
```

```
df_raw.rename(columns={
    'subject': 'TenBds',
    'category_name':'TheLoai',
```

```
    'area_name':'Huyen',
    'region_name':'Tinh',
    'price':'Gia',
    'rooms':'TongSoPhong',
    'property_legal_document':'GiayTo',
    'ward_name':'Phuong/Xa',
    'price_million_per_m2':'Gia/m2',
    'house_type':'LoaiHinhNhaO',
    'land_type':'LoaiHinhDat',
    'width':'ChieuNgang',
    'length':'ChieuDai',
    'toilets':'PhongVeSinh',
    'floors':'SoTang',
    'furnishing_sell':'NoiThat',
    'living_size':'DienTichSuDung',
    'commercial_type':'LoaiHinhVanPhong',
    'detail_address':'TenDuongCuThe',
    'direction':'HuongCuaChinh',
    'apartment_type':'LoaiHinhCanHo',
    'property_status':'TinhTrangBds',
    'street_number':'SoDuong',
    'block': 'TenPhanKhu/Lo/Block/Thap',
    'floornumber':'TangSo',
    'balconydirection':'HuongBanCong',
    'unitnumber':'MaLo',
    'apartment_feature':'DacDiemCanHoChungCu',
    'size':'DienTich(m2)',
    'street_name':'TenDuong'
    }, inplace=True)
```

Step 2:

Trường	Mô tả
TheLoai	Loại hình bất động sản, ví dụ như nhà ở, đất, căn hộ/Chung cư, văn phòng (Mặt bằng kinh doanh), phòng trọ.
Huyen	Tên huyện vị trí bất động sản.
Tinh	Tên tỉnh vị trí bất động sản.
Gia	Giá của bất động sản tính theo VNĐ.
TongSoPhong	Tổng số phòng trong bất động sản.
GiayTo	Loại giấy tờ pháp lý của bất động sản gồm có: Đã có sổ, đang chờ sổ, giấy tờ khác, hợp đồng đặt cọc, sổ chung, sổ hồng riêng.
DienTich (m^2)	Diện tích của bất động sản tính bằng mét vuông.
LoaiHinhNhaO	Loại hình nhà ở gồm có: Nhà mặt phố, nhà ngõ/hẻm, nhà biệt thự, nhà phố liền kề.
LoaiHinhDat	Loại hình đất gồm có: Đất thổ cư, đất nền dự án, đất nông nghiệp, đất công nghiệp.
ChieuNgang	Chiều ngang của bất động sản.
ChieuDai	Chiều dài của bất động sản.
PhongVeSinh	Số lượng phòng vệ sinh trong bất động sản.
SoTang	Số tầng của bất động sản.
NoiThat	Loại nội thất của bất động sản gồm có: nội thất cao cấp, nội thất đầy đủ, hoàn thiện cơ bản, bàn giao thô.
DienTichSuDung	Diện tích sử dụng của bất động sản.
LoaiHinhVanPhong	Loại hình văn phòng gồm có: Shophouse, officetel, văn phòng, mặt bằng kinh doanh.

TenDuongCuThe	Tên đường cụ thể nơi bất động sản tọa lạc.
HuongCuaChinh	Hướng của cửa chính của bất động sản, như Đông, Tây, Nam, Bắc .
LoaiHinhCanHo	Loại hình căn hộ, như chung cư, căn hộ dịch vụ, penthouse.
TinhTrangBds	Tình trạng bất động sản, như chưa bàn giao, đã bàn giao.
SoDuong	Số đường nơi bất động sản tọa lạc.
TenPhanKhu	Tên phân khu, lô, block, tháp nơi bất động sản tọa lạc.
TangSo	Số tầng cụ thể của căn hộ hoặc bất động sản.
HuongBanCong	Hướng của ban công của bất động sản gồm có: Đông, Tây, Nam, Bắc, Đông Bắc, Đông Nam, Tây Bắc, Tây Nam.

Detail information of the house price data

Step 2: Next, we will remove data with a null value rate greater than 55 percent. This is to improve data quality for analysis.

```
perce_null = df_c.isna().sum()*100/df_c.shape[0]
df_c= df_c.drop(columns=  perce_null[perce_null.values > 55].index )
```

Step 3: Check the outliers and remove them

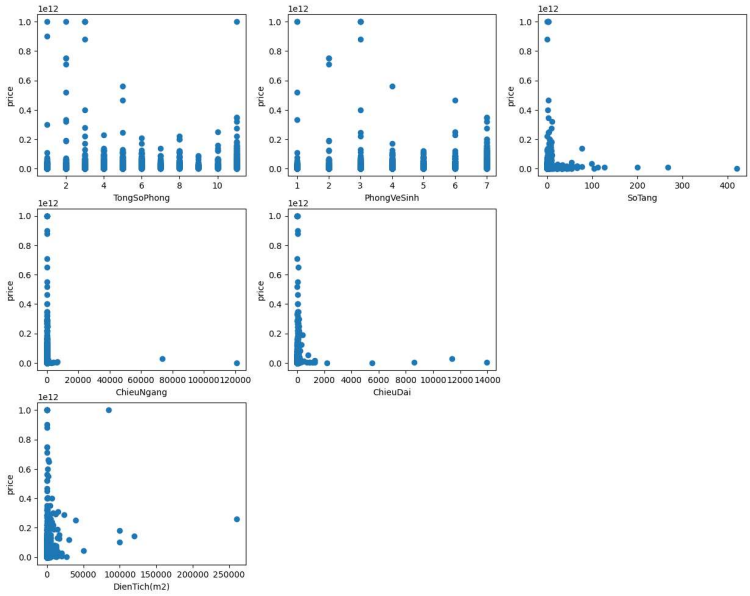
Outlier in the variable 'Gia': The max value (999999999999) is far beyond the average value (7894028777.748519) and even exceeds the highest actual value of real estate. The large gap between the 75th percentile value (7500000000) and the max value indicates the presence of unusual values in the distribution tail.

Outliers in the variables 'DienTich', 'ChieuNgang', 'ChieuDai': The max values of these variables (260000, 120797, 13947) are also very large compared to the average and 75th percentile values. This suggests that there are properties with abnormal sizes compared to most other properties.

Outlier in the variable 'SoTang': Although the max value (421) is not extremely large compared to the average, it still far exceeds the 75th percentile value of 5 and may be considered an outlier.

	Gia	TongSoPhong	DienTich(m2)	ChieuNgang	ChieuDai	PhongVeSinh	SoTang
count	40854.000000	32737.000000	40853.000000	23631.000000	20391.000000	24591.000000	18388.000000
mean	7894028777.748519	3.545804	129.764259	15.464205	19.299993	3.127852	3.701708
std	23489169657.313416	2.047753	1732.622733	922.677118	147.506917	1.598266	5.375369
min	400000.000000	1.000000	1.000000	1.200000	1.000000	1.000000	1.000000
25%	2700000000.000000	2.000000	45.000000	4.000000	11.000000	2.000000	2.000000
50%	4500000000.000000	3.000000	64.000000	4.800000	15.000000	3.000000	3.000000
75%	7500000000.000000	4.000000	95.000000	5.401450	20.000000	4.000000	5.000000
max	999999999999.000000	11.000000	260000.000000	120797.000000	13947.000000	7.000000	421.000000

Overview of data values of data numbers



Visualize digital data relative to price before remove outliers

Step 4: Fill category and numeric missing value

```python
q3_gia=df_c['Gia'].quantile(0.98)
df_c = df_c[(df_c['Gia'] < q3_gia)]
```

```python
q3=df_c['DienTich(m2)'].quantile(0.99)
df_c= df_c[df_c['DienTich(m2)'] <=q3]
```

```python
df_c = df_c[df_c['ChieuNgang'] <= 60000]
```

```python
df_c = df_c[df_c['ChieuDai'] <= 2000]
```

Remove the outliers

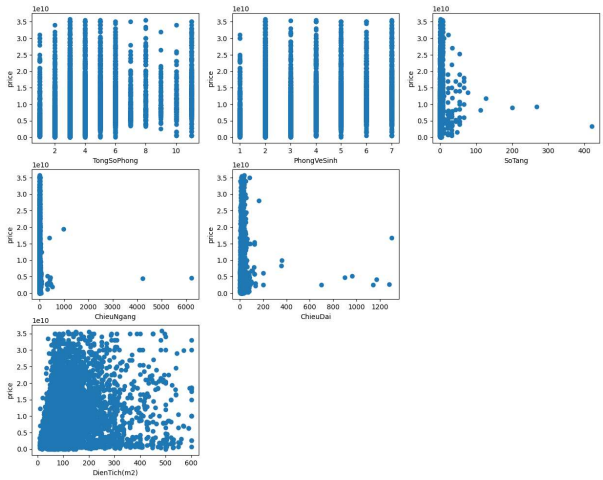
```python
# Fill the numeric missing values with mean values
for column in [
'ChieuNgang',
'ChieuDai',
'DienTich(m2)'
]:
    df_c[column].fillna(value=df_c[column].mean(),inplace=True)
```

```python
# Fill the category null values with mode value
for column in [
'TongSoPhong',
'Huyen',
'GiayTo',
'LoaiHinhNhaO',
'PhongVeSinh',
'NoiThat',
'PhongVeSinh',
'SoTang',

]:
    df_c[column].fillna(value=df_c[column].mode()[0],inplace=True)
```

Fill the missing values

Visuallize the data after cleaning:


1e10
price
TongSoPhong
PhongVeSinh
SoTang
ChieuNgang
ChieuDai
DienTich(m2)

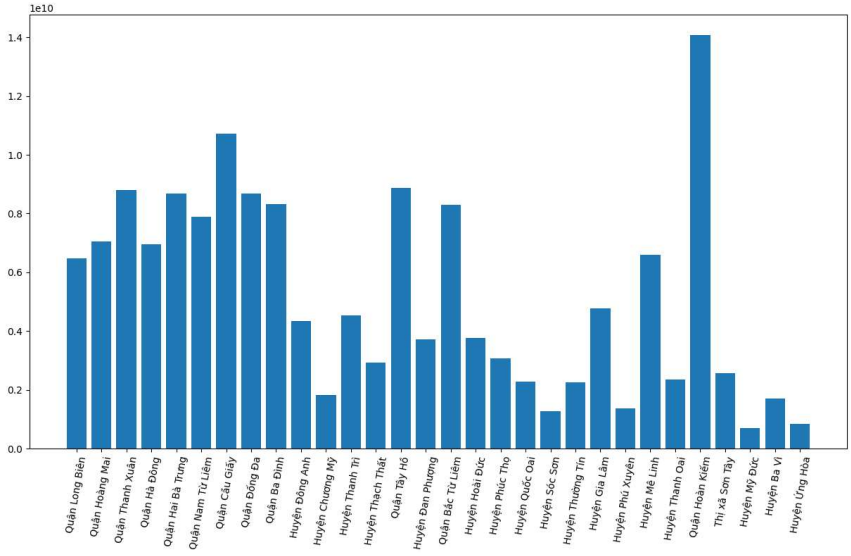
Chapter 3: Analyze the house price

3.1. Analyze the average house price between Ha Noi, Ho Chi Minh, Da Nang

Districts with the highest real estate prices: Hoàn Kiếm District has the highest average price, significantly surpassing other districts. Other central districts such as Tây Hồ, Cầu Giấy, Ba Đình, Đống Đa, and Hai Bà Trưng also have relatively high prices.

Districts with average real estate prices: Districts like Thanh Xuân, Hoàng Mai, and Long Biên have average prices that are neither too high nor too low.

Districts and counties with the lowest real estate prices: Some districts like Chương Mỹ, Quốc Oai, Sóc Sơn, Thường Tín, Phú Xuyên, and Mỹ Đức have lower prices compared to the central districts.

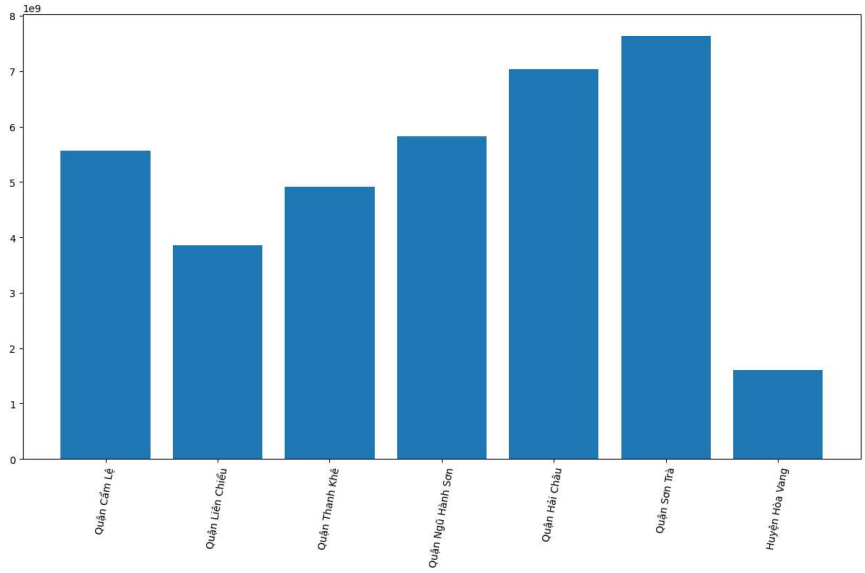


Compare the average house price in Ha Noi

Districts with the highest real estate prices: Hải Châu District has the highest average prices, significantly surpassing other districts. Ngũ Hành Sơn and Sơn Trà Districts also have relatively high prices, indicating these are prime areas in Đà Nẵng.

Districts with average real estate prices: Thanh Khê District and Cẩm Lệ District have average prices that are neither too high nor too low.

Districts and counties with the lowest real estate prices: Liên Chiểu District and Hòa Vang District have the lowest real estate prices.

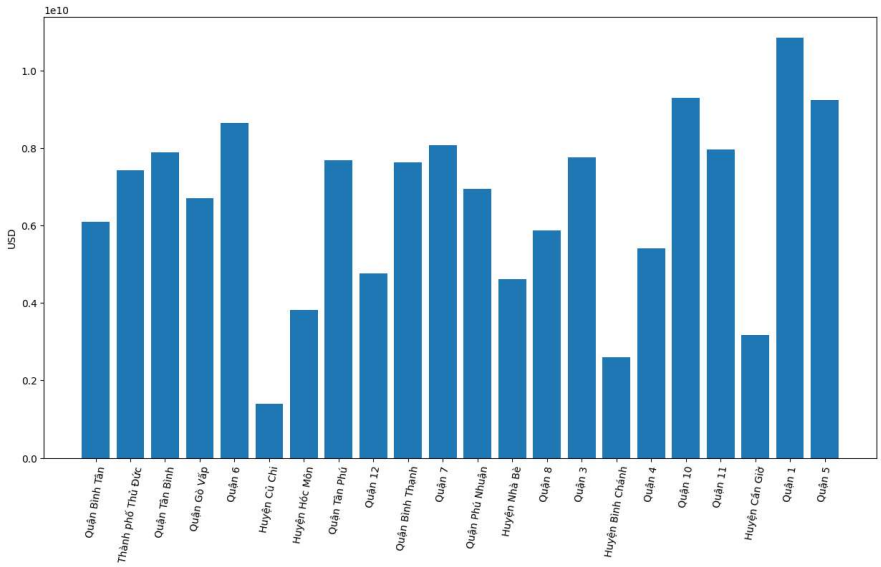


Compare the average house price in the districts of Da Nang

Districts with the highest real estate prices: District 1 has the highest average prices, significantly surpassing other districts. Other central districts like District 3, District 5, and District 10 also have relatively high prices.

Districts with average real estate prices: Districts such as Tân Bình, Bình Thạnh, Gò Vấp, and Tân Phú have average prices that are neither too high nor too low.

Districts and suburban districts with the lowest real estate prices: Suburban districts like Bình Chánh and Củ Chi have the lowest real estate prices.

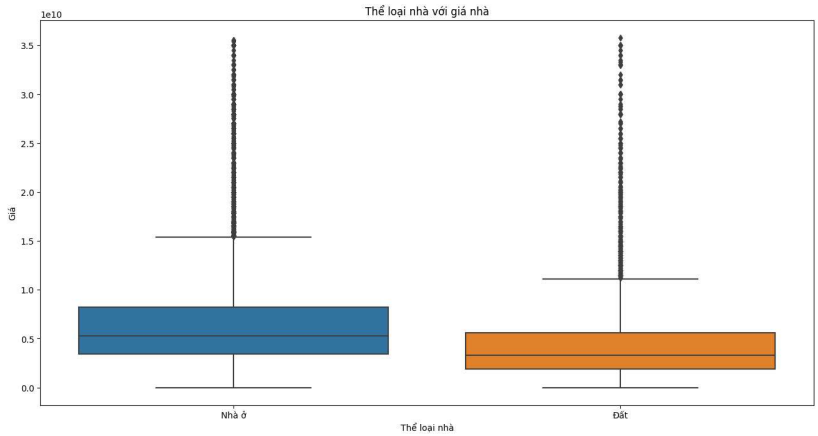


Compare the average house price in the districts of Ho Chi Minh

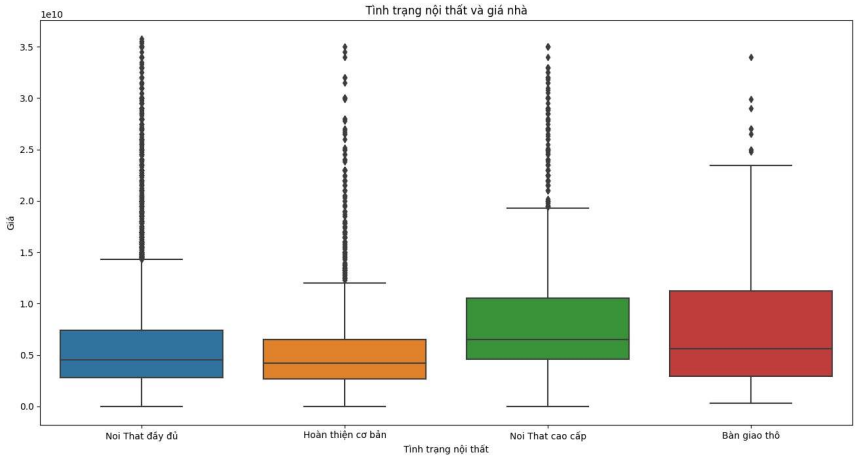
3.2. Analyze the affect of the features to the house price

The median line, Q1, Q3 are all higher than the land type, proving that housing prices are higher than land.

Land prices have greater dispersion, with many values far from the average value. This shows that land prices have greater fluctuations than housing prices.

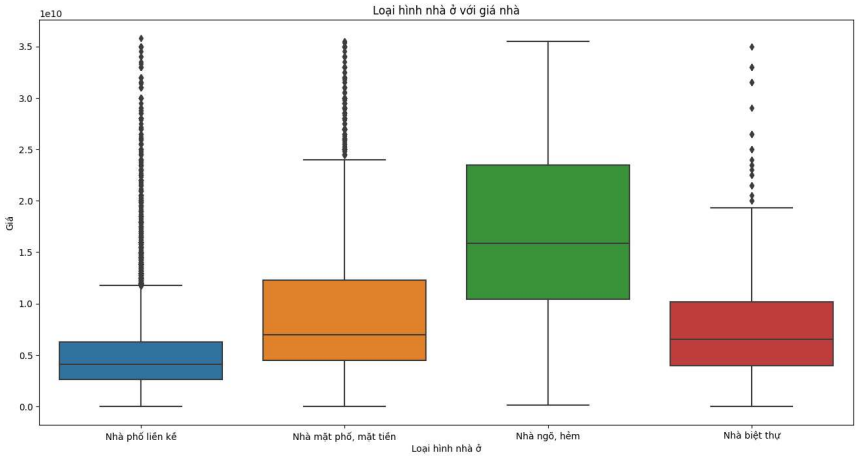


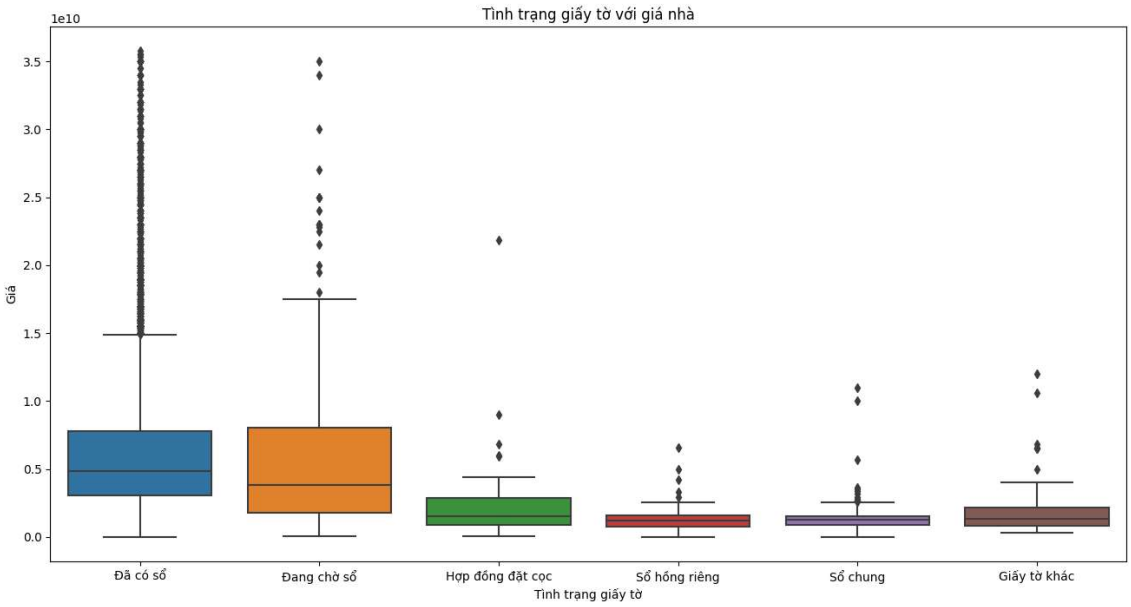
Fully furnished and basic finished interiors have the same average price. High-end interiors have a higher median than the previous two groups and a higher distribution of house prices as well.



“Nha ngo, hem”: Have the highest average price and also the widest price range distribution.

“Nha pho lien ke”: Have the lowest average price and the narrowest price range. This indicates that terraced houses are quite consistent in price and have little variation. However, they contain the most outliers among all four groups.




Tình trạng giấy tờ với giá nhà
1e10
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Giá
Đã có sổ
Đang chờ sổ
Hợp đồng đặt cọc
Sổ hồng riêng
Sổ chung
Giấy tờ khác
Tình trạng giấy tờ

Chapter 4: House prediction using linear regression

4.1. Predicting house price using linear regression

R-squared (R^2): R^2 is 0.58, It is very a low score.

```
from sklearn.metrics import r2_score
from sklearn.metrics import explained_variance_score

print(r2_score(y_test_lr, y_predictlr))
# print(explained_variance_score(y_test_lr, y_predictlr))
```

✓ 0.0s

```
0.5805718602196204
```

Root Mean Squared Error (RMSE): 3 523 026 975.35 . This is the average error of the model compared to the actual values. This value is quite high because house price data has a very large distribution, the lowest is 400,000 to the largest 27.89 billion.

There are 75% number of price has the “% sai lech” <=63 % . There is a big wide distribution in the “% sai lech”, the mean is 120 but the max value can reach at 67970

	Gia du doan	Gia thuc	% sai lech
count	3946.00000000	3946.00000000	3946.00000000
mean	6290685873.89037132	6265153185.91611767	120.23909388
std	4191680958.77650213	5440542698.39478302	1824.36459379
min	-8639878528.16618538	11000000.00000000	0.04330429
25%	3466390143.83381510	2850000000.00000000	14.15389734
50%	5932935807.83381462	4600000000.00000000	31.60129957
75%	8778263167.83381462	7500000000.00000000	63.43176017
max	26895919459.79660797	35800000000.00000000	67970.02808797

High percentage error at low values: For real estate properties with values close to zero, the percentage error can exceed 800%, indicating that the model struggles to accurately predict prices for low to very low-value properties. The high percentage error at low price levels may be due to the model lacking sufficient data for training in this segment or because these values are outliers.

More stability at higher prices: For properties with higher prices, the percentage error fluctuates less and is generally lower. This indicates that the model predicts more effectively when dealing with high-value properties.