Analyze and predict house price

Table of content

Chap 1: Overview of the nhatot.com website

1.1. Problem

nhatot. com is a famous real estate website , provide the features and price of house in VietNam

I will help customer understand the market and invest effective:

- Get understand overview of the price in three major province: Ha Noi, Ho Chi Minh, Da Nang
- Get understand which affect the house price
- Predict the house price base on linear regression model

Chap 2: Extract and handling data

2.1.1 Extract data and cleaning data

I will crawl data through gateway url

Example one row of the raw data:

Column Header	Value
ad_id	153400852
list_id	112848721
list_time	1712249766000
date	15 phút trước
account_id	23093184
account_oid	9757fd9ebb02faa9e6
account_name	Lương Thị Đức Hậu
state	accepted
subject	Bán nhà dân xây ph
body	Cần bán nhà 4 tầng,
category	1020
category_name	Nhà ở
area	81
area_name	Quận Long Biên
region	12
region_name	Hà Nội
company_ad	True
type	s
price	3290000000
price_string	3,29 tỷ
image	https://cdn.chotot
webp_image	https://cdn.chotot
videos	[]
number_of_images	3.0
avatar	https://cdn.chotot
rooms	4.0
property_legal_document	1.0
size	38.0
region_v2	12000
area_v2	12081

ward	43
ward_name	Phường Giang Biên
price_million_per_m2	86.578947
house_type	3.0
contain_videos	2
location	21.0557,105.9085
longitude	105.9085
latitude	21.0557
phone_hidden	True
owner	False
protection_entitlement	False
escrow_can_deposit	2
params	[{'id': 'size', 'va
zero_deposit	False
street_name	Việt Hưng
pty_jupiter	0
label_campaigns	NaN
ad_labels	NaN
pty_characteristics	[3]
seller_info.full_name	NaN
seller_info.avatar	NaN
seller_info.sold_ads	0
seller_info.live_ads	0
has_video	NaN
company_logo	NaN
land_type	NaN
streetnumber_display	NaN
width	NaN
length	NaN
toilets	4.0
floors	9.5
orig_list_time	4.0
furnishing_sell	4.0
living_size	NaN
special_display_images	2.0

```
special_display                          112.0
commercial_type                          ['https://cdn.chot
```

Remove unnecessary column and rename the value to have meaning:

Column Header	Value
TheLoai	Nhà ở
Huyen	Quận Long Biên
Tinh	Hà Nội
Gia	3290000000
TongSoPhong	4.0
GiayTo	Đã có sổ
DienTich(m2)	38.0
LoaiHinhNhaO	Nhà phố liền kề
LoaiHinhDat	NaN
ChieuNgang	4.0
ChieuDai	9.5
PhongVeSinh	4.0
SoTang	4.0
NoiThat	Noi That đầy đủ
DienTichSuDung	112.0
LoaiHinhVanPhong	NaN
TenDuongCuThe	NaN
HuongCuaChinh	NaN
LoaiHinhCanHo	NaN
TinhTrangBds	NaN
SoDuong	NaN
TenPhanKhu/Lo/Block/Thap	NaN
TangSo	NaN
HuongBanCong	NaN
MaLo	NaN
DacDiemCanHoChungCu	NaN

Detail of each categorical datafields:

- **TheLoai: '**Nhà ở', 'Đất', 'Căn hộ/Chung cư', 'Văn phòng, Mặt bằng kinh doanh', 'Phòng trọ’
- **GiayTo:** 'Đã có sổ', 'Sổ hồng riêng', 'Sổ chung', 'Đang chờ sổ', 'Hợp đồng đặt cọc', 'Giấy tờ khác’
- **LoaiHinhNhaO:** 'Nhà phố liền kề', 'Nhà mặt phố, mặt tiền', 'Nhà biệt thự', 'Nhà ngõ, hẻm’
- **LoaiHinhDat:** 'Đất thổ cư', 'Đất nền dự án', 'Đất công nghiệp', 'Đất nông nghiệp’
- **HuongCuaChinh** and **HuongBanCong: '**Đông Nam', 'Đông Bắc', 'Tây Bắc', 'Tây Nam', 'Đông', 'Tây', 'Nam', 'Bắc’
- **LoaiHinhCanHo**: 'Chung cư', 'Căn hộ dịch vụ, mini', 'Tập thể, cư xá', 'Penthouse', 'Duplex', 'Officetel’
- **TinhTrangBds**: 'Đã bàn giao', 'Chưa bàn giao’
- **NoiThat**: 'Noi That đầy đủ', 'Noi That cao cấp', 'Hoàn thiện cơ bản', 'Bàn giao thô’

We get overview of the number type in dataset:

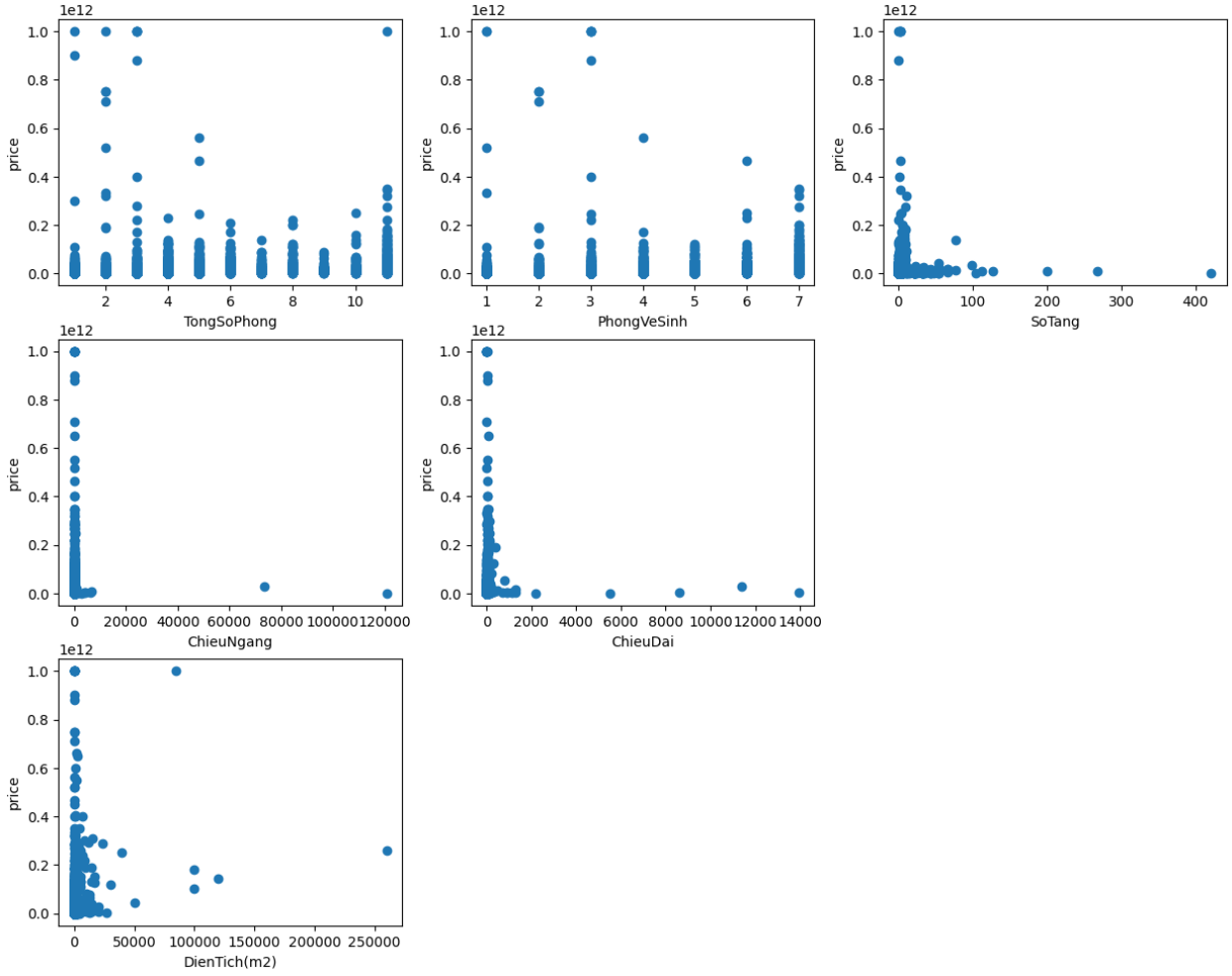
	Gia	TongSoPhong	DienTich(m2)	ChieuNgang	ChieuDai	PhongVeSinh	SoTang	DienTichSuDung	TangSo
count	40854.000	32737.000	40853.000	23631.000	20391.000	24591.000	18388.000	14818.000	1190.000
mean	7894028777.749	3.546	129.764	15.464	19.300	3.128	3.702	140.972	10.951
std	23489169657.314	2.048	1732.623	922.677	147.507	1.598	5.375	174.649	9.047
min	400000.000	1.000	1.000	1.200	1.000	1.000	1.000	1.000	1.000
25%	2700000000.000	2.000	45.000	4.000	11.000	2.000	2.000	50.000	4.000
50%	4500000000.000	3.000	64.000	4.800	15.000	3.000	3.000	96.000	9.000
75%	7500000000.000	4.000	95.000	5.401	20.000	4.000	5.000	180.000	15.000
max	999999999999.000	11.000	260000.000	120797.000	13947.000	7.000	421.000	7000.000	110.000

We can see the field **Gia**, **DienTich(m2)**, **ChieuNgang, ChieuDai, SoTang, DienTichSuDung ,TangSo** there are many too big and strange number in each fields, like `9999999999999` in field G**ia** or **260000** m2 in **DienTich(m2)**

This is the percentage of null value in my dataset, i will remove the categorical which contain more than 55% null values

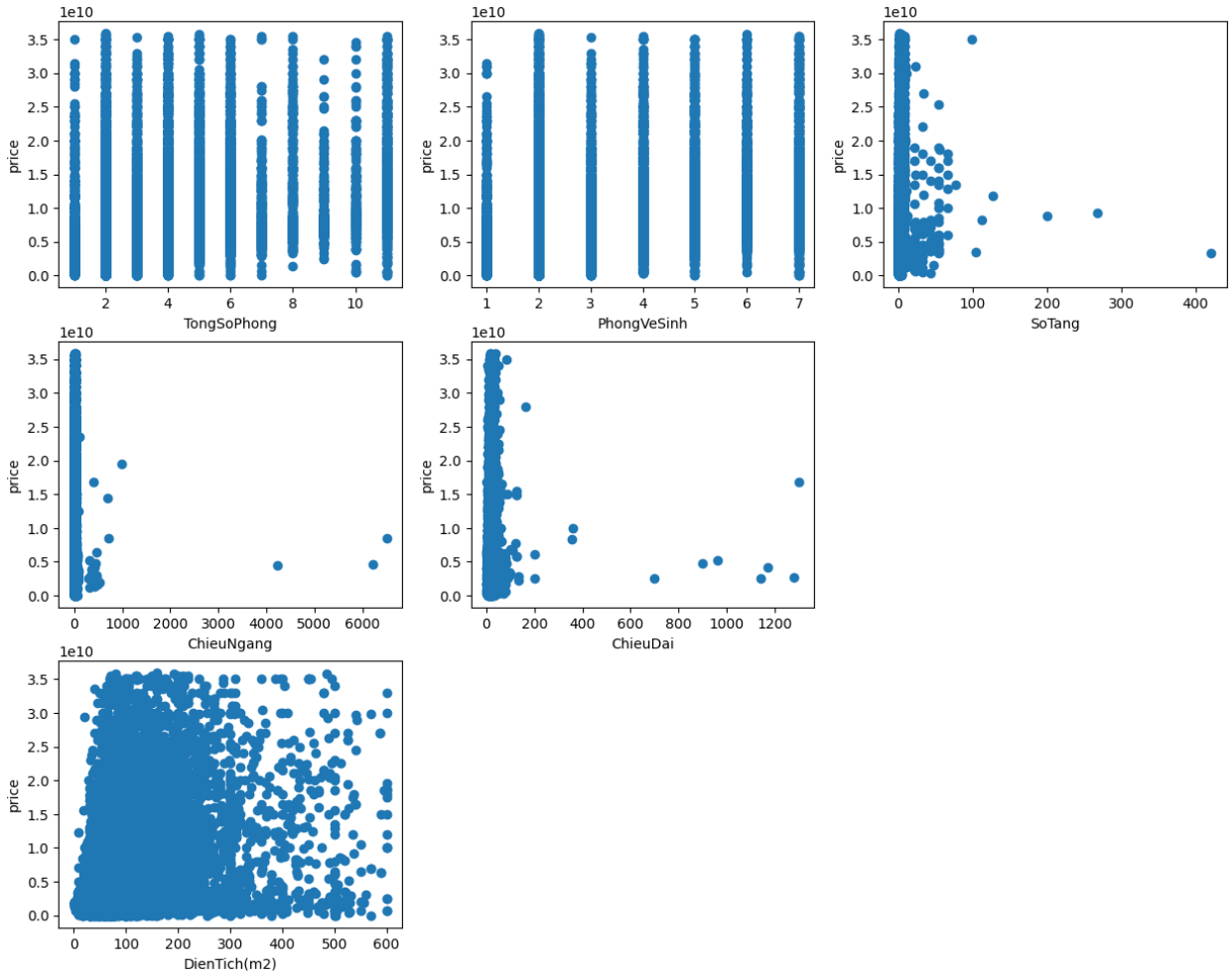
TheLoai	0.000
Huyen	0.002
Tinh	0.000
Gia	0.000
TongSoPhong	19.868
GiayTo	4.357
DienTich(m2)	0.002
LoaiHinhNhaO	35.032
LoaiHinhDat	81.326
ChieuNgang	42.157
ChieuDai	50.088
PhongVeSinh	39.808
SoTang	54.991
NoiThat	54.612
DienTichSuDung	63.729
LoaiHinhVanPhong	98.808
TenDuongCuThe	87.137
HuongCuaChinh	72.632
LoaiHinhCanHo	84.836
TinhTrangBds	84.836
SoDuong	86.902
TenPhanKhu/Lo/Block/Thap	94.860
TangSo	97.087
HuongBanCong	95.961
MaLo	96.125
DacDiemCanHoChungCu	97.156

Here is the visualization of each categorical:



⇒ So i will remove the outlier because it can affect the analysis and prediction.

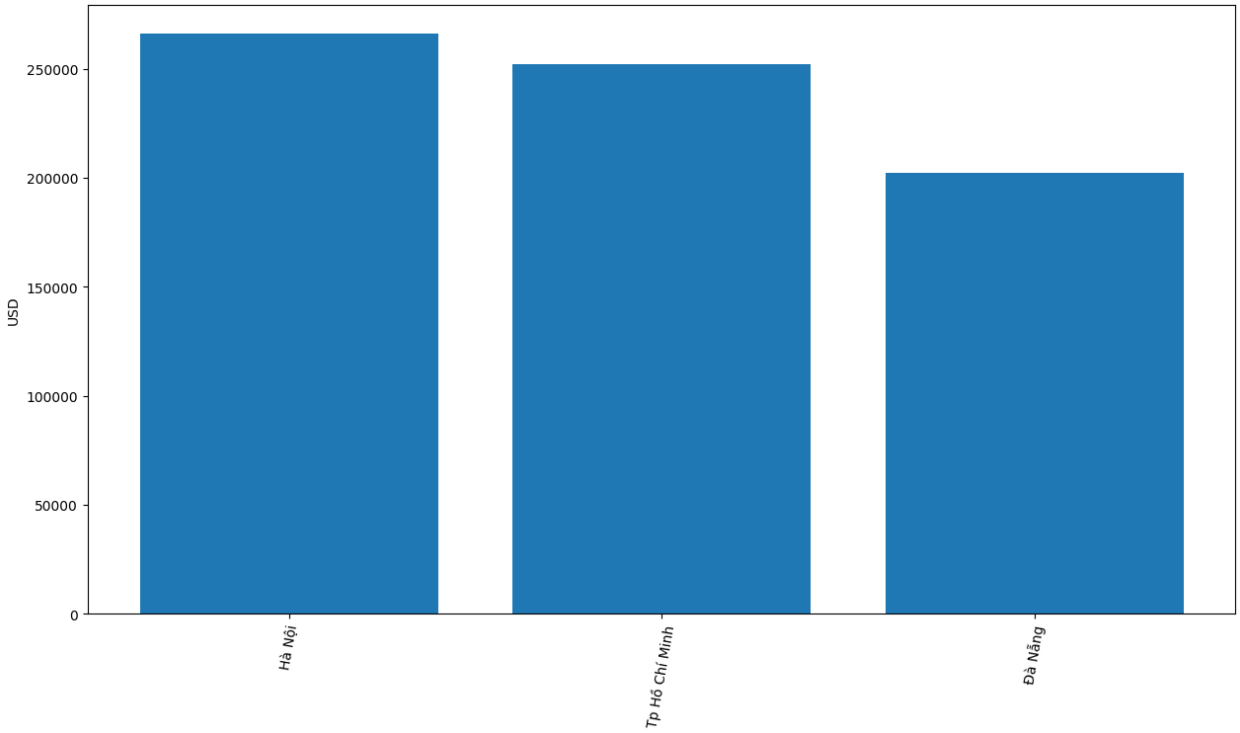
The visualization after cleaning outlier



Chap 3: Analyze data

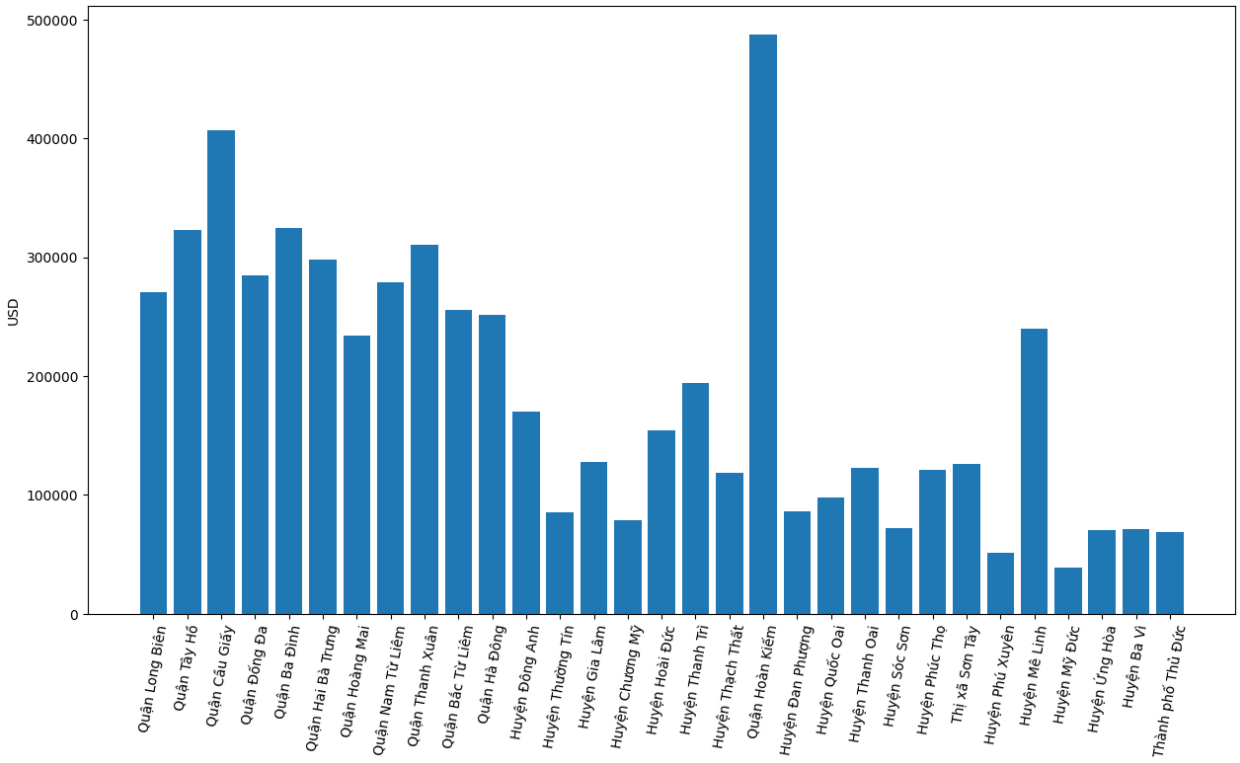
3.1. Compare the data in three big city

Here is the bar diagram compare average price (compare in use price because the number is very big) in three biggest in viet nam



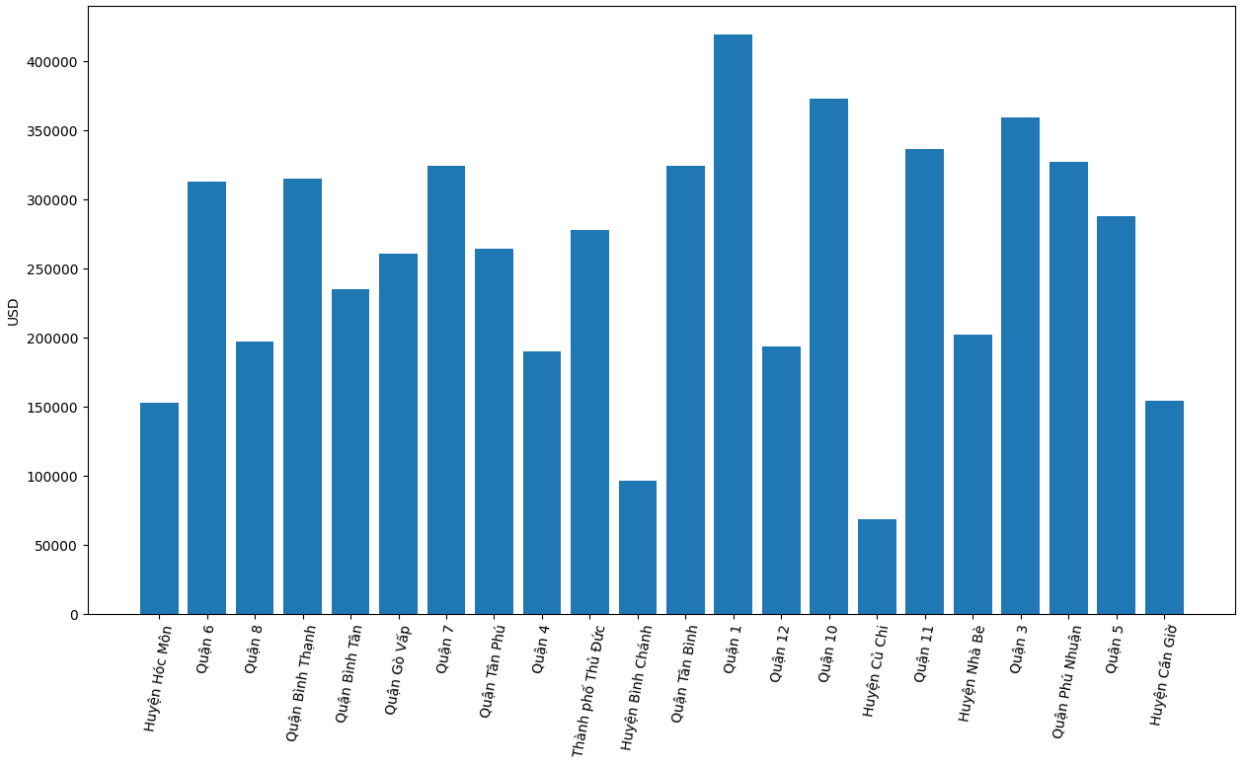
We can see Hà Nội is the city have biggest price and Đà Nẵng have lowest price.

Here is the bar diagram compare average price (compare in use price because the number is very big) in Ha Noi



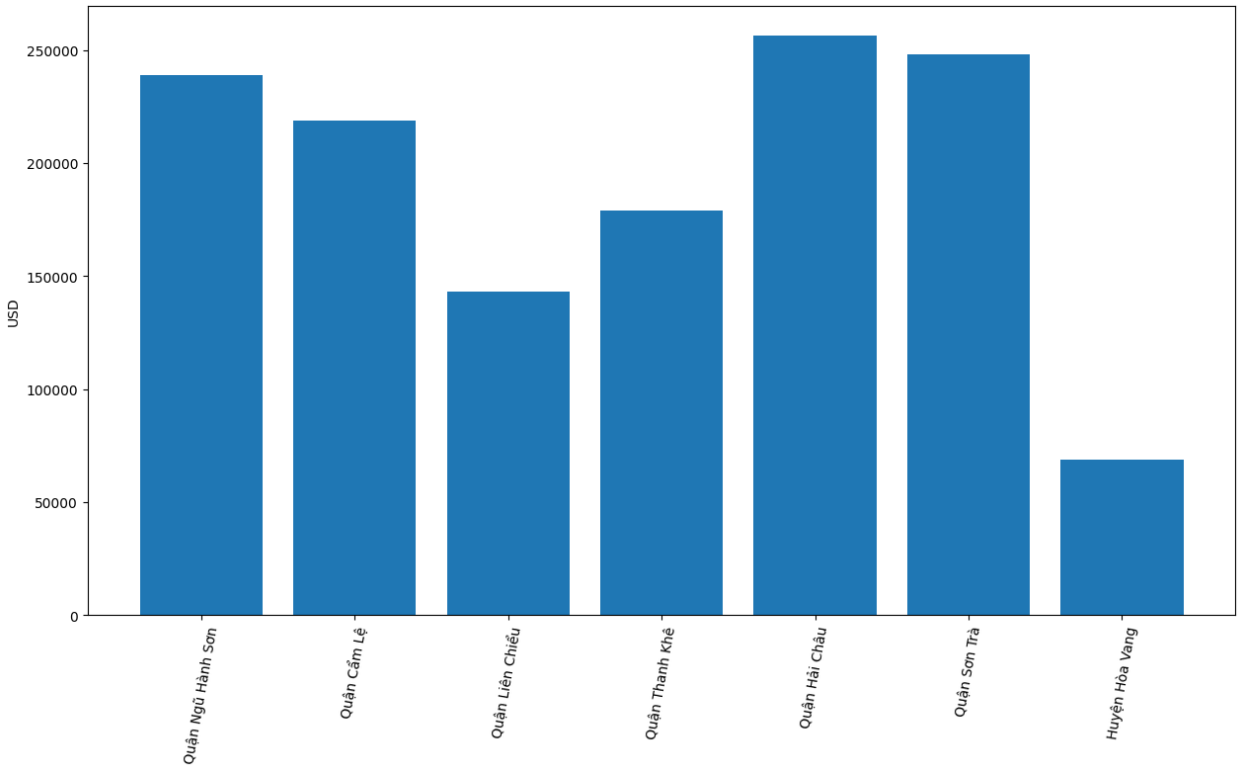
The average price Quận hoàng kiếm**,** Quân cầu giấy are very high. Specially, Quan hoang Kiem with about 500000 USD , This place is central of Ha Noi

Here is the bar diagram compare average price (compare in use price because the number is very big) in HCM



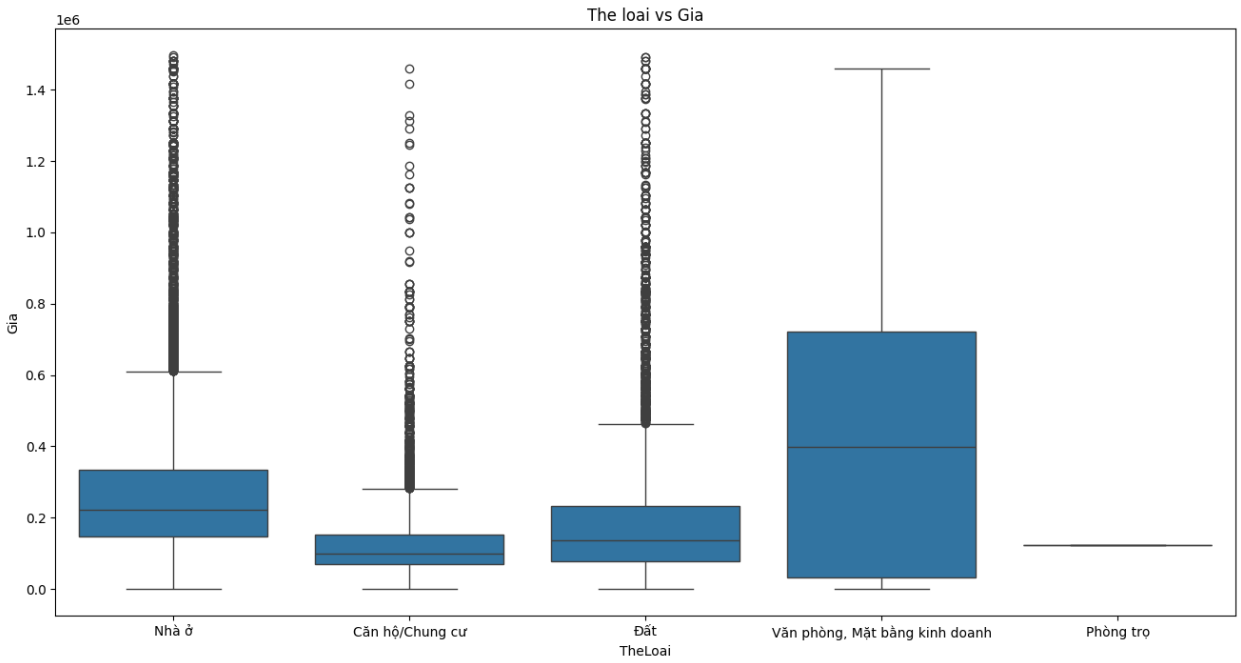
⇒ Quận 1 has the highest price, overall, HCM has the price very uniform

Here is the bar diagram compare average price (compare in use price because the number is very big) in Da Nang



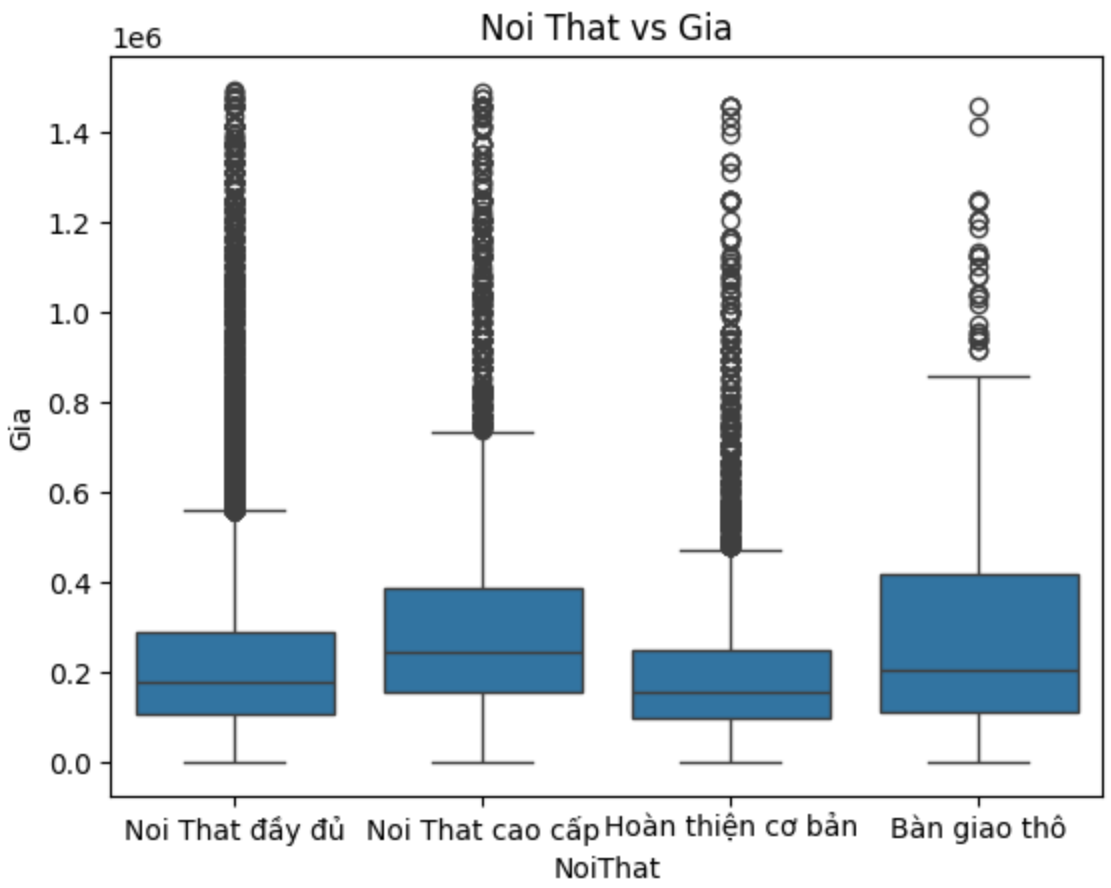
⇒ Quận Ngũ Hành Sơn , Quận hải châu, quận sơn trà have the same average price, represent the important central in Da Nang

Here is the box diagram compare average price between four type of The Loai



- *Nhà ở* has prices range from about 0.1 to 0.35 millions.There are many outliers above 0.6 million, indicating some high priced house
- *Căn hộ/ chung cư* has the smallest prices range , about 0.05 to 0.15 million. There are many outliers above 0.2 million
- *Đất* has the prices range mostly from 0.05 to 0.2 million. There are many outlier above 0.4 million
- *Văn phòng, Mặt bằng kinh doanh* has prices range widest from 0.05 to 0.7 million

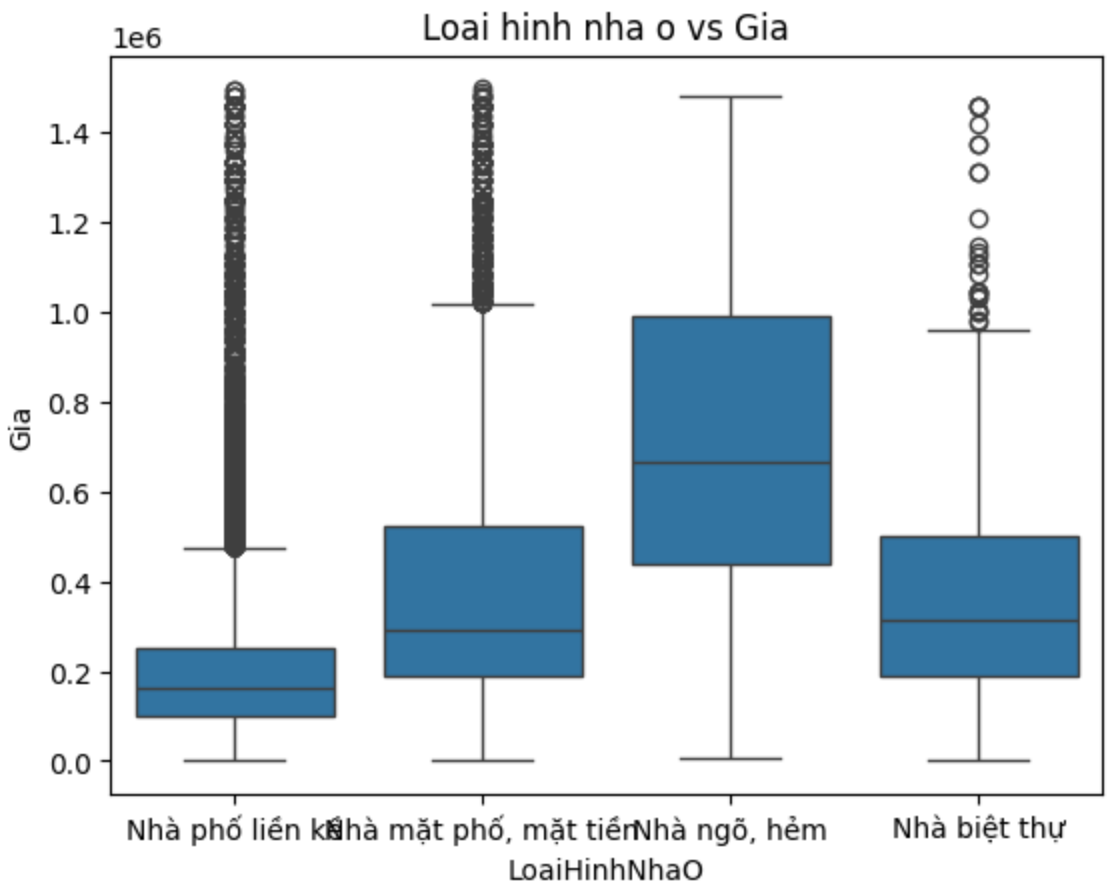
Here is the bar diagram compare average price (compare in use price because the number is very big) between four type of NoiThat



⇒ we can see *Nội thất cao cấp* are the type which has the range prices a little bit higher than others, from about 0.2 to 0.4 million.

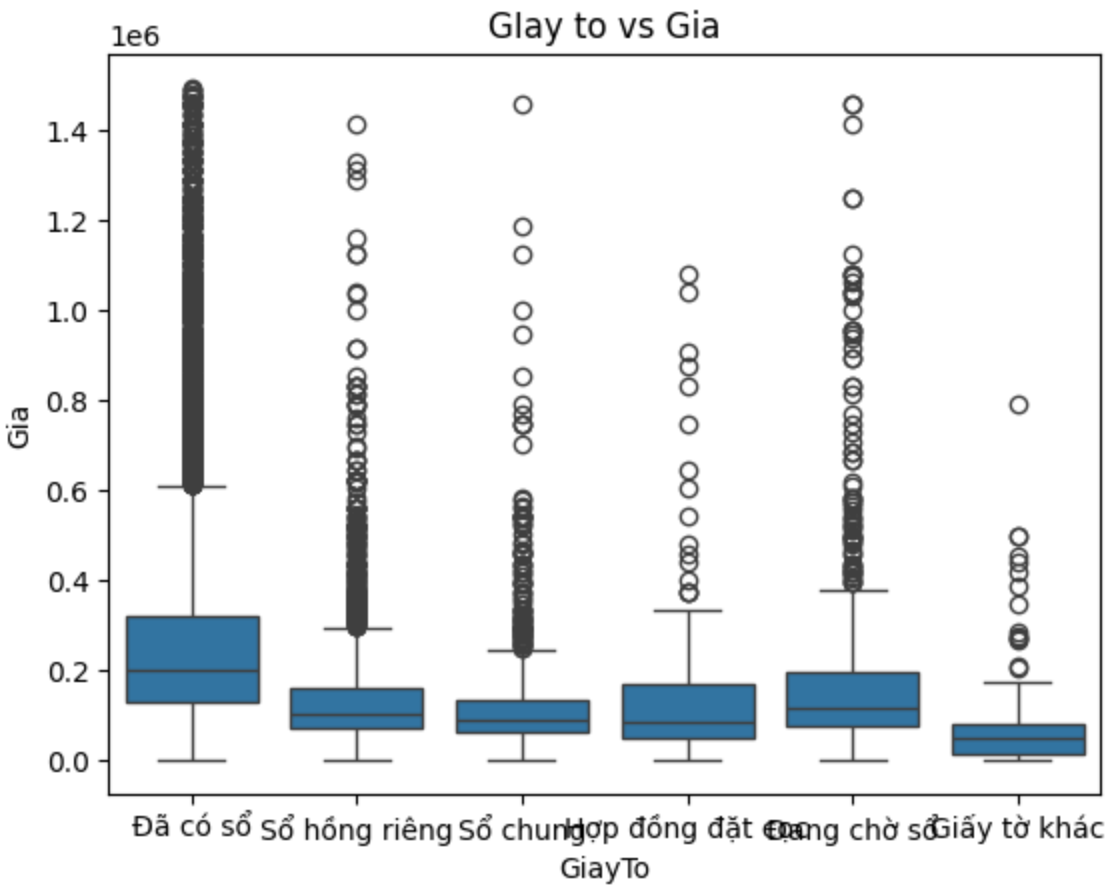
Overview all types have the same median values, about approximately 0.2 millions

Here is the bar diagram compare average price (compare in use price because the number is very big) between four type of LoaiHinhNhaO



- *Nhà Ngõ hẻm* is the type of house which has highest prices range. Which is from about 0.4 to 1 million.
- *Nhà Biệt Thư* and *Nhà Mặt phố* has the same house range prices about 0.2 to 0.5 million, they also have the same median about 0.3 million.

Here is the bar diagram compare average price (compare in use price because the number is very big) between four type of GiayTo



- *Đã có sổ* has the highest range prices, from about 0.15 to 0.4 million. There are many outliers above 0.6 million.⇒ This is the type which has highest house price
- Others *GiayTo* has the same median prices about 0.1 million

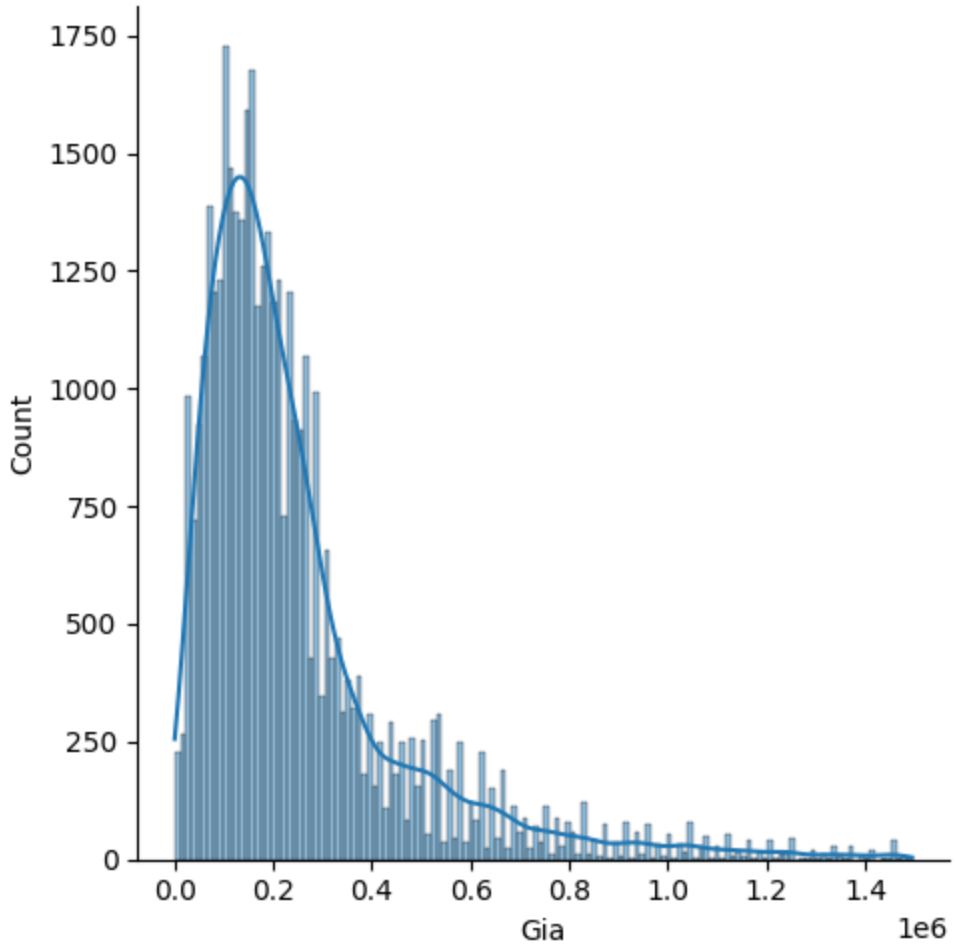
Chap 4: Predict the house price

4.1. Feature transformation

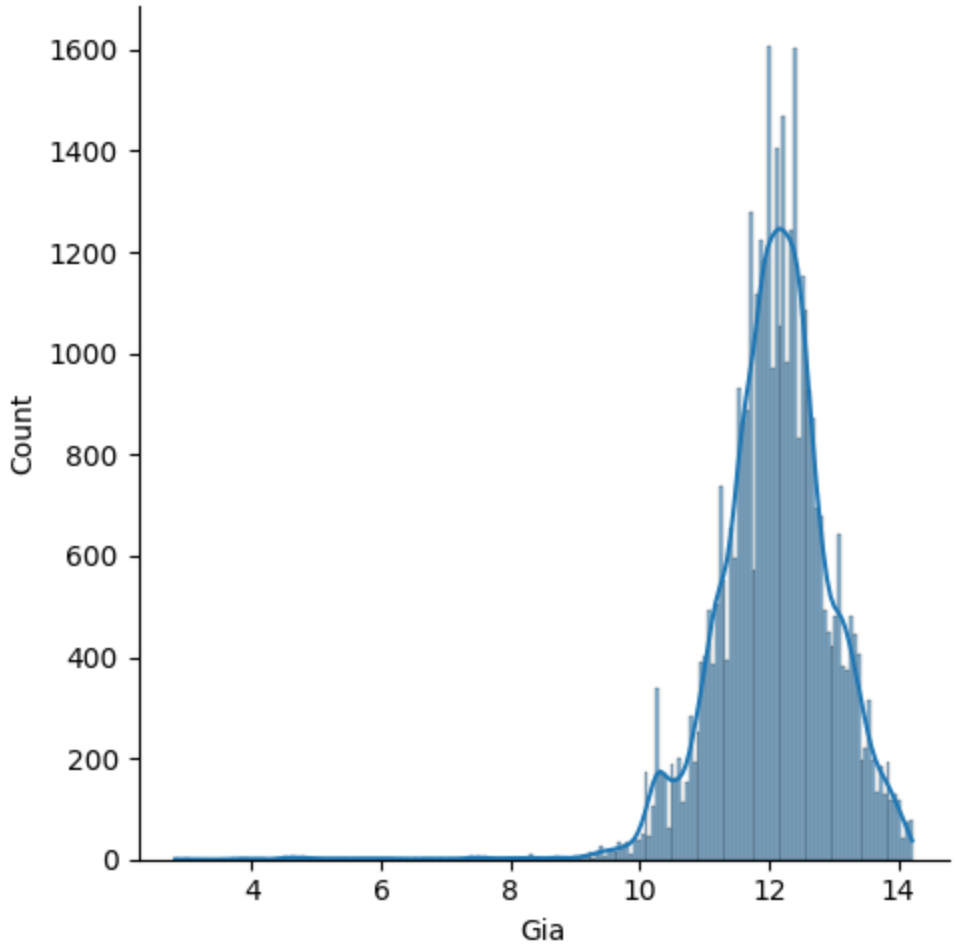
Convert each of the numeric value with has skew value has bigger than 0.5, by using logarithm

Transform the house price using logarithm to make it has the uniform distribution

Here is the histogram diagram of the price before transformation



Here is the histogram diagram of the price after transformation



Encode Categories

Convert the categories to the format that model can training, using One-hot coding:

Overview of dataset

	TongSoPhong	DienTich(m2)	ChieuNgang	ChieuDai	PhongVeSinh	SoTang	TheLoai_Căn hộ/Chung cư	TheLoai_Nhà ở	TheLoai_Phòng trọ	TheLoai_Văn phòng, Mặt bằng kinh doanh	TheLoai_Đất	Huyen_Huyện Ba Vì	Huye Bi
0	1.39	3.64	1.39	2.25	1.39	1.39	0.00	1.00	0.00	0.00	0.00	0.00	
1	2.40	4.98	2.74	2.96	0.69	2.08	0.00	1.00	0.00	0.00	0.00	0.00	
2	1.39	3.53	2.74	2.96	1.39	1.79	0.00	1.00	0.00	0.00	0.00	0.00	
3	1.10	4.50	2.74	2.96	0.69	0.69	1.00	0.00	0.00	0.00	0.00	0.00	
4	1.39	4.01	2.74	2.96	1.10	1.39	0.00	1.00	0.00	0.00	0.00	0.00	

Scaling

Using Z-score normalization, transform the value to make the mean value = 0 and std = 1

Overview of dataset

	TongSoPhong	DienTich(m2)	ChieuNgang	ChieuDai	PhongVeSinh	TheLoai_Nhà ở	Huyen_Huyện Bình Chánh	Huyen_Huyện Chương Mỹ	Huyen_Huyện Cần Giờ	Huyen_Huyện Củ Chi	Huyen_Huyện Gia Lâm
0	0.46383656	-0.82535167	-0.46745516	-0.79826314	0.72496029	0.00000000	-0.18698809	-0.02330233	-0.01164879	-0.07103056	-0.03083237
7	-0.14815244	-0.73215042	-0.05756623	-0.92929500	0.72496029	0.00000000	-0.18698809	-0.02330233	-0.01164879	-0.07103056	-0.03083237
29	-0.14815244	-0.97478044	-0.93214891	-0.67395389	0.72496029	0.00000000	-0.18698809	-0.02330233	-0.01164879	-0.07103056	-0.03083237
40	-0.14815244	-1.23081530	0.30909192	-1.91193928	0.13684971	0.00000000	-0.18698809	-0.02330233	-0.01164879	-0.07103056	-0.03083237
46	0.46383656	-1.05437095	-0.21577682	-1.27609975	1.18113429	0.00000000	-0.18698809	-0.02330233	-0.01164879	-0.07103056	-0.03083237

Training and Evualitaion

I am using linear regression to train model

Here is the R_score: -3.0380283688779e+22

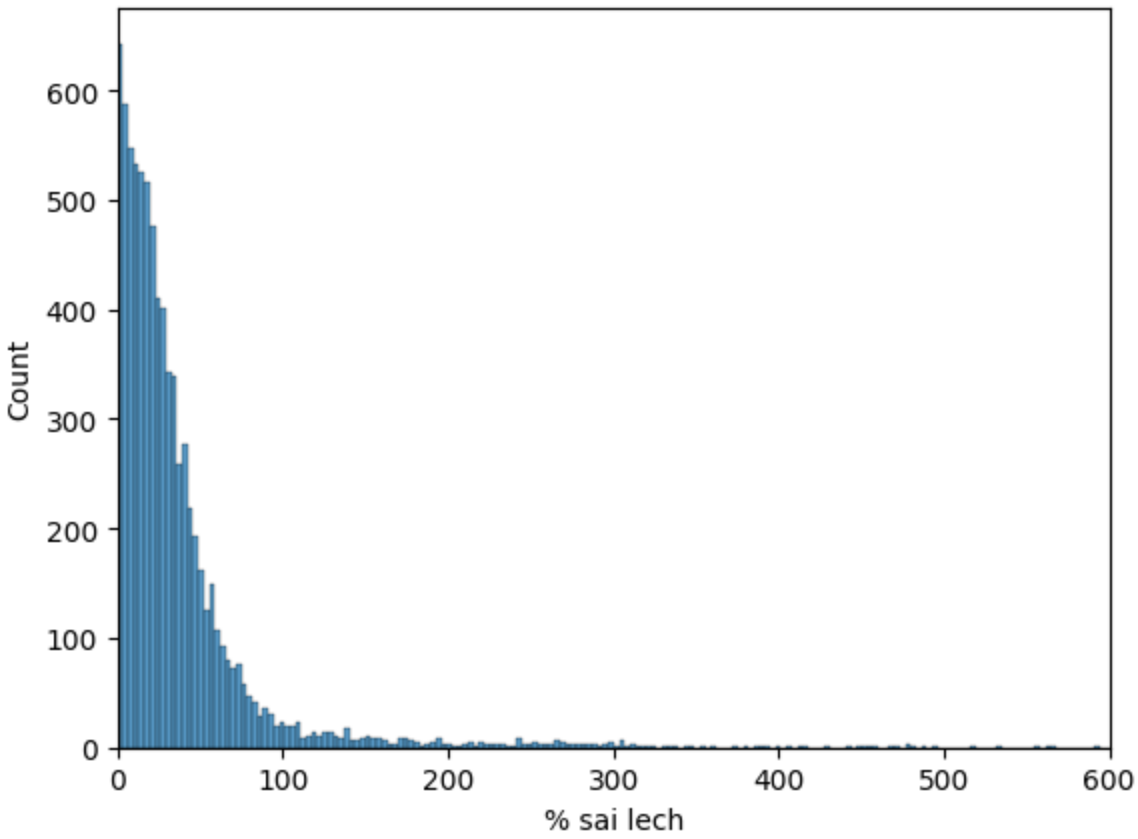
Here is the **root mean squared error(RMSE)**: 139128.62678864226

Here is the overview of the test price, real price and the percentage of deviation=

	Gia du doan	Gia thuc	% sai lech
count	7925.00000000	7925.00000000	7925.00000000
mean	221459.81808094	247141.48271256	173.43004630
std	162513.14483855	220011.10119299	3162.37906101
min	0.00000000	56.25000000	0.00263368
25%	115118.86682459	110416.66666667	11.00281783
50%	180467.57715986	185416.66666667	23.91405213
75%	276812.62183673	291666.66666667	43.53343693
max	2135704.40878021	1491666.66666666	145154.20344575

- The R score is very bad
- THe RMSE number is can acceptable, max value of price is 1 495 833 meanwhile the RMSE only 139 128
- 75 % of the data has percentage of deviation below 43% , 50% has percentage below 23% and 25 % has percentage below 11%

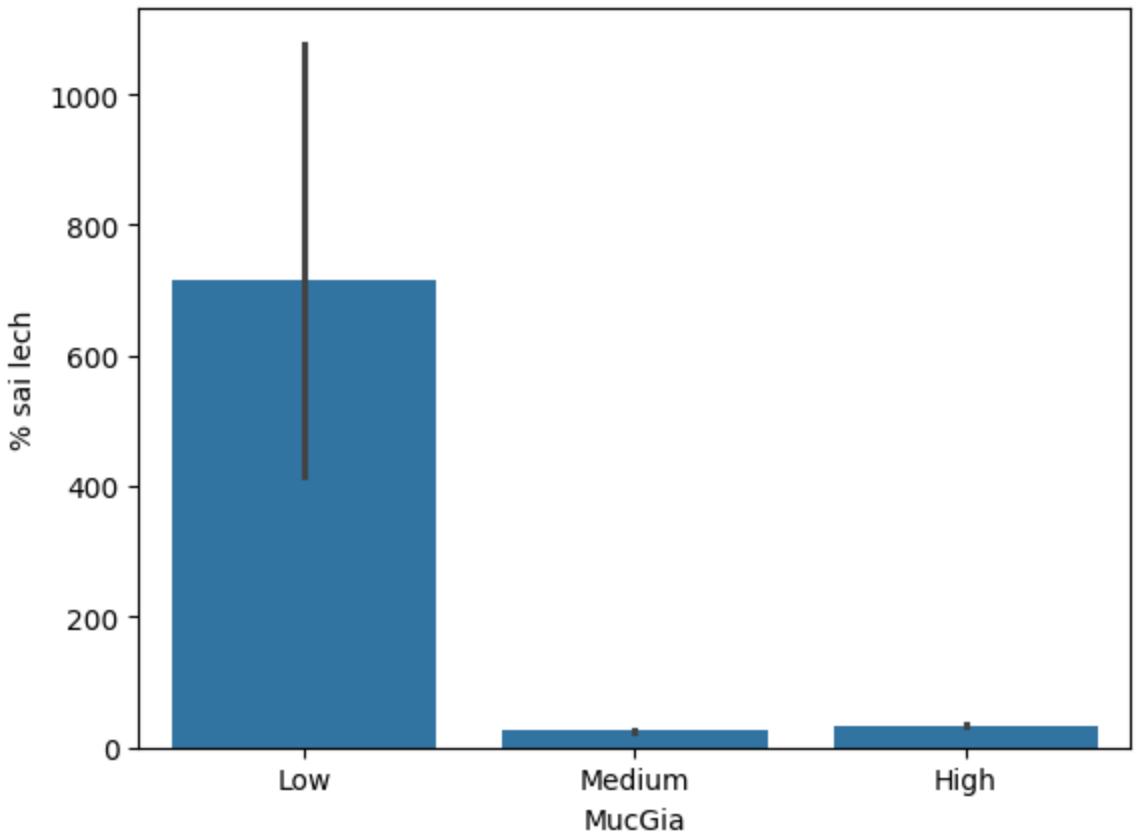
The diagram show the distribution of the percentage of deviation



⇒ The frequency percentage focus on below 75%.

I divide price to three type : price ≤= 100000 is low, price ≤ 400000 is medium and > 400000 is High

The diagram show the average of % sai lech between MucGia



We can see the model best fit which the price > 100000 USD