VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF ECONOMICS AND LAW

FINAL PROJECT

SUBJECT: MACHINE LEARNING

Predicting House Prices in Ho Chi Minh City Using Machine Learning

LECTURER	**:**	**M.S. Phan Huy Tam**
STUDENT	**:**	**232TC6701**
CLASS	**:**	**Nguyen Quoc Huy**
STUDENT ID	**:**	**K214142068**

Ho Chi Minh City, June 5 th, 2024

Contents

Project Summary

Forecasting home prices is an important financial decision, not only for those working in the real estate industry but also for potential home buyers. Because I live and work in Ho Chi Minh City, I decided to apply the Linear Regression, Random Forest Regression and Decision Tree Regression models to find the best house price prediction model in Ho Chi Minh City based on data taken from the ChoTot site.
The ultimate goal of the project is to build a machine learning model to help home buyers and sellers in Vietnam find a fair price for their homes.

Index term: House price prediction, Linear Regression, Random Forest Regression, Decision Tree Regression.

I. Introduction:

With property prices rising quickly all over the world, especially in Vietnam, house price prediction has become a hot topic in the real estate sector. However, some research models have very big mistakes and produce outcomes with low accuracy due to numerous restrictions. This is a mistake because there are numerous elements that affect the price of a house, including its location, orientation and other characteristics.

The study's goal is to determine which machine learning model is more effective at predicting Ho Chi Minh City real estate values. The model can forecast the price with the least amount of inaccuracy based on the location, size, number of bedrooms, toilets, and other factors collected from the ChoTot page (Cho Tot is a Vietnamese online classifieds site for real estate, vehicles, jobs, and secondhand (Cho Tot is a Vietnamese online marketplace for homes, cars, recruitment, used electronics, pets, and home services).

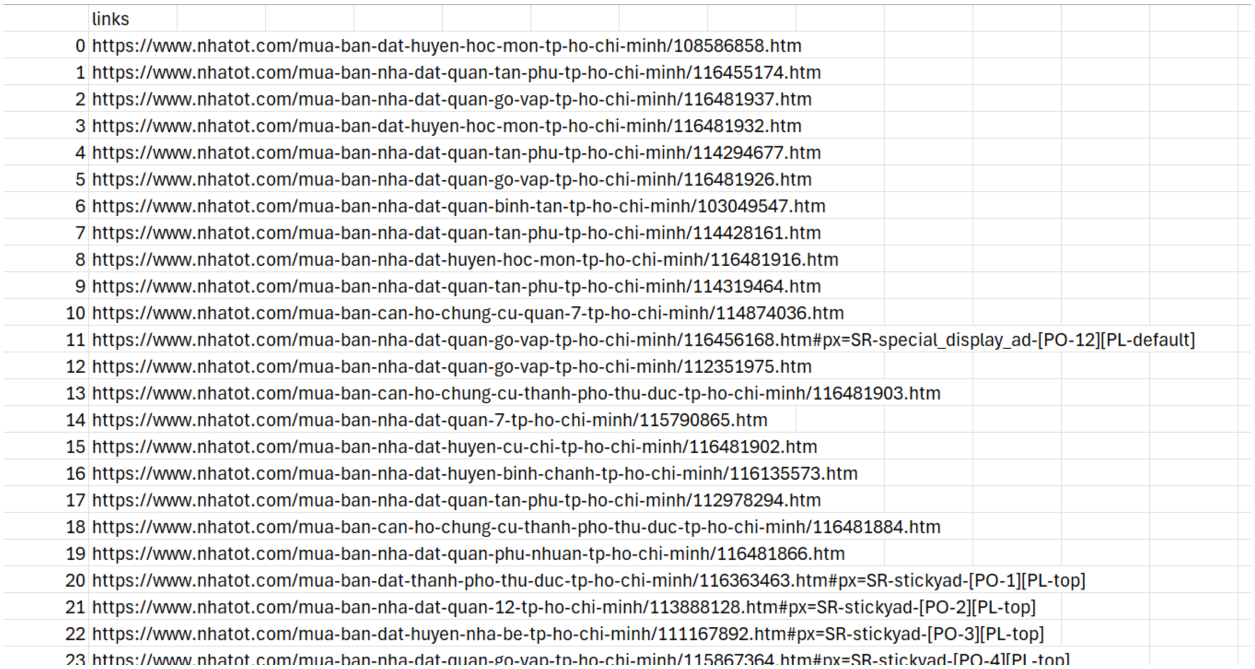
The difference of this study: Currently in Vietnam, there are not many predictions and data on house prices collected in Cho Tot. Furthermore, research documents demonstrate their low accuracy.

Disadvantages: The accuracy of the data has not been confirmed because it is taken from the ChoTot website. There is a lot of noise and outliers in the data, which makes prediction difficult and error-prone. In addition, there are many factors that affect home prices beyond the variables I included in the article.

II. Data Mining

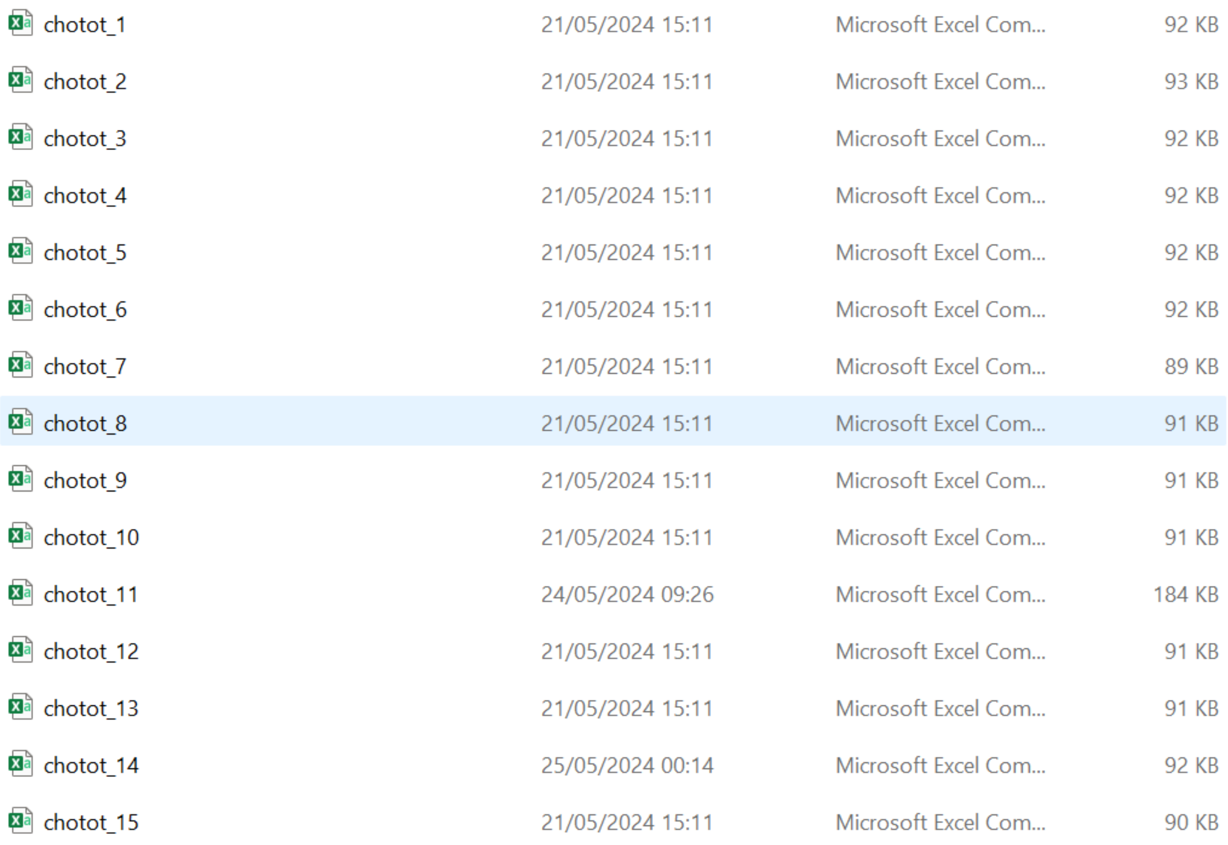
Data mining steps for a dataset containing house prices in Ho Chi Minh City, posted on the online Cho Tot site.

Step 1: I used Beautifulsoup to collect data but Cho Tot blocked me from getting data this way. So I used Selenium to get the link on the Cho Tot website. Then save it as a file "chotot.csv" containing the saved links.



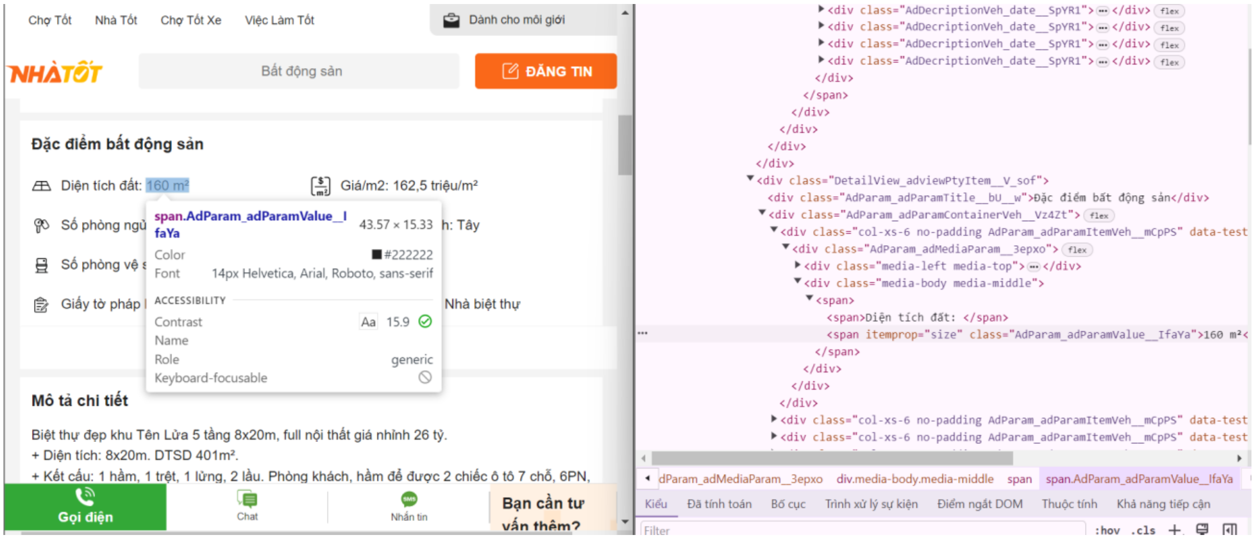
	links
0	https://www.nhatot.com/mua-ban-dat-huyen-hoc-mon-tp-ho-chi-minh/108586858.htm
1	https://www.nhatot.com/mua-ban-nha-dat-quan-tan-phu-tp-ho-chi-minh/116455174.htm
2	https://www.nhatot.com/mua-ban-nha-dat-quan-go-vap-tp-ho-chi-minh/116481937.htm
3	https://www.nhatot.com/mua-ban-dat-huyen-hoc-mon-tp-ho-chi-minh/116481932.htm
4	https://www.nhatot.com/mua-ban-nha-dat-quan-tan-phu-tp-ho-chi-minh/114294677.htm
5	https://www.nhatot.com/mua-ban-nha-dat-quan-go-vap-tp-ho-chi-minh/116481926.htm
6	https://www.nhatot.com/mua-ban-nha-dat-quan-binh-tan-tp-ho-chi-minh/103049547.htm
7	https://www.nhatot.com/mua-ban-nha-dat-quan-tan-phu-tp-ho-chi-minh/114428161.htm
8	https://www.nhatot.com/mua-ban-nha-dat-huyen-hoc-mon-tp-ho-chi-minh/116481916.htm
9	https://www.nhatot.com/mua-ban-nha-dat-quan-tan-phu-tp-ho-chi-minh/114319464.htm
10	https://www.nhatot.com/mua-ban-can-ho-chung-cu-quan-7-tp-ho-chi-minh/114874036.htm
11	https://www.nhatot.com/mua-ban-nha-dat-quan-go-vap-tp-ho-chi-minh/116456168.htm#px=SR-special_display_ad-[PO-12][PL-default]
12	https://www.nhatot.com/mua-ban-nha-dat-quan-go-vap-tp-ho-chi-minh/112351975.htm
13	https://www.nhatot.com/mua-ban-can-ho-chung-cu-thanh-pho-thu-duc-tp-ho-chi-minh/116481903.htm
14	https://www.nhatot.com/mua-ban-nha-dat-quan-7-tp-ho-chi-minh/115790865.htm
15	https://www.nhatot.com/mua-ban-nha-dat-huyen-cu-chi-tp-ho-chi-minh/116481902.htm
16	https://www.nhatot.com/mua-ban-nha-dat-huyen-binh-chanh-tp-ho-chi-minh/116135573.htm
17	https://www.nhatot.com/mua-ban-nha-dat-quan-tan-phu-tp-ho-chi-minh/112978294.htm
18	https://www.nhatot.com/mua-ban-can-ho-chung-cu-thanh-pho-thu-duc-tp-ho-chi-minh/116481884.htm
19	https://www.nhatot.com/mua-ban-nha-dat-quan-phu-nhuan-tp-ho-chi-minh/116481866.htm
20	https://www.nhatot.com/mua-ban-dat-thanh-pho-thu-duc-tp-ho-chi-minh/116363463.htm#px=SR-stickyad-[PO-1][PL-top]
21	https://www.nhatot.com/mua-ban-nha-dat-quan-12-tp-ho-chi-minh/113888128.htm#px=SR-stickyad-[PO-2][PL-top]
22	https://www.nhatot.com/mua-ban-dat-huyen-nha-be-tp-ho-chi-minh/111167892.htm#px=SR-stickyad-[PO-3][PL-top]
23	https://www.nhatot.com/mua-ban-nha-dat-quan-go-vap-tp-ho-chi-minh/115867364.htm#px=SR-stickyad-[PO-4][PL-top]

Step 2: Because the data set containing the link is too large, it should be divided into many small sets.



chotot_1	21/05/2024 15:11	Microsoft Excel Com...	92 KB
chotot_2	21/05/2024 15:11	Microsoft Excel Com...	93 KB
chotot_3	21/05/2024 15:11	Microsoft Excel Com...	92 KB
chotot_4	21/05/2024 15:11	Microsoft Excel Com...	92 KB
chotot_5	21/05/2024 15:11	Microsoft Excel Com...	92 KB
chotot_6	21/05/2024 15:11	Microsoft Excel Com...	92 KB
chotot_7	21/05/2024 15:11	Microsoft Excel Com...	89 KB
chotot_8	21/05/2024 15:11	Microsoft Excel Com...	91 KB
chotot_9	21/05/2024 15:11	Microsoft Excel Com...	91 KB
chotot_10	21/05/2024 15:11	Microsoft Excel Com...	91 KB
chotot_11	24/05/2024 09:26	Microsoft Excel Com...	184 KB
chotot_12	21/05/2024 15:11	Microsoft Excel Com...	91 KB
chotot_13	21/05/2024 15:11	Microsoft Excel Com...	91 KB
chotot_14	25/05/2024 00:14	Microsoft Excel Com...	92 KB
chotot_15	21/05/2024 15:11	Microsoft Excel Com...	90 KB

Step 3: Run each data set containing the link. Then get what you need to find such as: Address, size, rooms,...



Đặc điểm bất động sản

Diện tích đất: 160 m²

Giá/m2: 162,5 triệu/m²

Số phòng ngủ

h: Tây

Số phòng vệ s

Giấy tờ pháp l

Nhà biệt thự

span.AdParam_adParamValue__IfaYa 43.57 × 15.33

Color ■#222222

Font 14px Helvetica, Arial, Roboto, sans-serif

ACCESSIBILITY

Contrast Aa 15.9 ✓

Name

Role generic

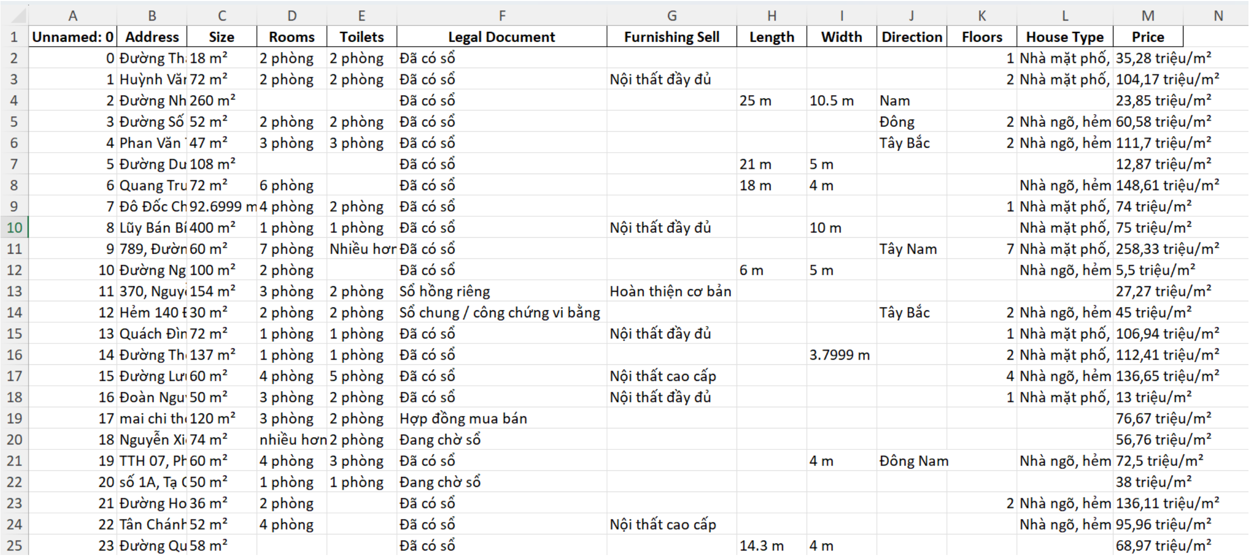
Keyboard-focusable ⃠

Mô tả chi tiết

Biệt thự đẹp khu Tên Lửa 5 tầng 8x20m, full nội thất giá nhỉnh 26 tỷ.
+ Diện tích: 8x20m. DTSD 401m².
+ Kết cấu: 1 hầm, 1 trệt, 1 lửng, 2 lầu. Phòng khách, hầm để được 2 chiếc ô tô 7 chỗ, 6PN,

```
              ▶<div class="AdDecriptionVeh_date__SpYR1">⋯</div> flex
              ▶<div class="AdDecriptionVeh_date__SpYR1">⋯</div> flex
              ▶<div class="AdDecriptionVeh_date__SpYR1">⋯</div> flex
              ▶<div class="AdDecriptionVeh_date__SpYR1">⋯</div> flex
             </div>
            </span>
           </div>
          </div>
         </div>
        </div>
       ▼<div class="DetailView_adviewPtyItem__V_sof">
         <div class="AdParam_adParamTitle__bU__w">Đặc điểm bất động sản</div>
        ▼<div class="AdParam_adParamContainerVeh__Vz4Zt"> flex
         ▼<div class="col-xs-6 no-padding AdParam_adParamItemVeh__mCpPS" data-test
          ▼<div class="AdParam_adMediaParam__3epxo"> flex
           ▶<div class="media-left media-top">⋯</div>
           ▼<div class="media-body media-middle">
            ▼<span>
               <span>Diện tích đất: </span>
...            <span itemprop="size" class="AdParam_adParamValue__IfaYa">160 m²<
             </span>
           </div>
          </div>
         </div>
        ▶<div class="col-xs-6 no-padding AdParam_adParamItemVeh__mCpPS" data-test
        ▶<div class="col-xs-6 no-padding AdParam_adParamItemVeh__mCpPS" data-test
```

Step 4: Merge the newly retrieved data files into file "datachotot.xlsx".



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Unnamed: 0	Address	Size	Rooms	Toilets	Legal Document	Furnishing Sell	Length	Width	Direction	Floors	House Type	Price	
2	0	Đường Th	18 m²	2 phòng	2 phòng	Đã có sổ					1	Nhà mặt phố,	35,28 triệu/m²	
3	1	Huỳnh Văı	72 m²	2 phòng	2 phòng	Đã có sổ	Nội thất đầy đủ				2	Nhà mặt phố,	104,17 triệu/m²	
4	2	Đường Nh	260 m²			Đã có sổ		25 m	10.5 m	Nam			23,85 triệu/m²	
5	3	Đường Số	52 m²	2 phòng	2 phòng	Đã có sổ				Đông	2	Nhà ngõ, hẻm	60,58 triệu/m²	
6	4	Phan Văn	47 m²	3 phòng	3 phòng	Đã có sổ				Tây Bắc	2	Nhà ngõ, hẻm	111,7 triệu/m²	
7	5	Đường Du	108 m²			Đã có sổ		21 m	5 m				12,87 triệu/m²	
8	6	Quang Tru	72 m²	6 phòng		Đã có sổ		18 m	4 m			Nhà ngõ, hẻm	148,61 triệu/m²	
9	7	Đô Đốc Ch	92.6999 m	4 phòng	2 phòng	Đã có sổ					1	Nhà mặt phố,	74 triệu/m²	
10	8	Lũy Bán Bí	400 m²	1 phòng	1 phòng	Đã có sổ	Nội thất đầy đủ		10 m			Nhà mặt phố,	75 triệu/m²	
11	9	789, Đườn	60 m²	7 phòng	Nhiều hơı	Đã có sổ				Tây Nam	7	Nhà mặt phố,	258,33 triệu/m²	
12	10	Đường Ng	100 m²	2 phòng		Đã có sổ		6 m	5 m			Nhà ngõ, hẻm	5,5 triệu/m²	
13	11	370, Nguy	154 m²	3 phòng	2 phòng	Sổ hồng riêng	Hoàn thiện cơ bản						27,27 triệu/m²	
14	12	Hẻm 140 Ē	30 m²	2 phòng	2 phòng	Sổ chung / công chứng vi bằng				Tây Bắc	2	Nhà ngõ, hẻm	45 triệu/m²	
15	13	Quách Đìn	72 m²	1 phòng	1 phòng	Đã có sổ	Nội thất đầy đủ				1	Nhà mặt phố,	106,94 triệu/m²	
16	14	Đường Th	137 m²	1 phòng	1 phòng	Đã có sổ			3.7999 m		2	Nhà mặt phố,	112,41 triệu/m²	
17	15	Đường Lư	60 m²	4 phòng	5 phòng	Đã có sổ	Nội thất cao cấp				4	Nhà ngõ, hẻm	136,65 triệu/m²	
18	16	Đoàn Ngu	50 m²	3 phòng	2 phòng	Đã có sổ	Nội thất đầy đủ				1	Nhà mặt phố,	13 triệu/m²	
19	17	mai chi th	120 m²	3 phòng	2 phòng	Hợp đồng mua bán							76,67 triệu/m²	
20	18	Nguyễn Xi	74 m²	nhiều hơn	2 phòng	Đang chờ sổ							56,76 triệu/m²	
21	19	TTH 07, Ph	60 m²	4 phòng	3 phòng	Đã có sổ			4 m	Đông Nam		Nhà ngõ, hẻm	72,5 triệu/m²	
22	20	số 1A, Tạ (	50 m²	1 phòng	1 phòng	Đang chờ sổ							38 triệu/m²	
23	21	Đường Ho	36 m²	2 phòng		Đã có sổ					2	Nhà ngõ, hẻm	136,11 triệu/m²	
24	22	Tân Chánh	52 m²	4 phòng		Đã có sổ	Nội thất cao cấp					Nhà ngõ, hẻm	95,96 triệu/m²	
25	23	Đường Qu	58 m²			Đã có sổ		14.3 m	4 m				68,97 triệu/m²	

Problem: There is a lot of data missing here and there will be some columns that do not affect the overall output price, so we will have to process and directly quantify the data columns to find any data that is retained to build build the training process

Data before cleaning: The data set includes 12 variables and 16,479 observations.

Variable name	Description	Data type
Address	Address of the house building, in Ho Chi Minh City	Object
Size	Actual area on pink book, unit: million/m2.	Object
Rooms	Number of bedrooms.	Object
Toilets	Number of toilets.	Object
Legal Document	Legal documents of the house, whether in dispute or not, legal or not.	Object
Furnishing Sell	Is the house furnished or not?	Object
Length	Length of the house	Object
Width	Width of the house	Object
Direction	Direction of the house	Object
Floors	What floor is the house located on?	Float64
House Type	Characteristics of the house	Object
Price	Selling price of the house	Object

Data after cleaning: The data set includes 11 variables and 10,263 observations.

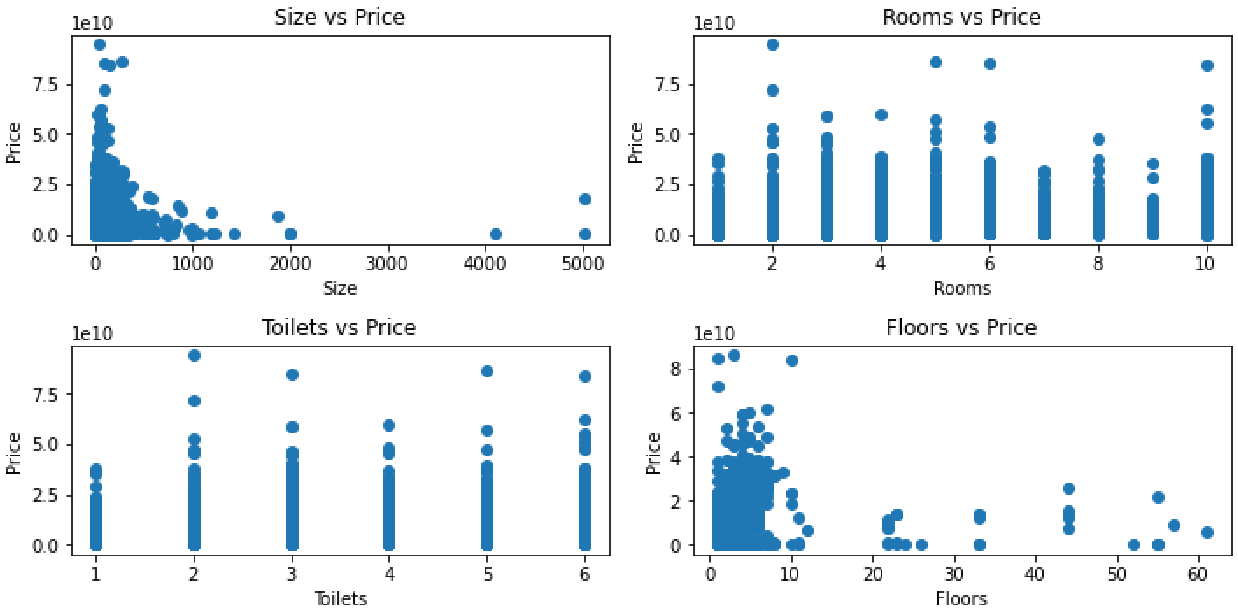
Variable name	Description	Data type

Address	Address of the house building, in Ho Chi Minh City	Object
District	Name of district/city	Object
Size	Actual area on pink book, unit: million/m2.	Float64
Rooms	Number of bedrooms.	Int32
Toilets	Number of toilets.	Int32
Legal Document	Legal documents of the house, whether in dispute or not, legal or not.	Object
Furnishing Sell	Is the house furnished or not?	Object
Direction	Direction of the house	Object
Floors	What floor is the house located on?	Float64
House Type	Characteristics of the house	Object
Price	Selling price of the house	Float64
USD	Convert VND to USD at the exchange rate of 25470	Float64

Note: Because the "Floors" Nan data is large, I do not drop the Nan data of this column, so I cannot convert non-finite values (NA or inf) to integers.

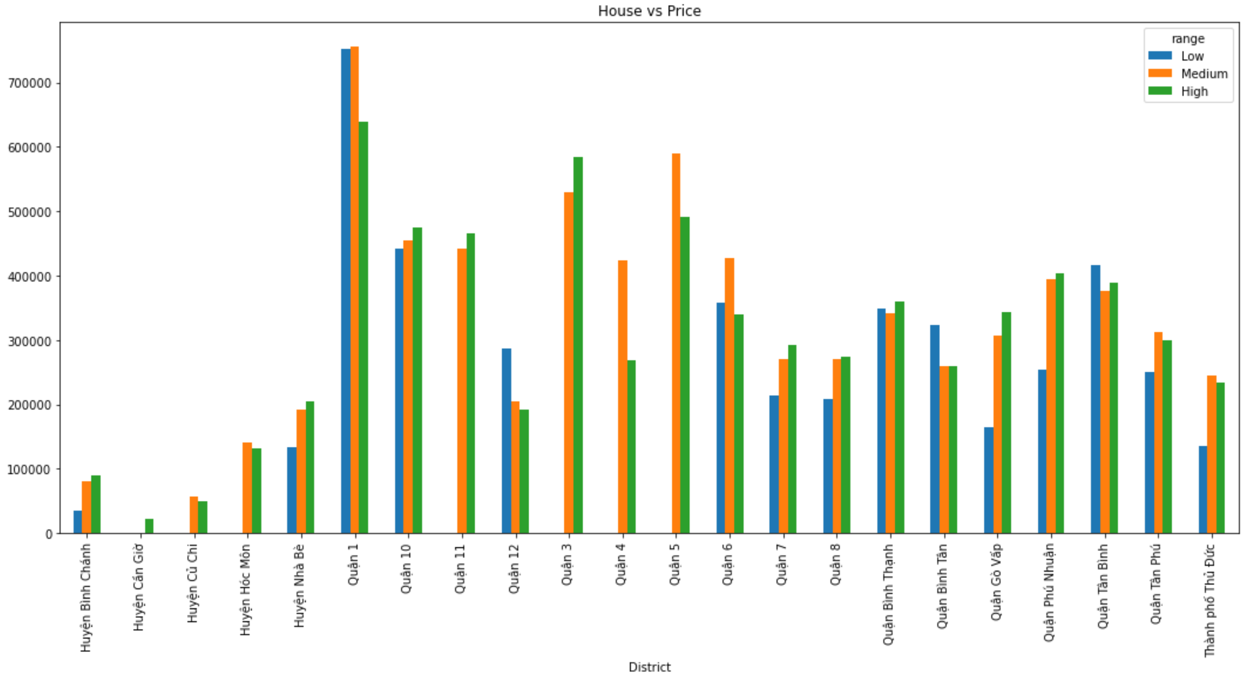
III. Data Visualization

Visualize numeric value data



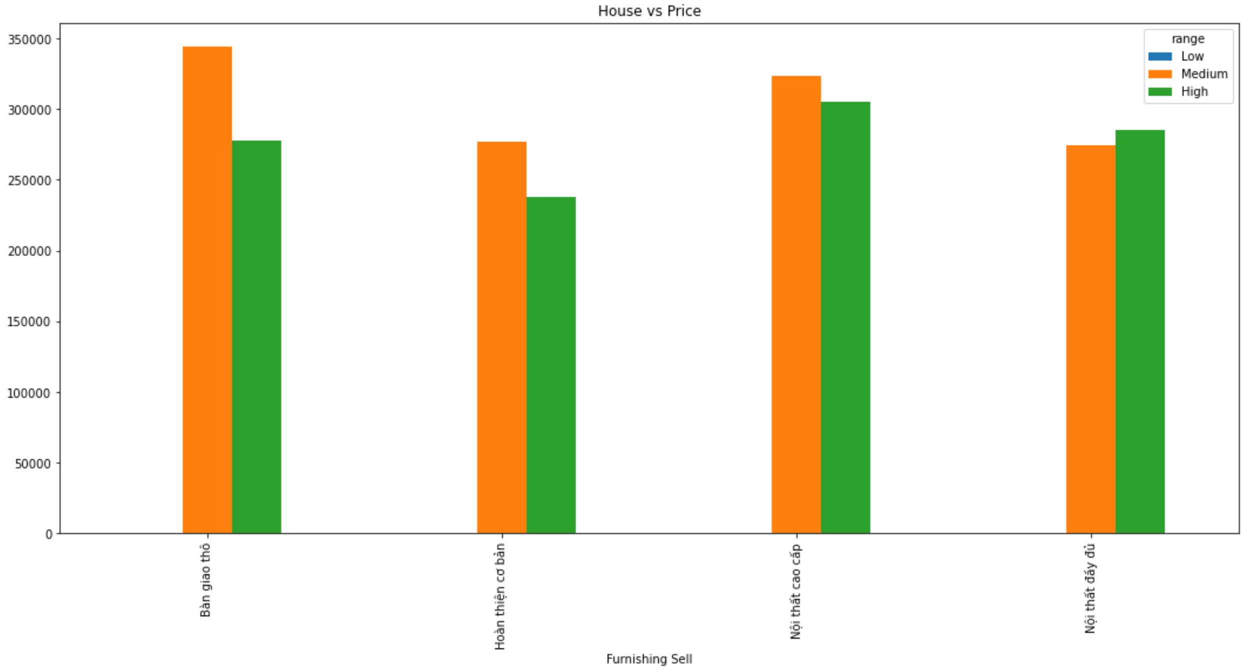
Conclusion: Bedrooms and toilets do not affect the price much.

Correlation between district and price



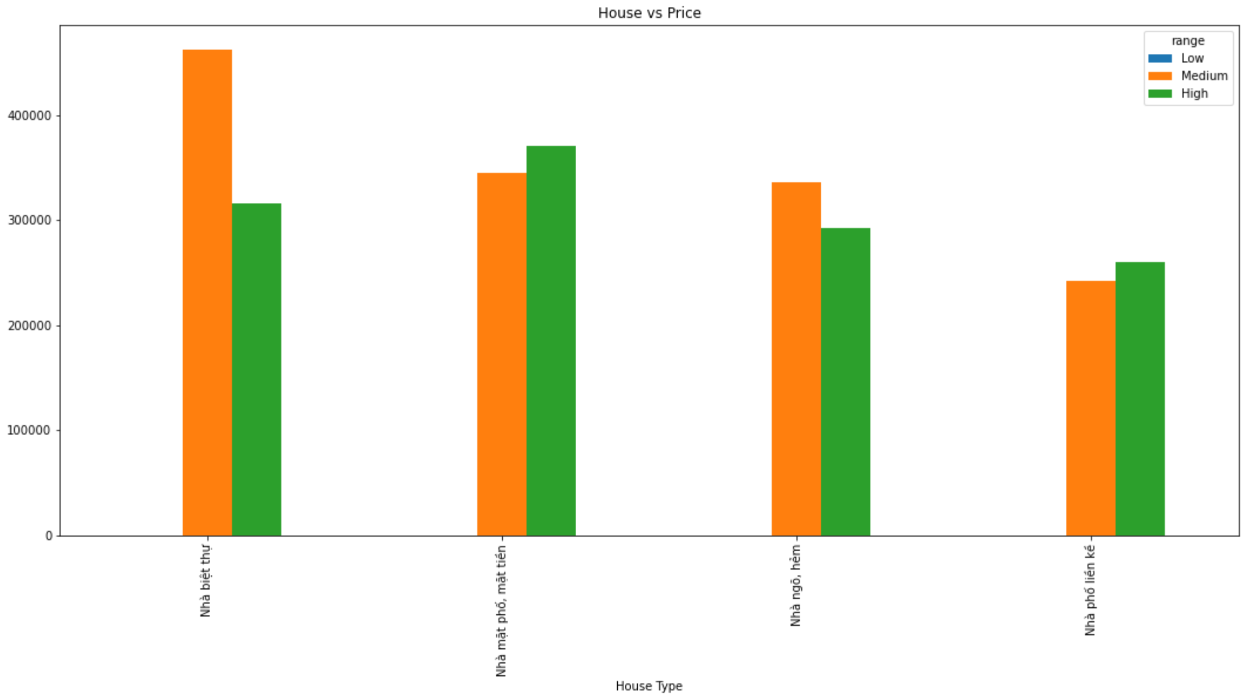
Conclusion: House prices are highest in District 1 and District 2 and lowest in Binh Chanh District. This shows that there is a difference in house prices across counties.

Correlation between furnishing sell and price:



Conclusion: High-end furniture is preferred in high-priced apartments.

Correlation between house type and price:

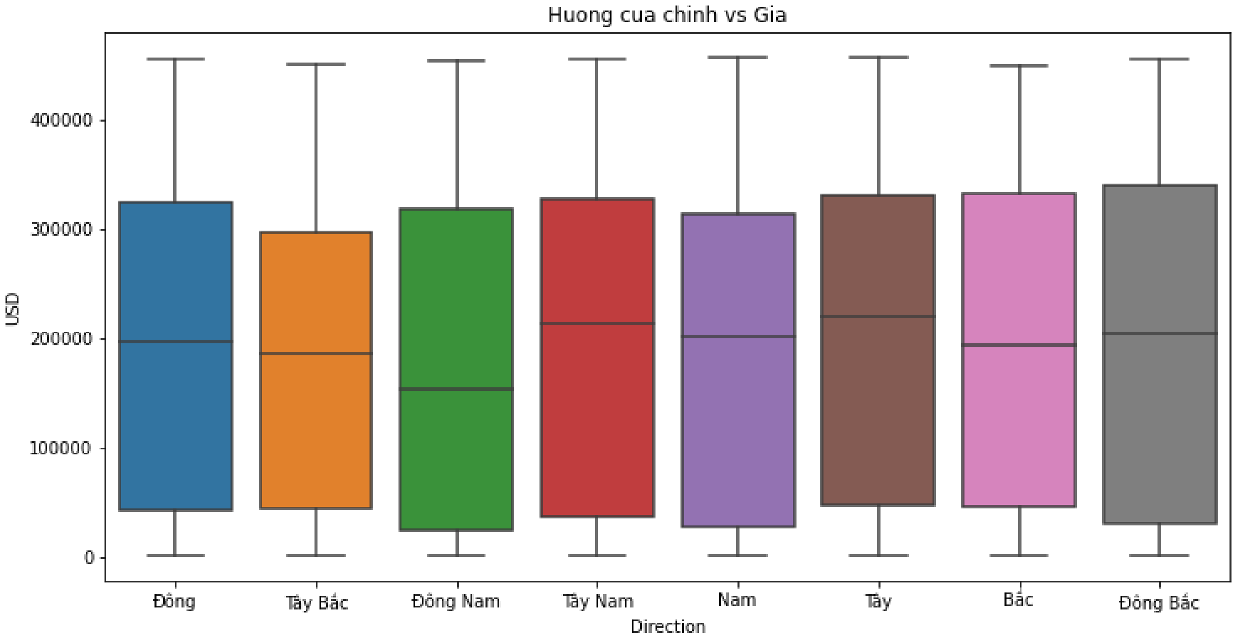


Conclusion:

High-priced apartments are often "street-front houses"

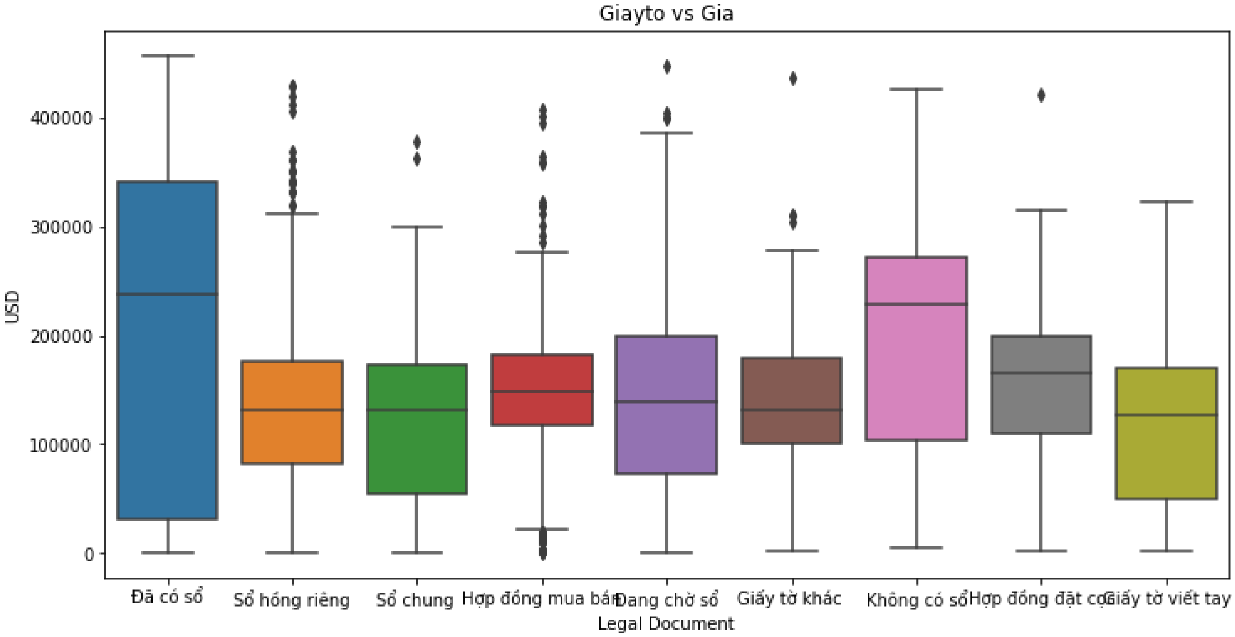
Apartments with average prices are often "villas"

Correlation between direction and price:



Conclusion: There is no big difference between the directions

Correlation between legal document and price:



Conclusion: In general, the average price of "Legal Documents" does not differ too much. "Already have books" accounts for the majority.

IV. Modeling

Because our data still has many NaN values, I decided to delete all columns with NaN above 50% and delete all rows that appear NaN. Because the variables used in the analysis are not completely appropriate. So for the category variables, I'll use Python's pseudo function to separate them. Finally, the variables used in the model are described as follows

Describing the variables included in the model:

Variable name	Description	Data type
Size	Actual area on pink book, unit: million/m2.	Float64
Rooms	Number of bedrooms.	Int32
Toilets	Number of toilets.	Int32
Floors	What floor is the house located on?	Float64
District_Huyện Củ Chi	Yes/No	Float64

District_Huyện Hóc Môn	Yes/No	Float64
District_Huyện Nhà Bè	Yes/No	Float64
District_Quận 1	Yes/No	Float64
District_Quận 10	Yes/No	Float64
District_Quận 11	Yes/No	Float64
District_Quận 12	Yes/No	Float64
District_Quận 3	Yes/No	Float64
District_Quận 4	Yes/No	Float64
District_Quận 5	Yes/No	Float64
District_Quận 6	Yes/No	Float64
District_Quận 7	Yes/No	Float64
District_Quận 8	Yes/No	Float64
District_Quận Bình Thạnh	Yes/No	Float64
District_ Bình Tân	Yes/No	Float64
District_Gò Vấp	Yes/No	Float64
District_ Phú Nhuận	Yes/No	Float64
District_Tân Bình	Yes/No	Float64
District_Tân Phú	Yes/No	Float64
District_ Thành phố Thủ Đức	Yes/No	Float64
Legal Document_Giấy tờ viết tay	Yes/No	Float64
Legal Document_Không có sổ	Yes/No	Float64
Legal Document_Sổ chung	Yes/No	Float64
Legal Document_Đang chờ sổ	Yes/No	Float64
Legal Document_Đã có sổ	Yes/No	Float64
House Type_Nhà mặt phố, mặt tiền	Yes/No	Float64
House Type_Nhà ngõ, hẻm	Yes/No	Float64
House Type_Nhà phố liền kề	Yes/No	Float64

I use StandardScaler to transform data such that its distribution will have a mean value 0 and standard deviation of 1

4.1 Model selection

The models selected for forecasting are: Linear Regression, Decision Tree Regression, Random Forest Regression.

4.1.1 Linear Regression

It is an algorithm that is used for estimating the real values (cost of houses, number of calls, complete deals and so forth) in view of continuous variable(s). Here, we try to find a best fit line which can get us the relationship between independent and dependent variables.

4.1.2 Decision Tree Regression

It is a tree-based model and is a supervised learning algorithm which can be used regression models here the nodes are decision points having conditions the results of which then extends the tree into more nodes

4.1.3 Random Forest Regression

Forest is a kind of democratic collection of many decision trees, where to tackle the problem of overfitting of a single Decision tree we now do voting, and the most voted class wins and is the result for your target observation.

4.2 Evaluation

4.2.1 MAE

Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as the sum of absolute errors divided by the sample size. (Wikipedia)

$$MAE = \frac{\sum_{t=1}^{n} \left| \varepsilon_t \right|}{n} = \frac{\sum_{t=1}^{n} \left| Y_t - \hat{Y}_t \right|}{n}$$

In our case these continuous variables are listing price value and predicted price value of the house property.

4.2.2 MSE

In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate. (Wikipedia)

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

4.3.3 RMSE

MSE sometimes increases the actual error, making it difficult to realize and understand the actual error amount. This problem is resolved by the RMSE measure, which is obtained by simply taking the square root of MSE.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

The goal to choose the best model is RMSE and MSE, the smaller the MAE, the smallerthe MAE because they said errors are the differences between the predicted values (values predicted by our regression model) and the actual values of a variable.

4.3 Model comparison with RMSE, MSE, MAE

The goal to choose the best model is RMSE and MSE, the smaller the MAE, the smaller the MAE because they said errors are the differences between the predicted values (values predicted by our regression model) and the actual values of a variable.

MODEL	MAE	MSE	RMSE
Linear	**170627**	**56236334071**	**237142**
Decision Tree	**7395**	**1382022616**	**37175**
Random Forest	**58306**	**8752429250**	**93554**

As in the previous discussion the evaluation ratio of each model is equal to its evaluation MSE, MAE, RMSE. The smaller evaluation ratio, the higher accuracy of the model’s prediction.Table shows that, when applied to test data, Decision Tree Regression outperforms Random Forest and linear regression in terms of prediction accuracy.

It can be concluded that Decision Tree Regression is the best model to forecast house prices using this dataset

V. Conclusion

This research paper's major objective is to assist home sellers and purchasers in avoiding being overpriced or underpriced in light of machine learning models. Based on empirical evidence, Decision Tree Regression is the most effective model for predicting property prices in Ho Chi Minh City, as evidenced by its minimal RSME, MAE, and MSE values. The majority of the forecast's variables are important for projecting home values. When it comes to forecasting, random forest and linear models are not very accurate.

Besides, the price of a house posted on Cho Tot sometimes does not reflect its true value and inadvertently causes this assessment to be overvalued or undervalued. The fact that when buying/selling a house depends on many other situations such as: land price fluctuations, how is the real estate market, etc. The results of this study are for reference only. However, this will be the most basic thing when you want to buy / sell a house.

REFERENCES

[1] Thamarai, M., Malarvizhi, S. J. I. J. o. I. E., & Business, E. (2020). House Price Prediction Modeling Using Machine Learning. *12*(2).

[2] Zaki, J., Nayyar, A., Dalal, S., Ali, Z. H. J. C., practice, c., & experience. (2022). House price prediction using hedonic pricing model and machine learning techniques. *34*(27), e7342.

[3] Sharma, M., Chauhan, R., Devliyal, S., & Chythanya, K. R. (2024). *House Price Prediction Using Linear and Lasso Regression.* Paper presented at the 2024 3rd International Conference for Innovation in Technology (INOCON).

[4] Verma, A., Nagar, C., Singhi, N., Dongariya, N., & Sethi, N. (2022). *Predicting House Price in India Using Linear Regression Machine Learning Algorithms.* Paper presented at the 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM).

[5] Chen, K., Lu, W., & Yan, Y. (2024). *Write A Code Using Linear Regression and Neural Layered Structure To Predict The House Price.* Paper presented at the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023).

[6] Hùng, M. (2022). BÁO CÁO ĐỒ ÁN CUỐI KÌ - KHOA HỌC DỮ LIỆU ỨNG DỤNG.[online] GitHub. Available at: https://github.com/HungTrinhIT/FinalProject-Datascience .