Housing Price Linear Regression

Le Ngoc Duy, Le Duc Khang, Le Minh Ngoc,
Nguyen Nhat Minh, Doan Van Nguyen

1 Introduction

This study uses a dataset sourced from alonhadat.com.vn, one of the leading real estate platforms in Vietnam. The dataset encompasses diverse property listings across Da Nang, capturing key details that influence pricing. It includes data on property features, such as size, type, and structural specifics, as well as locational attributes like district.

Accurate predictions of housing prices play a vital role in supporting buyers, investors, and policymakers in making informed decisions. House prices depend on various factors such as property characteristics, location all of which can significantly impact the housing market's overall properties. Predicting these prices not only aids individuals in finding desirable properties but also assists economists in understanding broader market dynamics and supports government agencies in implementing timely real estate policies.

Leveraging this dataset allows us to develop a predictive model that can account for the complex interactions between property characteristics, location, and market dynamics. By applying multiple linear regression and other machine learning algorithms, this study aims to establish a reliable model for estimating housing prices in Da Nang, contributing to a better understanding of local market trends and serving as a practical tool for potential buyers and investors.

2 Data

2.1 Data

After thoroughly reviewing and exploring several real estate websites, our team decided to select data from *alonhadat.com.vn.*The data for this study was sourced from *alonhadat.com.vn*, a popular real estate website in Vietnam, providing comprehensive information on the real estate market in Da Nang. The dataset includes house prices and various detailed attributes of each property, grouped into the following key categories:

1. **Property Information**: This includes usable area (in m^2), number of bedrooms, bathrooms, and floors. The dataset also categorizes property types, such as townhouses, villas, apartments, or land plots, allowing the model to differentiate among property types when predicting prices.
2. **Location Factors**: Data on the geographic location of each property is included, specifying the district, ward, and street within Da Nang. Proximity to public amenities such as schools, hospitals, shopping centers, and transport hubs is also recorded, helping to analyze how these factors influence housing prices.

2.2 Method

The data here is being scraped using a general-purpose web crawler Selenium. The dataset adopts real estate data from four different websites in Da Nang scatter in all districts and we have collected over 10000 rows of information throughout four websites. However, the data collected isn't synchronized and then needed further clarification and modification.

The retrieved data is being reviewed and then reorganized the columns by using Python. Some libraries are being implemented to help in this process, such as Geopy, Mathplotlib, Numpy, Pandas, PymySQL, Seaborn. The data is clarified by finding generic columns then remove the unnecessary ones, leaving 'Prices', 'Area', 'toFace', 'Type', 'Certificate', 'Width', 'Length', 'Floors', 'Latitude', 'Longtitude', 'Street', 'Ward', 'District', 'Distancetobeach', 'Distancetocenter' and 'Distancetoairport'.

3 Metrics and Procedure

3.1 Metrics

The evaluation of the regression models used in housing price prediction relies on several key metrics that measure the accuracy and reliability of the predictions. The first and primary metric used is the Mean Squared Error (MSE), which calculates the average squared difference between the actual housing prices and the predicted values generated by the model. MSE is expressed as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. MSE provides a comprehensive view of the model's performance by penalizing larger errors more heavily, which can highlight issues if the model is far off in certain predictions. Additionally, the Root Mean Squared Error (RMSE) is employed, as it is simply the square root of MSE and thus presents the error in the same unit as the housing price, making it more interpretable:

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

Mean Absolute Error (MAE) is also utilized, as it calculates the average absolute difference between predicted and actual prices without squaring the errors, reducing the impact of outliers on the overall error measurement and offering a straightforward interpretation of average deviation from actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

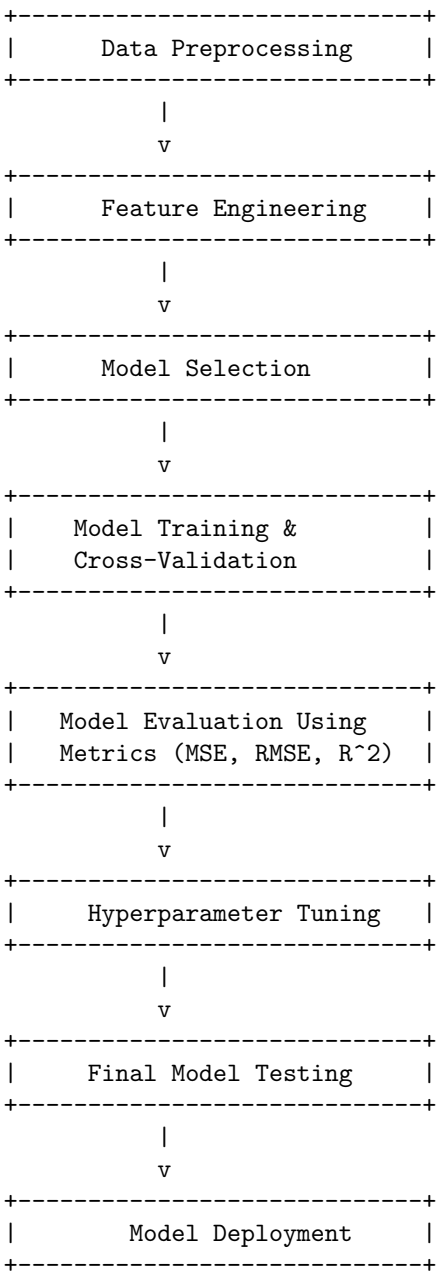
Another critical metric is the R^2 value, also known as the Coefficient of Determination, which represents the proportion of variance in the dependent variable (house price) that can be explained by the independent variables in the model. An R^2 value close to 1 indicates a high level of explanatory power by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

where $\bar{y}$ is the mean of the observed data.

3.2 Procedure

The modeling procedure for predicting housing prices involves a series of carefully structured steps, beginning with data preprocessing, followed by feature engineering, model training, and finally, model evaluation. The following flowchart illustrates these steps in the process:


Data Preprocessing
Feature Engineering
Model Selection
Model Training &
Cross-Validation
Model Evaluation Using
Metrics (MSE, RMSE, R^2)
Hyperparameter Tuning
Final Model Testing
Model Deployment

Initially, the housing dataset undergoes exploratory data analysis to understand variable distributions, outliers, and missing values, which are addressed to ensure data consistency. Variables with missing values, such as `LotFrontage` and `GarageCars`, are either imputed or removed if they do not significantly impact the model. Outliers that could skew results are identified and removed to maintain model accuracy. Scaling and normalization of features are also performed to standardize the dataset. Following data preprocessing, feature engineering is conducted to enhance the model's predictive power. New features are created by combining or transforming existing ones, particularly those with skewed distributions, which are corrected using logarithmic transformations. One-hot encoding is applied to categorical variables to enable the model to interpret these features in a numerical format.

With the data prepared, a baseline multiple linear regression model is implemented as a reference. Further models, including Ridge and Lasso regression, are then explored, leveraging regularization to mitigate overfitting. Cross-validation, specifically K-fold cross-validation, is employed during model training to ensure a reliable model, with the dataset divided into K subsets, allowing each subset to serve as a validation set in rotation. Model performance is rigorously evaluated using MSE, RMSE, MAE, R^2, and Adjusted R^2, and hyperparameters are fine-tuned using grid search to optimize results. The final model is then tested on a held-out test set, ensuring reliability, and prepared for deployment.

4 Results & Analysis

Firstly, we will analyze the price data. From Fig. 2, we can see that there is a right-skewed normal distribution with the majority of data clustering around 6-8 billion VND. Since the outliers were too much, we decided to place a cap on the prices for stability and avoid the impact of these outliers, although this would have some consequence on the higher end of values. The average price was 7.8 billion VND and the standard deviation was relatively large(11 billion VND for the non-capped prices and 7.4 for the capped prices).

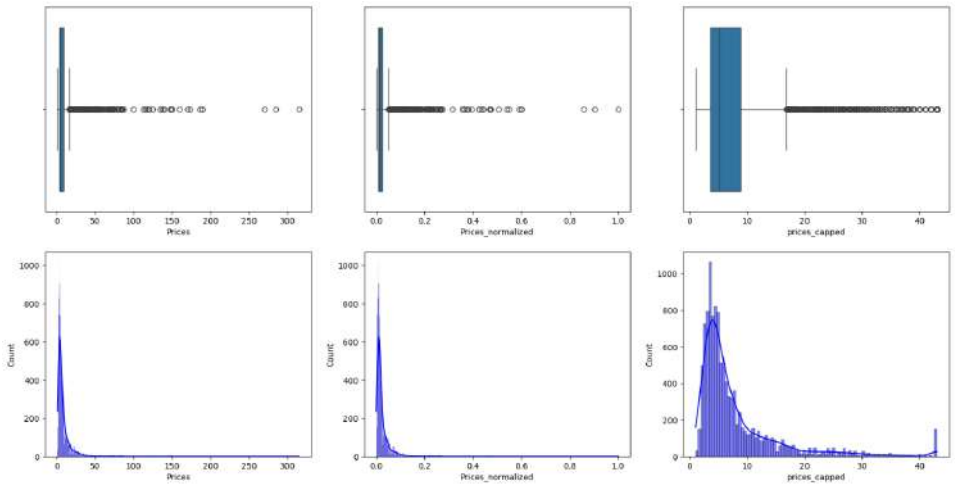


Fig. 1: Price distribution

This same approach was then used for the other numerical values where the outliers were too much to handle(shown in box-plot). Also, there are various spikes at the area, width, length attributes and they do not follow a normal distribution.

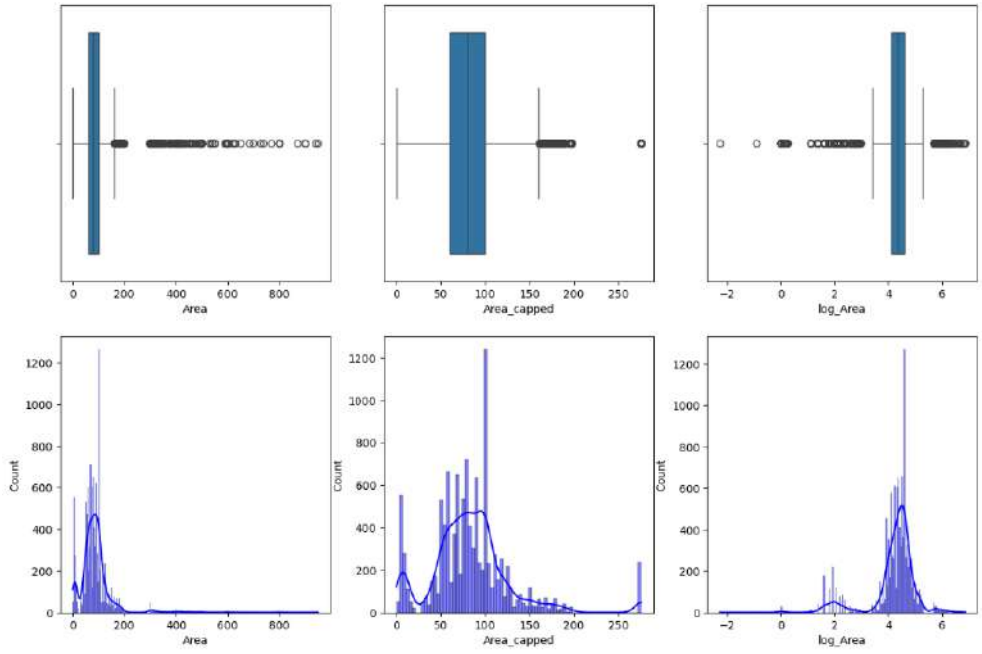


Fig. 2: Area distribution

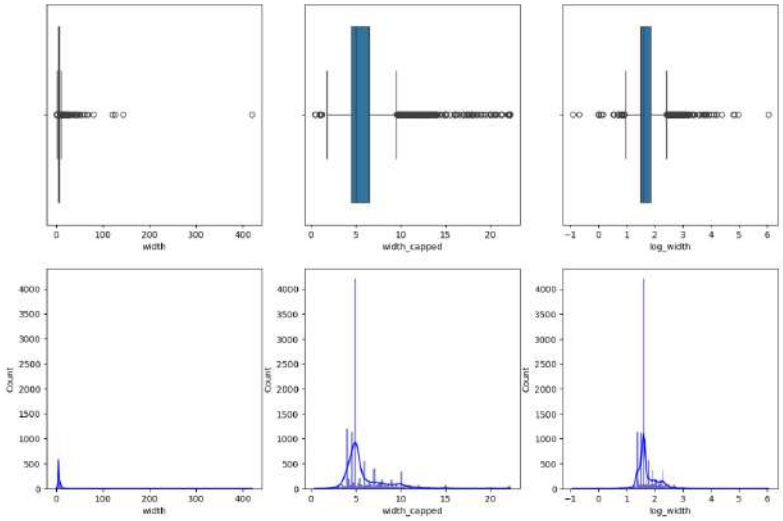


Fig. 3: Width distribution

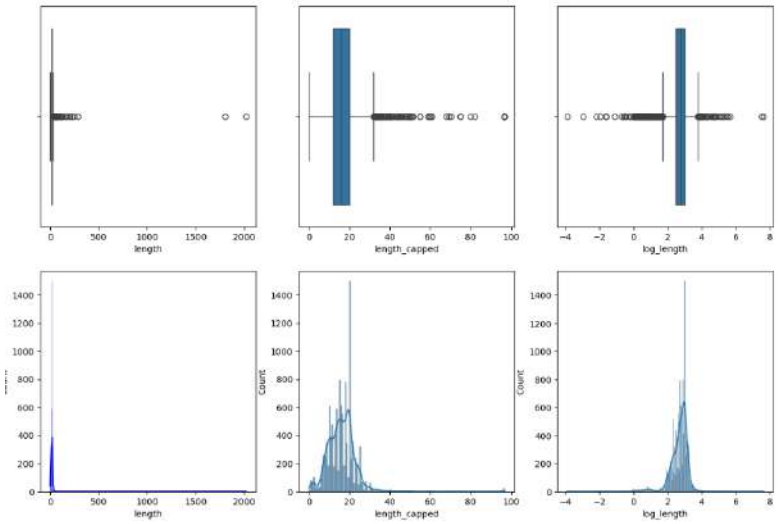


Fig. 4: Length distribution

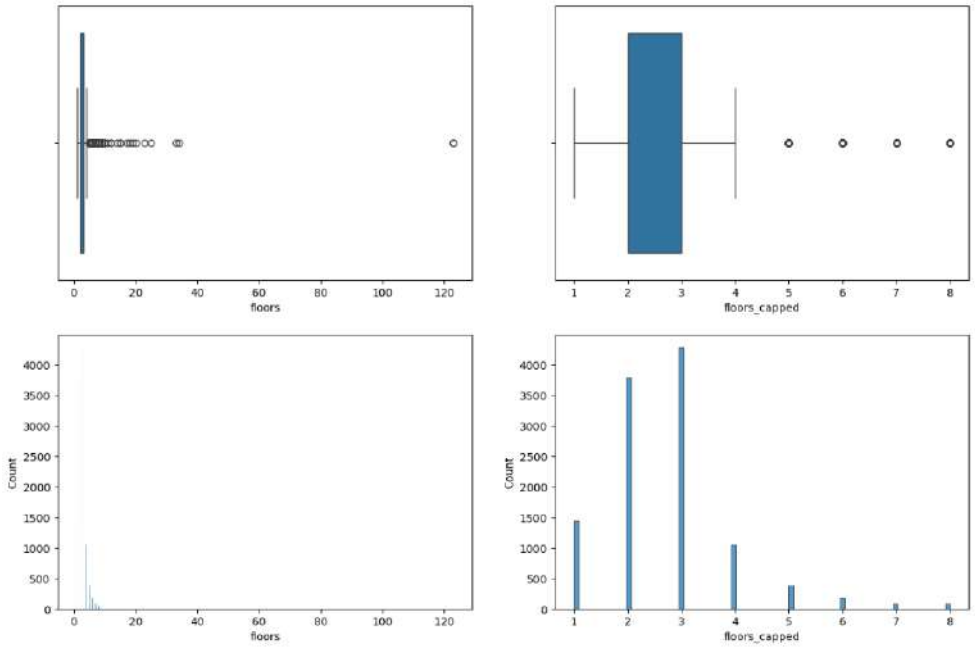


Fig. 5: Floors distribution

Since this project focuses more on houses, we doubt that there would be a 120-story house, thus we decided to cap it as well.

However, the scatter plot of these values when put up with price seems to not distribute well, this might be due to our data.

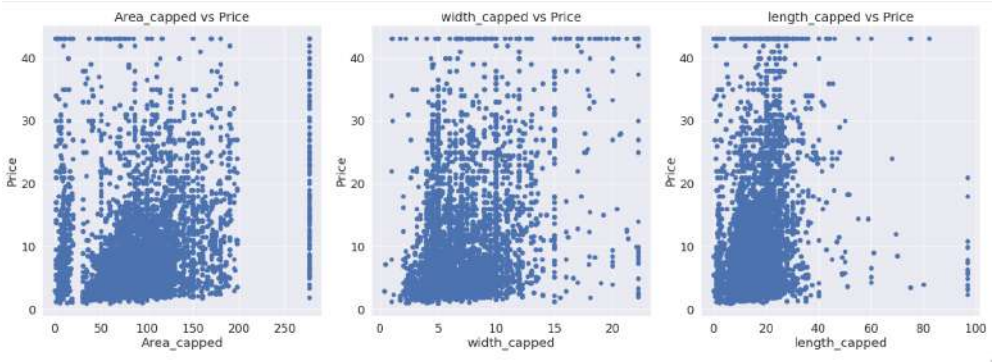


Fig. 6: Scatter plot of price vs other aspects

Regarding the categorical values, there were some notable aspects as well. As some of them seem to have an influence on the price.

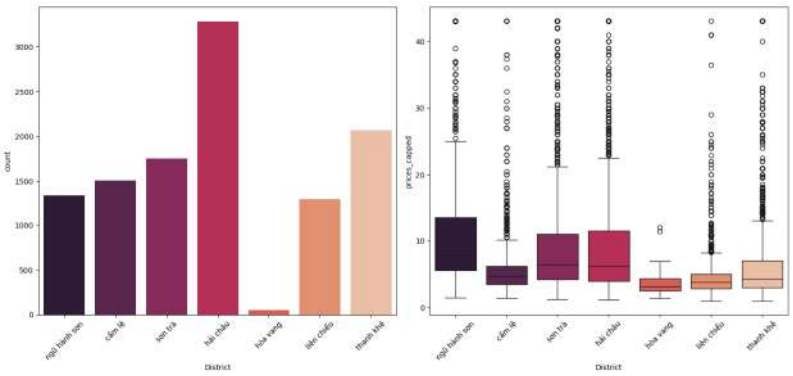


Fig. 7: Districts and prices based on districts

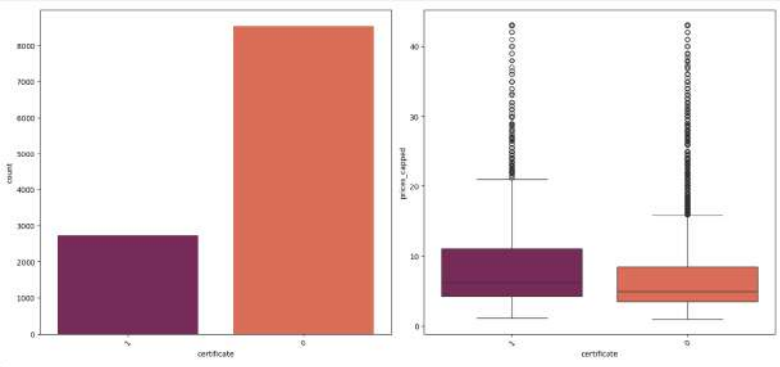


Fig. 8: Prices based on having certificate

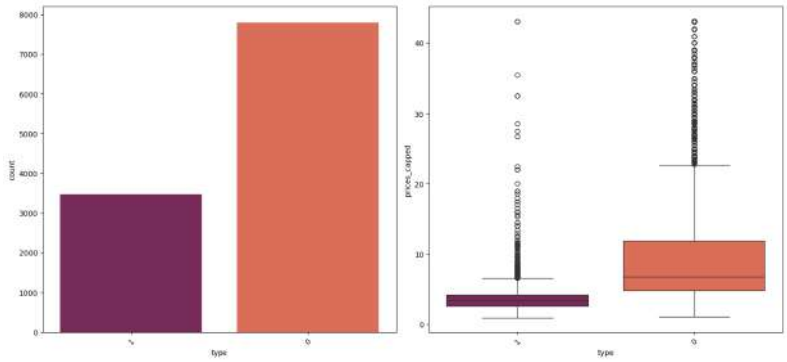


Fig. 9: Prices based on type of house(alley/front)

The categorical values have been mapped and put to use in the model since they have an impact on price as prices were higher on the houses in differing areas, differing type of houses.

For further analysis, let us look at the correlation matrix to see how the distinct column variables all correlated to each other, Fig. 11 will show the correlation matrix of independent and dependent variables. It can be seen that they do have some relationship with each other, confirming that the dependent variables have a good correlation with the price of houses and they do affect the price trend.

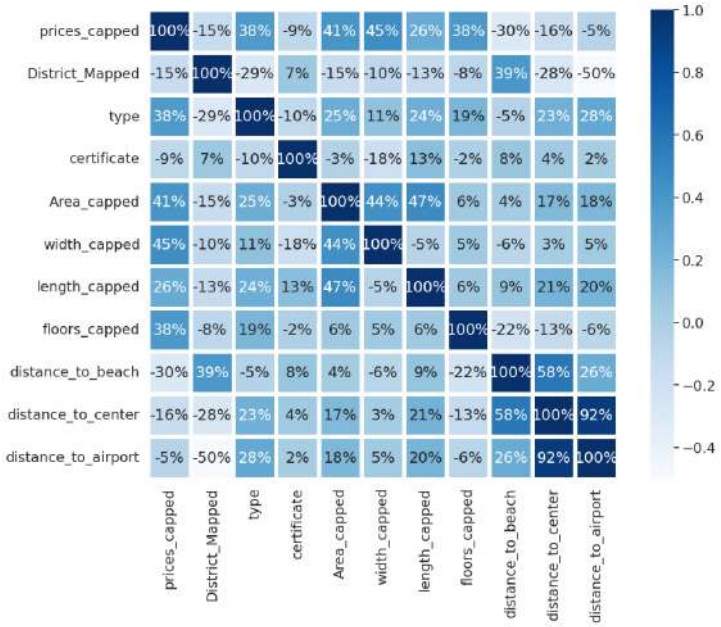


Fig. 10: Correlation

4.1 Prediction results of the model

In this project, we have used the scikit-learn library to split the data into training and testing sets then fit the data. After importing the training data and configuring the model, the model is instantiated and trained, then it is tested with the testing set. The observation of this model is plotted in Fig. 12 and Fig. 13 using differing methods to scale the data.

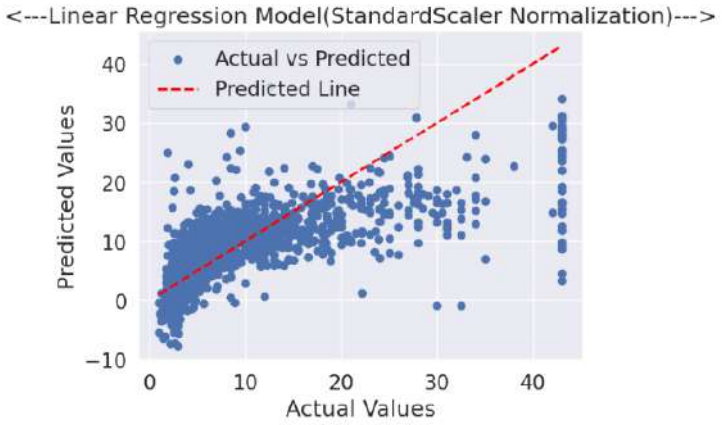


Fig. 11: Comparison between predicted and true values(StandardScaler)

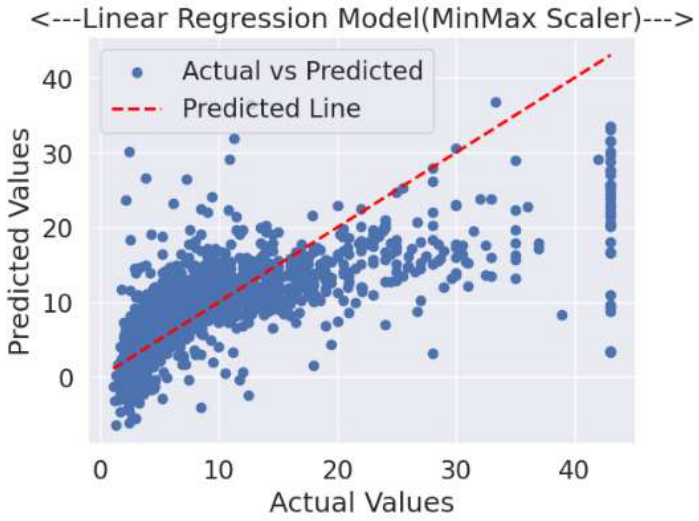


Fig. 12: Comparison between predicted and true values(MinMaxScaler)

From what we can observe, there was definitely trade off at higher end of the values, this has already been mentioned above. However, we could say that the lower end did pretty well as a large cluster is there and the model captured a good portion of it. But to apply to real-world scenarios might not be suitable.

As for specific value on how much of the variability the model captured and how much the predicted values deviate from the true value, we have calculated the MAE, MSE, RMSE, and R^2 value. As for the model that used standard scaler, the results are the following:

Mean Absolute Error (MAE):	3.32
Mean Squared Error (MSE):	27.98
Root Mean Squared Error (RMSE):	5.29
R-squared (R^2):	0.48

As for the model with min max scaler, the results are relatively the same:

Mean Absolute Error (MAE):	3.27
Mean Squared Error (MSE):	26.77
Root Mean Squared Error (RMSE):	5.17
R-squared (R^2):	0.51

4.2 Regression Analysis

4.2.1 Cross-Validation

To better analyze the regression, first, we have applied cross-validation to observe how stable the model is. The results are as follow:

	R2 Score	RMSE Score	MAE Score
Fold 1	0.5441	5.1329	3.2751
Fold 2	0.5680	5.0645	3.1971
Fold 3	0.5080	5.3520	3.2962
Fold 4	0.5402	5.0263	3.2730
Fold 5	0.5613	4.7878	3.1618
Fold 6	0.5078	5.0029	3.3214
Fold 7	0.5407	4.9518	3.2702
Fold 8	0.5133	5.2681	3.2183
Average Score	0.5354	5.0733	3.2517

Table 1: Cross-Validation Results for R2, RMSE, and MAE Scores for model with Standard Scaler

So we can say that they perform quite stable as it did not deviate too much from the original scores calculated in the section above.

4.2.2 Linear Assumptions

After that, we proceeded to check for the linear assumptions to make sure that the model's results are reliable and see if it might have produce biased/inefficient estimates.

	R2 Score	RMSE Score	MAE Score
Fold 1	0.5409	5.1224	3.1747
Fold 2	0.5171	5.5719	3.4675
Fold 3	0.5099	4.7241	3.1682
Fold 4	0.5024	5.0578	3.1953
Fold 5	0.5184	5.1179	3.2897
Fold 6	0.5469	4.9508	3.2947
Fold 7	0.5369	5.4547	3.4232
Fold 8	0.5576	4.7456	3.1442
Average Score	0.5288	5.0931	3.2697

Table 2: Cross-Validation Results for R2, RMSE, and MAE Scores of the model with MinMax Scaler

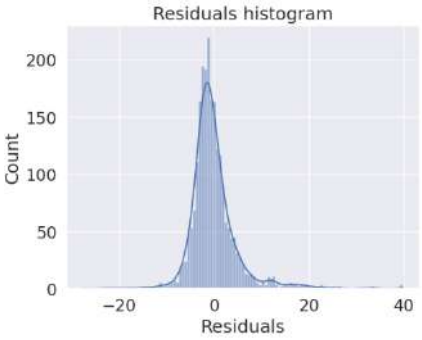


Fig. 13: Residuals histogram

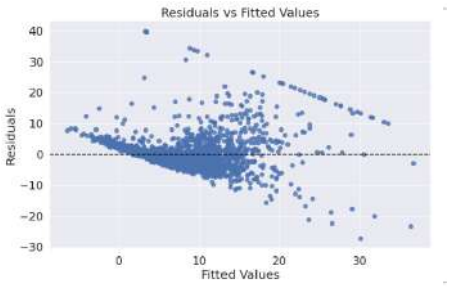


Fig. 14: Residuals analysis

From Fig. 15, it is observed that there might be a slight pattern here, which is evidence that there are some non-linear relationships. In addition, this leads to violating the homoscedasticity since there is not a constant variance of errors. But from Fig. 14, we say that there is normality of errors, which is a good sign despite some skewness and outliers to the right. The residuals may not be perfectly normal but is close enough for this dataset.

Features	**VIF Factor**
District_Mapped	6.147777
type	3.915470
certificate	4.021829
Area_capped	7.537361
width_capped	6.735955
length_capped	8.324573
floors_capped	3.039810
distance_to_beach	31.113650
distance_to_center	155.320217
distance_to_airport	103.937400

Table 3: VIF Factors for Features

From Table 3, we can see that the VIF factor for distancing attributes is quite high as they capture spatial information, maybe we could transform or delete a columns to reduce multicollinearity. Other than that, the rest follows a low and moderate VIFs factor.

4.2.3 ANOVA and t test

Regarding ANOVA test, after dividing the group with the categorical values, in this case is the *District* column. With the scipy.stats library, we have ran the one way ANOVA test with $foneway()$, we recorded a really low p-value but upon further reading, there are assumptions that must be satisfied between each groups so we decided to move on to the $kruskal()$ function which is possible to test even without the assumptions having to be true. The kruskal test returned a p-value of 0.0 so these tests suggest a statistically significant difference in mean/median prices across the districts.

In the case of the t-test, we conducted the t-test in respect to the type of house(being in the alley or not), it returned a p-value of 0, suggesting a significant difference in prices between front houses and the alley ones. We also conducted the t-test with respect to having certificate or not and again, the p-value was $2.13e-19$,which suggests the same thing for houses with certificate or not

5 Limitation and Prospects

This study has certain limitations. Firstly, the research did not employ more extensive prediction mechanisms, which may result in better prediction accuracy. Secondly, the dataset used in this paper focuses on Da Nang housing and cannot represent the trend of the whole real estate in Da Nang. Thirdly, social and natural factors, which can have an impact on housing prices every year, make the model not always applicable to the housing price prediction of all cities.

In the future, more advanced prediction models could be used to train and test the housing price prediction model. Moreover, a broader dataset should be used to predict house prices in different cities. Furthermore, data should be updated dynamically, and the latest datasets should be adopted to collect big data to achieve better real-time dynamic housing price prediction.

6 Conclusion

This study investigated a better prediction model for Da Nang housing price prediction based on linear regression and machine learning, and verified the reliability of the model. The study selected the Da Nang housing price dataset and used supervised multiple linear regression and machine learning algorithms to predict real estate prices. After transforming the data into a tensor, the price data was analyzed using histograms to confirm its availability. To further analyze the relationship between independent and dependent variables, a matrix graph was used for correlation analysis, and abnormal and missing data were optimized. Finally, the model was trained and tested. The experimental results showed that the real housing price scatterplots are clustered and

distributed on both sides of the predicted housing price, further proving the reliability of the proposed model.

However, this paper also has certain limitations. Other algorithms were not used for comparison, and the universality of the model was not further verified. In the future, more advanced deep learning models and more extensive and dynamically updated datasets should be adopted for research to achieve real-time stable prediction of housing prices. The main significance of this study is to provide suggestions for housing price prediction and provide a better reference for investors. Overall, these results offer a guideline for providing a better multiple regression model for the prediction of housing prices and give buyers a better reference.