Price Advisor

The Model for assessing real estate prices using data from amenities, location, convenience, public transportation, etc.

Problem Statement

Many homeowners face challenges in accurately valuating their properties when looking to sell. The absence of reliable and data-driven valuation tools often leads to overpricing or underpricing, potentially resulting in delayed sales or missed opportunities. There is a need for an accessible and accurate tool that empowers homeowners to set optimal asking prices and maximize their returns within a reasonable time frame.

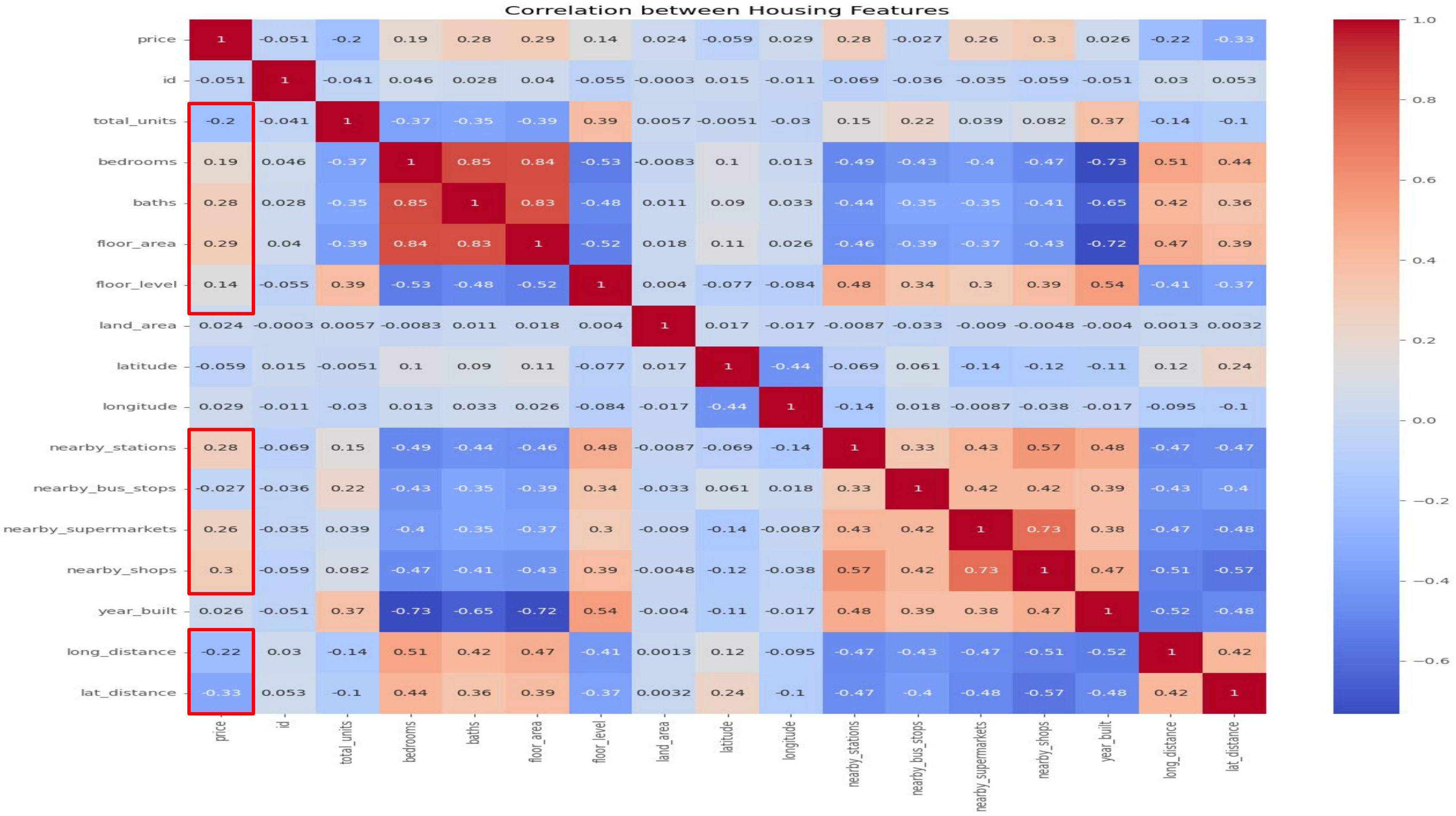
Questions:

1. How can the Model ensure accurate property valuations while considering various factors like location, size, facilities, and market trends?
2. What features and functionalities should be incorporated to create a user-friendly experience for homeowners seeking property valuations?

Data : Bangkok, Nonthaburi, Samut prakan Price Housing

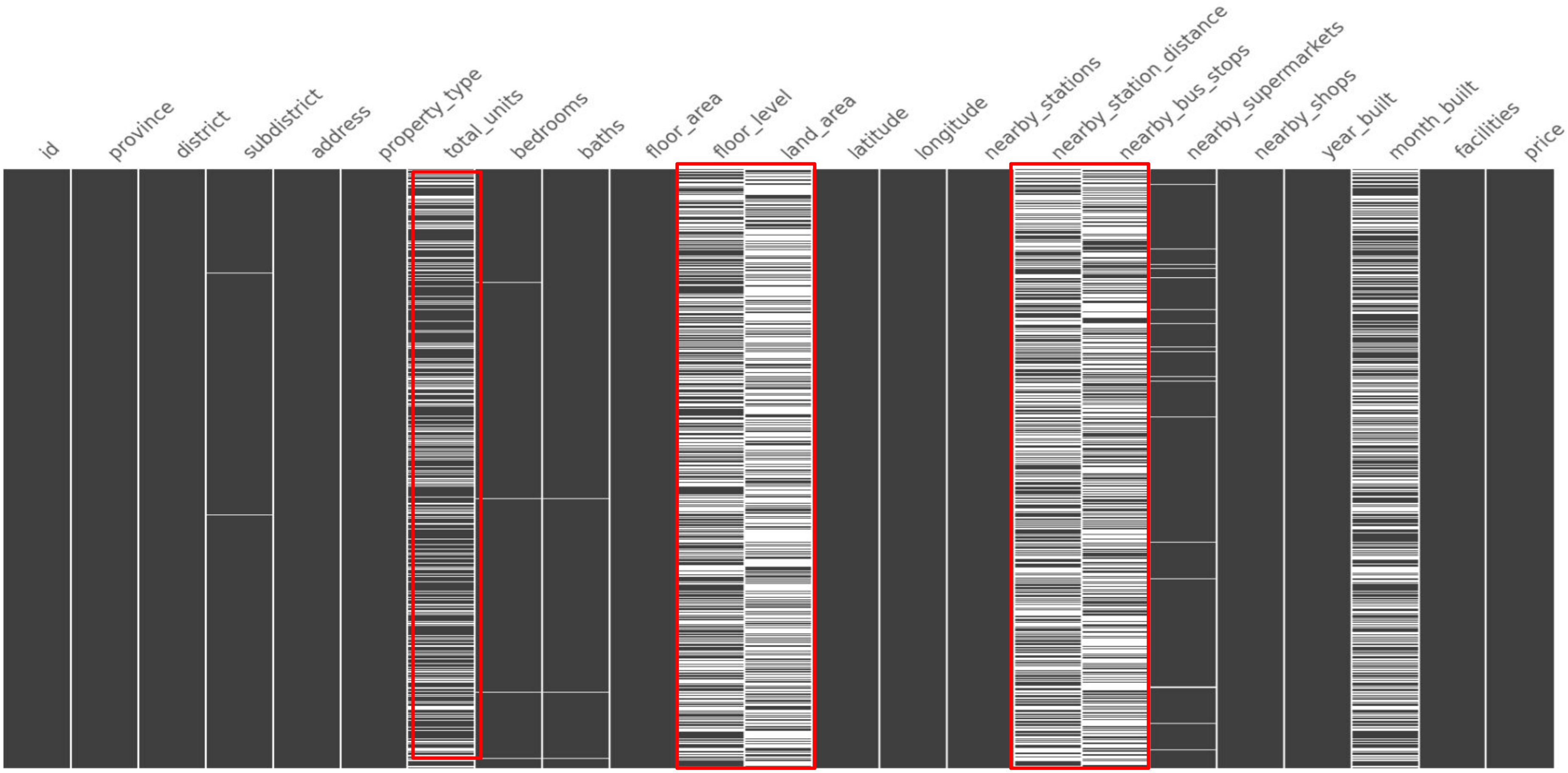
id	int	train.json	ID of selling item
province	string	train.json	province name: this dataset only includes Bangkok,Samut Prakan and Nonthaburi
district	string	train.json	district name
subdistrict	string	train.json	subdtistrict name
address	string	train.json	address e.g. street name, area name, soi number
property_type	string	train.json	type of the house: Condo, Townhouse or Detached House
total_units	float	train.json	the number of rooms/houses that the condo/village has
bedrooms	int	train.json	the number of bedrooms
baths	int	train.json	the number of baths
floor_level	int	train.json	floor level of the room
floor_area	float	train.json	total area of inside floor [m^2]
land_area	float	train.json	total area of the land [m^2]
latitude	float	train.json	latitude of the house
longitude	float	train.json	longitude of the house
nearby_stations	string	train.json	district name
nearby_station_distance	list	train.json	list of (station name, distance[m]). Each station name consists of station ID, station name, and Line such as "E4 Asok BTS"
nearby_shops	int	train.json	the number of nearby shops
nearby_supermarkets	int	train.json	the number of nearby supermarkets
nearby_shops	int	train.json	the number of nearby shops
year_built	int	train.json	year built
month_built	string	train.json	month built: January-December
long_distance	float	train.json	The distance(longtitude) from the building to the median point
lat_distance	float	train.json	The distance(lattitude) from the building to the median point
price	float	train.json	TARGET VALUE selling price

Exploratory Data Analysis and Cleaning Data



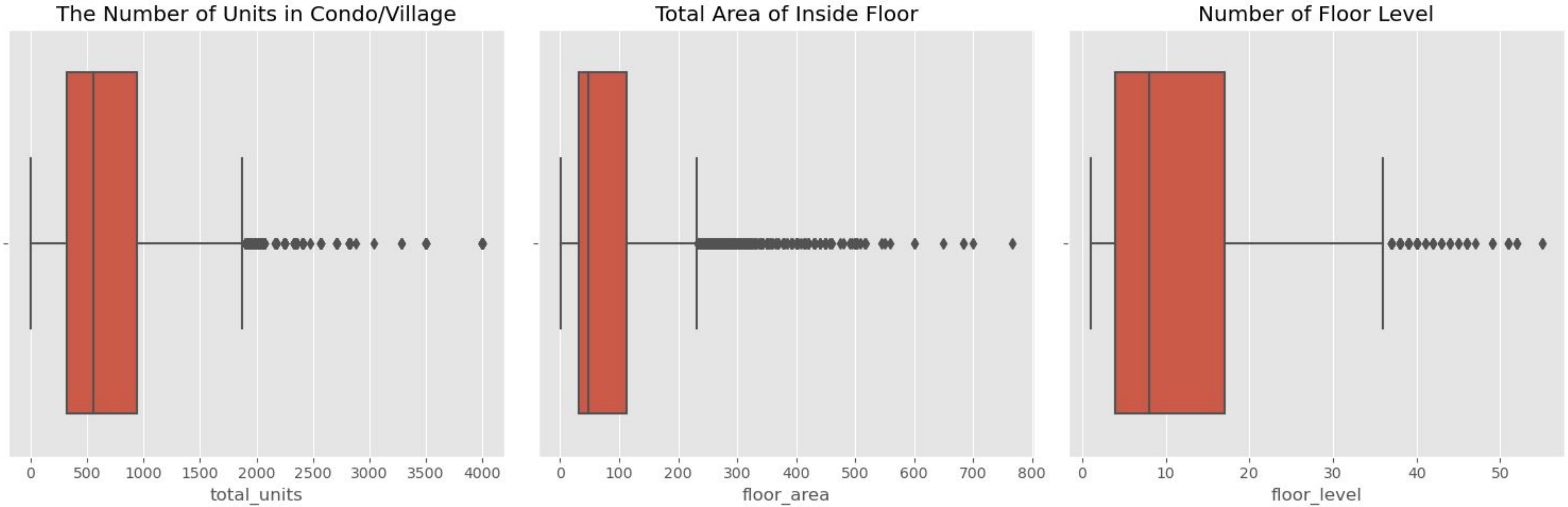
The relationships between the data, it was observed that features such as Total unit, bedrooms, baths, nearby_stations, floor_area, nearby_shops, floor_level, long_distance, and lat_distance have an influence on the price. However, this influence is not considered strong, with an average correlation of approximately 0.2-0.3

Exploratory Data Analysis and Cleaning Data



Exploratory Data Analysis and Cleaning Data

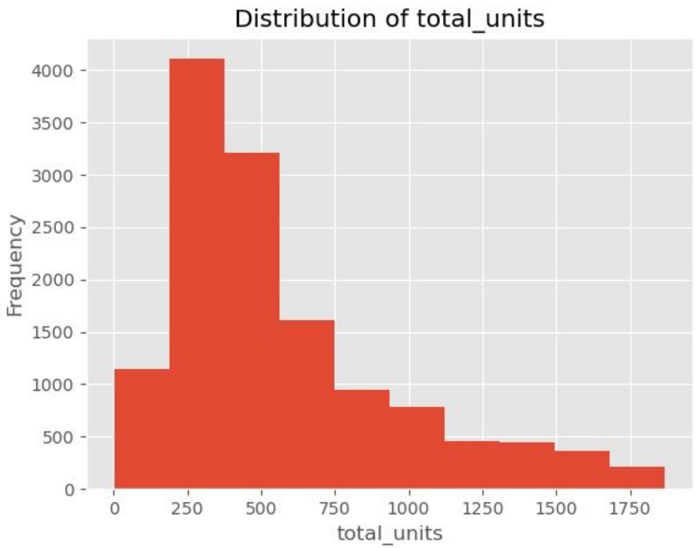
found that there are 3 columns that need to drop outliers.



CLEANING DATA

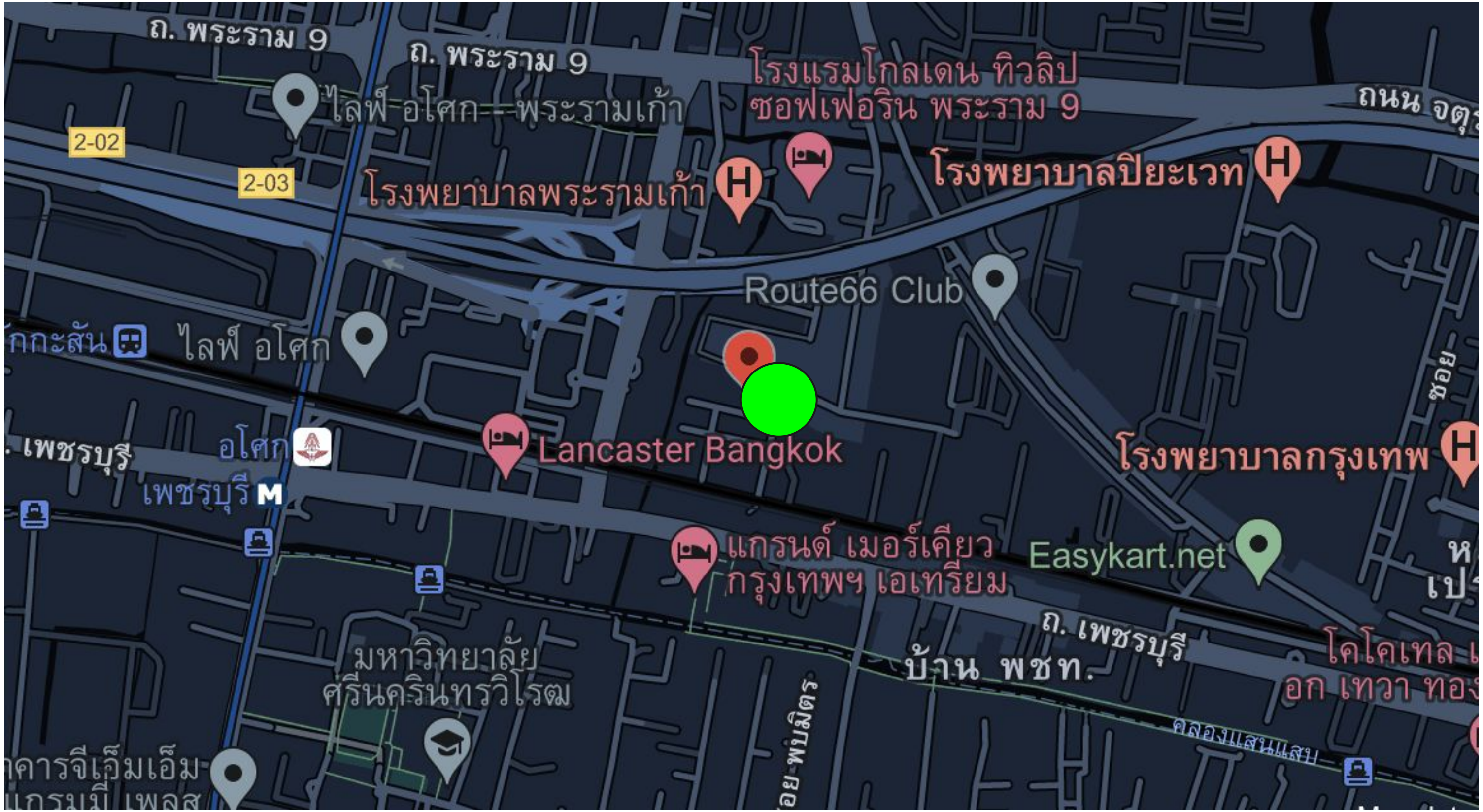
we have null values in several columns. I have decided to replace them with the average value for each property type.

- Number of Total unit
- Number of Bedroom
- Number of Bathroom
- Number of Floor level
- Number of Near by Supermarkets

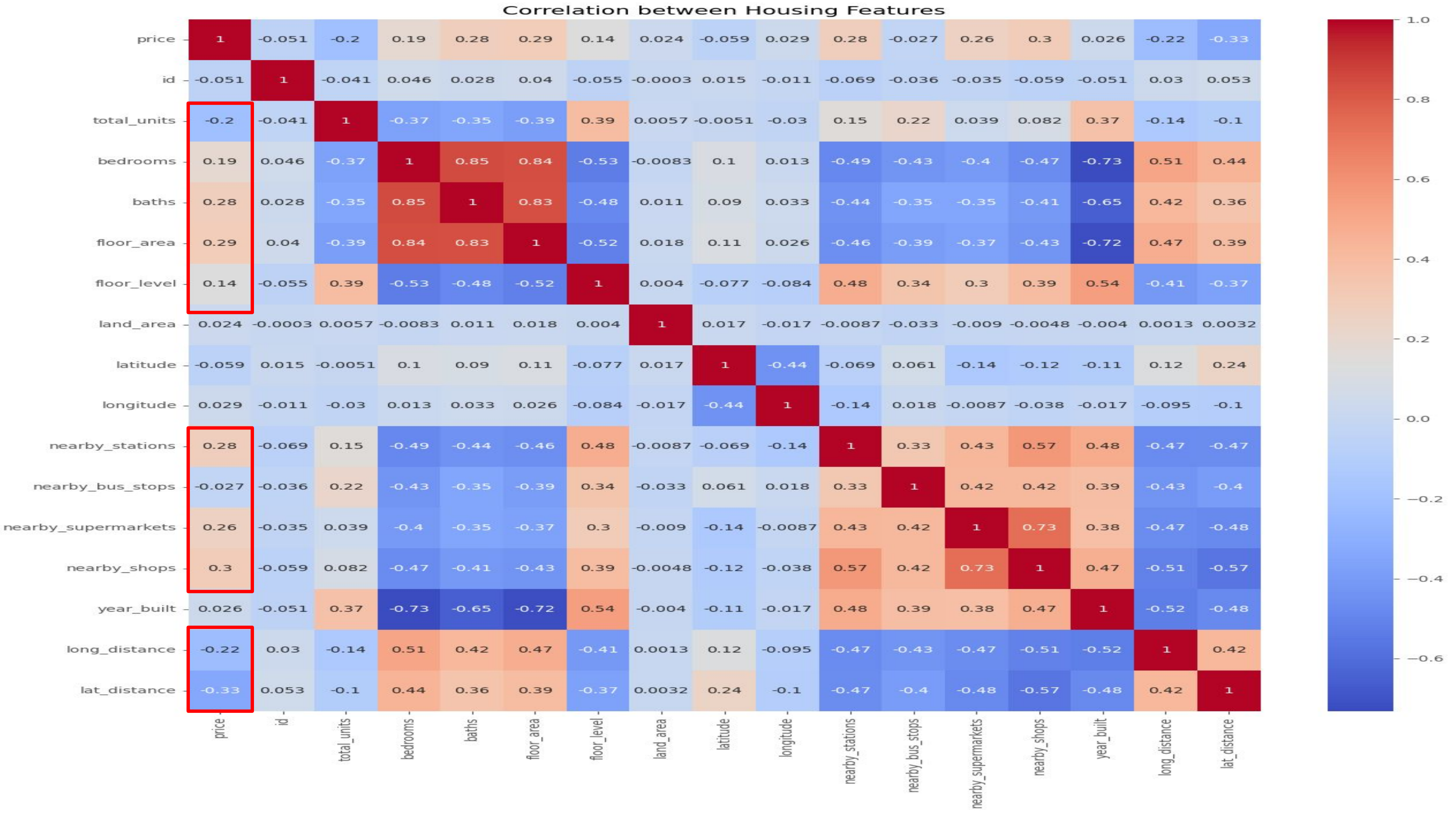


**The replacement should not significantly alter the distribution of the original data.

Median of Latitude and Longtitude



Exploratory Data Analysis and Cleaning Data



Model Preprocessing

I chose the features that the heatmap showed some degree of correlation, even though it's not very strong

Numeric data : Number of Bedroom , Number of Bathroom , Number of nearby station , Floor area , Number of Floor level, Number of nearby shop , Number of nearby station, Long distance , Lat distance

bedrooms	baths	nearby_stations	floor_area	nearby_shops	floor_level	long_distance	lat_distance
2.0	2.0	2	66	20	10.000000	0.013664	0.028139
1.0	1.0	3	49	20	8.000000	0.004237	0.008179
1.0	1.0	2	34	20	4.000000	0.005526	0.024688
3.0	3.0	0	170	4	1.752613	0.142748	0.071604
3.0	2.0	1	120	15	1.695214	0.077057	0.115766

Model Preprocessing

I chose the features that the heatmap showed some degree of correlation, even though it's not very strong

Categorical data : Province, District, Property type

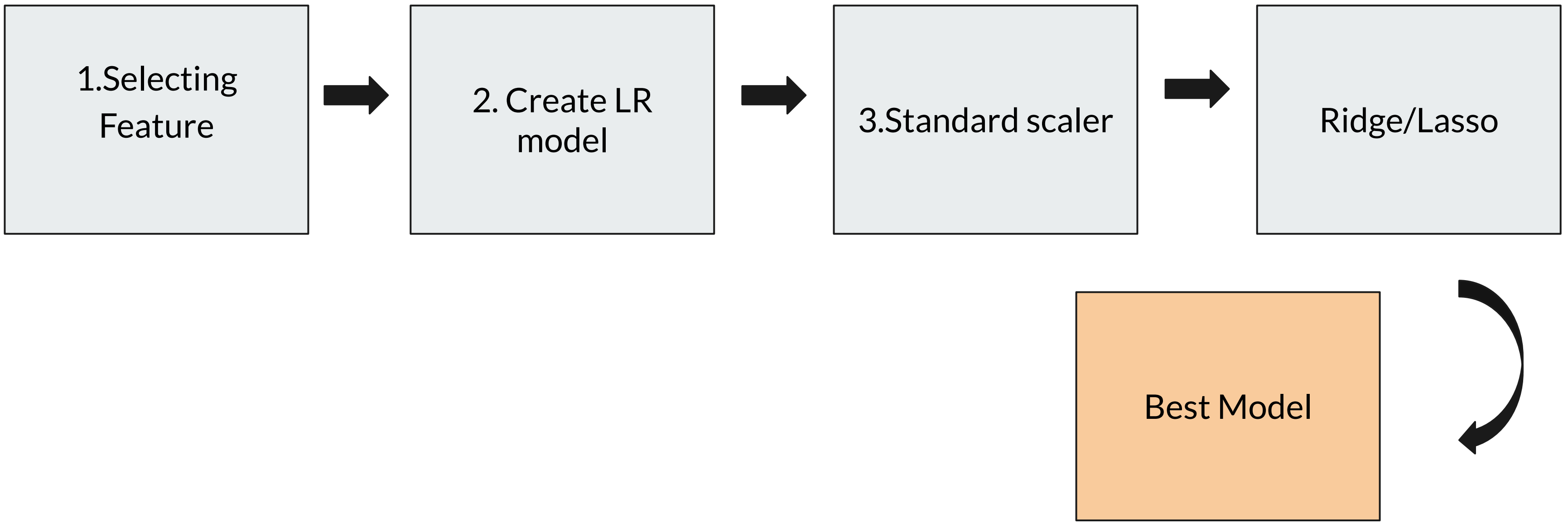
** Categorical data such as Province, District, and Property type need to be converted into numerical values before building the model. This can be achieved using one-hot encoding

district_Thung Khru	district_Wang Thonglang	district_Watthana	district_Yan Nawa	property_type_Detached House
0	0	1	0	0
0	0	1	0	0
0	0	0	0	0
0	0	0	0	1
0	0	0	0	0

Model Processing

Timeline of model processing

1. Selecting features that influence property prices.
2. Create Linear Regression Model .
3. The RMSE is high. Choose to standardize the scale to be consistent using the StandardScaler.
4. Experiment with Ridge and Lasso to find the best model.



Model Evaluation

Metrics Used for Model Evaluation

1. R2 (R-squared)
2. RMSE (Root Mean Square Error)

MODEL	R2 Score	RMSE(BAHT)
Linear Regression	0.62623	1,338,441
Linear Regression with StandardScaler	0.66444	1,251,252
Ridge	0.66434	1,251,502
Lasso	0.66429	1,251,592

Performance of Model

An RMSE (Root Mean Square Error) of 1,251,502 indicates that the model is making predictions with an average error of approximately 1.25 million baht. In general, a lower RMSE suggests better prediction accuracy. It's important to consider this value in the context of the overall price range of the houses being predicted. Keep in mind that RMSE is a measure of the model's predictive accuracy relative to the scale of the prices in the dataset.

Question

1. How can the Model ensure accurate property valuations while considering various factors like location, size, facilities, and market trends?
2. What features and functionalities should be incorporated to create a user-friendly experience for homeowners seeking property valuations?

Conclusions and Recommendations

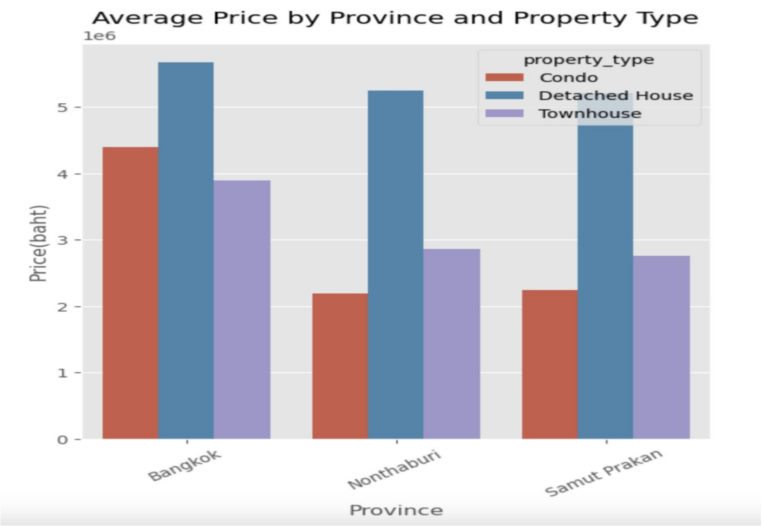
1. The model will be able to accurately assess house prices if we select appropriately sized features that are relevant to the price. Therefore, I recommend these features that will allow the model to calculate prices accurately

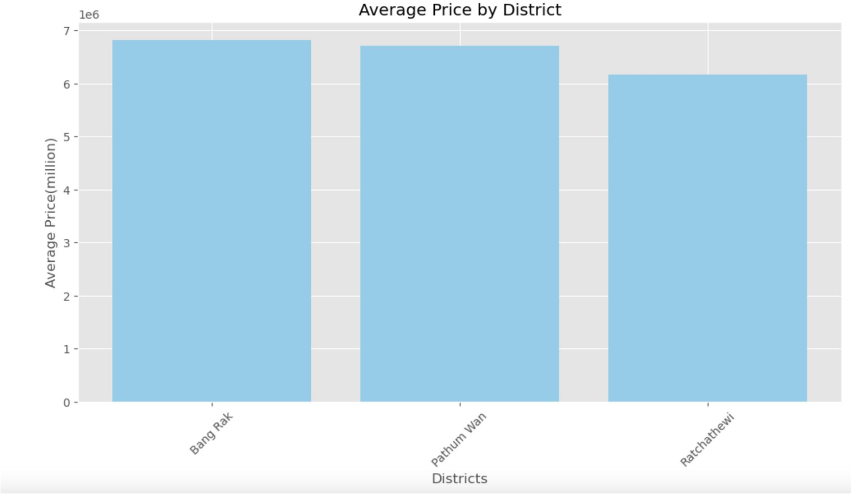
- Number of Bedroom , Number of Bathroom , Floor area , Number of Floor level, Number of Nearby shop , Number of Nearby station, Longtitude distance from Median , Lat distance from Median distance , Province, District, Property type

As I mentioned earlier, when using Ridge regression, the model performs the best and is the most accurate based on all the experiments conducted

Conclusions and Recommendations

Recommendation: If your house has a large floor area and is located close to shops, there is a higher likelihood that the price of your house will be higher. Additionally, houses in Bangkok tend to have the highest prices compared to the other two provinces





THANK YOU