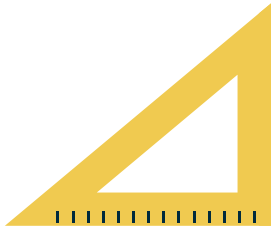
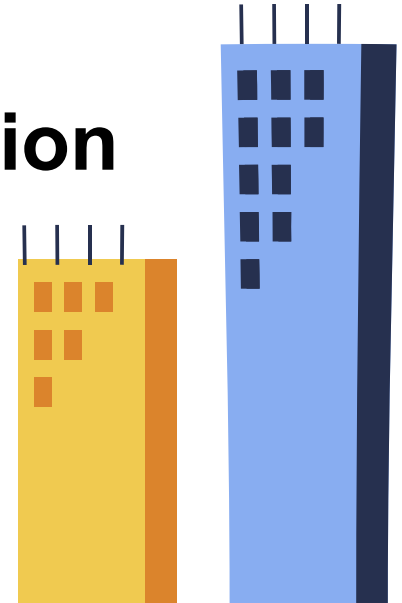


Project 2

Bangkok House Price Prediction



Oct 31, 2023 : DSI 03 : Yodchanon Nakwongwankul

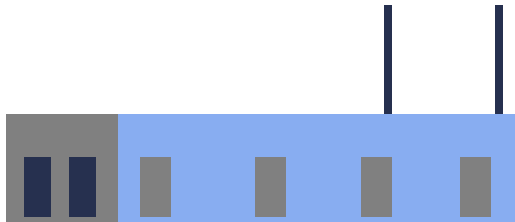
Problem Statement

“..AP Thai CMO is thinking of launching new project in 2024, aiming to found accommodation in Bangkok and suburban areas with data-driven pricing strategy. What he needs are

(1) To see which factors are significant to real estates price?

(2) An acceptably accurate price prediction model.

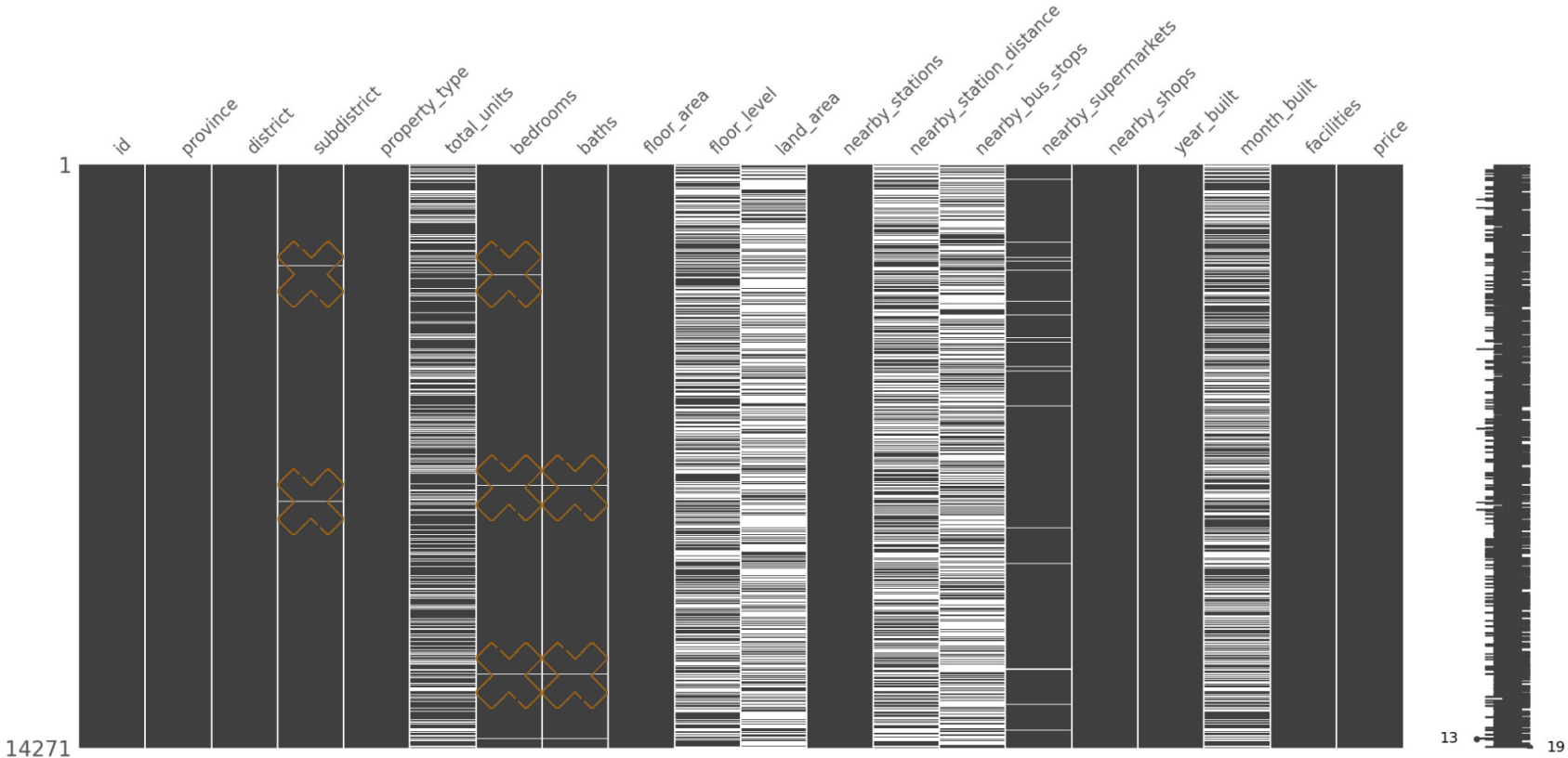
So, he requests his Data Science team to deliver the result within 2 weeks..”



Dataset Overview

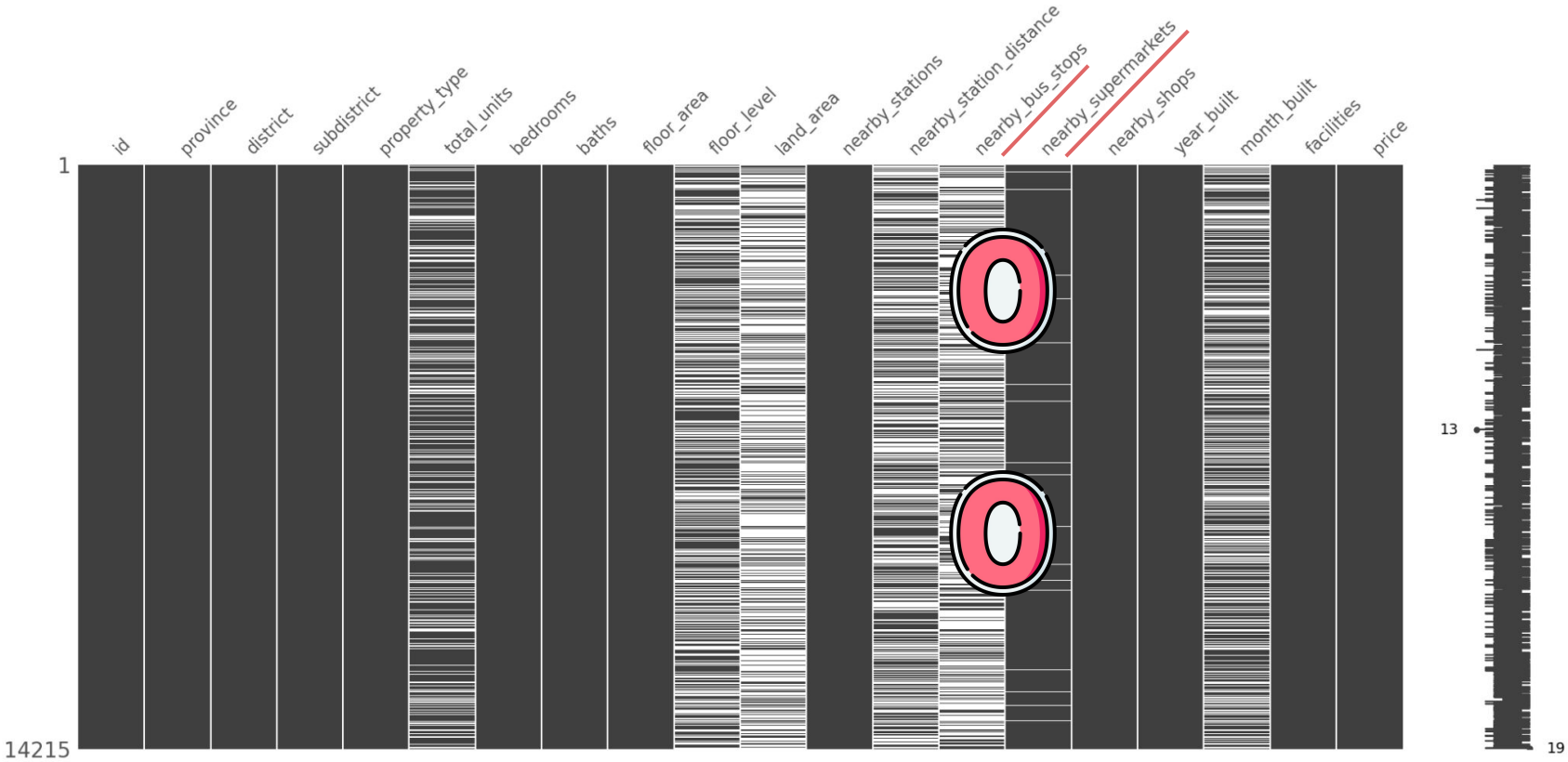
- 14,271 as number of observation
- 23 features from which 10 features having null value
- Having 3 property types : Condo-65%, Townhouse-20% and Detached House-15%
- `‘id’`, `‘address’`, `‘longitude’` and `‘latitude’` features will be excluded from this analysis

Observations contain missing value in columns `'subdistrict'`, `'bedrooms'` and `'baths'` were dropped

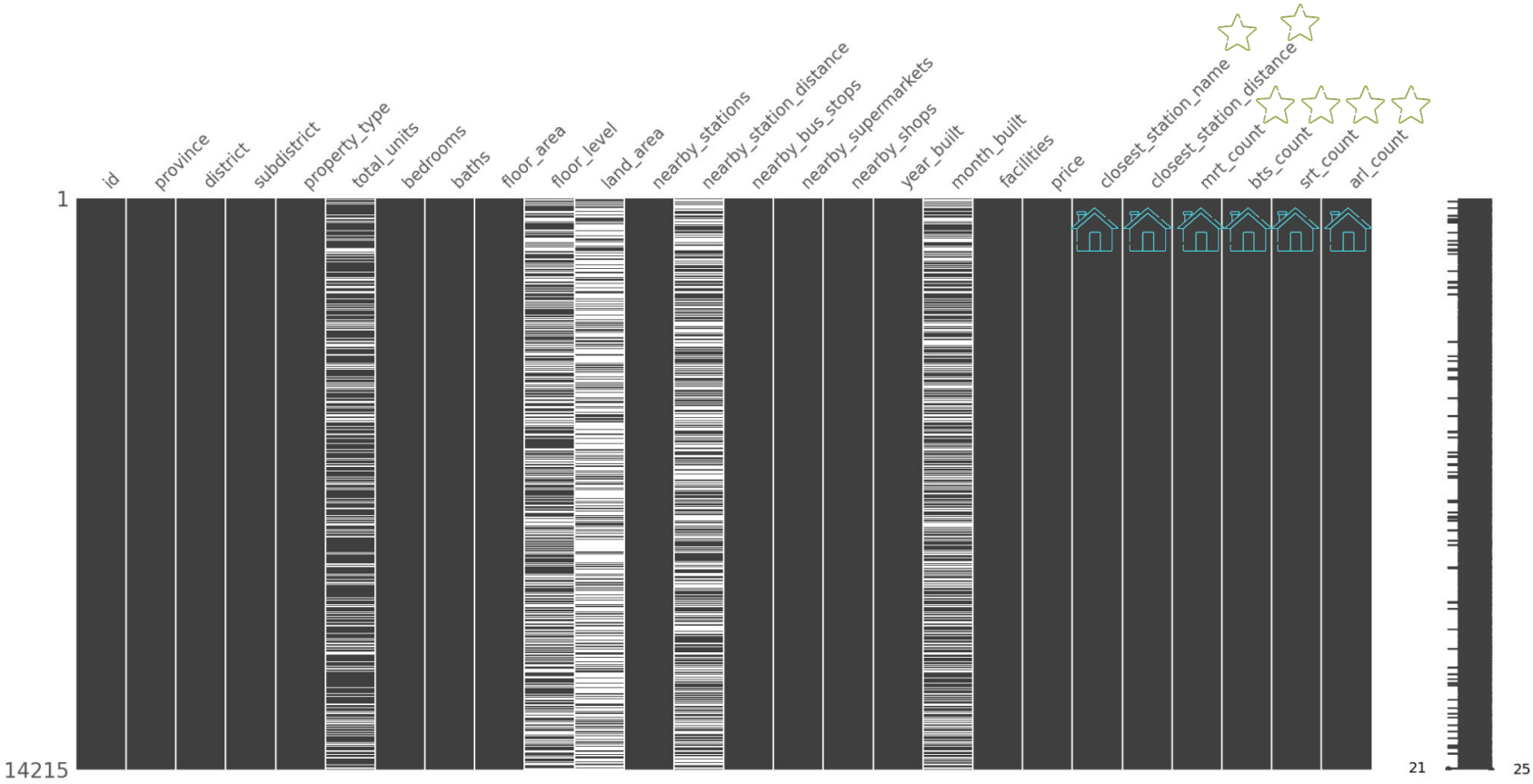


0.4% of total observations was dropped

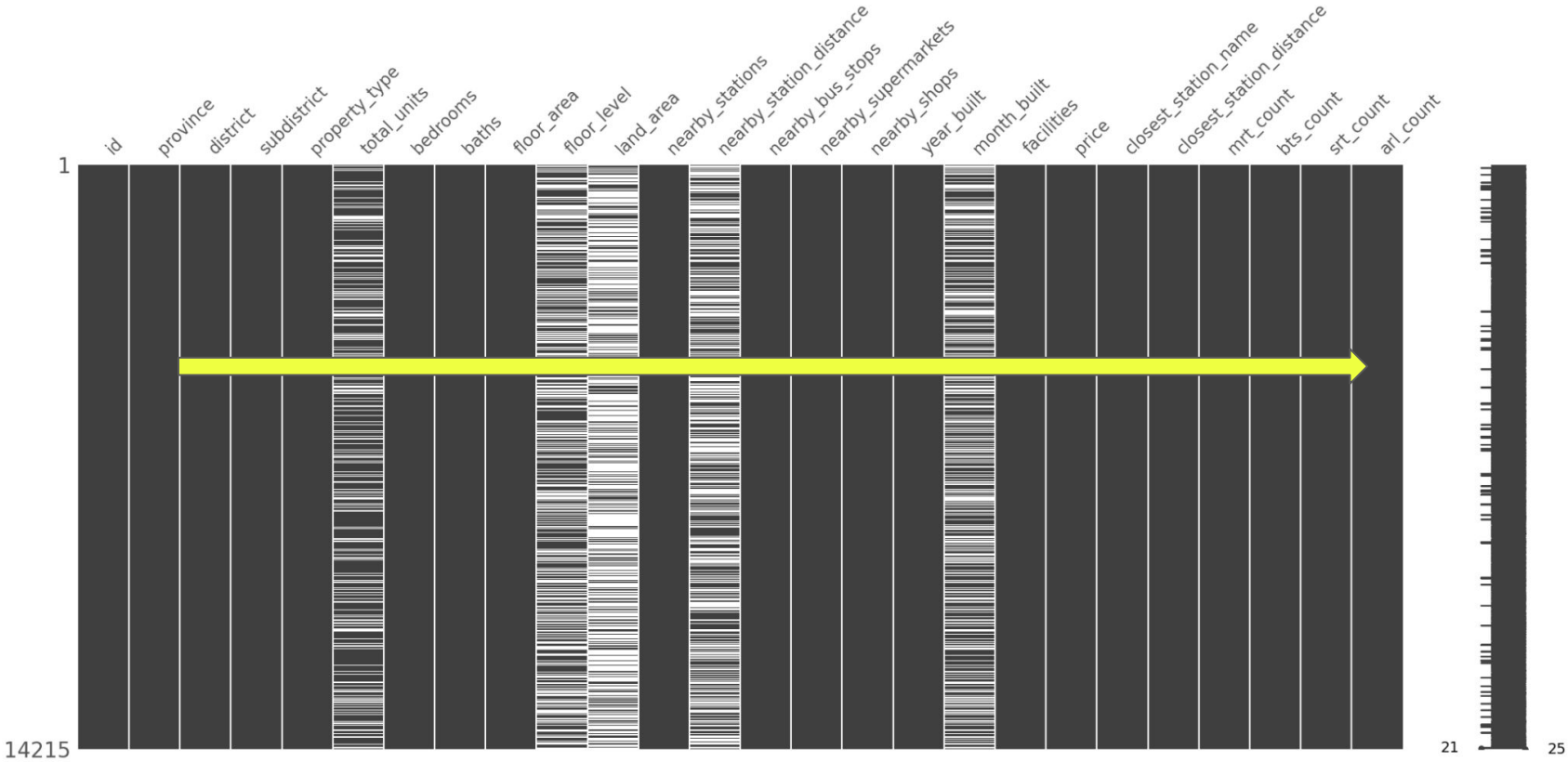
Replace Missing Value with zero 0 in `'nearby_bus_stops'` and `'nearby_supermarkets'` since the minimum values in those columns are above zero. Thus, the missing values are assumed to be zero.



6 numeric features were added to DataFrame which were extracted from `'nearby_station_distance'`

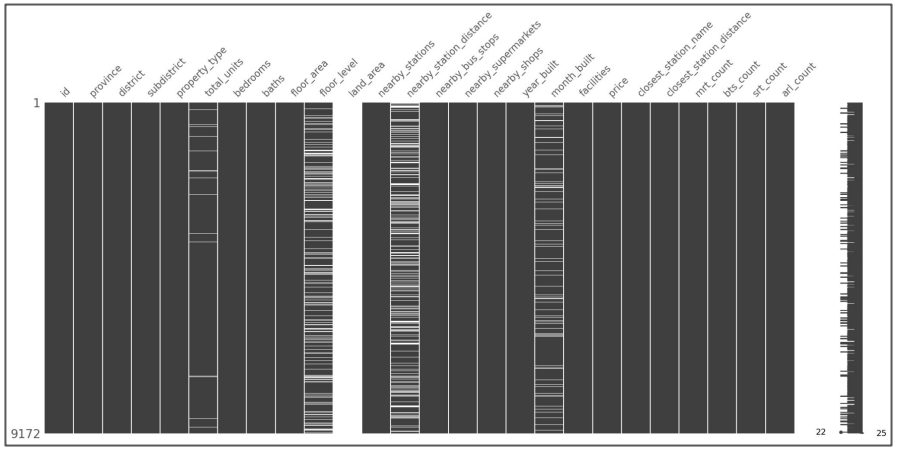


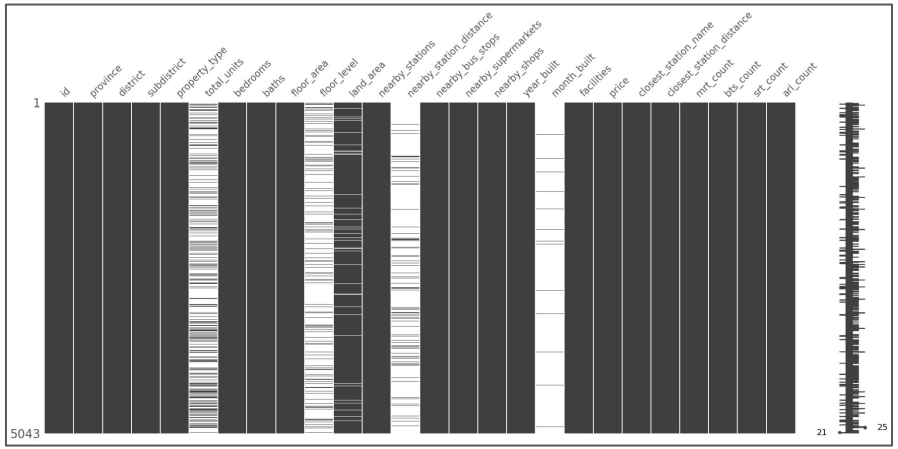
The remaining missing value seems to have a pattern in horizontal trend



Divide DataFrame into Condo-DataFrame and Non-Condo DataFrame for ease of missing value dealing

** Townhouse and Detached House are grouped into Non-Condo DataFrame

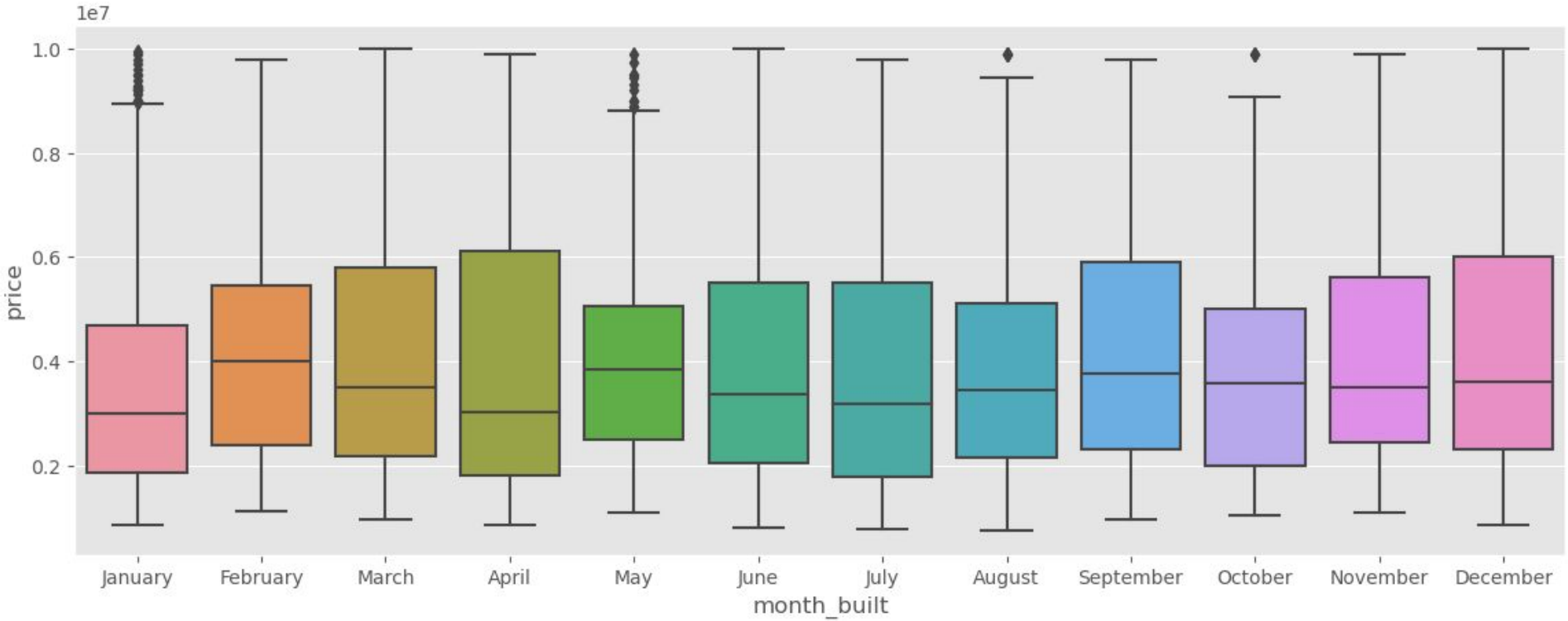




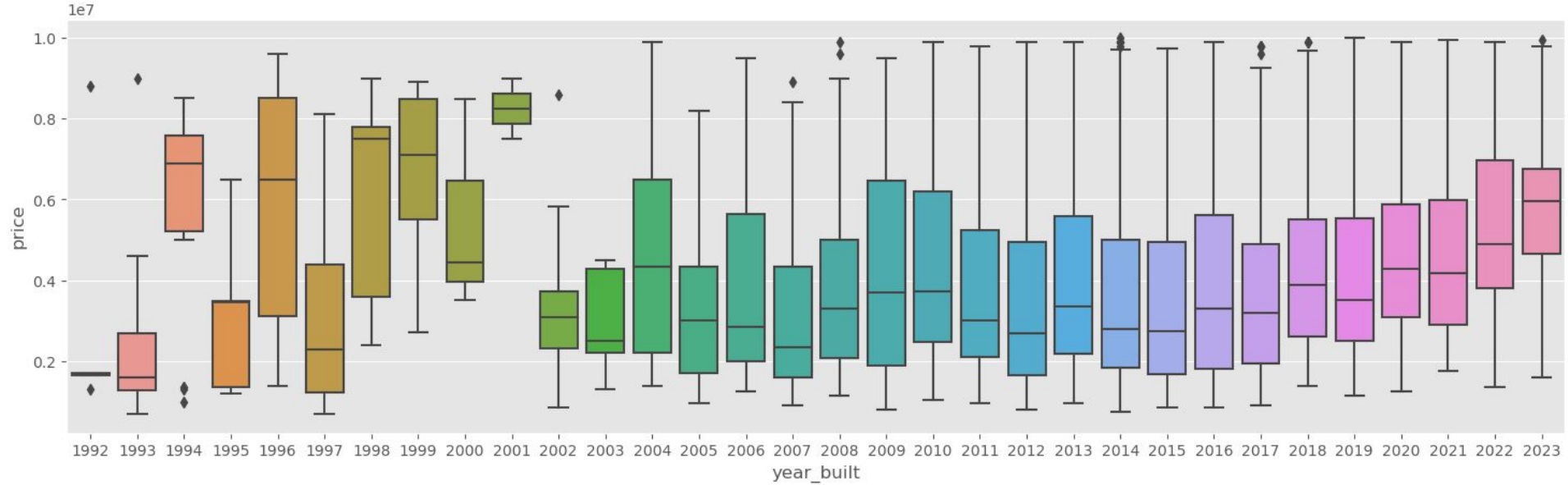
Condo DataFrame → Impute by subdistrict grouped median
→ Drop some columns
→ Dummify string columns
→ Standardize the features scale

Non-Condo DataFrame → Drop some columns
→ Dummify string columns
→ Scale Standardize

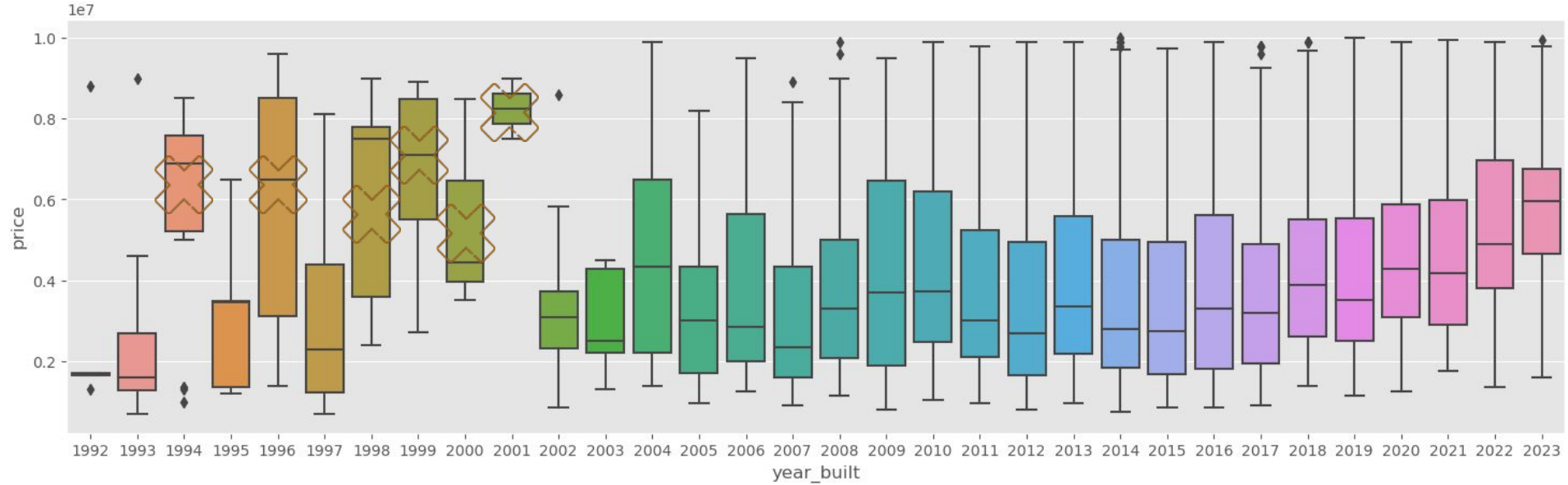
condo_df : `'month_built'` column was dropped as the distributions of `'price'` in each month are not different.



condo_df : `'price'` from `'year_built'` 2002 seems to get higher gradually but before `'year_built'` 2002 having high fluctuation, so some observations were dropped in order the get the distribution in the same trend.



condo_df : `'price'` from `'year_built'` 2002 seems to get higher gradually but before `'year_built'` 2002 having high fluctuation, so some observations were dropped in order the get the distribution in the same trend.



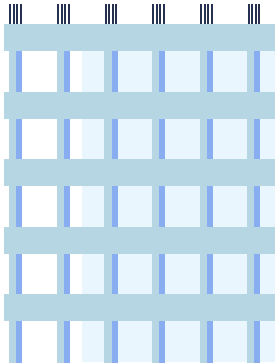
61 out of 9183 condo observations ,0.66%, was dropped

Modeling

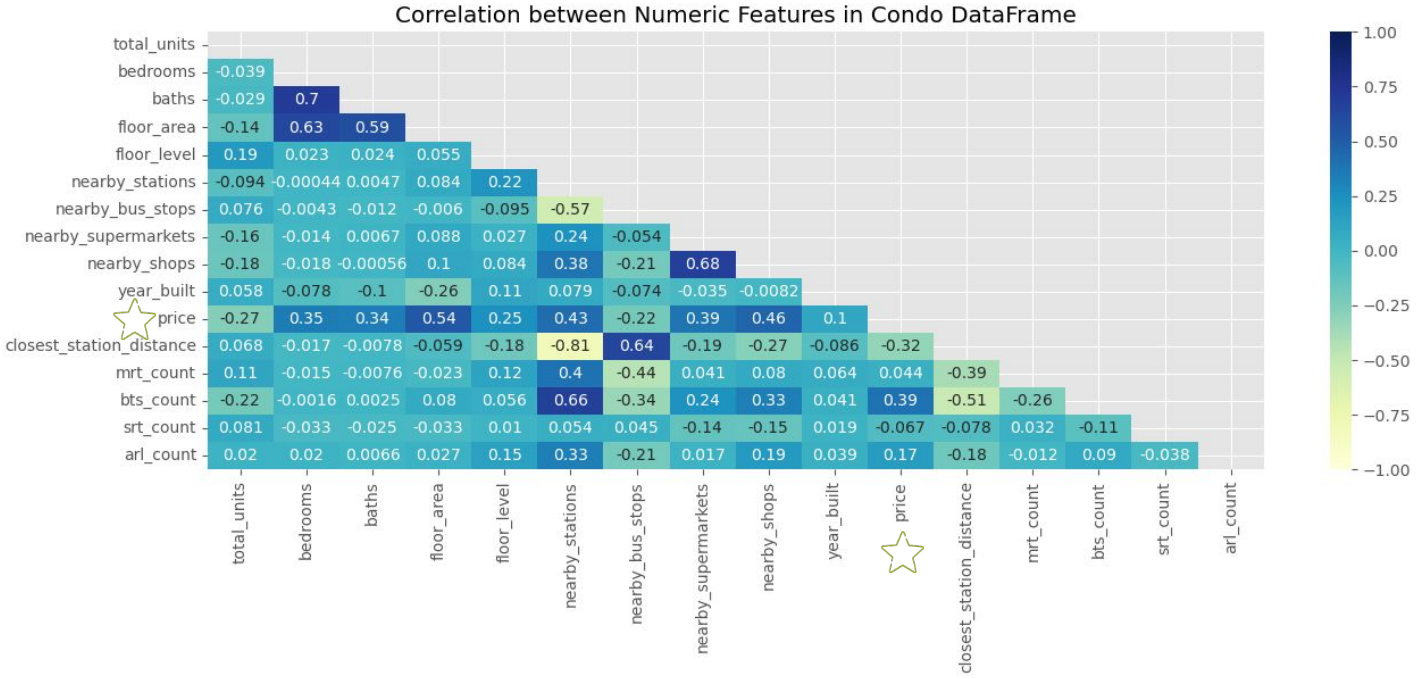
- Use LassoCV Model to predict the price for Condo and Non-condo property types

- To eliminate Multicollinearity between features when a large number of columns are added to the model

- To prevent overfitting incident as the number of features is high



condo_df : ‘floor_area’ has strongest positive correlation with ‘price’
‘closet_station_distance’ has strongest negative correlation with ‘price’



	price
price	1.000000
floor_area	0.538680
nearby_shops	0.459276
nearby_stations	0.429618
bts_count	0.390942
nearby_supermarkets	0.390459
bedrooms	0.348176
baths	0.335953
floor_level	0.247946
arl_count	0.174884
year_built	0.104080
mrt_count	0.043958
srt_count	-0.066988
nearby_bus_stops	-0.216240
total_units	-0.267610
closest_station_distance	-0.342243

condo_df : LassoCV Model (1)

- 17 features used

- R^2 Score - Training : 0.67913
- RMSE - Training : 1,236,913.73 THB

- R^2 Score - Test : 0.59483
- RMSE - Test : 1,394,385.29 THB

	Coefficient(LassoCV)
floor_area	1.054034e+06
year_built	4.779500e+05
floor_level	3.514049e+05
nearby_shops	3.417091e+05
bts_count	2.785592e+05
nearby_supermarkets	2.506112e+05
nearby_stations	1.061332e+05
baths	8.798380e+04
arl_count	6.744622e+04
mrt_count	3.147875e+04
bedrooms	3.101113e+04
srt_count	-0.000000e+00
nearby_bus_stops	-4.197816e+04
closest_station_distance	-7.588427e+04
total_units	-2.594722e+05
province_Nonthaburi	-3.014386e+05
province_Samut Prakan	-3.224617e+05

condo_df : LassoCV Model (2)

- 188 features used, `'district'` and `'closest_station_name'` are considered
- Drop `'province'`

- R^2 Score - Training : 0.82149
- RMSE - Training : 922,563.70 THB

- R^2 Score - Test : 0.75651
- RMSE - Test : 1,080,946 THB

Top 5 positive factors to `'price'`

	Coefficient(LassoCV)
floor_area	902392.427581
closest_station_name_NO STATION	749624.678145
district_Watthana	569997.869994
year_built	463330.139981
district_Khlong Toei	395114.679163

Bottom 5 negative factors to `'price'`

	Coefficient(LassoCV)
closest_station_name_PP15 Bang Son MRT	-97097.602070
closest_station_name_PP03 Sam Yaek Bang Yai MRT	-112795.468275
closest_station_name_PP13 Yaek Tiwanon MRT	-122762.164325
closest_station_name_PP10 Bang Krasor MRT	-176240.598474
closest_station_distance	-954337.285492

condo_df : LassoCV Model (3)

- 250 features used, `'facilities'` is considered

- R^2 Score - Training : 0.83660
- RMSE - Training : 882,666.25 THB

- R^2 Score - Test : 0.76722
- RMSE - Test : 1,056,890.35 THB

Top 5 positive factors to `'price'`

	Coefficient(LassoCV)
floor_area	933445.702345
closest_station_name_NO STATION	645681.134843
district_Watthana	524577.149013
year_built	401975.753450
district_Khlong Toei	342879.909802

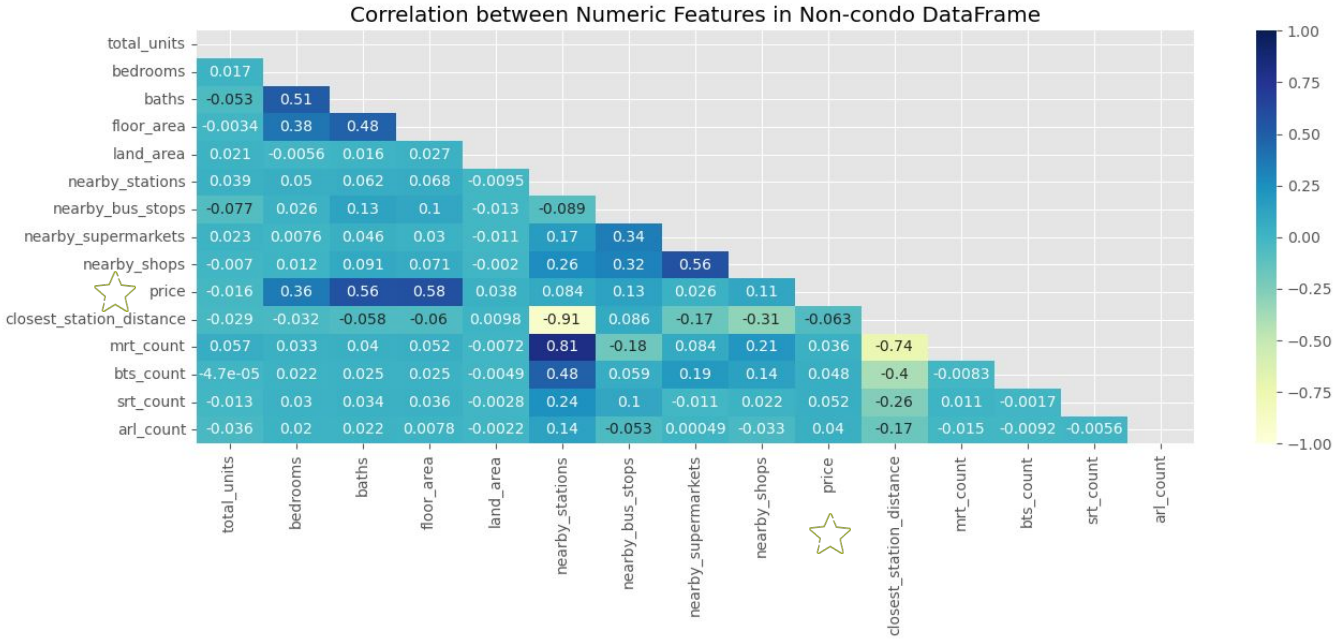
Bottom 5 negative factors to `'price'`

	Coefficient(LassoCV)
total_units	-107237.369214
closest_station_name_PP03 Sam Yaek Bang Yai MRT	-135356.272369
closest_station_name_PP13 Yaek Tiwanon MRT	-147282.383687
closest_station_name_PP10 Bang Krasor MRT	-161931.682796
closest_station_distance	-876662.493607

condo_df : Model 3 is best performing model

	LassoCV Model 1	LassoCV Model 2	LassoCV Model 3
Number of Features	17	188	250
R^2 - Score : Training Set	0.67913	0.82149	0.83660
RMSE : Training Set (THB)	1,236,913.73	922,563.70	882,666.25
R^2 - Score : Test Set	0.59483	0.75651	0.76722
RMSE : Test Set (THB)	1,394,385.29	1,080,946	1,056,890.35

non_condo_df : `'floor_area'` has strongest positive correlation with `'price'`



	price
price	1.000000
floor_area	0.579923
baths	0.562876
bedrooms	0.356604
nearby_bus_stops	0.134385
nearby_shops	0.108877
nearby_stations	0.084206
srt_count	0.052115
bts_count	0.047740
arl_count	0.039918
land_area	0.037896
mrt_count	0.035675
nearby_supermarkets	0.025911
total_units	-0.016065
closest_station_distance	-0.063402

non_condo_df : LassoCV Model(4)

- 17 features used

- R^2 Score - Training : 0.58887
- RMSE - Training : 1,368,795.91 THB

- R^2 Score - Test : 0.54628
- RMSE - Test : 1,425,545.88 THB

	Coefficient(LassoCV)
baths	769282.390392
floor_area	488845.551418
nearby_stations	120082.061558
nearby_shops	117020.931938
nearby_bus_stops	87550.149099
land_area	42364.944936
srt_count	26418.669983
arl_count	18342.710354
bts_count	-0.000000
total_units	-0.000000
closest_station_distance	0.000000
bedrooms	-18590.317664
mrt_count	-23186.592932
nearby_supermarkets	-34131.981625
province_Samut Prakan	-165775.103390
province_Nonthaburi	-187804.210543
property_type_Townhouse	-832889.936857

non_condo_df : LassoCV Model(5)

- 143 features used, `'district'`, `'closest_station_name'` and `'property_type'` are considered
- Drop `'province'`

- R^2 Score - Training : 0.66545
- RMSE - Training : 1,234,739.77 THB

- R^2 Score - Test : 0.61654
- RMSE - Test : 1,310,524.22 THB

Top 5 positive factors to `'price'`

	Coefficient(LassoCV)
floor_area	902392.427581
closest_station_name_NO STATION	749624.678145
district_Watthana	569997.869994
year_built	463330.139981
district_Khlong Toei	395114.679163

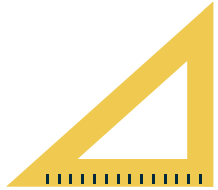
Bottom 5 negative factors to `'price'`

	Coefficient(LassoCV)
closest_station_name_PP15 Bang Son MRT	-97097.602070
closest_station_name_PP03 Sam Yaek Bang Yai MRT	-112795.468275
closest_station_name_PP13 Yaek Tiwanon MRT	-122762.164325
closest_station_name_PP10 Bang Krasor MRT	-176240.598474
closest_station_distance	-954337.285492

non_condo_df : Model 6 is best performing model

	LassoCV Model 4	LassoCV Model 5	LassoCV Model 6
Number of Features	17	143	206
R^2 - Score : Training Set	0.58887	0.66545	0.69549
RMSE : Training Set (THB)	1,368,795.91	1,234,739.77	1,1178,004.21
R^2 - Score : Test Set	0.54628	0.61654	0.64561
RMSE : Test Set (THB)	1,425,545.88	1,310,524.22	1,259,870.66

Recommendation and Summary



For Condo, the lassoCV model 3 is the best model with highest R2-Score and the overall average predicted price error is around 1,056,890 THB. The total area of the condo is the most positive factor when it increases by 1 meter square, the predicted price will increase around +933,445 THB while the distance from the closest railway station is the most negative factor when it increases by 1 meter, the predicted price will decrease around -876,662 THB

For Non-condo, Detached house and Townhouse, the lassoCV model 6 is the best model with highest R2-Score and the overall average predicted price error is around 1,259,870 THB.The number of baths is the most positive factor when it increases by 1 unit, the predicted price will increase around +636,543 THB while the ‘townhouse’ property type is the most negative factor which give the sales price cheaper than that of detached house around -802,908 THB in the same environment.