Bangkok Housing Price Prediction

For buyers and sellers

Problems

Do you ever have a problem with housing price

when trying move to a new home ?

Housing price in Bangkok is normally fluctuated due to many factors, some factors are considered to be positive for the price,

but is that always true ?

The data:

Bangkok Housing Price

Column	Type	Description
id	int	ID of selling item
province	string	province name: this dataset only includes Bangkok,Samut Prakan and Nonthaburi
district	string	district name
subdistrict	string	subdtistrict name
address	string	address e.g. street name, area name, soi number
property_type	string	type of the house: Condo, Townhouse or Detached House
total_units	float	the number of rooms/houses that the condo/village has
bedrooms	int	the number of bedrooms
baths	int	the number of baths
floor_area	float	total area of inside floor [m^2]
floor_level	int	floor level of the room
land_area	float	total area of the land [m^2]
latitude	float	latitude of the house
longitude	float	longitude of the house
nearby_stations	int	the number of nearby stations (within 1km)
nearby_station_distance	list	list of (station name, distance[m]). Each station name consists of station ID, station name, and Line such as "E4 Asok BTS"
nearby_bus_stops	int	the number of nearby bus stops
nearby_supermarkets	int	the number of nearby supermarkets
nearby_shops	int	the number of nearby shops
year_built	int	year built
month_built	string	month built: January-December
facilities	list	list of facilities
price	float	selling price

Considering the information …

Useless

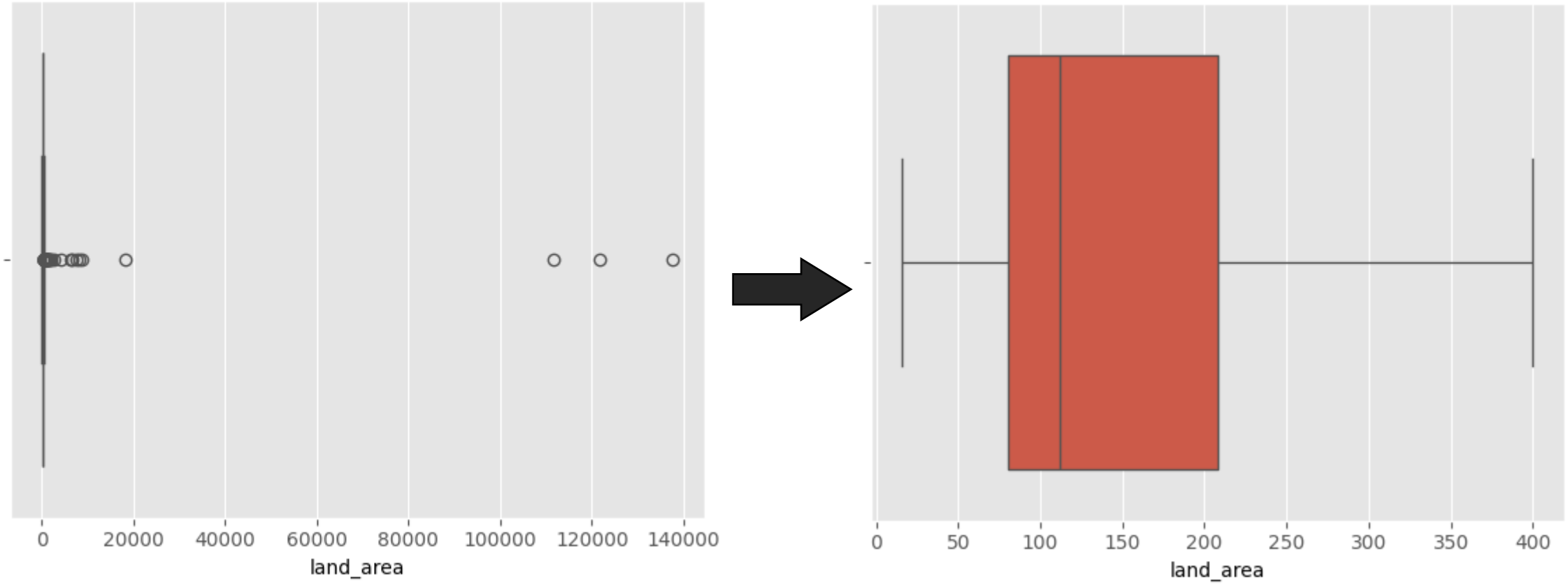
- Id
- Month built
- Address

Redundant

- Province
- Latitude
- Longitude

Column	Type	Description
district	string	district name
subdistrict	string	subdtistrict name
property_type	string	type of the house: Condo, Townhouse or Detached House
total_units	float	the number of rooms/houses that the condo/village has
bedrooms	int	the number of bedrooms
baths	int	the number of baths
floor_area	float	total area of inside floor [㎡]
floor_level	int	floor level of the room
land_area	float	total area of the land [㎡]
nearby_stations	int	the number of nearby stations (within 1km)
nearby_station_distance	list	list of (station name, distance[m]). Each station name consists of station ID, station name, and Line such as "E4 Asok BTS"
nearby_bus_stops	int	the number of nearby bus stops
nearby_supermarkets	int	the number of nearby supermarkets
nearby_shops	int	the number of nearby shops
year_built	int	year built
facilities	list	list of facilities

Outliers - land_area (cutoff at 400 m^2)

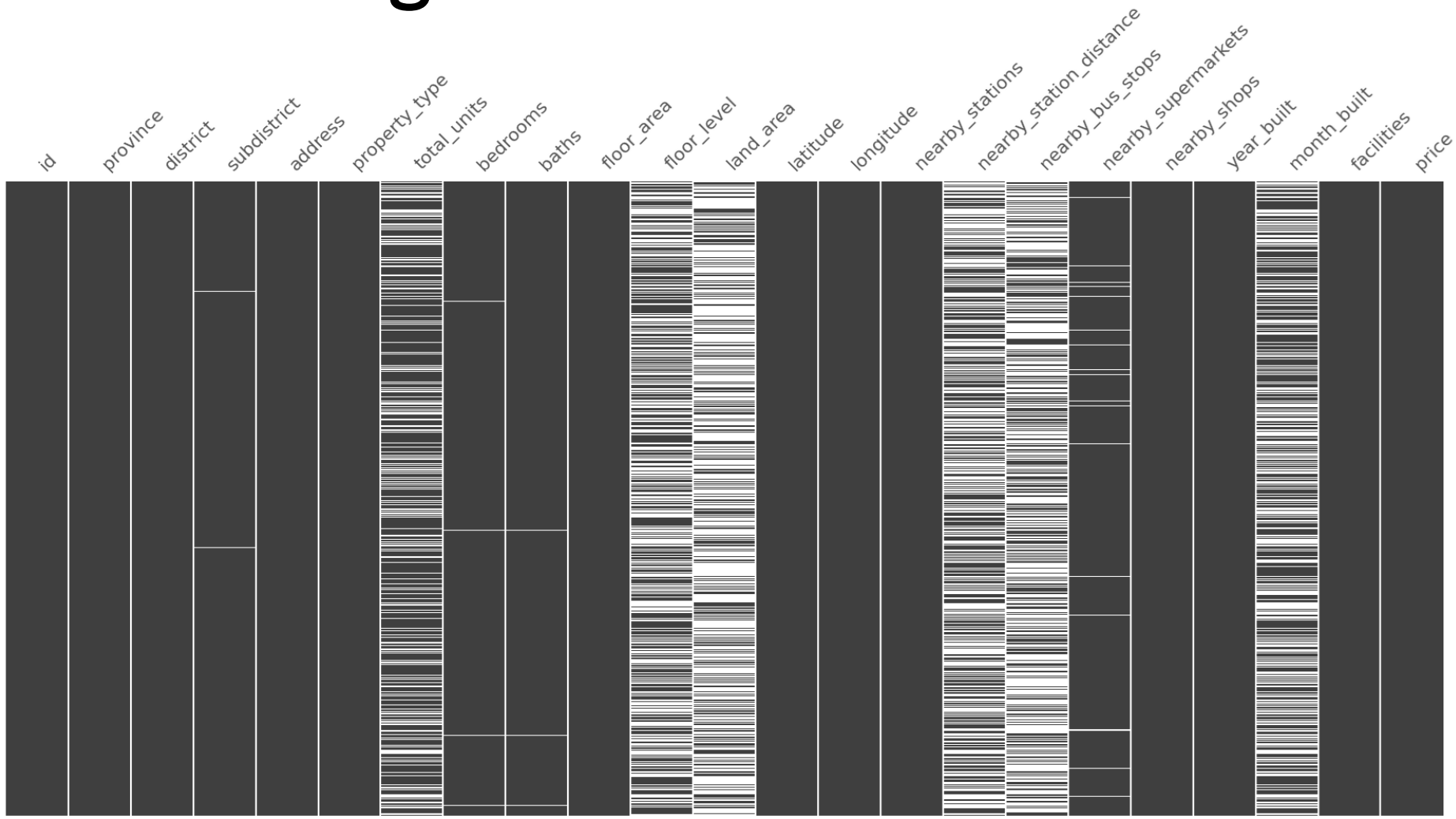


Value corrections

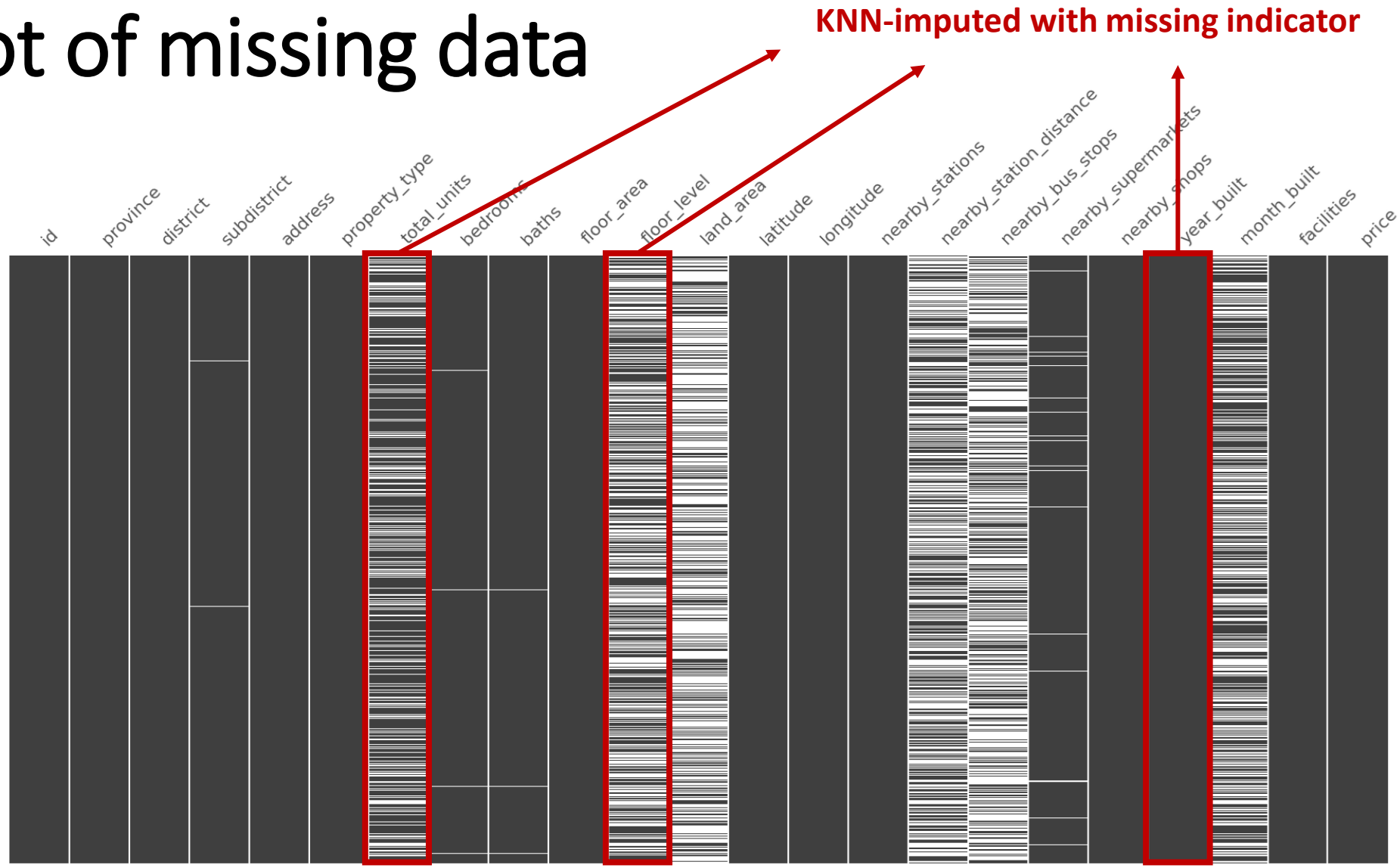
There are some incorrect values in subdistrict that need to be replaced.

dist_subdist	replacement
Bang Kapi_624 Condolette Ladprao	Bang Kapi_Khlong Chan
Bang Kapi_Sathorn Happy Land	Bang Kapi_Khlong Chan
Bang Kho Laem_StarView Rama 3	Bang Kho Laem_Bang Khlo
Bang Khun Thian_Smart Condo Rama 2	Bang Khun Thian_Samae Dam
Bang Rak_ASHTON Silom	Bang Rak_Suriyawong
Bang Rak_M Silom	Bang Rak_Suriyawong
Bangkok Yai_IDEO Thaphra Interchange	Bangkok Yai_Wat Tha Phra
Chatuchak_Supalai Park Ratchayothin	Chatuchak_Lat Yao
Chatuchak_	Chatuchak_Chan Kasem
Din Daeng_Lumpini Suite Dindaeng-Ratchaprarop	Din Daeng_Din Daeng
Din Daeng_The Kris Express 2	Din Daeng_Din Daeng
Din Daeng_The Kris Extra 5	Din Daeng_Din Daeng
Huai Khwang_Chateau In Town Ratchada 20	Huai Khwang_Sam Sen Nok
Huai Khwang_Noble Revolve Ratchada	Huai Khwang_Huai Khwang
Khlong San_Villa Sathorn	Khlong San_Khlong Ton Sai
Lak Si_Plum Condo Chaengwattana Station	Lak Si_Talat Bang Khen
Pathum Wan_CU Terrace	Pathum Wan_Wang Mai
Phra Khanong_Whizdom The Exclusive	Phra Khanong_Bang Chak
Phra Khanong_	Phra Khanong_Bang Chak
Ratchathewi_Life Asoke - Rama 9	Ratchathewi_Makkasan

A lot of missing data



A lot of missing data

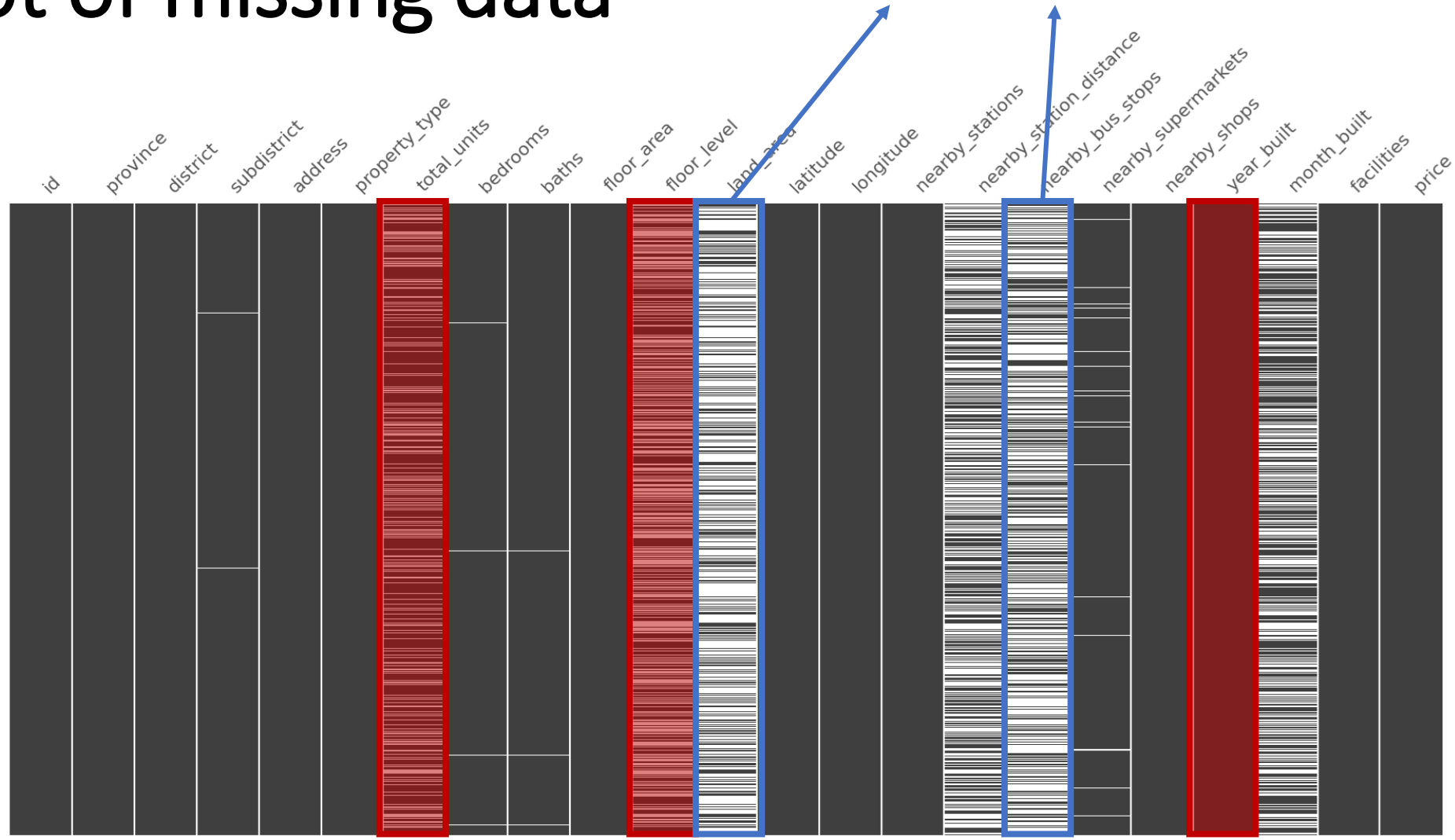


* There are 0 in year_built, replace it with empty value and then impute

KNN-imputed with missing indicator

A lot of missing data

Mean-imputed with missing indicator

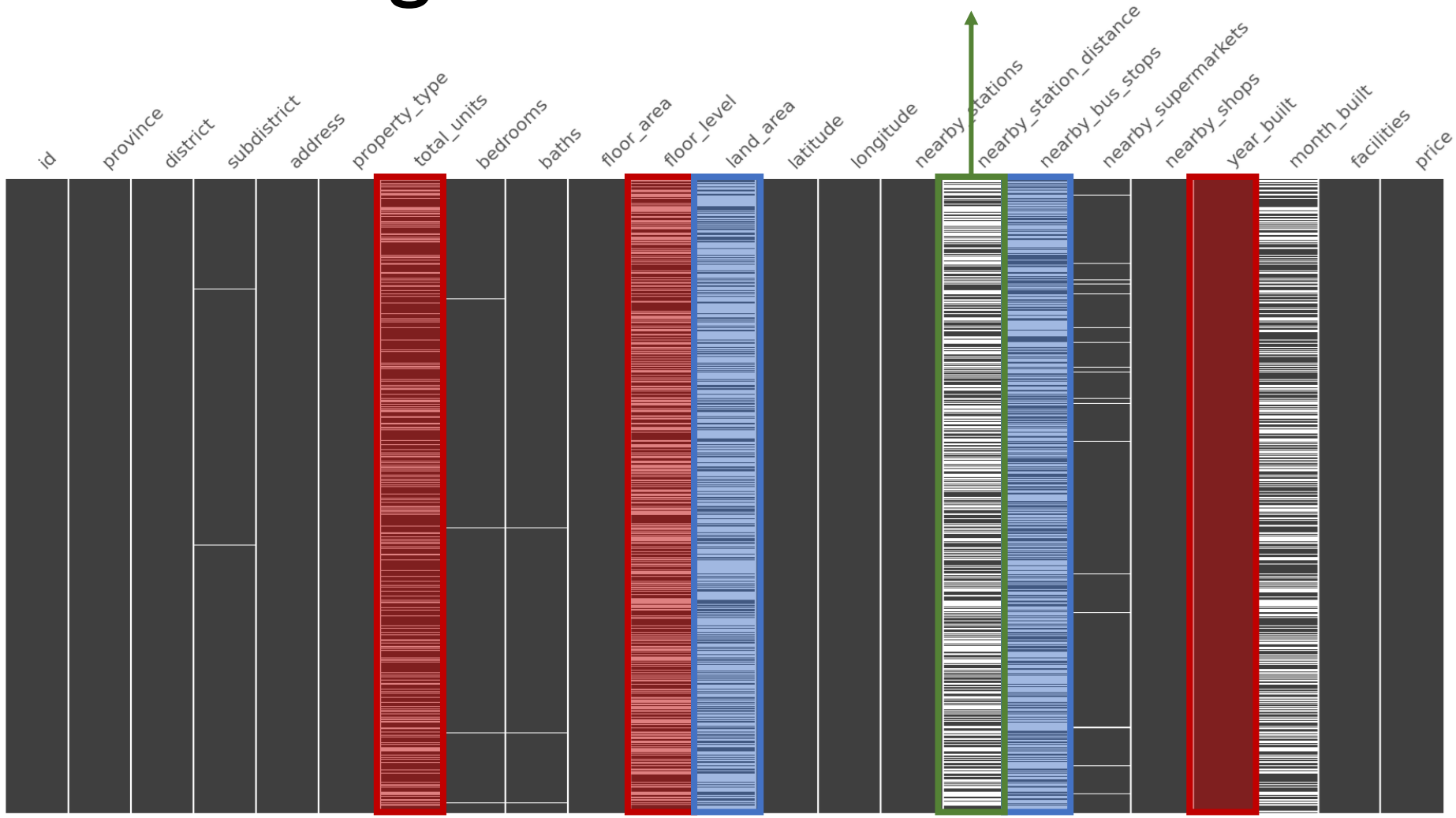


A lot of missing data

KNN-imputed with missing indicator

Mean-imputed with missing indicator

Custom transformation + scaling



A lot of missing data

KNN-imputed with missing indicator

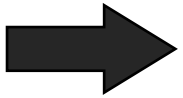
Mean-imputed with missing indicator

Custom transformation + scaling

For nearby_station_distance

Custom transformation + scaling

	nearby_station_distance
0	[[E7 Ekkamai BTS, 270], [E6 Thong Lo BTS, 800]]
1	[[BL22 Sukhumvit MRT, 720], [BL21 Phetchaburi ...
2	[[E5 Phrom Phong BTS, 650], [BL23 Queen Siriki...
3	None
4	[[PP09 Yaek Nonthaburi 1 MRT, 10]]
...	...

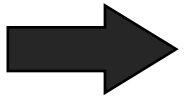


	E7 Ekkamai BTS	E6 Thong Lo BTS	BL22 Sukhumvit MRT	BL21 Phetchaburi MRT
0	-1.118517	0.890493	0.000000	0.000000
1	0.000000	0.000000	0.587246	1.080022
2	0.000000	0.000000	0.000000	0.000000
3	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	0.000000
...	...	...	...	...

Dummified

Because prediction technique needs all data to be numbers …

	property_type
0	Condo
1	Condo
2	Condo
3	Detached House
4	Townhouse
...	...
14266	Condo
14267	Townhouse
14268	Detached House
14269	Townhouse
14270	Condo



	property_type_Detached House	property_type_Townhouse
0	0	0
1	0	0
2	0	0
3	1	0
4	0	1
...	...	...
14266	0	0
14267	0	1
14268	1	0
14269	0	1
14270	0	0

Other transformations & summary

Column	Description	Transformation
floor_area	total area of inside floor [㎡]	log + polynomial degree 2 + scaled
floor_level	floor level of the room	log + polynomial degree 2 + scaled
land_area	total area of the land [㎡]	log + polynomial degree 2 + scaled
total_units	the number of rooms/houses that the condo/village has	Polynomial degree 2 + scaled
bedrooms	the number of bedrooms	Polynomial degree 2 + scaled
baths	the number of baths	Polynomial degree 2 + scaled
nearby_stations	the number of nearby stations (within 1km)	polynomial degree 2 + scaled
nearby_supermarkets	the number of nearby supermarkets	polynomial degree 2 + scaled
nearby_shops	the number of nearby shops	polynomial degree 2 + scaled
year_built	year built	polynomial degree 2 + scaled
facilities	list of facilities	Converted to facilities count - polynomial degree 2 + scaled
nearby_bus_stops	the number of nearby bus stops	scaled
nearby_station_distance	list of (station name, distance[m]). Each station name consists of station ID, station name, and Line such as "E4 Asok BTS"	Dummified-like (to scaled distance)
dist_subdist	district + subdistrict name concatenated with some value corrections	Dummified
property_type	type of the house: Condo, Townhouse or Detached House	Dummified

Result

Prediction model that

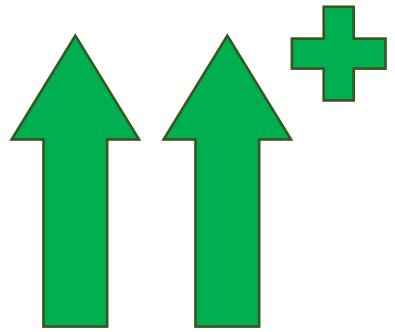
- Can explain **79.8%** of data
- The average error of **952K** THB approximately

Submission and Description	Public Score
submission.csv Complete · now · Ridge	960635

Other findings

The following factors are significantly affect housing price positively.

- Floor area
- Number of bathrooms
- Number of nearby stations (within 1km)
- Total area of the land
- Year built
- Number of bedrooms



Other findings

Top 10 district/subdistrict that will have higher housing price according to the model.

District	Subdistrict
Bangkok Yai	Wat Arun
Ratchathewi	Thung Phaya Thai
Bang Rak	Maha Phruettharam
Bang Rak	Suriyawong
Watthana	Khlong Tan Nuea
Pathum Wan	Rong Mueang
Khlong Toei	Khlong Tan
Watthana	Khlong Toei Nuea
Khlong Toei	Khlong Toei
Pathum Wan	Lumphini

Extra

- XGBoost model with **~85%** score
- Average error **~771K**
- Total of 49 features

Column	Transformation
bedrooms	polynomial degree 2 + scaled
baths	polynomial degree 2 + scaled
nearby_stations	polynomial degree 2 + scaled
nearby_supermarkets	polynomial degree 2 + scaled
nearby_shops	polynomial degree 2 + scaled
nearby_station_distance	converted to distance mean + polynomial degree 2 + scaled
facilities	converted to facilities count - polynomial degree 2 + scaled
floor_area	scaled
floor_level	scaled
land_area	scaled
latitude	scaled
longitude	scaled
nearby_bus_stops	scaled
year_built	scaled
property_type	dummified

submission.csv

Complete · 1d ago · XGBoost, fix scaling and missing ind + optimize numbers of X

808159