

# Bangkok & Metropolitan Area

## Price Prediction

Data Science Immersive Course  
Project 2



# Problem Statement



X

Real Estate Consulting Firm

Given that  
**we're real estate consulting firm whose clients are leading real estate developers in Bangkok and Metropolitan Area**

As real estate market is bouncing back from Covid financial crisis. **Our clients trying to explore possibility in launch projects in 2,500 prospective areas** (Test Data) given from client and client marketing team wants to know proper pricing in those areas that reflects **from existing estate data**. (Train Data)

So that, the marketing team could get glimpse of margin, worthiness of investment for each projects based on predicted data to ranking priority of prospective areas.

**Prospect Audience:** Client Marketing Team

## Scenario



PRUKSA



SANSIRI

PROPERTY PERFECT

Real Estate Developers

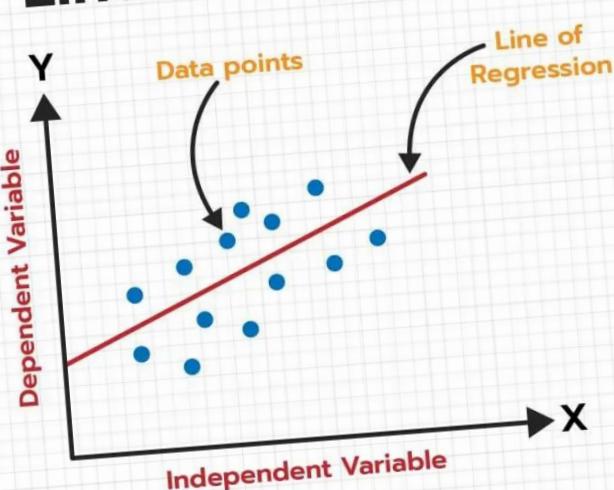
# Our Plan

Find out ➤

1. **Price Prediction Model** for 2,500 potential project location from existing real estate data
2. **Crucial Price Indicating Factor** that matter in real estate industry (i.e. premise features)
3. **Recommendation** for further studies and Business strategy/execution suggestion

1. **Price Prediction Model** for 2,500 potential project location from existing real estate data

## Linear Regression



## Simple Linear Regression

Approach for modelling the relationship between a **scalar** response and one explanatory variable (Price of Estate)

Assumptions associated with a linear regression model:

1. **Linearity:** The relationship between X and the mean of Y is linear.
2. **Independence:** Observations are independent of each other.
3. **Normality:** For any fixed value of X, Y is normally distributed.
4. **Equality of Variance:** If the residuals do not fan out in a triangular fashion that means that the equal variance assumption is met.
5. **Homoscedasticity:** The variance of residual is the same for any value of X.

\*With L1 regularization (Lasso) if needed in a model



2  
Datasets

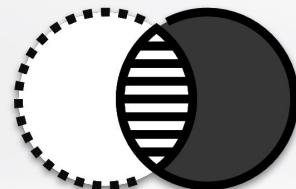
Train set (23 Features)  
Test (All Features w/o Price)

Period: Year 1993-2022

14,271 rows, 23 columns  
2,500 rows, 22 columns



Modeling  
& Prediction



Feature Selection  
Feature Engineering



## EDA

Exploratory  
Data Analysis



### Data Cleansing

Data Imputation  
(i.e. filling/dropping missing values, filtering out odd ones)



# Exploratory Data Analysis: Matrix

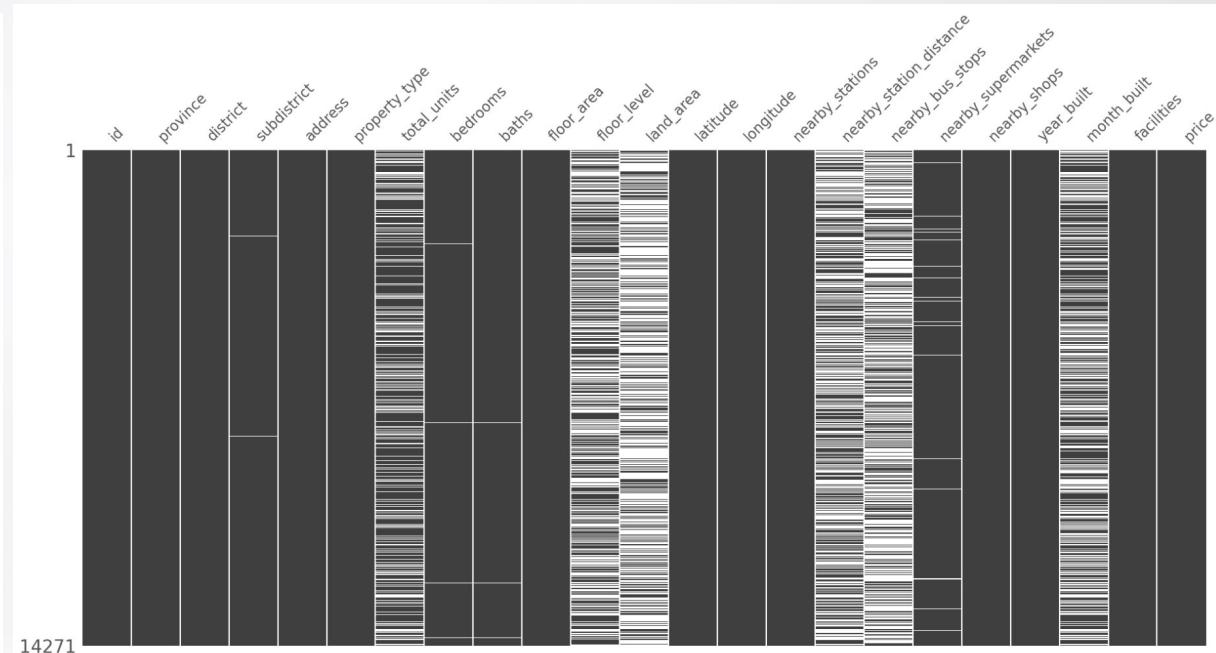
RangeIndex: 14271 entries, 0 to 14270

Data columns (total 23 columns):

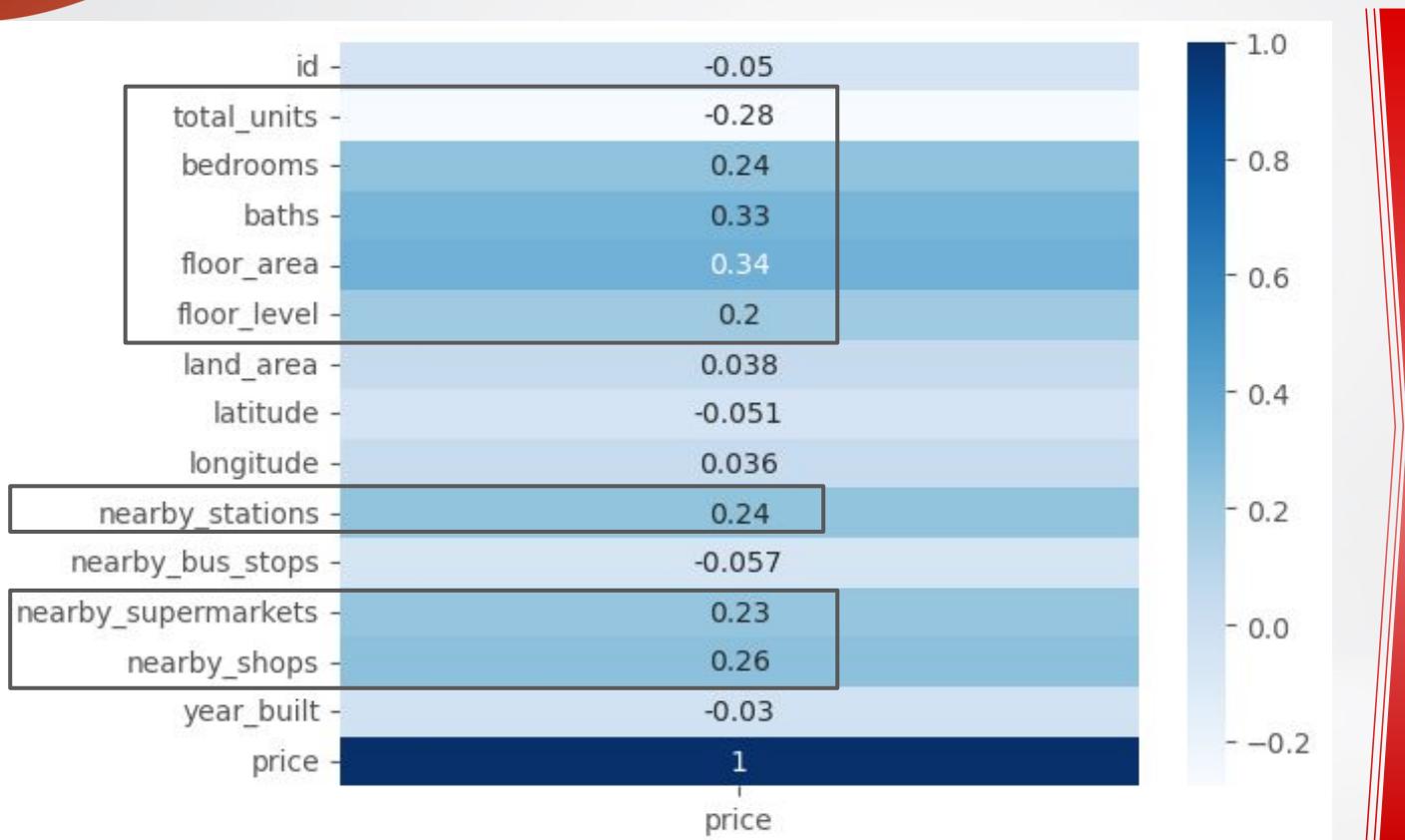
#	Column	Non-Null Count	Dtype
0	id	14271	non-null int64
1	province	14271	non-null object
2	district	14271	non-null object
3	subdistrict	14260	non-null object
4	address	14271	non-null object
5	property_type	14271	non-null object
6	total_units	10509	non-null float64
7	bedrooms	14228	non-null float64
8	baths	14236	non-null float64
9	floor_area	14271	non-null int64
10	floor_level	8093	non-null float64
11	land_area	4917	non-null float64
12	latitude	14271	non-null float64
13	longitude	14271	non-null float64
14	nearby_stations	14271	non-null int64
15	nearby_station_distance	7228	non-null object
16	nearby_bus_stops	6009	non-null float64
17	nearby_supermarkets	13885	non-null float64
18	nearby_shops	14271	non-null int64
19	year_built	14271	non-null int64
20	month_built	8397	non-null object
21	facilities	14271	non-null object
22	price	14271	non-null int64

dtypes: float64(9), int64(6), object(8)

memory usage: 2.5+ MB



\*Missing values will be fill with avg. of each features

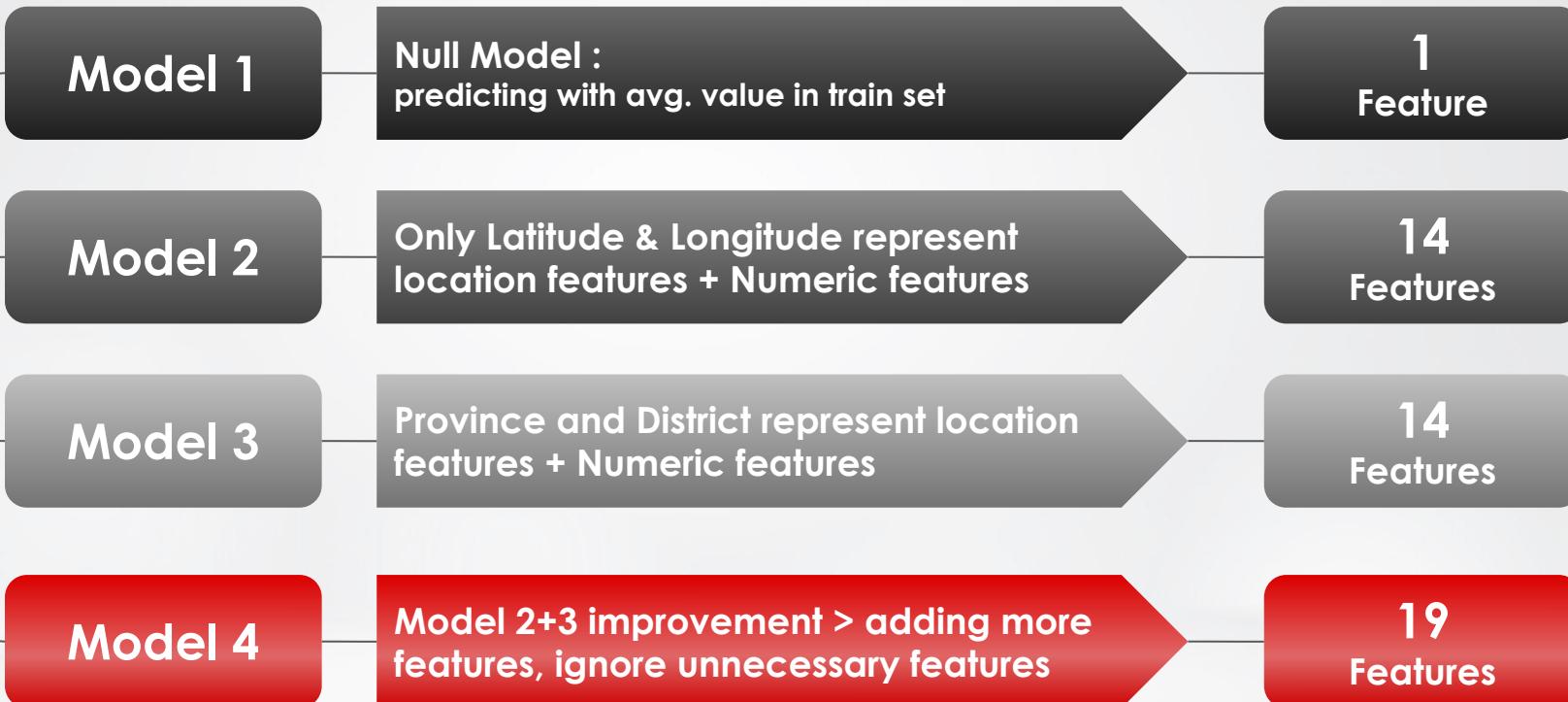
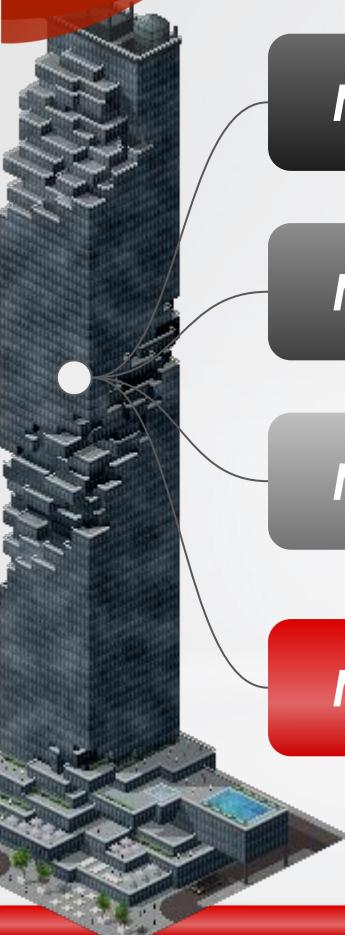


**>10% correlation  
To price**

Total\_units  
Bedrooms  
Baths  
Floor\_area  
Floor\_level  
Nearby\_stations  
Nearby\_supermarkets  
Nearby\_shops

8 Numeric Features

\*Red color indicates  
features that has missing  
values



# Model 1 : Null Model

## Features

id

province

district

sub-district

address

property type

total units

bedrooms

baths

floor area

floor level

land area

latitude

longitude

nearby stations

nearby station  
distance

nearby bus stop

nearby  
supermarket

nearby shops

year built

month built

facilities

price

## Regularization

L1: Lasso

L2: Ridge

# Model 2 : Lat & Long for location features + Numeric F.

## Features

id

province

district

sub-district

address

D

property type

total units

bedrooms

baths

floor area

floor level

land area

latitude

longitude

nearby stations

nearby station  
distance

nearby bus stop

nearby  
supermarket

nearby shops

year built

month built

C

facilities

price

## Regularization

L1: Lasso

L2: Ridge

Elastic Net

## Feature Engineering

D

Dummify

C

Count (iterate through items)

# Model 3 : Province & District features + Numeric F.

## Features

id

D

province

D

district

sub-district

address

D

property type

total units

bedrooms

baths

floor area

floor level

land area

latitude

longitude

nearby stations

nearby station  
distance

nearby bus stop

nearby  
supermarket

nearby shops

year built

month built

C

facilities

price

## Regularization

L1: Lasso

L2: Ridge

Elastic Net

## Feature Engineering

D

Dummify

C

Count (iterate through items)

# Model 4 : Model 2+3 Improvement

## Features

id

D

province

D

district

sub-district

address

D

property type

total units

bedrooms

baths

floor area

NEW

floor level

NEW

land area

latitude

longitude

nearby stations

C

nearby station  
distance

nearby bus stop

nearby  
supermarket

nearby shops

year built

month built

C

facilities

price

## Regularization

L1: Lasso

L2: Ridge

Elastic Net

## Feature Engineering

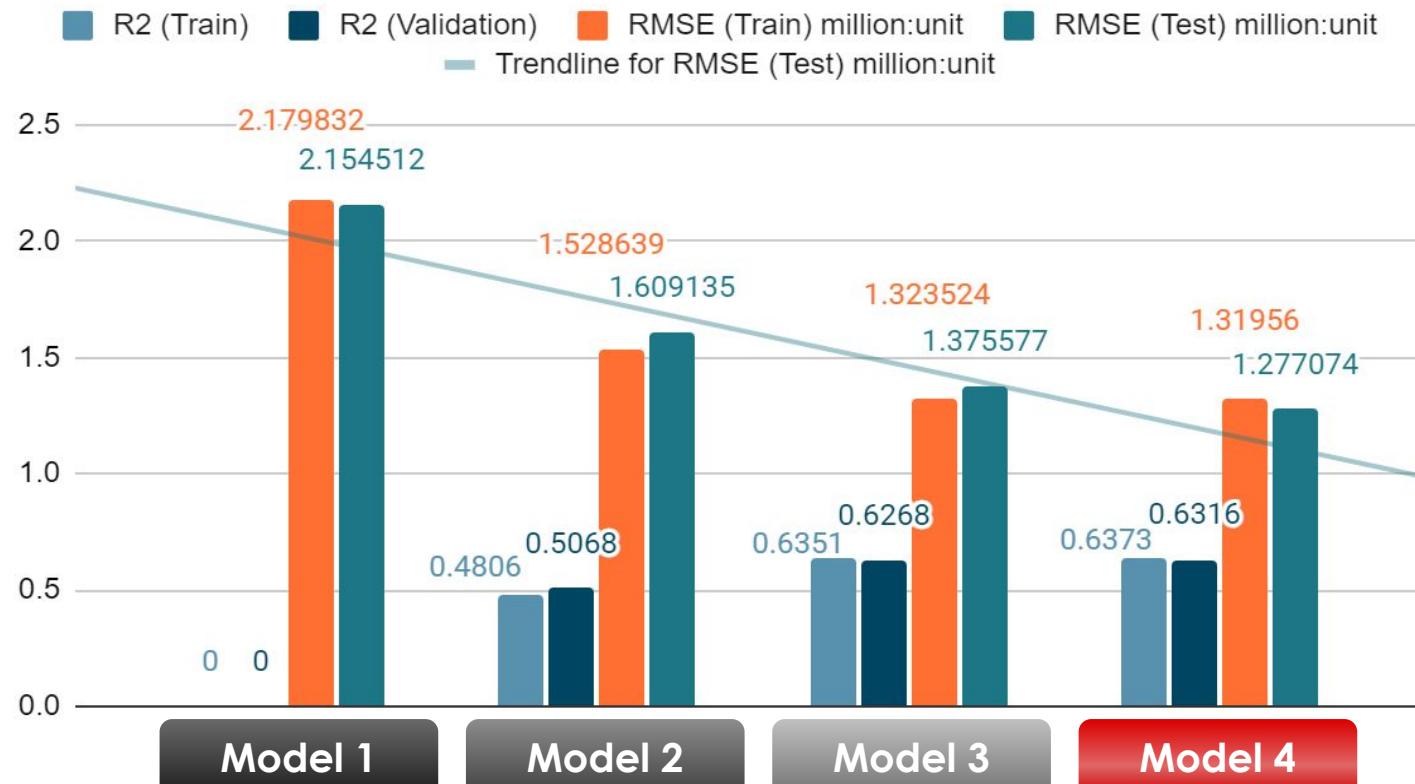
D

Dummify

C

Count (iterate through items)

## Model Comparison



## After deploying 4 models,

**Model#4** is surely perform best but might be unrealistic in term of generalization trading off with preciseness improvement from **Model#3 — (i.e. 19 features vs 14 features)**

**Whilst Model#3** has drastically different score from **Model#2** given that they have equal no. of features which proof that Province and Subdistrict are better representative of location features than Latitude and Longitude or the latter 2 are needed engineering before fitting with the model.

**For further studies,** We could dive deep into Latitude & Longitude feature engineering for better prediction

## Vital Price Determiners

### Greater, better (+ve)

Floor Area (Usable Area)  
Bedroom  
Bathroom  
Year Built  
Nearby shops  
Facility  
BTS > MRT > ARL > SRT nearby counts

### Greater, worsen (-ve)

Total Units



## Strategy Recommendation

**Real Estate Developers** may focus on floor area (usable area) > land area and try not to squeeze in too much total units as it will be effect the price. In Addition, customer convenience for living (facility, transportation and shop nearby) also plays key roles in marking up price for real estate project.

# Thank you