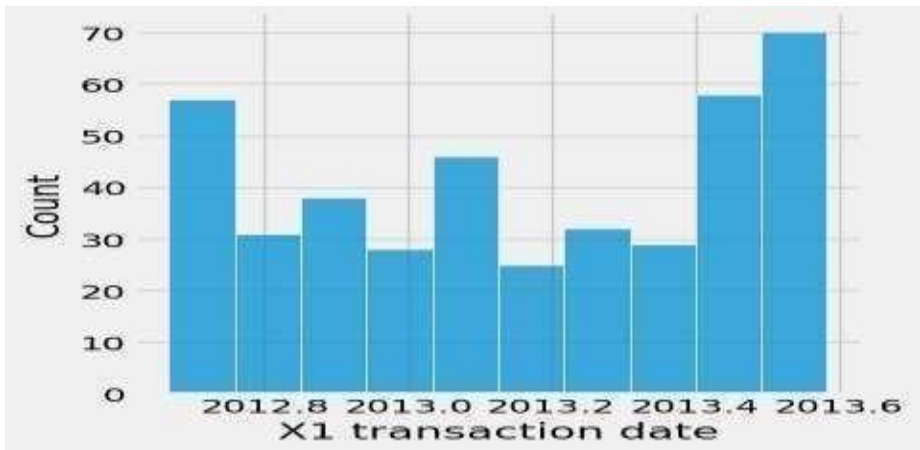


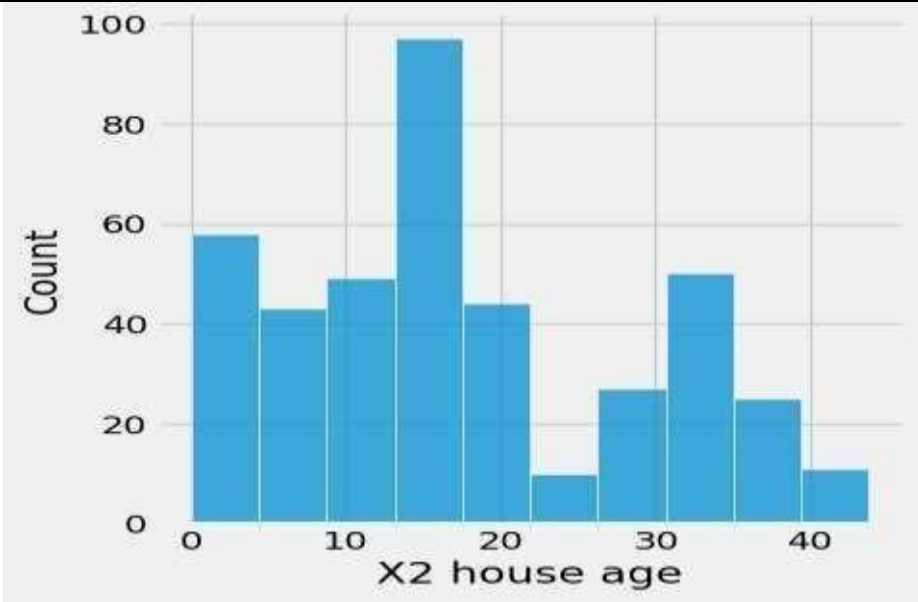
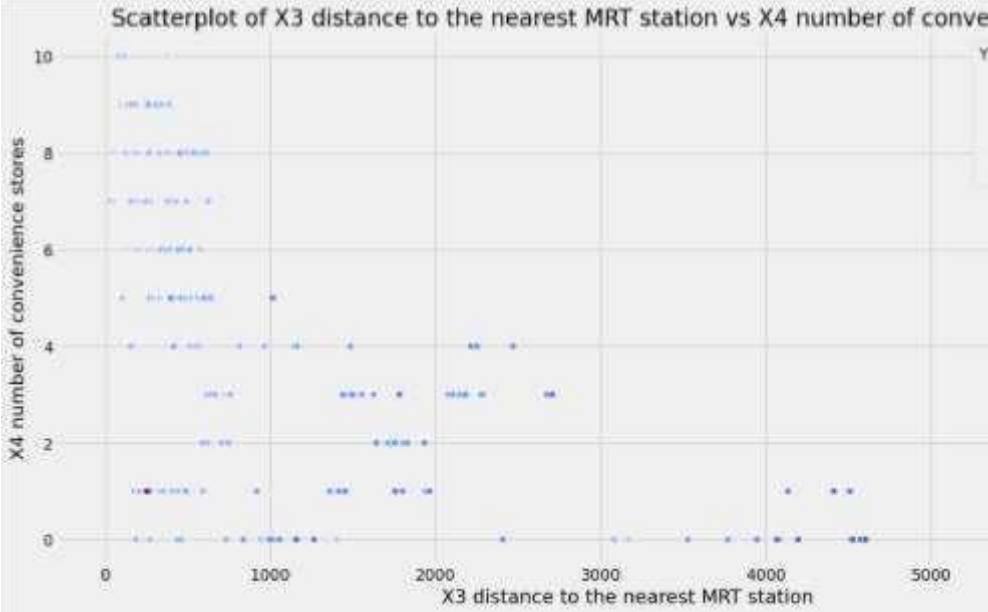
Data Collection and Preprocessing Phase

Date	8 July 2024
Team ID	740138
Project Title	Identification Of Methodology Used In Real Estate Property Valuation
Maximum Marks	6 Marks

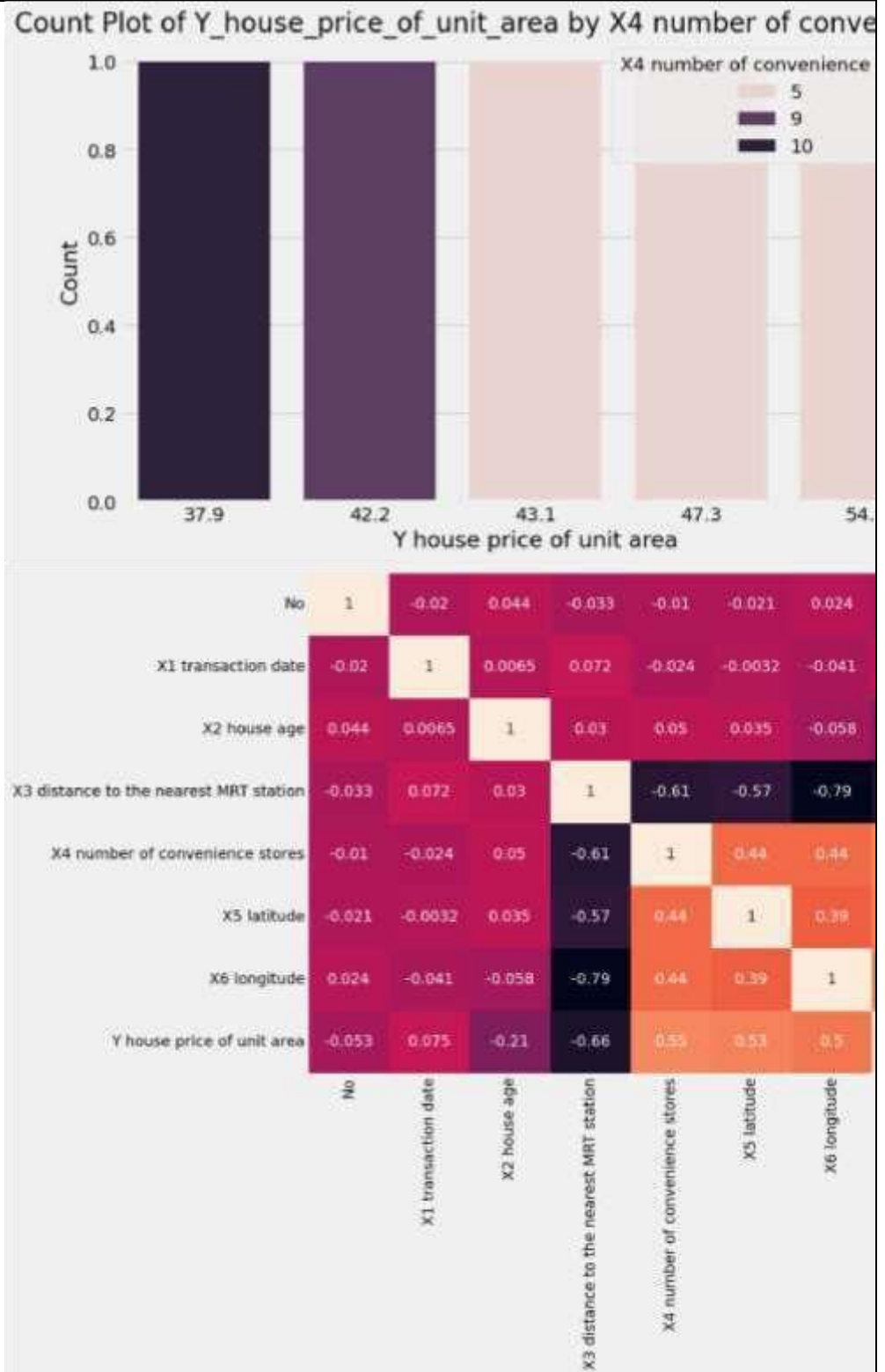
Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

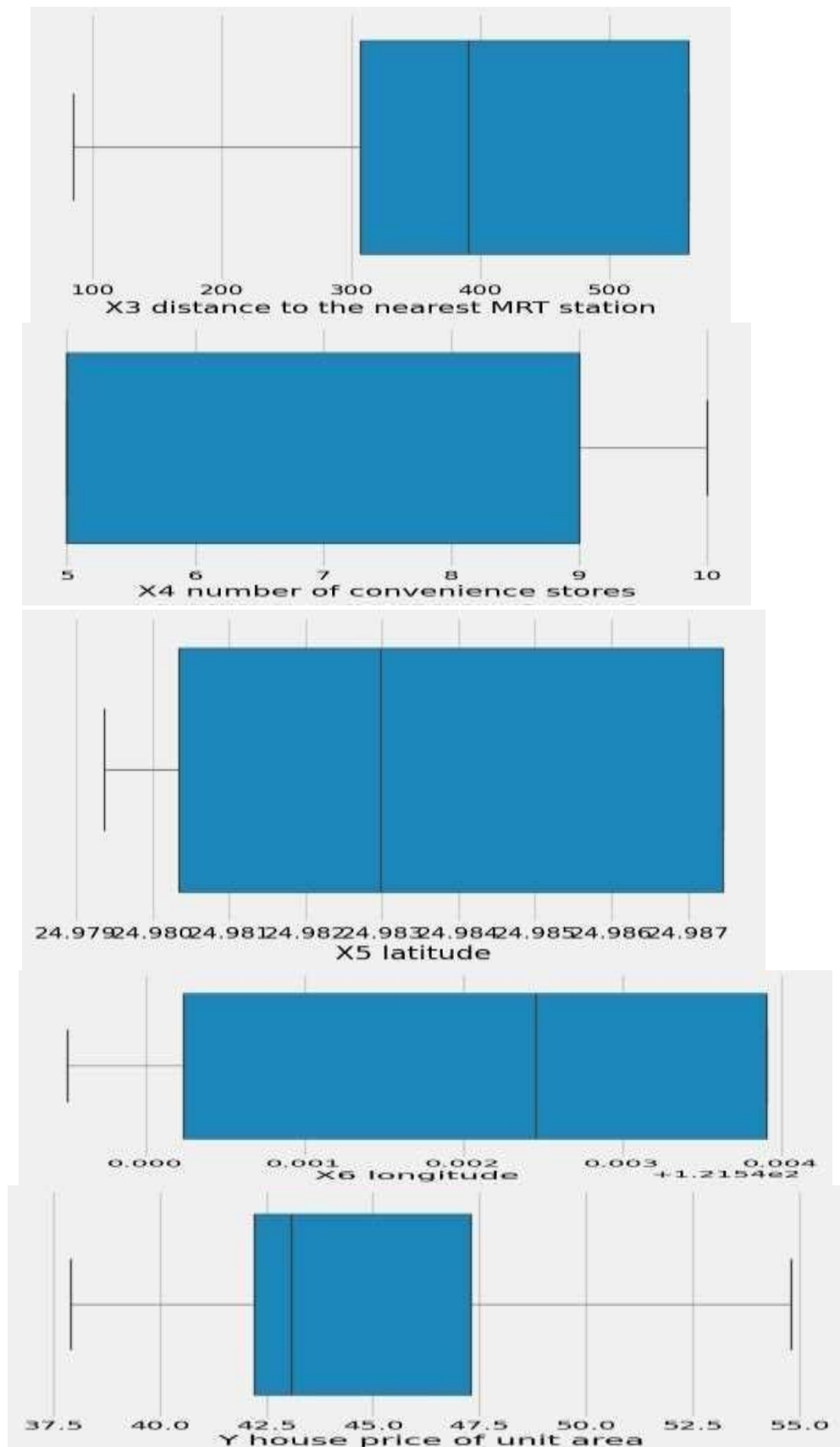
Section	Description																																																																
Data Overview	<div>Dimension: 331rowx 8columns</div> <div>Descriptive statistics:</div> <div><table><tr><th></th><th>X1 transaction date</th><th>X2 floor age</th><th>X3 distance to the nearest M50 station</th><th>X4 number of convenience stores</th><th>X5 latitude</th><th>X6 longitude</th><th>X7 house price of unit area</th></tr><tr><td>mean</td><td>2012.833333</td><td>16.000000</td><td>287.257143</td><td>4.000000</td><td>30.500000</td><td>121.500000</td><td>45.000000</td></tr><tr><td>std</td><td>0.000000</td><td>0.000000</td><td>196.244719</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>0.000000</td></tr><tr><td>min</td><td>2012.833333</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>30.500000</td><td>121.500000</td><td>0.000000</td></tr><tr><td>25%</td><td>2012.833333</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>30.500000</td><td>121.500000</td><td>0.000000</td></tr><tr><td>50%</td><td>2012.833333</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>30.500000</td><td>121.500000</td><td>0.000000</td></tr><tr><td>75%</td><td>2012.833333</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>30.500000</td><td>121.500000</td><td>0.000000</td></tr><tr><td>max</td><td>2012.833333</td><td>0.000000</td><td>0.000000</td><td>0.000000</td><td>30.500000</td><td>121.500000</td><td>0.000000</td></tr></table></div>		X1 transaction date	X2 floor age	X3 distance to the nearest M50 station	X4 number of convenience stores	X5 latitude	X6 longitude	X7 house price of unit area	mean	2012.833333	16.000000	287.257143	4.000000	30.500000	121.500000	45.000000	std	0.000000	0.000000	196.244719	0.000000	0.000000	0.000000	0.000000	min	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000	25%	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000	50%	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000	75%	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000	max	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000
		X1 transaction date	X2 floor age	X3 distance to the nearest M50 station	X4 number of convenience stores	X5 latitude	X6 longitude	X7 house price of unit area																																																									
mean	2012.833333	16.000000	287.257143	4.000000	30.500000	121.500000	45.000000																																																										
std	0.000000	0.000000	196.244719	0.000000	0.000000	0.000000	0.000000																																																										
min	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000																																																										
25%	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000																																																										
50%	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000																																																										
75%	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000																																																										
max	2012.833333	0.000000	0.000000	0.000000	30.500000	121.500000	0.000000																																																										
Univariate Analysis	 <p>A histogram showing the distribution of transaction dates (X1). The x-axis is labeled 'X1 transaction date' and ranges from 2012.8 to 2013.6. The y-axis is labeled 'Count' and ranges from 0 to 70. The bars represent the frequency of transactions for different date intervals.</p> <table><tr><th>Transaction Date Range</th><th>Count</th></tr><tr><td>2012.8 - 2012.9</td><td>58</td></tr><tr><td>2012.9 - 2013.0</td><td>31</td></tr><tr><td>2013.0 - 2013.1</td><td>38</td></tr><tr><td>2013.1 - 2013.2</td><td>28</td></tr><tr><td>2013.2 - 2013.3</td><td>46</td></tr><tr><td>2013.3 - 2013.4</td><td>25</td></tr><tr><td>2013.4 - 2013.5</td><td>32</td></tr><tr><td>2013.5 - 2013.6</td><td>29</td></tr><tr><td>2013.6 - 2013.7</td><td>58</td></tr><tr><td>2013.7 - 2013.8</td><td>70</td></tr></table>	Transaction Date Range	Count	2012.8 - 2012.9	58	2012.9 - 2013.0	31	2013.0 - 2013.1	38	2013.1 - 2013.2	28	2013.2 - 2013.3	46	2013.3 - 2013.4	25	2013.4 - 2013.5	32	2013.5 - 2013.6	29	2013.6 - 2013.7	58	2013.7 - 2013.8	70																																										
Transaction Date Range	Count																																																																
2012.8 - 2012.9	58																																																																
2012.9 - 2013.0	31																																																																
2013.0 - 2013.1	38																																																																
2013.1 - 2013.2	28																																																																
2013.2 - 2013.3	46																																																																
2013.3 - 2013.4	25																																																																
2013.4 - 2013.5	32																																																																
2013.5 - 2013.6	29																																																																
2013.6 - 2013.7	58																																																																
2013.7 - 2013.8	70																																																																

	 <p>A histogram showing the distribution of house age (X2). The x-axis is labeled 'X2 house age' and ranges from 0 to 45. The y-axis is labeled 'Count' and ranges from 0 to 100. The distribution is unimodal and slightly right-skewed, with a peak count of approximately 95 for houses aged 15-20 years.</p> <table border="1"><thead><tr><th>House Age Range</th><th>Count</th></tr></thead><tbody><tr><td>0-5</td><td>58</td></tr><tr><td>5-10</td><td>43</td></tr><tr><td>10-15</td><td>49</td></tr><tr><td>15-20</td><td>95</td></tr><tr><td>20-25</td><td>44</td></tr><tr><td>25-30</td><td>10</td></tr><tr><td>30-35</td><td>27</td></tr><tr><td>35-40</td><td>50</td></tr><tr><td>40-45</td><td>25</td></tr><tr><td>45-50</td><td>11</td></tr></tbody></table>	House Age Range	Count	0-5	58	5-10	43	10-15	49	15-20	95	20-25	44	25-30	10	30-35	27	35-40	50	40-45	25	45-50	11
House Age Range	Count																						
0-5	58																						
5-10	43																						
10-15	49																						
15-20	95																						
20-25	44																						
25-30	10																						
30-35	27																						
35-40	50																						
40-45	25																						
45-50	11																						
Bivariate Analysis	 <p>A scatterplot titled 'Scatterplot of X3 distance to the nearest MRT station vs X4 number of convenience stores'. The x-axis is labeled 'X3 distance to the nearest MRT station' and ranges from 0 to 5000. The y-axis is labeled 'X4 number of convenience stores' and ranges from 0 to 10. The plot shows a negative correlation, with a regression line and a shaded confidence interval. The data points are scattered, with a higher density at lower distances and higher numbers of convenience stores.</p> <table border="1"><thead><tr><th>X3 distance to the nearest MRT station</th><th>X4 number of convenience stores</th></tr></thead><tbody><tr><td>0</td><td>10</td></tr><tr><td>1000</td><td>5</td></tr><tr><td>2000</td><td>4</td></tr><tr><td>3000</td><td>3</td></tr><tr><td>4000</td><td>2</td></tr><tr><td>5000</td><td>1</td></tr></tbody></table>	X3 distance to the nearest MRT station	X4 number of convenience stores	0	10	1000	5	2000	4	3000	3	4000	2	5000	1								
X3 distance to the nearest MRT station	X4 number of convenience stores																						
0	10																						
1000	5																						
2000	4																						
3000	3																						
4000	2																						
5000	1																						

Multivariate Analysis



Handled
Outliers and
Anomalies



Data Preprocessing Code Screenshots

Loading
Data

```

In [444]:
df = pd.read_csv('11000001RealEstateValuationDataSet.csv')

df.head()
Out[444]:
   No  X1 transaction date  X2 house age  X3 distance to the nearest MRT station  X4 number of convenience stores  X5 latitude  X6 longitude  Y house price of unit area
0  1  2012.017            30.0          81.070000          50.000000          23.96336          121.04234          37.8
1  2  2012.017            18.5          59.544750          50.000000          23.96919          121.04919          41.2
2  3  2012.003            12.1          61.984400          50.000000          23.96936          121.04977          32.1
3  4  2012.006            19.4          59.134400          50.000000          23.96733          121.04935          36.3
4  5  2012.008            20.0          59.566600          50.000000          23.97707          121.04545          35.1

...

490  410  2012.008            12.7          60.023300          50.000000          23.97036          121.05030          33.4
491  411  2012.067             6.6          60.000000          50.000000          23.97433          121.04119          35.9
492  412  2012.228            13.0          59.999900          50.000000          23.97733          121.03986          40.8
493  413  2012.000             8.1          59.000100          50.000000          23.96619          121.04607          33.0

```

Finding
& Handling
Missing Data

```

In [ ]: df.isnull(inplace=True)

In [ ]: df.info()
Out[ ]:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414 entries, 0 to 413
Data columns (total 8 columns):
 #   Column                                     Non-Null Count  Dtype
---  --
 0   No                                         414 non-null    int64
 1   X1 transaction date                       414 non-null    float64
 2   X2 house age                             414 non-null    float64
 3   X3 distance to the nearest MRT station    414 non-null    float64
 4   X4 number of convenience stores           414 non-null    int64
 5   X5 latitude                              414 non-null    float64
 6   X6 longitude                             414 non-null    float64
 7   Y house price of unit area               414 non-null    float64
dtypes: float64(6), int64(2)
memory usage: 26.8 KB

In [ ]: df.isnull().any()
Out[ ]:
No: False
X1 transaction date: False
X2 house age: False
X3 distance to the nearest MRT station: False
X4 number of convenience stores: False
X5 latitude: False
X6 longitude: False
Y house price of unit area: False
dtype: bool

```

Data
Transformat
ion

-

Feature
Engineering

Attached the code in final submission

Save
Processed
Data

```

In [ ]: import pickle
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor
rf_model = RandomForestRegressor()
scaler = StandardScaler()
with open('price.pkl', 'wb') as f:
    pickle.dump(rf_model, f)
with open('scale.pkl', 'wb') as f:
    pickle.dump(scaler, f)

from google.colab import files
files.download('price.pkl')

from google.colab import files
files.download('scale.pkl')

from google.colab import files
files.download('/content/drive/MyDrive/ dataset/real estate valuation data set.csv')

```