
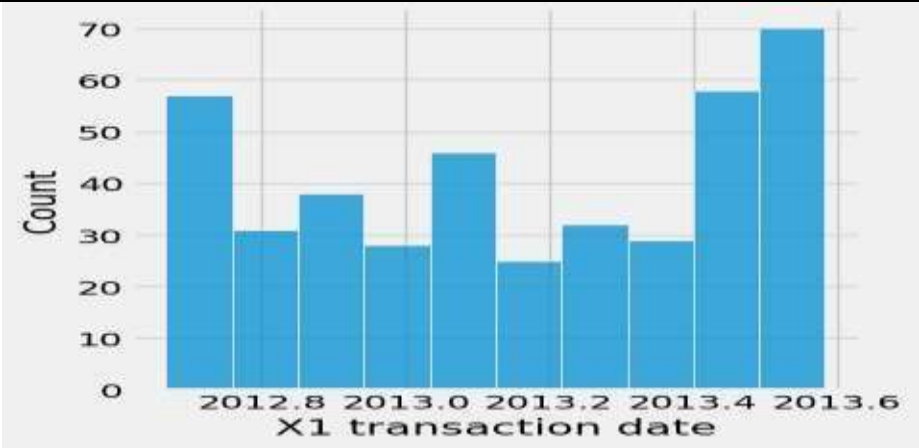


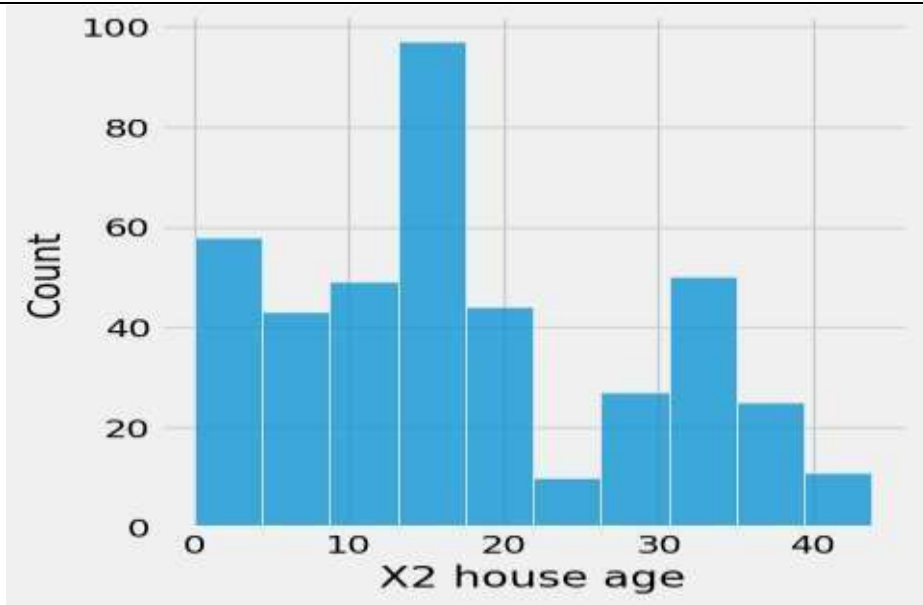
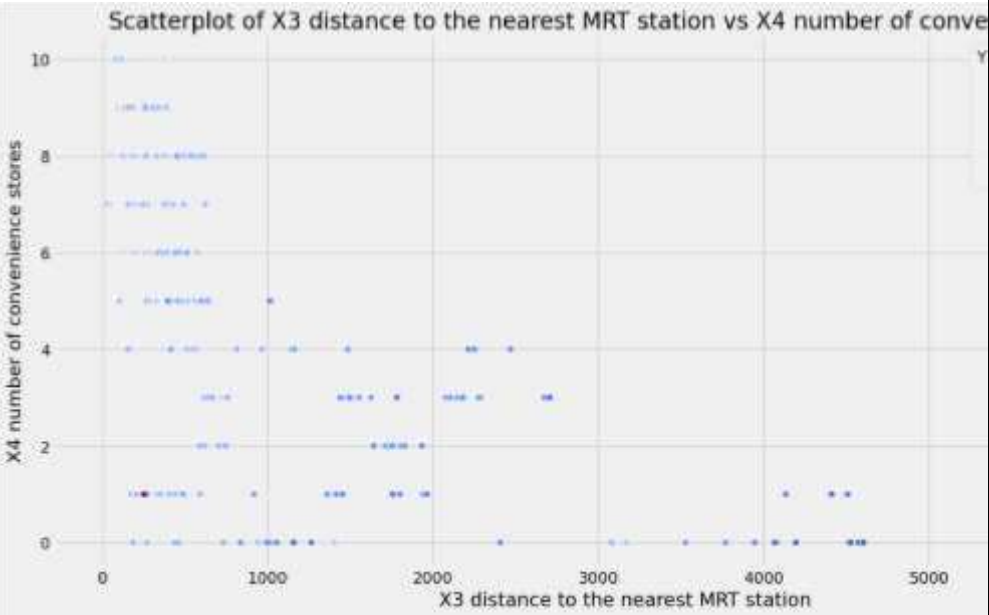
Data Collection and Preprocessing Phase

Date	8 July 2024
Team ID	739991
Project Title	Identification Of Methodology Used In Real Estate Property Valuation
Maximum Marks	6 Marks

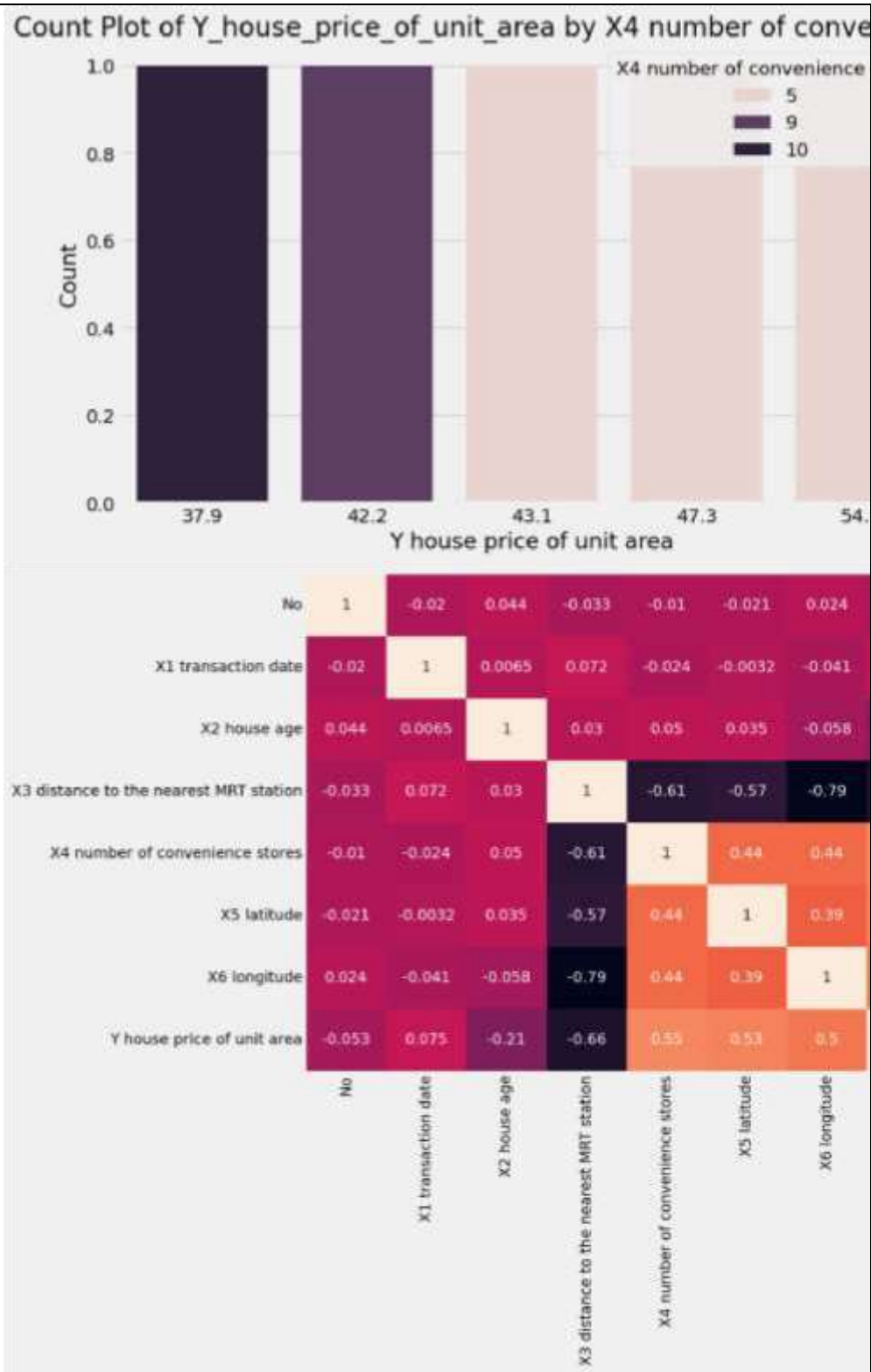
Data Exploration and Preprocessing Template

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

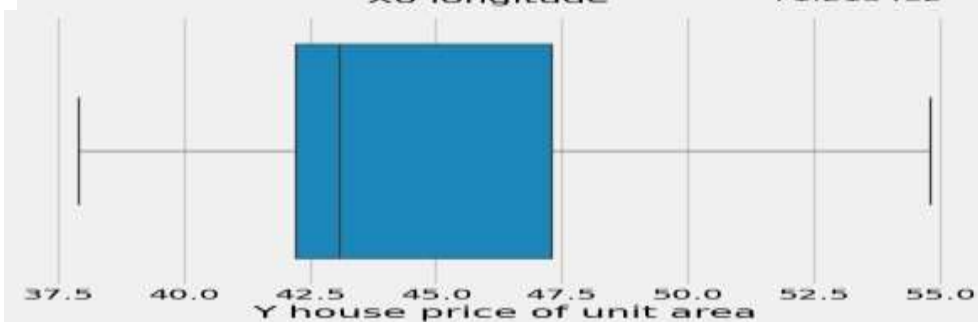
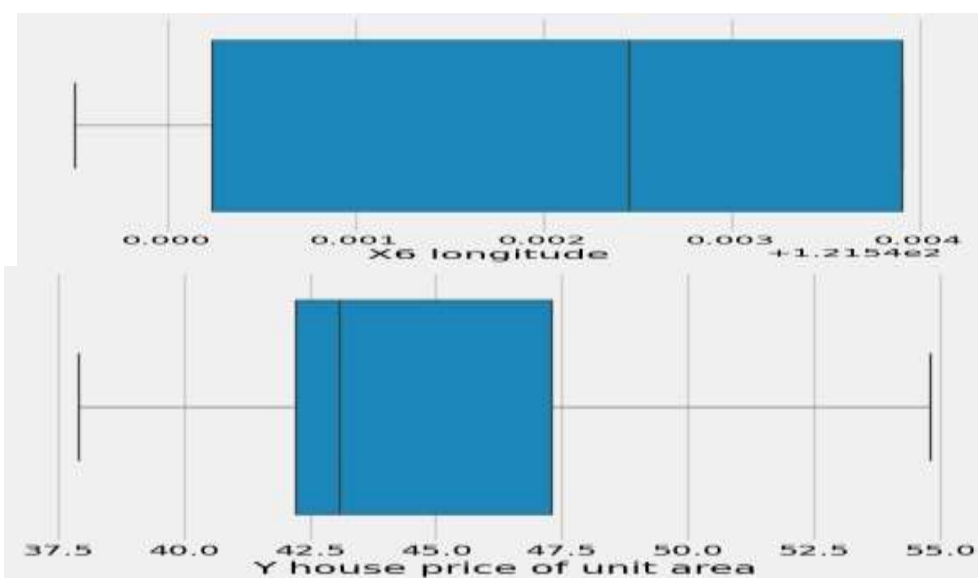
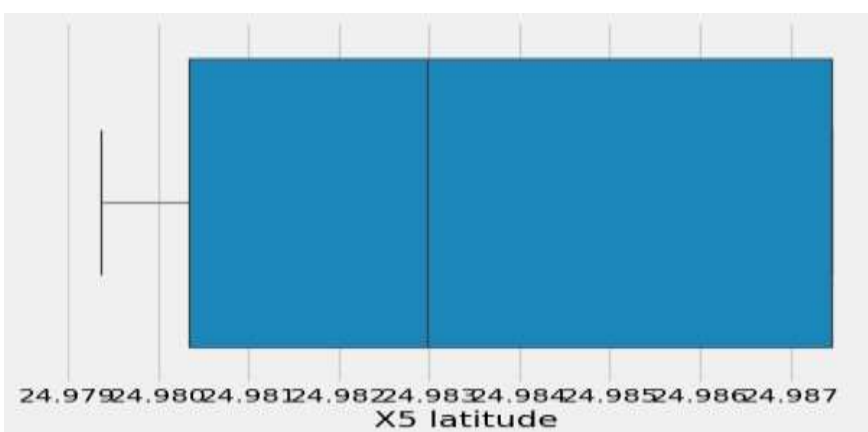
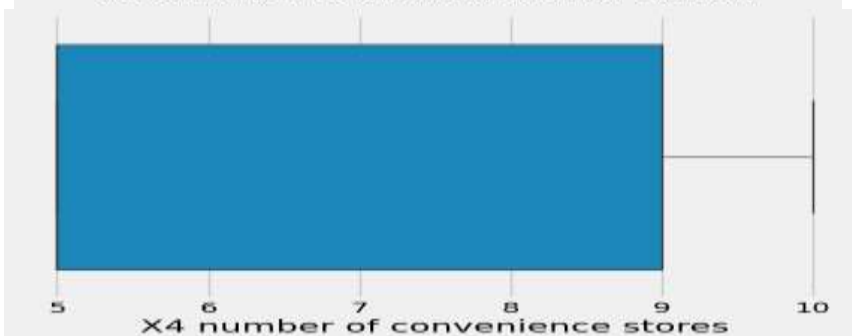
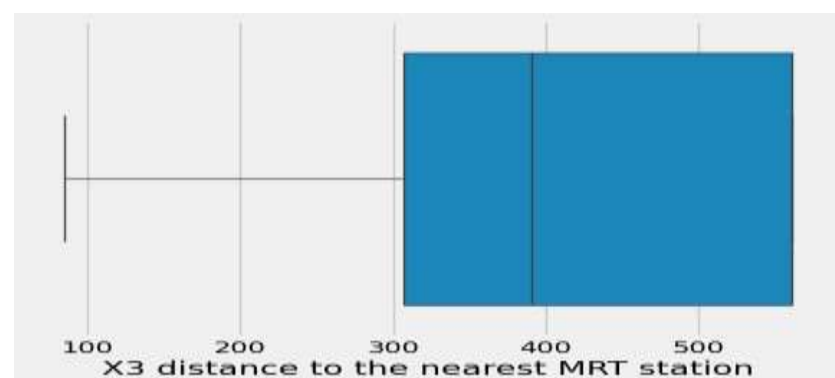
Section	Description
Data Overview	<div>Dimension: 331rowsx 8columnns <u>Descriptive</u> <u>statistics:</u></div> <div></div>
Univariate Analysis	<div></div>

	 <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 between 15 and 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 dense cluster of points at low distances (below 1000) and low numbers of stores (below 4). There are several outliers at higher distances, notably one point at approximately (4500, 10) and another at (4500, 8).</p>																						

Multivariate Analysis



Handled  
Outliers and  
Anomalies



## Data Preprocessing Code Screenshots

Loading  
Data

```

1) Read data.csv
2) dt.head()
3) dt.info()
4) dt.describe()

```

#	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	2012.017	30.0	84.07002	0	23.96258	121.54024	37.8
1	2012.017	19.5	396.59470	0	24.96258	121.54024	42.2
2	2013.000	13.0	541.98440	0	25.00746	121.54024	47.3
3	2013.000	13.3	541.98440	0	24.96776	121.54024	34.8
4	2013.000	3.0	396.59440	0	24.97707	121.54024	40.5

dt.info() output:

```

RangeIndex: 414 entries, 0 to 413
Data columns (total 8 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   X1 transaction date                  414 non-null    float64
1   X2 house age                        414 non-null    float64
2   X3 distance to the nearest MRT station 414 non-null    float64
3   X4 number of convenience stores      414 non-null    int64
4   X5 latitude                         414 non-null    float64
5   X6 longitude                        414 non-null    float64
6   Y house price of unit area          414 non-null    float64
dtypes: float64(6), int64(1)
memory usage: 26.8 KB

```

dt.describe() output:

```

X1 transaction date    2013.000    45
X2 house age           16.45466
X3 distance to the nearest MRT station  84.07002
X4 number of convenience stores         0
X5 latitude            24.97437
X6 longitude           121.54024
Y house price of unit area            40.9

```

Finding  
& Handling  
Missing  
Data

```

1) dt.dropna(inplace=True)
2) dt.info()
3) dt.isnull().any()

```

dt.info() output:

```

RangeIndex: 414 entries, 0 to 413
Data columns (total 8 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                ---
0   X1 transaction date                  414 non-null    float64
1   X2 house age                        414 non-null    float64
2   X3 distance to the nearest MRT station 414 non-null    float64
3   X4 number of convenience stores      414 non-null    int64
4   X5 latitude                         414 non-null    float64
5   X6 longitude                        414 non-null    float64
6   Y house price of unit area          414 non-null    float64
dtypes: float64(6), int64(1)
memory usage: 26.8 KB

```

dt.isnull().any() output:

```

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
dtypes: bool

```

Data  
Transformat  
ion

Feature  
Engineering

Attached the code in final submission

Save  
Processed  
Data

```

1) 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)

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

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

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

```

