


Data Collection and Preprocessing Phase

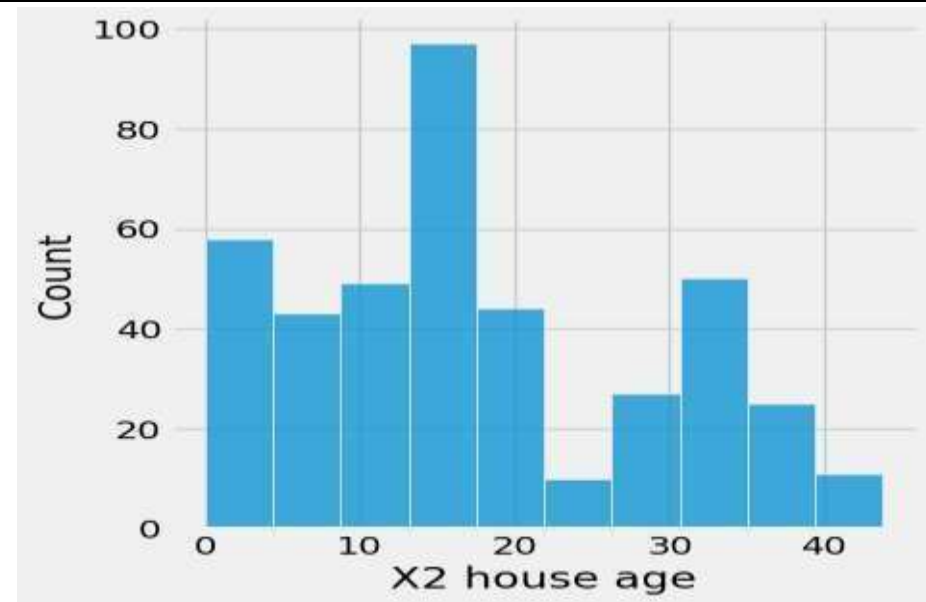
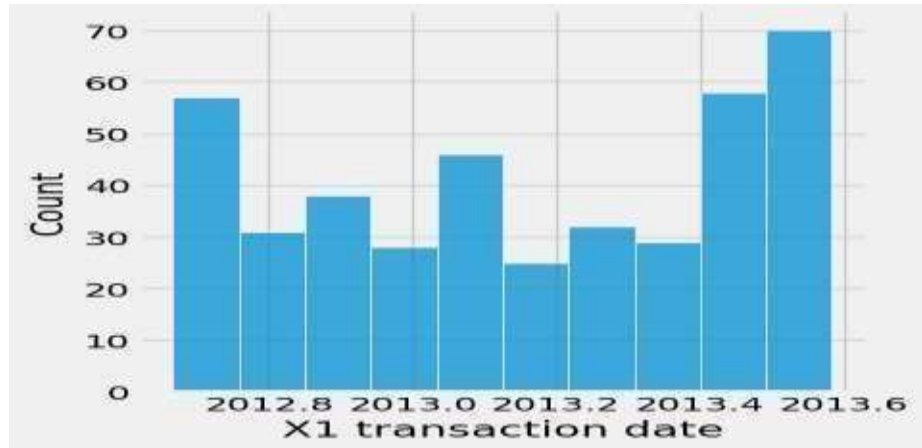
Date	8 July 2024
Team ID	739996
Project Title	Identification Of Methodology Used In Real Estate Property Valuation
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

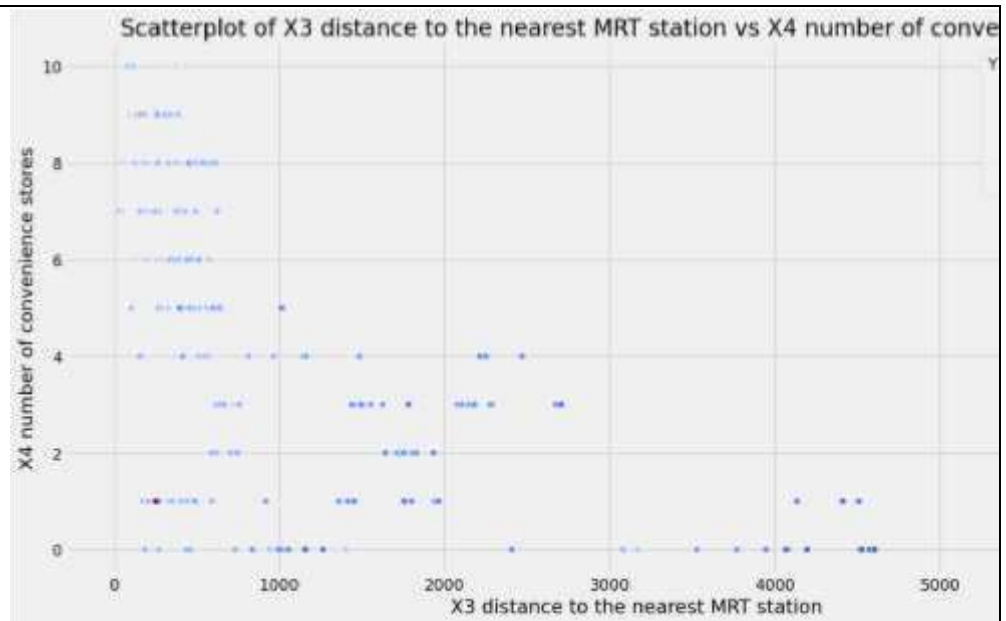
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	<div>Dimension: 331rowx 8column</div> <div>Descriptive statistics:</div> <div></div>

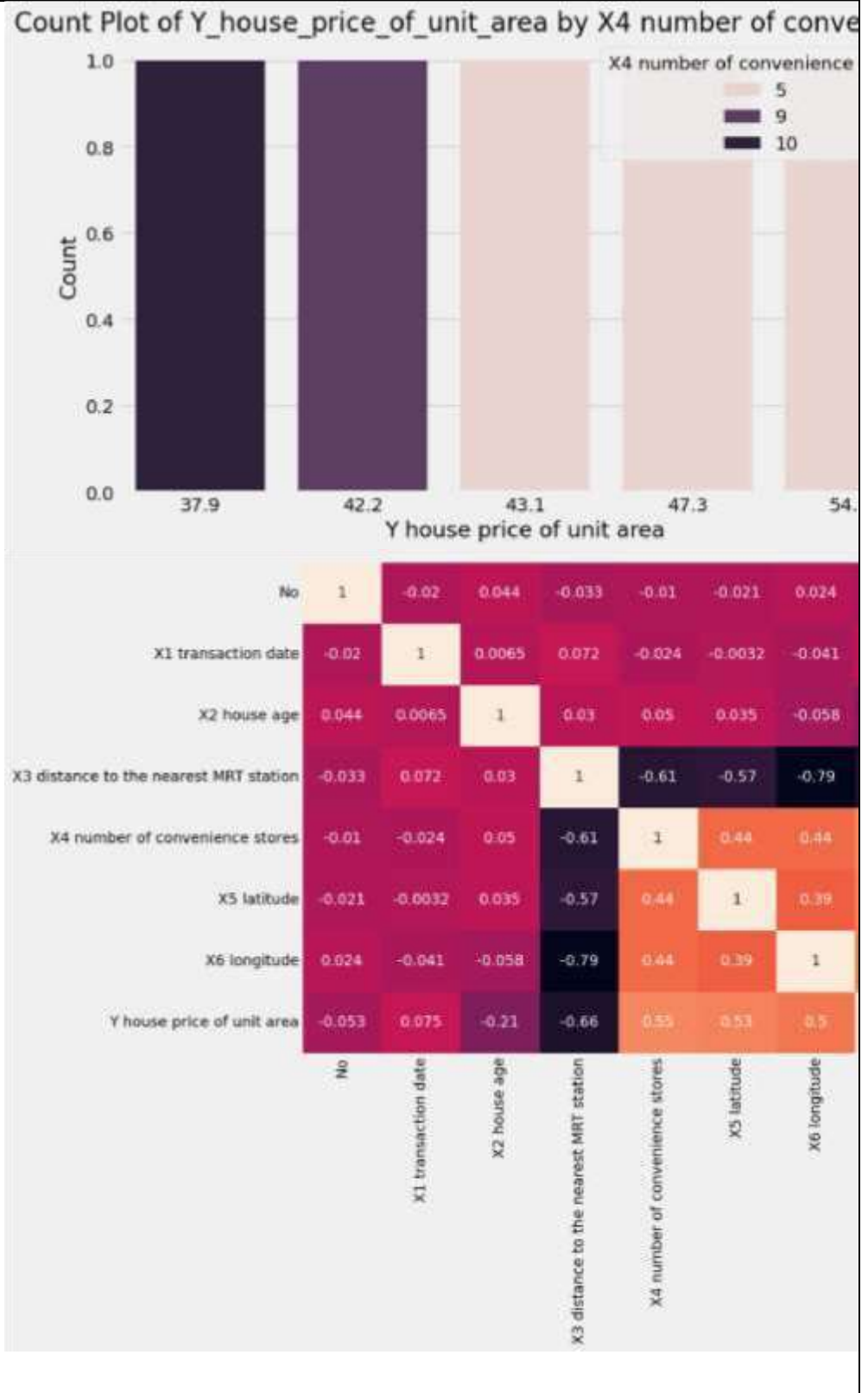
Univariate
Analysis



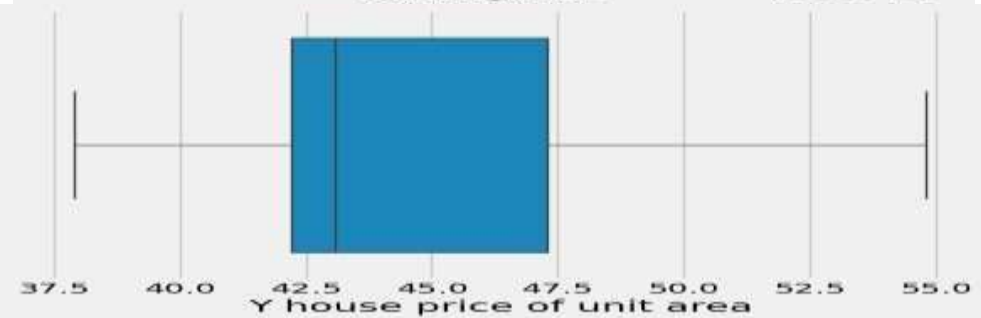
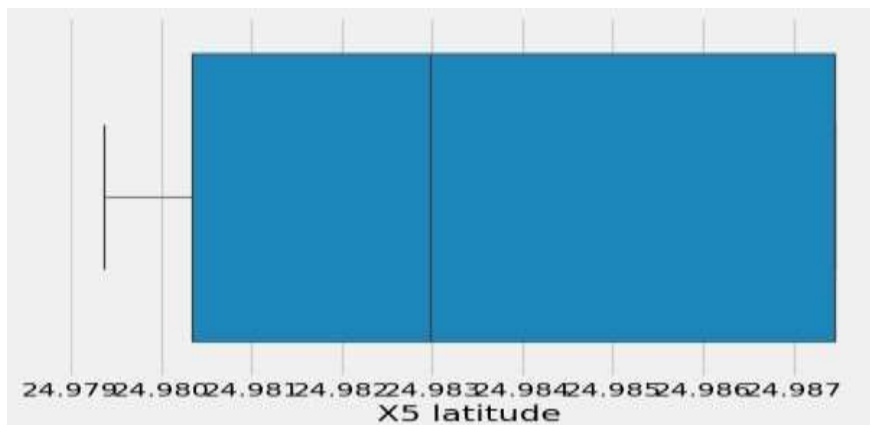
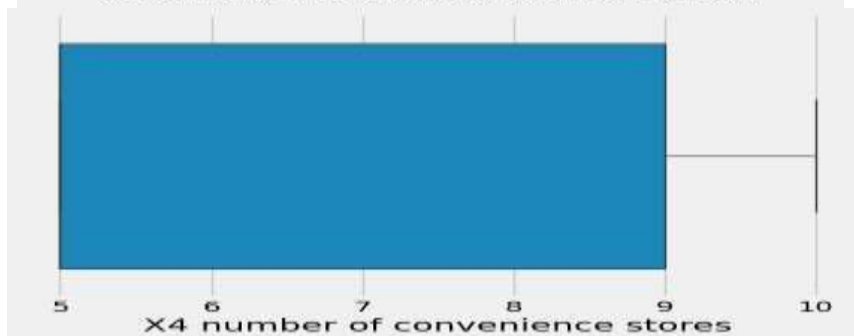
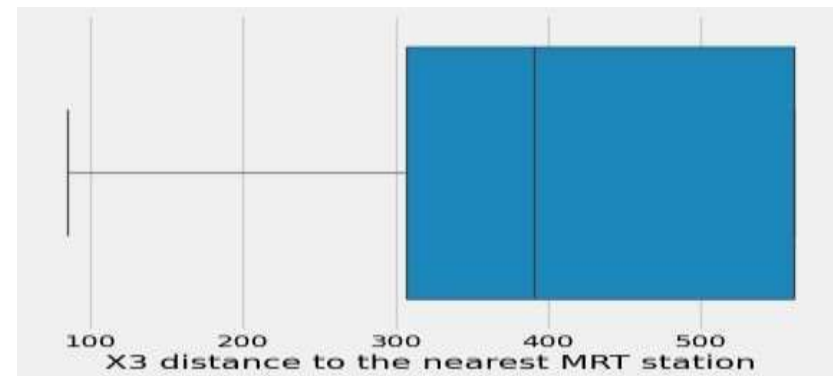
Bivariate
Analysis



Multivariate Analysis



Handled
Outliers and
Anomalies



Data Preprocessing Code Screenshots

Loading Data

```

1 |> drop_unused_vars(1:ncol(my_data)/2) # Remove every column with only one value
2 |> my_data %>%
3 |> summarise(
4 |>   ID = transaction_id, ID_base = age, ID_distance = the nearest ID station, ID_number = transaction status, ID_latitude = ID longitude, ID_longitude = ID base price of unit time
5 |> )
6 |>
7 |>
8 |>
9 |>
10 |>
11 |>
12 |>
13 |>
14 |>
15 |>
16 |>
17 |>
18 |>
19 |>
20 |>
21 |>
22 |>
23 |>
24 |>
25 |>
26 |>
27 |>
28 |>
29 |>
30 |>
31 |>
32 |>
33 |>
34 |>
35 |>
36 |>
37 |>
38 |>
39 |>
40 |>
41 |>
42 |>
43 |>
44 |>
45 |>
46 |>
47 |>
48 |>
49 |>
50 |>
51 |>
52 |>
53 |>
54 |>
55 |>
56 |>
57 |>
58 |>
59 |>
60 |>
61 |>
62 |>
63 |>
64 |>
65 |>
66 |>
67 |>
68 |>
69 |>
70 |>
71 |>
72 |>
73 |>
74 |>
75 |>
76 |>
77 |>
78 |>
79 |>
80 |>
81 |>
82 |>
83 |>
84 |>
85 |>
86 |>
87 |>
88 |>
89 |>
90 |>
91 |>
92 |>
93 |>
94 |>
95 |>
96 |>
97 |>
98 |>
99 |>
100 |>

```

ID	ID_base	ID_distance	ID_number	ID_latitude	ID_longitude	ID_base price of unit time
1	2013.217	33.0	34.07002	0	14.94004	121.54204
2	2013.217	19.5	35.04470	0	14.94004	121.54204
3	2013.200	13.5	34.10440	0	14.94004	121.54204
4	2013.200	13.5	34.10440	0	14.94004	121.54204
5	2013.200	13.5	34.10440	0	14.94004	121.54204
6	2013.200	13.5	34.10440	0	14.94004	121.54204
7	2013.200	13.5	34.10440	0	14.94004	121.54204
8	2013.200	13.5	34.10440	0	14.94004	121.54204
9	2013.200	13.5	34.10440	0	14.94004	121.54204
10	2013.200	13.5	34.10440	0	14.94004	121.54204
11	2013.200	13.5	34.10440	0	14.94004	121.54204
12	2013.200	13.5	34.10440	0	14.94004	121.54204
13	2013.200	13.5	34.10440	0	14.94004	121.54204
14	2013.200	13.5	34.10440	0	14.94004	121.54204
15	2013.200	13.5	34.10440	0	14.94004	121.54204
16	2013.200	13.5	34.10440	0	14.94004	121.54204
17	2013.200	13.5	34.10440	0	14.94004	121.54204
18	2013.200	13.5	34.10440	0	14.94004	121.54204
19	2013.200	13.5	34.10440	0	14.94004	121.54204
20	2013.200	13.5	34.10440	0	14.94004	121.54204
21	2013.200	13.5	34.10440	0	14.94004	121.54204
22	2013.200	13.5	34.10440	0	14.94004	121.54204
23	2013.200	13.5	34.10440	0	14.94004	121.54204
24	2013.200	13.5	34.10440	0	14.94004	121.54204
25	2013.200	13.5	34.10440	0	14.94004	121.54204
26	2013.200	13.5	34.10440	0	14.94004	121.54204
27	2013.200	13.5	34.10440	0	14.94004	121.54204
28	2013.200	13.5	34.10440	0	14.94004	121.54204
29	2013.200	13.5	34.10440	0	14.94004	121.54204
30	2013.200	13.5	34.10440	0	14.94004	121.54204
31	2013.200	13.5	34.10440	0	14.94004	121.54204
32	2013.200	13.5	34.10440	0	14.94004	121.54204
33	2013.200	13.5	34.10440	0	14.94004	121.54204
34	2013.200	13.5	34.10440	0	14.94004	121.54204
35	2013.200	13.5	34.10440	0	14.94004	121.54204
36	2013.200	13.5	34.10440	0	14.94004	121.54204
37	2013.200	13				

Finding & Handling Missing Data

```
[ ] dt.dropna(inplace=True)

dt.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 8 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   X1 transaction date                   418 non-null    int64
1   X2 house age                          418 non-null    float64
2   X3 distance to the nearest MRT station 418 non-null    float64
3   X4 number of convenience stores       418 non-null    int64
4   X5 latitude                           418 non-null    float64
5   X6 longitude                           418 non-null    float64
6   Y house price of unit area            418 non-null    float64
dtypes: float64(6), int64(2)
memory usage: 26.8 KB

dt.isnull().any()

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
dtypes: bool
```

Data
Transformat
ion

Feature Engineering	Attached the code in final submission
---------------------	---------------------------------------

Save
Processed
Data

```
[.] 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')
```