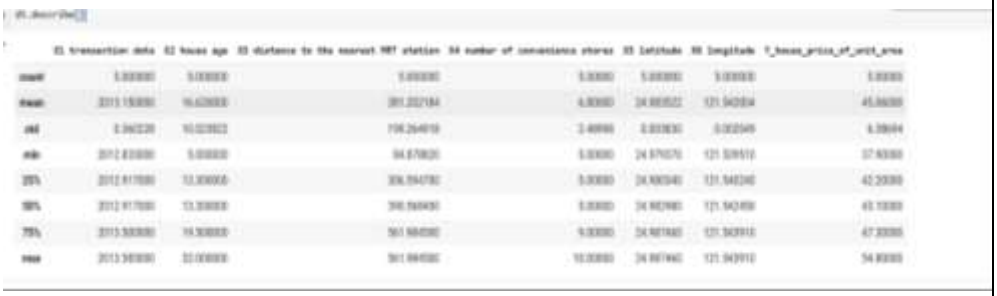
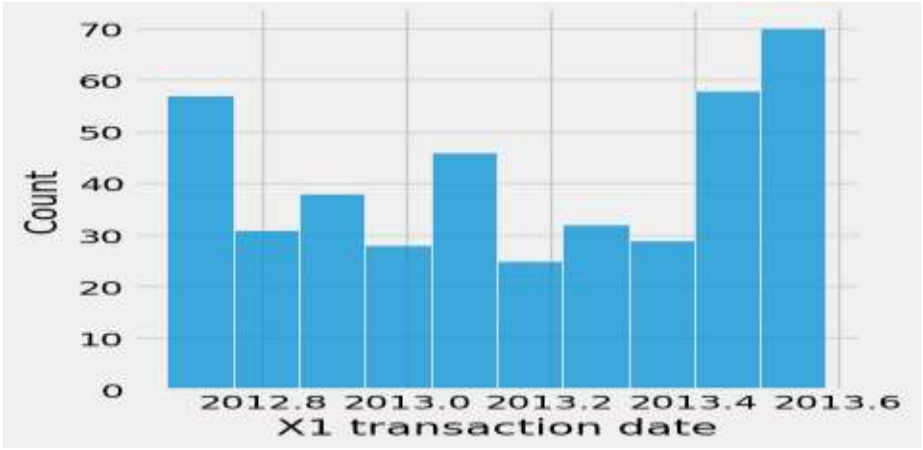
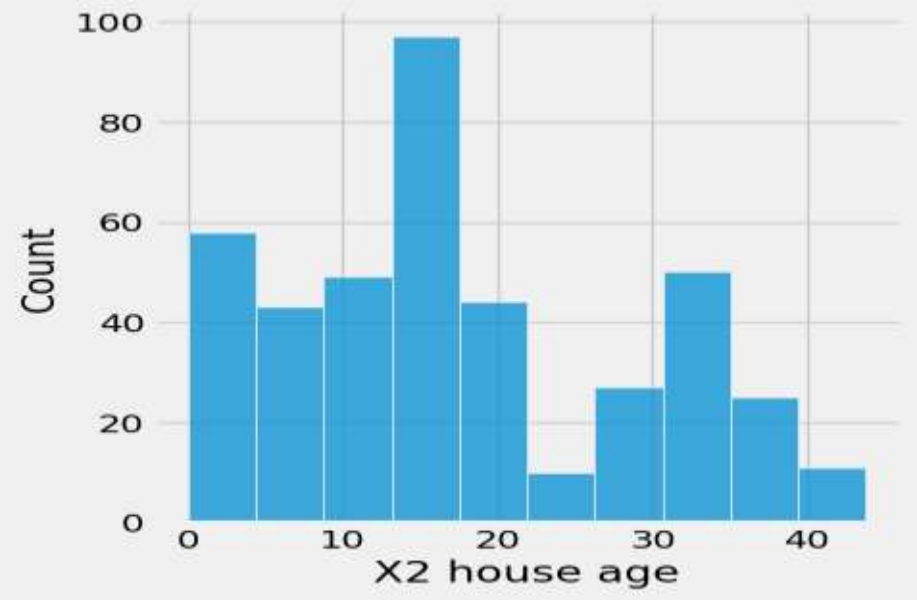
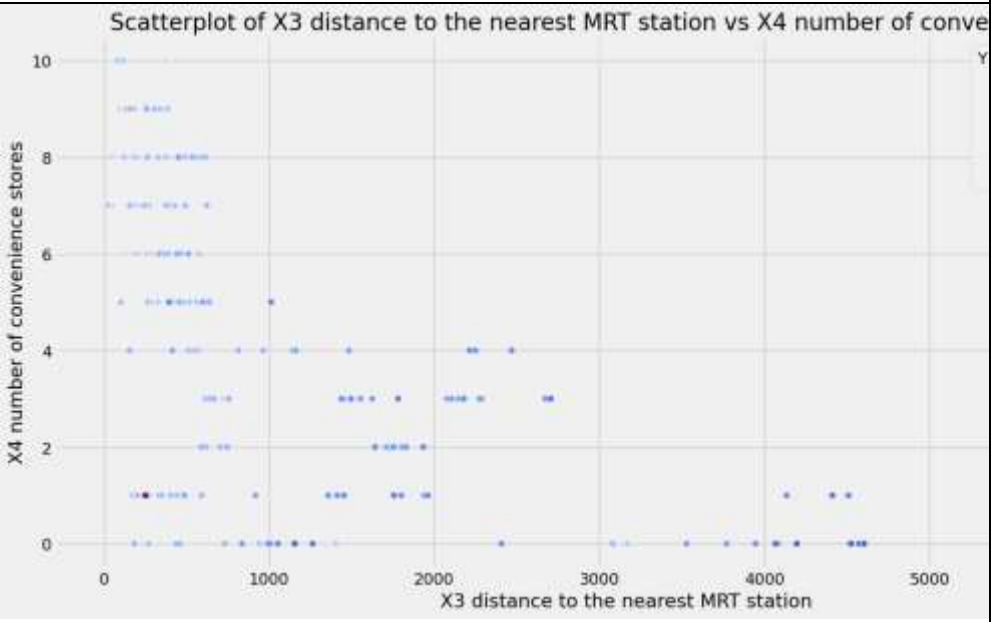
Data Collection and Preprocessing Phase

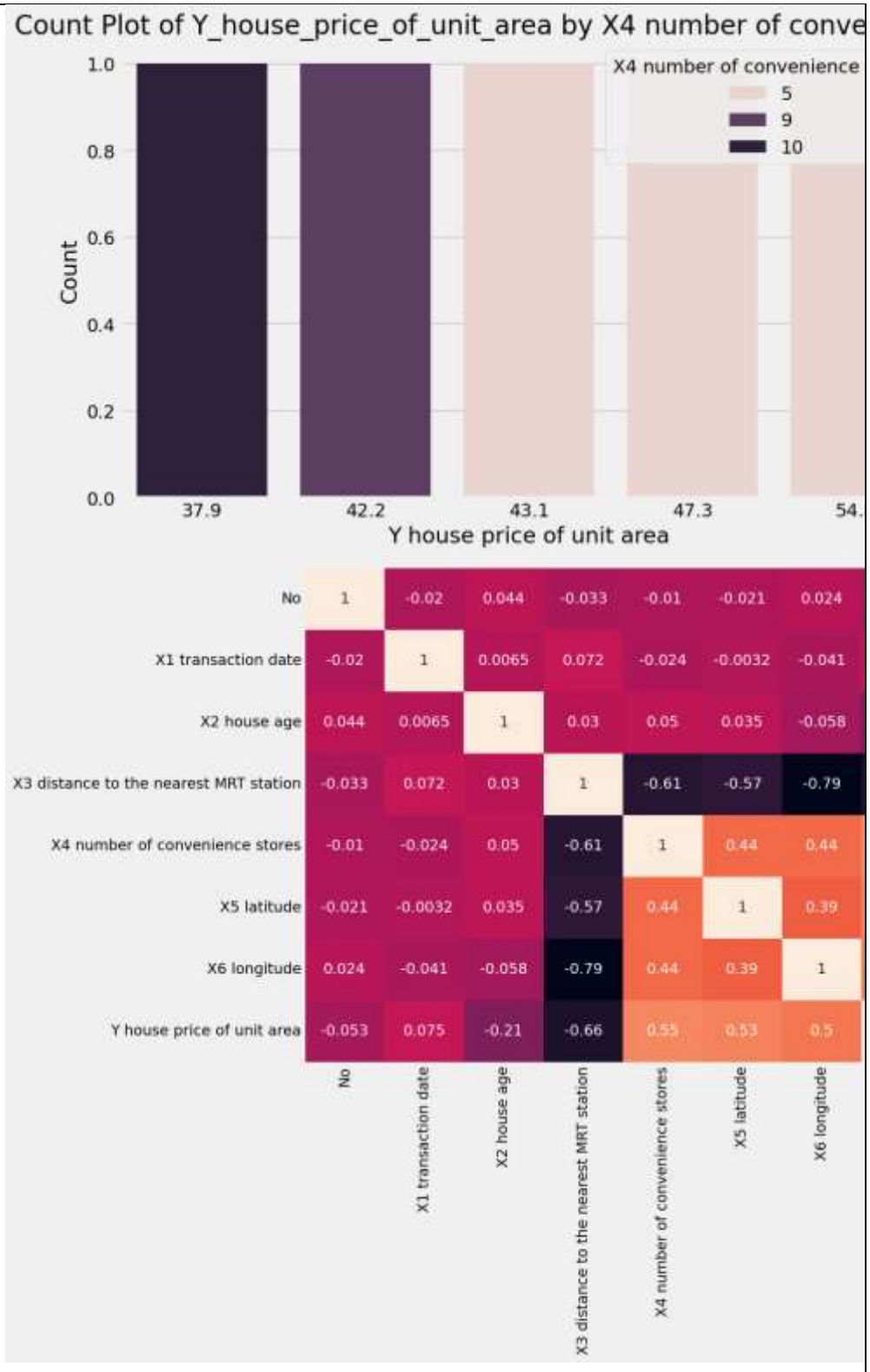
Date	8 July 2024
Team ID	739996
Project Title	Identification Of Methodology Used In Real Estate Property Valuation
Maximum Marks	6 Marks

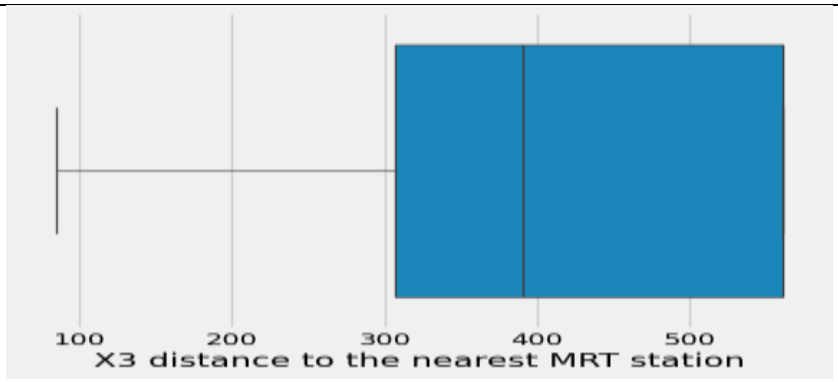
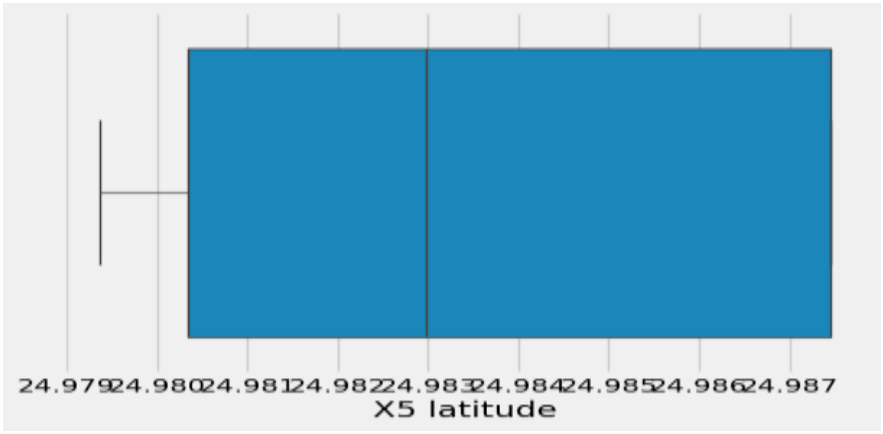
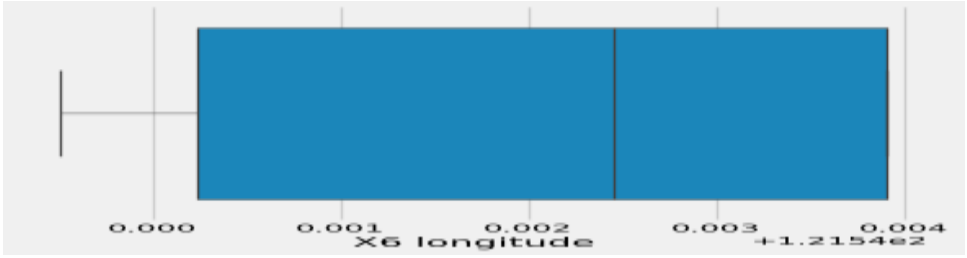
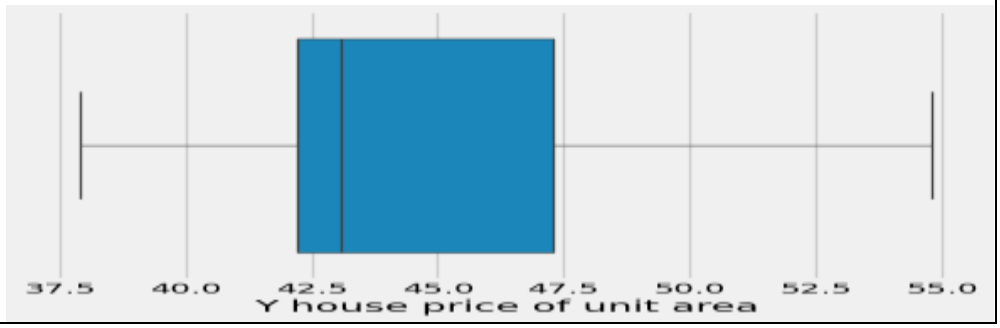
Data Exploration and Preprocessing Template

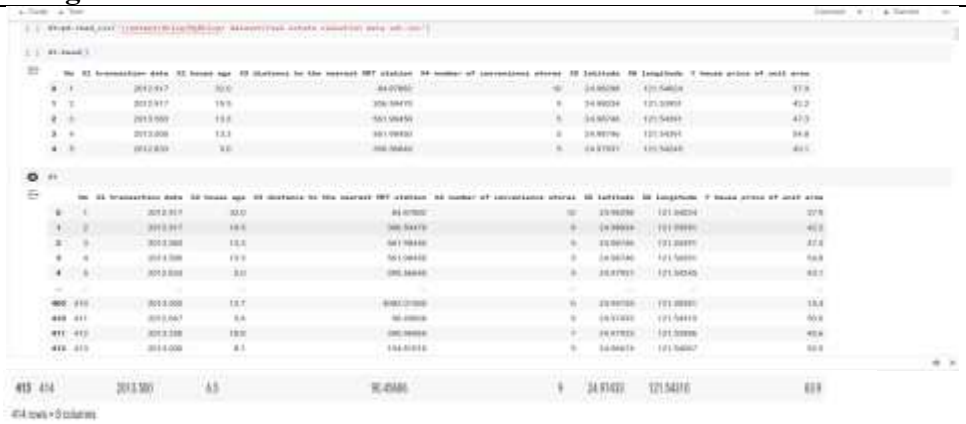
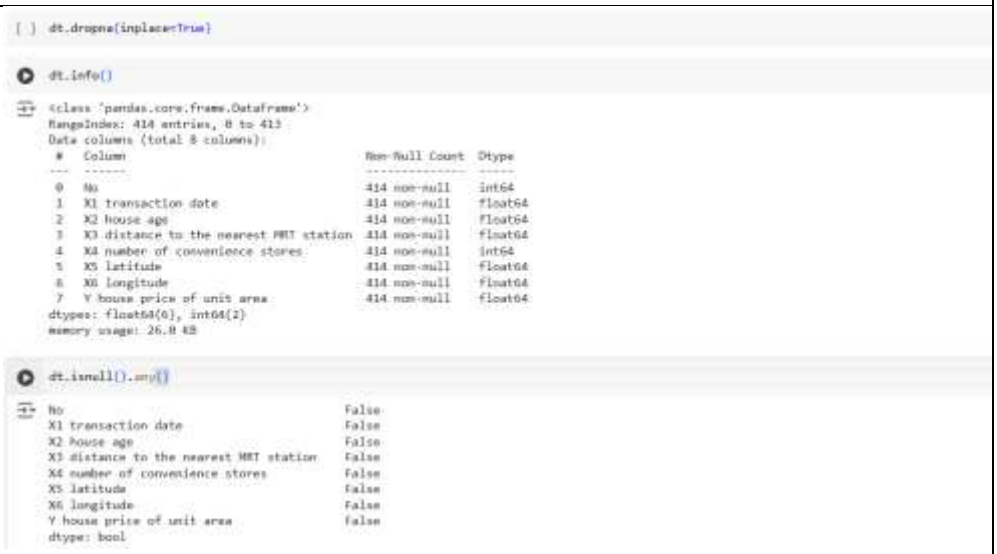
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	Dimension: 3008rows x 6columns Descriptive statistics: 
Univariate Analysis	

	
Bivariate Analysis	

Multivariate Analysis	

Handled Outliers and Anomalies	    

Data Preprocessing Code Screenshots	
Loading Data	
Finding &Handling Missing Data	
Data Transformation	-
Feature Engineering	Attached the code in final submission
Save Processed Data	

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

dt.info()

<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: 26.0 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
dtype: bool
```

```
[ ] 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')
```