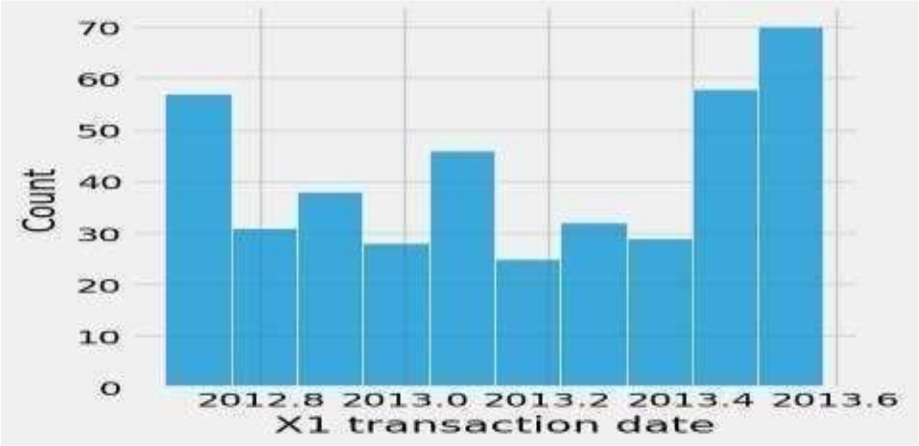
Data Collection and Preprocessing Phase

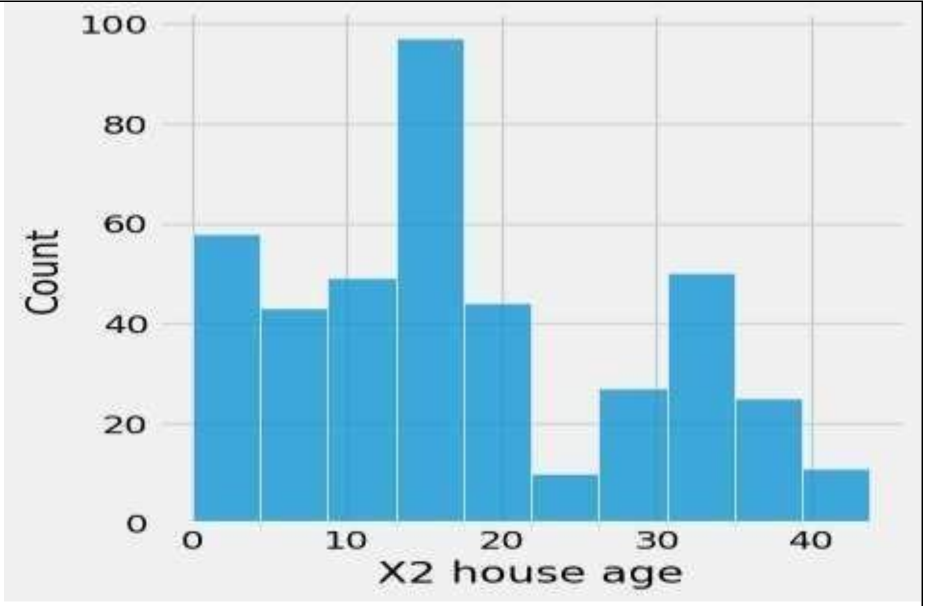
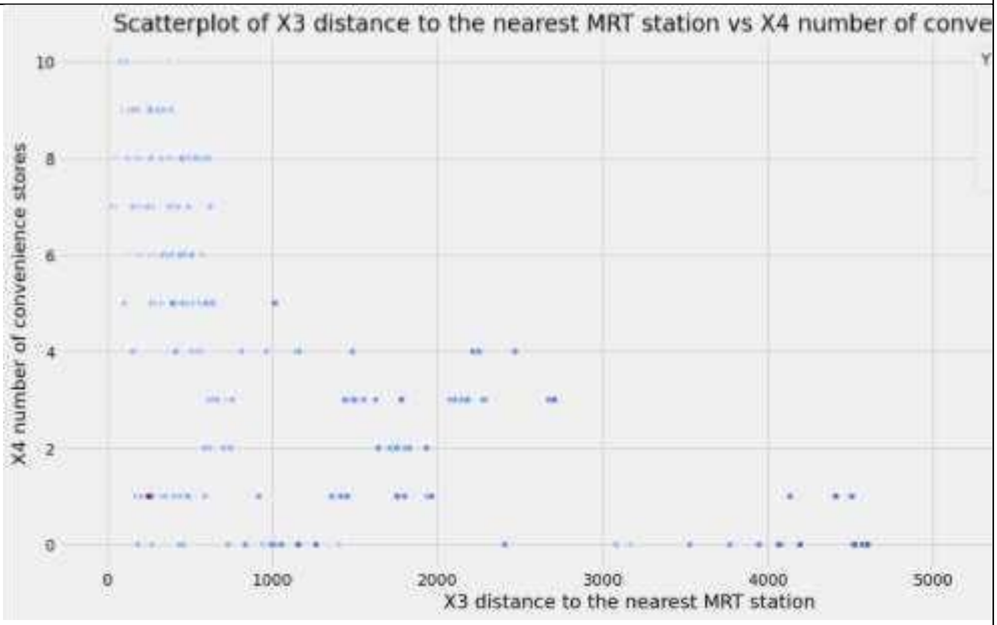
Date	8 July 2024
Team ID	740138
Project Title	Identification Of Methodology Used In Real Estate Property Valuation
Maximum Marks	6 Marks

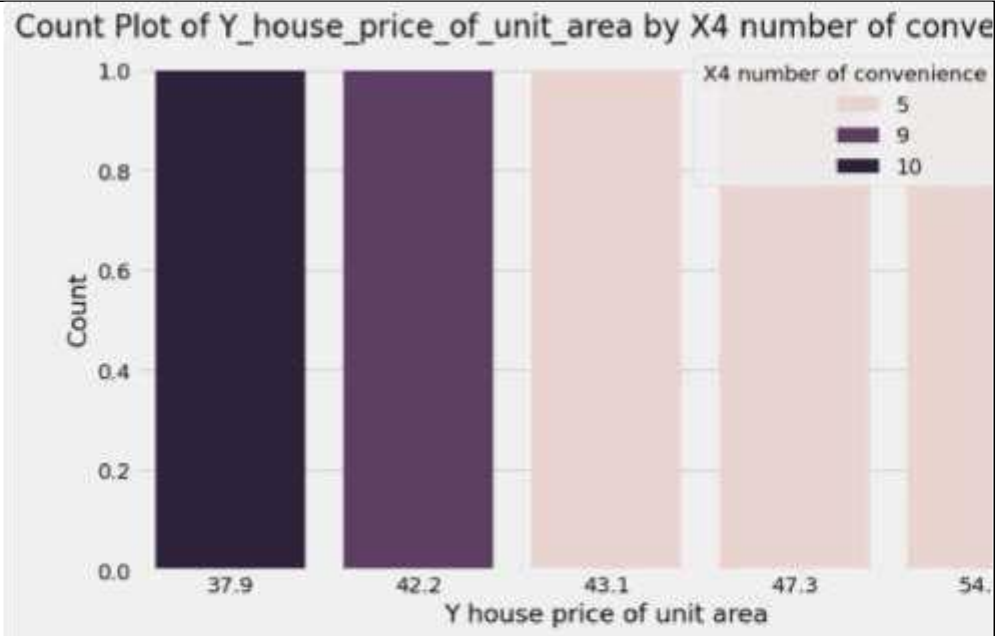
Data Exploration and Preprocessing Template

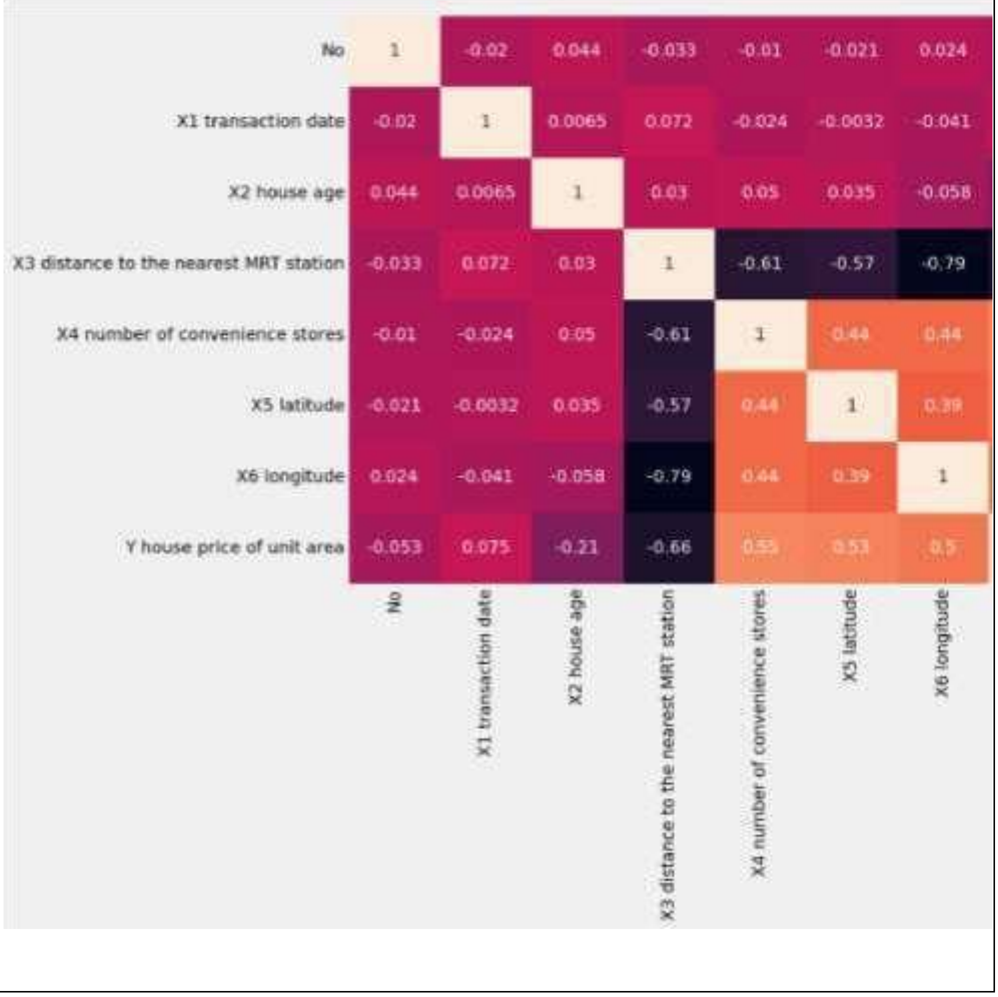
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	Dimension: 331rowsx 8columns Descriptive statistics:
Univariate Analysis	

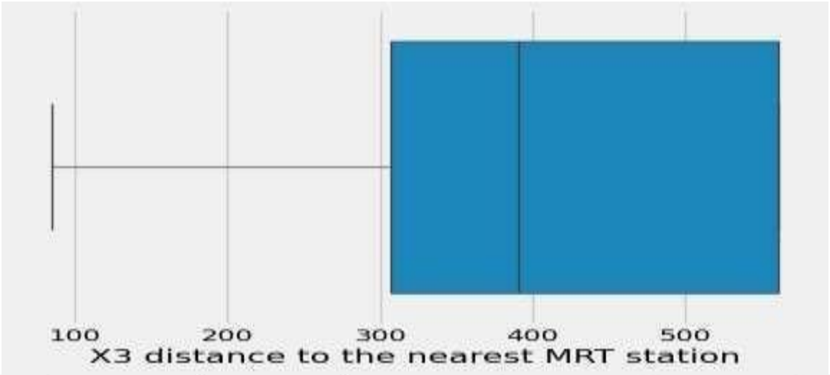
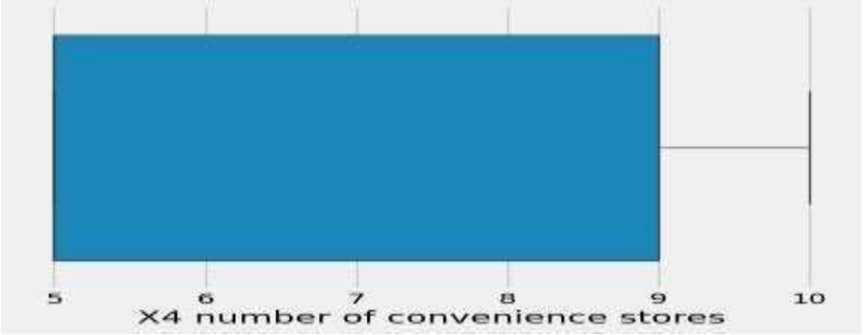
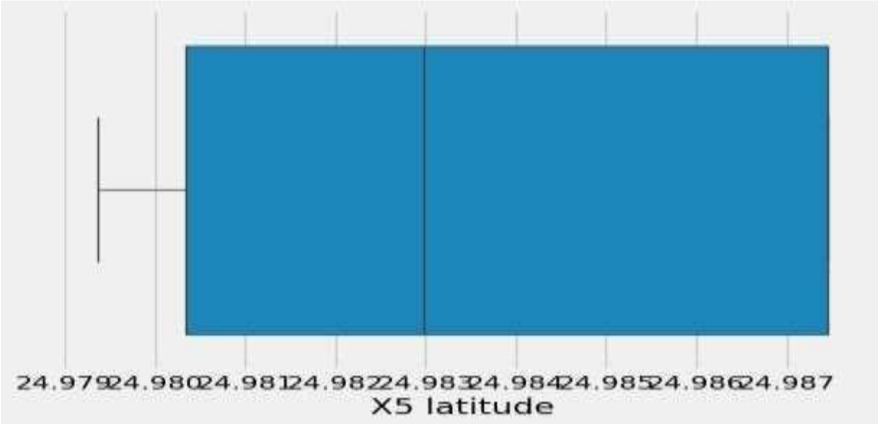
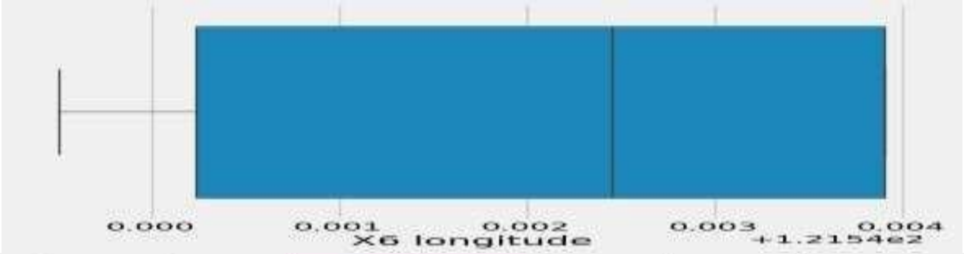
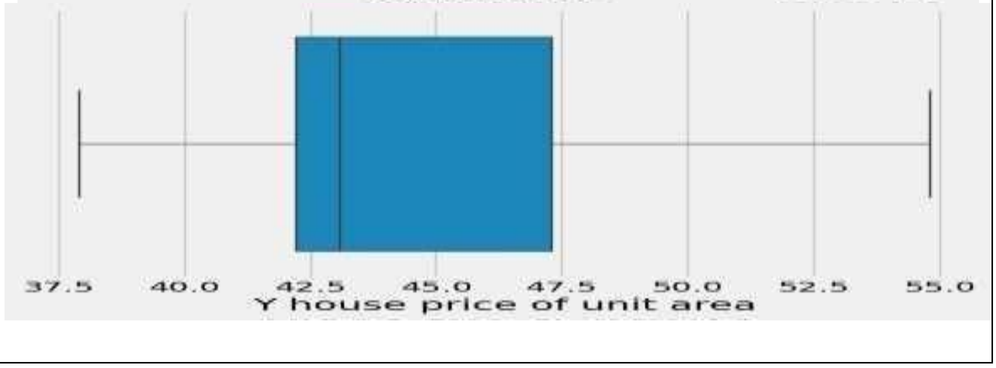
	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y_house_price_of_unit_area
count	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
mean	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
std	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
min	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
50%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
75%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
max	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

	
Bivariate Analysis	

<table>
<tr>
<td>Multivariate Analysis</td>
<td>


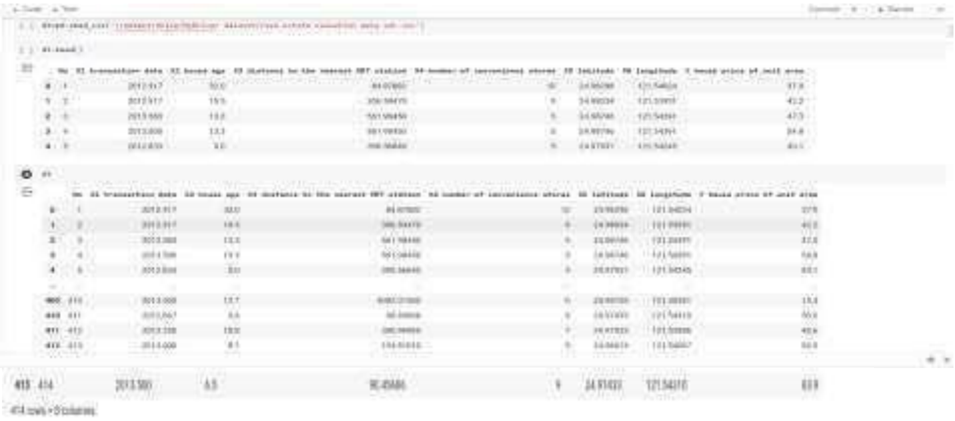


</td>
</tr>
</table>

Handled Outliers and Anomalies	    

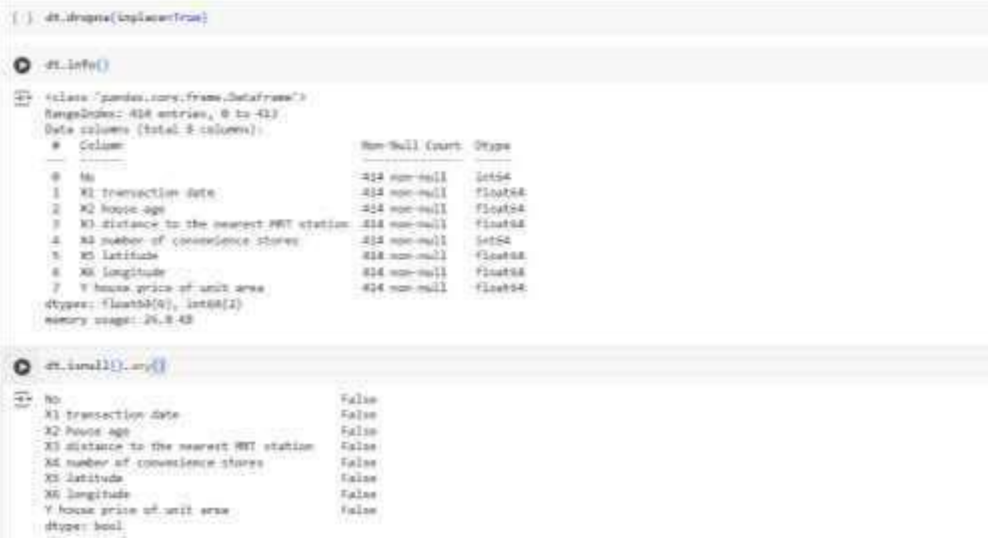
Data Preprocessing Code Screenshots

Loading Data



```
[ ] dt=pd.read_csv('/content/drive/MyDrive/ dataset/real estate valuation data set.csv')

[ ] dt.head()

   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             2012.917          32.0                                84.87882                               10     24.98298     121.54024                        37.9
1   2             2012.917          19.5                               306.59470                                9     24.98034     121.53951                        42.2
2   3             2013.583          13.3                               561.98450                                5     24.98746     121.54391                        47.3
3   4             2013.500          13.3                               561.98450                                5     24.98746     121.54391                        54.8
4   5             2012.833           5.0                               390.56840                                5     24.97937     121.54245                        43.1

dt

     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             2012.917          32.0                                84.87882                               10     24.98298     121.54024                        37.9
1     2             2012.917          19.5                               306.59470                                9     24.98034     121.53951                        42.2
2     3             2013.583          13.3                               561.98450                                5     24.98746     121.54391                        47.3
3     4             2013.500          13.3                               561.98450                                5     24.98746     121.54391                        54.8
4     5             2012.833           5.0                               390.56840                                5     24.97937     121.54245                        43.1
...
409  410            2013.000          13.7                              4082.01500                                0     24.94155     121.50381                        15.4
410  411            2012.667           5.6                                90.45606                                9     24.97433     121.54310                        50.0
411  412            2013.250          18.8                               390.96960                                7     24.97923     121.53986                        40.6
412  413            2013.000           8.1                               104.81010                                5     24.96674     121.54067                        52.5
413  414            2013.500           6.5                                90.45606                                9     24.97433     121.54310                        63.9
414 rows × 8 columns
```

Finding &Handling Missing 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
```

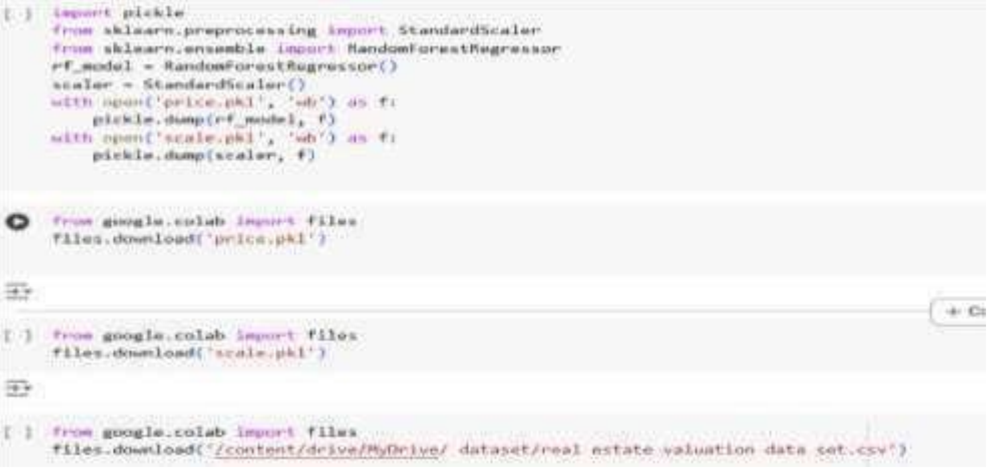
Data Transformation

-

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')
```