```
import pandas as pd
df = pd.read_csv('Real estate.csv')
```

```
print("first five")
print(df.head())
print("last three")
print(df.tail(3))
```

```
first five
   X1 transaction date  X2 house age  X3 distance to the nearest MRT station  \
0             2012.917          32.0                                84.87882
1             2012.917          19.5                               306.59470
2             2013.583          13.3                               561.98450
3             2013.500          13.3                               561.98450
4             2012.833           5.0                               390.56840

   X4 number of convenience stores  X5 latitude  X6 longitude  \
0                             10.0     24.98298     121.54024
1                              9.0     24.98034     121.53951
2                              5.0     24.98746     121.54391
3                              5.0     24.98746     121.54391
4                              5.0     24.97937     121.54245

   Y house price of unit area
0                        37.9
1                        42.2
2                        47.3
3                        54.8
4                        43.1
last three
     X1 transaction date  X2 house age  \
412             2013.000           8.1
413             2013.500           6.5
414             2013.167           1.9

     X3 distance to the nearest MRT station  X4 number of convenience stores  \
412                               104.81010                               5.0
413                                90.45606                               9.0
414                               355.00000                               NaN

     X5 latitude  X6 longitude  Y house price of unit area
412     24.96674     121.54067                        52.5
413     24.97433     121.54310                        63.9
414     24.97293     121.54026                        40.5
```

```
rows, columns = df.shape
print(f"it has {rows} rows and {columns} columns")
```

```
it has 415 rows and 7 columns
```

```
print("mean")
print(df.mean())
```

```
mean
X1 transaction date                       2013.149014
X2 house age                                17.674458
X3 distance to the nearest MRT station    1082.129338
X4 number of convenience stores              4.094203
X5 latitude                                 24.969039
X6 longitude                               121.533378
Y house price of unit area                  37.986265
dtype: float64
```

```
print("median")
print(df.median())
```

```
median
X1 transaction date                       2013.16700
X2 house age                                16.10000
X3 distance to the nearest MRT station     492.23130
X4 number of convenience stores              4.00000
X5 latitude                                 24.97110
X6 longitude                               121.53863
Y house price of unit area                  38.50000
dtype: float64
```

```
print("standard deviation")
print(df.std())
```

```
standard deviation
X1 transaction date                          0.281628
X2 house age                                11.405161
X3 distance to the nearest MRT station    1261.092057
X4 number of convenience stores              2.945562
X5 latitude                                  0.012397
X6 longitude                                 0.015332
Y house price of unit area                  13.590608
dtype: float64
```

```
print("five point summary")
print(df.describe())
```

```
five point summary
       X1 transaction date  X2 house age  \
count           415.000000    415.000000
mean           2013.149014     17.674458
std               0.281628     11.405161
min            2012.667000      0.000000
25%            2012.917000      8.950000
50%            2013.167000     16.100000
75%            2013.417000     28.100000
max            2013.583000     43.800000

       X3 distance to the nearest MRT station  \
count                              415.000000
mean                              1082.129338
std                               1261.092057
min                                 23.382840
25%                                289.324800
50%                                492.231300
75%                               1452.760000
max                               6488.021000

       X4 number of convenience stores  X5 latitude  X6 longitude  \
count                       414.000000   415.000000    415.000000
mean                          4.094203    24.969039    121.533378
std                           2.945562     0.012397      0.015332
min                           0.000000    24.932070    121.473530
25%                           1.000000    24.963010    121.528570
50%                           4.000000    24.971100    121.538630
75%                           6.000000    24.977450    121.543300
max                          10.000000    25.014590    121.566270

       Y house price of unit area
count                  415.000000
mean                    37.986265
std                     13.590608
min                      7.600000
25%                     27.700000
50%                     38.500000
75%                     46.600000
max                    117.500000
```

```
q1 = df.quantile(0.25)
q3 = df.quantile(0.75)
iqr = q3-q1
print(f"IQR: {iqr}")
```

```
IQR: X1 transaction date                        0.50000
X2 house age                                 19.15000
X3 distance to the nearest MRT station    1163.43520
X4 number of convenience stores              5.00000
X5 latitude                                  0.01444
X6 longitude                                 0.01473
Y house price of unit area                  18.90000
dtype: float64
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 415 entries, 0 to 414
Data columns (total 7 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   X1 transaction date                     415 non-null    float64
 1   X2 house age                            415 non-null    float64
 2   X3 distance to the nearest MRT station  415 non-null    float64
```

```
 3   X4 number of convenience stores         414 non-null    float64
 4   X5 latitude                             415 non-null    float64
 5   X6 longitude                            415 non-null    float64
 6   Y house price of unit area              415 non-null    float64
dtypes: float64(7)
memory usage: 22.8 KB
```

```
df['X22'] = df['X2 house age'] * 365
df.head()
```

	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	X22
0	2012.917	32.0	84.87882	10.0	24.98298	121.54024	37.9	11680.0
1	2012.917	19.5	306.59470	9.0	24.98034	121.53951	42.2	7117.5
2	2013.583	13.3	561.98450	5.0	24.98746	121.54391	47.3	4854.5
3	2013.500	13.3	561.98450	5.0	24.98746	121.54391	54.8	4854.5

```
print(df.isnull().sum())
```

```
X1 transaction date                       0
X2 house age                              0
X3 distance to the nearest MRT station    0
X4 number of convenience stores           1
X5 latitude                               0
X6 longitude                              0
Y house price of unit area                0
X22                                       0
dtype: int64
```

```
df = df.drop(columns=['X22'])
```

```
df.head()
```

	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	2012.917	32.0	84.87882	10.0	24.98298	121.54024	37.9
1	2012.917	19.5	306.59470	9.0	24.98034	121.53951	42.2
2	2013.583	13.3	561.98450	5.0	24.98746	121.54391	47.3
3	2013.500	13.3	561.98450	5.0	24.98746	121.54391	54.8

```
new_data = [
    {'X1 transaction date': '2024-10-01', 'X2 house age': 5, 'X3 distance to the nearest MRT station': 300,
     'X4 number of convenience stores': 2, 'X5 latitude': 25.0478, 'X6 longitude': 121.5319, 'Y house price of unit area': 30},

    {'X1 transaction date': '2024-10-02', 'X2 house age': 10, 'X3 distance to the nearest MRT station': 150,
     'X4 number of convenience stores': 5, 'X5 latitude': 25.0330, 'X6 longitude': 121.5645, 'Y house price of unit area': 40},

    {'X1 transaction date': '2024-10-03', 'X2 house age': 1, 'X3 distance to the nearest MRT station': 600,
     'X4 number of convenience stores': 1, 'X5 latitude': 25.0420, 'X6 longitude': 121.5022, 'Y house price of unit area': 25}
]
```

```
new_df = pd.DataFrame(new_data)
df = pd.concat([df,new_df],ignore_index=True)
print(df)
```

```
     X1 transaction date  X2 house age  X3 distance to the nearest MRT station  \
0              2012.917          32.0                                84.87882
1              2012.917          19.5                               306.59470
2              2013.583          13.3                               561.98450
3                2013.5          13.3                               561.98450
4              2012.833           5.0                               390.56840
..                  ...           ...                                     ...
413              2013.5           6.5                                90.45606
414            2013.167           1.9                               355.00000
415          2024-10-01           5.0                               300.00000
416          2024-10-02          10.0                               150.00000
417          2024-10-03           1.0                               600.00000
```

```
     X4 number of convenience stores  X5 latitude  X6 longitude  \
0                               10.0     24.98298     121.54024
1                                9.0     24.98034     121.53951
2                                5.0     24.98746     121.54391
3                                5.0     24.98746     121.54391
4                                5.0     24.97937     121.54245
..                               ...          ...          ...
413                              9.0     24.97433     121.54310
414                              NaN     24.97293     121.54026
415                              2.0     25.04780     121.53190
416                              5.0     25.03300     121.56450
417                              1.0     25.04200     121.50220

     Y house price of unit area
0                          37.9
1                          42.2
2                          47.3
3                          54.8
4                          43.1
..                          ...
413                        63.9
414                        40.5
415                        30.0
416                        40.0
417                        25.0

[418 rows x 7 columns]
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 7 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   X1 transaction date                     418 non-null    object
 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         417 non-null    float64
 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), object(1)
memory usage: 23.0+ KB
```

```
df = df.iloc[:-3]
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 415 entries, 0 to 414
Data columns (total 7 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   X1 transaction date                     415 non-null    object
 1   X2 house age                            415 non-null    float64
 2   X3 distance to the nearest MRT station  415 non-null    float64
 3   X4 number of convenience stores         414 non-null    float64
 4   X5 latitude                             415 non-null    float64
 5   X6 longitude                            415 non-null    float64
 6   Y house price of unit area              415 non-null    float64
dtypes: float64(6), object(1)
memory usage: 22.8+ KB
```

```
print(df)
```

```
     X1 transaction date  X2 house age  X3 distance to the nearest MRT station  \
0              2012.917          32.0                                84.87882
1              2012.917          19.5                               306.59470
2              2013.583          13.3                               561.98450
3                2013.5          13.3                               561.98450
4              2012.833           5.0                               390.56840
..                  ...           ...                                     ...
410            2012.667           5.6                                90.45606
411             2013.25          18.8                               390.96960
412              2013.0           8.1                               104.81010
413              2013.5           6.5                                90.45606
414            2013.167           1.9                               355.00000

     X4 number of convenience stores  X5 latitude  X6 longitude  \
0                               10.0     24.98298     121.54024
1                                9.0     24.98034     121.53951
```

```
2                               5.0     24.98746     121.54391
3                               5.0     24.98746     121.54391
4                               5.0     24.97937     121.54245
..                              ...          ...           ...
410                             9.0     24.97433     121.54310
411                             7.0     24.97923     121.53986
412                             5.0     24.96674     121.54067
413                             9.0     24.97433     121.54310
414                             NaN     24.97293     121.54026

     Y house price of unit area
0                          37.9
1                          42.2
2                          47.3
3                          54.8
4                          43.1
..                          ...
410                        50.0
411                        40.6
412                        52.5
413                        63.9
414                        40.5

[415 rows x 7 columns]
```

```
df.loc[df['Y house price of unit area'] > 110 , 'Y house price of unit area']= 110
print(df)
```

```
     X1 transaction date  X2 house age  X3 distance to the nearest MRT station  \
0              2012.917          32.0                                84.87882
1              2012.917          19.5                               306.59470
2              2013.583          13.3                               561.98450
3                2013.5          13.3                               561.98450
4              2012.833           5.0                               390.56840
..                  ...           ...                                     ...
410            2012.667           5.6                                90.45606
411             2013.25          18.8                               390.96960
412              2013.0           8.1                               104.81010
413              2013.5           6.5                                90.45606
414            2013.167           1.9                               355.00000

     X4 number of convenience stores  X5 latitude  X6 longitude  \
0                               10.0     24.98298     121.54024
1                                9.0     24.98034     121.53951
2                                5.0     24.98746     121.54391
3                                5.0     24.98746     121.54391
4                                5.0     24.97937     121.54245
..                               ...          ...           ...
410                              9.0     24.97433     121.54310
411                              7.0     24.97923     121.53986
412                              5.0     24.96674     121.54067
413                              9.0     24.97433     121.54310
414                              NaN     24.97293     121.54026

     Y house price of unit area
0                          37.9
1                          42.2
2                          47.3
3                          54.8
4                          43.1
..                          ...
410                        50.0
411                        40.6
412                        52.5
413                        63.9
414                        40.5

[415 rows x 7 columns]
```

```
df.loc[df['Y house price of unit area'] > 60 , 'Y house price of unit area']= 110
print(df)
```

```
     X1 transaction date  X2 house age  X3 distance to the nearest MRT station  \
0              2012.917          32.0                                84.87882
1              2012.917          19.5                               306.59470
2              2013.583          13.3                               561.98450
3                2013.5          13.3                               561.98450
4              2012.833           5.0                               390.56840
..                  ...           ...                                     ...
410            2012.667           5.6                                90.45606
411             2013.25          18.8                               390.96960
412              2013.0           8.1                               104.81010
```

```
413              2013.5           6.5                                90.45606
414            2013.167           1.9                               355.00000

     X4 number of convenience stores  X5 latitude  X6 longitude  \
0                               10.0     24.98298     121.54024
1                                9.0     24.98034     121.53951
2                                5.0     24.98746     121.54391
3                                5.0     24.98746     121.54391
4                                5.0     24.97937     121.54245
..                               ...          ...          ...
410                              9.0     24.97433     121.54310
411                              7.0     24.97923     121.53986
412                              5.0     24.96674     121.54067
413                              9.0     24.97433     121.54310
414                              NaN     24.97293     121.54026

     Y house price of unit area
0                          37.9
1                          42.2
2                          47.3
3                          54.8
4                          43.1
..                          ...
410                        50.0
411                        40.6
412                        52.5
413                       110.0
414                        40.5

[415 rows x 7 columns]
```

```
result = df.loc[df['Y house price of unit area'] <=20,['X5 latitude','X6 longitude']]
print(result)
```

```
     X5 latitude  X6 longitude
8       24.95095     121.48458
40      24.94155     121.50381
41      24.94297     121.50342
48      24.94684     121.49578
49      24.94925     121.49542
55      24.94968     121.53009
73      24.94155     121.50381
83      24.96056     121.50831
87      24.94297     121.50342
93      24.94920     121.53076
113     24.96172     121.53812
116     24.94375     121.47883
117     24.93885     121.50383
155     24.94155     121.50381
156     24.94883     121.52954
162     24.94297     121.50342
170     24.94741     121.49628
176     24.94867     121.49507
180     24.94898     121.49621
183     24.94155     121.50381
226     24.94155     121.50381
229     24.94890     121.53095
231     24.94235     121.50357
232     24.95032     121.49587
249     24.95743     121.47516
251     24.94960     121.53018
255     24.95095     121.48458
298     24.94155     121.50381
309     24.94883     121.52954
320     24.93885     121.50383
329     24.93885     121.50383
330     24.94935     121.53046
331     24.94826     121.49587
347     24.95719     121.47353
384     24.94297     121.50342
409     24.94155     121.50381
```

```
print(df.isnull().sum())
```

```
X1 transaction date                       0
X2 house age                              0
X3 distance to the nearest MRT station    0
X4 number of convenience stores           1
X5 latitude                               0
X6 longitude                              0
Y house price of unit area                0
dtype: int64
```

```
average = df['X4 number of convenience stores'].mean()
df['X4 number of convenience stores'].fillna(average,inplace=True)
print(df)
```

```
     X1 transaction date  X2 house age  X3 distance to the nearest MRT station  \
0              2012.917          32.0                                84.87882
1              2012.917          19.5                               306.59470
2              2013.583          13.3                               561.98450
3                2013.5          13.3                               561.98450
4              2012.833           5.0                               390.56840
..                  ...           ...                                     ...
410            2012.667           5.6                                90.45606
411             2013.25          18.8                               390.96960
412              2013.0           8.1                               104.81010
413              2013.5           6.5                                90.45606
414            2013.167           1.9                               355.00000

     X4 number of convenience stores  X5 latitude  X6 longitude  \
0                         10.000000     24.98298     121.54024
1                          9.000000     24.98034     121.53951
2                          5.000000     24.98746     121.54391
3                          5.000000     24.98746     121.54391
4                          5.000000     24.97937     121.54245
..                              ...          ...          ...
410                        9.000000     24.97433     121.54310
411                        7.000000     24.97923     121.53986
412                        5.000000     24.96674     121.54067
413                        9.000000     24.97433     121.54310
414                        4.094203     24.97293     121.54026

     Y house price of unit area
0                          37.9
1                          42.2
2                          47.3
3                          54.8
4                          43.1
..                          ...
410                        50.0
411                        40.6
412                        52.5
413                       110.0
414                        40.5

[415 rows x 7 columns]
<ipython-input-32-d1dc276343f2>:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignm
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting valu

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].me
```

```
df['X4 number of convenience stores'].fillna(average,inplace=True)
```

```
df.isnull()
```

	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	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...
410	False	False	False	False	False	False	False
411	False	False	False	False	False	False	False
412	False	False	False	False	False	False	False
413	False	False	False	False	False	False	False
414	False	False	False	False	False	False	False

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 415 entries, 0 to 414
Data columns (total 7 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   X1 transaction date                     415 non-null    object
 1   X2 house age                            415 non-null    float64
 2   X3 distance to the nearest MRT station  415 non-null    float64
 3   X4 number of convenience stores         415 non-null    float64
 4   X5 latitude                             415 non-null    float64
 5   X6 longitude                            415 non-null    float64
 6   Y house price of unit area              415 non-null    float64
dtypes: float64(6), object(1)
memory usage: 22.8+ KB
```

```
df.isnull().sum()
```

	0
X1 transaction date	0
X2 house age	0
X3 distance to the nearest MRT station	0
X4 number of convenience stores	0
X5 latitude	0
X6 longitude	0
Y house price of unit area	0

```
mean_distance = df['X3 distance to the nearest MRT station'].mean()
std_distance = df['X3 distance to the nearest MRT station'].std()
```

```
df['Z_score_normalize'] = (df['X3 distance to the nearest MRT station']-mean_distance)/std_distance
print(df[['Z_score_normalize','X3 distance to the nearest MRT station']])
```

```
     Z_score_normalize  X3 distance to the nearest MRT station
0            -0.790783                                84.87882
1            -0.614971                               306.59470
2            -0.412456                               561.98450
3            -0.412456                               561.98450
4            -0.548383                               390.56840
..                 ...                                     ...
410          -0.786361                                90.45606
411          -0.548064                               390.96960
412          -0.774979                               104.81010
413          -0.786361                                90.45606
414          -0.576587                               355.00000

[415 rows x 2 columns]
```

```
min_dis = df['X3 distance to the nearest MRT station'].min()
max_dis = df['X3 distance to the nearest MRT station'].max()
df['min_max_nomalize'] = (df['X3 distance to the nearest MRT station']-min_dis)/(max_dis-min_dis)
print(df[['X3 distance to the nearest MRT station', 'min_max_nomalize']])
```

```
     X3 distance to the nearest MRT station  min_max_nomalize
0                                  84.87882          0.009513
1                                 306.59470          0.043809
2                                 561.98450          0.083315
3                                 561.98450          0.083315
4                                 390.56840          0.056799
..                                      ...               ...
410                                90.45606          0.010375
411                               390.96960          0.056861
412                               104.81010          0.012596
413                                90.45606          0.010375
414                               355.00000          0.051297

[415 rows x 2 columns]
```

```
max_abs_dis = df['X3 distance to the nearest MRT station'].abs().max()
j = len(str(int(max_abs_dis)))
```

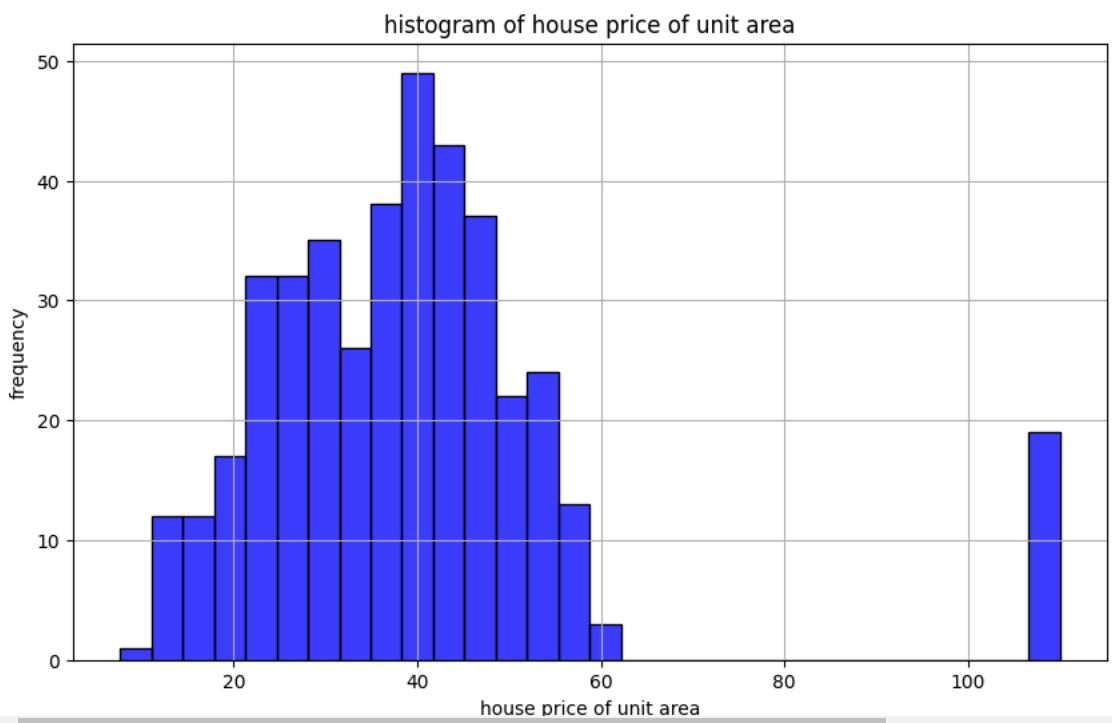
```
df['decimal_scaled']= (df['X3 distance to the nearest MRT station']/(10**j))
print(df[['X3 distance to the nearest MRT station','decimal_scaled']])
```

```
     X3 distance to the nearest MRT station  decimal_scaled
0                                  84.87882        0.008488
1                                 306.59470        0.030659
2                                 561.98450        0.056198
3                                 561.98450        0.056198
4                                 390.56840        0.039057
..                                      ...             ...
410                                90.45606        0.009046
411                               390.96960        0.039097
412                               104.81010        0.010481
413                                90.45606        0.009046
414                               355.00000        0.035500

[415 rows x 2 columns]
```

```
import seaborn as sns
import matplotlib.pyplot as plt
```

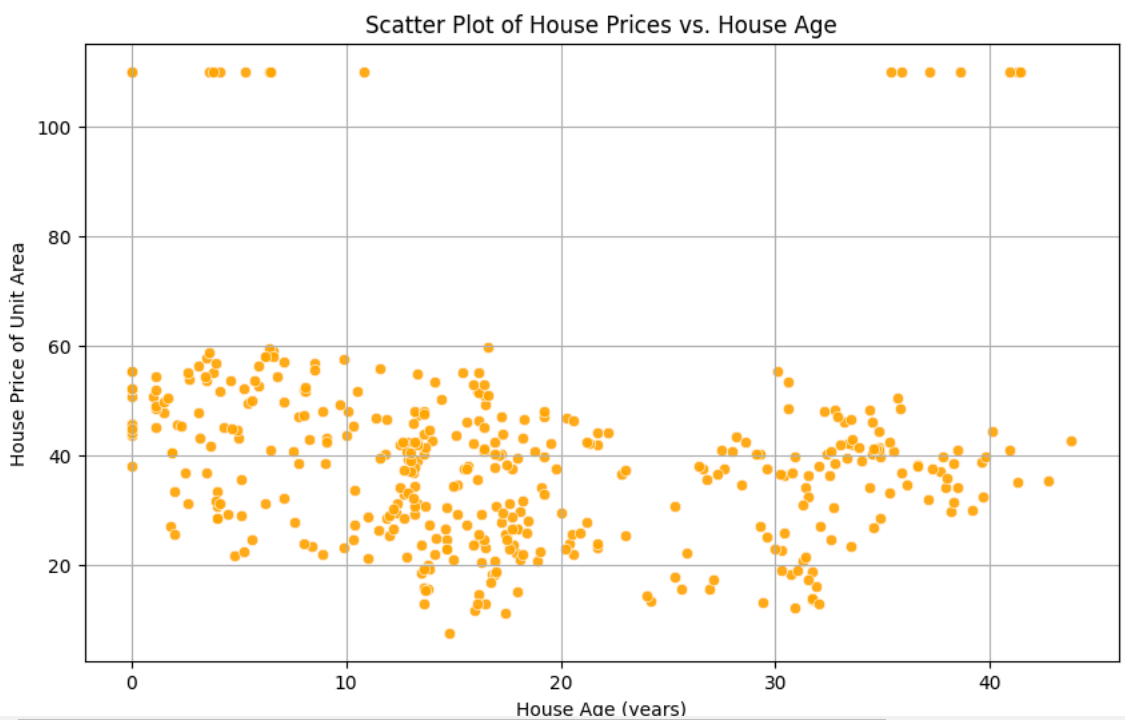
```
plt.figure(figsize=(10,6))
sns.histplot(df['Y house price of unit area'],bins=30,color='blue')
plt.title("histogram of house price of unit area")
plt.xlabel("house price of unit area")
plt.ylabel("frequency")
plt.grid()
plt.show()
```



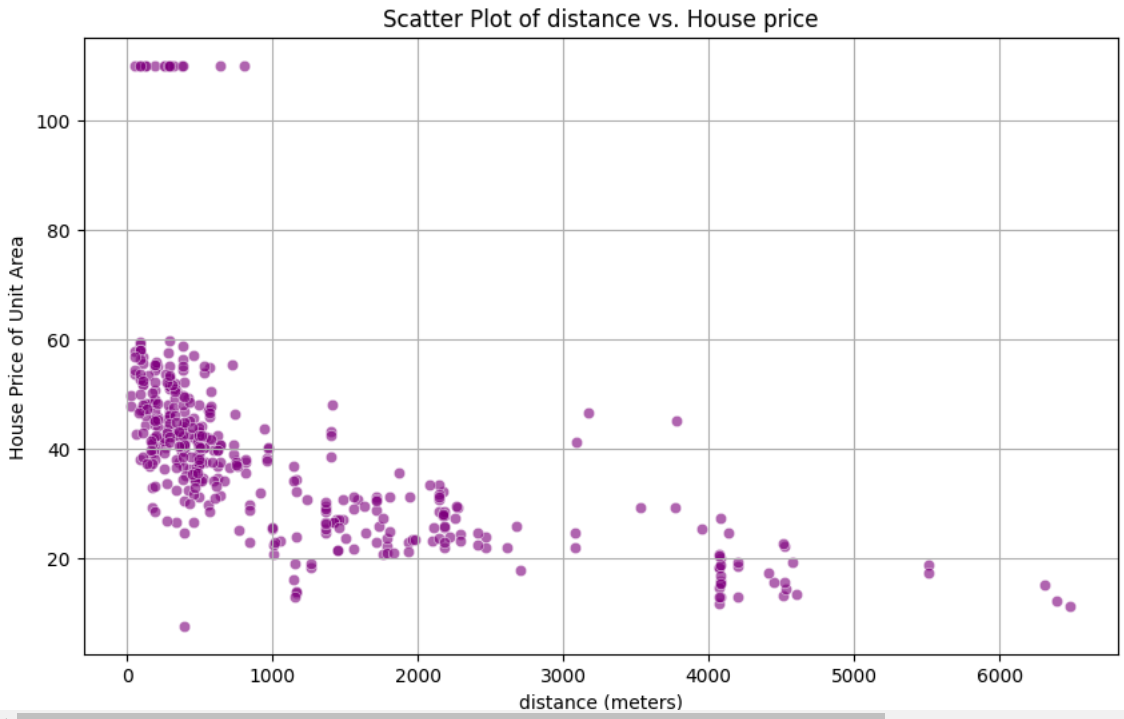

```
plt.figure(figsize=(10,6))
sns.boxplot(df['Y house price of unit area'],color='green')
plt.title("histogram of house price of unit area")
plt.ylabel("house price of unit area")
plt.grid()
plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/categorical.py:640: FutureWarning: SeriesGroupBy.grouper is deprecated and will be remov
  positions = grouped.grouper.result_index.to_numpy(dtype=float)
```


```
plt.figure(figsize=(10, 6))
sns.scatterplot(x=df['X2 house age'], y=df['Y house price of unit area'], color='orange', alpha=0.9)
plt.title('Scatter Plot of House Prices vs. House Age')
plt.xlabel('House Age (years)')
plt.ylabel('House Price of Unit Area')
plt.grid()
plt.show()
```




```
plt.figure(figsize=(10, 6))
sns.scatterplot(x=df['X3 distance to the nearest MRT station'], y=df['Y house price of unit area'], color='purple', alpha=0.6)
plt.title('Scatter Plot of distance vs. House price')
plt.xlabel('distance (meters)')
plt.ylabel('House Price of Unit Area')
plt.grid()
plt.show()
```


Scatter Plot of distance vs. House price
House Price of Unit Area
distance (meters)
100
80
60
40
20
0
1000
2000
3000
4000
5000
6000