## TASK - 1 | Downloaded and loaded Zameen.com property data from https://www.kaggle.com/datasets/huzzefakhan/zameencom-property-data-pakistan

In [ ]:
```
import pandas as pd
data = pd.read_csv('Property_with_Feature_Engineering.csv')
```

## TASK-2 | Describe the data properties of each column,

1. Datatype of each column
2. Missing values in each column
3. Null values in each column
4. Outliers in each column

In [ ]:
```
# 1.
data.dtypes
```

Out[ ]:
```
property_id        int64
location_id        int64
page_url          object
property_type     object
price              int64
price_bin         object
location          object
city              object
province_name     object
locality          object
latitude         float64
longitude        float64
baths              int64
area              object
area_marla       float64
area_sqft        float64
purpose           object
bedrooms           int64
date_added        object
year               int64
month              int64
day                int64
agency            object
agent             object
dtype: object
```

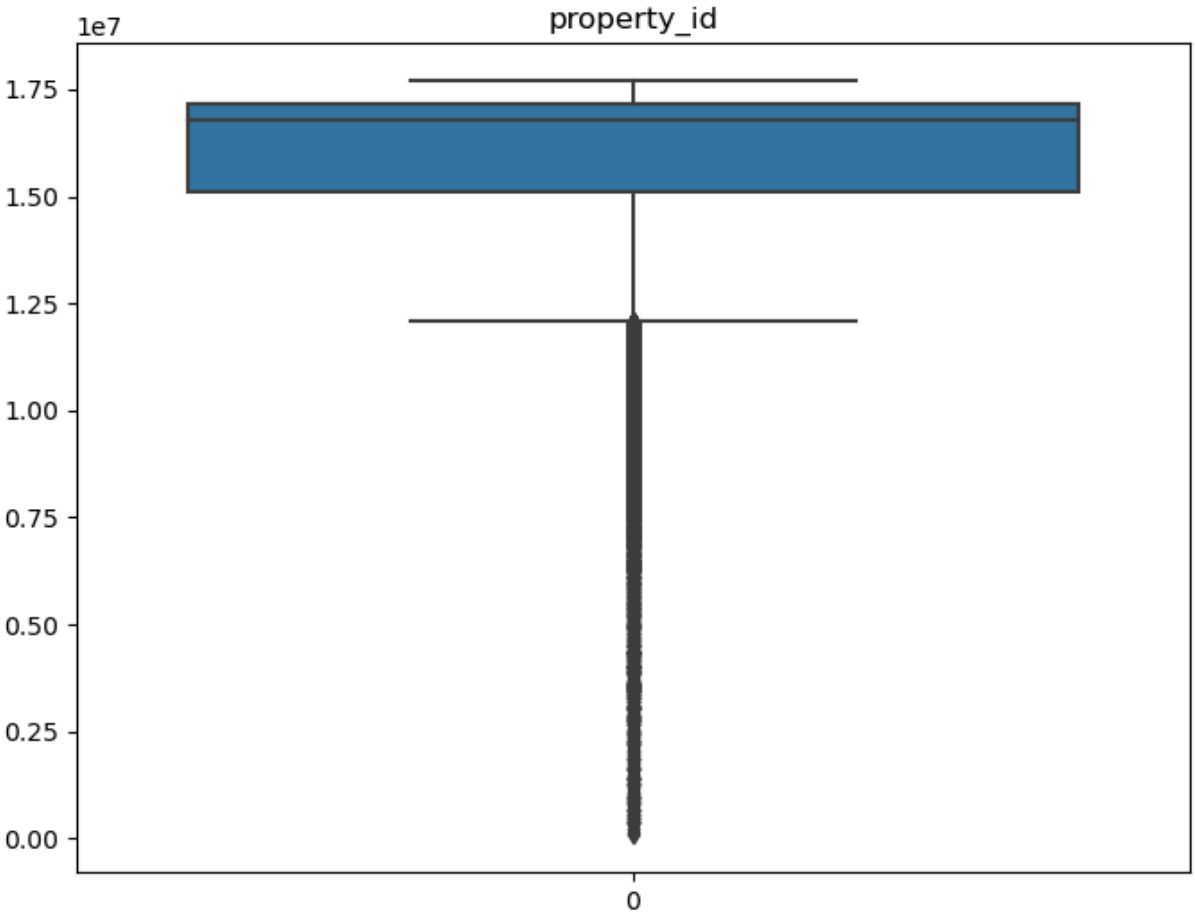
In [ ]:
```
# 2.
data.isnull().sum()
```

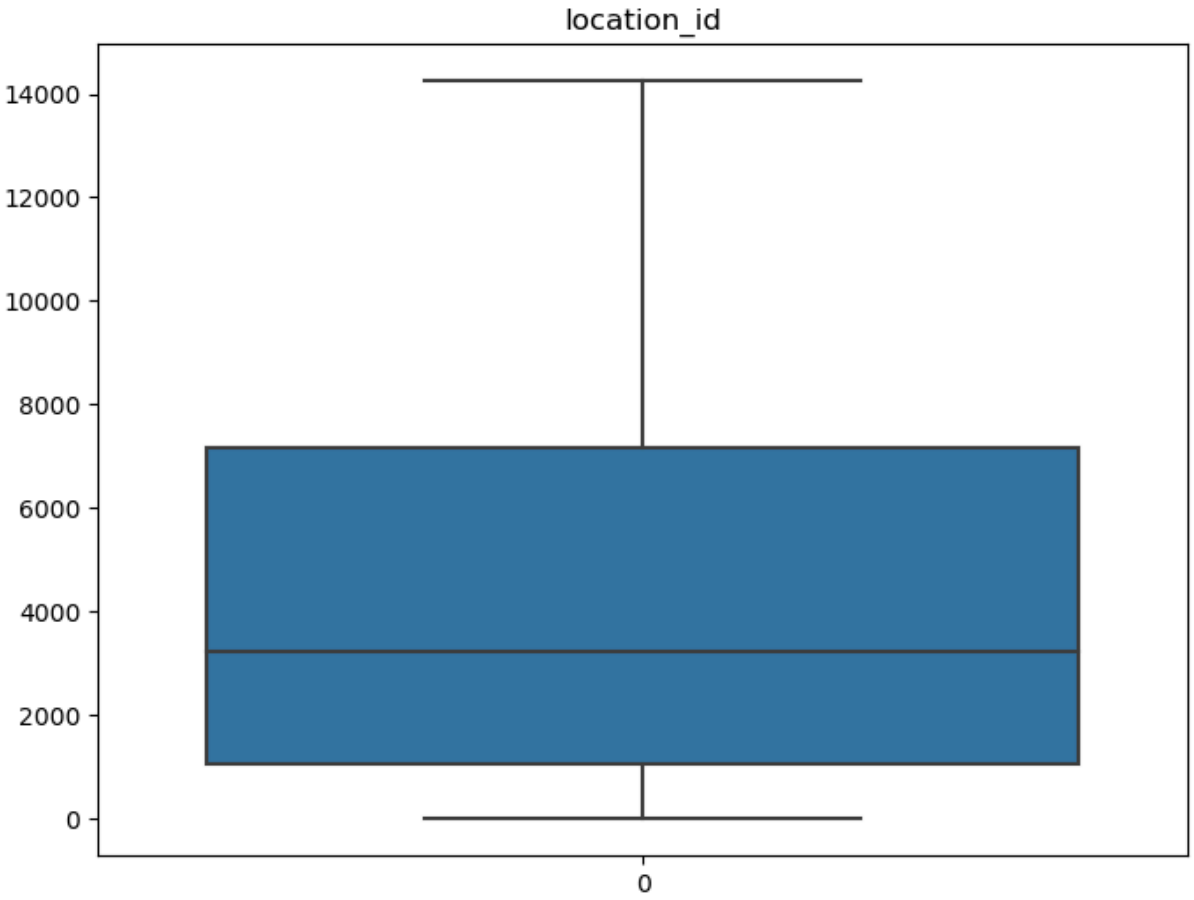
Out[ ]:
```
property_id          0
location_id          0
page_url             0
property_type        0
price                0
price_bin            0
location             0
city                 0
province_name        0
locality             0
latitude             0
longitude            0
baths                0
area                 0
area_marla           0
area_sqft            0
purpose              0
bedrooms             0
date_added           0
year                 0
month                0
day                  0
agency           47379
agent            47380
dtype: int64
```

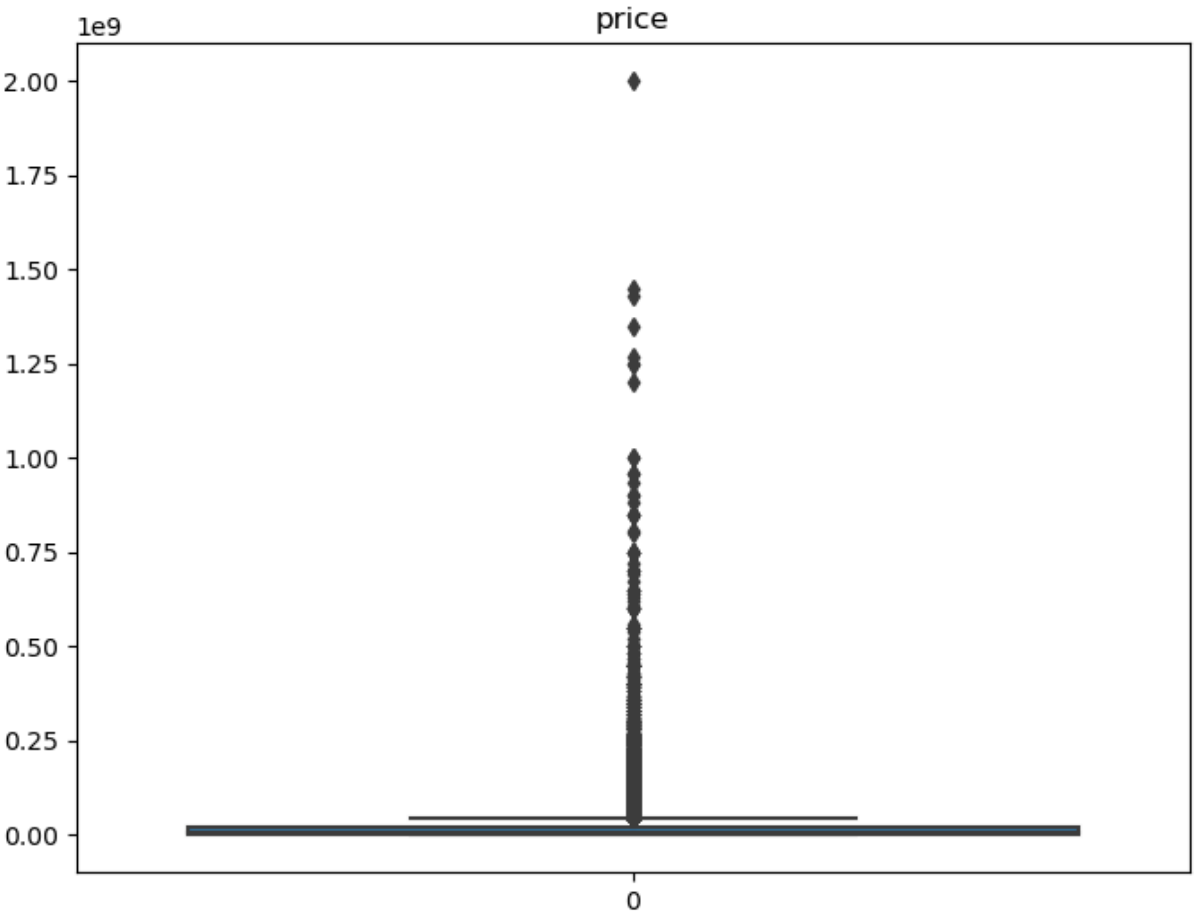
In [ ]:
```
#3.
data.isna().sum()
```

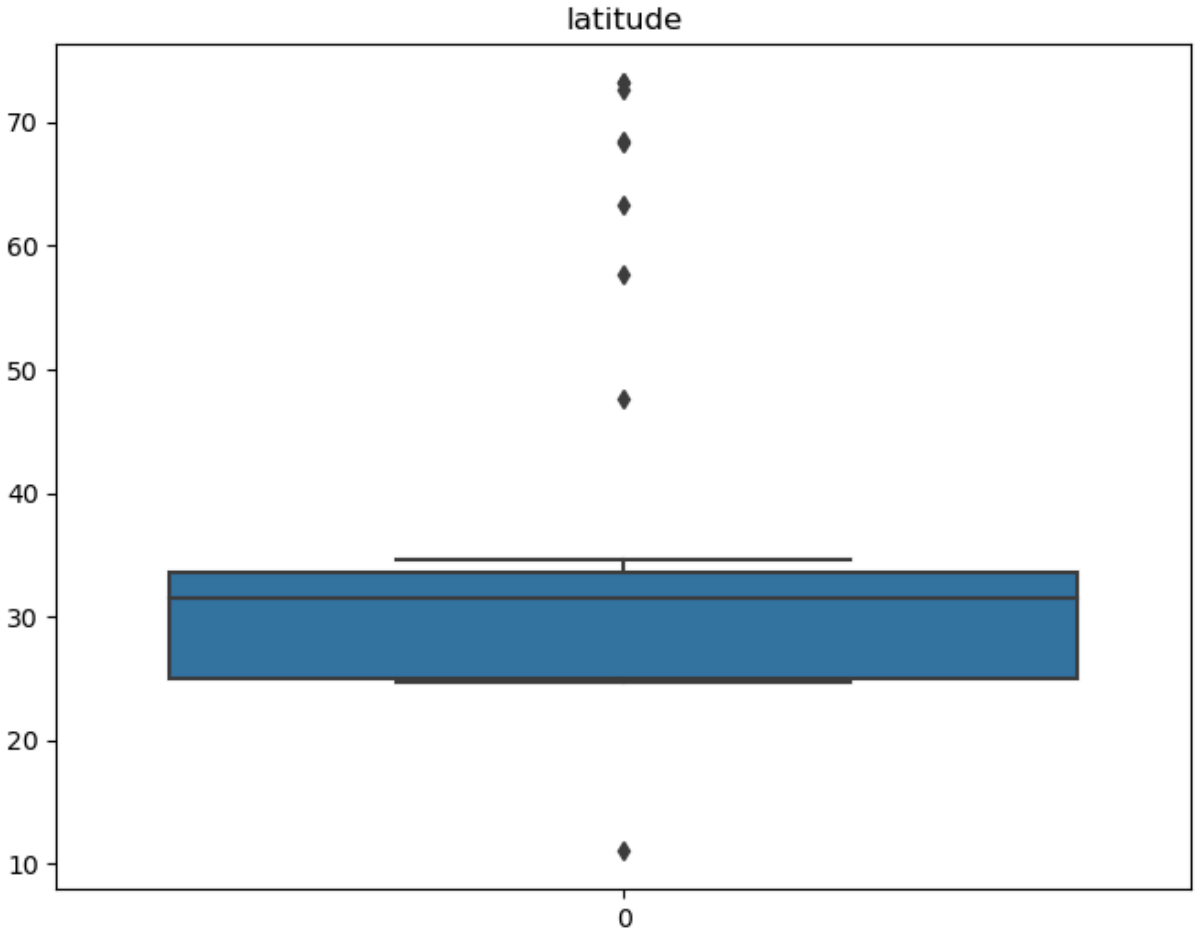
Out[ ]:
```
property_id          0
location_id          0
page_url             0
property_type        0
price                0
price_bin            0
location             0
city                 0
province_name        0
locality             0
latitude             0
longitude            0
baths                0
area                 0
area_marla           0
area_sqft            0
purpose              0
bedrooms             0
date_added           0
year                 0
month                0
day                  0
agency           47379
agent            47380
dtype: int64
```

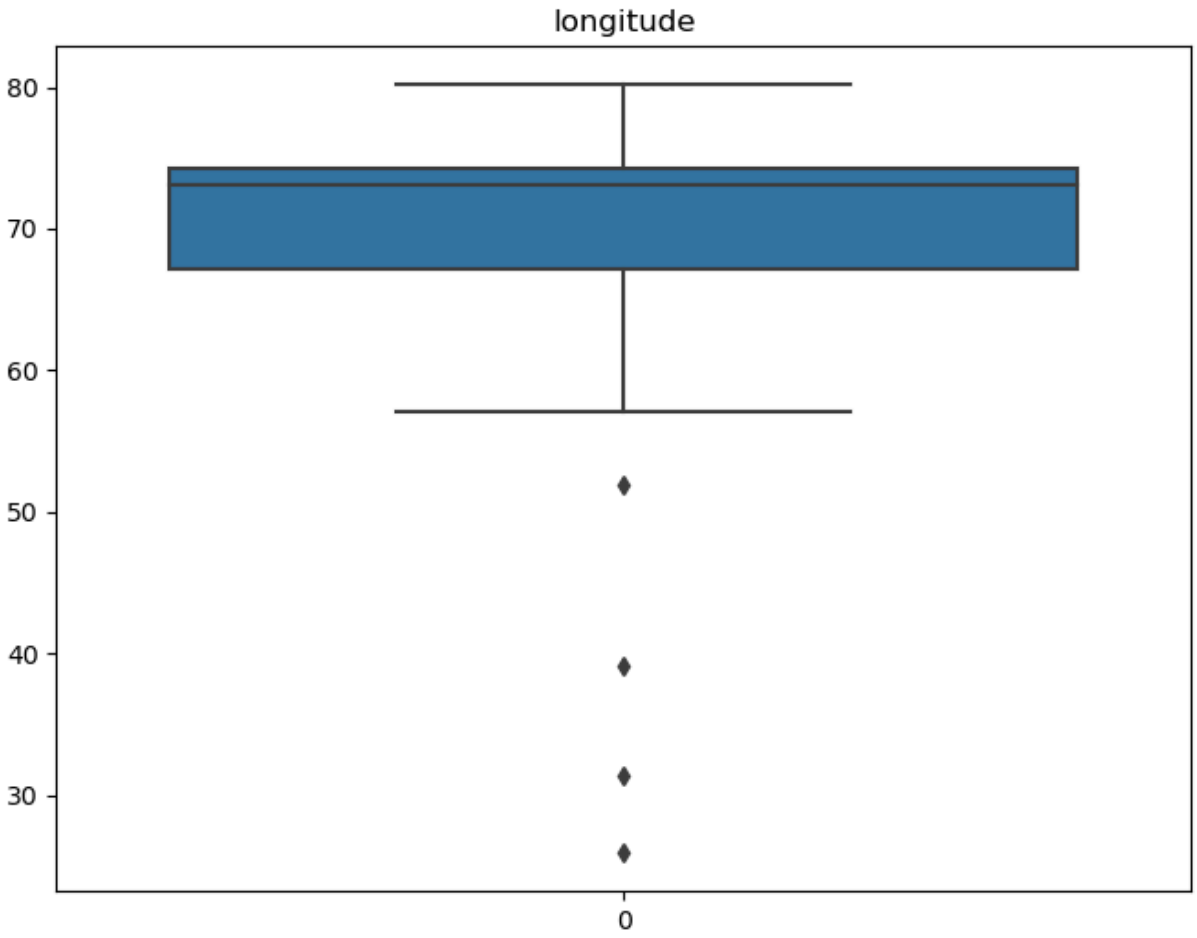
In [ ]:
```
# 4. Box plotting for numerical columns to visualize outliers.
import seaborn as sns
import matplotlib.pyplot as plt

columns = data.select_dtypes(include=['int64', 'float64']).columns

for col in columns:
    plt.figure(figsize=(8,6))
    sns.boxplot(data[col])
    plt.title(col)
    plt.show()
```

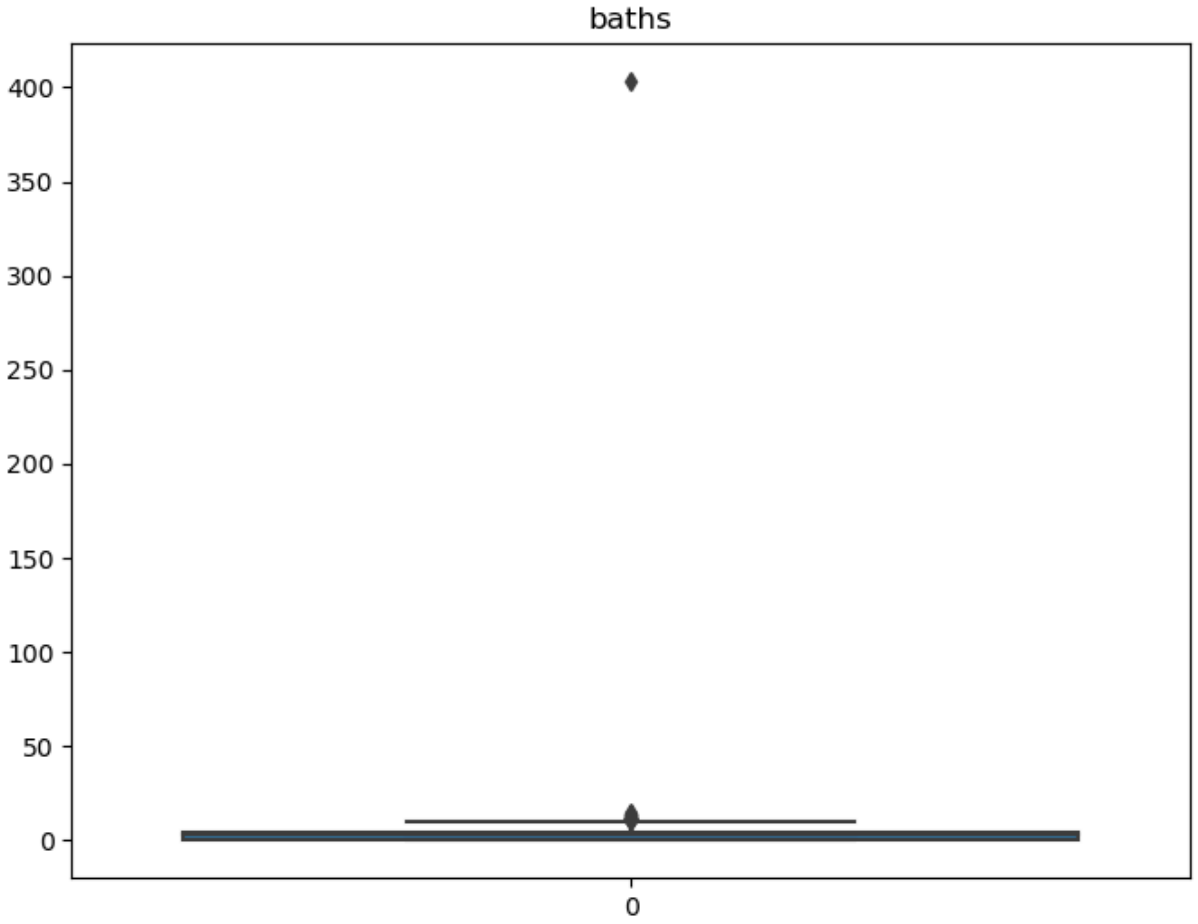


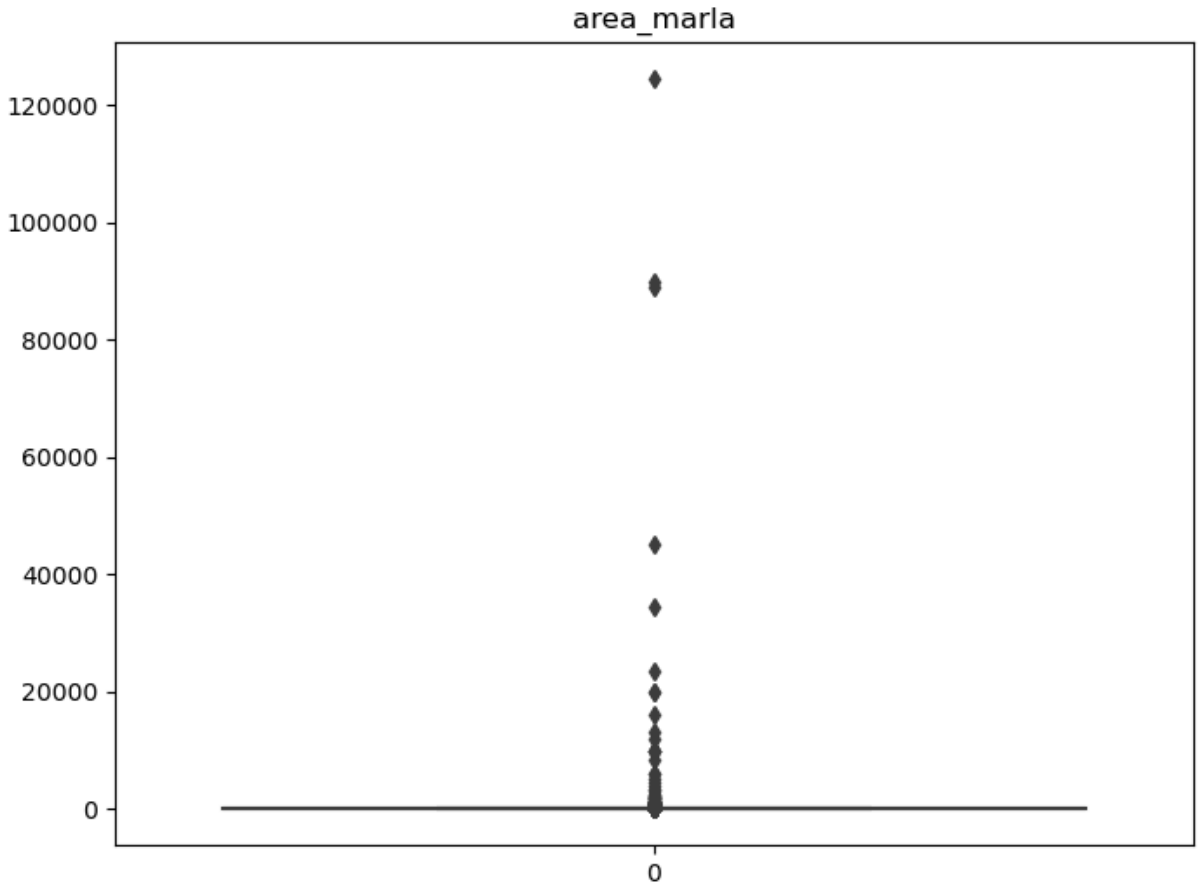

location_id
14000
12000
10000
8000
6000
4000
2000
0
0

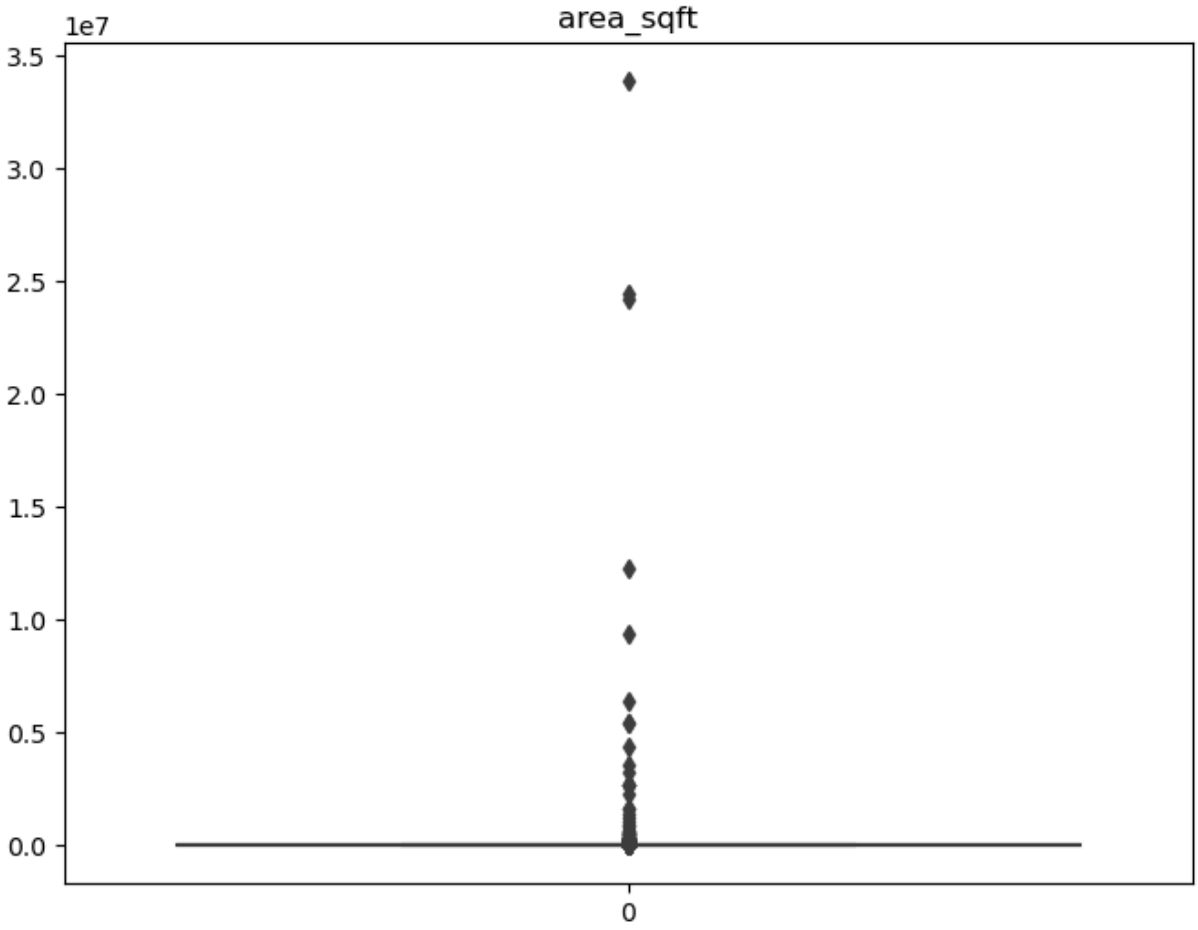

price
1e9
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0

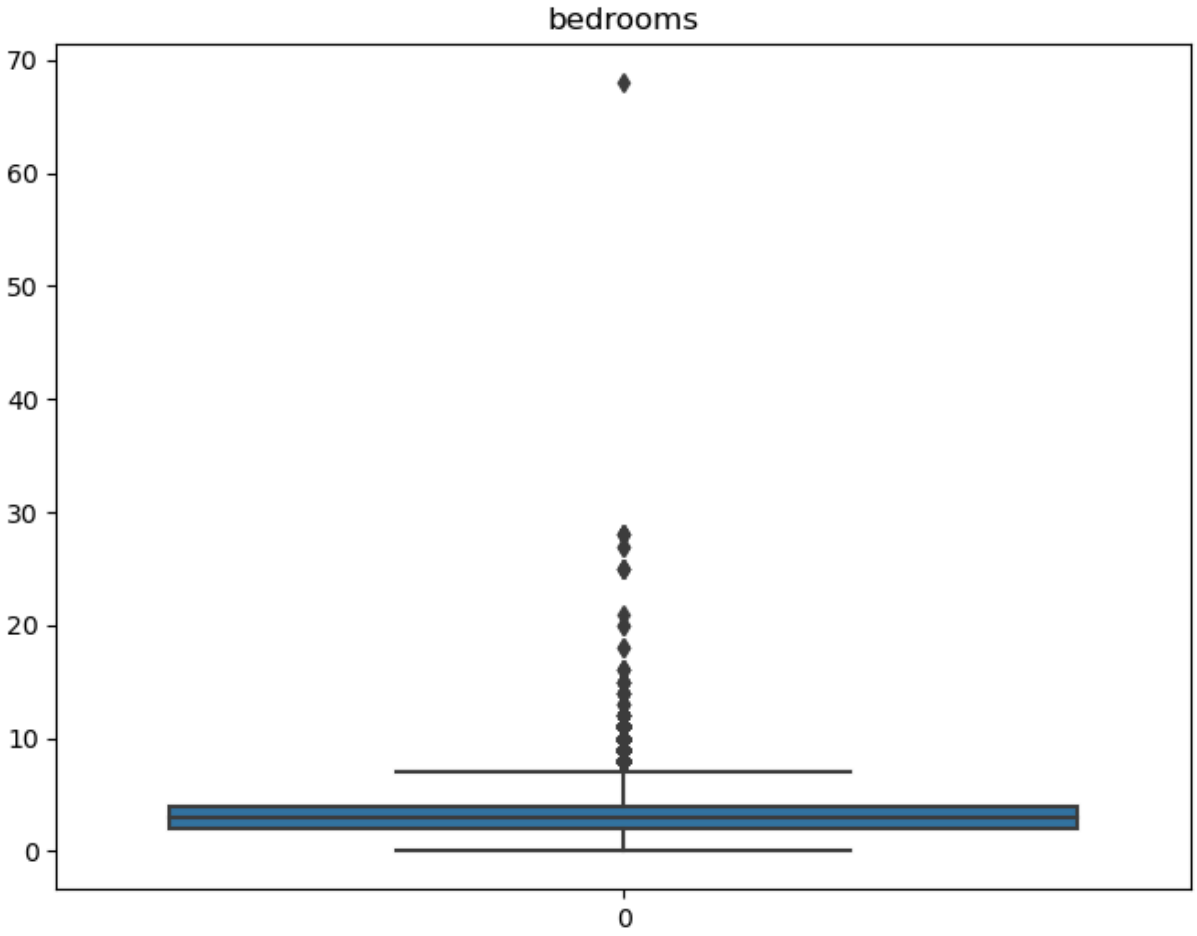

latitude
70
60
50
40
30
20
10
0

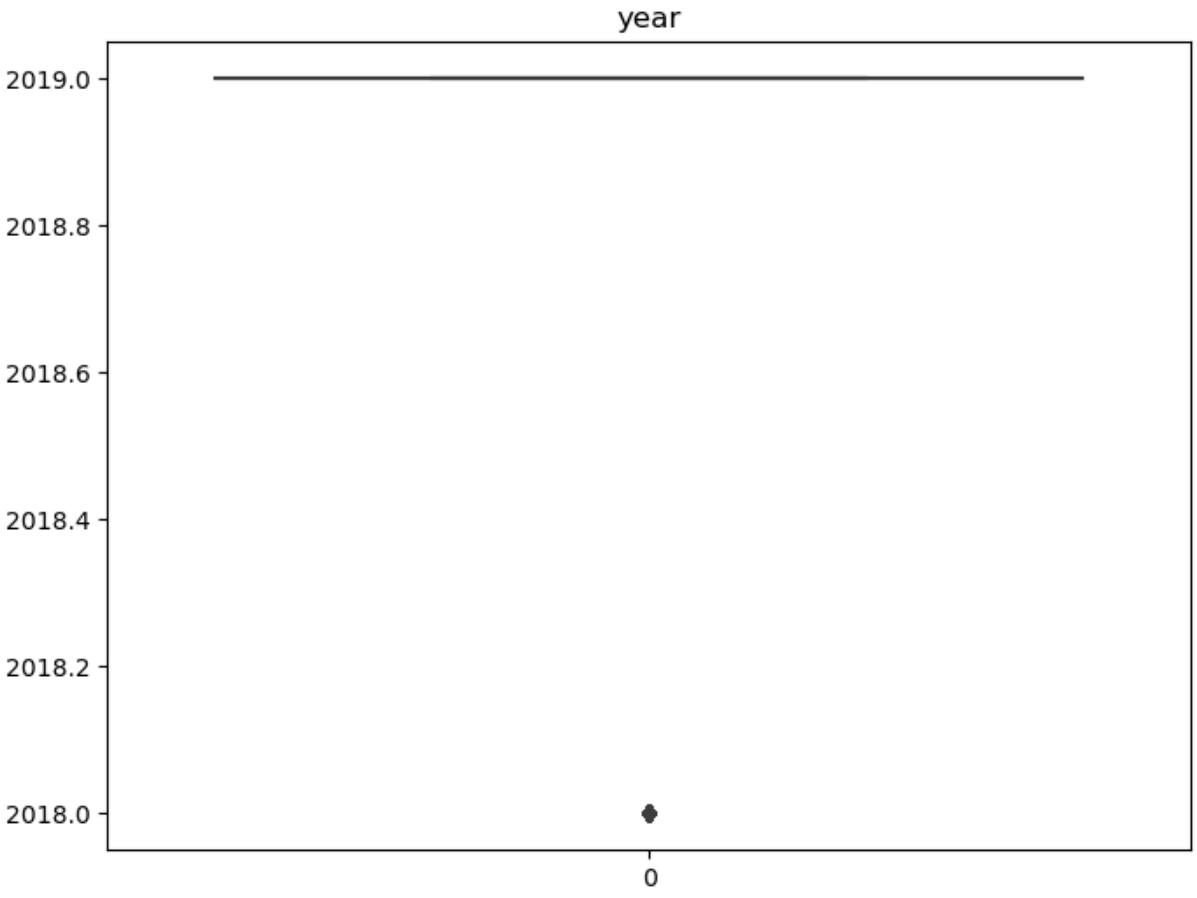

longitude
80
70
60
50
40
30
0

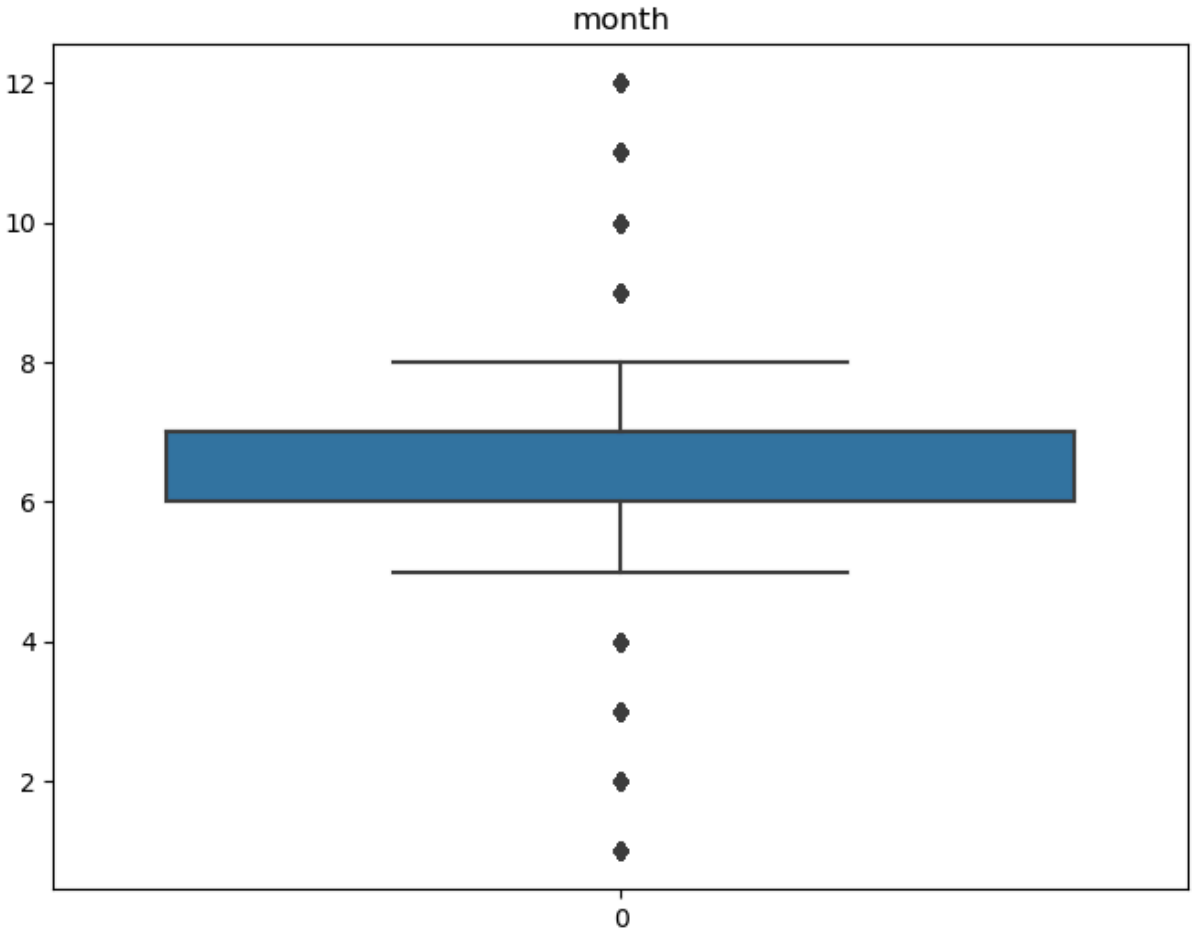

baths
400
350
300
250
200
150
100
50
0
0

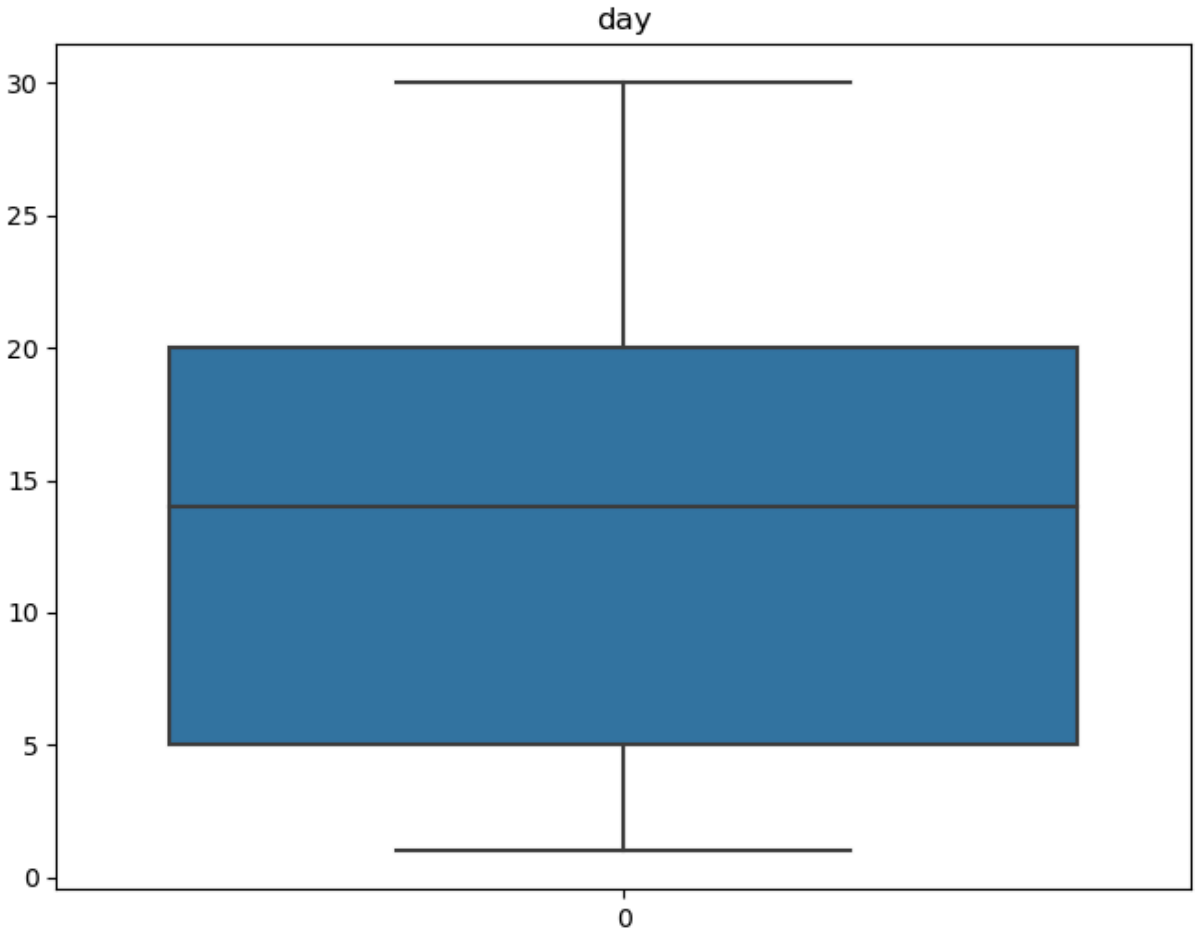

area_marla
120000
100000
80000
60000
40000
20000
0
0


area_sqft
1e7
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0


bedrooms
70
60
50
40
30
20
10
0
0


year
2019.0
2018.8
2018.6
2018.4
2018.2
2018.0
0


month
12
10
8
6
4
2
0




## TASK - 3 | Handle null values by replacing them with suitable values.

In [ ]:
```
import pandas as pd
data.fillna(0, inplace=True)
```

## TASK - 4 | Suppose you have to predict the cost of a house. For this purpose, select the appropriate columns that will help you develop a machine learning model. Save the selected columns dataset in a separate CSV file.

In [ ]:
```
import pandas as pd
selected_columns = ['property_id','location_id','page_url','property_type','price',
selected_data = data[selected_columns]
selected_data.to_csv('results_of_selected_house.csv', index=False)
```

## TASK - 5 | List down the descriptive variables and target variable.

In [ ]:
```
X = selected_data.drop('price', axis=1)
Y = selected_data['locality']
```

## TASK - 6 | Describe the statistics of the new data.

In [ ]:
```
import pandas as pd

selected_data = pd.read_csv('results_of_selected_house.csv')
numerical_stats = selected_data.describe()
categorical_stats = selected_data.describe(include=['object'])

print("Summary statistics for numerical variables:")
print(numerical_stats)

print("\nFrequency counts for categorical variables:")
print(categorical_stats)
```

```
Summary statistics for numerical variables:
        property_id    location_id         price
count  1.913930e+05  191393.000000  1.913930e+05
mean   1.573170e+07    4224.580350  1.644655e+07
std    2.215249e+06    3719.125201  3.416412e+07
min    8.657500e+04       1.000000  0.000000e+00
25%    1.511867e+07    1057.000000  8.000000e+04
50%    1.676385e+07    3233.000000  7.300000e+06
75%    1.715282e+07    7182.000000  1.800000e+07
max    1.769386e+07   14246.000000  2.000000e+09

Frequency counts for categorical variables:
                                                 page_url property_type  \
count                                              191393        191393
unique                                             191393             7
top     https://www.zameen.com/Property/lahore_model_t...         House
freq                                                    1        118915

       price_bin     location     city province_name  \
count     191393       191393   191393        191393
unique         4         1536        5             3
top          Low  DHA Defence  Karachi        Punjab
freq       50175        26161    60484         90714

                          locality
count                       191393
unique                        1619
top     DHA Defence, Lahore, Punjab
freq                         11208
```

## TASK - 7 | Compute the covariance and correlation matrices among descriptive variables.

In [ ]:
```
import pandas as pd

selected_data = pd.read_csv('results_of_selected_house.csv')

# Identify non-numeric columns
non_numeric_columns = selected_data.select_dtypes(exclude=['float64', 'int64']).col
```

```
# Drop non-numeric columns
selected_data_numeric = selected_data.drop(columns=non_numeric_columns)

# Calculate covariance matrix
covariance_matrix = selected_data_numeric.cov()

# Calculate correlation matrix
correlation_matrix = selected_data_numeric.corr()

# Print the covariance matrix
print("Covariance Matrix:")
print(covariance_matrix)

# Print the correlation matrix
print("\nCorrelation Matrix:")
print(correlation_matrix)
```

```
Covariance Matrix:
              property_id   location_id         price
property_id  4.907330e+12 -9.074539e+07 -3.003616e+12
location_id -9.074539e+07  1.383189e+07 -1.033500e+10
price       -3.003616e+12 -1.033500e+10  1.167187e+15

Correlation Matrix:
             property_id  location_id     price
property_id     1.000000    -0.011014 -0.039687
location_id    -0.011014     1.000000 -0.081339
price          -0.039687    -0.081339  1.000000
```

## TASK - 8 | Group the data by city, location, and area.

In [ ]:
```
import pandas as pd

selected_data = pd.read_csv('Property_with_Feature_Engineering.csv')
grouped_data = selected_data.groupby(['city', 'location', 'area'])

# Sort the grouped data by a column, e.g., 'property_id', in descending order
sorted_grouped_data = grouped_data.apply(lambda x: x.sort_values(by='property_id',

# Iterate over the top 10 groups
for group, group_data in sorted_grouped_data.head(10).groupby(level=[0, 1, 2]):
    print("City:", group[0])
    print("Location:", group[1])
    print("Area:", group[2])
    print("Property ID:", group_data['property_id'].mean())
    print("\n")
```

```
City: Faisalabad
Location: 204 Chak Road
Area: 12 Marla
Property ID: 17355575.0


City: Faisalabad
Location: 204 Chak Road
Area: 19 Marla
Property ID: 10053844.0


City: Faisalabad
Location: 204 Chak Road
Area: 2.5 Marla
Property ID: 15029742.0


City: Faisalabad
Location: 204 Chak Road
Area: 3.3 Marla
Property ID: 12038300.0


City: Faisalabad
Location: 204 Chak Road
Area: 4.5 Marla
Property ID: 14375798.0


City: Faisalabad
Location: 204 Chak Road
Area: 5 Marla
Property ID: 12739735.0
```

## TASK - 9 | Count the total values of each item for all attributes.

In [ ]:
```python
import pandas as pd
selected_data = pd.read_csv('Property_with_Feature_Engineering.csv')

for column in selected_data.columns:
    print("Attribute:", column)
    print(selected_data[column].value_counts())
    print("\n")
```

```
Attribute: property_id
property_id
347795      1
17055095    1
17054555    1
17054578    1
17054672    1
           ..
15844370    1
15844668    1
15844823    1
15845395    1
17468660    1
Name: count, Length: 191393, dtype: int64


Attribute: location_id
location_id
1483     2955
1448     1818
329      1763
9030     1695
1447     1595
         ...
9357        1
13645       1
4017        1
3552        1
3216        1
Name: count, Length: 4321, dtype: int64


Attribute: page_url
page_url
https://www.zameen.com/Property/lahore_model_town_6_kanal_excellent_house_for_sale_i
n_model_town-347795-8-1.html
1
https://www.zameen.com/Property/islamabad_pwd_housing_scheme_well_built_house_availa
ble_in_good_location-17055095-424-4.html
1
https://www.zameen.com/Property/islamabad_d_12_well_built_portion_available_in_good_
location-17054555-160-4.html
1
https://www.zameen.com/Property/islamabad_d_12_well_built_portion_available_in_good_
location-17054578-160-4.html
1
https://www.zameen.com/Property/dha_defence_dha_defence_phase_1_upper_portion_availa
ble_in_dha_1-17054672-376-4.html
1

..
https://www.zameen.com/Property/rawalpindi_munawar_colony_house_is_available_for_sal
e-15844370-6034-1.html
1
https://www.zameen.com/Property/bahria_town_rawalpindi_bahria_town_phase_4_1636_squa
re_feet_apartment_is_available_for_sale_in_bahria_heights__bahria_town_phase_4_rawal
```

```
pindi-15844668-3041-1.html    1
https://www.zameen.com/Property/rawalpindi_adiala_road_double_storey_house_is_availa
ble_for_sale-15844823-478-1.html
1
https://www.zameen.com/Property/bahria_town_rawalpindi_bahria_town_phase_8_house_ava
ilable_for_sale-15845395-3048-1.html
1
https://www.zameen.com/Property/i_10_i_10_2_i_10_2_upper_portion_for_rent_good_house
-17468660-3421-4.html
1
Name: count, Length: 191393, dtype: int64


Attribute: property_type
property_type
House            118915
Flat              40157
Upper Portion     18475
Lower Portion     11693
Room               1029
Farm House          725
Penthouse           399
Name: count, dtype: int64


Attribute: price
price
35000        3415
45000        2921
25000        2740
50000        2733
15000000     2730
             ...
8424000         1
219900000       1
8175000         1
369000          1
40              1
Name: count, Length: 2116, dtype: int64


Attribute: price_bin
price_bin
Low          50175
High         48112
Medium       46978
Very High    46128
Name: count, dtype: int64


Attribute: location
location
DHA Defence               26161
Bahria Town Rawalpindi     9278
Bahria Town Karachi        8548
Bahria Town                8244
```

```
Gulistan-e-Jauhar          5877
                          ...
Iqbal Colony                  1
PTV Colony                    1
Raheemabad                    1
Abu Alkhair Road              1
Abid Road                     1
Name: count, Length: 1536, dtype: int64


Attribute: city
city
Karachi       60484
Lahore        58736
Islamabad     40195
Rawalpindi    22898
Faisalabad     9080
Name: count, dtype: int64


Attribute: province_name
province_name
Punjab               90714
Sindh                60484
Islamabad Capital    40195
Name: count, dtype: int64


Attribute: locality
locality
DHA Defence, Lahore, Punjab                                                  112
08
DHA Defence, Karachi, Sindh                                                  109
27
Bahria Town Rawalpindi, Rawalpindi, Punjab                                    92
78
Bahria Town Karachi, Karachi, Sindh                                           85
48
Bahria Town, Lahore, Punjab                                                   65
11
                                                                             ...
Federal Government Employees Housing Foundation, Islamabad, Islamabad Capital
1
Fane Road, Lahore, Punjab
1
Munir Garden, Lahore, Punjab
1
Aiza Garden, Islamabad, Islamabad Capital
1
Abid Road, Lahore, Punjab
1
Name: count, Length: 1619, dtype: int64


Attribute: latitude
latitude
```

```
24.805045    2850
31.462493    2049
31.471571    1818
33.698137    1726
25.020961    1686
             ...
33.672149       1
33.682607       1
33.723193       1
25.004884       1
33.644189       1
Name: count, Length: 8091, dtype: int64


Attribute: longitude
longitude
67.064323    2850
74.445906    1814
72.978215    1726
67.321172    1686
74.409342    1591
             ...
73.013493       1
73.064640       1
73.002754       1
73.120704       1
72.959656       1
Name: count, Length: 8594, dtype: int64


Attribute: baths
baths
0      48130
3      39161
2      29006
4      22258
6      18796
5      17547
1       6530
7       5989
8       2205
10       983
9        763
12        10
11         7
13         3
14         3
403        1
15         1
Name: count, dtype: int64


Attribute: area
area
1 Kanal        25452
5 Marla        24239
```

```
10 Marla       21875
8 Marla        10814
4 Marla         7528
               ...
416 Kanal          1
61.7 Kanal         1
8.6 Kanal          1
5.8 Kanal          1
122.5 Kanal        1
Name: count, Length: 352, dtype: int64


Attribute: area_marla
area_marla
20.0      25458
5.0       24239
10.0      21875
8.0       10814
4.0        7528
          ...
8320.0        1
1234.0        1
172.0         1
116.0         1
2450.0        1
Name: count, Length: 351, dtype: int64


Attribute: area_sqft
area_sqft
5445.02       25458
1361.25       24239
2722.51       21875
2178.01       10814
1089.00        7528
              ...
2265128.32        1
335957.73         1
46827.17          1
31581.12          1
667014.95         1
Name: count, Length: 351, dtype: int64


Attribute: purpose
purpose
For Sale    127018
For Rent     64375
Name: count, dtype: int64


Attribute: bedrooms
bedrooms
3     52643
2     35065
5     27120
```

```
4     26050
0     24959
6     12867
1      5784
7      3246
8      1595
9       861
10      671
11      463
12       29
14        8
15        8
16        4
13        4
28        4
25        4
18        2
27        2
20        2
21        1
68        1
Name: count, dtype: int64


Attribute: date_added
date_added
07-03-2019    10400
07-17-2019     8769
07-04-2019     6815
06-27-2019     6639
07-02-2019     6623
               ...
09-05-2018        3
12-25-2018        1
08-06-2018        1
06-23-2019        1
08-27-2018        1
Name: count, Length: 148, dtype: int64


Attribute: year
year
2019    179084
2018     12309
Name: count, dtype: int64


Attribute: month
month
7     88069
6     50566
5     14132
4      9641
3      6414
2      5440
1      4800
```

```
10     2706
12     2656
11     2589
9      2509
8      1871
Name: count, dtype: int64


Attribute: day
day
3     15734
5     13928
4     12911
17    12509
18    12280
6     10449
20     8857
19     8468
2      8307
10     7433
27     6645
16     6464
11     5555
1      4904
9      4642
23     4332
28     4276
25     4265
12     4182
14     4137
15     4030
21     3920
26     3628
29     3537
13     3461
22     3349
24     3231
30     2514
7      2406
8      1039
Name: count, dtype: int64


Attribute: agency
agency
Mash Allah Estate & Builders    821
Real Investment Consultants     794
Future Planners                 561
Lahore Grande Estate            463
Arham Estate                    429
                               ...
Great Deal Associates             1
Al Imran Real Estate              1
Dial 4 Property                   1
Savul Estate                      1
Al Basit Estate & Advisors        1
```

```
Name: count, Length: 5923, dtype: int64


Attribute: agent
agent
Azam Ali                                                                 797
Boez Ayub                                                                787
Muhammad Imran                                                           597
Kashif                                                                   400
Daud Ahmad(Co-CEO), Shafique Arshad Waince(Co-CEO), Zafar Iqbal Bajwa (CEO)    375
                                                                        ...
Mian Muhammad Farooq                                                       1
miss sana Gul                                                              1
Rageeb                                                                     1
Orient Associates                                                          1
TALHA   MIAN AHMAD                                                         1
Name: count, Length: 11352, dtype: int64
```

## TASK - 10 | Encode categorical values of 'property_type' and 'province_name' features with numbers.

In [ ]:
```
from sklearn.preprocessing import LabelEncoder

selected_data = pd.read_csv('Property_with_Feature_Engineering.csv')
label_encoder = LabelEncoder()
selected_data['property_type_encoded'] = label_encoder.fit_transform(selected_data[
selected_data['province_name_encoded'] = label_encoder.fit_transform(selected_data[
print(selected_data.head())
```

```
   property_id  location_id  \
0       347795            8
1       482892           48
2       555962           75
3       562843         3821
4       686990         3522

                                          page_url property_type      price  \
0  https://www.zameen.com/Property/lahore_model_t...         House  220000000
1  https://www.zameen.com/Property/lahore_multan_...         House   40000000
2  https://www.zameen.com/Property/eden_eden_aven...         House    9500000
3  https://www.zameen.com/Property/gulberg_2_gulb...         House  125000000
4  https://www.zameen.com/Property/allama_iqbal_t...         House   21000000

   price_bin           location    city province_name  \
0  Very High         Model Town  Lahore        Punjab
1  Very High        Multan Road  Lahore        Punjab
2        Low               Eden  Lahore        Punjab
3  Very High            Gulberg  Lahore        Punjab
4       High  Allama Iqbal Town  Lahore        Punjab

                          locality  ...   purpose  bedrooms  date_added  \
0        Model Town, Lahore, Punjab  ...  For Sale         0  07-17-2019
1       Multan Road, Lahore, Punjab  ...  For Sale         5  10-06-2018
2             Eden, Lahore, Punjab  ...  For Sale         3  07-03-2019
3          Gulberg, Lahore, Punjab  ...  For Sale         8  04-04-2019
4  Allama Iqbal Town, Lahore, Punjab  ...  For Sale         6  04-04-2019

   year  month  day                  agency                   agent  \
0  2019      7   17  Real Biz International              Usama Khan
1  2018     10    6             Khan Estate         mohsinkhan and B
2  2019      7    3         Shahum Estate 2  Babar Hameed, Raja Omar
3  2019      4    4                     NaN                     NaN
4  2019      4    4                     NaN                     NaN

  property_type_encoded  province_name_encoded
0                     2                      1
1                     2                      1
2                     2                      1
3                     2                      1
4                     2                      1

[5 rows x 26 columns]
```