```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

pd.options.display.max_columns = None
```

Task 2

```
df = pd.read_csv("House_Price_India.csv")
df.head(5)
```

```
            id   Date  number of bedrooms  number of bathrooms  living
area  \
0  6762810145  42491                   5                 2.50
3650
1  6762810635  42491                   4                 2.50
2920
2  6762810998  42491                   5                 2.75
2910
3  6762812605  42491                   4                 2.50
3310
4  6762812919  42491                   3                 2.00
2710

   lot area  number of floors  waterfront present  number of views  \
0      9050               2.0                   0                4
1      4000               1.5                   0                0
2      9480               1.5                   0                0
3     42998               2.0                   0                0
4      4500               1.5                   0                0

   condition of the house  grade of the house  \
0                       5                  10
1                       5                   8
2                       3                   8
3                       3                   9
4                       4                   8

   Area of the house(excluding basement)  Area of the basement  Built
Year  \
0                                   3370                   280
1921
1                                   1910                  1010
1909
2                                   2910                     0
1939
3                                   3310                     0
2001
4                                   1880                   830
```

```
1929

   Renovation Year  Postal Code  Lattitude  Longitude
living_area_renov  \
0                0       122003    52.8645   -114.557
2880
1                0       122004    52.8878   -114.470
2470
2                0       122004    52.8852   -114.468
2940
3                0       122005    52.9532   -114.321
3350
4                0       122006    52.9047   -114.485
2060

   lot_area_renov  Number of schools nearby  Distance from the airport
\
0            5400                         2                         58

1            4000                         2                         51

2            6600                         1                         53

3           42847                         3                         76

4            4500                         1                         51


     Price
0  2380000
1  1400000
2  1200000
3   838000
4   805000

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14620 entries, 0 to 14619
Data columns (total 23 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   id                                     14620 non-null  int64
 1   Date                                   14620 non-null  int64
 2   number of bedrooms                     14620 non-null  int64
 3   number of bathrooms                    14620 non-null  float64
 4   living area                            14620 non-null  int64
 5   lot area                               14620 non-null  int64
 6   number of floors                       14620 non-null  float64
 7   waterfront present                     14620 non-null  int64
```

```
 8   number of views                            14620 non-null  int64
 9   condition of the house                     14620 non-null  int64
 10  grade of the house                         14620 non-null  int64
 11  Area of the house(excluding basement)      14620 non-null  int64
 12  Area of the basement                       14620 non-null  int64
 13  Built Year                                 14620 non-null  int64
 14  Renovation Year                            14620 non-null  int64
 15  Postal Code                                14620 non-null  int64
 16  Lattitude                                  14620 non-null  float64
 17  Longitude                                  14620 non-null  float64
 18  living_area_renov                          14620 non-null  int64
 19  lot_area_renov                             14620 non-null  int64
 20  Number of schools nearby                   14620 non-null  int64
 21  Distance from the airport                  14620 non-null  int64
 22  Price                                      14620 non-null  int64
dtypes: float64(4), int64(19)
memory usage: 2.6 MB
```

```
df.dtypes
```

```
id                                           int64
Date                                         int64
number of bedrooms                           int64
number of bathrooms                        float64
living area                                  int64
lot area                                     int64
number of floors                           float64
waterfront present                           int64
number of views                              int64
condition of the house                       int64
grade of the house                           int64
Area of the house(excluding basement)        int64
Area of the basement                         int64
Built Year                                   int64
Renovation Year                              int64
Postal Code                                  int64
Lattitude                                  float64
Longitude                                  float64
living_area_renov                            int64
lot_area_renov                               int64
Number of schools nearby                     int64
Distance from the airport                    int64
Price                                        int64
dtype: object
```

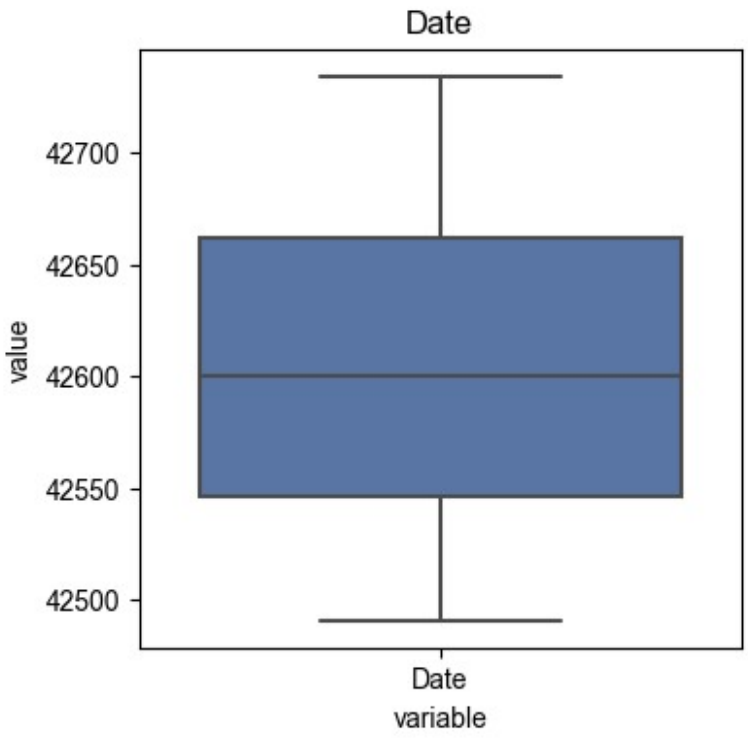
```
df.nunique()
```

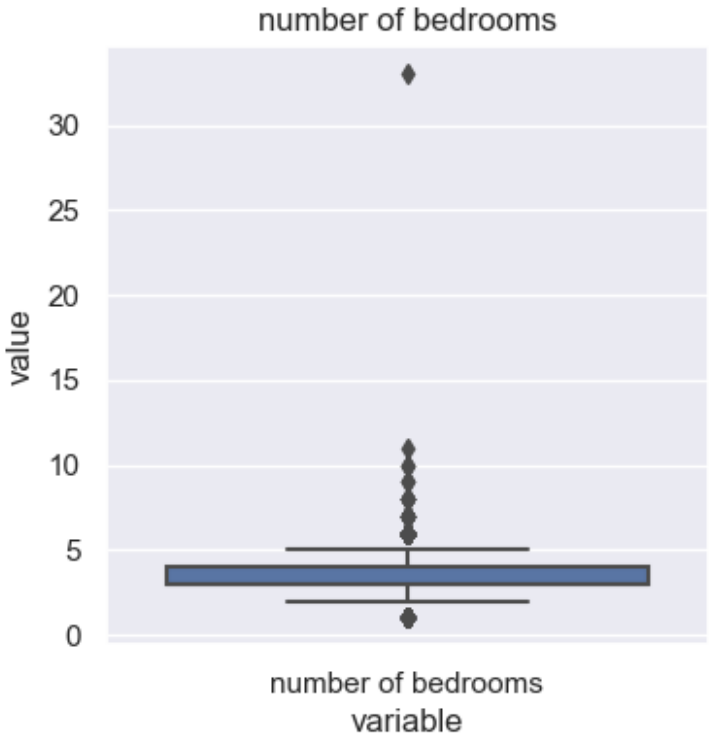
```
id                                         14620
Date                                         241
number of bedrooms                            12
```

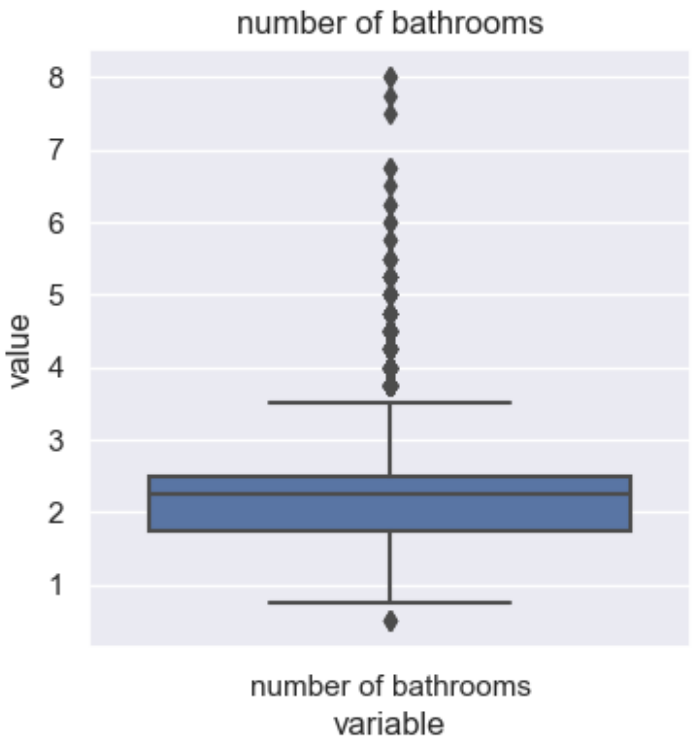
```
number of bathrooms                          29
living area                                 865
lot area                                   7451
number of floors                              6
waterfront present                            2
number of views                               5
condition of the house                        5
grade of the house                           10
Area of the house(excluding basement)       781
Area of the basement                        280
Built Year                                  116
Renovation Year                              68
Postal Code                                  70
Lattitude                                  4662
Longitude                                   716
living_area_renov                           665
lot_area_renov                             6835
Number of schools nearby                      3
Distance from the airport                    31
Price                                      2901
dtype: int64
```

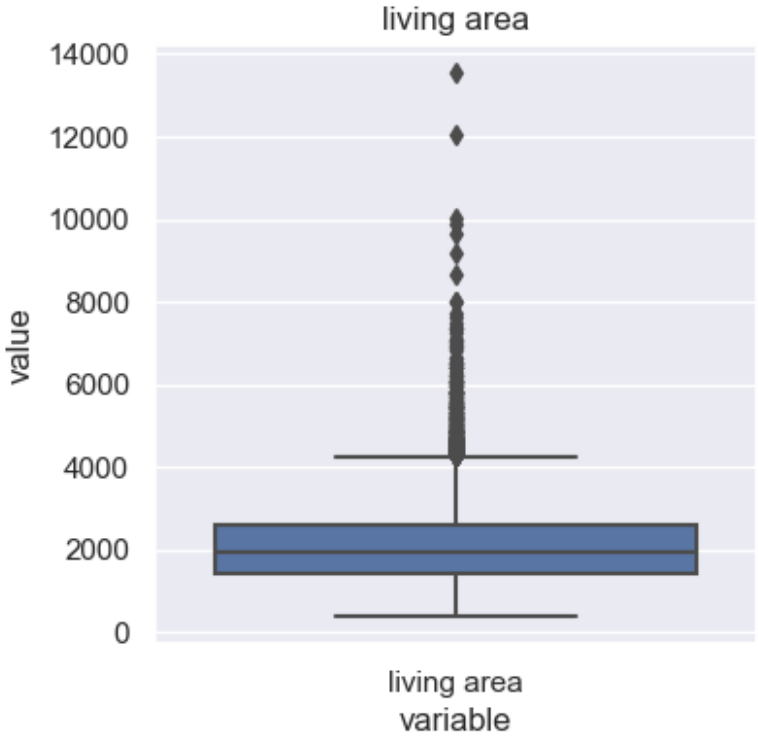
Task 3

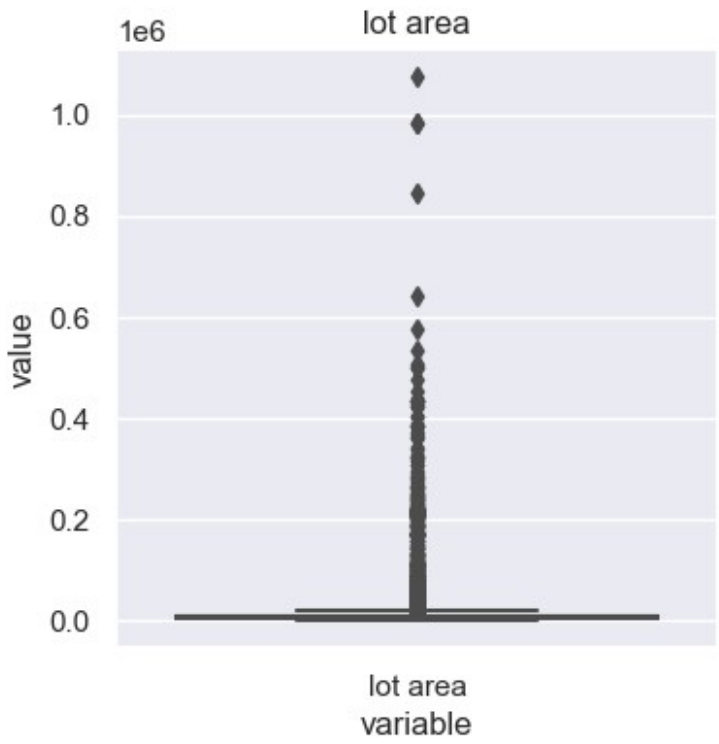
```
# univariate anaylsis using box plot
for col in ["Date", "number of bedrooms", "number of bathrooms",
"living area", "lot area", "Area of the house(excluding basement)"]:
    plt.figure(figsize=(4,4))
    plt.title(col)
    dfM = df.loc[:, [col]]
    sns.set(rc={'figure.figsize':(4,4)})
    sns.boxplot(x="variable", y="value", data=pd.melt(dfM))
```

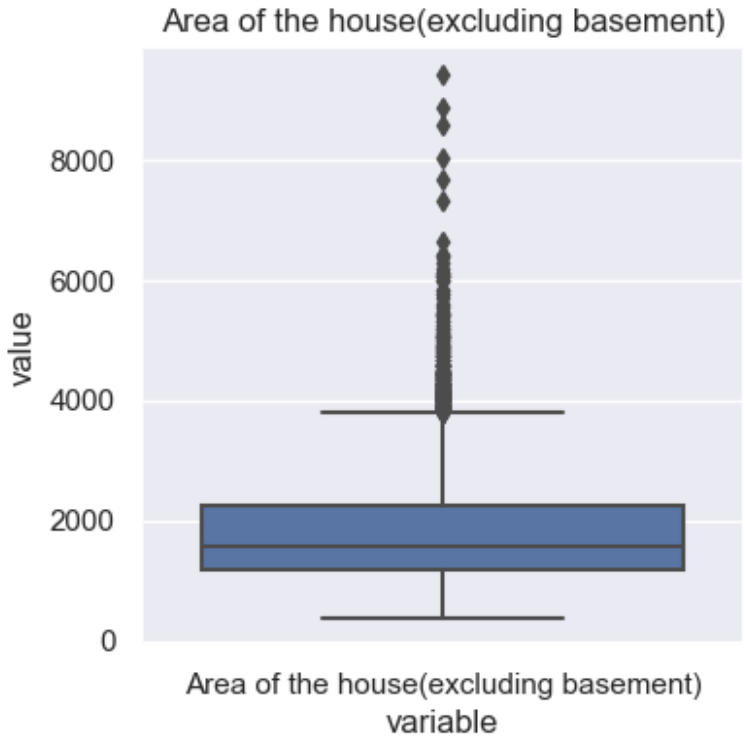

Date
42700
42650
42600
42550
42500
value
Date
variable


number of bedrooms
30
25
20
15
10
5
0
value
number of bedrooms
variable


number of bathrooms
8
7
6
5
4
3
2
1
value
number of bathrooms
variable


living area
14000
12000
10000
8000
6000
4000
2000
0
value
living area
variable


lot area
1e6
1.0
0.8
0.6
0.4
0.2
0.0
value
lot area
variable


Area of the house(excluding basement)
8000
6000
4000
2000
0
value
Area of the house(excluding basement)
variable

```
df.columns
```

```
Index(['id', 'Date', 'number of bedrooms', 'number of bathrooms',
       'living area', 'lot area', 'number of floors', 'waterfront
present',
       'number of views', 'condition of the house', 'grade of the
house',
       'Area of the house(excluding basement)', 'Area of the
basement',
       'Built Year', 'Renovation Year', 'Postal Code', 'Lattitude',
       'Longitude', 'living_area_renov', 'lot_area_renov',
       'Number of schools nearby', 'Distance from the airport',
'Price'],
      dtype='object')
```

```
#bivariate analysis
sns.pairplot(df.drop(columns=["id", "Longitude", "Lattitude", "Postal
Code", "condition of the house", "number of views"]))
```

```
/opt/homebrew/anaconda3/envs/MLEnv/lib/python3.10/site-packages/
seaborn/axisgrid.py:118: UserWarning: The figure layout has changed to
tight
  self._figure.tight_layout(*args, **kwargs)
```

```
<seaborn.axisgrid.PairGrid at 0x16d17f610>
```

Task 4

```
df.describe()
```

```
                 id           Date  number of bedrooms  number of
bathrooms  \
count  1.462000e+04   14620.000000         14620.000000
14620.000000
mean   6.762821e+09   42604.538646             3.379343
2.129583
std    6.237575e+03      67.347991             0.938719
0.769934
```

```
min    6.762810e+09  42491.000000            1.000000
0.500000
25%    6.762815e+09  42546.000000            3.000000
1.750000
50%    6.762821e+09  42600.000000            3.000000
2.250000
75%    6.762826e+09  42662.000000            4.000000
2.500000
max    6.762832e+09  42734.000000           33.000000
8.000000

        living area      lot area  number of floors  waterfront
present  \
count  14620.000000  1.462000e+04      14620.000000
14620.000000
mean    2098.262996  1.509328e+04          1.502360
0.007661
std      928.275721  3.791962e+04          0.540239
0.087193
min      370.000000  5.200000e+02          1.000000
0.000000
25%     1440.000000  5.010750e+03          1.000000
0.000000
50%     1930.000000  7.620000e+03          1.500000
0.000000
75%     2570.000000  1.080000e+04          2.000000
0.000000
max    13540.000000  1.074218e+06          3.500000
1.000000

       number of views  condition of the house  grade of the house  \
count     14620.000000            14620.000000        14620.000000
mean          0.233105                3.430506            7.682421
std           0.766259                0.664151            1.175033
min           0.000000                1.000000            4.000000
25%           0.000000                3.000000            7.000000
50%           0.000000                3.000000            7.000000
75%           0.000000                4.000000            8.000000
max           4.000000                5.000000           13.000000

       Area of the house(excluding basement)  Area of the basement  \
count                           14620.000000          14620.000000
mean                             1801.783926            296.479070
std                               833.809963            448.551409
min                               370.000000              0.000000
25%                              1200.000000              0.000000
50%                              1580.000000              0.000000
75%                              2240.000000            580.000000
max                              9410.000000           4820.000000
```

```
         Built Year  Renovation Year    Postal Code     Lattitude  \
count  14620.000000     14620.000000   14620.000000  14620.000000
mean    1970.926402        90.924008  122033.062244     52.792848
std       29.493625       416.216661      19.082418      0.137522
min     1900.000000         0.000000  122003.000000     52.385900
25%     1951.000000         0.000000  122017.000000     52.707600
50%     1975.000000         0.000000  122032.000000     52.806400
75%     1997.000000         0.000000  122048.000000     52.908900
max     2015.000000      2015.000000  122072.000000     53.007600

          Longitude  living_area_renov  lot_area_renov  \
count  14620.000000       14620.000000    14620.000000
mean    -114.404007        1996.702257    12753.500068
std        0.141326         691.093366    26058.414467
min     -114.709000         460.000000      651.000000
25%     -114.519000        1490.000000     5097.750000
50%     -114.421000        1850.000000     7620.000000
75%     -114.315000        2380.000000    10125.000000
max     -113.505000        6110.000000   560617.000000

       Number of schools nearby  Distance from the airport
Price
count              14620.000000               14620.000000
1.462000e+04
mean                   2.012244                  64.950958
5.389322e+05
std                    0.817284                   8.936008
3.675324e+05
min                    1.000000                  50.000000
7.800000e+04
25%                    1.000000                  57.000000
3.200000e+05
50%                    2.000000                  65.000000
4.500000e+05
75%                    3.000000                  73.000000
6.450000e+05
max                    3.000000                  80.000000
7.700000e+06
```

Task 5

```
median = df.median()
df.fillna(median)

               id   Date  number of bedrooms  number of bathrooms  \
0      6762810145  42491                   5                 2.50
1      6762810635  42491                   4                 2.50
2      6762810998  42491                   5                 2.75
3      6762812605  42491                   4                 2.50
```

```
4      6762812919  42491                   3                   2.00
...           ...    ...                 ...                    ...
14615  6762830250  42734                   2                   1.50
14616  6762830339  42734                   3                   2.00
14617  6762830618  42734                   2                   1.00
14618  6762830709  42734                   4                   1.00
14619  6762831463  42734                   3                   1.00

       living area  lot area  number of floors  waterfront present  \
0             3650      9050               2.0                   0
1             2920      4000               1.5                   0
2             2910      9480               1.5                   0
3             3310     42998               2.0                   0
4             2710      4500               1.5                   0
...            ...       ...               ...                 ...
14615         1556     20000               1.0                   0
14616         1680      7000               1.5                   0
14617         1070      6120               1.0                   0
14618         1030      6621               1.0                   0
14619          900      4770               1.0                   0

       number of views  condition of the house  grade of the house  \
0                    4                       5                  10
1                    0                       5                   8
2                    0                       3                   8
3                    0                       3                   9
4                    0                       4                   8
...                ...                     ...                 ...
14615                0                       4                   7
14616                0                       4                   7
14617                0                       3                   6
14618                0                       4                   6
14619                0                       3                   6

       Area of the house(excluding basement)  Area of the basement  \
0                                       3370                   280
1                                       1910                  1010
2                                       2910                     0
3                                       3310                     0
4                                       1880                   830
...                                      ...                   ...
14615                                   1556                     0
14616                                   1680                     0
14617                                   1070                     0
14618                                   1030                     0
14619                                    900                     0

       Built Year  Renovation Year  Postal Code  Lattitude  Longitude
\
0            1921                0       122003    52.8645   -114.557
```

```
1             1909                0       122004    52.8878   -114.470

2             1939                0       122004    52.8852   -114.468

3             2001                0       122005    52.9532   -114.321

4             1929                0       122006    52.9047   -114.485

...            ...              ...          ...        ...        ...

14615         1957                0       122066    52.6191   -114.472

14616         1968                0       122072    52.5075   -114.393

14617         1962                0       122056    52.7289   -114.507

14618         1955                0       122042    52.7157   -114.411

14619         1969             2009       122018    52.5338   -114.552


       living_area_renov  lot_area_renov  Number of schools nearby  \
0                   2880            5400                         2
1                   2470            4000                         2
2                   2940            6600                         1
3                   3350           42847                         3
4                   2060            4500                         1
...                  ...             ...                       ...
14615               2250           17286                         3
14616               1540            7480                         3
14617               1130            6120                         2
14618               1420            6631                         3
14619                900            3480                         2

       Distance from the airport    Price
0                             58  2380000
1                             51  1400000
2                             53  1200000
3                             76   838000
4                             51   805000
...                          ...      ...
14615                         76   221700
14616                         59   219200
14617                         64   209000
14618                         54   205000
14619                         55   146000

[14620 rows x 23 columns]
```