```
pip install pandas
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pandas in /usr/local/lib/python3.9/dist-packages (1.5.3)
Requirement already satisfied: numpy>=1.20.3 in /usr/local/lib/python3.9/dist-packages (from pandas) (1.22.4)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.9/dist-packages (from pandas) (2022.7.1)
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.9/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.9/dist-packages (from python-dateutil>=2.8.1->pandas) (1.16.0)
```

```
pip install seaborn
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: seaborn in /usr/local/lib/python3.9/dist-packages (0.12.2)
Requirement already satisfied: pandas>=0.25 in /usr/local/lib/python3.9/dist-packages (from seaborn) (1.5.3)
Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in /usr/local/lib/python3.9/dist-packages (from seaborn) (3.7.1)
Requirement already satisfied: numpy!=1.24.0,>=1.17 in /usr/local/lib/python3.9/dist-packages (from seaborn) (1.22.4)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (1.0.7
Requirement already satisfied: importlib-resources>=3.2.0 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seabo
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (8.4.0)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (2
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (1.4.
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (0.11.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (3.0.9
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (23.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.9/dist-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (4.39
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.9/dist-packages (from pandas>=0.25->seaborn) (2022.7.1)
Requirement already satisfied: zipp>=3.1.0 in /usr/local/lib/python3.9/dist-packages (from importlib-resources>=3.2.0->matplotlib!=3.6.1
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.9/dist-packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.1->s
```

```
pip install matplotlib
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: matplotlib in /usr/local/lib/python3.9/dist-packages (3.7.1)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (1.4.4)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (1.0.7)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (4.39.3)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (23.1)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (8.4.0)
Requirement already satisfied: importlib-resources>=3.2.0 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (5.12.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (3.0.9)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (0.11.0)
Requirement already satisfied: numpy>=1.20 in /usr/local/lib/python3.9/dist-packages (from matplotlib) (1.22.4)
Requirement already satisfied: zipp>=3.1.0 in /usr/local/lib/python3.9/dist-packages (from importlib-resources>=3.2.0->matplotlib) (3.15
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.9/dist-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
```

```
pip install numpy
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: numpy in /usr/local/lib/python3.9/dist-packages (1.22.4)
```

Load the dataset

```
import pandas as pd
import numpy as np
```

Double-click (or enter) to edit

```
df=pd.read_csv('/content/drive/MyDrive/House.csv')
```

```
df
```

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	..
0	6762810145	42491	5	2.50	3650	9050	2.0	0	4	5	.
1	6762810635	42491	4	2.50	2920	4000	1.5	0	0	5	.
2	6762810998	42491	5	2.75	2910	9480	1.5	0	0	3	.
3	6762812605	42491	4	2.50	3310	42998	2.0	0	0	3	.
4	6762812919	42491	3	2.00	2710	4500	1.5	0	0	4	.
...	...	...	...	...	...	...	...	...	...	...	.
14615	6762830250	42734	2	1.50	1556	20000	1.0	0	0	4	.
14616	6762830339	42734	3	2.00	1680	7000	1.5	0	0	4	.
14617	6762830618	42734	2	1.00	1070	6120	1.0	0	0	3	.
14618	6762830709	42734	4	1.00	1030	6621	1.0	0	0	4	.
14619	6762831463	42734	3	1.00	900	4770	1.0	0	0	3	.

```
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
```

Visualization

Univariate Analysis

Double-click (or enter) to edit

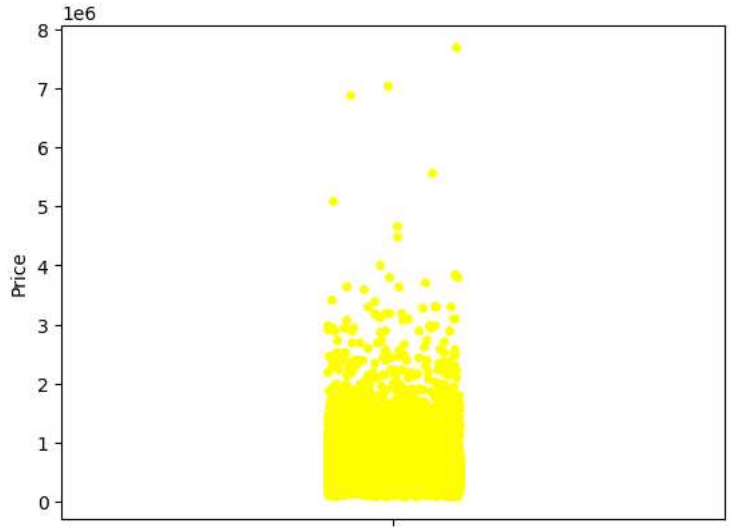
```
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

```
plt.hist(df['Price'])
```

```
(array([1.2916e+04, 1.4260e+03, 1.9100e+02, 6.0000e+01, 1.9000e+01,
        2.0000e+00, 2.0000e+00, 1.0000e+00, 1.0000e+00, 2.0000e+00]),
 array([  78000.,  840200., 1602400., 2364600., 3126800., 3889000.,
        4651200., 5413400., 6175600., 6937800., 7700000.]),
 <BarContainer object of 10 artists>)
```


```
sns.stripplot(y=df['Price'],color='yellow')
```

```
<Axes: ylabel='Price'>
```



Double-click (or enter) to edit

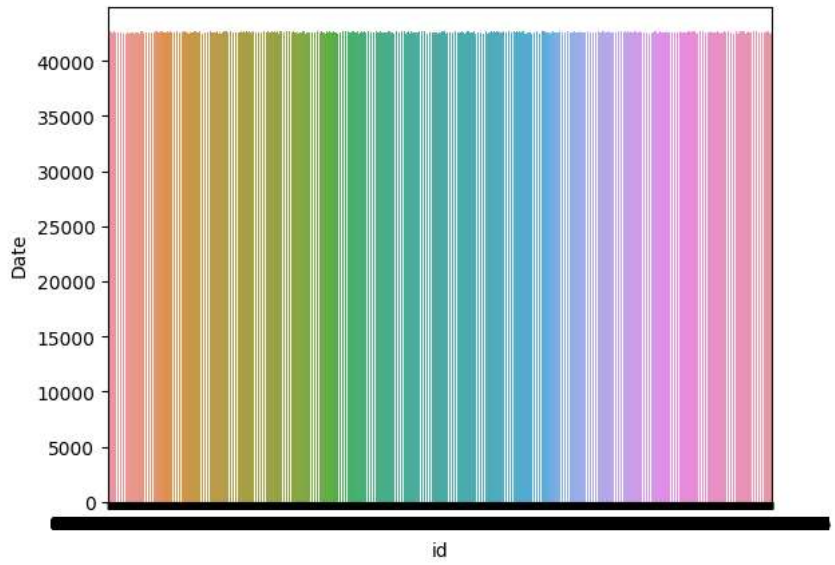
Bi-Variate Analysis

```
rate=pd.read_csv('/content/drive/MyDrive/House.csv')
rate.plot(x='id',y='Price',kind='scatter',color='indigo');
plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

```
sns.barplot(data=rate,x='id',y='Date');
plt.show
```

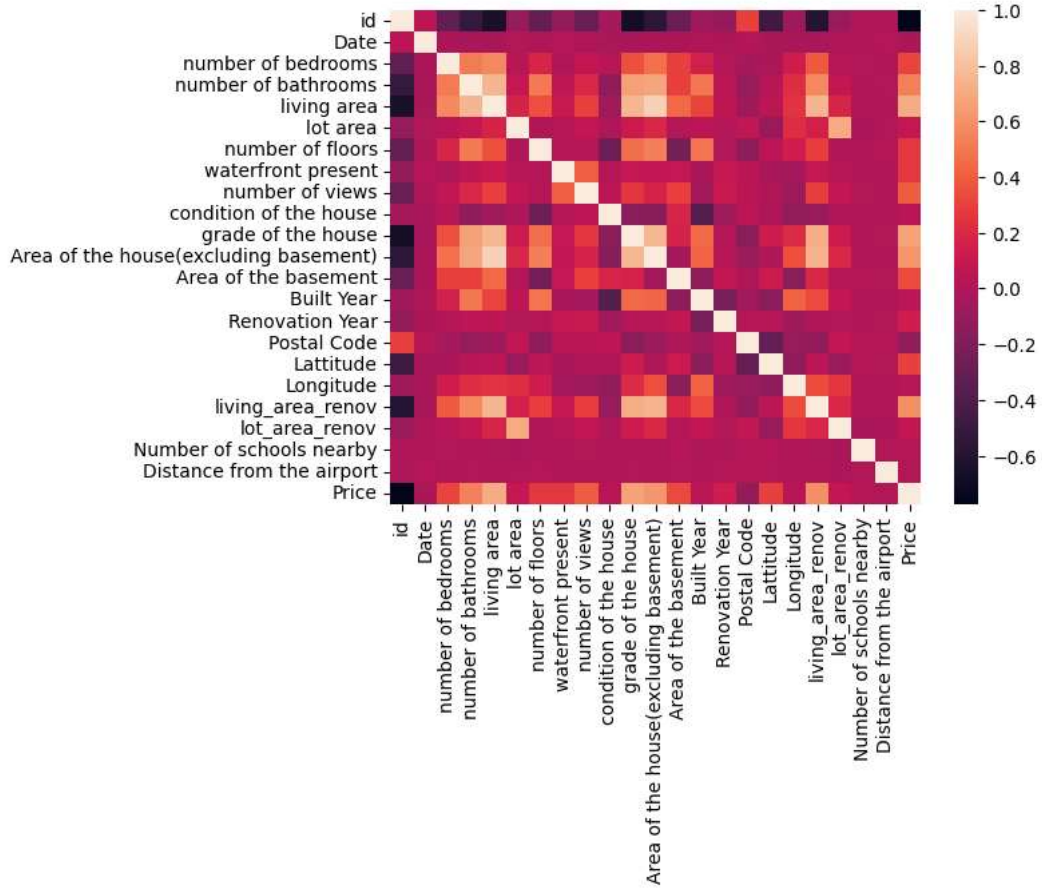
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



Multi-Variate Analysis

```
sns.heatmap(rate.corr(),annot=False)
```

```
<Axes: >
```



Descriptive Statitics

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

```
import warnings
warnings.filterwarnings('ignore')
```

```
data='/content/drive/MyDrive/House.csv'
df=pd.read_csv(data)
```

```
df.shape
```

```
(14620, 23)
```

```
df.head()
```

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	...	Bu Y
0	6762810145	42491	5	2.50	3650	9050	2.0	0	4	5	...	1
1	6762810635	42491	4	2.50	2920	4000	1.5	0	0	5	...	1
2	6762810998	42491	5	2.75	2910	9480	1.5	0	0	3	...	1
3	6762812605	42491	4	2.50	3310	42998	2.0	0	0	3	...	2
4	6762812919	42491	3	2.00	2710	4500	1.5	0	0	4	...	1

5 rows × 23 columns

```
df.describe()
```

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	1
count	1.462000e+04	14620.000000	14620.000000	14620.000000	14620.000000	1.462000e+04	14620.000000	
mean	6.762821e+09	42604.538646	3.379343	2.129583	2098.262996	1.509328e+04	1.502360	
std	6.237575e+03	67.347991	0.938719	0.769934	928.275721	3.791962e+04	0.540239	
min	6.762810e+09	42491.000000	1.000000	0.500000	370.000000	5.200000e+02	1.000000	
25%	6.762815e+09	42546.000000	3.000000	1.750000	1440.000000	5.010750e+03	1.000000	
50%	6.762821e+09	42600.000000	3.000000	2.250000	1930.000000	7.620000e+03	1.500000	
75%	6.762826e+09	42662.000000	4.000000	2.500000	2570.000000	1.080000e+04	2.000000	
max	6.762832e+09	42734.000000	33.000000	8.000000	13540.000000	1.074218e+06	3.500000	

8 rows × 23 columns

```
df.describe(include='all')
```

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors
count	1.462000e+04	14620.000000	14620.000000	14620.000000	14620.000000	1.462000e+04	14620.000000
mean	6.762821e+09	42604.538646	3.379343	2.129583	2098.262996	1.509328e+04	1.502360
std	6.237575e+03	67.347991	0.938719	0.769934	928.275721	3.791962e+04	0.540239
25%	6.762815e+09	42546.000000	3.000000	1.750000	1440.000000	5.010750e+03	1.000000

Mean

```
mean=df['waterfront present'].mean()
print(mean)
```

```
0.007660738714090287
```

Median

```
median=df['waterfront present'].median()
print(median)
```

```
0.0
```

Mode

```
mode=df['waterfront present'].mode()
print(mode)
```

```
0    0
Name: waterfront present, dtype: int64
```

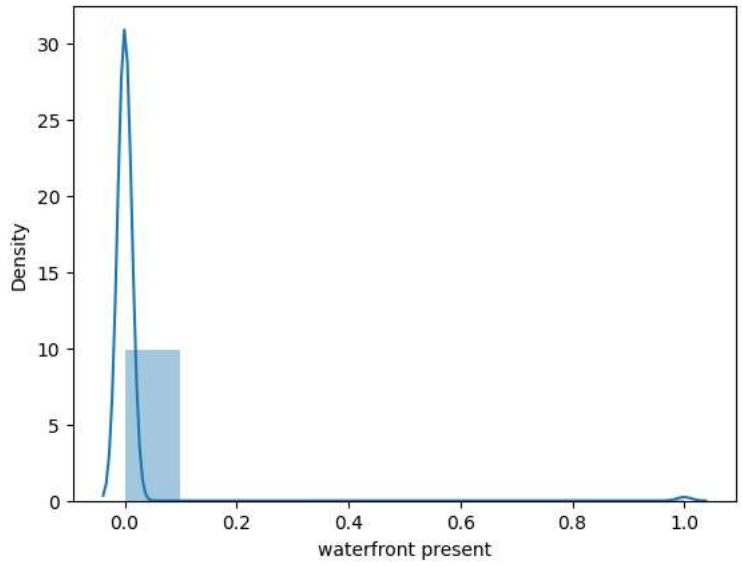
Observation

We can see that *Mean>Median>Mode*.So,the distribution of *waterfront present* is positively skewed.I will plot its ditribution to confirm the same. *bold text*

Plot the distribution

```
data=df['waterfront present']
sns.distplot(data,bins=10,hist=True,kde=True,label='waterfront present')
```

```
<Axes: xlabel='waterfront present', ylabel='Density'>
```



Check for missing values

```
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
```

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

```
id                                    0
Date                                  0
number of bedrooms                    0
number of bathrooms                   0
living area                           0
lot area                              0
number of floors                      0
waterfront present                    0
number of views                       0
condition of the house                0
grade of the house                    0
Area of the house(excluding basement)    0
Area of the basement                  0
Built Year                            0
Renovation Year                       0
Postal Code                           0
Lattitude                             0
Longitude                             0
living_area_renov                     0
lot_area_renov                        0
Number of schools nearby              0
Distance from the airport             0
Price                                 0
dtype: int64
```

```
updated_df=df.dropna(axis=1)
```

```
updated_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
```

```
updated_df=df.dropna(axis=0)
```

```
updated_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
```

```
updated_df=df
updated_df['waterfront present']=updated_df['waterfront present'].fillna(updated_df['waterfront present'].mean())
updated_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
```

```
updated_df=df
updated_df['waterfront present missing']=updated_df['waterfront present'].isnull()
from sklearn.impute import SimpleImputer
my_imputer=SimpleImputer(strategy='median')
data_new=my_imputer.fit_transform(updated_df)
updated_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14620 entries, 0 to 14619
Data columns (total 24 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
 23  waterfront present missing            14620 non-null  bool
dtypes: bool(1), float64(4), int64(19)
memory usage: 2.6 MB
```