VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Name- Utkarsh Jain

Reg. No- 21BCE3095

Campus-Vellore

Email Id- utkarsh.jain2021@vitstudent.com

ASSIGNMENT - 2

1. Download the dataset: Dataset(Downloaded in file)

2. Load the dataset.

Utkarsh Jain

21BCE3095

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

```
df=pd.read_csv('/content/House Price India.csv')
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	...	Built Year	Renovation Year	Postal Code	Lattitude	Longitude	living_area_renov	lot_area_
0	6762810145	42491	5	2.50	3650	9050	2.0	0	4	5	...	1921	0	122003	52.8645	-114.557	2880	
1	6762810635	42491	4	2.50	2920	4000	1.5	0	0	5	...	1909	0	122004	52.8878	-114.470	2470	
2	6762810998	42491	5	2.75	2910	9480	1.5	0	0	3	...	1939	0	122004	52.8852	-114.468	2940	
3	6762812605	42491	4	2.50	3310	42998	2.0	0	0	3	...	2001	0	122005	52.9532	-114.321	3350	
4	6762812919	42491	3	2.00	2710	4500	1.5	0	0	4	...	1929	0	122006	52.9047	-114.485	2060	

5 rows × 23 columns

21BCE3095

```
[16] df.shape
```

```
(14620, 23)
```

```
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
```

21BCE3095

```
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
```

2. Perform the below visualization

>>Univariate Analysis

21BCE3095

Double-click (or enter) to edit

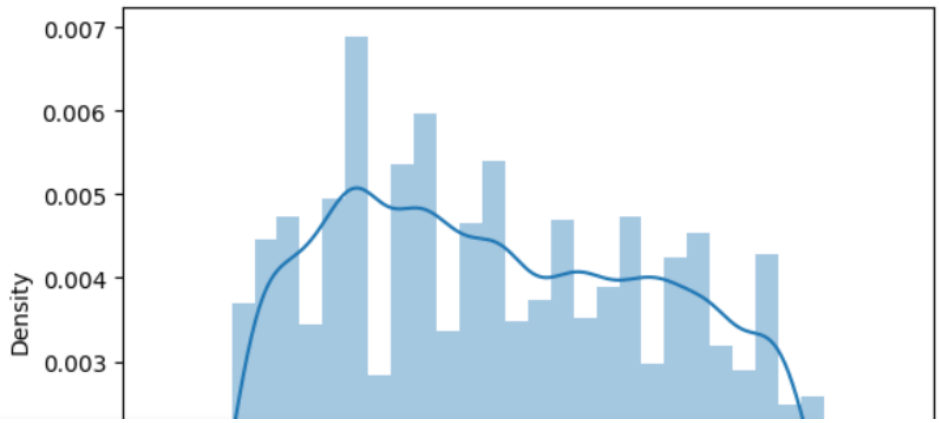
```
sns.distplot(df.lot_area_renov)
```

```
<ipython-input-19-fbd8f64c04a5>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df.lot_area_renov)
<Axes: xlabel='lot_area_renov', ylabel='Density'>
```

21BCE3095

```
sns.distplot(df.Date)
```

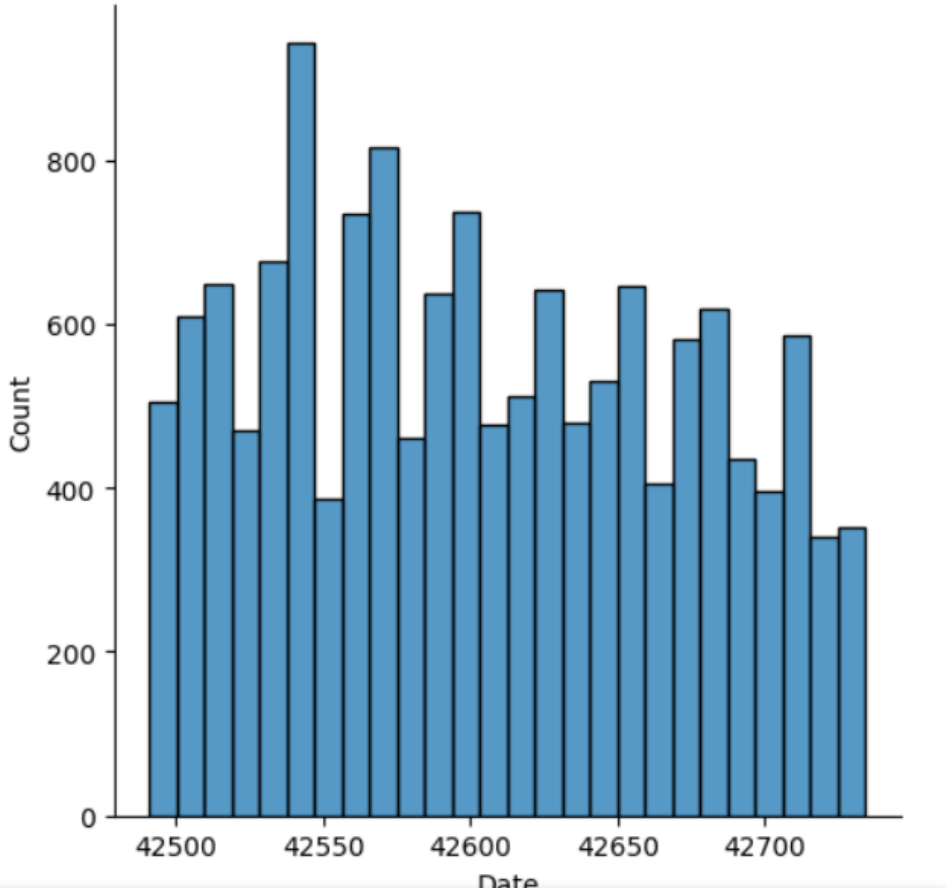
```
<ipython-input-20-82d9cb3bf0f8>:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df.Date)
<Axes: xlabel='Date', ylabel='Density'>
```



21BCE3095

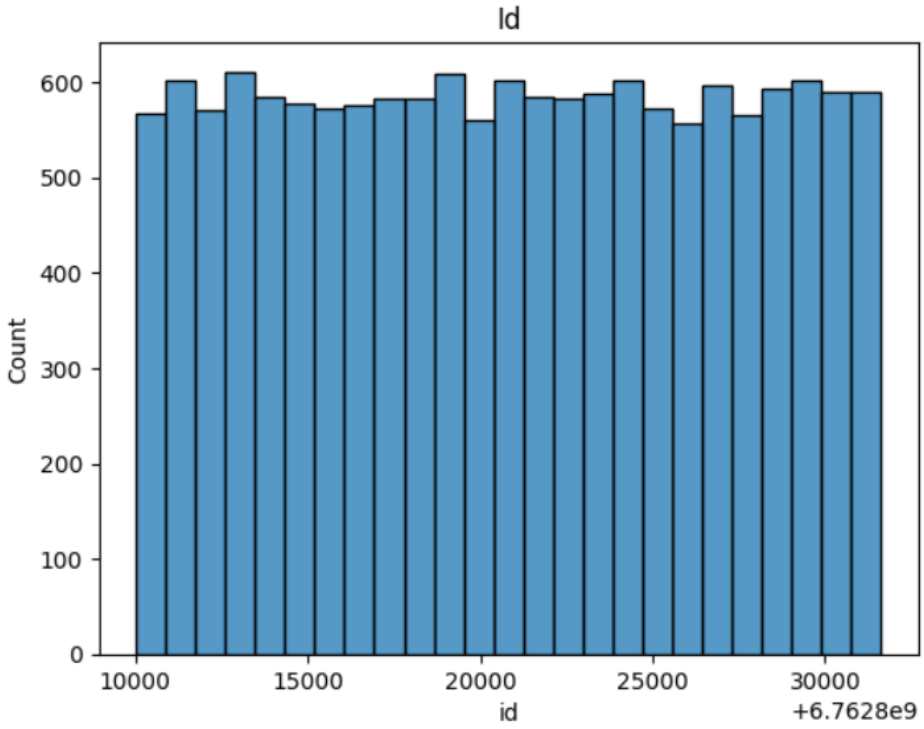
```
sns.displot(df.Date)
```

```
<seaborn.axisgrid.FacetGrid at 0x7d51461c82b0>
```



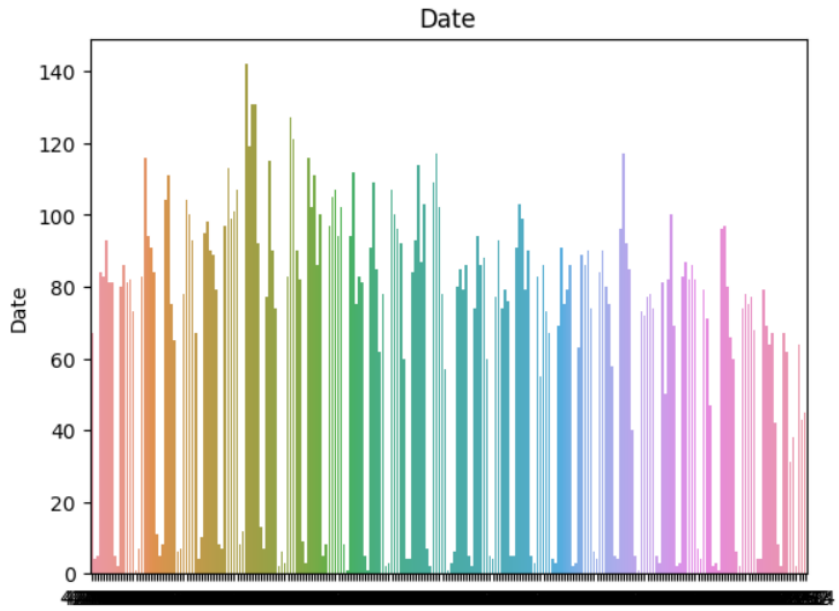
```
[22] sns.histplot(df.id)
     plt.title('Id')
```

```
Text(0.5, 1.0, 'Id')
```



```
sns.barplot(x=df.Date.value_counts().index,y=df.Date.value_counts())
plt.title("Date")
```

```
Text(0.5, 1.0, 'Date')
```

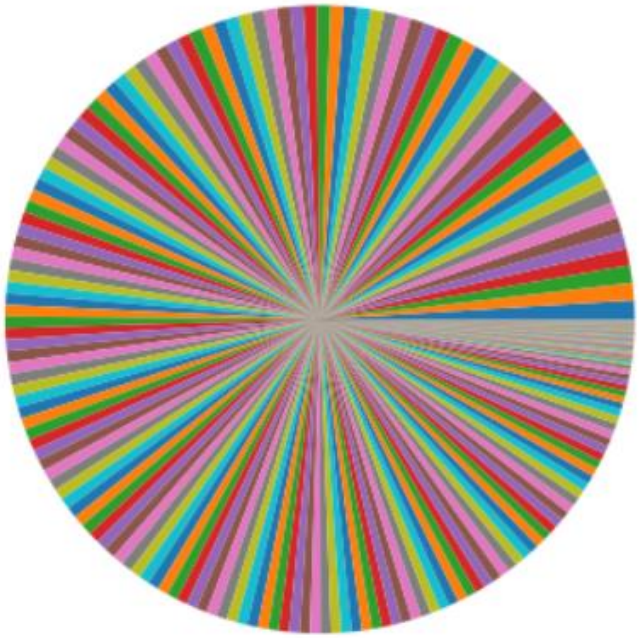


21BCE3095

```
plt.pie(df.Date.value_counts())
```

```
Text(1.0992134236172717, -0.04159145753152115, ''),
Text(1.0992832978403189, -0.039701776903725804, ''),
Text(1.099349923476552, -0.03781197894978788, ''),
Text(1.0994133003290807, -0.03592206925441221, ''),
Text(1.0994734282106142, -0.03403205340263377, ''),
Text(1.0995303069434634, -0.032141936979801326, ''),
Text(1.099583936359541, -0.030251725571560816, ''),
Text(1.0996343163003623, -0.028361424763839835, ''),
Text(1.099675733046743, -0.026707342587915683, ''),
Text(1.0997092528046615, -0.025289509600090875, ''),
Text(1.099740944524109, -0.023871634573704267, ''),
Text(1.099770808152405, -0.022453719865676968, ''),
Text(1.0997988436399069, -0.021035767832998972, ''),
Text(1.0998250509400118, -0.019617780832721345, ''),
Text(1.0998494300091555, -0.018199761221954252, ''),
Text(1.0998719808068127, -0.016781711357859143, ''),
Text(1.0998927032954975, -0.015363633597650682, ''),
Text(1.0999115974407632, -0.013945530298588932, ''),
Text(1.0999259458876498, -0.012763759757962725, ''),
Text(1.099936510446337, -0.011818332671541821, ''),
Text(1.0999462623755796, -0.010872896853775951, ''),
Text(1.0999552016681726, -0.009927453003149955, ''),
Text(1.099963328317512, -0.00898200181815462, ''),
Text(1.099970642317594, -0.008036543997286138, ''),
```

[24]
```
Text(1.0999917717262788, -0.004254660325165501, ''),
Text(1.0999950224110264, -0.0033091798025278096, ''),
Text(1.0999969271082544, -0.0026000677679425023, ''),
Text(1.0999979429457905, -0.0021273257930733737, ''),
Text(1.0999987556146094, -0.001654583425288917, ''),
Text(1.099999365114561, -0.0011818407519042597, ''),
Text(1.099999771445533, -0.0007090978602345863, ''),
Text(1.0999999746074502, -0.0002363548375951198 4, '')])
```



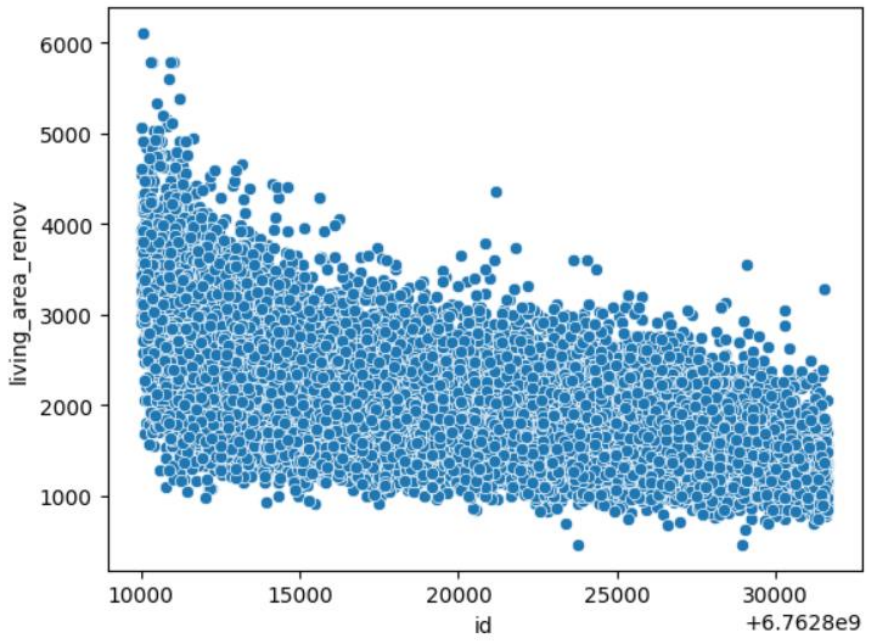
21BCE3095

>>Bivariante Analysis

21BCE3095

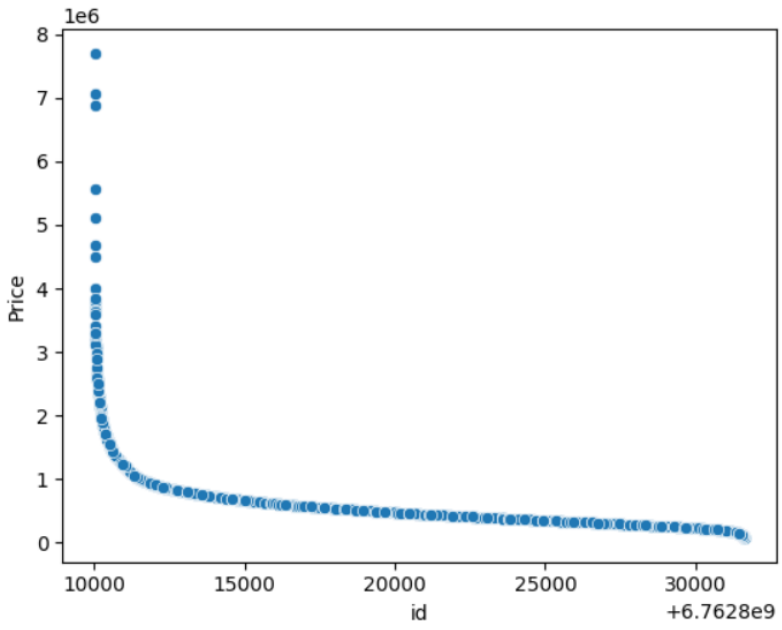
```
sns.scatterplot(x=df.id,y=df.living_area_renov)
```

```
<Axes: xlabel='id', ylabel='living_area_renov'>
```



```
[26] sns.scatterplot(x=df.id,y=df.Price)
```

```
<Axes: xlabel='id', ylabel='Price'>
```

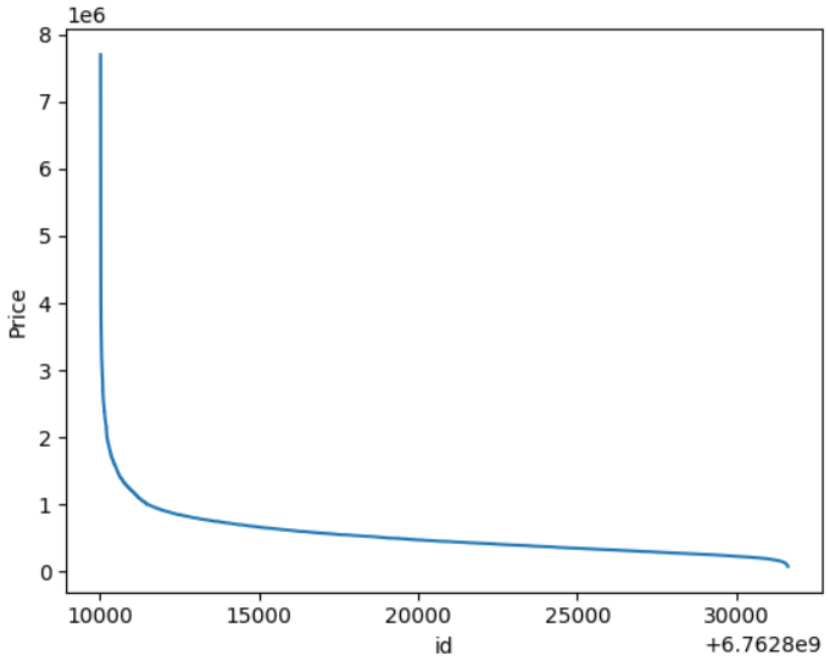


21BCE3095

21BCE3095

```
sns.lineplot(x=df.id,y=df.Price)
```

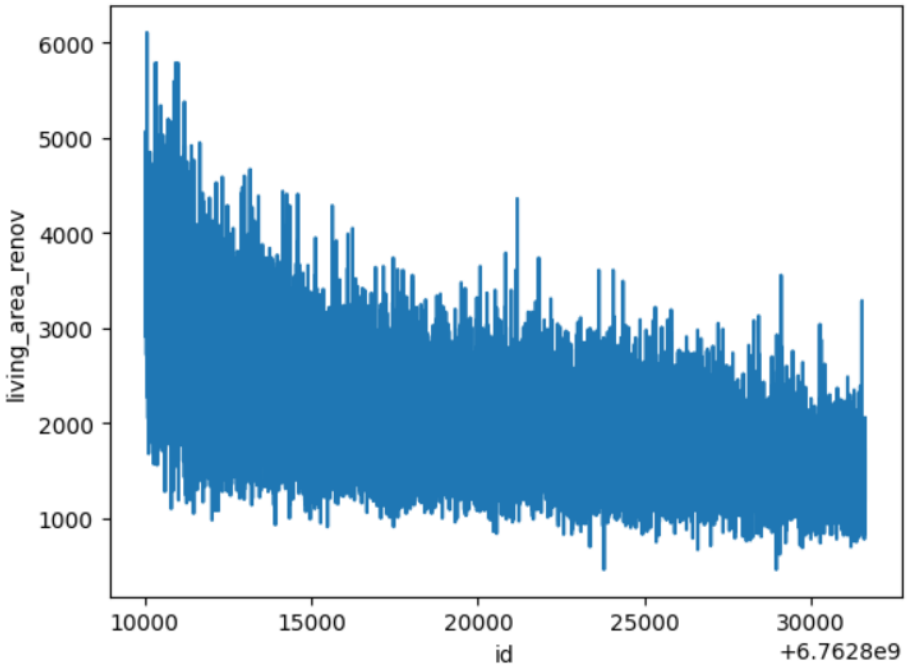
```
<Axes: xlabel='id', ylabel='Price'>
```



21BCE3095

```
sns.lineplot(x=df.id,y=df.living_area_renov)
```

```
<Axes: xlabel='id', ylabel='living_area_renov'>
```



>> Multivariante Analysis

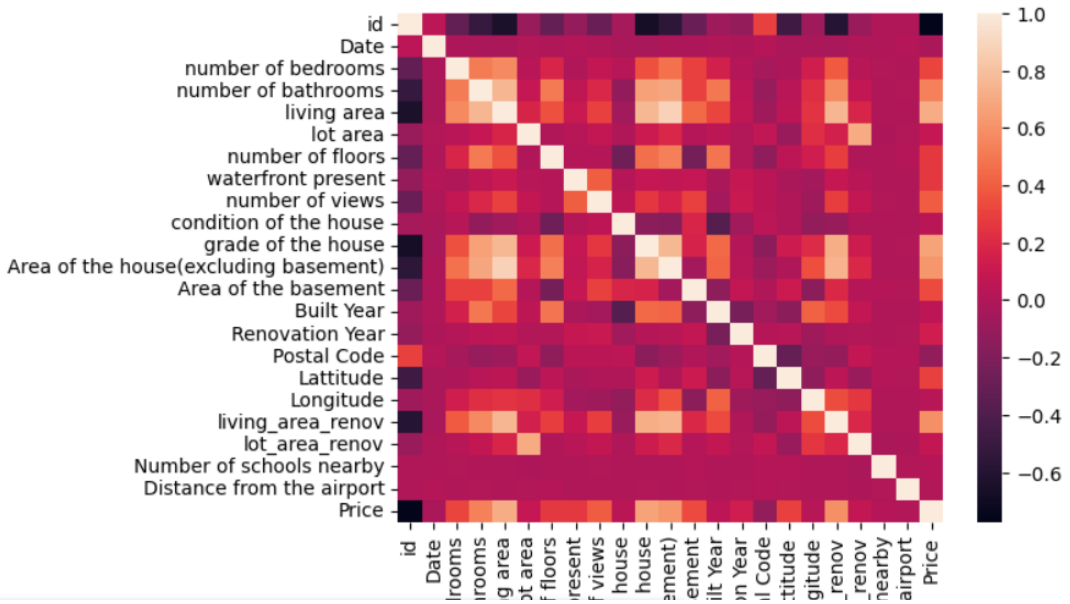
21BCE3095

```
df.corr()
```

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	...	Built Year	Renovation Year	Posta Cod
id	1.000000	0.045966	-0.329034	-0.516909	-0.648127	-0.100269	-0.312305	-0.112937	-0.293004	-0.045061	...	-0.068645	-0.109155	0.29470
Date	0.045966	1.000000	-0.015663	-0.026485	-0.021958	0.004392	-0.010335	0.012006	-0.004782	-0.027402	...	-0.005869	-0.011636	0.01824
number of bedrooms	-0.329034	-0.015663	1.000000	0.509784	0.570526	0.034416	0.177294	-0.006257	0.078665	0.026597	...	0.152954	0.016132	-0.04415
number of bathrooms	-0.516909	-0.026485	0.509784	1.000000	0.753517	0.080806	0.502924	0.060104	0.183789	-0.128232	...	0.498127	0.049669	-0.10554
living area	-0.648127	-0.021958	0.570526	0.753517	1.000000	0.174420	0.354743	0.105837	0.287728	-0.063358	...	0.309602	0.059400	-0.08030
lot area	-0.100269	0.004392	0.034416	0.080806	0.174420	1.000000	-0.004138	0.026282	0.078308	-0.008548	...	0.051615	0.006848	0.07013
number of floors	-0.312305	-0.010335	0.177294	0.502924	0.354743	-0.004138	1.000000	0.016316	0.020153	-0.269928	...	0.481565	0.006705	-0.12978
waterfront present	-0.112937	0.012006	-0.006257	0.060104	0.105837	0.026282	0.016316	1.000000	0.400206	0.018644	...	-0.024226	0.085865	0.03831
number of views	-0.293004	-0.004782	0.078665	0.183789	0.287728	0.078308	0.020153	0.400206	1.000000	0.052533	...	-0.055357	0.102944	0.03926
condition of the house	-0.045061	-0.027402	0.026597	-0.128232	-0.063358	-0.008548	-0.269928	0.018644	0.052533	1.000000	...	-0.381718	-0.062126	0.04533

```
[30] sns.heatmap(df.corr())
```

```
<Axes: >
```



3.Perform descriptive statistics on the dataset.

21BCE3095

```
df.isnull().any()
```

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

```
df.describe()
```

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	...	Built Year	Renovation Year
count	1.462000e+04	14620.000000	14620.000000	14620.000000	14620.000000	1.462000e+04	14620.000000	14620.000000	14620.000000	14620.000000	...	14620.000000	14620.000000
mean	6.762821e+09	42604.538646	3.379343	2.129583	2098.262996	1.509328e+04	1.502360	0.007661	0.233105	3.430506	...	1970.926402	90.924008
std	6.237575e+03	67.347991	0.938719	0.769934	928.275721	3.791962e+04	0.540239	0.087193	0.766259	0.664151	...	29.493625	416.216661
min	6.762810e+09	42491.000000	1.000000	0.500000	370.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000	...	1900.000000	0.000000
25%	6.762815e+09	42546.000000	3.000000	1.750000	1440.000000	5.010750e+03	1.000000	0.000000	0.000000	3.000000	...	1951.000000	0.000000
50%	6.762821e+09	42600.000000	3.000000	2.250000	1930.000000	7.620000e+03	1.500000	0.000000	0.000000	3.000000	...	1975.000000	0.000000
75%	6.762826e+09	42662.000000	4.000000	2.500000	2570.000000	1.080000e+04	2.000000	0.000000	0.000000	4.000000	...	1997.000000	0.000000
max	6.762832e+09	42734.000000	33.000000	8.000000	13540.000000	1.074218e+06	3.500000	1.000000	4.000000	5.000000	...	2015.000000	2015.000000

8 rows × 23 columns

21BCE3095

```
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
```