Sayandeep Chaudhuri 21BCE2501 sayandeep.chaudhuri2021@vitstudent.ac.in

```
import numpy as np
import pandas as pd
import sklearn.preprocessing as sk
import matplotlib.pyplot as plt
```

```
df=pd.read_csv('/content/House Price India.csv')
df
```

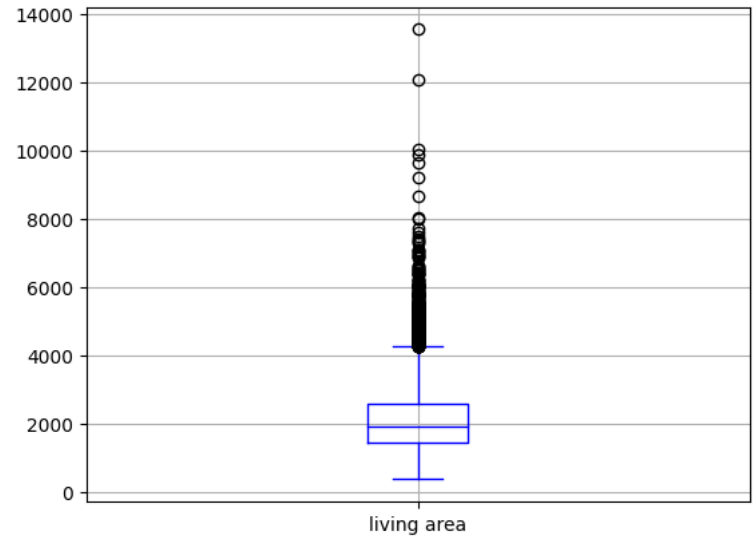
	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	co
0	6762810145	42491	5	2.50	3650	9050	2.0	0	4	
1	6762810635	42491	4	2.50	2920	4000	1.5	0	0	
2	6762810998	42491	5	2.75	2910	9480	1.5	0	0	
3	6762812605	42491	4	2.50	3310	42998	2.0	0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

14620 rows × 23 columns

Uni-variate analysis:

```
df.boxplot(column=['living area'],color='blue')
```

```
<Axes: >
```




```
import seaborn as sns
sns.distplot(df['living area'])
```

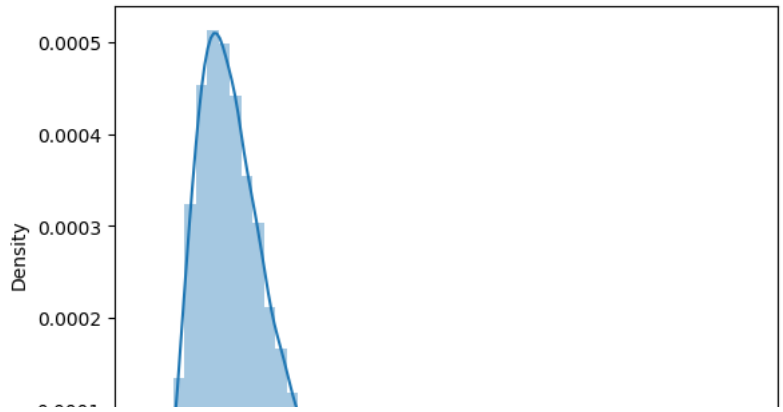
```
<ipython-input-5-03f4bd5ffec4>:2: UserWarning:

`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['living area'])
<Axes: xlabel='living area', ylabel='Density'>
```

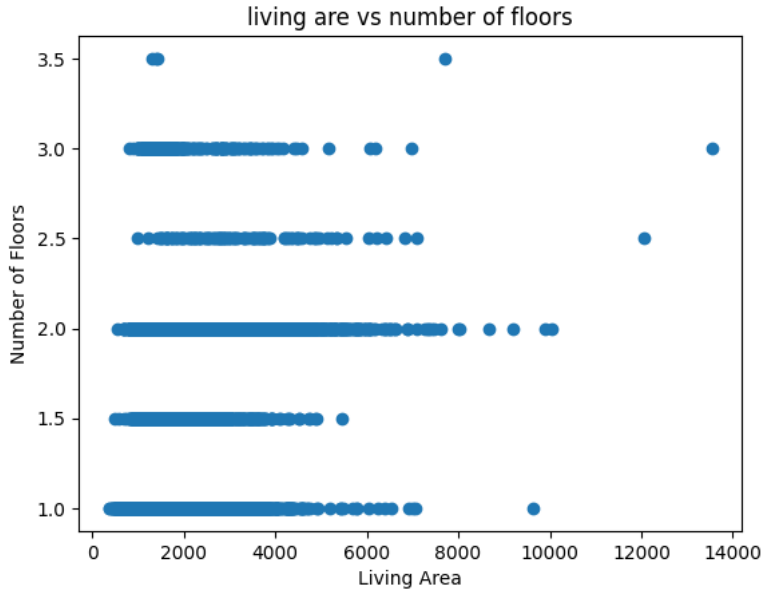


Double-click (or enter) to edit

Bi variate Analysis:

```
plt.scatter(df['living area'],df['number of floors'])
plt.title('living are vs number of floors')
plt.xlabel('Living Area')
plt.ylabel('Number of Floors')
```

```
Text(0, 0.5, 'Number of Floors')
```

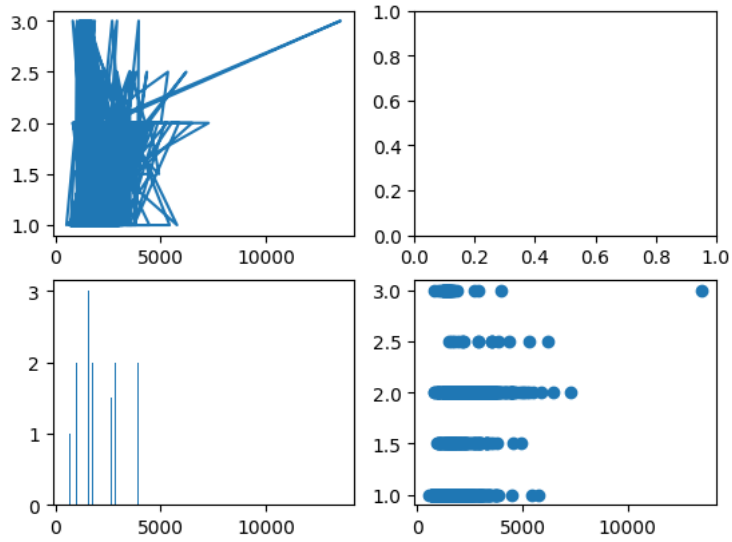



```
plt.scatter(df['living area'],df['number of floors'])
```

```
<matplotlib.collections.PathCollection at 0x7d1fc28d23b0>
```


```
x=df['living area'][:1000]
y=df['number of floors'][:1000]
fig,ax=plt.subplots(2,2)
ax[0,0].plot(x,y)
ax[1,0].bar(x,y)
ax[1,1].scatter(x,y)
```

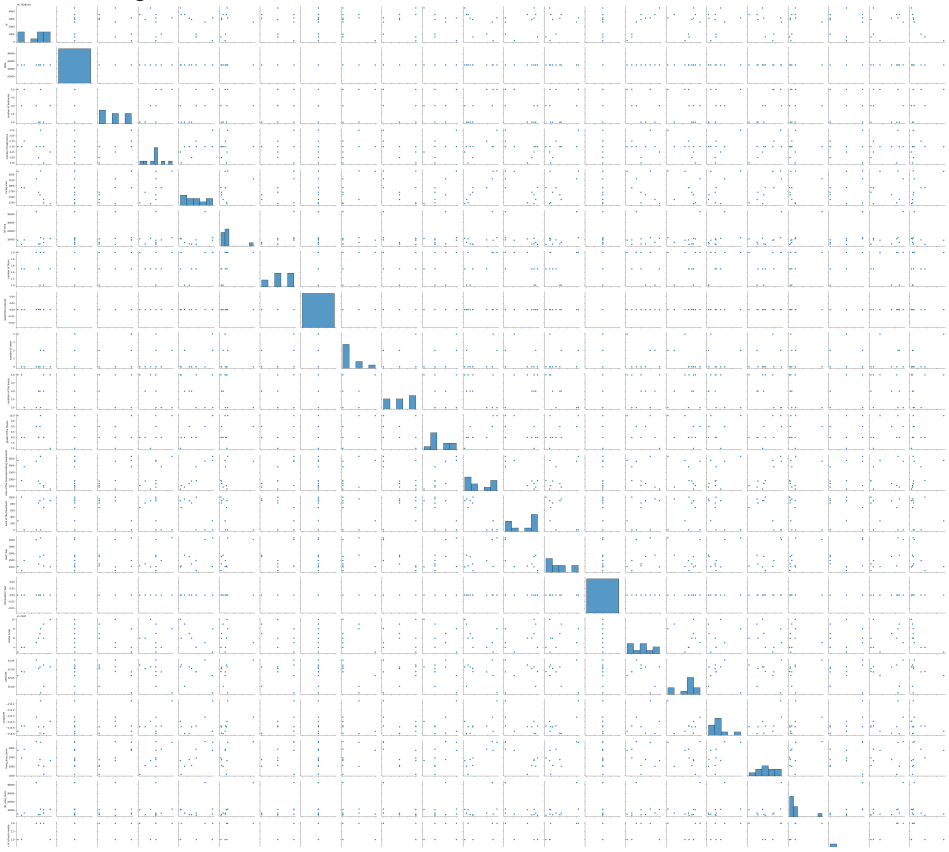
```
<matplotlib.collections.PathCollection at 0x7d1fc26031c0>
```



Multivariate analysis:

```
sns.pairplot(df.head(10))
```

```
<seaborn.axisgrid.PairGrid at 0x7d1fc2700400>
```



```
sns.heatmap(df.head(5).corr())
```

```
<Axes: >
```

Descriptive Analysis

```
df.describe()
```

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	n
count	1.462000e+04	14620.000000	14620.000000	14620.000000	14620.000000	1.462000e+04	1462
mean	6.762821e+09	42604.538646	3.379343	2.129583	2098.262996	1.509328e+04	
std	6.237575e+03	67.347991	0.938719	0.769934	928.275721	3.791962e+04	
min	6.762810e+09	42491.000000	1.000000	0.500000	370.000000	5.200000e+02	
25%	6.762815e+09	42546.000000	3.000000	1.750000	1440.000000	5.010750e+03	
50%	6.762821e+09	42600.000000	3.000000	2.250000	1930.000000	7.620000e+03	
75%	6.762826e+09	42662.000000	4.000000	2.500000	2570.000000	1.080000e+04	
max	6.762832e+09	42734.000000	33.000000	8.000000	13540.000000	1.074218e+06	

8 rows × 23 columns

Handling null values

```
df.fillna(method='pad')
```

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

14620 rows × 23 columns