

ASSIGNMENT - 1

Data Visualization

Assignment Date	21 April , 2023
Student Name	Balakumaran L.
Student Roll Number	420620104305
Maximum marks	2 Mark

1. Download the dataset:

<https://www.kaggle.com/datasets/mohamedafsal007/house-price-dataset-of-india?resource=download>

2. Load the dataset:

```
In [1]: import pandas as pd

In [2]: import matplotlib.pyplot as plt
```

Load the dataset

```
In [3]: #reading the csv data set

In [7]: dataset=pd.read_csv(r"C:\Users\indira\Downloads\archive.zip")

In [9]: dataset

Out[9]:
```

	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	Latitude	Longitude	ln
0	6782810145	42491	5	2.50	3650	9050	2.0	0	4	5	...	1921	0	122003	52.8645	-114.557	
1	6782810635	42491	4	2.50	2920	4000	1.5	0	0	5	...	1909	0	122004	52.8878	-114.470	
2	6782810998	42491	5	2.75	2910	9480	1.5	0	0	3	...	1939	0	122004	52.8852	-114.458	
3	6782812605	42491	4	2.50	3310	42998	2.0	0	0	3	...	2001	0	122005	52.9532	-114.321	
4	6782812919	42491	3	2.00	2710	4500	1.5	0	0	4	...	1929	0	122006	52.9047	-114.485	
...
14615	6782830250	42734	2	1.50	1558	20000	1.0	0	0	4	...	1957	0	122066	52.8191	-114.472	
14616	6782830339	42734	3	2.00	1680	7000	1.5	0	0	4	...	1998	0	122072	52.5075	-114.393	
14617	6782830618	42734	2	1.00	1070	6120	1.0	0	0	3	...	1982	0	122056	52.7289	-114.507	
14618	6782830709	42734	4	1.00	1030	6621	1.0	0	0	4	...	1955	0	122042	52.7157	-114.411	
14619	6782831463	42734	3	1.00	900	4770	1.0	0	0	3	...	1999	2009	122018	52.5338	-114.552	

14620 rows x 23 columns



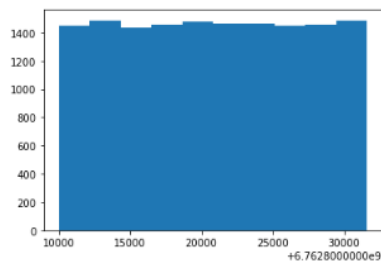
3.Perform the Below Visualizations:

Visualizations

Univariate Analysis

```
In [29]: plt.hist(dataset['id'])
```

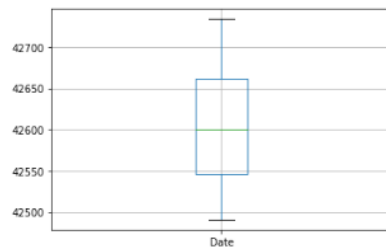
```
Out[29]: (array([1448., 1487., 1432., 1459., 1479., 1460., 1466., 1449., 1457.,  
1483.]),  
array([6.76281002e+09, 6.76281218e+09, 6.76281434e+09, 6.76281650e+09,  
6.76281866e+09, 6.76282082e+09, 6.76282298e+09, 6.76282514e+09,  
6.76282730e+09, 6.76282946e+09, 6.76283162e+09]),  
<BarContainer object of 10 artists>)
```



Bi variate Analysis

```
In [16]: dataset.boxplot(column="Date")
```

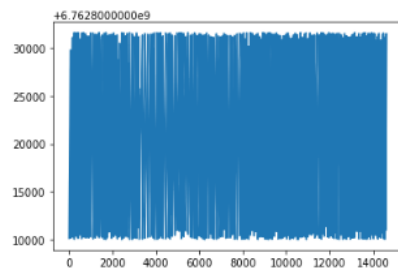
```
Out[16]: <AxesSubplot:>
```



Multi variate Analysis

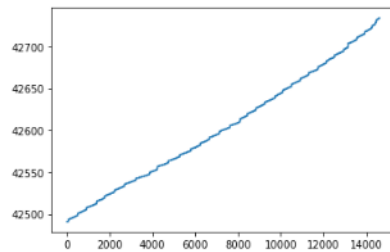
```
In [30]: plt.plot(dataset['id'])
```

```
Out[30]: [<matplotlib.lines.Line2D at 0x23594fd45e0>]
```



```
In [37]: plt.plot(dataset['Date'])
```

```
Out[37]: [ <matplotlib.lines.Line2D at 0x235951ec400> ]
```



4. Perform descriptive statistics on the dataset:

```
In [38]: dataset.describe()
```

```
Out[38]:
```

	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	...	
count	1.462000e+04	14620.000000	14620.000000	14620.000000	14620.000000	1.462000e+04	14620.000000	14620.000000	14620.000000	14620.000000	...	14620
mean	6.762821e+09	42604.538646	3.379343	2.129583	2098.262906	1.509328e+04	1.502380	0.007981	0.233105	3.430506	...	1971
std	6.237575e+03	87.347991	0.938719	0.769934	928.275721	3.791982e+04	0.540239	0.087193	0.796259	0.984151	...	21
min	6.762810e+09	42491.000000	1.000000	0.500000	370.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000	...	1901
25%	6.762815e+09	42546.000000	3.000000	1.750000	1440.000000	5.010750e+03	1.000000	0.000000	0.000000	3.000000	...	195
50%	6.762821e+09	42600.000000	3.000000	2.250000	1930.000000	7.620000e+03	1.500000	0.000000	0.000000	3.000000	...	1971
75%	6.762826e+09	42662.000000	4.000000	2.500000	2570.000000	1.080000e+04	2.000000	0.000000	0.000000	4.000000	...	1991
max	6.762832e+09	42734.000000	33.000000	8.000000	13540.000000	1.074218e+06	3.500000	1.000000	4.000000	5.000000	...	2011

8 rows x 23 columns

```
In [39]: dataset["id"].mean()
```

```
Out[39]: 6762820830.52565
```

```
In [41]: dataset["number of views"].median()
```

```
Out[41]: 0.0
```

```
In [44]: dataset["Price"].mode()
```

```
Out[44]: 0 450000  
Name: Price, dtype: int64
```

5. Handle the Missing values:

Handle the missing values

```
In [45]: dataset.isnull().any()
```

```
Out[45]:
```

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
Latitude	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	



```
In [46]: dataset.isnull().sum()
```

```
Out[46]: 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
Latitude              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
```

```
In [47]: dataset.skew()
```

```
Out[47]: id                -0.000802
Date                0.143747
number of bedrooms   2.663257
number of bathrooms  0.556663
living area          1.538337
lot area            10.155206
number of floors      0.586158
waterfront present   11.294672
number of views       3.409219
condition of the house 1.018018
grade of the house    0.777584
Area of the house(excluding basement) 1.436446
Area of the basement  1.609744
Built Year           -0.472049
Renovation Year       4.359764
Postal Code           0.227735
Latitude             -0.523831
Longitude             0.873803
living_area_renov     1.081959
lot_area_renov        7.774206
Number of schools nearby -0.022519
Distance from the airport 0.006114
Price                4.269298
dtype: float64
```

```
In [ ]:
```



