**ASSIGNMENT-2**

**G PAVAN KUMAR**

**21BCE9495**

**VIT-AP UNIVERSITY**

download datset

```
!pip install opendatasets
```

```
Collecting opendatasets
  Downloading opendatasets-0.1.22-py3-none-any.whl (15 kB)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from opendatasets) (4.66.1)
Requirement already satisfied: kaggle in /usr/local/lib/python3.10/dist-packages (from opendatasets) (1.5.16)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from opendatasets) (8.1.7)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (1.16.0)
Requirement already satisfied: certifi in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (2023.7.22)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (2.8.2)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (2.31.0)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (8.0.1)
Requirement already satisfied: urllib3 in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (2.0.4)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (6.0.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->kaggle->opendatasets) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in /usr/local/lib/python3.10/dist-packages (from python-slugify->kaggle->opendat
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle->opendata
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle->opendatasets) (3.4)
Installing collected packages: opendatasets
Successfully installed opendatasets-0.1.22
Requirement already satisfied: opendatasets in /usr/local/lib/python3.10/dist-packages (0.1.22)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from opendatasets) (4.66.1)
Requirement already satisfied: kaggle in /usr/local/lib/python3.10/dist-packages (from opendatasets) (1.5.16)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from opendatasets) (8.1.7)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (1.16.0)
Requirement already satisfied: certifi in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (2023.7.22)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (2.8.2)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (2.31.0)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (8.0.1)
Requirement already satisfied: urllib3 in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (2.0.4)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from kaggle->opendatasets) (6.0.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->kaggle->opendatasets) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in /usr/local/lib/python3.10/dist-packages (from python-slugify->kaggle->opendat
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle->opendata
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle->opendatasets) (3.4)
```

```
import pandas as pd
import opendatasets as od
od.download("https://www.kaggle.com/datasets/mohamedafsal007/house-price-dataset-of-india")
```

```
Please provide your Kaggle credentials to download this dataset. Learn more: http://bit.ly/kaggle-creds
Your Kaggle username: gogulapavan
Your Kaggle Key: ··········
Downloading house-price-dataset-of-india.zip to ./house-price-dataset-of-india
100%|██████████| 480k/480k [00:00<00:00, 101MB/s]
```

Load dataset

```
df=pd.read_csv("/content/house-price-dataset-of-india/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 | ... | 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.shape
```

```
(14620, 23)
```

```
df["Built Year"].isnull()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
14615    False
14616    False
14617    False
14618    False
14619    False
Name: Built Year, Length: 14620, dtype: bool
```

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

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-1-629914d6d1c7> in <cell line: 1>()
----> 1 df.isnull().sum()

NameError: name 'df' is not defined
```

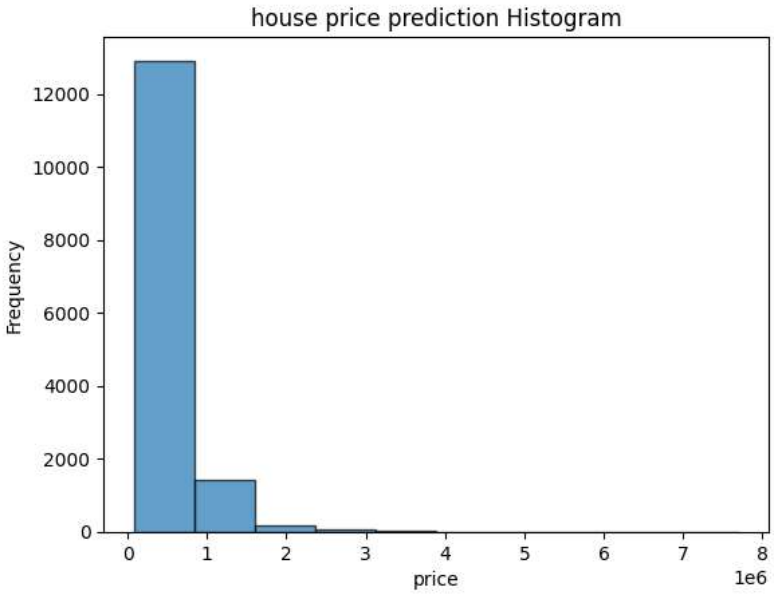
SEARCH STACK OVERFLOW

Double-click (or enter) to edit

UNIVARIATE ANALYSIS

```
import matplotlib.pyplot as plt
import numpy as np
from matplotlib import rcParams
import seaborn as sns

plt.hist(df["lot_area_renov"], bins=10, edgecolor='k', alpha=0.7)
plt.xlabel('number of bedrooms')
plt.ylabel('Frequency')
plt.title('house price prediction Histogram')

plt.show()
```


```
plt.hist(df["Price"], bins=10, edgecolor='k', alpha=0.7)
plt.xlabel('price')
plt.ylabel('Frequency')
plt.title('house price prediction Histogram')

plt.show()
```




```
df['Price'].describe()
```

```
count    1.462000e+04
mean     5.389322e+05
std      3.675324e+05
min      7.800000e+04
25%      3.200000e+05
50%      4.500000e+05
75%      6.450000e+05
max      7.700000e+06
Name: Price, dtype: float64
```

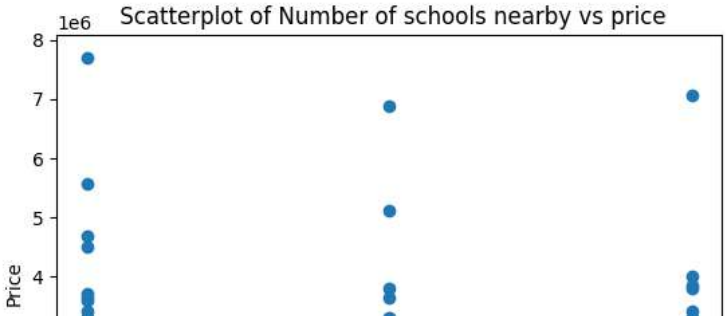
```
df['lot_area_renov'].describe()
```

```
count     14620.000000
mean      12753.500068
std       26058.414467
min         651.000000
25%        5097.750000
50%        7620.000000
75%       10125.000000
max      560617.000000
Name: lot_area_renov, dtype: float64
```

**Bi variate analysis**

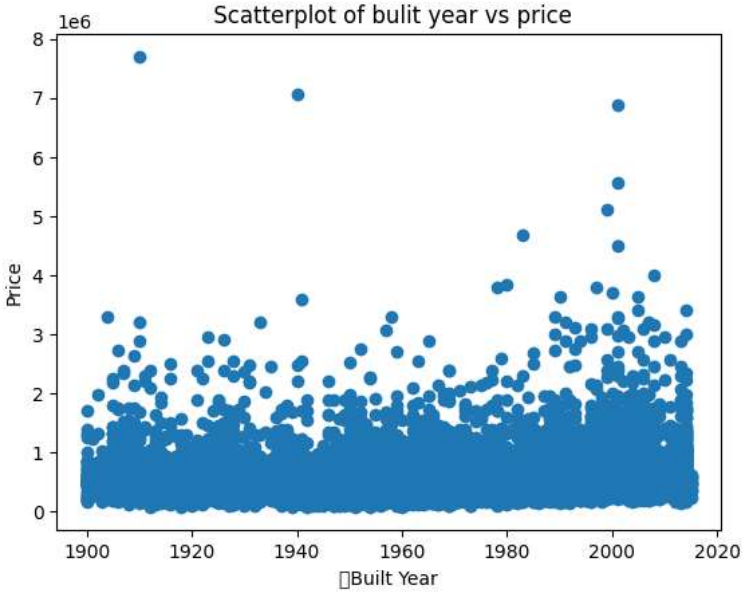
**Scatter plot**

```
plt.scatter(x=df["Number of schools nearby"],y=df["Price"])
plt.xlabel('Number of schools nearby')
plt.ylabel('Price')
plt.title(' Scatterplot of Number of schools nearby vs price')

plt.show()
```

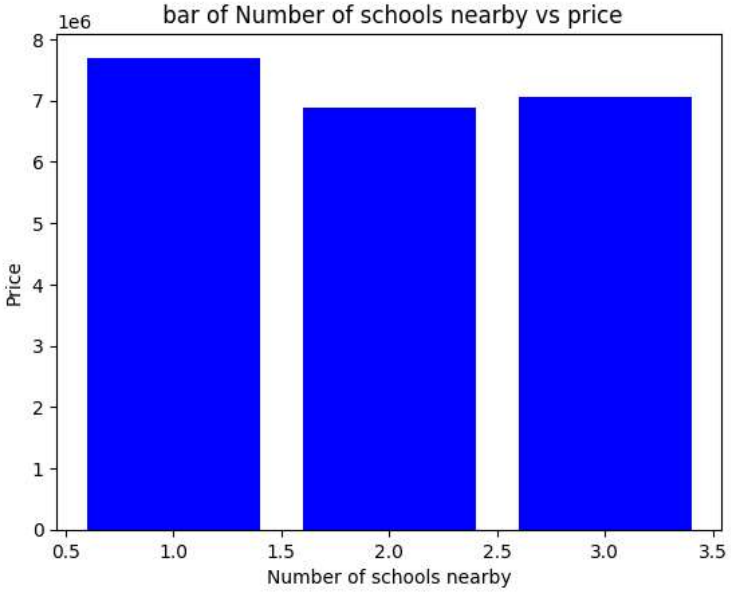



```
plt.scatter(x=df["Built Year"],y=df["Price"])
plt.xlabel('    Built Year')
plt.ylabel('Price')
plt.title(' Scatterplot of bulit year vs price')

plt.show()
```

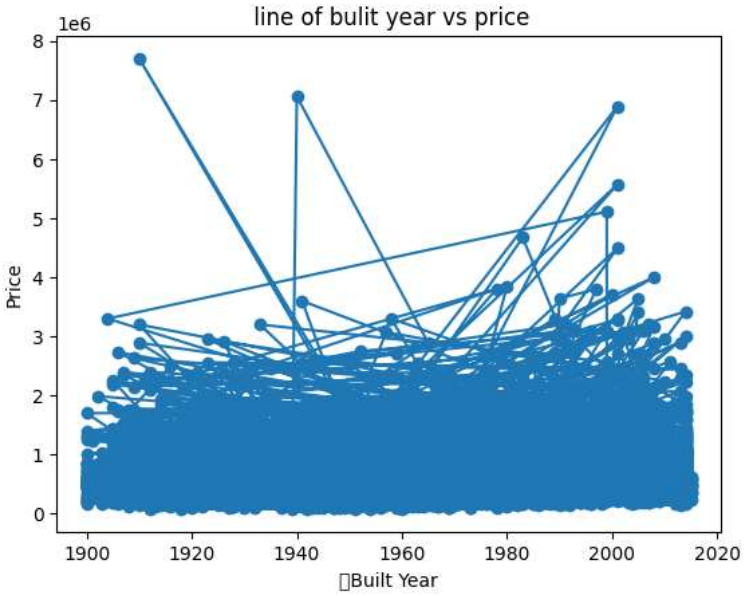
```
/usr/local/lib/python3.10/dist-packages/IPython/core/pylabtools.py:151: UserWarning: Glyph 9 (   ) missi
  fig.canvas.print_figure(bytes_io, **kw)
```




```
plt.bar(df["Number of schools nearby"],df["Price"],color="blue")
plt.xlabel('Number of schools nearby')
plt.ylabel('Price')
plt.title(' bar of Number of schools nearby vs price')

plt.show()
```



```
plt.plot(df["Built Year"],df["Price"],marker='o', linestyle='-')
plt.xlabel('    Built Year')
plt.ylabel('Price')
plt.title(' line of bulit year vs price')

plt.show()
```




MULTIVARIATE ANALYSIS:

```
sns.pairplot(df)
```

```
df.describe()
```

| | 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 |
| **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

HANDLING MISSING VALUES:

```
df.isnull()
```

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | False | False | False | False | False | False | False | False | False | False | ... | False |
| **1** | False | False | False | False | False | False | False | False | False | False | ... | False |
| **2** | False | False | False | False | False | False | False | False | False | False | ... | False |
| **3** | False | False | False | False | False | False | False | False | False | False | ... | False |
| **4** | False | False | False | False | False | False | False | False | False | False | ... | False |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

```
df.dropna()
```

| | 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 | . |

14620 rows × 23 columns

```
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
```

```
df['Built Year'].isnull()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
14615    False
14616    False
14617    False
14618    False
```

```
14619    False
Name: Built Year, Length: 14620, dtype: bool
```