import pandas as pd import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

path=" /content/House Price India.csv" df=pd.read\_csv(path)

# load the dataset

df.info()

df.head()

 <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

## id Date

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **bedrooms** |  |  |  | **floors** |  | **views** |
| **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 |

**number**

**of**

**number of bathrooms**

**living area**

**lot area**

**number**

**of**

**waterfront**

**present**

**number**

**of**

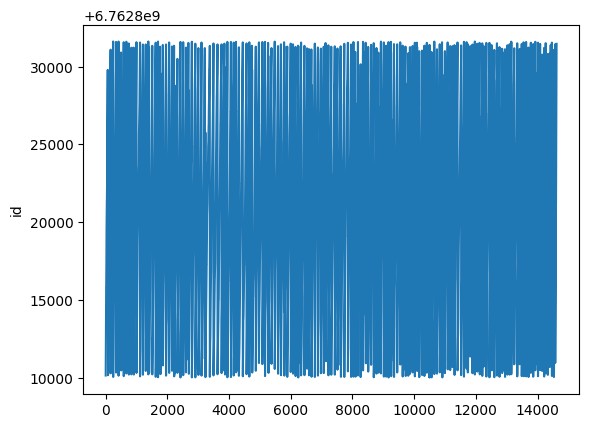
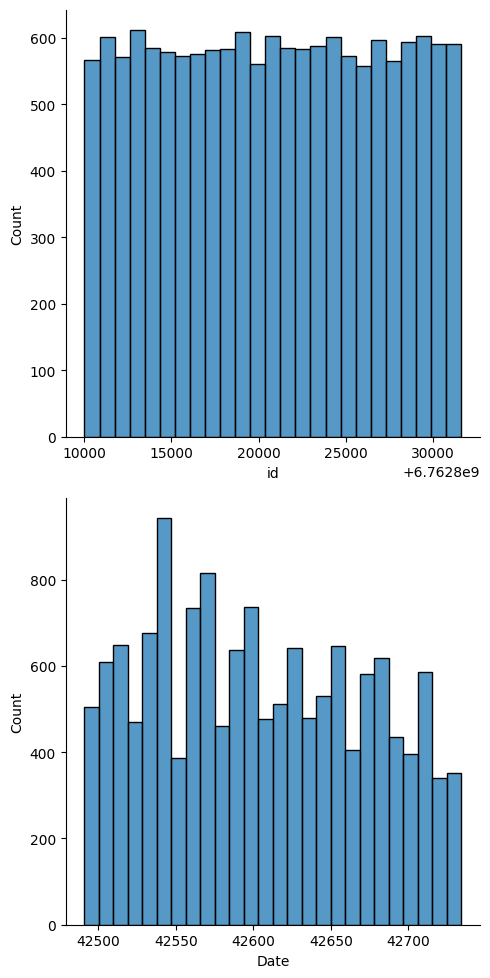
5 rows × 23 columns

# Univariate

sns.displot(df.id)

sns.displot(df.Date)

<seaborn.axisgrid.FacetGrid at 0x7fe53944bc40>



# Bi-variate

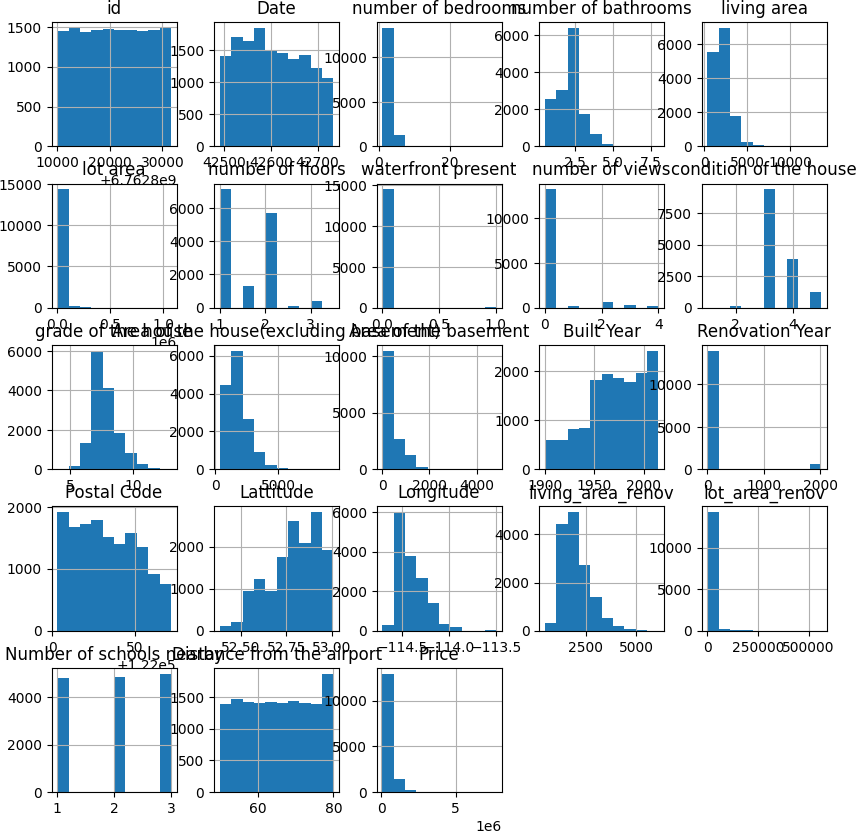
sns.lineplot(df.id)

<Axes: ylabel='id'>

# Multivariate

df.hist(figsize=(10,10))

array([[<Axes: title={'center': 'id'}>, <Axes: title={'center': 'Date'}>,



<Axes: title={'center': 'number of bedrooms'}>,

<Axes: title={'center': 'number of bathrooms'}>,

<Axes: title={'center': 'living area'}>], [<Axes: title={'center': 'lot area'}>,

<Axes: title={'center': 'number of floors'}>,

<Axes: title={'center': 'waterfront present'}>,

<Axes: title={'center': 'number of views'}>,

<Axes: title={'center': 'condition of the house'}>], [<Axes: title={'center': 'grade of the house'}>,

<Axes: title={'center': 'Area of the house(excluding basement)'}>,

<Axes: title={'center': 'Area of the basement'}>,

<Axes: title={'center': 'Built Year'}>,

<Axes: title={'center': 'Renovation Year'}>], [<Axes: title={'center': 'Postal Code'}>,

<Axes: title={'center': 'Lattitude'}>,

<Axes: title={'center': 'Longitude'}>,

<Axes: title={'center': 'living\_area\_renov'}>,

<Axes: title={'center': 'lot\_area\_renov'}>],

[<Axes: title={'center': 'Number of schools nearby'}>,

<Axes: title={'center': 'Distance from the airport'}>,

<Axes: title={'center': 'Price'}>, <Axes: >, <Axes: >]], dtype=object)

# Perform Descriptive Statistics on the Dataset

df.mean()

df.median()

norm\_df=pd.DataFrame(np.random.normal

(size=100000))

norm\_df.plot(kind="density",

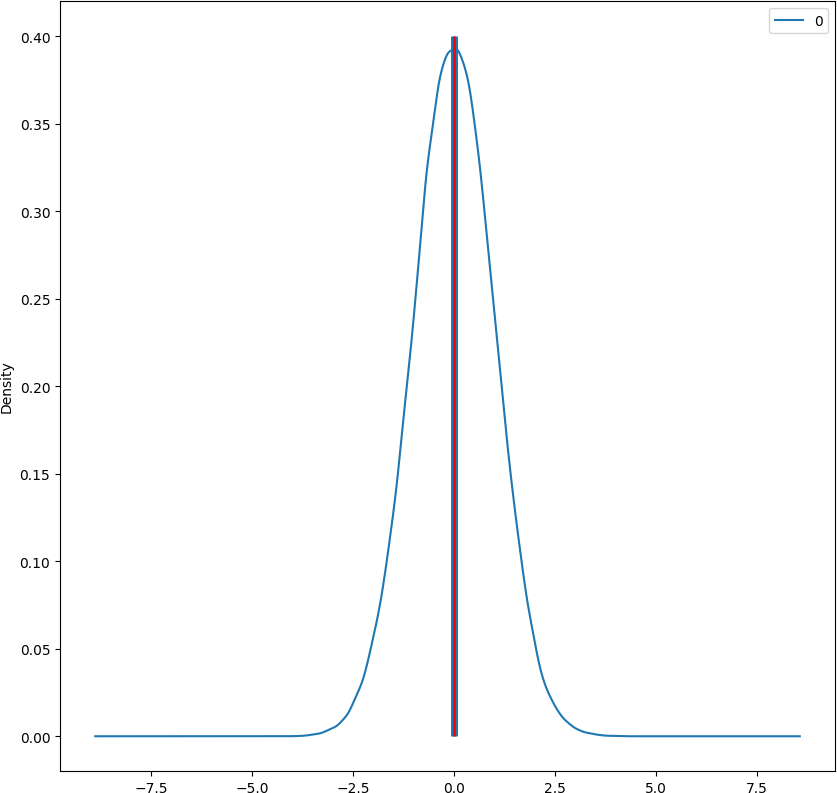
figsize=(10,10)); plt.vlines(norm\_df.mean(),

ymin=0, ymax=0.4,

linewidth=5.0); plt.vlines(norm\_df.median(),

ymin=0, ymax=0.4,

linewidth =2.0, color="red");



Handle the Missing value

df=pd.DataFrame(df) df.isnull()

**id Date**

**number**

**of**

**number of living lot**

**bathrooms**

**area area**

**number**

**of**

**waterfront**

**present**

**number**

**of**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | **bedrooms** |  |  |  | **floors** |  | **views** |
| **0** | False | False | False | False | False | False | False | False | False |
| **1** | False | False | False | False | False | False | False | False | False |
| **2** | False | False | False | False | False | False | False | False | False |
| **3** | False | False | False | False | False | False | False | False | False |
| **4** | False | False | False | False | False | False | False | False | False |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **14615** | False | False | False | False | False | False | False | False | False |
| **14616** | False | False | False | False | False | False | False | False | False |
| **14617** | False | False | False | False | False | False | False | False | False |
| **14618** | False | False | False | False | False | False | False | False | False |
| **14619** | False | False | False | False | False | False | False | False | False |

14620 rows × 23 columns

[Colab paid products](https://colab.research.google.com/signup?utm_source=footer&utm_medium=link&utm_campaign=footer_links) - [Cancel contracts here](https://colab.research.google.com/cancel-subscription)