from google.colab import files

uploaded = files.upload()



Choose Files House Price India.csv.zip

House Price India.csv.zip(application/x-zip-compressed) - 491826 bytes, last modified: 4/9/2025 - 100% done

Saving House Price India csv zip to House Price India csv zip

import zipfile
import os

with zipfile.ZipFile("House Price India.csv.zip", 'r') as zip\_ref: zip\_ref.extractall("house\_price\_data")

os.listdir("house\_price\_data")

→ ['House Price India.csv']

import pandas as pd

df = pd.read\_csv("house\_price\_data/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
<b>0</b> 6762810145	42491	5	2.50	3650	9050	2.0	0	4
<b>1</b> 6762810635	42491	4	2.50	2920	4000	1.5	0	0
<b>2</b> 6762810998	42491	5	2.75	2910	9480	1.5	0	0
<b>3</b> 6762812605	42491	4	2.50	3310	42998	2.0	0	0
<b>4</b> 6762812919	42491	3	2.00	2710	4500	1.5	0	0

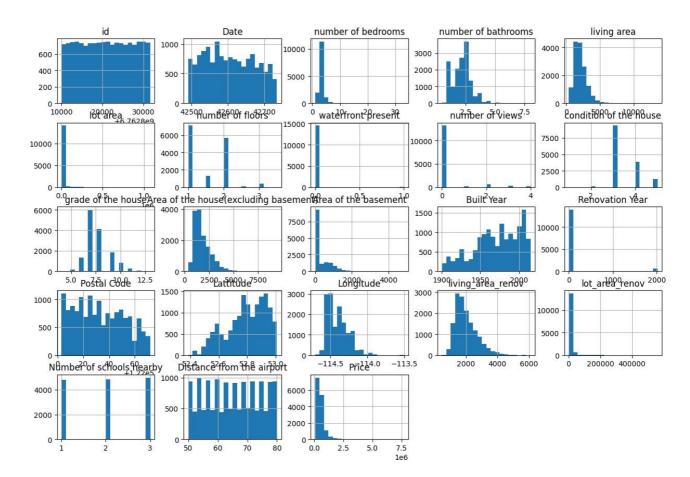
5 rows × 23 columns



import seaborn as sns
import matplotlib.pyplot as plt

```
df.select_dtypes(include='number').hist(bins=20, figsize=(15,10))
plt.suptitle("Univariate Analysis - Histograms")
plt.show()
```

## Univariate Analysis - Histograms



```
categorical_cols = df.select_dtypes(include='object').columns
for col in categorical_cols:
    plt.figure(figsize=(6,4))
    sns.countplot(data=df, x=col)
```

```
plt.title(f'Univariate - {col}')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

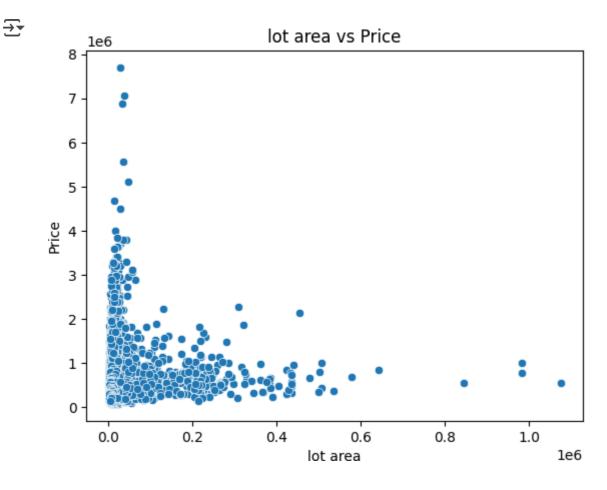
plt.figure(figsize=(10,8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Multivariate Analysis - Correlation Heatmap")
plt.show()
```



```
Multivariate Analysis - Correlation Heatmap
                                                                                                                                                    1.0
                                       id -1.000.050.3±0.5±0.6±0.1±0.3±0.1±0.2±0.0±0.6±0.5±0.2±0.0±0.1±0.2±0.4±0.0±0.6±0.0±0.0±0.0±0.7
                                    Date -0.051.000.020.030.020.060.010.010.060.030.030.020.020.010.010.020.020.020.030.060.060.010.03
                   number of bedrooms -0.330.021.000.510.570.030.180.010.080.030.350.470.300.150.020.040.010.140.390.030.090.010.31
                                                                                                                                                   - 0.8
                  number of bathrooms - 0.5-2.0.3.511.000.750.080.500.060.180.130.660.660.290.500.050.110.030.220.570.080.000.010.53
                              living area - 0.650.020.570.751.000.170.350.110.290.000.760.850.440.310.060.080.050.240.760.180.000.000.71
                                                                                                                                                    0.6
                                 lot area -0.100.000.030.080.171.000.000.030.080.010.110.180.020.050.010.070.090.220.150.7140.010.000.08
                        number of floors -0.310.010.180.500.350.001.000.020.020.020.270.460.530.240.480.010.130.050.130.290.010.010.020.26
                     waterfront present -0.1D.010.0D.060.110.030.021.000.400.020.080.070.090.020.090.040.020.090.030.000.000.000.26
                                                                                                                                                    - 0.4
                       number of views -0.290.000.080.180.290.080.020.401.000.050.250.160.290.060.100.040.060.080.280.070.010.000.40
                 condition of the house -0.090.030.030.130.060.030.270.020.051.000.150.170.180.350.060.050.060.120.160.060.040.000.04
                     grade of the house -0.6 0.030.350.660.7 0.110.460.080.250.151.000.7 0.170.440.010.150.120.200.7 20.120.000.000.
Area of the house(excluding basement) -0.550.020.470.680.880.180.530.070.160.170.761.000.050.420.030.080.000.350.740.190.000.000
                  Area of the basement -0.290.020.300.290.440.020.240.090.290.180.170.050.010.140.080.010.110.150.200.010.010.010.000.33
                                                                                                                                                   - 0.0
                               Built Year -0.070.010.150.500.310.050.480.020.060.3 0.440.420.141.000.230.060.140.410.330.070.060.000.05
                        Renovation Year -0.130.010.020.050.060.010.010.090.190.060.010.030.080.231.000.020.030.080.000.010.000.010.13
                             Postal Code -0.290.020.040.110.080.070.130.040.040.050.150.080.010.060.021.000.310.100.110.080.010.010.12
                                                                                                                                                    -0.2
                               Lattitude -0.440.020.010.030.050.090.050.020.060.000.120.000.110.140.030.311.000.130.050.090.010.010.30
                              Longitude -0.070.020.140.220.240.220.130.050.080.120.200.350.150.410.080.160.131.000.340.260.010.000.02
                                                                                                                                                    -0.4
                       living_area_renov -0.640.030.390.570.760.150.290.090.280.100.720.740.200.330.060.110.050.341.000.190.060.010
                         lot_area_renov -0.090.000.030.080.18077-0.010.030.070.000.120.190.010.070.010.080.090.260.191.000.030.010.080
             -0.6
              Price -0.770.030.310.530.710.080.260.260.400.040.670.620.330.050.130.120.300.020.550.080.010.001.00
                                                               lot area
                                                    number of bedrooms
                                                       number of bathrooms
                                                           living area
                                                                        waterfront present
                                                                           number of views
                                                                                    grade of the house
                                                                                            Area of the basement
                                                                                                        Postal Code
                                                                                                             Lattitude
                                                                                                                Longitude
                                                                    number of floors
                                                                                condition of the house
                                                                                        Area of the house(excluding basement)
                                                                                                Built Year
                                                                                                    Renovation Year
                                                                                                                     living_area_renov
                                                                                                                             Number of schools nearby
                                                                                                                         lot_area_renov
                                                                                                                                 Distance from the airpor
```

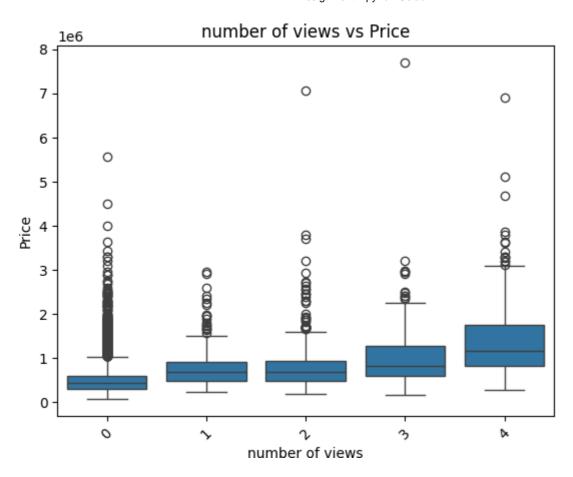
```
if 'lot area' in df.columns and 'Price' in df.columns:
    sns.scatterplot(data=df, x='lot area', y='Price')
    plt.title('lot area vs Price')
```

```
plt.show()
else:
   print("Check: Columns names 'lot area' or 'Price' do not exist")
```



```
if 'number of views' in df.columns and 'Price' in df.columns:
    sns.boxplot(data=df, x='number of views', y='Price')
    plt.title('number of views vs Price')
    plt.xticks(rotation=45)
    plt.show()
else:
    print("Check: Column names 'number of views' or 'Price' do not exist")
```





desc\_stats = df.describe(include='all')
print(desc\_stats)

$\overline{\mathbf{T}}$		id	Date	number of bedroom	ms number of bathrooms	\	
	count	1.462000e+04	14620.000000	14620.00000	14620.000000		
	mean	6.762821e+09	42604.538646	3.37934	13 2.129583		
	std	6.237575e+03	67.347991	0.93871	0.769934		
	min 6.762810e+09 25% 6.762815e+09		42491.000000	1.00000	0.500000		
			42546.000000	3.00000	1.750000		
	50%	6.762821e+09	42600.000000	3.00000	2.250000		
	75%	6.762826e+09	42662.000000	4.00000	2.500000		
	max	6.762832e+09	42734.000000	33.00000	8.00000		
		living area	lot area	number of floors	waterfront present \		
	count	14620.000000	1.462000e+04	14620.000000	14620.000000		
	mean	2098.262996	1.509328e+04	1.502360	0.007661		
	std	928.275721	3.791962e+04	0.540239	0.087193		
	min	370.000000	5.200000e+02	1.000000	0.000000		
	25%	1440.000000	5.010750e+03	1.000000	0.000000		
	50%	1930.000000	7.620000e+03	1.500000	0.000000		
	75%	2570.000000	1.080000e+04	2.000000	0.000000		
	max 13540.000000		1.074218e+06	3.500000	1.000000		
		number of vie	ws condition	of the house	Built Year \		
	count	14620.0000	00	14620.000000	14620.000000		
	mean	0.2331	05	3.430506	1970.926402		
	std	0.7662	59	0.664151	29.493625		
	min	0.0000	00	1.000000	00000 1900.000000		
	25% 0.000000			3.000000	1951.000000		

```
1975.000000
50%
               0.000000
                                         3.000000
75%
               0.000000
                                         4.000000
                                                          1997.000000
max
               4.000000
                                         5.000000
                                                          2015.000000
       Renovation Year
                            Postal Code
                                             Lattitude
                                                             Longitude
          14620.000000
                                                         14620.000000
                           14620.000000
                                          14620.000000
count
              90.924008
                          122033.062244
                                             52.792848
                                                           -114.404007
mean
std
             416.216661
                              19.082418
                                              0.137522
                                                              0.141326
min
               0.000000
                          122003.000000
                                             52.385900
                                                          -114.709000
25%
               0.000000
                          122017.000000
                                             52.707600
                                                           -114.519000
50%
               0.000000
                          122032.000000
                                             52.806400
                                                          -114.421000
75%
               0.000000
                          122048.000000
                                             52.908900
                                                           -114.315000
            2015.000000
                          122072.000000
                                             53.007600
                                                           -113.505000
max
                                             Number of schools nearby
       living area renov
                            lot area renov
             14620.000000
                              14620.000000
                                                          14620.000000
count
              1996.702257
                              12753.500068
mean
                                                               2.012244
std
               691.093366
                              26058.414467
                                                               0.817284
               460.000000
                                                               1.000000
min
                                651.000000
25%
              1490,000000
                               5097,750000
                                                               1,000000
50%
              1850.000000
                               7620.000000
                                                               2.000000
75%
                              10125.000000
              2380.000000
                                                               3.000000
              6110.000000
                             560617.000000
                                                               3.000000
max
       Distance from the airport
                                            Price
count
                      14620.000000
                                     1.462000e+04
mean
                         64.950958
                                     5.389322e+05
std
                                     3.675324e+05
                          8.936008
                         50.000000
                                     7.800000e+04
min
25%
                         57.000000
                                     3.200000e+05
50%
                                     4.500000e+05
                         65.000000
75%
                         73.000000
                                     6.450000e+05
```

missing\_values = df.isnull().sum()
print("Missing values:\n", missing\_values)

## → Missing values:

O .	
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

https://colab.research.google.com/drive/1AjBbfymRS4rMaDJQD0gDyaoaowx8CR3d?usp=sharing