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 zin to House Price India csv zin

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()

 $\overline{\mathbf{T}}$

id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views
6762810145	42491	5	2.50	3650	9050	2.0	0	4
6762810635	42491	4	2.50	2920	4000	1.5	0	0
6762810998	42491	5	2.75	2910	9480	1.5	0	0
6762812605	42491	4	2.50	3310	42998	2.0	0	0
6762812919	42491	3	2.00	2710	4500	1.5	0	0
	6762810145 6762810635 6762810998 6762812605	6762810145 42491 6762810635 42491 6762810998 42491 6762812605 42491	id Date bedrooms 6762810145 42491 5 6762810635 42491 4 6762810998 42491 5 6762812605 42491 4	id Date bedrooms of bedrooms number of bathrooms 6762810145 42491 5 2.50 6762810635 42491 4 2.50 6762810998 42491 5 2.75 6762812605 42491 4 2.50	id Date bedrooms of bedrooms number of bathrooms living area 6762810145 42491 5 2.50 3650 6762810635 42491 4 2.50 2920 6762810998 42491 5 2.75 2910 6762812605 42491 4 2.50 3310	id Date bedrooms of bedrooms number of bathrooms living area lot area 6762810145 42491 5 2.50 3650 9050 6762810635 42491 4 2.50 2920 4000 6762810998 42491 5 2.75 2910 9480 6762812605 42491 4 2.50 3310 42998	id Date of bedrooms number of bathrooms living area lot area of floors 6762810145 42491 5 2.50 3650 9050 2.0 6762810635 42491 4 2.50 2920 4000 1.5 6762810998 42491 5 2.75 2910 9480 1.5 6762812605 42491 4 2.50 3310 42998 2.0	id Date of bedrooms number of bathrooms living area lot area of floors waterfront present 6762810145 42491 5 2.50 3650 9050 2.0 0 6762810635 42491 4 2.50 2920 4000 1.5 0 6762810998 42491 5 2.75 2910 9480 1.5 0 6762812605 42491 4 2.50 3310 42998 2.0 0

5 rows × 23 columns

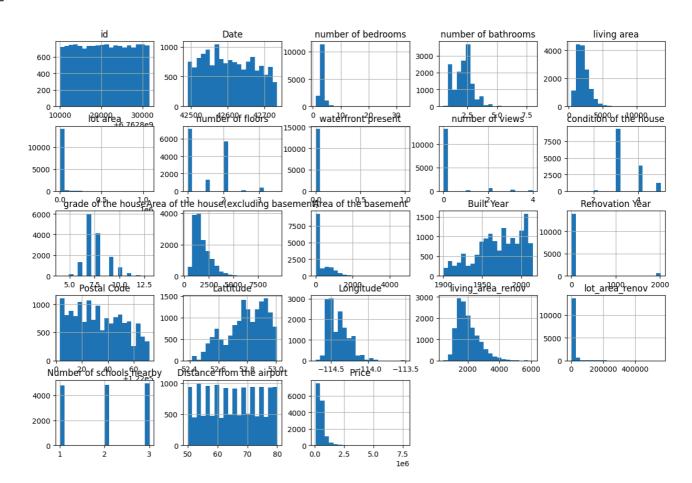


import seaborn as sns
import matplotlib.pyplot as plt

```
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()

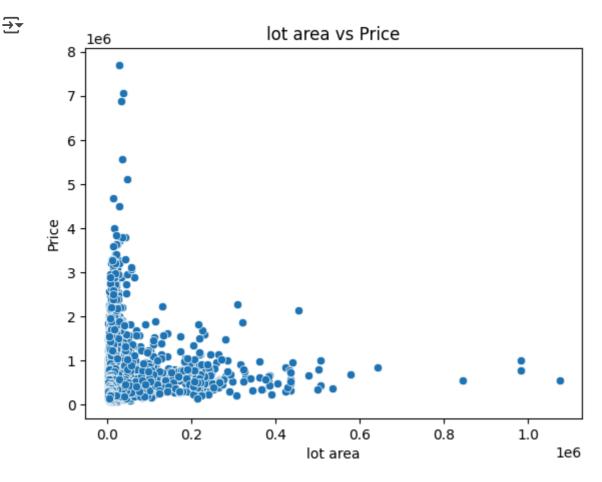
plt.figure(figsize=(10,8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Multivariate Analysis - Correlation Heatmap")
plt.show()
```



```
Multivariate Analysis - Correlation Heatmap
                                                                                                                                                 1.0
                                      id -1.000.050.350.550.650.160.310.110.290.050.670.570.290.070.110.290.450.070.660.090.060.000.000.70
                                    Date -0.051.000.020.030.020.090.010.010.000.030.020.020.010.010.020.020.020.030.010.010.020.020.020.020.030.000.010.03
                   number of bedrooms -0.330.02<mark>1.000.510.57</mark>0.030.180.010.080.030.350.470.300.150.020.040.010.14<mark>0.39</mark>0.030.090.000.31
                                                                                                                                                - 0.8
                  number of bathrooms -0.520.030.511.000.750.080.500.060.180.120.660.680.290.500.050.110.030.220.570.080.000.010.53
                             living area -0.650.020.570.751.000.170.350.110.290.000.760.880.440.310.060.080.050.240.760.180.000.000.71
                                                                                                                                                 0.6
                                lot area -0.100.000.030.080.171.000.000.030.080.010.110.180.020.050.010.070.090.220.150.710.010.000.08
                       number of floors -0.310.010.180.500.350.001.000.020.020.270.460.530.240.480.010.130.050.130.290.010.010.020.26
                     waterfront present -0.1 0.010.010.060.110.030.021.000.400.020.080.070.090.020.090.040.020.050.090.030.000.000.20
                                                                                                                                                - 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 -0.050.030.030.130.060.030.270.020.051.000.150.170.180.350.060.050.060.120.160.060.030.000.04
                     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.
                                                                                                                                                 - 0.2
Area of the basement -0.290.020.300.290.440.020.240.090.290.180.170.051.000.140.080.010.110.150.200.010.010.010.000.33
                                                                                                                                                 - 0.0
                               Built Year -0.070.010.150.500.310.050.480.020.060.3 0.440.420.14.000.230.060.140.410.330.070.060.000.05
                        Renovation Year -0.130.010.020.050.060.010.010.090.190.060.010.030.080.231.000.020.030.080.010.000.010.000.010.13
                            Postal Code -0.290.020.040.130.080.070.130.040.040.050.150.080.030.060.021.0000.3310.160.130.080.010.010.12
                                                                                                                                                 -0.2
                               Lattitude -0.440.020.010.030.050.090.050.020.000.000.120.000.110.140.030.311.000.130.050.090.010.010.30
                              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 -0.6(0.030.390.570.760.150.290.090.280.1(0.720.740.200.330.000.110.050.341.000.190.000.010
                         lot_area_renov -0.090.000.030.080.18<mark>0.73</mark>0.010.030.070.000.120.190.010.070.010.080.090.260.191.0000.030.010.080
             Price - 0.770.030.310.530.710.080.260.260.400.040.670.620.330.050.130.120.300.020.5 £0.080.010.001.00
                                                              lot area
                                                   number of bedrooms
                                                      number of bathrooms
                                                          living area
                                                                  number of floors
                                                                      waterfront present
                                                                          number of views
                                                                              condition of the house
                                                                                  grade of the house
                                                                                      Area of the house(excluding basement)
                                                                                          Area of the basement
                                                                                              Built Year
                                                                                                  Renovation Year
                                                                                                      Postal Code
                                                                                                              Longitude
                                                                                                          Lattitude
                                                                                                                          Number of schools nearby
                                                                                                                   living area renov
                                                                                                                      lot_area_renov
                                                                                                                              Distance from the airpor
```

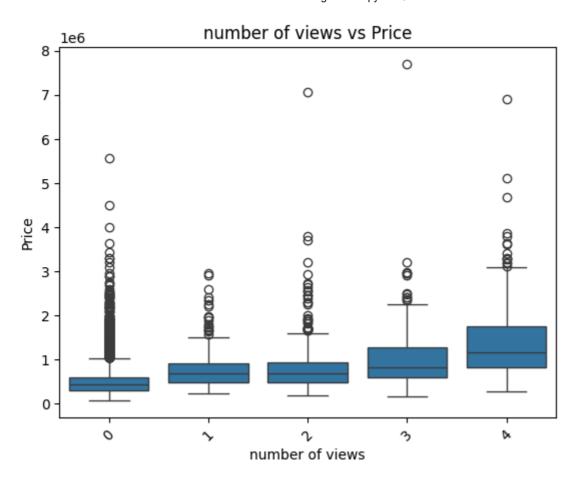
```
if 'lot area' in df.columns and 'Price' in df.columns:
    sns.scatterplot(data=df, x='lot area', y='Price')
    plt.title('lot area vs Price')
```

```
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)

\rightarrow		id	Date	number of bedro	oms number of bat	hrooms \
	count	1.462000e+04	14620.000000	14620.000	000 14620.	000000
	mean	6.762821e+09	42604.538646	3.379	343 2.	129583
	std	6.237575e+03	67.347991	0.938	719 0.	769934
	min	6.762810e+09	42491.000000	1.000	000 0.	500000
	25%	6.762815e+09	42546.000000	3.000	000 1.	750000
	50%	6.762821e+09	42600.000000	3.000	000 2.	250000
	75% 6.762826e+09		42662.000000	4.000	000 2.	500000
	max	6.762832e+09	42734.000000 33.000000		000 8.	000000
		living area	lot area	number of floor	s waterfront pres	ent \
	count	14620.000000	1.462000e+04	14620.00000	0 14620.000	000
	mean	2098.262996	1.509328e+04	1.50236	0.007	661
	std	928.275721	3.791962e+04	0.54023	9 0.087	193
	min	370.000000	5.200000e+02	1.00000	0.000	000
	25%	1440.000000	5.010750e+03	1.00000	0.000	000
	50%	1930.000000	7.620000e+03	1.50000	0.000	000
	75%	2570.000000	1.080000e+04	2.00000	0.000	000
	max	13540.000000	1.074218e+06	3.50000	0 1.000	000
		number of vie	ws condition	of the house \hdots .	. Built Year \	
	count	14620.000000		14620.000000	. 14620.000000	
	mean	0.2331	05	3.430506	. 1970.926402	
	std	0.766259		0.664151	. 29.493625	
	min	0.00000		1.000000	. 1900.000000	
	25%	0.0000	00	3.000000	. 1951.000000	

			7.001g/iii/io/it/iipy/iib	Joines	
50%	0.000000		3.000000	1975.000000	
75%	0.000000		4.000000	1997.000000	
max	4.000000		5.000000	2015.000000	
	Renovation Year	Postal Code		O	\
count	14620.000000	14620.00000			
mean	90.924008	122033.06224	4 52.792848	-114.404007	
std	416.216661	19.08241	8 0.137522	0.141326	
min	0.000000	122003.00000	52.385900	-114.709000	
25%	0.000000	122017.00000	52.707600	-114.519000	
50%	0.000000	122032.00000	52.806400	-114.421000	
75%	0.000000	122048.00000	52.908900	-114.315000	
max	2015.000000	122072.00000	53.007600	-113.505000	
	living_area_reno			schools nearby	
count	14620.00000			14620.000000	
mean	1996.70225		9068	2.012244	
std	691.09336			0.817284	
min	460.00000	0 651.00	9000	1.000000	
25%	1490.00000	0 5097.750	9000	1.000000	
50%	1850.00000	0 7620.00	9000	2.000000	
75%	2380.00000	0 10125.00	0000	3.000000	
max	6110.00000	0 560617.00	9000	3.000000	
	Distance from th	•	Price		
count		20.000000 1.4			
mean			389322e+05		
std			675324e+05		
min		50.000000 7.8			
25%		57.000000 3.			
50%			500000e+05		
75%		73.000000 6.4	450000e+05		

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

→ Missing values:

6	
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	a