

DA ASSIGNMENT – 3

NAME : KARUNAKARAN

REGISTER NO : 911720104027

Load the dataset :

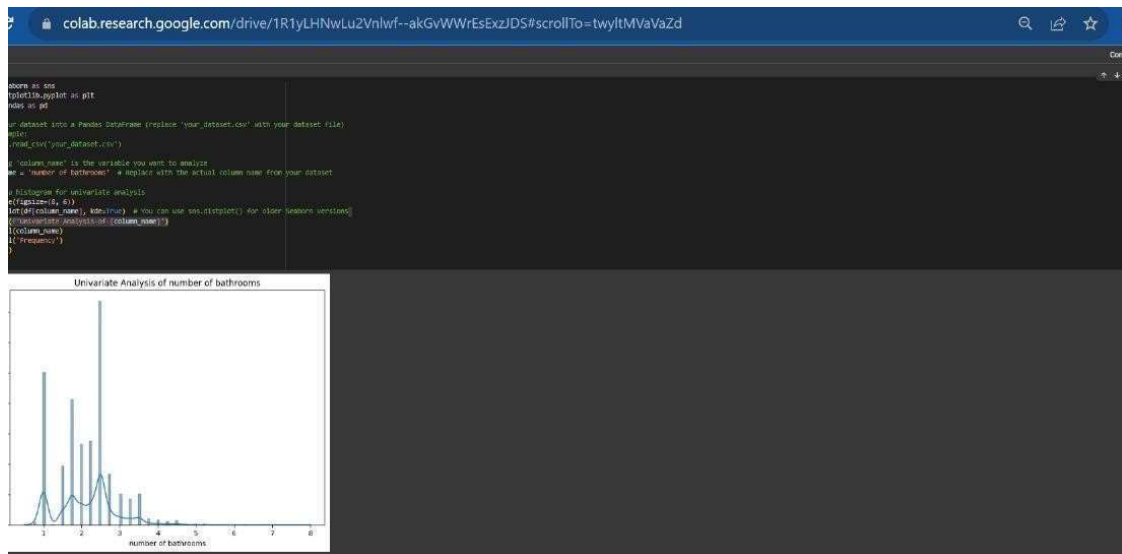
```
colab.research.google.com/drive/1R1yLHNwLu2Vnlwf--akGvWWrEsExzJDS

Untitled0.ipynb
File Edit View Insert Runtime Tools Help Last edited on October 5
+ Code + Text
Connect
(x)
import pandas as pd
df=pd.read_csv("/content/drive/MyDrive/House Price India.csv")
df.head()
```

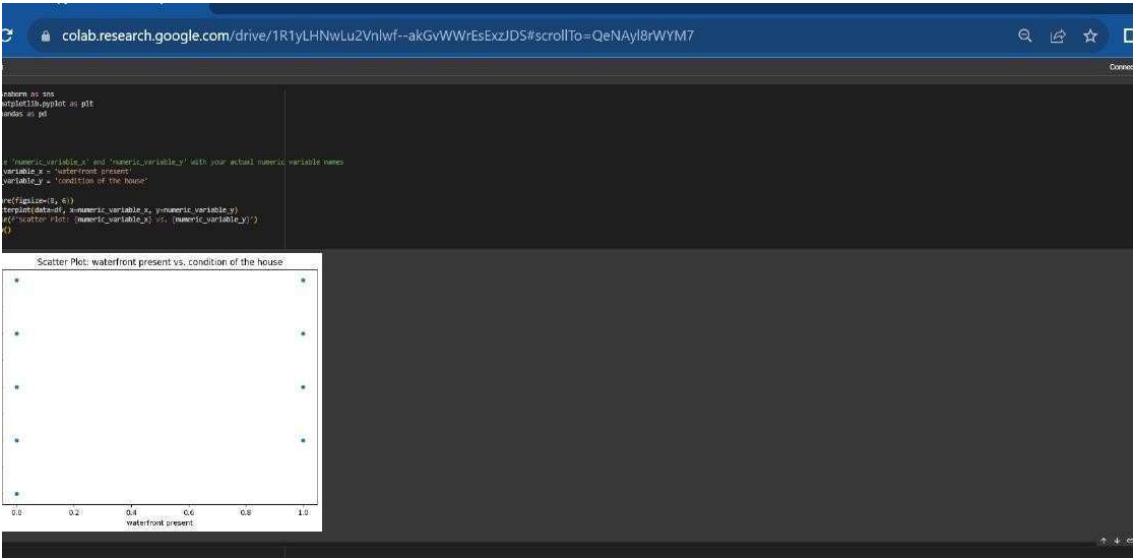
	id	Date	number of bedrooms	number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	...	Built Year	Renovation Year	Postal Code	Latitude	Longitude
0	6762810145	42491	5	2.50	3650	9050	2.0	0	4	5	...	1921	0	122003	52.8645	-114.557
1	6762810635	42491	4	2.50	2920	4000	1.5	0	0	5	...	1909	0	122004	52.8878	-114.470
2	6762810998	42491	5	2.75	2910	9480	1.5	0	0	3	...	1939	0	122004	52.8852	-114.468
3	6762812605	42491	4	2.50	3310	42998	2.0	0	0	3	...	2001	0	122005	52.9532	-114.321
4	6762812919	42491	3	2.00	2710	4500	1.5	0	0	4	...	1929	0	122006	52.9047	-114.485

5 rows x 23 columns

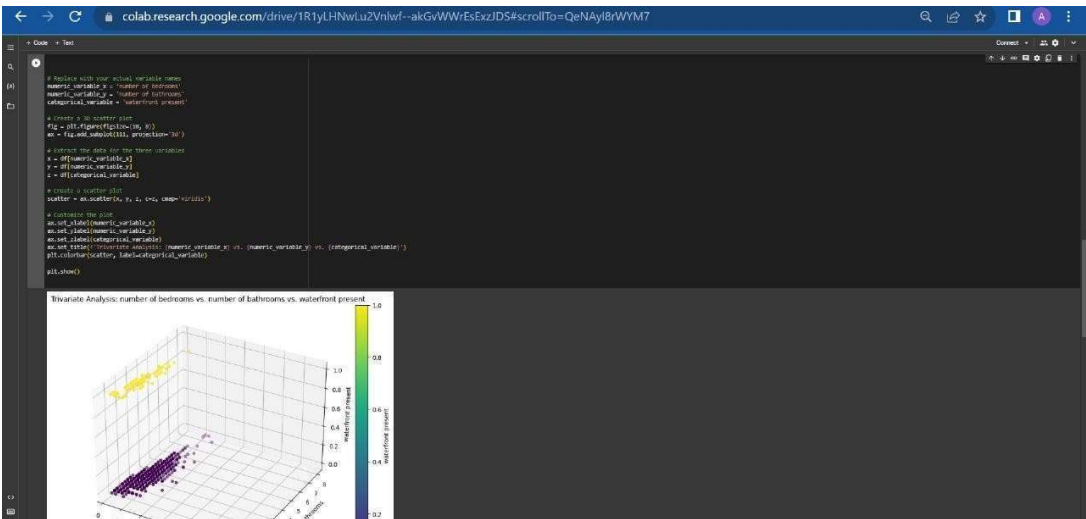
Univariate Analysis :



Bi - Variate Analysis :



Multi-Variate Analysis :



Descriptive statistics on the dataset :

```

+ Code + Test
  • Run interactive statistics for numerical columns
    query_name = de.devises[1]
    print(query_name)

    id      date  number of bedrooms  number of bathrooms \
count  1.46288e+04  14628.000000  14628.00000  14628.00000
mean   6.76222e+03  6762.222222  3.779210  3.122023
std    6.23797e+03  6237.971091  6.238712  6.709354
min    6.76222e+03  6762.000000  3.000000  3.000000
max    6.76222e+03  6762.000000  3.000000  3.750000
std    6.76222e+03  6762.000000  3.000000  3.750000
min    6.76222e+03  6762.000000  3.000000  3.000000
max    6.76222e+03  6762.000000  3.000000  3.750000

    living area  lot area  number of floors  underwater percent \
count  14628.00000  1.46288e+04  14628.00000  14628.00000
mean   2006.222729  1.589120e+04  1.589230  0.001151
std    1906.000000  1.200000e+04  1.000000  0.000000
min    179.000000  5.200000e+03  1.000000  0.000000
max    179.000000  5.200000e+03  1.000000  0.000000
std    179.000000  5.200000e+03  1.000000  0.000000
min    179.000000  5.200000e+03  1.000000  0.000000
max    179.000000  5.200000e+03  1.000000  0.000000

    number of votes  conviction of the house  ...  null rate \
count  14628.00000  14628.00000  ...  14628.00000
mean   8.223340  8.433986  ...  3776.530402
std    8.223340  8.433986  ...  3776.530402
min    8.000000  8.000000  ...  3776.530402
max    8.000000  8.000000  ...  3776.530402
std    8.000000  8.000000  ...  3776.530402
min    8.000000  8.000000  ...  3776.530402
max    8.000000  8.000000  ...  3776.530402

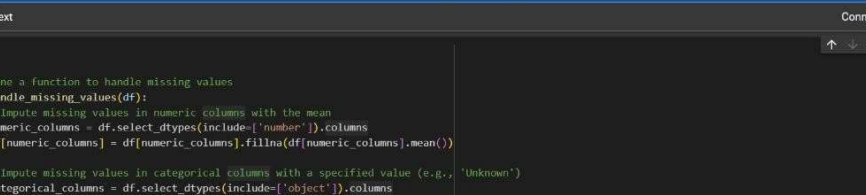
    demolition time  postal code  latitude  longitude \
count  14628.00000  14628.00000  14628.00000  14628.00000
mean   365.222222  11070.000000  42.187500  18.744166
std    442.766667  11.000000  0.139292  0.144166
min    6.000000  11070.000000  42.187500  18.744166
max    6.000000  11070.000000  42.187500  18.744166
std    6.000000  11070.000000  42.187500  18.744166
min    6.000000  11070.000000  42.187500  18.744166
max    6.000000  11070.000000  42.187500  18.744166

    living area  house  lot area  review  number of schools nearby \
count  14628.00000  14628.00000  14628.00000  14628.00000
mean   1916.202729  1916.202729  1916.202729  1916.202729
std    681.893346  681.893346  681.893346  681.893346
min    189.000000  189.000000  189.000000  189.000000
max    1916.202729  1916.202729  1916.202729  1916.202729
std    681.893346  681.893346  681.893346  681.893346
min    189.000000  189.000000  189.000000  189.000000
max    1916.202729  1916.202729  1916.202729  1916.202729

    distance from the airport  price \
count  14628.00000  14628.00000
mean   66.958816  66.958816
std    66.958816  66.958816
min    66.958816  66.958816
max    66.958816  66.958816
std    66.958816  66.958816
min    66.958816  66.958816
max    66.958816  66.958816

```

Handle the Missing values :



The screenshot displays a Jupyter Notebook environment. The top toolbar includes icons for file operations, search, and user profile. Below the toolbar, there are tabs for '+ Code' and '+ Text'. The main area shows a Python script with the following content:

```
# Define a function to handle missing values
def handle_missing_values(df):
    # Impute missing values in numeric columns with the mean
    numeric_columns = df.select_dtypes(include=['number']).columns
    df[numeric_columns] = df[numeric_columns].fillna(df[numeric_columns].mean())

    # Impute missing values in categorical columns with a specified value (e.g., 'Unknown')
    categorical_columns = df.select_dtypes(include=['object']).columns
    df[categorical_columns] = df[categorical_columns].fillna("Unknown")

    return df

# Apply the function to handle missing values
df = handle_missing_values(df)

# Check if there are any remaining missing values
missing_values_count = df.isnull().sum().sum()
if missing_values_count == 0:
    print("All missing values have been handled.")
else:
    print(f"There are still {missing_values_count} missing values in the dataset.")

# Save the cleaned dataset to a new file (optional)
df.to_csv('cleaned_dataset.csv', index=False) # Replace with your desired file name
```

At the bottom of the notebook, a status bar indicates: "All missing values have been handled."