Zoya Zahra_ML-DL-3_Selenium_Data_Scrapping_Code_Documentation

Assignment:

Scrape 50+ entries (Selenium + Crawl4AI)
Build preprocessing pipeline (clean → normalize → encode → save)
Submit final dataset

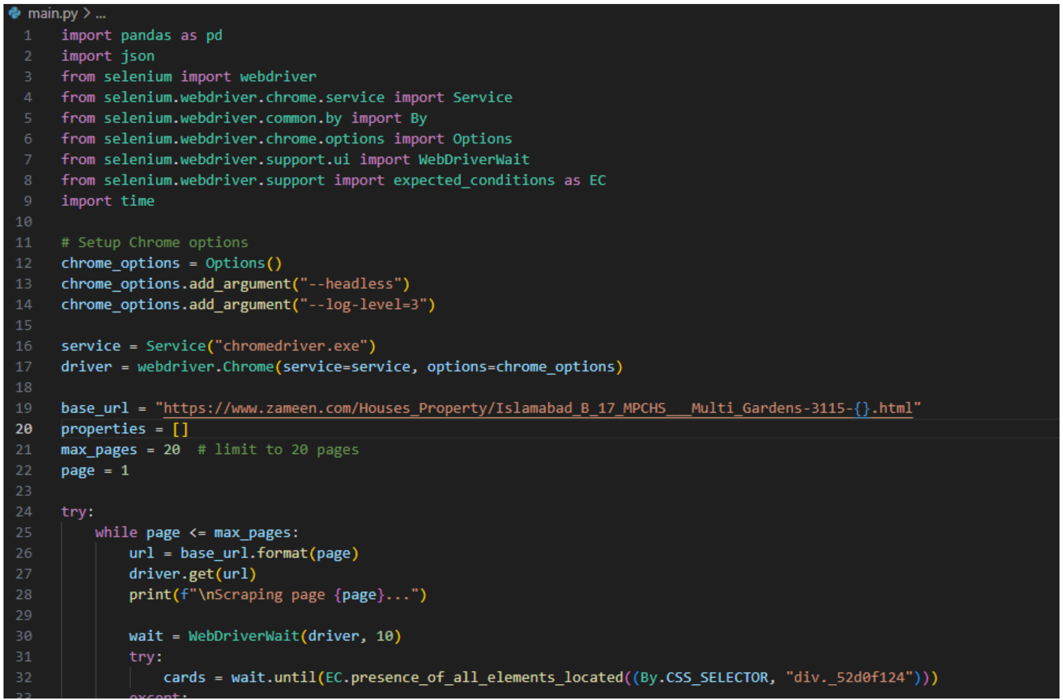
DATA COLLECTION USING SELENIUM:

```
import pandas as pd
import json
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import time

# Setup Chrome options
chrome_options = Options()
chrome_options.add_argument("--headless")
chrome_options.add_argument("--log-level=3")

service = Service("chromedriver.exe")
driver = webdriver.Chrome(service=service, options=chrome_options)

base_url =
"https://www.zameen.com/Houses_Property/Islamabad_B_17_MPCHS___Multi_Gardens-
3115-{}.html"
properties = []
max_pages = 20  # limit to 20 pages
page = 1

try:
    while page <= max_pages:
        url = base_url.format(page)
        driver.get(url)
        print(f"\nScraping page {page}...")

        wait = WebDriverWait(driver, 10)
        try:
            cards = wait.until(EC.presence_of_all_elements_located((By.CSS_SELECTOR,
"div._52d0f124")))
        except:
            print(f"No listings found on page {page}. Stopping...")
            break
        for card in cards:
            try:
                title = card.find_element(By.XPATH,
".//ancestor::li//a[@class='d870ae17']").get_attribute("title")
            except:
                title = None


            try:
    # Get price using full XPath
                price = card.find_element(By.XPATH, ".//ancestor::li//span[@aria-
label='Price']").text.strip()
            except:
                price = None

            try:
                area = card.find_element(By.CSS_SELECTOR, 'span[aria-label="Area"]
span').text.strip()
            except:
                area = None

            try:
                beds = card.find_element(By.CSS_SELECTOR, 'span[aria-
label="Beds"]').text.strip()
            except:
                beds = None

            try:
                baths = card.find_element(By.CSS_SELECTOR, 'span[aria-
label="Baths"]').text.strip()
            except:
                baths = None

            try:
                location = card.find_element(By.CSS_SELECTOR, 'div[aria-
label="Location"]').text.strip()
            except:
                location = None
            listing = {
                "Title": title,
                "Price": price,
                "Area": area,
                "Bedrooms": beds,
                "Bathrooms": baths,
                "Location": location
            }

            # Add to list
            properties.append(listing)

            # Print to terminal
            print(listing)

        page += 1
        time.sleep(1)  # Polite delay

finally:
    driver.quit()

# Save to CSV
df = pd.DataFrame(properties)
df.to_csv("zameen_listings.csv", index=False)
print("\nData saved to zameen_listings.csv")

# Save to JSON
with open("zameen_listings.json", "w", encoding="utf-8") as f:
    json.dump(properties, f, ensure_ascii=False, indent=4)
print("Data saved to zameen_listings.json")

print(f"Total listings scraped: {len(properties)}")
```

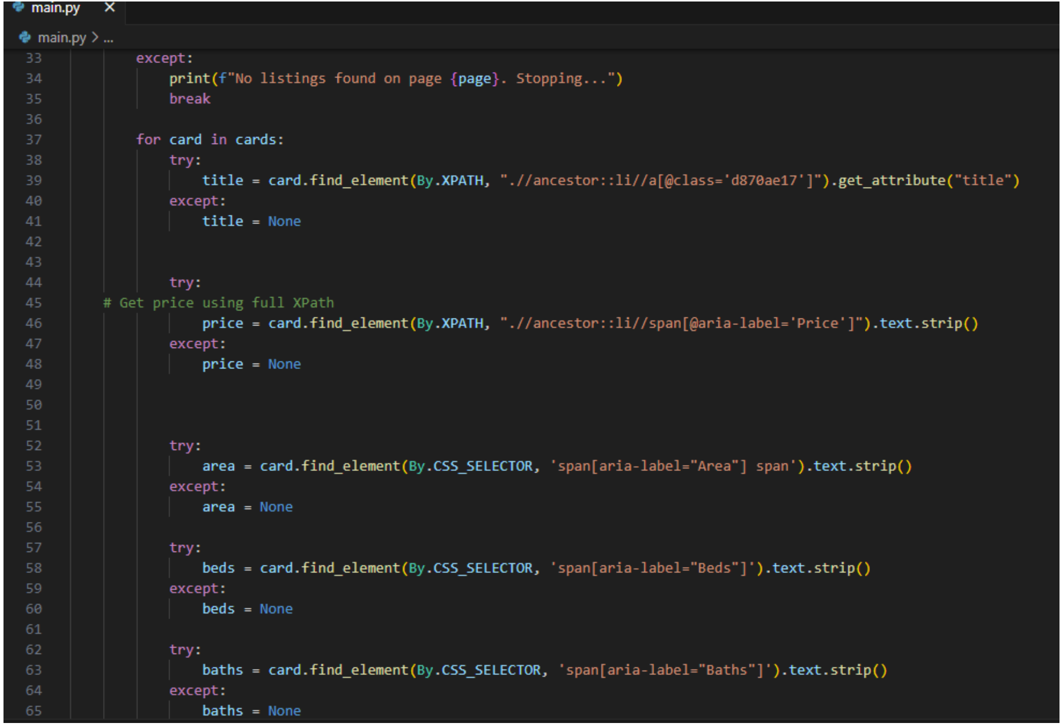
EXPLANATION:

The above code scrapes property listings from Zameen.com using Selenium. Code targets first 20 pages for the given URL. It opens each page in headless Chrome and waits for all property cards to load. For each card, the code works by extracting the title and the price through XPath. The bedrooms, bathrooms, and location are grabbed using CSS selectors. Each listing is stored as a dictionary which is appended to a list. After scraping all pages, the data is saved in form of CSV and JSON.

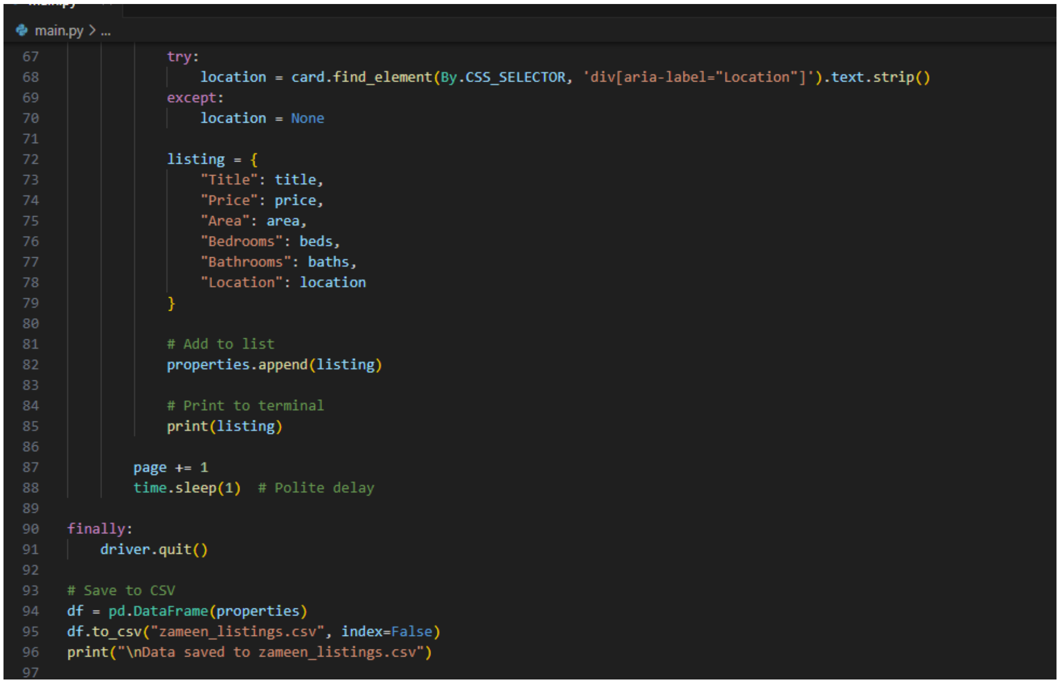
CODE SNIPPETS:



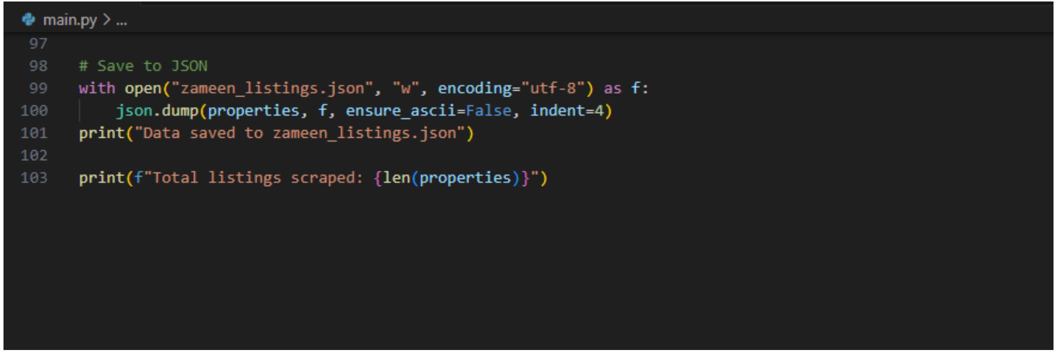
```python
import pandas as pd
import json
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import time

# Setup Chrome options
chrome_options = Options()
chrome_options.add_argument("--headless")
chrome_options.add_argument("--log-level=3")

service = Service("chromedriver.exe")
driver = webdriver.Chrome(service=service, options=chrome_options)

base_url = "https://www.zameen.com/Houses_Property/Islamabad_B_17_MPCHS___Multi_Gardens-3115-{}.html"
properties = []
max_pages = 20  # limit to 20 pages
page = 1

try:
    while page <= max_pages:
        url = base_url.format(page)
        driver.get(url)
        print(f"\nScraping page {page}...")

        wait = WebDriverWait(driver, 10)
        try:
            cards = wait.until(EC.presence_of_all_elements_located((By.CSS_SELECTOR, "div._52d0f124")))
```



```python
        except:
            print(f"No listings found on page {page}. Stopping...")
            break

        for card in cards:
            try:
                title = card.find_element(By.XPATH, ".//ancestor::li//a[@class='d870ae17']").get_attribute("title")
            except:
                title = None


            try:
    # Get price using full XPath
                price = card.find_element(By.XPATH, ".//ancestor::li//span[@aria-label='Price']").text.strip()
            except:
                price = None



            try:
                area = card.find_element(By.CSS_SELECTOR, 'span[aria-label="Area"] span').text.strip()
            except:
                area = None

            try:
                beds = card.find_element(By.CSS_SELECTOR, 'span[aria-label="Beds"]').text.strip()
            except:
                beds = None

            try:
                baths = card.find_element(By.CSS_SELECTOR, 'span[aria-label="Baths"]').text.strip()
            except:
                baths = None
```

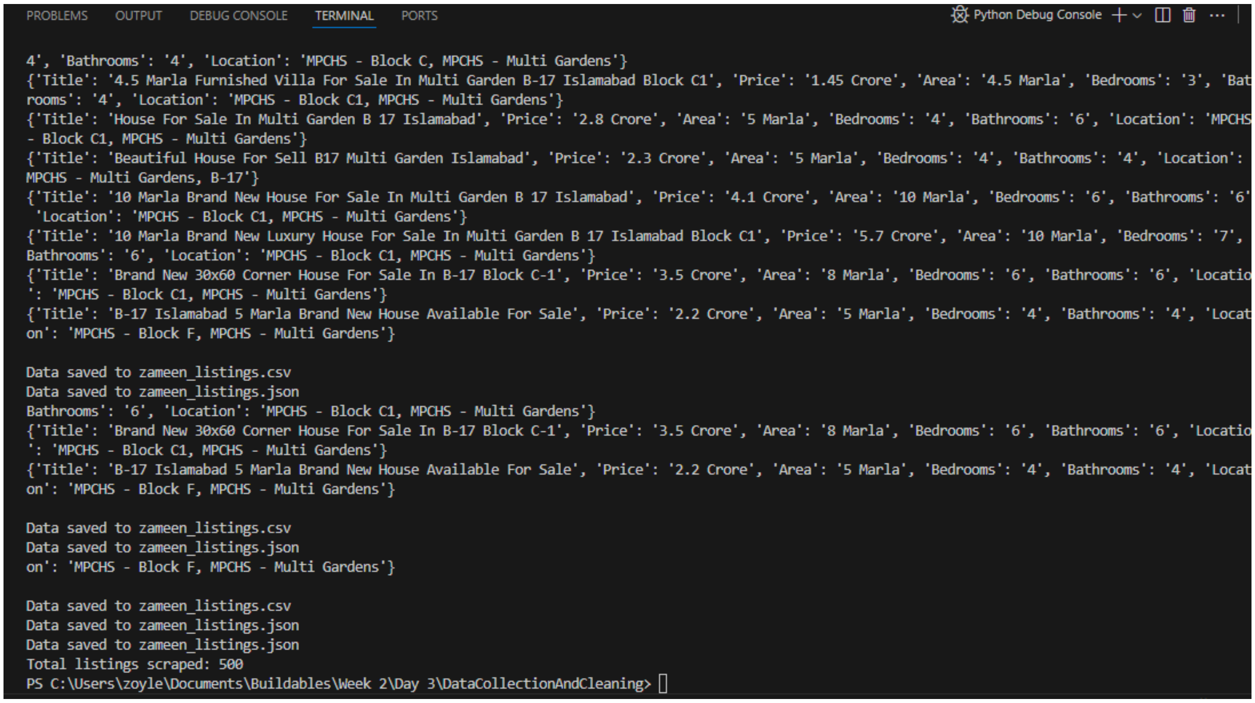


```python
            try:
                location = card.find_element(By.CSS_SELECTOR, 'div[aria-label="Location"]').text.strip()
            except:
                location = None

            listing = {
                "Title": title,
                "Price": price,
                "Area": area,
                "Bedrooms": beds,
                "Bathrooms": baths,
                "Location": location
            }

            # Add to list
            properties.append(listing)

            # Print to terminal
            print(listing)

        page += 1
        time.sleep(1)  # Polite delay

finally:
    driver.quit()

# Save to CSV
df = pd.DataFrame(properties)
df.to_csv("zameen_listings.csv", index=False)
print("\nData saved to zameen_listings.csv")
```



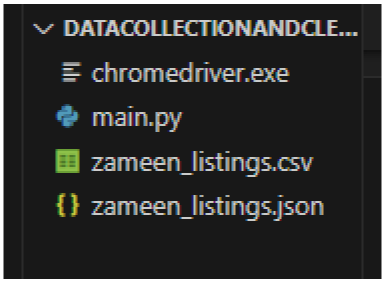
```python

# Save to JSON
with open("zameen_listings.json", "w", encoding="utf-8") as f:
    json.dump(properties, f, ensure_ascii=False, indent=4)
print("Data saved to zameen_listings.json")

print(f"Total listings scraped: {len(properties)}")
```

CODE OUTPUT:



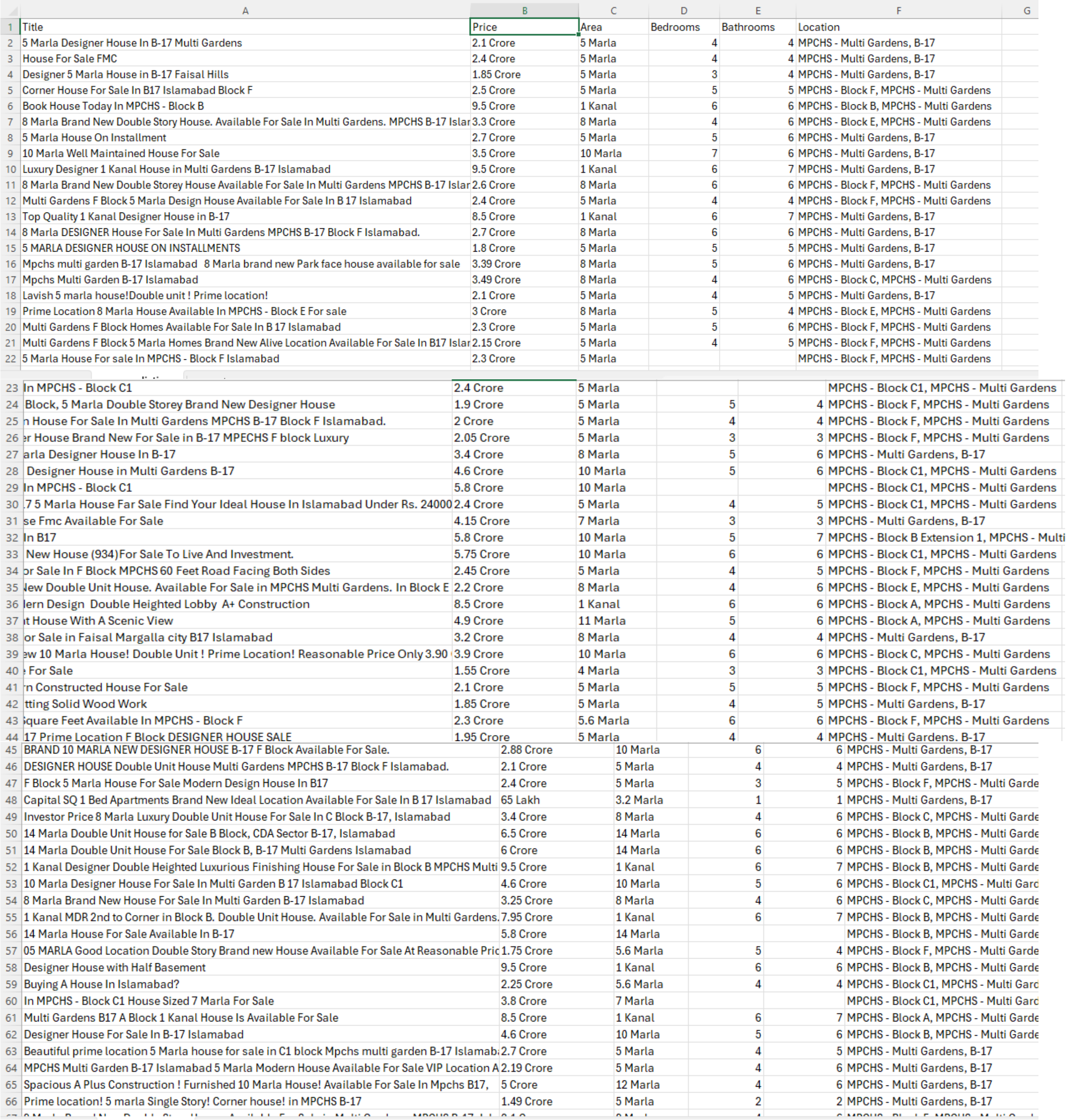
```
4', 'Bathrooms': '4', 'Location': 'MPCHS - Block C, MPCHS - Multi Gardens'}
{'Title': '4.5 Marla Furnished Villa For Sale In Multi Garden B-17 Islamabad Block C1', 'Price': '1.45 Crore', 'Area': '4.5 Marla', 'Bedrooms': '3', 'Bat
rooms': '4', 'Location': 'MPCHS - Block C1, MPCHS - Multi Gardens'}
{'Title': 'House For Sale In Multi Garden B 17 Islamabad', 'Price': '2.8 Crore', 'Area': '5 Marla', 'Bedrooms': '4', 'Bathrooms': '6', 'Location': 'MPCHS
- Block C1, MPCHS - Multi Gardens'}
{'Title': 'Beautiful House For Sell B17 Multi Garden Islamabad', 'Price': '2.3 Crore', 'Area': '5 Marla', 'Bedrooms': '4', 'Bathrooms': '4', 'Location':
MPCHS - Multi Gardens, B-17'}
{'Title': '10 Marla Brand New House For Sale In Multi Garden B 17 Islamabad', 'Price': '4.1 Crore', 'Area': '10 Marla', 'Bedrooms': '6', 'Bathrooms': '6'
 'Location': 'MPCHS - Block C1, MPCHS - Multi Gardens'}
{'Title': '10 Marla Brand New Luxury House For Sale In Multi Garden B 17 Islamabad Block C1', 'Price': '5.7 Crore', 'Area': '10 Marla', 'Bedrooms': '7',
Bathrooms': '6', 'Location': 'MPCHS - Block C1, MPCHS - Multi Gardens'}
{'Title': 'Brand New 30x60 Corner House For Sale In B-17 Block C-1', 'Price': '3.5 Crore', 'Area': '8 Marla', 'Bedrooms': '6', 'Bathrooms': '6', 'Locatio
': 'MPCHS - Block C1, MPCHS - Multi Gardens'}
{'Title': 'B-17 Islamabad 5 Marla Brand New House Available For Sale', 'Price': '2.2 Crore', 'Area': '5 Marla', 'Bedrooms': '4', 'Bathrooms': '4', 'Locat
on': 'MPCHS - Block F, MPCHS - Multi Gardens'}

Data saved to zameen_listings.csv
Data saved to zameen_listings.json
Bathrooms': '6', 'Location': 'MPCHS - Block C1, MPCHS - Multi Gardens'}
{'Title': 'Brand New 30x60 Corner House For Sale In B-17 Block C-1', 'Price': '3.5 Crore', 'Area': '8 Marla', 'Bedrooms': '6', 'Bathrooms': '6', 'Locatio
': 'MPCHS - Block C1, MPCHS - Multi Gardens'}
{'Title': 'B-17 Islamabad 5 Marla Brand New House Available For Sale', 'Price': '2.2 Crore', 'Area': '5 Marla', 'Bedrooms': '4', 'Bathrooms': '4', 'Locat
on': 'MPCHS - Block F, MPCHS - Multi Gardens'}

Data saved to zameen_listings.csv
Data saved to zameen_listings.json
on': 'MPCHS - Block F, MPCHS - Multi Gardens'}

Data saved to zameen_listings.csv
Data saved to zameen_listings.json
Data saved to zameen_listings.json
Total listings scraped: 500
PS C:\Users\zoyle\Documents\Buildables\Week 2\Day 3\DataCollectionAndCleaning> 
```

FILES CREATED:



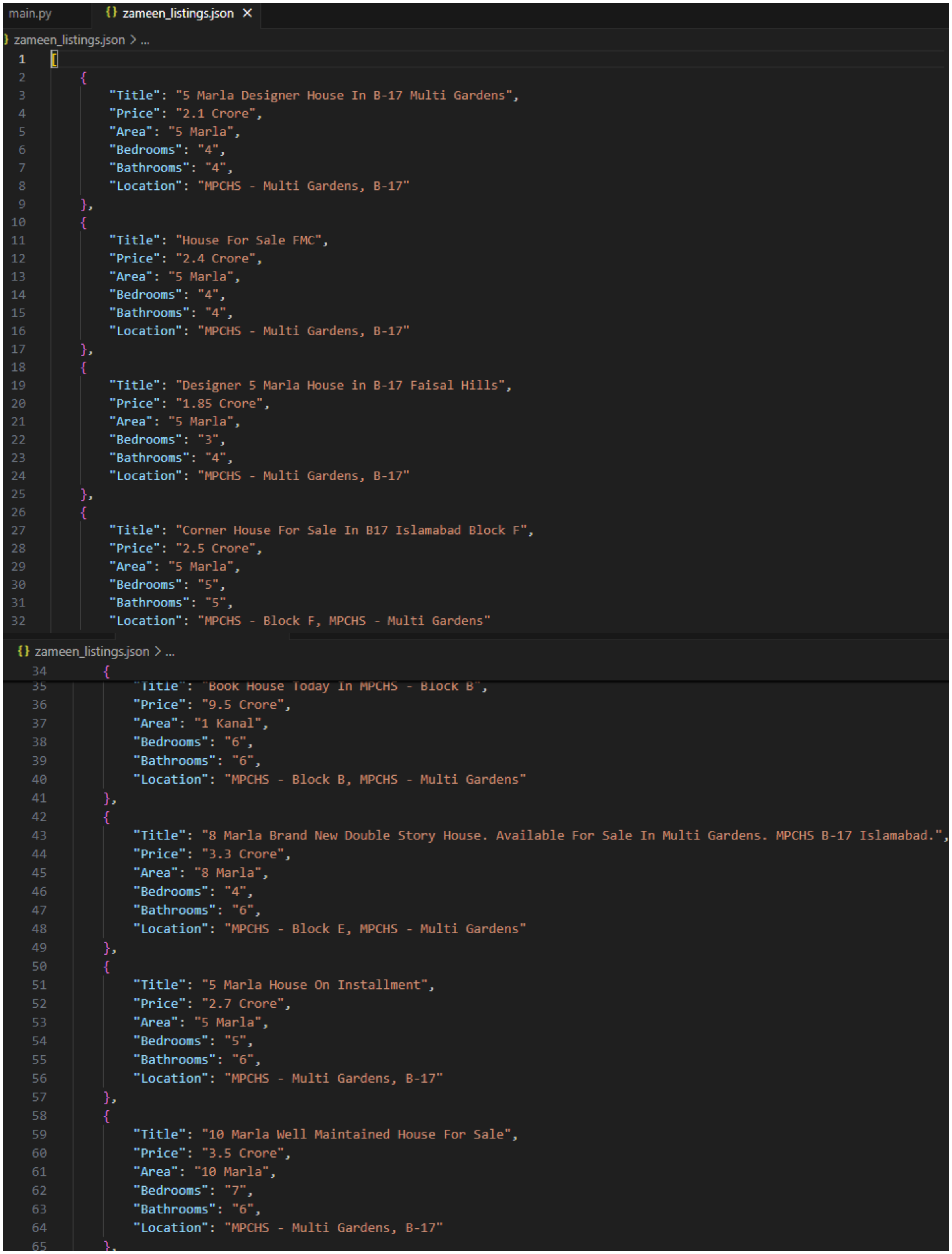
DATACOLLECTIONANDCLE...
- chromedriver.exe
- main.py
- zameen_listings.csv
- zameen_listings.json

CSV FILE:



	A	B	C	D	E	F	G
1	Title	Price	Area	Bedrooms	Bathrooms	Location	
2	5 Marla Designer House In B-17 Multi Gardens	2.1 Crore	5 Marla	4	4	MPCHS - Multi Gardens, B-17	
3	House For Sale FMC	2.4 Crore	5 Marla	4	4	MPCHS - Multi Gardens, B-17	
4	Designer 5 Marla House in B-17 Faisal Hills	1.85 Crore	5 Marla	3	4	MPCHS - Multi Gardens, B-17	
5	Corner House For Sale In B17 Islamabad Block F	2.5 Crore	5 Marla	5	5	MPCHS - Block F, MPCHS - Multi Gardens	
6	Book House Today In MPCHS - Block B	9.5 Crore	1 Kanal	6	6	MPCHS - Block B, MPCHS - Multi Gardens	
7	8 Marla Brand New Double Story House. Available For Sale In Multi Gardens. MPCHS B-17 Islar	3.3 Crore	8 Marla	4	6	MPCHS - Block E, MPCHS - Multi Gardens	
8	5 Marla House On Installment	2.7 Crore	5 Marla	5	6	MPCHS - Multi Gardens, B-17	
9	10 Marla Well Maintained House For Sale	3.5 Crore	10 Marla	7	6	MPCHS - Multi Gardens, B-17	
10	Luxury Designer 1 Kanal House in Multi Gardens B-17 Islamabad	9.5 Crore	1 Kanal	6	7	MPCHS - Multi Gardens, B-17	
11	8 Marla Brand New Double Storey House Available For Sale In Multi Gardens MPCHS B-17 Islar	2.6 Crore	8 Marla	6	6	MPCHS - Block F, MPCHS - Multi Gardens	
12	Multi Gardens F Block 5 Marla Design House Available For Sale In B 17 Islamabad	2.4 Crore	5 Marla	4	4	MPCHS - Block F, MPCHS - Multi Gardens	
13	Top Quality 1 Kanal Designer House in B-17	8.5 Crore	1 Kanal	6	7	MPCHS - Multi Gardens, B-17	
14	8 Marla DESIGNER House For Sale In Multi Gardens MPCHS B-17 Block F Islamabad.	2.7 Crore	8 Marla	6	6	MPCHS - Multi Gardens, B-17	
15	5 MARLA DESIGNER HOUSE ON INSTALLMENTS	1.8 Crore	5 Marla	5	5	MPCHS - Multi Gardens, B-17	
16	Mpchs multi garden B-17 Islamabad 8 Marla brand new Park face house available for sale	3.39 Crore	8 Marla	5	6	MPCHS - Multi Gardens, B-17	
17	Mpchs Multi Garden B-17 Islamabad	3.49 Crore	8 Marla	4	6	MPCHS - Block C, MPCHS - Multi Gardens	
18	Lavish 5 marla house!Double unit ! Prime location!	2.1 Crore	5 Marla	4	5	MPCHS - Multi Gardens, B-17	
19	Prime Location 8 Marla House Available In MPCHS - Block E For sale	3 Crore	8 Marla	5	4	MPCHS - Block E, MPCHS - Multi Gardens	
20	Multi Gardens F Block Homes Available For Sale In B 17 Islamabad	2.3 Crore	5 Marla	5	6	MPCHS - Block F, MPCHS - Multi Gardens	
21	Multi Gardens F Block 5 Marla Homes Brand New Alive Location Available For Sale In B17 Islar	2.15 Crore	5 Marla	4	5	MPCHS - Block F, MPCHS - Multi Gardens	
22	5 Marla House For sale In MPCHS - Block F Islamabad	2.3 Crore	5 Marla			MPCHS - Block F, MPCHS - Multi Gardens	
23	In MPCHS - Block C1	2.4 Crore	5 Marla			MPCHS - Block C1, MPCHS - Multi Gardens	
24	Block, 5 Marla Double Storey Brand New Designer House	1.9 Crore	5 Marla	5	4	MPCHS - Block F, MPCHS - Multi Gardens	
25	n House For Sale In Multi Gardens MPCHS B-17 Block F Islamabad.	2 Crore	5 Marla	4	4	MPCHS - Block F, MPCHS - Multi Gardens	
26	r House Brand New For Sale in B-17 MPECHS F block Luxury	2.05 Crore	5 Marla	3	3	MPCHS - Block F, MPCHS - Multi Gardens	
27	arla Designer House In B-17	3.4 Crore	8 Marla	5	6	MPCHS - Multi Gardens, B-17	
28	Designer House in Multi Gardens B-17	4.6 Crore	10 Marla	5	6	MPCHS - Block C1, MPCHS - Multi Gardens	
29	In MPCHS - Block C1	5.8 Crore	10 Marla			MPCHS - Block C1, MPCHS - Multi Gardens	
30	.7 5 Marla House Far Sale Find Your Ideal House In Islamabad Under Rs. 24000	2.4 Crore	5 Marla	4	5	MPCHS - Block C1, MPCHS - Multi Gardens	
31	se Fmc Available For Sale	4.15 Crore	7 Marla	3	3	MPCHS - Multi Gardens, B-17	
32	In B17	5.8 Crore	10 Marla	5	7	MPCHS - Block B Extension 1, MPCHS - Multi	
33	New House (934)For Sale To Live And Investment.	5.75 Crore	10 Marla	6	6	MPCHS - Block C1, MPCHS - Multi Gardens	
34	or Sale In F Block MPCHS 60 Feet Road Facing Both Sides	2.45 Crore	5 Marla	4	5	MPCHS - Block F, MPCHS - Multi Gardens	
35	lew Double Unit House. Available For Sale in MPCHS Multi Gardens. In Block E	2.2 Crore	8 Marla	4	6	MPCHS - Block E, MPCHS - Multi Gardens	
36	lern Design Double Heighted Lobby A+ Construction	8.5 Crore	1 Kanal	6	6	MPCHS - Block A, MPCHS - Multi Gardens	
37	t House With A Scenic View	4.9 Crore	11 Marla	5	6	MPCHS - Block A, MPCHS - Multi Gardens	
38	or Sale in Faisal Margalla city B17 Islamabad	3.2 Crore	8 Marla	4	4	MPCHS - Multi Gardens, B-17	
39	ew 10 Marla House! Double Unit ! Prime Location! Reasonable Price Only 3.90	3.9 Crore	10 Marla	6	6	MPCHS - Block C, MPCHS - Multi Gardens	
40	For Sale	1.55 Crore	4 Marla	3	3	MPCHS - Block C1, MPCHS - Multi Gardens	
41	rn Constructed House For Sale	2.1 Crore	5 Marla	5	5	MPCHS - Block F, MPCHS - Multi Gardens	
42	tting Solid Wood Work	1.85 Crore	5 Marla	4	5	MPCHS - Multi Gardens, B-17	
43	square Feet Available In MPCHS - Block F	2.3 Crore	5.6 Marla	6	6	MPCHS - Block F, MPCHS - Multi Gardens	
44	17 Prime Location F Block DESIGNER HOUSE SALE	1.95 Crore	5 Marla	4	4	MPCHS - Multi Gardens, B-17	
45	BRAND 10 MARLA NEW DESIGNER HOUSE B-17 F Block Available For Sale.	2.88 Crore	10 Marla	6	6	MPCHS - Multi Gardens, B-17	
46	DESIGNER HOUSE Double Unit House Multi Gardens MPCHS B-17 Block F Islamabad.	2.1 Crore	5 Marla	4	4	MPCHS - Multi Gardens, B-17	
47	F Block 5 Marla House For Sale Modern Design House In B17	2.4 Crore	5 Marla	3	5	MPCHS - Block F, MPCHS - Multi Garde	
48	Capital SQ 1 Bed Apartments Brand New Ideal Location Available For Sale In B 17 Islamabad	65 Lakh	3.2 Marla	1	1	MPCHS - Multi Gardens, B-17	
49	Investor Price 8 Marla Luxury Double Unit House For Sale In C Block B-17, Islamabad	3.4 Crore	8 Marla	4	6	MPCHS - Block C, MPCHS - Multi Garde	
50	14 Marla Double Unit House for Sale B Block, CDA Sector B-17, Islamabad	6.5 Crore	14 Marla	6	6	MPCHS - Block B, MPCHS - Multi Garde	
51	14 Marla Double Unit House For Sale Block B, B-17 Multi Gardens Islamabad	6 Crore	14 Marla	6	6	MPCHS - Block B, MPCHS - Multi Garde	
52	1 Kanal Designer Double Heighted Luxurious Finishing House For Sale in Block B MPCHS Multi	9.5 Crore	1 Kanal	6	7	MPCHS - Block B, MPCHS - Multi Garde	
53	10 Marla Designer House For Sale In Multi Garden B 17 Islamabad Block C1	4.6 Crore	10 Marla	5	6	MPCHS - Block C1, MPCHS - Multi Garc	
54	8 Marla Brand New House For Sale In Multi Garden B-17 Islamabad	3.25 Crore	8 Marla	4	6	MPCHS - Block C, MPCHS - Multi Garde	
55	1 Kanal MDR 2nd to Corner in Block B. Double Unit House. Available For Sale in Multi Gardens.	7.95 Crore	1 Kanal	6	7	MPCHS - Block B, MPCHS - Multi Garde	
56	14 Marla House For Sale Available In B-17	5.8 Crore	14 Marla			MPCHS - Block B, MPCHS - Multi Garde	
57	05 MARLA Good Location Double Story Brand new House Available For Sale At Reasonable Pric	1.75 Crore	5.6 Marla	5	4	MPCHS - Block F, MPCHS - Multi Garde	
58	Designer House with Half Basement	9.5 Crore	1 Kanal	6	6	MPCHS - Block B, MPCHS - Multi Garde	
59	Buying A House In Islamabad?	2.25 Crore	5.6 Marla	4	4	MPCHS - Block C1, MPCHS - Multi Garc	
60	In MPCHS - Block C1 House Sized 7 Marla For Sale	3.8 Crore	7 Marla			MPCHS - Block C1, MPCHS - Multi Garc	
61	Multi Gardens B17 A Block 1 Kanal House Is Available For Sale	8.5 Crore	1 Kanal	6	7	MPCHS - Block A, MPCHS - Multi Garde	
62	Designer House For Sale In B-17 Islamabad	4.6 Crore	10 Marla	5	6	MPCHS - Block B, MPCHS - Multi Garde	
63	Beautiful prime location 5 Marla house for sale in C1 block Mpchs multi garden B-17 Islamab	2.7 Crore	5 Marla	4	5	MPCHS - Multi Gardens, B-17	
64	MPCHS Multi Garden B-17 Islamabad 5 Marla Modern House Available For Sale VIP Location A	2.19 Crore	5 Marla	4	6	MPCHS - Multi Gardens, B-17	
65	Spacious A Plus Construction ! Furnished 10 Marla House! Available For Sale In Mpchs B17,	5 Crore	12 Marla	4	6	MPCHS - Multi Gardens, B-17	
66	Prime location! 5 marla Single Story! Corner house! in MPCHS B-17	1.49 Crore	5 Marla	2	2	MPCHS - Multi Gardens, B-17	

JSON FILE:

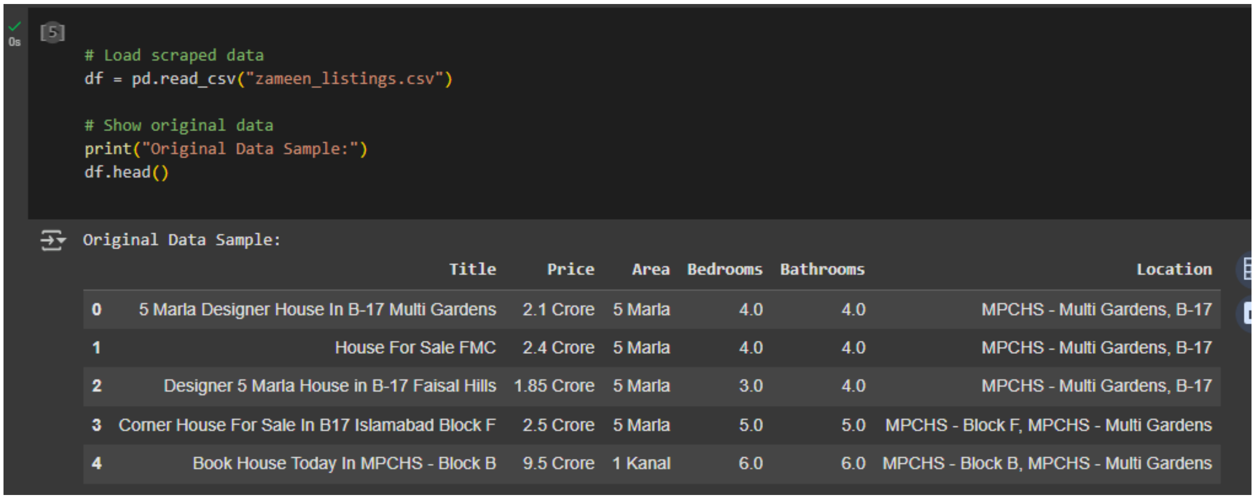


```
[
    {
        "Title": "5 Marla Designer House In B-17 Multi Gardens",
        "Price": "2.1 Crore",
        "Area": "5 Marla",
        "Bedrooms": "4",
        "Bathrooms": "4",
        "Location": "MPCHS - Multi Gardens, B-17"
    },
    {
        "Title": "House For Sale FMC",
        "Price": "2.4 Crore",
        "Area": "5 Marla",
        "Bedrooms": "4",
        "Bathrooms": "4",
        "Location": "MPCHS - Multi Gardens, B-17"
    },
    {
        "Title": "Designer 5 Marla House in B-17 Faisal Hills",
        "Price": "1.85 Crore",
        "Area": "5 Marla",
        "Bedrooms": "3",
        "Bathrooms": "4",
        "Location": "MPCHS - Multi Gardens, B-17"
    },
    {
        "Title": "Corner House For Sale In B17 Islamabad Block F",
        "Price": "2.5 Crore",
        "Area": "5 Marla",
        "Bedrooms": "5",
        "Bathrooms": "5",
        "Location": "MPCHS - Block F, MPCHS - Multi Gardens"
```

```
    {
        "Title": "Book House Today In MPCHS - Block B",
        "Price": "9.5 Crore",
        "Area": "1 Kanal",
        "Bedrooms": "6",
        "Bathrooms": "6",
        "Location": "MPCHS - Block B, MPCHS - Multi Gardens"
    },
    {
        "Title": "8 Marla Brand New Double Story House. Available For Sale In Multi Gardens. MPCHS B-17 Islamabad.",
        "Price": "3.3 Crore",
        "Area": "8 Marla",
        "Bedrooms": "4",
        "Bathrooms": "6",
        "Location": "MPCHS - Block E, MPCHS - Multi Gardens"
    },
    {
        "Title": "5 Marla House On Installment",
        "Price": "2.7 Crore",
        "Area": "5 Marla",
        "Bedrooms": "5",
        "Bathrooms": "6",
        "Location": "MPCHS - Multi Gardens, B-17"
    },
    {
        "Title": "10 Marla Well Maintained House For Sale",
        "Price": "3.5 Crore",
        "Area": "10 Marla",
        "Bedrooms": "7",
        "Bathrooms": "6",
        "Location": "MPCHS - Multi Gardens, B-17"
    },
```

DATA CLEANING AND PREPROCESSING:

Viewing Raw Data Head:



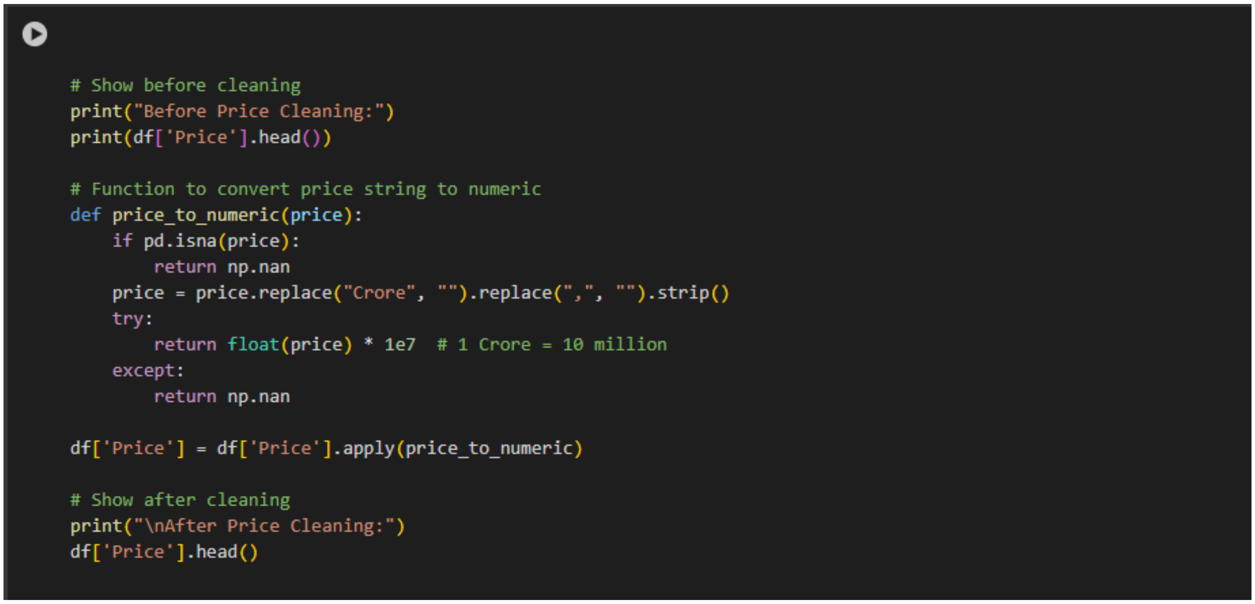
```
# Load scraped data
df = pd.read_csv("zameen_listings.csv")

# Show original data
print("Original Data Sample:")
df.head()
```

```
Original Data Sample:
```

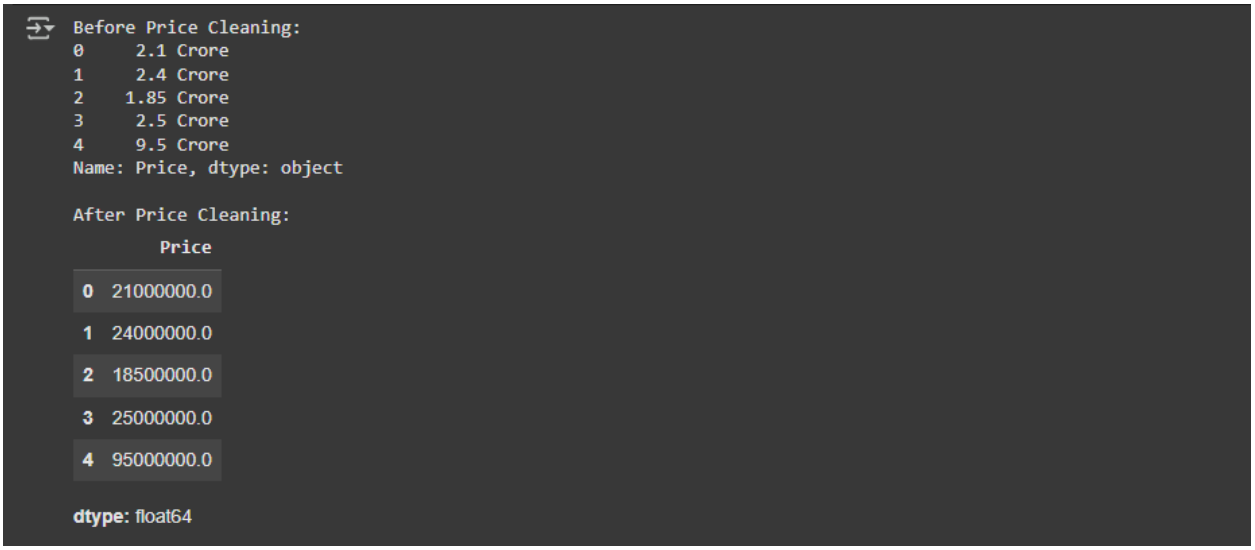
	Title	Price	Area	Bedrooms	Bathrooms	Location
0	5 Marla Designer House In B-17 Multi Gardens	2.1 Crore	5 Marla	4.0	4.0	MPCHS - Multi Gardens, B-17
1	House For Sale FMC	2.4 Crore	5 Marla	4.0	4.0	MPCHS - Multi Gardens, B-17
2	Designer 5 Marla House in B-17 Faisal Hills	1.85 Crore	5 Marla	3.0	4.0	MPCHS - Multi Gardens, B-17
3	Corner House For Sale In B17 Islamabad Block F	2.5 Crore	5 Marla	5.0	5.0	MPCHS - Block F, MPCHS - Multi Gardens
4	Book House Today In MPCHS - Block B	9.5 Crore	1 Kanal	6.0	6.0	MPCHS - Block B, MPCHS - Multi Gardens

Modifying Price Column:

The price column consisted of values such as 1 crore 0.2 crores. Converted them into Numbers. Removed the string "crore" from the price column.



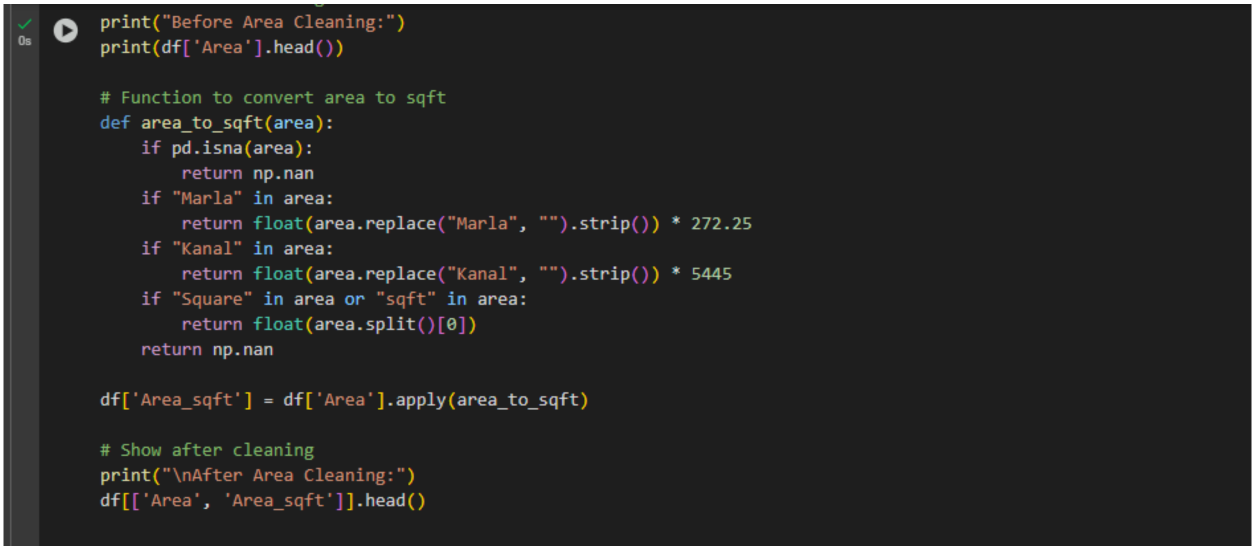
```
# Show before cleaning
print("Before Price Cleaning:")
print(df['Price'].head())

# Function to convert price string to numeric
def price_to_numeric(price):
    if pd.isna(price):
        return np.nan
    price = price.replace("Crore", "").replace(",", "").strip()
    try:
        return float(price) * 1e7  # 1 Crore = 10 million
    except:
        return np.nan

df['Price'] = df['Price'].apply(price_to_numeric)

# Show after cleaning
print("\nAfter Price Cleaning:")
df['Price'].head()
```



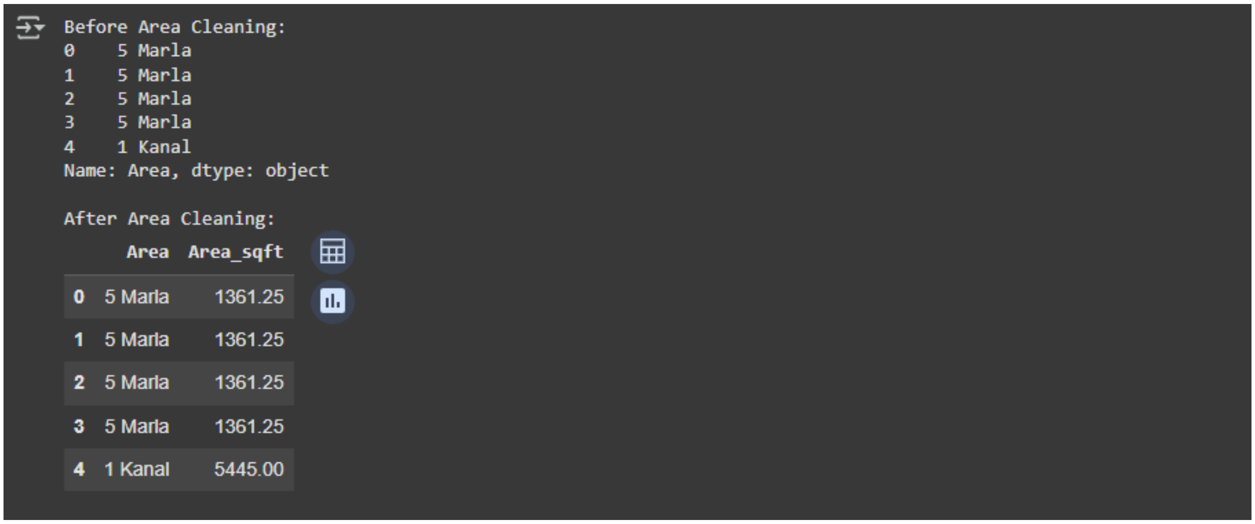
```
Before Price Cleaning:
0     2.1 Crore
1     2.4 Crore
2    1.85 Crore
3     2.5 Crore
4     9.5 Crore
Name: Price, dtype: object

After Price Cleaning:
        Price
0  21000000.0
1  24000000.0
2  18500000.0
3  25000000.0
4  95000000.0

dtype: float64
```

Cleaning Area Column:

The Area Column consisted of values such as 1 Marla and 0.5 Kanals. In order to make a consistent area representation, the marlas and kanals were converted into square feet. The strings Marla and Kanals were removed from the area column.



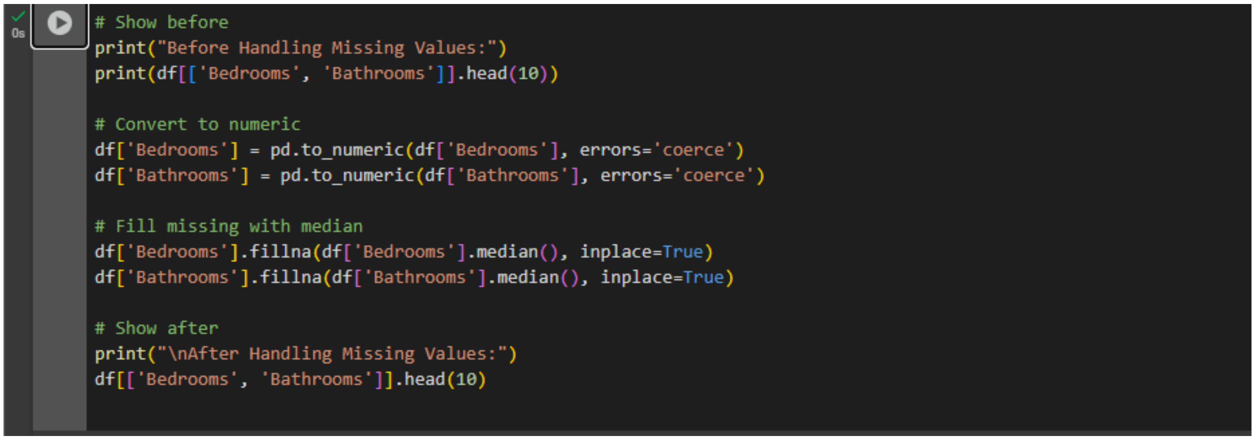
```
print("Before Area Cleaning:")
print(df['Area'].head())

# Function to convert area to sqft
def area_to_sqft(area):
    if pd.isna(area):
        return np.nan
    if "Marla" in area:
        return float(area.replace("Marla", "").strip()) * 272.25
    if "Kanal" in area:
        return float(area.replace("Kanal", "").strip()) * 5445
    if "Square" in area or "sqft" in area:
        return float(area.split()[0])
    return np.nan

df['Area_sqft'] = df['Area'].apply(area_to_sqft)

# Show after cleaning
print("\nAfter Area Cleaning:")
df[['Area', 'Area_sqft']].head()
```



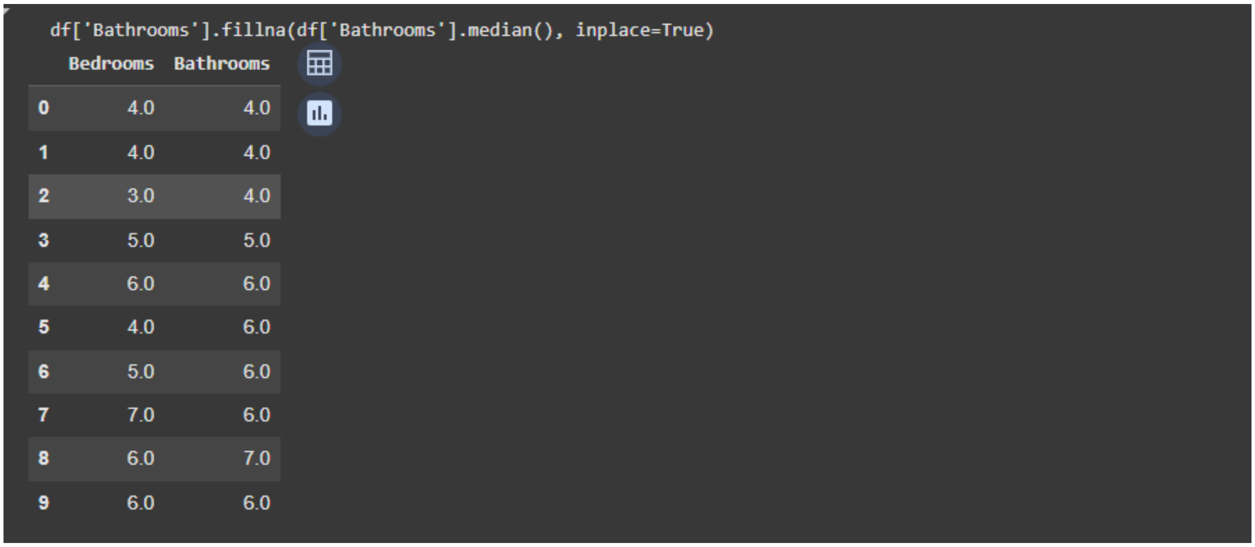
```
Before Area Cleaning:
0    5 Marla
1    5 Marla
2    5 Marla
3    5 Marla
4    1 Kanal
Name: Area, dtype: object

After Area Cleaning:
       Area  Area_sqft
0   5 Marla    1361.25
1   5 Marla    1361.25
2   5 Marla    1361.25
3   5 Marla    1361.25
4   1 Kanal    5445.00
```

Handling Missing Bedrooms and Bathroom Numbers:

The bathrooms and Bedrooms columns were converted into type numeric and then the missing values were replaced with the median of the values for bathrooms and bedrooms.



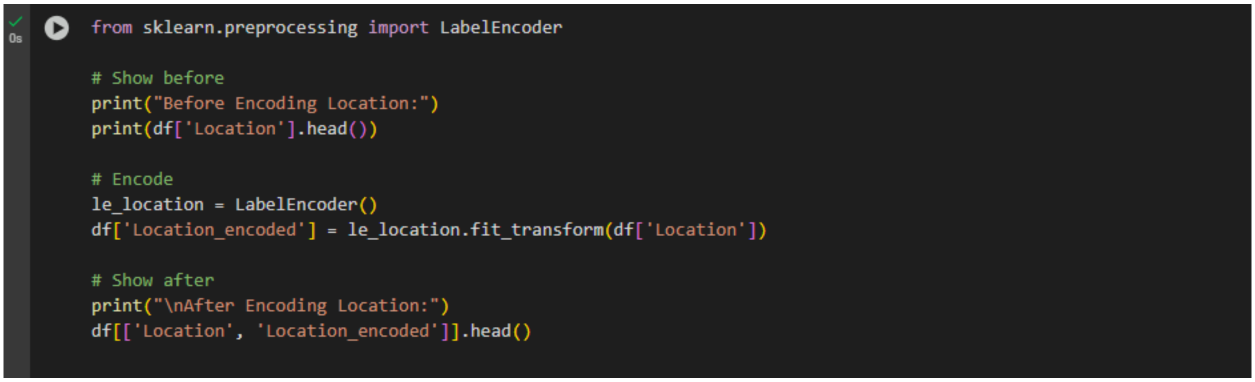
```python
# Show before
print("Before Handling Missing Values:")
print(df[['Bedrooms', 'Bathrooms']].head(10))

# Convert to numeric
df['Bedrooms'] = pd.to_numeric(df['Bedrooms'], errors='coerce')
df['Bathrooms'] = pd.to_numeric(df['Bathrooms'], errors='coerce')

# Fill missing with median
df['Bedrooms'].fillna(df['Bedrooms'].median(), inplace=True)
df['Bathrooms'].fillna(df['Bathrooms'].median(), inplace=True)

# Show after
print("\nAfter Handling Missing Values:")
df[['Bedrooms', 'Bathrooms']].head(10)
```



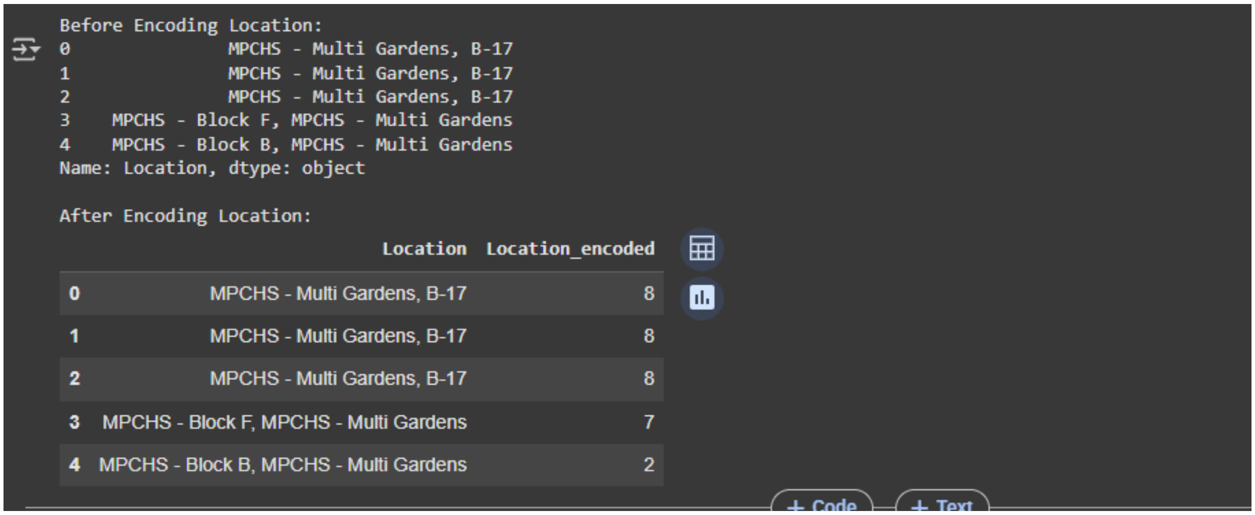
```
df['Bathrooms'].fillna(df['Bathrooms'].median(), inplace=True)
```

	Bedrooms	Bathrooms
0	4.0	4.0
1	4.0	4.0
2	3.0	4.0
3	5.0	5.0
4	6.0	6.0
5	4.0	6.0
6	5.0	6.0
7	7.0	6.0
8	6.0	7.0
9	6.0	6.0

Encoding the Location:



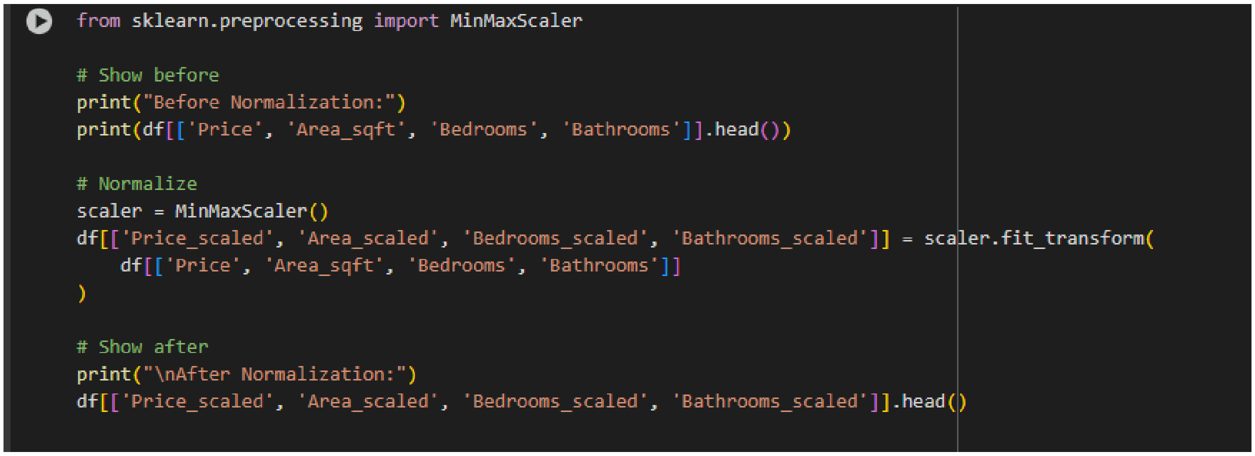
```
from sklearn.preprocessing import LabelEncoder

# Show before
print("Before Encoding Location:")
print(df['Location'].head())

# Encode
le_location = LabelEncoder()
df['Location_encoded'] = le_location.fit_transform(df['Location'])

# Show after
print("\nAfter Encoding Location:")
df[['Location', 'Location_encoded']].head()
```



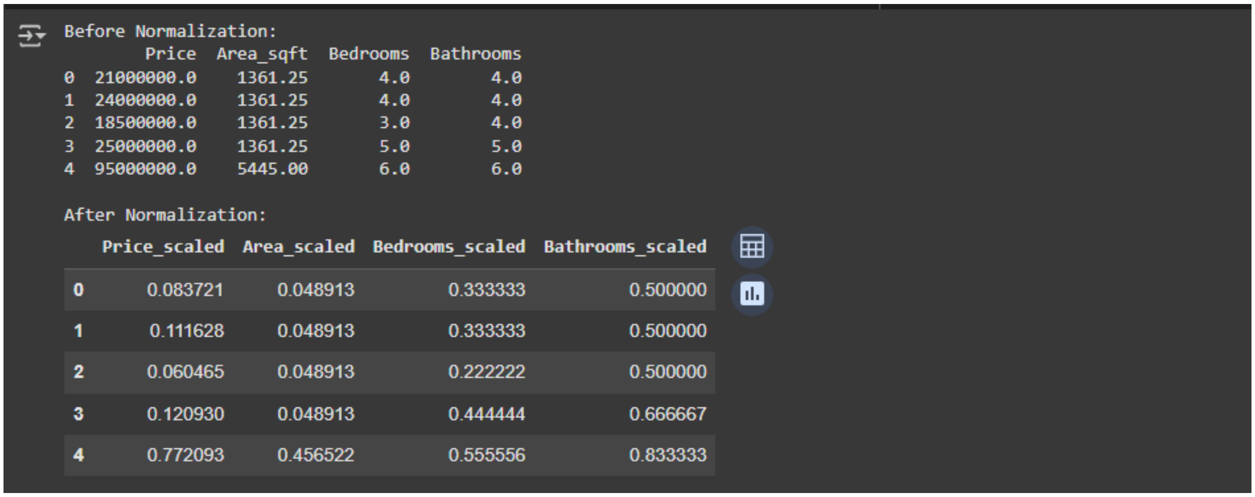
```
Before Encoding Location:
0              MPCHS - Multi Gardens, B-17
1              MPCHS - Multi Gardens, B-17
2              MPCHS - Multi Gardens, B-17
3    MPCHS - Block F, MPCHS - Multi Gardens
4    MPCHS - Block B, MPCHS - Multi Gardens
Name: Location, dtype: object

After Encoding Location:
```

	Location	Location_encoded
0	MPCHS - Multi Gardens, B-17	8
1	MPCHS - Multi Gardens, B-17	8
2	MPCHS - Multi Gardens, B-17	8
3	MPCHS - Block F, MPCHS - Multi Gardens	7
4	MPCHS - Block B, MPCHS - Multi Gardens	2

Normalizing the Columns Area, Price, Bedrooms and Bathrooms:



```
from sklearn.preprocessing import MinMaxScaler

# Show before
print("Before Normalization:")
print(df[['Price', 'Area_sqft', 'Bedrooms', 'Bathrooms']].head())

# Normalize
scaler = MinMaxScaler()
df[['Price_scaled', 'Area_scaled', 'Bedrooms_scaled', 'Bathrooms_scaled']] = scaler.fit_transform(
    df[['Price', 'Area_sqft', 'Bedrooms', 'Bathrooms']]
)

# Show after
print("\nAfter Normalization:")
df[['Price_scaled', 'Area_scaled', 'Bedrooms_scaled', 'Bathrooms_scaled']].head()
```

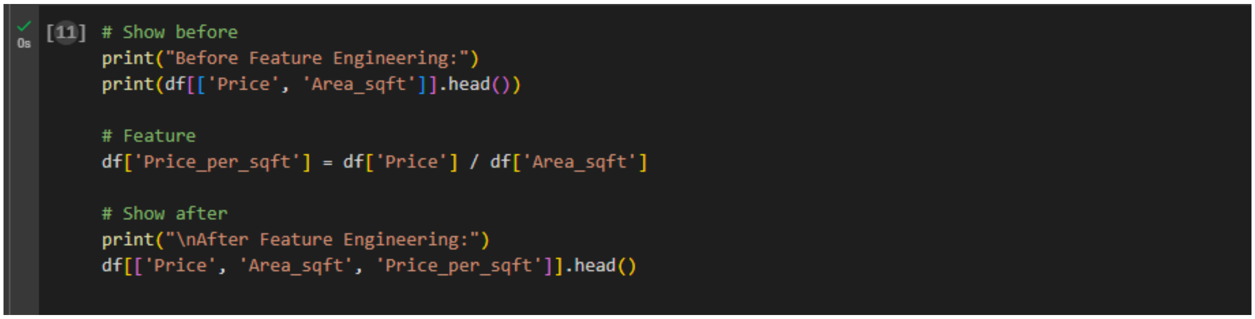


```
Before Normalization:
        Price  Area_sqft  Bedrooms  Bathrooms
0  21000000.0    1361.25       4.0        4.0
1  24000000.0    1361.25       4.0        4.0
2  18500000.0    1361.25       3.0        4.0
3  25000000.0    1361.25       5.0        5.0
4  95000000.0    5445.00       6.0        6.0

After Normalization:
```

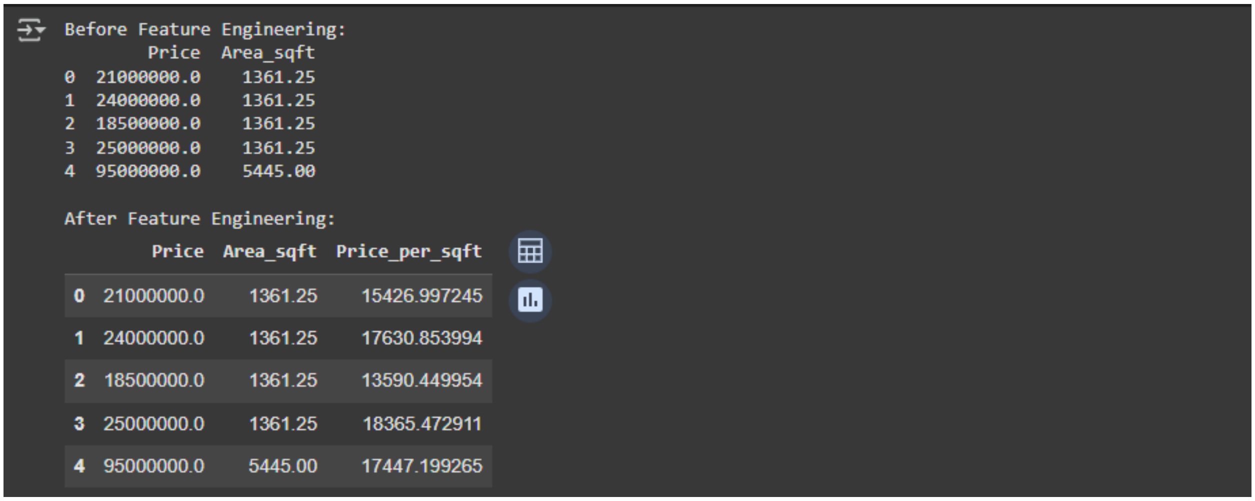
	Price_scaled	Area_scaled	Bedrooms_scaled	Bathrooms_scaled
0	0.083721	0.048913	0.333333	0.500000
1	0.111628	0.048913	0.333333	0.500000
2	0.060465	0.048913	0.222222	0.500000
3	0.120930	0.048913	0.444444	0.666667
4	0.772093	0.456522	0.555556	0.833333

Feature Engineering:

I created another column named Area_sqft in which I calculated the Price per Square feet of area. This gives a better understanding and comparison of the prices.



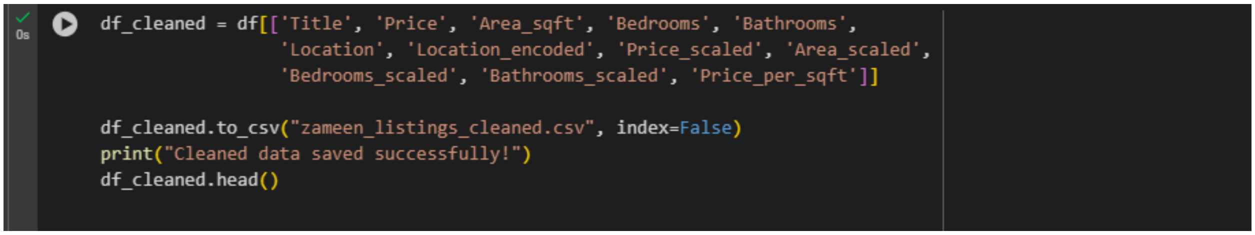
```
[11] # Show before
     print("Before Feature Engineering:")
     print(df[['Price', 'Area_sqft']].head())

     # Feature
     df['Price_per_sqft'] = df['Price'] / df['Area_sqft']

     # Show after
     print("\nAfter Feature Engineering:")
     df[['Price', 'Area_sqft', 'Price_per_sqft']].head()
```



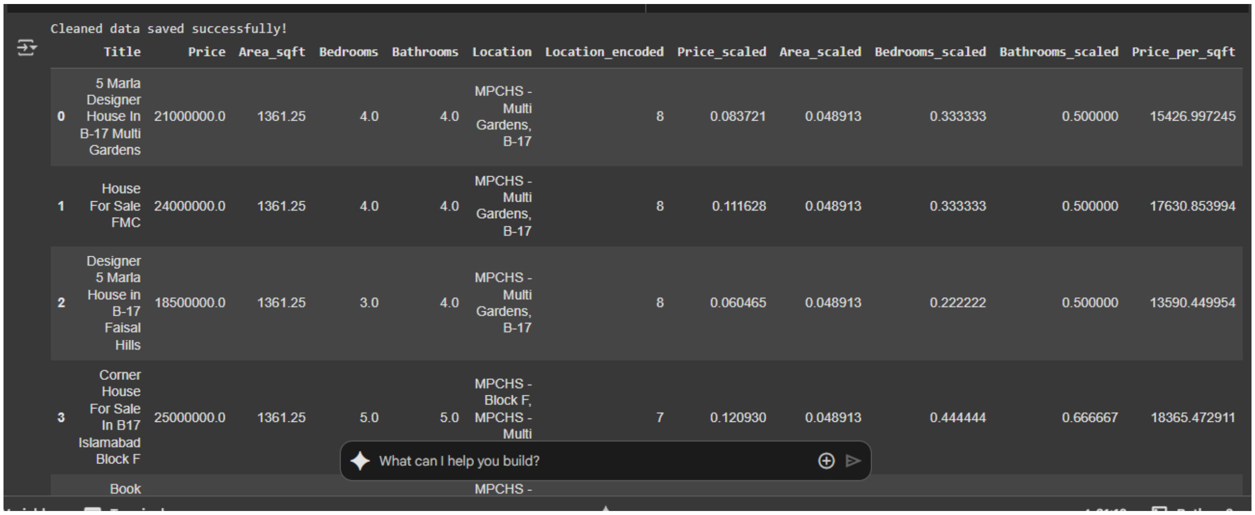
```
Before Feature Engineering:
        Price  Area_sqft
0  21000000.0    1361.25
1  24000000.0    1361.25
2  18500000.0    1361.25
3  25000000.0    1361.25
4  95000000.0    5445.00

After Feature Engineering:
```

	Price	Area_sqft	Price_per_sqft
0	21000000.0	1361.25	15426.997245
1	24000000.0	1361.25	17630.853994
2	18500000.0	1361.25	13590.449954
3	25000000.0	1361.25	18365.472911
4	95000000.0	5445.00	17447.199265

Cleaned Dataset:



```python
df_cleaned = df[['Title', 'Price', 'Area_sqft', 'Bedrooms', 'Bathrooms',
                 'Location', 'Location_encoded', 'Price_scaled', 'Area_scaled',
                 'Bedrooms_scaled', 'Bathrooms_scaled', 'Price_per_sqft']]

df_cleaned.to_csv("zameen_listings_cleaned.csv", index=False)
print("Cleaned data saved successfully!")
df_cleaned.head()
```

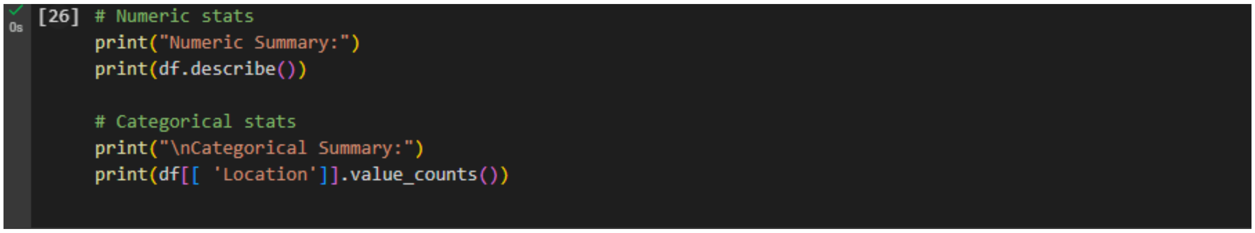


```
Cleaned data saved successfully!
```

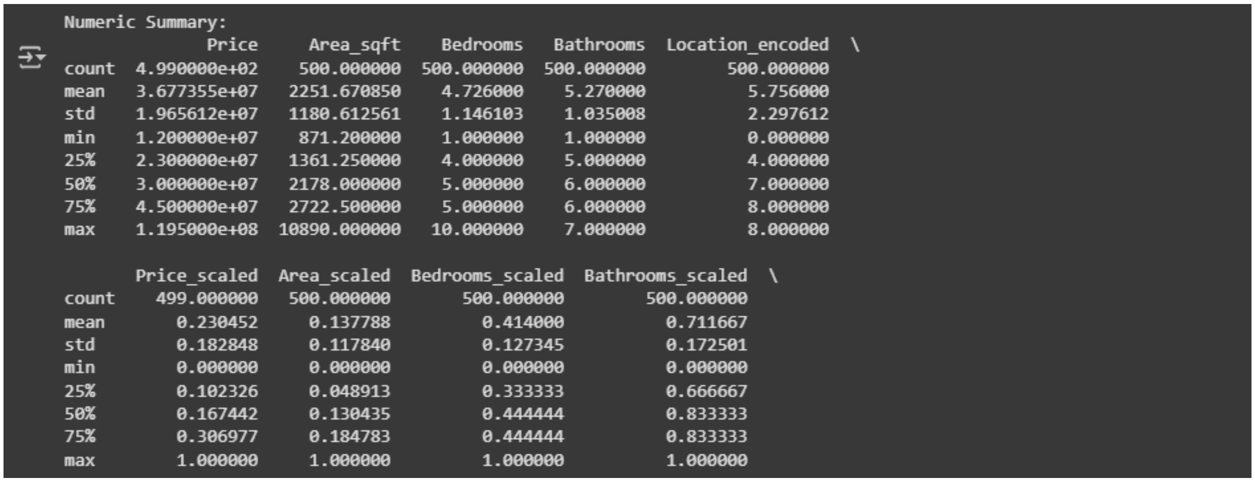
	Title	Price	Area_sqft	Bedrooms	Bathrooms	Location	Location_encoded	Price_scaled	Area_scaled	Bedrooms_scaled	Bathrooms_scaled	Price_per_sqft
0	5 Marla Designer House In B-17 Multi Gardens	21000000.0	1361.25	4.0	4.0	MPCHS - Multi Gardens, B-17	8	0.083721	0.048913	0.333333	0.500000	15426.997245
1	House For Sale FMC	24000000.0	1361.25	4.0	4.0	MPCHS - Multi Gardens, B-17	8	0.111628	0.048913	0.333333	0.500000	17630.853994
2	Designer 5 Marla House in B-17 Faisal Hills	18500000.0	1361.25	3.0	4.0	MPCHS - Multi Gardens, B-17	8	0.060465	0.048913	0.222222	0.500000	13590.449954
3	Corner House For Sale In B17 Islamabad Block F	25000000.0	1361.25	5.0	5.0	MPCHS - Block F, MPCHS - Multi	7	0.120930	0.048913	0.444444	0.666667	18365.472911
	Book					MPCHS -						

DATA VISUALIZATION:

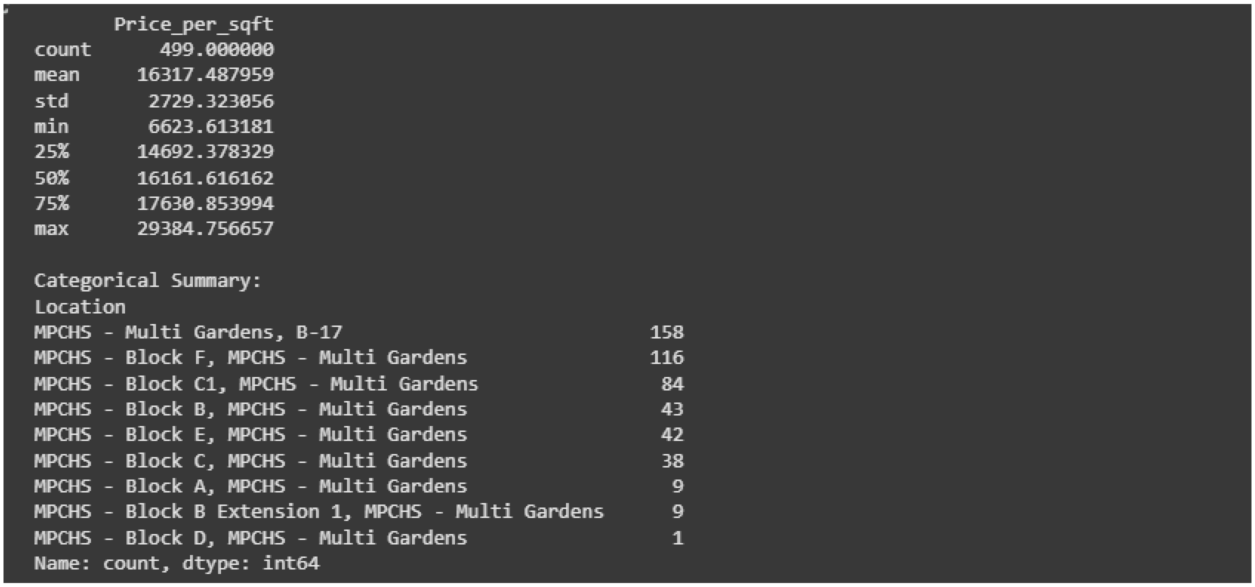
Having a look at the data Description:



```
[26] # Numeric stats
     print("Numeric Summary:")
     print(df.describe())

     # Categorical stats
     print("\nCategorical Summary:")
     print(df[[ 'Location']].value_counts())
```

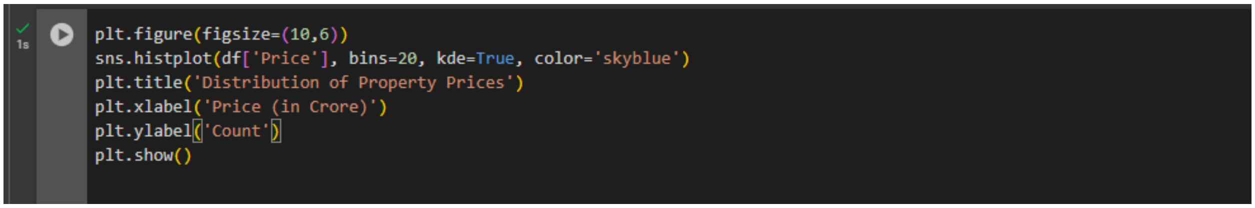


```
Numeric Summary:
              Price     Area_sqft    Bedrooms   Bathrooms  Location_encoded  \
count  4.990000e+02    500.000000  500.000000  500.000000        500.000000
mean   3.677355e+07   2251.670850    4.726000    5.270000          5.756000
std    1.965612e+07   1180.612561    1.146103    1.035008          2.297612
min    1.200000e+07    871.200000    1.000000    1.000000          0.000000
25%    2.300000e+07   1361.250000    4.000000    5.000000          4.000000
50%    3.000000e+07   2178.000000    5.000000    6.000000          7.000000
75%    4.500000e+07   2722.500000    5.000000    6.000000          8.000000
max    1.195000e+08  10890.000000   10.000000    7.000000          8.000000

       Price_scaled  Area_scaled  Bedrooms_scaled  Bathrooms_scaled  \
count    499.000000   500.000000       500.000000        500.000000
mean       0.230452     0.137788         0.414000          0.711667
std        0.182848     0.117840         0.127345          0.172501
min        0.000000     0.000000         0.000000          0.000000
25%        0.102326     0.048913         0.333333          0.666667
50%        0.167442     0.130435         0.444444          0.833333
75%        0.306977     0.184783         0.444444          0.833333
max        1.000000     1.000000         1.000000          1.000000
```



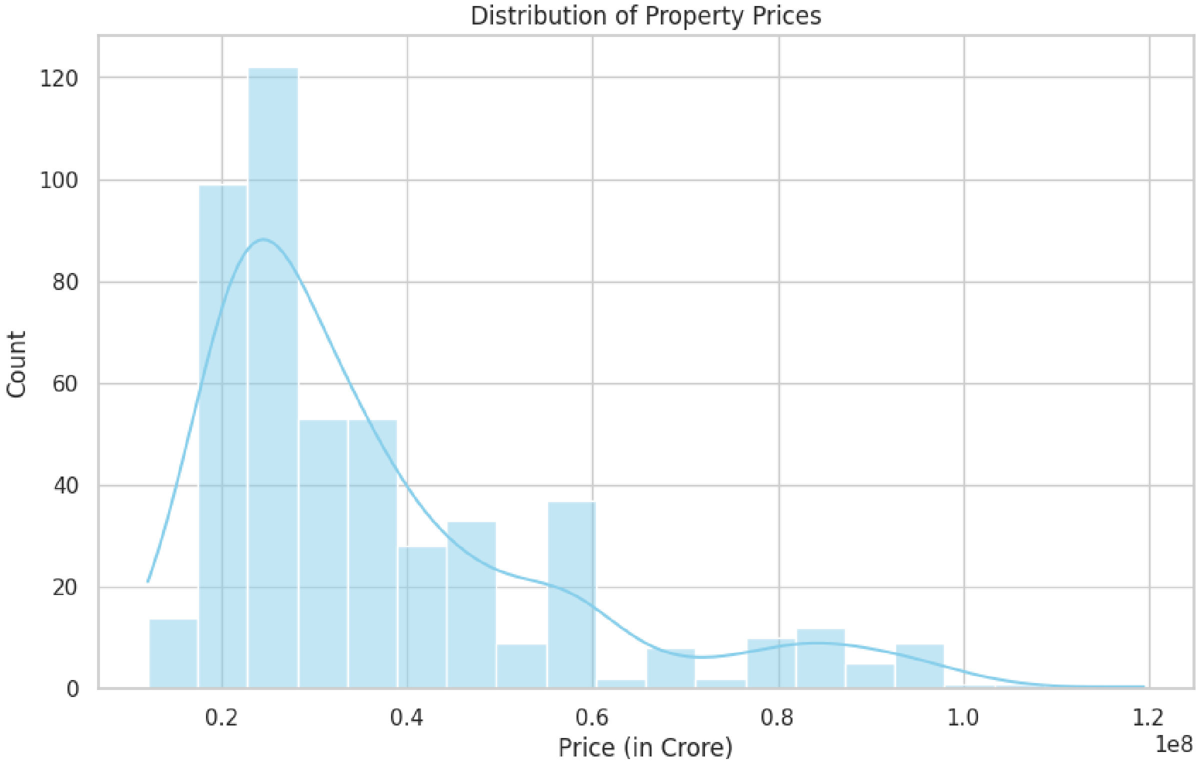
```
       Price_per_sqft
count      499.000000
mean     16317.487959
std       2729.323056
min       6623.613181
25%      14692.378329
50%      16161.616162
75%      17630.853994
max      29384.756657

Categorical Summary:
Location
MPCHS - Multi Gardens, B-17                              158
MPCHS - Block F, MPCHS - Multi Gardens                   116
MPCHS - Block C1, MPCHS - Multi Gardens                   84
MPCHS - Block B, MPCHS - Multi Gardens                    43
MPCHS - Block E, MPCHS - Multi Gardens                    42
MPCHS - Block C, MPCHS - Multi Gardens                    38
MPCHS - Block A, MPCHS - Multi Gardens                     9
MPCHS - Block B Extension 1, MPCHS - Multi Gardens         9
MPCHS - Block D, MPCHS - Multi Gardens                     1
Name: count, dtype: int64
```

Price vs Number of property:

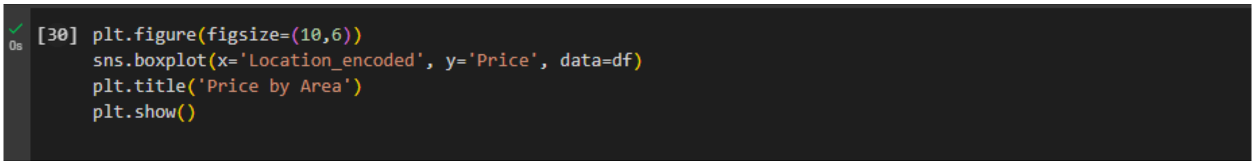


```
plt.figure(figsize=(10,6))
sns.histplot(df['Price'], bins=20, kde=True, color='skyblue')
plt.title('Distribution of Property Prices')
plt.xlabel('Price (in Crore)')
plt.ylabel('Count')
plt.show()
```

The graph shows that most of the property prices range between 0.2 to 0.4 crores. Avery few are listed at more that 1 crore.

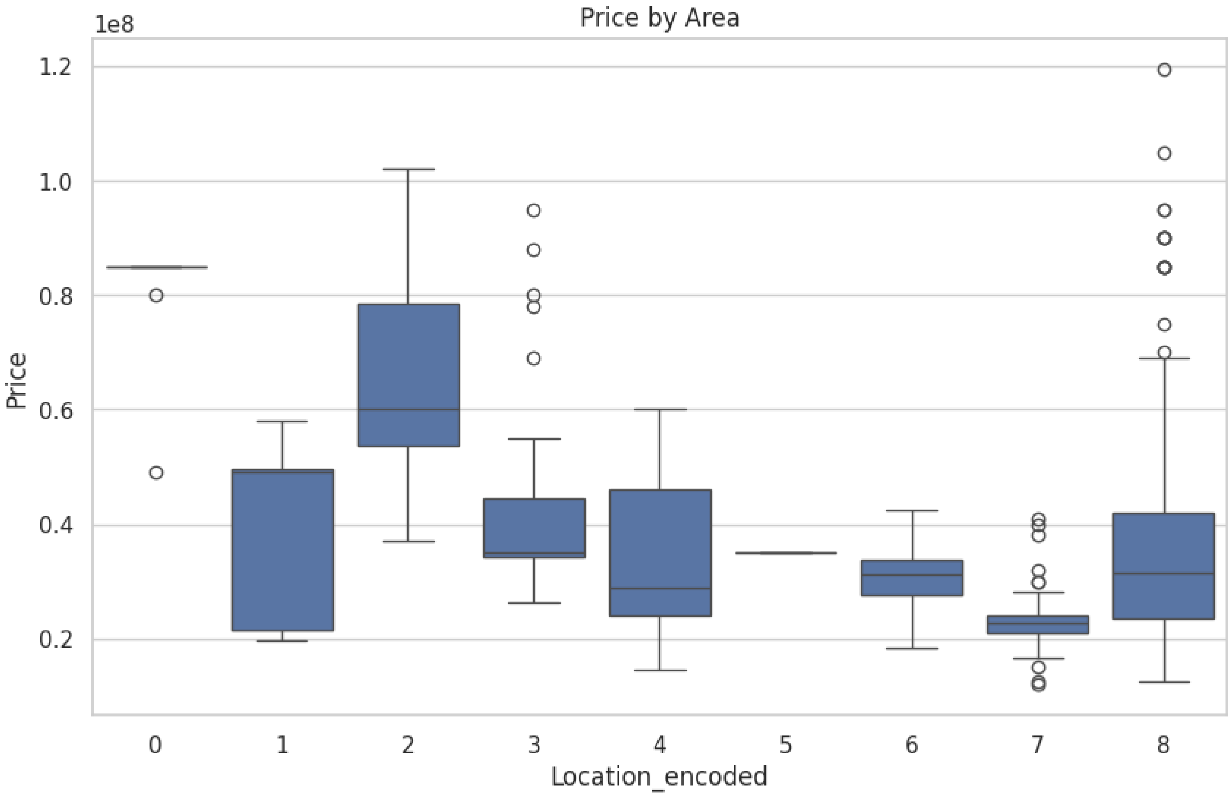


Location VS Price Graph:

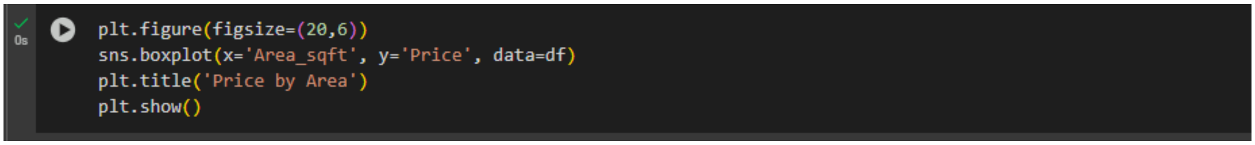


```
plt.figure(figsize=(10,6))
sns.boxplot(x='Location_encoded', y='Price', data=df)
plt.title('Price by Area')
plt.show()
```

The graph shows that the area encoded with the value 2 has the highest Mean Price value. As for Location 8, It holds the maximum property price value.

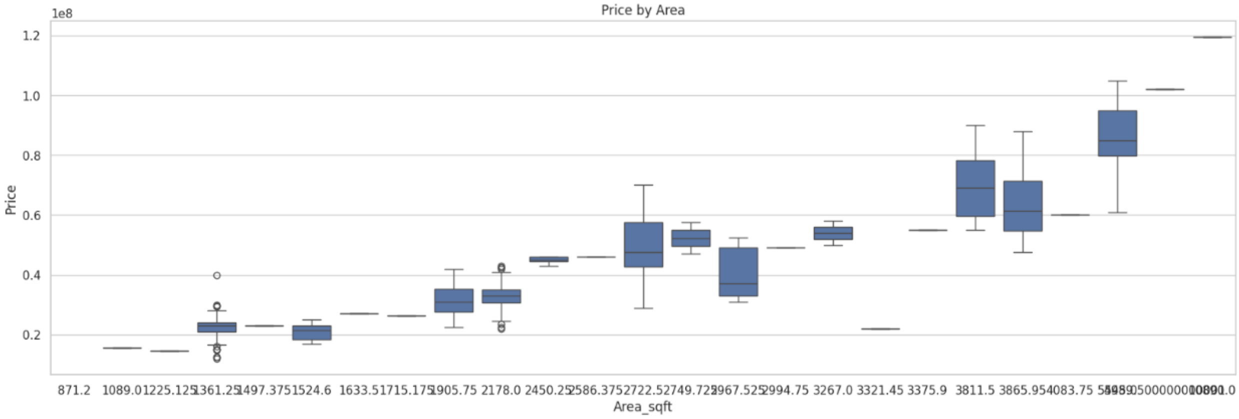


Area in sqft Vs Price Graph:

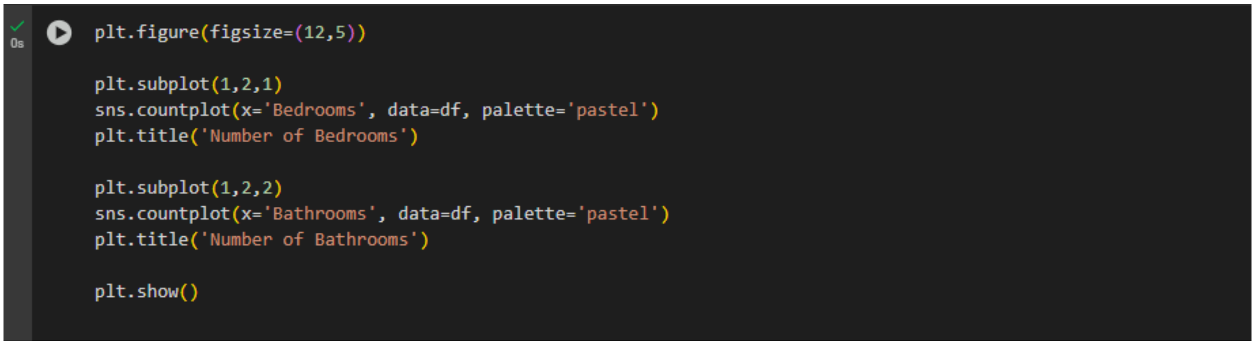


```
plt.figure(figsize=(20,6))
sns.boxplot(x='Area_sqft', y='Price', data=df)
plt.title('Price by Area')
plt.show()
```

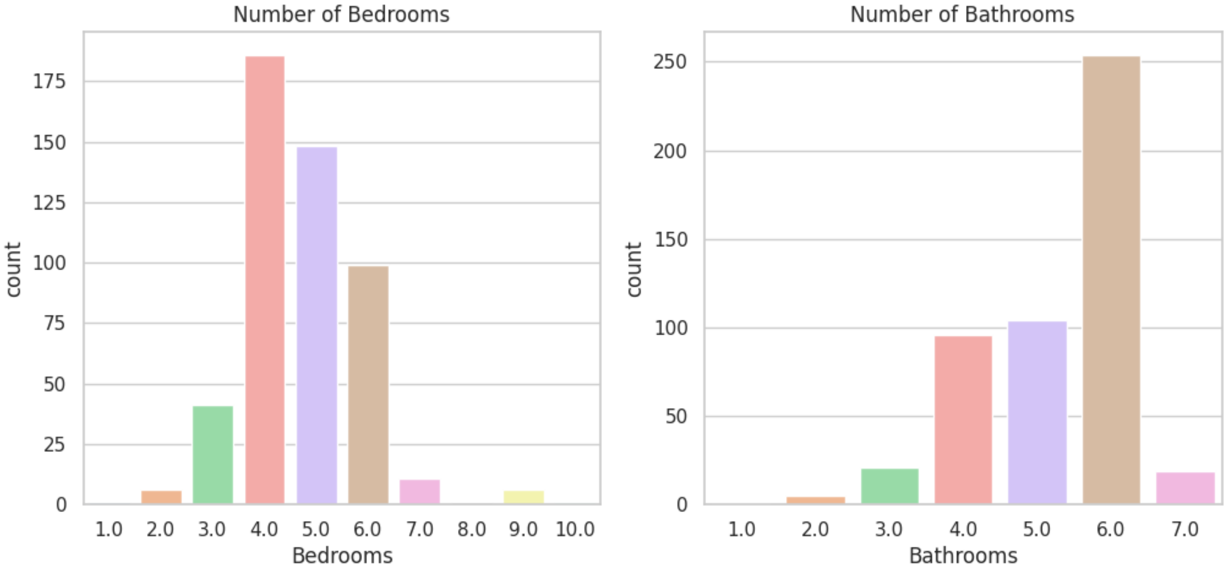
The graph shows that with the increase in the area, The prices also increases significantly.



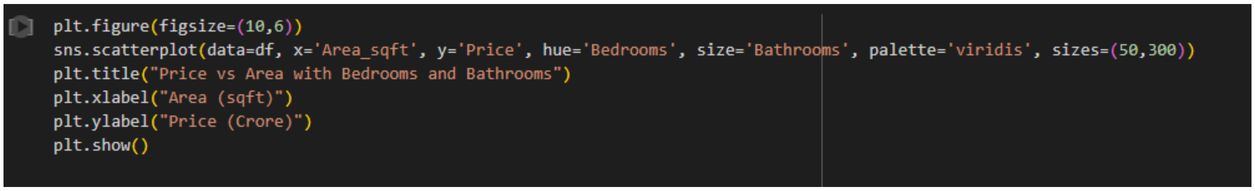
Looking at the bedroom and bathroom stats:



```
plt.figure(figsize=(12,5))

plt.subplot(1,2,1)
sns.countplot(x='Bedrooms', data=df, palette='pastel')
plt.title('Number of Bedrooms')

plt.subplot(1,2,2)
sns.countplot(x='Bathrooms', data=df, palette='pastel')
plt.title('Number of Bathrooms')

plt.show()
```

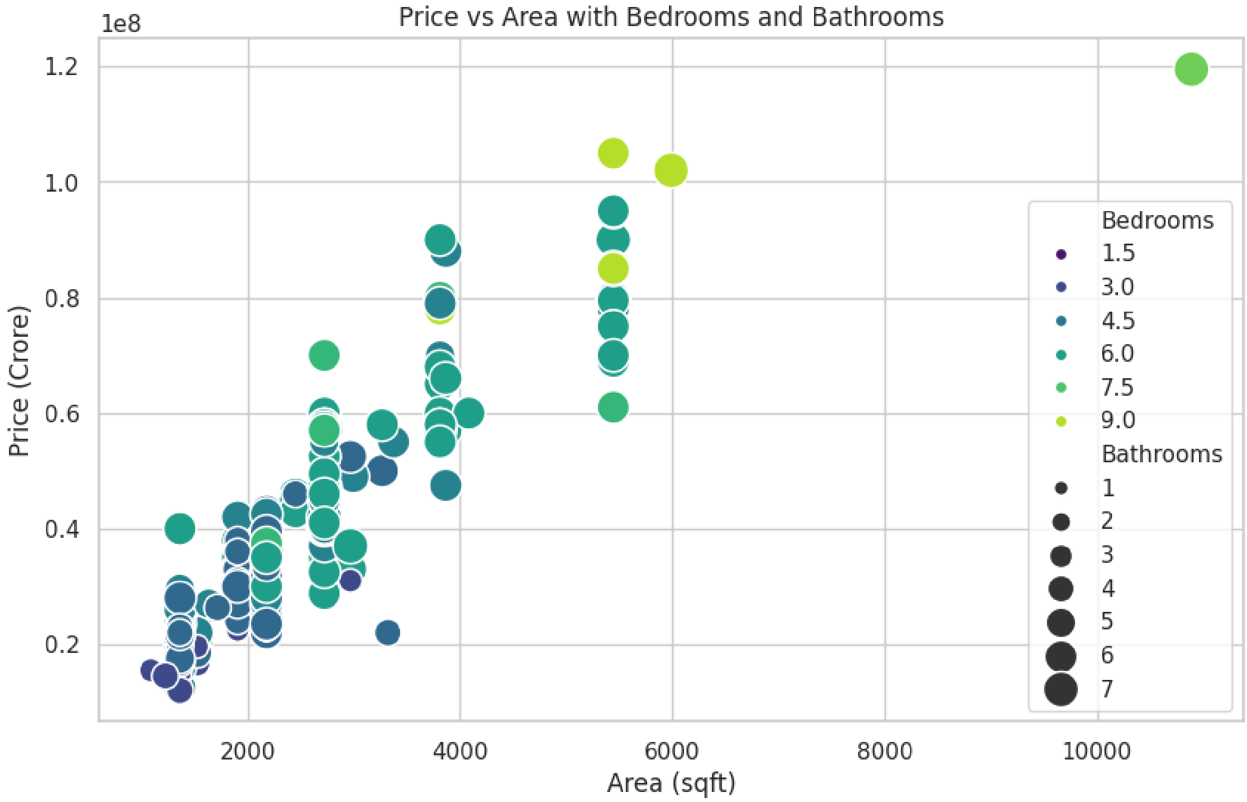


Price Vs Area with Bedrooms and Bathrooms graph:

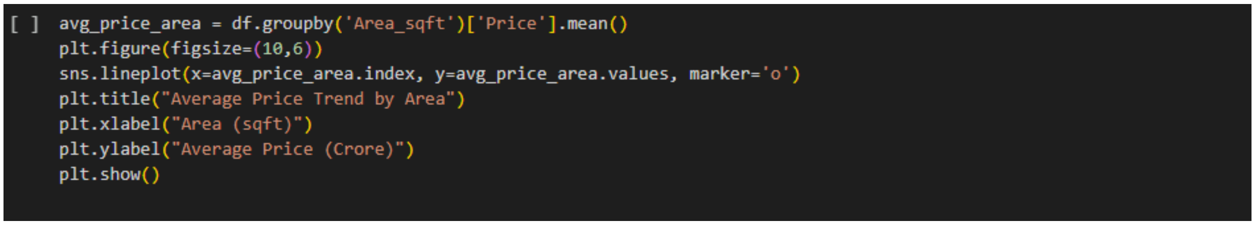


```
plt.figure(figsize=(10,6))
sns.scatterplot(data=df, x='Area_sqft', y='Price', hue='Bedrooms', size='Bathrooms', palette='viridis', sizes=(50,300))
plt.title("Price vs Area with Bedrooms and Bathrooms")
plt.xlabel("Area (sqft)")
plt.ylabel("Price (Crore)")
plt.show()
```

The graph shows that with the increase in Area the price Increases. Also for the same Area in Square Feet, The number of Beddrooms also play a role in increasing the price of the house.

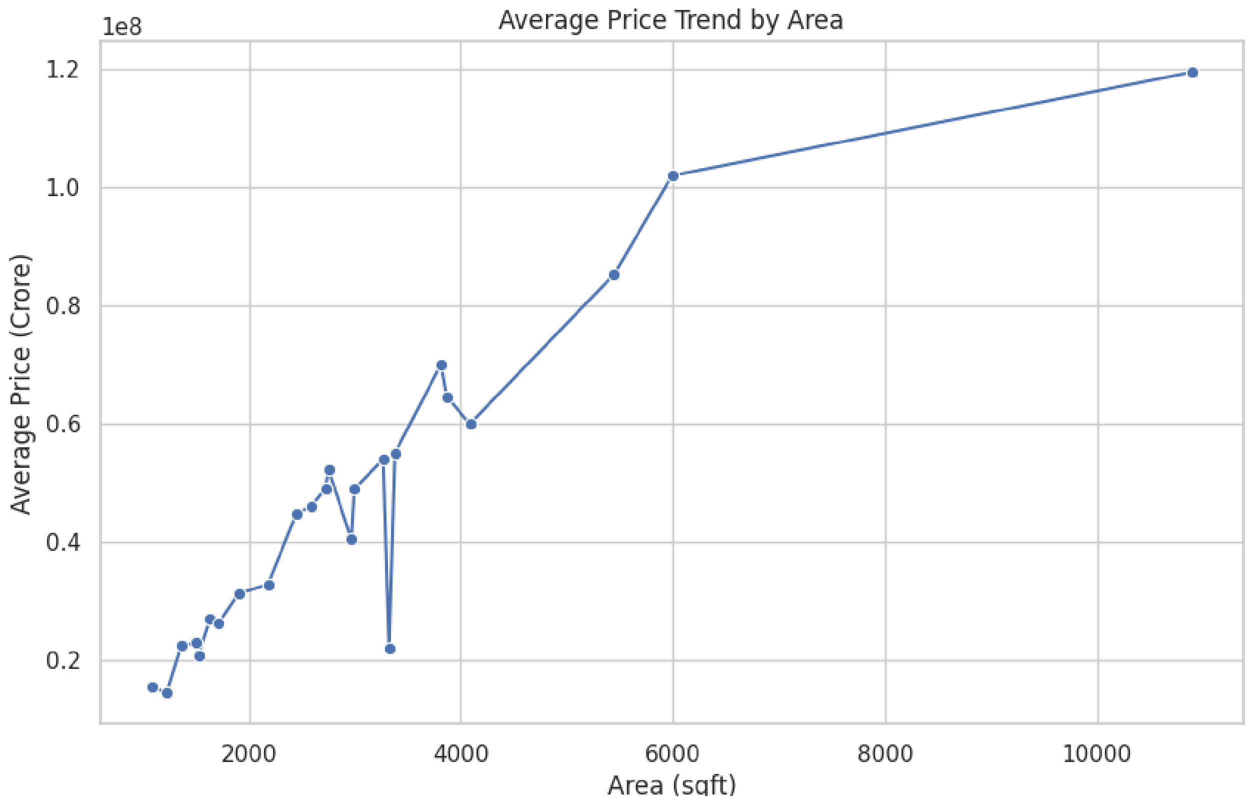


Average Price Vs Area graph:

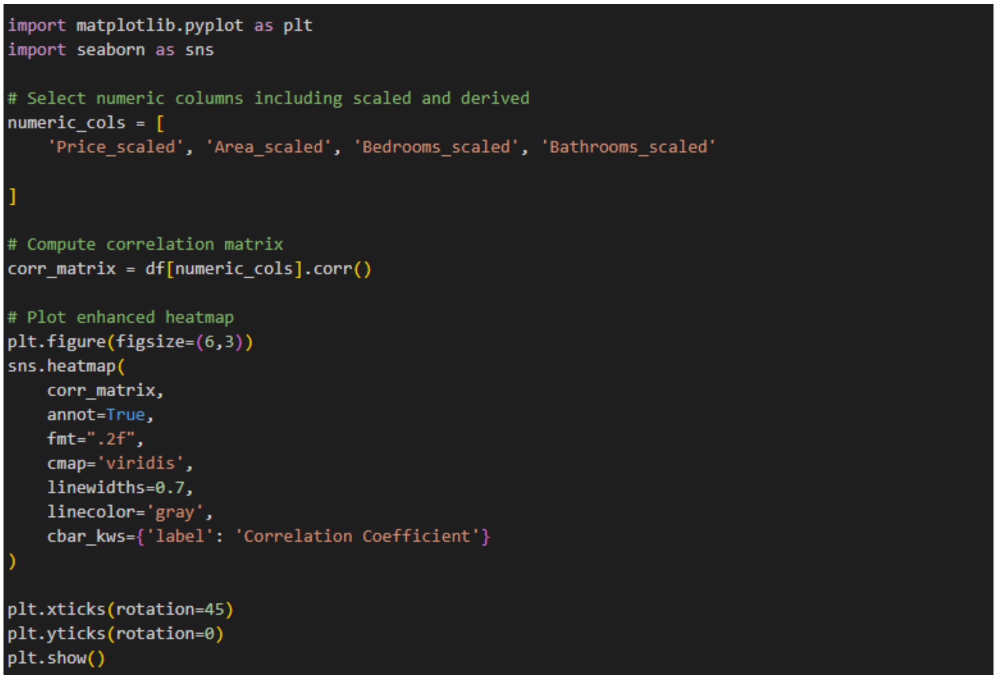


```
[ ]  avg_price_area = df.groupby('Area_sqft')['Price'].mean()
     plt.figure(figsize=(10,6))
     sns.lineplot(x=avg_price_area.index, y=avg_price_area.values, marker='o')
     plt.title("Average Price Trend by Area")
     plt.xlabel("Area (sqft)")
     plt.ylabel("Average Price (Crore)")
     plt.show()
```

The graph shows that with the increase in Area, the price increases significantly.

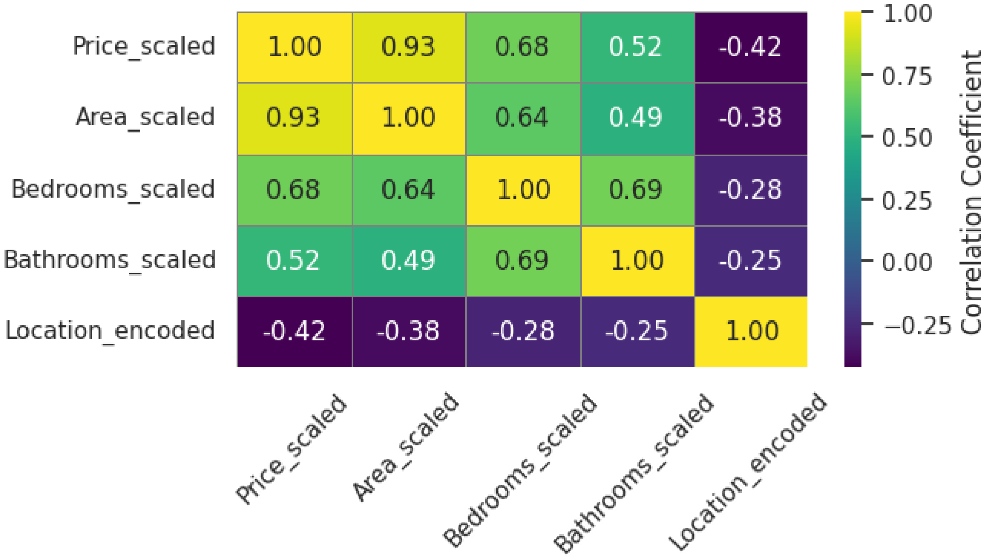


Correlation Matrix:



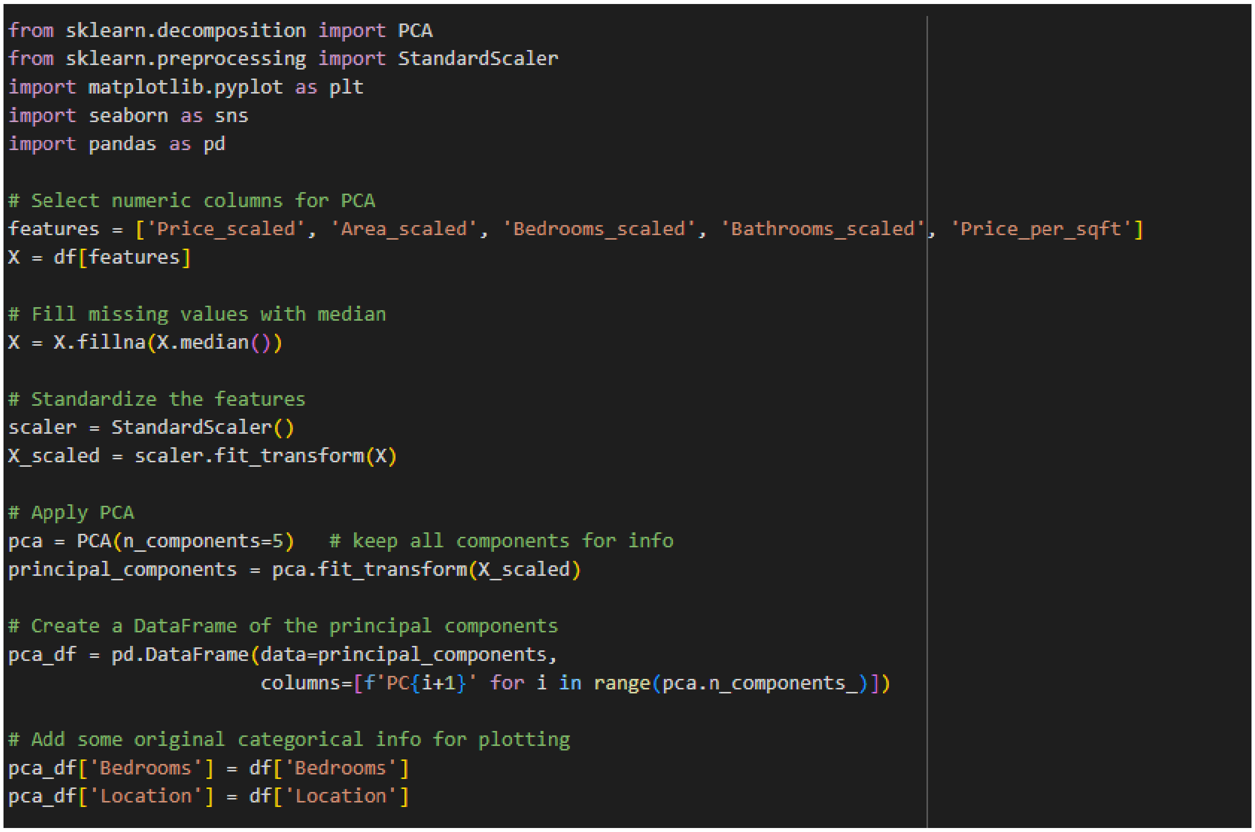
```
import matplotlib.pyplot as plt
import seaborn as sns

# Select numeric columns including scaled and derived
numeric_cols = [
    'Price_scaled', 'Area_scaled', 'Bedrooms_scaled', 'Bathrooms_scaled'

]

# Compute correlation matrix
corr_matrix = df[numeric_cols].corr()

# Plot enhanced heatmap
plt.figure(figsize=(6,3))
sns.heatmap(
    corr_matrix,
    annot=True,
    fmt=".2f",
    cmap='viridis',
    linewidths=0.7,
    linecolor='gray',
    cbar_kws={'label': 'Correlation Coefficient'}
)

plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.show()
```

The correlation Matrix shows that the price is highly correlated to the area, the Number of bedrooms and then bathrooms. The fact that Price is inversely correlated to Location is due to the encoding of the places from 1 to 8. The location with the encoding 8 holds the properties with the highest values. So the location and the prices are inversely proportional according to my understanding.

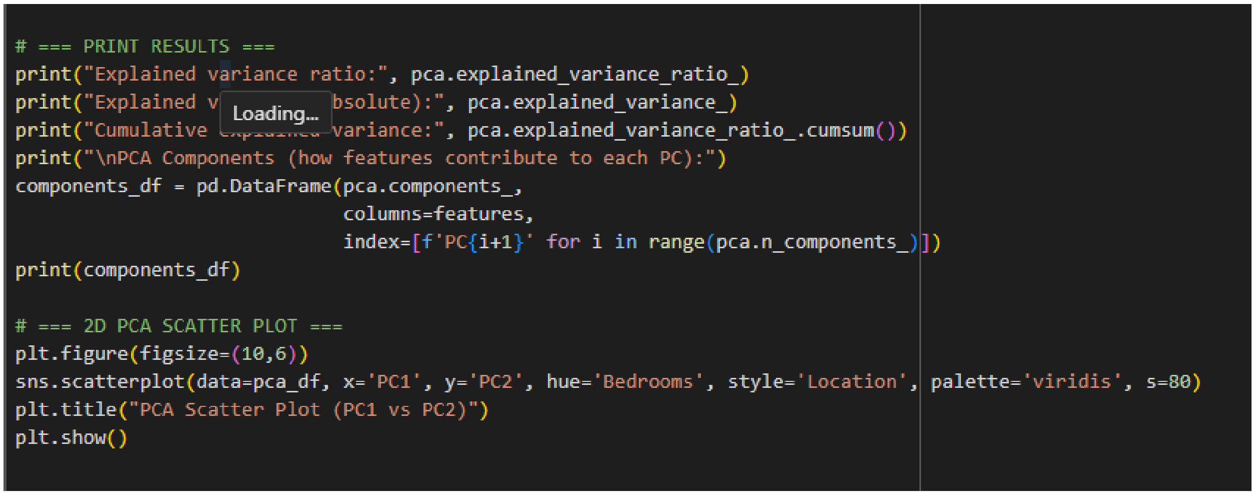


	Price_scaled	Area_scaled	Bedrooms_scaled	Bathrooms_scaled	Location_encoded
Price_scaled	1.00	0.93	0.68	0.52	-0.42
Area_scaled	0.93	1.00	0.64	0.49	-0.38
Bedrooms_scaled	0.68	0.64	1.00	0.69	-0.28
Bathrooms_scaled	0.52	0.49	0.69	1.00	-0.25
Location_encoded	-0.42	-0.38	-0.28	-0.25	1.00

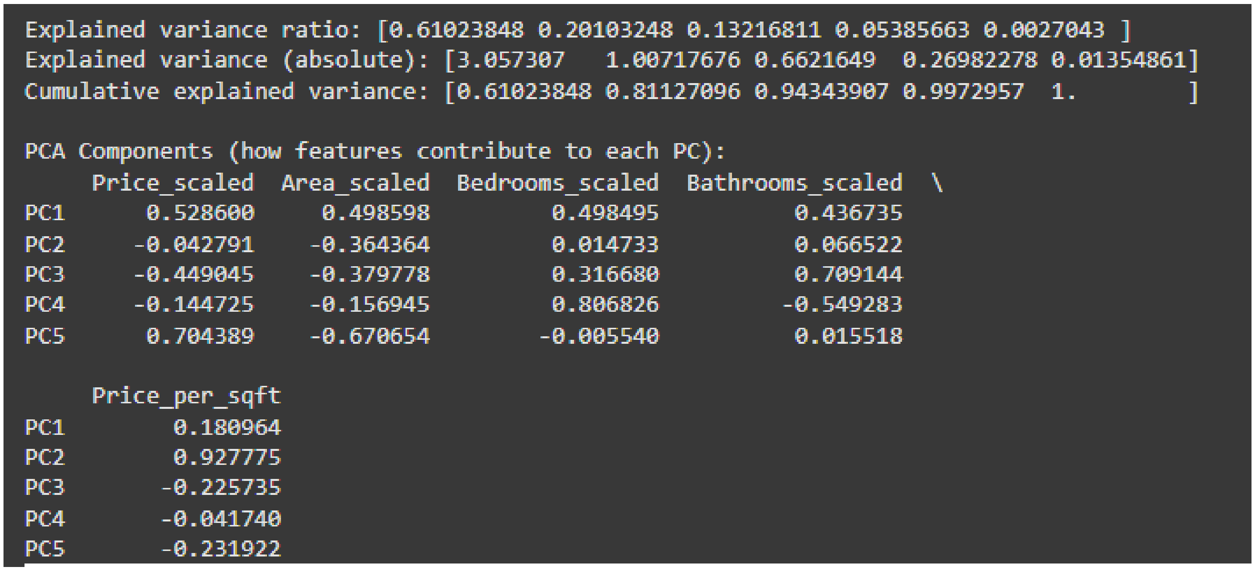
Principal Component Analysis and Dimensionality Reduction:



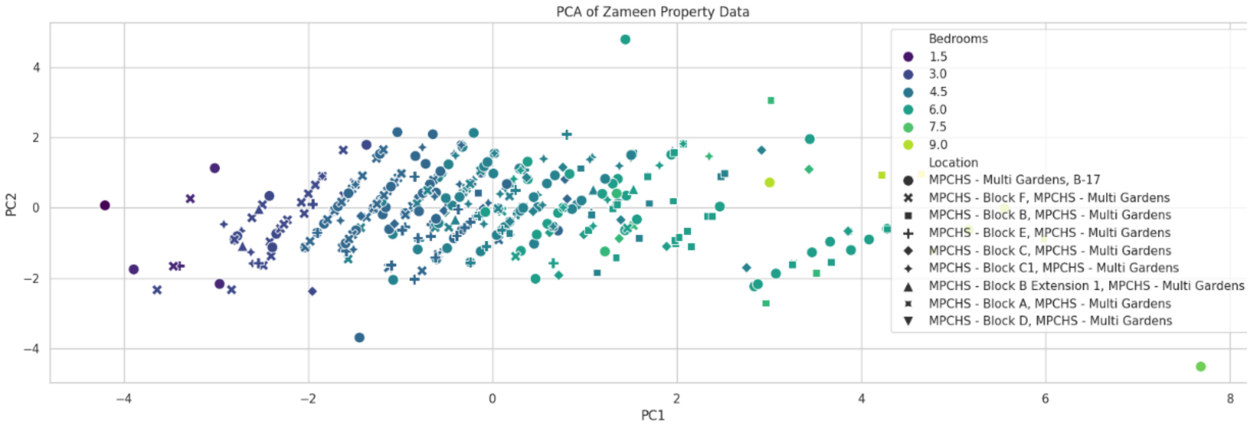
```python
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Select numeric columns for PCA
features = ['Price_scaled', 'Area_scaled', 'Bedrooms_scaled', 'Bathrooms_scaled', 'Price_per_sqft']
X = df[features]

# Fill missing values with median
X = X.fillna(X.median())

# Standardize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Apply PCA
pca = PCA(n_components=5)   # keep all components for info
principal_components = pca.fit_transform(X_scaled)

# Create a DataFrame of the principal components
pca_df = pd.DataFrame(data=principal_components,
                      columns=[f'PC{i+1}' for i in range(pca.n_components_)])

# Add some original categorical info for plotting
pca_df['Bedrooms'] = df['Bedrooms']
pca_df['Location'] = df['Location']
```



```python
# === PRINT RESULTS ===
print("Explained variance ratio:", pca.explained_variance_ratio_)
print("Explained v Loading... bsolute):", pca.explained_variance_)
print("Cumulative explained variance:", pca.explained_variance_ratio_.cumsum())
print("\nPCA Components (how features contribute to each PC):")
components_df = pd.DataFrame(pca.components_,
                             columns=features,
                             index=[f'PC{i+1}' for i in range(pca.n_components_)])
print(components_df)

# === 2D PCA SCATTER PLOT ===
plt.figure(figsize=(10,6))
sns.scatterplot(data=pca_df, x='PC1', y='PC2', hue='Bedrooms', style='Location', palette='viridis', s=80)
plt.title("PCA Scatter Plot (PC1 vs PC2)")
plt.show()
```



```
Explained variance ratio: [0.61023848 0.20103248 0.13216811 0.05385663 0.0027043 ]
Explained variance (absolute): [3.057307   1.00717676 0.6621649  0.26982278 0.01354861]
Cumulative explained variance: [0.61023848 0.81127096 0.94343907 0.9972957  1.        ]

PCA Components (how features contribute to each PC):
     Price_scaled  Area_scaled  Bedrooms_scaled  Bathrooms_scaled  \
PC1      0.528600     0.498598         0.498495          0.436735
PC2     -0.042791    -0.364364         0.014733          0.066522
PC3     -0.449045    -0.379778         0.316680          0.709144
PC4     -0.144725    -0.156945         0.806826         -0.549283
PC5      0.704389    -0.670654        -0.005540          0.015518

     Price_per_sqft
PC1        0.180964
PC2        0.927775
PC3       -0.225735
PC4       -0.041740
PC5       -0.231922
```



PC1 alone explains ~61% of the total variance.

PC1 + PC2 explains ~81% → already a strong dimensionality reduction.

PC1 + PC2 + PC3 explains ~94% → which is excellent as almost all the data’s structure is preserved.

PC4 and PC5 contribute very little.

PC1 is the Price, PC2 is the area and PC3 is the Number of bedrooms.

This means that our dataset can be safely reduced from 5D → 3D without losing much information.