Name: Muhammad Saad

Cohort 5 AWS

Report On EDA project

---

1. Introduction

This report presents an exploratory data analysis (EDA) of property listings scraped from Zameen.com — Pakistan's leading online real estate platform. The primary goal is to derive actionable insights related to pricing trends, area preferences, property types, and neighborhood dynamics across Pakistan. These insights are intended to guide real estate investors and stakeholders in making data-driven decisions.

---

2. Objective

To investigate the core question:

What factors influence property prices in Pakistan's real estate market?

Specific sub-objectives include:

- Identifying pricing trends across major cities
- Analyzing listing distribution and preferences
- Evaluating relationships between area, price, and property characteristics
- Detecting price-efficiency hotspots (price per sqft)

---

3. Dataset Overview

The dataset contains **18,031 listings** and includes the following features:

- **Basic Info:** Title, City, Type, Area, Price, Purpose, Location
- **Descriptive:** Description text, Bedrooms, Bathrooms

---

4. Data Cleaning & Feature Engineering

Tasks Performed:

- **Price Standardization:** Converted non-numeric price formats (e.g., "2.5 Crore") to PKR values using a currency conversion logic.
- **Area Conversion:** Standardized all property area values to square feet using conversion units for Marla, Kanal, Sq. Yd., etc.
- **Missing Values:** Dropped rows with missing critical fields such as price or area.
- **New Feature:** Created Price_per_Sqft by dividing cleaned price by area.

Code of cleaning data

```
# Loading Excel
df = pd.read_excel("Scarped Zameen.com.xlsx")

# Selecting important columns
df = df[['Title', 'City', 'Type', 'Area', 'Price', 'Purpose',
'Location', 'Description', 'Bedrooms', 'Bathrooms']].copy()

# Price cleaning
def clean_price(price_str):
    if isinstance(price_str, str):
        price_str = price_str.replace('PKR', '').replace('\n',
'').replace(',', '').strip()
        if 'Arab' in price_str:
            return float(re.findall(r'\d+\.?\d*', price_str)[0]) *
1e9
        elif 'Crore' in price_str:
            return float(re.findall(r'\d+\.?\d*', price_str)[0]) *
1e7
        elif 'Lakh' in price_str:
            return float(re.findall(r'\d+\.?\d*', price_str)[0]) *
1e5
        elif 'Thousand' in price_str:
            return float(re.findall(r'\d+\.?\d*', price_str)[0]) *
1e3
        elif re.match(r'^\d+\.?\d*$', price_str):
            return float(price_str)
    return np.nan

df['Price_Clean'] = df['Price'].apply(clean_price)

# Area cleaning
unit_conversion = {
```

```
    'Sq. Yd.': 9.0,
    'Sq. Ft.': 1.0,
    'Marla': 272.25,
    'Kanal': 5445.0,
    'Sq. Meter': 10.7639
}

def convert_area(area_str):
    if isinstance(area_str, str):
        match = re.match(r'(\d+\.?\d*)\s*([A-Za-z. ]+)',
area_str.strip())
        if match:
            size = float(match.group(1))
            unit = match.group(2).strip()
            factor = unit_conversion.get(unit, np.nan)
            if not np.isnan(factor):
                return size * factor
    return np.nan

df['Area_Sqft'] = df['Area'].apply(convert_area)

# Converting beds & baths to numbers
df['Bedrooms'] = pd.to_numeric(df['Bedrooms'], errors='coerce')
df['Bathrooms'] = pd.to_numeric(df['Bathrooms'], errors='coerce')

# Price per square foot
df['Price_per_Sqft'] = df['Price_Clean'] / df['Area_Sqft']

# View summary
print(df[['Price_Clean', 'Area_Sqft', 'Price_per_Sqft']].describe())
```

Handling Missing Values

```
# Cleaning the 'Price' column
# We convert prices like 'PKR 25,000,000' → 25000000.0 (float)
# This method works for numeric-only values, not for "Crore/Lakh" text-based formats
df['Price_Clean'] = df['Price'].astype(str).str.replace(',', '').str.extract(r'(\d+\.?\d*)').astype(float)

#  Cleaning the 'Area' column
# Similar cleaning: extract numeric area value from strings like '1,200 Sq. Ft.' → 1200.0
df['Area_Sqft'] = df['Area'].astype(str).str.replace(',', '').str.extract(r'(\d+\.?\d*)').astype(float)

# Checking for missing values after cleaning
# Helps us understand which columns still have NaNs before analysis
print("Missing values after cleaning:\n")
print(df.isnull().sum())

#  Dropping rows with missing essential fields
# We remove listings with no price, area, bedroom, or bathroom info — they're not useful for analysis
df.dropna(subset=['Price_Clean', 'Area_Sqft', 'Bedrooms', 'Bathrooms'], inplace=True)
```

Removing Outliers

```
# Outlier removal using IQR
Q1 = df['Price_Clean'].quantile(0.25)
Q3 = df['Price_Clean'].quantile(0.75)
IQR = Q3 - Q1

df = df[(df['Price_Clean'] >= Q1 - 1.5 * IQR) & (df['Price_Clean'] <= Q3 + 1.5 * IQR)]
```

---

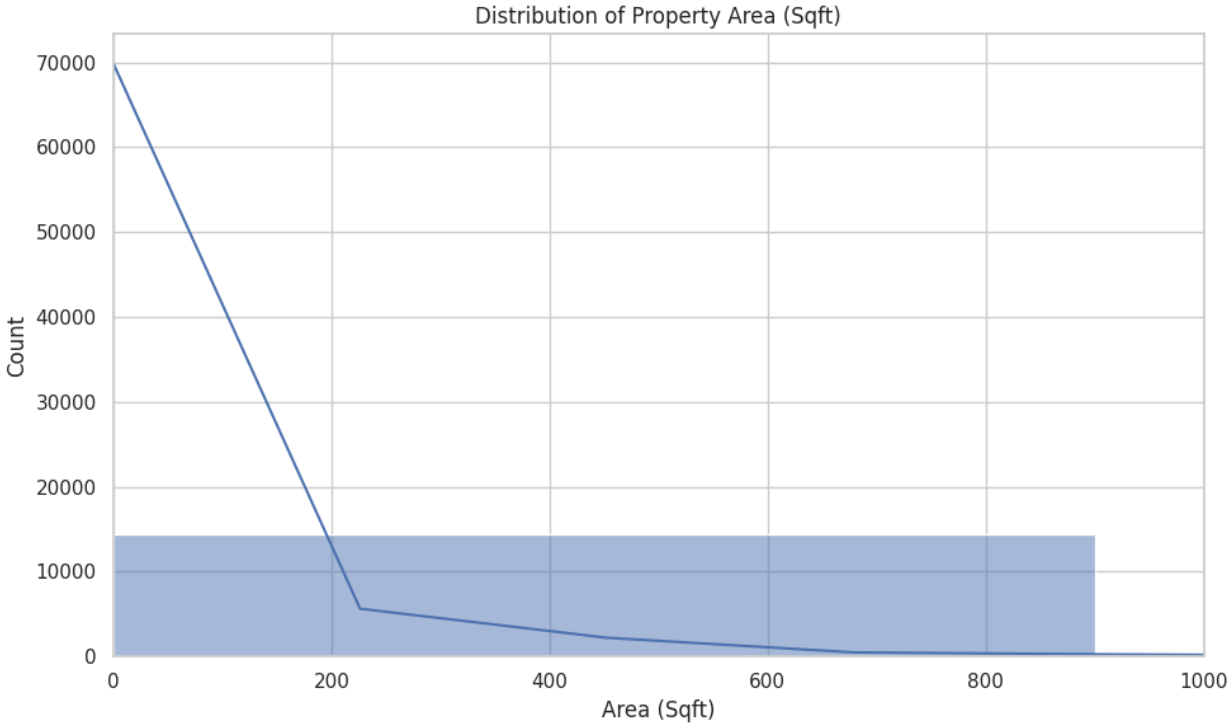
5. Univariate Analysis

Plot 1: Distribution of Property Prices

- **Purpose:** Understand how property prices are distributed.
- **Method:** Histogram + KDE curve.
- **Insight Expected:** Are most listings affordable, mid-range, or luxury?

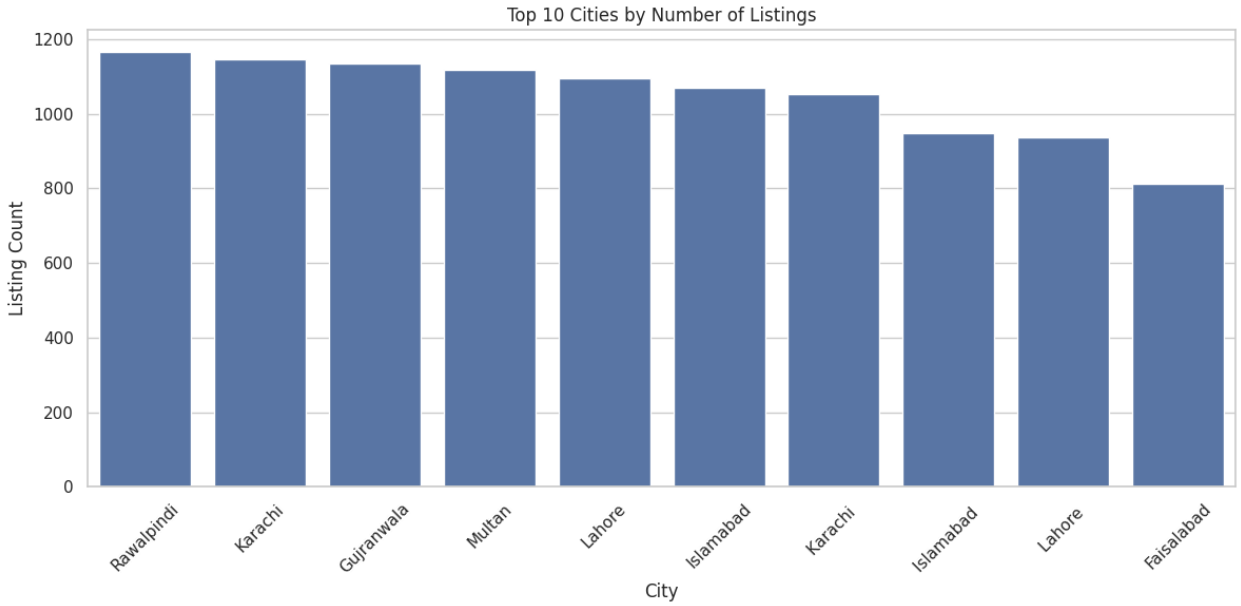
Plot 2: Distribution of Property Area (Sqft)

- **Purpose:** Check how property sizes vary.
- **Method:** Histogram + KDE.**Insight Expected:** Are most properties small (e.g., apartments) or large (houses/plots)?



Plot 3: Top 10 Cities by Number of Listings

- **Purpose:** Identify the most active real estate markets on Zameen.com.
- **Method:** Barplot.
- **Insight Expected:** Which cities (like Lahore, Karachi, Islamabad) have the most property listings?



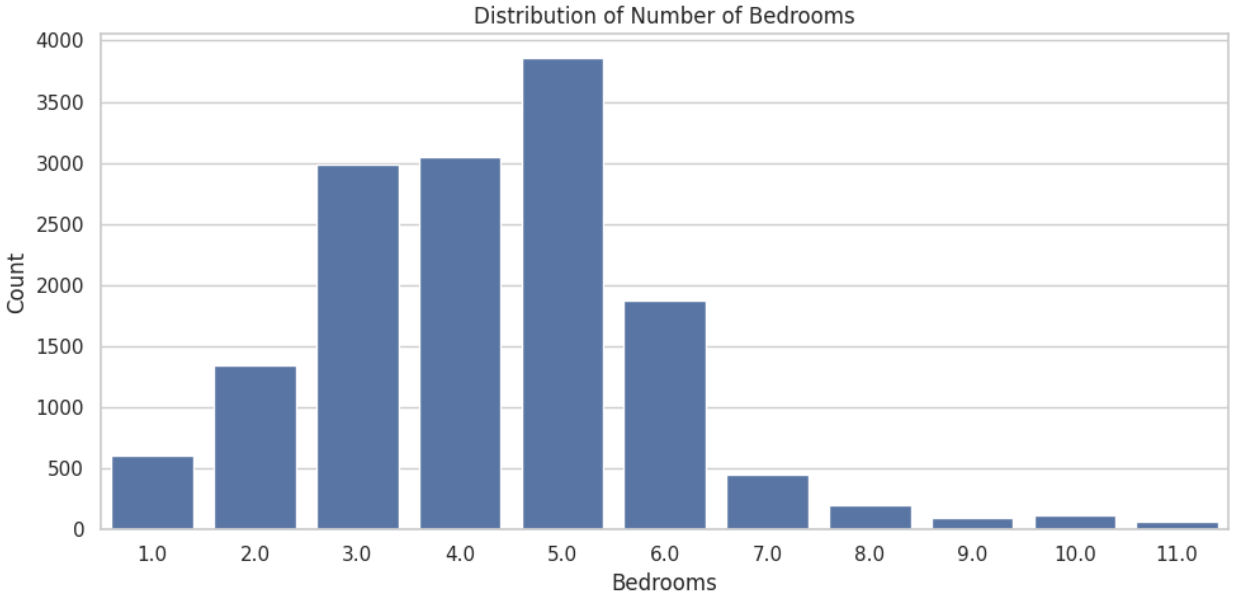
Plot 4: Listings by Property Type

- **Purpose:** See which property types are most commonly listed.
- **Method:** Barplot.

Insight Expected: Are there more Houses, Plots, Flats, Commercial listings?

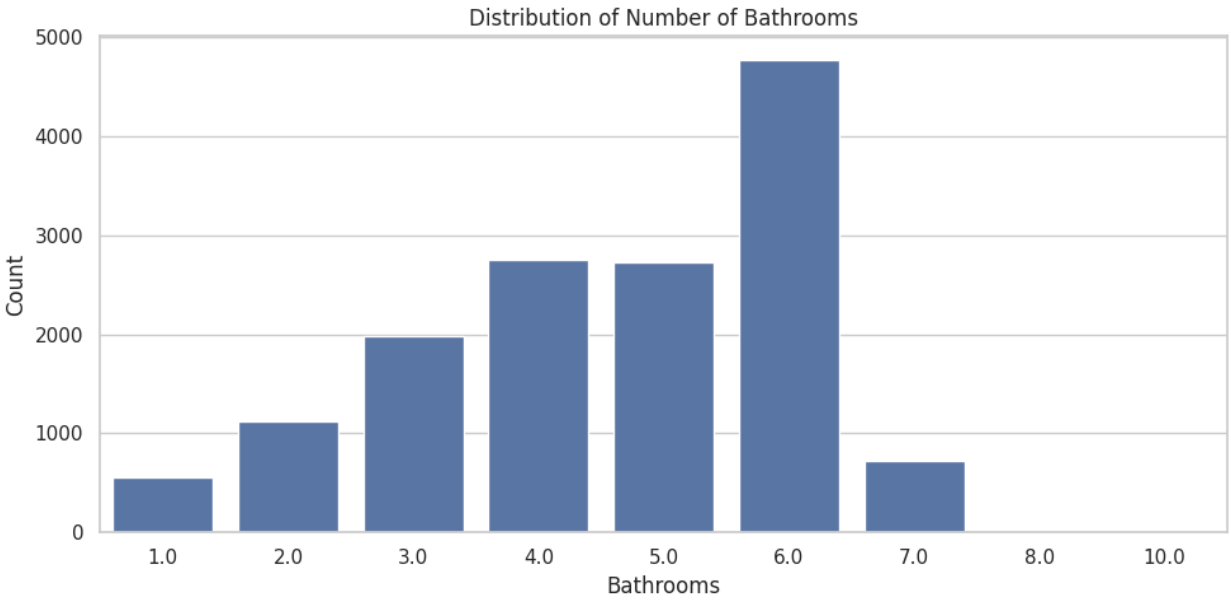
Plot 5: Distribution of Bedrooms

- **Purpose:** Examine how many bedrooms are typical in listings.
- **Method:** Countplot (barplot of counts).
- **Insight Expected:** Are 2-bed or 3-bed properties most common?



Plot 6: Distribution of Bathrooms

- **Purpose:** Similar to bedrooms, see the distribution of bathrooms.
- **Method:** Countplot.
- **Insight Expected:** What are the typical bathroom configurations? (e.g., 1, 2, or 3+)



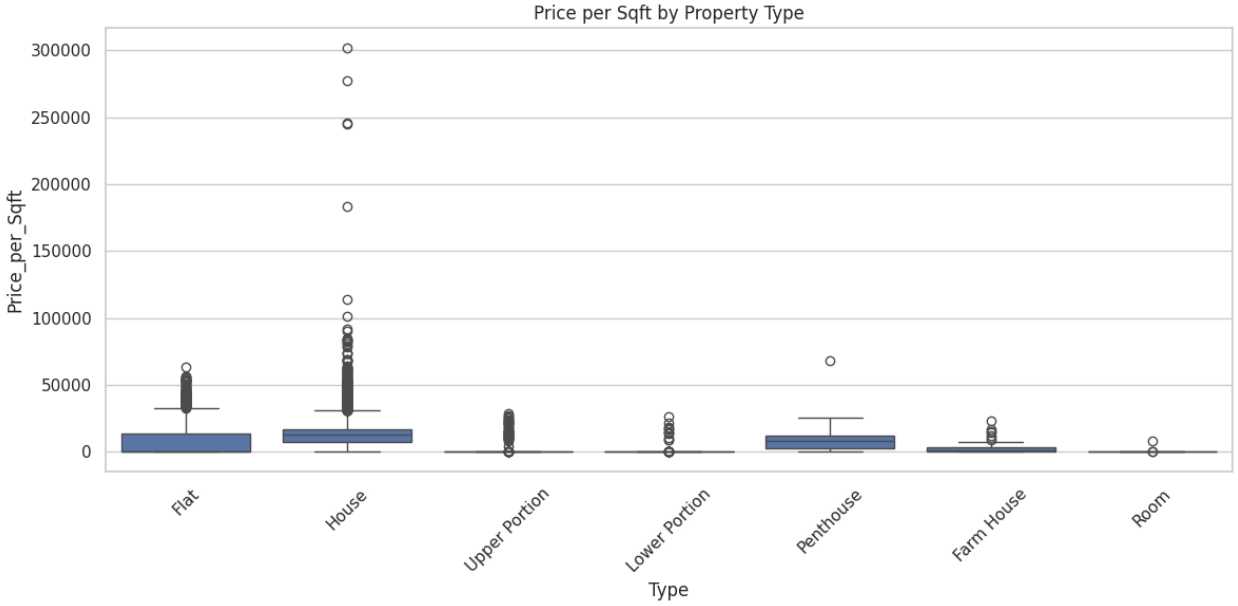
---

6. Bivariate Analysis

Plot 1: Price per Sqft by Property Type

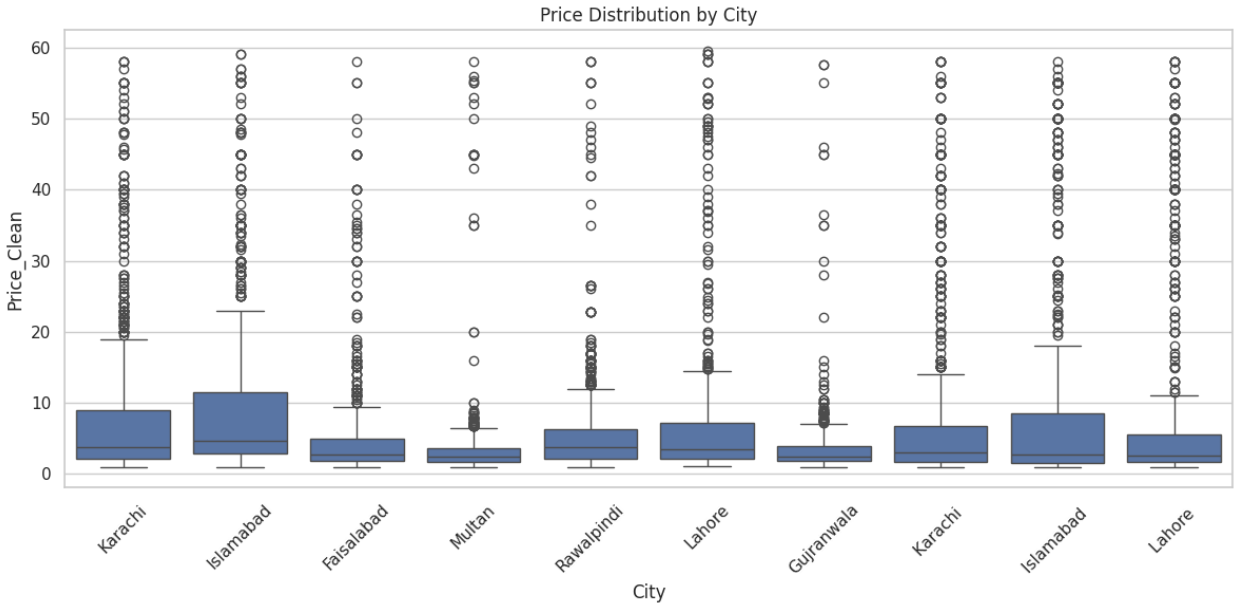
- **Variables Analyzed**: Type (categorical) vs Price_per_Sqft (numerical)
- Method: Boxplot
- Purpose: Shows price per square foot variation by property type.

- Insight Expected: Are houses costlier per sqft than plots or flats? Outliers and median values are visible.



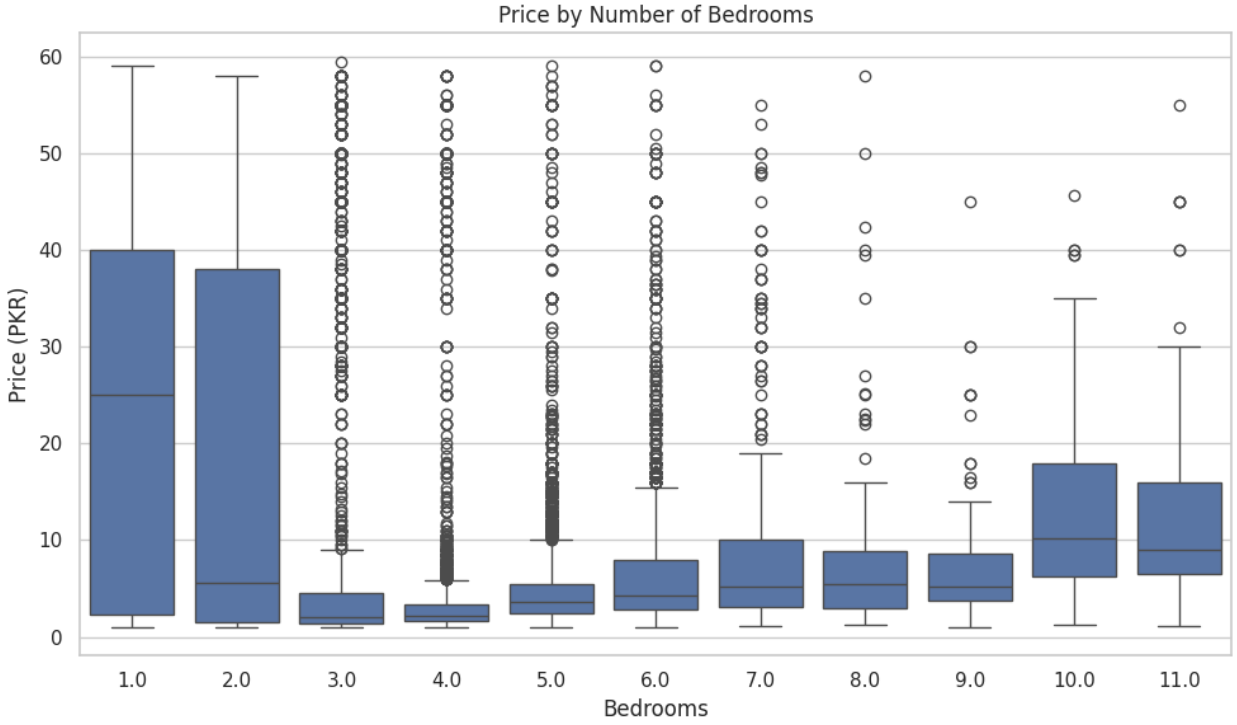
Plot 2: Price Distribution by City

- **Variables Analyzed**: City (categorical) vs Price_Clean (numerical**)**
- **Method:** Boxplot (limited to top 10 cities)
- **Purpose:** Compares property prices across top cities.
- **Insight Expected:** Which cities are more expensive? Are there large price disparities within a city?



Plot 3: Bedrooms vs Price

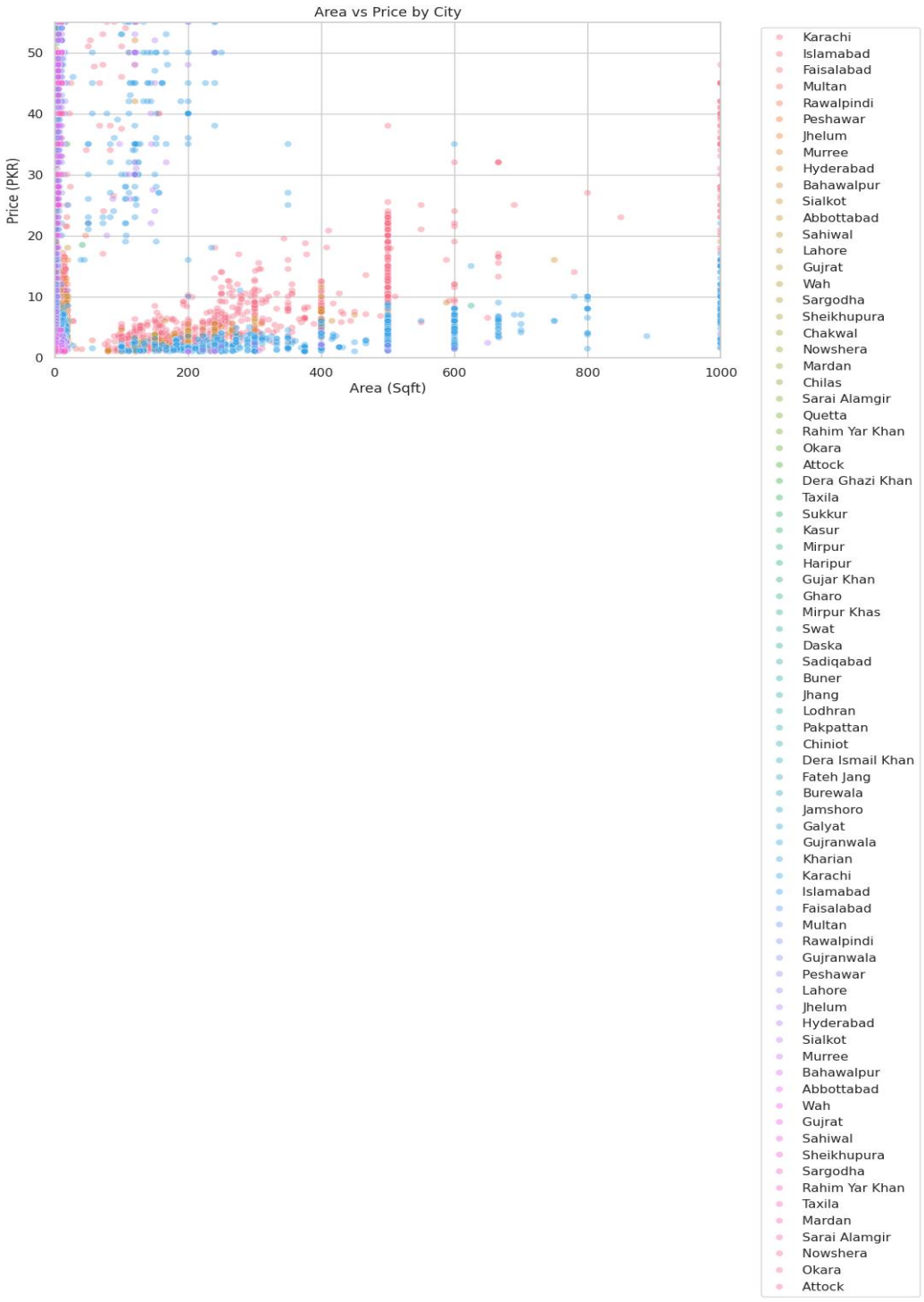
- **Variables Analyzed**: Bedrooms (numerical/categorical) vs Price_Clean
- **Method:** Boxplot
- **Purpose:** Shows how price varies with number of bedrooms.
- **Insight Expected:** Do more bedrooms mean higher prices? Is there a consistent pattern?



Plot 4: Area vs Price Scatter Plot (Colored by City)

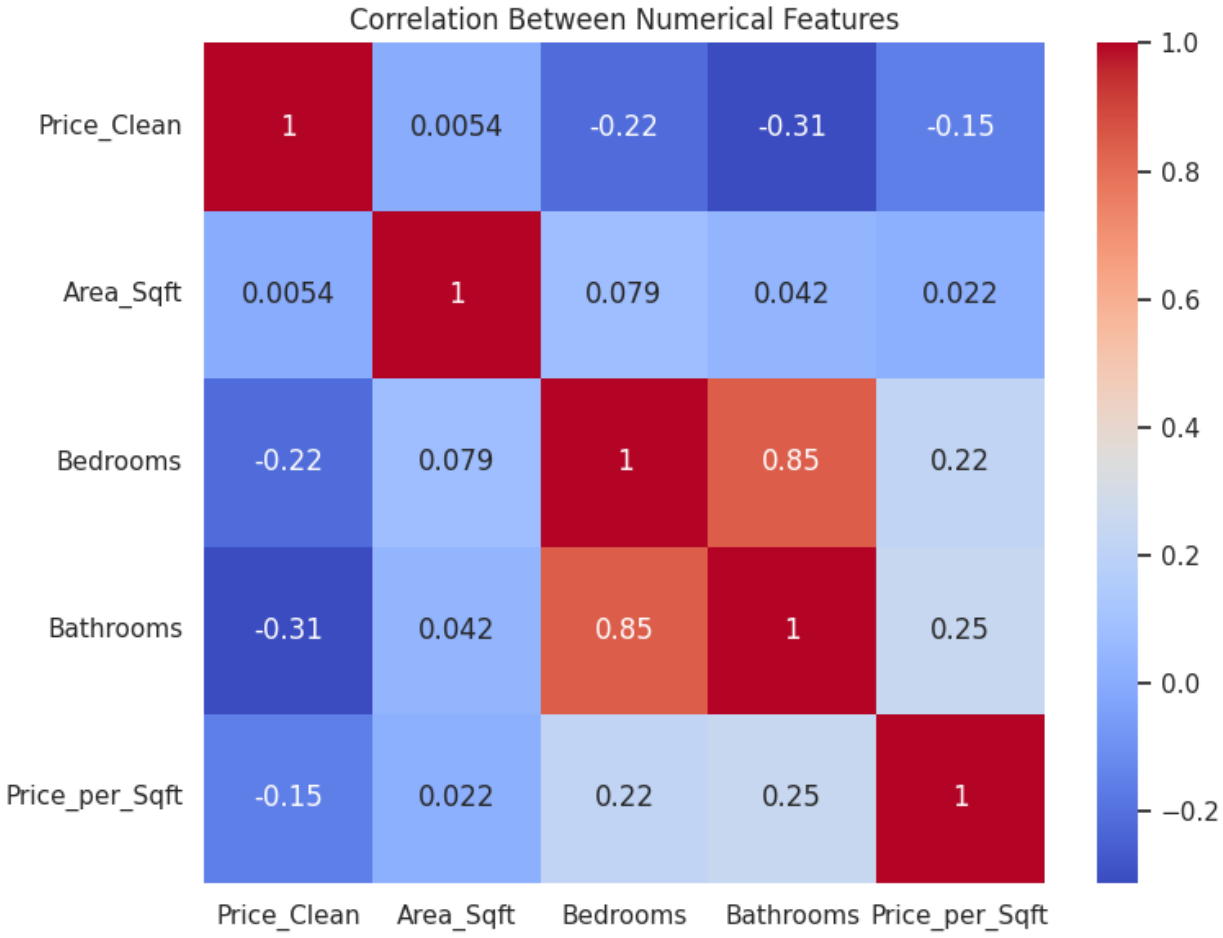
- **Variables Analyzed:** Area_Sqft vs Price_Clean, colored by City
- **Method:** Scatterplot with color hue
- **Purpose:** Visualize correlation between area and price, broken down by city.

- **Insight Expected:** Are larger properties always more expensive? Which city has higher price-per-sqft?



Plot 5: Correlation Heatmap

- **Variables Analyzed:** Numerical features only — Price_Clean, Area_Sqft, Bedrooms, Bathrooms, Price_per_Sqft
- **Method:** Correlation matrix (heatmap)
- **Purpose:** Quantify strength and direction of linear relationships between numeric features.
- **Insight Expected:**
    - Is area positively correlated with price?
    - Are bedrooms and bathrooms strong predictors of price?
    - Is price-per-sqft negatively correlated with area?
    -



---

7. Key Insights

Step 1: Calculate Median Price per City

- You group the data by City and calculate the **median** of Price_Clean.

Step 2: Print Sorted List

- You sort the cities in **descending order of median price**.
- This allows you to identify the **most expensive cities** for real estate.

Step 3: Plot Top 10 Cities by Median Price

- You create a **horizontal bar chart** for better readability.
- Uses the viridis color palette for a visually appealing plot.
- **Insight Expected:** Which 10 cities have the **highest median** real estate prices?
-

-

8. Recommendations

- Target premium zones (DHA, Bahria) for high-margin investments.
- Use price_per_sqft instead of raw price for valuation comparisons.
- Apply outlier detection to remove skewed listings from decision-making.

- Incorporate rental yield and listing age in future analyses.

---

9. Conclusion

The EDA highlights that **city, property type, and area** are the strongest influencers of real estate pricing in Pakistan. Insights from this project support more accurate property valuation, optimized marketing, and smarter investment targeting.