Real Estate Market Insights in Pakistan: Exploratory Data Analysis of Zameen.com Listings

1. Introduction

The real estate sector in Pakistan is a major area of interest for investors, homeowners, and policymakers alike. The aim of this project is to conduct a comprehensive exploratory data analysis (EDA) on property listings obtained from Zameen.com, focusing on understanding the key factors that drive property prices.This analysis excludes rental listings and focuses solely on properties listed for sale, in line with investor interests in purchase-based investment opportunities.
This study is particularly targeted toward real estate investors, to help identify price determinants, luxury indicators, and city-wise trends across Pakistan.

2. Objective

- To extract actionable insights from real estate listing data.
- To identify which property features are most influential in driving price.
- To understand how different cities and property types vary in value.
- To prepare the dataset for possible modeling or investment strategy formulation.

3. Dataset Overview

Original Shape: 18,255 rows × 59 columns
Source: Zameen.com scraped data
Key Variables: Price, Area, Bedrooms, Bathrooms, Type, City, Built in Year

Many columns were either completely null or contained inconsistent values, requiring extensive preprocessing.

4. Data Cleaning & Preprocessing

Step-by-Step Cleaning:

a) Handling Price

- Original values included text like “PKR 1.5 Crore”.
- Created convert_price() function to:

    - Strip “PKR”
    - Detect units like Lakh, Crore, Thousand
    - Convert to numeric PKR value.
- Filled missing prices using median imputation due to right skew (Skewness ≈ 6.3).

b) Handling Area

- Original values contained units like Marla, Kanal, Sq. Ft.
- Created convert_area() to:
    - Normalize all areas to square feet.
- Filled missing areas using median (Skewness ≈ 51.2).

c) Bedrooms & Bathrooms

- Converted to numeric using pd.to_numeric() with coercion.
- Filled missing values with median for Bedrooms and Bathrooms.

d) Built in Year

- Cleaned extreme years (e.g., 202122) by filtering [1950, current_year].
- Filled missing with median.
- Created new feature Property Age = current_year - built_in_year.

e) Categorical Fixes

- Fixed 'Purpose' (e.g., "For" → "For Rent")
- Standardized city, type, and province using fuzzy matching (fuzzywuzzy).
- Filtered the dataset to retain only 'For Sale' listings, as the analysis is intended for real estate investors and focuses on sale price dynamics. All 'For Rent' entries were excluded.

f) Location Column

- Split 'Location' into:
    - Area Name
    - City_Location
    - Province
- Dropped rows with missing location and removed non-quantifiable text columns (Title, Description, etc.).

5. Feature Engineering

- Property Age: From "Built in Year"
- Age Category:

    - 0–5 yrs → New
    - 6–15 yrs → Moderate
    - 16–30 yrs → Old
    - 30+ yrs → Very Old
- Is Luxury (Binary): This feature is designed to flag high-end properties. The indicator was constructed using a combination of conditions:

  - Price >= 200,000,000 PKR
  - Bedrooms >= 5
  - Area >= 4,500 sqft
  - Located in known luxury localities such as DHA Defence, Clifton, F-7, Bahria Town, etc.

6. Outlier Detection & Handling

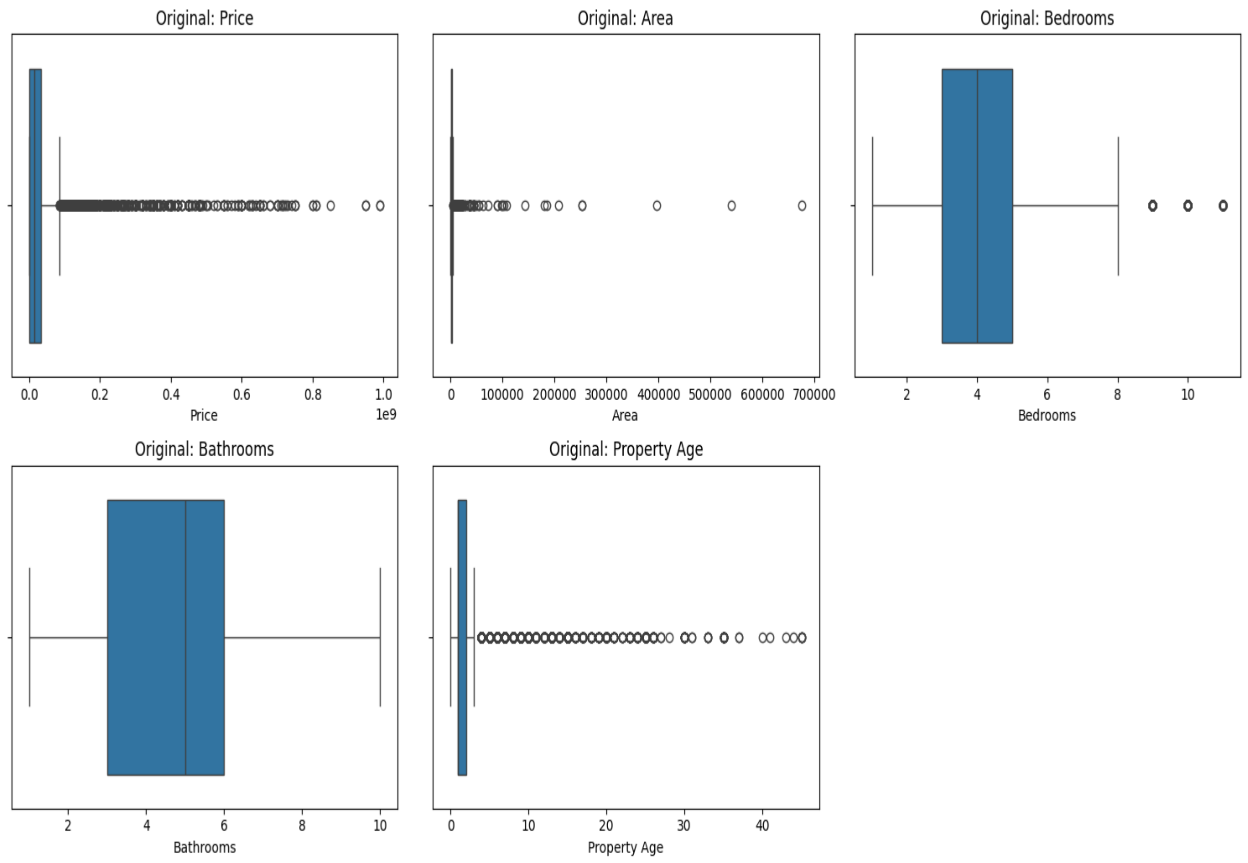
Outliers were retained to reflect true market variability, especially for luxury properties. We used IQR method for detection:

Feature	Outliers
Area	1770
Price	1108
Bedrooms	216
Bathrooms	1
Property Age	4772

I removed outliers only from Area and Price using manual thresholds based on local real estate norms (e.g., 1 marla = 272.25 sq ft), as statistical methods (like IQR) flagged many valid but rare properties as outliers.

```
# 3. Define minimum and maximum thresholds based on real-world logic
min_area_sqft = 272.25       # 1 marla
max_area_sqft = 50000        # Very large but still possible (e.g., big farmhouses)
min_price = 100000           # Rs. 1 lakh (below this is unrealistic)
max_price = 1_000_000_000     # Rs. 100 crore (upper limit)

# 4. Apply filters to remove only extreme/unrealistic outliers
df = df[
   (df['Area'] >= min_area_sqft) & (df['Area'] <= max_area_sqft) &
   (df['Price'] >= min_price) & (df['Price'] <= max_price)
]
```



Shape after removing unrealistic outliers: (14608, 11)

7. Univariate Analysis

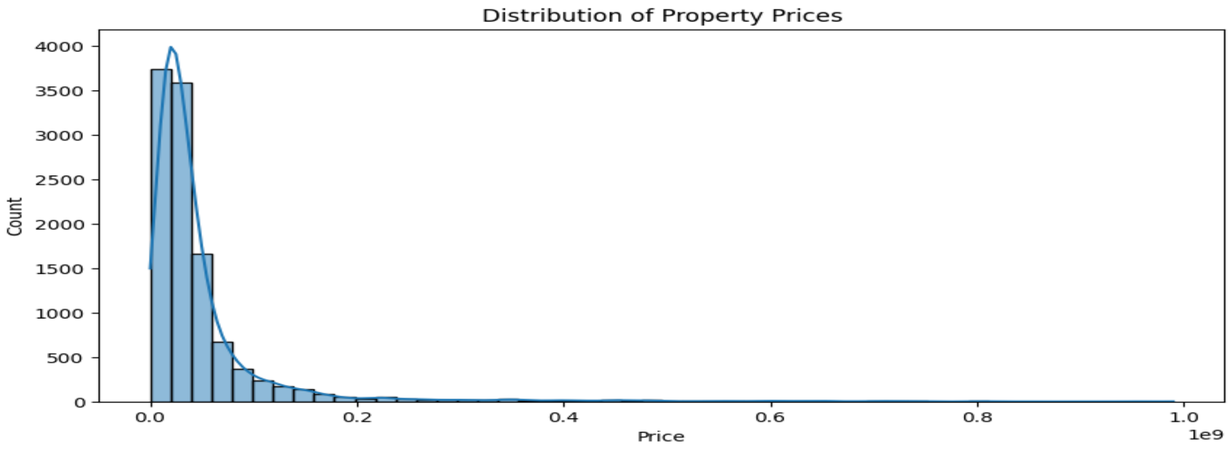
Key Findings:

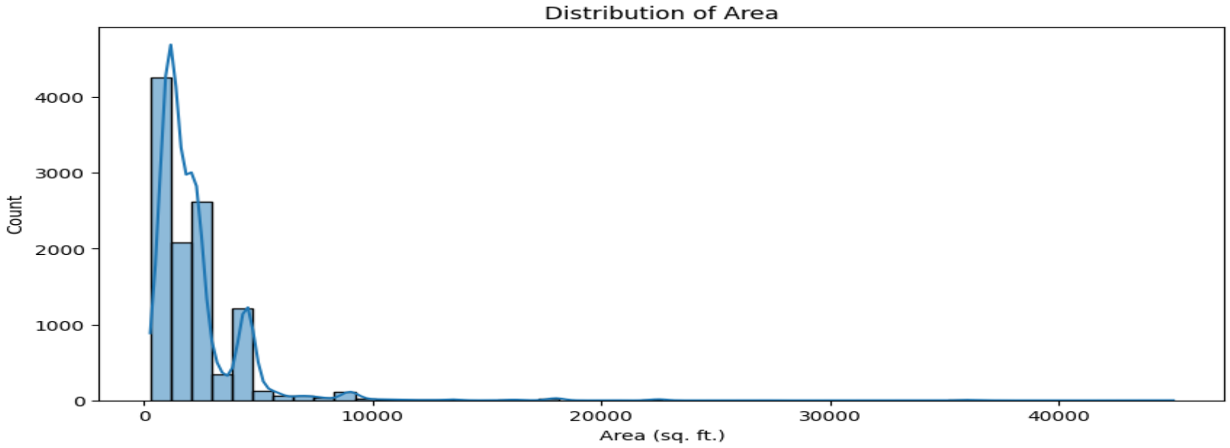
Property Prices

- Right-skewed: Most properties are low-priced, with a few high-end outliers.
- Indicates high price variability; log transformation was essential.

Area (sq. ft.)

- Highly skewed: Majority under 10,000 sq. ft., with extreme outliers above 300,000 sq. ft.
- Reflects mostly standard-sized homes with few large luxury properties.

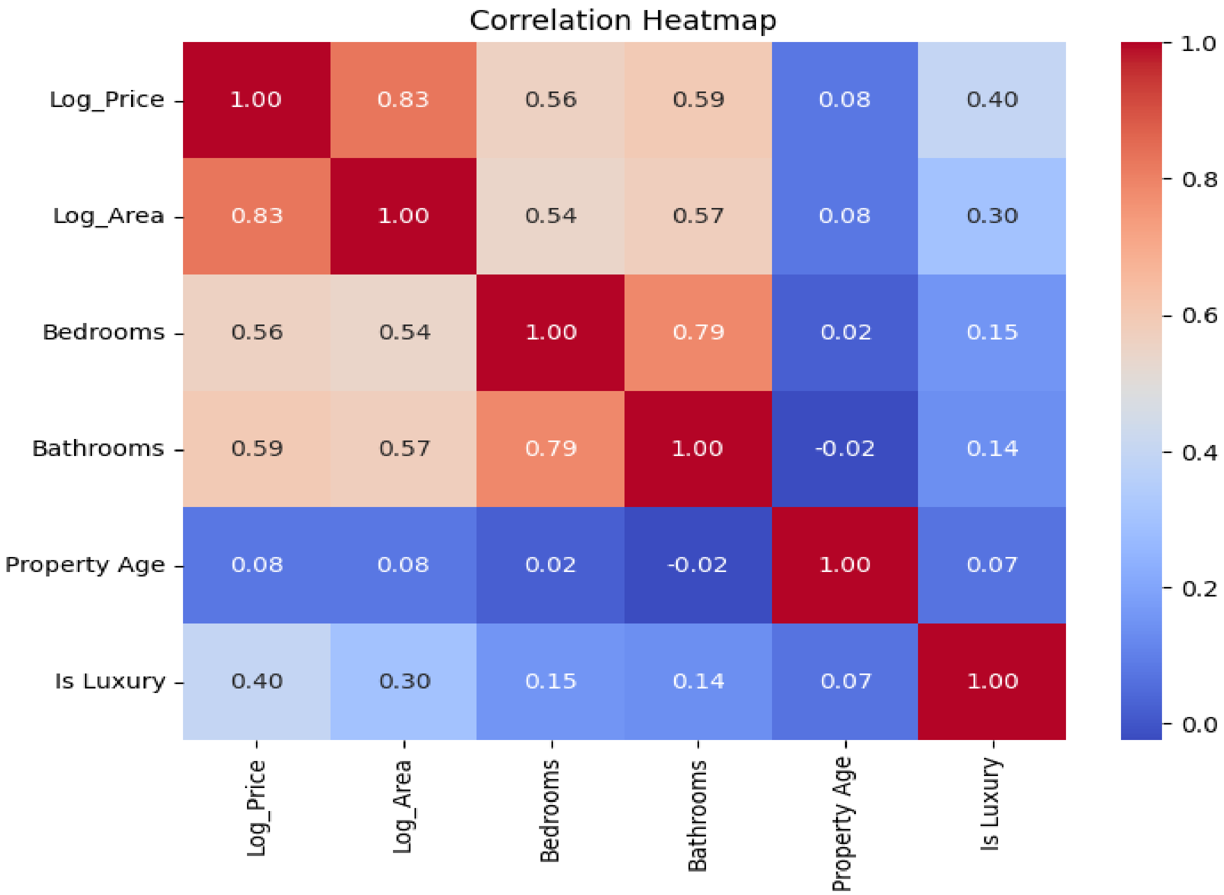




8. Bivariate Analysis

a) Correlation Heatmap

To understand linear relationships, a correlation heatmap was computed using log-transformed values for Price and Area. log transformation on Area and Price to reduce skewness and make their distributions more normal, which improves the accuracy of visualizations and statistical analysis.

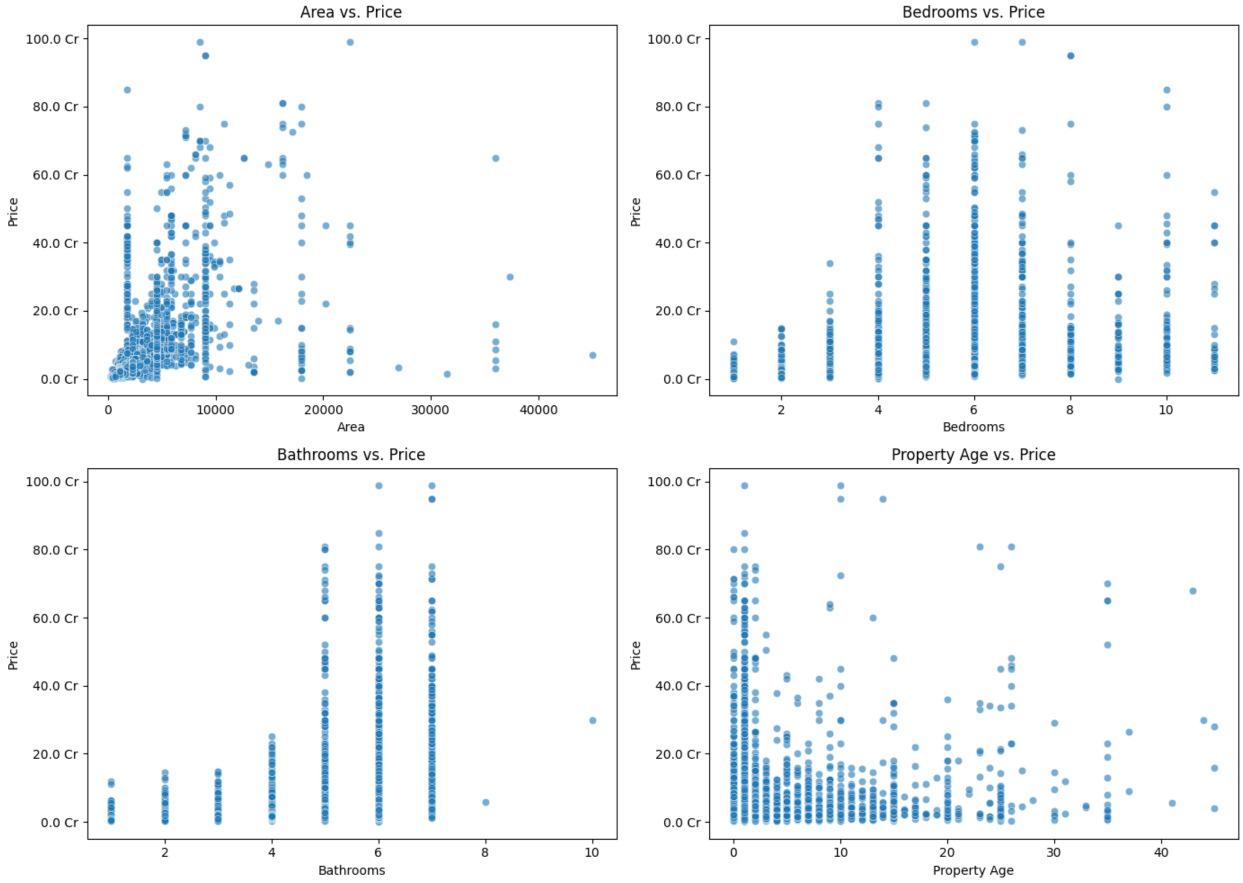


Feature Correlation with Log_Price, Key Takeaway

Log_Area	0.83	Strongest positive correlation with price.
Bathrooms	0.59	Strongly correlated with price.
Bedrooms	0.56	Moderate-to-strong correlation.
Is_Luxury	0.40	Indicates luxury classification impacts price.
Property Age	0.08	Very weak correlation with price.

- Area, Bathrooms, and Bedrooms are the primary drivers of price.
- Property Age has little influence — some older properties may still fetch high prices due to location or type.
- Luxury Status moderately affects price, which aligns with box plot outliers in major cities and for farm houses/penthouses.

b) Scatter Plots: Numerical vs Price



Area vs. Price

- Positive trend: Larger area generally correlates with higher price.
- Many listings cluster around smaller area values, but some large-area properties fetch extremely high prices.
- A few high-area, low-price outliers suggest underpriced listings or rural/less prime locations.

Bedrooms vs. Price

- Peak concentration around 5–6 bedrooms, with moderate prices.
- Beyond 6 bedrooms, no strong upward trend — possibly due to decreasing demand or niche luxury use.
- Large bedroom counts do not always guarantee high prices.

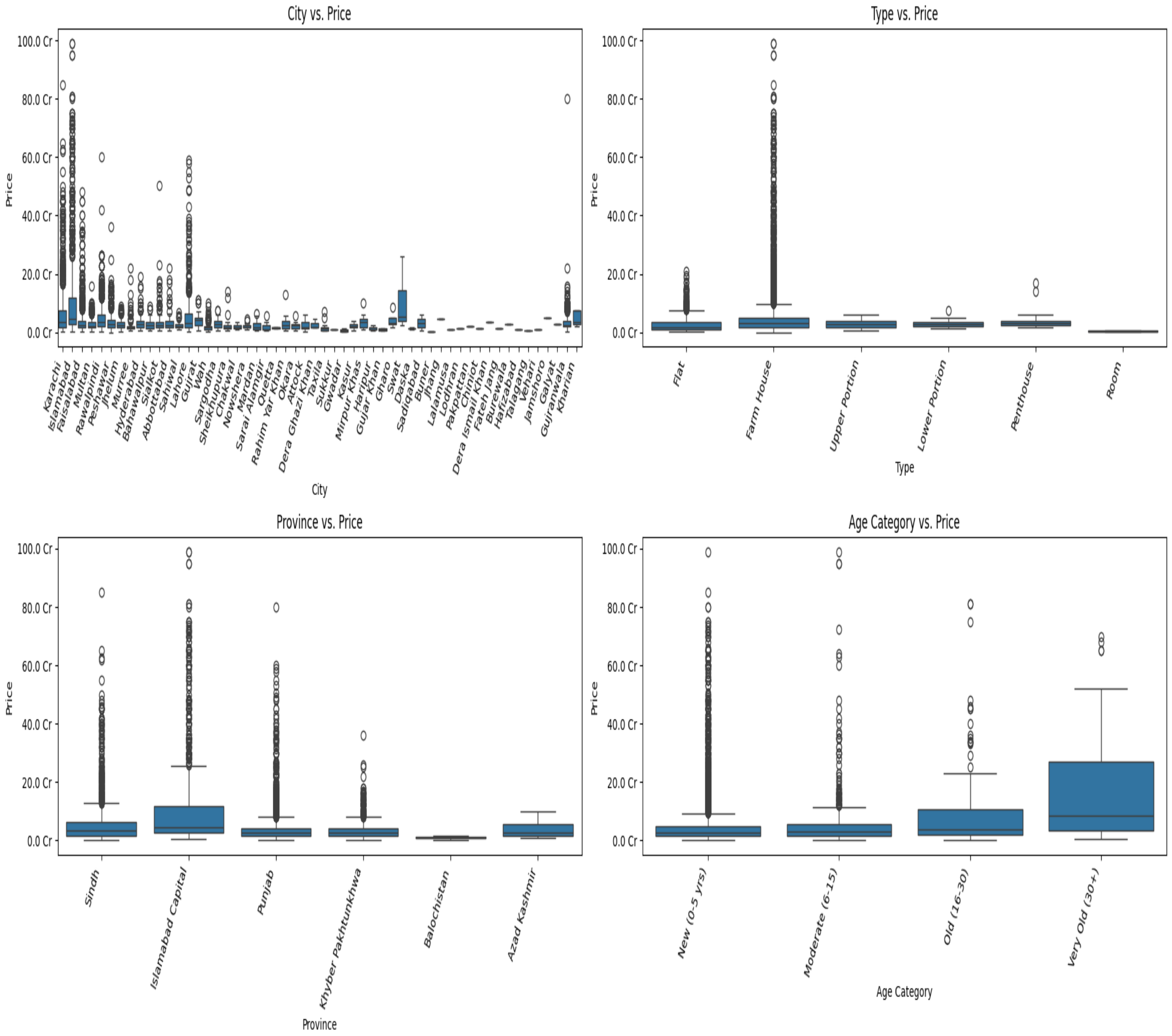
Bathrooms vs. Price

- Similar to bedrooms, 4–6 bathrooms have higher price densities.
- Price peaks around 5–6 bathrooms, but again, more bathrooms beyond that doesn’t necessarily mean higher price.

Property Age vs. Price

- Negative trend: Newer properties tend to be higher priced.
- Many properties aged <10 years cluster in the higher price range.
- Older properties (>20 years) exist in both low and high price ranges — possibly due to location or renovation status.

c) Boxplots: Categorical vs Price



City vs. Price

High-Value Cities:

- Karachi, Islamabad, and Lahore show the highest median prices and the widest price range.
- Islamabad shows many high-priced outliers, indicating presence of luxury listings.

Low-Value Cities:

- Cities like Dera Ghazi Khan, Mianwali, Swabi, etc., show much lower median prices and less variance.

Price Distribution:

- Large number of outliers in major cities suggest that some luxury properties greatly skew price statistics.

Type vs. Price

- Farm Houses and Penthouses generally have the highest median prices.
- Flats, Portions, and Rooms are relatively cheaper with tight interquartile ranges.
- Farm Houses show a high number of outliers, again pointing to luxury listings.

Province vs. Price

- Islamabad Capital Territory shows the highest median price and broad range, indicating expensive properties.
- Sindh and Punjab follow, with large spreads.
- Balochistan and Azad Kashmir have the lowest price ranges and medians.

Age Category vs. Price

- Very Old (30+ yrs) properties surprisingly show high median and upper-range prices.
- New (0–5 yrs) and Moderate (6–15 yrs) properties generally have tighter distributions and fewer high-value listings.
- This suggests some very old properties may be in prime locations or highly valuable heritage/luxury homes.

9. Insights & Recommendations

Key Insights

Price Drivers

- Log-transformed area (Log_Area) has the strongest correlation with price (0.83), confirming that property size is the most significant driver of price.
- Bathrooms (0.59) and bedrooms (0.56) also contribute notably to property value, with higher counts generally associated with higher prices — though diminishing returns are observed beyond 6+ rooms.
- The Is_Luxury indicator (0.40) confirms that properties in elite locations with premium features command significantly higher prices.

- Property age shows weak correlation (0.08), meaning location and amenities often outweigh age in determining price.

City & Region Trends

- Islamabad, Karachi, and Lahore exhibit the highest median prices and widest price spreads, indicating a strong mix of standard and luxury markets.
- Cities like Dera Ghazi Khan, Swabi, and Mianwali show lower price points and narrower ranges, suggesting affordability but limited appreciation potential.
- Farmhouses and Penthouses command the highest price premiums among property types, but with high variance, which implies risk.

Outlier Handling

- IQR detection alone flagged over 1,700 outliers in area and 1,100 in price, but many of these were valid listings (e.g., large farmhouses).
- By applying real-world cutoffs (e.g., 1 marla = 272.25 sq ft), only truly unrealistic or mistyped entries were removed — improving data integrity without sacrificing high-value insights.

Age Category vs Price

- Surprisingly, very old (30+ yrs) properties showed high prices in certain areas, likely due to prime location, historical value, or renovation.
- Newer properties (0–5 yrs) clustered tightly around higher prices, indicating investor confidence and buyer preference for fresh inventory.

💡 Recommendations for Investors

Invest in:

- Focus on High-Demand Cities: Invest in properties located in Islamabad, Karachi, and Lahore, as these cities show consistently higher median prices and broader investment potential due to urban development, demand, and amenities.
- Target Mid-Sized Properties: Properties with 4–6 bedrooms and bathrooms offer the best balance between affordability and return on investment. These are ideal for families and have strong rental and resale potential.
- Explore Renovated or Well-Located Older Properties: Some very old properties in high-value areas still command strong prices. These can be profitable renovation or redevelopment opportunities.

- Prioritize 'Luxury' Features and Localities: Listings with features like large area (4500+ sqft), 5+ bedrooms, and premium locations (e.g., DHA, Clifton, Bahria Town, F-6/F-7) consistently fall in the high-price range and may offer long-term value retention.

Be Cautious of:

- Listings with very high area but low price these may indicate mispriced, rural, or less desirable locations.
- Properties with excessive bedrooms (7+) and large area but no significant price lift, these may have limited market demand.

Data-Driven Strategy:

- Use Log_Area and bathroom count as core filters when identifying undervalued or overperforming listings.
- Focus on cities with high variance for flipping/investment opportunities, and low variance cities for stable rental income.

10. Conclusion & Next Steps

Summary

This exploratory analysis of 18,255 real estate listings from Zameen.com has revealed meaningful patterns in property pricing dynamics across Pakistan. The project involved extensive data cleaning, including handling missing values, unit conversions, log transformations, and manual outlier filtering — all essential for producing a high-quality, investor-ready dataset.

Key insights emerged from both univariate and bivariate analyses, helping uncover influential drivers such as property area, number of bathrooms, luxury classification, and city.

Key Takeaways:

- Handling missing values using domain-appropriate strategies (e.g., median imputation for skewed features) ensured data completeness without distortion.
- Bigger isn't always better — while area impacts price, location, layout, and category often have a stronger influence.
- Luxury tags and prime city zones (like DHA, Clifton, Bahria Town, and F-7 Islamabad) lead to premium pricing, often outweighing traditional factors like age.
- Manual outlier removal, based on local real estate logic (e.g., 1 marla ≈ 272.25 sq ft), helped retain valid high-end listings while filtering out clearly unrealistic data.

Next Steps

- Use this cleaned data for predictive modeling (e.g., price prediction).
- Build a dashboard for dynamic exploration by investors.
- Add time dimension if longitudinal data is available (e.g., price trends).