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

1. Problem Statement

The Pakistani real estate market is vast, dynamic, and complex. Investors, both local and international, seek data-driven insights to navigate pricing trends, assess property value, and identify high-performing neighborhoods.

Key Business Question:

What are the main factors driving property prices across cities in Pakistan, and how can investors use this data to make better decisions?

2. Data Understanding & Preprocessing

The original dataset includes 18254 rows and 59 columns. After dropping unnecessary columns, the remaining dataset includes the following columns

❖ Dataset Features:

- City
- Location
- Price (in PKR)
- Area (sq. yd, marla, kanal, sqft)
- Type
- Parking spaces
- Servant quarters
- Store rooms
- Bedrooms
- Bathrooms
- Purpose
- Built in Year

Steps Taken:

- Imported data using `pandas`

- Checked for duplicates using `.duplicated()` no duplicates were found
- Cleaned `price`, `area`, and
- Converted `bedrooms/bathrooms` columns into numeric data type and imputed NaN where '-' was entered
  - Stripped symbols like "PKR" and newline characters/ white spaces
  - Removed 16 rows for which prices and area were missing, after which the dataset has 18239 rows
  - Converted area to square feet using conversion function (where units like marla/kanal exist)
  - Converted prices into Rs. Which were originally expressed in arabs, crores, lakhs, thousands

---

□ 3. Missing Values Treatment

- Missing values identified using `.isnull().sum()`

- Dropped columns where missing data percentage was more than 45% (Parking spaces, Servant quarters, Store rooms)

- Imputation Strategy:
  - Mode imputation for categorical variables (`Type`, `City`)
  - Median for `Area`, `Price`
  - Built in year column was standardized by imputing the invalid entries outside the range 1980-2024 NaN

🔄 4. Data Cleaning & Consistency

- Standardized city names using `FuzzyWuzzy` (e.g., "Lahor", "lahore" → "Lahore") and applied functions such as lowercase and strip
- Checked for outliers in numerical columns like price and area using the **IQR method and no outliers were detected**
- Unified inconsistent property types (e.g., "Flat", "Apartment") by converting the column into lowercase and removed white spaces using lowercase and strip functions

---

5. Feature Engineering

Created the following new features:

- `price_per_sqft` = `price` / `area_sqft`
- `region` = extracted from `location`

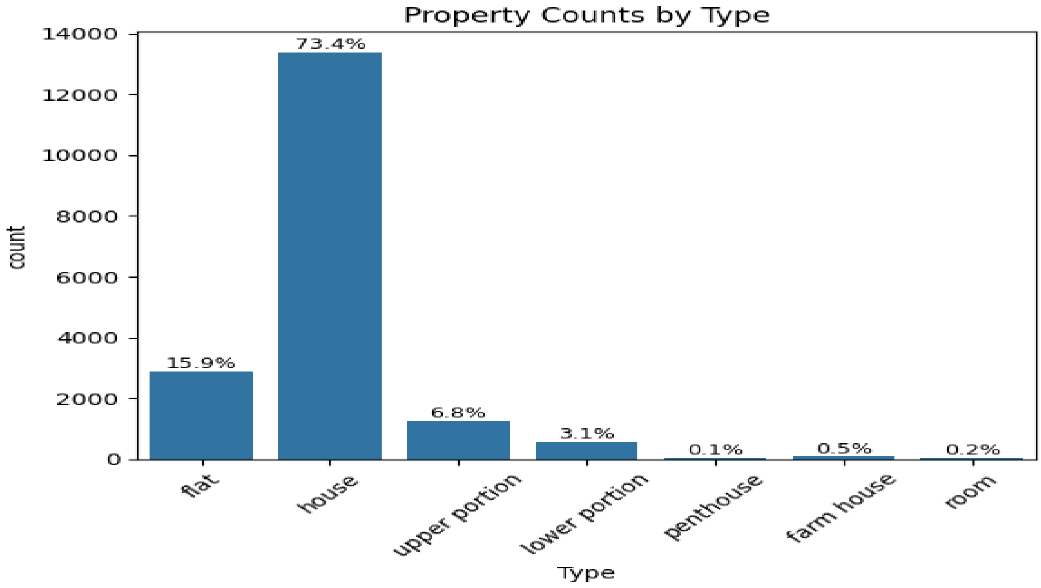
- `price_category`: Binned into ‘Low’, ‘Mid’, ‘High’, ‘Luxury’
- `property_age_group`: based on `Built in year` column
- `log_price`: for better visualization of skewed data

---

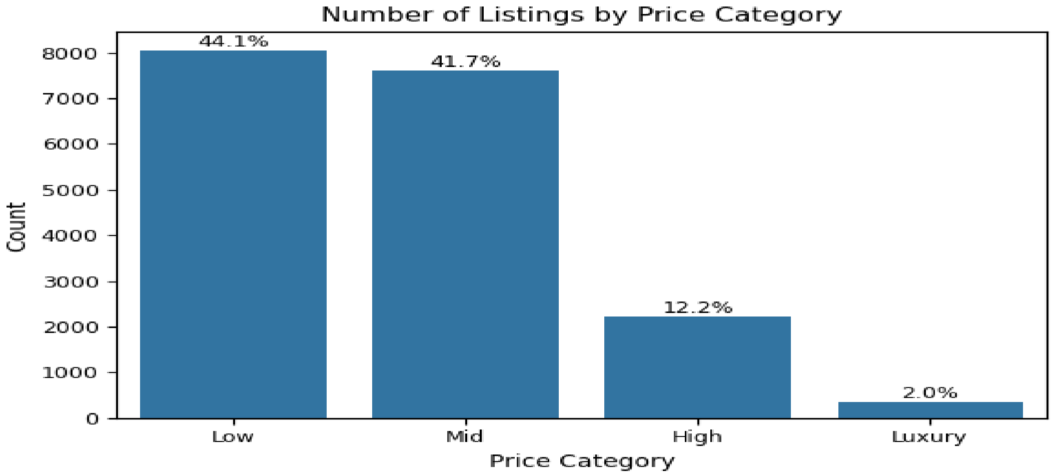
6. Univariate & Bivariate Analysis

Univariate:

- Most listings are Houses i.e. 73.4%

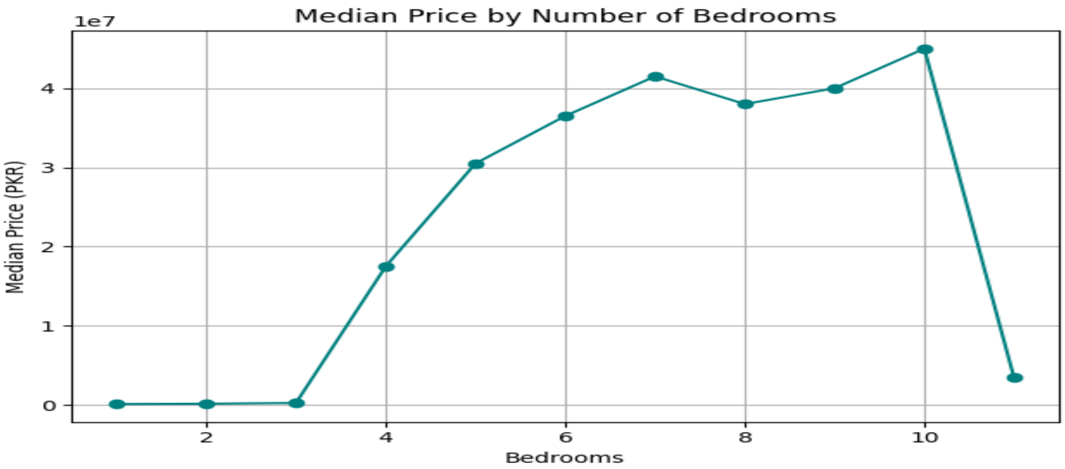


- Low (44.1%) and mid-price (41.7%) category listings dominate
-

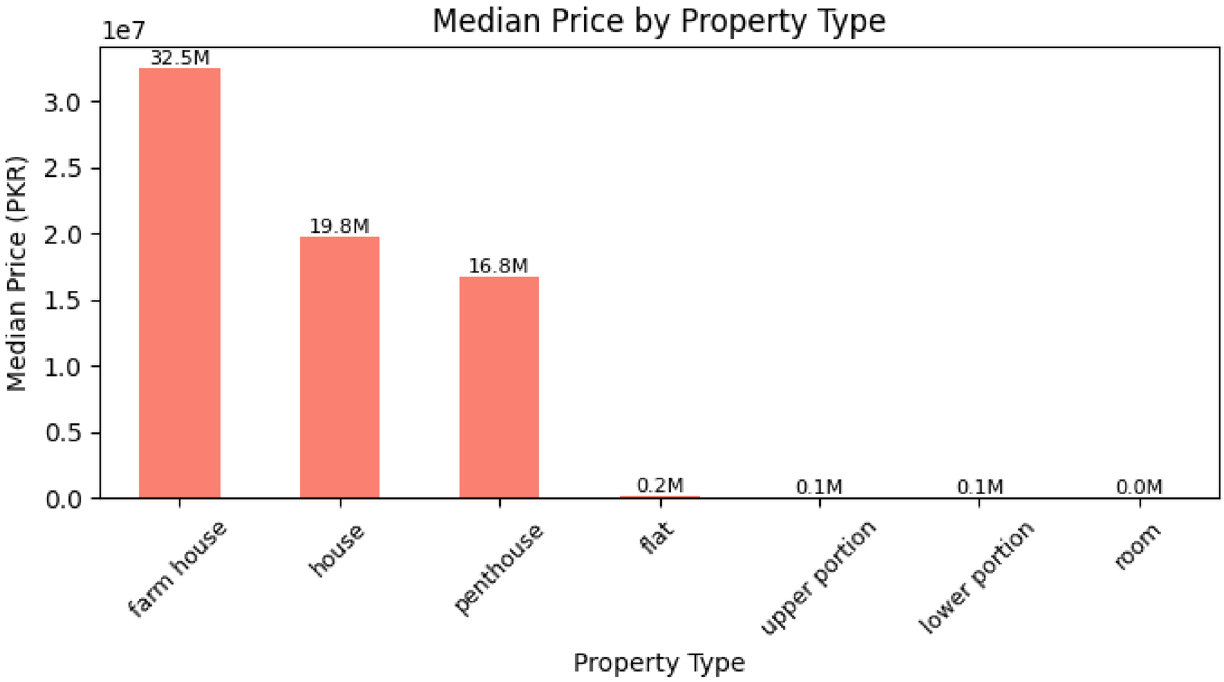


-

- When the number of bedrooms exceed 2, the median prices increase rapidly and drops sharply after 10

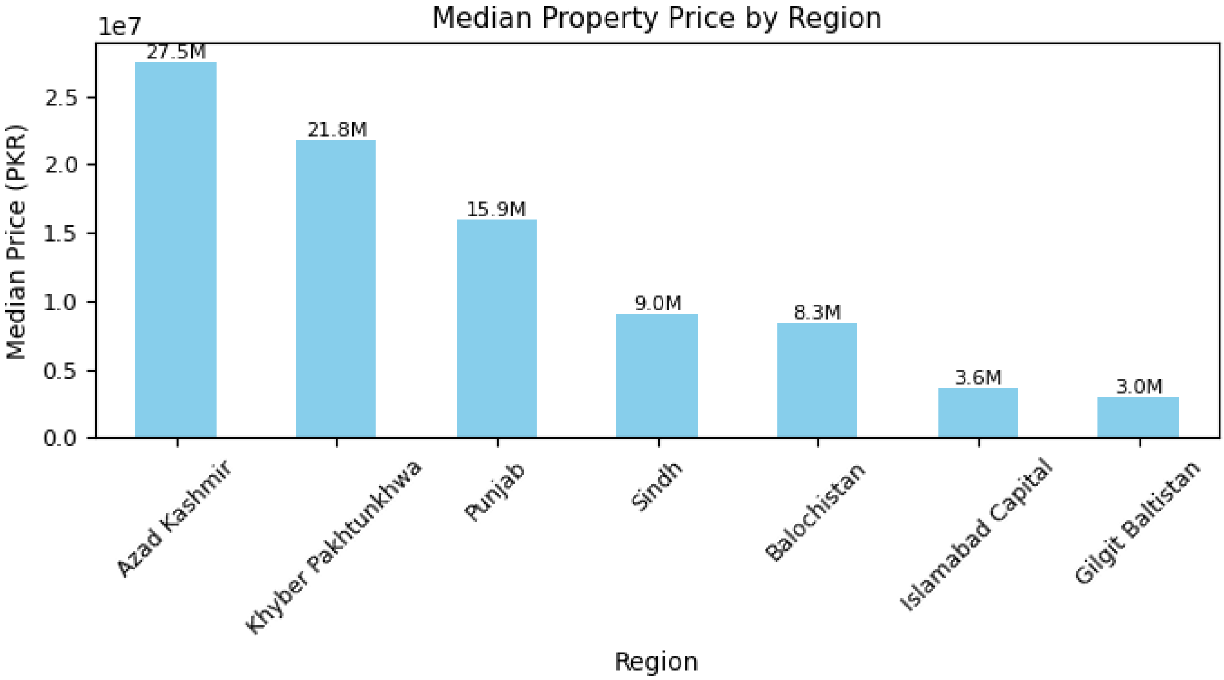


- Area distribution is non-uniform; many listings around 1200–1500 sqft
- **Farm houses** dominate listings by average prices

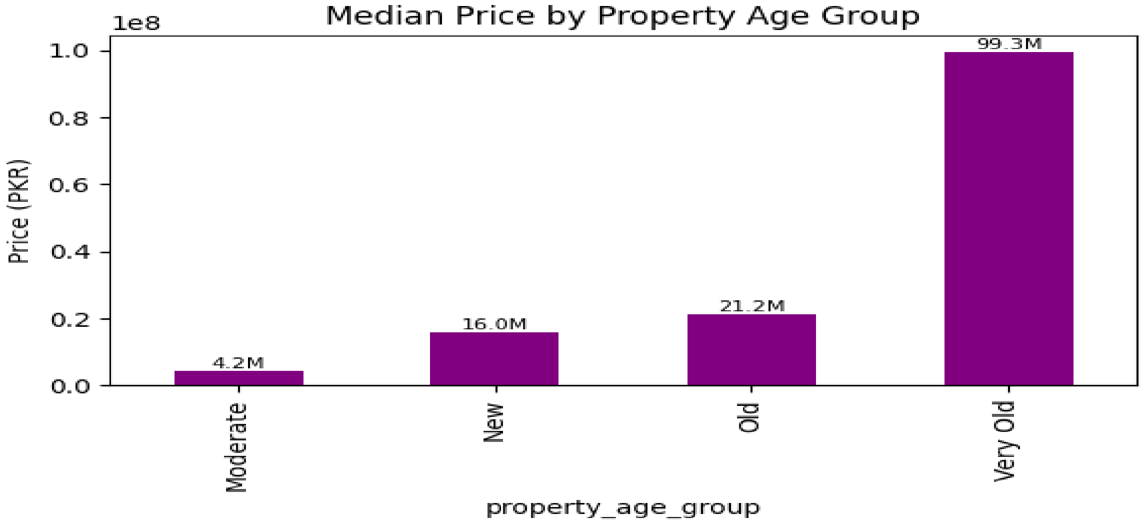


Bivariate:

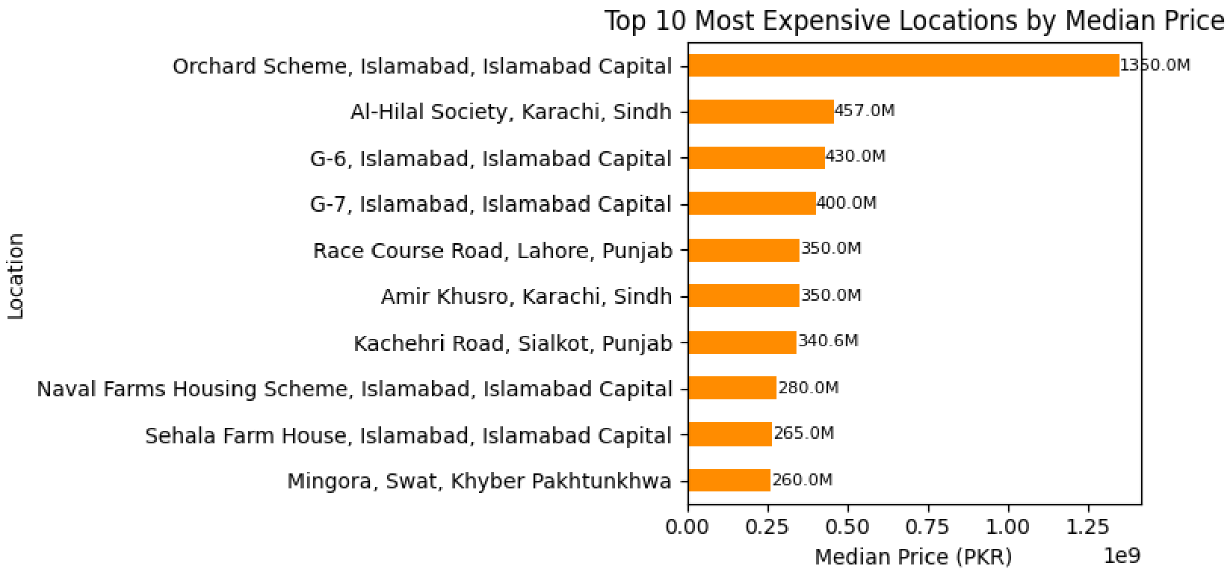
- **Azad Kashmir** has the highest property prices on average, and **Gilgit Baltistan** has the lowest



- **Positive correlation** between price and area (0.68)Very old properties command the highest median prices (99.3M) due to their prime locations and larger plots in well-established urban areas.



- Price per sqft varies significantly by **city and property type**
- Islamabad has the most expensive locations



Multivariate Analysis

Key Observations from the Correlation Heatmap:

1. Strong Relationships:

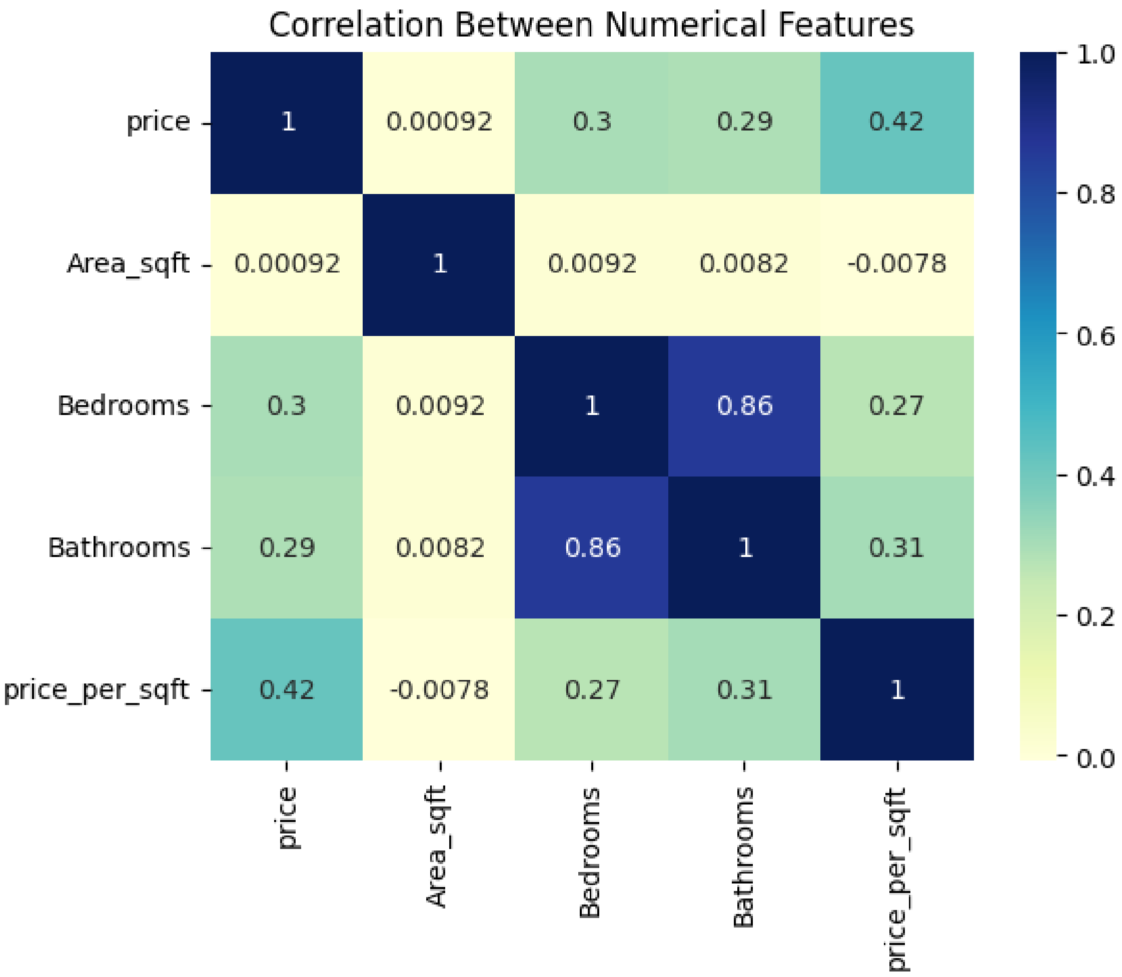
- **Bedrooms & Bathrooms (r = 0.86):**
    - Very high positive correlation. This is expected, as more bedrooms typically mean more bathrooms in larger homes.

2. Moderate Relationships:

- **Price & Price per Sqft (r = 0.42):**
    - Moderate positive correlation. Higher-priced properties tend to also have a higher price per sqft, but not always—this may vary by location and type.
- **Price & Bedrooms (r = 0.30)** and **Price & Bathrooms (r = 0.29):**
    - Properties with more bedrooms and bathrooms tend to be more expensive, but the relationship is moderate, not strong.

3. Weak or Negligible Relationships:

- **Price & Area_sqft (r ≈ 0.0009):**
  - Surprisingly **very weak correlation**. This suggests that area alone does not drive price in your dataset—likely because **location and type** of property play a much stronger role.
- **Price per Sqft & Area_sqft (r ≈ -0.0078):**
  - Essentially **no relationship**. This implies that price per sqft varies significantly regardless of total area—again, probably due to location.



Top 3 Cities by Listings (Descending order):

- Islamabad
- Lahore
- Karachi

Most Expensive per Sqft:

- DHA Defence (Karachi)
- Gulberg (Lahore)
- Bahria Town (Islamabad)
-

7. Key Insights & Recommendations

Recommendations

- **Focus on High-Value Locations for Investment**

Properties in premium areas like **DHA Defence (Karachi)**, **Gulberg (Lahore)**, and **Bahria Town (Islamabad)** command the **highest price per sqft**. Investors seeking strong returns should prioritize these locations, where demand and prestige elevate value.

- **Leverage the Premium on Older Properties in Urban Centers**

Very old properties often have the **highest median prices**, likely due to their presence in **well-established neighborhoods** with better infrastructure and larger plots. Developers can benefit from **renovating or redeveloping** such assets.

- **Prioritize Price per Sqft Over Total Area in Valuations**

The correlation between **`area_sqft` and price is nearly zero (r ≈ 0.0009)**, indicating that **size alone does not determine value**. `Price_per_sqft` (r = 0.42) is a **stronger and more reliable metric**, especially for cross-city or cross-type comparisons.

- **Highlight Functional Features in Listings (Bedrooms & Bathrooms)**

Bedrooms and bathrooms have **moderate correlation with price (r = 0.30 and 0.29)** but a **very strong correlation with each other (r = 0.86)**. Listings with **balanced room-to-bathroom ratios** are more attractive to buyers and should be emphasized in marketing.

- **Optimize Pricing Based on Property Type and City**

Since **price per sqft varies significantly** across both **cities and property types**, pricing strategies should reflect **local market conditions**. For instance, **farmhouses consistently command higher average prices** due to exclusivity and land size.

- **Use Price Categories for Targeted Marketing**

Segmenting listings into '**Low**, **Mid**, **High**, and **Luxury**' enables **better targeting of buyers** with specific budget ranges and expectations. This also helps in designing suitable **financing and promotional strategies**.

- **Consider Regional Differences for Diversified Investment**

The wide regional disparity — with **Azad Kashmir** having the **highest average prices** and **Gilgit Baltistan** the lowest — suggests that developers and investors should consider **geographic diversification** to balance **risk and return**.

What Drives Property Prices in Pakistan?

What Drives Property Prices in Pakistan?

- Location – Prime areas like DHA, Gulberg, and Bahria Town have higher prices.
- Property Type – Farmhouses and houses are priced higher than apartments.
- Bedrooms & Bathrooms – More rooms moderately increase price.
- Price per Sqft – A stronger value indicator than total area.
- Property Age – Older properties in urban centers often cost more.
- Regional Differences – Prices vary widely across cities and provinces.
- Area Size – Shows weak correlation with price; not a key driver.

In summary, property prices in Pakistan are driven by the interplay of location desirability, size, type, and age of the property, with location and property size being the most influential factors.

8. Conclusion & Next Steps

This exploratory analysis reveals **pricing patterns**, **regional trends**, and **listing inconsistencies** in the Pakistani real estate market. The insights derived can help:

- **Investors** prioritize locations offering higher value per sqft
- **Developers** optimize listings by aligning with local market trends
- **Zameen.com** improve data consistency and user filters

Next Steps:

- Apply predictive models (e.g., regression) to estimate price based on features
- Automate outlier detection
- Integrate geospatial analysis for mapping hotspots