

Exploratory Data Analysis (EDA) Report Of

Zameen.com properties by
Somia

Overview

This report presents a comprehensive exploratory data analysis of a Pakistan-wide real estate dataset containing property listings from major cities including Karachi, Lahore, Islamabad, Rawalpindi, and Faisalabad.

The goal of this analysis is to understand the patterns that drive property prices, identify top-performing areas, handle missing values effectively, and extract meaningful trends through descriptive statistics and visual exploration.

Objectives of the Analysis

This study focuses on answering key questions:

What are the most influential factors driving property prices in Pakistan?

Dataset Summary

The dataset includes thousands of property listings with attributes such as: City, Location, Area, Price, Bedrooms, Bathrooms, Kitchens, Store Rooms, Property Type (House, Plot, Flat, Commercial), Purpose (Sale/Rent), Floors, Parking Spaces. Initial inspection shows the dataset contains missing values across several numerical variables, especially for Built-in Year, Floors, and Parking Spaces, which required careful handling before analysis.

Methodology

This analysis includes:

1. Descriptive Statistics

Summary of distributions for numerical features

Range, mean, median, and variability insights

Identification of skewness and outliers

Missing Value Handling

Median imputation for skewed variables

Removal of unusable or invalid records

Standardization of area units into Marla

3. Visual Explorations

Distribution plots (Price, Area, Bedrooms, Bathrooms)

City-wise and area-wise bar charts

Purpose distribution (Sale vs Rent)

Property type distribution

Correlation heatmap

Trends across top cities and areas

4. Insights & Findings

Top cities by price

Highest-demand areas

Relationship between structure, area, and price

Detailed Descriptive Statistics & Missing Value Handling

Descriptive Statistics

A detailed statistical summary was generated for all major numerical features, including Price, Area (converted to Marla), Bedrooms, Bathrooms, Floors, Parking Spaces, and Built-in Year.

Key observations include:

Price distribution is highly right-skewed, indicating the presence of luxury or high-end properties significantly above the average.

Area (in Marla) also shows a wide range, reflecting the diversity of property sizes across cities.

Bedrooms and bathrooms follow a moderate distribution, with most properties having 3–5 bedrooms and 3–4 bathrooms.

Built-in Year contains older properties mixed with newly built ones, showing a broad historical spread of development.

Several features display outliers, particularly Price, Area, and Floors, which is expected in real estate data.

These statistics help in identifying central tendencies, spread, and irregularities before moving into deeper analysis.

Handling Missing Values

The dataset contained missing values across several fields, especially:

Built-in Year

Bathrooms

Bedrooms

Numerical Columns

Median imputation was used for skewed columns like Price, Area, Bedrooms, and Bathrooms since median is more robust against extreme values.

Invalid values (such as 0 for Built-in Year) were converted to NaN and replaced using median.

Columns containing non-numeric characters were cleaned and converted using `to_numeric(errors='coerce')`.

Area Standardization

Area values originally came in Kanal, Marla, Sq Yd.
To maintain consistency:

Kanal converted to Marla ($\times 20$)

Sq Yd converted to Marla ($\div 30.25$)

Marla kept as baseline unit

This conversion allowed consistent comparison across listings.

Visual Explorations

Several visualizations were created to understand the data better:

Histograms for Bedrooms, Bathrooms, Floors, Area, and Price

Boxplots to identify outliers

City-wise distribution of property prices

Bar charts for top areas and property types

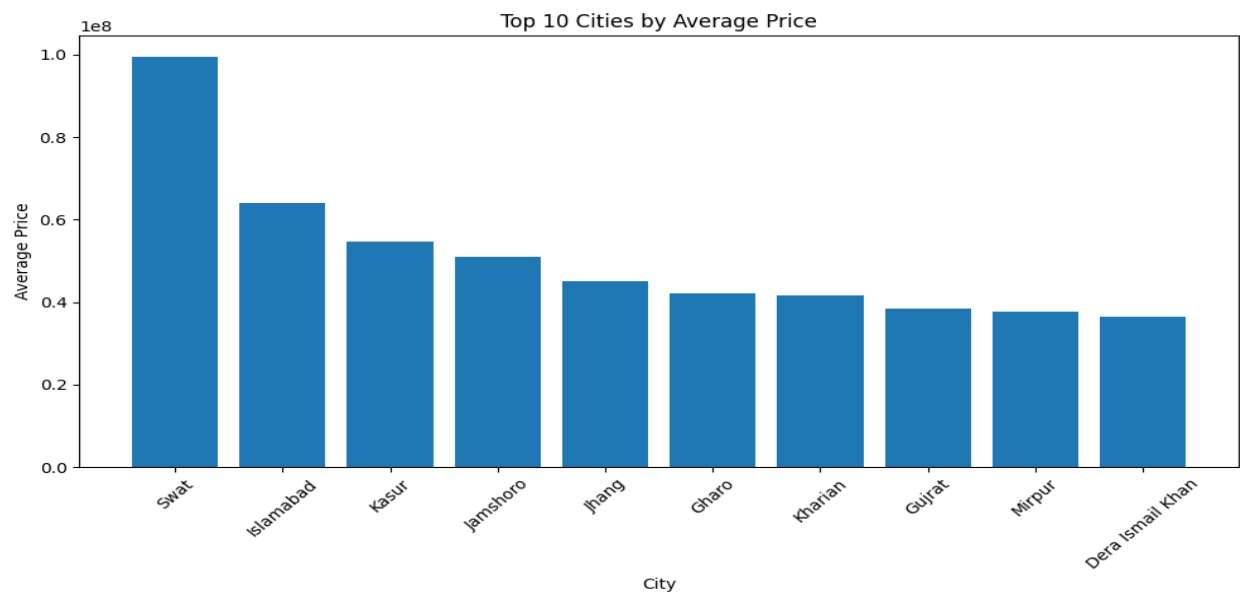
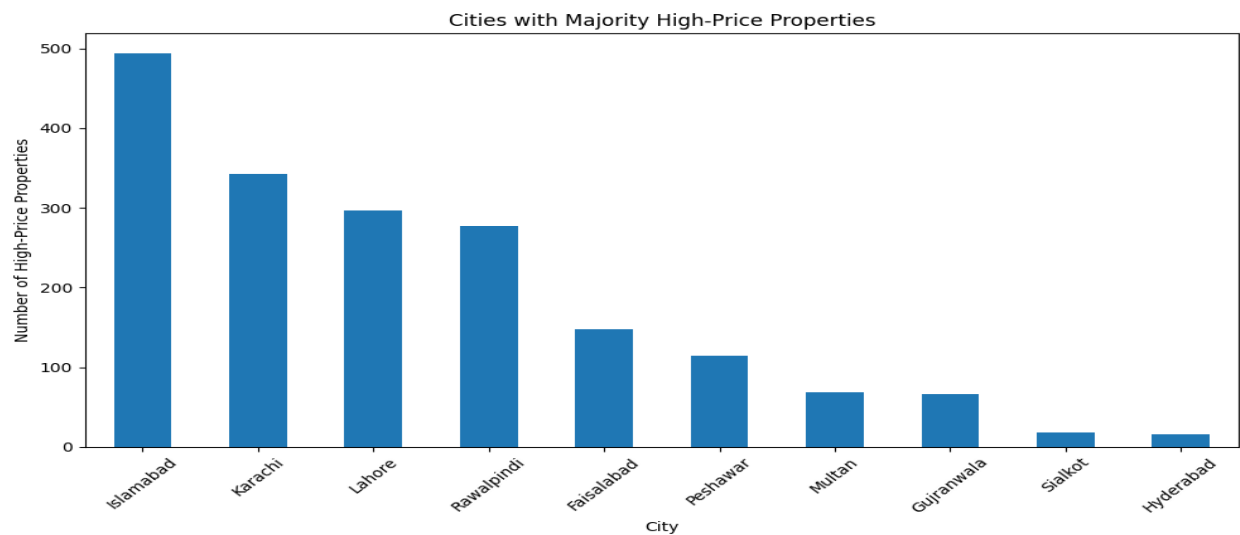
Heatmap to inspect correlations between numerical features

These visuals helped highlight patterns not visible through statistics alone

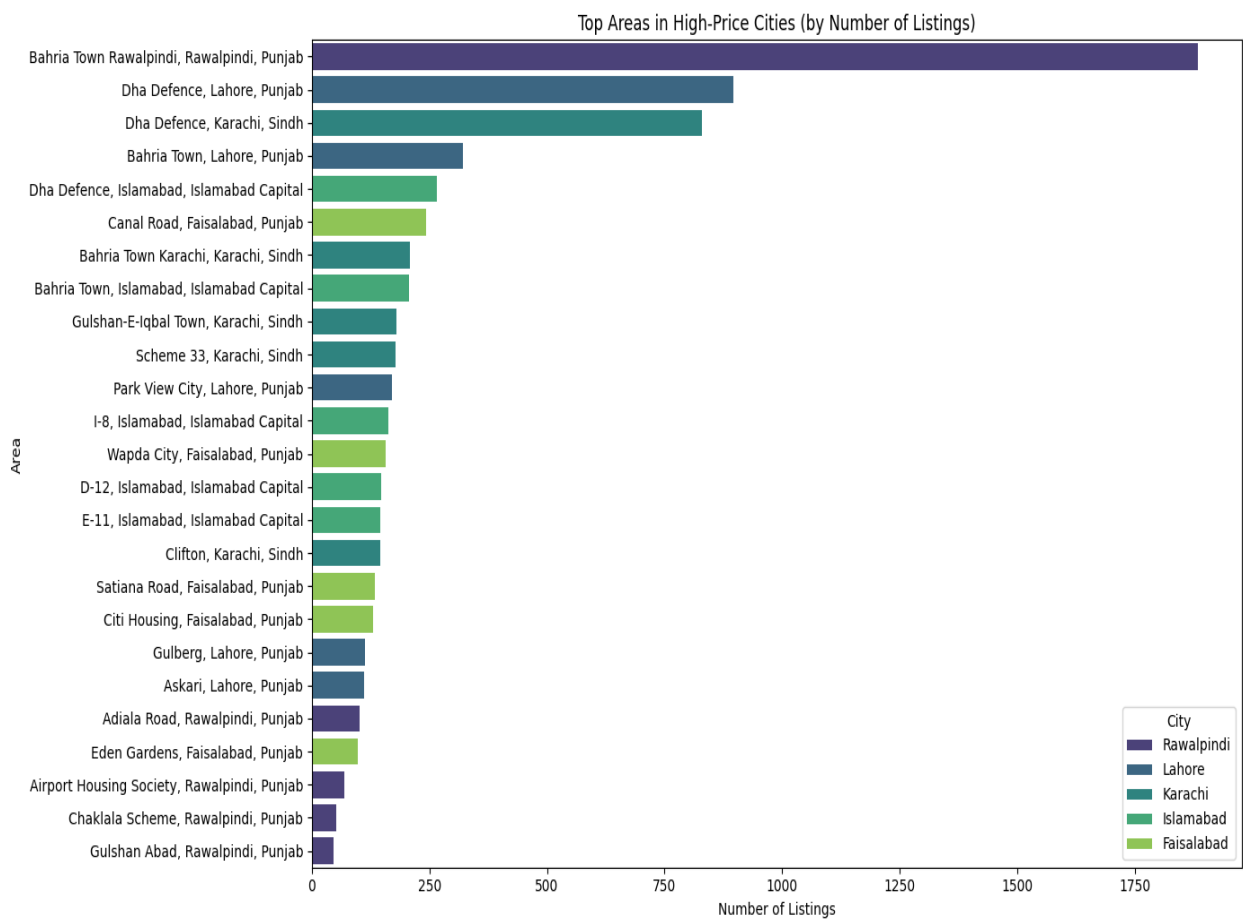
Key Findings

Top Cities by Highest Prices and Average Price

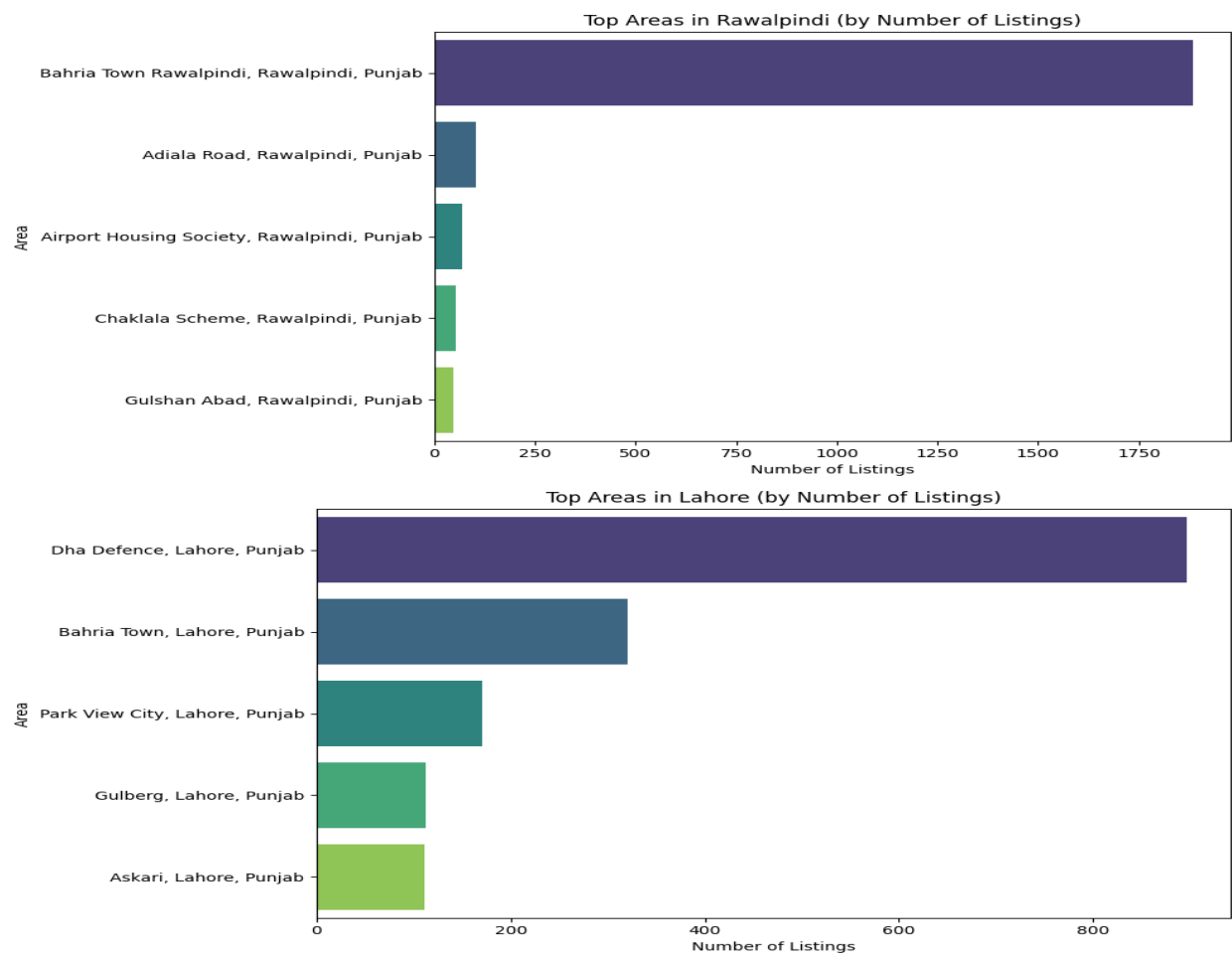
Cities with the most expensive property listings:

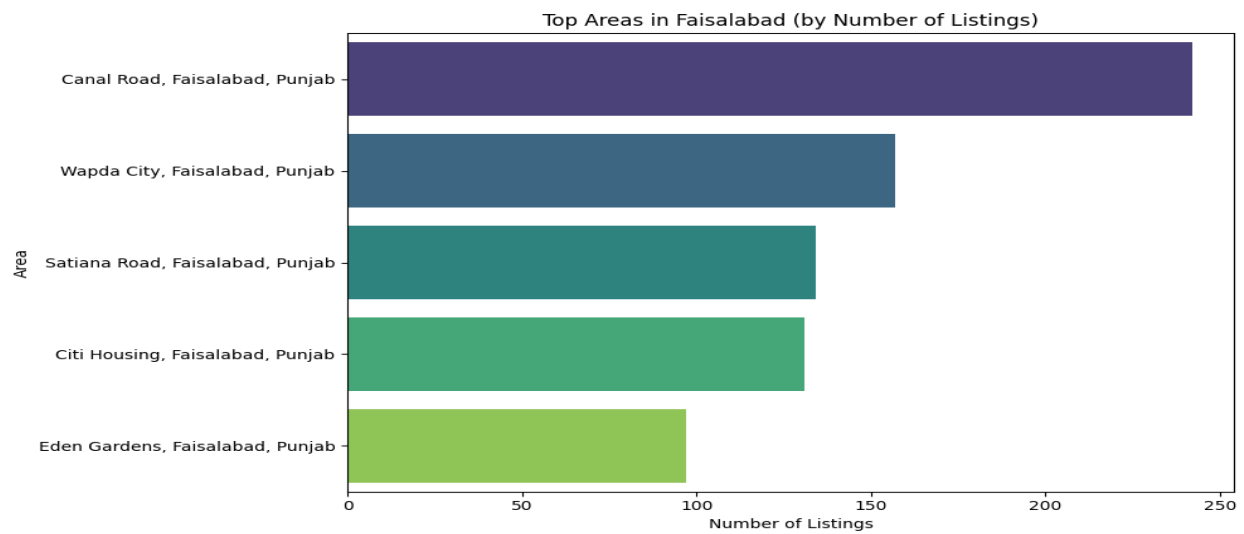
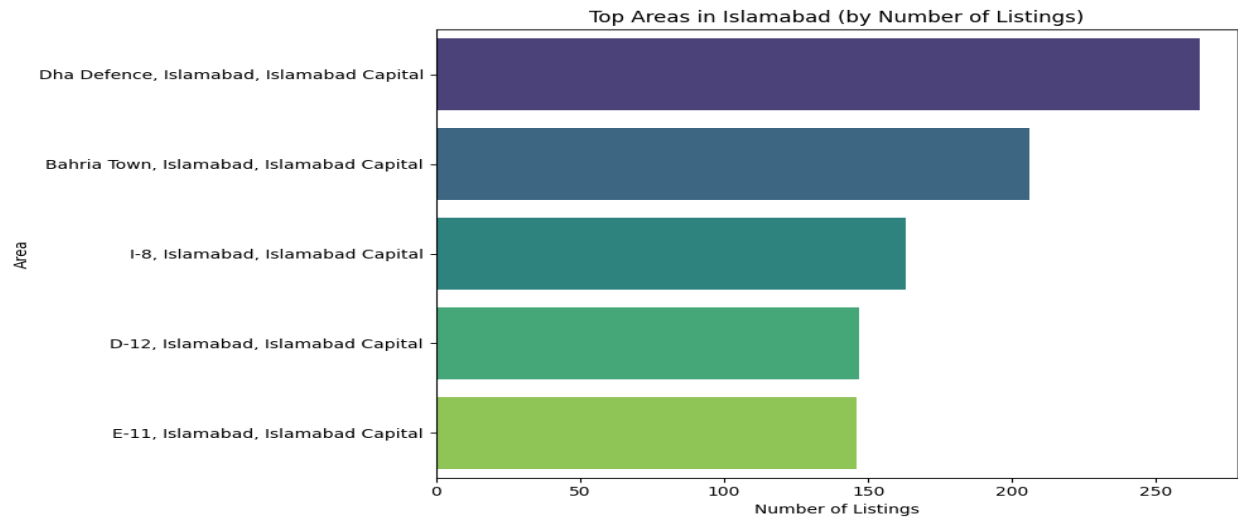
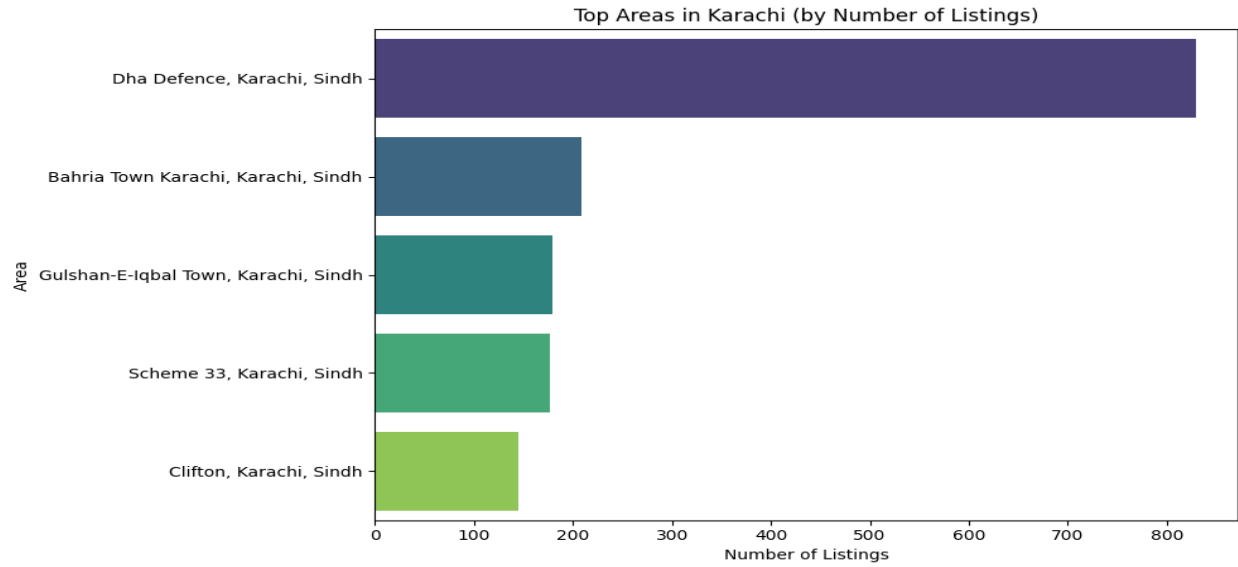


Top High-Demand Areas



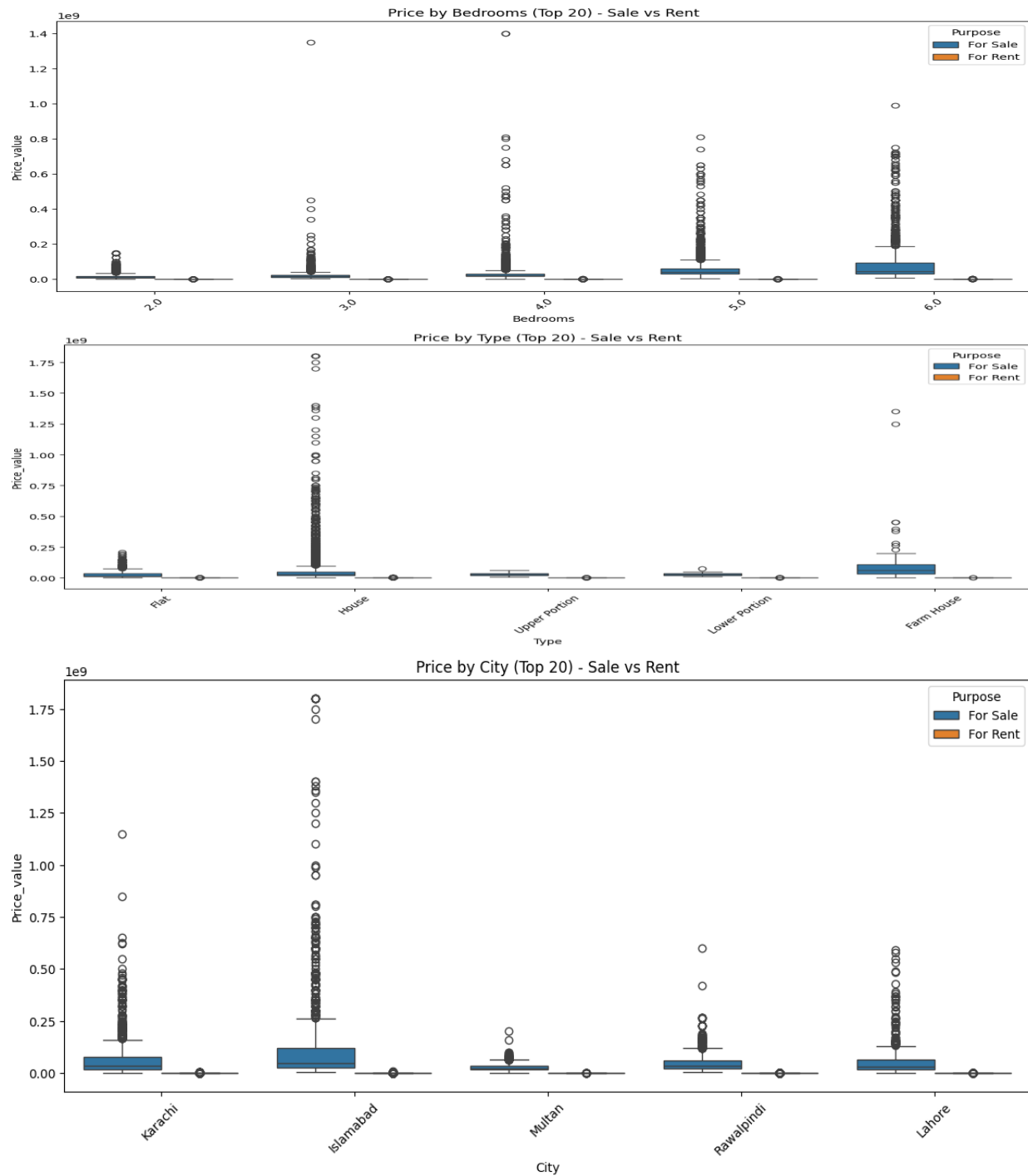
Top High-Demand Areas in Top Cities





Purpose Distribution (Sale vs Rent)

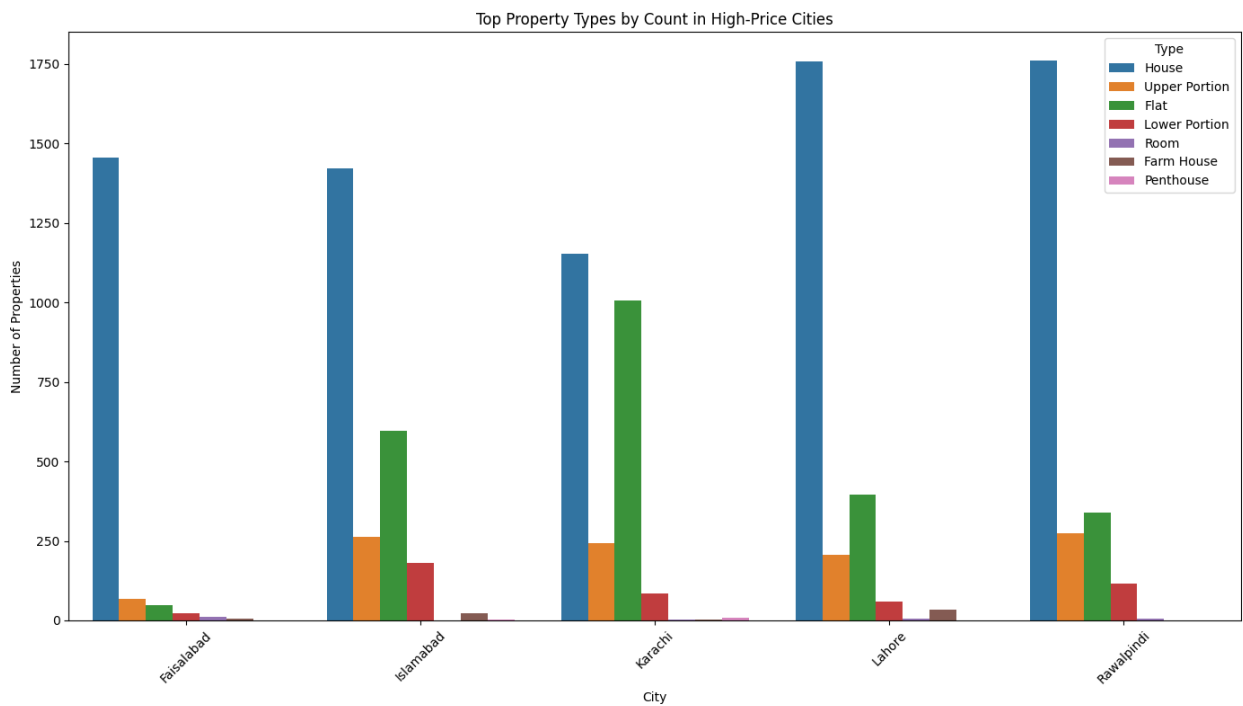
Rental properties are low compare to Sale properties and Most high orices proeptrties are around 6 bedromms ,farmhouses in cities like Karachi islamad



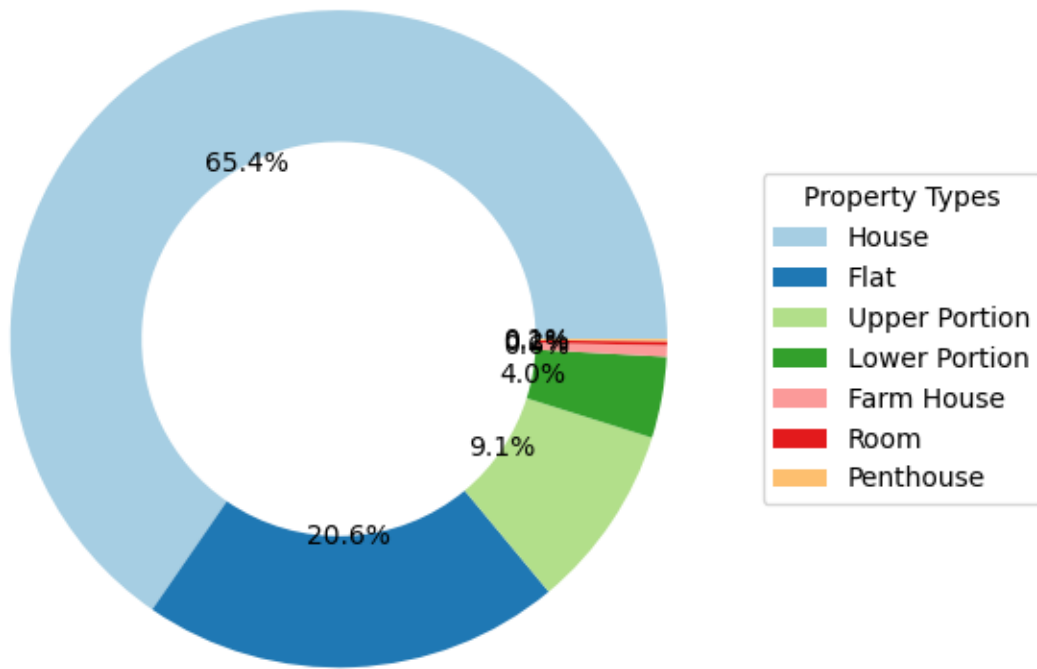
Property Type Trends

The most common property types overall:

- 1. Houses
- 2. Upper portion
- 3. Flats

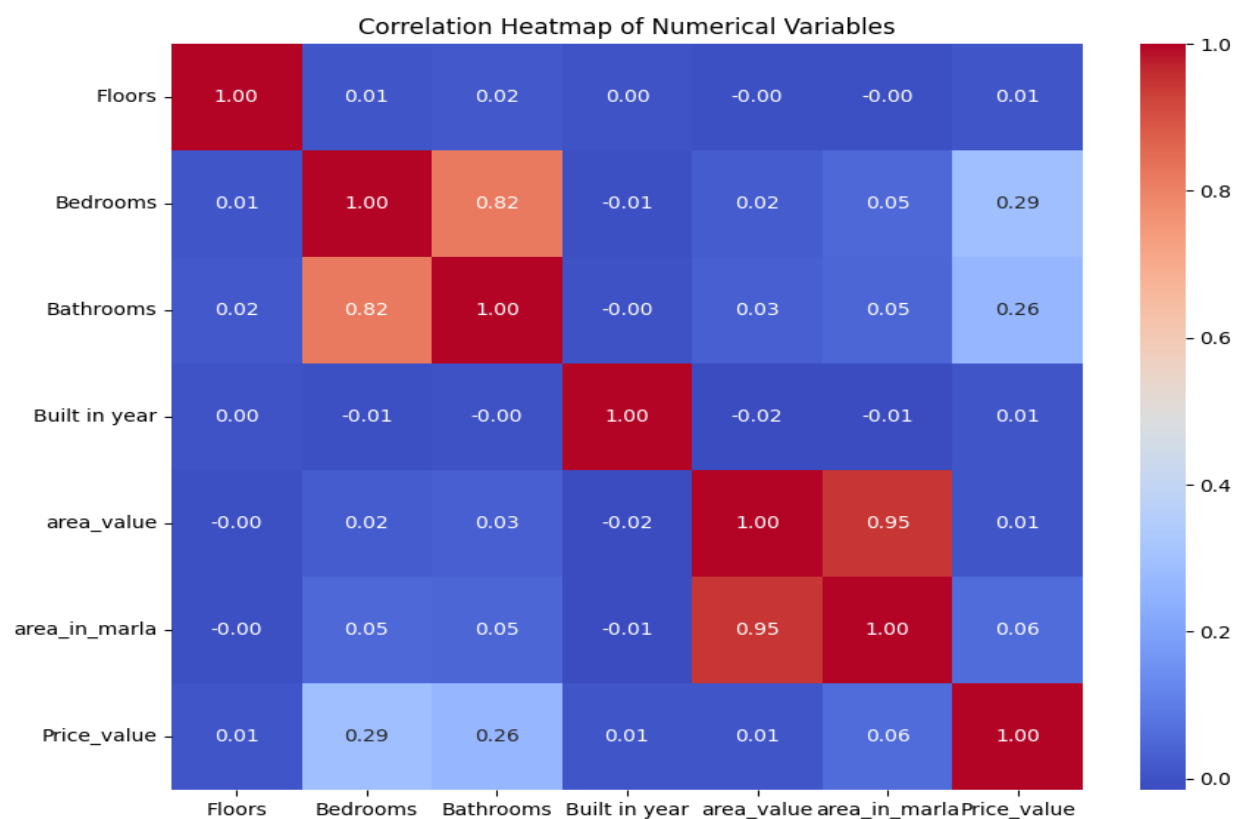


Distribution of Properties by Type

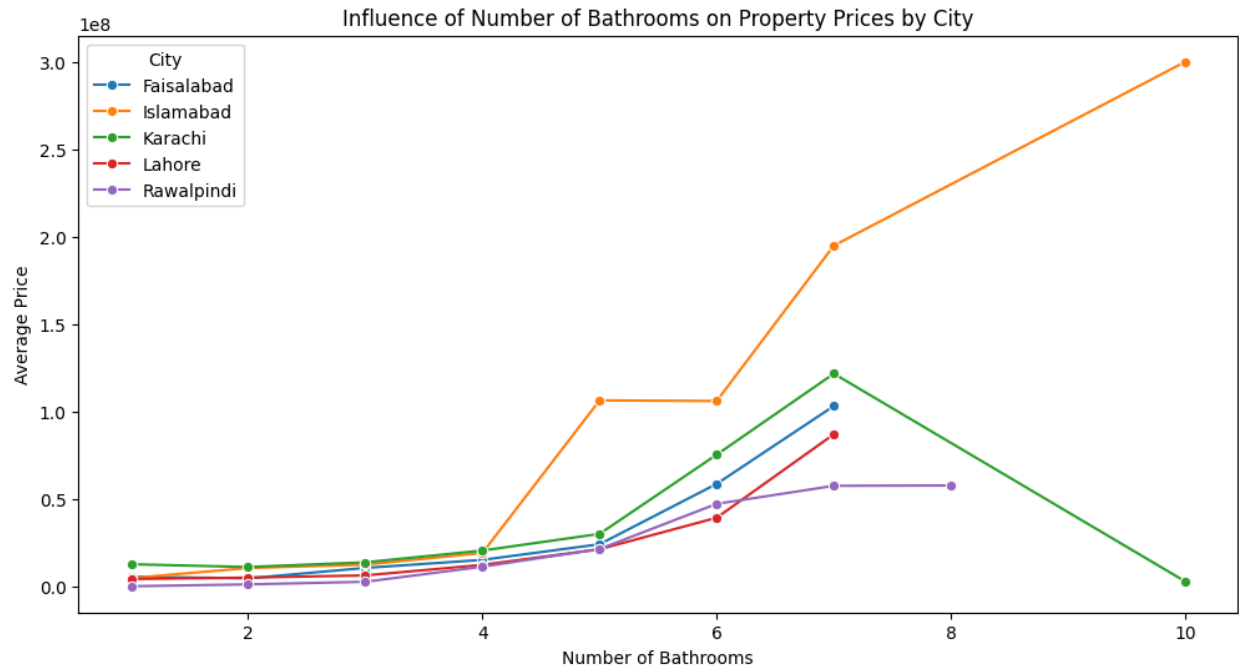
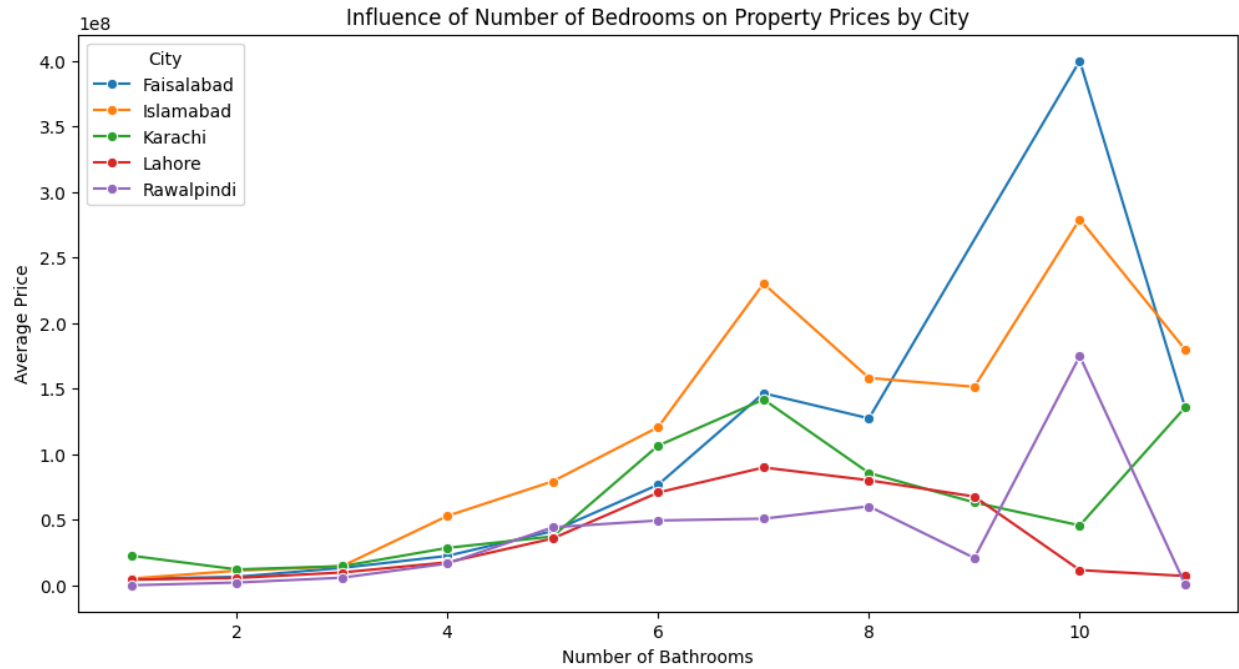


Correlation Analysis

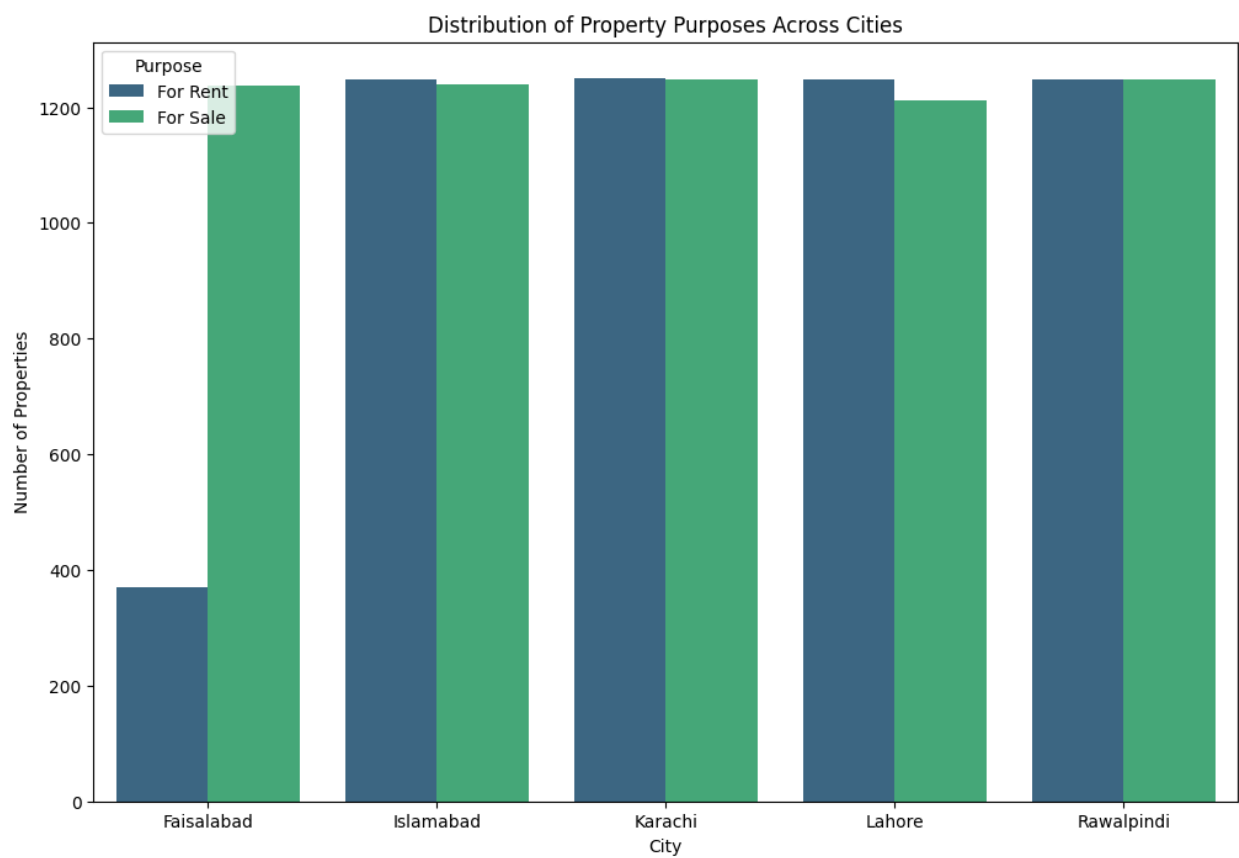
Bedrooms and bathrooms show a moderate positive correlation with price.



Influence of Number of Bathrooms on Property Prices by City



Distribution of Property Purposes Across Cities



Summary

Property prices in Pakistan are mainly driven by **location, area size, and bedroom count**. Major cities such as **Islamabad, Karachi, and Lahore** show the highest price levels, with premium areas like **Bahria Town and DHA Karachi** leading the market.

After cleaning and standardizing the dataset, the analysis confirms that **Houses are the most common property type**, followed by plots and apartments. Overall, **location and property size** emerge as the strongest contributors to price variation across cities.