REPORT

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

Project Objective:

To extract actionable insights from property listings on Zameen.com -such as pricing trends, neighborhood comparisons, and listing quality -that can help real estate investors make informed decisions.

1. Problem Statement

Property prices are influenced by broader economic trends and specific location-based factors. Key drivers include a comparative analysis of different neighborhoods and housing societies to identify areas with strong appreciation potential. The intrinsic quality of a listing, such as the number of bedrooms, bathrooms, and available amenities, also significantly impacts its value. By understanding these dynamics, real estate investors can identify profitable opportunities. In essence, a combination of market forces, location, and property-specific characteristics dictates the price of real estate in Pakistan.

2. Data Understanding & Preprocessing

Dataset file: Scarped Zameen.com.xlsx Total rows: 18255 Detected key columns (heuristic): price -> Price, area-> Area, city -> City, beds -> Bedrooms, baths -> Bathrooms, property_type -> Type, date -> None

For Data Understanding, we performed an initial exploration of the `df` DataFrame. `df.head()` gives a quick preview of the top rows, column names, and data types. The `df.info()` command provides a summary of the DataFrame's structure, including non-null counts and data types, which is useful for identifying missing values. Finally, `df.describe(include='all')` generates a comprehensive statistical summary for both numerical and categorical columns, revealing central tendencies, dispersion, and value frequencies. Together, these commands offer a foundational understanding of the dataset's characteristics and quality before further analysis.

1. This line uses the `drop_duplicates()` method from the pandas library to identify and remove any rows that are identical across all columns

2. Then, cleaning and converting the 'Price' and 'Area' columns in your DataFrame to numeric types.

3. Missing Value Treatment

Then, fill in the gaps in data by replacing missing values in numerical columns with the median and missing values in object columns with the mode. This is a common technique in data preprocessing to ensure that missing data doesn't interfere with subsequent analysis or modeling

4. Data Cleaning & Consistency

In this step, we use fuzzy string matching to standardize the city names in 'City' column, which is useful for ensuring that variations of the same city name are treated as the same city for analysis.

Next is removing outliers, using the IQR method, it identifies and removes data points in the 'Price' column that are considered outliers based on the IQR method. This helps to create a more robust dataset for analysis by reducing the influence of extreme values.

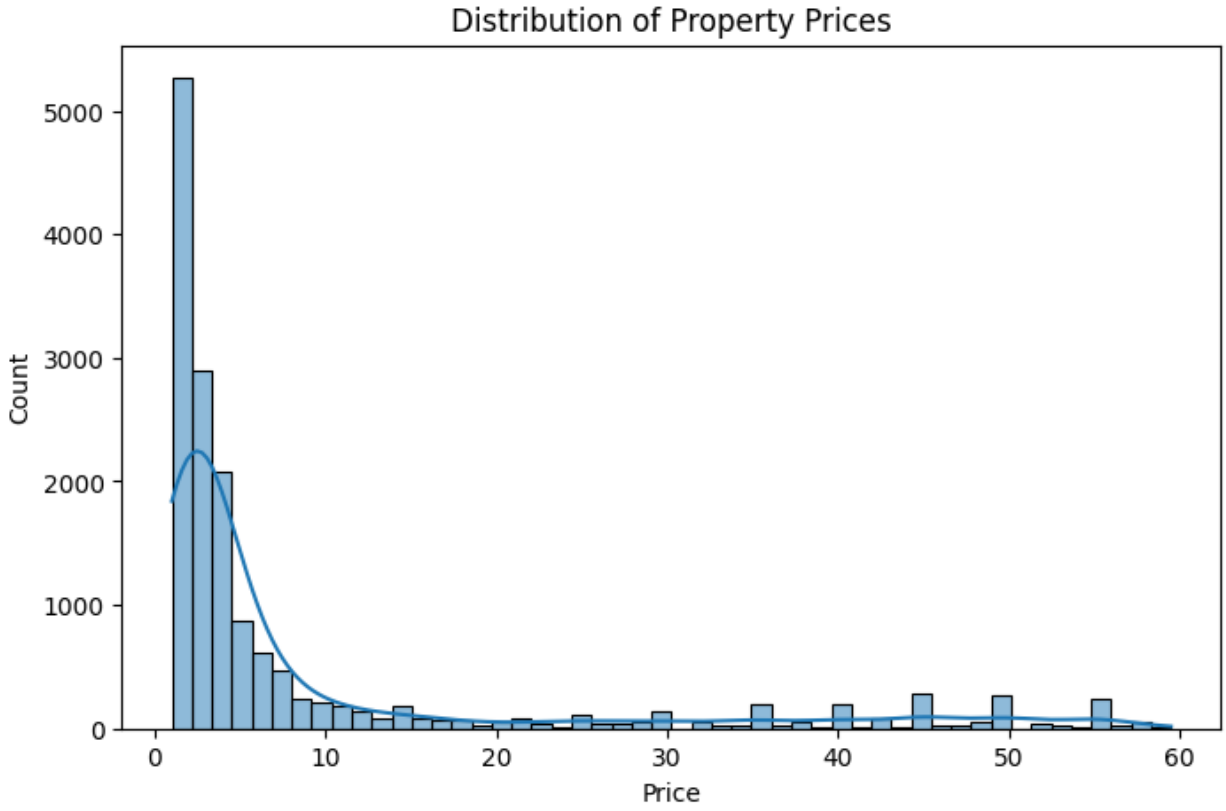
5. Feature Engineering

Then perform feature engineering on the code by creating new features based on existing ones. These new features ('Price_per_sqft', 'Year', and 'Month') can be valuable for further analysis, allowing you to explore relationships and trends that might not be apparent from the original data.

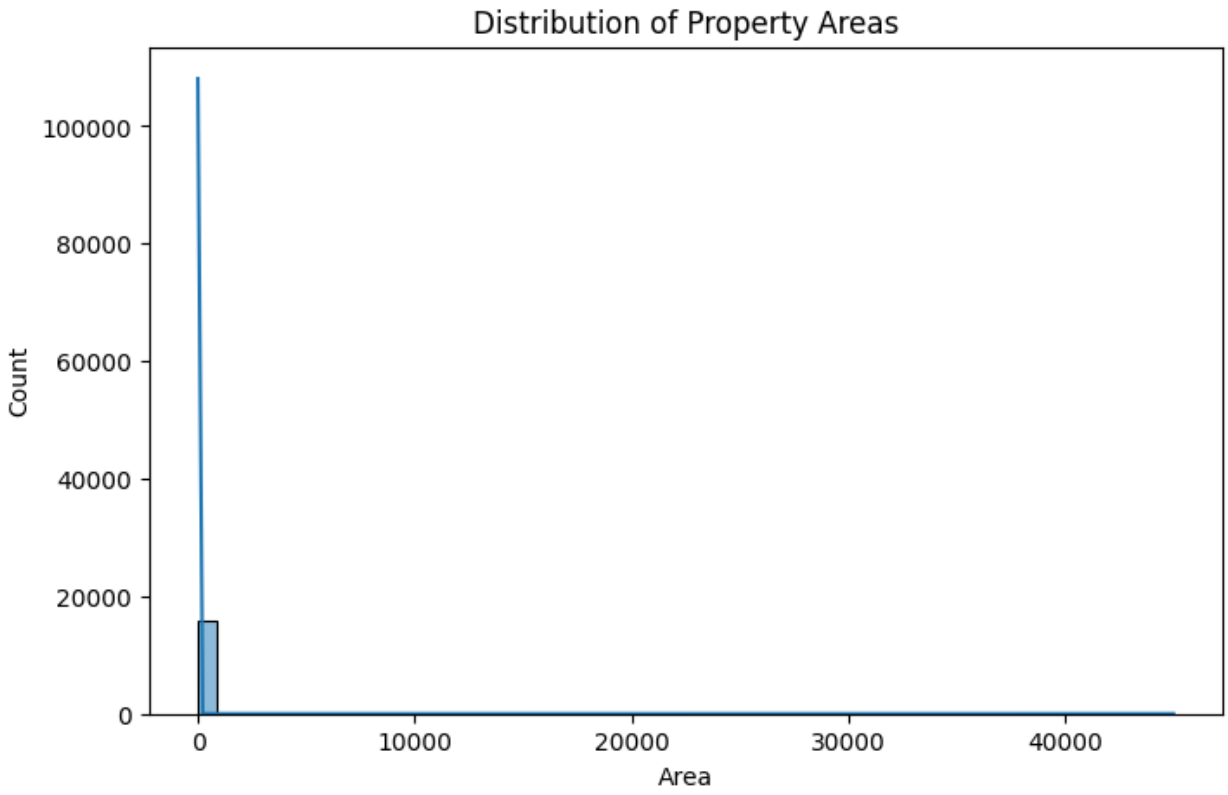
6. Univariate & Bivariate Analysis

The univariate and bivariate analysis to understand the distributions and relationships within your real estate data. Here's a breakdown:

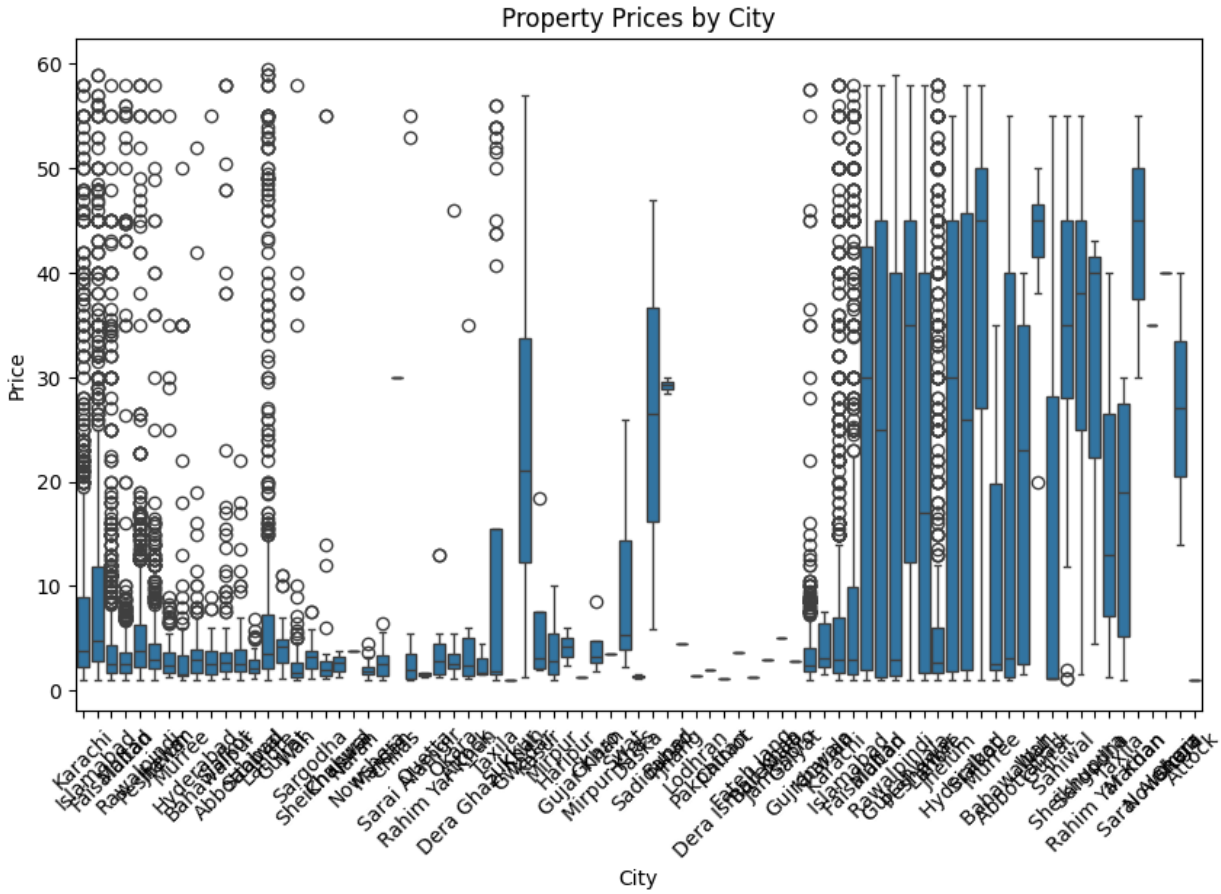
1. **Distribution of Property Prices:** It generates a histogram with a Kernel Density Estimate (KDE) overlay to visualize the distribution of property prices (`df['Price']`). This helps to see the frequency of different price ranges and the overall shape of the distribution.



2. **Distribution of Property Areas:** Similar to the price distribution, it creates a histogram with a KDE for the property areas (`df['Area']`). This shows the distribution of property sizes.



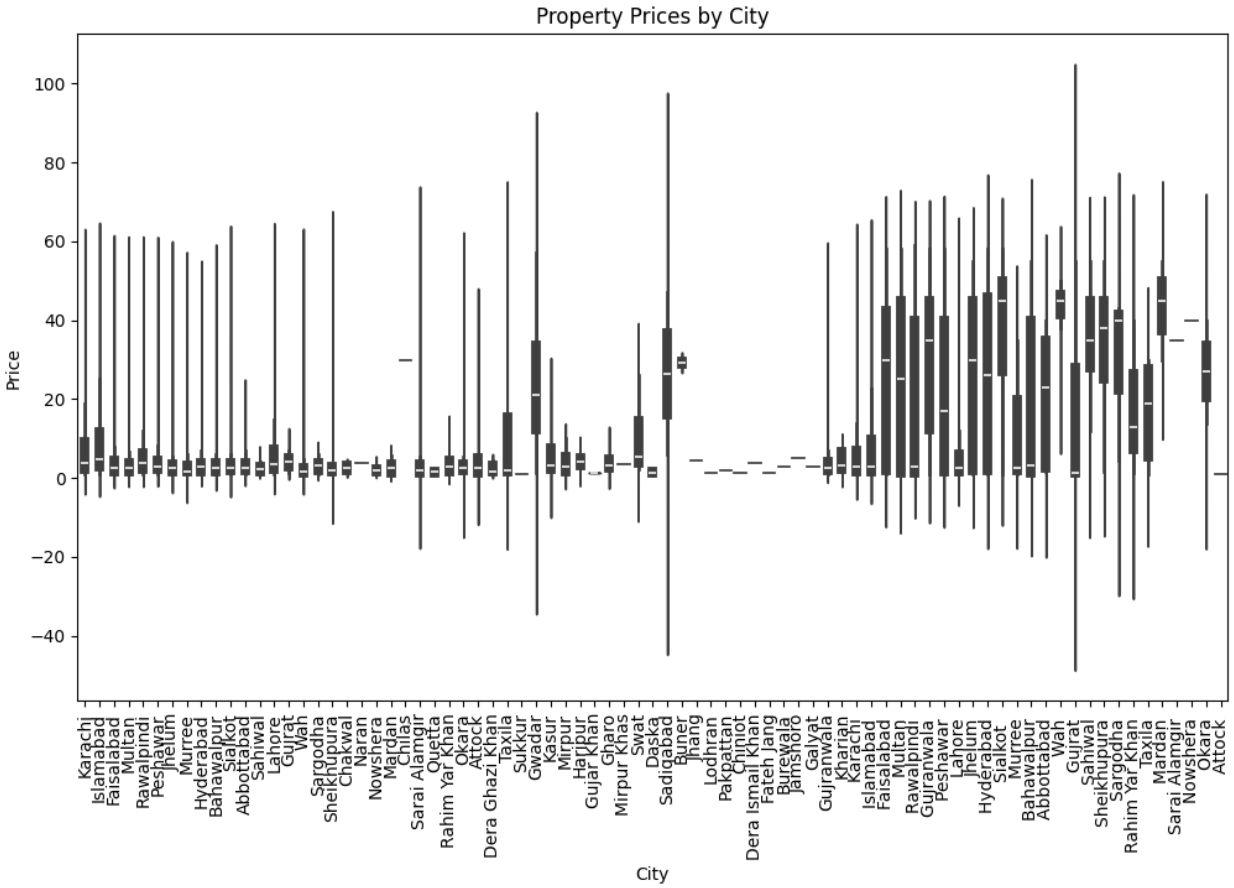
3. **Property Prices by City:** A box plot is generated to compare the distribution of property prices across different cities (`'City'` vs. `'Price'`). Box plots are useful for visualizing the median, quartiles, and potential outliers for each city. The x-axis labels are rotated for better readability.



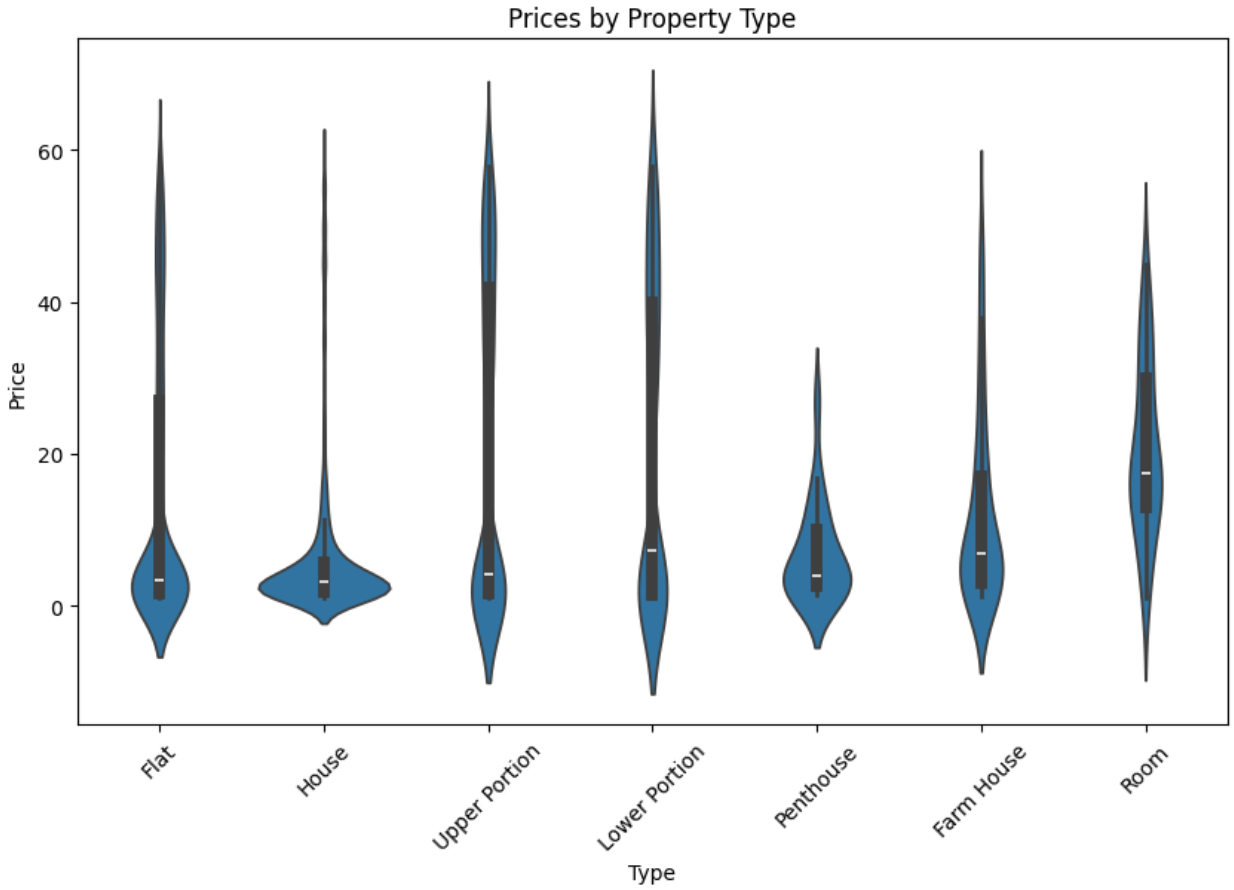
4. **Correlation Heatmap:** A heatmap is generated to visualize the correlation between numerical columns in the DataFrame (`df.corr(numeric_only=True)`). The `annot=True` argument displays the correlation values on the heatmap, and `cmap="coolwarm"` sets the color scheme. This helps identify which numerical features are strongly correlated with each other.

The violin plots or box plots to visualize the relationship between property `Price` and different categorical variables: `City`, `Type` (property type), and `Bedrooms`. Here's a breakdown:

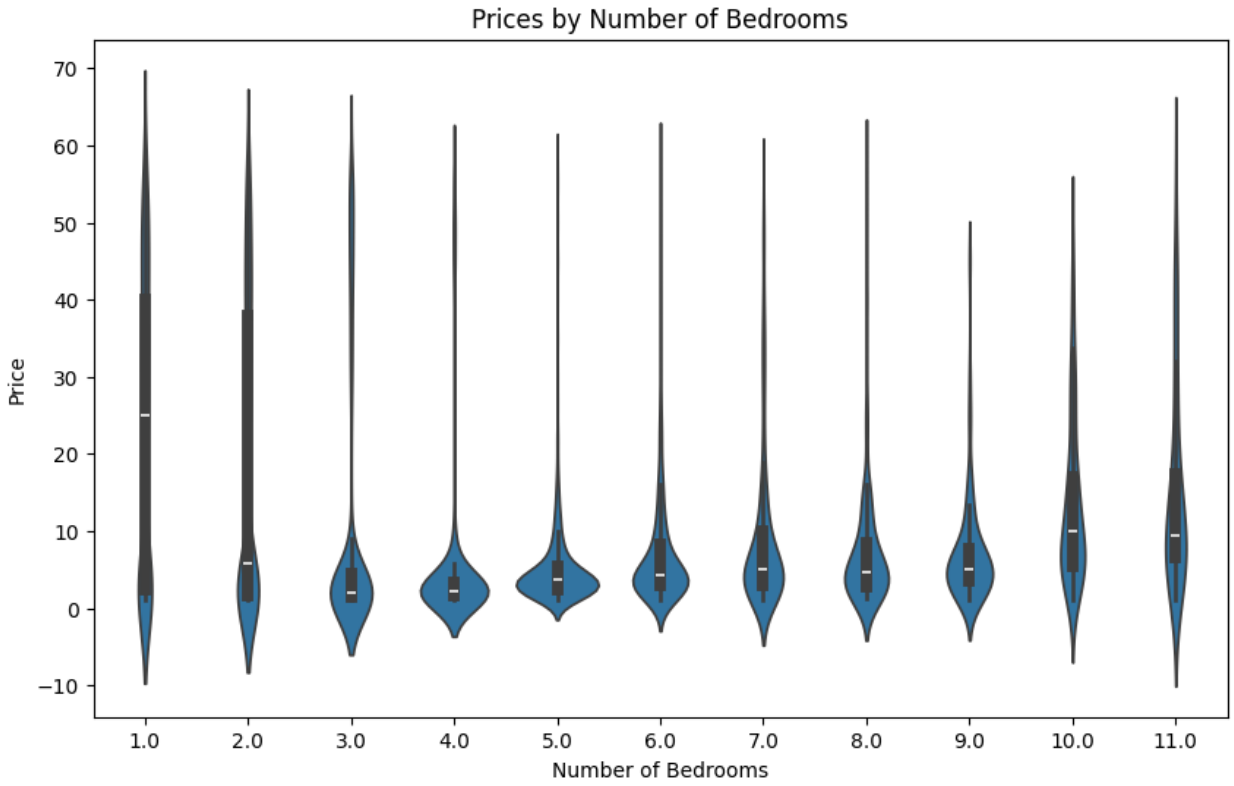
1. **Price vs. City:** It creates a violin plot showing the distribution of property prices for each city. This helps to compare price ranges and distributions across different locations. The city names on the x-axis are rotated for better readability.



2. **Price vs. Property Type:** This part generates a violin plot (or box plot) to visualize the price distribution for different property types. It allows you to see how the price varies depending on whether the property is a house, apartment, etc. The property types on the x-axis are rotated.



3. **Price vs. Bedrooms:** This section aims to plot the price distribution based on the number of bedrooms. It first attempts to convert the 'Bedrooms' column to a numeric type, handling any non-numeric values by converting them to `NaN`. It then drops rows with `NaN` values in the numeric 'Bedrooms' column for plotting. If there's data remaining, it generates a violin plot to show how prices relate to the number of bedrooms. If the conversion or data cleaning results in an empty dataset for plotting, it prints a message indicating that the plot could not be created.



7. Insights & Recommendations

Insights:

1. **Price and Area Distribution:** The histograms show the distribution of property prices and areas. The price distribution appears to be right-skewed, indicating that most properties are in the lower price range, with fewer properties at higher price points. The area distribution shows the typical sizes of properties in the dataset.
2. **Relationship between Price and City/Property Type:** The box plots for Price by City and Price by Property Type clearly demonstrate that both the location (city) and the type of property significantly influence the price. Some cities have much higher median prices and wider price ranges than others, and different property types (e.g., houses vs. flats) also show distinct price characteristics.

Recommendations:

1. **Target Specific Cities/Property Types:** Investors should use the box plots to identify cities and property types that align with their investment goals. For example, if seeking higher returns, focus on cities with higher median prices. If looking for affordability, explore cities with lower price ranges. Similarly, analyze which property types offer the best potential based on price distribution and market demand.
2. **Consider Area as a Key Factor:** The correlation heatmap shows the relationship between numerical features. The positive correlation between Area and Price suggests that larger properties tend to be more expensive. Investors should consider the price per square foot (a

feature engineered in a previous step) in conjunction with the total price and area to evaluate the value proposition of a property.
3. **Investigate Outliers:** While outliers were removed in a previous step based on the IQR method, it's worth noting that extreme values in the original data might represent luxury properties or unique market conditions. Depending on the investment strategy, further investigation into these high-value properties could be beneficial.

8. Conclusions:

Summary:

Based on the data analysis, it is evident that property prices are heavily influenced by both location (City) and property type. The data exhibits a right-skewed price distribution, with a positive correlation between property area and price. Significant missing data in many feature columns highlights the importance of thorough listing evaluation.

Suggestions for Stakeholders (Investors):

1. **Strategic Targeting:** Investors should strategically target specific cities and property types that align with their risk tolerance and return expectations, leveraging the observed price variations.
2. **Value Assessment:** Utilize the price per square foot metric alongside total price and area to assess the relative value of properties, especially in areas with diverse property sizes.
3. **Due Diligence on Listing Quality:** Prioritize listings with comprehensive information and conduct thorough due diligence for listings with significant missing data to ensure a clear understanding of the property's true value and potential.