Real Estate Price Analysis in Pakistan Zameen.com

Introduction

This report aims to extract actionable insights from property listings on Zameen.com to empower real estate investors in Pakistan with data-driven decision-making. The primary objective is to identify and understand the key factors that influence property prices across various cities and property types in the Pakistani real estate market. By analyzing the provided dataset, we seek to uncover the underlying drivers of property valuation, enabling investors to make more informed choices regarding buying, selling, or investing in properties.

1. Data Description

The dataset used for this analysis was obtained from Zameen.com, a popular real estate portal in Pakistan. It comprises property listings and includes various details about each property.

The initial dataset (df) contains information on 18255 rows and 59 columns, as shown by the df.shape output.

For this analysis, a subset of the most relevant features was selected to create a cleaned DataFrame, new_df. After removing duplicate entries and selecting these columns, the new_df contains information on 17974 rows and 14 columns, as indicated by the new_df.info() output and the duplicate removal steps.

The key features utilized in the new_df DataFrame are:

- **Title:** The title of the property listing.
- **City:** The city where the property is located. The dataset covers properties in various cities across Pakistan, including major metropolitan areas and smaller towns, as seen in the unique values of the 'City' column.
- **Area:** The size of the property, originally in various units (Sq. Yd., Marla, Kanal) and converted to a standardized unit (sqft) for analysis.
- **Price:** The listed price of the property, originally in a text format with currency and units (PKR, Crore, Lakh, Thousand) and converted to a numerical format (PKR) for analysis.
- **Type:** The type of property being listed (e.g., Flat, House, Upper Portion, Lower Portion, Penthouse, Farm House, Room). The Type.unique() output shows the different property types present.
- **Bedrooms:** The number of bedrooms in the property.
- **Bathrooms:** The number of bathrooms in the property.
- **Description:** A textual description of the property.
- **Location:** The specific location or neighborhood of the property within the city.

These selected features provide essential information for exploring the relationships between property characteristics and their prices in the Pakistani real estate market.

2. Methodology

This section details the steps taken to prepare the dataset for analysis, including cleaning, transformation, handling of outliers, and feature engineering.

Data Cleaning

Initial inspection revealed missing values and duplicate entries. The following steps were taken to address these issues:

a) Handling Missing Values:

'Title' and 'Type': Missing values were imputed using the mode (most frequent value) as these are categorical/short text columns where the mode represents the most common property type or title. 'Location': Missing location data was forward-filled (ffill). This approach assumes that a property's location is likely to be similar to the preceding listing, which is a reasonable assumption for sequentially scraped data from a specific area.

'Price' and 'Area': Missing numerical values were filled with the median of their respective columns. The median is preferred over the mean for skewed distributions (which are common for price and area data) as it is less affected by extreme values.

'Bedrooms' and 'Bathrooms': Missing values were imputed using the mode. These represent counts, and the mode maintains the realistic integer nature of the data.

'Description': Missing descriptions were replaced with the placeholder 'No description provided' to avoid nulls and indicate the absence of a detailed description.

Handling Duplicate Rows: Exact duplicate rows were identified and removed using the drop_duplicates() function. This ensures that each listing in the dataset is unique, preventing overrepresentation of certain properties in the analysis.

b) Data Transformation

To ensure consistency and enable numerical analysis, the 'Price' and 'Area' columns were transformed:

- 'Price': The 'Price' column, which contained text with currency and units (PKR, Crore, Lakh, Thousand), was converted into a consistent numerical format (PKR). A custom function was applied to extract the numerical value and multiply it by the appropriate factor based on the unit mentioned (Crore, Lakh, Thousand).
- 'Area': The 'Area' column, which contained values in different units (Sq. Yd., Marla, Kanal, Sqft), was converted into a single consistent unit (square feet). A custom function parsed the unit and applied the relevant conversion factor (1 Sq. Yd. = 9 Sqft, 1 Marla = 272.25 Sqft, 1 Kanal = 5445 Sqft).

c) Outlier Handling

- Outliers in numerical columns ('Area', 'Price', 'Bedrooms', 'Bathrooms') were addressed using the Interquartile Range (IQR) method.

- IQR Capping: For each numerical column, the first quartile (Q1) and third quartile (Q3) were calculated. The IQR was determined as Q3 - Q1. Values falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR were identified as outliers. Instead of deleting these rows, the outliers were capped at the lower and upper bounds calculated by the IQR method. This approach was chosen over deletion to retain as much data as possible, as extreme values in real estate data often represent genuinely large or expensive properties rather than errors.

d) Feature Engineering

Several new features were created to provide additional insights and improve the potential for future modeling:

- 'Price_Numeric': A numerical representation of the 'Price' after transformation, used for quantitative analysis.
- 'Price_per_sqft': Calculated by dividing 'Price_Numeric' by the standardized 'Area'. This metric allows for comparison of property values on a per-unit-area basis, which is useful for understanding value independent of size.
- 'Total_Rooms': Created by summing the 'Bedrooms' and 'Bathrooms' counts. This feature provides a combined measure of the property's size in terms of usable rooms.
- 'Price_Log': The natural logarithm of 'Price_Numeric' (using np.log1p for handling zero values). This transformation was applied to address the skewness observed in the 'Price_Numeric' distribution, making it more suitable for analyses that assume normality.
- 'City_Standardized': The 'City' column was standardized using fuzzy matching with a predefined list of common city names. This step aimed to correct for potential variations or typos in city names, ensuring consistency in geographical analysis. A similarity score threshold of 80 was used to determine a match.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the distribution of key variables, identify relationships between features, and visualize trends in property prices.

Univariate Analysis

- Price Distribution: The histogram of 'Price_Numeric' showed a heavily right-skewed distribution, indicating that most properties are concentrated at lower price points, with a long tail extending towards very high-priced properties. The log transformation of the price ('Price_Log') resulted in a distribution that is much closer to normal, which is often beneficial for modeling.
- Area Distribution: The histogram of 'Area' also exhibited a right-skewed distribution, similar to price, suggesting a prevalence of smaller properties and fewer very large ones.

Bivariate Analysis

- Correlation Heatmap: The correlation heatmap revealed the relationships between numerical features:
- 'Price_Numeric' shows a moderate positive correlation with 'Bedrooms' (0.36), 'Bathrooms' (0.40), and 'Total_Rooms' (0.40). This suggests that properties with more rooms tend to have higher prices. The correlation between 'Price_Numeric' and 'Area' (-0.007) is very weak, almost negligible. This is counterintuitive and might be influenced by the diverse units originally present in the 'Area' column and the outlier capping applied, or it could genuinely reflect that in some markets, the number of rooms is a stronger price driver than the total area.
- 'Price_per_sqft' has a strong positive correlation with 'Price_Numeric' (0.75), indicating that properties with a higher price per unit area are generally more expensive. This metric seems to capture value more effectively than raw area alone.
- 'Bedrooms' and 'Bathrooms' are highly correlated with each other (0.83) and with 'Total_Rooms' (0.96 and 0.95 respectively), which is expected as 'Total_Rooms' is their sum.
- Price Variation by Categorical Features:
- City: The violin plot of 'Price_Numeric' by 'City' clearly demonstrates significant price variations across different cities. Some cities show a wider range of prices and potentially higher median prices than others, highlighting the importance of location in determining property value.
- Property Type: The violin plot of 'Price_Numeric' by 'Type' shows distinct price distributions for different property types. Houses appear to have a broader price range and higher potential maximum prices compared to Flats, Upper Portions, or Lower Portions.
- Bedrooms: The box plot of 'Price_Numeric' by 'Bedrooms' indicates a general trend of increasing median price with an increasing number of bedrooms. The spread of prices also tends to increase for properties with more bedrooms.

4. Insights & Recommendations

Based on the exploratory data analysis of the Zameen.com property listings, several key insights regarding the drivers of property prices in Pakistan have been identified. These insights translate into actionable recommendations for real estate investors aiming to make informed decisions.

a) Key Insights from EDA

- Price Distribution Characteristics: The analysis revealed that property prices in the dataset are heavily skewed towards the lower end, with a smaller number of high-value properties. This suggests a broad base of affordable properties and a premium segment at the higher end of the market. The logarithmic transformation of price (Price_Log) helped to normalize this distribution, which is important for certain types of statistical modeling.
- Area vs. Rooms as Price Drivers: While 'Area' (in sqft) shows a surprisingly weak correlation with 'Price_Numeric' (-0.007), the number of 'Bedrooms' (0.36), 'Bathrooms' (0.40), and the combined 'Total_Rooms' (0.40) exhibit more significant

positive correlations with price. This indicates that for this dataset, the number of functional rooms within a property appears to be a stronger determinant of price than the overall built-up area.

- Price per Square Foot as a Value Indicator: The 'Price_per_sqft' metric shows a strong positive correlation with 'Price_Numeric' (0.75). This suggests that properties with a higher price per unit of area command a premium. This metric is valuable for comparing properties of different sizes on a standardized basis and can highlight properties in prime locations or with superior features that justify a higher per-square-foot value.
- Significant City-Based Price Variations: The violin plots clearly illustrate substantial differences in property price distributions across various cities in Pakistan. Major metropolitan areas likely have higher price ranges and potentially higher average prices compared to smaller cities. Understanding these city-specific market dynamics is crucial for targeted investment.
- Property Type Influence on Price: The analysis showed that 'Type' of property significantly impacts the price distribution. For example, 'House' properties tend to have a wider price range and higher potential maximum prices compared to 'Flat' or 'Portion' types. This highlights the need to consider the property type and its typical market value in a given location.
- Outlier Significance: The presence of significant outliers in 'Area' and 'Price' suggests the existence of properties with exceptionally large sizes or high values. While capped for analysis, these outliers likely represent the luxury segment or unique properties with specific characteristics that command premium prices.

b) **Actionable Recommendations for Real Estate Investors**

- Prioritize Room Count over Raw Area: Given the stronger correlation of price with the number of bedrooms and bathrooms than with total area, investors should pay close attention to the room configuration of a property. A property with an optimal number of bedrooms and bathrooms for the target market might offer better value or demand than one with a large area but an inefficient layout or fewer rooms.
- Utilize Price per Square Foot for Relative Valuation: When comparing properties, especially those of different sizes, use the 'Price_per_sqft' metric as a key indicator of value. A higher 'Price_per_sqft' might suggest a more desirable location, better quality construction, or premium features. Conversely, a lower 'Price_per_sqft' could indicate a potential undervaluation or a less prime location.
- Develop City-Specific Investment Strategies: Recognize that the Pakistani real estate market is not monolithic. Prices and demand vary significantly by city. Investors should conduct thorough research on the specific cities or even neighborhoods they are interested in. Analyze local market trends, economic factors, and future development plans that could impact property values in those areas.
- Align Investment with Property Type Market Dynamics: Understand the typical price points, target demographics, and investment potential for different property types (Houses, Flats, Portions, etc.) in your chosen location. Your investment strategy should align with the market characteristics of the property type you are considering.

- Investigate High-Value Properties (Outliers): While the analysis focused on general trends, the outliers in price and area represent a specific segment of the market. Investors interested in the luxury market or unique investment opportunities should investigate these properties further to understand the specific features or locations that contribute to their high value.
- Beyond Basic Features: Location is Paramount: Although detailed location-based features were not numerically engineered in this phase, the city-level analysis strongly emphasizes the importance of location. Future, more granular analysis should consider factors like neighborhood reputation, proximity to schools, hospitals, transportation hubs, and commercial areas, as these significantly influence desirability and price. Investors should evaluate properties not just on their intrinsic characteristics but also on their surrounding environment and accessibility.
- Seek Additional Data for Deeper Analysis: To build more robust predictive models and gain a competitive edge, consider incorporating additional data points beyond the basic listing details. This could include information on the age of the property, specific amenities (e.g., security, parking, communal facilities), construction quality, and local market supply and demand data.

5. Conclusion & Next Steps

a) Summary of Learnings

This exploratory data analysis of Zameen.com property listings in Pakistan has provided valuable insights into the factors influencing property prices. The data cleaning and transformation steps were crucial in standardizing the 'Price' and 'Area' columns, making them suitable for quantitative analysis. Handling outliers through capping helped to mitigate their impact while retaining valuable data points representing the diversity of the market. Feature engineering, particularly the creation of 'Price_per_sqft' and 'Total_Rooms', provided more insightful metrics for understanding property value.

b) Key findings from the analysis include: The distribution of property prices and areas is highly skewed, indicating a concentration of properties at lower values and sizes, with a long tail of luxury or large properties.The number of bedrooms and bathrooms, as well as the total number of rooms, show a stronger positive correlation with property price compared to the total area in square feet. This suggests that the functional layout and number of rooms are significant drivers of value.The 'Price_per_sqft' metric is a strong indicator of overall price and is useful for comparing the relative value of properties regardless of their size.Property prices vary significantly across different cities and property types, emphasizing the importance of location and property characteristics in determining market value.

c) Future Work Building upon this foundational analysis, the following steps are recommended for further research and model development:

d) Advanced Feature Engineering:

Location-Based Features: Extract more granular features from the 'Location' column. This could involve using geocoding to obtain latitude and longitude, calculating distances to key amenities (e.g., schools, hospitals, commercial centers, transportation hubs), or categorizing locations based on neighborhood reputation or development status.

Text Analysis of Description: Apply natural language processing (NLP) techniques to the 'Description' column to extract features related to property quality, amenities mentioned, or marketing language that might correlate with price.

Time-Based Features: If listing dates were available, analyze price trends over time or incorporate features related to how long a property has been on the market.

e) **Predictive Modeling:**Develop regression models (e.g., Linear Regression, Ridge, Lasso, Decision Trees, Random Forests, Gradient Boosting) to predict property prices based on the engineered features.Evaluate model performance using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.Consider using cross-validation to ensure model robustness and generalizeability.

f) **Clustering Analysis:**Perform clustering on properties based on their features (e.g., price, area, number of rooms, location) to identify distinct market segments. This could help in understanding different property clusters and their characteristics.

g) **Gathering Additional Data:**If possible, incorporate external data sources such as local economic indicators, demographic data, crime rates, or information on nearby infrastructure projects, as these factors can significantly influence property values.

h) **Summary:**

- Data Analysis Key Findings
- The initial dataset contained 18255 rows and 59 columns, which was reduced to 17974 rows and 9 relevant columns after cleaning and duplicate removal.
- Missing values were handled using mode, ffill, median, or placeholder values depending on the column type.
- 'Price' and 'Area' columns were successfully transformed into consistent numerical formats (PKR and sqft, respectively).
- Outliers in numerical features were addressed using the IQR capping method to retain data while mitigating extreme values.
- New features like 'Price_Numeric', 'Price_per_sqft', and 'Total_Rooms' were engineered to enhance analysis.
- Univariate analysis revealed that 'Price_Numeric' and 'Area' distributions were heavily right-skewed. Bivariate analysis showed that 'Bedrooms' (0.36), 'Bathrooms' (0.40), and 'Total_Rooms' (0.40) had a moderate positive correlation with 'Price_Numeric', while 'Area' had a very weak correlation (-0.007).
- 'Price_per_sqft' showed a strong positive correlation with 'Price_Numeric' (0.75).
- Significant price variations were observed across different cities and property types.
- Insights or Next Steps
- Future analysis should focus on advanced feature engineering, particularly extracting granular location-based features and applying NLP to property descriptions, as location and specific amenities are likely strong price drivers.
- Develop predictive models using various regression techniques to forecast property prices based on the engineered features, and potentially perform clustering analysis to identify distinct market segments.