Final Project Report

Analyzing the factors
influencing housing prices in India

Master of Science
In
INFORMATION SYSTEMS

By
Pratik Mane
Keerthi Jayram
Shashank Mysore
Yash Jivani
Saurabh Khatri

Under the guidance of
Prof: Jonathan Lee

Class: Data Science I

Introduction

The main objective of this research initiative is to explore the several factors that influence housing decisions in India. We aim to analyze complex datasets using advanced techniques and uncover the subtle aspects of Indian residents' needs and preferences. By doing so, we hope to gain a comprehensive understanding of the numerous factors that impact housing choices.

Who can benefit?

Understanding the complex needs and preferences of Indian residents can lead to developing more customized and responsive solutions. This can benefit architects, developers, and housing agencies by enabling them to design and provide housing options that better align with the diverse requirements of the population.

Why data analysis?

Data analysis in the real estate sector enables the recognition of patterns and trends, offering insights into sought-after localities, the correlation between property prices and specific features (such as the number of bedrooms or furnishing status), and the diverse factors influencing the overall market dynamics. Armed with data analysis, buyers, sellers, and real estate professionals are empowered to make more informed decisions.

How to improve decision-making?

- Informed Policy Formulation: The analysis will provide valuable insights for policymakers, enabling them to create housing policies that better align with the actual needs of the population.
- Optimized Urban Planning: Data-driven insights from the analysis will contribute to more effective urban planning by considering factors that influence housing choices.
- Tailored Housing Solutions: Developers and housing agencies will gain a deeper understanding of specific preferences, allowing them to offer more customized and suitable housing solutions.

Data

The given data represents information about different real estate properties, including details such as area, number of bedrooms (BHK), number of bathrooms, furnishing status, locality, parking availability, price, status (ready to move or almost ready), transaction type (new property or resale), property type (apartment or builder floor), and price per square foot.

Source

The dataset is sourced from Kaggle. This dataset contains information about real estate properties, including details such as area, number of bedrooms (BHK), number of bathrooms, furnishing status, locality, parking availability, price, and other relevant features. It covers various areas in Delhi like Rohini and Lajpat Nagar and includes diverse types of properties like apartments and builder floors. The structure of the dataset indicates that it can be utilized for tasks such as predicting property prices, analyzing market trends, or obtaining valuable insights into the real estate scenario in these specific regions.
URL : https://www.kaggle.com/datasets/aemyjutt/indianhousesdataanalysis

Variables

In the context of our analysis, the dependent variable (DV) or the explained variable is "Price." This is the variable we aim to understand, predict, or explain.
On the other hand, the independent variables (IV) or explaining variables include "Type," "Transaction," "Status," "Parking," "Locality," "Furnishing," "Bathroom," "BHK" (number of bedrooms), and "Area." These independent variables are factors that we believe may influence or contribute to variations in the dependent variable, Price.

Data Cleaning

In the initial phase of data cleaning, a meticulous assessment was conducted to identify and address any missing values present in our dataset. The examination revealed that numerous rows across various columns were incomplete. Moreover, some columns exhibited a mix of string and integer data types. To ensure the highest level of data cleanliness, tailored strategies were employed based on the nature of these data types.

Specifically, columns pertaining to bathrooms, parking availability, and square footage were found to have missing values in the form of integers. To remedy this, a statistically sound approach was applied, involving the imputation of these missing values by replacing them with the mean value derived from each respective column. This method is particularly effective in maintaining the integrity of the dataset while addressing missing numerical values.

However, columns related to furnishing status and property type consisted of string data types, with instances of missing values. In such cases, an approach that ensures effective resolution while adhering to the most frequent occurrences (mode) was implemented. The missing values in these string columns were replaced using mode substitution, thereby preserving the overall distribution of categorical data.

As part of our cleansing process for quality assurance, it was important to check for the presence of any duplicate rows within the dataset. Detecting and addressing duplicate entries is crucial in ensuring the accuracy and reliability of the data.

Descriptive statistics:

In analyzing the statistics for the output column, specifically the house prices, we observe that the mean price is 21 million, while the median is 14 million. This suggests a right-skewed distribution as the mean is higher than the median. The presence of houses with prices exceeding 100 million indicates the existence of outliers in the dataset, specifically 19 such houses. Interestingly, most of these expensive houses are categorized as ready to move and fall under the classification of new properties.

Addressing the outliers in the price values becomes essential for improving the accuracy of the model. Outliers, which are data points significantly deviating from the norm, can distort statistical analyses and model predictions. Removing these outliers can enhance the robustness of the model results.

The mean value being higher than the median is a common characteristic in datasets with right-skewed distributions, as observed in house prices. In such cases, the median is often considered a more reliable measure of central tendency, as it is less sensitive to extreme values.

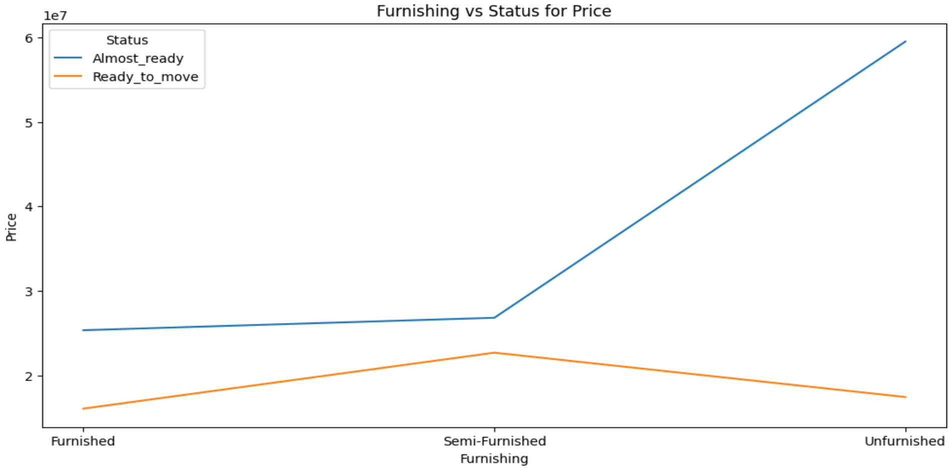
Surprisingly, houses with ten bedrooms do not appear in the results, suggesting that the number of bedrooms may not have a substantial impact on the price. This finding could influence feature selection in the model-building process, focusing on variables that exhibit a more pronounced influence on house prices.

Shifting the focus to the area of houses, the mean value is 1500, but the presence of a maximum value of 24,000 suggests a considerable spread in the data. A box plot distribution (Presented in Ipynb File) reveals that a few data points have exceptionally high values for the area, indicating potential outliers. These extreme values may need scrutiny, as they can skew the interpretation of the data and influence model performance.

Exploratory Analysis

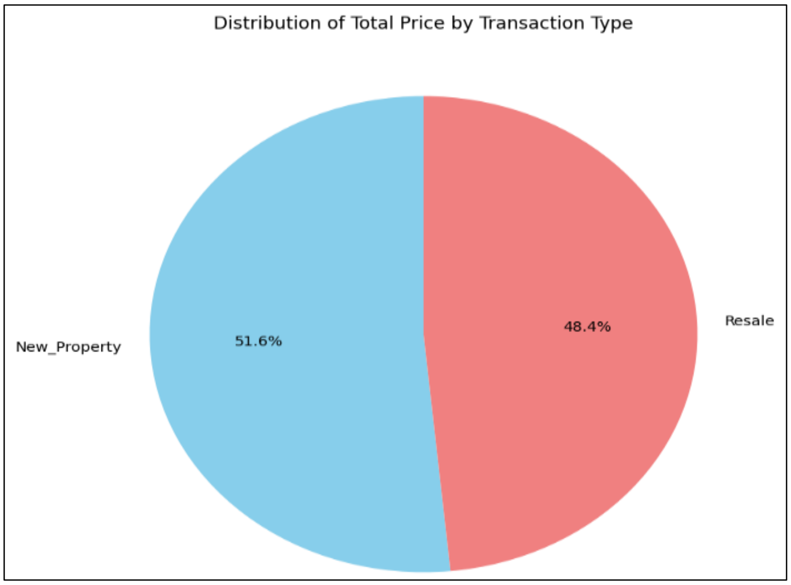
Q.1) How does Furnishing vs Status of the house affect the price of the house?

After looking at the following chart, there is a good relation between Furnishing and Status. The houses that are Almost Ready but unfurnished have the highest price as compared to the houses that are ready to move but are not furnished. Similar trends are seen for Semi Furnished and Furnished houses. Semi-furnished houses have a maximum price for Ready to Move houses followed by Unfurnished and Furnished. For Almost ready houses Unfurnished houses have the highest average prices followed by semi-furnished and then furnished. With the help of this data set, there are chances that as we are calculating the mean count of properties that were unfurnished and almost ready to use is more which can cause it to be highest among all.



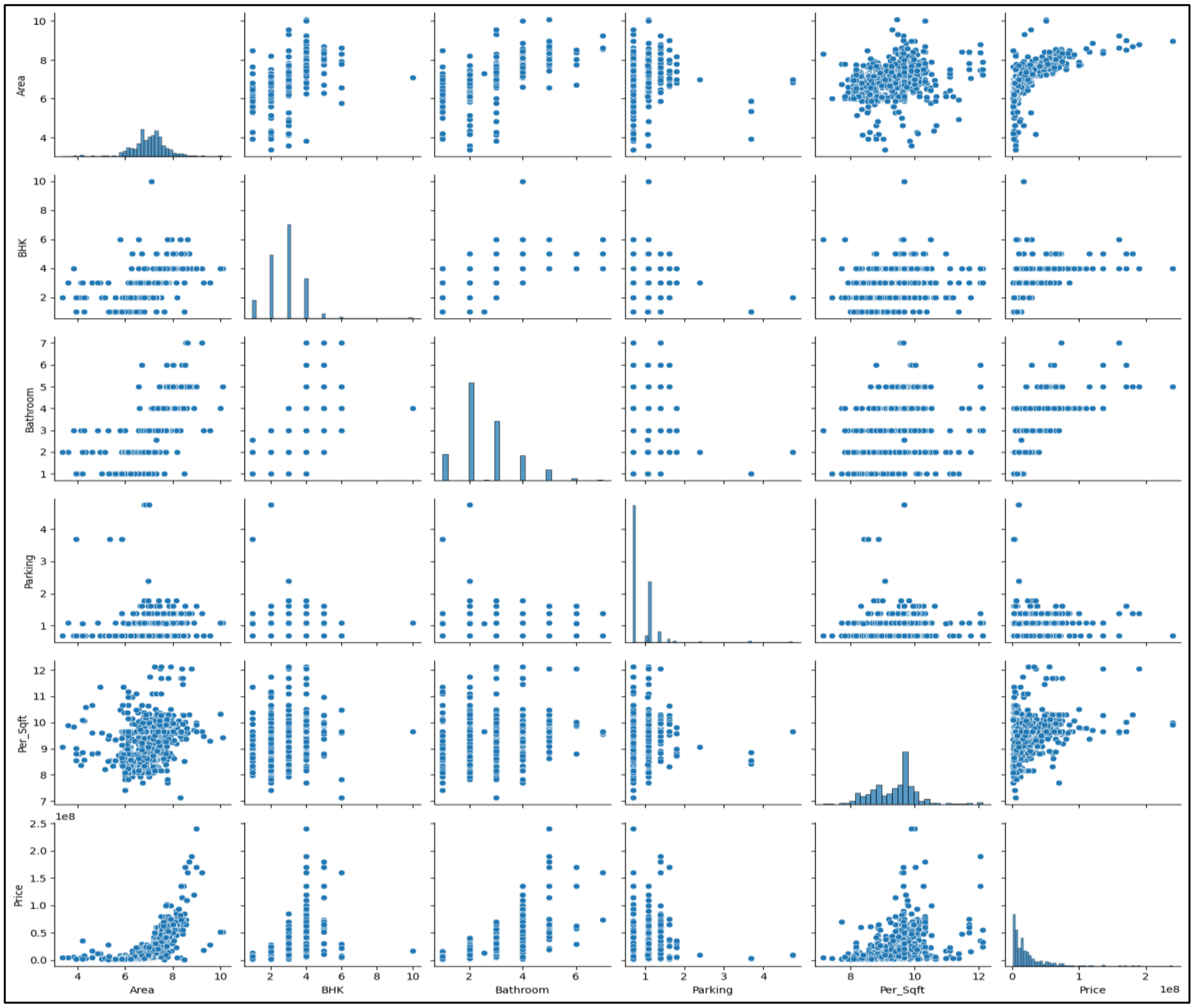
Q.2) What is the distribution of the total price by transaction type?

To find this we have plotted a Pie chart, transaction is divided into two categories New Property and Resale, our aim here is to find which one of them has a higher average price. The Pie chart makes it clearer that the new houses have a higher average price than the resale houses.



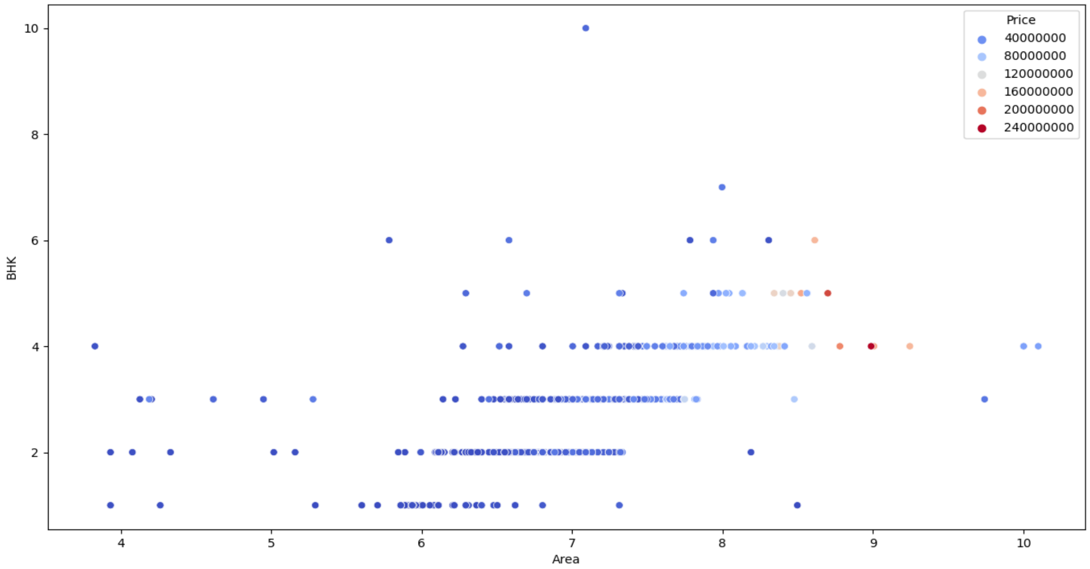
Q.3) Does the number of rooms affect the price of a house?

We are plotting a pair plot to answer this question along with multiple other questions. The pair plot displays the relationship between six variables: area, BHK, bathroom, parking, price per square foot, and total price. A distinct scatter plot is created for every pair of variables, with the distribution of each variable displayed on the diagonal plots. Looking at the graph, the BHK and Prices have a strong positive correlation. Along with BHK, Area, Bathroom, parking, and per square foot also have a strong positive correlation with prices.



Q.4) Does the BHK per area affect the price of the apartment, or are prices for BHK standard and not related to the area of BHK?

The analysis aimed to determine how the number of bedrooms (BHK) per unit area affects apartment prices by introducing a new metric, the "BHK Ratio," derived from dividing BHK by the area. Using a scatter plot, the study visually demonstrated a significant positive correlation (0.41) between the BHK Ratio and house prices. The subsequent plot affirmed the trend, indicating that larger BHK units, in terms of spatial occupancy, generally command higher prices. Outliers in the data were acknowledged, suggesting unique cases that may warrant further investigation due to their potential impact on statistical analyses.



Q.5) How does the Furnishing variable in the dataset is related to the Price of the house

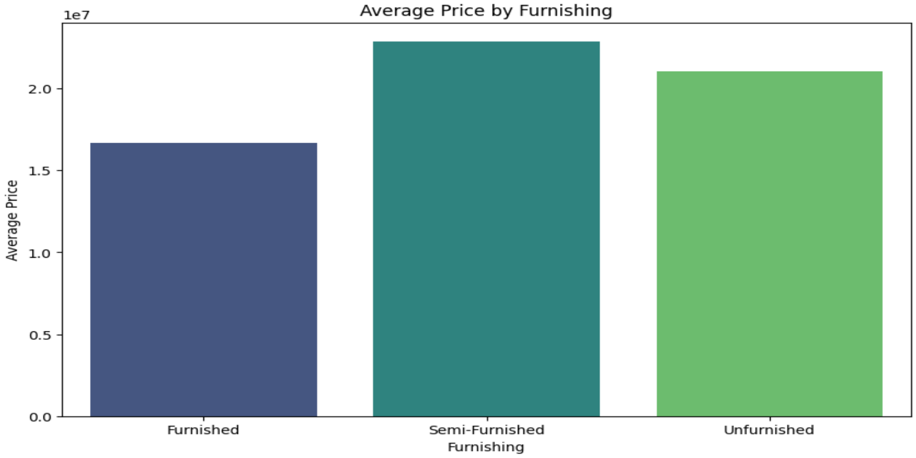
To answer this question, we had to find the number of distinct records in the Furnishing column. We found out that there are 3 distinct categories in the furnishing column which are "semi–furnished, unfurnished, and furnished". We mapped these and with the help of Get Dummies, we created 3 different columns for these categories. Then we again performed Correlation and Heatmap to find out the relation of these three categories with the Price of the house. Please refer to the heatmap shown below to see the relation between furnished, semi-furnished, and unfurnished. The heatmap shows Unfurnished houses are negatively related to the price of the house, and semi-furnished houses are positively related to the Price of the house though the relation is very weak. Almost ready houses have a strong relation with the Price of the house. The following bar chart shows the average price of a house per Furnishing. The graph shows the average price of semi-furnished is highest followed by unfurnished and furnished, the reason for this is the number of records present in the dataset for unfurnished.

Q.6) Are there any outliers in the dataset and what is the spread of the data?

According to the boxplot shown below it explains that the spread of data in Area, BHK, Per_Sqft and bathroom is good, although we can see some outliers in all

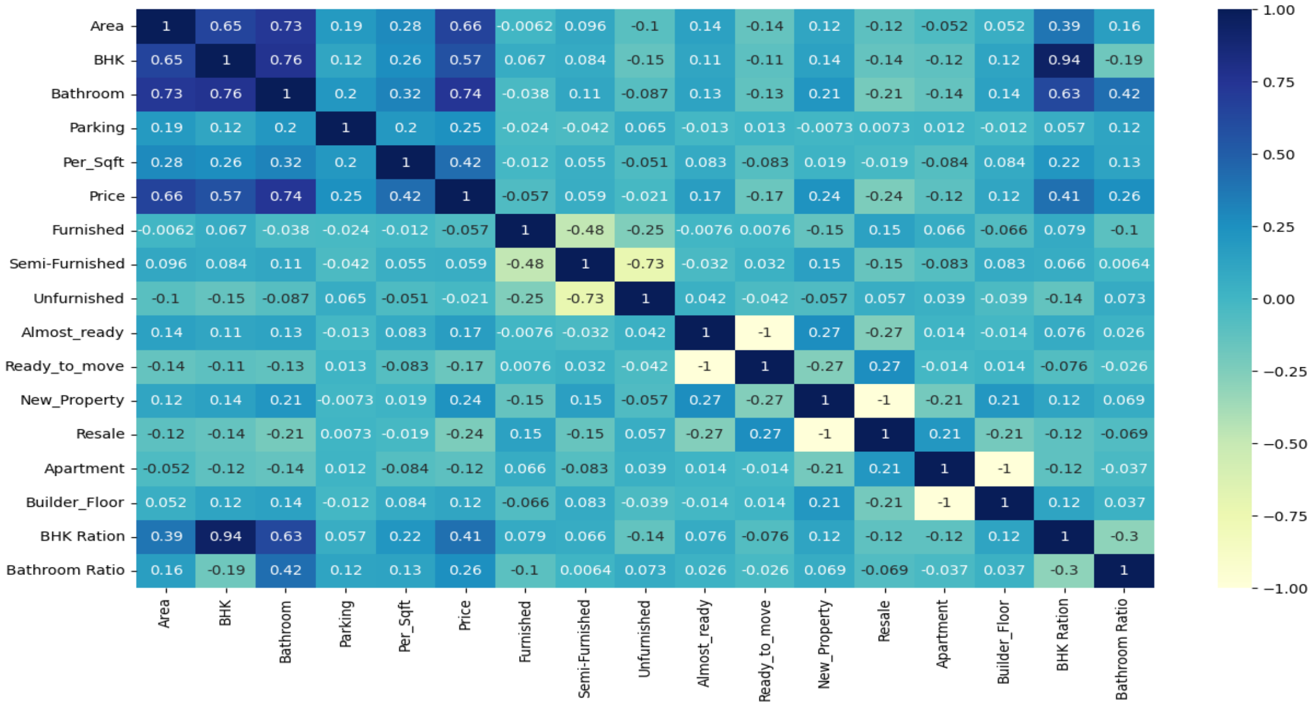
A variable has any relationship with the variable you are trying to predict in the next step.

While doing the analysis and trying to find the correlation, we got to know that Bathroom is more closely related to price than any other variable, so understand that we performed feature engineering on bathroom variable, we divided bathroom variable with number of bedrooms to find if the correlation of bathroom per bedroom is related to price or not as shown in the heat map we can see bathroom ratio is related to Price variable

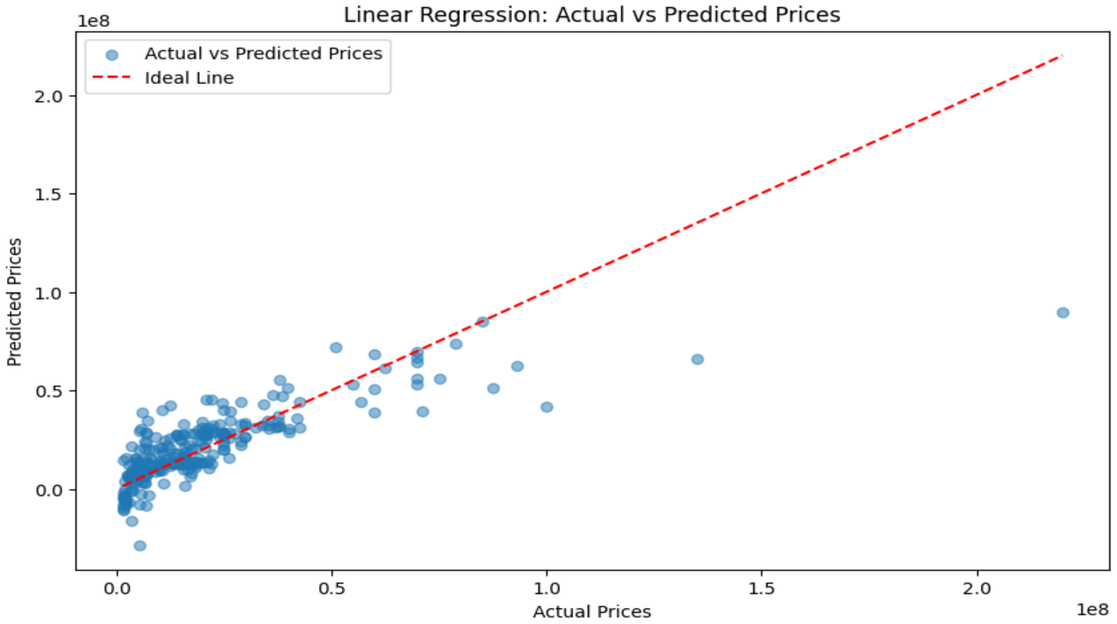


- Does a variable have any relationship with the variable you are trying to predict in the next step?

While doing the analysis and trying to find the correlation, we got to know that Bathroom is more closely related to price than any other variable, so understand that we performed feature engineering on bathroom variable, we divided bathroom variable with number of bedrooms to find if the correlation of bathroom per bedroom is related to price or not as shown in the heat map we can see bathroom ratio is related to Price variable



Results and Conclusion



The linear regression model applied to our dataset displayed an accuracy of 61% as indicated in the preceding slide. This level of accuracy, falling within the range of 60-70%, can be considered suboptimal for a machine-learning model. As a result, we decided to explore alternative approaches and employed random forest regressors.

Upon utilizing random forest regressors, we achieved a significantly improved accuracy rate of 88.65%. This outcome shows that this algorithm is better suited for our dataset than linear regression.

Furthermore, upon conducting further analysis of our data, we observed strong correlations between prices and certain variables such as the number of bathrooms (bathroom), number of bedrooms (BHK), area size, and price per square foot (per square feet). These findings suggest that these factors have a substantial influence on property prices within our dataset.