**[Bengaluru Home Price Prediction For Real Estate Purpose]**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**Nazeer Mastan Basha - AP21110011037**

**Bhanu Bharath - AP21110010995**

**Brahmendra Reddy - AP21110011015**

**Navadeep - AP21110011002**

**Sai Dheeraj - AP21110011244**

**A picture containing text

Description automatically generated**

Under the Guidance of

**Prof . Murali Krishna Yendluri**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**[April 15, 2024]**

# Table of Contents

[Table of Contents 2](#_1fob9te)

[1. Justification of the Project 3](#_3znysh7)

[2. Exploratory Data Analysis 4](#_2et92p0)

[1. Performed Data Overview : 4](#_wycl4bjf1g6)

[2. Performed Data Cleaning : 4](#_viyzl6e6z5cu)

[3. Visualization Enhancements : 4](#_ujy2xaa0p8rn)

[3. Data Pre-processing 6](#_tyjcwt)

[4. Feature Selection and Feature Generation 7](#_1t3h5sf)

[4.1 Feature Generation 7](#_6y19c1idnq2a)

[4.2 Feature Selection 8](#_w79stc920q27)

[5. Machine learning algorithm 10](#_4d34og8)

# Justification of the Project

**Problem**: Forecasting house prices in Bangalore is difficult owing to factors such as location, property type, amenities, and market trends. Traditional approaches may be subjective or time-consuming.

**Data Availability** : Bengaluru, being a large metropolitan city in India, provides an abundance of data on property listings, historical prices, property features, demographics, and economic factors. This data provides an ideal foundation for data science applications.

**Demand for smart Decision Making** : In Bengaluru, homebuyers, sellers, and real estate brokers need data-driven insights to help them make smart property transactions. Predictive models can be useful in this area.

**Complexity of Market Factors** : The real estate market in Bengaluru is influenced by various factors like location, amenities, size, infrastructure, and economic trends. Data science techniques can help in analyzing and understanding these complex relationships to predict home prices accurately.Predictive models can help in this situation.

**Data Science Suitability** : Data science provides a data-driven approach to developing models that analyze these characteristics and estimate housing values accurately and effectively.

# Exploratory Data Analysis

We have done a lot of Exploratory on the data of Bengaluru Houses. The dataset consists of 13320 tuples including the missing values with 9 attributes or columns. The columns are

-----------------------------------------------------------------------------------------------------

Index(['area\_type', 'availability', 'location', 'size', 'society',

'total\_sqft', 'bath', 'balcony', 'price'],

dtype='object')

-----------------------------------------------------------------------------------------------------

## **Performed Data Overview** :

Gathering basic information: Use df.info() to learn about data types, missing values, and the amount of entries for each feature.

To gain a sense of the data's content, use df.head() and df.tail() to display the first and final few rows, respectively.

Descriptive Statistics: To summarize numerical characteristics, call df.describe() (mean, median, standard deviation, minimum, maximum).

## Performed Data Cleaning :

Missing values identification and solved some with central tendencies and dropped out unwanted columns. Those columns are :

['area\_type','society','balcony','availability']

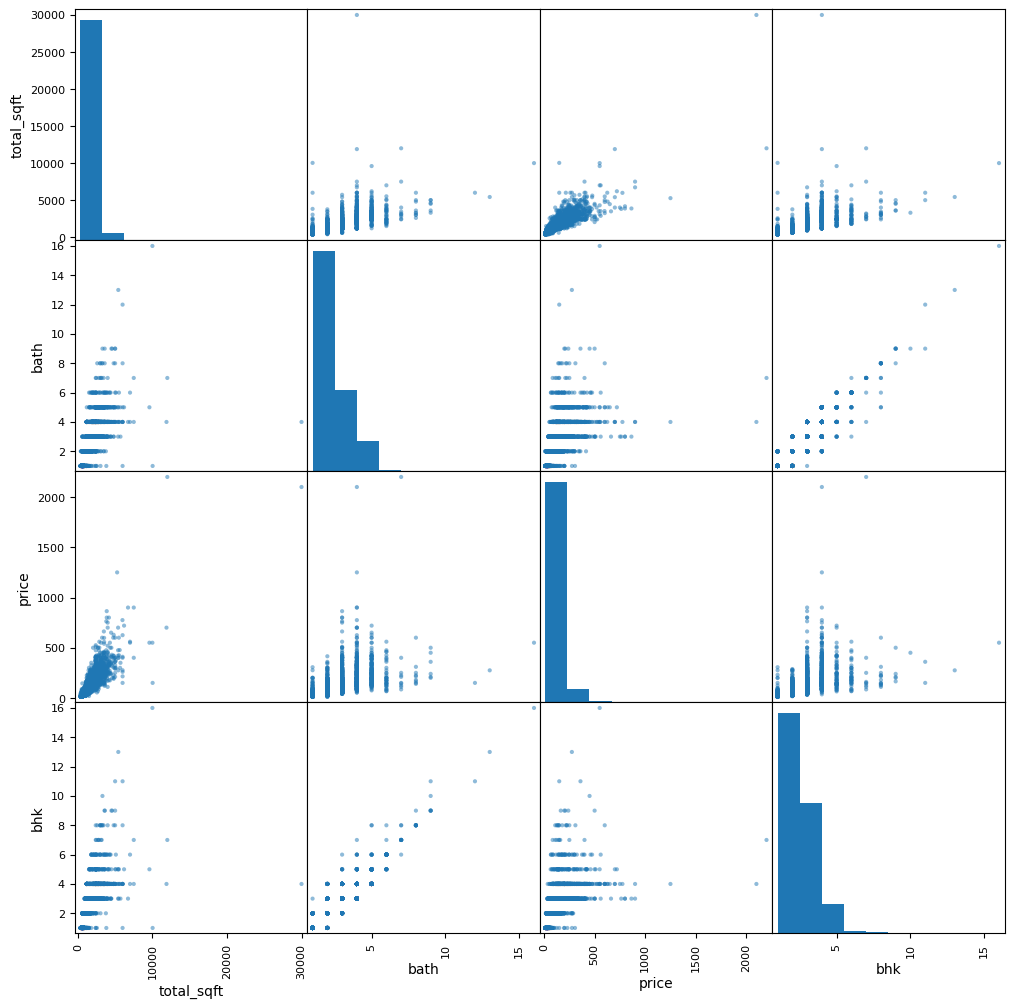
Analyzing outliers to determine if they are actual data points or mistakes. Depending on domain expertise, you may need to cap or winsorize extreme numbers.

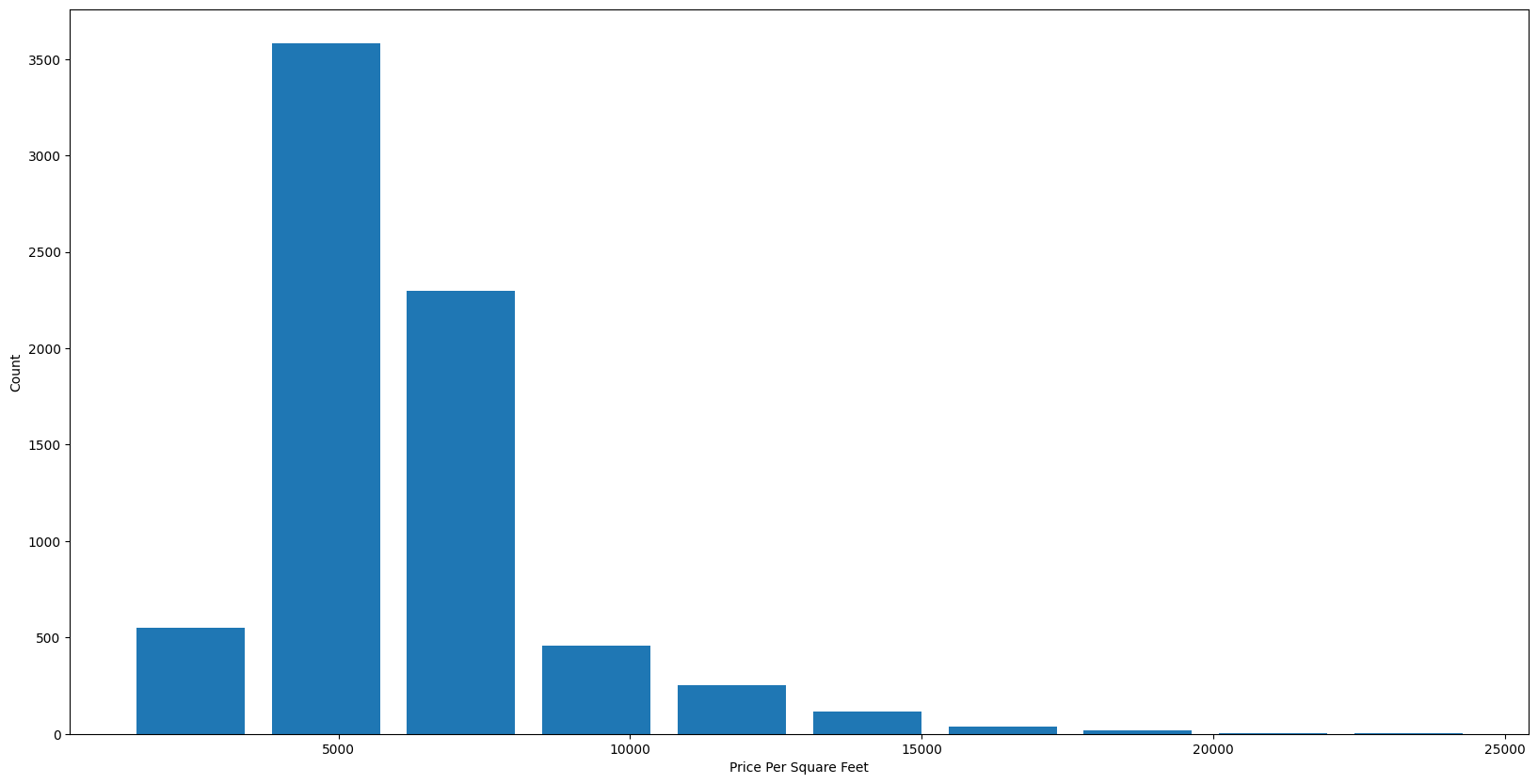
## Visualization Enhancements :

Scatterplots: To investigate potential links, create scatterplots of price and other numerical parameters (for example, area and total\_sqft).

Boxplots by Category: Create boxplots of price or other numerical data grouped by category features (e.g., location, BHK) to see how prices differ across categories.

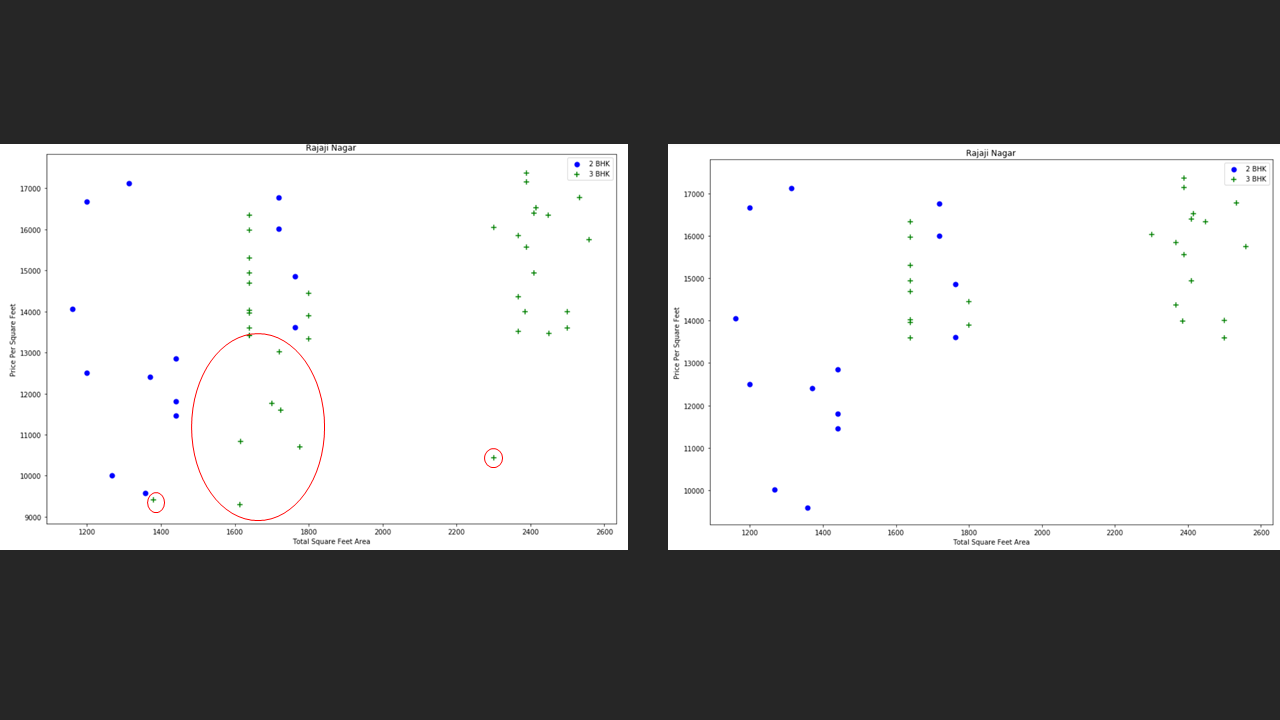
Heatmaps using seaborn: Consider utilizing heatmaps from seaborn to visualize relationships between various features in more detail.





# Data Pre-processing

Feature Scaling: Standardized or normalized numerical features with packages such as sklearn.preprocessing. StandardScaler or sklearn.preprocessing.MinMaxScaler. This guarantees that numerical features have a consistent scale, minimizing biases towards features with greater ranges.



Categorical Encoding: With packages like sklearn.preprocessing, you may encode categorical features using approaches such as one-hot encoding or label encoding.OneHotEncoder or sklearn.preprocessing.LabelEncoder. This sets up category characteristics for the machine learning algorithm.

Handle Missing Values : As said before we have performed column dropping and also done central tendency measures. Removed the unreliable data from the dataset which are less than the mean value. total\_sqft can be a range (e.g. 2100-2850). For such cases we can just take the average of min and max value in the range.

The following is cell output before pre-processing

location 1

size 16

total\_sqft 0

bath 73

price 0

dtype: int64

The following is cell output after pre-processing

location 0

size 0

total\_sqft 0

bath 0

price 0

dtype: int64

# Feature Selection and Feature Generation

Feature Selection and Feature Generation both are the techniques of feature engineering which is an important stage in the pipeline that includes converting raw data into features that are more suited to training a machine learning model.

It is simply the art and science of generating, choosing, and modifying data to improve model performance.

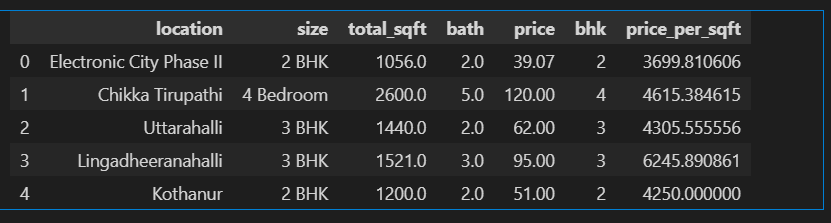
## 4.1 Feature Generation

We have added 2 new features which are most needed for the real estate broker. They are BHK(Bedrooms Hall Kitchen) and Price Per Square Feet.

Added up some values randomly and again removed the outliers from that.

For example total square feet of the house is 2100 Sq.ft and total no.of bedrooms are 43 which is not even possible.

Added the 2nd feature called Price Per Square Feet so that we can predict the price with the help of the total\_sqft and price\_per\_sqft

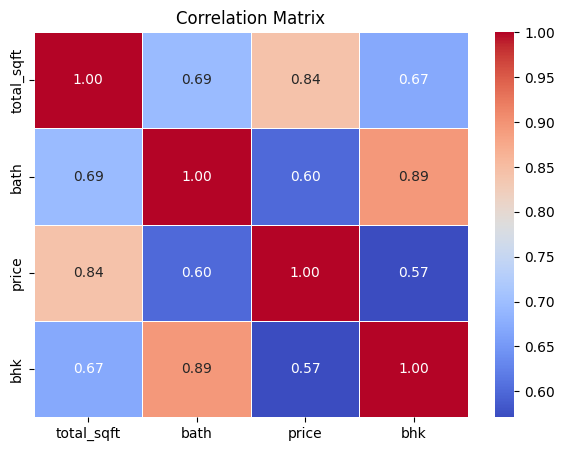


## 4.2 Feature Selection

Correlation analysis: Used techniques such as the Pearson correlation coefficient to identify factors that have a strong link with the target variable (price). Consider preserving only the most important ones.

Filtering methods: Used libraries such as sklearn.feature\_selection.SelectKBest or sklearn.feature\_selection.f\_classif will rate features based on statistical tests and preserve the top K features.

Wrapper methods: Investigate strategies such as recurrent feature elimination (RFE), which iteratively removes features and evaluates model performance before picking the best feature subset.



We have performed the Correlation analysis on the dataset. Using the heatmap of seaborn we visualized the correlation between the features. We almost noticed a very good kind of relationship among all the features

# Machine learning algorithm

As the project aim is prediction. We have used a regression algorithm. Linear Regression performed much well than other ML algorithms. The Accuracy score was 85+ so we assume it predicts well.

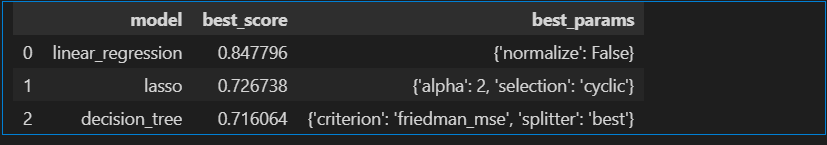
Advantages of Linear Regression for Prediction:

Interpretability: Linear regression models are reasonably straightforward to understand. The coefficients of the fitted line represent the strength and direction of the association between each independent variable and its dependent variable.

Simple: Linear regression is a well-known and computationally efficient approach. It's an excellent starting point for many prediction problems, particularly when the connection between variables is expected to be linear.

Baseline Performance: Even if a more complicated model is eventually adopted, linear regression can serve as a performance benchmark for comparison.

We have also compared its performance with other ML algorithms .



We used Grid Search CV for finding the best model.Also made some trails on the test data to find how our model is working.

