Predicting House Prices Using Machine Learning

A Comparative Analysis Algorithms and Features

Prepared by: Miral Naik

Index

- Problem Statement
- Dataset Overview
- Exploratory Data Analysis & Visualization
- Data Preprocessing & Feature Engineering
- Model Building
- Results & Insights

Problem Statement

- This is the data of a Real Estate company where they are trying to find Property Price evaluations as per the location and need to develop a predictive machine learning model.

➢ Overview of the Problem Statement

- We are creating a machine learning model to predict property prices using data provided by a real estate company. This involves analyzing information about the location and features of properties to estimate their value.
- We're developing this model to help the real estate company accurately evaluate property prices. By using machine learning, we can process large amounts of data and identify patterns that influence property prices. This will enable the company to make more informed decisions when buying, selling, or valuing properties.

Dataset Overview

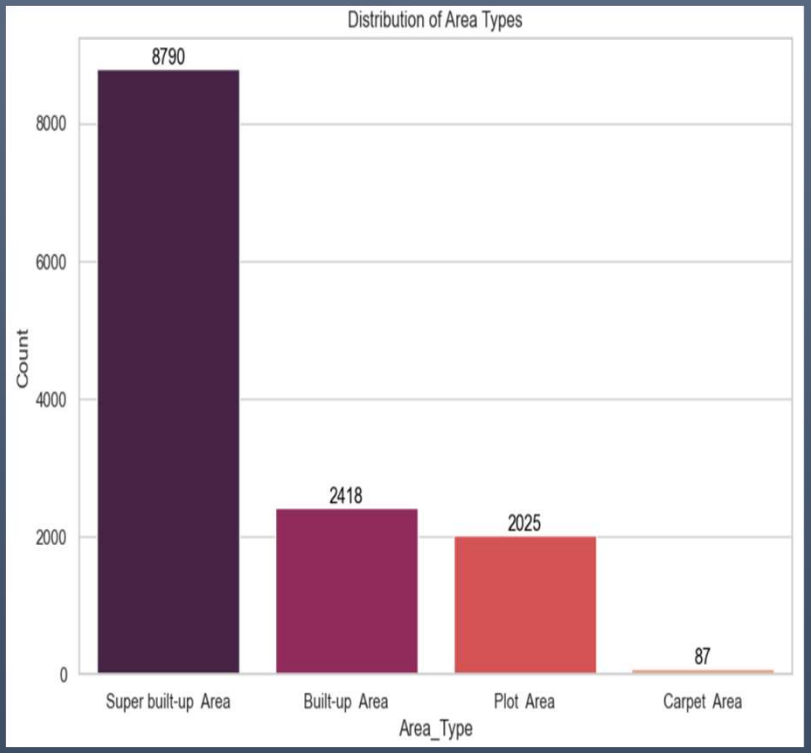
- There are 13320 records and 9 features in the dataset.
- The features include area type, availability, location, size, society, total sqft, bath, balcony, and price.
- The data types of the columns are integers (float64) and objects.
- Integers are used for numerical information such as bath and balcony.
- Objects are used for categorical information such as area type, availability, location, size, society, total sqft, and price.

Overview of the features

- Area Type: The Type of Area of Property
- Availability: Earliest time to move in the property, availability for possession.
- Location: Locality or Area in the city
- Size: Property Type (Like 3BHK, 4BHK)
- Society: The property in the society or not
- Total Sqft: Area of property
- Bath: No of Bathroom in that particular Property
- Balcony: No of Balcony
- Price: Price of the property (target Column)

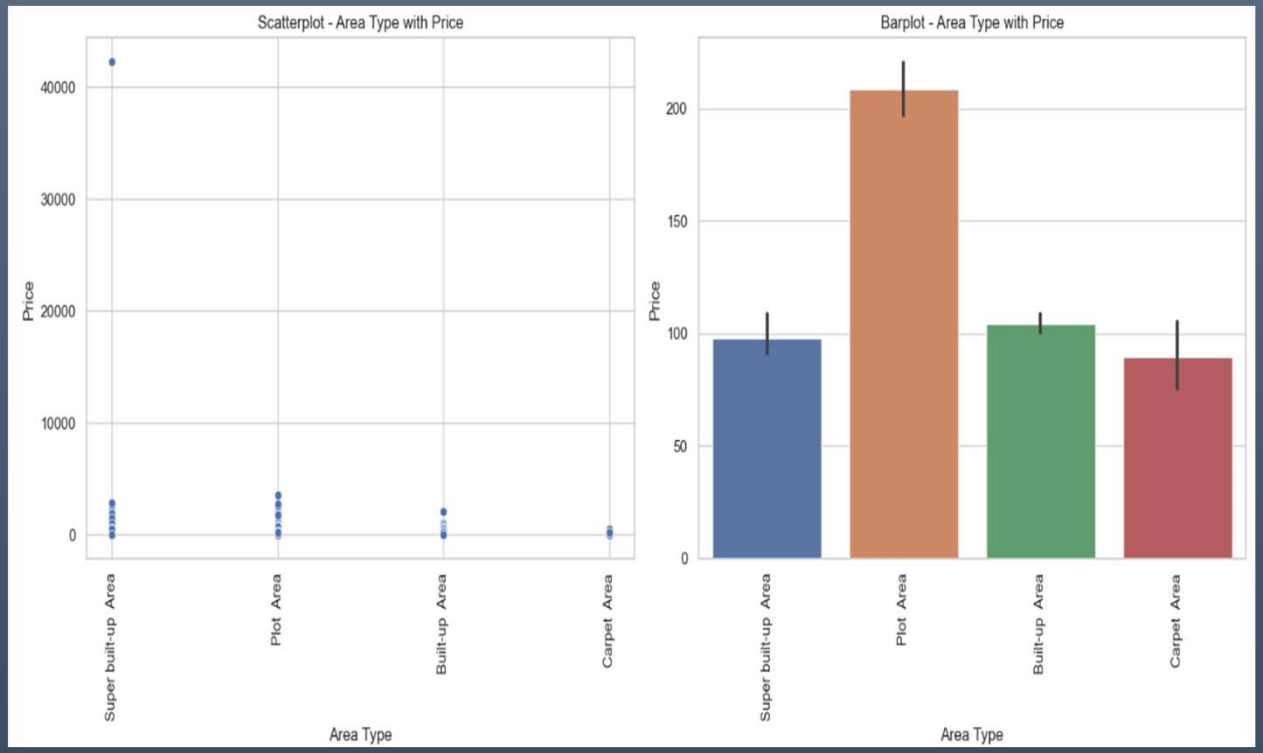
Exploratory Data Analysis & Visualization

Area Type



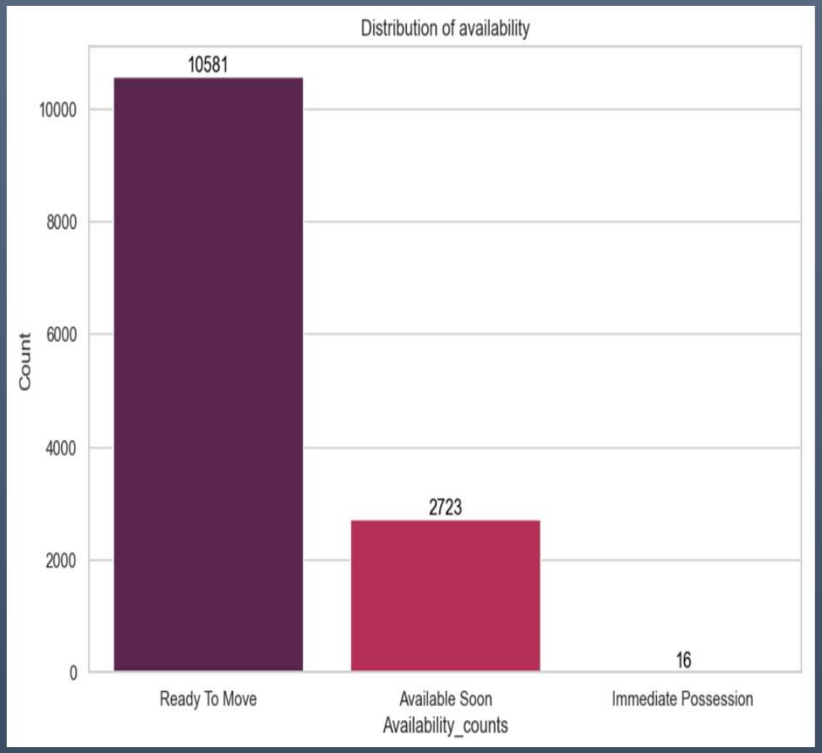
- "Super built-up area" is the most common
- "Built-up area" and "plot area" are less common but still visible.
- "Carpet area" is the rarest of all.

Area Type with Price



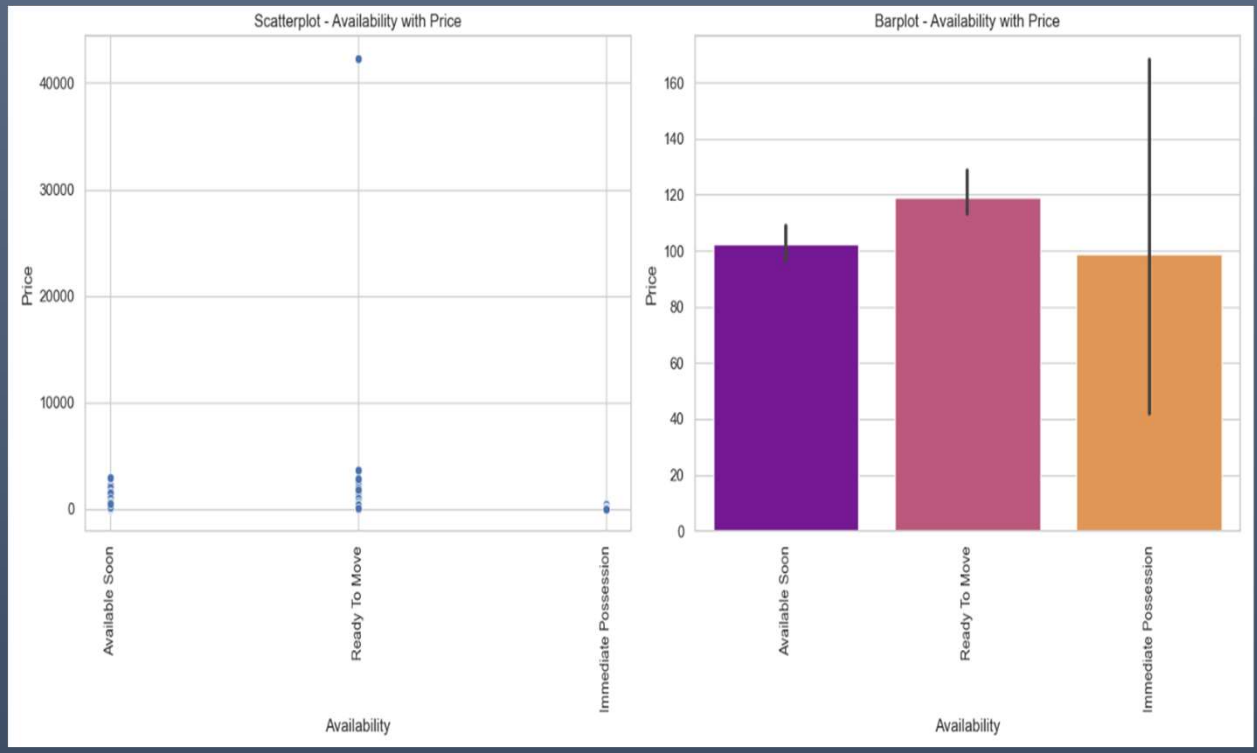
- The "area type" doesn't seem to have a significant impact on the price.
- Therefore, it might not provide useful information for predicting prices accurately in our model.
- So we can drop it from analysis.

Availability



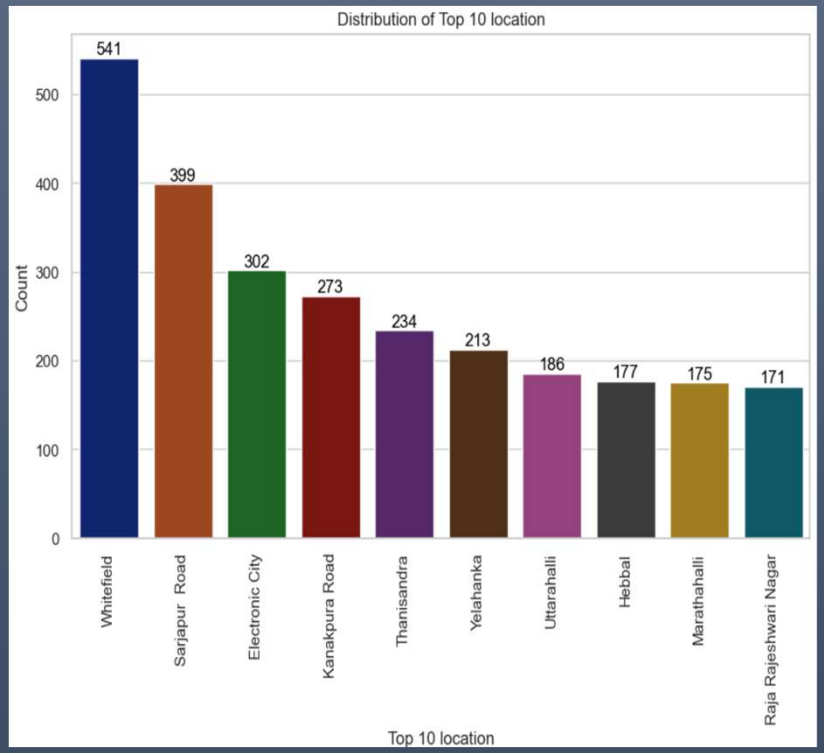
- The majority of properties are listed as "Ready to Move", followed by "Available Soon”.

Availability with Price



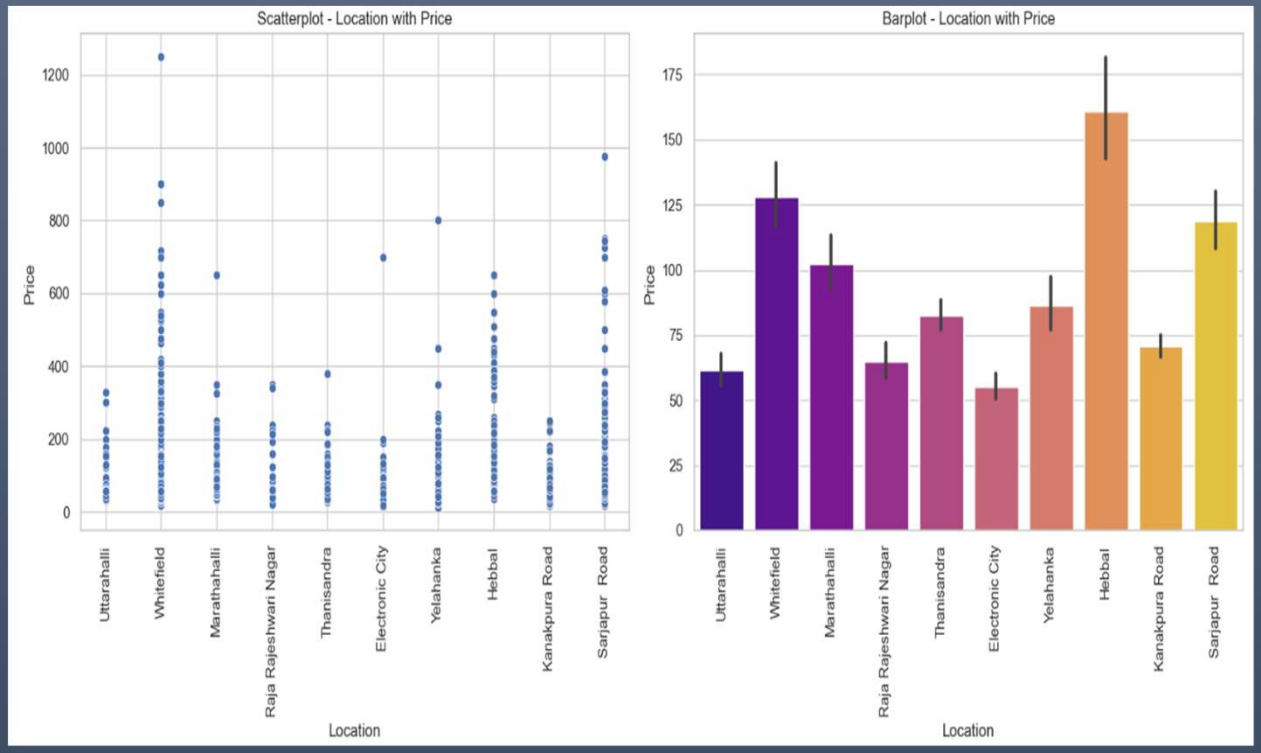
- The prices of properties categorized as "Available Soon" and "Ready to Move" seem to be very similar.
- This means that the availability of properties doesn't affect their prices differently.
- So, it might make sense to remove the availability from analysis.

Location



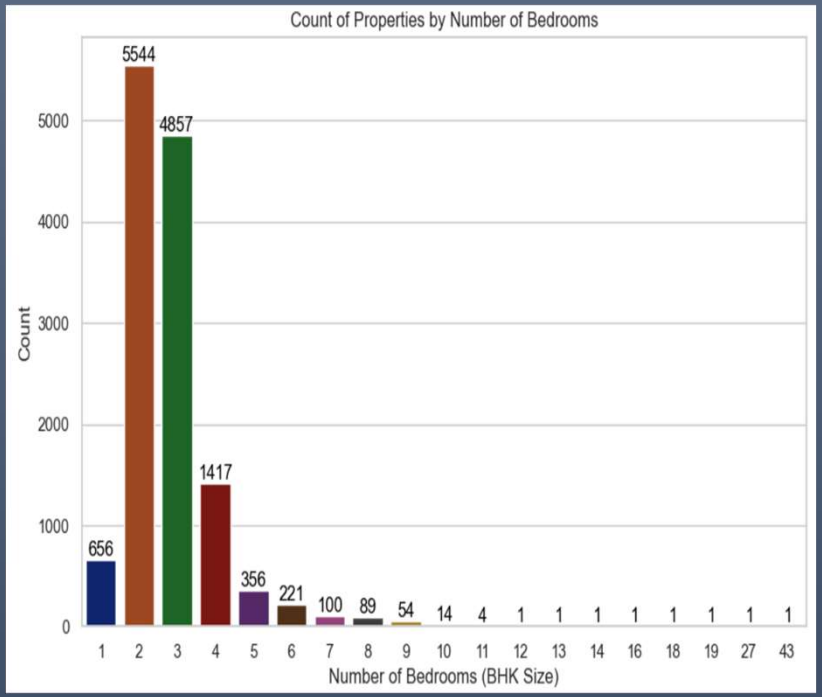
- Whitefield has the highest frequency of properties listed, followed by Sarjapur Road and Electronic City.
- There is a missing value in Location.

Location with Price



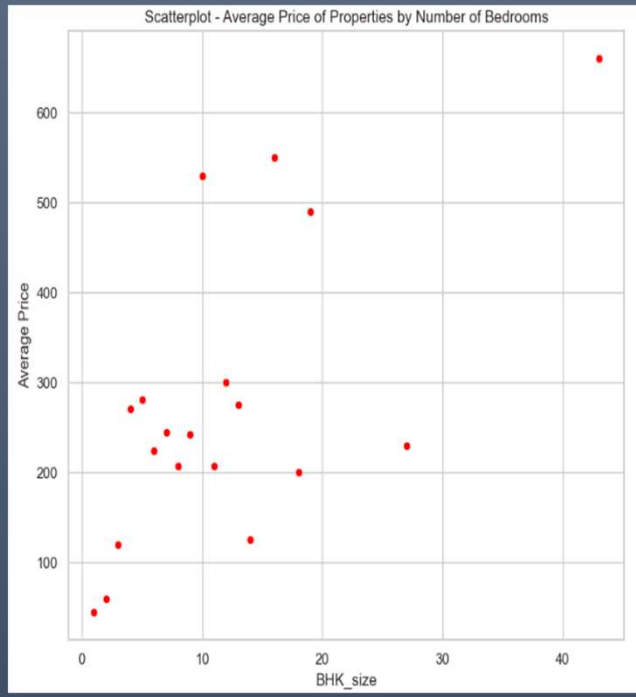
- Hebbal, Sarjapur Road and Whitefield are the top 3 locations that have the highest prices for houses.

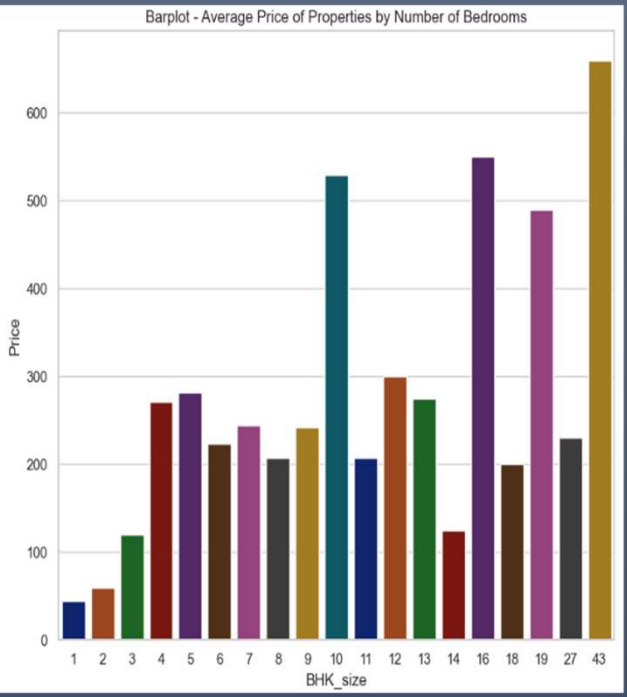
Size



- In Bengaluru, People prefer 2 BHK and 3 BHK houses the most
- It's unlikely for a house to have 27 or 43 bedrooms. So we need to check them.
- There are 16 missing values in Size.

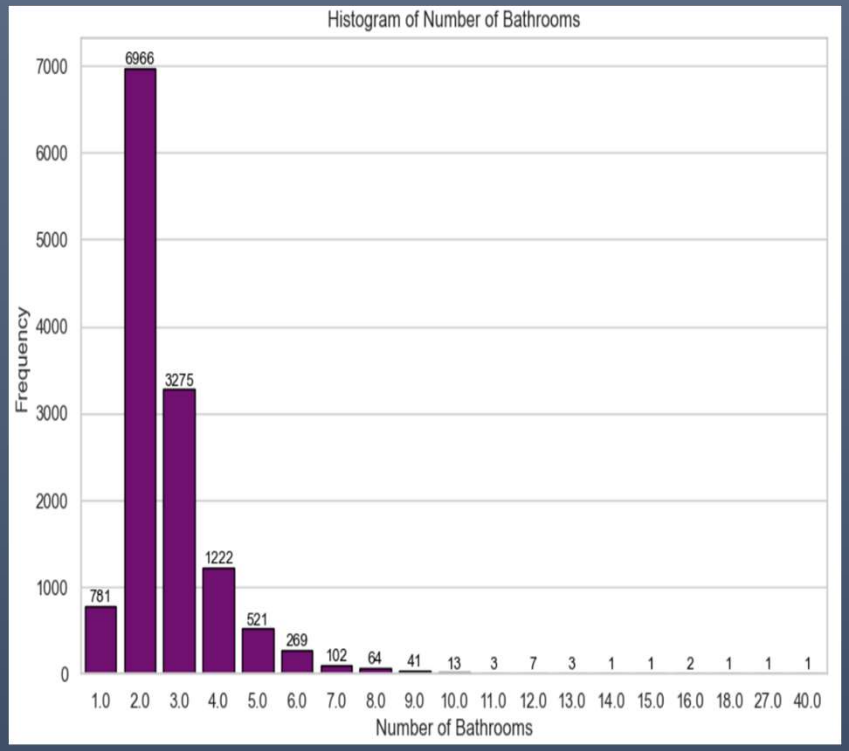
Size with Price





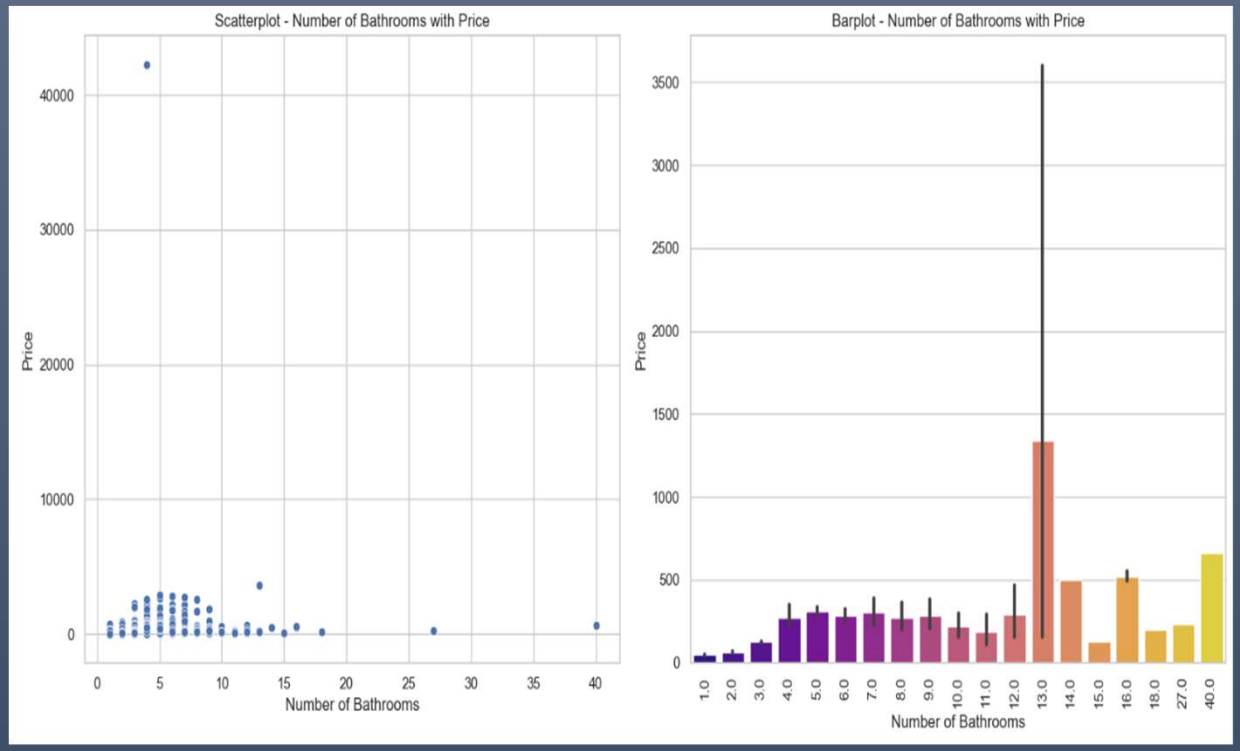
- Properties with more bedrooms (larger BHK sizes) tend to have higher prices than houses with fewer bedrooms.

Bath



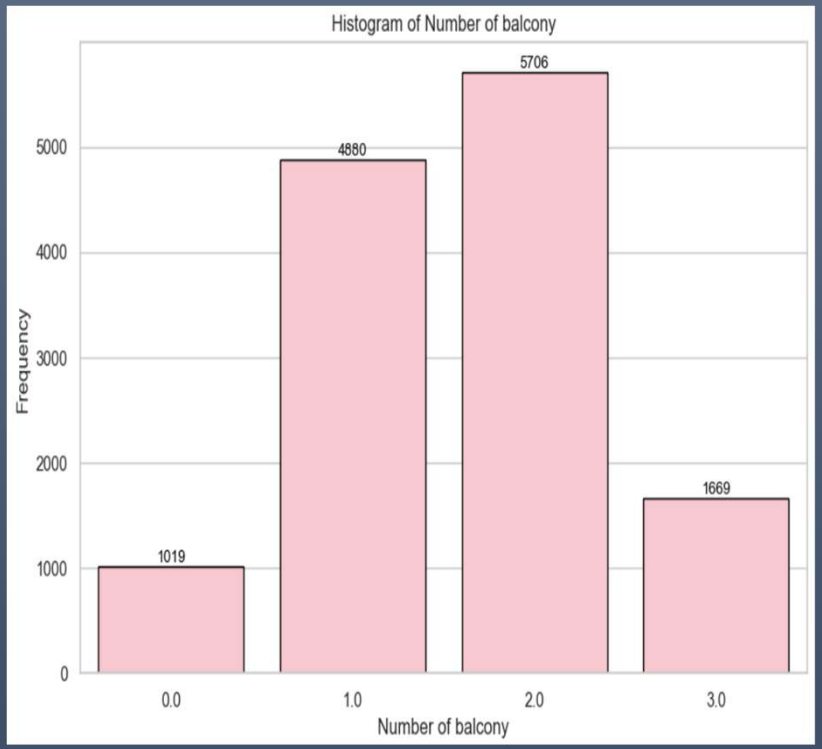
- The majority of properties have 2 bathrooms, followed by 3 bathrooms and then 4 bathrooms.
- There are more properties with fewer bathrooms compared to properties with more bathrooms.
- There are 73 missing values in Bath column.

Bath with Price



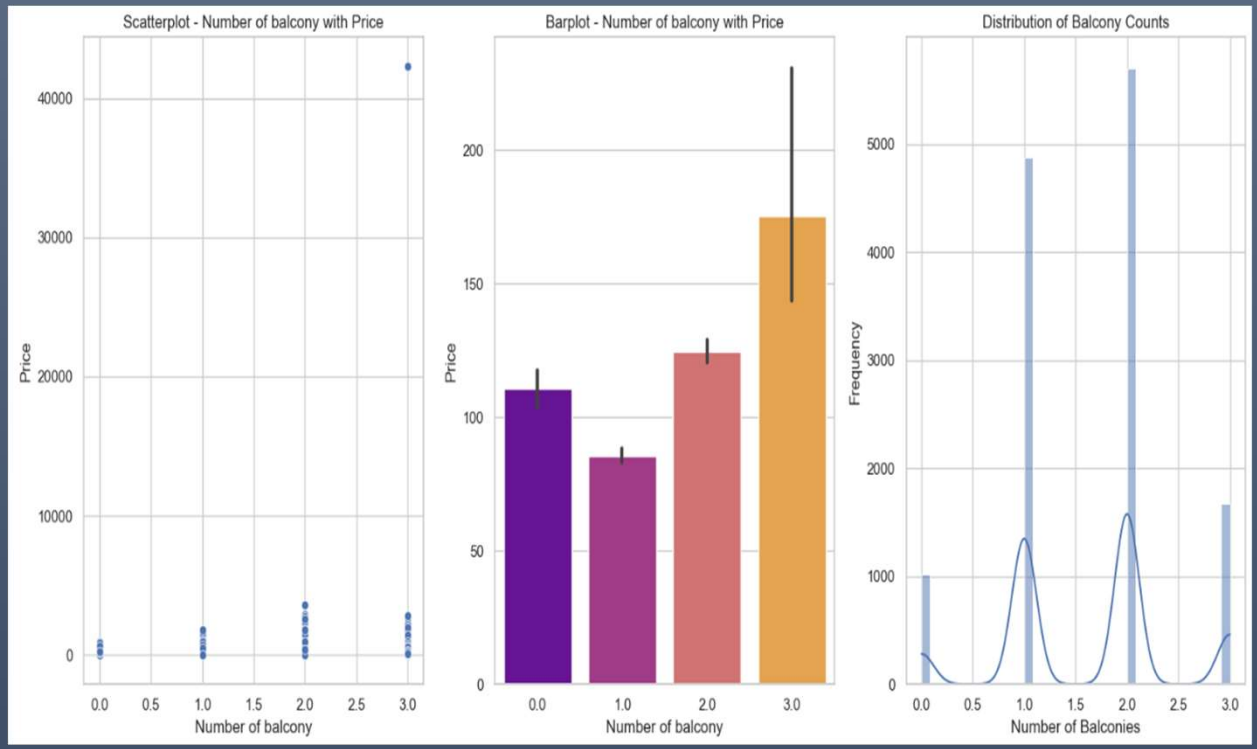
- Properties with a higher number of bathrooms tend to have higher prices.
- However, there are some outliers where properties with fewer bathrooms have higher prices.

Balcony



- The majority of properties have either 1 or 2 balconies, with a significant number of properties having 0 balconies as well.
- There are relatively fewer properties with 3 balconies.
- There are 609 missing values in Balcony.

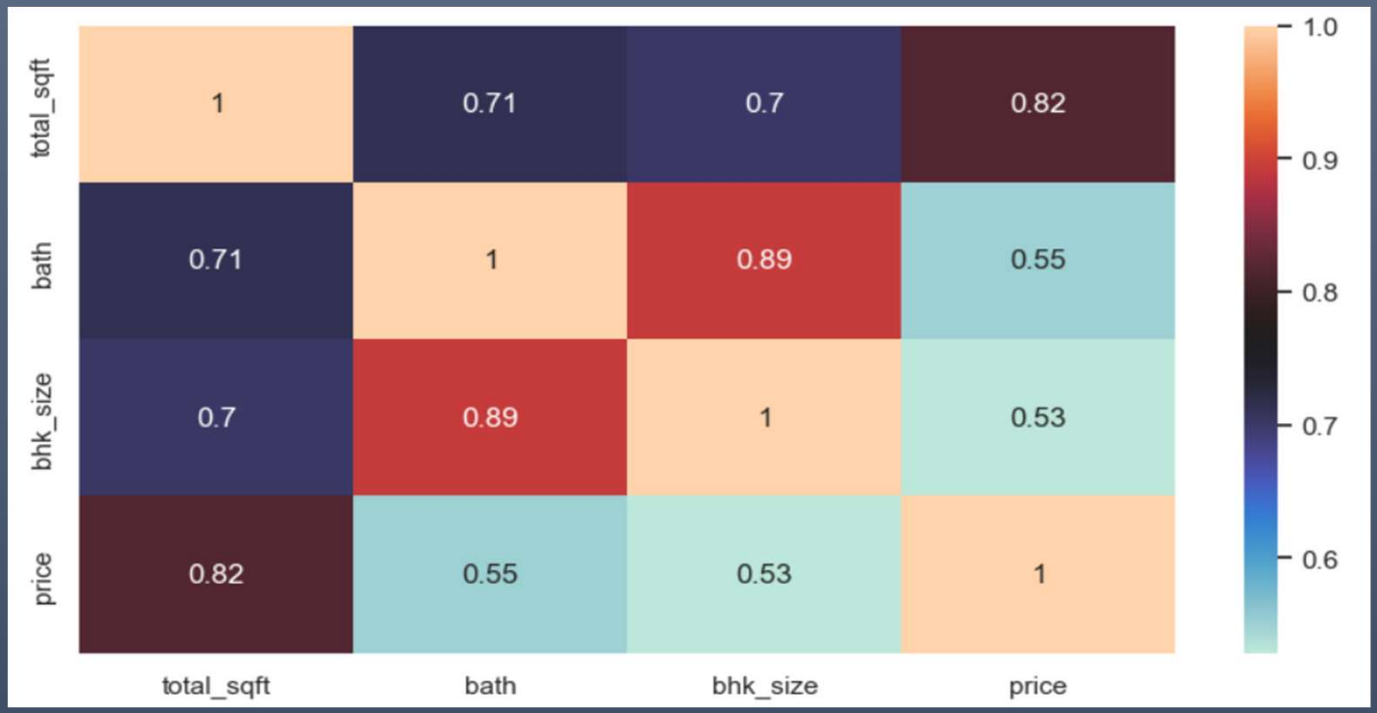
Balcony with Price



- The prices of properties with 2 and 3 balconies appear to be very similar.
- So the number of balconies may not have a significant impact on the price of properties.
- So, it might make sense to remove the Balcony from analysis.

Data Preprocessing
&
Feature Engineering

Correlation for Numerical data using Heatmap



- Price is strongly related to the total sqft (82% correlation), meaning as the size of the property increases, the price also tends to go up.
- The correlation with the number of bathrooms and bedrooms (55% and 53% respectively) is moderate, suggesting that properties with more bathrooms and bedrooms generally have higher prices, but not as strongly as with total sqft.

➢ Missing Value Treatment

- The dataset contains 13,320 rows with missing values in some columns, including 1 in 'Location', 16 in 'Size', 73 in 'Bath', 609 in 'Balcony', and 5,502 in 'Society' (about 41% of the data).
- So the Imputation method was performed using mode for categorical columns (Location, Size) and median for numerical columns (Bath, Balcony).
- The Society column was dropped due to its high percentage(41%) of missing values.

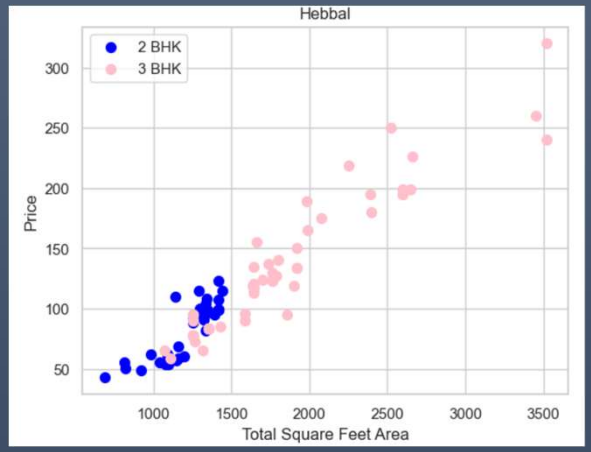
➢ Feature Engineering & Encoding Part

- I created an additional column like “price_per_sqft” to easily compare property prices based on their size. This helps identify outliers, understand market trends, and make informed decisions related to property investment or pricing strategies.
- Also, we know that “location” is an important factor in determining the price. Also, It is a categorical column. So we have to convert it into numerical format. So I used the One-Hot Encoding Technique.
- But One-Hot Encoding can lead to a high-dimensional dataset(because It has a lot of unique values), known as the “curse of dimensionality,” which may require additional processing or dimensionality reduction techniques. Here, the dimensionality reduction technique reduces the number of locations.
- One way to make the dataset smaller is to group locations that don't have many properties and label them as "others." We only keep the locations with a lot of properties. This reduces the number of categories significantly. we need to deal with, making our dataset simpler.

➢ Handling Outliers

- There are a huge amount of outliers present in 3 columns “Location”, “Size” and “Bath”. It will make an impact on model prediction so we have to handle these outliers.
- I have removed outliers from “Location” columns using Mean and Standard deviations based on the 'price_per_sqft' column, grouping by the 'location' column. It calculates the mean and standard deviation for each location group and filters out rows that fall outside one standard deviation from the mean price_per_sqft of that location.

- The same location a 2bhk house costs more than a 3bhk house with the same area.
- I have removed outliers from “Size” column based on the 'price_per_sqft' column, considering the 'location' and 'bhk_size' columns for grouping and calculating the and standard deviation . Outliers are identified using the mean of 'price_per_sqft' for the previous BHK size if certain conditions are met.



- I have removed outliers from “bath” columns based on the number of bathrooms compared to the number of bedrooms, removes excessive bathroom counts, and creates a new DataFrame without the 'price_per_sqft' column for further analysis.

Model Building

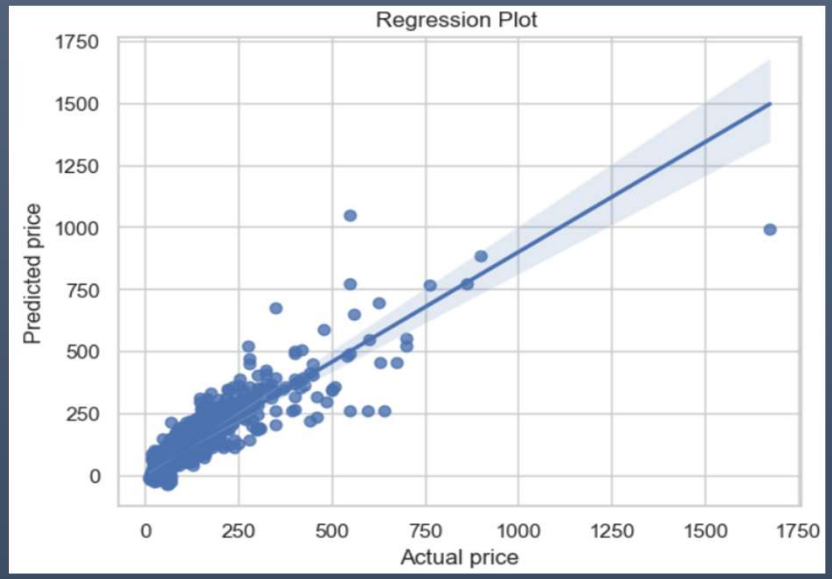
- I have Split the data into test and train in a 20:80 ratio
- For choosing the final model I have used the regression models such as Linear Regression Model, Lasso Regression Model, Ridge Regression Model, Decision Tree Regression Model and Random Forest Regression Model.
- I Checked their performance by different metrics like MSE, MAE, R-Squared, Adjusted R-Squared and Accuracy Score.
- Based on these metrics, the **Linear Regression model** appears to perform the best overall, as it has the lowest MSE, lowest MAE, highest R-squared, highest Adjusted R-squared, and highest Accuracy Score among the models provided.

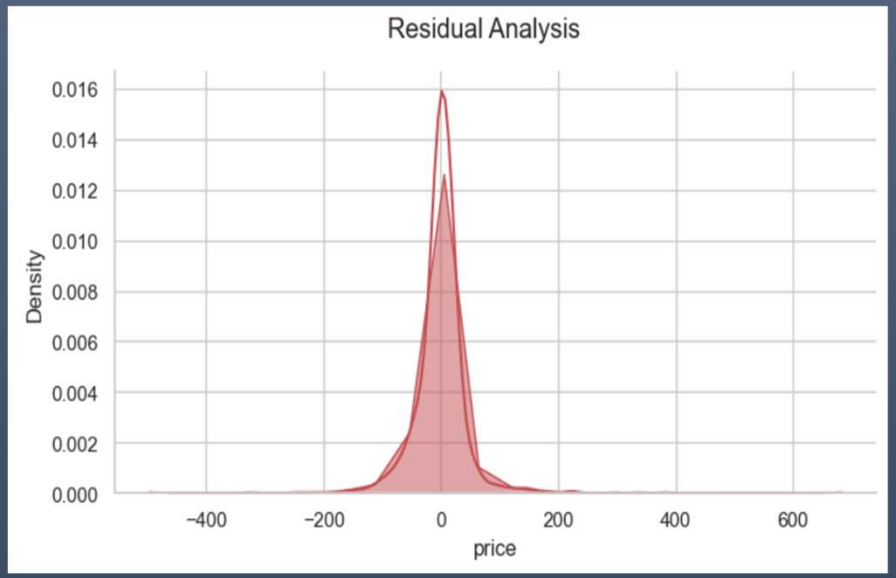
Model Name	MSE	MAE	R-Squared	Adjusted R-Squared	Accuracy Score
Linear Regression Model	2343.49	26.52	0.79	0.75	79.07
Lasso Regression Model	3718.94	33.8	0.67	0.6	66.79
Ridge Regression Model	2386.19	26.38	0.79	0.74	78.69
Decision Tree Regression Model	5407.55	29.74	0.52	0.42	51.71
Random Forest Regression Model	3704.41	26.5	0.67	0.6	66.92

Results & Insights

✓ Result

- Here, The Linear Regression Model is our final model.
- The accuracy and precision of the model's predictions, as demonstrated by the regression plot and residual analysis, suggest that the model performs well in predicting property prices.





- Here, points clustered closely around the diagonal line. So We can say that model is predicting accurately.
- Here we are able to find best fit line as well.

- The residuals in the residual analysis plot are normally distributed around zero, it suggests that the model's predictions are accurate and unbiased.

✓ Insights

- “Location”, “Total_Sqft”, “Bath” and “Size” are important features for predicting the prices of the houses.
- Hebbal, Sarjapur Road and Whitefield are the top 3 locations that have the highest prices for houses.
- The total square feet of a house can potentially increase its value.
- Houses with a higher number of bathrooms tend to have higher prices.
- Houses with more bedrooms (larger BHK sizes) tend to have higher prices than houses with fewer bedrooms.
- Predicting house prices has practical applications across various sectors, helping stakeholders to make informed decisions related to real estate investment, homeownership, mortgage lending, market analysis, urban planning, and risk management.

Thank You!