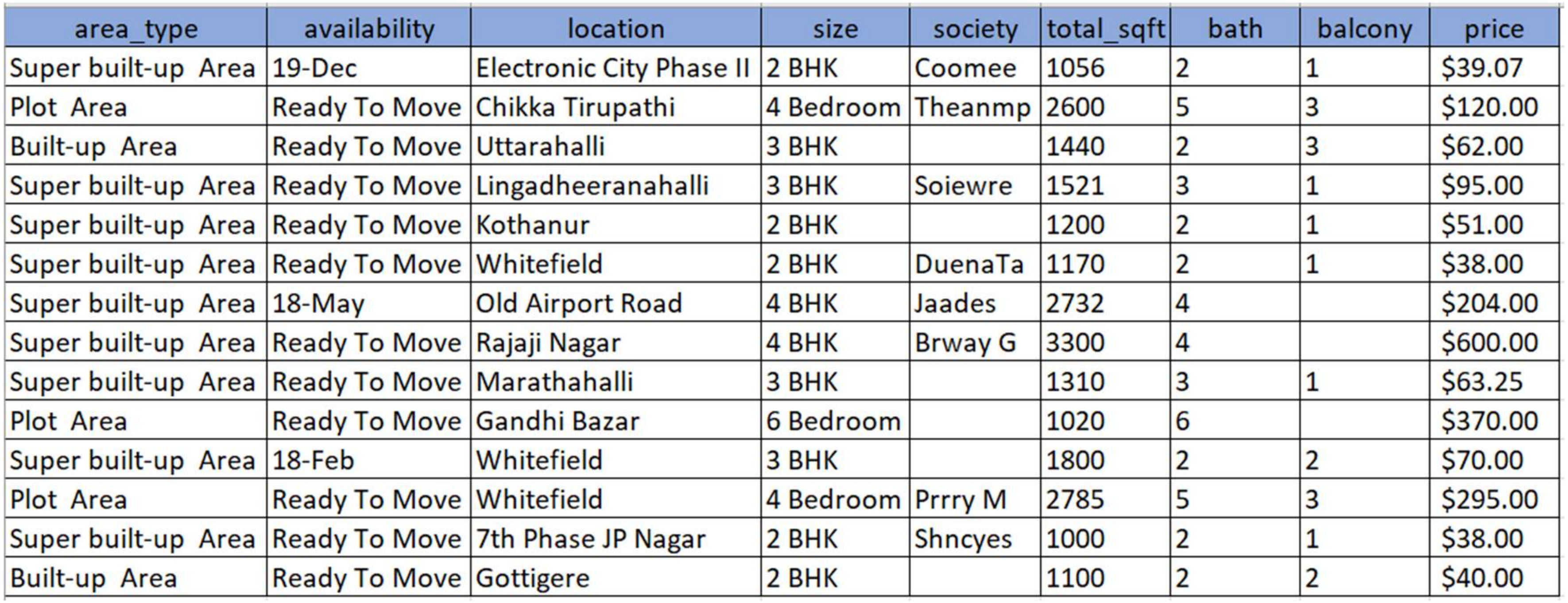
HOUSE PRICE PREDICTION

Made By: Priyanka Kolhe

OBJECTIVE

Develop a robust machine learning model for house price prediction that accurately estimates the market value of residential properties based on relevant features, with the aim of providing valuable insights for real estate professionals, homeowners, and potential buyers.

DATASET



area_type	availability	location	size	society	total_sqft	bath	balcony	price
Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2	1	$39.07
Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5	3	$120.00
Built-up Area	Ready To Move	Uttarahalli	3 BHK		1440	2	3	$62.00
Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3	1	$95.00
Super built-up Area	Ready To Move	Kothanur	2 BHK		1200	2	1	$51.00
Super built-up Area	Ready To Move	Whitefield	2 BHK	DuenaTa	1170	2	1	$38.00
Super built-up Area	18-May	Old Airport Road	4 BHK	Jaades	2732	4		$204.00
Super built-up Area	Ready To Move	Rajaji Nagar	4 BHK	Brway G	3300	4		$600.00
Super built-up Area	Ready To Move	Marathahalli	3 BHK		1310	3	1	$63.25
Plot Area	Ready To Move	Gandhi Bazar	6 Bedroom		1020	6		$370.00
Super built-up Area	18-Feb	Whitefield	3 BHK		1800	2	2	$70.00
Plot Area	Ready To Move	Whitefield	4 Bedroom	Prrry M	2785	5	3	$295.00
Super built-up Area	Ready To Move	7th Phase JP Nagar	2 BHK	Shncyes	1000	2	1	$38.00
Built-up Area	Ready To Move	Gottigere	2 BHK		1100	2	2	$40.00

DATA DICTIONARY

- **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**: Area of property
- **Bathroom Nos**: No of Bathroom in that particular Property
- **Balcony**: No of Balcony
- **Price**: Price of the property (target Column)

ABOUT DATA

The data consists of 13320 rows, 9 columns.

The data has 1 numerical variables:

- Price

The data has 8 categorical variables:

- area_type, availability, location, size, total_sqft, balcony, society, bath

It has 8 independent variables:

- area_type, availability, location, size, total_sqft, balcony, society, bath

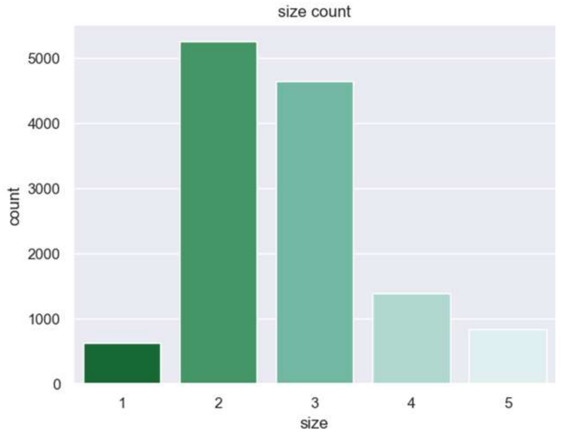
Dependent variable: price

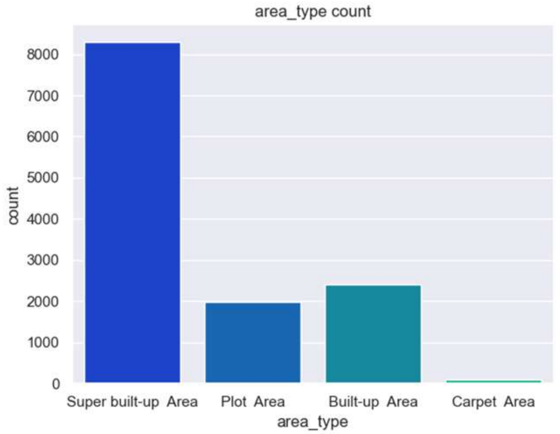
CHALLENGES

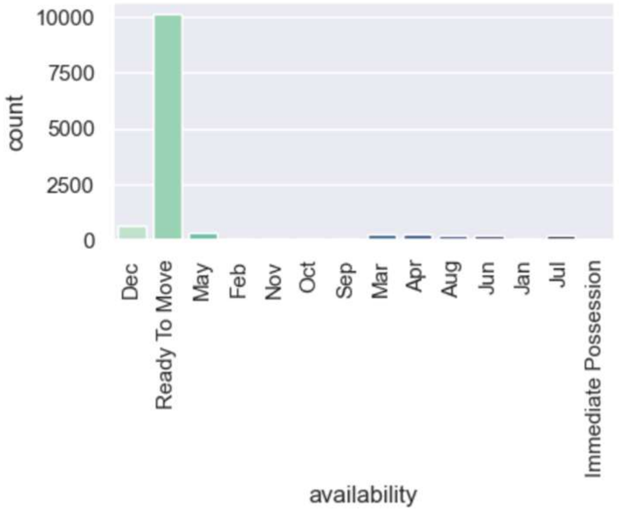
- While accessing the file through Jupyter Notebook an error occurred. There was an issue with decoding a byte sequence as UTF-8.
    - Solution: Opened the CSV file with Notepad and saved it with encoding UTF-8.
- Missing values were there to handle.
- Duplicate values were there to handle.
- The column *price* had some special characters and whitespaces. These got handled for this column.
- The column *total_sqft* had some values in the range, some were present with the character value. Those were handled for this column.
- The column *size* had mentioned bedroom for some values instead of BHK. That got handled for this column.
- The column l*ocation* needed significant cleaning due to various issues with its values and whitespaces.

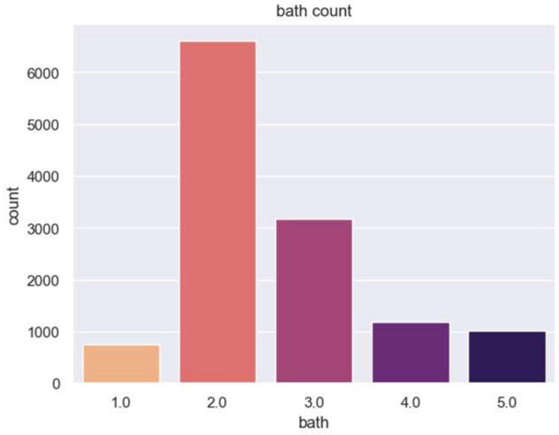
ANALYSIS

- Houses with the 'area_type' of 'super built-up area' have a higher count compared to all other categories.
- In the 'size' feature, 2 BHK and 3 BHK have a higher count than other sizes of houses.
- Houses with 1 or 2 balconies have a higher count compared to other categories.


size count

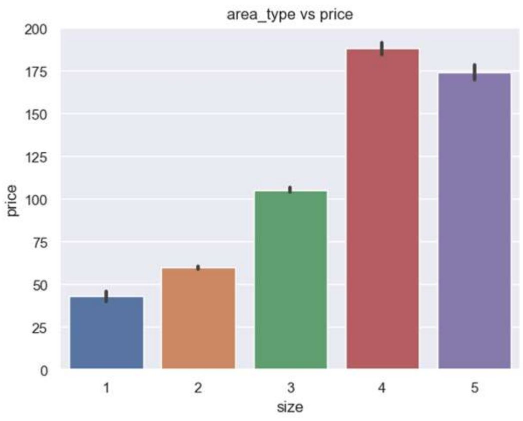

area_type count

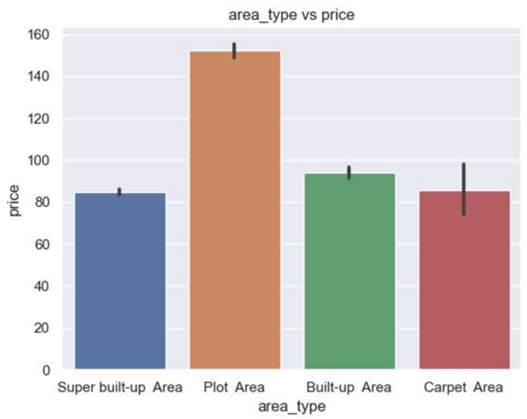


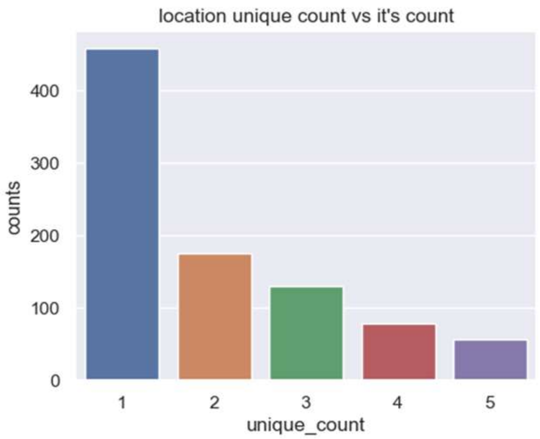

bath count

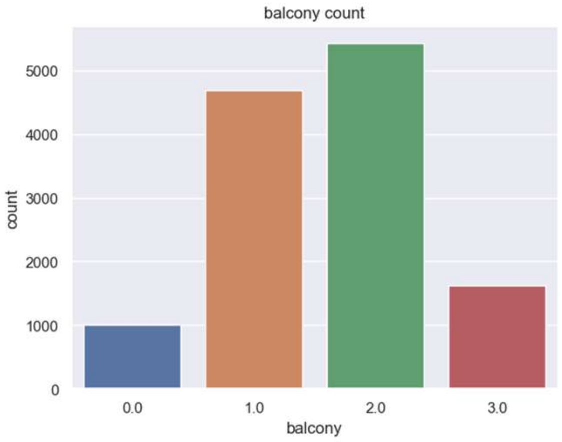
ANALYSIS (Contd.)

- In this dataset, locations with a count of 1 are more prominent than locations with higher counts.
- The area type 'Plot Area' has a higher price compared to other area types.
- 4BHK houses have a higher price compared to other sizes of houses present in this dataset.



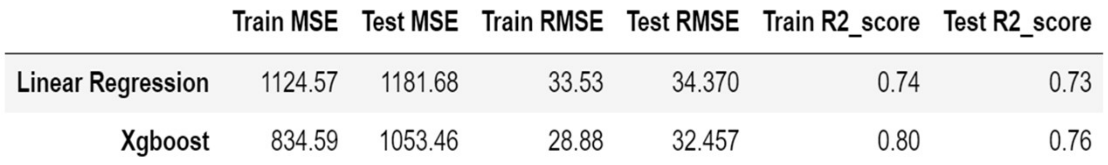






CONCLUSION

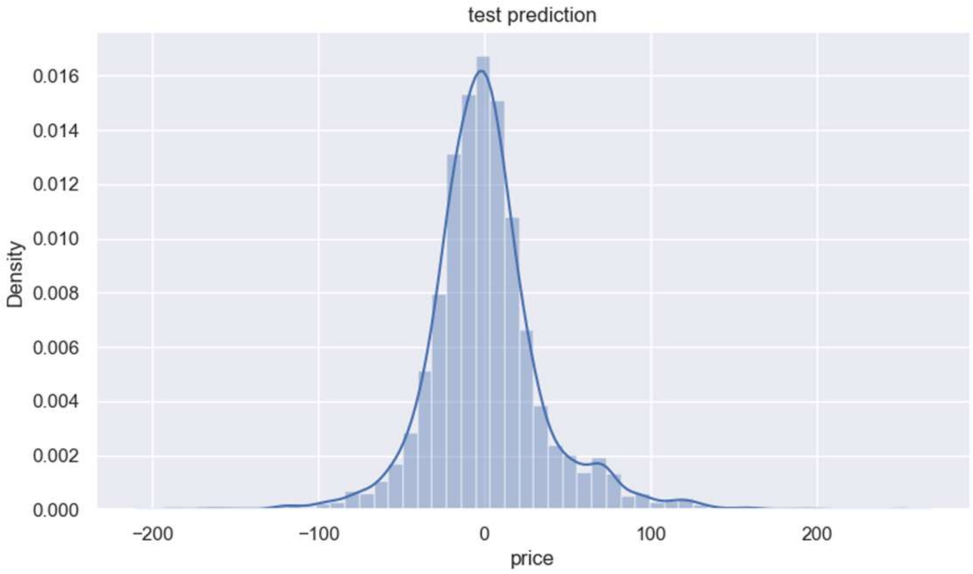
- The two models have been implemented for this dataset: one is a 'Linear Regression' model and the other is an 'XGBoost' model.
- For Linear Regression, the train R2 score is 0.74, and the test R2 score is 0.73.
- For XGBoost, the train R2 score is 0.80, and the test R2 score is 0.76.
- If we compare the results, XGBoost is providing the best results.

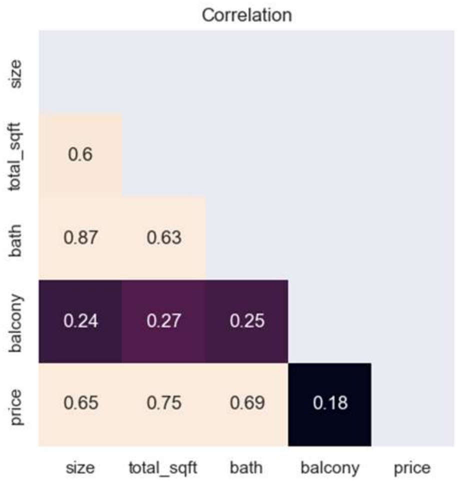


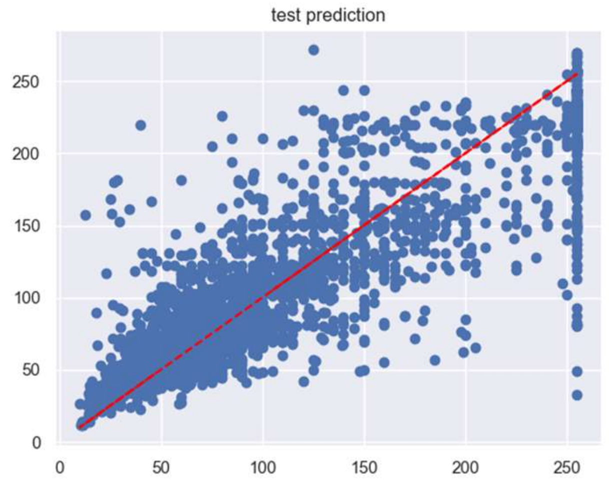
	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2_score	Test R2_score
Linear Regression	1124.57	1181.68	33.53	34.370	0.74	0.73
Xgboost	834.59	1053.46	28.88	32.457	0.80	0.76

CONCLUSION (Contd.)

- The train and test outputs exhibit a linear relationship with their predictions
- The residuals of the train and test data are normally distributed.
- There is no residual autocorrelation; the Durbin-Watson value is approximately 1.99, close to 2.
- The following independent variables exhibit correlations: 'bath' and 'size', 'total_sqft' and 'size', and 'bath' and 'total_sqft’.
- The variable 'total_sqft' is highly correlated with the dependent variable 'price'.


test prediction
Density
price


Correlation
size
total_sqft
bath
balcony
price
size
total_sqft
bath
balcony
price


test prediction