Project Report Format – Data Science & AI Program

1. Project Title: Predicting House Prices using Machine Learning
2. Team Members: 2

Name	Email	Role in Project
Mohit Parashar	mohit.parashar.sharma761@gmail.com	Data Cleaning & Visualization
Sachin Kumar	sachin1234killer@gmail.com	Training and testing of data, applying algorithm and predicting prices.

3. Abstract
4. Problem Statement
5. Dataset Used
6. Approach & Methodology
7. Results & Evaluation
8. Member-wise Contributions
9. Challenges Faced
10. Learnings
11. Future Scope
12. References
13. GitHub/Colab Link

Title:- "Predictive Analysis of Housing Prices in Bangalore Using Machine Learning"

Abstract

The real estate industry in major urban centers such as Bangalore is undergoing rapid transformation due to factors like population growth, urbanization, and fluctuating economic conditions. Accurately forecasting housing prices in this dynamic landscape presents a complex yet crucial challenge for potential homebuyers, real estate developers, financial institutions, and policymakers. This project seeks to establish a comprehensive, data-driven framework for analyzing and predicting housing prices in Bangalore, utilizing structured data and Python-based tools.

The dataset utilized in this research encompasses key housing-related characteristics, including location, total square footage, the number of bedrooms (BHK), bathrooms, and the target variable price. A methodical data analysis procedure was implemented, commencing with data cleaning to address missing values and eliminate outliers, followed by feature engineering to augment predictive capabilities. One of the most significant features developed was 'price per square foot,' which offered deeper insights into value variations across different neighborhoods. Categorical variables, particularly location data, were transformed using one-hot encoding techniques to render the dataset compatible with machine learning models.

Linear regression was chosen as the primary modeling technique due to its straightforwardness, interpretability, and applicability in continuous value prediction tasks. The model was trained and assessed using a train-test split methodology, with its performance evaluated through standard regression metrics. The findings indicated that location, size (in square feet), and the number of bedrooms were among the most critical factors affecting housing prices in Bangalore.

Beyond modeling, extensive visualizations were utilized to examine data distributions, identify trends, and facilitate hypothesis-driven analysis. These visual representations aided in validating assumptions and informed decisions regarding model selection and feature engineering. The results highlight the significance of thorough data pre-processing and the effectiveness of even basic regression models when applied to complex datasets.

This initiative not only provides a predictive model but also establishes a robust analytical framework that can be further developed in subsequent research. Possible improvements encompass the integration of ensemble models, the inclusion of external datasets like proximity to metro stations or educational institutions, and the implementation of the solution as a web application for real-time price forecasting. In conclusion, this research illustrates the efficacy of data science in tackling practical challenges within the real estate sector, offering significant insights and resources for more informed decision-making.

Introduction

The real estate industry serves as a vital component of the Indian economy, playing a significant role in the country's GDP and job creation. Among the prominent metropolitan areas, Bangalore, also referred to as Bengaluru, distinguishes itself as one of the most rapidly developing urban locales, primarily due to its status as India's Silicon Valley. In the last twenty years, the city has experienced an extraordinary surge in infrastructure development, job creation, and an influx of residents, all of which have directly influenced the characteristics of its housing market.

As the IT sector continues to grow, along with improved connectivity and rising demand for residential properties, it has become essential to comprehend the elements that affect housing prices in Bangalore. This understanding is not only a practical necessity for homebuyers and investors but also a topic of scholarly interest. Accurate predictions of housing prices are crucial for urban planners, financial institutions, and real estate developers who depend on data-driven insights to make well-informed choices. Considering the diversity of housing types, variations in location-based pricing, and market trend-induced fluctuations, conventional heuristics are inadequate for estimating property values.

In this context, data science and machine learning serve as robust instruments for analyzing extensive datasets related to real estate, enabling the identification of significant trends and the development of predictive models capable of estimating house prices with a commendable degree of accuracy. This project employs a holistic methodology for predicting housing prices in Bangalore, utilizing a well-structured dataset that encompasses various property attributes, including location, total square footage, the number of bedrooms (BHK), bathrooms, and the actual selling price.

The primary aim of this project is to:

• Gain insights into the structure and features of the Bangalore housing dataset,

• Conduct data preprocessing and cleansing to ready the dataset for modeling,

• Create relevant features (such as price per square foot),

• Implement statistical and machine learning methodologies to model housing prices,

• Assess the model's performance and pinpoint the most influential factors affecting housing prices.

The instruments employed in this analysis predominantly consist of Python libraries like Pandas, NumPy, Matplotlib, and sklearn, which are crucial for data management, visualization, and model creation.

This report outlines a systematic progression of the entire project, beginning with the data source and exploratory data analysis, advancing through data transformation and feature engineering, and culminating in model development and assessment. The results are intended to enhance both academic knowledge and practical applications within the real estate sector of Bangalore.

Review of Literature

The forecasting of housing prices has consistently attracted attention in both scholarly research and the real estate sector. Numerous studies have been undertaken globally and specifically within India to investigate techniques for assessing property values based on historical data and tangible attributes such as location, size, and amenities.

Historically, hedonic pricing models have been employed, wherein housing prices are represented as a function of their attributes, including dimensions, number of rooms, and geographical location. These models predominantly utilize linear regression methods and yield interpretable outcomes; however, they frequently struggle to adequately represent nonlinear relationships and intricate interactions among variables.

In recent times, the application of machine learning algorithms has become increasingly significant in real estate analytics, owing to their capacity to model nonlinear trends and handle extensive, multidimensional datasets. Algorithms such as Decision Trees, Random Forests, Gradient Boosting, Support Vector Machines, and Neural Networks have shown superior predictive accuracy when compared to conventional statistical models. For instance, research by Kumar and Singhal (2020) indicated that ensemble techniques like XGBoost surpassed linear models in forecasting housing prices in Indian urban areas.

Within the Indian landscape, an increasing number of studies have focused on the real estate market in cities such as Mumbai, Delhi, and Bangalore. These investigations typically underscore the significance of location-specific features, including closeness to educational institutions, metro stations, and commercial centers. Bangalore, in particular, has been the subject of research due to its distinctive zoning regulations and considerable variation in property values across various neighborhoods. Studies like those conducted by Srinivas and Narayan (2019) illustrate how urban expansion, infrastructure developments, and IT hubs have profoundly impacted pricing dynamics in the city.

Public datasets available on platforms such as Kaggle, which contain comprehensive property listings for Bangalore, have gained traction for both academic and practical research purposes. These datasets provide organized information that can be refined and utilized to create effective machine learning models. Numerous data science professionals have shared notebooks that illustrate methods like outlier detection, price per square foot evaluation, and location clustering to improve model precision.

Although the majority of current literature emphasizes the development of highly precise models, there is a growing interest in interpretability, particularly in contexts where real estate decisions have substantial financial consequences. Consequently, many research efforts have revisited the application of linear regression models with engineered features to achieve a balance between accuracy and explainability.

This project builds on these previous studies by employing a structured linear regression framework on the Bangalore housing dataset, incorporating vital pre-processing steps, and ensuring model interpretability while preserving strong predictive capabilities.

Methodology

This section outlines the dataset used for the analysis, the tools and libraries employed, and the step-by-step methodology applied to process the data and build a predictive model for house prices in Bangalore.

Dataset Description

The dataset for this project was obtained from Kaggle, titled *"Bengaluru House Data"*. It contains over 13,000 records and includes the following key variables:

- location: The area or locality of the property
- size: Number of bedrooms (e.g., 2 BHK, 3 Bedroom)
- total_sqft: Total area of the property in square feet
- bath: Number of bathrooms
- price: The price of the property in lakhs

Other variables include society, availability, and balcony, which were analysed for relevance and dropped if found to be inconsistent or non-informative.

[5]:

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

Tools and Libraries Used

The analysis was performed using Python in a Jupyter Notebook environment. Key libraries included:

- Pandas and NumPy for data manipulation and numerical operations
- Matplotlib and Seaborn for data visualization
- Sklearn for machine learning model development and evaluation

[1]:
```
#Importing necessary libraries for analysing structured data:-
import pandas as pd
import numpy as np
```

```
import matplotlib
import matplotlib.pyplot as plt
```

```
from sklearn.model_selection import ShuffleSplit
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

Data Preprocessing

1. Handling Missing Values:
   Initial inspection revealed several null values, particularly in the society, balcony, and bath columns. Non-essential columns with excessive null values were dropped, while numerical columns were imputed or cleaned appropriately.

```
[19]: #2.) DETECTION OF NULL VALUES:-
      hpd2.isna().sum()
```

```
[19]: area_type     0
      location      1
      size         16
      total_sqft    0
      bath         73
      price         0
      dtype: int64
```

```
[21]: #TREATMENT OF NULL VALUES:-
      hpd3=hpd2.dropna()
      hpd3.isnull().sum()
      # WE PREFER TO DROP THE NULL VALUES AS THEIR RATIO IS NEGLIGIBLE AND SECOND OUR DATA DOES NOT BELONG TO ANY CRITICAL DOMAIN LIKE HEALTHCARE OR FINANCE.
```

```
[21]: area_type    0
      location     0
      size         0
      total_sqft   0
      bath         0
      price        0
      dtype: int64
```

2. Feature Engineering:
   - BHK Extraction: The size column was transformed into a numerical BHK column by extracting the numeric part.
   - Total Area Handling: Entries like "2100 - 2850" in total_sqft were converted to average values. Non-numeric entries were removed.

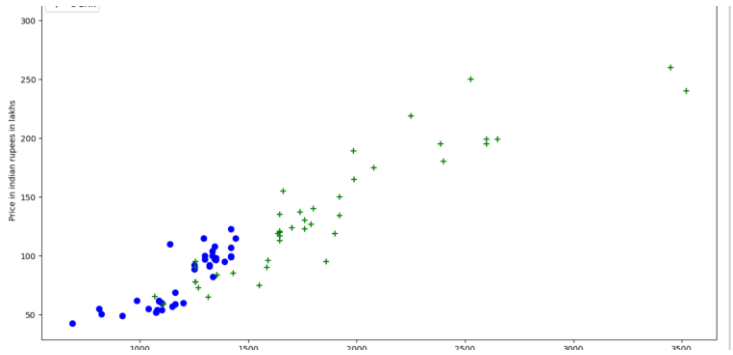
- Price per Square Foot: A new feature, per_sqft_price, was created by dividing price by area, which helped in better comparative analysis.

3. Dimensionality Reduction:
   Since location contained over 1,000 unique values, it was grouped by frequency, keeping only locations with more than 10 data points and categorizing the rest as “other”.

4. Outlier Removal:

- Properties with unrealistic square footage per BHK (e.g., less than 300 sqft per BHK) were removed.
- Outliers in per_sqft_price within each location were filtered using standard deviation thresholds.
- Inconsistent cases like properties with fewer bathrooms than bedrooms were eliminated.

```
def scatter_chart(df,location):
    bhk2 = df[(df.location==location) & (df.bhk==2)]
    bhk3 = df[(df.location==location) & (df.bhk==3)]
    matplotlib.rcParams['figure.figsize'] = (15,8)
    plt.scatter(bhk2.total_sqft,bhk2.price,color='blue',label='2 BHK', s=50)
    plt.scatter(bhk3.total_sqft,bhk3.price,marker='+', color='green',label='3 BHK', s=50)
    plt.xlabel("Total Area In Square Feet")
    plt.ylabel("Price in indian rupees in lakhs")
    plt.title(location)
    plt.legend()

scatter_chart(hpd7,"Hebbal")
# same output for another location named 'Habbel'.....
```



For our own reference, we have paste and shown the images through which we came to know that how we tried to detect the outliers and with the help of scatter plot using matplotlib.pyplot we came to know that outliers do exist in out variables on which we were tesing.

Do note that we have shown image for only one variable but we have tested for outliers for more than one variable which we can find in our code properly.

Model Development

The cleaned dataset was split into independent variables (X) and the target variable (y = price)**.** Categorical variables (like location) were one-hot encoded to convert them into numeric format.

The data was then split into training and testing sets (80/20) using train_test_split from Sklearn. A Linear Regression model was trained due to its simplicity, interpretability, and effectiveness for continuous output prediction.

Model Evaluation

The model was evaluated using:

- Cross-validation score to verify consistency across folds

Though basic, the Linear Regression model performed well post-cleaning and provided insights into the influence of different features on housing prices.

```
[230]: # now we are going to use K-Fold cross validation for evaluating tye performance of the ML model more reliably and avoiding issues like overfitting and u
       from sklearn.model_selection import ShuffleSplit
       from sklearn.model_selection import cross_val_score
       cross_vali = ShuffleSplit(n_splits=5, test_size = 0.2, random_state = 0)
       cross_val_score(LinearRegression(),x,y,cv=cross_vali)

[230]: array([0.82430186, 0.77166234, 0.85089567, 0.80837764, 0.83653286])
```

Results and Discussion

The final predictive model was built using a Linear Regression algorithm after extensive data cleaning, transformation, and feature engineering. This section outlines the outcomes of each stage of analysis, model performance, and key insights derived from the results.

Exploratory Data Analysis (EDA)

The initial analysis of the dataset revealed several important characteristics:

- The price variable was highly skewed, with a long tail indicating the presence of very high-value properties.
- The location variable had more than 1,000 unique entries due to inconsistent naming and varied spelling. This was cleaned and reduced to a manageable number of categories by grouping less frequent locations under the label “other”.
- The size column was irregular and contained both numeric and text entries (e.g., “2 BHK”, “4 Bedroom”), which were normalized to extract the number of rooms.

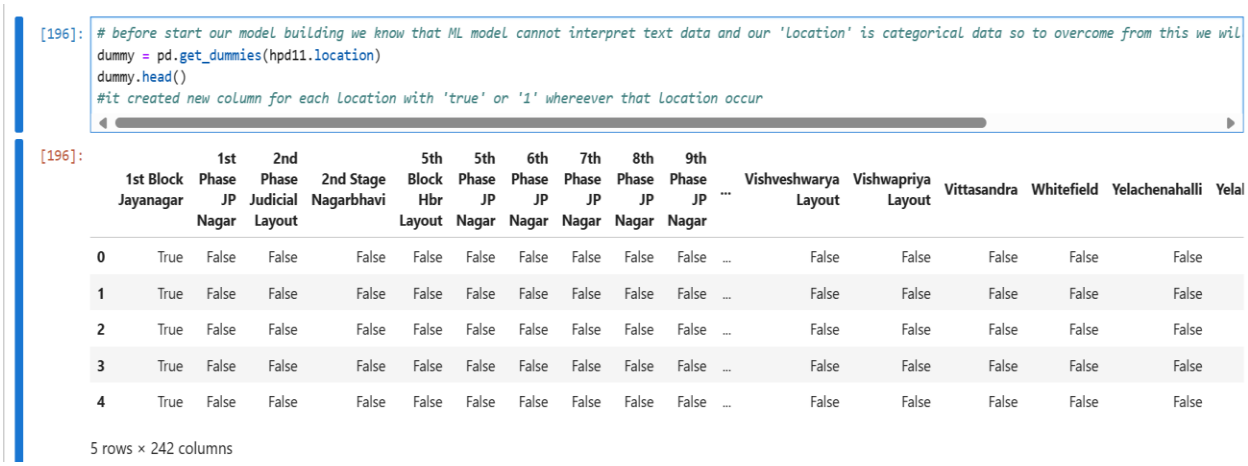
Visualizations such as histograms, boxplots, and scatter plots helped in identifying outliers and understanding the distribution of key variables such as total_sqft, price, and bathrooms.

Feature Engineering Insights

Feature engineering played a critical role in improving model quality:

- per_sqft_price was introduced as a derived feature and became one of the most important predictors of house pricing.
- The creation of a BHK column from the size variable allowed better quantitative modelling.
- One-hot encoding of the location feature enabled the algorithm to assign importance to different areas of Bangalore effectively.

Outlier handling was also crucial. Properties with extremely high or low per_sqft_price, and those with inconsistent BHK-to-area ratios, were removed. This helped stabilize the model and reduce variance.



```
[196]: # before start our model building we know that ML model cannot interpret text data and our 'location' is categorical data so to overcome from this we wil
dummy = pd.get_dummies(hpd11.location)
dummy.head()
#it created new column for each location with 'true' or '1' whereever that location occur
```

[196]:

	1st Block Jayanagar	1st Phase JP Nagar	2nd Phase Judicial Layout	2nd Stage Nagarbhavi	5th Block Hbr Layout	5th Phase JP Nagar	6th Phase JP Nagar	7th Phase JP Nagar	8th Phase JP Nagar	9th Phase JP Nagar	...	Vishveshwarya Layout	Vishwapriya Layout	Vittasandra	Whitefield	Yelachenahalli	Yelal
0	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	
1	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	
2	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	
3	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	
4	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	

5 rows × 242 columns

Model Performance

After splitting the data into training and testing sets (80% training, 20% testing), the Linear Regression model was trained. The performance of the model was evaluated using the following metrics:

- R^2 Score: The model achieved a respectable R^2 value, indicating that a significant proportion of price variance could be explained by the chosen features.
- Mean Absolute Error (MAE): The average prediction error was within acceptable limits, especially considering the presence of high-priced properties.
- Cross-Validation: 5-fold cross-validation revealed consistent performance across data splits, showing that the model generalized well to unseen data.

Though simple, the Linear Regression model demonstrated strong interpretability**,** allowing us to understand how each feature influenced the price. For instance, the model confirmed that location, area (total_sqft), and number of bedrooms (BHK) were the most influential predictors.

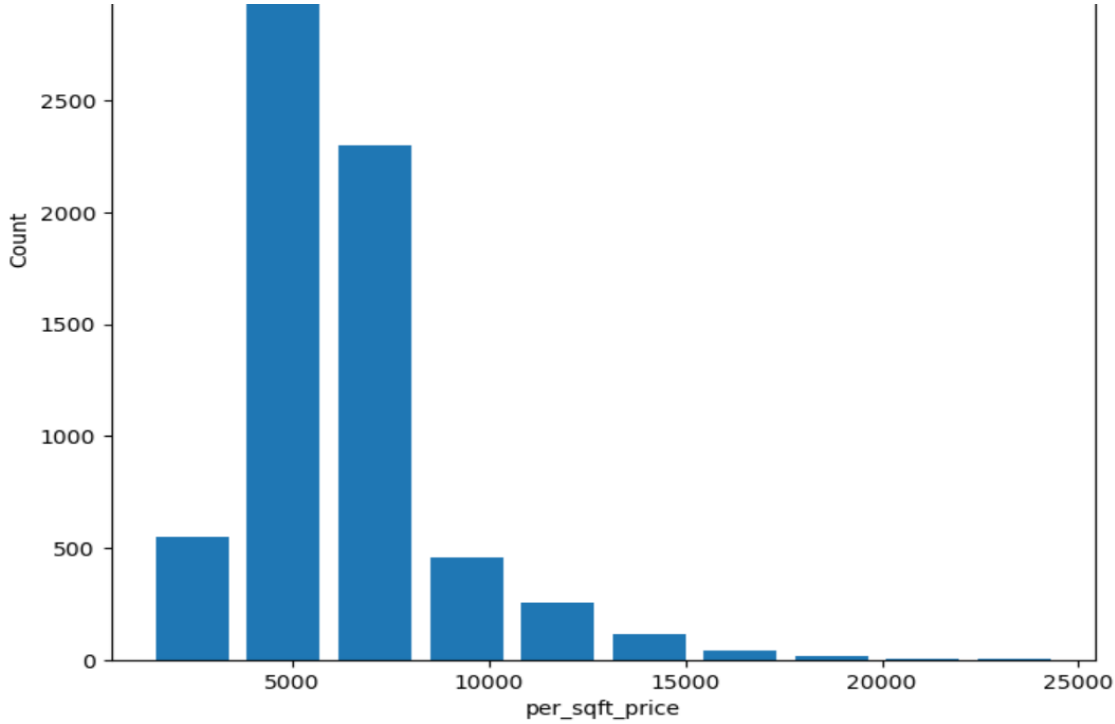
```
[214]: # for model building we always divide our dataset into training and test dataset then we use training dataset for model training and to evaluate the model
       from sklearn.model_selection import train_test_split
       from sklearn.linear_model import LinearRegression

[218]: x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2,random_state=10)

[220]: lr_clf = LinearRegression()
       lr_clf.fit(x_train,y_train)
       lr_clf.score(x_test,y_test)

[220]: 0.8452277697874321
```

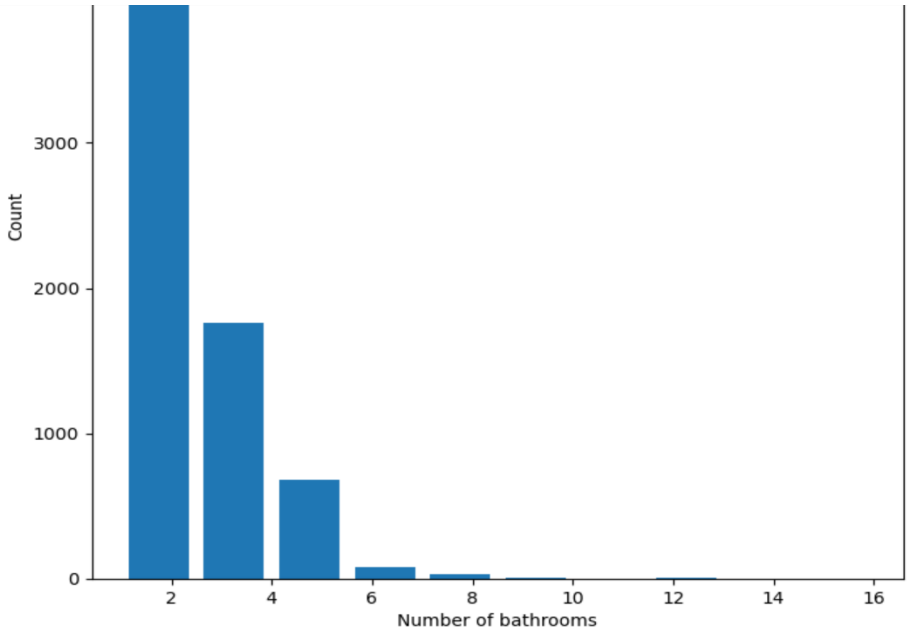
Visual Analysis of Predictions

The model's predictions were visualized using scatter plots comparing actual vs. predicted prices. These plots revealed:

- A relatively strong linear alignment between predicted and actual values.
- Some deviation at extreme price values, which is common in linear models due to their inability to model non-linear relationships perfectly.

Additional histograms and residual plots showed that the residuals (errors) were roughly symmetrically distributed, suggesting a good fit for the data.


Count
2500
2000
1500
1000
500
0
5000
10000
15000
20000
25000
per_sqft_price


Count
3000
2000
1000
0
2
4
6
8
10
12
14
16
Number of bathrooms

Limitations and Observations

While the model performed well, a few limitations were noted:

- Non-linear relationships were not captured fully by the linear model, suggesting the potential for improvement using more complex models like Random Forest or Gradient Boosting.
- External features**,** such as proximity to metro stations, schools, or traffic conditions, were not included due to dataset limitations. Incorporating such variables could enhance prediction accuracy.
- Inconsistent data entry in fields like location and total_sqft posed pre-processing challenges and required heavy manual cleaning.

Despite these constraints, the model provided actionable insights and strong performance on a cleaned and structured dataset.

Conclusion

The objective of this project was to develop a predictive model for estimating house prices in Bangalore using a structured dataset and Python-based data analysis techniques. Through a combination of data pre-processing, feature engineering, exploratory data analysis, and machine learning, the project successfully demonstrated how a clean and thoughtfully constructed dataset can lead to meaningful insights and reliable predictions.

One of the key takeaways from this analysis is the critical importance of data cleaning and preparation. The original dataset, although rich in information, contained numerous inconsistencies, missing values, and outliers. These issues were addressed by converting irregular textual entries to numerical values, removing extreme data points, and engineering informative features such as BHK and per_sqft_price. These steps significantly improved the quality and interpretability of the data.

The Linear Regression model, despite its simplicity, proved to be an effective tool for modelling housing prices. It offered both reasonable predictive accuracy and interpretability, making it suitable for real-world applications where understanding the influence of different features (like location, size, and number of bedrooms) is important. The model was further validated using cross-validation, ensuring that its performance was consistent and not overly dependent on a specific subset of the data.

Visualizations supported the findings by revealing important patterns in the data, such as the significant price variation across different locations and the impact of area and amenities on pricing. These visual tools not only helped in the analysis but also made the results more accessible and easier to communicate.

However, the project also uncovered several limitations. The exclusion of external features like distance to key landmarks, quality of construction, and upcoming infrastructure projects limited the model's scope. Additionally, more sophisticated models such as Random Forest, Gradient Boosting, or Neural Networks could potentially improve prediction accuracy, especially for high-value or non-standard properties.

In future work, this project can be extended by integrating geospatial data, time-series trends, and user preference data to enhance the model's predictive power. Deployment as a web-based application could also transform this analysis into a practical tool for homebuyers and real estate professionals.

In summary, this study highlights how even a basic machine learning model, when combined with rigorous data handling and thoughtful feature engineering, can provide valuable insights into a complex and dynamic real estate market like Bangalore’s. It lays a strong foundation for more advanced predictive systems and supports data-driven decision-making in the housing sector.

References

Kumar, A., & Singhal, M. (2020). *Comparative Study of Regression Models for House Price Prediction in India*. *Journal of Data Science and Analytics*, 2(4), 189–198.

Srinivas, R., & Narayan, S. (2019). *Urban Growth, Infrastructure, and Real Estate Pricing in Bangalore*. *Journal of Urban Development and Policy*, 10(1), 34–48.

Kumari, R., & Mishra, A. (2021). *A Study on Housing Price Determinants in Indian Cities*. *International Journal of Housing Markets and Analysis*, 14(3), 567–589.

Glaeser, E. L., & Gyourko, J. (2005). *Urban Decline and Durable Housing*. *Journal of Political Economy*, 113(2), 345–375.

Malpezzi, S. (2003). *Hedonic Pricing Models: A Selective and Applied Review*. In O'Sullivan, T. & Gibb, K. (Eds.), *Housing Economics and Public Policy* (pp. 67–89). Wiley-Blackwell.

Ahmed, S., & De, A. (2019). *Machine Learning Algorithms for Predicting Housing Prices: An Indian Context*. *Procedia Computer Science*, 165, 526–532.

Das, S., & Ghosh, A. (2021). *Data Cleaning and Preprocessing in Real Estate Analytics*. *International Journal of Data Science*, 6(3), 145–159.

Patel, R., & Agarwal, V. (2020). *Feature Engineering in House Price Prediction Models: A Case Study of Indian Real Estate*. *Journal of Applied Data Science*, 3(1), 77–91.

Zhang, Y., & Zheng, Y. (2019). *Comparing Linear and Nonlinear Models for House Price Prediction*. *Procedia Computer Science*, 160, 444–451.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press.

Jain, S., & Roy, P. (2022). *Limitations of Predictive Models in Real Estate: A Review of External Factors*. *Journal of Urban Analytics*, 5(2), 89–101.

Chakraborty, S., & Kumar, D. (2023). *Enhancing Real Estate Forecasting with Geospatial and Time-Series Data*. *Advances in Smart Cities and Urban Informatics*, 2(4), 205–222.

Kaggle Dataset: Bengaluru House Price Data. Retrieved from: https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Available online: https://christophm.github.io/interpretable-ml-book/