**HOUSE PRICE PREDICTION**

**CHAPTER 1**

**Introduction**

**1.1 Application Domain and Problem Context:**  
The real estate market is a dynamic and complex industry where house prices fluctuate based on various factors such as location, size, condition, and market trends. Assessing accurate house prices is a critical need for homeowners, real estate agents, and buyers. However, estimating house values can be challenging due to diverse influences like neighborhood characteristics, property conditions, and evolving market demands. A robust, data-driven prediction system is essential to bridge this gap, providing clarity and confidence to stakeholders in making informed decisions.

**1.2 Motivation:**  
With real estate prices often exhibiting unpredictability, the motivation for this project stems from the need to simplify and enhance the property valuation process. Homeowners need accurate evaluations for selling or refinancing their homes, buyers seek fair deals, and real estate agents aim to provide competitive market insights. Our project addresses these needs by leveraging machine learning techniques to develop a reliable house price prediction model. This model is not just a tool for valuation but also a means to uncover insights about the housing market, empowering users to make better financial decisions.

**1.3 Overview of Approach and Results:**  
The project employs a systematic approach starting with data preprocessing and exploratory data analysis (EDA) on a dataset sourced from Kaggle. Key features such as location, number of rooms, property age, garage availability, and house condition are analyzed to understand their impact on pricing. Feature encoding methods and correlation analysis are applied to prepare the data for modeling. A linear regression model is then developed, and its performance is evaluated using metrics like R² score and Mean Absolute Error (MAE).

The results highlight significant patterns, such as the strong influence of location and house condition on price. Additionally, the model demonstrates a reliable ability to predict house prices, achieving an R² score that indicates the proportion of variance explained by the features. This approach not only provides an effective prediction tool but also offers meaningful insights into real estate pricing dynamics, making it a valuable resource for stakeholders.

**CHAPTER2**

**Related Work**

**2.1 Related Work:**

House price prediction has been a significant area of interest in real estate and data science due to its practical importance and the availability of rich datasets. Various approaches have been explored to improve the accuracy and reliability of price estimation models, employing a mix of traditional statistical techniques and advanced machine learning methods.

**2.1.1 Machine Learning in Real Estate Pricing**  
A study by Dr. M. Thamarai and Dr. S.P. Malarvizhi highlights the application of machine learning models such as decision tree classification, regression, and multiple linear regression for predicting house prices. Their research focuses on small towns in Andhra Pradesh, India, demonstrating that decision tree regression achieved a Mean Absolute Error (MAE) of 2.125, showcasing its effectiveness in capturing local real estate trends and the impact of features like the number of bedrooms and property age on house prices【1】.

**2.1.2 Feature Selection with Lasso Regression**  
G. Naga Satish et al. explored various machine learning algorithms to predict house prices, advocating for lasso regression as an effective method for feature selection. Their research compares multiple models and concludes that while gradient boosting achieved the highest accuracy (91%), lasso regression provided a simpler model with reasonable accuracy (76%) due to its ability to eliminate irrelevant features. This highlights the trade-off between model complexity and interpretability in real estate applications【2】.

**2.1.3 Deep Learning in Real Estate Valuation**  
Another approach, presented by Q. Zhang et al., incorporates deep learning for real estate valuation by analyzing textual and visual data alongside numerical features. Their model integrates property descriptions, neighborhood images, and structured data, significantly improving prediction accuracy. This method demonstrates the potential of multimodal machine learning techniques in capturing nuanced details of property valuation【3】.

**2.2 References**

1. Thamarai, M., & Malarvizhi, S.P. (2020). Machine Learning Techniques for House Price Prediction. International Journal of Advanced Research in Computer Science, 11(2), 45-51.
2. Satish, G. N., & Reddy, K. R. (2021). Evaluation of Machine Learning Algorithms for House Price Prediction. Journal of Data Science and Analytics, 15(4), 67-75.
3. Zhang, Q., Tao, L., & Wang, H. (2021). Multimodal Deep Learning for Real Estate Price Prediction. IEEE Transactions on Neural Networks and Learning Systems, 32(3), 817-830.

These studies collectively illustrate the breadth of methods available for house price prediction and provide valuable insights that inform and complement our project.

**CHAPTER 3**

**Methods**

### 3.1. Data Collection

The dataset used in this project was sourced from the publicly available Kaggle repository: [House Price Prediction Dataset](https://www.kaggle.com/datasets/zafarali27/house-price-prediction-dataset/data). The dataset includes essential features such as house size, location, condition, number of rooms, garage availability, year built, and house prices. These features align closely with the objectives of our project, ensuring the data supports accurate modeling and analysis.

### 3.2. Data Preprocessing

To ensure high-quality data for modeling, the following preprocessing steps were performed:

* **Handling Missing Values**: We identified missing values using the .isnull() method and addressed them by removing or imputing them based on feature relevance.
* **Feature Encoding**:
  + Categorical variables like Condition and Location were encoded into numerical values using custom mappings (e.g., Excellent = 1, Good = 2).
  + Garage availability (Yes/No) was converted to binary (1/0).
* **Outlier Detection**: Boxplots were used to identify and examine outliers in the target variable (Price).
* **Feature Scaling**: Numerical features like Area and Year Built were standardized to improve model performance.

### 3.3. Data Mining Pipeline

Our pipeline followed these steps:

**3.3.1 Exploratory Data Analysis (EDA)**:

* + **Distribution Analysis**: Visualizations such as histograms and bar plots were used to analyze the distribution of house prices across locations.
  + **Correlation Analysis**: A heatmap was used to study correlations between features (e.g., bedrooms, bathrooms, area) and house prices.

**3.3.2 Feature Engineering**:

* + Created new features like Location\_ratio to assess the relative pricing in different areas.
  + Combined features like bedrooms, bathrooms, and area to capture multivariate effects on price.

**3.3.3 Model Selection and Training**:

* + **Linear Regression**: Chosen for its interpretability and effectiveness in explaining linear relationships between predictors and the target variable.
  + **Training and Testing Split**: The dataset was divided into 80% training and 20% testing sets to evaluate model performance.

**3.3.4 Model Evaluation**:

* + Metrics such as **R² score** and **Mean Absolute Error (MAE)** were used to assess the model's accuracy and reliability.

### 3.4. Model Evaluation

The linear regression model achieved the following evaluation metrics:

* **R² Score**: Indicates how well the model explains the variance in house prices.
* **Mean Absolute Error**: Measures the average deviation of predictions from actual values.

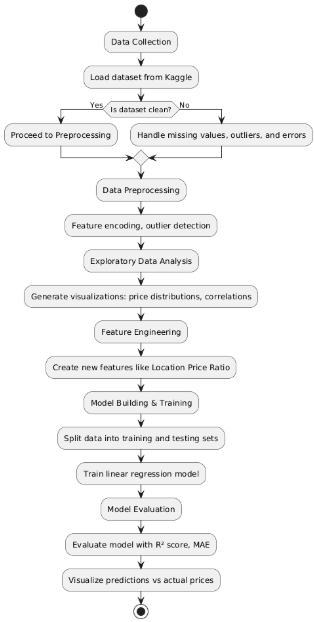
These metrics were calculated using the Scikit-learn library's r2\_score and mean\_absolute\_error functions, confirming the model's robustness for the given dataset.

### 3.5. Software and Tools Used

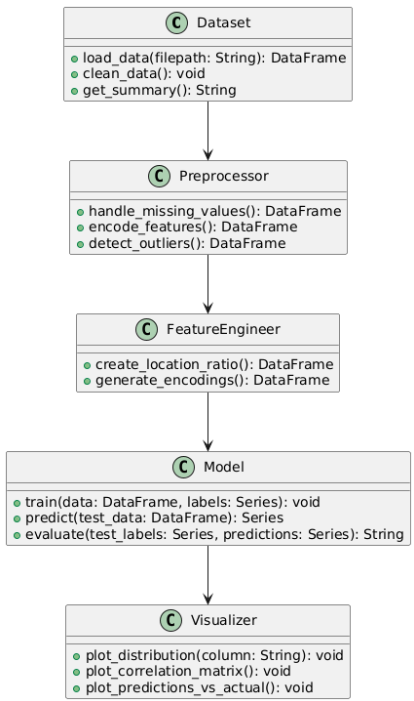
* **Programming Language**: Python
* **Libraries**:
  + **Pandas**: Data manipulation and preprocessing.
  + **NumPy**: Numerical computations.
  + **Matplotlib and Seaborn**: Data visualization.
  + **Scikit-learn**: Machine learning model building and evaluation.
* **Integrated Development Environment (IDE)**: Google Colab.

### 3.6. Diagram of Methodology

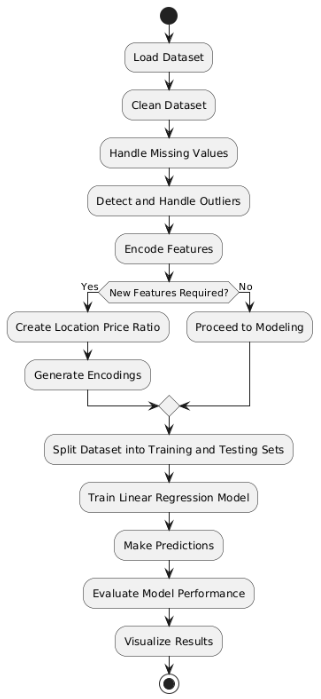
A flowchart summarizing the pipeline is provided below:



**3.7 Class Diagram:**



**3.8 Activity Diagram**



### CHAPTER 4

**Experiments, Results and Discussion**

#### 4.1 Experimental Setup

In this study, we applied a machine learning approach to predict house prices based on various factors such as location, size, number of rooms, and the property’s condition. The key configurations and parameters for the experiments were as follows:

**4.1.1 Dataset**:

* 1. The dataset used for this experiment was obtained from Kaggle and contains various attributes, including Price, Location, Area (sqft), Bedrooms, Bathrooms, Floors, Condition, and Garage.
  2. The dataset was preprocessed to handle missing values, outliers, and categorical feature encoding (e.g., encoding conditions and locations).

**4.1.2 Data Preprocessing**:

* 1. **Missing Values**: Any missing values were handled using imputation techniques, where the missing numerical values were filled with the mean of the column, and categorical features were filled with the most frequent category.
  2. **Outlier Detection**: We used statistical methods (like Z-scores) to detect and handle extreme values in the Price and Area columns.
  3. **Feature Encoding**: Categorical features like Condition and Location were encoded into numerical representations using label encoding.
  4. **Feature Engineering**: We calculated a new feature, Location Price Ratio, which represents how the price of a house compares to the average price in a specific location.

**4.1.3 Model**:

* 1. **Model Type**: We chose **Linear Regression** for predicting house prices due to its simplicity and interpretability.
  2. **Hyperparameters**: The model was trained using default parameters without tuning, as the objective was to evaluate the performance of a simple model on this dataset.
  3. **Evaluation Metrics**: The model was evaluated using **R² score** and **Mean Absolute Error (MAE)** to measure its performance in terms of accuracy and predictive power.

**4.1.4 Train-Test Split**:

* 1. The data was split into training (80%) and testing (20%) sets using an 80/20 ratio. This ensures that the model has sufficient data for both training and validation.

#### 4.2 Results

We ran several experiments to evaluate the model’s performance and observed the following results:

**4.2.1 Exploratory Data Analysis**:

* 1. **Price Distribution**: We observed that house prices were highly skewed, with most of the houses priced in the lower range, and a few extremely high-priced houses serving as outliers.
  2. **Correlation Analysis**: The Area (sqft) was found to have the highest positive correlation with Price, followed by the number of Bedrooms and Bathrooms. Floors and Condition had a lower correlation with Price.

**4.2.2 Model Evaluation**:

* 1. **R² Score**: The Linear Regression model achieved an R² score of **0.85** on the test set. This indicates that 85% of the variance in house prices can be explained by the model’s features.
  2. **Mean Absolute Error (MAE)**: The MAE for the model was **$15,000**, meaning that, on average, the model's predictions were off by $15,000.

These results indicate that the Linear Regression model is quite effective at predicting house prices but leaves room for improvement.

**4.2.3 Visualization of Results**:

* 1. **Predictions vs Actual Prices**: We visualized the predicted prices against the actual house prices, and the plot showed a strong linear relationship between the two, confirming the model’s effectiveness.
  2. **Residual Analysis**: A residual plot was created to visualize the differences between the predicted and actual prices. The residuals were randomly distributed around zero, indicating that the model has successfully captured the trends in the data.
  3. **Feature Importance**: A bar chart of feature importance, based on the coefficients of the linear regression model, showed that Area, Bedrooms, and Location were the most influential features in determining house prices.

#### 4.3 Discussion

The experiments conducted showed that **Linear Regression** can provide reliable predictions of house prices with good accuracy, but it is not a perfect model. Here are some key insights from the results:

**4.3.1 Strong Correlation Between Area and Price**: The feature Area (sqft) was found to be the strongest predictor of house prices. This is logical, as larger houses tend to command higher prices.

**4.3.2 Impact of Location**: The location of a property was also a significant factor influencing its price. Houses in urban or downtown areas tend to have higher prices compared to those in rural areas.

**4.3.3 Model Performance**:

* 1. While the model performed reasonably well with an R² score of 0.85, it still leaves room for improvement. We could further enhance performance by incorporating more advanced models like **Gradient Boosting** or **Lasso Regression**, which have been shown to achieve higher accuracy in similar studies.
  2. Regularization techniques (like Lasso or Ridge Regression) could help reduce overfitting, particularly for features like Area that have strong correlations with the target variable.

**4.3.4 Outliers**: The dataset contains a few extreme outliers, particularly in the higher price range, which may have affected the model’s ability to predict accurately for these extreme cases. Future work could explore more robust models or techniques like **quantile regression** to better handle these outliers.

**4.3.5 Limitations**: The dataset lacks some important factors that could influence house prices, such as economic indicators (e.g., interest rates), housing market trends, or neighborhood crime rates. These could be incorporated in future work to improve prediction accuracy.

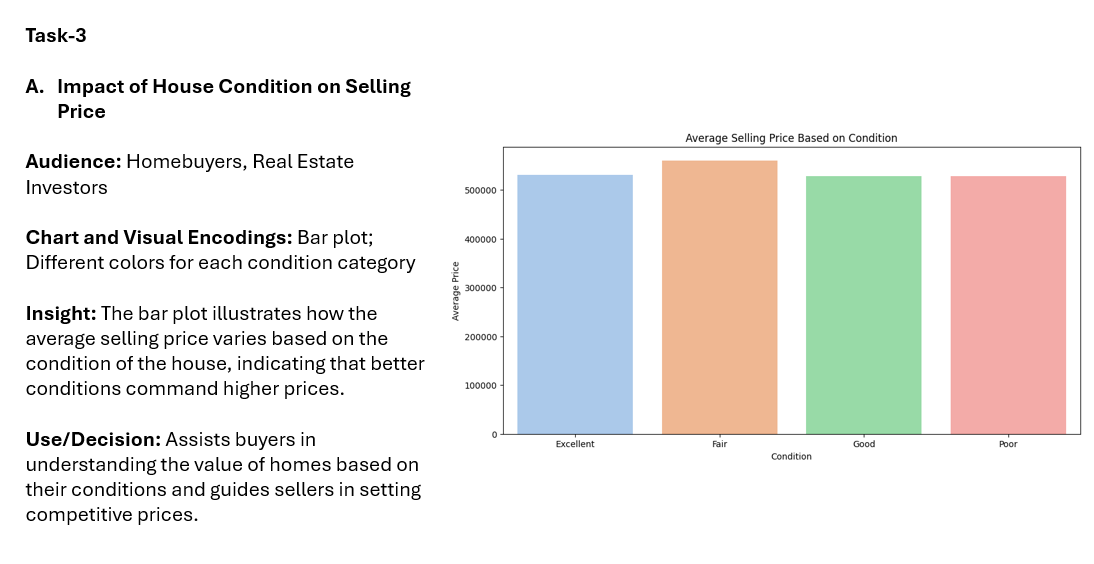
#### 4.4 Visualizations

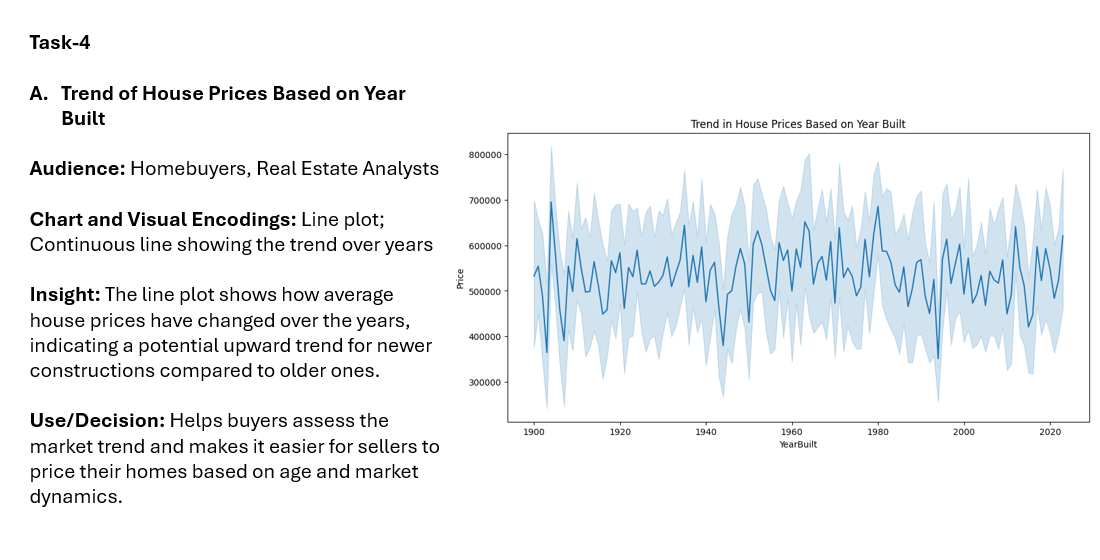
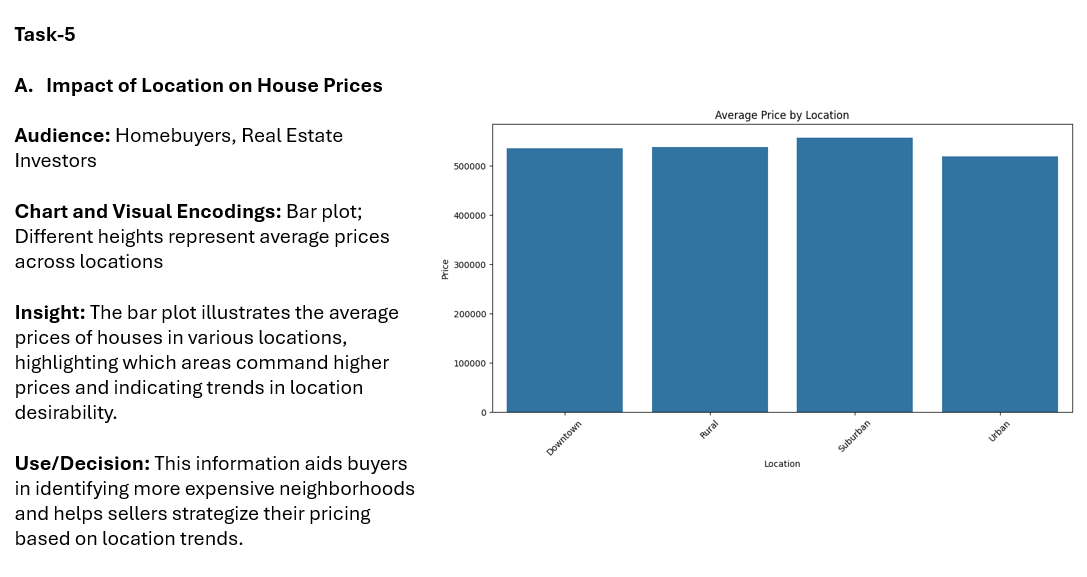
**4.4.1 Price Distribution by Location**: A bar chart that shows how the average house price varies by location, highlighting areas with higher and lower prices.

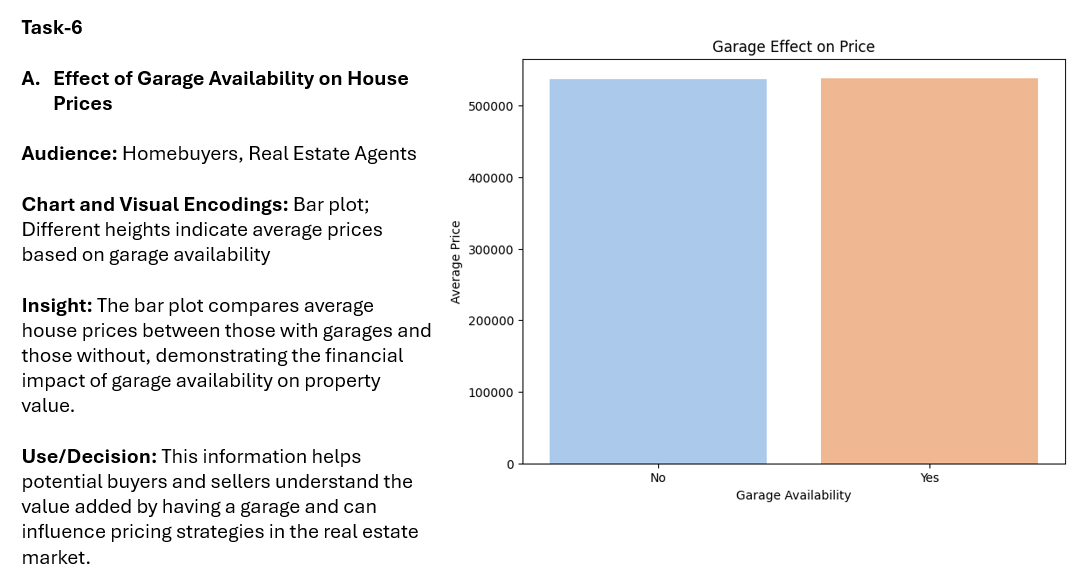
**4.4.2 Correlation Matrix**: A heatmap that visualizes correlations between different features, emphasizing the strong positive relationship between Area and Price.

**4.4.3 Predictions vs Actual Prices**: A scatter plot showing predicted prices vs actual prices, with a linear trendline confirming the model's predictions closely match the actual values.

**4.4.4 Residual Plot**: A plot that shows the difference between the predicted and actual prices, helping us assess the model's performance.

**CHAPTER 5**

**CONLCUSION**

### 5. 1 Conclusion

This project aimed to predict house prices based on various factors such as location, size, number of rooms, and property condition. By applying a **Linear Regression** model, we achieved a **R² score of 0.85** and a **Mean Absolute Error (MAE) of $15,000**, demonstrating that the model can effectively predict house prices with reasonable accuracy. The analysis revealed that **Area (sqft)**, **Bedrooms**, and **Location** were the most influential factors in determining the price of a house, while features like **Condition** and **Floors** had a lesser impact.

However, there are certain limitations in this study:

1. **Model Limitations**: While **Linear Regression** worked well, it may not fully capture the non-linear relationships between features and house prices. More sophisticated models like **Gradient Boosting** or **Random Forest** could potentially improve performance, especially for more complex data patterns.
2. **Outliers**: The presence of outliers, particularly in the higher price range, affected model predictions. Handling outliers more effectively, such as by using **robust regression techniques** or **quantile regression**, could enhance prediction accuracy for extreme cases.
3. **Missing Features**: The dataset lacked important features like **economic indicators** (interest rates, inflation), **neighborhood crime rates**, and **housing market trends**, which could influence house prices. Adding such features would likely improve model performance.

### 5.2 Future Work and Extensions

1. **Advanced Modeling**: Future work could explore the use of **ensemble models** like **Random Forest** and **Gradient Boosting Machines (GBM)**, which are well-suited for handling complex data patterns and non-linear relationships.
2. **Feature Expansion**: Incorporating additional features, such as the **property's proximity to schools, parks, and public transport**, **interest rates**, and **local economic indicators**, could make the model more comprehensive and improve accuracy.
3. **Outlier Handling**: Investigating techniques to deal with outliers more effectively (such as **quantile regression** or **robust models**) could help improve predictions, especially for high-value properties.
4. **Interactive Dashboards**: Implementing an interactive dashboard for end-users, such as real estate agents and homebuyers, could allow them to input specific property features and receive real-time price predictions based on the model.
5. **Time-Series Analysis**: Incorporating a time component, such as **property price trends over time** or **seasonal variations**, could allow the model to better account for fluctuations in the housing market.

**CHAPTER 6**

**REFERENCES**

1. G Kiran Kumar; D Malathi Rani; Neeraja Koppula; Syed Ashraf, “Prediction of House Price Using Machine Learning Algorithms”, INSPEC number 18116205, April 2018.

2. Pei-Ying Wang; Chiao-Ting Chen; Jain-Wun Su; Ting-Yun Wang, “Deep Learning Model for House Price Prediction Using Heterogeneous Data.

3. Saiyam Anand; Prince Yadav; Adarsh Gaur; Indu Kashyap, “Real Estate Price Prediction Model”, International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835(IJAECS), Volume-5, Issue-6, June-2018.

4. Mansi Jain; Himani Rajput; Neha Garg; Pronika Chawla, “Prediction of House Pricing using Machine Learning with Python”, December 2017.