SUMMER TRAINING/INTERNSHIP

PROJECT REPORT

(Term June-July 2025)

(Bengaluru House Price Prediction)

Submitted by

Hrituparna Ghosh-12316169

Puravanshi Attri-12315036

Sneha Srivastava-12316057

Hridyansh Isher-12316175

Raghvendra Pratap Singh-12314171

Course Code ……PETV76………………

Under the Guidance of

(Dr. Jaffar Amin Chacket and Dr.Sandip Kaur)

School of Computer Science and Engineering

**DECLARATION**

We Hrituparna, Sneha, Hridyansh, Puravanshi, Raghvendra. student of LPU under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 Signature

Registration No. 12316169,12316057,12316175,12315036,12314171

**CERTIFICATE**

This is to certify that Hrituparna Ghosh bearing Registration no. 12316169, Sneha Srivastava bearing Registration no.12316057 , Puravanshi Attri bearing Registration no. 12315036, Hridyansh Isher bearing Registration no. 12316175, Raghvendra Pratap Singh bearing Registration no. 12314171 has completed ...PETV76-........ <Course Code> project titled, **“Bengaluru House Price Prediction”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date:

**ACKNOWLEDGEMENT**

we would like to express my heartfelt gratitude to my teachers for their invaluable guidance and unwavering support throughout our journey. Their dedication to education and commitment to nurturing our growth have profoundly influenced our understanding and passion for learning, inspiring me to pursue our goals with confidence and determination.

We are also deeply thankful for our friends, whose encouragement and camaraderie have made this experience enjoyable and enriching. Their unwavering support, thoughtful insights, and meaningful discussions have inspired us to push our boundaries and strive for excellence in every endeavour and challenge we faced along the way.

Additionally, we owe a special debt of gratitude to our family for their love and encouragement. Their belief in our abilities has been our greatest motivation, providing us with the strength to overcome obstacles and persevere through difficult times.

Finally, we want to thank everyone for being our pillars of support and for believing in us every step of the way. Your contributions have helped us grow into the person We are today, and We are forever grateful.

**TABLE OF CONTENTS**

Chapter 1: Introduction

Chapter 2: Training Overview

Chapter 3: Project Details

Chapter 4: Implementation

Chapter 5: Results and Discussion

Chapter 6. Conclusion

Chapter 7. Future scope

Chapter 8. References

**Introduction**

This project delves into the **housing market of Bengaluru**, focusing on the analysis of key parameters affecting property prices across the city. Using a real-world dataset, the project applies **data cleaning and exploratory data analysis (EDA)** techniques to extract meaningful insights about house pricing trends, property size, and configuration (such as BHK).

The dataset is first thoroughly cleaned by handling missing values, converting non-standard formats like square footage ranges, and engineering features like **number of bedrooms (BHK)**. This preprocessing ensures a consistent and reliable foundation for deeper analysis.

The visualizations include **bar plots, distribution charts, and correlation heatmaps** that highlight trends such as the most common BHK configurations, average prices per square foot, and the relationship between property size and price. These graphs help in decoding the underlying patterns of real estate dynamics in one of India’s fastest-growing tech hubs.

Special attention is given to **outlier detection**, with boxplots used to identify properties with unusually high prices or sizes. Additionally, price-per-square-foot metrics allow for better comparability across different property types and sizes, providing a normalized view of affordability.

By presenting these insights in an interactive and structured format, this project serves as a useful tool for **home buyers, real estate analysts, and data enthusiasts**. It demonstrates how simple data processing and visual storytelling can make complex real estate data more accessible, ultimately helping to make informed decisions in the property market.

* Problem **Statement**

In rapidly urbanizing cities like Bengaluru, the real estate sector plays a critical role in both economic development and individual financial planning. However, the housing market is highly complex, with property prices influenced by multiple interdependent factors such as location, number of bedrooms, total area, and market demand. For buyers, investors, and analysts, making informed decisions is challenging due to inconsistent data, lack of transparency, and the absence of centralized, easy-to-understand insights into price trends and configurations. The availability of real estate data is increasing, but the raw format in which it is often presented—full of missing values, inconsistent formatting, and scattered information—makes it difficult to derive actionable insights.

One of the main problems lies in the lack of a streamlined process to clean, organize, and analyze this housing data. Inaccuracies such as non-standard formats in square footage or missing entries in key columns like size and location further complicate the analysis. Moreover, the inability to easily compare configurations such as 1BHK, 2BHK, or 3BHK homes in different regions of the city limits the effectiveness of real estate assessments. Without reliable tools for data exploration and visualization, both casual homebuyers and industry professionals often struggle to understand market trends or spot investment opportunities.

Another challenge is the presentation of real estate insights in a format that is accessible and interactive. While statistical models and raw numbers may be valuable to data scientists, they are often inaccessible to end users who may lack technical expertise. This creates a gap between available data and its practical use in decision-making, especially when it comes to visually identifying outliers, comparing property values, or observing locality-based trends over time.

This project addresses these issues by developing a structured workflow using Python for data cleaning and exploratory analysis, and Power BI for building an interactive dashboard. By converting raw real estate data into clean, interpretable, and visual insights, the project aims to simplify property analysis for a wide range of users. The Power BI dashboard presents key metrics such as average price, BHK-wise distributions, and area-based trends through intuitive charts and filters. This approach not only aids in informed decision-making but also promotes data-driven understanding of the housing market in Bengaluru, making real estate analysis both efficient and user-friendly.

* **Objectives**

The objective of this study is to apply Python-based data analysis and Power BI data visualization techniques to extract valuable insights from the Bengaluru housing market. The project aims to understand and interpret trends in property pricing, configuration, and size by analysing a real-world dataset. By cleaning and processing the data using Python libraries such as Pandas and NumPy, the project ensures accuracy and consistency in the dataset, which is essential for generating meaningful analysis. Key preprocessing steps included handling missing values, dropping irrelevant columns, and converting non-standard formats such as range-based square footage and mixed-format size entries. A new feature, BHK (number of bedrooms), was derived to further structure the data.

Following the data cleaning process, the study focuses on identifying patterns and variations in property prices across different localities and configurations. By analyzing factors such as total square footage, price per square foot, and BHK count, the project uncovers important trends that help explain how these variables influence real estate pricing in Bengaluru. This analysis also allows for the detection of outliers and abnormal listings, which can be crucial in assessing the reliability of market values. Through visualization techniques using Matplotlib and Seaborn, the project highlights distribution patterns, correlations, and price behaviors across different property types.

To make these findings accessible and interactive, the project also includes the development of a comprehensive Power BI dashboard. This dashboard visually presents key metrics and trends using elements such as bar charts, pie charts, line graphs, and KPI cards. It allows users to filter data by area, BHK type, or pricing category, supporting dynamic exploration and interpretation of the housing data. The dashboard serves as a practical tool for stakeholders such as home buyers, investors, and real estate analysts to gain a clearer understanding of Bengaluru’s real estate landscape.

Ultimately, this project aims to simplify complex property data through visual storytelling and structured analysis. By combining data cleaning, exploratory analysis, and dashboard visualization, the study contributes to more informed decision-making in the real estate sector, helping users identify pricing trends, evaluate affordability, and recognize potential investment opportunities within the Bengaluru housing market.

## **Training overview**

**Tools and Technologies used**

Throughout the project, a combination of programming tools and data visualization platforms were used to clean, analyze, and present real estate data effectively:

* **Python** – Primary programming language for data processing and exploratory analysis.
* **Pandas** – Used for data manipulation, cleaning, and feature engineering.
* **NumPy** – Supported numerical operations and efficient handling of arrays.
* **Matplotlib & Seaborn** – For creating detailed plots such as histograms, boxplots, and heatmaps to visualize trends.
* **Jupyter Notebook** – Environment used to develop and document the entire data analysis workflow interactively.
* **Power BI** – Business intelligence tool used to build an interactive dashboard to visualize real estate trends dynamically.
* **CSV Files** – Used as the format for importing raw data and exporting cleaned data for use in Power BI.

**Areas covered during project**

The training program covered both technical and analytical aspects of real estate data interpretation. The major areas covered were:

* **Data Collection & Loading**: Reading and understanding the structure of the housing dataset.
* **Data Preprocessing**: Handling missing data, converting inconsistent values (e.g., range in square footage), and dropping irrelevant features such as ‘society’ and ‘availability’.
* **Feature Engineering**: Creating new metrics such as **BHK** (from size) and standardizing values like **total square footage**.
* **Exploratory Data Analysis (EDA)**: Using Python libraries to understand price distributions, detect outliers, and analyze relationships between variables (e.g., price vs. size).
* **Interactive Dashboard Design**: Leveraging **Power BI** to build a comprehensive dashboard showing:
  + Top 10 costliest and cheapest properties
  + BHK-wise distribution
  + Area-wise average prices
  + Monthly and yearly price trends
  + Boxplots for outlier detection
  + KPIs for total listings, average price, average area, and total BHK
* **Storytelling Through Visualization**: Structuring visual elements for clarity, insights, and interactivity to support decision-making.

**Weekly Work Summary**

**Week1**- Loaded the raw dataset, explored column meanings, and identified missing values and formatting issues.

**Week2**- Cleaned the data using Python by removing irrelevant fields, converting formats, and extracting BHK values. Performed EDA using Seaborn and Matplotlib to analyze price trends and distributions.

**Week3**- Imported the cleaned data into Power BI and built an interactive dashboard with charts,

filters, and KPIs to visualize housing insights effectively.

|  |
| --- |
|  |

**Title of the Project:**

Bengaluru House Price Prediction

**Problem Statement**

The Bengaluru real estate market is highly dynamic, with property prices influenced by multiple factors such as location, built-up area, number of bedrooms, age of the property, and access to public amenities. This project aims to develop a predictive model using regression techniques to estimate housing prices based on these variables. Through data preprocessing, analysis, and model building, the goal is to identify key pricing drivers and trends. The final insights will be visualized using Power BI to help buyers and developers make informed decisions.

**Scope**

The scope of this project includes developing a machine learning-based regression model to predict housing prices in Bengaluru using historical housing data. The dataset comprises features such as property location, built-up area, number of bedrooms, property age, and proximity to public amenities. The project involves data cleaning, exploratory data analysis (EDA), model building, evaluation, and visualization of insights through interactive Power BI dashboards. The insights aim to support informed decision-making for homebuyers, real estate developers, and investors.

**Objectives**

* To clean, preprocess, and explore housing data from Bengaluru for meaningful insights.
* To identify key factors that influence property prices in the city.
* To develop and evaluate multiple regression models (e.g., Linear Regression, Decision Tree Regressor) for accurate price prediction.
* To visualize neighborhood-wise pricing trends and major pricing influencers using Power BI.
* To assist buyers and developers in understanding market dynamics and making informed decisions.

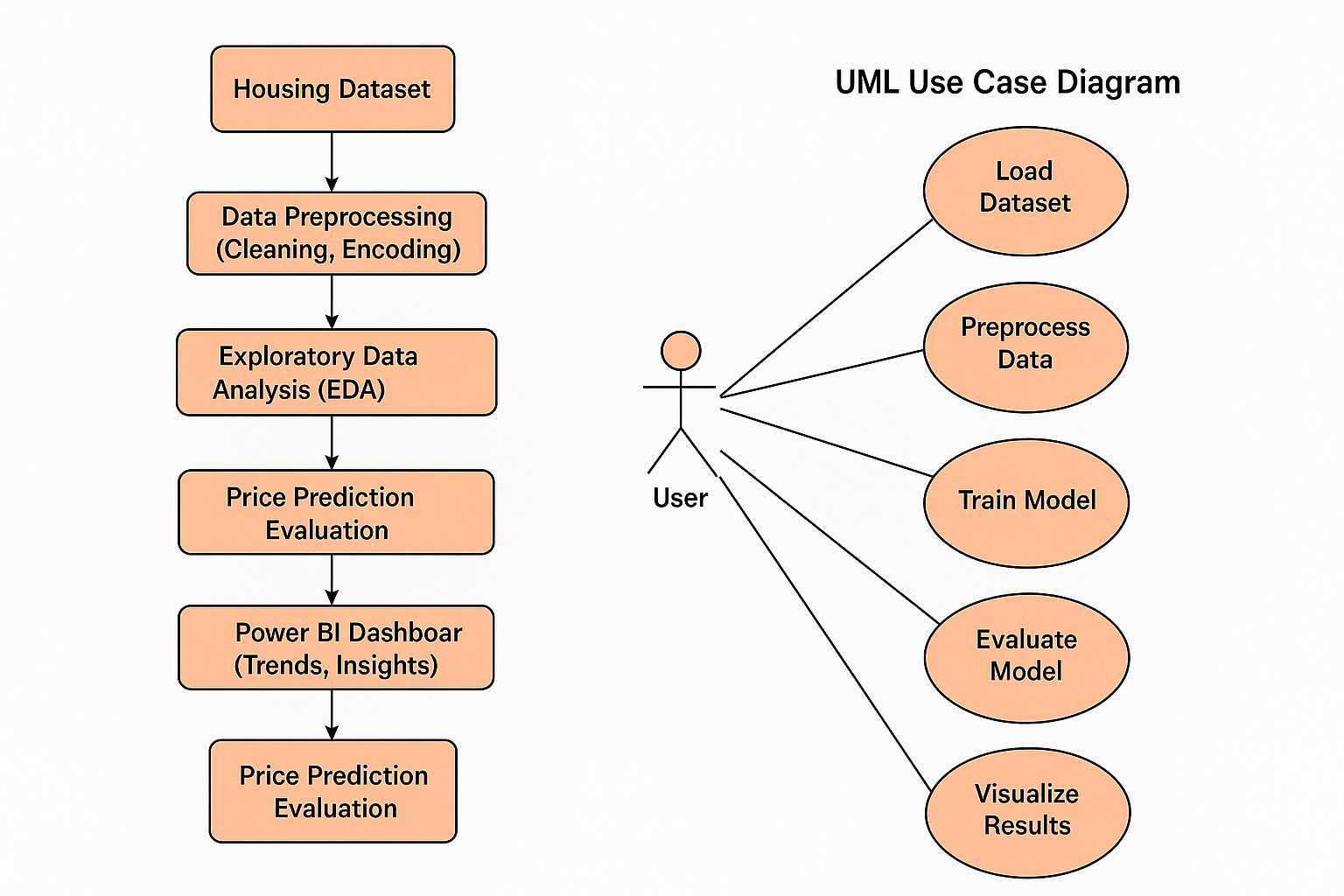
**System Requirements**

**1. Hardware Requirements:**

* Processor: Intel Core i5 or higher
* RAM: Minimum 8 GB (16 GB recommended for faster processing)
* Storage: At least 10 GB of free disk space
* Graphics: Basic integrated graphics (for Power BI and data visualization)

**2. Software Requirements:**

* Operating System: Windows 10/11, macOS, or Linux
* Programming Language: Python (3.7 or above)
* Python Libraries:
  + pandas – for data manipulation
  + numpy – for numerical computations
  + matplotlib, seaborn – for data visualization
  + scikit-learn – for model building and evaluation
  + statsmodels – for statistical analysis (optional)
* **IDE/Notebook:**
  + Jupyter Notebook or VS Code (with Python support)
* **Data Visualization Tool:**
  + Microsoft Power BI (Desktop version for development)
* **Data flow / UML Diagrams**



**Tools Used**

To develop and analyze the Bengaluru House Price Prediction model, a combination of programming tools, libraries, and visualization platforms were used:

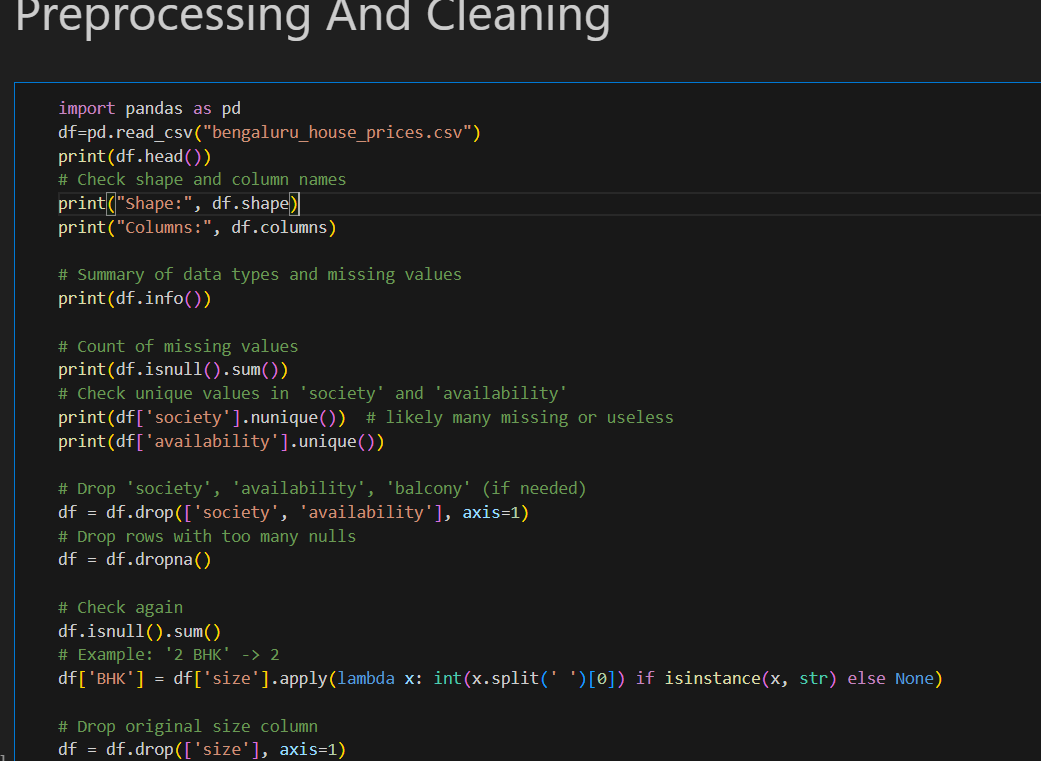
* **Python:** The primary programming language used for data preprocessing, analysis, and model building due to its vast ecosystem and ease of use in data science tasks.
* **Jupyter Notebook / VS Code**: Interactive environments used for developing and testing code in an organized, modular format.
* **Pandas and NumPy**: Essential Python libraries for data manipulation, transformation, and numerical operations**.**
* **Matplotlib and Seaborn:** Visualization libraries used for Exploratory Data Analysis (EDA) to understand patterns, trends, and relationships between variables.
* **Scikit-learn:** Machine learning library used for implementing regression models, model training, and evaluation using metrics like RMSE and R² score**.**
* **Power BI**: A powerful data visualization and business intelligence tool used to build interactive dashboards showcasing neighborhood pricing trends, model performance, and key pricing influencers.

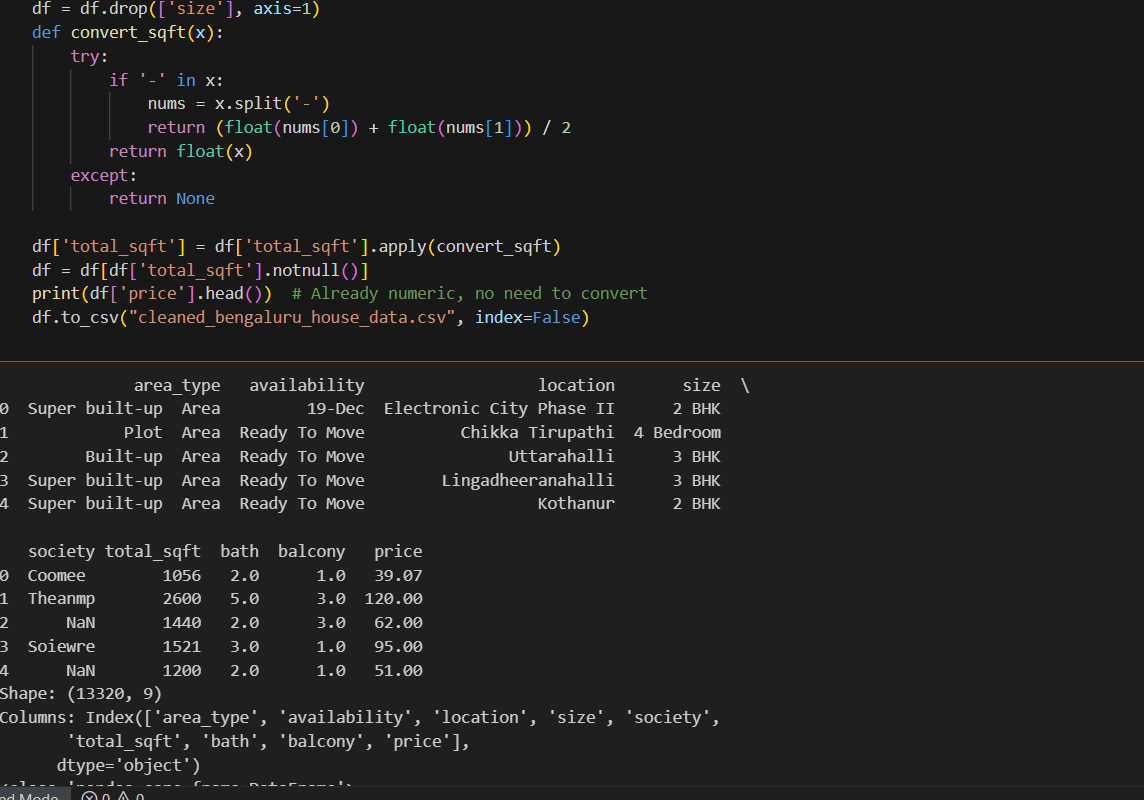
**Methodology**

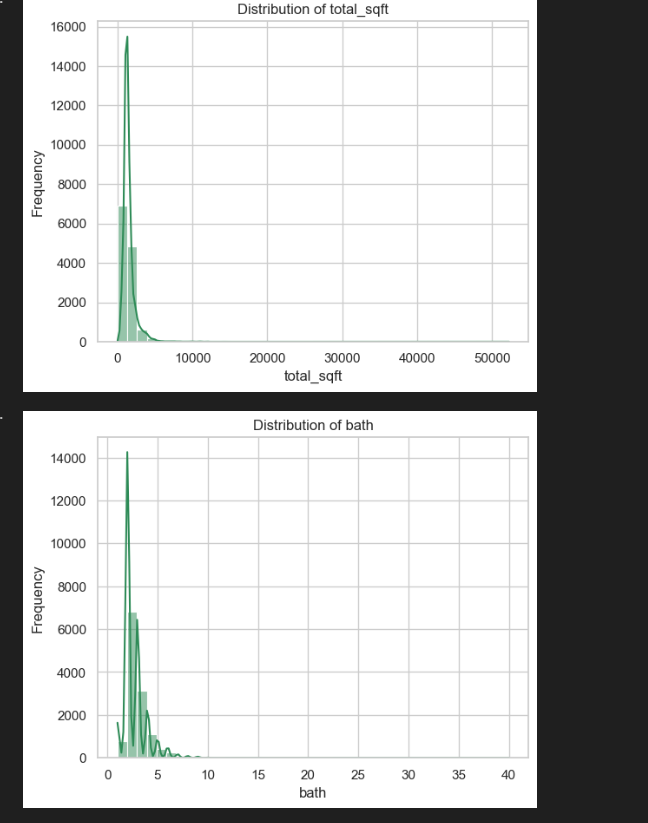
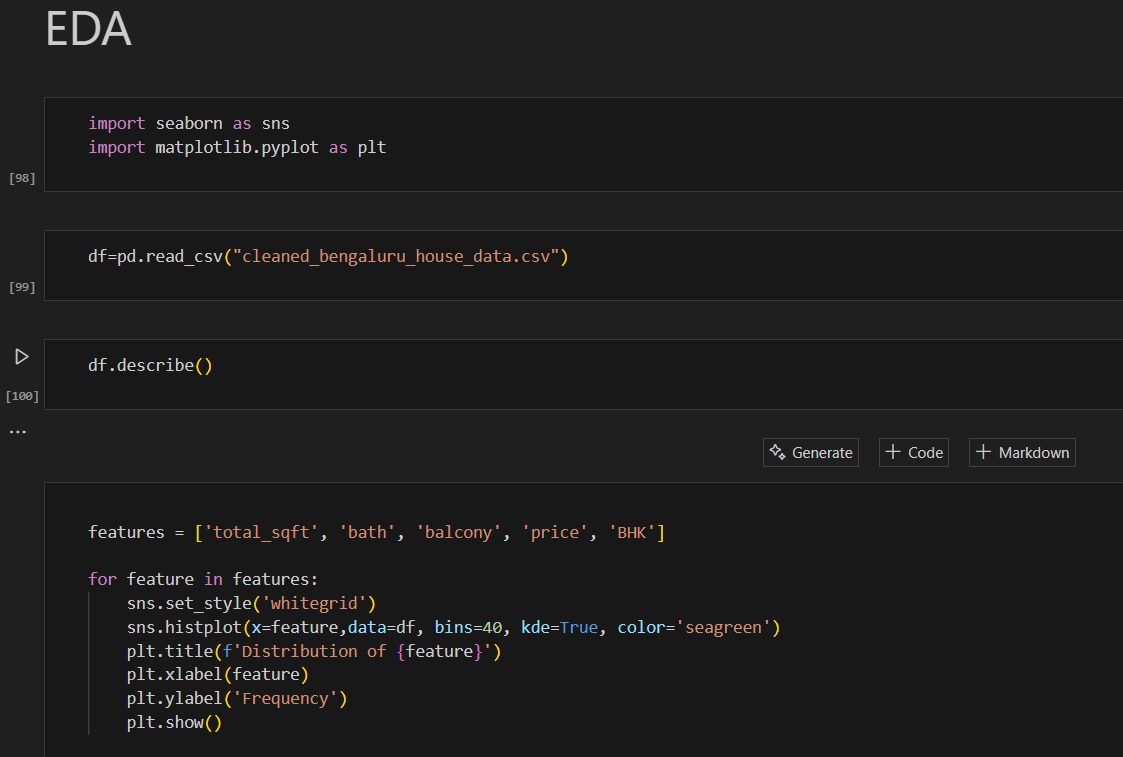
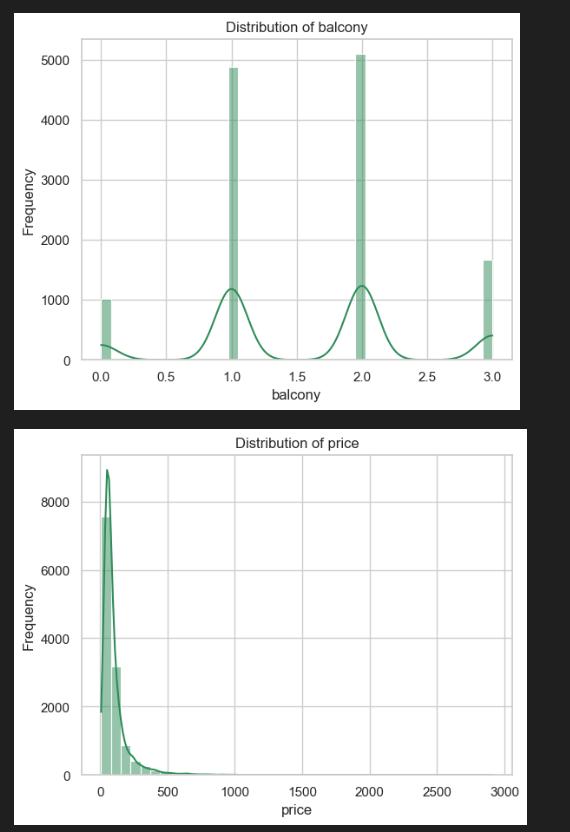
The project follows a structured and systematic approach to build an accurate predictive model for estimating housing prices in Bengaluru:

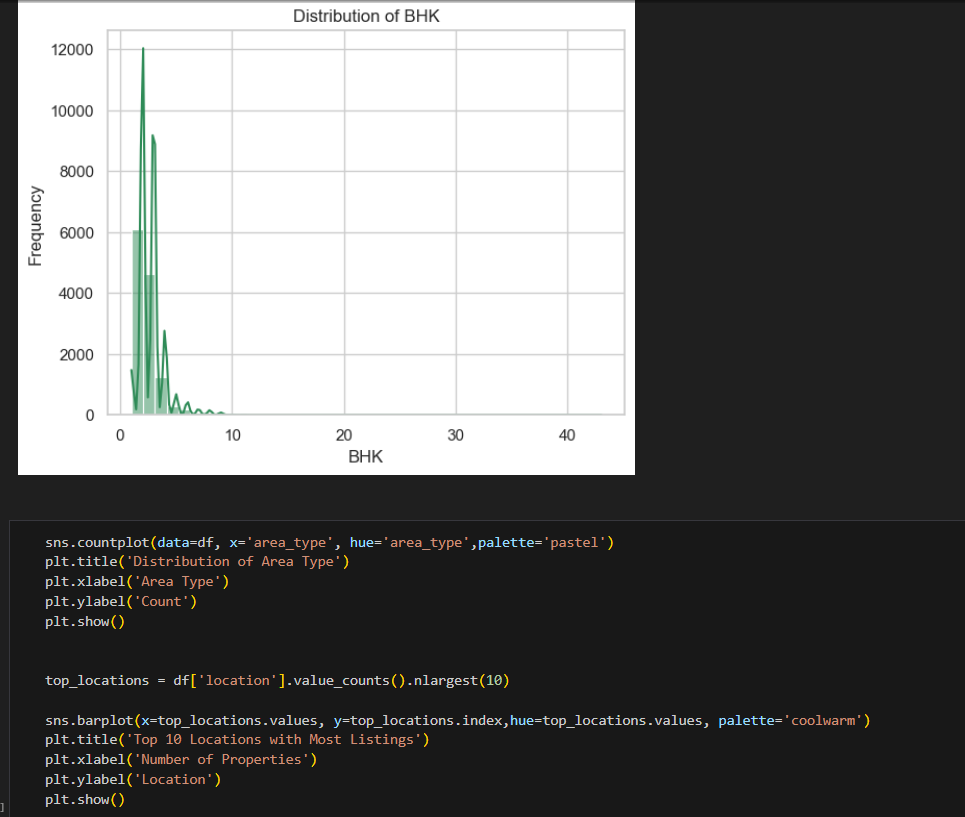
1. **Data Collection**Housing data was gathered, including features such as location, built-up area, number of bedrooms, age of the property, and access to public amenities.
2. **Data Preprocessing**
   * Missing and inconsistent values were identified and appropriately handled.
   * Categorical variables (e.g., location) were encoded using techniques like one-hot encoding.
   * Outliers and duplicate entries were removed to improve model performance.
   * **Numerical features were scaled if required.**
3. **Exploratory Data Analysis (EDA)**
   * Visualizations were created to explore feature distributions, correlations, and data trends.
   * Relationships between price and factors such as area, number of bedrooms, and location were analyzed.
   * Neighborhood-level price variations were studied.
4. **Model Building**Multiple regression algorithms were implemented to estimate house prices:
   * Linear Regression
   * Decision Tree Regressor
   * Random Forest Regressor
   * Lasso and Ridge Regression  
     Models were trained using an 80/20 train-test split and tuned for best performance.
5. **Model Evaluation**Models were evaluated using standard performance metrics:
   * R² Score: Measures how well the model explains variance in price.
   * Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE): Indicate prediction error.
6. **Data Visualization and Reporting (Power BI)**
   * Interactive dashboards were developed to visualize:
     + Price trends by location
     + Key pricing influencers
     + Distribution of high- and low-priced properties
     + Comparison of actual vs. predicted prices  
       These dashboards enable intuitive insights for homebuyers and developers.
7. **Conclusion and Insights**The final model results were interpreted to identify the most significant factors affecting housing prices in Bengaluru. These insights support better real estate decision-making and investment planning.

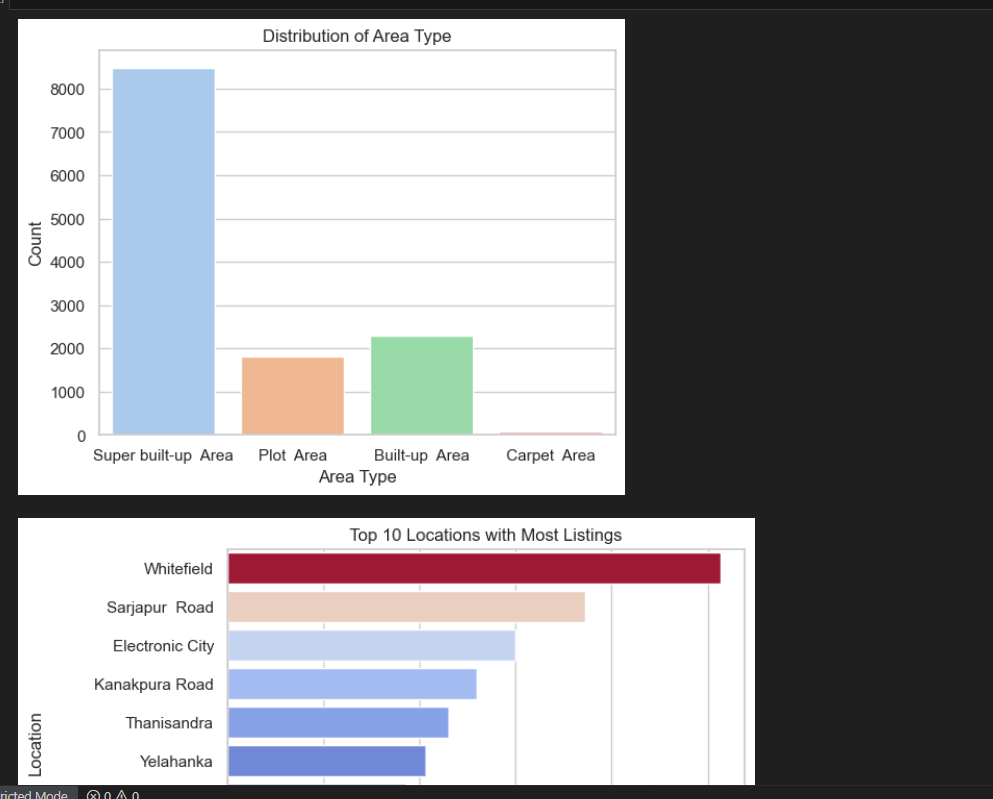
**Screenshots**

****

****

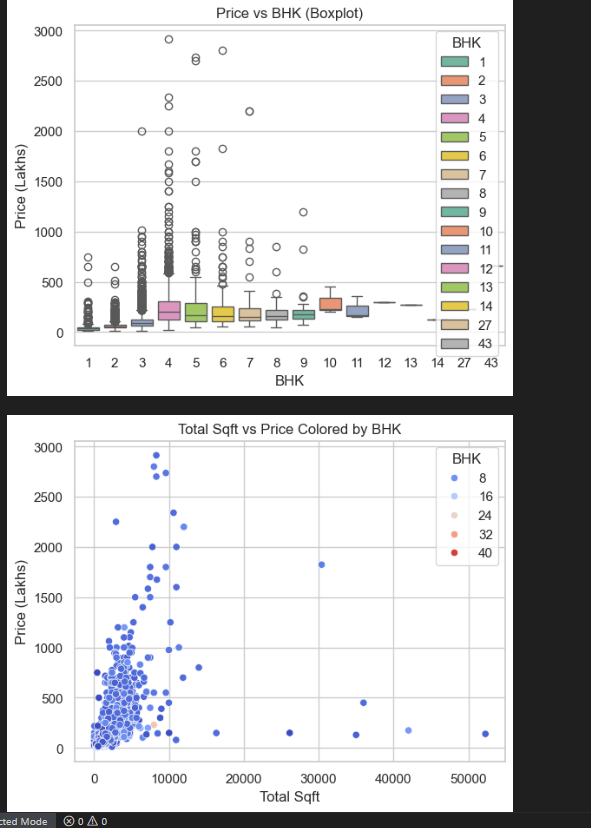
**** ****

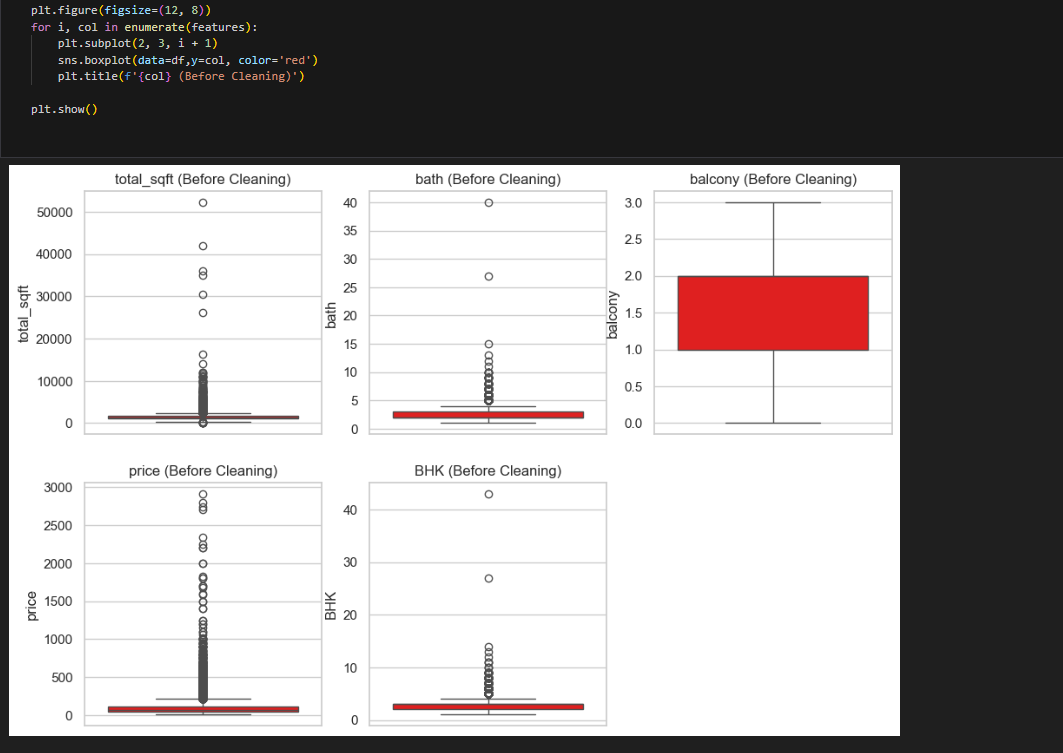
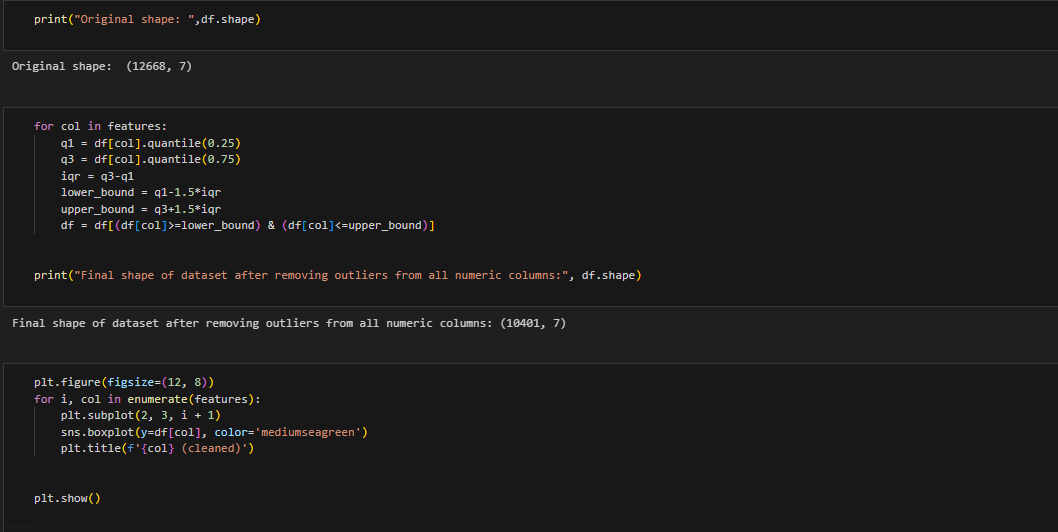
****

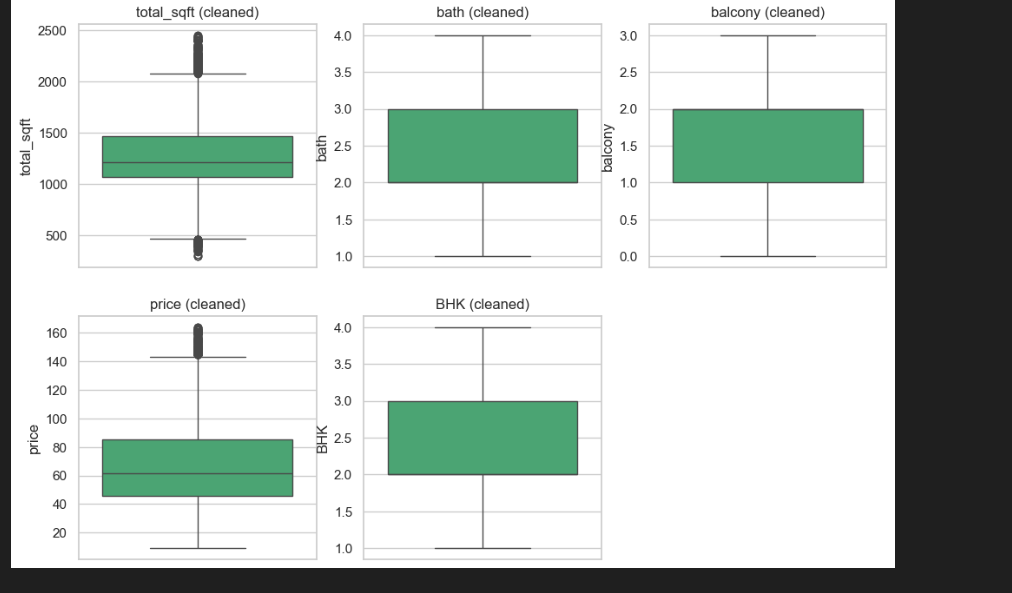
****

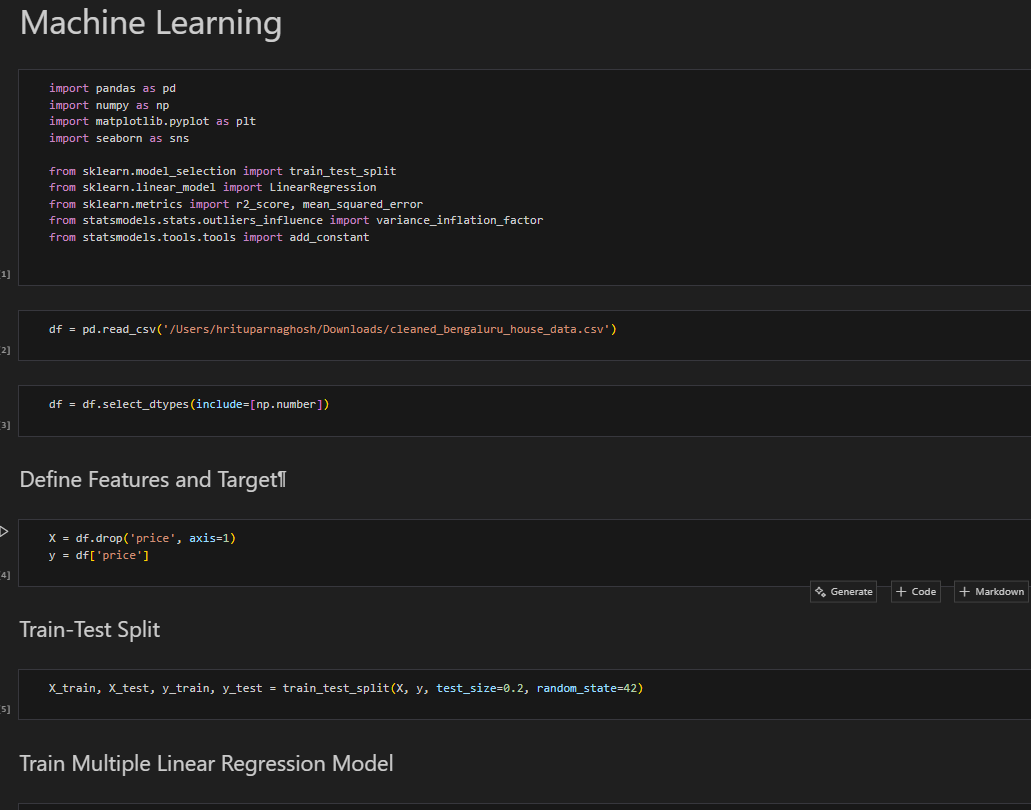
****

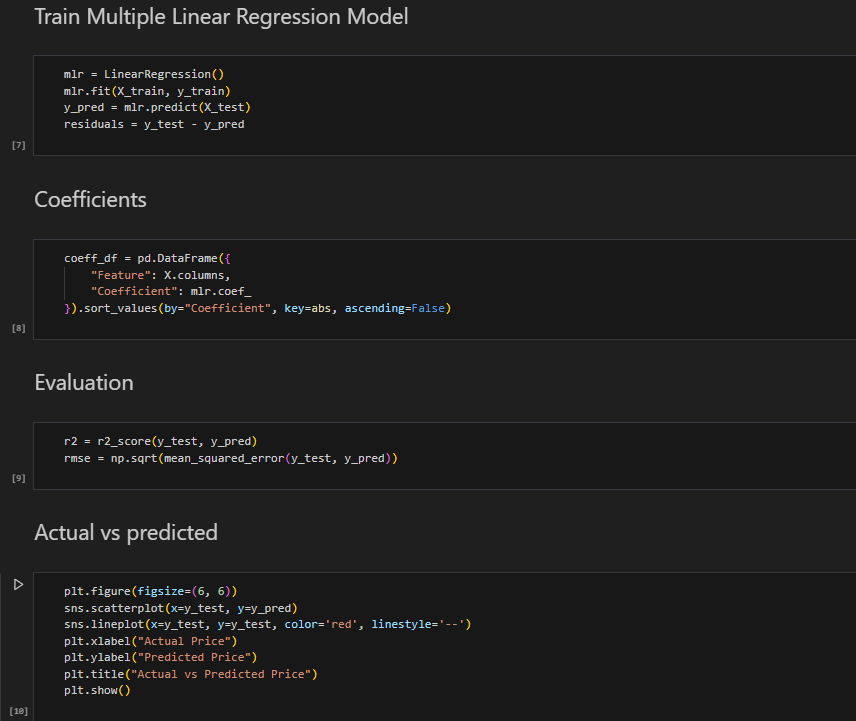
****

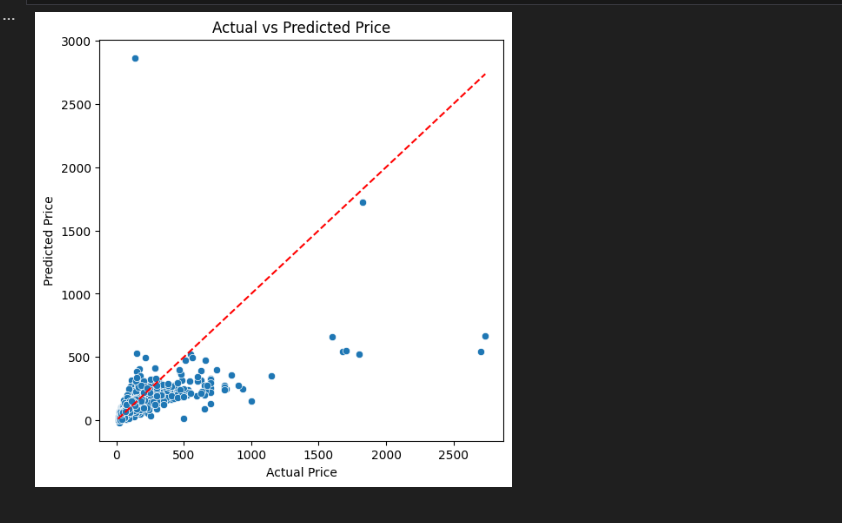
****

**** ****

****

****

****

****

**Results and Discussion**

**output**

The project successfully converted a raw, inconsistent housing dataset into a clean and structured form using Python. Through exploratory data analysis (EDA), key patterns were uncovered—such as the dominance of 2BHK and 3BHK properties, variations in price per square foot, and significant outliers in both pricing and area. Visualizations created using Seaborn and Matplotlib helped in understanding the distribution of total square footage, BHK-wise pricing, and overall market trends.

The cleaned dataset was then used to build an interactive Power BI dashboard. This dashboard includes:

KPI cards showing average price, total listings, and average BHK

Bar charts comparing prices by BHK

Area-wise pricing comparisons

Filters and slicers for dynamic data exploration

The combination of Python-based backend analysis and Power BI’s frontend visualization resulted in a comprehensive and accessible platform for exploring Bengaluru’s real estate trends.

**Challenges Faced**

One of the major challenges was the inconsistency and incompleteness of the dataset. Columns such as total\_sqft included values in different formats—single numbers, ranges, and even text—which required custom functions for conversion. Similarly, the size column had to be processed to extract meaningful numerical values like the BHK count.

Another challenge involved ensuring that visualizations were meaningful and not skewed by outliers, which were common in both pricing and property size. Careful filtering and use of appropriate plots helped mitigate this.

Finally, while integrating the cleaned dataset into Power BI, some formatting adjustments were necessary to ensure data compatibility and smooth visualization rendering.

**Learnings**

This project provided a hands-on understanding of the end-to-end data analysis pipeline—from loading raw data, cleaning and transforming it using Python, performing EDA, and finally, presenting the results in an interactive Power BI dashboard.

**Learnings include:**

Practical application of data cleaning and feature engineering techniques using Python

Mastery of Seaborn and Matplotlib for trend analysis and visual storytelling

Designing a user-friendly Power BI dashboard to make insights accessible to non-technical users

Understanding how structured analysis can support data-driven decisions in the real estate domain

Overall, the project demonstrated the importance of combining coding and visualization tools to turn complex datasets into clear, actionable insights.

**Conclusion**

This project successfully demonstrated how raw and inconsistent real estate data can be transformed into meaningful insights through the combined use of Python and Power BI. Starting with a dataset containing missing values, non-standard formats, and redundant fields, the project applied effective data cleaning and feature engineering techniques to prepare the data for analysis. Exploratory Data Analysis (EDA) using Python libraries like Pandas, Seaborn, and Matplotlib helped uncover key patterns in property prices, BHK distribution, and area-wise trends across Bengaluru.

The development of an interactive Power BI dashboard further enhanced the accessibility of these insights. Users can now explore average prices, compare BHK configurations, and filter by locality or size—all within a visually engaging and user-friendly interface. This dashboard serves as a practical tool for prospective homebuyers, investors, and analysts looking to make informed decisions in Bengaluru’s dynamic housing market.

Through this project, a strong understanding of the real estate domain was combined with technical skills in data science and business intelligence. The successful integration of backend analysis with frontend visualization proves the value of data-driven decision-making in solving real-world problems. Ultimately, the project showcases how structured analysis and interactive dashboards can simplify complex data and empower users with actionable insights.