**INTERNSHIP REPORT**

**Submitted By**

**Bhargav Bhatt**

**(*226090307004*)**

**To**

**COMPUTER ENGINEERING DEPARTMENT**

**C U SHAH (GOVT.) POLYTECHNIC – SURENDRANAGAR**

****

**GUJARAT TECHNOLOGICAL UNIVERSITY – AHMEDABAD**

**JUNE-AUGUST – 2024**

**INDEX**

|  |  |
| --- | --- |
| Sr No | Content |
| 1 | Student Registration Form |
| 2 | Attendance Sheet |
| 3 | Daily Log with Topic Learned (Day or Week wise) |
| 4 | Feedback from Industry |
| 5 | Completion Certificate from Industry |
| 6 | Certificate from Institute |
| 7 | Project Profile |
| 8 | Project Introduction |
| 9 | Project advantages and limitations |
| 10 | Project applications |
| 11 | Technical Details |
| 12 | System Requirements |
| 13 | Functional and Non-Functional requirements of project |
| 14 | List of Users and Use case Diagrams for each user of project |
| 15 | Data Dictionary of project |
| 16 | Sample coding of project |
| 17 | Screen shot of various output pages of project |
| 18 | Testing strategies of project |
| 19 | Conclusion and Future work |
| 20 | References |
| 21 | PPTs of Project presentation |
| 22 | Screen shot of Project Poster |
| 23 | Source code, Database, PPTs, Final report of Project in DVD(s) |

**Project Profile**

* Project Title :- House Price Pridiction
* Goal of Project :- Provide robust analytical framework that utilizes machine learning techniques to understand and predict residential property prices. By examining a comprehensive dataset containing property characteristics
* Project Guide : Mr. S. S. Parmar Sir
* Platform
* Front End : StreamLit (python)
* Backend : Python
* Data Dictionary : CSV file
* Documentation Tool : Microsoft Office Word

**Project Introduction**

The project focuses on analyzing residential property prices to gain insights into the factors influencing the housing market. By examining various variables such as property characteristics, location, market trends, and economic indicators, this project aims to provide a comprehensive understanding of the factors driving home prices and identify key trends and patterns. The project utilizes a dataset containing information on residential properties, including features such as square footage, number of bedrooms and bathrooms, location attributes (e.g., neighborhood, proximity to amenities), and corresponding sale prices. By analyzing these variables, the project seeks to uncover correlations and trends that can help stakeholders understand the dynamics of the housing market and make informed decisions regarding real estate investments, pricing strategies, and policy development. Through exploratory data analysis and statistical modeling techniques, the project aims to identify the key factors influencing home prices. This includes assessing the impact of property characteristics (e.g., size, amenities), location factors (e.g., proximity to schools, transportation), and market trends (e.g., supply and demand, interest rates) on property values. Additionally, the project explores the relationships between economic indicators (e.g., GDP growth, employment rates) and home prices to understand the broader macroeconomic influences on the housing market. The outcomes of this project have implications for real estate professionals, investors, policymakers, and prospective homeowners. Real estate professionals can utilize the findings to better understand market trends, advise clients, and develop effective pricing strategies. Investors can gain insights into factors that drive property value appreciation and identify lucrative investment opportunities. Policymakers can use the findings to inform housing policies, zoning regulations, and urban planning initiatives. Prospective homeowners can make more informed decisions regarding property purchases, taking into account factors that influence home prices and long-term value appreciation

**Project Advantages And Limitation’s**

* **Advantages** :-

1. Informed Decision-Making : Provides real estate professionals, investors, and policymakers with actionable insights into factors that drive property values, aiding in more informed decision-making.
2. Market Understanding : Enhances understanding of market dynamics by analyzing trends and correlations between property characteristics, location, and economic indicators.
3. Investment Opportunities : Helps investors identify lucrative opportunities by highlighting which factors most significantly impact property appreciation and market trends
4. Pricing Strategies : Offers real estate professionals valuable insights for setting competitive and accurate pricing strategies based on market data.
5. Comprehensive Analysis : Utilizes a broad range of variables, including property features, location attributes, and economic indicators, for a well-rounded analysis of housing market factors.

* **Limitation’s** :-

1. Data Quality and Availability : The accuracy of insights depends on the quality and completeness of the dataset. Incomplete or inaccurate data can lead to misleading conclusions.
2. Dynamic Market Conditions : Housing markets are influenced by many external factors (e.g., economic downturns, natural disasters) that may not be fully captured in the dataset or analysis.
3. Model Limitations : Predictive models may not capture all nuances of the housing market and could be influenced by unobserved variables or changes in market conditions.
4. External Factors : Factors such as changes in government policy, shifts in societal preferences, or unforeseen events (e.g., pandemics) can impact the housing market in ways not accounted for in the analysis.

**Project Applications**

* **Real Estate Investment:**
* **Investment Decisions:** Investors can use insights from the analysis to identify areas with high growth potential and make informed investment decisions.
* **Risk Assessment:** Analyze factors that could pose risks to property value, helping investors mitigate potential losses .
* **Real Estate Marketing and Sales :**
* **Targeted Marketing:** Real estate agents can tailor marketing strategies based on the factors most appealing to buyers in different segments .
* **Pricing Strategies:** Agents can use the analysis to set competitive and accurate prices, enhancing sales effectiveness .
* **Homebuyer Guidance :**
* **Buying Decisions:** Prospective homeowners can understand factors that influence property values, aiding in selecting homes with better long-term value .
* **Financial Institutions:**
* **Mortgage Lending:** Banks and financial institutions can use the analysis to assess the risk and potential return on mortgage loans.
* **Property Valuation:** Accurate property valuations based on analysis can enhance loan underwriting processes .
* **Real Estate Technology Solutions:**
* **Property Platforms:** Online real estate platforms can integrate analysis to provide users with detailed insights into property values and market conditions.
* **Automated Valuation Models (AVMs):** Enhance AVMs with data-driven insights for more accurate property valuations.

**Technical Details**

* **Frontend Technologies :-**

1. **Streamlit :**

* Streamlit is employed to build the user interface, offering a seamless and interactive experience for users. It facilitates the integration of data visualization and user input directly from Python scripts, making it ideal for developing interactive data applications.

1. **Visualization Libraries (Matplotlib, Seaborn):**

* These libraries are essential for creating visual representations of the data, allowing users to easily interpret complex datasets through a variety of charts and graphs. They enhance the project's ability to present insights effectively.
* **Backend Technologies :-**

1. **Python:**

* Python serves as the core programming language for the project, providing a versatile platform for data processing, analysis, and model development. Its extensive ecosystem of libraries supports efficient data manipulation and machine learning.

1. **Scikit-learn:**

* Scikit-learn is used for building and evaluating machine learning models that predict property prices based on various features. It offers a range of algorithms and tools to support the development of robust predictive models.

1. **Data Handling Libraries (Pandas, NumPy):**

* Pandas and NumPy are used for managing and processing data efficiently. They enable complex data manipulations and numerical computations, which are crucial for analyzing large datasets.

**System Requirements**

* **Hardware Requirements :-**
* **Processor:** Intel Core i5 or equivalent
* A mid-range processor is recommended to handle data processing and model computations efficiently.
* **RAM:** 8 GB or more
* Sufficient RAM is needed to handle large datasets and ensure smooth operation during data analysis and model training
* **Storage:** 256 GB SSD or larger
* An SSD provides fast read and write speeds, which are beneficial for data loading and processing tasks.
* **Software Requirement:-**
* **Operating System:** Windows 10, macOS, or Linux
* The application is cross-platform and can be run on any major operating system that supports Python.
* **Python:** Version 3.7 or later
* The project is developed in Python, so a compatible Python version is required.
* **Web Browser:**
* A modern web browser such as Google Chrome, Mozilla Firefox, or Microsoft Edge is required to access the Streamlit application interface.
* **Jupyter Notebook:**
* Required for data visualization and analysis, enabling users to explore data, visualize results, and run models interactively.

**Functional requirements**

* **Data Input and Processing:**
* The system shall allow users to upload datasets containing residential property information for analysis.
* The system shall clean and preprocess the input data, handling missing values and outliers.
* **Data Analysis and Modeling:**
* The system shall perform exploratory data analysis (EDA) to identify trends and patterns in the data.
* The system shall allow users to select features for building machine learning models.
* The system shall train and evaluate predictive models to estimate property prices.
* **Data Visualization:**
* The system shall generate visualizations such as scatter plots, histograms, and heatmaps to represent data relationships.
* The system shall update visualizations dynamically based on user input and data changes.
* **User Interface:**
* The system shall provide an interactive web-based interface for users to interact with the data and models.
* The system shall include components such as sliders, buttons, and input fields to capture user input.

**Non-Functional Requirements**

* **Performance:**
* The system shall process and analyze datasets within a reasonable time frame to provide timely insights.
* The system shall support concurrent users without significant degradation in performance.
* **Scalability:**
* The system shall be capable of handling increasing amounts of data and users without significant performance loss.
* **Reliability:**
* The system shall ensure data integrity and accuracy throughout the analysis process.
* The system shall handle errors and exceptions gracefully, providing meaningful error messages to users.
* **Portability:**
* The system shall be compatible with major operating systems (Windows, macOS, Linux) and web browsers.

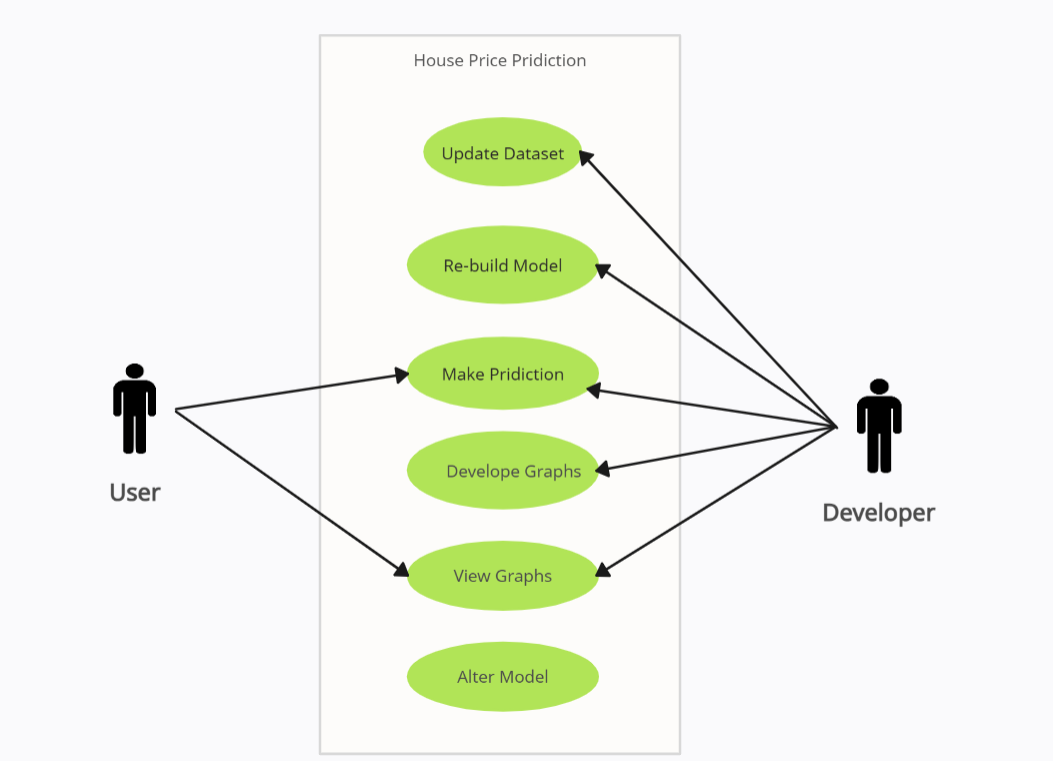
**Users of System**

* **Real Estate Professionals**
* Real estate agents, brokers, and appraisers who use the system to analyze market trends and set pricing strategies.
* **Investors**
* Individuals or organizations interested in using data-driven insights to make informed investment decisions.
* **Homebuyers**
* Prospective homebuyers seeking to understand market dynamics and identify favorable properties.
* **Data Analysts/Researchers**
* Professionals and researchers conducting market studies or academic research on housing trends.
* **System Administrators**
* Personnel responsible for maintaining and managing the application, ensuring its smooth operation.

**Use Case Diagram**

* **User** use the UI to input data into the model. For analysis and visualization, they navigate through Jupyter Notebook, where they can explore the data, run models, and visualize results.
* **Developer** manage access to both the Streamlit UI and Jupyter Notebook, ensuring that users can perform their tasks without technical hindrances. They are also responsible for maintaining the system's reliability and security.

Use Case Diagram of House Price Pridiction Project

****

**Data Dictionary**

* The data dictionary provides a detailed description of each column in the CSV dataset used for analyzing residential property prices. This dataset includes various attributes of residential properties that are used to explore trends and patterns in the housing market.

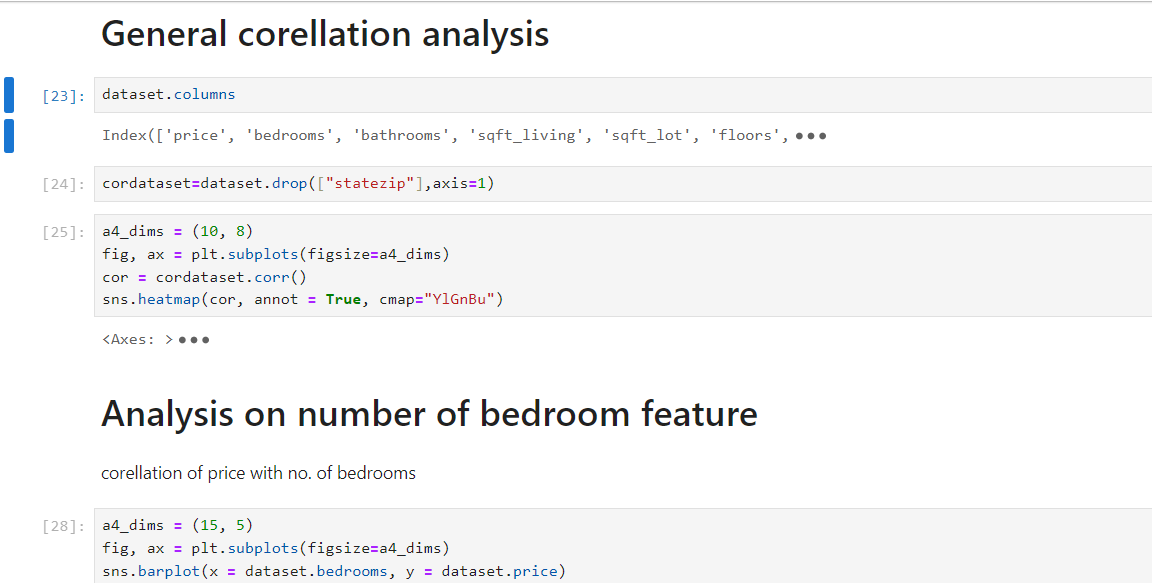
|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| date | Date | The date associated with the property listing or sale. |
| price | Float | The sale price of the property. |
| bedrooms | Integer | The number of bedrooms in the property. |
| bathrooms | Float | The number of bathrooms in the property, including partial bathrooms. |
| sqft\_living | Integer | The total living area of the property in square feet. |
| sqft\_lot | Integer | The total area of the lot or land on which the property is situated, measured in square feet. |
| floors | Float | The number of floors in the property. |
| waterfront | Integer | A binary indicator (0 or 1) representing whether the property has a waterfront view. |
| view | Integer | A rating of the property's view, with higher values indicating a better view. |
| condition | Integer | A rating of the overall condition of the property, with higher values indicating better condition. |
| sqft\_above | Integer | The square footage of the property that is above ground level. |
| sqft\_basement | Integer | The square footage of the property's basement, if applicable. |
| yr\_built | Integer | The year the property was originally built. |
| yr\_renovated | Integer | The year the property was last renovated, if applicable. |
| street | String | The street address or location of the property. |
| city | String | The city where the property is located. |
| statezip | String | The state and ZIP code of the property. |
| country | String | The country where the property is located. |

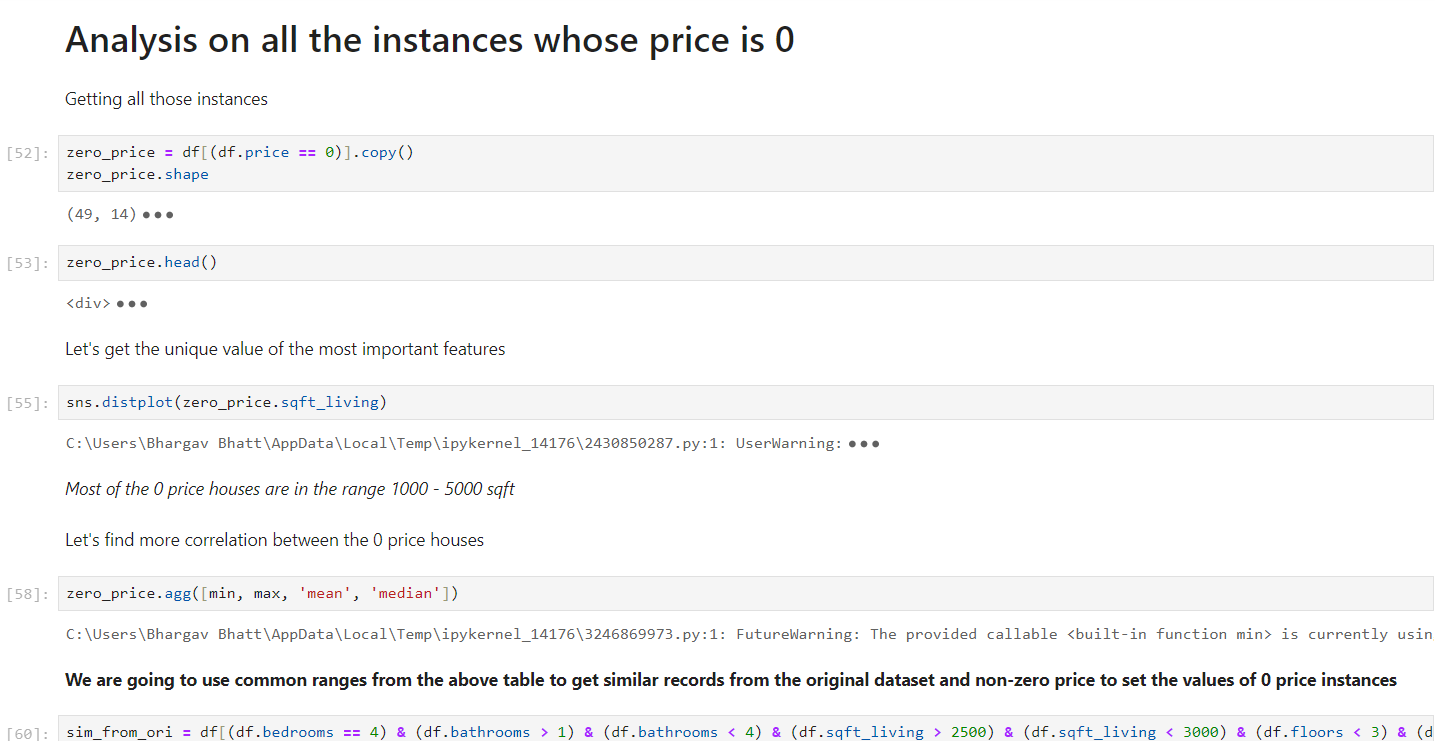
* **Sample Data :-**
* **What is CSV?**
* CSV stands for **Comma-Separated Values**. It's a simple file format used to store tabular data, such as a spreadsheet or database, in plain text. Each line in a CSV file corresponds to a row in the table, and each value in the row is separated by a comma . here the above data is of csv file

**Sample Coding**

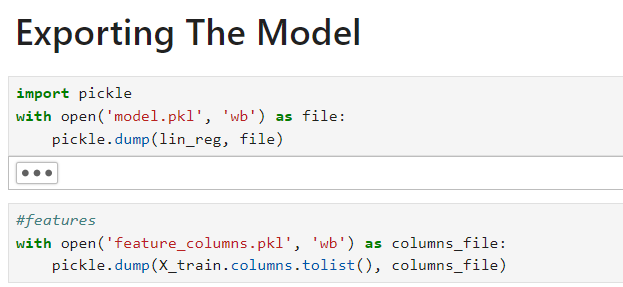
* Loading Dataset :-



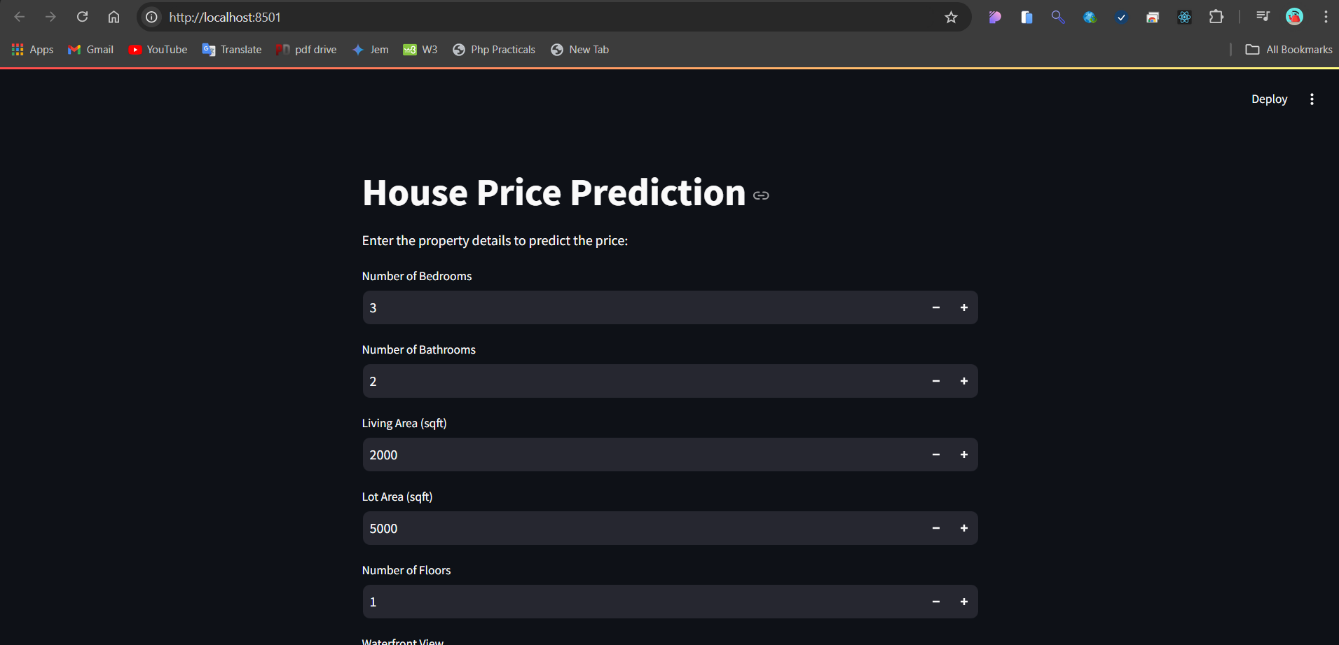
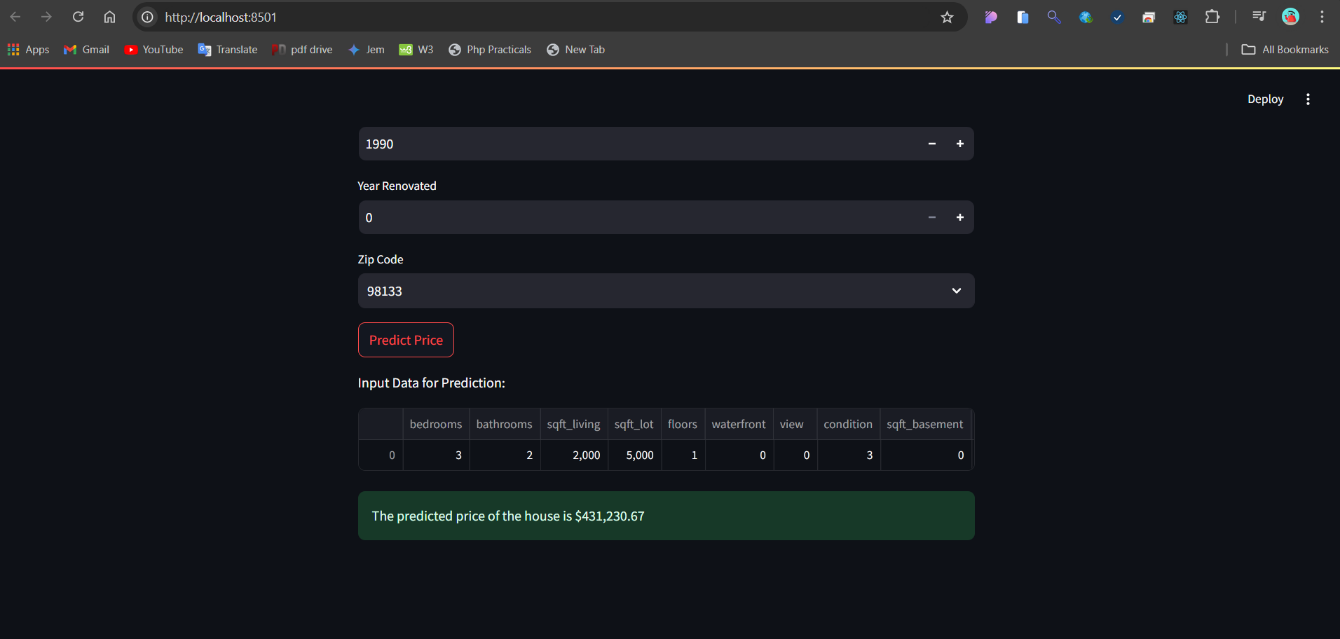
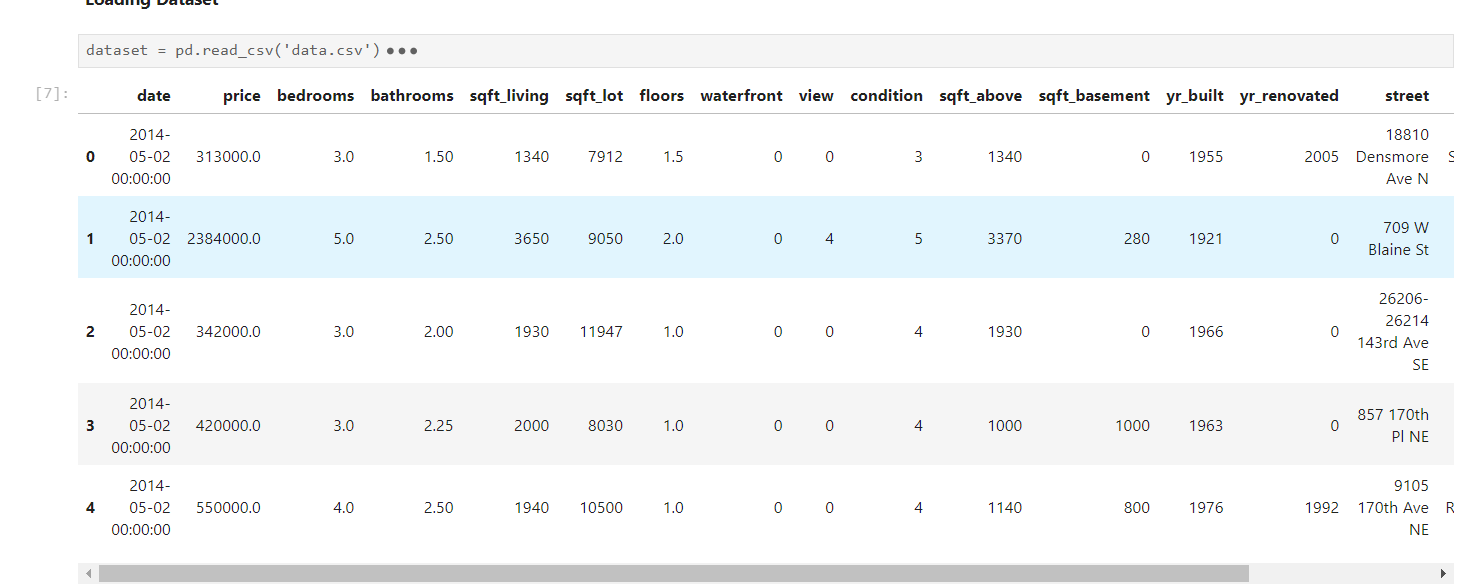
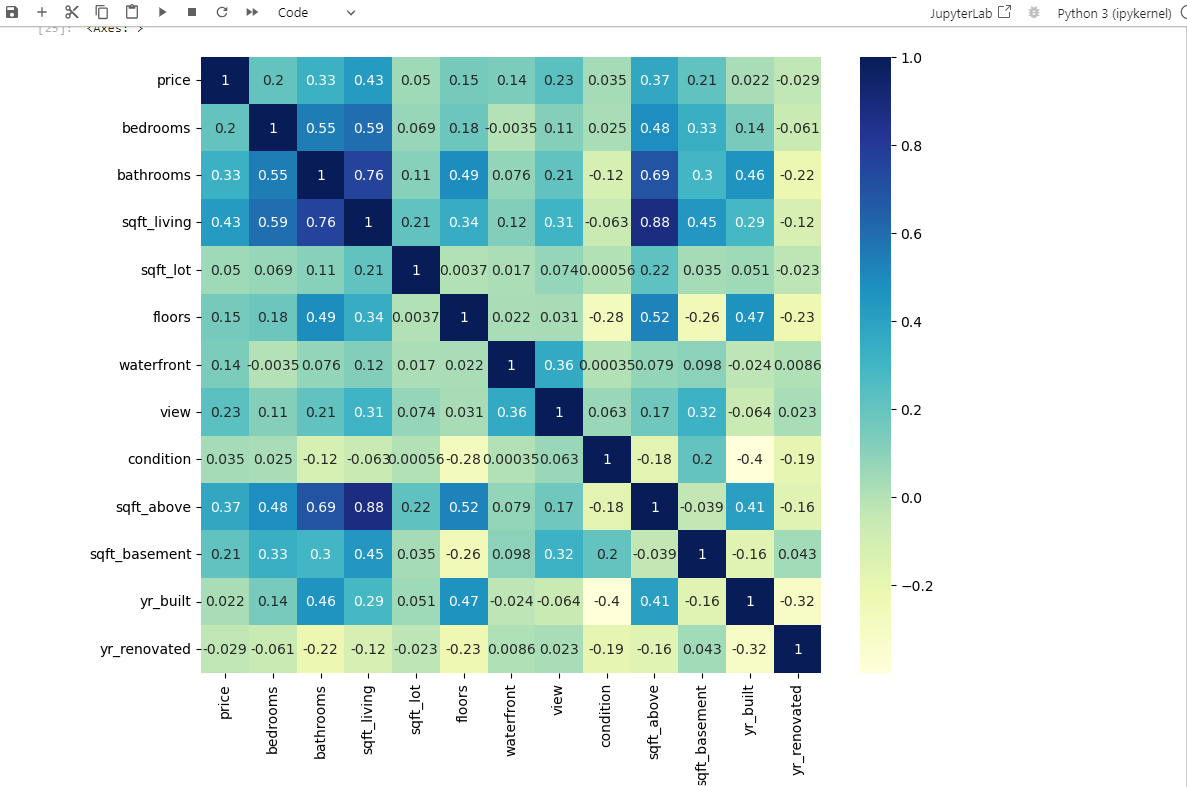
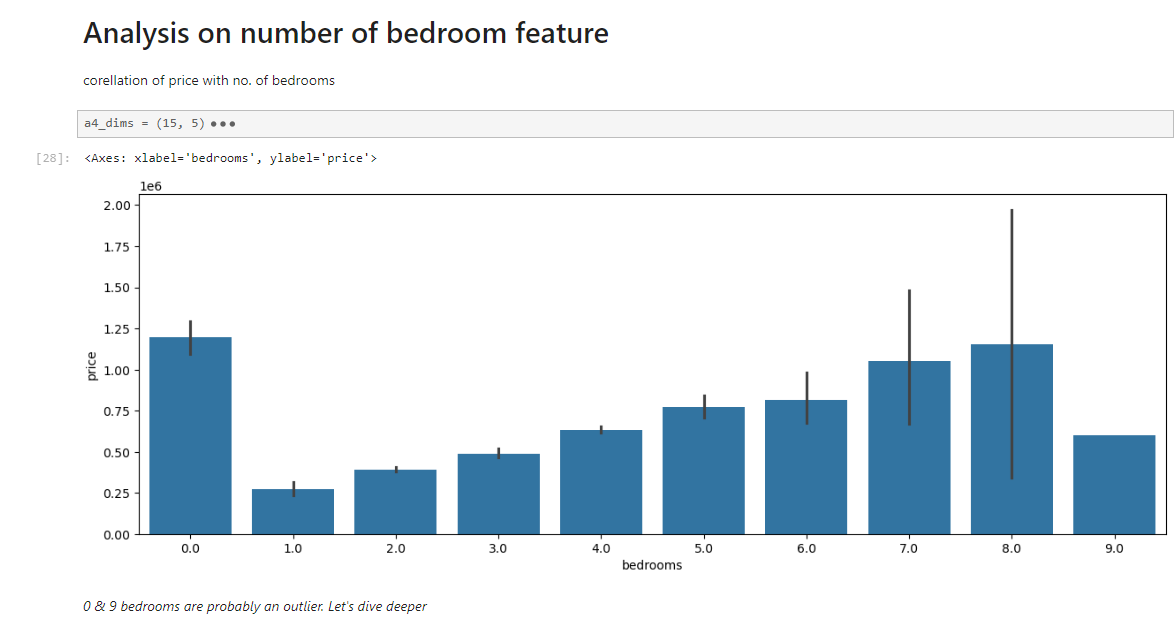
* General Correlation Analysis
* Analysis on Zero Price data

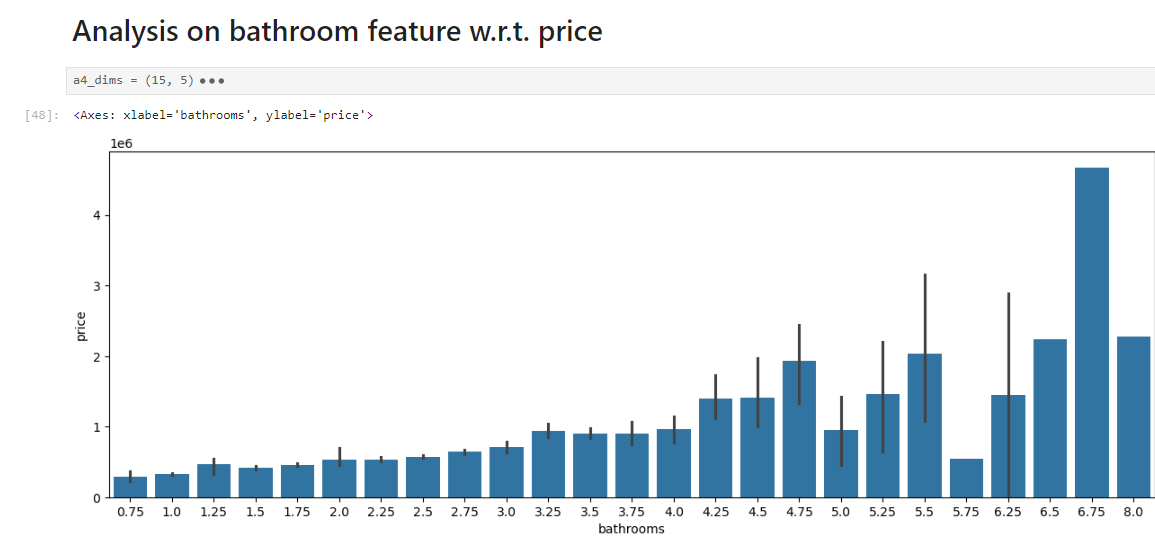


* Building The Model :- Linear Regression
* Exporting the Model :-



**Outputs**

* **Feature Entry** :-
* **Pridiction :-**
* **Dataset :-**
* **General Corelation :-**
* **Analysis on Number Of bedrooms w.r.t Price :-**
* **Analysis on Number Of Bathrooms w.r.t Price :-**



**Testing Strategies**

The testing strategies for the residential property price analysis project focus on validating the performance and reliability of the linear regression model. These strategies ensure that the model provides accurate predictions and insights based on the dataset. Below are the key testing strategies employed in the project :

* **Data Splitting :-**
* **Training and Testing Sets:**
  + The dataset is divided into training and testing sets, typically with an 80/20 or 70/30 split. The training set is used to fit the linear regression model, while the testing set evaluates its performance.
* **Validation Set:**
  + Optionally, a separate validation set can be used to fine-tune model parameters and prevent overfitting, especially in more complex models or when additional hyperparameters are involved.
* **Model Evaluation Metrics**
* **Mean Absolute Error (MAE):**
  + Measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as the average over the test sample of the absolute differences between prediction and actual observation.
* **Mean Squared Error (MSE):**
  + Represents the average of the squared differences between predicted and actual values. It penalizes larger errors more than MAE, making it sensitive to outliers.
* **R-squared (R²):**
  + Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R² value indicates a better fit of the model
* **Residual Analysis**
* **Residual Plots:**
  + Plotting the residuals (the differences between actual and predicted values) helps identify patterns or non-random distribution, which can indicate model inadequacies or the presence of outliers.

**Conclusion & Future Work**

* **Conclusion**

The residential property price analysis project successfully developed and implemented a linear regression model to predict housing prices based on various property characteristics and economic indicators. Through detailed exploratory data analysis and rigorous testing, the model has provided valuable insights into the factors influencing property prices, including the impact of location, property size, and market trends.

The integration of the Streamlit UI allowed for seamless interaction with the model, enabling users to input data and receive real-time predictions. This user-friendly interface, combined with robust statistical analysis, has made the project a practical tool for stakeholders such as real estate professionals, investors, and policymakers. By understanding the key drivers of property values, these stakeholders can make informed decisions about investments, pricing strategies, and policy development.

Overall, the project demonstrates the potential of data-driven approaches in real estate analysis, offering a comprehensive understanding of market dynamics and supporting more strategic decision-making.

* **Future Work**

While the project has achieved significant milestones, there are several areas for future exploration and enhancement:

1. **Incorporation of Additional Data Sources :**

* Integrating more diverse datasets, such as crime rates, school quality, and public transportation access, could provide a more holistic view of the factors affecting property prices.

1. **Advanced Modeling Techniques :**

* Exploring more sophisticated machine learning models, such as decision trees, random forests, or neural networks, could improve prediction accuracy and capture non-linear relationships between variables.

1. **Dynamic and Real-Time Data Updates :**

* Implementing a system for real-time data updates and analysis would allow the model to adapt to rapidly changing market conditions and provide more timely insights.

1. **User Interface Enhancements :**

* Further refining the UI to include data visualization features, such as interactive graphs and maps, would enhance user engagement and provide more accessible insights.

**References**

* Websites :-

* <https://Kaggle.com> :- for datasets of project
* <https://Chatgpt.com> :- for documentation’s
* <https://www.geeksforgeeks.org/machine-learning/> :- for learning
* <https://www.ibm.com/topics/machine-learning> :- for learning
* <https://streamlit.io/cloud> :- for UI development
* <https://share.streamlit.io/> :- Deployed Project
* <https://google.com> :- for query’s