**HOUSE PRICE PREDICTION USING MACHINE LEARNING**

|  |  |
| --- | --- |
| **Date** | **31-10-2023** |
| **Team ID** | **935** |
| **Project Name** | **House Price Prediction Using ML** |

**Introduction:**

The task at hand is to develop a machine learning model that can accurately predict house prices based on a set of relevant features. House price prediction is a common problem in the real estate industry and has a wide range of applications, from helping buyers make informed decisions to assisting real estate professionals in setting competitive prices for listings.

**LITERATURE SURVEY**

**1.Predicting Housing Market Trends Using Twitter Data by M Velthorst [ 2019]**

In this comprehensive report, we delve into the intriguing realm of predicting housing market trends utilizing Twitter data. Our project involved the meticulous collection of textual data from Twitter over the course of a month, specifically targeting tweets that discussed upward or forward trends in house prices. The insights derived from this analysis are particularly invaluable for prospective homebuyers.

Our data collection and analysis were conducted in Dutch, and encompassed text mining, machine learning, and an exploration of the average house prices inferred from the Twitter data. We adopted a 'bag of words' approach to process the text data. To provide a holistic perspective, we employed a range of models, including the utilization of Google search index data and the hedonic model, which serves as a sentiment indicator.

**2.House Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia by TD Phan[2018 ]**

This study focuses on predicting housing prices in Melbourne using machine learning techniques. It begins with an introduction emphasizing the importance of accurate price predictions in real estate decisions and policy-making. After exploring related work in the field, the study delves into data preparation and descriptive findings, providing insights into the dataset. Data reduction and model selection techniques are employed to improve prediction accuracy, and several machine learning models are tested, including Support Vector Machine (SVM) with both Stepwise and PCA-selected features. The results reveal that Polynomial Regression and Regression Trees perform well, while SVM shows promise, particularly with PCA-selected features. Model performance, runtime, and interpretability are discussed, leading to recommendations for further investigation and potential improvements in housing price predictions for Australian cities.

**3. Machine Learning based Predicting House Prices using Regression Techniques by J Manasa[2020 ]**

"In this report, we present a concise overview of our journal on house price prediction in an urban area of Bengaluru using regression techniques. The dataset used for this analysis was sourced from a machine learning hackathon platform. Our analysis encompasses the application of various regression models, including Ridge Regression, Support Vector Machine (SVM), k-Nearest Neighbor(KNN), and Boosting algorithms.These models were thoroughly examined and evaluated for their performance in predicting house prices in the urban area. After careful analysis, we found that the Ridge and Lasso regression techniques outperformed the others, making them the optimal choices for our predictive model. All of our analytical work was conducted within a Jupyter Notebook environment, providing a transparent and reproducible framework for our research.

**4. House Price Forecasting using Machine Learning[2020] by** **Kuvalekar**

Regression, a powerful machine learning technique, leverages existing data to establish correlations between the target parameter (house price) and various independent variables (e.g., number of rooms, area, location), allowing for reliable property valuation within a given area.

The model encompasses key features including bedrooms, bathrooms, area, floor, property age, zip code, latitude, and longitude. Notably, it also incorporates air quality and crime rate as influential factors, providing a unique edge as these are often omitted in conventional datasets. Higher values of these features correlate with reduced property prices, enriching the predictive accuracy. Implementation of this model is executed in Python, harnessing the Scikit-learn library's Decision Tree Regressor. Further, the Grid Search CV method optimizes the decision tree's depth. The model is seamlessly integrated into a user interface through Flask, a versatile Python framework. Additionally, this paper highlights the extensive research in the field, showcasing studies by various researchers who have explored diverse algorithms and methodologies for property price prediction. These contributions emphasize the importance of incorporating contextual features and employing advanced algorithms to enhance prediction accuracy. Looking ahead, this system aims to conduct a thorough comparative analysis between predicted prices and those sourced from established real estate websites. Moreover, future plans involve expanding the dataset to encompass a wider array of cities and states across India, making the system even more comprehensive and informative. The integration of Gmap functionality to provide insights into nearby amenities is poised to further augment property valuations. This paper signifies a significant advancement in real estate prediction models, promising enhanced accuracy and practical applicability in the dynamic real estate market.

**5. Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning[**]2018]**by Viktrovinch**

This research paper presents a robust methodology for predicting house sale prices using machine learning and regression techniques. Leveraging a Kaggle dataset, the authors conducted extensive data preprocessing, employing data imputation and feature engineering. They harnessed Python libraries like Pandas, NumPy, Scikit-learn, and XG Boost to implement a diverse range of models, including Lasso regression, Elastic Net, gradient boosting, and neural networks. Innovative additions to their approach include a unique "residual regressor" for case-specific predictions and logit transformation to address imbalanced data in tree-based models. Feature selection, normalization, and polynomial feature engineering were pivotal in improving model accuracy. The research culminated in an ensemble of models, which, after careful tuning, achieved a top 1% ranking in a Kaggle competition. The paper's insights and techniques offer valuable guidance for accurate house price prediction and highlight the importance of feature engineering and innovative model training approaches in real-world machine learning applications.

**Problem Statement**:

Develop a Machine learning model that can predict House prices with a high level of Accuracy

**Design Thinking Process Approach :**

**Empathize:**

Before diving into solving the problem, it's crucial to empathize with the users and understand their needs. In this case, our primary users are potential homebuyers and real estate professionals. We need to gather insights into what factors are most important to them when considering house prices and how accurate predictions can benefit them.

**Actions:**

- Conduct surveys or interviews with potential users to gather their perspectives.

- Analyse historical real estate market trends to identify critical pricing factors.

- Seek feedback from domain experts in the real estate industry.

**Define:**

Based on our understanding of the problem and the users' needs, we will define clear objectives and success criteria for our project.

**Objectives:**

- Develop a machine learning model that achieves a Mean Absolute Error (MAE) of less than $X on the test data.

- Create a user-friendly web application for users to input house details and receive price predictions.

**Idea:**

Brainstorm potential solutions and approaches to address the problem. This phase involves thinking creatively and considering various algorithms and techniques for house price prediction.

**Prototype**

Create a prototype of the machine learning model and the user interface for price prediction.

**Test**

Evaluate the model's performance using appropriate metrics and gather feedback from users.

**Implement**

Once the prototype meets the defined objectives and receives positive feedback, proceed with full implementation.

**Iterate**

Continuous improvement is essential. Gather user feedback and iterate on the model and interface to enhance accuracy and usability.

**Phases of development:**

The term "phases of development" typically refers to different stages or factors that influence the pricing of houses. These phases can include various aspects such as:

Initial Development: Refers to the early stages of construction and infrastructure development in a particular area. New developments can significantly impact house prices.

Infrastructure Growth: The expansion of local amenities like schools, hospitals, transportation networks, and shopping centers can lead to increased property values.

Economic Factors: Economic growth, job opportunities, and overall prosperity in an area can drive housing demand and influence prices.

Market Trends: Market trends, both regional and national, play a crucial role. If there's a housing boom or recession, it can affect prices.

Regulatory Changes: Changes in laws or regulations, such as zoning laws or tax policies, can impact property values.

Demographic Shifts: Changes in the population, such as migration patterns and population density, can influence housing demand and subsequently prices.

Supply and Demand: The balance between the number of available houses (supply) and the number of people looking to buy (demand) directly affects prices.

Understanding these phases and their interplay is essential in predicting house prices accurately, as each factor can have a varying degree of impact at different stages of development. Machine learning models and data analysis techniques are often used to analyse these factors and predict housing market trends.

**Dataset Used:**

<https://www.kaggle.com/datasets/vedavyasv/usa-housing>

**Description :**

The dataset available at the provided Kaggle link, titled "USA\_Housing," contains information related to housing prices in the United States. Here's a general description of what you can expect from this dataset.

# **Dataset Overview:**

**'Avg. Area Income'** The average income of residents in the area where the house is located.

**'Avg. Area House Age':** The average age of houses in the same locality.

**'Avg. Area Number of Rooms':** The average number of rooms in houses nearby.

**'Avg. Area Number of Bedrooms':** The average number of bedrooms in houses nearby.

**'Area Population'**: The population of the area where the house is located.

**'Price'**: The selling price of the house.

**'Address':** The address of the house.

# **Dataset Features:**

**Numerical Features:**

Avg. Area Income,' 'Avg. Area House Age,' 'Avg. Area Number of Rooms,' 'Avg. Area Number of Bedrooms,' 'Area Population,' and 'Price' are numerical features representing various aspects of the houses and the areas they are located in.

**Categorical Feature:**

The 'Address' feature is likely categorical, representing the location of the houses. However, in most machine learning models, addresses need to be transformed into numerical values for analysis.

## **Potential Use Case:**

This dataset is ideal for regression analysis, specifically predicting house prices based on different factors like income, house age, number of rooms and bedrooms, and area population. Researchers and data scientists can use this dataset to develop machine learning models aimed at accurately predicting house prices in the USA. The dataset provides a diverse set of features, making it suitable for exploring various regression algorithms and techniques for predicting real estate

**Data Pre processing**:

In the context of house price prediction, data pre processing refers to the process of preparing and cleaning the raw data to make it suitable for analysis and modaling. Several steps are involved in data pre processing for house price prediction

**Data Collection:**

Gathering relevant data from various sources, such as real estate databases, government records, or online platforms, to create a comprehensive dataset.

**Categorical Data Encoding:**

Converting categorical variables into numerical representations suitable for machine learning algorithms. Common methods include one-hot encoding and label encoding.

**Handling Outliers:**

Identifying and addressing outliers in the data. Outliers can significantly affect the model's performance, so they may be removed or transformed using appropriate techniques.

**Data Splitting:**

Dividing the dataset into training and testing sets. The training set is used to train the machine learning model, while the testing set is used to evaluate its performance.

**Normalization of the Target Variable:**

In some cases, especially in regression tasks, the target variable (house prices) might be transformed to follow a normal distribution. Techniques like logarithmic transformation can be applied to achieve this.

**Feature extraction Techniques**

Feature extraction in the context of house price prediction involves selecting or creating relevant features from the raw data that can effectively capture patterns and relationships, enhancing the predictive power of machine learning models. Here are some common feature extraction techniques for house price prediction:

**Area-related Features:**

**Total Area:**

The total area of the property is a fundamental feature. For houses, this includes both indoor and outdoor spaces.

**Living Area**:

The area specifically used for living spaces, excluding garages or storage areas.

**Number of Bedrooms and Bathrooms**:

The count of bedrooms and bathrooms in the house is a crucial factor.

**Lot Size:**

The size of the land on which the house is built.

Location-based Features:

**Neighbourhood Characteristics:**

Features describing the neighbourhood, such as crime rates, school quality, proximity to amenities, and public transportation access.

Geographical Features: Latitude, longitude, and distance from key locations like city centres or business

**Time-related Features:**

Year of Construction: The age of the property can be an important factor influencing its price.

**Renovation History:**

Information about recent renovations or upgrades to the property.

**Amenities and Facilities:**

Swimming Pool, Garden, Garage: The presence or absence of these amenities can affect the house price.

Energy Efficiency: Features related to energy-efficient appliances, solar panels, or insulation.

**Market Trends and Economic Indicators:**

Historical Price Trends: Price trends of similar properties in the area over a specific period.

Economic Indicators: Local economic data, such as employment rates and GDP growth, can influence property prices.

**Derived Features:**

Price per Square Foot: Calculating the price per square foot can provide a standardized measure for comparison.

Age of the Property: Deriving the age of the property from the year of construction can be a relevant feature

Sentiment Analysis: Analysing online reviews and sentiments about the neighbourhood or the property itself can provide additional context.

**Accessibility Features:**

Distance to Public Transportation: Proximity to bus stops, train stations, or subway lines.

Walkability Score: A measure of how easily amenities can be accessed on foot.

Demographic Data: Information about the demographics of the area's population, such as average income and education level.

**Choice of Machine learning Algorithm**

**Linear Regression Basics:**

Linear regression is a straightforward method for predicting numerical outcomes, making it ideal for tasks like house price prediction.

**Simplicity and Interpretability**:

Its simplicity allows for easy interpretation of the relationship between input features and house prices, making it accessible to a wide audience, including non-experts.

**Computational Efficiency:**

Linear regression is computationally efficient, enabling it to handle large datasets and a high number of features without significant processing time.

**Feature Importance Analysis:**

The model provides insights into feature importance, helping identify which factors have a substantial impact on house prices

**Benchmark Model:**

Linear regression serves as a benchmark for comparing the performance of more complex algorithms, providing a foundational understanding of the problem.

**Non-Linearity Consideration:**

However, it assumes a linear relationship, which might not hold true for all datasets. In cases of highly nonlinear relationships, more advanced models are necessary for accurate predictions

**Regularization Techniques:**

Advanced versions of linear regression, like Lasso and Ridge regression, incorporate regularization methods, enhancing the model's accuracy and preventing overfitting.

However, its applicability depends on the linearity of the relationship, and for more complex scenarios, considering advanced techniques is crucial.Top of Form

**Model Training:**

Gather and clean historical house price data, converting categorical features and handling missing values.

Choose key features like square footage and bedrooms. Split data for training and testing.

Train the model, adjusting coefficients through optimization methods to minimize errors, using metrics such as Mean Squared Error for evaluation.

Fine-tune hyper parameters like learning rate, ensuring optimal performance.

Deploy the model for real-time predictions, assisting buyers and sellers in making informed decisions based on accurate house price estimates.

Continuous monitoring allows adaptation to market changes, ensuring long-term reliability.

**Evaluation metrics:**

House prices are predicted using linear regression through evaluation metrics such as Mean Squared Error (MSE) or Root Mean Squared Error (RMSE). Here's how these metrics work:

**Mean Squared Error (MSE):**

MSE measures the average squared difference between predicted and actual house prices. Lower MSE indicates better accuracy. The formula is MSE = Σ(yᵢ - ŷᵢ)² / n, where yᵢ is the actual price, ŷᵢ is the predicted price, and n is the number of data points.

**2. Root Mean Squared Error (RMSE):**

RMSE is the square root of MSE, providing an interpretable measure in the same unit as house prices. RMSE = √(Σ(yᵢ - ŷᵢ)² / n). Smaller RMSE values signify more accurate predictions.

**3. R-squared (R²) Score:**

R² measures the proportion of the variance in the dependent variable (house prices) that is predictable from the independent variables (features). R² ranges from 0 to 1, where 1 indicates a perfect fit. Higher R² values signify a better model fit.

By using these evaluation metrics, linear regression models are assessed for accuracy and reliability in predicting house prices, ensuring effective decision-making in real estate transactions

**Innovative Techniques:**

In our house price prediction project, consider these unique and cutting-edge techniques:

**Deep Learning Architectures**: Utilize neural networks like CNNs or RNNs to capture intricate patterns in data, enhancing prediction accuracy.

**Transfer Learning**: Adapt pre-trained models for related tasks, leveraging existing knowledge to boost predictions.

**Time Series Forecasting**: Apply ARIMA or Prophet for precise predictions, especially if your data has a temporal component.

**Geospatial Analysis**: Incorporate spatial techniques like GIS data analysis to understand location-based price patterns.

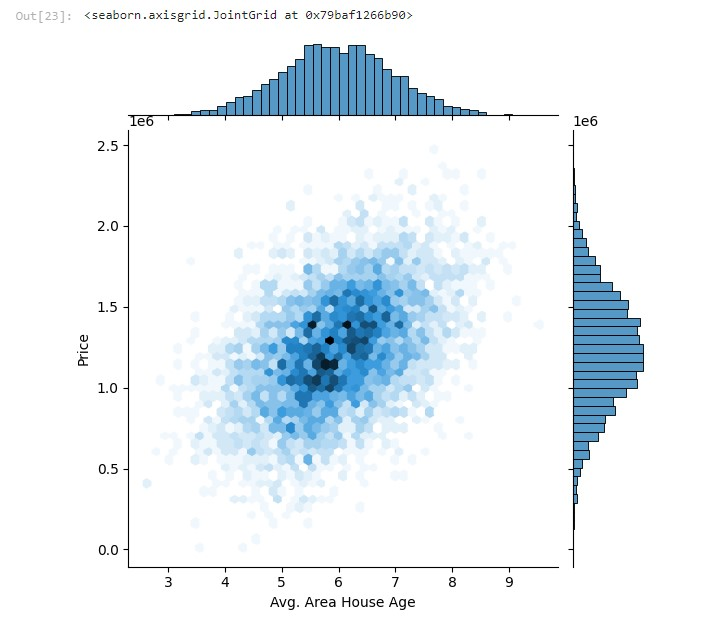
**GANs for Data Augmentation**: Generate synthetic data with GANs to augment limited datasets and improve model generalization.

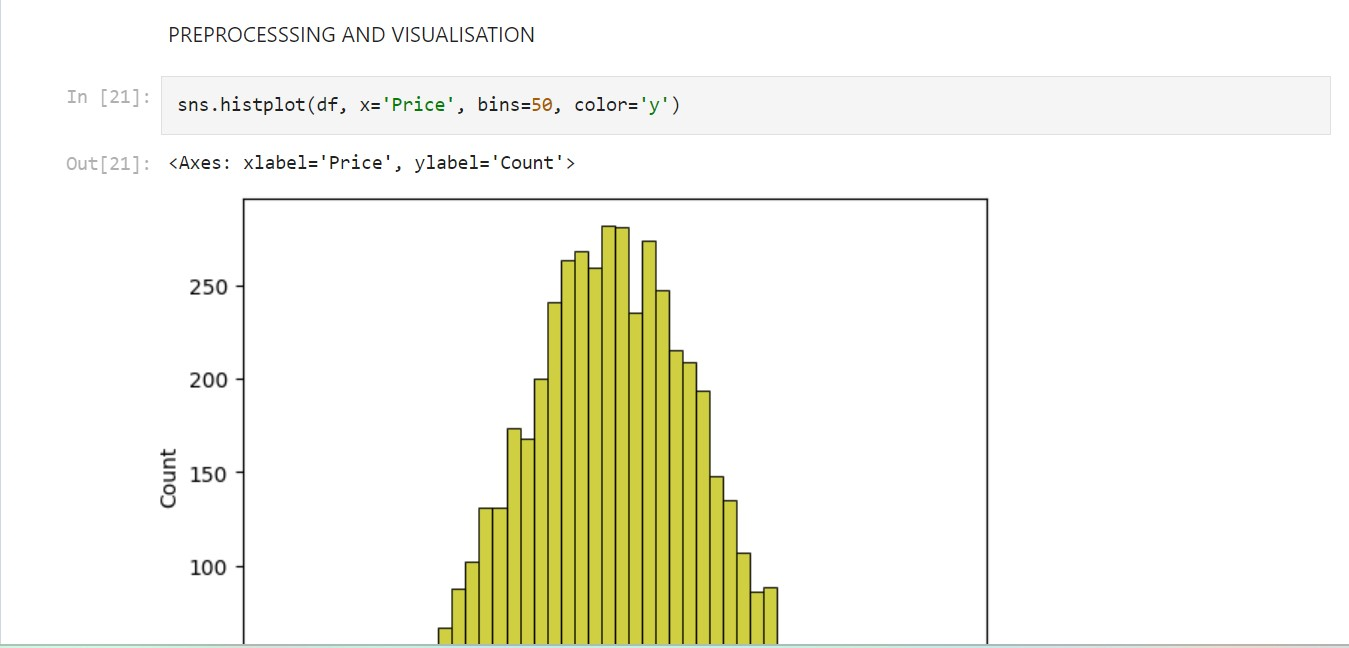
**Explainable AI (XAI):** Use SHAP or LIME to make predictions interpretable, enhancing transparency and trust in your model.

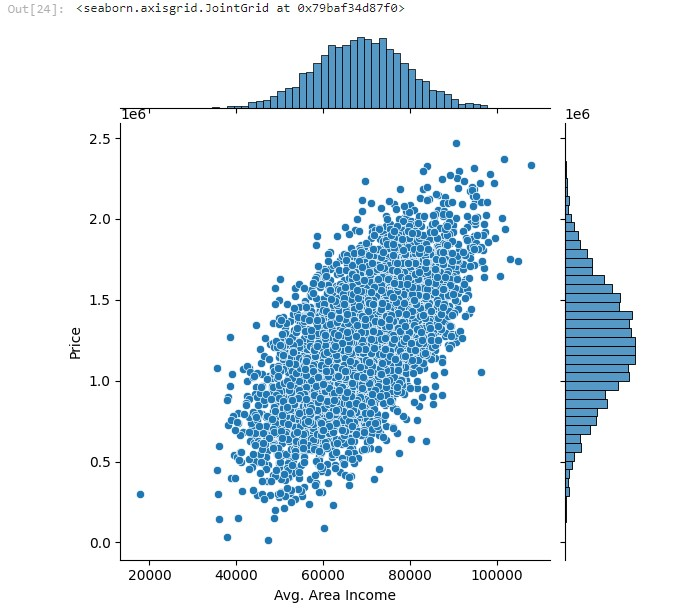
**Federated Learning**: Implement this technique for privacy-preserving model training across decentralized data sources.

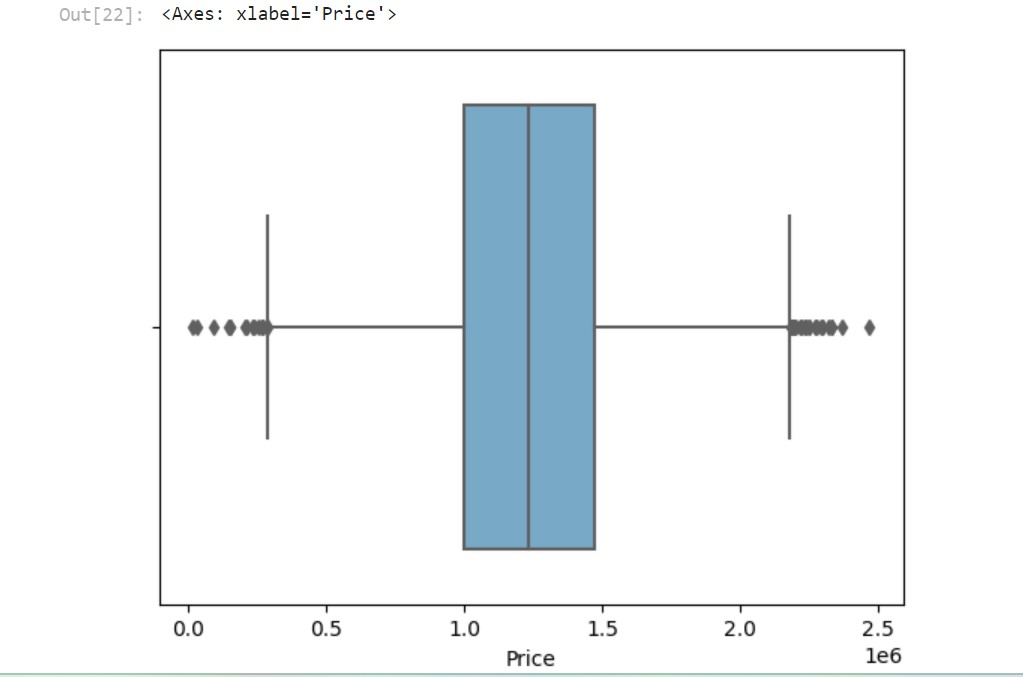
**Block chain for Data Integrity**: Ensure data integrity and immutability using block chain, crucial for maintaining data trustworthiness.

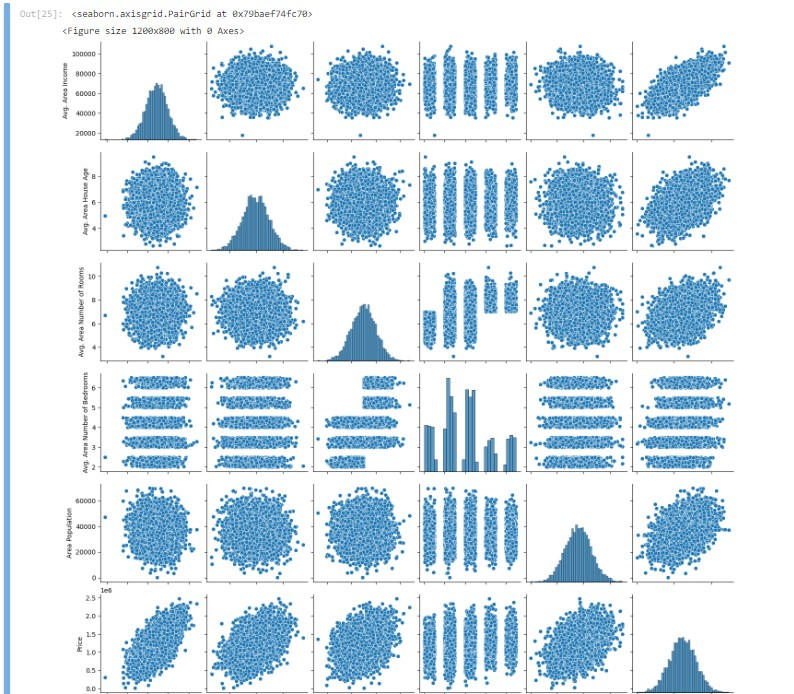
**Data Pre processing and Visualization**

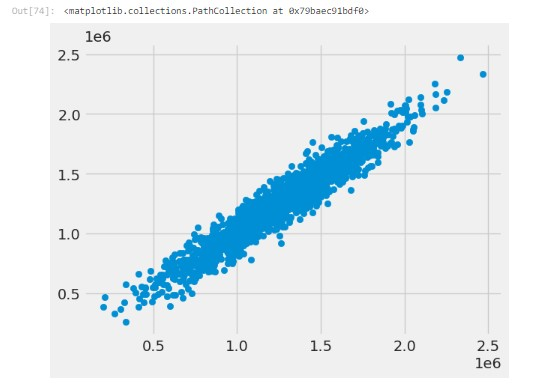


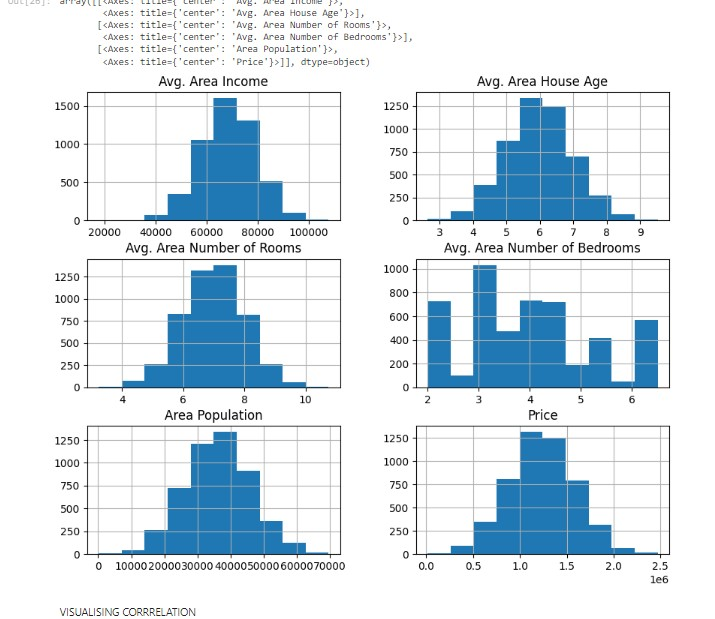


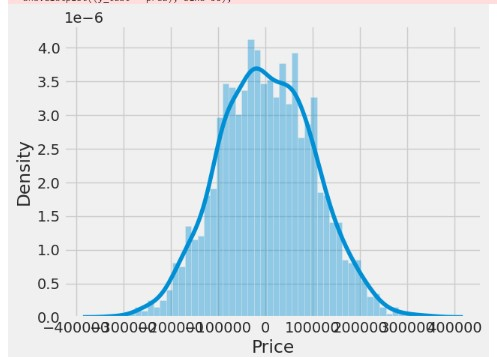












**Conclusion**

In conclusion, for our house price prediction project, incorporating cutting-edge techniques such as deep learning architectures, transfer learning, time series forecasting, geospatial analysis, GANs for data augmentation, explainable AI methods, federated learning, and block chain for data integrity can significantly enhance the accuracy, interpretability, and robustness of your predictive model. By leveraging these innovative approaches, your project can achieve more precise predictions, even with limited or complex data, while ensuring transparency, privacy, and trustworthiness in the process.