**House price prediction (phase-5)**

## Problem definition:

The problem is to predict house prices using machine learning techniques. The objective is to develop a model that accurately predicts the prices of houses based on a set of features such as location, square footage, number of bedrooms and bathrooms, and other relevant factors. This project involves data preprocessing, feature engineering, model selection, training, and evaluation.

## Abstract:

The House Price Prediction project aims to leverage machine learning techniques to create a robust and accurate model for predicting the prices of residential properties. The project revolves around the comprehensive analysis of various features, including but not limited to location, square footage, number of bedrooms and bathrooms, and other pertinent factors, to develop a predictive model that assists both buyers and sellers in making informed decisions.

The project's success hinges on a multi-stage process, commencing with meticulous data preprocessing to clean and organize the dataset. Subsequently, feature engineering techniques will be applied to extract valuable insights from the data, potentially creating new features that can enhance prediction accuracy.

Model selection will be a critical aspect of the project, with a focus on exploring diverse machine learning algorithms such as linear regression, decision trees, random forests, and more, to determine the best-fit model for the task. Rigorous training will be conducted on historical housing data to optimize the chosen model's performance.

Evaluation of the model's effectiveness will be carried out using various metrics, including mean squared error, root mean squared error, and R-squared, among others, to gauge its predictive capabilities. The project's ultimate goal is to develop a dependable house price prediction model that can provide valuable insights into the real estate market, aiding both buyers and sellers in making well-informed decisions.

## Design thinking:

**1.Data Source:** We choose a dataset from Kaggle which contains information about house including features like location, square footage, number of bedrooms, average number of rooms, and price.

1. **Avg. Area Income**: This column represents the average income of residents in the area where the property is located.
2. **Avg. Area House Age**: This column represents the average age of houses in the same area.
3. **Avg. Area Number of Rooms**: This column indicates the average number of rooms in houses in the area.
4. **Avg. Area Number of Bedrooms:** This column represents the average number of bedrooms in houses in the area.
5. **Area Population:** This column represents the population of the area where the property is situated.
6. **Price:** This column contains the price of the property.
7. **Address:** This column provides the address of each property, including city or town names, postal codes, and unit or apartment numbers.
8. **Data Preprocessing:**

Clean and preprocess the data, handle missing values, and convert categorical features into numerical representations.

1. **Check for missing values**: *isnull()* function to identify missing values in the dataset and *sum()* to count them for each column.
2. **Handle missing values:** Fill missing values with the mean of the respective columns using the *fillna()* method.
3. **Convert categorical features into numerical representations:** One hot Encoding is the best way to convert categorical data into binary vectors. This maps the values to integer values.

## 2. Feature Selection:

For predicting house prices **feature variable** as X which contains **'Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',' Avg. Area Number of Bedrooms', 'Area Population'** and target variable as Y which contains **‘Price’.**

## 3. Model Selection:

A suitable regression algorithm for predicting house prices is **Linear Regression**. Linear Regression predicts the final output-dependent value based on the given independent features. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable.

## 4. Model Training:

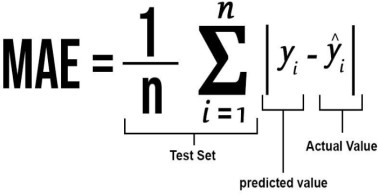
Train the selected model using the preprocessed data. Feed the training data (features and labels) into the model. The model will adjust its internal parameters to learn the underlying patterns in the data. During training, the algorithm chosen tries to minimize a certain loss function, which measures the difference between its predictions and the true labels in the training data.

## 5. Evaluation:

To evaluate the model's performance using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

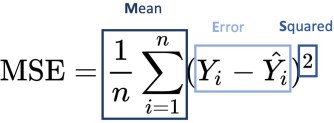
## MAE:

Mean Absolute Error (MAE) is a straightforward metric for assessing the accuracy of a regression model. It provides a clear and interpretable measure of how well the model predicts continuous numerical value.



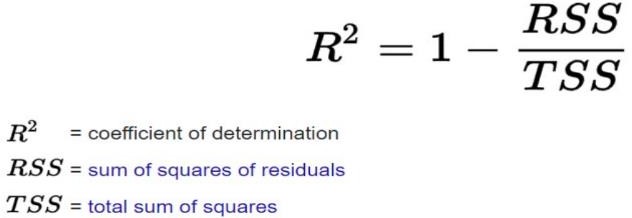
## MSE:

MSE stands for "Mean Squared Error." It is another common metric used for evaluating

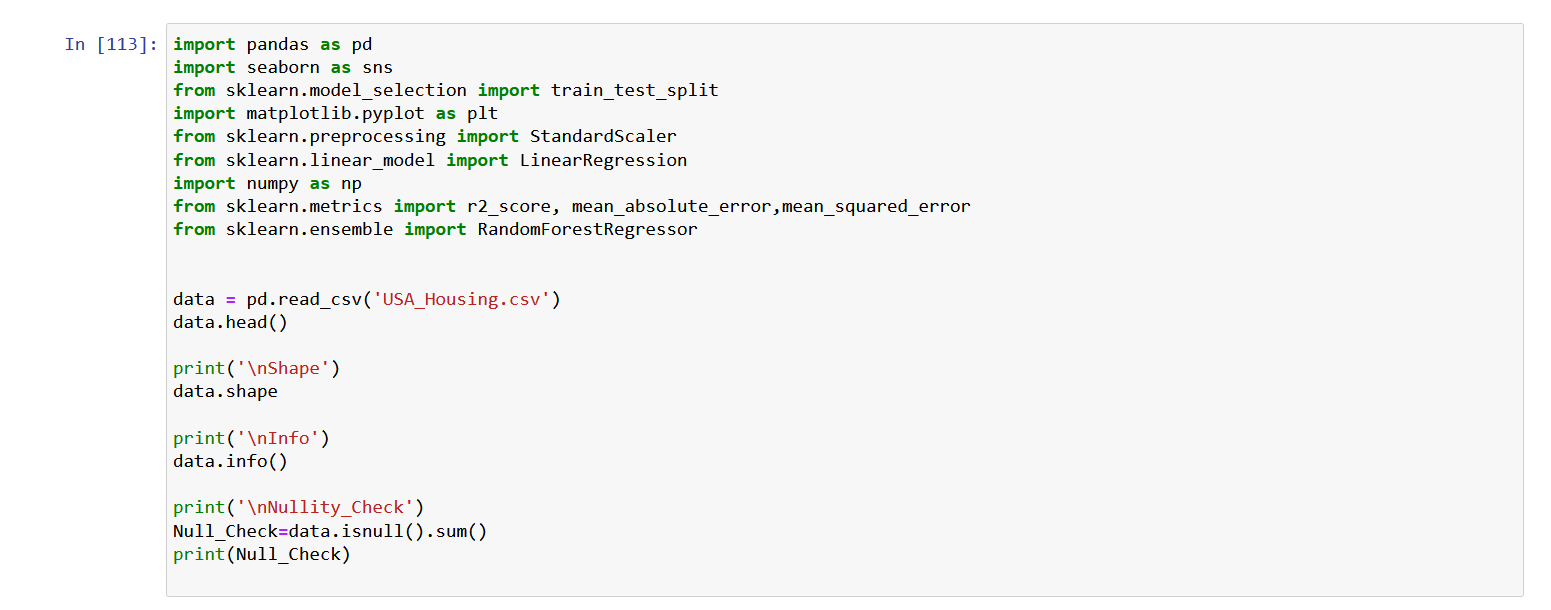
the performance of regression models, especially in the context of predictive modeling. MSE quantifies the average of the squared differences between predicted values and actual (true) values in a dataset.

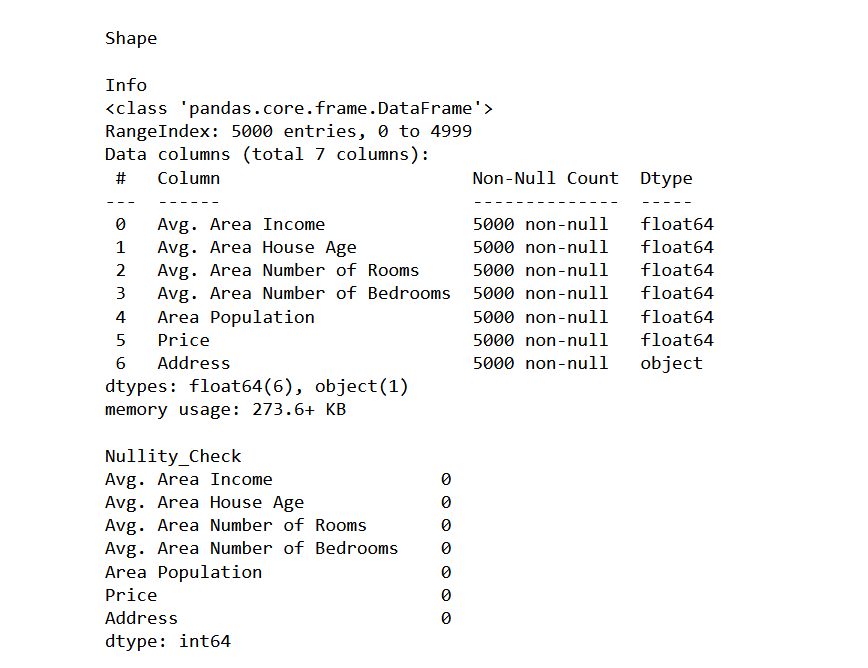
## R-squared:

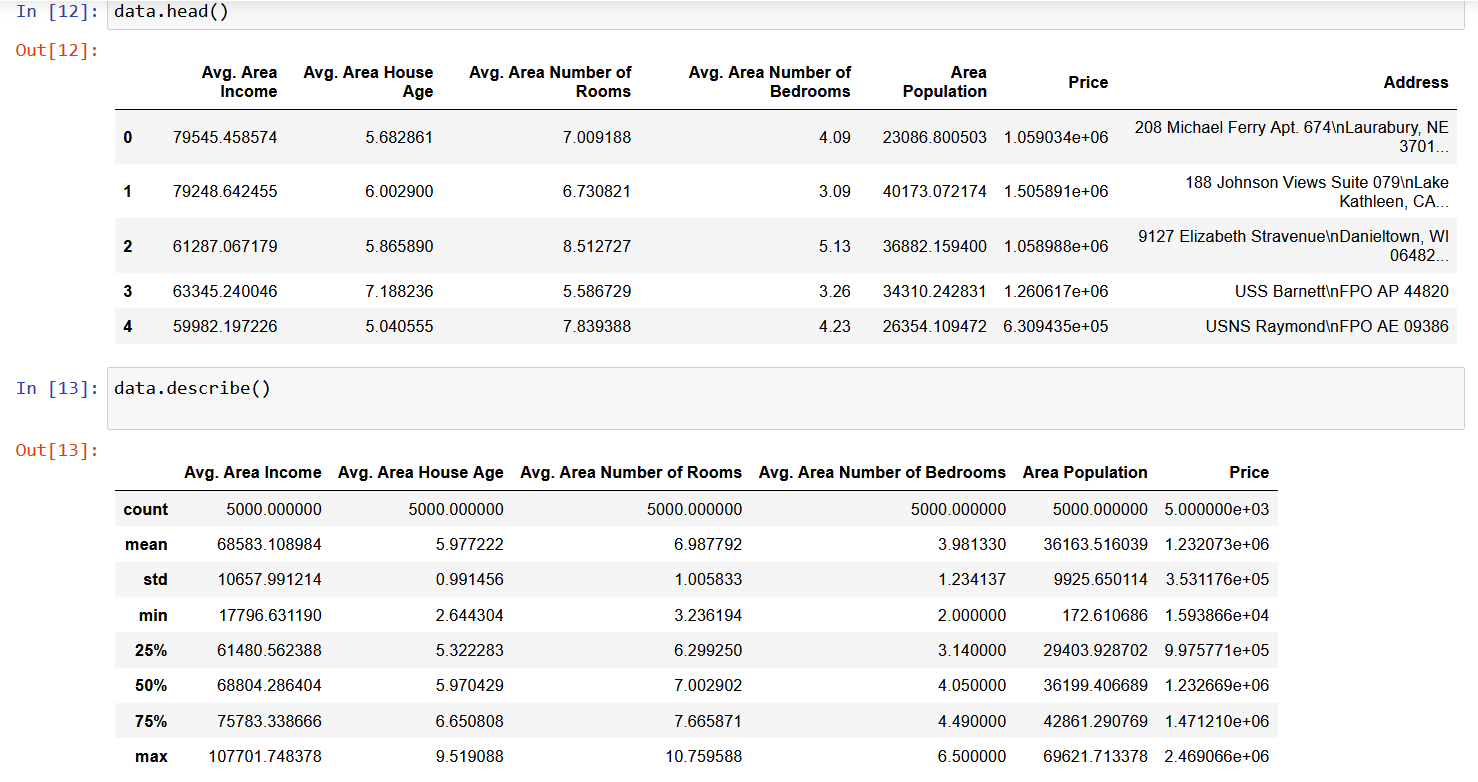
R-squared (R²), also known as the coefficient of determination, is a statistical measure used to assess the goodness of fit of a regression model. It quantifies the proportion of the variance in the dependent variable (target variable) that is explained by the independent variables (features) in the model. In other words, R-squared indicates how well the model fit the data.



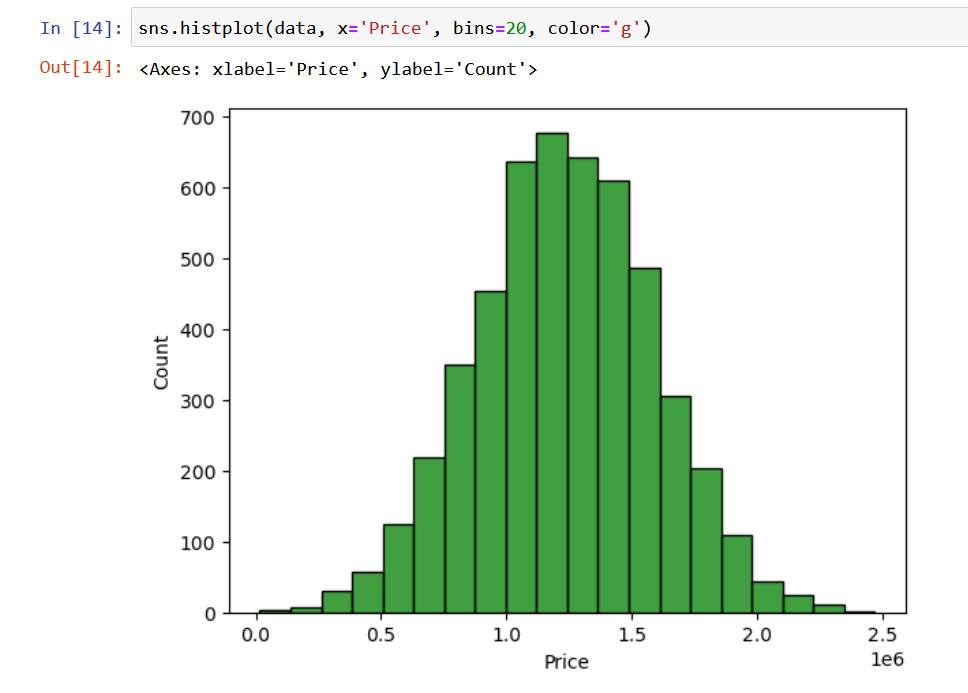
**Implementations in Jupyter Notebook:**

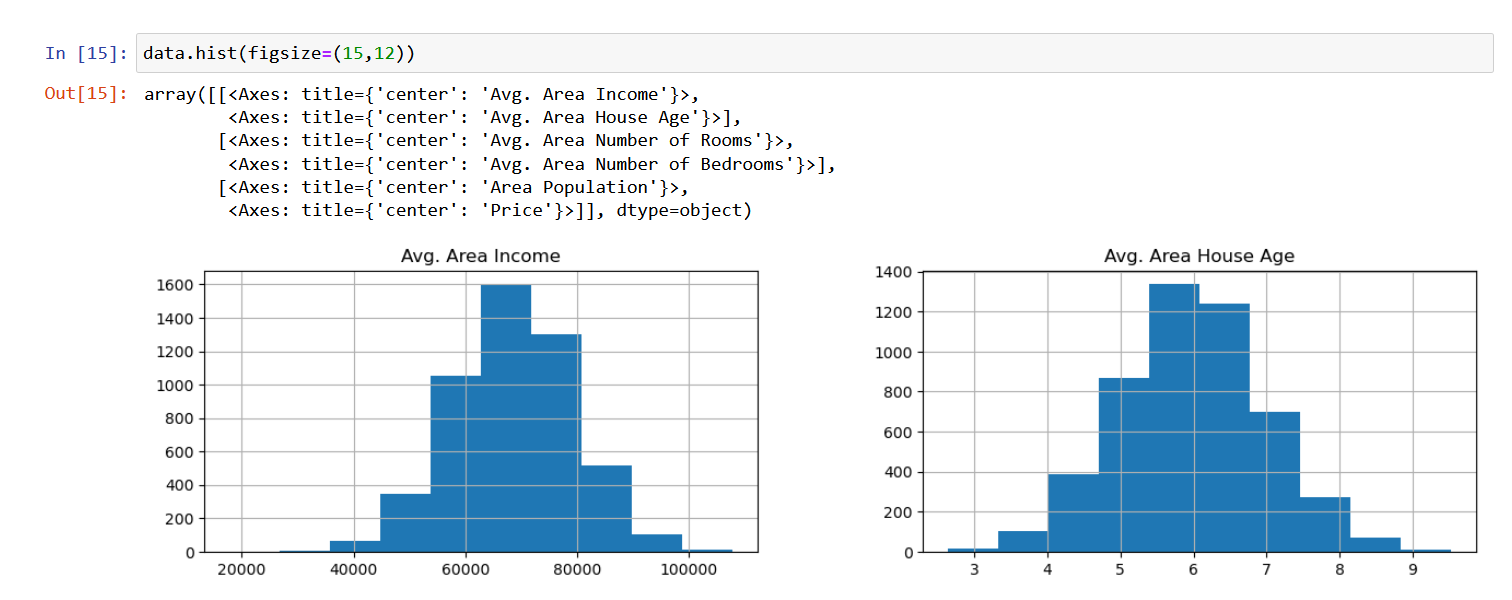


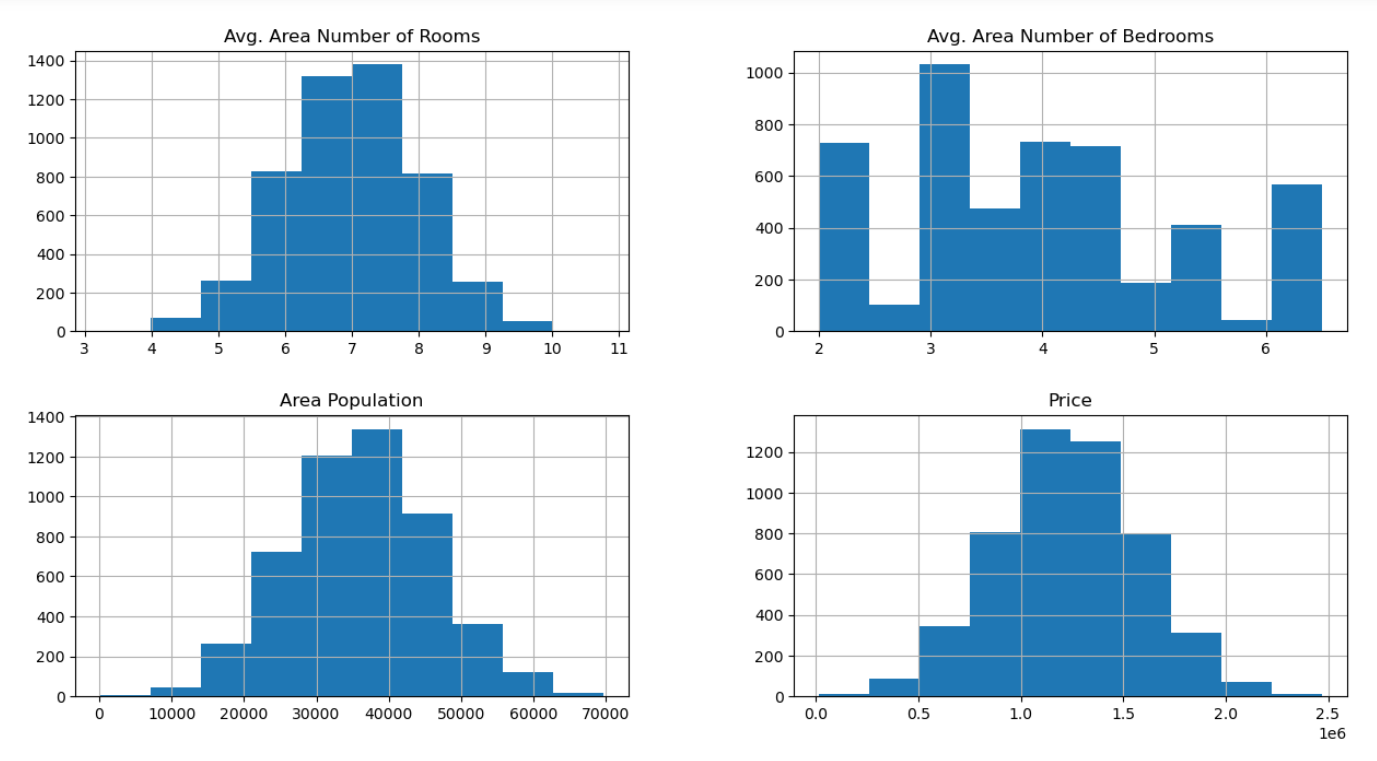
**Output:**

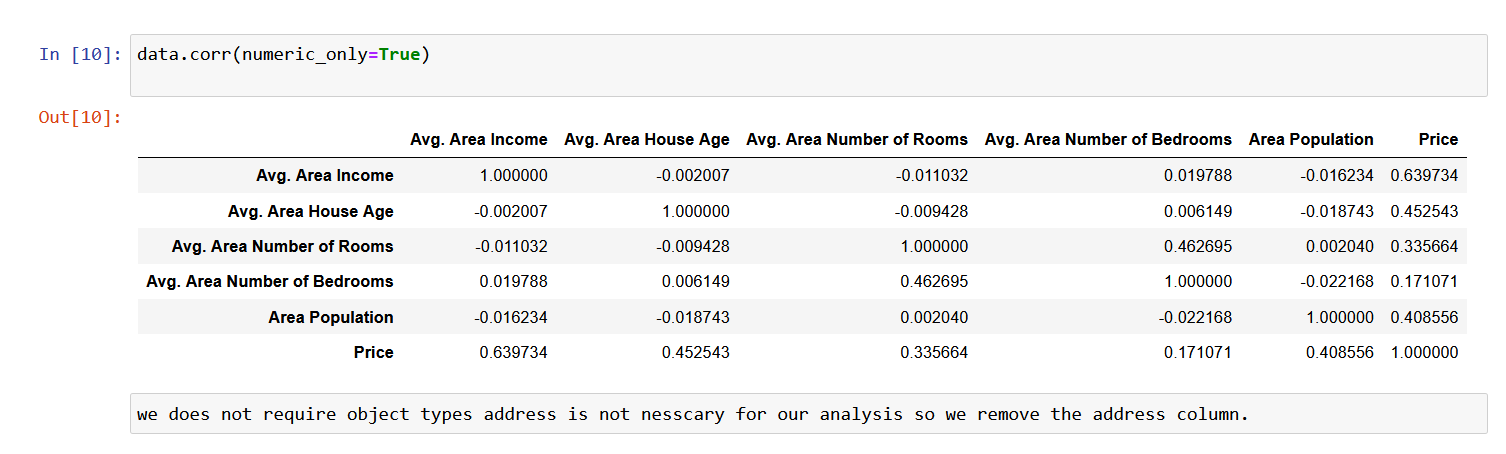


**There is no null values in the cells. So, it does need to drop or fill the null values**.

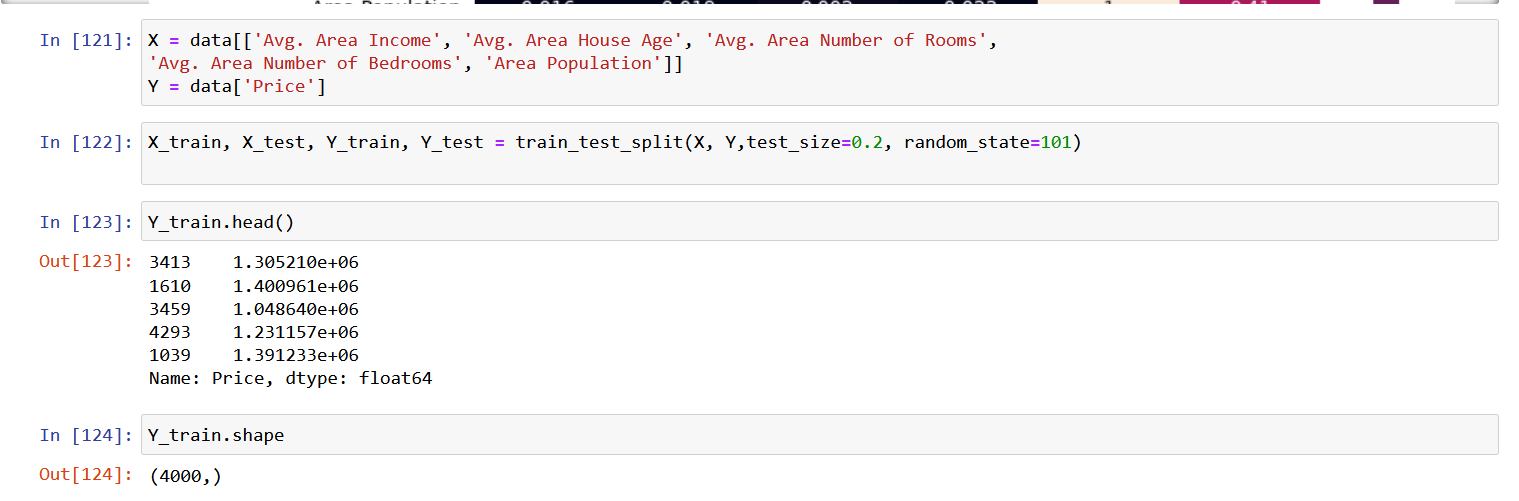




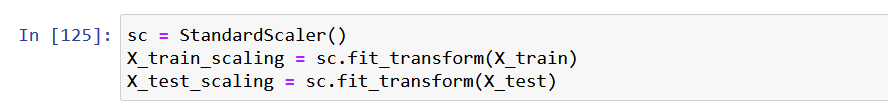




**Dividing dataset & Train, Test, split the data**



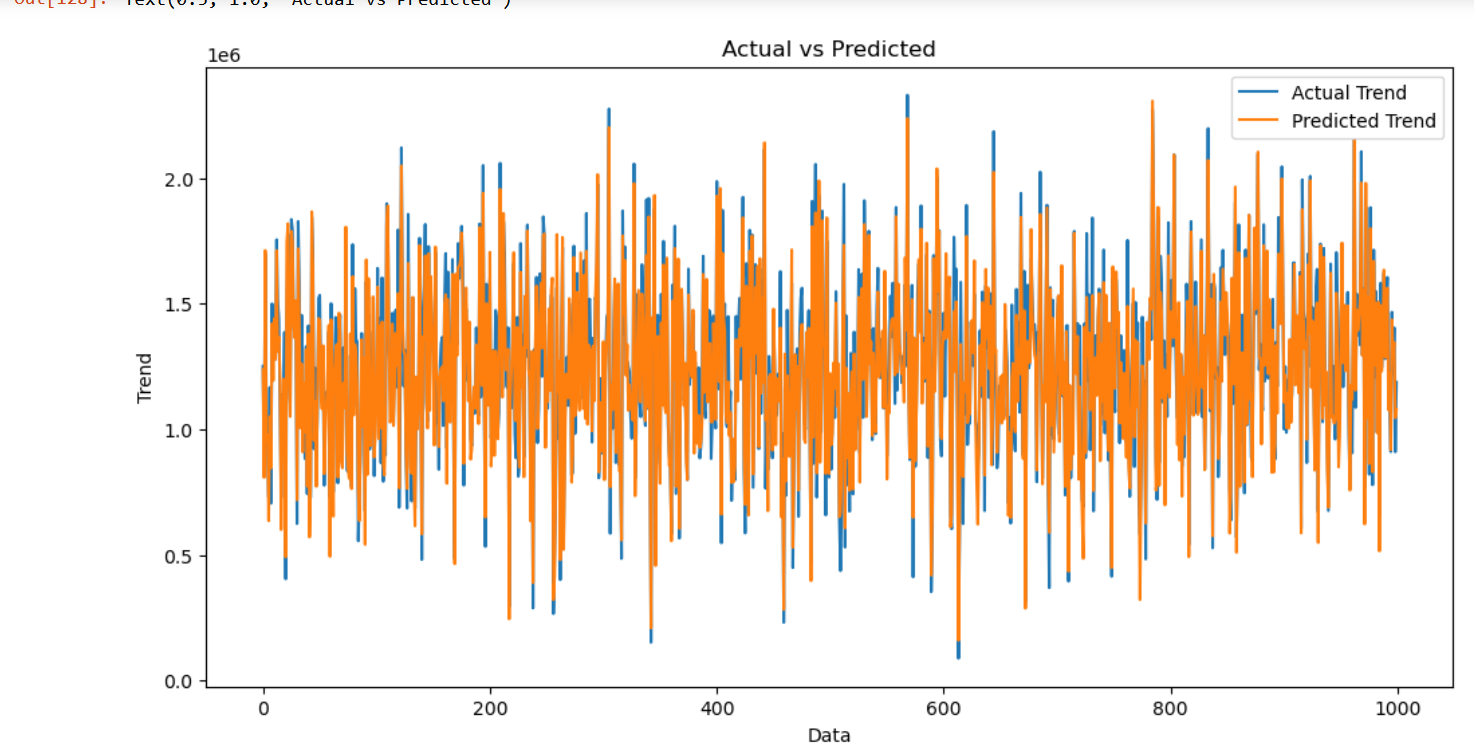
# **Standardizing the data**

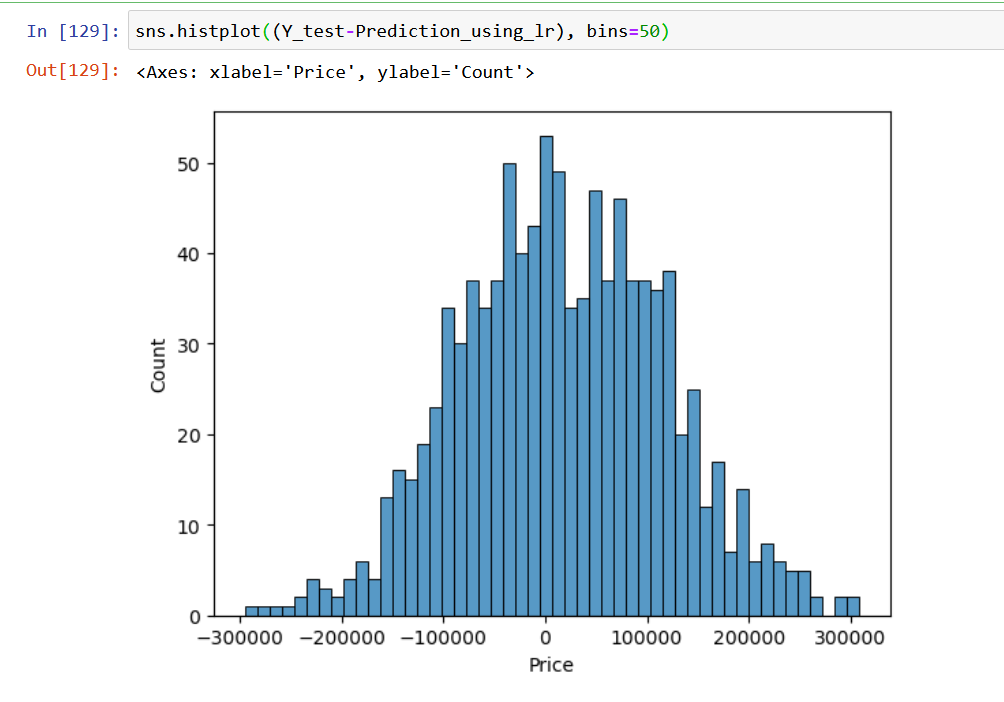


**Model building and Evaluation using Linear Regression algorithm**

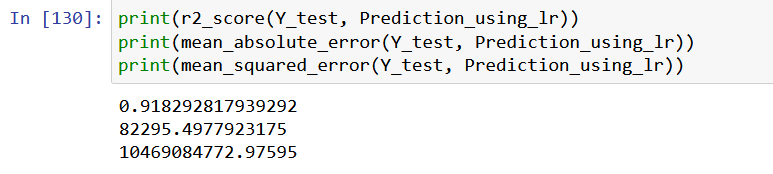
**Linear regression** uses a linear relationship between an independent and dependent variable to forecast future events. It is a statistical technique for predictive analysis used in data science and machine learning.





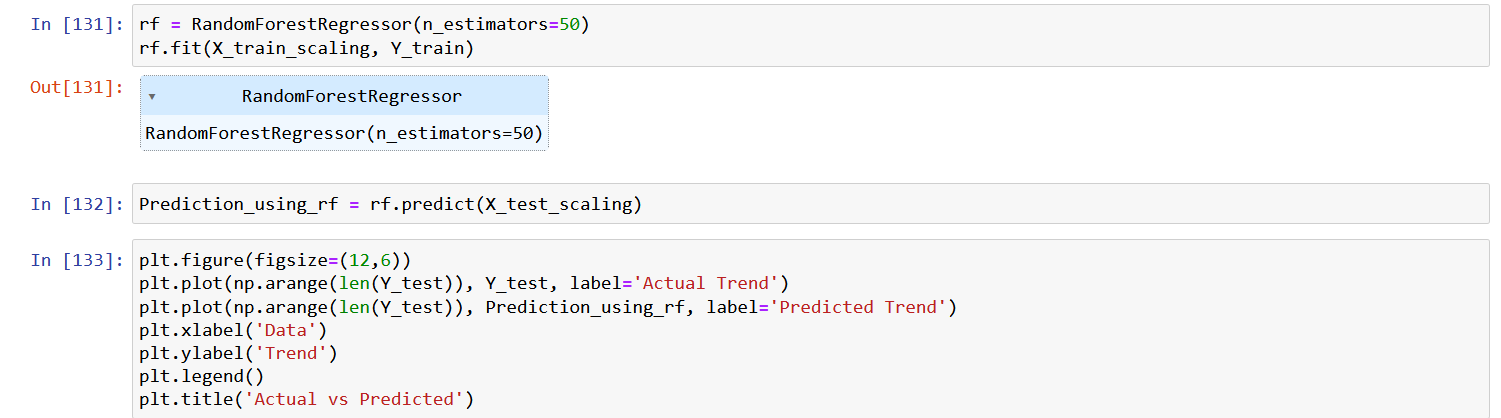
**Testing view in histogram graphical view**

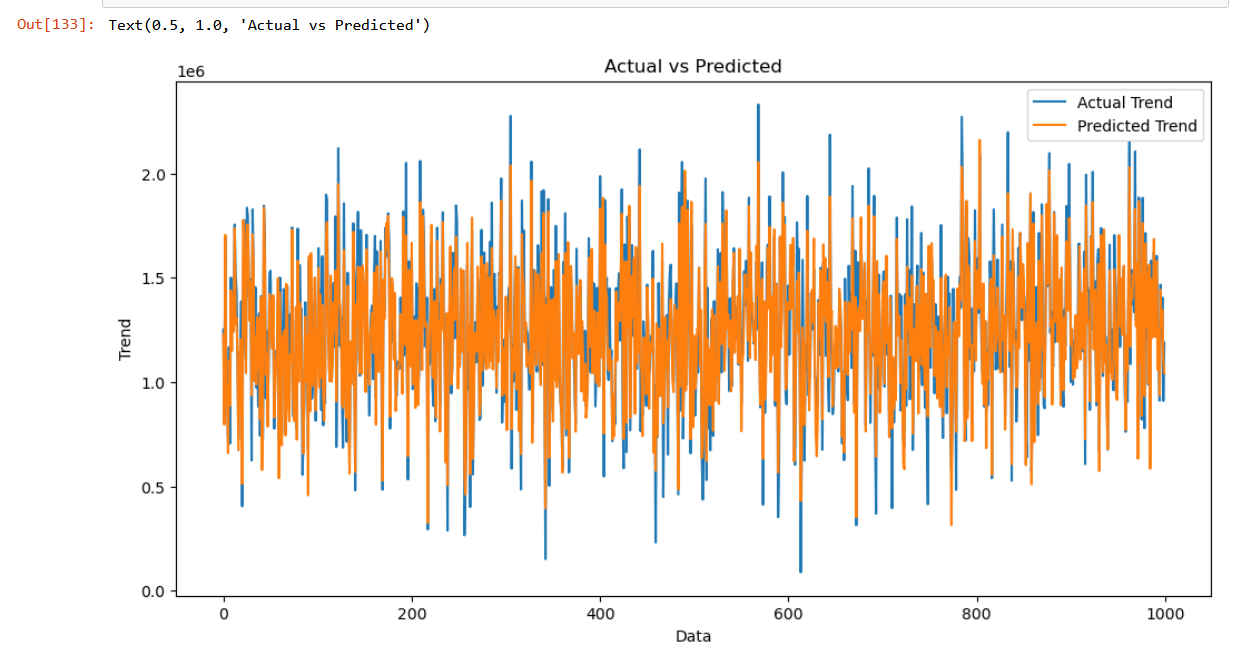
**Our prediction Score value**

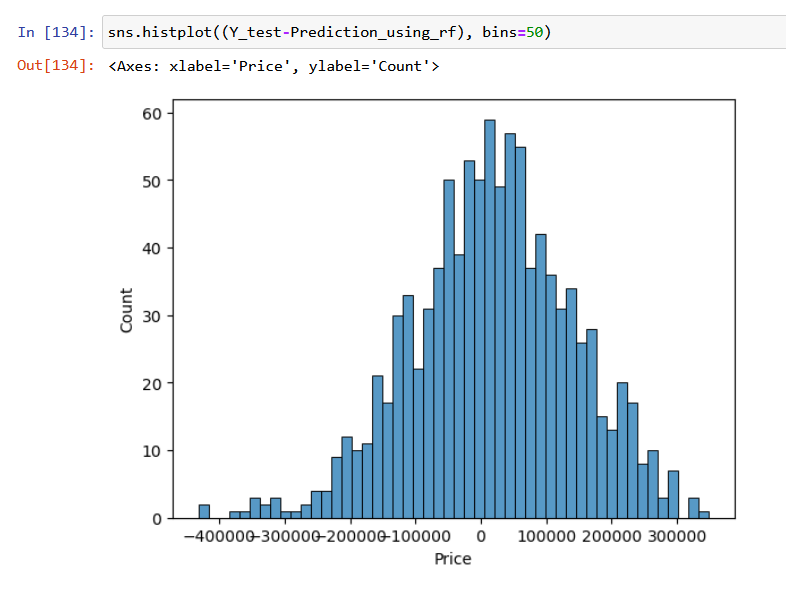
****

**Model building and Evaluation using Random Forest Regressor algorithm**

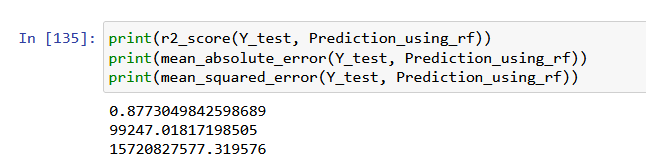
**Random Forest** is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging

****





**Prediction Score using random forest regressor**



Both are nearly give the r2\_score, absolute error and squared error. And the both graphical view for actual vs predicted values are also nearly same.

## CONCLUSION :

House price prediction plays a crucial role in the real estate industry, offering invaluable insights for both buyers and sellers. Data quality is paramount, with key variables such as property location, size, and comparable sales data forming the foundation for accurate predictions. Machine learning models, including regression algorithms and deep learning techniques, are commonly employed, with feature engineering to enhance predictive capabilities. Data splitting and rigorous evaluation metrics, such as mean squared error and root mean squared error, help gauge model performance. Guarding against overfitting and incorporating domain knowledge are essential, as house prices fluctuate over time and can be influenced by local market conditions. Furthermore, ethical considerations, like addressing potential biases, are vital to ensure fair and unbiased predictions, making this task an intricate blend of data science, real estate expertise, and responsible AI application.

House price prediction involves a multifaceted approach that goes beyond mere statistical modeling. Consideration of geospatial data, which includes factors like neighborhood safety, school quality, and accessibility to public services, can provide a more comprehensive understanding of a property's value. Recognizing seasonal trends and integrating historical data allow for a nuanced view of price fluctuations over time. Moreover, external economic factors and their influence on the real estate market should not be underestimated. Keeping an eye on interest rates, inflation, and employment rates can yield valuable insights for predictions. Property-specific features and conditions, such as architectural style, age, and maintenance, add another layer of complexity to the prediction process. Lastly, awareness of local regulations and their potential impact on property values is crucial, as zoning laws, tax policies, and other governing rules can significantly affect housing prices. Therefore, a holistic approach to house price prediction combines data analysis, domain knowledge, and a deep understanding of the broader economic and regulatory landscape.

**Dataset link:**

[**https://www.kaggle.com/datasets/vedavyasv/usa-housing**](https://www.kaggle.com/datasets/vedavyasv/usa-housing)

**Github link:**

https://github.com/kumarsnkl2003/house\_price\_prediction\_phase-5.git