HOME.LLC ASSESSMENT

“Evaluating Key Factors Influencing US Home Prices: A Data Science Approach”

Define the Problem:

- **Objective**: Build a data science model that explains how key factors have influenced US home prices over the last 20 years using the S&P Case-Schiller Home Price Index as a proxy.

Procedure:

- **Importing Libraries and Loading Data.**

Libraries like numpy, pandas, matplotlib, and seaborn are imported for data manipulation and visualization.

train_test_split and metrics like r2_score and mean_squared_error from sklearn are used for model evaluation.

The dataset is loaded from a CSV file and indexed by the 'DATE' column for time series analysis.

- **Data Preprocessing**

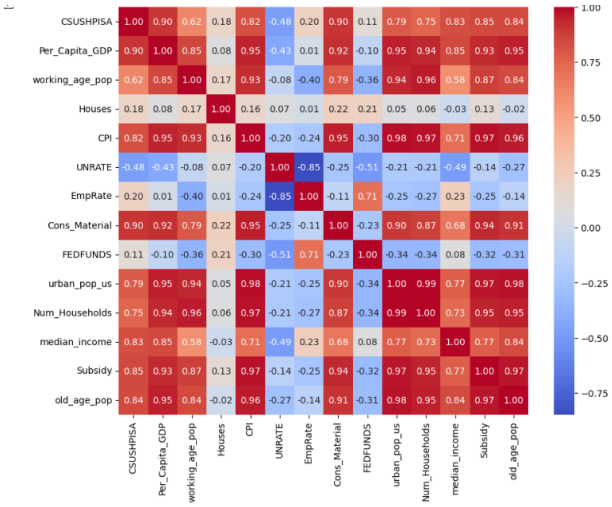
The 'Year' and 'Month' columns are dropped as they are redundant once the data is indexed by 'DATE'.

Summary statistics provide insights into the central tendency and dispersion of the data.
The correlation matrix helps in understanding how different features relate to each other and to the target variable.

	CSUSHPISA	Per_Capita_GDP	working_age_pop	Houses	CPI	UNRATE	EmpRate	Cons_Material	FEDFUNDS	urban_pop_us	Num_Households	median_income	Subsidy	old_age_pop
count	252.000000	252.000000	2.520000e+02	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000	252.000000
mean	177.877885	57049.416667	1.993721e+08	6.039683	227.634710	6.001190	69.613343	209.573222	1.319008	81.185714	120770.761905	68776.666667	34.177714	13.952381
std	42.147185	4166.690070	6.962554e+06	1.955341	28.868851	1.985615	2.071874	45.902963	1.542732	1.115463	6496.223206	4512.996725	6.274362	1.535125
min	117.144000	50091.000000	1.825653e+08	3.300000	177.700000	3.500000	60.195798	142.000000	0.050000	79.400000	109297.000000	63350.000000	24.183000	12.300000
25%	146.768500	54205.833333	1.952725e+08	4.575000	205.750000	4.600000	67.798555	183.225000	0.120000	80.300000	116011.000000	65760.000000	29.512000	12.500000
50%	170.175500	55677.000000	2.014882e+08	5.500000	228.997000	5.550000	70.272035	206.300000	0.715000	81.100000	121084.000000	66780.000000	33.283000	13.600000
75%	194.020500	60008.250000	2.054529e+08	7.000000	246.482750	7.200000	71.374867	223.300000	1.935000	82.100000	126224.000000	72090.000000	37.550000	15.100000
max	304.755000	65979.666667	2.075097e+08	12.200000	298.990000	14.700000	72.333310	353.015000	5.260000	83.100000	131202.000000	78250.000000	48.021000	17.100000

- **Visualizing Correlations**

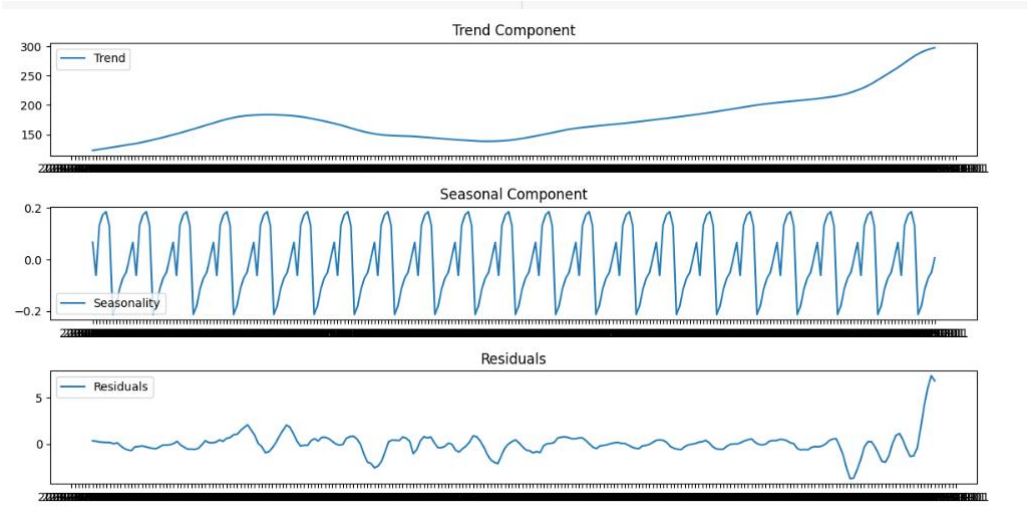
A heatmap is used to visualize the correlation matrix. This helps in identifying strong relationships between features and the target variable, as well as any potential multicollinearity.



- **Time Series Analysis**

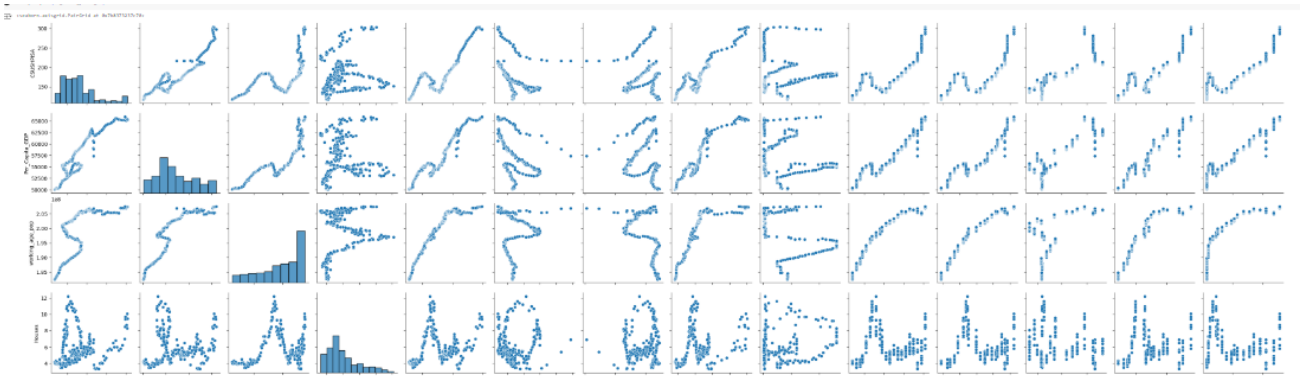
1. Decomposition**:** The time series is decomposed into trend, seasonal, and residual components to understand underlying patterns.

Autocorrelation (ACF) and Partial Autocorrelation (PACF): These plots help in identifying the presence of autocorrelation and the number of lags to consider in time series models.

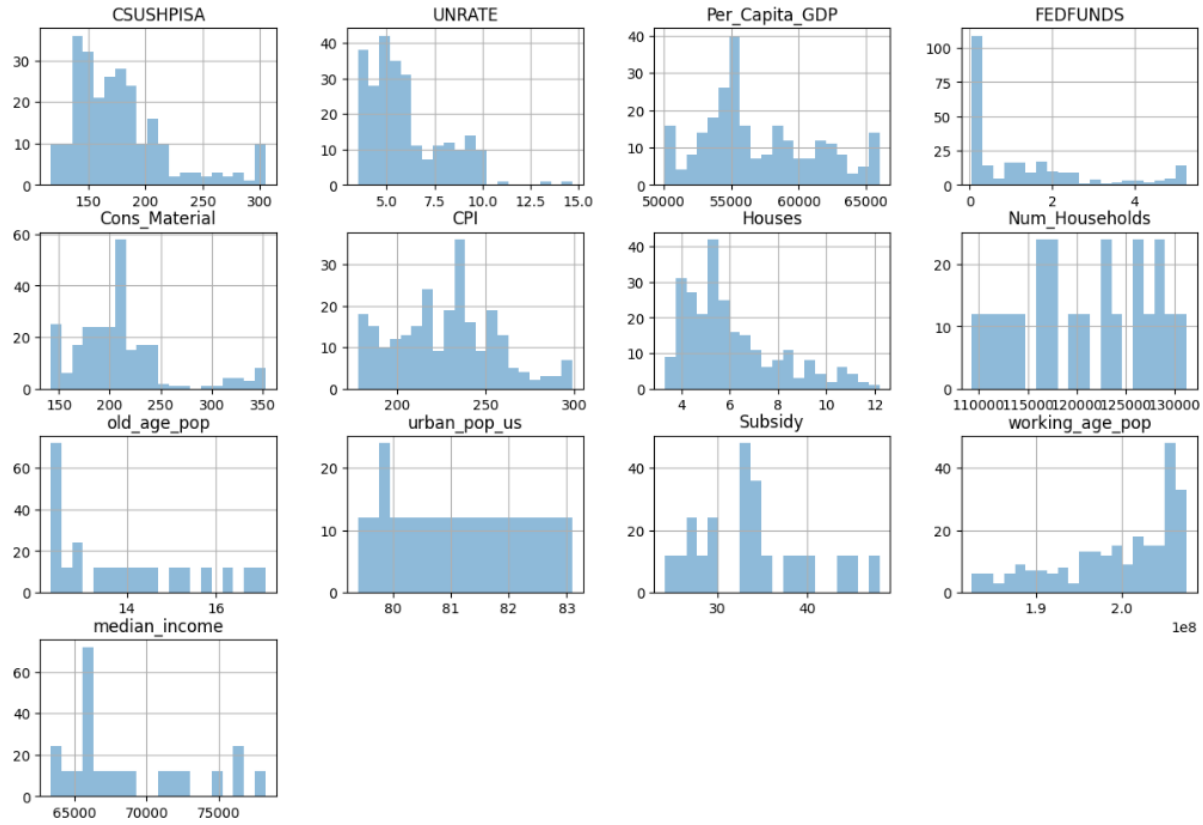


Exploratory Data Analysis.

- Pair **plot:** Shows scatter plots for each pair of features to explore relationships.

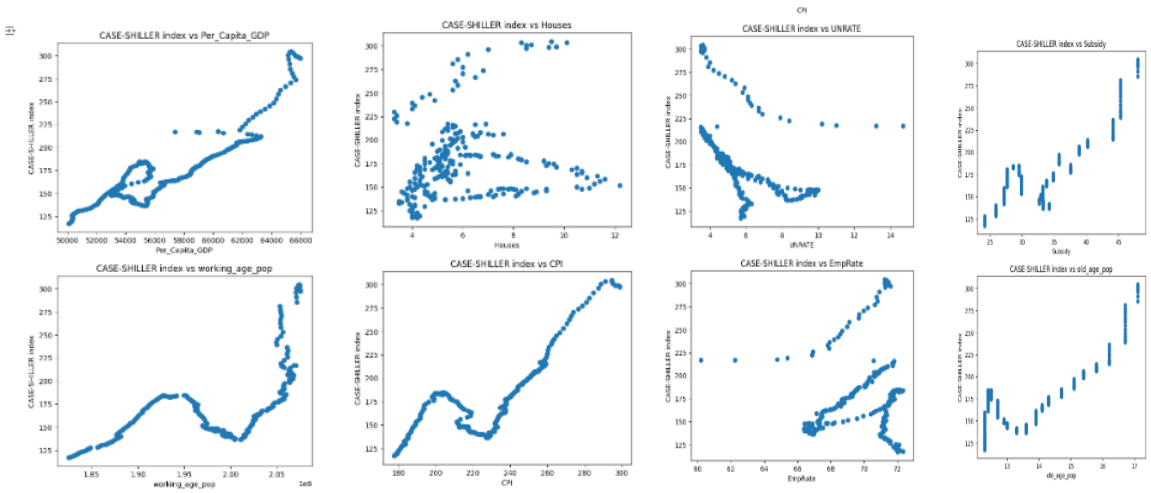


- Histograms**:** Visualize the distribution of each feature.



- **Feature Selection and Mo del Training**

Feature Analysis: Scatter plots and correlations are used to assess the relationship between each feature and the target variable.



Feature Selection: Features with low correlation are removed to simplify the model.

Data Splitting and Standardization: The data is split into training and testing sets, and features are standardized to ensure consistent scaling.

Model Training and Evaluation

Training: Each model is trained using the training data (X_train_scaled and y_train).

Prediction: Predictions are made on the test data (X_test_scaled).

Evaluation Metrics:

- Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values. Lower values indicate better performance.
- R-squared: Represents the proportion of variance explained by the model. Higher values indicate better performance.

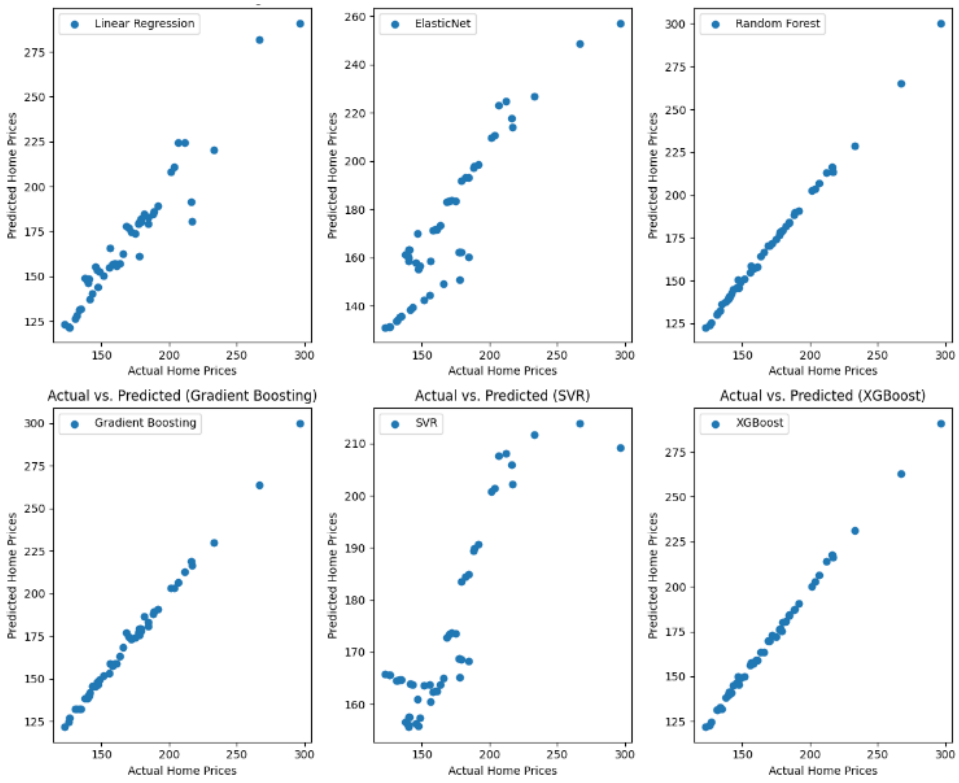
Coefficients and Intercept:

- For linear models, the coefficients and intercept are printed to understand the influence of each feature.
- For non-linear models, feature importance is displayed to see which features have the most impact on predictions.

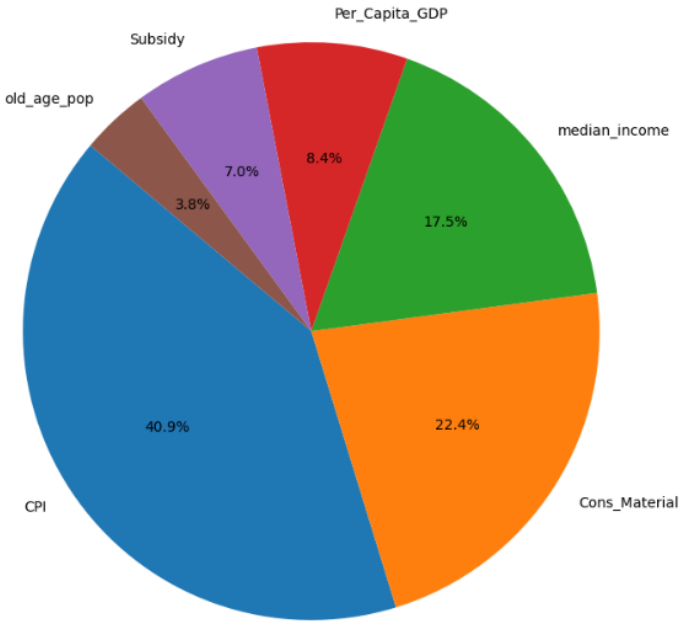
Best Model Selection:

The model with the lowest MSE is identified as the best model.

Visualization: Scatter plots of actual vs. predicted values are created for each model to visually assess how well each model performs.



Factors that has High Feature Importance



- **Consumer Price Index (CPI):**

CPI measures the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services.

It is a direct indicator of inflation and the cost of living. High CPI indicates that prices are rising, which can lead to higher home prices as the cost-of-living increases and people seek to invest in property as a hedge against inflation.

- **Materials (Cons_Material):**

This factor likely represents the cost of materials used in construction. High material costs can directly impact the cost of building new homes, leading to increased home prices.

When the cost of construction materials rises, builders may pass these costs onto buyers, driving up home prices.

CONCLUSION:

The analysis demonstrates how various economic factors, including CPI and material costs, significantly impact home prices. By leveraging models that provide feature importance, we can identify and quantify the influence of each factor on the housing market. Understanding these relationships helps in making informed decisions about real estate investments and policymaking.

Among the evaluated models, **Random Forest** showed comparatively high accuracy. Training and predicting home prices using Random Forest would result in more accurate price predictions due to its ability to handle complex interactions between features and its robustness against overfitting. Therefore, Random Forest is a reliable choice for predicting home prices based on the identified key factors.