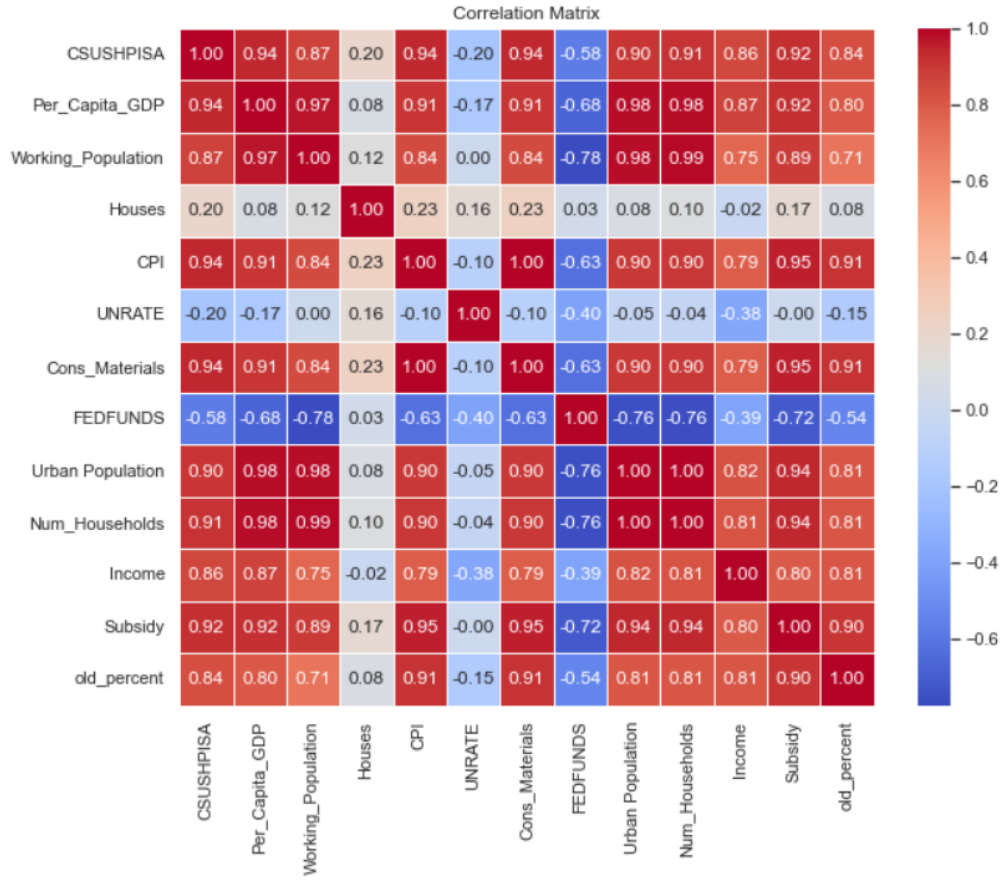
# Comprehensive Analysis of US House Price Prediction

**Introduction:**
This project aims to predict US house prices based on various economic indicators and demographic factors. We aim to explore the relationships between these variables and the Case-Shiller Home Price Index (CSUSHPISA). The analysis includes data preprocessing, correlation analysis, feature selection, and the application of linear regression, Random Forest, and XGBoost models.

**Data Preparation:**
We begin by importing necessary libraries and loading the prepared dataset, excluding unnecessary columns like month and year. The dataset includes variables such as Per Capita GDP, Working Population, Houses, CPI, Unemployment Rate (UNRATE), Federal Funds Rate (FED FUNDS), Urban Population, Number of Households, Income, Subsidy, and old_percent. The synthesis of these variables into a single data frame forms the foundation for our comprehensive analysis.

**Correlation Analysis:**
We analyze the correlation among variables using a correlation matrix and visualize it with a heatmap. Key observations include:

- A negative correlation between the unemployment rate and home prices.
- The unexpectedly low correlation for the number of new houses.
- The impact of the Great Recession on various plots.

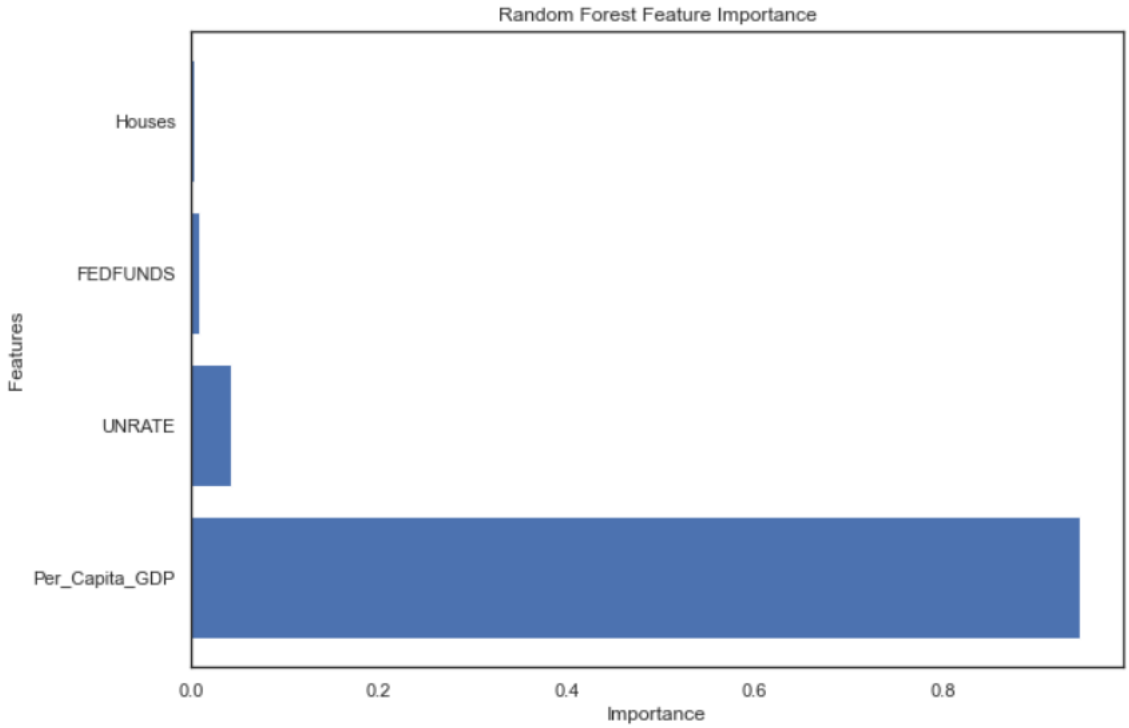


**Multicollinearity Handling:**
Due to the high correlation between some independent variables, we drop certain columns to address multicollinearity. The decision is based on retaining columns with a higher correlation to the target variable.

**Model Building:**
The dataset is scaled, and training and validation sets are divided. A Linear Regression model is deployed, and its performance is evaluated using the R-squared score, which f for the validation set is 0.909. Surprising signs in the regression coefficients prompt an in-depth examination.

To enhance prediction accuracy, we employ Random Forest and XGBoost models. Both models exhibit high R-squared scores, validating their predictive capabilities. Feature importance plots are generated to understand the impact of variables on house prices.

- R-squared for the validation set in case of Random Forest is 0.995
- R-squared for the validation set in case of XGBoost is 0.985




**Insights and Interpretation:**
The analysis highlights the importance of considering non-linear relationships, as evidenced by the disparities between correlation and regression coefficients. The impact of variables on home prices is discussed, including the expected negative effects of the unemployment rate and the unexpected sign in the regression coefficient.

**Future Considerations:**
Potential variables impacting home prices are identified, although data availability limitations hinder a comprehensive analysis. Areas for further exploration include net immigration, marriage rate, average house size, land availability, tax rate, and active listings.

**Conclusion:**
In conclusion, this project thoroughly analyzes US house price prediction, incorporating various models and insights into the relationships between economic indicators and the Case-Shiller Home Price Index. The comparison of linear and non-linear models offers a nuanced understanding of predictive accuracy. Future work could involve acquiring additional data and refining models for more robust predictions.