Q. Find publicly available data for key factors that influence US home prices nationally. Then, build a data science model that explains how these factors impacted home prices over the last 20 years. Use the S&P Case-Schiller Home Price Index as a proxy for home prices: fred.stlouisfed.org/series/CSUSHPISA.

Ans

The Major Key factors that influences US home prices are:

* Economic Change
* Population
* Personal Income
* Employment
* Supply and Demand
* Mortgage Interest rates
* Labor Force Participation Rate
* Locations like Highways, Schools, Hospitals, etc.,
* Government Laws

# Tools used:

* Programming Languages: Python
* Data Analysis Libraries: NumPy, pandas, matplotlib, seaborn
* Machine Learning Libraries: scikit-learn
* Data Visualization: Matplotlib, Seaborn

# Data Collected

* Target (S&P/Case-Shiller U.S. National Home Price Index.)
* Population (Population includes resident population plus armed forces overseas.)
* Personal Income (Income that persons receive in return for their provision of labor, land, and capital used in current production and the net current transfer payments that they receive from business and from government.)
* Mortgage Rate (A mortgage rate is the interest rate charged for a home loan.(Percentage))
* Employment- (Population Ratio (emratio))
* Building Construction issued permit in US (Total Units)
* Labor Force Participation Rate (The participation rate is the percentage of the population that is either working or actively looking for work.)
* Median Sales Price. (Median Sales Price of Houses Sold for the United States.(US Dollars))
* Producer Price Index -Cement Manufacturing
* Producer Price Index by Industry: Concrete Brick
* New Privately-Owned Housing Construction Completed: (Total units in thousands)
* New Privately-Owned Housing Units Under Construction: Total Units in thousands

# Feature Selection

In our analysis, we identified several key features and their correlations with the target variable, represented by the S&P Case-Shiller Home Price Index.

|  |  |
| --- | --- |
| **Feature** | **Correlation with Home Price Index** |
| MSPUS | 0.980146 |
| PPI\_Cement | 0.955074 |
| GDP | 0.860633 |
| income | 0.946608 |
| PPI\_Concrete | 0.985072 |
| population | 0.789974 |
| total\_emp\_cons | 0.511130 |
| new\_private\_hw\_under | 0.633935 |

|  |  |
| --- | --- |
| **Feature** | **Correlation with Home Price Index** |
| all\_Const\_Emp | 0.563479 |
| home\_ow\_rate | 0.209294 |
| monthly\_supply | 0.178504 |
| permit | 0.234189 |
| house\_st | -0.001372 |
| new\_private\_house | -0.068468 |
| unemployed\_rate | -0.249211 |
| IPI\_Cement | -0.260807 |
| p\_saving\_rate | -0.286001 |
| emratio | -0.538776 |
| mortgage\_rate | -0.730709 |
| labor\_percent | -0.788313 |

The positive correlation values indicate a direct relationship with home prices, while negative values suggest an inverse relationship. Features with higher absolute correlation values have a larger impact on home prices.

# Model Selection and Cross-Validation

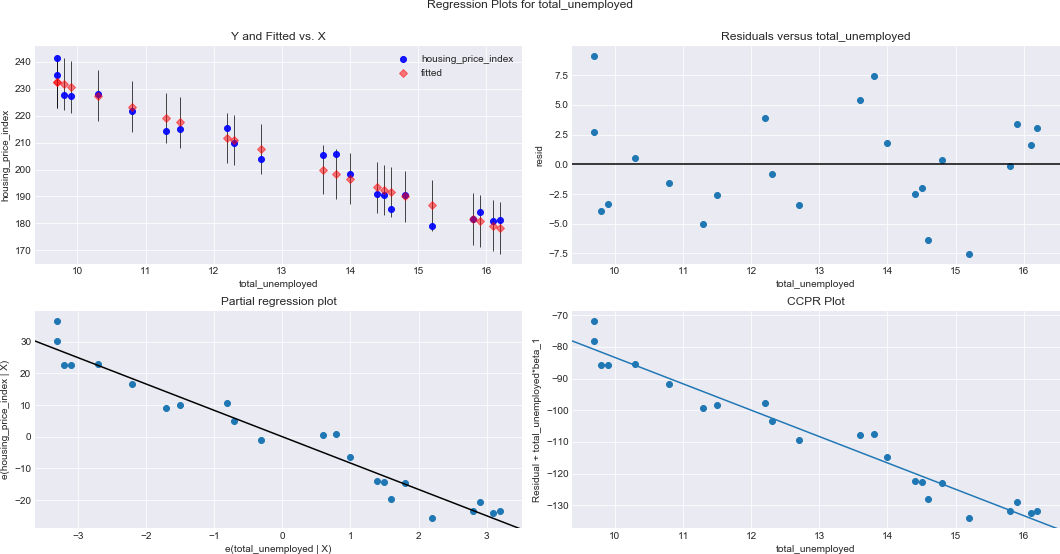
In this project, we employed the **Lasso regression** model due to indications of significant collinearity in the dataset. The Lasso regression model is known for its ability to handle collinearity by applying L1 regularization, which encourages sparsity in feature coefficients.

To optimize the Lasso model's performance and select the best regularization hyperparameter (alpha), we utilized **cross-validation**.

* Optimal alpha: **0.0576**.

The cross-validation process resulted in the following **R-squared** scores for different folds:

* Fold 1: 0.9952
* Fold 2: 0.9939
* Fold 3: 0.9933
* Fold 4: 0.9926
* Fold 5: 0.9932



* **mean R-squared** : **0.9937**
* **standard deviation R squared** : **0.0009**

# Best Features With non-zero Coefficients

Here are the features and their respective coefficients obtained from our Lasso regression model:

|  |  |
| --- | --- |
| **Feature** | **Coefficient** |
| p\_saving\_rate | -0.923351 |
| IPI\_Cement | 0.196152 |
| PPI\_Concrete | 16.893031 |
| population | -8.945503 |
| new\_private\_hw\_under | 4.770749 |
| GDP | 18.499826 |
| house\_st | 2.103615 |
| labor\_percent | 1.501053 |
| all\_Const\_Emp | 10.887265 |
| monthly\_supply | 3.212566 |
| unemployed\_rate | 7.251776 |
| MSPUS | 28.943474 |

These coefficients represent the impact of each feature on the prediction of home prices. Positive coefficients indicate a direct relationship with home prices, while negative coefficients suggest an inverse relationship.