Report on Factors Influencing US Home Prices

Problem Statement-

The objective of this report is to analyze the key factors that influence home prices in the United States on a national scale. To accomplish this, we collected and analyzed various economic and demographic data over the last 20 years to understand the relationships and impacts of different variables on home prices.

Data Collection and Preparation-

We began by collecting relevant data from the Federal Reserve Economic Data (FRED). The datasets used in this analysis are as follows:

1. **Case-Shiller Home Price Index (df_hp):** This dataset provides information on the national home price index.
2. **Unemployment Rate Data (df_unemp):** This dataset contains data on the national unemployment rate.
3. **Gross Domestic Product Data (df_gdp):** This dataset offers information on the GDP of the United States.
4. **Personal Disposable Income Data (df_rdpi):** This dataset provides data on personal disposable income in the country.
5. **Population Data (df_population):** This dataset gives insights into the population growth in the United States.
6. **Homeownership Rate Data (df_homeown):** This dataset contains data on the national home ownership rate.
7. **New House Permit Data (df_hpermit):** This dataset provides information on permits for new house construction.
8. **Privately Owned Housing Data (df_private_owned_housing):** This dataset contains data on privately owned housing units.
9. **Federal Funds Rate Data (df_fed_funds):** This dataset provides information on the federal funds rate.

After collecting the data, we performed several data preparation steps:

- We filtered the data to include only the last 20 years, focusing on more recent trends.
- We adjusted the date data type to ensure consistency across datasets.
- For datasets with quarterly data (GDP and homeownership rate), we interpolated the data to create monthly values for better alignment.
- We merged all datasets based on the date column to create a unified dataset for analysis.

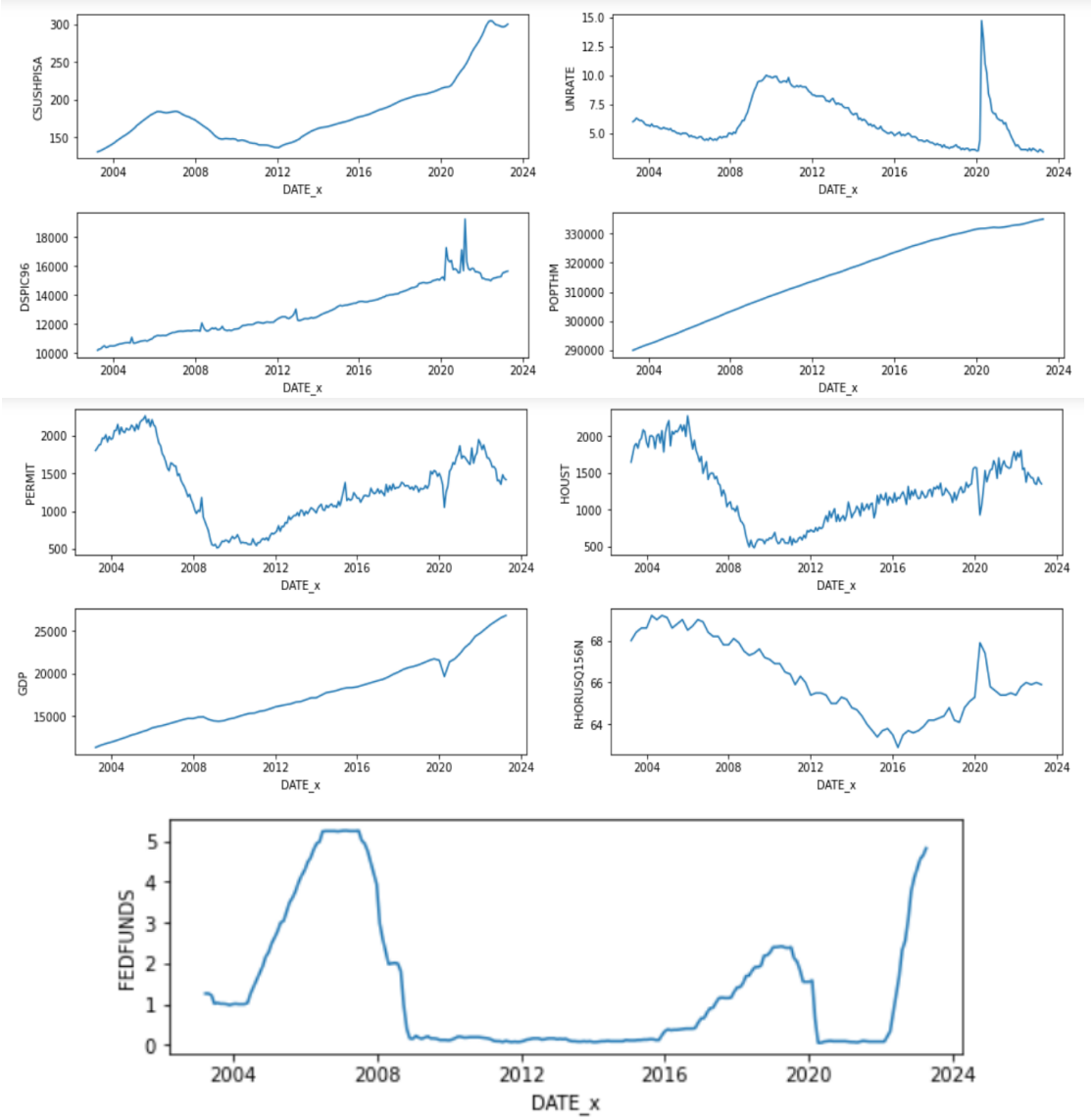
Exploratory Data Analysis (EDA)-

In the EDA phase, we conducted the following analyses:

- Checked for null values, and fortunately, there were no missing data points.
- Explored the distribution of each variable.

	CSUSHPISA	UNRATE	DSPIC96	POPTHM	PERMIT	HOUST	FEDFUNDS	GDP	RHORUSQ156N
count	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000
mean	183.246315	5.971784	12911.736515	315461.800830	1305.041494	1244.759336	1.356929	17479.196398	66.166390
std	43.417141	2.055691	1739.220123	13624.981944	475.866918	456.576180	1.630804	3864.048016	1.814167
min	130.884000	3.400000	10181.400000	290024.000000	513.000000	478.000000	0.050000	11312.766000	62.900000
25%	148.409000	4.500000	11538.600000	303926.000000	980.000000	917.000000	0.120000	14549.105667	64.666667
50%	173.830000	5.400000	12431.500000	316535.000000	1291.000000	1206.000000	0.400000	16699.551000	65.900000
75%	200.655000	7.300000	14275.900000	328364.000000	1644.000000	1561.000000	2.130000	20260.389667	67.900000
max	304.817000	14.700000	19213.900000	334880.000000	2263.000000	2273.000000	5.260000	26798.605000	69.200000

- Conducted Time - Series analysis by plotting line graphs for each factor with respect to time:



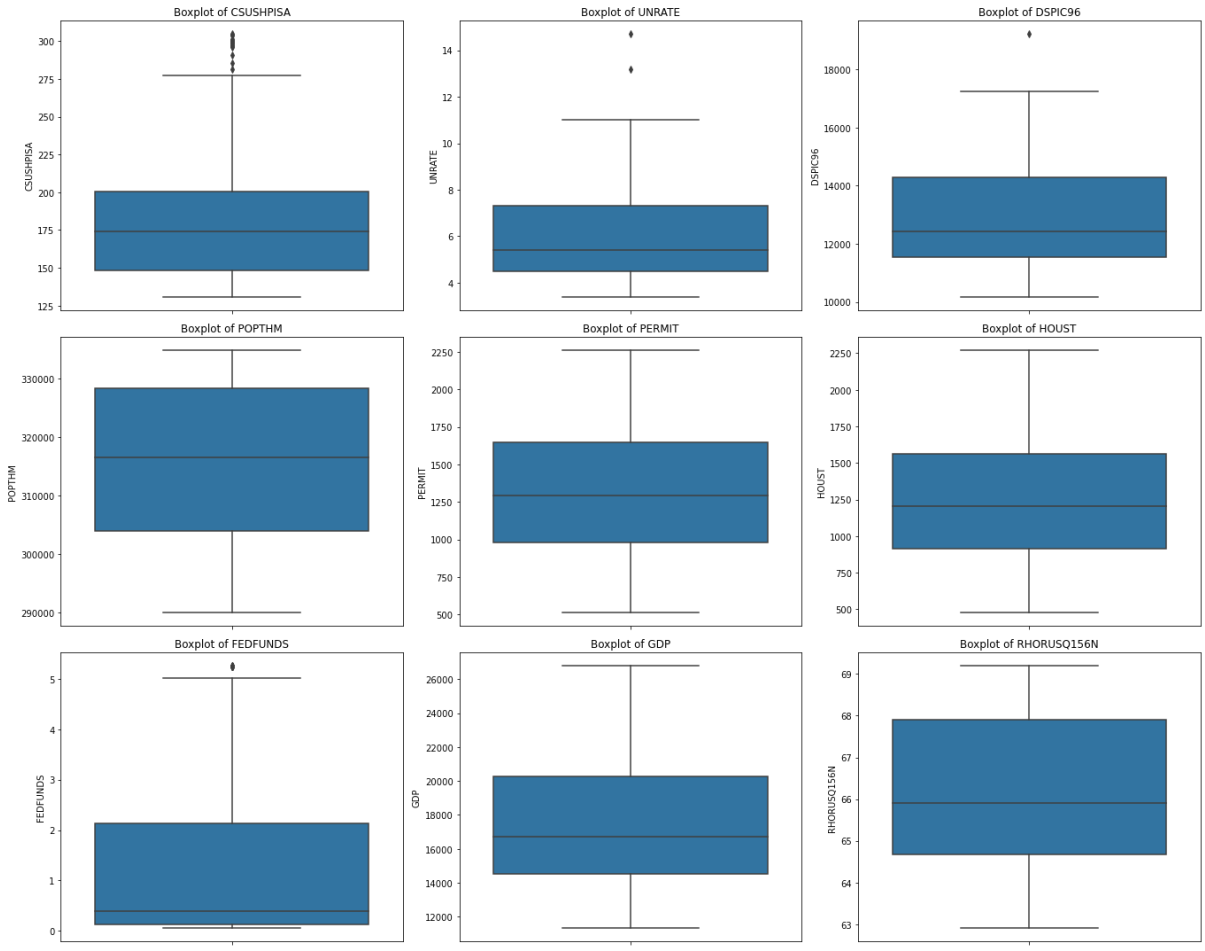
- Home prices declined from 2007 to 2012 but have been steadily increasing, with a substantial jump in prices since 2020.
- Unemployment rates has increased during 2008 and spiked during 2020.
- Daily disposable income saw significant fluctuations in 2020 and after.
- Population shows a positive trend.

- Permits for new housing construction decreased around 2008 and 2020.
- Privately owned housing units followed a similar pattern as permits.
- GDP experienced a downturn around 2020.
- Homeownership rates declined until 2016 and then began increasing, with peaks around 2020 and 2021.
- Federal funds rates decreased around 2008, increased from 2016 to 2020.

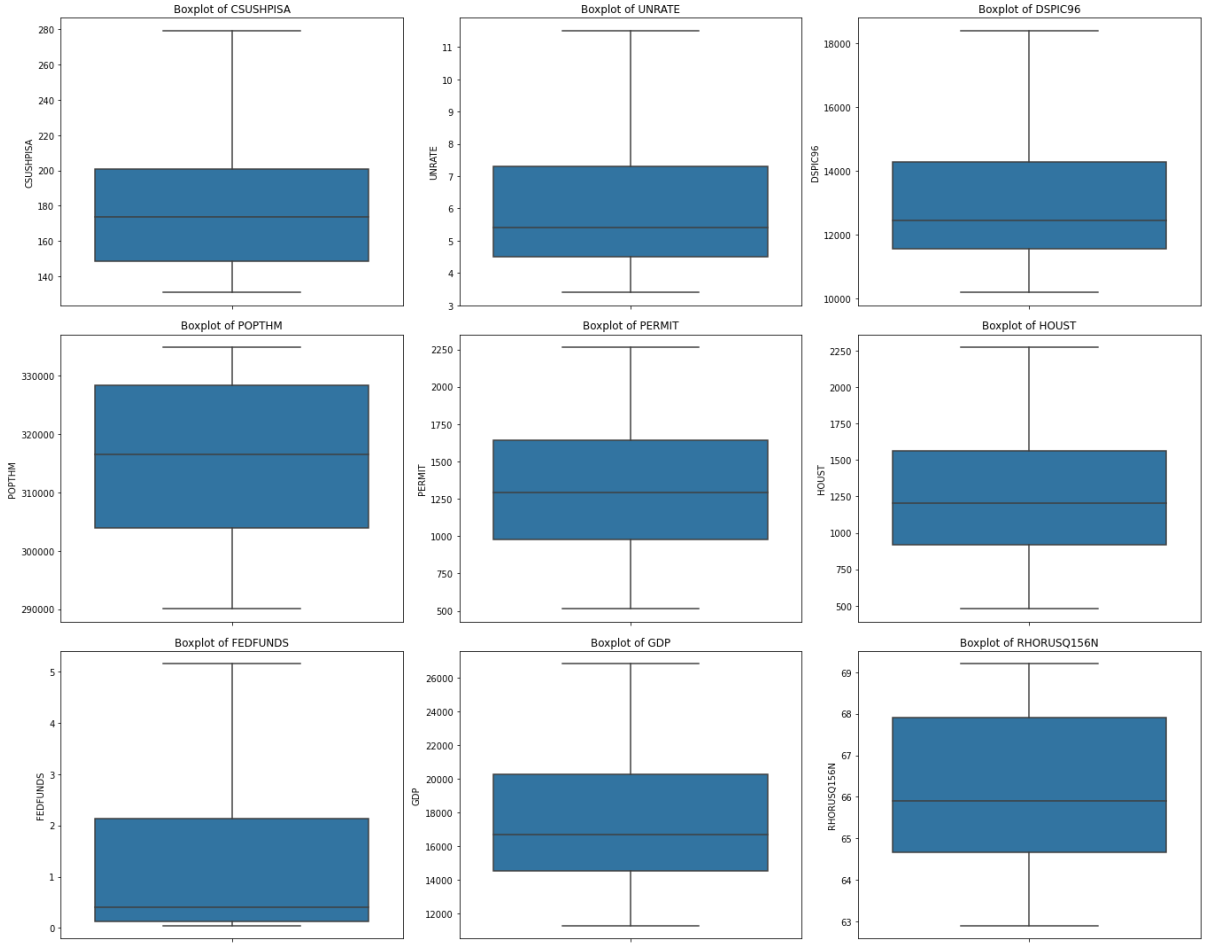
Outlier Removal-

We also examined box plots and identified outliers in home prices, unemployment rates, daily disposable income, and federal funds rates. These outliers were treated to ensure the robustness of the analysis.

Boxplot before outlier removal-

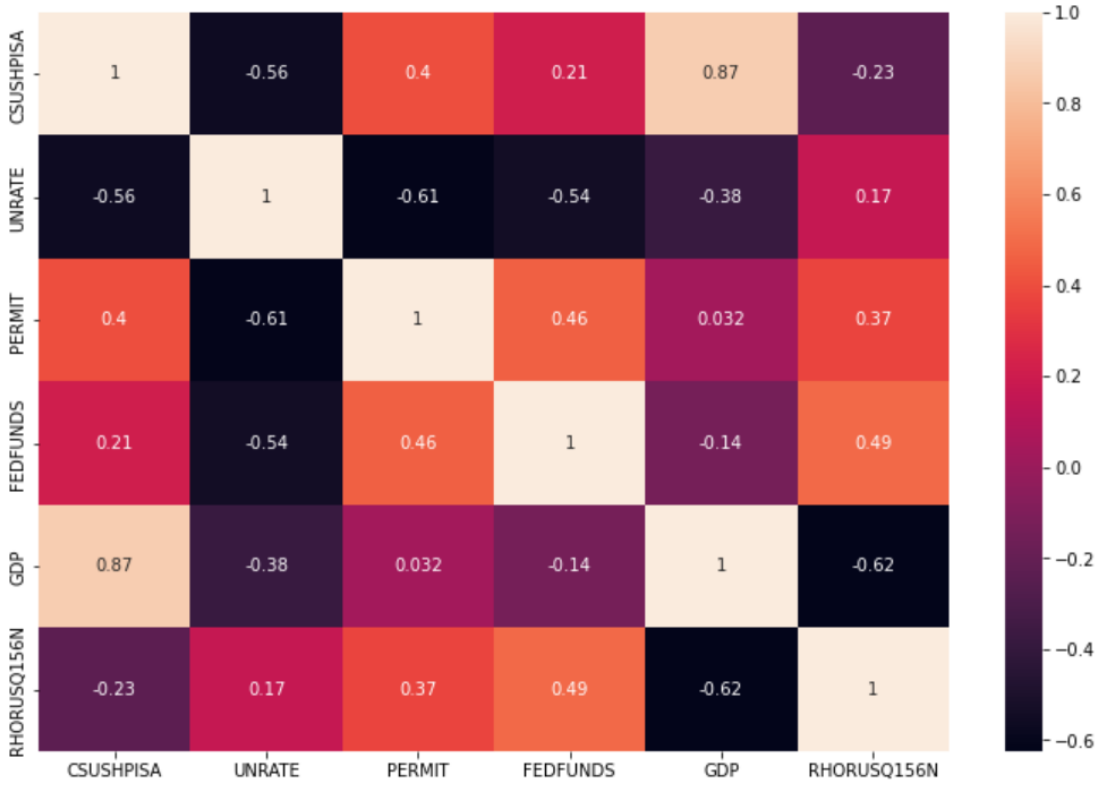


Boxplot after Outlier Removal-



Heatmap-

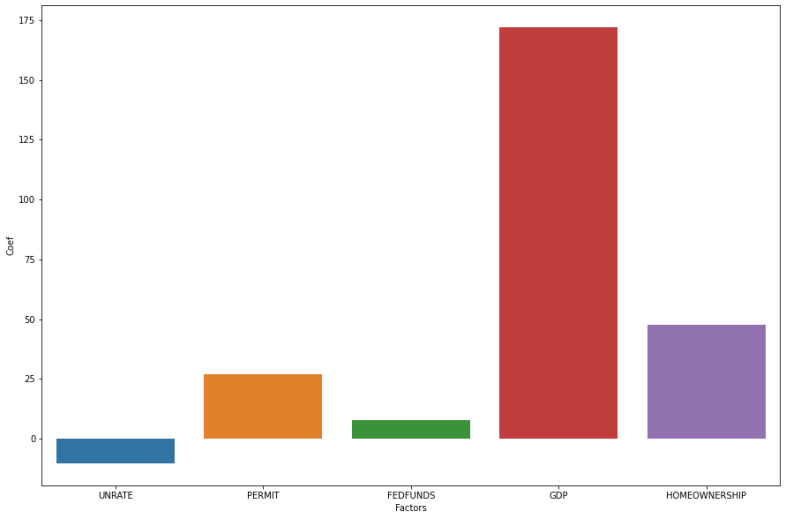
Additionally, we checked for multicollinearity using a heatmap and observed high correlations between GDP, daily disposable income, and population, as well as between New House Permit and privately owned housing units. To address multicollinearity, we dropped daily disposable income, population, and New House Permit, which improved model stability.



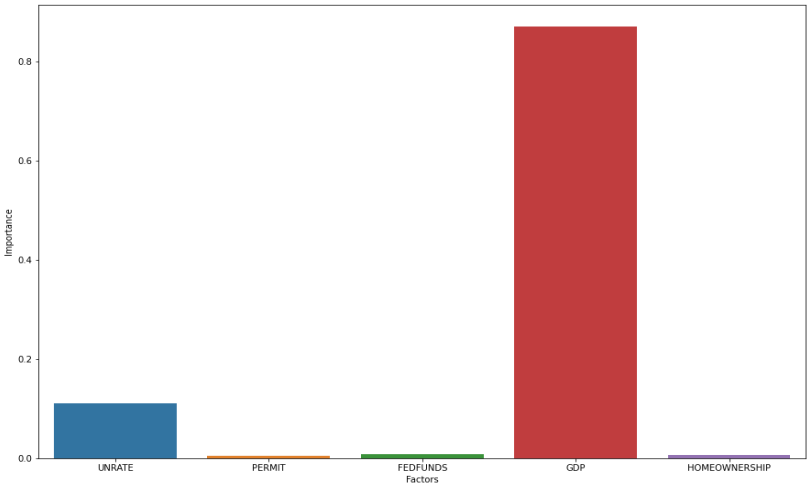
Regression Modelling

For the regression analysis, we separated the independent and dependent variables, with home prices as the dependent variable (y) and the remaining factors as independent variables (X). We scaled the independent data using Min-Max scaling.
We split the data into training and testing sets and applied two regression models:

1. **Linear Regression**
    - Training MSE: 51.503262869326655
    - Training R2: 0.9706806696044044
    - Test MSE: 52.17480649834704
    - Test R2: 0.9569528287493424
    - Coefficient: [-10.37978413, 26.87717444, 7.69221594, 172.03681507, 47.78419082]



2. **Random Forest Regression**
    - Training MSE: 2.325654233610436
    - Training R2: 0.9986760717464805
    - Test MSE: 4.384726079762049
    - Test R2: 0.996382352573771
    - Feature Importance: [0.11054038, 0.00498099, 0.00725238, 0.87003537, 0.00719089]



Conclusion-

The coefficients in a Linear Regression model indicate how each independent variable affects the dependent variable.

1. **"UNRATE" (Unemployment Rate):**
    - Coefficient: -10.37978413
    - Effect: For each one-unit increase in the unemployment rate, house prices in the USA are estimated to decrease by approximately 10.38 units. Higher unemployment rates are associated with lower house prices.
2. **"PERMIT" (Housing Permits Issued):**
    - Coefficient: 26.87717444
    - Effect: For each one-unit increase in the number of housing permits issued, house prices in the USA are estimated to increase by approximately 26.88 units. More housing permits being issued is associated with higher house prices.
3. **"FEDFUNDS" (Federal Funds Rate or Interest Rate):**
    - Coefficient: 7.69221594
    - Effect: For each one-unit increase in the Federal Funds Rate (interest rate), house prices in the USA are estimated to increase by approximately 7.69 units. This suggests that higher interest rates are associated with higher house prices.
4. **"GDP" (Gross Domestic Product):**
    - Coefficient: 172.03681507
    - Effect: For each one-unit increase in GDP, house prices in the USA are estimated to increase by approximately 172.04 units. A stronger economy, as indicated by higher GDP, is associated with higher house prices.
5. **"HOMEOWNERSHIP" (Homeownership Rate):**
    - Coefficient: 47.78419082
    - Effect: For each one-unit increase in the homeownership rate, house prices in the USA are estimated to increase by approximately 47.78 units. A higher rate of homeownership is associated with higher house prices.

The Random Forest model, "GDP" stands out as the most important predictor of house prices, while "UNRATE," "PERMIT," "FEDFUNDS," and "HOMEOWNERSHIP" have comparatively lower importance scores. These scores indicate the relative contribution of each independent variable to the model's Predictions.

Model Comparison-

	Model	MSE Train	R2 Train	MSE Test	R2 Test
0	linear_reg	51.5033	0.970681	52.1748	0.956953
1	random_forest	2.32565	0.998676	4.38473	0.996382

Random Forest Regression appears to be the more effective model for predicting the target variable compared to Linear Regression. It exhibits better performance in terms of both training and testing error metrics, suggesting its suitability for the predictive task.