FACTORS AFFECTING USA NATIONAL HOME PRICES

TEAM 9

Avinav Sahni

Mansi Pathak

Han Pham

Takahiro Sasaki

Komal Ghazanfar

Yang Gao

Overview

- Summary Statistic
- Correlation Analysis
- Time Series
- Pairplot
- Histograms and Kernel
- Violin Plot
- CASE-SHILLER Index vs. Features
- Absolute Correlation
- Table Analysis
- Testing and Split
- Model Evaluation Metrics
- Hyperparameter and Grid Search
- PCA of Economic Indicator
- Conclusions

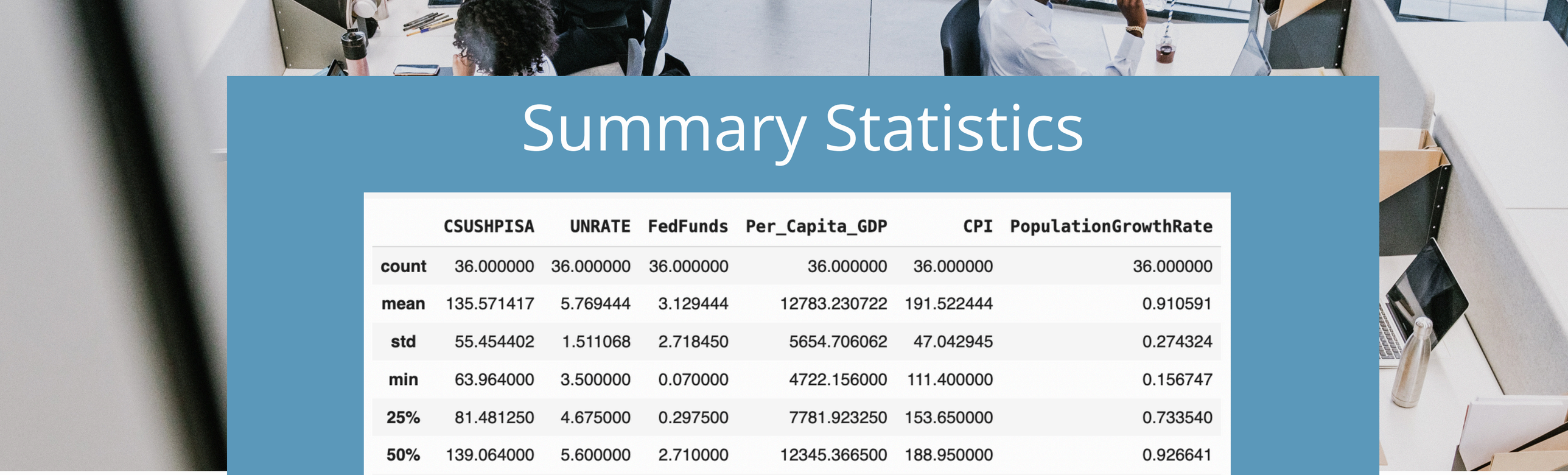
Introduction

Welcome to our presentation on the "Analysis of Factors Affecting USA National Home Prices." This project is part of our coursework, aimed at applying the concepts and techniques learned in class to real-world data analysis problems.

Objective:
The primary objective of this project is to explore and analyze the factors influencing national home prices in the USA using the Mortgage.csv dataset from Kaggle. We aim to identify key trends, patterns, and relationships within the data.

Approach:

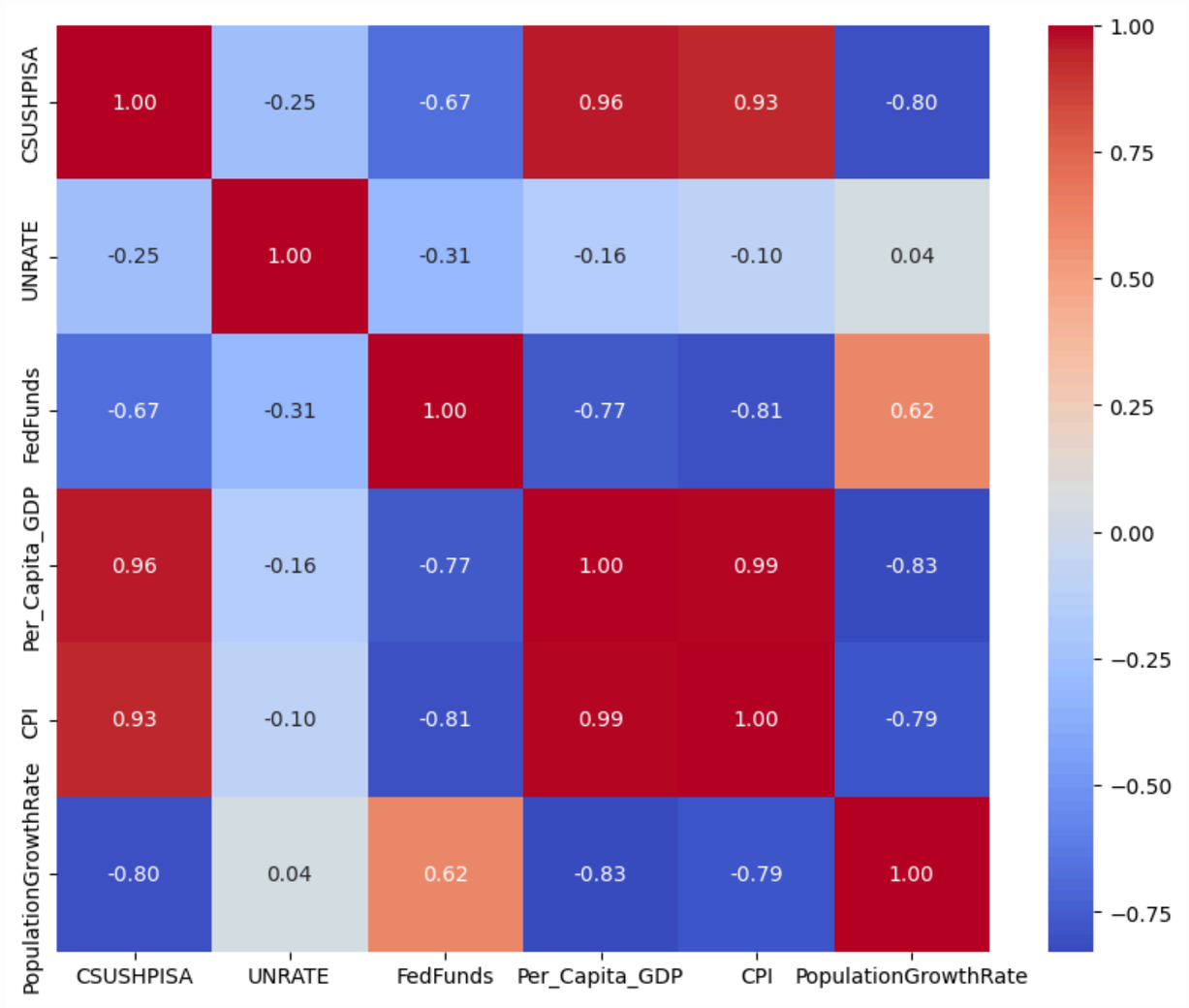
- **Descriptive Statistics**: We begin by generating descriptive statistics to understand the distribution and key characteristics of the dataset.
- **Data Visualization:** Using Plotly, we create visualizations to explore and present the relationships within the data, providing a clearer understanding of the influencing factors.
- **Advanced Techniques:** We apply machine learning techniques, specifically Random Forest for regression, to predict house prices based on the identified factors. Additionally, we explore the use of Reinforcement Learning for decision-making processes where relevant.



Summary Statistics

	CSUSHPISA	UNRATE	FedFunds	Per_Capita_GDP	CPI	PopulationGrowthRate
count	36.000000	36.000000	36.000000	36.000000	36.000000	36.000000
mean	135.571417	5.769444	3.129444	12783.230722	191.522444	0.910591
std	55.454402	1.511068	2.718450	5654.706062	47.042945	0.274324
min	63.964000	3.500000	0.070000	4722.156000	111.400000	0.156747
25%	81.481250	4.675000	0.297500	7781.923250	153.650000	0.733540
50%	139.064000	5.600000	2.710000	12345.366500	188.950000	0.926641
75%	174.167250	6.600000	5.470000	16785.576250	232.446000	1.134323
max	285.829000	9.800000	9.120000	25029.116000	282.599000	1.386886

These statistics show that major indicators in the economy fluctuated significantly over time. House prices, federal funds rate, and per capita GDP vary widely, indicating economic boom and recession. Due to moderate variability in unemployement, CP and population growth rates, employement and population dynamics appear stable despite an economic shift.



Correlation Matrix

Correlation Analysis

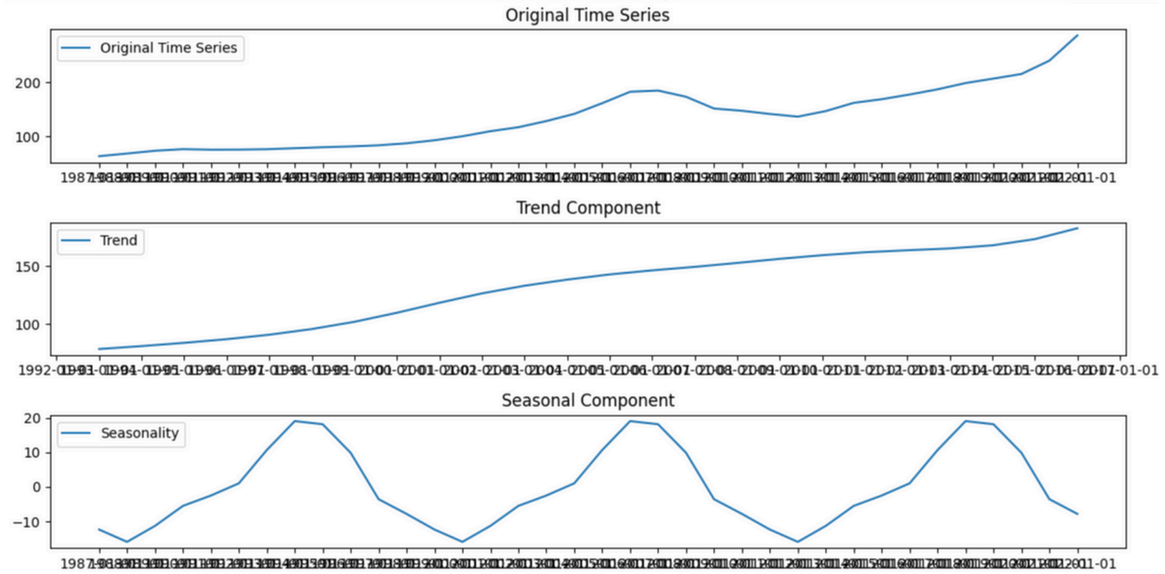
```
# Correlation matrix
corr_matrix = us_house_price_df.corr()
corr_matrix
```

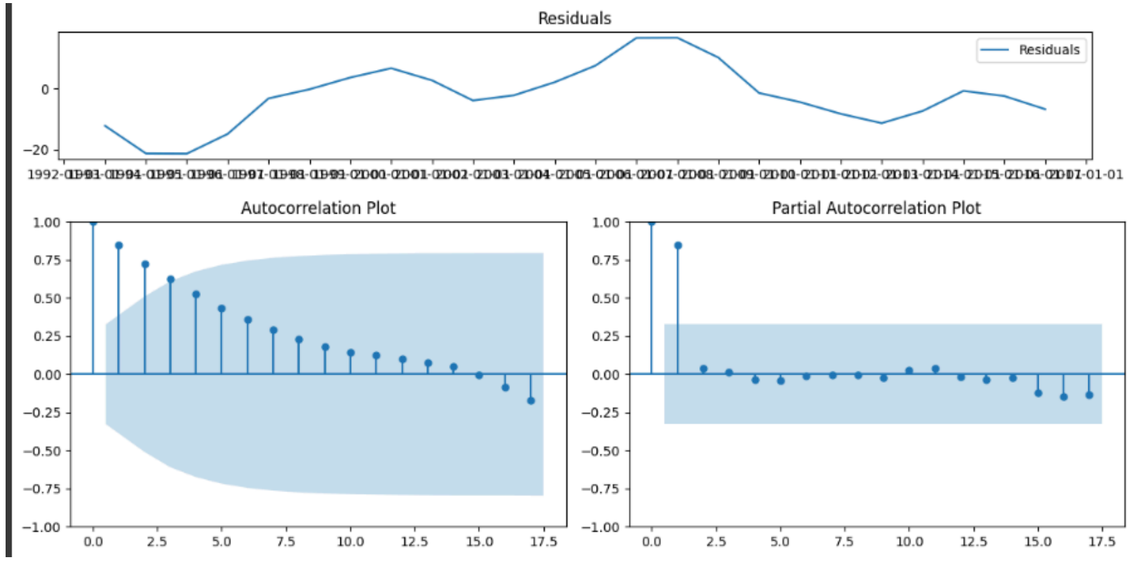
	CSUSHPISA	UNRATE	FedFunds	Per_Capita_GDP	CPI	PopulationGrowthRate
CSUSHPISA	1.000000	-0.250971	-0.673685	0.957521	0.934890	-0.801698
UNRATE	-0.250971	1.000000	-0.313140	-0.160969	-0.095973	0.040410
FedFunds	-0.673685	-0.313140	1.000000	-0.770299	-0.812384	0.620417
Per_Capita_GDP	0.957521	-0.160969	-0.770299	1.000000	0.991344	-0.828332
CPI	0.934890	-0.095973	-0.812384	0.991344	1.000000	-0.790102
PopulationGrowthRate	-0.801698	0.040410	0.620417	-0.828332	-0.790102	1.000000

The correlation matrix reveals several key relationships between economic indicators

- Higher house price = better economic conditions
  - High correlation between capita GDP and CPI
- Unemployment rates show weak inverse correlations with economic growth indicator
- Higher federal funds rate correlated with lower per capita GDP and CPI
- Per capita GDP and CPI show a near-perfect positive correlation
  - infer as economic growth drives higher consumer prices

Time Series Decomposition





Original Time Series:
Show an overall house price index over time. There is an noticeable upward trend, indicating that house prices have generally increased over the period. There are also fluctuations that suggest some seasonality and variability

Trend Component:
The graph reveals teh long-term movement in teh house prices, and smoothing out short-term fluctuations. It clearly shows a steady increase over the years, reflecting a long-term growth house prices. The dip around mid-200s, which likely is due to the housing market crash during the financial crisis, followed by a recovered and continued growth

Seasonal Component:
The seasonal component shows repeating pattersn within each year. It captures the cyclical nature of data, highlight the periods within the year where house prices tend to increase or decrease. The graph has consistent peaks and troughs, indicating regulat seasonal variations in house prices

Time Series Decomposition

Residuals:
The residuals represent the random noise left after removing the trend and seasonal components. Ideally, they should resemble white noise with no discernible pattern. However, the presence of patterns suggests additional underlying factors influencing house prices that are not captured by the trend and seasonal components alone.

Overall Analysis:
The decomposition analysis reveals the components influencing house prices. The increasing trend reflects long-term growth, while the seasonal component shows regular intra-year fluctuations. Residuals suggest additional variability, possibly due to external factors. This helps in understanding the time series structure and forecasting future values.

```
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Time Series Decomposition
decomposition = seasonal_decompose(us_house_price_df['CSUSHPISA'], model='additive', period=12)
trend = decomposition.trend
seasonal = decomposition.seasonal
residual = decomposition.resid

# Plot Time Series Components
plt.figure(figsize=(12, 8))

plt.subplot(4, 1, 1)
plt.plot(us_house_price_df['CSUSHPISA'], label='Original Time Series')
plt.legend()
plt.title('Original Time Series')

plt.subplot(4, 1, 2)
plt.plot(trend, label='Trend')
plt.legend()
plt.title('Trend Component')

plt.subplot(4, 1, 3)
plt.plot(seasonal, label='Seasonality')
plt.legend()
plt.title('Seasonal Component')

plt.subplot(4, 1, 4)
plt.plot(residual, label='Residuals')
plt.legend()
plt.title('Residuals')

plt.tight_layout()
plt.show()

# Autocorrelation and Partial Autocorrelation Plots
plt.figure(figsize=(12, 4))

# Autocorrelation Plot
plt.subplot(1, 2, 1)
plot_acf(us_house_price_df['CSUSHPISA'], lags=17, ax=plt.gca(), title='Autocorrelation Plot')

# Partial Autocorrelation Plot
plt.subplot(1, 2, 2)
plot_pacf(us_house_price_df['CSUSHPISA'], lags=17, ax=plt.gca(), title='Partial Autocorrelation Plot')

plt.tight_layout()
plt.show()
```

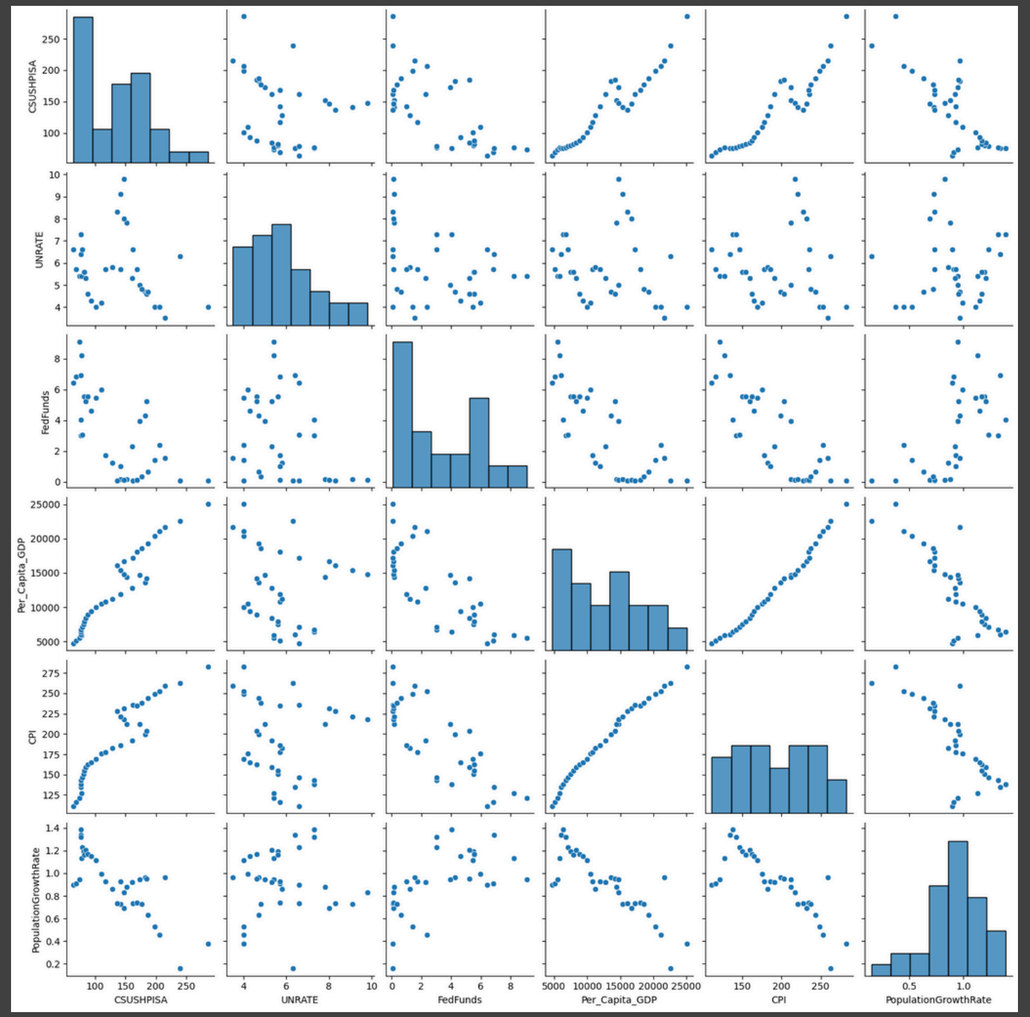
Pair Plot Visualization

CSUSHPISA (House Price Index):
• Positive linear relationship with Per Capita GDP and CPI
• Higher house prices associated with higher Per Capita GDP and consumer prices
• Negative relationship with Population Growth Rate
• Higher house prices associated with lower population growth rates
• Scatter plots with UNRATE and FedFunds show no clear linear relationships, suggesting complex interactions

UNRATE (Unemployment Rate):
• Scatter plots with other variables show no clear linear relationships, indicating weak or complex interactions
• Histogram indicates an unimodal distribution with values clustered around the mean

FedFunds (Federal Funds Rate):
• Negative relationship with Per Capita GDP and CPI
• Higher federal funds rates associated with lower economic growth and consumer prices
• Positive relationship with Population Growth Rate
• Higher federal funds rates associated with higher population growth rates
• Histogram shows a bimodal distribution, indicating two distinct periods of federal funds rates.

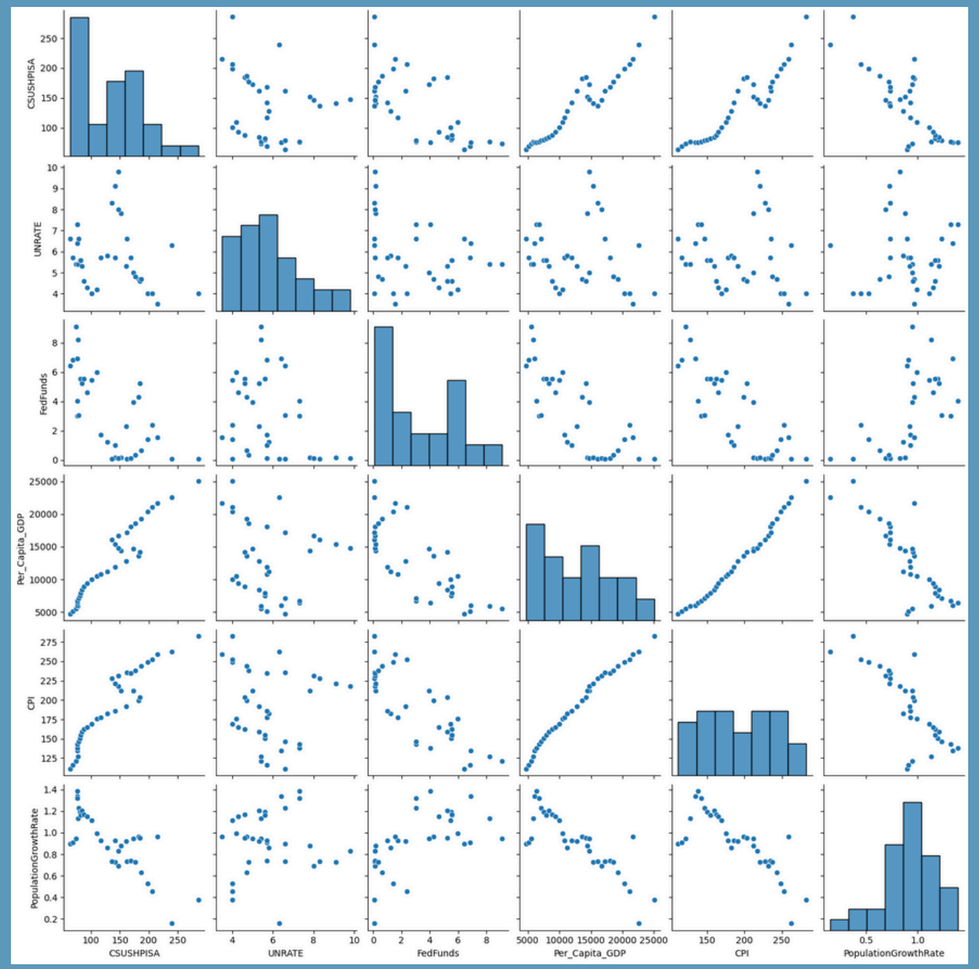
Per Capita GDP:
·Strong positive linear relationship with CPI
·Higher Per Capita GDP associated with higher consumer prices
·Negative relationship with Population Growth Rate
·Higher Per Capita GDP associated with lower population growth rates
·Histogram indicates a rightskewed distribution, with more lower GDP values



Pair Plot Visualization

CPI (Consumer Price Index):
·Strong correlation with Per Capita GDP
·Higher consumer prices associated with higher Per Capita GDP
·Negative relationship with Population Growth Rate
·Higher consumer prices associated with lower population growth rates
·Histogram shows a unimodal distribution, peaking around the mean value

Population Growth Rate:
·Negative relationships with CSUSHPISA, Per Capita GDP, and CPI
·Higher population growth rates associated with lower house prices, GDP, and consumer prices
·Positive relationship with FedFunds
·Higher population growth rates associated with higher federal funds rates
·Histogram indicates a rightskewed distribution, with values clustered towards the lower end

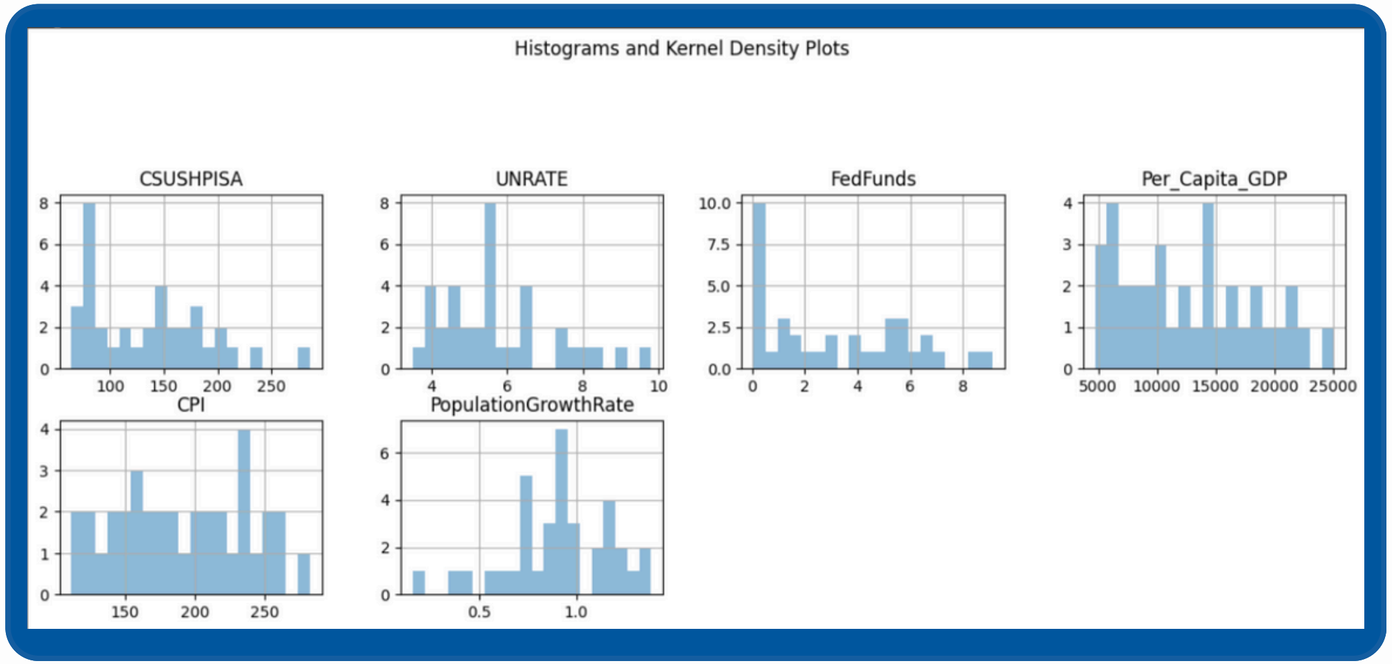


The pair plot highlights strong positive correlations between house prices, per capita GDP, and consumer prices, as well as negative relationships between population growth rates and these variables. The federal funds rate shows distinct patterns with economic growth and population dynamics, indicating its role in monetary policy.

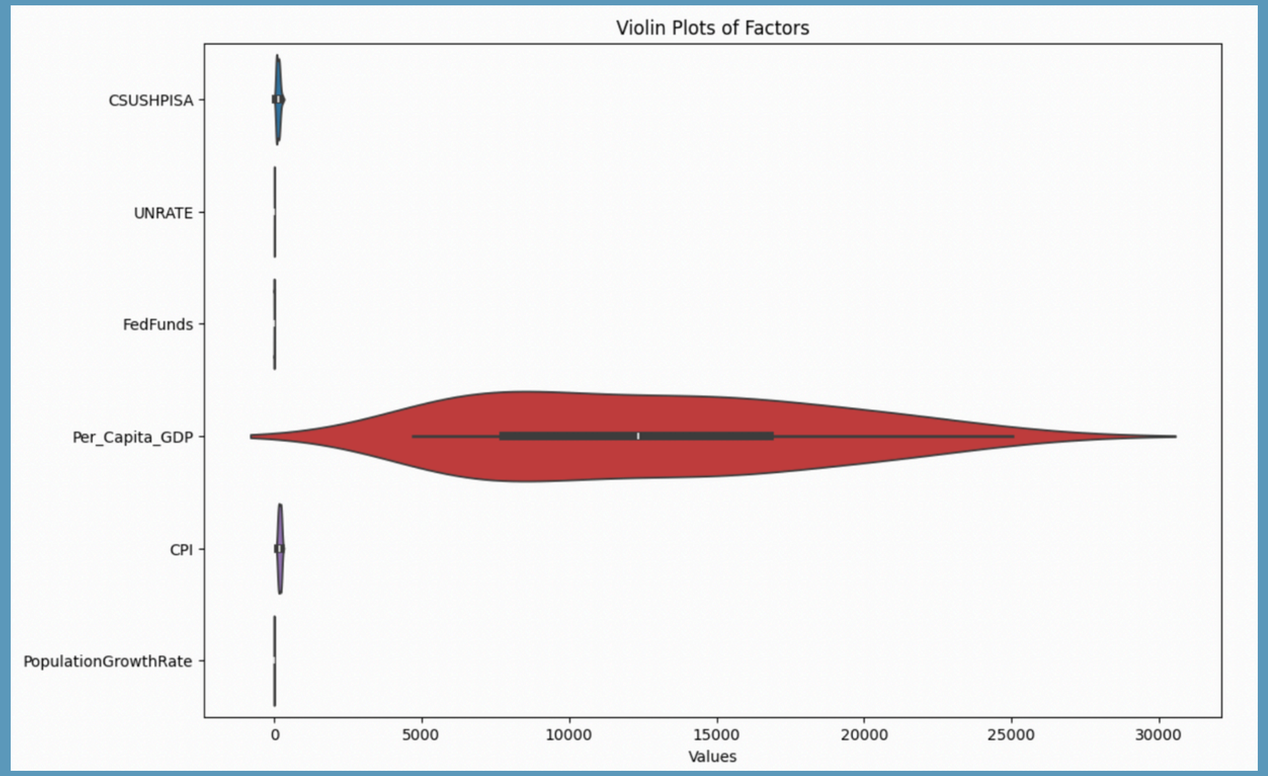
Histograms and Kernel Density

- House Price Index (CSUSHPISA): Right-skewed distribution, higher house prices are less common
- Unemployment Rate (UNRATE): Bimodal distribution with peaks around 4% and 6%, indicating distinct economic periods
- Federal Funds Rate (FedFunds): Heavily right-skewed, reflecting periods of low interest rates
- Per Capita GDP: Wide spread with a slight right skew, most values between 10,000 and 15,000
- Consumer Price Index (CPI): Fairly uniform distribution, indicating varied inflation levels
- Population Growth Rate: Right-skewed, most values below 1.0, high growth rates are rare
- These patterns highlight the variability and asymmetries in the data.

```
plt.figure(figsize=(13, 8))
us_house_price_df[factors].hist(bins=20, alpha=0.5, layout=(4, 4), figsize=(15, 10))
plt.suptitle('Histograms and Kernel Density Plots', y=1.02)
plt.show()
```



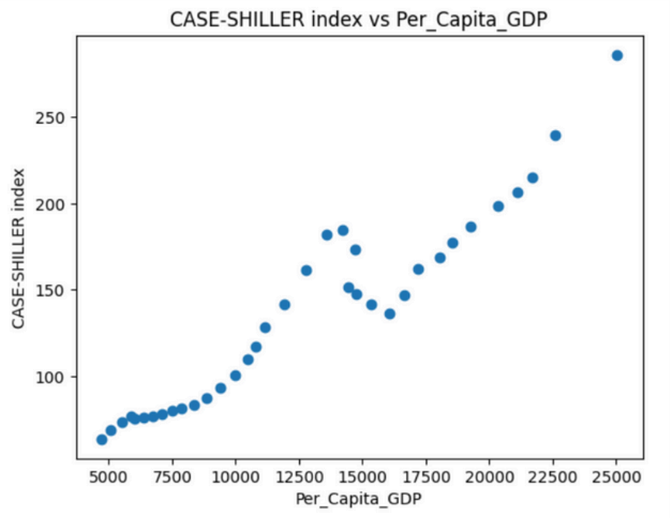
Violin Plots



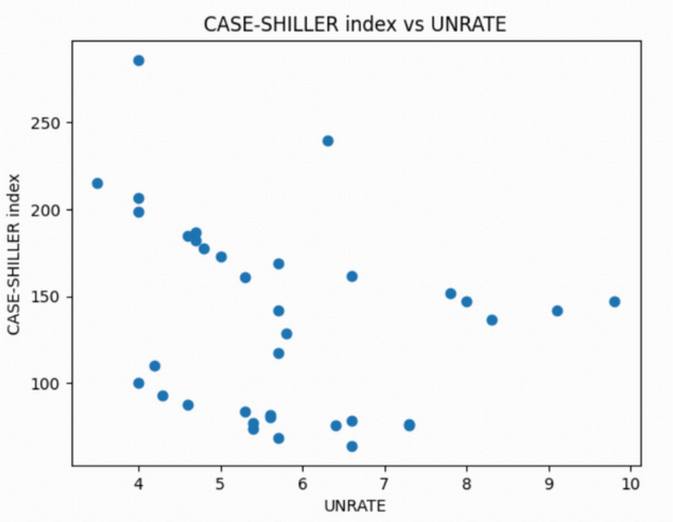

```
# Violin Plots
plt.figure(figsize=(12, 8))
sns.violinplot(data=us_house_price_df[factors], orient='h')
plt.title('Violin Plots of Factors')
plt.xlabel('Values')
plt.show()
```

- House Price Index: Concentration of lower values with fewer higher values; right-skewed distribution.
- Unemployment Rate (UNRATE): Bimodal distribution; suggests two distinct economic periods.
- Federal Funds Rate (FedFunds): Heavily skewed towards lower values; reflects frequent periods of low interest rates.
- Per Capita GDP: Broad distribution; most values concentrated around the median.
- Consumer Price Index (CPI): Fairly uniform distribution; indicates varied inflation levels.
- Population Growth Rate: Skewed towards lower values; low growth rates are more common.

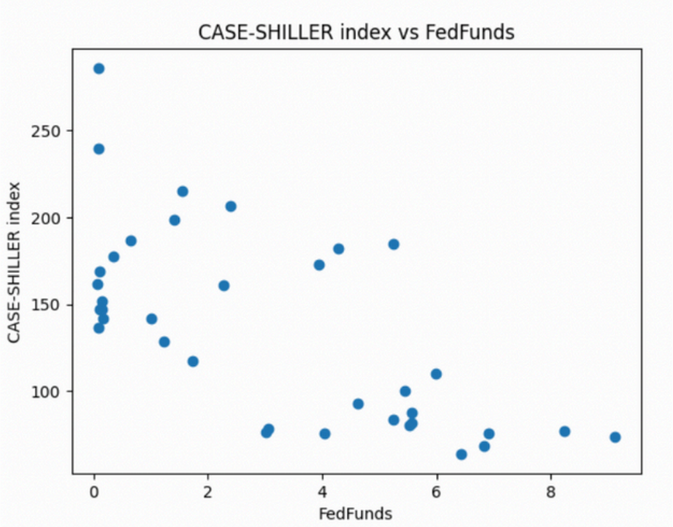
CASE-SHILLER index vs. features



CASE-SHILLER index vs Per_Capita_GDP:
A strong positive relationship is visible between per capita GDP and house prices. As per capita GDP increases, the CASE-SHILLER index also rises, suggesting that higher economic prosperity (as measured by per capita GDP) is associated with higher house prices.



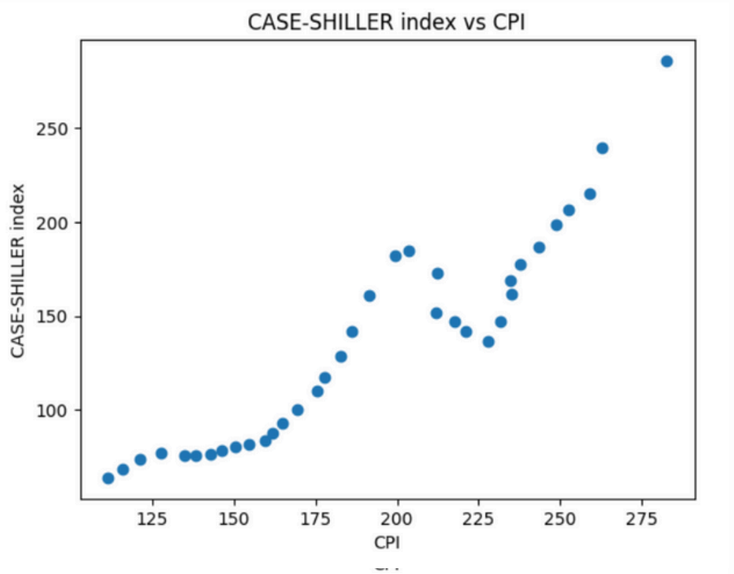
CASE-SHILLER index vs UNRATE (Unemployment Rate):
The scatter plot shows a generally negative relationship. Higher unemployment rates tend to be associated with lower house prices. However, the relationship is not perfectly linear, and there is considerable spread in the data, indicating other factors may also influence house prices.



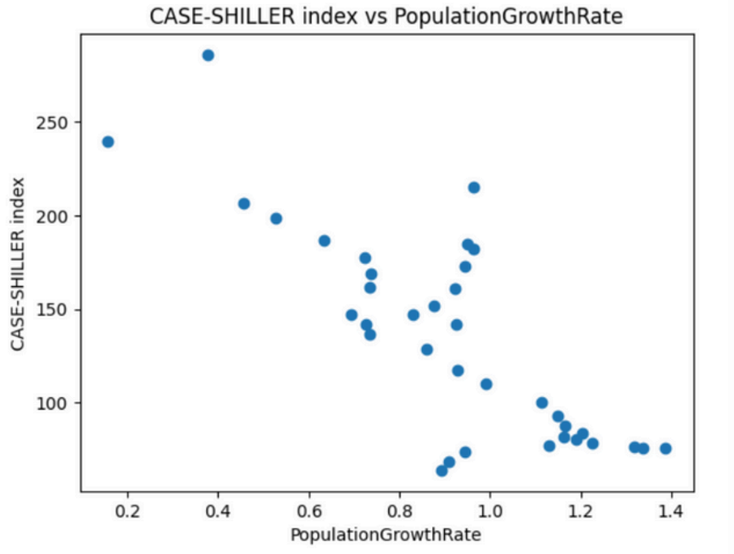
CASE-SHILLER index vs FedFunds (Federal Funds Rate):
This plot indicates a negative relationship between the federal funds rate and the CASE-SHILLER index. As the federal funds rate increases, the house prices tend to decrease. This inverse relationship suggests that higher interest rates might suppress house price growth, likely due to higher borrowing costs.

CASE-SHILLER index vs. features

CASE-SHILLER index vs CPI (Consumer Price Index):
The scatter plot shows a clear positive relationship. As the CPI increases, indicating higher inflation, the CASE-SHILLER index also increases. This suggests that inflationary pressures are associated with higher house prices, possibly due to increased costs of materials and labor.



CASE-SHILLER index vs PopulationGrowthRate:
There is a negative relationship between population growth rate and the CASE-SHILLER index. Higher population growth rates are associated with lower house prices. This could indicate that areas with rapidly growing populations might experience more affordable housing prices or other factors affecting housing demand and supply.



Absolute Correlation

- Unemployment Rate (UNRATE): Weakest relationship with CASE-SHILLER index; minimal impact on house prices.
- Federal Funds Rate (FedFunds): Moderate to strong inverse relationship; higher rates associated with lower house prices.
- Population Growth Rate: Moderate to strong inverse relationship; higher growth associated with lower house prices.
- Consumer Price Index (CPI): Very high correlation with CASE-SHILLER index; higher inflation linked to higher house prices.
- Per Capita GDP: Very high correlation with CASE-SHILLER index; economic prosperity linked to higher house prices.
- Conclusion: Economic growth and inflation are significant drivers of house prices, while unemployment plays a lesser role.

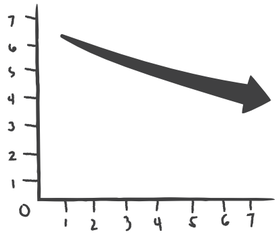
```
correlations = X.apply(lambda column: np.abs(column.corr(y)))

# Sort correlations in ascending order
sorted_correlations = correlations.sort_values()

# Display features with lower correlation
print("Features with Lower Correlation to Target:")
print(sorted_correlations)
```

```
Features with Lower Correlation to Target:
UNRATE                  0.250971
FedFunds                0.673685
PopulationGrowthRate    0.801698
CPI                     0.934890
Per_Capita_GDP          0.957521
dtype: float64
```

Table Analysis

Economic Growth and Inflation:

- The consistent rise in Per Capita GDP and CPI indicates economic growth paired with inflation.
- Higher GDP per capita generally reflects increased economic output and prosperity.
- The rising CPI shows that prices for goods and services have increased, affecting consumer purchasing power.



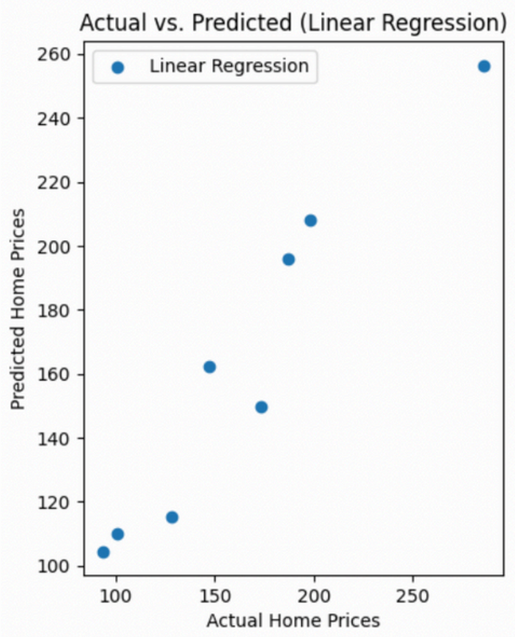
Declining Population Growth:

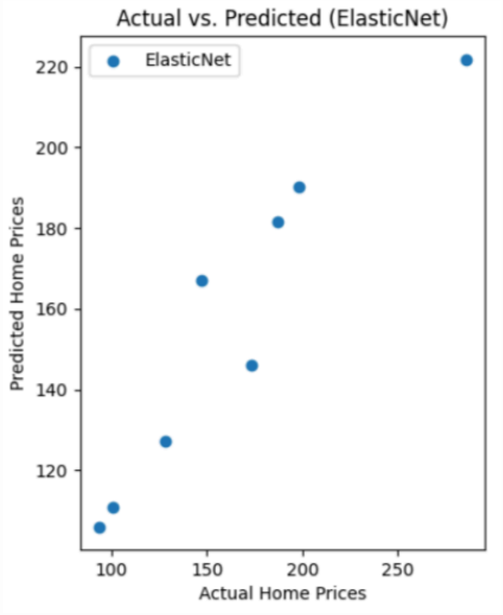
- The noticeable decline in the population growth rate over the years is significant.
- A declining growth rate can affect labor force availability, housing markets, and social services.
- This trend may require policies to address aging populations and workforce sustainability.

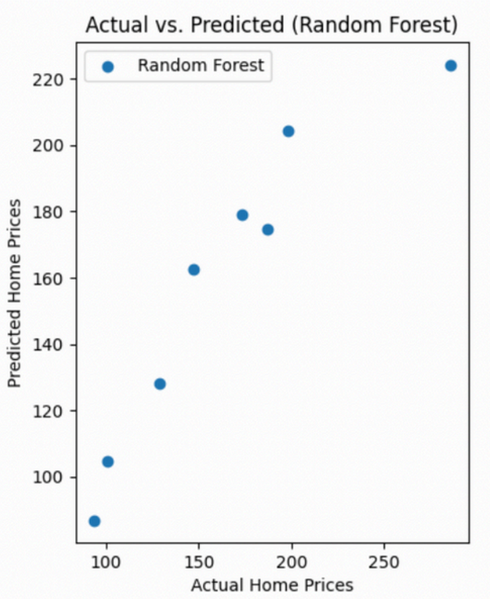
From 1987 to 2022, Per Capita GDP rose from 4,722.156 to 25,029.116, indicating economic growth, while the CPI increased from 111.4 to 282.599, reflecting inflation. The population growth rate declined from 0.8938 to 0.3776, suggesting slower growth due to demographic changes.

DATE	Per_Capita_ GDP	CPI	Population Growth Rate
1/1/1987	4722.156	111.4	0.893829
1/1/1988	5073.372	116	0.907999
1/1/1989	5511.253	121.2	0.944406
1/1/1990	5872.701	127.5	1.129651
1/1/1991	6035.178	134.7	1.336261
1/1/1992	6363.102	138.3	1.386886
1/1/1993	6729.459	142.8	1.31868
1/1/1994	7115.652	146.3	1.226296
1/1/1995	7522.289	150.5	1.190787
1/1/1996	7868.468	154.7	1.163412
1/1/1997	8362.655	159.4	1.20396
1/1/1998	8866.48	162	1.165715
1/1/1999	9411.682	164.7	1.14834
1/1/2000	10002.179	169.3	1.112769
1/1/2001	10470.231	175.6	0.989741
1/1/2002	10783.5	177.7	0.927797
1/1/2003	11174.129	182.6	0.859482
1/1/2004	11923.447	186.3	0.925484
1/1/2005	12767.286	191.6	0.921713
1/1/2006	13599.16	199.3	0.964254
1/1/2007	14215.651	203.437	0.951055
1/1/2008	14706.538	212.174	0.945865
1/1/2009	14430.902	211.933	0.876651
1/1/2010	14764.61	217.488	0.829617
1/1/2011	15351.448	221.187	0.726787
1/1/2012	16068.805	227.842	0.7336
1/1/2013	16648.189	231.679	0.69286
1/1/2014	17197.738	235.288	0.733362
1/1/2015	18063.529	234.747	0.736217
1/1/2016	18525.933	237.652	0.724676
1/1/2017	19280.084	243.618	0.632644
1/1/2018	20328.553	248.859	0.526435
1/1/2019	21104.133	252.718	0.455381
1/1/2020	21706.513	259.037	0.964348
1/1/2021	22600.185	262.65	0.156747
1/1/2022	25029.116	282.599	0.377565

Testing and Splits







Linear Regression:

- Performance:
- Strong performance with an R-squared of 0.92
- Lowest MSE of 279.80
- Explains approximately 92% of the variance in house prices
- Coefficients:
- Per Capita GDP: Highest positive influence on house prices
- CPI: Significant negative influence on house prices
- Population Growth Rate: Small positive impact on house prices

ElasticNet:

Performance:

- Moderate performance with an R-squared of 0.79
- MSE of 702.63

Coefficients:

- All coefficients are much smaller compared to Linear Regression
- Signs of coefficients differ
- CPI has a positive impact, contrary to Linear Regression

Random Forest:

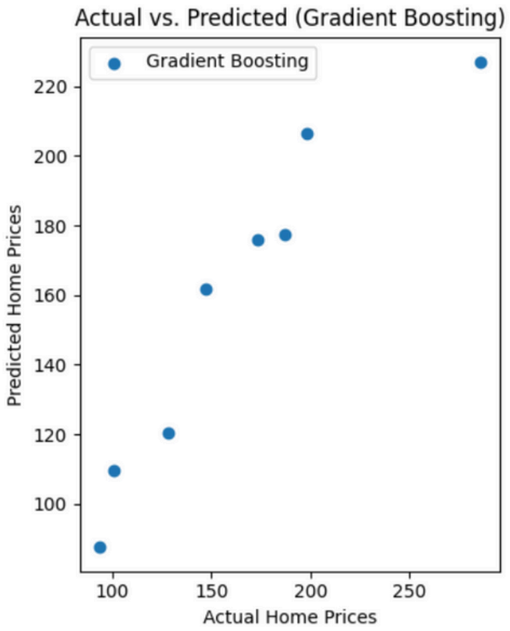
Performance:

- Good performance with an R-squared of 0.84
- MSE of 542.02

Feature Importance:

- Per Capita GDP: Most important feature
- CPI: Second most important feature
- Population Growth Rate: Least influence, consistent with other models

Testing and Splits



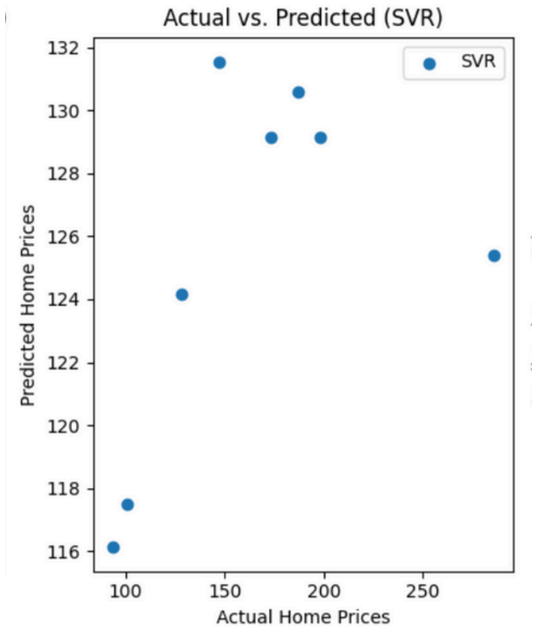
Gradient Boosting:

Performance:

- Similar to Random Forest with an R-squared of 0.85
- MSE of 501.47

Feature Importance:

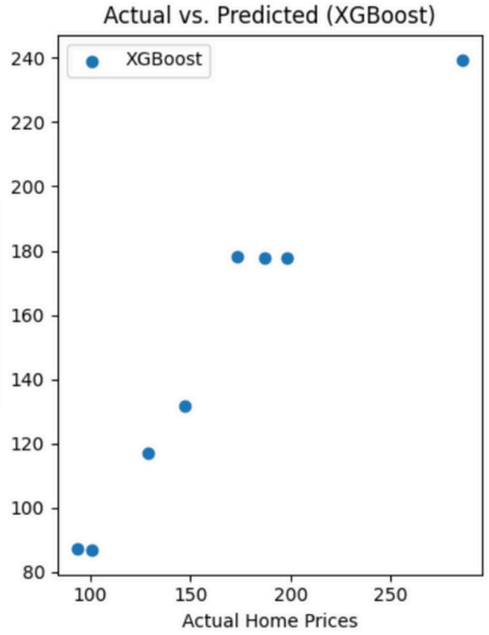
- Strongly dominated by Per Capita GDP, the most crucial factor
- CPI and Population Growth Rate have much lower importance



Support Vector Regression (SVR):

Performance:

- Performs poorly with a negative R-squared (-0.35)
- Very high MSE of 4585.34
- Indicates SVR is not suitable for this dataset, possibly due to sensitivity to scaling or inappropriate kernel choice



XGBoost:

Performance:

- Very strong performance with an R-squared of 0.88
- MSE of 406.35
- One of the top-performing models

Feature Importance:

- Heavily relies on Per Capita GDP (almost 99% importance)
- Negligible importance given to CPI and Population Growth Rate

Model Evaluation Metrics

Model	Mean Squared Error	R-squared
Linear Regression	279.79749180923886	0.9177594798023773
ElasticNet	702.6281537386947	0.7934774021907173
Random Forest	544.7008379086651	0.8398967768723563
Gradient Boosting	433.3156657160015	0.8726360785872975
SVR	4585.340471550712	-0.34776328131148637
XGBoost	406.3464077297732	0.8805631182179582

Model Coefficients and Features Importance

Model	Per_Capita_GDP	CPI	PopulationGrowthRate	Intercept
Linear Regression	93.12022449868712	-42.83788322077985	2.8072396206076373	127.40928571428573
ElasticNet	17.1293775852441	15.706827997375441	-8.660703675425227	127.40928571428572
Random Forest	0.5128924462680821	0.45235338962504085	0.03475416410687721	nan
Gradient Boosting	0.5870501794745152	0.38857999115361785	0.024369829371866998	nan
SVR	nan	nan	nan	nan
XGBoost	0.9899574518203735	0.00015177231398411095	0.009890790097415447	nan

Significant Insights:

- Linear Regression Performance: Linear Regression performs exceptionally well, with the lowest MSE and highest R-squared, making it the most reliable model for this dataset.
- XGBoost Effectiveness: XGBoost also shows strong performance with a low MSE and high R-squared, indicating it is a robust model, second only to Linear Regression.
- Per Capita GDP: Across all models, Per Capita GDP emerges as the most crucial feature, having the highest impact on predicting house prices.
- Model Variation: Different models assign varying levels of importance to CPI and Population Growth Rate, indicating that feature importance can be model-dependent.
- SVR Performance: SVR significantly underperforms compared to other models, indicating it is not suitable for this dataset.

Hyperparamenter and Grid Search

	Per_Capita_GDP	CPI	PopulationGrowthRate
Per_Capita_GDP	1.000000	0.991344	-0.828332
CPI	0.991344	1.000000	-0.790102
PopulationGrowthRate	-0.828332	-0.790102	1.000000

	Per_Capita_GDP	CPI	PopulationGrowthRate
Per_Capita_GDP	1.000000	0.999743	-0.758816
CPI	0.999743	1.000000	-0.759331
PopulationGrowthRate	-0.758816	-0.759331	1.000000

	Per_Capita_GDP	CPI	PopulationGrowthRate
Per_Capita_GDP	1.000000	0.996825	-0.641270
CPI	0.996825	1.000000	-0.644444
PopulationGrowthRate	-0.641270	-0.644444	1.000000

Significant Insights:

Linear Regression performs exceptionally well, with the lowest MSE and highest R-squared, making it the most reliable model for this dataset. XGBoost also shows strong performance, indicating it is a robust model, second only to Linear Regression. Across all models, Per Capita GDP emerges as the most crucial feature, having the highest impact on predicting house prices. Different models assign varying importance to CPI and Population Growth Rate, showing feature importance can be model-dependent. SVR significantly underperforms compared to other models, indicating it is not suitable for this dataset.

Key Insights:

All methods show a very strong positive correlation between Per Capita GDP and CPI (0.99+), indicating they move closely together. Population Growth Rate consistently shows negative correlations with both, with Pearson showing the strongest negative correlation (-0.83 with GDP and -0.79 with CPI). Spearman and Kendall correlations indicate slightly weaker negative relationships, suggesting higher population growth rates are associated with lower economic growth and inflation.

Hypertuning and Grid Search

R-squared Values:
- Indicate model performance.

Ridge Regression
- Achieved an R-squared of 0.88
- Shows strong performance.

Lasso Regression:
- Achieved an R-squared of 0.89
- Performed slightly better than Ridge Regression
- Handles feature selection effectively by shrinking some coefficients to zero.

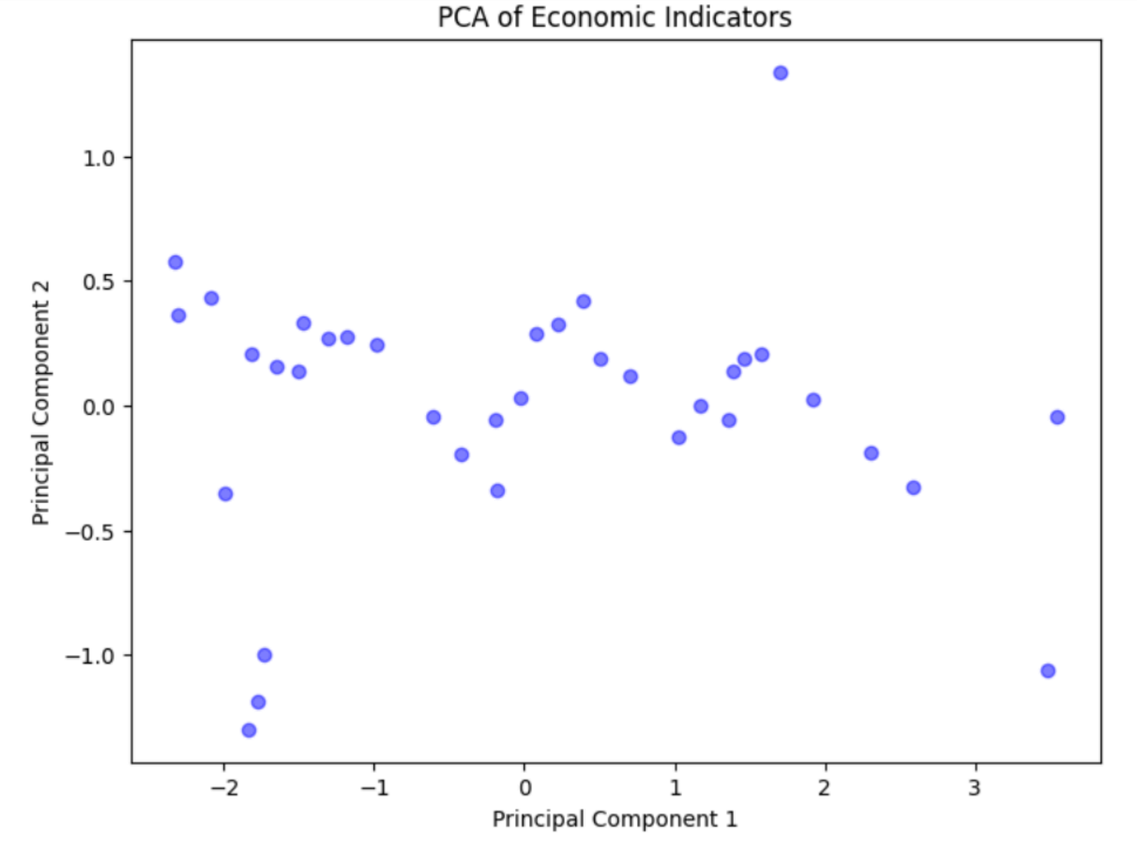
ElasticNet:
- Achieved a lower R-squared of 0.79
- Did not perform as well as Ridge and Lasso Regression in this instance.

```python
from sklearn.linear_model import Ridge, Lasso, ElasticNet

ridge = Ridge()
lasso = Lasso()
elastic_net = ElasticNet()

models = [ridge, lasso, elastic_net]
for model in models:
    model.fit(X_train_scaled, y_train)
    y_pred = model.predict(X_test_scaled)
    print(f"{model.__class__.__name__} R^2: {model.score(X_test_scaled, y_test)}")
```

```
Ridge R^2: 0.8757121578067654
Lasso R^2: 0.8864765302303425
ElasticNet R^2: 0.7934774021907173
```

Overall, Lasso Regression emerged as the best performer among the three, balancing regularization and feature selection to achieve the highest explanatory power for the dataset.

PCA of Economic Indicator



Key Insights and Analysis:

PCA Transformation:

- Reduces the dimensionality of economic indicators data to two principal components.
- Captures the most significant patterns in the data.

Explained Variance Ratio:

- Indicates how much of the total variance is captured by each principal component.

Scatter Plot:

- Shows the distribution of data points along the two principal components.
- Highlights patterns and potential clusters.

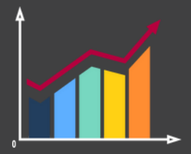
Data Distribution:

- Data points spread out along both principal components.
- Noticeable spread in Principal Component 1, indicating variability in the underlying economic indicators.

Summary:

- PCA effectively summarizes the dataset into two main directions of variance.
- Simplifies the complexity of the dataset while retaining most of its informative structure.

Conclusion



Comprehensive Analysis
The comprehensive analysis provided a deep understanding of how various economic indicators influence house prices.

Techniques Employed
By employing a range of statistical and machine learning techniques, the project effectively identified the most significant predictors and the best-performing models.

Insights Gained
The insights gained from the correlation analyses, visualizations, and model evaluations can inform economic policy and investment decisions in the housing market.

PCA Simplification
The PCA further simplified the data's complexity, making it easier to interpret the underlying patterns.

Overall Success
Overall, the project successfully achieved its objective of analyzing and predicting house prices based on economic indicators.

HULT
INTERNATIONAL
BUSINESS SCHOOL

THANK YOU