US HOME PRICE ANALYSIS

SUBMITTED BY: **SHAMBHABI DHAR**
dharshambhabi@gmail.com

OBJECTIVE

This project aims to identify key factors influencing U.S. home prices over the last 20 years and build a data science model to explain their impact. Using the S&P Case-Shiller Home Price Index as a proxy for home prices, publicly available data will be analysed to understand these trends and relationships.

METHODOLOGY

The methodology entails mainly 3 phases: 1. **Data Collection**, 2. **Data Preprocessing**, 3. **Data Analysis.**

1. DATA COLLECTION

Through various literature reviews and web surfing, I identified some key factors affecting home prices in the US. These factors include:

- Home Prices
- Inflation
- Unemployment rate
- Interest rates
- Economic activity
- Population

Data Source:

- The data was collected from the Federal Reserve Economic Data (FRED) database using the fredapi library.
- An API key was registered to access the data.

Key Factors and FRED Series IDs:

- Home Prices: S&P Case-Schiller Home Price Index (CSUSHPINSA)
- Inflation: Consumer Price Index for All Urban Consumers: All Items (CPIAUCSL), 10-Year Breakeven Inflation Rate (T10YIE)
- Unemployment Rate: Unemployment Rate (UNRATE)
- Interest Rates: 30-Year Fixed Rate Mortgage Average in the United States (MORTGAGE30US), Federal Funds Rate (FEDFUNDS)
- Economic Activity: Personal Consumption Expenditures (PCE), Real Gross Domestic Product (GDP), Median Household Income in the United States (MEHOINUSA646N)
- Population: Population Level (CNP16OV)

Fetching Data:

- Data for each key factor was fetched from FRED and stored in a dictionary.
- The dictionary was converted to a DataFrame.
- The data was restricted to the last 20 years (since January 2004) and saved as merged_housing_data.csv.

2. DATA PREPROCESSING

- The date column was parsed as a datetime index.
- Missing values were handled using the forward-fill method.
- The data was normalised using the StandardScaler from the sklearn.preprocessing module to ensure each feature had a mean of 0 and a standard deviation of 1.
- The target variable was defined as the S&P Case-Schiller Home Price Index.
- Other factors were defined as features.
- The dataset was checked for NaN and infinite values.
- Rows with such values were removed.

3. DATA ANALYSIS

- **Correlation Analysis:** Correlation coefficient was computed for each feature in relation to housing prices. A high positive or negative correlation signified a strong linear relationship between the feature and housing prices. A heatmap of the correlation matrix was plotted for visual representation.
- **Linear Regression**: The attributes were treated as independent variables and home prices as the dependent variable in a multiple linear regression model. Mean Squared Error (MSE) and R-squared (R^2) were computed to assess model performance. The model's coefficients were analysed to determine the significance of each article on the cost of housing. While negative coefficients implied a negative impact while positive coefficients indicated a positive impact.
- **One Way ANOVA:** One-Way ANOVA was performed to compare home prices across different economic indicators. The F-statistic and p-value were computed to determine the significance of each feature. A low p-value from the ANOVA test indicated that the feature significantly influences housing prices
- **T-Test & F-Test:** T-tests and F-tests showed significant differences in feature values before and after 2020 (i.e., pre and post covid time as there were significant economic changes during that period across the globe) , highlighting the impact of economic changes on home prices.
- **Random Forest:** To assess feature importance scores, a Random Forest model was used. This method provided insights into the relative influence of each characteristic on house prices by quantifying its contribution to the model's prediction performance.

RESULTS AND INTERPRETATIONS:

1. <u>CORRELATION ANALYSIS:</u>

- **Strong Positive:** Personal Consumption Expenditures (0.9050), Real GDP (0.9062), Median Household Income (0.8957), Consumer Price Index (0.8738), Population Level (0.7539) – all strongly associated with higher home prices.
- **Significant Negative:** Unemployment Rate (-0.5666) – associated with lower home prices.

2. <u>LINEAR REGRESSION:</u>

- **Strong Positive Impacts**: Real GDP (1.6368) and Median Household Income (0.6274) significantly increase home prices.
- **Negative Impacts**: Consumer Price Index (-0.4821) and Population Level (-1.0126) decrease home prices.
- Model Accuracy: High, with low Mean Squared Error (0.0265) and high R-squared (0.9731), indicating **high prediction accuracy and excellent fit**.

3. <u>ANOVA:</u>

All features show highly significant effects on the dependent variable. Notable results include:

- **Consumer Price Index:** F-statistic: 17,582.47, p-value: 0.0
- **Real GDP:** F-statistic: 17,262.56, p-value: 0.0
- **Personal Consumption Expenditures:** F-statistic: 14,947.67, p-value: 0.0

These indicate strong and statistically significant relationships.

4. <u>T-TEST:</u>

Notable T-test results indicate significant effects:

- **Consumer Price Index:** T-statistic: -95.90, p-value: 0.0
- **Personal Consumption Expenditures:** T-statistic: -98.62, p-value: 0.0
- **Real GDP:** T-statistic: -98.73, p-value: 0.0
- **Median Household Income:** T-statistic: -99.51, p-value: 0.0

These features show very strong and statistically significant effects.

5. <u>F-TEST:</u>

Notable F-Test results:

- **Real GDP:** F-statistic: 1.1118, p-value: 0.0118 (significant effect).
- **Median Household Income:** F-statistic: 5.4827, p-value: 1.11e-16 (highly significant effect).
- **Population Level:** F-statistic: 14.5918, p-value: 1.11e-16 (highly significant effect).

Other features show no significant impact.

6. RANDOM FOREST:

The Random Forest model shows:

- **Mean Squared Error (MSE):** 1.2599e-06 (extremely low, indicating high prediction accuracy).
- **R-squared (R^2):** 0.99999872 (nearly perfect fit, explaining 99.9999% of variance).

These results indicate a highly accurate and well-fitting model.

Notable feature importances:

- **Consumer Price Index:** 0.7859 (most influential feature).
- **Population Level:** 0.0976 (moderate importance).
- **Real GDP:** 0.0515 (notable but smaller impact).

Other features contribute minimally, with the 10-Year Breakeven Inflation Rate being the least important.

Therefore, it can be said after multiple analyses that:

- Inflation (CPI), Economic indicators (income, GDP, expenditure) and demographics (population) play a significant role in influencing housing prices positively.
- Unemployment Rate has a negative impact on housing pricing showing the relationship between housing market and labour market conditions (as demand decreases, price falls).
- Interest rates also have some effect on the housing market however it is not very significant.
- A combination of all these parameters can be used to determine the fluctuations in the US housing market and help in making informed decisions with regard to real estate economics.

LIMITATIONS

The following can be some of the limitations of the data science model:

1. **Data Quality and Completeness:** Despite forward-filling, missing data points can still affect model accuracy. The data might miss short-term fluctuations.
2. **Feature Selection:** The model may not include all potential influencing factors (e.g., local real estate conditions, policy changes) as they are only based on my understanding.

3. **Model Assumptions:** Linear Regression assumes a linear relationship, which may not capture complex non-linear patterns. While Random Forests can handle non-linearity, they can be less interpretable.
4. **Overfitting:** Random Forest may overfit the training data.

CONCLUSIONS

In this project, I have used multiple parameters to look at US home prices for the last 20 years. I collected data from open sources, handling missing values and other inconsistent data. I identified a few key factors using techniques like regression and correlation: inflation rates, interest rates, economic activities(GDP, income) and demographics, with unemployment rates having a negative effect. The results provide insight into housing patterns, however the model may have its own share of limitations, including missing data and arbitrary feature selections, though I have tried to overcome it to the best of my capabilities.