US Home Prices Modeling Project Report (2000–2025)

🌍 Project Objective

To model and analyze the key economic, demographic, and housing-related factors influencing US home prices over the past two decades using publicly available data and machine learning techniques.

📊 Data Collection & Preprocessing

- **Data Sources** *(Public & Official)*:
    - S&P Case-Shiller Home Price Index (FRED)
    - 30-Year Fixed Rate Mortgage Average (FRED)
    - Unemployment Rate (FRED)
    - Median Household Income (US Census Bureau)
    - Housing Starts (FRED)
    - Consumer Price Index (FRED)
    - Population Estimates (US Census Bureau)
- **Datasets Used**:
    - S&P Case-Shiller Home Price Index (Target)
    - Mortgage Rates
    - Unemployment Rate
    - Median Household Income
    - Housing Starts
    - Consumer Price Index (CPI)
    - Population Estimates (2000–2025)
- **Key Data Decisions**:
    - Chose S&P Case-Shiller HPI as the target due to its reliability and coverage.
    - Used monthly granularity across all datasets for uniformity.
    - Handled missing values using forward-fill/backward-fill techniques.
    - ROI analysis was not performed due to the absence of cost or revenue data in the dataset.
- **Preprocessing Steps**:
    - Converted all date columns to a unified datetime format.
    - Merged datasets chronologically on a monthly basis.

- Created a unified dataframe `master_df1` with 243 rows and 13 features.

- **Visualization Tools**:

  I prioritized more statistically rigorous visualizations using Python (Seaborn/Matplotlib) to ensure interpretability and precision. Tableau was explored, but Python visuals were found more suitable for this modeling-based analysis.

📊 Exploratory Data Analysis (EDA)

- **Key Findings**:
  - Strong positive correlation between CPI and Home Prices.
  - Unemployment and Mortgage Rates showed weaker correlation.
  - HPI demonstrated an upward trend, impacted by events like the 2008 recession and COVID-19.
  - Seasonal patterns in Housing Starts; CPI and Unemployment showed economic cycles.
- **Visualization Techniques**:
  - Correlation heatmap
  - Trend analysis with line plots
  - Distribution check with histograms and boxplots
  - Applied rolling averages for smoothing volatility

🎙 Feature Engineering

- **New Features Created**:
  - `HPI_rolling`: 12-month moving average of HPI
  - `Post_COVID`: Binary feature marking post-March 2020 period
  - `Log_Unemployment`: Log transformation to handle skewness in unemployment rate
  - `CPI_Growth`: Monthly % change in CPI
  - `Mortgage_Rate_Level`: Categorized mortgage rates into Low, Medium, High
- **Encoding**:
  - `Mortgage_Rate_Level` transformed using `OrdinalEncoder`

⚖ Modeling

- **Train-Test Split**:

- Time-based chronological split (80% train, 20% test) to preserve temporal patterns
- Avoided random split to maintain causality and avoid data leakage

- **Models Trained**:
  - Linear Regression: baseline model
  - Random Forest Regressor: ensemble model chosen for robustness
- **Performance Metrics**:

Model	R^2 Score	RMSE
Linear Regression	0.9992	0.0102
Random Forest	0.9998	0.0023

Residual Analysis

- **Linear Regression**:
  - Slight funnel shape in residuals → mild heteroscedasticity
  - Histogram slightly skewed but still bell-shaped
- **Random Forest**:
  - Residuals tightly centered around zero with no visible pattern
  - Histogram appears normally distributed

Conclusion: Random Forest outperformed Linear Regression in terms of error structure and fit.

Feature Importance

- **Top Predictors from Random Forest**:

Feature	Importance
Home_Price_Index	0.347
HPI_rolling	0.270
CPI	0.206
Population_Monthly	0.120

- These four features accounted for over **90%** of model predictive power.
- **Negligible Features**:
  - Mortgage_Rate, Housing_Starts, Month, CPI_Growth, Mortgage_Rate_Level
  - Low importance attributed to high collinearity or minimal monthly variation

---

🔁 Re-training with Top Features

- **Why**: To simplify the model without compromising accuracy
- **Action**: Re-trained Random Forest using top 4 features only
- **Outcome**:
  - R^2 Score: 0.9997
  - RMSE: 0.0042

💡 Decision Justification: Acceptable performance drop; model stays reliable while being easier to understand

---

💾 Model Exporting

- Final model exported as: `final_random_forest_model.pkl`
- Format: `joblib` or `pickle`
- Reusable in production pipelines or Flask-based apps

📄 Deliverables

- **Jupyter Notebook**: Includes EDA, modeling, and evaluation
- **Final Report**: This document
- **Model File**: `final_random_forest_model.pkl`

📊 Final Thoughts & Takeaways

- The project highlighted major economic patterns that influenced US housing prices.
- Historical pricing, CPI, and population growth were key influencers.
- Creating new features like rolling averages and transformations made the model perform noticeably better.
- Using a time-based split was essential to keep the evaluation realistic and prevent data leakage.
- Random Forest gave the most accurate results and helped explain which features mattered most.