US House Price Estimation Project Documentation

1. Introduction

The **US House Price Prediction** project is focused on predicting the trends and values of house prices across the United States based on key economic and housing indicators. The project aims to use historical data to forecast future house prices, which could help real estate investors, homebuyers, and policymakers understand market dynamics and make better-informed decisions.

1.1 Problem Statement

The core problem of this project is to predict the price trend of homes in the United States using various indicators that influence housing prices. These include economic factors such as the GDP, interest rates, unemployment rates, and housing supply factors such as housing starts and mortgage rates.

1.2 Objective

The goal is to build a predictive model that can forecast housing prices with reasonable accuracy, based on historical data and market indicators. The project uses machine learning techniques to understand and predict how different economic factors impact the housing market.

2. Data Collection & Sources

2.1 Data Source

The dataset used in this project is sourced from publicly available datasets, including:

- **S&P CoreLogic Case-Shiller U.S. National Home Price Index (CSUSHPISA)** – This dataset tracks the average price of homes in the U.S. over time, adjusted for inflation.
- **Federal Reserve Economic Data (FRED)** – This dataset contains macroeconomic indicators such as GDP, unemployment rates, and mortgage rates that affect housing prices.

These sources provide comprehensive data on U.S. housing prices, economic indicators, and housing market metrics over multiple years, which forms the basis for building a predictive model.

2.2 Data Collection Process

Data was collected over the course of several weeks to ensure that all variables relevant to housing prices were included:

1. **Housing Prices**: The CSUSHPISA index, which tracks U.S. house prices.
2. **Economic Indicators**: Includes GDP, interest rates, inflation, unemployment rates, and real estate loans.
3. **Housing Market Indicators**: Includes data on housing starts, building permits, and home sales.

The data was collected in various formats and unified into a clean, standardized dataset using **Python** libraries like **Pandas**.

2.3 Data Cleaning & Preprocessing

- **Removing Duplicates**: Any duplicate rows or entries were identified and removed.
- **Handling Missing Values**: Missing data points were filled using techniques such as forward filling, backward filling, or imputation based on the mean/median values.
- **Data Standardization**: Features with different scales (e.g., GDP in billions, mortgage rate in percentage) were standardized using the StandardScaler from **sklearn** to ensure that the model performed effectively.
- **Data Merging**: Data from multiple sources (e.g., FRED, S&P Case-Shiller) was merged based on common columns such as **Date** or **Region**.

3. Feature Engineering

3.1 Key Features

The main feature we aim to predict is **CSUSHPISA (S&P/Case-Shiller U.S. National Home Price Index)**, which represents the housing price index. However, several other features play a key role in influencing the prediction:

1. **GDP** (Gross Domestic Product): Indicates the overall economic health of the U.S., which affects housing prices.
2. **UNRATE** (Unemployment Rate): Higher unemployment typically reduces demand for housing.
3. **MORTGAGE30US** (30-Year Fixed Mortgage Rate): Affects affordability and demand for housing.
4. **HOUST** (Housing Starts): Represents the number of new housing units being built and can indicate the state of the housing market.
5. **PERMIT** (Building Permits): Indicates future housing supply and market conditions.

These features were selected based on their economic relevance and impact on housing prices.

3.2 Feature Transformation

- **Log Transformation**: Used for features like GDP and housing prices to normalize the data and reduce skewness.
- **Time-based Features**: For time-series prediction, date-based features (such as **Year** and **Month**) were extracted to capture seasonality or trends.
- **Lag Features**: Previous month's or quarter's data was used as features to predict the next period’s housing prices.

4. Exploratory Data Analysis (EDA)

4.1 Visualization Tools

To understand the relationships between various features and the target variable (house prices), the following Python libraries were used:

- **Matplotlib**: For creating basic charts like scatter plots and line plots.
- **Seaborn**: For creating advanced visualizations like heat maps and pair plots.
- **Pandas**: For handling and manipulating data.
- **NumPy**: For numerical operations such as calculating correlations and statistics.

4.2 Data Visualizations

Visualizations helped in gaining insights into how features interact and their correlation with the target feature (**CSUSHPISA**):

- **Scatter Plots**: Used to understand the relationship between individual features (e.g., **MORTGAGE30US** vs **CSUSHPISA**).
- **Heat Map**: Displayed the correlation matrix between all features, helping identify which features are strongly correlated with house prices.
- **Box Plot**: Identified outliers in features such as **MORTGAGE30US** and **GDP**.
- **Histogram**: Visualized the distribution of house prices, showing whether the prices are normally distributed or skewed.
- **Pair Plot**: Used to visualize relationships between multiple features and identify potential patterns.
- **Line Plot**: To show housing price trends over time.
- **Stacked Area Chart**: Used in Power BI dashboard to show the contributions of different features over time.

5. Machine Learning Model

5.1 Model Selection

The chosen model for predicting housing prices is the **Random Forest Regressor**. Random forests are ensemble learning methods that combine the predictions of multiple decision trees to improve accuracy and reduce overfitting. It is particularly useful for handling large datasets with complex relationships between features.

5.2 Model Implementation

Using **scikit-learn**:

- **train_test_split**: Split the data into training and testing datasets.
- **StandardScaler**: Applied to scale the features for better performance of the model.
- **RandomForestRegressor**: The main algorithm used to fit the model and predict house prices.
- **Evaluation Metrics**:
    - **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual house prices.

- **R² Score**: Measures how well the model explains the variance in the target variable (house prices).

5.3 Model Training and Testing

- The dataset was split into 80% training data and 20% testing data.
- The model was trained on the training dataset, and its performance was evaluated on the testing dataset.

5.4 Model Performance

- **R² Score**: A high R² score (close to 1) indicates that the model is able to explain a large portion of the variance in the house prices.
- **Mean Squared Error (MSE)**: A lower MSE indicates that the model's predictions are closer to the actual house prices.

6. Dashboard (Power BI)

6.1 Dashboard Visualizations

The interactive **Power BI** dashboard was built to visualize the predictions and underlying factors driving the house price trends:

- **Pie Chart**: Displayed the proportion of housing starts, GDP, mortgage rates, etc., contributing to house price changes.
- **Bar Chart**: Showed year-on-year changes in the average housing price across different regions.
- **Gauge Chart**: Indicated the current house price relative to the maximum observed value, providing a sense of how "hot" or "cold" the market is.
- **Stacked Area Chart**: Displayed how different economic indicators contribute to the housing price trend over time.
- **Slicers**: Allowed filtering by **Region** and **Year** to explore different subsets of the data.
- **Table**: Displayed detailed predictions of house prices along with important features for each period.

6.2 Slicers and Filters

To allow users to customize their views, slicers were added:

- **Region**: Enables users to filter housing prices based on geographic regions (e.g., **San Francisco**, **New York**, etc.).
- **Year**: Allows users to see data for specific years or time periods.

6.3 Drill-through Features

The Power BI dashboard supports drill-through features, allowing users to click on a specific region or year to view detailed insights and trends about housing prices, GDP, unemployment rates, etc.

7. Conclusion

The project successfully predicts U.S. house prices using machine learning techniques. By leveraging Random Forest Regressor, the model is able to take into account multiple factors influencing housing prices, such as GDP, mortgage rates, and housing supply metrics. The **Power BI** dashboard offers an intuitive interface for exploring and visualizing the predictions and the factors contributing to house price changes.

7.1 Key Insights

- The **mortgage rate** and **housing starts** are among the most influential factors in determining housing prices.
- Historical trends and economic indicators such as GDP growth and unemployment rates can provide early warnings about changes in the housing market.

8. Future Work

8.1 Model Improvements

- **Incorporating More Features**: Including features such as population density, crime rates, and infrastructure projects can provide a more comprehensive view of the housing market