**Home LLC Assignment**

**Problem Statement:**

Find publicly available data for key factors that influence US home prices nationally. Then, build a data science model that explains how these factors impacted home prices over the last 20 years.

Use the S&P Case-Schiller Home Price Index as a proxy for home prices: fred.stlouisfed.org/series/CSUSHPISA.

**To build a data science model explaining how various factors have impacted U.S. home prices over the last 20 years, we'll follow these steps:**

1. **Data Collection**: Finding key factors which influence US home prices nationally and extracting data for 20 years
2. **Data Cleaning & Preprocessing**: merging datasets, missing value, duplicate values etc.
3. **Exploratory Data Analysis**: Analyze the data to find patterns and insights
4. **Model Selection & Training**: Choose and train a suitable model
5. **Model Evaluation**: Assess the performance of the model
6. **Prediction & Insight generation**: Interpret the results to understand the impact of different factors on home prices.

**STEP 1: Data Collection**

**Sources used to gather data**

* CASE-SCHILLER Home Price Index - <https://fred.stlouisfed.org/series/CSUSHPISA>
* Unemployment rate - <https://fred.stlouisfed.org/series/UNRATE>
* Interest rates - <https://fred.stlouisfed.org/series/FEDFUNDS>
* New Constructed units - <https://fred.stlouisfed.org/series/COMPUTSA>
* Income - <https://fred.stlouisfed.org/series/DSPIC96>
* Construction price index - <https://fred.stlouisfed.org/series/WPUSI012011>
* Per Capita GDP - <https://fred.stlouisfed.org/series/A939RX0Q048SBEA>
* percent urban population - <https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?end=2021&locations=US&start=2001>
* Total households - <https://fred.stlouisfed.org/series/TTLHH>
* Consumer Price Index : <https://fred.stlouisfed.org/series/MEDCPIM158SFRBCLE>
* Working Age Population: <https://fred.stlouisfed.org/series/LFWA64TTUSM647S>

The data above gives us an overview of various factors that could impact U.S. home prices. Here's a brief summary of each dataset:

* ConstructionPriceIndex: Contains data about the construction price index over time
* HousingSubsidiesFederal: Details about federal housing subsidies
* Income: Information on income levels
* InterestRate: Data on interest rates
* NewConstructedUnits: Numbers of newly constructed housing units
* PerCapitaGDP: Data on per capita GDP
* TotalHouseholds: Information about the total number of households
* UNRATE: The unemployment rate data
* UrbanPopulation: Data on urban population
* HousePriceIndex: Shows the house price index over the years

**STEP 2: Data Cleaning & Preprocessing**

After reading all the dataset (csv) files in jupyter notebook, I worked on the following:

* Correcting date format
* Checking missing values
* Checking duplicate values
* Handling any other anomalies

The rest of the steps were performed in the jupyter notebook. It is attached as a pdf document along with this word document.