Exploring the Impact of Supply-Demand Factors on US Home Prices: A 20-Year Analysis

Finding the influence of Supply and Demand Factors on US Home Prices: A 20-Year Data Science Study Utilizing Machine Learning

Key Supply-Demand factors that influence US home prices.

Home Prices (Response Variable):

Use the S&P Case-Schiller Home Price Index as the proxy for home prices.
Source: S&P Dow Jones Indices LLC
https://fred.stlouisfed.org/series/CSUSHPISA#0

Economic Indicators (GDP):

Gross domestic product (GDP) is the total monetary or market value of all the finished goods and services produced within a country's borders
Source: U.S. Bureau of Economic Analysis (BEA)
https://www.bea.gov/data/gdp/gross-domestic-product

Economic Indicators (Unemployment Rate):

The percentage of people in the labour force that are unemployed.
Source: Federal Reserve Bank of St. Louis
https://fred.stlouisfed.org/series/UNRATE

Mortgage Rates:

A mortgage is an agreement between you and a lender that gives the lender the right to take your property if you fail to repay the money you've borrowed plus interest.
Source: FRED website
https://fred.stlouisfed.org/series/REAINTRATREARAT10Y

Housing Supply and Demand:

The relationship between the availability of housing and the demand for housing in a particular area or market.
Source: USHMC-Housing-Demand
https://www.huduser.gov/portal/ushmc/hd_home_sales.html

Population Growth:

Population growth is the increase in the number of people in a population or dispersed group.
Source: Federal Reserve Bank of St. Louis
https://fred.stlouisfed.org/series/POPTHM

Inflation Rates:

The rate at which prices increase over time, resulting in a fall in the purchasing value of money.
Source: Federal Reserve Bank of St. Louis
https://fred.stlouisfed.org/series/T10YIE#0

Employment-Population Ratio:

The employment-to-population ratio measures the number of workers currently employed against the total working-age population of a region.

Source: Federal Reserve Bank of St. Louis

https://fred.stlouisfed.org/series/EMRATIO

New House Sold:

Number of new houses sold every month in US.

Source: USHMC-Housing-Demand

https://www.huduser.gov/portal/ushmc/hd_home_sales.html

Average Sales Price for US:

Average Price of House sold in US over a period of time.

Source: USHMC-Housing-Demand

https://www.huduser.gov/portal/ushmc/hd_home_sales.html

Date:

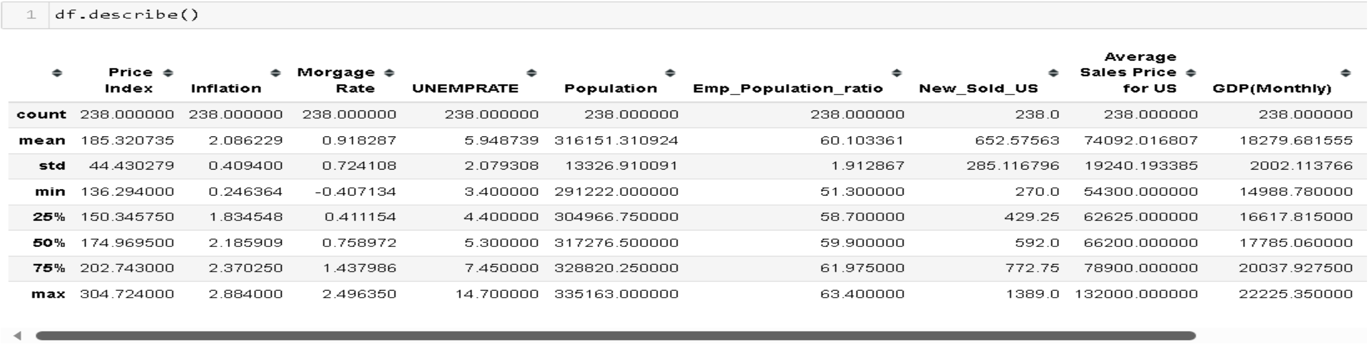
The Date of the observation. (2003-2023)

Note: All the data are Collected from above mention source for duration 01-09-2003 and 01-09-2023.

Exploratory Data Analysis (Patterns or correlations):

Exploratory data analysis was employed to uncover insights into the correlation between various supply and demand factors and the U.S. National Home Price Index. The collected datasets were scrutinized, revealing key findings and visualizations that effectively portray the trends and relationships of each factor with the overall trajectory of home prices in the United States.

We have 239 rows and 10 columns in our dataset. Most of the data are Float, there is no missing value in the dataset, we converted the data type of date-to-date type and then further made 2 new features month and year, we dropped original date column.



```
df.describe()
```

	Price Index	Inflation	Morgage Rate	UNEMPRATE	Population	Emp_Population_ratio	New_Sold_US	Average Sales Price for US	GDP(Monthly)
count	238.000000	238.000000	238.000000	238.000000	238.000000	238.000000	238.0	238.000000	238.000000
mean	185.320735	2.086229	0.918287	5.948739	316151.310924	60.103361	652.57563	74092.016807	18279.681555
std	44.430279	0.409400	0.724108	2.079308	13326.910091	1.912867	285.116796	19240.193385	2002.113766
min	136.294000	0.246364	-0.407134	3.400000	291222.000000	51.300000	270.0	54300.000000	14988.780000
25%	150.345750	1.834548	0.411154	4.400000	304966.750000	58.700000	429.25	62625.000000	16617.815000
50%	174.969500	2.185909	0.758972	5.300000	317276.500000	59.900000	592.0	66200.000000	17785.060000
75%	202.743000	2.370250	1.437986	7.450000	328820.250000	61.975000	772.75	78900.000000	20037.927500
max	304.724000	2.884000	2.496350	14.700000	335163.000000	63.400000	1389.0	132000.000000	22225.350000

From this code we get many information such as:

1.All column have 238 values so no missing value present

2.The std is high for mean in (Chance of Outlier)

- Price Index (Label)
- New_Sold_US

3.All the minimum values are possible.

4.The difference between min,25%,50%,75% and max is not normal for: -

- Mortgage rate
- UnEmployement
- New house sold us
- Average Sale price

5.The mean value is higher than median (50%), which means data is right skewed
- Price Index (Label)
- Unemployment
- Average Sale price
And left skewed in
- Inflation

Inflation: Inflation have 0.1125 correlation with respect to Pricing index. There is a weak positive relationship between the Inflation and CSUSHPISA. This suggests that as the rise in Inflation may have a slight positive impact on home prices.

Mortgage rate: Mortgage rate have -0.0642 correlation with respect to Pricing index. There is a weak negative relationship between the Mortgage rate and CSUSHPISA. This suggests that as the rise in Mortgage rate may have a slight negative impact on home prices.

Unemployment: Unemployment have -0.5408 correlation with respect to Pricing index. There is a moderate negative relationship between the Unemployment and CSUSHPISA. This suggests that as the rise in Unemployment have a negative impact on home prices.

Population: Population have 0.67614 correlation with respect to Pricing index. There is a high positive relationship between the Population and CSUSHPISA. This suggests that as the rise in Population have a high positive impact on home prices.

Employment by Population ratio: EPR have -0.063 correlation with respect to Pricing index. There is a weak negative relationship between the EPR and CSUSHPISA. This suggests that as the rise in EPR may have a slight negative impact on home prices.

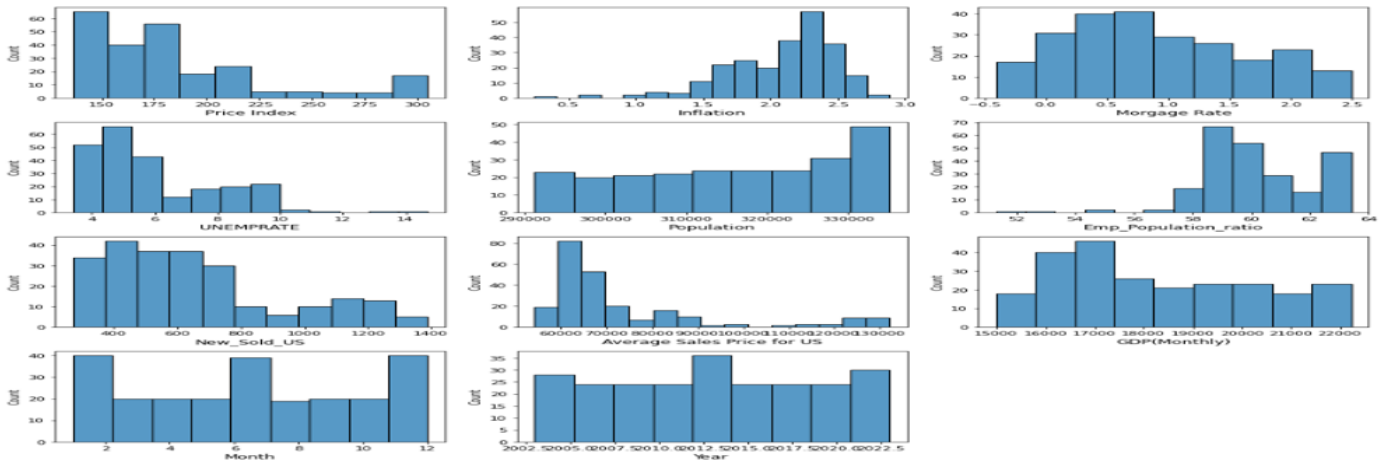
Number of New House sold: Number of New House sold have 0.1921 correlation with respect to Pricing index. There is a positive relationship between the Number of New House sold and CSUSHPISA. This suggests that as the rise in Number of New House sold may have a positive impact on home prices.

Average Sale Price of House US: Average Sale Price of House US have 0.9639 correlation with respect to Pricing index. There is a very high positive relationship between the Average Sale Price of House US and CSUSHPISA. This suggests that as the rise in Average Sale Price of House US have a very high positive impact on home prices.

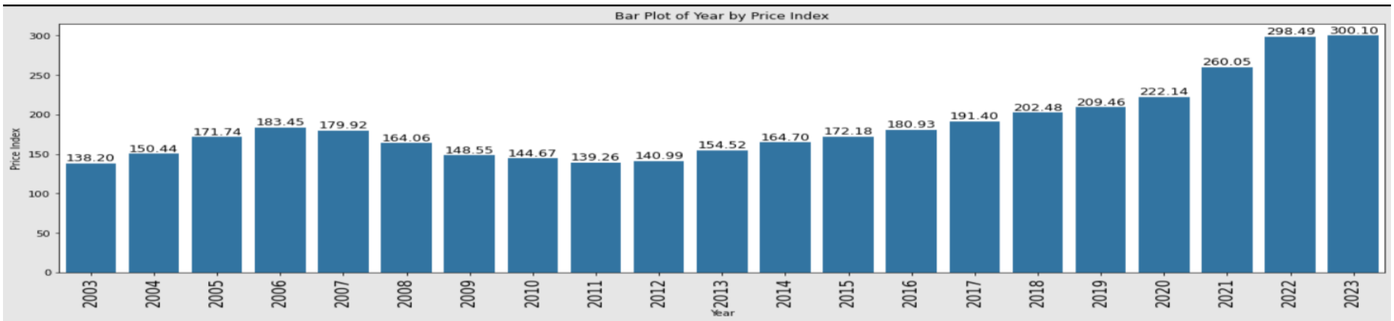
GDP: GDP have 0.8469 correlation with respect to Pricing index. There is a high positive relationship between the GDP and CSUSHPISA. This suggests that as the rise in GDP have a high positive impact on home prices.

Year: Year have 0.7490 correlation with respect to Pricing index. There is a high positive relationship between the Year and CSUSHPISA. This suggests that as the rise in Year have a high positive impact on home prices.

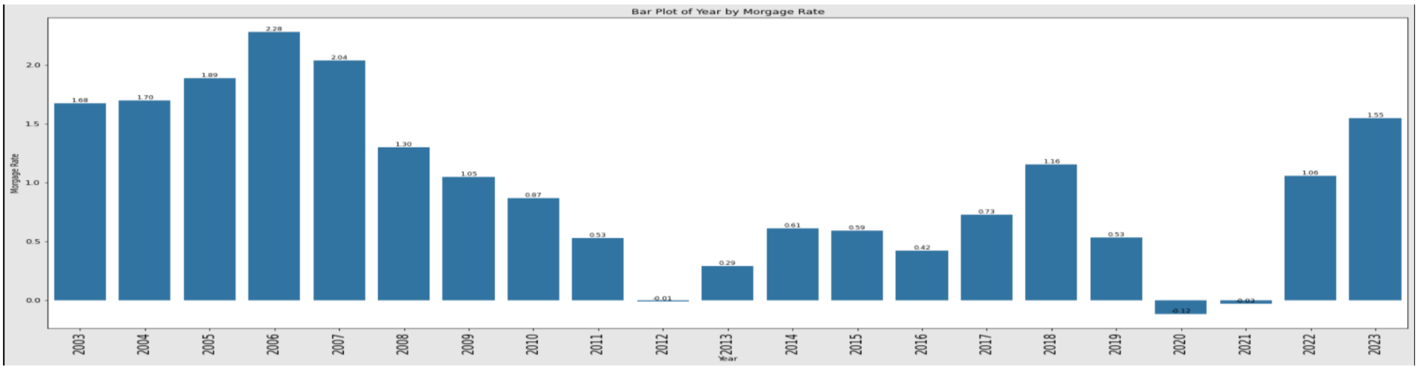
Graphical Analysis



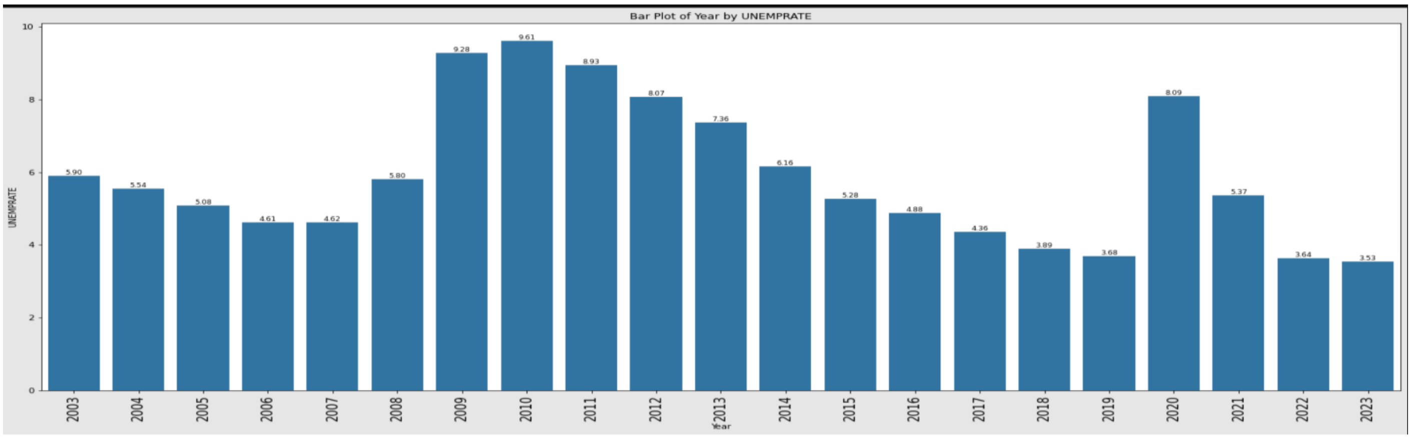
From the above graph we can see the distribution of each feature as in initial Analysis we found out that Price index, inflation, Unemployment rate, Emp_Population_ratio, Average Sale price data is skewed and it is found in the graphs as well.



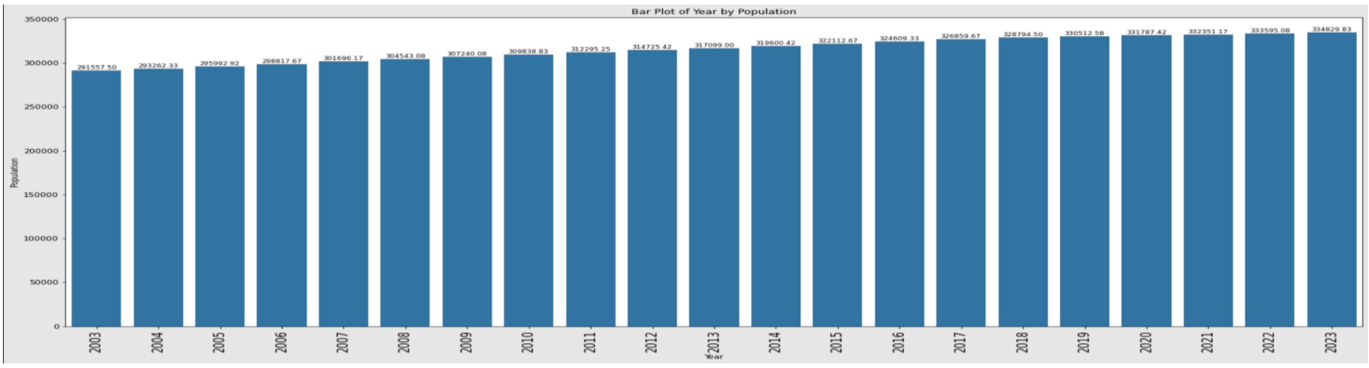
In the above graph we can see the Price Index Changes with Year as we can see that in it was increasing in 2003 but it starts falling from 2008. We very well know about recession Phase, then after 2012 it started Increasing. The price index is taken as 100 for year 2000.



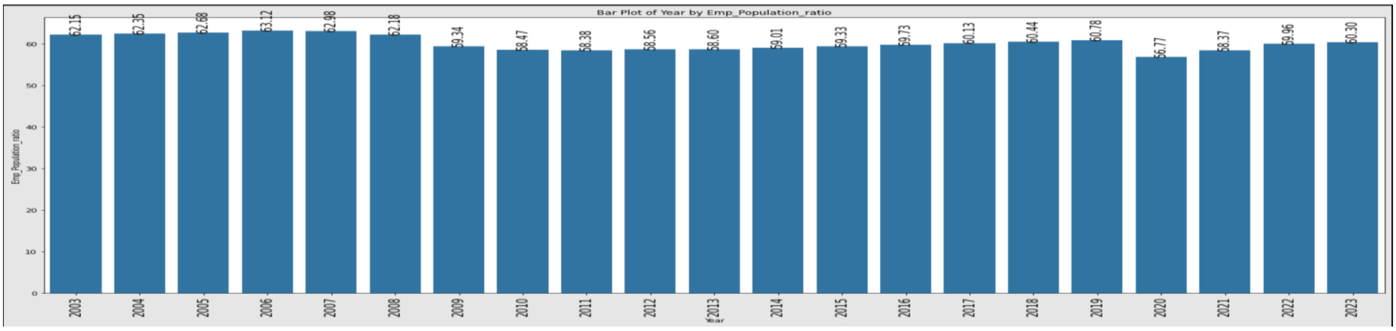
From the above graph we can see that the Mortgage rate was highest in 2006 and goes down in 2012 due to recession and then it started rising but at the time of Covid it even goes negative in 2020 which we can observe in the graphical analysis of Mortgage rate over the time period.



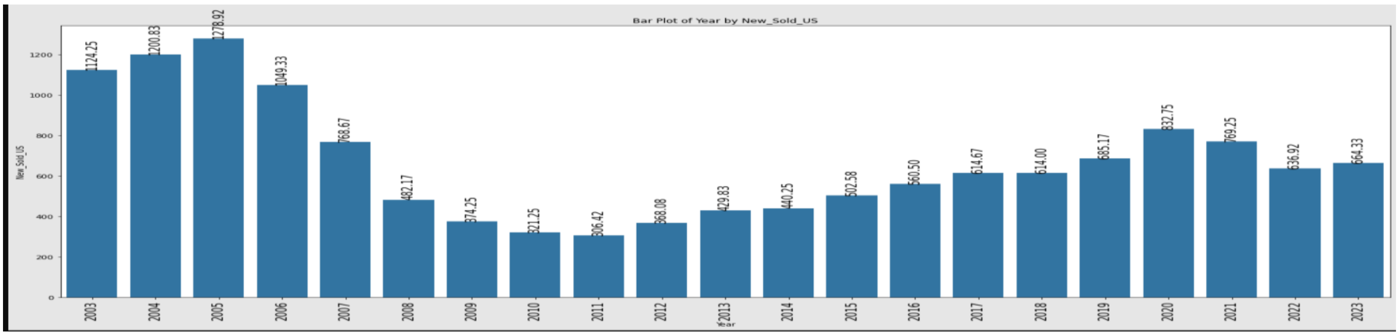
From the above graph we can see that the Unemployment rate was highest in 2010 due to recession which is 9.01 and goes down in 2019 and then it started rising at the time of Covid it reaches high in 2020 which is 8.09 we can observe in the graphical analysis of Unemployment rate over the time period, as we can see the Unemployment is almost opposite to Mortgage rate.



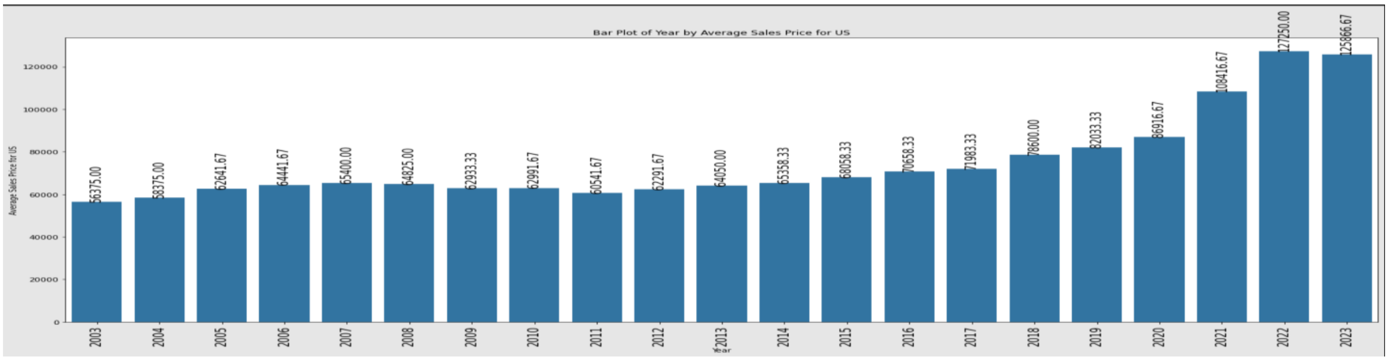
From the above graph we can see that Population Increases over time with no abnormalities as the time increase the population increase gradually.



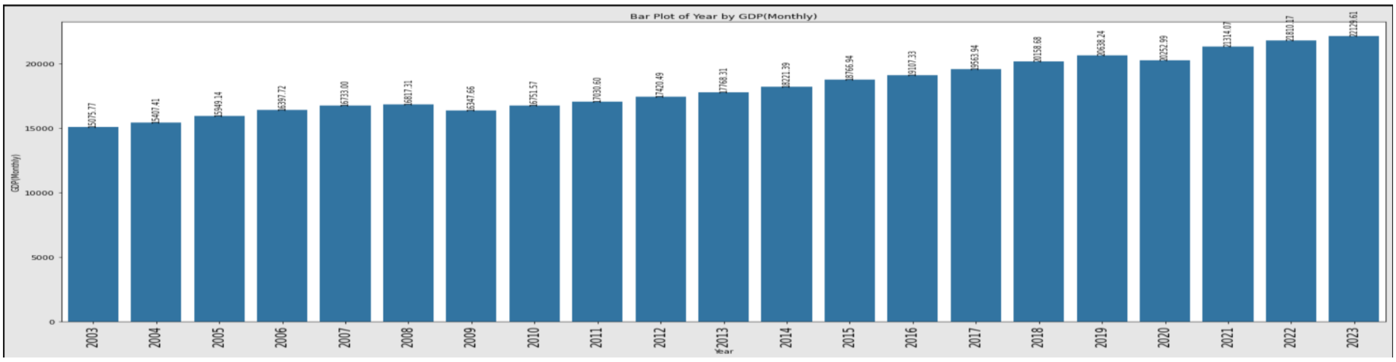
From the above graph we can see that ratio of employment population as it decreased in years 2010- 2020 as this year were bad in term of economy for the US



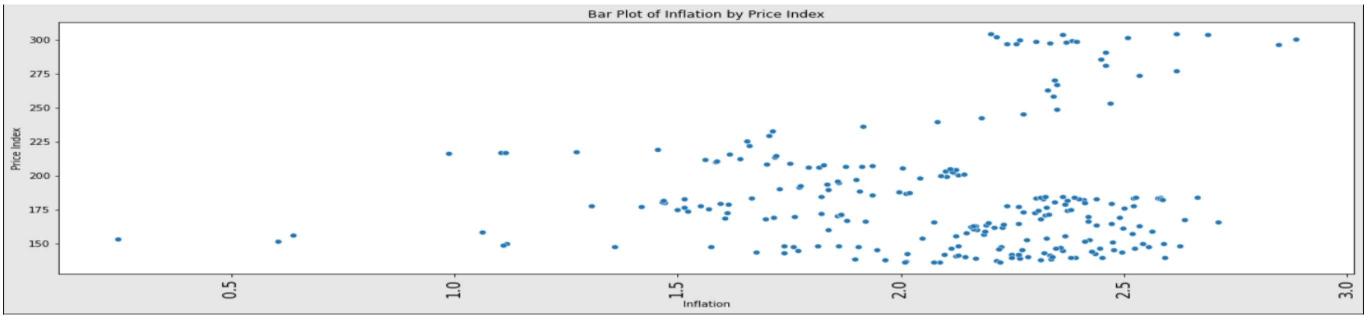
From the above graph we can see that new house sold in US increase from 2003 to 2005 but later dropped to lowest 2010 as US hit the recession and then started increasing, in year 2020 we can se the demand is rising as people use to invest this time in real state for better return later year, we can see drop Due to some factor.



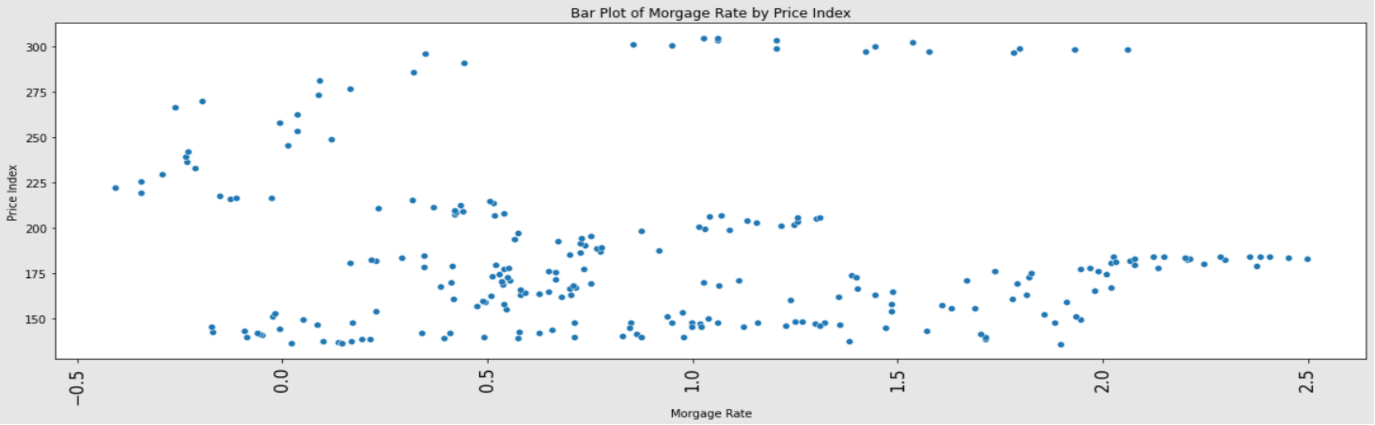
From the above graph we can see that Average price of House kept on increasing and in recent year 2021-2022 the growth was very high and then dropped a bit in year as 2023 as in 2020 people invested in real estate and price rises rapidly and in year 2023 it dropped due to high supply low demand.



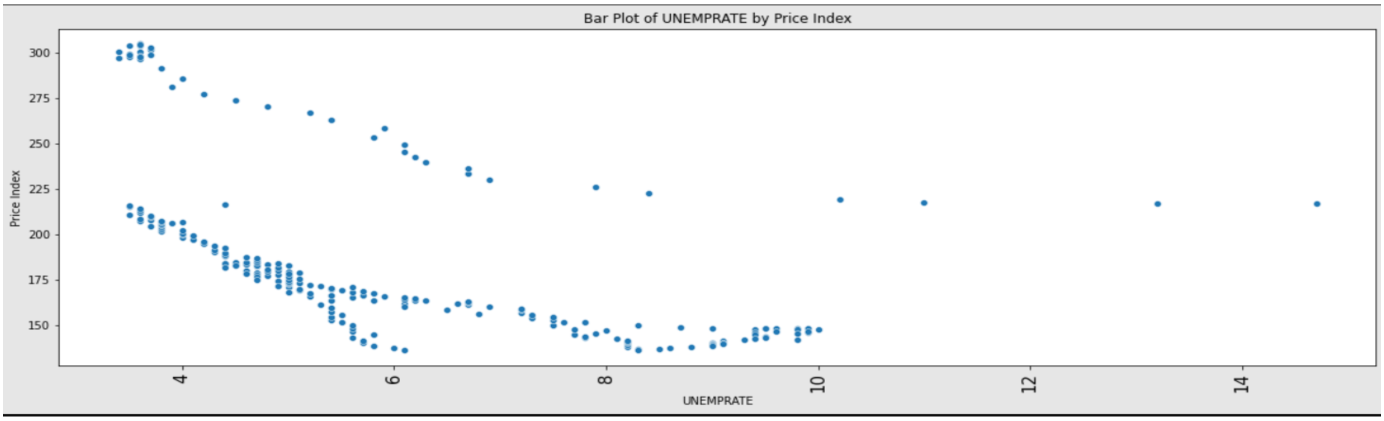
From the above graph we can see that The GDP of US dropped in 2009 as US was Hit by recession and then again in year 2020. From the above analysis we can say that GDP of US grows with time but if some reason such as recession and pandemic occur it can drop.



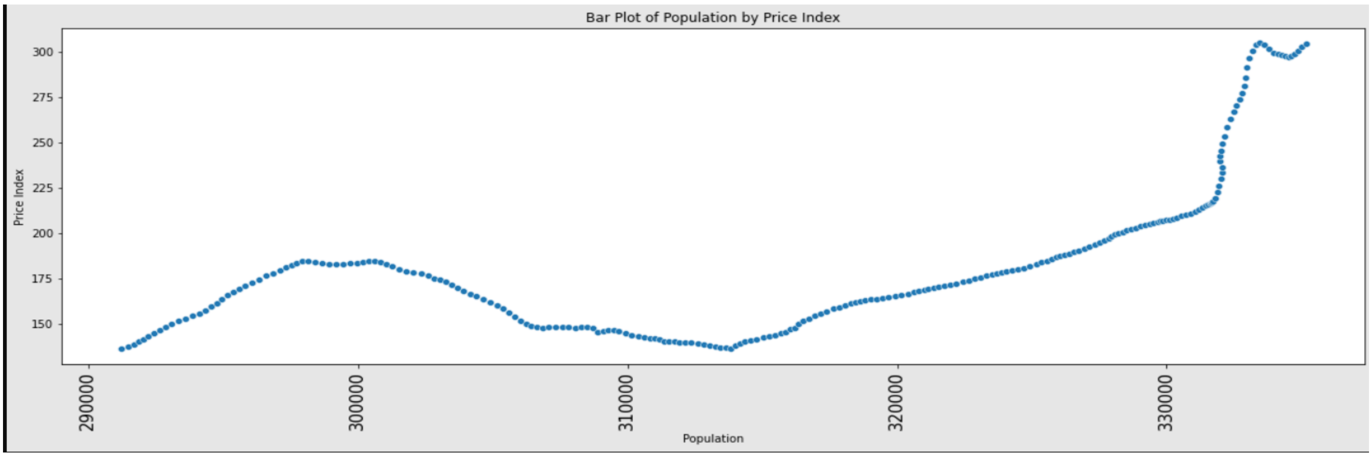
From the above graph we can see that for low inflation Price Index is also low but for the higher price index we can't say anything as the graph is very random and scattered as the inflation rate increases, but for higher price Index we can say that inflation is high and it is one of the key factors.



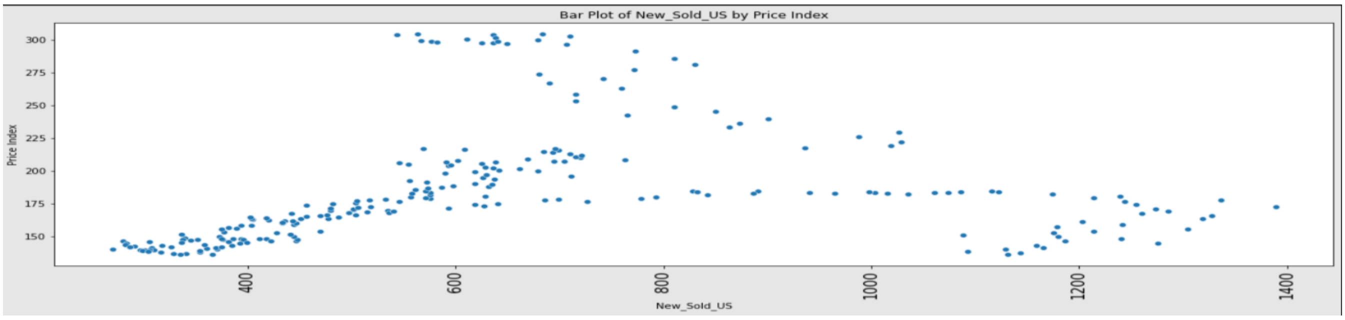
From the above graph between price index and Mortgage rate we can see that there is no fix trend which state that there is very less correlation between these two features.



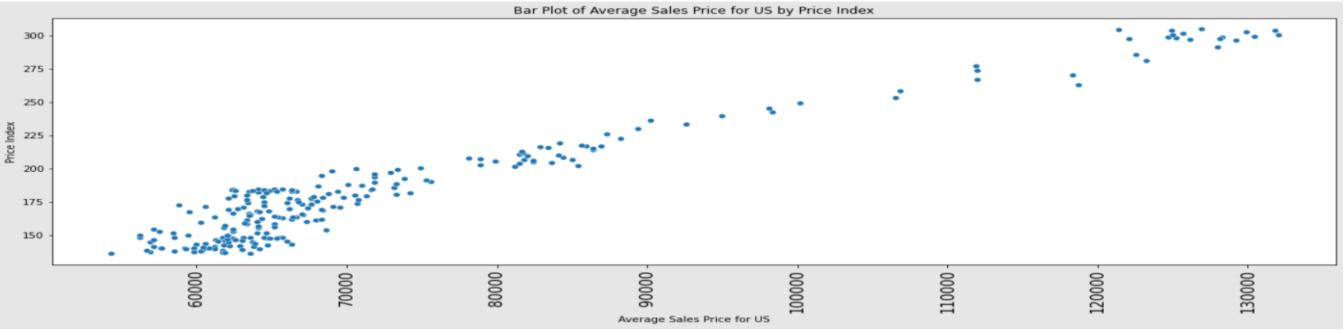
From the above graph we can see that as unemployment rate increases the Price index decreases, we can see that for low unemployment rate the Price Index is quite high.



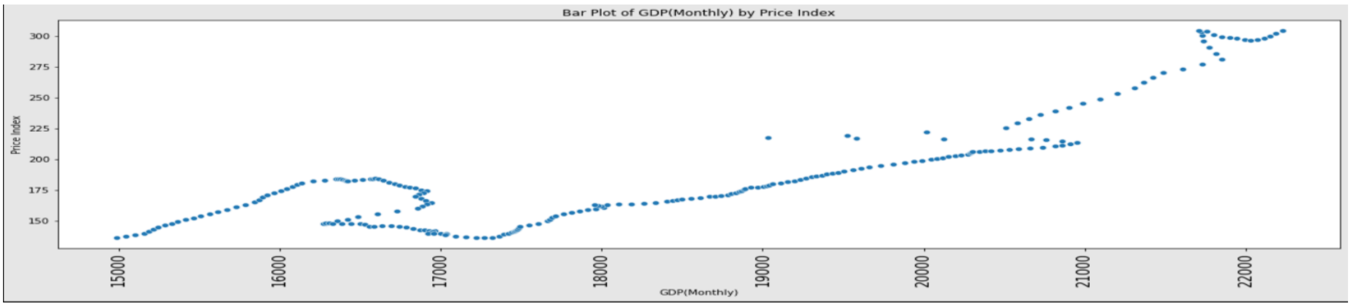
From the above graph we can see that as the population increases the Price index increase not perfectly but to some extent, there might be some other factor which might impact as well.



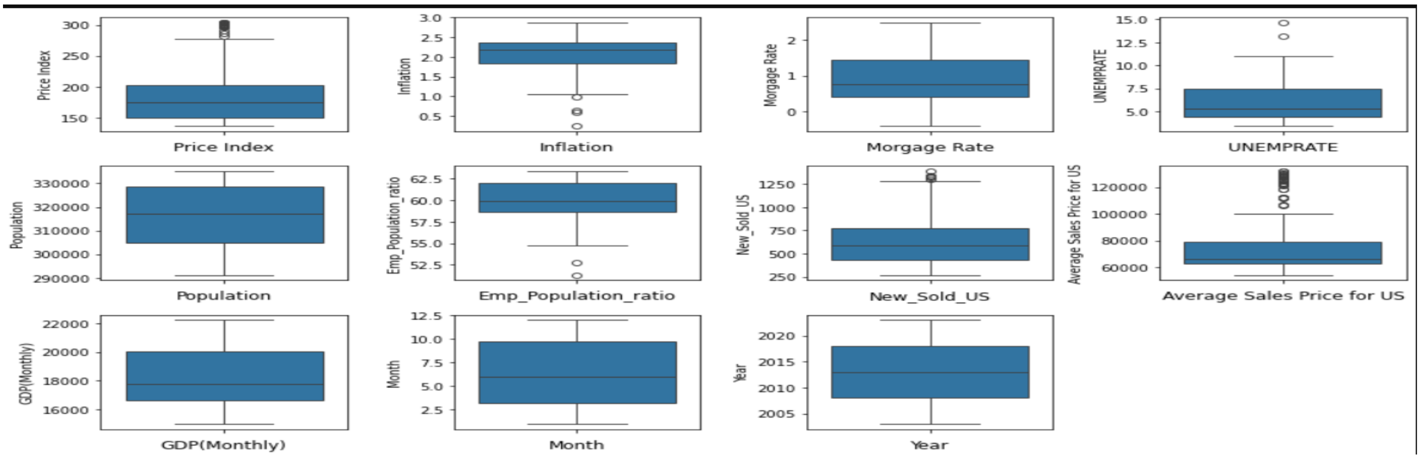
From the above graph we can see that there is no proper trend found but initially Price index was increasing as Number of house sold increased after a point randomness increases.



From the above graph we can see that as the Average price of house increases the Price Index also increases, the trend is linear and the correlation is strong between them.



From the above graph we can see that with increase in GDP the Price index increases to some extent. But there are some exceptions as well in the graph we can notice some irregularity.



From the above graph we can see that there are some outliers in few features but as our dataset is small, we will leave it.

Model Building

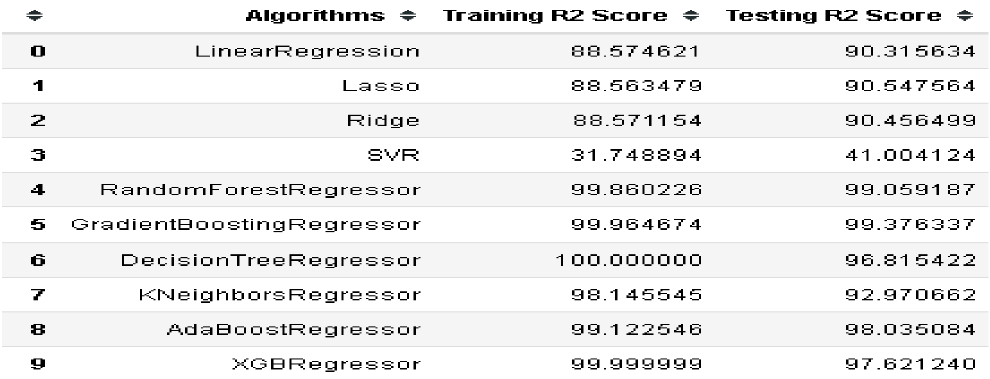
To build the data science model, we are going to use Many models and later we will find out the best model for predicting continuous numerical values based on the relationship between independent variables (features) and the dependent variable (target).

The first step in building the model was to prepare the data. We selected the relevant features Inflation, Mortgage rate, Unemployment rate, Employment Population ratio, Number of new houses sold, Average sale price of house for Us, GDP and The target variable we aimed to predict was ‘CSUSHPISA’ (S&P Case-Shiller U.S. National Home Price Index).

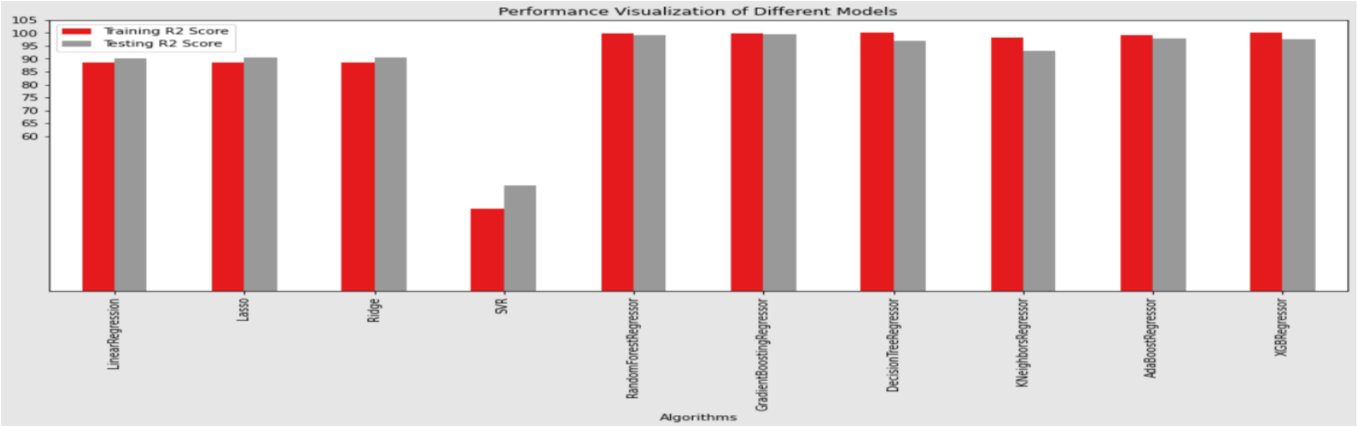
Once the data was prepared, we split it into training and testing sets using a 75:25 ratio, where 75% of the data was used for training the model, and 25% was reserved for evaluating its performance.

Next, we defined a dictionary of candidate models, including Linear Regression, Ridge, Lasso, SVR, Decision Tree, Random Forest, Gradient Boosting, KNeighbors Regressor, AdaBoost Regressor, XGB Regressor. These models represent different algorithms with varying complexities and learning capabilities.

We will start by selecting the best random state followed by train test split, then we made a function to evaluate the best performance of the model, we build 10 model to solve our problem.



⬍	Algorithms ⬍	Training R2 Score ⬍	Testing R2 Score ⬍
0	LinearRegression	88.574621	90.315634
1	Lasso	88.563479	90.547564
2	Ridge	88.571154	90.456499
3	SVR	31.748894	41.004124
4	RandomForestRegressor	99.860226	99.059187
5	GradientBoostingRegressor	99.964674	99.376337
6	DecisionTreeRegressor	100.000000	96.815422
7	KNeighborsRegressor	98.145545	92.970662
8	AdaBoostRegressor	99.122546	98.035084
9	XGBRegressor	99.999999	97.621240



To select the best performing model, we used Hyper parameter tuning with cross-validation with four folds. This technique helps assess the models' performance on different subsets of the training data. We used the mean R square (r2) as the evaluation metric, where higher values indicate better performance.

Based on the Hyper parameter tuning results, we identified Gradient Boosting Regressor as the best model with the highest R2 score. Finally, we evaluated the model's performance on the testing set by making predictions and calculating the R2 score.

In summary, our approach involved selecting relevant features, splitting the data, trying multiple regression models, performing Hyper parameter tuning with cross-validation for model selection, and evaluating the chosen model on the testing set. The Gradient Boosting Regressor model showed the best performance, and we used its coefficients to understand the impact of each feature on the predicted target variable.

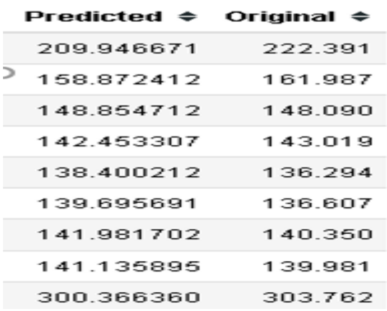
Model Evaluation:

To evaluate the performance of our model, we used two key metrics R-squared score. This metrics provide insights into the accuracy and reliability of the model's predictions.

We used the R-squared score, which measures the proportion of variance in the target variable that can be explained by the model. It ranges from 0 to 1, with higher values indicating a better fit. The R-squared score helps us understand how well the independent variables (features) explain the variation in the dependent variable (target).

Based on our evaluation, Gradient Boosting Regressor model performed well. **R-squared score was 0.993, suggesting that approximately 99.30%** of the variation in the target variable can be explained by the model.

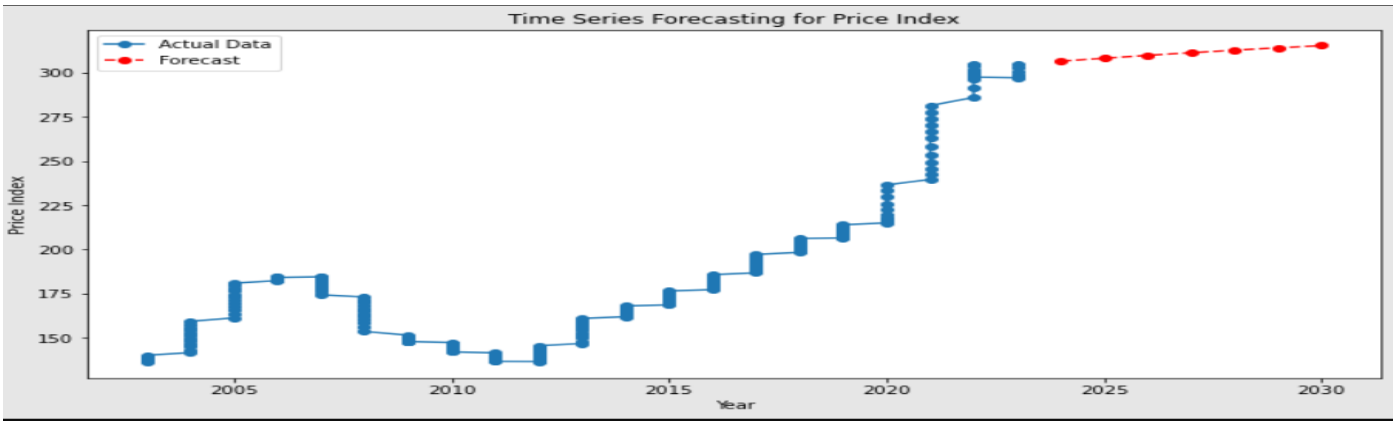
Analysing the coefficients of Gradient Boosting Regressor model provides insights into the importance and impact of each feature on the predicted target variable.



Predicted ⇕	Original ⇕
209.946671	222.391
158.872412	161.987
148.854712	148.090
142.453307	143.019
138.400212	136.294
139.695691	136.607
141.981702	140.350
141.135895	139.981
300.366360	303.762

In the above figure we can see the performance of the model and its result.

Additionally, we can forecast the future Price index using the ARIMA model using the previously available data of Year and Price Index. This is known as Time Series Forecasting



As form the above graph we can see the forecast for next 7 year in yellow points and it values are

Forecast for 2024: 306.4083863758612
Forecast for 2025: 308.1160230408474
Forecast for 2026: 309.7223053278464
Forecast for 2027: 311.23324898373687
Forecast for 2028: 312.65451269918304
Forecast for 2029: 313.9914193012055
Forecast for 2030: 315.2489756878963