**An Analysis of the US National Home Price Index for the president of property building of the US Federal Government**

Anish Sriniketh, senior analyst of property building for the United States Government

**Executive Summary**

The US government wants to be able to keep the US National Home Price Index as high as possible, so the economy thrives. The analysis that was done show that the factors that significantly affect the US National Home Price Index are the Consumer Confidence Index, the PPI Constraint Matrix, Inflation, Corporate Bond Yield, the Monthly Home Supply Index, the Share of the Working Population, GDP per Capita, Quarterly Real GDP, and Quarterly GDP Growth Rate. The ideal values of the factors are 138.4 for the Consumer Confidence Index, 143.8 for the PPI Constraint Matrix, 5.6% for Inflation, 6.75% for the Corporate Bond Yield, 12.2 for the Monthly Home Supply Index, 65.08% for the Share of Working Population, 37860 for the GDP per Capita, 19202.31 for the Quarterly Real GDP, and 7.55% for the Quarterly GDP Growth Rate. With these values, we can expect the US National Home Price Index to be 540.99 and we are 95% confident that the US National Home Price Index is between -631.259 and 1713.243. Only the share of the working population and the GDP per Capita have negative relationships with the US National Home Price Index, so the lowest values are ideal. For all the other factors, the greatest values are ideal as their relationships with the US National Home Price Index are positive. The model is very good in terms of being a quality predictor for the US National Home Price Index. The model explains 95.21% of the total variability of the US National Home Price Index and 95.01% of the variability of the US National Home Price Index that is affected by the other factors in the model. A regression model was used for this analysis. The process to make an ideal model took five regression models because we took out some factors that were not quality predictors for the US National Home Price Index.

**An Analysis of the US National Home Price Index**

The US government wants to be able to keep the US National Home Price Index as high as possible, so the economy thrives. The analysis that was done show that the factors that significantly affect the US National Home Price Index are the Consumer Confidence Index, the PPI Constraint Matrix, Inflation, Corporate Bond Yield, the Monthly Home Supply Index, the Share of the Working Population, GDP per Capita, Quarterly Real GDP, and Quarterly GDP Growth Rate. The ideal values of the factors are 138.4 for the Consumer Confidence Index, 143.8 for the PPI Constraint Matrix, 5.6% for Inflation, 6.75% for the Corporate Bond Yield, 12.2 for the Monthly Home Supply Index, 65.08% for the Share of Working Population, 37860 for the GDP per Capita, 19202.31 for the Quarterly Real GDP, and 7.55% for the Quarterly GDP Growth Rate. With these values, we can expect the US National Home Price Index to be 540.99 and we are 95% confident that the US National Home Price Index is between -631.259 and 1713.243. Only the share of the working population and the GDP per Capita have negative relationships with the US National Home Price Index, so the lowest values are ideal. For all the other factors, the greatest values are ideal as their relationships with the US National Home Price Index are positive. The model is very good in terms of being a quality predictor for the US National Home Price Index. The model explains 95.21% of the total variability of the US National Home Price Index and 95.01% of the variability of the US National Home Price Index that is affected by the other factors in the model.

The equation for developed for the US National Home Price Index is US National Home Price Index = 3982.465 + 0.373696\*Consumer Confidence Index – 0.4283\*PPI Constraint Matrix + 2.090451\*Inflation (%) + 5.664044\*Corporate Bond Yield + 5.672642\*Monthly Home Supply Index – 54.3206\*Share of Working Population – 0.00926\*GDP per Capita + 0.016705\*Quarterly Real GDP + 1.911611\*Quarterly GDP Growth Rate. (See Appendix A)

The greatest coefficient is the share of the working population which means that for every percent increase of the share of working population, the National Home Price Index decreases by 54.3206 points. The second greatest coefficient is the monthly home supply index which means that for every point increase of the monthly home supply index, the National Home Price Index increases by 5.672642 points. The third greatest coefficient is the corporate bond yield which means that for every percent increase of the corporate bond yield, the National Home Price Index increases by 5.66 points. These factors have the most significant effect on the US National Home Price Index for one point or percent change on the factors. (See Appendix A)

The factor with the greatest impact to the US National Home Price Index is the GDP per capita of the country with an impact of 255.956 on the Home Price Index. The factor with the second greatest impact is the percent share of the working population with an impact of 120.389 on the Home Price Index. The factor with the third greatest impact is the Quarterly Real GDP with an impact of 95.635 on the Home Price Index. The factor with the fourth greatest impact is the Monthly Home Supply Index with an impact of 50.487 on the Home Price Index. The factor with the fourth greatest impact is the Monthly Home Supply Index with an impact of 50.487 on the Home Price Index. These four factors have the largest effect on changing the value of the US National Home Price Index. (See Appendix A)

The model is very good in terms of being a quality predictor for the US National Home Price Index. The model explains 95.21% of the total variability of the US National Home Price Index and 95.01% of the variability of the US National Home Price Index that is affected by the other factors in the model. The coefficients of the model are also quality predictors of having significant relationships with the US National Home Price Index. We are at least 96.18% confident in all the other factors having a significant relationship with the US National Home Price Index that is not zero. Also, with many of the factors we are essentially 100% confident of this. Also, the residuals plots of the final model were all random and had no distinct pattern, so there was no need of manipulating or combining any of the variable. (See Appendix B)

The final model we built was the fifth regression model of the process. In the first regression model we had a total of thirteen factors that could have affected the US National Home Price Index. The first step was to take out the CPI All Items because we were only 10.54% that this factor was a quality predictor, and this was the factor we were least confident in being a quality predictor. In the second regression, we took out the Median Household Income because we were only 17.17% that this was a quality predictor, and this was the factor we were least confident in being a quality predictor in this regression. In the third regression, we took out the Unemployment Rate because we were only 44.02% confident that this was a quality predictor, and this was the factor we were least confident in being a quality predictor in this regression. In the fourth regression, we took out the Monthly Mortgage Interest Rate because we were only 74.80% confident that this was a good predictor, and this was the factor we were least confident in being a quality predictor in this regression. This led us to the fifth and final regression model where we were at least 96.18% confident that all the factors were quality predictors. None of the Residual Plots throughout the process had a distinct pattern, so we decided not to manipulate or combine any of the variables. (See Appendix C)

Our data was varied. Many of the variables had some skewness to the data, while others, such as Inflation, followed a normal distribution. Some variables had relatively low variance, such as the Corporate Bond Yield and the Share of the Working Population, while others, such as Median Household Income, had relatively high variance. The PPI Constraint Matrix had the largest correlation with the US National Home Price Index, while the next two highly correlated factors, CPI All Items and Monthly Mortgage Interest Rate, were not included in the final model. The Share of the Working Population had the lowest range even though the percentages were relatively high ranging from 65.08% to 67.30%. There were two major outliers in the Quarterly GDP Growth Rate data; however, we did not take those out because we wanted the entire data to obtain the most accurate model. The structure of the data for each variable was due to the nature of the variable. For example, variables, such as Inflation and Corporate Bond Yield, were percentages with low ranges and low variances, while other variables, such as GDP per capita, were large numbers with relatively high variance. The US National Home Price Index data has a high variance and is skewed to the right. There was multicollinearity with a few of the variables. CPI All Items was highly correlated with the PPI Constraint Matrix and the Quarterly Real GDP. CPI All Items was the first factor taken out of the model in the modeling process. (See Appendix D)

The data are from Kaggle, and the dataset is called Economic Factors affecting Housing Price Index. The data consists of thirteen factors that affect the US National Home Price Index and the US National Home Price Index. In total, the data is fourteen columns with 224 observations in each column. The data consist of is listed at the end of the month every month from January 2002 to January 2022. The data was missing median household income from 2021 and 2022 so we took the data out and looked at the data from 2002 to 2020. Also, a few of the Share of Working Population datapoints were missing in 2021 and 2022, but this was fixed when we took out the data because of the missing Median Household Income values. (See Appendix E)

**Appendix A - Model and Interpretation**

US National Home Price Index = 3982.465 + 0.373696\*Consumer Confidence Index – 0.4283\*PPI Constraint Matrix + 2.090451\*Inflation (%) + 5.664044\*Corporate Bond Yield + 5.672642\*Monthly Home Supply Index – 54.3206\*Share of Working Population – 0.00926\*GDP per Capita + 0.016705\*Quarterly Real GDP + 1.911611\*Quarterly GDP Growth Rate

Intercept: This means that if all the factors in the equation are equal to zero, then the US National Home Price would equal 3982.465.

Consumer Confidence Index: Each point increase in consumer confidence index increases the US National Home Price Index by 0.373696. The total impact of the consumer confidence index is 42.377 as the range is 113.4.

PPI Constraint Matrix: Each point increase in PPI Constraint Matrix decreases the US National Home Price Index by 0.4283. The total impact of the PPI Constraint Matrix is 44.629 as the range is 104.2.

Inflation (%): Each percent increase in Inflation increases the US National Home Price Index by 2.090451. The total impact of the consumer confidence index is 16.090 as the range is 7.697%.

Corporate Bond Yield (%): Each percent increase in Corporate Bond Yield increases the US National Home Price Index by 5.644044. The total impact of the Corporate Bond Yield is 26.019 as the range is 4.61%.

Monthly Home Supply Index: Each percent increase in the monthly home supply index increases the US National Home Price Index by 5.672642. The total impact of the monthly home supply index is 50.487 as the range is 8.9.

Share of Working Population (%): Each percent increase in the share of the working population decreases the US National Home Price Index by 54.3026. The total impact of the share of the working population is 120.389 as the range is 2.217.

GDP per Capita: Each point increase in GDP per capita decreases the US National Home Price Index by .00926. The total impact of the GDP per capita is 255.956 as the range is 27641.

Quarterly Real GDP: Each point increase in Quarterly Real GDP increases the US National Home Price Index by .016705. The total impact of the Quarterly Real GDP is 95.635 as the range is 5724.954.

Quarterly GDP Growth Rate (%): Each percent increase in Quarterly GDP Growth Rate increases the US National Home Price Index by 1.911611. The total impact of Quarterly GDP Growth Rate is 31.513 as the range is 16.485.

**Appendix B - Model Statistical Analysis**



The significance of the f-statistic being 4.8069\*10^-131 in this ANOVA test means that we are essentially 100% confident that there is a significant relationship between the US National Housing Price Index and the other factors.

***P-Values***



* Intercept: The intercept coefficient has a P-Value of 4.1 \* 10^-68 meaning that we are essentially 100% confident that the intercept is significant and not equal to zero.
* Consumer Confidence Index: The Consumer Confidence Index coefficient has a P-Value of 1.51 \* 10^-14 meaning that we are essentially 100% confident that the Consumer Confidence Index is significant and not equal to zero.
* PPI Constraint Matrix: The PPI Constraint Matrix coefficient has a P-Value of .006775 meaning that we are 99.32% confident that the PPI Constraint Matrix is significant and not equal to zero.
* Inflation (%): The Inflation coefficient has a P-Value of 4.55\*10^-5 meaning that we are essentially 100% confident that the PPI Constraint Matrix is significant and not equal to zero.
* Corporate Bond Yield (%): The Corporate Bond Yield coefficient has a P-Value of .001937 meaning that we are 99.81% confident that the Corporate Bond Yield is significant and not equal to zero.
* Monthly Home Supply Index: The Monthly Home Supply Index coefficient has a P-Value of 6.8\*10^-19 meaning that we are essentially 100% confident that the Monthly Home Supply Index is significant and not equal to zero.
* Share of Working Population (%): The Share of Working Population coefficient has a P-Value of 1.27\*10^-59 meaning that we are essentially 100% confident that Share of Working Population is significant and not equal to zero.
* GDP per Capita: The GDP per Capita coefficient has a P-Value of 1.62\*10^-5 meaning that we are essentially 100% confident that GDP per Capita is significant and not equal to zero.
* Quarterly Real GDP: The Quarterly Real GDP coefficient has a P-Value of .038187 meaning that we are 96.18% confident that the Quarterly Real GDP is significant and not equal to zero.
* Quarterly GDP Growth Rate (%): Quarterly GDP Growth Rate coefficient has a P-Value of 1.63\*10^-5 meaning that we are essentially 100% confident that Quarterly GDP Growth Rate is significant and not equal to zero.

***SUMMARY STATISTICS***



R-Squared: The R-squared of .952485 means that the model explains 95.25% of the variability of the US National Home Price Index.

Adjusted R-Squared: The Adjusted R-squared of .94954 means that the model explains 94.95% of the variability of the US National Home Price Index that is affected by the other variables in the model, which is more relevant for us because we are using the model to predict the future.

Standard Error: The standard error of 8.016635 means that 95% of the actual values of the variables are within two standard errors from the predicted model values in the regression.

***RESIDUAL PLOTS***

Consumer Confidence Index: This residual plot has no pattern and is random, so we did not try to change the variable. There is some variability in the residuals.

PPI Constraint Matrix: This residual plot has no pattern and is random, so we did not try to change the variable. The variability of the residuals is very low.

Inflation (%): This residual plot has no pattern and is random, so we did not try to change the variable. There is some variability in the residuals.

Corporate Bond Yield (%): This residual plot has no pattern and is random, so we did not try to change the variable. The variability of the residuals is very low.

Corporate Bond Yield (%): This residual plot has no pattern and is random, so we did not try to change the variable. The variability of the residuals is very low.

Share of Working Population (%): This residual plot has no pattern and is random, so we did not try to change the variable. The residual plot has many points at the same x-values because there is low variability in the share of working population data.

GDP per Capita: This residual plot has no pattern and is random, so we did not try to change the variable. The variability of the residuals is very low.

Quarterly Real GDP: This residual plot has no pattern and is random, so we did not try to change the variable. The variability of the residuals is very low.

Quarterly Real GDP: This residual plot has no pattern and is random, so we did not try to change the variable. The variability of the residuals is very low.

***NORMAL PROBABILITY PLOT***

The data is close to normally distributed as the normal probability plot is close to a straight, diagonal line until the end where the slope slightly increases.

**Appendix C - Model Development**



In the first regression the P-Value of the CPI All Items coefficient was 0.894598, which means that we were only 10.54% confident that the CPI All Items is significant and not equal to zero, so we took that out of the model because it was not a quality predictor. The R-squared and adjusted R-squared were relatively high at .952485 and .949544, so the model was able to predict 95.25% of the total variability of the US National Home Price Index and 94.95% of the variability of US National Home Price Index that is affected by the other variables in the model.

In the second regression the P-Value of the Median Household Income coefficient was 0.828274, which means that we were only 17.17% confident that the Median Household Income is significant and not equal to zero, so we took that out of the model because it was not a quality predictor. The R-squared slightly decreased and adjusted R-squared slightly increased and were relatively high at .952481 and .949779. The difference in the values became less making the model more accurate.



In the third regression the P-Value of the Unemployment Rate coefficient was 0.559821, which means that we were only 44.02% confident that the Unemployment Rate is significant and not equal to zero, so we took that out of the model because it was not a quality predictor. The R-squared slightly decreased and adjusted R-squared slightly increased and were relatively high at .95247 and .950004. The difference in the values became less making the model more accurate.

In the fourth regression the P-Value of the Monthly Mortgage Interest Rate coefficient was 0.252048, which means that we were only 74.80% confident that the Monthly Mortgage Interest Rate is significant and not equal to zero, so we took that out of the model because it was not a quality predictor. The R-squared slightly decreased and adjusted R-squared slightly increased and were relatively high at .952394 and .950159. The difference in the values became less making the model more accurate.

In the fifth regression all the P-Values were less than .05 meaning that all the dependent variable significant and not equal to zero, so they were quality predictors. The R-squared slightly increased and adjusted R-squared slightly decreased and were relatively high at .952099 and .950085. The difference in the values increased slightly; however, the P-Values were all below 0.05 and both the R-squared values were above 0.95, so this is the best model for the data.

**Appendix D - Data Analysis**

**Unemployment Rate (%)**

**Table

Description automatically generated Chart, histogram

Description automatically generated**

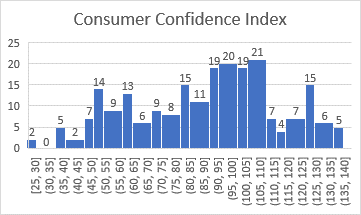
**Chart, scatter chart

Description automatically generated**

The data is skewed to the right as most of the values are between 3.5 and 5.5, which is the very low end of the data.

**Consumer Confidence Index**

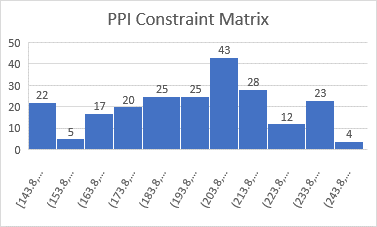
**A picture containing table

Description automatically generated** 

This data has slight skewness to the left but has very high variability.

**PPI Constraint Matrix**

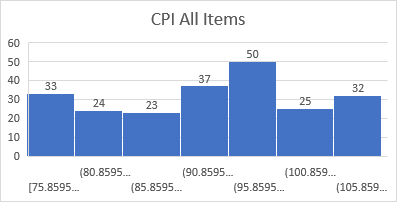
**Table

Description automatically generated with low confidence** 

The PPI Constraint Matrix has a relatively strong correlation of decreasing as the CSUSHPISA increases.

**CPI All Items**

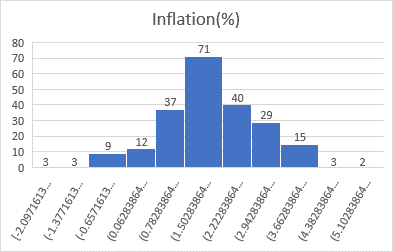
Table

Description automatically generated with low confidence 

The data is close to normally distributed and has a strong correlation of decreasing as CSUSHPISA increases.

**Inflation (%)**

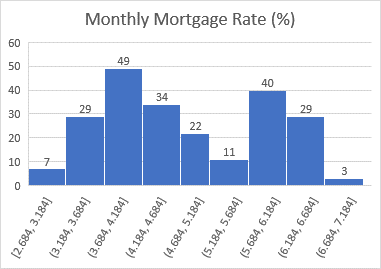
A picture containing table

Description automatically generated 

The data follows a normal distribution with very low variance.

**Monthly Mortgage Interest Rate**

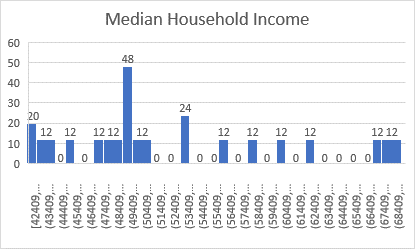
**Table

Description automatically generated** 

The data is range is very low because it is a percentage, so it seems to have a strong correlation with the response variable.

**Median Household Income**

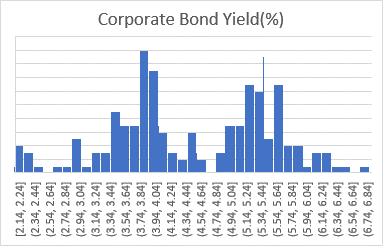
**A picture containing table

Description automatically generated** 

The variance of this variable is very high, and the data is highly skewed to the right.

**Corporate Bond Yield (%)**

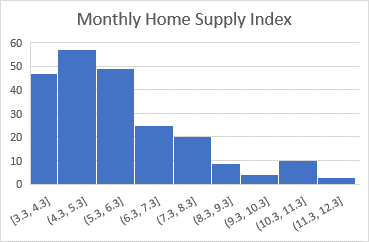
Table

Description automatically generated with low confidence 

The data looks highly correlated with the response variable and the variance is low as the data is stored as a percentage.

**Monthly Home Supply Index**

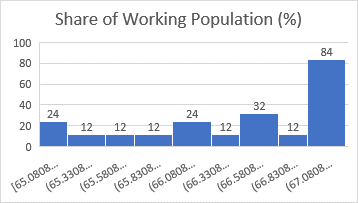
**Table

Description automatically generated with low confidence** 

The data is highly skewed to the right but does not have a high variance.

**Share of Working Population (%)**

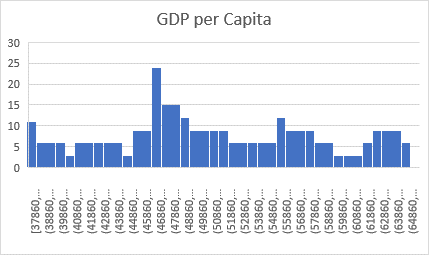
Table

Description automatically generated with medium confidence 

The data is highly skewed to the left and has a very low variance.

**GDP per Capita**

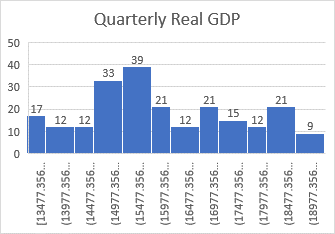
Table

Description automatically generated with medium confidence

The variance is very high, while the data closely follows a normal distribution.

**Quarterly Real GDP**

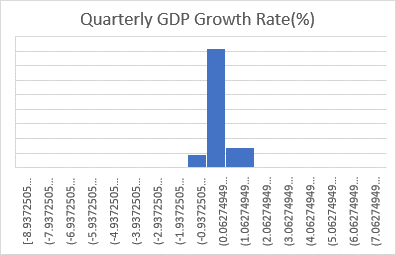
A picture containing table

Description automatically generated 

The variance is very high, while the data closely follows a normal distribution.

**Quarterly GDP Growth Rate**

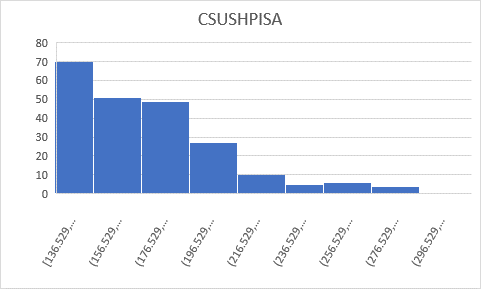
**Table

Description automatically generated with medium confidence** 

The variance is very low, however there are two major outliers.

**US National Home Price Index (Response Variable)**





The data is highly skewed to the right and the variance is relatively high.

**Correlation Matrix**



**Appendix E – Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Removed Data** | **Why** | **Impact** |
| **Median Household Income** | All rows with null values | Incomplete Data | Removed 24 rows |
| **Share of Working Population (%)** | All rows with null values | Incomplete Data | Removed 7 rows |

After data cleaning:

|  |  |
| --- | --- |
| Columns | 14 |
| Observations | 224 |

This model is representative of economic factors affecting the US National Housing Price Index from 2002-2020.

This data was sourced from Kaggle, an online database platform. This data can be accessed with this link:

<https://www.kaggle.com/code/sagarvarandekar/economic-factors-affecting-housing-price-index/data>