Problem Statement:
Find publicly available data for key *supply-demand* 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.

Before we get in to coding part lets try to understand what are the possible supply and demand factors that might influence the problem statement Here are some of the most relevant factors influencing supply and demand in the US housing market:

Supply Factors:

1.Housing Inventory: You can find data on the number of homes listed for sale in the US. Websites like Zillow, Realtor.com, and Redfin provide such data.

2. Construction Data: Look for data on new housing construction, including building permits, housing starts, and completions. The U.S. Census Bureau and the National Association of Home Builders provide relevant data.

3.Land Availability: Check for data on available land for residential development. Local planning departments and real estate associations may have this information.

4.Labor and Material Costs: Economic data on labor and material costs can impact the supply of housing. Economic research organizations and government agencies like the Bureau of Labor Statistics provide relevant data.

5. Regulatory Environment: Regulations affecting home construction and zoning laws can significantly influence supply. Local government websites and housing advocacy groups might have this information.

Demand Factors:

1.Population Growth: You can find data on population growth at the national and local levels from the U.S. Census Bureau.

2.Economic Indicators: Look for data on employment rates, income levels, and GDP growth, as these factors affect people's ability to buy homes. The Bureau of Labor Statistics and the Bureau of Economic Analysis provide such data.

3. Interest Rates: Data on mortgage interest rates can be obtained from sources like the Federal Reserve or financial news websites.

4. Consumer Confidence: Surveys like the Consumer Confidence Index can give insights into consumer sentiment regarding the housing market.

5. Demographic Trends: Analyze demographic data, such as age distribution and household formation trends. The U.S. Census Bureau offers detailed demographic data.

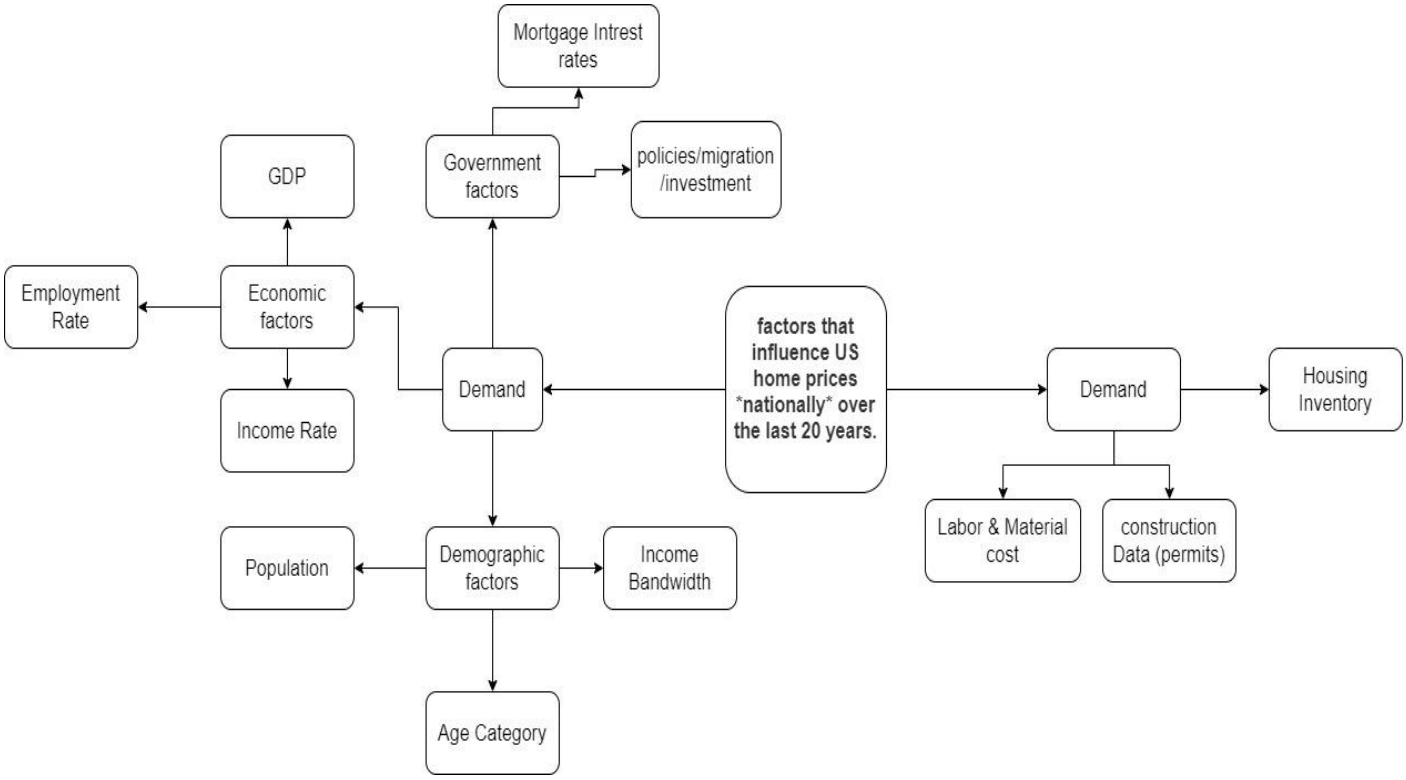
6. Government Policies: Government policies, such as first-time homebuyer incentives or tax credits, can influence demand. Government websites and news sources can provide information on these policies.

7. Rental Market: Data on the rental market, such as vacancy rates and rental prices, can impact housing demand. Real estate websites and the U.S. Census Bureau offer relevant data.

8.Migration Patterns: Migration patterns can affect demand for housing in specific regions. The Census Bureau and state-level government agencies often track migration data.

Out of all above discussed factors , i will be considering publicly available data from FRED ECONOMIC DATA (“fred.stlouisfed.org/series/CSUSHPISA.”) and URBAN INSTITUTE (“https://www.urban.org/”)

factors that influence US home prices *nationally* over the last 20 years.



Demographic

- population

Income-age distribution

- average expenditure 25-34
- average expenditure 35-44
- average expenditure 45-54
- avg-expenditure-55-64

Mortgages

- HCAI_GOVT
- HCAI_GSE
- HCAI_PP

- MORTGAGE30US

Health of the economy

- GDP
- CPI
- private_job_gains
- personal_saving_rate
- UNRATE - Unemployment rate
- unrate_construction - Unemployment rate in construction industry

Construction Industry

- employees_construction
- industrial_production_cement
- pvt_owned_house_under_const
- residential_const_val
- producer_price_index_concrete_brick

Housing industry

- houses-for-sale-to-sold - Number of houses for sale vs number of houses got sold
- home-ownership-rate
- house_units_completed - Number of new house units completed in a given month
- retail_sales_home_furnishing_stores - Sales of home furnishing stores

Infrastructure and permits

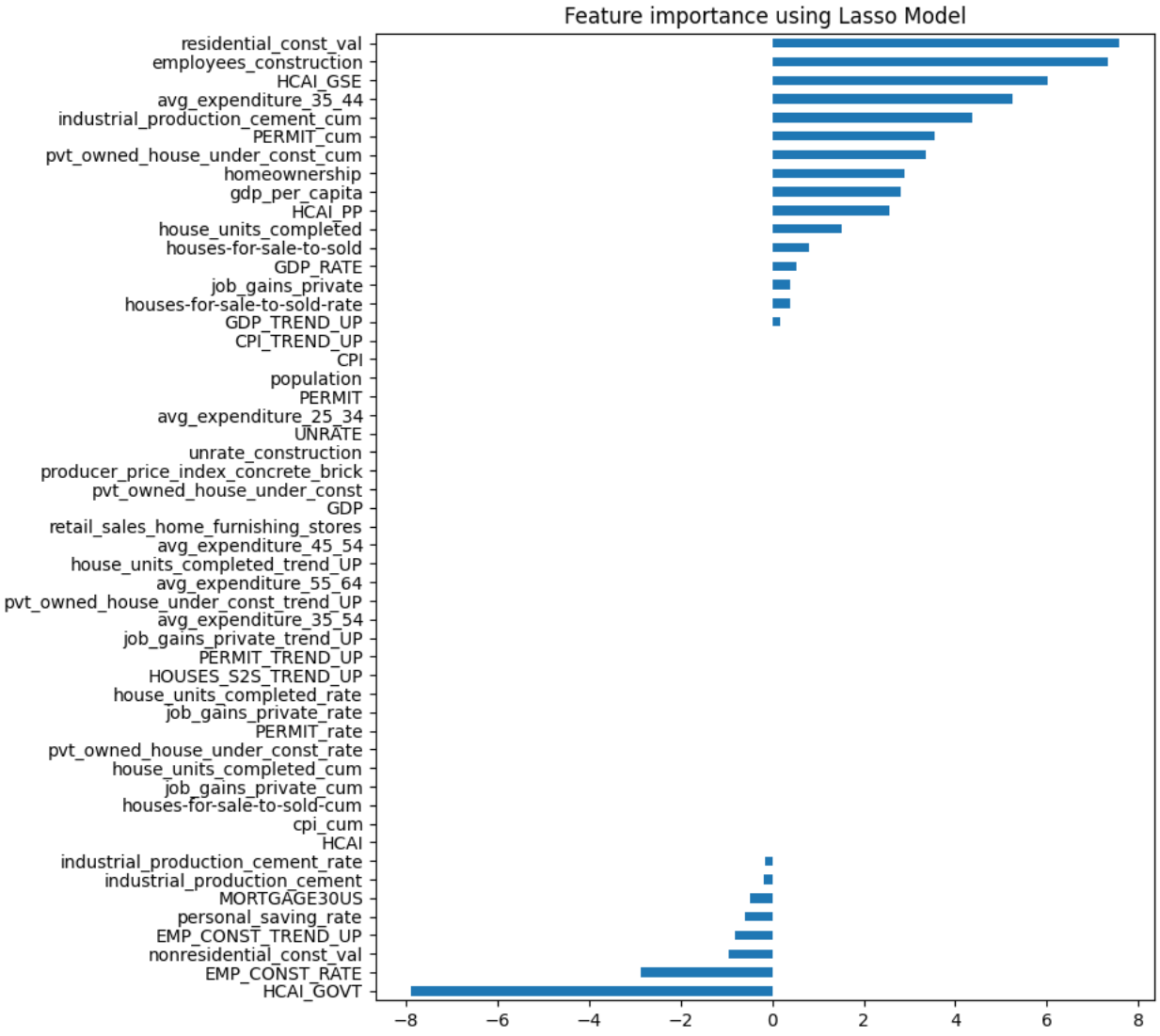
- nonresidential_const_val
- Permits

We've encountered features in both annual and quarterly frequencies, and in order to standardize our data, we've transformed them into a monthly frequency. This transformation assumes an equal distribution of changes in feature values over each month. Considering that the average duration for the completion of residential buildings in the U.S. falls within the range of 8 to 12 months, we've chosen a window size of 12 months for our analysis.

We are particularly interested in capturing the rate of change and trends within these features. While some features don't exhibit a linear correlation with our target variable, we've observed that their derivatives have a notable impact on prices. To account for these derivatives, we've implemented several techniques, including cumulative sums, rolling sums (typically spanning 12 months), rate of change over the past 12 months, and the introduction of a categorical variable denoting trends as 'UP' for uptrends and 'DOWN' for downtrends."

Feature finalizing

Best alpha using built-in LassoCV: 0.01471



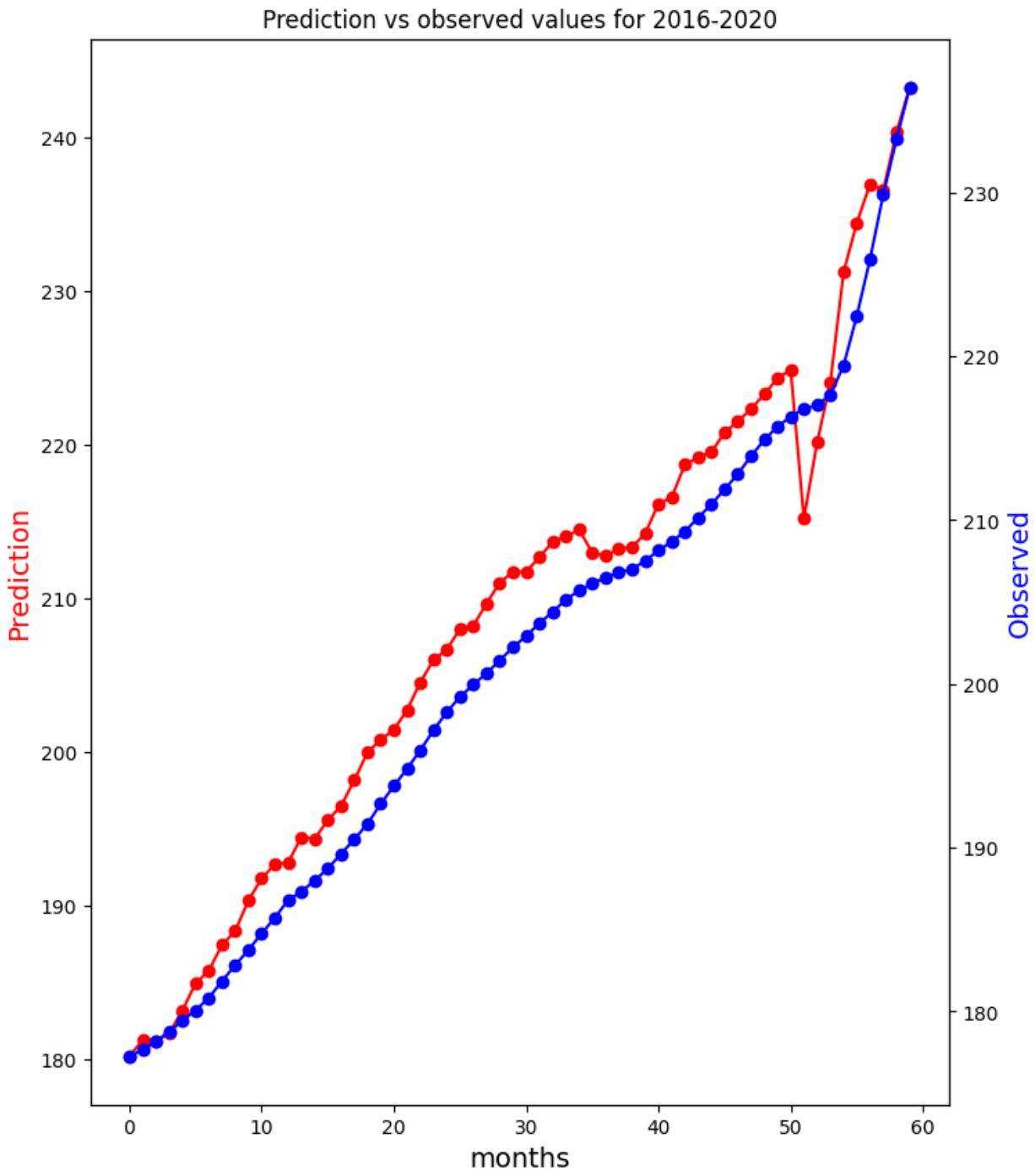
Features with zero coef are eliminated

```
coef[coef==0]
```

```
avg_expenditure_55_64                   -0.0
avg_expenditure_45_54                   -0.0
avg_expenditure_25_34                    0.0
GDP                                      0.0
pvt_owned_house_under_const              0.0
PERMIT                                   0.0
population                               0.0
UNRATE                                  -0.0
unrate_construction                      0.0
retail_sales_home_furnishing_stores      0.0
producer_price_index_concrete_brick     -0.0
avg_expenditure_35_54                    0.0
HCAI                                    -0.0
cpi_cum                                  0.0
houses-for-sale-to-sold-cum              0.0
job_gains_private_cum                    0.0
house_units_completed_cum                0.0
pvt_owned_house_under_const_rate        -0.0
PERMIT_rate                             -0.0
job_gains_private_rate                  -0.0
house_units_completed_rate              -0.0
HOUSES_S2S_TREND_UP                     -0.0
PERMIT_TREND_UP                          0.0
job_gains_private_trend_UP               0.0
pvt_owned_house_under_const_trend_UP     0.0
house_units_completed_trend_UP           0.0
dtype: float64
```

```
((coef[coef!=0]).sort_values(ascending=False))
```

```
residential_const_val                7.567529
employees_construction               7.338326
HCAI_GSE                             6.014372
avg_expenditure_35_44                5.260055
industrial_production_cement_cum     4.381133
PERMIT_cum                           3.546097
pvt_owned_house_under_const_cum      3.343789
homeownership                        2.885602
gdp_per_capita                       2.811986
HCAI_PP                              2.557347
house_units_completed                1.512792
houses-for-sale-to-sold              0.799481
GDP_RATE                             0.522801
job_gains_private                    0.400222
houses-for-sale-to-sold-rate         0.387155
GDP_TREND_UP                         0.169724
CPI_TREND_UP                         0.013554
CPI                                  0.012670
industrial_production_cement_rate   -0.141455
industrial_production_cement        -0.175783
MORTGAGE30US                        -0.478568
personal_saving_rate                -0.603005
EMP_CONST_TREND_UP                  -0.818836
nonresidential_const_val            -0.961708
EMP_CONST_RATE                      -2.875535
HCAI_GOVT                           -7.892093
dtype: float64
```



METRICS

MSE = 56.691090505213886 RMSE = 7.529348610949946 R2 = 0.7418

the R^2 is 0.7347. This means that the independent variables in your regression model explain approximately 73.47% of the variance in the dependent variable. It indicates a reasonably good fit of the model to the data.

In summary, MSE and RMSE measure the accuracy of a regression model, with RMSE providing a more interpretable error metric. R-squared (R^2) quantifies how well the model explains the variation in the dependent variable, with higher values indicating a better fit. In your specific case, the model has an MSE of 58.2454, an RMSE of 7.6319, and an R^2 of 0.7347, indicating that the model is reasonably accurate and explains a significant portion of the variance in the dependent variable.
NOTE: data considered from 2000-2021 as per publicly available data