Home Price Index Prediction Solution

Created	@November 11, 2023 8:09 PM
Author	Ⓗ Hao Quan
Tags	

Overview

The S&P CoreLogic Case-Shiller Home Price Indices are the leading measures of U.S. residential real estate prices, tracking changes in the value of residential real estate nationally. The objective of this project is to predict the future values of this home price index using relevant indicators from the FRED database.

Data Collection

To ensure that we do not violate the no-look-ahead principle, for every piece of data we acquire from the FRED database, we select the first release version, which is not seasonally adjusted, without incorporating future values into the averaging process.

Target indicator (Dependent variable)

`home_price`: S&P/Case-Shiller U.S. National Home Price Index (CSUSHPISA)

Potentially useful indicators

- Federal Funds Effective Rate (FEDFUNDS)
- Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis(GS10)
- 30-Year Fixed Rate Mortgage Average in the United States (MORTGAGE30US)
- Personal Saving Rate (PSAVERT)

- Consumer Price Index for All Urban Consumers: All Items in U.S. City Average (CPIAUCSL)
- Consumer Price Index for All Urban Consumers: All Items Less Food and Energy in U.S. City Average(CPILFESL)
- Unemployment Rate (UNRATE)

Data Preprocessing

Data Summary

- Size of the dataset
  - `home_price`: 584 monthly datapoints, from `1975-01-01` to `2023-08-01`
  - Other indicators: we look at the same time range as `home_price`
  - Variables type: all `float64` numbers

Datetime alignment

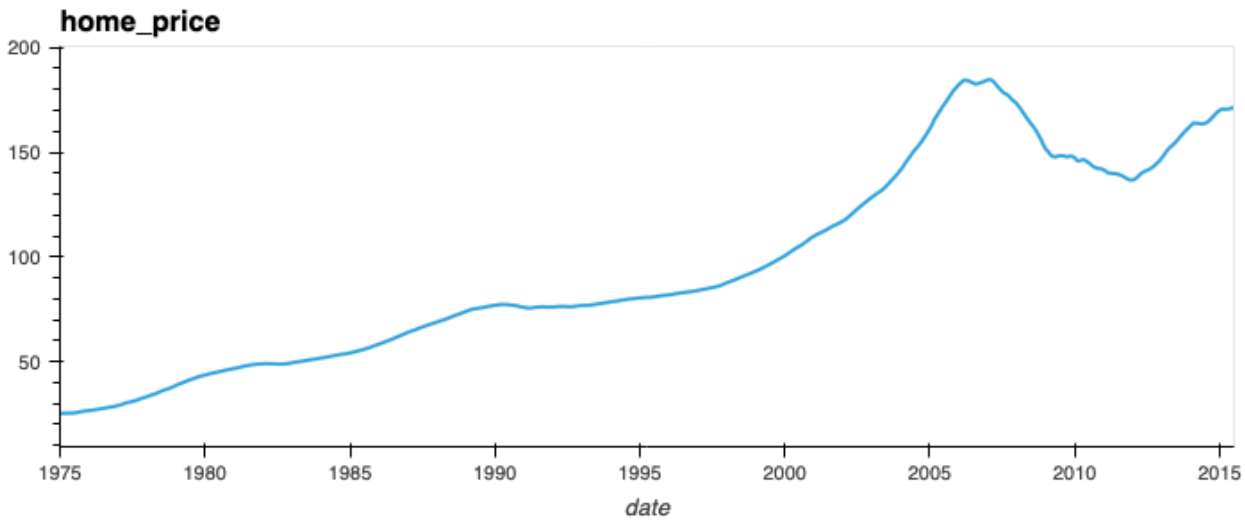
- All indicators except mortgage rate are released monthly on the first day of each month.
- For mortgage rate, we merge it into the dataframe of all the other indicators by left join using the datetime index in the backward direction (to ensure no lookahead bias).
  - e.g. we align mortgage rate on `1974-12-27` with the monthly data on `1975-01-01`

Train Test split

- We follow a 5-to-1 split of the whole time series dataset
  - Train `1975-01-01` to `2015-07-01`
  - Test: `2015-08-01` to `2023-08-01`
- For the following exploratory data analysis, we only look at data in the training set.

Other issues

- No missing value observed
- Outliers
    - No obvious outlier for home price (maybe except 2008 credit crisis)

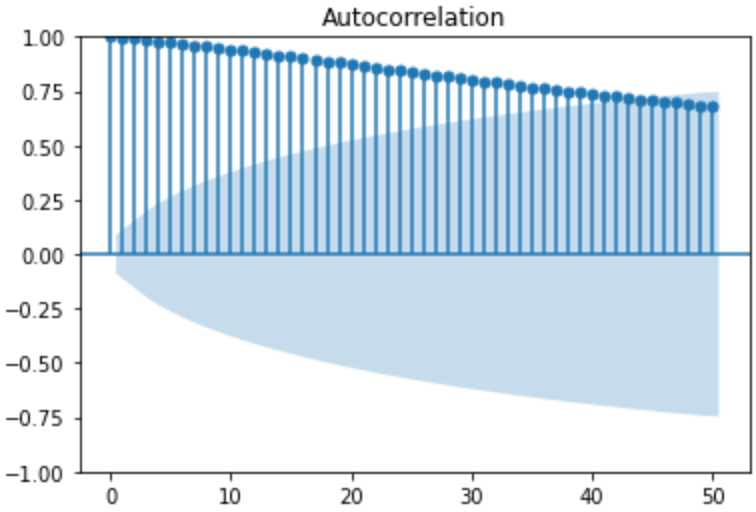


Feature Engineering

Home Price Index

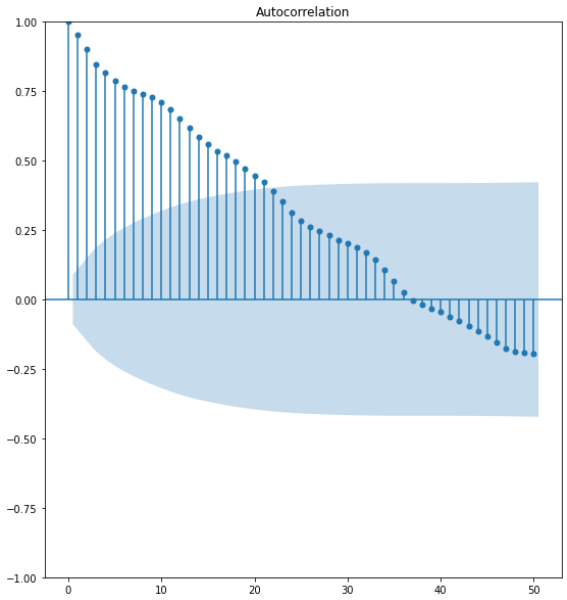
- Trend

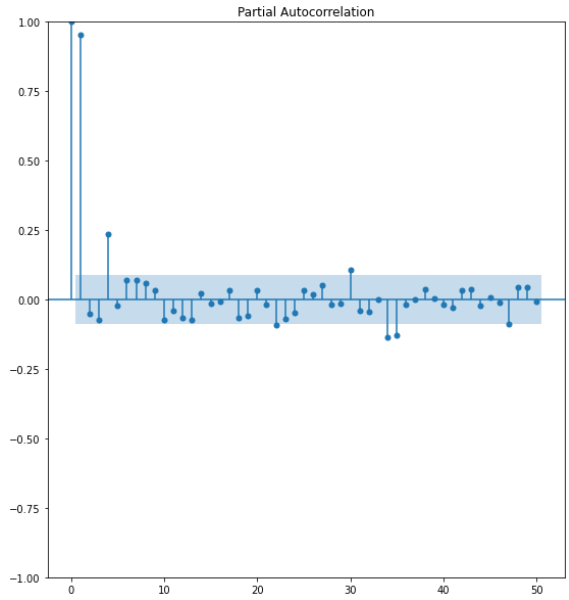
  From the home price plot, we can see a general upward trend over time, which makes the series non-stationary. This is also reflected in the autocorrelation plot.



To address the autocorrelation, we take the log difference or the percentage change. In fact, these two look almost the same for home prices. For the remaining part of the project, we consider the log difference as the dependent variable in the model.

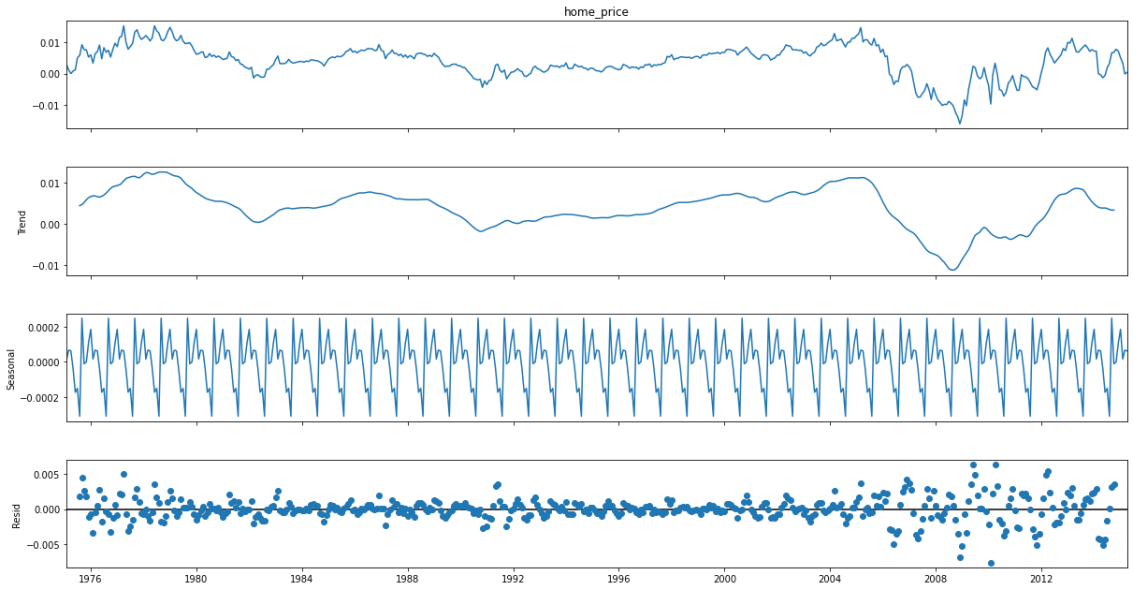
Now we check the autocorrelation and partial correlation. The autocorrelation decreases, and the partial correlation roughly shows lag-1 as the only significant autocorrelation.



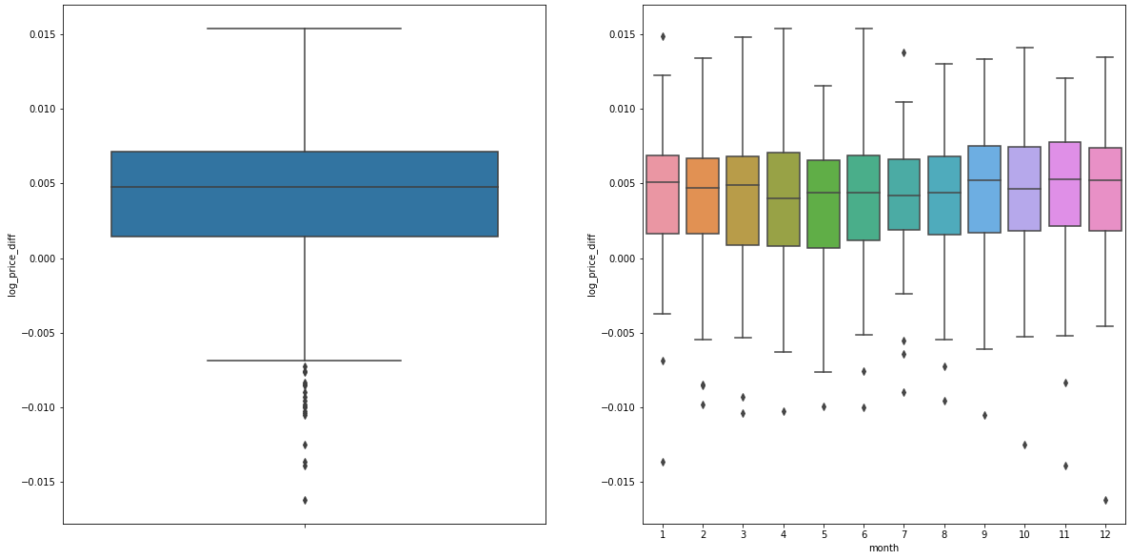


The result of the Adjusted Dickey Fuller test also provides a p-value of 0.039765, indicating that we can reject the null hypothesis of the existence of a unit root in the series. Looking at the log difference plot, the mean appears to fluctuate around a positive mean very close to 0, without any clear upward or downward trend, which is consistent with test result as well.

- Seasonality
  - There are certainly cycles in the time series, but there is no clear repeated seasonal pattern, and there is a large increase in variance after 2005.
  - In the following STL decomposition plot of the home price log difference, the increasing variance of residuals confirms the observation above.



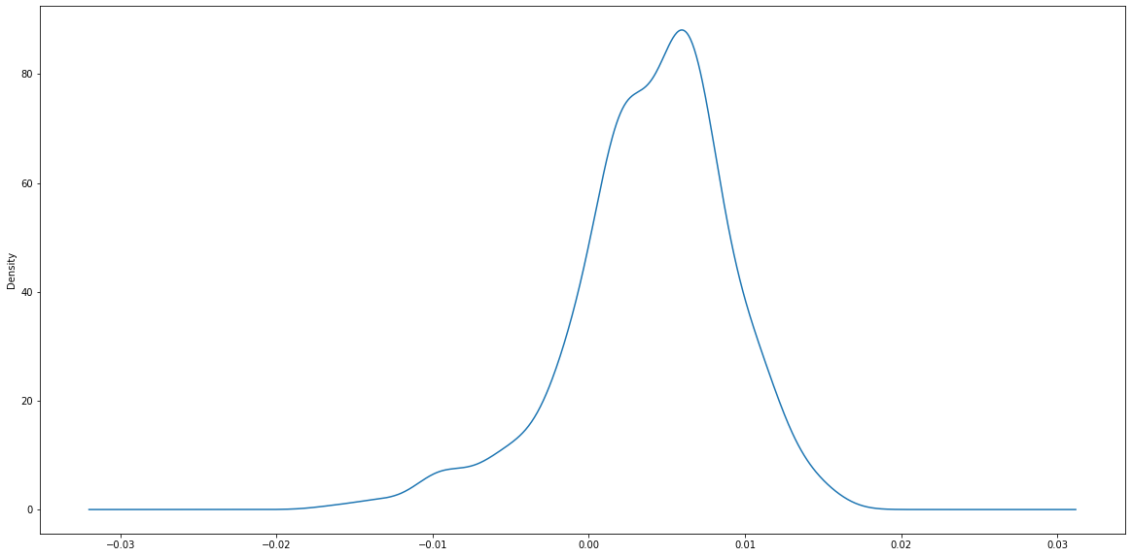
- From the monthly box plot below, we do not see a clear monthly pattern either (maybe the median is slightly higher from September to December at year end).

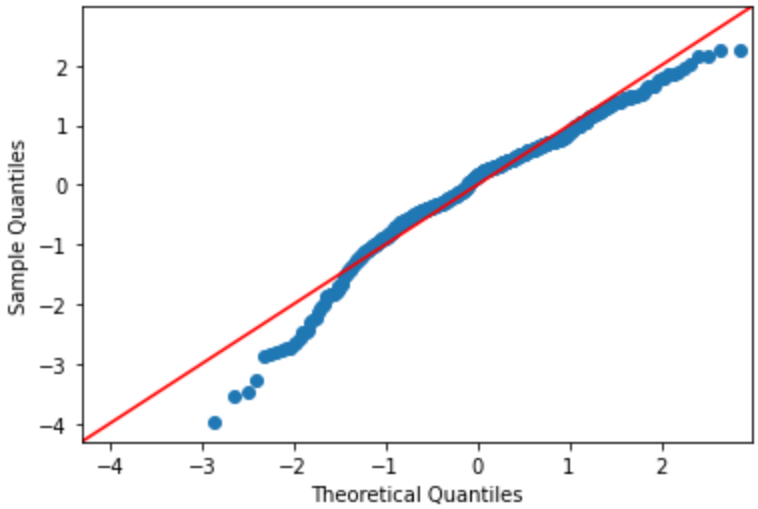


Note that there are several **outliers** in the log difference plot; however, these mostly occur in singular events like the 2008 credit crisis and COVID-19. Typical approaches include removing outliers or imputing with statistical mean, interpolation, or regression. However, these singular events are real historical events that happened in reality and had, and still have, a huge impact on the economy, not to mention home prices. So, the naive practice of typical approaches would simply rewrite history, which we do not want to do. To deal

with issues like this, we need a more sophisticated approach to model large tail events. However, I will not consider this for the sake of the current project and will leave it for future research.

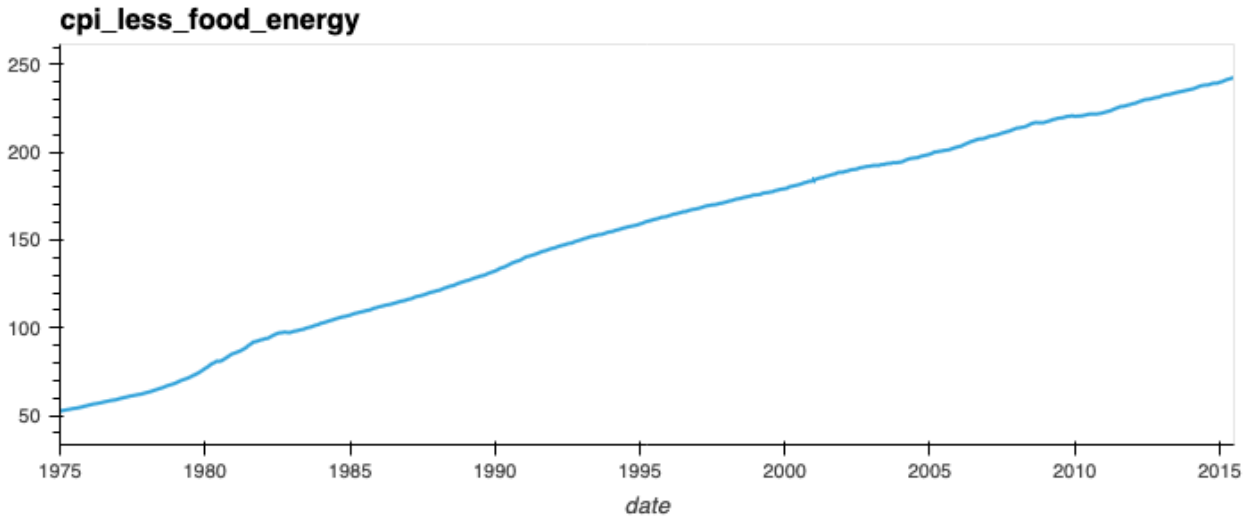
- From the distribution plot and QQ-plot against a Gaussian distribution, now the distribution of the dependent variable looks a lot more similar to a Gaussian distribution (there is a deviation from the Gaussian on the left tail, but we ignore this issue for the moment).



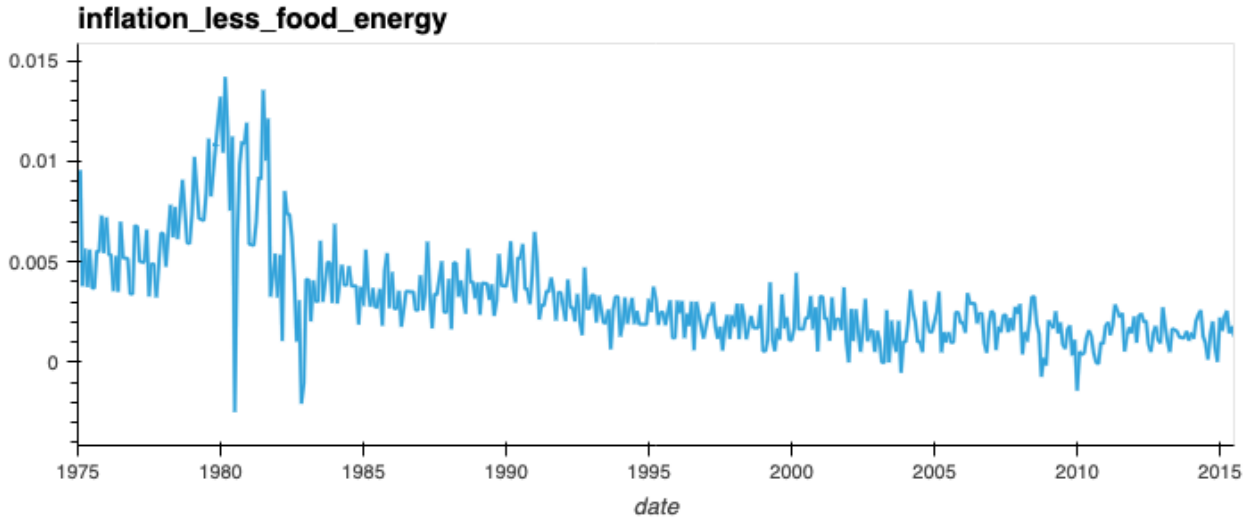


CPI

The CPI is clearly trending upwards, but it does not provide much meaningful information. From an economic perspective, it does not reveal much about the actual cost of living on its own either.



However, inflation, the percentage change of CPI, offers a lot more information than CPI.



Rates

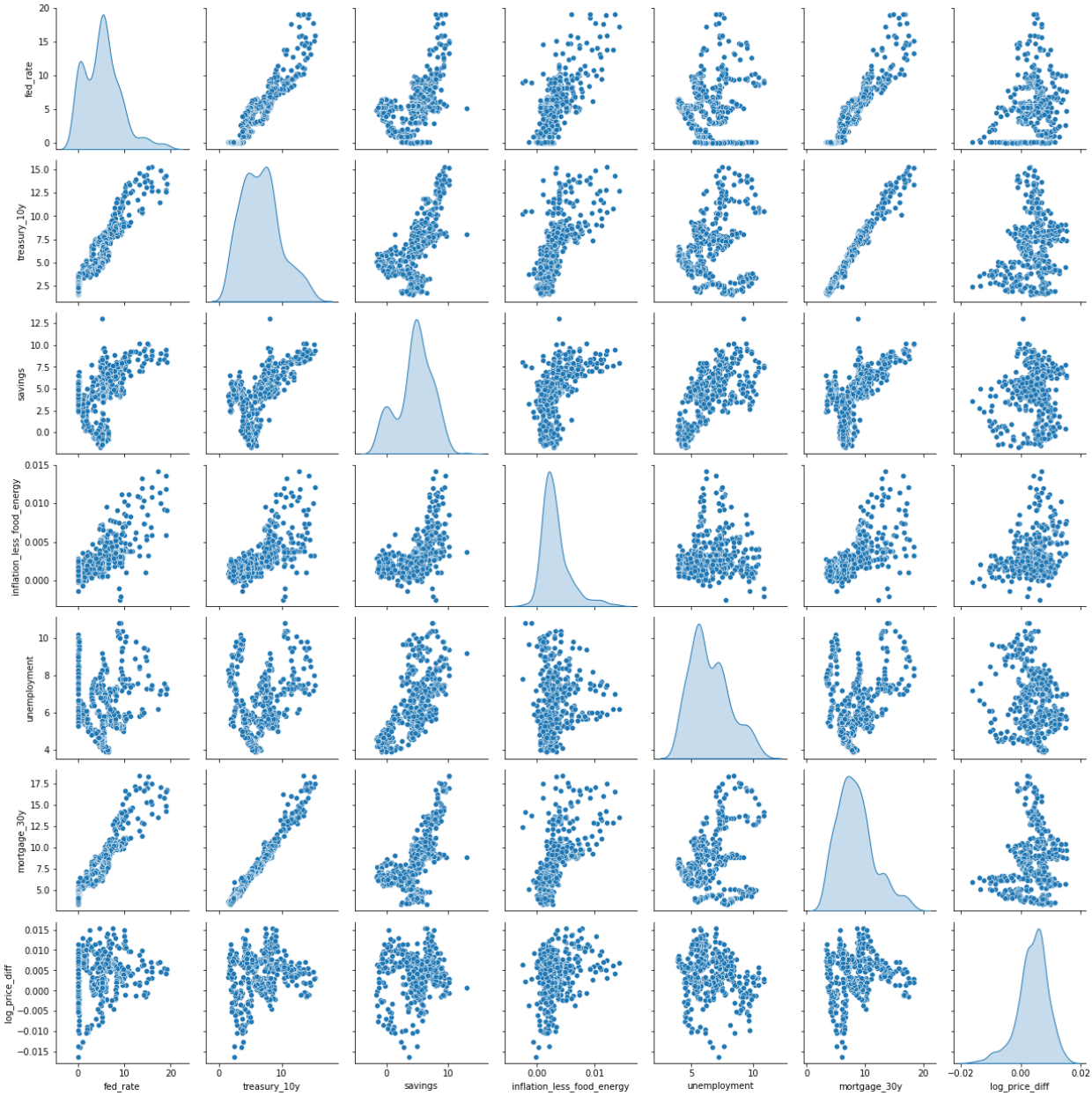
We also considered several rate data, which are presented in percentage units, which is a larger scale than the log difference of home price. To avoid the potential problem of a

high condition number in solving linear equations, we convert these rates to decimals by dividing them by 100.

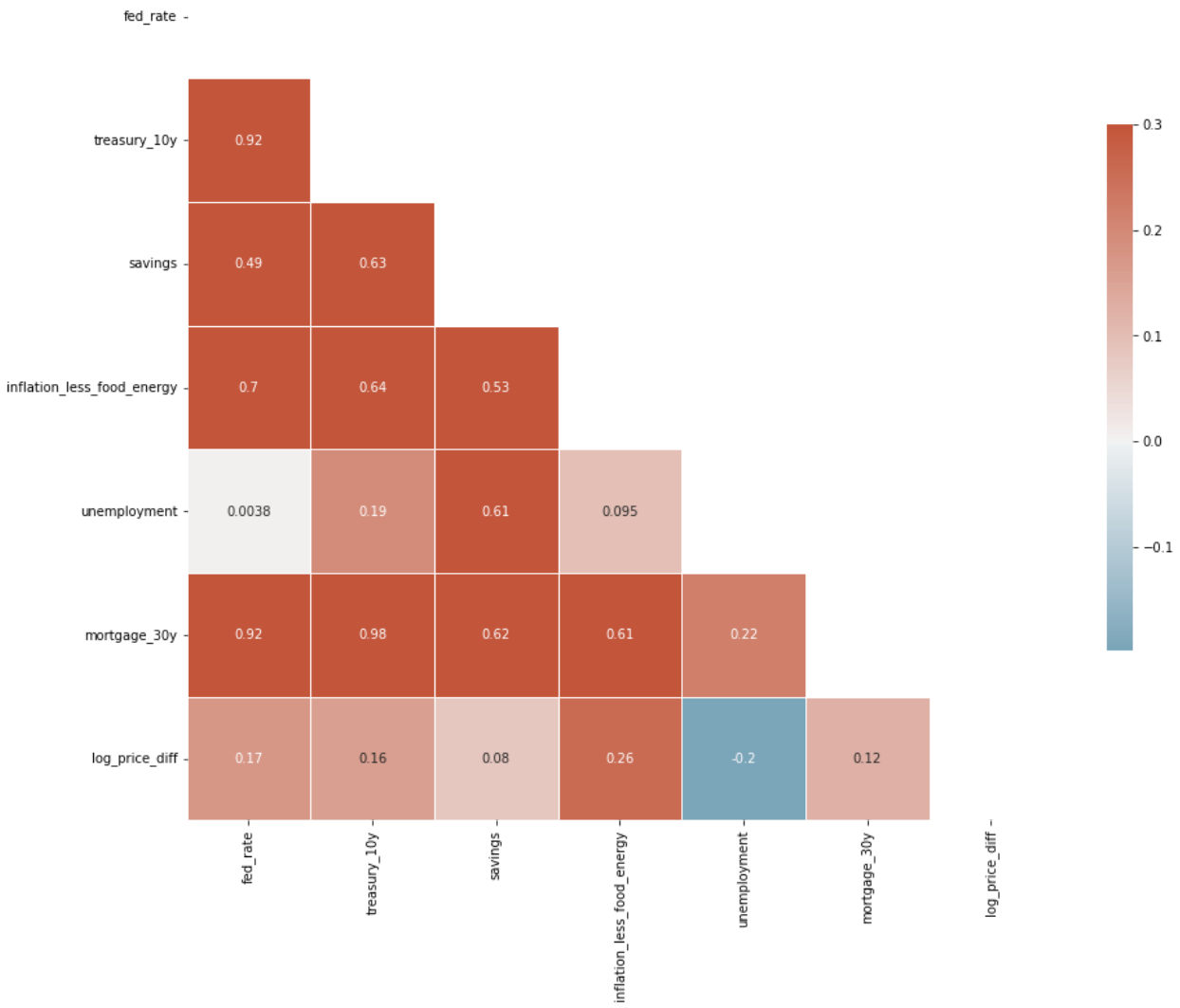
Pairwise plot

From the pairwise scatter plot, we see clear linear relationships among

- Federal rate
- Yield on 10-year treasury
- Mortgage rate
- (savings rate, not as clear as the first three)



From the correlation map below, we see that indeed these factors above are highly correlated. Thus, we should only consider one of them to avoid multicollinearity. Since we are predicting home price, the mortgage rate is a directly related index and makes more sense to be considered as a predictor in the model.

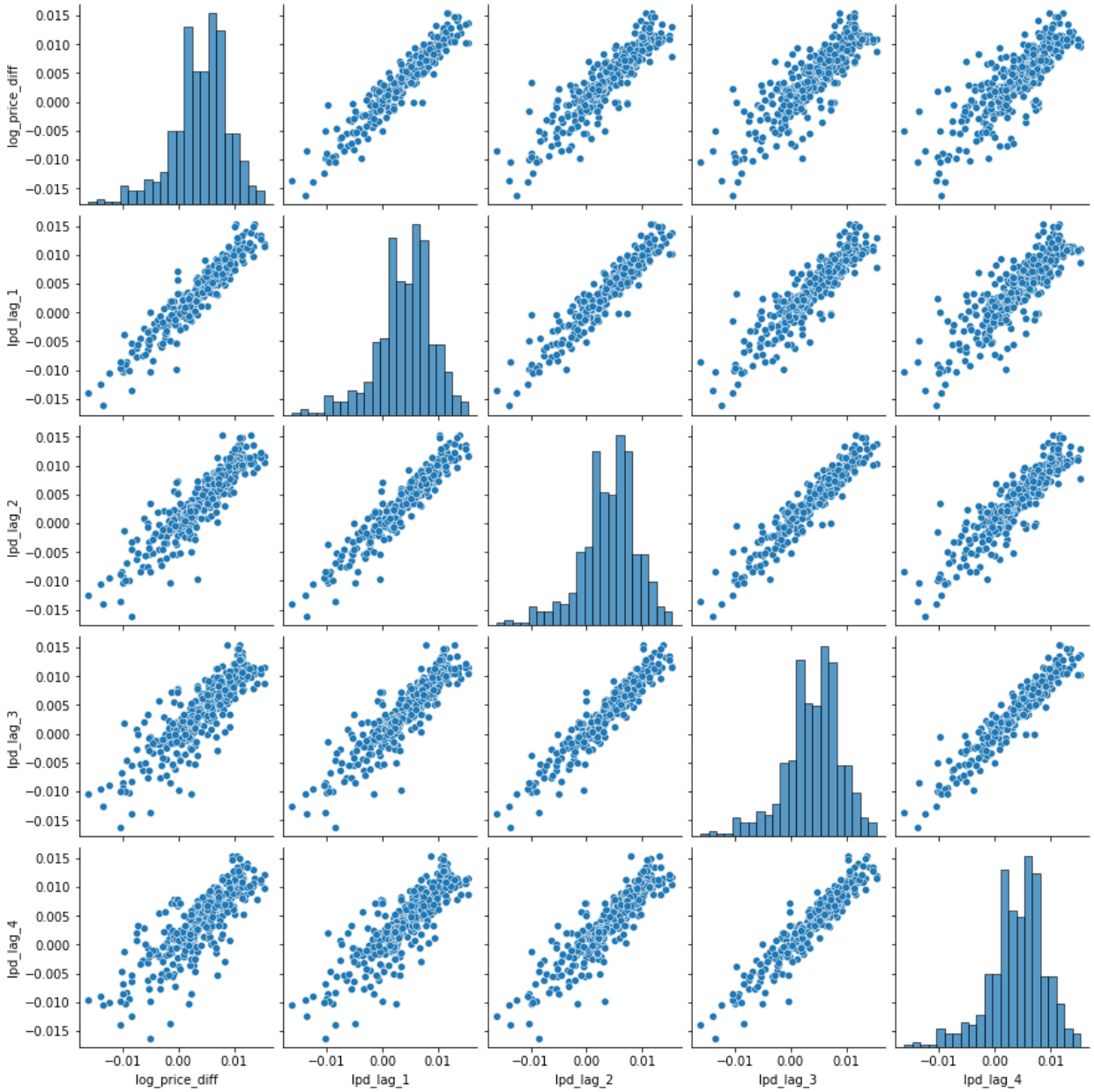


The observation that inflation and unemployment are correlated with the home price change aligns with empirical experience. With high inflation, most assets become expensive, leading to a rise in home prices. With a high unemployment rate, people lose their jobs and may face financial difficulties and pessimism about the future. This results in a decrease in the number of buyers in the market and a subsequent drop in prices.

Lags

Although we applied the log transform and took the difference, there is still autocorrelation in the series. This suggests that we should examine some lags of the log price difference. Based on the observations from the partial correlation plot (excluding effects from other lags), we will consider lags from order 1 to order 4.

From the pairwise plot of the log price difference and its different lags, we can see clear linear relationships for every pair. This makes sense as the S&P home price is a composite index of national house markets, which changes slowly on a monthly basis. Therefore, we will only consider one of them in the model, specifically lag 1.



Model

Predictors

Based on our analysis and feature engineering, we consider the following indicators as predictors

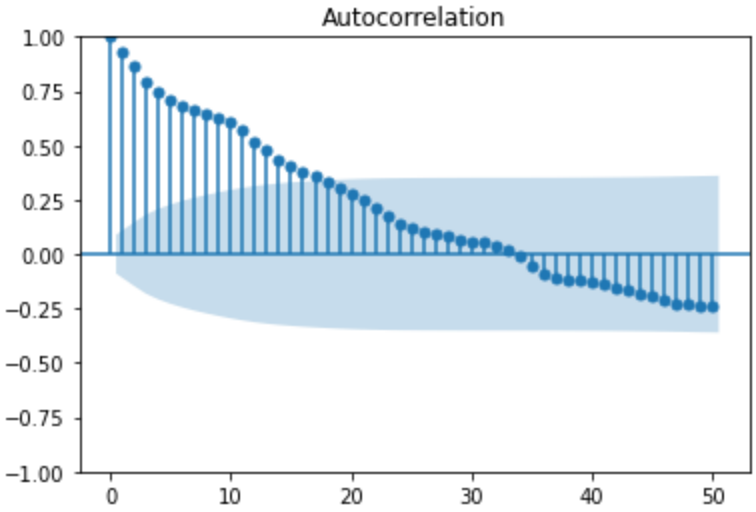
- savings rate
- inflation
- unemployment
- mortgage rate
- lag 1 of log price difference

Linear Regression without lag

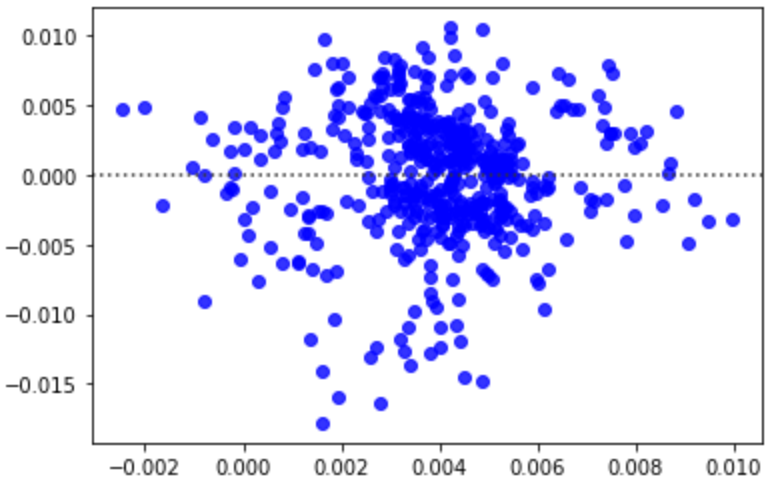
We first try linear regression without time lag, and the result is as follows.

```
Results: Ordinary least squares
============================================================================
Model:                 OLS               Adj. R-squared:       0.123
Dependent Variable:    log_price_diff    AIC:                  -3818.7219
Date:                  2023-11-11 18:14  BIC:                  -3797.7908
No. Observations:      486               Log-Likelihood:       1914.4
Df Model:              4                 F-statistic:          18.04
Df Residuals:          481               Prob (F-statistic):   8.14e-14
R-squared:             0.130             Scale:                2.2420e-05
----------------------------------------------------------------------------
                           Coef.  Std.Err.    t    P>|t|   [0.025  0.975]
----------------------------------------------------------------------------
Intercept                  0.0086   0.0012  7.3023 0.0000  0.0062  0.0109
savings                    0.0365   0.0132  2.7647 0.0059  0.0106  0.0625
inflation_less_food_energy  0.4545   0.1183  3.8403 0.0001  0.2219  0.6870
unemployment              -0.1053   0.0184 -5.7289 0.0000 -0.1415 -0.0692
mortgage_30y              -0.0104   0.0094 -1.1012 0.2714 -0.0289  0.0082
----------------------------------------------------------------------------
Omnibus:                43.729          Durbin-Watson:             0.145
Prob(Omnibus):          0.000           Jarque-Bera (JB):          58.038
Skew:                   -0.688          Prob(JB):                  0.000
Kurtosis:               3.986           Condition No.:             556
============================================================================
```

R^2 is only about 0.12, not explaining the variance of log price difference well. p-value for mortgage rate is large, indicating its coefficient is not statistically significant different from 0.



Autocorrelation of residual



Residual plot against fitted value of linear model

There is still autocorrelation in the residual, so we try autoregressive model taking lags into consideration.

Regression with Lag

- ARMA - 4 features

SARIMAX Results

Dep. Variable:	log_price_diff	No. Observations:	486
Model:	ARIMA(1, 0, 0)	Log Likelihood	2449.015
Date:	Sat, 11 Nov 2023	AIC	-4884.030
Time:	18:18:47	BIC	-4854.727
Sample:	02-01-1975	HQIC	-4872.517
	- 07-01-2015		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0007	0.003	0.214	0.830	-0.006	0.007
savings	-0.0030	0.008	-0.367	0.713	-0.019	0.013
inflation_less_food_energy	0.1411	0.041	3.472	0.001	0.061	0.221
unemployment	0.0261	0.037	0.712	0.476	-0.046	0.098
mortgage_30y	0.0027	0.028	0.098	0.922	-0.052	0.057
ar.L1	0.9272	0.013	70.031	0.000	0.901	0.953
sigma2	2.502e-06	1.04e-07	24.119	0.000	2.3e-06	2.71e-06

Ljung-Box (L1) (Q):	2.56	Jarque-Bera (JB):	265.53
Prob(Q):	0.11	Prob(JB):	0.00
Heteroskedasticity (H):	2.26	Skew:	-0.15
Prob(H) (two-sided):	0.00	Kurtosis:	6.61

After adding Lag-1 using the ARIMA model, AIC and BIC decrease by 1000 compared to the first linear model. The p-values of Lag 1 and inflation are very low, suggesting the statistical significance of these two predictors. Since the p-values of the constant term and mortgage rate are very high, we will next remove these two in ARIMA.

- ARMA - 3 features

SARIMAX Results

Dep. Variable:	log_price_diff	No. Observations:	486
Model:	ARIMA(1, 0, 0)	Log Likelihood	2340.385
Date:	Sat, 11 Nov 2023	AIC	-4670.770
Time:	18:20:51	BIC	-4649.839
Sample:	02-01-1975	HQIC	-4662.547
	- 07-01-2015		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
savings	-0.0091	0.010	-0.901	0.367	-0.029	0.011
inflation_less_food_energy	0.7501	0.031	24.483	0.000	0.690	0.810
unemployment	0.0253	0.016	1.584	0.113	-0.006	0.057
ar.L1	0.9191	0.017	54.589	0.000	0.886	0.952
sigma2	3.826e-06	2.23e-07	17.126	0.000	3.39e-06	4.26e-06

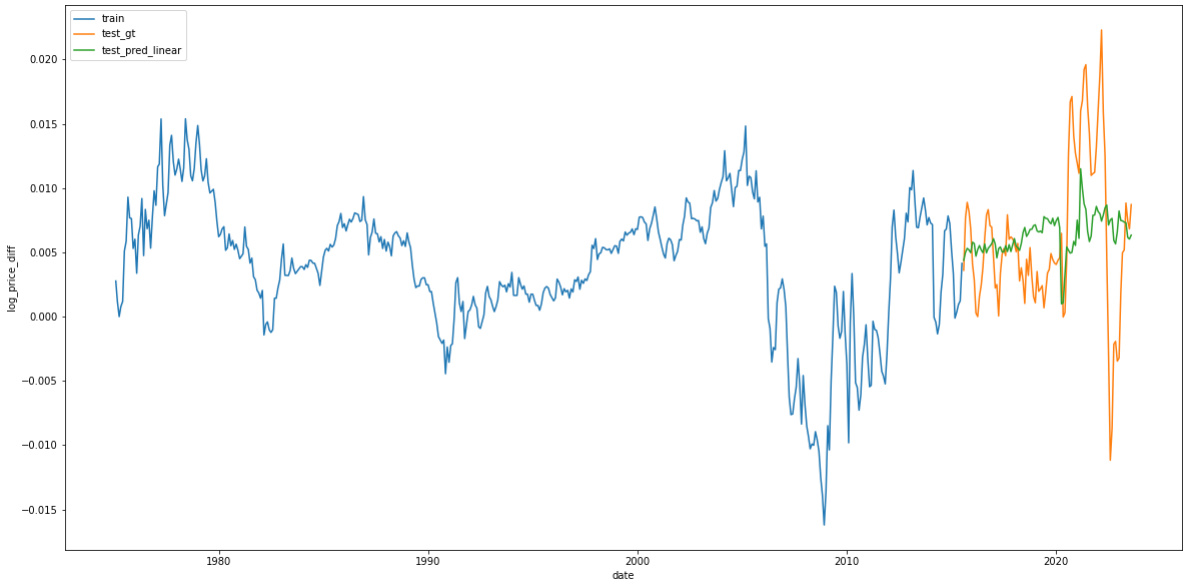
Ljung-Box (L1) (Q):	4.99	Jarque-Bera (JB):	206.18
Prob(Q):	0.03	Prob(JB):	0.00
Heteroskedasticity (H):	1.02	Skew:	0.13
Prob(H) (two-sided):	0.91	Kurtosis:	6.18

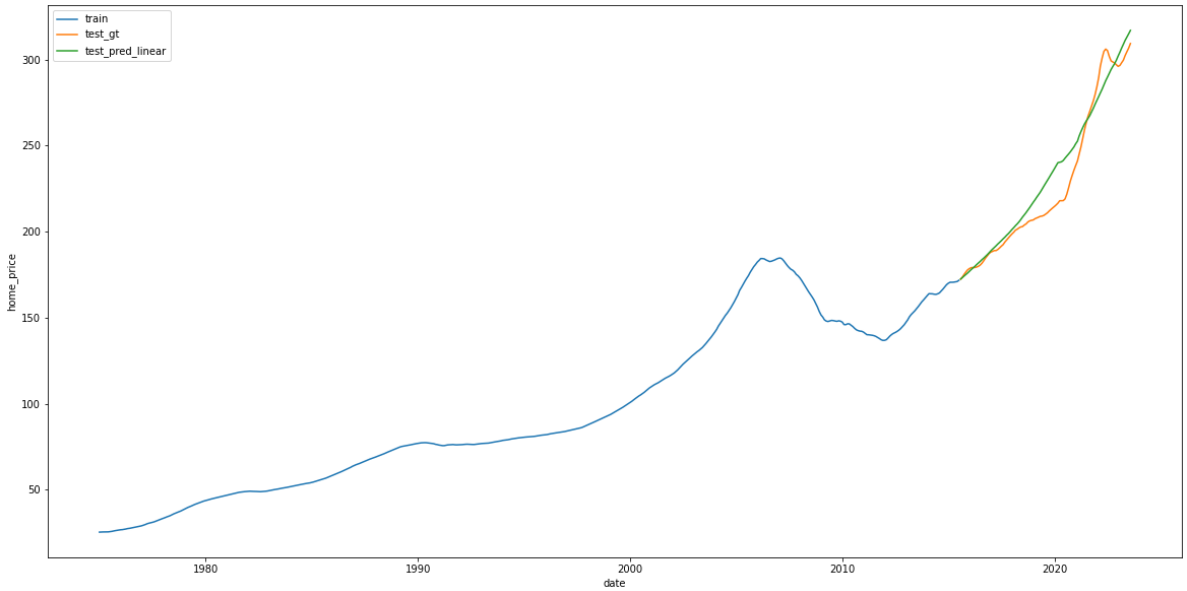
Performance Assessment

We tested all three models on the test set from `2015-08-01` to `2023-08-01`.

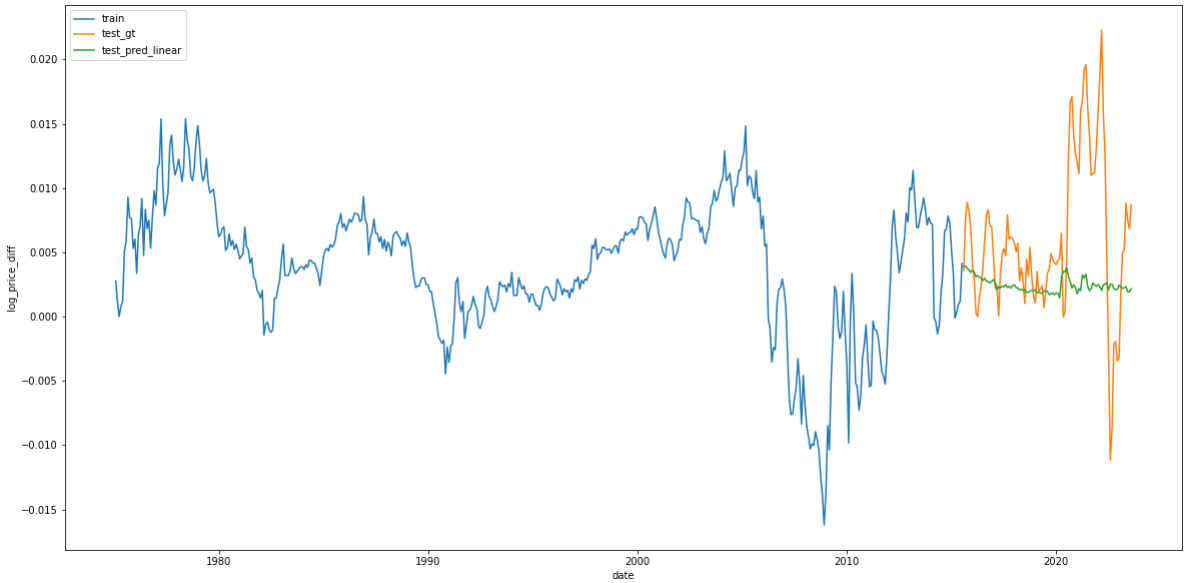
	Linear Model	ARMA - 4	ARMA - 3
RMSE - log diff	0.0059	0.0070	0.0068
MAE - log diff	0.0045	0.0051	0.0051
RMSE - price index	11.0803	45.4754	45.0763
MAPE - price index	8.3703	32.5156	32.8244

- Linear Model
  - Although there is a significant difference in the log difference prediction with a much smaller variance, the model captures the upward trend after being transferred into the price range.

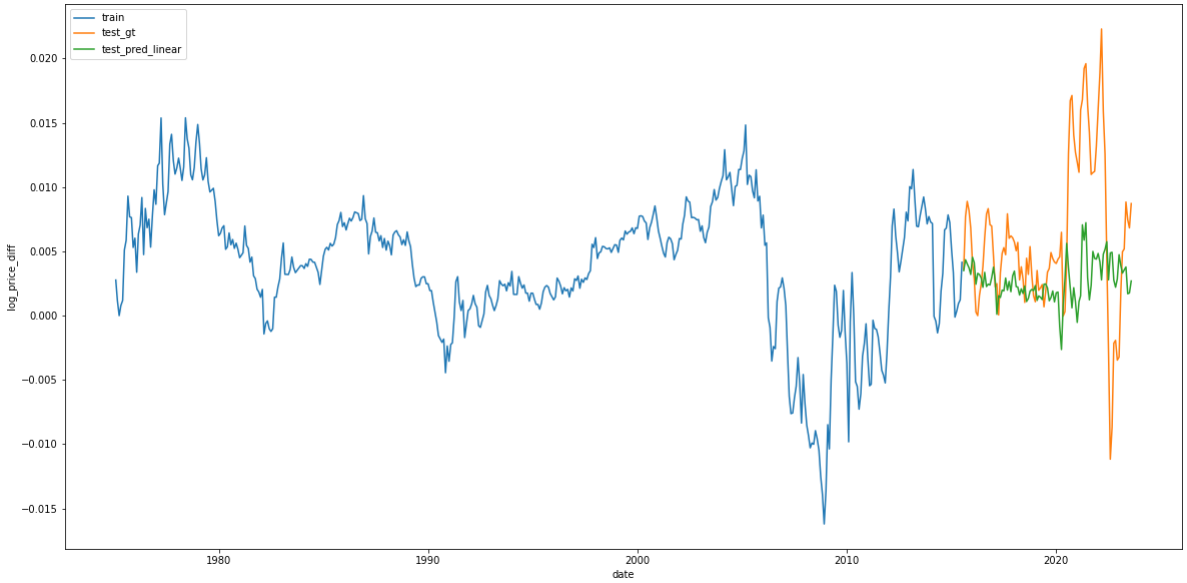




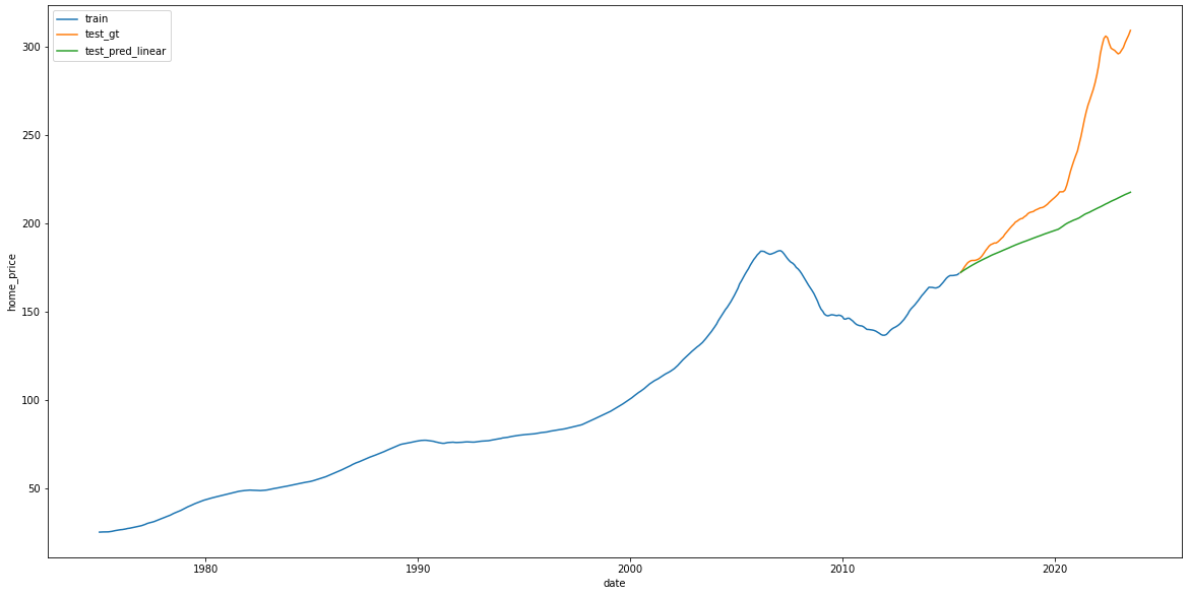
- AR
  - Both autoregressive models perform poorly on the test data, but the AR model with 3 features performs better with a larger variance.



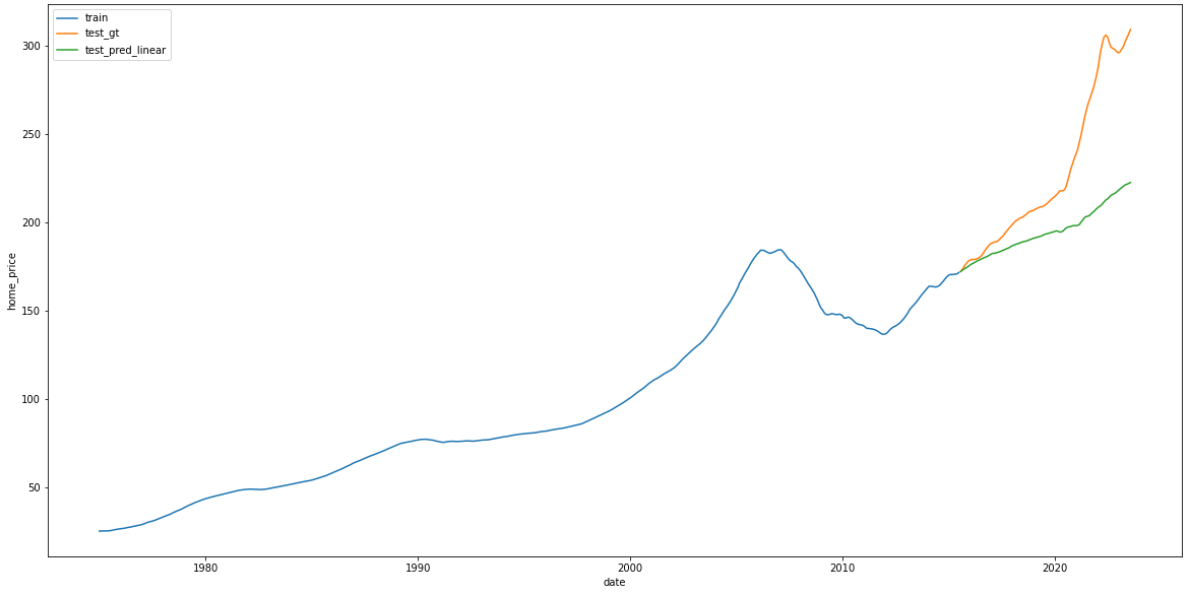
AR - 4 features



AR - 3 features



AR - 4 features



AR - 3 features

Conclusion

The indicators considered in this project were selected based on fundamental economic reasoning, as the dependent variable, the home price index, is a macro-economic statistic.

Through exploratory data analysis, correlation analysis, and modeling, we found that inflation and the unemployment rate are more correlated with changes in home prices. These indicators are useful for predicting the home price index in linear models.

Macro-economic indexes, like the home price index, typically change slowly over time. Therefore, it is more meaningful to capture the long-term trend and focus on explaining changes. Linear models and autoregressive models are good options for this purpose.

Based on the assessment of model performance, we found that though having autocorrelation, the linear model performs the best. It accurately captures price changes and achieves the smallest prediction errors. Autoregressive models give too much weight to previous time points, making them slower in capturing price changes.