ARIMA MODEL FOR INFLATION RATE FORECASTING

Muhammad Iqbal Rustan

MIND MAP

Use Case

Data Understanding

Data Preparation

Exploratory Data Analysis

Modeling

Forecasting

USE CASE

Objective Statement

- Get insight about how is the value of inflation descriptively
- Get insight how is the rate of inflation in the future

Challenges

- Large amounts of data are required
- Inappropriate model selection can affect the accuracy of predictions

Expected Outcome

- Know how is the value of inflation descriptively
- Know how is the rate of inflation in the future

Methodology

- Descriptive statistics
- AutoRegressive Integrated Moving Average (ARIMA)

Source Data

- www.bi.go.id

DATA UNDERSTANDING

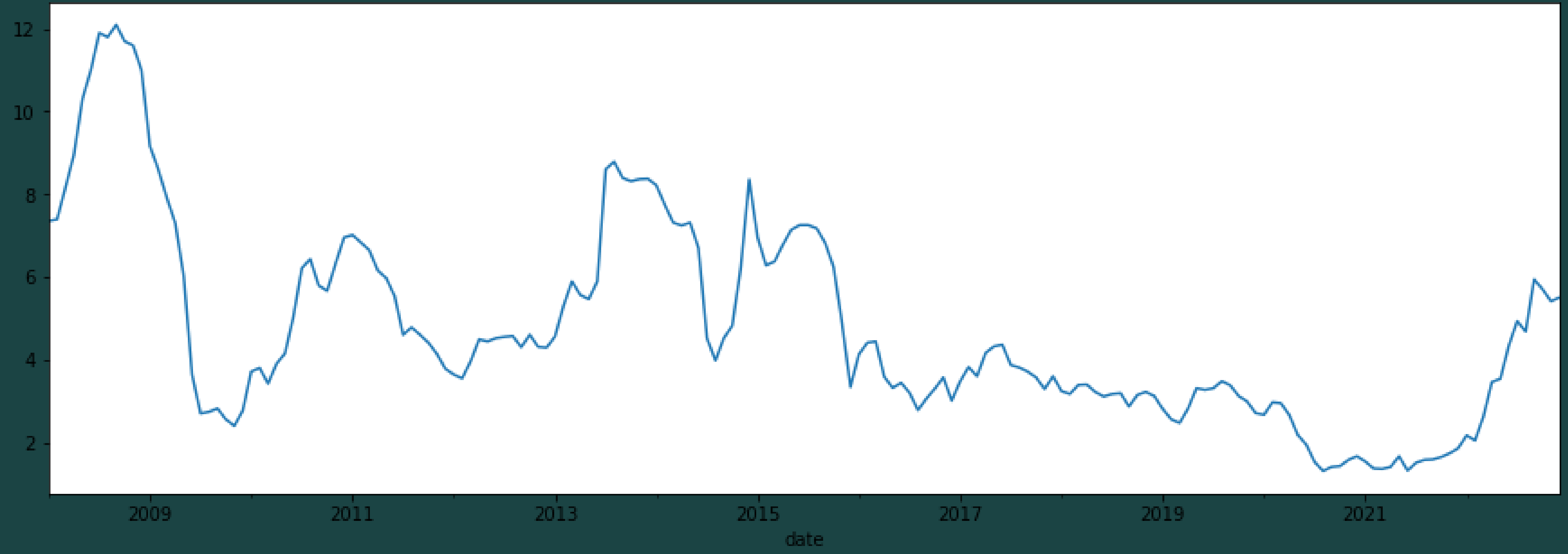
- Indonesian monthly inflation data from 01 August 2008 to 01 December 2022
- Source data : www.bi.go.id
- The dataset has 2 columns and 181 rows
- Data dictionary :
  - date : date, month and year of inflation
  - data : inflation rate

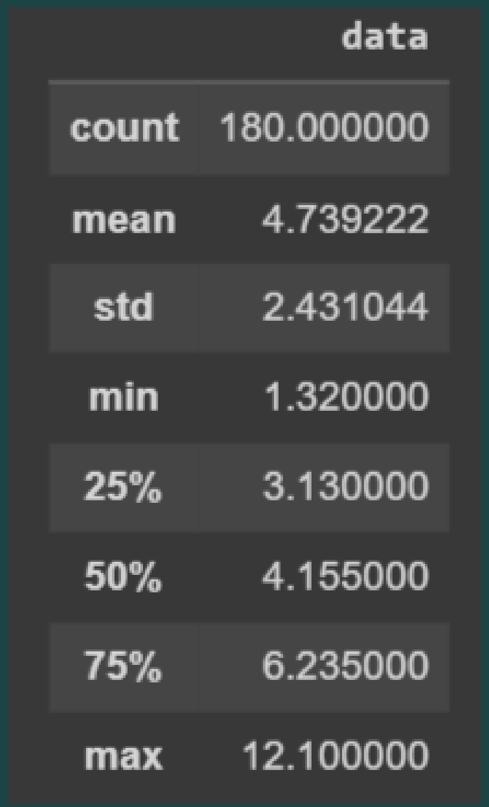
DATA PREPARATION

Code used :

- Python version : 3.7.15
- Packages : pandas, numpy, math, matplotlib, pmdarima, statmodels.api

EXPLORATORY DATA ANALYSIS





	data
count	180.000000
mean	4.739222
std	2.431044
min	1.320000
25%	3.130000
50%	4.155000
75%	6.235000
max	12.100000

From the plot above, we can see that inflation rate mean from 2008 to 2021 is 4.74%. The highest inflation occurred in 2008 worth 12.1% then the lowest inflation occurred in 2021 worth 1.32% and after that there is an uptrend until now

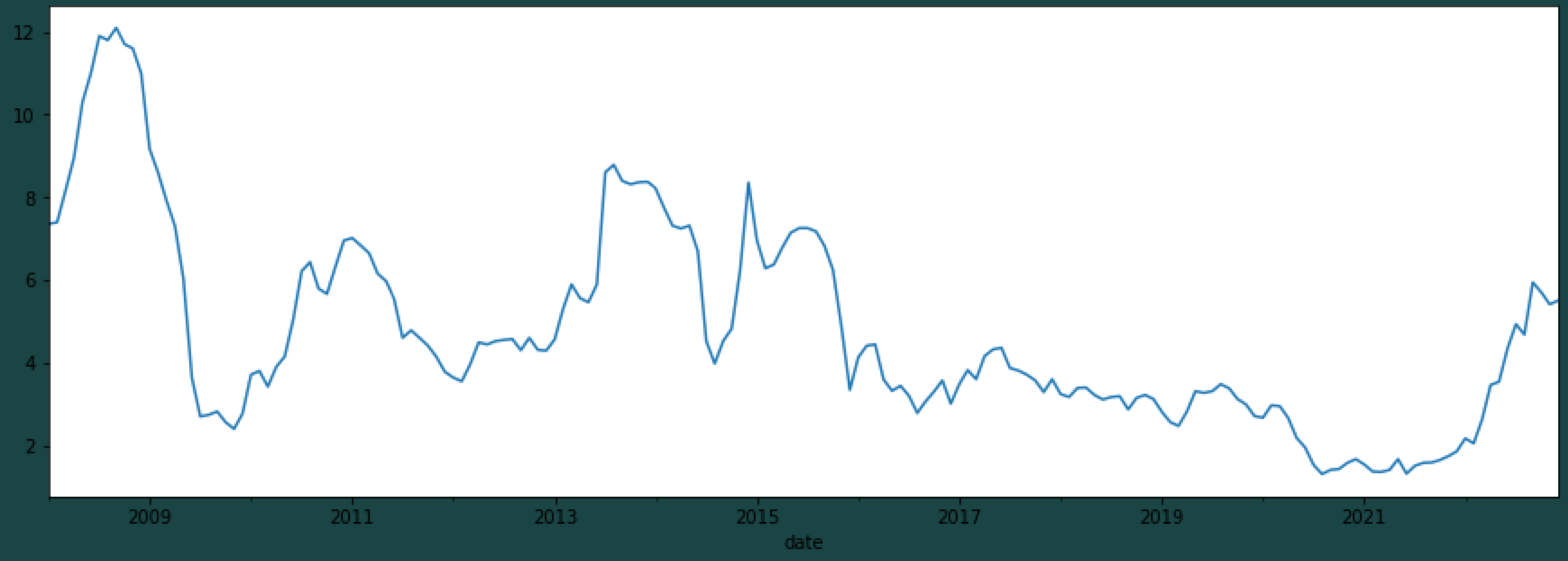
MODELING

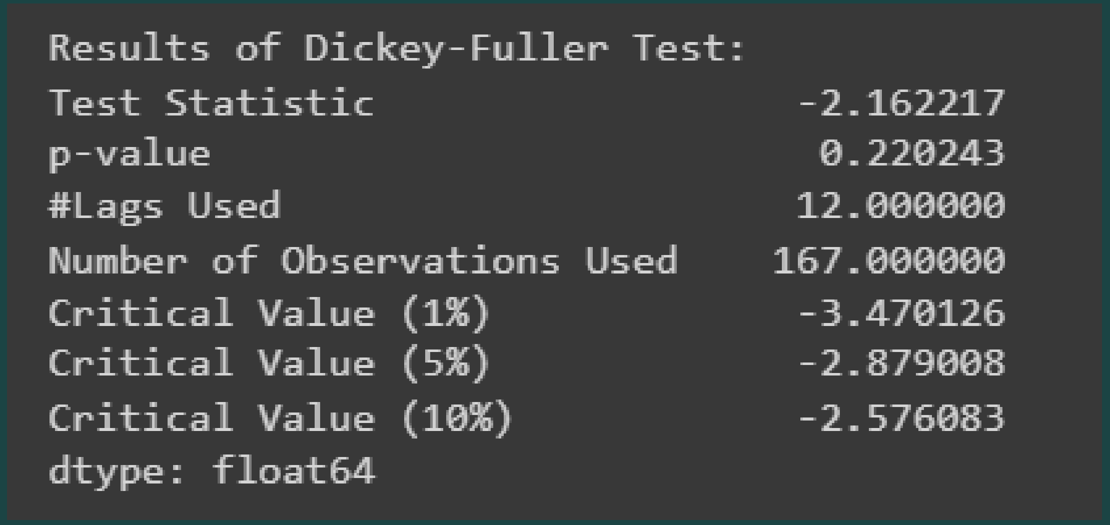
What is ARIMA ?

AutoRegressive Integrated Moving Average (ARIMA) is a time series forecasting model that incorporates autocorrelation measures to model temporal structures within the time series data to predict future values. The autoregression part of the model measures the dependency of a particular sample with a few past observations

MODELING

Model Identification





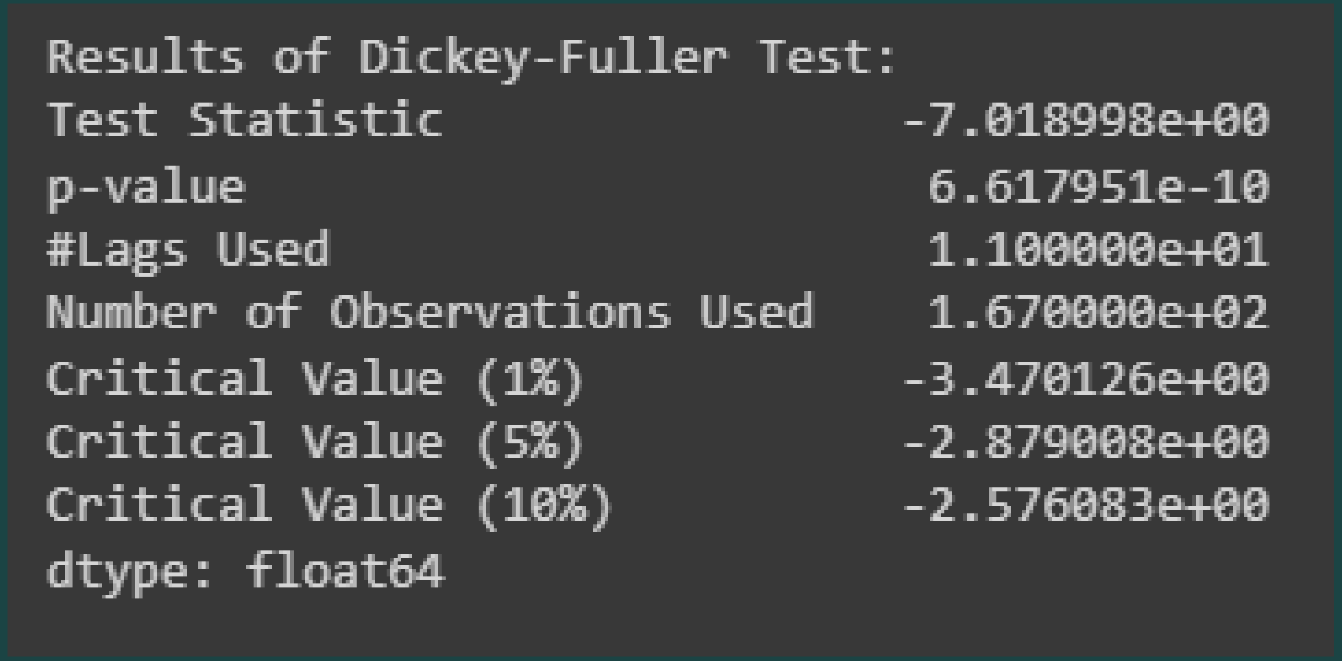
```
Results of Dickey-Fuller Test:
Test Statistic                  -2.162217
p-value                          0.220243
#Lags Used                      12.000000
Number of Observations Used    167.000000
Critical Value (1%)             -3.470126
Critical Value (5%)             -2.879008
Critical Value (10%)            -2.576083
dtype: float64
```

Based on the data pattern, the data has not formed a stationary pattern, also we can see that p-value 0.22 is bigger than critical values 5%, so we can conclude that our data is **not stationary**.

MODELING

Differencing

Differencing at its simplest, involves taking the difference of two adjacent data points. The purpose of differencing is to make the time series stationary but we should be careful to not over-difference the series. An over differenced series may still be stationary, which in turn will affect the model parameters.



```
Results of Dickey-Fuller Test:
Test Statistic                 -7.018998e+00
p-value                         6.617951e-10
#Lags Used                      1.100000e+01
Number of Observations Used     1.670000e+02
Critical Value (1%)            -3.470126e+00
Critical Value (5%)            -2.879008e+00
Critical Value (10%)           -2.576083e+00
dtype: float64
```

From adf test results, we can see that the time series reaches stationarity after one orders of differencing (p value < 5%)

MODELING

ACF and PACF

Autocorrelation Function (ACF) is a measure of the correlation between the the time series (ts) with a lagged version of itself. ACF is used to determine the q parameter in the arima model

Partial Autocorrelation Function (PACF) is measures the correlation between the ts with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons. PACF is used to determine the p parameter in the arima model

MODELING

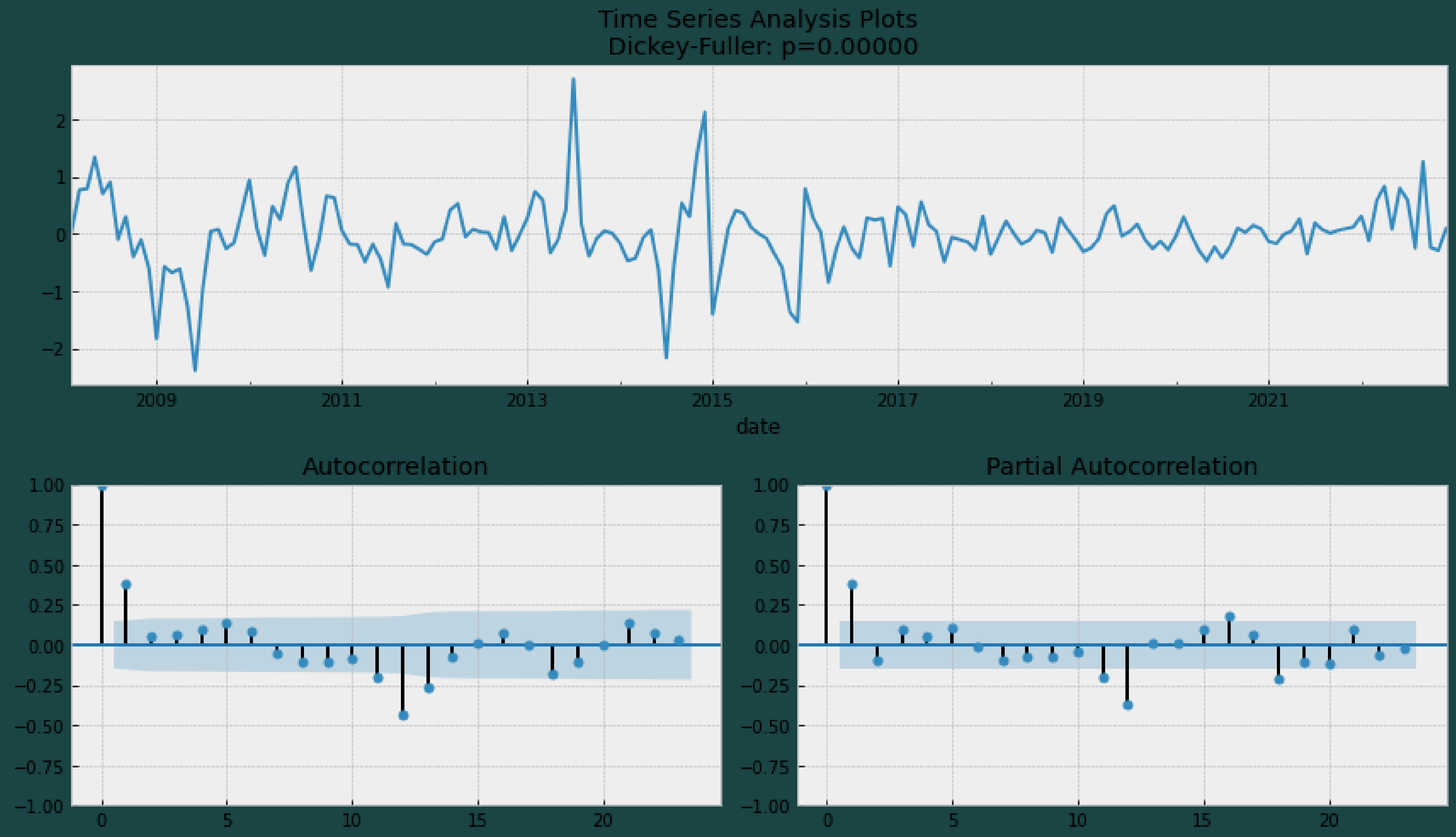
ACF and PACF

ARIMA Models are specified by three order parameters (p, d, q), where :

- p : Lag value where the Partial Autocorrelation (PACF) graph cuts off or drops to 0 for the 1st instance. (order of the AR term)
- d : Number of times differencing is carried out to make the time series stationary.
- q : Lag value where the Autocorrelation (ACF) graph crosses the upper confidence interval for the 1st instance. (order of the MA term)

MODELING

Parameter Estimation



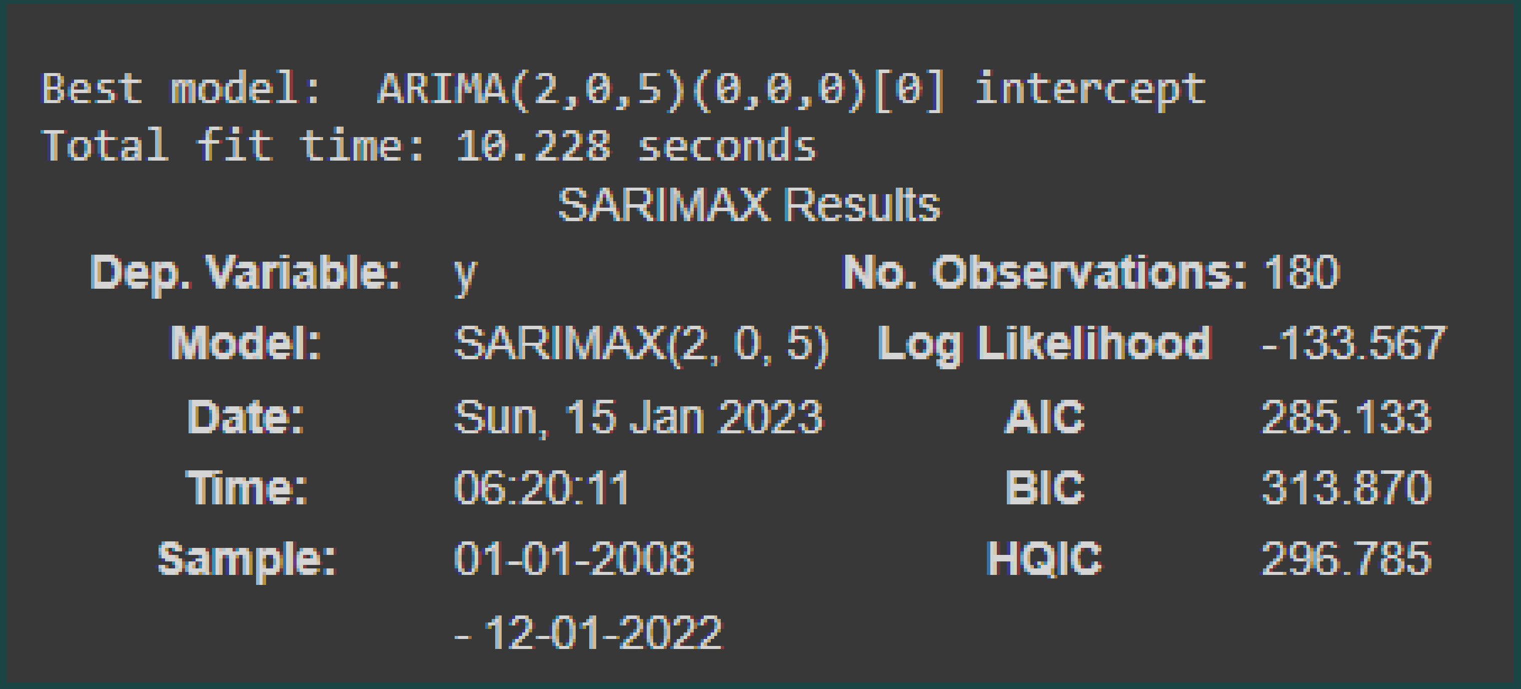
From the plots, the following ARIMA model sequences are selected from the previous criteria: :

- p = 2
- d = 1
- q = 3

So the estimated parameters that are formed is ARIMA (2, 1, 3)

MODELING

Parameter Estimation

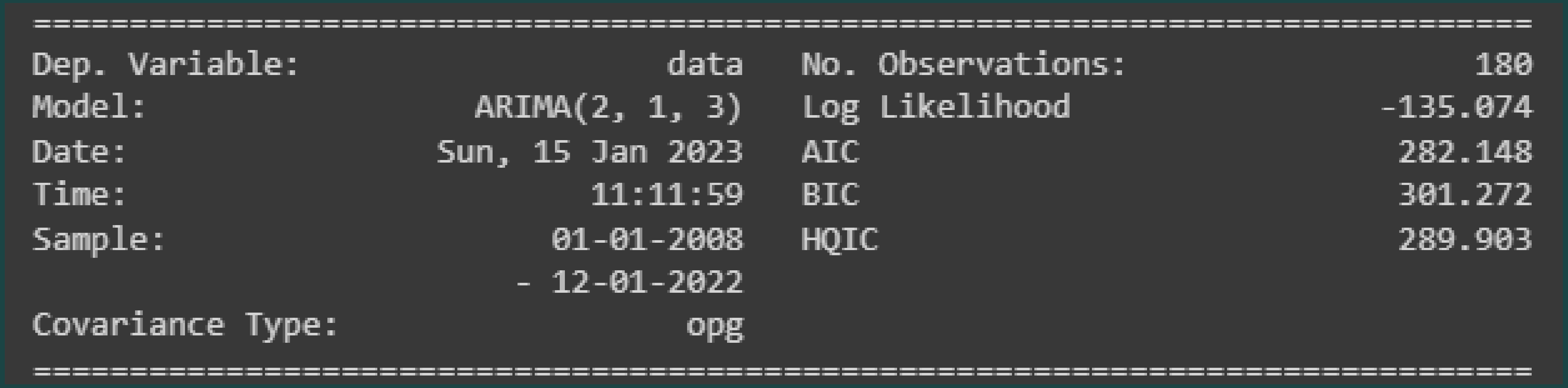


```
Best model:  ARIMA(2,0,5)(0,0,0)[0] intercept
Total fit time: 10.228 seconds
```

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	180
Model:	SARIMAX(2, 0, 5)	**Log Likelihood**	-133.567
Date:	Sun, 15 Jan 2023	**AIC**	285.133
Time:	06:20:11	**BIC**	313.870
Sample:	01-01-2008 - 12-01-2022	**HQIC**	296.785

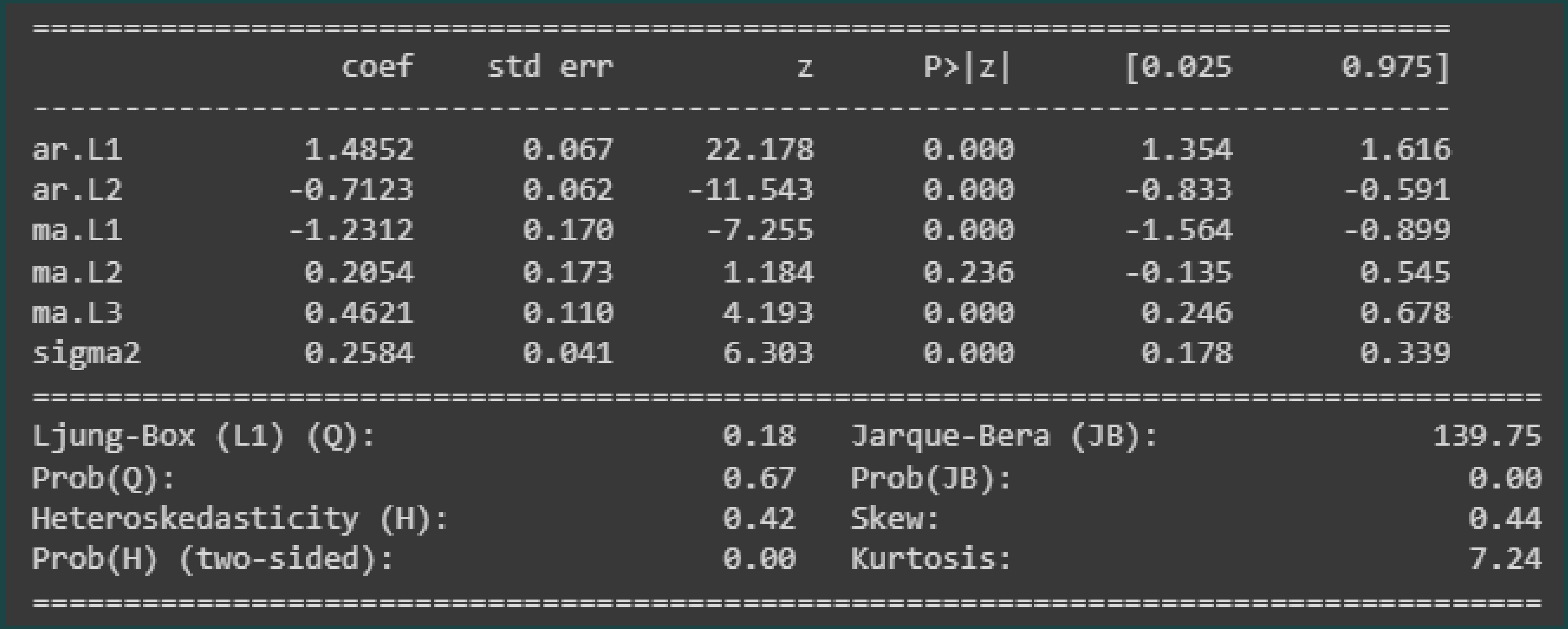
pmdarima packages also provide an arima estimator that can determine the best model for our data, but if we compare the AIC values, the model that we manually specify has a smaller AIC value so we will use the previous model, ARIMA (2,1,3)



```
==============================================================================
Dep. Variable:                   data   No. Observations:                  180
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -135.074
Date:                Sun, 15 Jan 2023   AIC                            282.148
Time:                        11:11:59   BIC                            301.272
Sample:                    01-01-2008   HQIC                           289.903
                         - 12-01-2022
Covariance Type:                  opg
==============================================================================
```

MODELING

Diagnostic Test



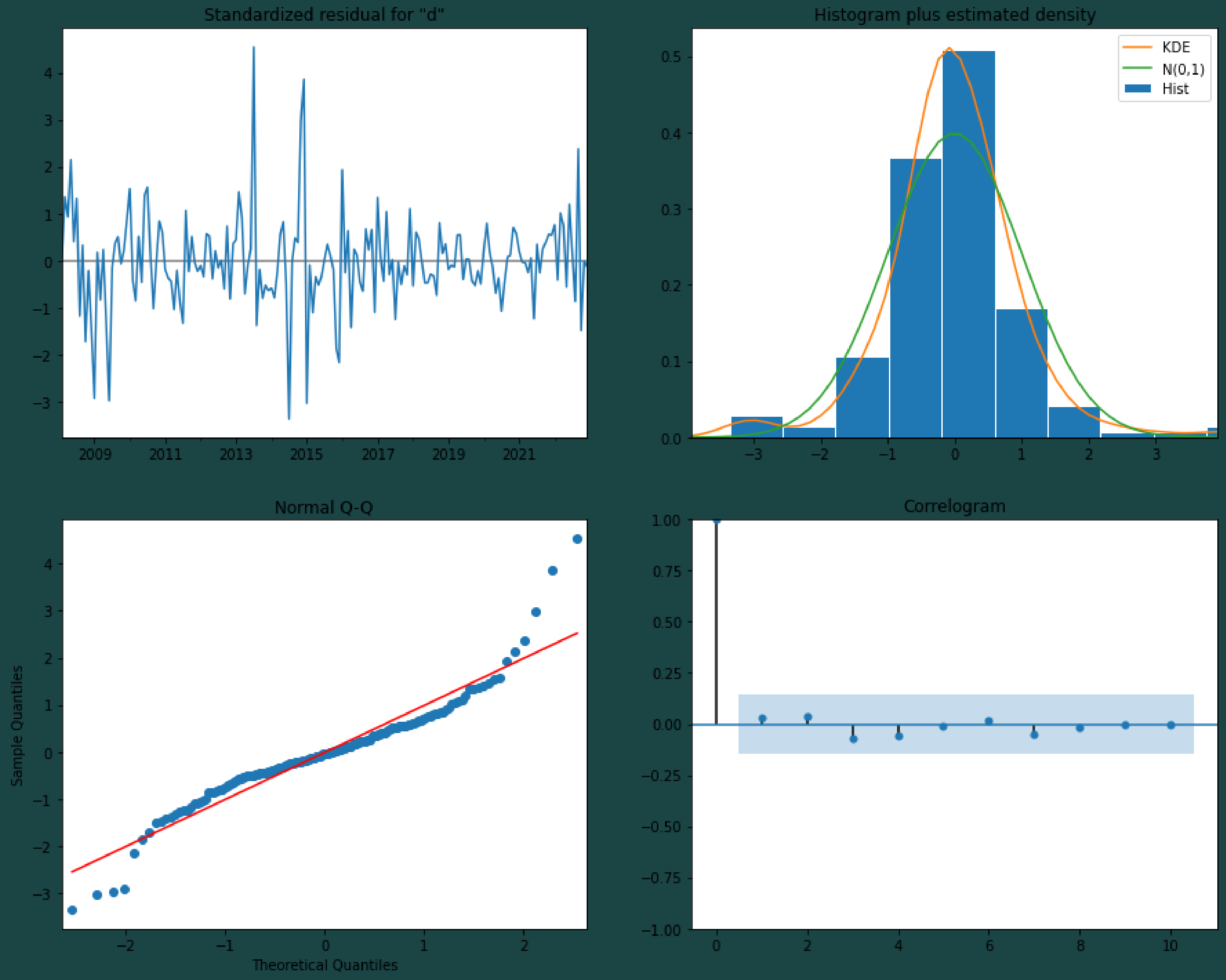
```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.4852      0.067     22.178      0.000       1.354       1.616
ar.L2         -0.7123      0.062    -11.543      0.000      -0.833      -0.591
ma.L1         -1.2312      0.170     -7.255      0.000      -1.564      -0.899
ma.L2          0.2054      0.173      1.184      0.236      -0.135       0.545
ma.L3          0.4621      0.110      4.193      0.000       0.246       0.678
sigma2         0.2584      0.041      6.303      0.000       0.178       0.339
===================================================================================
Ljung-Box (L1) (Q):                   0.18   Jarque-Bera (JB):               139.75
Prob(Q):                              0.67   Prob(JB):                         0.00
Heteroskedasticity (H):               0.42   Skew:                             0.44
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.24
===================================================================================
```

The P>|z| column informs us of the significance of each feature weight. Here, each weight has a p-value close to 0, so we can conclude that the parameters are already significant.

We can also see that the ljung box value is greater than the critical value (0.18 > 0.05) so that the model meets the residual white noise assumption

MODELING

Diagnostic Test



The model diagnostic suggests that the model residual is normally distributed based on the following:

- In the top right plot, the red KDE line follows closely with the N(0,1) line. Where, N(0,1) is the standard notation for a normal distribution .This is a good indication that the residuals are normally distributed

MODELING

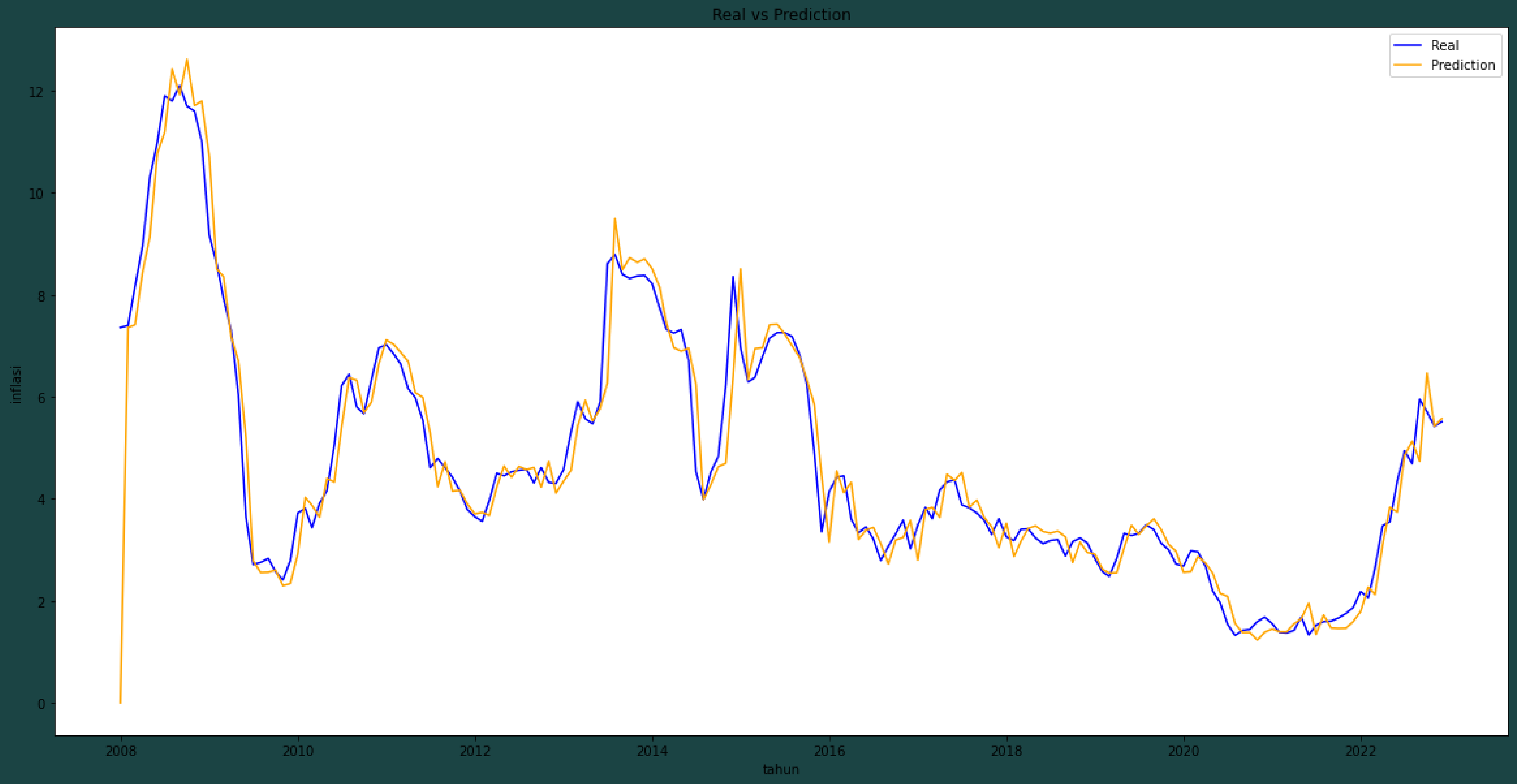
Diagnostic Test

- The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution. Again, this is a strong indication that the residuals are normally distributed.
- The residuals over time (top left plot) don't display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

Those observations lead us to conclude that our model produces a satisfactory fit that could help us understand our time series data and forecast future values

FORECASTING

In - Sample Forecasting

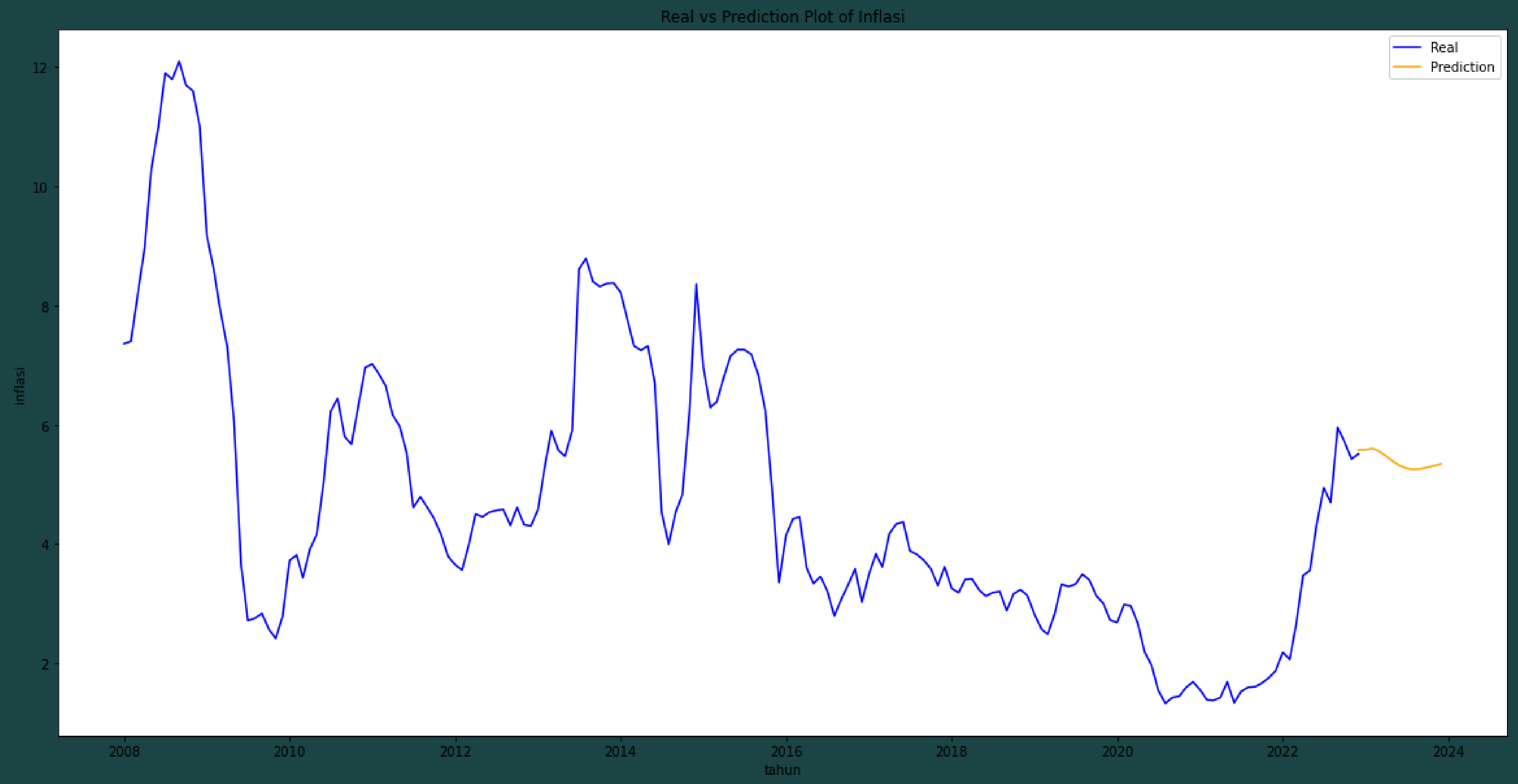


In Sample Forecasting is model forecasts values for the existing data points of the time series.

This plot are the results of in sample forecasting with a MAPE value of 8.75% and an RMSE of 0.75 which means the forecasting model has very good accuracy

FORECASTING

Out - Sample Forecasting

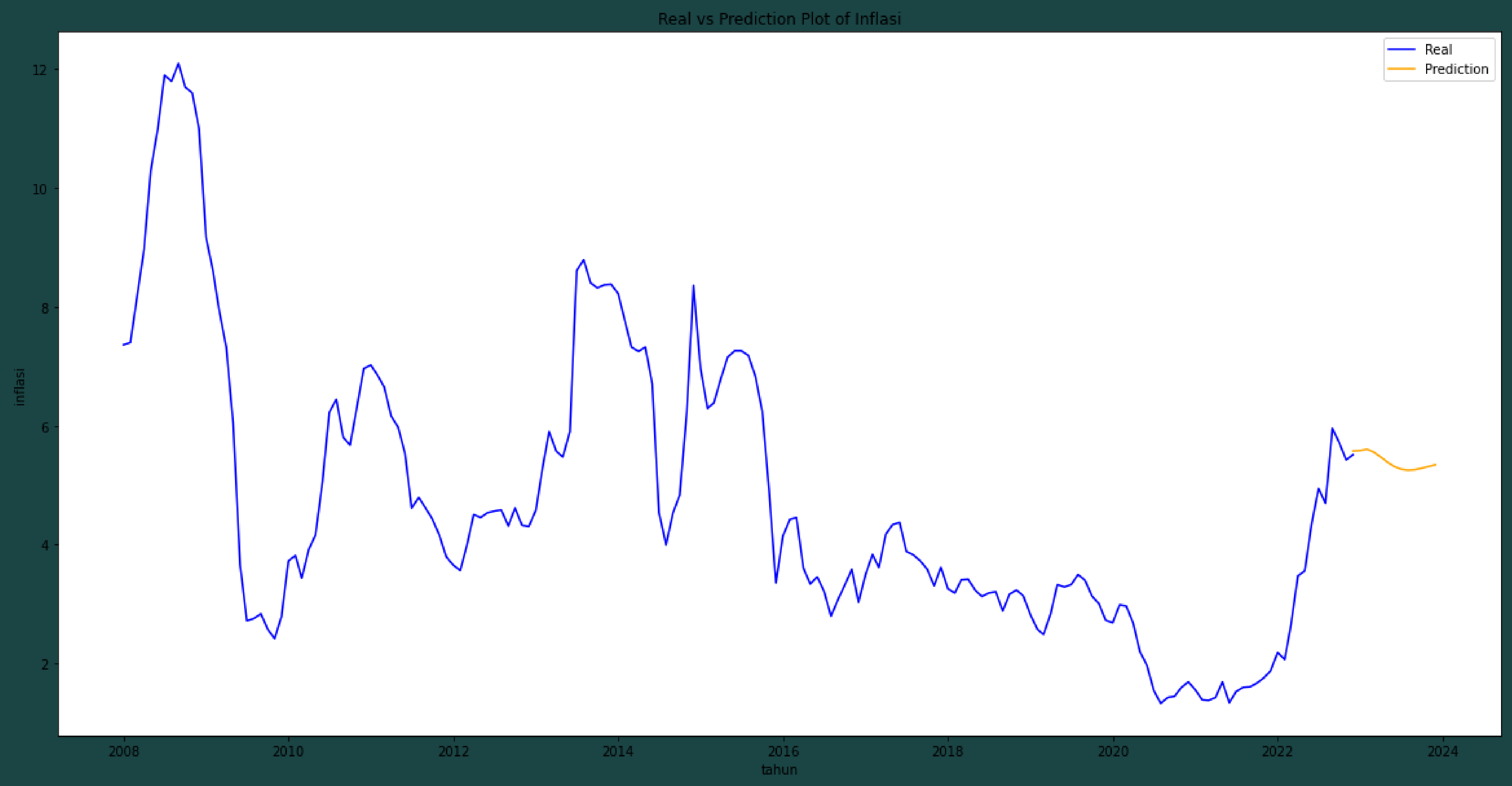


Out of Sample Forecasting is model forecasts values for the future data points.

From the results of out of sample forecasting, we can see that the inflation rate in the coming year will fluctuate with mean value is 5.39%.

FORECASTING

Out - Sample Forecasting



Out of Sample Forecasting is model forecasts values for the future data points.

From the results of out of sample forecasting, we can see that the inflation rate in the coming year will fluctuate with mean value is 5.39%.

THANK YOU

Muhammad Iqbal Rustan

github.com/muhiqbalrustan

iqbal.jr47@gmail.com