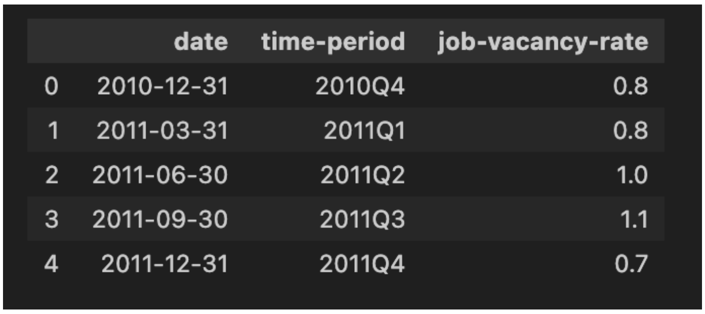
HWII - Workshop II

University of Luxembourg

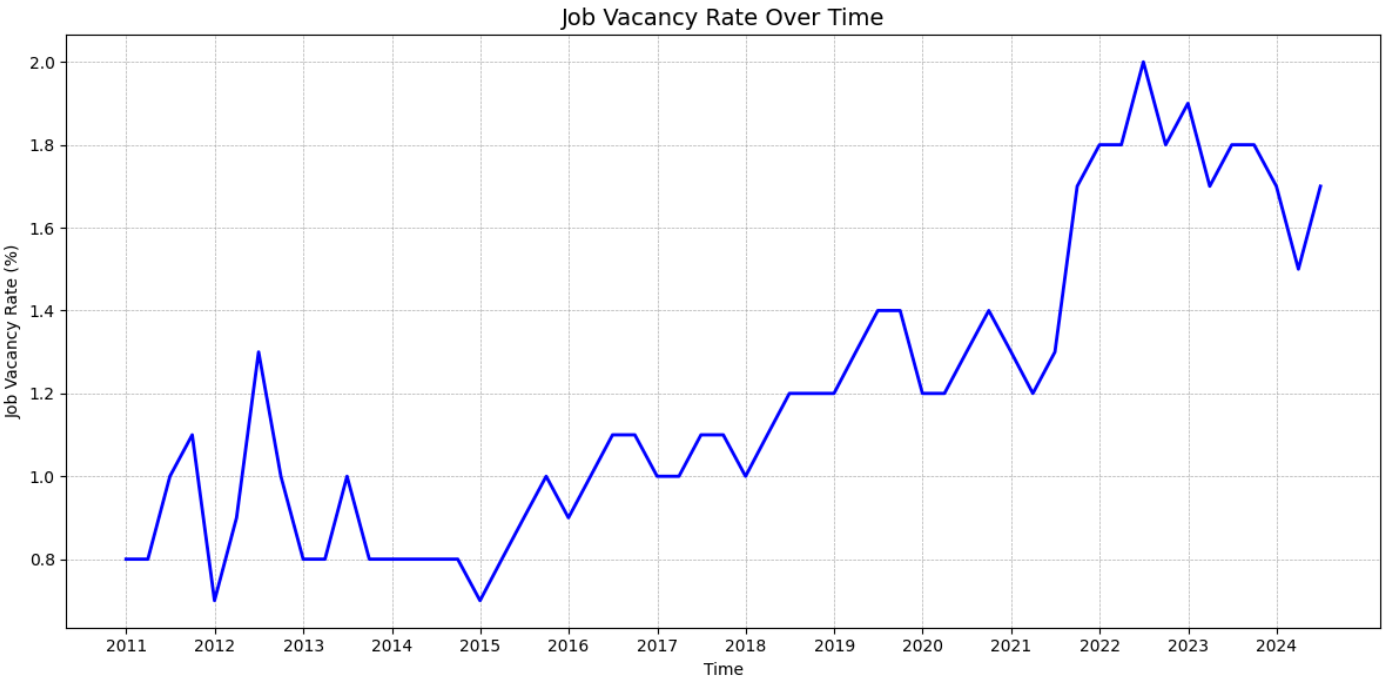
Zahra Hashemi 0231846534

Dataset

The dataset is derived from the European Central Bank Datacenter. It consists of data from winter 2010 until summer 2024, reporting “job vacancy rate” in the Euro area. The dataset reports the job vacancy rate seasonally.



	date	time-period	job-vacancy-rate
0	2010-12-31	2010Q4	0.8
1	2011-03-31	2011Q1	0.8
2	2011-06-30	2011Q2	1.0
3	2011-09-30	2011Q3	1.1
4	2011-12-31	2011Q4	0.7



POINT 1 [STATIONARITY & TRANSFROMATIONS]:

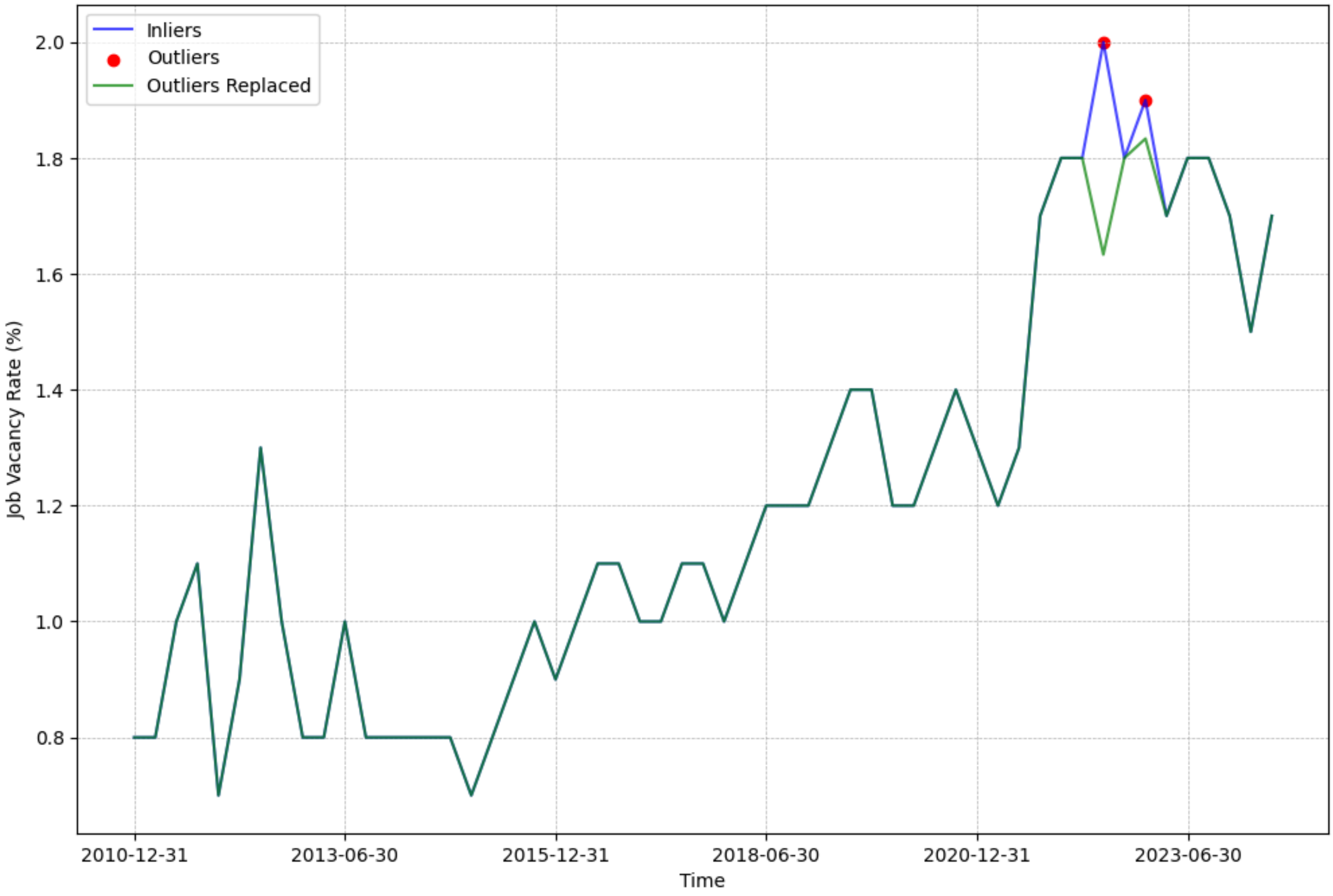
For data cleaning, no missing values were found.
However for outliers, two methods were used to detect outliers:

Outliers removing using IRQ

With this method, no outliers were found. So we went for another approach for detecting outliers.

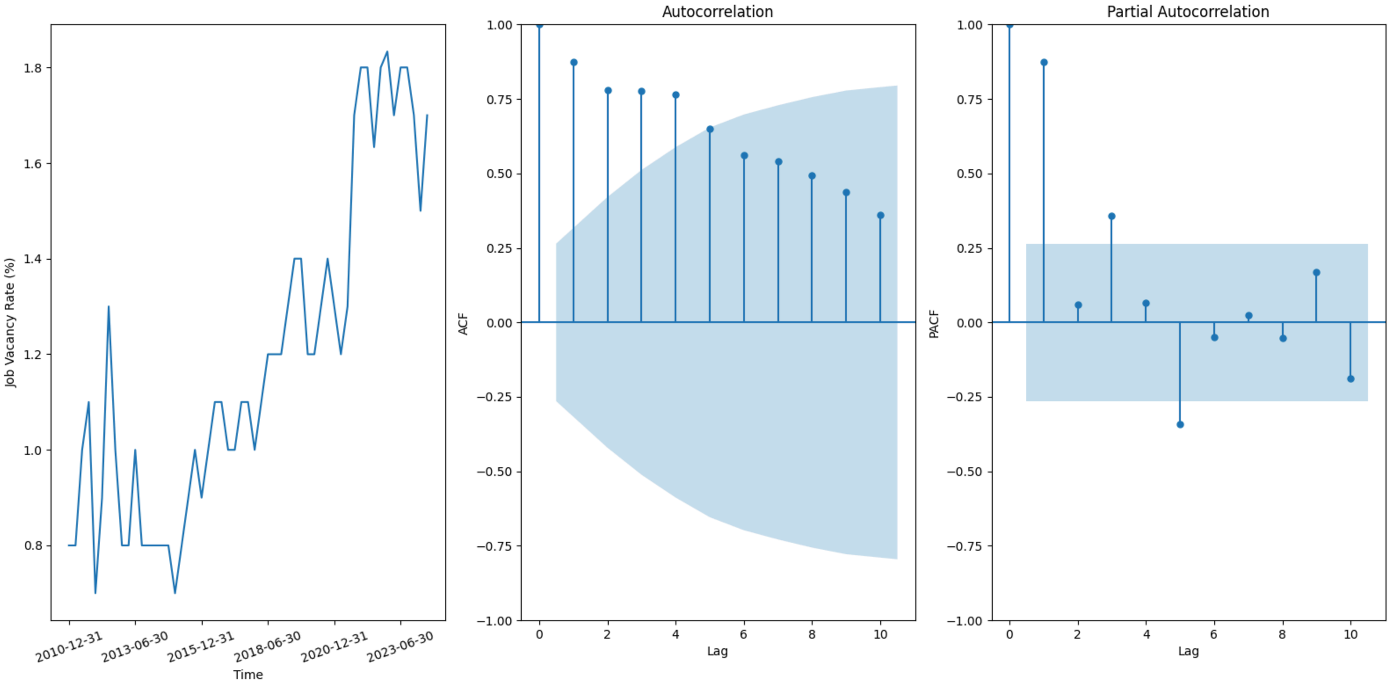
Outliers removing using Isolation forest

Isolation Forest is an algorithm for data anomaly detection using binary trees.
Using this method, 2 anomalies were detected and replaced by the rolling method with the window size of 6.



Data being stationary or not

After plotting Autocorrelation and Partial Autocorrelation functions, we can start interpreting them:



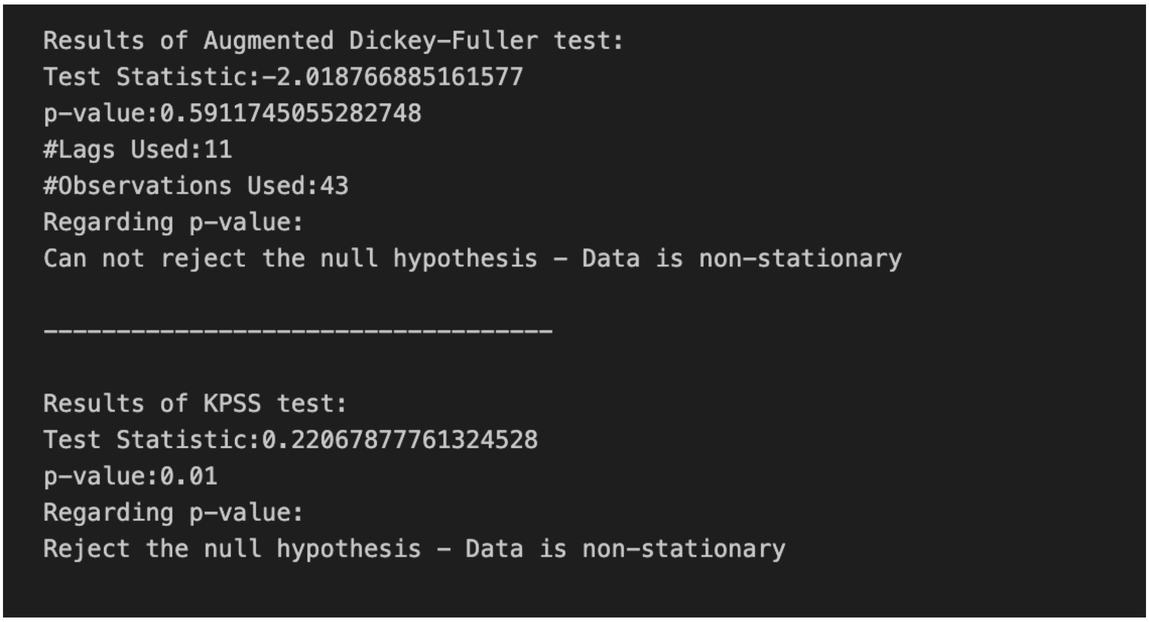
From the job vacancy plot, it is almost obvious that we have a global upward trend in our data. So, it is likely non-stationary.

In ACF, the coefficients gradually decrease as the lag increases, but they remain significant up to a certain lag, indicating some level of persistence or trend in the series.

For PACF, we can see that it doesn't drop to zero quickly, even, it bounces (gets close and far to zero as the lag increases). So it suggests non-stationary data.

Statistical tests

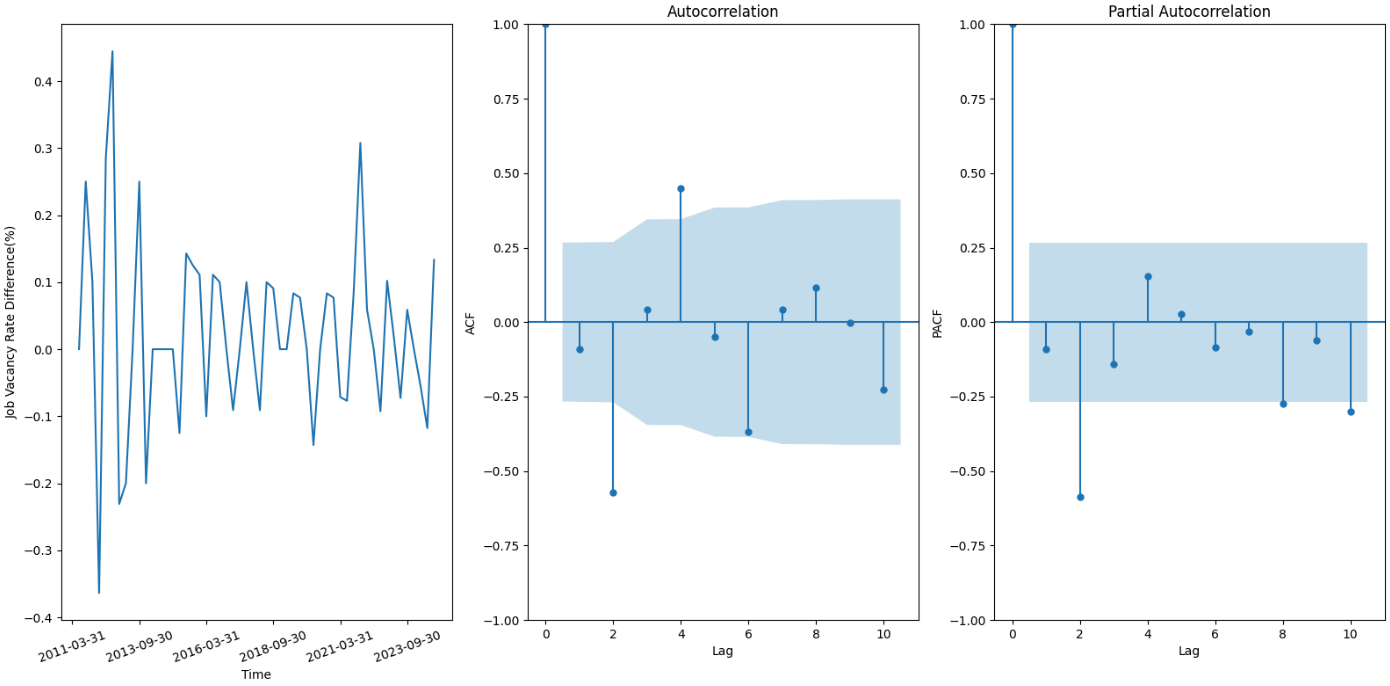
After implementing Augmented Dickey-Fuller test and KPSS test, regarding the p-values obtained, it suggests that the data is non-stationary.

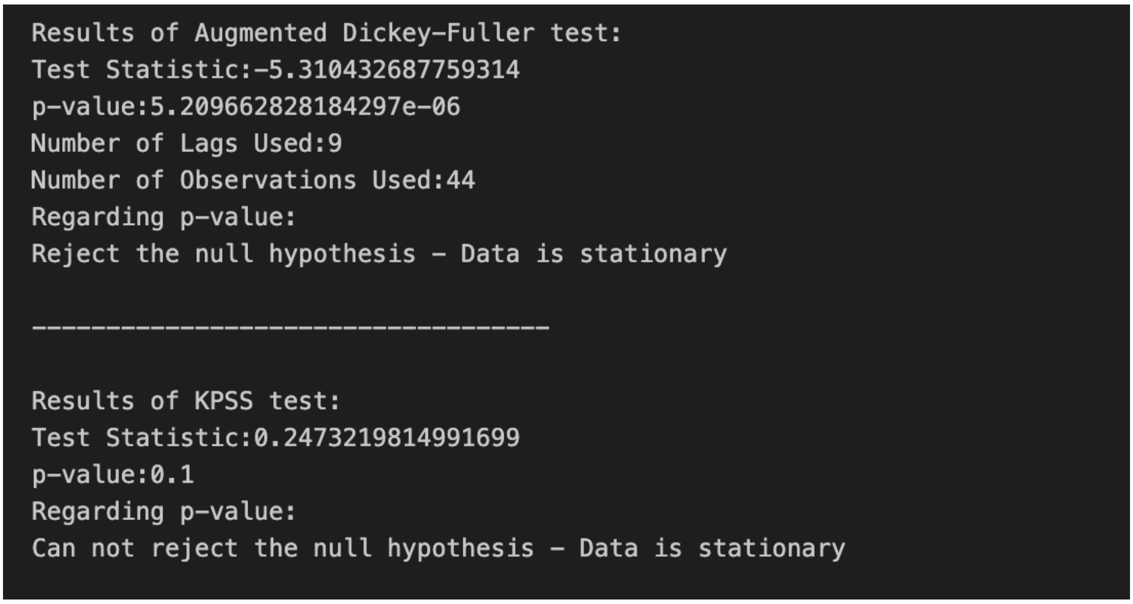


```
Results of Augmented Dickey-Fuller test:
Test Statistic:-2.018766885161577
p-value:0.5911745055282748
#Lags Used:11
#Observations Used:43
Regarding p-value:
Can not reject the null hypothesis - Data is non-stationary

-----------------------------------

Results of KPSS test:
Test Statistic:0.22067877761324528
p-value:0.01
Regarding p-value:
Reject the null hypothesis - Data is non-stationary
```

Transformation

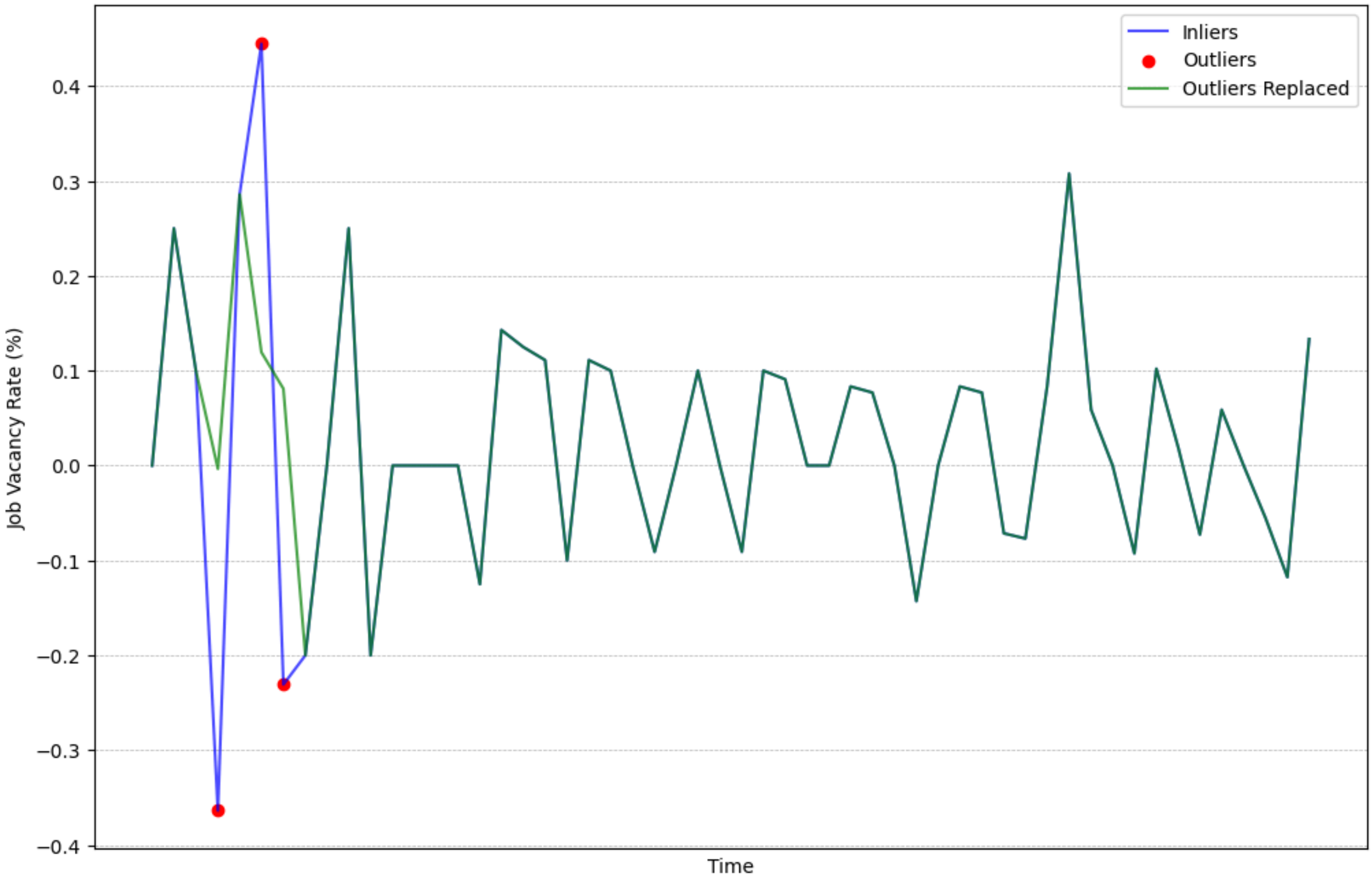
So a transformation is needed to convert the data into stationary data. The first and easy transformation that comes to head is **first-order differencing**, subtracting each data point from the previous one to remove trends.





```
Results of Augmented Dickey-Fuller test:
Test Statistic:-5.310432687759314
p-value:5.209662828184297e-06
Number of Lags Used:9
Number of Observations Used:44
Regarding p-value:
Reject the null hypothesis - Data is stationary

-----------------------------------

Results of KPSS test:
Test Statistic:0.2473219814991699
p-value:0.1
Regarding p-value:
Can not reject the null hypothesis - Data is stationary
```

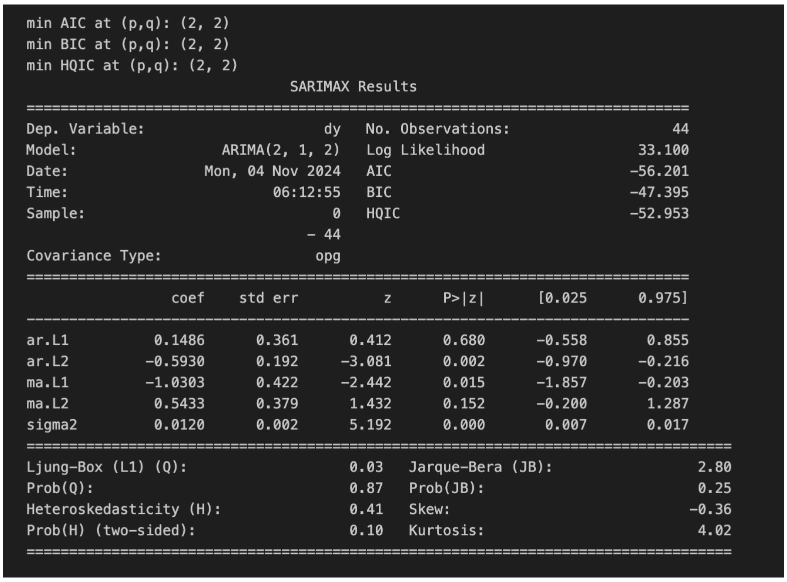
And this is the differenced data after detecting (isolation forest) and smoothing the outliers.



POINT 2 [ARIMA MODEL]:

To choose the best parameters for p, q and d:
(p - AR model lags - number of pas values used to predict the current observation
d - differencing order - number of times the series has been differenced
q - MA model lags - number of past errors used to predict the current observation)

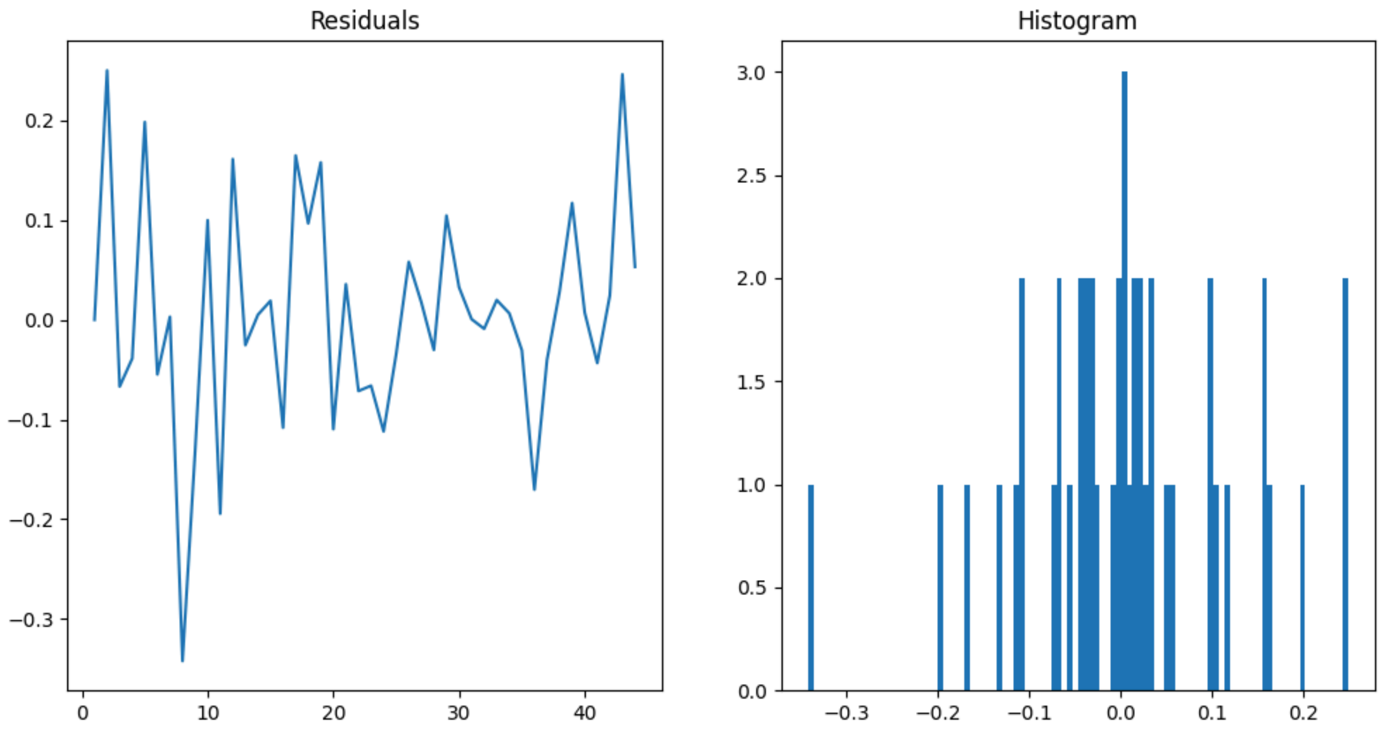
The function **arma_order_select_ic** was used to find the best p and q. d also equals to 1, since the data was differenced one time to make it stationary.

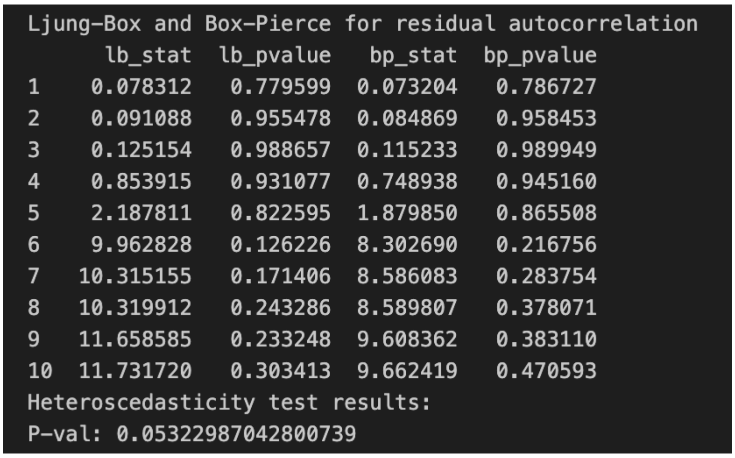


```
min AIC at (p,q): (2, 2)
min BIC at (p,q): (2, 2)
min HQIC at (p,q): (2, 2)
                               SARIMAX Results
==============================================================================
Dep. Variable:                     dy   No. Observations:                   44
Model:                 ARIMA(2, 1, 2)   Log Likelihood                  33.100
Date:                Mon, 04 Nov 2024   AIC                            -56.201
Time:                        06:12:55   BIC                            -47.395
Sample:                             0   HQIC                           -52.953
                                 - 44
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.1486      0.361      0.412      0.680      -0.558       0.855
ar.L2         -0.5930      0.192     -3.081      0.002      -0.970      -0.216
ma.L1         -1.0303      0.422     -2.442      0.015      -1.857      -0.203
ma.L2          0.5433      0.379      1.432      0.152      -0.200       1.287
sigma2         0.0120      0.002      5.192      0.000       0.007       0.017
===================================================================================
Ljung-Box (L1) (Q):                   0.03   Jarque-Bera (JB):                 2.80
Prob(Q):                              0.87   Prob(JB):                         0.25
Heteroskedasticity (H):               0.41   Skew:                            -0.36
Prob(H) (two-sided):                  0.10   Kurtosis:                         4.02
===================================================================================
```

We can see all the suggested values for (p, q) is (2, 2).

The ARIMA model is built with the parameters of “order=(p, d, q) = (2, 1, 2)”





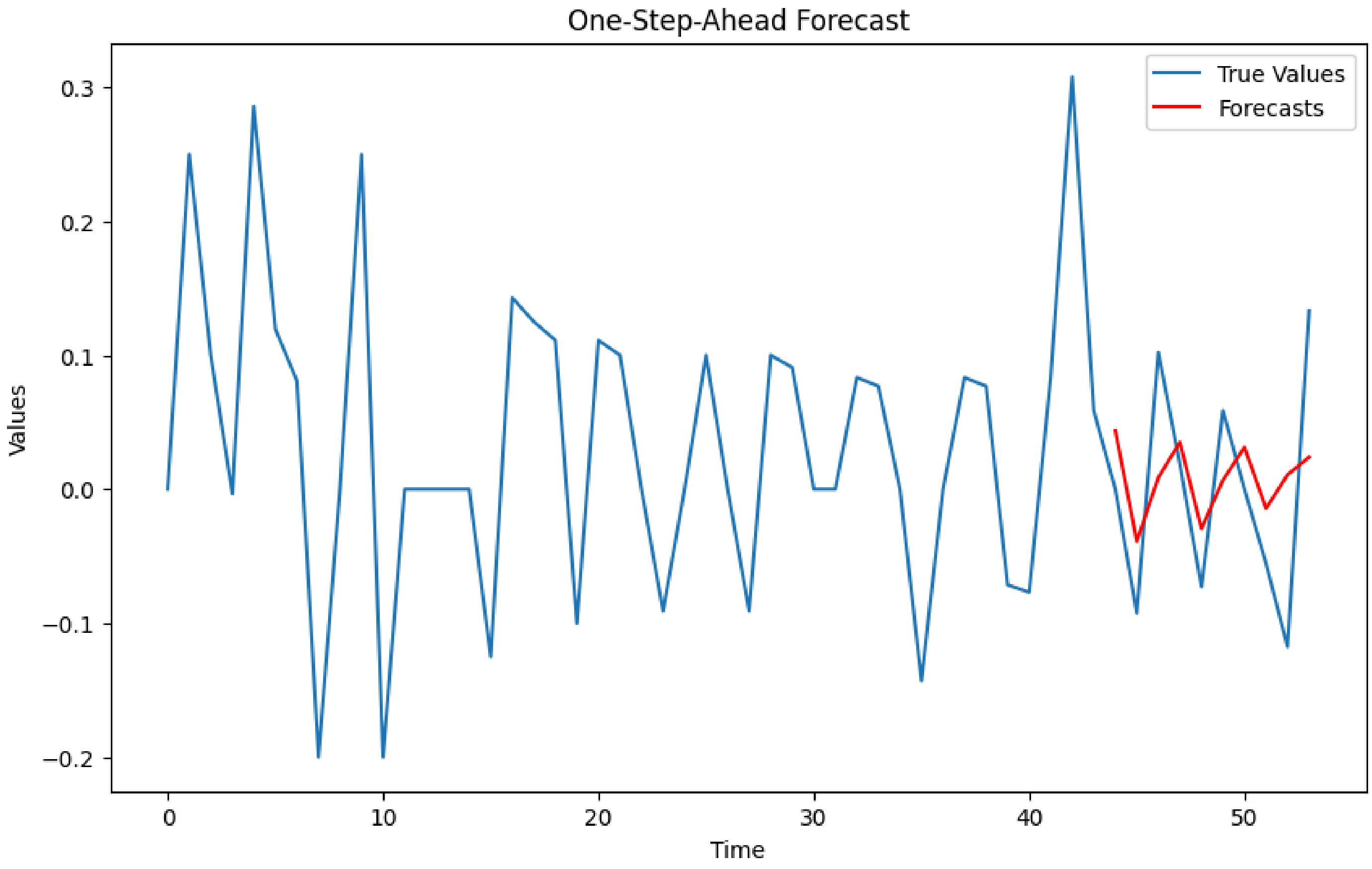
```
Ljung-Box and Box-Pierce for residual autocorrelation
      lb_stat  lb_pvalue   bp_stat  bp_pvalue
1    0.078312   0.779599  0.073204   0.786727
2    0.091088   0.955478  0.084869   0.958453
3    0.125154   0.988657  0.115233   0.989949
4    0.853915   0.931077  0.748938   0.945160
5    2.187811   0.822595  1.879850   0.865508
6    9.962828   0.126226  8.302690   0.216756
7   10.315155   0.171406  8.586083   0.283754
8   10.319912   0.243286  8.589807   0.378071
9   11.658585   0.233248  9.608362   0.383110
10  11.731720   0.303413  9.662419   0.470593
Heteroscedasticity test results:
P-val: 0.05322987042800739
```

After plotting residuals, we can see that the residuals bounce randomly around zero. And its histogram follows almost a normal distribution.

Besides, the p-values of the test are above 0.05, which suggests that there is no significant autocorrelation between residuals, indicating a good model fit.

POINT 3 [FORECASTS]:

After implementing the one-step-ahead forecast regarding the expanding of train data at each iteration, you can find the results below.



RMSE: 0.07036165147492941