Multi variate time series analysis

By Samir Abou Jaoude

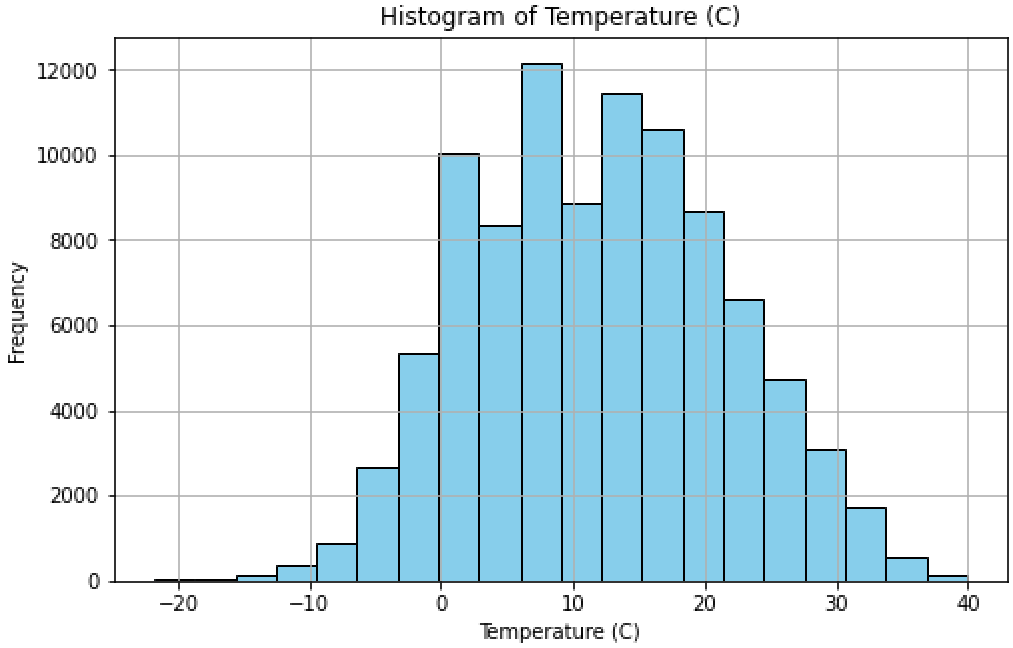
Same data as the univariate, we are still trying to predict the monthly average temperature but this time with other variables.

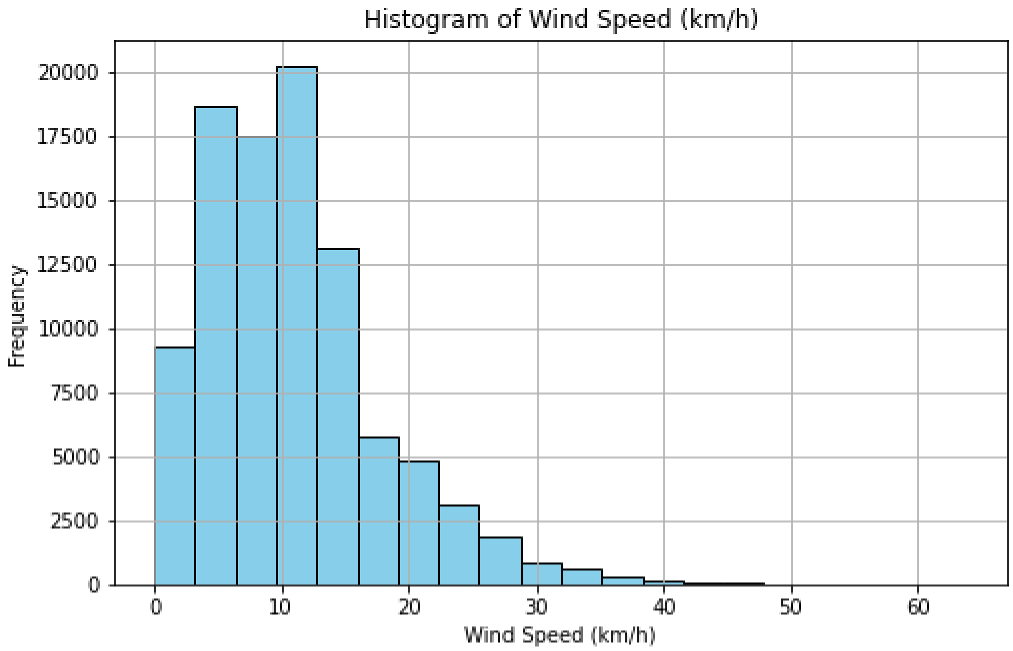
Those are our columns:

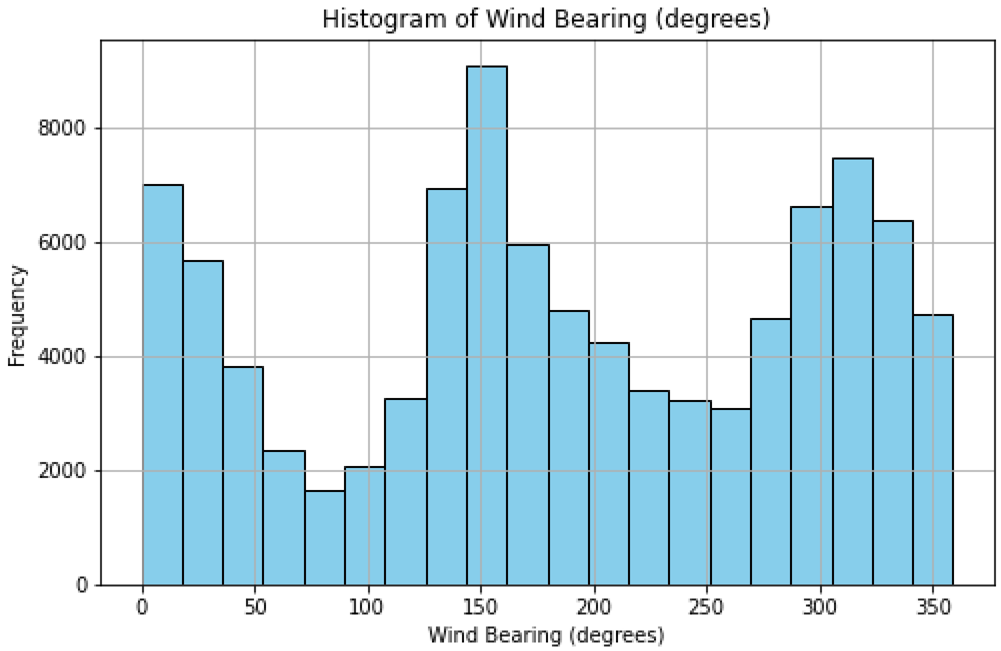
We have temperature (our dependent variable), wind speed, wind bearing and visibility.

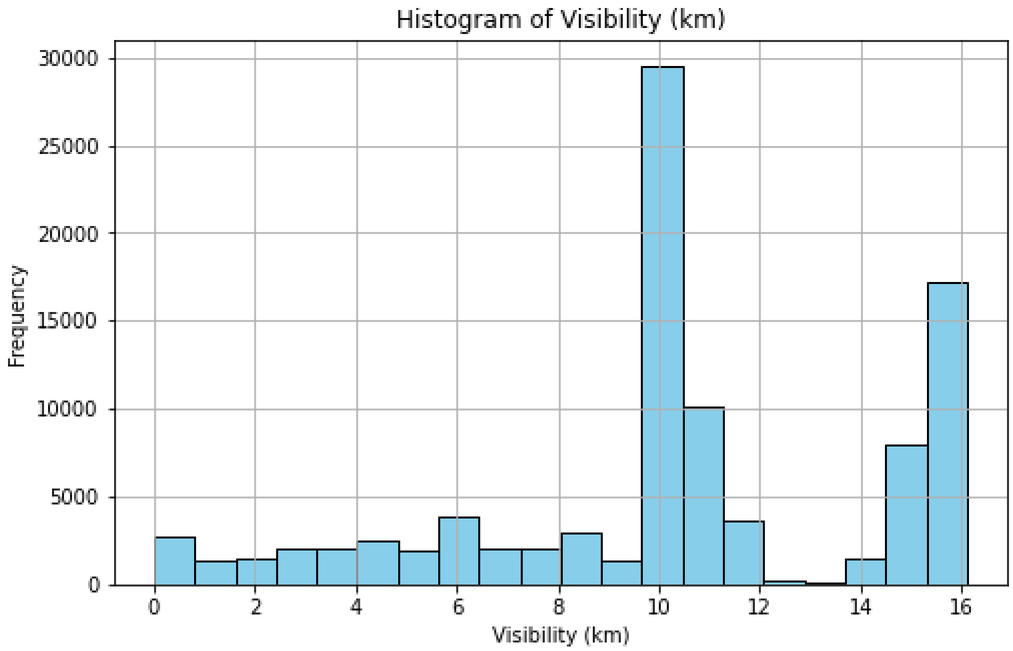
Data Visualization

That’s the distribution of our variables:

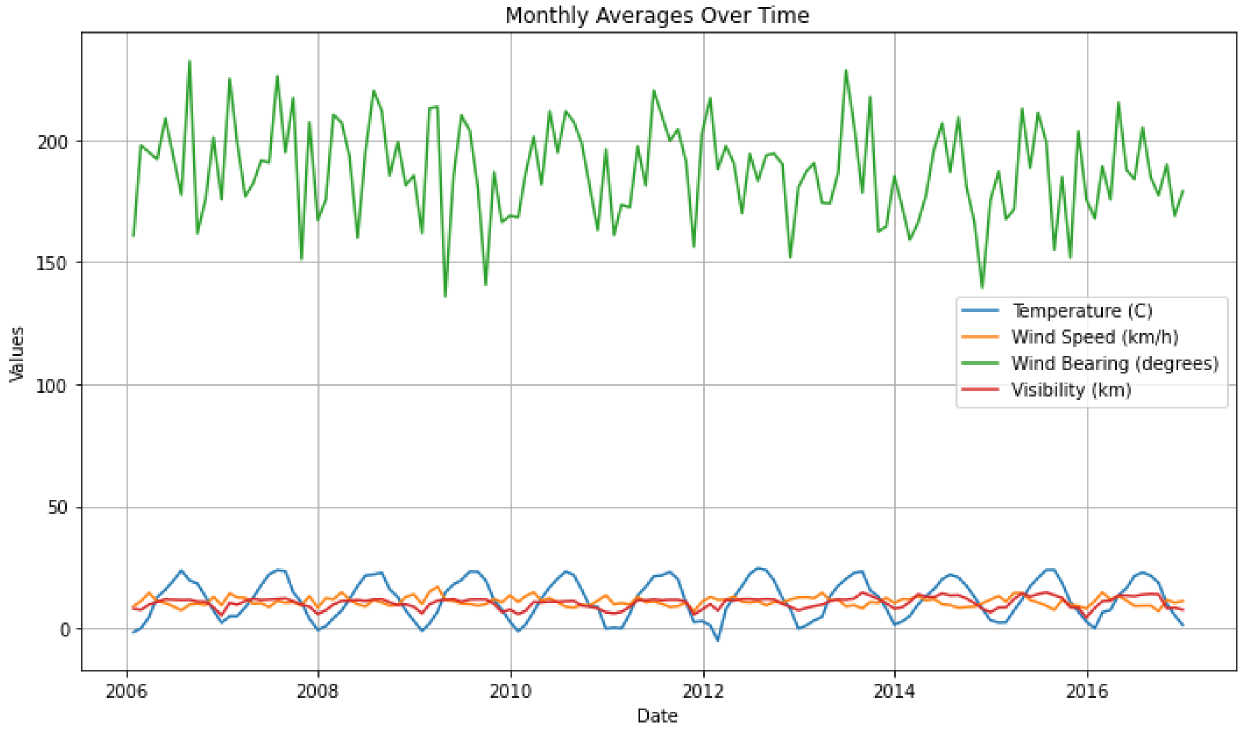



Histogram of Wind Speed (km/h)
Frequency
20000
17500
15000
12500
10000
7500
5000
2500
0
0
10
20
30
40
50
60
Wind Speed (km/h)

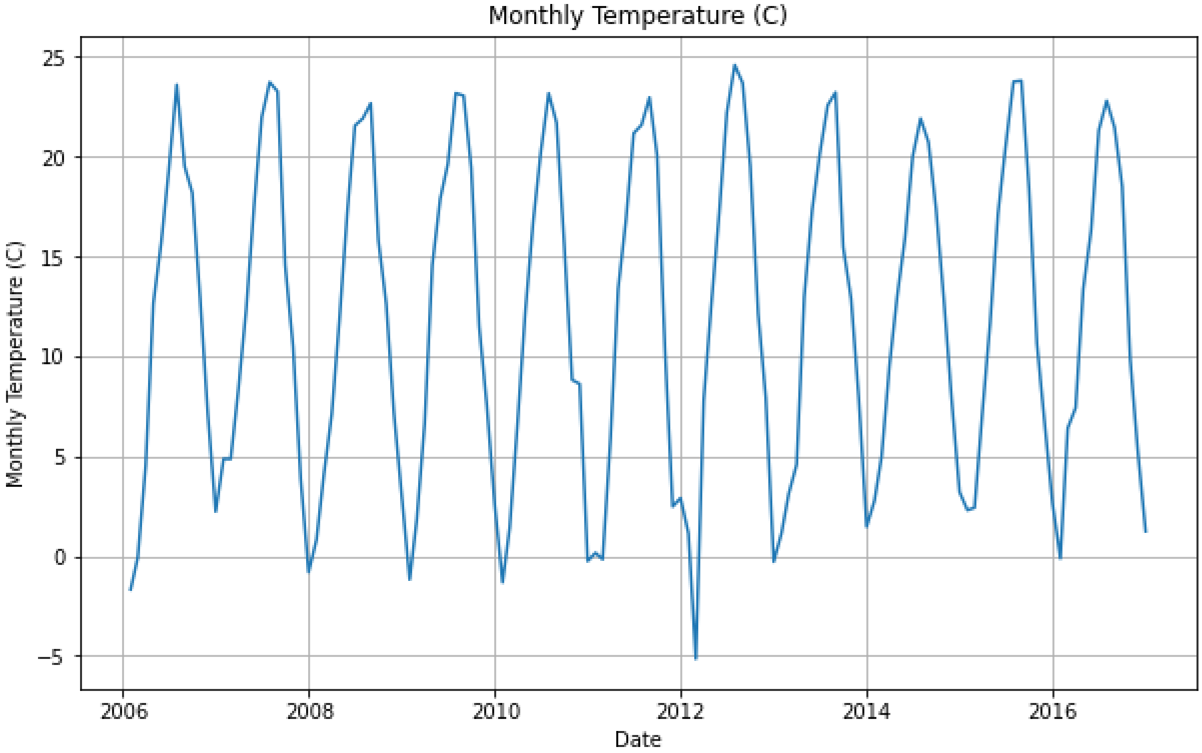

Histogram of Wind Bearing (degrees)
Frequency
8000
6000
4000
2000
0
0
50
100
150
200
250
300
350
Wind Bearing (degrees)

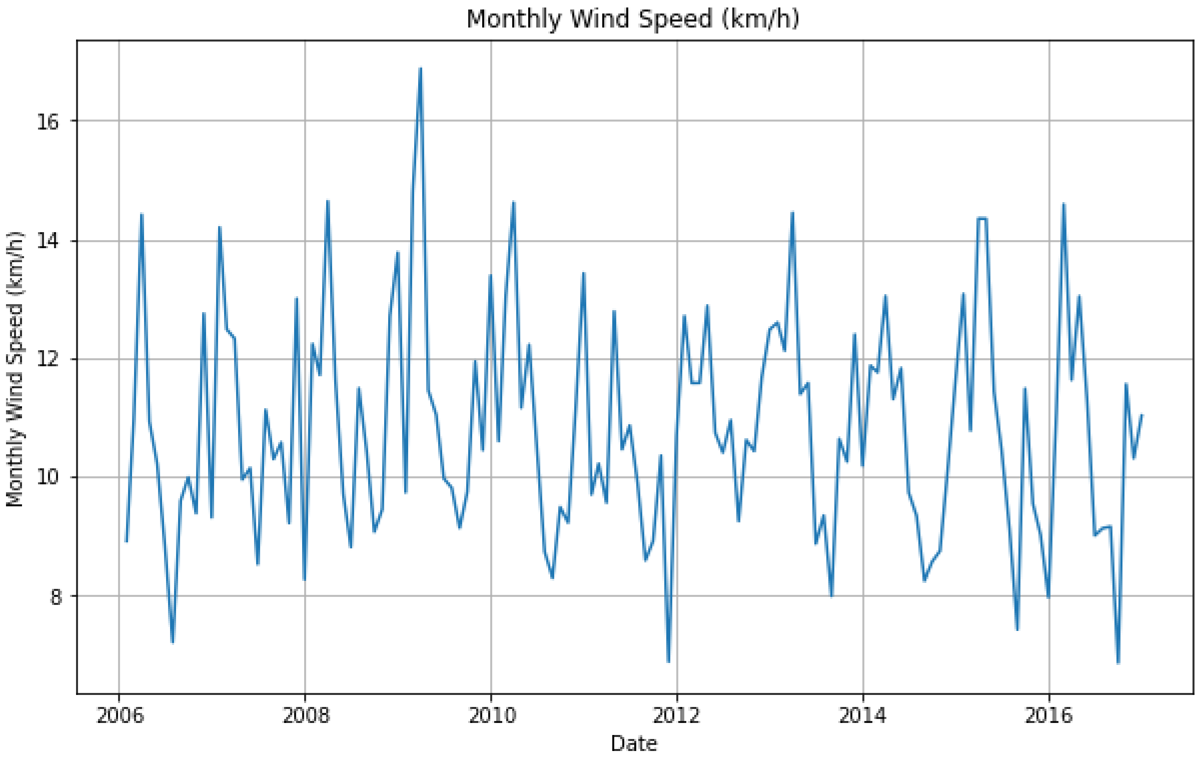


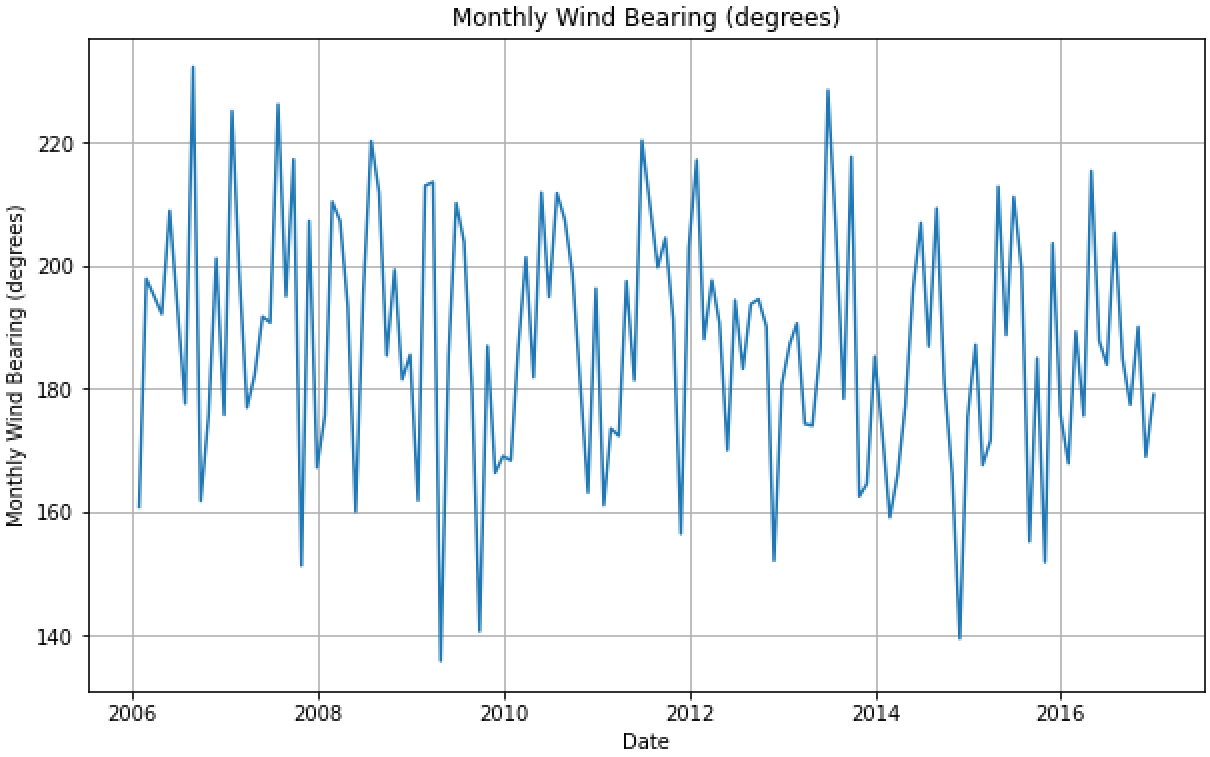
That’s the time series plot for the 5 variables:

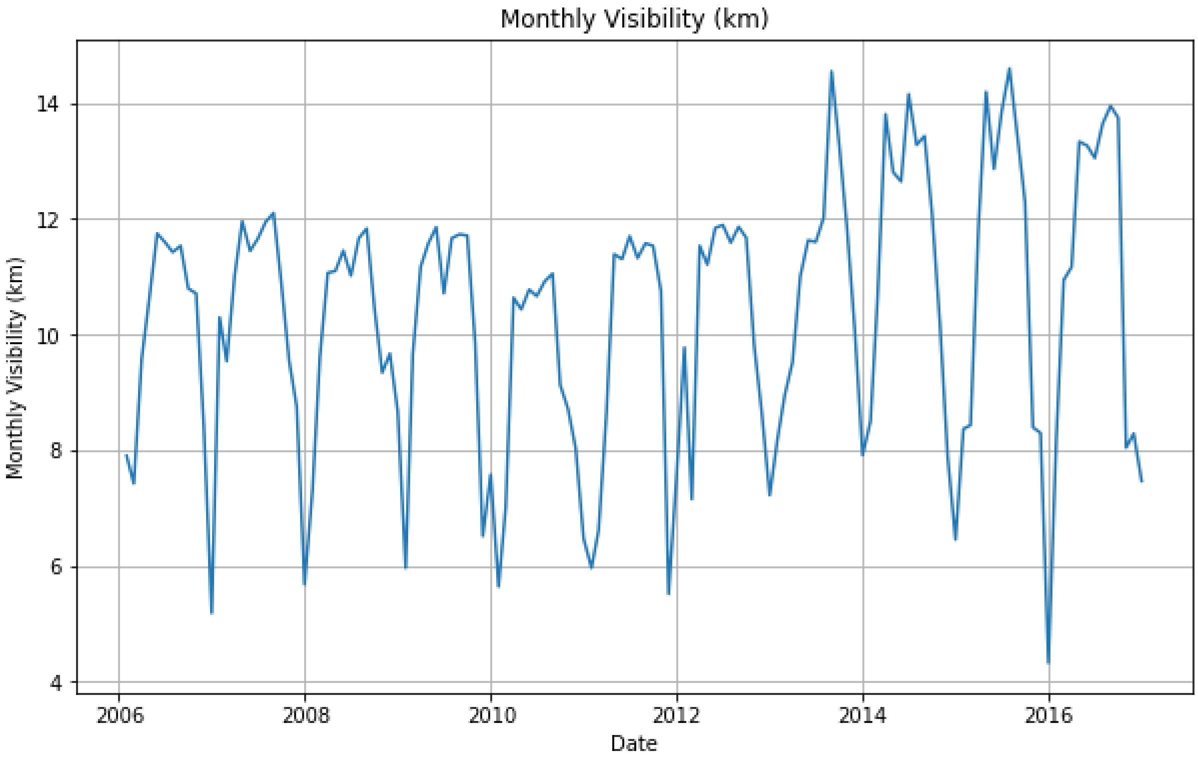


Now we will plot each variable individually.


Monthly Temperature (C)
Monthly Temperature (C)
25
20
15
10
5
0
−5
2006
2008
2010
2012
2014
2016
Date


Monthly Wind Speed (km/h)
Monthly Wind Speed (km/h)
16
14
12
10
8
2006
2008
2010
2012
2014
2016
Date


Monthly Wind Bearing (degrees)
Monthly Wind Bearing (degrees)
220
200
180
160
140
2006
2008
2010
2012
2014
2016
Date



We can see based on these plots that we have yearly seasonality. We can also see that these variables have approximately constant mean and variance. They do not follow any trend. We will later check the stationary of each variable.

Granger Causality

My y variable is temperature, those are the results of the granger causality test between y and each other variable:

Granger Causality Test Results:
Wind Speed (km/h): p-value = 0.009158267882606536
Wind Bearing (degrees): p-value = 0.09758380729073765
Visibility (km): p-value = 0.00915913478747074

Interpretation:

Wind Speed (km/h): The p-value (0.0092) is less than the 0.05 threshold, indicating strong evidence against the null hypothesis. Thus, there is statistical evidence to suggest that past values of 'Wind Speed (km/h)' Granger cause 'Temperature (C)'.

Wind Bearing (degrees): The p-value (0.0976) is greater than 0.05, which means there isn't sufficient evidence to reject the null hypothesis. In other words, there's no significant Granger causality found between 'Wind Bearing (degrees)' and 'Temperature (C)' at the specified significance level.

Visibility (km): Similar to 'Wind Speed (km/h)', the p-value (0.0092) is less than 0.05, indicating strong evidence against the null hypothesis. Thus, there is statistical evidence to suggest that past values of 'Visibility (km)' Granger cause 'Temperature (C)'.

In summary:

'Wind Speed (km/h)' and 'Visibility (km)' have p-values below 0.05, suggesting a significant causal relationship with 'Temperature (C)' based on the Granger causality test.

However, 'Wind Bearing (degrees)' does not show a significant causal relationship with 'Temperature (C)' based on the p-value obtained (above 0.05).

So I removed Wind Bearing from the dataset.

We end up with temperature, visibility and wind speed.

Stationarity test

Results of Dickey-Fuller Test for temperature:

```
Test Statistic                  -2.427507
p-value                          0.134123
Lags Used                       12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
```

Results of Dickey-Fuller Test for wind speed:

```
Test Statistic                  -2.096256
p-value                          0.246023
Lags Used                       12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
```

Results of Dickey-Fuller Test for visibility:

```
Test Statistic                  -0.998010
p-value                          0.753989
Lags Used                       13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
```

Based on the test results the data is not stationary at level.

Co-integration test:

At first we split the data into training and testing set, we take the first 80% of observations as train and the last 20 % as test. So we have 105 training observations and 27 test observations.

We will use Johansen co-integration test using trace and using maximum Eigenvalue.

Since I am using python that’s the code for the cointegration test:

```
select_coint_rank(train_ecm, det_order = 0, k_ar_diff = 6,
                  method = 'trace', signif=0.05)
```

I will explain what each parameter means and then say how I chose its value.

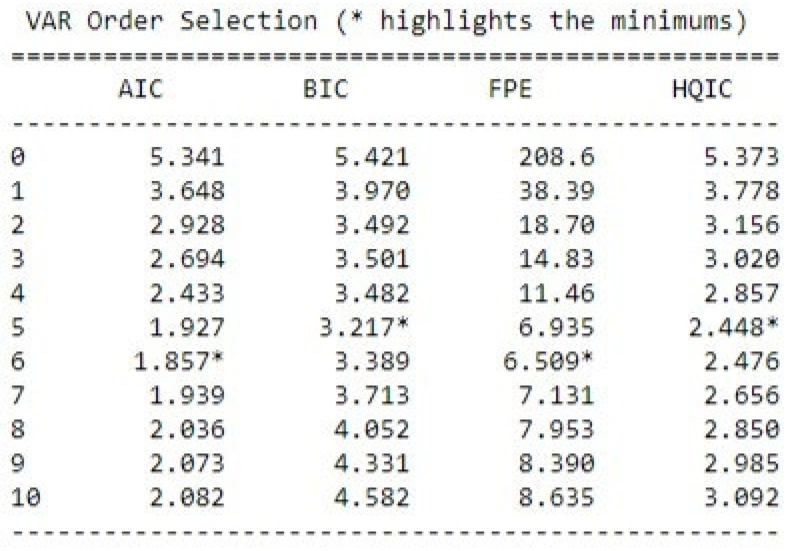
det_order = 0: This parameter refers to the deterministic order in the cointegration test. It specifies the number of deterministic terms (like a constant or a trend) to include in the model. A value of 0 means no deterministic terms, i.e. , just the variables themselves without any additional trend or constant.

k_ar_diff = 6: This parameter represents the maximum number of autoregressive differences to consider during the test. It defines the maximum order of differencing used in the autoregressive process. This parameter helps in identifying the order of integration of the time series.

Why I chose k_ar_diff = 6:

We used the select_order method of the VAR model to determine the optimal lag order for the VAR model.(I used 10 lags)

Those are the results:



```
 VAR Order Selection (* highlights the minimums)
==================================================
       AIC         BIC         FPE         HQIC
--------------------------------------------------
0        5.341       5.421       208.6       5.373
1        3.648       3.970       38.39       3.778
2        2.928       3.492       18.70       3.156
3        2.694       3.501       14.83       3.020
4        2.433       3.482       11.46       2.857
5        1.927      3.217*       6.935      2.448*
6       1.857*       3.389      6.509*       2.476
7        1.939       3.713       7.131       2.656
8        2.036       4.052       7.953       2.850
9        2.073       4.331       8.390       2.985
10       2.082       4.582       8.635       3.092
--------------------------------------------------
```

The optimal lag length is 6 based on AIC, that’s why we chose 6 as the k_ar_diff parameter value.

Now let’s check the results of the cointegrated tests starting using trace.

Johansen cointegration test using trace test statistic with 5% significance level:

```
=====================================
r_0 r_1 test statistic critical value
-------------------------------------
  0   3          72.24          29.80
  1   3          13.30          15.49
-------------------------------------
```

Stage 1:

H0 : k=0

H1 : k>0

If H0 cannot be rejected, stop testing, and k = 0. If null is rejected, perform next test.

Stage 2:

H0 : k <=1

H1 : k>1

If H0 cannot be rejected, stop testing, and k <=1. If null is rejected, perform next test.

Interpretation:

For r=0 (stage 0) the test statistic (72.24) > critical value (29.8) ➔ we reject H0

So there is at least one co-integrated relationship.

For r=1 (stage 1) the test statistic (13.3) < critical value (15.49) ➔ We fail to reject H0, r<=1, suggesting the presence of at most 1 cointegrated relationship.

Now let's perform the co-integrated test using maximum eigen value

Johansen co-integration test using maximum eigenvalue test statistic with 5% significance level:

```
=====================================
r_0 r_1 test statistic critical value
-------------------------------------
  0   1          58.94          21.13
  1   2          13.29          14.26
-------------------------------------
```

Considers the null hypothesis that the co-integrating rank is k.

against the alternative hypothesis that the co-integrating rank is k+1.

H0 : r = k

H1 : r = k+1

Interpretation:

At stage 1 for r=0 , we reject the null hypothesis because 58.94 > 21.13 ➔ r=1

At stage 2 for r=1, we fail to reject the null hypothesis because 13.29 <14.26 ➔r=1

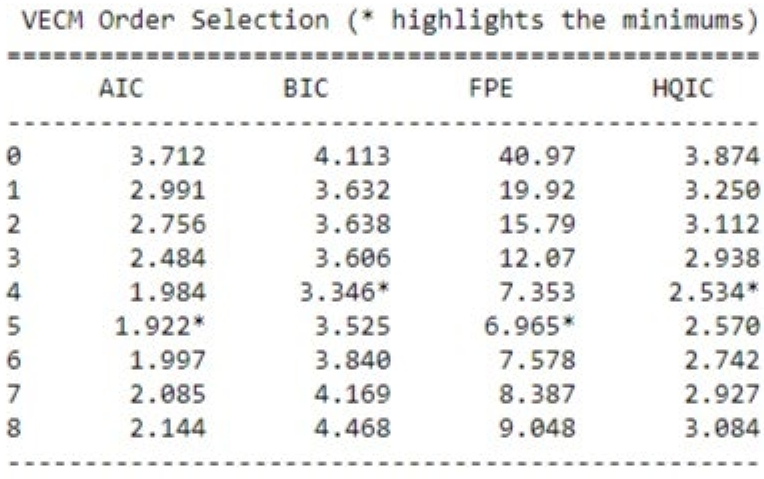
So we have 1 co-integrating relationship.

Based on both methods we find that the best rank is 1.

So we will use VECM model.

VECM MODEL

Now before fitting the VECM model we want to check the optimal lag order for vecm, to do this we use vecm order selection.



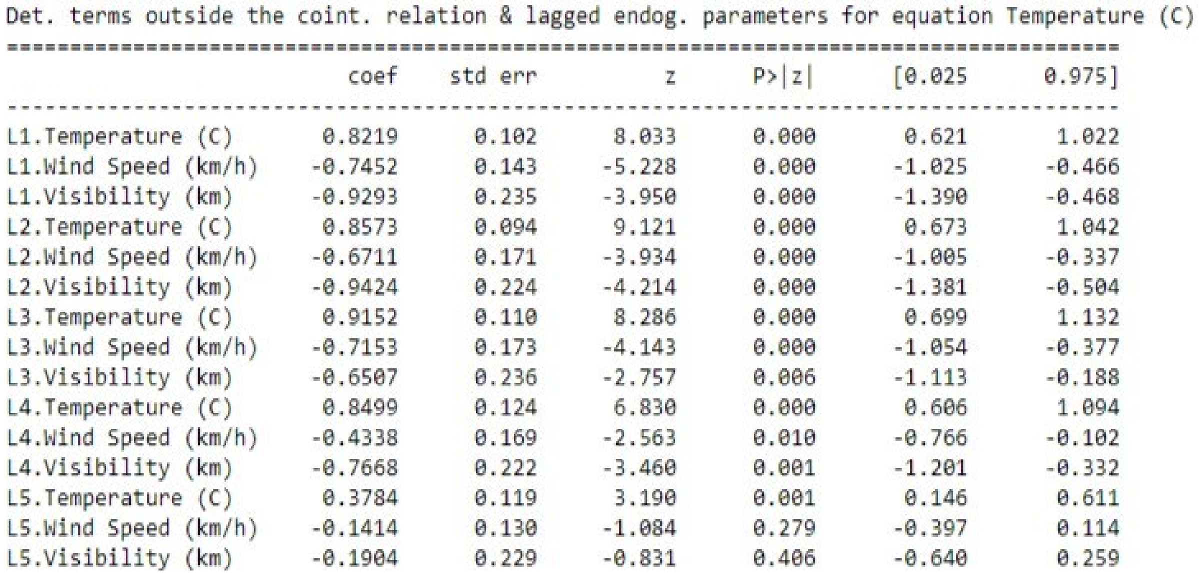
```
 VECM Order Selection (* highlights the minimums)
==================================================
      AIC         BIC         FPE         HQIC
--------------------------------------------------
0       3.712       4.113       40.97       3.874
1       2.991       3.632       19.92       3.250
2       2.756       3.638       15.79       3.112
3       2.484       3.606       12.07       2.938
4       1.984      3.346*       7.353      2.534*
5      1.922*       3.525      6.965*       2.570
6       1.997       3.840       7.578       2.742
7       2.085       4.169       8.387       2.927
8       2.144       4.468       9.048       3.084
--------------------------------------------------
```

Based on AIC the optimal lag is 5 for the vecm model.

That’s our model:

model = VECM(train_ecm, k_ar_diff=5, coint_rank=1, deterministic='ci')

Summary of our model:



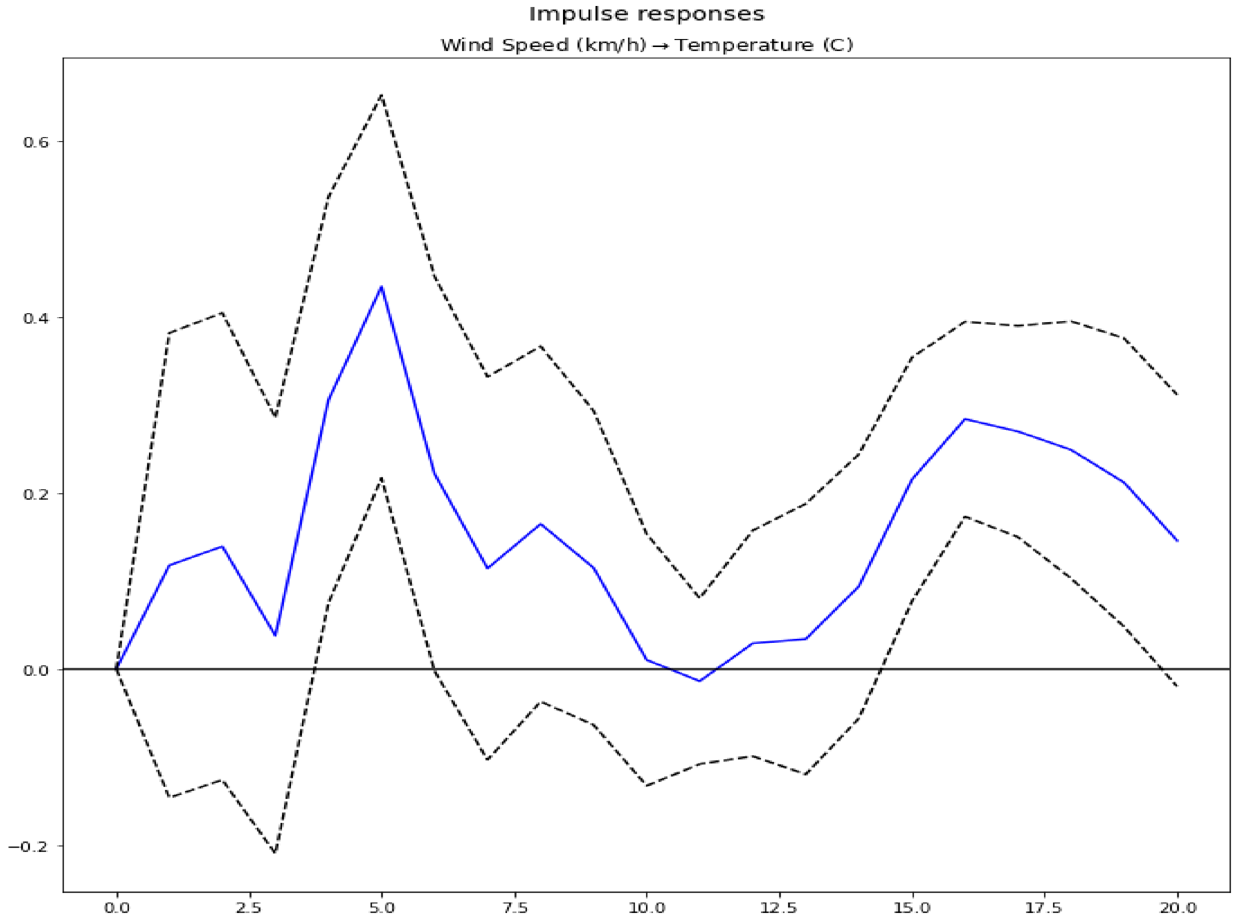
```
Det. terms outside the coint. relation & lagged endog. parameters for equation Temperature (C)
=========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------
L1.Temperature (C)       0.8219      0.102      8.033      0.000       0.621       1.022
L1.Wind Speed (km/h)    -0.7452      0.143     -5.228      0.000      -1.025      -0.466
L1.Visibility (km)      -0.9293      0.235     -3.950      0.000      -1.390      -0.468
L2.Temperature (C)       0.8573      0.094      9.121      0.000       0.673       1.042
L2.Wind Speed (km/h)    -0.6711      0.171     -3.934      0.000      -1.005      -0.337
L2.Visibility (km)      -0.9424      0.224     -4.214      0.000      -1.381      -0.504
L3.Temperature (C)       0.9152      0.110      8.286      0.000       0.699       1.132
L3.Wind Speed (km/h)    -0.7153      0.173     -4.143      0.000      -1.054      -0.377
L3.Visibility (km)      -0.6507      0.236     -2.757      0.006      -1.113      -0.188
L4.Temperature (C)       0.8499      0.124      6.830      0.000       0.606       1.094
L4.Wind Speed (km/h)    -0.4338      0.169     -2.563      0.010      -0.766      -0.102
L4.Visibility (km)      -0.7668      0.222     -3.460      0.001      -1.201      -0.332
L5.Temperature (C)       0.3784      0.119      3.190      0.001       0.146       0.611
L5.Wind Speed (km/h)    -0.1414      0.130     -1.084      0.279      -0.397       0.114
L5.Visibility (km)      -0.1904      0.229     -0.831      0.406      -0.640       0.259
```

This model got us an AIC = 939

Impulse Response Function

Wind Speed --> Temperature:

This plot shows the response of the temperature variable to a shock in the wind speed variable. It illustrates the impact of a one standard deviation shock in wind speed on the temperature variable over time.
It indicates the dynamics of how changes in wind speed influence temperature fluctuations.

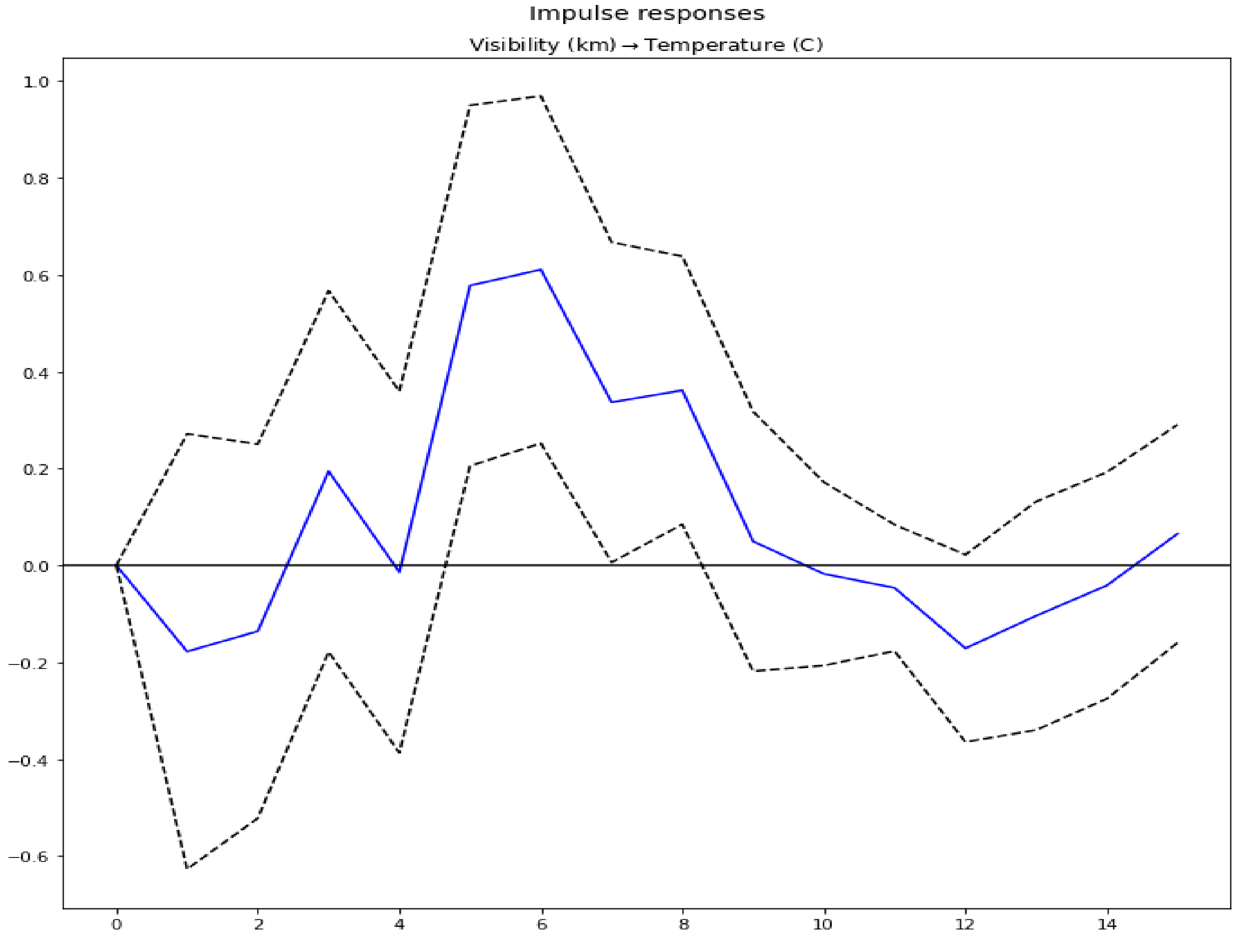


Interpretation of the plot:

One standard deviation positive shock to wind speed leads to a 0.17 standard deviation increase in the temperature in the second time period after the shock. In time period 5 the shock further causes a 0.42 standard deviation increase in temperature. The effects of the shock start to reduce after the fifth time period and keeps on decreasing till reaching 0 at time 12 and then it repeats itself because we have seasonal monthly data.

Visibility --> Temperature:

This plot represents the response of the temperature variable to a shock in the vis ibility variable.
It demonstrates how a one-standard-deviation shock in visibility affects the temp erature variable over time.
It provides insights into how changes in visibility impact temperature changes.



Interpretation:

One standard deviation negative shock to visibility leads to a -0.18 standard devia tion decrease in the temperature in the first time period after the shock. In time period 3 the shock causes a 0.2 standard deviation increase in temperature. The effects of the shock keep’s on increasing till reaching its peak at period 6 and then again starts to decrease till period 12, then it repeats itself because we have seasonal monthly data.

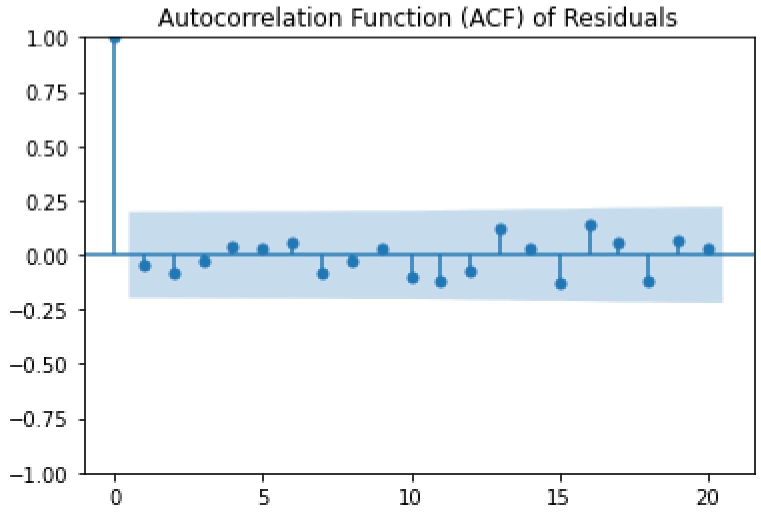
Model Diagnostics

Residual Auto Correlation:

Those are the results of Durbin Watson test:
Temperature (C) : 2.09
Wind Speed (km/h) : 2.08
Visibility (km) : 2.08

The values are close to 2 indicating no significant auto correlation

ACF plot of residuals:



The plot also shows no autocorrelation.

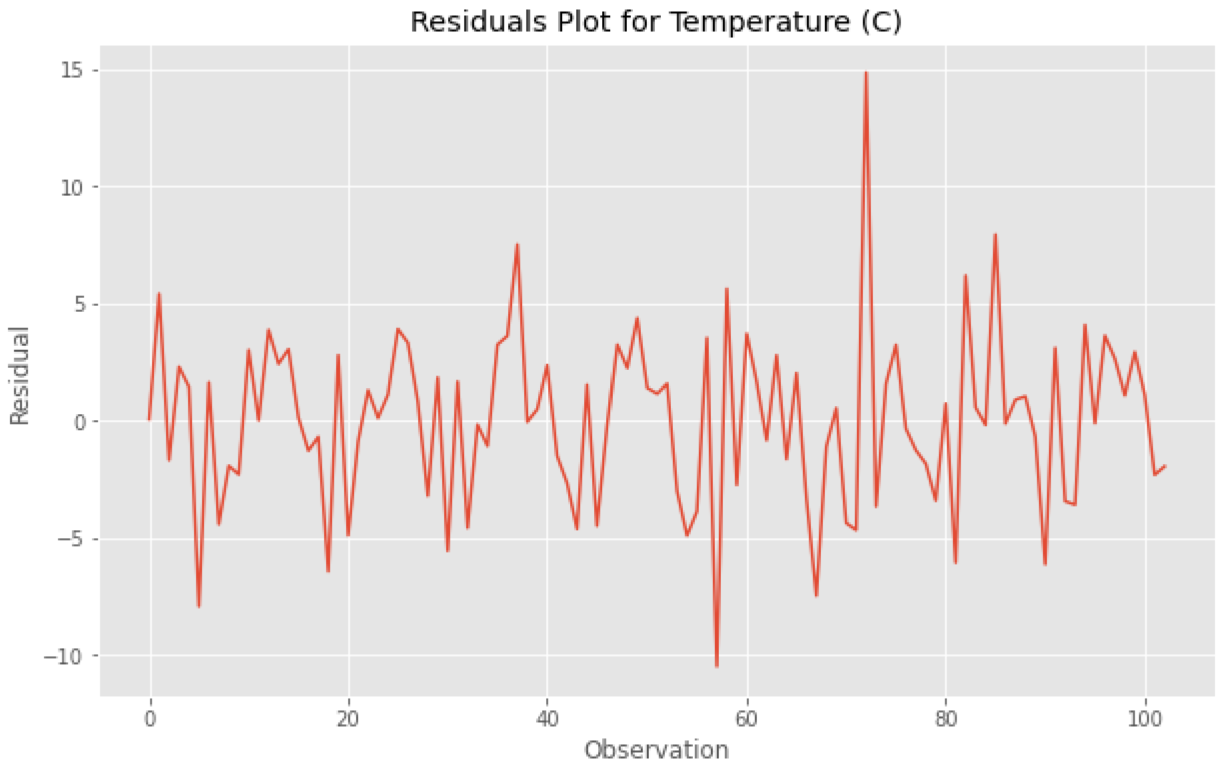
Ljung box test:

```
   lb_stat    lb_pvalue
1  0.223578   0.636327
2  0.872034   0.646607
3  0.936704   0.816563
4  1.117309   0.891516
```

```
5   1.211394   0.94378
6   1.549605   0.956132
7   2.29332    0.94184
8   2.366903   0.967649
9   2.431449   0.982671
10  3.59338    0.963832
11  5.25159    0.918372
12  5.78756    0.92641
13  7.46621    0.876556
14  7.54924    0.911464
15  9.39355    0.856057
16  11.6996    0.764385
17  12.0418    0.797598
18  13.67341   0.750116
19  14.2632    0.768117
```

We fail to reject H0(P values > 0.05) ➔ The residuals are independently Distributed.

Residual plot



Heteroscedasticity test

Using ARCH test:

```
ARCH test statistic: 13.547087984335015
p-value: 0.19467029998865074
```

The p-value of approximately 0.195 obtained from the ARCH test indicates that there isn't strong evidence to reject the null hypothesis of no conditional heteroscedasticity in the residuals at the 5% significance level. This suggests that the residuals may have constant variance (homoscedasticity) and do not exhibit significant conditional heteroscedasticity.

Stationarity of Residuals:

```
ADF test statistic: -7.744002260096455
p-value for adf: 1.0436638591954426e-11
KPSS test statistic: 0.11380994501613659
p-value for KPSS: 0.1
```

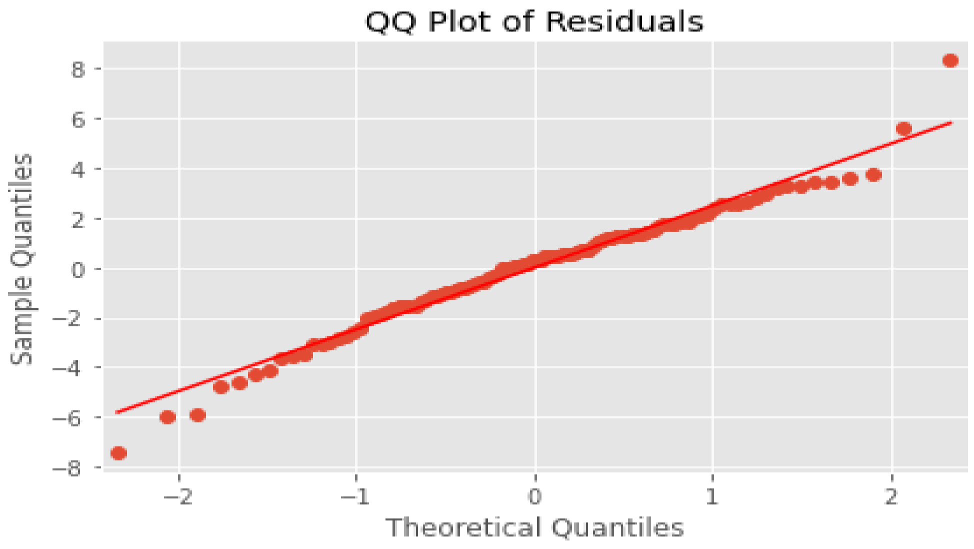
Based on ADF the residuals are stationary (p-value < 0.05) and based on KPSS the residuals are stationary (p-value>0.05)

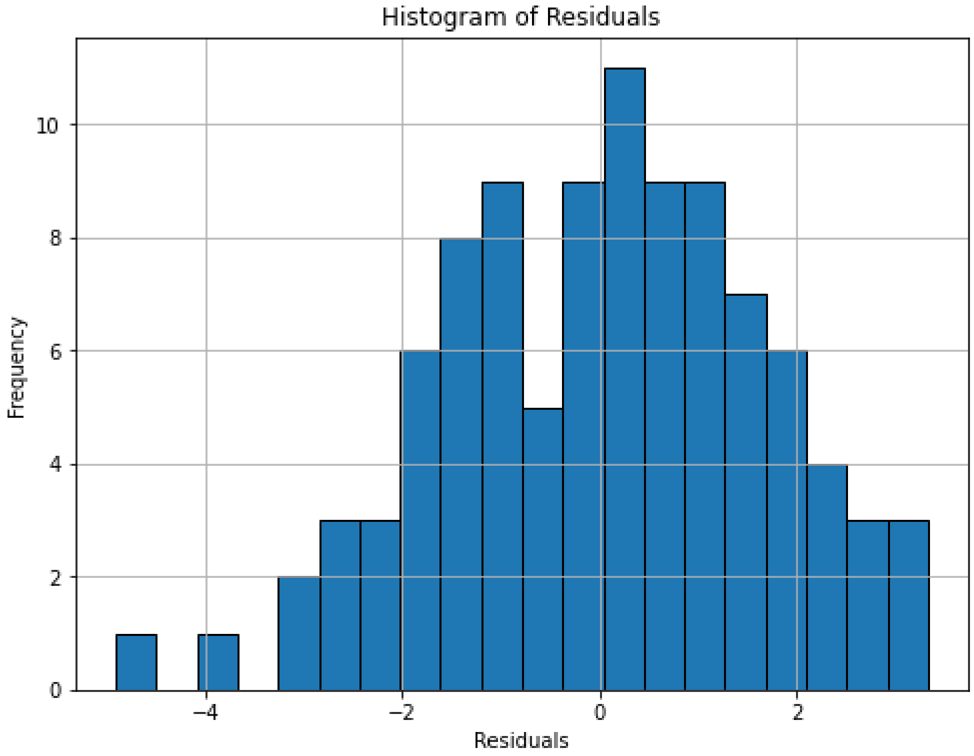
Normality of Residuals:

```
Jarque-Bera test statistic: 1.3314653236756469
p-value: 0.5138968771142847
```

```
Shapiro-Wilk Test Statistic: 0.9905331134796143
p-value: 0.7139136791229248
```

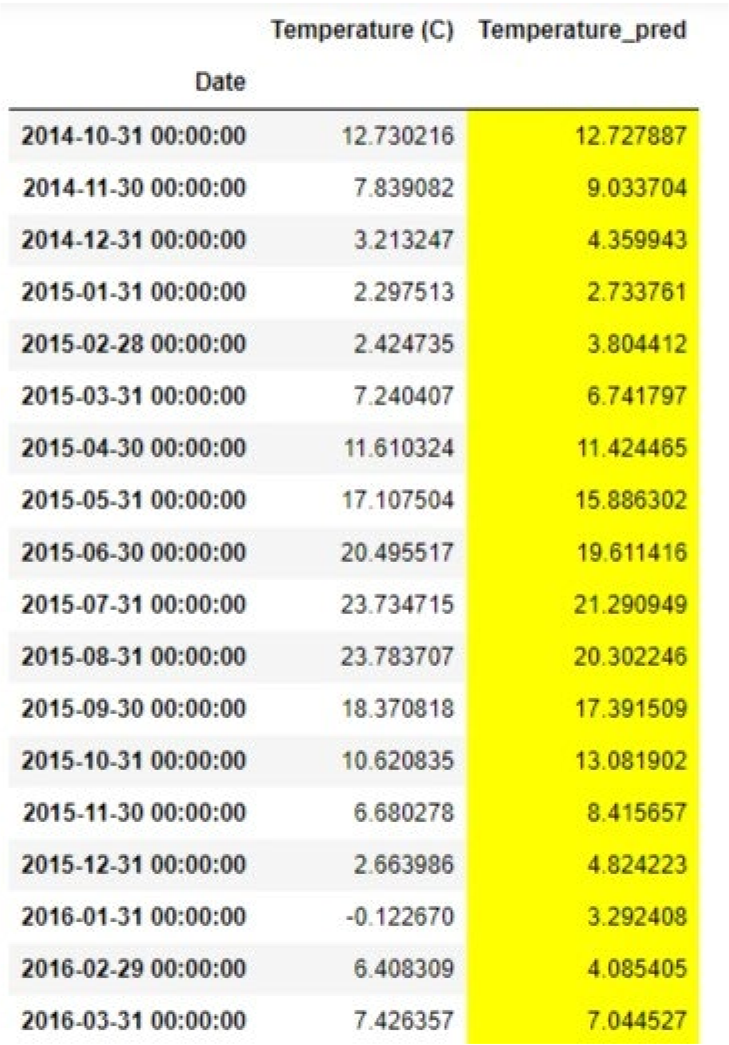
Based on the results from both tests the residuals appear to be normally distributed (fail to reject H0)





Now we will test the model on the testing set.

Those are the results between the actual data and predicted data:

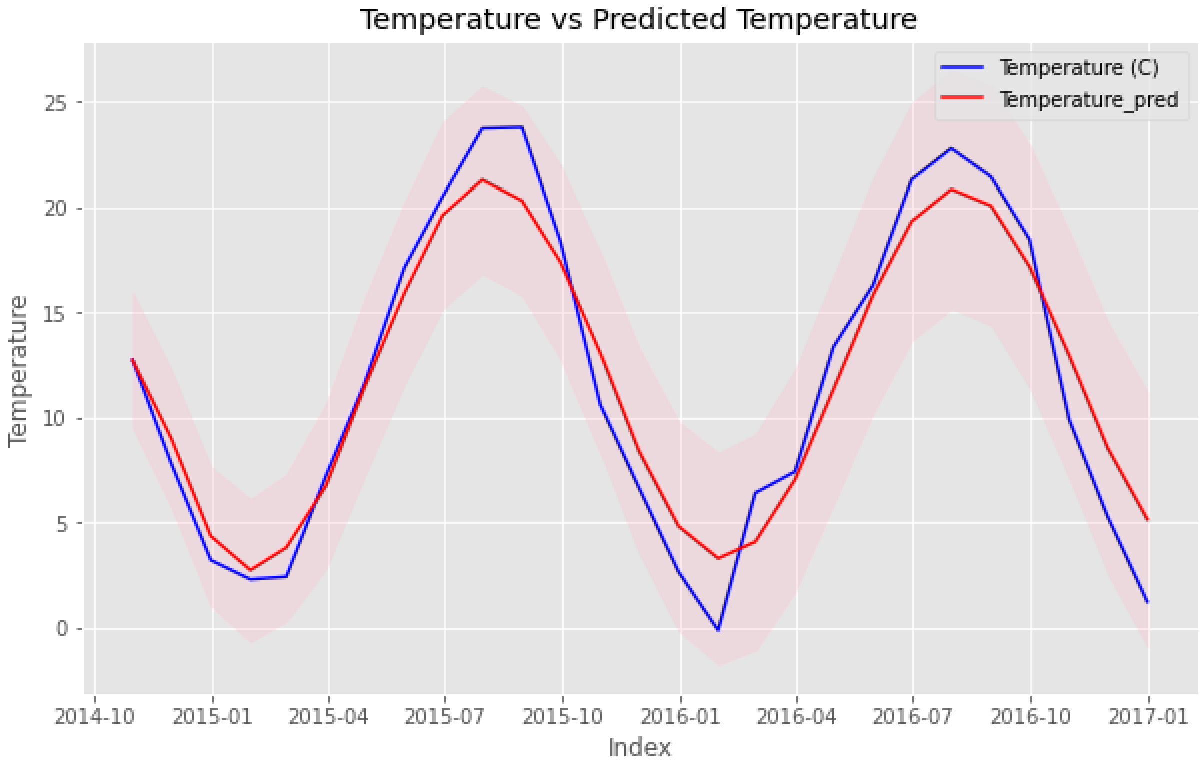


	Temperature (C)	Temperature_pred
Date		
2014-10-31 00:00:00	12.730216	12.727887
2014-11-30 00:00:00	7.839082	9.033704
2014-12-31 00:00:00	3.213247	4.359943
2015-01-31 00:00:00	2.297513	2.733761
2015-02-28 00:00:00	2.424735	3.804412
2015-03-31 00:00:00	7.240407	6.741797
2015-04-30 00:00:00	11.610324	11.424465
2015-05-31 00:00:00	17.107504	15.886302
2015-06-30 00:00:00	20.495517	19.611416
2015-07-31 00:00:00	23.734715	21.290949
2015-08-31 00:00:00	23.783707	20.302246
2015-09-30 00:00:00	18.370818	17.391509
2015-10-31 00:00:00	10.620835	13.081902
2015-11-30 00:00:00	6.680278	8.415657
2015-12-31 00:00:00	2.663986	4.824223
2016-01-31 00:00:00	-0.122670	3.292408
2016-02-29 00:00:00	6.408309	4.085405
2016-03-31 00:00:00	7.426357	7.044527

Model Metrics:

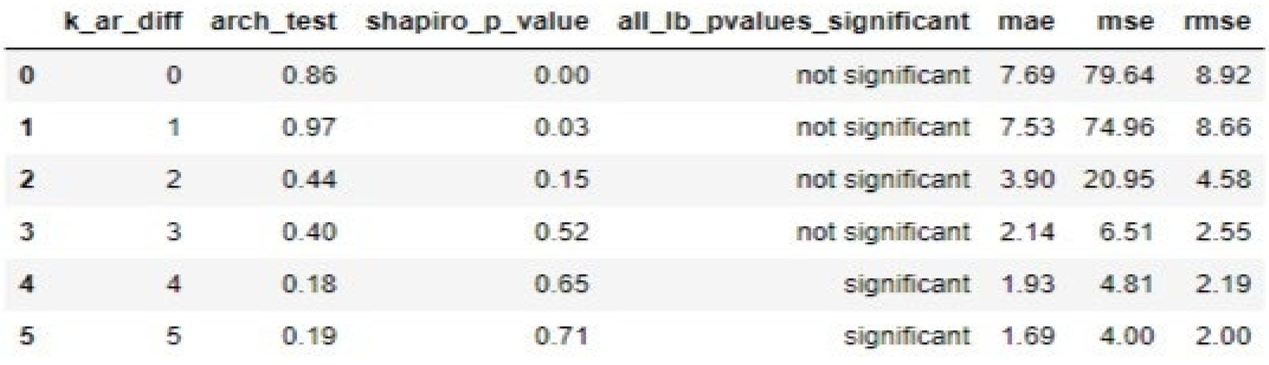
MAE = 1.6895655892222592
MSE = 4.003793032912269
RMSE = 2.000948033536171

Temperature vs Predicted temperature plot:



We can see that the fit is very good. The observed values fall within the confidence interval of predicted values, the model's forecasts are accurate within the given level of confidence. It signifies that the observed data is well within the expected range of variability as estimated by the forecasting model, considering the uncertainty associated with the predictions.

That was the first model with lag = 5, we can also try all the vecm models with lags <5 and 1 co-integrated relationship, I will do the tests for each model and return them in a data frame with their results.



	k_ar_diff	arch_test	shapiro_p_value	all_lb_pvalues_significant	mae	mse	rmse
0	0	0.86	0.00	not significant	7.69	79.64	8.92
1	1	0.97	0.03	not significant	7.53	74.96	8.66
2	2	0.44	0.15	not significant	3.90	20.95	4.58
3	3	0.40	0.52	not significant	2.14	6.51	2.55
4	4	0.18	0.65	significant	1.93	4.81	2.19
5	5	0.19	0.71	significant	1.69	4.00	2.00

Those are the results of the 6 models, we did the arch test for heteroskedasticity, shapiro test for normality and the 4th column (all_lb_pvalues_significant) is the result of the ljung box test.

Conclusion

We can see that the models that have lag 0,1,2,3 are not independently distributed; they exhibit serial correlation (results of the 4th column).

So we end up with two models:

-model with 4 lags

-model with 5 lags (that we did earlier)

Both passes the diagnostic tests, so let's choose based on metrics.

We can see that model with 5 lags have lower rmse, mae, mse and AIC than the model with 4 lags so our best model is the ecm model with 5 lags and 1 co-integrated relationship.

Expression of Gratitude and Request for Feedback

Thank you for taking the time to review my work and investing your valuable attention in it. I sincerely appreciate your consideration and would greatly benefit from your feedback and remarks. Your insights are immensely valuable to me as they contribute significantly to refining and advancing my projects. I am open to any suggestions or observations you may have, as they would greatly assist in enhancing the quality and effectiveness of my work. Once again, thank you for your time and consideration.