Univariate Time Series Analysis

By Samir Abou Jaoude

Problem Statement:

The goal of this project is to develop an accurate and reliable predictive model for forecasting monthly temperatures. Temperature prediction is essential for various sectors, including agriculture, energy management, and urban planning. Accurate forecasts enable better decision-making, resource allocation, and preparation for temperature-related events.

Addressing the Problem

In this project, the aim is to predict monthly temperatures through an exhaustive exploration of diverse forecasting models, encompassing Simple Moving Average (SMA), Exponential Moving Average (EMA), Holt-Winters, and Seasonal Autoregressive Integrated Moving Average (SARIMA). The endeavor begins with meticulous data collection and preprocessing, ensuring the quality and consistency of historical monthly temperature data. Subsequently, an in-depth Exploratory Data Analysis (EDA) is conducted to unveil inherent patterns, trends, and potential outliers. Addressing the critical issues of stationarity and seasonality, the data undergoes transformations and statistical tests to render it suitable for time series analysis. The model selection process involves the progressive implementation of forecasting techniques, starting with simpler models and culminating in the application of SARIMA. Training and validation phases follow, with a keen eye on hyperparameter tuning to optimize model performance. The results are comprehensively compared, assessing the strengths and limitations of each model in capturing the multifaceted aspects of monthly temperature variations. This methodology ensures a holistic

exploration of forecasting techniques, providing valuable insights into their efficacy in addressing the inherent challenges of temperature prediction.

Data Exploration and Preprocessing

That's my date column, we have hourly data, the first date here is 2006-04-01 where 2006 is the year, 04 is the month and 01 is the day.

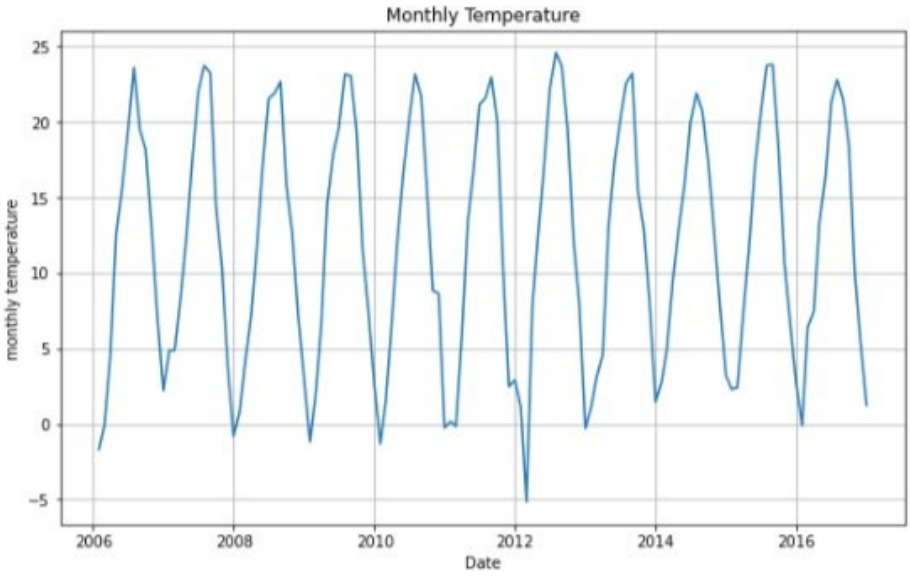
```
0         2006-04-01 00:00:00.000 +0200
1         2006-04-01 01:00:00.000 +0200
2         2006-04-01 02:00:00.000 +0200
3         2006-04-01 03:00:00.000 +0200
4         2006-04-01 04:00:00.000 +0200
                     ...
96448     2016-09-09 19:00:00.000 +0200
96449     2016-09-09 20:00:00.000 +0200
96450     2016-09-09 21:00:00.000 +0200
96451     2016-09-09 22:00:00.000 +0200
96452     2016-09-09 23:00:00.000 +0200
```

We will convert the dates into monthly data by aggregating the average temperature values on a monthly basis.

That's our new data:

```
Date
2006-01-31    -1.674283
2006-02-28    -0.061285
2006-03-31     4.533468
2006-04-30    12.625872
2006-05-31    15.665315
                 ...
2016-08-31    21.433998
2016-09-30    18.478465
2016-10-31     9.904579
2016-11-30     5.289074
2016-12-31     1.239628
Freq: M, Name: Temperature (C)
```

Visualyzing the data

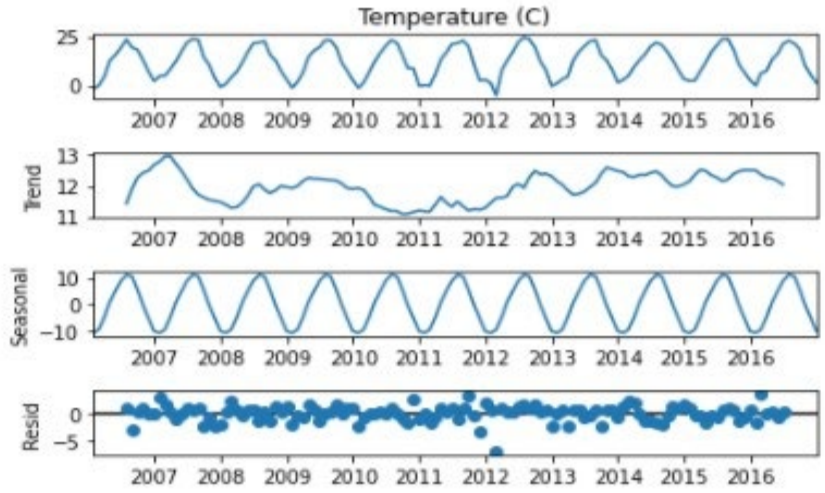


We can clearly see that we have a seasonal pattern every 12 months.

Decomposition of the data:

We will use an additive decomposition because the seasonal component is relatively constant regardless of changes in the trend.

Our data here is aggregated by month. The period we want to analyze is by year so we will set the period to 12.



Now we want to perform stationarity test:

Results of Dickey-Fuller Test:

```
Test Statistic                  -2.42750727
p-value                          0.13412270
Lags Used                       12.00000000
Number of Observations Used    119.00000000
Critical Value (1%)             -3.48653461
Critical Value (5%)             -2.88615099
Critical Value (10%)            -2.57989609
```

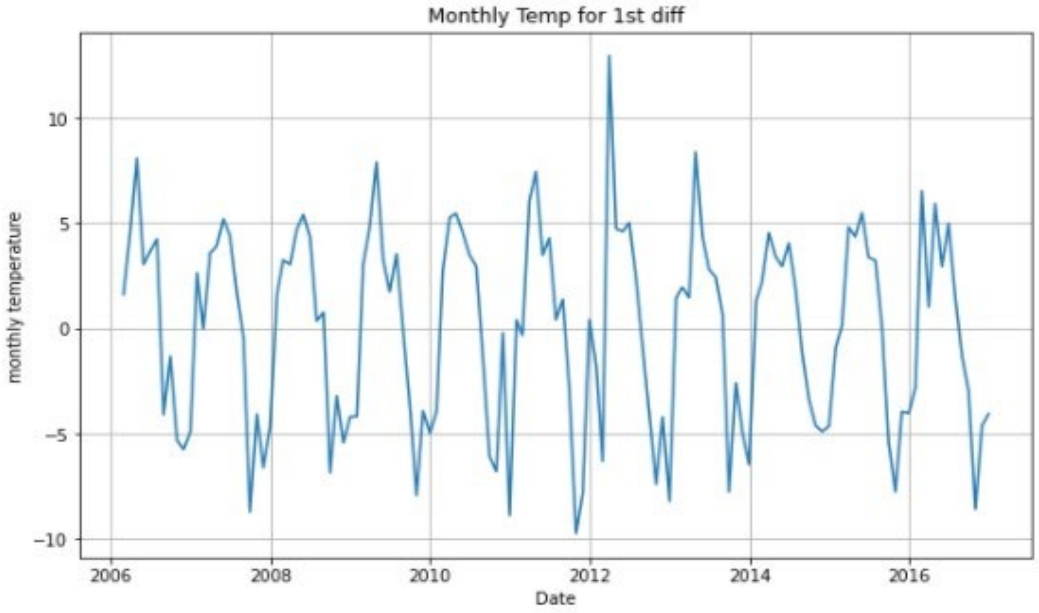
p-value > 0.05 ➔ The data is not stationary

Let's try to make it stationary by performing the first difference on the data and we got those results:

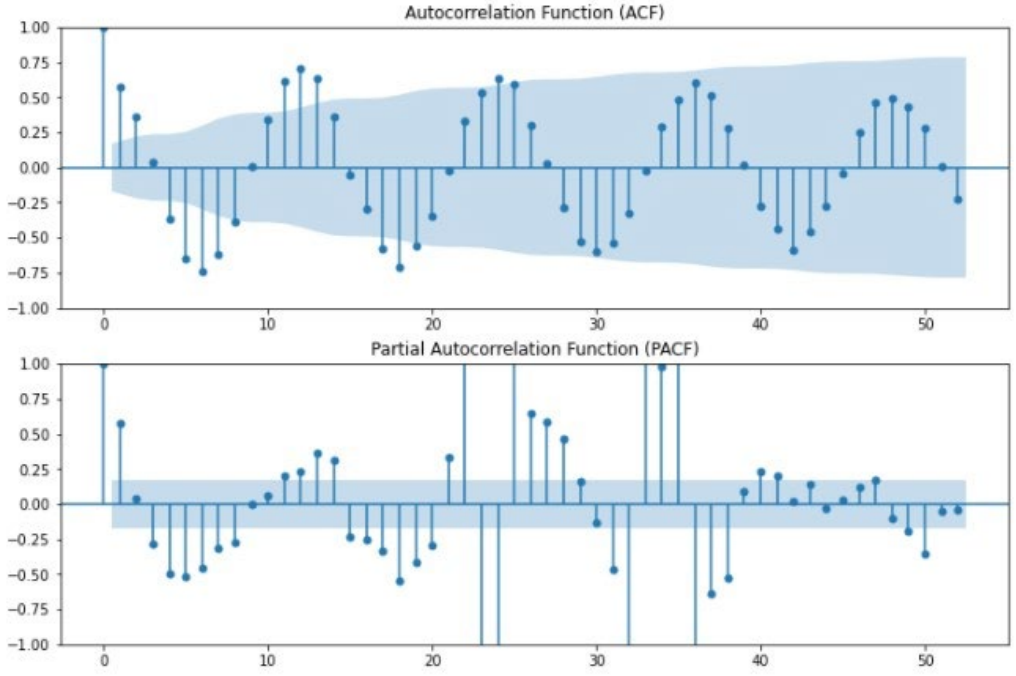
```
Results of Dickey-Fuller Test:
Test Statistic                -8.572765e+00
p-value                        8.113539e-14
Lags Used                      1.100000e+01
Number of Observations Used    1.190000e+02
Critical Value (1%)           -3.486535e+00
Critical Value (5%)           -2.886151e+00
Critical Value (10%)          -2.579896e+00
```

p-value < 0.05 we reject H0, the data is stationary.

We can also visualize the data after differencing



ACF and PACF plot for first difference data.



Based on the ACF and PACF we can see that we still have seasonality after differencing.

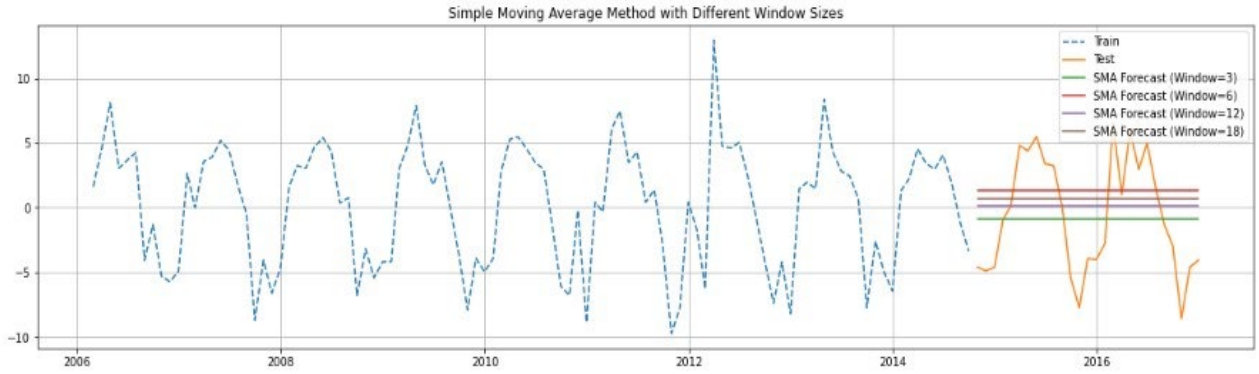
I will try some models before doing the seasonal difference.

We split the data into train and test. The first 80% of observations will be in our training set and the remaining 20% of values will be for our test set.

Starting with Simple Moving Average.

Simple Moving Average

That's the plot for this model:



	Method	MA_Window	RMSE
0	Simple moving average forecast	3	4.39000000
1	Simple moving average forecast	6	4.77000000
2	Simple moving average forecast	12	4.44000000
3	Simple moving average forecast	18	4.57000000

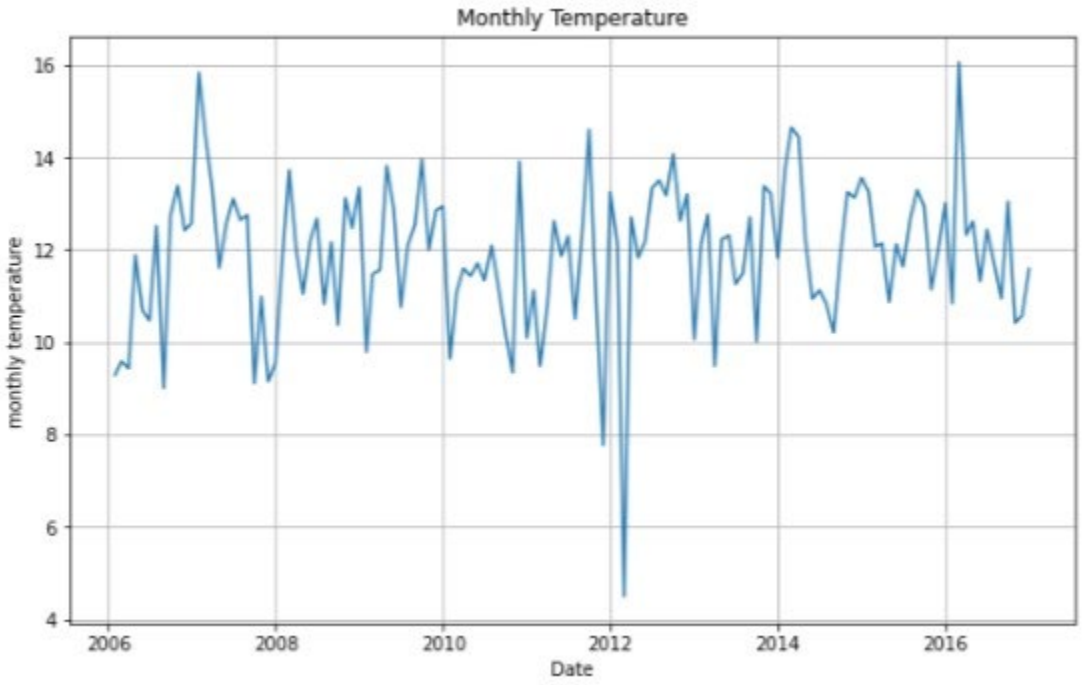
In the above plot the blue line corresponds to the train data, the orange line corresponds to the test data and the others are the forecasted results for window size = 3,6,12,18.

The best RMSE was for the window size = 3.

Let’s also try SMA after de-seasonalizing our data, we just need to subtract the seasonal component from the original data.

Deseason_temp = df – decomposition.seasonal

That’s the plot after seasonal differencing.



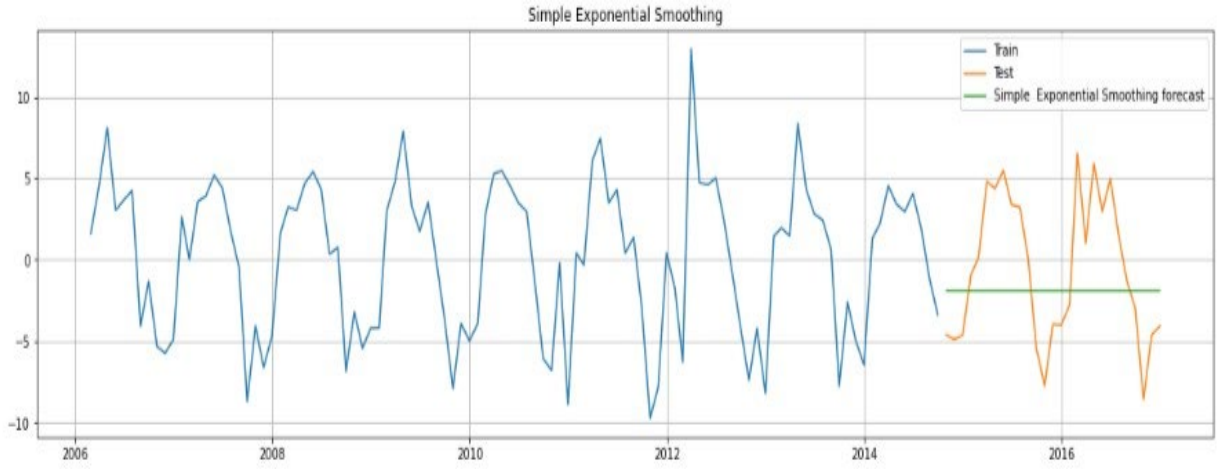
We can see that the seasonality have been removed.

We did the ADF test and we can conclude that the series is stationary.

We got an RMSE score equal to 1.18.

Simple Exponential Smoothing:

Now we will fit a simple exponential model on the differenced data with a smoothing level = 0.2



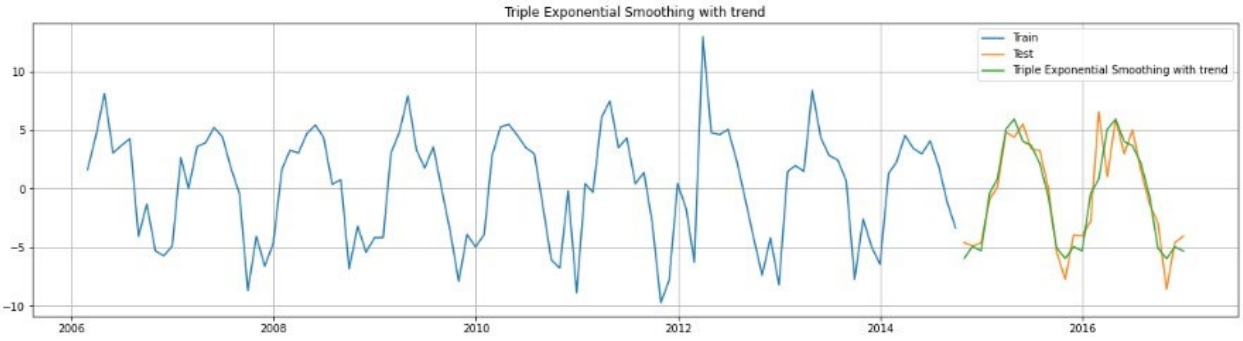
We got an RMSE value equal to 4.47.

We can also see that the fit is not that good (it's not able to detect the seasonal pattern), to try to get a better fit we will use the triple exponential smoothing method.

Triple Exponential Smoothing

We will fit two models:

first model: with trend and with seasonality



As we can see the fit is a lot better than before since Holtwinters takes into consideration the seasonality.

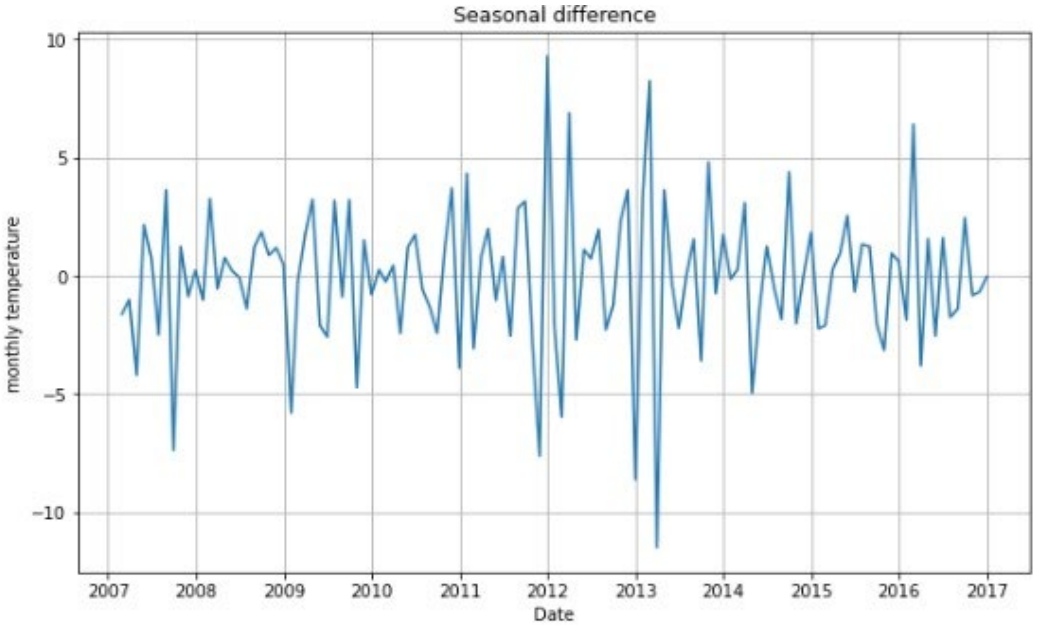
We got an RMSE = 1.79 and an AIC = 188

Second model: without trend and with seasonality

The fit is also very good we got an RMSE = 1.79 just like the previous but the AIC is slightly better than before equal to 184.

SARIMA:

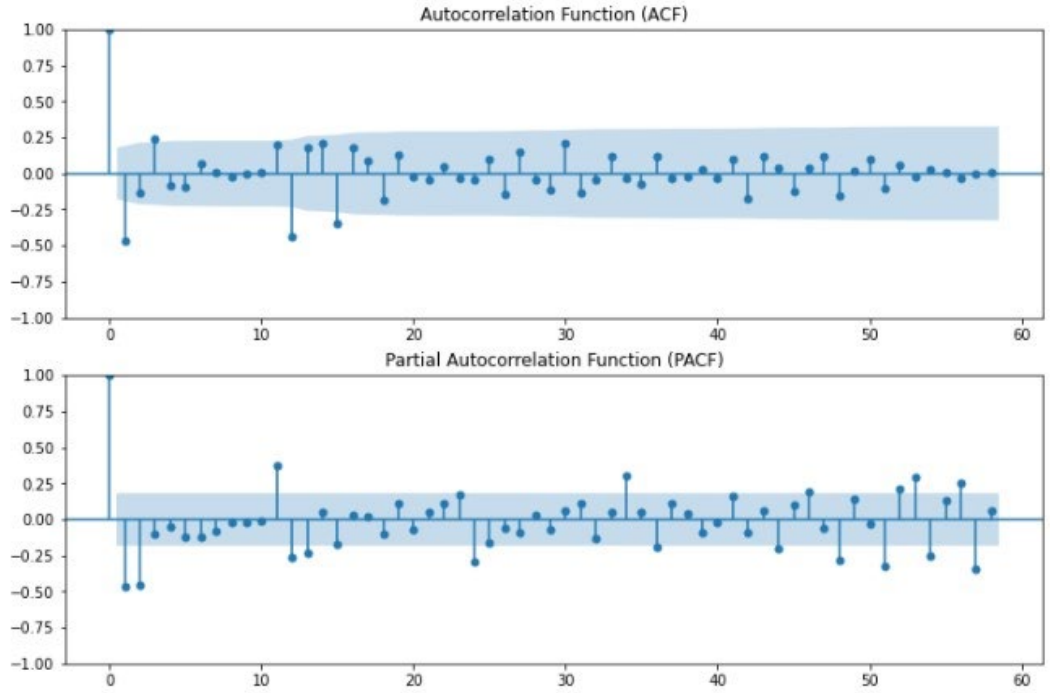
Before modeling we will do seasonal differencing (12) on the already differenced data



That’s the shape of the new data as we can see the seasonality have been removed and we did the ADF test, we got a p value < 0.05 so it’s stationary.

Let’s visualize the ACF and PACF and try to conclude p,q,P,Q.

We already know that d=1 and D=1.



We have some decaying spikes in the PACF plot at k=12,24,36

then at k =48 we have a higher one, now based on the ACF plot we have a seasonal spike at k=12 so I will take Q=1.

We can start with p=0, q=1, P=0, Q=1.

p, d, q, P, D, Q, m = 0,1,1,0,1,1,12

We got insignificant coefficients (p-values > 0.05)

Dep. Variable:	Temperature (C)	**No. Observations:**	105
Model:	SARIMAX(0, 1, 1)x(0, 1, 1, 12)	**Log Likelihood**	-198.010
Date:	Sat, 11 Nov 2023	**AIC**	402.021
Time:	21:32:07	**BIC**	409.586
Sample:	01-31-2006	**HQIC**	405.074
	- 09-30-2014		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9486	0.075	-12.660	0.000	-1.095	-0.802
ma.S.L12	-0.9992	33.102	-0.030	0.976	-65.878	63.879
sigma2	3.1525	104.100	0.030	0.976	-200.880	207.185

Ljung-Box (L1) (Q):	1.75	**Jarque-Bera (JB):**	5.80
Prob(Q):	0.19	**Prob(JB):**	0.05
Heteroskedasticity (H):	0.74	**Skew:**	-0.53
Prob(H) (two-sided):	0.41	**Kurtosis:**	3.63

Let’s try some other model:

p, d, q, P, D, Q, m = 0,1,1,1,1,1,12

Again we got insignificant coefficients.

Another model:

p, d, q, P, D, Q, m = 1,1,1,0,1,1,12

Again insignificant parameters.

Another way to choose a model is doing grid search that returns the model with the lowest AIC.

Those are the parameters:

p_values = range(0, 3) # Autoregressive order (0,1,2)

d_values = [1] # Differencing order

q_values = range(0, 3) # Moving average order (0,1,2)

P_values = range(0, 3) # Seasonal autoregressive order (0,1,2)

D_values = [1] # Seasonal differencing order

Q_values = range(0, 3) # Seasonal moving average order (0,1,2)

m_values = [12] # Seasonal period

We got those results:

```
Best Parameters: (1, 1, 1, 0, 1, 1, 12)
Best AIC: 401.4548848593519
```

Again we got insignificant coefficients.

Since we have difficulty finding a model with significant coefficients, another approach I took is again doing a grid search with the same parameters as above but this time I analyzed the summary of each model and chose those who are significant (p-value < 0.05).

We have in total 3 x 1 x 3 x 3 x 1 x 3 x 1 = 81 models.

We end up with 8 models:

p, d, q, P, D, Q, m =0, 1, 1, 2, 1, 0, 12

p, d, q, P, D, Q, m =2, 1, 0, 2, 1, 0, 12

p, d, q, P, D, Q, m =2, 1, 1, 1, 1, 0, 12

p, d, q, P, D, Q, m =2, 1, 0, 1, 1, 0, 12

p, d, q, P, D, Q, m =1, 1, 0, 2, 1, 0, 12

p, d, q, P, D, Q, m =0, 1, 1, 0, 1, 0, 12

p, d, q, P, D, Q, m =2, 1, 0, 0, 1, 0, 12

p, d, q, P, D, Q, m =1, 1, 0, 1, 1, 0, 12

All those models have significant coefficients. Now we will choose the best model based on AIC and RMSE but first we need to do some hypothesis testing.

We need to check:

Normality of residuals: using shapiro and jarque-bera

Random residuals: using runs test

Residuals are independently distributed: using box test

We will also check ACF and PACF plots for residuals.

Those are the models sorted ascendingly based on AIC and we can also see the values of the tests for each model.

Model	AIC	RMSE	Shapiro P-value	Jarque-Bera P-value	Runs Test P-value	Residuals Distribution
(0, 1, 1, 2, 1, 0, 12)	414.33875205	1.55000000	0.50633270	0.74223677	0.20499331	Normal
(2, 1, 0, 2, 1, 0, 12)	419.11260684	2.36000000	0.19486429	0.71589151	0.80293915	Normal
(2, 1, 1, 1, 1, 0, 12)	428.02853076	2.15000000	0.56593883	0.98642572	0.06254306	Normal
(2, 1, 0, 1, 1, 0, 12)	428.38976196	2.08000000	0.66946447	0.88501400	0.78603642	Normal
(1, 1, 0, 2, 1, 0, 12)	436.01431826	2.57000000	0.13651660	0.44107602	0.35107267	Normal
(0, 1, 1, 0, 1, 0, 12)	441.82821716	1.77000000	0.78352141	0.89122370	0.06254306	Normal
(2, 1, 0, 0, 1, 0, 12)	442.21740880	1.58000000	0.95555717	0.90096511	0.29016386	Normal
(1, 1, 0, 1, 1, 0, 12)	445.33892344	1.88000000	0.41415390	0.78800608	0.92113383	Normal

As we can every model satisfies the normality (Shapiro p-value > 0.05, Jarque-Bera p-value>0.05) and the randomness tests (Runs test p-value

> 0.05). We still have to do the box test and plot the ACF and PACF for residuals.

Let's start with analyzing each model starting with the first one.

Model 1:

```
  lb_stat    lb_pvalue
1 6.55301120 0.01047068
2 8.58883595 0.01364451
3 11.0580400 0.01141614
4 11.1720027 0.02469746
5 12.3641184 0.03012518
6 12.5287189 0.05116120
```

Those are the p values of the box test for the first 6 lags of the first model, we can see that we have p-values <0.05 so we won't choose model 1 for now.

Model 2:

Results of box test:

lb_stat lb_pvalue

1 1.94275585 0.16336962

2 2.80482591 0.24600265

3 3.23708122 0.35649621

4 3.36689171 0.49840408

5 4.55765205 0.47220704

6 5.04707182 0.53779015

7 6.02217635 0.53716233

8 7.40493499 0.49363835

9 7.51568290 0.58359016

10 7.6749973 0.66054869

11 7.6769126 0.74191985

12 8.6638331 0.73134363

13 10.104566 0.68536245

14 10.134456 0.75228855

15 12.866352 0.61262009

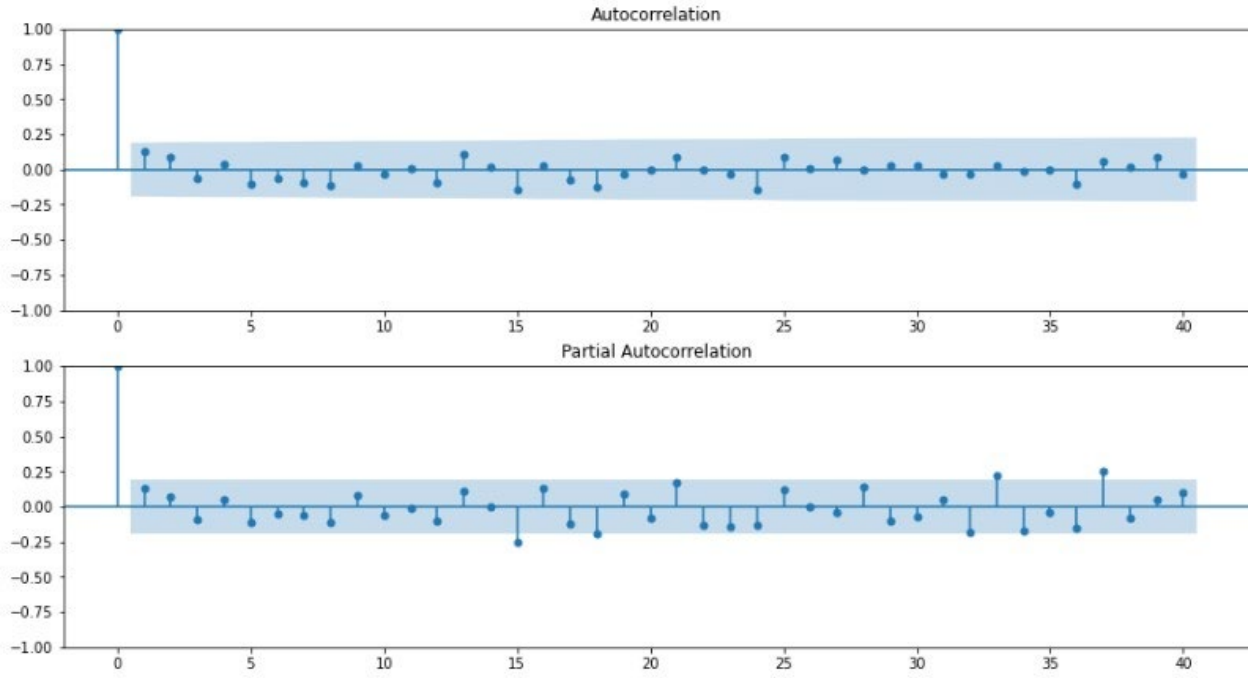
16 12.945328 0.67675094

17 13.584227 0.69628301

18 15.487829 0.62824408

P values are above 0.05 which is good.

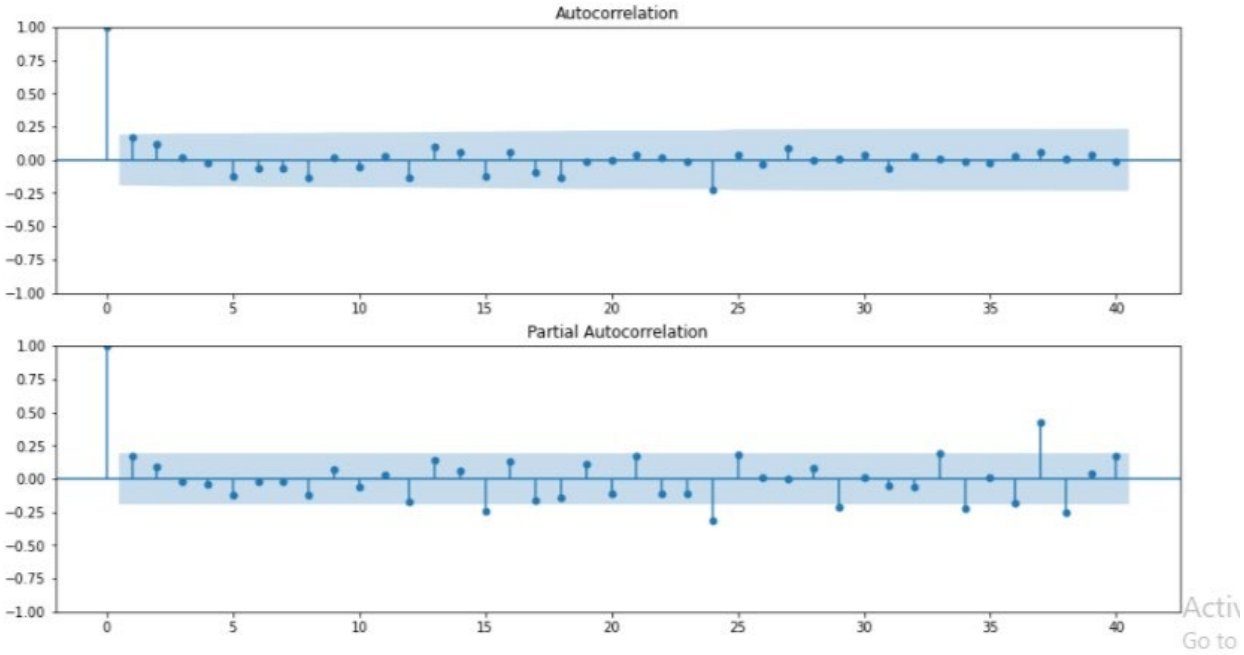
Let’s visualize ACF and PACF of residuals.



Model 2 is fine the lags are not that high (barely outside the box).

Model 3:

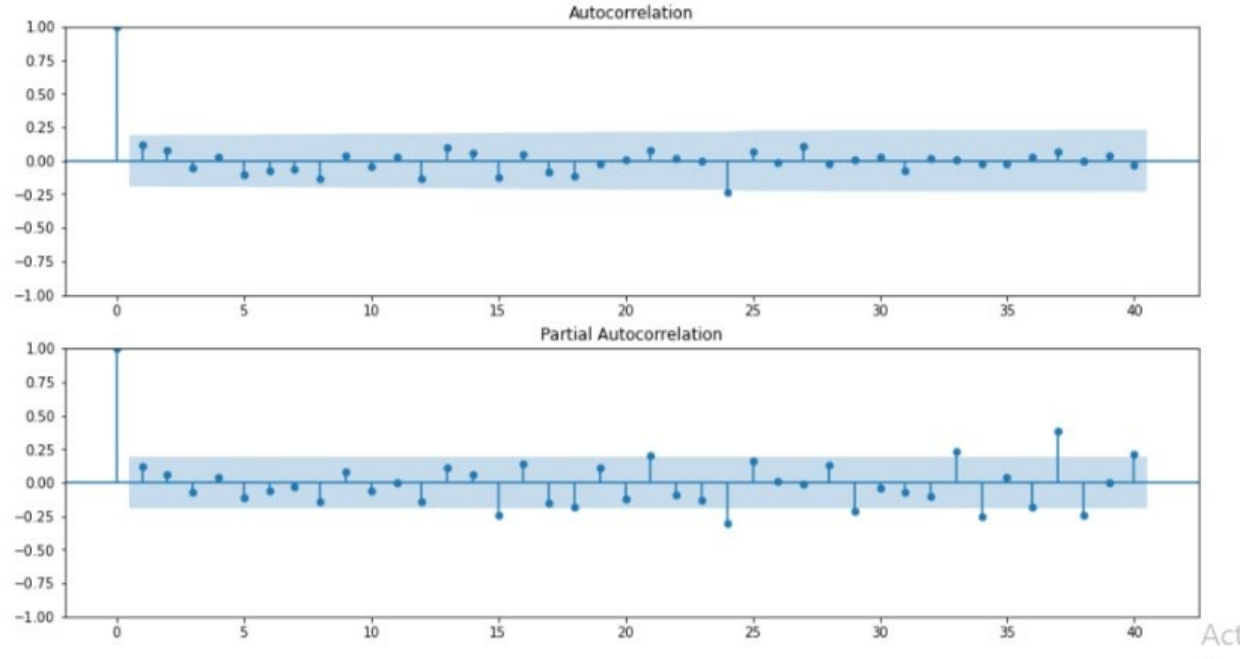
Box test is fine (p-values > 0.05)



We won't use model 3 since we have some spikes in the PACF.

Model 4:

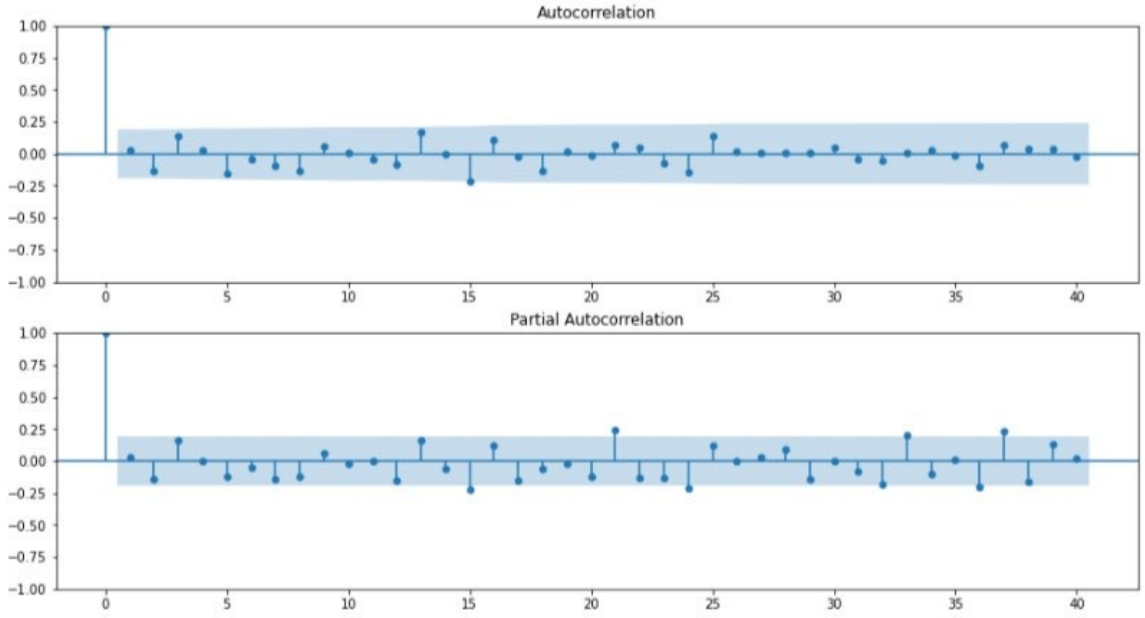
Box test results were good



We won't use model 4 since we have some spikes in the PACF.

Model 5:

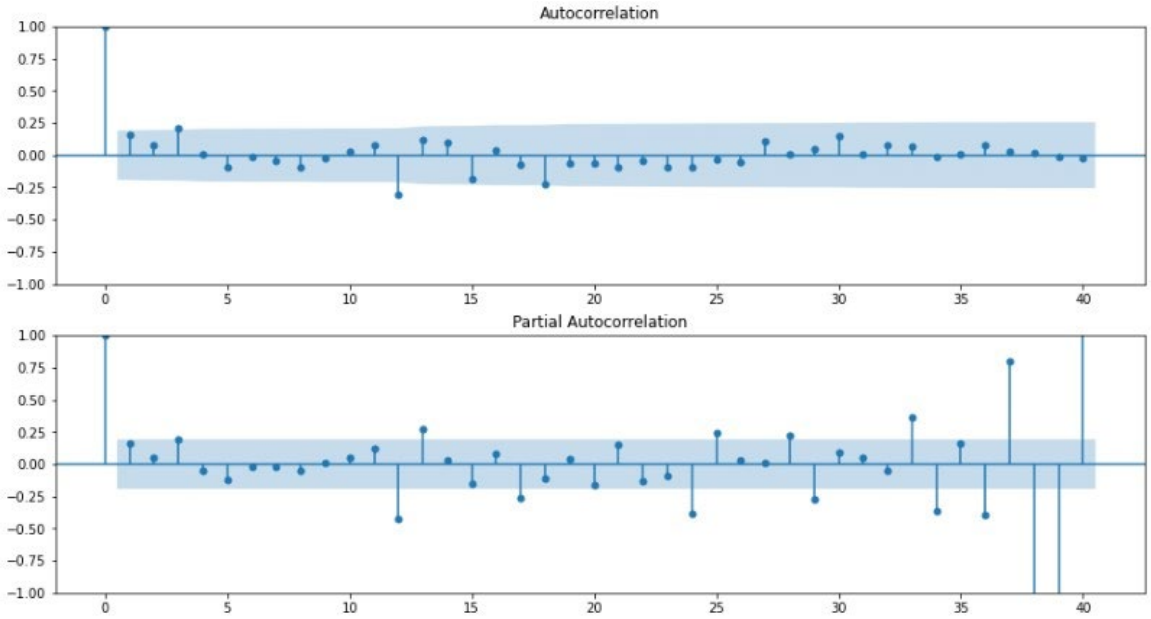
Again box test results are fine. Let's look at the plots.



We have some small spikes outside the blue box but not a big deal. We will keep model 5 for now.

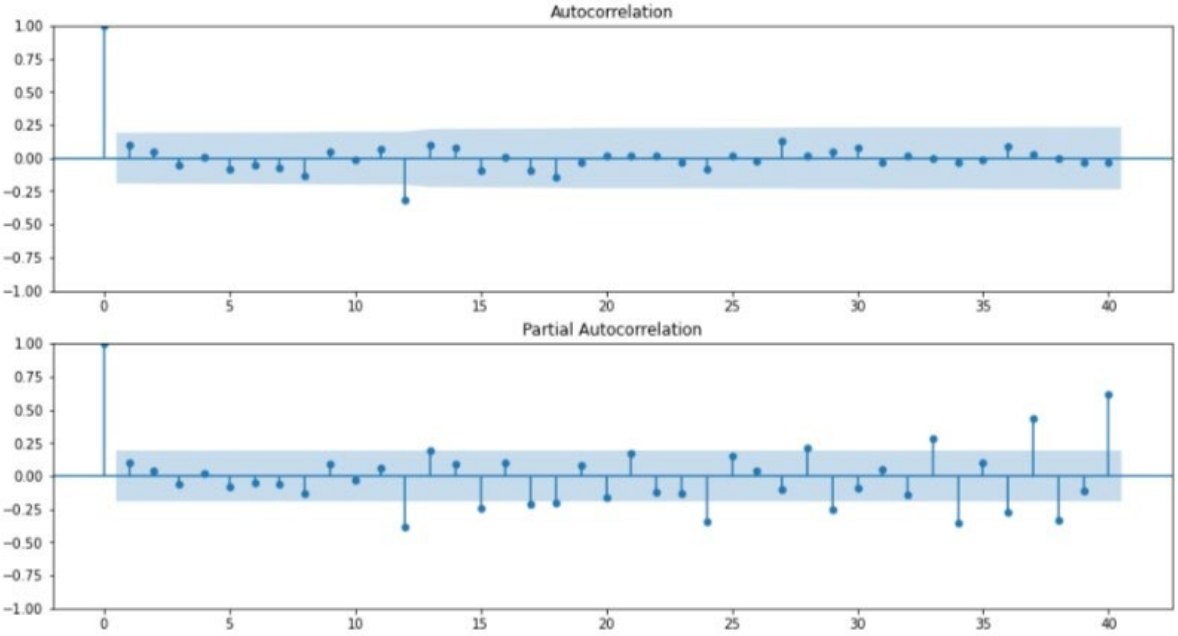
Model 6:

Box test results are not that great, we have p values lower than 0.05, we can also see from the PACF that we have very high spikes.



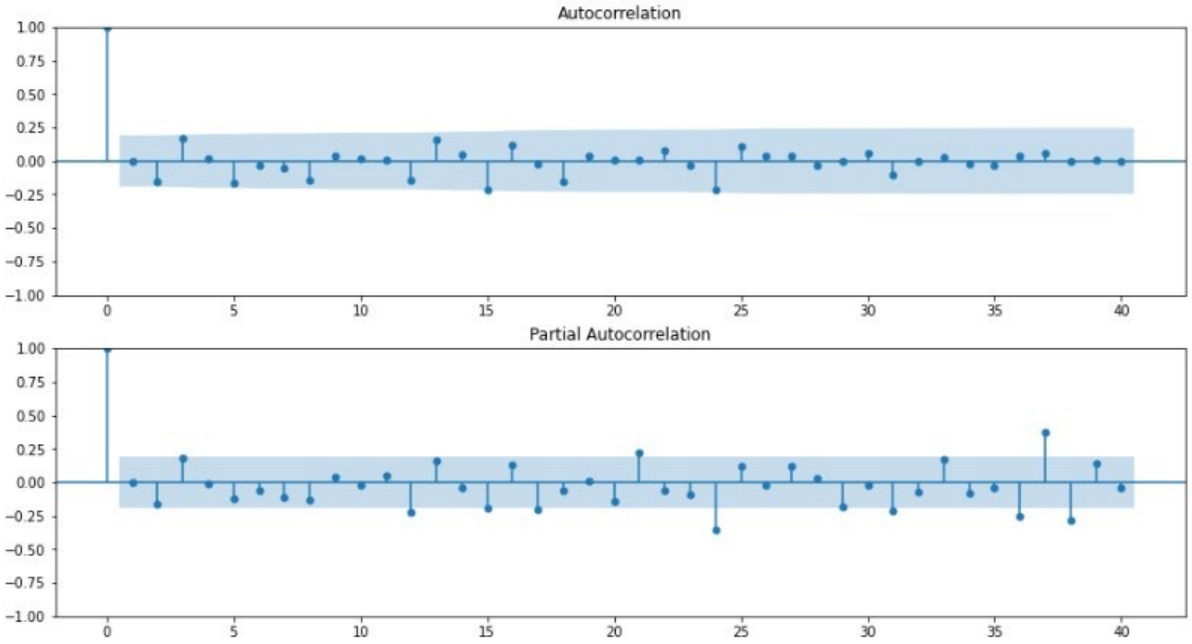
We won't use model 6 since we have high spikes in the PACF.

Model 7:



We won't use model 7 since we have high spikes in the PACF.

Model 8:



We won't use model 8 since we have high spikes in the PACF.

So we end up with two models: model 2 and model 5. (All assumptions satisfied for both models)

Model 2: (2,1,0,2,1,0,12), AIC = 419, RMSE = 2.36

Model 5: (1,1,0,2,1,0,12), AIC = 436, RMSE = 2.57

We will choose model 2 since it has better metrics than model 5.

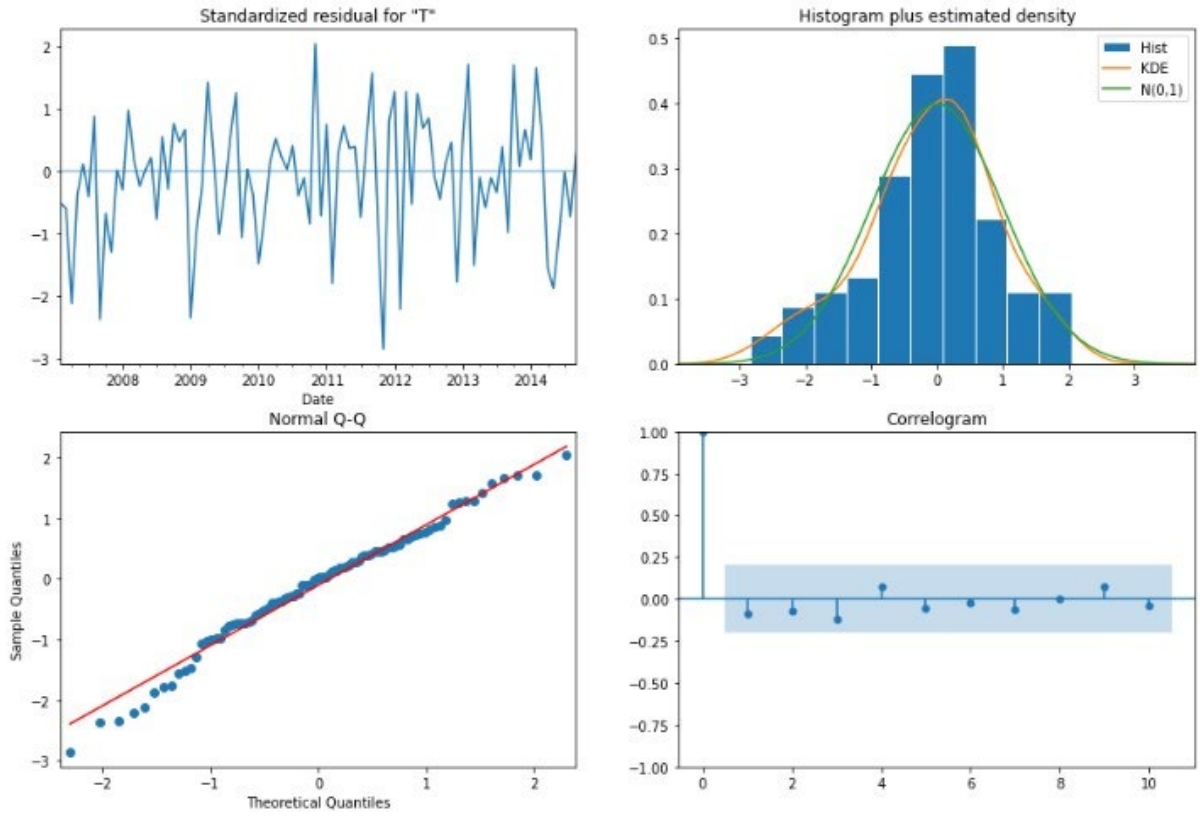
That's the summary of mode 2:

Dep. Variable:	Temperature (C)	No. Observations:	105
Model:	SARIMAX(2, 1, 0)x(2, 1, 0, 12)	Log Likelihood	-204.556
Date:	Sat, 11 Nov 2023	AIC	419.113
Time:	22:47:21	BIC	431.722
Sample:	01-31-2006	HQIC	424.202
	- 09-30-2014		
Covariance Type:	opg		

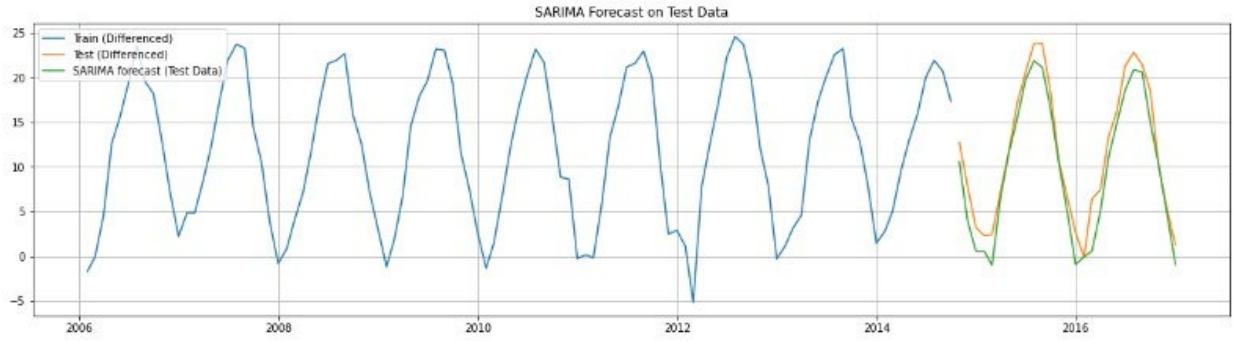
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.6380	0.090	-7.106	0.000	-0.814	-0.462
ar.L2	-0.4337	0.083	-5.197	0.000	-0.597	-0.270
ar.S.L12	-0.6137	0.096	-6.387	0.000	-0.802	-0.425
ar.S.L24	-0.4188	0.112	-3.754	0.000	-0.638	-0.200
sigma2	4.5947	0.685	6.706	0.000	3.252	5.938

Ljung-Box (L1) (Q):	0.72	Jarque-Bera (JB):	2.55
Prob(Q):	0.40	Prob(JB):	0.28
Heteroskedasticity (H):	1.07	Skew:	-0.41
Prob(H) (two-sided):	0.86	Kurtosis:	3.07

Plots of residuals:



And that's the fit of model 2



It fits great. It's able to catch the seasonal pattern.

Conclusion:

The best model using SARIMA after doing trend differencing and seasonal differencing is model 2 (p, d, q, P, D, Q, m = 2,1,0,2,1,0,12)

With an AIC = 419 and RMSE = 2.36

We also got great results using Holtwinters (triple exp smoothing) without trend and with seasonality (RMSE = 1.79 and an AIC = 188)

So those were our best 2 models, both had good fits but the better one was using Holtwinters since it has lower RMSE and lower AIC, plus it's less complex.

Thank You!