Time Series Analysis: Forecasting Temperature Based On Air Quality

DATS 6450 : Time Series Analysis & Modeling

Instructor : Dr. Reza Jafari

Student: Arathi Nair

Date: December 8th, 2022

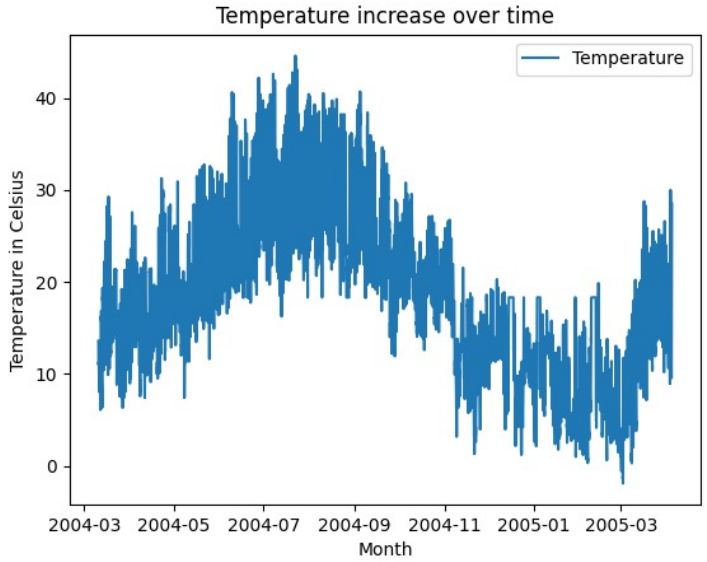
Introduction

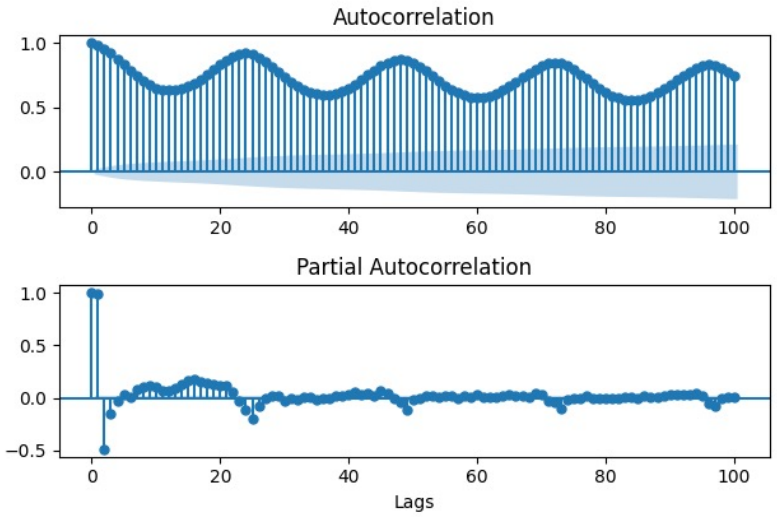
Dataset: Air Quality Data Set (UCI Machine Learning Repository)
Description: 9358 instances of hourly averaged responses from 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device located in a significantly polluted area within an Italian city.

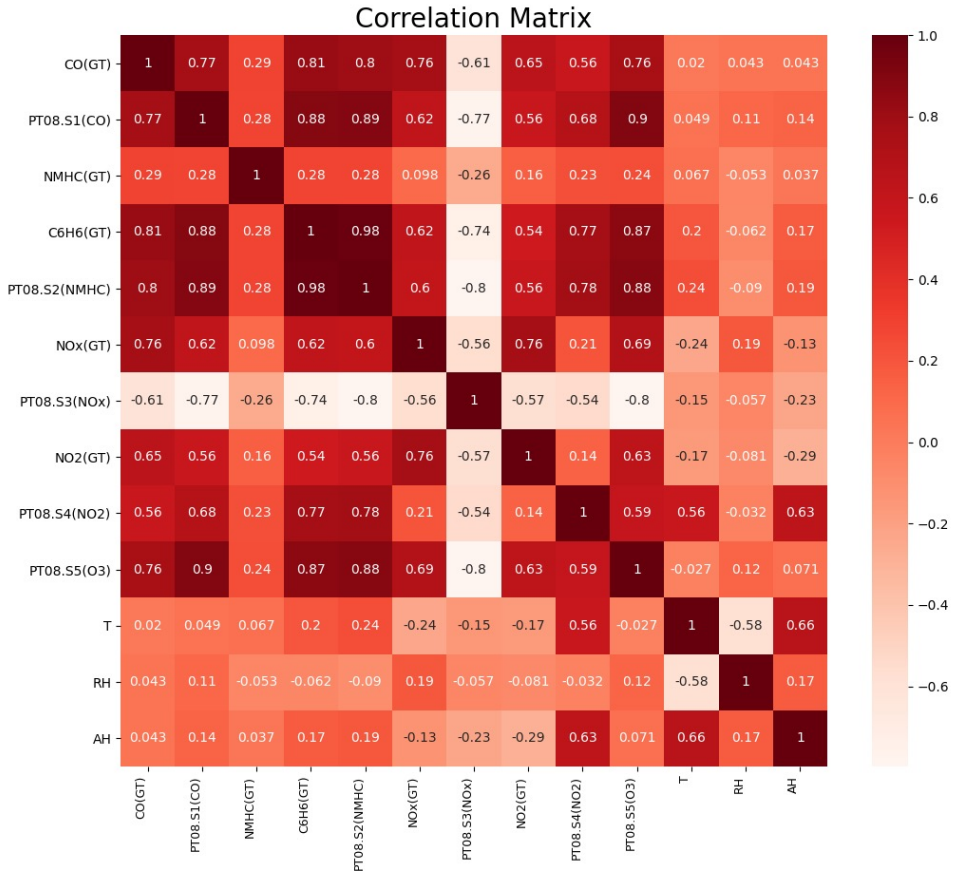
ATTRIBUTE	DESCRIPTION
DATE	Date (DD/MM/YYYY)
TIME	Time (HH.MM.SS)
CO(GT)	True hourly averaged concentration CO in mg/m^3
PT08.S1(CO)	PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
NMHC(GT)	True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3
C6H6(GT)	True hourly averaged Benzene concentration in microg/m^3
PT08.S2(NMHC)	PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
NOX(GT)	True hourly averaged NOx concentration in ppb(reference analyzer)

ATTRIBUTE	DESCRIPTION
PT08.S3(NOx)	PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
NO2(GT)	True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
PT08.S4(NO2)	PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
PT08.S5(O3)	PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
T	Temperature in °C
RH	Relative Humidity (%)
AH	AH Absolute Humidity

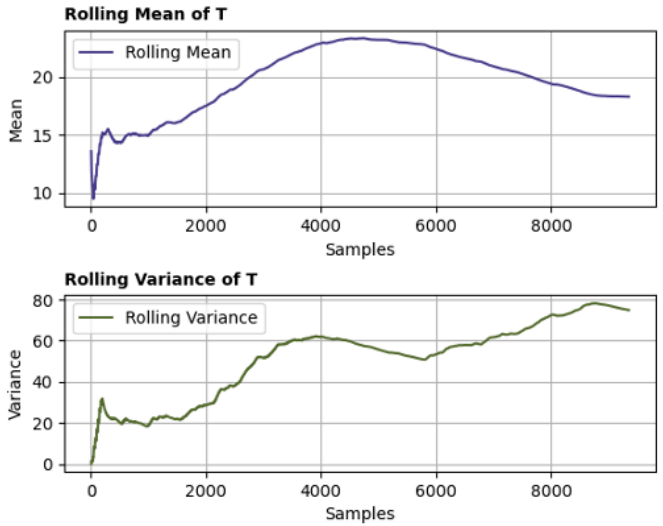
Description

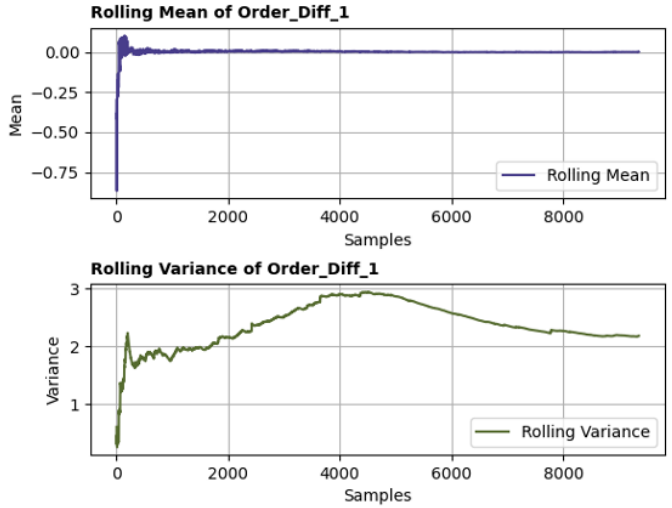

Temperature increase over time
Temperature
Temperature in Celsius
Month
40
30
20
10
0
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03


Autocorrelation
Partial Autocorrelation
Lags
1.0
0.5
0.0
−0.5
0
20
40
60
80
100


Correlation Matrix
CO(GT) 1 0.77 0.29 0.81 0.8 0.76 -0.61 0.65 0.56 0.76 0.02 0.043 0.043
PT08.S1(CO) 0.77 1 0.28 0.88 0.89 0.62 -0.77 0.56 0.68 0.9 0.049 0.11 0.14
NMHC(GT) 0.29 0.28 1 0.28 0.28 0.098 -0.26 0.16 0.23 0.24 0.067 -0.053 0.037
C6H6(GT) 0.81 0.88 0.28 1 0.98 0.62 -0.74 0.54 0.77 0.87 0.2 -0.062 0.17
PT08.S2(NMHC) 0.8 0.89 0.28 0.98 1 0.6 -0.8 0.56 0.78 0.88 0.24 -0.09 0.19
NOx(GT) 0.76 0.62 0.098 0.62 0.6 1 -0.56 0.76 0.21 0.69 -0.24 0.19 -0.13
PT08.S3(NOx) -0.61 -0.77 -0.26 -0.74 -0.8 -0.56 1 -0.57 -0.54 -0.8 -0.15 -0.057 -0.23
NO2(GT) 0.65 0.56 0.16 0.54 0.56 0.76 -0.57 1 0.14 0.63 -0.17 -0.081 -0.29
PT08.S4(NO2) 0.56 0.68 0.23 0.77 0.78 0.21 -0.54 0.14 1 0.59 0.56 -0.032 0.63
PT08.S5(O3) 0.76 0.9 0.24 0.87 0.88 0.69 -0.8 0.63 0.59 1 -0.027 0.12 0.071
T 0.02 0.049 0.067 0.2 0.24 -0.24 -0.15 -0.17 0.56 -0.027 1 -0.58 0.66
RH 0.043 0.11 -0.053 -0.062 -0.09 0.19 -0.057 -0.081 -0.032 0.12 -0.58 1 0.17
AH 0.043 0.14 0.037 0.17 0.19 -0.13 -0.23 -0.29 0.63 0.071 0.66 0.17 1
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Stationarity

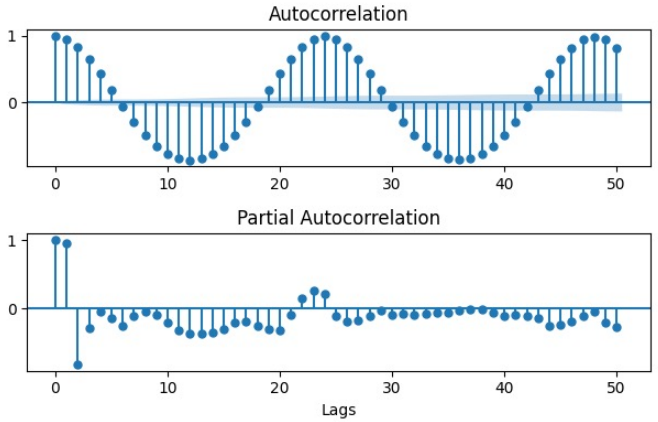




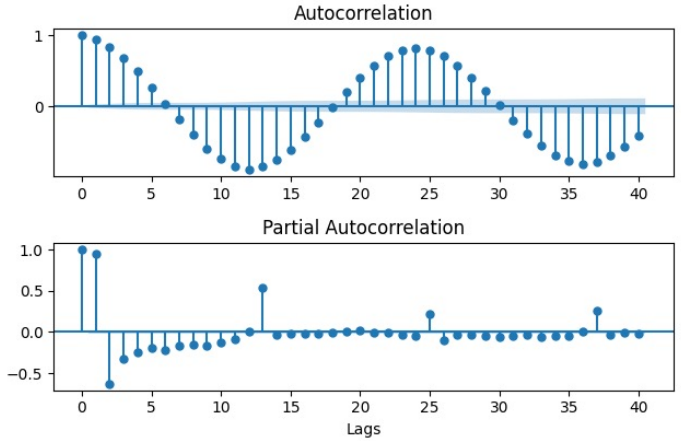

```
ADF Statistic for T: -3.621766
p-value: 0.005358
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
Results of KPSS Test for T:
Test Statistic            7.205413
p-value                   0.010000
Lags Used                55.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
```

```
ADF Statistic for Order_Diff_1: -18.451598
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
Results of KPSS Test for Order_Diff_1:
Test Statistic           0.002747
p-value                  0.100000
Lags Used                9.000000
Critical Value (10%)     0.347000
Critical Value (5%)      0.463000
Critical Value (2.5%)    0.574000
Critical Value (1%)      0.739000
```

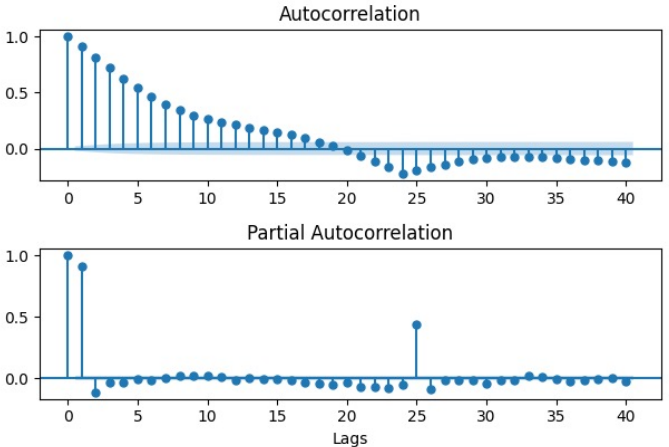
First Order Differencing

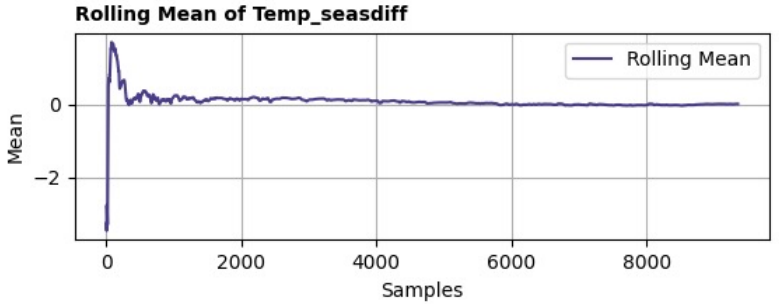


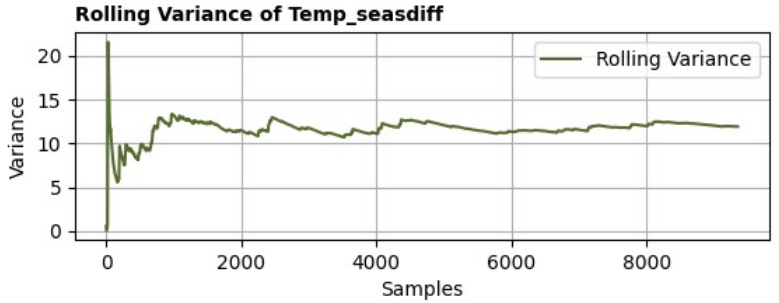
Seasonal Differencing 12 hours



Seasonal Differencing 24 hours

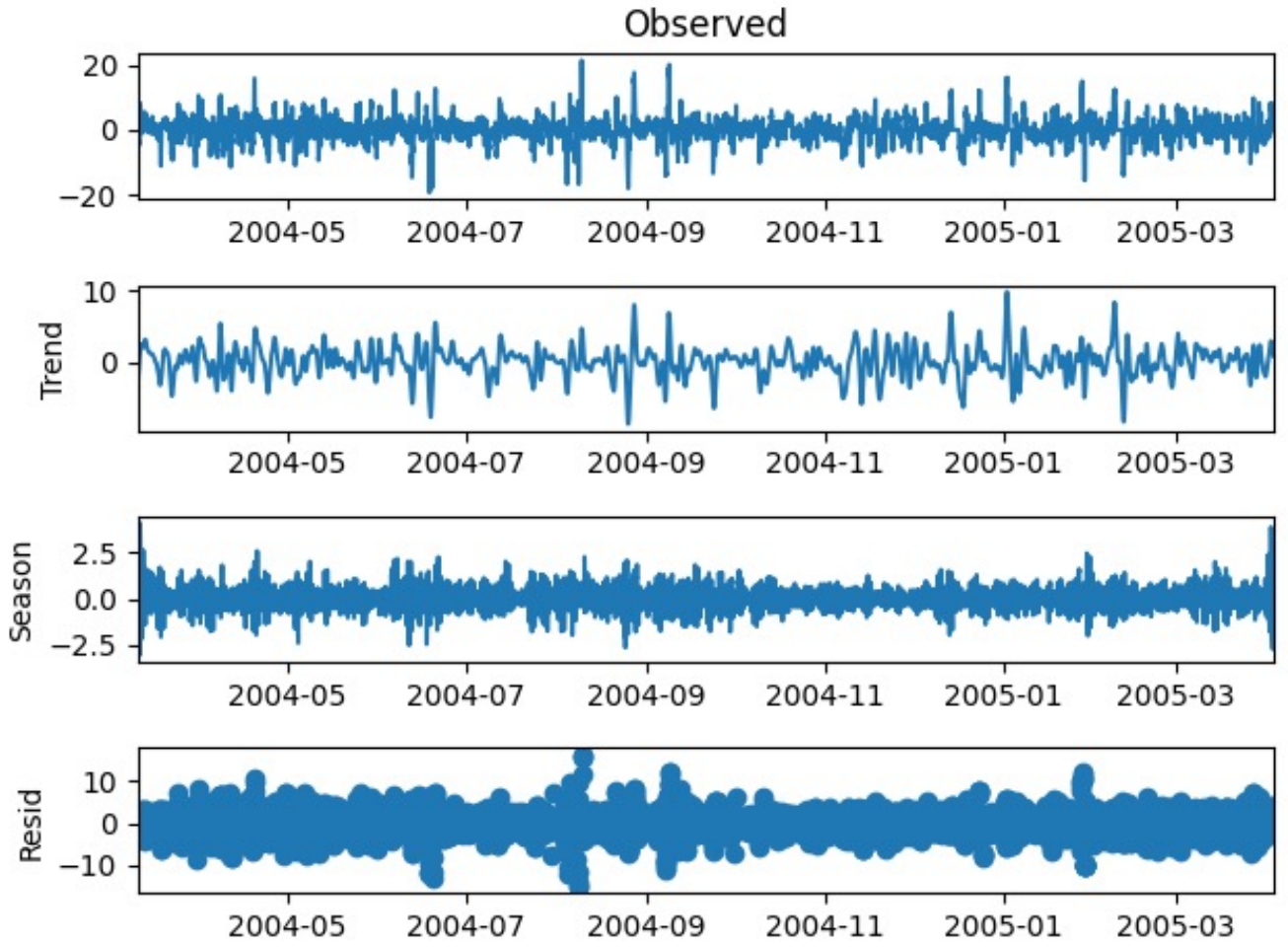


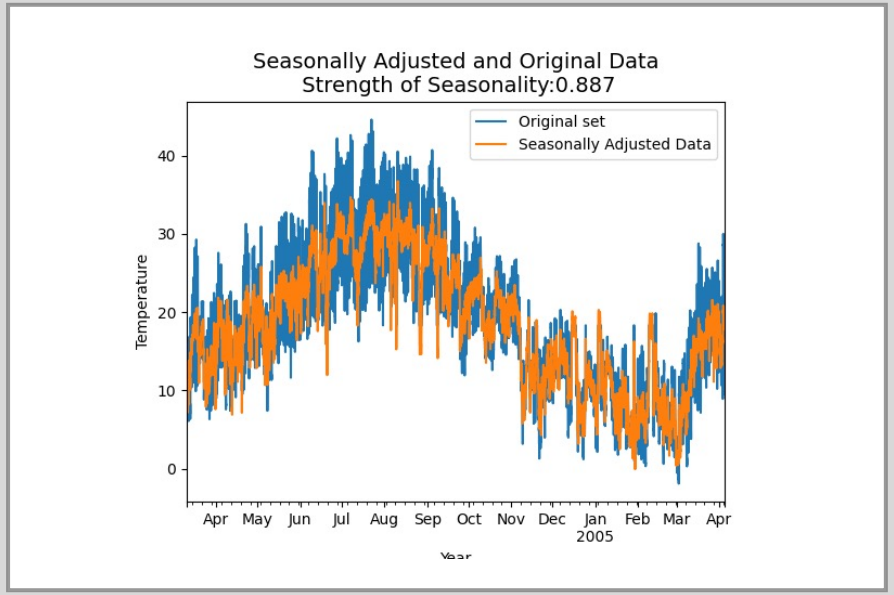


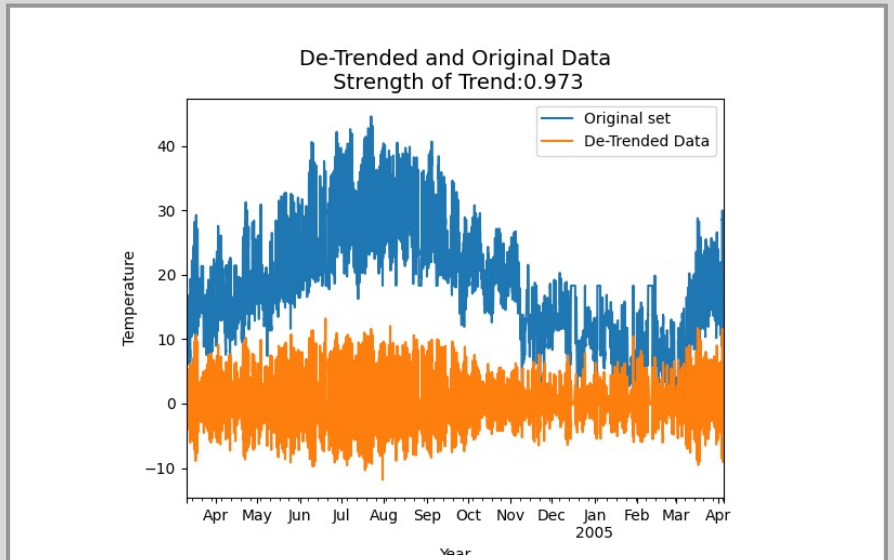



```
ADF Statistic for Temp_seasdiff: -15.993908
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
Results of KPSS Test for Temp_seasdiff:
Test Statistic            0.048617
p-value                   0.100000
Lags Used                51.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

Time Series Decomposition


Observed
20
0
−20
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
Trend
10
0
Season
2.5
0.0
−2.5
Resid
10
0
−10


Seasonally Adjusted and Original Data
Strength of Seasonality:0.887
Original set
Seasonally Adjusted Data
Temperature
40
30
20
10
0
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
2005
Year


De-Trended and Original Data
Strength of Trend:0.973
Original set
De-Trended Data
Temperature
40
30
20
10
0
−10
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
2005
Year

Feature Selection

CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOX(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	RH	AH	AIC	BIC	Adj R^2
1	1	1	1	1	1	1	1	1	1	1	1	3.215e+04	3.224e+04	0.934
1	1	0	1	1	1	1	1	1	1	1	1	3.215e+04	3.223e+04	0.934
1	0	0	1	1	1	1	1	1	1	1	1	3.215e+04	3.222e+04	0.934
1	0	0	1	1	1	0	1	1	1	1	1	3.215e+04	3.221e+04	0.934
1	0	0	1	1	0	0	1	1	1	1	1	3.215e+04	3.221e+04	0.934
0	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**4.345e+04**	**4.347e+04**	**0.959**

Multicollinearity

OLS Regression Results

Dep. Variable:	Temperature	R-squared:	0.934
Model:	OLS	Adj. R-squared:	0.934
Method:	Least Squares	F-statistic:	8769.
Date:	Sat, 20 Nov 2021	Prob (F-statistic):	0.00
Time:	12:56:59	Log-Likelihood:	-16062.
No. Observations:	7485	AIC:	3.215e+04
Df Residuals:	7472	BIC:	3.224e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.6912	0.677	23.161	0.000	14.363	17.019
CO(GT)	-0.1883	0.042	-4.434	0.000	-0.272	-0.105
PT08.S1(CO)	0.0002	0.000	0.642	0.521	-0.000	0.001
NMHC(GT)	-0.0002	0.000	-0.601	0.548	-0.001	0.001
C6H6(GT)	-0.1846	0.022	-8.352	0.000	-0.228	-0.141
PT08.S2(NMHC)	0.0079	0.001	10.304	0.000	0.006	0.009
NOx(GT)	0.0007	0.000	2.270	0.023	9.35e-05	0.001
PT08.S3(NOx)	-0.0002	0.000	-0.728	0.466	-0.001	0.000
NO2(GT)	0.0067	0.001	5.569	0.000	0.004	0.009
PT08.S4(NO2)	0.0023	0.000	9.505	0.000	0.002	0.003
PT08.S5(O3)	-0.0034	0.000	-18.383	0.000	-0.004	-0.003
RH	-0.3523	0.002	-184.058	0.000	-0.356	-0.349
AH	13.9758	0.121	115.793	0.000	13.739	14.212

Condition Number: 72725.11054355717

OLS Regression Results

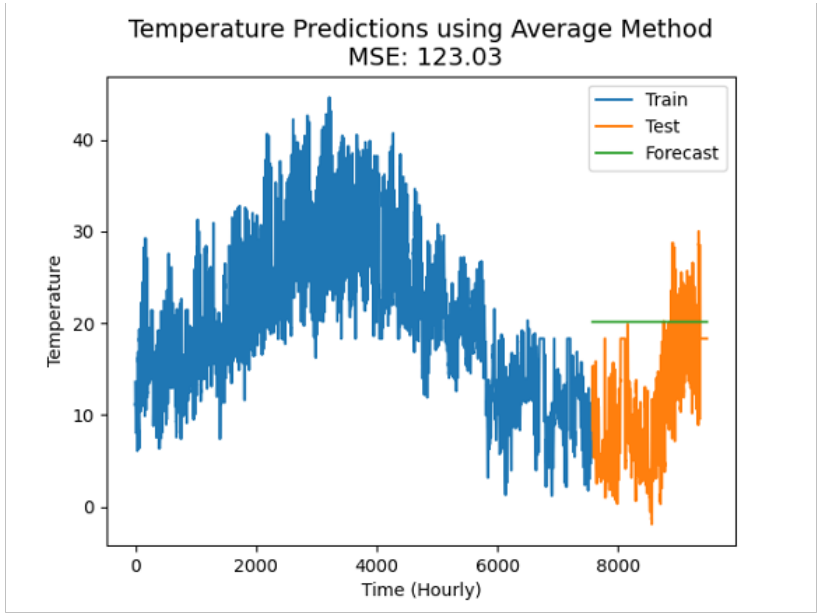
Dep. Variable:	Temperature	R-squared (uncentered):	0.959
Model:	OLS	Adj. R-squared (uncentered):	0.959
Method:	Least Squares	F-statistic:	4.408e+04
Date:	Sat, 20 Nov 2021	Prob (F-statistic):	0.00
Time:	16:56:36	Log-Likelihood:	-21719.
No. Observations:	7485	AIC:	4.345e+04
Df Residuals:	7481	BIC:	4.347e+04
Df Model:	4		
Covariance Type:	nonrobust		

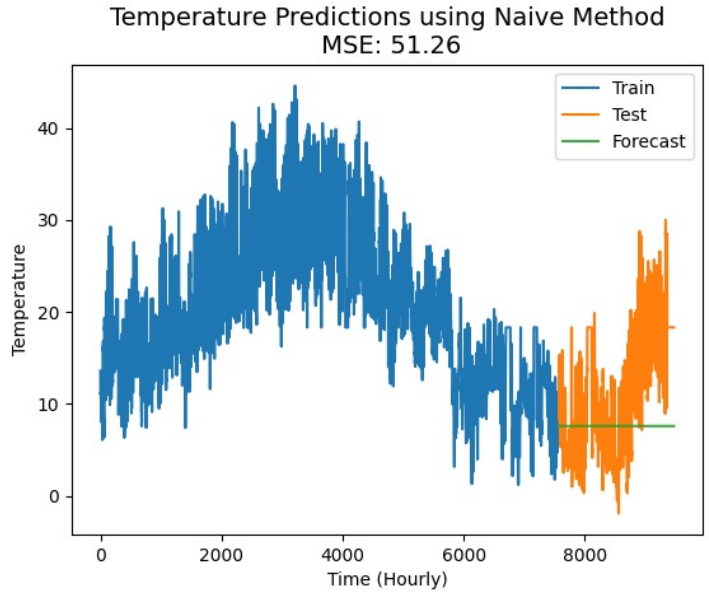
	coef	std err	t	P>|t|	[0.025	0.975]
C6H6(GT)	-0.1622	0.009	-18.339	0.000	-0.180	-0.145
NO2(GT)	0.0914	0.001	66.659	0.000	0.089	0.094
RH	-0.2703	0.003	-100.762	0.000	-0.276	-0.265
AH	22.2275	0.117	189.651	0.000	21.998	22.457

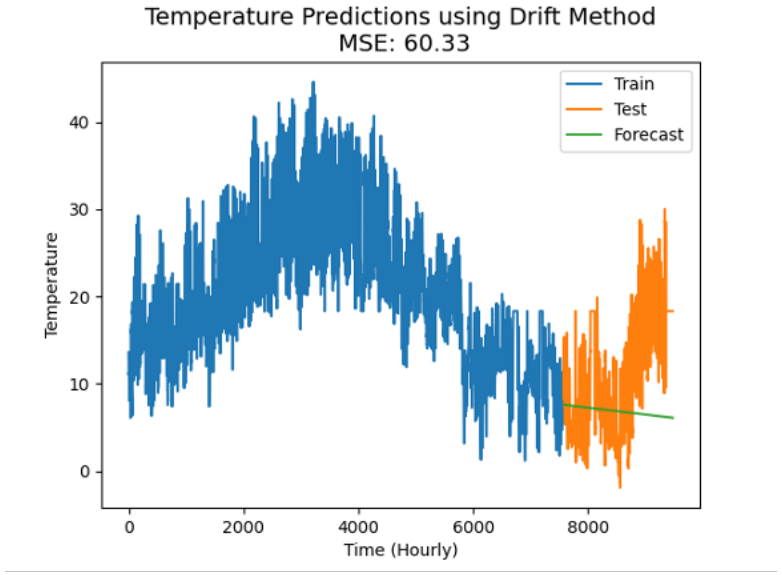
Omnibus:	243.481	Durbin-Watson:	0.259
Prob(Omnibus):	0.000	Jarque-Bera (JB):	388.145
Skew:	0.303	Prob(JB):	5.19e-85
Kurtosis:	3.936	Cond. No.	281.

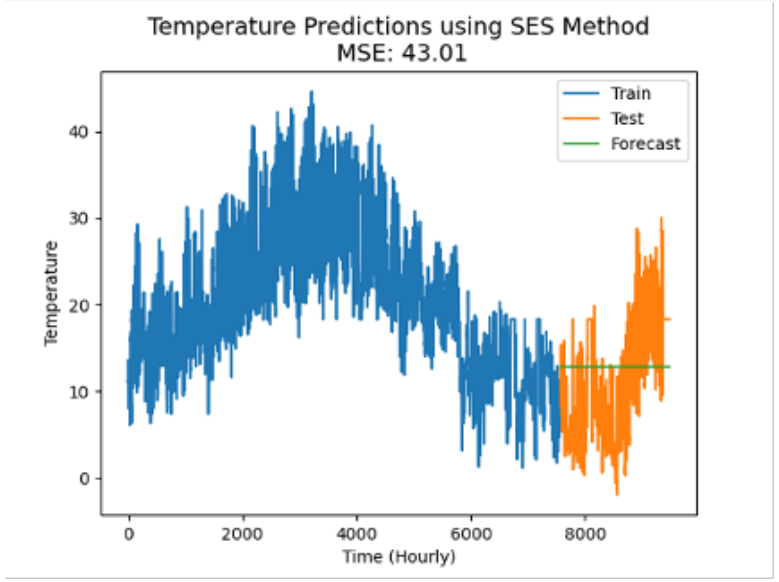
SingularValues:
[10662.43663072 2009.22267335 506.93555842 38.16176761]

Base Models

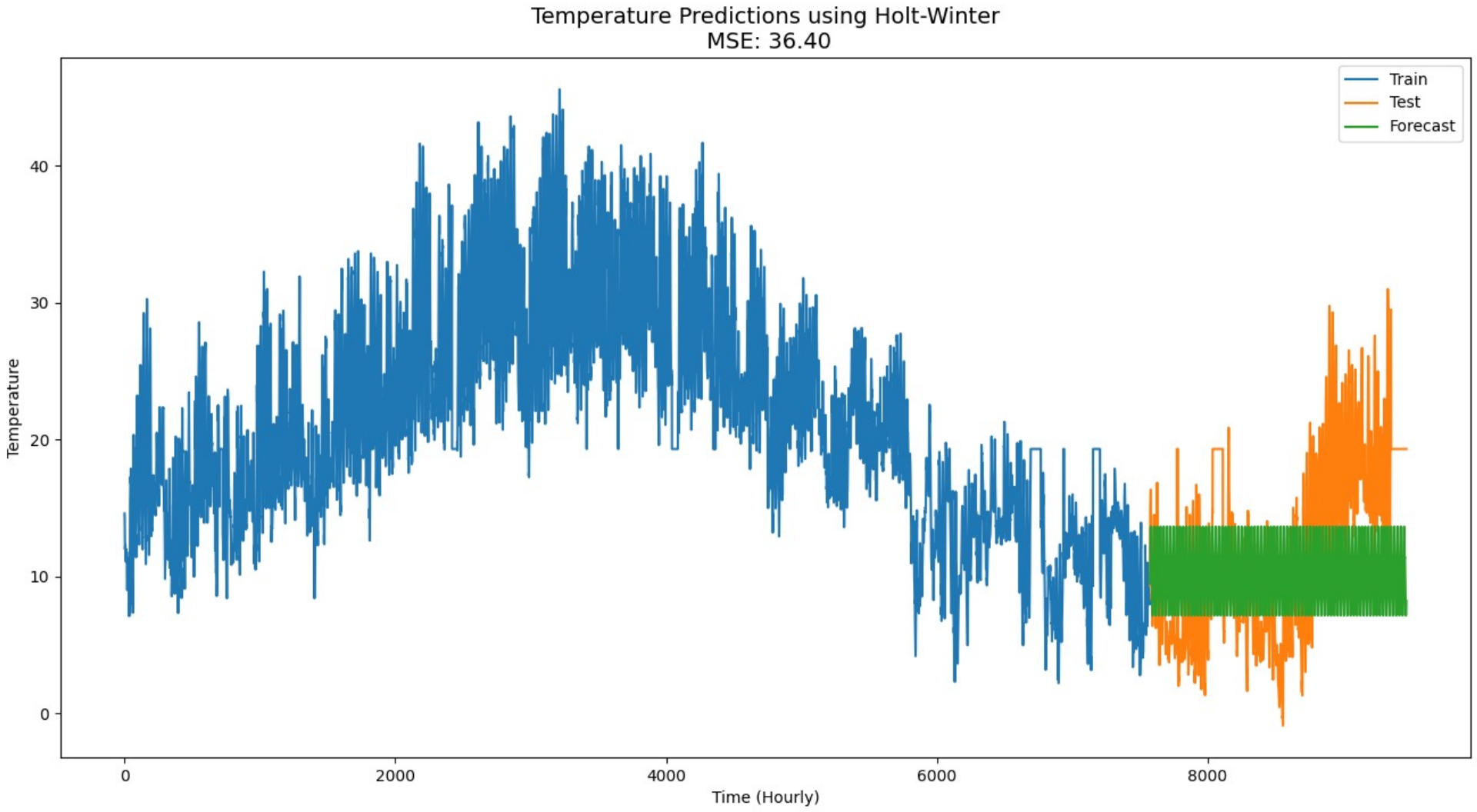

Temperature Predictions using Average Method
MSE: 123.03
Train
Test
Forecast
Temperature
Time (Hourly)
0
10
20
30
40
0
2000
4000
6000
8000


Temperature Predictions using Naive Method
MSE: 51.26
Train
Test
Forecast
Temperature
Time (Hourly)
0
10
20
30
40
0
2000
4000
6000
8000


Temperature Predictions using Drift Method
MSE: 60.33
Train
Test
Forecast
Temperature
Time (Hourly)
0
10
20
30
40
0
2000
4000
6000
8000


Temperature Predictions using SES Method
MSE: 43.01
Train
Test
Forecast
Temperature
Time (Hourly)
0
10
20
30
40
0
2000
4000
6000
8000

Holt-Winters Method



Multiple Linear Regression Model

Observation for t-test:

- ŷ = b0 + b1(C6H6(GT)) + b2 (NO2(GT)) + b3(RH) + b4(AH)
- All t-test values are significant and satisfy the criteria of p <0.05

Observation for F-test:

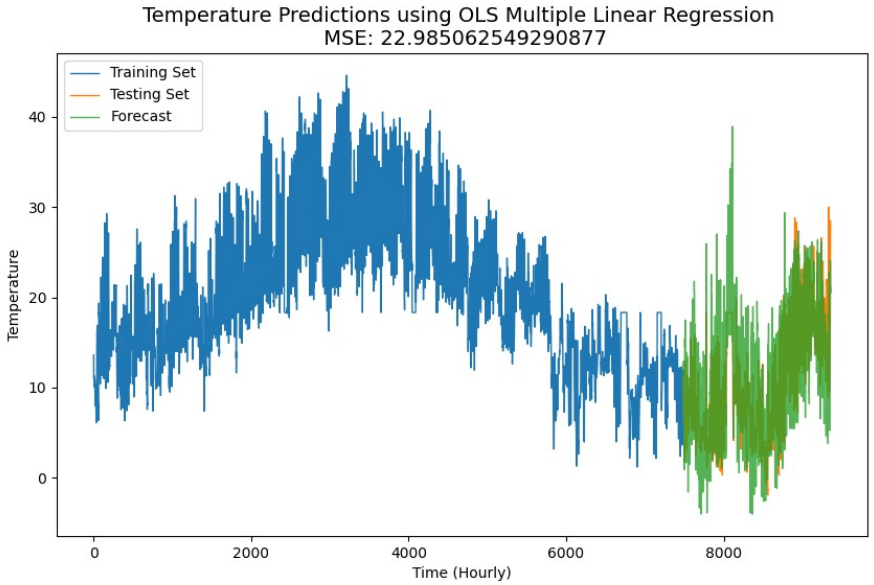
- F-test value is 4.431 e+04 which is significant with a p <0.01
- Reject the null hypothesis and conclude stepwise regression model is a better fit

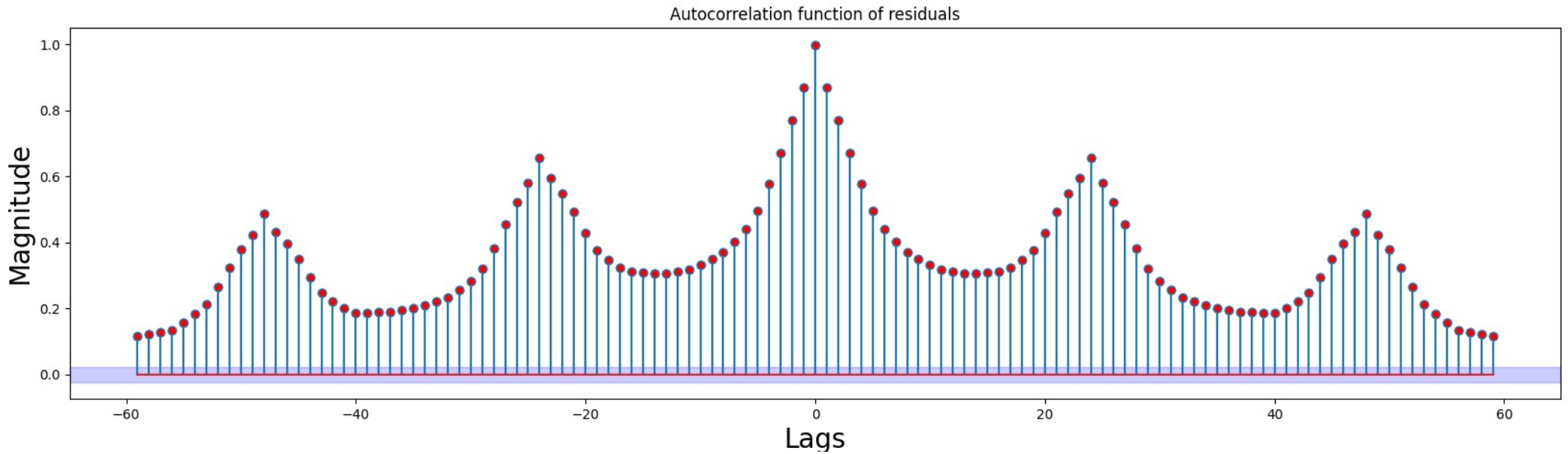
OLS Regression Results

Dep. Variable:	Temperature	R-squared (uncentered):	0.959
Model:	OLS	Adj. R-squared (uncentered):	0.959
Method:	Least Squares	F-statistic:	4.431e+04
Date:	Sun, 05 Dec 2021	Prob (F-statistic):	0.00
Time:	09:27:26	Log-Likelihood:	-21968.
No. Observations:	7576	AIC:	4.394e+04
Df Residuals:	7572	BIC:	4.397e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
C6H6(GT)	-0.1628	0.009	-18.563	0.000	-0.180	-0.146
NO2(GT)	0.0903	0.001	66.767	0.000	0.088	0.093
RH	-0.2704	0.003	-101.784	0.000	-0.276	-0.265
AH	22.3269	0.115	193.797	0.000	22.101	22.553

Omnibus:	248.301	Durbin-Watson:	0.254
Prob(Omnibus):	0.000	Jarque-Bera (JB):	395.865
Skew:	0.305	Prob(JB):	1.09e-86
Kurtosis:	3.939	Cond. No.	279.

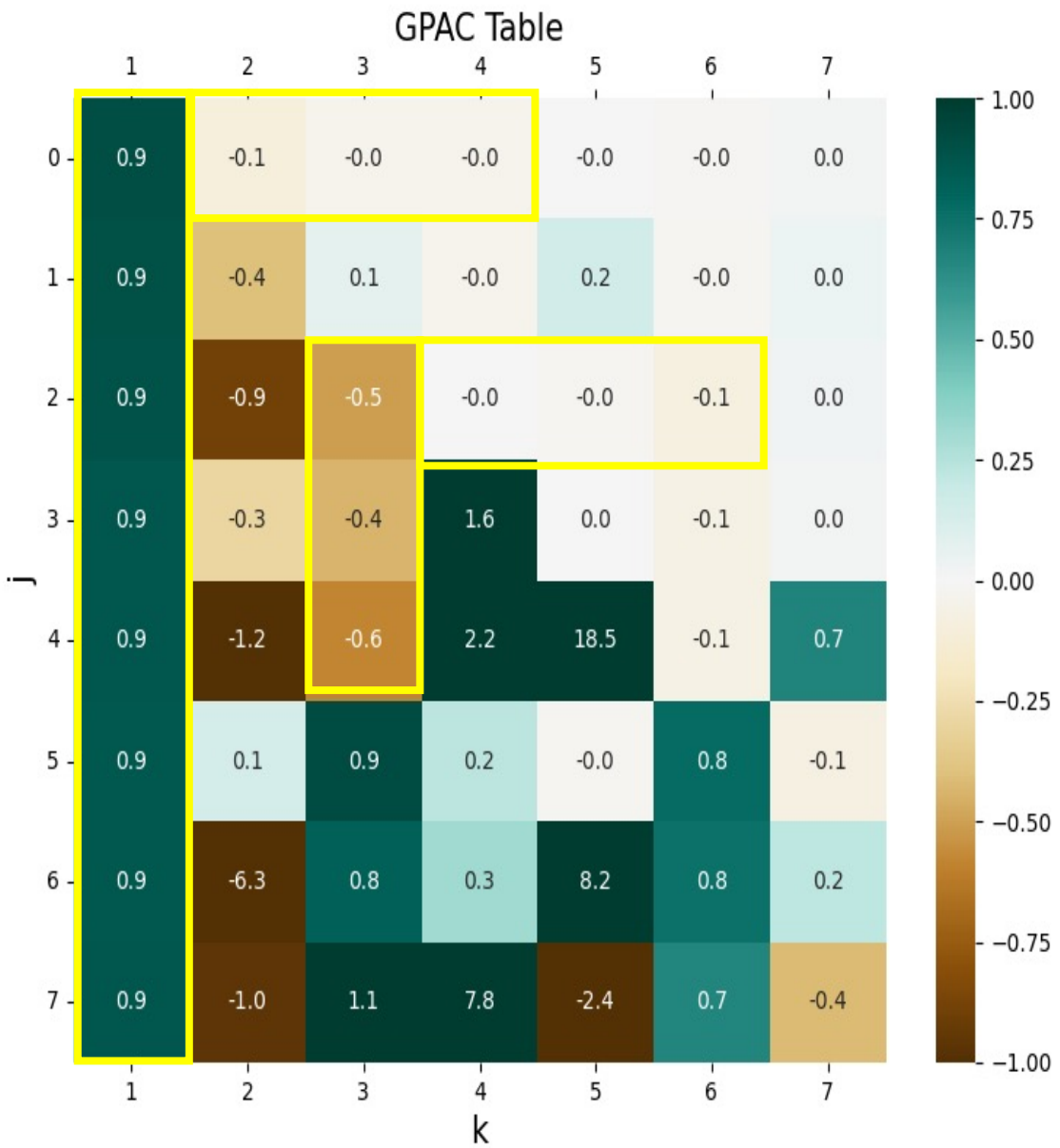


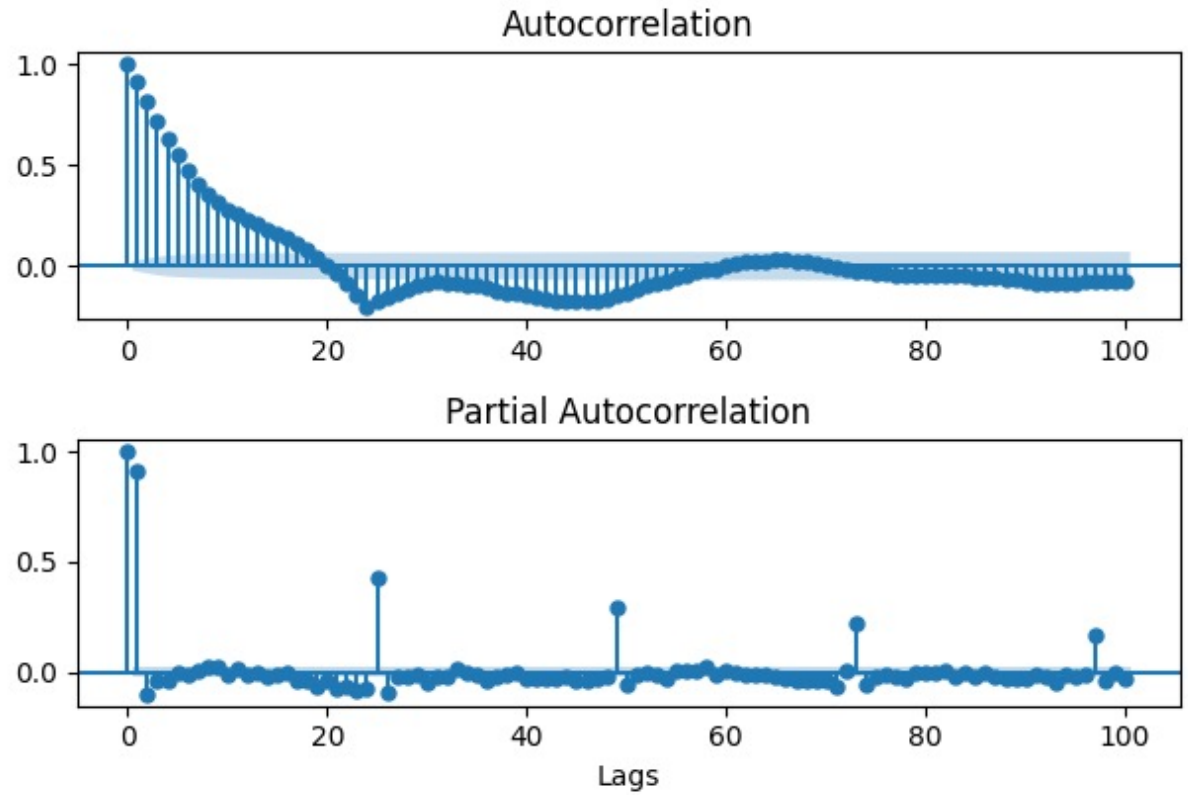



```
The Residual errors are NOT RANDOM for this model

The Critical value:  7855.107012083004
The Q value:  [16169.37832241]
The P value:  [0.]
The variance of the residuals are:  17
The mean of the residuals are:  -1
```

Model Order Determination





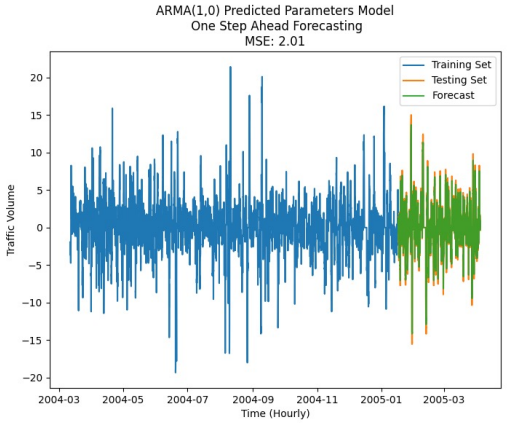
Estimated model parameters:

- ARMA(1,0)
- ARMA(3,2)
- SARIMA(2,0,0)(2,0,0)24

Levenberg Marquardt Algorithm Model Estimation

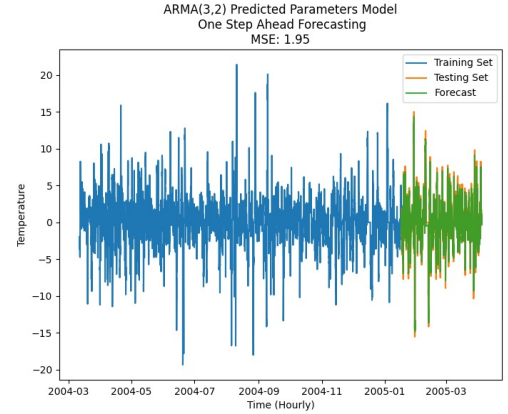
ARMA(1,0)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0041	0.184	-0.022	0.982	-0.364	0.356
ar.L1	0.9112	0.003	313.486	0.000	0.905	0.917
sigma2	2.0144	0.010	194.609	0.000	1.994	2.035



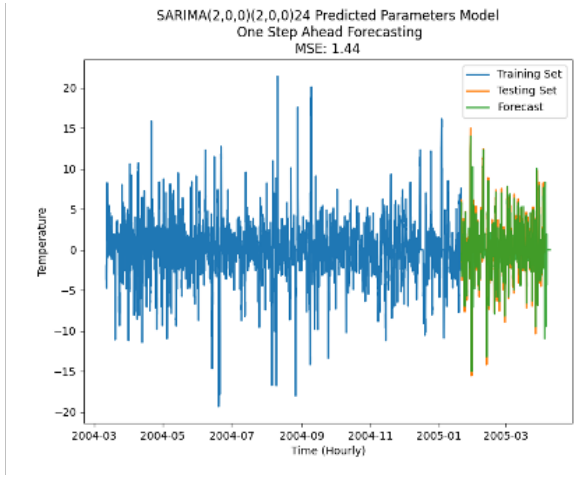
ARMA(3,2)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0082	0.153	-0.053	0.958	-0.309	0.292
ar.L1	0.5443	0.071	7.634	0.000	0.405	0.684
ar.L2	0.8801	0.005	175.749	0.000	0.870	0.890
ar.L3	-0.5253	0.061	-8.583	0.000	-0.645	-0.405
ma.L1	0.4811	0.074	6.501	0.000	0.336	0.626
ma.L2	-0.5188	0.074	-7.024	0.000	-0.664	-0.374
sigma2	1.9446	0.012	159.399	0.000	1.921	1.969



SARIMA(2,0,0)(2,0,0)24

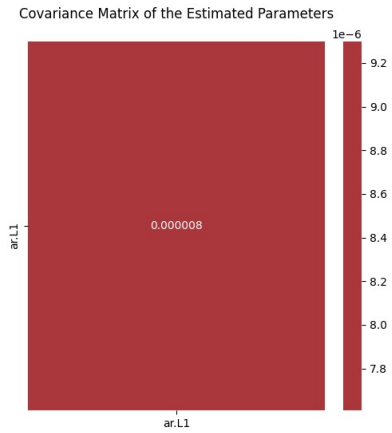
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0047	0.107	-0.044	0.965	-0.215	0.206
ar.L1	1.0609	0.008	135.298	0.000	1.046	1.076
ar.L2	-0.1301	0.008	-16.451	0.000	-0.146	-0.115
ar.S.L24	-0.6019	0.005	-111.320	0.000	-0.612	-0.591
ar.S.L48	-0.3002	0.006	-53.102	0.000	-0.311	-0.289



Diagnostic Analysis

Model	Q-Score	Q-Critical	Result	Mean of the Residuals	Biased/ Unbiased	Variance of error	Variance of residual errors	Variance of forecast error
ARMA(1,0)	1693.91	41.63	Not Random	0.00	No	1.817	2.016	22.157
ARMA(3,2)	1542.27	36.19	Not Random	0.00	No	1262.178	1.948	22.228
SARIMA(2,0,0)(2,0,0)24	143.94	40.28	Not Random	0.00	No	2.475	1.438	22.762

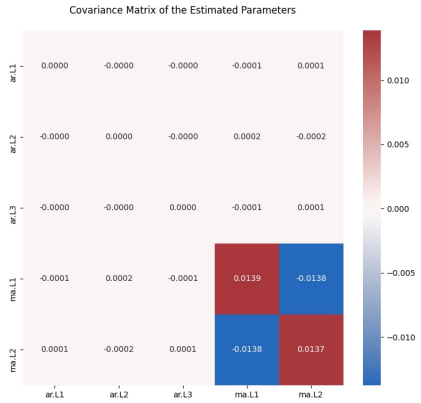
ARMA(1,0)



Zero-Pole Cancellation

$$\frac{(1-0)}{(1+0.9112)}$$

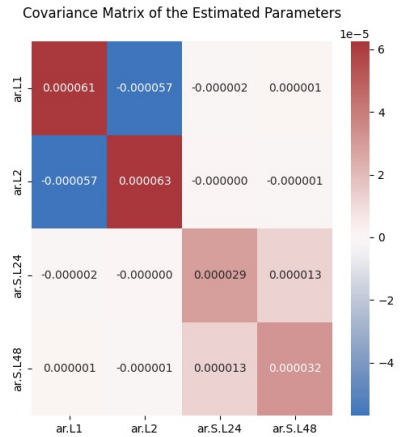
ARMA(3,2)



Zero-Pole Cancellation

$$\frac{(1-0.0378)(1+0.2496)}{(1+0.8992)(1+0.2692)(1-0.2517)}$$

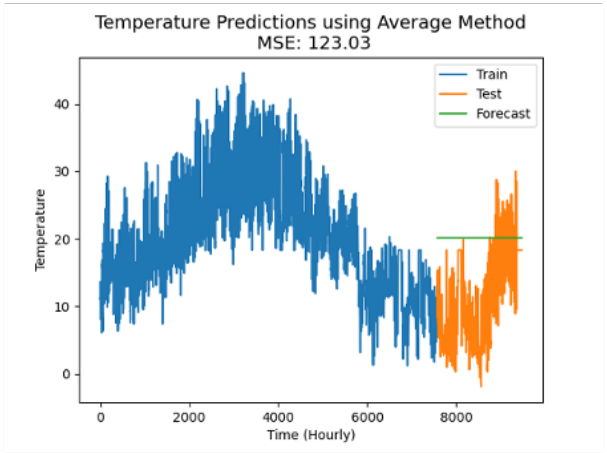
SARIMA(2,0,0)(2,0,0)24

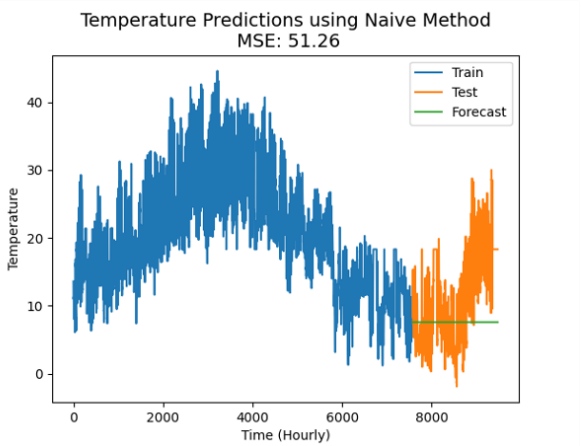


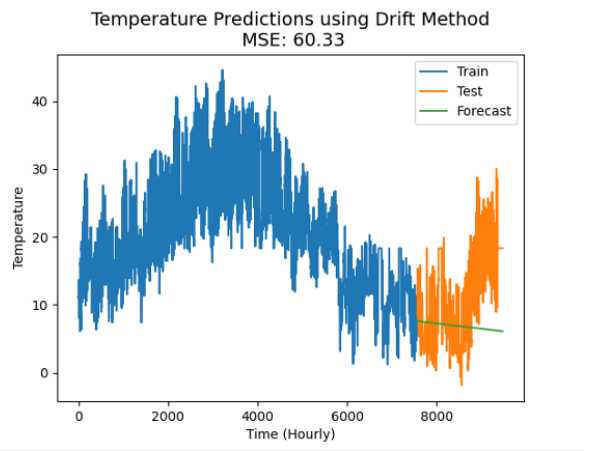
Zero-Pole Cancellation

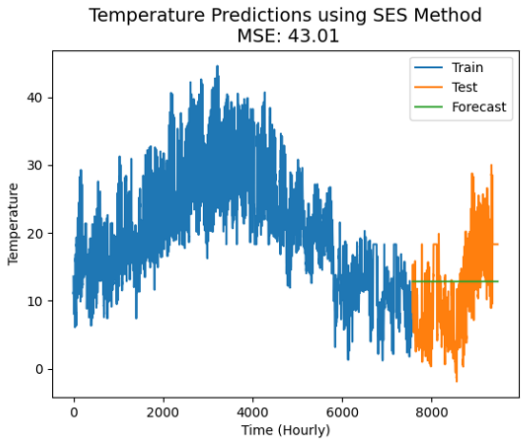
$$\frac{(1-0)}{(1+0.0287)(1+0.7970)(1+0.2927)(1+0.0249)}$$

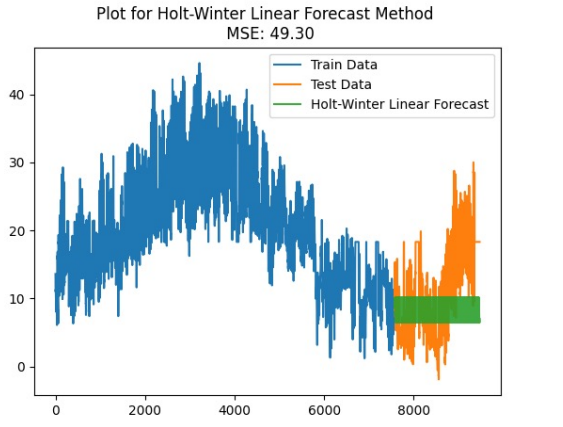
Final Model Selection

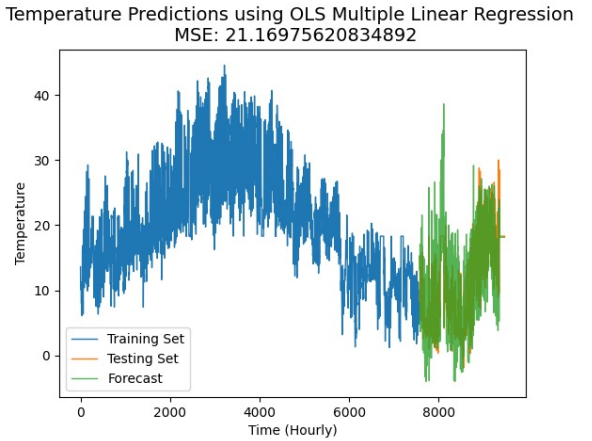


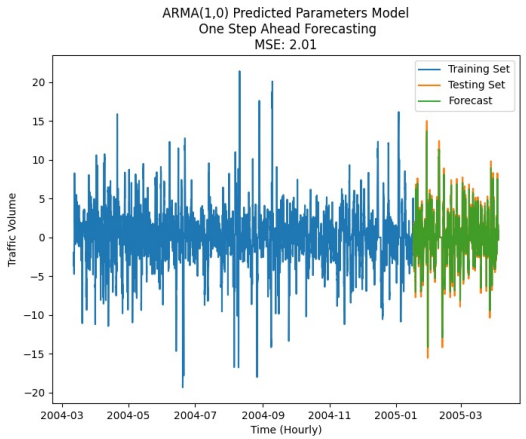


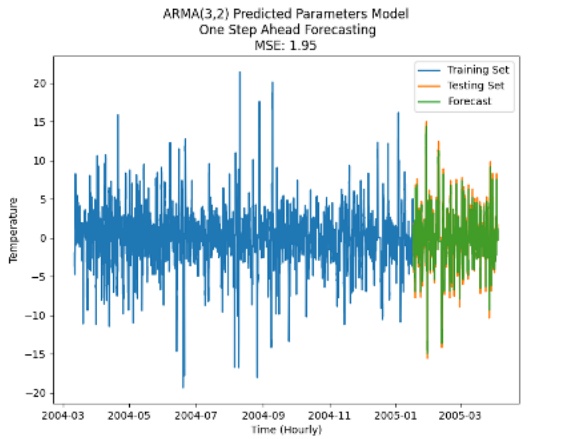


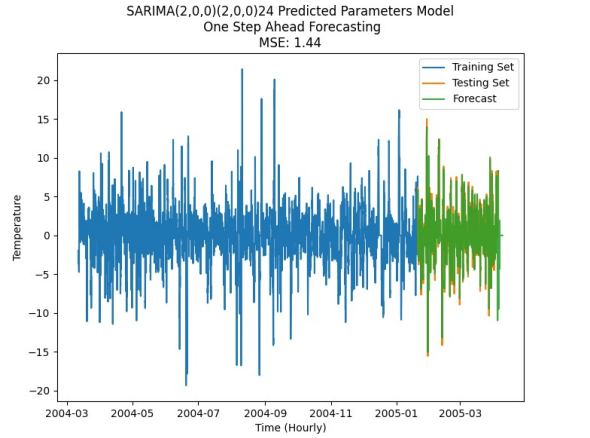






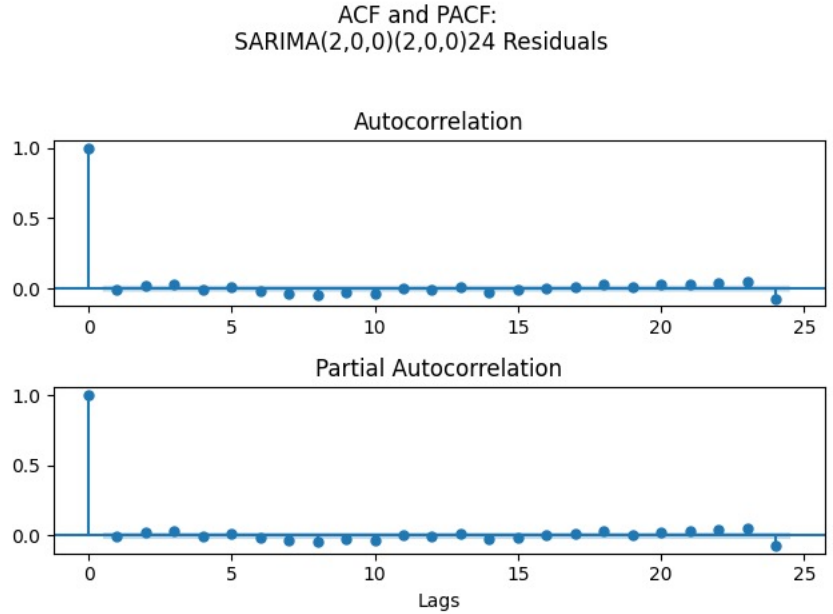






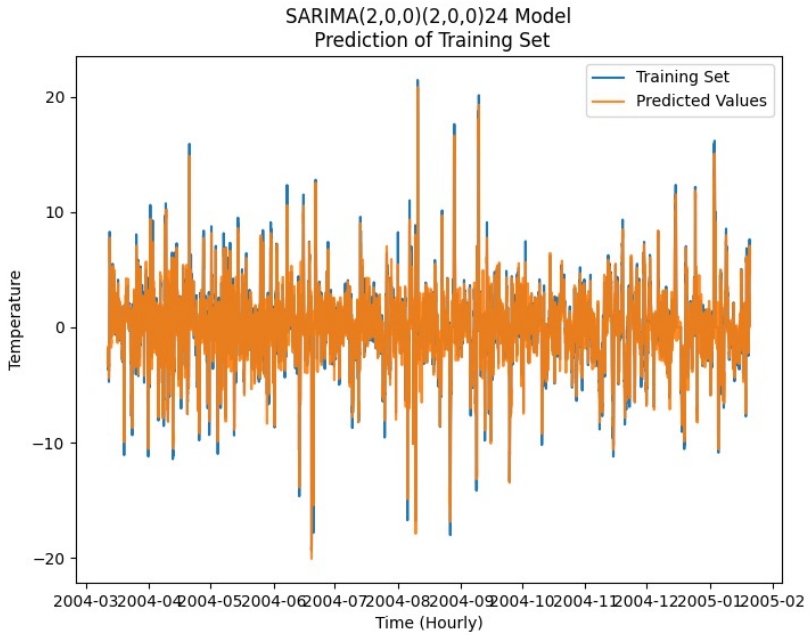
MODEL	RSME
Average	11.09
Naive	7.16
Drift	7.77
SES	6.56
Holt- Winter	7.02
OLS	4.6
ARMA(1,0)	1.41
ARMA(3,2)	1.39
SARIMA(2,0,0)(2,0,0)24	1.19

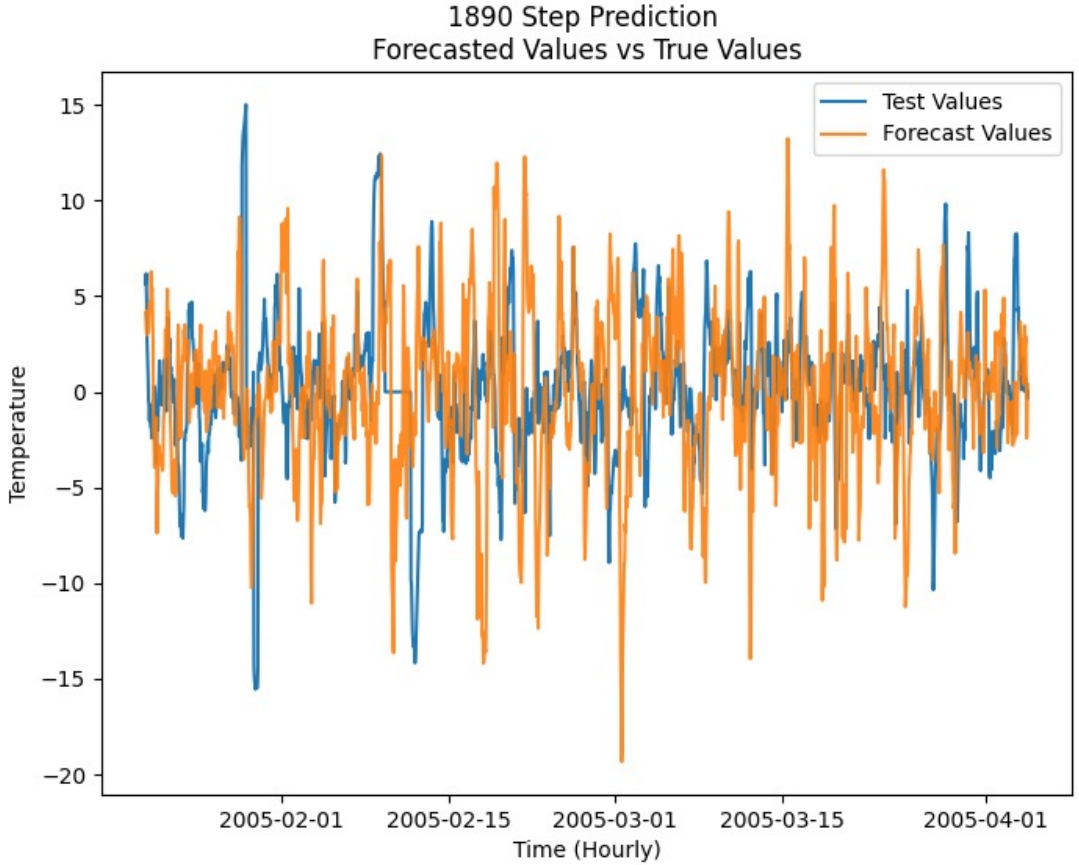
Forecasting



Forecast Function:

$y(t+h) = -1.0609\,(t+h-1) + + 0.1301\,(t+h-2) + 0.6019\,(t+h-24) + 0.3002\,(t+h-48) + e(t)$





Conclusion

Summary

- Highly seasonal dataset
- Poor predictive of base models
- Holt winter model and multiple regression provided insight into seasonality but did not yield accurate results
- ARMA and SARIMA models were significant improvements on the prior models and provide necessary input and insight into time series data
- The Seasonal ARIMA model provided greater flexibility for seasonality adjustments but needs more improvement

THANK YOU