SYMBIOSIS STATISTICAL INSTITUTE

SEMESTER-2

BATCH 2023-2024

TIME SERIES ANALYSIS REPORT

GROUP NO - 09

NAME	PRN
Ashmi Datta	23060641006
Anoushka Dhanpal	23060641014
Anuja Borse	23060641050
Tanushree Roy	23060641044
Sanskruti	23060641041

ABSTRACT:

Weather forecasting is crucial for various sectors, including agriculture, transportation, and disaster management. In this study, we present a comprehensive dataset tailored for developers aiming to train models specifically for weather forecasting in the Indian climate. The dataset spans four years, from 2013 to 2017, and encompasses weather data collected in the city of Delhi. Four key parameters, namely mean temperature, humidity, wind speed, and mean pressure, are included in the dataset.

This dataset is designed to facilitate time series analysis, offering researchers and developers an opportunity to explore the intricacies of weather patterns and trends in the Indian subcontinent. The availability of long-term, localized data such as this can contribute significantly to the development of accurate and reliable weather forecasting models, thereby enhancing preparedness and decision-making in various sectors affected by weather conditions.

INTRODUCTION:

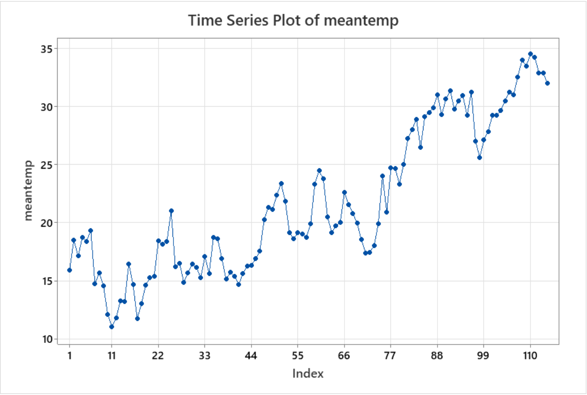
This dataset comprises four key parameters crucial for weather forecasting: mean temperature, humidity, wind speed, and mean pressure. Collected over a period of four years, from 2013 to 2017, the data originates from observations in the city of Delhi, one of India's most significant urban centers. Delhi's climate exhibits substantial variability across seasons, making it an ideal location for studying the nuances of the Indian weather patterns

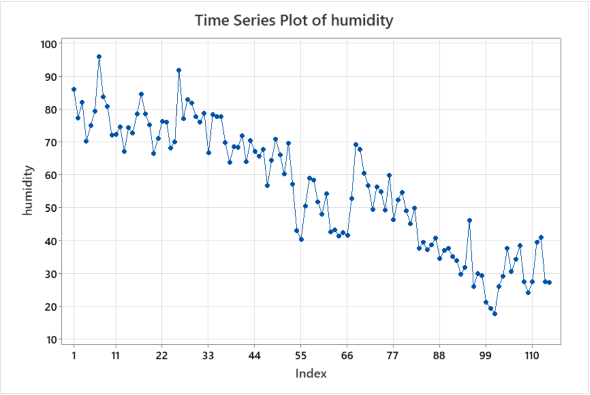
We try to identifying the suitable ARIMA model by conducting the analysis of Autocorrelation function (ACF) and Partial autocorrelation function (PACF) to the selected differenced series and present the forecasting.

PROBLEM IDENTIFICATION:

Despite advancements in weather forecasting, accurately predicting future weather conditions, particularly mean temperature and humidity, remains a challenging task due to the complex and dynamic nature of the Indian climate.

ANALYZE:



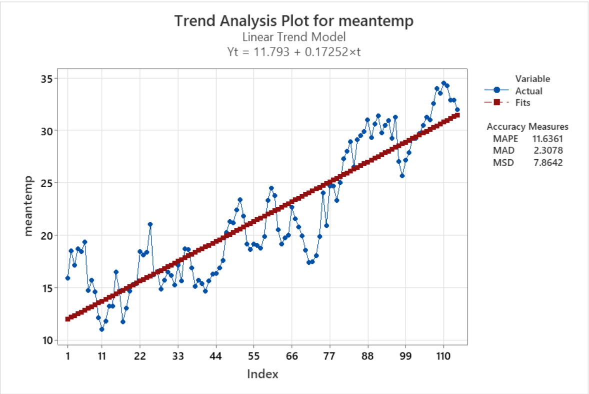


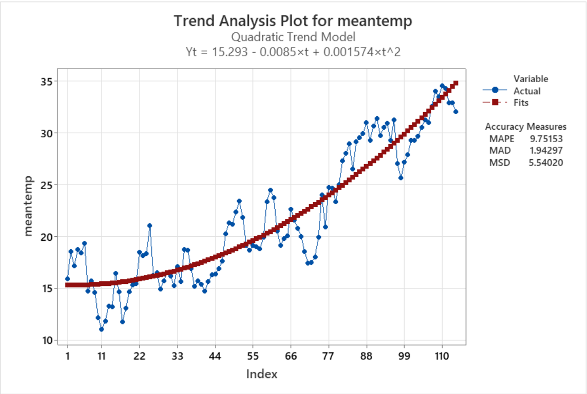
TREND ANALYSIS:

The trend analysis was conducted using a linear trend model and a quadratic trend model on the variables "meantemp" and “humidity” over a dataset of 114 observations.

For Linear Trend Model of:

Trend Analysis for meantemp





Methodology:

- Model Type: Linear Trend Model Trend Model

- Data: humidity

- Length: 114 data points points

- Missing Values: 0

- Fitted Trend Equation:

Yt = 86.55 - 0.5267×t - 0.001455×t^2

Methodology:

- Model Type: Quadratic

- Data: humidity

- Length: 114 data

- Missing Values: 0

- Fitted Trend Equation:

Yt = 83.31 - 0.3594×t

Accuracy Measures:

MAPE: 11.6361%

MAD: 2.3078

MSD: 7.8642

- MAPE: 9.75153%

- MAD: 1.94297

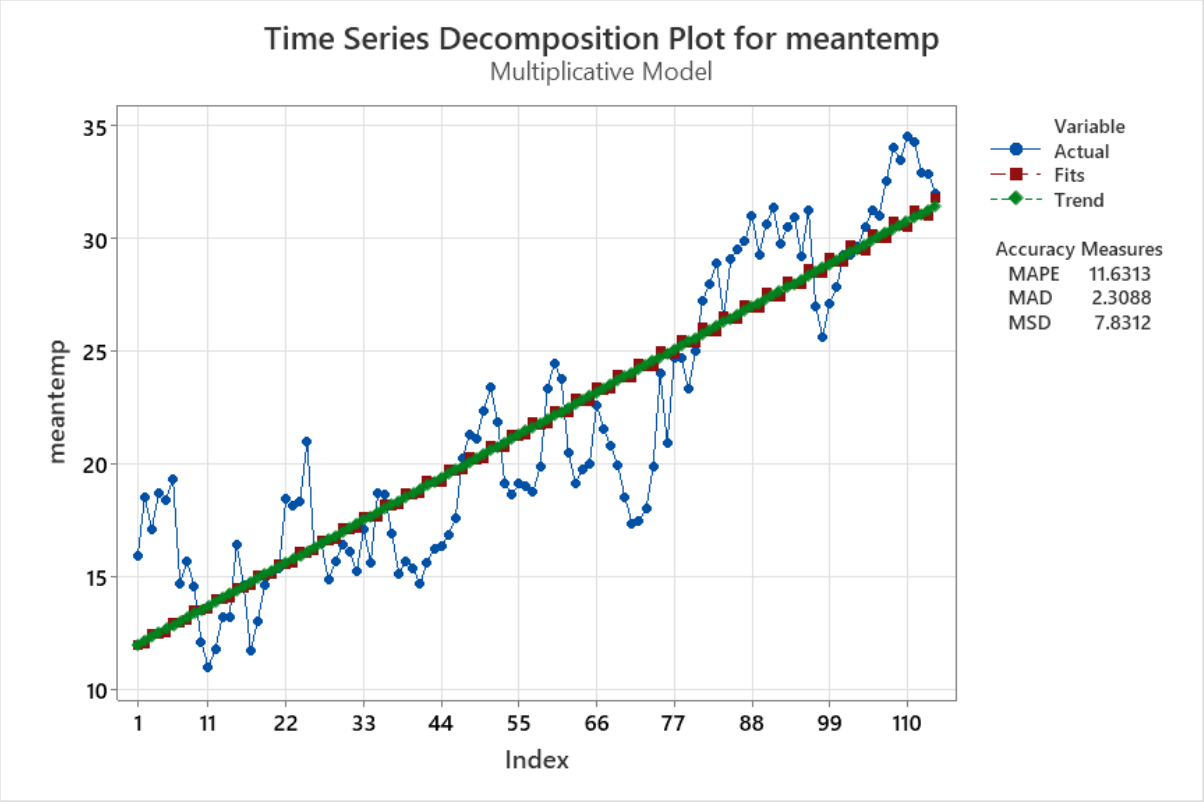
- MSD: 5.54020

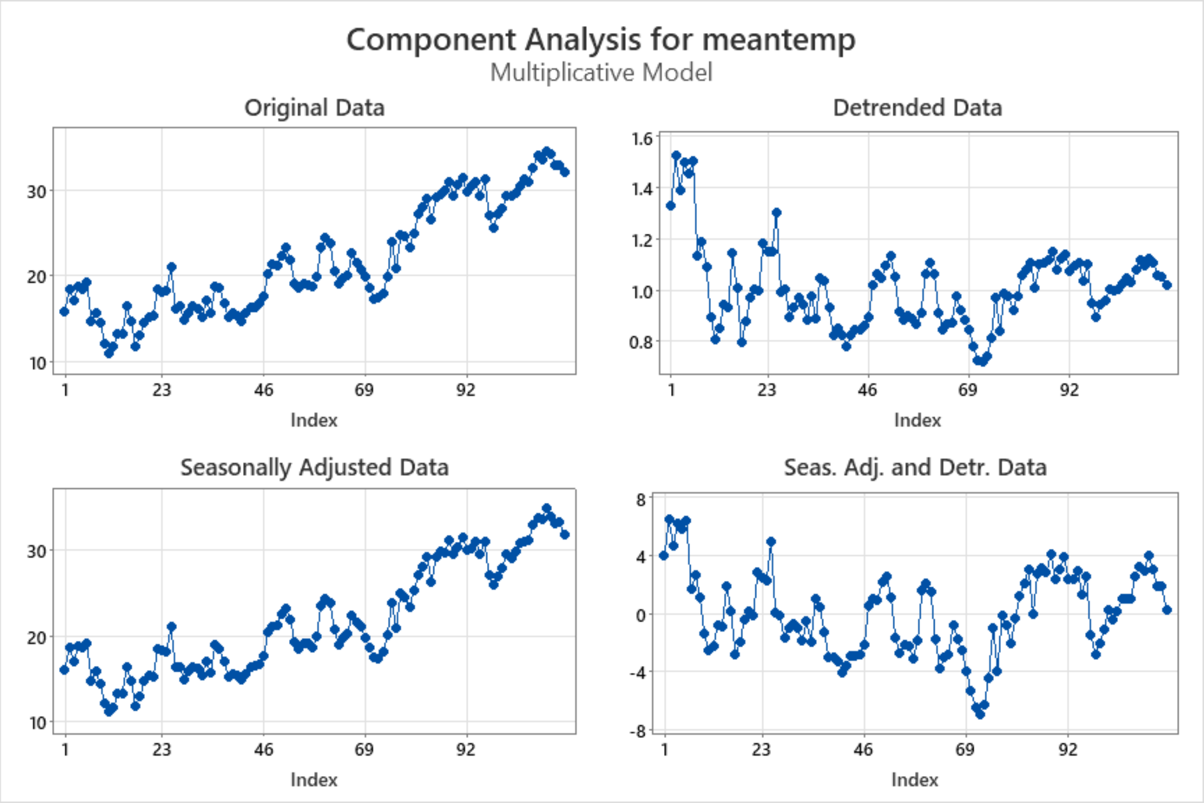
Conclusion:

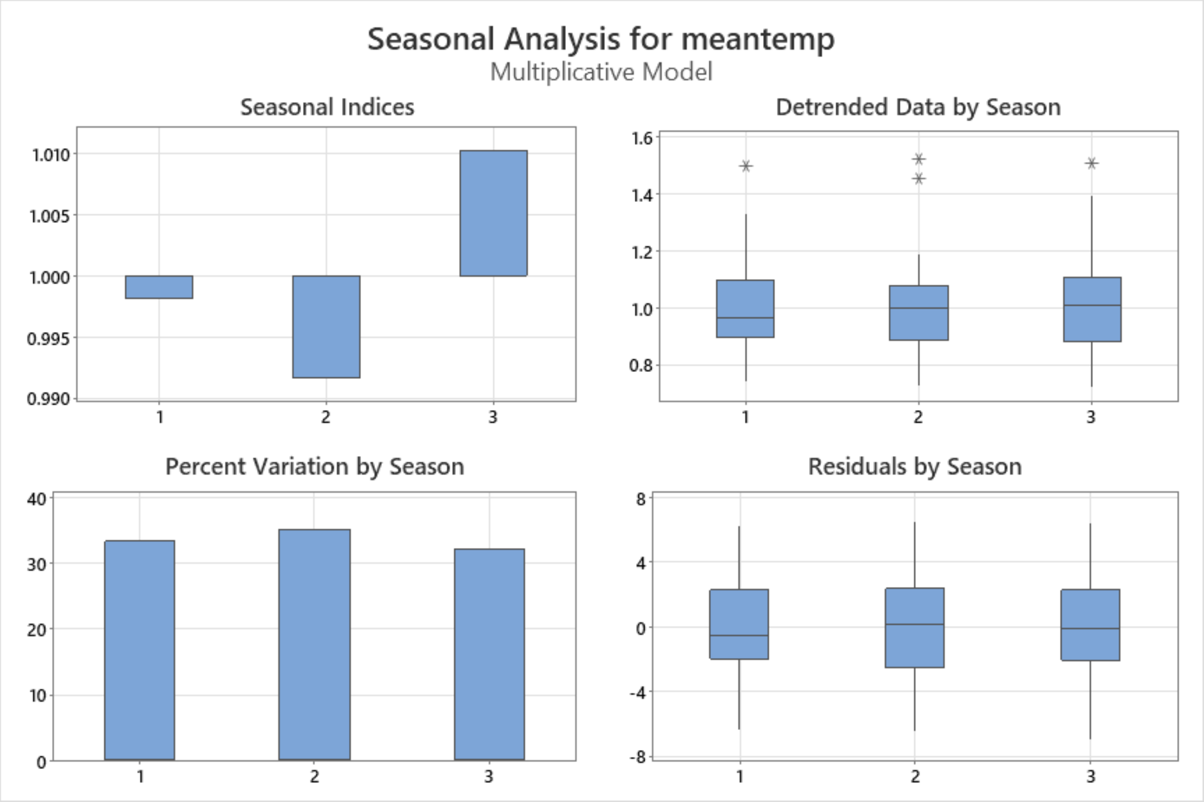
- The fitted trend equation indicates that the mean temperature (meantemp) increases by 0.17252 units per unit increase in time (t), with an intercept of 11.793.
- The fitted trend equation indicates that the mean temperature (meantemp) is modelled as a quadratic function of time (t), with coefficients of -0.0085 for t and 0.001574 for t^2, along with an intercept of 15.293.
- We consider quadratic trend model is a better fit for the model as it gives comparatively lower values for the MAPE, MAD and MSD.

DECOMPOSTION :

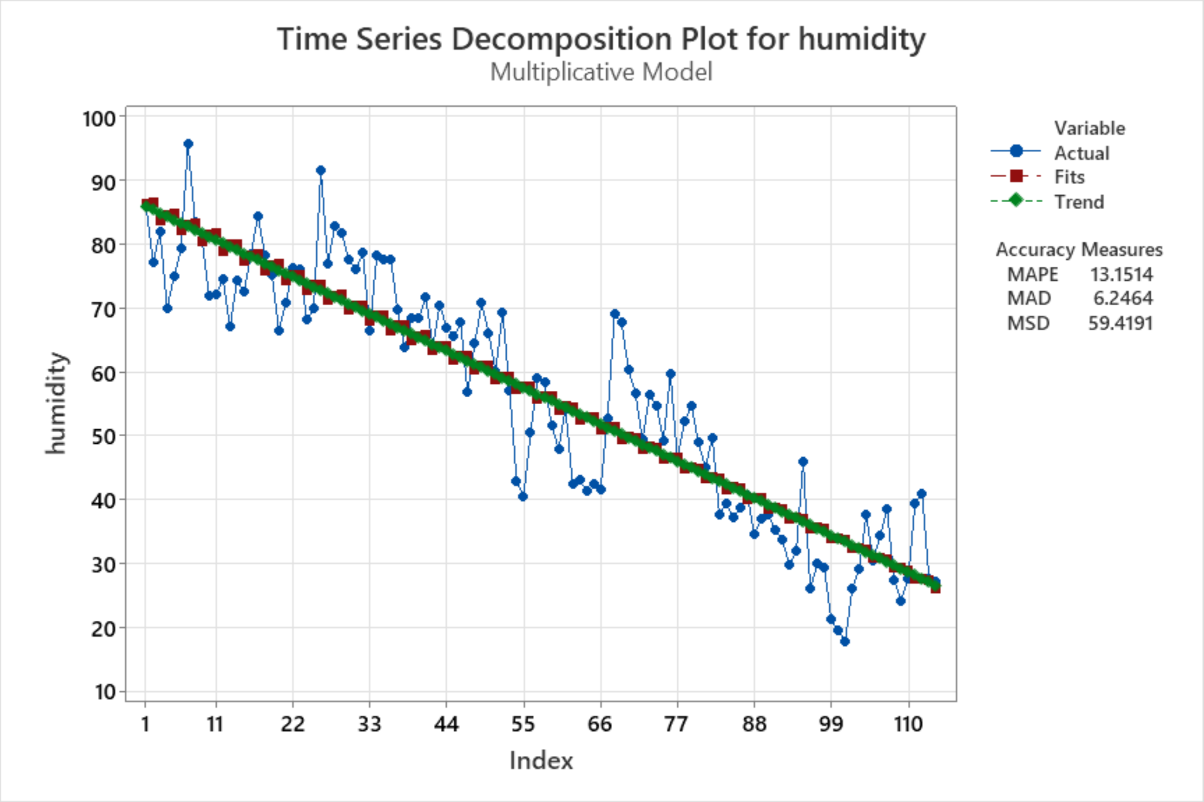
Time Series Decomposition for meantemp:

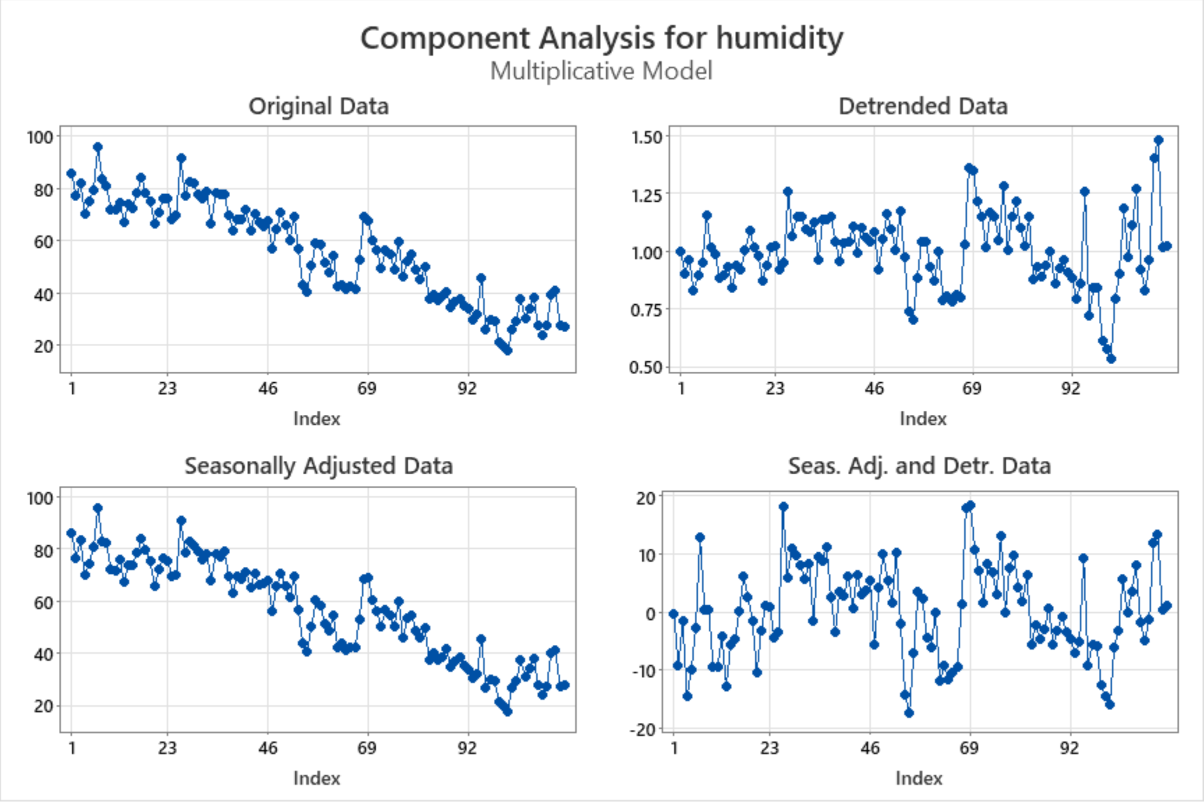


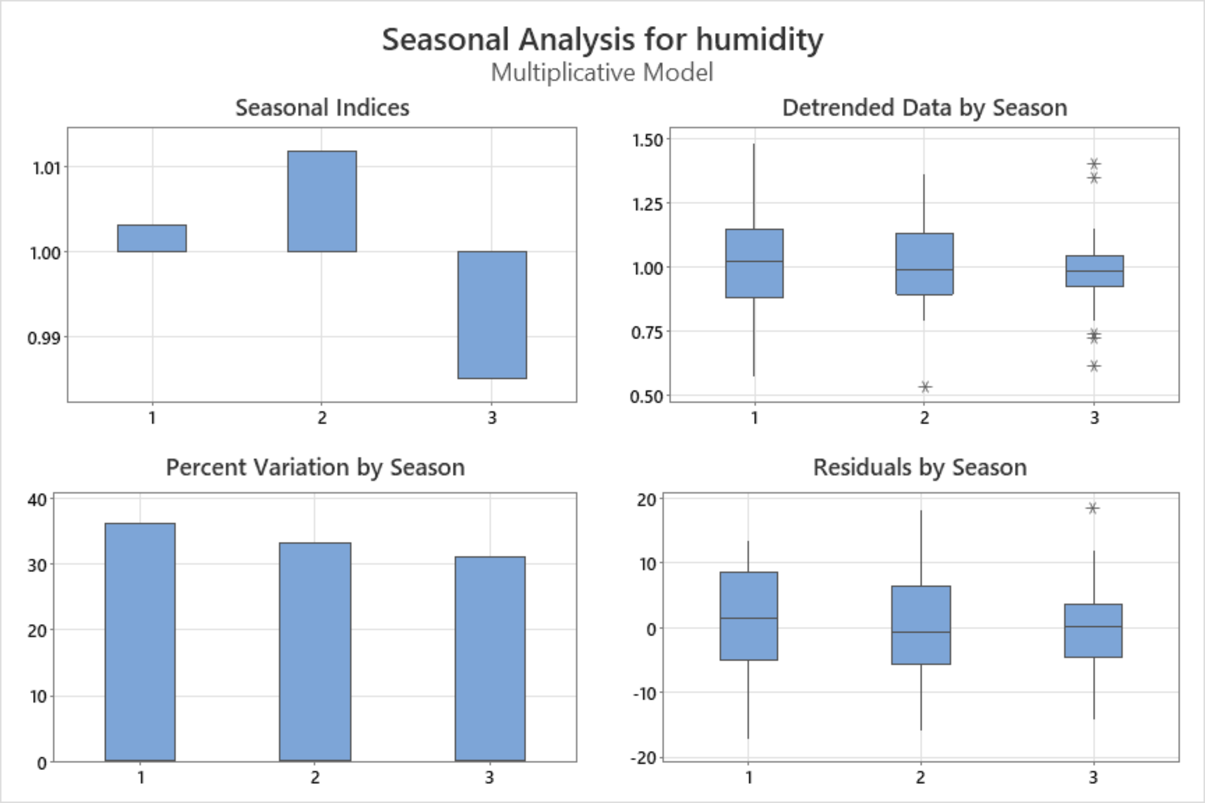

Component Analysis for meantemp
Multiplicative Model
Original Data
10
20
30
1
23
46
69
92
Index
Detrended Data
0.8
1.0
1.2
1.4
1.6
1
23
46
69
92
Index
Seasonally Adjusted Data
10
20
30
1
23
46
69
92
Index
Seas. Adj. and Detr. Data
-8
-4
0
4
8
1
23
46
69
92
Index


Seasonal Analysis for meantemp
Multiplicative Model
Seasonal Indices
0.990
0.995
1.000
1.005
1.010
1
2
3
Detrended Data by Season
0.8
1.0
1.2
1.4
1.6
1
2
3
Percent Variation by Season
0
10
20
30
40
1
2
3
Residuals by Season
-8
-4
0
4
8
1
2
3

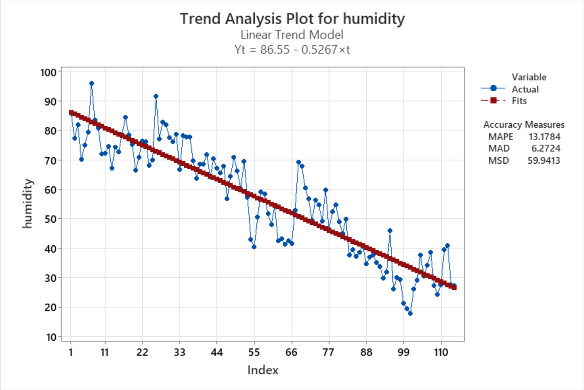
Time Series Decomposition for humidity

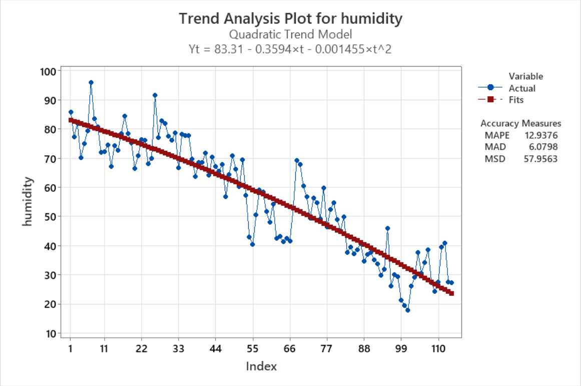






Trend Analysis for humidity





Methodology:

- Model Type: Linear Trend Model

Methodology:

- Model Type: Quadratic Trend Model

- Data: humidity

- Length: 114 data points

- Missing Values: 0

- Fitted Trend Equation:

Yt = 86.55 - 0.5267×t

Accuracy Measures:

MAPE: 13.1784%

MAD: 6.2724

MSD: 59.9413

- Data: humidity

- Length: 114 data points

- Missing Values: 0

- Fitted Trend Equation:

Yt = 83.31 - 0.3594×t - 0.001455×t^2

- MAPE: 12.9376%

- MAD: 6.0798

- MSD: 57.9563

Conclusion:

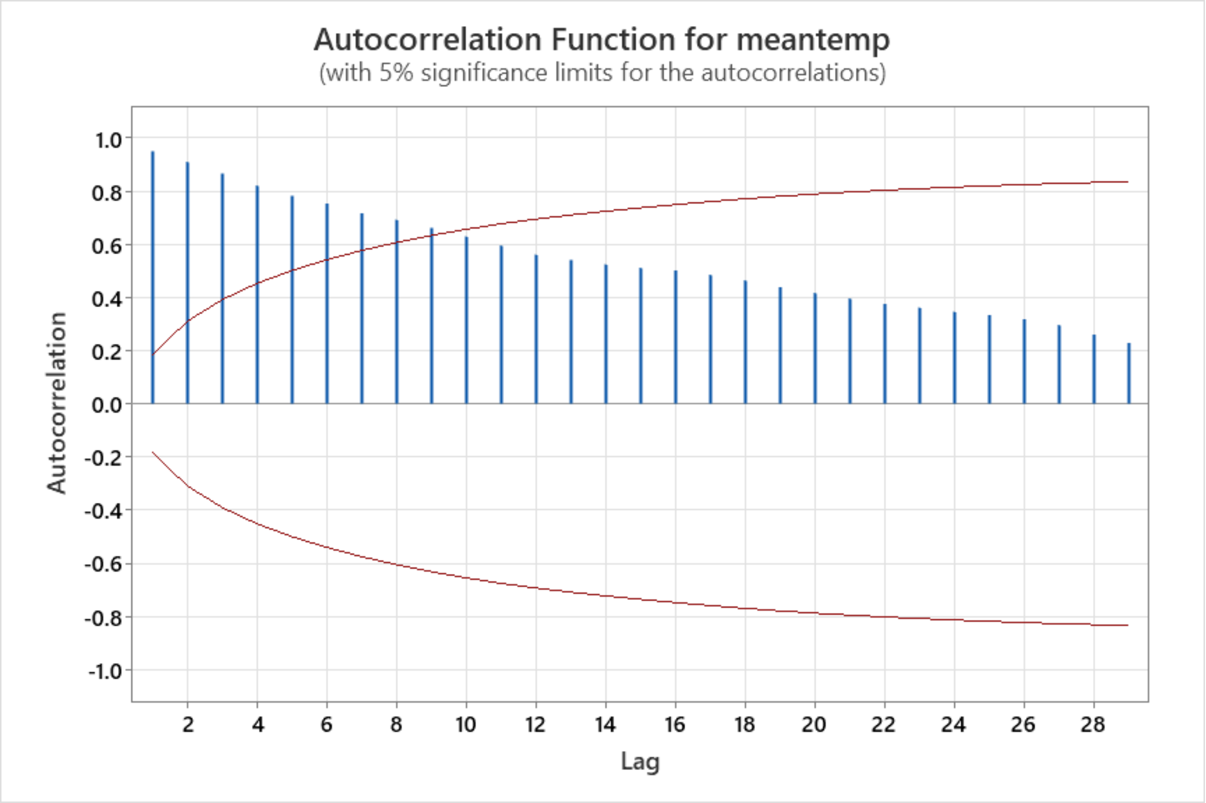
- The fitted linear trend equation indicates that the humidity (Yt) decreases by 0.5267 units per unit increase in time (t), with an intercept of 86.55.
- The fitted trend equation indicates that the humidity (Yt) is modelled as a quadratic function of time (t), with coefficients of -0.3594 for t and -0.001455 for t^2, along with an intercept of 83.31.
- We consider quadratic trend model is a better fit for the model as it gives comparatively lower values for the MAPE, MAD and MSD.

ACF (AUTOCORRELATION FUNCTION)

Autocorrelations

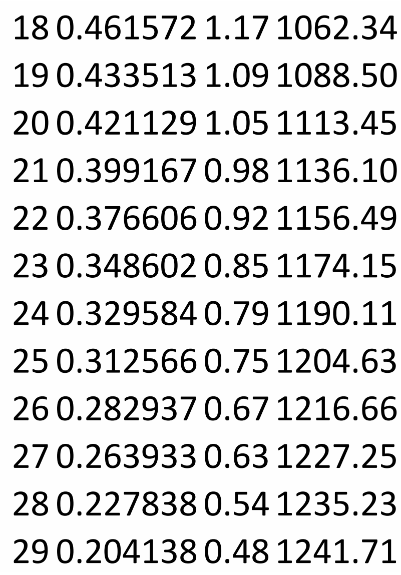
Lag	ACF	T	LBQ
1	0.949444	10.14	105.49

2	0.908762	5.80	203.00
3	0.865330	4.38	292.21
4	0.819139	3.58	372.87
5	0.782145	3.09	447.09
6	0.752520	2.75	516.43
7	0.716584	2.46	579.89
8	0.690539	2.26	639.38
9	0.661451	2.07	694.48
10	0.628552	1.90	744.72
11	0.594050	1.74	790.03
12	0.560192	1.60	830.71
13	0.540706	1.51	868.99
14	0.523183	1.43	905.19
15	0.510253	1.37	939.97
16	0.500632	1.32	973.79
17	0.483643	1.26	1005.68
18	0.462620	1.19	1035.16
19	0.438627	1.11	1061.94
20	0.415685	1.04	1086.25
21	0.394860	0.98	1108.42
22	0.375453	0.93	1128.68
23	0.360414	0.88	1147.55
24	0.344299	0.84	1164.97
25	0.333688	0.81	1181.52
26	0.317242	0.76	1196.64
27	0.294539	0.70	1209.83
28	0.259812	0.62	1220.21
29	0.228753	0.54	1228.35

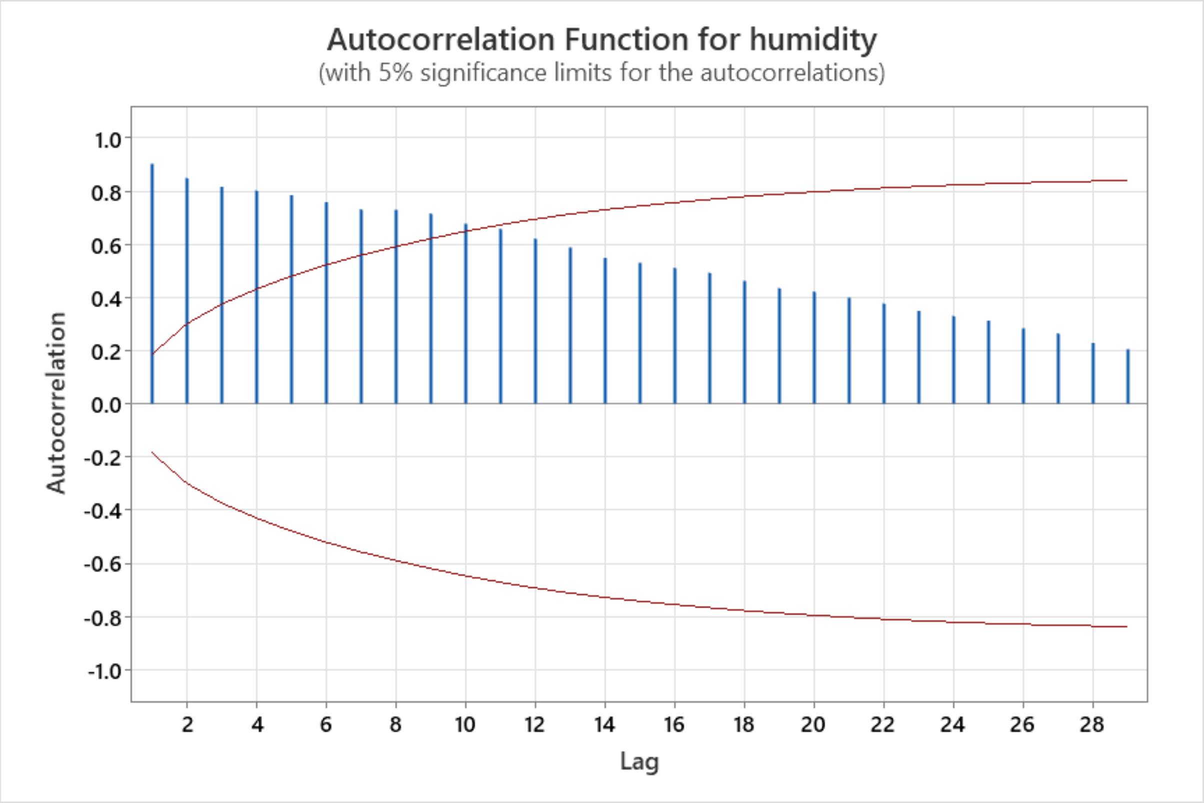


Autocorrelations

Lag	ACF	T	LBQ
1	0.902860	9.64	95.39
2	0.848477	5.59	180.40
3	0.816012	4.32	259.73
4	0.801922	3.68	337.03
5	0.784392	3.24	411.68
6	0.757862	2.88	482.01
7	0.730779	2.59	548.01
8	0.729316	2.45	614.37
9	0.714650	2.28	678.69
10	0.677229	2.07	737.00
11	0.656571	1.93	792.35
12	0.620674	1.77	842.30
13	0.587952	1.63	887.56
14	0.548786	1.49	927.38
15	0.529708	1.41	964.86
16	0.509964	1.34	999.96
17	0.492258	1.27	1032.99



18	0.461572	1.17	1062.34
19	0.433513	1.09	1088.50
20	0.421129	1.05	1113.45
21	0.399167	0.98	1136.10
22	0.376606	0.92	1156.49
23	0.348602	0.85	1174.15
24	0.329584	0.79	1190.11
25	0.312566	0.75	1204.63
26	0.282937	0.67	1216.66
27	0.263933	0.63	1227.25
28	0.227838	0.54	1235.23
29	0.204138	0.48	1241.71



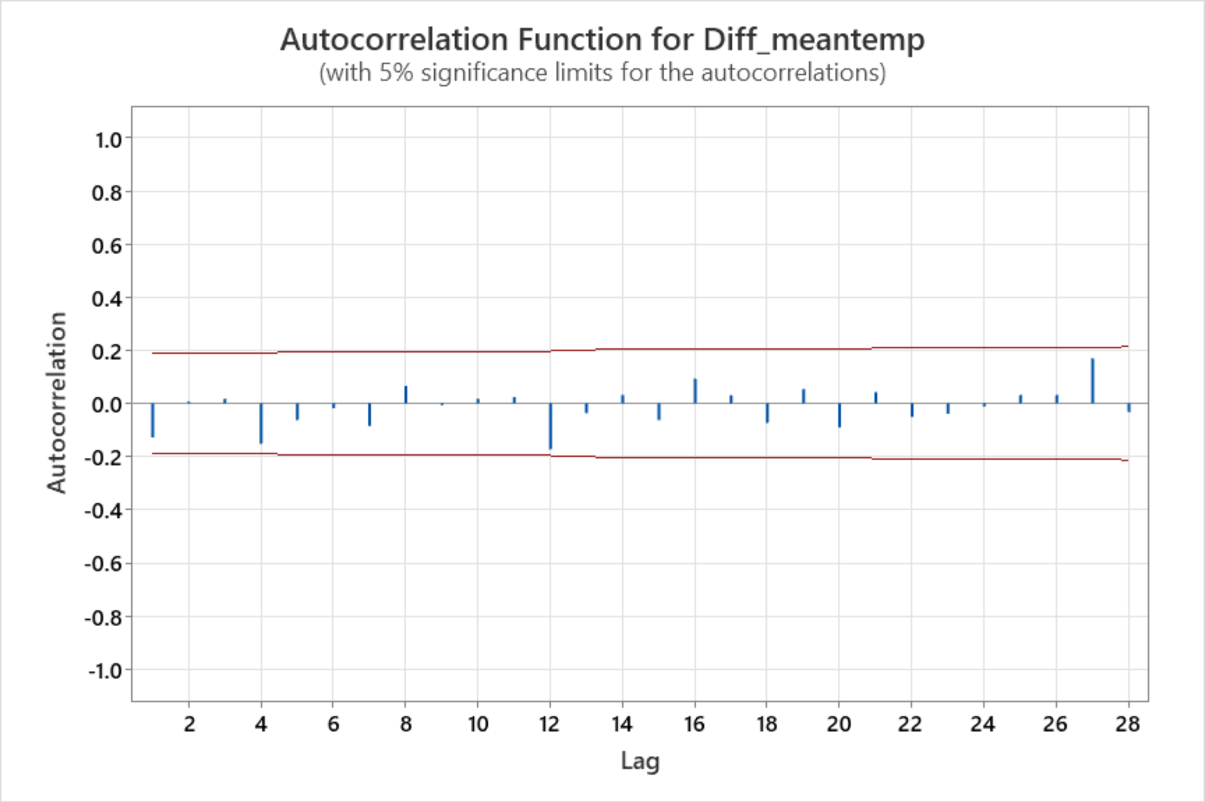
The ACF for meantemp and humidity are showing trends
that is:
shorter lags are showing large positive correlation because observations closer in time tend to have similar values
And correlation taper of slowly as the lags increase

Here we apply differencing to get better ACF Graph as follows:

Autocorrelations

Lag	ACF	T	LBQ
1	-0.128573	-1.37	1.92
2	0.006146	0.06	1.92
3	0.017680	0.18	1.96
4	-0.152264	-1.59	4.72
5	-0.062814	-0.64	5.20
6	-0.018417	-0.19	5.24
7	-0.085002	-0.87	6.13
8	0.065402	0.66	6.65
9	-0.006356	-0.06	6.66
10	0.016636	0.17	6.69
11	0.023213	0.23	6.76
12	-0.173332	-1.75	10.63
13	-0.036606	-0.36	10.80
14	0.031796	0.31	10.94
15	-0.062491	-0.61	11.45
16	0.093064	0.91	12.61
17	0.030011	0.29	12.74
18	-0.073144	-0.71	13.47
19	0.053200	0.51	13.86
20	-0.090710	-0.87	15.01
21	0.042179	0.40	15.26
22	-0.050630	-0.48	15.63

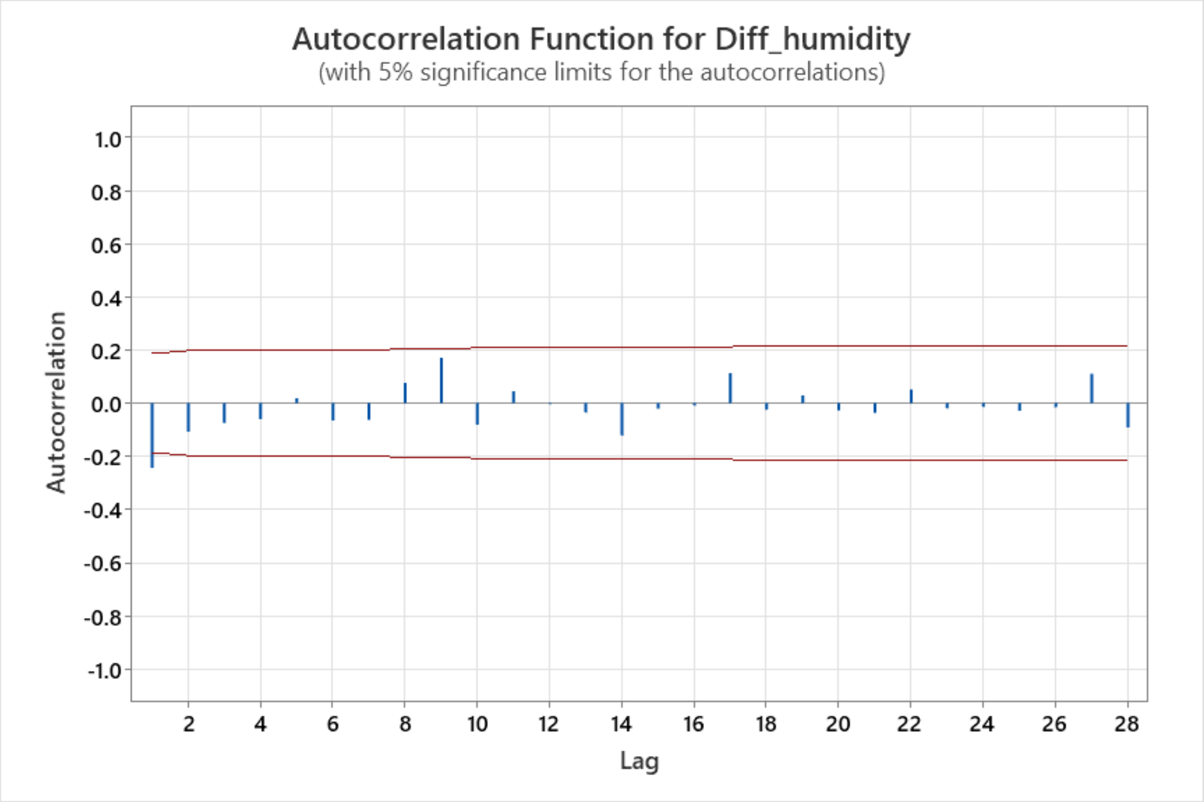
23	-0.039899	-0.38	15.86
24	-0.012225	-0.12	15.88
25	0.031567	0.30	16.02
26	0.031942	0.30	16.18
27	0.169194	1.61	20.50
28	-0.033209	-0.31	20.67



Autocorrelations

Lag	ACF	T	LBQ
1	-0.243509	-2.59	6.88
2	-0.107664	-1.08	8.24
3	-0.074756	-0.74	8.90
4	-0.060230	-0.60	9.33
5	0.017977	0.18	9.37

6	-0.065861	-0.65	9.90
7	-0.062632	-0.62	10.38
8	0.076189	0.75	11.09
9	0.170645	1.66	14.73
10	-0.081135	-0.77	15.56
11	0.044731	0.42	15.82
12	-0.003685	-0.03	15.82
13	-0.035667	-0.34	15.99
14	-0.121585	-1.15	17.93
15	-0.021094	-0.20	17.99
16	-0.008664	-0.08	18.00
17	0.112256	1.05	19.70
18	-0.025162	-0.23	19.79
19	0.028781	0.27	19.90
20	-0.027913	-0.26	20.01
21	-0.036275	-0.33	20.20
22	0.051278	0.47	20.57
23	-0.019193	-0.18	20.63
24	-0.013948	-0.13	20.65
25	-0.029140	-0.27	20.78
26	-0.016378	-0.15	20.82
27	0.109665	1.01	22.64
28	-0.091415	-0.83	23.91



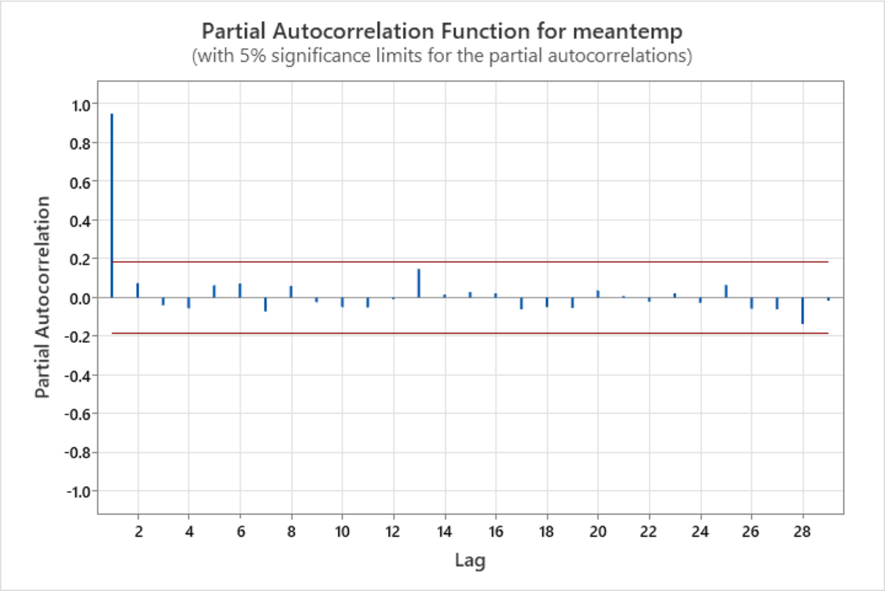
On differencing we can see that the given data is not random as it has at least one significant lag.

When the data is not random it is good indication for the use of time series analysis or incorporate lags into a regression analysis to model the appropriately.

PACF FOR MEANTEMP:

Partial Autocorrelations

Lag	PACF	T
1	0.949444	10.14
2	0.074256	0.79
3	-0.040003	-0.43
4	-0.055786	-0.60
5	0.063662	0.68
6	0.073172	0.78
7	-0.070734	-0.76
8	0.060593	0.65
9	-0.023831	-0.25
10	-0.048645	-0.52
11	-0.050335	-0.54
12	-0.008355	-0.09
13	0.147862	1.58
14	0.016244	0.17
15	0.028822	0.31
16	0.022088	0.24
17	-0.060128	-0.64
18	-0.048810	-0.52
19	-0.053992	-0.58
20	0.036194	0.39
21	0.007384	0.08
22	-0.021060	-0.22
23	0.022071	0.24
24	-0.027706	-0.30
25	0.065245	0.70
26	-0.057450	-0.61
27	-0.060714	-0.65
28	-0.135925	-1.45
29	-0.015403	-0.16

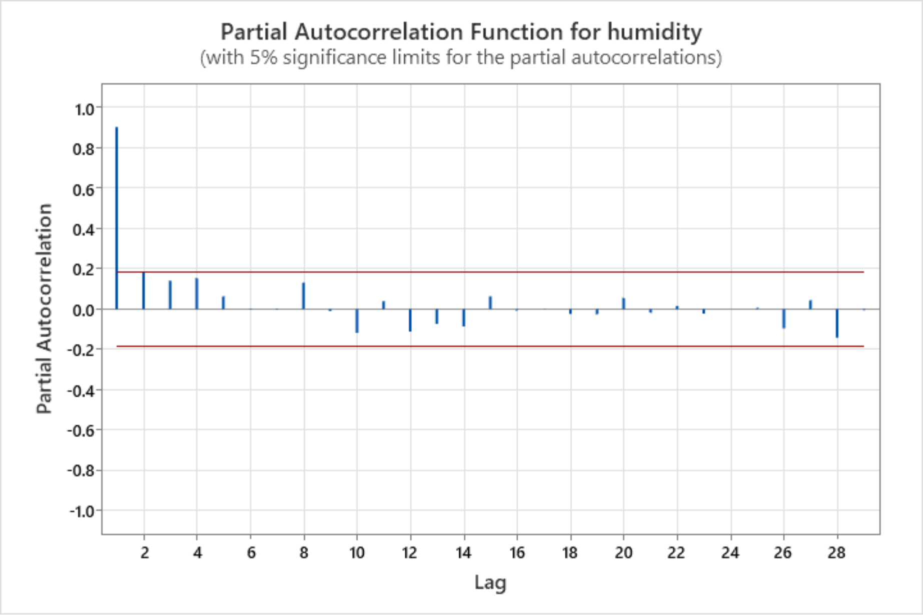

Partial Autocorrelation Function for meantemp
(with 5% significance limits for the partial autocorrelations)

- The PACF graph's x-axis to find the range of lag values. The partial autocorrelation values are represented on the y-axis while the lag values are represented on the x-axis. The typical range of lag values, where n is the length of the time series, is 0 to n-1.
- The partial autocorrelation values are taken into consideration to be significant when they reach the significance level. In the PACF graph, a horizontal line is typically used to indicate it.
- Any partial autocorrelation value greater than the 95% significance threshold is deemed significant when using a 95% confidence level, which is the standard.
- In figures the PACF has 1 significant lags followed by a drop in PACF values and they become insignificant. With 1 significant PACF lags we can say that the series is an AR(1) process. The lags of AR are determined by the number of significant lags of PACF.

PACF FOR HUMIDITY:

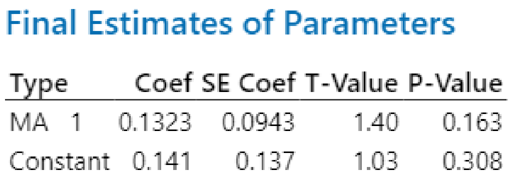
Partial Autocorrelations

Lag	PACF	T
1	0.902860	9.64
2	0.180264	1.92
3	0.141446	1.51
4	0.153827	1.64
5	0.064080	0.68
6	-0.003627	-0.04
7	-0.003810	-0.04
8	0.131920	1.41
9	-0.008933	-0.10
10	-0.117231	-1.25
11	0.039450	0.42
12	-0.110271	-1.18
13	-0.073503	-0.78
14	-0.086137	-0.92
15	0.063355	0.68
16	-0.006667	-0.07
17	-0.001345	-0.01
18	-0.024248	-0.26
19	-0.025013	-0.27
20	0.055902	0.60
21	-0.017227	-0.18
22	0.015479	0.17
23	-0.022389	-0.24
24	0.000664	0.01
25	0.005935	0.06
26	-0.095255	-1.02
27	0.043809	0.47
28	-0.140678	-1.50
29	-0.004589	-0.05



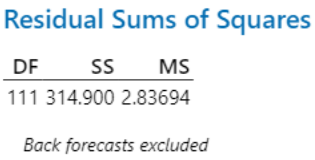
- Lags 1, 2, 3, 4, 8, 11, 15, 20, 26, 27 exhibit relatively strong positive correlations with the original series.
- Lags 10, 12, 13, 14, 28 demonstrate moderate negative correlations.
- Many lags have PACF values close to zero, indicating little direct correlation with the original series.
- the PACF graph's x-axis to find the range of lag values. The partial autocorrelation values are represented on the y-axis while the lag values are represented on the x-axis. The typical range of lag values, where n is the length of the time series, is 0 to n-1.
- The partial autocorrelation values are taken into consideration to be significant when they reach the significance level. In the PACF graph, a horizontal line is typically used to indicate it.
- Any partial autocorrelation value greater than the 95% significance threshold is deemed significant when using a 95% confidence level, which is the standard.
- process. In the figures the PACF has 1 significant lags followed by a drop in PACF values and they become insignificant. With 1 significant PACF lags we can say that the series is an AR(1) process. The lags of AR are determined by the number of significant lags of PACF.

ARIMA MODEL FOR MEANTEMP(0,1,1):



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
MA 1	0.1323	0.0943	1.40	0.163
Constant	0.141	0.137	1.03	0.308



Residual Sums of Squares

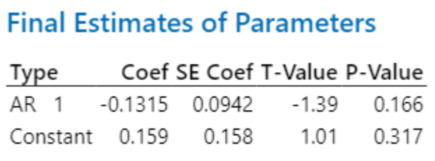
DF	SS	MS
111	314.900	2.83694

Back forecasts excluded

Use the mean square error (MS) to determine how well the model fits the data. Smaller values indicate a better fitting model

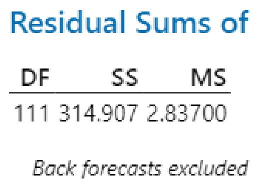
The mean square error is 2.8369 for this model

ARIMA MODEL FOR MEANTEMP(1,1,0):



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	-0.1315	0.0942	-1.39	0.166
Constant	0.159	0.158	1.01	0.317



Residual Sums of Squares

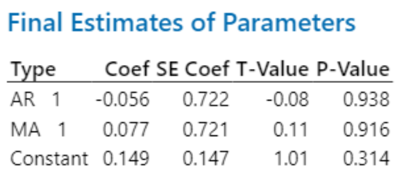
DF	SS	MS
111	314.907	2.83700

Back forecasts excluded

Use the mean square error (MS) to determine how well the model fits the data. Smaller values indicate a better fitting model

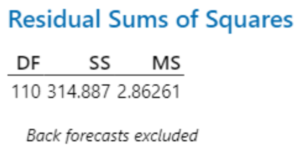
The mean square error is 2.83700 for this model

ARIMA MODEL FOR MEANTEMP(1,1,1):



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	-0.056	0.722	-0.08	0.938
MA 1	0.077	0.721	0.11	0.916
Constant	0.149	0.147	1.01	0.314



Residual Sums of Squares

DF	SS	MS
110	314.887	2.86261

Back forecasts excluded

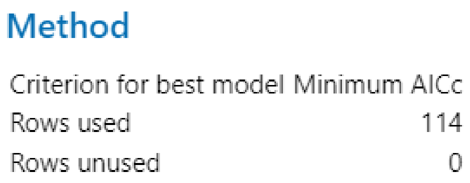
Use the mean square error (MS) to determine how well the model fits the data. Smaller values indicate a better fitting model

The mean square error is 2.86261 for this model

INTERPRETATION:

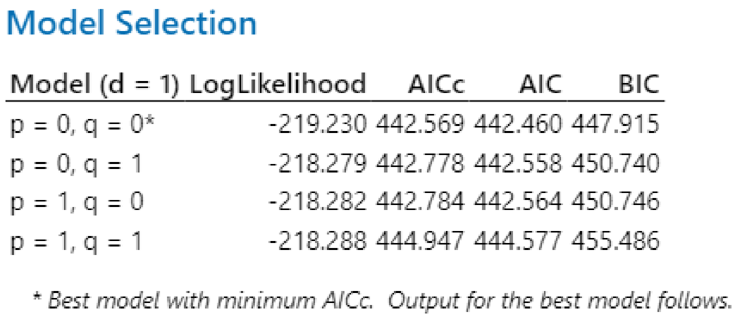
ARIMA(0,1,1) has the lowest MS value and hence best model.

FORECAST WITH BEST ARIMA MODEL FOR MEANTEMP:



Method

Criterion for best model	Minimum AICc
Rows used	114
Rows unused	0

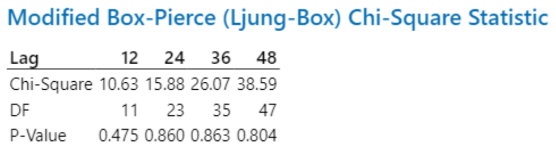


Model Selection

Model (d = 1)	LogLikelihood	AICc	AIC	BIC
p = 0, q = 0*	-219.230	442.569	442.460	447.915
p = 0, q = 1	-218.279	442.778	442.558	450.740
p = 1, q = 0	-218.282	442.784	442.564	450.746
p = 1, q = 1	-218.288	444.947	444.577	455.486

** Best model with minimum AICc. Output for the best model follows.*

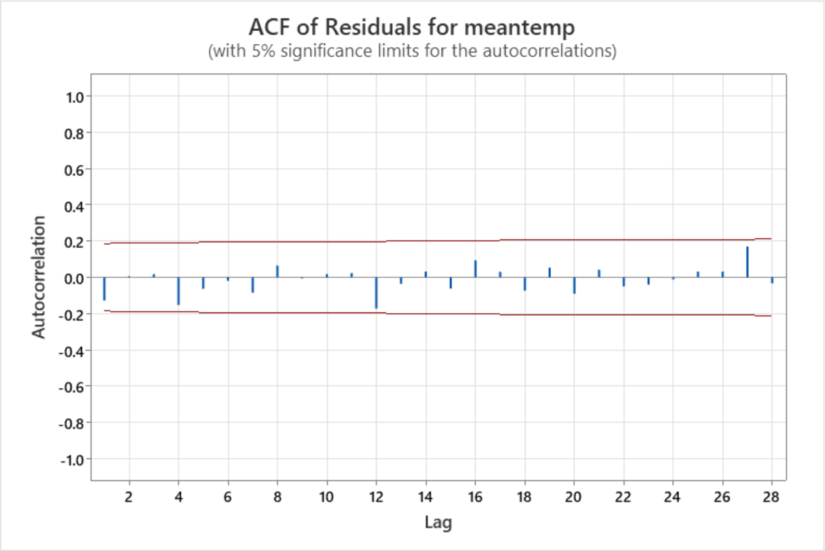
- Use AIC, AICc and BIC to compare different models. Smaller values are desirable.
- Use tests and plots to assess how well the model fits the data. By default, the ARIMA results are for the model with the best value of AICc.
- The ARIMA(0, 1,0) has the best value of AICc. The ARIMA results that follow are for the ARIMA(0, 1, 0) model.

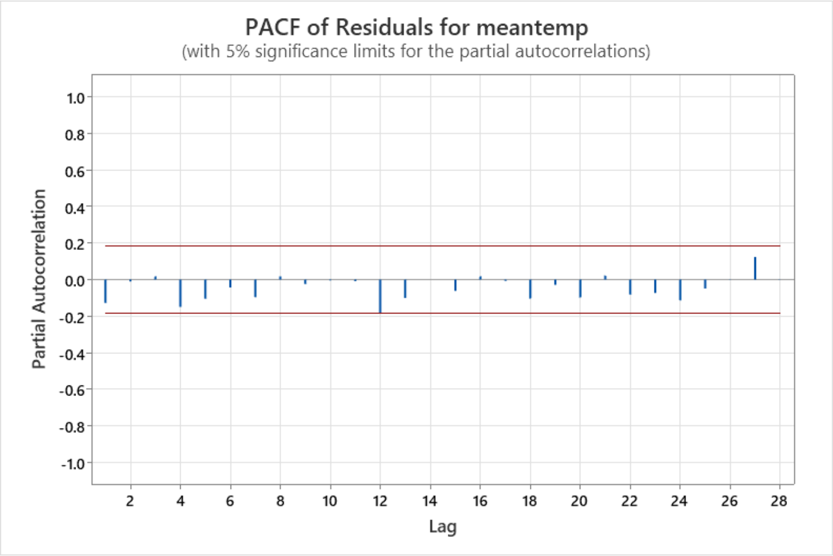


Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

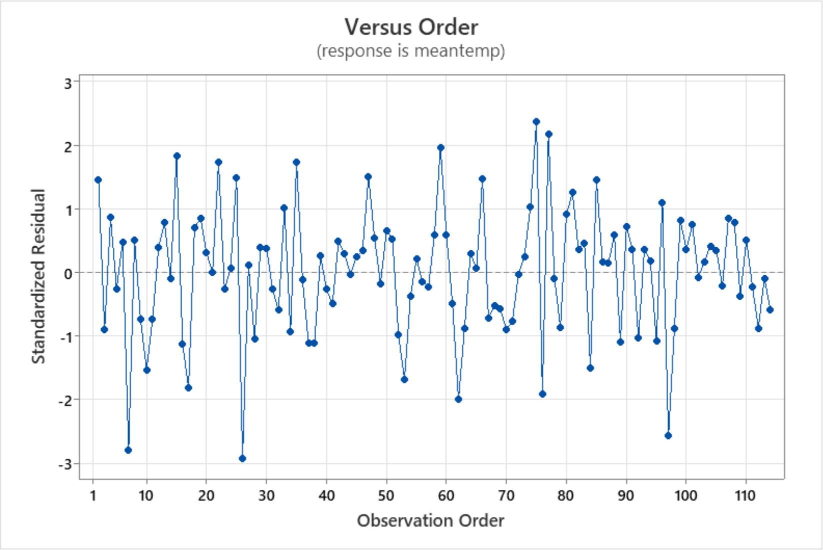
Lag	12	24	36	48
Chi-Square	10.63	15.88	26.07	38.59
DF	11	23	35	47
P-Value	0.475	0.860	0.863	0.804

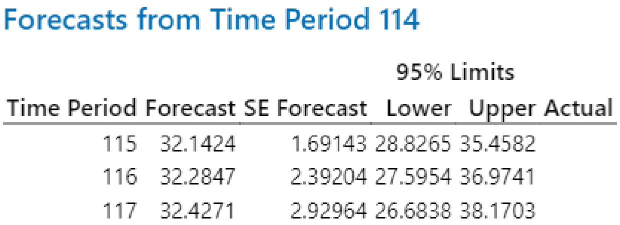
- To determine whether the residuals are independent, compare the p-value to the significance level for each chi square statistic. Usually, a significance level (denoted as α or alpha) of 0.05 works well. If the p-value is greater than the significance level, you can conclude that the residuals are independent and that the model meets the assumption.
- In these results, the p-values for the Ljung-Box chi-square statistics are all greater than 0.05





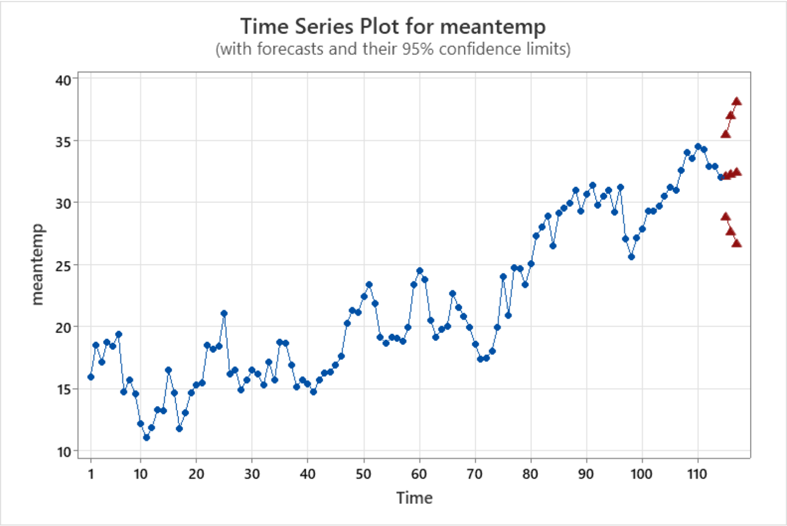
None of the correlations for the autocorrelation function of the residuals or the partial autocorrelation function of the residuals are significant. We can conclude that the model meets the assumption that the residuals are independent.



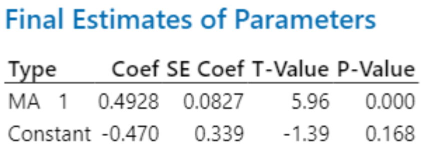


Forecasts from Time Period 114

			95% Limits		
Time Period	Forecast	SE Forecast	Lower	Upper	Actual
115	32.1424	1.69143	28.8265	35.4582	
116	32.2847	2.39204	27.5954	36.9741	
117	32.4271	2.92964	26.6838	38.1703	

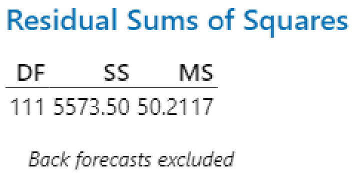


ARIMA MODEL FOR HUMIDITY(0,1,1):



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
MA 1	0.4928	0.0827	5.96	0.000
Constant	-0.470	0.339	-1.39	0.168



Residual Sums of Squares

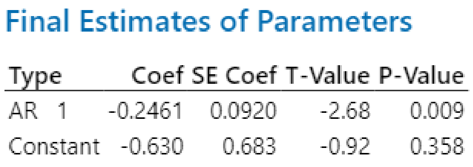
DF	SS	MS
111	5573.50	50.2117

Back forecasts excluded

Use the mean square error (MS) to determine how well the model fits the data. Smaller values indicate a better fitting model

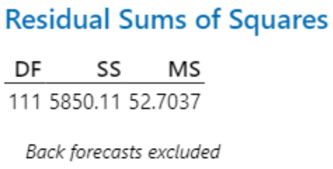
The mean square error is 50.2117 for this model.

ARIMA MODEL FOR HUMIDITY(1,1,0):



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	-0.2461	0.0920	-2.68	0.009
Constant	-0.630	0.683	-0.92	0.358



Residual Sums of Squares

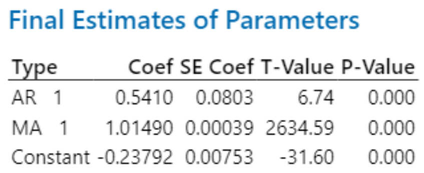
DF	SS	MS
111	5850.11	52.7037

Back forecasts excluded

Use the mean square error (MS) to determine how well the model fits the data. Smaller values indicate a better fitting model

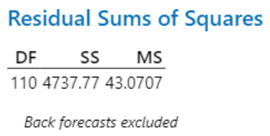
The mean square error is 52.7037 for this model.

ARIMA MODEL FOR HUMIDITY(1,1,1):



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	0.5410	0.0803	6.74	0.000
MA 1	1.01490	0.00039	2634.59	0.000
Constant	-0.23792	0.00753	-31.60	0.000



Residual Sums of Squares

DF	SS	MS
110	4737.77	43.0707

Back forecasts excluded

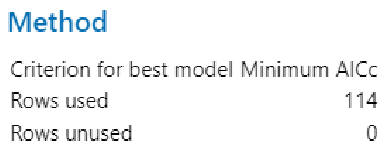
Use the mean square error (MS) to determine how well the model fits the data. Smaller values indicate a better fitting model

The mean square error is 43.0707 for this model.

INTERPRETATION:

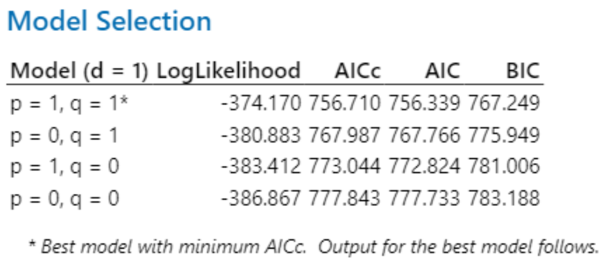
ARIMA(1,1,1) has the lowest MS value and hence best model.

FORECAST WITH BEST ARIMA MODEL FOR HUMIDITY:



Method

Criterion for best model	Minimum AICc
Rows used	114
Rows unused	0



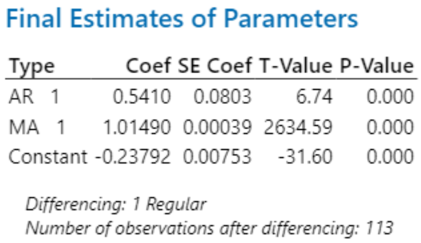
Model Selection

Model (d = 1)	LogLikelihood	AICc	AIC	BIC
p = 1, q = 1*	-374.170	756.710	756.339	767.249
p = 0, q = 1	-380.883	767.987	767.766	775.949
p = 1, q = 0	-383.412	773.044	772.824	781.006
p = 0, q = 0	-386.867	777.843	777.733	783.188

* *Best model with minimum AICc. Output for the best model follows.*

- Use AIC, AICc and BIC to compare different models. Smaller values are desirable.

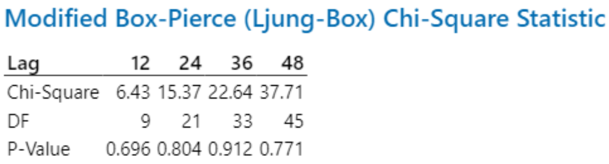
- Use tests and plots to assess how well the model fits the data. By default, the ARIMA results are for the model with the best value of AICc.
- The ARIMA(1, 1,1) has the best value of AICc. The ARIMA results that follow are for the ARIMA(1, 1, 1) model.



Final Estimates of Parameters

Type	Coef	SE Coef	T-Value	P-Value
AR 1	0.5410	0.0803	6.74	0.000
MA 1	1.01490	0.00039	2634.59	0.000
Constant	-0.23792	0.00753	-31.60	0.000

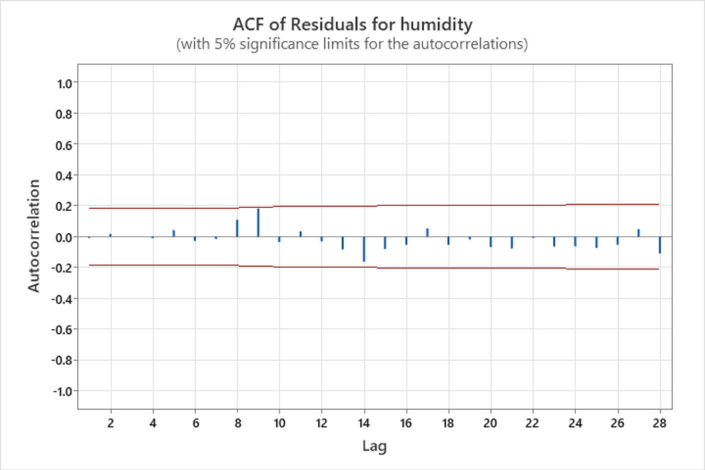
Differencing: 1 Regular
Number of observations after differencing: 113

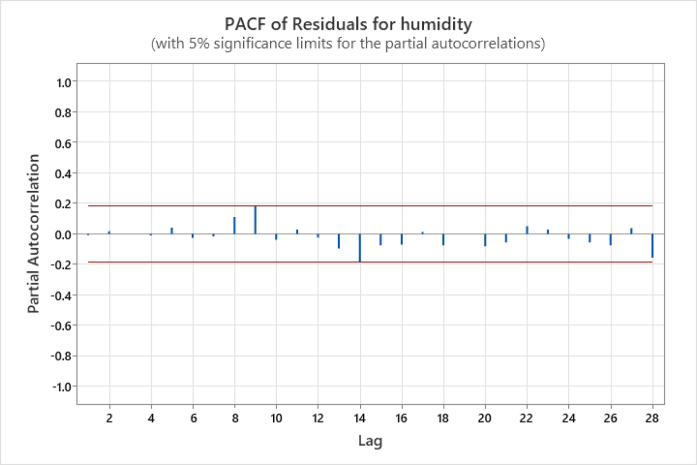


Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

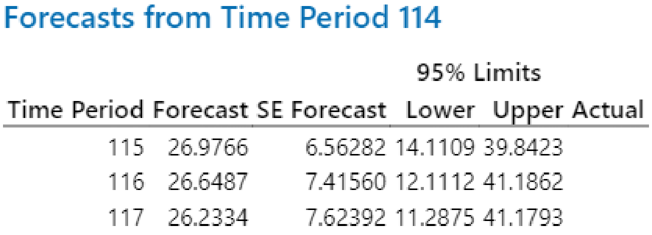
Lag	12	24	36	48
Chi-Square	6.43	15.37	22.64	37.71
DF	9	21	33	45
P-Value	0.696	0.804	0.912	0.771

- To determine whether the residuals are independent, compare the p-value to the significance level for each chi square statistic. Usually, a significance level (denoted as α or alpha) of 0.05 works well. If the p-value is greater than the significance level, you can conclude that the residuals are independent and that the model meets the assumption.
- In these results, the p-values for the Ljung-Box chi-square statistics are all greater than 0.05



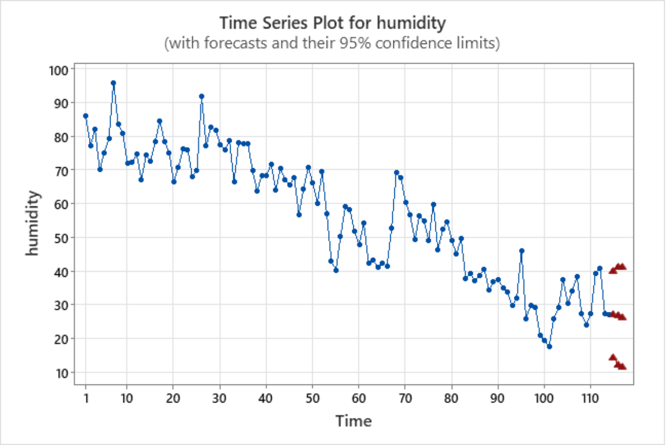


None of the correlations for the autocorrelation function of the residuals or the partial autocorrelation function of the residuals are significant. We can conclude that the model meets the assumption that the residuals are independent.



Forecasts from Time Period 114

			95% Limits		
Time Period	Forecast	SE Forecast	Lower	Upper	Actual
115	26.9766	6.56282	14.1109	39.8423	
116	26.6487	7.41560	12.1112	41.1862	
117	26.2334	7.62392	11.2875	41.1793	



CONCLUSION:

From the trend analysis it can be concluded that meantemp is expected to rise and humidity is going to decrease.

From the ACF and PACF plots for differences it is concluded that ARIMA(1,1,1) model is best fit for number of humidity when value of MS is taken into consideration and for mean temp ARIMA(0,1,0) is the best fit.

Tools recommended for Analysis:

Time Series Analysis in Minitab.

Minitab provides several functionalities that make it suitable for time series analysis:

1. Time Series Plotting: Minitab allows users to easily create time series plots to visualize trends, seasonality, and other patterns in the data. These plots are essential for understanding the underlying structure of the time series.

2. Forecasting Tools: Minitab offers forecasting tools such as exponential smoothing, ARIMA modelling, and seasonal decomposition, enabling users to generate forecasts based on historical data.

3. Autocorrelation Analysis: Minitab includes tools for autocorrelation analysis, which helps identify dependencies and patterns within the time series data.

4. Statistical Analysis: Minitab's wide range of statistical tools can be applied to time series data to perform hypothesis testing, identify outliers, and assess the significance of trends and patterns.

Overall, Minitab provides a comprehensive suite of tools and capabilities that support the analysis of time series data, making it a justified choice for users looking to perform such analyses.