

OPIM 5671 - Data Mining and Business Intelligence

"Forecasting Insight: Leveraging Temperature Forecasting for Strategic Planning"

**Group 1**

Sai Harish Boddu

John Gebhard

Tejaswini Gudisi

Tejesh Hawaligi

Srinivas Minnakanti

Bhuvaneshwar Reddy Muddam

**Table of Contents**

Introduction ……..................................................................................................................3

Problem Statement …….......................................................................................................3

Dataset Description ..............................................................................................................3

Data Manipulation and Imputation .......................................................................................4

Data Preparation in SAS.......................................................................................................5

Time Series Exploration........................................................................................................6

Exponential Smoothing.........................................................................................................10

ARIMA..................................................................................................................................12

Independent Variable Pre-Whitening....................................................................................14

Best Model …………............................................................................................................24

Best Fit Model Description ..................................................................................................25

Best Fit Model Generation Process – After Pre-whitening...................................................27

Model Fit Comparison...........................................................................................................36

Summary and Inferences........................................................................................................37

References..............................................................................................................................38

**Introduction**

Understanding and predicting future weather patterns is a vital function for planning, resource allocation, disaster preparedness, and overall public safeness. It’s estimated that the annualized benefit of investing in public weather forecasts is about $31.5 billion1. In addition to being financially beneficial, knowing future temperatures is extremely practical. Industries such as construction, agriculture, or energy are incredibly reliant on forecasting future weather patterns. Predicting severe weather events like tornadoes, hurricanes, or heavy rains can truly mean the difference between life and death. Additionally, understanding the long-term effects of changing temperature conditions may empower lawmakers to enact legislation or develop strategies to mitigate trends in the data. Understanding this importance was the reasoning behind our decision to construct a model using a small sample size of data taken from Houston, Texas.

**Problem Statement**

Our goal for this project is to construct an accurate yet relatively low complex model to predict the average temperature for the Houston region. Our data consists of over 4 years of collected weather metrics with a granularity of days. Our plan of attack consists of analyzing components such as trends, seasonality, and potential anomalies to develop a comprehensive model capable of predicting the ever-changing temperature patterns. We will do this by exploring various modeling techniques, including, but not limited to, exponential smoothing, ARIMA, and ARIMAX. In order to assess each model's accuracy, metrics such as MAPE and RMSE will be compared. For models with similar accuracy, AIC and SBC values will be used to choose the best model relative to its complexity.

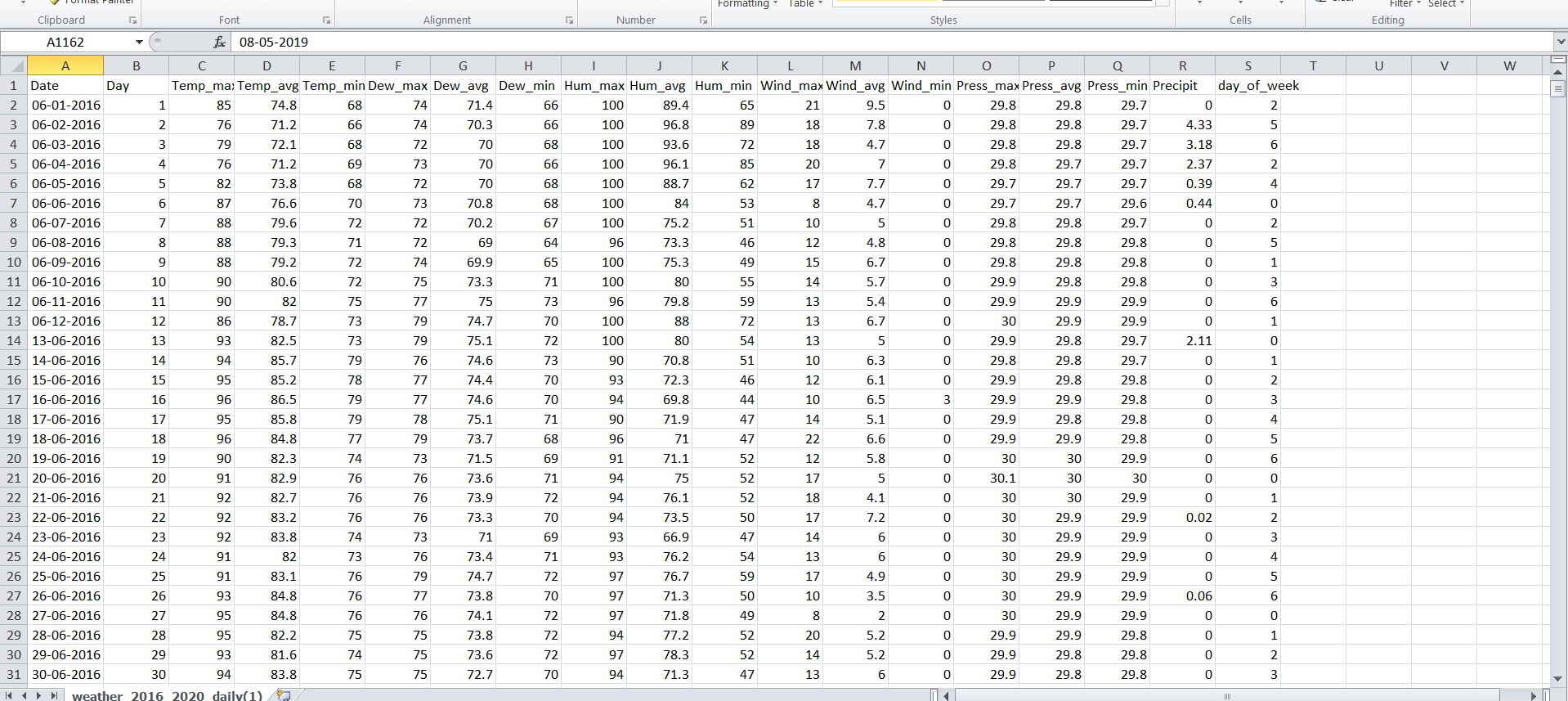
**Dataset Description**

Our data is derived from Kaggle2. The dataset contains 1,552 rows from June 1, 2016, to August 30, 2020. The table consists of 19 columns, including the time value of “Date.” Additionally, two fields track the day of the month and the day of the week. There is a field containing the precipitation amount for each day. The remaining 15 columns contain five weather-related variables: temperature (our dependent variable), dew, humidity, wind, and pressure. Each variable has a daily minimum, maximum, and average value. A fully exhaustive list including each variables unit is below:

* Date with a granularity of days
* Day of the week (1-7)
* Day of the month (1-31\*)
* Precipitation in Inches
* Temperature in ℉ (Min, Max, and Avg)
* Dew point in ℉ (Min, Max, and Avg)
* Wind in Miles per Hour (Min, Max, and Avg)
* Humidity in % (Min, Max, and Avg)
* Pressure in Hg (Min, Max, and Avg)

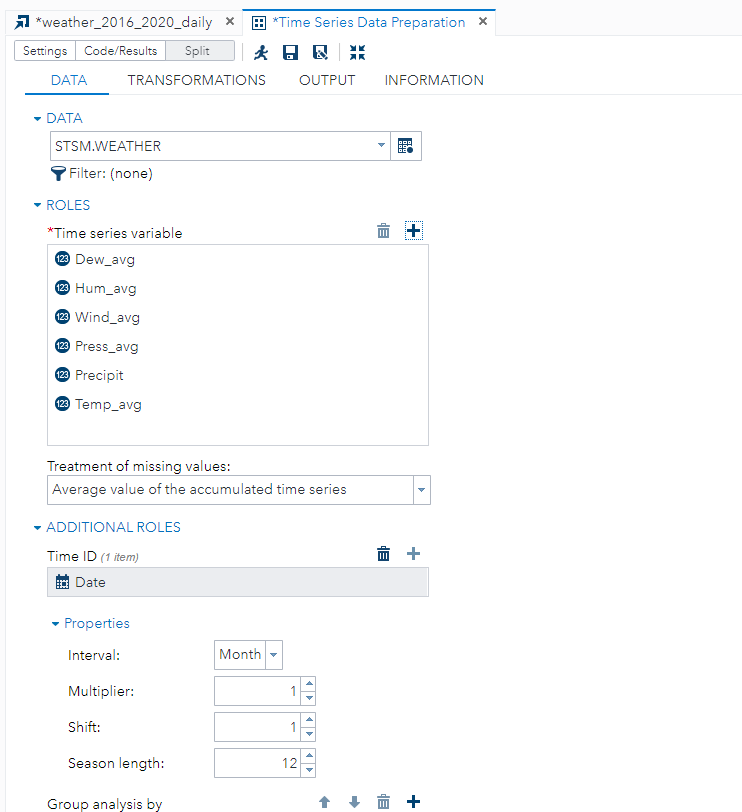
**Data Manipulation and Imputation**

When encountering inconsistent date formats in the dataset, we standardized all dates to the MM/DD/YYYY format for consistency and ease of analysis. This ensured uniformity in the dataset and facilitated further data processing and analysis without ambiguity.

****

**Data Preparation in SAS**

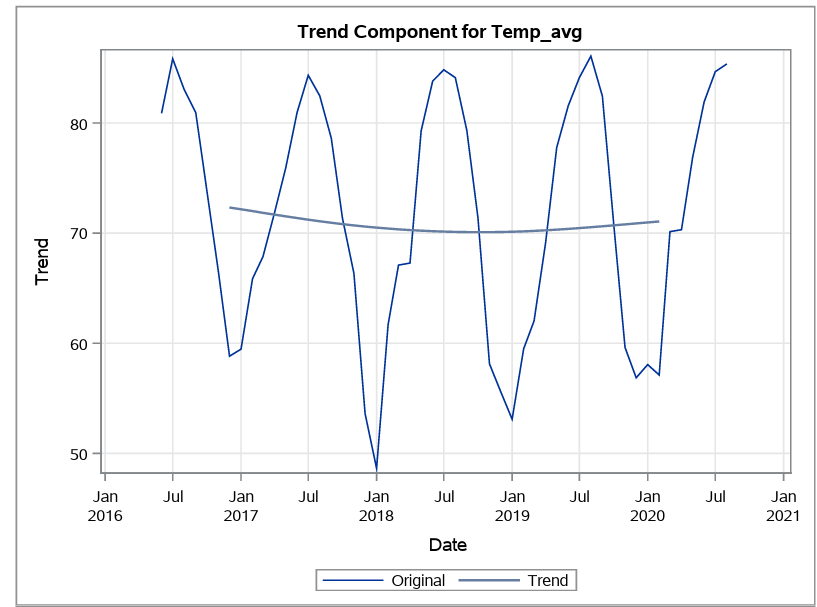
We utilized the data aggregation capabilities within SAS, specifically leveraging the time series data preparation module, to streamline and consolidate our daily dataset into monthly data. This process allowed us to condense and organize our data effectively for further analysis and exploration.

****

**Time Series Exploration**

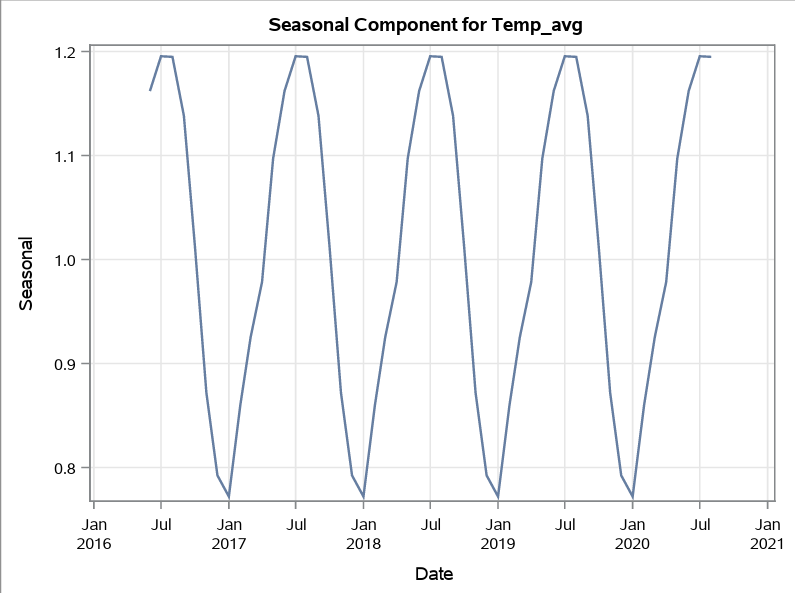
Below are the insights after performing the time series exploration, with the dependent variable as ‘Temp\_avg’ and the Time ID as ‘Date’.

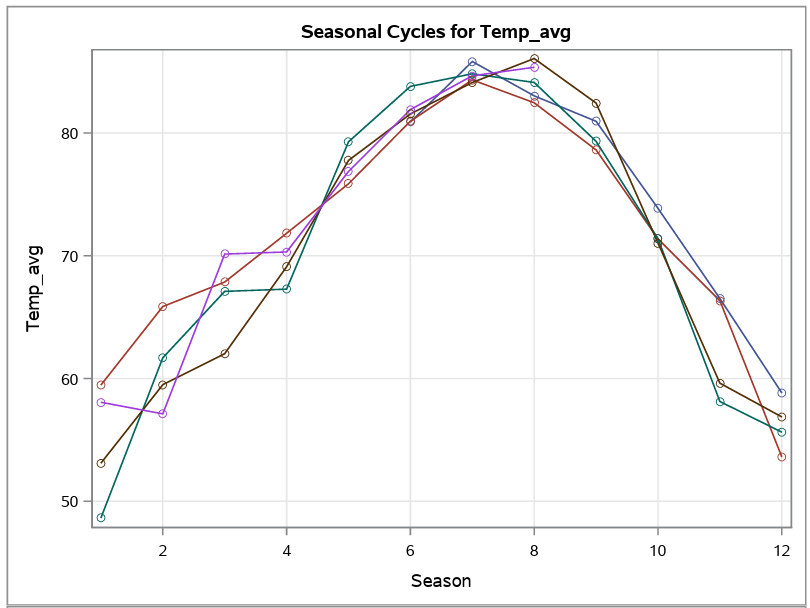
**Trend Component:**

****

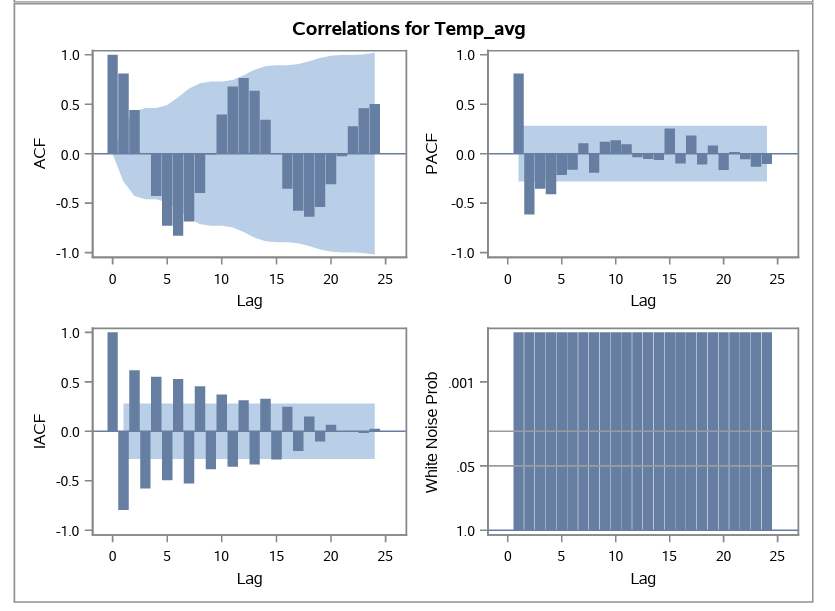
The analysis of the weather report dataset indicates the absence of a trend component, as evidenced by a nearly horizontal line when plotting the average temperature over time. Without a discernible upward or downward trend, the data appears to fluctuate around a constant mean value. This suggests that over the observed period, there is no systematic increase or decrease in average temperatures. Instead, the temperatures remain relatively stable without any long-term directional movement. The absence of a trend component simplifies the modeling process, allowing the focus to remain on capturing seasonal variations and other cyclic patterns present in the data. This characteristic of a straight horizontal line reinforces the stationary nature of the time series, where the statistical properties of the data remain constant over time, facilitating more accurate forecasting and analysis.

**Seasonality Component:**

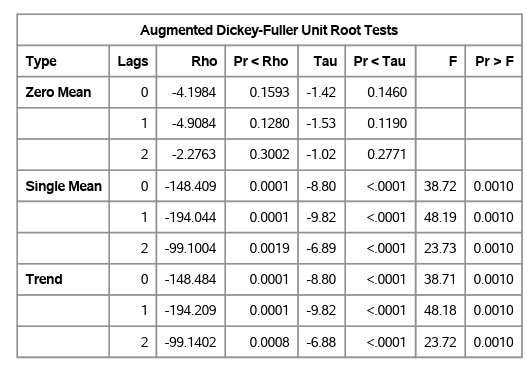
****

****

The weather dataset analysis reveals distinct spikes in average temperatures during July and August, marking the peak of summer. These spikes reflect seasonal temperature fluctuations driven by longer days and increased solar radiation. Their recurrence underscores the predictable nature of seasonal patterns, which is crucial for sectors like agriculture and tourism. Forecasters enhance weather predictions by accurately capturing these trends, aiding in resource management and decision-making. Accounting for such seasonal variations improves our understanding and adaptation to changing weather dynamics, emphasizing the significance of incorporating seasonal adjustments in temperature forecasting models.

****

The autocorrelation function (ACF) and partial autocorrelation function (PACF) analyses of the weather forecasting dataset reveal significant spikes in autocorrelation up to lag 6 for ACF and lag 4 for PACF. These spikes indicate temporal solid dependencies in the data, particularly within these lag intervals, suggesting that past observations significantly influence current temperatures. Furthermore, the inverse autocorrelation function (IACF) shows a continuous decline, reaching 0 by lag 25, indicating a diminishing correlation between observations as the lag increases. This suggests that beyond lag 25, the influence of past observations on current temperatures becomes negligible. Moreover, the white noise probability distribution suggests that the residuals are randomly distributed, indicating that any remaining patterns in the data have been adequately captured by the model, further enhancing its predictive capability for weather forecasting.

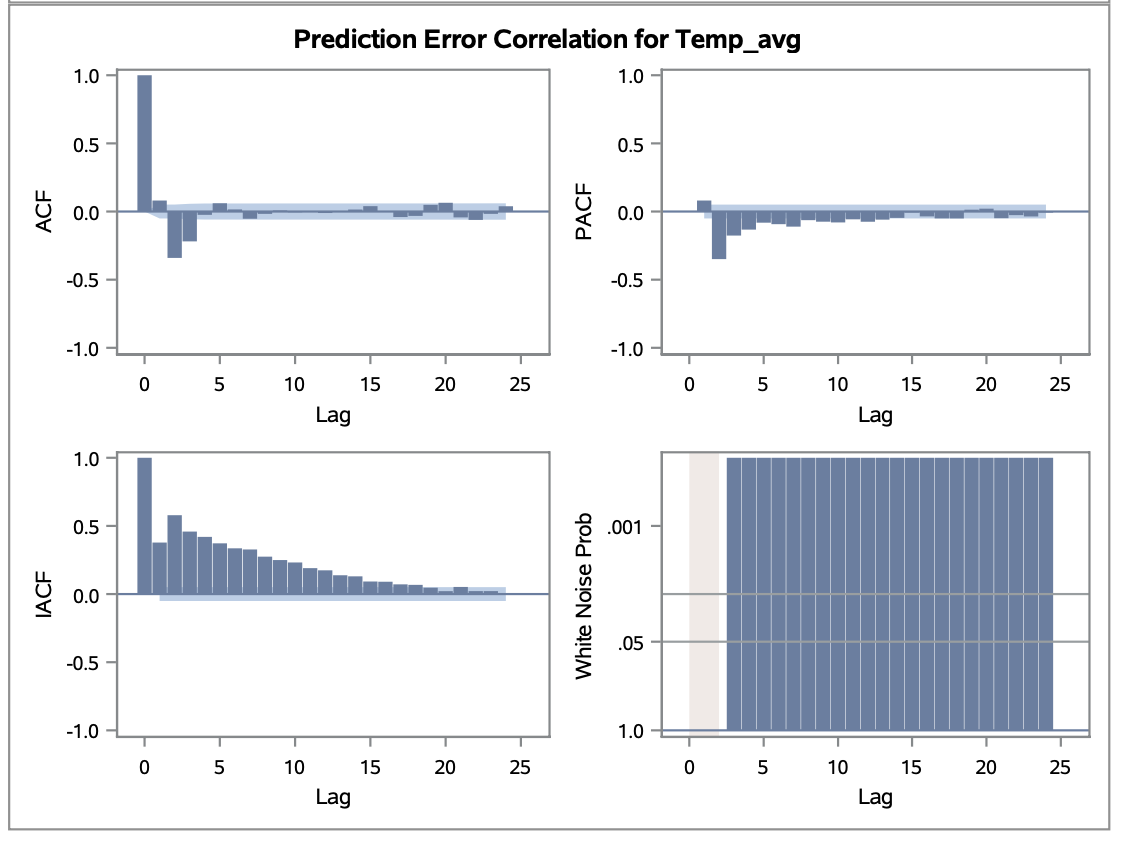
****

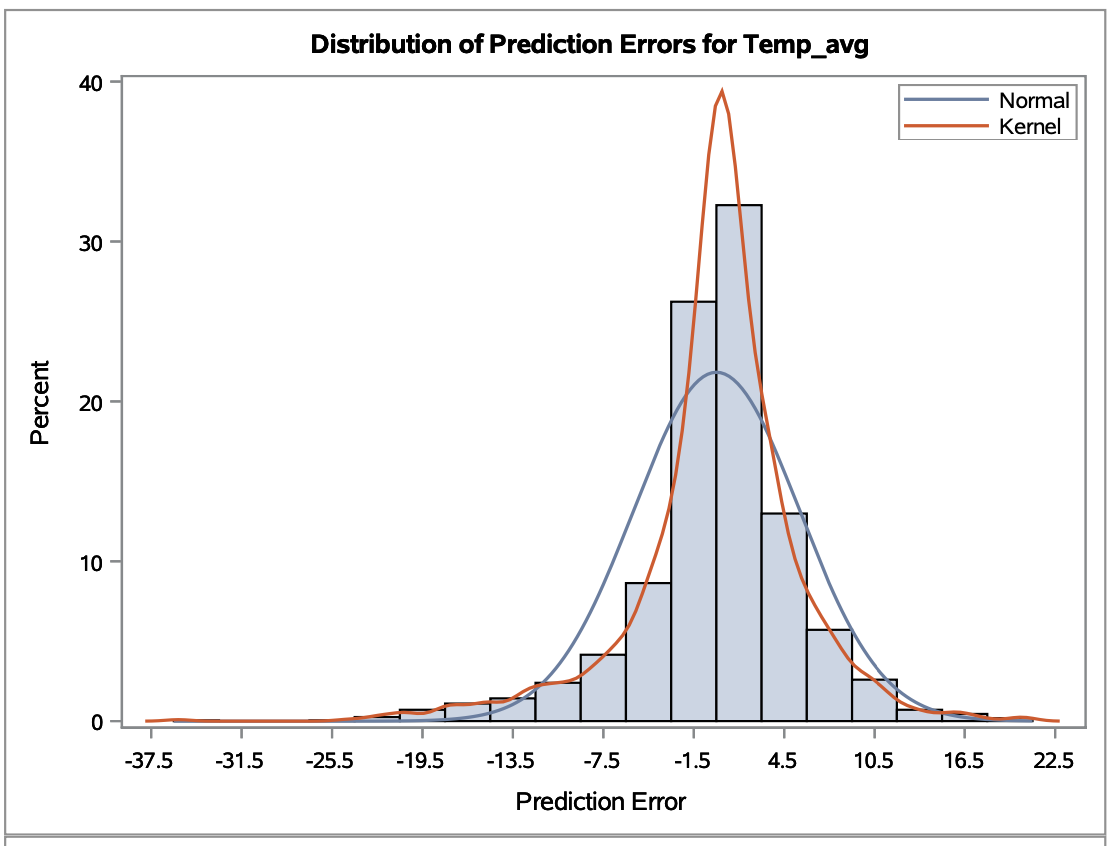
The augmented Dickey-Fuller test results in the rejection of the null hypothesis for the weather forecasting dataset, indicating evidence of stationarity in the data. This suggests that the average temperature series does not possess a unit root, signifying stability over time. Stationarity is critical for time series analysis, facilitating accurate modeling and forecasting. The observed stationarity implies that the statistical properties of the temperature series remain consistent over time, enabling reliable predictions and insights into seasonal variations and other temporal patterns.

**Performing Exponential Smoothing**

In instances where there is no discernible trend but evident seasonality in time series data, such as in temperature forecasting, initiating the modeling process with additive seasonal exponential smoothing proves advantageous. This method effectively captures the recurrent seasonal patterns while maintaining simplicity. By emphasizing recent observations and adjusting for historical seasonal fluctuations, additive exponential smoothing provides accurate forecasts without the need to account for long-term trends. Its iterative updating of smoothed values and seasonal components enables the model to adapt to evolving data patterns, resulting in reliable temperature predictions. Through validation using metrics like Mean Absolute Percentage Error (MAPE), the model’s efficacy in capturing seasonal variations can be confirmed, paving the way for robust temperature forecasting.

**Additive Seasonal Exponential smoothing:**

****

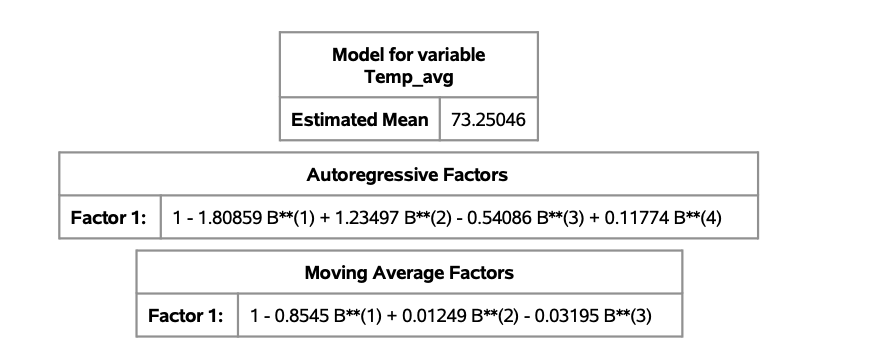
****

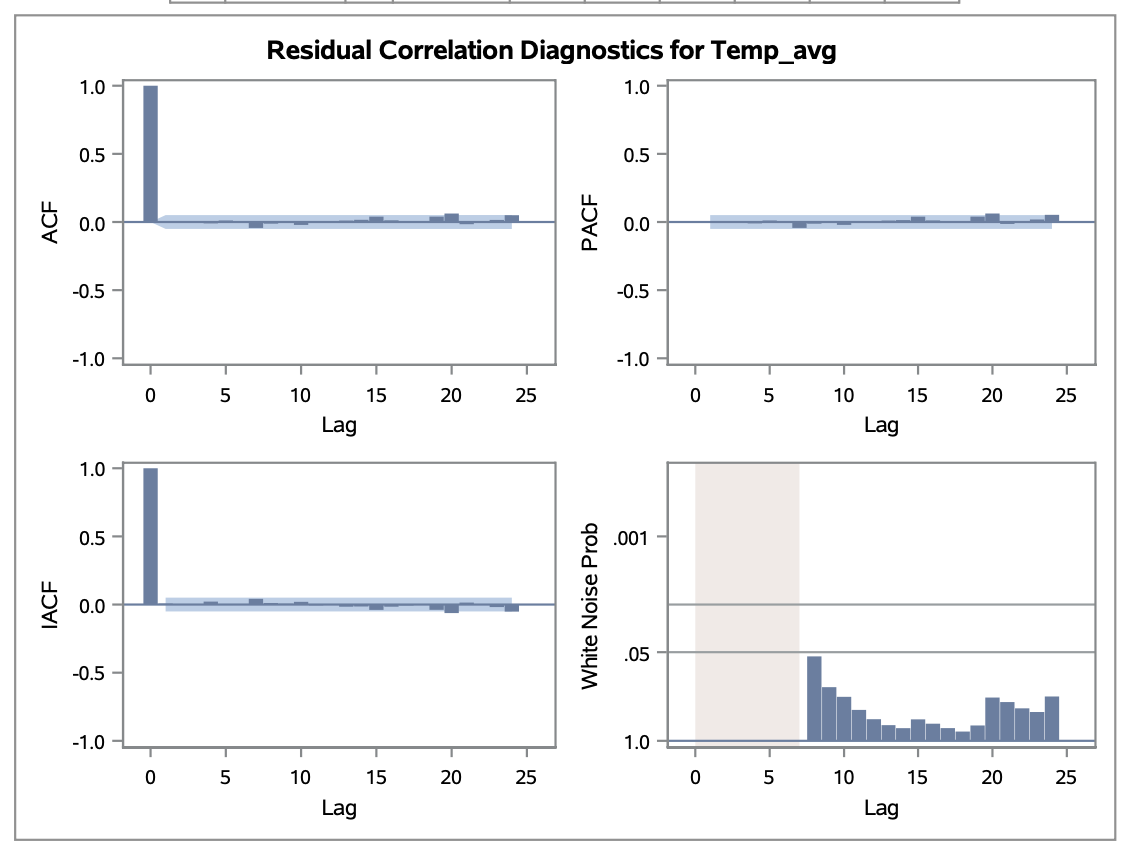
**Observation:**

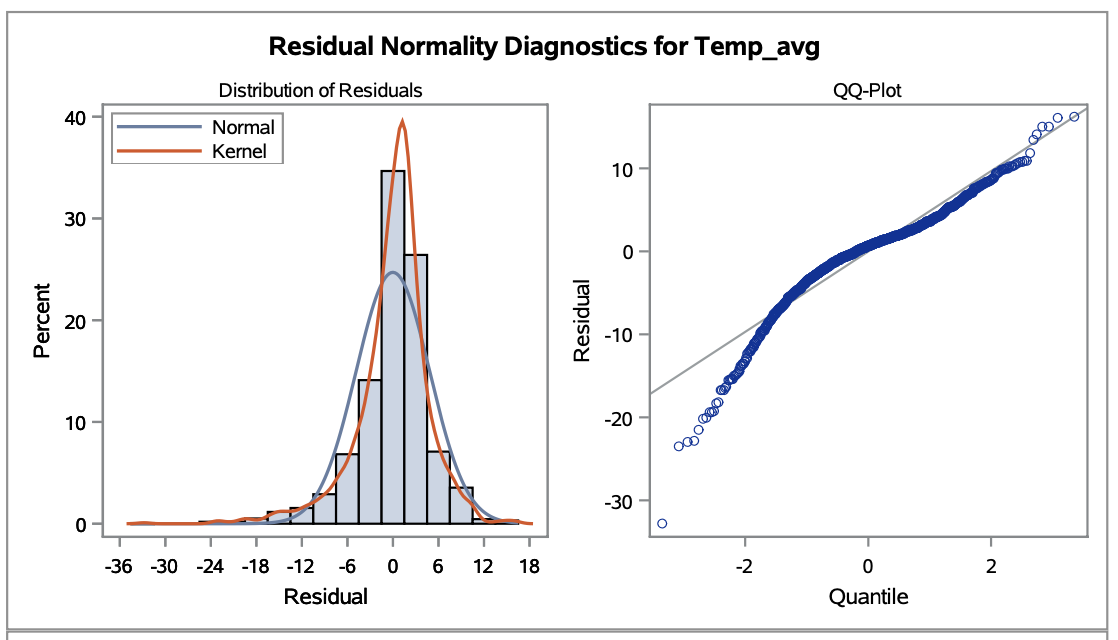
In exponential smoothing for time series forecasting, poorly distributed white noise undermines predictive accuracy by disrupting the model's ability to capture underlying data dynamics. Significant spikes at lag 2 in autocorrelation and partial autocorrelation functions suggest strong correlations, hindering accurate forecasts.

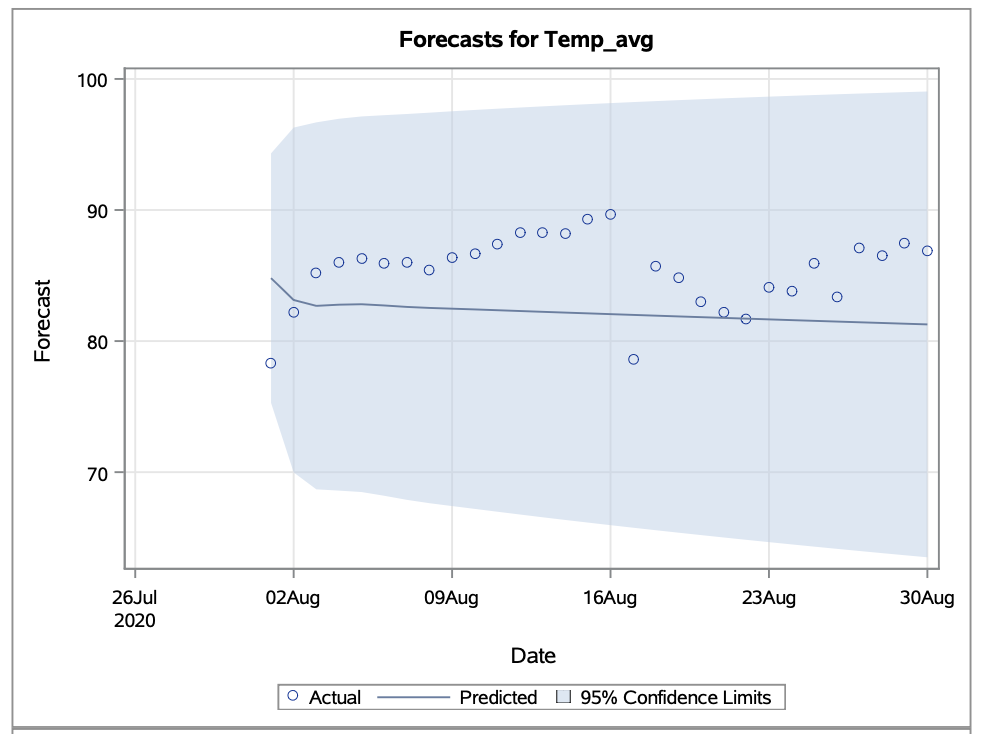
In response to increased data complexity, we're incorporating ARIMA alongside exponential smoothing for comprehensive forecasting. Through iterative refinement, we aim to enhance predictive accuracy despite white noise and data intricacies.

**ARIMA:**









**Observation:**

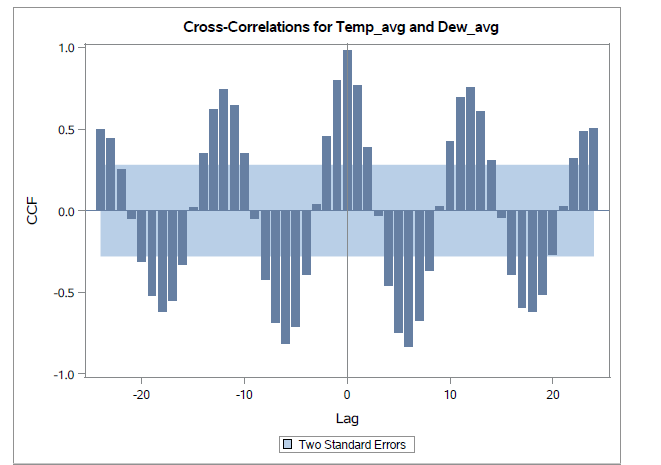
We implemented ARIMA (4,3) modeling, successfully reducing white noise to below 0.05, indicating effective distribution. Notably, residual correlation graphs (ACF, PACF, IACF) showed no significant spikes within the significant region. Additionally, residual normality diagnostics confirmed a normal distribution. This comprehensive analysis validates the efficacy of ARIMA (4,3) in mitigating white noise and accurately capturing underlying data dynamics, enhancing forecasting precision. However, to refine our model further for greater accuracy, we aim to explore ARIMAX.

For performing ARIMAX, we initially perform **Pre-whitening**.

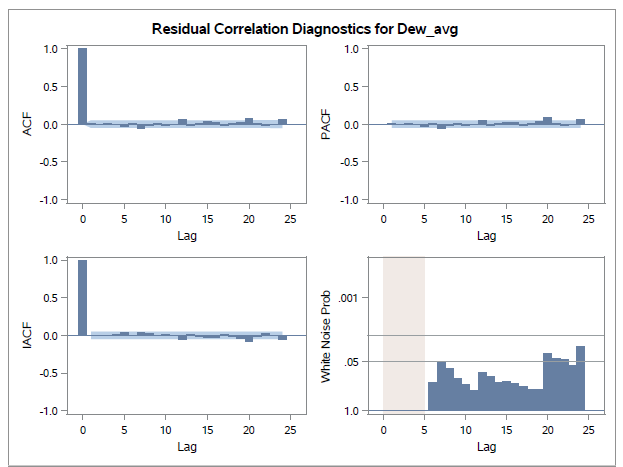
**Independent variable Pre-whitening:**

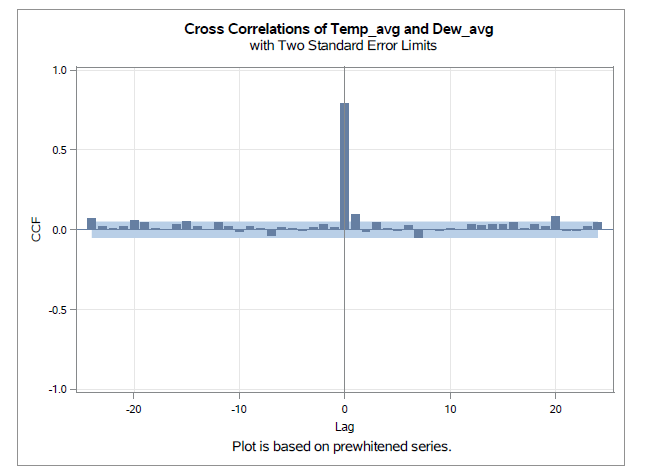
**Dew\_Avg Prewhitening:**

There was a cross-correlation between Temp\_avg and Dew\_avg, but it was difficult to tell if the autocorrelation in both time series contributed to this relationship or if it was a true association. It's possible that the correlation value found at this stage doesn't fairly depict the underlying relationship.



After applying pre-whitening techniques to eliminate autocorrelation, we revisited the cross-correlation analysis between the average temperature and the average dew point. At this stage, any observed cross-correlations are more likely to reflect a true relationship since the influence of autocorrelation from within each individual data series has been neutralized. This refined analysis offers a clearer view of the interaction between temperature and dew point, free from the distortions of internal series correlation.

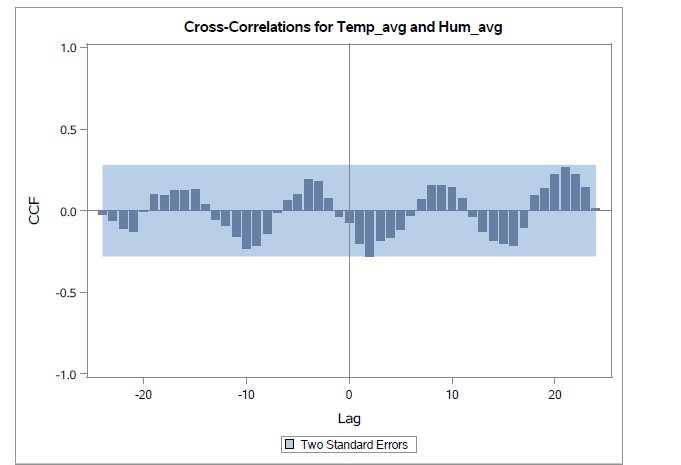




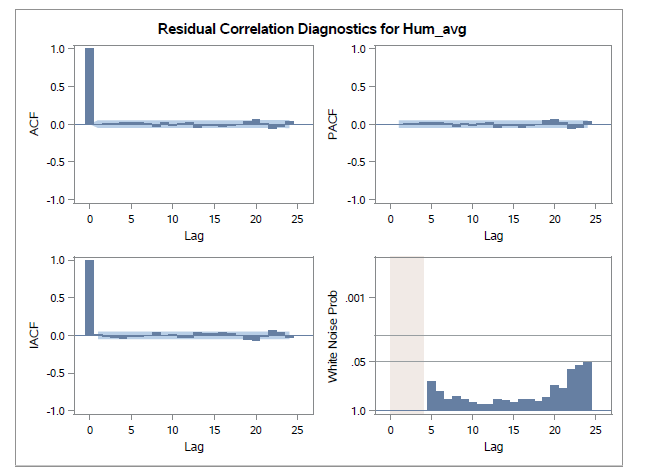
Our Pre-whitening analysis shows significant cross-correlation at lag 0 between Temp\_avg and Dew\_avg.The above cross-correlation plot provides insights into the temporal dynamics between temperature and dew point, which can be useful for weather forecasting and climate modeling.

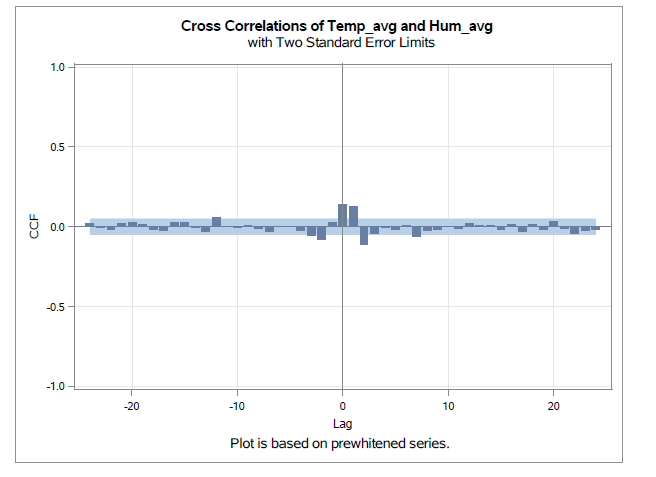
**Hum\_Avg Prewhitening:**

The plot displays the cross-correlations between average temperature and average humidity (denoted as Temp\_avg and Hum\_avg). The graph indicates that there appears to be a lag-dependent pattern of association, with certain lags exhibiting greater connections than others. Additional statistical analysis would be necessary to determine the precise nature of the correlations—whether positive or negative—and their significance.



We reevaluate the cross-correlation after the Pre-whitening procedure, which eliminates autocorrelation. As autocorrelation does not skew any cross-correlations identified at this point, they are more likely to be legitimate.

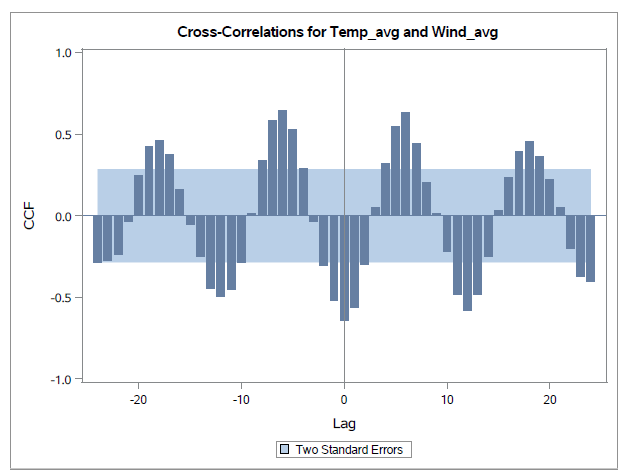




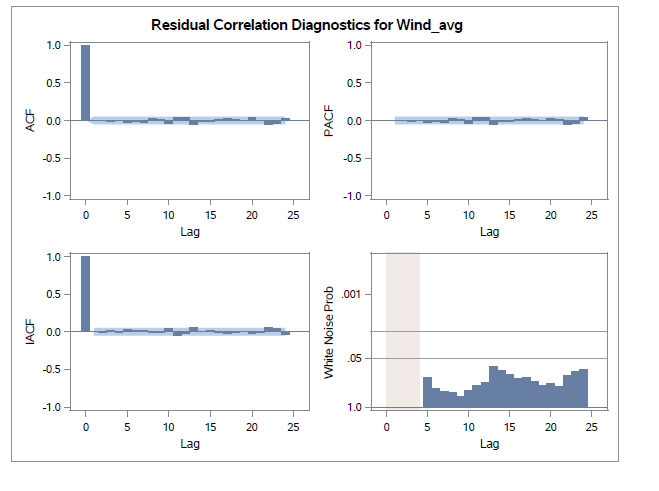
Our Pre-whitening analysis revealed significant cross-correlations at lag 0 between Temp\_avg and Hum\_avg, indicating immediate and short-term relationships.

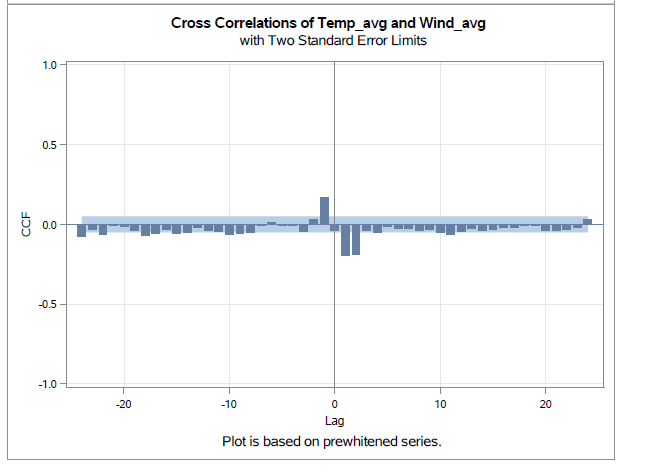
**Wind\_Avg Prewhitening:**

The cross-correlation plot shows the relationship between average temperature (Temp\_avg) and average wind speed (Wind\_avg) with lag times. The bars represent the cross-correlation coefficients at different lags. This initial comparison of average temperature and wind, without pre-whitening, helps us to grasp their fundamental correlation, while also taking into account the existing autocorrelation within each time series dataset.



After applying Pre-whitening, that removes autocorrelation we reevaluate the cross-correlation. At this point, any identified cross-correlations are more credible, as the influence of autocorrelation has been minimized, allowing for a truer understanding of the relationship between the variables.

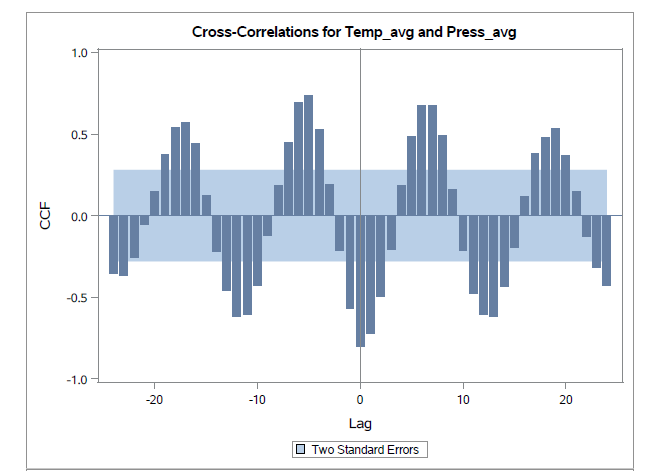




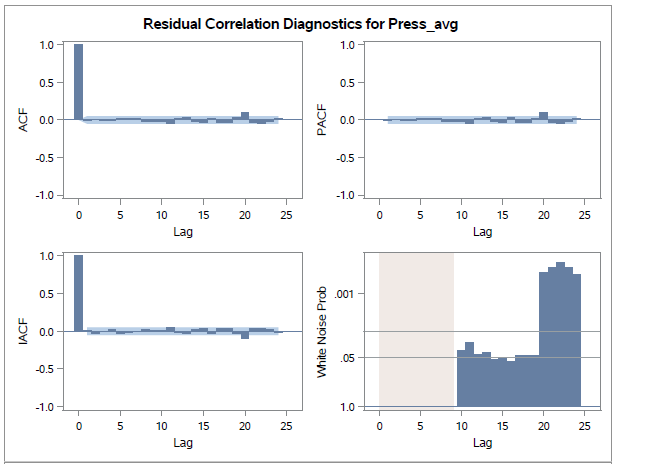
Following Pre-whitening, we found that there was a significant cross-correlations between Temp\_avg and Wind\_avg at lag 1. This implies that there is a significant relationship between Temp\_avg and the Wind\_avg with a lag effect.

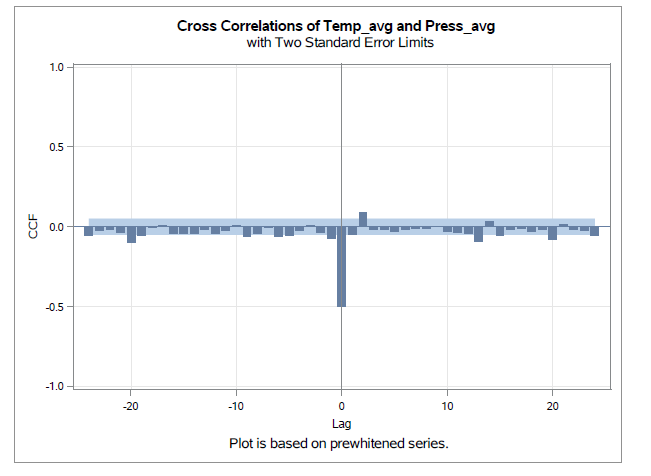
**Press\_AvgPrewhitening:**

Initially we computed the cross correlation between Temp\_Avg and Press\_Avg without using pre-whitening. Taking into account the impacts of the inherent autocorrelation in each time series, this helps in our understanding of the two variables' raw relationship.



We reassess the cross-correlation after the Pre-whitening process, which eliminates autocorrelation between Temp\_avg and Press\_avg. As autocorrelation does not skew any cross-correlations identified at this point, they are more likely to be real.

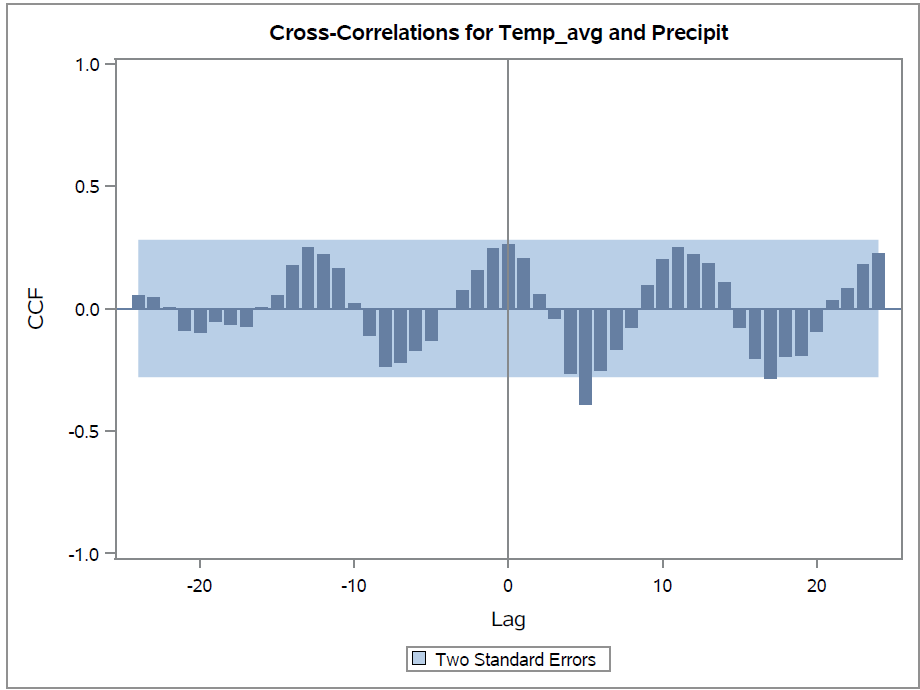




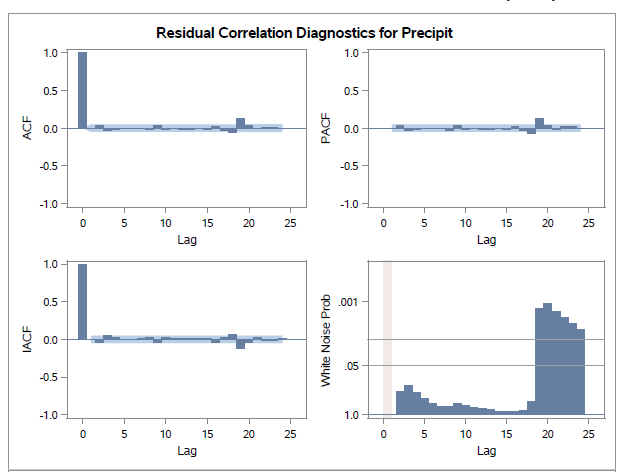
Our Pre-whitening analysis shows significant cross-correlation at lag 0 between Temp\_avg and Dew\_avg. The above cross-correlation plot provides insights into the temporal dynamics between temperature and Pressure, which can be useful for weather forecasting and climate modeling.

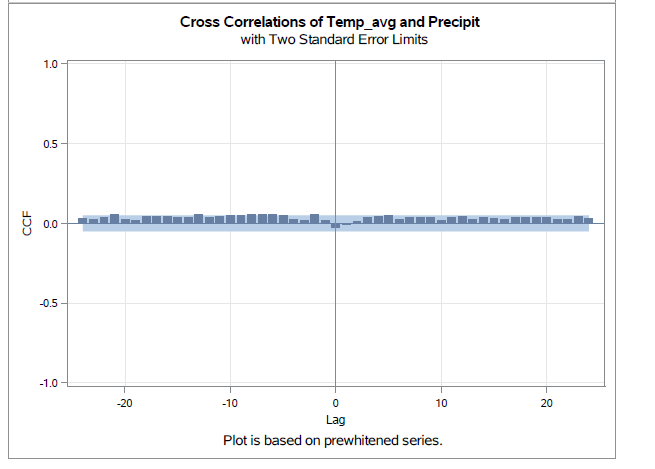
**Precipit Prewhitening:**

In the initial analysis, there appeared to be a correlation between Temp\_avg and Precipitation. Yet, this perceived relationship could be skewed by the individual time series' autocorrelation, complicating the ability to determine clear-cut conclusions



After conducting the pre-whitening process, which eliminates autocorrelation from the time series data, we reassess the cross-correlation between the variables. This step ensures that any detected cross-correlations are more reliable, as they are less likely to be influenced by the internal correlations of the individual series, providing a clearer insight into the genuine interactions between temperature and precipitation as depicted in the plot.





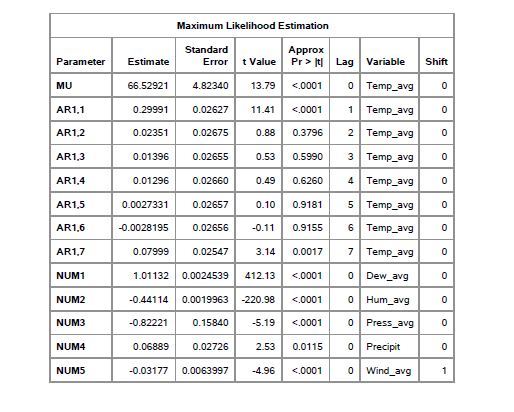
After applying Pre-whitening, we see there is no significant cross-correlations between Temp\_avg and Precipitation. This suggests that there may not be a significant relationship between Temperature average and Precipitation within the analyzed time period.

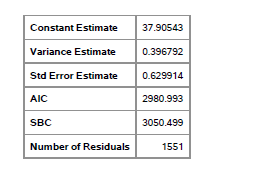
While the absence of significant cross-correlations may indicate that Precipitation does not have a substantial impact on Temperature Average during the observed time frame.

**Best Model:**

We chose to employ a more advanced modeling technique called ARIMAX to improve the accuracy of our predictive analysis. Before implementing ARIMAX, we conducted pre-whitening procedures on all independent variables: dew, humidity, pressure, wind speed, and precipitation. Notably, wind speed showed a lag 1 effect, while the other independent variables demonstrated lag 0 effects except the precipitation.

Afterward, our finalized ARIMAX model, labeled as (7,0,0: 0,0,0), had a sophisticated parameter configuration. Impressively, this model produced predictions with minimal errors, as measured by the Mean Absolute Percentage Error (MAPE). Furthermore, comprehensive evaluation metrics confirmed the superior performance of our ARIMAX model compared to simpler alternatives and other models within the ARIMAX framework based on accuracy, AIC, and SBC scores.





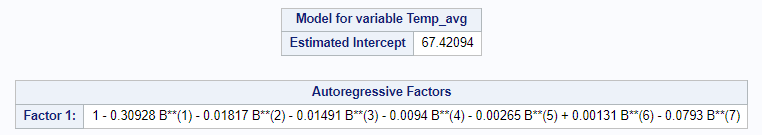
**Best Fit Model Description:**

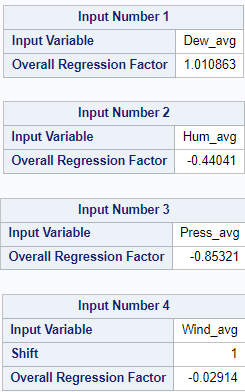
**ARIMAX Components (7,0,0):**

* **p (Autoregressive Order):** 7
* **d (Integration Order):** 0
* **q (Moving Average Order):** 0

In our model, we use ARIMA to understand how temperatures change over time. The AR part looks at how past temperatures affect current ones, while the MA part looks at how past errors in temperature predictions influence the current temperature.

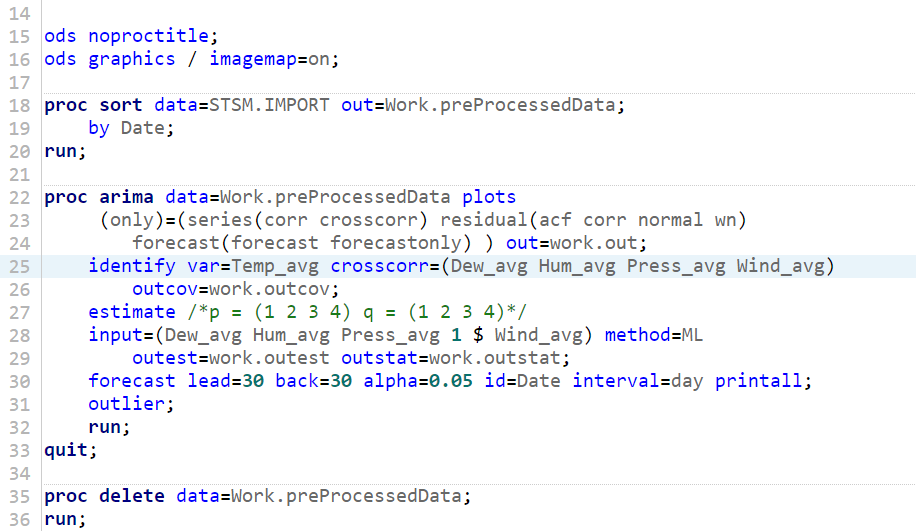
The autoregressive factors are given by:



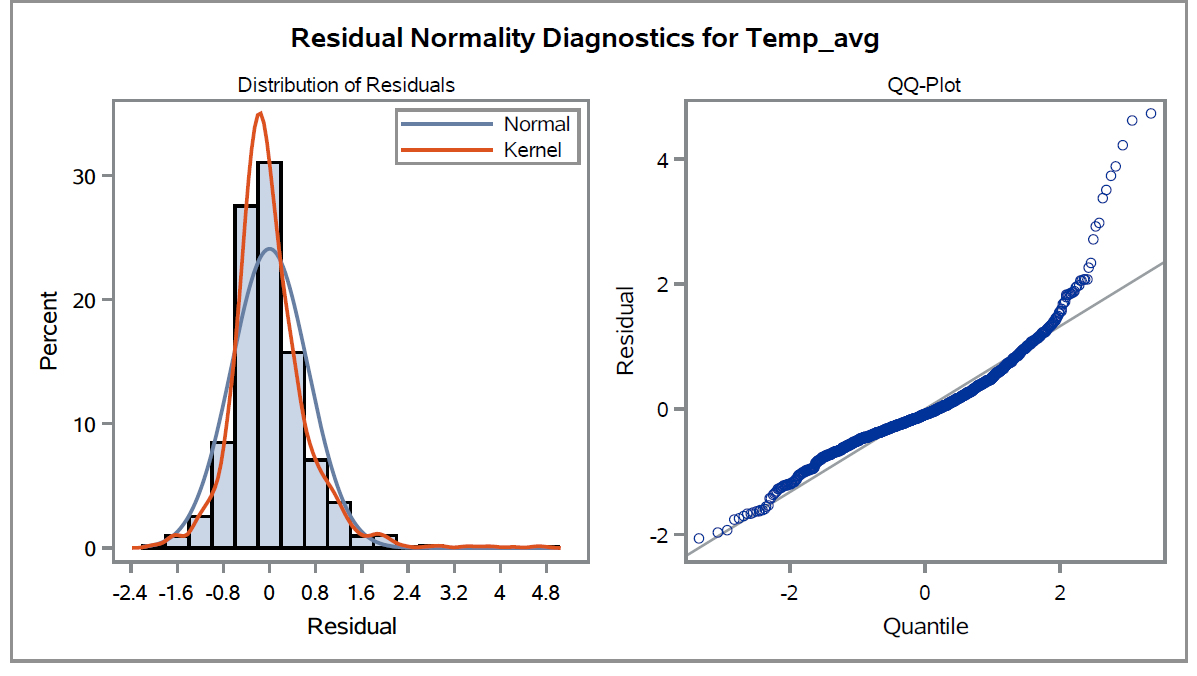


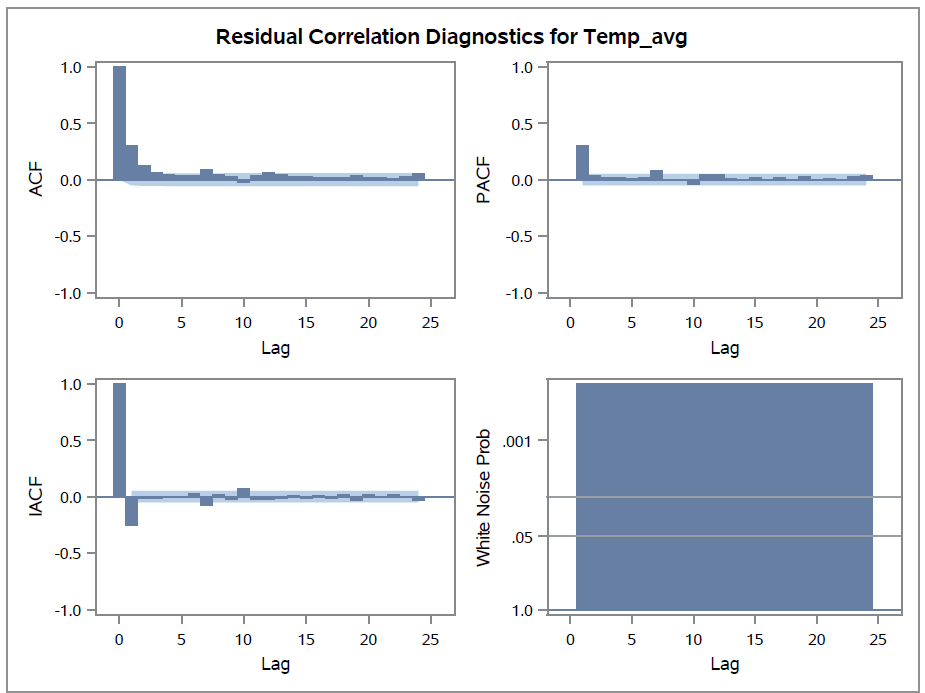
**Process of deriving the best-fit model- After Pre-Whitening:**

**1. Initial Model with ARMAX(0,0):**



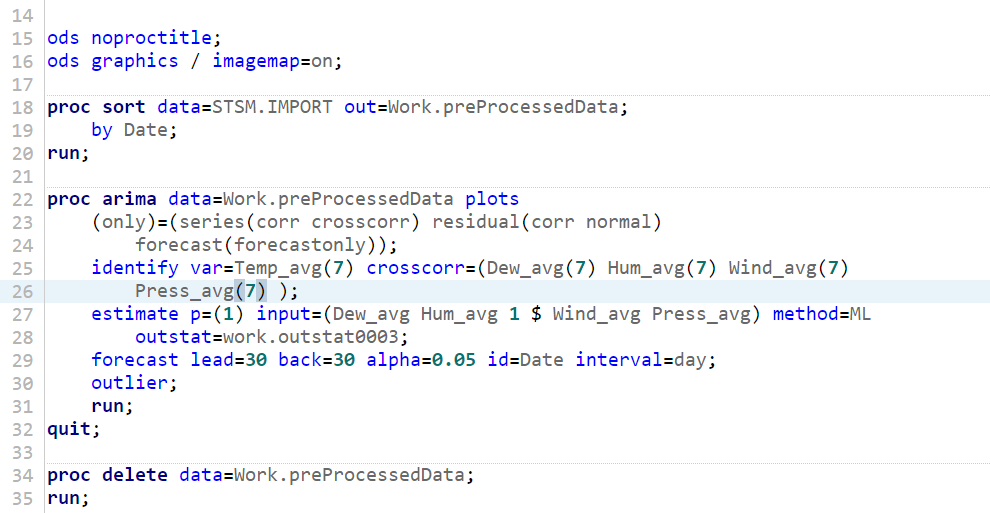
The independent variable Wind is considered with a lag 1 in line 28 of the code as we observed in the Pre-Whitening.



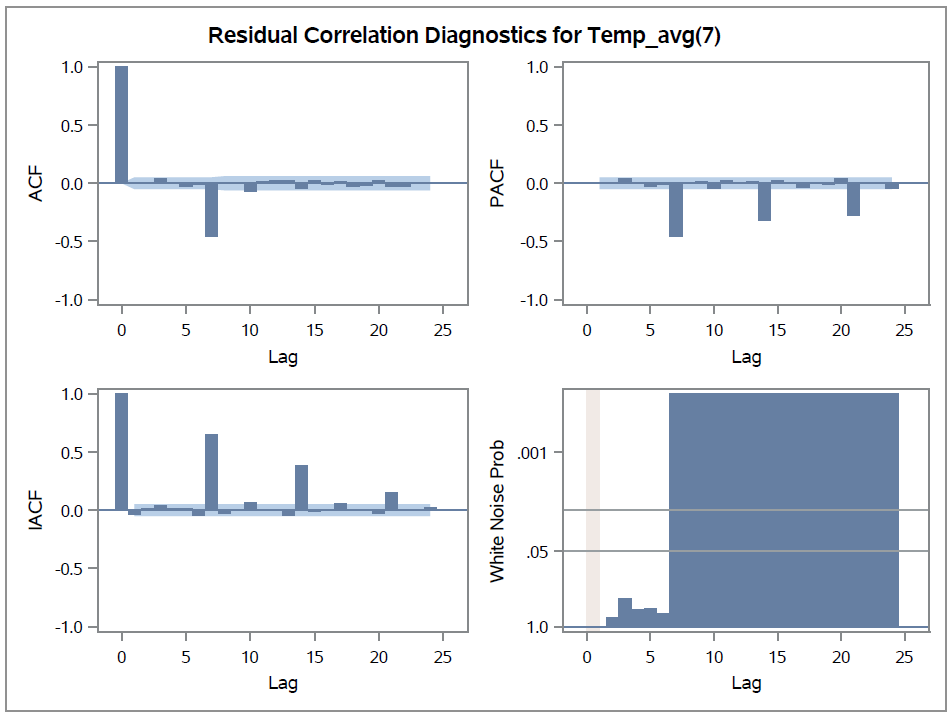


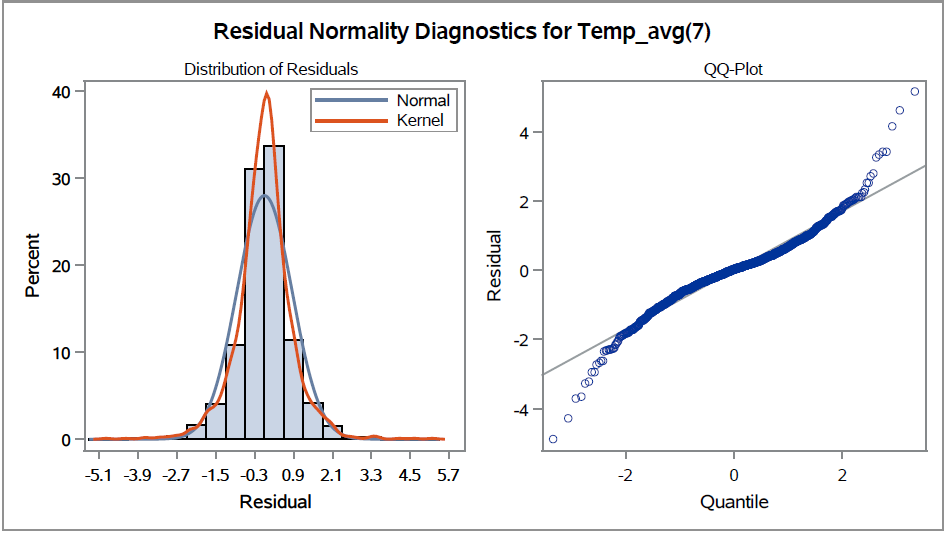
When we created an ARMAX(0,0) model, taking into account all independent variables and the lag effect on wind, it failed the white noise test. Additionally, examining the autocorrelation function (ACF), we noticed a gradual decrease, with significant spikes at lag 1 in both the partial autocorrelation function (PACF) and the inverse autocorrelation function (IACF). Furthermore, there was a notable spike at lag 7 across all ACF, PACF, and IACF plots. Consequently, for our subsequent model, we opted for p = 1 and D = 1.

**2. ARIMAX (p=1,d=0,q=0 ; P=0,D=1,Q=0)**



**Output:**

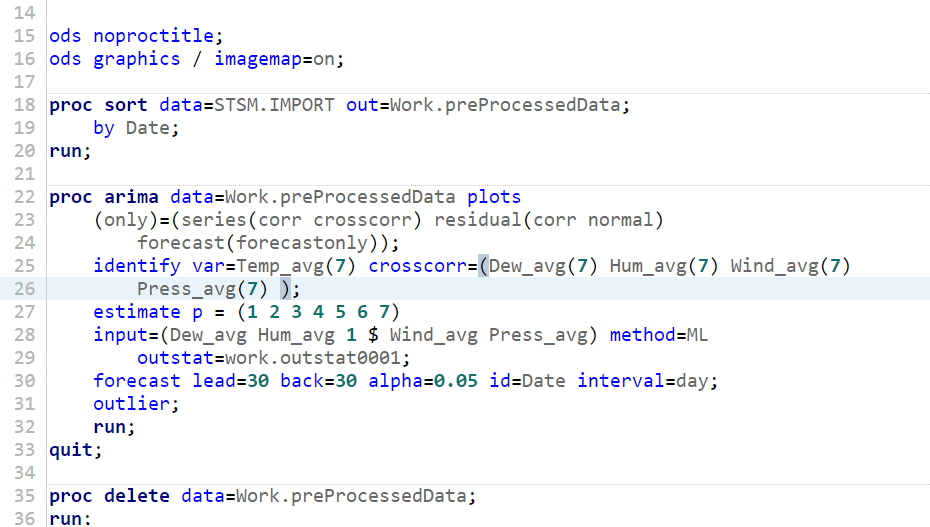




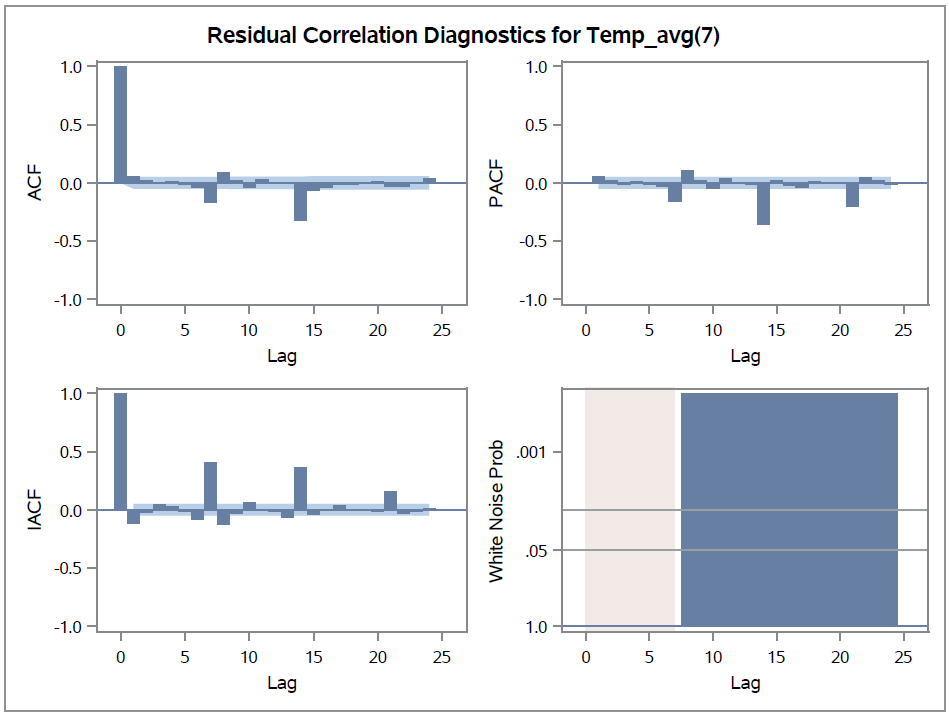
**Observation:**

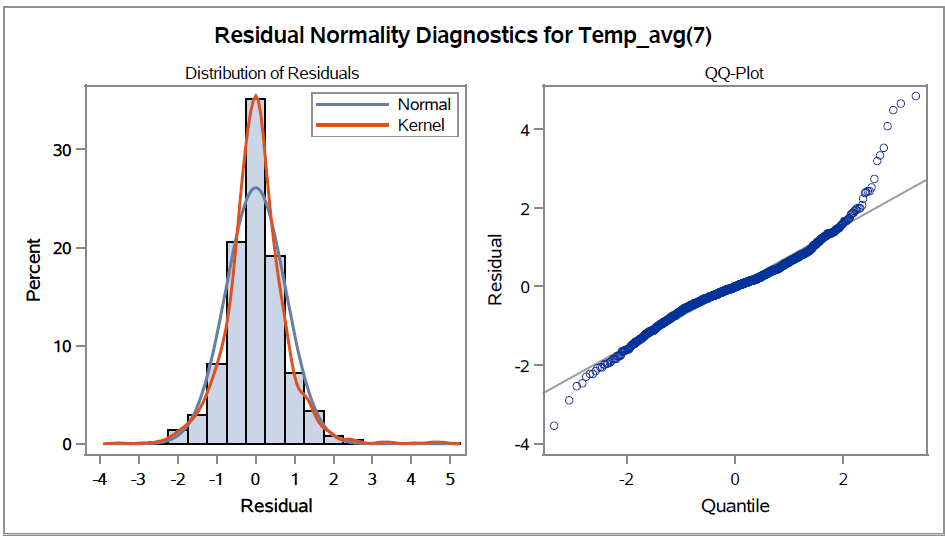
Despite adjusting the model parameters to p=1 and D=1, the white noise probability test continued to indicate failure, suggesting that there is still an observable signal within the data that can be modeled. Additionally, the significant spike at lag 7 in the partial autocorrelation function (PACF) indicated a persistent pattern in the data. Therefore, we decided to further refine our model by increasing the autoregressive order to p=7 while maintaining D=1.

**3. ARIMAX (p=7,d=0,q=0; P=0,D=1,Q=0)**

****

**Output:**

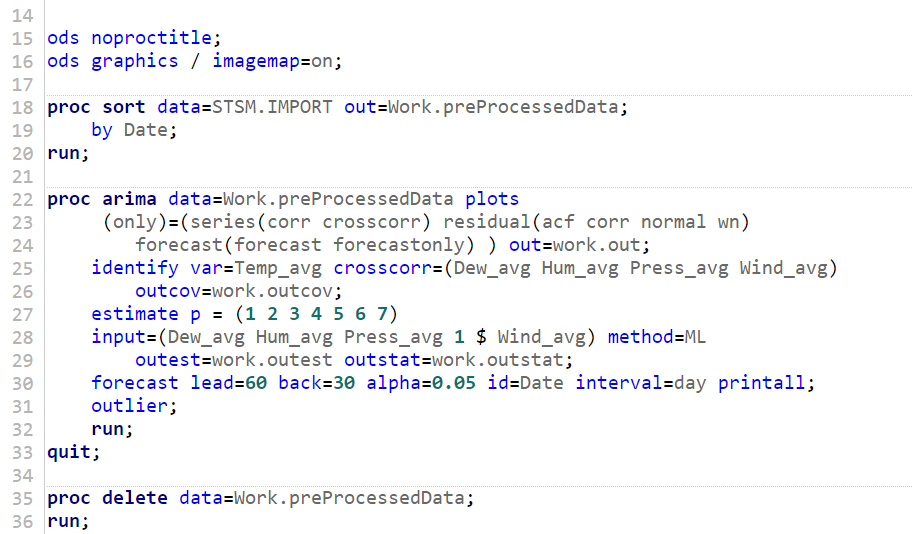
****

****

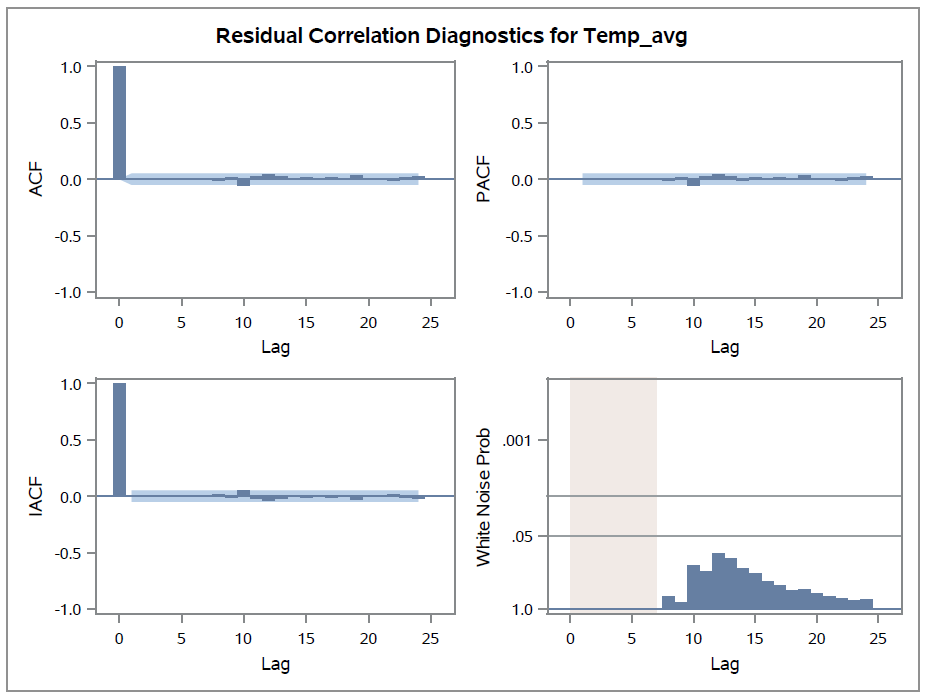
**Observation:**

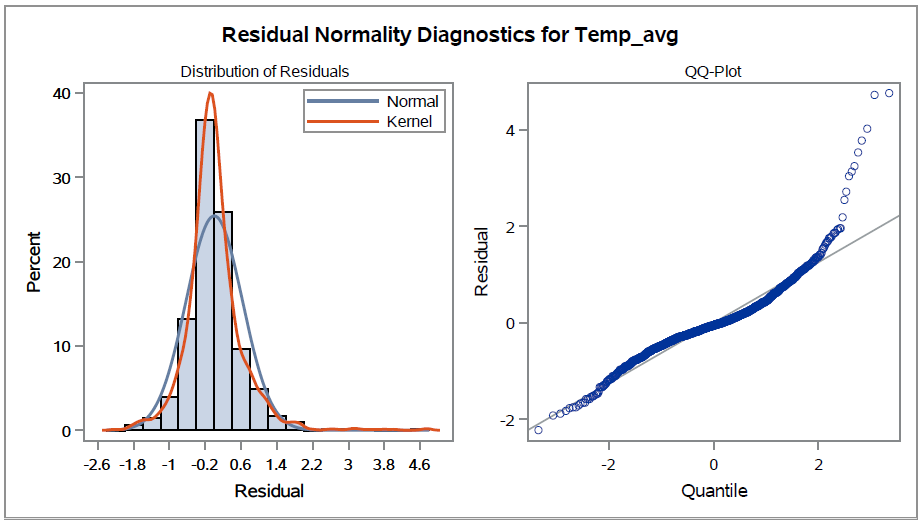
Despite our efforts, the white noise probability test persisted in indicating failure, suggesting ongoing signal presence within the data. Notably, significant peaks at various lags, particularly lag 7, were observed in both the partial autocorrelation function (PACF) and inverse autocorrelation function (IACF) plots. Consequently, we proceeded with further modeling by eliminating seasonality (setting D=1) and focusing solely on the autoregressive component with p=7. We then re-ran the model to refine our analysis.

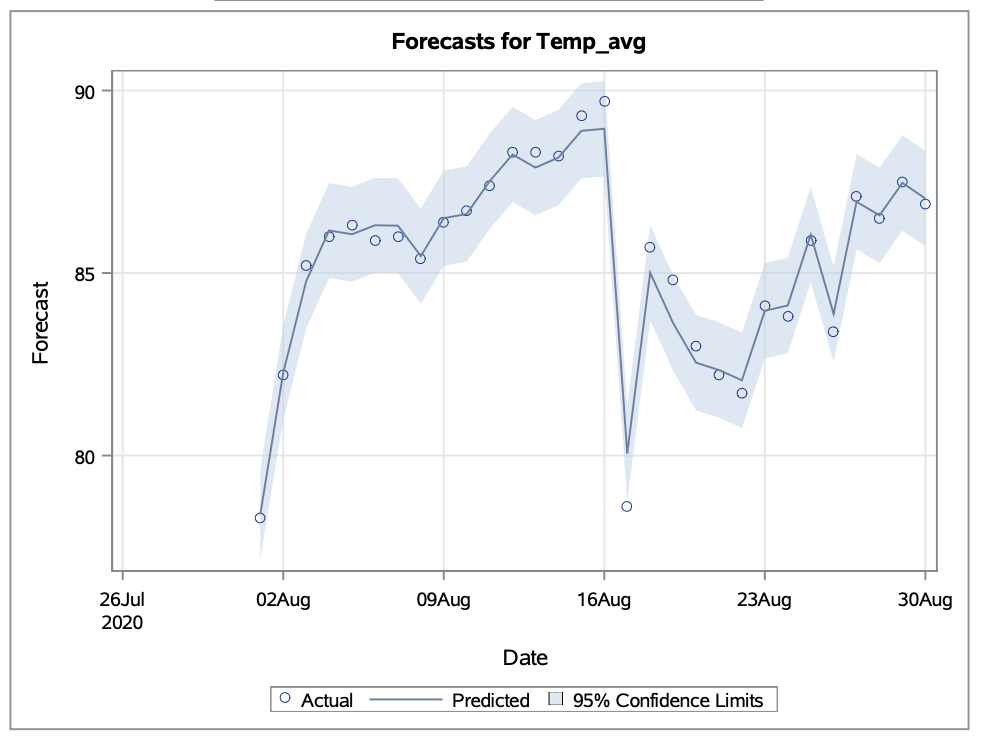
**4. ARIMAX (p=7,d=0,q=0 ; P=0,D=0,Q=0)**

****

**Output:**

****



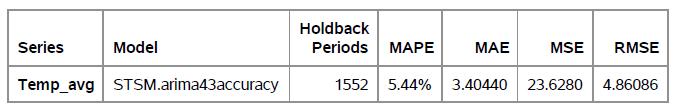


**Observation:**

The model at a significance level of 7 has proven its adequacy through the successful completion of the white noise test, affirming its ability to accurately discern underlying data patterns with minimal residual noise. Additionally, our concluding forecasting model has exhibited exceptional effectiveness, boasting an impressively low error rate of 0.38%. And from the residual graph, it is evident that the error is normally distributed.

**Model Fit Comparison:**

Now, we calculate the accuracy and fit statistics of the models that we have tried. Although the ARIMAX(1,0,0;0,1,0) and ARIMAX(7,0,0;0,1,0) did not pass the White Noise Test, let's have a look at the accuracy and fit statistics.









We see a minute difference in the Error metrics in all the above models. Clearly, the ARMAX(7,0,0;0,0,0) is the best model with the least error which implies highest accuracy.

Let's have a look at AIC and SBC scores also.

| **Model** | **p** | **d** | **q** | **P** | **D** | **Q** | **AIC** | **SBC** | **MAPE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Additive seasonal ESM | - | - | - | - | - | - | 5245.14 | 5255.82 | 5.94 |
| ARIMA | 4 | 0 | 3 | 0 | 0 | 0 | 9320 | 9362.78 | 5.44% |
| ARIMAX | 1 | 0 | 0 | 0 | 1 | 0 | 3907.51 | 3944.91 | 0.47% |
| ARIMAX | 7 | 0 | 0 | 0 | 1 | 0 | 3579.45 | 3648.90 | 0.41% |
| ARIMAX | 7 | 0 | 0 | 0 | 0 | 0 | 2980.99 | 3050.50 | 0.38% |

We can clearly see that the Model ARMAX(7,0,0;0,0,0) which has the best MAPE also has the best fit statistics i.e, the AIC and SBC values of this model are low compared to all other models.

**Summary and Inferences:**

In conclusion of our analysis, we were quite pleased with our constructed model. Our starting data demonstrated strong seasonality as well as having no gaps to fill in. We had to do some slight cleaning of the data by reformatting the date syntax. Additionally, we removed extraneous columns including min and max components, day of the week, and day of the month. Our dataset also passed the ADF test without the need for differencing. The strongest ARIMAX model achieved a very high accuracy score with relatively low AIC and SBC scores. Our residuals were distributedly quite normally with some right skewness. The model showed zero lag effect for dew, humidity, pressure, and precipitation. It did however demonstrate a lag 1 effect for wind. This means that today’s windspeed would influence the average temperature for tomorrow. There are many complex factors to weather, but strong winds certainly play a part. Significant winds might blow in warmer or colder air from a neighboring region. One component of our data that we did not observe was trend. This makes sense given the very small sample size that was used, relative to what is needed for a significant trend in temperature. Our model would have looked much differently if that was the case.

**References**

1

<https://www.kaggle.com/datasets/srinuti/residential-power-usage-3years-data-timeseries?select=weather_2016_2020_daily.csv>

2

National Academies of Sciences, Engineering, and Medicine. 2010. When Weather Matters: Science and Services to Meet Critical Societal Needs. Washington, DC: The National Academies Press. <https://doi.org/10.17226/12888>.