OPIM 5671

Data Mining and Business Intelligence

Group 1: Amazon Sales forecasting model

(*Team members: Ekaterina Burkhanova, Neehar Namjoshi, Pooja Shah)*

Table of Contents

1. Executive Summary:

The raw dataset contained historical sales data for 45 Amazon stores located in different geographies. Every store contained 98 departments and each of the departments has got a certain weekly sale. The data contains 3 years of data for the time horizon: 2019-02-05 to 2021-11-01 split at weekly level for 45 Amazon stores. We will be evaluating time - series forecasting to determine sales for one of these stores: Store 1.

By understanding the sales trends, the store can work on optimizing inventories for lean periods and jack up the inventories where the sales would shoot up. Also, stores can run promotional events to gain an uplift in sales in case the projections show feeble sales in the coming weeks.

We have run several ARIMA and ARIMAX models with mean weekly sales as dependent variable, and temperature, fuel price, CPI, unemployment, and holiday as independent variables. We have run ARIMA(0,2), ARIMAX(2,0), ARIMAX(1,0), ARIMAX(2,2), ARIMAX(0,2), ARIMAX(0,0), ARIMAX(1,1), ARIMAX(0,1) models.

We found out that ARIMAX(2,0) is best model because looking at the accuracy and statistics fit, ARIMAX(2,0) has the lowest MAPE at 2.2% than any other models. In addition, it has no significant autocorrelation exists and has high degree of white noise for residuals, whereas other models appear to either have autocorrelation or lower residual white noise.

In summary, ARIMAX(2,0) is the optimal model for business and shows the decline of the sales trend for the forward time horizon. It is in sync with the trend captured for the last 3 years of sales. Given that the macroeconomic indicators show a positive trend, the forecast shows a decline in sales, hence, the recommendation for the business would be investigating the factors leading to drop; department level sales trends can be understood to find critical focus areas; and other indicators like quality, delivery of service, price points can be evaluated to understand if there are other factors influencing the sales.

2. Data Overview:

2.1 Explaining Amazon dataset:

The dataset contains historical sales data for 45 Amazon stores located in different geographies. Every store contains 98 departments and each of the departments and each of the departments has got a certain weekly sale. The data has 3 years of data for the time horizon: 2019-02-05 to 2021-11-01 slip weekly level of 45 Amazon stores.

By understanding the sales trends, the store can work on optimizing inventories for lean periods and jack up the inventories where the sales would shoot up. Also, stores can run promotional events to gain an uplift in sales in case the projections show feeble sales in the coming weeks.

Amazon sales dataset source:

https://data.world/revanthkrishnaa/amazon-uk-sales-forecasting-2018-2021/workspace/project-summary?agentid=revanthkrishnaa&datasetid=amazon-uk-sales-forecasting-2018-2021

2.2 Column summary:

Column Name	Column Description
Store	The store number
Dept	The department number
Date	The week
Weekly_Sales	Sales for the given department in the given store
Temperature	Average temperature in the region
Fuel_Price	Cost of fuel in the region
CPI	The consumer price index
Unemployment	The unemployment rate
IsHoliday	Whether the week is a special holiday week
Mean	The average of weekly sales in each store

2.3 Data cleansing:

We cleaned the data in the Excel sheet, changed the format of the "Date" from dd/mm/yyyy to mm/dd/yyyy, and deleted four columns – "Type", "Size", "Total_MarkDown", and duplicated column "Date" as they were not necessarily needed for the forecasting. The data has 45 Amazon stores, therefore, we used only "1" segment in the "Store" column.

Accumulation function: Also, we have taken aggregation of this weekly data which is for week's mean sales and not total actual sales. This is to ensure that the impact of price changes, promotions, etc. is not factored in and the data is smooth.

3. Data Exploration:

After data cleansing, mean weekly aggregated sales of Amazon store 1 is chosen as the dependent variable. Next, we had 5 independent variables: Temperature, Fuel_Price, CPI, Unemployment, and IsHoliday. All of these variables will be further explored below.

3.1 Exploring dependent variable:

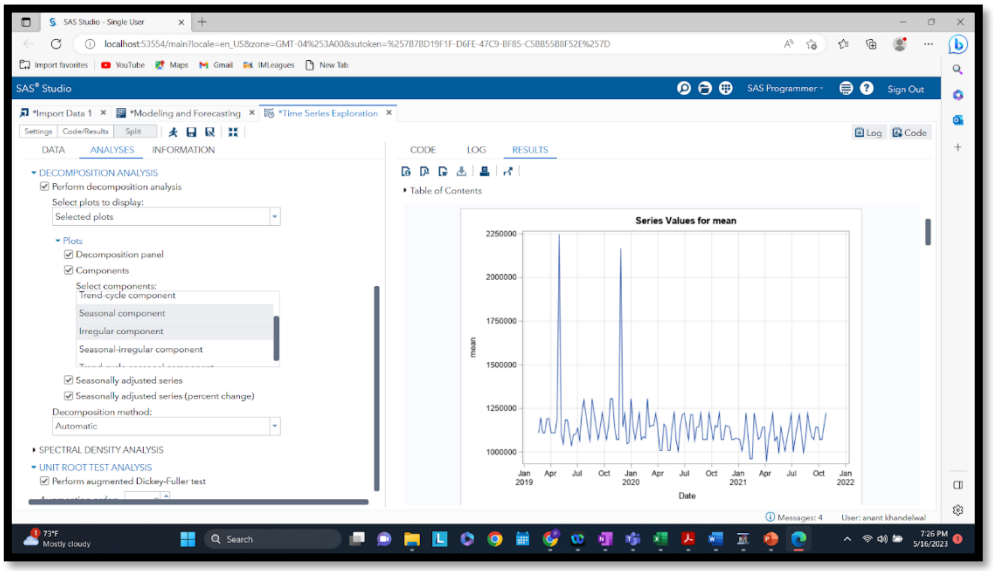


Fig. 1 - Mean Sales Time Series

From just the time series plot above, generally a trend and seasonality component is seen for aggregated weekly mean sales for Amazon Store 1. But further decomposition is required to extract the signals into the different components. Lastly, the time series consists of data points for 3 years broken up into aggregated weekly mean sales.

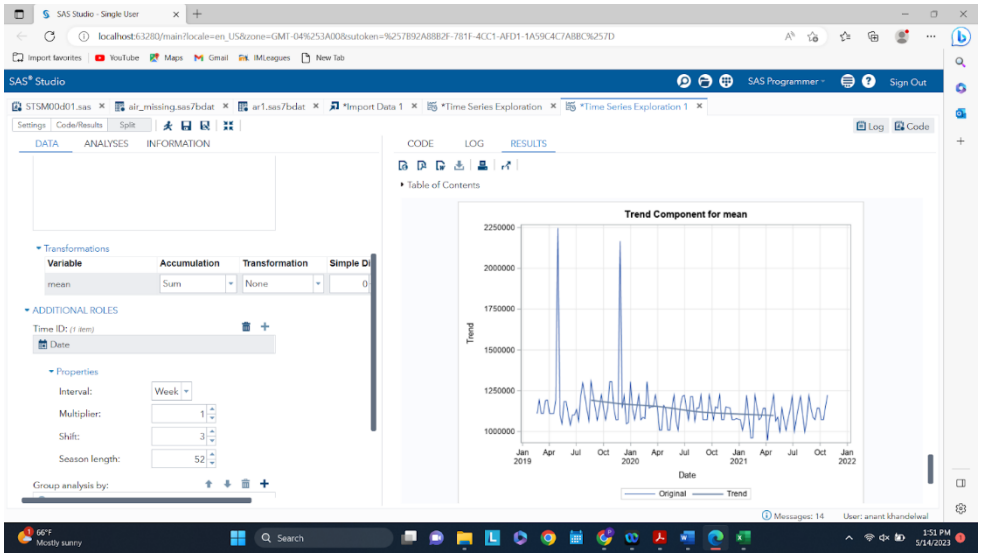


Fig. 2 - Trend Component of Mean Sales

Clearly, a trend is seen in the decomposed trend component of the time series. A declining trend is seen, meaning that from Jan 2019 - Jan 2022, the aggregated weekly mean sales for Amazon store 1 are declining. This is something that will have to be investigated further.

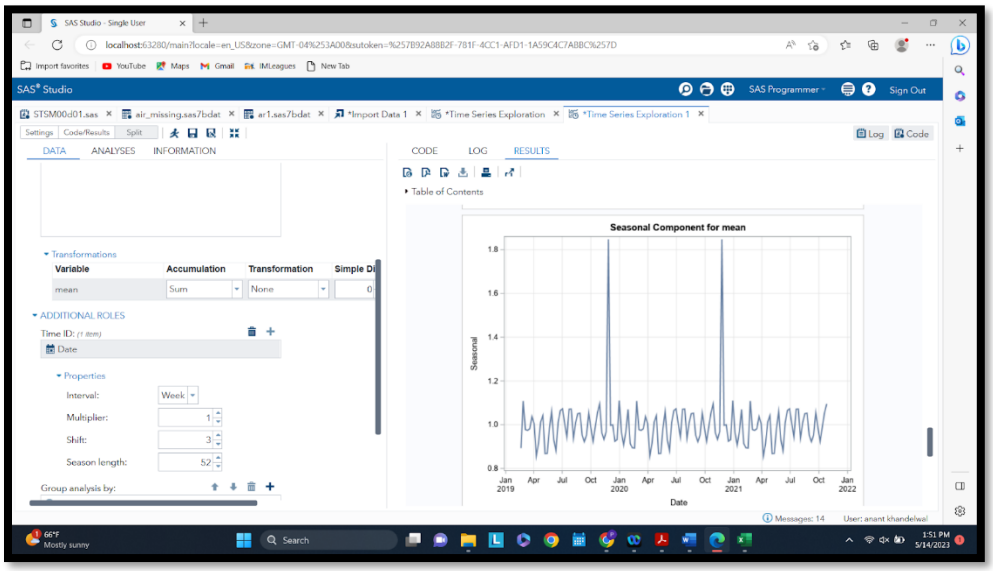


Fig. 3 - Seasonality Component of Mean Sales

Next, above we see the decomposed time series, showing the seasonal component. A relatively clear seasonality is shown in the plot above, with a high point seen between November and December for both the year of 2020 and 2021. A low point is generally seen in April of each year and the overall pattern repeats itself. Therefore, the time series data does consist of a seasonal component.

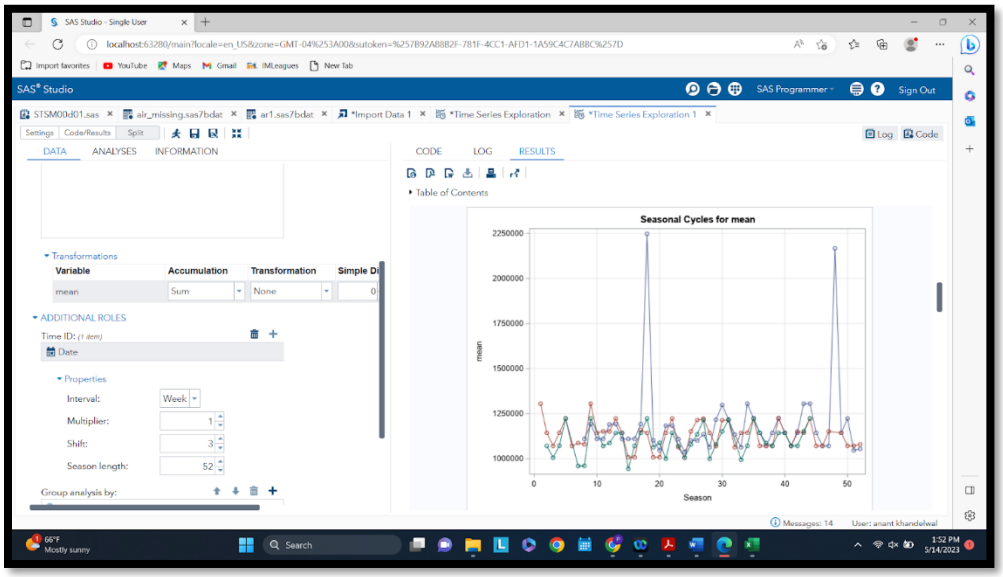


Fig. 4 - Seasonal Cycle Plot

To further solidify the fact that a seasonal component and pattern exists, the seasonal cycle plot is examined. As seen from this plot above, similar to the seasonal component of mean plot, a general pattern or seasonality is seen. Therefore, the time series does indeed have a seasonal component.

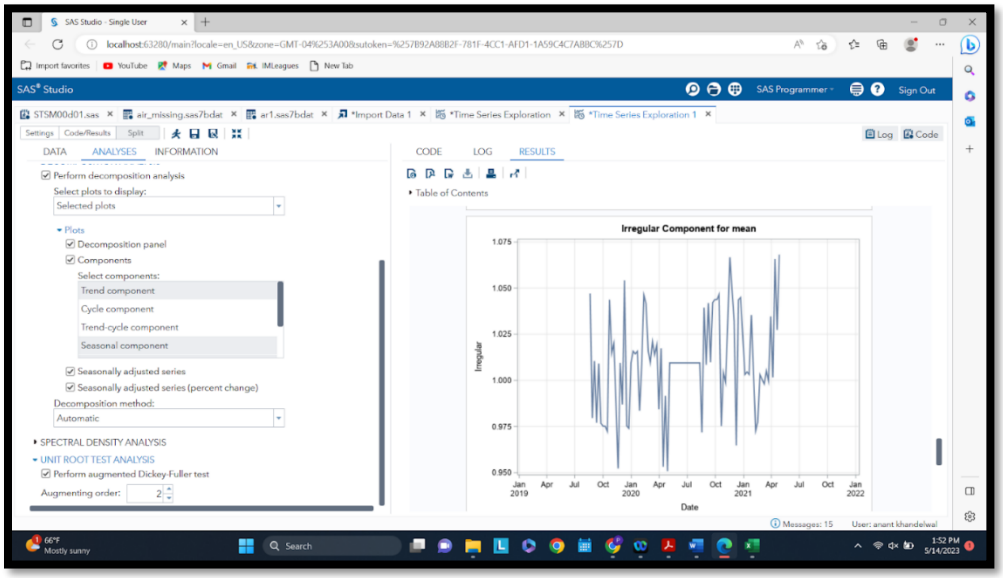


Fig. 5 - Irregular Component of Mean Sales

Finally, examining the irregular component after decomposing the dependent variable. No trend or seasonality can really be seen or extracted from this.

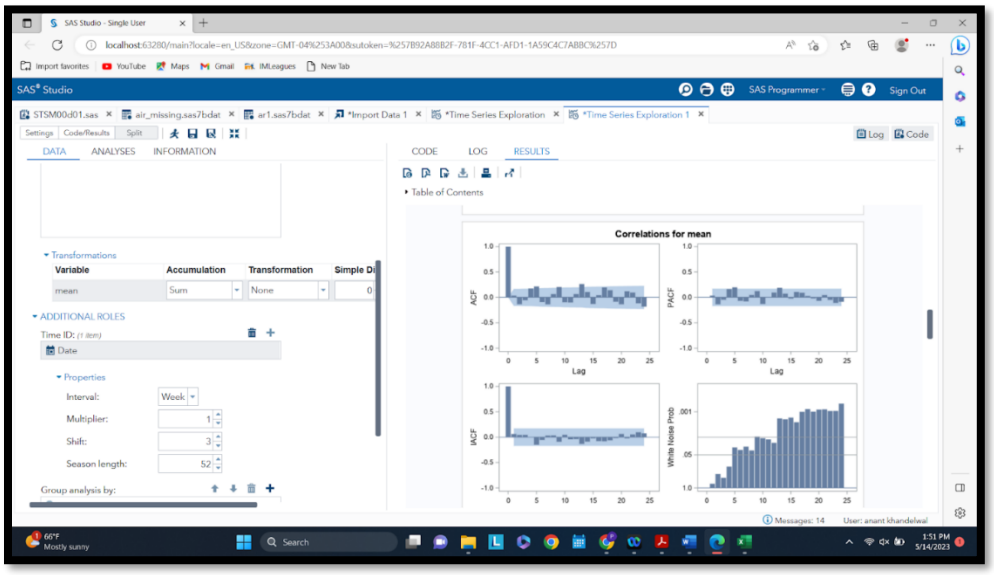


Fig. 6 - Correlations for Mean

Further examining the ACF, PACF, IACF, and the White Noise Test, a signal is present in the dependent variable of the dataset. Therefore, it is fair to assume that trend and seasonal components have to be extracted to help

describe the given time series. However, further modeling and analysis will be required to completely extract the signal, by looking at tests and metrics, such as the white noise test.

3.2 Exploring independent variables:

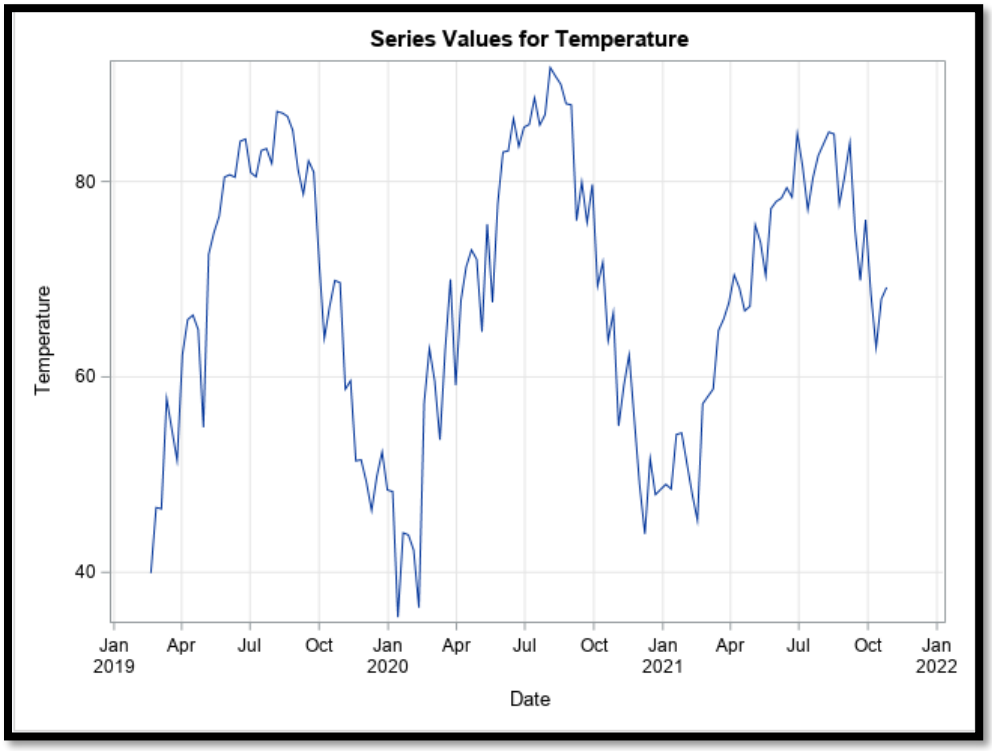


Fig. 7 - Temperature Time Series

Having looked at the dependent variable aggregated weekly mean sales of the Amazon store, the attention will now shift to exploring 5 of the given independent variables that may have an impact on the target variable. The independent variables will first be explored further to see if they are also time series, in which case pre-whitening will have to be done. Lastly, cross-correlations will have to be run between the dependent variable and the independent variables before moving on to modeling.\

3.2.1 Checking for time series

From the plot above, it can be seen that a general seasonal pattern is seen. But the trend seems quite flat. Hence, it can be concluded that temperature is a time series, as expected. As months change in a year, the temperature generally experiences a seasonal shift as well. Below are the decompositions of the time series for temperature to further examine the components.

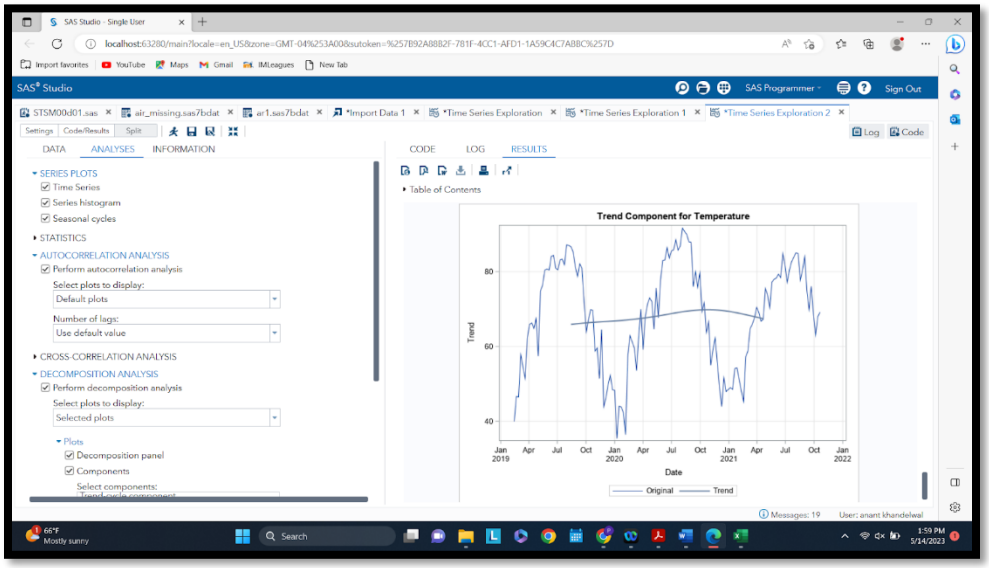


Fig. 8 - Trend Component of Temperature

As mentioned earlier, the trend seems quite flat, so the temperature time series does not consist of a trend component.

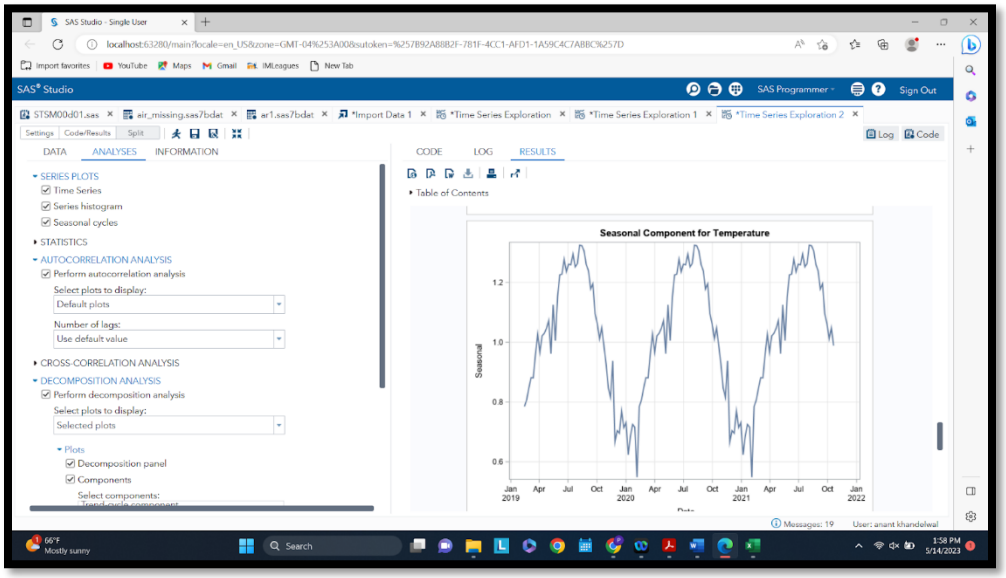


Fig. 9 - Seasonality Component of Temperature

A very clear and distinct pattern or seasonality is depicted in this component plot. Therefore, it can be concluded that the temperature plot consists of only a seasonal component.

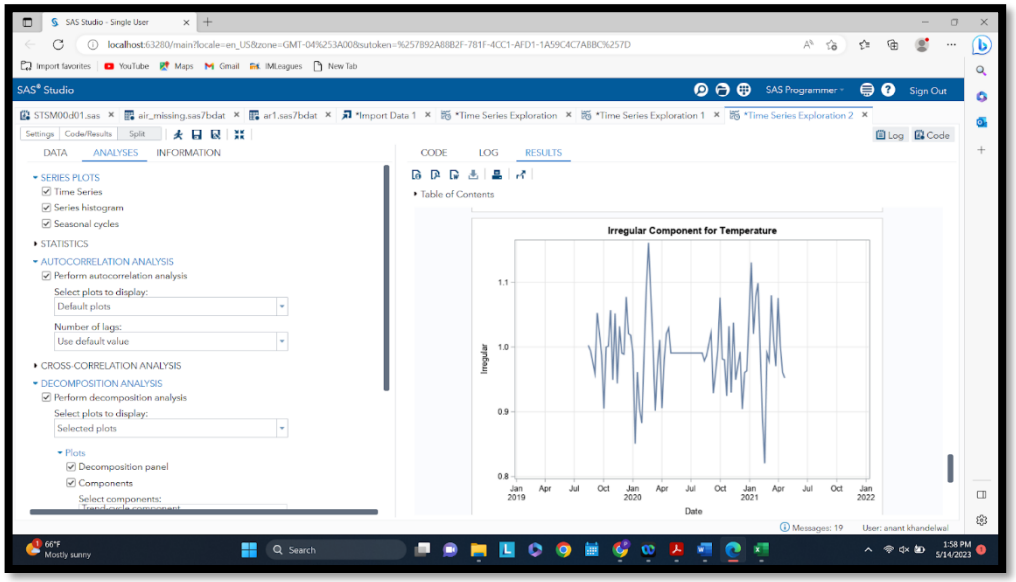


Fig. 10 - Irregular Component of Temperature

Finally, looking at the irregular component for temperature, it can be seen that it is indeed quite random and no patterns emerge.

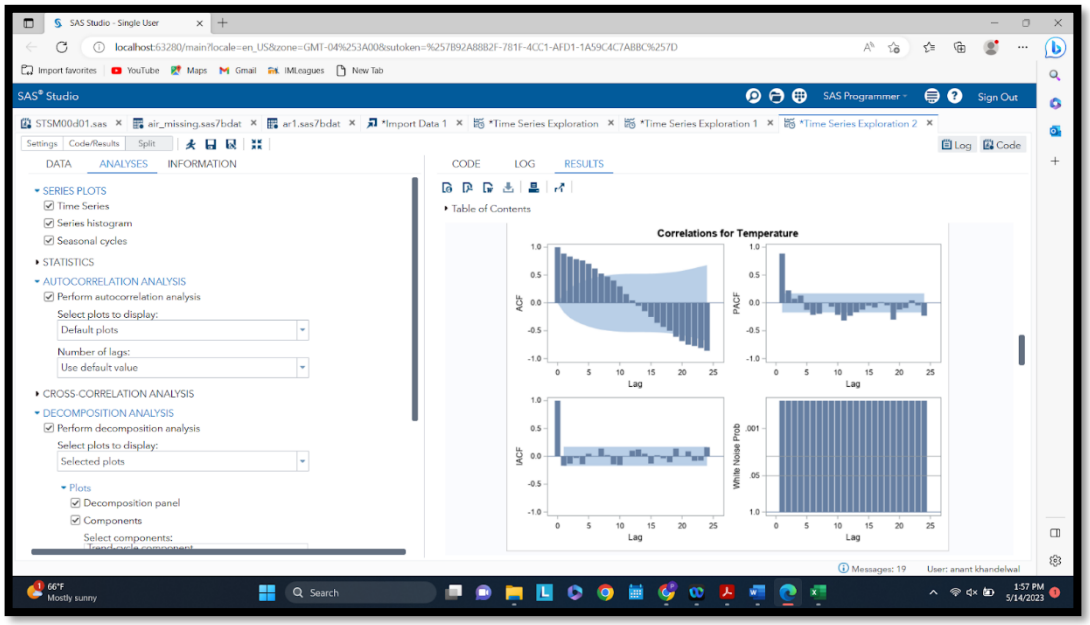


Fig. 11 - Correlations for Temperature

Lastly, examining the ACF, PACF, IACF, and the White Noise Test, a signal is present in the temperature variable of the data Therefore, the seasonal component does capture a good representation of the time series. Further, pre-whitening analysis will have to be conducted to see if temperature has ordinary or dynamic regressors to the dependent variable.

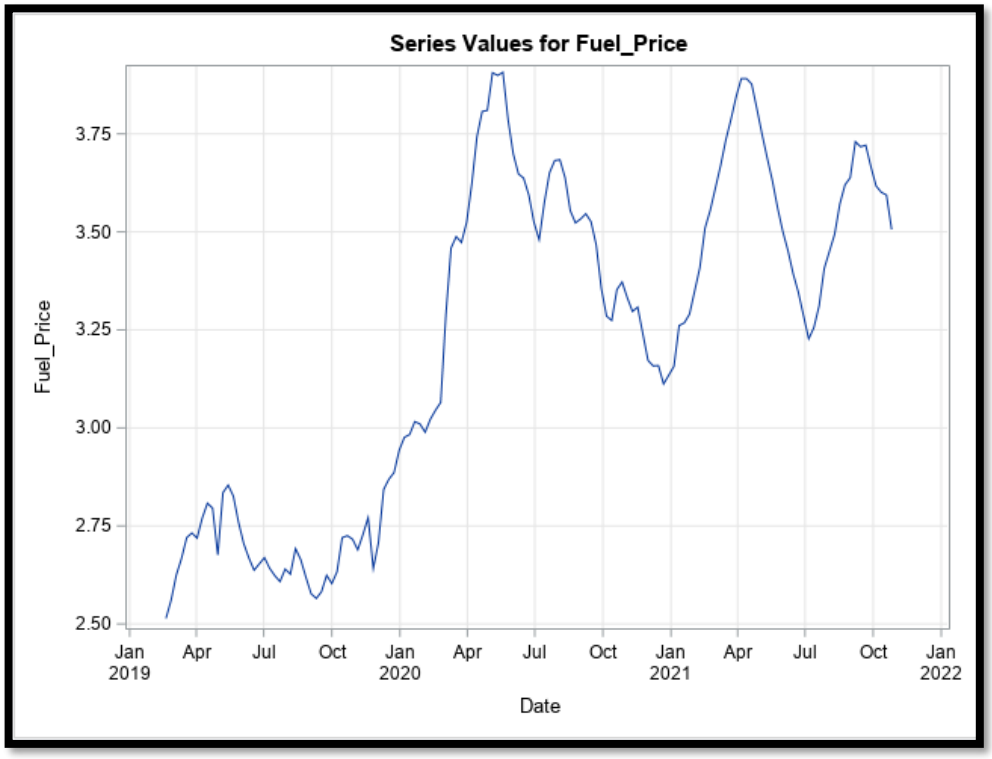


Fig. 12 - Fuel Price Time Series

From the plot above, it can be seen that this is a general trend. But no explicit seasonal pattern is seen. Hence, it can be concluded that fuel price is a time series for the given time period of three years from 2019 - 2022. Below are the decompositions of the time series for fuel prices to further examine the components.

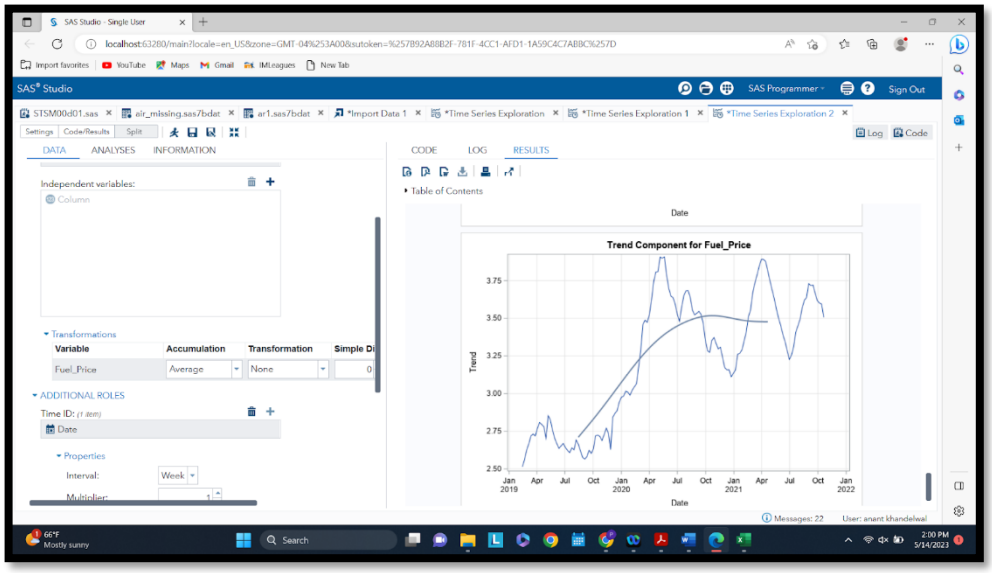


Fig. 13 - Trend Component of Fuel Price

A very clear trend component is seen for fuel price. It is an increasing trend from 2019 - 2023, with a slight drop at the end. Overall, the fuel prices are increasing for the given time period of this analysis.

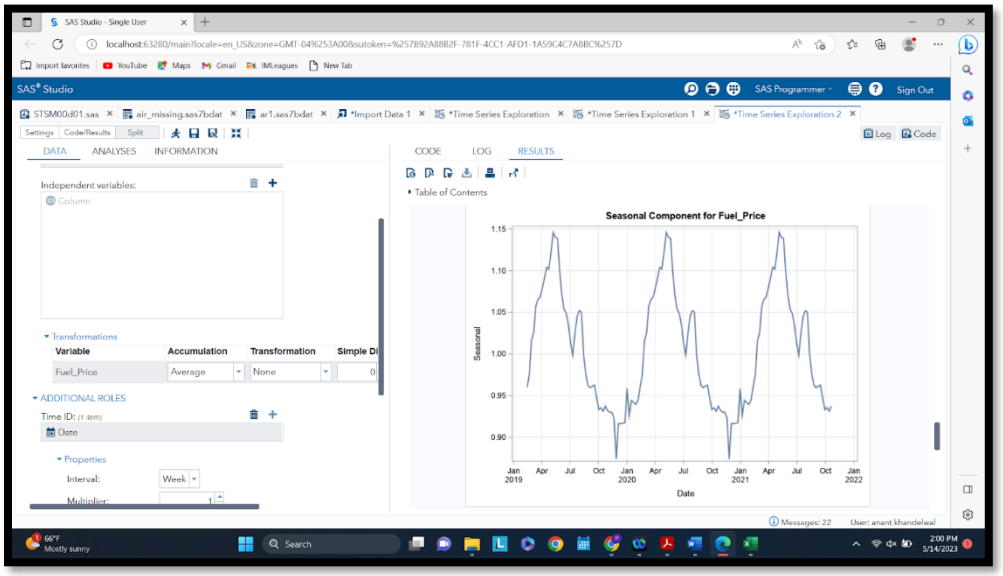


Fig. 14 - Seasonality Component of Fuel Price

A very clear and distinct pattern or seasonality is depicted in this component plot. Therefore, it can be concluded that the fuel prices plot consists of a seasonal component for the given time period as well.

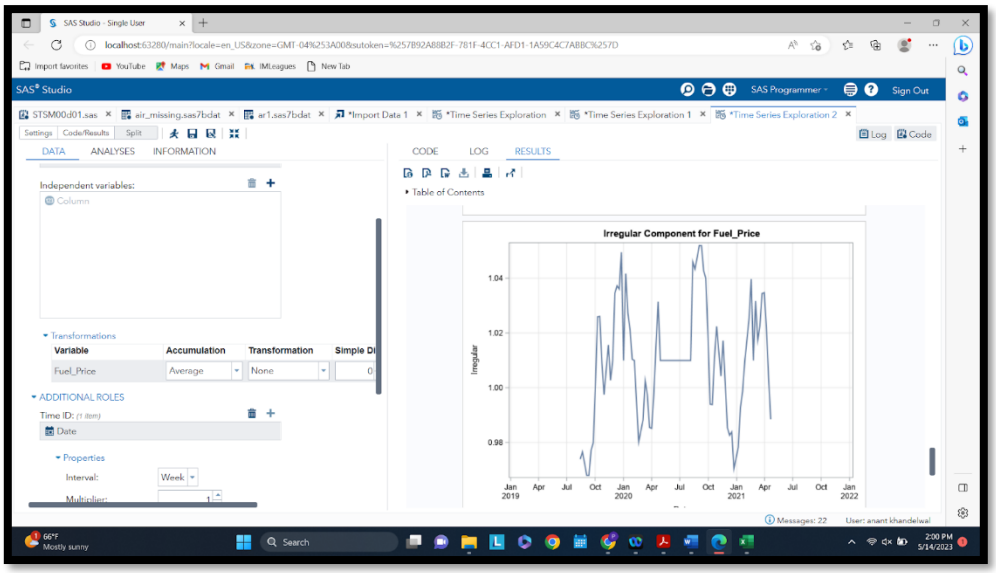


Fig. 15 - Irregular Component of Fuel Price

Finally, looking at the irregular component for fuel prices, it can be seen that it is indeed quite random and no patterns emerge.

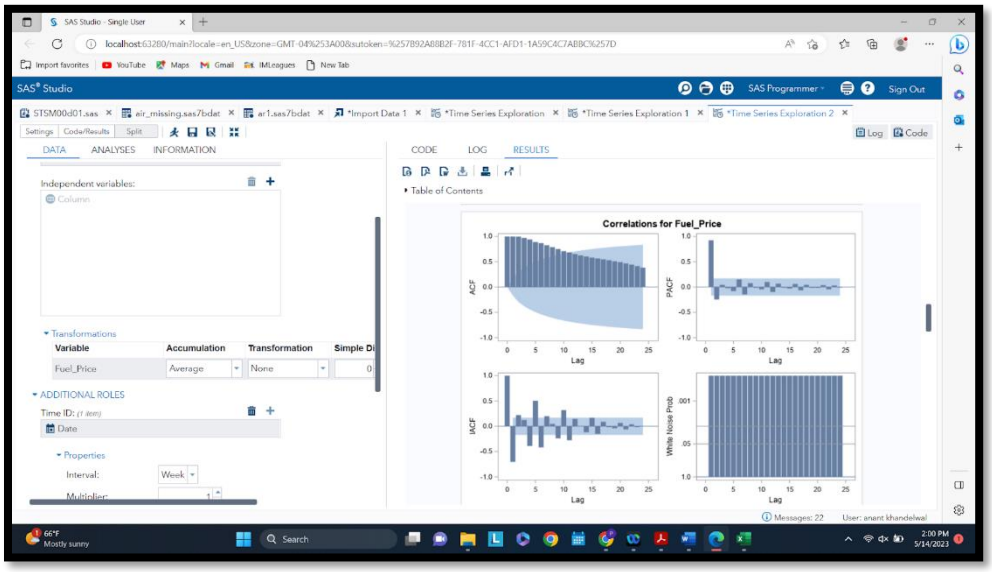


Fig 16. - Correlations for Fuel Price

Lastly, examining the ACF, PACF, IACF, and the White Noise Test, a signal is present in the fuel price variable of the data Therefore, the trend and seasonal components do capture a good representation of the time series. Further, pre-whitening analysis will have to be conducted to see if temperature has ordinary or dynamic regressors to the dependent variable.

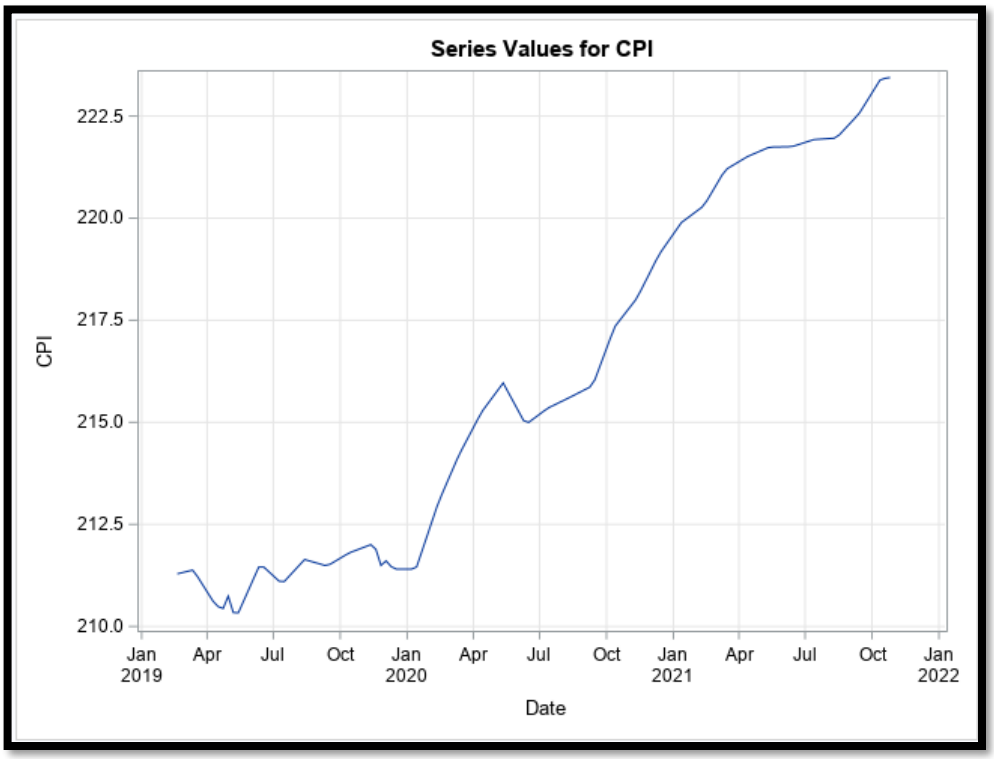


Fig. 17 - CPI Time Series

From the plot above, it can be seen that there is a general trend. But no explicit seasonal pattern is seen. Hence, it can be concluded that CPI is a time series for the given time period of three years from 2019 - 2022. CPI is the consumer price index, meaning that for the given time period, items are getting more expensive as compared to the previous year or baseline. Below are the decompositions of the time series for CPI to further examine the components.

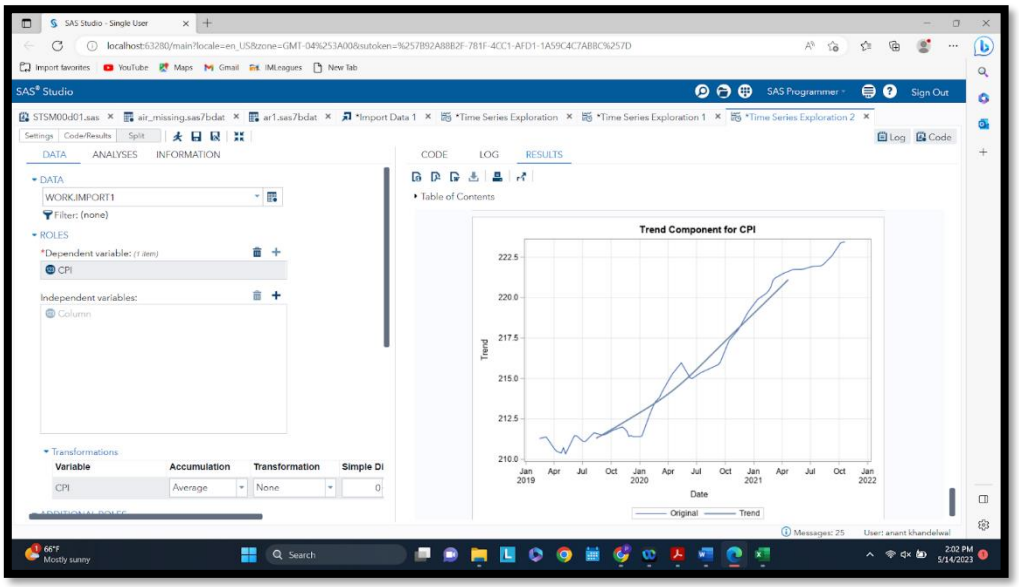


Fig. 18 - Trend Component of CPI

A very clear trend component is seen for CPI. It is an increasing trend from 2019 - 2023.

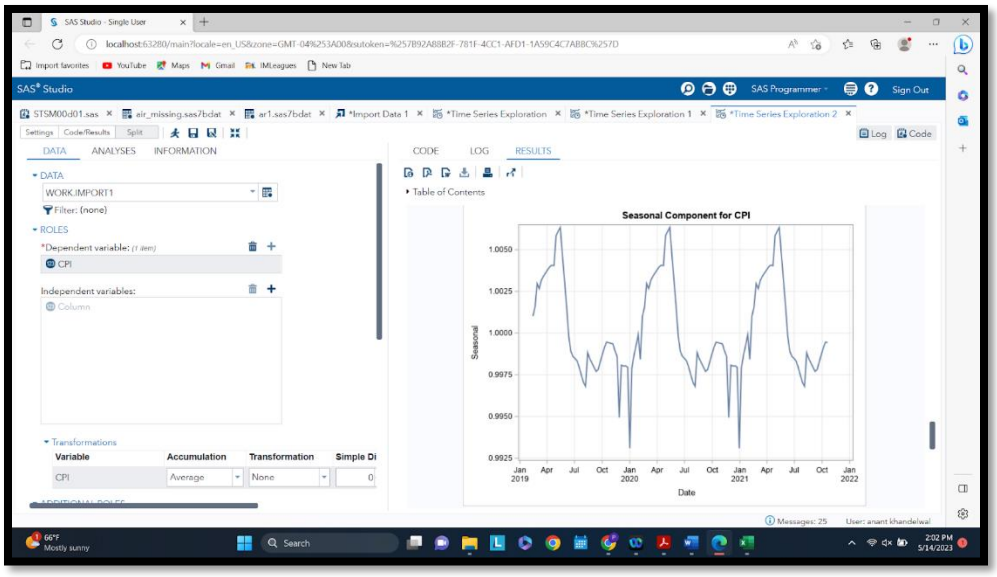


Fig. 19 - Seasonality Component of CPI

A very clear and distinct pattern or seasonality is depicted in this component plot. Therefore, it can be concluded that the CPI plot consists of a seasonal component for the given time period as well.

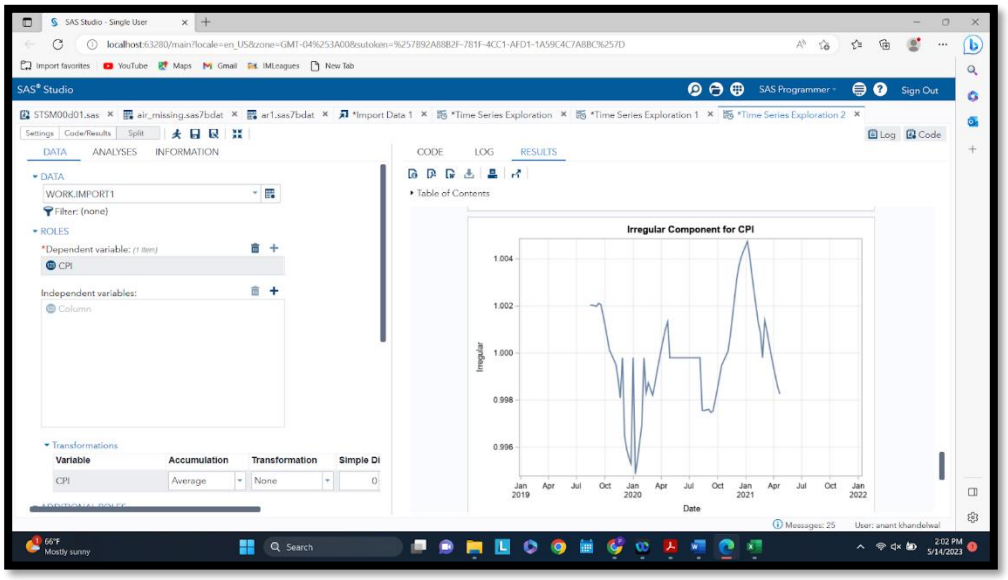


Fig. 20 - Irregular Component of CPI

Finally, looking at the irregular component for CPI, it can be seen that it is indeed quite random, and no patterns emerge.

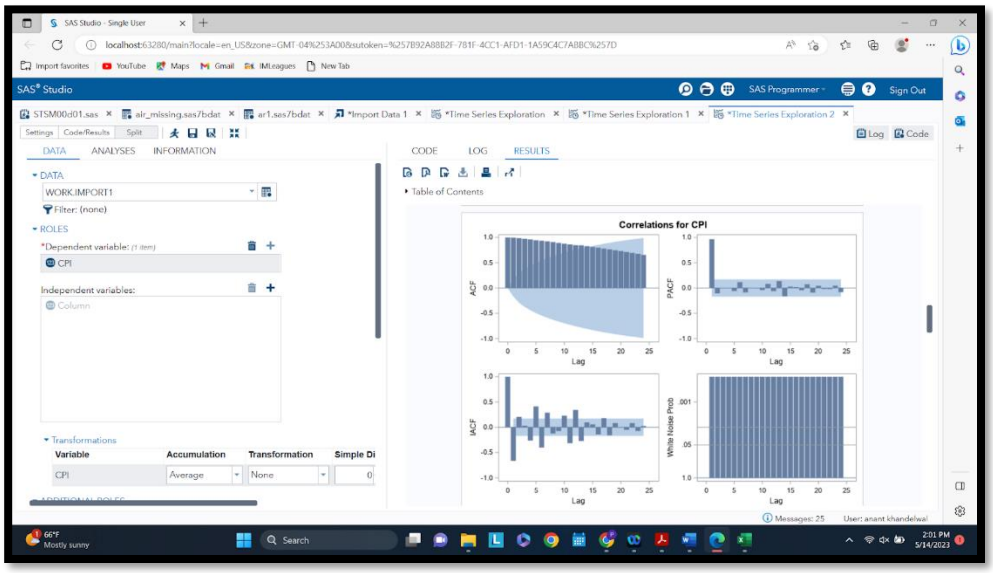


Fig. 21 - Correlations for CPI

Lastly, examining the ACF, PACF, IACF, and the White Noise Test, a signal is present in the CPI variable of the data Therefore, the trend and seasonal components do capture a good representation of the time series. Further, pre-whitening analysis will have to be conducted to see if temperature has ordinary or dynamic regressors to the dependent variable.

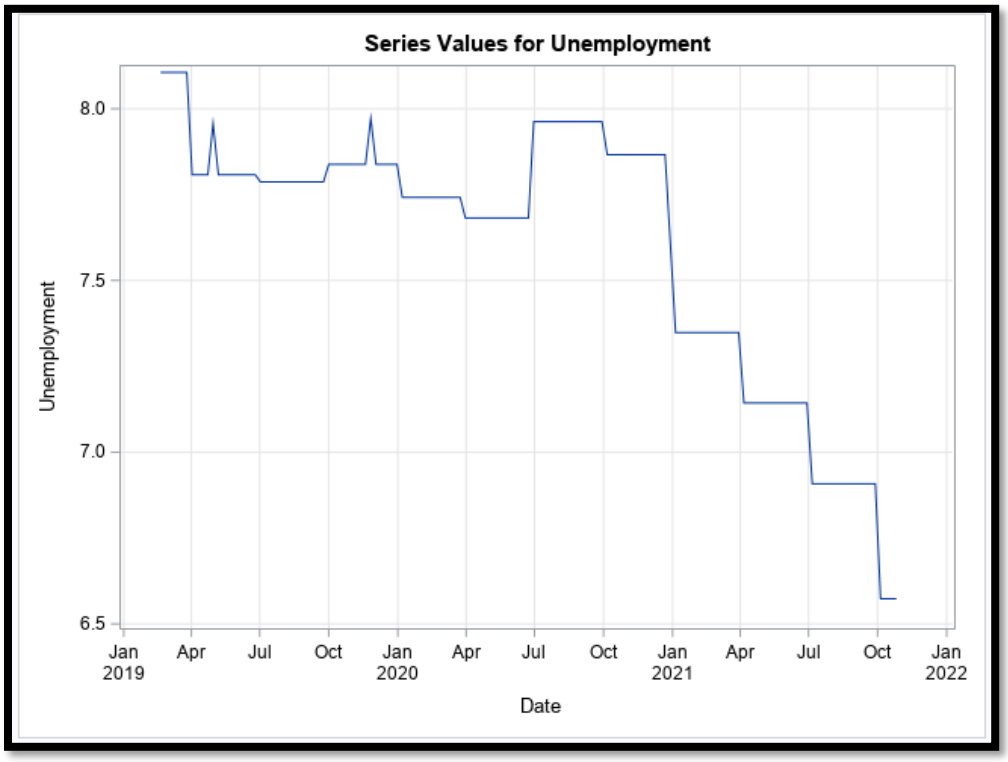


Fig. 22 - Unemployment Time Series

From the plot above, it can be seen that there is a general declining trend. But no explicit seasonal pattern is seen. Hence, it can be concluded that unemployment is a time series for the given time period of three years from 2019 - 2022. Unemployment is a macroeconomic measure of how an economy is doing, with a decreasing unemployment rate, the economy is generally doing well, meaning there is more disposable income and therefore more cash injections into the economy. This will be a very important independent variable for our analysis and below are the decompositions of the time series for unemployment to further examine the components.

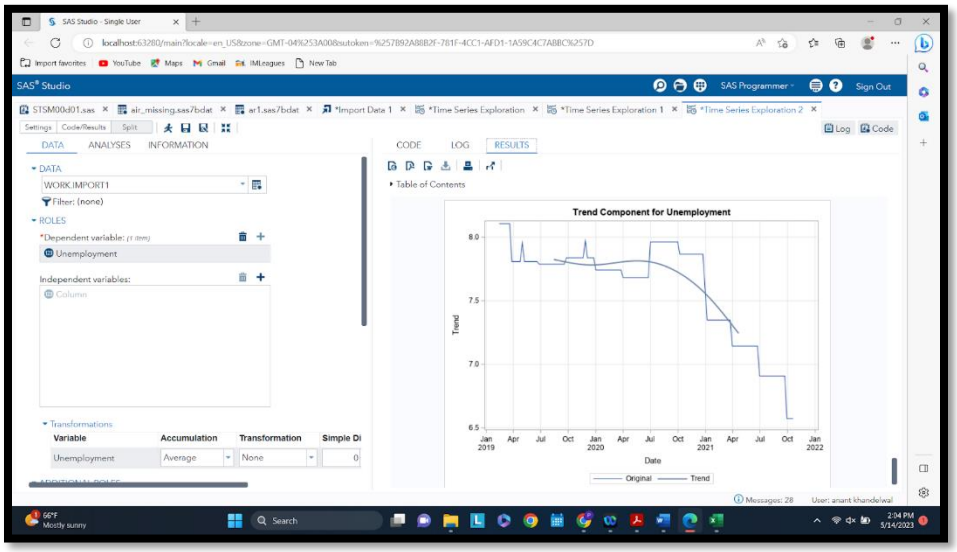


Fig. 23 - Trend Component of Unemployment

A very clear trend component is seen for unemployment. It is a decreasing trend from 2019 - 2023.

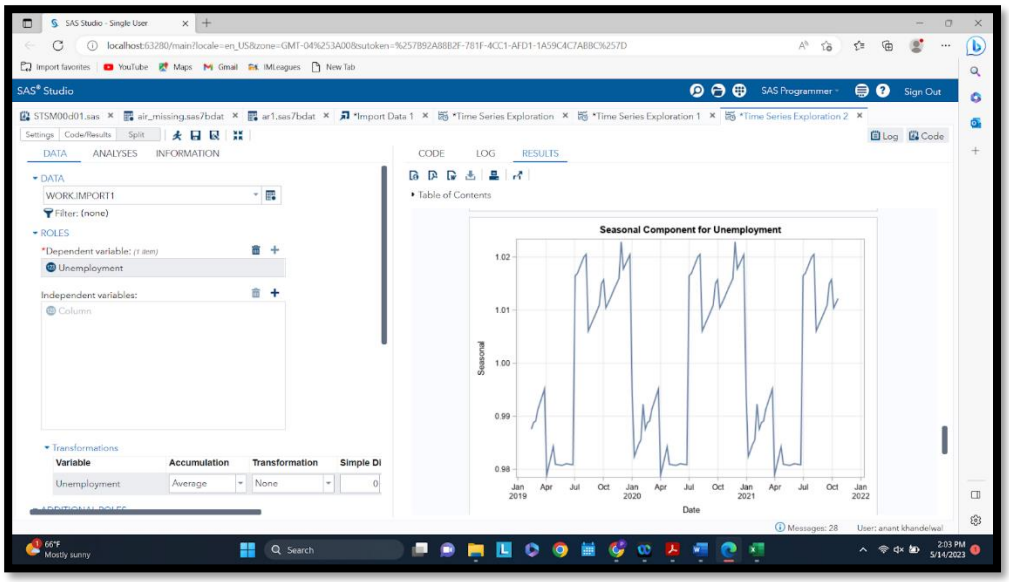
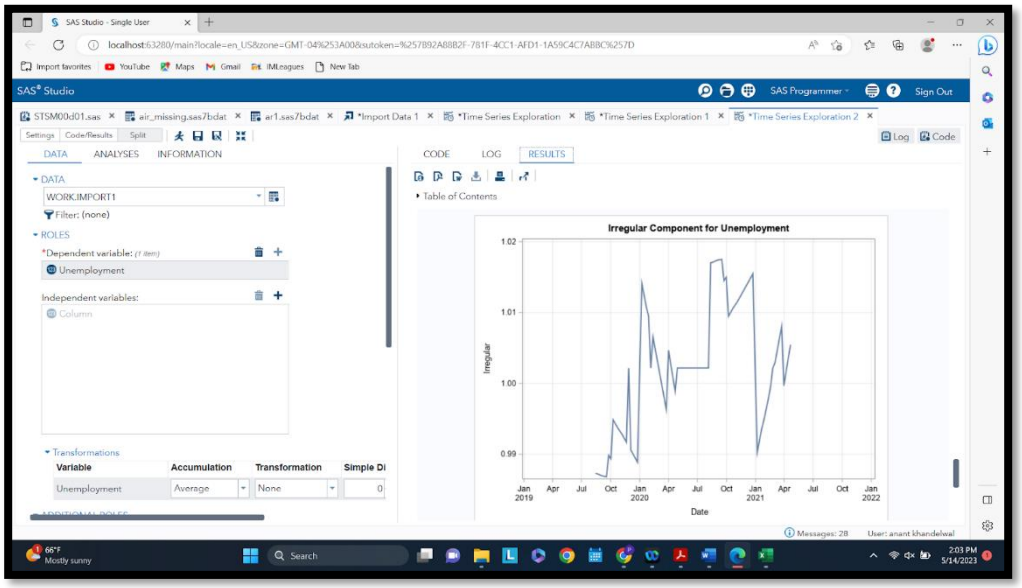


Fig. 24 - Seasonality Component of Unemployment

A distinct but not convincing pattern is seen for unemployment. While the pattern is repeating, it seems to be jumping from high to low to high values rather than having some smoothness to them. But a further look into this during modeling will help identify if a seasonal component exists in this time series.
Fig. 25 - Irregular Component of Unemployment



Finally, looking at the irregular component for unemployment, it can be seen that it is indeed quite random and no patterns emerge.

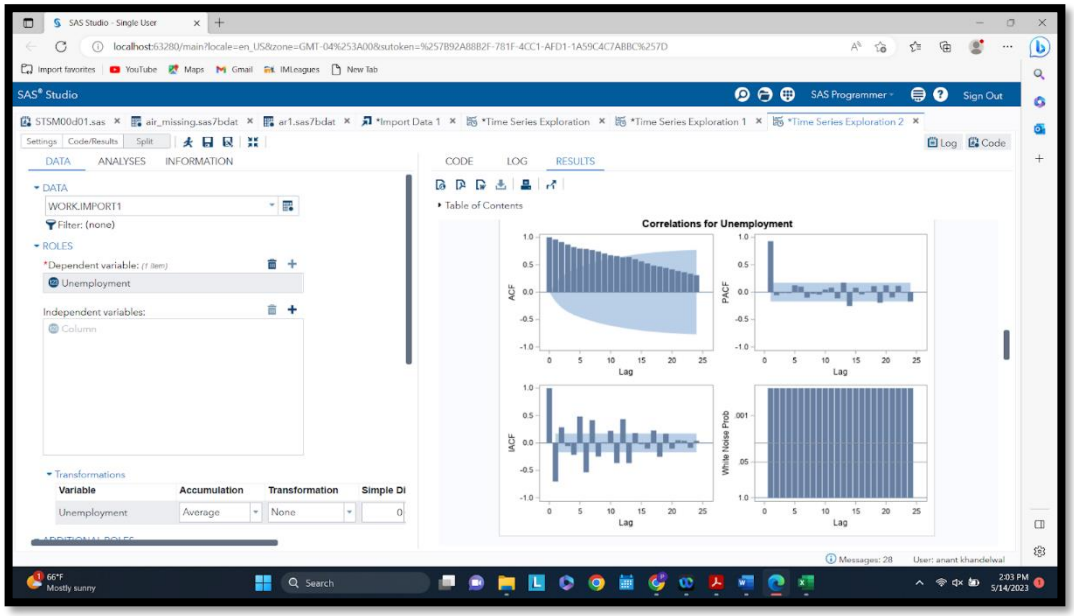


Fig 26. - Correlations for Unemployment

Lastly, examining the ACF, PACF, IACF, and the White Noise Test, a signal is present in the unemployment variable of the data. Therefore, the trend and seasonal components do capture a good representation of the time series. Further, pre-whitening analysis will have to be conducted to see if temperature has ordinary or dynamic regressors to the dependent variable.

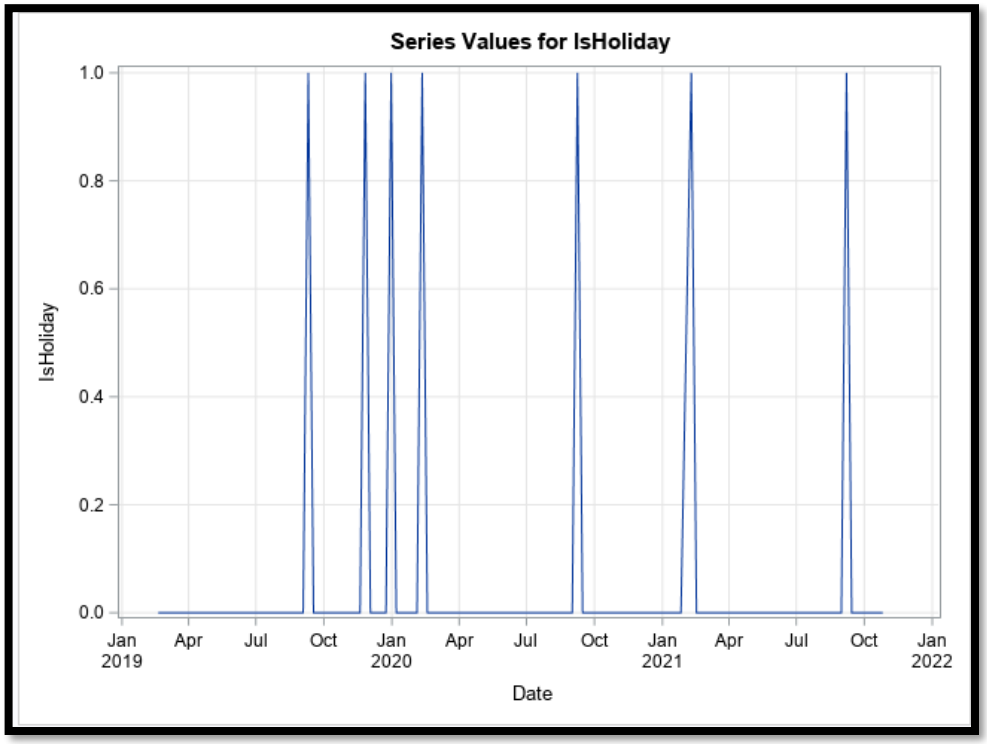


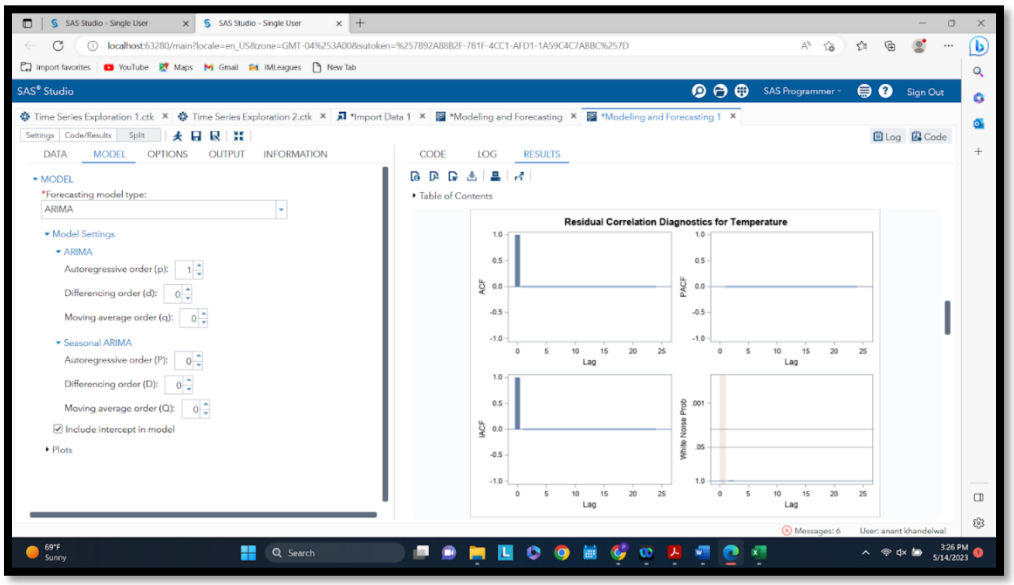
Fig. 27 - IsHoliday Time Series

IsHoliday is a categorical independent variable that gives insight into whether or not a day is a holiday or not in the given time period 2019 - 2023. By looking at the plot above, it can be seen quite clearly that the values only fluctuate between 0 and 1. Therefore, isHoliday is not a time series and will be treated simply as a categorical exogenous variable.

3.2.2 Pre-Whitening & Cross-Correlations

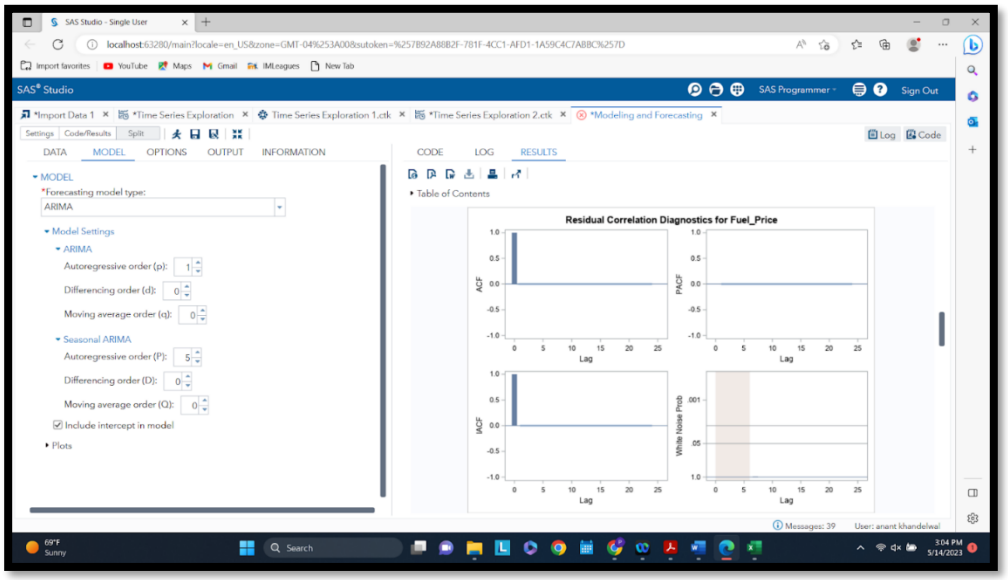
After exploring the independent variables, the conclusion was that four of them are indeed time series. Different models were run on each of four independent variables until only white noise is left in the residual correlation plots. This means that the signal is extracted for the independent variables. The plots are shown below and the corresponding model used.

Temperature - ARMA (1,0)



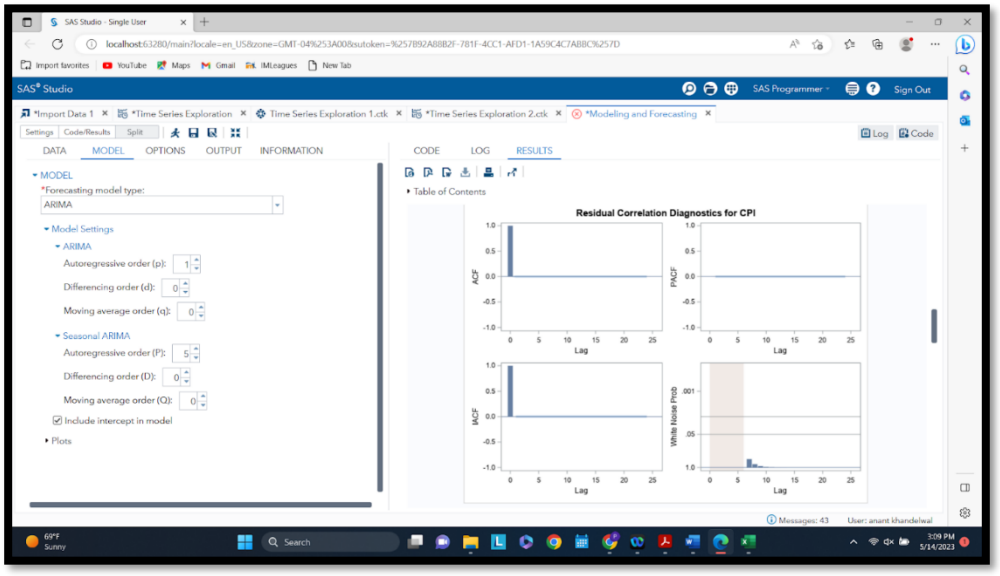
Residuals show that the model ARIMA (1,0) consists of only white noise and therefore the signal is extracted. This is now ready to be used with the dependent variable to check for ordinary and dynamic regressors.

Fuel Price - ARMA (1,0)



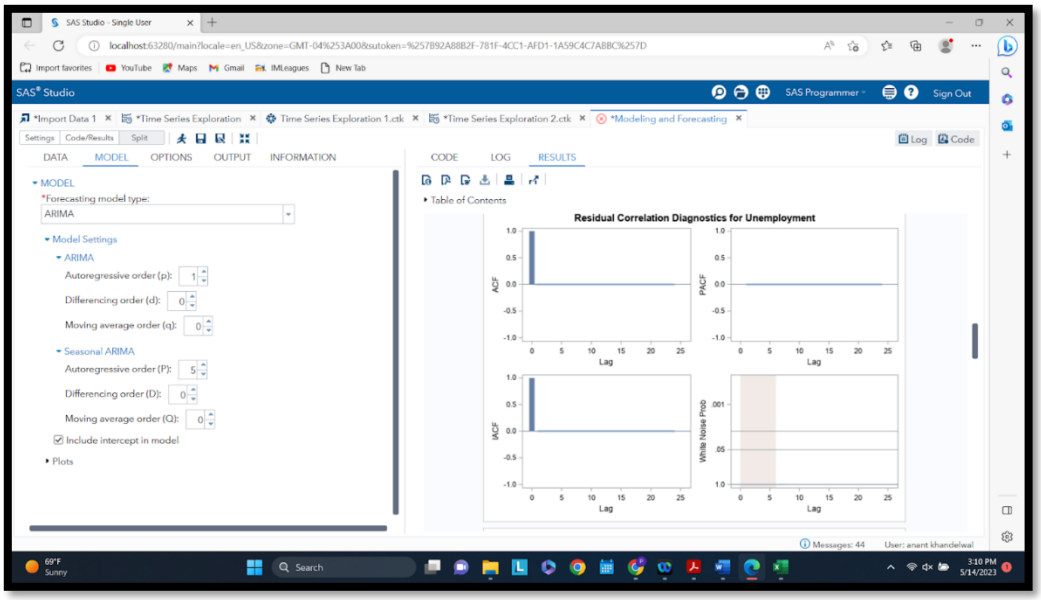
Residuals show that the model ARIMA (1,0) consists of only white noise and therefore the signal is extracted. This is now ready to be used with the dependent variable to check for ordinary and dynamic regressors.

CPI - ARMA (1,0):



Residuals show that the model ARIMA (1,0) consists of only white noise and therefore the signal is extracted. This is now ready to be used with the dependent variable to check for ordinary and dynamic regressors.

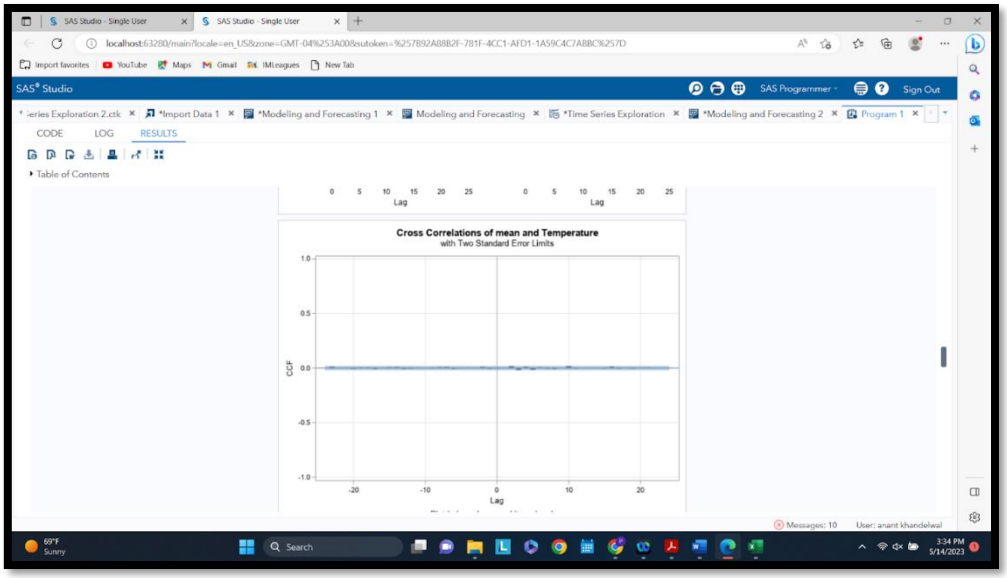
Unemployment - ARMA (1,0):



Residuals show that the model ARIMA (1,0) consists of only white noise and therefore the signal is extracted. This is now ready to be used with the dependent variable to check for ordinary and dynamic regressors.

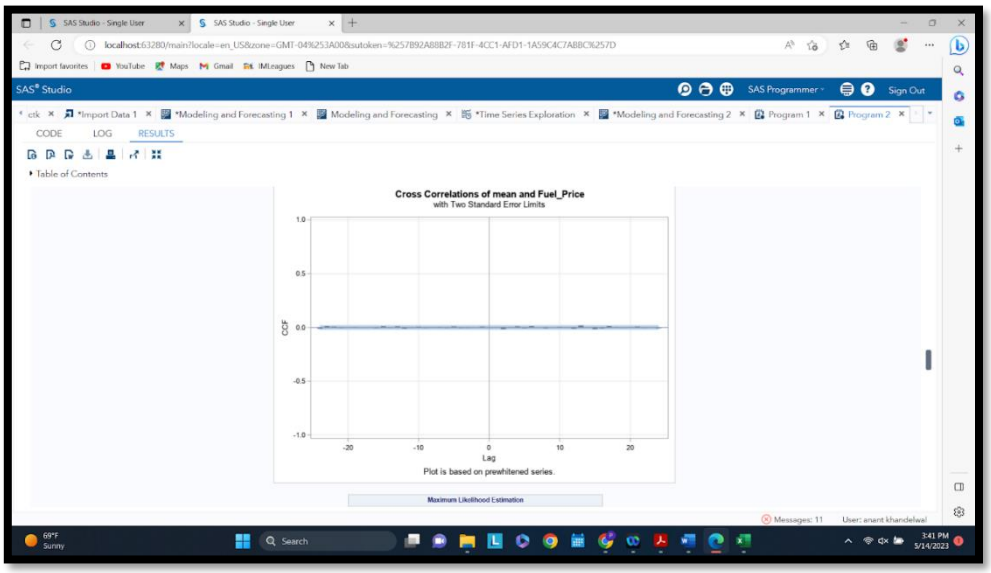
Now that the independent variables are modeled, it is time to run it on the dependent variable to check the cross-correlations between the two to find if there are any ordinary and/or dynamic regressors or simply lagged effects.

Plot 1 - Cross-Correlation between Temperature and Aggregated Mean Weekly Sales



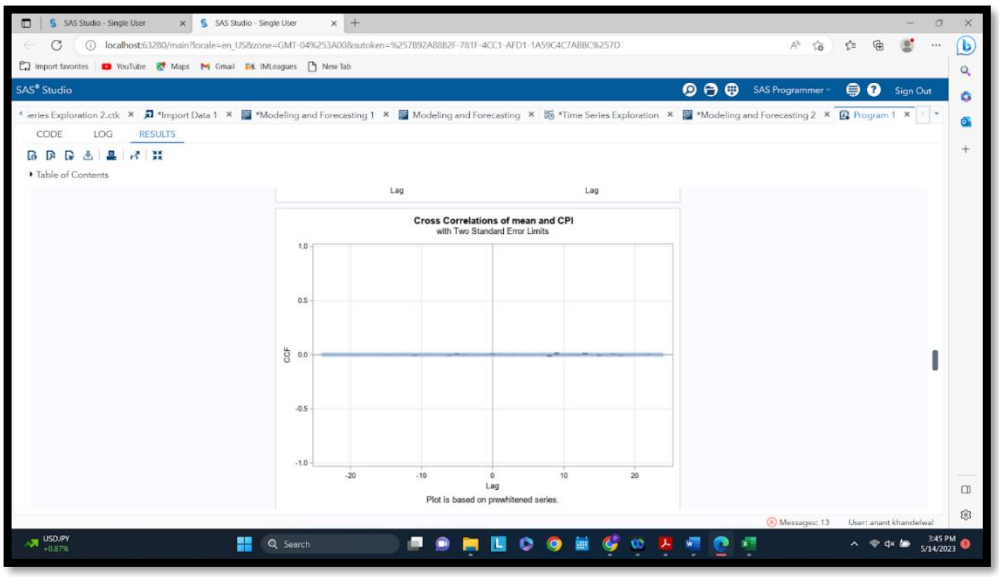
From the cross-correlation plot above, it can be seen that temperature does not have any ordinary or dynamic regressors to the dependent variable. So it can be omitted in our analysis.

Plot 2 - Cross-Correlation between Fuel Price and Aggregated Mean Weekly Sales



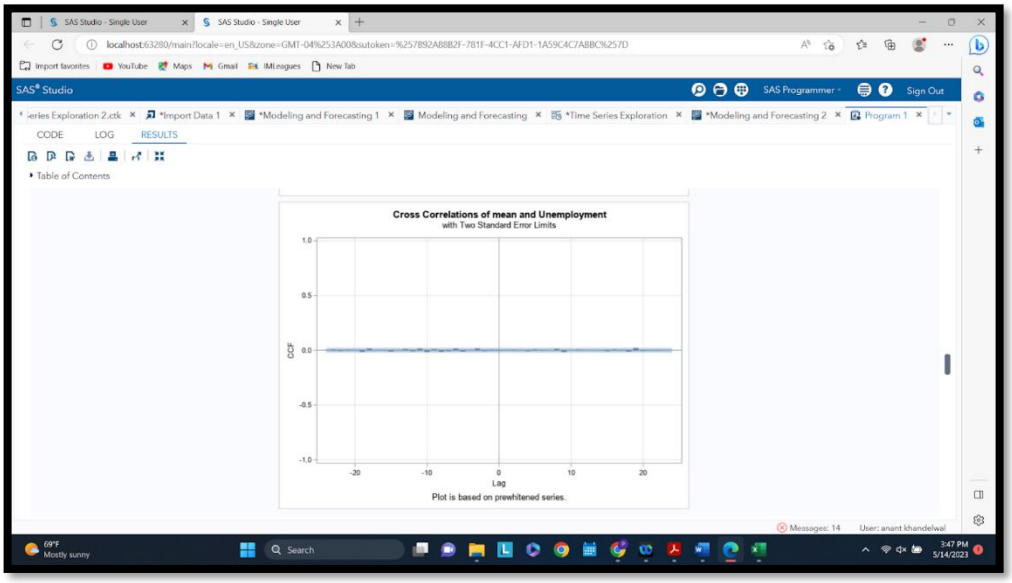
From the cross-correlation plot above, it can be seen that fuel price does not have any ordinary or dynamic regressors to the dependent variable. So it can be omitted in our analysis.

Plot 3 - Cross-Correlation between CPI and Aggregated Mean Weekly Sales



From the cross-correlation plot above, it can be seen that CPI does not have any ordinary or dynamic regressors to the dependent variable. So it can be omitted in our analysis.

Plot 4 - Cross-Correlation between Unemployment and Aggregated Mean Weekly Sales



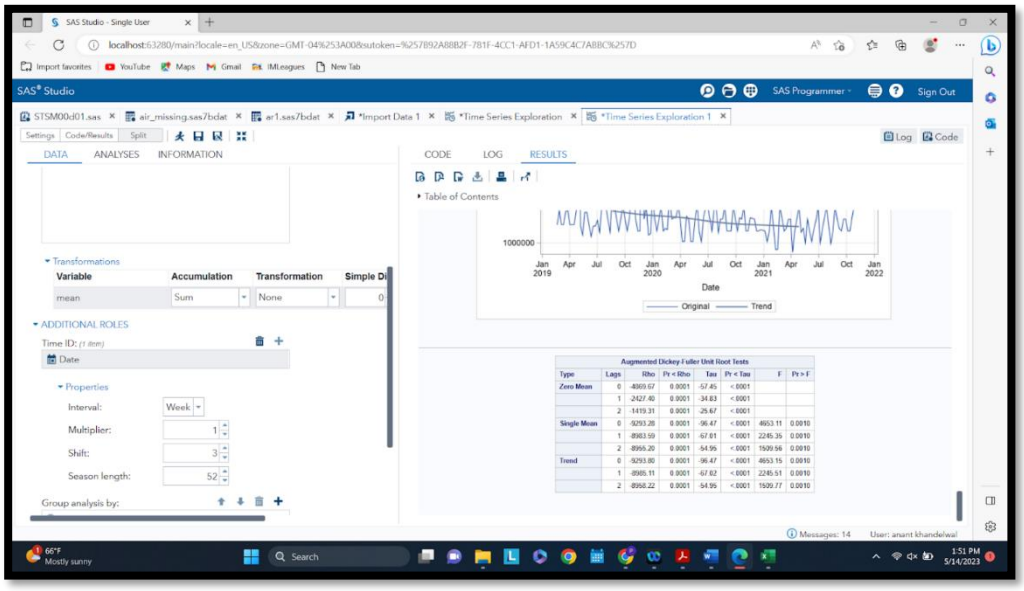
From the cross-correlation plot above, it can be seen that unemployment does not have any ordinary or dynamic regressors to the dependent variable. So it can be omitted in our analysis.

Therefore, it can be seen that the four independent variables: temperature, fuel price, CPI, and unemployment do not show any cross-correlation with the dependent variable and are statistically insignificant. All spikes are contained within the 95% confidence interval and therefore all four independent variables can be omitted from our analysis.

3.3 Checking for stationarity:

Lastly, prior to moving on to modeling and forecasting the aggregate weekly sales for the amazon store. First, the Augmented Dickey-Fuller (ADF) test must be used to check for stationarity of the dataset. This is a statistical test to verify that the data is stationary, which allows the use of ARIMA models that are very powerful and accurate. The ADF test is as follows: the null hypothesis is that the data is not stationary and the alternative hypothesis is that the data set is stationary. Therefore if the p-values for Pr< Rho and Pr > F are significant, ie. < 0.05, the data is stationary.

Stationary Test Table:



The dependent variable clearly shows a trend and hence the focus should be on rows corresponding to the 'Trend' type. Examining these rows, it is seen that the Pr< Rho and Pr > F are 0.0001 and 0.0010, respectively. This means that the test is significant and the data is stationary, therefore ARIMA models can be used for modeling and forecasting purposes.

4. Modeling and forecasting:

There were 2 important factors to be considered before kick starting modeling:

I) Variables or parameters to be considered for model generation.
II) Understanding trend and seasonality in data to ensure correct models are used and optimized.

4.1 Finalizing key independent variables:

To start creating models, it was critical to select the variables to be used to generate forecasts that would help predict sales trends based on key input parameters. Based on the variables explored in the previous section, we concluded on using the below set of variables:

1. Stochastic variables (*Numeric*):
    a. Temperature
    b. CPI
    c. Unemployment
    d. Fuel price

2. Deterministic variables (*Categorical*) :
    a. IsHoliday

4.2 Determining right set of models:

Amazon data set explored in the initial section shows there is definite seasonality in sales of store 1. Also, the trend graph shows there is a declining trend from 2019 to 2021 for the store considering sales for all the departments aggregated.

Also, the correlation graph shows that the data is not white noise and hence there is some systematic variation that can be captured using the models. Here, the PACF and IACP do who some spikes, thought not very significant. Hence, we tried to consider different ARIMA models to capture auto-regressive and moving average models.

4.3 Modeling:

There were 2 sets of models generated – ARIMA and ARIMAX using a **HOLDOUT** sample of **12 months** and a **forecast** of **24 months** as below:

4.3.1 ARIMA:

Model 1: ARIMA(0,2):

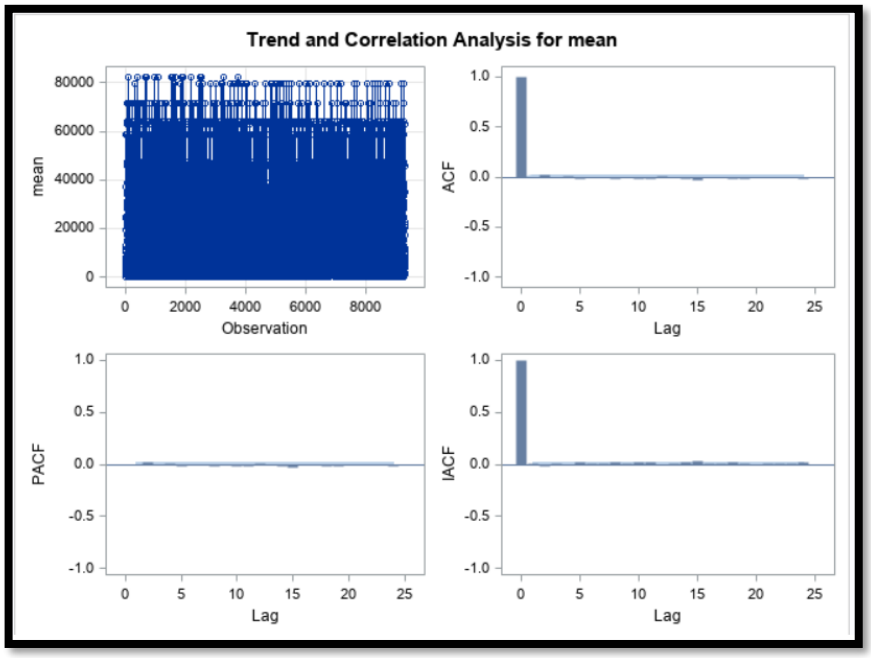
The first model created was an ARIMA model with parameters as below:

p = 0

q = 2

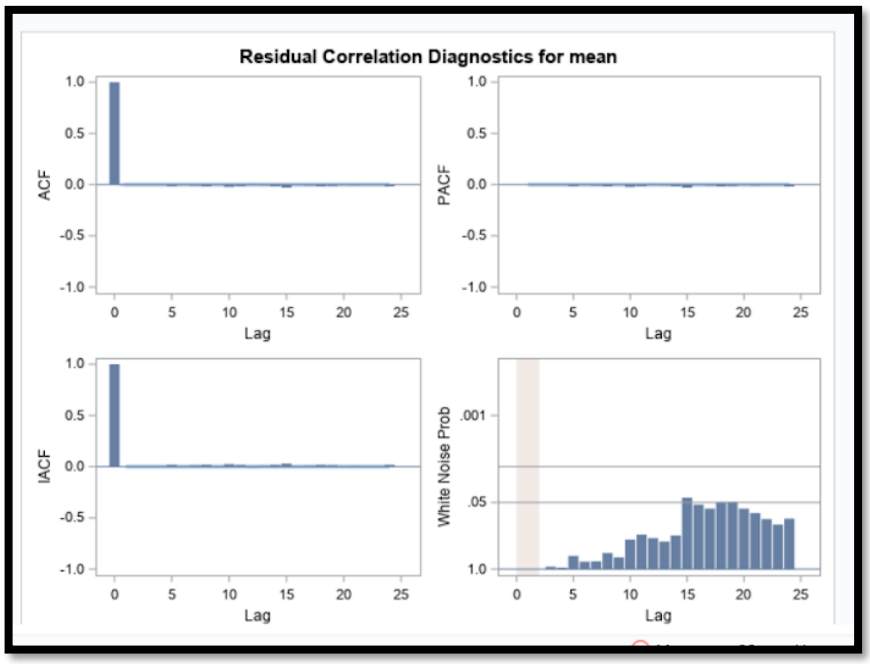
With the same, below were the outputs:

I) **Correlation for mean**:



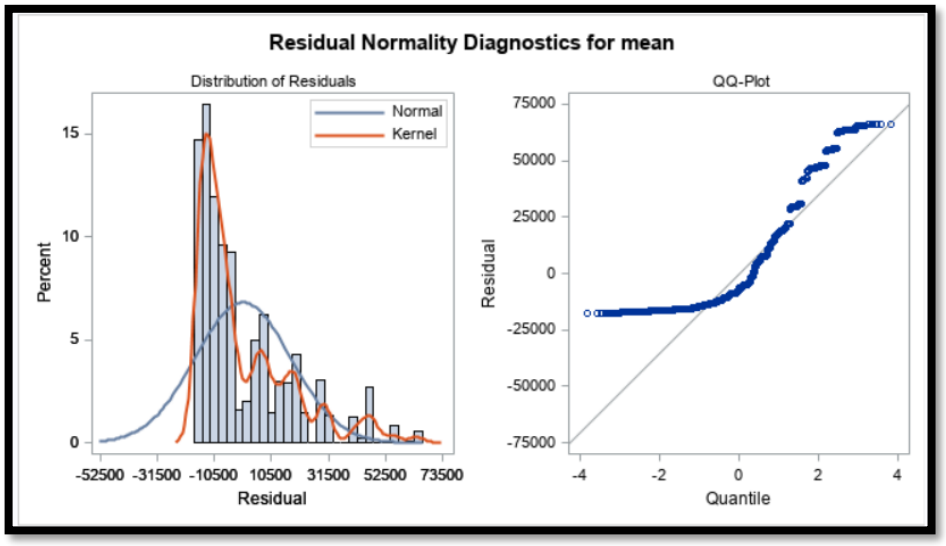
Through ACF graph, it is evident that there is no autocorrelation. Also, there are no spikes in the PACF and the IACF graphs.

II) Residual plots:



The results show that the white noise test fails for the residuals with some spikes above the threshold and hence, the model is not able to capture all the systematic variation existing in the dataset well. Also, there is no auto-correlation that exists in the residual plot and hence is great.

III) Residual normality check:



The residual graph indicates that the residuals are not normal but right skewed. Ideally, we would expect residuals to show normality and hence this model shows some skewness against that.

IV) Autocorrelation check for residuals:

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.10	4	0.7165	0.000	0.000	0.001	0.004	-0.015	-0.001
12	12.59	10	0.2474	-0.010	-0.016	-0.004	-0.022	-0.015	0.008
18	26.38	16	0.0489	-0.007	-0.016	-0.029	-0.004	-0.008	-0.016
24	30.65	22	0.1034	-0.012	0.001	-0.007	-0.001	-0.001	-0.016
30	41.90	28	0.0443	-0.031	-0.001	-0.006	-0.001	-0.015	-0.000
36	59.82	34	0.0040	-0.028	-0.010	-0.007	0.010	-0.023	-0.020
42	65.54	40	0.0066	-0.006	0.007	-0.016	-0.010	0.013	0.004
48	69.27	46	0.0149	0.005	0.013	-0.012	-0.002	-0.003	0.007

As indicated in the graph, the table above also indicates that there is no auto-correlation for the residuals and all values are insignificant.

V) Parameter estimates:

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	16737.0	185.08752	90.43	<.0001	0
MA1,1	0.0015602	0.01038	0.15	0.8805	1
MA1,2	-0.01682	0.01038	-1.62	0.1053	2

Constant Estimate	16736.99
Variance Estimate	3.0843E8
Std Error Estimate	17562.15
AIC	207734.7
SBC	207756.1
Number of Residuals	9280

Correlations of Parameter Estimates			
Parameter	MU	MA1,1	MA1,2
MU	1.000	0.000	0.000
MA1,1	0.000	1.000	-0.003
MA1,2	0.000	-0.003	1.000

The parameter estimates for MA1 at lag 1 and MA1 at lag 2 are large as they are above the threshold of 0.001 and hence clearly indicate that the model is not able to predict the sales trends well.

VI) Model and forecast outputs:

Model for variable mean	
Estimated Mean	16736.99

Moving Average Factors	
Factor 1:	1 - 0.00156 B**(1) + 0.01682 B**(2)

Note: Further warnings will not be printed.

Forecasts for variable mean						
Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
9269	16493.2205	17562.146	-17927.9530	50914.3940	9196.5579	-7296.6626
9270	16587.5271	17562.167	-17833.6883	51008.7424	6187.3897	-10400.1373
9271	16736.9922	17564.650	-17689.0895	51163.0739	1374.6497	-15362.3426
9272	16736.9922	17564.650	-17689.0895	51163.0739	21623.1847	4886.1925
9273	16736.9922	17564.650	-17689.0895	51163.0739	7808.4506	-8928.5416
9274	16736.9922	17564.650	-17689.0895	51163.0739	20727.8609	3990.8687
9275	16736.9922	17564.650	-17689.0895	51163.0739	11652.8274	-5084.1648
9276	16736.9922	17564.650	-17689.0895	51163.0739	63180.5682	46443.5760
9277	16736.9922	17564.650	-17689.0895	51163.0739	17139.3146	402.3224
9278	16736.9922	17564.650	-17689.0895	51163.0739	4091.5715	-12645.4207
9279	16736.9922	17564.650	-17689.0895	51163.0739	8052.0762	-8684.9161
9280	16736.9922	17564.650	-17689.0895	51163.0739	7436.0173	-9300.9749
9281	16736.9922	17564.650	-17689.0895	51163.0739	.	.
9282	16736.9922	17564.650	-17689.0895	51163.0739		

The forecast is shown for 12 months as that is the hold out period and residual is calculated only for that duration.

4.3.2 ARIMAX:

Model 2: ARIMAX(0,0)

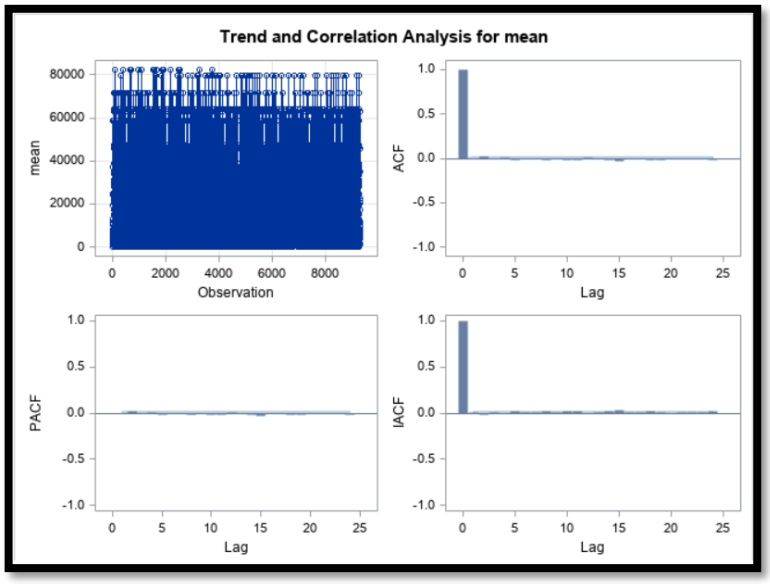
The second model was created using ARIMAX and had below parameters:

p = 0

q = 0

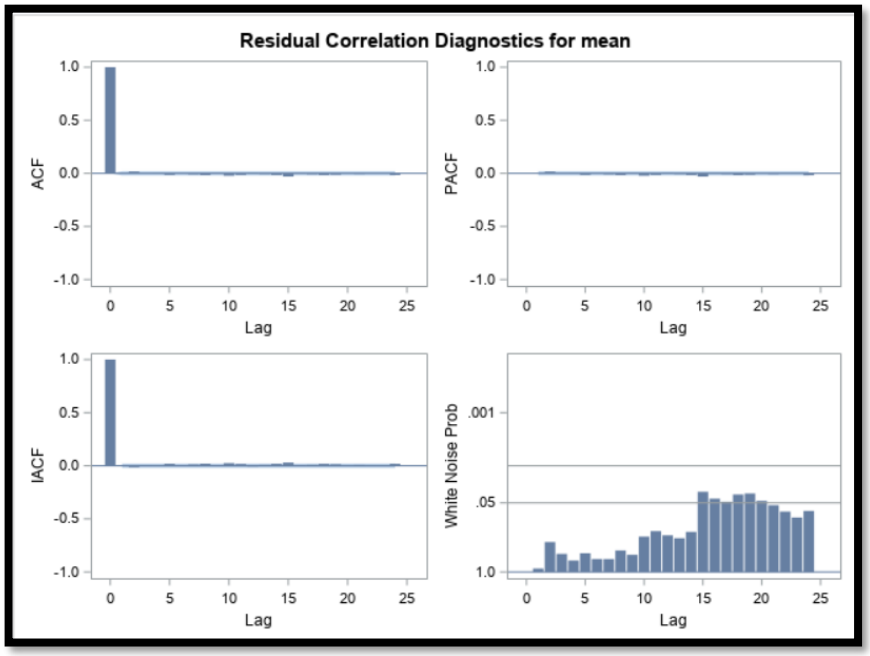
With the same, below were the outputs:

I) **Correlation for mean**:



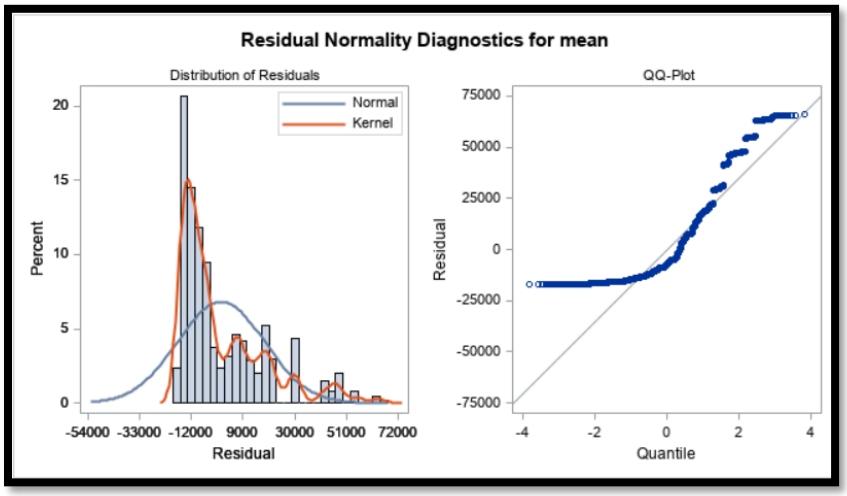
Through ACF graph, it is evident that there is no autocorrelation. Also, there are extremely small spikes in the PACF and the IACF graphs.

II) Residual plots:



The results show that the white noise test fails for the residuals with some spikes above the threshold and hence, the model is not able to capture all the systematic variation existing in the dataset well. Also, there is no auto-correlation that exists in the residual plot and hence is great.

III) Residual normality check:



The above plot indicates that the residual graph again is not normal and shows right skewness and kurtosis.

IV) Autocorrelation check for residuals:

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.79	6	0.5716	-0.002	0.017	0.001	0.003	-0.015	-0.001
12	15.74	12	0.2035	-0.010	-0.017	-0.004	-0.022	-0.015	0.007
18	30.29	18	0.0347	-0.008	-0.016	-0.029	-0.005	-0.009	-0.016
24	34.83	24	0.0709	-0.013	0.001	-0.007	-0.001	-0.002	-0.017
30	46.67	30	0.0268	-0.031	-0.002	-0.007	-0.001	-0.016	-0.000
36	65.07	36	0.0021	-0.029	-0.010	-0.008	0.009	-0.023	-0.020
42	70.71	42	0.0037	-0.007	0.006	-0.016	-0.010	0.013	0.004
48	74.42	48	0.0086	0.005	0.013	-0.012	-0.002	-0.004	0.007

The above table indicates that there is barely any auto – correlation existing amongst the residuals and hence is great.

V) Parameter estimates:

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	24775.7	26823.7	0.92	0.3557	0	mean	0
NUM1	14.77995	13.64044	1.08	0.2786	0	Temperature	0
NUM2	-59.84902	704.35320	-0.08	0.9323	0	Fuel_Price	0
NUM3	-41.05827	103.88360	-0.40	0.6927	0	CPI	0
NUM4	3.47377	890.60781	0.00	0.9969	0	Unemployment	0
NUM5	-417.73928	791.74373	-0.53	0.5978	0	IsHoliday	0

Constant Estimate	24775.67
Variance Estimate	3.0853E8
Std Error Estimate	17565.11
AIC	207740.8
SBC	207783.6
Number of Residuals	9280

Correlations of Parameter Estimates

Variable / Parameter	mean MU	Temperature NUM1	Fuel_Price NUM2	CPI NUM3	Unemployment NUM4	IsHoliday NUM5
mean MU	1.000	-0.211	0.595	-0.980	-0.883	-0.033
Temperature NUM1	-0.211	1.000	-0.243	0.181	0.177	0.147
Fuel_Price NUM2	0.595	-0.243	1.000	-0.703	-0.328	0.016
CPI NUM3	-0.980	0.181	-0.703	1.000	0.780	0.023
Unemployment NUM4	-0.883	0.177	-0.328	0.780	1.000	0.021
IsHoliday NUM5	-0.033	0.147	0.016	0.023	0.021	1.000

The parameter estimates indicate high values for the independent variables and hence indicate that the model needs to be fine-tuned further to generate better sales forecasts.

VI) Model and forecast outputs:

Model for variable mean	
Estimated Intercept	24775.67

Input Number 1	
Input Variable	Temperature
Overall Regression Factor	14.77995

Input Number 2	
Input Variable	Fuel_Price
Overall Regression Factor	-59.849

Input Number 3	
Input Variable	CPI
Overall Regression Factor	-41.0583

Input Number 4	
Input Variable	Unemployment
Overall Regression Factor	3.473774

Input Number 5	
Input Variable	IsHoliday
Overall Regression Factor	-417.739

The value for option LEAD= has been reduced to 12.

Forecasts for variable mean

Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
9269	16436.6230	17565.109	-17990.3584	50863.6045	9196.5579	-7240.0651
9270	16436.6230	17565.109	-17990.3584	50863.6045	6187.3897	-10249.2333
9271	16436.6230	17565.109	-17990.3584	50863.6045	1374.6497	-15061.9734
9272	16436.6230	17565.109	-17990.3584	50863.6045	21623.1847	5186.5617
9273	16436.6230	17565.109	-17990.3584	50863.6045	7808.4506	-8628.1725
9274	16436.6230	17565.109	-17990.3584	50863.6045	20727.8609	4291.2379
9275	16436.6230	17565.109	-17990.3584	50863.6045	11652.8274	-4783.7956
9276	16436.6230	17565.109	-17990.3584	50863.6045	63180.5682	46743.9451
9277	16436.6230	17565.109	-17990.3584	50863.6045	17139.3146	702.6916
9278	16436.6230	17565.109	-17990.3584	50863.6045	4091.5715	-12345.0515
9279	16436.6230	17565.109	-17990.3584	50863.6045	8052.0762	-8384.5469
9280	16436.6230	17565.109	-17990.3584	50863.6045	7436.0173	-9000.6057

The forecast is shown for 12 months as that is the hold out period and residual is calculated only for that duration.

Model 3: ARIMAX(1,0)

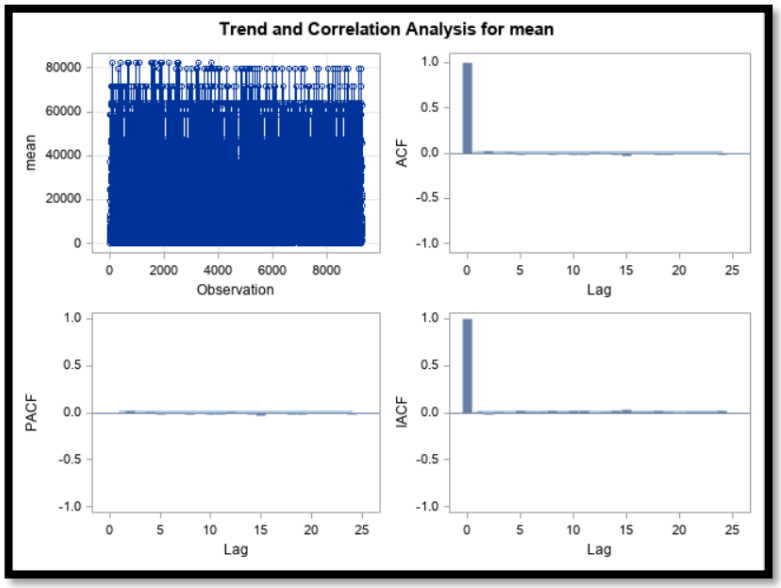
The third model was created using ARIMAX and had below parameters:

p = 1

q = 0

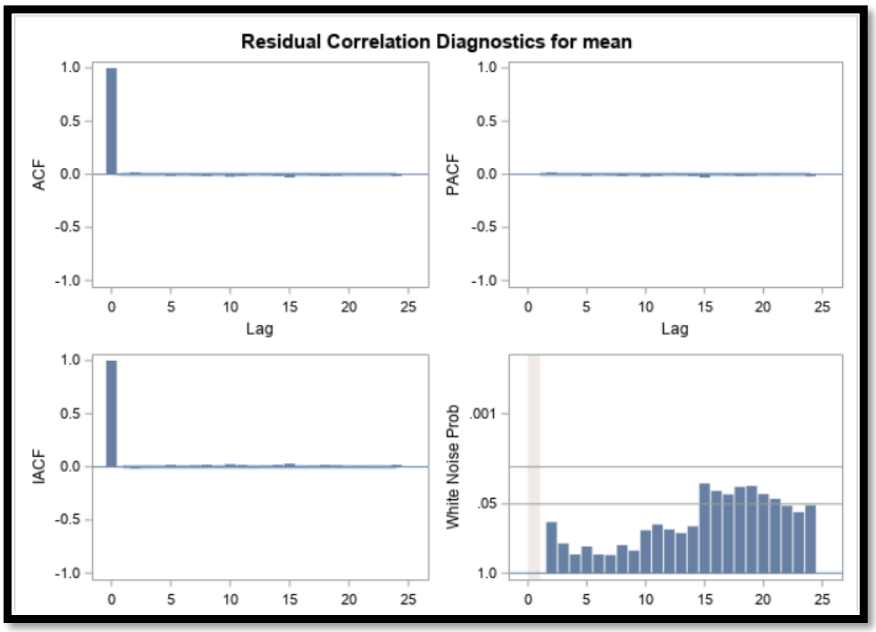
With the same, below were the outputs:

I) Correlation for mean:



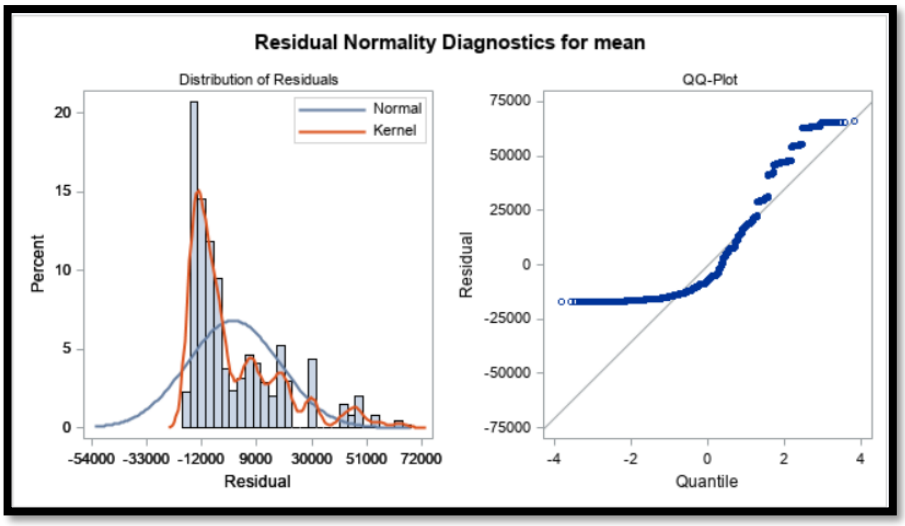
Through ACF graph, it is evident that there is no autocorrelation. Also, there are extremely small spikes in the PACF and the IACF graphs.

II) Residual plots:



The results show that the white noise test fails for the residuals with some spikes above the threshold and hence, the model is not able to capture all the systematic variation existing in the dataset well. Also, there is no autocorrelation that exists in the residual plot and hence is great.

III) Residual normality check:



The above plot indicates that the residual graph again is not normal and shows right skewness and kurtosis.

IV) Autocorrelation check for residuals:

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.75	5	0.4467	0.000	0.017	0.001	0.003	-0.015	-0.001
12	15.74	11	0.1509	-0.010	-0.017	-0.004	-0.022	-0.015	0.007
18	30.37	17	0.0238	-0.008	-0.016	-0.029	-0.005	-0.009	-0.016
24	34.93	23	0.0528	-0.013	0.000	-0.007	-0.001	-0.002	-0.017
30	46.80	29	0.0195	-0.031	-0.002	-0.007	-0.001	-0.016	-0.001
36	65.23	35	0.0014	-0.029	-0.010	-0.008	0.009	-0.023	-0.020
42	70.87	41	0.0026	-0.007	0.006	-0.016	-0.010	0.013	0.004
48	74.58	47	0.0064	0.005	0.013	-0.012	-0.002	-0.004	0.007

The above table indicates that there is barely any auto – correlation existing amongst the residuals and hence is great.

V) Parameter estimates:

Parameter	Estimate	Error	t Value	Pr > |t|	Lag	Variable	Shift
MU	24775.6	26776.6	0.93	0.3548	0	mean	0
AR1,1	-0.0018139	0.01038	-0.17	0.8613	1	mean	0
NUM1	14.78101	13.61650	1.09	0.2777	0	Temperature	0
NUM2	-59.84441	703.11566	-0.09	0.9322	0	Fuel_Price	0
NUM3	-41.05872	103.70116	-0.40	0.6922	0	CPI	0
NUM4	3.48427	889.04441	0.00	0.9969	0	Unemployment	0
NUM5	-417.65391	790.37254	-0.53	0.5972	0	IsHoliday	0

Constant Estimate	24820.54
Variance Estimate	3.0857E8
Std Error Estimate	17566.03
AIC	207742.8
SBC	207792.7
Number of Residuals	9280

Correlations of Parameter Estimates							
Variable Parameter	mean MU	mean AR1,1	Temperature NUM1	Fuel_Price NUM2	CPI NUM3	Unemployment NUM4	IsHoliday NUM5
mean MU	1.000	0.000	-0.211	0.595	-0.980	-0.883	-0.033
mean AR1,1	0.000	1.000	-0.000	-0.000	-0.000	-0.000	-0.002
Temperature NUM1	-0.211	-0.000	1.000	-0.243	0.181	0.177	0.147
Fuel_Price NUM2	0.595	-0.000	-0.243	1.000	-0.703	-0.328	0.016
CPI NUM3	-0.980	-0.000	0.181	-0.703	1.000	0.780	0.023
Unemployment NUM4	-0.883	-0.000	0.177	-0.328	0.780	1.000	0.021
IsHoliday NUM5	-0.033	-0.002	0.147	0.016	0.023	0.021	1.000

The parameter estimates indicate high values for the independent variables (>0.001) and hence indicate that the model needs to be fine-tuned further to generate better sales forecasts.

VI) Model and forecast outputs:

Model for variable mean	
Estimated Intercept	24775.6

Autoregressive Factors	
Factor 1:	1 + 0.00181 B**(1)

Input Number 1	
Input Variable	Temperature
Overall Regression Factor	14.78101

Input Number 2	
Input Variable	Fuel_Price
Overall Regression Factor	-59.8444

Input Number 3	
Input Variable	CPI
Overall Regression Factor	-41.0587

Input Number 4	
Input Variable	Unemployment
Overall Regression Factor	3.484273

Input Number 5	
Input Variable	IsHoliday
Overall Regression Factor	-417.654

Forecasts for variable mean						
Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
9269	16452.3400	17566.027	-17976.4411	50881.1210	9196.5579	-7255.7821
9270	16436.5822	17566.056	-17992.2554	50865.4199	6187.3897	-10249.1925
9271	16436.6108	17566.056	-17992.2269	50865.4485	1374.6497	-15061.9612
9272	16436.6108	17566.056	-17992.2269	50865.4484	21623.1847	5186.5739
9273	16436.6108	17566.056	-17992.2269	50865.4484	7808.4506	-8628.1602
9274	16436.6108	17566.056	-17992.2269	50865.4484	20727.8609	4291.2501
9275	16436.6108	17566.056	-17992.2269	50865.4484	11652.8274	-4783.7834
9276	16436.6108	17566.056	-17992.2269	50865.4484	63180.5682	46743.9574
9277	16436.6108	17566.056	-17992.2269	50865.4484	17139.3146	702.7039
9278	16436.6108	17566.056	-17992.2269	50865.4484	4091.5715	-12345.0392
9279	16436.6108	17566.056	-17992.2269	50865.4484	8052.0762	-8384.5346
9280	16436.6108	17566.056	-17992.2269	50865.4484	7436.0173	-9000.5934

The forecast is shown for 12 months as that is the hold out period and residual is calculated only for that duration.

Model 4: ARIMAX(0,1)

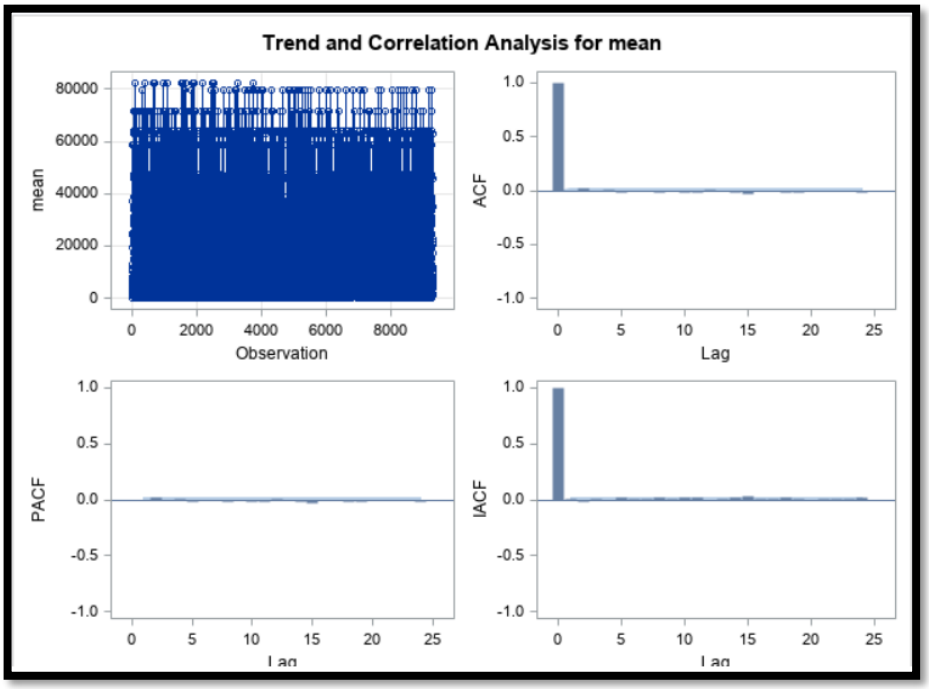
The third model was created using ARIMAX and had below parameters:

p = 0

q = 1

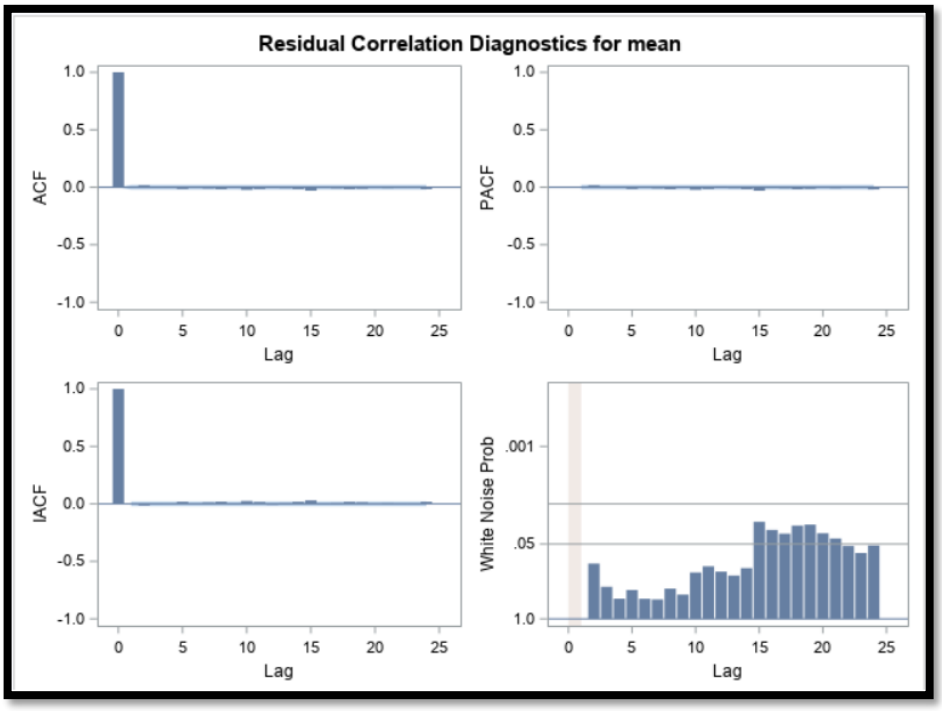
With the same, below were the outputs:

I) Correlation for mean:



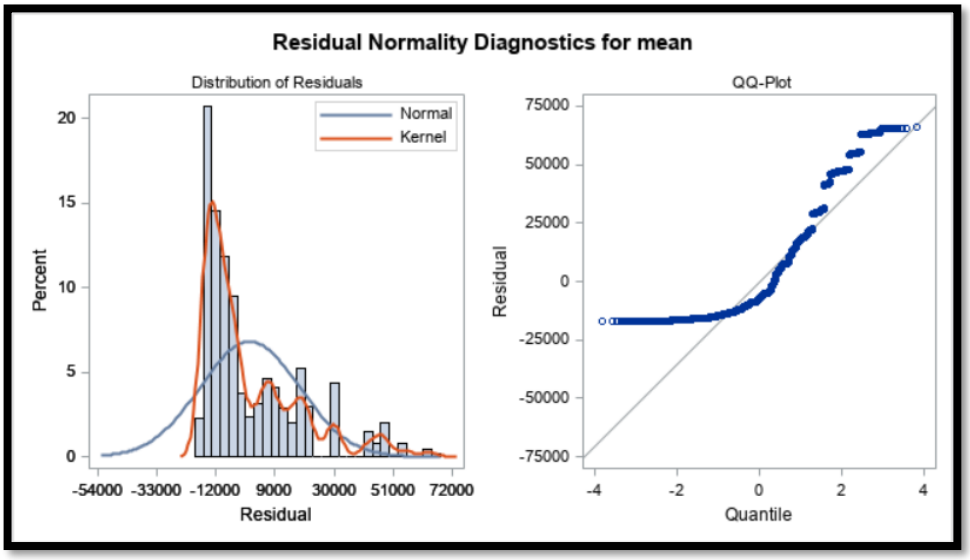
Through ACF graph, it is evident that there is no autocorrelation. Also, there are extremely small spikes in the PACF and the IACF graphs.

II) Residual plots:



The results show that the white noise test fails for the residuals with some spikes above the threshold and hence, the model is not able to capture all the systematic variation existing in the dataset well. Also, there is no autocorrelation that exists in the residual plot and hence is great.

III) Residual normality check:



The above plot indicates that the residual graph again is not normal and shows right skewness and kurtosis.

IV) Autocorrelation check for residuals:

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.75	5	0.4465	-0.000	0.017	0.001	0.003	-0.015	-0.001
12	15.74	11	0.1509	-0.010	-0.017	-0.004	-0.022	-0.015	0.007
18	30.36	17	0.0238	-0.008	-0.016	-0.029	-0.005	-0.009	-0.016
24	34.93	23	0.0529	-0.013	0.000	-0.007	-0.001	-0.002	-0.017
30	46.79	29	0.0196	-0.031	-0.002	-0.007	-0.001	-0.016	-0.001
36	65.23	35	0.0014	-0.029	-0.010	-0.008	0.009	-0.023	-0.020
42	70.87	41	0.0026	-0.007	0.006	-0.016	-0.010	0.013	0.004
48	74.57	47	0.0064	0.005	0.013	-0.012	-0.002	-0.004	0.007

The above table indicates that there is barely any auto – correlation existing amongst the residuals and hence is great.

V) Parameter estimates:

Parameter	Estimate	Error	t Value	Pr > |t|	Lag	Variable	Shift
MU	24775.6	26778.5	0.93	0.3549	0	mean	0
MA1,1	0.0017395	0.01038	0.17	0.8670	1	mean	0
NUM1	14.78097	13.61747	1.09	0.2777	0	Temperature	0
NUM2	-59.84472	703.16580	-0.09	0.9322	0	Fuel_Price	0
NUM3	-41.05870	103.70855	-0.40	0.6922	0	CPI	0
NUM4	3.48384	889.10776	0.00	0.9969	0	Unemployment	0
NUM5	-417.65779	790.42820	-0.53	0.5972	0	IsHoliday	0

Constant Estimate	24775.6
Variance Estimate	3.0857E8
Std Error Estimate	17566.03
AIC	207742.8
SBC	207792.7
Number of Residuals	9280

Correlations of Parameter Estimates							
Variable Parameter	mean MU	mean MA1,1	Temperature NUM1	Fuel_Price NUM2	CPI NUM3	Unemployment NUM4	IsHoliday NUM5
mean MU	1.000	-0.000	-0.211	0.595	-0.980	-0.883	-0.033
mean MA1,1	-0.000	1.000	0.000	0.000	0.000	0.000	0.002
Temperature NUM1	-0.211	0.000	1.000	-0.243	0.181	0.177	0.147
Fuel_Price NUM2	0.595	0.000	-0.243	1.000	-0.703	-0.328	0.016
CPI NUM3	-0.980	0.000	0.181	-0.703	1.000	0.780	0.023
Unemployment NUM4	-0.883	0.000	0.177	-0.328	0.780	1.000	0.021
IsHoliday NUM5	-0.033	0.002	0.147	0.016	0.023	0.021	1.000

The parameter estimates indicate high values for the independent variables (>0.001) and hence indicate that the model needs to be fine-tuned further to generate better sales forecasts.

VI) Model and forecast outputs:

Model for variable mean	
Estimated Intercept	24775.6

Moving Average Factors	
Factor 1:	1 - 0.00174 B**(1)

Input Number 1	
Input Variable	Temperature
Overall Regression Factor	14.78097

Input Number 2	
Input Variable	Fuel_Price
Overall Regression Factor	-59.8447

Input Number 3	
Input Variable	CPI
Overall Regression Factor	-41.0587

Input Number 4	
Input Variable	Unemployment
Overall Regression Factor	3.483844

Input Number 5	
Input Variable	IsHoliday
Overall Regression Factor	-417.658

Forecasts for variable mean						
Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
9269	16451.7407	17566.028	-17977.0422	50880.5235	9196.5579	-7255.1828
9270	16436.6113	17566.055	-17992.2237	50865.4462	6187.3897	-10249.2216
9271	16436.6113	17566.055	-17992.2237	50865.4462	1374.6497	-15061.9616
9272	16436.6113	17566.055	-17992.2237	50865.4462	21623.1847	5186.5734
9273	16436.6113	17566.055	-17992.2237	50865.4462	7808.4506	-8628.1607
9274	16436.6113	17566.055	-17992.2237	50865.4462	20727.8609	4291.2496
9275	16436.6113	17566.055	-17992.2237	50865.4462	11652.8274	-4783.7839
9276	16436.6113	17566.055	-17992.2237	50865.4462	63180.5682	46743.9569
9277	16436.6113	17566.055	-17992.2237	50865.4462	17139.3146	702.7033
9278	16436.6113	17566.055	-17992.2237	50865.4462	4091.5715	-12345.0398
9279	16436.6113	17566.055	-17992.2237	50865.4462	8052.0762	-8384.5351
9280	16436.6113	17566.055	-17992.2237	50865.4462	7436.0173	-9000.5939

The forecast is shown for 12 months as that is the hold out period and residual is calculated only for that duration.

Model 5: ARIMAX(1,1)

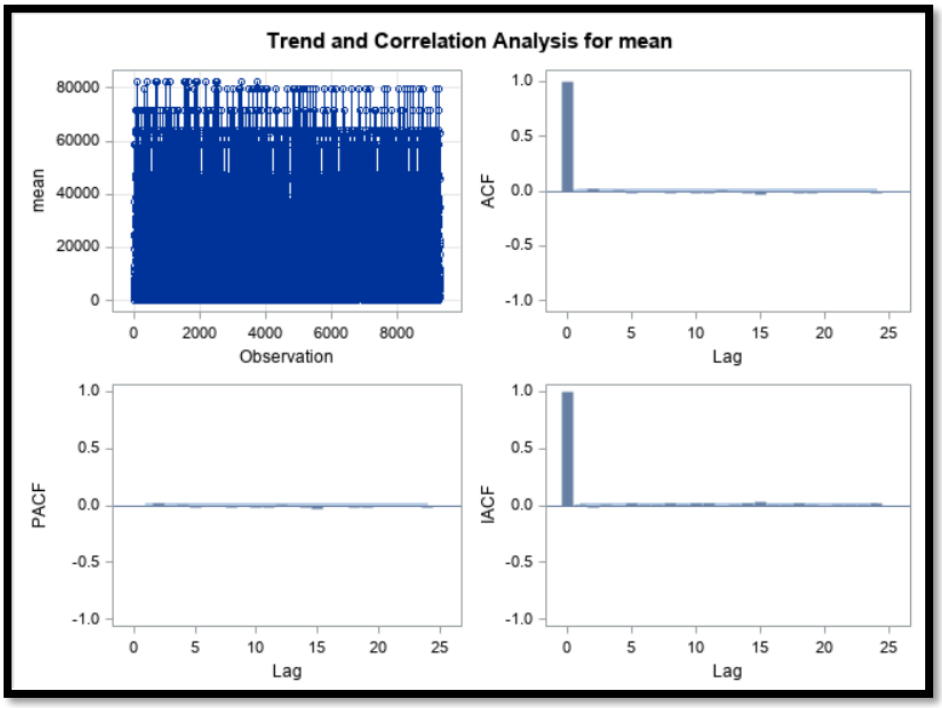
The fourth model was created using ARIMAX and had below parameters:

p = 1

q = 1

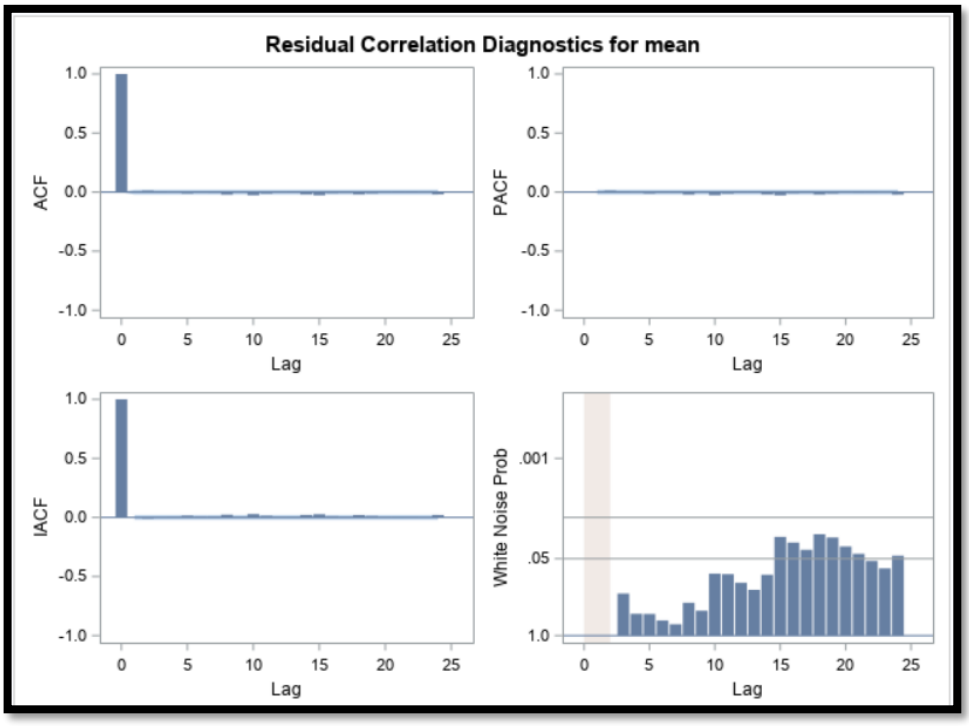
With the same, below were the outputs:

I) Correlation for mean:



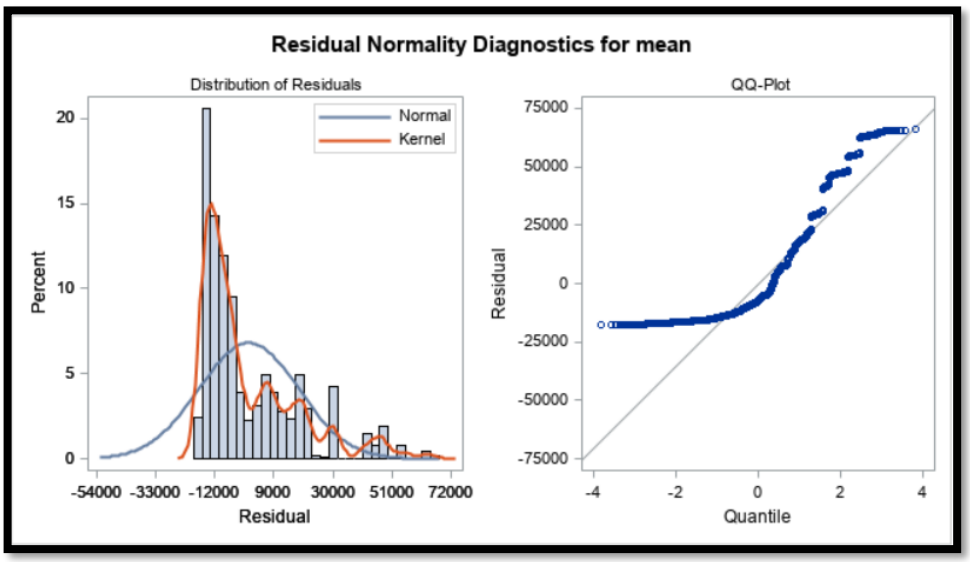
Through ACF graph, it is evident that there is no autocorrelation. Also, there are extremely small spikes in the PACF and the IACF graphs.

II) Residual plots:



The results show that the white noise test fails for the residuals with some spikes above the threshold and hence, the model is not able to capture all the systematic variation existing in the dataset well. Also, there is no autocorrelation that exists in the residual plot and hence is great.

III) Residual normality check:



The above plot indicates that the residual graph again is not normal and shows right skewness and kurtosis.

IV) Autocorrelation check for residuals:

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	3.03	4	0.5532	0.003	0.012	0.005	-0.001	-0.011	-0.005
12	15.12	10	0.1279	-0.006	-0.021	-0.000	-0.026	-0.012	0.004
18	29.77	16	0.0192	-0.005	-0.019	-0.026	-0.008	-0.006	-0.019
24	34.43	22	0.0443	-0.010	-0.002	-0.004	-0.004	0.001	-0.019
30	44.52	28	0.0247	-0.029	-0.004	-0.004	-0.003	-0.014	-0.003
36	61.86	34	0.0024	-0.027	-0.012	-0.006	0.007	-0.021	-0.021
42	67.36	40	0.0044	-0.005	0.004	-0.014	-0.011	0.014	0.002
48	70.57	46	0.0114	0.006	0.012	-0.011	-0.003	-0.002	0.006

The above table indicates that there is barely any auto – correlation existing amongst the residuals and hence is great.

V) Parameter estimates:

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	24769.5	26760.1	0.93	0.3546	0	mean	0
MA1,1	-0.96438	0.03346	-28.82	<.0001	1	mean	0
AR1,1	-0.96893	0.03087	-31.39	<.0001	1	mean	0
NUM1	14.77686	13.60810	1.09	0.2775	0	Temperature	0
NUM2	-59.46784	702.68252	-0.08	0.9326	0	Fuel_Price	0
NUM3	-41.04883	103.63724	-0.40	0.6920	0	CPI	0
NUM4	3.85778	888.49626	0.00	0.9965	0	Unemployment	0
NUM5	-414.13334	789.87792	-0.52	0.6001	0	IsHoliday	0

Constant Estimate	48769.22
Variance Estimate	3.085E8
Std Error Estimate	17564.04
AIC	207741.7
SBC	207798.8
Number of Residuals	9280

Correlations of Parameter Estimates								
Variable Parameter	mean MU	mean MA1,1	mean AR1,1	Temperature NUM1	Fuel_Price NUM2	CPI NUM3	Unemployment NUM4	IsHoliday NUM5
mean MU	1.000	-0.000	-0.000	-0.211	0.595	-0.980	-0.883	-0.033
mean MA1,1	-0.000	1.000	0.997	0.000	-0.000	0.000	0.000	0.000
mean AR1,1	-0.000	0.997	1.000	0.000	-0.000	0.000	0.000	-0.000
Temperature NUM1	-0.211	0.000	0.000	1.000	-0.243	0.181	0.177	0.147
Fuel_Price NUM2	0.595	-0.000	-0.000	-0.243	1.000	-0.703	-0.328	0.016
CPI NUM3	-0.980	0.000	0.000	0.181	-0.703	1.000	0.780	0.023
Unemployment NUM4	-0.883	0.000	0.000	0.177	-0.328	0.780	1.000	0.021
IsHoliday NUM5	-0.033	0.000	-0.000	0.147	0.016	0.023	0.021	1.000

The parameter estimates indicate low values for the independent variables (<0.001) and hence indicate that the model is good to generate sales forecasts.

VI) Model and forecast outputs:

Model for variable mean	
Estimated Intercept	24769.46

Autoregressive Factors	
Factor 1:	1 + 0.96893 B**(1)

Moving Average Factors	
Factor 1:	1 + 0.96438 B**(1)

Input Number 1	
Input Variable	Temperature
Overall Regression Factor	14.77686

Input Number 2	
Input Variable	Fuel_Price
Overall Regression Factor	-59.4678

Input Number 3	
Input Variable	CPI
Overall Regression Factor	-41.0488

Input Number 4	
Input Variable	Unemployment
Overall Regression Factor	3.857778

Model 6: ARIMAX(2,0)

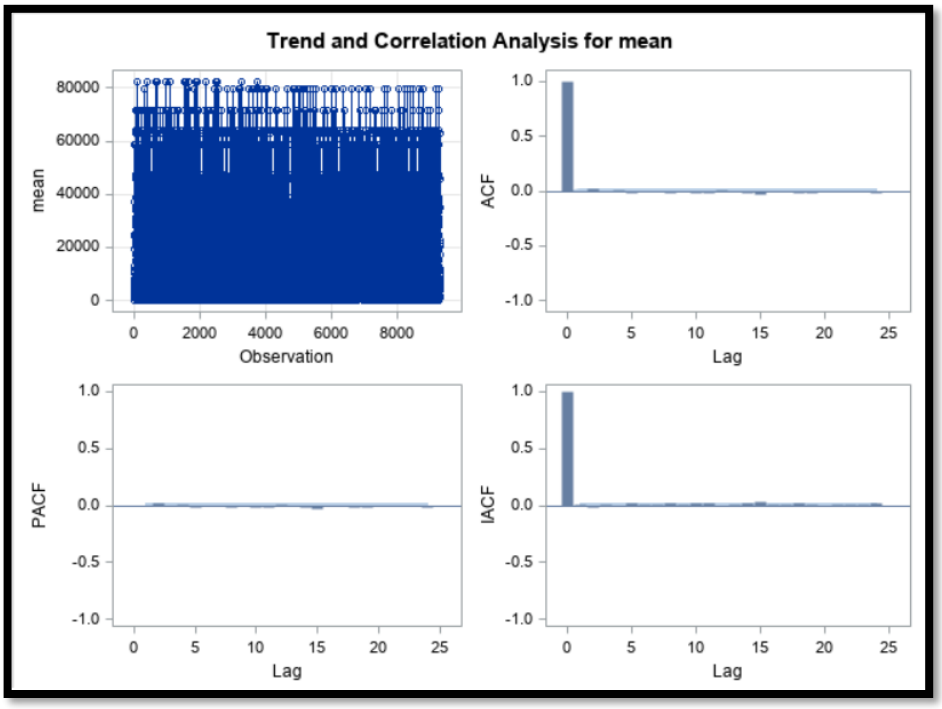
The fourth model was created using ARIMAX and had below parameters:

p = 2

q = 0

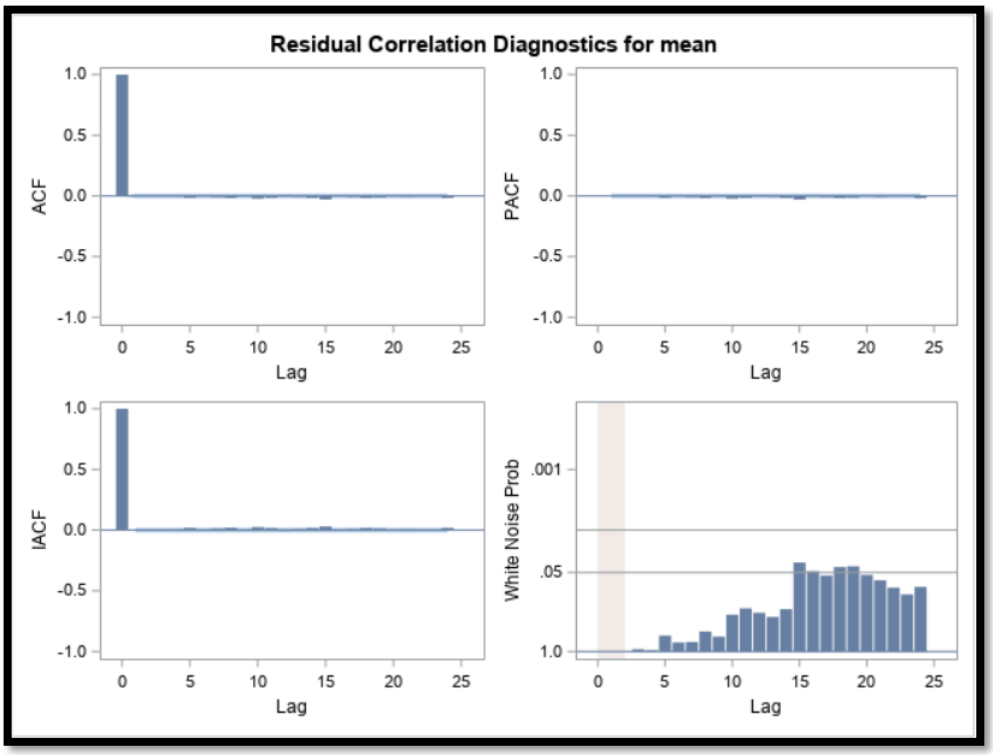
With the same, below were the outputs:

I) Correlation for mean:



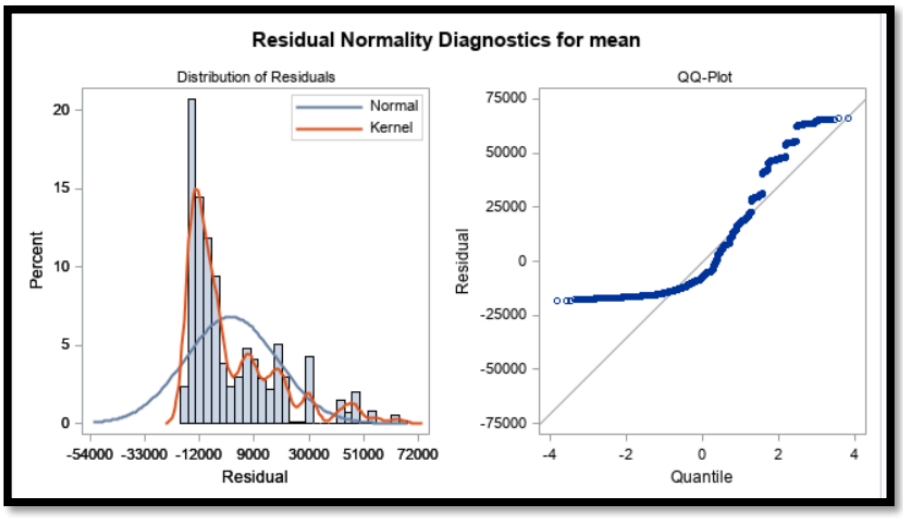
Through ACF graph, it is evident that there is no autocorrelation. Also, there are extremely small spikes in the PACF and the IACF graphs.

II) Residual plots:



The results show that the white noise test fails for the residuals with some spikes above the threshold and hence, the model is not able to capture all the systematic variation existing in the dataset well. Also, there is no autocorrelation that exists in the residual plot and hence is great.

III) Residual normality check:



The above plot indicates that the residual graph again is not normal and shows right skewness and kurtosis.

IV) Autocorrelation check for residuals:

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.13	4	0.7112	-0.000	-0.000	0.001	0.003	-0.015	-0.001
12	12.88	10	0.2303	-0.010	-0.017	-0.004	-0.022	-0.015	0.008
18	27.08	16	0.0406	-0.007	-0.016	-0.029	-0.005	-0.009	-0.016
24	31.52	22	0.0860	-0.012	0.001	-0.007	-0.001	-0.001	-0.017
30	43.03	28	0.0346	-0.031	-0.002	-0.006	-0.001	-0.016	-0.000
36	61.26	34	0.0028	-0.028	-0.010	-0.007	0.009	-0.023	-0.020
42	66.99	40	0.0048	-0.006	0.006	-0.016	-0.010	0.013	0.004
48	70.70	46	0.0111	0.005	0.013	-0.012	-0.002	-0.003	0.007

The above table indicates that there is barely any auto – correlation existing amongst the residuals and hence is great.

V) Parameter estimates:

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	24778.8	27225.6	0.91	0.3628	0	mean	0
AR1,1	-0.0017823	0.01038	-0.17	0.8637	1	mean	0
AR1,2	0.01662	0.01038	1.60	0.1094	2	mean	0
NUM1	14.75104	13.84460	1.07	0.2867	0	Temperature	0
NUM2	-59.17318	714.92192	-0.08	0.9340	0	Fuel_Price	0
NUM3	-41.07506	105.44077	-0.39	0.6969	0	CPI	0
NUM4	3.48564	903.94635	0.00	0.9969	0	Unemployment	0
NUM5	-415.40732	803.27905	-0.52	0.6051	0	IsHoliday	0

Constant Estimate	24411.08
Variance Estimate	3.0851E8
Std Error Estimate	17564.55
AIC	207742.2
SBC	207799.3
Number of Residuals	9280

Correlations of Parameter Estimates

Variable Parameter	mean MU	mean AR1,1	mean AR1,2	Temperature NUM1	Fuel_Price NUM2	CPI NUM3	Unemployment NUM4	IsHoliday NUM5
mean MU	1.000	0.000	0.000	-0.211	0.595	-0.980	-0.883	-0.033
mean AR1,1	0.000	1.000	0.002	-0.001	-0.000	-0.000	-0.000	-0.002
mean AR1,2	0.000	0.002	1.000	-0.001	0.000	-0.000	-0.000	-0.001
Temperature NUM1	-0.211	-0.001	-0.001	1.000	-0.243	0.181	0.177	0.147
Fuel_Price NUM2	0.595	-0.000	0.000	-0.243	1.000	-0.703	-0.328	0.016
CPI NUM3	-0.980	-0.000	-0.000	0.181	-0.703	1.000	0.780	0.023
Unemployment NUM4	-0.883	-0.000	-0.000	0.177	-0.328	0.780	1.000	0.021
IsHoliday NUM5	-0.033	-0.002	-0.001	0.147	0.016	0.023	0.021	1.000

The parameter estimates indicate high values for the independent variables (>0.001) and hence indicate that the model is not good to generate sales forecasts.

VI) Model and forecast outputs:

Model for variable mean	
Estimated Intercept	24778.81

Autoregressive Factors	
Factor 1:	1 + 0.00178 B**(1) - 0.01662 B**(2)

Input Number 1	
Input Variable	Temperature
Overall Regression Factor	14.75104

Input Number 2	
Input Variable	Fuel_Price
Overall Regression Factor	-59.1732

Input Number 3	
Input Variable	CPI
Overall Regression Factor	-41.0751

Input Number 4	
Input Variable	Unemployment
Overall Regression Factor	3.485642

Input Number 5	
Input Variable	IsHoliday
Overall Regression Factor	-415.407

Forecasts for variable mean						
Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
9269	16199.6174	17564.547	-18226.2619	50625.4967	9196.5579	-7003.0595
9270	16292.7363	17564.575	-18133.1976	50718.6703	6187.3897	-10105.3466
9271	16432.7708	17567.002	-17997.9208	50863.4624	1374.6497	-15058.1211
9272	16434.0691	17567.002	-17996.6226	50864.7607	21623.1847	5189.1156
9273	16436.3945	17567.003	-17994.2985	50867.0875	7808.4506	-8627.9439
9274	16436.4119	17567.003	-17994.2811	50867.1049	20727.8609	4291.4490
9275	16436.4506	17567.003	-17994.2424	50867.1435	11652.8274	-4783.6231
9276	16436.4508	17567.003	-17994.2422	50867.1438	63180.5682	46744.1174
9277	16436.4514	17567.003	-17994.2416	50867.1444	17139.3146	702.8632
9278	16436.4514	17567.003	-17994.2416	50867.1444	4091.5715	-12344.8799
9279	16436.4514	17567.003	-17994.2416	50867.1444	8052.0762	-8384.3753
9280	16436.4514	17567.003	-17994.2416	50867.1444	7436.0173	-9000.4341

The forecast is shown for 12 months as that is the hold out period and residual is calculated only for that duration.

Model 7: ARIMAX(0,2)

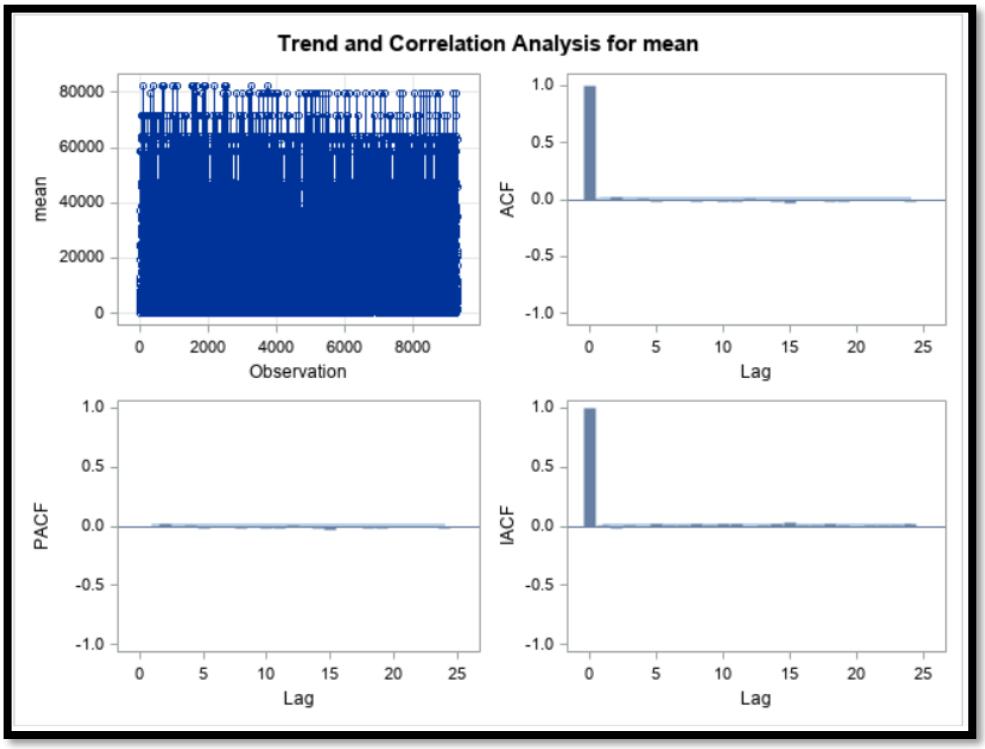
The fourth model was created using ARIMAX and had below parameters:

p = 0

q = 2

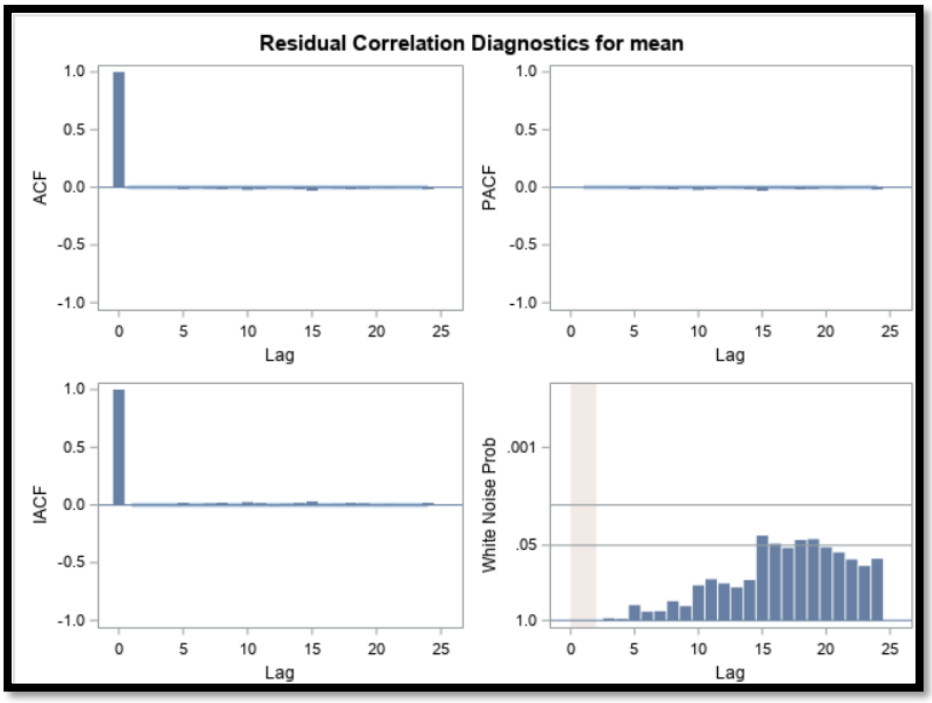
With the same, below were the outputs:

I) Correlation for mean:



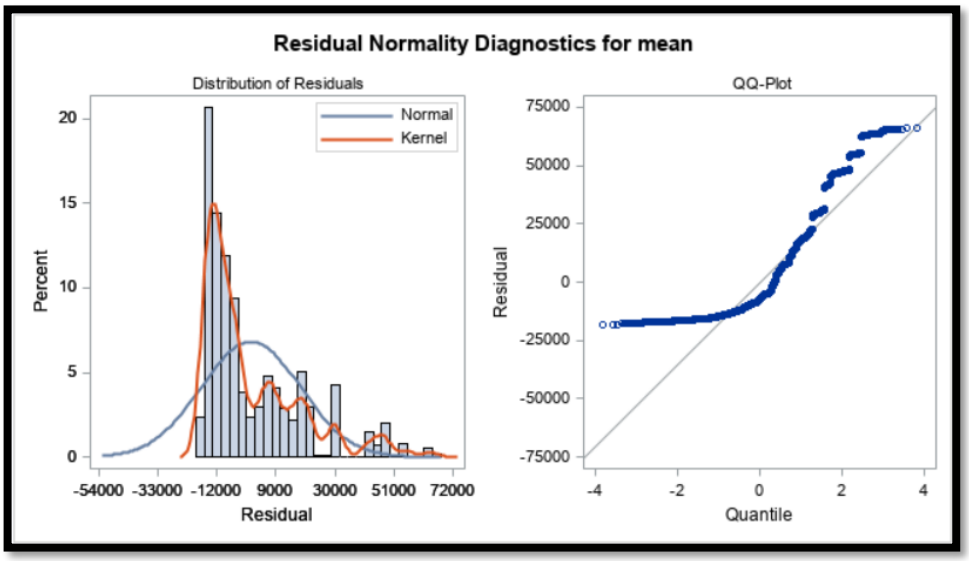
Through ACF graph, it is evident that there is no autocorrelation. Also, there are extremely small spikes in the PACF and the IACF graphs.

II) Residual plots:



The results show that the white noise test fails for the residuals with some spikes above the threshold and hence, the model is not able to capture all the systematic variation existing in the dataset well. Also, there is no autocorrelation that exists in the residual plot and hence is great.

III) Residual normality check:



The above plot indicates that the residual graph again is not normal and shows right skewness and kurtosis.

IV) Autocorrelation check for residuals:

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.15	4	0.7084	0.000	0.000	0.001	0.003	-0.015	-0.001
12	12.90	10	0.2291	-0.010	-0.017	-0.004	-0.022	-0.015	0.008
18	27.11	16	0.0403	-0.007	-0.016	-0.029	-0.005	-0.009	-0.016
24	31.55	22	0.0854	-0.012	0.001	-0.007	-0.001	-0.001	-0.017
30	43.08	28	0.0342	-0.031	-0.002	-0.006	-0.001	-0.016	-0.000
36	61.32	34	0.0028	-0.028	-0.010	-0.007	0.009	-0.023	-0.020
42	67.05	40	0.0047	-0.006	0.006	-0.016	-0.010	0.013	0.004
48	70.76	46	0.0110	0.005	0.013	-0.012	-0.002	-0.003	0.007

The above table indicates that there is barely any auto – correlation existing amongst the residuals and hence is great.

V) Parameter estimates:

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	24780.3	27216.2	0.91	0.3626	0	mean	0
MA1,1	0.0018215	0.01038	0.18	0.8608	1	mean	0
MA1,2	-0.01653	0.01038	-1.59	0.1114	2	mean	0
NUM1	14.75098	13.83978	1.07	0.2865	0	Temperature	0
NUM2	-59.15131	714.67207	-0.08	0.9340	0	Fuel_Price	0
NUM3	-41.08048	105.40398	-0.39	0.6967	0	CPI	0
NUM4	3.44082	903.63134	0.00	0.9970	0	Unemployment	0
NUM5	-415.46846	803.01082	-0.52	0.6049	0	IsHoliday	0

Constant Estimate	24780.26
Variance Estimate	3.0851E8
Std Error Estimate	17564.56
AIC	207742.2
SBC	207799.3
Number of Residuals	9280

Correlations of Parameter Estimates								
Variable Parameter	mean MU	mean MA1,1	mean MA1,2	Temperature NUM1	Fuel_Price NUM2	CPI NUM3	Unemployment NUM4	IsHoliday NUM5
mean MU	1.000	-0.000	-0.000	-0.211	0.595	-0.980	-0.883	-0.033
mean MA1,1	-0.000	1.000	-0.003	0.000	0.000	0.000	0.000	0.002
mean MA1,2	-0.000	-0.003	1.000	0.001	-0.000	0.000	0.000	0.001
Temperature NUM1	-0.211	0.000	0.001	1.000	-0.243	0.181	0.177	0.147
Fuel_Price NUM2	0.595	0.000	-0.000	-0.243	1.000	-0.703	-0.328	0.016
CPI NUM3	-0.980	0.000	0.000	0.181	-0.703	1.000	0.780	0.023
Unemployment NUM4	-0.883	0.000	0.000	0.177	-0.328	0.780	1.000	0.021
IsHoliday NUM5	-0.033	0.002	0.001	0.147	0.016	0.023	0.021	1.000

The parameter estimates indicate high values for the independent variables (>0.001) and hence indicate that the model is not good to generate sales forecasts.

VI) Model and forecast outputs:

Model for variable mean	
Estimated Intercept	24780.26

Moving Average Factors	
Factor 1:	1 - 0.00182 B**(1) + 0.01653 B**(2)

Input Number 1	
Input Variable	Temperature
Overall Regression Factor	14.75098

Input Number 2	
Input Variable	Fuel_Price
Overall Regression Factor	-59.1513

Input Number 3	
Input Variable	CPI
Overall Regression Factor	-41.0805

Input Number 4	
Input Variable	Unemployment
Overall Regression Factor	3.440819

Input Number 5	
Input Variable	IsHoliday
Overall Regression Factor	-415.468

Forecasts for variable mean

Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
9269	16203.7100	17564.561	-18222.1972	50629.6171	9196.5579	-7007.1521
9270	16294.3551	17564.590	-18131.6092	50720.3194	6187.3897	-10106.9654
9271	16436.4692	17566.989	-17994.1974	50867.1358	1374.6497	-15061.8196
9272	16436.4692	17566.989	-17994.1974	50867.1358	21623.1847	5186.7155
9273	16436.4692	17566.989	-17994.1974	50867.1358	7808.4506	-8628.0186
9274	16436.4692	17566.989	-17994.1974	50867.1358	20727.8609	4291.3917
9275	16436.4692	17566.989	-17994.1974	50867.1358	11652.8274	-4783.6418
9276	16436.4692	17566.989	-17994.1974	50867.1358	63180.5682	46744.0990
9277	16436.4692	17566.989	-17994.1974	50867.1358	17139.3146	702.8454
9278	16436.4692	17566.989	-17994.1974	50867.1358	4091.5715	-12344.8977
9279	16436.4692	17566.989	-17994.1974	50867.1358	8052.0762	-8384.3931
9280	16436.4692	17566.989	-17994.1974	50867.1358	7436.0173	-9000.4519

The forecast is shown for 12 months as that is the hold out period and residual is calculated only for that duration.

Model 8: ARIMAX(2,2)

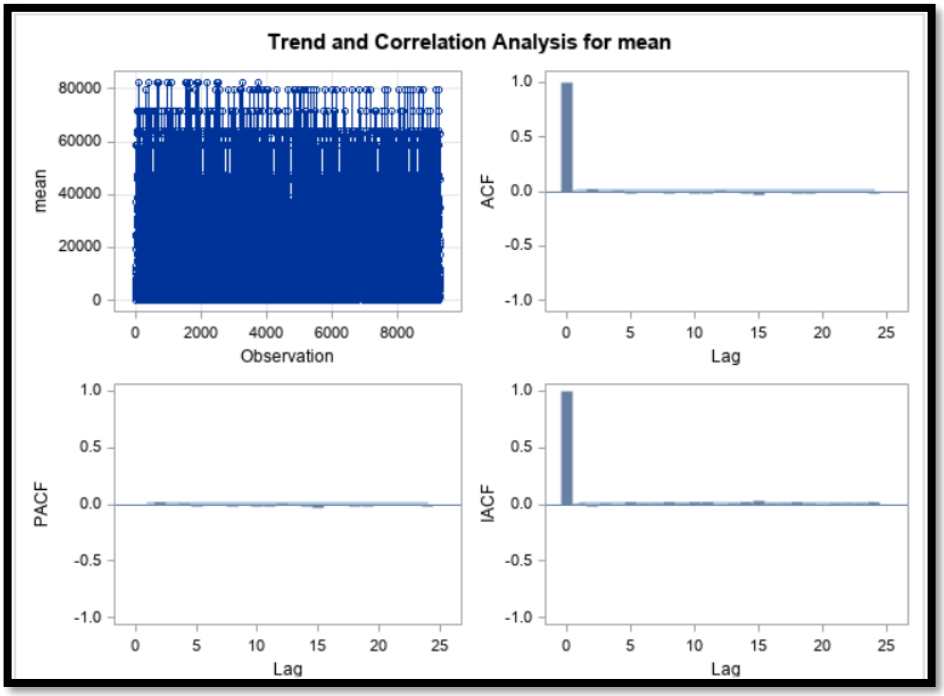
The fourth model was created using ARIMAX and had below parameters:

p = 2

q = 2

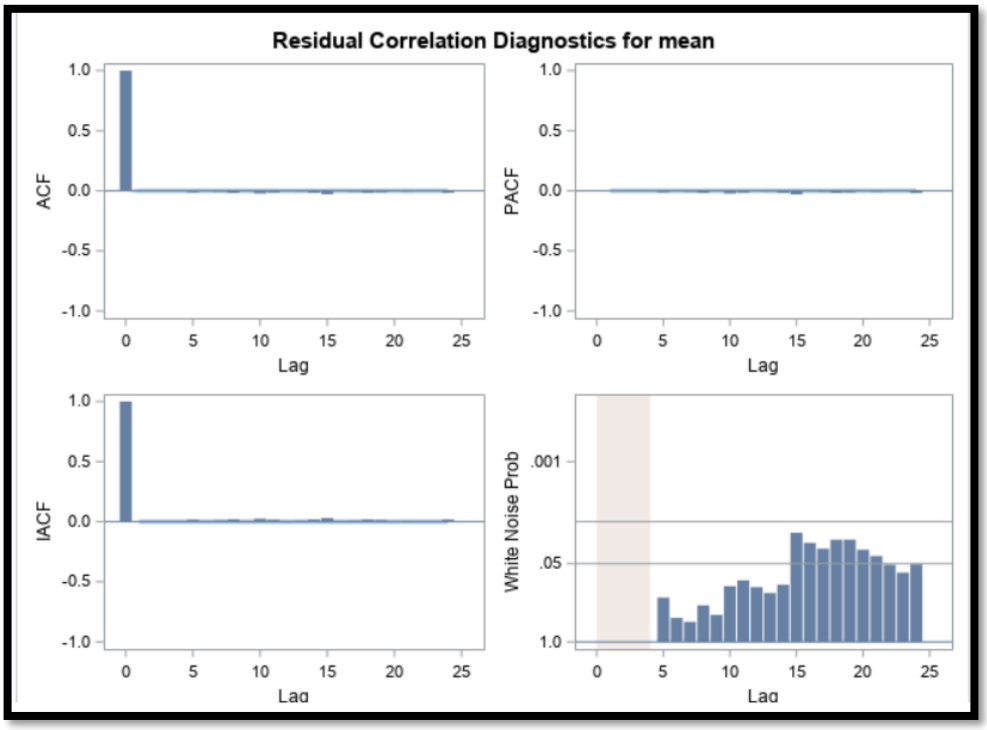
With the same, below were the outputs:

I) Correlation for mean:



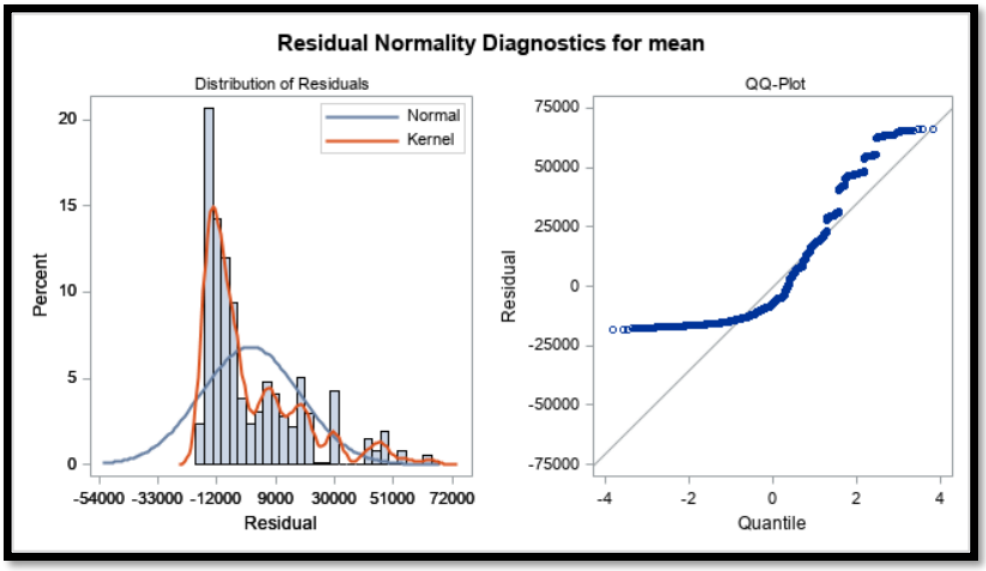
Through ACF graph, it is evident that there is no autocorrelation. Also, there are extremely small spikes in the PACF and the IACF graphs.

II) Residual plots:



The results show that the white noise test fails for the residuals with some spikes above the threshold and hence, the model is not able to capture all the systematic variation existing in the dataset well. Also, there is no autocorrelation that exists in the residual plot and hence is great.

III) Residual normality check:



The above plot indicates that the residual graph again is not normal and shows right skewness and kurtosis.

IV) Autocorrelation check for residuals:

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	1.84	2	0.3987	-0.001	0.001	0.005	-0.002	-0.013	-0.003
12	12.70	8	0.1227	-0.009	-0.018	-0.003	-0.022	-0.015	0.007
18	26.88	14	0.0199	-0.007	-0.016	-0.029	-0.005	-0.009	-0.016
24	31.32	20	0.0512	-0.012	0.001	-0.007	-0.001	-0.001	-0.017
30	42.70	26	0.0207	-0.031	-0.002	-0.006	-0.001	-0.015	-0.000
36	60.75	32	0.0016	-0.028	-0.010	-0.007	0.009	-0.023	-0.020
42	66.47	38	0.0029	-0.006	0.006	-0.016	-0.010	0.013	0.004
48	70.18	44	0.0073	0.005	0.013	-0.012	-0.002	-0.003	0.007

The above table indicates that there is barely any auto – correlation existing amongst the residuals and hence is great.

V) Parameter estimates:

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	24749.4	27210.9	0.91	0.3631	0	mean	0
MA1,1	-0.24797	0.66507	-0.37	0.7093	1	mean	0
MA1,2	0.22971	0.68375	0.34	0.7369	2	mean	0
AR1,1	-0.24912	0.66281	-0.38	0.7070	1	mean	0
AR1,2	0.24534	0.68365	0.36	0.7197	2	mean	0
NUM1	14.75598	13.83711	1.07	0.2862	0	Temperature	0
NUM2	-59.66141	714.53510	-0.08	0.9335	0	Fuel_Price	0
NUM3	-40.96491	105.38367	-0.39	0.6975	0	CPI	0
NUM4	4.37094	903.45650	0.00	0.9961	0	Unemployment	0
NUM5	-414.23889	802.83433	-0.52	0.6059	0	IsHoliday	0

Constant Estimate	24843.08
Variance Estimate	3.0857E8
Std Error Estimate	17566.2
AIC	207746
SBC	207817.3
Number of Residuals	9280

Correlations of Parameter Estimates										
Variable Parameter	mean MU	mean MA1,1	mean MA1,2	mean AR1,1	mean AR1,2	Temperature NUM1	Fuel_Price NUM2	CPI NUM3	Unemployment NUM4	IsHoliday NUM5
mean MU	1.000	0.000	0.000	0.000	0.000	-0.211	0.595	-0.980	-0.883	-0.033
mean MA1,1	0.000	1.000	0.696	1.000	0.701	-0.002	0.000	-0.000	-0.000	-0.001
mean MA1,2	0.000	0.696	1.000	0.697	1.000	-0.001	0.000	-0.000	-0.000	-0.000
mean AR1,1	0.000	1.000	0.697	1.000	0.702	-0.002	0.000	-0.000	-0.000	-0.001
mean AR1,2	0.000	0.701	1.000	0.702	1.000	-0.001	0.000	-0.000	-0.000	-0.000
Temperature NUM1	-0.211	-0.002	-0.001	-0.002	-0.001	1.000	-0.243	0.181	0.177	0.147
Fuel_Price NUM2	0.595	0.000	0.000	0.000	0.000	-0.243	1.000	-0.703	-0.328	0.016
CPI NUM3	-0.980	-0.000	-0.000	-0.000	-0.000	0.181	-0.703	1.000	0.780	0.023
Unemployment NUM4	-0.883	-0.000	-0.000	-0.000	-0.000	0.177	-0.328	0.780	1.000	0.021
IsHoliday NUM5	-0.033	-0.001	-0.000	-0.001	-0.000	0.147	0.016	0.023	0.021	1.000

The parameter estimates indicate high values for the independent variables (>0.001) and hence indicate that the model is not good to generate sales forecasts.

VI) Model and forecast outputs:

Model for variable mean	
Estimated Intercept	24749.45

Autoregressive Factors	
Factor 1:	1 + 0.24912 B**(1) - 0.24534 B**(2)

Moving Average Factors	
Factor 1:	1 + 0.24797 B**(1) - 0.22971 B**(2)

Input Number 1	
Input Variable	Temperature
Overall Regression Factor	14.75598

Input Number 2	
Input Variable	Fuel_Price
Overall Regression Factor	-59.6614

Input Number 3	
Input Variable	CPI
Overall Regression Factor	-40.9649

Input Number 4	
Input Variable	Unemployment
Overall Regression Factor	4.370944

Input Number 5	
Input Variable	IsHoliday
Overall Regression Factor	-414.239

Note: Further warnings will not be printed.

educed to 12.

Forecasts for variable mean						
Obs	Forecast	Std Error	95% Confidence Limits		Actual	Residual
9269	16180.7853	17566.198	-18248.3293	50609.8998	9196.5579	-6984.2273
9270	16352.8705	17566.209	-18076.2669	50782.0079	6187.3897	-10165.4808
9271	16394.2516	17568.434	-18039.2457	50827.7489	1374.6497	-15019.6020
9272	16426.1619	17568.592	-18007.6460	50859.9697	21623.1847	5197.0228
9273	16428.3647	17568.808	-18005.8670	50862.5964	7808.4506	-8619.9142
9274	16435.6448	17568.854	-17998.6763	50869.9658	20727.8609	4292.2161
9275	16434.3716	17568.882	-18000.0043	50868.7475	11652.8274	-4781.5442
9276	16436.4748	17568.891	-17997.9184	50870.8681	63180.5682	46744.0933
9277	16435.6385	17568.895	-17998.7629	50870.0399	17139.3146	703.6761
9278	16436.3629	17568.897	-17998.0415	50870.7673	4091.5715	-12344.7913
9279	16435.9772	17568.897	-17998.4285	50870.3829	8052.0762	-8383.9011
9280	16436.2510	17568.897	-17998.1552	50870.6572	7436.0173	-9000.2337

The forecast is shown for 12 months as that is the hold out period and residual is calculated only for that duration.

5. Model comparison:

5.1 Accuracy & fit statistics with and without hold out sample:

Considering the models created above, the selection of best model was done based on the Accuracy and fit statistics. The measures for all the 8 models created using a **HOLDOUT** sample of 12 months as below:

Models	SSE	n	p	MSE	MAPE	AIC	SBC
ARIMAX(2,0)	3790941127794	9280	8	408859052	2.22%	210334.4	210391.5
ARIMAX(1,0)	4244368791590	9280	7	457712584	4.02%	211380.5	211430.4
ARIMAX(2,2)	2883675746369	9280	10	311076132	4.14%	207808.5	207879.8
ARIMAX(0,2)	2861874675199	9280	8	308657752	4.30%	207734.3	207791.4
ARIMAX(0,0)	2861335610903	9280	6	308533061	4.58%	207740.8	207783.6
ARIMAX(1,1)	2862478387514	9280	8	308722863	4.69%	207736.2	207793.3
ARIMA(0,2)	2862062283406	9280	3	308511618	5.00%	207724.9	207746.3
ARIMAX(0,1)	2862621818212	9280	7	308705038	5.23%	207734.7	207784.7

Here, the model with best accuracy was ARIMAX(2,0). For the same, the MAPE stands at 2.22% which far better in comparison to the rest of the models. Hence, fit statistics is not needed to select the best model based on parsimony.

For this best model, below is the accuracy and fit statistics **without the HOLDOUT sample**:

Models	SSE	n	p	MSE	MAPE	AIC	SBC
ARIMAX(2,0)	2860535400000	9280	8	308513308.9	2.03%	207742.2	207799.3

5.2 Final Forecasting Parameters & Equation:

Models	Model Equation:
ARIMAX(2,0)	Y(t) + 0.00178Y(t-1) – 0.0162Y(t-2) = 24778.8 + 14.75T(t) – 59.1F(t) – 41.07C(t) + 3.4U(t) – 415H(t)
ARIMAX(1,0)	Y(t) + 0.0018139 Y(t-1) = 24775.6 + 14.78 T(t) – 59.84 F(t) – 41.05 C(t) + 3.48 U(t) – 417.65 H(t)
ARIMAX(2,2)	Y(t) + 0.2479 Y(t-1) – 0.2297 Y(t-2) = 24749 + 0.249 E(t-1) – 0.245 E(t-2) + 14.75 T(t) – 59.66 F(t) – 40.96 C(t) + 4.37 U(t) – 414.23 H(t)
ARIMAX(0,2)	Y(t) = 24780.3 - 0.0018215 E(t-1) +0.01653 E(t-2) + 14.7 T(t) - 59.15 F(t) - 41.08 C(t) + 3.44082 U(t) - 415.46 H(t)
ARIMAX(0,0)	Y(t) = 24775.7 + 14.7799 T(t) -59.84902 F(t) -41.0582 C(t)+3.47377 U(t)-417.73928 H(t)
ARIMAX(1,1)	Y(t) = 24769.5 -0.96893Y(t-1) + 0.96438 E(t-1) + 14.78 T(t) -59.472 F(t) -41.05 C(t) + 3.86 U(t) -414.13H(t)
ARIMA(0,2)	Y(t) = 16737.0 - 0.0015602 E(t-1) + 0.01682 E(t-2)
ARIMAX(0,1)	Y(t) = 24775.6 - 0.0017395 E(t-1) + 14.78097 T(t) - 59.84472 F(t) - 41.05870 C(t) + 3.48384 U(t) - 417.65779 H(t)

6. Conclusion:

In conclusion, the best model was the ARIMAX(2,0) model. It has a mean absolute percentage error of 2.22%, an AIC of 210334.4, and an SBC of 210391.5. The best model had the best MAPE of all the other models therefore the AIC and SBC were not considered, as these are based on fit statistics or training data. Moreover, the forecast shows a drop in sales in the forward time horizon following the trend depicted in fit data.

Finally, the business recommendations that yield from this analysis are as follows:

- Firstly, given that the macroeconomic indicators show a positive trend, the forecast shows a decline in sales and hence factors leading to the drop should be investigated.
- Secondly, department level sales trends can be understood to find critical focus areas.
- Finally, other indicators like quality, delivery of service, price points can be evaluated to understand if there are other factors influencing the sales.