Self Project On

Gurugram's Air Quality Index

Time Series Analysis

- **Conducted time series analysis on daily AQI data to identify seasonal trends.**
- **Filled missing values using linear interpolation for a complete dataset.**
- **Applied log and uniform transformations to stabilize variance and manage data range.**
- **Performed stationarity tests using differencing to remove unit roots.**
- **Identified ARIMA & SARIMA model parameters using ACF and PACF plots.**
- **Evaluated model performance based on AIC, BIC, and MSE to select optimal models.**
- **Validated ARIMA(5,1,0) and SARIMA models for accurate AQI forecasting.**

1- Data Description

This dataset provides a comprehensive record of Gurugram's daily Air Quality Index (AQI), spanning from April 2023 to March 2024. It comprises three key columns: Date and AQI.

The objective is to analyse temporal patterns and trends in Gurugram's daily Air Quality Index (AQI) data from April 2023 to March 2024. By employing time series analysis and forecasting techniques, the aim is to uncover insights for environmental management and public health interventions, facilitating informed decision-making to mitigate air pollution and enhance air quality in Gurugram.

No. of rows: 397
No. of columns: 2

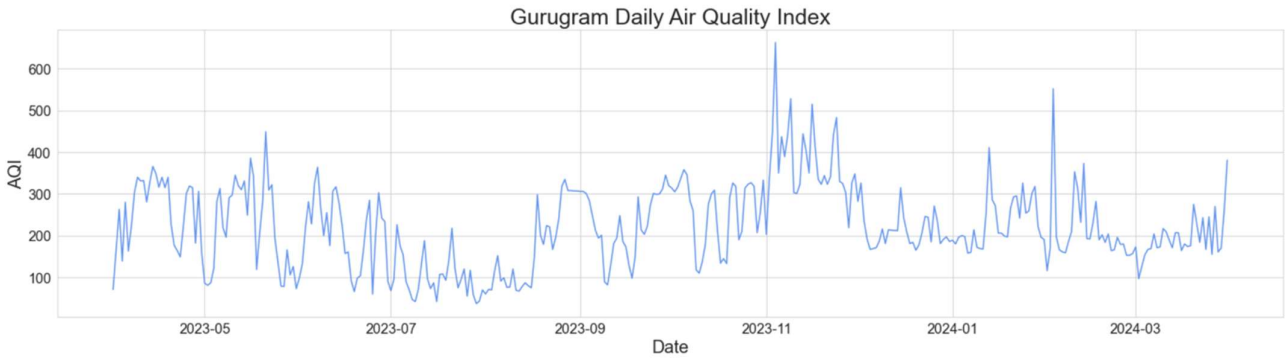
The Data looks as follows:

date	value
2023-04-01	71.0
2023-04-02	168.0
2023-04-03	263.0
2023-04-04	139.0
2023-04-05	280.0
2023-04-06	163.0
2023-04-07	224.0
2023-04-08	304.0
2023-04-09	340.0
2023-04-10	331.0

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 366 entries, 2023-04-01 to 2024-03-31
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   value   364 non-null    float64
dtypes: float64(1)
```

There were two null values in the dataset which has been filled with the linear interpolation

2- Data Visualization and Interpretations

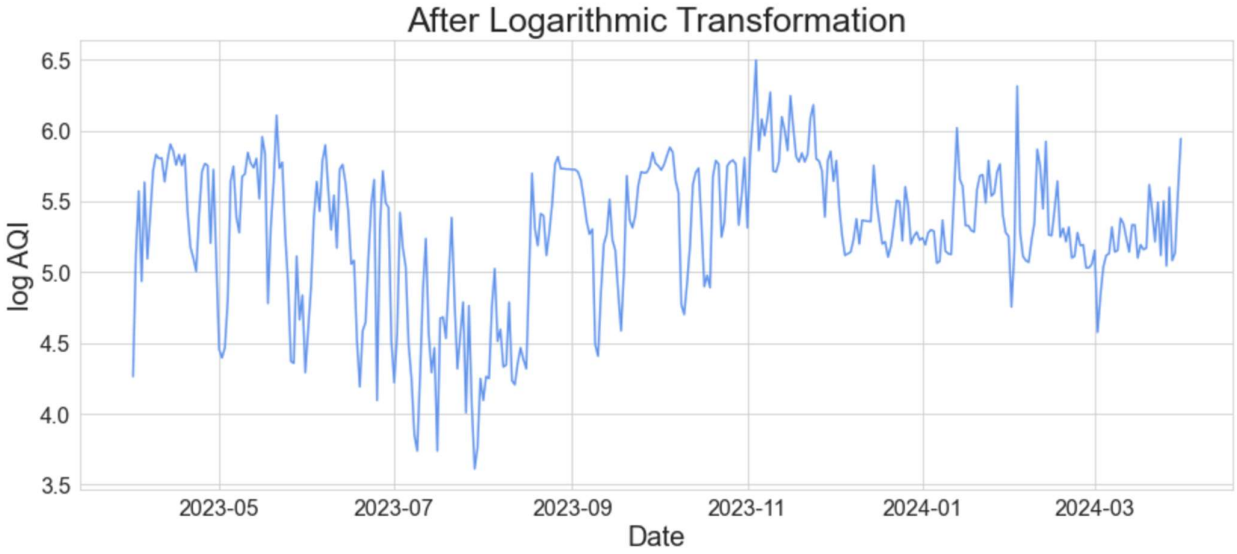


To transform the discrete data into a continuous format, two options are available:

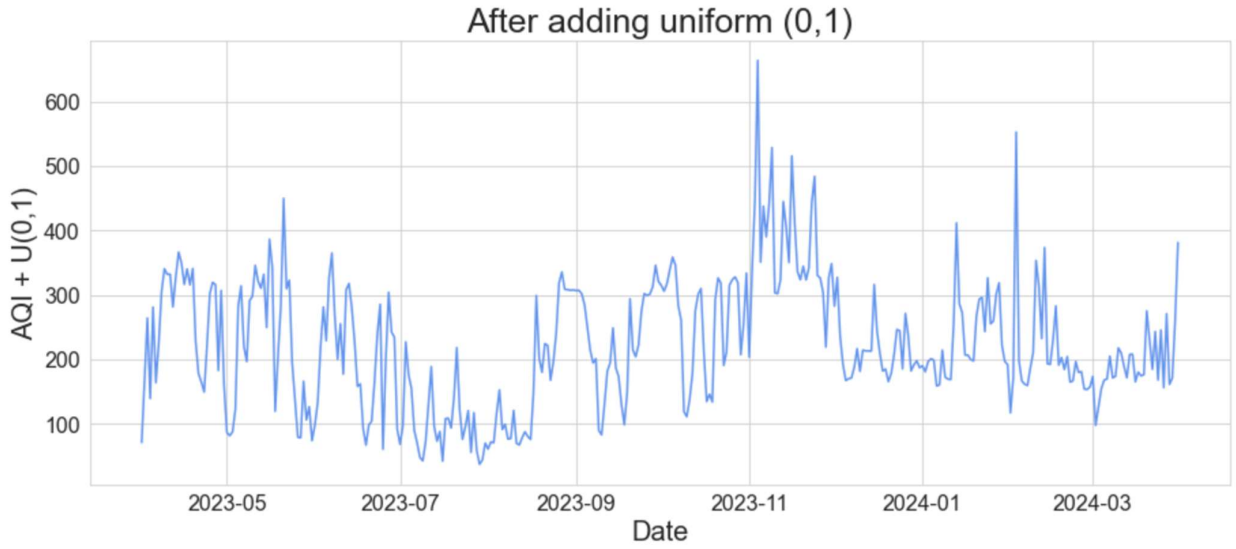
1) **Log transformation:** This method not only converts the data into a continuous form but also scales it to a more manageable range.

2) **Adding a number between 0 and 1 from a uniform distribution:** This approach introduces randomness to the data while maintaining continuity, thereby ensuring a smooth transition from discrete to continuous representation.

Log AQI Series plot:



AQI + U(0,1) Series plot:



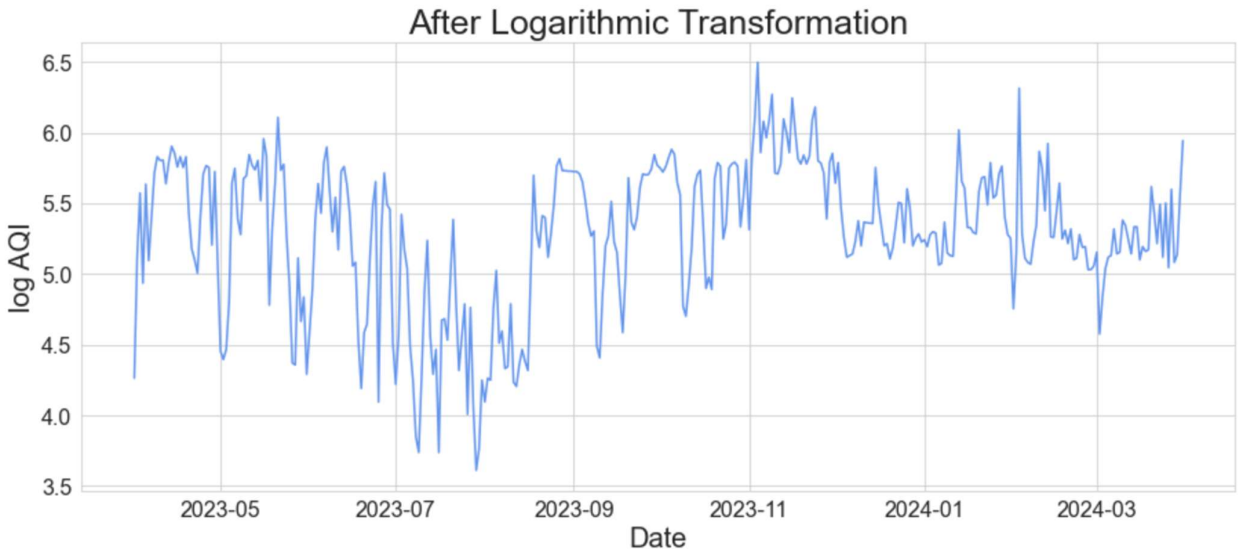
The log transformation has effectively compressed the range of the series, resulting in a more manageable and scaled representation of the data. The uniform transformed series appears to maintain its original characteristics, showing no discernible change.

The series plots reveal a dynamic pattern where the mean and variance exhibit fluctuations over time, indicating a departure from stationarity, which will be thoroughly examined in the next section.

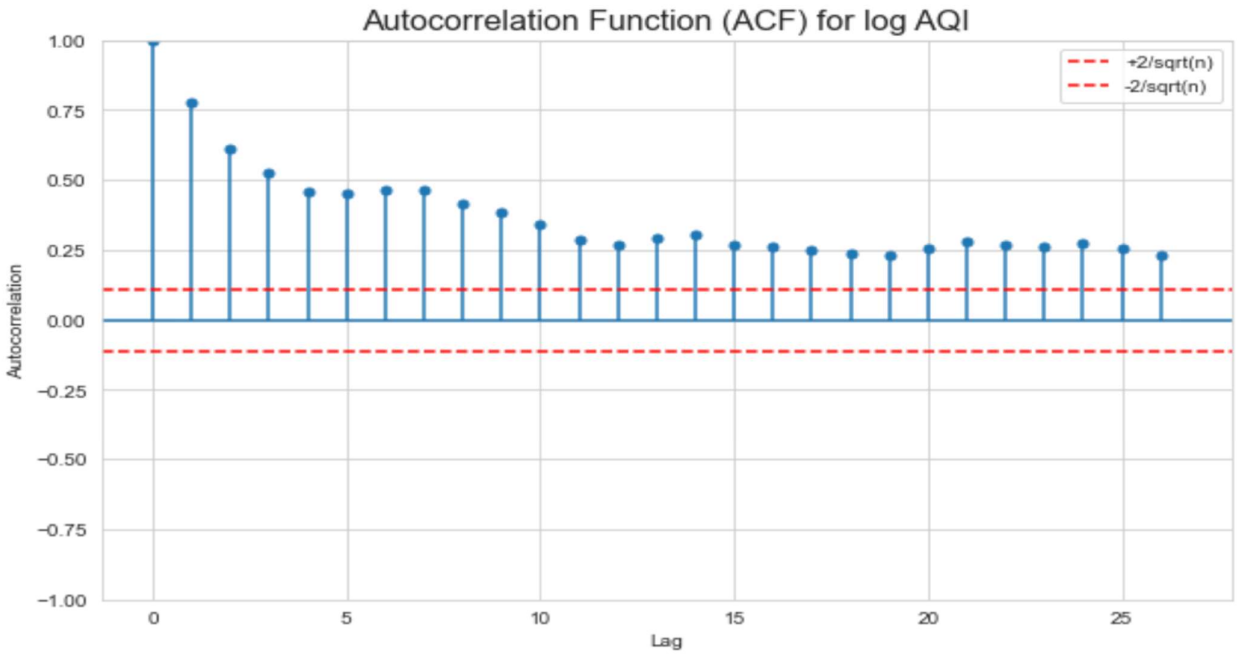
To validate a time series model, the dataset is typically divided into two parts: one for training and another for testing. In this context, the Air Quality Index (AQI) data for the last 42 days is used for the testing set.

3- Examining Stationarity

3.1- For Log transformed series

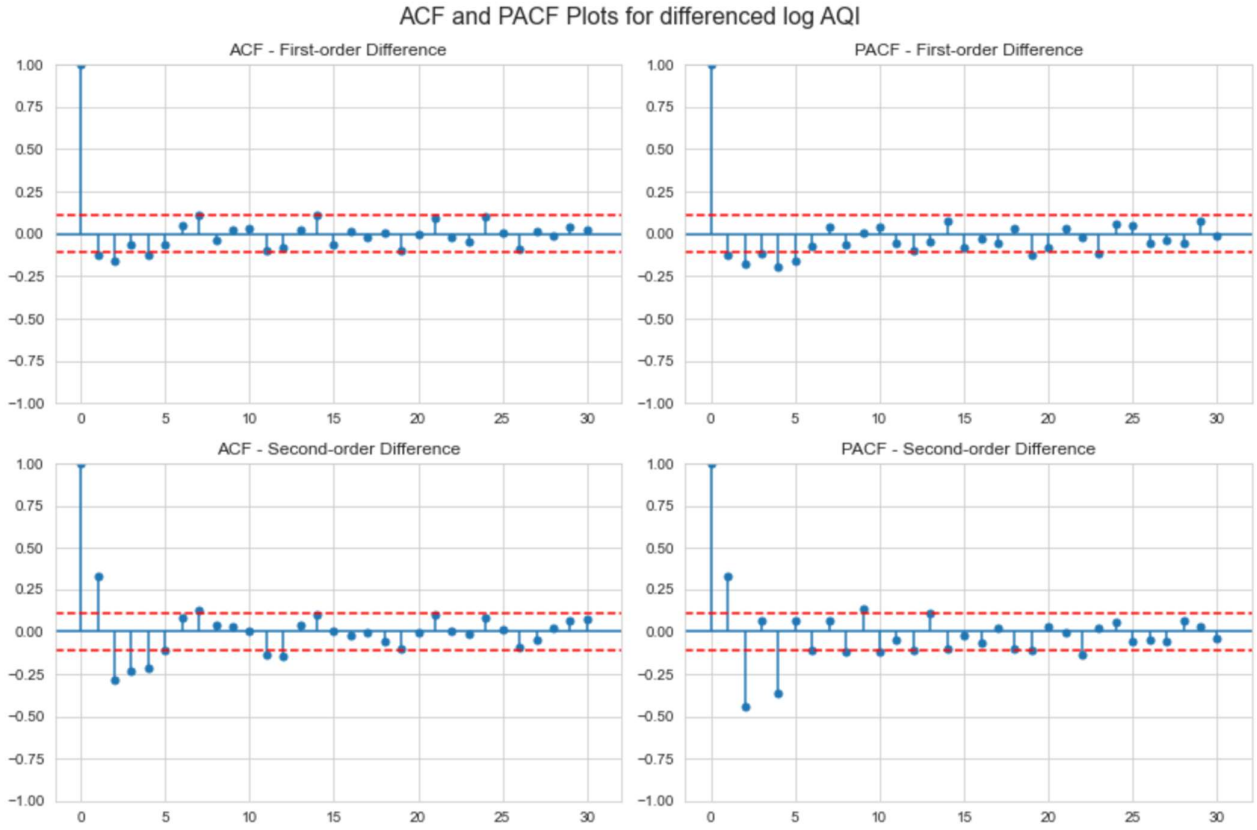


ACF Plot –



- The autocorrelation function (ACF) displays a slow decay, indicating the presence of a unit root and thus non-stationarity.
- Additionally, prominent spikes are evident at lags 7, 14, and 21, implying the existence of seasonality with a weekly period of 7 days. Addressing this seasonality will be a key focus during the identification of SARIMA models.

Differencing to achieve stationarity

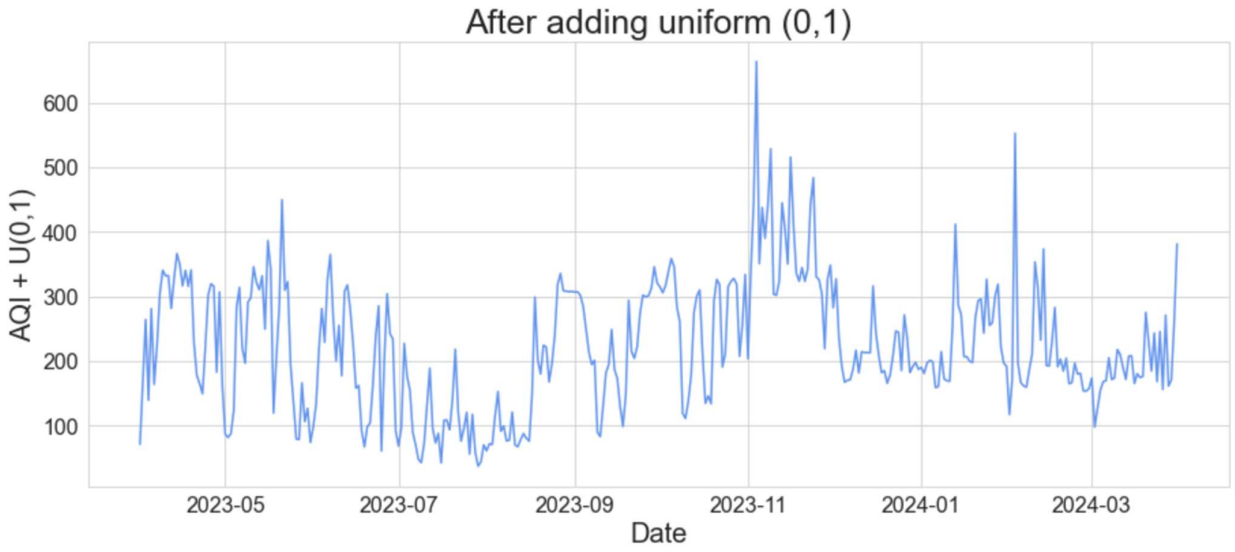


d = 1 : The autocorrelations for first order differenced time series are all small, demonstrating that stationarity has now been achieved.

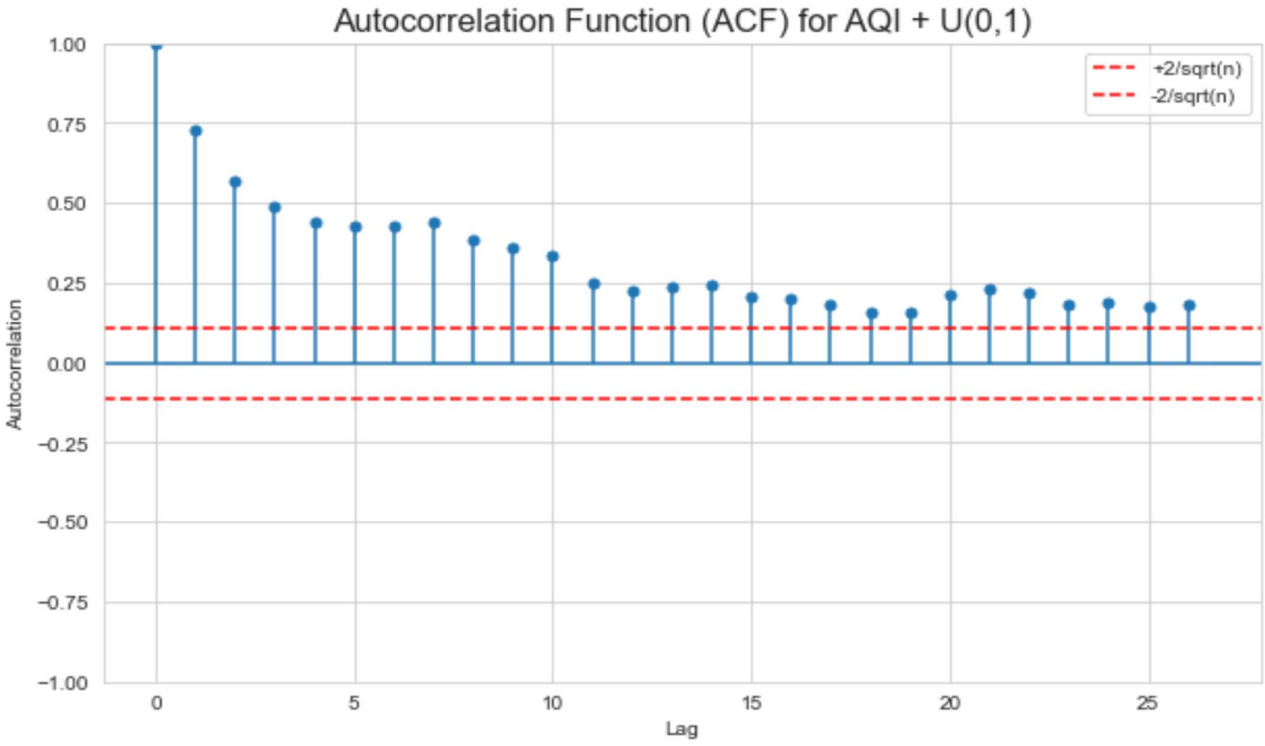
d = 2 : The autocorrelations for second order differenced time series has spikes at 2, 3 and 4, showing extra correlation that has emerged because of overdifferencing.

So, we can choose d = 1 for log AQI.

3.2- For uniform transformed series

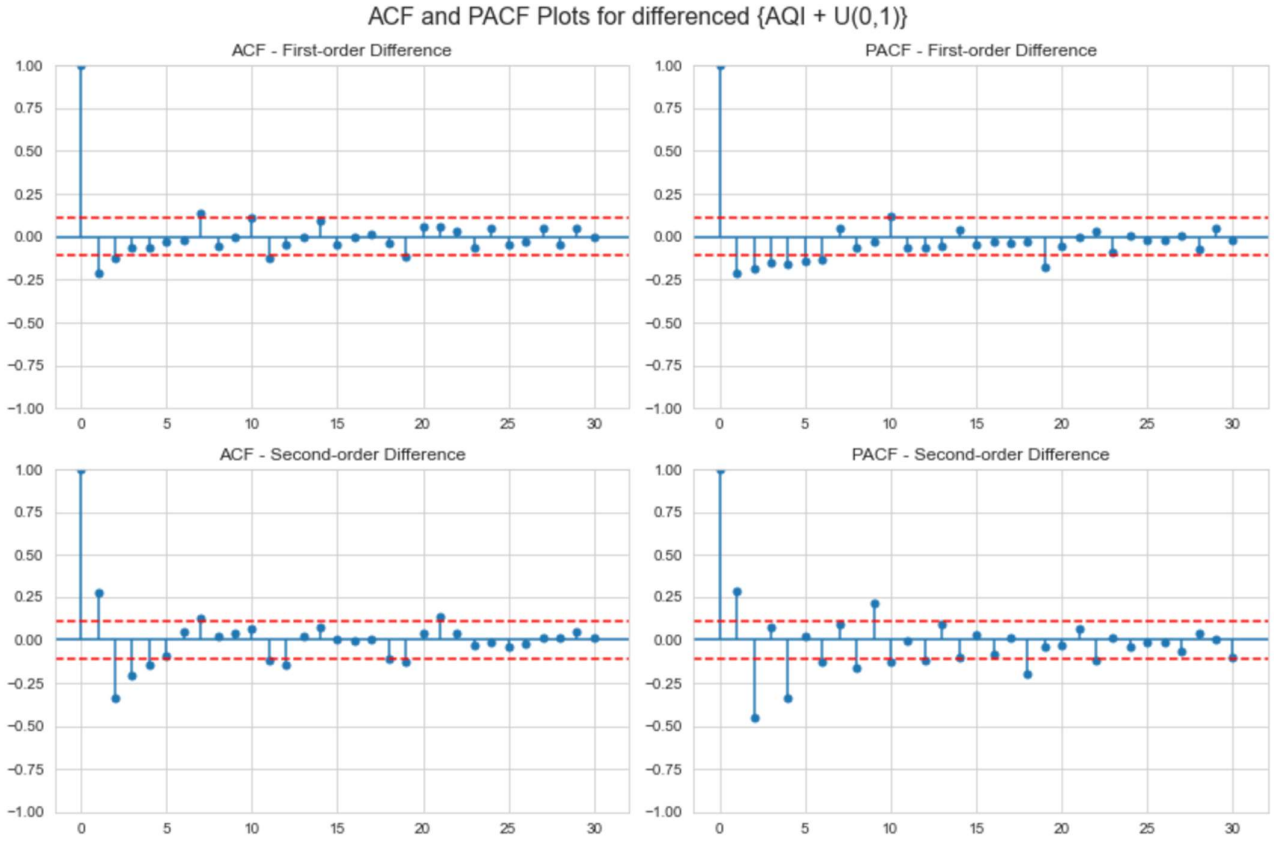


ACF Plot –



- The ACF in this case is very similar to that of log AQI, displaying a slow decay, indicating the presence of a unit root and thus non-stationarity.
- Spikes are also observed at lag 7, 14 and 21, indicating the presence of seasonality which will be examined later.

Differencing to achieve stationarity

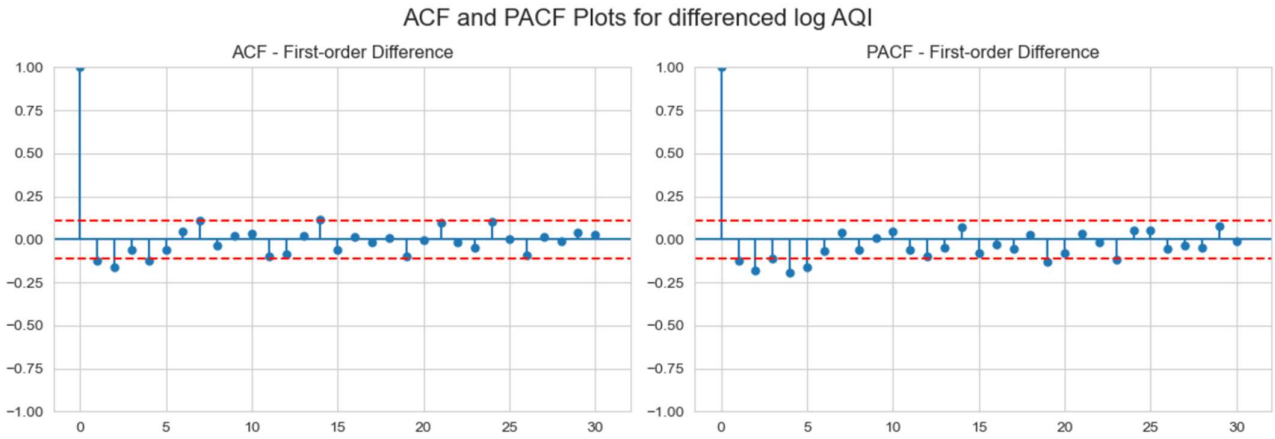


In this case of we can observe overdifferencing at d = 2. So, we will take **d = 1** in this case also.

4- Model Identification

4.1- For Log transformed series

4.1.1- ARIMA Models



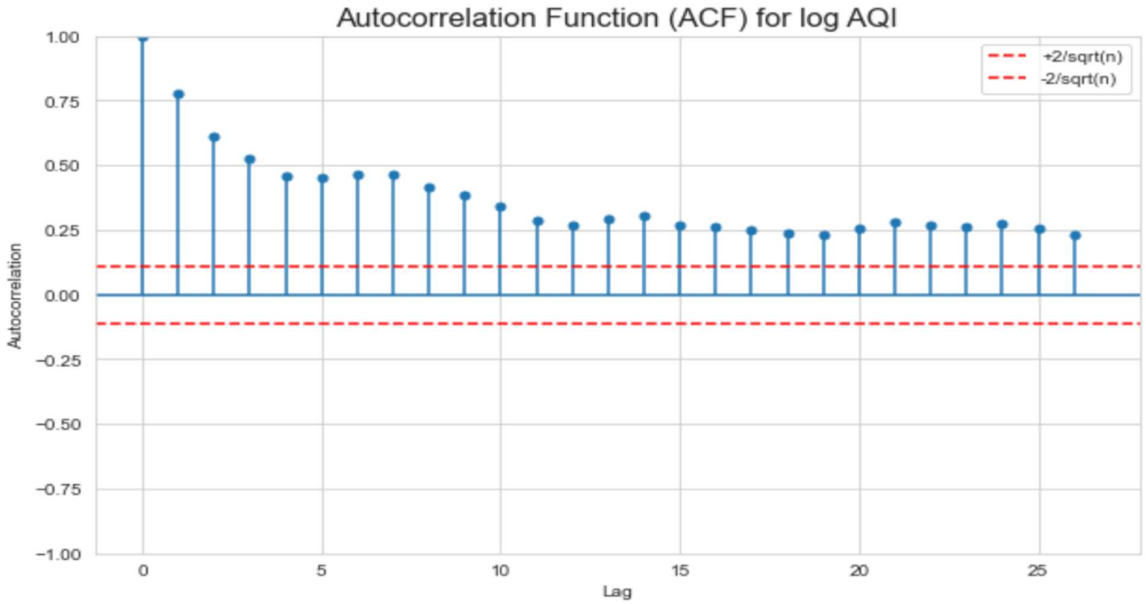
In the log transformed series we can see that

- ACF cuts of after lag 2 and PACF shows damped sine waves - ARIMA (0,1,1) and ARIMA (0,1,2)
- PACF cuts of after lag 5 and ACF shows damped sine waves - ARIMA(5,1,0).
- ACF tails of after lag 2 and PACF cuts of after lag 5 - ARIMA(5,1,2)

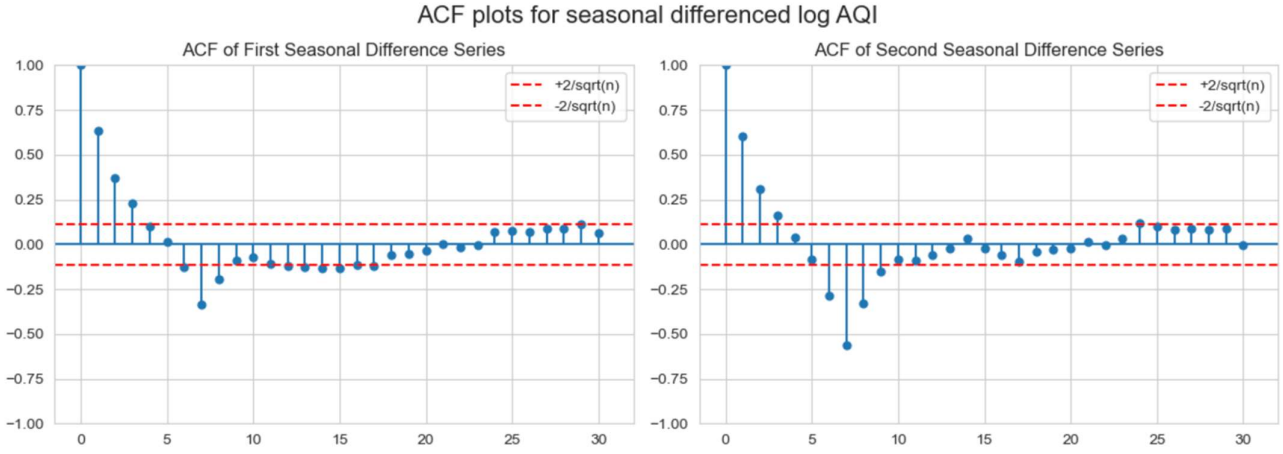
```
AIC and BIC for models identified for log transformed series:
          Model         AIC         BIC
2  ARIMA(5,1,0)  220.178937  242.844851
3  ARIMA(5,1,2)  220.405670  246.849237
1  ARIMA(0,1,2)  223.009486  234.342443
0  ARIMA(0,1,1)  244.951086  252.506391
```

Based on AIC and BIC, ARIMA(5,1,0) and ARIMA(0,1,2) is identified as a plausible model for log transformed series.

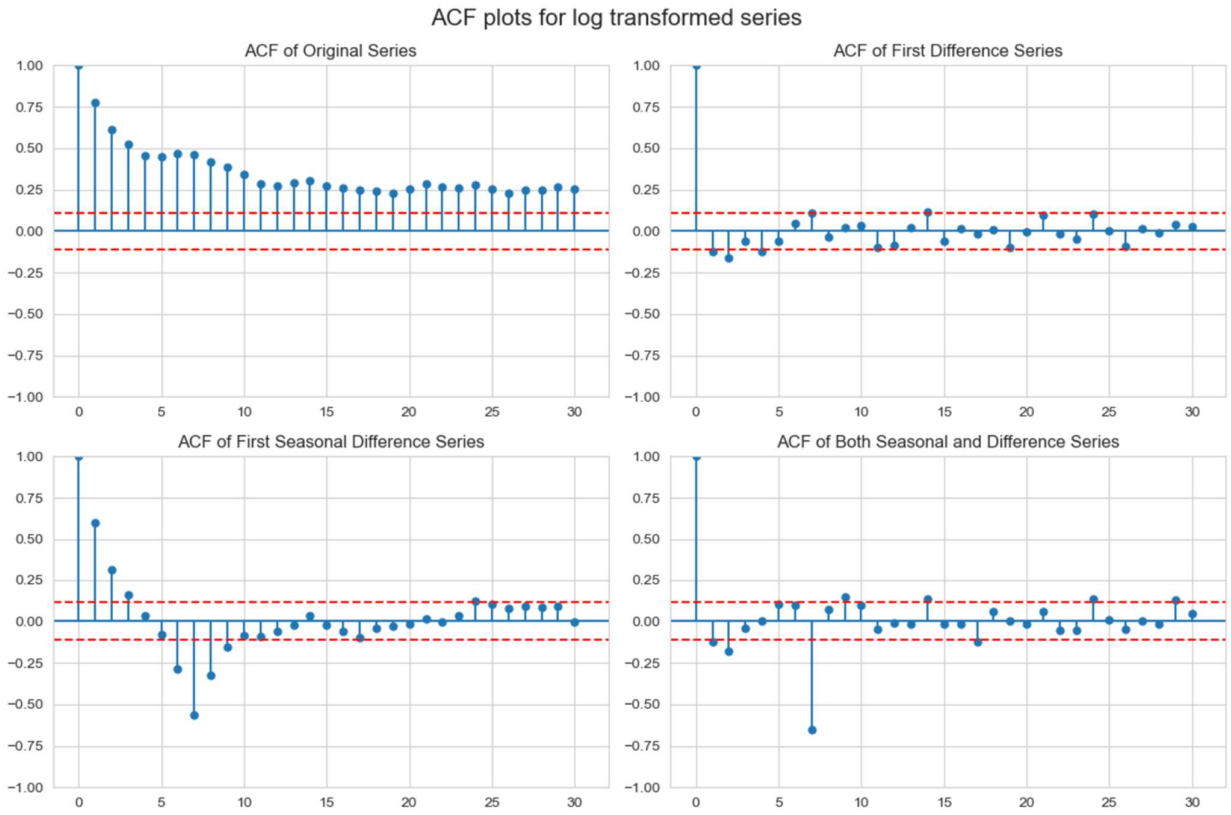
4.1.2- SARIMA Models



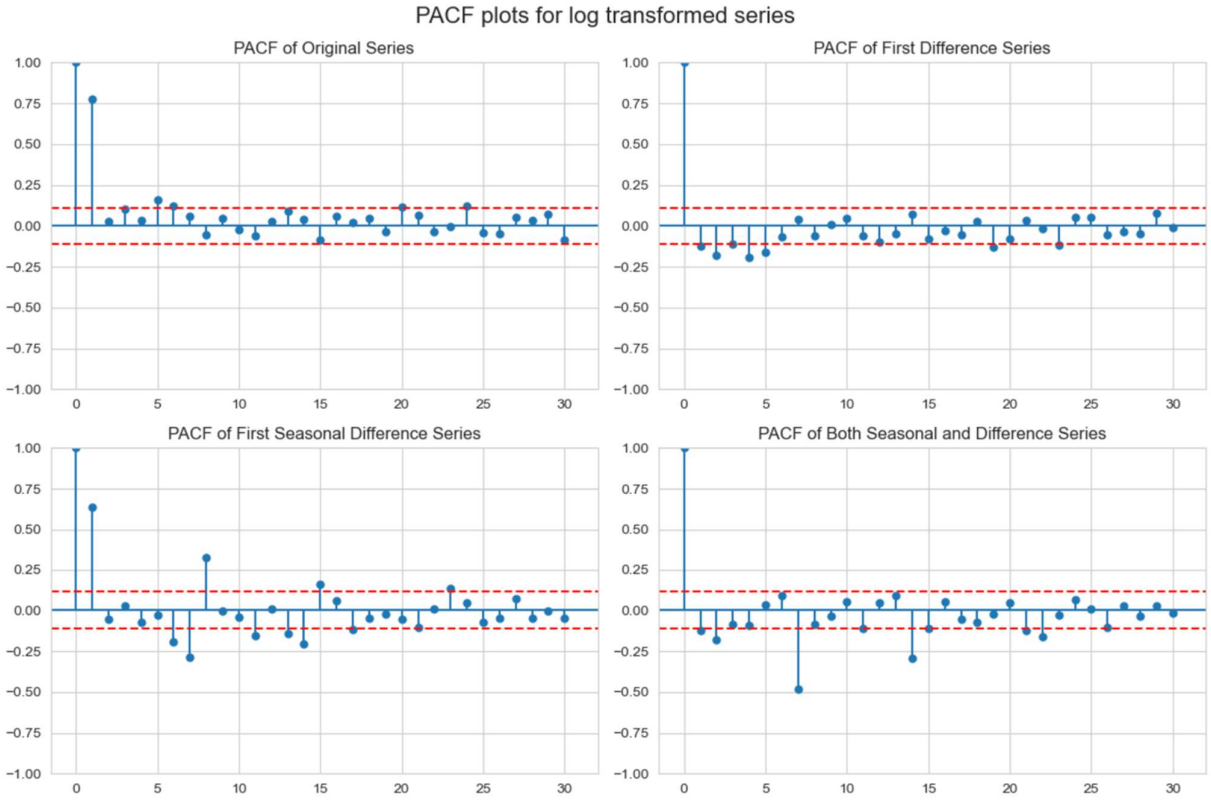
As previously noted, significant spikes at lags 7, 14, and 21 suggest a weekly seasonal pattern with a periodicity of 7 days. To address this and achieve stationarity, we will apply seasonal differencing with a periodicity of 7 days.



We observed an increase in spikes with second-order seasonal differencing, indicating that a first-order seasonal difference (D = 1) is more appropriate. However, the autocorrelations remained strong at lags 6, 7, and 8. To address this, we will apply normal differencing in addition to seasonal differencing to further reduce these autocorrelations.



We can see that seasonal differencing along with normal differencing has markedly reduced the autocorrelations throughout.



Based on the Partial Autocorrelation Function (PACF), the autocorrelation diminishes after lag 2, with noticeable spikes occurring at lags 7, 14, and 21. This suggests that the order of the autoregressive component `**p**` might be between 0 and 2, while the seasonal autoregressive component `**P**` could be in the range of 0 to 3, accounting for the observed periodicity.

The Autocorrelation Function (ACF) decreases significantly after lag 2, with a notable spike at lag 14. This observation suggests that the possible values for the moving average component **'q'** could range from 0 to 2, while the seasonal moving average component **'Q'** might be either 0 or 1.

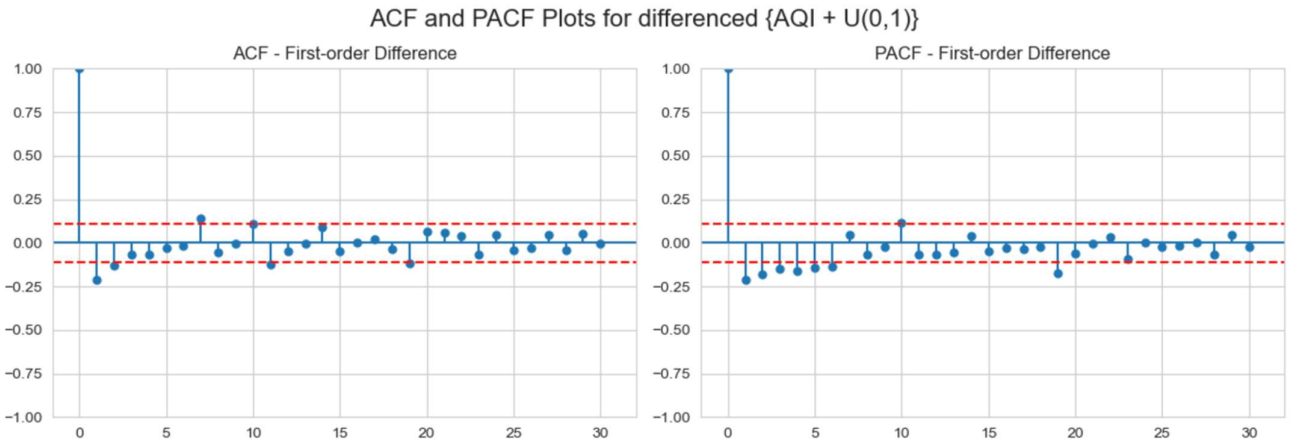
Considering these potential values for **p, q, P** and **Q**, we calculated the **AIC and BIC** for a variety of SARIMA models. The top five models with the best AIC and BIC scores are listed below:

	Model Order	AIC	BIC
37	SARIMA(1, 1, 1)(2, 1, 1, 7)	236.705872	259.240326
35	SARIMA(1, 1, 1)(1, 1, 1, 7)	236.846164	255.624875
47	SARIMA(1, 1, 2)(3, 1, 1, 7)	237.829918	267.875856
43	SARIMA(1, 1, 2)(1, 1, 1, 7)	237.884891	260.419344
59	SARIMA(2, 1, 1)(1, 1, 1, 7)	237.970894	260.505348

Based on AIC and BIC, plausible models are: SARIMA(1,1,1)(2,1,1) and SARIMA(1,1,1)(0,1,1). These model will diagnosed in the next chapter.

4.2- For uniform transformed series

4.2.1- ARIMA Models



In the uniform transformed series we can see that

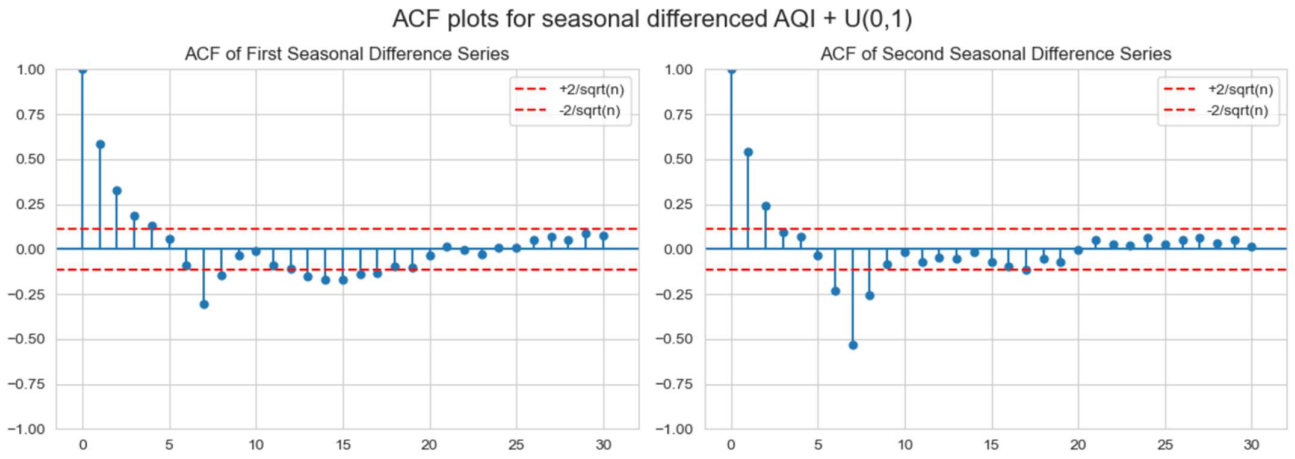
- ACF cuts of after lag 2 and PACF can we considered to tail off exponentially - ARIMA (0,1,1) and ARIMA (0,1,2)
- PACF cuts of after lag 5 or 6 and ACF is a mixture of exponential and damped sine waves. - ARIMA(5,1,0) and ARIMA (6,1,0).
- ACF decays exponentially from lag 1 and PACF also decays after lag 1 - ARIMA(1,1,1)

AIC and BIC for models identified for uniform transformed series:

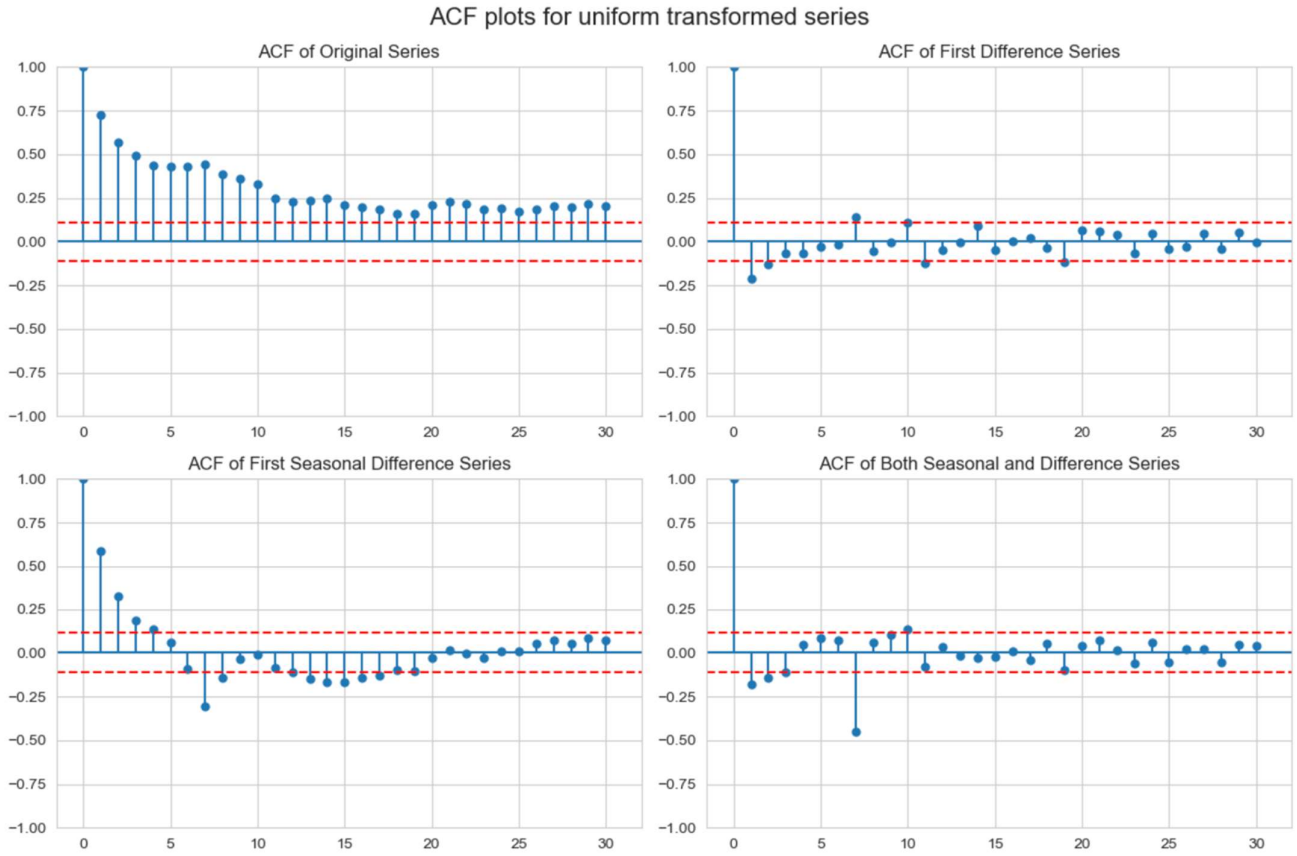
	Model	AIC	BIC
4	ARIMA(1,1,1)	3665.104228	3676.437185
1	ARIMA(0,1,2)	3671.268398	3682.601355
3	ARIMA(6,1,0)	3671.395084	3697.838651
2	ARIMA(5,1,0)	3675.530792	3698.196706
0	ARIMA(0,1,1)	3690.124494	3697.679798

Based on AIC and BIC, ARIMA(1,1,1) is identified as a plausible model for uniform transformed series.

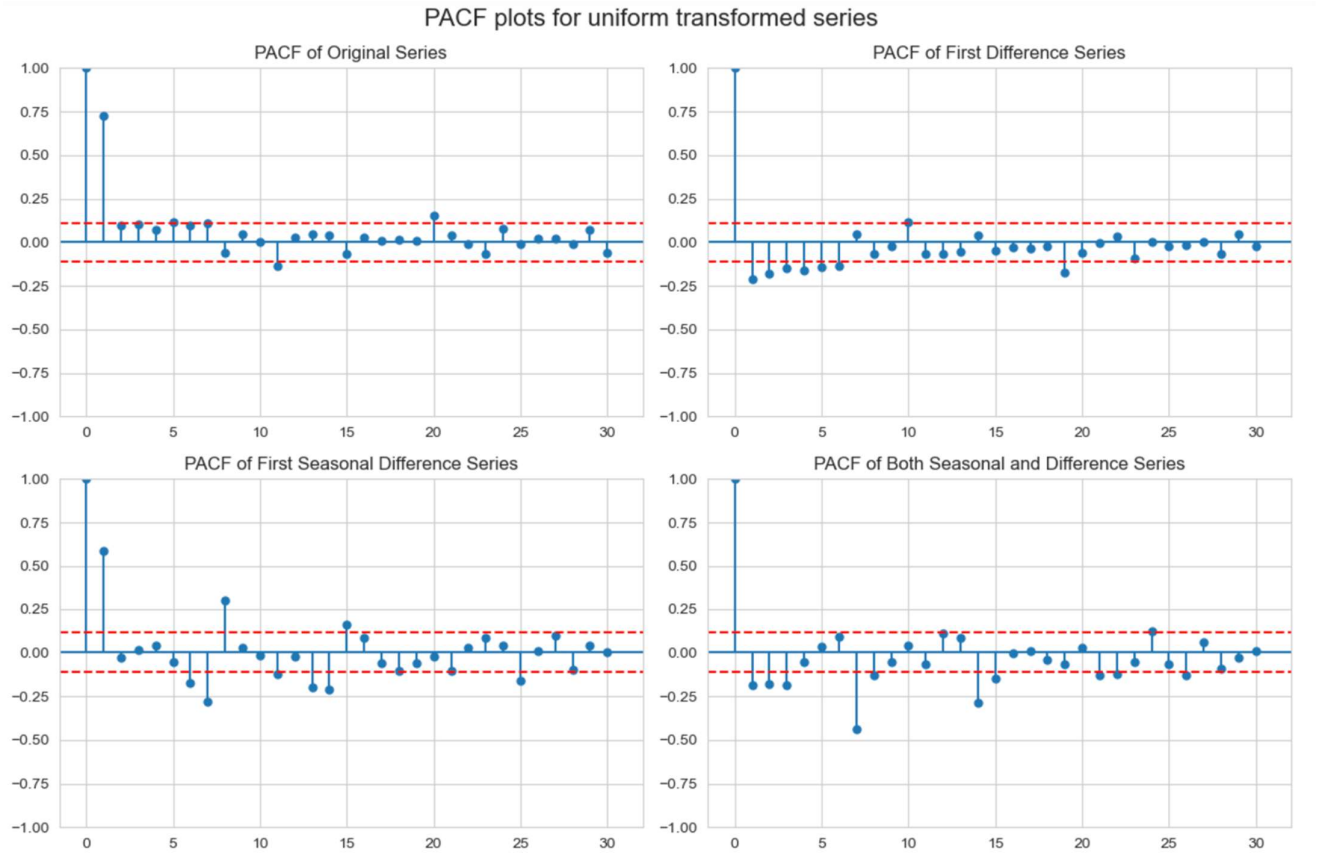
4.2.2- SARIMA Models



We observed an increase in spikes with second-order seasonal differencing, indicating that a first-order seasonal difference (D = 1) is more appropriate.



We can see that seasonal differencing along with normal differencing has markedly reduced the autocorrelations throughout.



Based on the Partial Autocorrelation Function (PACF), the autocorrelation diminishes after lag 3, with noticeable spikes occurring at lags 7 and 14. This suggests that the order of the autoregressive component `**p**` might be between 0 and 3, while the seasonal autoregressive component `**P**` could be in the range of 0 to 2, accounting for the observed periodicity.

The Autocorrelation Function (ACF) decreases significantly after lag 2, with a notable spike at lag 14. This observation suggests that the possible values for the moving average component **'q'** could range from 0 to 2, while the seasonal moving average component **'Q'** might be either 0 or 1.

Considering these potential values for **p, q, P** and **Q**, we calculated the **AIC and BIC** for a variety of SARIMA models. The top five models with the best AIC and BIC scores are listed below:

	Model Order	AIC	BIC
27	SARIMA(1, 1, 1)(1, 1, 1, 7)	3607.721138	3626.499849
25	SARIMA(1, 1, 1)(0, 1, 1, 7)	3609.061208	3624.084177
45	SARIMA(2, 1, 1)(1, 1, 1, 7)	3609.217828	3631.752281
29	SARIMA(1, 1, 1)(2, 1, 1, 7)	3609.336077	3631.870531
33	SARIMA(1, 1, 2)(1, 1, 1, 7)	3609.514650	3632.049104

Based on AIC and BIC, plausible models are: SARIMA(1,1,1)(1,1,1) and SARIMA(1,1,1)(0,1,1). These model will diagnosed in the next chapter.

5- Model Diagnostics

5.1- For Log transformed series

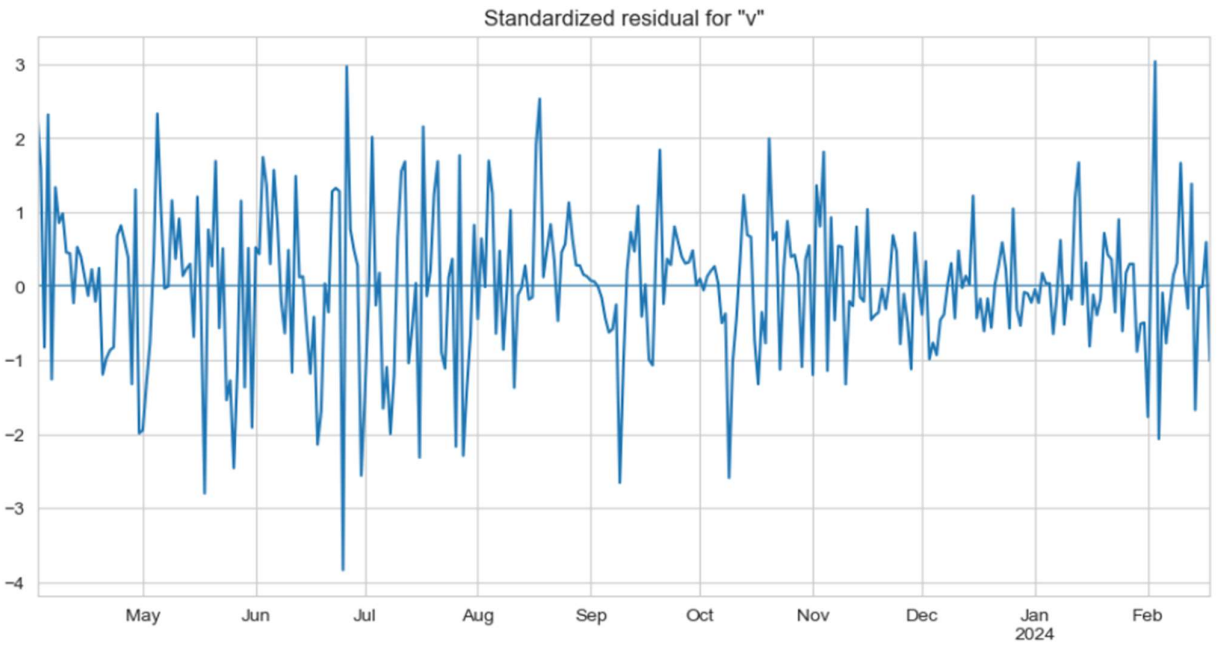
IMA(0, 1, 2)

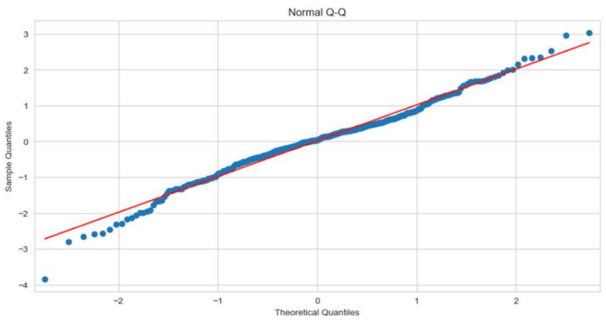
Dep. Variable:	value	**No. Observations:**	324
Model:	ARIMA(0, 1, 2)	**Log Likelihood**	-108.505
Date:	Fri, 19 Apr 2024	**AIC**	223.009
Time:	00:30:56	**BIC**	234.342
Sample:	04-01-2023	**HQIC**	227.533
	- 02-18-2024		
Covariance Type:	opg		

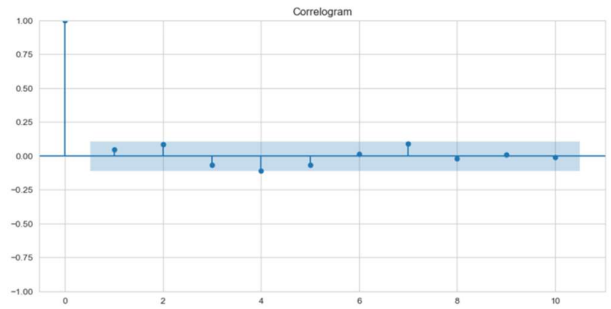
	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.2675	0.044	-6.042	0.000	-0.354	-0.181
ma.L2	-0.3186	0.048	-6.647	0.000	-0.413	-0.225
sigma2	0.1145	0.008	14.953	0.000	0.099	0.130

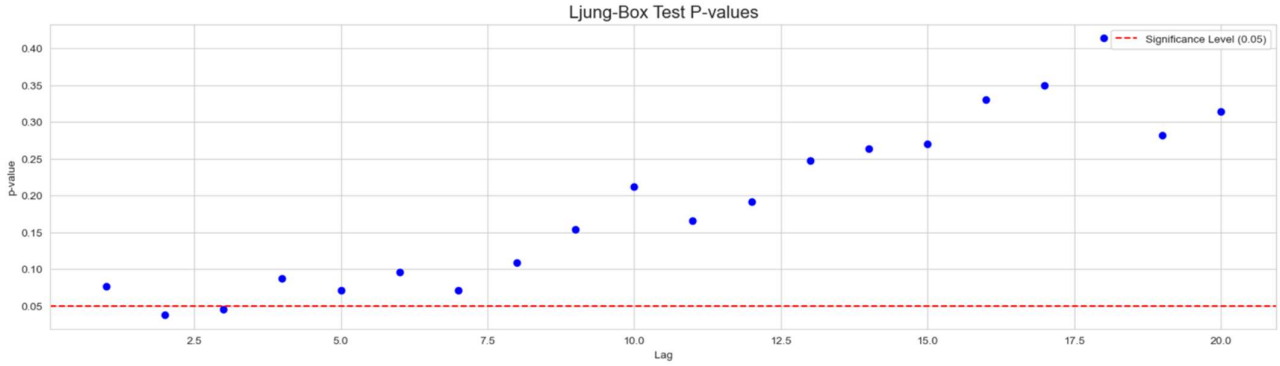
We can see that all the coefficients are coming out to be significant.

Residual Diagnostics









While most of the individual autocorrelations fall within the two-standard-error bounds, some values including $r2(\hat{a})$, $r4(\hat{a})$ and $r7(\hat{a})$, are close to these bounds. There is a suspicion of some lack of fit. This is confirmed by examining the p-values of the Ljung and Box test shown in the bottom graph. We note that some of the p-values at smaller lags are below or near the 5% level indicating some lack of fit.

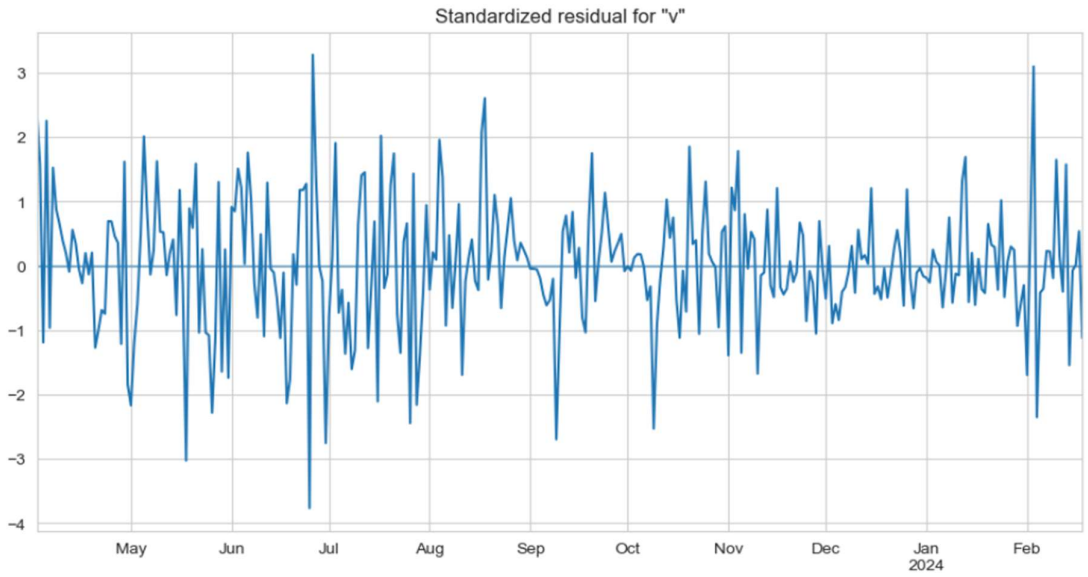
ARIMA(5, 1, 0)

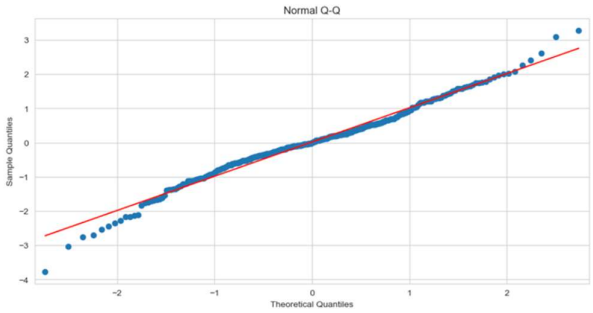
Dep. Variable:	value	**No. Observations:**	324
Model:	ARIMA(5, 1, 0)	**Log Likelihood**	-104.089
Date:	Fri, 19 Apr 2024	**AIC**	220.179
Time:	00:30:54	**BIC**	242.845
Sample:	04-01-2023	**HQIC**	229.227
	- 02-18-2024		
Covariance Type:	opg		

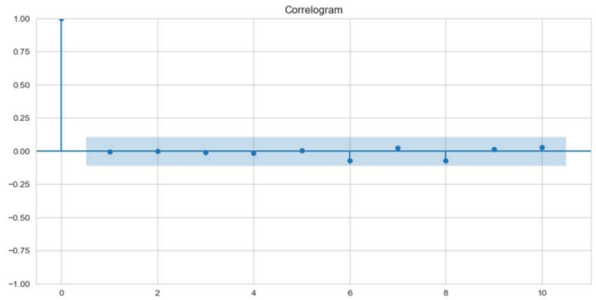
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2237	0.041	-5.434	0.000	-0.304	-0.143
ar.L2	-0.2631	0.056	-4.714	0.000	-0.373	-0.154
ar.L3	-0.1846	0.050	-3.705	0.000	-0.282	-0.087
ar.L4	-0.2280	0.051	-4.437	0.000	-0.329	-0.127
ar.L5	-0.1666	0.048	-3.464	0.001	-0.261	-0.072
sigma2	0.1114	0.007	15.323	0.000	0.097	0.126

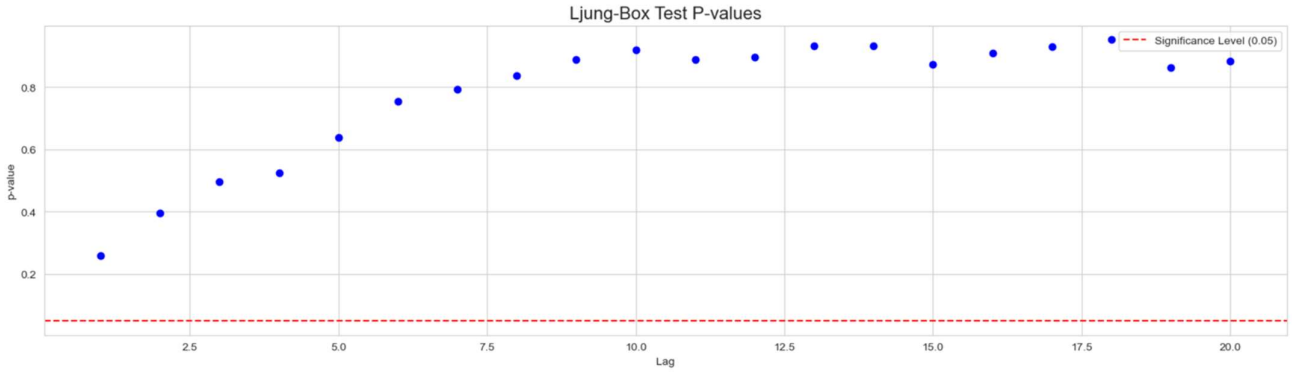
We can see that all the coefficients are coming out to be significant.

Residual Diagnostics:









The graph of the residual autocorrelation function shows no large values for this model. This is also reflected in the p-values of the Ljung and Box statistic shown at the bottom of the graph. These diagnostic checks show a clear improvement over the IMA(0, 1, 2) model examined above. The graph of the standardized residuals and the normal Q--Q plot reveal that outliers are present, however.

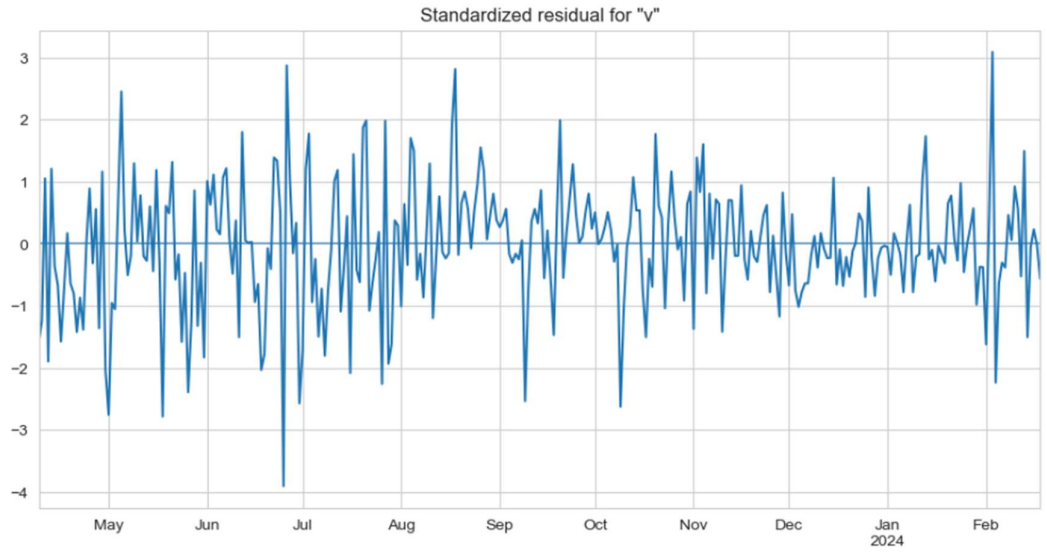
SARIMA(1, 1, 1)(2, 1, 1)

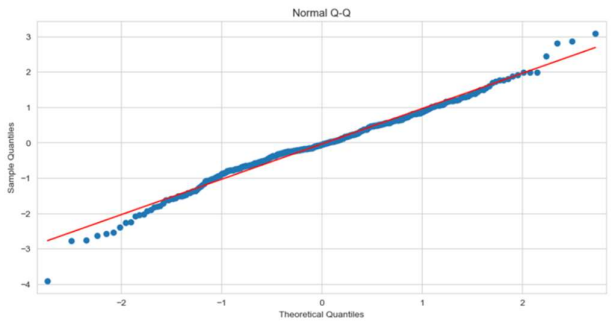
Dep. Variable:	value	**No. Observations:**	324
Model:	SARIMAX(1, 1, 1)x(2, 1, 1, 7)	**Log Likelihood**	-112.353
Date:	Fri, 19 Apr 2024	**AIC**	236.706
Time:	00:31:00	**BIC**	259.240
Sample:	04-01-2023	**HQIC**	245.708
	- 02-18-2024		
Covariance Type:	opg		

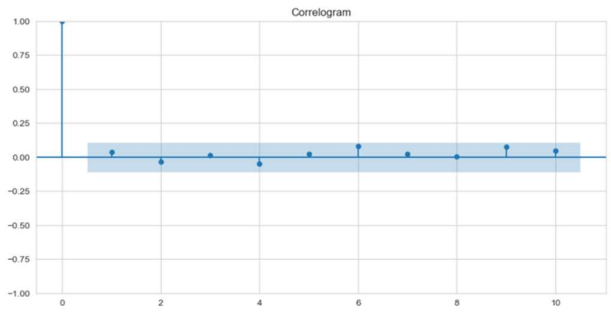
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6791	0.055	12.238	0.000	0.570	0.788
ma.L1	-0.9502	0.035	-26.890	0.000	-1.019	-0.881
ar.S.L7	0.1349	0.062	2.166	0.030	0.013	0.257
ar.S.L14	0.0924	0.065	1.432	0.152	-0.034	0.219
ma.S.L7	-0.9940	0.342	-2.910	0.004	-1.664	-0.324
sigma2	0.1103	0.036	3.089	0.002	0.040	0.180

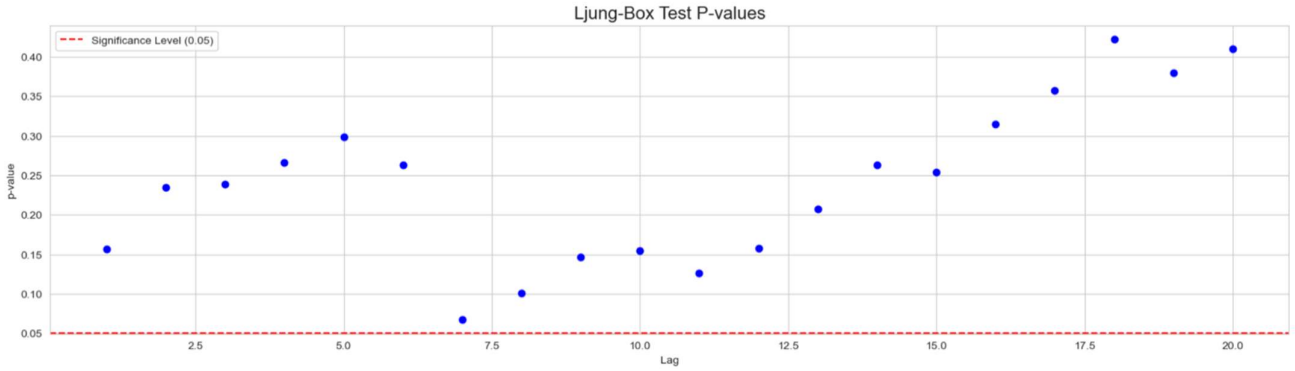
We can see that all the coefficients are coming not out to be significant. Model inadequacy can be suspected.

Residual Diagnostics:









The graph of the residual autocorrelation function shows no large values for this model. This is also reflected in the p-values of the Ljung and Box statistic shown at the bottom of the graph. P-value at lag 7 is very close to 5% level.

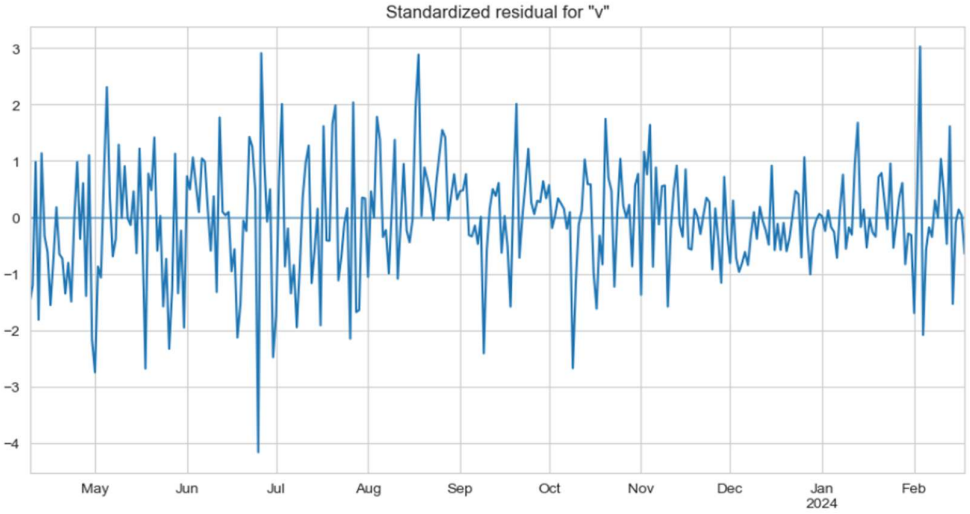
SARIMA(1, 1, 1)(0, 1, 1)

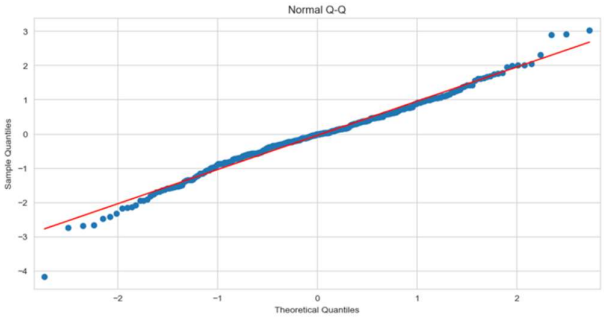
Dep. Variable:	value	**No. Observations:**	324
Model:	SARIMAX(1, 1, 1)x(0, 1, 1, 7)	**Log Likelihood**	-114.995
Date:	Fri, 19 Apr 2024	**AIC**	237.989
Time:	00:31:41	**BIC**	253.012
Sample:	04-01-2023	**HQIC**	243.991
	- 02-18-2024		
Covariance Type:	opg		

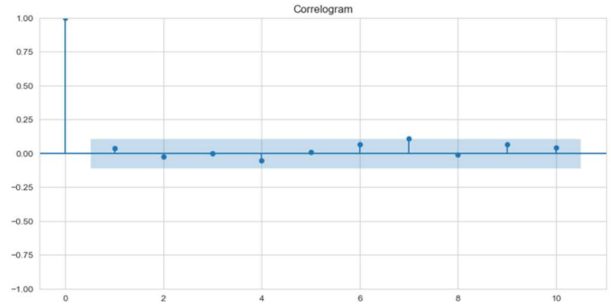
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6538	0.061	10.679	0.000	0.534	0.774
ma.L1	-0.9295	0.040	-23.096	0.000	-1.008	-0.851
ma.S.L7	-0.9075	0.040	-22.490	0.000	-0.987	-0.828
sigma2	0.1159	0.008	15.200	0.000	0.101	0.131

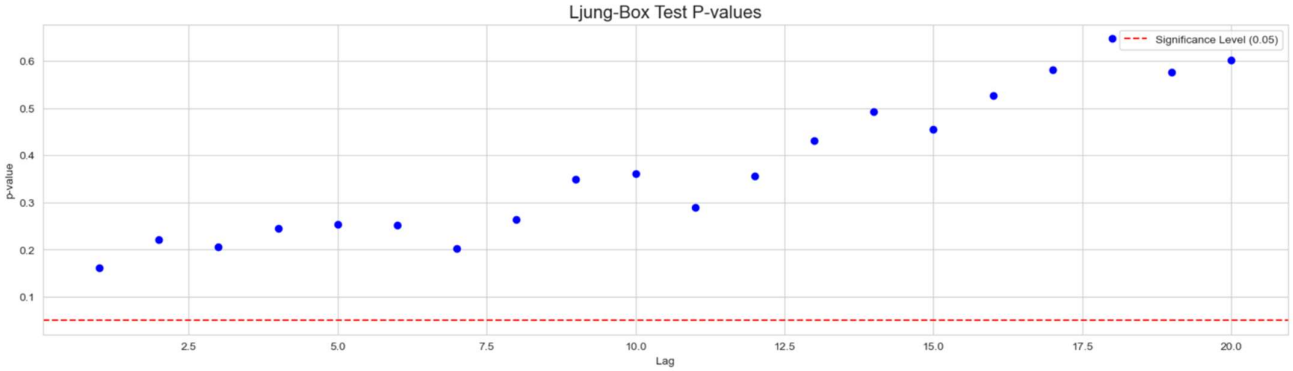
We can see that all the coefficients are coming out to be significant.

Residual Diagnostics:









The graph of the residual autocorrelation function shows no large values for this model. This is also reflected in the p-values of the Ljung and Box statistic shown at the bottom of the graph. These diagnostic checks show a clear improvement over the SARIMA(1, 1, 1)(2, 1, 1) model examined above. The graph of the standardized residuals and the normal Q--Q plot reveal that outliers are present, however.

5.2- For uniform transformed series

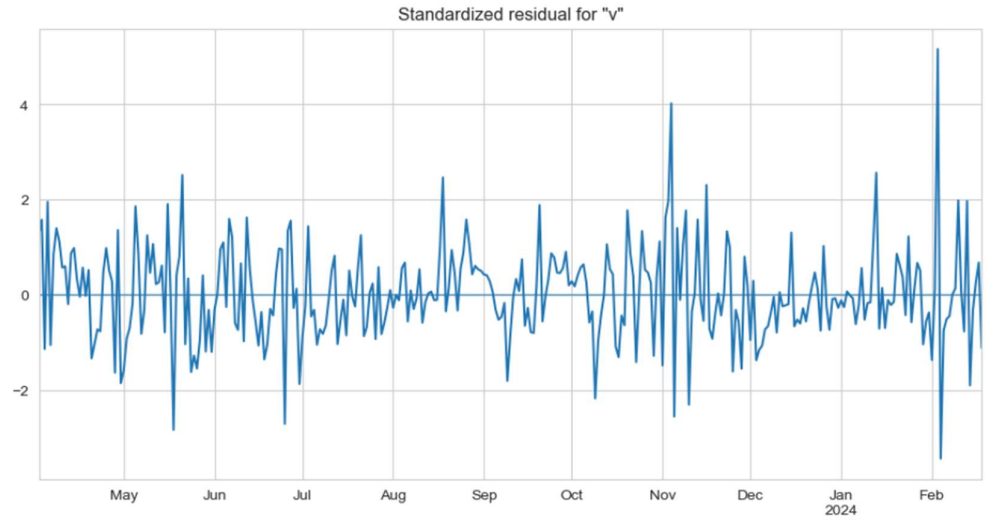
ARIMA(1, 1, 1)

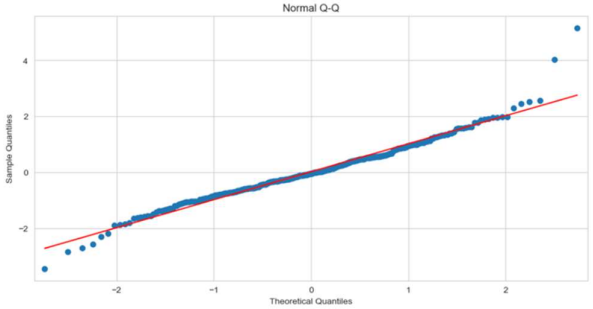
Dep. Variable:	value	**No. Observations:**	324
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-1829.552
Date:	Fri, 19 Apr 2024	**AIC**	3665.104
Time:	00:31:44	**BIC**	3676.437
Sample:	04-01-2023	**HQIC**	3669.628
	- 02-18-2024		
Covariance Type:	opg		

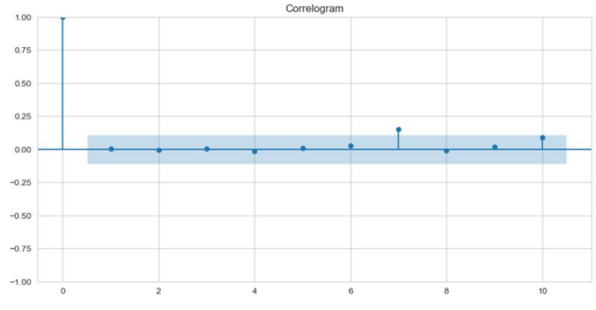
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5511	0.057	9.671	0.000	0.439	0.663
ma.L1	-0.9060	0.038	-23.554	0.000	-0.981	-0.831
sigma2	4857.9152	255.696	18.999	0.000	4356.761	5359.069

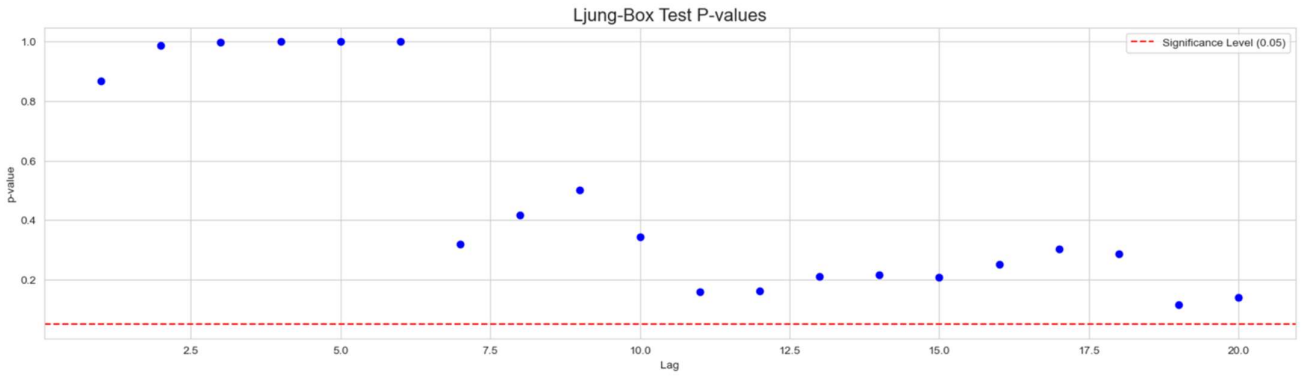
We can see that all the coefficients are coming out to be significant.

Residual Diagnostics:









The graph of the residual autocorrelation function shows a large spike at lag 7 for this model. This is also reflected in the p-values of the Ljung and Box statistic shown at the bottom of the graph which are very low for large lags. The normal Q--Q plot reveal serious departure from normality.

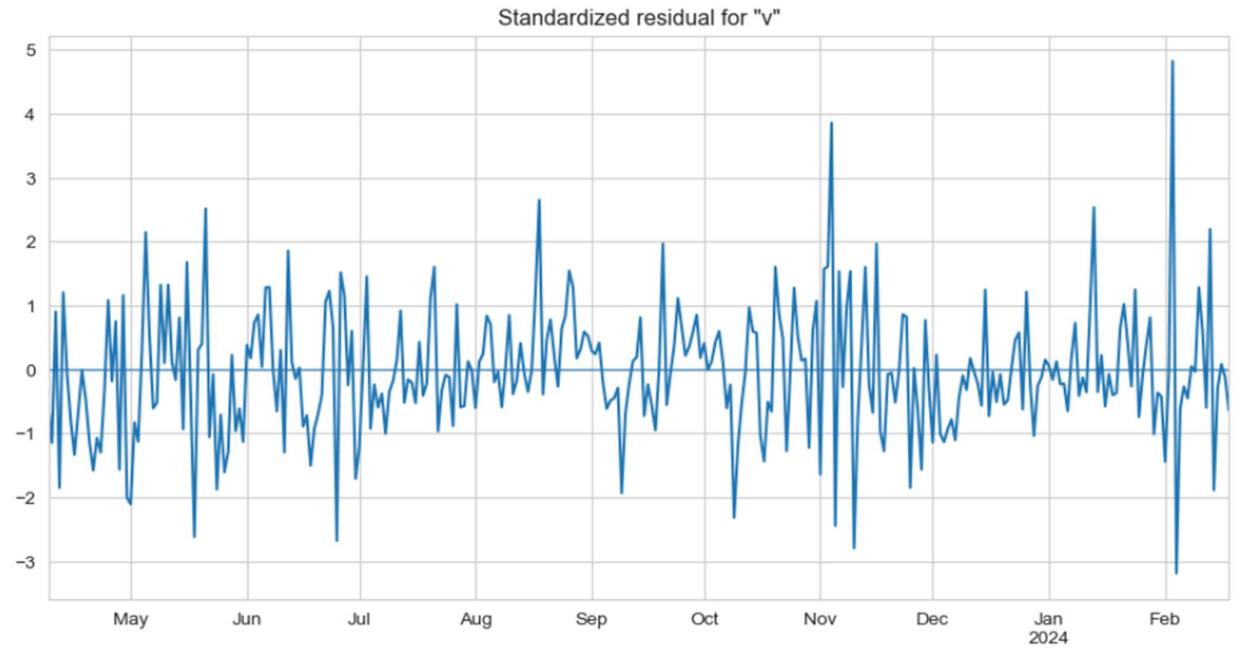
SARIMA(1, 1, 1)(0, 1, 1)

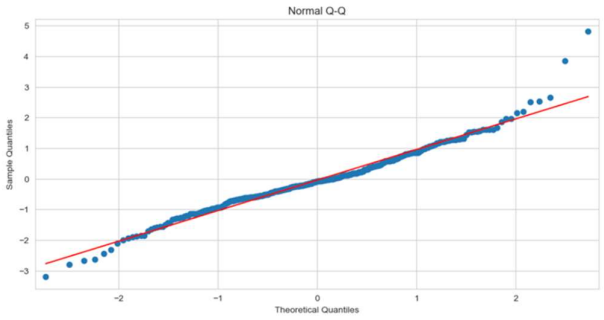
Dep. Variable:	value	**No. Observations:**	324
Model:	SARIMAX(1, 1, 1)x(0, 1, 1, 7)	**Log Likelihood**	-1800.531
Date:	Fri, 19 Apr 2024	**AIC**	3609.061
Time:	00:32:33	**BIC**	3624.084
Sample:	04-01-2023	**HQIC**	3615.063
	- 02-18-2024		
Covariance Type:	opg		

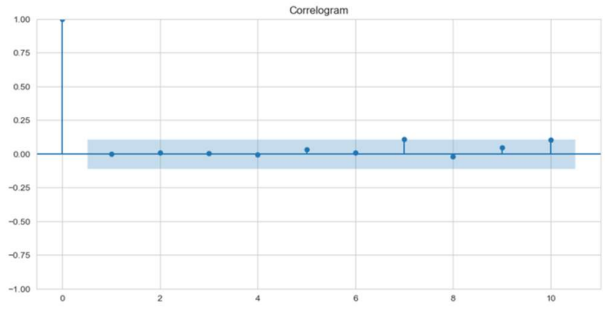
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5528	0.067	8.300	0.000	0.422	0.683
ma.L1	-0.8962	0.048	-18.652	0.000	-0.990	-0.802
ma.S.L7	-0.9183	0.033	-27.571	0.000	-0.984	-0.853
sigma2	4973.9851	283.154	17.566	0.000	4419.014	5528.956

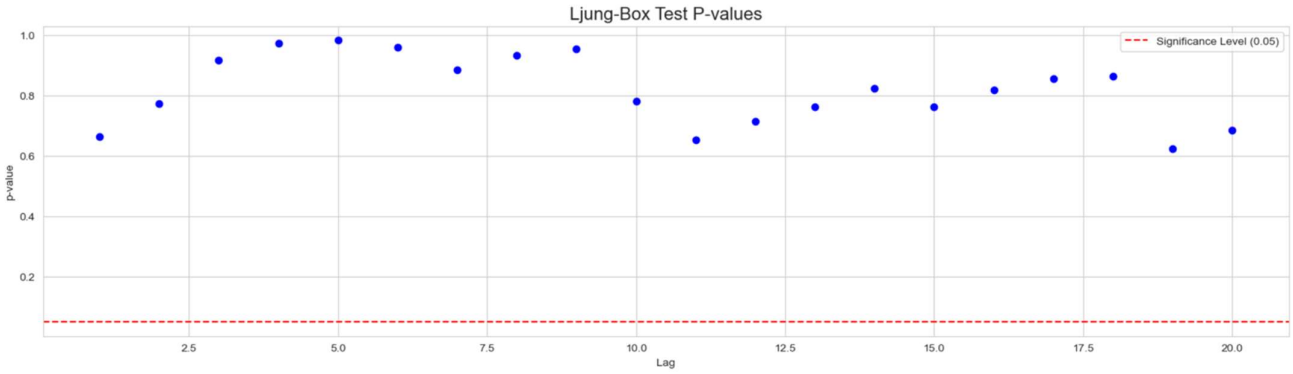
We can see that all the coefficients are coming out to be significant.

Residual Diagnostics:









In this model also graph of the residual autocorrelation function shows a large spike at lag 7 for this model. But the p-values of the Ljung and Box statistic shown at the bottom of the graph are well above the 5% level. The normal Q--Q plot reveal serious departure from normality (on the right side).

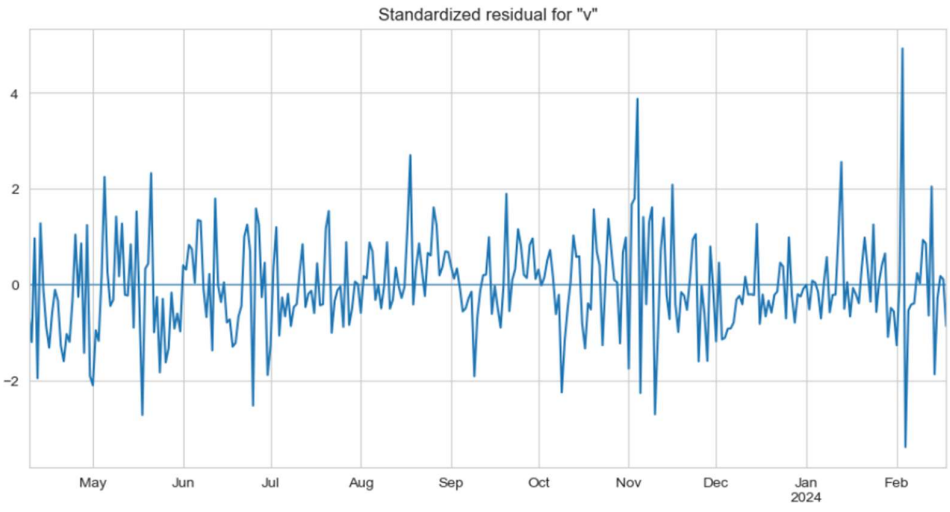
SARIMA(1, 1, 1)(1, 1, 1)

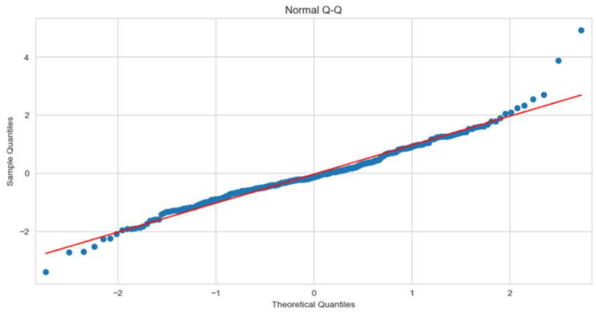
Dep. Variable:	value	**No. Observations:**	324
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 7)	**Log Likelihood**	-1798.861
Date:	Fri, 19 Apr 2024	**AIC**	3607.721
Time:	00:32:51	**BIC**	3626.500
Sample:	04-01-2023	**HQIC**	3615.223
	- 02-18-2024		
Covariance Type:	opg		

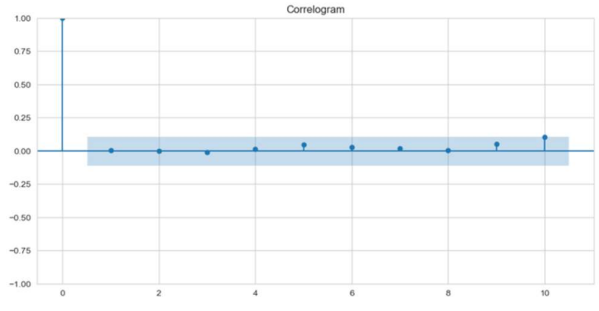
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5846	0.060	9.812	0.000	0.468	0.701
ma.L1	-0.9259	0.040	-22.921	0.000	-1.005	-0.847
ar.S.L7	0.1524	0.074	2.052	0.040	0.007	0.298
ma.S.L7	-0.9947	0.264	-3.766	0.000	-1.512	-0.477
sigma2	4757.9166	1104.292	4.309	0.000	2593.543	6922.290

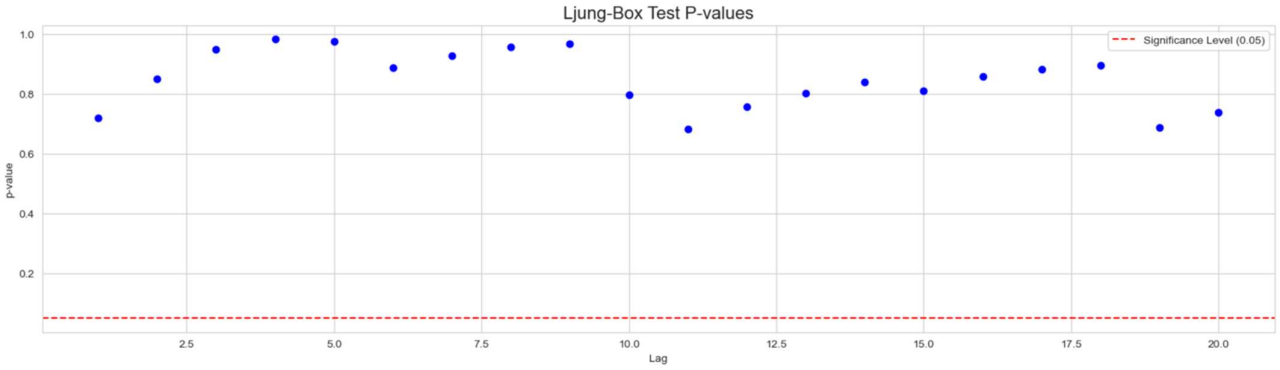
We can see that all the coefficients are coming out to be significant.

Residual Diagnostics:









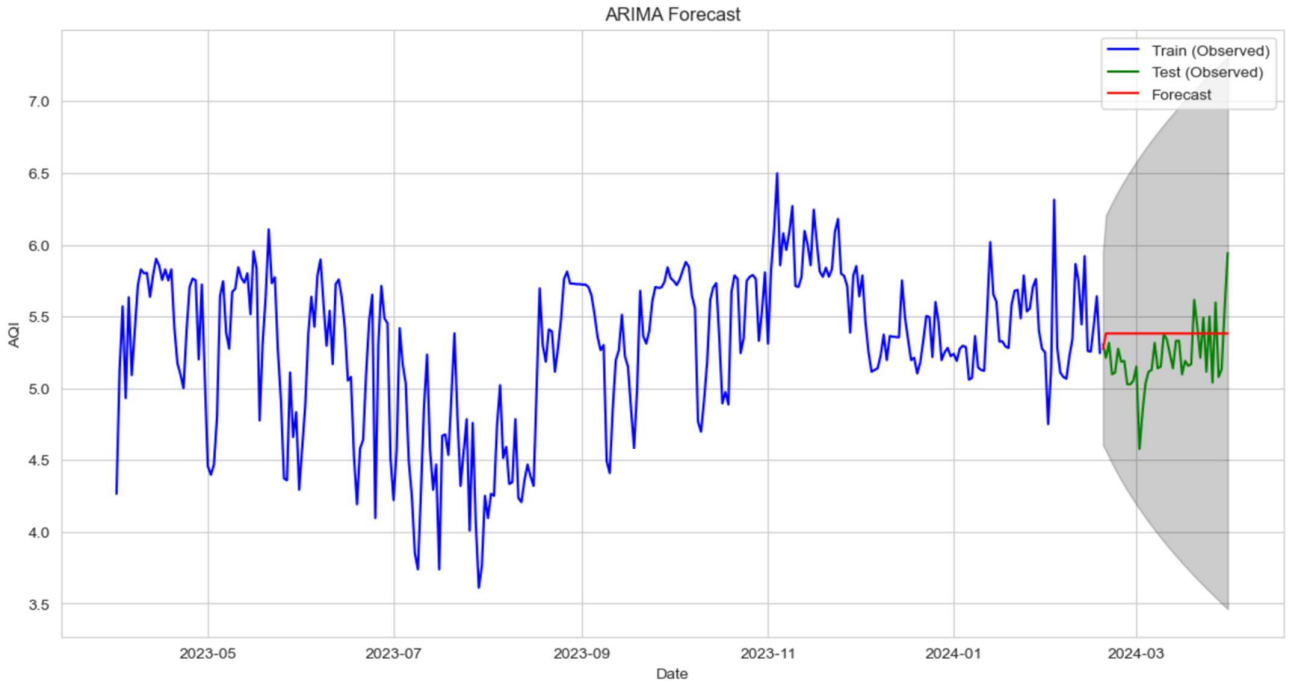
The graph of the residual autocorrelation function shows no large values for this model. This is also reflected in the p-values of the Ljung and Box statistic shown at the bottom of the graph. These diagnostic checks show a clear improvement over the SARIMA(1, 1, 1)(0, 1, 1) model examined above. The graph of the standardized residuals and the normal Q--Q plot reveal that outliers are present, however.

6- Forecasting

6.1- For Log transformed series

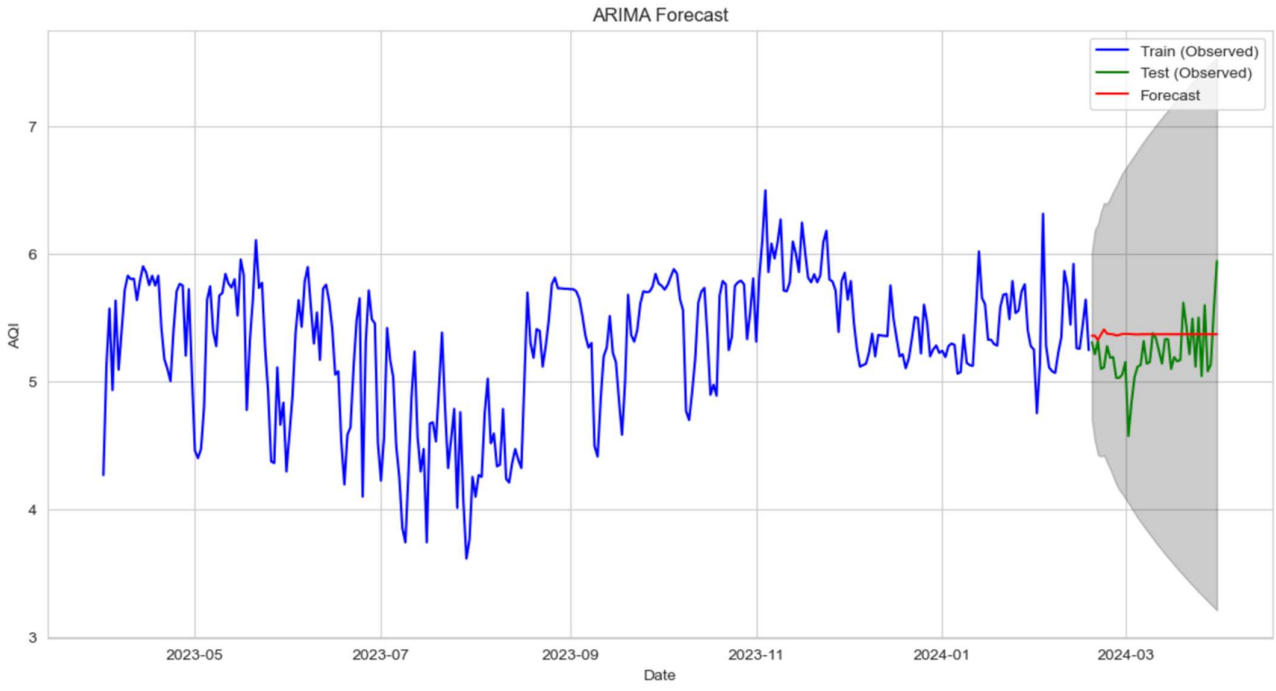
IMA(0, 1, 2)

```
Test MSE: 0.07486630184304605
```



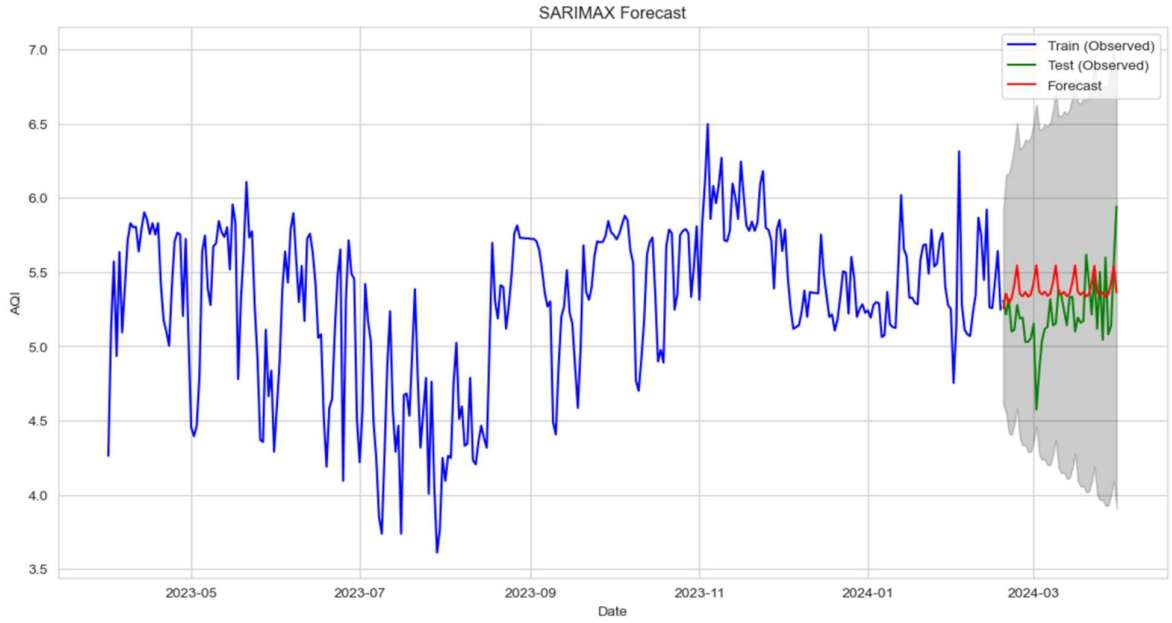
ARIMA(5, 1, 0)

```
Test MSE: 0.07146354662222733
```



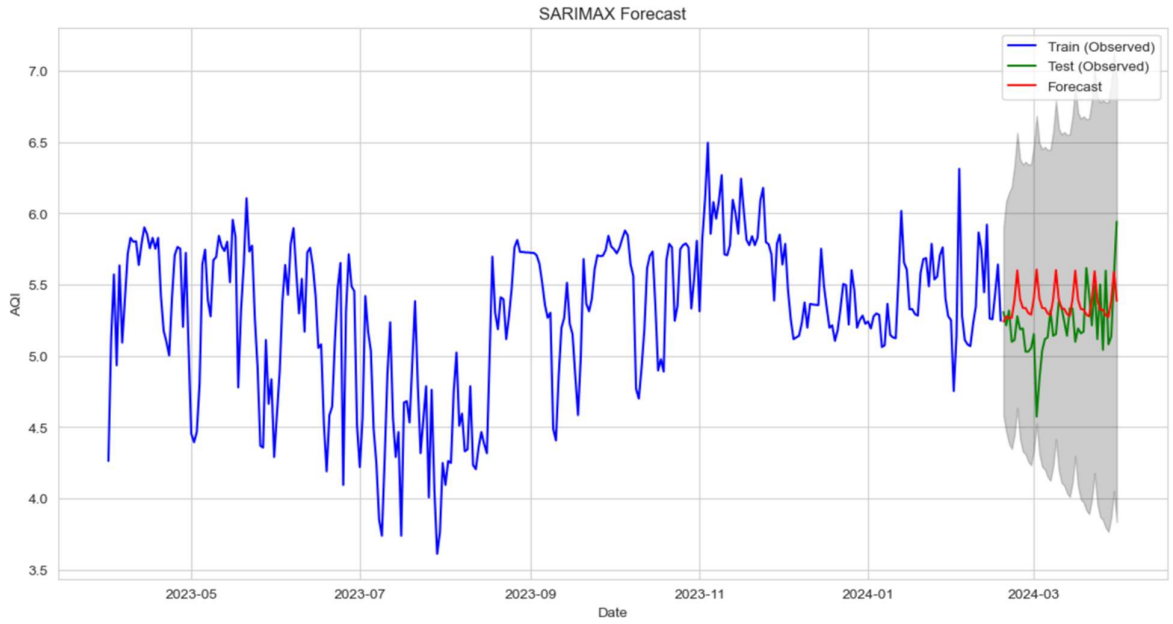
SARIMA(1, 1, 1)(2, 1, 1)

```
Test MSE: 0.08493966863732194
```



SARIMA(1, 1, 1)(0, 1, 1)

```
Test MSE: 0.08667206078582232
```



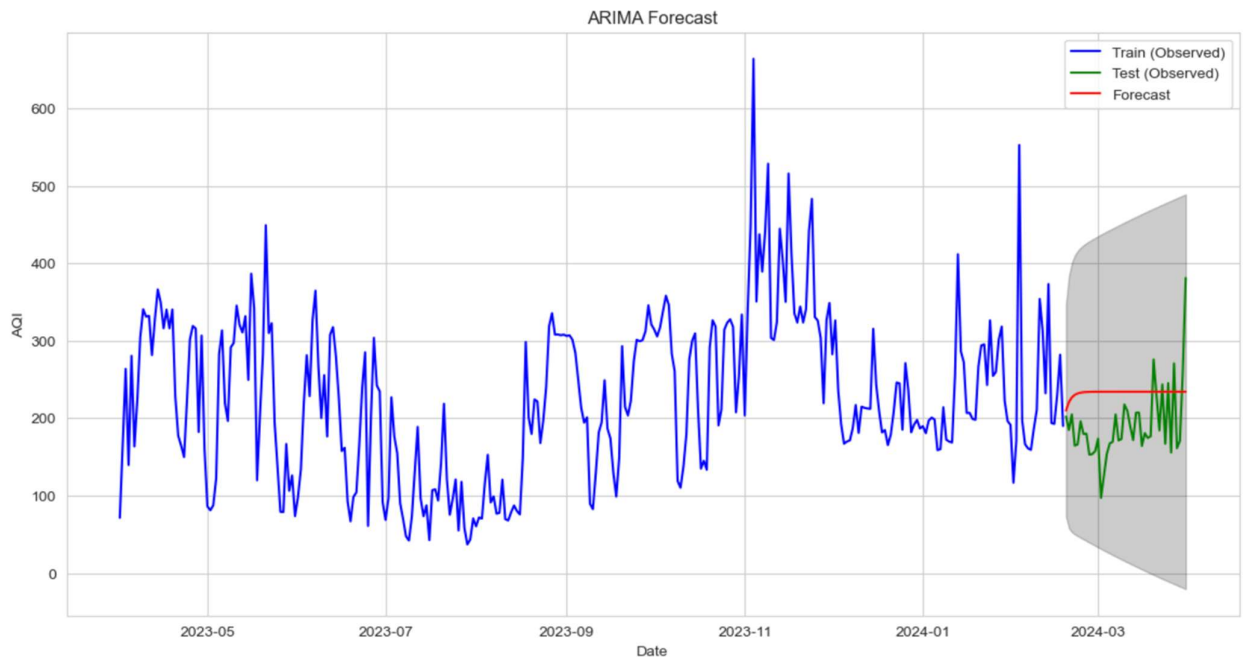
Summary:

Model	AIC	BIC	MSE (Test)
IMA(0, 1, 2)	223.009	234.342	0.07487
ARIMA(5, 1, 0)	220.179	242.845	0.07146
SARIMA(1, 1, 1)(2,1,1)	236.706	259.240	0.08494
SARIMA(1, 1, 1)(0,1,1)	237.989	253.012	0.08667

6.2- For uniform transformed series

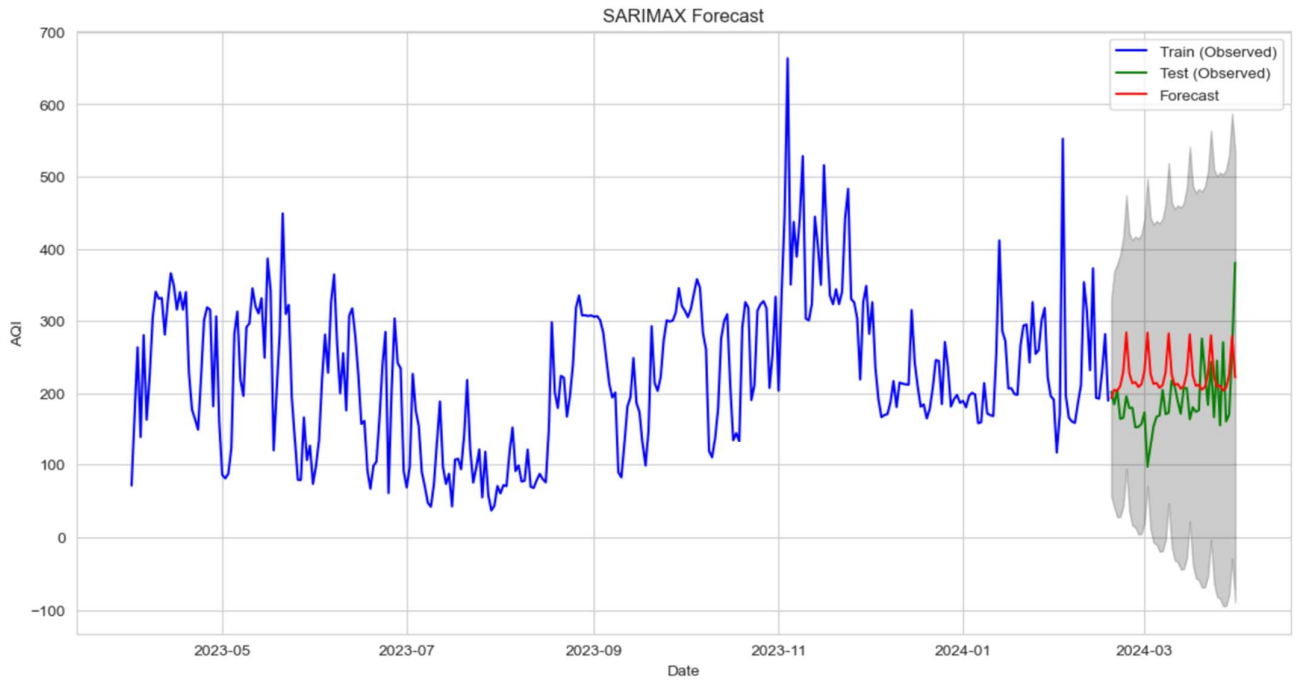
ARIMA(1, 1, 1)

```
Test MSE: 3930.7297054443084
```



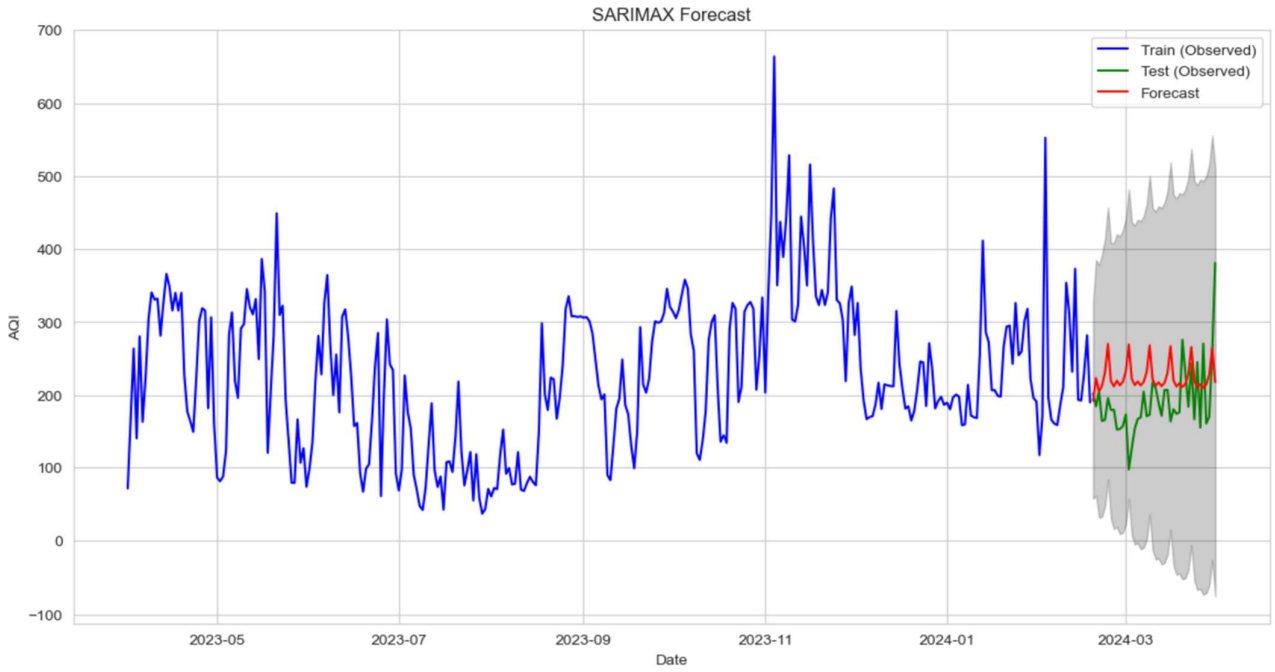
SARIMA(1, 1, 1)(0, 1, 1)

```
Test MSE: 4063.433101086319
```



SARIMA(1, 1, 1)(1, 1, 1)

```
Test MSE: 3814.9729894683287
```



Summary:

Model	AIC	BIC	MSE (Test)
ARIMA(5, 1, 0)	3665.104	3676.437	3930.7297
SARIMA(1, 1, 1)(0,1,1)	3609.061	3264.084	4063.4331
SARIMA(1, 1, 1)(1,1,1)	3607.721	3626.500	3814.9730

AQI	Associated Health Impacts
Good (0–50)	Minimal Impact
Satisfactory (51–100)	May cause minor breathing discomfort to sensitive people
Moderate (101–200)	May cause breathing discomfort to the people with lung disease such as asthma and discomfort to people with heart disease, children and older adults
Poor (201–300)	May cause breathing discomfort to people on prolonged exposure and discomfort to people with heart disease with short exposure
Very Poor (301–400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases
Severe (401-500)	**May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity**