Nonseasonal Hawaiian Airlines Time Series Analysis

Seasonal Atlantic Storms Count

Rohan Sharma

Department of Mathematical Sciences, Stevens Institute of Technology, Hoboken, NJ

Project Supervisor: Dr. Hadi Safari Katesari

Year: Spring 2025

Course: Time Series Analysis - I - MA 641

Time Series Analysis Project

Hawaiian Airlines January 2014 to December 2023 - Departure Delay Prediction

Introduction and Motivation

Departure Delay is a very common phenomenon in the airline industry. Especially in a country like America, where weather can change any hour, which is usually considered to be the biggest contributor either in arrival or departure delay.

The main goal of this project is to predict how long flights will be delayed when they depart. I am focusing on using time series data to make these predictions.

I have always been an aviation enthusiast, or what some people call an "avgeek." By analyzing data from airlines like Hawaiian Airlines, I hope to contribute valuable insights back to the avgeek community that has taught me so much. It feels great to be able to give back in this way, and I am looking forward to seeing how data can make a difference in aviation.

Data Description

Date Range : 1/1/2014 to 12/1/2023

Datasource Description: Dataset has the details of Hawaiian Airlines, from 1st January 2014 to 1st December 2023 with the following columns departure and arrival delay in minutes, originating airport, Destination, which day of the week the flight was on. Our target variable in this case is, mainly two Columns departure delay and arrival delay

Departure Delay (DEP_DELAY): This column shows how many minutes Hawaiian Airlines was delayed from its scheduled departure time. It's a key factor in understanding how often flights leave late and by how much.

Data Source:
https://www.kaggle.com/datasets/oleksiimartusiuk/bts-january-2024-commercial-flights-data/data

https://www.kaggle.com/code/dongxu027/airline-delays-eda-deep-dive-lessons-learned/notebook

https://www.kaggle.com/code/argxgd/flight-delay-exploratory-data-analysis/notebook

After taking data from these multiple sources and combining for doing feature engineering we create our base dataset with target variable Departure Delay - average delay in minutes in a month as we are doing on a monthly basis.

Some Data Analysis:

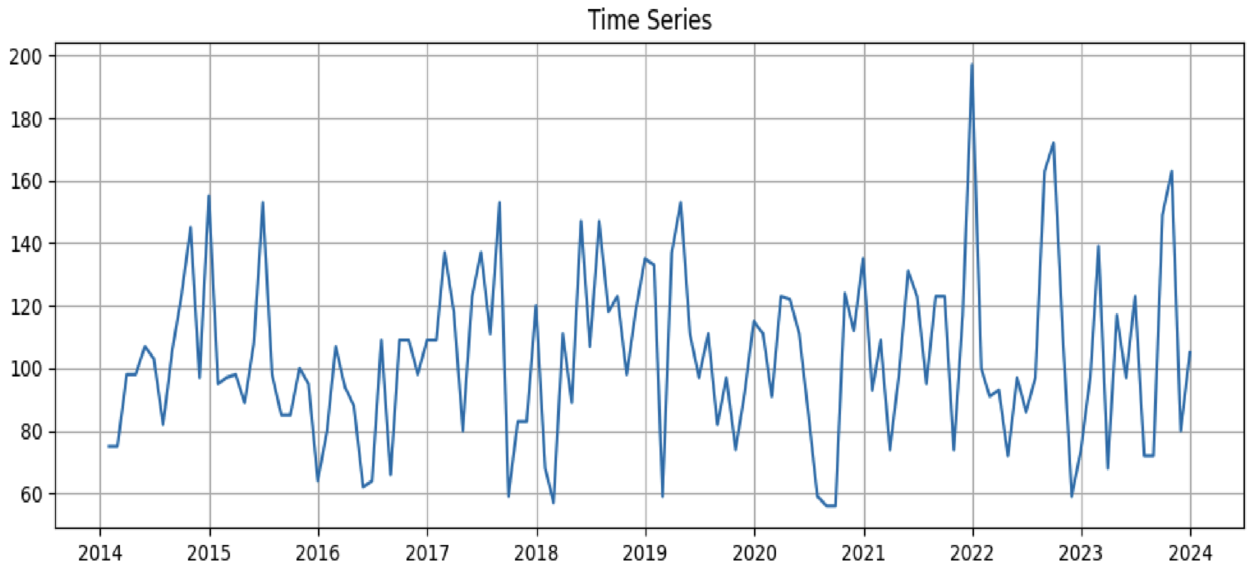

Figure.1 Time Series Analysis over years for Departure Delay

Time series over time from 2014 to 2023 for the mean of departure delay of Hawaiian Airlines in minutes.

Hence in our project we would be taking into account departure delay as a target variable to predict in our time series prediction project.

ACF and PACF Plots

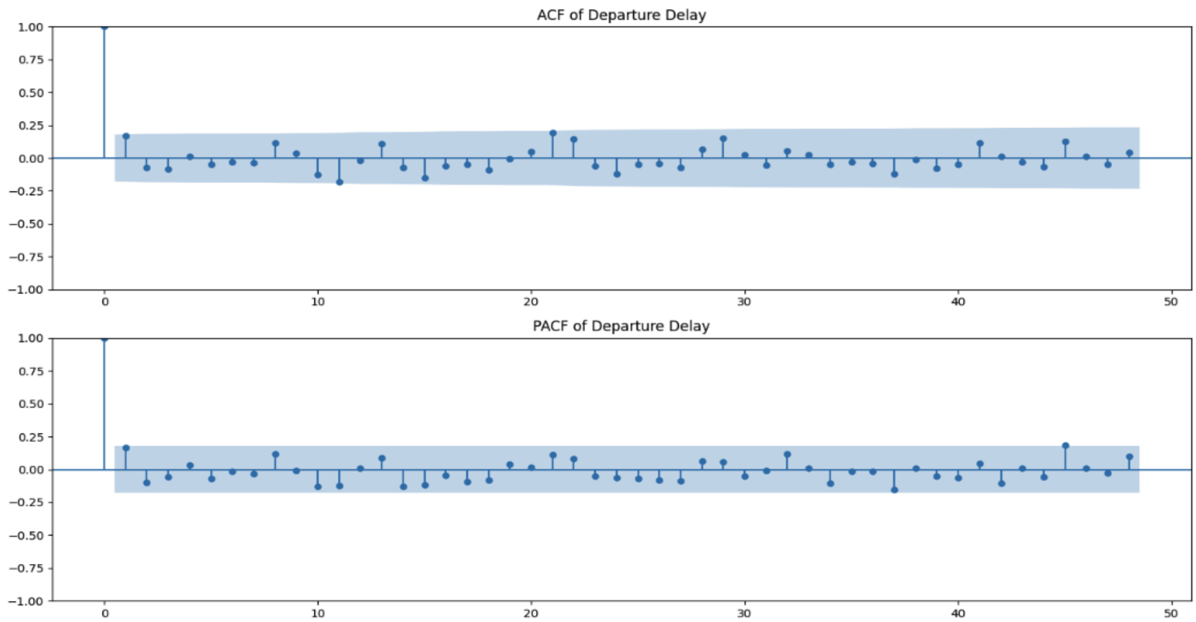

Figure.2 ACF and PACF of Departure Delay

Looking at the ACF and PACF plots the time series looks stationary, there is no seasonal pattern as we are taking data yearly we can confirm non seasonality from lags 12,24,36 and 48.

Stationarity Test (ADF Test)

To check if the data is stationary, I ran the Augmented Dickey-Fuller (ADF) test on departure delay.

Departure Delay:

```
Augmented Dickey-Fuller (ADF) Test

ADF Test Statistic: -9.17547965004152

p value: 2.326442218810547e-15

Reject the null hypothesis,the data is stationary
```

The ADF statistic was -`9.17547965004152`, and the p-value was extremely small. Since the p-value is less than 0.05, we rejected the null hypothesis. This means the departure delay data is stationary.

3. Finding Models:

After carefully looking at the ACF and PACF plots, I decided to run **for loops** for finding optimum p and q values and selecting the model with the lowest AIC and BIC value.

```
order          AIC          BIC
0  (0, 1, 2)  1133.783627  1142.120998
1  (1, 1, 1)  1134.315420  1142.652791
2  (2, 1, 1)  1135.285213  1146.401707
3  (1, 1, 2)  1135.738403  1146.854897
4  (2, 0, 2)  1135.798142  1152.523093
5  (0, 1, 1)  1136.092553  1141.650800
```

4. Parameter Redundancy:

Even with the lowest AIC and BIC values the residual analysis graphs does not seem to give us the optimal solution, hence after trying with multiple permutation and combination the best optimum models turns out to be of order

```
model_arma = ARIMA(ts_data, order=(2, 0, 2))
```

5. Residual Analysis and Results:

```
print(f"AIC : {results_arma.aic}")
print(f"BIC : {results_arma.bic}")
```

```
AIC : 1135.7981423634965
BIC : 1152.5230928201888
```

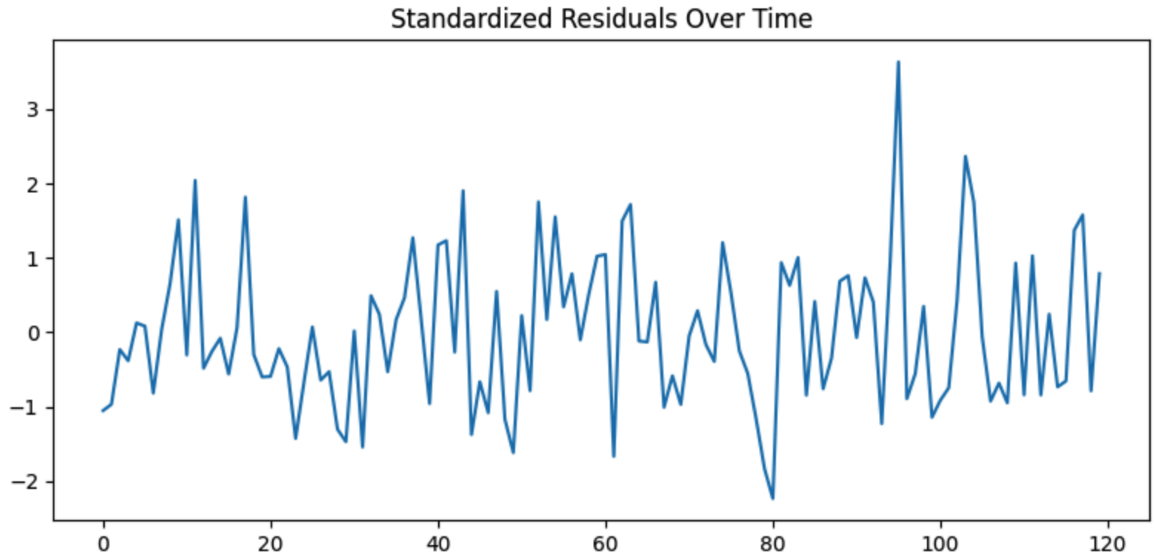

Figure.3 Standardized Residuals of ARIMA(2,0,2)

After looking at the standardised residuals plot we tend to see the residuals revolving around a constant mean of 0, though there are certain spiks at some intervals it could be because of possible noise in the data(extreme departure delay in minutes on some days because of some non controllable reasons)

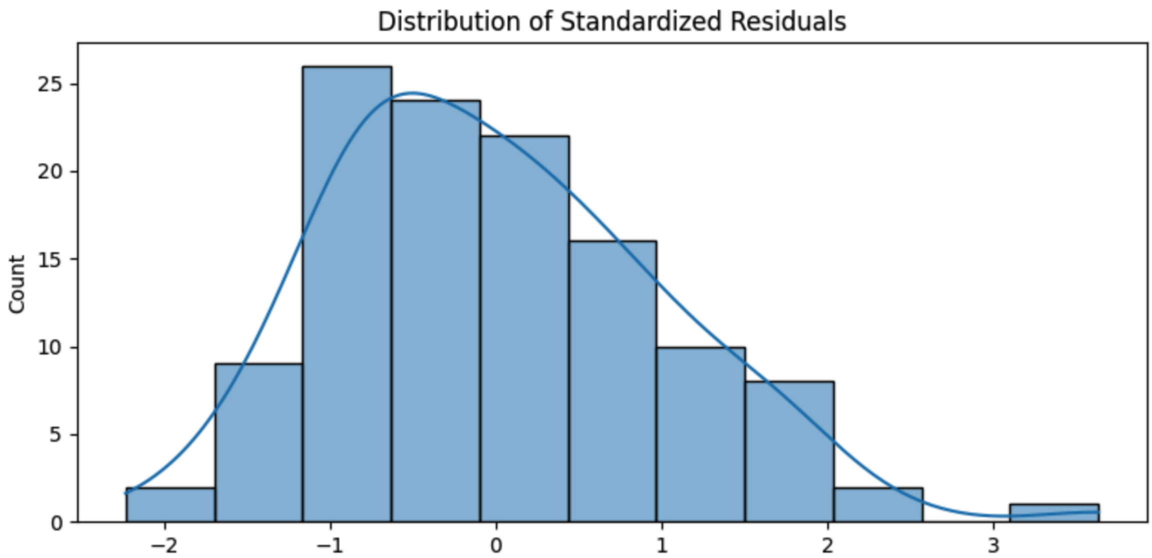

Figure.4 Histogram of Standardized Residuals

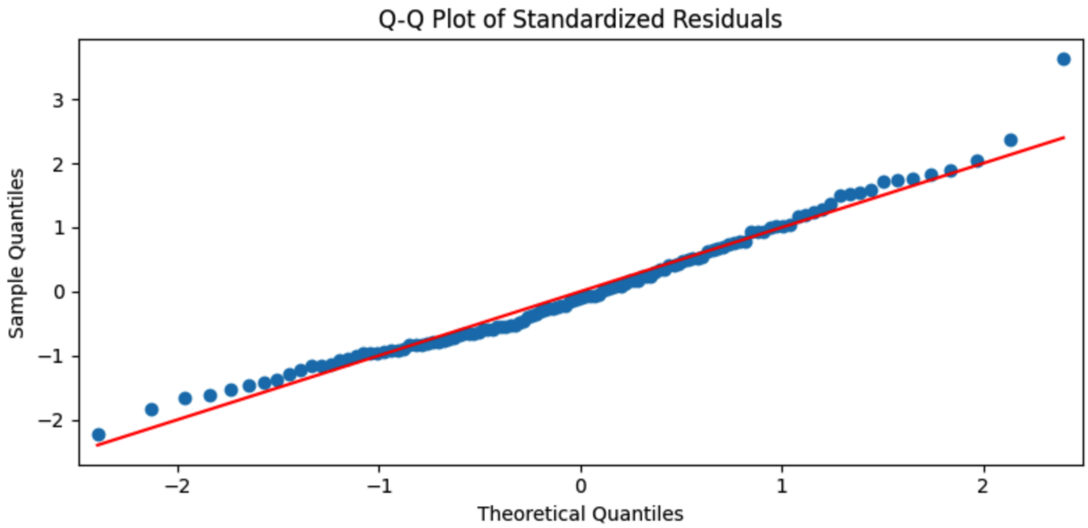

Figure.5 QQ Plot of Residuals

After looking at the QQ Plot and Histogram of Residuals the residuals do look normal, though the Histograms look right skewed but we will confirm this with the help of Shapiro Wilk test to test the normality of the residuals.

```
Shapiro-Wilk test: W = 0.9773518083821952, p-value =
0.04046057180093303
The samples are not normally distributed
```

Shapiro Wilk test REJECTS NORMALITY

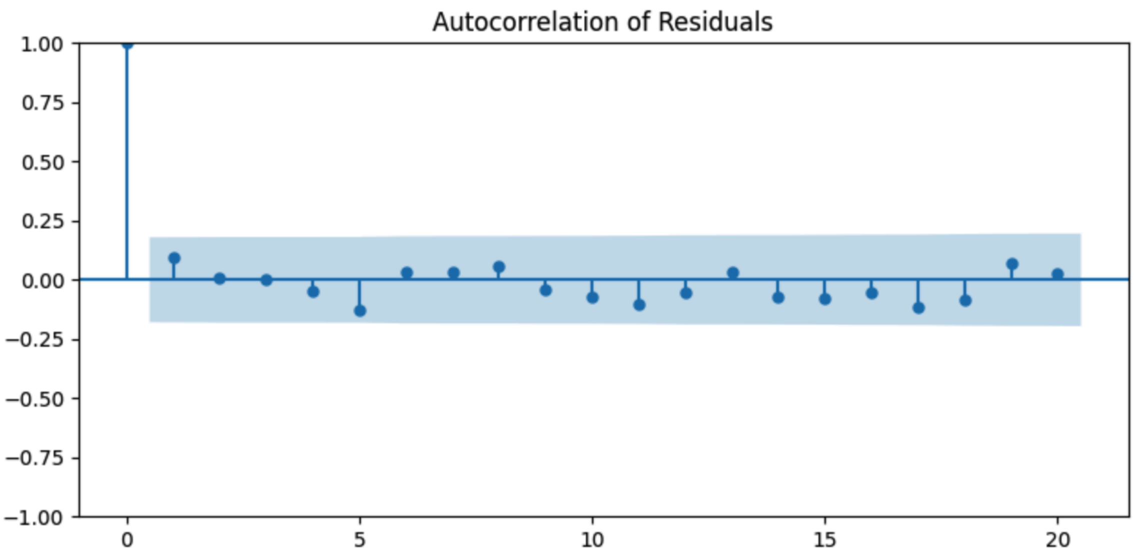

Figure.6 ACF of Residuals

Looking at the plot of residuals we can confirm that the ACF of residuals is a **WHITE NOISE** as all the lags lie inside the significant line, signifying **there is NO AUTOCORRELATION among the residuals.**

We confirm this again with Ljung's box test for all the 20 lags.

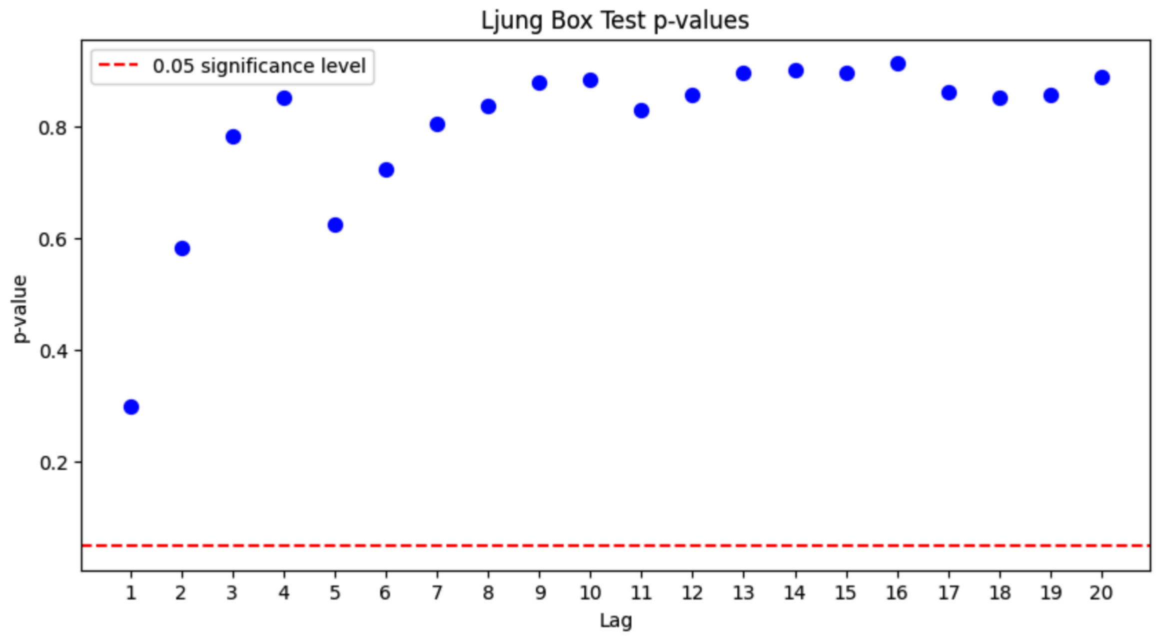

Figure.7 Ljung's Box Test p - values

As for all the 20 lags we have p values greater than 0.05 (significant line) , meaning and it confirms that **there is NO AUTOCORRELATION** again among the residuals.

P - values for first 20 lags

```
1     0.300254
2     0.582530
3     0.781723
4     0.851261
5     0.623961
6     0.723484
7     0.805789
8     0.835601
9     0.878496
10    0.883556
11    0.829590
12    0.857358
13    0.895063
14    0.899719
15    0.894972
16    0.912457
17    0.862433
18    0.852568
19    0.857106
20    0.888285
```

6. Forecasting:

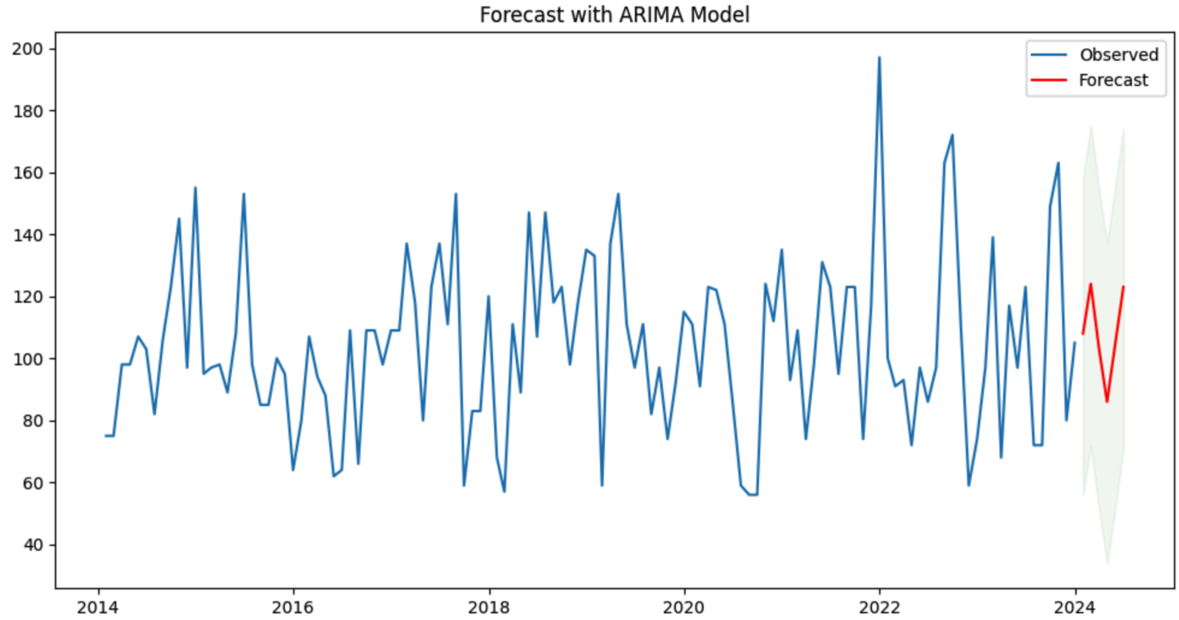

Figure.8 Forecasting with ARIMA (2,0,2) Model

As we can see from the graph that the forecasting values are not a flat line usually the case in many cases, we can say that our model is performing well capturing the upward and downward trends.

The values for the forecast for 6 months are as follows -

	Forecast	Lower Bound	Upper Bound
2024-01-31	108.0	56.0	158.0
2024-02-29	124.0	72.0	175.0
2024-03-31	103.0	51.0	154.0
2024-04-30	86.0	34.0	137.0
2024-05-31	105.0	52.0	156.0
2024-06-30	123.0	71.0	174.0

==

Atlantic Storm January 2010 to November 2022 - Total Number of Storms Prediction

Introduction and Motivation

Storms have always fascinated me not just because of their intensity, but because of the patterns they leave behind over time.

This project is a chance to let the data speak and when it comes to something as dynamic as storms, there's no shortage of stories waiting to be told.

There is clearly a trend where the number of Atlantic storms increases in the months of August, September, October, November and sometimes December.

Data Description

Date Range : 1/1/2010 to 11/1/2022

Datasource Description: The dataset tracks named storms from 1950 onward,
Though for our model we consider data from 2010, to get a better sense of seasonality and storm patterns over time,
target variable: total number of storms per month.

We are grouping storms month-wise across all years. I can start to see which months are historically storm heavy and which are relatively calm. This is one of those small steps that adds big value.

Data Source:
https://www.kaggle.com/datasets/thedevastator/atlantic-named-storms-maximum-wind-speeds-1950-p

Some Data Analysis:

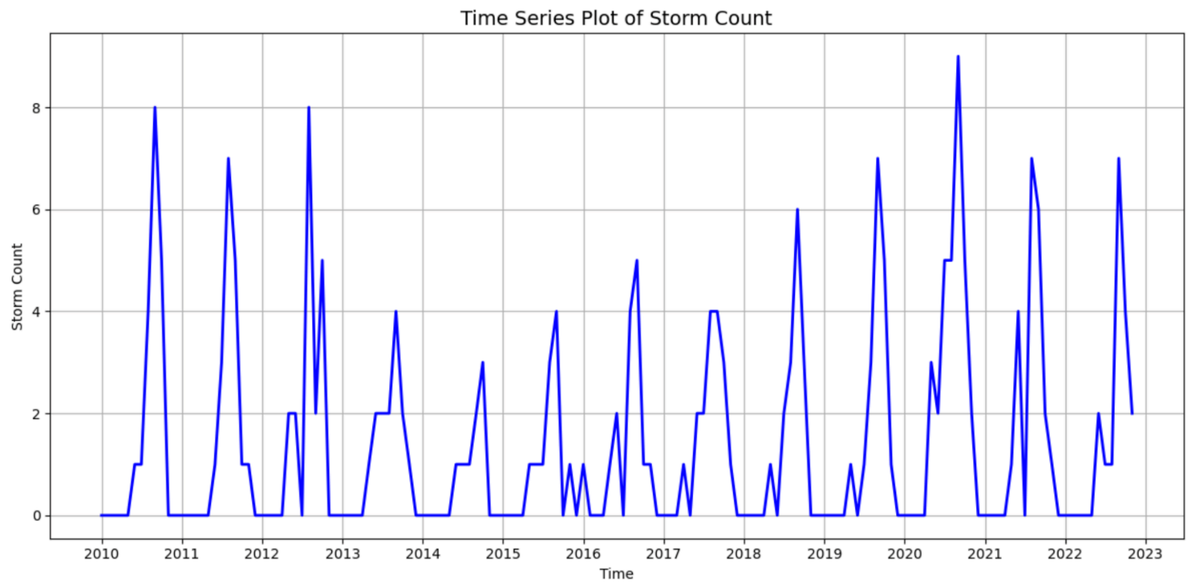

Figure.9 Time Series Analysis over years for Total Storms

Time series over time from 2010 to 2022 for the mean of Total Storms in count.

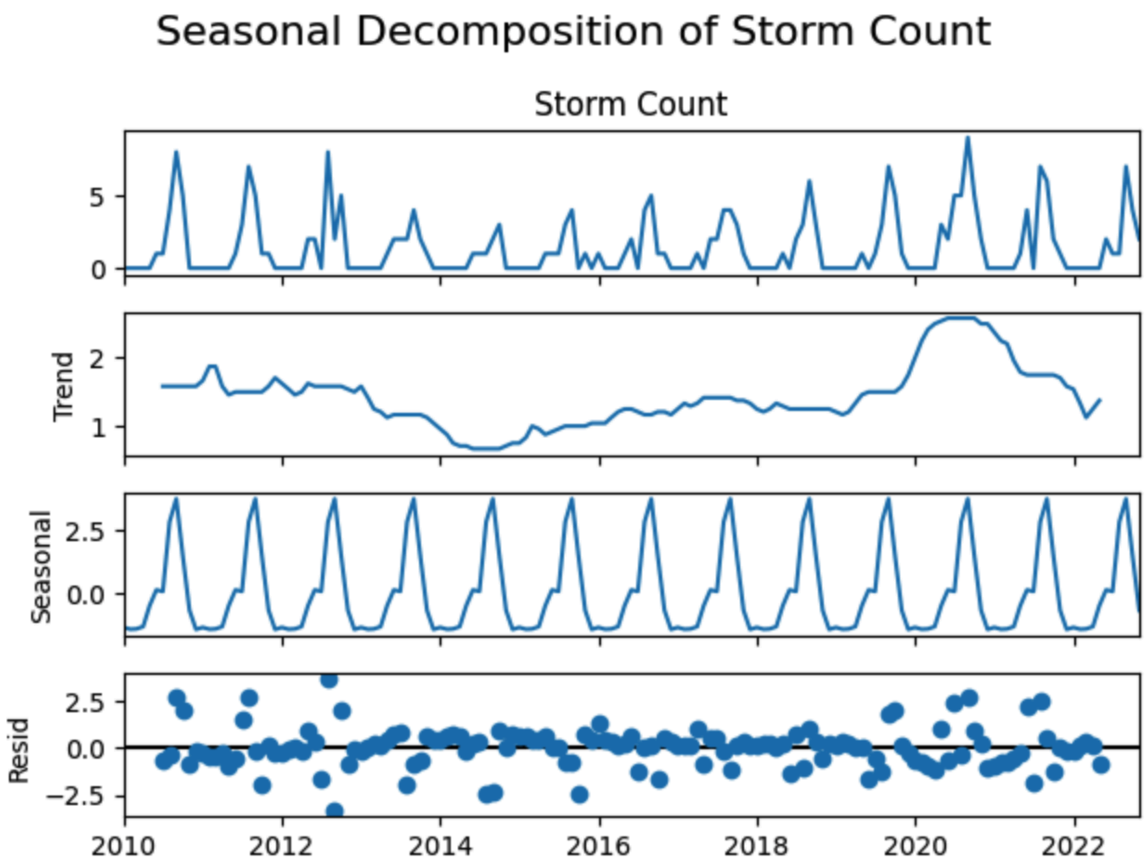

Figure.10 Seasonal Decomposition of Total Storms Data

We also see from the Seasonal Decomposition Graph, that **there is a clear cut seasonal pattern observed in terms of the total number of storms.**

Hence in our project we would be taking into account Total Storms as a target variable to predict in our time series prediction project.

ACF and PACF Plots

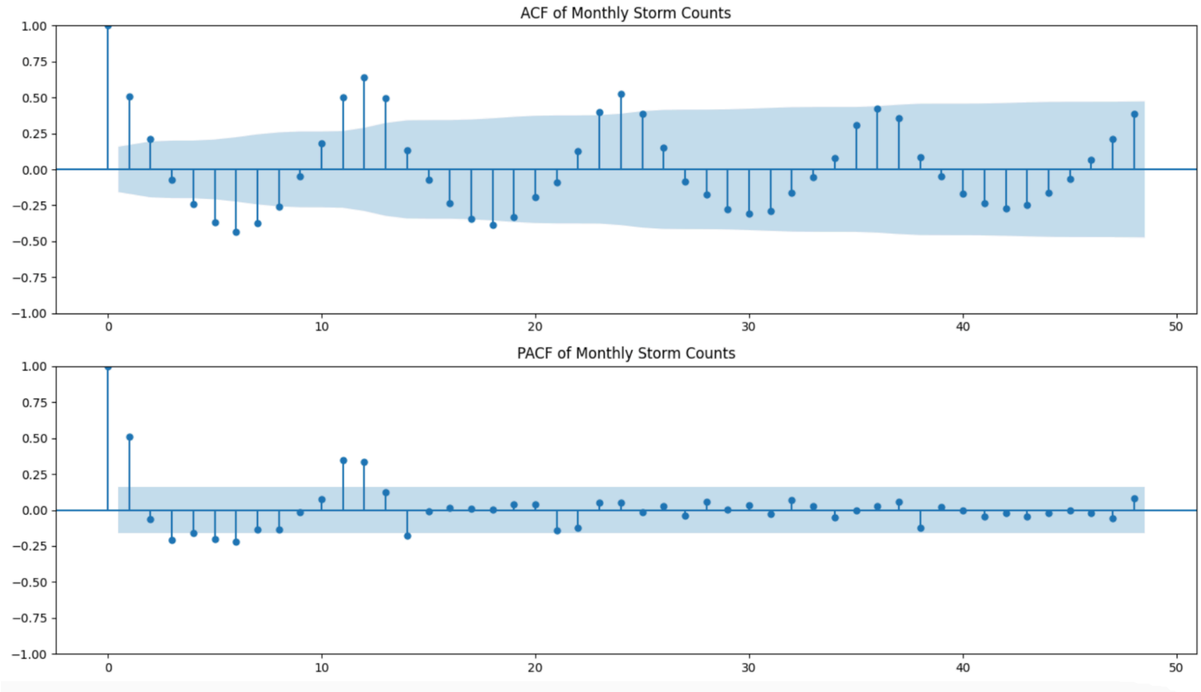

Figure.11 ACF and PACF of Total Storms Count before differencing

Looking at the ACF and PACF plots the time series does stationary, **there is a seasonal pattern as we are taking data yearly we can confirm seasonality from lags 12,24,36 which seems to be significant as we see from the ACF Plot**

Stationarity Test (ADF Test)

To check if the data is stationary, I ran the Augmented Dickey-Fuller (ADF) test on Total Storms.

Total Storms:

```
Augmented Dickey-Fuller (ADF) Test

ADF Test Statistic: -1.6445

p value: 0.4599


Fail to reject the null,the data is not stationary
```

The ADF statistic was `-1.6445`, and the p-value is `0.4599`. Since the p-value is greater than 0.05, we do not reject the null hypothesis. This means the total storm count variable is **not stationary**.

After doing **differencing of order 1**, we see the differenced series first.

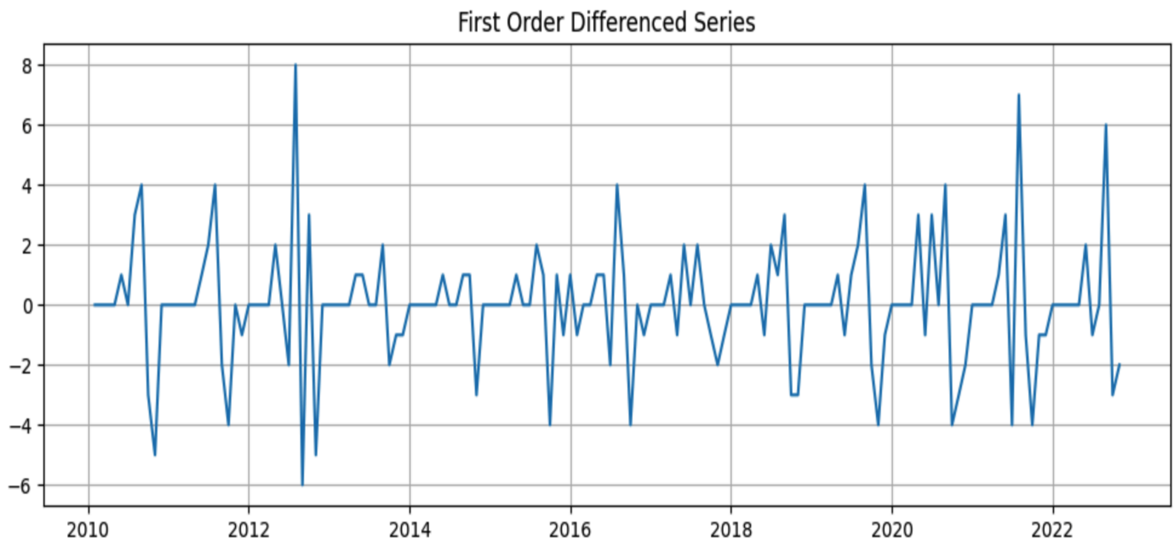

Figure.12 First Order Differenced Series

Now we check stationarity using the ADF test again on the differenced data just to confirm if the target variable is still stationary or not.

```
ADF Test on First Differenced Series:
ADF Statistic: -6.6073226742066025
p-value: 6.507240026478501e-09

 Reject the null hypothesis,the data is stationary
```

The first ordered differenced series makes our target variable stationary.

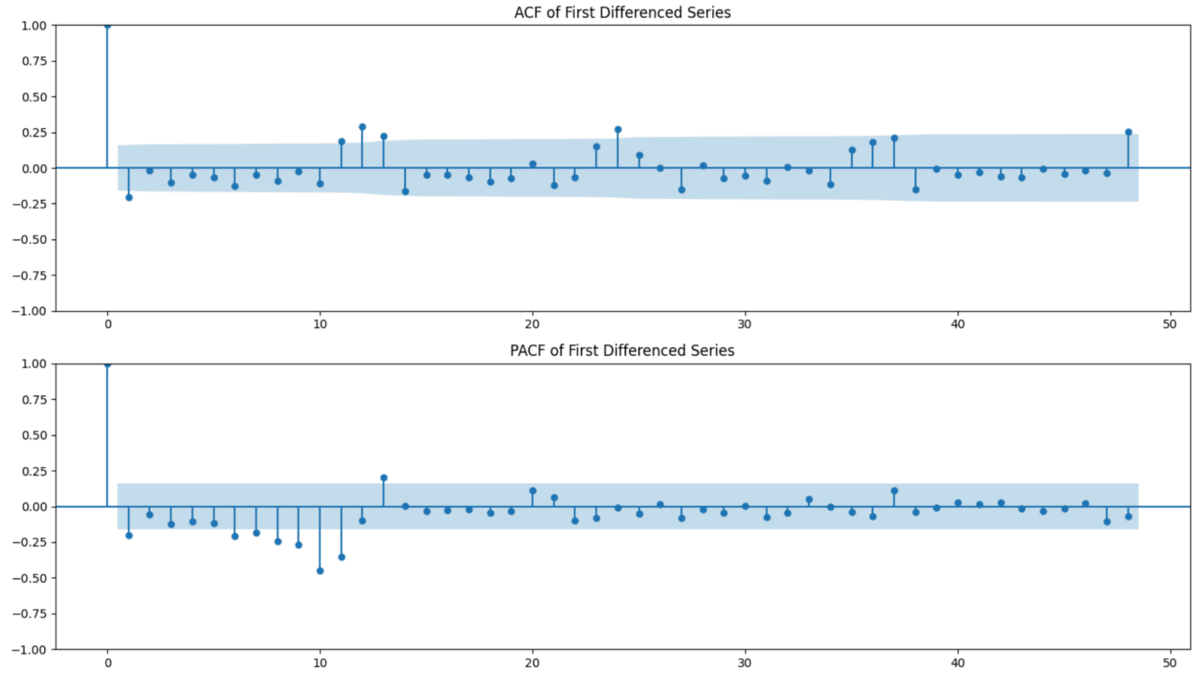

Figure.13 ACF and PACF of Total Storms Count after differencing

We can clearly see seasonality in the differenced ACF and PACF plot, in the ACF Plot we can see the significant lags at lag 12,24 and 48 thus confirming the seasonality

3. Finding Models:

After carefully looking at the ACF and PACF plots, I decided to run **for loops** for finding optimum p d, q, P,D,Q - p,d,q signifying non seasonal values and P,D,Q signifying seasonal values and selecting the model with the lowest AIC and BIC value. I am running the **SARIMA Model** because of the seasonality pattern observed in our data.

	ord	s_ord	AIC	BIC
0	(2, 0, 2)	(0, 1, 1, 12)	433.039188	450.104311
1	(2, 0, 1)	(0, 1, 1, 12)	433.291369	447.551520
2	(1, 0, 2)	(0, 1, 1, 12)	434.847923	449.068858
3	(2, 0, 2)	(1, 1, 1, 12)	435.015219	454.924528
4	(0, 0, 2)	(0, 1, 1, 12)	435.106966	446.483714
5	(2, 1, 2)	(0, 1, 1, 12)	435.331663	452.349355

4. Parameter Redundancy:

The first model with the lowest AIC and BIC value is the best model as I tried with all the top 5 models and the best one in terms of parameters redundancy and residual analysis turns out to be -

```
model_sarima = SARIMAX(ts_data, order=(2, 0, 2), seasonal_order=(0,1,1,
12))
```

5. Residual Analysis and Results:

```
print(f"AIC : {results_sarima.aic}")
print(f"BIC : {results_sarima.bic}")
```

```
AIC : 494.63781716549715
BIC : 512.3727795111047
```

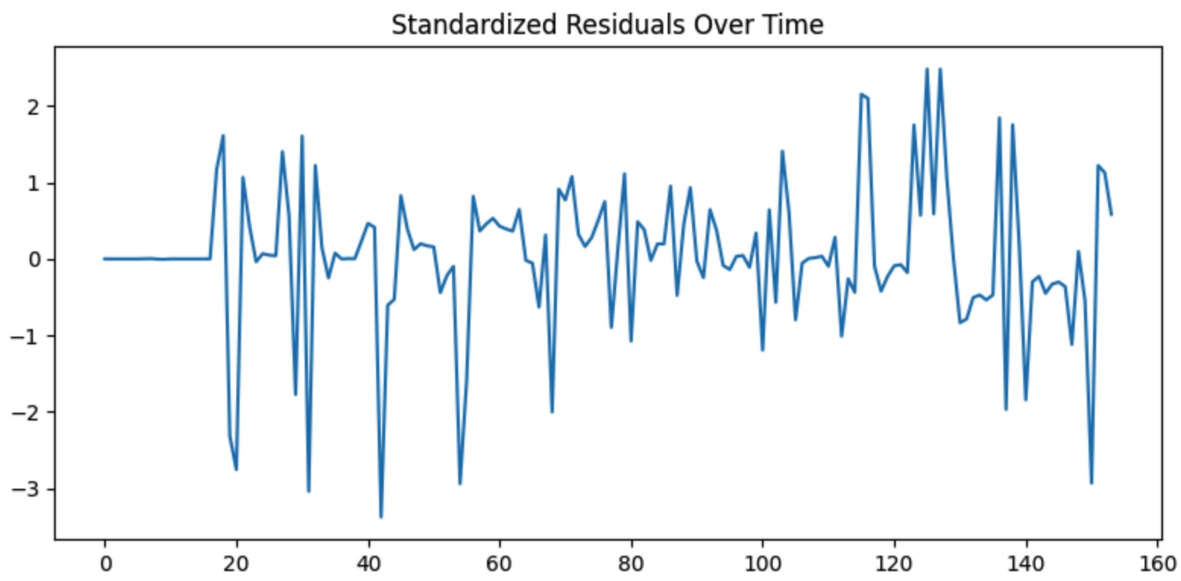

Figure.14 Standardized Residuals of **SARIMA** `(2, 0, 2) (0, 1, 1, 12) Model`

After looking at the standardised residuals plot we tend to see the residuals revolving around a constant mean of 0, though there are certain spiks at some intervals.

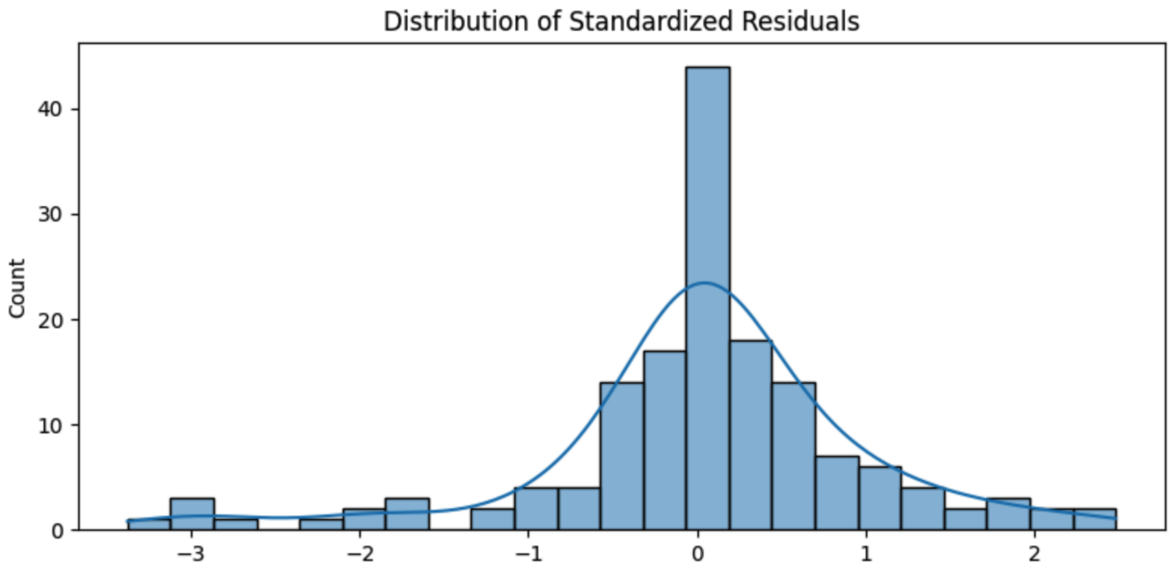

Figure.15 Histogram of Standardized Residuals

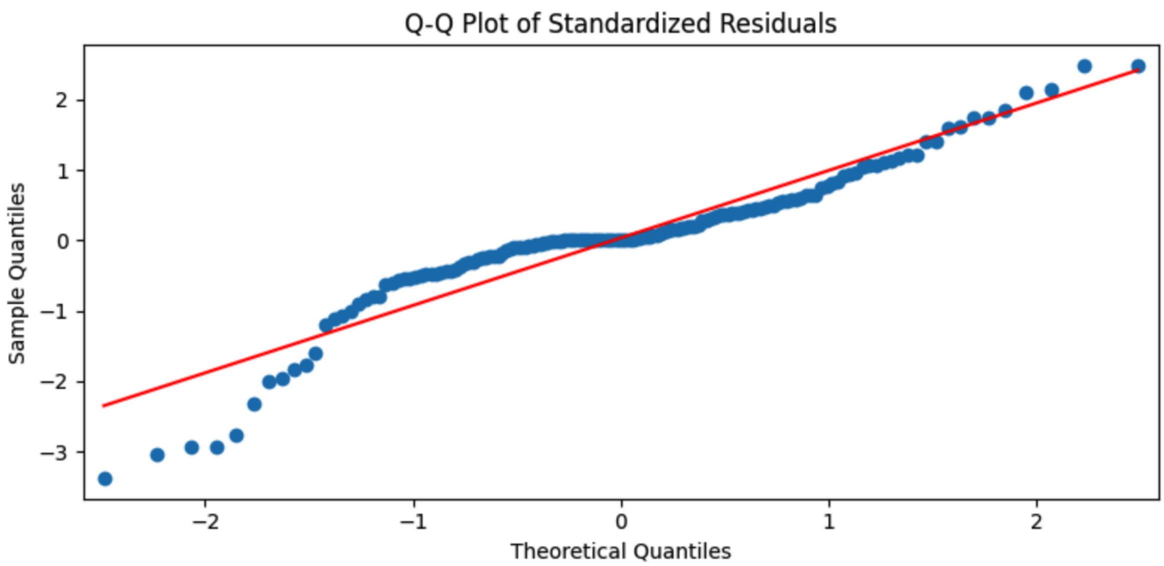

Figure.16 QQ Plot of Residuals

After looking at the Histogram of Residuals the residuals do look normal but QQ Plot doesnt seem to show normality, but we will confirm this with the help of Shapiro Wilk test to test the normality of the residuals.

```
Shapiro-Wilk test: W = 0.9104301739935857, p-value =
3.886311323477132e-08
The samples are not normally distributed
```

Shapiro Wilk test REJECTS NORMALITY

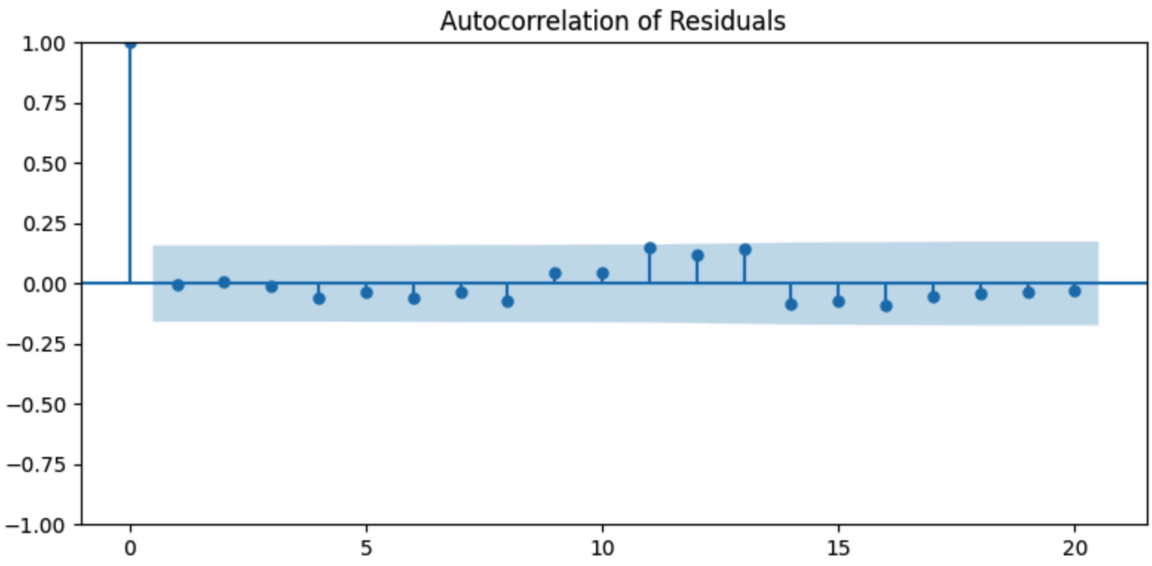

Figure.17 ACF of Residuals

Looking at the plot of residuals we can confirm that the ACF of residuals is a **WHITE NOISE** as all the lags lie inside the significant line, signifying **there is NO AUTOCORRELATION among the residuals.** Though at lag 11,13 it seems to be close but it does not cross the significant line and it could be close because of the noise in the data.

We confirm this again with Ljung's box test for all the 20 lags.

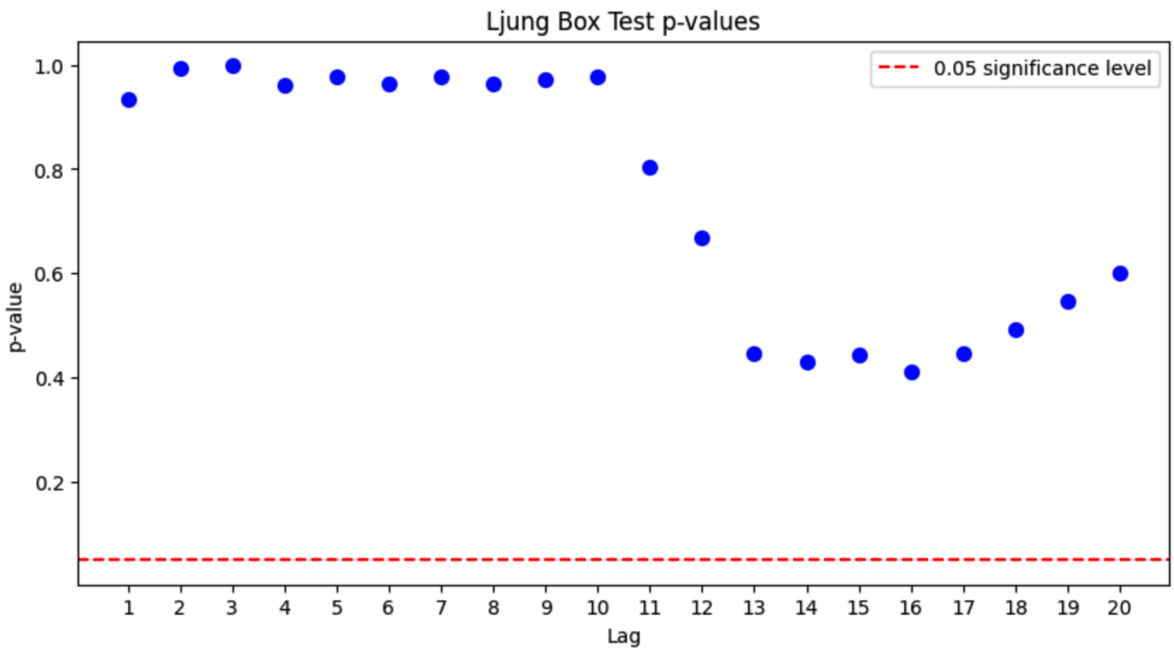

Figure.18 Ljung's Box Test p - values

As for all the 20 lags we have p values greater than 0.05 (significant line) , meaning and it confirms that **there is NO AUTOCORRELATION** again among the residuals.

P - values for first 20 lags

1	0.935359
2	0.994213
3	0.998455
4	0.962250
5	0.976132
6	0.963477
7	0.977784
8	0.962569
9	0.970810
10	0.976258
11	0.804547
12	0.669372
13	0.446756
14	0.431262
15	0.442936
16	0.411577
17	0.446407
18	0.492954
19	0.545915
20	0.600153

6. Forecasting:

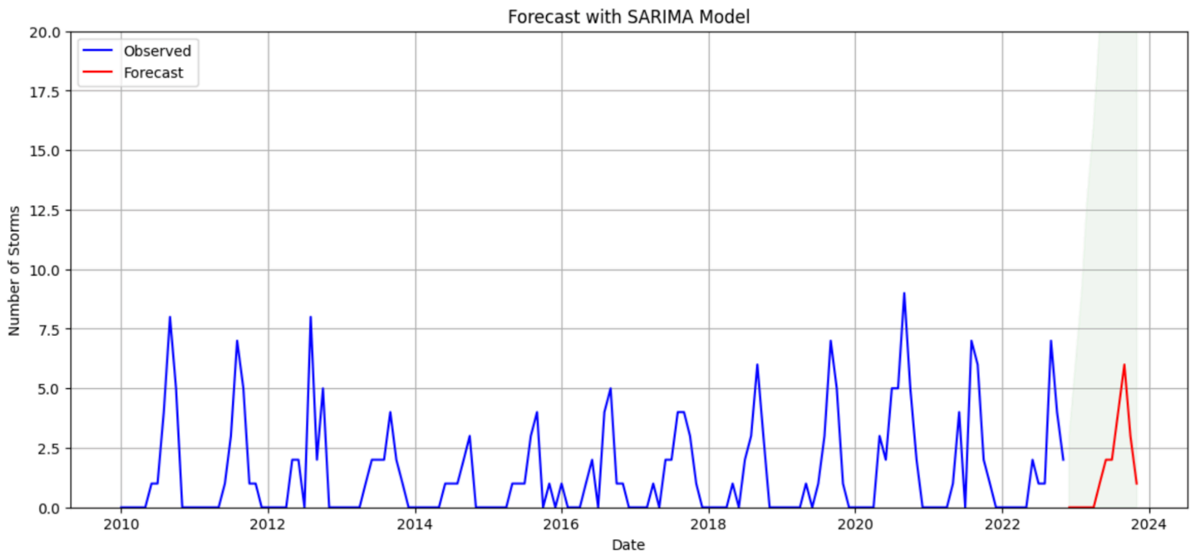

Figure.19 Forecasting with **SARIMA** `(2, 0, 2) (0, 1, 1, 12)` **Model**

As we can see from the graph that the forecasting values are not a flat line usually the case in many cases, we can say that our model is performing well capturing the upward and downward trends.

The values for the forecast for 12 months are as follows -

	Forecast	Lower Bound	Upper Bound
2022-12-31	0.0	0.0	3.0
2023-01-31	0.0	0.0	6.0
2023-02-28	0.0	0.0	9.0
2023-03-31	0.0	0.0	13.0
2023-04-30	0.0	0.0	16.0
2023-05-31	1.0	0.0	20.0
2023-06-30	2.0	0.0	25.0
2023-07-31	2.0	0.0	28.0
2023-08-31	4.0	0.0	34.0
2023-09-30	6.0	0.0	39.0
2023-10-31	3.0	0.0	40.0
2023-11-30	1.0	0.0	41.0

==