TIME SERIES ANALYSIS ON WALMART SALES DATA

VISHNU M

23M0025

PROJECT REPORT

May 2024

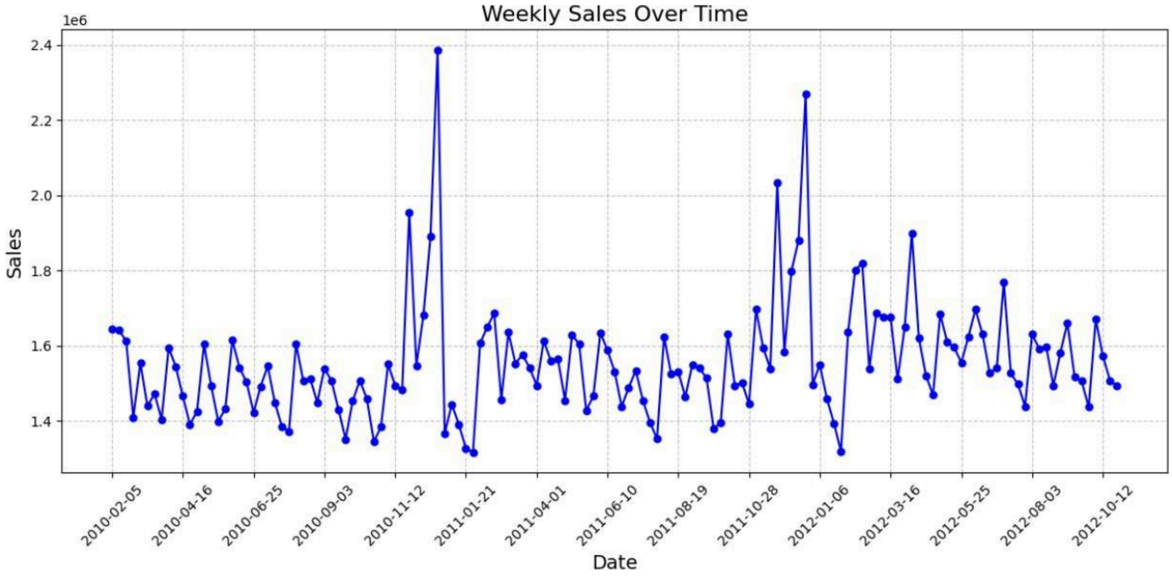
Dataset Description:

This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in the data Walmart_Store_sales. The columns are

- Store - the store number
- Date - the week of sales
- Weekly_Sales - sales for the given store
- Holiday_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
- Temperature - Temperature on the day of sale
- Fuel_Price - Cost of fuel in the region
- CPI – Prevailing consumer price index
- Unemployment - Prevailing unemployment rate
- Holiday Events
 Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
 Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
 Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
 Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

There are total 5 stores and corresponding to each store there are 144 datapoints.

STEP 1: Visualization of the dataset:

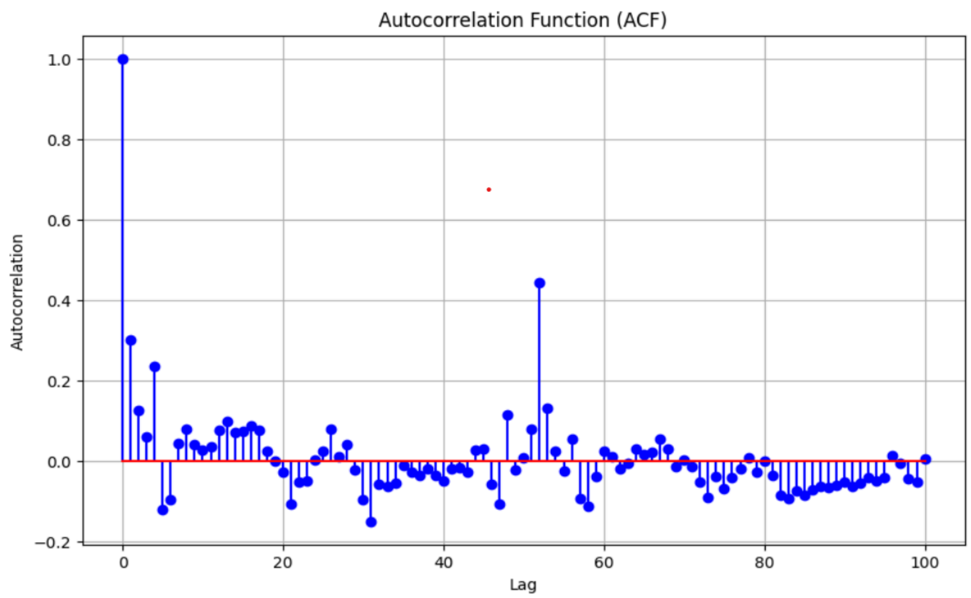


Observation:

- We can see that in this weekly data, the sales reach a peak at the beginning of each month which is quite obvious since customers tend to shop at a large scale in the beginning of the month. (seasonality)
- There is a slight increasing trend over the years. (Trend)
- We can observe that there is a huge peak at the end of year 2010 and 2011 which can be explained by increased sales during holiday season. (Cyclicity)

CHECKING FOR STATIONARITY:

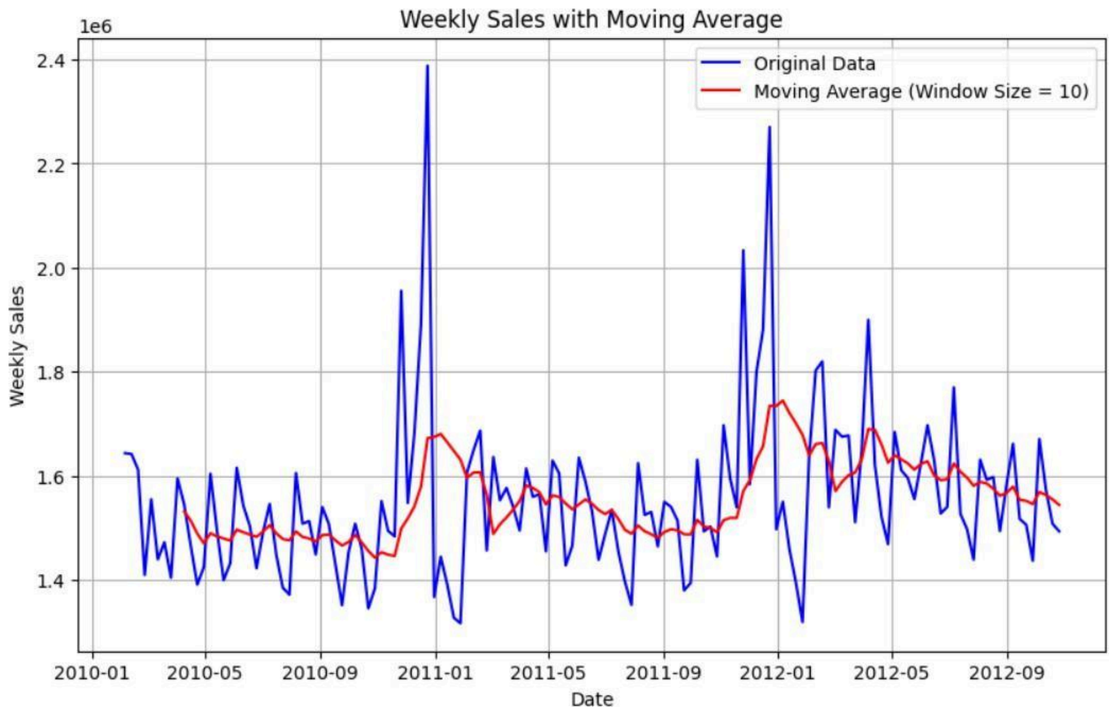
ACF plot for the sales:



Observations:

- The ACF plot does not seem to be exponentially decreasing that is it does not seem to be stationary.
- Although there seems to be some sort of pattern in the ACF plot of the original series which hints that there might be some periodicity in the original time series plot.

Moving Average of the weekly sales data:



Observation:

- We can see that the mean is clearly non constant over time which again indicates that the original series in non-stationary.

MODEL ASSUMPTIONS:

We assume that the model will be in the following form

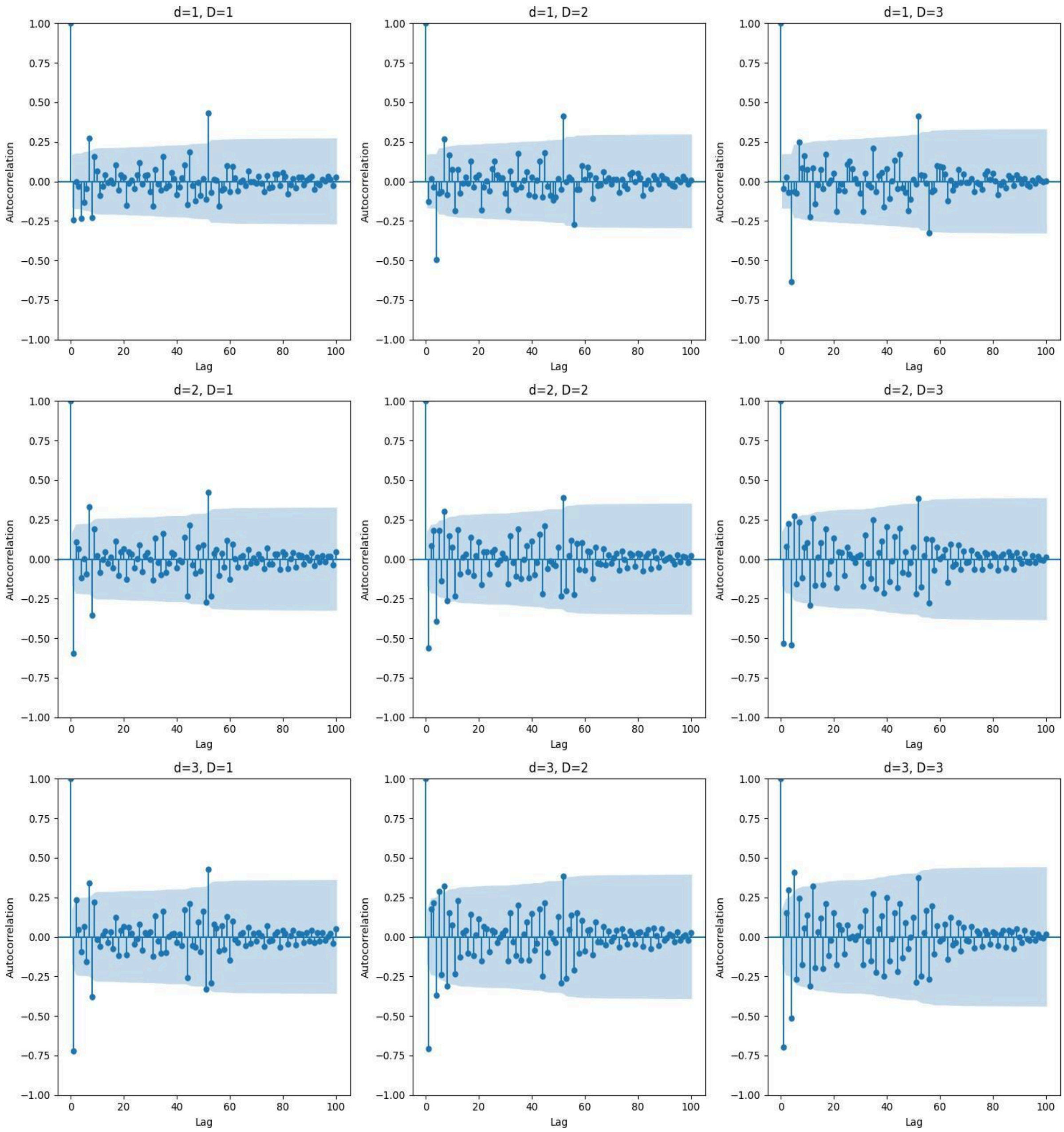
$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_t = \theta_q(B)\Theta_Q(B^s)a_t$$

STEP 2: MAKING THE SERIES STATIONARY

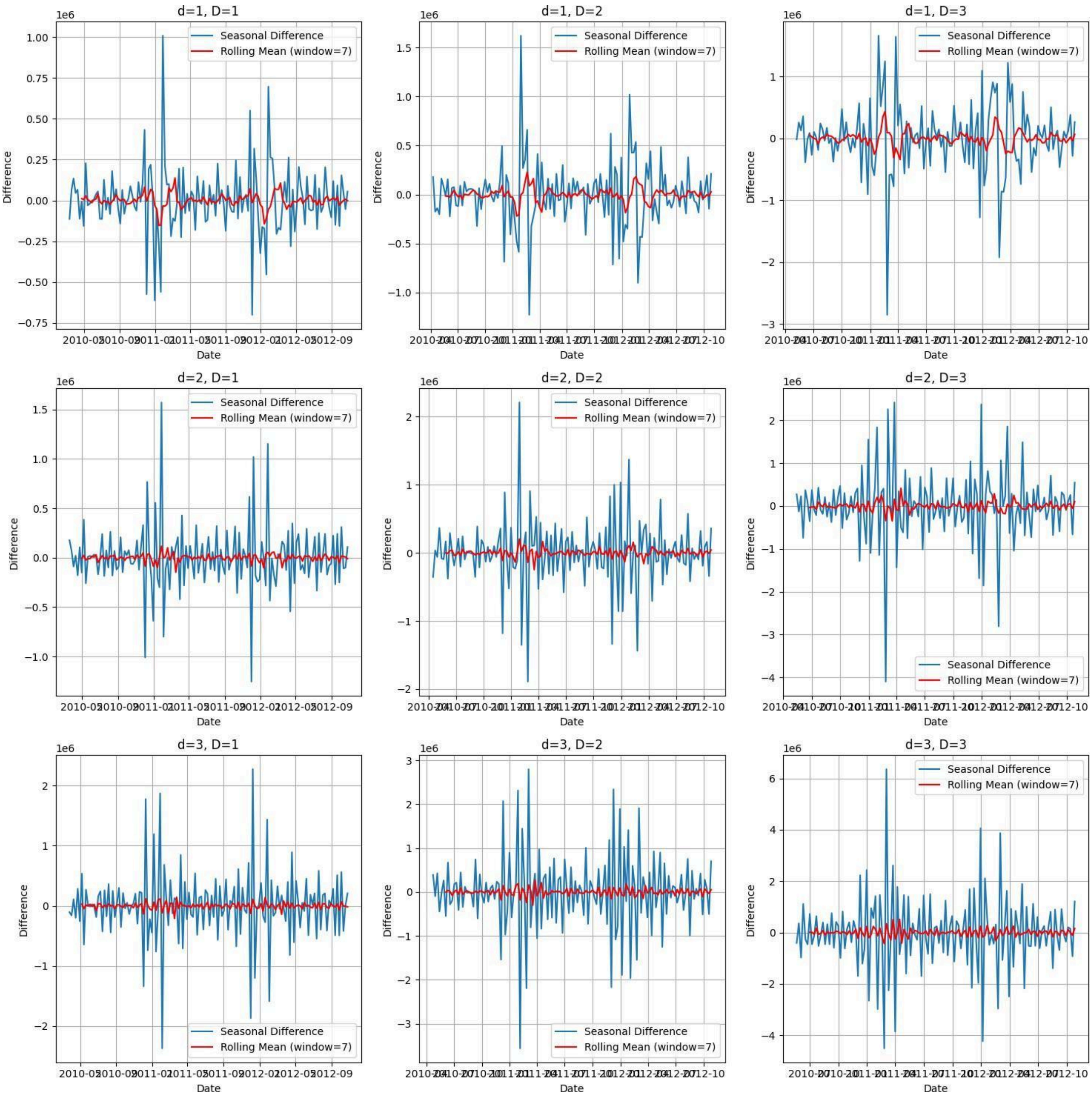
Method: We have already noted that the series has a seasonal component in it hence the series requires seasonal differencing in addition to normal differencing.

So we proceed to take a matrix of values for d and D where they can take values 1,2 or 3. We apply normal and seasonmal differencing to the series using the pair (d,D) and check which pair of differencing make the series closest to stationary.

ACF plot of the differenced series for different values of (d,D):



Moving Average plots of the differenced series:

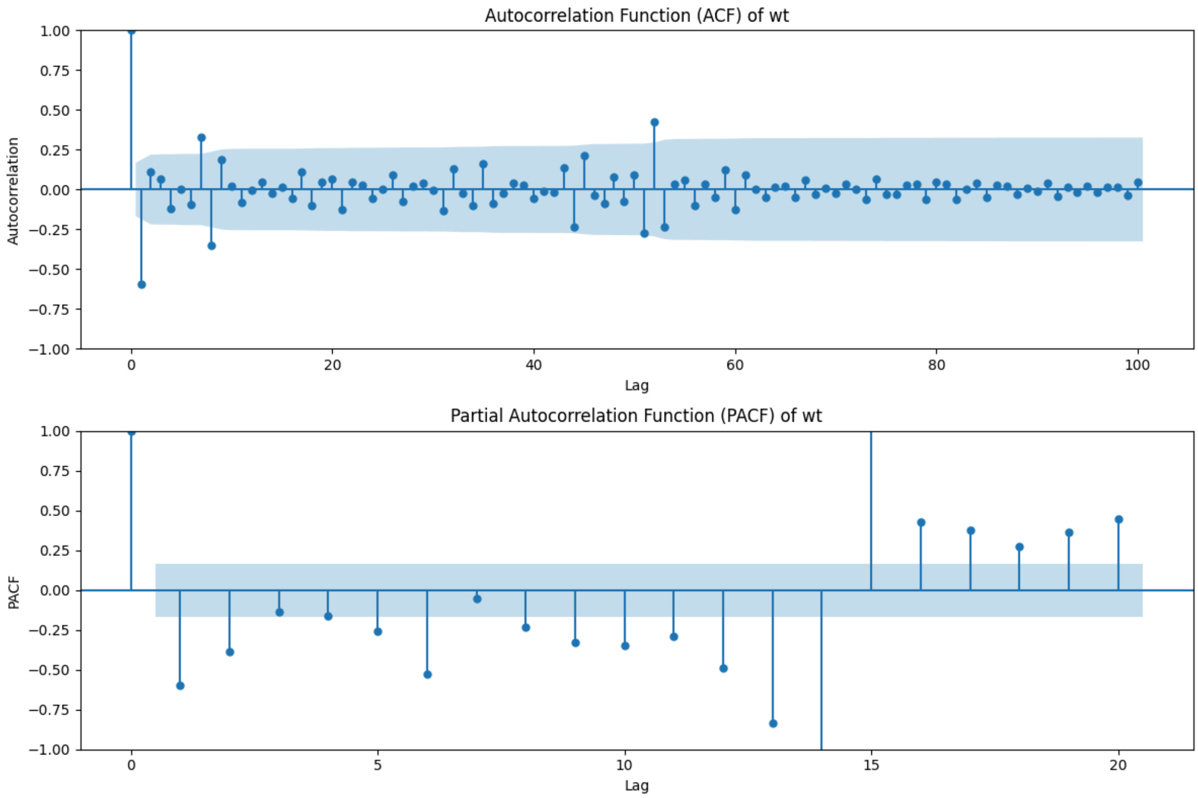


Observations:

- We shortlist the pairs (1,1) and (2,1) for our series on seeing the ACF plots of the differenced series since they show a significant decreasing behaviour in comparision to others.
- The fact that as values of d and D increases model complexity increases too is also kept in mind while choosing the pairs.
- On seeing the Moving Average plots we see that the pair (2,1) shows better result as compared to (1,1) since (1,1) shows higher fluctuations.
- Finally we proceed to choose the pair (2,1) and use this in our final model.

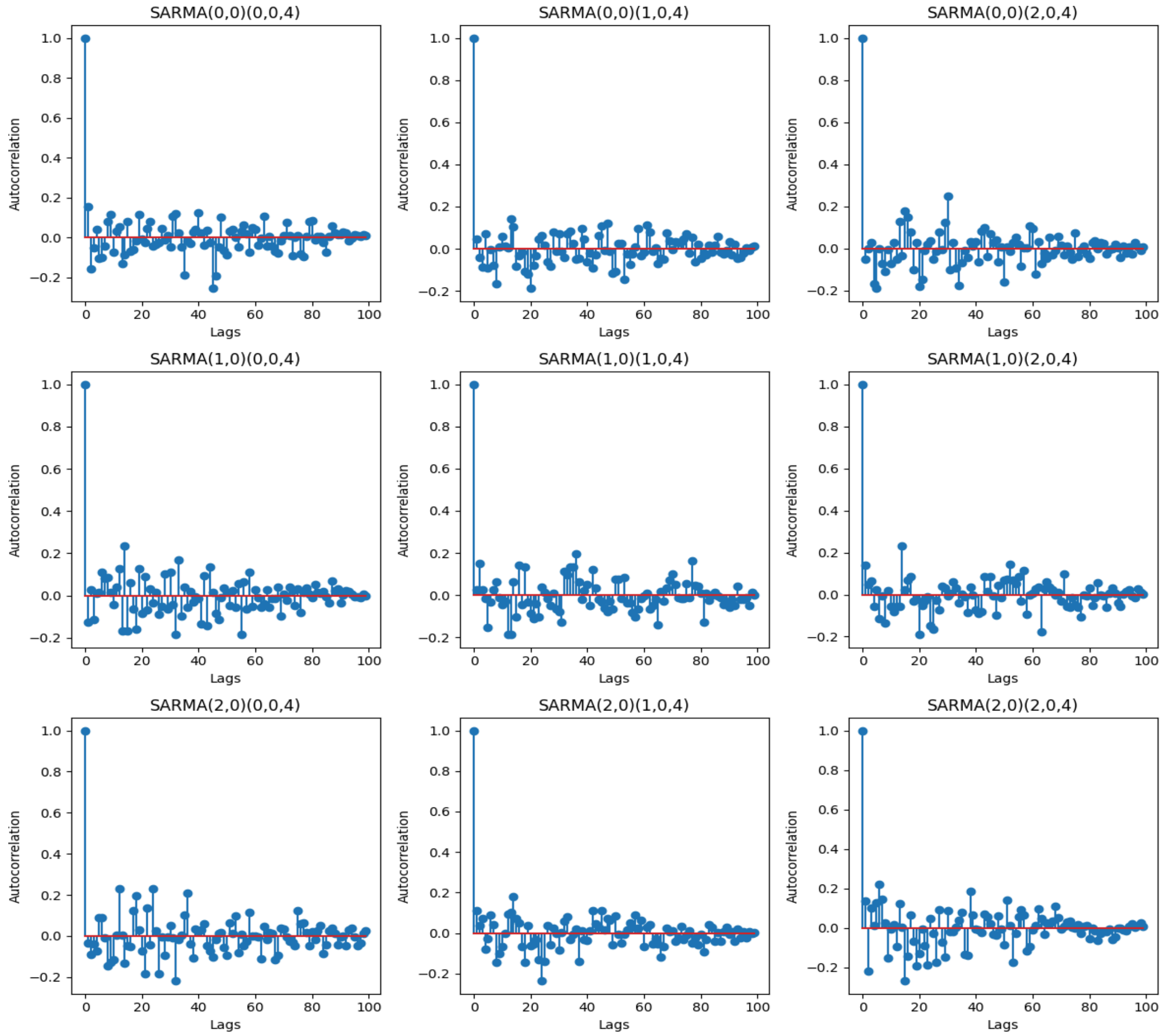
Note: We denote the final differenced series as W_t.

STEP 3: PACF and ACF plot of w_t



STEP 4: Fitting an SARMA model to W_t

At first we tried to compare the theoratical autocorrelation values with the observed aultocorrelation values to find the values of p,P,q,Q. Here is the ACF for some well known SARMA models:



Observaltion:

- Here none of the theoratical ACF matches with the observed ACF.

Note:

- Here we generate the theoratical ACF plots by simulation, using the assumption that random error follows Gaussian distribution
- So it is very tedious job to compare theoratical ACF and observed ACF . So we follow the next method.

STEP 5: Finding values of p,q,P and Q

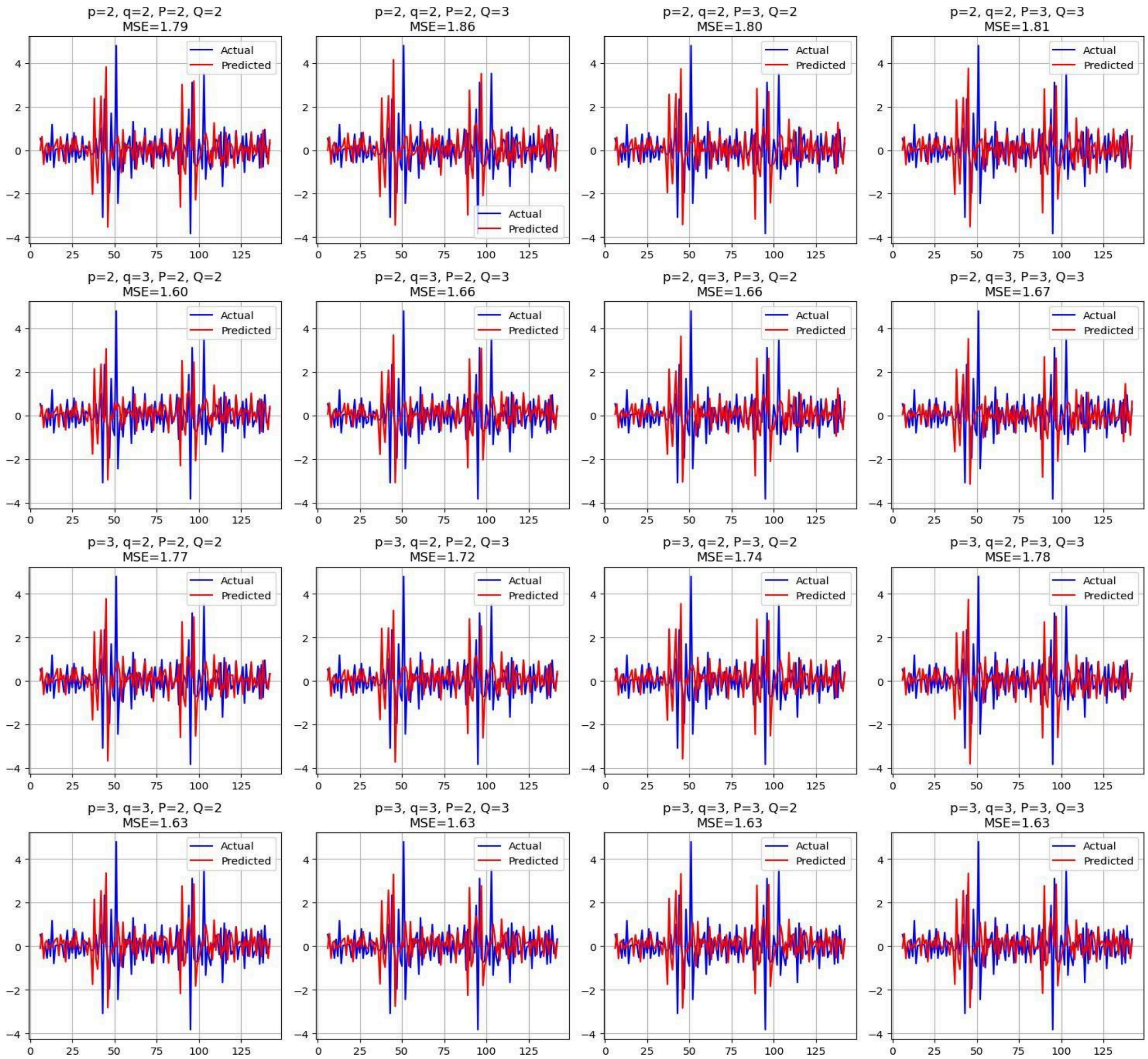
Method: We first consider various tuples of (p,q,P,Q) where p,q,P,Q= 2,3. We fit a SARMA((p,0,q)x(P,0,D,4)) model to W_t for the different tuples. We predict the values of Wt using each model and then overlay the plot of predicted values of W_t over observed value of W_t to check how well do they match.

We further compute the MSE values as a measure of fitness.

Note:

- We first standardised the data before starting our analysis to interpret the graphs more clearly since the sales is in order of 10^6.
- We take the value of s as 4 since we have weekly data and there seems to be monthly seasonality.

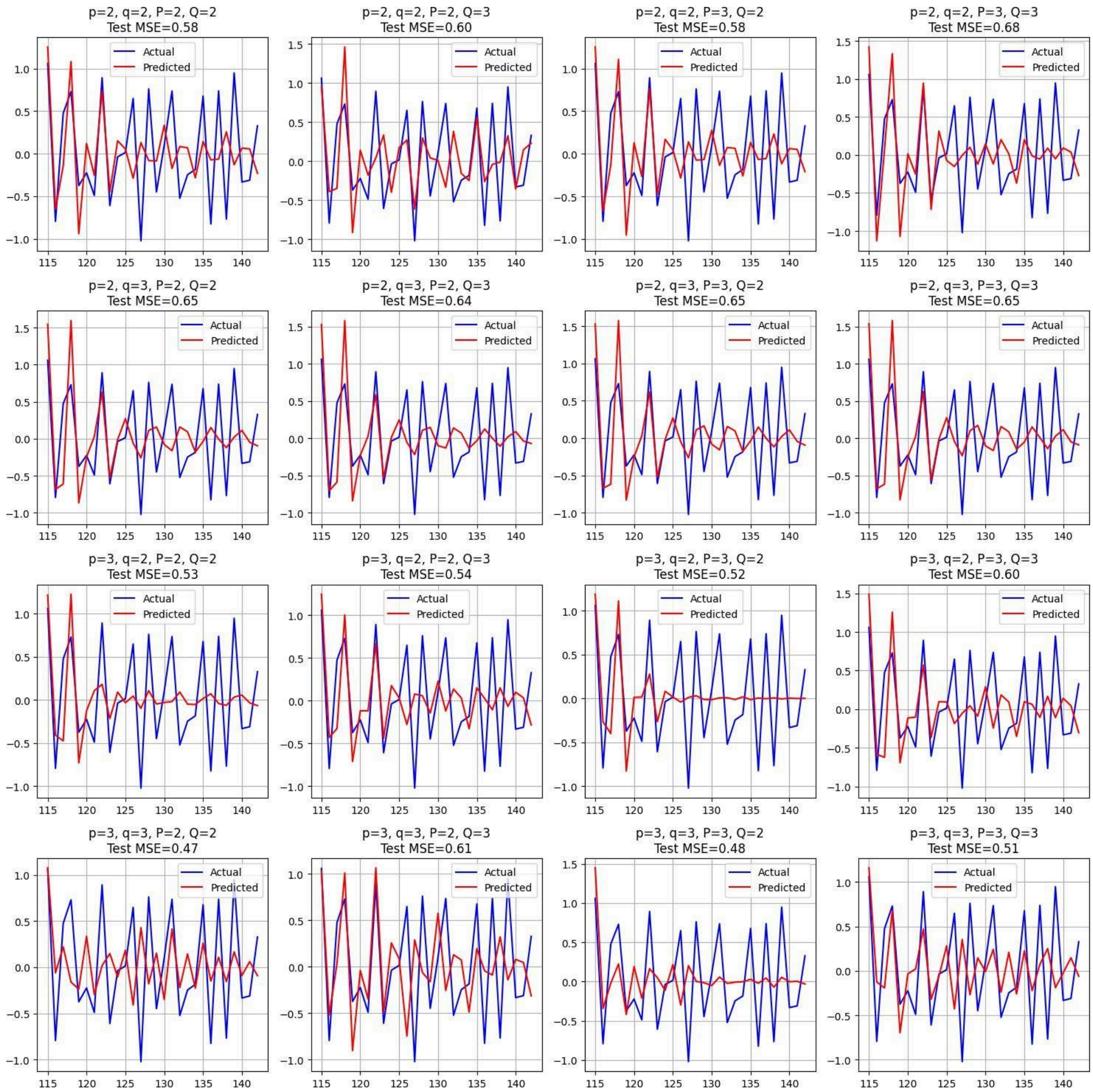
Predicted vs Actual values of W_t:



TRAIN AND TEST SPLITTING:

To study the model better we divide the dataset into training and test data. Then we fit the model using each tuple on the training data and forecast the W_t values for the test period. Then we obverlay the forecasted values with observed test period values and compute the MSE corresponding to test period.

Train: test ratio= 4:1



Observations:

- We shortlist two tuples by combined analysis of graph and test MSE values as follows:
  (3,3,2,2) and (3,3,2,3).
- We denote (3,3,2,2) as model A and (3,3,2,3) as model B.
- We further shortlist our final model after residual analysis.

STEP 6: Model Fitting

Before fitting this model we standardrize W_t.

Model A:

Now the results of the fitting of Model A are as follows.

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                         Weekly_Sales   No. Observations:                 137
Model:             SARIMAX(3, 0, 3)x(2, 0, [1, 2], 4)   Log Likelihood                -111.254
Date:                              Sat, 04 May 2024   AIC                            244.507
Time:                                      02:31:41   BIC                            276.627
Sample:                                           0   HQIC                           257.560
                                              - 137
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.2751      0.129     -9.885      0.000      -1.528      -1.022
ar.L2         -1.1493      0.153     -7.489      0.000      -1.450      -0.849
ar.L3         -0.5539      0.145     -3.820      0.000      -0.838      -0.270
ma.L1          0.0858      0.162      0.531      0.596      -0.231       0.403
ma.L2         -0.0978      0.126     -0.779      0.436      -0.344       0.148
ma.L3         -0.5457      0.160     -3.408      0.001      -0.860      -0.232
ar.S.L4       -1.0754      0.249     -4.319      0.000      -1.563      -0.587
ar.S.L8       -0.1457      0.207     -0.704      0.481      -0.551       0.260
ma.S.L4        0.2528      0.361      0.701      0.483      -0.454       0.959
ma.S.L8       -0.7059      0.264     -2.671      0.008      -1.224      -0.188
sigma2         0.2824      0.040      7.075      0.000       0.204       0.361
===================================================================================
Ljung-Box (L1) (Q):                   1.43   Jarque-Bera (JB):                37.04
Prob(Q):                              0.23   Prob(JB):                         0.00
Heteroskedasticity (H):               0.72   Skew:                            -0.50
Prob(H) (two-sided):                  0.27   Kurtosis:                         5.34
===================================================================================
```

Model B:

Now the results of the fitting of Model A are as follows.

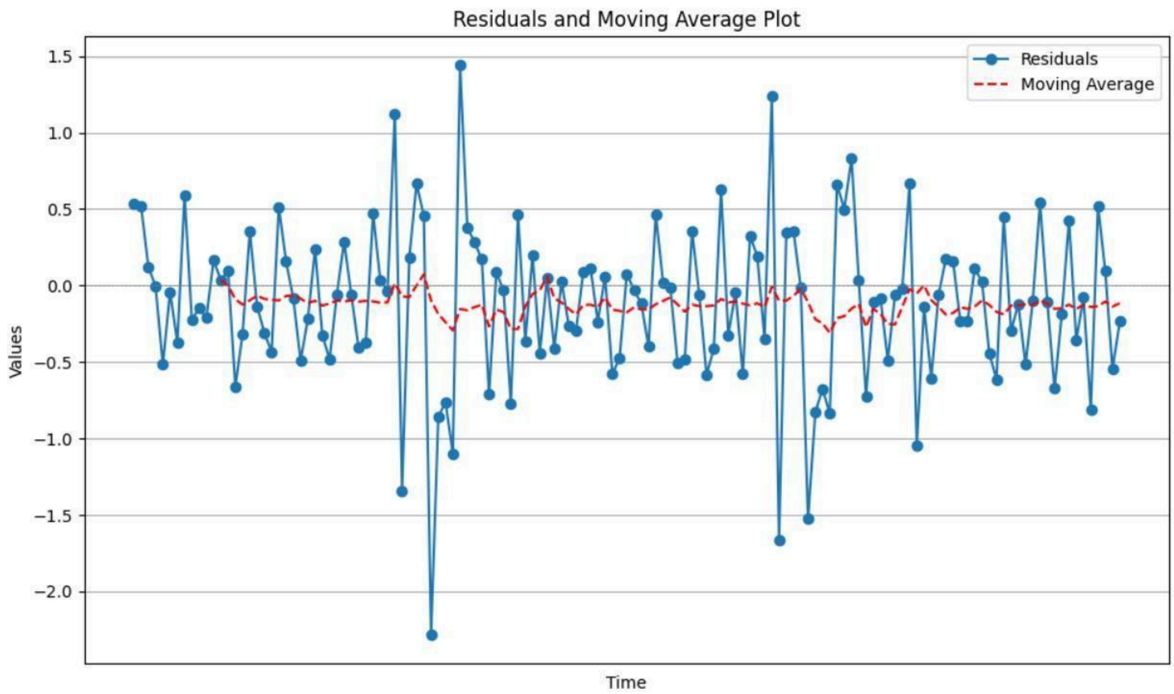
```
                                     SARIMAX Results
=========================================================================================
Dep. Variable:                      Weekly_Sales   No. Observations:                  137
Model:             SARIMAX(3, 0, 3)x(2, 0, 3, 4)   Log Likelihood                -111.609
Date:                           Sat, 04 May 2024   AIC                            247.218
Time:                                   02:32:34   BIC                            282.257
Sample:                                        0   HQIC                           261.457
                                           - 137
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.2923      0.134     -9.609      0.000      -1.556      -1.029
ar.L2         -1.1715      0.153     -7.674      0.000      -1.471      -0.872
ar.L3         -0.5577      0.131     -4.260      0.000      -0.814      -0.301
ma.L1          0.1224      0.166      0.735      0.462      -0.204       0.449
ma.L2         -0.1093      0.121     -0.901      0.368      -0.347       0.128
ma.L3         -0.5788      0.160     -3.615      0.000      -0.893      -0.265
ar.S.L4       -1.3187      0.513     -2.572      0.010      -2.324      -0.314
ar.S.L8       -0.7181      0.491     -1.461      0.144      -1.681       0.245
ma.S.L4        0.4997      0.639      0.782      0.434      -0.753       1.752
ma.S.L8       -0.3088      0.351     -0.880      0.379      -0.997       0.379
ma.S.L12      -0.4698      0.444     -1.059      0.290      -1.340       0.400
sigma2         0.2813      0.031      9.063      0.000       0.220       0.342
===================================================================================
Ljung-Box (L1) (Q):                   0.98   Jarque-Bera (JB):                29.70
Prob(Q):                              0.32   Prob(JB):                         0.00
Heteroskedasticity (H):               0.74   Skew:                            -0.38
Prob(H) (two-sided):                  0.32   Kurtosis:                         5.15
===================================================================================
```

STEP 6: Residual Analysis

Model A:

At first we will check $E[a_t] = 0$ or not. So observe the following plot.

Residual Plot for Model A



Observation:

- Here the moving average line is fluctuating around zero with a low standard deviation. Hence we can conclude that $E[a_t] = 0$

Now we will conduct **Ljung-Box test.**

Ho : Our model is adequate H1: Model is not adequate.

Test Statistic:

$$\tilde{Q} = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}r_k^2(\hat{a})$$

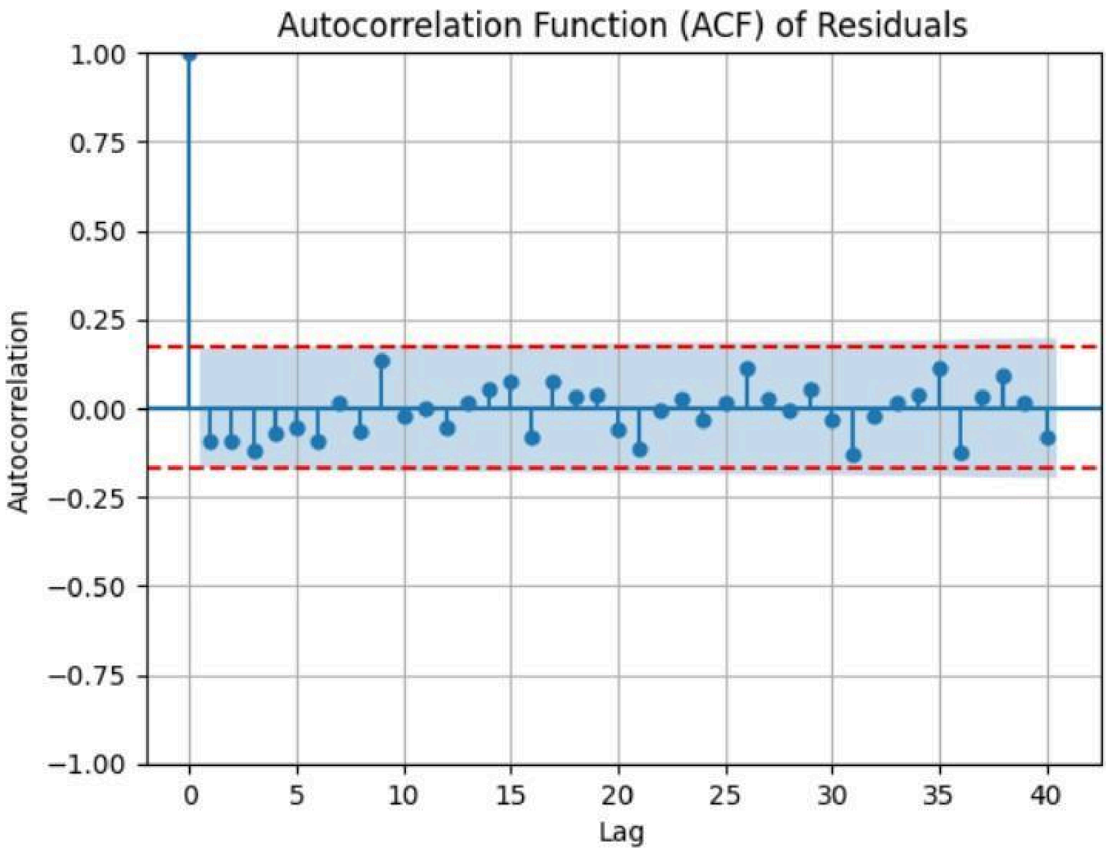
Under H0 this follows $X^2(K-p-q)$

Here the value of the test statistic is 93.823 and the corresponding p value is 0.03<0.05

So model A is adequate for fitting.

Now we will observe autocorrelation plot of a_t

Autocorrelation plot of a_t



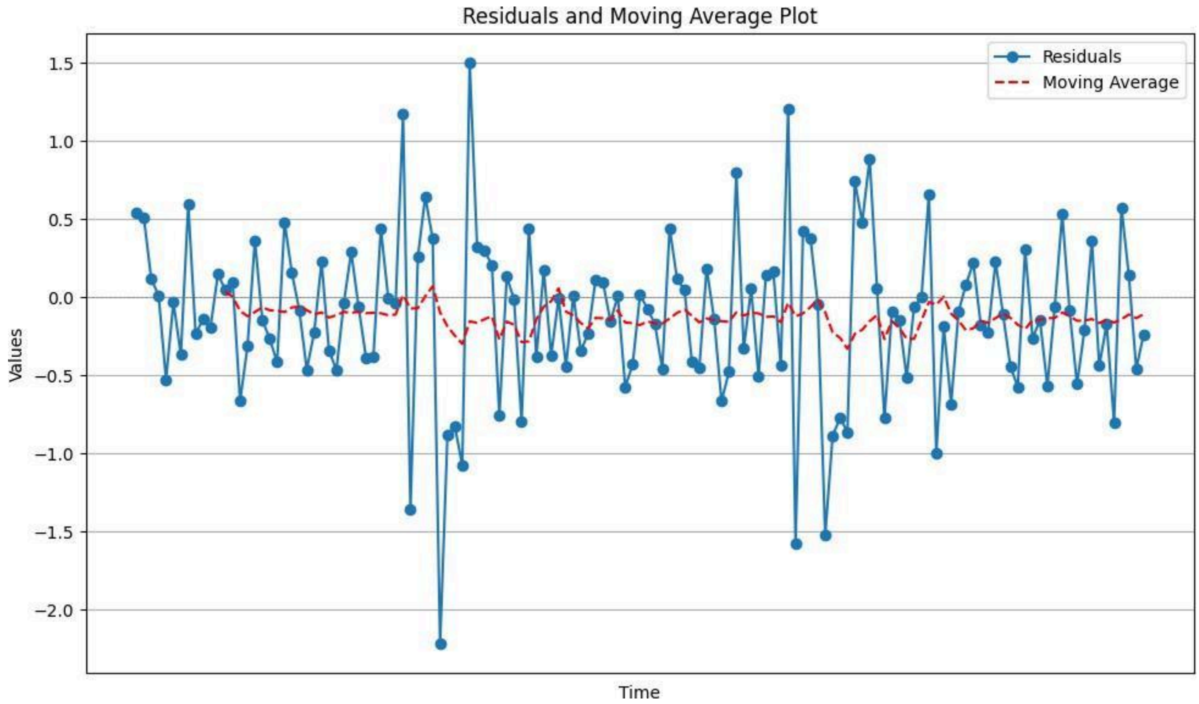
Observation:

- Here all the autocorrelations lie within -2/n and 2/n . Hence a_t ‘s are independent

Model B:

At first we will check $E[a_t] = 0$ or not. So observe the following plot.

Residual Plot for Model B



Observation:

- Here the moving average line is fluctuating around zero with a low standard deviation. Hence we can conclude that $E[a_t] = 0$

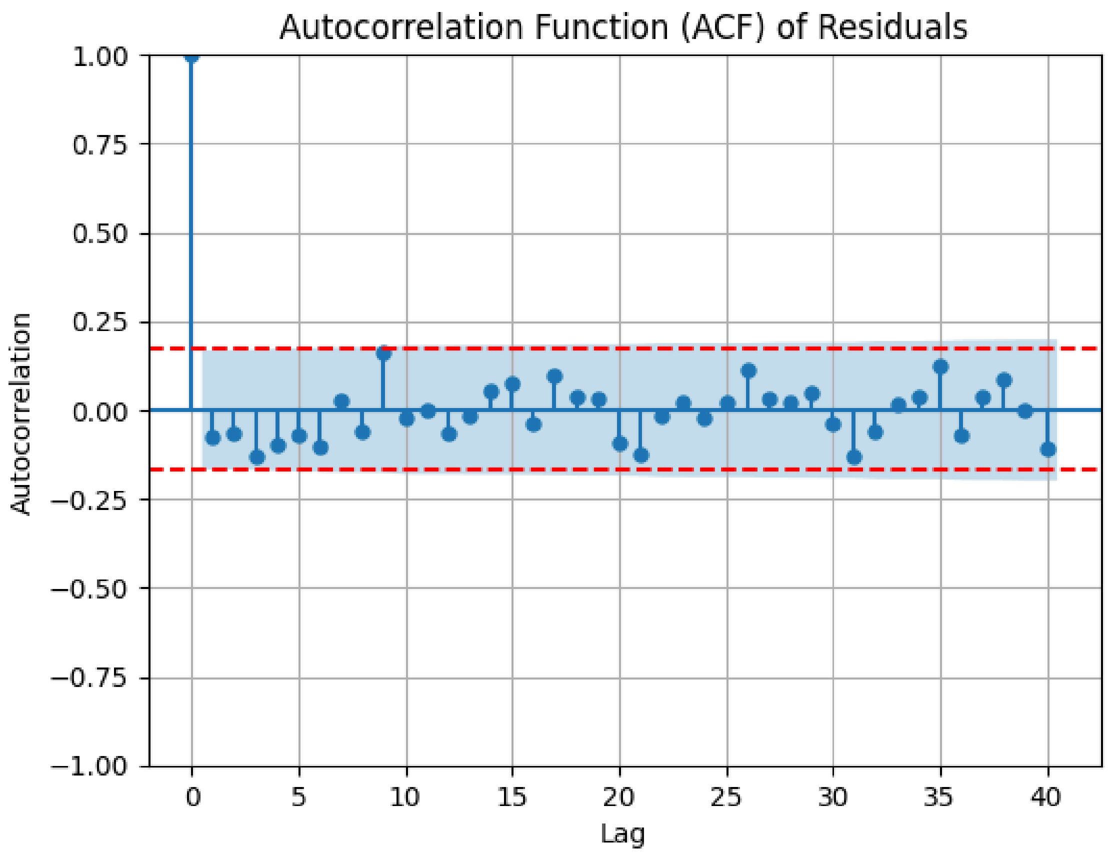
Now we will conduct **Ljung-Box test.**

Ho : Our model is adequate H1: Model is not adequate.

Here the value of the test statistic is 96.83 and the corresponding p value is 0.0186<0.05

So model B is adequate for fitting.

Now we will observe autocorrelation plot of a_t



Observation:

- Here all the autocorrelations lie within -2/n and 2/n . Hence a_t ‘s are independent

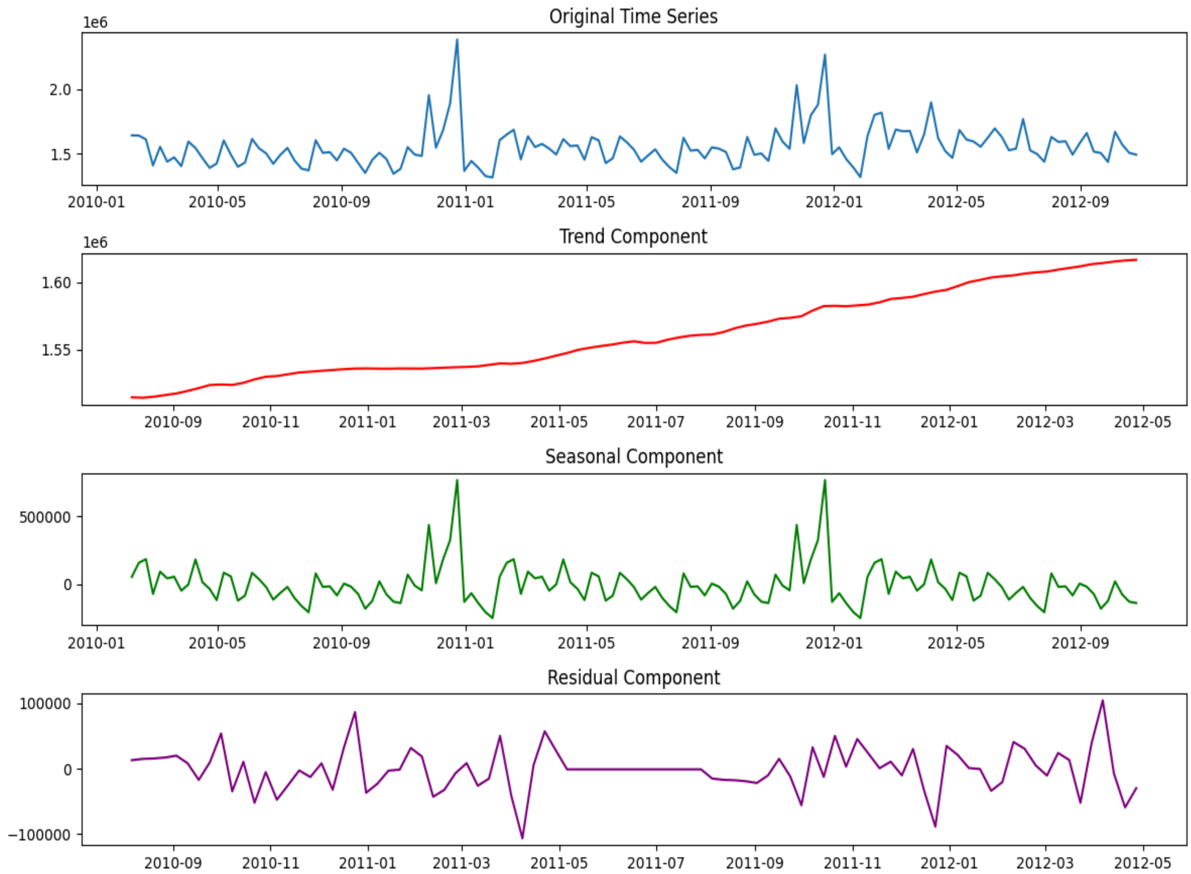
STEP 7 : Time Series Components

Here we consider the addive model

yt=St+Tt+Rt, where yt is the data, St is the seasonal component, Tt is the trend-cycle component, and Rt is the remainder component, all at period t.

Here is the result of our time series decomposition

Time Series Decomposition:



Conclusion:

- Here both the models are good as both the models satisfies the basic assumptions of related to errors.
- But the quality of forecast may increase if we include the all the explanatory variables in our model.