Sales analysis and forcasting using ARIMA

Members:

Objective

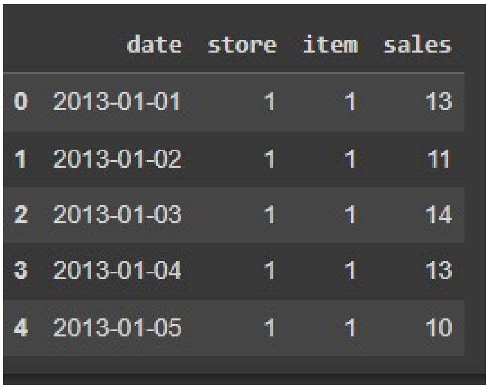
Employ Autoregressive Integrated Moving Average (ARIMA) to analyze 5 years of sales data for a store item. This can help us predict sales for the next 3 months more accurately. This information can be valuable for making smart business decisions.

Outline of Notebook

- Dataset
- Exploratory Data Analysis
- Data preprocessing
- Model Building
- Forecasting
- Result Analysis and Conclusion

Dataset

We have used one of the publicly available inventory sales datasets for our project. The dataset contains 5 years of store-item sales data containing date, store id, item id, and sales as features. For this project, we have analyzed sales for a particular item in a specific store.

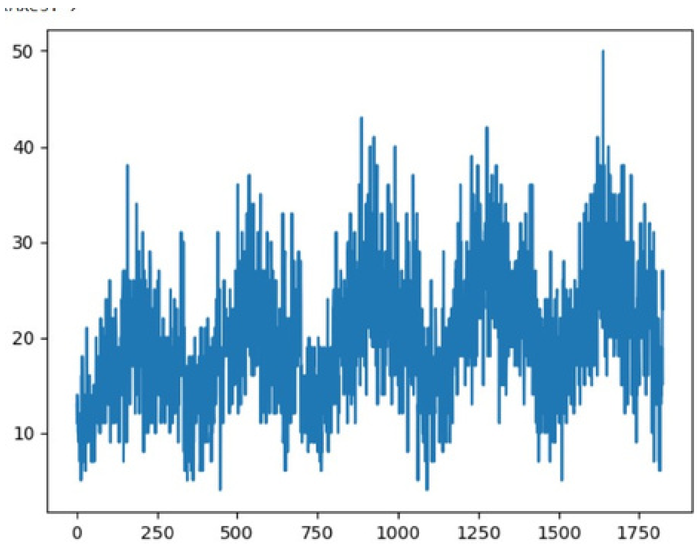


	date	store	item	sales
0	2013-01-01	1	1	13
1	2013-01-02	1	1	11
2	2013-01-03	1	1	14
3	2013-01-04	1	1	13
4	2013-01-05	1	1	10

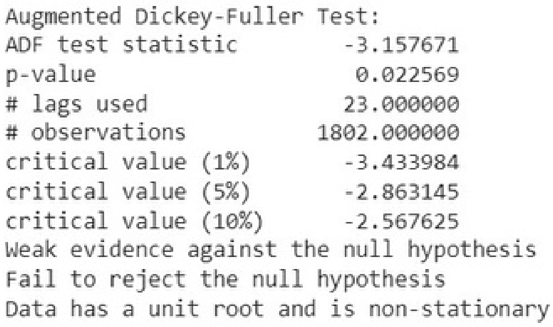
Exploratory Data Analysis

Checked whether the dataset contains missing values but none were found.

Plotted the sales data with time stamps.

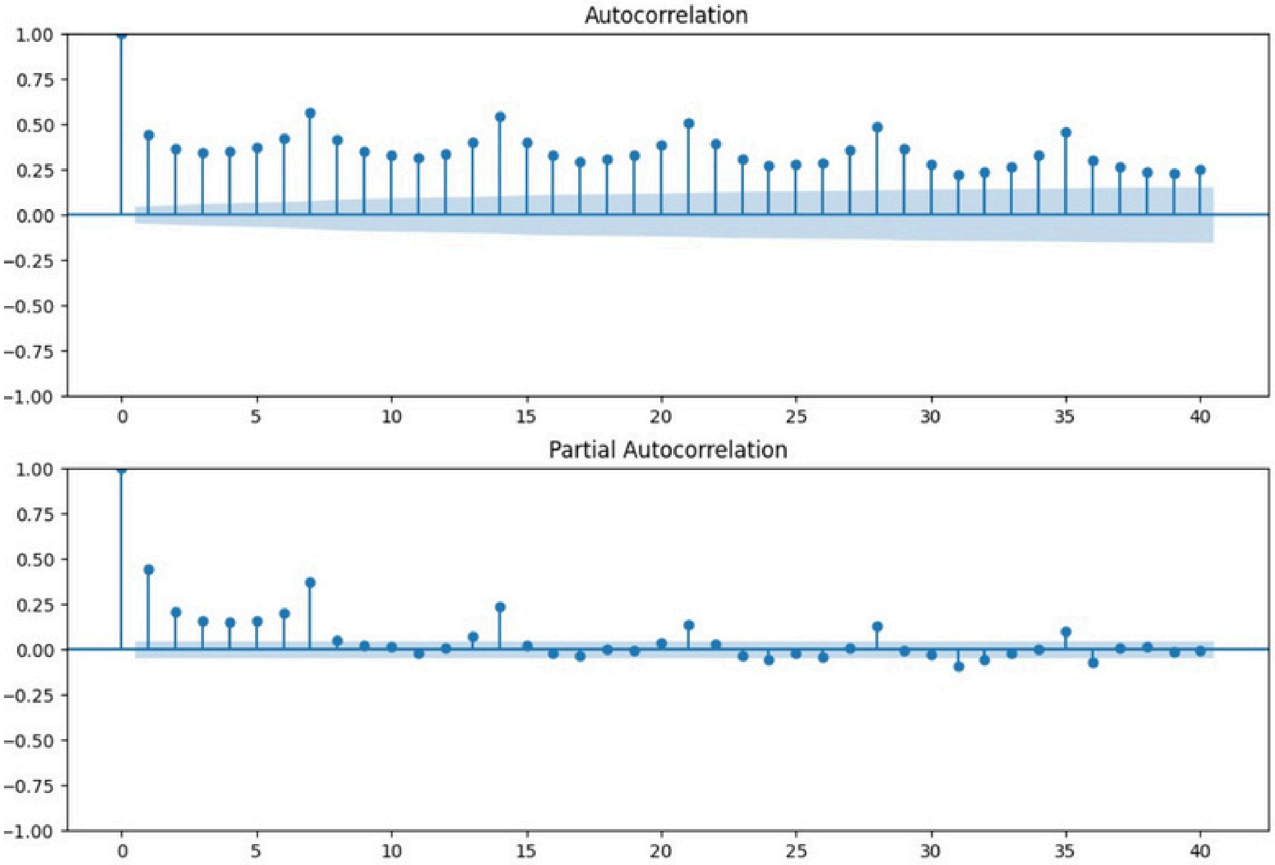


The above plot clearly shows data is non stationary, as both the mean and variance appear to be dependent on time. To confirm our observations we go on to perform Augmented Dickey-Fuller Test.



```
Augmented Dickey-Fuller Test:
ADF test statistic        -3.157671
p-value                    0.022569
# lags used               23.000000
# observations          1802.000000
critical value (1%)       -3.433984
critical value (5%)       -2.863145
critical value (10%)      -2.567625
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

The smaller p-value, the more likely it's stationary. Here our p-value is 0.022. It's actually not bad, if we use a 5% Critical Value(CV), this series would be considered stationary. But as we just visually found an upward trend, we want to be more strict, we use 1% CV.


Autocorrelation

Partial Autocorrelation

The trends in ACF and PACF plot shows that the data is not stationary.

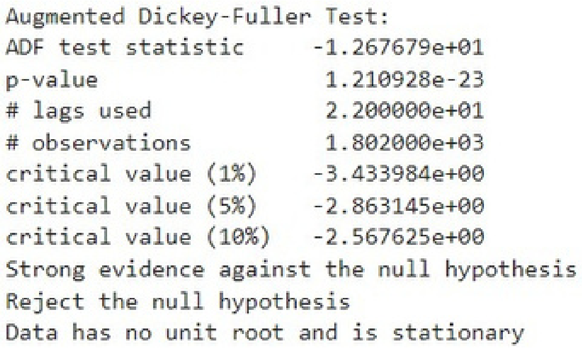
Data preprocessing

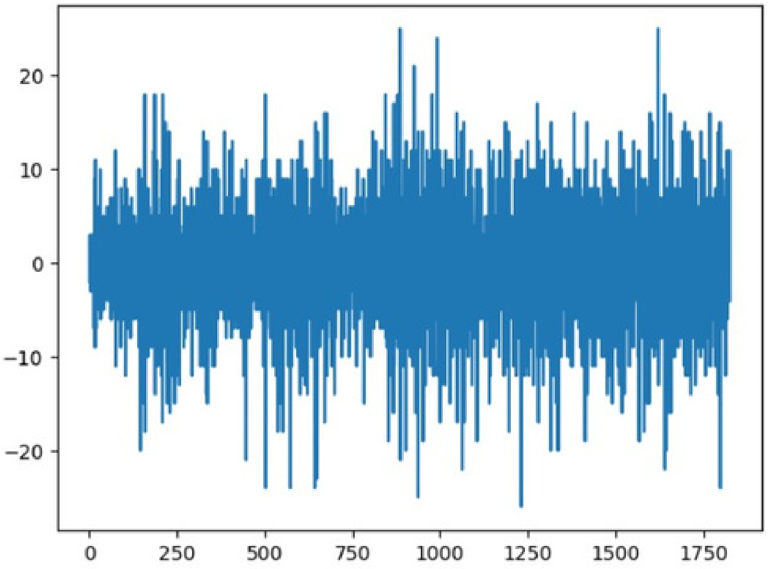
To make the data stationary we perform the first difference

(wherein we subtracted the current value with its first lag value.)

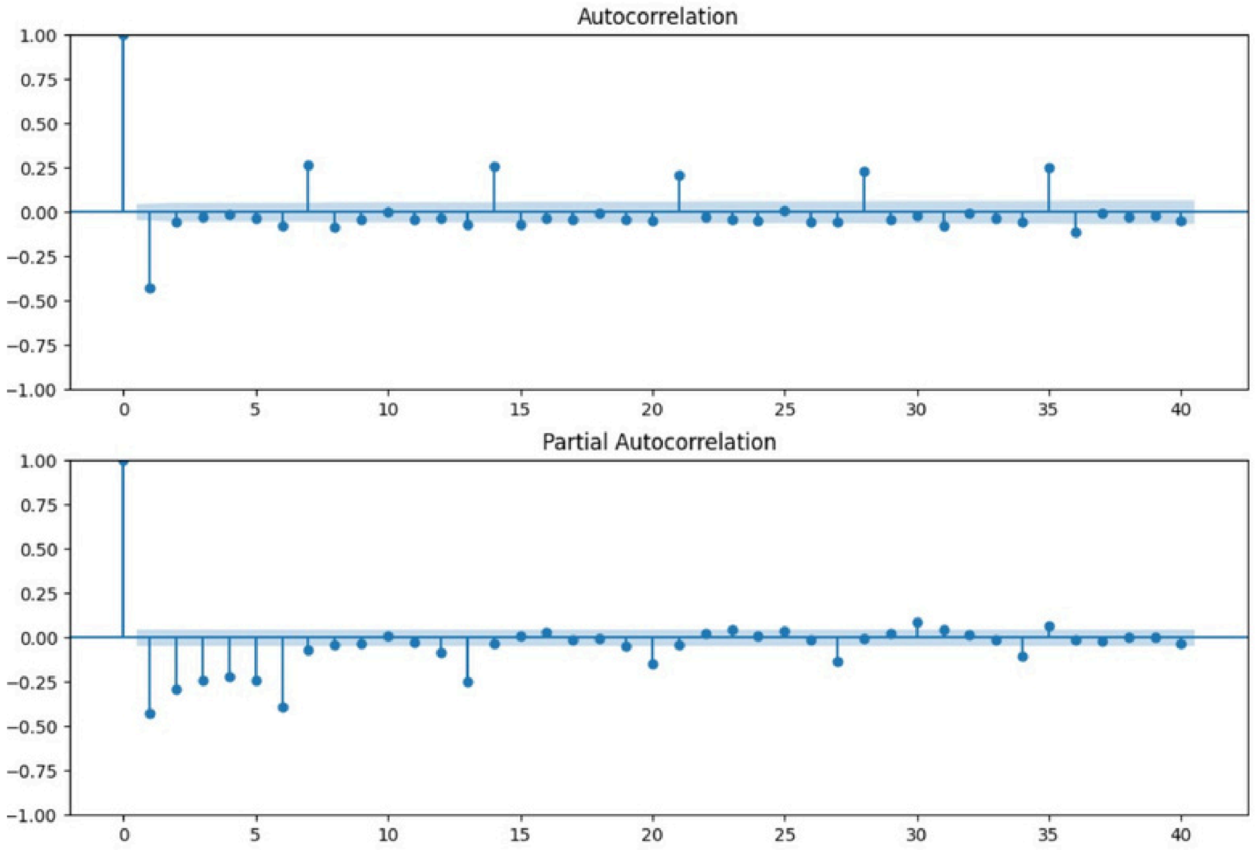
$$\nabla x_t = x_t - x_{t-1}.$$

Now, we perform ADF Test and plot the ACF and PACF.


```
Augmented Dickey-Fuller Test:
ADF test statistic     -1.267679e+01
p-value                 1.210928e-23
# lags used             2.200000e+01
# observations          1.802000e+03
critical value (1%)    -3.433984e+00
critical value (5%)    -2.863145e+00
critical value (10%)   -2.567625e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```



After first difference, the dataset becomes stationary which is evident by the small p-value and the plot of the data.



Here we can see the acf and pacf both has a recurring pattern every 7 periods. Indicating a weekly pattern exists. Any time you see a regular pattern like that in one of these plots, you should suspect that there is some sort of significant seasonal effect.

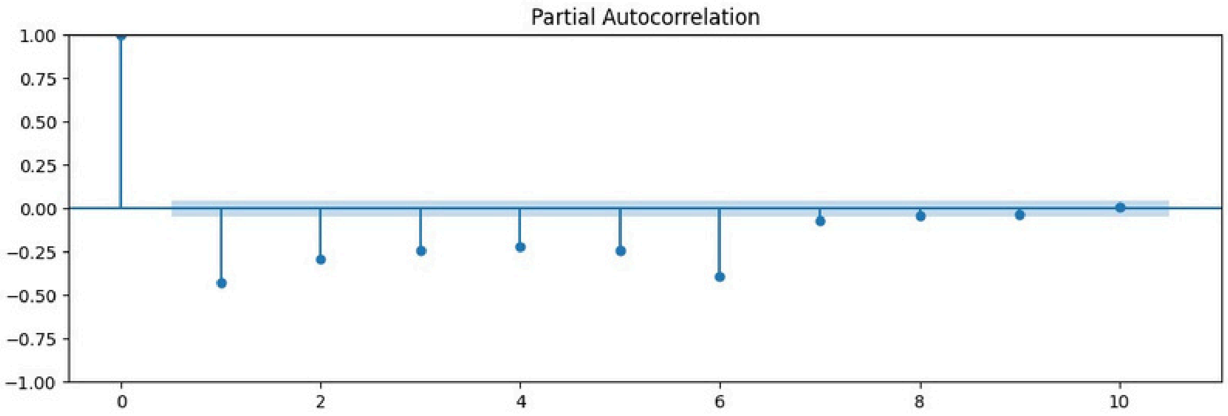
Model building

We broke down the ARIMA Model into simple steps:

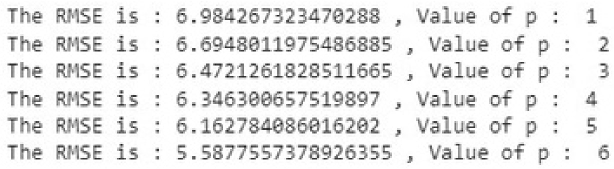
1 . Making the data stationary by differencing. (I)

This was already performed in the Data preprocessing part.

2. Fitting an AR model . (AR)



From the PACF plot, we can see significant spikes at lags 1 to 6 because of the significant PACF value. In contrast, for everything within the blue band, we don't have evidence that it's different from zero.

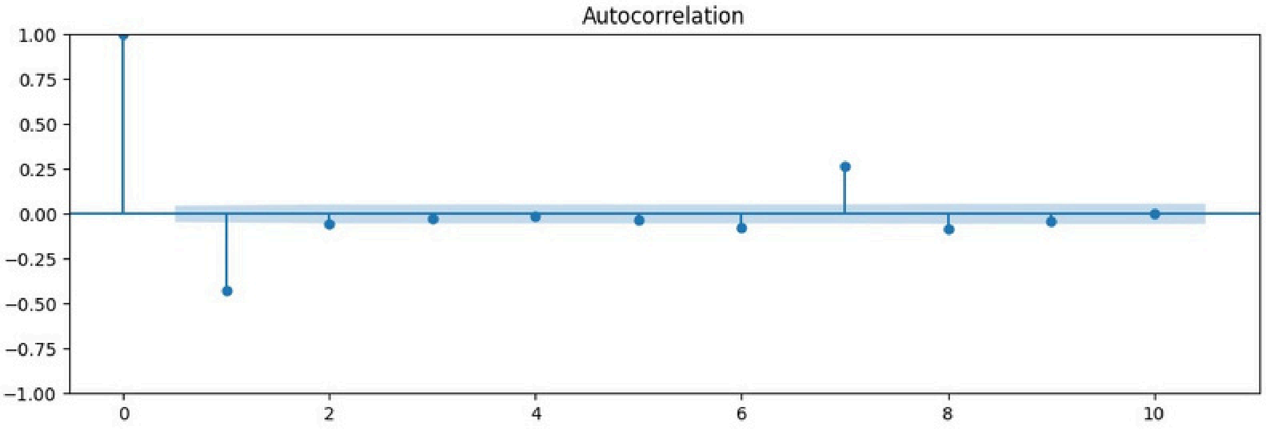


```
The RMSE is : 6.984267323470288 , Value of p :  1
The RMSE is : 6.6948011975486885 , Value of p :  2
The RMSE is : 6.4721261828511665 , Value of p :  3
The RMSE is : 6.346300657519897 , Value of p :  4
The RMSE is : 6.162784086016202 , Value of p :  5
The RMSE is : 5.5877557378926355 , Value of p :  6
```

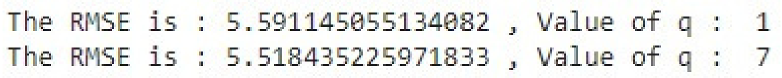
It was found that Root Mean Square Error (RMSE) is less in p=6, hence we chose AR(6)

3. Fitting an MA model on the residuals. (MA)

We generated residuals by the difference of predictions from AR(6) and the original data.



From the ACF plot, we analyzed all spikes higher than the blue area, i.e. q=7 and q=1

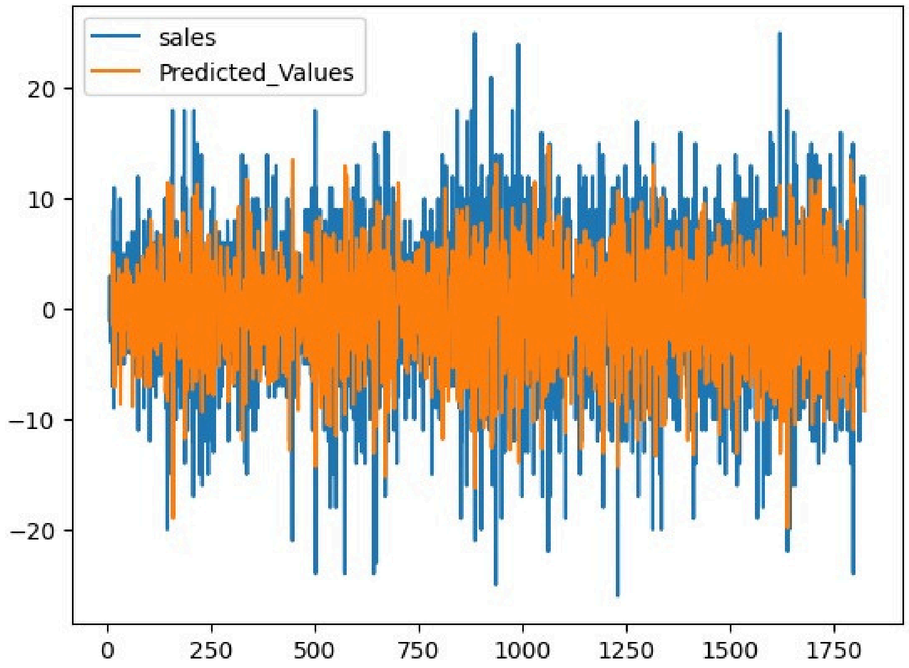


```
The RMSE is : 5.591145055134082 , Value of q :  1
The RMSE is : 5.518435225971833 , Value of q :  7
```

Since RMSE for q=7 is lower, we chose MA(7).

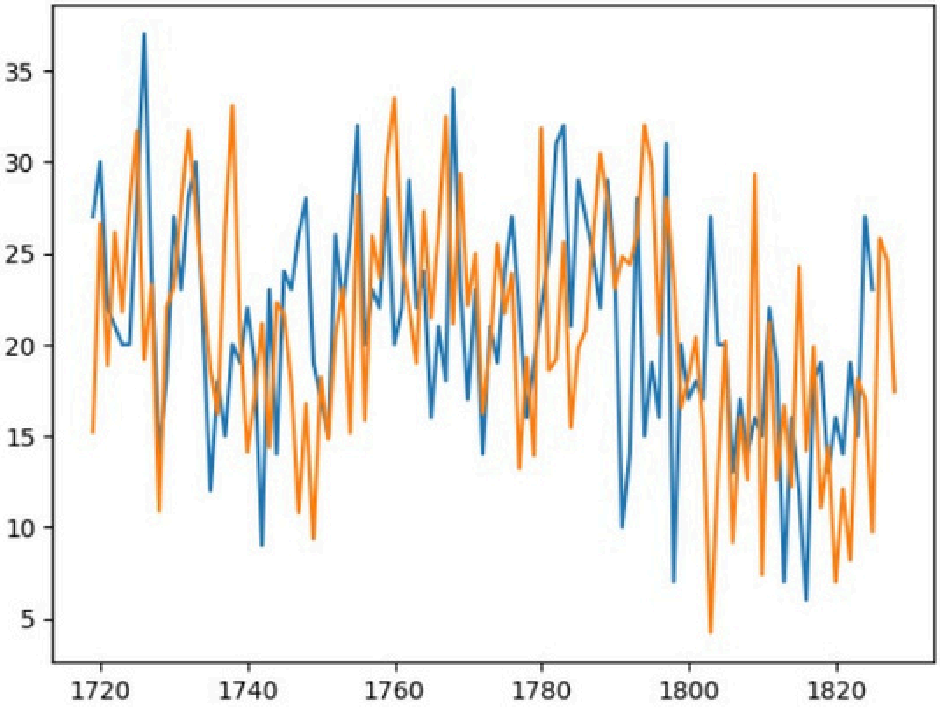
4. Getting Back the Original data

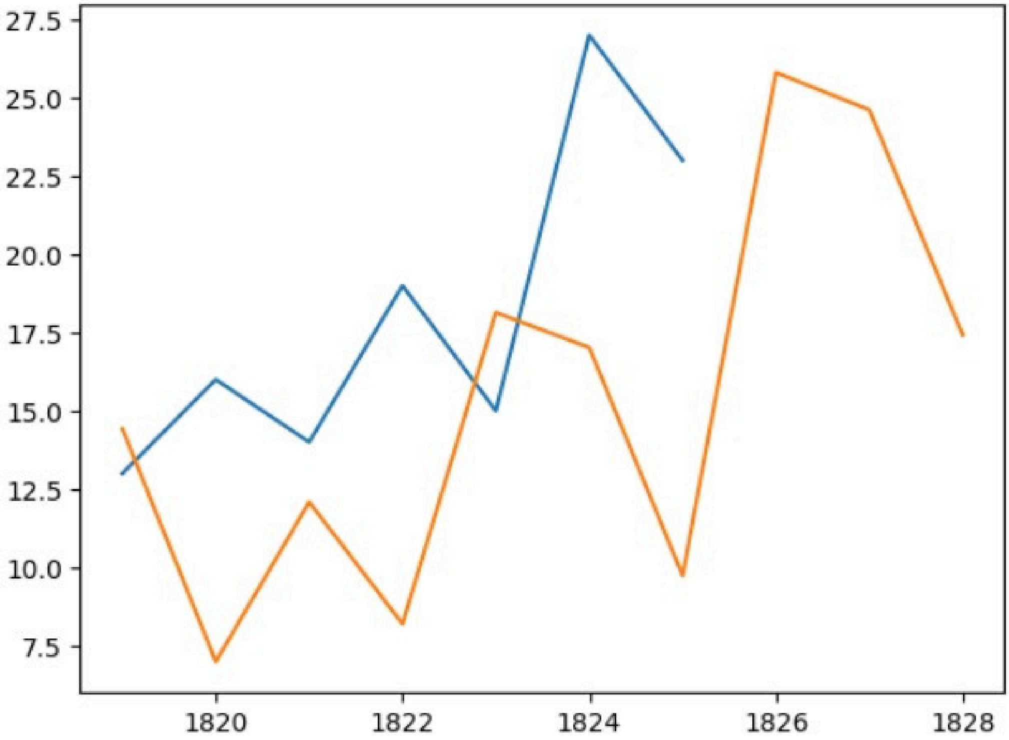
Now we added the residuals obtained from MA(7) to the prediction of AR(6). We reversed the steps performed for differencing.



Forecasting

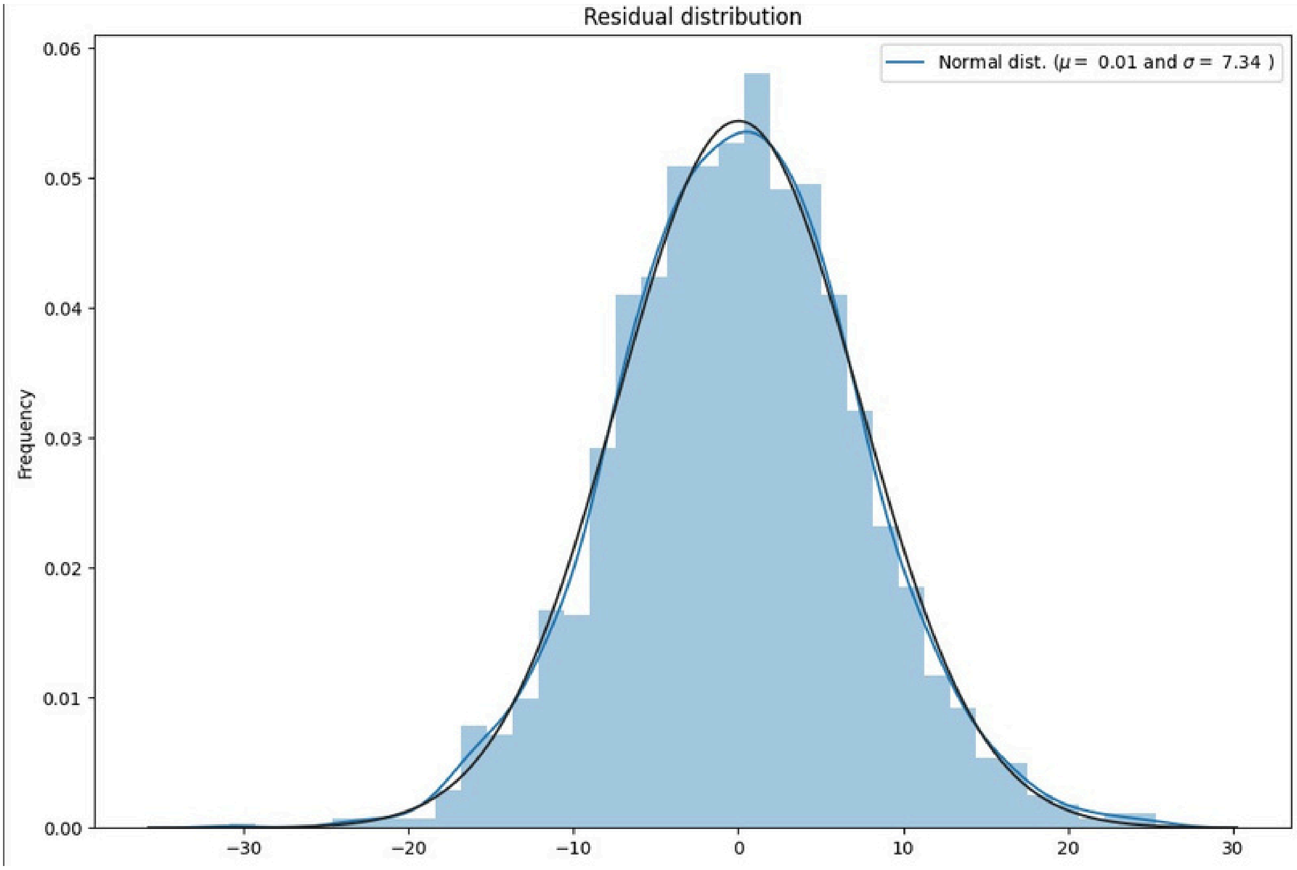
We have used the model proposed above for 3 months of future inventory sales forecasting. The graphs below gives an idea about the future predictions.



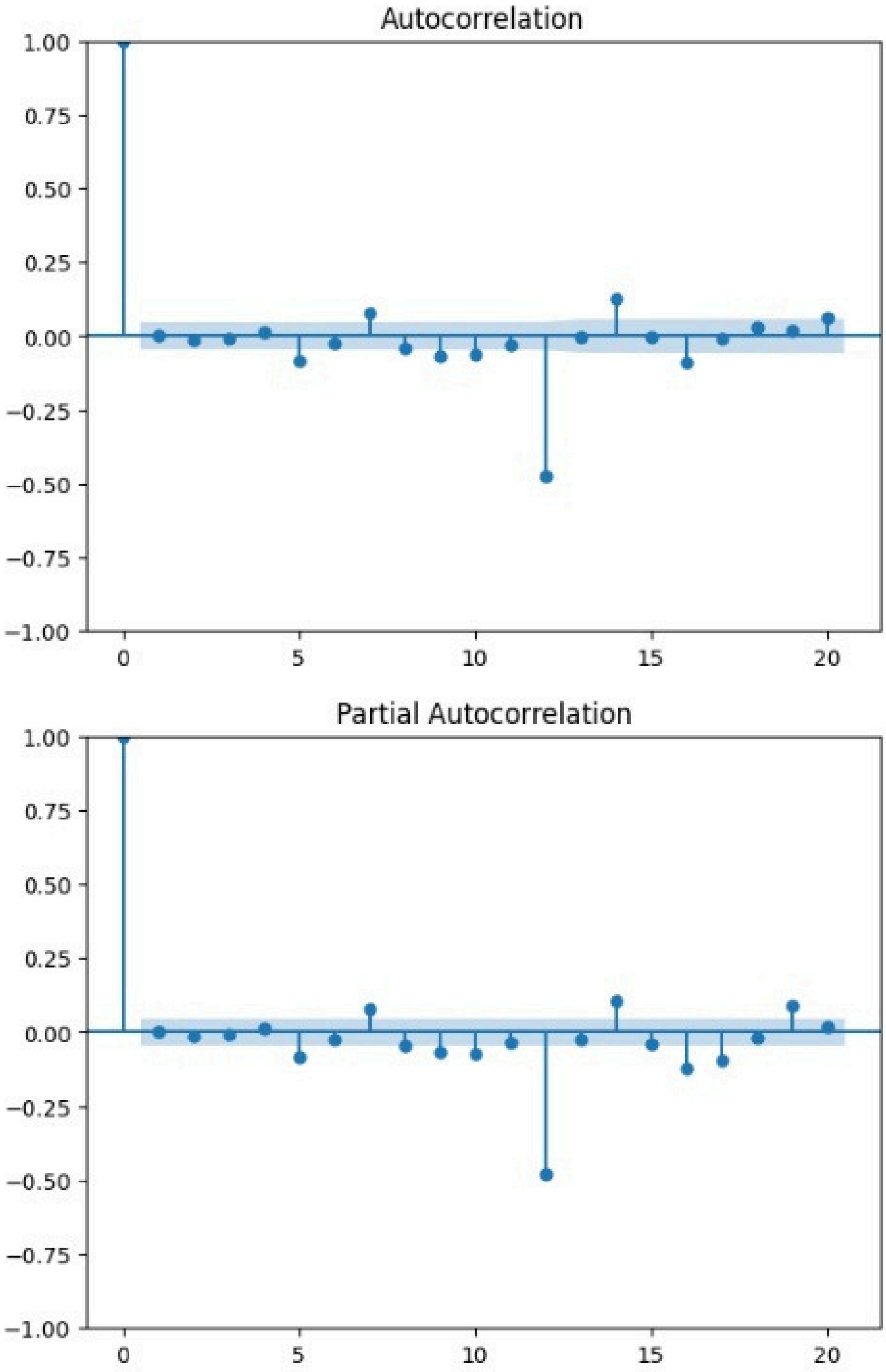


Result Analysis and Conclusion

To analyze our model, we can plot the residual distribution. As we know that if the residual distribution is normal and equivalent to white noise, the model is a good fit. Also, the ACF and PACF should not have significant terms. Ensuring that the residual is iid.



The plot appears to be somewhat normal distribution but is clearly not.



But here we can clearly see that a recurring correlation exists in both ACF and PACF. So we need to incorporate seasonality to make our model more robust and accurate.