RTSM PROJECT:

Pratham Kadam	**22IE3FP35**
Sushant Jha	**22CH3FP09**
Mayank Sony	**22CH3FP28**
Manish Vaghmashi	**22CH3FP04**
Mayur Tank	**22MT3FP26**
Harshit Pathak	**22IE10026**

AMERICAN EXPRESS STOCK PRICE PREDUTION

Objective

Employ Autoregressive Integrated Moving Average (ARIMA) to analyze 4 years of AMEX daily closing price data. This can help us more accurately predict the AMEX stock price for the next 12 months. This information can be valuable for making smart investment decisions.

Outline of the Notebook

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

About the data

The AMEX Bank dataset contains daily Open, High, Low, Close, Adjusted Close, and Volume data for the past 4 years. The dataset was obtained from the Yahoo finance website.

```
df.head()
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2020-01-01	1276.099976	1280.000000	1270.599976	1278.599976	1244.189819	1836849
1	2020-01-02	1279.000000	1288.000000	1279.000000	1286.750000	1252.120483	3068583
2	2020-01-03	1282.199951	1285.000000	1263.599976	1268.400024	1234.264282	5427775
3	2020-01-06	1260.000000	1261.800049	1236.000000	1240.949951	1207.552856	5445093
4	2020-01-07	1258.900024	1271.449951	1252.250000	1260.599976	1226.674072	7362247

```
df.tail()
```

	Date	Open	High	Low	Close	Adj Close	Volume
1054	2024-04-03	1472.099976	1495.650024	1471.400024	1482.300049	1482.300049	22792193
1055	2024-04-04	1504.000000	1530.000000	1504.000000	1527.599976	1527.599976	44467533
1056	2024-04-05	1539.000000	1554.500000	1530.150024	1549.550049	1549.550049	29527951
1057	2024-04-08	1554.949951	1557.250000	1541.550049	1546.599976	1546.599976	10241470
1058	2024-04-09	1554.849976	1554.849976	1540.300049	1548.550049	1548.550049	10942247

The data ranges from 1st January 2020 upto 9th April 2024

```
df.shape
```

```
(1059, 6)
```

The data frame contains 1059 rows and 6 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1059 entries, 2020-01-01 to 2024-04-09
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Open       1059 non-null   float64
 1   High       1059 non-null   float64
 2   Low        1059 non-null   float64
 3   Close      1059 non-null   float64
 4   Adj Close  1059 non-null   float64
 5   Volume     1059 non-null   int64
dtypes: float64(5), int64(1)
memory usage: 57.9 KB
```

All the columns are numerical and non-null, as there as 1059 non-null values.

```
df.describe()
```

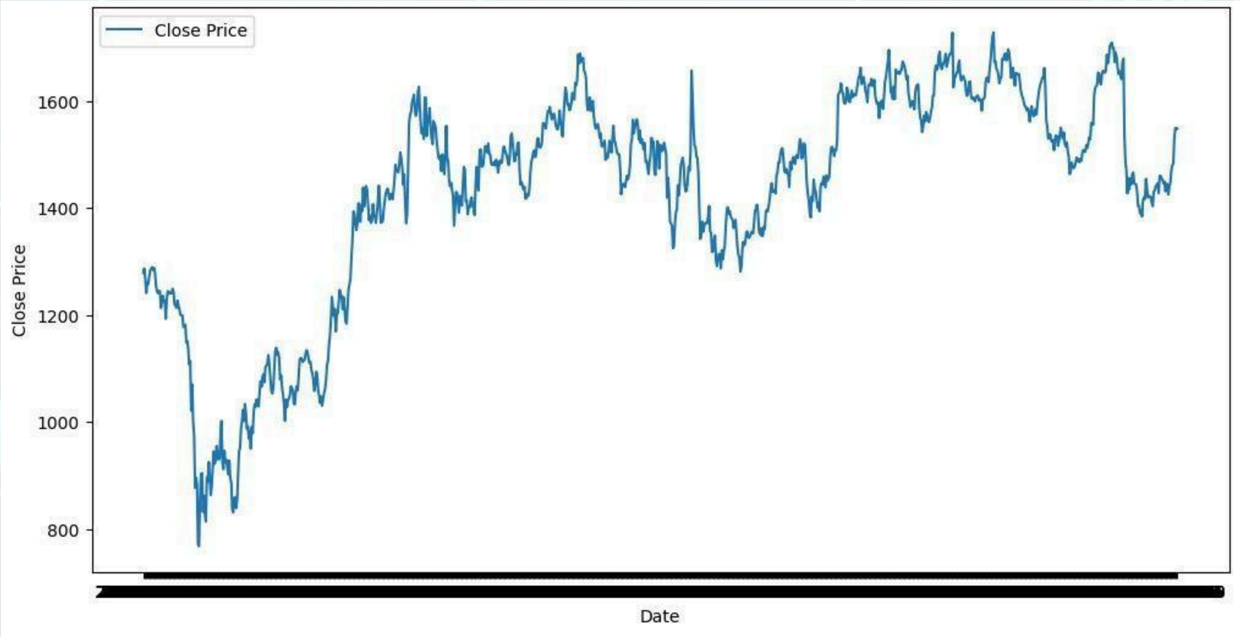
	Open	High	Low	Close	Adj Close	Volume
count	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000	1.059000e+03
mean	1430.502030	1444.242869	1415.737585	1430.071202	1407.335192	1.229766e+07
std	204.900963	203.603692	207.292238	205.420139	209.626972	9.189217e+06
min	770.450012	810.000000	738.750000	767.700012	747.039307	5.484040e+05
25%	1363.125000	1379.000000	1350.775024	1366.599976	1337.882019	6.207358e+06
50%	1484.000000	1495.000000	1467.550049	1482.650024	1454.269897	9.774854e+06
75%	1585.450012	1598.000000	1567.650024	1582.100036	1558.975036	1.594037e+07
max	1723.449951	1757.500000	1713.800049	1728.199951	1728.199951	8.670560e+07

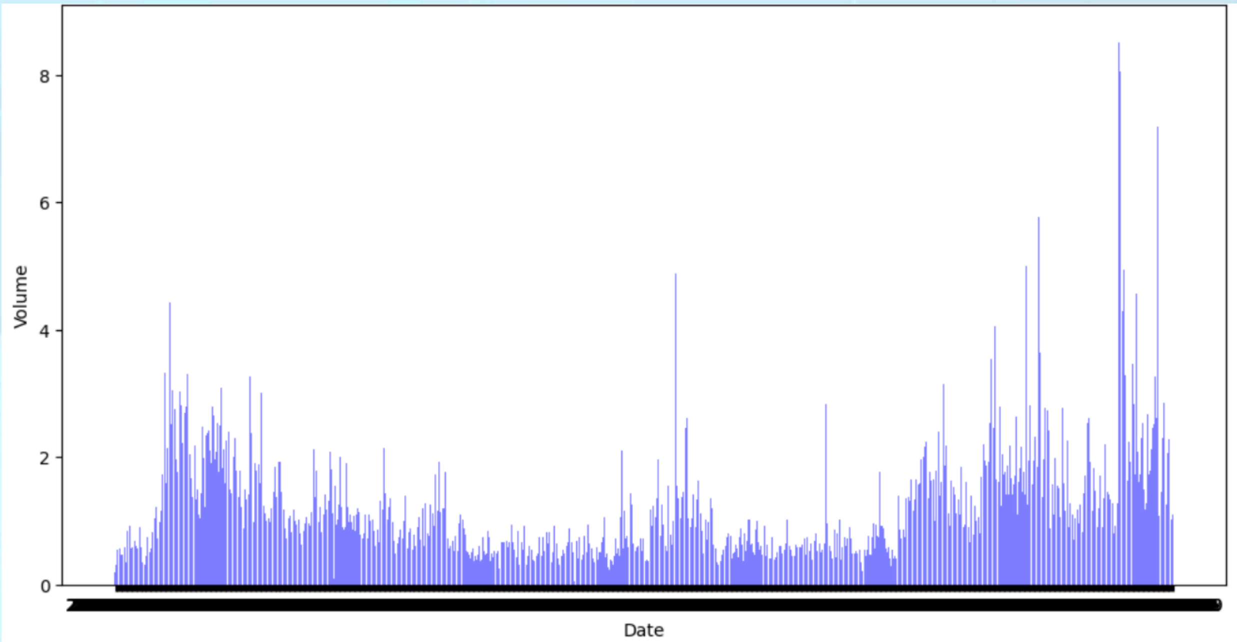
The above image shows us the mean, min, max count, and percentile statistics

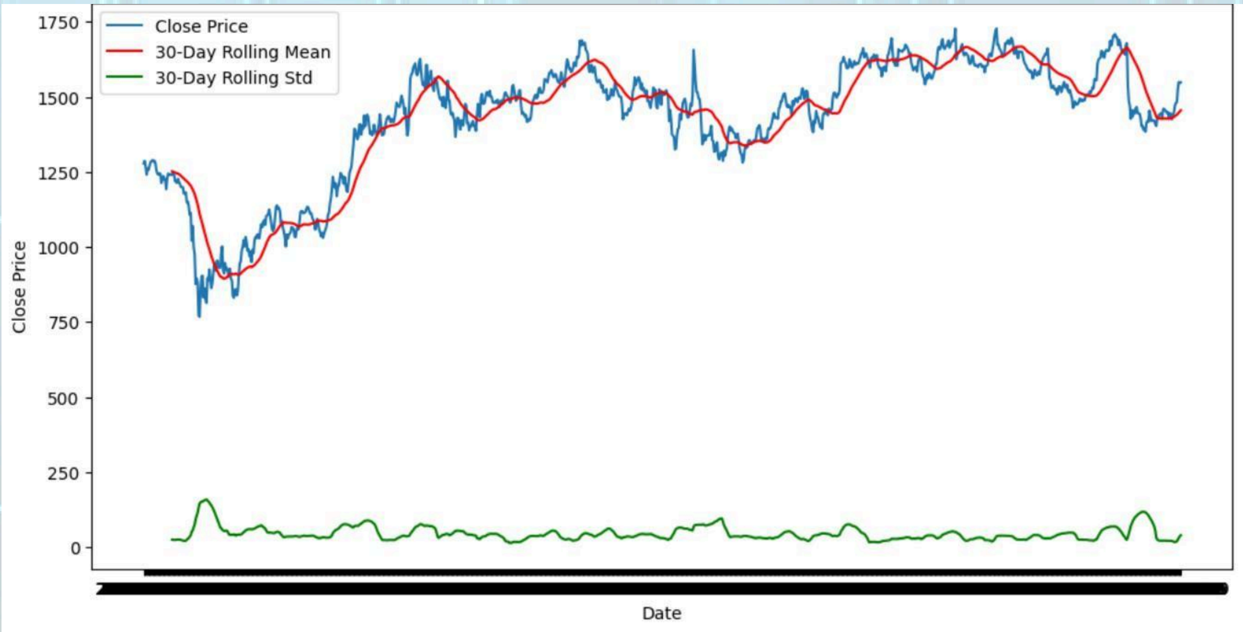
---

Exploratory Data Analysis

Plotted the daily closing price and daily trading volume data





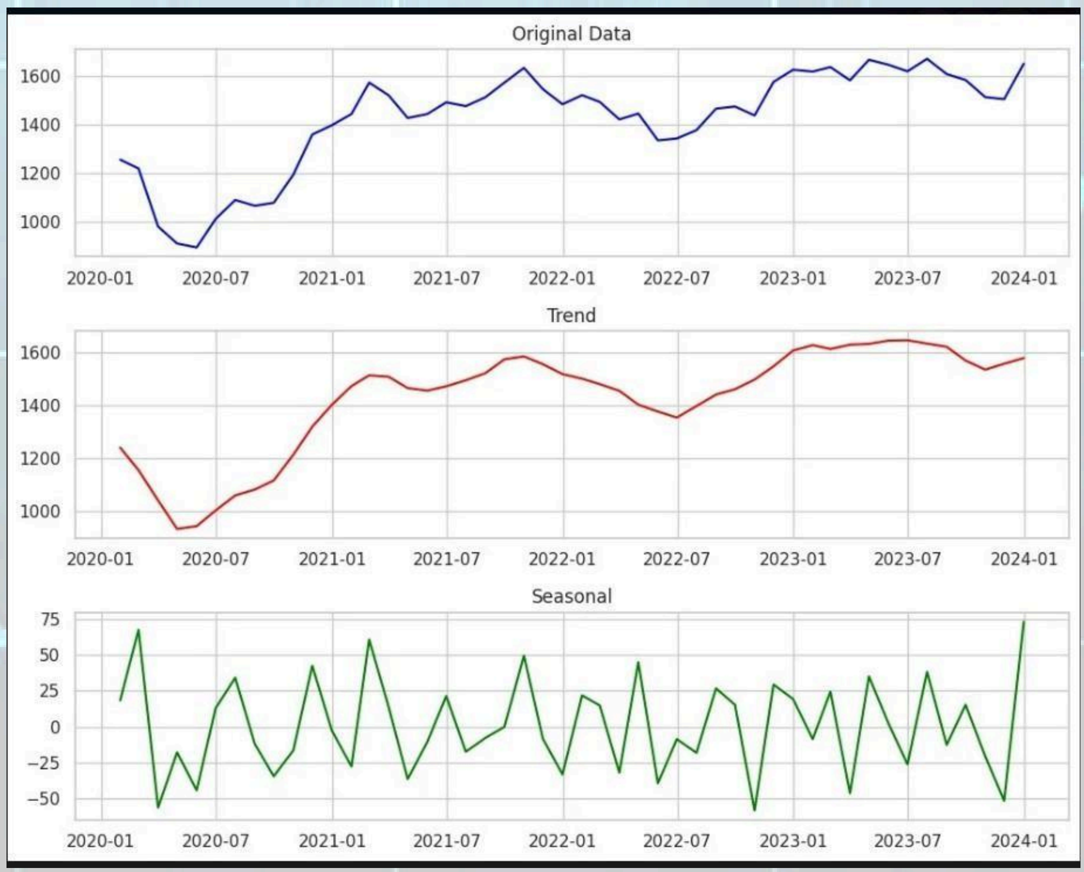


The above plot clearly shows that the data is non-stationary, as both the mean and variance appear to be dependent on time. To confirm our observations, we went on to perform the Augmented DickeyFuller Test, from which we found that the p values were 0.69 > 0.05; therefore, the series was nonstationary.

```
p-value: 0.6956054153750227
ADF Statistic: -1.1478799284910495
Critical Values:
    1%: -3.5778480370438146
    5%: -2.925338105429433
    10%: -2.6007735310095064
```

Data Preprocessing

We have decomposed the original data into Trend and Seasonal parts; we can see a clear upward trend here.



To make the data stationary, we perform the first difference (wherein we subtract the current value from its first lag value).

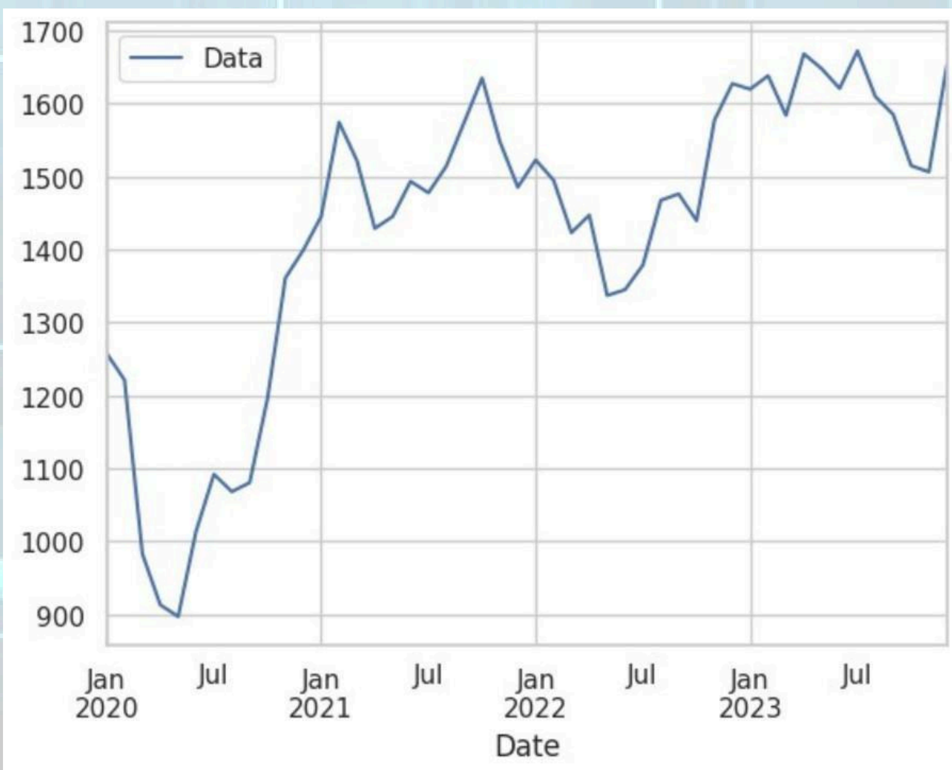
$$\nabla x_t = x_t - x_{t-1}.$$

Now, we again performed the ADF Test

```
p-value: 0.030435702226763305
ADF Statistic: -3.0505512414001186
Critical Values:
    1%: -3.6155091011809297
    5%: -2.941262357486514
    10%: -2.6091995013850418
```

The p-value is 0.03, and the critical value is 5%; hence, we can consider the data stationary.

We have resampled the data with the sampling interval taken to be a month, as it’s better to predict the monthly closing price as we don’t exactly know which day is working and which isn’t.



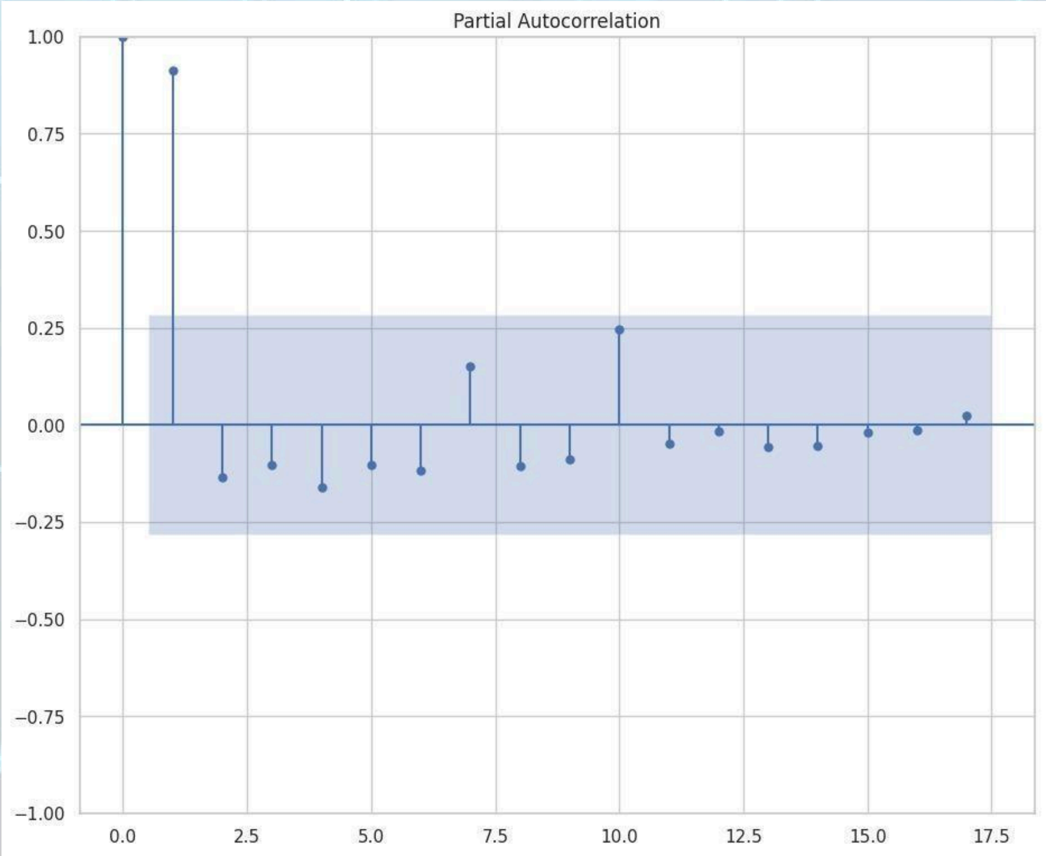
Monthly closing price data.

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.

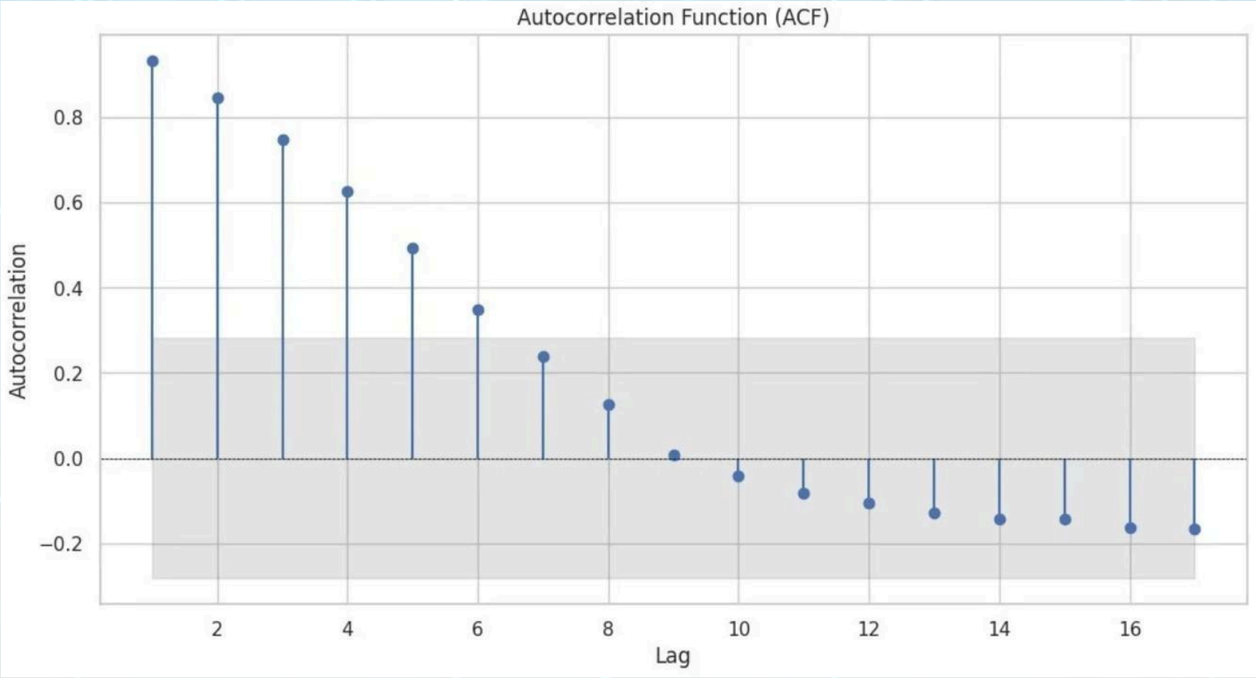


2. Fitting an AR model. (AR)

From the PACF plot, we can see a significant spike at lag 1 because of the significant PACF value. In contrast, we don't have evidence that everything within the blue band is different from zero.

p=1, hence we chose AR(1)

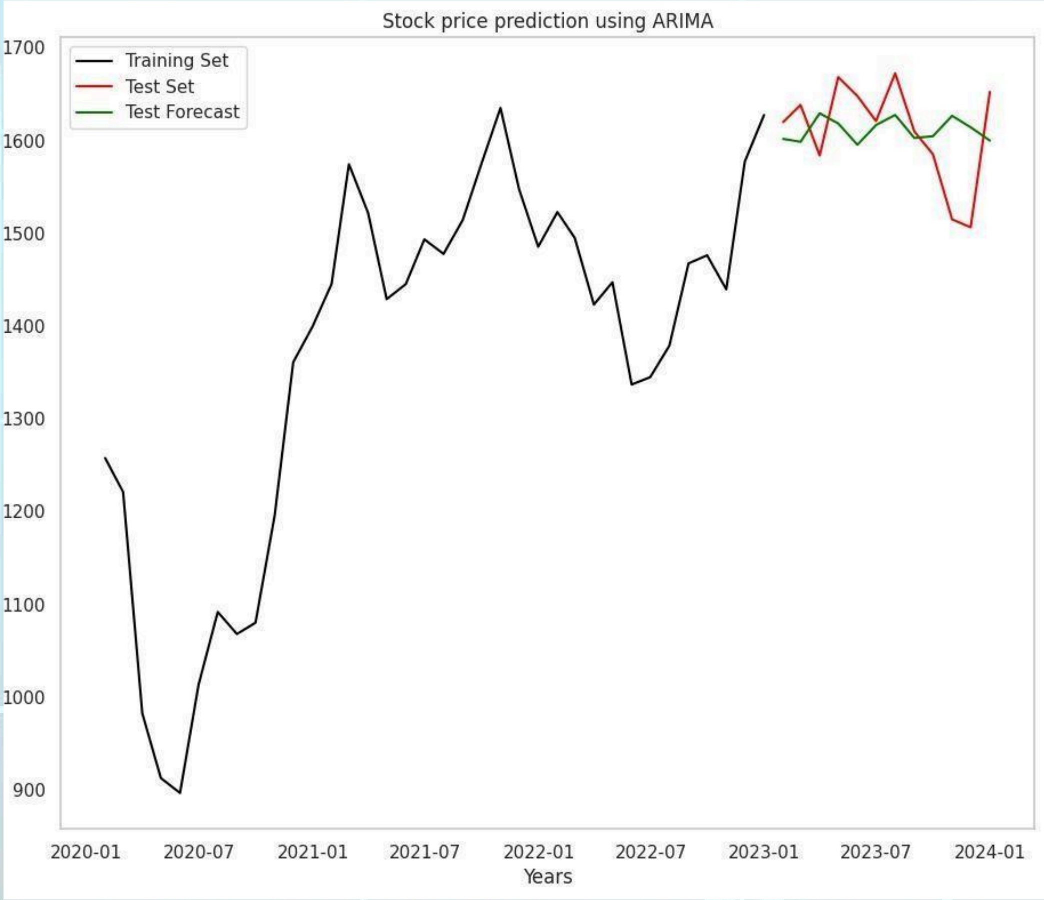
3. Fitting an MA model on the residuals. (MA)
We generated residuals by the difference of predictions from AR(1) and the original data.

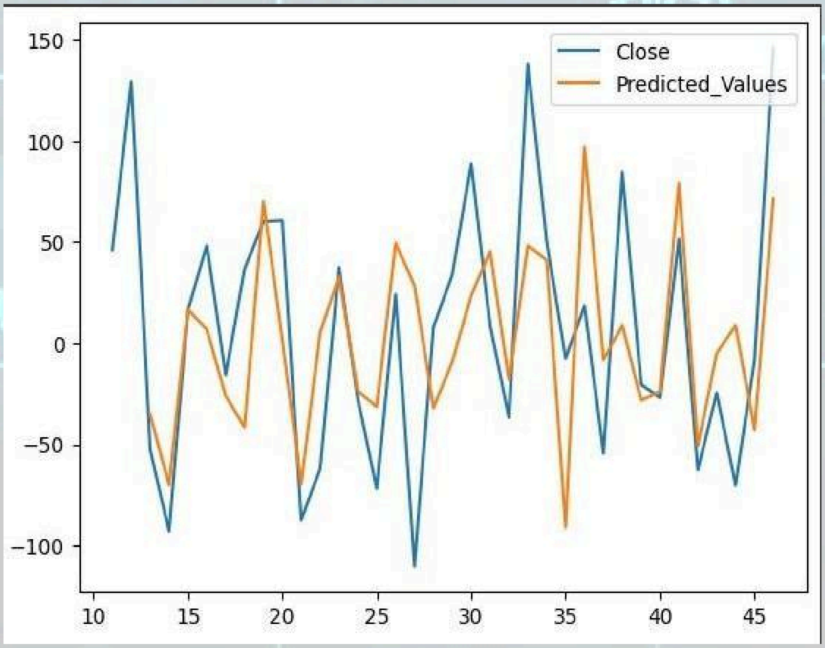


q=5, hence we chose MA(5)

Forecasting

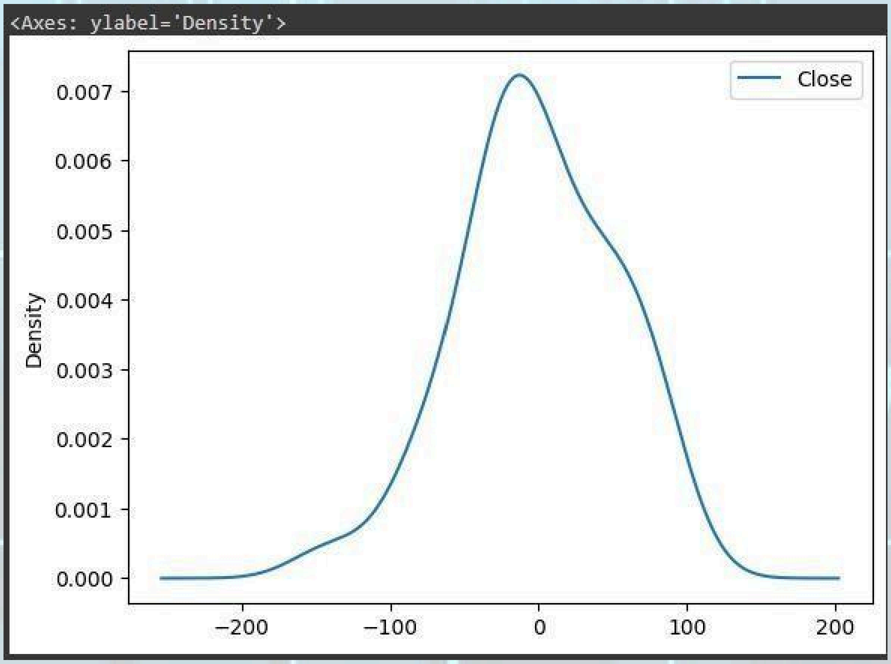
We have used 3 years of data to train the model(2020-22) and predicted closing prices for 12 months. The graph below gives an idea of future predictions.



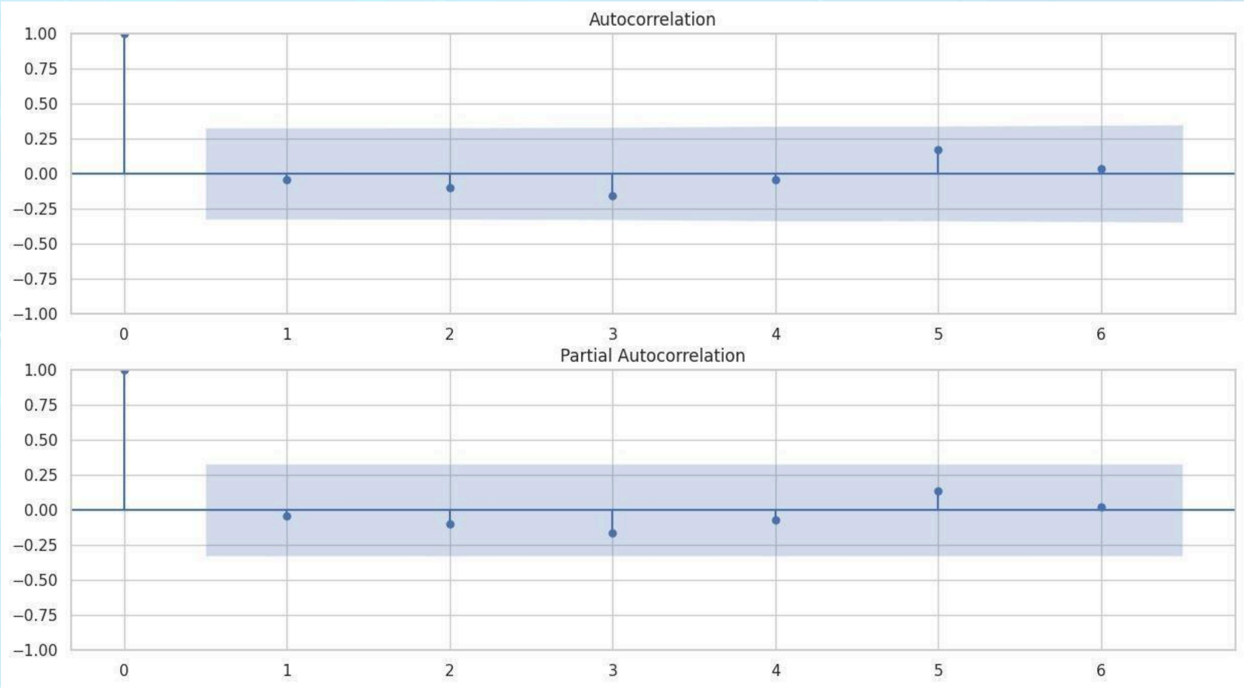


Result & Analysis

To analyze our model, we can plot the residual distribution. 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 an approximate normal distribution.



We can clearly see that all the following values of ACF and PACF lie within the confidence interval hence, our model provides a good prediction.