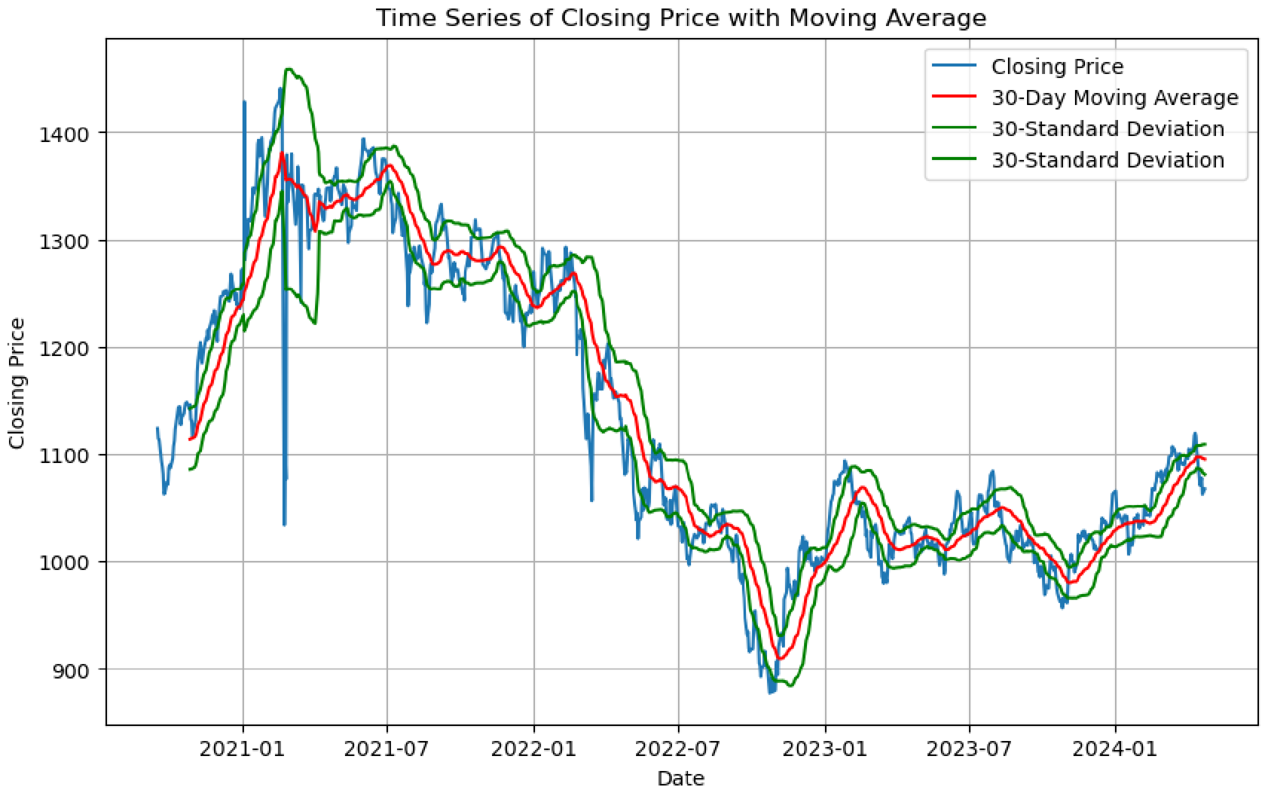
FTSE ESG Emerging Index Prediction

I. Data analysis.

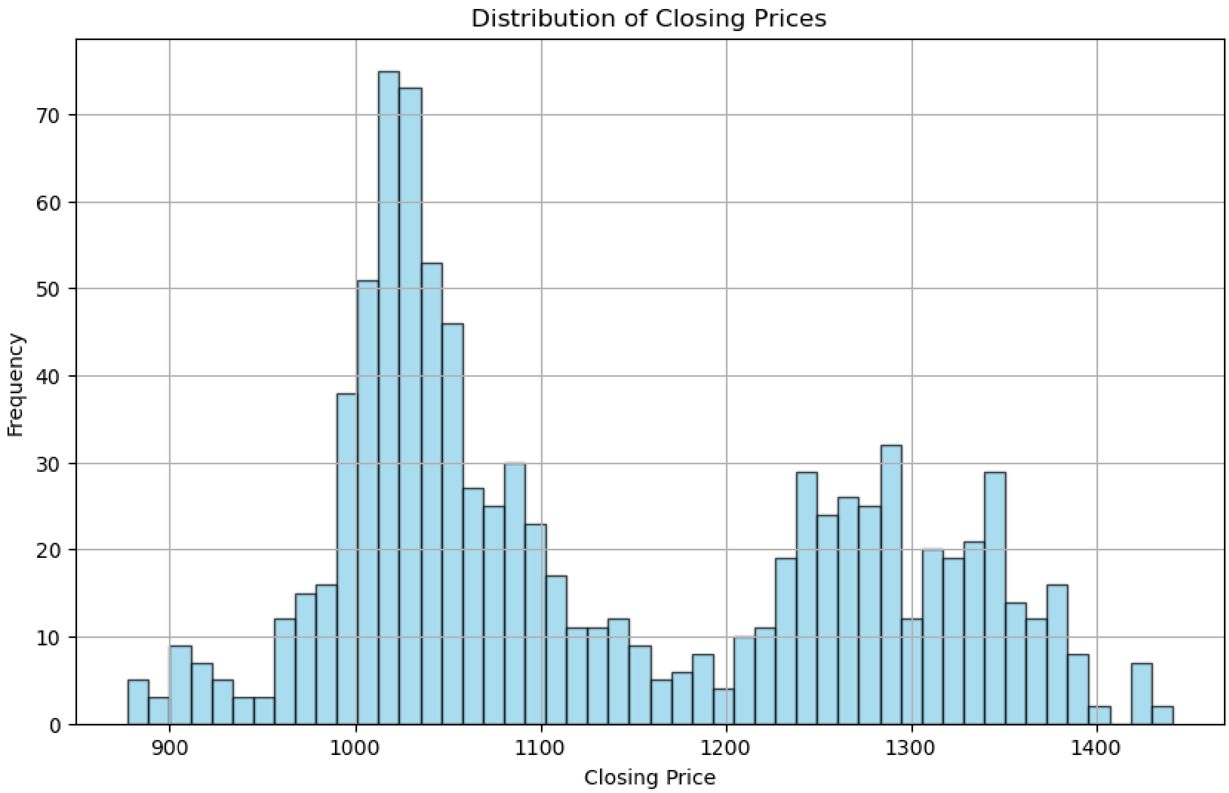
a. Data cleaing. Firstly I dealt with some erroneous data. The first thing that I have noticed was the fictitious 30th of February 2024 date. Another issue was the 1st of January 2023, which was a non trading, essentially just a duplicate of the following day (2-Jan-23). What led me to discovering these abnormalities was creating a Column and populating it with the log returns of the Closing Price series.

b. Managing missing Data. Because I do not find that the columns which display missing data are necessary for the following analysis and because the missing values would be hard to impute without resorting to looking up the data in other sources, I decided to drop the columns containing missing values and use only the following columns= Exchange Date, Close, Net. By differentiating the Closing Price series I also computed another series called Returns, and by apply the natural logarithm to the division between Close at time t to Close at time t-1 I also calculated the Log-Returns series. These columns are equivalent or near equivalent to the column %Chg, but they are not scaled as percentages and have better precision, thus they are more useful in time series analysis. The cleaning and the additional columns were done in excel.

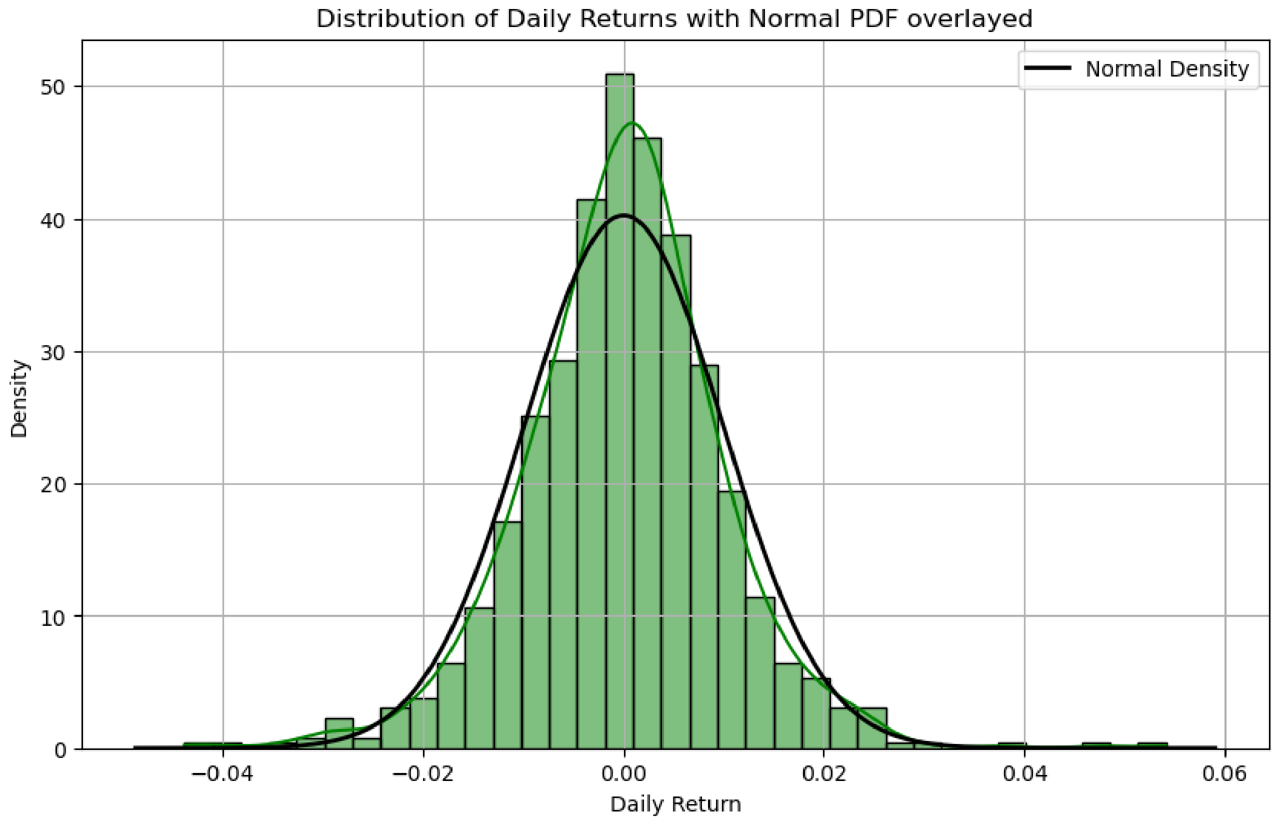
c. Trends in the data. In my opinion there is no apparent trend in the data. Both the Trend and Volatility appear to be stochastic, though we can identify different regimes. Initially the series was strongly trending upward, but after a period of extremely high volatility it started to trend downward for a while, until the trend reversed and towards the end of the Closing Prices series we can observe a mildly upward trend.

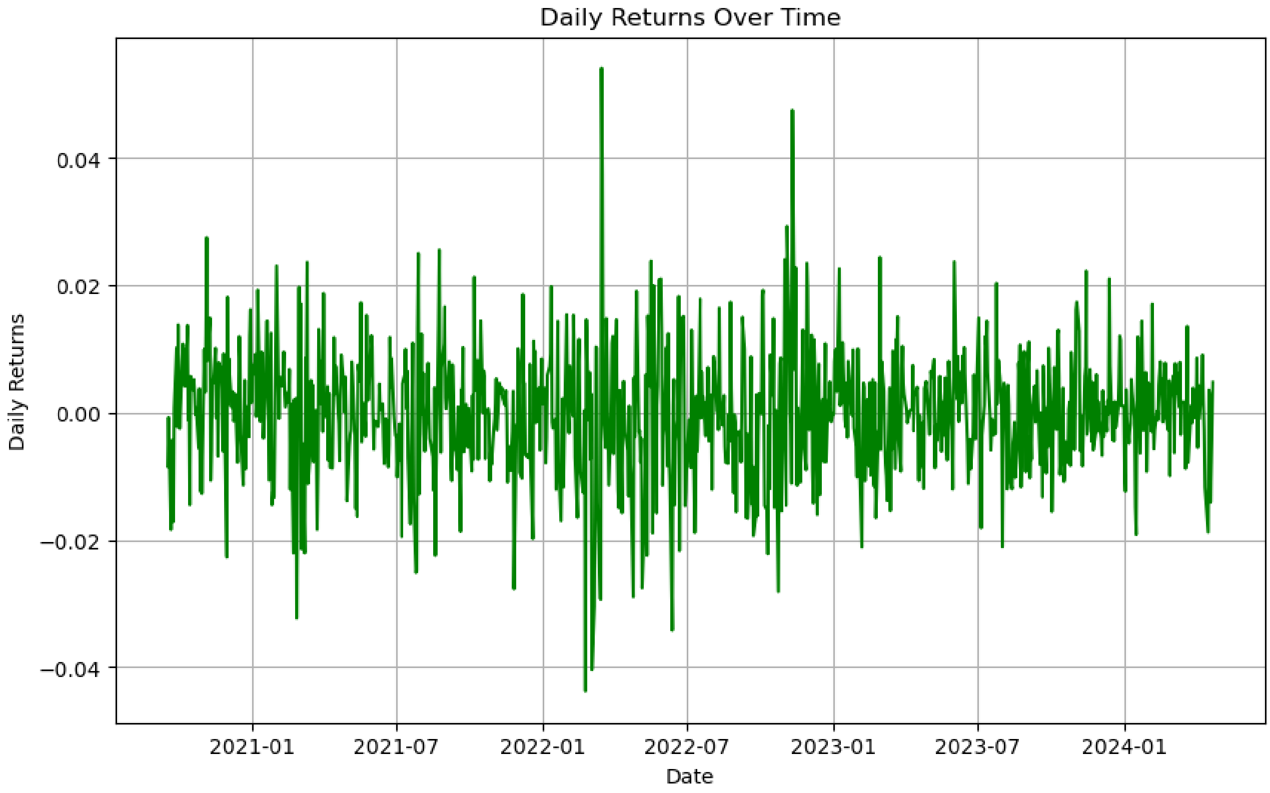


The distribution of Closing Prices plotted as a histogram does not reveal much either. The data appears to be slightly skewed and split in the middle where the mean should be and.



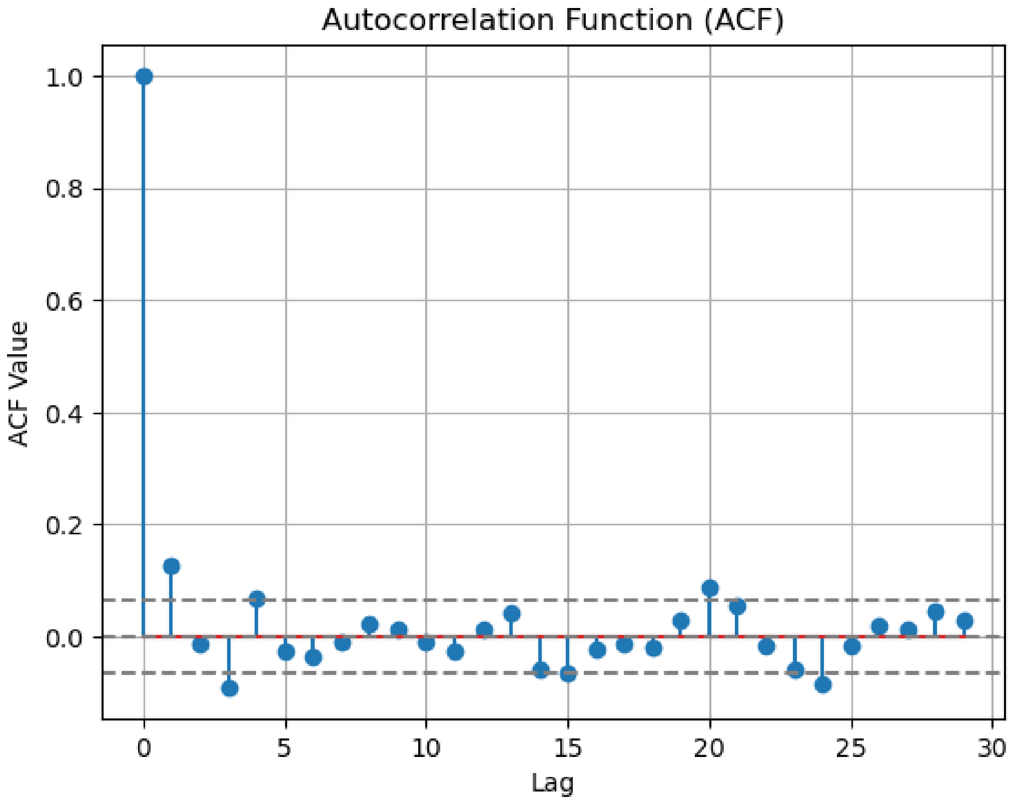
Using traditional Time Series analysis and forecasting methods we should work with the differenced or the Returns series in order to conduct our analysis of the data. As expected, the Daily Returns appear very obviously as a Normal Distribution on a histogram plot and as a Stationary White Noise centered closely to 0. The mean of the Returns Series is *0.000027* and its standard deviation is *0.009915*. As expected, the tails of the Returns Series are in fact slightly "fatter" than what we would expect from a Normal Distribution with the above-mentioned parameters, as events that determine extreme volatility happen more often in practice.

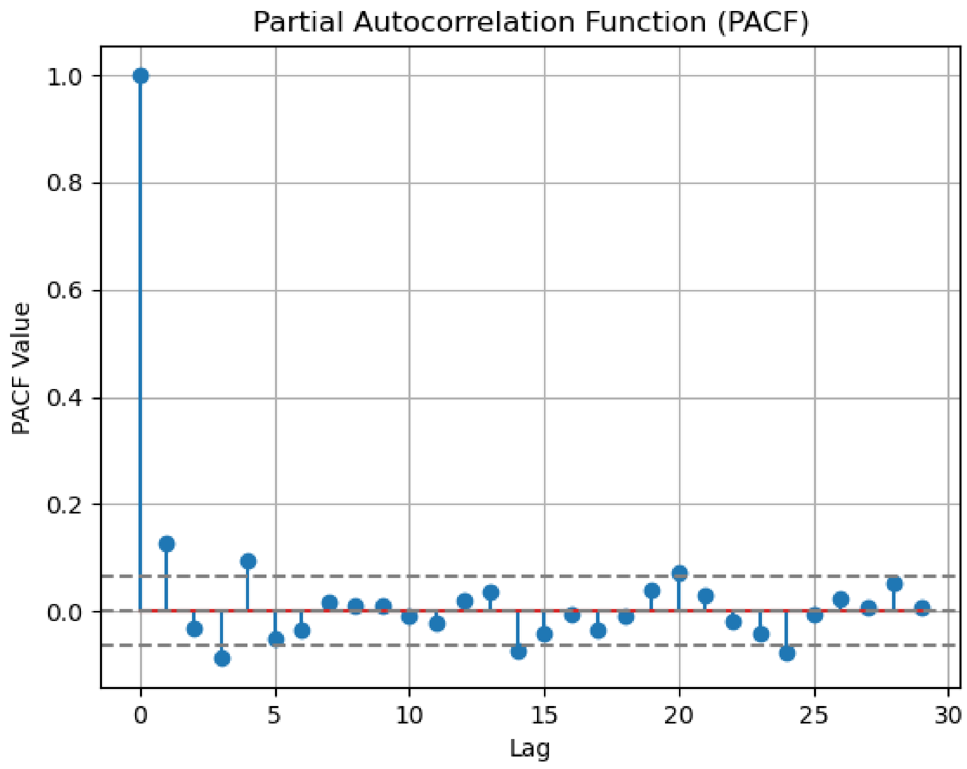

Distribution of Daily Returns with Normal PDF overlayed
Normal Density
Density
0
10
20
30
40
50
−0.04
−0.02
0.00
0.02
0.04
0.06
Daily Return


Daily Returns Over Time
Daily Returns
−0.04
−0.02
0.00
0.02
0.04
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
2024-01
Date

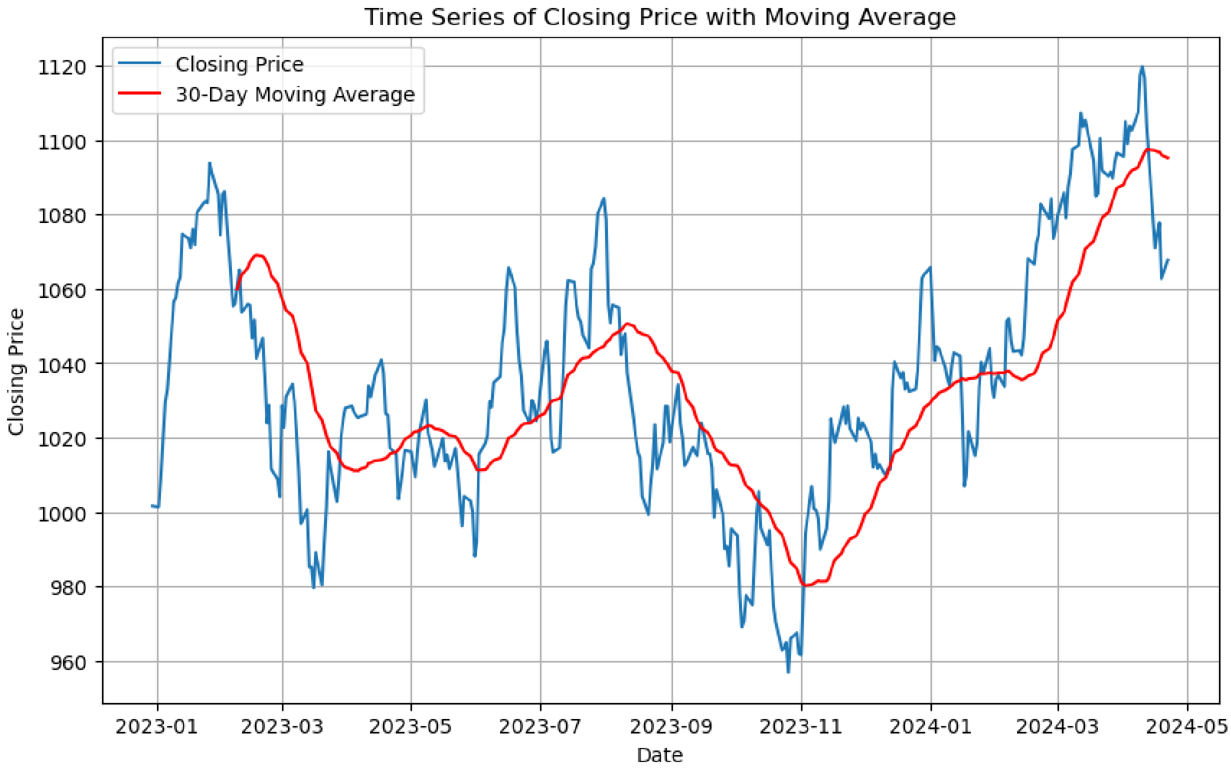
If we want to determine if there are any correlative effects between the different lags of the series, and their magnitude, we can use the Autocorrelation and the Partial Autocorrelation Function. After running these analyses, I could conclude that the lags of order 1,3,4,14,15,20 and 24 are statistically significant. Below are the coefficients of correlation of each lag with the lag[0].

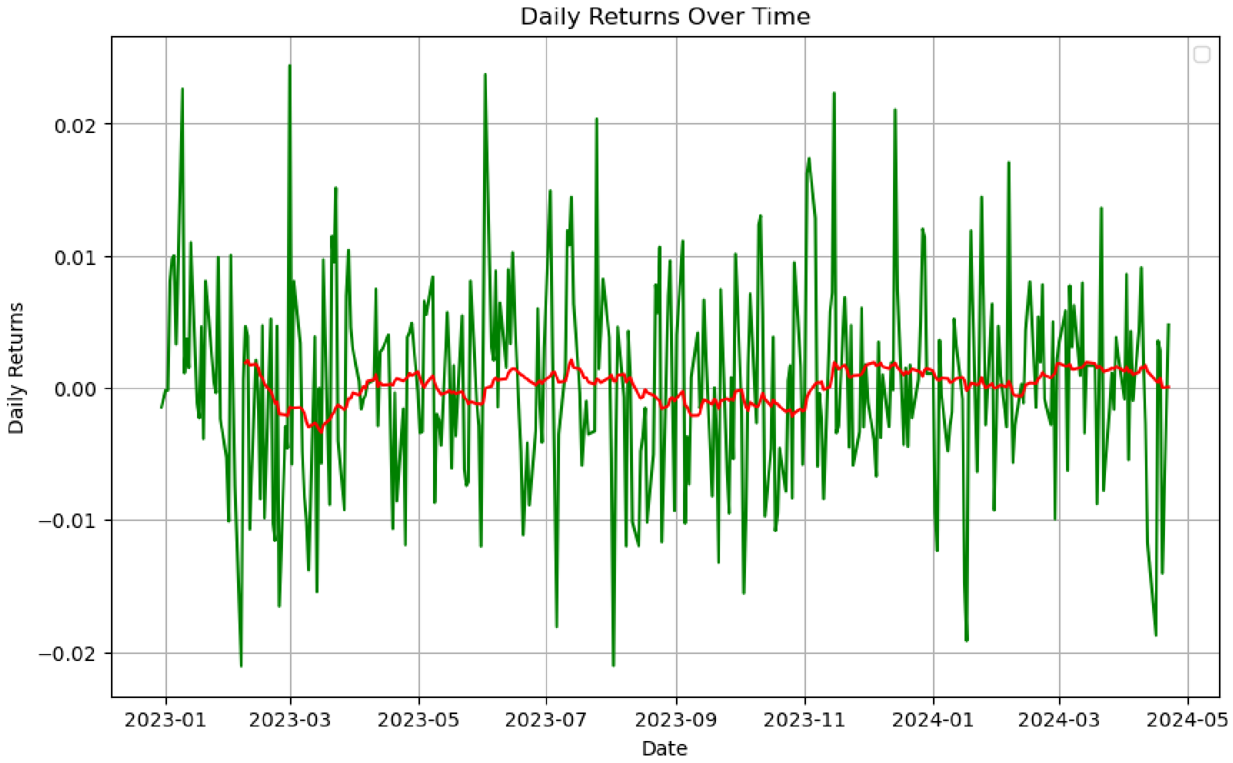
```
lag:  0  | acf_value:  1.0                          lag:  0  | pacf_value:  1.0
lag:  1  | acf_value:  0.12566813420112938          lag:  1  | pacf_value:  0.12580210875784703
lag:  3  | acf_value:  -0.09018036918979201         lag:  3  | pacf_value:  -0.08625321394604549
lag:  4  | acf_value:  0.06923601555082655          lag:  4  | pacf_value:  0.09385877937591967
lag:  15  | acf_value:  -0.06443002219897259       lag:  14  | pacf_value:  -0.07377406158640555
lag:  20  | acf_value:  0.08856511403922186        lag:  20  | pacf_value:  0.07336321609606812
lag:  24  | acf_value:  -0.08304489267274524       lag:  24  | pacf_value:  -0.0783977714387478
```





I believe that it is not useful to base our forecasting on values that are too far into the past to be relevant, likewise the lookback period for training and testing the model should ideally not find itself too far into the past. Analysing the latest 240 days of this series should be sufficient. I have chosen this period because it is reasonably safe to assume that we are finding ourselves in a mildly uptrending regime with low volatility at the current moment.


Time Series of Closing Price with Moving Average
Closing Price
30-Day Moving Average
Closing Price
1120
1100
1080
1060
1040
1020
1000
980
960
Date
2023-01
2023-03
2023-05
2023-07
2023-09
2023-11
2024-01
2024-03
2024-05


Daily Returns Over Time
Daily Returns
0.02
0.01
0.00
−0.01
−0.02
Date
2023-01
2023-03
2023-05
2023-07
2023-09
2023-11
2024-01
2024-03
2024-05

II. Modelling and Forecasting

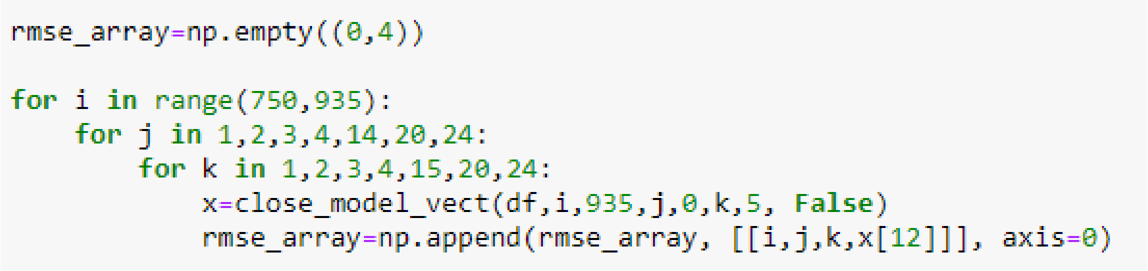
Based on the information that the AC and the PAC functions could provide I have concluded that an ARIMA model with the parameters p and q selected in such a thoughtful manner would be the best approach given that the AR component can capture the trend following influences, whereas the MA component can capture the mean-reverting effects of a time series. I believe that this class of models is the most appropriate for this kind of problem, where we would want to predict series with reasonable accuracy while at the same time avoiding overfitting to the data.

I have used the ACF to determine the optimal order of the Moving Average component and the PACF to determine the one for the Autoregressive component.

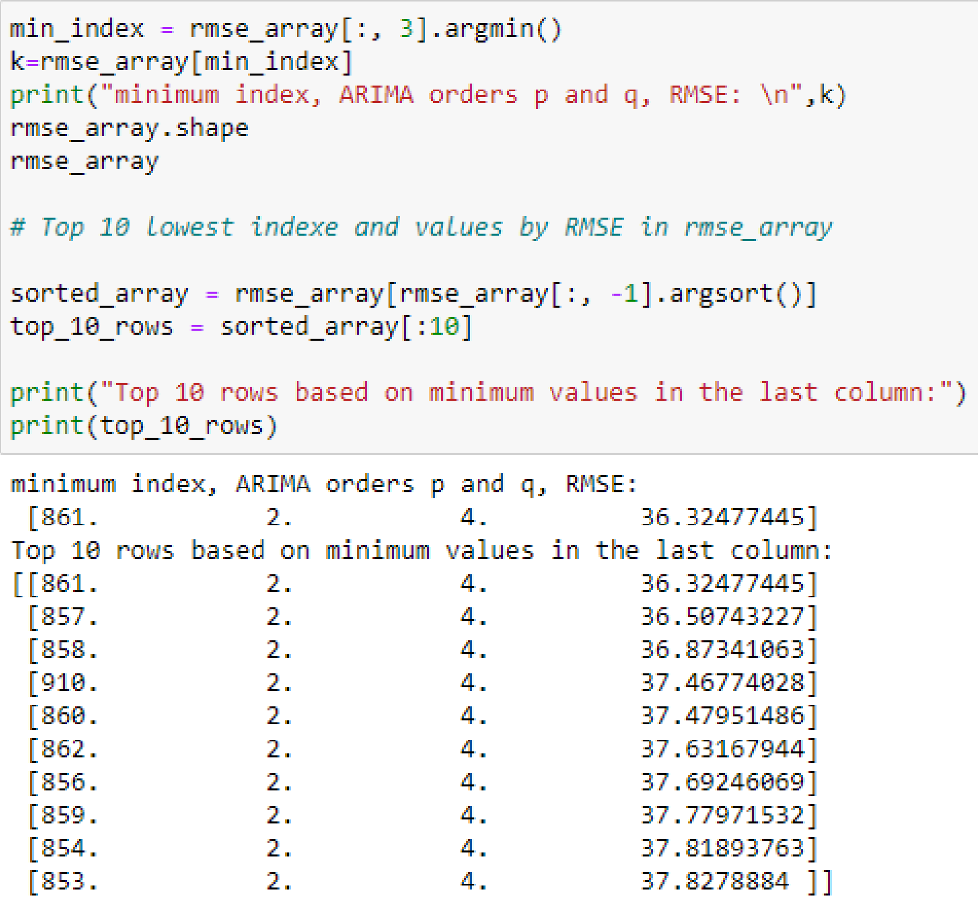
As mentioned above, I have determined that in the case of the MA order, the lags 1,3,4,15,20 and 24 are statistically significant (we can infer that by inspecting the ACF), thus I will chose one of these values for p in ARIMA(p,d,q). In the case of the AR component, by inspecting the PACF we arrive at similar conclusions, where the lags 1,3,4,14,20,24 are significant, thus the order q will take on one of these values.

III. Lookback Period

Given that the assignment asks for a 5 day forecast, I will split my data such that the “testing set” will also be 5 days, and the training set will be of a variable length. Then I will attempt to compute RMSE between the realized actual closing prices, and the forecasted values. The forecasted values will be computed by multiplying the lag[0] with the first predicting return+1, then this value forecast[0] or “lag[-1]” will be similarly multiplied with 1+the next forecasted return.



```
rmse_array=np.empty((0,4))

for i in range(750,935):
    for j in 1,2,3,4,14,20,24:
        for k in 1,2,3,4,15,20,24:
            x=close_model_vect(df,i,935,j,0,k,5, False)
            rmse_array=np.append(rmse_array, [[i,j,k,x[12]]], axis=0)
```

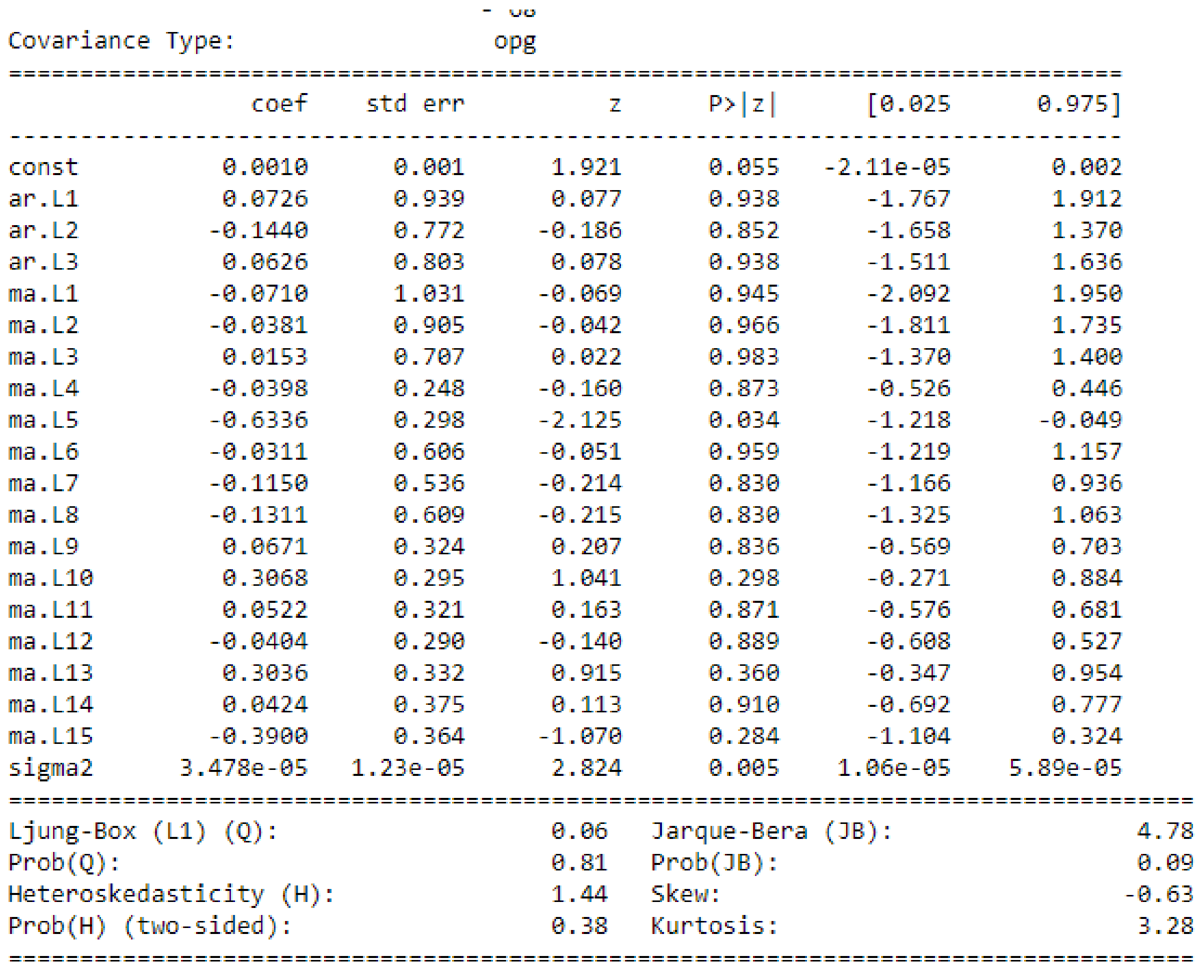


```python
min_index = rmse_array[:, 3].argmin()
k=rmse_array[min_index]
print("minimum index, ARIMA orders p and q, RMSE: \n",k)
rmse_array.shape
rmse_array

# Top 10 lowest indexe and values by RMSE in rmse_array

sorted_array = rmse_array[rmse_array[:, -1].argsort()]
top_10_rows = sorted_array[:10]

print("Top 10 rows based on minimum values in the last column:")
print(top_10_rows)
```

```
minimum index, ARIMA orders p and q, RMSE: 
 [861.           2.            4.          36.32477445]
Top 10 rows based on minimum values in the last column:
[[861.           2.            4.          36.32477445]
 [857.           2.            4.          36.50743227]
 [858.           2.            4.          36.87341063]
 [910.           2.            4.          37.46774028]
 [860.           2.            4.          37.47951486]
 [862.           2.            4.          37.63167944]
 [856.           2.            4.          37.69246069]
 [859.           2.            4.          37.77971532]
 [854.           2.            4.          37.81893763]
 [853.           2.            4.          37.8278884 ]]
```

In the following code section I am forecasting the ARIMA model of order (p,d,q), where d=0, p=k and q=j, and the lookback period is variable as both the starting and ending time index can be variable as well. In this case I have chosen to choose 800 and 920 respectively to border the range, but I have gone over many different value combinations starting from 600 and ending with 935 (I cannot compare any forecast after this value with any other value in the dataset).

After testing multiple order of p and q in ARIMA(p,d,q) over a great number of ranges, and computing the minimum RMSE, I have concluded that the model with the lowest error is one of the order ARIMA(3,0,15) with the lookback period of 68 days. The following is the model summary:



```
                                  - 08
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0010      0.001      1.921      0.055   -2.11e-05      0.002
ar.L1          0.0726      0.939      0.077      0.938      -1.767      1.912
ar.L2         -0.1440      0.772     -0.186      0.852      -1.658      1.370
ar.L3          0.0626      0.803      0.078      0.938      -1.511      1.636
ma.L1         -0.0710      1.031     -0.069      0.945      -2.092      1.950
ma.L2         -0.0381      0.905     -0.042      0.966      -1.811      1.735
ma.L3          0.0153      0.707      0.022      0.983      -1.370      1.400
ma.L4         -0.0398      0.248     -0.160      0.873      -0.526      0.446
ma.L5         -0.6336      0.298     -2.125      0.034      -1.218     -0.049
ma.L6         -0.0311      0.606     -0.051      0.959      -1.219      1.157
ma.L7         -0.1150      0.536     -0.214      0.830      -1.166      0.936
ma.L8         -0.1311      0.609     -0.215      0.830      -1.325      1.063
ma.L9          0.0671      0.324      0.207      0.836      -0.569      0.703
ma.L10         0.3068      0.295      1.041      0.298      -0.271      0.884
ma.L11         0.0522      0.321      0.163      0.871      -0.576      0.681
ma.L12        -0.0404      0.290     -0.140      0.889      -0.608      0.527
ma.L13         0.3036      0.332      0.915      0.360      -0.347      0.954
ma.L14         0.0424      0.375      0.113      0.910      -0.692      0.777
ma.L15        -0.3900      0.364     -1.070      0.284      -1.104      0.324
sigma2      3.478e-05   1.23e-05      2.824      0.005    1.06e-05    5.89e-05
===================================================================================
Ljung-Box (L1) (Q):                   0.06   Jarque-Bera (JB):                 4.78
Prob(Q):                              0.81   Prob(JB):                         0.09
Heteroskedasticity (H):               1.44   Skew:                            -0.63
Prob(H) (two-sided):                  0.38   Kurtosis:                         3.28
===================================================================================
```

The forecasted returns are:

[0.01528827, 0.00257943, 0.0051864, 0.00948583, 0.00419364]

The forecasted Closing Prices are:

[1084.13496881, 1086.93141591, 1092.56867634, 1102.93259589, 1107.55789904]

date	value	team
4/23/2024	1084.134969	Badea Andrei Carol
4/24/2024	1086.931416	
4/25/2024	1092.568676	
4/26/2024	1102.932596	
4/29/2024	1107.557899	