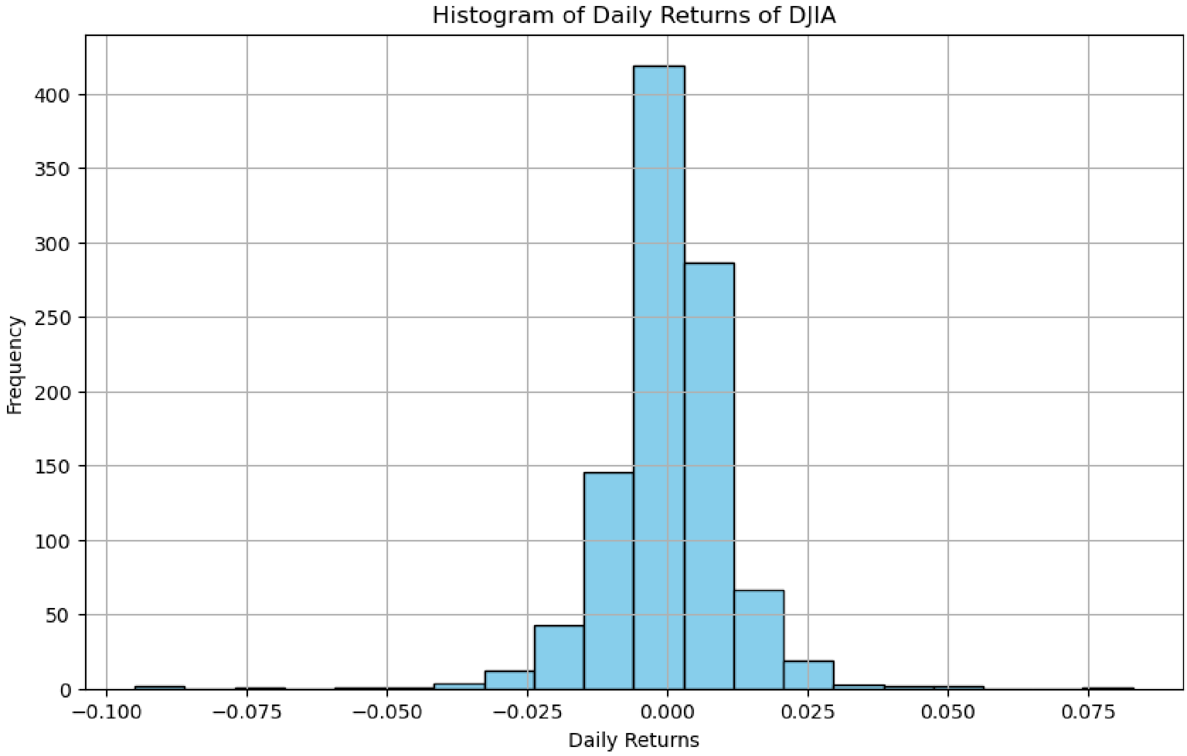
Final Exam Report

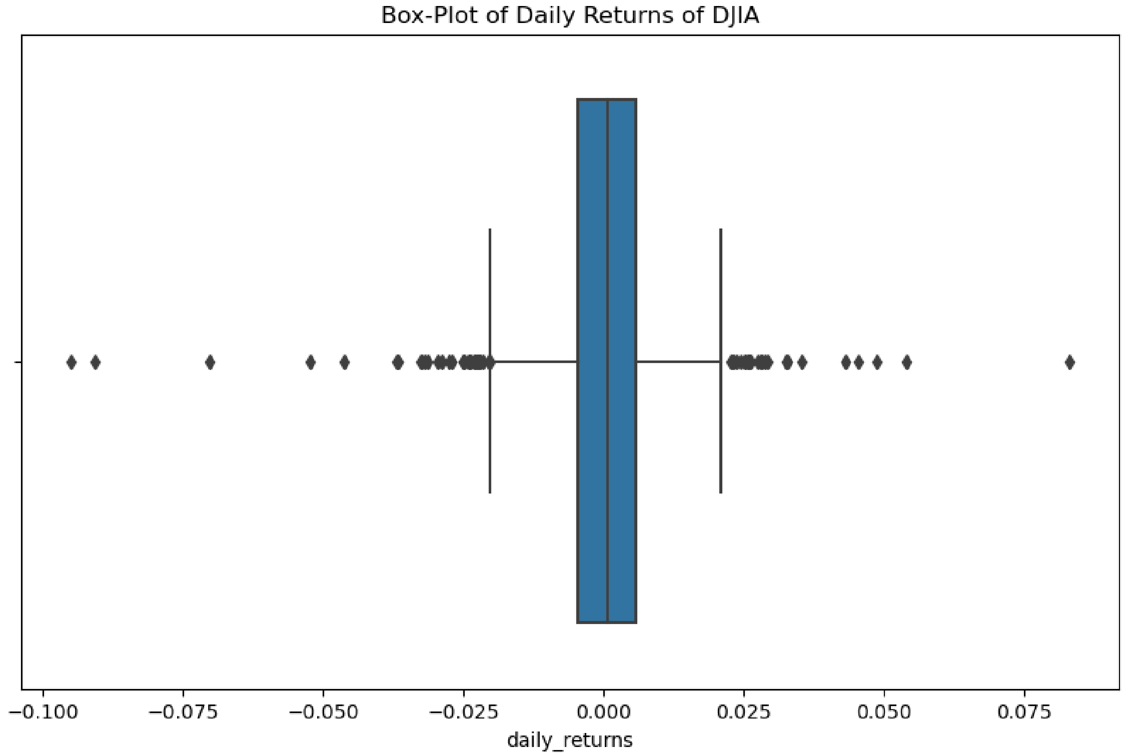
For the Final Exam, I chose the Dow Jones Industrial Average data. Starting with Question 1, I tried to do some basic cleaning of the data, by converting the datetime format in the format, to the datetime format implemented within pandas. Since the instruction was to take in account 4 years of data, the starting date was taken as **2019/01/01** and ending date was taken as **2023/01/01.**

Below is the histogram plot generated for the daily returns within the time limits of the data.



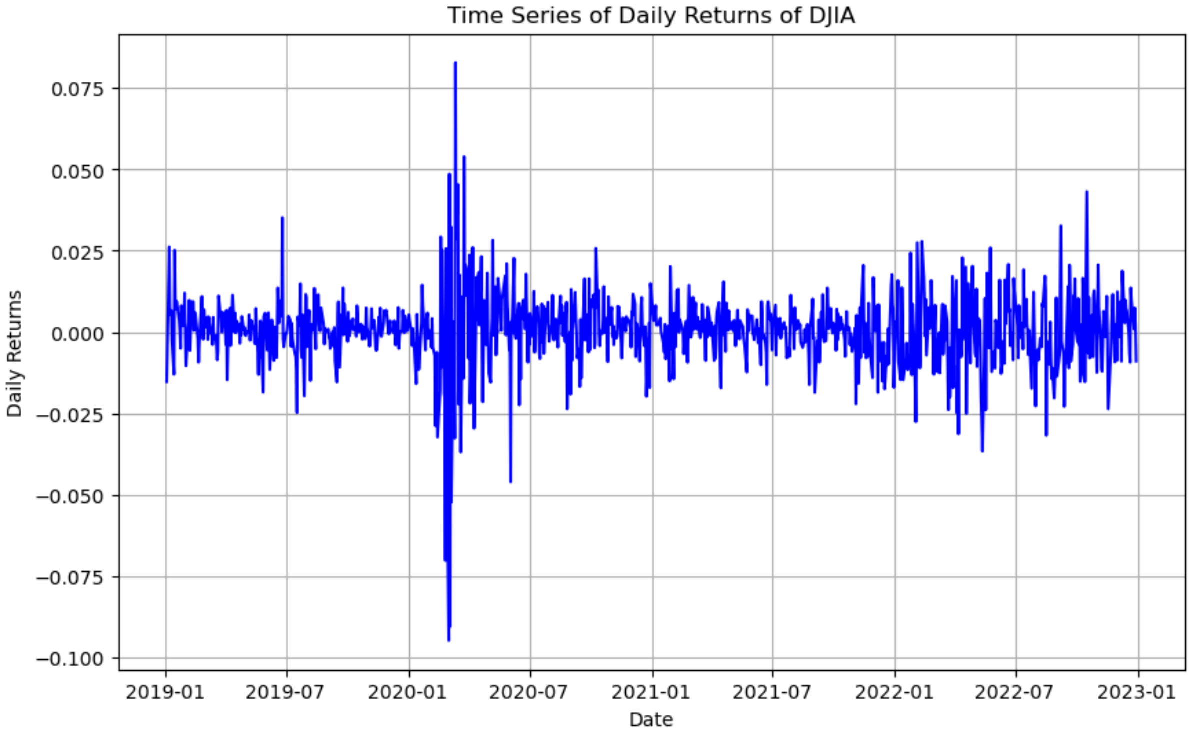
We can observe that there were some extreme returns values of -0.100 and 0.075, which accounted for further investigation.

Below is the Box-Plot of the daily returns within the specified time limit.



We can observe the median to be around 0.000 and there are lot of datapoints which lies outside the whiskers of the box-plot, which shows the 25 percentile and 75 percentile of the daily returns.

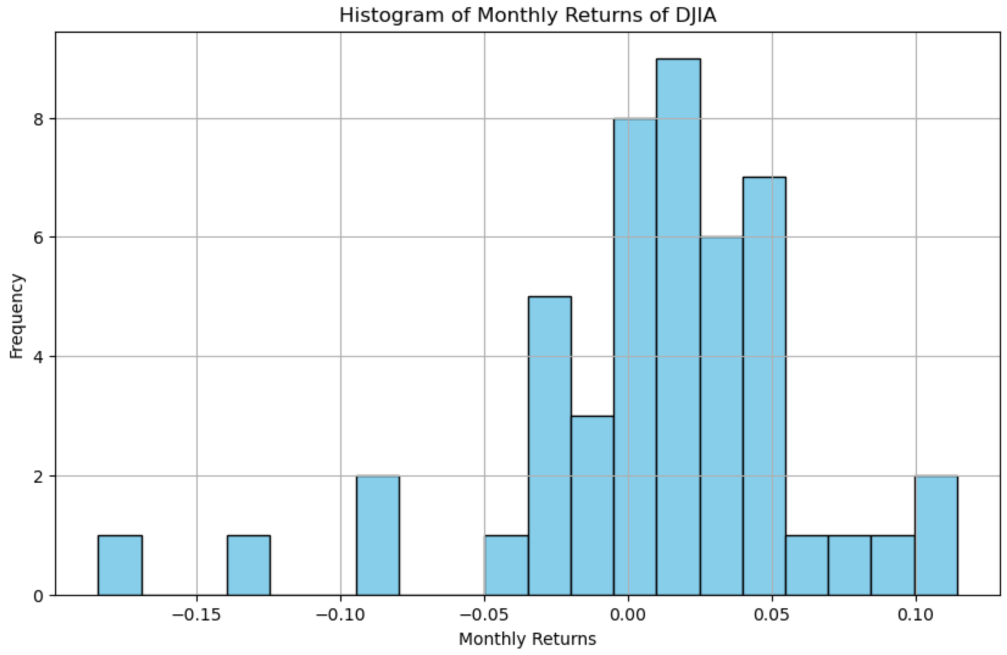
Below is the time-series plot for the daily returns.

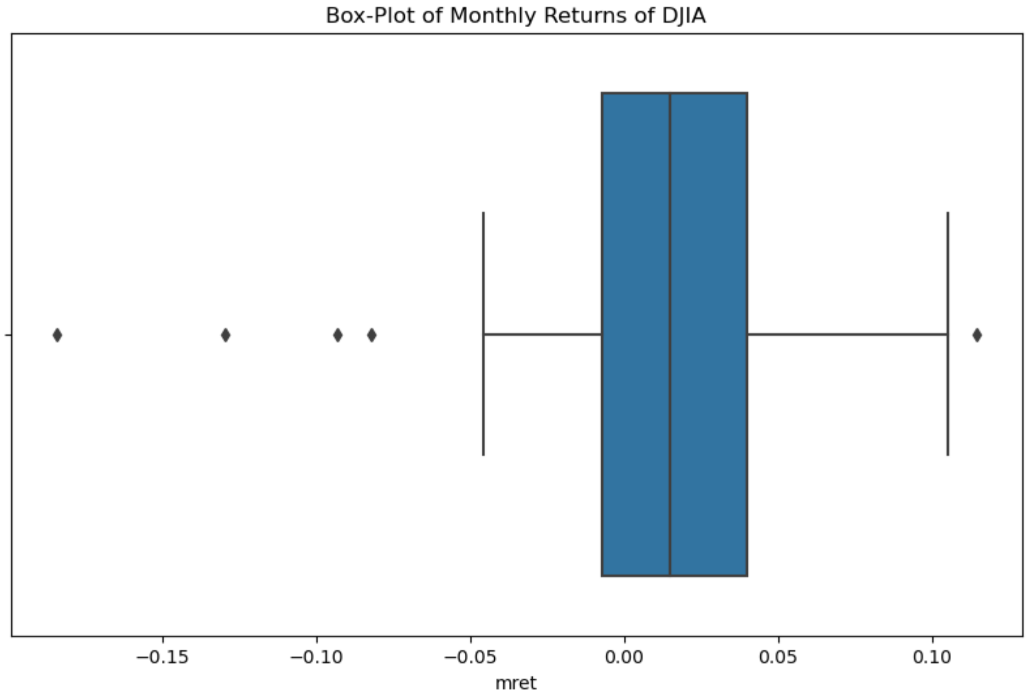


You can see here, that after 2020, there was a considerable movement in the daily returns of the index. Just before 2019-07 and after 2022-07 also had events which could be considered huge movements away from mean.

Now in order to calculate the monthly returns, it was required to rebalance the daily returns, and index the data for the period of one month. The procedure is described in the notebook.

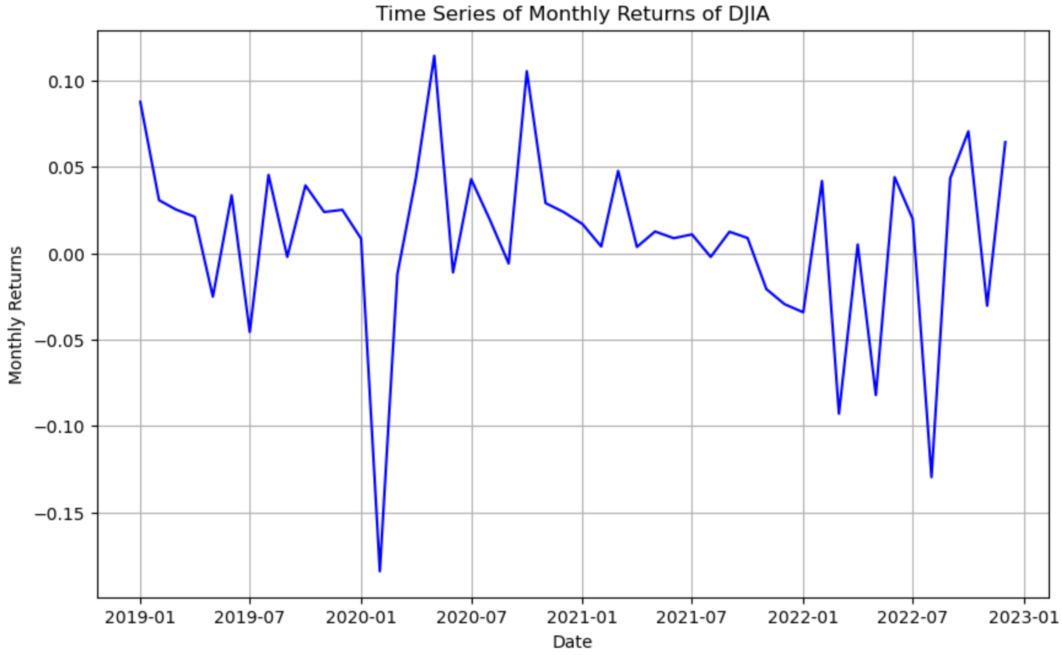
Below is the histogram plot and box-plot for the monthly returns

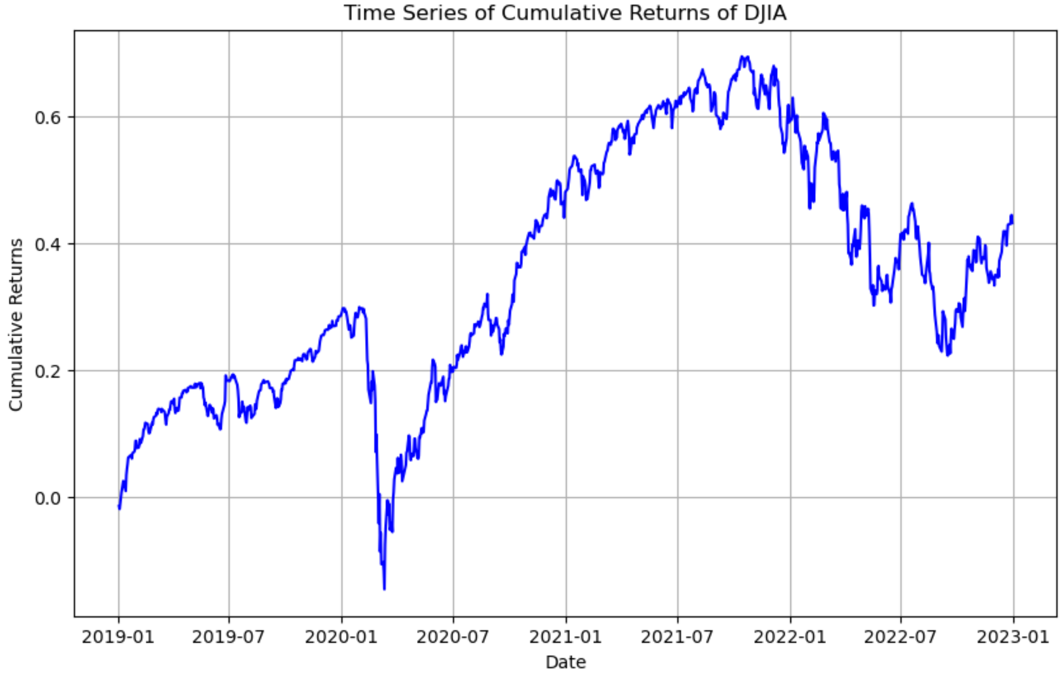




Plotting the monthly returns, reduces the deviation of the daily returns from its normality assumption. The primary benefit of using monthly returns data instead of daily return data is that with monthly data, returns are at least *approximately* normally distributed (or, at the very least, the simplifying assumption of normality is much less crazy for monthly returns than it is for daily returns).

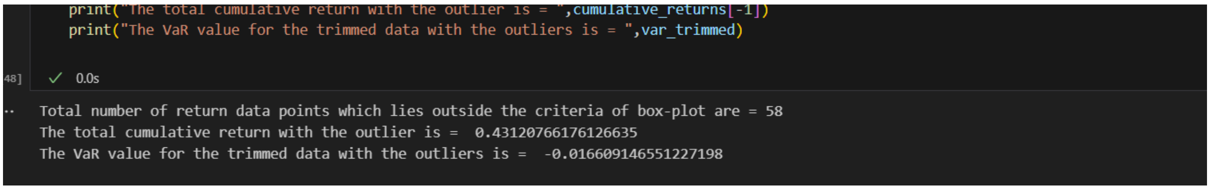
Below is the time series plot of monthly returns and cumulative returns





Cumulative returns show that is 1000 dollars invested at the start of period, how much their values are going to be at the end of period. It can be shown that if people would have invested around lowest points of 2020-01 in Dow Jones index, they could have earned heavy returns by 2021-07.

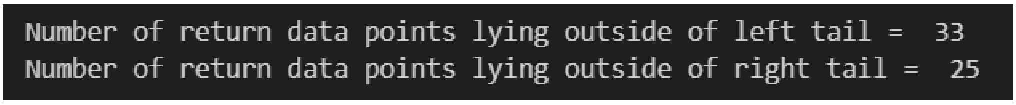
For Question 2, The box-plot method was used to detect the outliers for the daily returns. The difference between the first quantile and third quantile, called the Inter Quantile Range was used to determine the thresholds. Total number of outliers found in the data were 58.



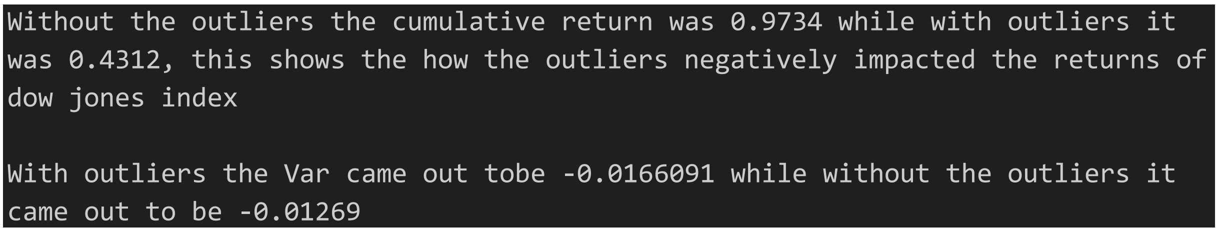
```
    print("The total cumulative return with the outlier is = ",cumulative_returns[-1])
    print("The VaR value for the trimmed data with the outliers is = ",var_trimmed)

48]   ✓ 0.0s

··  Total number of return data points which lies outside the criteria of box-plot are = 58
    The total cumulative return with the outlier is =  0.43120766176126635
    The VaR value for the trimmed data with the outliers is =  -0.016609146551227198
```



```
Number of return data points lying outside of left tail =  33
Number of return data points lying outside of right tail =  25
```



```
Without the outliers the cumulative return was 0.9734 while with outliers it
was 0.4312, this shows the how the outliers negatively impacted the returns of
dow jones index

With outliers the Var came out tobe -0.0166091 while without the outliers it
came out to be -0.01269
```

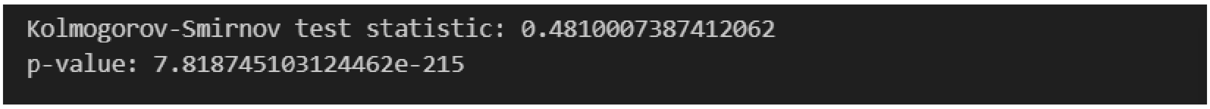
Value at Risk (VaR) and the Cumulative returns with and without outliers gives accurate description of the impact of outliers.

For Question 3, I used the Kolmogrov-Smirnov test. This test used to test the null hypothesis that values generated (empirical distribution) follows or derived from normal distribution. The KS test, calculated the residuals between the normal cumulative density function and the empirical cumulative distribution. The KS test statistic and the p-values, helps us to either accept or reject the null hypothesis.

H_0 = Empirical distribution follow normal distribution.

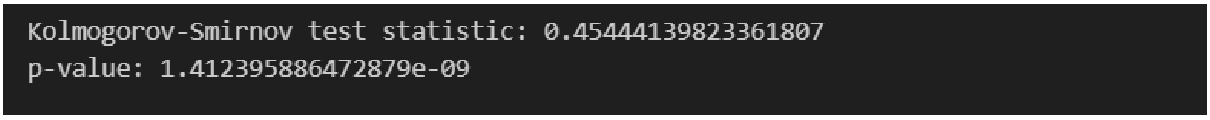
H_1 = Empirical distribution doesn't follows normal distribution.

Below is the Output for KS test for daily returns



```
Kolmogorov-Smirnov test statistic: 0.4810007387412062
p-value: 7.818745103124462e-215
```

Below is output for KS test for monthly returns

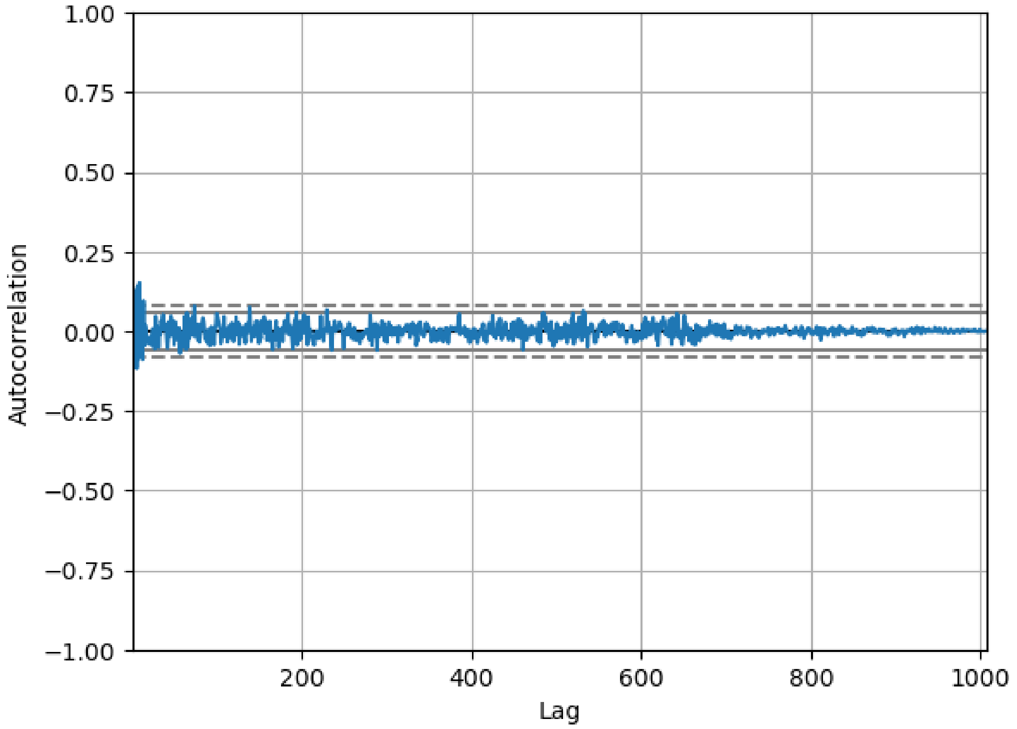


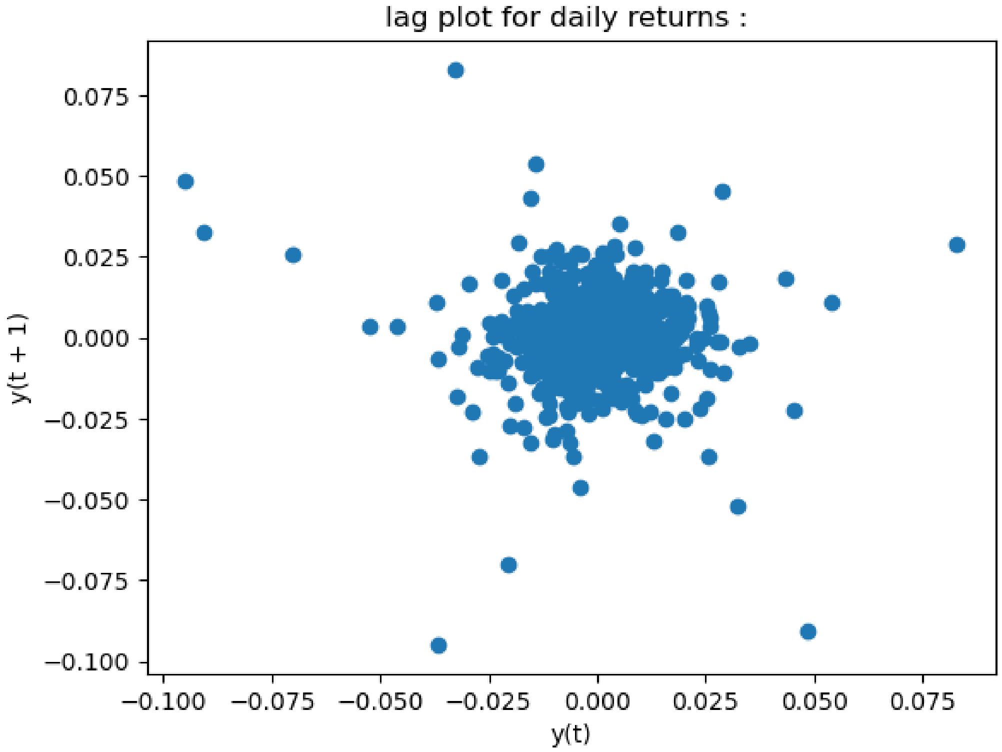
```
Kolmogorov-Smirnov test statistic: 0.45444139823361807
p-value: 1.412395886472879e-09
```

We can reject the null hypothesis here at 0.05 significance. Interesting thing is, the KS test statistic for monthly returns is less compare to the daily returns. This shows a slight resemblance of the monthly returns with normal distribution.

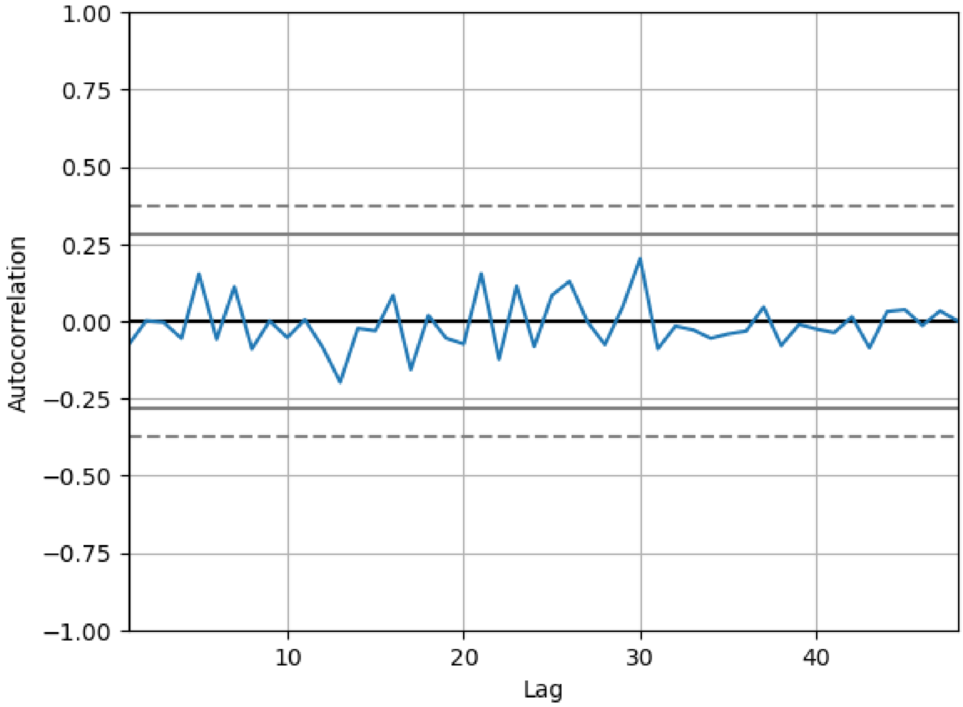
For question 4, the lag plots and auto correlation plots both were done to test for daily returns and monthly returns.

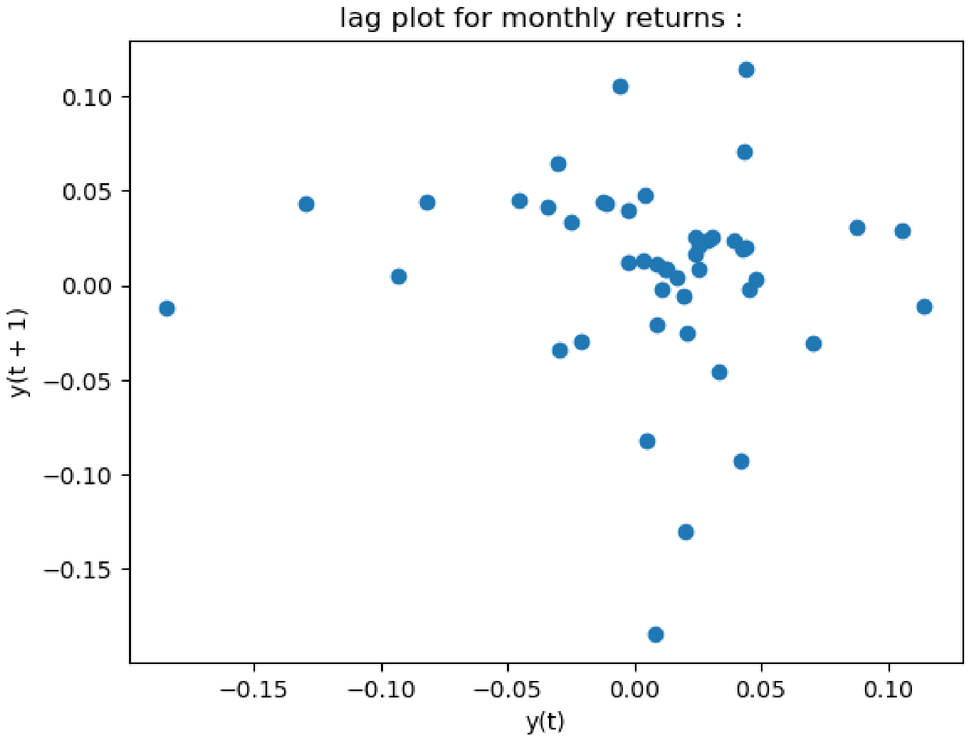
Below is autocorrelation plot and lag plot for daily returns





And below are the same plots done for monthly returns





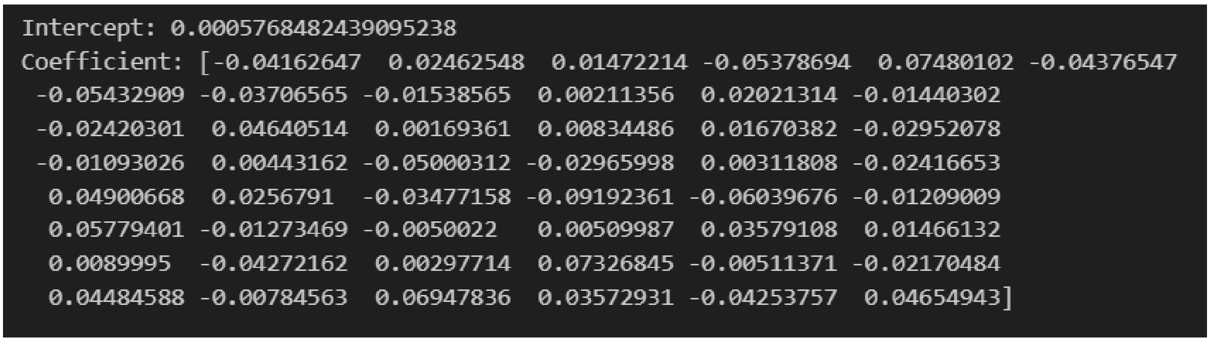
The bounds for the Autocorrelation are used to test the null hypothesis that an autocorrelation coefficient is 0. The null hypothesis is rejected if the sample autocorrelation is outside the bounds. The usual level of the test is 0.05, so one can expect to see about 1 out of 20 sample autocorrelations outside the test bounds simply by chance.

In the autocorrelation plot for daily returns of dow-jones industrial average index, we see it that it happens to be out of bounds. While we don't see that for monthly returns. This suggest that the daily returns series is not white noise.

We can also see the mean reverting nature from the autocorrelation plots, as the returns tend to oscillate around some fixed level. If the series wanders without returning repeatedly to some fixed level, then the series should not be modelled as a stationary process.

For Question 5, All daily returns for all the constituent elements were calculated. (Please refer to the code for more information).

The output for the linear regression is described as below –



```
Intercept: 0.0005768482439095238
Coefficient: [-0.04162647  0.02462548  0.01472214 -0.05378694  0.07480102 -0.04376547
 -0.05432909 -0.03706565 -0.01538565  0.00211356  0.02021314 -0.01440302
 -0.02420301  0.04640514  0.00169361  0.00834486  0.01670382 -0.02952078
 -0.01093026  0.00443162 -0.05000312 -0.02965998  0.00311808 -0.02416653
  0.04900668  0.0256791  -0.03477158 -0.09192361 -0.06039676 -0.01209009
  0.05779401 -0.01273469 -0.0050022   0.00509987  0.03579108  0.01466132
  0.0089995  -0.04272162  0.00297714  0.07326845 -0.00511371 -0.02170484
  0.04484588 -0.00784563  0.06947836  0.03572931 -0.04253757  0.04654943]
```

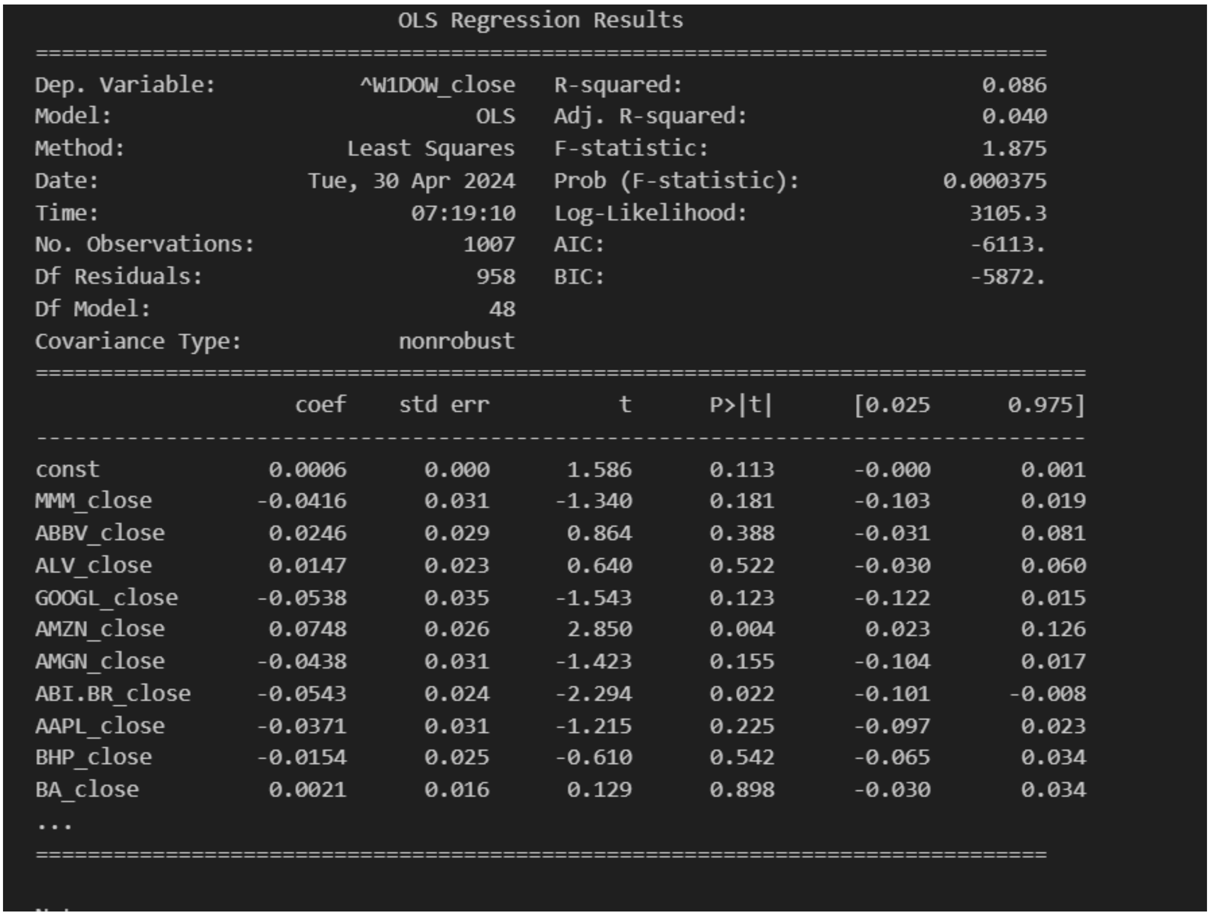
The coefficients are for the individual constituents.

Below is the list of some indexes which are sorted according to their coefficients-

[('AMZN_close', 0.07480101815653349), ('RY_close', 0.07326844898491602), ('V_close', 0.06947836159930447), ('NESN.SW_close', 0.05779400636546913), ('JNJ_close', 0.04900668353658187), ('DIS_close', 0.046549433799398035), ('CSCO_close', 0.046405135211922716), ('SIE.DE_close', 0.044845877778383954), ('PEP_close', 0.03579108092504038), ('7203.T_close', 0.035729307355126146), ('JPM_close', 0.025679101835676524), ('ABBV_close', 0.024625475701573512), ('BP_close', 0.02021313843790049), ('DD_close', 0.016703818574702285), ('ALV_close', 0.014722140472797858), ('PFE_close', 0.014661322911285826), ('PM_close', 0.008999499105116516), ('KO_close', 0.008344863750168319), ('ORCL_close', 0.005099870177774122), ('GE_close', 0.004431616954324817), ('INTC_close', 0.0031180791649885616), ('ROG_close', 0.0029771353625981513), ('BA_close', 0.002113556929379501), ('C_close', 0.0016936114026423733), ('NVDA_close', -0.005002204044181354), ('SMSN.L_close', -0.005113707728688258), ('TSM_close', -0.007845630220038348), ('META_close', -0.010930262570839679), ('MSFT_close', -0.012090086445346317), ('NOVN_close', -0.012734687841094938), ('BATS.L_close', -0.014403020415010053), ('BHP_close', -0.015385649856402062), ('SAN_close', -0.021704835654807), ('IBM_close', -0.02416652968655013), ('CVX_close', -0.024203013674792256), ('XOM_close', -0.029520783397979623), ('HSBA.L_close', -0.02965998197099175), ('MA_close', -0.034771581847966915), ('AAPL_close', -0.03706565013679959), ('MMM_close', -0.04162646563537461), ('WMT_close', -0.04253757362930558), ('PG_close', -0.04272162160858022), ('AMGN_close', -0.04376547116404851), ('GSK_close', -0.05000311503065762), ('GOOGL_close', -0.05378693783465815), ('ABI.BR_close', -0.05432908672335938), ('MRK_close', -0.0603967629137254), ('MCD_close', -0.09192361495066559)]

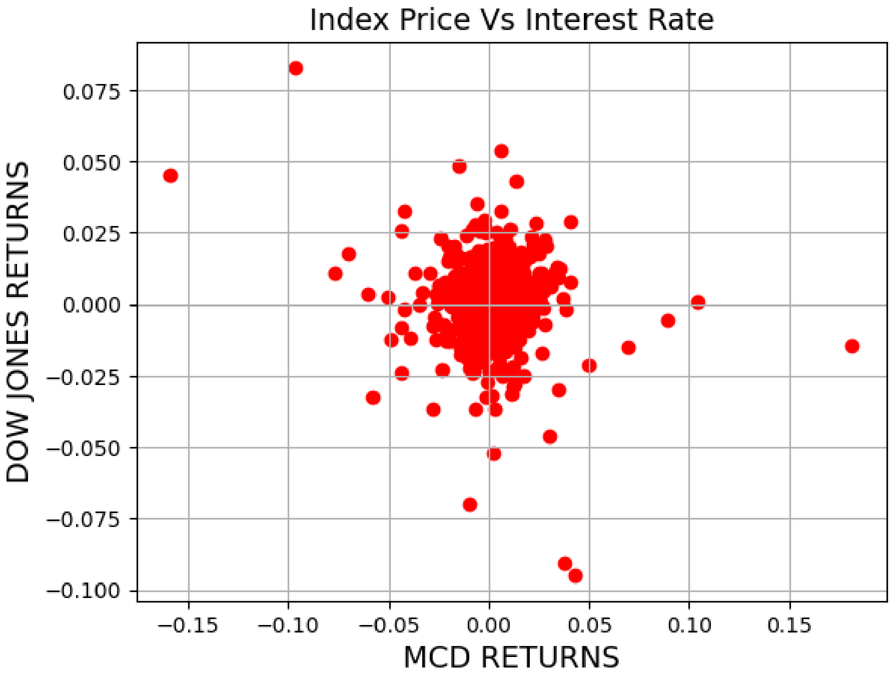
It can be seen that Amazon and Mcdonalds had one of the highest impacts on the dow jones industrial average. It makes sense because, during covid-19, businesses like amazon and mcd were highly affected due to lock down

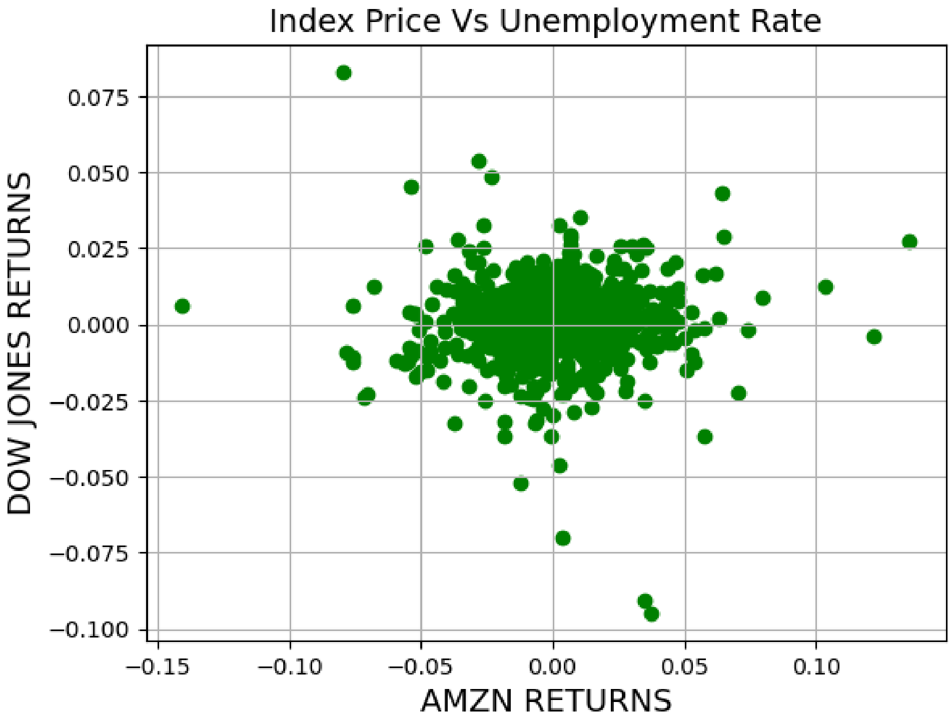
Below is an extra OLS analysis done for the dow jones index –



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            ^W1DOW_close   R-squared:                       0.086
Model:                             OLS   Adj. R-squared:                  0.040
Method:                  Least Squares   F-statistic:                     1.875
Date:                 Tue, 30 Apr 2024   Prob (F-statistic):           0.000375
Time:                         07:19:10   Log-Likelihood:                 3105.3
No. Observations:                 1007   AIC:                            -6113.
Df Residuals:                      958   BIC:                            -5872.
Df Model:                           48
Covariance Type:             nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.0006      0.000      1.586      0.113      -0.000       0.001
MMM_close       -0.0416      0.031     -1.340      0.181      -0.103       0.019
ABBV_close       0.0246      0.029      0.864      0.388      -0.031       0.081
ALV_close        0.0147      0.023      0.640      0.522      -0.030       0.060
GOOGL_close     -0.0538      0.035     -1.543      0.123      -0.122       0.015
AMZN_close       0.0748      0.026      2.850      0.004       0.023       0.126
AMGN_close      -0.0438      0.031     -1.423      0.155      -0.104       0.017
ABI.BR_close    -0.0543      0.024     -2.294      0.022      -0.101      -0.008
AAPL_close      -0.0371      0.031     -1.215      0.225      -0.097       0.023
BHP_close       -0.0154      0.025     -0.610      0.542      -0.065       0.034
BA_close         0.0021      0.016      0.129      0.898      -0.030       0.034
...
==============================================================================
```

Also, the residual plots for amazon and McDonald's –





When there are many potential predictor variables, often we wish to find a subset of them that provide a parsimonious regression model. It is more appropriate to use a model selection criterion such as AIC or BIC. The plots for these are available in the jupyter notebook.

Conclusion –

In the period of 4 years Dow jones industrial average saw a huge number of outlier events. The impact was majorly because of the Covid-19 pandemic and businesses like Amazon and McDonald’s, whose revenues were greatly impacted by the lockdown impacted dow jones index. Other enterprises can be referred to in the list above. I think best time to invest in the index was during the covid and some good strategies could have been generated referring to the autocorrelation plots.