# Analysis of Dow Jones Industrial Average Returns using ARIMA

**STAT 485: Applied Time Series Analysis**
**Simon Fraser University**

**Instructor –** Dr. Gary Parker
**TA –** Alex Wang

**Team 8**

**Student Name and Student ID:**

Inayat Lakhani - 301449154

Isabah Tabenda Hasan - 301435392

Romeesa Jabbar - 301361128

Vibhuti Gandhi - 301431087

**December 08, 2022**

# Executive Summary

**Goal:**

The main goal of our project, analysis of Dow Jones Industrial Average Returns using ARIMA, is to understand the stochastic trends of DJI and more forecasts.

**Method & Findings:**

After conducting an initial exploratory data analysis, we followed the Box-Jenkins approach and fitted an ARIMA(2,0,2) model through the DJI daily stock returns since it had the lowest AICc score amongst the different ARIMA models. We found that the data is stationary. In the Box Jenkins approach, we initially tried the ACF, PACF and EACF but they were not very helpful and we were not able to find appropriate parameters for AR and MA. Therefore, to determine values for the parameters of a model we resorted to using Maximum Likelihood Estimation and the Conditional sum-of-squares methods.

We noticed that the forecasts were failing to accurately capture the fluctuations in our data. This was due to the fact that the DJI daily returns are uncorrelated. We figured that this must be caused because of volatility clustering. We then modeled volatility using a GARCH model and made the case that DJI daily stock returns saw fluctuations at the same time window as spikes in volatility. Finally, to implement ARIMA modeling we introduce correlation by taking the square of the DJI daily stock returns.

We were then able to fit an ARIMA (4,1,2) model by following the Box-Jenkins approach again. We compared and contrasted volatility in the ARIMA model with returns and ARIMA with returns squared model and found out that the latter were able to get better forecasts (though they were only marginally better).

**Limitations:**

Further analysis can be conducted using ARCH/GARCH models instead of ARIMA models as they are better suited to modeling financial data. They can capture the fluctuations and the heteroskedasticity in the stock market to a greater degree than ARIMA models. However, these models are beyond the scope of this report.

**Conclusion:**

Since forecasts generated by ARIMA(4,1,2) are not significantly better than those of ARIMA(2,0,2) and following the principle of parsimony, we decide to stick with ARIMA(2,0,2) for our analysis.

# Report

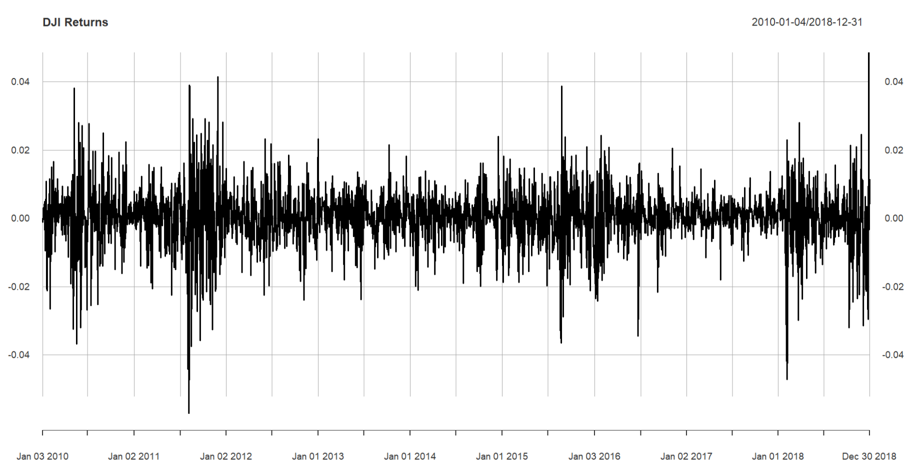
# 1 Introduction

## 1.1 Background

The Dow Jones Industrial Average (DJI), also known as the Dow 30, is the second oldest stock market index in the US and the most watched one in the world (Ganti, 2022). It is a price-weighted measure of 30 public-owned U.S. blue-chip companies with consistent stable earnings trade on the New York Stock Exchange and Nasdaq (Ganti, 2022). The selected companies cover all industries except utilities and transportation. It is a measure that helps investors determine the overall direction of the stock market.

## 1.2 Data Preparation

We are using the data set of Dow Jones Industrial Average (DJI) and sourcing it from Yahoo! Finance. We have used the stock data from January 1, 2010 to January 1, 2019. We have not included the period in which the major recession and the pandemic took place to capture a stable period. We calculate the daily returns by taking the difference of the log values of DJI close prices. Below is the time series plot of DJI daily returns over time.




*(Appendix A, Figure 1)*

## 1.3 Goal

Our goal for this project is to:

- Understand the stochastic trends of daily returns of Dow Jones Industrial Average stock.
- Fit a suitable model to fit the daily returns using the Box-Jenkins approach and make predictions.
- Visualize volatility across the time series using a GARCH model.
- Transform the data to make better predictions.

## 2 Exploratory Data Analysis

- We analyzed the daily returns and found that our data has rapid fluctuations and displays non-seasonality since it shows no clear patterns or trends.
- After our exploratory data analysis, we concluded that our dataset plots have a non-zero mean and a non constant variance. The value of the mean is around 0.0003546006 which is very close to zero. *(Appendix A, Figure 1)*
- The histograms follow a normal distribution pattern since it is symmetric. *(Appendix A, Figure 2)*
- The boxplot shows a symmetric distribution as the median is in the middle of the box. However, there are multiple occurrences of outliers. *(Appendix A, Figure 3)*
- The violin plot shows the symmetric distribution as the histogram, the wider section displays the high probability and thinner sections show low probability of observations. *(Appendix A, Figure 4)*
- According to the Q-Q plot, the data is normally distributed as most of the data points fall on the 45-degree reference line but these data points deviate at the tails. The deviation is symmetric with fat tails and excess kurtosis which suggest that more data is distributed towards the extremes of the distribution and less data is distributed towards the center. *(Appendix A, Figure 5)*
- The correlation values of Yt vs Yt-1 = -0.03577669, Yt vs Yt-2 = 0.009759712, Yt vs Yt-3 = -0.02361557. We observe that our data has very weak correlation. *(Appendix A, figure 6)*

## 3 Methodology

Based on the figures and values computed on the Arima Modelling R script, we inferred different graphs and plots to find statistical significance and make conclusions. All graphs, tables and plots have been added to our appendix.

### 3.1 Testing for Stationarity

Based on the Box-Jenkins approach, the time series has to be stationary. A stationary time series is without a trend, the mean and variance are constant. Hence, the behavior of the stationary process does not change over time. Stationarity detection is based on visualizing the plots and graphs. Plots that have prominent seasonality and increasing variance are non-stationary. The DJI returns have visible patterns or trends. We ran the Augmented Dickey Fuller test to corroborate our findings and they matched. Hence, we infer that our data is stationary *(Appendix A, Figure 1)*

### 3.2 Differencing

Differencing is a way to change non-stationary time series to a stationary time series by computing the differences between consecutive observations of the time series. We found out in 3.1 that our dataset is stationary so we would not require to use a differencing method on it. Therefore, the d term in our ARIMA(p,d,q) would be 0.

### 3.3 Analysis of PACF plot

p is the number of autoregressive terms in an Auto Regressive (AR) model and the function (PACF) can be used to visualize it through a plot.This can be done by studying the PACF plot and trying to find if there is any significant spikes that surpass the 95% confidence interval mark in the plot. With the spikes in respective lags, the number p can be determined. In the initial lag there is little correlation. The PACF plot of DJI daily returns show significant spikes in lag 5, 14, 19 and 25. We do not see an exponential decay. *(Appendix A, Figure 7)*

### 3.4 Analysis of ACF plot

q is the number of lagged forecast errors in Moving Average (MA) model and the function Autocorrelation Function (ACF) can be used to visualize it through a plot. In order to determine the order of q, the same method as above can be applied. Any significant spike has to be examined in the ACF plot. This plot also has little correlation in the initial lag. If a spike is discovered in respective lags, the order will represent that number. The ACF plot of DJI daily returns show significant spikes in lag 5, 14 and 25. We do not see an exponential decay. *(Appendix A, Figure 7)*

### 3.5 Analysis of EACF plot

The ACF and PACF are not very productive mediums to verify ARIMA/ARMA models. After a definite number of lags, neither the true ACF nor the true PACF will be cut off completely. Hence, the extended autocorrelation function (EACF) is one method suggested to estimate the orders of an ARIMA(p,d,q) model. The EACF table shows a triangular pattern of zeros with the top left zero occurring in the p-th row and q-th column (with row and column labels starting from 0). In the EACF table, corresponding to p and q in the first row and column, (0,0,0) can be spotted. Hence, The sample EACF also suggests that a white noise model is suitable for our data. *(Appendix A, Figure 8)*

### 3.6 Model Fitting

The ACF, PACF and EACF were not very helpful and we were not able to find appropriate parameters for AR and MA. To determine values for the parameters of a model we use the Maximum Likelihood Estimation(MLE) and the Conditional sum-of-squares methods.

MLE is the estimation method that determines values of the model parameters that maximize the likelihood of the statistical model (Barlett, 2018). Unlike the least squares, all of the information is used and is generalizable to some extent. It acts as a joint probability density function where observed data is fixed and unknown parameter values are yet to be obtained (Chan & Cryer).

Conditional sum-of-squares (CSS) is the method that determines the sum of squares of the fitted innovations obtained from the maximum lag which helps to compare different fits. The previous innovations are then considered zero (Chan & Cryer).

After fitting different ARIMA models, we arrived at ARIMA(2,0,2) *(Appendix A, figure 12)* since it had the lowest AICc score. Our model has the following parameters - AR1 = 0.2233, AR2 = 0.7267, ma1=-0.2627, ma2= -0.7144. The variance of the residuals is 8.023e-05, AIC scores =-14918.9, AICc score =-14918.86 , BIC score =-14884.55. *( Appendix A, Figure 10)*

### 3.7 Model Diagnostics

We used R script to conduct model diagnostics of the ARIMA (2,0,2) by analyzing the residuals.

- Plot of residuals over time does not present any seasonality/trends. This shows that it resonates more with a white noise process as the variables are uncorrelated. *(Appendix A, Figure 11)*
- The ACF plots show lags at 5, 14 and 25. There are significant lags and portray the unpredictable nature of the stock market. This leads to the failure of white noise assumption. *(Appendix A, figure 7)*
- The residuals (and hence the error terms) of the histogram are normally distributed with approximately mean 0 and standard deviation 1. *(Appendix A, figure 11)*

### 3.8 Forecasting

We forecasted ARIMA(2,0,2) and we concluded that our predictions were not significantly different from each other and they failed to capture the fluctuations of daily returns. *(Appendix A, figure 13)*

## 4. Volatility Clustering

Benoit B. Mandelbrot described Volatility Clustering as financial data that tends to follow the trend that "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes." (Mandelbrot, 1967) As volatility tends to cluster, a low volatility day in the present can signal a low volatility day with a better performance also in the future - these effects can compound over months and years. This can be visualized by the rolling 1-month and 3-month estimate of annualized volatility *(Appendix B, figure 1).* This clustering effect can have a significant impact on evaluating DJI as a favorable stock to invest in.

As we discussed before the DJI daily returns are uncorrelated and have a non-constant variance. These factors allude to the fact that there exists volatility in our data. To visualize this volatility we trained a Generalized AutoRegressive Conditional Heteroskedasticity (GARCH(1,1)) variance series. *(Appendix B, figure 2)* Upon analyzing the DJI returns volatility we see that DJI daily stock returns see fluctuations at the same time-windows as spikes in volatility. We used the coefficients from the GARCH(1,1) model to plot bands of conditional standard deviation along with deviations from its mean. This helps corroborate our findings in the previous sections and explains how the existence of volatility is causing our ARIMA(2,0,2) model to fail to accurately capture the fluctuations in the DJI returns. *(Appendix B, figure 2)*

# 5. ARIMA Modelling with adjusted data
## 5.1 Transforming the data
We found that the predictions made by ARIMA(2,0,2) model failed to capture the fluctuations in daily returns due to volatility clustering. Therefore, we transformed the data by taking the square of daily returns. Since the returns themselves are uncorrelated, taking the square of returns allows us to introduce correlation (Granger & Ding, 1995). *(Appendix C, figure 1)*
## 5.2 Exploratory Data Analysis
- The histogram is no longer normally distributed. It is skewed to the right as the peak of the graph lies to the left of the center. *(Appendix C, figure 2)*
- The box plot clearly indicates the skewness of the data. The distribution is right skewed and positive. There are a huge number of outliers affecting the overall time series. *(Appendix C, figure 3)*
- The Q-Q plot shows deviation from the straight line at the upper-end while the lower end follows the line. The curve has a longer tail to the right making it positively right skewed. *(Appendix C, figure 5)*
- The Violin plot shows a wider section to the left representing higher probability of observations while the thinner sections to the right represents a lower probability. *(Appendix C, figure 4)*
- The correlation values of Yt vs Yt-1 = 0.2536075, Yt vs Yt-2 = 0.3510526, Yt vs Yt-3 = 0.2401968. We observe that our data has better correlation than before. *(Appendix C, figure 6)*

## 5.3 Stationarity and Differencing
We adjusted the new dataset and tested for stationarity and differencing. Similar to the previous dataset, we also did analysis of PACF, ACF, EACF as well as model fitting, model diagnostics and forecasting. We followed the similar approach as before but came up with a new ARIMA(4,1,2) model that makes better predictions. We inferred by visually interpreting the plot that our initial ARIMA (2,0,2) model was stationary. However, our new model ARIMA (4,1,2) is not stationary with noticeable trends *(Appendix C, figure 11).* Since the data is not stationary, differencing method would be applied to the series. In this case the differencing order is 1. This difference here infers the sequence of changes from one period to the next period.

## 5.4 Analysis of PACF, ACF & EACF
The ACF and PACF plots are much more decipherable compared to the previous plots. Both ACF and PACF are gradually decaying with lags. However, for both the plots tailing off or cutting off is still not too clear. It can be observed that the correlation at initial lags are significant and positive. The ACF compared to the PACF has a lot more significant spikes. Gradual decaying can be observed for ACF. Significant spikes in lags 1,2,3,4 can be observed in the PACF plot initially for the order of 4 for p. Observing the EACF table outputs, ARIMA(2,1,3), (4,1,2), (3,1,2) seem appropriate. *(Appendix C, figure 7 & 8)*

### 5.5 Model Fitting

To fit our ARIMA model, we used both ML and CSS again. We arrived at ARIMA(4,1,2). *(Appendix C, figure 11)* Our model has the following parameters - AR1 = -0.7364, AR2 = 0.3084, MA1= -0.1358, MA2= -0.8066. The variance of the residuals is 3.213e-08, AIC scores = -32612.66, AICc score = -32612.61, BIC score =-32572.58. *( Appendix C , Figure 9)*

### 5.6 Model Diagnostics

We used R script to conduct model diagnostics of the ARIMA (4,1,2) by analyzing the residuals.

- Plot of residuals over time does not present any seasonality/trends. This shows that it resonates more with a white noise process as the variables are uncorrelated. *(Appendix C, figure 10)*
- The ACF plots show significant 16 and 23 lags but compared to the previous ACF plot of residuals by ARIMA(2,0,2) on daily returns, the significant lags are much smaller. *(Appendix C, figure 7)*
- The residuals (and hence the error terms) of the histogram look like they are distributed with a slight right skew which could be ignored. *(Appendix C, figure 10)*

### 5.7 Forecasting

Our forecast initially was failing to capture the huge fluctuations in the data due to volatility clustering. Therefore, we adjusted the data to make better predictions with ARIMA (4,1,2). The new forecasts do a slightly better job at capturing the volatility of the time series in comparison to our previous forecast. The predictions of the forecast have changed and we do not see a flat line anymore. *(Appendix C, figure 12)*

## 6. Conclusions

We tried to fit ARIMA (2,0,2) but it failed to capture the predictions. It didn't provide good predictions due to volatility clustering and non-constant variance. We then transformed the data using the squared daily returns of DJI with ARIMA (4,1,2) model. Our predictions got slightly better. The principle of parsimony states that a simpler model with fewer parameters is favored over more complex models with more parameters, provided the models fit the data similarly well ("Parsimony principle", n.d.) Since forecasts generated by ARIMA(4,1,2) are not significantly better than those of ARIMA(2,0,2), we decide to stick with ARIMA(2,0,2) for our analysis. Further analysis can be conducted using ARCH/GARCH models as they are better suited to modeling financial data. They can capture the fluctuations and the heteroskedasticity in the stock market to a greater degree than ARIMA models. However, we understand that ARCH/GARCH models are beyond the scope of this report.

**APPENDIX A (Daily returns):**

**Figure 1:** *The graph below shows the time series of the DJI daily returns from January 1, 2010 to January 1, 2019.*

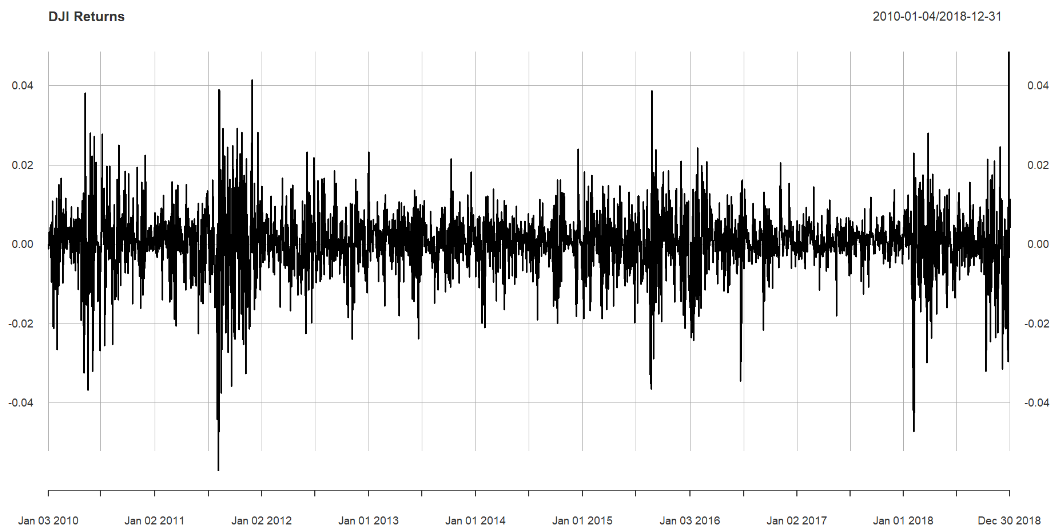



**Figure 2:** *The histogram below represents the distributions of DJI daily returns.*

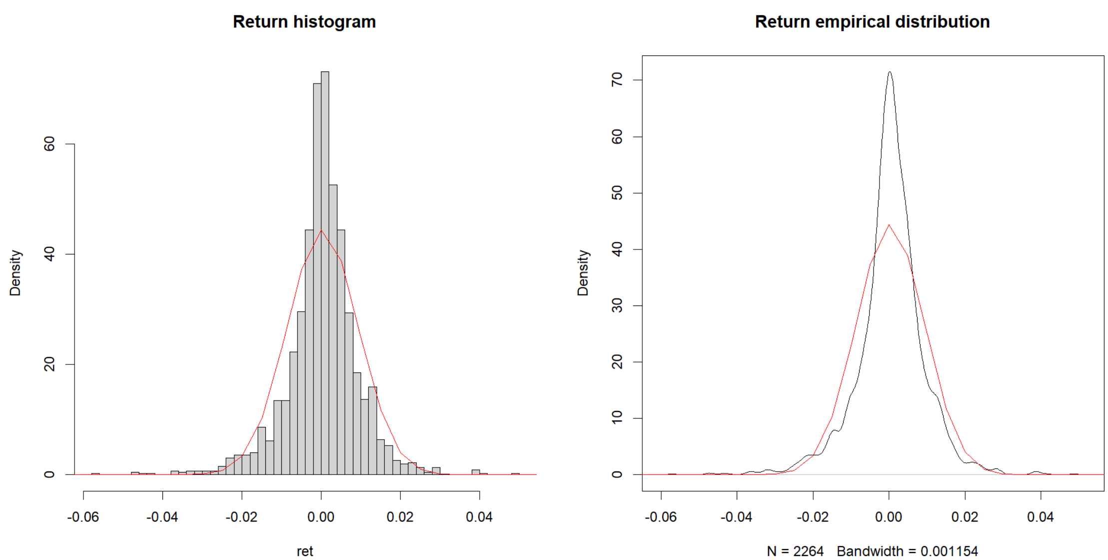
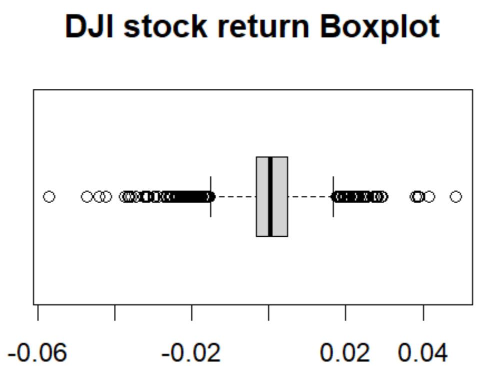


**Figure 3:** *The boxplot on the left shows the distribution of DJI daily returns.*
**Figure 4:** *The violin plot on the right shows the distribution of DJI daily returns.*



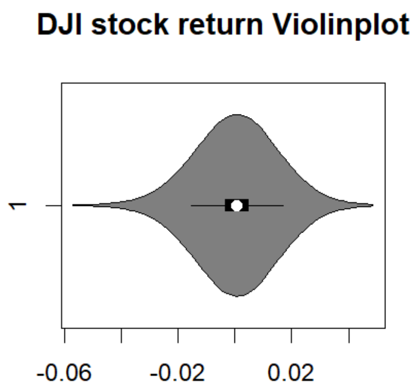




**Figure 5:** *The Q-Q plot below shows the normality of the distribution of DJI daily returns.*

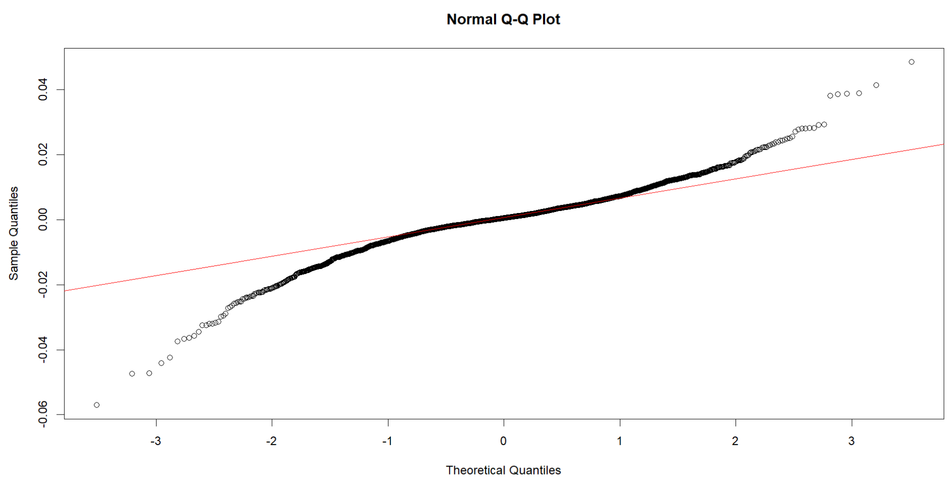



**Figure 6:** *These three plots below show the correlation between Yt and Yt-1, Yt-2, Yt-3 for DJI daily returns.*

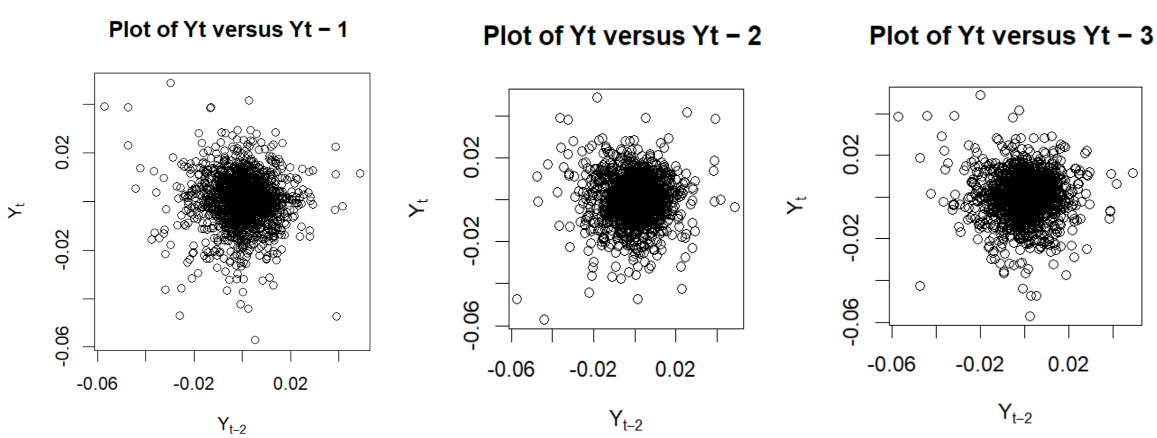


**Figure 7:** *The plot on left shows the ACF and on the right shows the PACF of DJI daily returns.*

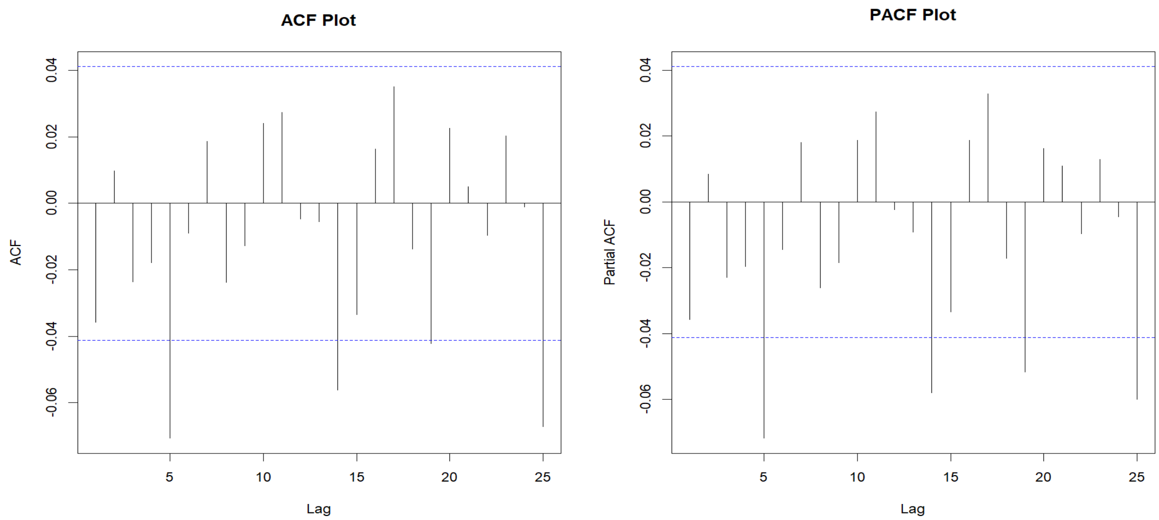



**Figure 8:** *The table below shows the EACF of DJI daily returns.*

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 o o o o x o o o o o o  o  o  x
1 x o o o x o o o o o o  o  o  o
2 x x o o x o o o o o o  o  o  o
3 x x x o x o o o o o o  o  o  o
4 x x x x x o o o o o o  o  o  o
5 x x x o o o o x o o o  o  o  o
6 x o x x o x o o o o o  o  o  o
7 x o x o o x o o o o o  o  o  o
```

**Figure 9:** *The table below displays different ARIMA models and their AICc scores respectively.*

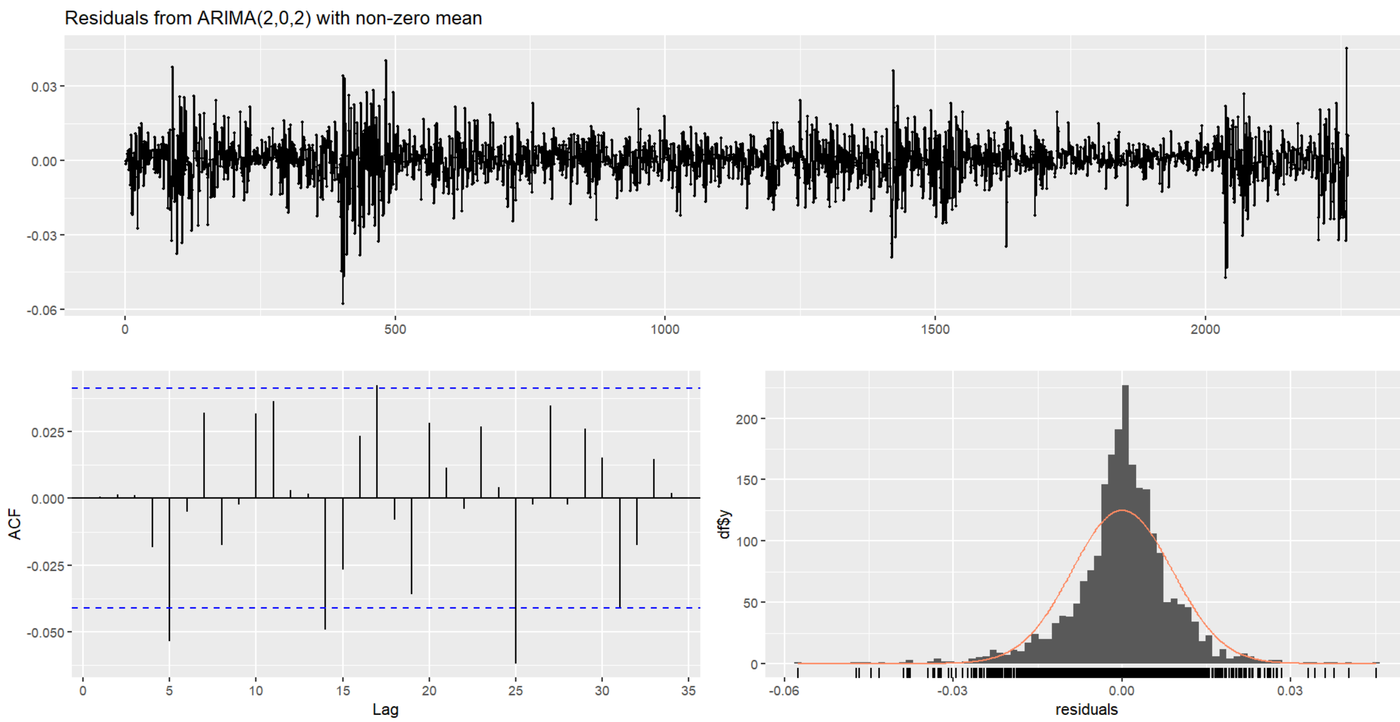
| ***Model*** | ***AICc Score*** |
|---|---|
| ARIMA(2,0,2) with non-zero mean | -14918.86 |
| ARIMA(0,0,0) with non-zero mean | -14914.29 |
| ARIMA(1,0,0) with non-zero mean | -14915.18 |
| ARIMA(0,0,1) with non-zero mean | -14915.13 |
| ARIMA(3,0,1) with non-zero mean | -14918.16 |
| ARIMA(3,0,2) with non-zero mean | -14916.9 |
| ARIMA(2,0,3) with non-zero mean | -14916.92 |
| ARIMA(1,0,3) with non-zero mean | -14918.13 |

**<u>Figure 10:</u>** *Model Parameters for ARIMA(2,0,2) Model:*

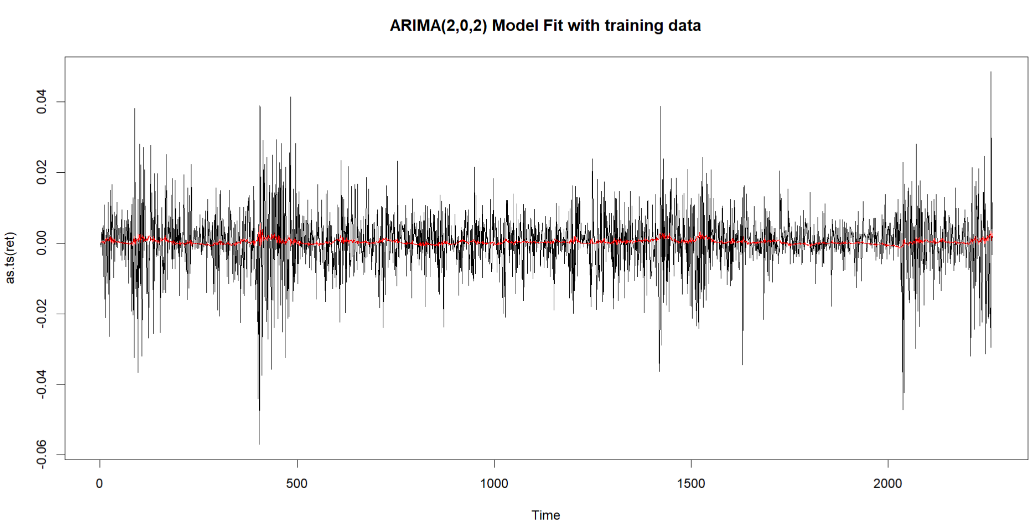
```
Coefficients:
         ar1     ar2      ma1      ma2   mean
      0.2233  0.7267  -0.2627  -0.7144  4e-04
s.e.  0.1950  0.1862   0.2029   0.1982  1e-04

sigma^2 = 8.023e-05:  log likelihood = 7465.45
AIC=-14918.9   AICc=-14918.86   BIC=-14884.55
```

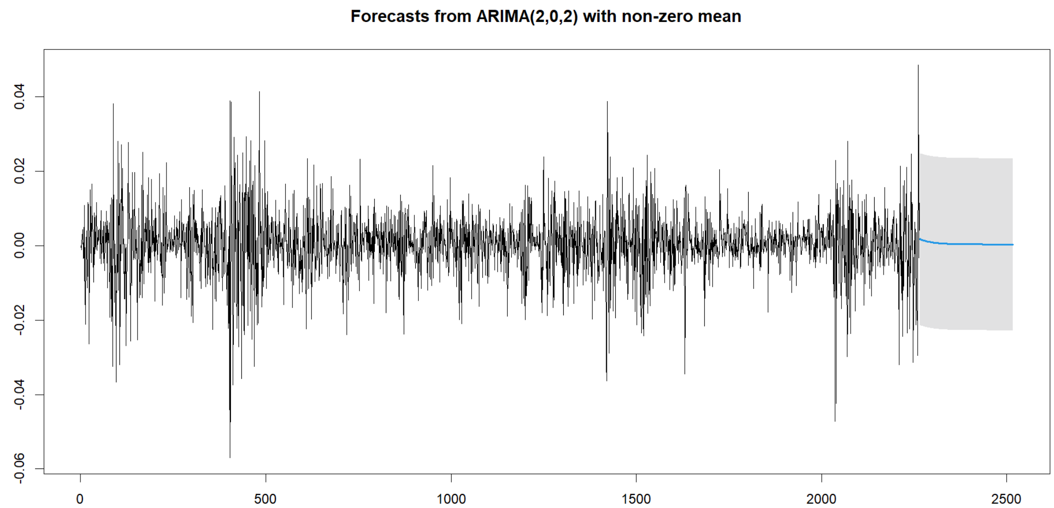
**<u>Figure 11:</u>** *The plot below shows residual analysis of ARIMA(2,0,2) of DJI daily returns.*




**<u>Figure 12:</u>** *The plot shows model fitting of training data with ARIMA(2,0,2) of DJI daily returns.*

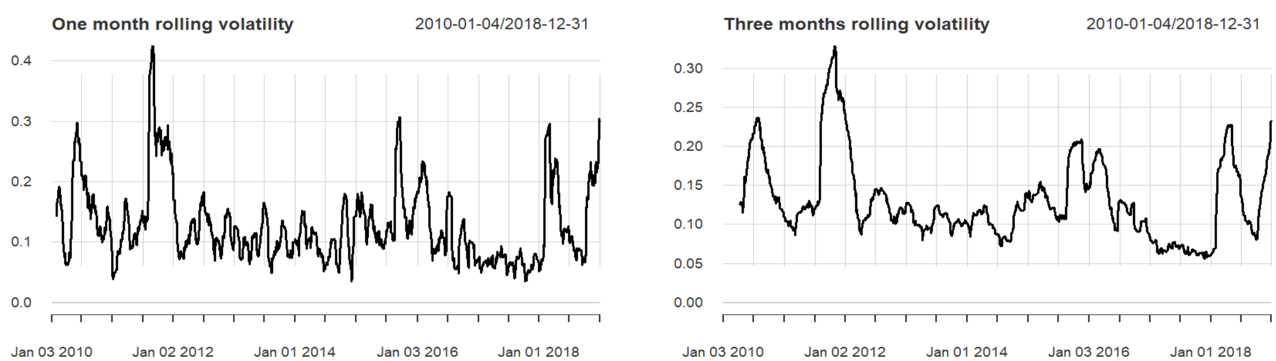


**<u>Figure 13:</u>** *The plot below shows forecasts of ARIMA(2,0,2) of DJI daily returns.*




## APPENDIX B (Volatility Clustering):

**<u>Figure 1:</u>** *The plots below show the rolling 1 month and 3 month estimate of annualized volatility of DJI daily returns.*




**<u>Figure 2:</u>** *The following plot shows a variance series using the GARCH(1,1) model.*

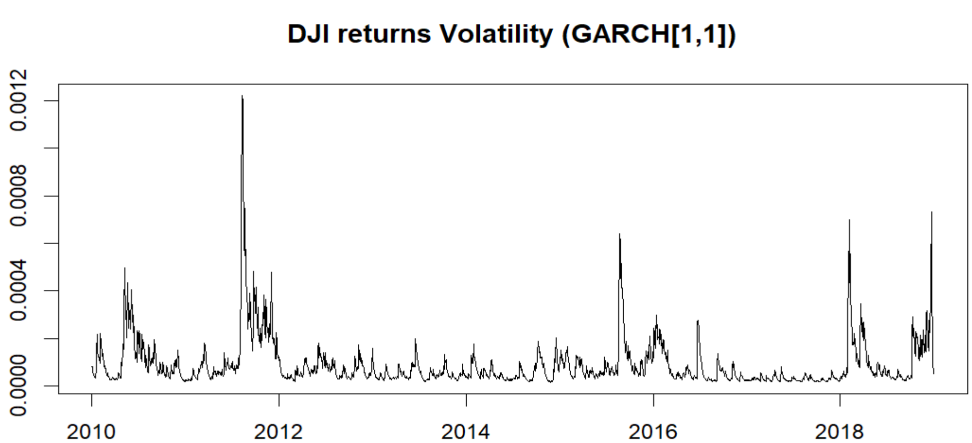
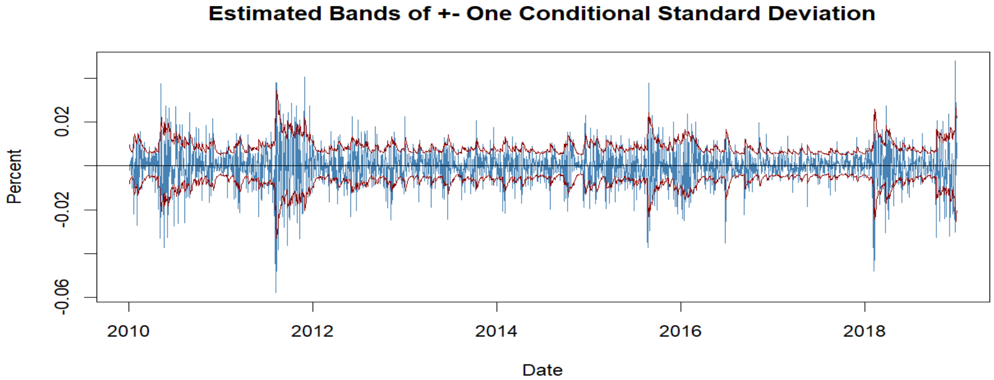


**Figure 3:** *The following plot shows bands of conditional standard deviation along the DJI daily returns*




**APPENDIX C (Squared Daily Returns):**

**Figure 1:** *The graph below shows the time series of the DJI daily returns from January 1, 2010 to January 1, 2019.*

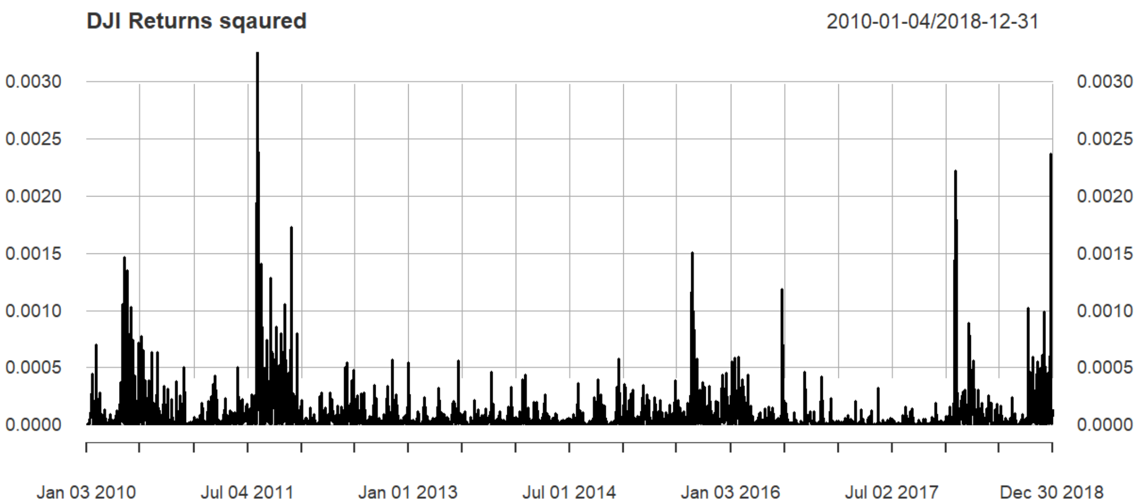



**Figure 2:** *The histogram below represents the distributions of squared DJI daily returns*

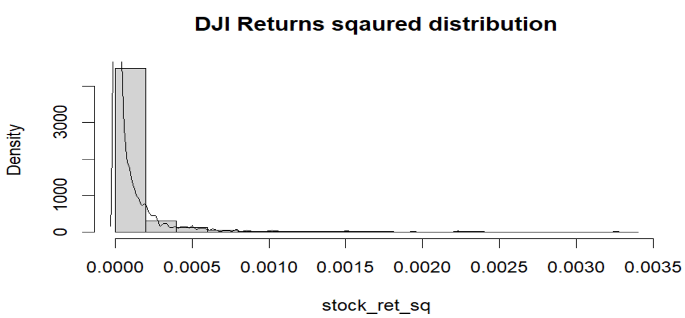
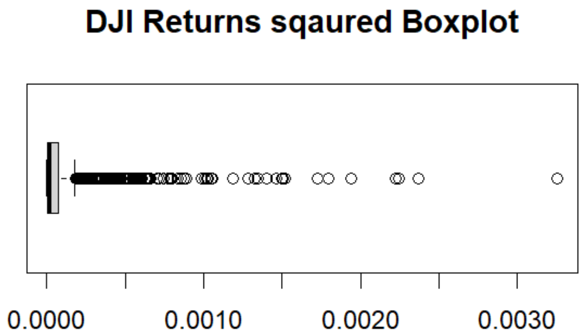


**Figure 3:** *The boxplot on the left shows the distribution of  squared DJI daily returns.*
**Figure 4:** *The violin plot on the right shows the distribution of squared DJI daily returns*



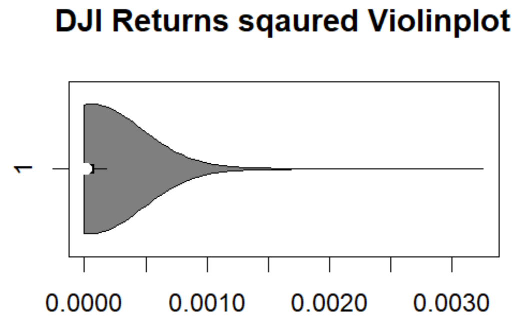
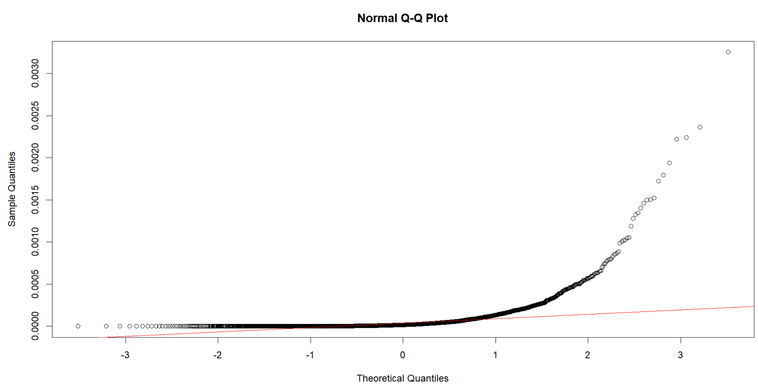
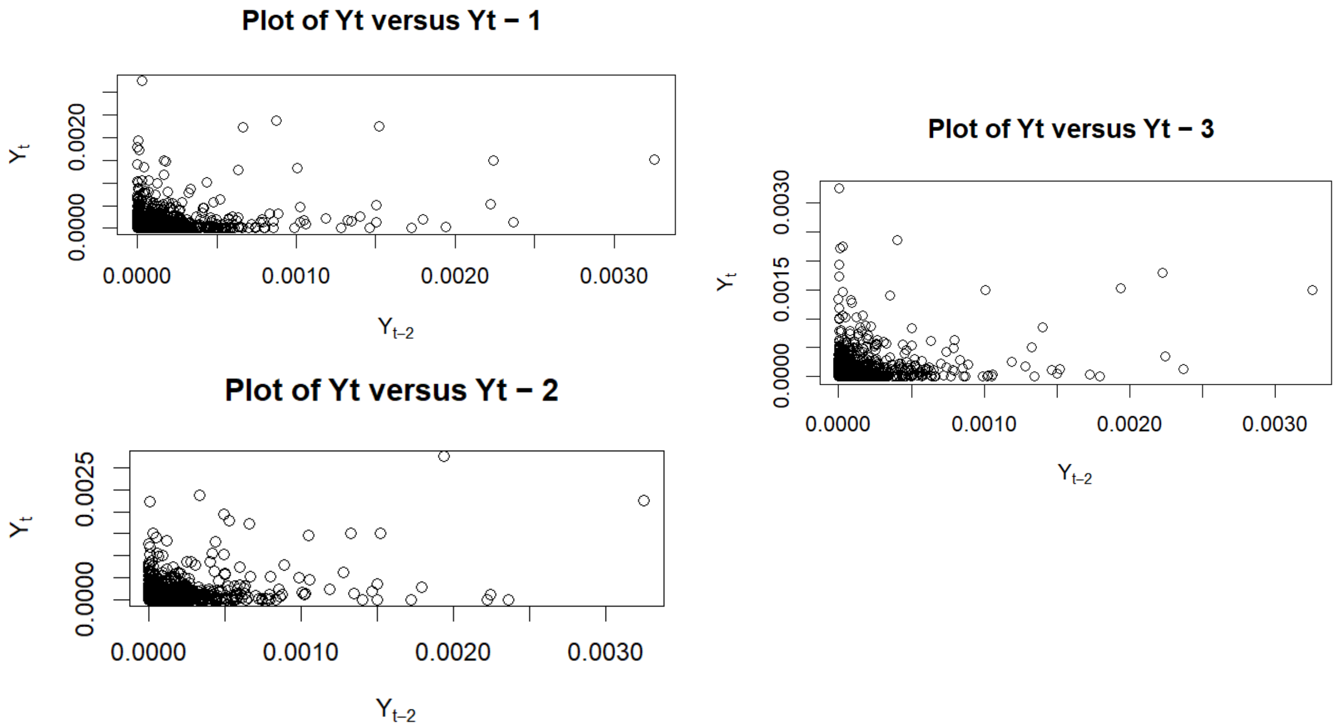



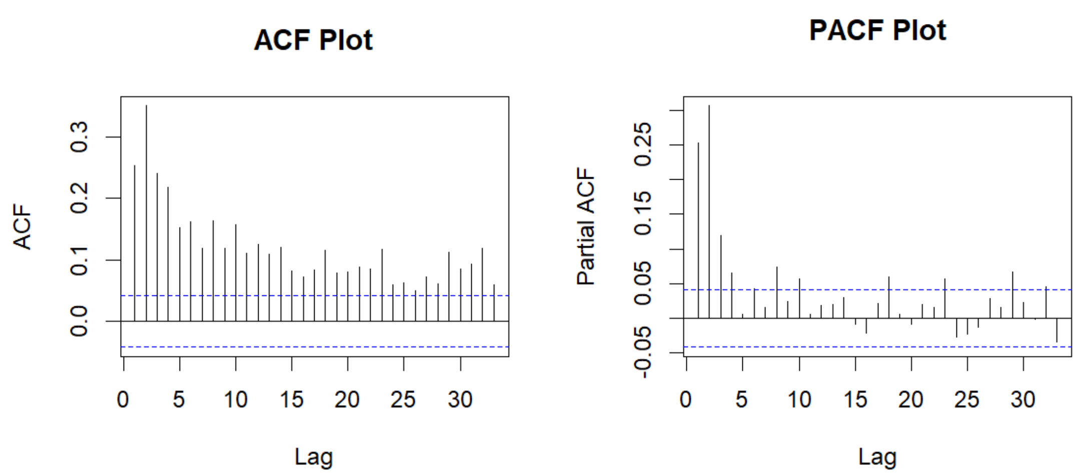

**Figure 5:** *The Q-Q plot shows the normality of the distribution of squared DJI daily returns.*



**<u>Figure 6:</u>** The three plots below show the correlation between Yt and Yt-1, Yt-2, Yt-3 *for DJI returns squared.*




**<u>Figure 7:</u>** *The plot on left shows the ACF and on the right shows the PACF of DJI daily returns squared.*



**<u>Figure 8:</u>** *The table below shows the EACF of DJI daily returns squared.*

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x x x x x x x x x x o  o  o  o
2 x x x o o o o o o o o  o  o  o
3 x x o x o o o o o o o  o  o  o
4 x x x o o o o o o o o  o  o  o
5 x x x x o x o o o o o  o  o  o
6 x x o x o x o o o o o  o  o  o
7 x x x x x o o o o o o  o  o  o
```

**<u>Figure 9:</u>** *Model Parameters for ARIMA(4,1,2) Model:*

```
Best model: ARIMA(4,1,2)
Series: stock_ret_sq
ARIMA(4,1,2)
Coefficients:
          ar1     ar2     ar3     ar4      ma1      ma2
      -0.7364  0.3084  0.2706  0.0987  -0.1358  -0.8066
s.e.   0.0928  0.0319  0.0366  0.0226   0.0912   0.0870

sigma^2 = 3.213e-08:  log likelihood = 16313.33
AIC=-32612.66   AICc=-32612.61   BIC=-32572.58
```

**<u>Figure 10:</u>** *The plot below shows residual analysis of ARIMA(4,1,2) DJI daily returns squared.*

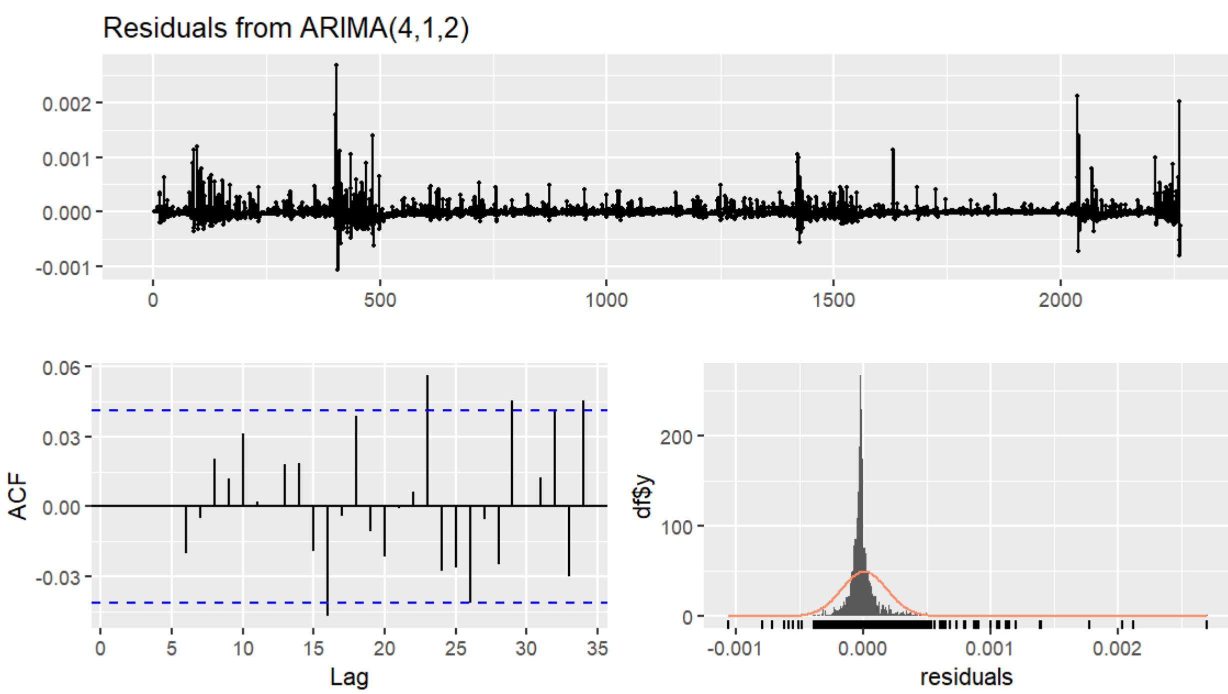


**Figure 11:***The plot shows model fitting of training data of ARIMA(4,1,2) DJI daily returns squared.*

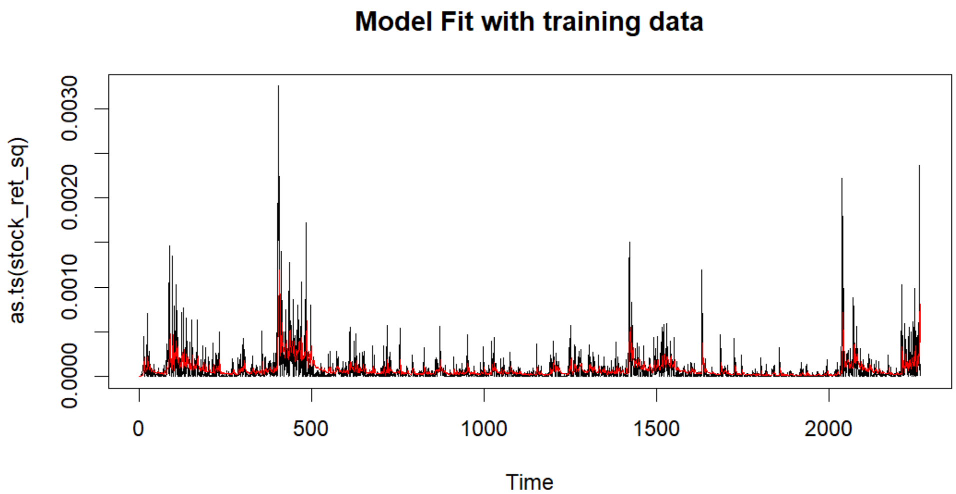
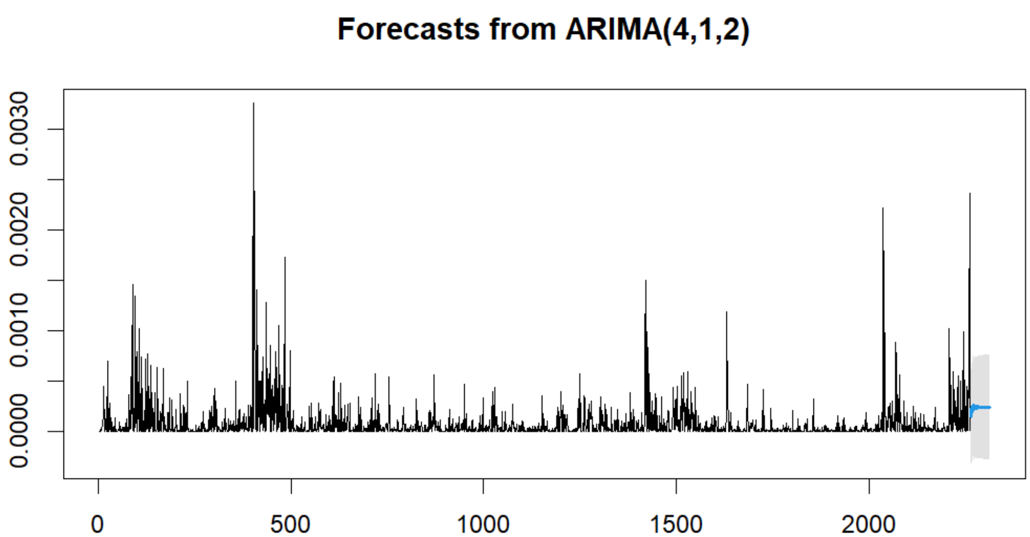



**Figure 12:** *The plot below shows forecasts of ARIMA(4,1,2) of DJI daily returns squared.*



## References:


Barlett, J.B. (January 3, 2018). *Probability concepts explained: Maximum likelihood estimation.* Medium. https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

Cryer, J. D., & Chan, K.-S. (n.d.). *Time Series Analysis*. Springer New York. https://doi.org/10.1007/978-0-387-75959-3

Ganti, A. (October 5, 2022). *What Is the Dow Jones Industrial Average (DJIA)?* Investopedia. https://www.investopedia.com/terms/d/djia.asp

Granger, C. W. J., & Ding, Z. (1995). Some Properties of Absolute Return: An Alternative Measure of Risk. *Annales d'Économie et de Statistique*, *40*, 67–91. https://doi.org/10.2307/20076016

Mandelbrot, B. (1967). The Variation of Some Other Speculative Prices. *The Journal of Business*, *40*(4), 393–413. http://www.jstor.org/stable/2351623

*Parsimony principle*. (n.d.) ClubVita. https://www.clubvita.net/glossary/parsimony-principle#:~:text=The%20parsimony%20principle%20for%20a,fit%20the%20data%20similarly%20well