Financial modelling and analysis of EBAY closing stock prices from year 2013 to 2018

Project on Modelling Financial Data

MM905

Sonal Jain
201961083
MSc Data Analytics

MM905: Project on Modelling Financial Data

Table of Contents

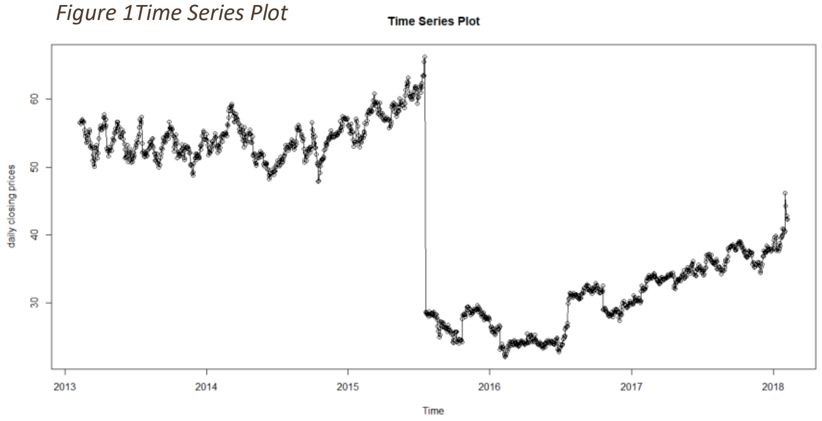
Table of Figures:

Introduction:

The aim of this project is to do time series analysis and modelling of financial data. The financial data chosen for this report is the closing stock prices of stock of company EBAY taken from S&P500 dataset from Kaggle- https://www.kaggle.com/camnugent/sandp500.

PART 1: Stationarity of time series data

Figure 1Time Series Plot



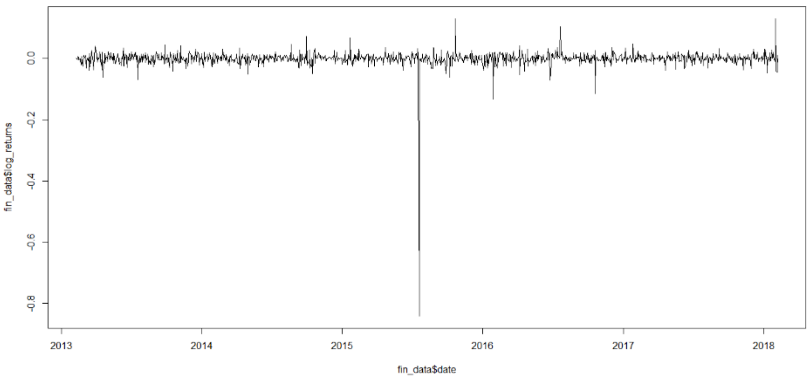
The best way to study the stationarity of a time series data is by plotting it. The graph alongside shows the daily closing stock prices of the stock EBAY from year 2013 to 2018. The distribution clearly shows that the data is not stationary as its mean and variance varies at different points, in other words the mean and variance of the distribution is not constant.

Another very important method of checking the stationarity of data is by tests for white noise. "Ljung-Box" test which is a one-tail statistical test was conducted on this dataset and p value was found to be smaller than 0.05. 0.05 here is the significance level or alpha which leads to rejecting the null hypothesis of non- rejection of white noise, in other words this data distribution is not white noise.

Log Returns:

To do a forecast it is essential to have some similarity between historic data and future. The above data is not suitable as the mean changes over time hence it needs to be transformed into log returns. Below plot shows the distribution of log returns of the data. It can be seen that the mean is constant overtime (except for the drop between year 2015 and 2016, but despite this drop the mean remains constant throughout the distribution) and there is invariance in the data. Log returns are defined by –

Figure 3 Log Returns



$$r_t = \log P_t - \log P_{t-1} = \log (1 + P_t - P_{t-1} / P_{t-1})$$

where P are closing stock prices and t is time.

"Ljung-Box" test of white noise was also performed on log returns. The p-value was found to be smaller than the significant level of 0.05 and hence the null hypothesis was rejected or in other words the distribution fails the white noise test at 0.05 level of significance. Hence, from looking at both the above mentioned distributions it can be inferred that the data is not stationary.

PART 2: Distribution of Returns and Value At Risk.

Test for Normality of closing stock prices

To study the distribution of returns we perform tests for normality. Mean of the closing prices of the stock EBAY and standard deviation were found to be greater than zero. To test normality of returns, Kolmogorov-Smirnov Test is used to generate a p-value. The p-value for this distribution is found to be smaller than 0.05 which means the distribution of the returns of EBAY closing stock price is significantly different from normal distribution under level 0.05. This is also confirmed by skewness and kurtosis. The value for skewness is negative and kurtosis is not equal to 3 which shows that the data does not follow a normal distribution as it does not meet the conditions required for a normal distribution. In practice it is difficult to get a normal distribution but characteristics such mean, variance, skewness, kurtosis are sufficient for analysis.

Test of Normality of log returns.

Similar approach as that mentioned above was used to test the normality of log returns and similar conclusions were drawn. The mean and variance were greater than zero, the skewness was negative, the kurtosis was greater than 3 and the Kolmogorov-Smirnov Test performed on log returns gave the p-value smaller than 0.05 which means the distribution of the returns of EBAY closing stock price is significantly different from normal distribution under level 0.05. The distribution of log returns does not follow normal distribution.

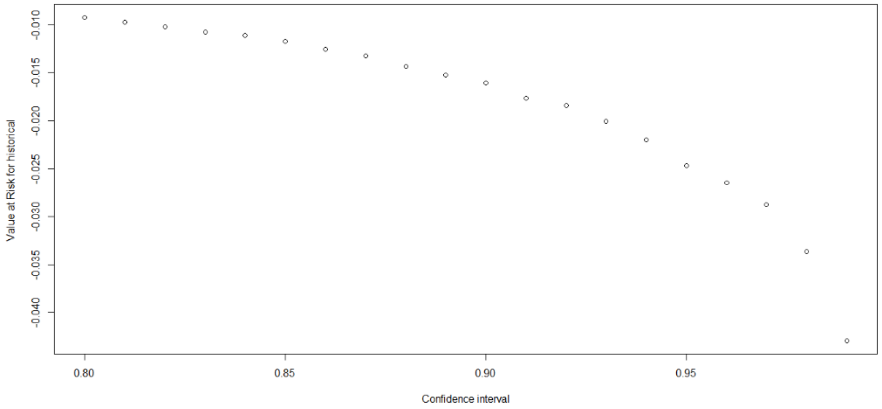
Value At Risk:

Value at Risk for different confidence intervals were calculated using the library called 'Performance Analytics' in R. Historical and Gaussian methods were used at confidence intervals ranging from 80 to 99%.

• Historical (non-parametric) method

In a set of returns for which sufficiently long history exists, the per-period Value at Risk is simply the quantile of the period negative returns: VaR = quantile(−r, p) = −quantile(r, 1 − p). This method it is based on the empirical distribution of the returns.

Figure 5 VaR using Historical Method

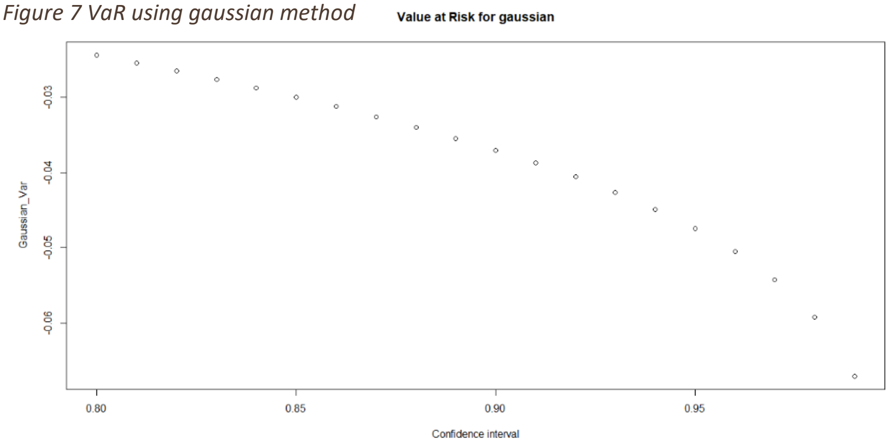


The negative signs of Value at risk denote the amount of loss. The plot alongside shows the value at risk at various confidence intervals while using Historical non parametric method. It can be said that Value At Risk ranges from -0.05 to -0.010

- Gaussian method

Parametric VaR often do a better job, because it accounts for the tails of the distribution by more precisely estimating shape of the distribution tails of the risk 1 quantile.

Figure 7 VaR using gaussian method

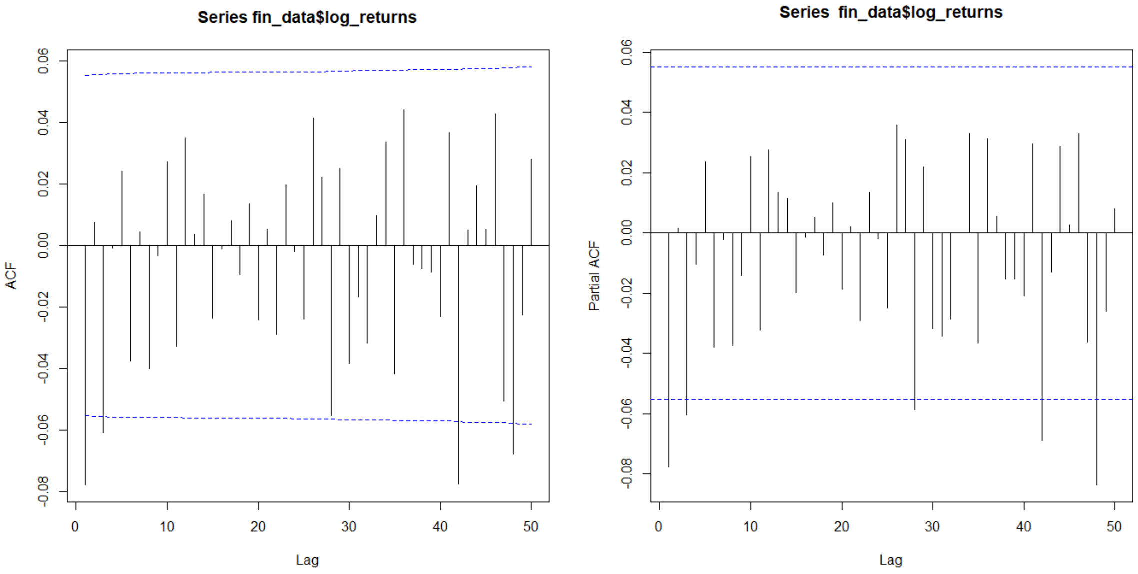


The plot alongside shows the value at risk at various confidence intervals ranging from 80 to 99 %. while using Gaussian parametric method. It can be said that VaR ranges from -0.03 to -0.07

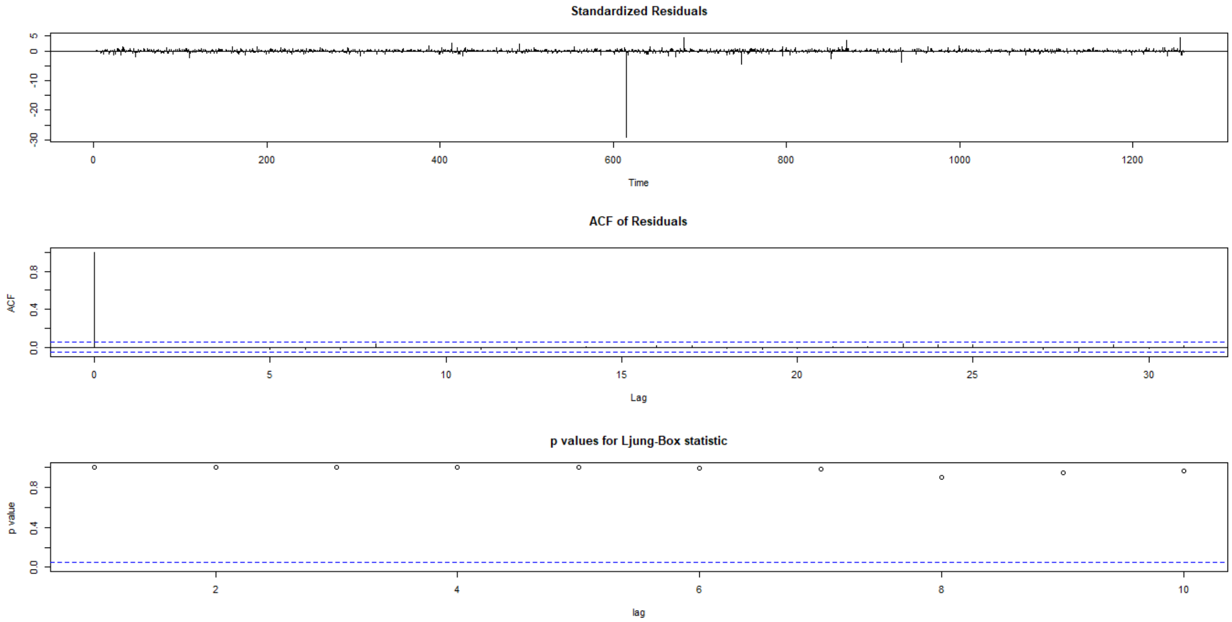
PART 3: Time series model

Time series model is fitted for log returns of EBAY stock. To begin making a model ACF (Auto correlation Function) and PACF (Partial Correlation Function) plots are plotted. ARIMA model requires the parameters of AR and MA model. The order of MA is identified using ACF and the order of AR-Autoregressive Model is identified using PACF. Below plots show the cut of lags and significant values at many points i.e values that are outside the significance level of (-0.05 and 0.05). It can be seen that points 1,3,28,42 and 48 are significant for MA and points 1,3,42,48 are significant for AR. The order of the AR is determined using the function 'ar' and 'order'. The best fit for this data is AR (3) model. The best ARIMA model was found to be of order 3,0,1. 0 here denotes the differentiated value used for stochasticity.

Figure 9 ACF and PACF of Log Returns



The ‘tsdiag’ function in R is used to study the fitted time series model and it produces the following plots. Even though higher orders of AR and MA such as (28,0,3) till the highest (48,0,48) gave better outcomes they weren’t significantly greater than that of (3,0,1), which was checked using ‘LRT’ and ‘loglik’. Hence the decision was made to use less parameters for the fitted model. These show the standardized residuals, ACF of residuals and p value for white noise test.

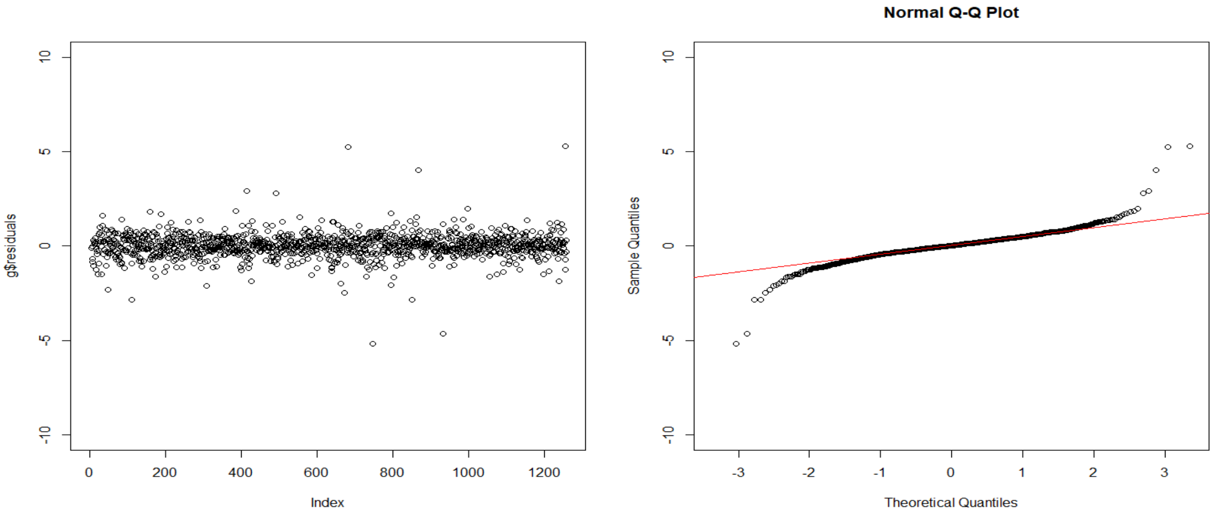


The residuals have a constant mean around 0 which is the case for normal distribution. The ACF of residuals show white noise with no significant values that are outside the interval of (-0.05 and 0.05) and the p values in the third plot are all greater than 0.05 suggesting that the residuals follow white noise. The above factors show that the ARIMA Model (3,0,1) is a good fit for this distribution of log returns.

PART 4: Model for volatility of log returns.

Volatility can be seen as a market measure for risk. Low volatility means returns have stayed close to the average and high volatility indicate a large fluctuation over time period. Volatility model is made with the help of ‘garch’ function in R. Volatility mainly studies the residuals of the returns. The order of the model depends on the data but the ideal garch order for a model is (1,1). This particular data required high order to make it normal (3,3) order was used , high orders such as (5,5) gave even better measures for volatility. The plots below represent the volatility of log return in order (5,5).

Figure 10 Volatility of Log Returns



PART 5 : Summary and Possible Implications.

The data of closing stock prices of EBAY was not normally distributed over the years from 2013 to 2018. The data was not also not stationary which was seen from the test for white noise. For the purpose of analysis the data was transformed into log returns. Log returns were also tested for normality and stationarity. ADF (Augmented Dickey-Fuller Test) was also performed which also showed that the data was not stationary and same with KPSS Test for Level Stationarity. Even though the log returns were not stationary the characteristics of the data such as skewness , variance, mean and kurtosis help in further analysis. The autocorrelation and partial autocorrelation function for the log returns showed many signification points outside the interval of -0.05 and 0.05 indicating that the data isn't white noise. Value at Risk for both closing stock prices and log returns were calculated for Confidence Interval ranging from 80 to 99%. The VaR ranged from -0.01 to -0.07.

A model is fitted for log returns to make the data a normal distribution. AR(1) model is considered a special case for white noise. The order of AR model is determined using PACF plot. The PACF plot showed many significant values outside the interval range (3,28,42,48). As per the rule of the graph the parameter selected for the AR model should be the one which is the last significant value and there are no more significant values after that, but for this particular distribution the cut off lags after 3 were 28, 42 and 48. After checking possible combinations it was seen that the difference between AR order 3 and others was not significantly higher and as it is not good practice to include too many parameters in the model the AR of order 3 was chosen. Order of MA is determined in a similar way as that of AR, the only difference is that the plot that is referred to for the purpose of finding significant points is ACF which has the 'ci.type' = 'MA'. As was the case with AR , different cut off lags were found for here as well and they were all tested against one another using the function 'LRT' and 'loglik'. The higher order parameter did not show significant increase in value as compared to lower ones and hence the order of MA was decided to be 1. Which made the ARIMA model of order (3,0,1). 0 here denotes the differentiated value used for stochasticity. When the summary of the fitted model was checked it was found that the p-value was greater than 0.05 and the residual plots showed the results that the distribution after fitting the ARIMA model was similar to the distribution of white noise.

The model for volatility of log returns was constructed using the function ‘garch’. Fitting an appropriate model using ‘garch’ for the residuals is done to make the residuals normal inn distribution. This is done by determining an order of the model. The ideal order is (1,1) but for this particular data this ideal order didn’t fit hence higher orders were tried. The significant results were obtained from order (3,3) and the best were obtained using order (5,5).

Possible Implications of Results:

It can be said that the ARIMA model and the ‘garch’ model are well suited for future predictions of stock closing prices for the stock of EBAY. They can be implemented for further analysis and evaluation with the help of log return distribution. The order of ARIMA model gives good performance on residuals as they are proven to be similar to that of white noise and the volatility model is able to fit the residuals into a normal distribution as well. From the values of risk and quantiles calculated from the closing prices a clear impression of the stock prices can be made, the VaR is at max a negative 0.1 in other words the value of loss at the time of risk could not be more than 0.1%.

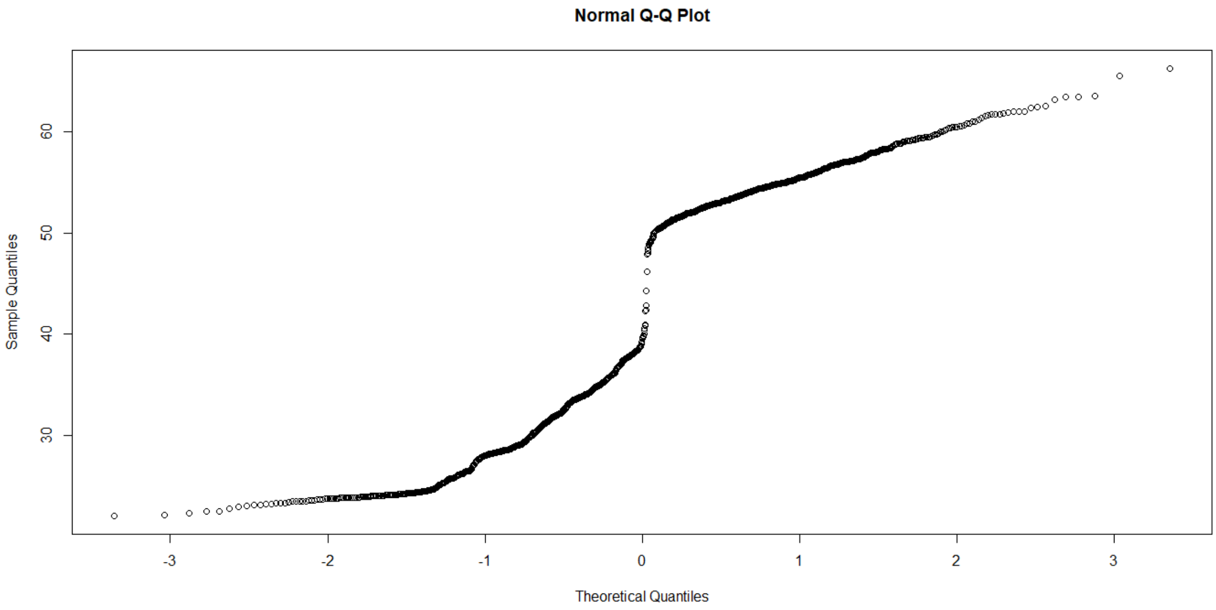
REFERENCES

Ruey S. Tsay. (2010) Analysis of Financial Time Series, 3rd Edition. Wiley

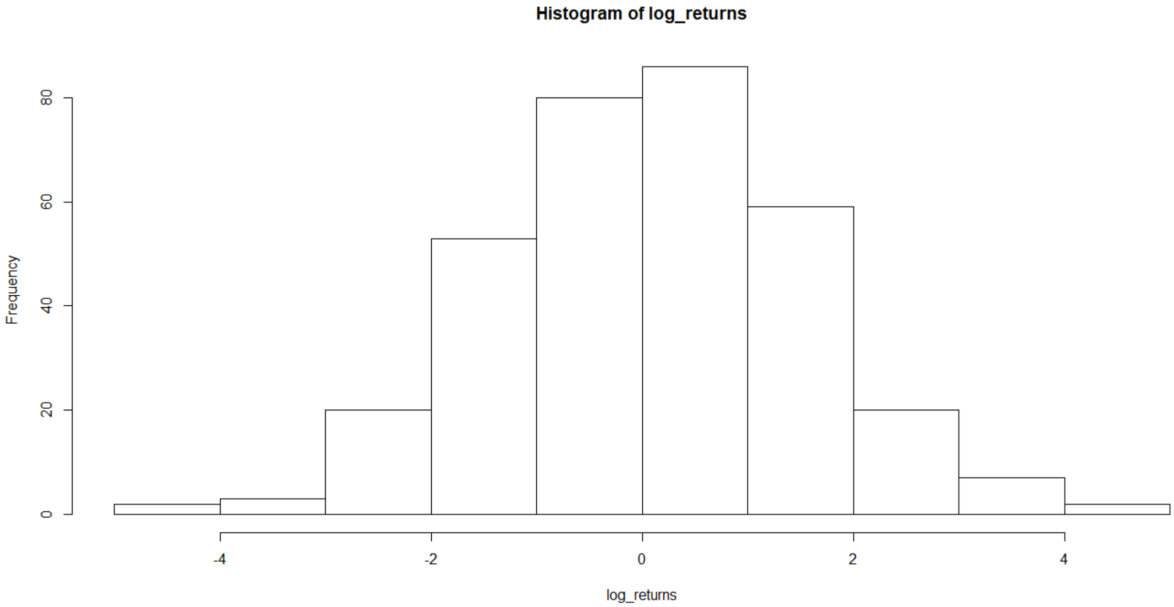
www.rdocumentation.org

APPENDIX:

QQPLOT of closing prices



Histogram of log returns when trend is removed.



```
finData = read.csv("all_stocks_5yr.csv")


fin_data = finData[finData$Name == "EBAY",]
sapply(fin_data,function(x)sum(is.na(x))) #no missing data or NA values.
fin_data$date = as.Date(fin_data$date, "%Y-%m-%d") #changing date format.
row.names(fin_data) = fin_data$date
qqnorm(fin_data$close)


#Log returns of closing price
plot(fin_data$close~fin_data$date, type = 'o', main='Time Series Plot'
    , ylab='daily closing prices', xlab='Time') #time series plot.
log_returns = diff(log(fin_data$close))      #calculating log return on clsoing prices
decompose(log_returns)
hist(log_returns)                            #histogram of returns.
quantile(log_returns,probs=c(0.01,0.025,0.05)) # vaR at 99%, 97.5% and 95%
fin_data$log_returns = c(c(0), log_returns)
plot(fin_data$log_returns~fin_data$date, type = "l")


#Testing Stationarity for closing price
library(stats)
library(TSA)
acf(fin_data$close)
Box.test(fin_data$close,type="Ljung-Box")
adf.test(fin_data$close)
kpss.test(fin_data$close)



#Testing Stationarity for LogReturns
acf(fin_data$log_returns)
```

```
Box.test(fin_data$log_returns,type="Ljung-Box")

adf.test(fin_data$log_returns)

kpss.test(fin_data$log_returns)


#Testing Normality for Data

m = mean(fin_data$close)

m

stdev = var(fin_data$close)

stdev

skewness(fin_data$close)

kurtosis(fin_data$close)

shapiro.test(fin_data$close)

x=rnorm(120,m, sqrt(stdev))

ks.test(fin_data$close,x)

#p-value is smaller than 0.05 the disturbution of data is significanlty different from normal distribution.


#Testing Normality for LogReturns

m = mean(fin_data$log_returns)

stdev = var(fin_data$log_returns)

skewness(fin_data$log_returns)

kurtosis(fin_data$log_returns)

shapiro.test(fin_data$log_returns)

x=rnorm(120,m, sqrt(stdev))

ks.test(fin_data$log_returns,x)

#pvalue is greater than 0.05 the distributions of returs is not significanly different from normal distribution.


#Value at Risk for LogReturns

#Negative signs show amount of loss.
```

```r
library(PerformanceAnalytics)


confidence_intervals = seq(0.80,0.99,0.01) #CI from 80% to 99%

historical_VaR=c()

Modified_VaR =c()

Gaussian_Var=c()

for (i in confidence_intervals) {

  val=VaR(fin_data[,"log_returns"], p=i, method="historical")

  historical_VaR=c(historical_VaR, val)

  val = VaR(fin_data[,"log_returns"], p=i, "modified")

  Modified_VaR = c(Modified_VaR , val)

  val=VaR(fin_data[,"log_returns"], p=i, method="gaussian")

  Gaussian_Var=c(Gaussian_Var, val)

}

plot(historical_VaR~confidence_intervals, type = "p", ylab = "Value at Risk for historical ", xlab = "Confidence interval")

plot(Modified_VaR~confidence_intervals, type = "p", ylab = "Value at Risk for modified", xlab = "Confidence interval")

plot(Gaussian_Var~confidence_intervals, type = "p", main = "Value at Risk for gaussian", xlab = "Confidence interval")


#Identifying and Fitting the best Time Series Model

library(tseries)

#1. Identify white noise

Box.test(fin_data$log_returns,  type="Ljung-Box")

#pvalue is less than 0.05 hence the data is not white noise.


#2. ACF and PACF

w=acf(fin_data$log_returns, lag.max = 50, ci.type = 'ma')

w
```

```
#PACF
x=pacf(fin_data$log_returns,50)
x
ord=ar(fin_data$log_returns)
ord$aic
ord$order
# possible AR VALUES= 3,42,48
#possible MA VALUES= 1,3,42,48
z=arima(fin_data$log_returns, order = c(3,0,1))
tsdiag(z)

#Testing Log Returns Volatility
par(mfrow=c(1,2))
g = garch(fin_data$log_returns, order=c(5,5))
summary(g)
plot(g$residuals, ylim=c(-10,10))
qqnorm(g$residuals,ylim=c(-10,10)); qqline(g$residuals, col=2, )
```