```r
suppressMessages(library(tidyverse))
suppressMessages(library(lubridate))
suppressMessages(library(tidytext))
suppressMessages(library(textdata))
suppressMessages(library(dplyr))
suppressMessages(library(quantmod))
suppressMessages(library(fGarch))
```

```r
df <- read.csv("data/stock_tweets.csv")  %>%
  filter(Stock.Name == 'TSLA') %>%
  mutate(Tweet.ID = row_number()) %>%
  dplyr::select(Tweet.ID, Date, Tweet)
dim(df)
```

```
## [1] 37422     3
```

```r
names(df)
```

```
## [1] "Tweet.ID" "Date"     "Tweet"
```

```r
df$Date <-  ymd(substr(df$Date, 1, 10))
tweets <- data.frame(df)
```

```r
# Sentiment analysis
map_bing_sentiment <- function(sentiment) {
  ifelse(sentiment %in% c("positive"), 1, ifelse(sentiment %in% c("negative"), -1, 0))
}

map_nrc_sentiment <- function(sentiment) {
  nrc_positive_sentiments <- c("positive", "anticipation", "surprise", "trust", "joy")
  nrc_negative_sentiments <- c("negative", "anger", "disgust", "fear", "sadness")
  ifelse(sentiment %in% nrc_positive_sentiments, 1,
         ifelse(sentiment %in% nrc_negative_sentiments, -1, 0))
}

tweet_tokens <- tweets %>%
  unnest_tokens(word, Tweet)

sentiments <- get_sentiments("bing") %>% mutate(sentiment_score = map_bing_sentiment(sentiment))
#sentiments <- get_sentiments("afinn") %>% mutate(sentiment_score = value)
#sentiments <- get_sentiments("nrc") %>% mutate(sentiment_score = map_nrc_sentiment(sentiment))

tweets_sentiment <- tweet_tokens %>%
  inner_join(sentiments, by = "word", relationship = "many-to-many")  %>%
  distinct(Tweet.ID, Date, word, .keep_all = TRUE)

tweets_sentiment_summary <- tweets_sentiment %>%
  group_by(Tweet.ID, Date) %>%
  summarise(sentiment_score = sum(sentiment_score, na.rm = TRUE), .groups = "drop")

daily_sentiment <- tweets_sentiment_summary  %>%
  group_by(Date) %>%
  summarise(daily_sentiment = mean(sentiment_score))
```

Most positive Tweet

```r
most_pos_twid <-
  tweets_sentiment_summary[which.max(tweets_sentiment_summary$sentiment_score),"Tweet.ID"]
tweets[most_pos_twid$Tweet.ID,]$Tweet
```

```
## [1] "Love my S Plaid more every day since purchased in June. It’s the smartest, most fun &amp; full
```

```r
max(tweets_sentiment_summary$sentiment_score)
```

```
## [1] 9
```

Most negative Tweet

```r
most_neg_twid <-
  tweets_sentiment_summary[which.min(tweets_sentiment_summary$sentiment_score),"Tweet.ID"]
tweets[most_neg_twid$Tweet.ID,]$Tweet
```

```
## [1] "Whenever there is big trouble and bad news at @Tesla, @elonmusk is doing a publicity stunt to d
```

```r
min(tweets_sentiment_summary$sentiment_score)
```

```
## [1] -9
```

Most positive day

```r
daily_sentiment[which.max(daily_sentiment$daily_sentiment),]
```

```
## # A tibble: 1 x 2
##   Date       daily_sentiment
##   <date>               <dbl>
## 1 2021-12-25            1.65
```

Most negative day

```r
daily_sentiment[which.min(daily_sentiment$daily_sentiment),]
```

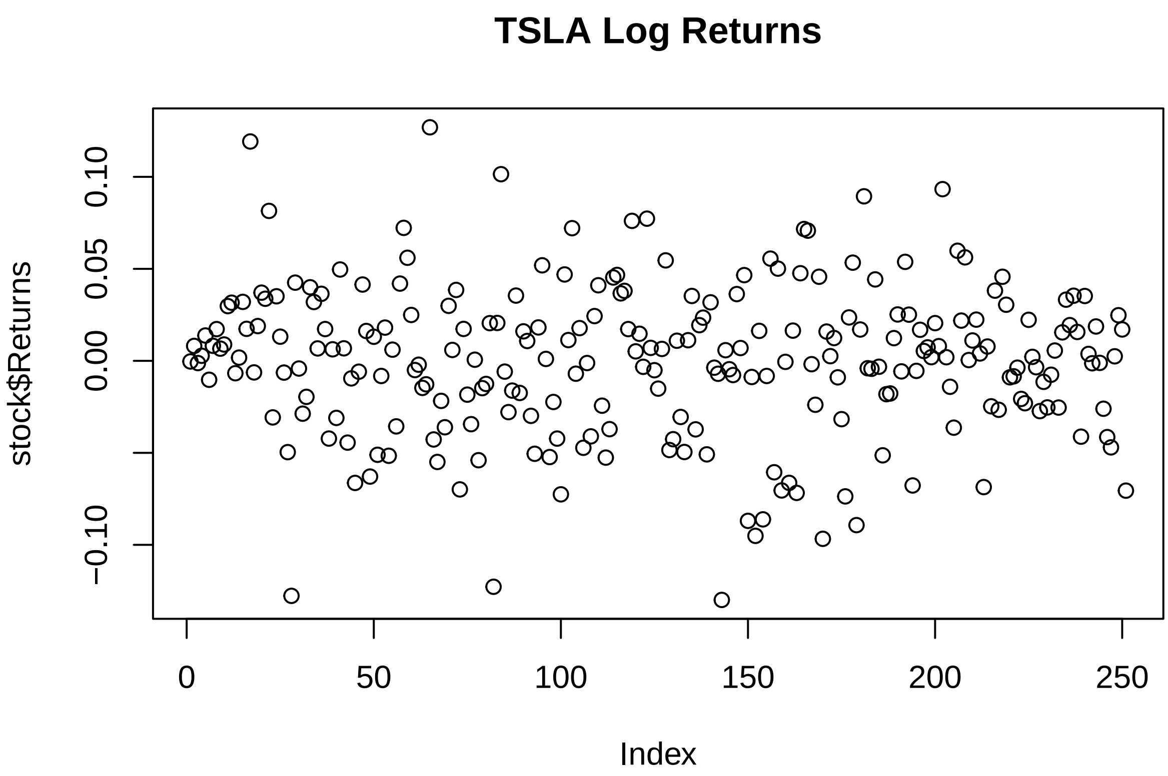
```
## # A tibble: 1 x 2
##   Date       daily_sentiment
##   <date>               <dbl>
## 1 2022-07-07          -0.575
```

```r
df <- read.csv("data/stock_yfinance_data.csv") %>%
  filter(Stock.Name == 'TSLA') %>%
  dplyr::select(Date, Adj.Close)
df$Date <- as.Date(df$Date)
head(df)
```

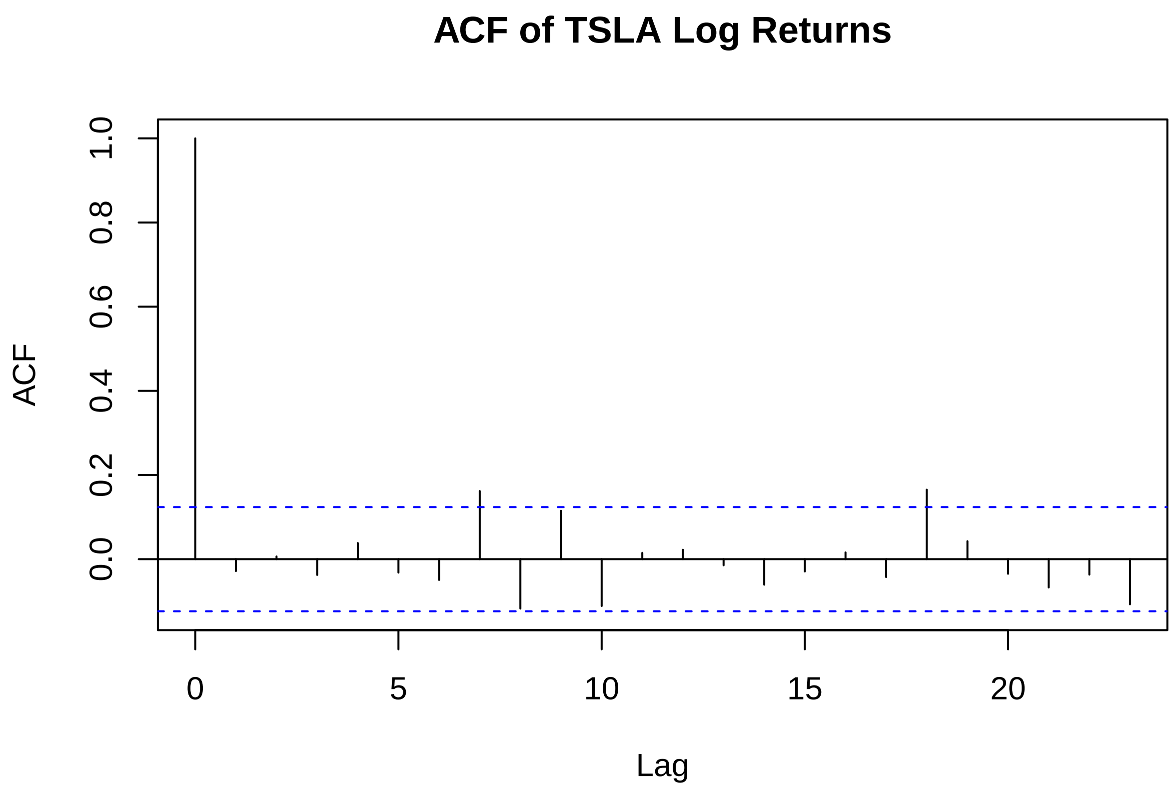
```
##          Date Adj.Close
## 1 2021-09-30  258.4933
## 2 2021-10-01  258.4067
## 3 2021-10-04  260.5100
## 4 2021-10-05  260.1967
## 5 2021-10-06  260.9167
## 6 2021-10-07  264.5367
```

```r
df$Returns <- c(diff(log(df$Adj.Close)), NA)
stock <- data.frame(df)
stock <- stock %>% na.omit()
```

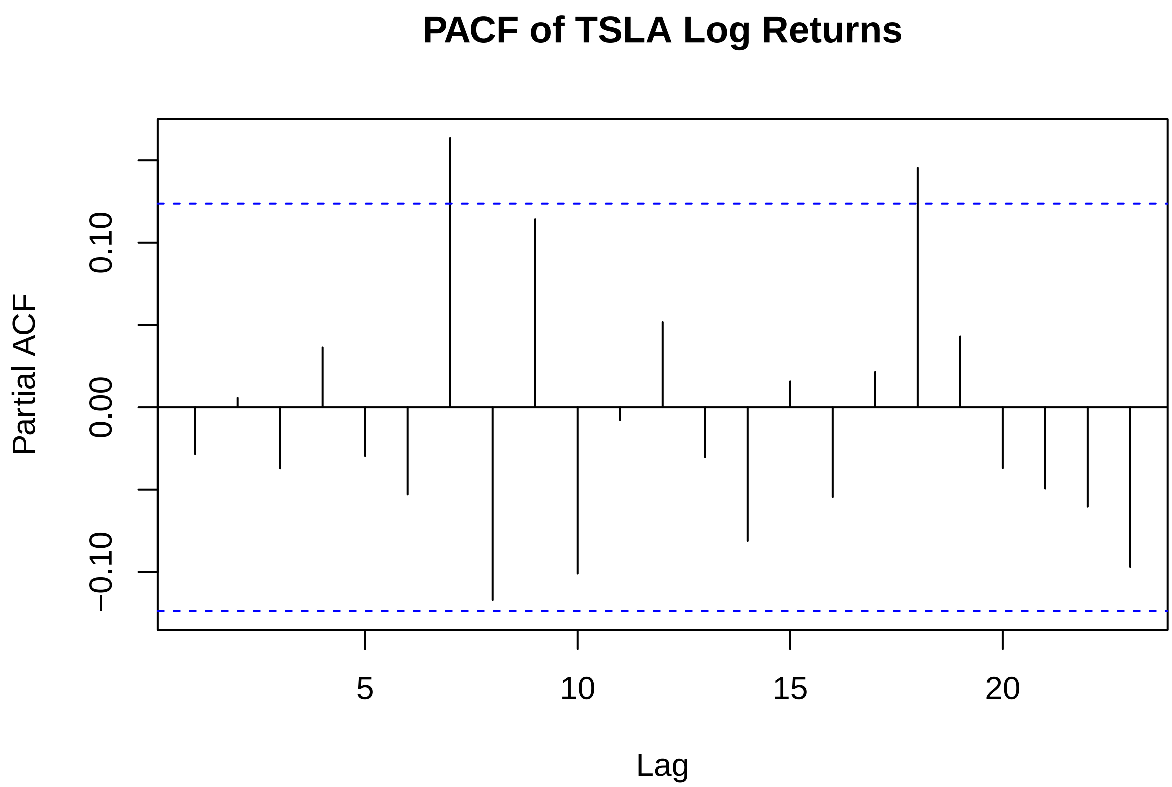
```r
plot(stock$Returns, main="TSLA Log Returns")
```



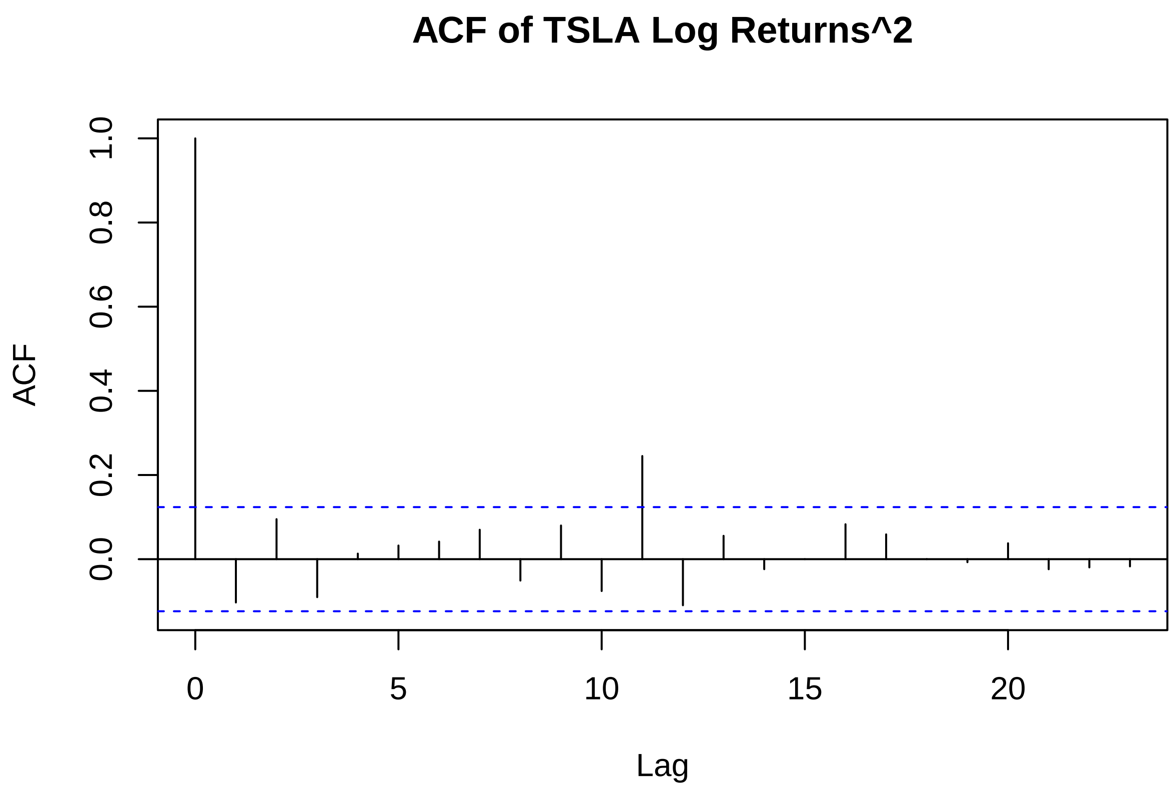

```r
par(mfrow=c(1,1))
acf(stock$Returns, main="ACF of TSLA Log Returns", na.action = na.pass)
```



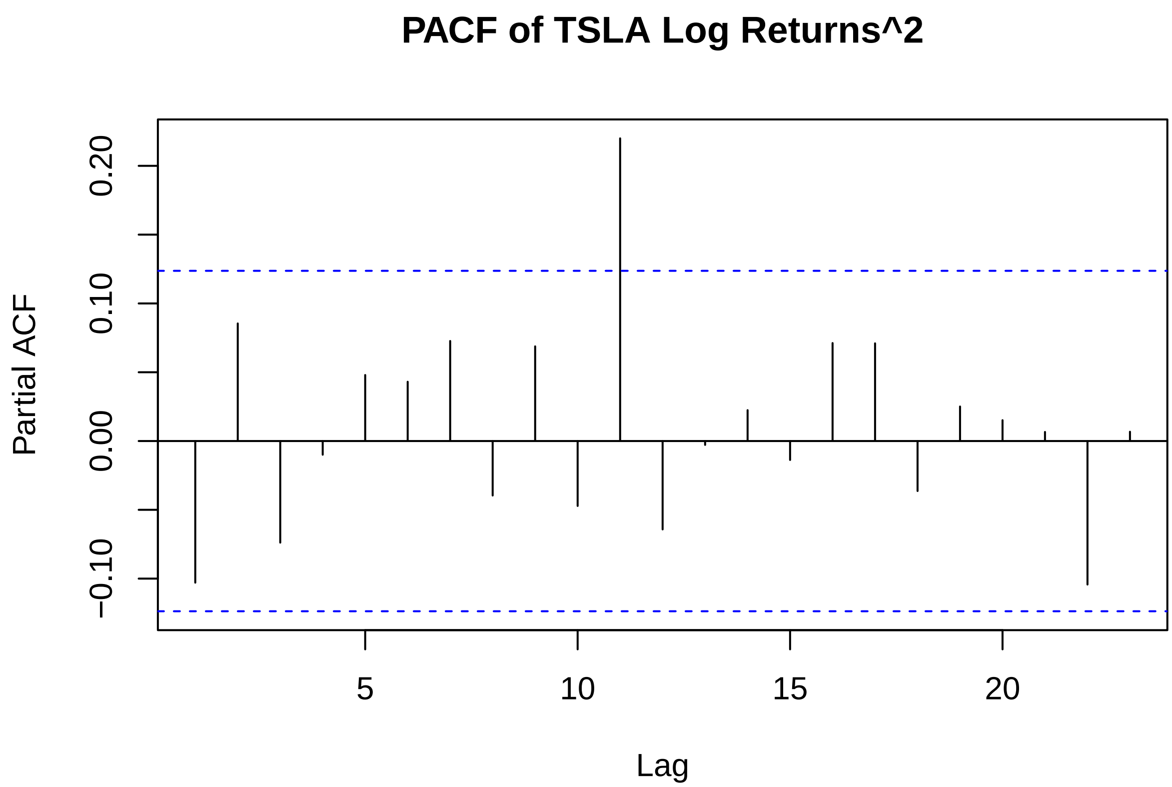

```
pacf(stock$Returns, main="PACF of TSLA Log Returns", na.action = na.pass)
```




```
acf(stock$Returns^2, main="ACF of TSLA Log Returns^2", na.action = na.pass)
```




```r
pacf(stock$Returns^2, main="PACF of TSLA Log Returns^2", na.action = na.pass)
```




```
suppressWarnings(library(forecast))

arma_rt_squared <- auto.arima(stock$Returns^2, max.p = 5, max.q = 5, max.order = 10,
                              stationary = T, seasonal = F, trace = F,
                              stepwise = F, approximation = F)
summary(arma_rt_squared)
```

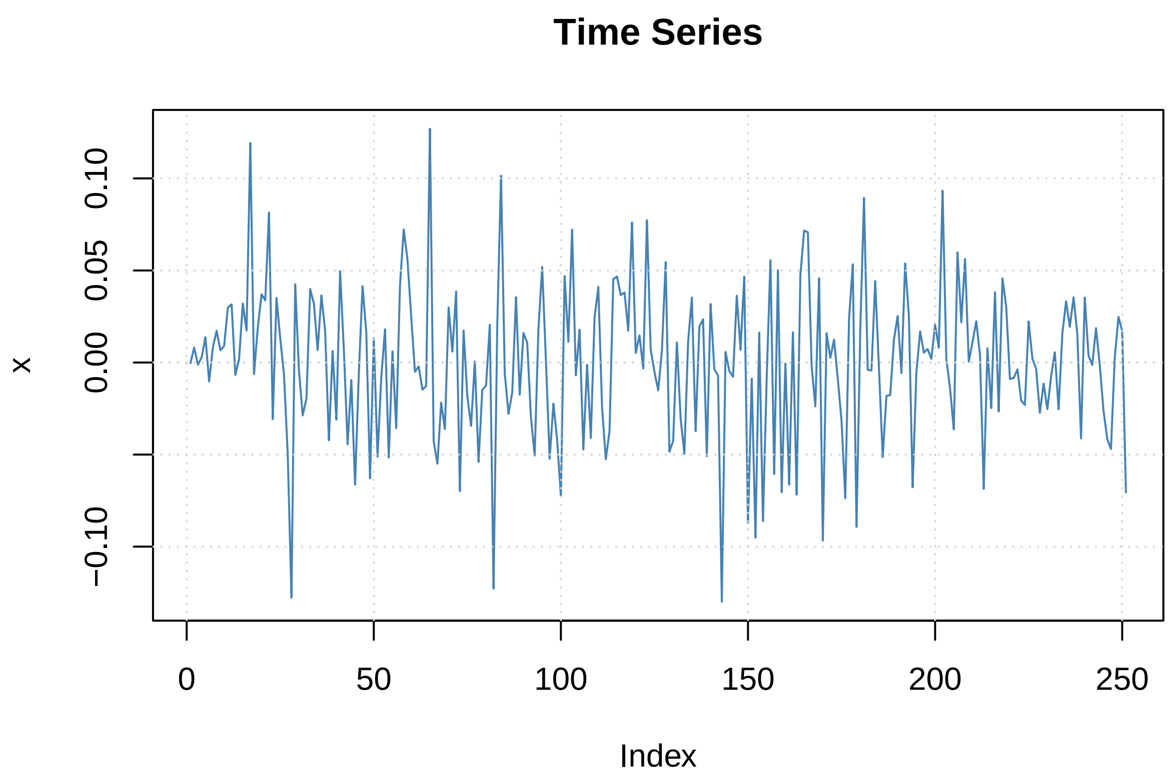
```
## Series: stock$Returns^2
## ARIMA(1,0,1) with non-zero mean
##
## Coefficients:
##           ar1     ma1    mean
##       -0.6196  0.5105  0.0017
## s.e.   0.1958  0.2102  0.0002
##
## sigma^2 = 7.706e-06:  log likelihood = 1122.91
## AIC=-2237.83   AICc=-2237.66   BIC=-2223.72
##
## Training set error measures:
##                         ME        RMSE         MAE       MPE     MAPE      MASE
## Training set -4.434692e-07 0.002759389 0.001733483 -17226.44 17257.83 0.7116065
##                    ACF1
## Training set 0.01881578
```

```r
stock_ts <- ts(stock$Returns,
               start=c(2021,9),
               frequency=365)
garch_model <- garchFit(~ garch(1,1), data=stock_ts, trace=FALSE)
summary(garch_model)
```

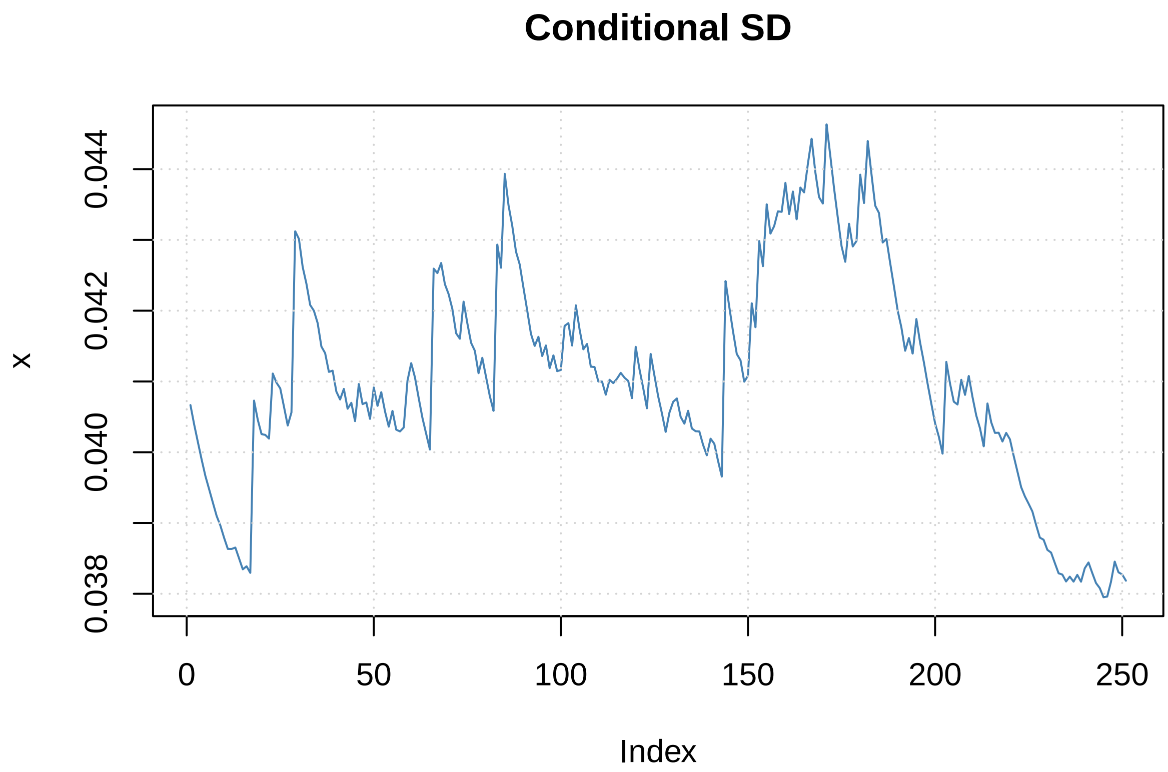
```
##
## Title:
##  GARCH Modelling
##
## Call:
##  garchFit(formula = ~garch(1, 1), data = stock_ts, trace = FALSE)
##
## Mean and Variance Equation:
##  data ~ garch(1, 1)
## <environment: 0x152d91598>
##  [data = stock_ts]
##
## Conditional Distribution:
##  norm
##
## Coefficient(s):
##         mu       omega      alpha1       beta1
## 0.00034657  0.00007856  0.01438639  0.93899717
##
## Std. Errors:
##  based on Hessian
##
## Error Analysis:
##         Estimate  Std. Error  t value Pr(>|t|)
## mu     3.466e-04   2.557e-03    0.136    0.892
## omega  7.856e-05   8.856e-05    0.887    0.375
## alpha1 1.439e-02   1.997e-02    0.720    0.471
## beta1  9.390e-01   5.494e-02   17.090   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log Likelihood:
##  448.0723    normalized:  1.785149
##
## Description:
##  Tue Apr 16 10:43:49 2024 by user:
##
##
## Standardised Residuals Tests:
##                                 Statistic     p-Value
##  Jarque-Bera Test   R    Chi^2  11.8733436 0.002640804
##  Shapiro-Wilk Test  R    W       0.9837818 0.005887743
##  Ljung-Box Test     R    Q(10)  17.7126029 0.060009507
##  Ljung-Box Test     R    Q(15)  19.0600447 0.211025838
##  Ljung-Box Test     R    Q(20)  27.7444493 0.115588155
##  Ljung-Box Test     R^2  Q(10)  11.4358694 0.324582474
##  Ljung-Box Test     R^2  Q(15)  30.6866295 0.009672734
```

```
##  Ljung-Box Test     R^2  Q(20)  33.2456537 0.031704058
##  LM Arch Test       R    TR^2   21.5516296 0.042863209
##
## Information Criterion Statistics:
##       AIC       BIC       SIC      HQIC
## -3.538425 -3.482243 -3.538922 -3.515816
```

```r
par(mfrow=c(1,1))
plot(garch_model, which = 1)
```




```r
plot(garch_model, which = 2)
```




```r
par(mfrow=c(1,1))
```

```r
combined_data <- left_join(stock, daily_sentiment, by = "Date")
cor(combined_data$daily_sentiment, combined_data$Returns, use = "complete.obs")
```

Combine data from stock, daily_sentiment

```
## [1] -0.01235564
```

```r
model <- lm(Returns ~ daily_sentiment, data = combined_data)
summary(model)
```

```
##
## Call:
## lm(formula = Returns ~ daily_sentiment, data = combined_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.129903 -0.023383  0.001996  0.023012  0.126955
##
## Coefficients:
```

```
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.0006787  0.0037512   0.181    0.857
## daily_sentiment -0.0019186  0.0098397  -0.195    0.846
##
## Residual standard error: 0.04081 on 249 degrees of freedom
## Multiple R-squared:  0.0001527,  Adjusted R-squared:  -0.003863
## F-statistic: 0.03802 on 1 and 249 DF,  p-value: 0.8456
```

```r
suppressMessages(library(vars))
df <- combined_data[, c("Returns", "daily_sentiment")]
# VARselect
lag.select <- VARselect(df,
                        lag.max = 30,
                        type = "both")
optimal.lags <- lag.select$selection['AIC(n)']

# Fit the VAR model
var.model <- VAR(df, p = optimal.lags)

summary(var.model)
```

```
##
## VAR Estimation Results:
## =========================
## Endogenous variables: Returns, daily_sentiment
## Deterministic variables: const
## Sample size: 249
## Log Likelihood: 441.396
## Roots of the characteristic polynomial:
## 0.5583 0.3913 0.2375 0.2375
## Call:
## VAR(y = df, p = optimal.lags)
##
##
## Estimation results for equation Returns:
## ========================================
## Returns = Returns.l1 + daily_sentiment.l1 + Returns.l2 + daily_sentiment.l2 + const
##
##                      Estimate Std. Error t value Pr(>|t|)
## Returns.l1          -0.021839   0.063820  -0.342   0.7325
## daily_sentiment.l1 -0.014608   0.010277  -1.421   0.1565
## Returns.l2           0.025587   0.065023   0.394   0.6943
## daily_sentiment.l2  0.020740   0.010126   2.048   0.0416 *
## const               -0.001562   0.004358  -0.358   0.7204
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 0.04078 on 244 degrees of freedom
## Multiple R-Squared: 0.02158, Adjusted R-squared: 0.005541
## F-statistic: 1.345 on 4 and 244 DF,  p-value: 0.2537
##
##
## Estimation results for equation daily_sentiment:
```

```
## ================================================
## daily_sentiment = Returns.l1 + daily_sentiment.l1 + Returns.l2 + daily_sentiment.l2 + const
##
##                     Estimate Std. Error t value Pr(>|t|)
## Returns.l1           1.31367    0.38965   3.371 0.000869 ***
## daily_sentiment.l1  0.16335    0.06275   2.603 0.009799 **
## Returns.l2           0.78679    0.39700   1.982 0.048618 *
## daily_sentiment.l2  0.15634    0.06183   2.529 0.012081 *
## const                0.18730    0.02661   7.040 1.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Residual standard error: 0.249 on 244 degrees of freedom
## Multiple R-Squared: 0.1198,  Adjusted R-squared: 0.1054
## F-statistic: 8.302 on 4 and 244 DF,  p-value: 2.708e-06
##
##
##
## Covariance matrix of residuals:
##                     Returns daily_sentiment
## Returns           0.0016627      -0.0001715
## daily_sentiment -0.0001715       0.0619805
##
## Correlation matrix of residuals:
##                   Returns daily_sentiment
## Returns           1.00000        -0.01689
## daily_sentiment -0.01689         1.00000
```

```r
suppressMessages(library(dplyr))

combined_data <- combined_data %>%
  arrange(Date) %>%
  mutate(
    Returns_l1 = lag(Returns, 1),
    Returns_l2 = lag(Returns, 2),
    daily_sentiment_l1 = lag(daily_sentiment, 1),
    daily_sentiment_l2 = lag(daily_sentiment, 2)
  )

model <- lm(Returns ~ Returns_l1 + Returns_l2 + daily_sentiment + daily_sentiment_l1 + daily_sentiment_l

summary(model)
```

```
##
## Call:
## lm(formula = Returns ~ Returns_l1 + Returns_l2 + daily_sentiment +
##     daily_sentiment_l1 + daily_sentiment_l2, data = combined_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.125592 -0.025207  0.002669  0.024002  0.128883
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.001044   0.004789  -0.218   0.8277
## Returns_l1          -0.018204   0.065415  -0.278   0.7810
## Returns_l2           0.027764   0.065669   0.423   0.6728
## daily_sentiment     -0.002767   0.010505  -0.263   0.7925
## daily_sentiment_l1 -0.014156   0.010439  -1.356   0.1763
## daily_sentiment_l2  0.021173   0.010278   2.060   0.0405 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04085 on 243 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.02186,    Adjusted R-squared:  0.001734
## F-statistic: 1.086 on 5 and 243 DF,  p-value: 0.3686
```

```
model <- lm(daily_sentiment ~ Returns + Returns_l1 + Returns_l2 + daily_sentiment_l1 + daily_sentiment_l

summary(model)
```

```
##
## Call:
## lm(formula = daily_sentiment ~ Returns + Returns_l1 + Returns_l2 +
##     daily_sentiment_l1 + daily_sentiment_l2, data = combined_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87008 -0.12520  0.00037  0.14850  0.60489
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.18714    0.02666   7.019 2.22e-11 ***
## Returns             -0.10313    0.39161  -0.263  0.79250
## Returns_l1           1.31141    0.39049   3.358  0.00091 ***
## Returns_l2           0.78943    0.39788   1.984  0.04837 *
## daily_sentiment_l1  0.16185    0.06313   2.564  0.01096 *
## daily_sentiment_l2  0.15848    0.06247   2.537  0.01182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2494 on 243 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:   0.12,  Adjusted R-squared:  0.1019
## F-statistic:  6.63 on 5 and 243 DF,  p-value: 8.299e-06
```

Conclusion: Returns can't be predicted based on current or l1 lagged values of daily_sentiment, or l1/l2 lagged values of Returns.

Conclusion: daily_sentiment can't be predicted based on current Return but can be explained by l1/l2 lagged values of Returns as well as l1/l2 lagged values of daily_sentiment.