IE5374 - Sec 03 - Group 5: Bitcoin Tweets and Price Analysis

Abhishek Taware

Libraries

```
library(devtools)
library(tidyr)
library(magrittr)
library(lubridate)
library(nonlinearTseries)
library(stringr)
library(lemon)
library(knitr)
library(dplyr)
library(ggplot2)
library(forecast)
library(tidytext)
library(factoextra)
library(lubridate)
library(wordcloud)
```

Data Preprocessing

Loading the datasets

```
Bitcoin_tweets <- read.csv("C:\\Bitcoin_tweets.csv")
Bitcoin_price <- read.csv("C:\\BTC-USD.csv")
```

Viewing the data for Tweets

```
head(Bitcoin_tweets,5)
```

```
##                                             user_name   user_location
## 1                                       DeSota Wilson     Atlanta, GA
## 2                                           CryptoND
## 3                                          Tdlmatias London, England
## 4                                  Crypto is the future
## 5 Alex Kirchmaier ðŸ‡¦ðŸ‡¹ðŸ‡¸ðŸ‡ª #FactsSuperspreader          Europa
##
```

```
## 1 Biz Consultant, real estate, fintech, startups, posts are not the view of my employer, RTs are not
## 2                        ðŸ\230Ž BITCOINLIVE is a Dutch platform aimed at informing the general publi
## 3               IM Academy : The best #forex, #SelfEducation, #Cryptocurrency trading education plat
## 4                 I will post a lot of buying signals for BTC trading, so I hope that you block me. I
## 5                           Co-founder @RENJERJerky | Forbes 30Under30 | Innovation Economist, Lateral
##          user_created user_followers user_friends user_favourites user_verified
## 1 2009-04-26 20:05:09           8534         7605            4838         False
## 2 2019-10-17 20:12:10           6769         1532           25483         False
## 3 2014-11-10 10:50:37            128          332             924         False
## 4 2019-09-28 16:48:12            625          129              14         False
## 5 2016-02-03 13:15:55           1249         1472           10482         False
##                   date
## 1 2021-02-10 23:59:04
## 2 2021-02-10 23:58:48
## 3 2021-02-10 23:54:48
## 4 2021-02-10 23:54:33
## 5 2021-02-10 23:54:06
##
## 1            Blue Ridge Bank shares halted by NYSE after #bitcoin ATM announcement https://t.co/xaaZn
## 2 ðŸ\230Ž Today, that's this #Thursday, we will do a "ðŸŽ¬ Take 2" with our friend @LeoWandersleb, #I
## 3          Guys evening, I have read this article about BTC and would like to share with you all - I
## 4                                                    $BTC A big chance in a billion! Price: \\4872(
## 5       This network is secured by 9 508 nodes as of today. Soon, the biggest bears will recognise: #
##                                   hashtags              source is_retweet
## 1                                ['bitcoin']     Twitter Web App      False
## 2 ['Thursday', 'Btc', 'wallet', 'security'] Twitter for Android      False
## 3                                            Twitter Web App      False
## 4        ['Bitcoin', 'FX', 'BTC', 'crypto']             dlvr.it      False
## 5                                    ['BTC']     Twitter Web App      False
```

Viewing the data for Price

```
head(Bitcoin_price,5)
```

```
##         Date     Open     High      Low    Close Adj.Close      Volume
## 1 2021-01-01 28994.01 29600.63 28803.59 29374.15  29374.15 40730301359
## 2 2021-01-02 29376.46 33155.12 29091.18 32127.27  32127.27 67865420765
## 3 2021-01-03 32129.41 34608.56 32052.32 32782.02  32782.02 78665235202
## 4 2021-01-04 32810.95 33440.22 28722.76 31971.91  31971.91 81163475344
## 5 2021-01-05 31977.04 34437.59 30221.19 33992.43  33992.43 67547324782
```

Checking the dimensions for tweets

```
dim(Bitcoin_tweets)
```

```
## [1] 1944387      13
```

Checking the dimensions for price

```
dim(Bitcoin_price)
```

```
## [1] 347   7
```

Tweets data columns

```
colnames(Bitcoin_tweets)
```

```
##  [1] "user_name"        "user_location"    "user_description" "user_created"
##  [5] "user_followers"   "user_friends"     "user_favourites"  "user_verified"
##  [9] "date"             "text"             "hashtags"         "source"
## [13] "is_retweet"
```

Price data columns

```
colnames(Bitcoin_price)
```

```
## [1] "Date"      "Open"      "High"      "Low"       "Close"     "Adj.Close"
## [7] "Volume"
```

Checking all the null values in both the datasets

```
sum(Bitcoin_tweets[is.null(Bitcoin_tweets)])
```

```
## [1] 0
```

```
sum(Bitcoin_price[is.null(Bitcoin_price)])
```

```
## [1] 0
```

Summarizing the distinct data for the Tweets

```
Bitcoin_tweets %>% summarise_all(n_distinct)
```

```
##   user_name user_location user_description user_created user_followers
## 1    327177         56382           312495       323908          50836
##   user_friends user_favourites user_verified    date    text hashtags source
## 1        19572           90376            18 1530931 1902666   451621   1576
##   is_retweet
## 1          2
```

Subsetting Bitcoin Price data to contain data between 5th February 2021 and 26th November 2021

```
Bitcoin_price <- Bitcoin_price[Bitcoin_price$Date >= "2021-02-05" &    # Extract data frame subset
                  Bitcoin_price$Date <= "2021-11-26", ]
```

Cleaning the Data

```
# Apart from cleaning the data over here, we have also cleaned the data as and
# when needed by subsetting it so that there are no redundancies / dependencies.

Bitcoin_tweets <- Bitcoin_tweets[- grep("[btc]", Bitcoin_tweets$date),]
Bitcoin_tweets <- Bitcoin_tweets[!(is.na(Bitcoin_tweets$user_followers)),]
```

Creating separate columns for Date, Month and Time

```
Bitcoin_tweets$Date <- as.Date(Bitcoin_tweets$date)
Bitcoin_tweets$Date <- as.Date(Bitcoin_tweets$Date, format = "%Y-%m-%d")
Bitcoin_tweets$Month <- format(Bitcoin_tweets$Date, "%m")
Bitcoin_tweets$Time <- format(as.POSIXct(Bitcoin_tweets$date), format = "%H:%M:%S")
```

Writing the new data

```
write.csv(Bitcoin_tweets, "C:\\Users\\Abhishek's PC\\Downloads\\Bitcoins.csv")
write.csv(Bitcoin_price,"C:\\Users\\Abhishek's PC\\Downloads\\BTC-USD.csv")
```

Reading the cleaned data

```
Bitcoin_price <- read.csv("C:\\Users\\Abhishek's PC\\Downloads\\BTC-USD.csv")
Bitcoin_tweets <- read.csv("C:\\Users\\Abhishek's PC\\Downloads\\Bitcoins.csv")
```

Section 1: Probability Questions

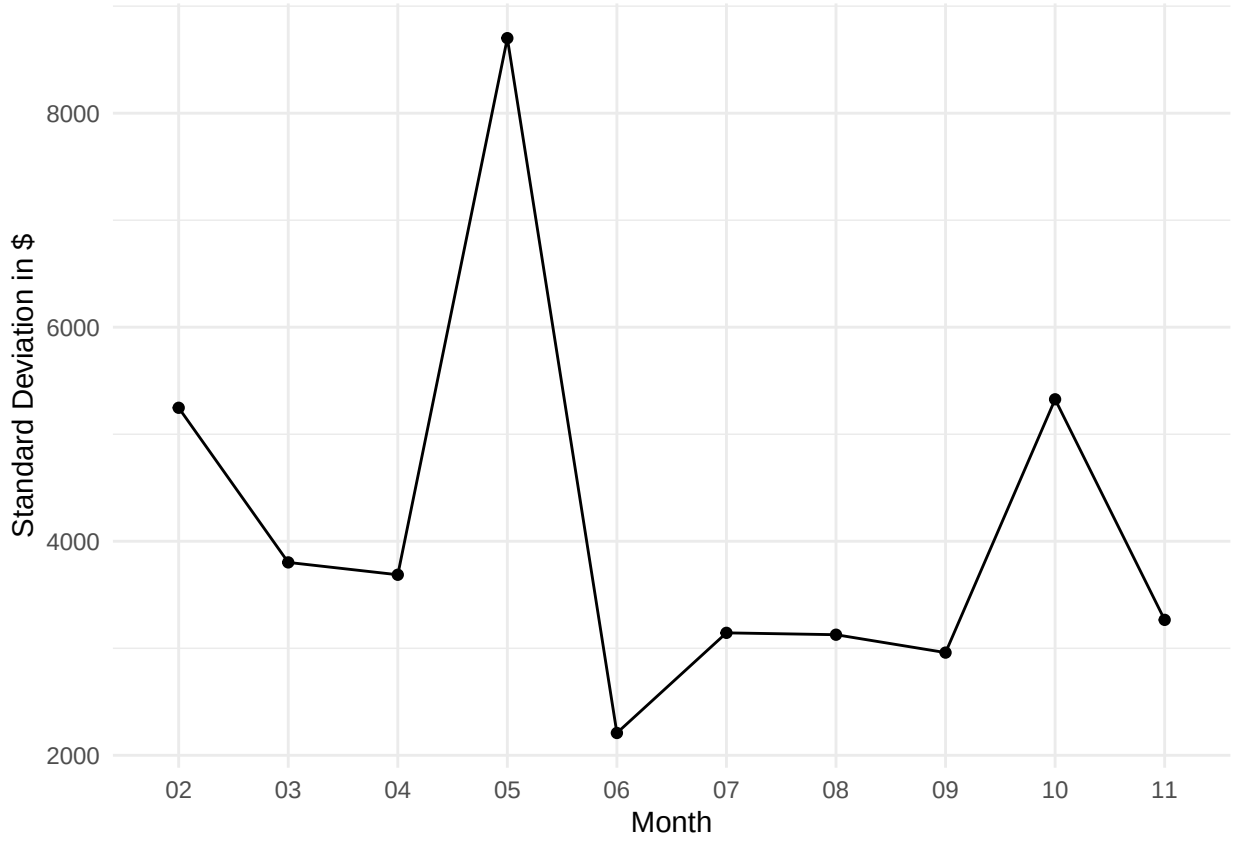
Q. What was the Month on Month Volatility of Bitcoin's Price Movement?

```
Bitcoin_price$Date <- as.Date(Bitcoin_price$Date, format = "%Y-%m-%d")
Bitcoin_price$Month <- format(Bitcoin_price$Date,"%m")

bit <- Bitcoin_price %>%
  select(Month, High, Volume) %>%
  drop_na() %>%
```

```
  group_by(Month) %>%
  summarise(avg = sum(High)/30, max_price = max(High),
            min_price = min(High), standard_deviation = sd(High),
            avg_volume = sum(Volume)/30)

ggplot(data = bit, aes(x=Month,y=standard_deviation, group = 1))+
  geom_line()+ geom_point() +
  theme_minimal()+
   xlab("Month") + ylab("Standard Deviation in $")
```



Conclusion

In trading, standard deviation is used as a measure of volatility of an instrument i.e how much it will move either up or down on an average. As we can see through the line graph above, bitcoin was highly volatile during the months of February, May and October which means that during these months, bitcoin had very high price movements due to which taking a trade during these Months would have been a risky affair.

Q. Is there any correlation between the number of followers a user has compared to the number of friends?

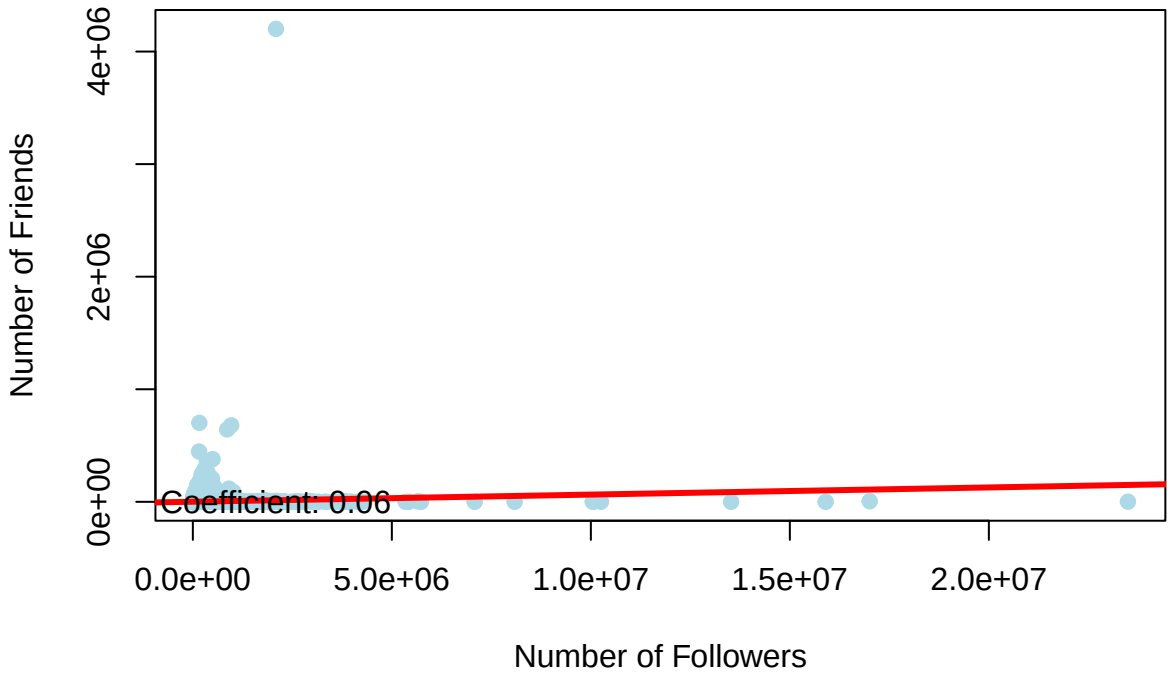
```
Corr_Data <- Bitcoin_tweets[c('user_name','user_followers', 'user_friends')]
```

```r
Corr_Data <- Corr_Data[!duplicated(Corr_Data[,c('user_name')]),]

x <- Corr_Data$user_followers
y <- Corr_Data$user_friends
plot(x, y, pch = 19, col = "lightblue", xlab="Number of Followers", ylab="Number of Friends")

# Regression line
abline(lm(y ~ x), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation Coefficient:", round(cor(x, y), 2)), x = 2500, y = 95)
```



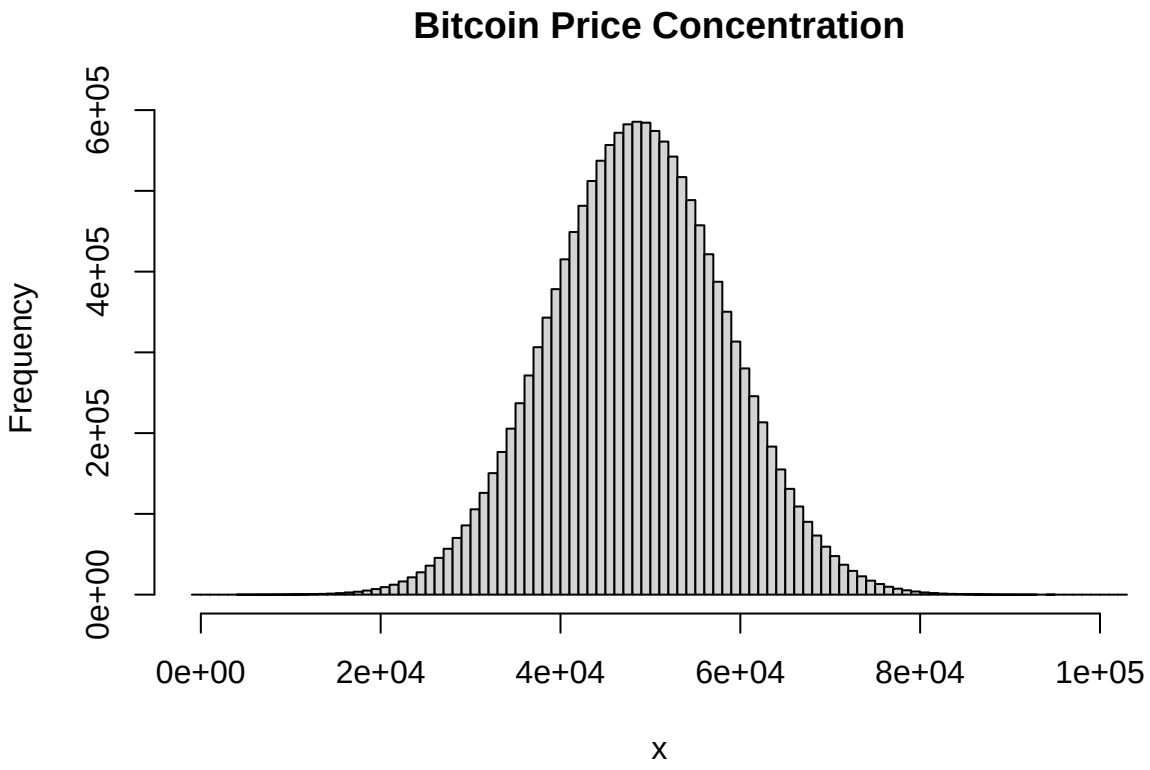
Conclusion

As we can see, the Correlation Coefficient for the entire dataset is 0.06 which means that there is absolutely no relation between the number of Followers a user has compared to number of friends a user has which can be a good datapoint to see if the accounts that are being used are just bot accounts or not. Since the correlation is low, we can say that most of the user accounts are real accounts.

Q. What has been the range in which bitcoin has traded the highest over the available data?

```
s <- sum(Bitcoin_price$Close)
m <- mean(Bitcoin_price$Close)
sd <- sd(Bitcoin_price$Close)

x <-  rnorm(s,mean=m,sd=sd)
hist(x, breaks = 100, main="Bitcoin Price Concentration")
```



Conclusion

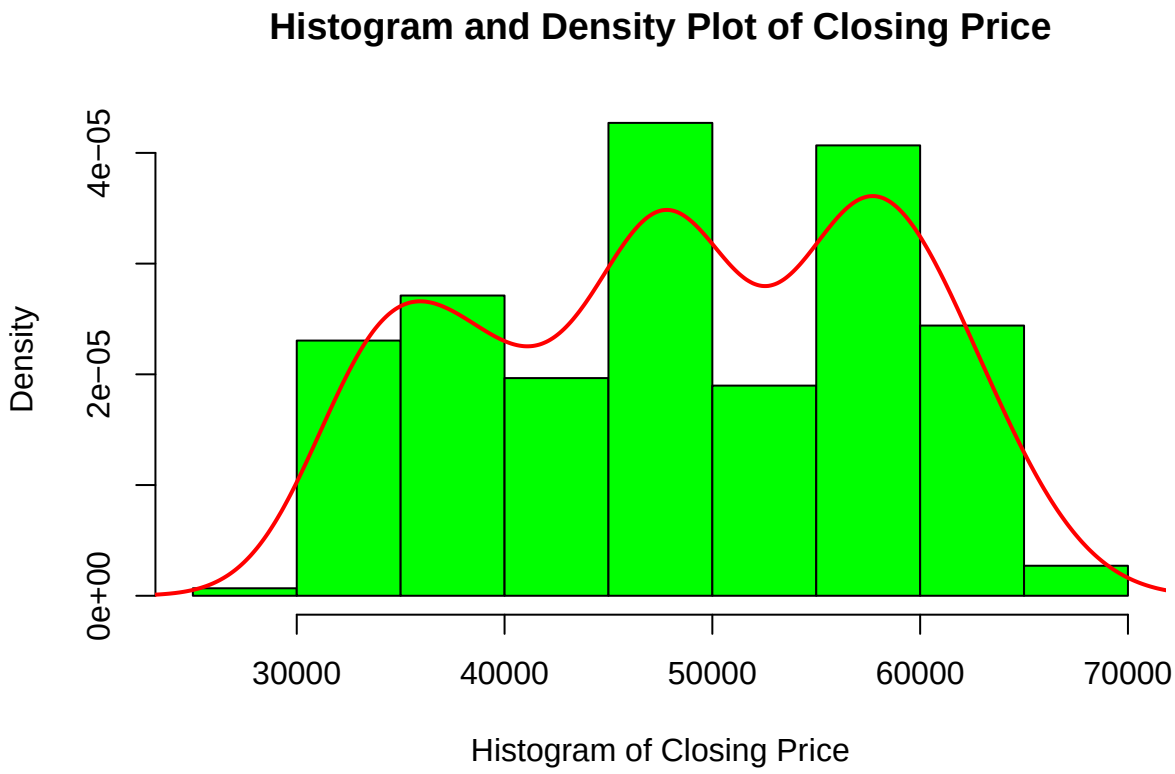
As we can see from the generated histogram, bitcoin more or less has stayed range bound between the range of $30,000 and $60,000 between February and November even though as per the previous graph it was quite volatile during the same period.

Q. What does the histogram and density plot say about Bitcoin's closing price?

```
hist(Bitcoin_price$Close,
     col="green",
     prob = TRUE,
```

```
     xlab = "Histogram of Closing Price",
     main = "Histogram and Density Plot of Closing Price")

lines(density(Bitcoin_price$Close),
      lwd = 2,
      col = "red",
      ylab = "Density Plot of Closing Price")
```



Conclusion

As we can see except for the price ranges of $45,000 - $50,000 and $55,000 - $60,000, the density of Bitcoin's closing price has remained mostly higher than what the histogram plots

Q. What is the probability of a tweet coming in a given month?

```
Tweet_Prob <- Bitcoin_tweets %>% group_by(Month) %>%
  summarize(No.of.tweets = n(),
  Probability.of.Tweet = as.numeric(n() / count(Bitcoin_tweets))) %>%
  arrange(Month)

Tweet_Prob
```

```
## # A tibble: 10 x 3
##    Month No.of.tweets Probability.of.Tweet
##    <int>        <int>                <dbl>
##  1     2        44443              0.0229
##  2     3         4140              0.00213
##  3     4        58060              0.0299
##  4     5        21782              0.0112
##  5     6       125795              0.0647
##  6     7       466079              0.240
##  7     8       488987              0.252
##  8     9        23510              0.0121
##  9    10       351796              0.181
## 10    11       359654              0.185
```

Conclusion

As we can see, July, August has the highest amount of Tweets compared to any other months hence it can be said that these 2 months have a high probability of tweets coming in the future

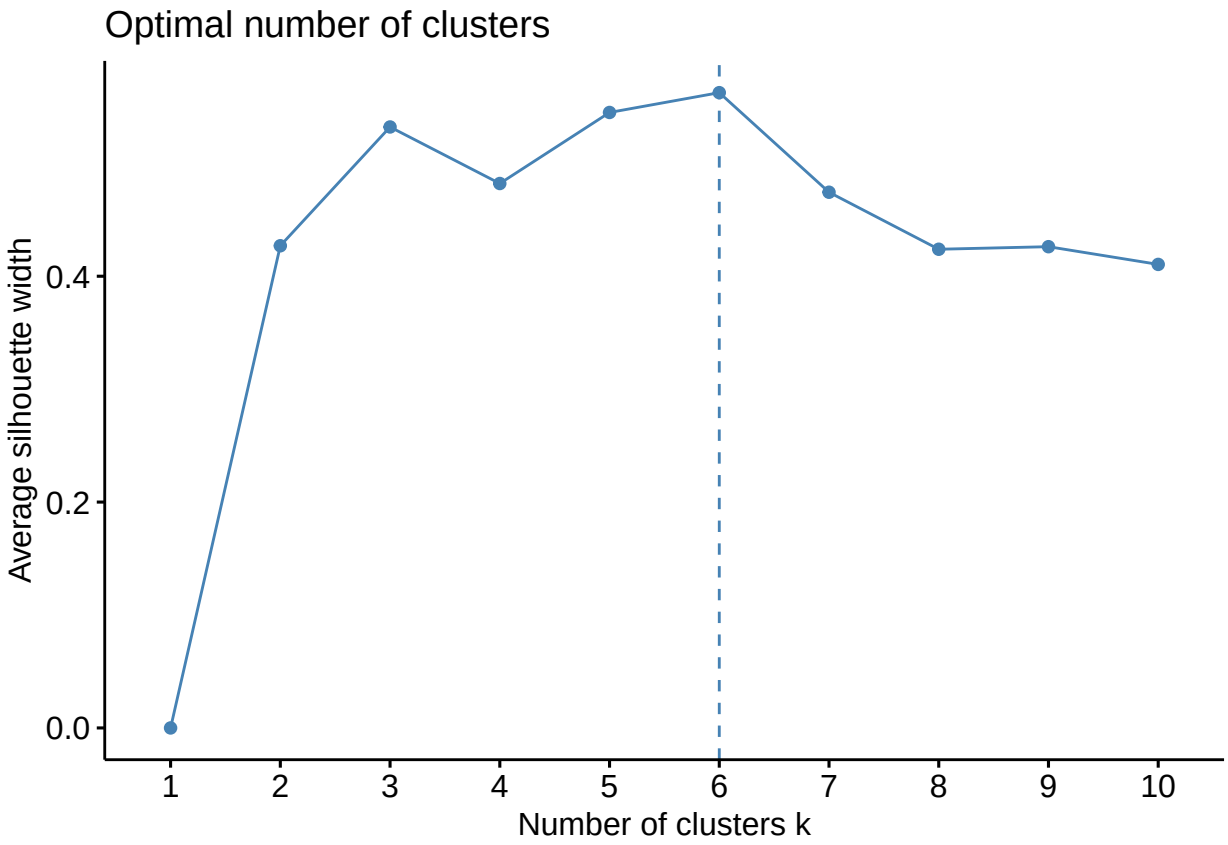
Section 2: Cluster Analysis

Q. Can we cluster bitcoins based on its opening and closing price?

```
coin <- Bitcoin_price%>%
  drop_na()%>%
  group_by(Month)%>%
  select(Open,Close,Month)

coin$Month <- as.numeric(coin$Month)

coin <- scale(coin)
set.seed(123)
fviz_nbclust(coin, kmeans,method = "silhouette")
```

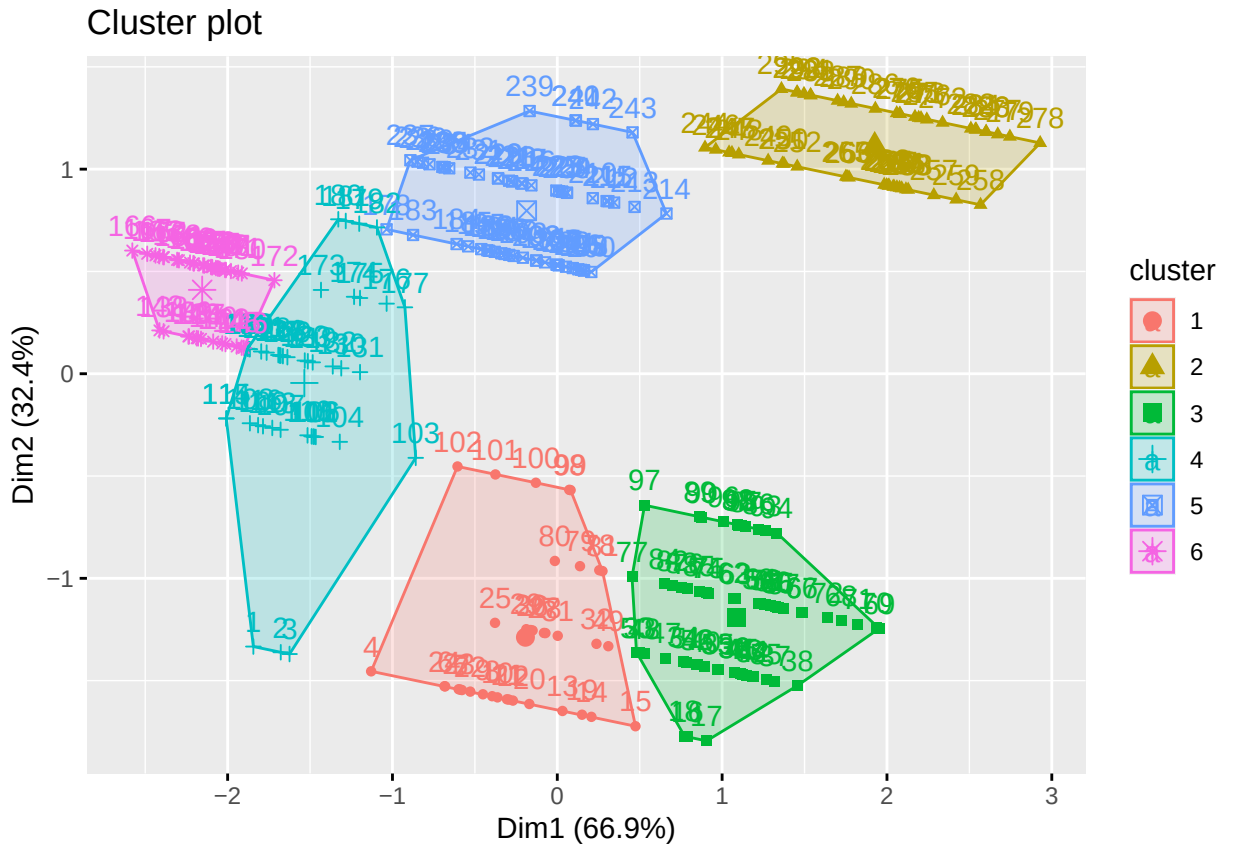



```
km <- kmeans(coin,6, nstart=25)
km
```

```
## K-means clustering with 6 clusters of sizes 36, 52, 63, 43, 62, 39
##
## Cluster means:
##           Open       Close       Month
## 1  0.003872866  0.02766725 -1.30217728
## 2  1.213662737  1.21139712  1.41815224
## 3  0.902204281  0.89019067 -0.99582109
## 4 -1.058586117 -1.07119751 -0.29827364
## 5 -0.227942282 -0.21789842  0.75520403
## 6 -1.549670017 -1.55126894  0.04805927
##
## Clustering vector:
##   [1] 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3
##  [38] 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [75] 3 3 3 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 4 4 4 4 4 4 4 4
## [112] 4 4 4 4 4 4 4 4 4 4 4 4 6 6 6 4 4 4 4 4 4 4 4 4 4 4 6 6 6 6 6 6 6 6 6 6
## [149] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 4 4 4 4 4 5 4 4 4 4 5 5 5
## [186] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
## [223] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [260] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 10.529593 15.249671 13.809367 13.813870 12.878920  3.006994
```

```
##  (between_SS / total_SS =  92.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz <- fviz_cluster(km,coin)
fviz
```



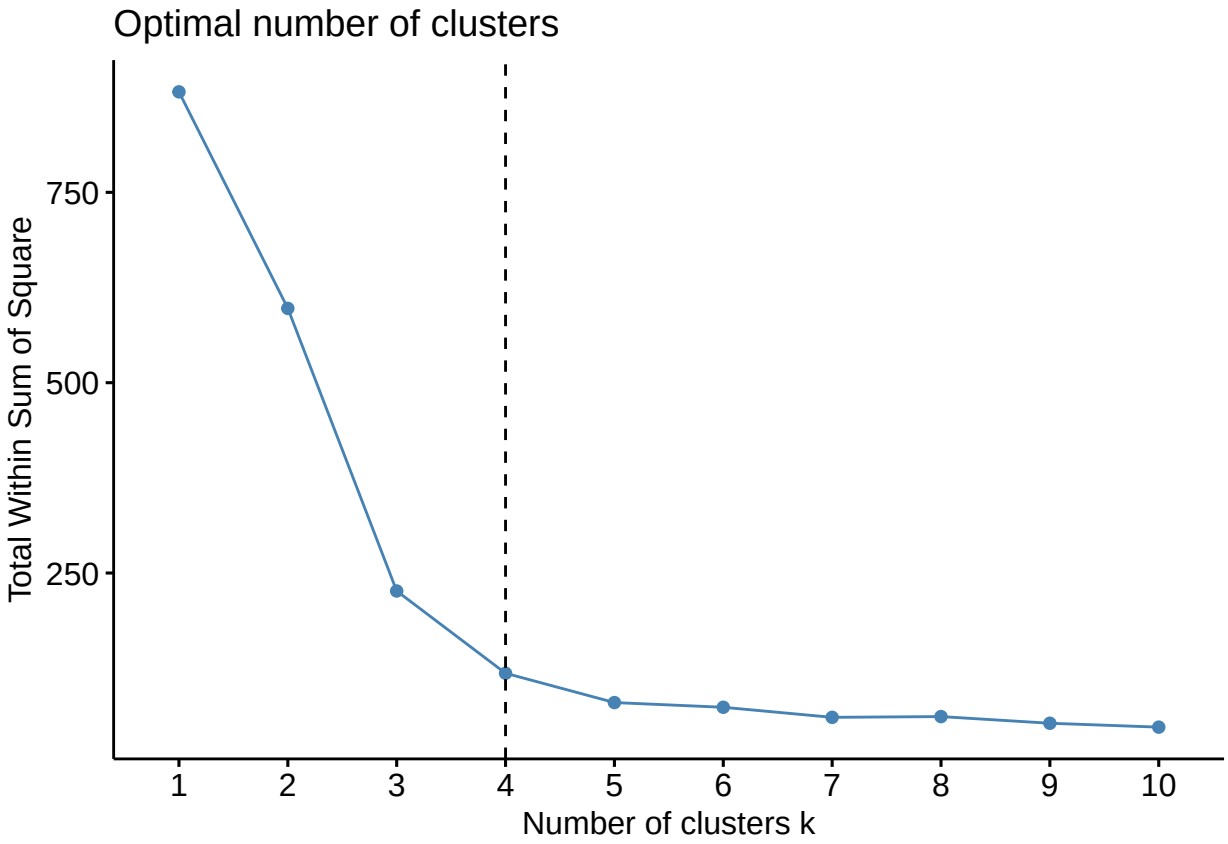
Conclusion

Using the Silhouette Method, we have clustered the Opening and Closing Prices of Bitcoins for any particular day into 6 clusters.

Q. Can we cluster bitcoins based on its day's highest and lowest price?

```
coin1 <- Bitcoin_price%>%
  drop_na()%>%
  select(High,Low)
coin1<-scale(coin)
```

```
set.seed(123)
fviz_nbclust(coin1, kmeans,method = "wss") + geom_vline(xintercept = 4,linetype = 2)
```




```
km <- kmeans(coin1,4, nstart=25)
km
```

```
## K-means clustering with 4 clusters of sizes 66, 96, 52, 81
##
## Cluster means:
##         Open      Close      Month
## 1 -0.2437805 -0.2387596  0.7033211
## 2  0.6080535  0.6048461 -1.1135413
## 3  1.2136627  1.2113971  1.4181522
## 4 -1.3011615 -1.2999968 -0.1637425
##
## Clustering vector:
##   [1] 4 4 4 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 4 4 4 4 4 4 4 4 4 4
## [112] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [149] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 4 4 4 1 1 1 1
## [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [260] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##
```

```
## Within cluster sum of squares by cluster:
## [1] 17.75073 54.77470 15.24967 30.56074
##  (between_SS / total_SS =  86.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz <- fviz_cluster(km,coin1)
fviz
```

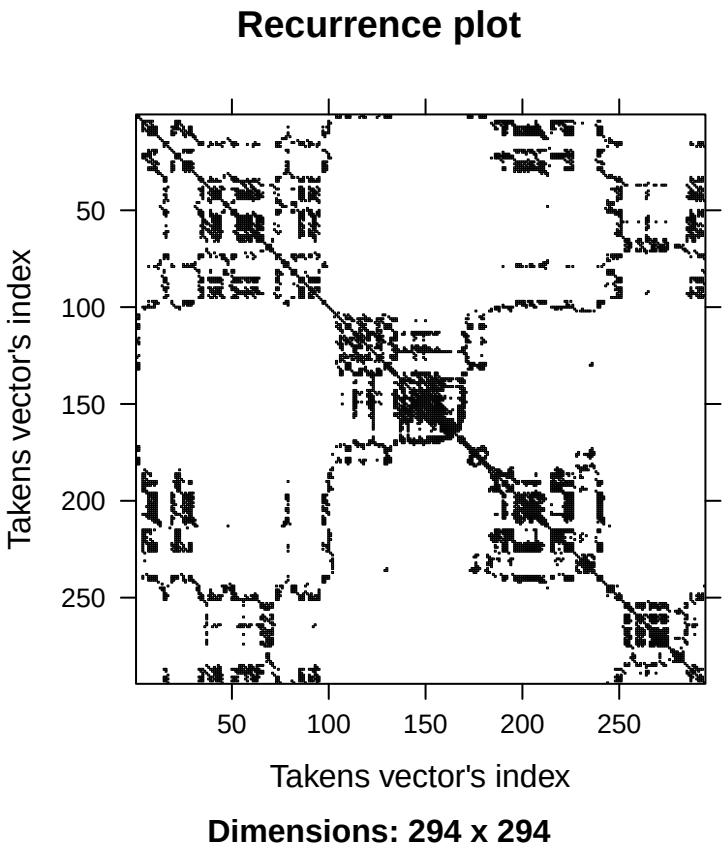
Conclusion

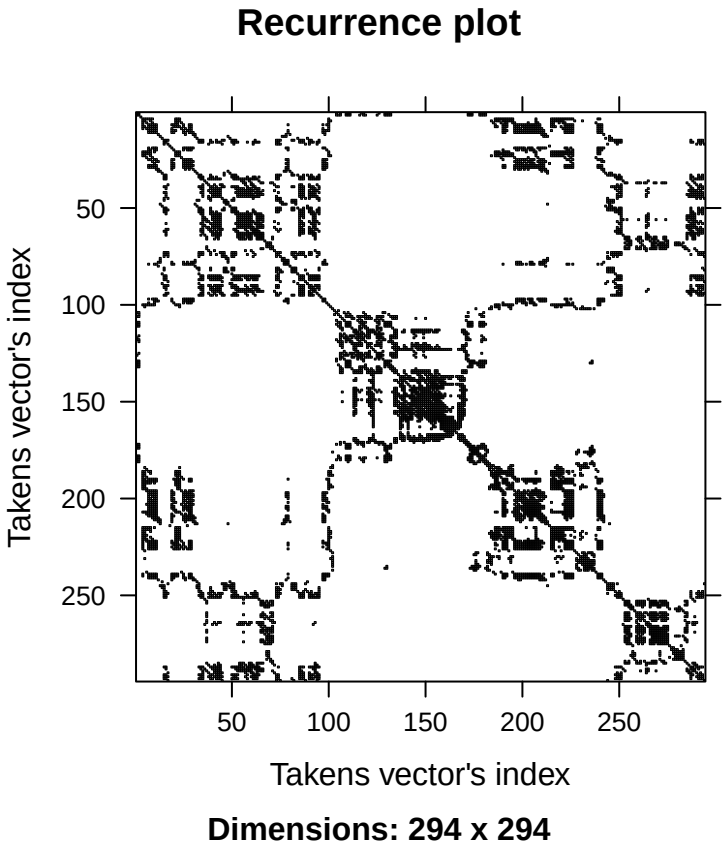
Since the dataset is small, using kmeans clustering, we have grouped the daily prices of bitcoins into 4 clusters to show if the opening price of a cluster comes in the same range as its closing price.

Section 3: Time Series Analysis

Q. What is RQA for Bitcoin's Closing Price in the given time period data available?

```
RQA_Data <- Bitcoin_price$Close
rqa.analysis=rqa(time.series = RQA_Data, embedding.dim=2, time.lag=1,
                 radius=2291,lmin=2,do.plot=FALSE,distanceToBorder=2)
plot(rqa.analysis)
```





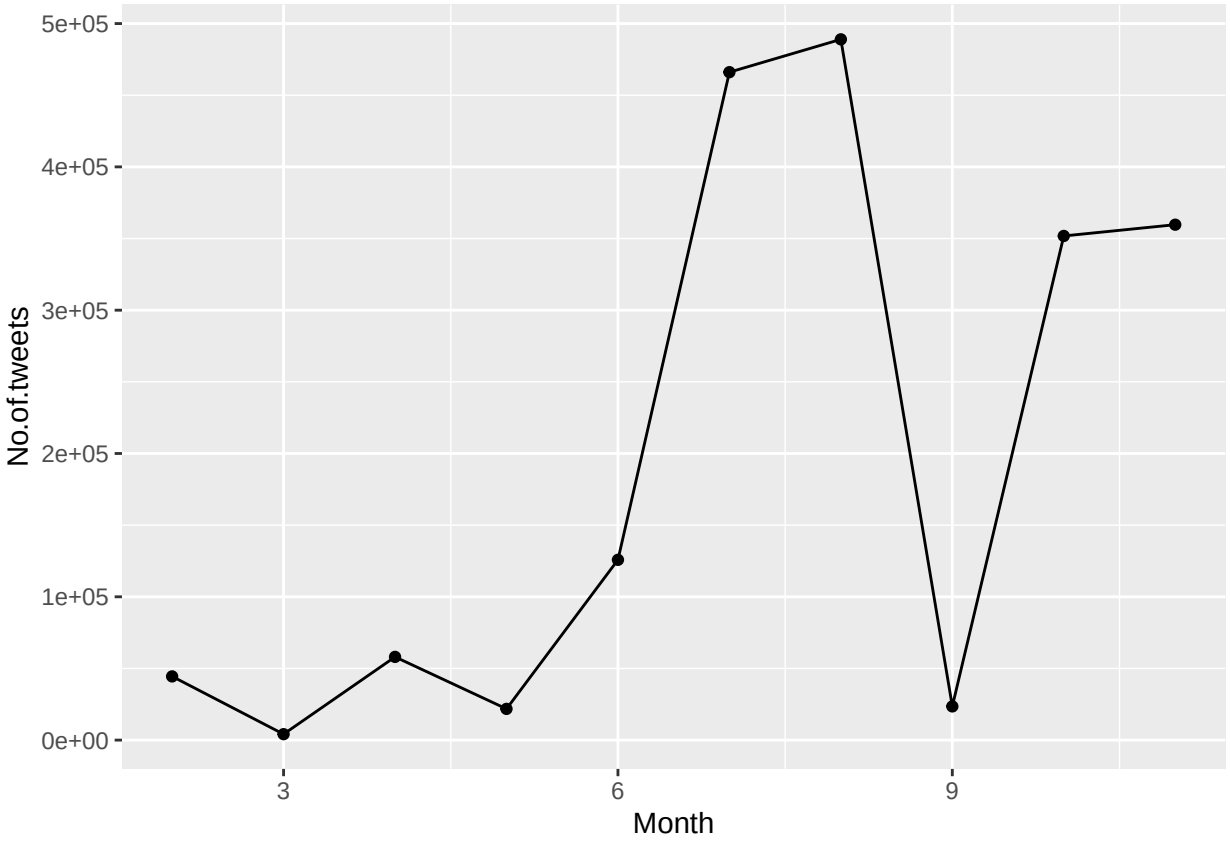
Conclusion

Here we chose the radius as 2291 which is it's ATR (Average True Range) in USD as on 10th December 2021 ATR is an indicator used in Financial Analysis to understand what is the price range of an instrument (Bitcoin in this case) for a given time range. It changes from day to day and gives a general idea of how much the instrument has moved up or down on an average in a given time range. Using the Vector index generated above, we can see that due to the volatility of Bitcoin, the price comes close to the plot line multiple times over the given set of data.

Q. What has been the trend at which the number of tweets that have occured since the inception of the dataset?

```
Monthly_trend <- Bitcoin_tweets %>% group_by(Month) %>% summarize(No.of.tweets = n()) %>% arrange(Month

ggplot(data=Monthly_trend, aes(x=Month, y=No.of.tweets, group=1)) +
  geom_line()+
  geom_point()
```



Conclusion

As we can see, the number of tweets containing the hashtags BTC or Bitcoin saw a drastic jump July onwards continuously except for the Month of September which again saw the Number of tweets come down to a regular level. Since this data was directly picked up from Twitter, there wouldn't be any inconsistency in the count of data available.

Section 4: Text Analysis

Preparing data for Text Analysis

```
# Subsetting the columns to contain only users and their tweets

Sample_Tweets <- Bitcoin_tweets[c(2,11)]

# Tokenizing the data to single word per row format
Tidy_tweets <- Sample_Tweets %>%
  unnest_tokens(word, text)

# Removing the stop words
Tidy_tweets <- Tidy_tweets %>% anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
# Removing additional words

dat <- data.frame(word=c("t.co","https","ðÿ","ðÿš","â","ï","1","à","bitcoin","btc","âž","10","âœ","ð","
                  var2=c("TWITTER","TWITTER","TWITTER","TWITTER","TWITTER","TWITTER","TWITTER","TWITTER
Tidy_tweets <- Tidy_tweets %>% anti_join(dat)
```

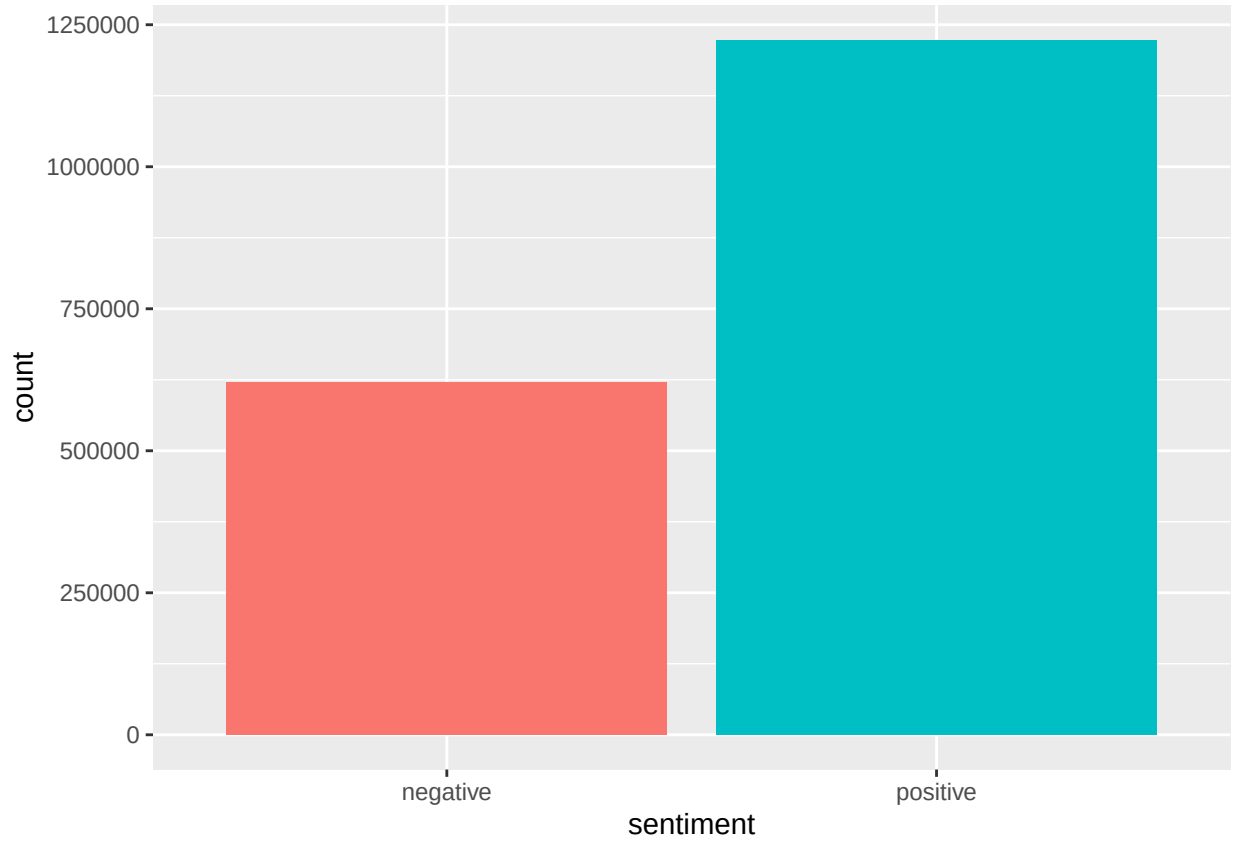
```
## Joining, by = "word"
```

How many keyword tweets that were available had a positive sentiment in favour of bitcoin compared to negative sentiment against bitcoins?

```
pos_neg <- Tidy_tweets %>% inner_join(get_sentiments('bing'))
```

```
## Joining, by = "word"
```

```
ggplot(pos_neg,aes(x=sentiment,fill=sentiment))+
  geom_bar()+
  guides(fill = F)
```

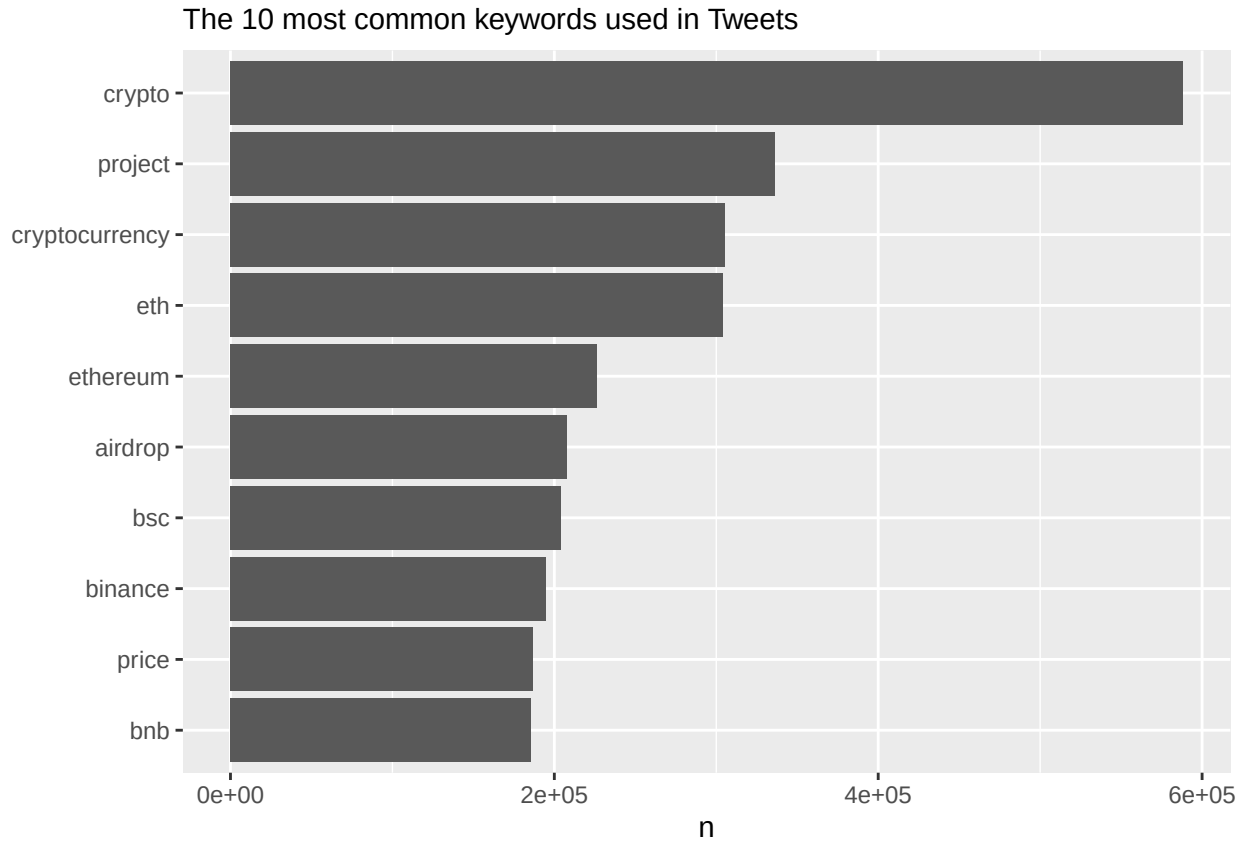


Conclusion

As we can see, the positive sentiment data is almost twice as compared to the negative sentiment data, we can say that approximately every 2 out of 3 tweets for bitcoins support the cryptocurrency.

Which were the most commonly used keywords in Tweets?

```
# Plotting the most popular words
Tidy_tweets %>%
  count(word, sort = TRUE) %>%
  filter(n > 600) %>%  slice(1:10) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word)) +
  geom_col() +
  labs(y = NULL, subtitle = "The 10 most common keywords used in Tweets")
```



Conclusion

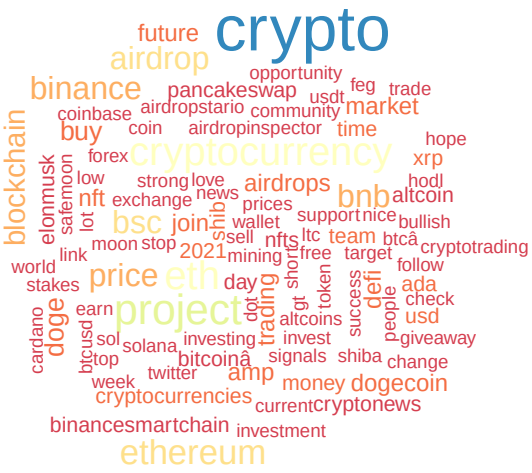
After cleaning the data using two different dictionaries (1 stop words and 1 custom), we can see that the keyword crypto is almost used twice as much as the next keyword in line i.e project. Along with bitcoin, names of some other cryptocurrencies such as Binance, Ethereum etc were also used in these tweets.

Q. What were the 100 most frequently used words in the dataset?

```
wordcloudData1 =
  Tidy_tweets%>%  group_by(word)%>%
  summarize(freq = n())%>%
  arrange(desc(freq))%>%
  ungroup()%>%
  data.frame()

set.seed(617)

wordcloud(words = wordcloudData1$word,wordcloudData1$freq,scale=c(2,0.5),max.words = 100,colors=brewer.
```



Conclusion

Using the Word Cloud, we wanted toshowcase which are the most used words in Tweets. As we can see even though the tweets are extracted using the hashtags #Bitcoin and #BTC, we can see that other cryptocurrencies such as Ripple, Binance, Ethereum, Doge, Shiba etc have been included as well.

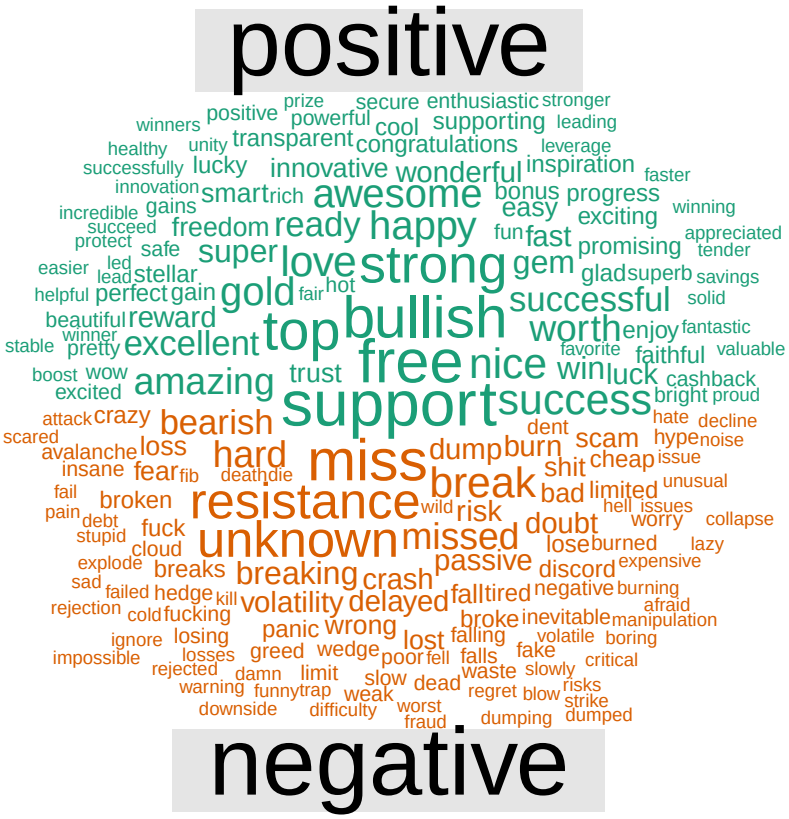
Q. Which positive and negative keywords were used in the Tweets?

```
wordcloudData2 =
  Tidy_tweets %>%  inner_join(get_sentiments('bing'))%>%
  count(sentiment,word,sort=T)%>%
  spread(key=sentiment,value = n,fill=0)%>%
  data.frame()
```

```
## Joining, by = "word"
```

```
rownames(wordcloudData2) = wordcloudData2[,'word']
wordcloudData2 = wordcloudData2[,c('positive','negative')]

set.seed(617)
comparison.cloud(term.matrix = wordcloudData2,scale = c(2,0.5),max.words = 200, rot.per=0)
```



Conclusion

Using the Word Cloud above, in the positive word cloud section, one keyword that stands out along with free, support, strong, gold etc is “bullish” which is a keyword which is normally used in the trading markets to indicate that people are highly positive about a particular stock / cryptocurrency / any other instrument. At the same time, for the negative keywords, words such as resistance, miss, bearish, risk, volatility etc are used which are keywords used in Technical Analysis to indicate that based on technical analysis, the price of bitcoin might go down.