Shashwat Pandya 17101A0038
Saurabh Naik 17101A0050

R Miniproject

Sentiment Analysis of r/wallstreetbets

- **Dataset Information**
- **Context**

Recently, members of the subreddit have influenced the stock market and driven prices of game stock (GME) stock up by around 2000%. So we believe sentiment analysis of the posts of the community might give us some interesting insights

We will implement the sentiment analysis processing part using syuzhet library In R while ggplot2 will be used for data visualization and presentation.

- **Objective**

As there it was a tumultuous period for the stock market in the US, the users of the subreddit took advantage of the situation and managed to make millions in profits while discussing over the subreddit. Analysis of the sentiments of the users during this period is bound to give insights into how the users managed to make the best of such an adverse situation.

- **Data Description**
  a. Title : The post title submitted by the user
  b. Score : The number of upvotes that the post received
  c. ID : Unique ID generated by reddit assigned to the post
  d. url : The url of the post as generated by reddit
  e. comms_num : The number of comments on the post
  f. body : The body of the post, the actual material posted by the original poster
  g. created: timestamp of the post recorded by reddit

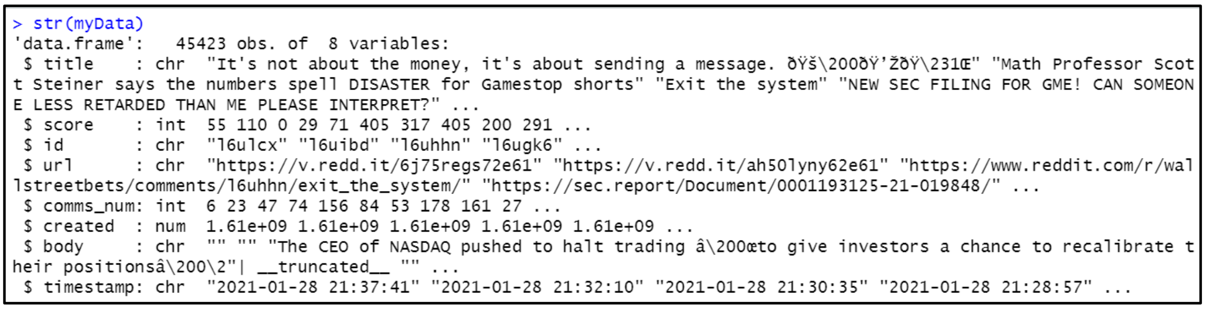
h. timestamp: timestamp of the post in real time format

The dataset was provided by Gabriel Preda on Kaggle with the intention of sentiment analysis of the posts during a period of interest

- **Data Exploration**

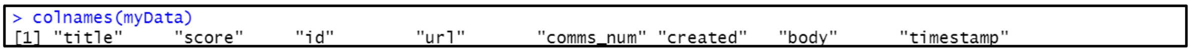
  The dataset contains multiple columns and to understand them better we use data exploration techniques to get a basic idea of the data

  a. Structure of the data in the dataset



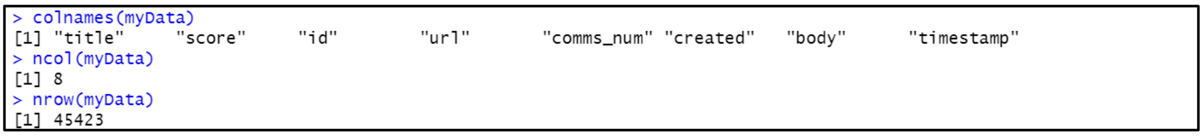
```
> str(myData)
'data.frame':	45423 obs. of  8 variables:
 $ title    : chr  "It's not about the money, it's about sending a message. ðŸš\200ðŸ’ŽðŸ\231Œ" "Math Professor Scot
t Steiner says the numbers spell DISASTER for Gamestop shorts" "Exit the system" "NEW SEC FILING FOR GME! CAN SOMEON
E LESS RETARDED THAN ME PLEASE INTERPRET?" ...
 $ score    : int  55 110 0 29 71 405 317 405 200 291 ...
 $ id       : chr  "l6ulcx" "l6uibd" "l6uhhn" "l6ugk6" ...
 $ url      : chr  "https://v.redd.it/6j75regs72e61" "https://v.redd.it/ah50lyny62e61" "https://www.reddit.com/r/wal
lstreetbets/comments/l6uhhn/exit_the_system/" "https://sec.report/Document/0001193125-21-019848/" ...
 $ comms_num: int  6 23 47 74 156 84 53 178 161 27 ...
 $ created  : num  1.61e+09 1.61e+09 1.61e+09 1.61e+09 1.61e+09 ...
 $ body     : chr  "" "" "The CEO of NASDAQ pushed to halt trading â\200œto give investors a chance to recalibrate t
heir positionsâ\200\2"| __truncated__ "" ...
 $ timestamp: chr  "2021-01-28 21:37:41" "2021-01-28 21:32:10" "2021-01-28 21:30:35" "2021-01-28 21:28:57" ...
```

  b. The names of the columns



```
> colnames(myData)
[1] "title"     "score"     "id"        "url"       "comms_num" "created"   "body"      "timestamp"
```

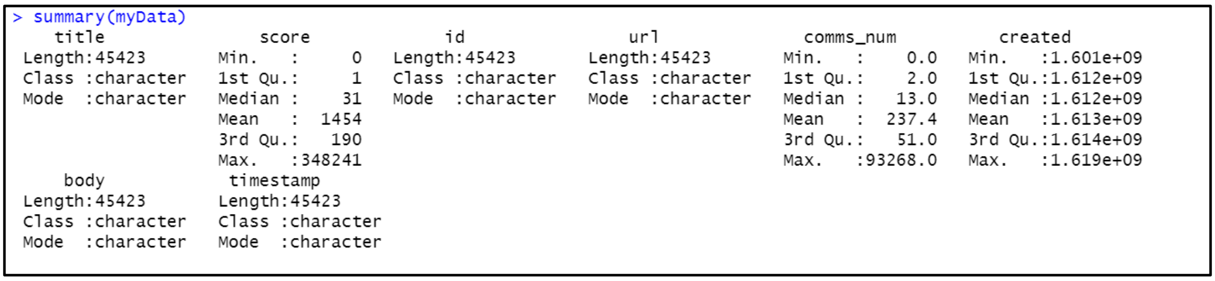
  c. Number of columns and rows



```
> colnames(myData)
[1] "title"     "score"     "id"        "url"       "comms_num" "created"   "body"      "timestamp"
> ncol(myData)
[1] 8
> nrow(myData)
[1] 45423
```

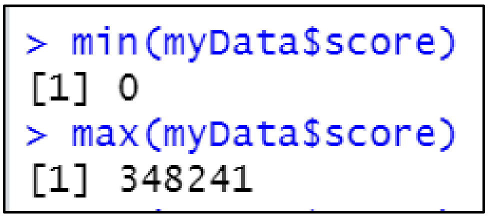
  d. Summary of the data

Shashwat Pandya 17101A0038
Saurabh Naik 17101A0050



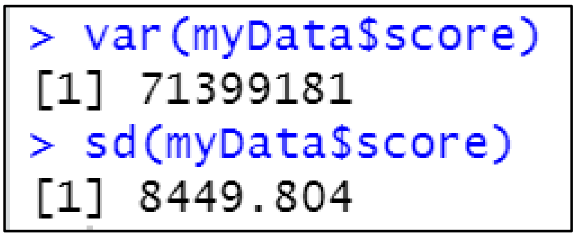
```
> summary(myData)
    title               score             id                url              comms_num          created         
 Length:45423       Min.   :     0   Length:45423       Length:45423       Min.   :    0.0   Min.   :1.601e+09  
 Class :character   1st Qu.:     1   Class :character   Class :character   1st Qu.:    2.0   1st Qu.:1.612e+09  
 Mode  :character   Median :    31   Mode  :character   Mode  :character   Median :   13.0   Median :1.612e+09  
                    Mean   :  1454                                         Mean   :  237.4   Mean   :1.613e+09  
                    3rd Qu.:   190                                         3rd Qu.:   51.0   3rd Qu.:1.614e+09  
                    Max.   :348241                                         Max.   :93268.0   Max.   :1.619e+09  
     body            timestamp        
 Length:45423       Length:45423      
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

e. Min, max of the score column



```
> min(myData$score)
[1] 0
> max(myData$score)
[1] 348241
```

f. Variance and standard deviation of the score column

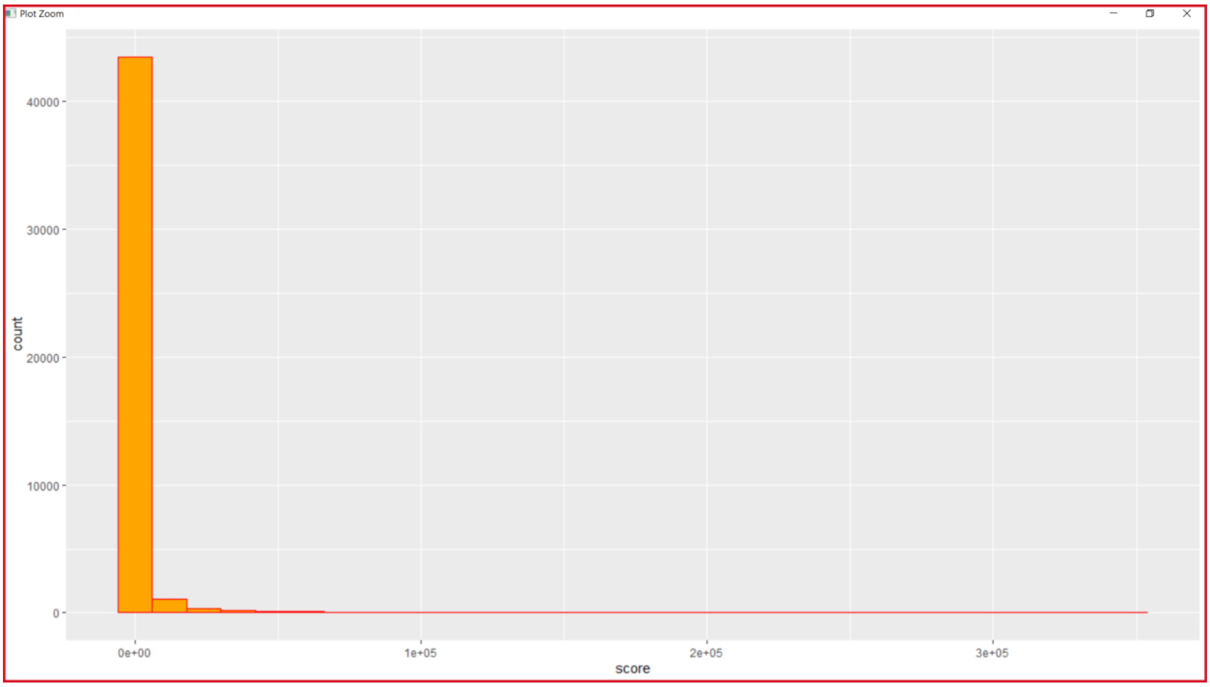


```
> var(myData$score)
[1] 71399181
> sd(myData$score)
[1] 8449.804
```

- ## Data Visualization

We create plots of various data columns to understand if a relationship exists between the columns

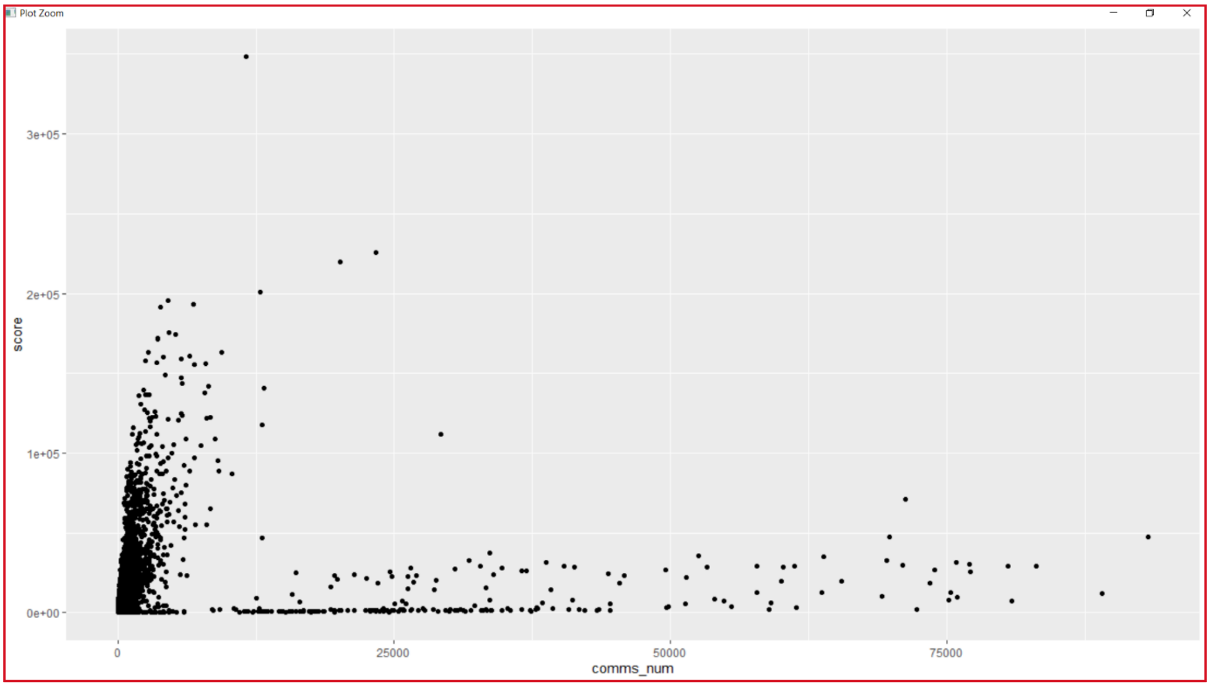
a. Score count histogram



This histogram depicts the number of posts that exceed a given score or number of upvotes. We can see that most of the posts are concentrated in the head of the graph, with a low score.

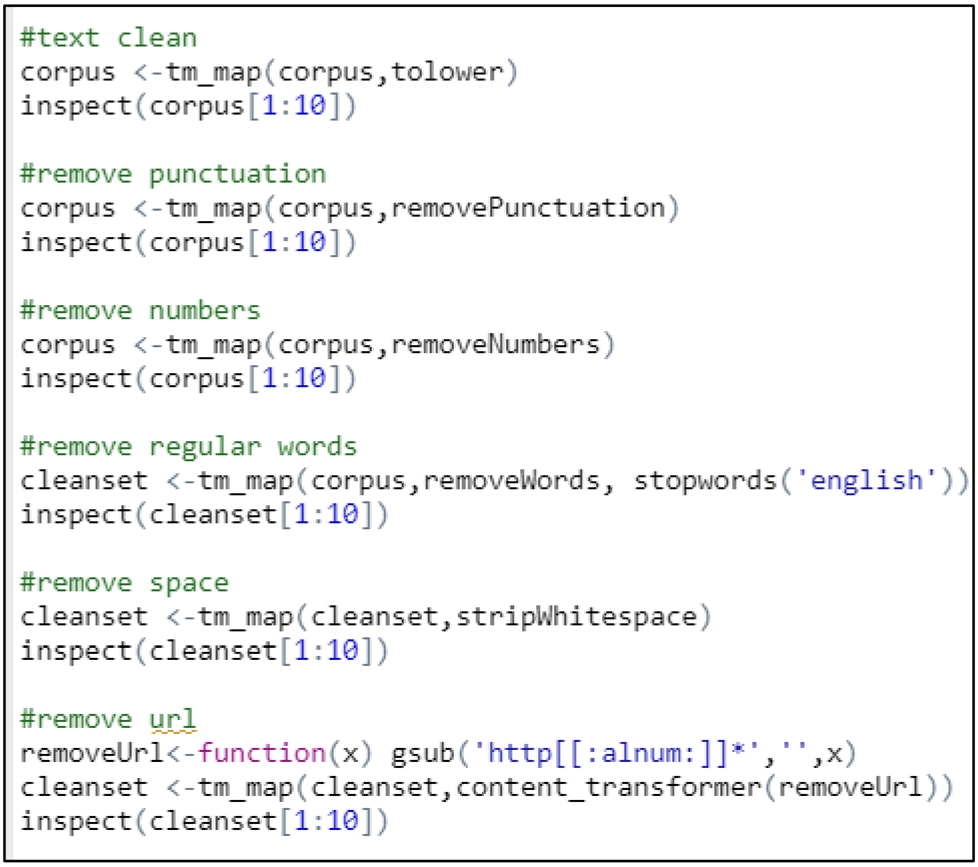
b. Score vs number of comments scatter plot

Shashwat Pandya 17101A0038
Saurabh Naik 17101A0050



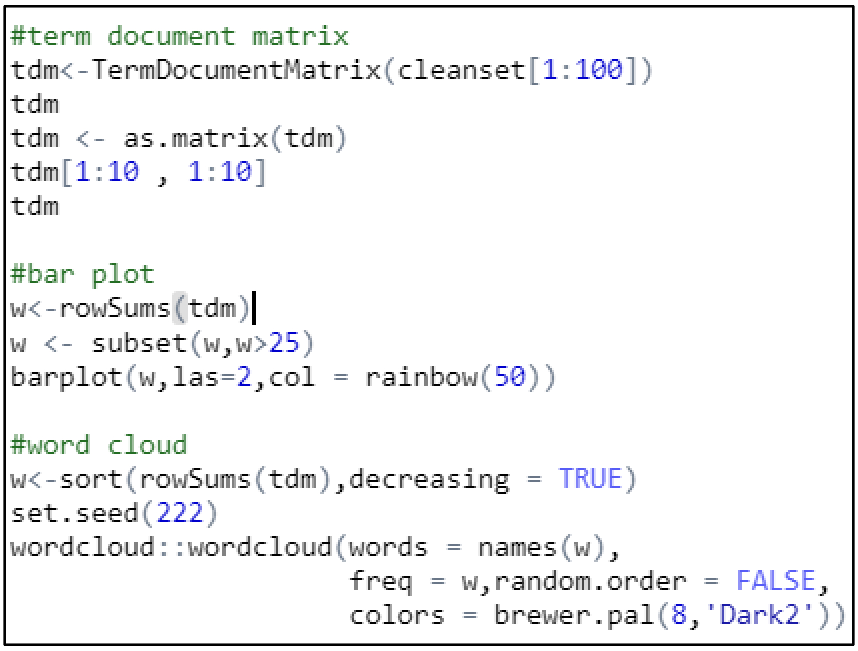
Here we can observe that as the number of comments increases, the score of the post decreases. This indicates that posts which have users with different perspectives do not reach higher scores.

- **Processed data, observations and visualization**
  Now that we have performed visualization on various columns of the dataset we decided to perform some analysis on the body column of our dataset that consists of actual material posted by the poster. But we observed that the data in this column consisted of many unnecessary components so we decided to perform some data cleaning on this column before performing visualization on it



```
#text clean
corpus <-tm_map(corpus,tolower)
inspect(corpus[1:10])

#remove punctuation
corpus <-tm_map(corpus,removePunctuation)
inspect(corpus[1:10])

#remove numbers
corpus <-tm_map(corpus,removeNumbers)
inspect(corpus[1:10])

#remove regular words
cleanset <-tm_map(corpus,removeWords, stopwords('english'))
inspect(cleanset[1:10])

#remove space
cleanset <-tm_map(cleanset,stripWhitespace)
inspect(cleanset[1:10])

#remove url
removeUrl<-function(x) gsub('http[[:alnum:]]*','',x)
cleanset <-tm_map(cleanset,content_transformer(removeUrl))
inspect(cleanset[1:10])
```

First we collected all data of the body column in corpus variable and used tm_map function for data cleaning. Data cleaning consisted of removing empty elements, removing URLs, removing stop words and removing numbers and punctuation marks.

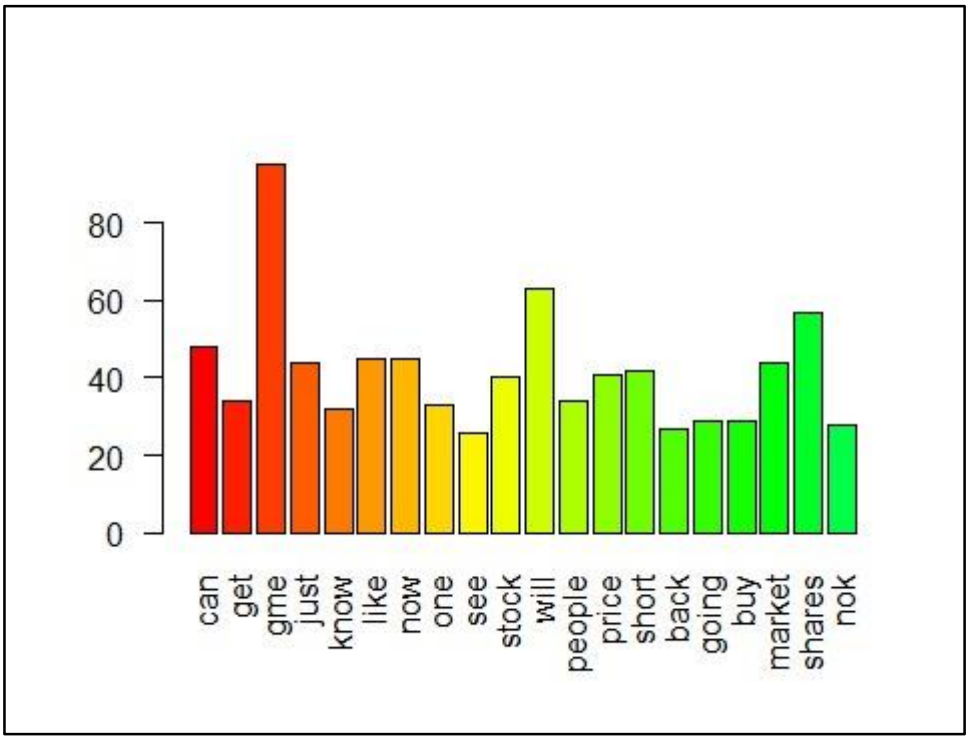


```
#term document matrix
tdm<-TermDocumentMatrix(cleanset[1:100])
tdm
tdm <- as.matrix(tdm)
tdm[1:10 , 1:10]
tdm

#bar plot
w<-rowSums(tdm)
w <- subset(w,w>25)
barplot(w,las=2,col = rainbow(50))

#word cloud
w<-sort(rowSums(tdm),decreasing = TRUE)
set.seed(222)
wordcloud::wordcloud(words = names(w),
                     freq = w,random.order = FALSE,
                     colors = brewer.pal(8,'Dark2'))
```

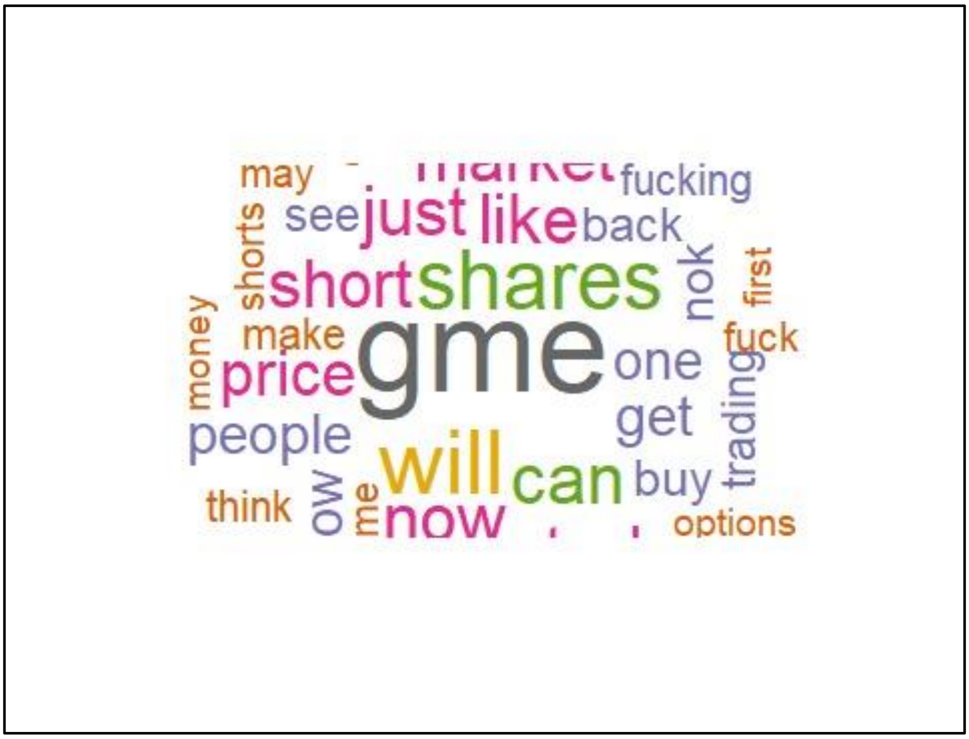
Now, the cleaned data is converted into a matrix by giving it into the TermDocumentMatrix and then a matrix having 10 rows and 10 columns are created in order to perform visualization on this clean data. This cleaned data consisting most frequently found words are displayed both using a bar plot as well as a word cloud.

Shashwat Pandya 17101A0038
Saurabh Naik 17101A0050

Bar plot of most frequently found words



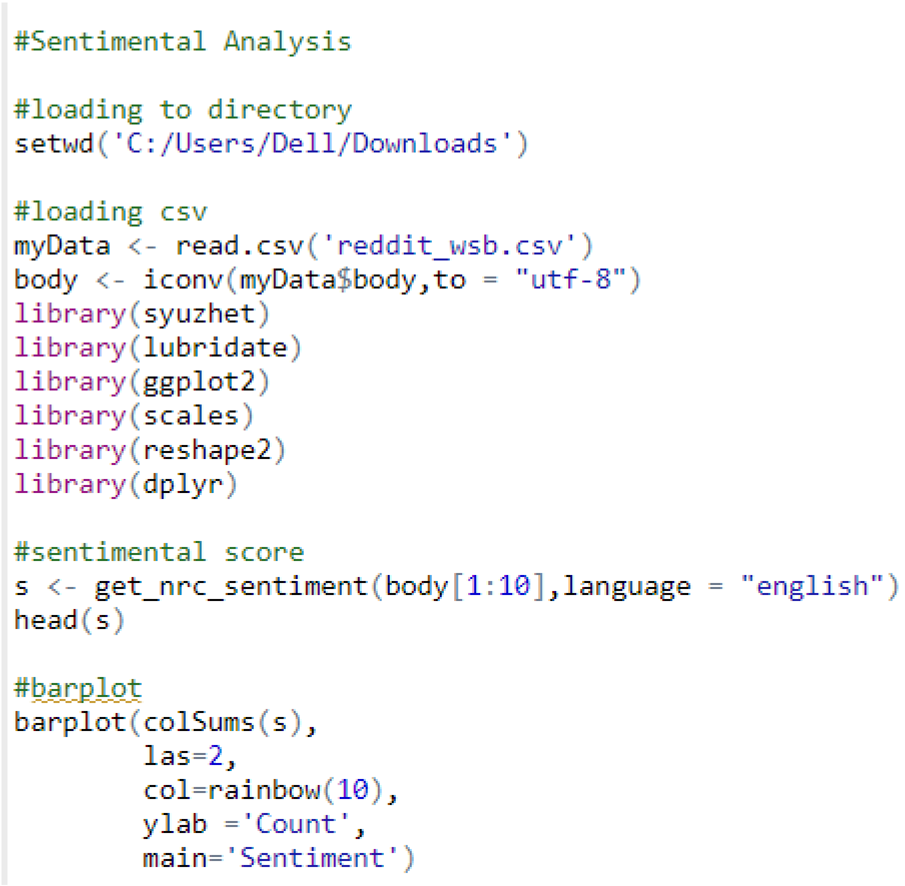
Word Cloud of most frequently found words



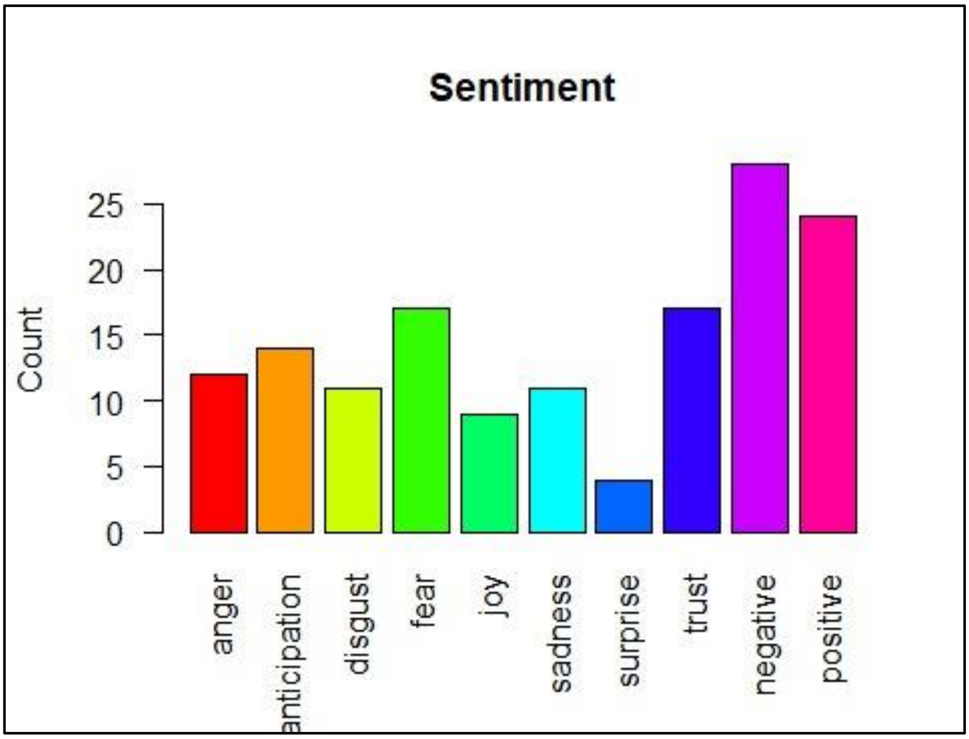
- **Sentiment analysis and visualization**

  Now that the cleaned data is obtain. We can perform sentimental analysis on it. Sentimental analysis is done using the package syuzhet. Based on the different moods like anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative and positive a bar plot is shown.

Shashwat Pandya 17101A0038
Saurabh Naik 17101A0050



```
#Sentimental Analysis

#loading to directory
setwd('C:/Users/Dell/Downloads')

#loading csv
myData <- read.csv('reddit_wsb.csv')
body <- iconv(myData$body,to = "utf-8")
library(syuzhet)
library(lubridate)
library(ggplot2)
library(scales)
library(reshape2)
library(dplyr)

#sentimental score
s <- get_nrc_sentiment(body[1:10],language = "english")
head(s)

#barplot
barplot(colSums(s),
        las=2,
        col=rainbow(10),
        ylab ='Count',
        main='Sentiment')
```


Sentiment
Count
anger
anticipation
disgust
fear
joy
sadness
surprise
trust
negative
positive