FIT3152 Assignment 1 report

By: Ko Ko Win

Load all the necessary libraries

```
library(ggplot2)
library(ggvis)
library(tidyverse)
library(lubridate)
library(dplyr)
library(feasts)
library(tsibble)
library(MASS)
library(reshape2)
library(reshape)
library(cowplot)
library(knitr)
```

Load the data set

```
rm(list = ls())
set.seed(31842305)
webforum = read.csv("webforum.csv",  header=T, na.strings='')
webforum = webforum[sample(nrow(webforum), 2000), ] #2000 rows
```

Data preprocessing

```
head(webforum)
```

```
##       ThreadID AuthorID       Date  Time  WC Analytic Clout Authentic  Tone
## 27193   310624    43162 2006-07-24 21:10  86    85.30 45.37     77.74 86.31
## 20046   161824    44490 2004-10-28 15:55  20    96.54 69.14     74.76  1.00
## 7238    255811     7160 2005-12-19 02:04  81    68.94 54.90     26.79 48.69
## 14066   652643   196951 2009-11-08 08:38 244    79.40 76.96      1.71 25.77
## 7196    833308   231141 2011-09-20 23:47 124    93.58 40.44     17.20  3.16
## 4268    265706    73999 2006-01-27 01:05  11    79.25 18.16     99.00 25.77
##       ppron     i   we  you shehe they posemo negemo anx anger  sad focuspast
## 27193 10.47  8.14 0.00 1.16  1.16 0.00   3.49   0.00   0  0.00 0.00      9.30
## 20046 10.00  5.00 0.00 0.00  0.00 5.00   0.00   5.00   0  0.00 0.00      0.00
## 7238   2.47  2.47 0.00 0.00  0.00 0.00   3.70   2.47   0  2.47 0.00      4.94
## 14066  2.46  0.41 0.41 0.00  0.41 1.23   2.87   2.87   0  1.23 0.00      3.69
## 7196   0.00  0.00 0.00 0.00  0.00 0.00   0.81   3.23   0  1.61 0.81      0.81
## 4268  18.18 18.18 0.00 0.00  0.00 0.00   0.00   0.00   0  0.00 0.00      0.00
##       focuspresent focusfuture
## 27193         5.81        0.00
## 20046        15.00        0.00
## 7238          6.17        0.00
## 14066        12.30        0.00
## 7196          9.68        0.81
## 4268          9.09        0.00
```

```
print(paste("The number of rows with 0 WC is: ", with(webforum, length(WC[WC == 0]))))
```

```
## [1] "The number of rows with 0 WC is:  10"
```

```
#remove rows with 0 WC
#webforum[webforum$WC==0,]
webforum = webforum[webforum$WC != 0,]
#remove AuthorID with -1
webforum = webforum[webforum$AuthorID != -1,]
```

```
lv_summary = summary(webforum[,5:23])
lv_summary
```

```
##        WC            Analytic         Clout         Authentic
##  Min.   :   1.0   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00
##  1st Qu.:  26.0   1st Qu.:39.45   1st Qu.:41.98   1st Qu.: 7.84
##  Median :  60.0   Median :64.77   Median :61.58   Median :29.19
##  Mean   : 104.7   Mean   :60.51   Mean   :59.33   Mean   :36.92
##  3rd Qu.: 124.0   3rd Qu.:84.09   3rd Qu.:81.23   3rd Qu.:61.34
##  Max.   :6848.0   Max.   :99.00   Max.   :99.00   Max.   :99.00
##       Tone           ppron              i                we
##  Min.   : 1.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000
##  1st Qu.:14.91   1st Qu.: 4.170   1st Qu.: 0.000   1st Qu.: 0.0000
##  Median :25.77   Median : 7.140   Median : 2.060   Median : 0.0000
##  Mean   :44.50   Mean   : 7.612   Mean   : 3.271   Mean   : 0.8932
##  3rd Qu.:80.64   3rd Qu.:10.420   3rd Qu.: 4.835   3rd Qu.: 1.1400
##  Max.   :99.00   Max.   :50.000   Max.   :50.000   Max.   :16.6700
##       you             shehe              they            posemo
##  Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 0.0000   Median : 0.000   Median : 2.420
##  Mean   : 1.401   Mean   : 0.6915   Mean   : 1.355   Mean   : 3.549
##  3rd Qu.: 1.800   3rd Qu.: 0.0000   3rd Qu.: 1.960   3rd Qu.: 4.440
##  Max.   :25.000   Max.   :16.6700   Max.   :23.530   Max.   :75.000
##      negemo             anx               anger               sad
##  Min.   :  0.000   Min.   : 0.0000   Min.   :  0.0000   Min.   : 0.0000
##  1st Qu.:  0.000   1st Qu.: 0.0000   1st Qu.:  0.0000   1st Qu.: 0.0000
##  Median :  1.350   Median : 0.0000   Median :  0.0000   Median : 0.0000
##  Mean   :  2.235   Mean   : 0.2729   Mean   :  0.9022   Mean   : 0.2832
##  3rd Qu.:  3.080   3rd Qu.: 0.0000   3rd Qu.:  1.1900   3rd Qu.: 0.0000
##  Max.   :100.000   Max.   :25.0000   Max.   :100.0000   Max.   :16.6700
##    focuspast       focuspresent     focusfuture
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 7.185   1st Qu.: 0.000
##  Median : 2.630   Median :10.340   Median : 0.000
##  Mean   : 3.417   Mean   :10.701   Mean   : 1.085
##  3rd Qu.: 5.080   3rd Qu.:13.505   3rd Qu.: 1.520
##  Max.   :50.000   Max.   :50.000   Max.   :22.220
```

Variables WC, Analytic,Clout, Authentic and Tone has realtively higher mean value compared to other variables which indicates that other vairables are highly dependant on those variables.

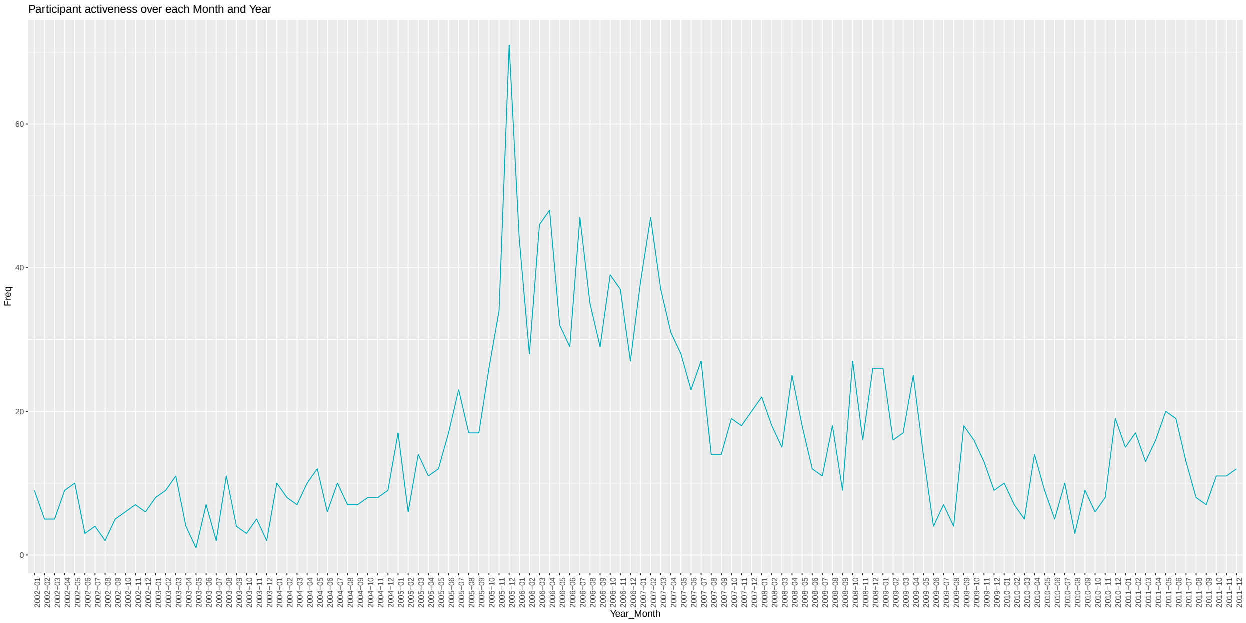
1a)

```
webforum$Date = as.Date(webforum$Date)
webforum$month_year = format(as.Date(webforum$Date), "%Y-%m")
```

```


by_yearmonth = as.data.frame(as.table(tapply(webforum$ThreadID, list(webforum$month_year), length)))
colnames(by_yearmonth) = c("Year_Month", "Freq")
```

```


#Plot to see the activeness
yearmonth_plot <- ggplot(data = by_yearmonth, aes(x=Year_Month, y= Freq, group=1))+
                  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
                  geom_line(color = "#00AFBB") + ggtitle("Participant activeness over each Month and Yea

yearmonth_plot
```

Question: How active are participants over the longer term(thatis,over months and/or years)?



Question: Are there periods where activity increases or decreases?

Answer: By looking at the plot we have done on the pariticiapant acitveness over the course of 9 years we can see that there seem to be an increase in number of participants at the start of every year. This maybe because, the participants have long holiday due to christmas and new year holiday. However, this is not the case for the year 2011 as we can see a decrease in pariticipant.

Throughout the period of 9 years, December 2005 has highest peak in number of participant comapared to rest of the years. By looking at the wikipedia on internet users in 2005 we can see that Americans and Europeans used the internet most in 2005[1]. On december 2005, Britain legalise same sex marriage[2] and famous footballer Roanldinho was named the best FIFA football player in the world on 2005 december as well [3]. It may be the causes of high number of interaction on the forum.

References: https://en.wikipedia.org/wiki/Global_Internet_usage [1]

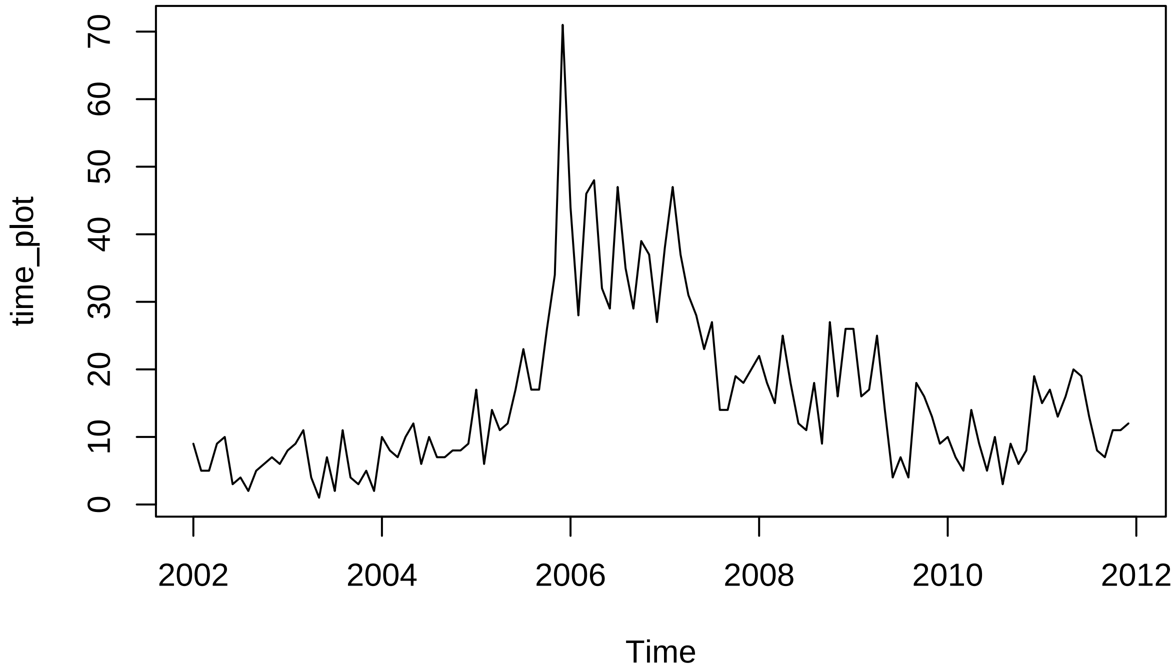
: https://www.neweurope.eu/article/key-events-2005/ [2]

: https://www.ronaldinho10.com/en_biographie_barcelone.htm [3]

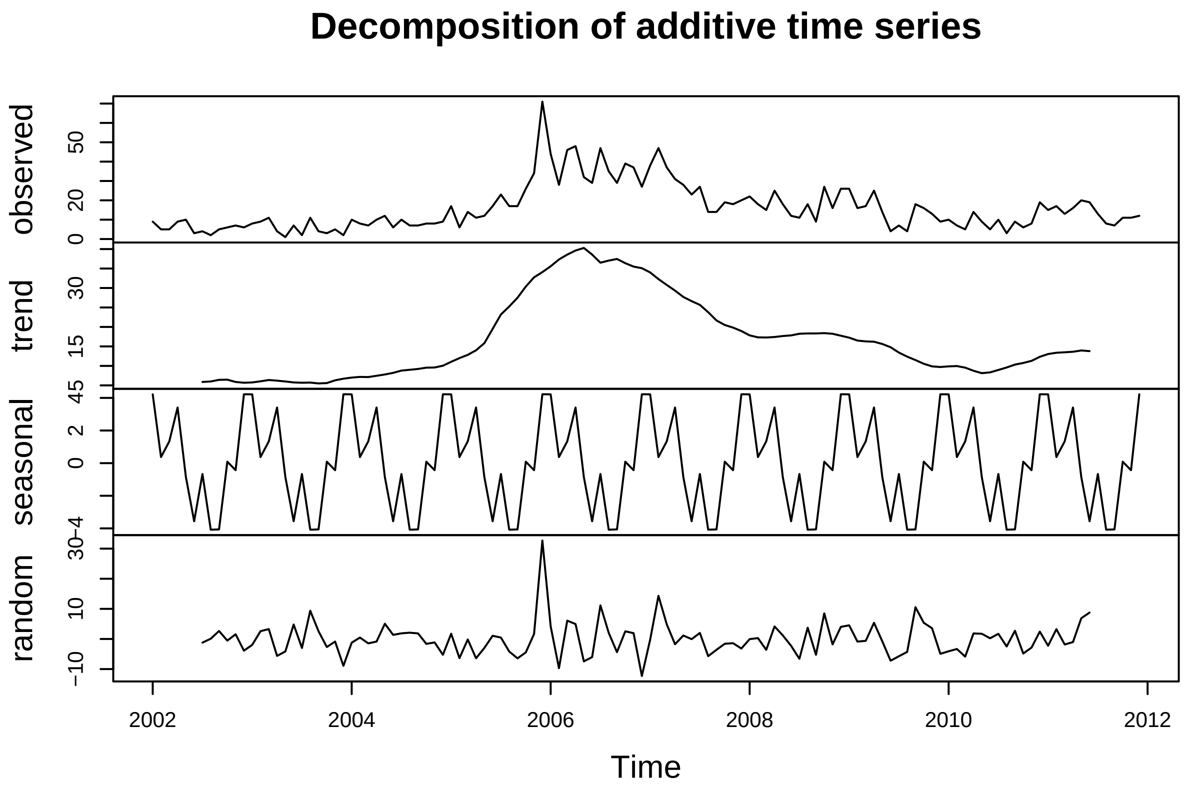
Is there a trend over time?

```
#remove.packages("igraph")
#remove.packages("igraphdata")
time_plot <- ts(by_yearmonth[,2], start = c(2002,01), frequency = 12)
plot(time_plot)
```

By looking at the trend in time-series decomposed plot it shows that number of threads have gradually increases from 2002 onwards until some where around 2006. However, starting from some month in 2006 the number of threads started to fall which shows that participants became less active starting from 2006. It maybe because participants have started to lose intrest in the topics being discussed in the forum or participants spend time on other social platfomr such as Twitter, Myspace or Facebook.



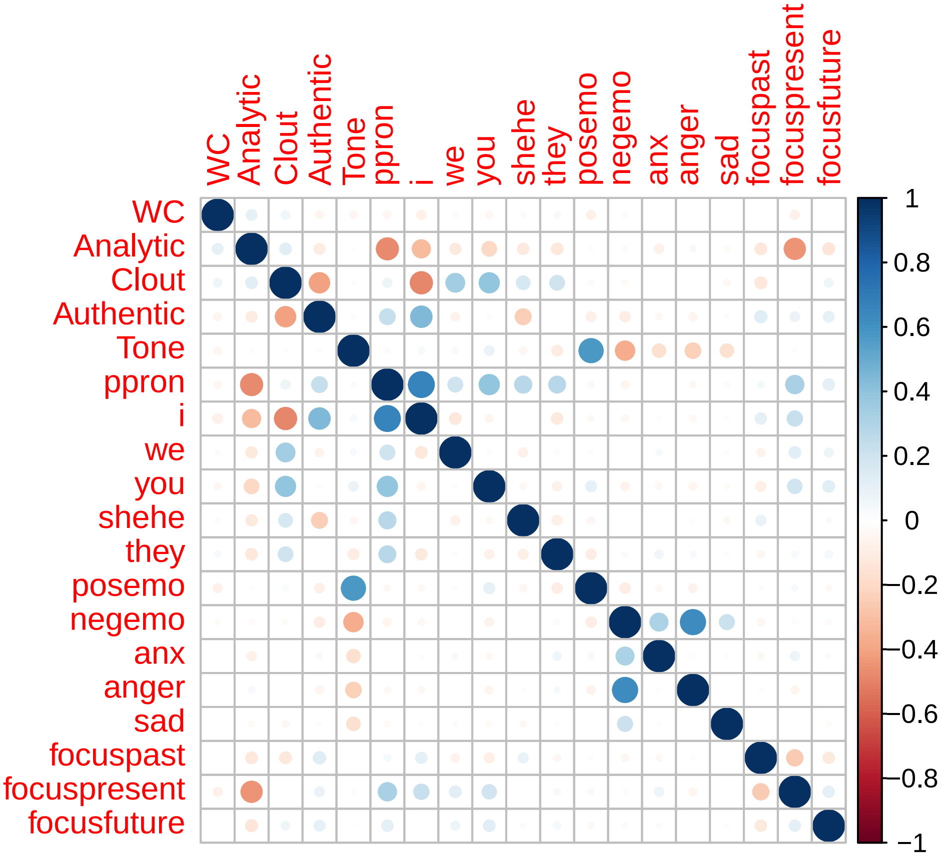
```
m <- decompose(time_plot)
plot(m)
```



2a)

```r
library(corrplot)
corr_data = webforum[,5:23]
D <- cor(corr_data[, c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19)])
corrplot(D, method = "circle")
```

Analysing the relationship between the variables are crucial as it will give us an insight on which variables are dependant to another variable. By looking at the correlation matrix we will now be able to tell if other linguistic variables are dependant to variables with high mean values such as WC,Analytic,Clout,Authentic and Tone.



Question: Is there a relationship between linguistic variables over the longer term?

Answer: The variable Analytic seems to have a relationship with most of the linguistic variables compared to the rest which shows that Analytical thinking is used in most of the Threads in the webforum. On the other hand, the word count of the text of the post (WC) shows very little to no relationship with other variables which means our initial assumption that other variables are dependent on WC is incorrect. Moreover, the variable Authentic has a high relationship with the variable "i" which indicates that when participant themselves the tone of the voice is authentic.

To be continued, the variable ppron shows a high relationship with pronouns (i,we,you,shehe,they). However, it has a positively high relationship with the pronoun "i" as both of the variables used the same word "I" to describe themselves.

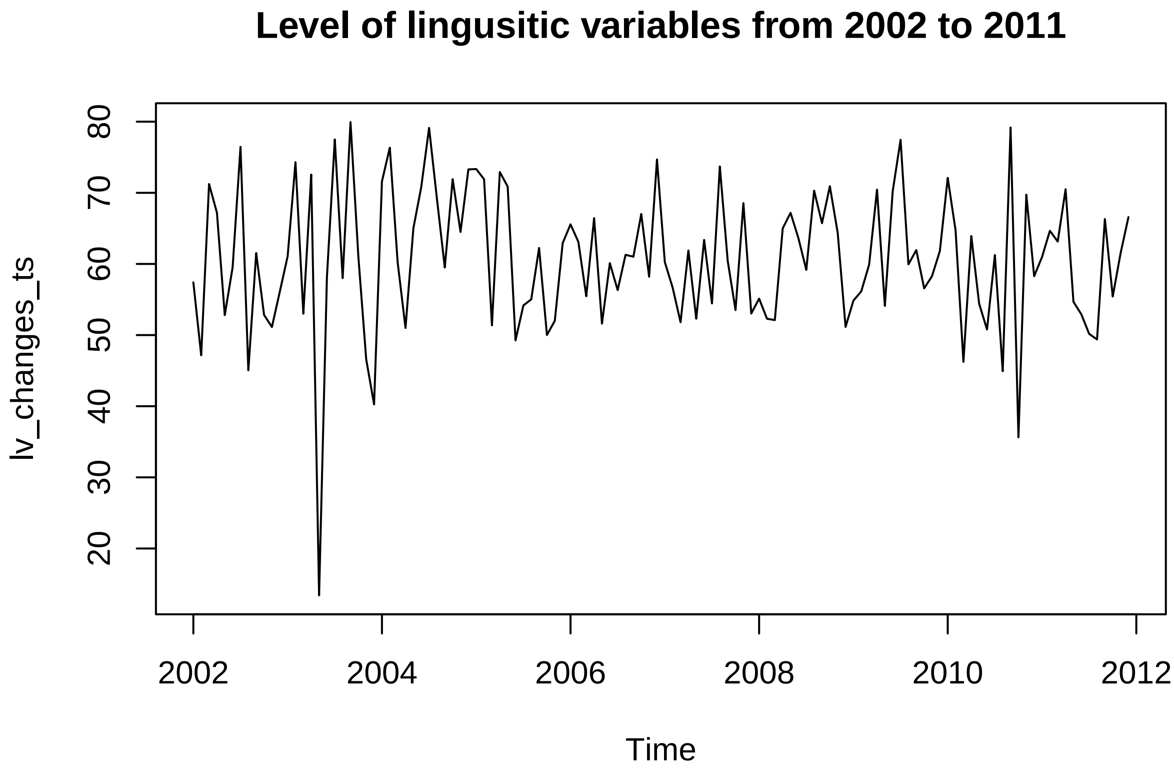
Lastly, the variable Tone has a relationship with certain variables such as (posemo,negemo,anx,anger,sad) which are used to express their emotions. Furthermore, variable Tone has a high positive relationship with variable posemo which indicates that the majority of emotional tones used in the forum are positive emotions. Moreover, variable negemo has a high positive relationship with variable anger which represents that the Thread which expresses negative emotions feels angry the most.

```
#Group by with each month and year and calculate the mean of chosen linguistic variables based on corre
lv_changes = as.data.frame(aggregate(webforum[,c(6,7,8,9,10,11,16,17,19)], list(webforum$month_year), m
names(lv_changes)[1] <- "Year_Month"
```

```
#reshape the data to make it more organize
lv_changes_new = melt(lv_changes, id=c("Year_Month"))
lv_changes_new$value = format(round(lv_changes_new$value, 2))
colnames(lv_changes_new) = c("Year_Month", "Linguistic_Varaibles", "Values")
```

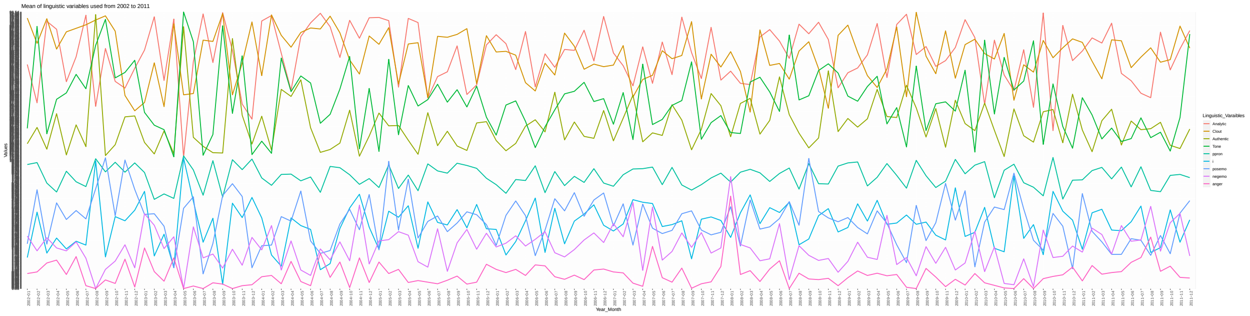
From the analysis above, the variables Analytic, Clout, Tone, Authentic, posemo, negemo,anger, ppron and i will be used to further investigate. Time-series to check the level of linguistic varaibles from 2002 to 2011.

```
lv_changes_ts = ts(lv_changes_new[,3], start = c(2002,01), end = c(2011,12), frequency = 12)
plot(lv_changes_ts, main = "Level of lingusitic variables from 2002 to 2011")
```



```
lv_plot = ggplot(data = lv_changes_new, aes(x= Year_Month, y= Values, group = Linguistic_Varaibles , co
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) + ggtitle("Mean of linguistic variables used
lv_plot
```

From the time series plot above we can see level of lingusitic variables have been fluctuating over the period of 9 years. Around year 2003 there is a sharp decrease in the level. The same significant have been spotted in year 2004 and around year 2011. We can say that overall the level of linguistic variables are not constant. Every year there are spike decreases and increases in the level used. Let us now further investigate for each linguistic variables



Question: Looking at the linguistic variables, do the levels of these change over the duration of the forum?

Answer: Variables Authentic, Tone, Clout and Analytic have higher usage compared to other variables. Analytic and Clout fluctuate closely over time while Authentic and Tone show the same behavior as well. Around 2003, Analytic and Clout seem to have a spike decreased while Authentic and Tone show an opposite trend.

```
#do t-test to prove
webforum$Year = year(webforum$Date)

wf2002 = webforum[webforum$Year == 2002,]
wf2011 = webforum[webforum$Year == 2011,]
```

```
t.test(wf2002$Analytic, wf2011$Analytic, conf.level = 0.95)
```

Lastly, over the period of 9 years variables negemo and anger has spike increase and decreased together. For instance, a significant increase is noticeable in 2007 May and 2008 Jan. The same pattern has been observed with variables posemo and i. To see the changes in linguistic variable we will perform t-test.

```
##
##  Welch Two Sample t-test
##
## data:  wf2002$Analytic and wf2011$Analytic
## t = -0.36931, df = 134.02, p-value = 0.7125
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -8.765844  6.007296
## sample estimates:
## mean of x mean of y
##  58.24338  59.62265
```

```
t.test(wf2002$Clout, wf2011$Clout, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  wf2002$Clout and wf2011$Clout
## t = 1.0364, df = 126.4, p-value = 0.302
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.367278 10.772891
## sample estimates:
## mean of x mean of y
##  64.34873  60.64593
```

```
t.test(wf2002$ppron, wf2011$ppron, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  wf2002$ppron and wf2011$ppron
## t = 0.80102, df = 133.51, p-value = 0.4245
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8075453  1.9068341
## sample estimates:
## mean of x mean of y
##  8.186620  7.636975
```

```
t.test(wf2002$anger, wf2011$anger, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  wf2002$anger and wf2011$anger
## t = -0.71399, df = 189.51, p-value = 0.4761
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6323302  0.2962269
## sample estimates:
## mean of x mean of y
## 0.9402817 1.1083333
```

H0: mu(2002) != mu(2011)

HA: mu(2002) = mu(2011)

We have performed a t-test with 95% confidence interval to support the analysis we did on the visualization. 4 variables Authentic, Clout,ppron, and anger were chosen. Is it because Authentic and Clout show the most fluctuation over the span of 9 years. Whereas, ppron shows a more consistent and stable pattern on the graph compared to other linguistic variables. Lastly, anger seems to be the linguistic variable with the lowest mean value where it has its own characteristic at a certain period of time.

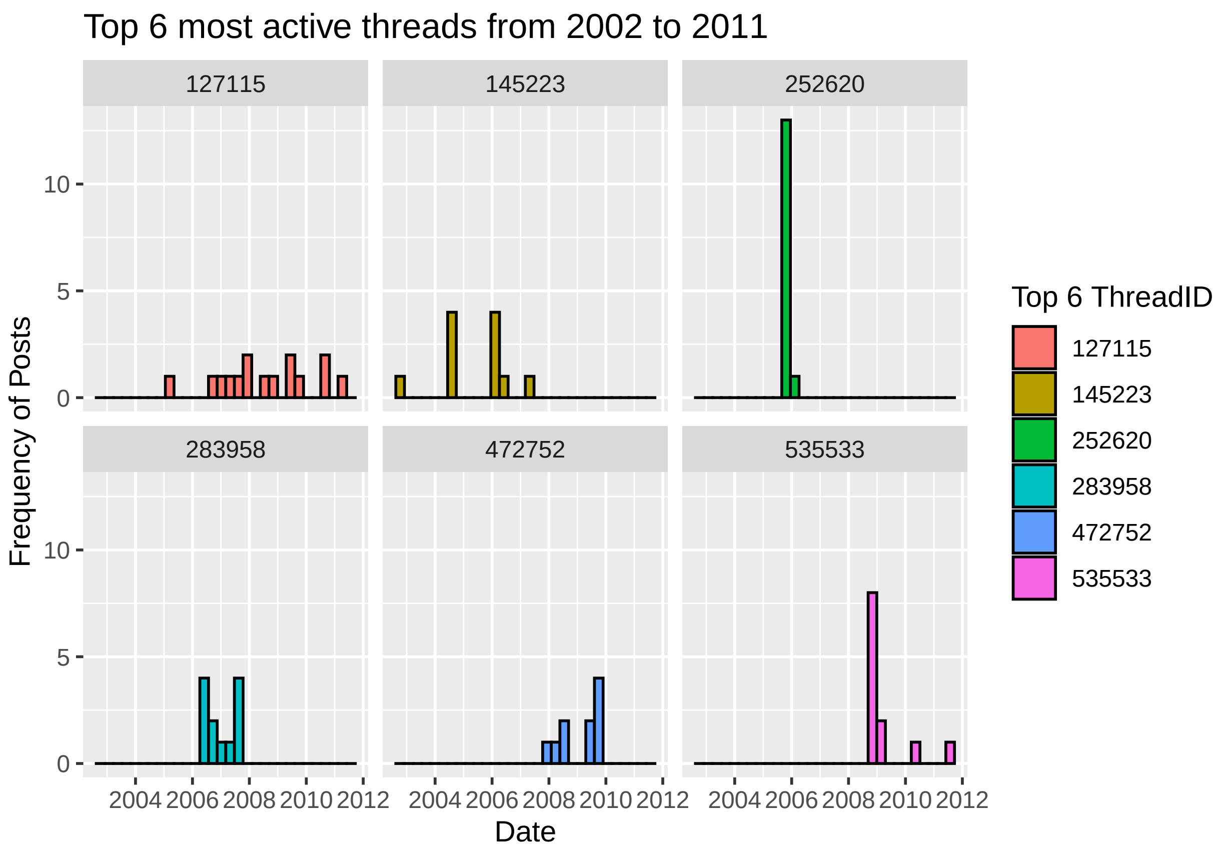
Based on our t-test we can see that p-values of all 4 of the linguistic variable are greater than 0.05 which states that we have insufficient evidence to reject the null hypothesis. Therefore, we can say that linguistic variables changes over time.

1b)

```
#take top 6 most active threads
top_threads = as.table(by(webforum$Date, webforum$ThreadID, length))
top_threads = sort(top_threads, decreasing = TRUE)

top6 = as.data.frame(head(top_threads,6))
colnames(top6) = c("ThreadID", "Freq")

top6 = webforum[(webforum$ThreadID %in% top6$ThreadID),]
```

```
#histogram for top6 thread
class(top6$ThreadID) = "String"
options(scipen=10000)
top6_plot = ggplot(data = top6) +
  geom_histogram(aes(x=as.Date(Date), fill=sprintf("%.0f",ThreadID)), color = "black") +
  ggtitle("Top 6 most active threads from 2002 to 2011") +
  xlab("Date") +
  ylab("Frequency of Posts") +
  scale_fill_discrete(name = "Top 6 ThreadID")+
  facet_wrap(~ sprintf("%.0f",ThreadID), nrow = 2)

top6_plot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Throughout the course of 9 years the number of Threads have increased. To further investigate on the language used by the threads I have chosen the top 6 most active threads from 2002 to 2011. ThreadID with ID's 127115,472752,283958,252620,535533, 145223 are the most active Threads among all. I have plotted the histogram to check the distribution over the 9 years. It can be seen that most of the threads start to be active from 2004 since the web become more popular from 2004 onwards. Now lets analyse the languages used by the top 6 Threads based on the correlation matrix relationship

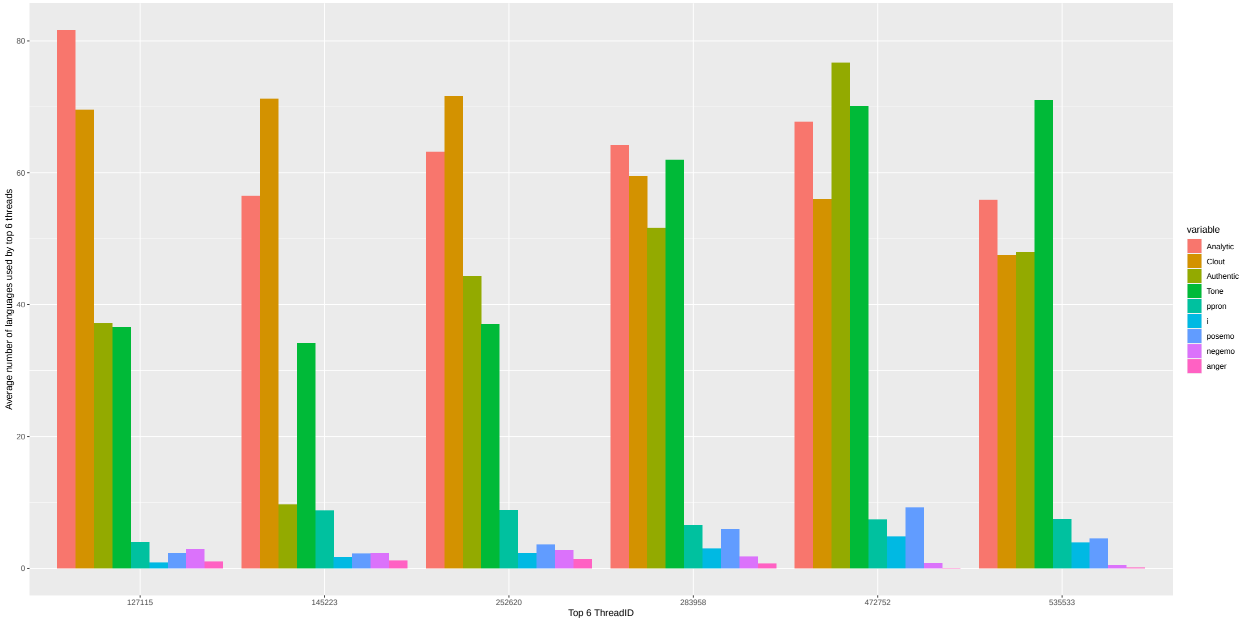
```
#Group by with ThreadID and calculate the mean of linguistic variable used by each thread
top6$ThreadID <- as.integer(top6$ThreadID)
top6_languages = as.data.frame(aggregate(top6[,5:23], by=list(top6$ThreadID), FUN = mean))
names(top6_languages)[1] <- "ThreadID"
```

Since there are variables that are dependent on some we will use those to further analyse

```
top6_language = top6_languages[,c(1,3,4,5,6,7,8,13,14,16)]
top6_language = melt(top6_language, id="ThreadID")
```

Bar chart for languages used by each of the top 6 threads

```
options(scipen=10000)
top6_lang = ggplot(top6_language, aes(x=sprintf("%.0f",ThreadID), y=value, fill=variable)) + geom_bar(p
top6_lang
```



Question: Using the relevant linguistic variables, is it possible to see whether or not particular threads are happier or more optimistic than other threads, or the forum in general, at different periods in time.

Answer: I have used variables that are dependent on each other based on the correlation matrix from part A. By analyzing the visualization above we can see that the top 6 threads have a different proportion of linguistic variables Analytic, Clout, Authentic, and Tone used. Overall, we can see that the linguistic variable posemo referring to positive emotions is used more compared to negemo referring to negative emotion which indicates that the top 6 threads are more positive rather than negative. Moreover, apart from Analytic, Clout, Authentic, and Tone, we can see that the linguistic variable ppron referring to personal pronouns is used more compared to other variables which means in each threads participant either write as first person, second person, or third person.

To further investigate the plot, thread number 127115 has the highest number of Analytical which means participants in that thread used more analytical thinking than the other threads. However, thread number 127115 also happens to have a higher proportion of negative emotions compared to positive emotions. That particular thread has the lowest amount of positive emotion compared to other threads.

On the other hand, thread number 472752 has the highest proportion of Authenitc compared to other threads which indicate in that thread the authentic tone of voice is being used the most. Surprisingly, that particular thread also seems to have the highest amount of positive emotions compared to other threads and the proportion of negative emotion and anger is lesser compared to other threads. To conclude, we can see that threads that used the linguistic variable Authentic referring to the authentic tone of voice more than the average tends to be more positive rather than negative.

T-test will be conducted to see if the above statment is correct on data of posemo and negemo for ThreadID 127115 and 472752

posemo data : H0 = mu(472752) >= mu(127115)

```
#extract data
d_127 = webforum[webforum$ThreadID == 127115,]
d_472 = webforum[webforum$ThreadID == 472752,]
```

```
#t-test for negemo for ThreadID 127115 and 472752
t.test(d_127$negemo,d_472$negemo, "less", conf.level = 0.95)
```

negemo data: H0 = mu(127115) >= mu(472752)

```
##
##  Welch Two Sample t-test
##
## data:  d_127$negemo and d_472$negemo
## t = 2.1914, df = 19.278, p-value = 0.9795
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 3.770725
## sample estimates:
## mean of x mean of y
##  2.931333  0.823000
```

```
#t-test for posemo for ThreadID 127115 and 472752
t.test(d_127$posemo,d_472$posemo, "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  d_127$posemo and d_472$posemo
## t = -1.4749, df = 9.2016, p-value = 0.9132
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -15.55692       Inf
## sample estimates:
## mean of x mean of y
##  2.297333  9.243000
```

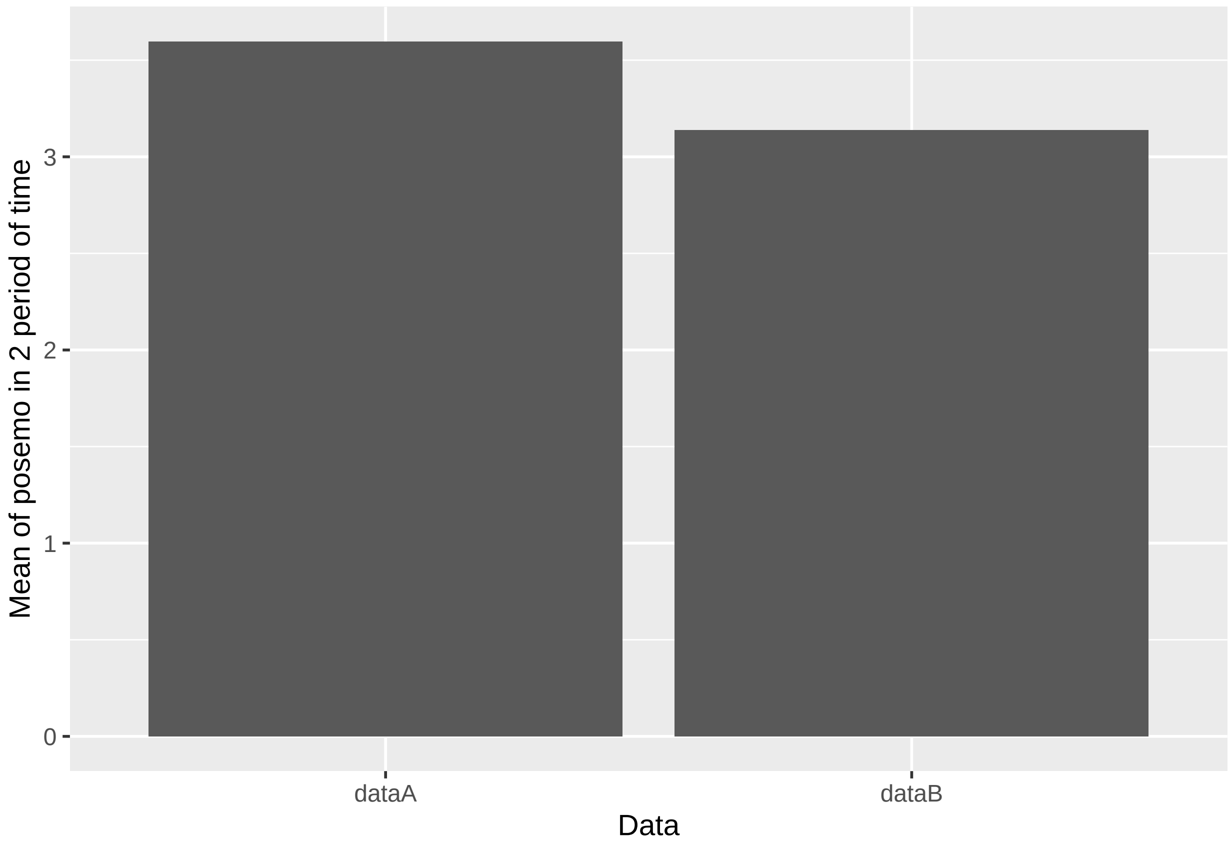
Since, p-values we obtained are larger than 0.05, we have insufficient evidence to reject the null hypothesis. Thus, we can conclude that ThreadID 472752 has highest poesmo proportion and ThreadID 127115 has highest negemo proportion.

```
#filter thread with year more than 2004 and less than 2009
dataA = filter(webforum, Date >= "2004-01-01" & Date < "2009-01-01")
```

```
#filter thread with year more than 2009
dataB = filter(webforum, Date >= "2009-01-01")
```

Is there specific period of time where the thread is more optimistic? I have chosen from 2004 on wards because based on the multi-variable bar chart for frequency of posts, the threads starts to be more active from 2004 on wards.

```
agg_dataA <- mean(dataA$posemo)
agg_dataB <- mean(dataB$posemo)
new_df = data.frame(data = c("dataA", "dataB"), avg_posemo = c(agg_dataA, agg_dataB))
ggplot(new_df, aes(x=data, y=avg_posemo)) +geom_bar(position = "dodge", stat = "identity")+xlab("Data")
```



We can see that average number of poesemo variables have changes from dataA to dataB in 2 period of time. Now lets do a t-test to prove the change.

H0 = mu(dataA) != mu(dataB)

```
t.test(dataA$posemo, dataB$posemo,conf.level = 0.95)
```

HA = mu(dataA) = mu(dataB)

```
##
##  Welch Two Sample t-test
##
## data:  dataA$posemo and dataB$posemo
## t = 1.7258, df = 865.78, p-value = 0.08474
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06289896  0.97927223
## sample estimates:
## mean of x mean of y
##  3.596421  3.138234
```

We have performed a t-test for the linguistic variable posemo referring to expressing positive emotion with a 95% confidence interval from 2004 to 2008 for the first data (dataA) and from 2009 onwards for the second data(dataB). The p-value we obtained is 0.08474 which is greater than 0.05 and indicates that we have insufficient evidence to reject the null hypothesis. Thus, we can say that at different points in time there are threads that are more optimistic than other threads.

1c)

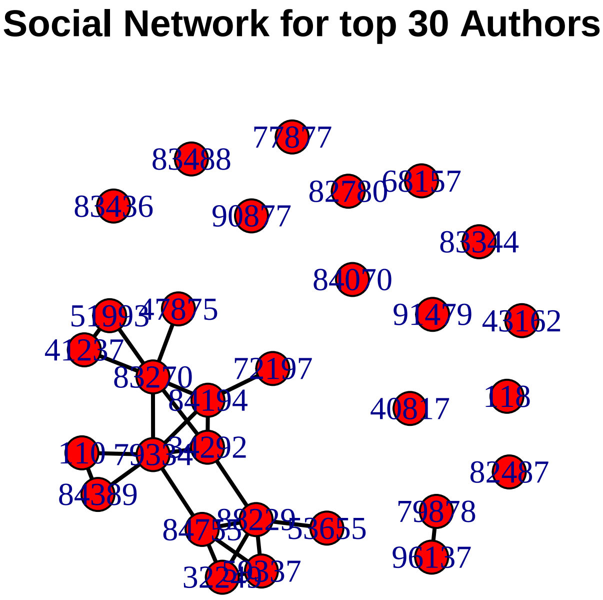
```
by_auth = as.data.frame(as.table(tapply(webforum$AuthorID, list(webforum$month_year), length)))
colnames(by_auth) = c("Year_Month", "Freq")
```

```
# 2006 is chosen because it has the maximum amount of authors
most_month = filter(webforum, Date >= "2006-03-01" & Date < "2006-08-01")
#choose author posted on the thread more than 2 times
topauth = as.table(by(most_month$Date, most_month$AuthorID, length))
topauth = sort(topauth, decreasing = TRUE)
top9_author = as.data.frame(head(topauth,30))
names(top9_author)[1] = "AuthorID"

#extract ThreadID and AuthorID of top9 authors
top9_author_new = most_month[(most_month$AuthorID%in%top9_author$AuthorID),]
top_9_author = top9_author_new[,1:2]
```

```
options(repos="https://cran.rstudio.com")
install.packages("pscl", repos = "https://cran.rstudio.com")
install.packages(c("igraph","igraphdata"))
library(igraph)
library(igraphdata)
```

```
#following code was taken from Lecture 5,slide 78
g <- make_empty_graph(directed = FALSE)

for (i in 1:nrow(top9_author)){
  g <- add_vertices(g, 1, name = as.character(top9_author$AuthorID[i]))
}

for(k in unique(top_9_author$ThreadID)){
  temp = top_9_author[(top_9_author$ThreadID == k),]
  if(nrow(temp) > 1){
    Edgelist = as.data.frame(t(combn(temp$AuthorID,2)))
    colnames(Edgelist) = c("P1", "P2")

    for(i in 1:nrow(Edgelist)){
    g <- add_edges(g, c( as.character(Edgelist$P1[i]), as.character(Edgelist$P2[i])))
  }
  }




}
```

```
g <- simplify(g)
plot(g, main = "Social Network for top 30 Authors", vertex.color = "red", edge.width = 2, edge.color =
```



```
#checking number of vertices and edges
vcount(g) #30
```

```
## [1] 30
```

```
ecount(g) #24
```

```
## [1] 24
```

```
#check simple or not
is.simple(g) #TRUE bcoz we simplify it
```

```
## [1] TRUE
```

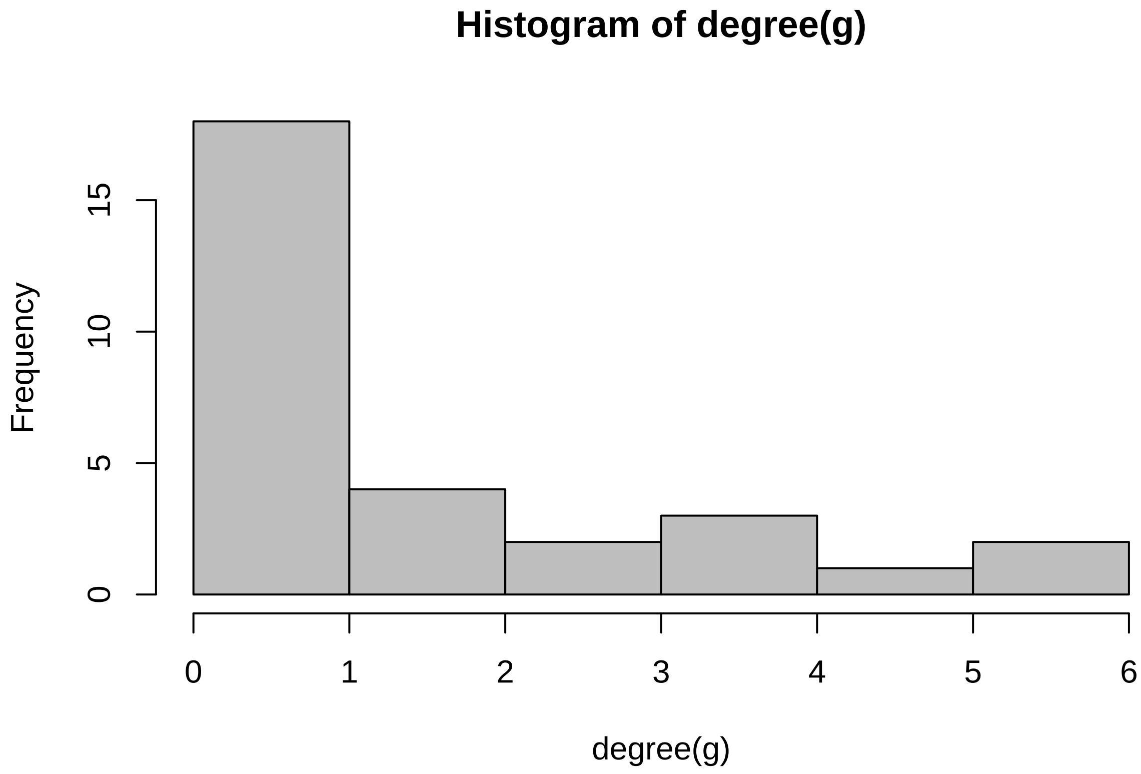
```
diameter(g) #4
```

```
## [1] 4
```

```
average.path.length(g) #2.339623
```

```
## [1] 2.339623
```

```
hist(degree(g), breaks = 5, col = "grey")
```



By doing some analysis we can see that year 2006 has the most amount of authors posting on threads compared to other years. Therefore, the date range from March 2006 to July 2006 was chosen as the period of time to analyse the social network for 30 authors who have posted 2 or more times to multiple threads. I have chosen 5 months period of time instead of 1 month because the thread may continue for months.Looking at the graph above it has 30 vertices representing each author and 24 edges. The social network has a diameter of 4 which means the longest distance between any 2 vertices is 4. Moreover, the graph has an average path length of 2.34 (2 d.p) which means average distance between any two vertices is 2.34. Furthermore, the histogram for the degree distribution was separated into 5, by observing the plot we can see that the histogram is right-skewed which shows that most vertices have few edges while some have no edges. The graph is a connected graph, while there are other vertices not connected and separating from the main group.

2c)

```
#following code was taken from tutorial 5
describe_vertices <- function(net, order_col) {
      if (missing(order_col)) {
        order_col = 1
      }
      degree = as.table(degree(net))
      betweenness = as.table(betweenness(net))
      closeness = as.table(closeness(net))
      eigenvector = as.table(evcent(net)$vector)
      vertex_tab = as.data.frame(rbind(degree, betweenness, closeness, eigenvector))
      vertex_tab = t(vertex_tab)
      vertex_tab = round(vertex_tab, 4)
      cat("VERTEX CHARACTERISTICS: ", "\n")
      return(vertex_tab[order(-vertex_tab[,order_col]), ])
}

vertex_char_table = as.data.frame(describe_vertices(g))
```

```
## VERTEX CHARACTERISTICS:
```

```
#Calculating graph clique
table(sapply(cliques(g), length))
```

```
##
##  1  2  3  4
## 30 24 10  2
```

```
#Finding which vertices have largest clique
cliques(g)[sapply(cliques(g), length) == 4]
```

```
## [[1]]
## + 4/30 vertices, named, from 0f0f662:
## [1] 88229 32249 59337 84755
##
## [[2]]
## + 4/30 vertices, named, from 0f0f662:
## [1] 79334 34292 84194 83270
```

Question: Identify the most important author in the social network you created

Answer: By observing the calculation above we can say that author number 79334 is the most important author in the social network. Author number 79334 rank first in measure of Degree, Betweeness and Eigenvector. However, author number 79334 ranked the second highest for closeness centrality measure while author number 96137 and 79878 ranked the most highest in closeness centrality measure. Furthermore, author number 83270 has same number of degree mesure as author 79334 but author number 79334 outweigh in betweenness, closeness and eigenvector .As a result of that, i have investiaged the authors invloved in the largest clique in the social network and it was seen that author number 79334 is invloved in the largest clique which gives a strong supporting evidence that author number 79334 is the most important author. Select authors with degree more than or equal 3 and degree less than 3.

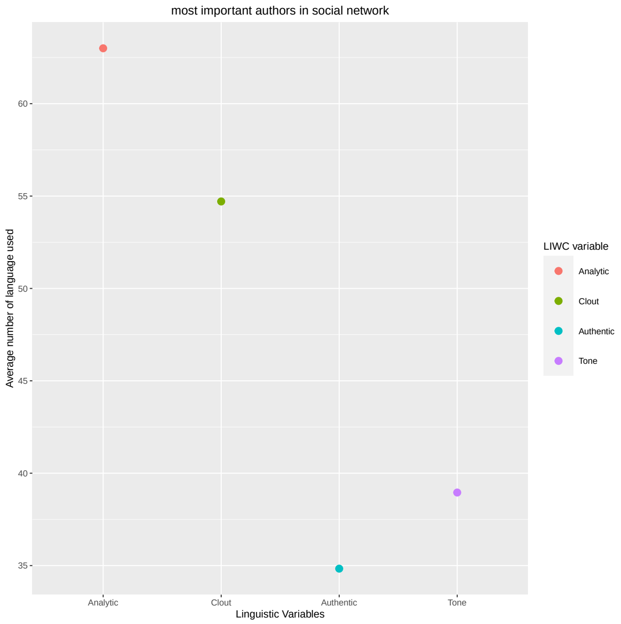
```
top = as.numeric(unlist(rownames(vertex_char_table[vertex_char_table$degree >= 3,])))

bottom = as.numeric(unlist(rownames(vertex_char_table[vertex_char_table$degree < 3,])))

top_df =  webforum[webforum$AuthorID %in% top,]
bottom_df = webforum[webforum$AuthorID %in% bottom,]
```

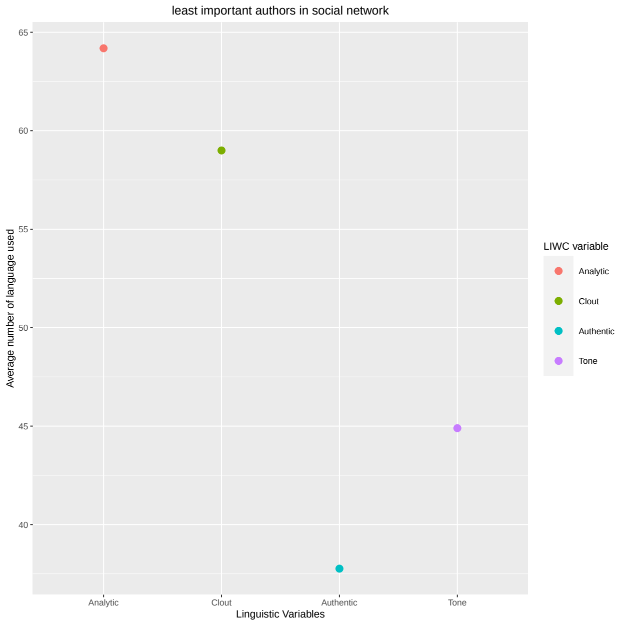
Calculate the mean of Analytic, Clout, Authentic and Tone. These variables are chosen because they have the highest mean values compared to other linguistic variables

```
top_social_auth = top_df %>%
              dplyr:: summarise(
                Analytic = mean(Analytic),
                Clout = mean(Clout),
                Authentic = mean(Authentic),
                Tone = mean(Tone)
              )

bottom_social_auth = bottom_df %>%
              dplyr:: summarise(
                Analytic = mean(Analytic),
                Clout = mean(Clout),
                Authentic = mean(Authentic),
                Tone = mean(Tone)
              )
```

```
top_social_auth <- top_social_auth %>% mutate(id = row_number())
topsocial_plot = melt(top_social_auth, id="id")
bottom_social_auth <- bottom_social_auth %>% mutate(id = row_number())
bottomsocial_plot = melt(bottom_social_auth, id="id")
```

```
p1 = ggplot(topsocial_plot, aes(x=variable, y = value, color=factor(variable))) + geom_point() + labs(ti
 theme(plot.title=element_text(hjust = 0.5))

p2 = ggplot(bottomsocial_plot, aes(x=variable, y = value, color=factor(variable))) + geom_point() + labs
 theme(plot.title=element_text(hjust = 0.5))

plot_grid(p1, p2)
```





Question: Looking at the language they use, can you observe any difference between them and other members of their social network

Answer: Among the 30 authors in the social network i have filter them based on the degree. I grouped authors having degree more than or equal to 3 as an important author. On the other hand, authors having less than 3 degree will be grouped as less important authors.

By observing the languages used by the most important authors and least importants author there seems to be a difference in language used by those 2 author groups from 2002 to 2011. We can see that average value of languages used for Analytic, Clout, Authentic and Tone is higher for least important author from our social network graph. Thus, we can conclude by saying that least important authors shows more analytical thinking, power , authentic tone of voice and emotional tone in their respective threads.To prove that there is a difference in the languages used we will perform a t-test.

Tone data: H0 = mu(least important author in social network) >= mu(most important author in social network)

```
t.test(bottom_df$Tone, top_df$Tone, "less", conf.level = 0.95)
```

Clout data : H0 = mu(most important author in social network) >= mu(least important author in social network)

```
##
##  Welch Two Sample t-test
##
## data:  bottom_df$Tone and top_df$Tone
## t = 1.5163, df = 233.54, p-value = 0.9346
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 12.41244
## sample estimates:
## mean of x mean of y
##  44.89576  38.95433
```

```
t.test(bottom_df$Clout, top_df$Clout, "less", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  bottom_df$Clout and top_df$Clout
## t = 1.3146, df = 212.87, p-value = 0.905
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf 9.68174
## sample estimates:
## mean of x mean of y
##  58.99794  54.70767
```

Since the p-values we obtained are >= 0.05 we have insuffficent evidence to reject the null hypothesis. Therefore, on average least important author used higher lingusitic variables comapred to most important author in our social network.