FIT 3152 Assignment 1

Xinyu Ma

28652703

Introduction:

The data set for this analysis is based on the user's post content, thread ID, author ID and posting time of a real online community. The data set contains a total of 29 variables, thread ID, author ID, date and time are character variables used for posting identification, WC is the number of words in a single post, and analytic, clout, authentic and tone are numeric variables that indicate the proportion of the feature words and sentences used in this post. The analysis tools used are R software and ggplot2, corrplot and dplyr and other R packages.

Analysis preparation:

First read the csv file into R and randomly sample 20,000 samples from it as the analysis object. In order to prevent differences in each sampling, we must first set the seeds in the system. Before the formal analysis, let's take a brief look at the situation of the 20,000 sample data extracted:

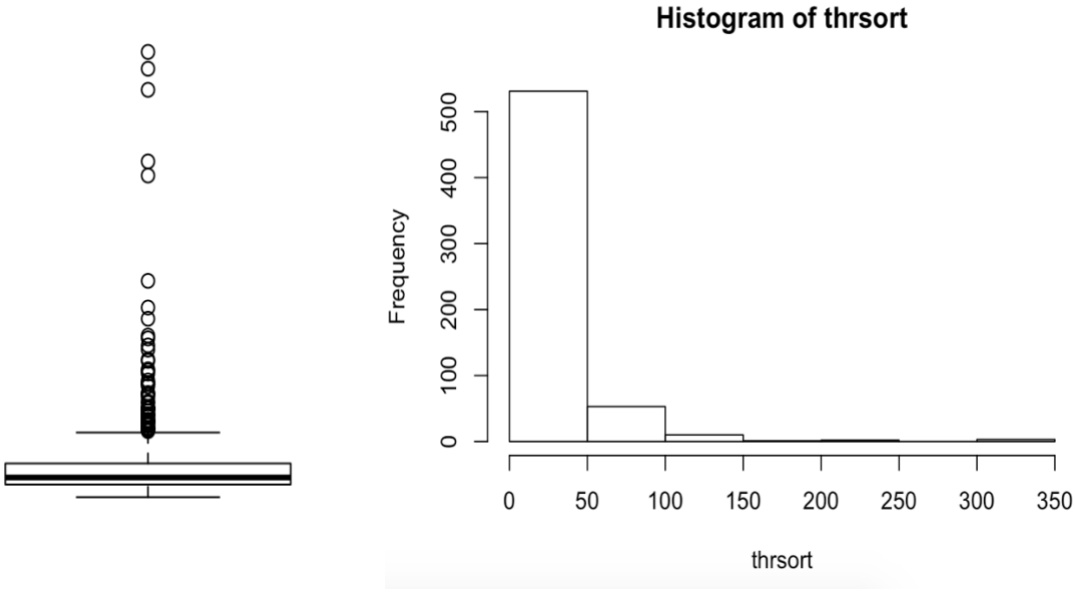
```
> webforum[which.min(as.Date(webforum$Date,"%Y-%m-%d")),]
     ThreadID AuthorID       Date  Time  WC Analytic Clout Authentic Tone ppron    i   we you shehe they number
5533    10133     1740 2002-01-16 23:56 161    26.74 77.19     15.49 2.94  9.94 1.86 1.86   0  0.62 5.59      0
     affect posemo negemo anx anger social family friend leisure money relig swear QMark
5533   2.48      0   2.48   0  0.62  12.42      0      0       0     0  1.86     0  0.62
> webforum[which.max(as.Date(webforum$Date,"%Y-%m-%d")),]  # find the earliest and latest dates recorded
     ThreadID AuthorID       Date  Time WC Analytic Clout Authentic  Tone ppron    i   we  you shehe they number
4476   853260   166362 2011-12-31 06:43 84    75.78 97.14     30.01 70.57  9.52 2.38 4.76 2.38     0    0      0
     affect posemo negemo anx anger social family friend leisure money relig swear QMark
4476   2.38   2.38      0   0     0  15.48      0      0       0     0     0     0     0
> summary(webforum)
    ThreadID         AuthorID             Date           Time             WC            Analytic
 Min.   : 10133   Min.   :    -1   2005-12-11:  132   04:52  :   39   Min.   :   0.0   Min.   : 0.00
 1st Qu.:245702   1st Qu.: 39864   2005-12-12:  121   07:10  :   38   1st Qu.:  28.0   1st Qu.:39.15
 Median :330904   Median : 79334   2005-12-15:  101   05:09  :   36   Median :  63.0   Median :63.30
 Mean   :375698   Mean   : 83354   2005-12-16:   88   06:50  :   36   Mean   : 103.1   Mean   :59.73
 3rd Qu.:486195   3rd Qu.:117404   2005-12-13:   76   04:58  :   35   3rd Qu.: 128.0   3rd Qu.:83.44
 Max.   :877240   Max.   :252309   2005-12-18:   64   07:16  :   35   Max.   :5298.0   Max.   :99.00
                                   (Other)   :19418   (Other):19781
     Clout          Authentic          Tone          ppron             i               we              you
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000
 1st Qu.:40.36   1st Qu.:10.08   1st Qu.:14.07   1st Qu.: 4.260   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.: 0.000
 Median :59.40   Median :30.91   Median :25.77   Median : 7.320   Median : 2.290   Median : 0.0000   Median : 0.000
 Mean   :57.80   Mean   :38.29   Mean   :44.17   Mean   : 7.647   Mean   : 3.344   Mean   : 0.8958   Mean   : 1.427
 3rd Qu.:79.13   3rd Qu.:62.53   3rd Qu.:79.41   3rd Qu.:10.502   3rd Qu.: 4.920   3rd Qu.: 1.1600   3rd Qu.: 1.960
 Max.   :99.00   Max.   :99.00   Max.   :99.00   Max.   :50.000   Max.   :50.000   Max.   :33.3300   Max.   :50.000
```

The analysis results show that the time span of the sample data is from January 2002 to December 2011. In terms of the overall data, the average length of users' posts is 103.1 characters, of which the words and sentences related to analysis thinking and those involving power, force and impact account for the most.

Formal analysis:

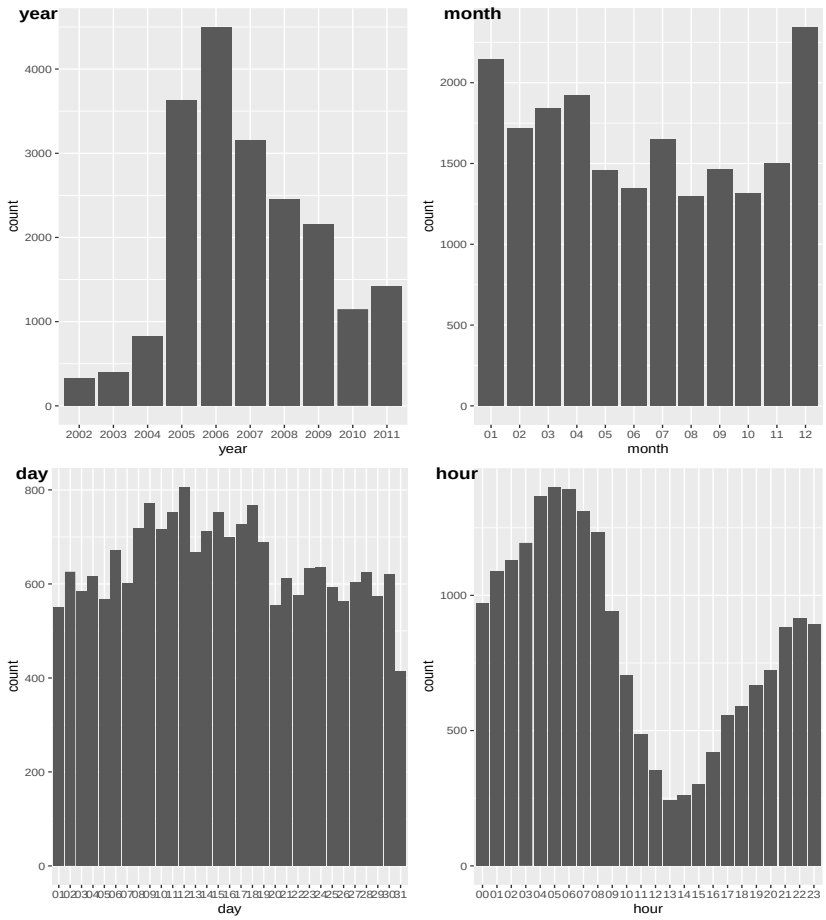
a:

In order to explore the activity of participants, we first count the total number of posts in each thread group in ten years and present them as a box plot and histogram:



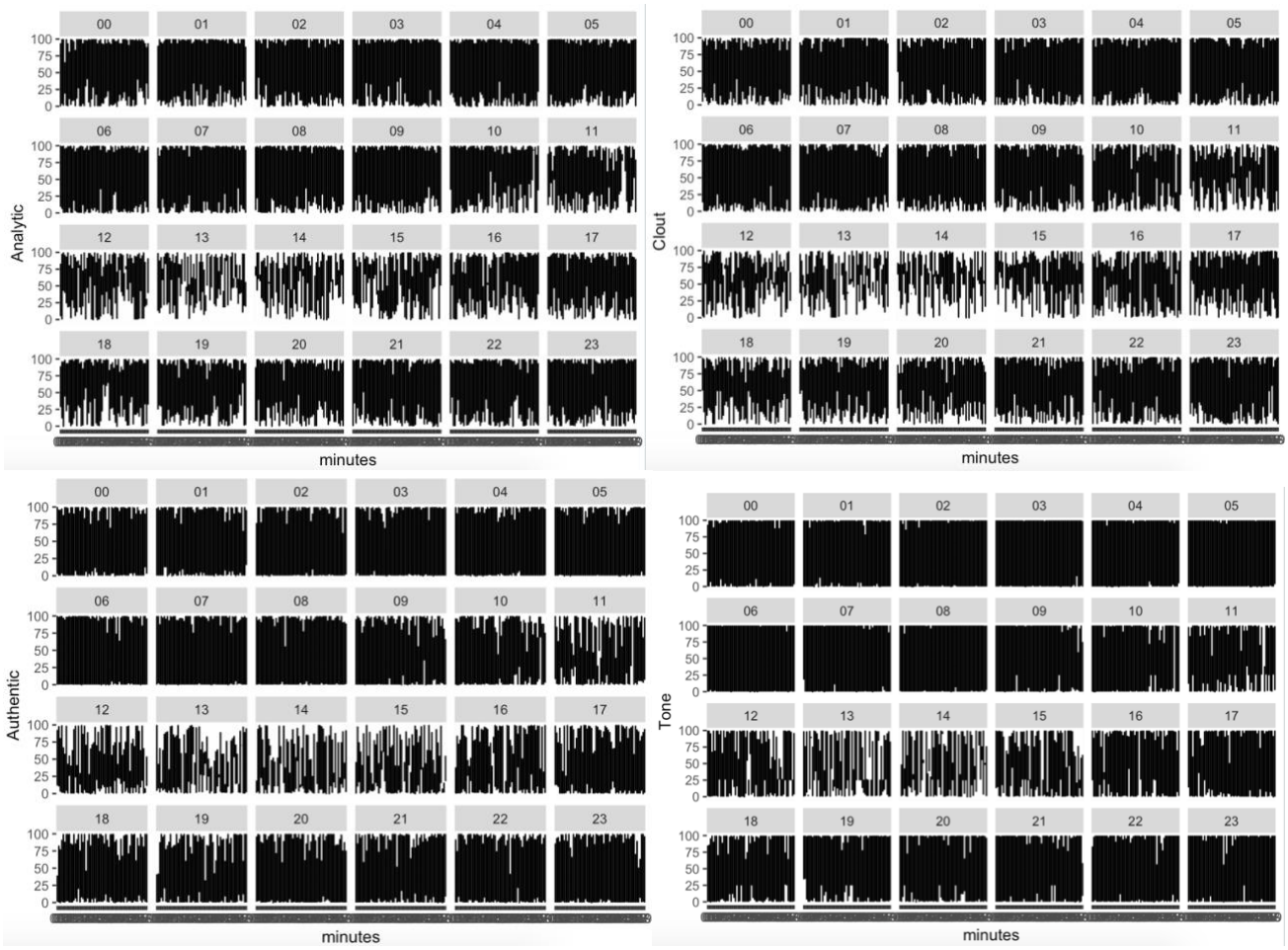
According to the box plot and histogram, the total number of posts in the majority of groups in the past ten years is concentrated around the median of 25. Only a small number of groups have extreme posts, reaching a level of more than 300, so overall, except for individual groups that are particularly active, most groups are moderately active.

Next, we will further explore the relationship between the number of posts in the forum and the time dimension such as year and month. Making a bar graph with the X axis as the time scale and y as the number of posts:

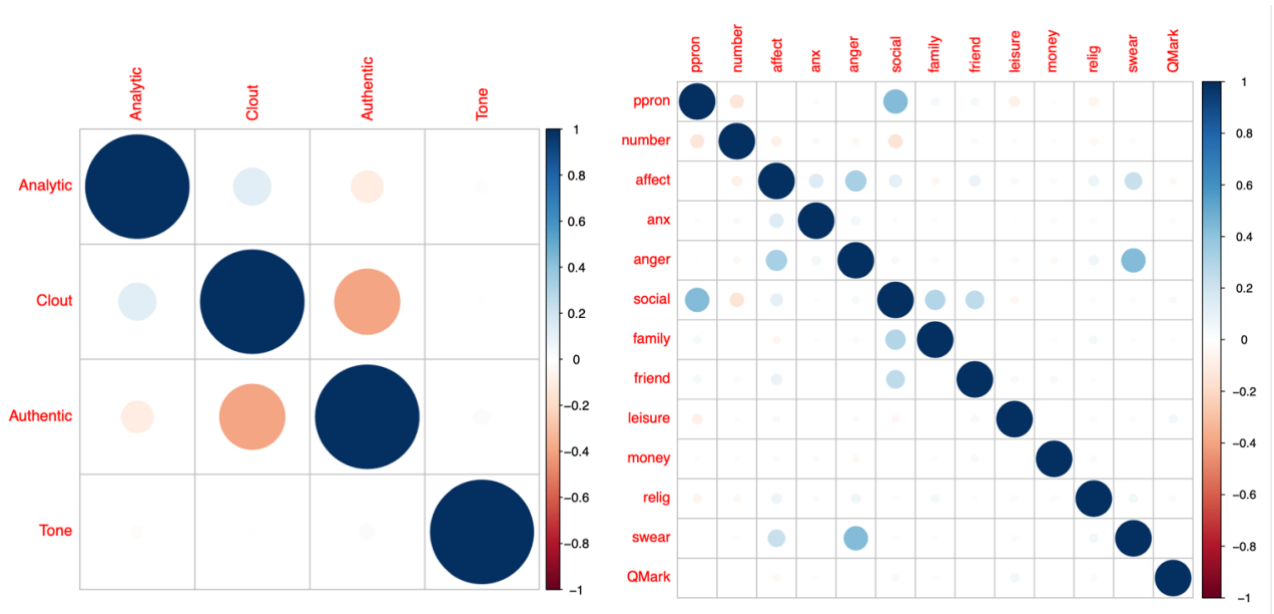


The bar graph above shows that the activity of the forum increased rapidly in 2005 and gradually declined after reaching its peak in 2006. In terms of months and dates, there is no particularly significant difference in the number of user posts, it was only a small increase at the beginning and end of the year (January and December). The most obvious change trend is the number of posts per day. It can be seen that users of this forum prefer to post from 4am to 7am, and then the activity gradually decreases until it reaches the lowest valley at noon.

Through a line chart of the changes of four linguistic variables such as analytic with time, it was found that these language variables also reached the peak of use in the early morning and then gradually fell to the lowest point of noon, and then slowly climbed until the peak of the use of the next morning.



Next, we explore whether there is a correlation between the proportions of various types of languages used, and make a correlation diagram of variables:



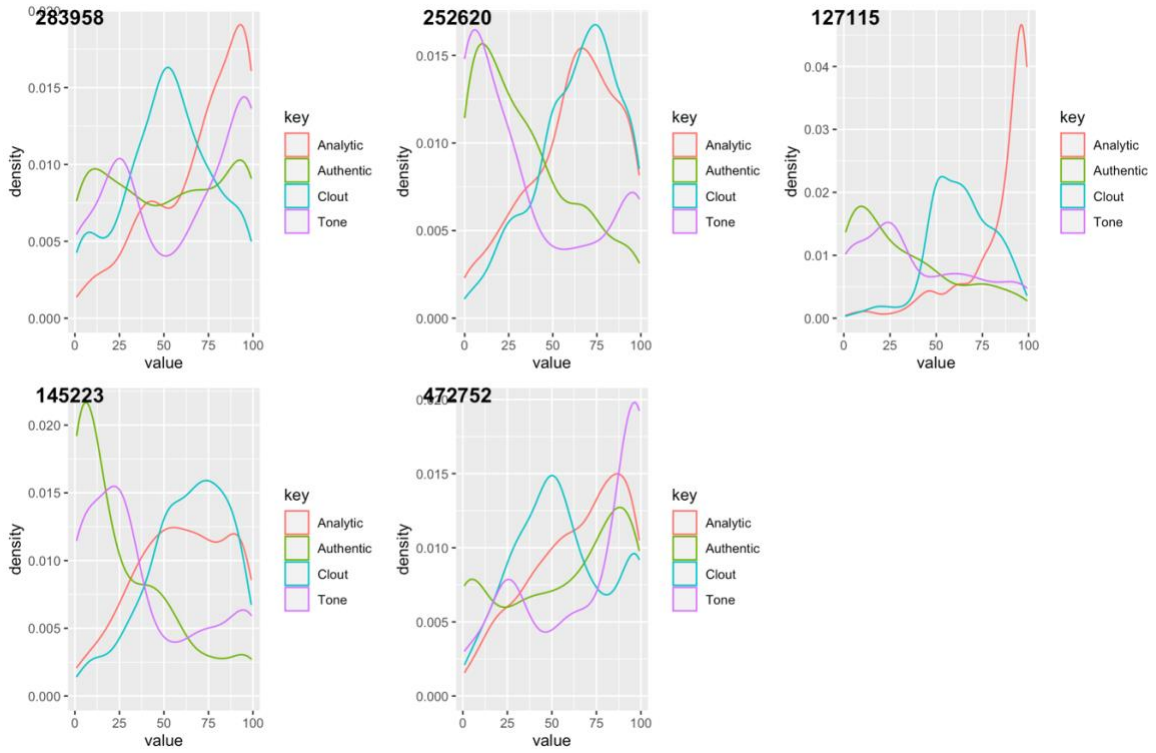
The former four variables are the variables that represent the proportion of the four language categories used (summary). The latter variables are the specific types of characters used in a post, such as personal pronouns, quantitative adverbs, etc. The darker the circle in the figure, the stronger the correlation between these variables. We can see there is an obvious negative correlation between authentic and clout, which means when people use authentic voice to express their thoughts, they will talk less about the impact of rights. Similarly, when discussing social activities, people will refer to personal pronouns more frequently. When they speak swear words, they will also have a clear tendency to be angry. In addition, there is no obvious correlation between other variables.

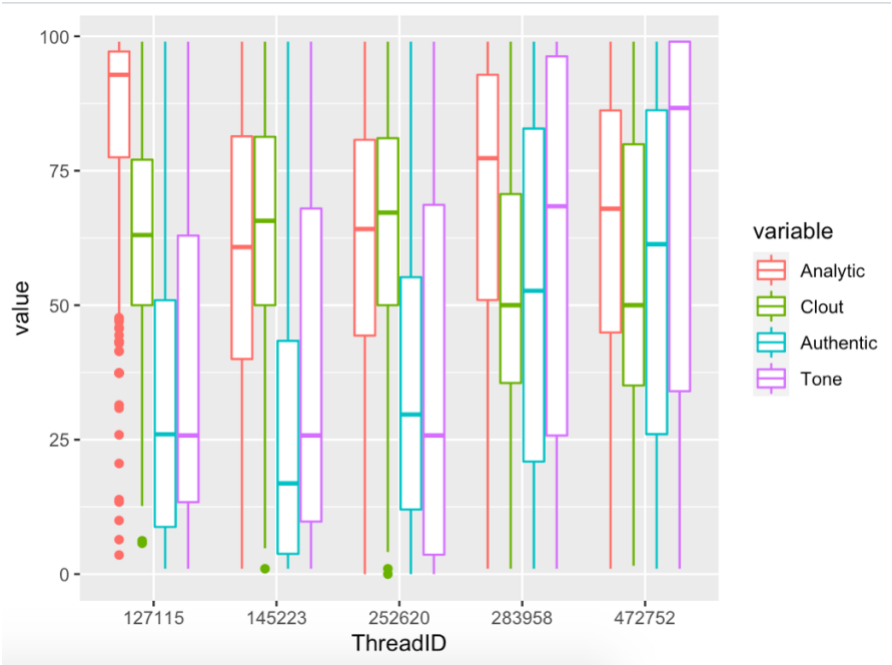
b:

In the process of exploring user activity in the previous questions, it has been found that most groups have a total number of posts in the range of 0 to 50 in ten years, and the degree of activity is average. However, the distribution of post frequency is more discrete, and there are also groups that are particularly active. The five groups with the largest number of posts in the figure are typical representatives:

```
> head(thrsort,5)
ThreadID
283958 252620 127115 145223 472752
   328    316    301    250    240
```

Because there are too many groups in the original sample and the data volume is too large, the analysis results are difficult to present with clear and beautiful graphics. Therefore, when comparing the language usage of different groups, the five most active groups in the sample are selected as representatives for analysis. We will use density maps and box plots to visually show the differences in the usage of various types of text in the five groups:

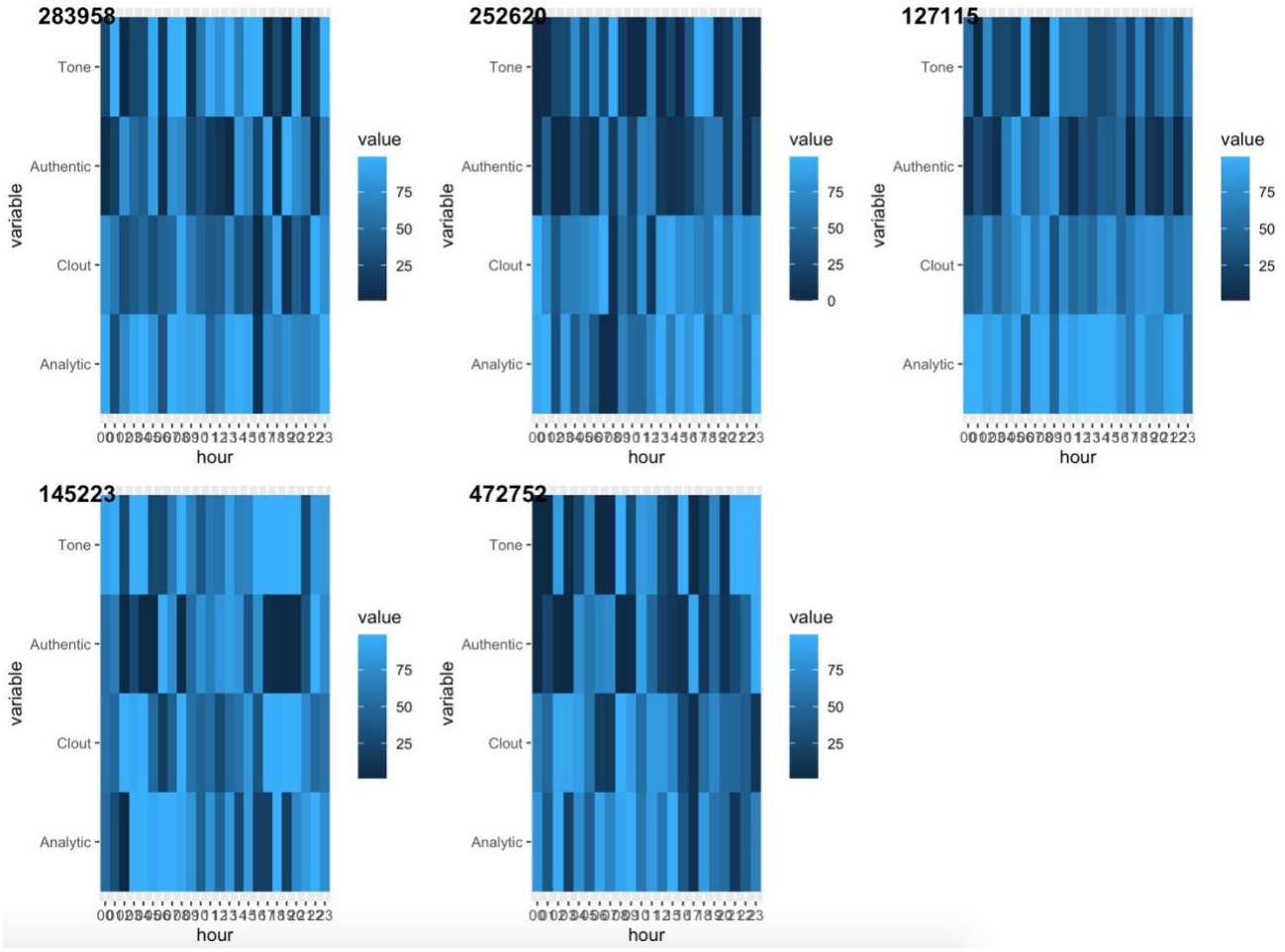




Regardless of which group, the words and sentences that represent analytical thinking are the most frequently used and the highest proportion within a single post. Correspondingly, users are less willing to post a lot of their own real voice messages in the forum, so the variable authentic has a higher density when its value is lower.

In addition, as far as the comparison of each group is concerned, although the first group is relatively balanced in the frequency and preferences of the four types of sentences, it is slightly more inclined to publish an argument with analytical thinking or emotional tone.
The second group is very distinctive, and the analytical arguments and discussion of the impact of power obviously occupy the vast majority of post content.
The third group also clearly shows that users in this group prefer to post analytical and research posts.
The fourth group of users is similar to the second group of users in the choice of language variables, but the frequency of use of preference variables in this group is not as strong as the second group.
The last group belongs to a group with a more comprehensive nature of posting content. This group has no obvious tendency for various topics, and the use of the four types of language features is not far from the frequency.

We will use heat maps of different language variables at various time points as the basis for analyzing the changes in language usage habits over time in the five groups:



First group: Except for the one hour at 4pm, users are more inclined to use analytical language, especially in the morning and afternoon. At this time, the percentage of analytical words in the post is higher than other categories of language variables. In addition, only posts from 1pm to 4pm will have more emotional tone.

Second group: This group of users mostly use analytical words and texts related to the discussion of power of rights, and the use of these two types of sentences in posts generally starting in the afternoon will gradually increase until it reaches the peak from 11pm to 1am midnight.

Third group: Users are very fond of using analytical sentences and posts that use more of these words and phrases are published at 12am to 2 am and 10am to 4pm. In addition, although there are not many posts with real voices, they will still reach a relatively small orgasm from 5am to 10am, and then gradually reduce the use.

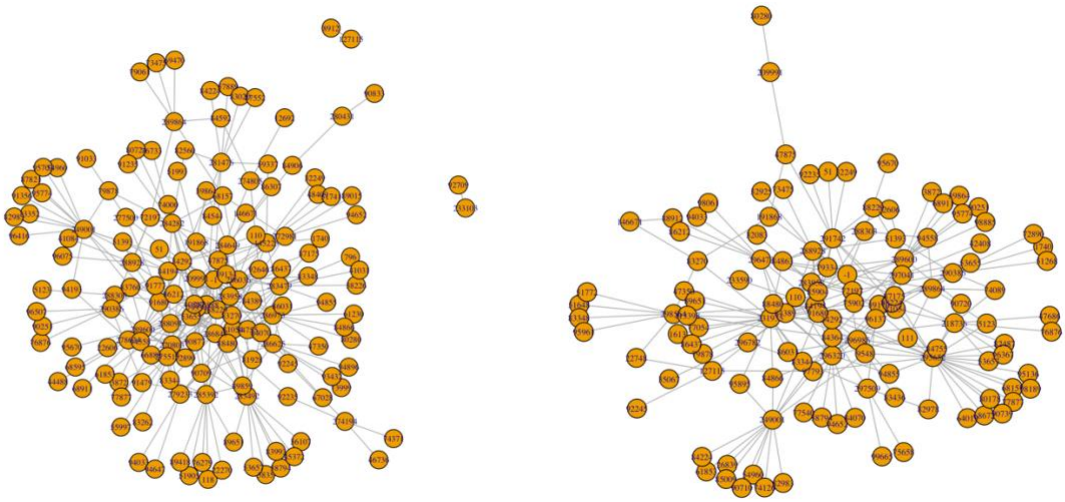
Forth group: After 2am, users will use more analytical sentences and sentences related to rights and power, and between 4pm to 8pm, they will reduce the use of analytical sentences and use more emotional-sounding words and sentences.

Fifth group: The user's preference for the four types of language variables is relatively balanced. In the early morning, they prefer to discuss topics such as influence and power. During the day, analytical comments on other things begin to increase gradually. After 9pm, the discussion and expression of emotions become the main content of the post.

c:

Since April 2006 was the month with the largest number of postings by the thread ID 283958, it was more representative, so we first selected this month as the SNA time node.

After filtering out all the samples of the current month, they are grouped and counted by group category and author ID, and then the obtained data set is converted into graphical objects in order to make a social network diagram. The social network graphs constructed by the sample data in April and May 2006 are as follows:



It can be seen that almost every circle is formed by the group as the central node, and the connection between the circles depends on the same user posting in different groups at the same time.
In April, although the forum users will also form several relatively independent communication circles, there are still certain associations between most circles, and there are more central nodes with greater influence and concentration.
The social network diagram in May showed that the number of active users who posted in the month decreased, and several important nodes were relatively scattered and the cohesion of other nodes was weakened, but the absorption of discrete subgroups was better than that of April.

Conclusion:

Through analysis, we can know that the forum reached its peak in 2006, and users were most active from 4am to 7am. Except for individual groups, most groups have average activity. In the analysis of language habits, the most active groups are used as the analysis and comparison objects. The results show that most users tend to post analysis-oriented posts, and the use of other types of languages varies according to the nature of the group. Finally, social networks of different months are established for the most active group samples. The analysis results show that for this group, the central node of the circle is mostly the discussion group, and the connection of different circles depends on the user's activity in multiple groups.

Appendix:

```
rm(list = ls())
```

```
set.seed(28652703)

########Part a
webforum<-read.csv("webforum.csv",header=T)
webforum<-webforum[sample(nrow(webforum),20000),]
attach(webforum)
webforum[which.min(as.Date(webforum$Date,"%Y-%m-%d")),]
webforum[which.max(as.Date(webforum$Date,"%Y-%m-%d")),]  # find the earliest and latest dates recorded
summary(webforum)
threadfrq<-with(webforum, table(ThreadID))
thrsort<-sort(threadfrq,decreasing = TRUE)    # count total posts of each group in descending order
thrsort
median(thrsort)    # median of posts of all groups
boxplot(thrsort)
hist(thrsort)
library(ggplot2)
ghour=ggplot(data = webforum) +geom_bar(mapping = aes(x=substring(Time,1,2)))+xlab("hour")
# count posts in every hour and plot
gmonth=ggplot(data = webforum) +geom_bar(mapping = aes(x=substring(Date,6,7)))+xlab("month")
#count posts in every month and plot
gyear=ggplot(data = webforum) +geom_bar(mapping = aes(x=substring(Date,1,4)))+xlab("year")
# count posts in every year and plot
gday=ggplot(data = webforum) +geom_bar(mapping = aes(x=substring(Date,9,10)))+xlab("day")
# count posts in everyday and plot
library(ggpubr)
ggarrange(gyear,gmonth,gday,ghour,ncol=2,nrow=2,labels=c("year","month","day","hour"))



ggplot(data = webforum,aes(x=substring(Time,4,5),y=Analytic))+
  geom_line()+facet_wrap( ~ substring(Time,1,2), ncol=6)+
  labs(x="minutes")
ggplot(data = webforum,aes(x=substring(Time,4,5),y=Clout))+
```

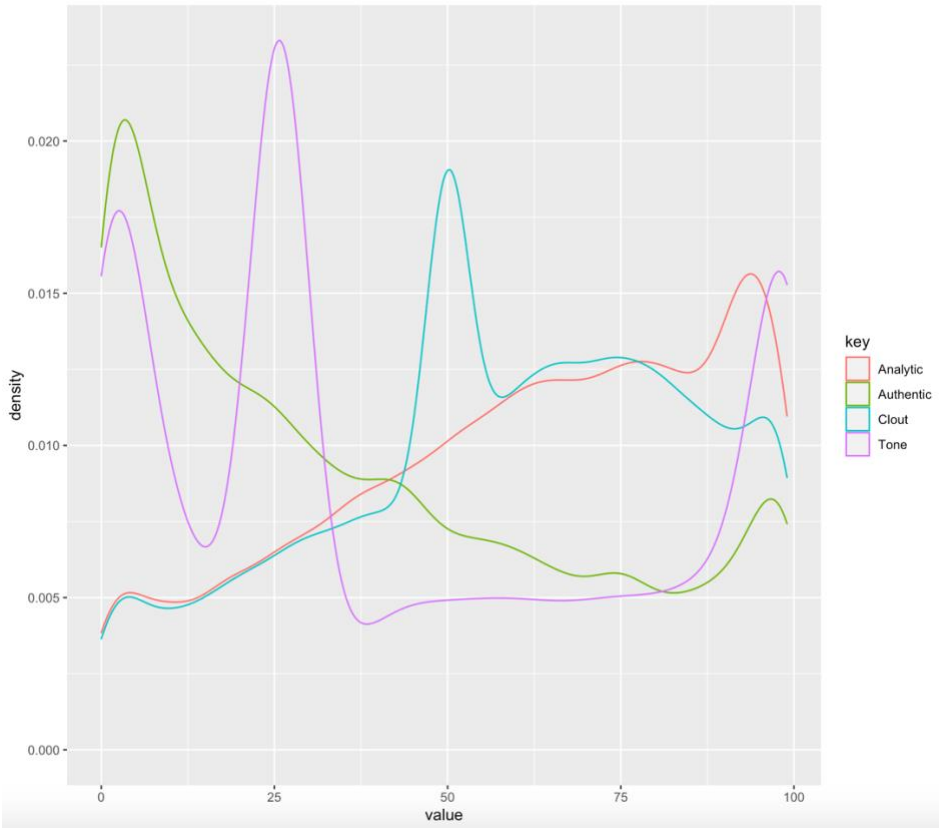
```
  geom_line()+facet_wrap( ~ substring(Time,1,2), ncol=6)+
  labs(x="minutes")
ggplot(data  =        webforum,aes(x=substring(Time,4,5),y=Authentic))+
  geom_line()+facet_wrap( ~ substring(Time,1,2), ncol=6)+
  labs(x="minutes")
ggplot(data  =        webforum,aes(x=substring(Time,4,5),y=Tone))+
  geom_line()+facet_wrap( ~ substring(Time,1,2), ncol=6)+
  labs(x="minutes")
# change of four linguistic variables in every hour
cor(webforum[,6:9])     # correlation coefficient between four variables
library(corrplot)
corrplot(cor(webforum[,6:9]))    # visualization of the correlation of four variables
corrplot(cor(webforum[,10:29][,-(2:6)][,-(4:5)]))    # visualization of the correlation
coefficients of other specific linguistic variables behind
library(tidyr)
library(magrittr)
language<- webforum[,6:9] %>% gather(key, value, Analytic:Tone)
ggplot(data  =        language)+
  geom_density(aes(value,  color = key), alpha = 0.5)
# density map of four variables


########Part b
# only the top five representative groups are selected as the analysis objects
head(thrsort,5)
# select the top 5 most frequently posted groups
library(dplyr)
aa<-webforum %>% filter( ThreadID %in% 283958 )
bb<-webforum %>% filter( ThreadID %in% 252620 )
cc<-webforum %>% filter( ThreadID %in% 127115 )
dd<-webforum %>% filter( ThreadID %in% 145223 )
ee<-webforum %>% filter( ThreadID %in% 472752 )   # filter samples from each group
to construct a sub-data set
tt<-as.data.frame(rbind(aa,bb,cc,dd,ee))


aal<- aa[,6:9] %>% gather(key, value, Analytic:Tone)
aaden<-ggplot(data         =         aal)+
  geom_density(aes(value,  color = key), alpha = 0.5)
bbl<- bb[,6:9] %>% gather(key, value, Analytic:Tone)
```

```
bbden<-ggplot(data = bbl)+
  geom_density(aes(value,  color = key), alpha = 0.5)
ccl<- cc[,6:9] %>% gather(key, value, Analytic:Tone)
ccden<-ggplot(data = ccl)+
  geom_density(aes(value,  color = key), alpha = 0.5)
ddl<- dd[,6:9] %>% gather(key, value, Analytic:Tone)
ddden<-ggplot(data = ddl)+
  geom_density(aes(value,  color = key), alpha = 0.5)
eel<- ee[,6:9] %>% gather(key, value, Analytic:Tone)
eeden<-ggplot(data = eel)+
  geom_density(aes(value,  color = key), alpha = 0.5)
ggarrange(aaden,bbden,ccden,ddden,eeden,ncol=3,nrow=2,labels=c("283958","2526
20","127115","145223","472752"))
# density plot of frequency of use of the four languages by each group
library(reshape)
tt2<-tt[,1:9][,-(2:5)]
tt2 = melt(tt2,id=1:1)
tt2$variable<-as.factor(tt2$variable)
tt2$ThreadID<-as.factor(tt2$ThreadID)
ggplot(tt2,aes(x=ThreadID,y=value))+
  geom_boxplot(aes(colour=variable ))
# box plot of the frequency of four languages by each group
# can be compared between groups or within groups

aa2<-aa[,1:9][,-(2:3)][,-3]
aa2<-melt(aa2,id=1:2)
aat<-ggplot(aa2,aes(x=substring(Time,1,2),y=variable,fill=value))+
  geom_tile()+xlab("hour")
bb2<-bb[,1:9][,-(2:3)][,-3]
bb2<-melt(bb2,id=1:2)
bbt<-ggplot(bb2,aes(x=substring(Time,1,2),y=variable,fill=value))+
  geom_tile()+xlab("hour")
cc2<-cc[,1:9][,-(2:3)][,-3]
cc2<-melt(cc2,id=1:2)
cct<-ggplot(cc2,aes(x=substring(Time,1,2),y=variable,fill=value))+
  geom_tile()+xlab("hour")
dd2<-dd[,1:9][,-(2:3)][,-3]
dd2<-melt(dd2,id=1:2)
ddt<-ggplot(dd2,aes(x=substring(Time,1,2),y=variable,fill=value))+
  geom_tile()+xlab("hour")
```

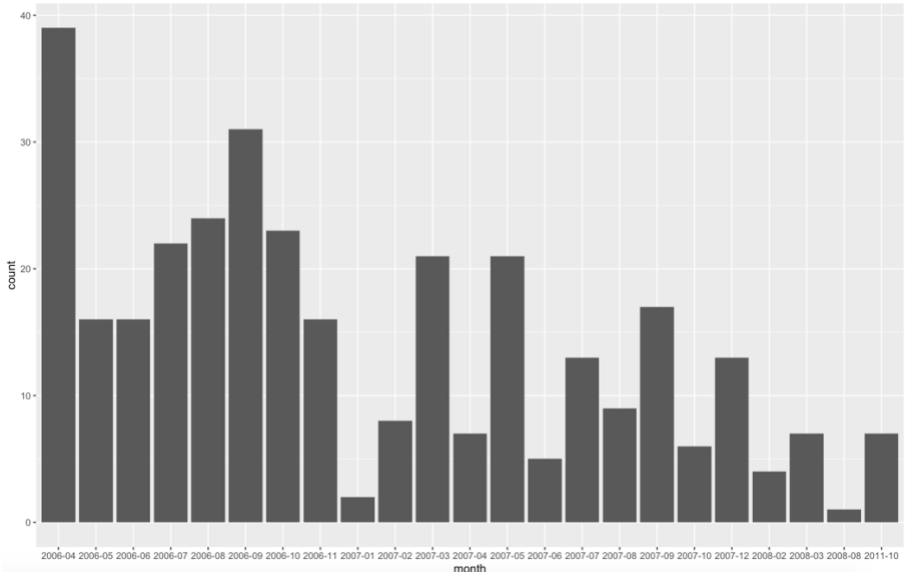
```
ee2<-ee[,1:9][,-(2:3)][,-3]
ee2<-melt(ee2,id=1:2)
eet<-ggplot(ee2,aes(x=substring(Time,1,2),y=variable,fill=value))+
  geom_tile()+xlab("hour")
ggarrange(aat,bbt,cct,eet,ddt,ncol=3,nrow=2,labels=c("283958","252620","127115",
"145223","472752"))
# heat map of four linguistic variables of each group over time


#######Part c
library(igraph)
library(igraphdata)
ggplot(data  =       aa)     +geom_bar(mapping =
aes(x=substring(Date,1,7)))+xlab("month")
# In April 2006, ThreadID283958 posted the most active posts, which was used as the
analysis object
wf<-webforum[,1:9][,-(4:5)]   # delete columns that do not need to be analyzed
wf1<-subset(wf,substring(Date,1,7)=="2006-04")    # filter out all current month
samples
wf1<-wf1%>%
  dplyr::group_by(ThreadID,AuthorID)%>%
  dplyr::summarize(count=n())          # count by groups
g1<-graph.data.frame(wf1,directed = FALSE)     # construct SNA objects
V(g1)
E(g1)    # vertex and edge sequence
is.simple(g1)
plot(g1,layout=layout.fruchterman.reingold,vertex.label.cex=0.6,vertex.size=8)
hist(degree(g1),     breaks        =        15,     col     =        "grey")

#2006-05
wf2<-subset(wf,substring(Date,1,7)=="2006-05")
wf2<-wf2%>%
  dplyr::group_by(ThreadID,AuthorID)%>%
  dplyr::summarize(count=n())
g2<-graph.data.frame(wf2,directed = FALSE)
V(g2)
E(g2)
is.simple(g2)
plot(g2,layout=layout.fruchterman.reingold,vertex.label.cex=0.6,vertex.size=8)
hist(degree(g2),     breaks        =        15,     col     =        "grey")
```

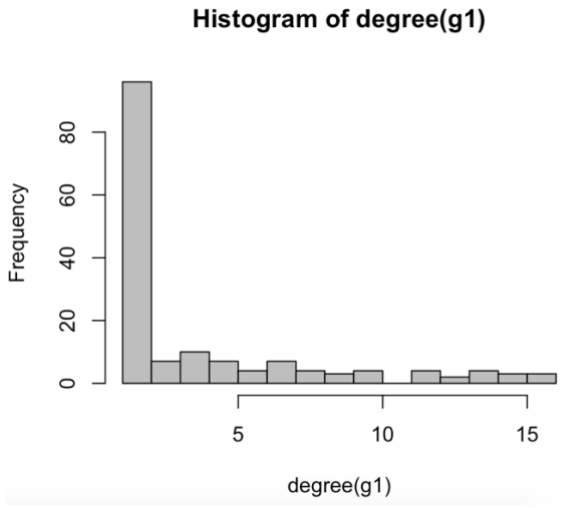
Additional support graphics generate by R:

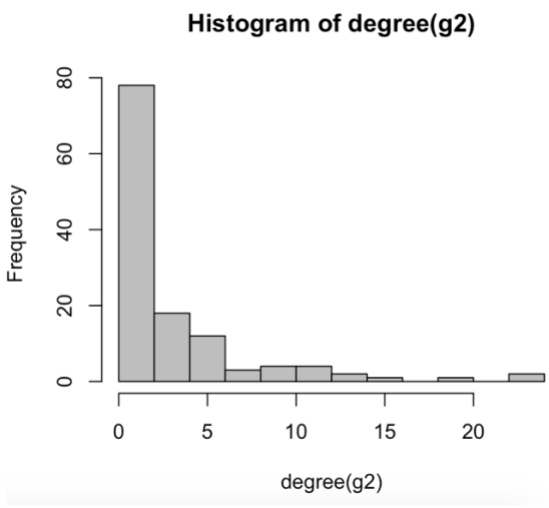
Density map of four variables



Part c other graphics




Histogram of degree(g1)
Frequency
0
20
40
60
80
5
10
15
degree(g1)


Histogram of degree(g2)
Frequency
0
20
40
60
80
0
5
10
15
20
degree(g2)