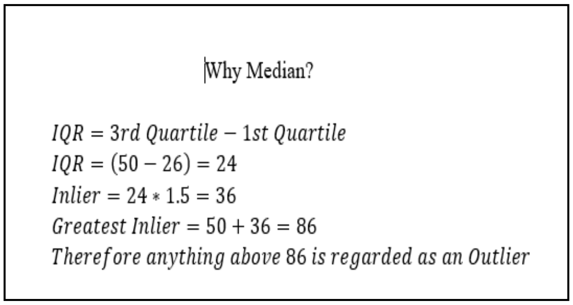
FIT3152 Assignment 1

The aim of this report is to provide an analysis of the webforum dataset which was given. The dataset contains x variables and is based on linguistic analysis of huge numbers of threads and posts between the year 2002 and 2011 which was conducted using Linguistic Inquiry and Word Count (LIWC). There are 3 questions to be addressed: (1) whether participants in an online forum who are communicating directly via threads uses similar language, (2) whether there exist seasonal changes in leisure over certain time period, and (3) whether anonymous presence affects the anger and negemo used in a thread.

1. Subset of the data to be analysed

Posts that has 0 Word Count (WC) is excluded because it contains images or diagrams and will not contribute to the report which analyses about language (text). Posts are then grouped by their threadID.

In grouping the threadID, number of authors in each thread is also taken into consideration (count number of authors in each thread and take average of it. Median is chosen as average because there exist outliers [see figure 1]. Threads which has lower number of authors than the median are not included). This is done to ensure that there are several different authors communicating with each other.



Why Median?

$IQR = 3rd\ Quartile - 1st\ Quartile$
$IQR = (50 - 26) = 24$
$Inlier = 24 * 1.5 = 36$
$Greatest\ Inlier = 50 + 36 = 86$
$Therefore\ anything\ above\ 86\ is\ regarded\ as\ an\ Outlier$

Figure 1: Why Median?

Next, we analyze number of posts in the threads and take the top 6 threads base on the graph in figure 2 (Thread with more than 200 posts). We then take the 6 threads (252620, 283958, 127115, 145223, 472752, and 532649) mentioned above and use it for this assignment

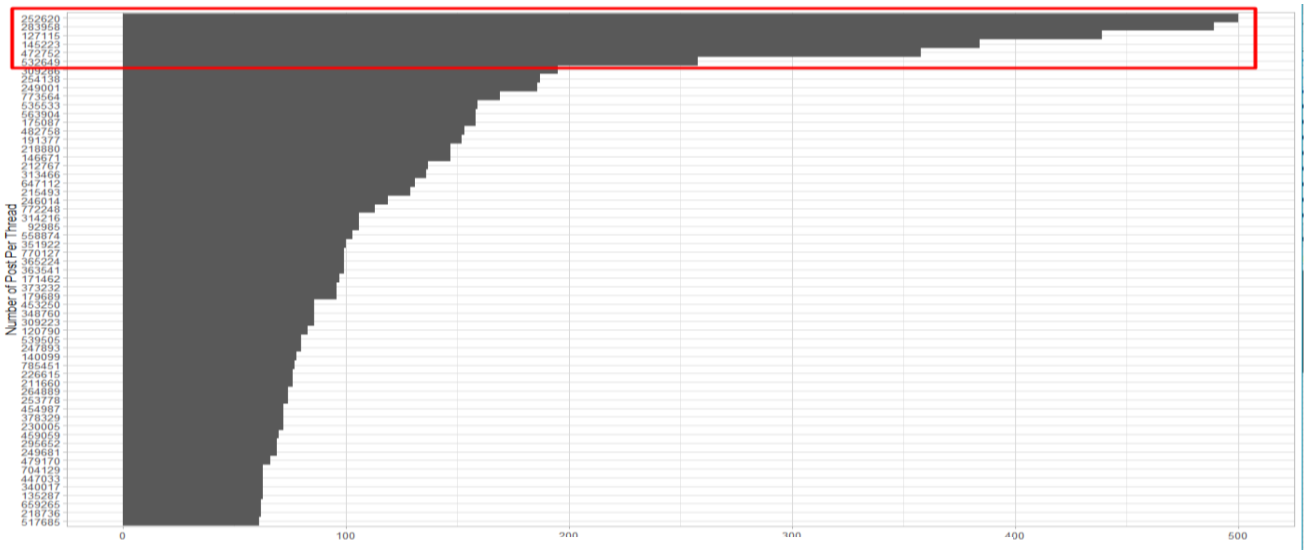


Figure 2: Number of Post in All Threads

2. Analysing a User's Language in Different Threads

Using the top 6 threads above, Analytic is chosen as the attribute to be analysed. Thread 127115 is chosen as the thread to be analysed as it has the top mean of analytic with 84.038 based on Figure 3 (shown by the color red).

Going further into thread 127115, year 2009 is chosen because it has the most number of post compared to the other years in this thread (refer to Figure 4).

Thread 127115 in 2009 was then analysed by AuthorID to get the author with the most number of posts. From Figure 5, author 47875 with the colour orange has most number of post. A conclusion can be derived that author 47875 uses analytic in thread 127115. Next step will be analysing author 47875's language usage in other thread.

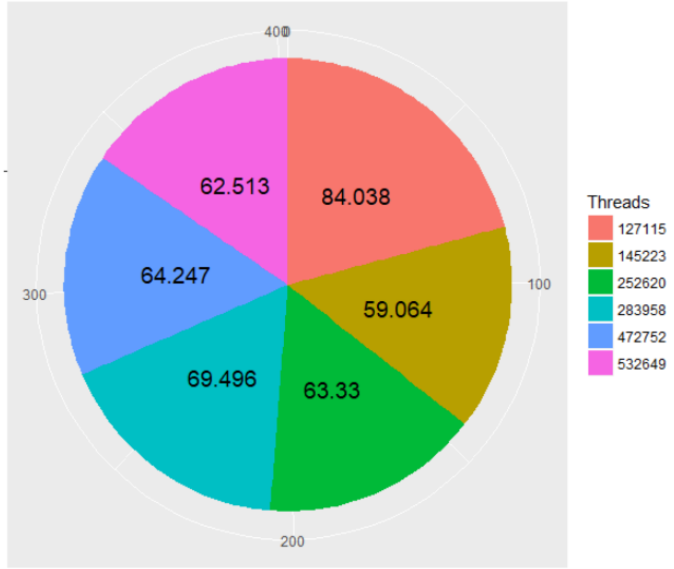


Figure 3: Mean of Analytics Between Top 6 Threads

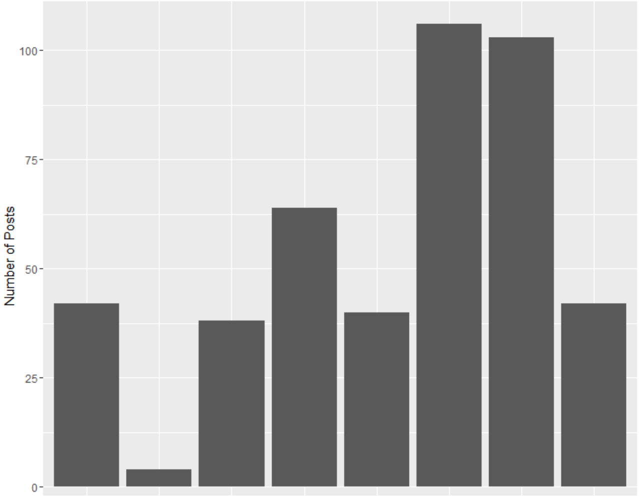


Figure 4: Number of Post by Year in Thread 127115

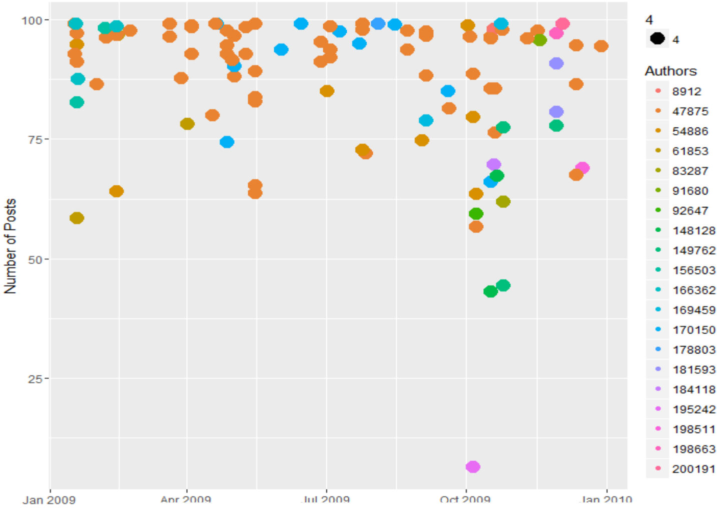
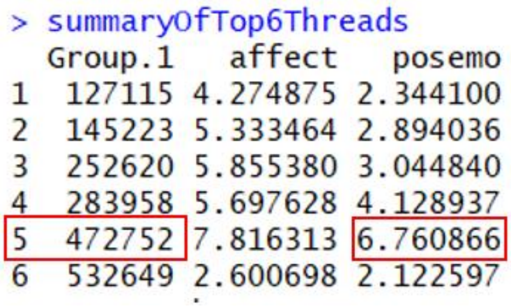


Figure 5: Posts in Thread 127115 by AuthorID

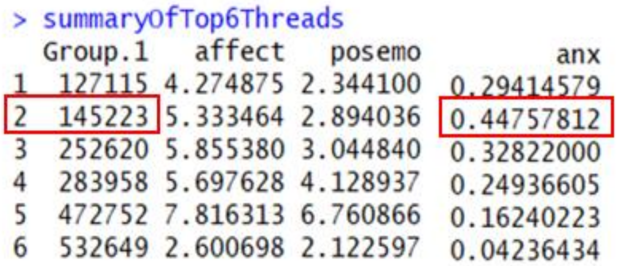
Author 47875 has posted in 3 different threads (472752, 145223, and 532649) which will then be analyzed except thread 532649 because author 47875 only posted once in that thread therefore it is assumed that the sample is not enough. After aggregating the top 6 threads, posemo (positive emotion) was found to be the significantly higher in thread 472752 compared to the other threads (refer to Figure 6) and anx (anxiety) was found to be significantly higher in thread 145223 compared to the other threads (refer to Figure 7).

T-test was also conducted to further support the fact that thread 472752 has higher posemo than the other threads in the top 6 threads (refer to appendix t.test 1) and thread 472752 has higher anx than the other threads in the top 6 threads (refer to appendix t.test 2).



```
> summaryOfTop6Threads
  Group.1   affect   posemo
1  127115 4.274875 2.344100
2  145223 5.333464 2.894036
3  252620 5.855380 3.044840
4  283958 5.697628 4.128937
5  472752 7.816313 6.760866
6  532649 2.600698 2.122597
```

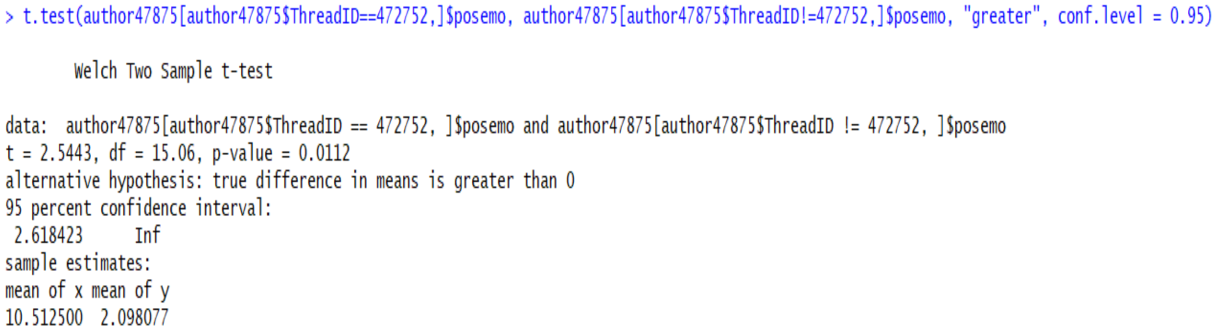
Figure 6: Summary of Top 6 Threads Showing Highest Posemo for Thread 472752



```
> summaryOfTop6Threads
  Group.1   affect   posemo        anx
1  127115 4.274875 2.344100 0.29414579
2  145223 5.333464 2.894036 0.44757812
3  252620 5.855380 3.044840 0.32822000
4  283958 5.697628 4.128937 0.24936605
5  472752 7.816313 6.760866 0.16240223
6  532649 2.600698 2.122597 0.04236434
```

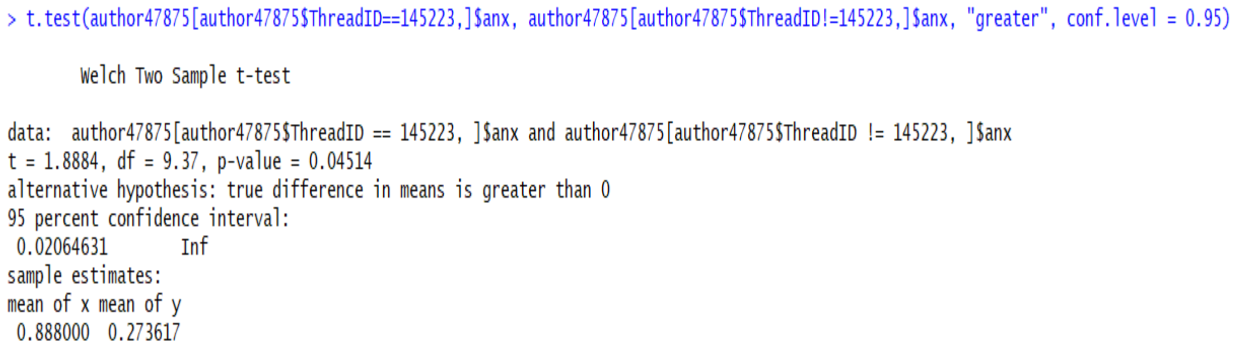
Figure 7: Summary of Top 6 Threads Showing Highest anx for Thread 145223

Another t.test with the confidence level of 0.95 is then conducted on both author 47875's post in the thread 472752 and 145223 to see whether his/her post still uses Analytic or has change according to the thread's theme/attribute. Referring to Figure 8 and Figure 9, p-value was less than the critical value (0.05) which as a result accepts the alternative hypothesis. The conclusion is author 47875's post in thread 472752 has higher positive emotions compared to his/her posts in other threads and author 47875's post in thread 145223 has higher anxiety compared to his/her posts in other threads which indicates that his/her language change according to the thread's theme/attribute.



```
> t.test(author47875[author47875$ThreadID==472752,]$posemo, author47875[author47875$ThreadID!=472752,]$posemo, "greater", conf.level = 0.95)

	Welch Two Sample t-test

data:  author47875[author47875$ThreadID == 472752, ]$posemo and author47875[author47875$ThreadID != 472752, ]$posemo
t = 2.5443, df = 15.06, p-value = 0.0112
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.618423      Inf
sample estimates:
mean of x mean of y 
10.512500  2.098077 
```

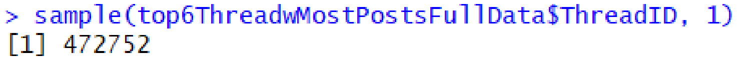
Figure 8: t.test of Author 47875's Post in Thread 472752 with His/Her Other Posts Regarding Posemo



```
> t.test(author47875[author47875$ThreadID==145223,]$anx, author47875[author47875$ThreadID!=145223,]$anx, "greater", conf.level = 0.95)

	Welch Two Sample t-test

data:  author47875[author47875$ThreadID == 145223, ]$anx and author47875[author47875$ThreadID != 145223, ]$anx
t = 1.8884, df = 9.37, p-value = 0.04514
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.02064631        Inf
sample estimates:
mean of x mean of y 
 0.888000  0.273617 
```

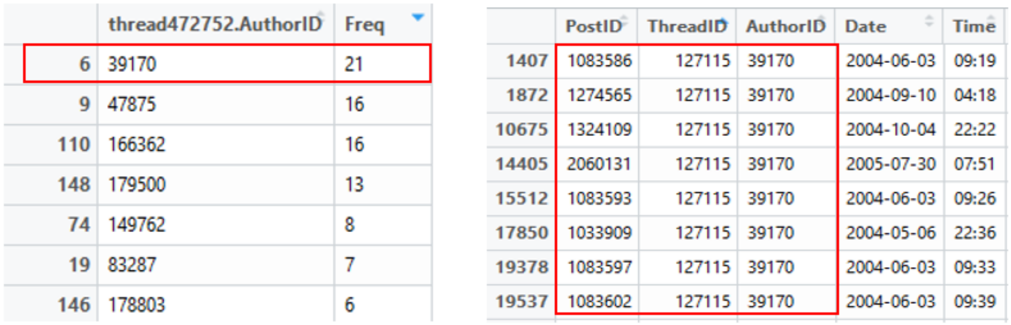
Figure 9: t.test of Author 47875's Post in Thread 145223 with His/Her Other Posts Regarding anx

To further support the conclusion above, a random thread is taken from the top 6 threads with most posts using random sampling. Thread 472752 was chosen then as the random thread (Figure 10). The same approach as in the analytic thread was chosen which is taking the author with most post in thread 472752 (author 39170) and seeing his/her posts in another thread to see his/her behavior regarding language usage (Figure 11). Author 39170 has posted in thread 127115 which is the analytic thread therefore the method before can be reused as well. T.test with confidence level of 0.95 was conducted to test the significance of analytics on average and the result was that the p-value was less than the critical value (0.05) which thus accepts the alternative hypothesis (Figure 12). This finding further supports the analysis that user's language in the posts changes according to the language used by others in the thread.



```
> sample(top6ThreadwMostPostsFullData$ThreadID, 1)
[1] 472752
```

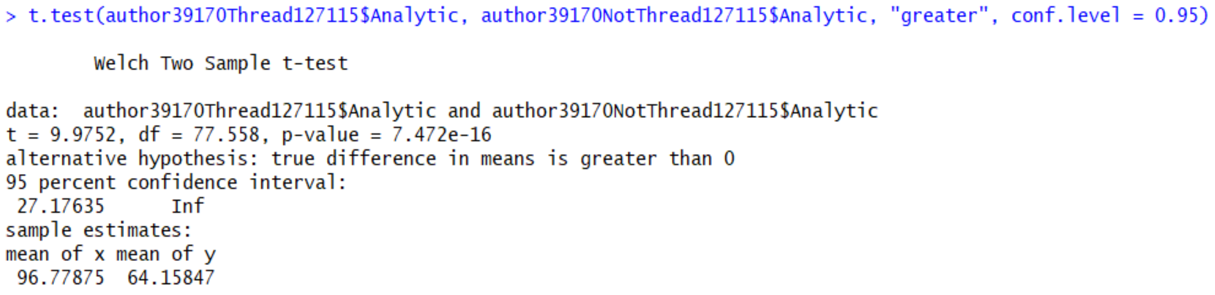
Figure 10: Random Sampling of Thread



	thread472752.AuthorID	Freq
6	39170	21
9	47875	16
110	166362	16
148	179500	13
74	149762	8
19	83287	7
146	178803	6

	PostID	ThreadID	AuthorID	Date	Time
1407	1083586	127115	39170	2004-06-03	09:19
1872	1274565	127115	39170	2004-09-10	04:18
10675	1324109	127115	39170	2004-10-04	22:22
14405	2060131	127115	39170	2005-07-30	07:51
15512	1083593	127115	39170	2004-06-03	09:26
17850	1033909	127115	39170	2004-05-06	22:36
19378	1083597	127115	39170	2004-06-03	09:33
19537	1083602	127115	39170	2004-06-03	09:39

Figure 11: Author with Most Post in Thread 472752 and His/Her Posts in Other Threads



```
> t.test(author39170Thread127115$Analytic, author39170NotThread127115$Analytic, "greater", conf.level = 0.95)

	Welch Two Sample t-test

data:  author39170Thread127115$Analytic and author39170NotThread127115$Analytic
t = 9.9752, df = 77.558, p-value = 7.472e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 27.17635      Inf
sample estimates:
mean of x mean of y 
 96.77875  64.15847 
```

Figure 12: t.test of Author 39170's Post in Thread 127115 With His/Her Other Posts

3. Analysing User’s Language Within a Time Period

In analysing the top 6 threads, thread 252620 only has posts on December 2005, January 2006, and December 2006 where in December 2005 it has a significant difference than the other time periods. (Refer to Figure 13). Therefore, thread 252620 was eliminated to prevent bias in the analysis

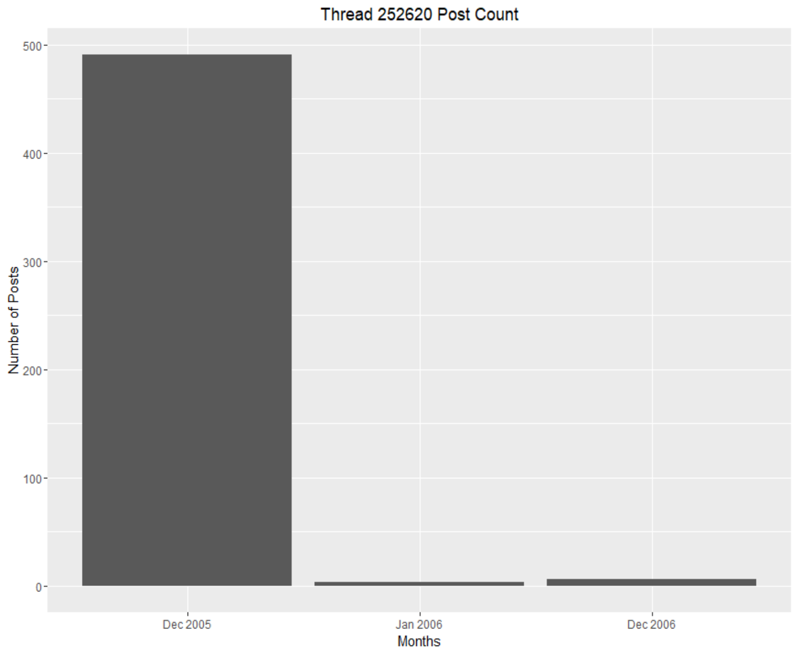
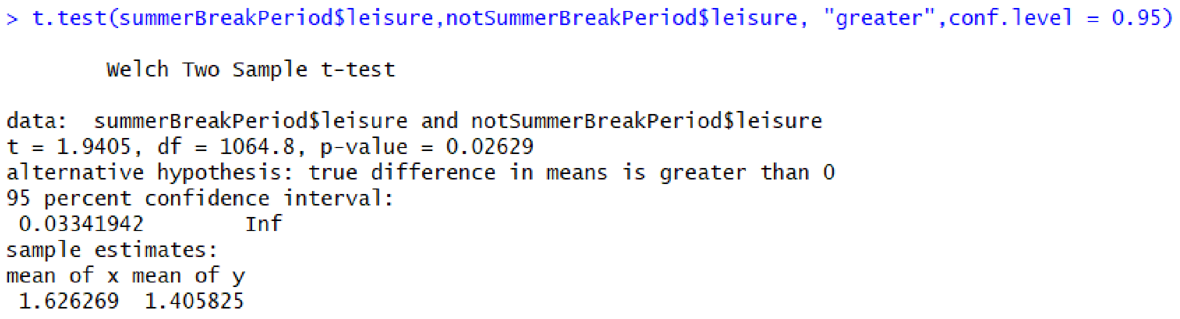


Figure 13: Thread 252620 Post Count

The aim of this section is to compare the number of posts that contains leisure between a certain time period with another time period. The time period to be compared is school summer break (except Australia and New Zealand) with the rest of the months. The demographics of people aged 18-29 using online forums are 23% out of 48% which shows that almost half of the forum users are from either college or university (Pew Research Center: Internet, Science & Tech, 2015). They should be active on the long holidays which falls in the month of June to September (Musiker Discovery Programs, 2013). This is the assumption that is taken for the analysis. T.test was performed to support the analysis that the number of posts containing leisure increases during the summer break period compared to the rest of the months (refer to figure 14).



```
> t.test(summerBreakPeriod$leisure,notSummerBreakPeriod$leisure, "greater",conf.level = 0.95)

        Welch Two Sample t-test

data:  summerBreakPeriod$leisure and notSummerBreakPeriod$leisure
t = 1.9405, df = 1064.8, p-value = 0.02629
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.03341942        Inf
sample estimates:
mean of x mean of y 
 1.626269  1.405825 
```

Figure 14: t.test of Summer Break vs Non-Summer Break

The result of the t-test above with the confidence level of 0.95 shows that the p-value (0.02629) is lower than the critical value which is 0.05. Therefore, we can reject the null hypothesis and accept the alternative hypothesis where the number of posts about leisure on summer break period is more than the rest of the months in average. To further support the statement above, years with more than 300 posts (chosen threshold) were chosen out of the top 6 threads without thread 252620 which was Year 2006 and 2009 (refer to Figure 15) and then t.test was performed on these 2 years (Figure 16 and 17). The results of the t.test was in agreement with the initial finding hence posts on summer break was proven to contain more leisure in average than posts on non-summer break.

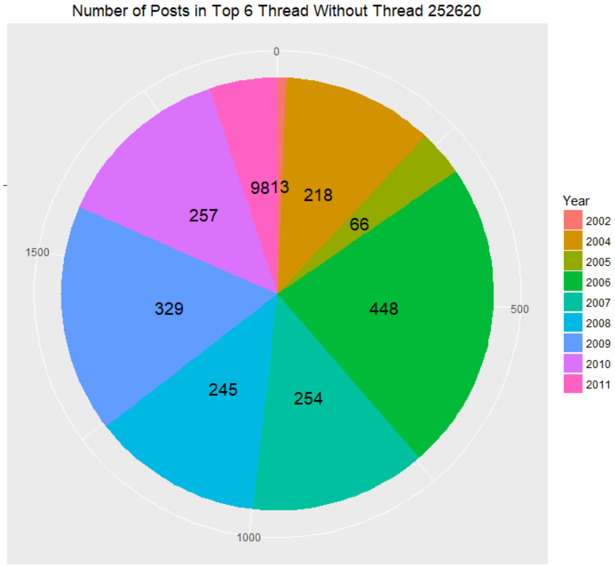
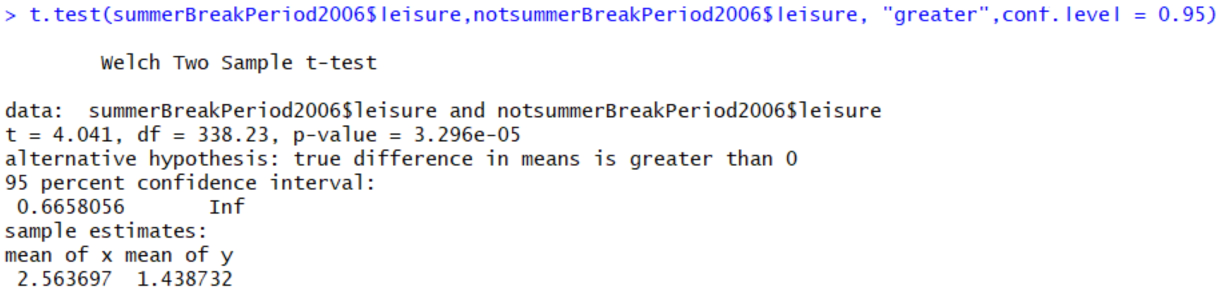
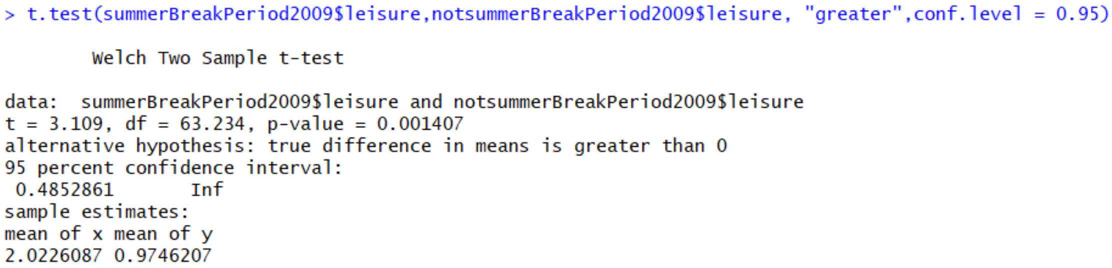


Figure 15: Pie Chart of Number of Posts in Top 6 Threads Without Thread 252620



```
> t.test(summerBreakPeriod2006$leisure,notsummerBreakPeriod2006$leisure, "greater",conf.level = 0.95)

        Welch Two Sample t-test

data:  summerBreakPeriod2006$leisure and notsummerBreakPeriod2006$leisure
t = 4.041, df = 338.23, p-value = 3.296e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6658056       Inf
sample estimates:
mean of x mean of y 
 2.563697  1.438732 
```

Figure 16: t.test on 2006 Summer Break vs Non-Summer Break on 5 Threads

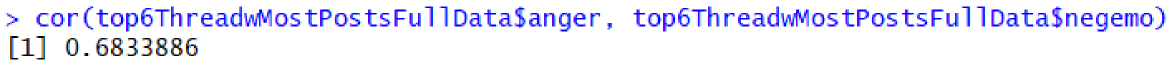


```
> t.test(summerBreakPeriod2009$leisure,notsummerBreakPeriod2009$leisure, "greater",conf.level = 0.95)

        Welch Two Sample t-test

data:  summerBreakPeriod2009$leisure and notsummerBreakPeriod2009$leisure
t = 3.109, df = 63.234, p-value = 0.001407
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.4852861       Inf
sample estimates:
mean of x mean of y 
2.0226087 0.9746207 
```

Figure 17: t.test on 2009 Summer Break vs Non-Summer Break on 5 Threads

4. Analysing Anonymous Effect on Other User’s Language

In analysing all of the attributes, negemo and anger were found to have the highest correlation among the top 6 threads’ attributes although it is only 0.68 or 68% (Figure 18). Therefore, anger and negemo will be used in further analysing anonymous effect on other user’s language. A multivariate graph was also created to give a visualization that anger and negemo does correlate with each other (meaning high anger leads to high negemo, refer to Figure 19).



```
> cor(top6ThreadwMostPostsFullData$anger, top6ThreadwMostPostsFullData$negemo)
[1] 0.6833886
```

Figure 18: Correlation Between Anger and Negemo Between Top 6 Most Posts Threads

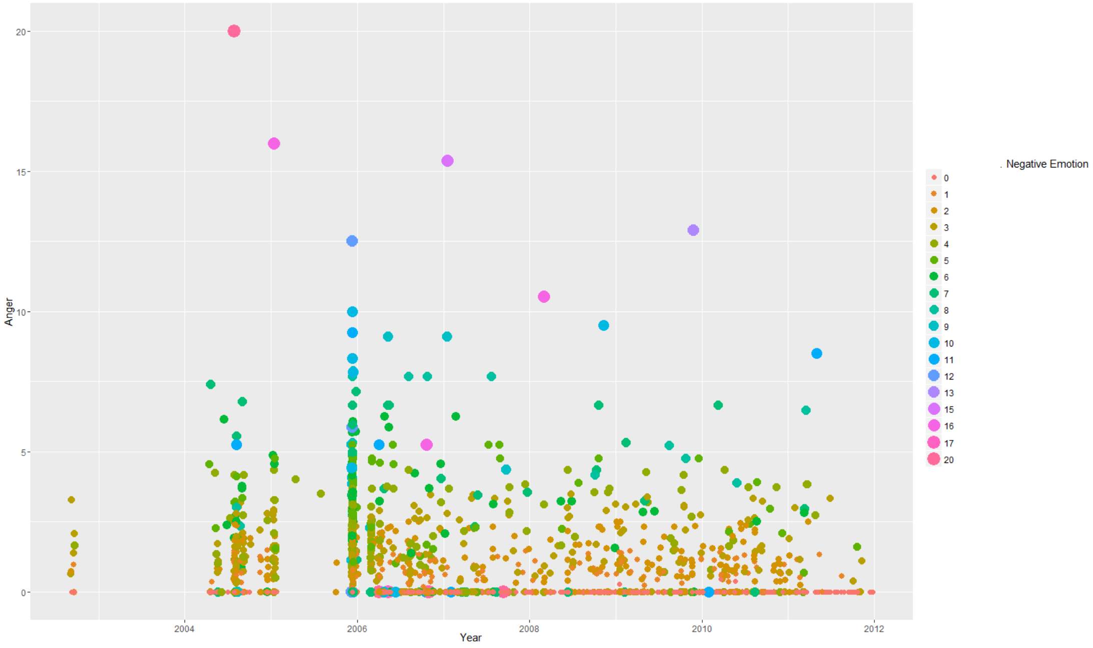


Figure 19: Correlation Between Anger and Negemo Between Top 6 Most Posts Threads

Using the top 6 threads without thread 252620 (to avoid bias due to concentrated posts in 3 months only), anonymous only posts were taken and then plotted against the year (Figure 20). The findings were that there 2qw an incline and decline in number of posts by anonymous during a certain period of time. Next step is comparing the overall posts containing anonymous against overall posts without anonymous in the top 6 threads without thread 252620.

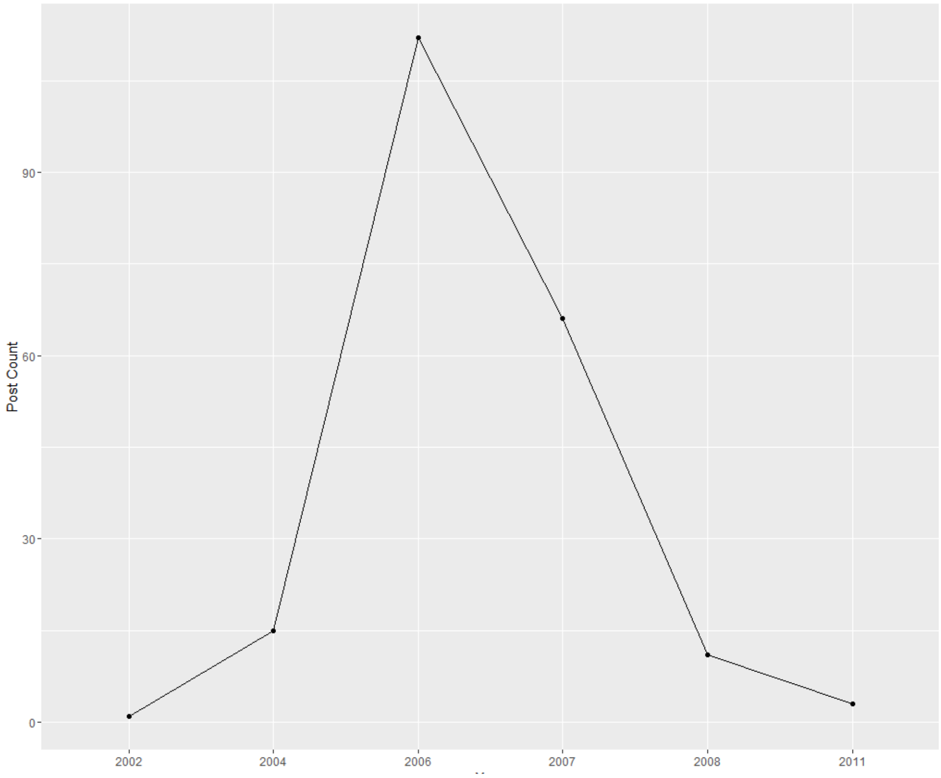
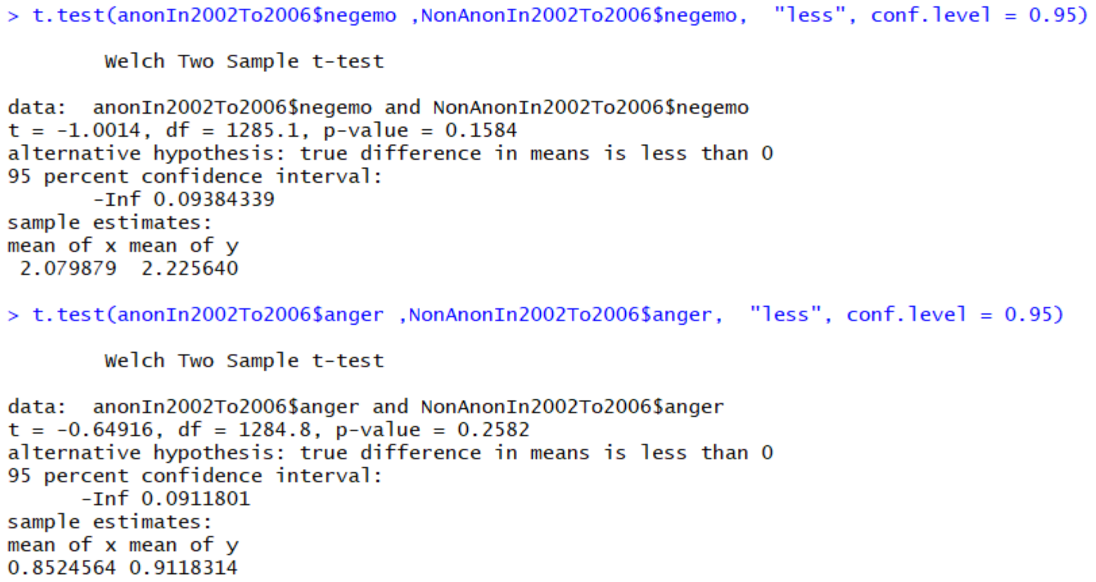


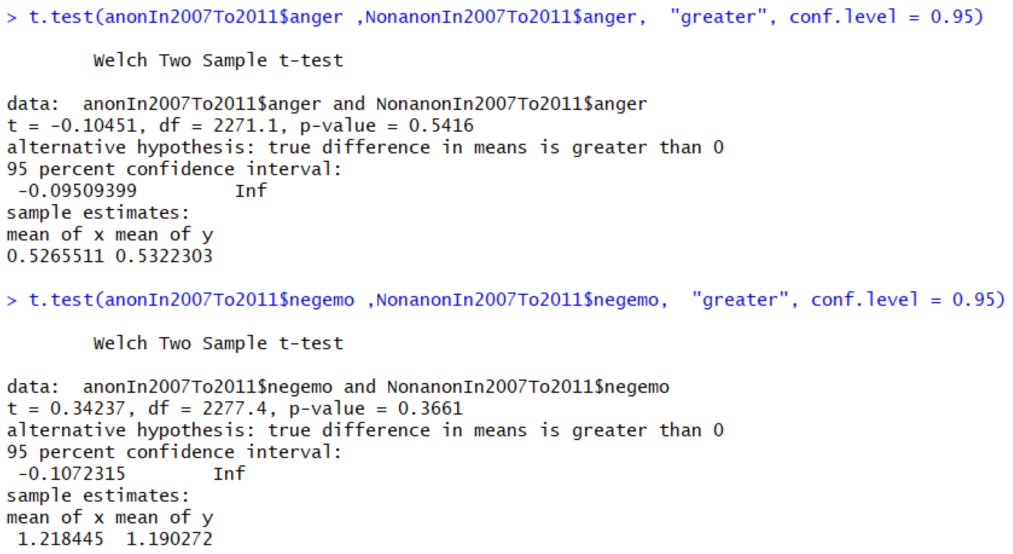
Figure 20: Number of Posts by Anonymous in The Top 6 Threads Without 252620

Then t.test were performed regarding negemo and anger to check if the decline or incline of the number of posts by anonymous affect the overall attributes of the top 6 threads. The t.test below (Figure 21 and 22) shows a result that the change in anonymous' posts doesn't affect the negemo and anger in the top 6 threads (p-value was less than critical value which is 0.05).



```
> t.test(anonIn2002To2006$negemo ,NonAnonIn2002To2006$negemo,  "less", conf.level = 0.95)

        Welch Two Sample t-test

data:  anonIn2002To2006$negemo and NonAnonIn2002To2006$negemo
t = -1.0014, df = 1285.1, p-value = 0.1584
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.09384339
sample estimates:
mean of x mean of y 
 2.079879  2.225640 

> t.test(anonIn2002To2006$anger ,NonAnonIn2002To2006$anger,  "less", conf.level = 0.95)

        Welch Two Sample t-test

data:  anonIn2002To2006$anger and NonAnonIn2002To2006$anger
t = -0.64916, df = 1284.8, p-value = 0.2582
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf 0.0911801
sample estimates:
mean of x mean of y 
0.8524564 0.9118314 
```

Figure 21: t.test on Anger and Negemo During 2002 Until 2006 Full Data vs Without Anonymous



```
> t.test(anonIn2007To2011$anger ,NonanonIn2007To2011$anger,  "greater", conf.level = 0.95)

        Welch Two Sample t-test

data:  anonIn2007To2011$anger and NonanonIn2007To2011$anger
t = -0.10451, df = 2271.1, p-value = 0.5416
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.09509399         Inf
sample estimates:
mean of x mean of y 
0.5265511 0.5322303 

> t.test(anonIn2007To2011$negemo ,NonanonIn2007To2011$negemo,  "greater", conf.level = 0.95)

        Welch Two Sample t-test

data:  anonIn2007To2011$negemo and NonanonIn2007To2011$negemo
t = 0.34237, df = 2277.4, p-value = 0.3661
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1072315        Inf
sample estimates:
mean of x mean of y 
 1.218445  1.190272 
```

Figure 22: t.test on Anger and Negemo During 2007 Until 2011 Full Data vs Without Anonymous

Conclusion

To summarize, all the analysis that have been mentioned above contribute to answer the two main questions related to the social and linguistic dynamics of an on-line community:

- Members communicate using similar language in a thread: This was proven by taking an author (author 47875) from an analytical thread and investigating his/her posts in another thread. The result was author 47875 followed the language used in thread 145223 (anx) and 472752 (posemo). Taking another author (author 39170) from a posemo thread, his/her post in another thread which was thread 127115 (analytical) also proves that the author followed the same language in the thread.
- There exist seasonal changes in leisure over certain time: The time period which was summer break (between June-September) are proven to increase posts about leisure. This was proven by using t.test on years with more than 300 posts (2006 and 2009). The results support the fact that mean of leisure increases in summer break than the other times in a year.
- The posts made by anonymous authors do not affect the language used by authors in a thread (in terms of anger and negemo). T.tests were conducted on period of time where anonymous users increase and decrease between full posts data against posts without anonymous. Results of the test are not significant enough to deduce that anonymous affects language.

Member/Task	Rico Jap	Reynald Nixon	Total
Preliminary analysis	45%	55%	100%
R Research and Coding	60%	40%	100%
Preparation of Graphics	55%	45%	100%
Analysis of Results	50%	50%	100%
Writing up the report	40%	60%	100%

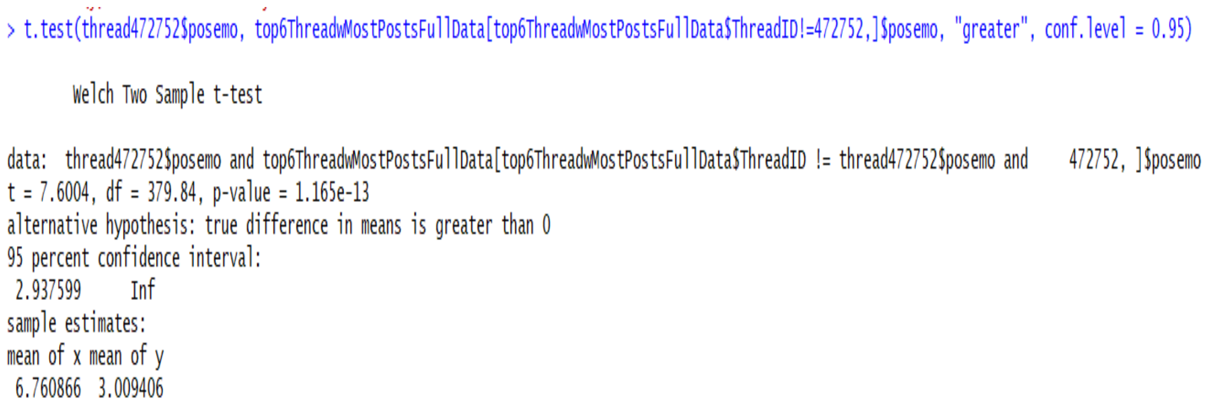
REFERENCES

Monash University Faculty of Information Technology: FIT1006. (n.d.). FIT1006 Business Information Analysis Revision Notes for FIT3152 Data Analytics (Slides 83/410) [online] Available at: FIT3152 Semester 2 2017 Moodle [Accessed 14 Sep. 2017]

Pew Research Center: Internet, Science & Tech. (2015). *Demographics of Online Discussion Forums*. [online] Available at: http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/2015-08-19_social-media-update_04/ [Accessed 15 Sep. 2017].

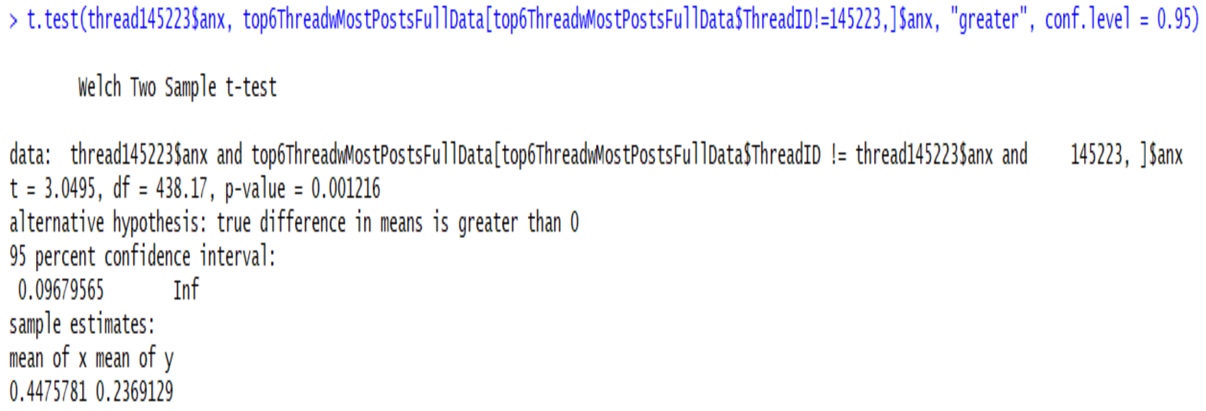
Musiker Discovery Programs, I. (2013). *Summer vacation around the world*. [online] Summer Discovery. Available at: https://www.summerdiscovery.com/blog/2013-11-15/summer-vacation-around-the-world [Accessed 15 Sep. 2017].

Appendix



```
> t.test(thread472752$posemo, top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID!=472752,]$posemo, "greater", conf.level = 0.95)

        Welch Two Sample t-test

data:  thread472752$posemo and top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID != thread472752$posemo and     472752, ]$posemo
t = 7.6004, df = 379.84, p-value = 1.165e-13
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.937599      Inf
sample estimates:
mean of x mean of y 
 6.760866  3.009406 
```

t.test 1: t.test for Posemo in thread 472752 against other threads in top 6 threads



```
> t.test(thread145223$anx, top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID!=145223,]$anx, "greater", conf.level = 0.95)

        Welch Two Sample t-test

data:  thread145223$anx and top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID != thread145223$anx and     145223, ]$anx
t = 3.0495, df = 438.17, p-value = 0.001216
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.09679565        Inf
sample estimates:
mean of x mean of y
0.4475781 0.2369129
```

t.test 2: t.test for anx in thread 145223 against other threads in top 6 threads

R Script:

```
install.packages("zoo")

install.packages("lubridate")

install.packages("ggplot2")

install.packages("plyr")

install.packages("dplyr")

install.packages("scales")


library(zoo)

library(lubridate)

library(ggplot2)

library(plyr)

library(dplyr)

library(scales)
```

```
specify_decimal <- function(x, k) trimws(format(round(x, k), nsmall=k))


                                        #Subset data

#read the webforum data and put it to data frame

webforum = read.csv("webforum.csv", header = TRUE)

#eliminate the 0 word count

wc0Webforum = webforum[!(webforum$WC==0),]

#make the date column as a date type

wc0Webforum$Date = as.Date(wc0Webforum$Date)

#find out how many posts per thread

countPerThread = aggregate(cbind(count = wc0Webforum$ThreadID) ~ wc0Webforum$ThreadID, data = wc0Webforum, FUN = function(x){NROW(x)})


#get the threads where the number of posts is more than the median

postsMoreThanMedian = countPerThread[countPerThread$count > median(countPerThread$count),]


#get the full data of each threadID according to PostsMoreThanMedian

wc0MrThMednWebforum = as.data.frame(wc0Webforum[(wc0Webforum$ThreadID %in% postsMoreThanMedian$'wc0Webforum$ThreadID'), ])


#get the number of authors per thread, so find out how many number of unique authors there are in each thread

nOfAuthPerThread = as.data.frame(as.table(by(wc0MrThMednWebforum, wc0MrThMednWebforum$ThreadID, function(df) length(unique(df$AuthorID)))))
```

```
#get the number of threads where there is more authors than the median number of authors in
a thread

nOfAuthPerThreadMoreThanMedian = nOfAuthPerThread[nOfAuthPerThread$Freq >
median(nOfAuthPerThread$Freq),]


#get the full data of each authorID according to nOfAuthPerThreadMoreThanMedian

wc0_MrThMednP_and_A_Webforum =
as.data.frame(wc0MrThMednWebforum[(wc0MrThMednWebforum$ThreadID %in%
nOfAuthPerThreadMoreThanMedian$wc0MrThMednWebforum.ThreadID), ])



#get the number of posts per thread

numOfPostsPerThread = as.data.frame(as.table(by(wc0_MrThMednP_and_A_Webforum,
wc0_MrThMednP_and_A_Webforum$ThreadID, function(df) length(df$PostID))))



#get the graph based on number of posts per thread

PostCountGraph = ggplot(numOfPostsPerThread,
aes(reorder(numOfPostsPerThread$wc0_MrThMednP_and_A_Webforum.ThreadID,
numOfPostsPerThread$Freq), numOfPostsPerThread$Freq)) + geom_bar(width = 1, stat=
"identity") + labs(x="Number of Post Per Thread",y="Thread ID") + theme_light() + coord_flip()


#GRAPH

dev.new()

PostCountGraph

#get the top 6 threads w most posts because the fifth and the sixth most posts have the same
value

top6ThreadwMostPosts =
numOfPostsPerThread[order(numOfPostsPerThread$Freq,decreasing=T)[1:6],]
```

```
#get the top 6 threads' full data

top6ThreadwMostPostsFullData =
as.data.frame(wc0_MrThMednP_and_A_Webforum[(wc0_MrThMednP_and_A_Webforum$Th
readID %in% top6ThreadwMostPosts$wc0_MrThMednP_and_A_Webforum.ThreadID),])
```

```
#Analysing Analytics

#get the thread which is the most analytical based on median since a couple of threads has
outliers such as thread 127115

meanOfAnalyticForEachTop6Thread =
as.data.frame(as.table(by(top6ThreadwMostPostsFullData,
top6ThreadwMostPostsFullData$ThreadID, function(df)
as.numeric(specify_decimal(mean(df$Analytic), 3)))))
```

```
#get the mean of analytics of the top 6 threads

#GRAPH

#bar graph for analytics for top 6 threads

bpFormeanOfAnalyticForEachTop6Thread = ggplot(meanOfAnalyticForEachTop6Thread,
aes(x="", y=meanOfAnalyticForEachTop6Thread$Freq,
fill=meanOfAnalyticForEachTop6Thread$top6ThreadwMostPostsFullData.ThreadID)) +
geom_bar(width = 1, stat= "identity")
```

```
#make a piechart

pieFormeanOfAnalyticForEachTop6Thread = bpFormeanOfAnalyticForEachTop6Thread +
coord_polar("y", start=0) + geom_text(aes(y = meanOfAnalyticForEachTop6Thread$Freq/2 +
c(0, cumsum(meanOfAnalyticForEachTop6Thread$Freq)[-
length(meanOfAnalyticForEachTop6Thread$Freq)]), label =
meanOfAnalyticForEachTop6Thread$Freq), size=5) + labs(x="", y="") +
scale_fill_discrete("Threads")
```

```
dev.new()

pieFormeanOfAnalyticForEachTop6Thread


#get the thread where threadID is 127115

thread127115 = top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID == 127115,]




#statistics of number of posts per year in thread 127115 since it is the most analytical thread

thread127115findTopPostPerYear = as.data.frame(as.table(by(thread127115, year(thread127115$Date), function(df) NROW(df$PostID))))


dev.new()

#bar graph based on thread127115findTopPostPerYear

ggplot(thread127115findTopPostPerYear, aes(x=thread127115findTopPostPerYear$year.thread127115.Date., y=thread127115findTopPostPerYear$Freq)) + geom_bar(stat = "identity") + labs(x="Year", y="Number of Posts")


#pick year 2009 to be analyzed since it has the most posts

thread127115in2009 = thread127115[thread127115$Date >= as.Date("2009-01-01") & thread127115$Date <= as.Date("2009-12-31"),]


#combine date and time columns

thread127115in2009$DateTime = as.POSIXct(paste(thread127115in2009$Date, thread127115in2009$Time), format = "%Y-%m-%d %H:%M")
```

```r
#GRAPH

dev.new()


#graph to show which authors have posted the most

qplot(thread127115in2009$DateTime, thread127115in2009$Analytic,
data=thread127115in2009, size=4, color=as.factor(thread127115in2009$AuthorID)) +
labs(x="Year", y="Number of Posts")  + scale_colour_discrete(name = "Authors")



#get the data for the author with most posts in thread 127115

author47875 = top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$AuthorID ==
47875,]


#the mean can be seen that author 47875 in thread 127115 is the most analytical and this
supports the thread 127115 that is the most analytical thread.

meanOfAnalyticOfAuthr47875 = as.data.frame(as.table(by(author47875,
author47875$ThreadID, function(df) as.numeric(specify_decimal(mean(df$Analytic), 3)))))


#to see which threads that the author 47875 has posted

author47875PostCountPerThread = as.data.frame(as.table(by(author47875,
author47875$ThreadID, function(df) NROW(df$PostID))))


#get the mean of all the attributes in the top 6 threads

summaryOfTop6Threads = aggregate(top6ThreadwMostPostsFullData[18:32],
list(top6ThreadwMostPostsFullData$ThreadID), mean)



#get the posts in thread 472752

thread472752 =
top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID==472752,]
```

```
#get the posts in thread 145223

thread145223 =
top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID==145223,]


#t-test to compare the posemo in thread 472752 to the rest of the thread

t.test(thread472752$posemo,
top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID!=472752,]$posemo
, "greater", conf.level = 0.95)

#t-test to compare the posemo in thread 472752 based on the posts that the author 47875
made to the rest of the posts that he/she made

t.test(author47875[author47875$ThreadID==472752,]$posemo,
author47875[author47875$ThreadID!=472752,]$posemo, "greater", conf.level = 0.95)


#t-test to compare the anxiety in thread 145223 to the rest of the thread

t.test(thread145223$anx,
top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID!=145223,]$anx,
"greater", conf.level = 0.95)

#t-test to compare the anxiety in thread 145223 based on the posts that the author 47875
made to the rest of the posts that he/she made

t.test(author47875[author47875$ThreadID==145223,]$anx,
author47875[author47875$ThreadID!=145223,]$anx, "greater", conf.level = 0.95)


#get the author 39170 which has the most posts in thread 472752

author39170 =
top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$AuthorID==39170,]


#t-test to compare the analytics in thread 127115 based on the posts that the author 47875
made to the rest of the posts that he/she made

t.test(author39170[author39170$ThreadID==127115,]$Analytic,
author39170[author39170$ThreadID!=127115,]$Analytic, "greater", conf.level = 0.95)
```

```
#Analysing Leisure

#we decided to remove thread 252620 because the thread is only active between 3 months only which is not suitable in choosing the most posts in a year

top5ThreadWout252620 = top6ThreadwMostPostsFullData[!(top6ThreadwMostPostsFullData$ThreadID==252620),]

#get the posts with thread 252620

thread252620 = top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$ThreadID == 252620,]

#get the post count of thread 252620 per month

thread252620PostCount = as.data.frame(as.table(by(thread252620, as.yearmon(thread252620$Date), function(df) NROW(df$PostID))))

#GRAPH

dev.new()

#plot the post count for th thread 252620 to show that the thread is focused on certain time frame

ggplot(thread252620PostCount, aes(x=thread252620PostCount$as.yearmon.thread252620.Date., y=thread252620PostCount$Freq)) + geom_bar(stat = "identity") + labs(x="Months",y="Number of Posts") + ggtitle("Thread 252620 Post Count")

#define summer break period

summerBreakPeriod = top5ThreadWout252620[month(top5ThreadWout252620$Date)==6 | month(top5ThreadWout252620$Date)==7 | month(top5ThreadWout252620$Date)==8 | month(top5ThreadWout252620$Date)==9,]
```

```
#define non summer break period

notSummerBreakPeriod = top5ThreadWout252620[month(top5ThreadWout252620$Date)!=6 |
month(top5ThreadWout252620$Date)!=7 | month(top5ThreadWout252620$Date)!=8 |
month(top5ThreadWout252620$Date)!=9,]


#perform a t-test on the summer break period and no summer break period

t.test(summerBreakPeriod$leisure,notSummerBreakPeriod$leisure, "greater",conf.level = 0.95)


#get the number of posts per year and the maximum is 2005 where it has the most number of
posts

noOfPostsPerYearTop5 = as.data.frame(as.table(by(top5ThreadWout252620,
year(top5ThreadWout252620$Date), function(df) length(df$PostID))))


#number of posts in top 6 thread without thread 252620

bpFornoOfPostsPerYearTop5 = ggplot(noOfPostsPerYearTop5, aes(x="",
y=noOfPostsPerYearTop5$Freq,
fill=noOfPostsPerYearTop5$year.top5ThreadWout252620.Date.)) + geom_bar(width = 1, stat=
"identity") +  ggtitle("Number of Posts in Top 6 Thread Without Thread 252620") +
scale_fill_discrete(name = "Year") + labs(x="", y="")

#make bpFornoOfPostsPerYearTop5 in pie chart

pieFornoOfPostsPerYearTop5 = bpFornoOfPostsPerYearTop5 + coord_polar("y", start=0) +
geom_text(aes(y = noOfPostsPerYearTop5$Freq/2 + c(0,
cumsum(noOfPostsPerYearTop5$Freq)[-length(noOfPostsPerYearTop5$Freq)]), label =
noOfPostsPerYearTop5$Freq), size=5)


#GRAPH

dev.new()

pieFornoOfPostsPerYearTop5
```

```
#take the summer break on 2006

summerBreakPeriod2006 = top5ThreadWout252620[top5ThreadWout252620$Date >=
as.Date("2006-06-01") & top5ThreadWout252620$Date <= as.Date("2006-09-30"),]

#take the period where it is not summer break on 2006

notsummerBreakPeriod2006  = top5ThreadWout252620[top5ThreadWout252620$Date >=
as.Date("2006-01-05") & top5ThreadWout252620$Date <= as.Date("2006-05-31"),]


#perform a t-test in 2006 on summer break vs not summer break period

t.test(summerBreakPeriod2006$leisure,notsummerBreakPeriod2006$leisure,
"greater",conf.level = 0.95)


#take the summer break on 2009

summerBreakPeriod2009 = top5ThreadWout252620[top5ThreadWout252620$Date >=
as.Date("2009-06-01") & top5ThreadWout252620$Date <= as.Date("2009-09-30"),]

#take the period where it is not summer break on 2009

notsummerBreakPeriod2009  = top5ThreadWout252620[top5ThreadWout252620$Date >=
as.Date("2009-01-05") & top5ThreadWout252620$Date <= as.Date("2009-05-31"),]


#perform a t-test in 2009 on summer break vs not summer break period

t.test(summerBreakPeriod2009$leisure,notsummerBreakPeriod2009$leisure,
"greater",conf.level = 0.95)



                                        #Analysing anonymous


#GRAPH

dev.new()
```

```
#Plot the anger and date and the color to be negemo level

qplot(top6ThreadwMostPostsFullData$Date, top6ThreadwMostPostsFullData$anger, size =
as.factor(round_any(top6ThreadwMostPostsFullData$negemo, 1)), color =
as.factor(round_any(top6ThreadwMostPostsFullData$negemo, 1))) +labs(x="Year", y="Anger")

#get the posts where it is made by anon

top6ThreadwMostPostsFullDataAnon =
top6ThreadwMostPostsFullData[top6ThreadwMostPostsFullData$AuthorID==-1,]

#get the posts without anonymous and also without thread 252620

top5ThreadwMostPostsFullDataWOutAnonWOut252620 =
top5ThreadWout252620[top5ThreadWout252620$AuthorID!=-1,]

#get the data of anonymous in top 6 threads without thread 252620

anonFreqWOut252620 =
top6ThreadwMostPostsFullDataAnon[top6ThreadwMostPostsFullDataAnon$ThreadID!=252620
,]


#frequency of anonymous posts in the top 6 threads without thread 252620

freqAnonTop6ThreadsWout252620 = as.data.frame(as.table(by(anonFreqWOut252620,
year(anonFreqWOut252620$Date), function(df) NROW(df$PostID) )))


#GRAPH

dev.new()

ggplot(freqAnonTop6ThreadsWout252620,
aes(x=freqAnonTop6ThreadsWout252620$year.anonFreqWOut252620.Date.,
y=freqAnonTop6ThreadsWout252620$Freq, group = 1)) + geom_point() + geom_line() +
labs(x="Year", y="Post Count")


#get the data of the top 6 threads between 2005 to end of 2006 including anonymous

anonIn2002To2006 = top5ThreadWout252620[top5ThreadWout252620$Date >=
as.Date("2002-01-01") & top5ThreadWout252620$Date <= as.Date("2006-12-31"),]
```

```
#get the data of the top 6 threads between 2005 to end of 2006 including anonymous

NonAnonIn2002To2006 = top5ThreadwMostPostsFullDataWOutAnonWOut252620[top5ThreadwMostPostsFullDataWOutAnonWOut252620$Date >= as.Date("2002-01-01") & top5ThreadwMostPostsFullDataWOutAnonWOut252620$Date <= as.Date("2006-12-31"),]


#get the data of the top 6 threads between 2007 to end of 2011 including anonymous

anonIn2007To2011 = top5ThreadWout252620[top5ThreadWout252620$Date >= as.Date("2007-01-01") & top5ThreadWout252620$Date <= as.Date("2011-12-31"),]

#get the data of the top 6 threads between 2007 to end of 2011 including anonymous

NonanonIn2007To2011 = top5ThreadwMostPostsFullDataWOutAnonWOut252620[top5ThreadwMostPostsFullDataWOutAnonWOut252620$Date >= as.Date("2007-01-01") & top5ThreadwMostPostsFullDataWOutAnonWOut252620$Date <= as.Date("2011-12-31"),]


#Performing t.tests.

t.test(anonIn2002To2006$negemo, NonAnonIn2002To2006$negemo, "less", conf.level = 0.95)

t.test(anonIn2002To2006$anger, NonAnonIn2002To2006$anger, "less", conf.level = 0.95)


t.test(anonIn2007To2011$anger, NonanonIn2007To2011$anger, "greater", conf.level = 0.95)

t.test(anonIn2007To2011$negemo, NonanonIn2007To2011$negemo, "greater", conf.level = 0.95)
```