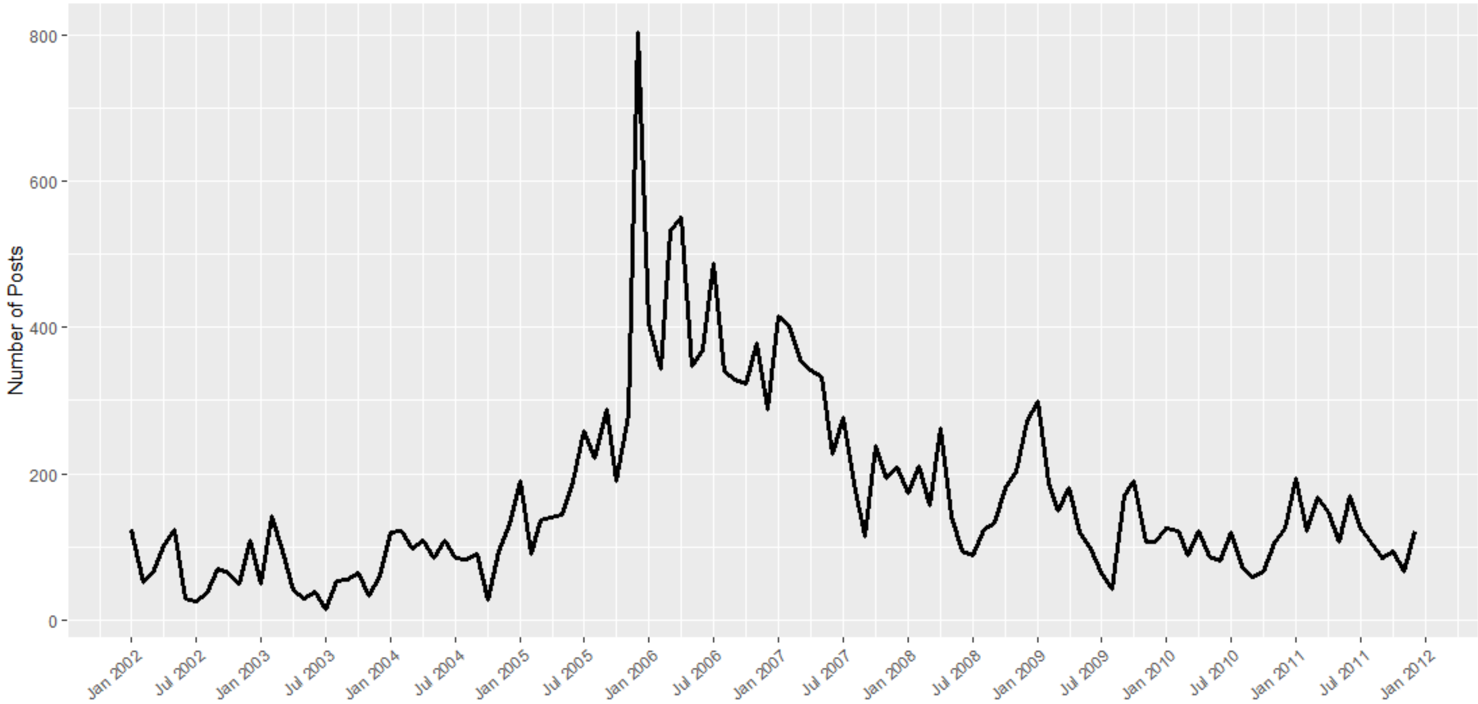
Question a1

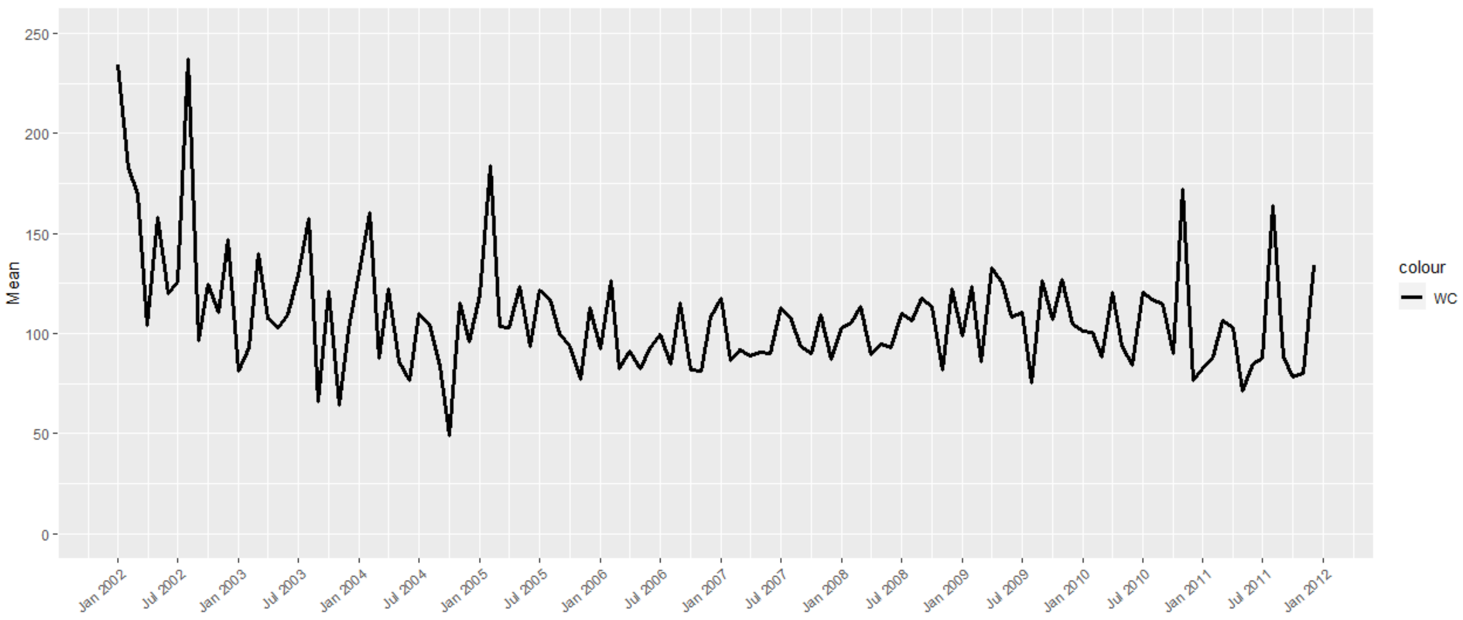
The activity of participants in the forum is analysed by grouping the data according to month and year and then counting the frequency of posts during those time periods. Initially, we also needed to convert the Date column in our dataframe into a date type. This produces the graph shown below.

Clear trends can be seen in the graph over time. For each year individually, we can see a gradual decline from January to November and then a significant increase in the number of posts around December which can likely be attributed to Christmas. Over the several years, we can see gradual increase in the number of posts from 2002 to 2005 after which there is a very large increase around the end of 2005. After this, we see a gradual decline over the years which can likely be attributed to the forum going out of popularity.

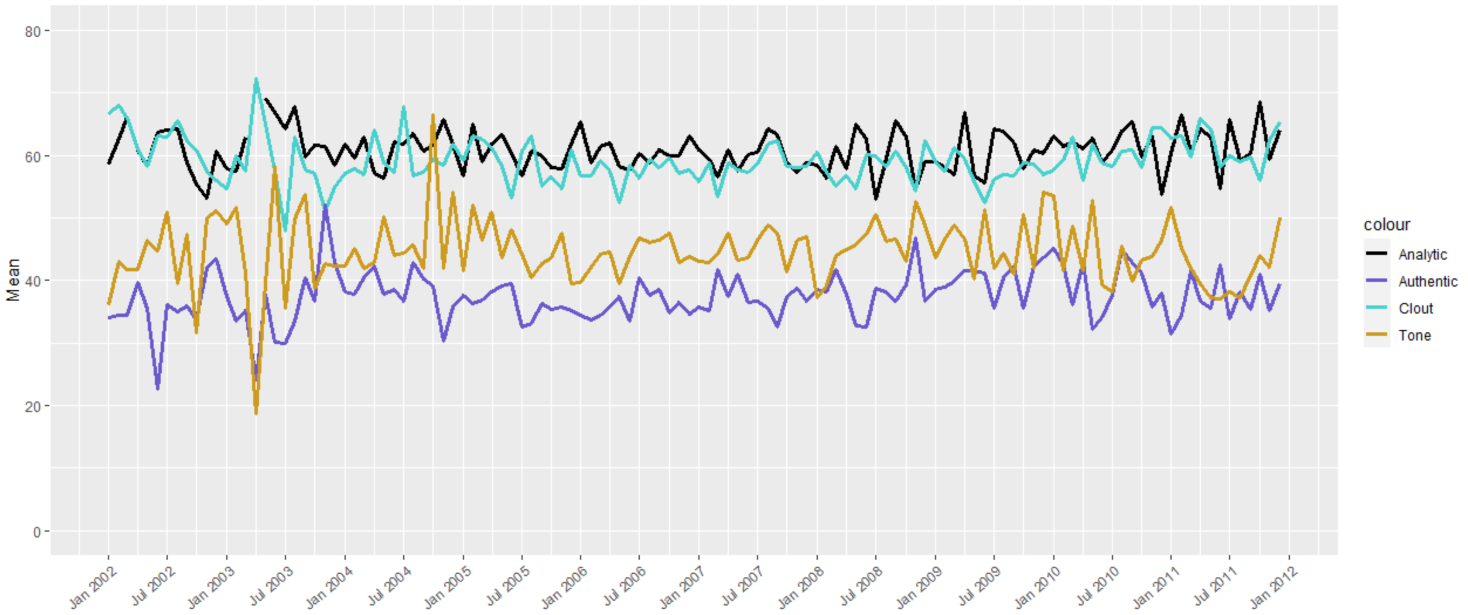


Question a2

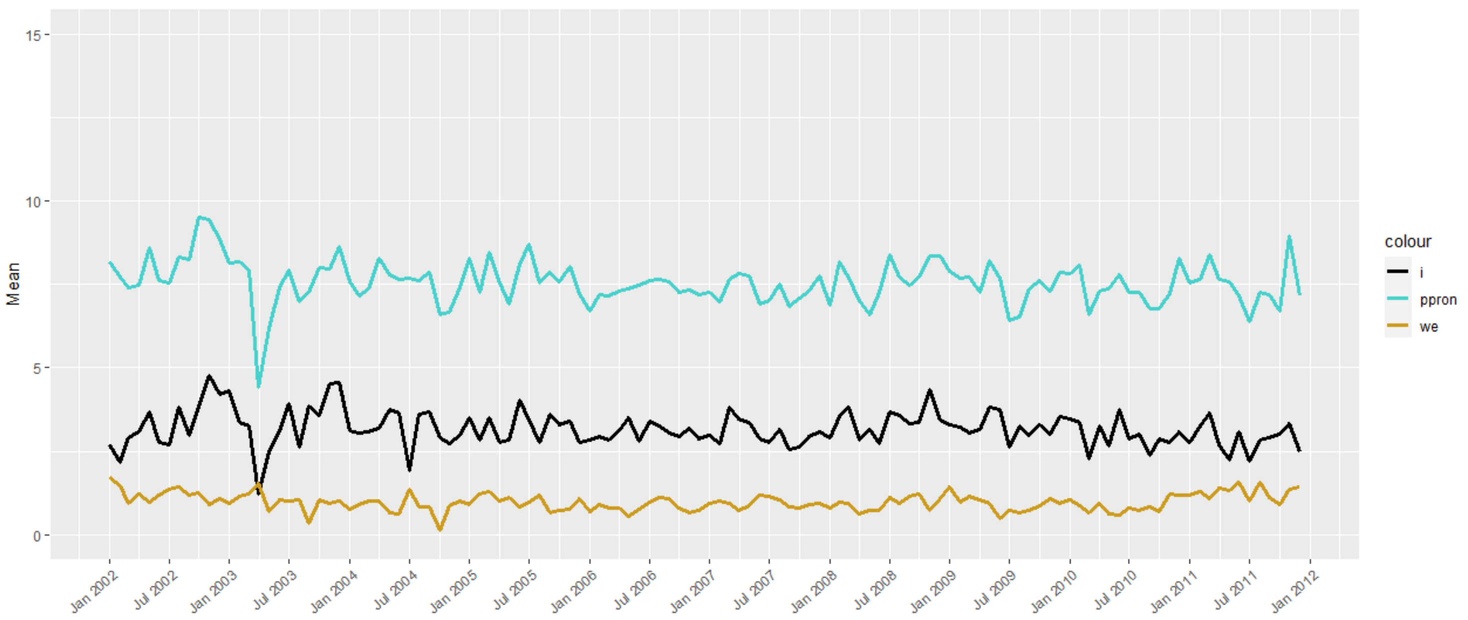
We can see a clear yearly pattern in the mean word count where we observe an alternating pattern where the mean word count increases one month and then falls the next. This repeats until we see a very large increase around December which can ultimately be attributed to the Christmas holidays and people sharing longer messages. Over the several years, we also observe that the mean word count ultimately alternates around a baseline of 100 words over the several years.



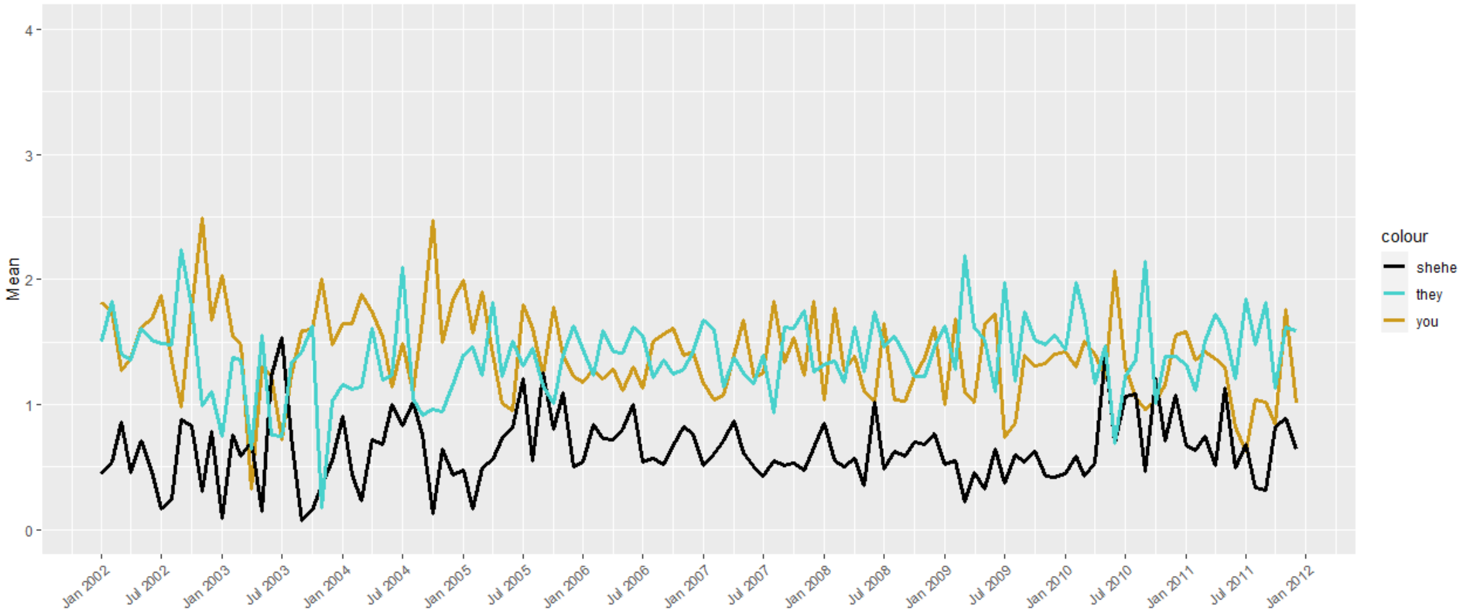
Here we compare four related linguistic variables; Analytic, Clout, Authentic and Tone. Firstly, we can see that over the several years, all the four linguistic variables closely follow each other and ultimately, when one variable increases, the rest follow. This behavior is expected as we would ultimately expect more authentic, analytic and tone-heavy posts to garner more clout.



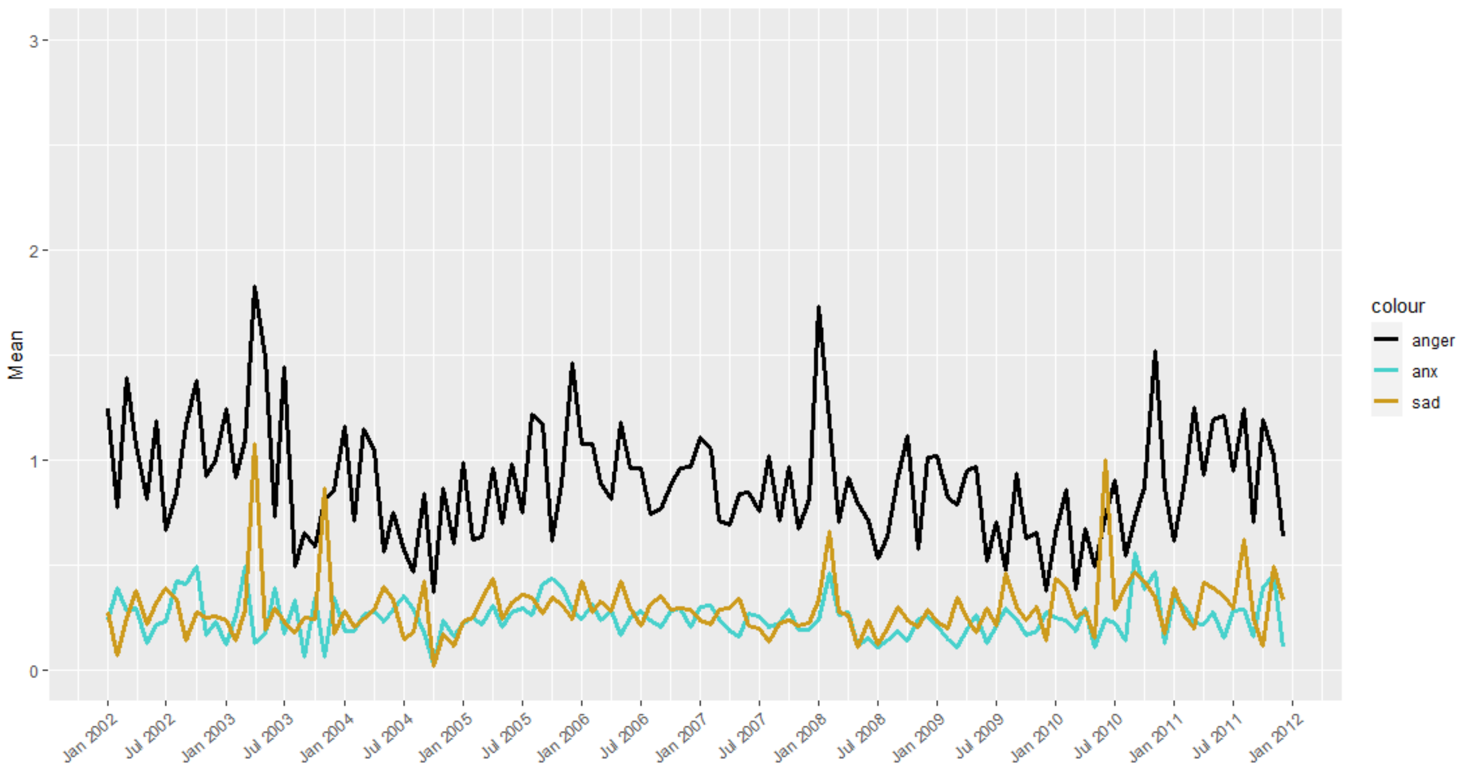
Here we compare the three related linguistic variables; ppron, i, and we. Firstly, we can see that over the several years, all the three linguistic variables closely follow each other and ultimately, when one variable increases, the rest follow. We see the variables i and ppron vary significantly over the years and have large peaks and troughs whereas the variable we has very minor changes and stays close to a baseline. One explanation for the very high relation between the we, ppron and i variables could be the fact that the these variables represent a similar subset of words that they represent.



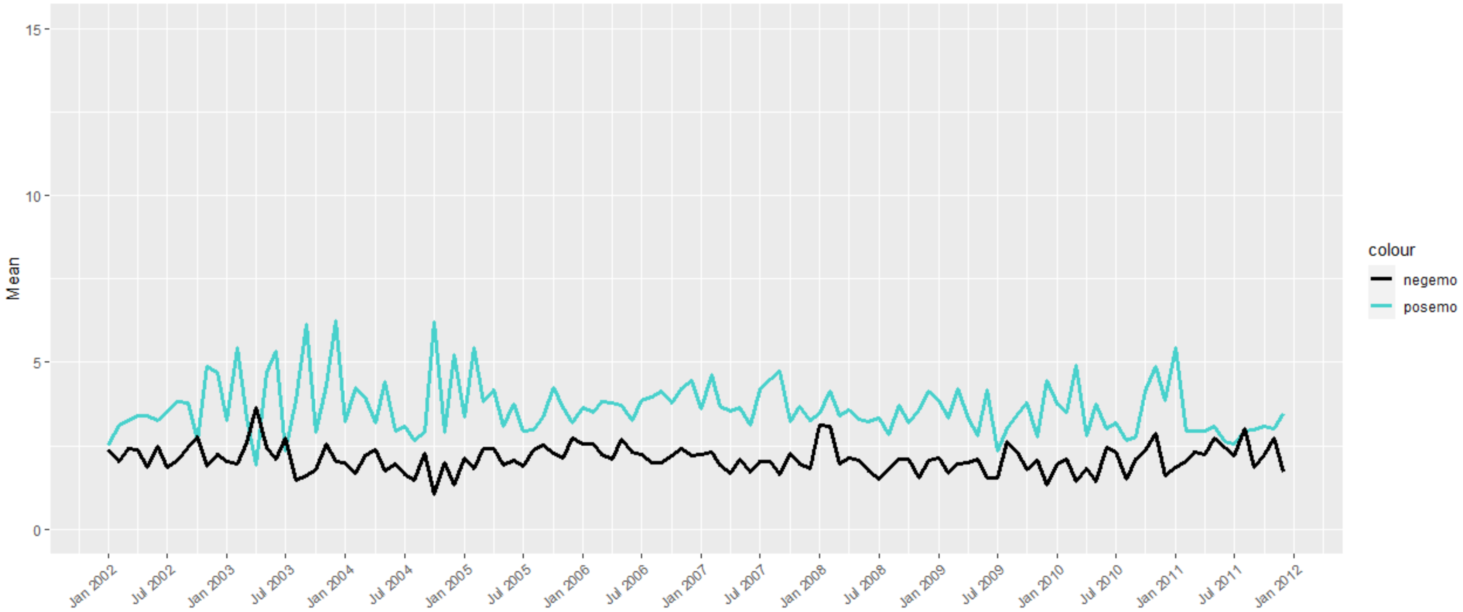
Here we compare three related linguistic variables; shehe, they and you. Firstly, we can see that over the several years, all the three linguistic variables are closely related to each other but share different relationships. Firstly, we can see that the shehe variable and the you variable have an inverse relationship where if one variable falls, the other increases. This relationship is also seen between the you and they variables where they share an inverse relationship. Finally, the shehe and they variables seem to have a positive relationship where an increase in one variables is followed by an increase in the other variable. This behavior is expected as we would ultimately expect posts talking about other people to contain more "shehe" and "they" words whereas we would expect conversation between two authors to contain more "you" words.



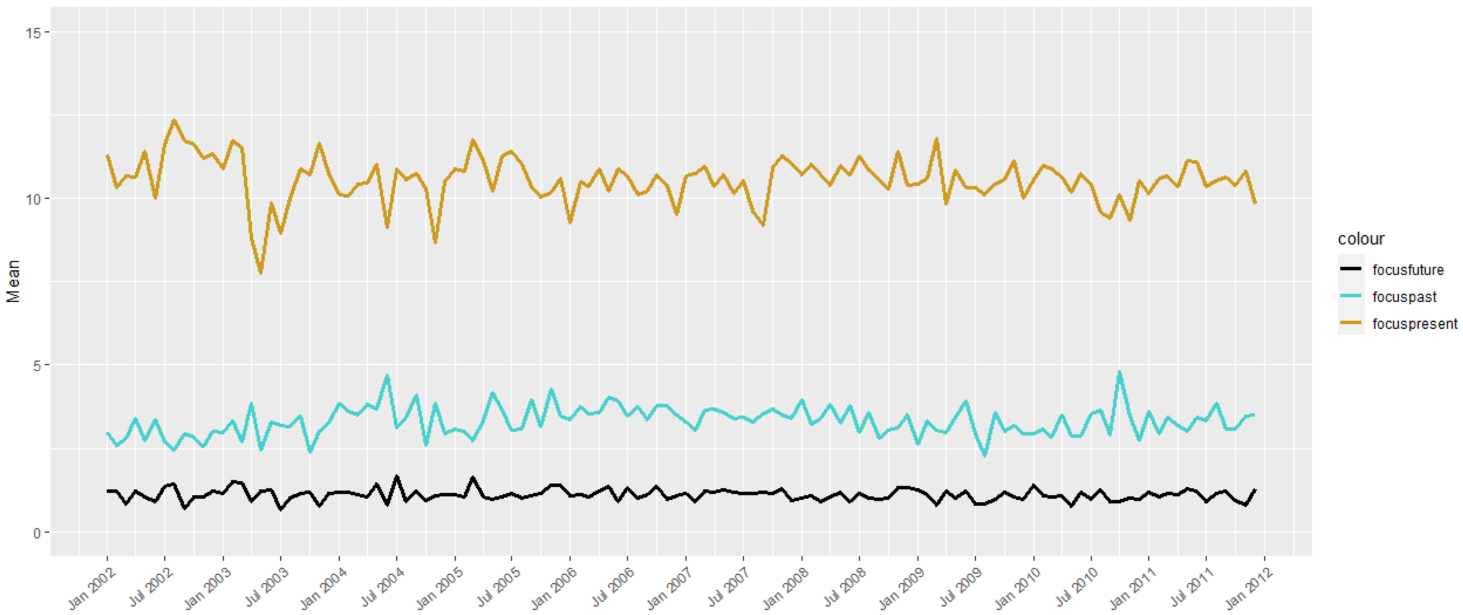
Here we compare the three related linguistic variables; anger, anx, and sad. Firstly, we can see that over the several years, the two variables have a strong relationship where an increase in one variable is usually followed by an increase in the variables. This behavior is expected as we would ultimately expect more posts with anger when there are more posts with sadness and anxiousness.



Here we compare the two related linguistic variables; negemo and posemo. Firstly, we can see that over the several years, the two variables have an inverse relationship where an increase in one variable is usually followed by a decrease in the variable. This behavior is expected as we would ultimately expect more negative emotion posts when there are less positive emotion posts and vice versa.



Here we compare three related linguistic variables; focusfuture, focuspast and focuspresent. Firstly, we can see that over the several years, all the two variables of focusfuture and focuspast maintain a steady baseline and ultimately vary insignificantly. We can see large peaks and troughs in the focuspresent variable throughout the years but it ultimately maintains a baseline as well. We can also see that the focuspresent variable has an inverse relationship with the focusppast and focusfuture variables where a peak in the focuspresent variable usually results in a trough in the other two variables. This behavior is expected as we would ultimately expect posts that have a higher focus on the present to have a lower focus on the future and past.



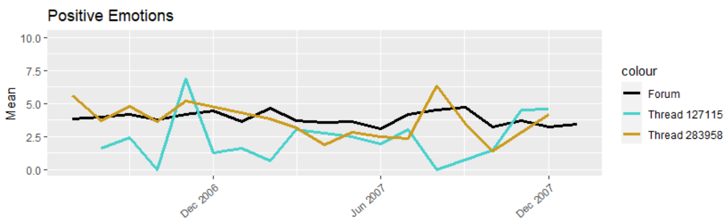
Question b1

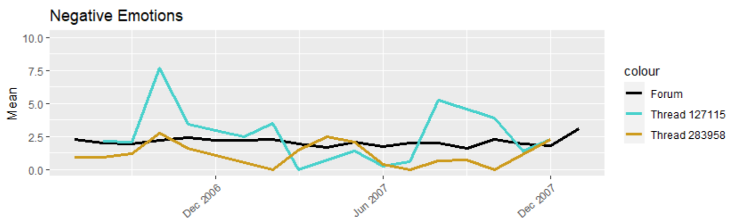
Thread 1 = thread 127115

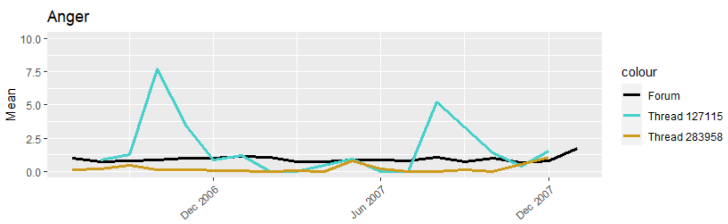
Thread 2 = thread 283958

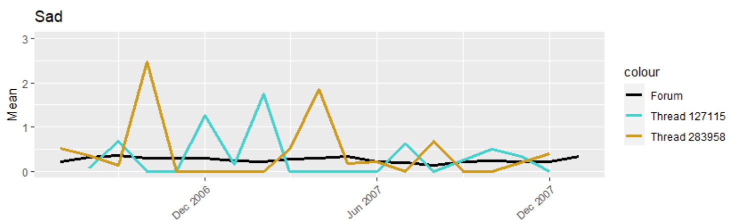
I used the variables of posemo, negemo, anger, sad and anx to determine the happiness and optimism of threads as these variables are mostly related to the happiness of a thread. Firstly, I picked out two threads to compare against each other and the forum as a whole. I did this by getting the number of posts each thread had and then choosing the threads with the highest number of posts as they would yield the most accurate results. Then I selected a timeframe between July 2006 and January 2008 as this interval contained most of the data for the threads and it matched the criteria of the question.

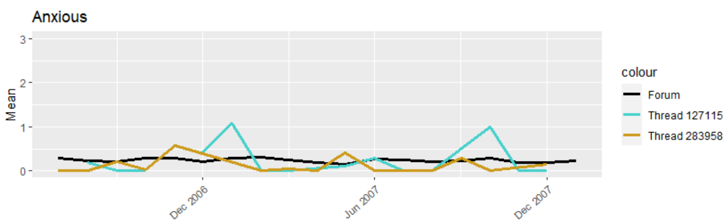
Thus, we can see in the plots that throughout the two year period both, Thread 1 and Thread 2, have slight lower positive emotions on average when compared to the forum. This pattern also holds for the anger, sad and anx variables, however, we can see that Thread 1 has higher negative emotions than the Forum while Thread 2 has lower negative emotions than the forum. Since negemo is the most distinctive variable, we also conduct a ttest to check if the negative emotions of Thread 1 are higher than the forum (p-value of 0.69) and if the negative emotions of Thread 1 are higher than Thread 2 (p-value of 0.019). A ttest to check if the negative emotions of Thread 2 are higher than the forum yields a p-value of 0.99. Ultimately, this means that Thread 1 very likely has a higher amount of negative emotions than Thread 2 but they both likely do not have a higher amount of negative emotions than the forum as a whole.











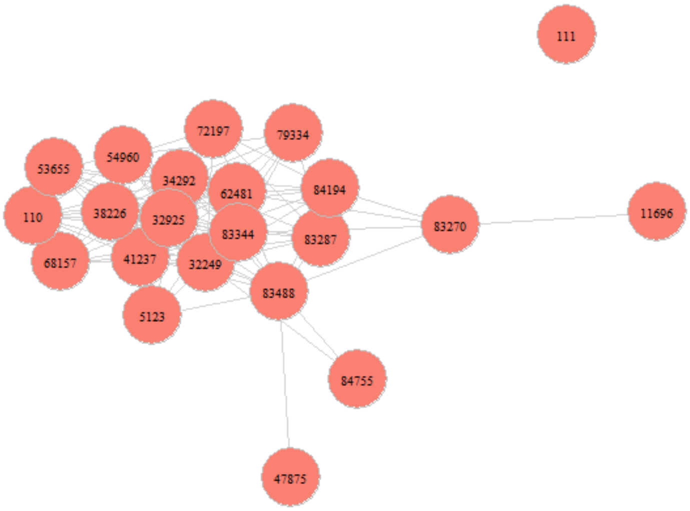

```
> t.test(thread_1$negemo, happy_means$negemo, alternative='greater')[3]
$p.value
[1] 0.6928686

> t.test(thread_2$negemo, happy_means$negemo, alternative='greater')[3]
$p.value
[1] 0.9992018

> t.test(thread_1$negemo, thread_2$negemo, alternative='greater')[3]
$p.value
[1] 0.01881687
```

Question c1

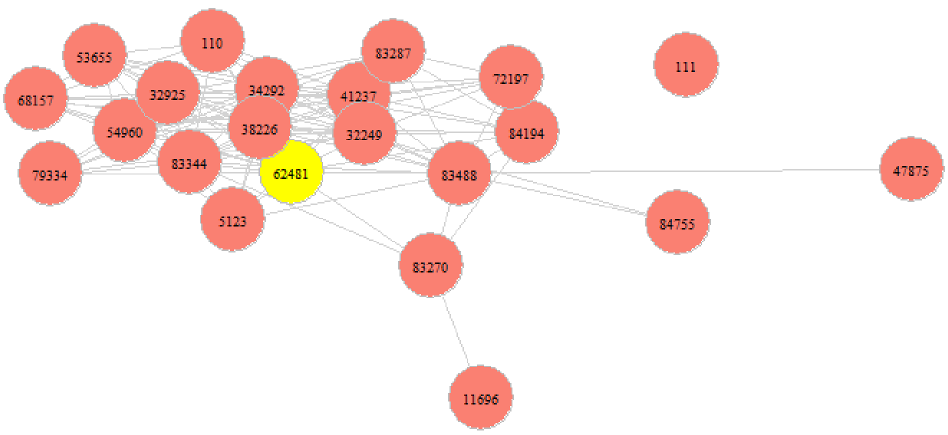
For this question, I first calculated the top 30 authors that made the most posts and then I calculated the most popular month for the posts made by these authors. Author -1 was excluded from these as they appear to be an invalid author (see investigation below). I then extracted this data from the webforum dataset and used this to create the social network of the authors. This results in the following social network. The creation of this social network was referenced form a website (RPubs, 2017).



Question c2

For this question, I used the betweenness, closeness and eigenvector centrality measures to determine that Author 62481 was the most important author in the social network as they had the highest betweenness score, the highest closeness score and the 10th highest eigenvector centrality score. Looking at the language used by this author, we can see that they used significantly more “they” (182% higher) and “we” (69% higher) words and also had significantly more anger (141% higher) and negative emotion (140% higher) in their posts when compared to the rest of the authors.

```
         WC    Analytic       Clout   Authentic        Tone       ppron
-13.1282677  -4.6305885  20.2282938 -43.2731604  -9.3240497  38.7662498
          i          we         you       shehe        they      posemo
-18.0642565  69.5346687 -26.8580329  39.3179330 182.5813362 -17.8990343
     negemo         anx       anger         sad   focuspast focuspresent
140.7351777 -38.7425904 141.2733000  42.7491231 -18.9882855    0.9564814
focusfuture
 23.9336998
```



-1 Author ID Investigation

There are no significant differences observed between posts by the regular authors and the posts by the Author ID equal to -1 which we can see in the calculation below which represents the percentage change of the variables between the mean and Author -1. Furthermore, we can confirm that the Author ID of -1 does not represent a unique individual as the number of posts with that Author ID is significantly higher (707) compared to the mean number of posts for all the other authors at 7.45 posts per author as seen below. Thus, we can infer that an AuthorID of -1 likely just represents a post that was later deleted by the author.

```
> diff
         WC     Analytic        Clout    Authentic         Tone        ppron            i           we          you
13.08523572  -2.27605566  -0.74792260   1.30748173   5.49375395   0.05058156  -1.28210531  -0.02873841  24.55356711
      shehe         they       posemo       negemo          anx        anger          sad    focuspast focuspresent
-6.33193865 -17.53312375  -6.74733216 -13.90597229 -12.00526774 -19.78169869  -7.29532439   9.71889966  -3.27438839
```

```
> #calculate average number of posts per author excluding author -1
> author_count = webforum %>% count(AuthorID)
> mean(author_count[-1, 2])
[1] 7.44616
> #get number of posts by author -1
> dim(author_neg_1)[1]
[1] 707
```

References

***RPubs - Bipartite/Two-Mode Networks in igraph*. (2017, October). Rpubs.com. Visited on 26 April 2022, at https://rpubs.com/pjmurphy/317838**

Code Appendix

Pre processing

```
#gathering data
rm(list = ls())
set.seed(31486347) # StudentID
webforum <- read.csv("webforum.csv")
webforum <- webforum [sample(nrow(webforum), 20000), ] # 20000 rows

#import all needed libraries
library(ggplot2)
library(dplyr)
library(igraph)
library(igraphdata)
library(gridExtra)

#Renaming row names
rownames(webforum) = NULL
```

Question a1

```
##QUESTION a1

#convert date column to Date type
webforum$Date = as.Date(webforum$Date, "%Y-%m-%d")

#add columns for month and year onto our dataset. Since converting to date reuqires the day, we just add the first day
webforum$yearmonth = format(webforum$Date, "%Y-%m-01")


#count number of posts for each group of month,Year
num_posts = webforum %>%
  group_by(yearmonth) %>%
  summarise(count_posts = n())

#convert yearmonth column to date data type
num_posts$yearmonth = as.Date(num_posts$yearmonth, "%Y-%m-%d")
View(num_posts)

#use ggplot2 to plot the timeseries data
graph = ggplot(num_posts, aes(x=yearmonth, y=count_posts)) +
  geom_line(size=1.1)+
  xlab("")+
  ylab("Number of Posts")+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=40, hjust=1))

graph
```

Question a2

```
##QUESTION a2
#calculate the mean for each linguistic variable for each year,month time period
linguistic_means = aggregate(webforum[, 5:23], list(webforum$yearmonth), mean)

names(linguistic_means)[1] = "yearmonth"

#convert yearmonth column to date data type
linguistic_means$yearmonth = as.Date(linguistic_means$yearmonth, "%Y-%m-%d")

#use ggplot2 to plot the timeseries data

#WordCount
graph = ggplot(linguistic_means, aes(x=yearmonth)) +
  geom_line(aes(y = WC, color = "WC"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ylim(0, 250)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black"))
graph

#Analytic, Clout, Authentic, Tone
graph = ggplot(linguistic_means, aes(x=yearmonth)) +
  geom_line(aes(y = Analytic, color = "Analytic"), size=1.1)+
  geom_line(aes(y = Clout, color = "Clout"), size=1.1)+
  geom_line(aes(y = Authentic, color = "Authentic"), size=1.1)+
  geom_line(aes(y = Tone, color = "Tone"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ylim(0, 80)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black", "slateblue3", "mediumturquoise", "goldenrod3"))

graph

#ppron, i, we
graph = ggplot(linguistic_means, aes(x=yearmonth)) +
  geom_line(aes(y = ppron, color = "ppron"), size=1.1)+
  geom_line(aes(y = i, color = "i"), size=1.1)+
  geom_line(aes(y = we, color = "we"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ylim(0, 15)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black", "mediumturquoise", "goldenrod3"))

graph

#you, shehe, they
graph = ggplot(linguistic_means, aes(x=yearmonth)) +
  geom_line(aes(y = you, color = "you"), size=1.1)+
  geom_line(aes(y = shehe, color = "shehe"), size=1.1)+
  geom_line(aes(y = they, color = "they"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ylim(0, 4)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black", "mediumturquoise", "goldenrod3"))

graph
```

```
#posemo, negeemo
graph = ggplot(linguistic_means, aes(x=yearmonth)) +
  geom_line(aes(y = posemo, color = "posemo"), size=1.1)+
  geom_line(aes(y = negemo, color = "negemo"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ylim(0, 15)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black","mediumturquoise"))

graph


#anx, anger, sad
graph = ggplot(linguistic_means, aes(x=yearmonth)) +
  geom_line(aes(y = anx, color = "anx"), size=1.1)+
  geom_line(aes(y = anger, color = "anger"), size=1.1)+
  geom_line(aes(y = sad, color = "sad"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ylim(0, 3)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black","mediumturquoise", "goldenrod3"))

graph



#focuspast, focuspresent, focusfuture
graph = ggplot(linguistic_means, aes(x=yearmonth)) +
  geom_line(aes(y = focuspast, color = "focuspast"), size=1.1)+
  geom_line(aes(y = focuspresent, color = "focuspresent"), size=1.1)+
  geom_line(aes(y = focusfuture, color = "focusfuture"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ylim(0, 15)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y")+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black","mediumturquoise", "goldenrod3"))

graph
```

Question b1

```
#QUESTION b1
#calculate number of posts for each thread and choose top 2 threads to compare against the whole forum
thread_count = webforum %>% count(ThreadID)
thread_count <- thread_count[order(-thread_count$n),]
head(thread_count)


#get dataframe containing mean posemo, negemo, anger, sad, anx for the whole forum
happy_means = aggregate(webforum[, 16:20], list(webforum$yearmonth), mean)

names(happy_means)[1] = "yearmonth"

#convert yearmonth column to date data type
happy_means$yearmonth = as.Date(happy_means$yearmonth, "%Y-%m-%d")
View(happy_means)

#calculate the same statistics for thread 127115 and thread 283958
#as these threads have the most posts over a large date range so they are more likely to be a better representation
#of the threads

thread_1 = webforum %>% filter(ThreadID == "127115")
thread_1 = aggregate(thread_1[, 16:20], list(thread_1$yearmonth), mean)
names(thread_1)[1] = "yearmonth"
thread_1$yearmonth = as.Date(thread_1$yearmonth, "%Y-%m-%d")
View(thread_1)

thread_2 = webforum %>% filter(ThreadID == "283958")
thread_2 = aggregate(thread_2[, 16:20], list(thread_2$yearmonth), mean)
names(thread_2)[1] = "yearmonth"
thread_2$yearmonth = as.Date(thread_2$yearmonth, "%Y-%m-%d")
View(thread_2)

#plot the variables relating to happiness for the two threads and the forum mean

#posemo
graph1 = ggplot() +
  geom_line(data=happy_means, aes(x=yearmonth, y=posemo, color = "Forum"), size=1.1)+
  geom_line(data=thread_1, aes(x=yearmonth, y=posemo, color = "Thread 127115"), size=1.1)+
  geom_line(data=thread_2, aes(x=yearmonth, y=posemo, color = "Thread 283958"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ggtitle("Positive Emotions")+
  ylim(0, 10)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y", limits=as.Date(c('2006-07-01','2008-01-01')))+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black","mediumturquoise", "goldenrod3"))
graph1

#negemo
graph2 = ggplot() +
  geom_line(data=happy_means, aes(x=yearmonth, y=negemo, color = "Forum"), size=1.1)+
  geom_line(data=thread_1, aes(x=yearmonth, y=negemo, color = "Thread 127115"), size=1.1)+
  geom_line(data=thread_2, aes(x=yearmonth, y=negemo, color = "Thread 283958"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ggtitle("Negative Emotions")+
  ylim(0, 10)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y", limits=as.Date(c('2006-07-01','2008-01-01')))+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black","mediumturquoise", "goldenrod3"))
graph2
```

```
#anger
graph3 = ggplot() +
  geom_line(data=happy_means, aes(x=yearmonth, y=anger, color = "Forum"), size=1.1)+
  geom_line(data=thread_1, aes(x=yearmonth, y=anger, color = "Thread 127115"), size=1.1)+
  geom_line(data=thread_2, aes(x=yearmonth, y=anger, color = "Thread 283958"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ggtitle("Anger")+
  ylim(0, 10)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y", limits=as.Date(c('2006-07-01','2008-01-01')))+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black","mediumturquoise", "goldenrod3"))
graph3

#sad
graph4 = ggplot() +
  geom_line(data=happy_means, aes(x=yearmonth, y=sad, color = "Forum"), size=1.1)+
  geom_line(data=thread_1, aes(x=yearmonth, y=sad, color = "Thread 127115"), size=1.1)+
  geom_line(data=thread_2, aes(x=yearmonth, y=sad, color = "Thread 283958"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ggtitle("Sad")+
  ylim(0, 3)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y", limits=as.Date(c('2006-07-01','2008-01-01')))+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black","mediumturquoise", "goldenrod3"))
graph4

#anxiety
graph5 = ggplot() +
  geom_line(data=happy_means, aes(x=yearmonth, y=anx, color = "Forum"), size=1.1)+
  geom_line(data=thread_1, aes(x=yearmonth, y=anx, color = "Thread 127115"), size=1.1)+
  geom_line(data=thread_2, aes(x=yearmonth, y=anx, color = "Thread 283958"), size=1.1)+
  xlab("")+
  ylab("Mean")+
  ggtitle("Anxious")+
  ylim(0, 3)+
  scale_x_date(date_break = "6 month", date_labels = "%b %Y", limits=as.Date(c('2006-07-01','2008-01-01')))+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  scale_color_manual(values=c("black","mediumturquoise", "goldenrod3"))
graph5




grid.arrange(graph1, graph2, graph3, graph4, graph5)



#using ttest() to determine if the threads have more positive emotions than the forum
t.test(thread_1$negemo, happy_means$negemo, alternative='greater')[3]
t.test(thread_2$negemo, happy_means$negemo, alternative='greater')[3]
t.test(thread_1$negemo, thread_2$negemo, alternative='greater')[3]
```

Question c1

```
#QUESTION c1
#get 30 authors that have the most posts, skip author id -1
author_top30 = webforum %>% count(AuthorID)
author_top30 = author_top30[order(-author_top30$n),]
author_top30 = author_top30[2:31, ]$AuthorID


#filter data to contain top30 authors during the most active month
network_data = webforum %>%
  filter(AuthorID %in% author_top30)

#get most active month period
toptime = network_data %>%
  count(yearmonth)

toptime = toptime[order(-toptime$n), ]
toptime = toptime[1, 1]


network_data = network_data %>%
  filter(yearmonth == toptime)

network_data = network_data[, 1:2]

#date being used = 2005-12-01


#creating network of authors
#this graph creation was referenced from https://rpubs.com/pjmurphy/317838
g = graph.data.frame(network_data,directed = FALSE)
V(g)$type <- bipartite_mapping(g)$type

bipartite_matrix <- as_incidence_matrix(g)

#Calculate AuthorID adjacency matrix
author_network <- t(bipartite_matrix) %*% bipartite_matrix
diag(author_network) <- 0

#plot network graph
author_network <- graph_from_adjacency_matrix(author_network,
                              mode = "undirected",
                              weighted = TRUE)


#customise network
V(author_network)$color <- "salmon"
V(author_network)$shape <- "circle"
E(author_network)$color <- "lightgray"
V(author_network)$label.color <- "black"
V(author_network)$label.cex <- 0.6
V(author_network)$frame.color <-  "gray"
V(author_network)$size <- 15

plot(author_network, layout = layout_with_graphopt)
```

Question c2

```
#QUESTION c2
#Identifying most important author based on closeness, betweenness
head(closeness(author_network)[order(closeness(author_network), decreasing = T)], 5)

head(betweenness(author_network)[order(betweenness(author_network), decreasing = T)], 5)

head(evcent(author_network)$vector[order(evcent(author_network)$vector, decreasing = T)], 10)

#most important author is AuthorID 62481
V(author_network)
V(author_network)[5]$color = "yellow"
plot(author_network, layout = layout_with_graphopt)

#getting language used by other others and our important autho
other_authors = webforum %>%
  filter(AuthorID %in% author_top30)

other_authors = other_authors %>%
  filter(yearmonth == toptime)



imp_author = other_authors %>%
  filter(AuthorID == '62481')

other_authors = other_authors %>%
  filter(AuthorID != '62481')



#calculate the mean for each linguistic variable for each year,month time period
imp_author = colMeans(imp_author[, 5:23])
other_authors = colMeans(other_authors[, 5:23])

(imp_author-other_authors)/other_authors * 100
```

Author ID -1 Investigation

```
#INVESTIGATING AUTHOR ID -1
author_neg_1 = webforum %>% filter(AuthorID == -1)
View(author_neg_1)

#comparing means for different categories
diff = (colMeans(author_neg_1[5:22]) - colMeans(webforum[5:22])) / colMeans(webforum[5:22]) * 100
diff

#comparing average number of posts vs author -1

#calculate average number of posts per author excluding author -1
author_count = webforum %>% count(AuthorID)
mean(author_count[-1, 2])

#get number of posts by author -1
dim(author_neg_1)[1]
```