A. Analyse activity and language on the forum over time. Some starting points:

A.1 Describe your data: How active are participants, and are there periods where this increases or decreases? Is there a trend over time?

Here, I calculated the number of threadID in each year to demonstrate the level of activity.

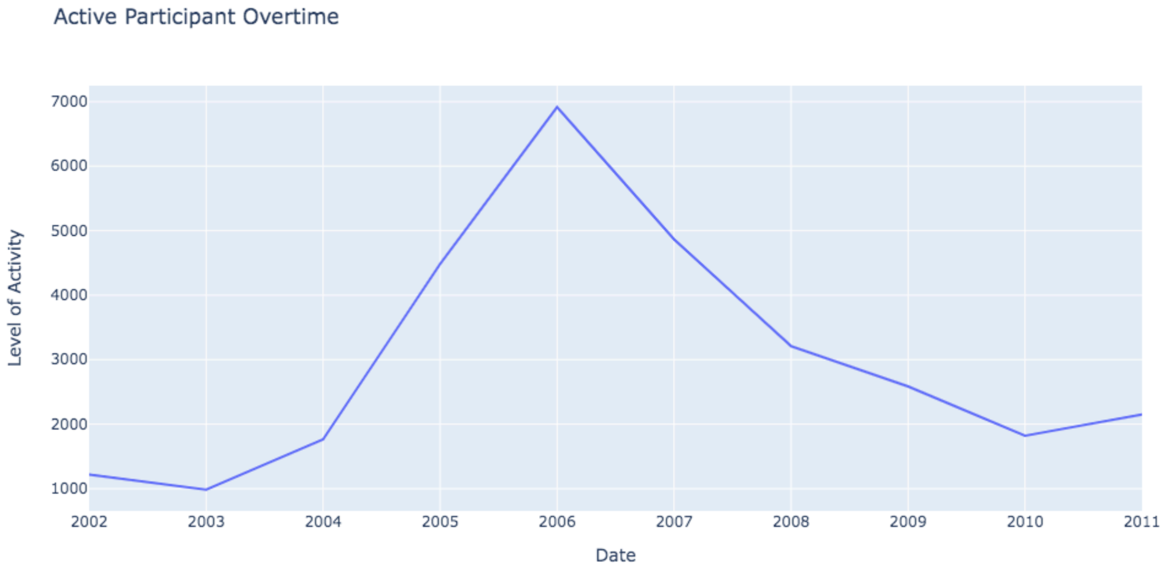


Figure 1: Level of active participants throughout the years. The level of activity is measured by the number of threads within each year.

Here you can see that the level of activity increased from the year 2002 to 2006 and decreased after the year 2006 (**Figure 1**). The peak is the highest around late 2005 / early 2006.

To investigate this further we will need to look at the daily activity (**Figure 2**). At a glance, there was a very sharp peak around December 2005 compared to the rest of the year.

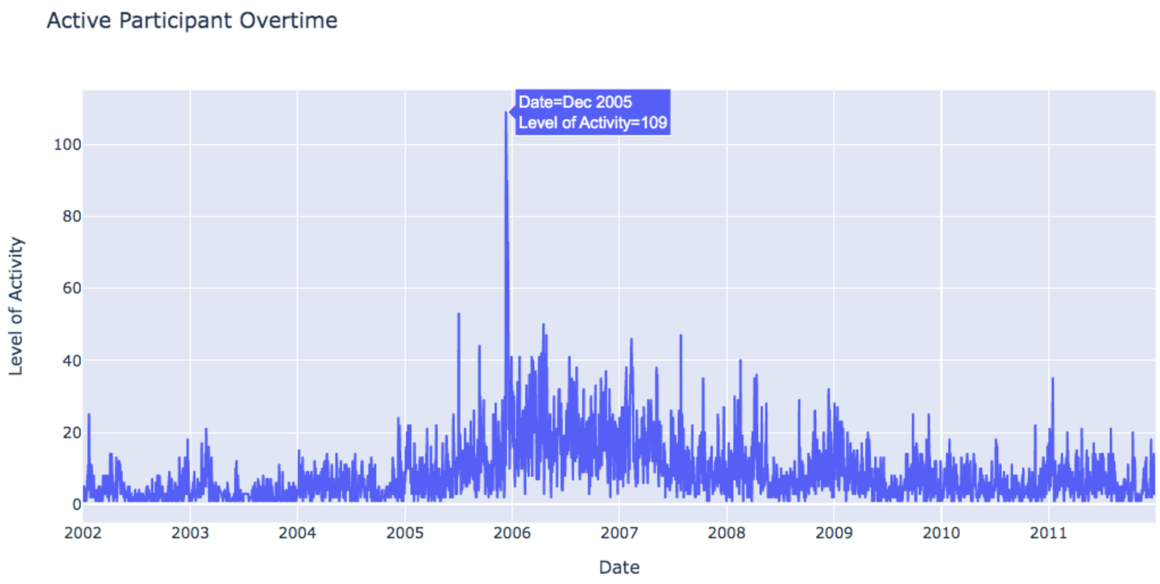


Figure 2: Level of active participants daily from 2002 to 2011. The level of activity is measured by the number of threads within each day.

When we zoom in further into december 2005 (**Figure 3**). We can see that the level of activity raised fairly quickly in December 11 and december 12 2005 was the highest peak throughout the years.

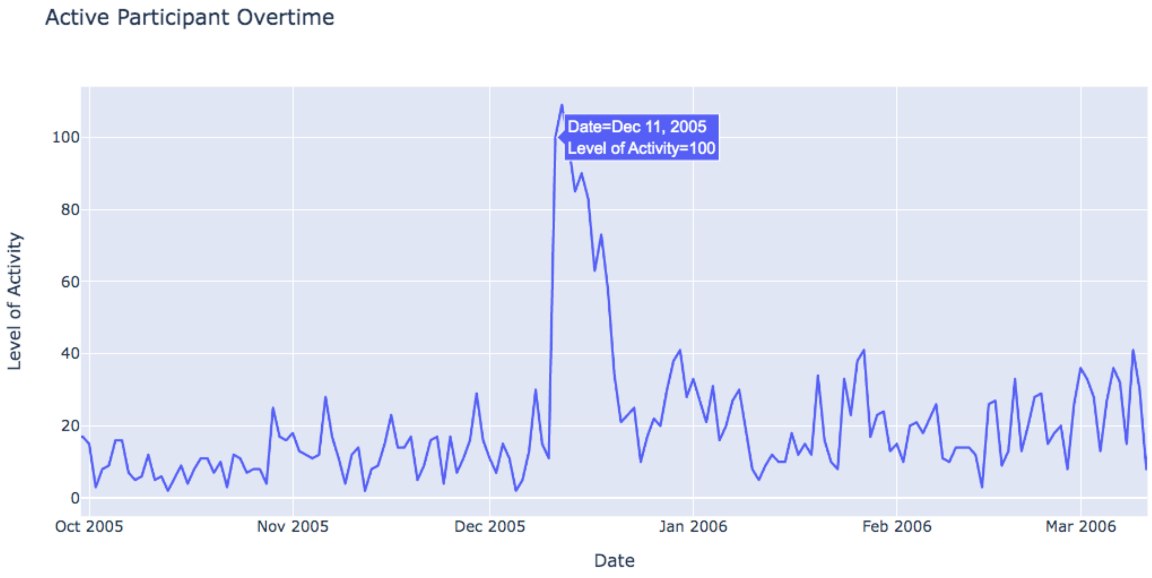


Figure 3: It is the same graph as figure 2. Here we zoomed in on December 2005.

By doing a little google search it turns out there were Cronulla riots happening during this time which could probably explain the spike during this time. Otherwise by looking at the daily line plot, it would have a constant pattern.

A.2 Looking at the linguistic variables, do these change over time? Is there a relationship between variables?

Based on the paper, linguistic variables are as follows, I will be focussing on these throughout the whole investigation:

- Analytic
- Clout
- Authentic
- Tone

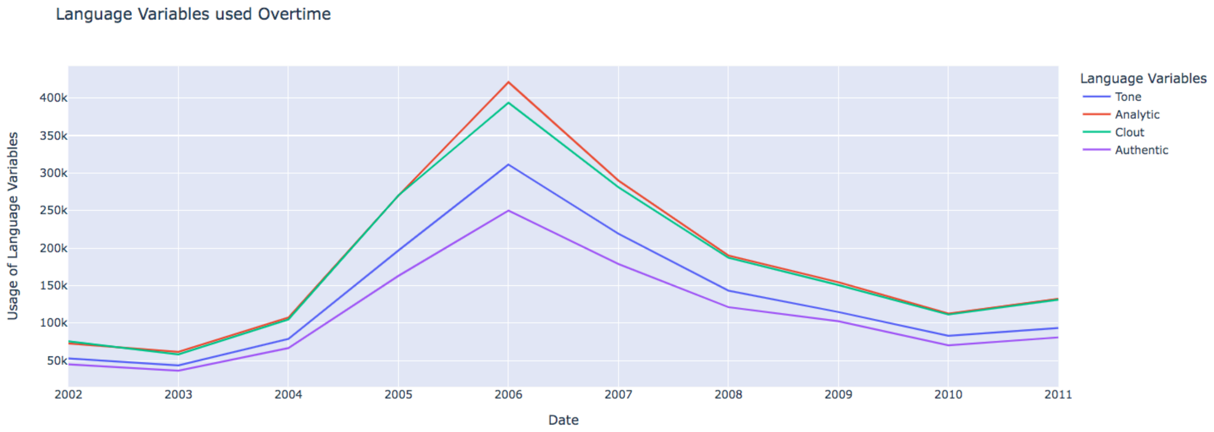


Figure 4: Usage of language variables (Tone, Analytic, Clout, Authentic) throughout the years.

For most of the time, the four language variables had a similar trend in usage over the years where it increased up to 2006 and decreased after 2006 (**Figure 4**).

It also appears that analytic and clout are commonly used together and they are the most dominant variables throughout the years. Authentic and emotional tones seem to be used together as they appear close together. However, further investigation is needed to confirm this.

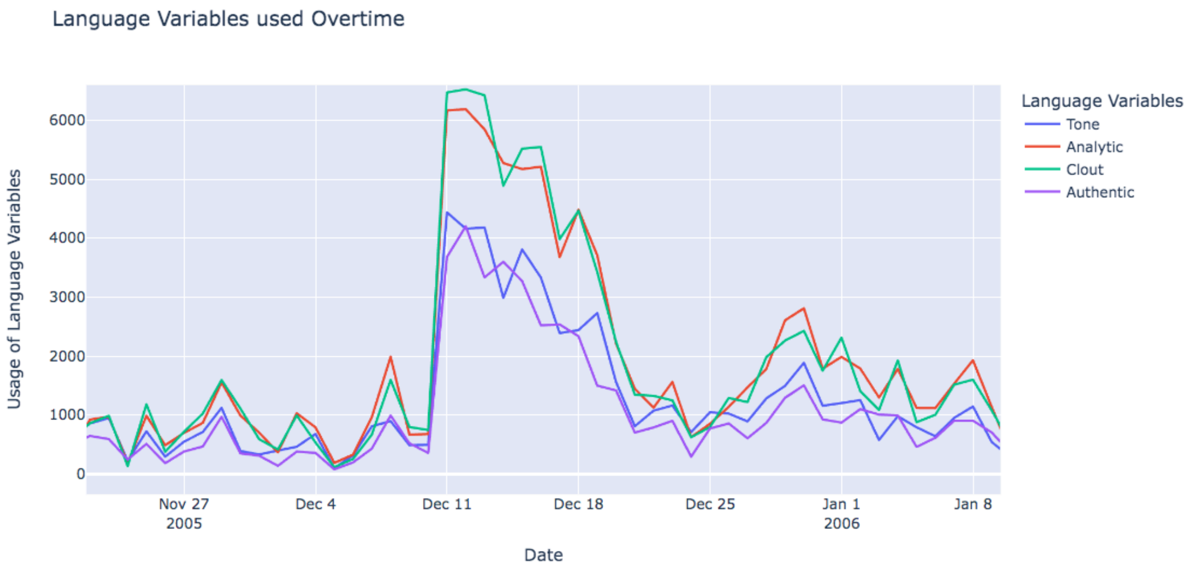


Figure 5: Usage of language variables on a daily basis. Here it is zoomed in on the month of December 2005.

Also, when we look at the usage of languages around december 2005 (**Figure 2005**). We can see that the dominant languages used by people during this time are mainly Clout and Analytical variables. Which is around the same time as the Cronulla riots which are demonstrations against ethinic violence. Which makes sense that you would expect a lot of analytical, powerful and impactful discussions during these times. The level of tone and authenticity also increased during this time which demonstrates that people were being empathetic to the ongoing riots and were genuine in their thoughts for the rough times.

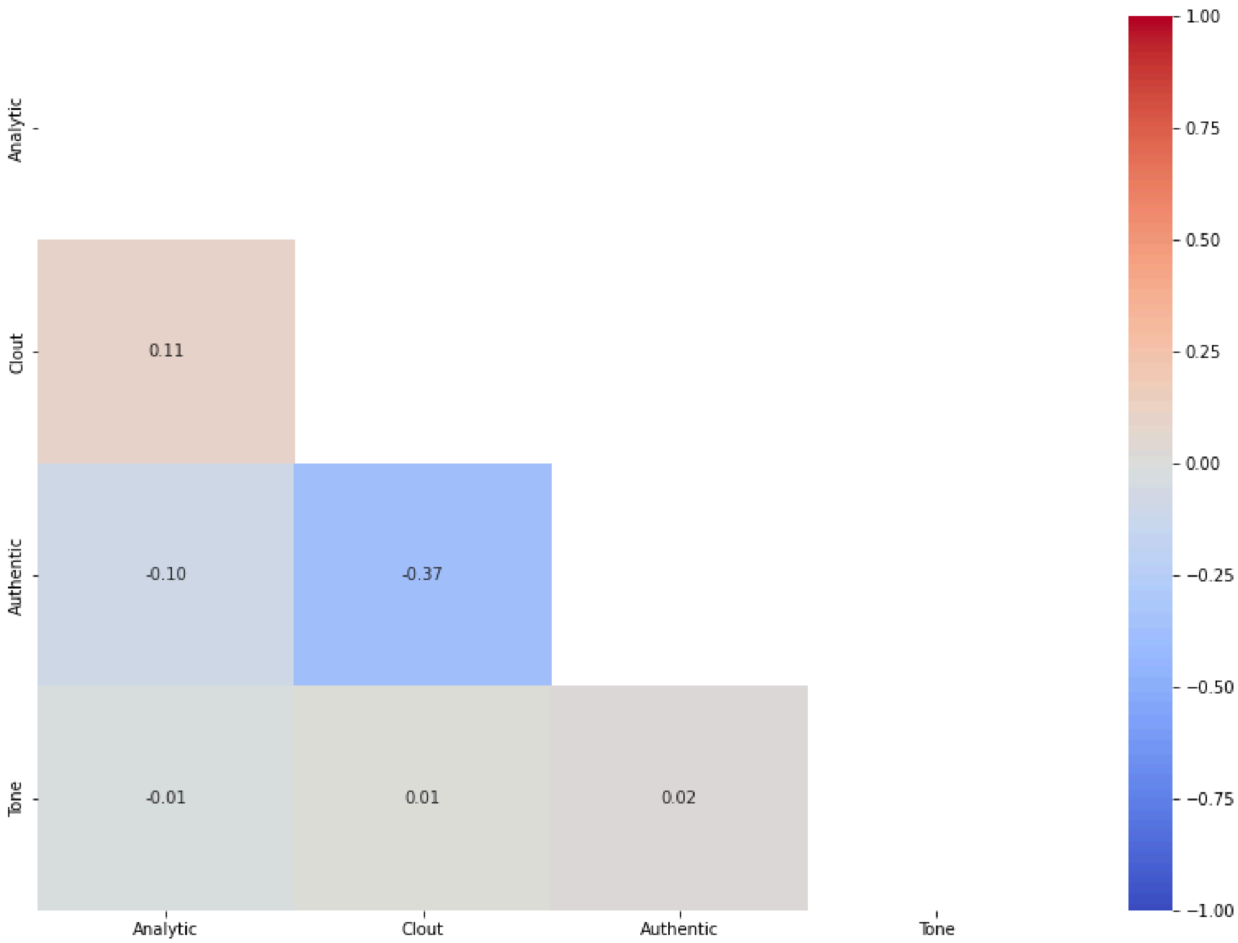


Figure 6: The results produced here is pearson correlation, I have tested out with kendall and spearman as well and they all produced similar results.

Based on the heatmap (**Figure 6**), There is a positive correlation between clout and analytical variable as observed in the line plot above (**Figure 4**). However, there is no correlation between tone and authentic variable as originally thought. Interestingly, there is a negative correlation between authenticity and clout variables. Would this mean that someone who often uses clout languages is less authentic with their words? Since clout carries a negative connotation, any other variables related to it are expected to be slightly negatively correlated.

b. Analyse the language used by groups. Some starting points:

B.1 Threads indicate groups of participants communicating on the same topic. Describe the threads present in your data.

Below is the distribution of participants in all threads, we can see that the majority of the threads would consist of a single person. The distribution is right skewed which is reflective of the box plot below (**Figure 8**).

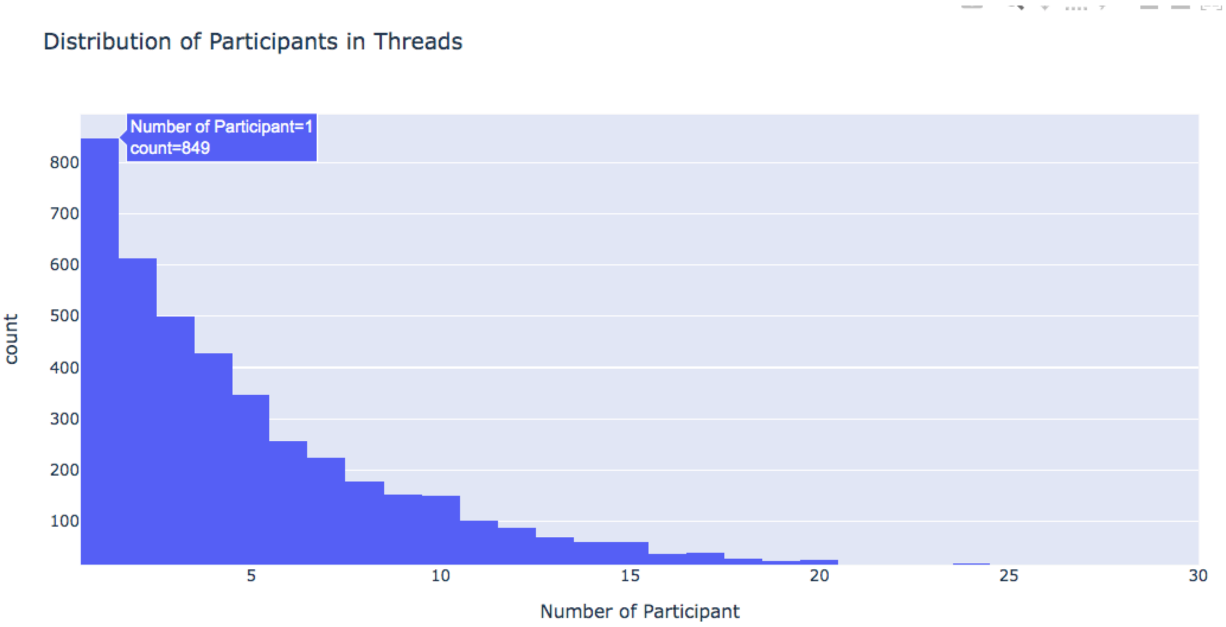


Figure 7: Distribution of participants in all threads. This was obtained by counting the number of participants in each thread and generating a histogram for it.

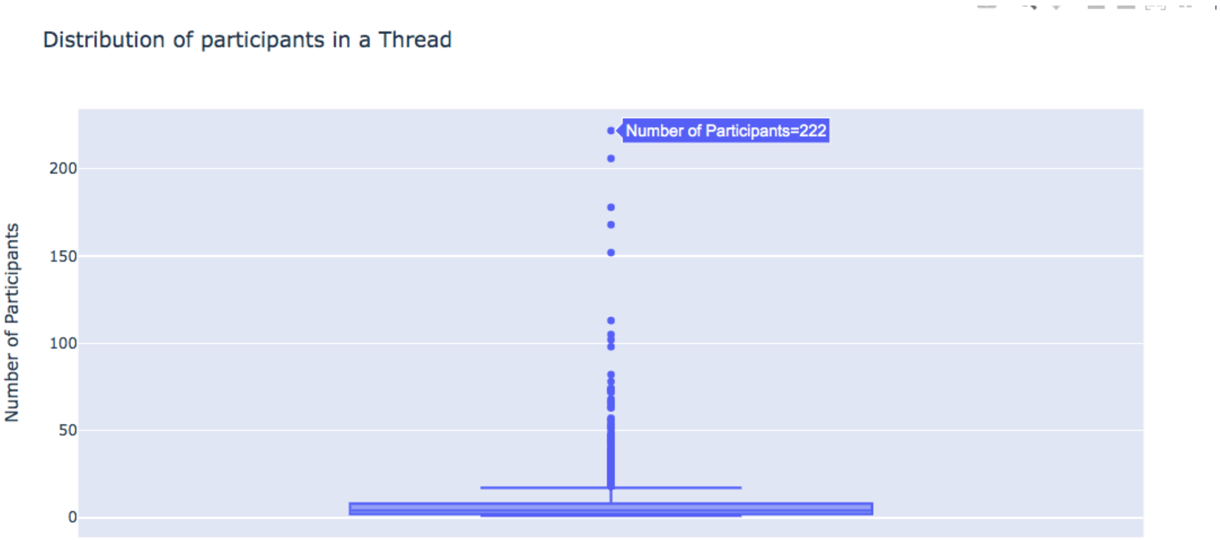


Figure 8: Above is the distribution of participants within a thread using a box plot.

As we can see that it is right skewed (**Figure 8**). If we zoom in as seen below. We can see that A single thread typically consists between 1 to 17 people. It rarely goes beyond 17 but it does happen. On average a single thread should consist of 4 people (**Figure 9**).

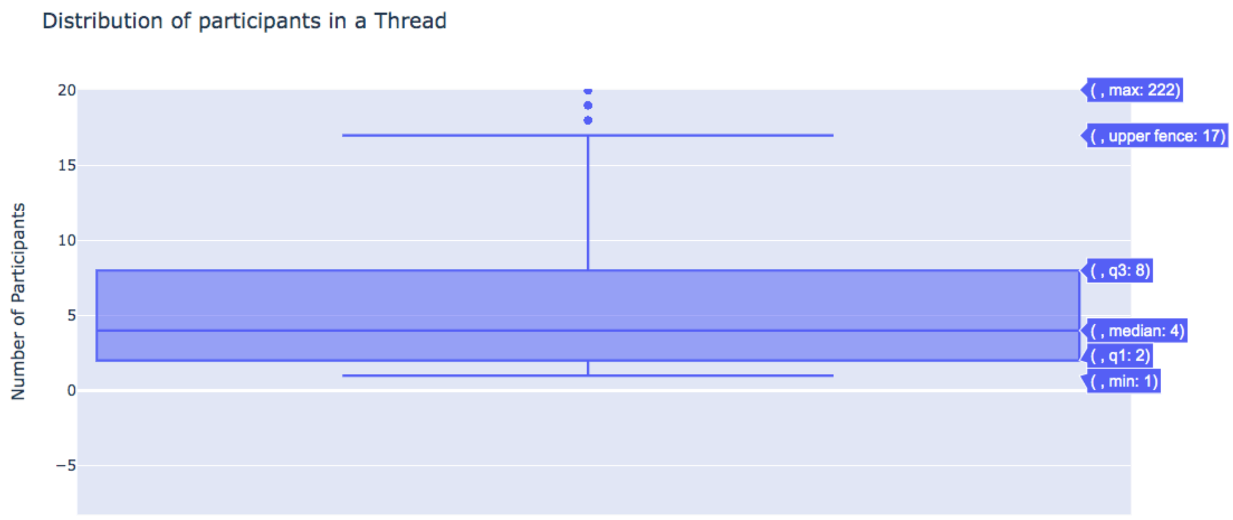
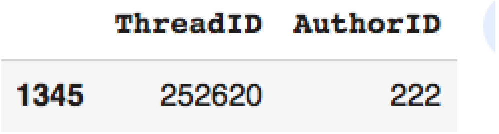


Figure 9: It is the same box plot as in **Figure 8**. Here we have zoomed into it further to have a better look at the box plot.

Back to the first picture of the box plot (**Figure 8**). We can see the highest number of participants involved in a single thread is 222 which rarely happens. When looked into it further it belongs to the threadID 252620.



	ThreadID	AuthorID
1345	252620	222

When looked further into this ID, it turns out the discussion was happening from 2005-12-07 to 2006-12-20 which lasted for about a year (**Figure 10**). We can see that the thread started a few days before the riot. we can see that it especially peaked around december 11 to 13 which is around the time the cronulla riot was happening and was a hot topic. That may have brought so many people together that it built a community which explains why a lot of people participated and it lasted for almost a year since the riot happened. Could have been a support group during the hard time.

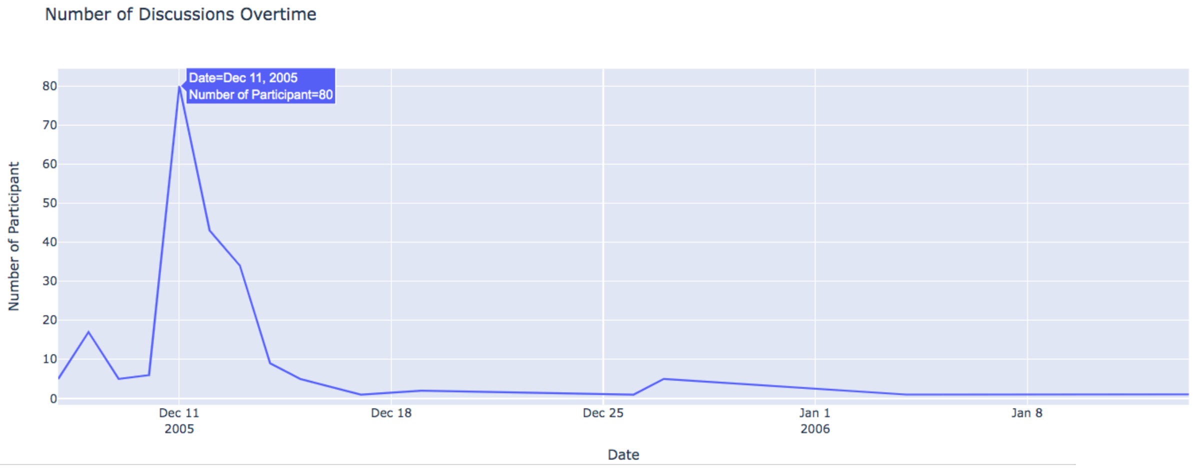


Figure 10: Number of discussion within the period of 2005-12-07 to 2006-12-20 where a single thread (252620) has the most number of participants discussing on a single topic.

B.2 By analysing the linguistic variables for all or some of the threads, is it possible to see a difference in the language used by different groups?

To answer this question, I had to look at the number of possible clusters in the dataset as it is possible that different threads are likely to use different language variables at different levels. Here I plotted k-means clustering ranging from 1 to 40 (**Figure 11**). It starts to plateau near 40 which indicates that there could be almost 40 different clusters. This means that there could be groups of 40 that uses the 4 language variables at different levels.

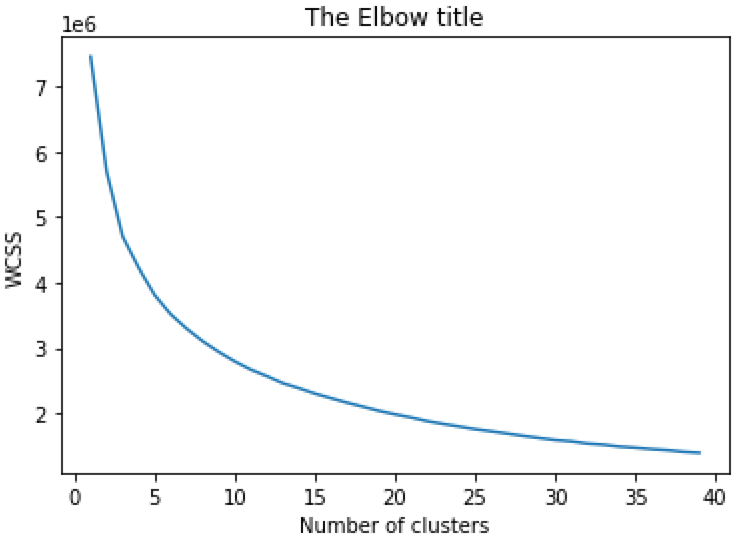


Figure 11: used Within Cluster Sum of Squares (WCSS) method to find the right number of clusters. WCSS helps to calculate the sum of squares distances of each data point from the centroid in which the process is iterated until we reach a minimum value for the sum distances.

We can further visualise these groups on a scatter plot. We can see that there are clear divisions between the groups which strengthens the idea that it is possible to see a difference in the language used by different groups. The black dots in the middle are the centroids of the group.

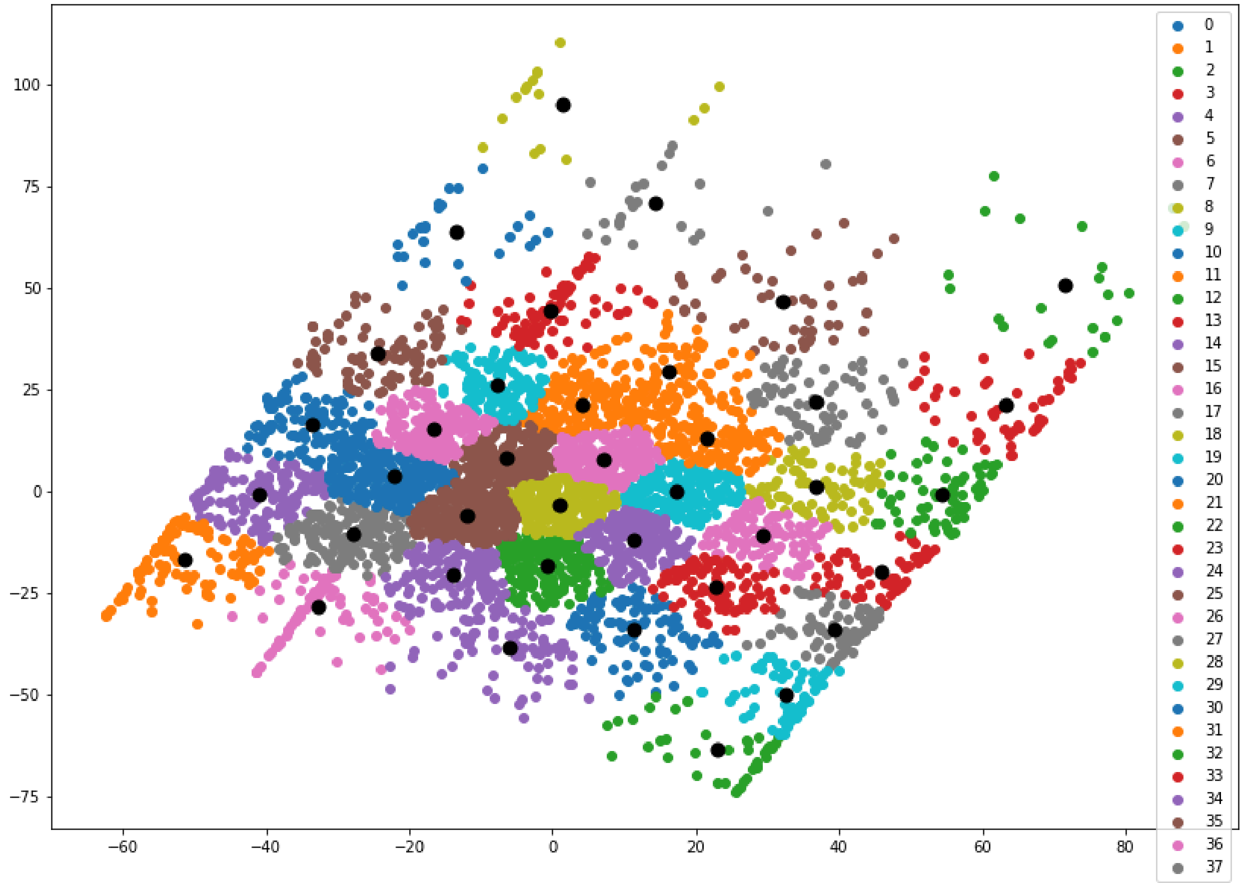


Figure 12: Here 38 groups of clusters were identified using language variables differently in each group.

B.3 Does the language used within threads (or between threads) change over time? How consistent or variable is the language used within threads?

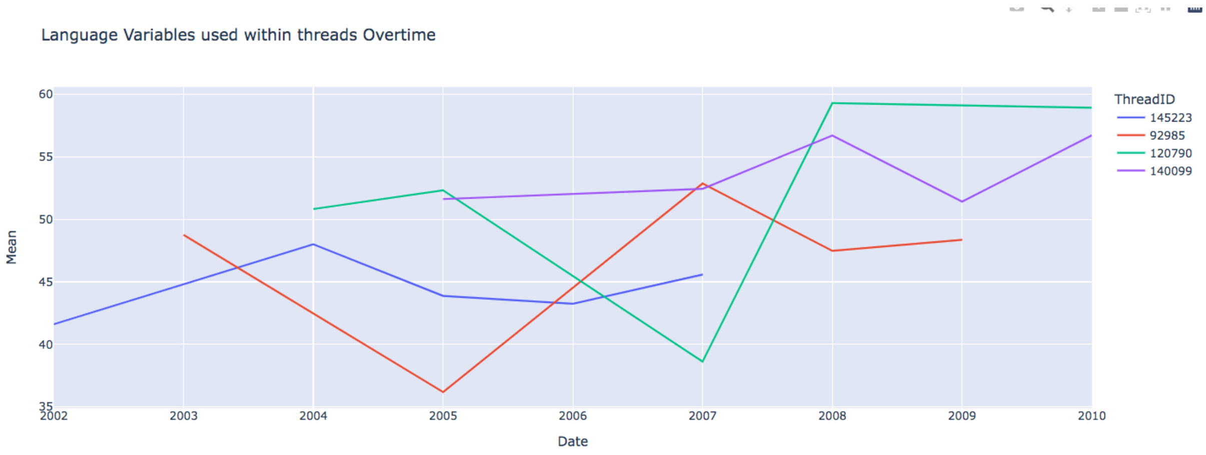


Figure 13: 5 threads were selected for this question and they were chosen because they have data for more than 5 years which allows us to see the changes of language over time. The mean was obtained by averaging the values of the 4 language variables together.

From the graph (**Figure 13**) we can see that the usage of language in each thread does change overtime and they are different from one another.

c. Challenge: Social networks online. We can think of participants posting to the same thread at similar times (for example during the same month) as forming a social network. When these participants also post to other threads over the same period, their social network extends.

- Can you define, graph and describe the social network that exists at a particular point in time, for example over one month? How does this change in the following months?
- Note: you only need to analyse a small portion of the social network over a short time period

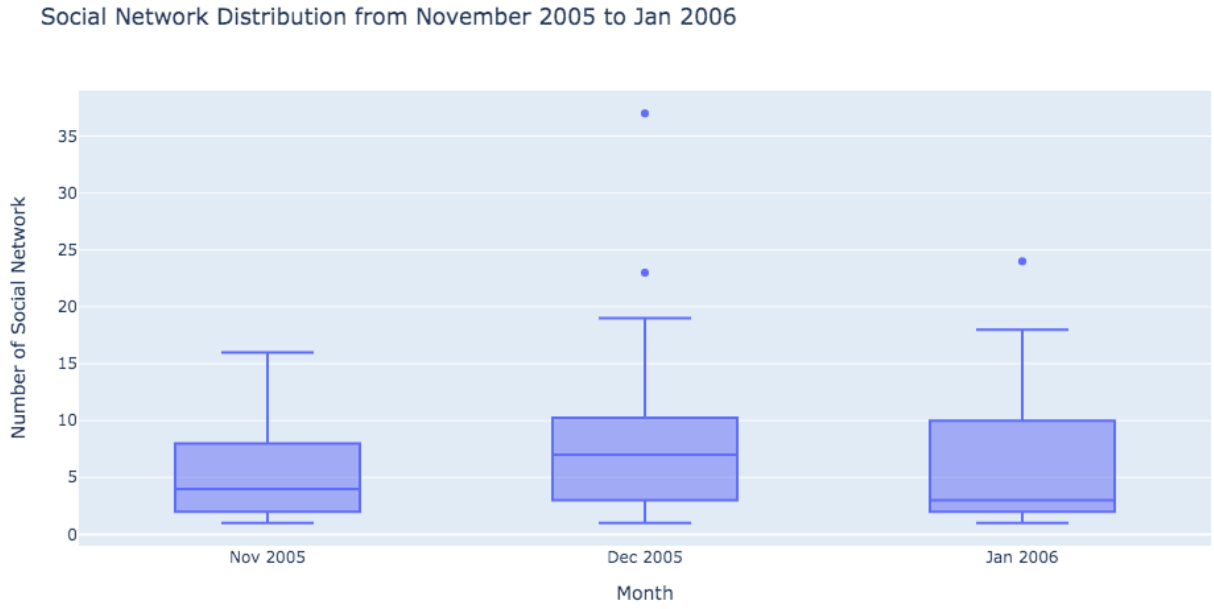


Figure 14: Number of social network (number of threadID) each individual is involved in is plotted by month from November 2005 to January 2006. This is to see the overall distribution of individuals that are involved in any social network from 2005 to January 2006.

Here we can see the distribution of the number of social groups individuals are involved in for these 3 months (**Figure 14**). You can see that the social network of individuals tend to increase from November to December 2005 and then declined in January 2006.

The rise in december could be due to the riot identified earlier on in december 11 and many support groups may have formed as result of that.

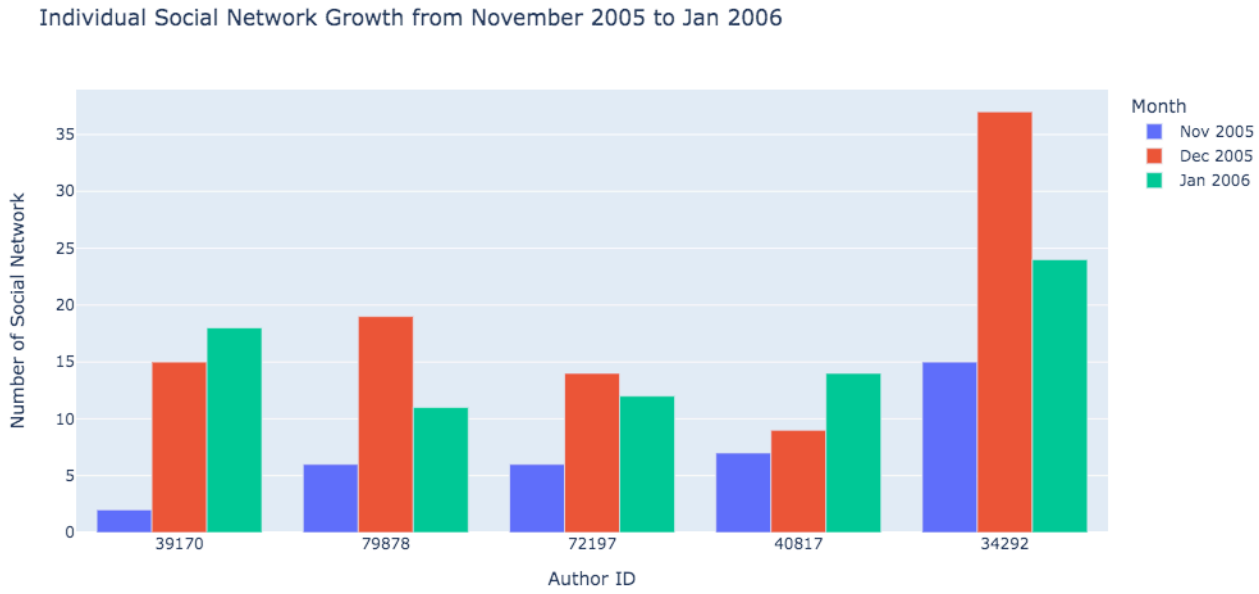


Figure 15: Looking at the changes in the number of social networks from the period (November 2005 to January 2006) in 5 individuals.

Here, 5 individuals were chosen (**Figure 15**). We looked into the changes in their social network from Nov 2005 to Jan 2006. Majority of the time the social network of these individuals increased from november to december. This is probably due to the riot caused in december 2005 which could have probably made people more active and join many social groups to discuss the particular topic. After December the number of social network declines, which could probably be due to the riot, have died down and they have moved on from the topic.

d. Reflection on your investigation. What did you first investigate? How did you then modify your research based on the results of your first investigation?

Initially, I was looking at the trends of active people which had shown a peak around late 2005 and early 2006. When zooming in further, the peak was a result of the spike around December 11 2005. This then probed me to want to know what happened during that date, which turned out to be a time where the cronulla riot was happening which was a big event of the time. When looking into the language variables at the time, the analytical thinking and clout variable was dominant at the time which is reflective of the riot that was ongoing. Because of this riot, there were 222 people participating in a single thread, which probably was having a discussion on the event that was happening since there was a peak in discussion of these groups of people during that time. It seems to may have affected the distribution to the number of networks at that time as well as seen in the **figure 14**, there was an increase in the number of social networks in december.

For question B initially was going to use the mean to see the changes in language variables usage overtime but the data doesn't make sense as to when using sum. Which is why sum was used instead as it is found to reflect the graph in **figure 1-3**.

Majority of the graph was created using plotly for interactivity but unfortunately I can't put it into docs but you can access them through this link which also contains the script that I have written for this analysis.