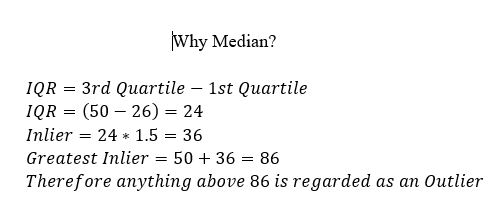
**FIT3152 Assignment 1**

The purpose of this assignment is to investigate whether participants in an online forum who are communicating directly via threads use similar language and whether the language used changes over time. The data is based on linguistic analysis of huge numbers of threads and posts and was conducted using Linguistic Inquiry and Word Count (LIWC).

**Sub setting the data**

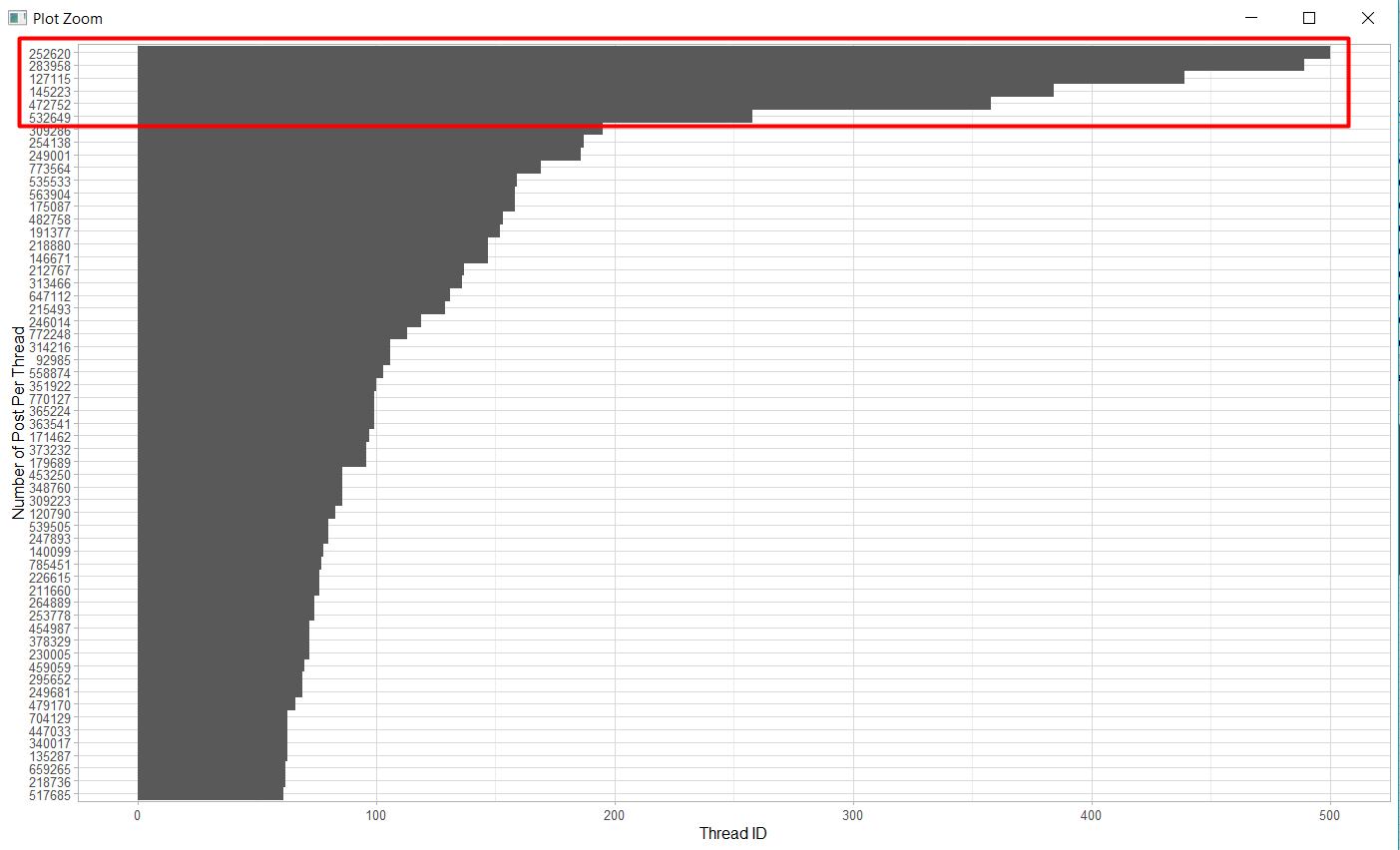
The first thing that we did is to first analyse the data as a whole and try to find out where can we make the data smaller. We then noticed the column in the data called ‘WC’ is the word count of each post. Some of the post has 0 word count, meaning the post probably contains images or diagrams. So we eliminates these posts that contains 0 word count.

Next we group the posts according to their threadID and then count the number of posts in each thread. In doing this we also take into account the number of authors in each thread (We count the number of authors in each thread and take the average of it. Here we choose median as average because there exist outliers. Then we eliminate the threads with number of authors lower than the median number of authors).



\*Statistics revision notes FIT1006 slide 83/410

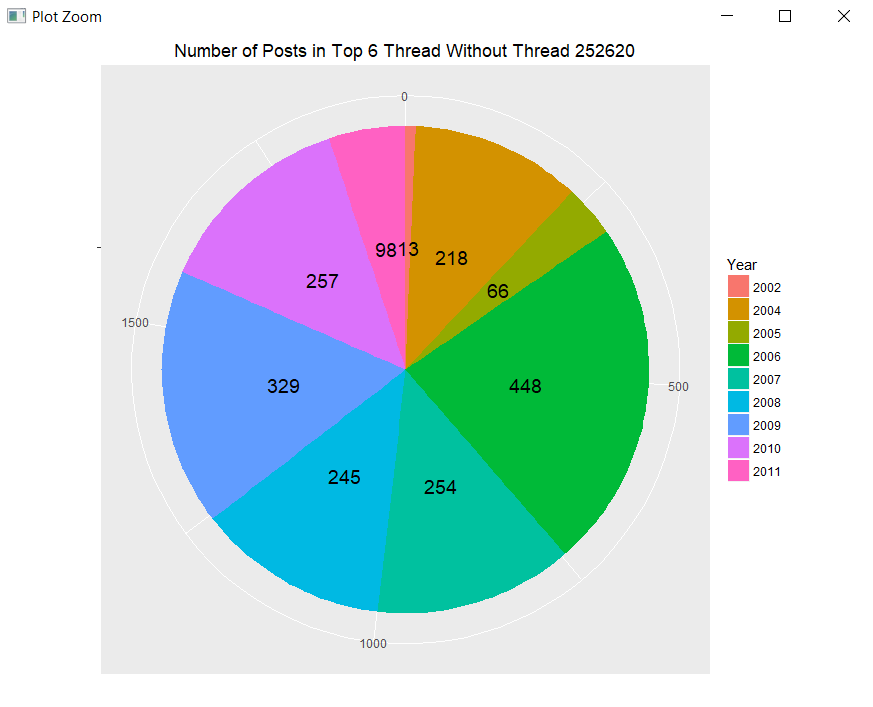
This is to make sure that there are several different authors in each thread and not only 2 people that keeps replying to each other. Next, we plot the number of post for every thread in the forum and we got the following graph:



Which gives us thread 252620, 283958, 127115, 145223, 472752, and 532649 as the top 6 threads with the most post (after thread 532649, the number of post is below 200 and we choose 200 as the threshold). We then take the 6 threads mentioned above and use it for this assignment.

**Analyzing Language Change Over the Time**

We take the top 6 threads and see that in thread 252620, it only contains post from the holiday period (summer and Christmas) in the world excluding Australia and New Zealand, therefore we exclude it to prevent bias. We choose 2006 and 2009 because it has the most post than the other years:



We did an average test by mean on 2006, 2009 and 2010 base on three season (January-May, June-September, and December-January). The reason we choose these 3 years is because on the pie chart above, these 3 years have the most posts. The results are the mean of either June-September or December-January will always be the greatest (meaning that many post about leisure are on that time period). This is because June-September is the summer break for students in the world (except Australia and New Zealand) and December-January is the Christmas and new year holiday in the world.