**Programming for Big Data**

**Name:** Melanie Dunne

**Student Number:** 10390914

**Github Repository:** https://github.com/MelanieDBS/CA4\_Big\_Data

**Requirements**

For this assignment we were asked to scrub a log file, organise the data into different holders, and analyse the data it contains. We had to show why and how they would be considered interesting.

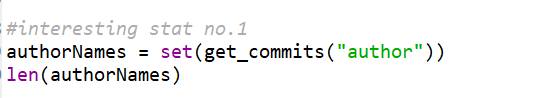
I will submit this document with the code in a zip on Moodle as well as on Github (link above).

**Issues with the code**

Although I had been working on this for a while and had the code in place, I had some issue with the code. The code will run for me, it tells me the shape of the data is 5255 x 422, but when I tried to turn it into a data frame so I could do statistics and visualisations of the set, there was no output. I said nothing was being held in the frame. Hence why I am unable to show the output of these within python. Nearly all of my outputs just brought back a 0.

I was able to upload a CSV version of this set onto Tableau and so that is where my interesting observations will be coming from.

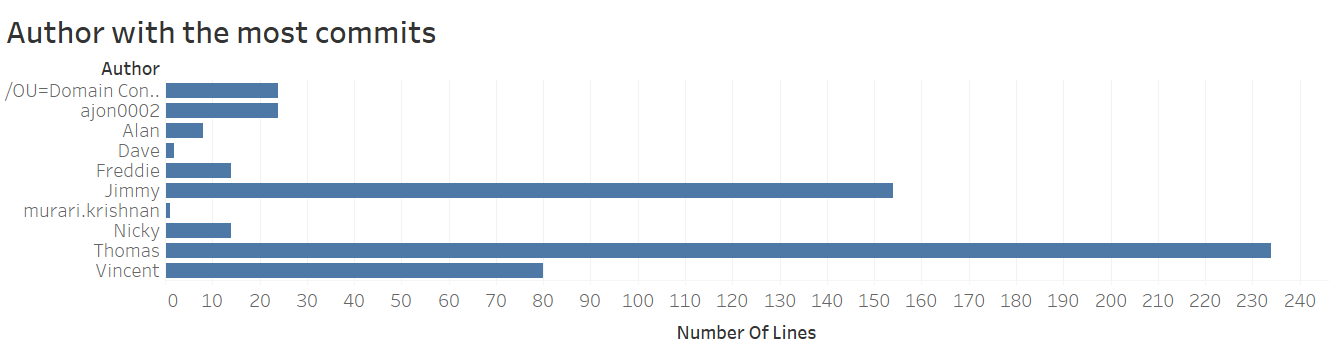
**Interesting observation no.1**

****

This code was implemented and what we saw once it was printed is that it consisted of a list of the individual names that were authors of these commits. Due to the issues that I expressed above I was unable to get an output, but looking at this in Tableau I was able to produce this table.



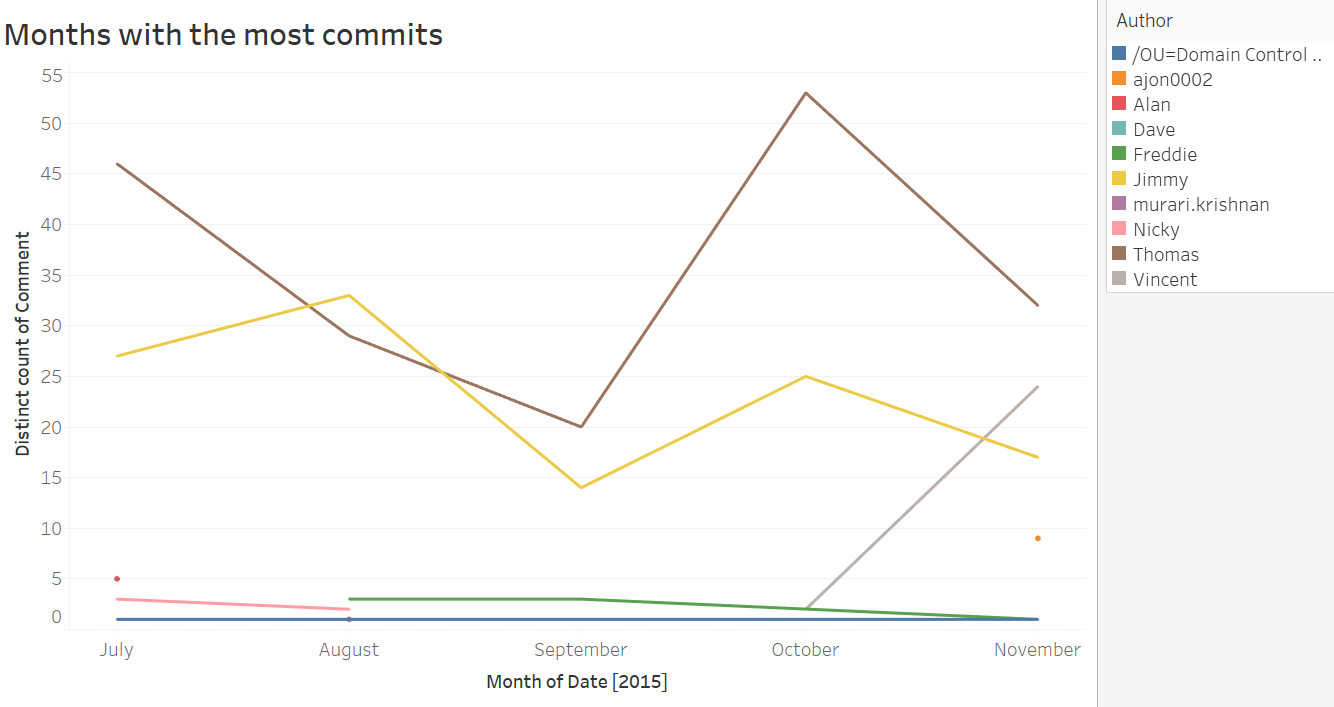
What I would have done next in python was find who had the most commits. In Tableau I was able to show that it was Thomas.



**Interesting Observation No. 2**

I initially tried to set up a barchart to show the author and the amount of comments made. I couldn’t execute due to my issues with the database but after playing on Tableau showed me that it would look a lot like the last barchart.

Instead here is a continuous line graph to show which months yielded the highest number of commits. I differentiated each line (and dot in some cases) by colour. We can see a definite dip in the summertime and a busy period around October.



**Interesting Observation No. 3**

Finally, going along the same theme of time, I have included a Ganntt table that I made up in Tableau. I think this is one of the clearest ways that we can see what time of day most of the commits were made. To note with this observation is that most were made during the standard 9-5 office hours.

