CA4 Programming for Big Data

Emma Rynhart

10125897

Module Code: B8IT105

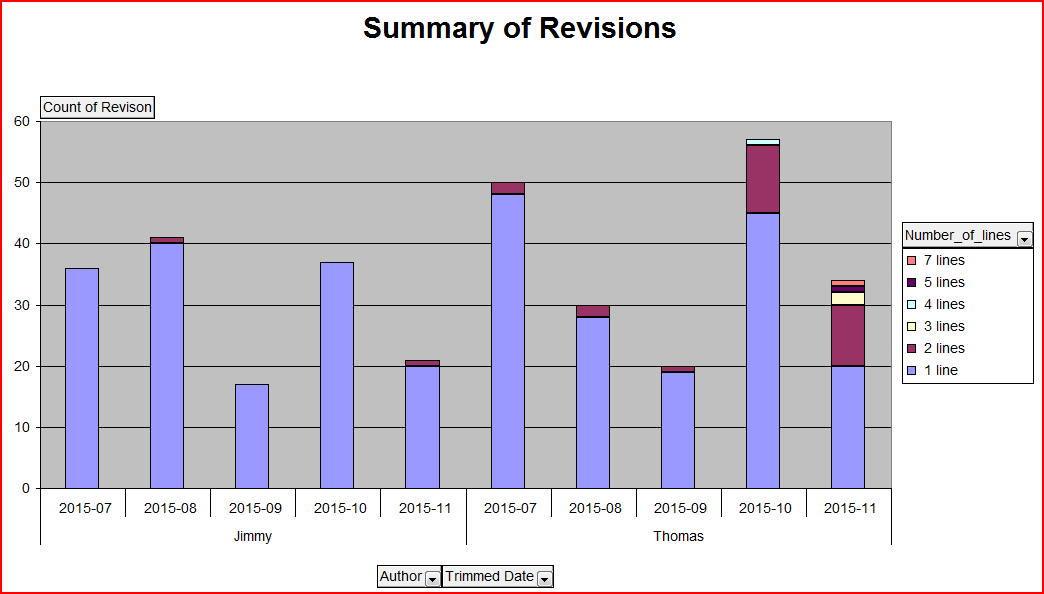
Module Title: Programming for Big Data

Lecturer: Darren Redmond  
Due Date: 07/05/2017

# Statistical Analytics Conclusions

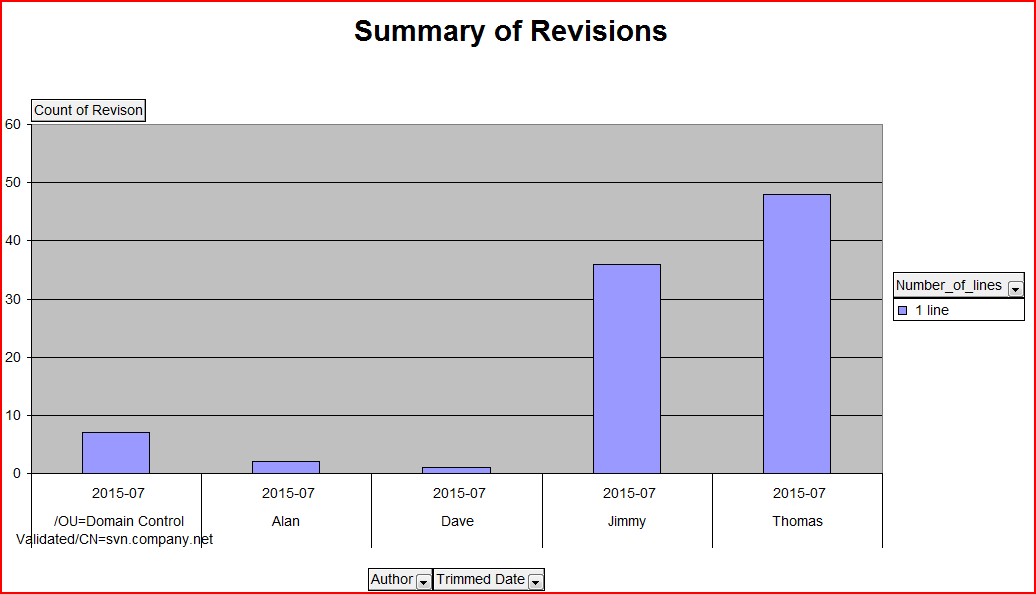
I encountered some technical issues when I began my analysis as I had created my first version of code in a newer version of Python. I managed to compile some data for analysis but I under a time constraint given my technical issues so didn’t get to evaluate it as much as I would have liked.

The first thing that struck me when I saw the data is that Jimmy and Thomas had computed the highest amount of revisions. What also struck me is that they both processed them within the same months with Thomas processing more at the end of the year.

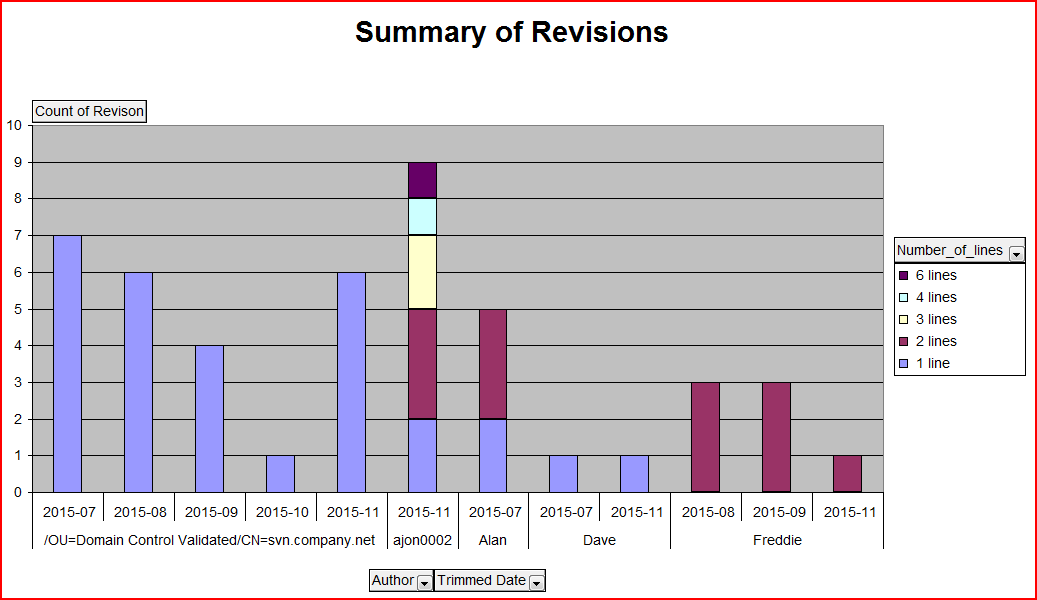


If you look at each month in isolation, Thomas and Jimmy are the highest

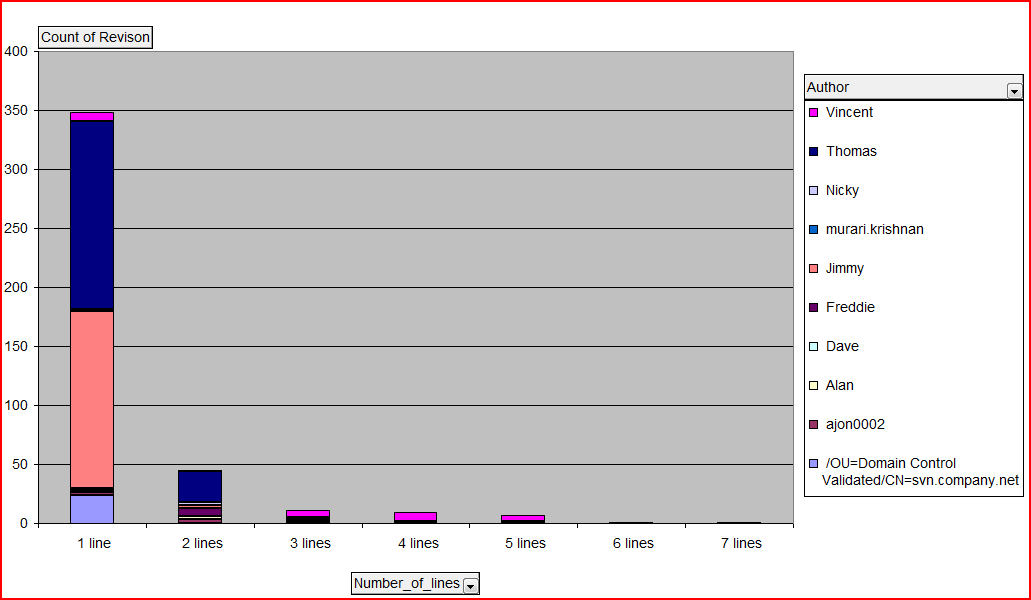
at committing one revision every month:



If you look at the people with the lowest revisions they seem to commit multiple revisions at a time instead on single revisions:



When looking at the total revisions instead of looking at them by month we get a different view of the data as Murari stands out as having the largest amounts of Revision.



If I had more time to refactor my code and extract the data correctly I would be interested in extracting the type of revision that was committed.

There may be a reason that Jimmy and Freddie do so many 1 line revisions e.g. correcting errors of misspelling the file etc. Similarly committing one line at a time could be more accurate and less errors instead of committing multiple lines at once.