Analysis of Changes_Pytong.log

MODULE: B8IT105 PROGRAMMING FOR BIG DATA

STUDENT NO: 10358601

IZABELA TYRNA

DUBLIN BUSINESS SCHOOL

26 NOVEMBER 2017, DUBLIN

The analysis of change_pyton.log was performed using Python programming language. The class Get_Data object was created to read through the file and to extract dates. The test file was created to make sure class object works properly.

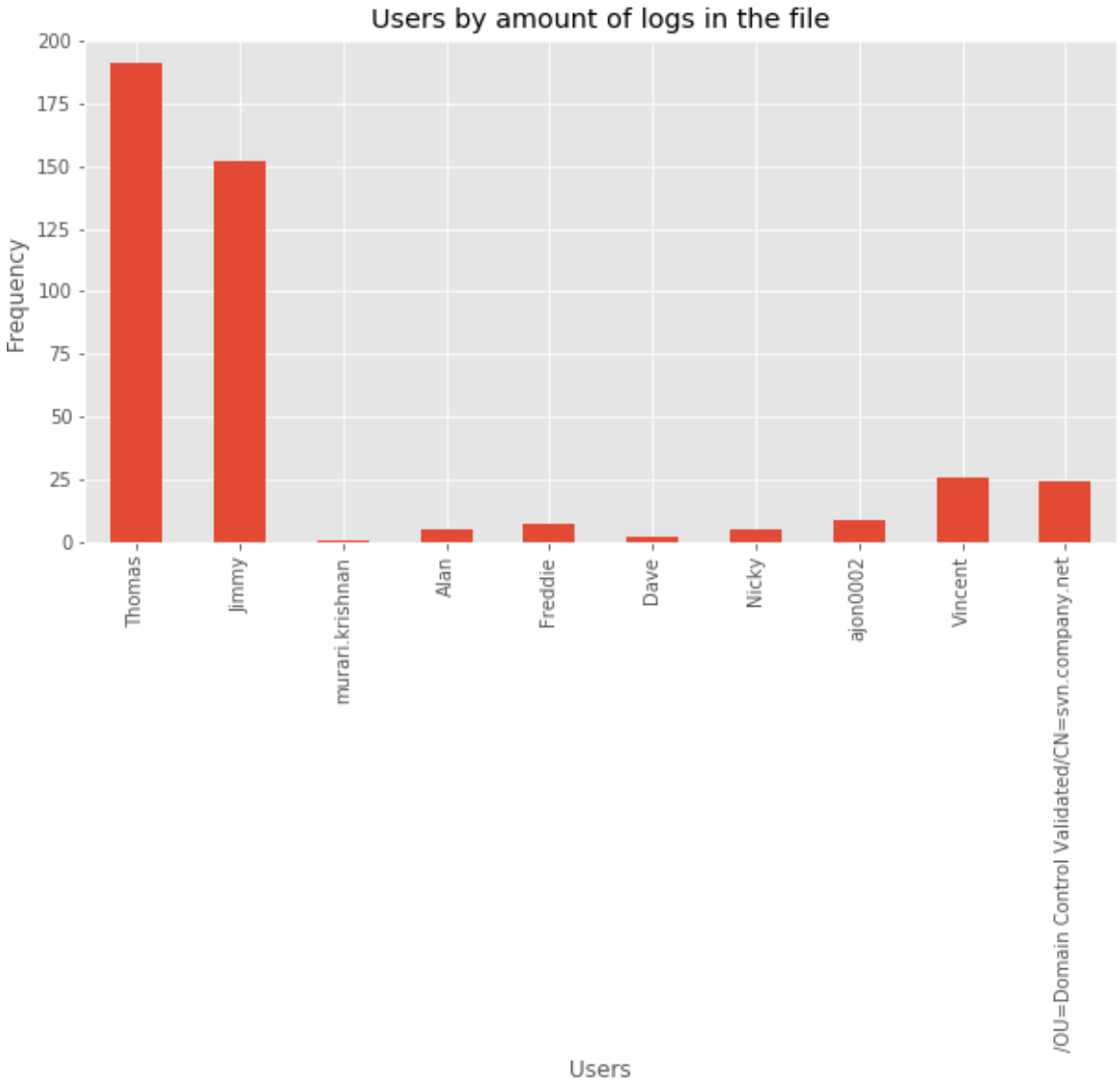
The file contains 5255 lines. The lines containing revision, author, date and number of lines are extracted from the file. The file contains 422 such lines.

Following that it should be noted that the file also contains 3009 lines with add, modify and delete user actions.

Analysis of the date shows that that the file was created between 13th of July 2017 and 27th of November 2017.

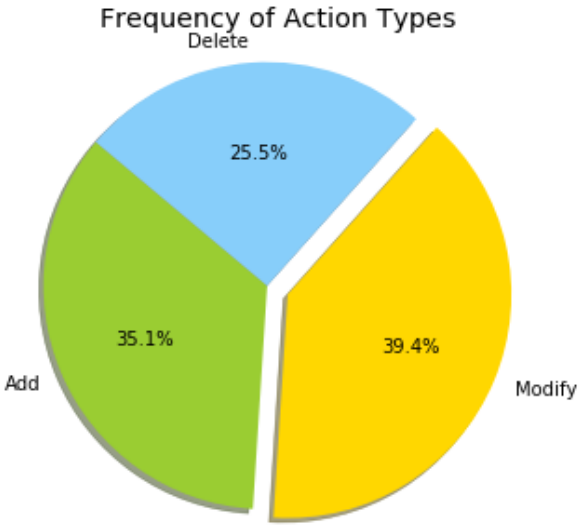
Analysis of unique values in author shows that there are 10 users in the file.

The chart below shows that Thomas and Jimmy were the most active users while other show much less activity.

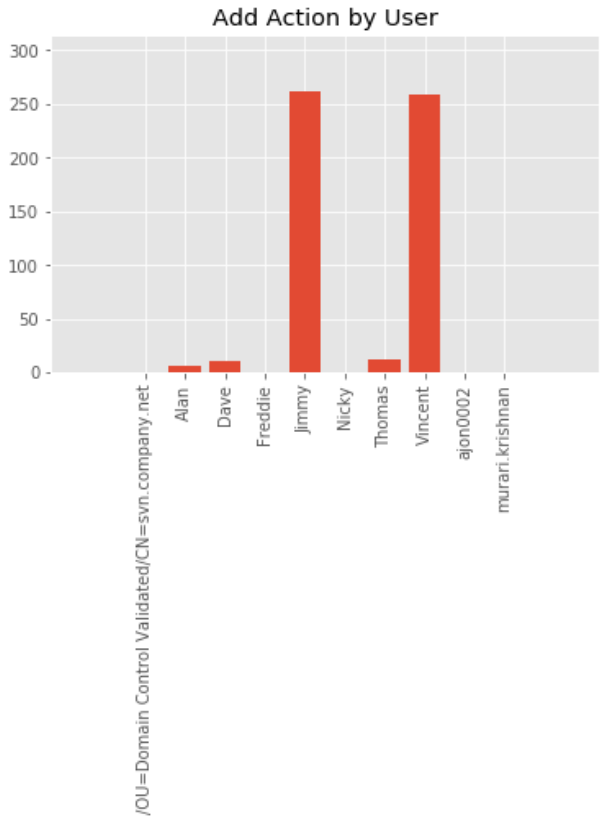


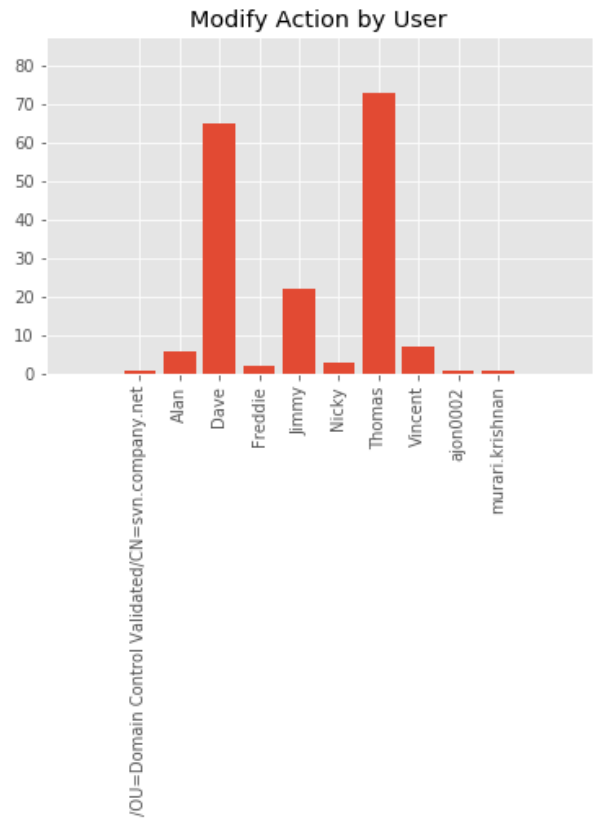
Further analysis of the file shows that add action was performed 1056 times, modify 1186 and delete 767. The pie chart represents the percentage of users' actions.

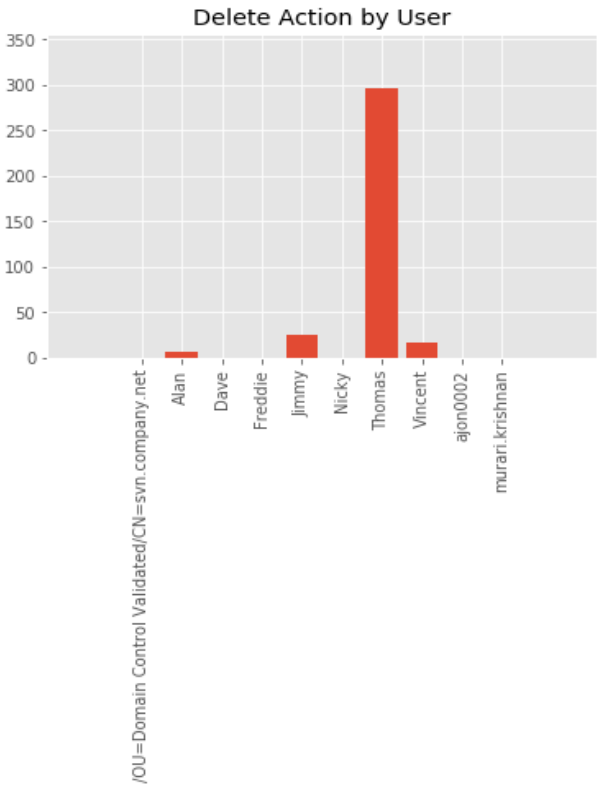
We can see that modify (accounting for 40%) is the most often performed action between 13th of July 2017 and 27th of November 2017.



The below bar charts were created to explore further users' behaviours. It can be seen how often users perform each of actions.







Jimmy and Vincent performed similar amount of add action while other users performed 10 or less.

Thomas and Dave made modifications between 65 to 75 times which is much more that other users.

Delete action was performed 300 times by Thomas. This means that almost half of all deletes were done by him. Other users performed delete actions 25 or less and 3 users did not complete delete action at all.

The descriptive statists show that average number of lines 1.3 and standard deviation is 0.8.

The max for number of line is 7

	add	delete	modify	number_of_lines
count	422	422	422	422
mean	2.502370	1.817536	2.810427	1.315166
std	19.730578	16.378173	6.079996	0.845672
Min	0	0	0	1
25%	0	0	1	1
50%	0	0	1	1
75%	0	0	1	1
max	262	297	73	7

Another interesting insight/metrics that can be extracted from the file are: time series forecasting, number of lines by user, average number of user actions between specific dates and days with the larges and smallest user activity etc.