Aneta Sokolowska

Student number: 10379935 | Date: 04/11/2018

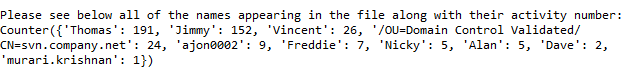
Programming for Big Data

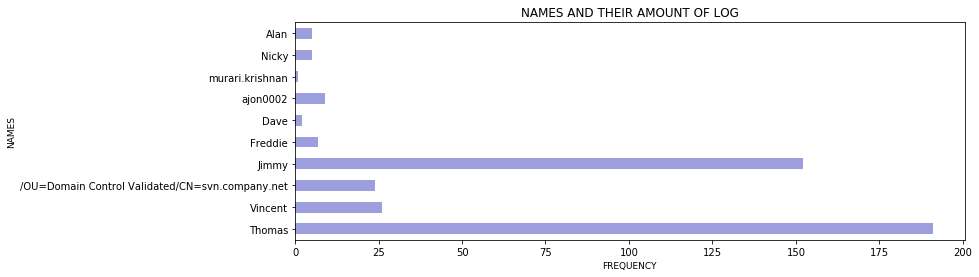
CA 4

The purpose of the ‘Assignment 4’ is to analyse dataset which initially was in text format and had 5000 lines. The dataset was cleaned and the amount of lines was reduced to 422 and put into csv file.

The task is to analyse the output and come up with 3 statistical pieces of information.

First of all, analysing the dataset and using the function counter I was able to check the frequency of names appearing in the file, the outcome is as per below:



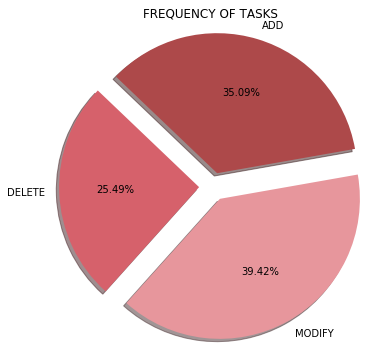
The following chart illustrates the above information:

From the chart ‘NAMES AND THEIR AMOUNT OF LOG’ we can see the below:

1. There are 10 names appearing in the log
2. The most active user is Thomas with just below 200 tasks,
3. Jimmy taking second place with around 150 tasks,
4. Murari.krishnan is taking the last place,
5. Dave is not any better than Murari.Krishnan

We can see that the tasks were created between 2015-07-13 and 2015-11-27.

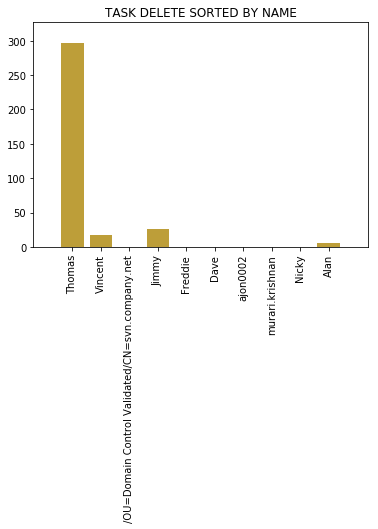
Analysing further we could see the frequency of each task. I have noticed that the file has A, D and M appearing in the file log. For the purpose of this assignment I assume that A is for Add, D is for Delete and M is for modify.



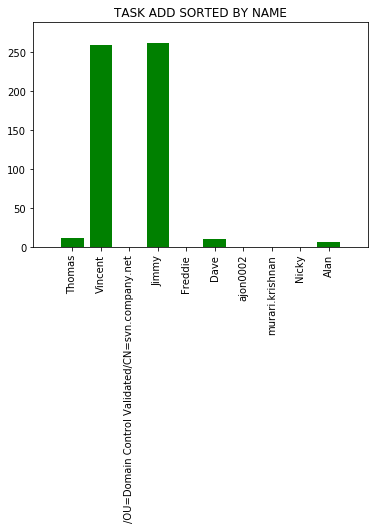
As per ‘FREQUENCY OF TASKS’ chart we can see the below:

1. The ‘modify’ task was the most frequent task with the result of 39.42% of all of the tasks,
2. The ‘add’ task took second place with the result of 35.09% of all of the tasks,
3. The ‘delete’ task was the least frequent task with the result of 25.49% of all of the tasks.

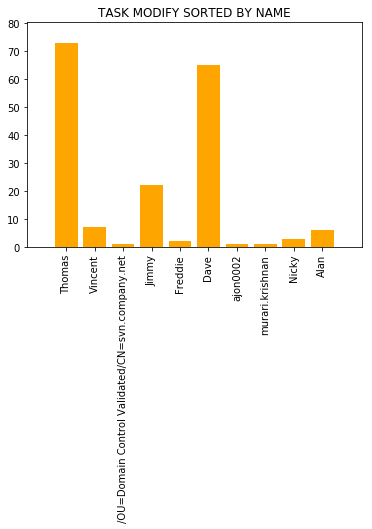
The next chart ‘TASK DELETE SORTED BY NAME’ shows that the person who deleted the highest number of lines was Thomas with the amount around 300. Jimmy deleted around 50 lines and Vincent and Alan around 20. The interesting fact is that the others didn’t delete any tasks. We could think that they have been out of office during the studied period but from the next charts we can see their activity which means they were in the office.



The next chart ‘TASK ADD SORTED BY NAME’ illustrate the amount of lines added per user. Here we can see that the most active were Vincent and Jimmy, with around 250 lines added.

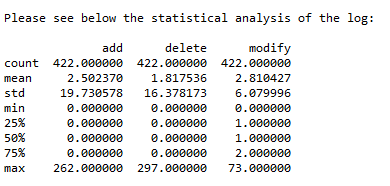


The next chart ‘TASK MODIFY SORTED BY NAME’ we can see that the active users here were Thomas, followed by Dave. Also we can see that every user in the file log have modified some lines.



Looking at the above graphs we could come to conclusion that there are few people in that team who deserve salary increase. Out of 10 user the following in my opinion deserve recognition: Thomas, Vincent, Jimmy and Dave.

We can also check the descriptive statistical analysis of the log per task (‘add’, ’delete’, ’modify’) which is as per below:



The max amount of task is also visible here:

1. We have 262 line that were added, and from previous charts we can see that this was done by Jimmy,
2. The maximum amount of deleted lines is 297, and again as per previous chart we know that this was done by Thomas,
3. The maximum amount of modified lines is 73, ad again as per previous chart we know that this was also done by Thomas.

We can see that the mean for ‘add’ is 2.5, the mean for ‘delete’ is 1.8 and the mean for ‘modify’ is 2.8. This can be also visible in the pie chart above.