**CA4 – Programming for Big Data – Statistical analytics conclusions**

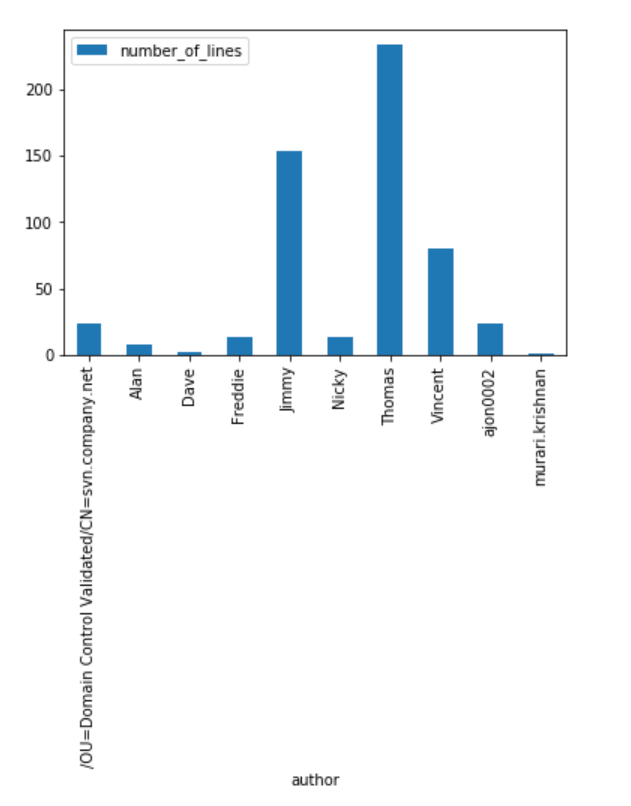
An analysis has been conducted to examine 422 different sets of commit objects to get insights and summary statistics on the performance of the employees.

The 422 objects were first saved in a list, then a data frame has been created from the relevant list. A data frame is a 2-dimensional labelled data structure with columns of potentially different types. The examined data frame shows the following columns: ‘author’, ‘changed\_path’, ‘comment’, ‘date’ and ‘number of lines’

To perform the analysis the following libraries have been used: pandas, numpy, matplotlib.

The analysis has been conducted in the following IDEs: Spyder and Jupiter Notebook (please see .py and .pipynb files)

First the total number of lines of code committed by each author has been calculated and the result was plotted. The chart shows that Thomas, Jimmy and Vincent were far the most productive in terms of line of codes written in the reference period with 234, 154 and 80 lines of code respectively.



The analysis was then oriented on getting insights on the time of the day when the releases were committed, the function describe() has been used

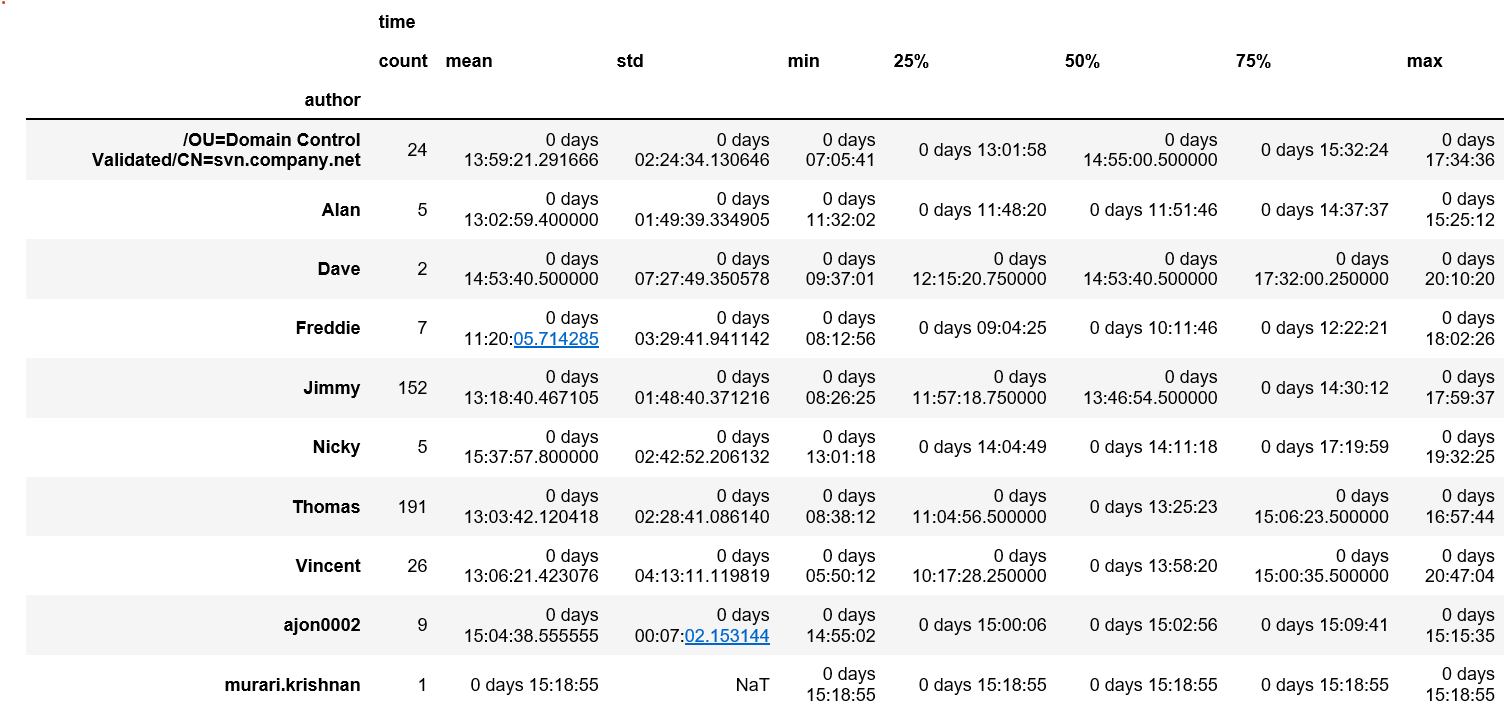
The result of describe() function is a data frame.

The data frame has been plotted and it gave some insights on the distribution of the time when each author commits his releases.

For examples the mean time when the releases were committed by Thomas was 13:03 with a standard deviation of 02:28 hours. The minimum time when Tomas committed was 08:38 and the maximum 16:57.

Instead the mean time when the releases were committed by Alan was 13:03 with a standard deviation of 01:49 hours.

The minimum time when Alan committed was 11:32 and the maximum 15:25, which is an indicator that tells us the Alan usually commits in a very specific range of time.



It has also been noticed that the mean time for the releases for each author doesn’t show a high level of variability, in this perspective a standard deviation of 04:13 hours (that is high compared to the others) for Vincent tells us that he usually he releases in a higher range of time compared to his colleagues.

A third piece of information has been considered relevant during the analysis, which is summary statistics on the project itself and on the employees as a whole.

It has been noticed that November 2015 was the most productive month in terms of total number of lines committed.

The most productive day in terms of lines committed was the 26th of November with a total of 30 number of lines released, this is a really high value considering that the average of lines committed per day is 5.55.

The second and the third most productive days were still in November: 2nd and 20th respectively

The average number of releases per day by all authors was also calculated, using the mean() function, the result showed 1.3 releases per day.

Please note that the analysis on the distribution of time of the day when the releases were committed can only been verified using the .pipynb file and Jupyter Notebook.