Programming for Big Data - B8IT105 – CA04

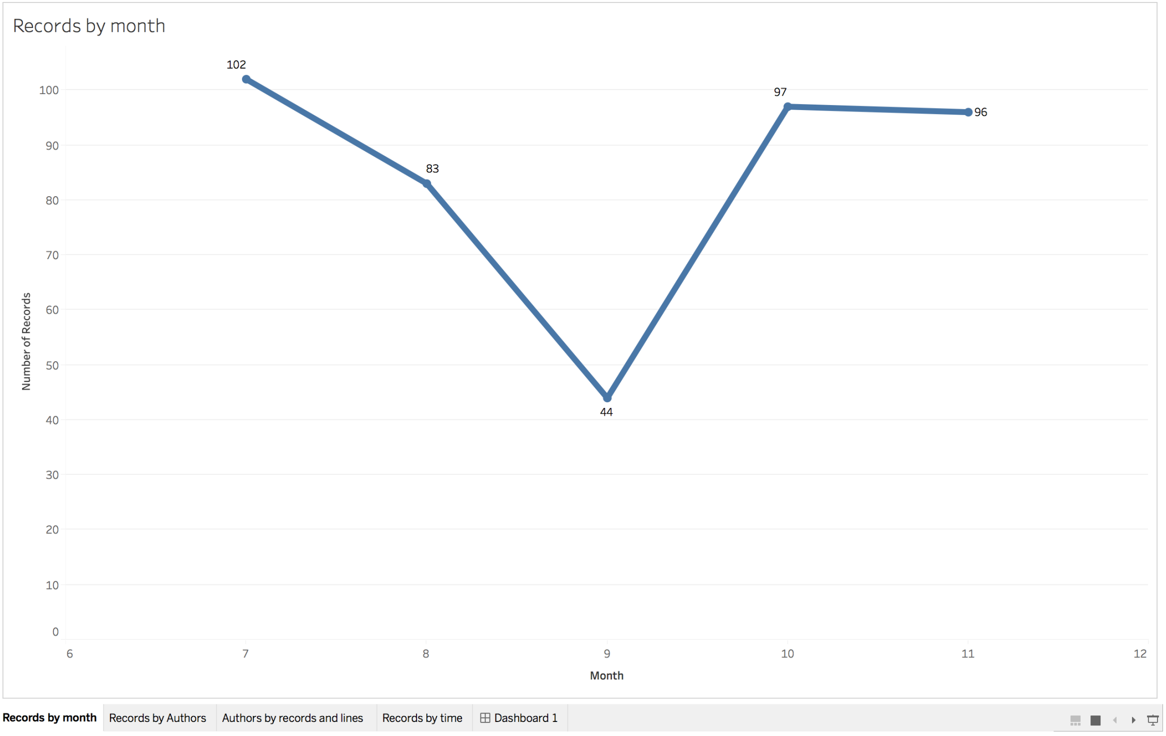
Elisandra Silva – 10347211

Report:

For our continuous assessment 4 we will be transforming a large dataset in text format, with precisely 5255 lines of text, into a readable information so we can analyse and take insights out of it, transporting it into a table format and creating a dashboard using Tableau to display our findings.
First thing that needs to be done is look at the data to find patterns so we can program functions in Python to read, separate and properly clean the data into relevant container objects. We ended up with a total of 422 different sets of commit objects.
By looking at the data, we can see that the first line of each commit contains a lot of relevant information that can be separated for further analysis, such as revision, author, date and time, and number of lines. So, this is how we want to parse our data by row:

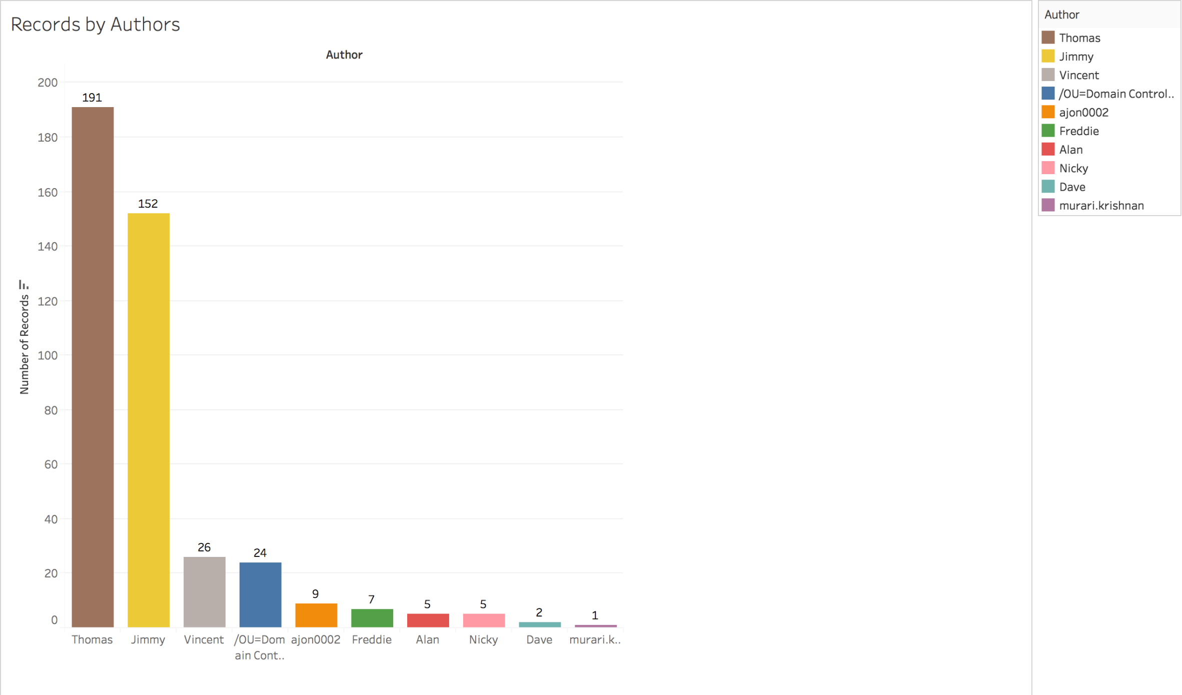
```
['r1551925', 'Thomas', '2015-11-27', '16:57:44', '1']
```

Next, we want to include all 422 objects that will look like the example above, into a table format. Transporting our new csv file to Tableau we can visualize some interesting facts about this data:

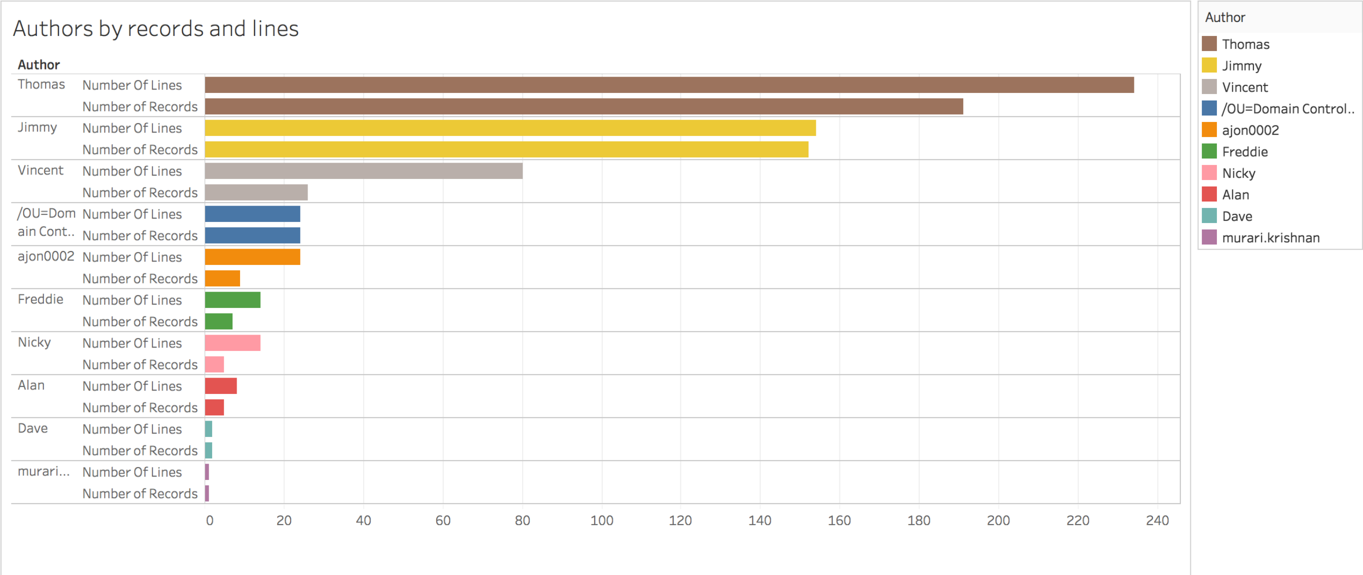


On the first chart that analyses **Number of Records by Month**, is possible to see that July was the month with the highest number of records (102), contrasting with September which only had 44 records in total.

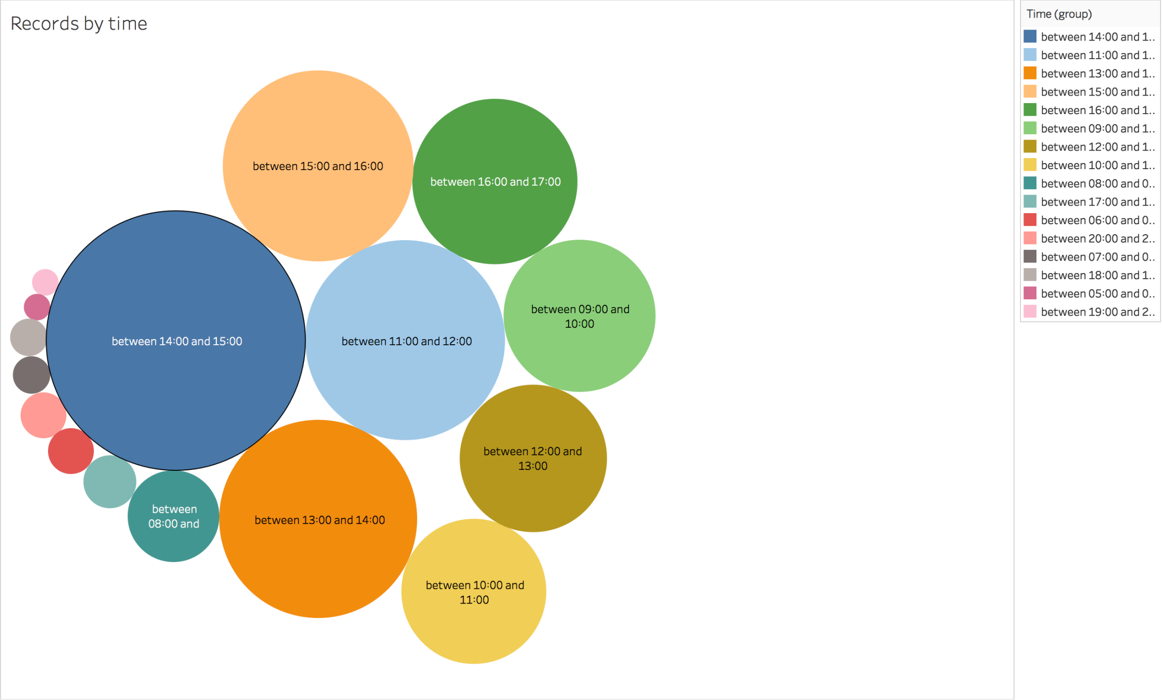
If we look at the **Records by Authors**, we can see the person that had the higher number of records is Thomas, with a total of 190 records over the analysed period:



Looking further at Authors' registers, is also possible to visualize that the person with the highest number of records is also the person appearing on the file with more lines in total per resgister:

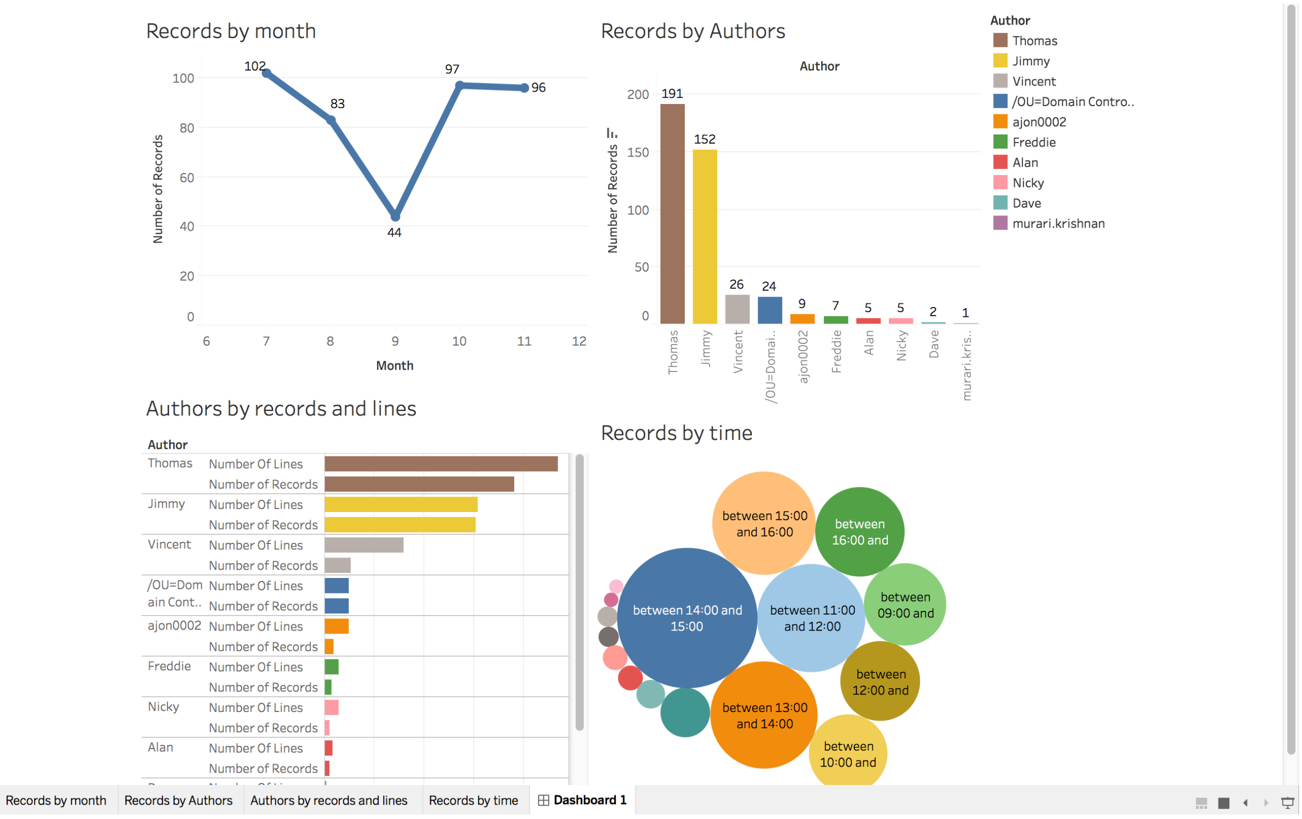


Grouping the time by hourly intervals, it was also possible to see at what time of the day we had more occurrences over the analysed period:



The chart above shows us that the busiest time of the day over the analysed period was between 14:00 and 15:00, with precisely 96 records, while the quietest times were between 05:00 and 06:00 as well as the time between 19:00 and 20:00, with 1 record each.

See complete dashboard bellow:


Records by month
Number of Records
100
80
60
40
20
0
102
83
44
97
96
6
7
8
9
10
11
12
Month
Records by Authors
Author
Number of Records
200
150
100
50
0
191
152
26
24
9
7
5
5
2
1
Thomas
Jimmy
Vincent
/OU=Domai..
ajon0002
Freddie
Alan
Nicky
Dave
murari.kris..
Author
Thomas
Jimmy
Vincent
/OU=Domain Contro..
ajon0002
Freddie
Alan
Nicky
Dave
murari.krishnan
Authors by records and lines
Author
Thomas
Number Of Lines
Number of Records
Jimmy
Number Of Lines
Number of Records
Vincent
Number Of Lines
Number of Records
/OU=Domain Cont..
Number Of Lines
Number of Records
ajon0002
Number Of Lines
Number of Records
Freddie
Number Of Lines
Number of Records
Nicky
Number Of Lines
Number of Records
Alan
Number Of Lines
Number of Records
Records by time
between 15:00 and 16:00
between 16:00 and
between 14:00 and 15:00
between 11:00 and 12:00
between 09:00 and
between 12:00 and
between 13:00 and 14:00
between 10:00 and
Records by month
Records by Authors
Authors by records and lines
Records by time
Dashboard 1