CA4 – Perform Analysis on 5000-line dataset

Michelle Carr
10032026
Submission date: Sunday 7th May

Using the python code attached in appendix 1, I ran a cleaning program on the change log file to tidy up the data and put it into a manageable format. I then output this to CSV for further analysis.

1. **Prolific authors over the measure time frame:**

I connected the CSV file to Tableau and produced the following graphics on authors and the number of commits over the time period measured. Live Tableau workbooks are in appendix 2.

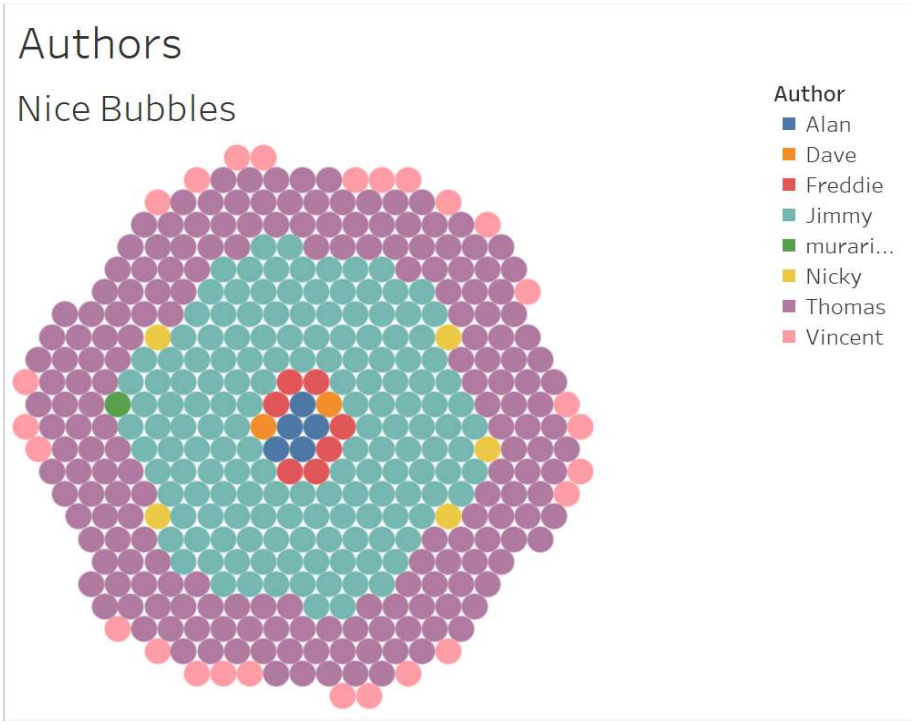

Fig 1

From Fig 1, It is easy to identify which authors (Jimmy & Thomas) have been the most prolific commits over the measured period. From an initial glance at this graphic it appears that Jimmy may have slightly more commits than Thomas.

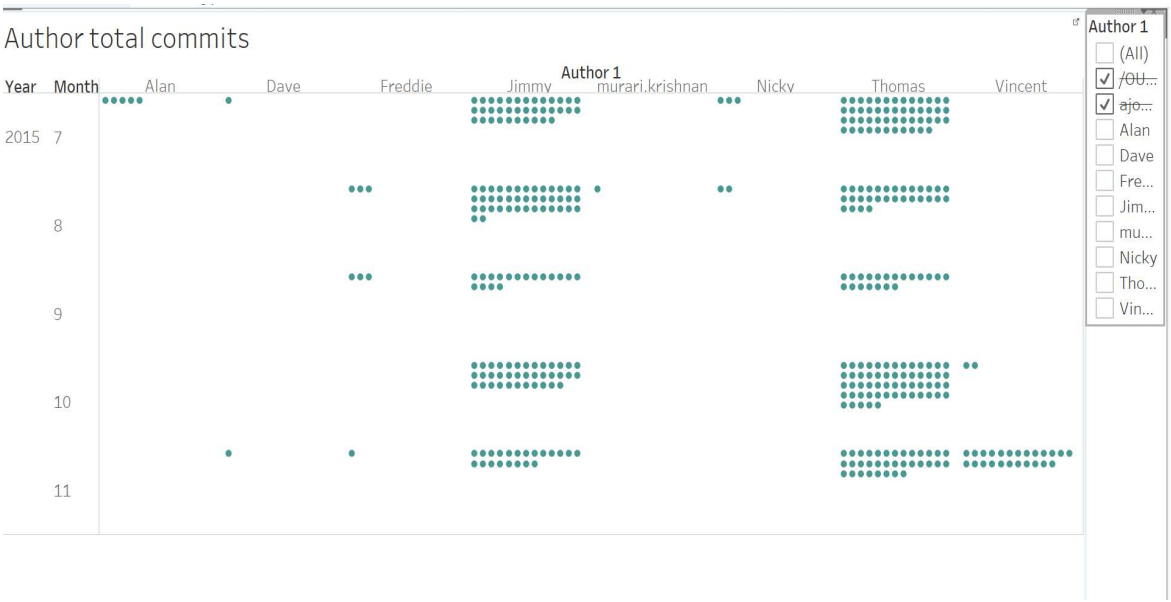

Fig 2

This is another visual representation of commit counts v authors. This is broken down by month on the x axis and provides slightly more details that the graphic in fig 1. Here it can be seen that Thomas appears to have a slight edge on Jimmy on the commit rates over the measured time- frame. This graphics although basic can be a nice starting point to identify trends that require further analysis.

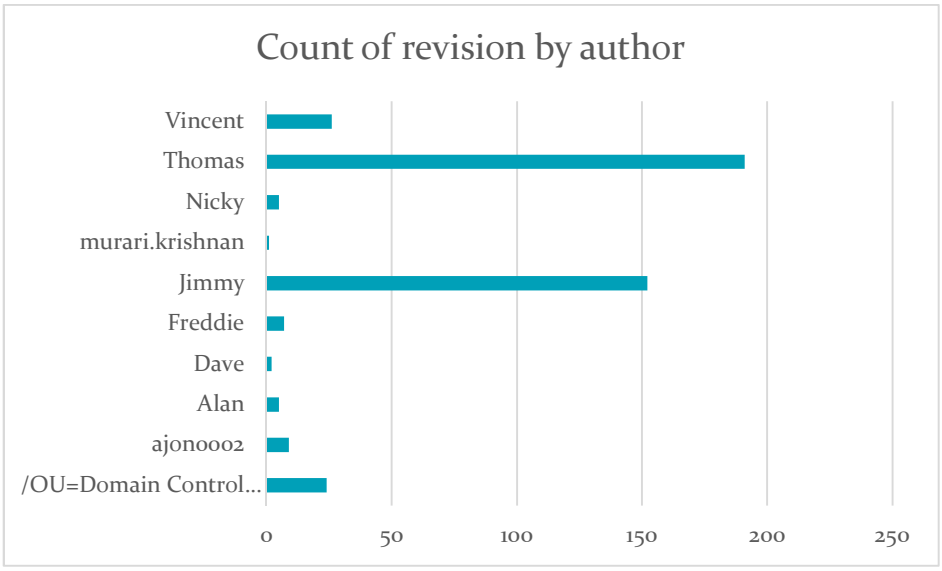

Fig 3 – beginning Excel and R analysis

Having looked at these graphics I used excel (worksheet in appendix 2) to sum the no of commits per authors:

Excel	
Author	***Count of commits***
Thomas	***191***
Jimmy	***152***
Vincent	***26***
ajon 002	***9***
Freddie	***7***
Alan	***5***
Nicky	***5***
Dave	***2***
Murari	***1***
/OU=Domain Control Validated/CN=svn.company.net	***24***

From this simple count calculation in excel it confirms what the visuals show above, that Thomas was the most profile author making commits over the measure period. The count of commits per author can also be carried out in R (script and console results in appendix 1) with all the above results confirmed:

R programming	
Author	**(Count of commits , Avg commit per day)**
/OU=Domain Control Validated/CN=svn.company.net	(24, 0.057)
ajon0002	(9, 0.021)
Dave	(2, 0.005)
Jimmy	(152, 0.360)
murari.krishnan	(1, 0.002)
Nicky	(5, 0.012)
Thomas	(191, 0.453)
Freddie	(7,0.17)
Vincent	(26,0.062)
Alan	(5,0.012)

Statistical analysis concludes that Thomas was the most profile author to commit of code over the measured period. Thomas is significantly higher than his nearest competitor of Jimmy.

The range between the highest committer Thomas and the lowest committer Murari is 190. It is clear from the above analysis that Thomas and Jimmy appear to be the most efficient workers in this group having a no of commits far exceeding the average per author.

On the surface this seems good for these employees however a high quantity or fast work rate does not always mean quality of work begin achieved. It would be interesting to see how the individual number of lines (one line v block of many lines) of code each of these two top committers (Thomas and Jimmy) are submitting on each commit vs some of the lower author commit rates.

2. **Can the number of lines submitted during each commit suggest a pattern among authors on quality of code submitted?**

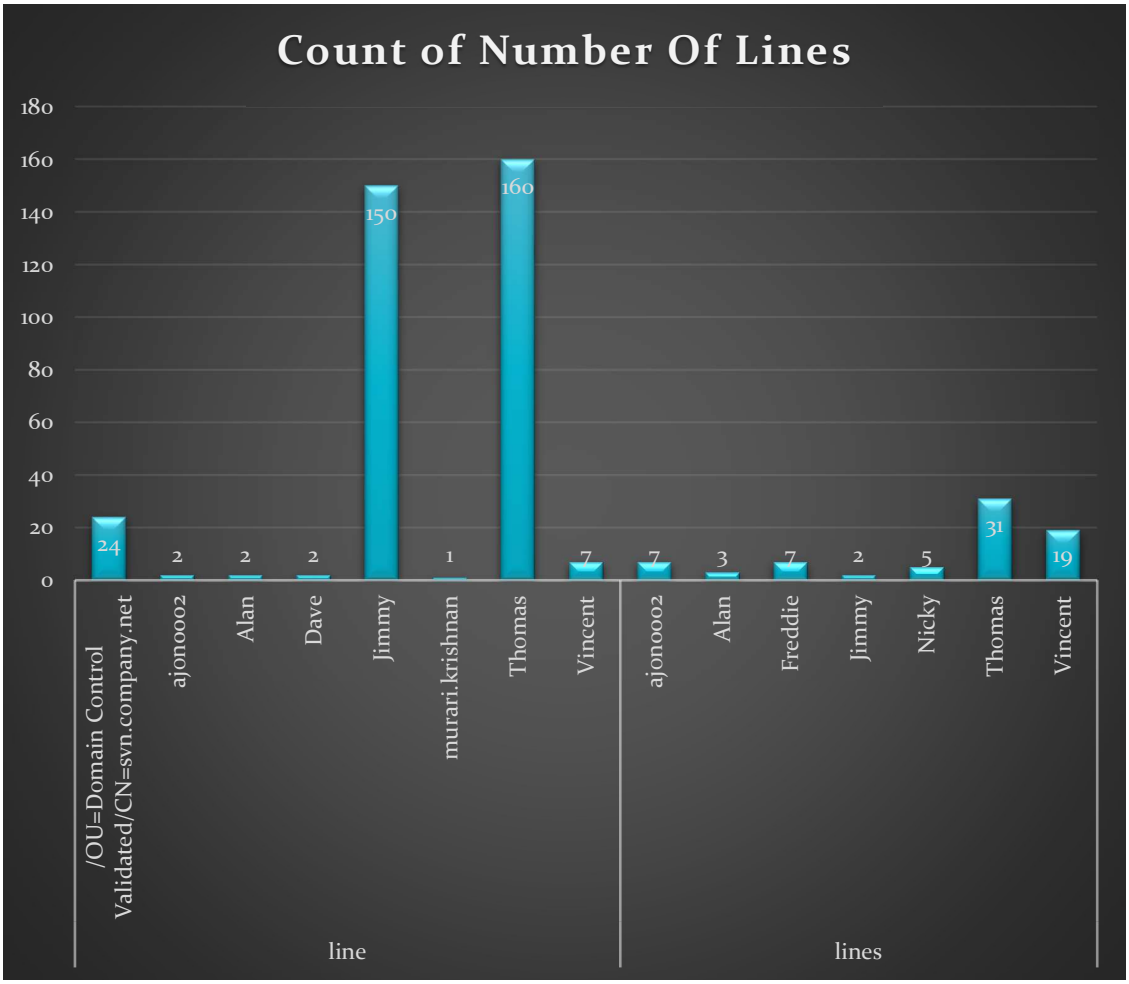

Fig 4

From the above graph produced in excel, the top committer Thomas has committed 160 instances of one line of code, versus only 31 instances of many lines of code. Jimmy as the second highest committer rate has committed 150 one instance lines and only 2 instances of many lines of code.

Thomas - lines of code	
	% submitted during commits
one line	78.53%
more than one line	16.23%

Jimmy - lines of code	
	% submitted during commits
one line	***98.68%***
more than one line	***1.32%***

Jimmy's initial analysis suggests a good work rate, however he only submitted less than 2% of more than one line of code. This may be normal enough in a development environment where constant and small revisions take place.

R Analysis of the lines adds weight this the inference where the proportion of lines (more than one line of code) is in line with Jimmy individual contributions of lines(% rate as calculated above) summitting one line of code at a time in present in all authors irrespective of commit rates:

R programming		
Value	**Line (1)**	**Lines (more than 1)**
frequency	348	74
Proportion	0.825	0.175

Inferences could be made here of the style of work environment that suits each of the developers/authors – that Thomas is capable of short and longer period of work whereas Jimmy many be easily distractible and only suited to very short intense period of work.

3. **Measure period July to November – does the traditional summer holiday period have any impact on workflow and committals of code.**

The time period (July to November 2015) has occurred during the traditional summer holiday timeframe on Ireland, assuming all these developers were working for an Irish company and located here does the number of commits reflect drop in productivity in the traditional summer holidays in Ireland? To achieve this, I imported the CSV into Tableau and split the date/time columns into independent columns (Year, Month, Time).
I then produced the following graphic, support workbook in appendix 2.

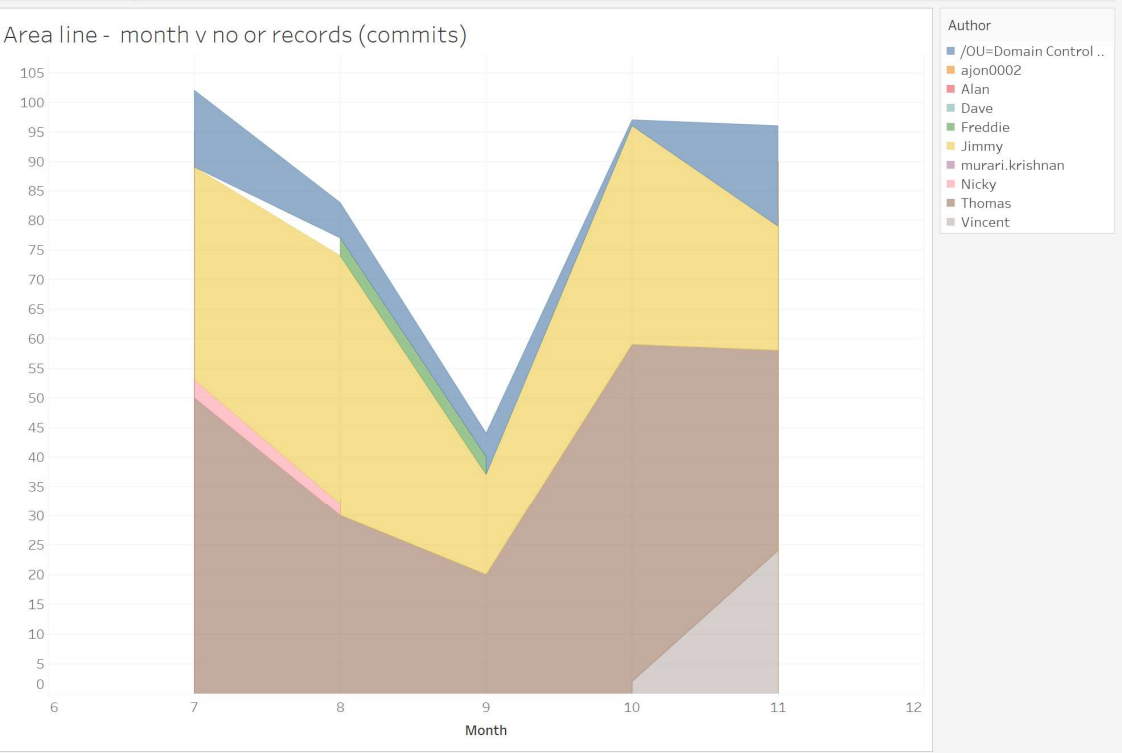

Fig 3

This graphic clearly shows a considerable dip in records (no of commits) during the month of September indicating a drop-in work efficiencies in that month. September would not be a traditional holiday month in Ireland, and you can infer form this that perhaps the developers favor this month for holidays. A significant drop in records in this month might also indicate the end of the current project and bench time before a new one begins but the records (commit rates) rise again in October suggesting that this is not the reason behind the drop in commits in September.

Conclusion

The statistics focused on in this report were mainly descriptive in nature. With further time and analysis more predictive analysis could be completed on this dataset to make more inferences. The above data set shows commits' over a time period in a quantitative fashion and does not account for human behaviors and qualitative factors that may have influenced the data. An example cold be that Thomas/Jimmy primary job role is mainly concerned with coding, whereas Murari/Nicky may be in a management role where they review and make slight changes to code. The context of the data gathering would play an influential role in using it in a more predictive fashion.

This small example hints at global issues in big data analysis namely surrounding privacy, security and information sharing. The inferences made in this report, are strongly focused on a subset of business intelligence and arguable could make employees feel "watched" if they knew such information on workflow rates could be use in HR matters like bonus calculations and KPI measurements. As a counter to this, if this information was used to make an work environment more friendly or tailored to individual need based on data like this then perhaps would ameliorate some of the employee concerns over data gathering.

Appendix 1 : Python & R script
Appendix 2: Tableau, Excel

END