**CA4 – Analysis of 5,000 Line Dataset**

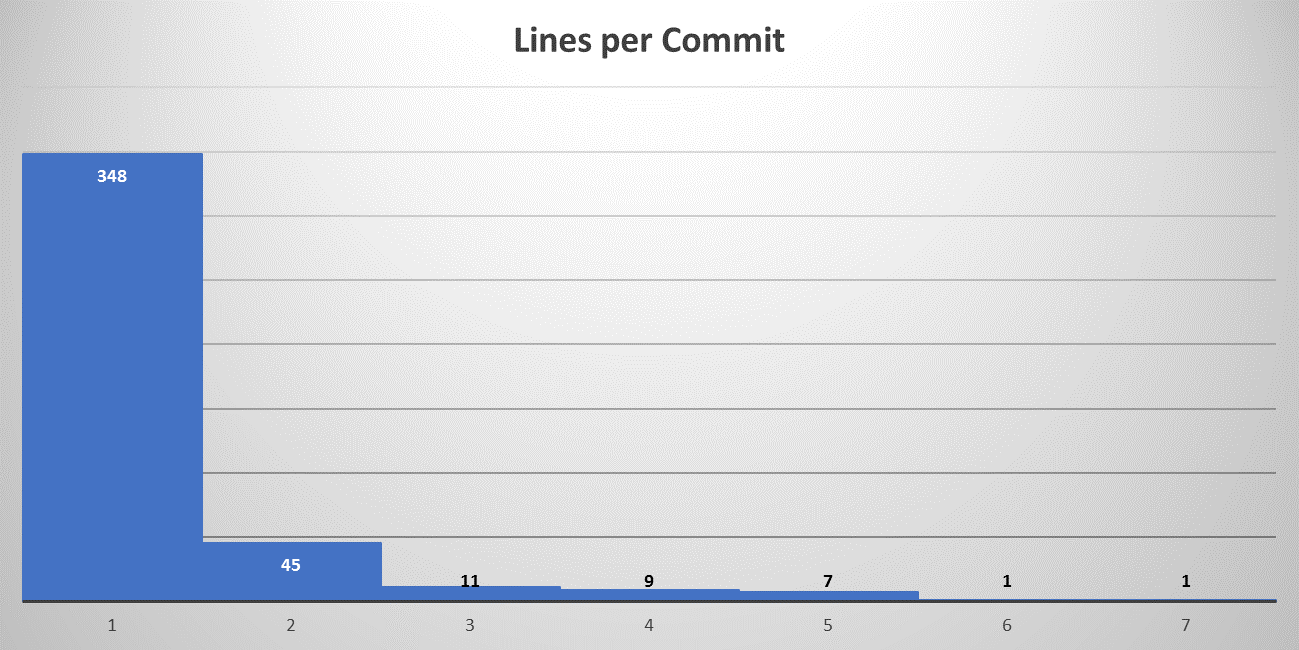
**The Data**

When the data was cleaned and exported as a .csv file, the fields generated were:

|  |  |
| --- | --- |
| Field Name | Description |
| revision | unique identifier code for each commit |
| author | name of the person/entity responsible for the commit |
| date | date of the commit |
| time | time of the commit |
| number\_of\_lines | number of lines in the commit |
| comment | additional notes on the commit |

**Analysis 1 – Number of Lines per Commit**

The number of lines per commit had an expected distribution where the shortest commits made up the largest proportion of the dataset. As the number of lines in the commit increased, the less frequent these types of occurrences were. Below is histogram to visualise this distribution:



This distribution could also be visualised by creating a pivot table of the data where the number\_of\_lines were applied to the rows and the count\_of\_commits were the table’s values:

|  |  |  |  |
| --- | --- | --- | --- |
| **Number\_of\_lines** | **Count of Commits** |  | **Proportion** |
| 1 | 348 |  | 82.5% |
| 2 | 45 |  | 10.7% |
| 3 | 11 |  | 2.6% |
| 4 | 9 |  | 2.1% |
| 5 | 7 |  | 1.7% |
| 6 | 1 |  | 0.2% |
| 7 | 1 |  | 0.2% |
| **Grand Total** | **422** |  |  |

By applying the two formulae below in Excel, it was also determined that of the 18 commits containing 4 or more lines, the author Vincent was responsible for two-thirds of these higher volume commits.

* =COUNTIFS(E:E,">=4")
* =COUNTIFS(E:E,">=4",B:B,"Vincent")

What this may indicate is that Vincent tends to be taking on more complex work than that of the other authors.

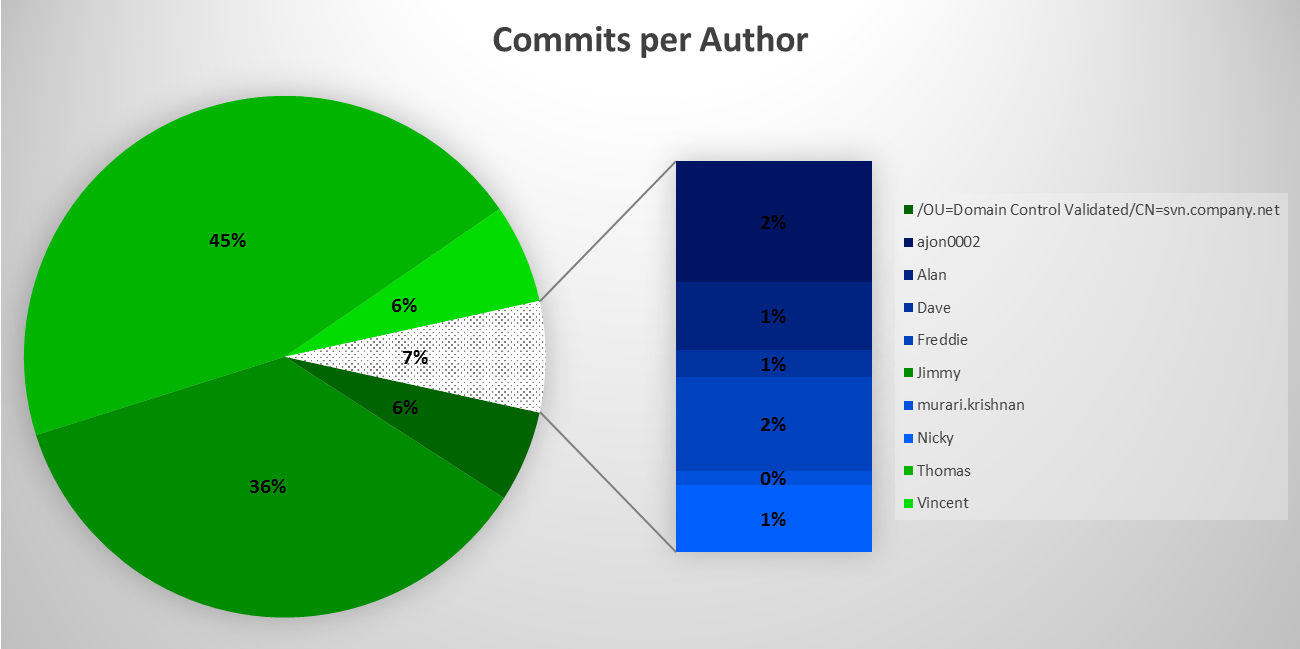
**Analysis 2 – Number of Commits per Author**

Of the ten authors listed in the dataset, there appeared to be two authors that were committing much more work than the others; Thomas and Jimmy.

|  |  |
| --- | --- |
| **Row Labels** | **Count of Commits** |
| /OU=Domain Control Validated/CN=svn.company.net | 24 |
| ajon0002 | 9 |
| Alan | 5 |
| Dave | 2 |
| Freddie | 7 |
| Jimmy | 152 |
| murari.krishnan | 1 |
| Nicky | 5 |
| Thomas | 191 |
| Vincent | 26 |
| **Grand Total** | **422** |

Again, this was determined by creating a pivot table of the data where the authors were applied to the rows and the count\_of\_commits were the table’s values.

On the next page is a pie chart to visualise this distribution of commits completed by author.

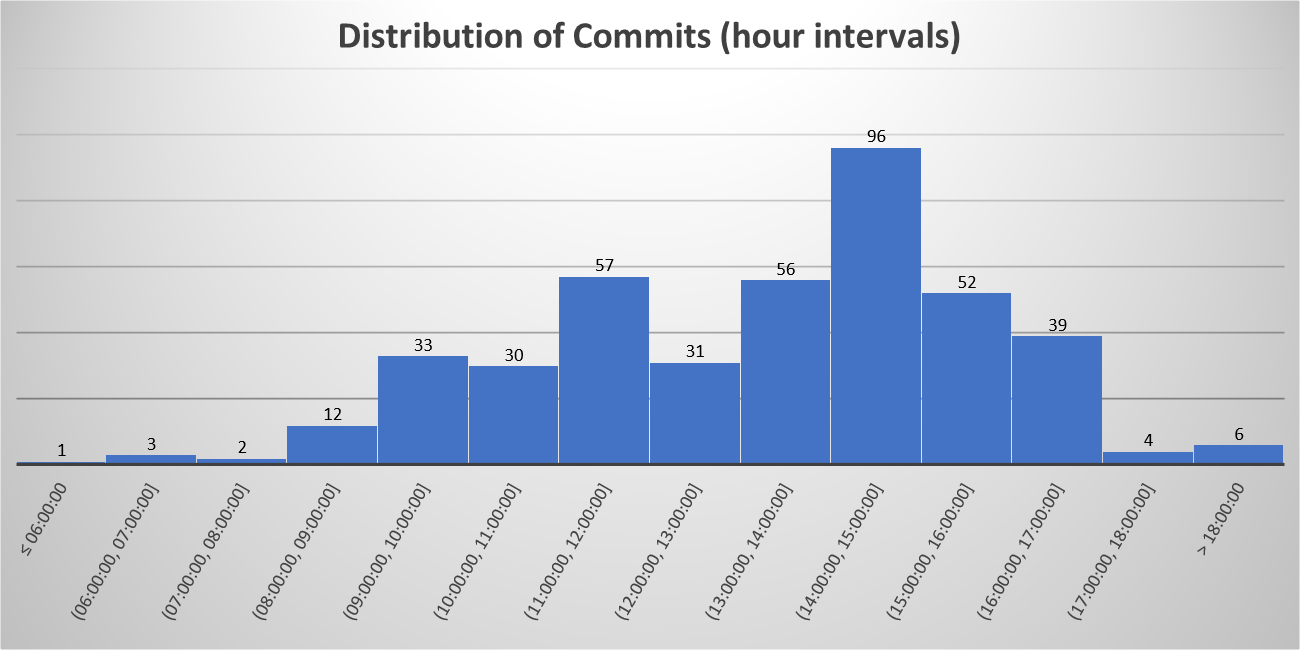


Again, applying the formulae below in Excel, it was determined that 160/191 of Thomas’ commits were one-line. The same applies for 150/152 of Jimmy’s commits also.

* =COUNTIFS(E:E,"1",B:B,"Thomas")
* =COUNTIFS(E:E,"1",B:B,"Jimmy")

**Analysis 3 – Hourly Interval Analysis**

The third analysis carried out was on the time of day that the commits were taking place. A histogram was created to visualise the number of commits that were submitted between 6am and 6pm during the day:



As you would expect, very few commits were made during ‘unsociable’ hours; 1 commit was made between 00:00-06:00 and 6 commits were made between 18:00-00:00.

The peak time for commits was between 14:00-15:00 which appears to indicate that the authors were much more motivated after lunch!

**Conclusion**

The three analyses detailed above were found to be the most interesting as it details the volume of the work, the main contributors and the times that the commits were made.

Other analyses were carried out looking at the month of the commit and the type of work by creating fields that contained certain text strings in the ‘comment’ field, eg. “phone”, “Android”, “test”.

However, these investigations did not provide as interesting material as the selected analyses.