* + - [**B8IT105 Programming for Big Data (B8IT105\_1718\_TM**](https://elearning.dbs.ie/course/view.php?id=7370)

**B8IT105 Programming for Big Data**

CA4 - Perform Analysis on a 5000 line dataset

Cathal O’Gorman (10361706)

December 2017

**The Data**

The original data source is a 5,255 line text file containing changes committed to a repository. This was arranged into columns by splitting on the relevant delimiter and   
saved as a CSV file. It was found that there were 422 commits completed. The following   
six row headers were added: Revision, Author, Date, Time, Number of Lines, and Comment.

**Analysis in Python**

Additional functions were added to find the number of commits by Author,  
and by Date. A third function found how many commits were of each line length.   
Tests were completed to verify sample observations. These findings were reproduced in Excel.

**Analysis in Excel**  
Commits by Author

The 422 commits were completed by 10 authors. It’s interesting to note that over 81% of these were submitted by either Thomas or Jimmy. The six lowest contributors completed just 6.87% of submissions between them.

|  |  |  |
| --- | --- | --- |
| **Author** | **Number of Commits** | **% Total** |
| **Thomas** | **191** | **45.26%** |
| **Jimmy** | **152** | **36.02%** |
| **Vincent** | **26** | **6.16%** |
| **/OU=Domain Control Validated/CN=svn.company.net** | **24** | **5.69%** |
| **ajon0002** | **9** | **2.13%** |
| **Freddie** | **7** | **1.66%** |
| **Alan** | **5** | **1.18%** |
| **Nicky** | **5** | **1.18%** |
| **Dave** | **2** | **0.47%** |
| **murari.krishnan** | **1** | **0.24%** |
| **Total** | **422** | **100.00%** |

Commits by Date / Date / Time

The commits were completed on 76 weekdays between 13th July and 27th November 2015.  
It’s interesting to note that the most activity took place on 4th August with 19 submissions, with 13th July the next highest at 14. Approximately half the commits relate to 19 dates, the other half over the remaining 57.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Date** | **Count of Date** | **% Total** |  | **Date** | **Sorted Count** | **% Total** |
| **13-Jul** | **14** | **3.32%** |  | **04-Aug** | **19** | **4.50%** |
| **14-Jul** | **7** | **1.66%** |  | **13-Jul** | **14** | **3.32%** |
| **15-Jul** | **13** | **3.08%** |  | **15-Jul** | **13** | **3.08%** |
| **16-Jul** | **9** | **2.13%** |  | **29-Oct** | **13** | **3.08%** |
| **17-Jul** | **7** | **1.66%** |  | **09-Oct** | **12** | **2.84%** |
| **20-Jul** | **1** | **0.24%** |  | **12-Nov** | **12** | **2.84%** |
| **21-Jul** | **6** | **1.42%** |  | **20-Nov** | **12** | **2.84%** |
| **22-Jul** | **2** | **0.47%** |  | **22-Oct** | **11** | **2.61%** |
| **23-Jul** | **10** | **2.37%** |  | **02-Nov** | **11** | **2.61%** |
| **24-Jul** | **9** | **2.13%** |  | **26-Nov** | **11** | **2.61%** |
| **27-Jul** | **3** | **0.71%** |  | **23-Jul** | **10** | **2.37%** |
| **28-Jul** | **2** | **0.47%** |  | **30-Jul** | **10** | **2.37%** |
| **29-Jul** | **3** | **0.71%** |  | **06-Oct** | **10** | **2.37%** |
| **30-Jul** | **10** | **2.37%** |  | **16-Jul** | **9** | **2.13%** |
| **31-Jul** | **6** | **1.42%** |  | **24-Jul** | **9** | **2.13%** |
| **04-Aug** | **19** | **4.50%** |  | **13-Aug** | **9** | **2.13%** |
| **05-Aug** | **5** | **1.18%** |  | **20-Aug** | **9** | **2.13%** |
| **06-Aug** | **2** | **0.47%** |  | **09-Sep** | **9** | **2.13%** |
| **07-Aug** | **2** | **0.47%** |  | **23-Oct** | **9** | **2.13%** |
| **10-Aug** | **4** | **0.95%** |  | **14-Jul** | **7** | **1.66%** |
| **12-Aug** | **7** | **1.66%** |  | **17-Jul** | **7** | **1.66%** |
| **13-Aug** | **9** | **2.13%** |  | **12-Aug** | **7** | **1.66%** |
| **14-Aug** | **4** | **0.95%** |  | **29-Sep** | **7** | **1.66%** |
| **17-Aug** | **2** | **0.47%** |  | **07-Oct** | **7** | **1.66%** |
| **19-Aug** | **5** | **1.18%** |  | **16-Nov** | **7** | **1.66%** |
| **20-Aug** | **9** | **2.13%** |  | **27-Nov** | **7** | **1.66%** |
| **21-Aug** | **4** | **0.95%** |  | **21-Jul** | **6** | **1.42%** |
| **24-Aug** | **2** | **0.47%** |  | **31-Jul** | **6** | **1.42%** |
| **25-Aug** | **3** | **0.71%** |  | **04-Sep** | **6** | **1.42%** |
| **26-Aug** | **3** | **0.71%** |  | **27-Oct** | **6** | **1.42%** |
| **27-Aug** | **2** | **0.47%** |  | **03-Nov** | **6** | **1.42%** |
| **28-Aug** | **1** | **0.24%** |  | **19-Nov** | **6** | **1.42%** |
| **03-Sep** | **2** | **0.47%** |  | **23-Nov** | **6** | **1.42%** |
| **04-Sep** | **6** | **1.42%** |  | **05-Aug** | **5** | **1.18%** |
| **07-Sep** | **2** | **0.47%** |  | **19-Aug** | **5** | **1.18%** |
| **08-Sep** | **5** | **1.18%** |  | **08-Sep** | **5** | **1.18%** |
| **09-Sep** | **9** | **2.13%** |  | **28-Oct** | **5** | **1.18%** |
| **10-Sep** | **3** | **0.71%** |  | **30-Oct** | **5** | **1.18%** |
| **16-Sep** | **3** | **0.71%** |  | **25-Nov** | **5** | **1.18%** |
| **17-Sep** | **1** | **0.24%** |  | **10-Aug** | **4** | **0.95%** |
| **18-Sep** | **1** | **0.24%** |  | **14-Aug** | **4** | **0.95%** |
| **22-Sep** | **1** | **0.24%** |  | **21-Aug** | **4** | **0.95%** |
| **25-Sep** | **3** | **0.71%** |  | **14-Oct** | **4** | **0.95%** |
| **28-Sep** | **1** | **0.24%** |  | **16-Oct** | **4** | **0.95%** |
| **29-Sep** | **7** | **1.66%** |  | **04-Nov** | **4** | **0.95%** |
| **01-Oct** | **3** | **0.71%** |  | **24-Nov** | **4** | **0.95%** |
| **02-Oct** | **3** | **0.71%** |  | **27-Jul** | **3** | **0.71%** |
| **06-Oct** | **10** | **2.37%** |  | **29-Jul** | **3** | **0.71%** |
| **07-Oct** | **7** | **1.66%** |  | **25-Aug** | **3** | **0.71%** |
| **08-Oct** | **3** | **0.71%** |  | **26-Aug** | **3** | **0.71%** |
| **09-Oct** | **12** | **2.84%** |  | **10-Sep** | **3** | **0.71%** |
| **13-Oct** | **1** | **0.24%** |  | **16-Sep** | **3** | **0.71%** |
| **14-Oct** | **4** | **0.95%** |  | **25-Sep** | **3** | **0.71%** |
| **15-Oct** | **1** | **0.24%** |  | **01-Oct** | **3** | **0.71%** |
| **16-Oct** | **4** | **0.95%** |  | **02-Oct** | **3** | **0.71%** |
| **22-Oct** | **11** | **2.61%** |  | **08-Oct** | **3** | **0.71%** |
| **23-Oct** | **9** | **2.13%** |  | **17-Nov** | **3** | **0.71%** |
| **27-Oct** | **6** | **1.42%** |  | **22-Jul** | **2** | **0.47%** |
| **28-Oct** | **5** | **1.18%** |  | **28-Jul** | **2** | **0.47%** |
| **29-Oct** | **13** | **3.08%** |  | **06-Aug** | **2** | **0.47%** |
| **30-Oct** | **5** | **1.18%** |  | **07-Aug** | **2** | **0.47%** |
| **02-Nov** | **11** | **2.61%** |  | **17-Aug** | **2** | **0.47%** |
| **03-Nov** | **6** | **1.42%** |  | **24-Aug** | **2** | **0.47%** |
| **04-Nov** | **4** | **0.95%** |  | **27-Aug** | **2** | **0.47%** |
| **05-Nov** | **1** | **0.24%** |  | **03-Sep** | **2** | **0.47%** |
| **12-Nov** | **12** | **2.84%** |  | **07-Sep** | **2** | **0.47%** |
| **16-Nov** | **7** | **1.66%** |  | **20-Jul** | **1** | **0.24%** |
| **17-Nov** | **3** | **0.71%** |  | **28-Aug** | **1** | **0.24%** |
| **18-Nov** | **1** | **0.24%** |  | **17-Sep** | **1** | **0.24%** |
| **19-Nov** | **6** | **1.42%** |  | **18-Sep** | **1** | **0.24%** |
| **20-Nov** | **12** | **2.84%** |  | **22-Sep** | **1** | **0.24%** |
| **23-Nov** | **6** | **1.42%** |  | **28-Sep** | **1** | **0.24%** |
| **24-Nov** | **4** | **0.95%** |  | **13-Oct** | **1** | **0.24%** |
| **25-Nov** | **5** | **1.18%** |  | **15-Oct** | **1** | **0.24%** |
| **26-Nov** | **11** | **2.61%** |  | **05-Nov** | **1** | **0.24%** |
| **27-Nov** | **7** | **1.66%** |  | **18-Nov** | **1** | **0.24%** |
| **Total** | **422** | **100.00%** |  | **Total** | **422** | **100.00%** |

Between 19.67% – 24.17% of commits were completed each month - except for September, which accounted for an interestingly low 10.43%

|  |  |  |
| --- | --- | --- |
| **Month** | **Commits** | **% Total** |
| **Jul** | **102** | **24.17%** |
| **Aug** | **83** | **19.67%** |
| **Sep** | **44** | **10.43%** |
| **Oct** | **97** | **22.99%** |
| **Nov** | **96** | **22.75%** |
| **Total** | **422** | **100.00%** |

Monday was an interestingly quiet day for submissions (12.56%), with Thursday more than making up for this. Submissions were 123% more likely on this day than on Monday. The remaining three days had from 18.01% - 22.51% of all, which is of less interest.

|  |  |  |
| --- | --- | --- |
| **Day** | **Count of Day** | **% Total** |
| **Mon** | **53** | **12.56%** |
| **Tue** | **80** | **18.96%** |
| **Wed** | **76** | **18.01%** |
| **Thu** | **118** | **27.96%** |
| **Fri** | **95** | **22.51%** |
| **Total** | **422** | **100.00%** |

The commits were completed in the 16 hours between 5am and 9pm. It’s interesting to note  
that 71.09% were submitted during the 5 hours centred around 11am – 5pm, where the relatively quiet hour after 12pm is excluded. Thou hour 2 - 3pm was by far the most popular for commits. The hours 5 – 8am and 5 – 9pm were the 7 least popular, accounting for a total of 3.79% of submissions. 8 – 9am was the next least popular.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hour** | **Count** | **% Total** |  | **Hour** | **Count** | **% Total** |
| **05** | **1** | **0.24%** |  | **14** | **96** | **22.75%** |
| **06** | **3** | **0.71%** |  | **11** | **57** | **13.51%** |
| **07** | **2** | **0.47%** |  | **13** | **56** | **13.27%** |
| **08** | **12** | **2.84%** |  | **15** | **52** | **12.32%** |
| **09** | **33** | **7.82%** |  | **16** | **39** | **9.24%** |
| **10** | **30** | **7.11%** |  | **9** | **33** | **7.82%** |
| **11** | **57** | **13.51%** |  | **12** | **31** | **7.35%** |
| **12** | **31** | **7.35%** |  | **10** | **30** | **7.11%** |
| **13** | **56** | **13.27%** |  | **8** | **12** | **2.84%** |
| **14** | **96** | **22.75%** |  | **17** | **4** | **0.95%** |
| **15** | **52** | **12.32%** |  | **6** | **3** | **0.71%** |
| **16** | **39** | **9.24%** |  | **20** | **3** | **0.71%** |
| **17** | **4** | **0.95%** |  | **7** | **2** | **0.47%** |
| **18** | **2** | **0.47%** |  | **18** | **2** | **0.47%** |
| **19** | **1** | **0.24%** |  | **5** | **1** | **0.24%** |
| **20** | **3** | **0.71%** |  | **19** | **1** | **0.24%** |
| **Total** | **422** | **100.00%** |  | **Total** | **422** | **100.00%** |

Commits by Thomas

Thomas completed 45.26% of submission and this is used to calculate his expected submissions  
for each month, given the overall amounts. We see he submitted the expected number of 20 for September, but favoured July and October over his peers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Month** | **Commits** | **% Total** | **Exp. Thomas** | **Thomas** | **% Exp** |
| **Jul** | **102** | **24.17%** | **46** | **50** | **108.70%** |
| **Aug** | **83** | **19.67%** | **38** | **30** | **78.95%** |
| **Sep** | **44** | **10.43%** | **20** | **20** | **100.00%** |
| **Oct** | **97** | **22.99%** | **44** | **57** | **129.55%** |
| **Nov** | **96** | **22.75%** | **43** | **34** | **79.07%** |
| **Total** | **422** | **100.00%** | **191** | **191** | **100.00%** |

Similarly, we see that Thomas had a greater preference for Friday than Monday relative to the   
group as a whole (71%).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Day** | **Commits** | **% Total** | **Exp. Thomas** | **Thomas** | **% Exp.** |
| **2** | **53** | **12.56%** | **24** | **18** | **75.00%** |
| **3** | **80** | **18.96%** | **36** | **37** | **102.78%** |
| **4** | **76** | **18.01%** | **34** | **35** | **102.94%** |
| **5** | **118** | **27.96%** | **53** | **46** | **86.79%** |
| **6** | **95** | **22.51%** | **43** | **55** | **127.91%** |
| **Grand Total** | **422** | **100.00%** | **191** | **191** | **100.00%** |

All of Thomas’ commits were made between 8am – 5pm. A much greater preference for  
4 – 5pm relative to his peers is observed, with submissions up over 25% from the previous hour, when -25% was expected.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Hour** | **Count** | **% Total** | **Exp. Thomas** | **Thomas** | **% Exp.** |
| **5** | **1** | **0.24%** | **0** | **0** | **0%** |
| **6** | **3** | **0.71%** | **1** | **0** | **0%** |
| **7** | **2** | **0.47%** | **1** | **0** | **0%** |
| **8** | **12** | **2.84%** | **5** | **7** | **140.00%** |
| **9** | **33** | **7.82%** | **15** | **25** | **166.67%** |
| **10** | **30** | **7.11%** | **14** | **15** | **107.14%** |
| **11** | **57** | **13.51%** | **26** | **30** | **115.38%** |
| **12** | **31** | **7.35%** | **14** | **12** | **85.71%** |
| **13** | **56** | **13.27%** | **25** | **13** | **52.00%** |
| **14** | **96** | **22.75%** | **43** | **37** | **86.05%** |
| **15** | **52** | **12.32%** | **24** | **23** | **95.83%** |
| **16** | **39** | **9.24%** | **18** | **29** | **161.11%** |
| **17** | **4** | **0.95%** | **2** | **0** | **0%** |
| **18** | **2** | **0.47%** | **1** | **0** | **0%** |
| **19** | **1** | **0.24%** | **0** | **0** | **0%** |
| **20** | **3** | **0.71%** | **1** | **0** | **0%** |
| **Total** | **422** | **100.00%** | **190** | **191** | **100.53%** |

Disclaimer: As Thomas made over 45% of commits, it would be more useful to compare his data  
to the rest of the group rather than the group as a whole. Observed contrasts are understated.

Commits by Line count

555 lines of comments were submitted over 422 commits, over 82% of them being just one line.

|  |  |  |  |
| --- | --- | --- | --- |
| **No. of Lines** | **Occurrences** | **% Total** | **Total Lines** |
| **1** | **348** | **82.46%** | **348** |
| **2** | **45** | **10.66%** | **90** |
| **3** | **11** | **2.61%** | **33** |
| **4** | **9** | **2.13%** | **36** |
| **5** | **7** | **1.66%** | **35** |
| **6** | **1** | **0.24%** | **6** |
| **7** | **1** | **0.24%** | **7** |
| **Total** | **422** | **100.00%** | **555** |