Student ID: 10011952

Student Name: Niamh Lynagh

Three interesting pieces of statistical information from the dataset. Discuss.

Describe the dataset

The dataset used for analysis consists of 5000 lines. It captures the script generated by the activity of numerous authors (or developers), that are submitting code changes to a shared repository of files. The script details the authors activity over a period of time. Activities recorded in the file is date stamped and all activities relate to committing to the shared repository. Each 'commit' is separated by a line of '-'.
In the script, you will see that there are 72 '-' separating each commit. When these are stripped out there are 422 commits in the line dataset. The information on the activities contain a revision number, author name, the date of the commit, the number of lines in the commit text, the files that were affected by the commit (and how they were affected; add, modify, delete) and a comment by the author.

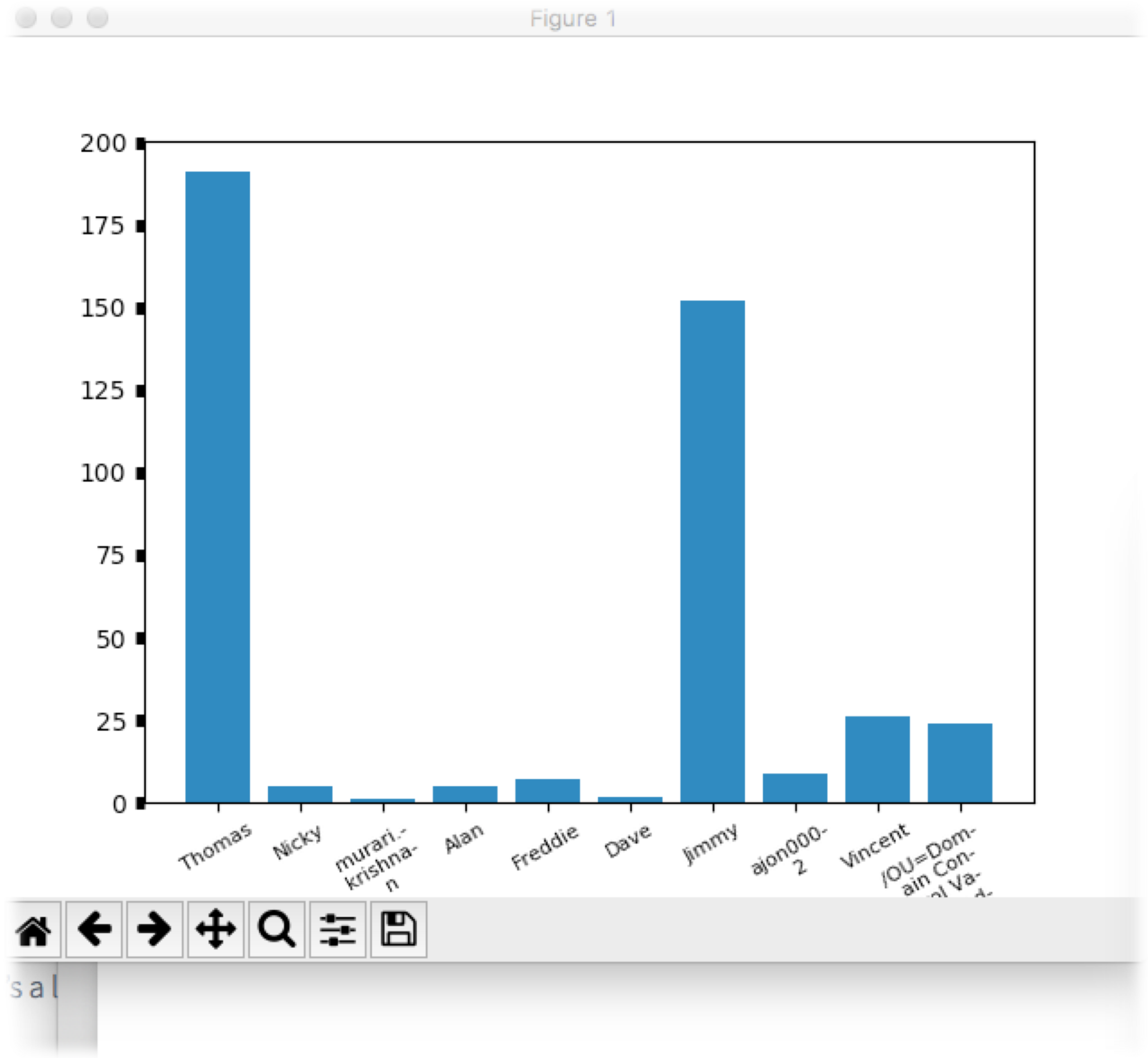
Describe the three pieces of interestingness

The code that isolates the three pieces of interestingness is contained in the commits.py file. The tests for the commits class are contained in a separate file called testCommits.py. The three pieces of interestingness are;

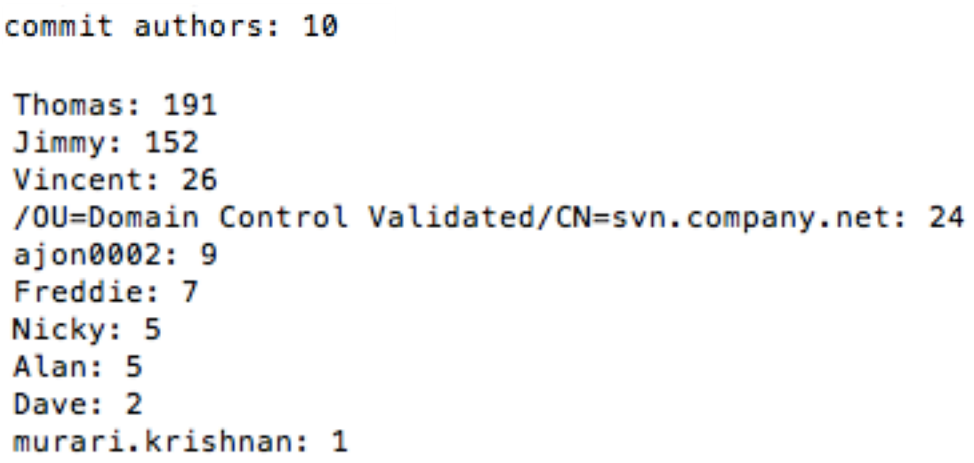
1. The names of the authors and the number of times they committed to the repository
2. The number of times that files are added, modified or deleted
3. The date of the first commit and the date of the last commit – and the number of days in between.

The authors add the commits to the file repository. This piece of interestingness lists the names of the authors and the number of times they have committed. It is an indicator of how actively they committed but does not provide information on the quantity or quality of the code committed. The list and graph were generated by the commits.py file.

The authors plotted

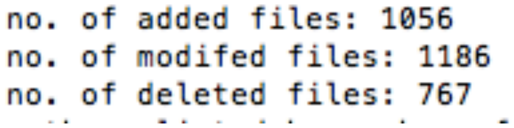


The total number of authors and their commits.



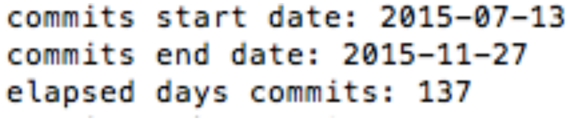
```
commit authors: 10

 Thomas: 191
 Jimmy: 152
 Vincent: 26
 /OU=Domain Control Validated/CN=svn.company.net: 24
 ajon0002: 9
 Freddie: 7
 Nicky: 5
 Alan: 5
 Dave: 2
 murari.krishnan: 1
```

The second piece of interestingness required interrogation of the changes in each commit to identify the number of files that are tagged with an ‘A’, ‘M’ or ‘D’ to represent how the file was treated by the commit.



```
no. of added files: 1056
no. of modifed files: 1186
no. of deleted files: 767
```

The third piece of interestingness interrogates the date stamp on the commits; the date of the first commit and the date of the last commit is extracted and the number of days in between are calculated. The calculation does include weekends and week days.



```
commits start date: 2015-07-13
commits end date: 2015-11-27
elapsed days commits: 137
```

What do the pieces of interestingness tell us?

The pieces of interestingness capture the baseline information about the dataset, this can direct the reader when deciding what additional analysis is needed to generate new pieces of interestingness from the dataset. The number of commits, the number authors, the number of times each author committed in the dataset, the first date and the last date of a commit in the dataset, the elapsed days in between.

Discuss other possible pieces of interestingness that could be interrogated to generate greater insights.

Other pieces of interestingness could include; identifying the dates that the authors committed, how many add, modify or delete actions did the authors commit, did any author work on weekends as well as week days? For instance, why does Thomas have 191 commits and Freddie only 7? It would be interesting to check the date that Freddie made his first commit, – maybe he joined the team later than Thomas or does he not commit his code regularly?

Further analysis would generate additional interesting information. What day of the week are authors most likely to commit their code? What do authors do when they commit – is it primarily adding, modifying or deleting code? Additionally, the commit comment could be interrogated to assess the quality of the comment and if they follow best practice – does it describe what the new code does or does it describe what the author did?