B8IT105 Programming for Big Data November 2017

Student: Emmet Feeney 10360392

**CA4 - Perform Analysis on a 5000 line dataset**

Assignment task:

* Transform a large dataset in text format - over 5000 lines of text
* Scrub (clean) the data and place it into the relevant holder/container objects
* Analyse the 422 objects that are in a list and come up with 3 interesting statistical pieces of information for this dataset with supporting evidence of "interestingness”
* Code for calculating the analysis should be documented and tested
* Test should be in a separate file runnable from the command line
* Statistical analytics conclusions should be in a word document explaining in approximately 500 words the information that you have gleamed from the dataset
* Submit code via github along with all documentation and tests

1. **Deconstructing the large dataset from log file into a csv file**
2. I opened the changes\_python.log file and saved it as a normal text file in NotePad++ under file name **changes\_python.txt**



1. Using Python and commit method (process\_changes.py) , I cleaned the large dataset of 5255 lines to 422 lines and output this to generate a csv file (changes.csv).

This program read in the input file (changes\_python.txt), identifies the separator between blocks of data and then picks specific values within the block to read in.

The following 6 variables were selected to write:

"revision,author,date,time,number\_of\_lines,comment”

Below are the python file (process\_changes.py) and the output csv file (changes.csv):



Below is the response received when ran the process\_changes.py program:

5255

422

{'comment': ['Renamed folder to the correct name'], 'number\_of\_lines': 1, 'author': 'Thomas', 'changed\_path': ['A /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client/res/drawable-xxxhdpi (from /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client/res/drawablw-xxxhdpi:1551688)', 'D /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client/res/drawablw-xxxhdpi', 'A /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client-bt/res/drawable-xxxhdpi (from /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client-bt/res/drawablw-xxxhdpi:1551922)', 'D /cloud/personal/client-international/android/branches/android-15.2-solutions/clients/client-bt/res/drawablw-xxxhdpi'], 'time': '16:57:44', 'date': '2015-11-27', 'revision': 'r1551925'}

Thomas

1. **Testing**

I also used the **test\_simple.py** file to do some tests on the data:

* Test that there were 5,255 lines in the file.
* Test that there were 422 objects.
* Testing to verify which user made different commits.

****

1. **Data Analytics**

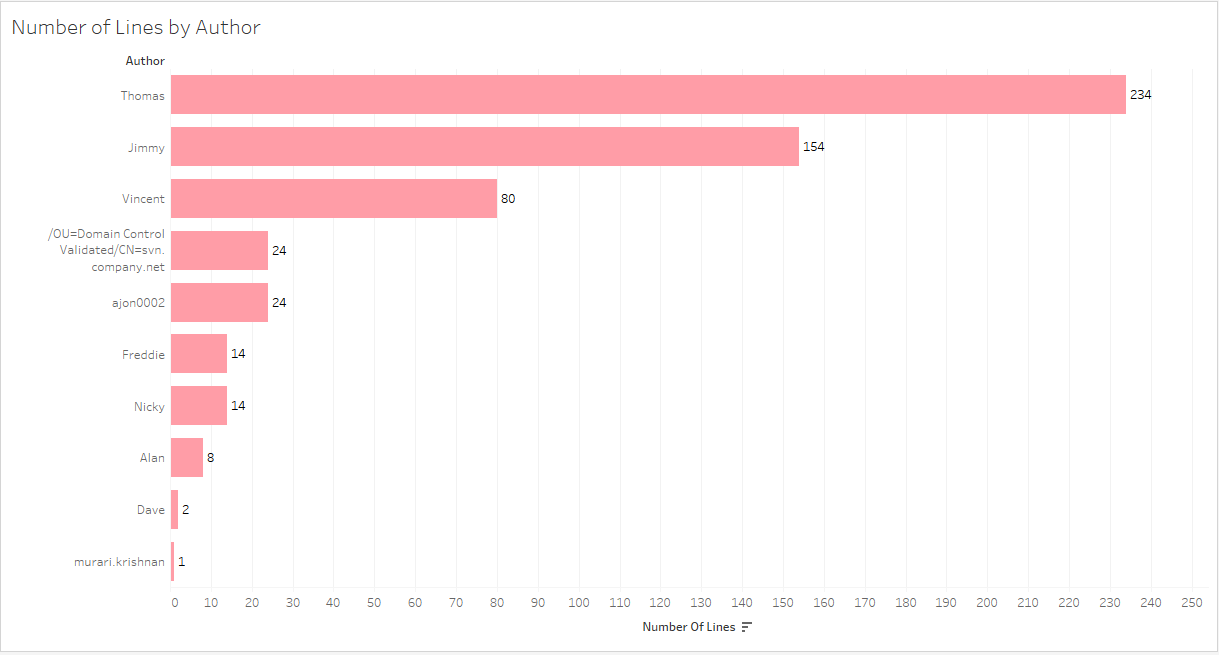
I uploaded the csv file “changes.csv” to Tableau software and created visualisations to give 3 interesting statistical insights from this data.

1. **Author Performance**

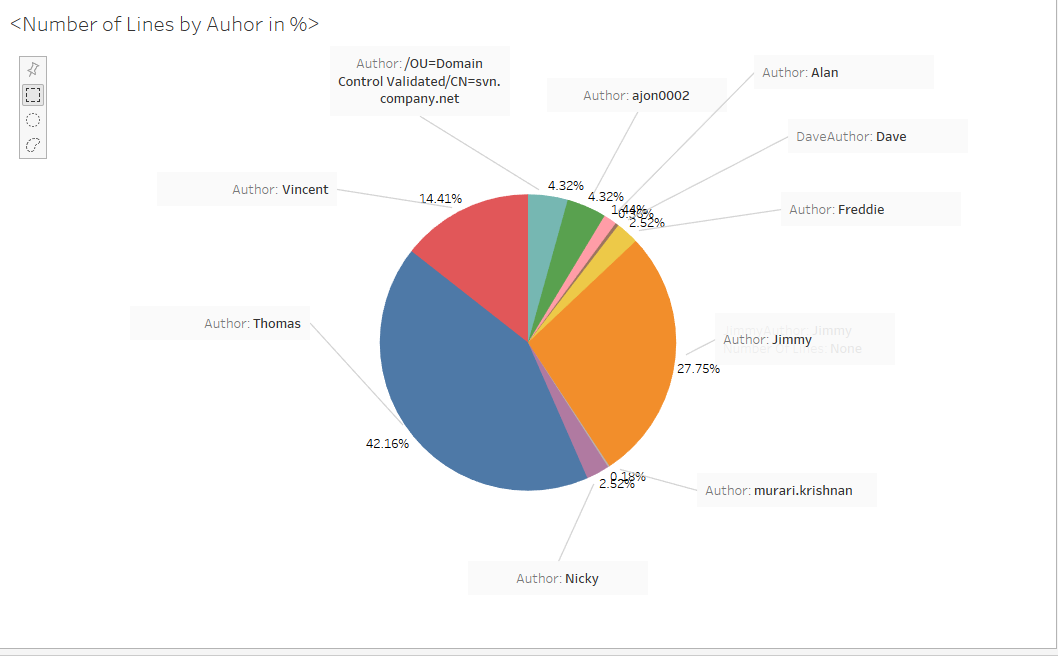
First I looked at the overall author performance in terms of total lines written.From the bar chart, we can see that there a high range in the number of lines written by each author, ranging from 234 lines by Thomas to just 1 line by murari.krishnan.

We determine that one of the authors is actually system generated lines

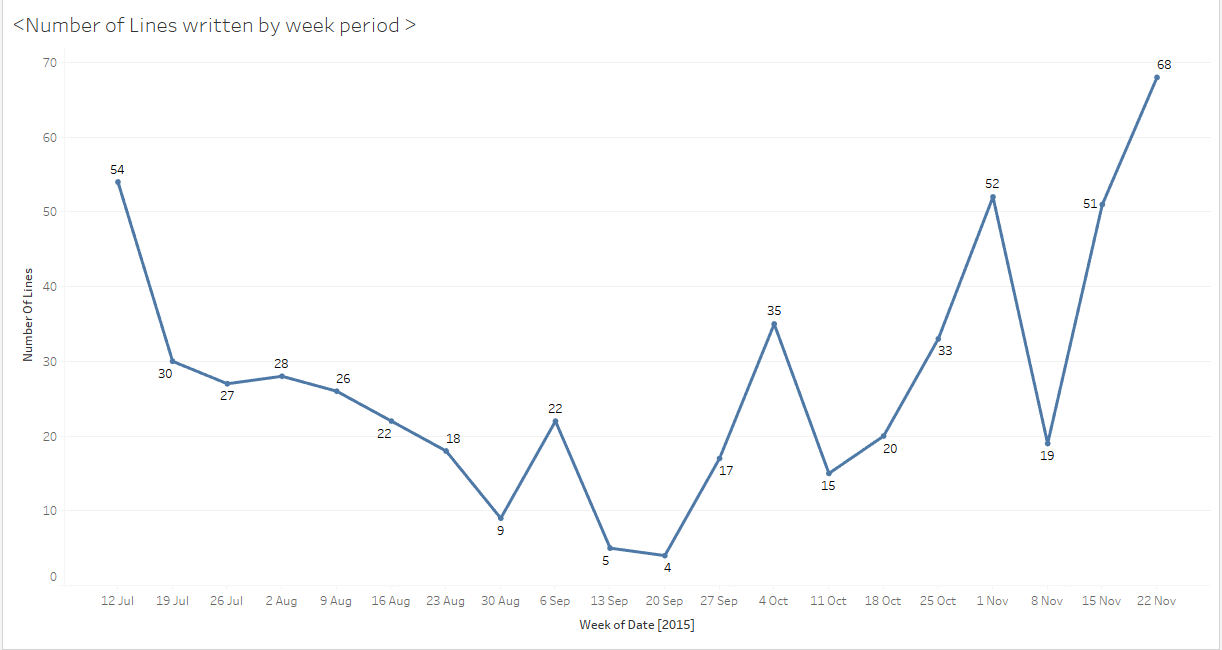
**/OU=Domain Control Validated/CN=svn.company.net**



You can also see the same dataset in the Pie Chart below, this time the number of lines by author is displayed as a percentage.

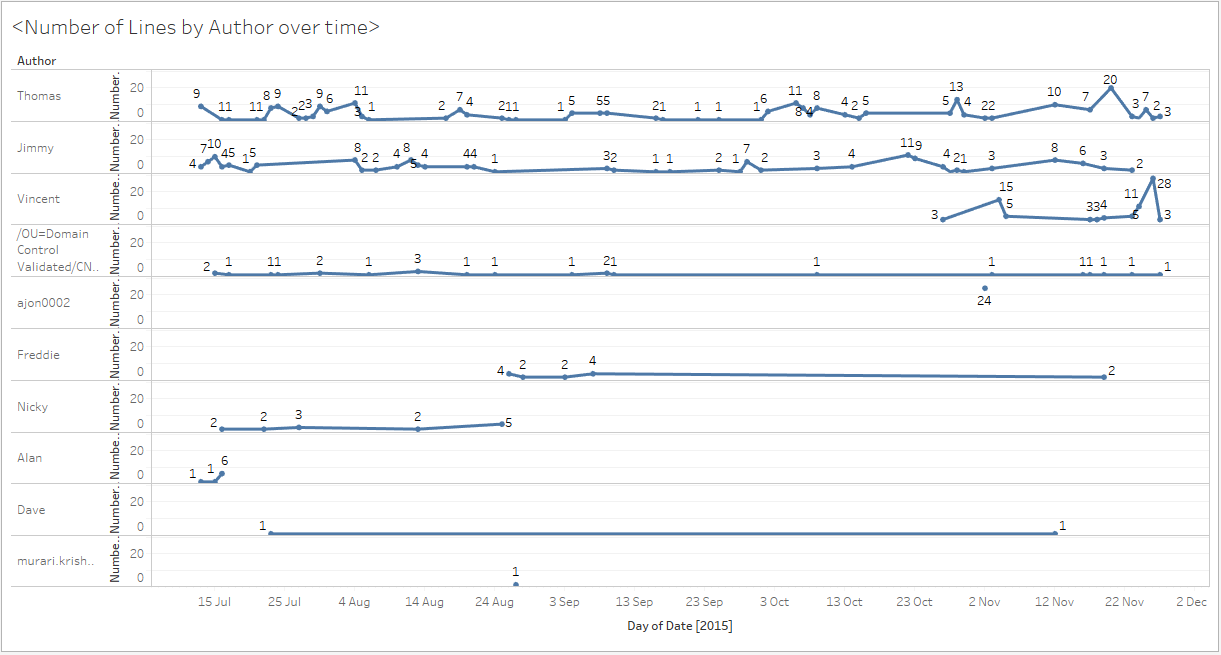


1. **Activity over time**

Next, I created line charts to visualise the activity over the time period. We can see the fluctuations in activity between 12 July and week ending 28 November. There is a noticeable downward trend in activity after the first week, with particularly low activity in week commencing 30 Aug, 13 Sep, 20 Sep. Similarly, there a spikes in activity, particularly week commencing 1 Nov, 15 Nov and 22 Nov. 

I also used line graph to visualise author activity weekly to see any interesting observations.

The highest by any author in a given week was 20 by Thomas on w/c 20th Nov.Vincent didn’t do any activity until late October. Nicky hasn’t any activity since end Aug, whilst Freddie only started activity since end Aug so perhaps Freddie replaced Nicky.

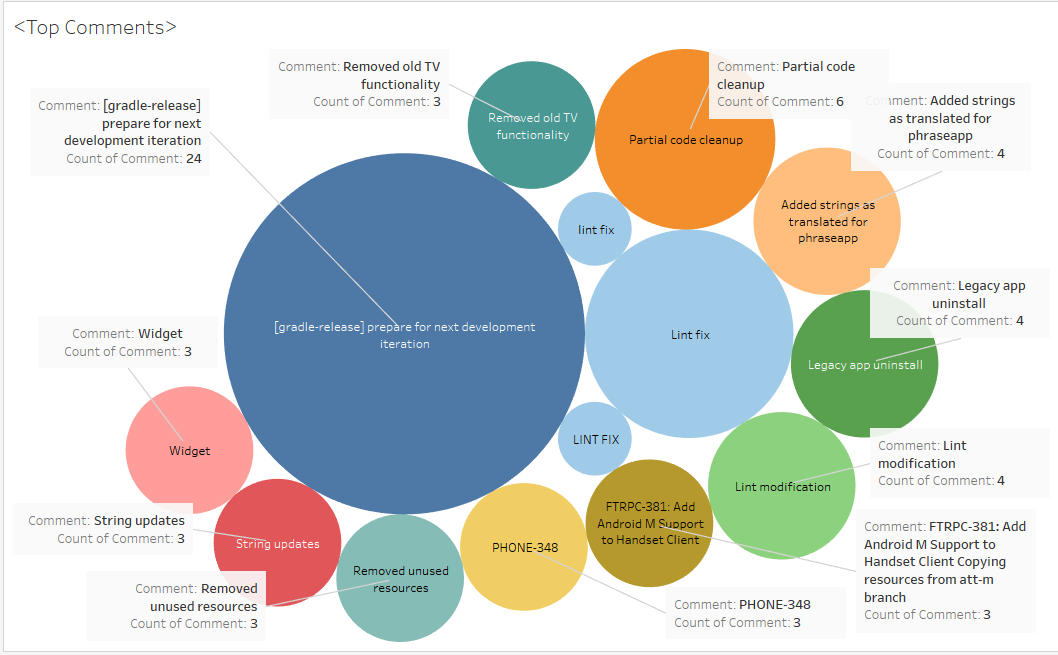


1. **Top Comments**

I also used the packed bubbles visualisation chart to show the top comments. There were 345 unique comments. I focused on the top ones only (those with a count of 3 or more).

You can see that the top 3 are :

1. “[gradle-release] prepare for next development iteration”, appearing 24 times
2. “Lint fix”, with 10 appearances
3. “Partial code clean up”, which appears on 6 occasions



**Concluding remarks**

The log file detail changes made to a system/ database by a number of users in 2015 between 12 July and week ending 28 November.

I copied the initial log file to text format and then parsed the file using Python program and output the 422 objects to a csv file. I used the unit test function in python to test the data.

I used Tableau to analyse the data in the csv file in order to identify 3 interesting pieces of statistical information.

This document and attached files are uploaded to DBS Moodle and also to Github.

<https://github.com/10360392>