## 

CA4 – Performance Analysis on a 5000 Line Dataset

# Task :

Assignment 4 is based on transforming a large dataset in text format - over 5000 lines of text.

You will need to scrub (clean) the data and place it into the relevant holder/container objects.

Once in these objects you will see that there are 422 different sets of commit objects.

So your task will be to analyse these 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'.

Overview -

Input : Log file for analysis

Process: Read log file stripping out spaces and compile data regarding the following variables - revision,author,date,time,number\_of\_lines,comment

Output: Output the cleaned data to csv file for statistical analysis

# Statistical Analysis :

Variable Name: author –

Measure Count

count 10.00

mean 42.20

std 69.29

min 1.00

25% 5.00

50% 8.00

75% 25.50

max 191.00

Interesting statistics – author

- there are 10 authors

- the average commits per author was 42 commits but noticeably there was high variance as the standard deviation was approx. 69 whilst the minimum number of commits by an author (murari.krishnan ) was 1 and the leading author (Thomas) had the maximum number of commits, 191.

See Python generated table and graph below for further details of author commit activity -

Thomas 191

Jimmy 152

Vincent 26

/OU=Domain Control Validated/CN=svn.company.net 24

ajon0002 9

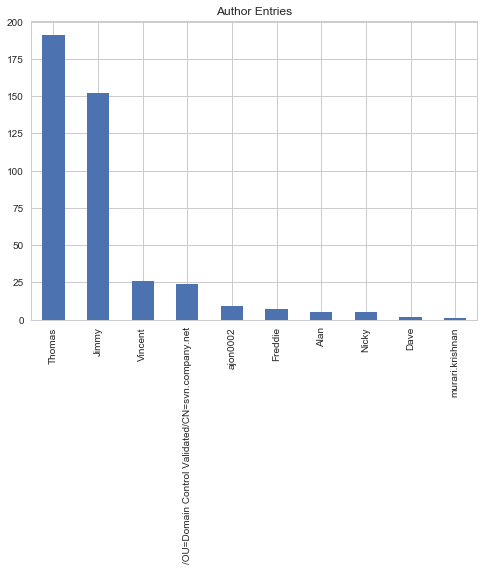
Freddie 7

Alan 5

Nicky 5

Dave 2

murari.krishnan 1



Variable Name: time

See Python generated table and graph below for further time Slot commit details -

# Time Slot Number of Commits

05:00:01 to 07:00:00: 4

19:00:01 to 21:00:00: 4

17:00:01 to 19:00:00: 6

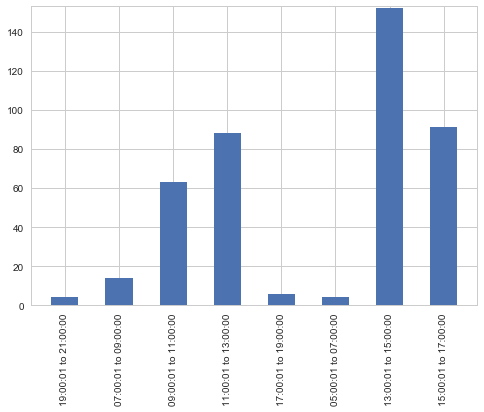
07:00:01 to 09:00:00: 14

09:00:01 to 11:00:00: 63

11:00:01 to 13:00:00: 88

15:00:01 to 17:00:00: 91

13:00:01 to 15:00:00: 152



Interesting statistics – time –

Time was divided into bins of 2 hourly slots

The plot above indicates that the busiest time slot by far for authors commits was between 13:00 and 15:00

The same number of commits (4) were completed by the early birds between 5:00 and 7:00 a.m. and the night owls between 19:00 and 21:00 p.m.

# Misc. Interesting Statistics –

# Author who committed most changes – also worked normal business hours

Thomas was the author who committed most changes. He had the by far the highest number of input changes committed , at 191, yet he never started to commit inputs before 8:30 a.m. (his earliest input was at 08:38:12 on 28/07/2015) and he never worked later than 5:00 p.m. (the latest he worked was 16:57:44 on 27/11/2015).

If we take it lunch time is between 1:00 p.m. to 2:00 p.m. it is also notable that he very rarely worked during lunch as he only committed 12 of his 191 commits (or approx. 6%) during this time period.

## Commit count per month

Total Number of commits input from: July 2015 – November 2015 = 422

Commits per Month : (Note - Graph generated in Excel)

July: 102 or 24.17 % ; Aug: 83 or 19.67% ; Sep: 44 or 10.43 %; Oct:97 or 22.99%; Nov: 96 or 22.75%