Student Name: Jean O’ Sullivan

Student Number: 10352794

Module Code: B8IT105

Module Title: B8IT105 Programming for Big Data

CA: CA4 - Perform Analysis on a 5000-line dataset

This dataset is a log of all the commits made to a remote repository over a period. Each commit log stores the following information 'revision', 'author', 'date', 'line count, ‘changes ‘and 'comment'. The 3 interesting statistical pieces of information I found about dataset were the author, comments and dates.

**Author**

Many authors were working on this project, my analysis showed 10 different users in total committed changes to the repository. The names of the authors were as follows: 'Thomas’, 'Vincent' '/OU=Domain Control Validated/CN=svn.company.net',

'Jimmy', 'Freddie', 'Dave’, 'ajon0002’, 'murari. krishnan’, ‘Nicky’, and 'Alan’.

The author who made the most changes to the repository was Thomas with a 191 commit where the Mean(average) was 42.2. That is a total of 148.8 more commits than the average. The mean, also referred to by statisticians as the *average,* is the most common statistic used to measure the center of a numerical data set.

The mean may not be a fair representation of the data, because the average is easily influenced by outliers (very small or large values in the data set that are not typical).

For instance, the author with the least commits was 'murari. krishnan' with just 1 commit. That is a total of 41.2 less than the Mean(average).

**Comments**

The most number of lines per comment was 7 and this was by the author Freddie. Where the Mean(average) number of comments per line was 1.3. That is a total of 5.7 lines more than the average comment in the repository.

Again, you can see that the mean was not fairly represented as the average was influence by outliers. As you can see from the data below the median was just 1. That is a total difference of 0.3 between the mean and median. The total difference between the median and the maximum amount of lines is 6.

**Dates**

The total number of dates where changes were made to the repository is 76. The most commits made on a day was 19, The date that most of the commits were made was 2015, 8, 4. There is no data to suggest that there is any significance to this date. One could suggest that perhaps a deadline of some sort had to be met as there were so many comment made on this day.

**Results**

List of authors: ['Thomas' 'Vincent' '/OU=Domain Control Validated/CN=svn.company.net'

'Jimmy' 'Freddie' 'Dave' 'ajon0002' 'murari. krishnan' 'Nicky' 'Alan']

No of authors: 10

Commits per Author

author

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

Alan 5

Dave 2

Freddie 7

Jimmy 152

Nicky 5

Thomas 191

Vincent 26

ajon0002 9

murari. krishnan 1

dtype: int64

Author most commits: Thomas

Author least commits: murari. krishnan

Maximum commits by author: 191

Commits mean: 42.2

Commits median: 8.0

Mean number of comment lines: 1.31516587678

Median number of comment lines: 1.0

Most number of lines per comment: 7

Least number of lines per comment: 1

Number of dates: 76

Most commits on a day: 19

Date of most commits: (2015, 8, 4)