**Student No:** 10393871 **Assignment:** Programming for Big Data - Continuous Assessment 4

**Assignment Detail:** Download Commit Log.txt file

Undertake an analysis of the Commit Log

Identify 3 statistical pieces of information with high degree of “interestingness”

After identifying the Commit Log file, I had to convert it to a csv file. After doing so it required some additional parsing as comment lines had not transposed correctly, and extended over additional columns. The comment detail did not appear overly important for this exercise - rather their 'interestingness' is deemed low. I imported columns A-F from 'changes.csv' into python for analysis.

After importing the log file as a dataframe using Pandas, I noted:

1. Authors
2. Dates
3. Times
4. Revisions

I grouped the data by Author and observed my first analysis point:

* 2 staff members - Thomas and Jimmy - are highly active in making revisions. I can also see that Vincent has just over 25 commits while everyone else has less than 10 entries.

I then had to manipulate the data to analyse the dates and times. The data had imported as ‘object’ rather than timestamp or date/time. As such, I stripped the date and time data and separated it out into year month day hours minutes and seconds.

* I then analysed the ‘authors’ or employees with the dates observed for when commits were logged. There is a number of employees on the log however only Thomas, Jimmy and Dave were logging commits over the entire period. There may be an issue with staff turnover within the organisation which appeared to impact the organisation’s productivity in September.

I then analysed the concentration of commits to hours during the day. Very few commits are occurring outside of the normal working hours.

From the above analysis of concentration of commits throughout the workday, we can see:

* Vincent appears to be logging the longest days, with commits occurring from early morning through to late evening.
* Thomas and Jimmy having logged the most commits over the period, are doing so during the standard working day of 8am - 6pm. They appear to be the most efficient of the staff.
* Dave's commit history indicates he starts after 10am and finishes around 7pm which is not unusal in itself however we know from previous analysis he has only 2 commits logged for the review period so we cannot perform a full analysis of his daily productivity yet. Dave may have some issues with his time management or might not be involved in logging commits as part of his normal job. Further information on Dave would be required to arrive at any conclusion.