**Dublin Business School**

**Course:** Data Analytics

**Module:** Programming For Big Data

**Student Name:** Ciaran Dunne

**Student Number:** 10393193

**Task Name** – Perform Analysis on 5000 line dataset

**Context**

* 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'
* You code for calculating the analysis should be documented and tested.
* Test should be in a separate file runnable from the command line.
* Your statistical analytics conclusions should be in a word document explaining in approximately 500 words the information that you have gleamed from the dataset.
* You will be required to submit your code via github along with all documentation and tests.

**GitHub Link** - <https://github.com/ciarandunne/Data-Analytics-DBS-Programming-For-Big-Data/tree/master>

The analysis took 3 main steps:

1. Gather the Data

From a text file which contained logs of commits to GitHub, structure the data into a useable format, in this case a csv file. This was done using Python and accompanying Python testing. The output was a 422 line csv file that can be used to analyse activity around GitHub commits. This could be done using a number of different programmes (eg. Spyder, Jupyter, Excel, Tableau), but for the purpose of this task I used Jypyter Notebook. By importing the csv file into Jupyter Notebook, we were able to start the analysis.

1. Prepare the Data

A number of changes needed to be made to prepare the data for analysis. This was a requirement for identifying “interestingness” using different approaches to visual analysis. This included:

* Make sure the csv is imported as a Dataframe
* Change unreadable names
* Restructure (separate) dates and times to support analysis
* Add day names to the Dataframe
* Change data types within the Dataframe for analysis

1. Get Insights

As part of the analysis I looked at the data from 6 different perspectives:

1. Comment detail per commit (averages)
2. Total commit volumes per author
3. Number of revisions made per month to understand peak and tough activity
4. Timing of commits (hour of the day)
5. Activity by day of the week
6. Activity outside of typical working hours (ie. 9am - 6pm)

A number of the above items (eg. levels of comment detail and weekend commits) while worth looking into, did not provide “interestingness” as activities were generally consistent. There were however three main areas which I found interesting that came out of the analysis:

**Interesting statistical pieces of information**

1. The team don't do much work outside of typical working hours.
   * There were no commits made on weekends.
   * There were a total of 26 commits outside of standard working hours over the 5 months (18 before 9am, and 8 after 5.30pm) – which in my opinion is quite low.

**Comment:** This might be a bit harsh, but if everyone in the team can get everything they need done between the hours of 9-5.30pm, Monday to Friday, over a period of 5 months, do they have enough work to do? Could there be an opportunity to increase the productivity of the team, or maybe identify some value add activities that might support performance improvement in other ways.

1. The volume of team member commits are not consistent across the team
   * There are some team members who are doing considerably more commits that the rest of the team.
   * Thomas and Jimmy make the vast majority of commits when compared to other team members.

**Comment:** Without understanding team member’s roles it’s hard to understand exactly what this indicates. However, if we assume that all team members have the same role, it is obvious that two team members are lifting heavier loads than the rest of the team. If there is a performance review process in place, this information could be used as feedback, and the analysis could be used to support cases for pay increases or bonuses. There may be a case where some team members have recently joined, or were on annual leave, so this could be looked at in more detail.

1. There are some peaks and troughs in terms of commit activity
   * The vast majority of commits are made in the middle of the day at around 2pm.
   * There is an even enough spread of commits over the days of the week
   * Each month had relatively consistent levels of activity, however, September had less than half the volume of commits as other months.

**Comment:** Depending on how the organisation is performing, this information could be used from an operational perspective. If there are operational constraints, the team could look to distribute commits more over the course of the day and avoid the 2pm bottleneck. From the perspective of monthly activity, there was significantly less activity in September - this might be worth looking into if it is already not understood. There could have been issues that needed to be resolved, and could be avoided in the future, or it could be down to planning. By looking at these peaks and troughs of activity, there might be opportunities to improve operational performance.