

**B8IT105 Programming for Big Data**

**CA4 – Analysis on Dataset**

*Student: Conor Geary*

*Student Number: 10337709*

*GitHub link:* [*https://github.com/ballyduffer/B8IT105\_CA4\_10337709*](https://github.com/ballyduffer/B8IT105_CA4_10337709)

**Assignment Brief:**

*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.*

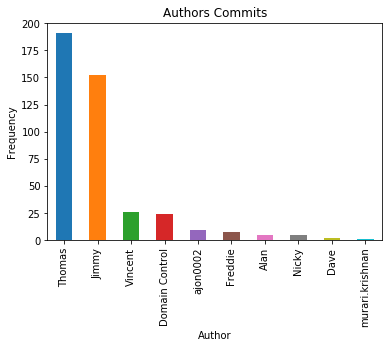
*You will be required to submit your code via github along with all documentation and tests.*

Workings:

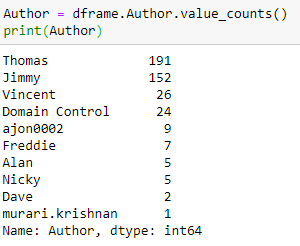
The text file the data was saved on was first read into python. This was then delimited to sort data into columns and exported as a csv. This was then analysed using Jupyter notebooks. Using this I transformed the data into a data frame and made some alterations to it. This included changing author names and adding columns for day name and hours. I then analysed this data on 3 different aspects; Revisions based on Authors, Days of the week and time. This analysis can be seen below. All workings and testing files can be found in the accompanying GitHub repository and zip file uploaded to moodle.

**Analysis:**

1. The graph below shows the number of commits made by each author.



From the graph it can be seen the Thomas made the most commits over the time period. 191 commits represents roughly 45% of the dataset and is over 7 times the number of commits by the 3rd largest contributor Vincent. Jimmy had the second largest amount with 152 - ~36% of all commits made. This highlights how Thomas and Jimmy are producing significantly more work than everyone else combined. Vincent, in 3rd had made only 2 more commits than the system generated “Domain Control”.

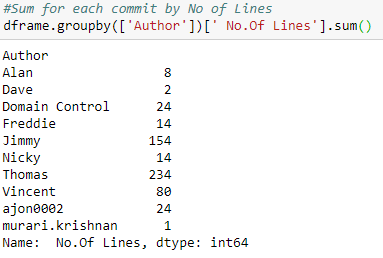


I also grouped Authors & Dates to see how many revisions were made by each per day (See Jupyter notebook for full table).



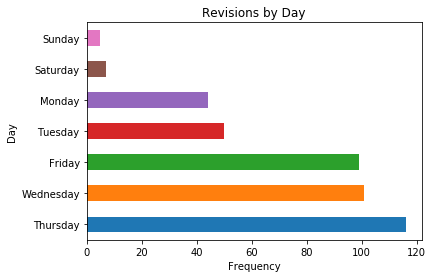


Interesting observations from this were that ajon0002, Jimmy, and murari made all their commits on one day. Alan, Dave and Freddie also made all the commits over a small number of days. This indicates these workers wouldn’t generally be working on this sort of stuff – hence why Thomas and Jimmy have such a high number of commits. This is confirmed by the number of lines in each commit.

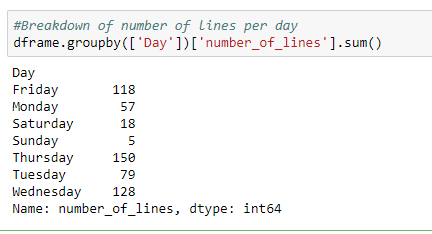


This indicates that Jimmy, Thomas and Vincent appear to be the main workers on this team.

1. The graph below shows number of commits by day:

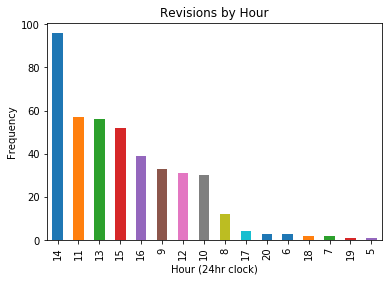


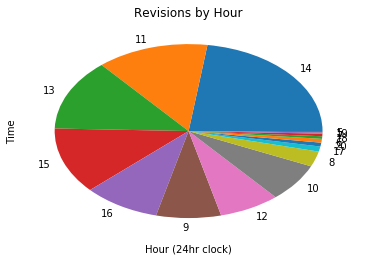
The graph shows how Thursday is the most productive day of the week for commits. This makes sense as workers would try to get as much done before the weekend. The graph shows that the amounts of commits is at least half of those at the end of the week, for Mondays and Tuesdays. This shows the workers are very unproductive at the weeks beginning. This data can be split further to check the number of lines of code are on average written a day:



I thought there may be a difference in days order here (i.e. less revisions on one day, but more lines of code per revision). However the above table disputes this, and instead follows the same form of revisions per day. This means Thursday has on average the most revisions, and the most lines of code.

1. The graph below shows commits done by time of day:





I originally graphed this data using a pie chart, but felt the bar chart was easier to read. I showed both for comparison. These graphs highlight how most work is done between 14:00 and 15:00 hours. This may show how workers try to get work done before clocking out for the evening. I would class the most productive time period as being from 11:00 to 16:00 hours. The 2nd and 3rd most productive hours are 11:00 and 13:00. This seems to indicate people trying to get work done before and after lunch breaks. This is again confirmed with the drop off in commits at 12:00 hours. There are no commits recorded from 20:00 to 05:00 hours. This would be expected. Some commits have been recorded past 5:00 & 6:00 hours however, showing some workers have come into work early. This could be due to issues, finishing work from the day previous or indicate the employee’s dedication to work. The hours from 8:00 to 10:00 appear to be the least productive of normal working hours; showing a slow start to the day for most workers.

Conclusion:

I have analysed the dataset on 3 different parameters gaining different insights from each. Thomas is the most productive of the workers, Thursday is on average the busiest day of the week, and most commits are performed from the hours of 14:00 to 15:00.