

Programming for Big Data –

CA04 Analysis of Dataset

**Student Name: Colum Kenny**Student No: 10330083  
May 2017  
Lecturer: Darren Redmond

# Understanding of Tasks

The purpose of this assignment is to:

1. Import a dataset using Python;
2. Cleanse the data and store each commit as a list;
3. Perform an analysis of cleansed data (using Python and R);
4. Test code in a separate file.

The files used for the analysis are:

1. simple\_copy.py;
2. test\_simple.py;
3. rscript\_data\_processing.R;
4. days.csv;
5. week\_days.csv.

# Analysis of Data

## Author Analysis

The purpose of this analysis is to determine how many commits each person made.. A number of steps were taken:

1. A blank list called “authors” was created;
2. Author names were stripped from each commit using a variable called “author”;
3. Using a while loop, a search was completed of the 422 names and, if not already in the “authors” list, appended to that list.

Each name in the list was then manually inputted using a “raw\_input” statement in Python (alternative analysis method chosen over running in R studio). Based on the input, Python searched through the original commits and counted each occurrence, returning the count and the percentage of total commits by that author.

Commits by person is detailed in Table 1.

Table 1: Commits by Person

|  |  |  |
| --- | --- | --- |
| Name | No. of Commits | % of Total Commits |
| Thomas | 191 | 45.26 |
| Jimmy | 152 | 36.02 |
| Vincent | 26 | 6.16 |
| ajon0002 | 9 | 2.13 |
| Freddie | 7 | 1.66 |
| Alan | 5 | 1.18 |
| Nicky | 5 | 1.18 |
| Dave | 2 | 0.47 |
| murari.krishnan | 1 | 0.24 |

The largest contributors were Thomas and Jimmy, with a combined no. of commits > 80%. Vincent was the 3rd highest with 26 commits.

Interestingly, “/OU=Domain Control Validated/CN=svn.company.net” returns 25 commits (5.69% of total). This domain validated certificate does not allow us to know the identity of the person making the changes.

## Dates Analysis

The first commit was completed on Monday, 13th July 2015. The final commit was completed on Friday, 27th November 2015.

In a similar set of steps to above, a blank list called “dates” was created, and date of commits (in format YYYY-MM-DD) was sliced from each commit. Each date was added to the “dates” list and sorted.

Analysis of dates was completed using R Studio. The “dates” list was written to a CSV file; titled “days.csv”.

The CSV file was imported into R Studio and analysis undertaken. The R script was named “rscript\_date\_processing.r”). A summary of the data provided by R is detailed in Figure 1.

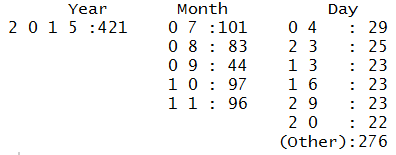


Figure 1: Summary of dates.csv file by R Studio

From the table, we know that most commits were completed in July (101). Interestingly, the lowest amount occurred September (44). Perhaps this is holiday season?

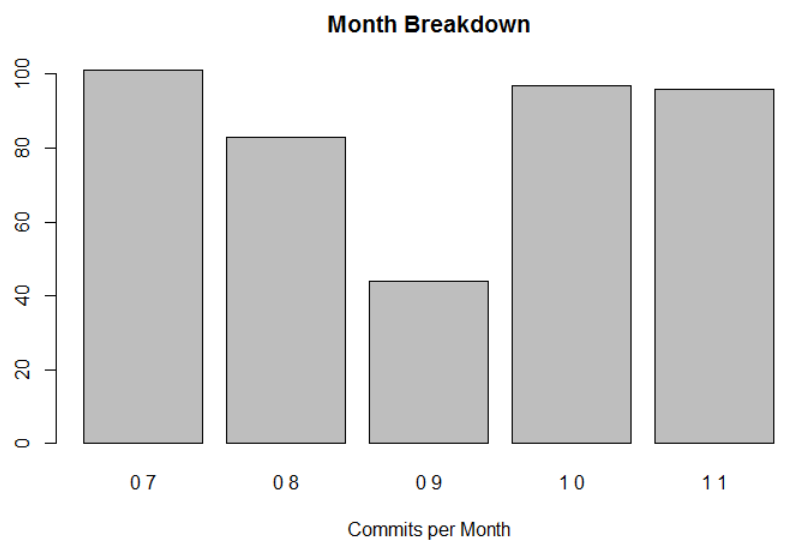


Figure 2: Monthly breakdown of commits

Generally, the most number of commits was in the region of 20-30 per day. A review of daily commits in Figure 3 does not present an obvious trend. It is noteworthy that after a day where a high number of commits we made (i.e. > 20), very few commits were completed the following day i.e. it’s rare that there would be two sequential days with a high number of commits.

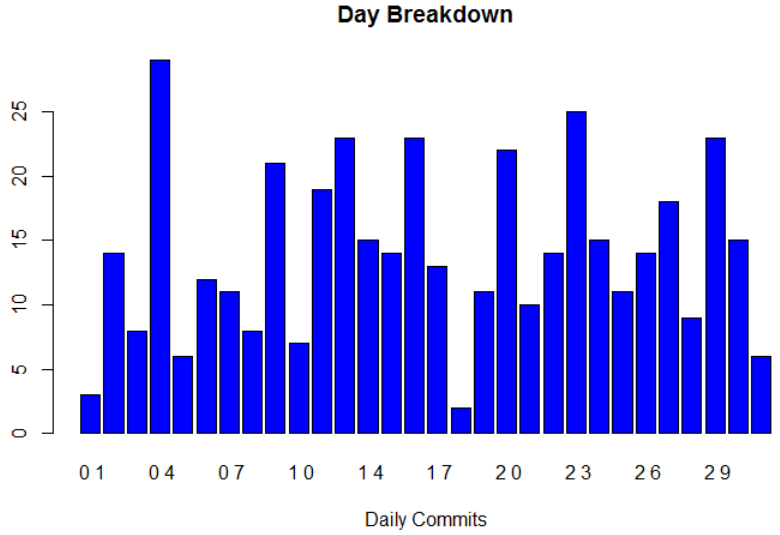


Figure 3: Daily breakdown of commits

## Week Day Analysis

In a similar set of steps to above, a blank list called “week\_days” was created, and the week-day of each commit was sliced and appended to the list.

Analysis of dates was completed using R Studio. The “week\_days” list was written to a CSV file; titled “week\_days.csv”.

The CSV file was imported into R Studio and analysis undertaken. A summary of the data provided by R is detailed in Figure 4.

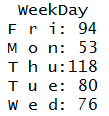


Figure 4: Summary of week\_days.csv file by R Studio

From the table above, we know that most commits were completed on Thursdays (118). It’s no surprise that the lowest number of commits occurred on Mondays. However, it is interesting that the second highest number of commits occurred on Fridays (generally seen as “winding down” day).

A breakdown of commits (by week day) is shown in Figure 5.

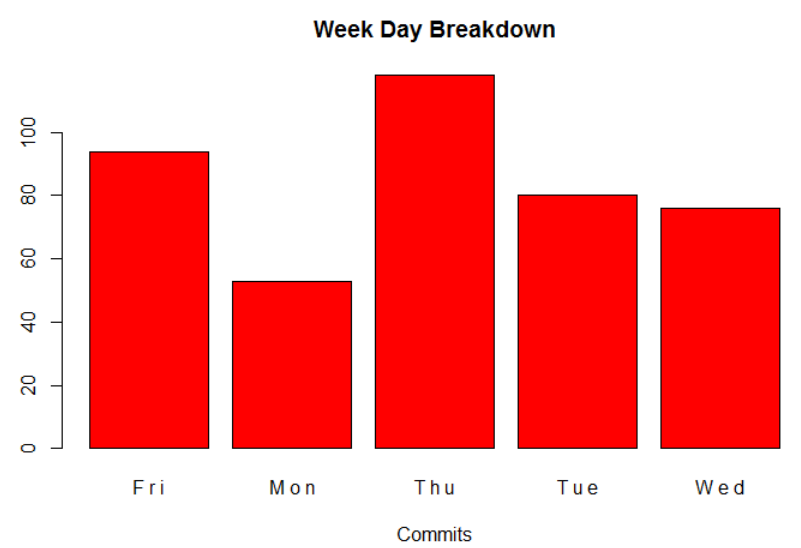


Figure 5: Week day breakdown of commits