Report Programming for Big Data

After using python to “clean” the data, which is called “Changes.csv”, I procedure to analyse the data using R:

The data has:

"revision", "date", "author", and "number\_of\_lines".

Using str() we can see the “revision” is an ID, “Date” is a format date, “Author” is

nominal and “number\_of\_lines” is nominal.

Even there are four columns, my analysis will be focus in “author” and “date”.

unique(my\_commits$author)

[1] "Thomas"

[2] "Vincent"

[3] "/OU=Domain Control Validated/CN=svn.company.net"

[4] "Jimmy"

[5] "Freddie"

[6] "Dave"

[7] "ajon0002"

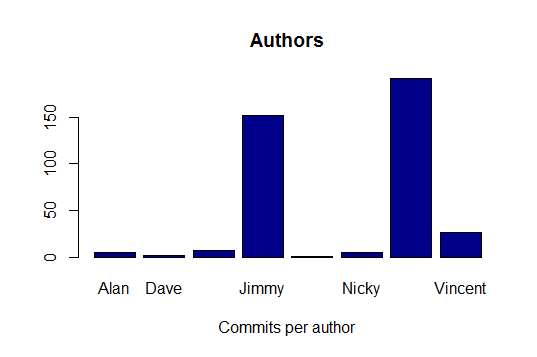
[8] "murari.krishnan"

[9] "Nicky"

[10] "Alan"

There are ten authors where the number three and seven are unusual names,

filtering those authors, we had the following analysis in the figure 1.1:



Thomas

Figure 1.1 Commitments per author.

|  |  |
| --- | --- |
| Author | Number of commitments |
| Thomas | 196 |
| Jimmy | 152 |
| Freddie | 7 |
| Dave | 2 |
| Alan | 5 |
| Nicky | 5 |
| Vincent | 26 |
| Murari.K | 1 |

Table 1.1 Author by commitments.

Thomas and Jimmy did more commitments through 2015, Murari just did one commitments, together Freddie, Dave, Alan and Nicky are close to Vincent.

Doing an analysis to the date. First, we must split the format due to its complexity:

"2015-11-27 16:57:44 +0000 (Fri, 27 Nov 2015)"

Using **splingr** library we can split the date in different columns, then converting it as a data frame to keep splitting until get years, month and days.

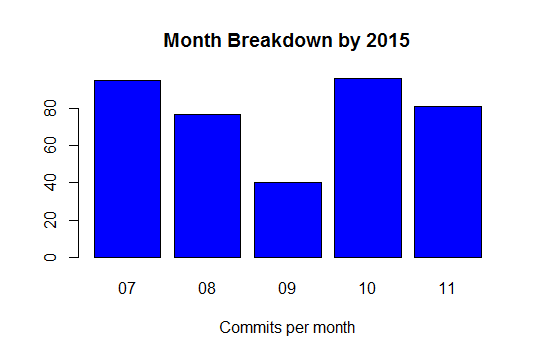


Figure 1.2

October was the month with more commitments (96), just the minimum difference with July.

July: 95.

August: 77.

September: 40.

October: 96.

November: 81.