ANALYSIS OF PARsed data

Douglas Nanka-Bruce email: 10365675@mydbs.ie 21-07-2017

CA-05 PROGRAMMING FOR BIG DATA

dublin business school | Castle House, Dublin 2, Ireland

2017

Table of Contents

[Table of Figures 1](#_Toc488448242)

[Parsing a log file into a csv file 2](#_Toc488448243)

[Authors and comments 2](#_Toc488448244)

[Number of comment lines by day of the week 3](#_Toc488448245)

[Number of comment lines by month of the year 4](#_Toc488448246)

[Number of comment lines by day of month 4](#_Toc488448247)

[Number of comment lines per month 5](#_Toc488448248)

[Number of comment lines by hour of day 6](#_Toc488448249)

# Table of Figures

[Figure 1: Parsing the date column into six sections using Excel’s Text to Columns function 2](#_Toc488448235)

[Figure 2: Number of comments and lines 2](#_Toc488448236)

[Figure 3: Dashboard of authors, comments and comment lines 3](#_Toc488448237)

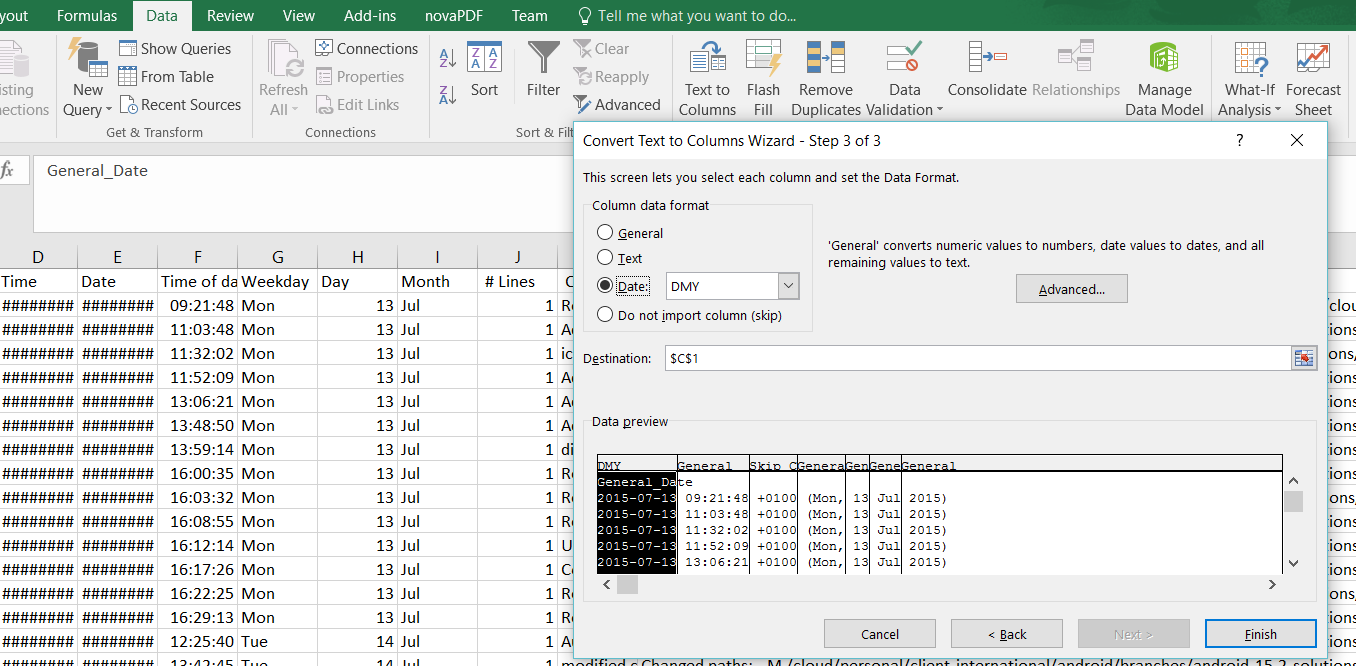
[Figure 4: Day of the week and number of lines 4](#_Toc488448238)

[Figure 5: Dashboard of number of lines by period 5](#_Toc488448239)

# Parsing a log file into a csv file

We use the Python program on a log file (named changes\_python.log) of 5255 lines and parse it through the commit method to achieve 422 lines (by using the process\_changes.py program) and writing to a csv file (saved as changes.csv). We test the first commit (using the simple.py file, and test\_simple.py file) on the first author named Thomas as follows: ‘author, ‘revision number’, ‘date’, and ‘number of lines’. We subsequently clean the csv file by parsing the ‘date’ column using Excel’s ‘Text to Columns’ wizard. An illustration is given in Figure 1.

Figure 1: Parsing the date column into six sections using Excel’s Text to Columns function

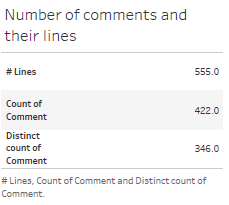


The cleaned Excel data is then exported into Tableau (version 10.3) to build some visualizations to reveal any interesting facts about the dataset.

# Authors and comments

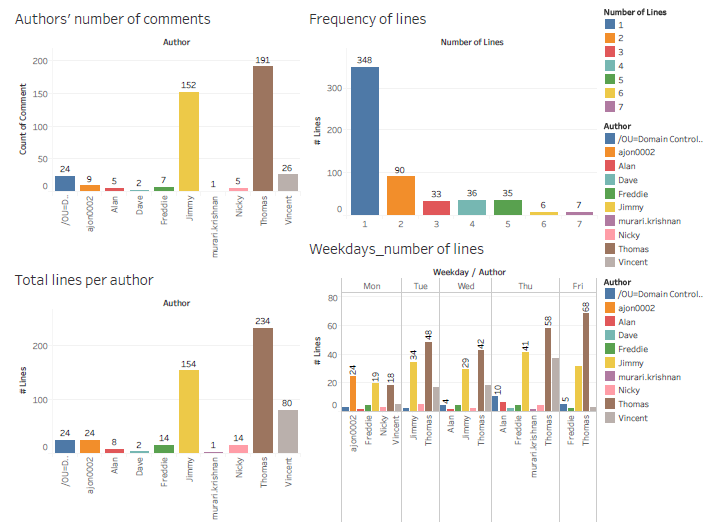
The csv file ‘changes.csv’ representing 422 comments written on 555 lines (because of multi-lined comments) is analyzed and summarized in Figure 2. The comments consist of 346 distinct comments implying that 76 of these have been repeated.

Figure : Number of comments and lines



Although, most of these comments are single-lined, some are between two to seven lines as shown in the frequency of lines chart in Figure 3. This gives a total number of 555 lines for the 422 comments. The total number of lines per author is from 1 (for ‘murari.krishnan’) to 234 (for ‘Thomas’). Figure 3 shows the detailed breakdown. It is worthy to note here that the 234 lines written by Thomas covers 191 comments.

Figure : Dashboard of authors, comments and comment lines



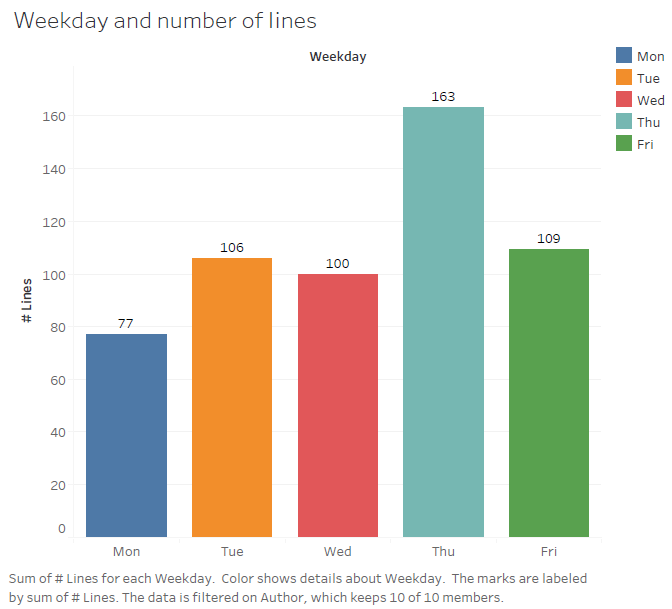
‘Jimmy’ has written 154 lines representing 152 comments and is in second place. This means most of his comments are a single-lined. The lines’ distribution indicates that 348 of the 555 are single lines written for comments. 90 are two lines, 33 for three lines, while four and five lines account for 36 and 35 respectively. The most lines written for a comment is seven and represents seven different occasions (see chart 2 of Figure 3). Comments with six lines are the least written and this happens six times.

# Number of comment lines by day of the week

The number of lines written each week day (i.e. Monday, Tuesday, Wednesday, Thursday or Friday) per author is also provided in Figure 3. The overall lines written per week day from the total of 555 is shown in Figure 4 and there are some interesting statistics as follows:

* The fewest number of lines are written on Mondays. The number given is 77.
* 106 lines are written on Tuesdays, making it the third busiest day.
* The second least comments are written on Wednesdays, a midweek, at 100.
* The most number of lines are written on Thursday at 163 lines
* Friday is also a busy day for comments (109 lines) but as will be shown later, this is all up to one author who does few comments on a Monday and does the opposite on Fridays.

Figure : Day of the week and number of lines



As Figure 3 has revealed, Thomas, who is by far the most prolific line writer, is at his busiest on Fridays writing 68 of 109 lines among all ten authors. Most of his lines are written on this day but he writes 58 lines on Thursday which has been established to the busiest line-writing day. Out of the 422 comments in total, ‘murari.krishnan’ authors just one and this is on a Thursday.

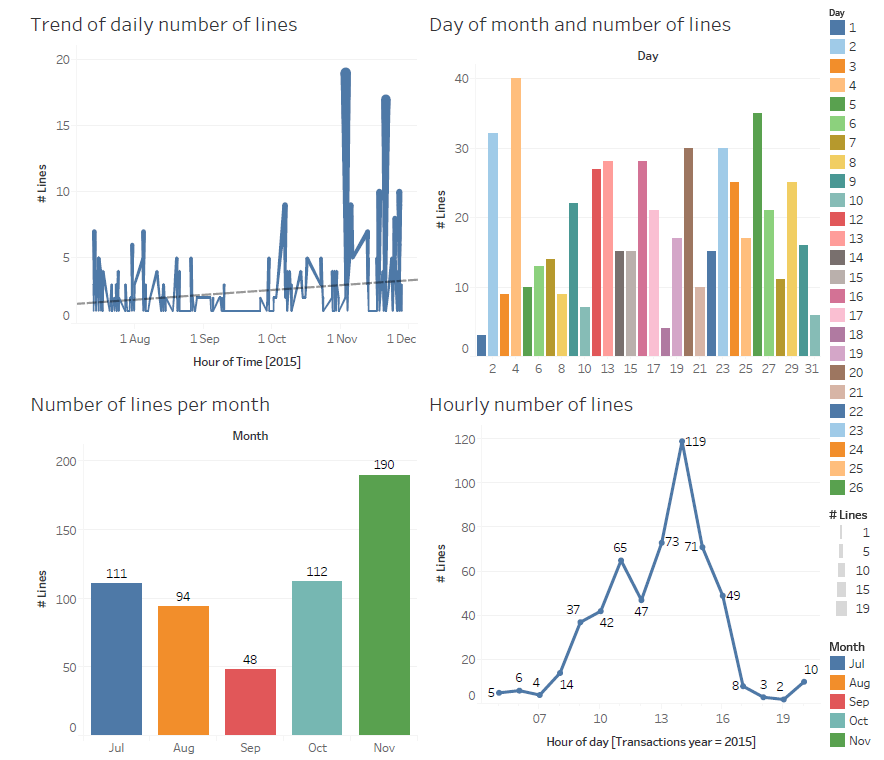
# Number of comment lines by month of the year

Figure 5 is a dashboard with four different charts analysing number of lines written over different types of period measure. The first is the daily number of lines written from July to November 2015. A trend line (of the form *‘# Lines = 0.0120 \* Hour - 505 : R2 = 0.0537 : p-value = 0.0003*’)reveals that there is a marginally significant increase in the number of lines written from the earlier period in July until when the last comment is authored towards the end of November. On November 02, 2015, the most lines, 15, are written. From 9th to 25th September, only one line per day representing a comment is written.

# Number of comment lines by day of month

When we look at the day of the month (Chart 2 of Figure 5), Day 4 of the month has the most lines written (40 lines) while only three lines are written on Day 1, which is half the number written on Day 31 (6 lines), however, 32 lines are written on Day 2. Similarly, in second place for most lines written, 35 lines are written on Day 26. To complete this section, 30 lines are written for both Day 20 and Day 23.

Figure : Dashboard of number of lines by period



# Number of comment lines per month

The period of data availability is from July 13th to November 27th. Quite fittingly, our prolific line writer, Thomas, is the author. The next analysis we undertake is to look at the number of lines produced per month. September has the least number of lines at only 48 compared to 111 for July (and 112 for October) or 190 lines for November, which is the highest for any month. The second lowest lines for any month is in August at 94. It is fair to say that, this follows an inverse quadratic shape where line authoring starts in earnest and then decreases in intensity before picking up again in October. The low R2 value (0.0537 or 5.37%) of the trend line in Figure 5 shows that a linear estimation is inappropriate as it only explains 5% of the variability between period and number of lines and, maybe, a quadratic estimation is more appropriate.

# Number of comment lines by hour of day

The last analysis to be discussed is Chart 4 of Figure 5 showing the number of lines written by hour of day. This is distributed from 5:00am to 8:00pm. The chart reveals that most lines are written between 1:00pm and 3:59pm. 73 lines are written between 1:00-1:59pm, 119 lines between 2:00-2:59pm and 71 lines are written between 3:00-3:59pm. Additionally, there are 65 lines written from 11:00-11:59am. This can be roughly interpreted as a bell-shaped curve where most lines of comment are written between 8:00am-5:00pm, representing the typical working day. The tails of the distribution represent the early morning and night comments that are authored.