5000 LINE DATASET ANALYSIS WITH PYTHON AND TABLEAU

Using Python, I was able to gain and confirm some initial information about the data set file 'changes_python.log' as follows:

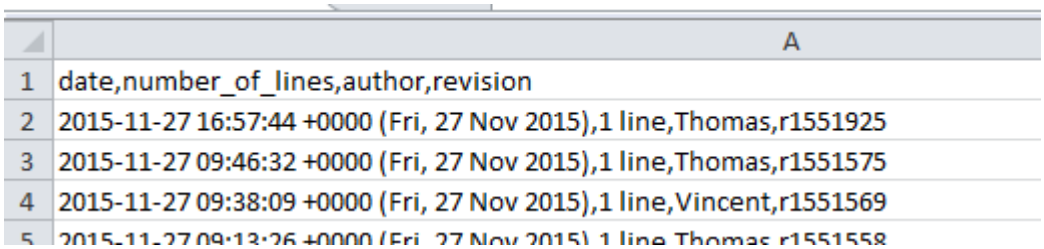
- How many lines of data in the file: 5,255
- How many commits: 422
- Details of the first commit: {'date': '2015-11-27 16:57:44 +0000 (Fri, 27 Nov 2015)', 'number_of_lines': '1 line', 'author': 'Thomas', 'revision': 'r1551925'}
- Total commits of each author: {'Thomas': 191, 'Jimmy': 152, 'murari.krishnan': 1, 'Alan': 5, 'Freddie': 7, 'Dave': 2, 'Nicky': 5, 'ajon0002': 9, 'Vincent': 26, '/OU=Domain Control Validated/CN=svn.company.net': 24}

In order to perform analysis on the data file I used python to export some useful information to a CSV file as follows:

- Details of every commit to one CSV file
- Details of every comment each author committed

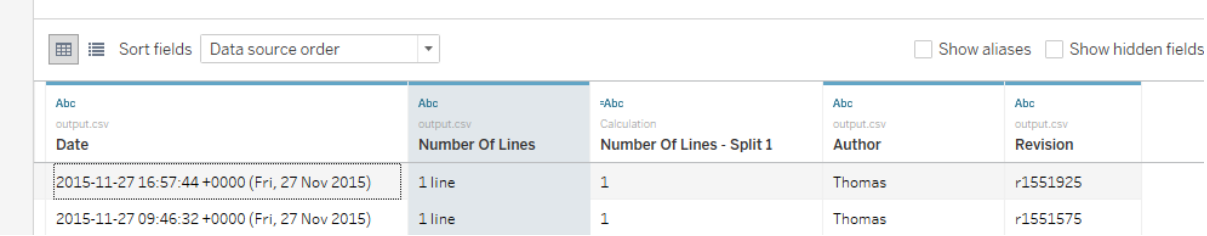
From there I used Tableau to analyse the data I had exported to CSV. I will describe the data munging process followed by the statistical results I produced.

Format of the commit information in CSV format:



	A
1	date,number_of_lines,author,revision
2	2015-11-27 16:57:44 +0000 (Fri, 27 Nov 2015),1 line,Thomas,r1551925
3	2015-11-27 09:46:32 +0000 (Fri, 27 Nov 2015),1 line,Thomas,r1551575
4	2015-11-27 09:38:09 +0000 (Fri, 27 Nov 2015),1 line,Vincent,r1551569
5	2015-11-27 09:13:26 +0000 (Fri, 27 Nov 2015),1 line,Thomas,r1551558

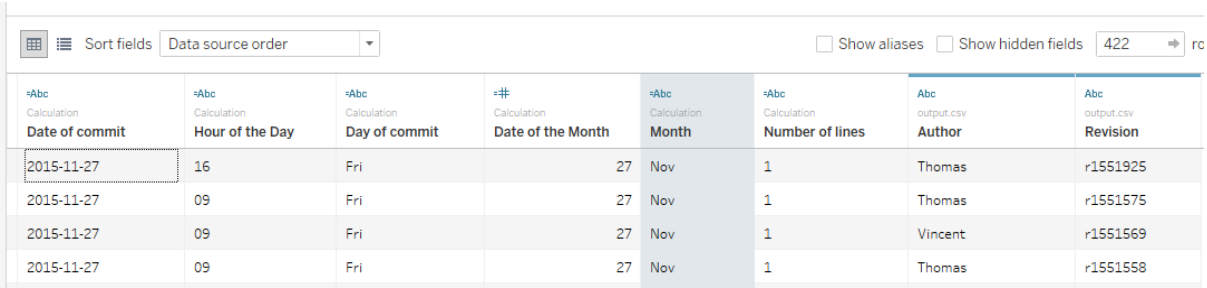
Loading this into Tableau I used the split function to get the information I wanted into columns:



Sort fields: Data source order — Show aliases — Show hidden fields

Abc output.csv Date	Abc output.csv Number Of Lines	=Abc Calculation Number Of Lines - Split 1	Abc output.csv Author	Abc output.csv Revision
2015-11-27 16:57:44 +0000 (Fri, 27 Nov 2015)	1 line	1	Thomas	r1551925
2015-11-27 09:46:32 +0000 (Fri, 27 Nov 2015)	1 line	1	Thomas	r1551575

Results as follows in Tableau:



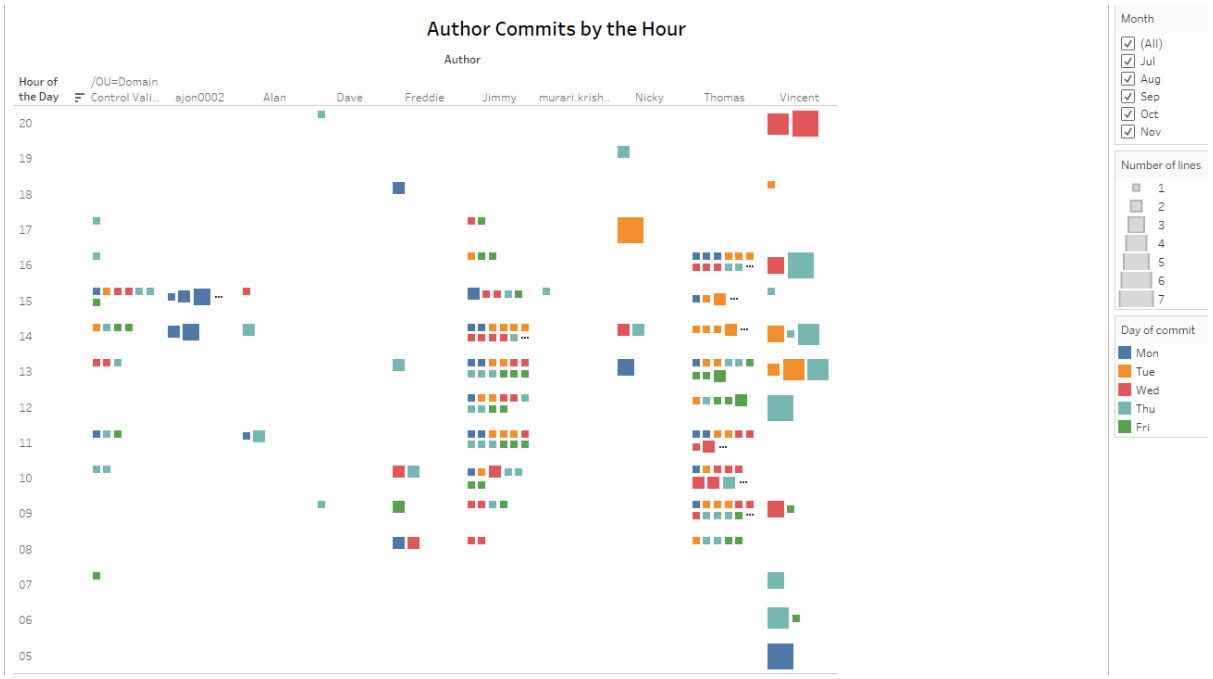
Sort fields: Data source order — Show aliases — Show hidden fields — 422

=Abc Calculation Date of commit	=Abc Calculation Hour of the Day	=Abc Calculation Day of commit	=# Calculation Date of the Month	=Abc Calculation Month	=Abc Calculation Number of lines	Abc output.csv Author	Abc output.csv Revision
2015-11-27	16	Fri	27	Nov	1	Thomas	r1551925
2015-11-27	09	Fri	27	Nov	1	Thomas	r1551575
2015-11-27	09	Fri	27	Nov	1	Vincent	r1551569
2015-11-27	09	Fri	27	Nov	1	Thomas	r1551558

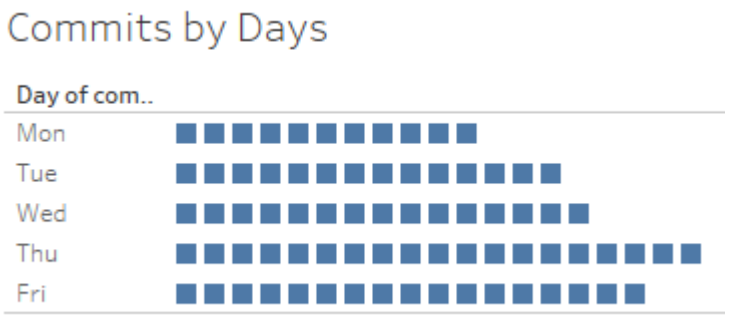
Happy with the data I went about creating a couple of worksheets and incorporated these into one Dashboard to display 3 pieces of statistically interesting information about the dataset.

The first worksheet I created is entitled “Author Commits by the Hour”. Tableau is a wonderful piece of software in that at first glance you can see which authors have committed more than others. Interesting finds from this Worksheet as follows:

- Jimmy and Thomas produce the most commits and do so during the hours of 8am to 5pm
- Vincent seems to include more detailed information with each commit than the others in the group
- Vincent seems to never sleep as he has commits as early in the day as 5am and as late as 8pm
- Ajon002 only has commits in the month of November, is he new to the team or project?

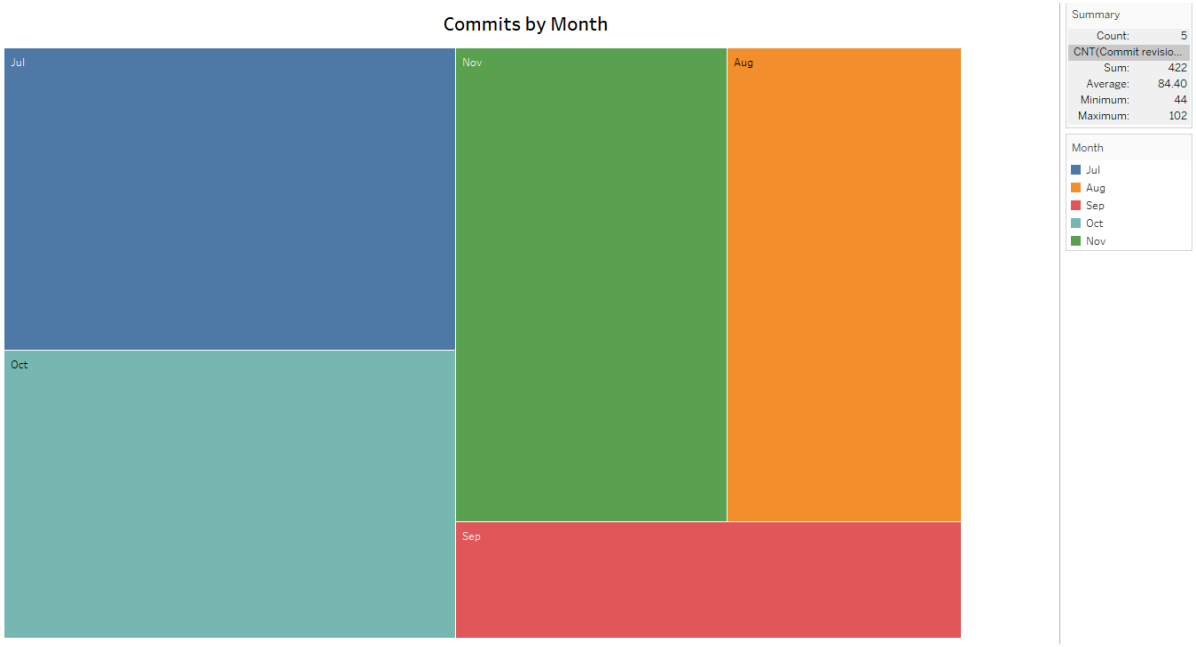


Another worksheet shows that Thursday is the busiest day for commits with 118 over the dataset, followed closely by Friday with 95. Monday is the quietest day for commits.

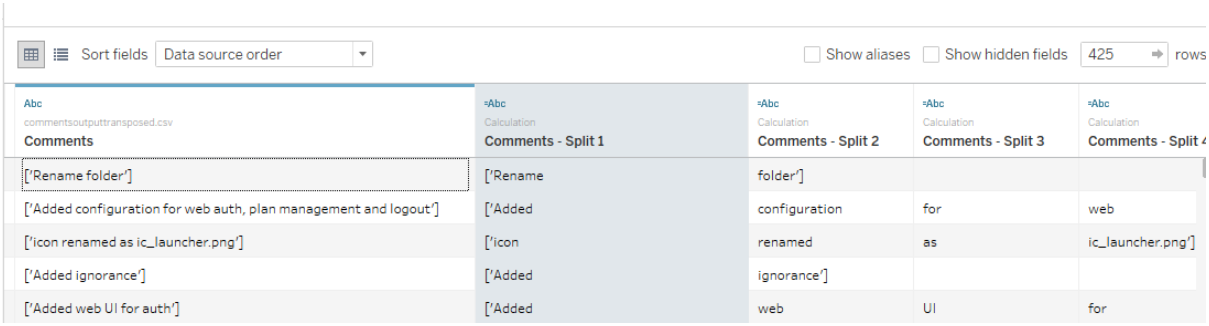


The “Commits by Month” worksheet gives a graphical representation of the total commits per month. Interesting stats as follows:

- July was the highest with 102 commits, with September the lowest with only 44. Where staff on holidays in September perhaps?
- Average commits per month was 84

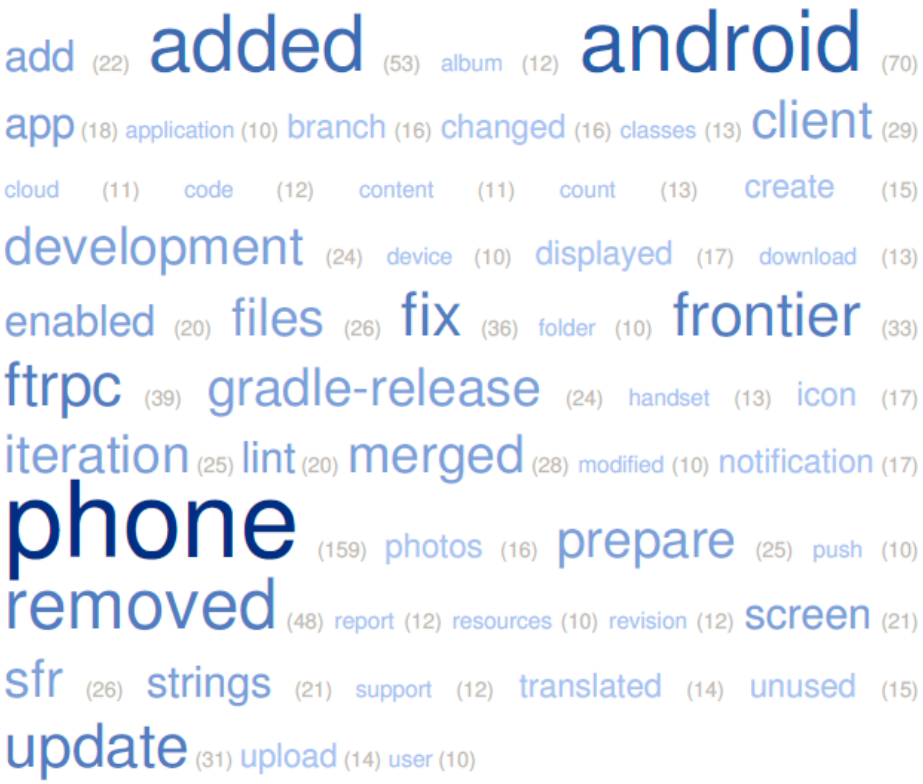


Finally I wanted to produce a word cloud from all the author comments I had extracted using Python to see where there any common trends in the teams work. However I had some difficulty using Tableau trying to split the file up.



I turned to the web and used a free word cloud maker from tagcrowd.com and created the below word cloud based on the comment CSV file I had extracted. Insights as follows:

- Phone is the most mentioned word with 159 instances of the word
- Software is Android it seems as it is mentioned 70 times
- There are more mentions of “added” than “removed”; 53 vs 48
- Translated was mentioned 14 times, does this software or project involve different spoken languages?
- Fix is mentioned 36 times
- Photos is mentioned 16 times and the word album 12 times
- The team seem to be using the Gradle build tool as gradle-release is mentioned 24 times



Final Dashboard