Programming for Big Data

Continues Assessment:

CA4 - Perform Analysis on a 5000-line dataset

Unit Leader:	**Darren Redmon**
Unit code:	**B8IT105**
Student:	**Andrea Guzman McMullan**

Author:	Andrea Guzmán McMullan
Effective date:	25th Nov 217

CONTENT

ASSESSMENT TASK

Assessment request

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.

Data Specification

We got a data set file with 5255 text lines downloaded from GitHub, with updates from various users during the period 13th Jul 2015 to 27th Nov 2017. the analysis of this data will be as follow:

Data Preparation

i. **Python code with and object** In Python **to read in the file** and fixing the data, using an object-oriented Programming.

ii. **Python code Test with and object** In Python **Test script** to test some functions: to ensure that my object-oriented code is working correctly testing the following aspects:
   a. Total number of lines in the data source file [5255]
   b. Total number of commit objects in the output file [422]
   c. The index committees of the committees, returns correct data checking commit 24 and 20

iii. The code will create a csv file called changes.txt

Cleaning Data

i. Having 422 objects that are in a list with the following attributes
   a. Review
   b. Author
   c. Date
   d. Time
   e. Numbers
   f. Comment

Data Transformations

i. In Excel I have renamed 24 lines where the author was: *'/OU=Domain Control Validated/CN=svn.company.net'*. as Unknow, from the CSV file. I am assuming that the user was not registered.

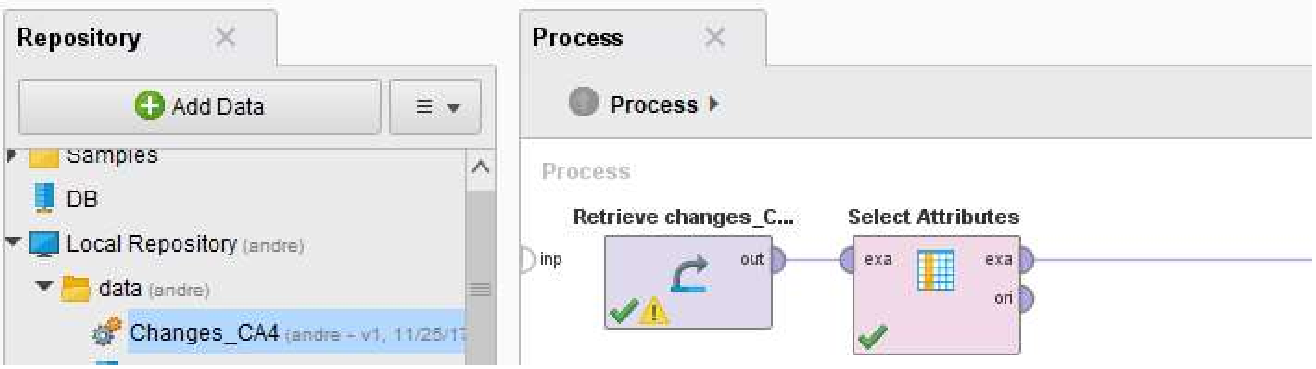
   A domain-validated or a low-authentication (1-factor) product is a server certificate delivered quickly and without real vetting. It does not guarantee the identity of the website's owner nor the actual existence of the organization!
   https://www.tbs-certificates.co.uk/FAQ/en/205.html

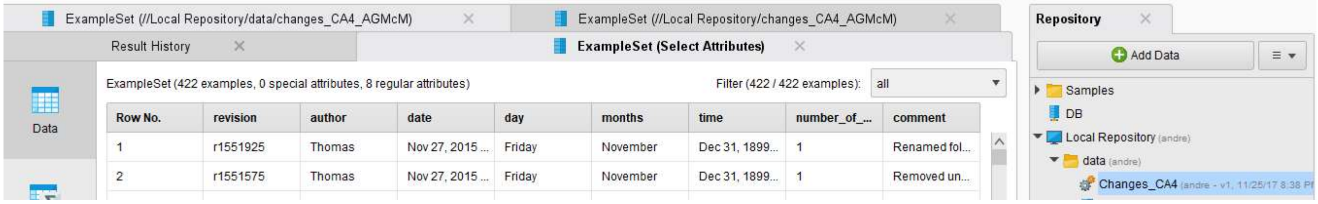
ii. In Excel I have split the date into the day of the week and the month. With the following formulas
   =TEXT(C2,"dddd")
   =TEXT(C2,"mmmm")

Interesting Statistical pieces

Rapidminer

I have used Rapidminer to analyse the data set already transformed in excel. I have downloaded in the Local repository the file called changes_CA4 as below …




ExampleSet (//Local Repository/data/changes_CA4_AGMcM)
ExampleSet (//Local Repository/changes_CA4_AGMcM)
Result History
ExampleSet (Select Attributes)
Data
Repository
Add Data
Samples
DB
Local Repository (andre)
data (andre)
Changes_CA4 (andre - v1, 11/25/17 8:38 P

ExampleSet (422 examples, 0 special attributes, 8 regular attributes)

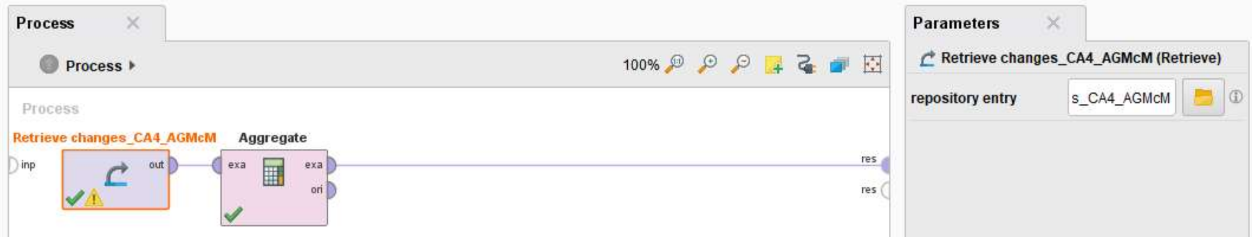
Filter (422 / 422 examples): all

Row No.	revision	author	date	day	months	time	number_of_...	comment
1	r1551925	Thomas	Nov 27, 2015 ...	Friday	November	Dec 31, 1899...	1	Renamed fol...
2	r1551575	Thomas	Nov 27, 2015 ...	Friday	November	Dec 31, 1899...	1	Removed un...

Number of times each user logged

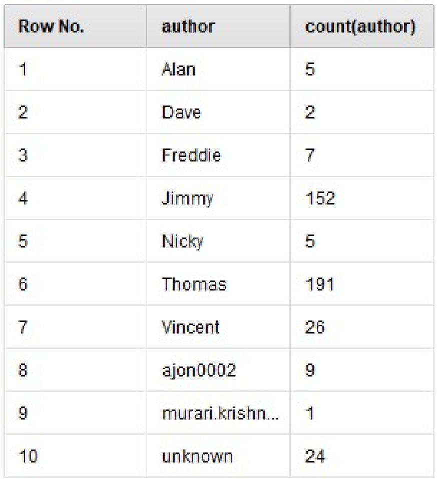
Here I am showing the process to generate the statistics and visualization of the number of times each user logged. I have retrieved the file and aggregate only the Author and count it.

Process



repository entry: s_CA4_AGMcM

Results - Data



Row No.	author	count(author)
1	Alan	5
2	Dave	2
3	Freddie	7
4	Jimmy	152
5	Nicky	5
6	Thomas	191
7	Vincent	26
8	ajon0002	9
9	murari.krishn...	1
10	unknown	24

Results - Chart

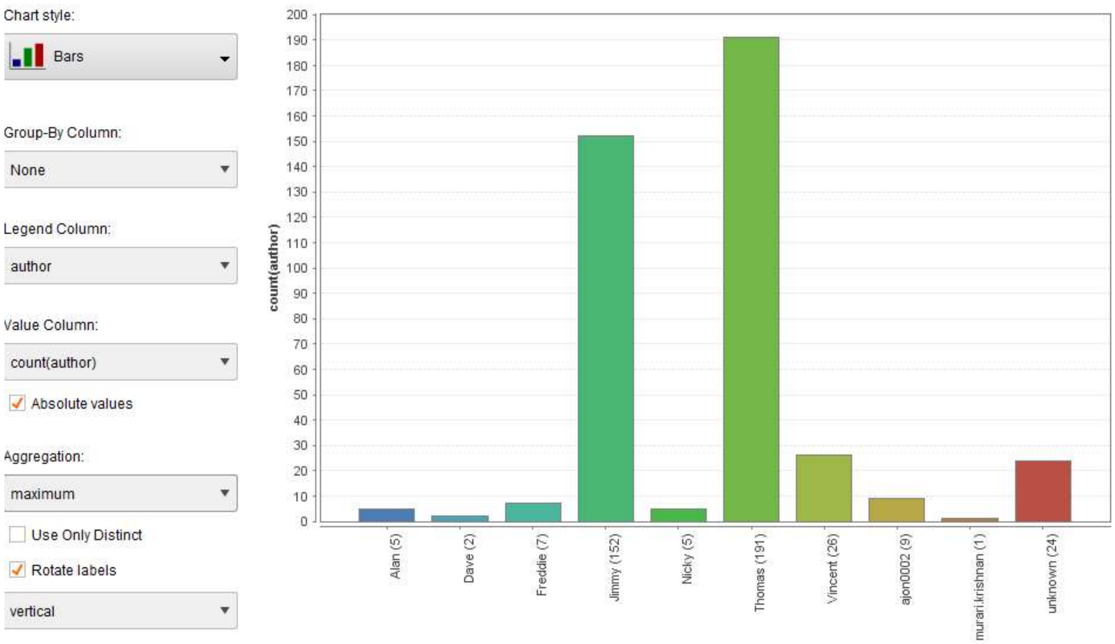


Chart style: Bars

Group-By Column: None

Legend Column: author

Value Column: count(author)

☑ Absolute values

Aggregation: maximum

☐ Use Only Distinct

☑ Rotate labels

vertical

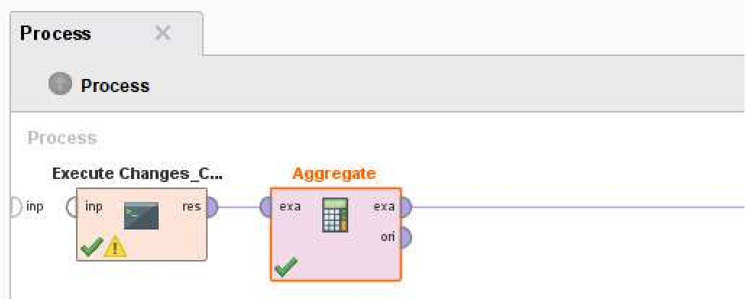
Conclusions

The Author that connect more frequently during this period time (13h Jul 2015 to 27th Nov 2017) is Thomas.

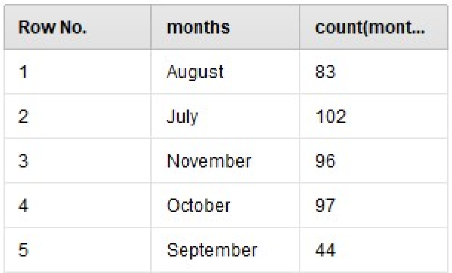
The Busiest / quietest Month

Here I am showing the process to generate the statistics and visualization of busiest and quietest time of the month. I have retrieved the file Changes_CA4 and aggregate only the month and count it.

Process



Data



Row No.	months	count(mont...
1	August	83
2	July	102
3	November	96
4	October	97
5	September	44

Results - Chart

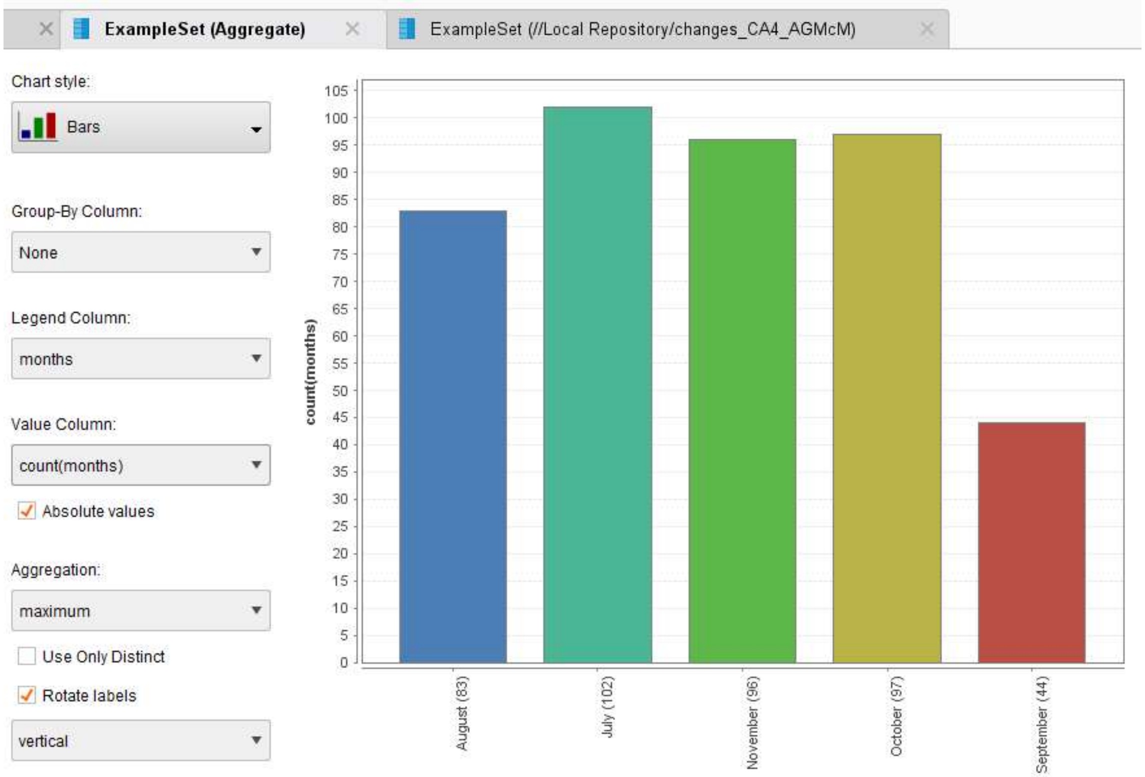


Chart style: Bars

Group-By Column: None

Legend Column: months

Value Column: count(months)

☑ Absolute values

Aggregation: maximum

□ Use Only Distinct

☑ Rotate labels

vertical

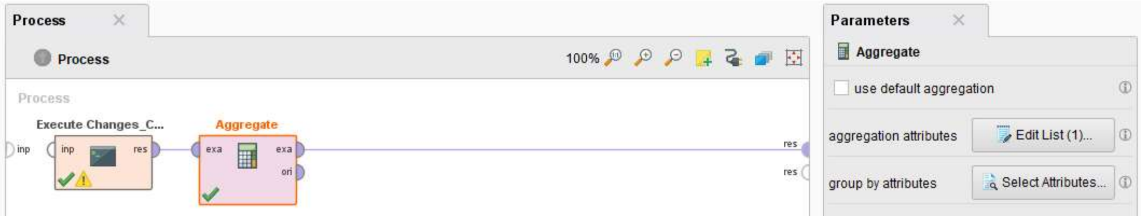
Conclusions

The **busiest** month during this period time (13h Jul 2015 to 27th Nov 2017) is **July** and the **quietest** is **September**.

Busiest day of the week

Here I am showing the process to generate the statistics and visualization the busiest day of the week. I have retrieved the file Changes_CA4 and aggregate only the day and count it.

Process

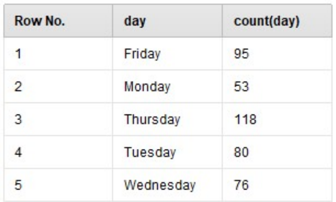


□ use default aggregation

aggregation attributes: Edit List (1)...

group by attributes: Select Attributes...

Data



Row No.	day	count(day)
1	Friday	95
2	Monday	53
3	Thursday	118
4	Tuesday	80
5	Wednesday	76

Results - Chart

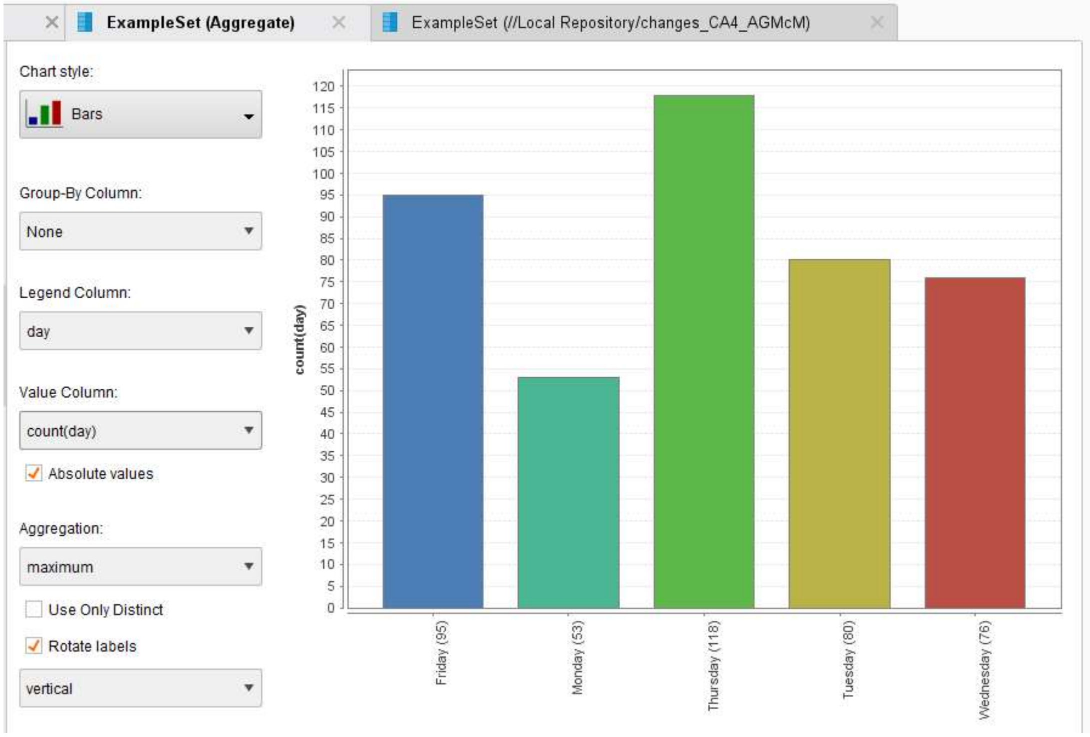


Chart style: Bars

Group-By Column: None

Legend Column: day

Value Column: count(day)

☑ Absolute values

Aggregation: maximum

□ Use Only Distinct

☑ Rotate labels

vertical

Conclusions

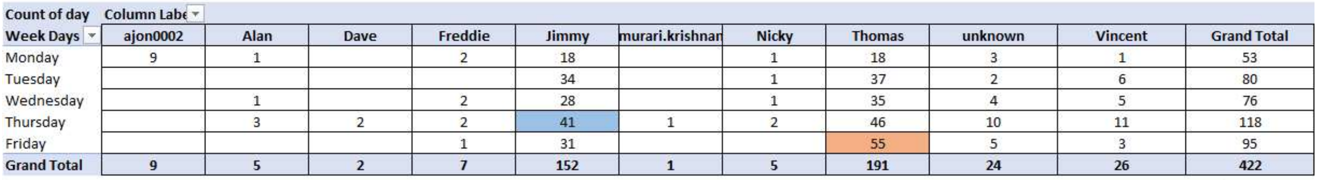
The Busiest day of the week was **Thursday** and interesting that nobody logs in at the weekend during this period time (13h Jul 2015 to 27th Nov 2017)

Posting day from Users

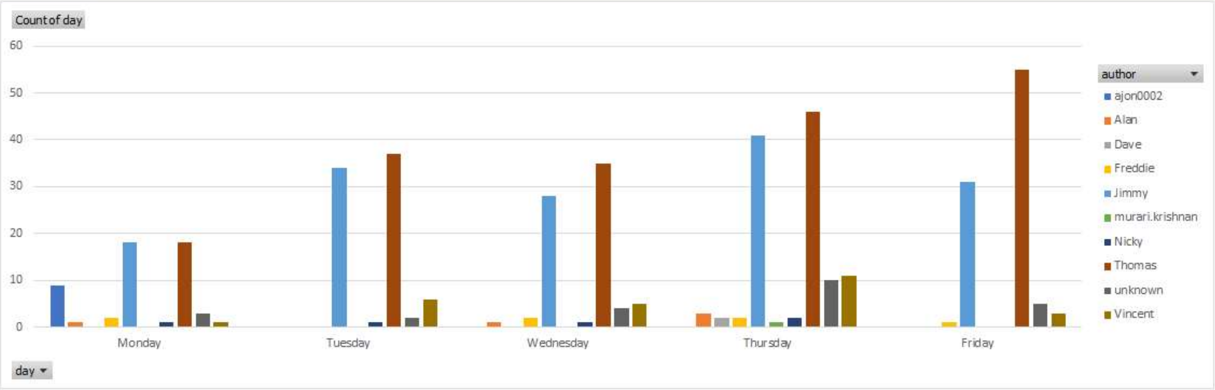
Excel

I have done an extra analysis in excel to find the days when the users that posted the most had log on.

As per the picture below Thomas logged on the most on Fridays and Jimmy on Thursdays.



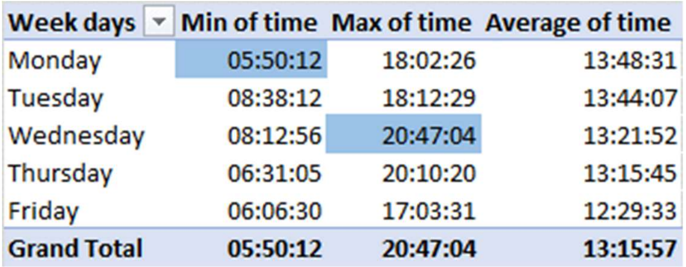
Count of day	Column Labe										
Week Days	ajon0002	Alan	Dave	Freddie	Jimmy	murari.krishnan	Nicky	Thomas	unknown	Vincent	Grand Total
Monday	9	1		2	18		1	18	3	1	53
Tuesday					34		1	37	2	6	80
Wednesday		1		2	28		1	35	4	5	76
Thursday		3	2	2	41	1	2	46	10	11	118
Friday				1	31			55	5	3	95
Grand Total	**9**	**5**	**2**	**7**	**152**	**1**	**5**	**191**	**24**	**26**	**422**



Busiest time of the day - Mean.

In excel also I have analysed the times. The time user posts the most in average is at 13:15.

Monday is when the users posted the earliest, Wednesday the users post the latest, during the period time of 13h Jul 2015 to 27th Nov 2017.



Week days	Min of time	Max of time	Average of time
Monday	05:50:12	18:02:26	13:48:31
Tuesday	08:38:12	18:12:29	13:44:07
Wednesday	08:12:56	20:47:04	13:21:52
Thursday	06:31:05	20:10:20	13:15:45
Friday	06:06:30	17:03:31	12:29:33
Grand Total	**05:50:12**	**20:47:04**	**13:15:57**

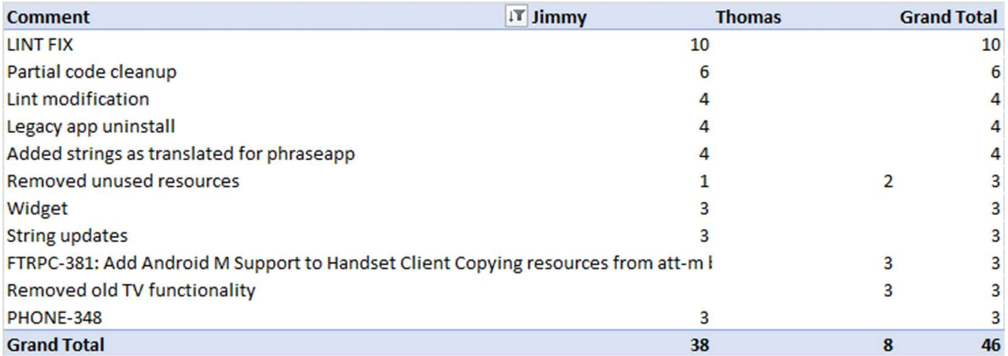
Which comment appears the most

We know that Jimmy and Thomas are the users that logged in the most, so only using these two Authors, I could analyse the most common comment in the data set.

I have done this analysis in Excel.

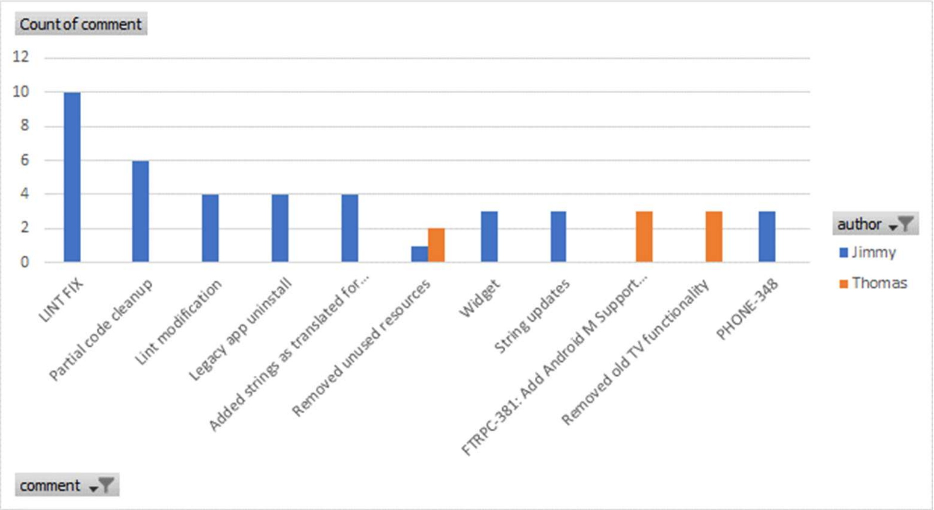
Data

This data from a Pivot table, filtering authors Jimmy and Thomas as the authors that logged more frequently.



Comment	Jimmy	Thomas	Grand Total
LINT FIX	10		10
Partial code cleanup	6		6
Lint modification	4		4
Legacy app uninstall	4		4
Added strings as translated for phraseapp	4		4
Removed unused resources	1	2	3
Widget	3		3
String updates	3		3
FTRPC-381: Add Android M Support to Handset Client Copying resources from att-m		3	3
Removed old TV functionality		3	3
PHONE-348	3		3
Grand Total	**38**	**8**	**46**

Results - Chart



Conclusions

The comment that appears the most is **LINT FIX**; this comment is for fixing errors. And the chart illustrates that the Author Jimmy is who get this the comment the most.