|  |  |
| --- | --- |
| CA 5 – PBD  B8IT105 | Mohammadbagher bagheri    B8IT105  Programming for Big Data |

For this project the log data file must be parsed, according to the [Business Dictionary](http://www.businessdictionary.com/definition/parsing.html) parsing is:

*“Breaking a data block into smaller chunks by following a set of rules, so that it can be more easily interpreted, managed, or transmitted by a computer.”*

After parsing the provided raw data (changes\_python.log) which slightly is over 5000 lines, the 422 lines in CSV file is produced. The produced CSV data needs to be scrubbed further.

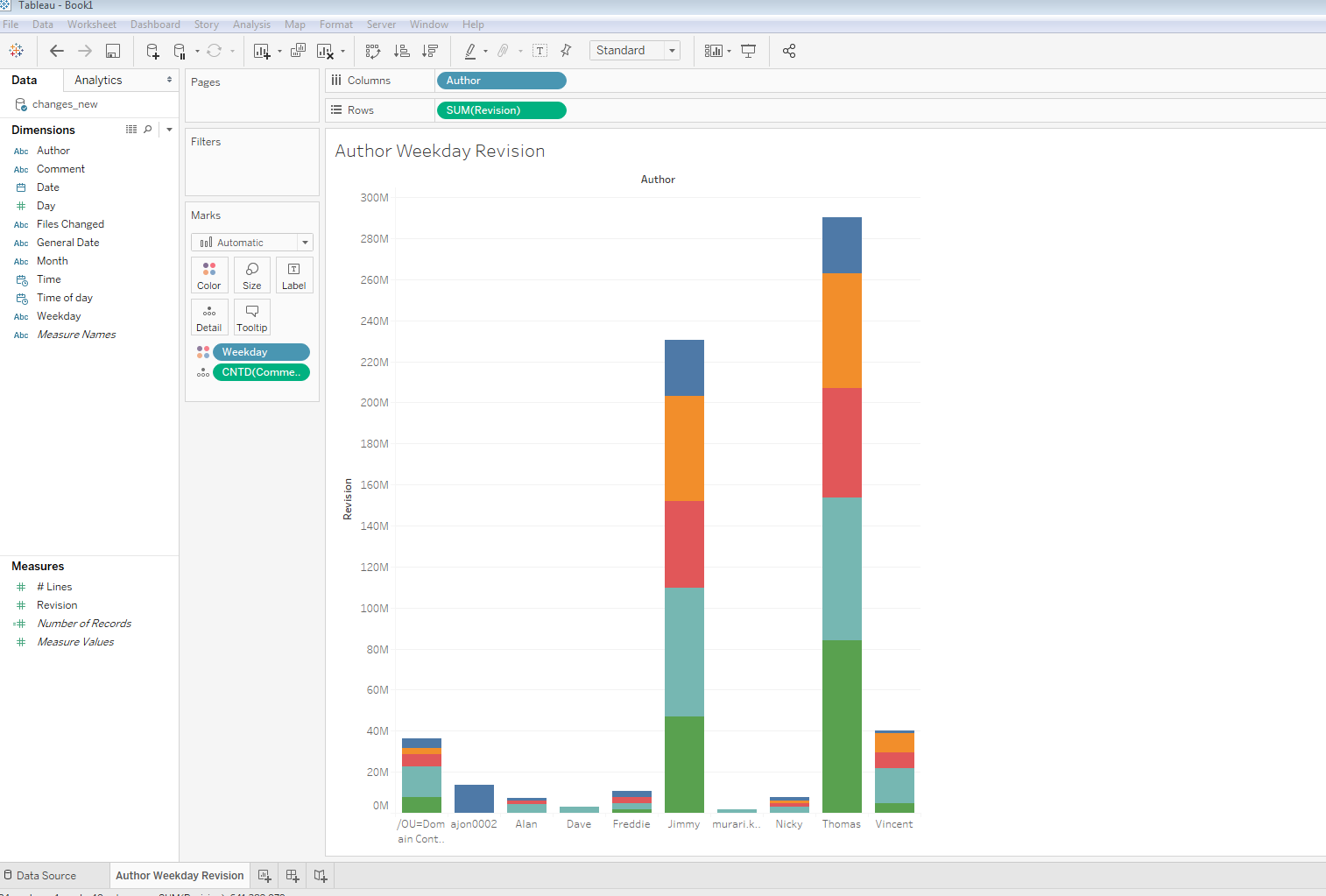
According to [techopedia](https://www.techopedia.com/definition/14651/data-scrubbing):

*“Data scrubbing refers to the procedure of modifying or removing incomplete, incorrect, inaccurately formatted, or repeated data in a database. The key objective of data scrubbing is to make the data more accurate and consistent.”*

After scrubbing and tiding up the data, it is now prepared to be visualised and analysed. As we have been practicing and using Tableau for data analysing and visualizing, I decided to analyse this set of data with Tableau.

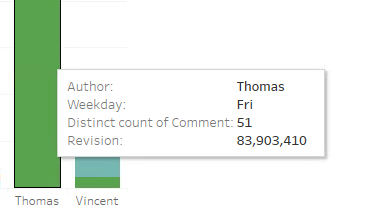
**Analysing CSV file in Tableau: Author Weekday Revision**

While processing the visualisation in Tableau, we can see that there are total number of 10 authors.

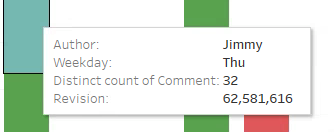


Analysing authors weekly revision indicate that Thomas is the author who has the highest revision, the number of revision of each day of the week also indicates that Thomas’s highest number of revision is Friday.

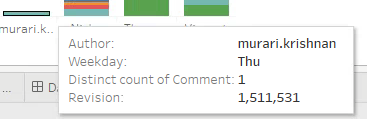
His overall number of weekly revision is about 290 million.



Jimmy has the second highest revision and Thursday is his highest revision day. His overall number of weekly revision is about 240 million.

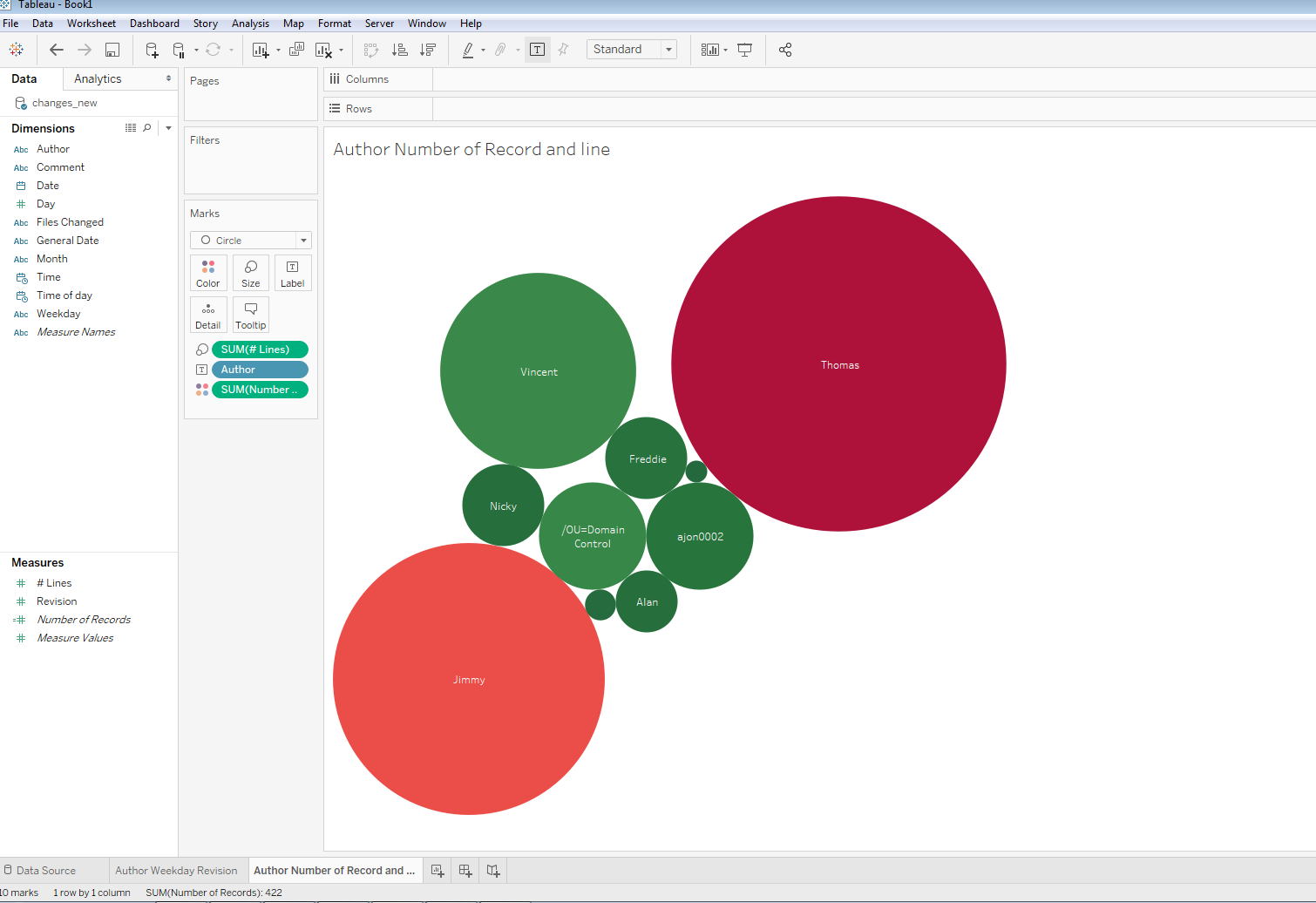


Muari Krishnan has the least number of revision and the only revision day for this author is Thursday. Mauri’s overall number of weekly revision is about 1.5 million.

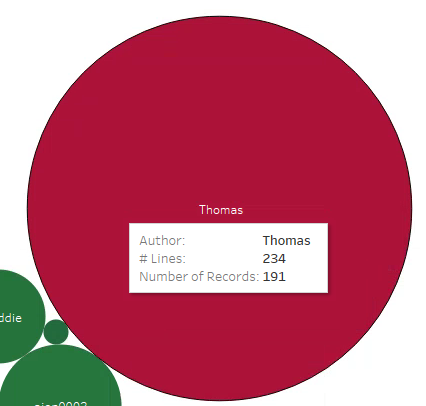


**Author’s Number of Record and Line**

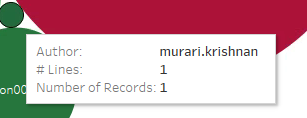
There is some more interesting discovery about this dataset.



Thomas has the highest number of records and number of line.

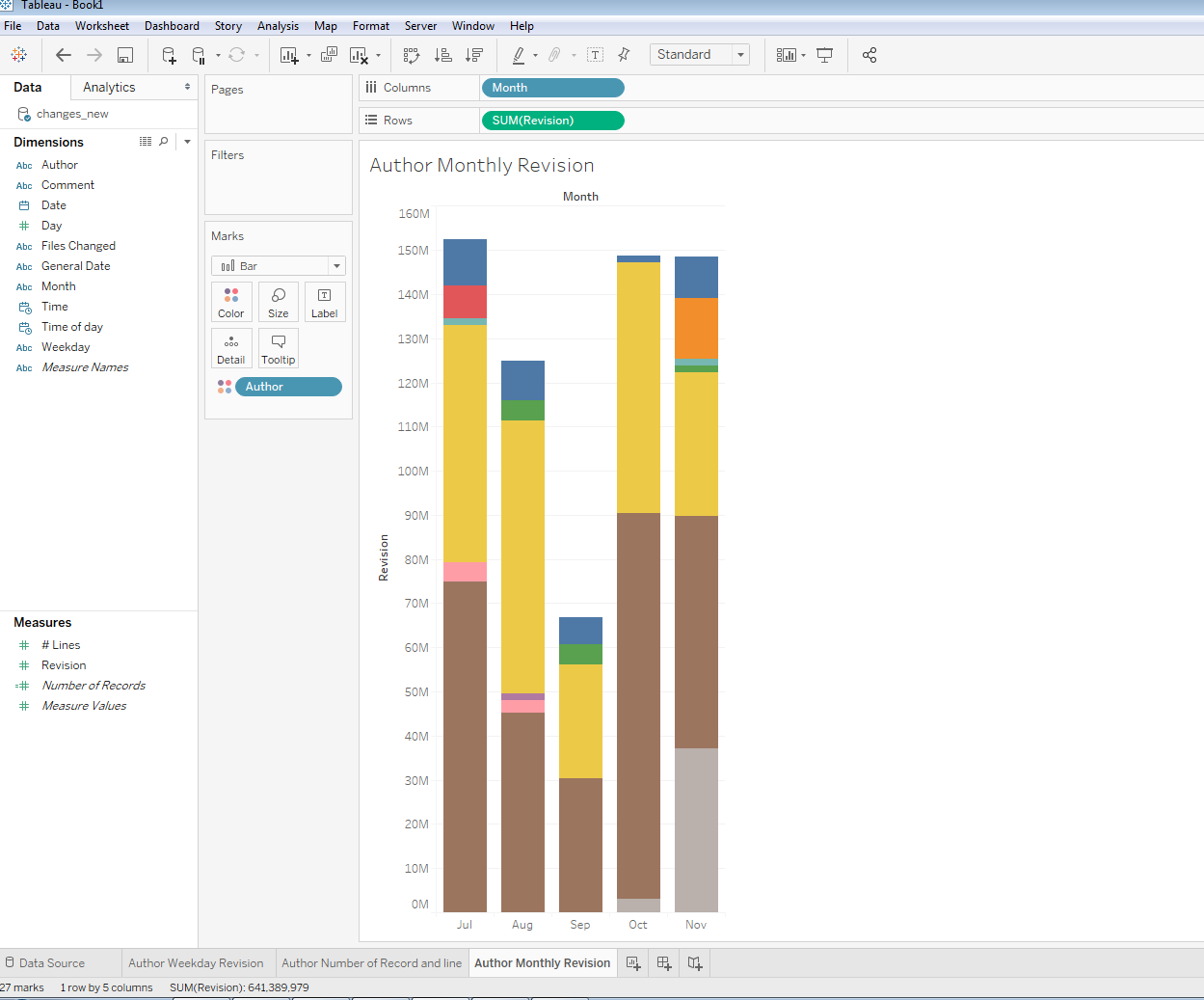


Murai has the least records (just 1) and 1 number of line.



**Author Monthly Revision**

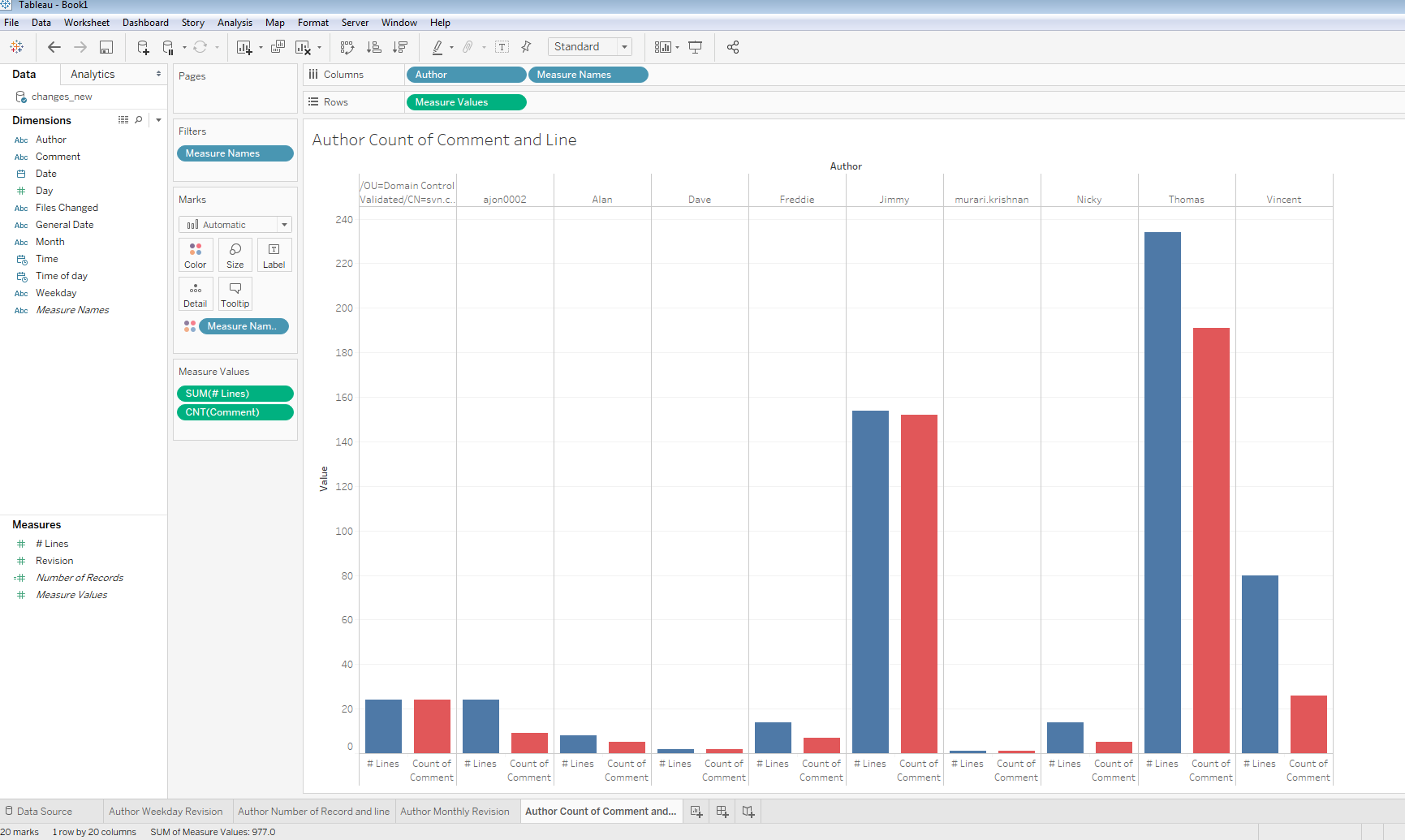
Next I visualised the monthly revision of the authors, it is interesting to see that in some month there are only 3 or 4 authors that their work has revisions.



July has the highest records of revision with over 150 million and September has the least records of revision with almost 65 million.

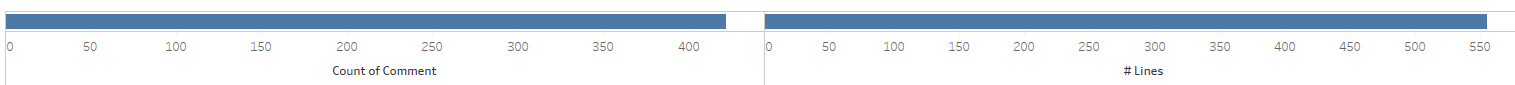
November is the month that there are 7 authors whose work has revisions which is the highest number of month.

Author Count of Comment and Number of Line



Further analysing of authors’ count of comments and number of lines produces another graph that indicates which author has the most and least comments. Thomas has 191 comments makes up 234 lines, Murai has 1 comment in 1 line and Dave is the second lowest comment and number of lines holder by 2 each.

Also one of the interesting discovery about number of overall comments and lines is that the number of lines is 555 and number of comments are 400, meaning that some of the comments must be multi lines.



Dashboard

In the dashboard below the four various visualization graph represent the analysed dataset with different variables.

