Vincent Obuseh

Programming for Big Data  Dublin Business School

CA 5 - Performing Analytics on a 5000 entry dataset.

Anlytics was performed on the 5000 entry dataset after it was cleaned and placed into the relevant container objects. Once in the objects, 422 different sets of commit objects were observed and analysed. Upon analysis, 3 interesting statistical pieces of information which are Revision, Author and Date were picked. Tableau, a powerful visualisation tool was used to gain insights on the downloaded csv file from my read\_changes.py python file. Information gained from the dataset are explained below.

1. **Number of Commits per Author:** From the dataset, looking at the author and number of records, I was able to tell which author had the highest number of commits. Thomas had the highest number of commits made, 191 in total while Murari Krishnan had the lowest with a total of just 1. See fig. 1 below.

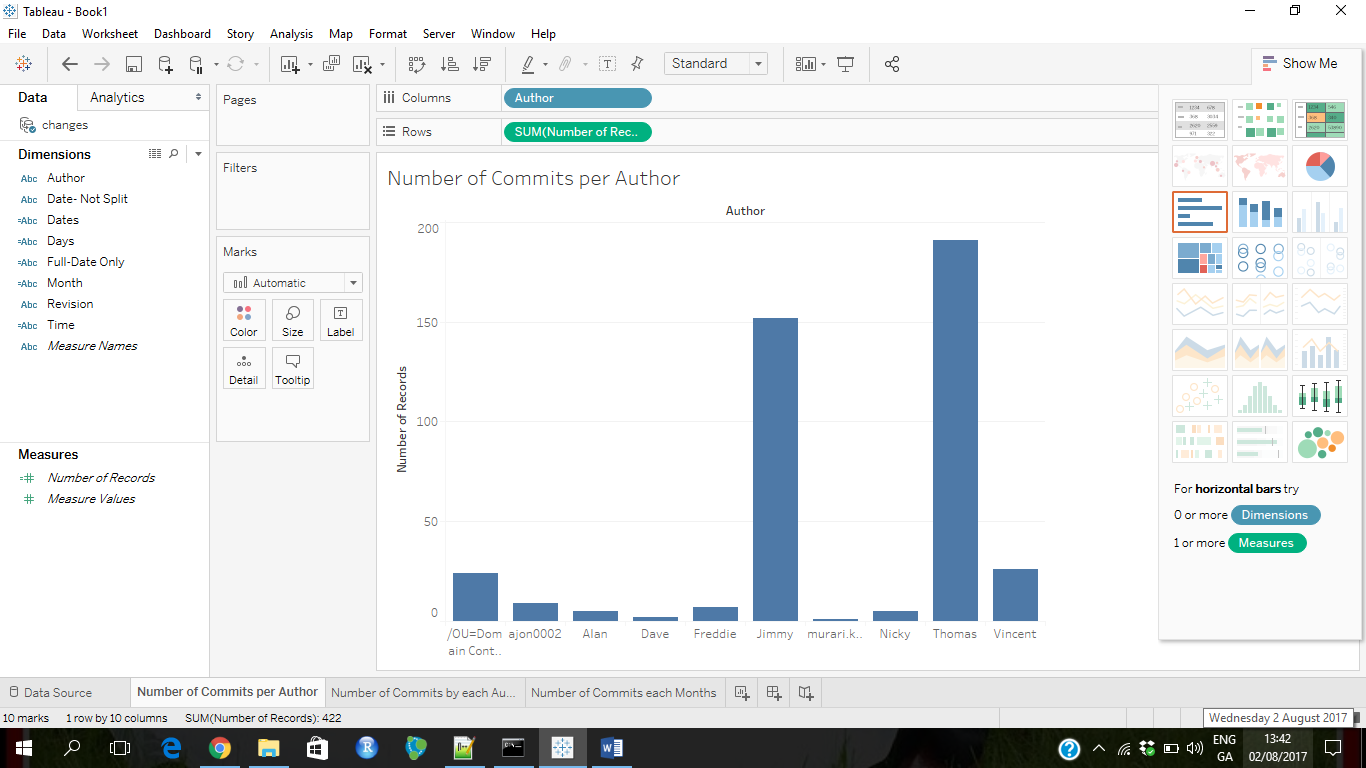


Fig. 1 Number of Commits per Author

1. **Number of Commits by each Author per Month:** It was also of interest to look at the number of commits made by the different authors per month. It was observed that some authors had no commits made in some months while it is also observed that the month of September had the lowest amount of commits. See fig 2 below.

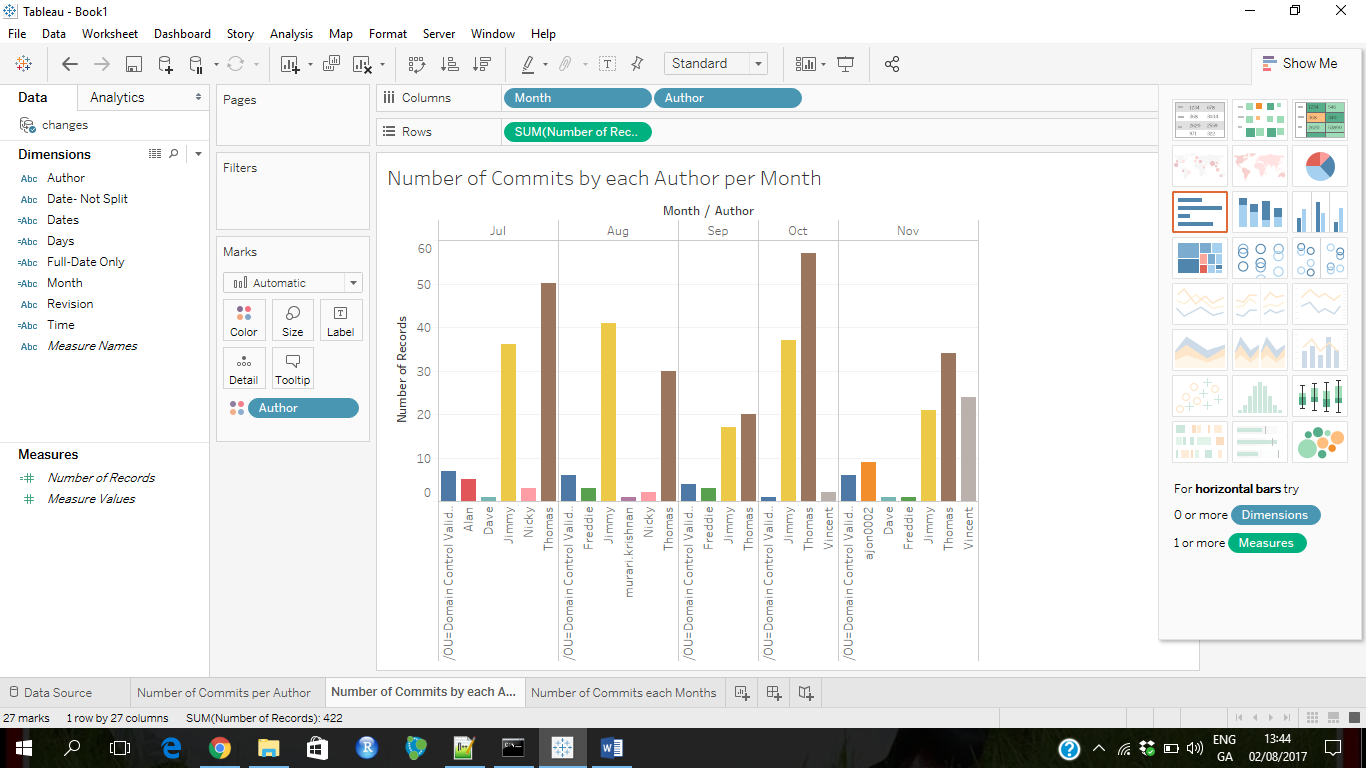


Fig. 2 Number of Commits by each Author per Month

1. **Number of Commits each Month:** From the dataset, number of commits each month was also considered. It was observed that July had the highest record of 102 commits made. This is very useful in that in could tell the month with the highest productivity and that is very useful information for any organisation. See fig.3 below.

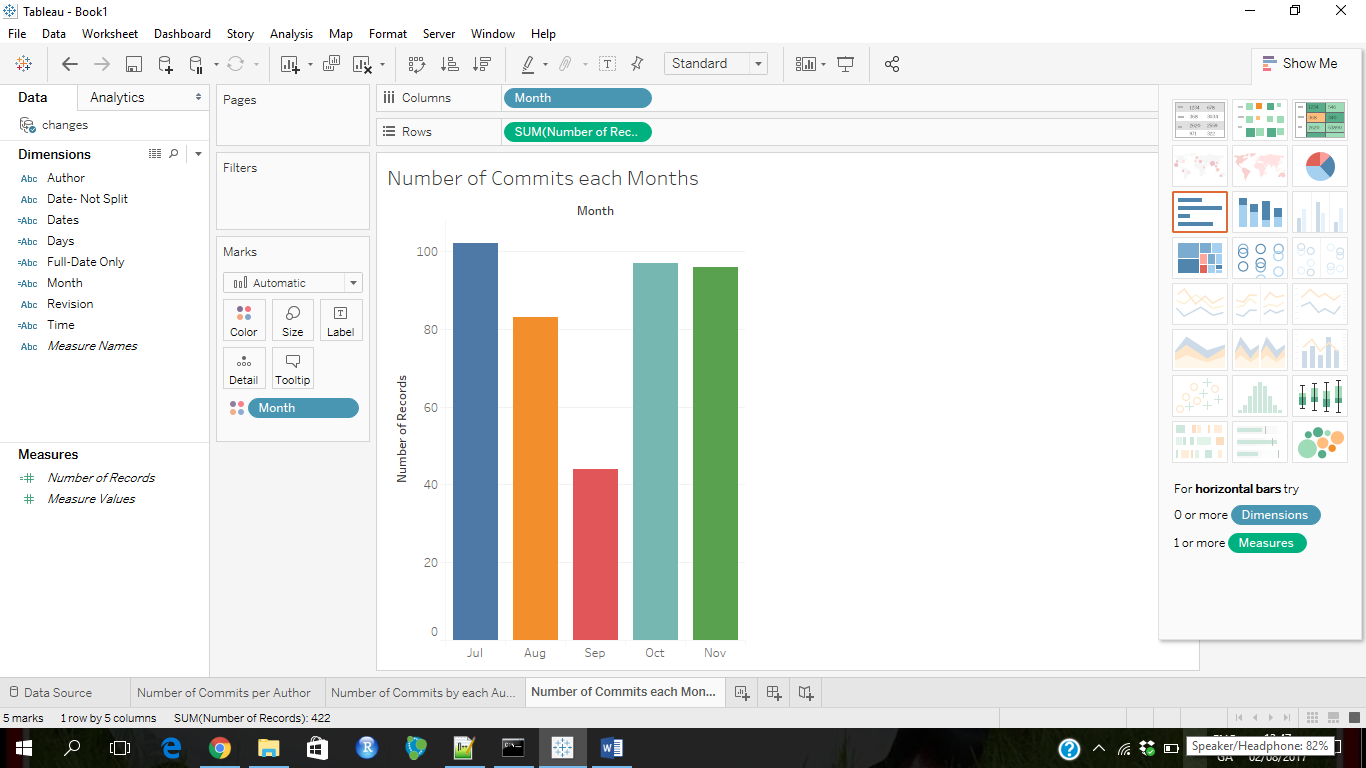


Fig. 3 Number of Commits each Month

**CONCLUSION**

Analysis of the dataset produced a substantial amount of information that is very crucial to the growth of any organisation. Information about a staff who is redundant could be obtained from the number of commits the staff makes. Also, organisations could see and compare productivity for each month.