Programming for Big Data

Perform analytics on a 5000 entry dataset

Soraia de Freitas Santos

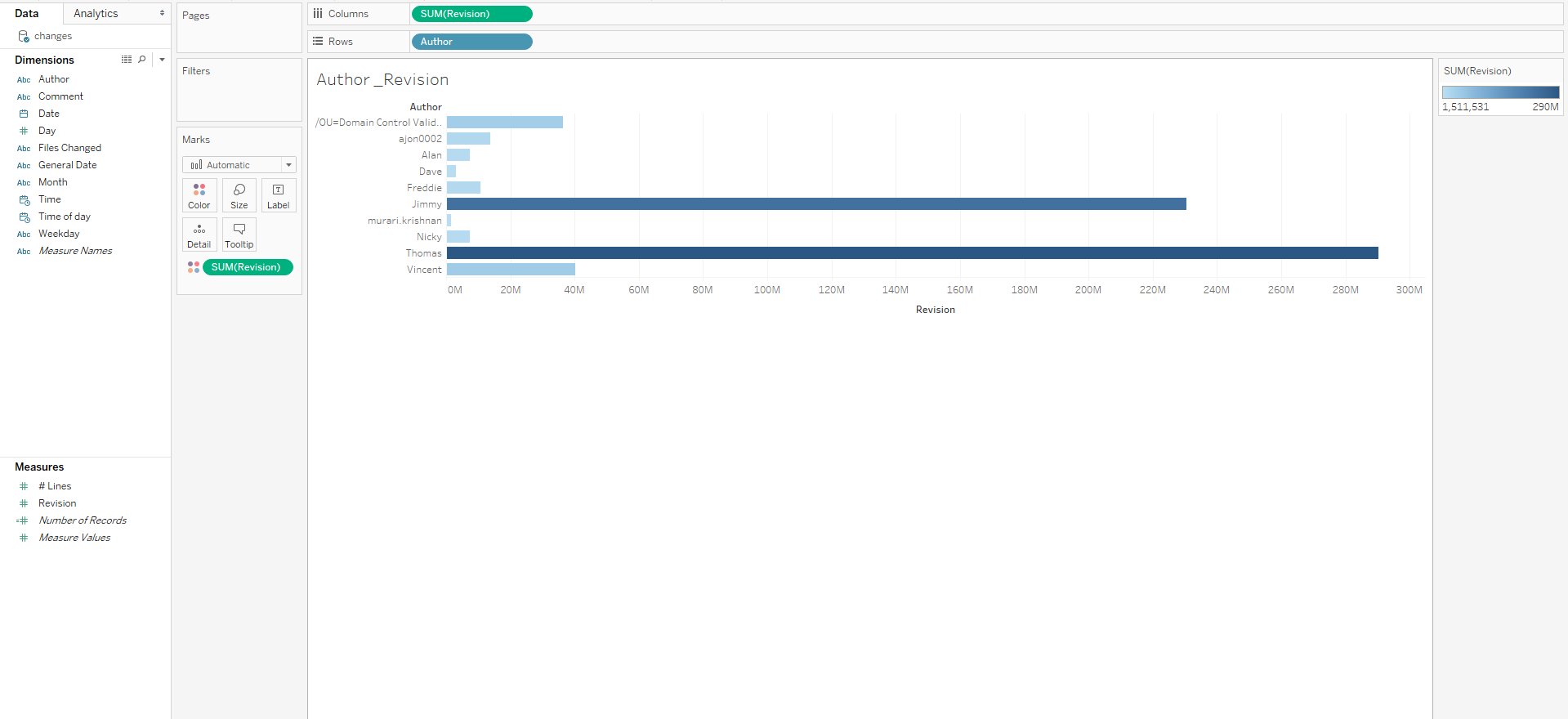
ID 10151853  CA\_05

When we received the file “changes\_python.log”, it was with 5,256 lines, so we split it in python and produce csv file. We scrub data further in CSV to be able to explore sufficiently, according to Wikipedia;

*"Data scrubbing is an error correction technique that uses a background task to periodically inspect main memory or storage for errors, then correct detected errors using redundant data in the form of different check sums or copies of data.”*

After all this process the csv file become 422 lines. After organising the data, we can have prepared to be visualized in the tableau tool, for more exploration visualization or analysing.

The data Bellow we can read this graphics, where nine author had their books reviewed, by the total of 1,511,531 reviewers. Who has the most review was the author, Thomas that got 290,256,683 of the revision.



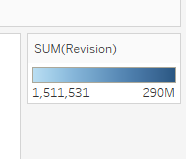
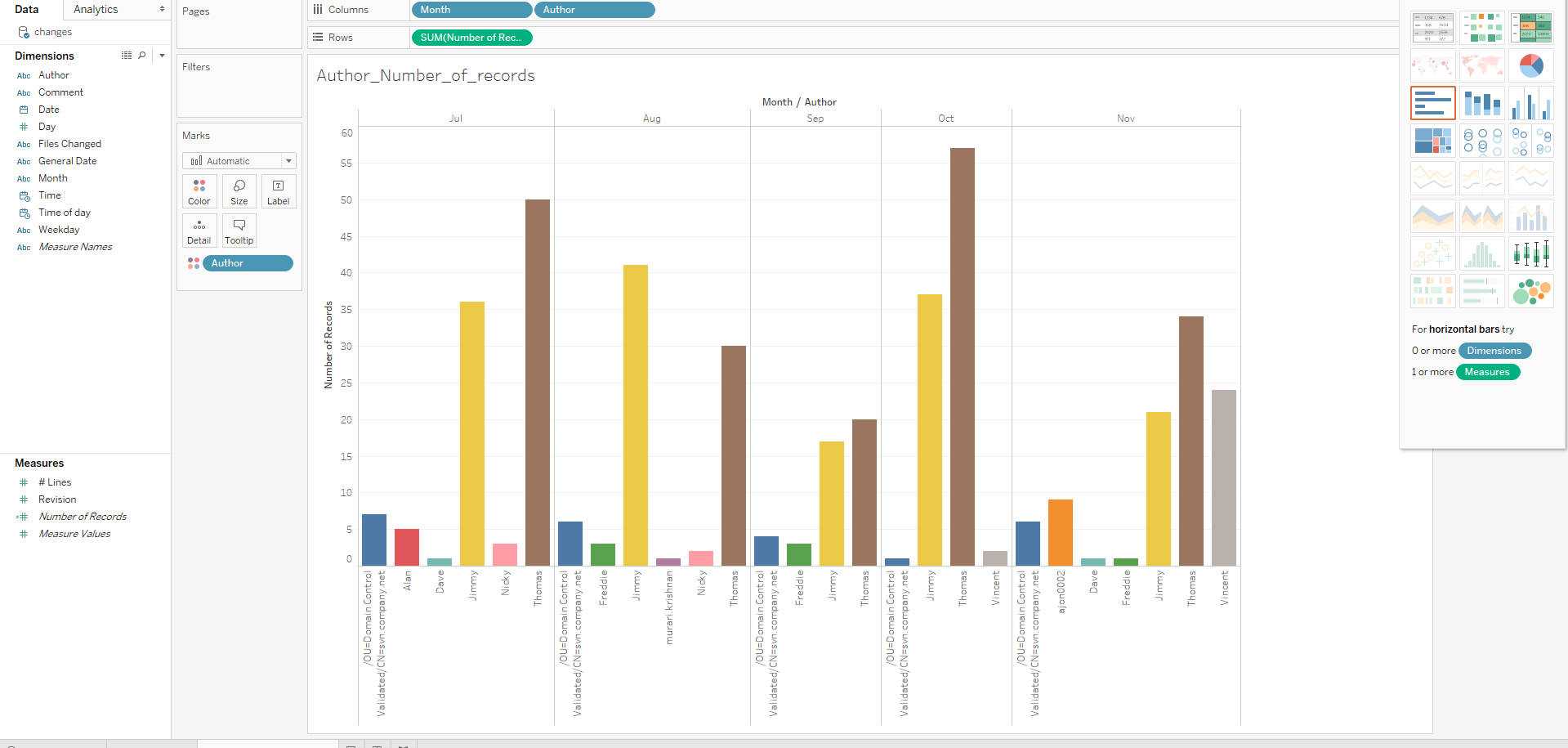


Tableau gave us a beautiful graphic of who had the most quantity of review.

**In this second visualization graph we can observe the author number, of records by month.**

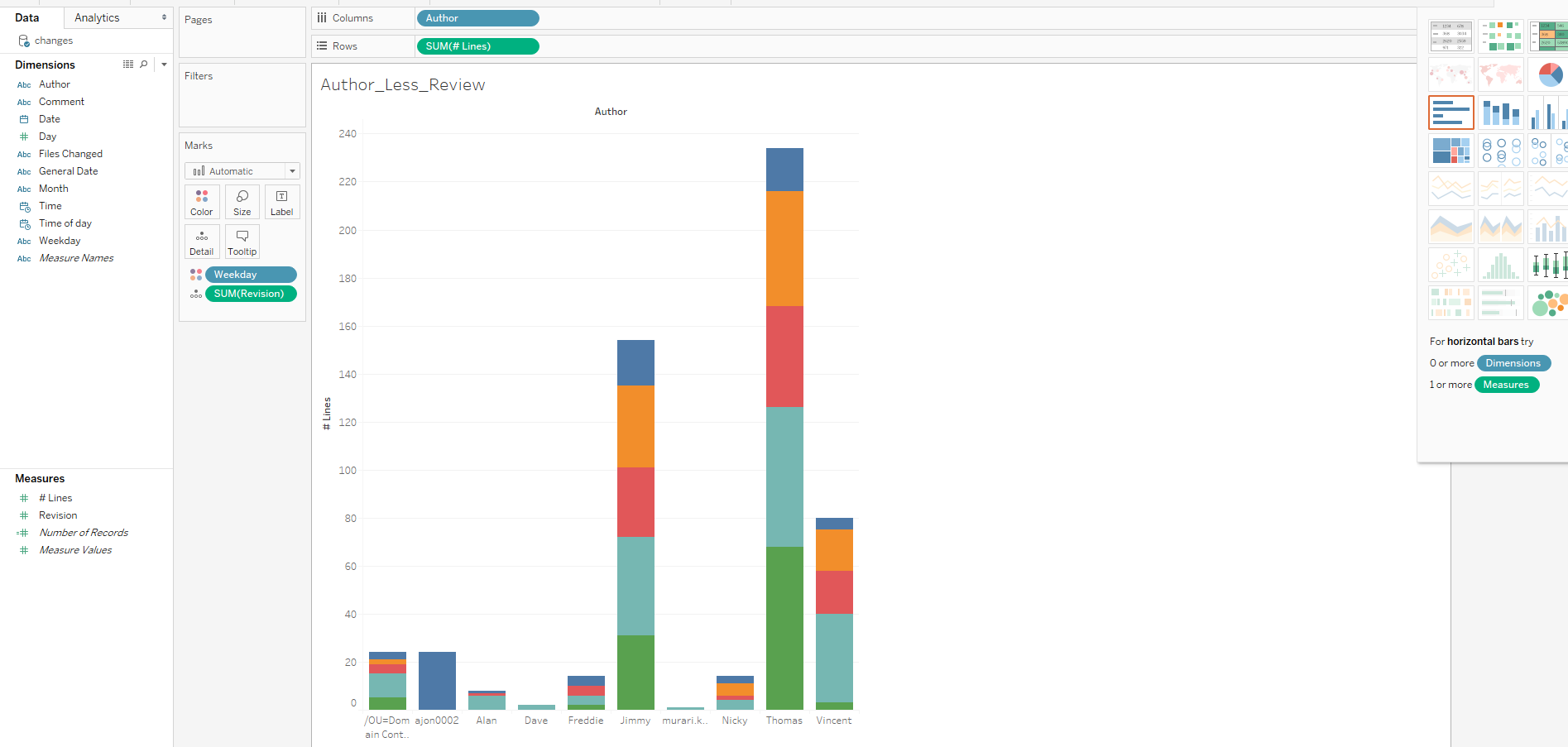


**Number of records in the month of June was:**

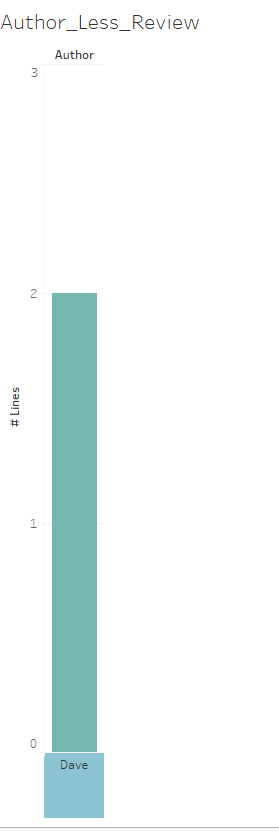
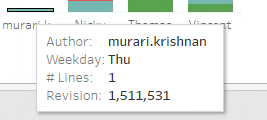
1. Thomas with 50 records
2. Jimmy with 36 records
3. The name of this author is not in the file just the domain, but he had 7 records
4. Alan was with just 5 record
5. Nicky had 3 records
6. Dave was in this month in the last place with just 1 record

But in August we had a difference, where Jimmy was in the first place with 41 records, followed by Thomas that had number of records of 30. In September and October Thomas occupied again the first place and again was followed by Jimmy, just in November the author Vicenti passed Jimmy, but with a small different and Thomas was always laddering the first place.

**This other graphic is about the author that had less review.**

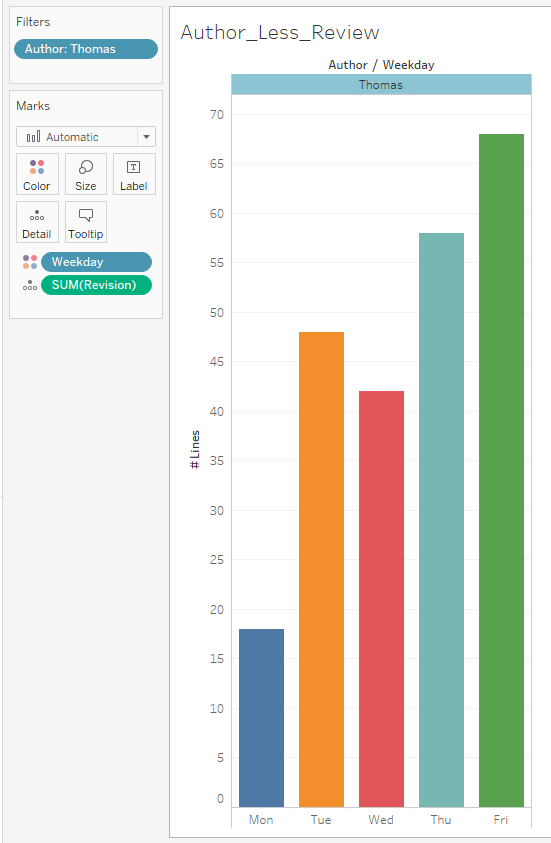


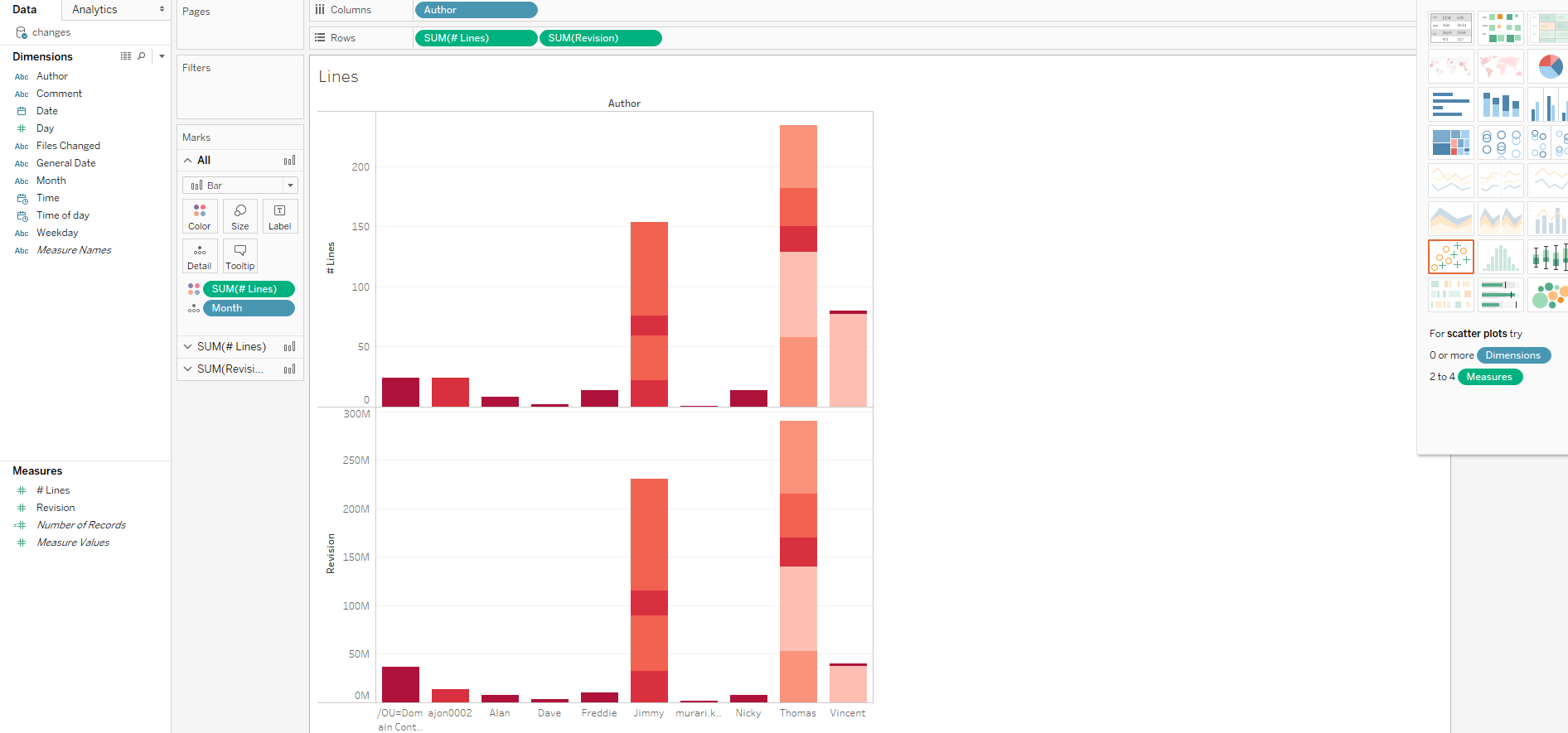
In the position of the author with less reviews of the week day, on Thursday, we had the author Dave and Murari, Dave got 3,040,167 revisions and Murari had 1,511,531, revision. See the graphic bellow.

Dave competitor, the author Thomas in his week day Friday had 83,903,410 reviews with this revision he got the first place for the week.

In the next graphic we can visualise, the quantity of lines that the authors readers had written to them. Here as well the author Thomas got the first place he had in October 71 lines and 87,348,824 revisions, it was the highest for that month, on another hands following him, we had in August Jimmy with 42 lines writing by his readers.





Now we can visualise the Dashboard. In the dashboard below the four various visualization graph represent the analysed dataset with different variables.

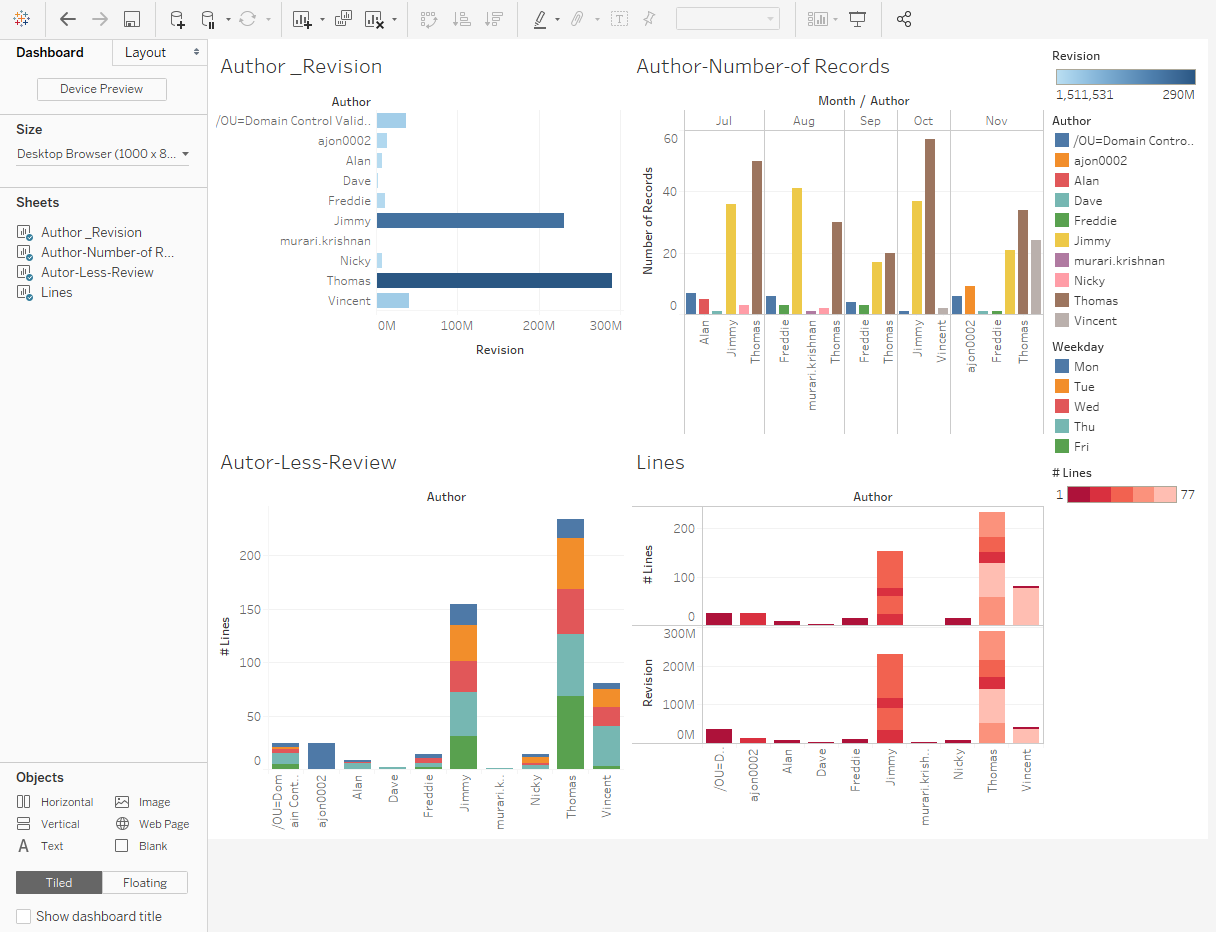
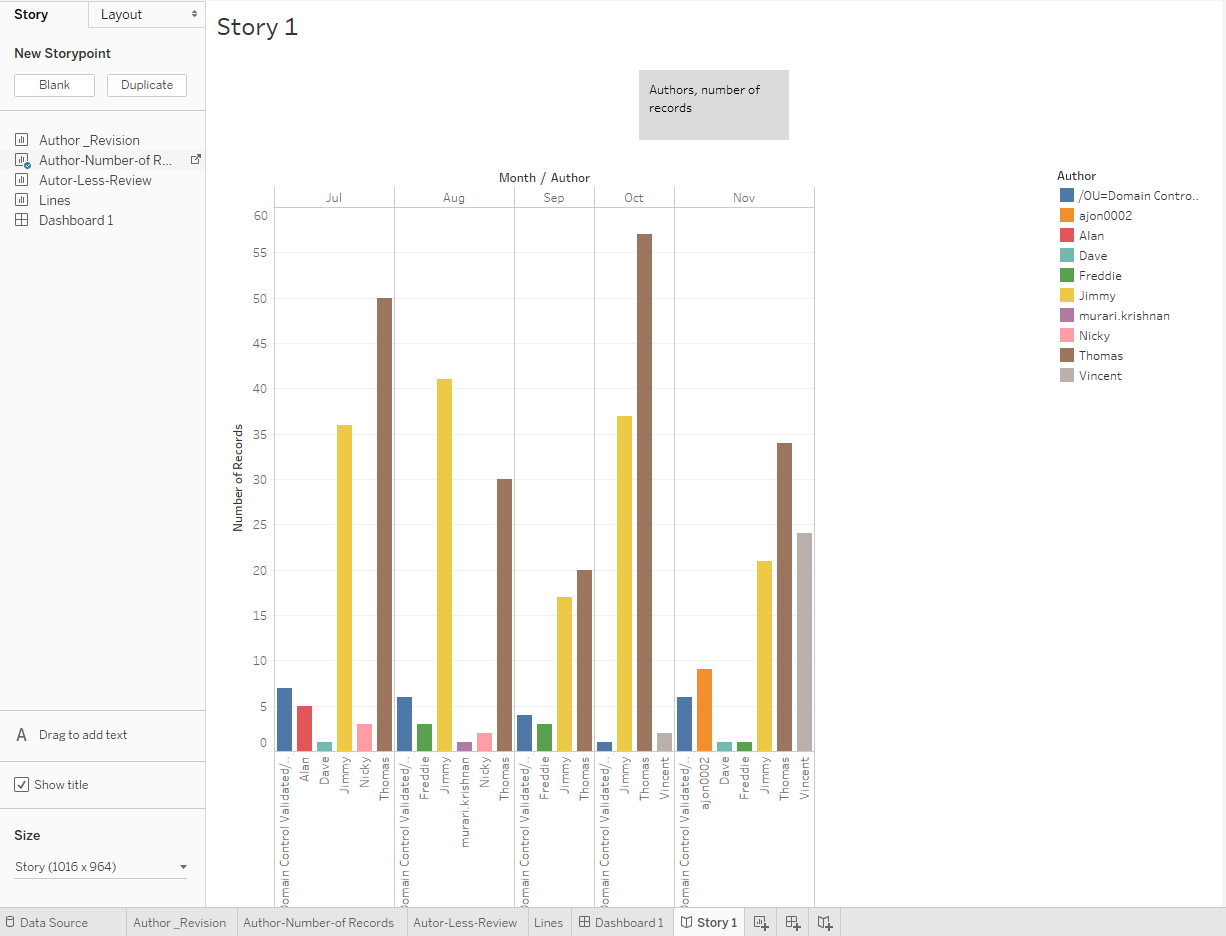


Tableau created this last graphic, where we can have a graphic of the story, where contains, authors number of records.



The result is that, Thomas in the total had the biggest number of reviews, records and lines measured by the people that likes his publications. With all these process was possible to measurable these 10 authors.