|  |
| --- |
| **HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS** |
| **assignment for PROGRAMMING for BIG DATA** |
|  |
| **Prepared by Changaiz Akhtar**  **Student ID 10171785**  **Date:02/08/2017** |
|  |
|  |

|  |
| --- |
|  |

**Contents Page No.**

**Introduction**………………………………………………………………………………………………. 2-2

**First Visualization** ……………………………………………………………………………………… 3-4

**Second Visualization** ………………………………………………………………………………… 5-6

**Third Visualization** ……………………………………………………………………………………… 7-7

**Forth Visualization** …………………………………………………………………………………….. 8-8

**Dash Board** …………………………………………………………………………………………………. 9-9

**Story Book.** …………………………………………………………………………………………………… 10-10

**Introduction**

Assignment is based on transferring a large dataset in text format over 5000 lines of text.

Need to clean the data and place into a relevant holder or container objects. Once in these objects there are 422 different sets of commit objects.

What I have to do analyse these objects and come up with three statistical pieces of information.

I received the file “changes\_python.log”, it was with 5,256 lines, so I split it in python and produce CSV file. I 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.”

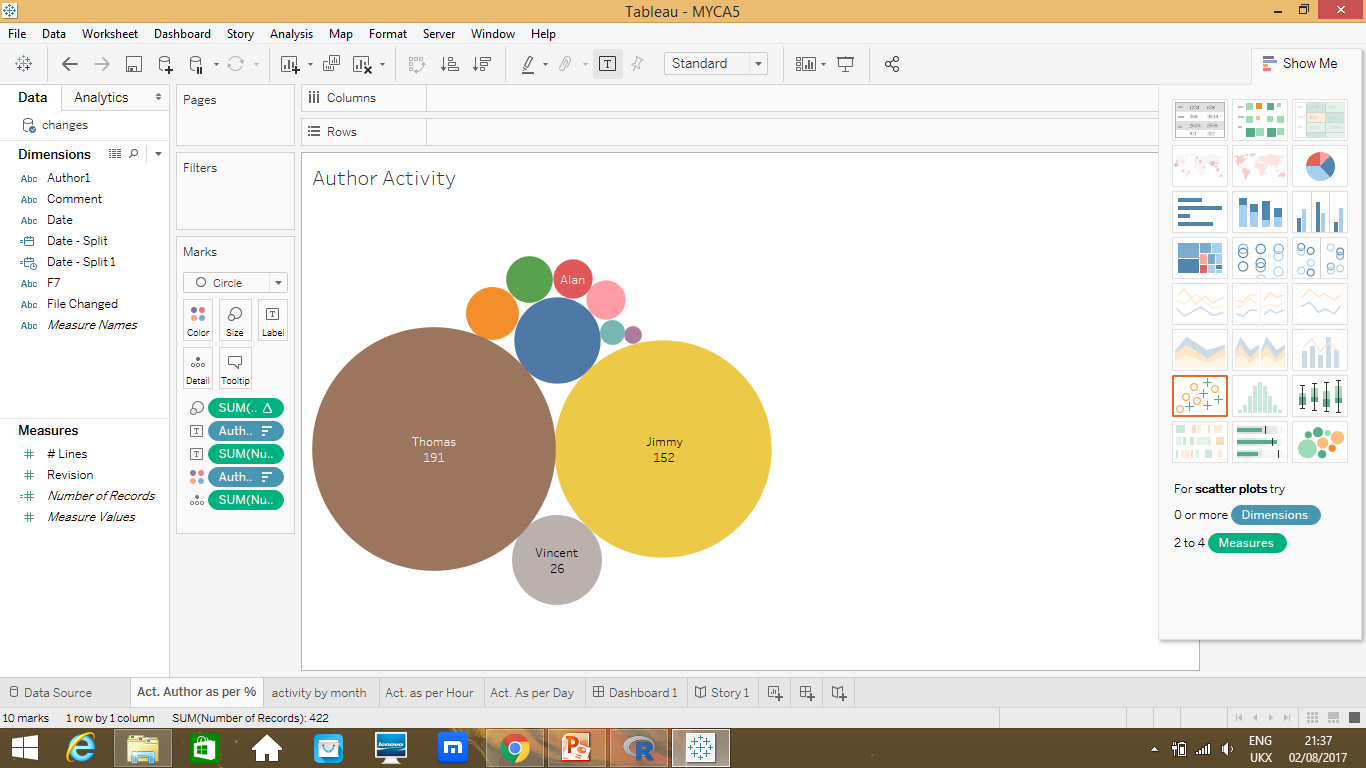
After all this process, the CSV file becomes 422 lines. After organising the data, I have prepared to be visualized in the tableau tool, for more exploration visualization and analysing.

After carefully read and analysed the data I suppose there are IT company workers who writing and changing blogs and these are called authors and I will used tableau to find some statistical interesting activities and explore the data with the help of grapes.

Four key activities of authors I will explore through statistical graphs.

* **Activities of authors as per number of records.**
* **Activities of authors as per month.**
* **Activities of authors as per day.**
* **Activities of authors as per hour.**

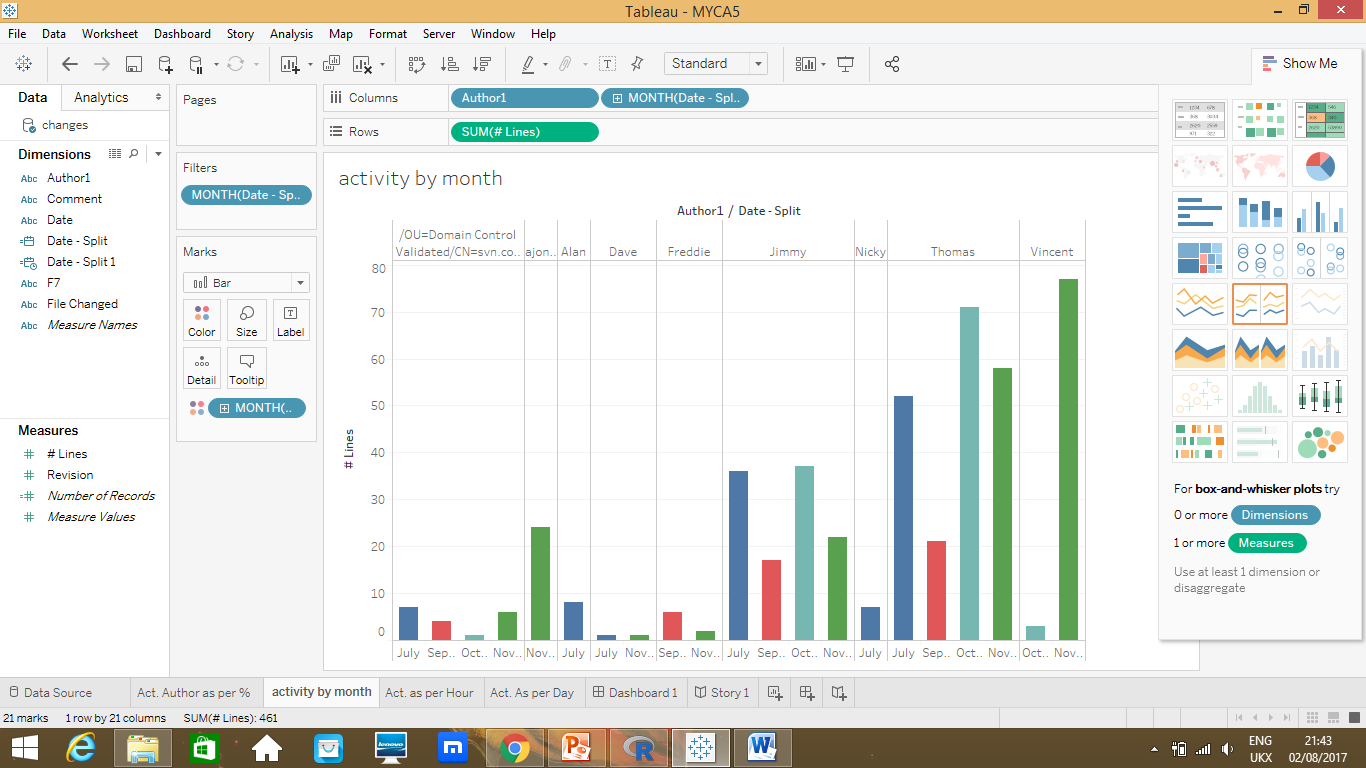
**FIRST VISUALIZATION**



**Comments:-**

* From the data set, I have selected the author’s activities by using packed bubbles where I used one dimension and two measures to find out the interesting statistical activity.
* Thomas has high number of 191 records and his activities are 45.26% following Jimmy 152 records with 36.02% and the lowest number of records by Murari with 0.24% activities.

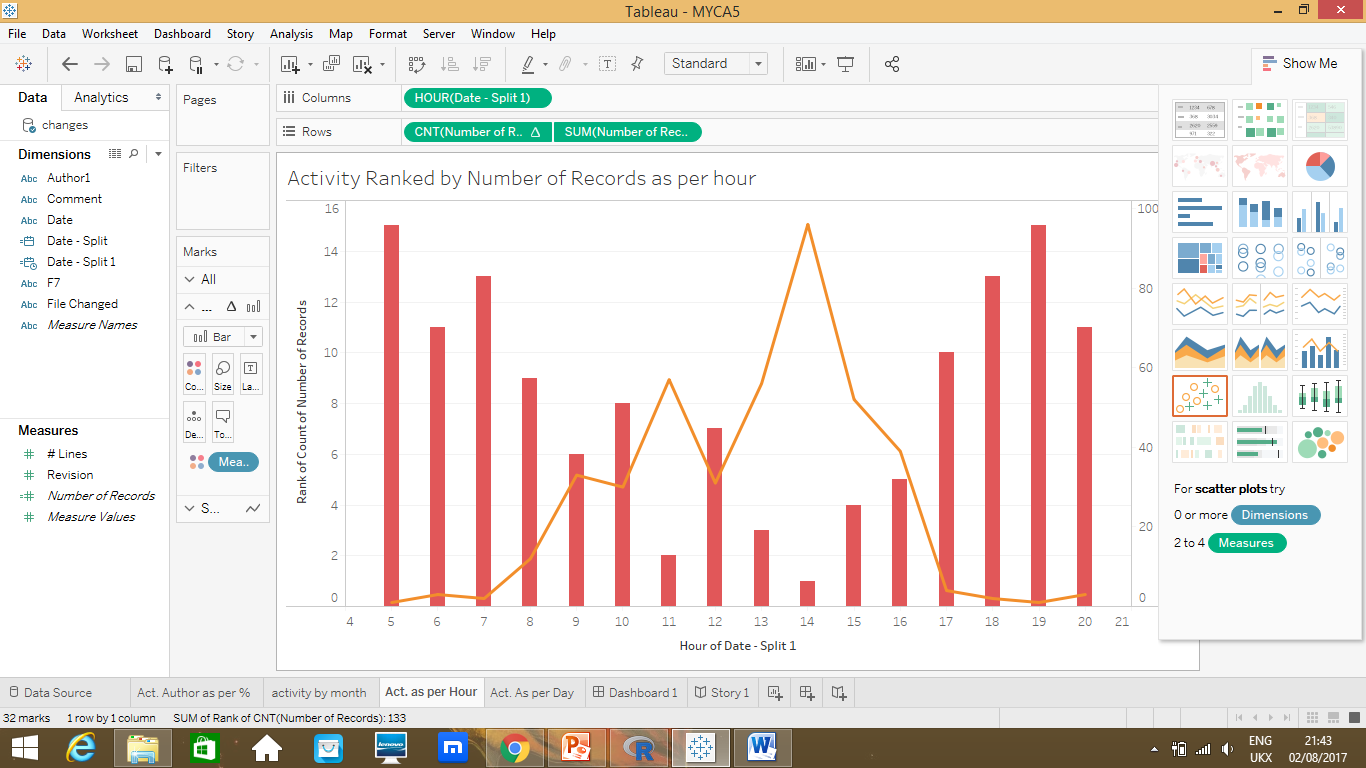
**SECOND VISUALIZATION**



**Comments:-**

* From the data set, I have selected the author’s activities as per month by using side-by-side bars, on x-axis months and y-axis number of lines written by each authors.
* As I can see from the above diagram, each authors activities on each month.
* For example, Vincent has written 77 lines in the month of November and 3 lines in the month of October.
* On the other hand, Thomas has written 58 lines in number , 71 lines in October,21 lines in September and 52 lines in July so from the above diagram Thomas has written highest number of lines as compared to others authors

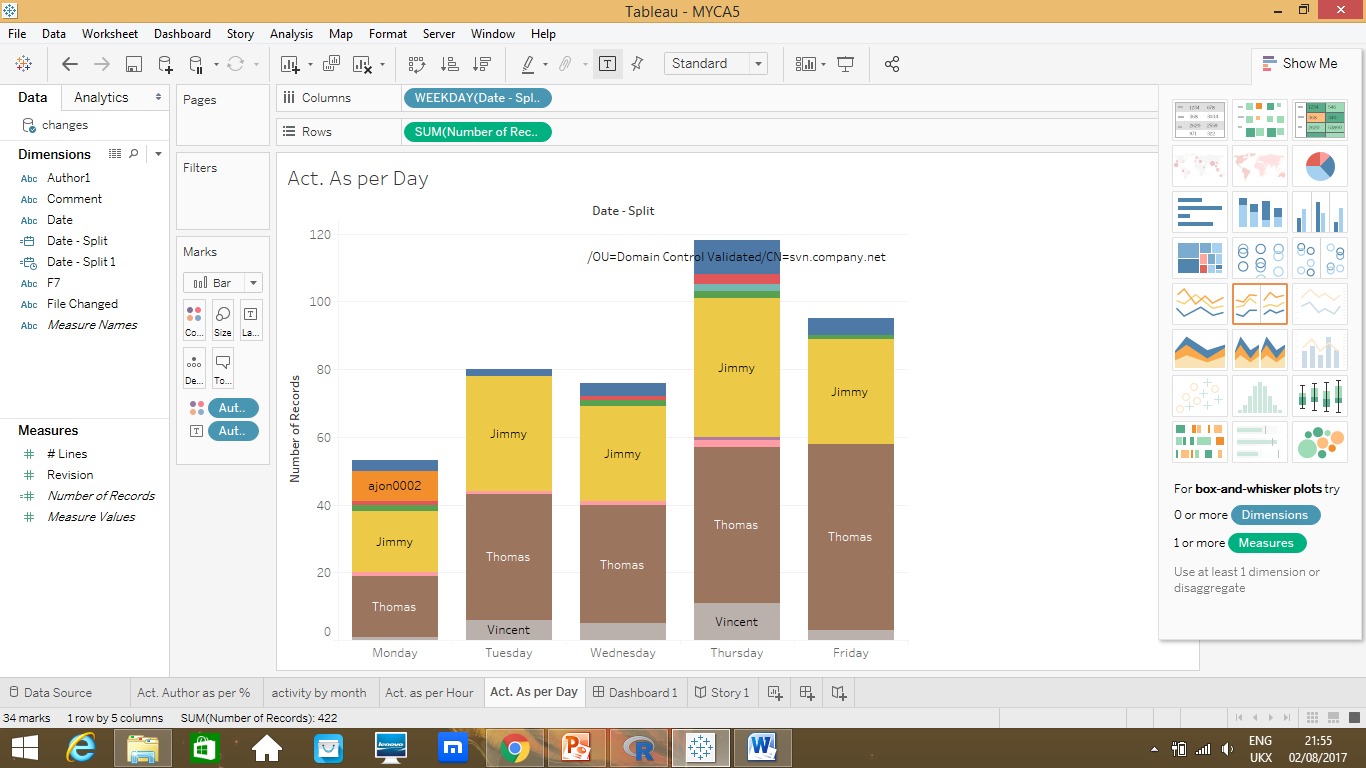
**THIRD VISUALIZATION**



**Comments:-**

* From the data set, I have selected the author’s activities as per hour by using dual combination try, on x-axis hours, which is started from 4am to 9pm and on the y-axis number of records. Here I want to see which hour was busy and which author has more records.
* From the above diagram, I can see the busy hours during the day are 13.00 to 15.00 hours.
* The highest records were count at 14.00 and are 96.

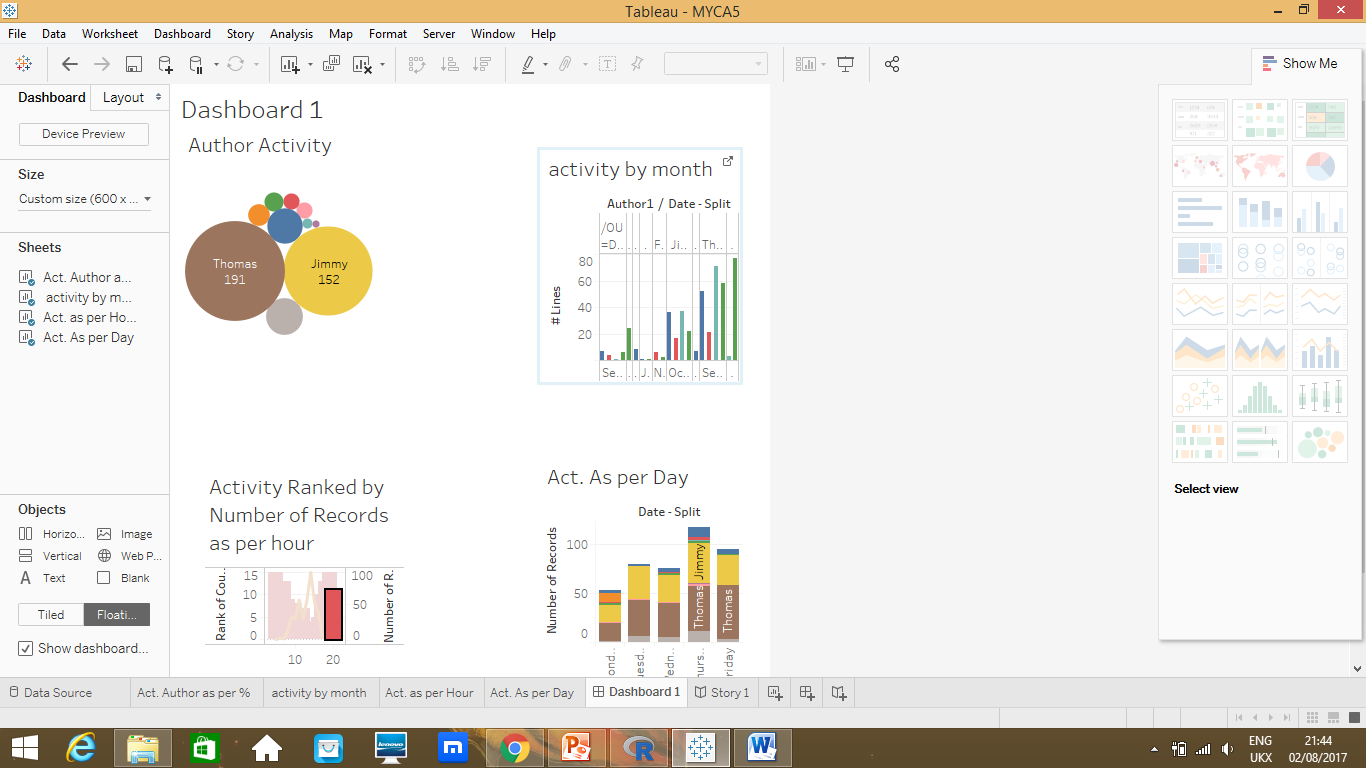
**FORTH VISUALIZATION**



**Comments:-**

* From the data set, here I have measured the authors activities as per day and I used stacked Bars, on x-axis days and on the y-axis are the number of records written by the authors..
* From the above diagram, I can see the busy day is the Thursday; in the week, second busiest day is Friday.
* Thomas and the Jimmy both are the top from all others authors.

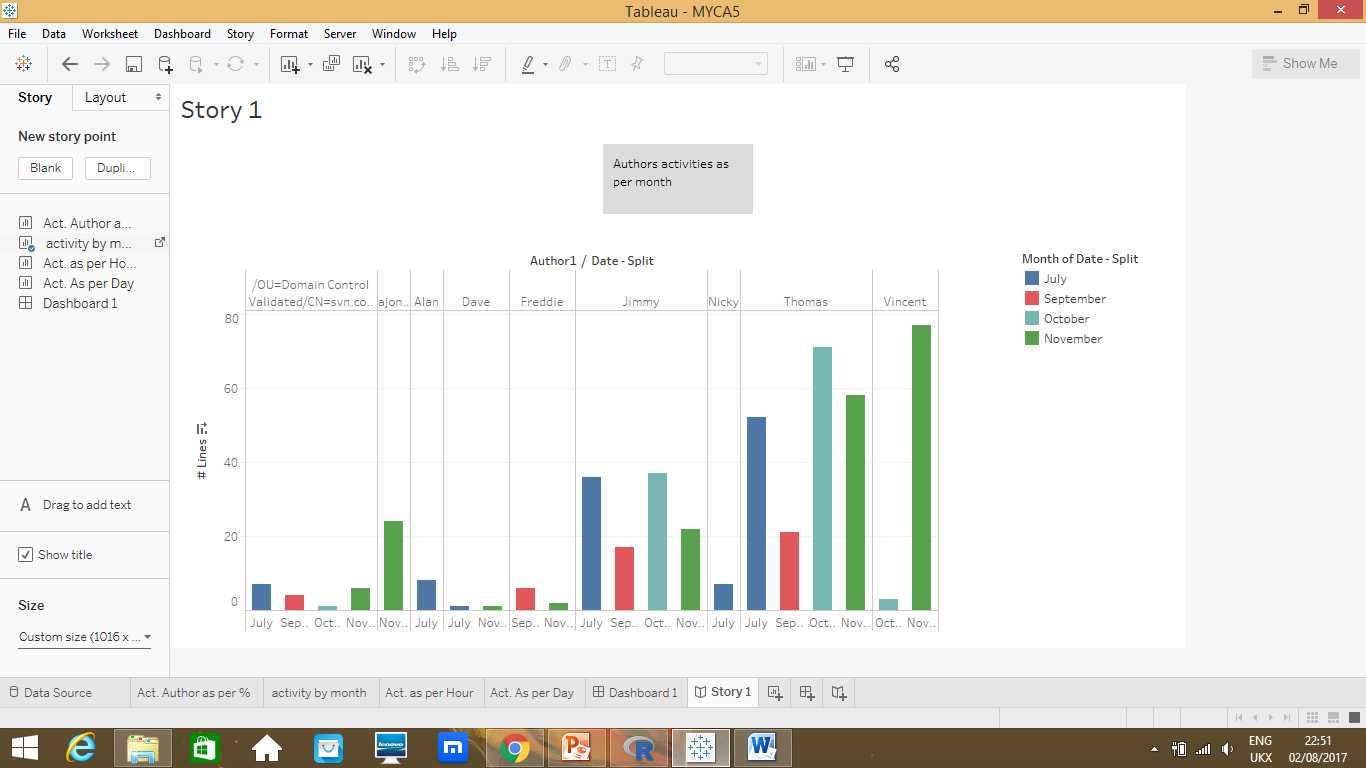
**DASHBOARD**



**Comments:-**

* In tableau I can visualize all the graphs created above on one single screen and can import and export to users very easily.
* I can visualize from single value or transaction to very large value on one screen.
* Here I can visualize author’s activities as per month, as per day and as per hour on one single screen.

**Story Book**



**Comments:-**

* In Tableau, we can tell our business or any other type of story through statistical graphs.
* Here I told the story of different authors and there activities as per month and I can see Vincent has high number of records.