****

Agnieszka (Aga) Wolska

Student #: 10361952

Programming for Big Data – CA 4

1. **Data extraction and preparation**

In assignment 4 a csv file was imported from Python code with class objects and testing. The file contained 421 lines (+ headers) of data describing activity of Github users that were committing changes to a project through a period of 5 months.

After running a program in Python, data preparation changes in Excel were made. While checking for data consistency possible errors were found ((“/OU=Domain Control Validated/CN=svn.company.net”). It seemed like automated domain changes were counted as changes made to a project by authors, so the rows containing it were removed from the file, leaving 398 rows of data. No other errors or duplicates were found.

In order to work and further prepare data for analysis I decided to add additional columns:

* *“Time of day”* – data was extracted from *“time”* column and I created an additional one that assigns a time that changes were made by authors (0-6 is “night”, 6-12 “morning”, 12-18 is “afternoon”, 18-0 is “evening”)
* All data was sorted based on revision number (progress in changes made by authors)
* Later on due to the fact that I decided to analyze and visualize data set in Tableau, the data was imported and saved as html file (Tableau does not work with csv files), as well as after importing data into the tool I decided to split first column – *“revision”* into two columns, in order to remove r from data, so that the tool could easily read whole data included in it as a numeric value.

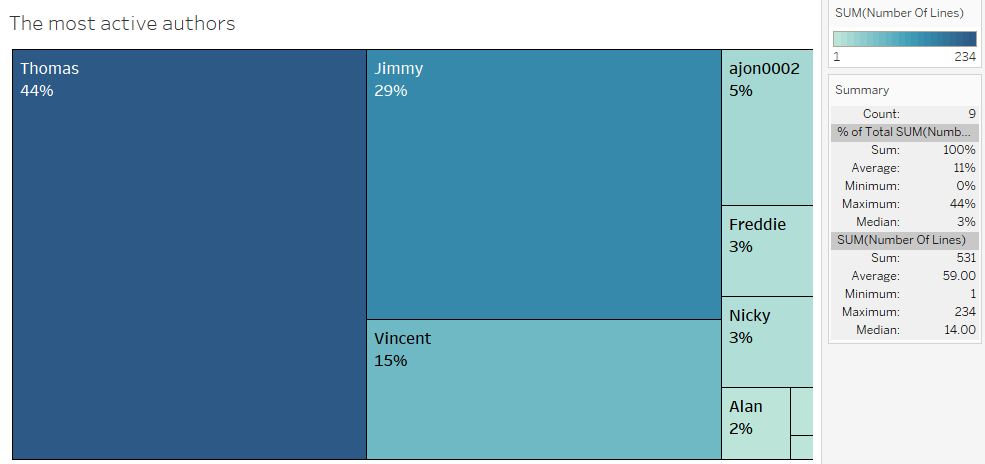
1. **Data analyze and visualization**

As mentioned further analyses and visualizations were made in Tableau. I decided to answer th0ree questions:

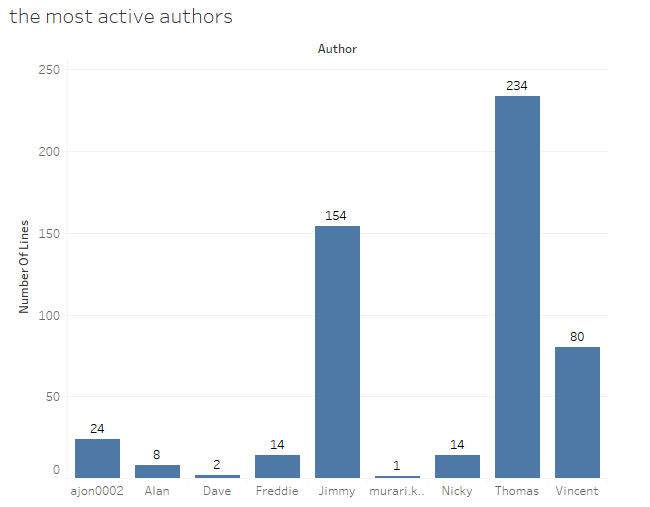
* Which Github user was the most active (i.e made the highest number of changes to the project?
* How time affected activeness of the users?
* Which users were contributing to the project and different stages?

In order to find an answer to that below graphs were prepared:

1. **The most active authors**



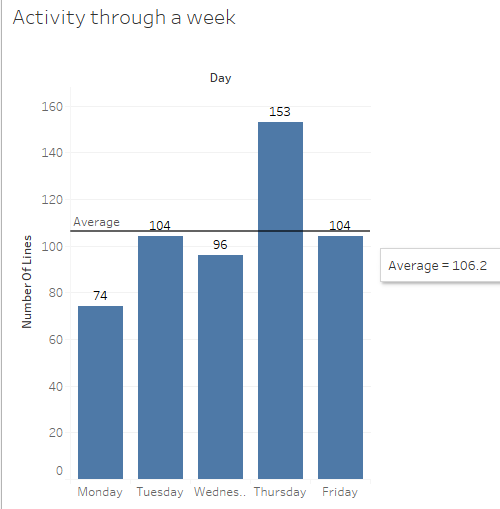
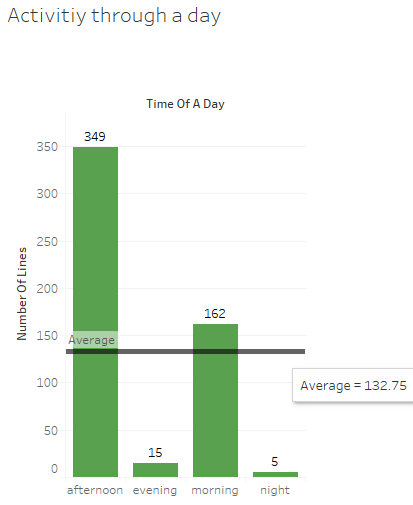
Pic. 1. Activity by user in percentage



Pic. 2. Activity by user with details in number of lines

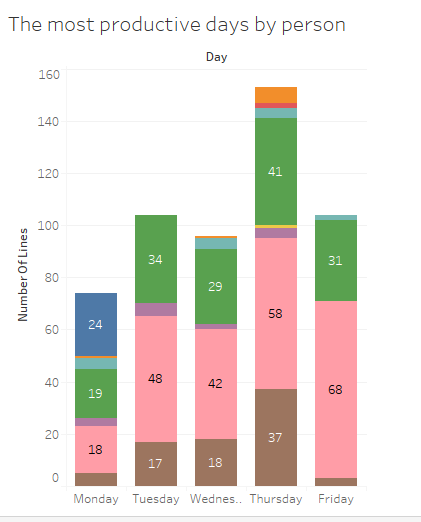
By using a treemap with marked percentage of contribution to the project based on number of lines that were added to it and bar chart that is showing the difference in activity. Within 9 authors we can say that with no doubt that Thomas was the top performer, followed by Jimmy and Vincent. Interestingly we can see that the range between the top performer and the least active user was 233 lines of code, with most lines created by Thomas (234) and the least created by murari.krishnan (1).

1. **The most active time of work**

Pic. 3. Activity based on a day Pic. 4. Activity based on time of a day

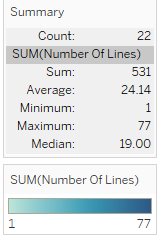
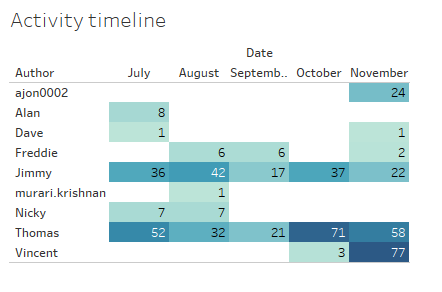
Based on visualizations above we can say that the most productive day in which most of the changes to the projects were made was Thursday (153 lines of code) in comparison to the least productive day - Monday (74 lines of code). Additionally we can see that the time of a day had a big impact on productivity of our authors who were most active during afternoon (between 12 – 6 p.m). Almost no changes were made in the evening and during a night (6 – 12p.m and 12 – 6 a.m).



Additionally we can observe by adding authors into our previous analyses that some of them, like Thomas, Vincent and Jimmy were working every day on the project, while other like Dave, Nicky or Alan were more selective and were only making any progress with it on certain days.

Pic. 5. Activity based on a day by author

1. **Activity of users through a project timeline**

****

Pic. 6. Activity of users through a project timeline

Moreover in order to discover how activity of authors changed throughout the project the above graph was prepared which shows activity of each user through 5 months time with details on their productivity (number of lines). We can see on the legend that the brightest colors are equal to the least number of changes made to the project and the darker the color gets, the more active user was. According to the graph, some users stopped working on the project in the early stages of It (like Alan), while others started participating just at the very end (ajon0002).

1. **Conclusion**

This analysis demonstrated how Python, Excel and data analytic tools like Tableau can be used together. In each of the steps of data analyses each tool was used to prepare a set of actions that enabled us to get an insight into the data and get a better understanding of it. All of the tools are very powerful and while in some cases they can be used by themselves through a data analytic process, I decided to focus on their strong points and use them all together for different purposes. In that way Python was used to format the data, Excel was mostly used in data preparation stage and Tableau which is a data visualization tool was used to visualize and analyze the data. Thanks to that we were able to answer three questions that were asked before performing the exercise. We discovered the most active authors with details of their contribution to the project. We understood how specific time of the day and week impacted the activity results. Moreover we could see how productivity and contribution of each author changed through a period of 5 months during which they were working on the project.