Michelle Fahy

10362430@mydbs.ie

Abstract

This report looks at the CSV file and visualisations for CA4 and discusses the interesting statistical information gathered from it.

CA4 Report

The data for this assignment is based on a work project for a phone application. The data comes from the Git Hub submissions from the members on the project in the form of a log.

The CSV file has the following information:

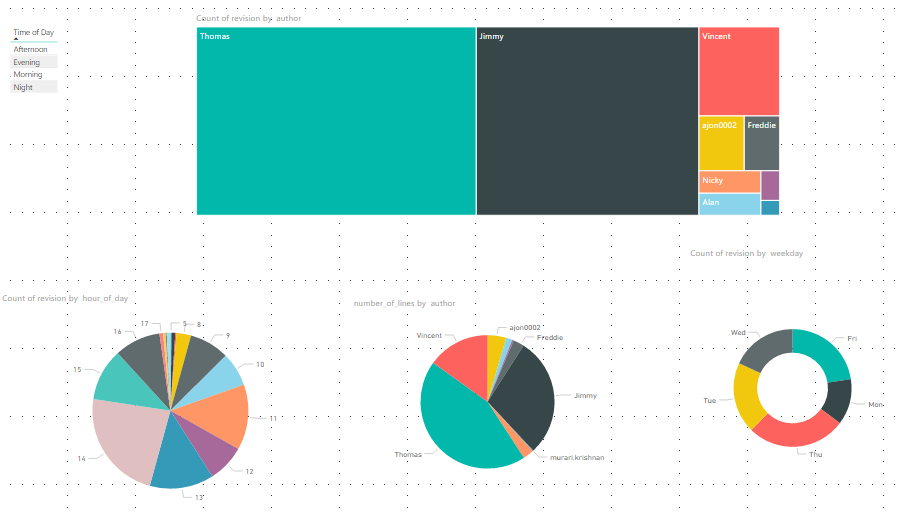
Revision number (unique ID), Author (employee name), date, time, number of lines, change type in which I took the first letter of each line – M, D, A (merge, delete, add) and comment. I have also created code to extract the hour of day and the weekday the work was processed on.

I have taken the CSV data and uploaded it into Power BI to create visualisations to help show the statistical information that may be of interest. While work with the data in Power BI I created an IF Statement formula to split the hours of day into Time of Day – Morning, Afternoon, Evening and Night. This was used to create a filter for viewing the data.

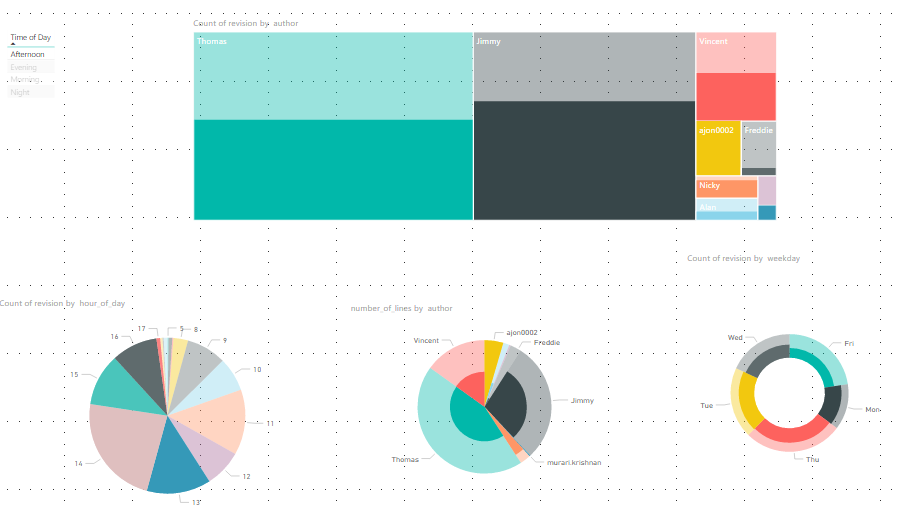
From looking at the visualisations created we can see at a glance (1a) that the main work has been processed by Thomas and Jimmy which would probably mean that they are the leads or main employees working on the projects. It may also mean that the others are not saving their work as often as they should be and may need to be looked into.

We can also see that the majority of the work is being carried out at around the 2pm mark and on a Thursday.

1a – Shows count of revision by author, Count of revision by hour of day, number of lines per author and count of revision by day of the week.



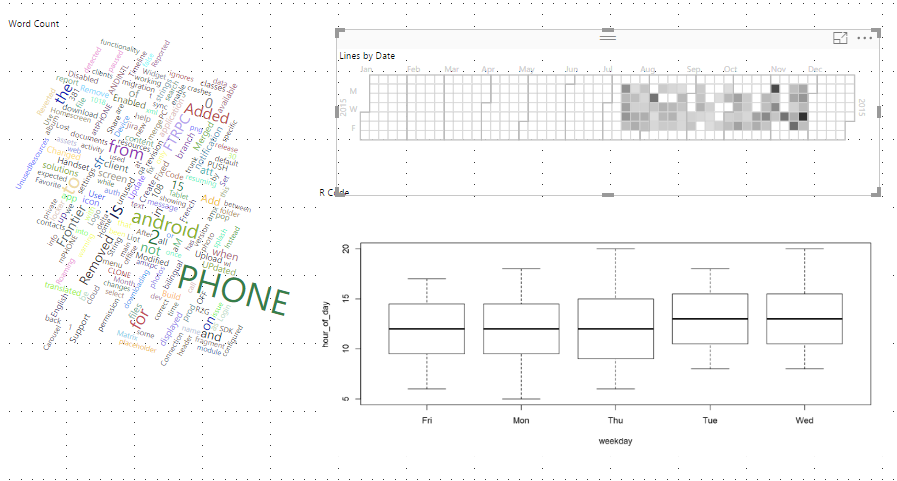
By applying a filter using time of the day image 1b shows us that the most changes to the log happen during the afternoon. We can see particularly that Jimmy saves most of his work during this time and this may need to be raised with him for future projects to ensure he saves his work more frequently.

1b

The next thing I looked at was the word count (2). Power BI has a Word Cloud that I used which I find is a very useful way to see the most commonly used words – in this example the words in the comments that the authors submitted when logging their changes. The more a word is used the bigger it appears in the Word Cloud. We can see that the most common word is Phone (which would make sense on a project for phone apps). More common words are Added, Removed, Android and Frontier.

I also looked at the lines by date and created a plot by using the R script visual. This highlights again that the majority of the commits are created on the Thursday.

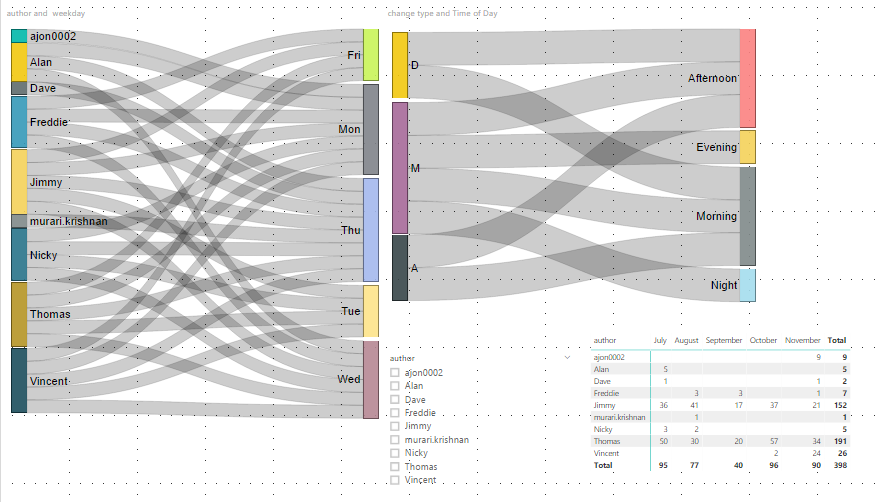
2 – Shows Word Cloud, Lines by date using the Calendar Visual and a box plot using R.



Finally, I made use of Sankey diagrams (3a) to show visualisations of the author to day of week and change type to time of day. Sankey diagrams are a form of flow diagram in which the width of the arrows is shown proportionally to the flow quantity. The items being connected are called *nodes* and the connections are called *links.* They are commonly used by Google.

The second diagram below show us that the most common change type was a merge (M) and these occurred at all times of the day.

3a – Shows the author to day of week and change type to time of day.



3b

