**CA4**

**Programming for Big Data**

**Elaine Treacy**

**10362944**

**Table of Contents**

Contents

[Assignment 3](#_Toc499558502)

[Input file 3](#_Toc499558503)

[Cleaning Data Using Python 3](#_Toc499558504)

[Analysis 4](#_Toc499558505)

[Rapidminer 4](#_Toc499558506)

[Power BI 5](#_Toc499558507)

[Excel 6](#_Toc499558508)

[Findings 7](#_Toc499558509)

[Commits by Author 7](#_Toc499558510)

[Commits by Month 9](#_Toc499558511)

[Number of Days of the week Authors carried out a commit 10](#_Toc499558512)

# Assignment

Assignment 4 is based on transforming a large dataset in text format - over 5000 lines of text. You will need to scrub (clean) the data and place it into the relevant holder/container objects. Once in these objects you will see that there are 422 different sets of commit objects.

So your task will be to analyse these 422 objects that are in a list and come up with 3 interesting statistical pieces of information for this dataset with supporting evidence of "interestingness'

You code for calculating the analysis should be documented and tested. Test should be in a separate file runnable from the command line.

Your statistical analytics conclusions should be in a word document explaining in approximately 500 words the information that you have gleamed from the dataset.

You will be required to submit your code via github along with all documentation and tests.

# Input file

A .html file was made available on moodle for analysis. The first step was to download and convert this file to a .txt format (i.e. process\_changes.txt) to allow me to proceed with the analysis.

As stated above this input file contains over 5000 lines of text representing commit actions made to Github by a number of users over a period of time (i.e. July – November 2015).

# Cleaning Data Using Python

Using python (process\_changes2.py) I cleaned the data into a readable .csv format (changes\_path.csv) with the following headings:

* Revision
* Author
* Date
* Time
* Number of Lines

Following the clean I was left with 422 different sets of commit objects.

I also noticed that one of the name of one of the author’s names appeared to be an error i.e. ‘/OU=Domain Control Validated/CN=svn.company.net’. There are 24 of these types of entries. Not knowing whether these errors relate to the other authors or not I decided to apply change the Author’s name for these entries to ‘Unknown’.

My next step was to carry out some simple tests on the process\_changes2.py file which I did using test\_simple.py. These tests simply looked to validate the number of commit objects and Authors names for various records.

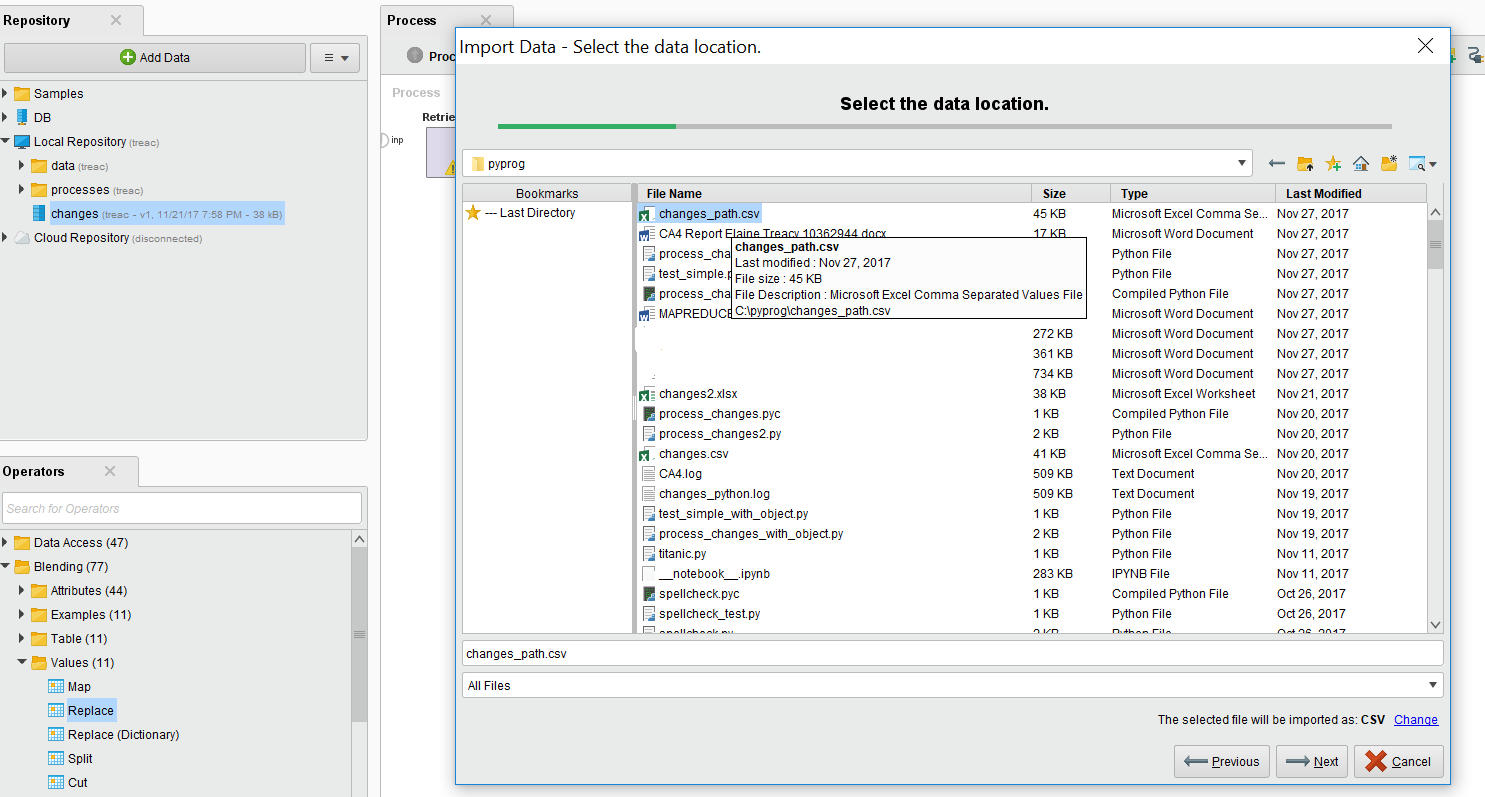
# Analysis

I decided to use a number of applications (i.e. Rapidminer, Power BI and Excel) to carry out my analysis as an exercise in improving my familiarity with these data analytical tools.

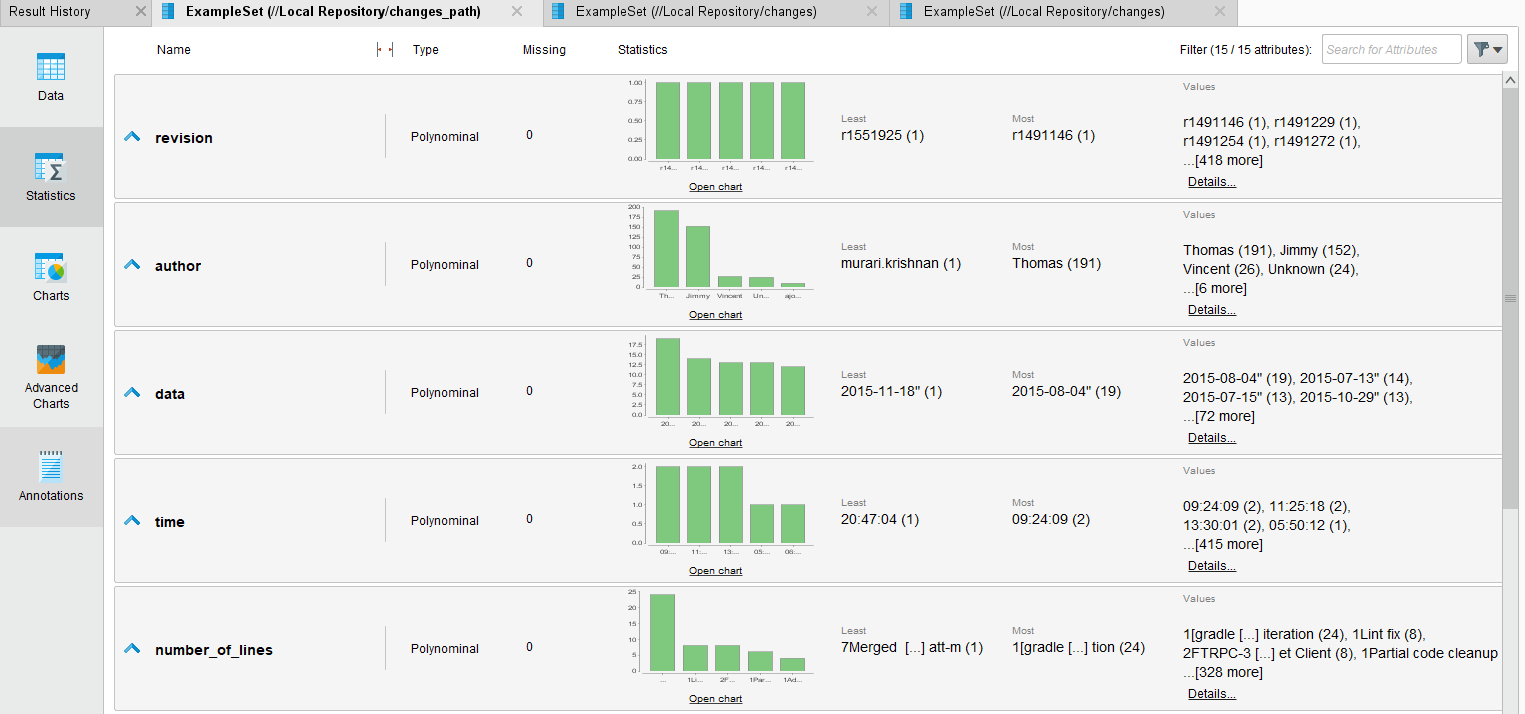
## Rapidminer

Prepping Rapidminer for analysis of the power\_changes.csv file

Import the process\_changes.csv file into Rapidminer and run through the various steps to clean the data. This was straightforward as I had already cleaned my data using python.



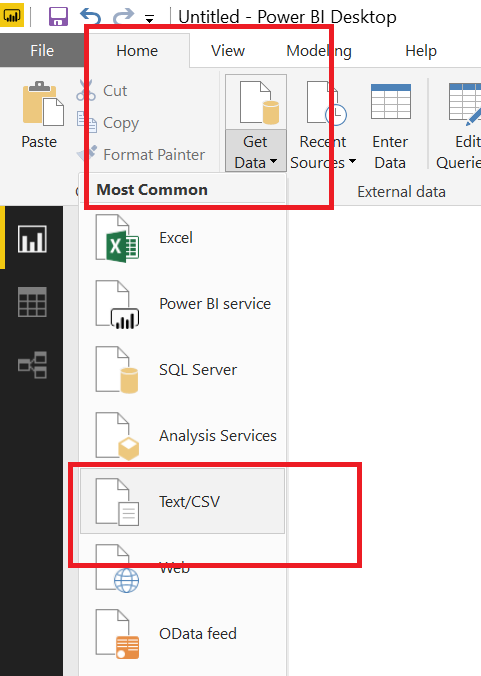
Click on the Statistics icon on the right-hand menu.



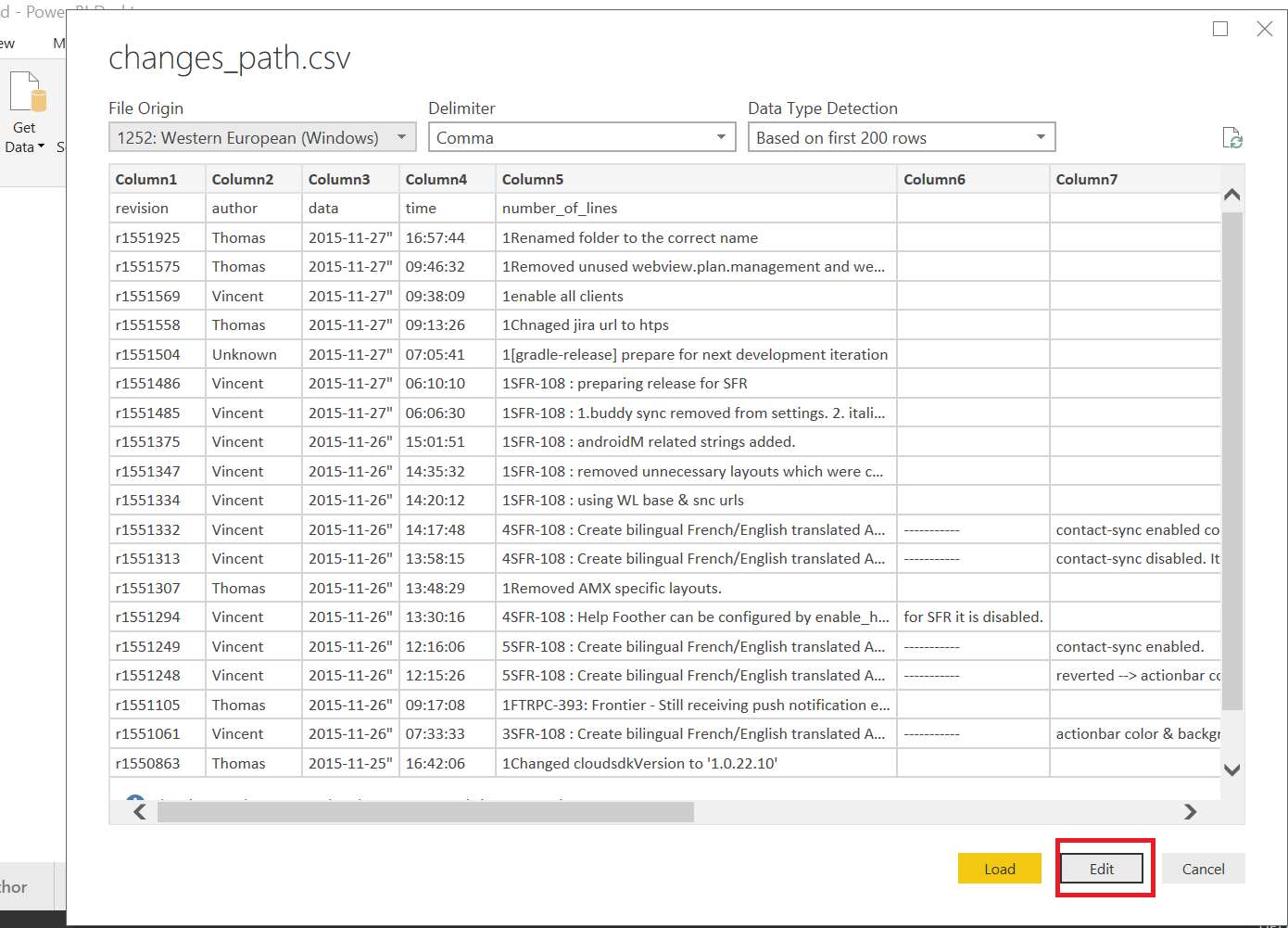
## Power BI

Prepping Power BI for analysis of the power\_changes.csv file

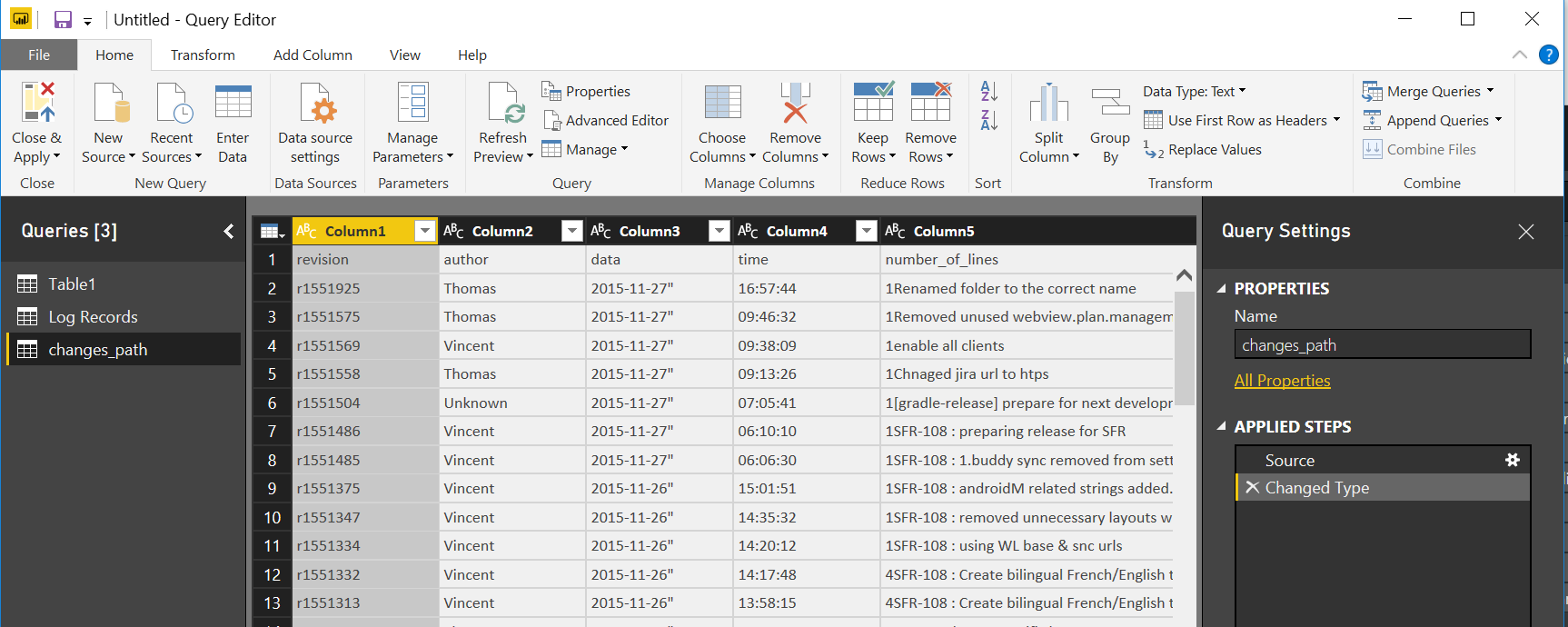
Open Power BI and import the data via the Home tab, Get Data and selecting text/CSV.



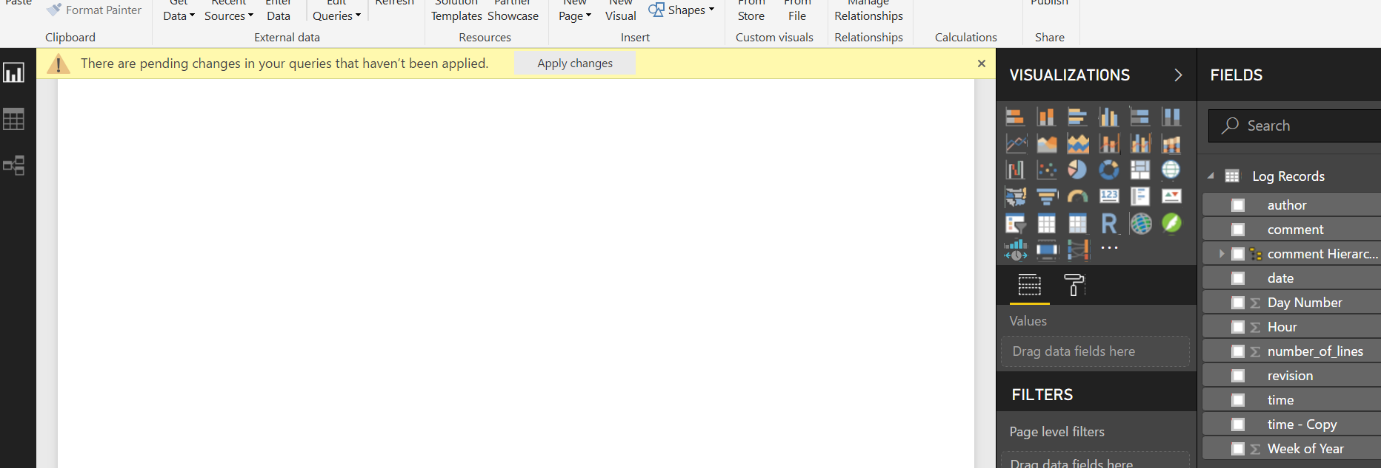
Within Power BI you can opt to carry out additional edits by selecting the Edit button



Which allows you to carry out such edits as splitting columns, duplicating columns and carrying out changes to these, e.g. changing date to week day number, etc. Once finished you click on ‘Close & Apply’ for the changes to load into Power BI.



After this the column headings will be listed on the right hand side for further analysis



## Excel

With excel I was able to create Pivot tables and charts as part of my analysis

# Findings

## Commits by Author

Below, from our various sources, we can see the number of commits by Author.

Even taking into account that the number of commits now assigned to ‘Unknown’ may actually have belonged to another author, Thomas is clearly the most prolific when it comes to the number of commits.

Murari Krishnan appears to be the author with the least number of commits, i.e. 1, followed very closely by Dave with 2.

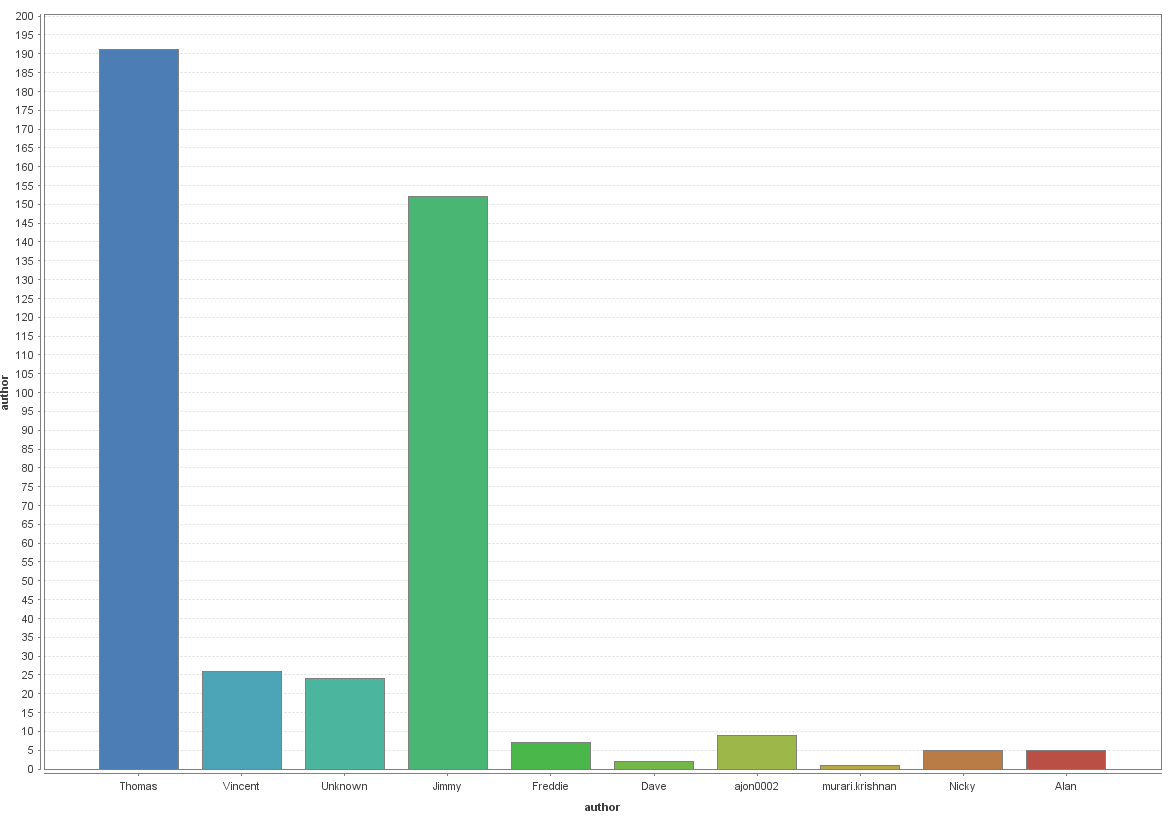
Interestingly, the charts below suggest that Tomas and Jimmy have spent significantly more time on github during the period than the other authors.

Taking ‘Unknown’ as being an author in their own right the average (mean) number of commits is 42.2, however we can see that this number is skewed significantly by Thomas and Jimmy’s numbers.

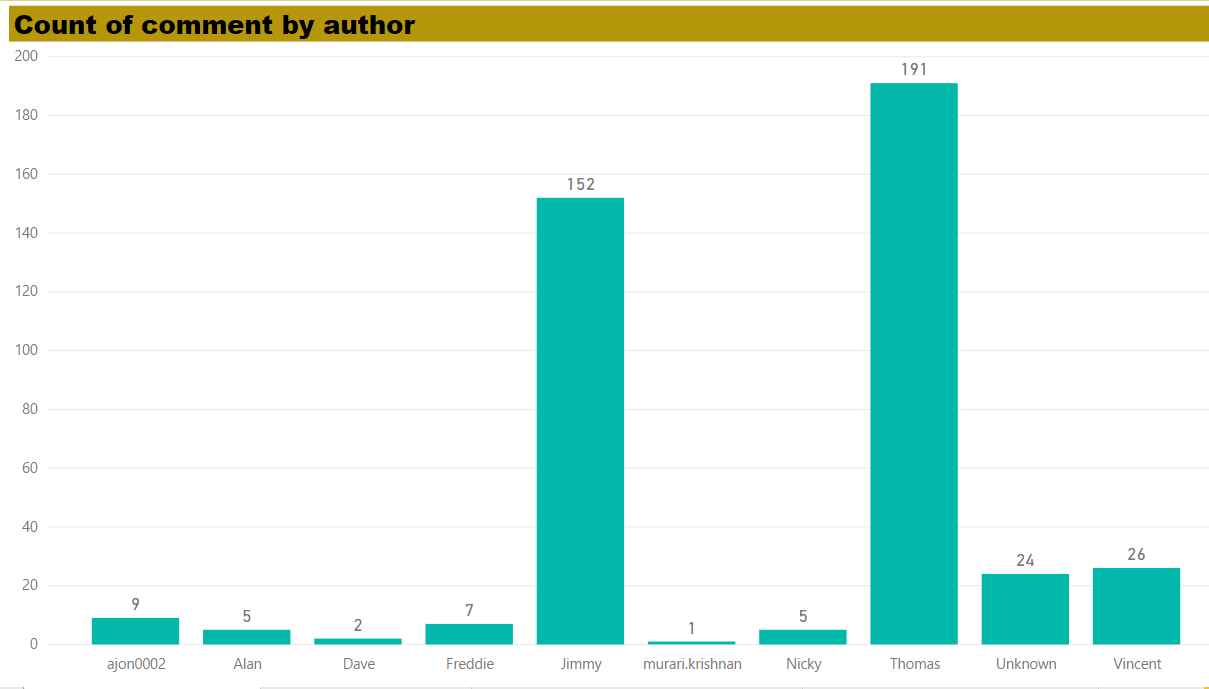
The median number of commits per author is 8. The mode number of commits per author is 5.

All these findings suggest that Thomas and Jimmy carried out an unrepresentatively large number of commits compared to rest of the authors.

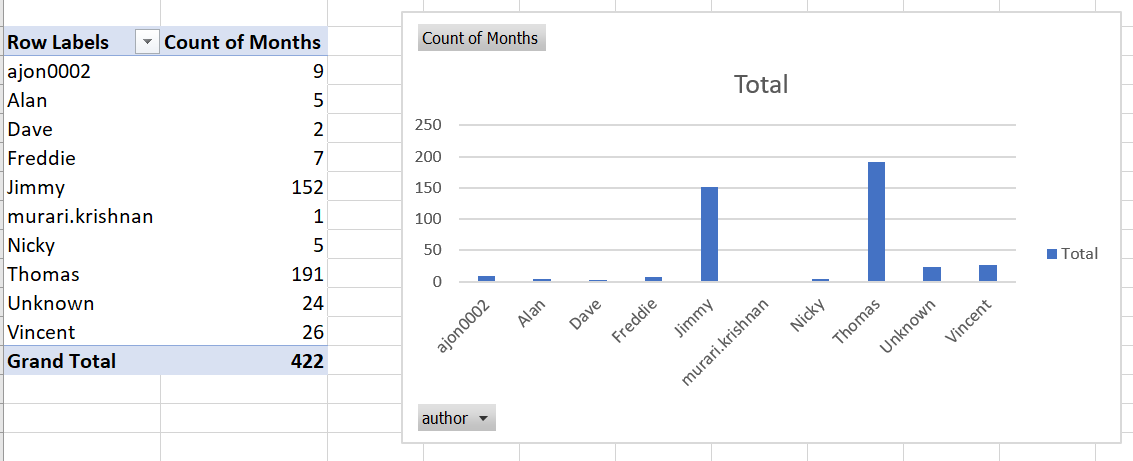
**Rapidminer Chart**



Power BI Chart



Excel



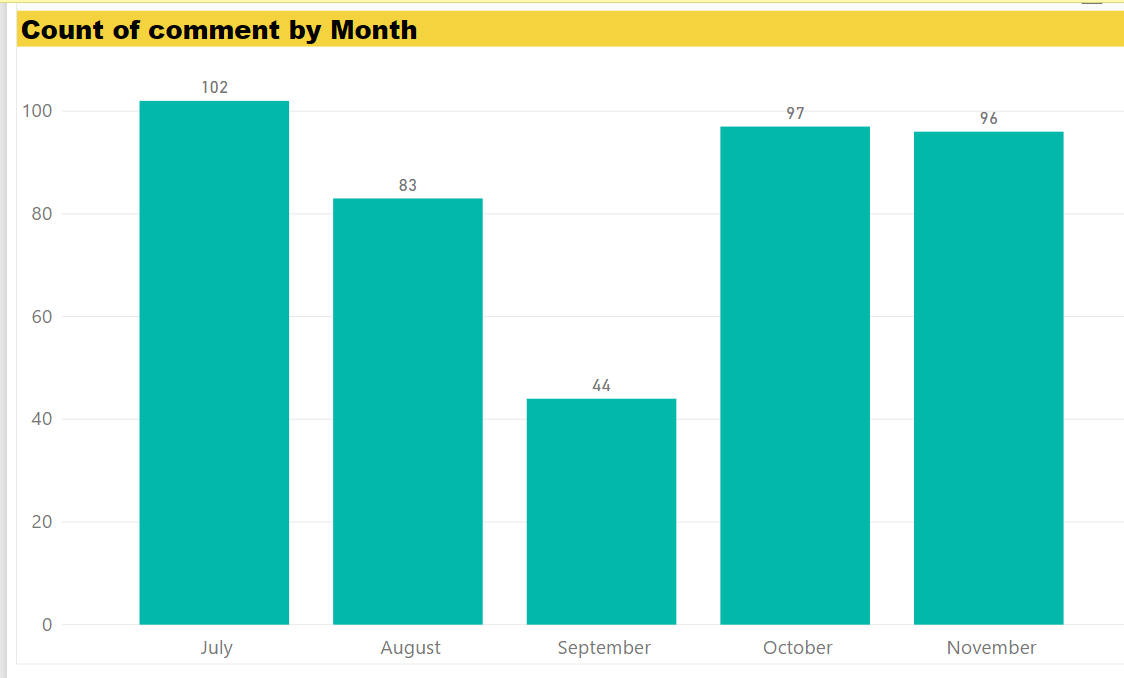
## Commits by Month

From the charts below, we can see that the busiest month, during the relevant period, in terms of the number of commits made by authors was July.

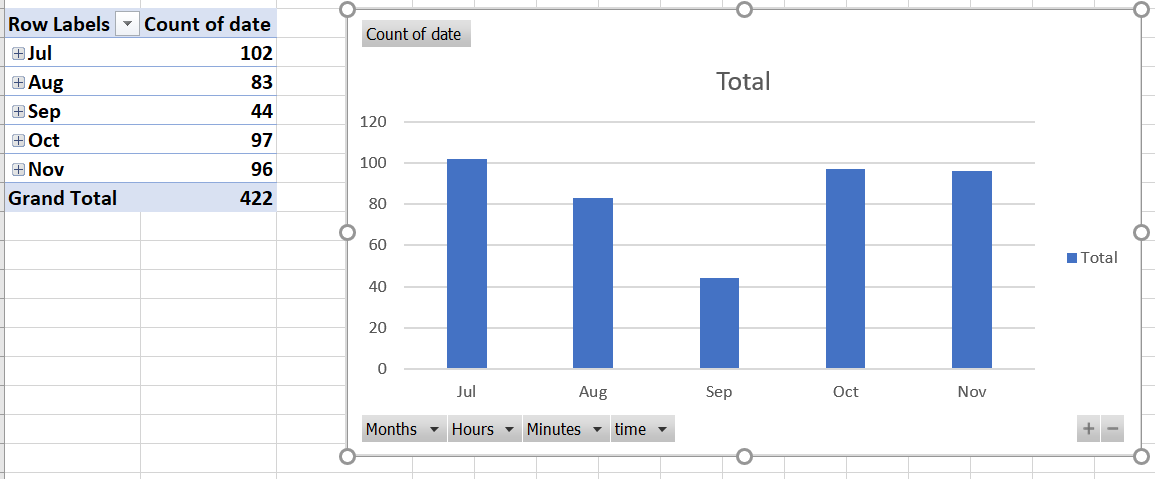
The quietest month was September. Assuming the authors in this case study are part time students from DBS this might be due to the fact that exams take place in September consequently there may not be as much practical work carried out in the lead up to and immediately after the exams.

The average (mean) number of commits a month is 84.4 and the median number of commits is 96.

**Power BI**



**Excel**



## Number of Days of the week Authors carried out a commit

From analysis carried out I can see that authors did not carry out commits during the weekend. Consequently I wanted to see whether authors tended to work all of the remaining days of the week.

From the Power BI chart below we can see that Jimmy, Thomas, Unknown and Vincent tended to submit commits on each day of the working week i.e. Monday to Friday while ajon0002, Dave and Murari.krishnan submitted commits on only one day of the working week. This may allude to different working styles and outside time commits of the various authors.

