|  |  |
| --- | --- |
| |  | | --- | |  | |
|  |
| Date: 7th May 2017 |
| CA4 |
| Programming for Big Data |
| Name: Paul Prew |
| Student Number: 10354828 |

# Contents

[Contents 0](#_Toc481999963)

[1. Background 1](#_Toc481999964)

[2. Methodology 2](#_Toc481999965)

[3.1 Analysis - By Authors 3](#_Toc481999966)

[3.2 Analysis – By Day 5](#_Toc481999967)

[3.3 Analysis – By File Change Action 6](#_Toc481999968)

# 1. Background

CA4 brief as follows:

Assignment is based on transforming a large dataset in text format.

Requirements are as follows:

* Scrub (clean) the data and place it into the relevant holder/container objects. There will be 422 different sets of commit objects.
* 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'.
* Code for calculating the analysis is to be documented and tested.
* Test should be in a separate file runnable from the command line.
* Statistical analytics conclusions should be in a word document explaining the information gleamed from the dataset.
* All program files and documentation to be uploaded to GitHub.

**URL location for GitHub repository is as follows:**

https://github.com/10354828/programming\_big\_data\_pp/tree/master/CA4

# 2. Methodology

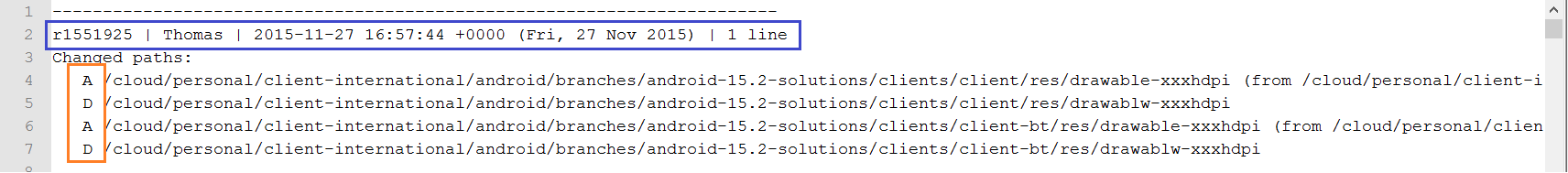
The data was scrubbed using a python program, to differentiate the data into 422 objects, which can be represented as rows on a spreadsheet.

Where possible additional data was derived from the data e.g. week number was derived from the date.

The data was mostly extracted from the title row. Additionally, for each row the number of files actioned for each revision number was calculated, and broken out by action type e.g. A, D, M, R.

This extract provides a count of the total number of file changes, broken out by action type. This also enables analysis on the number of individual files changed by author, and breakdown by action type.

Visualisation of analysis was done using charts in R Studio, and Microsoft Excel.



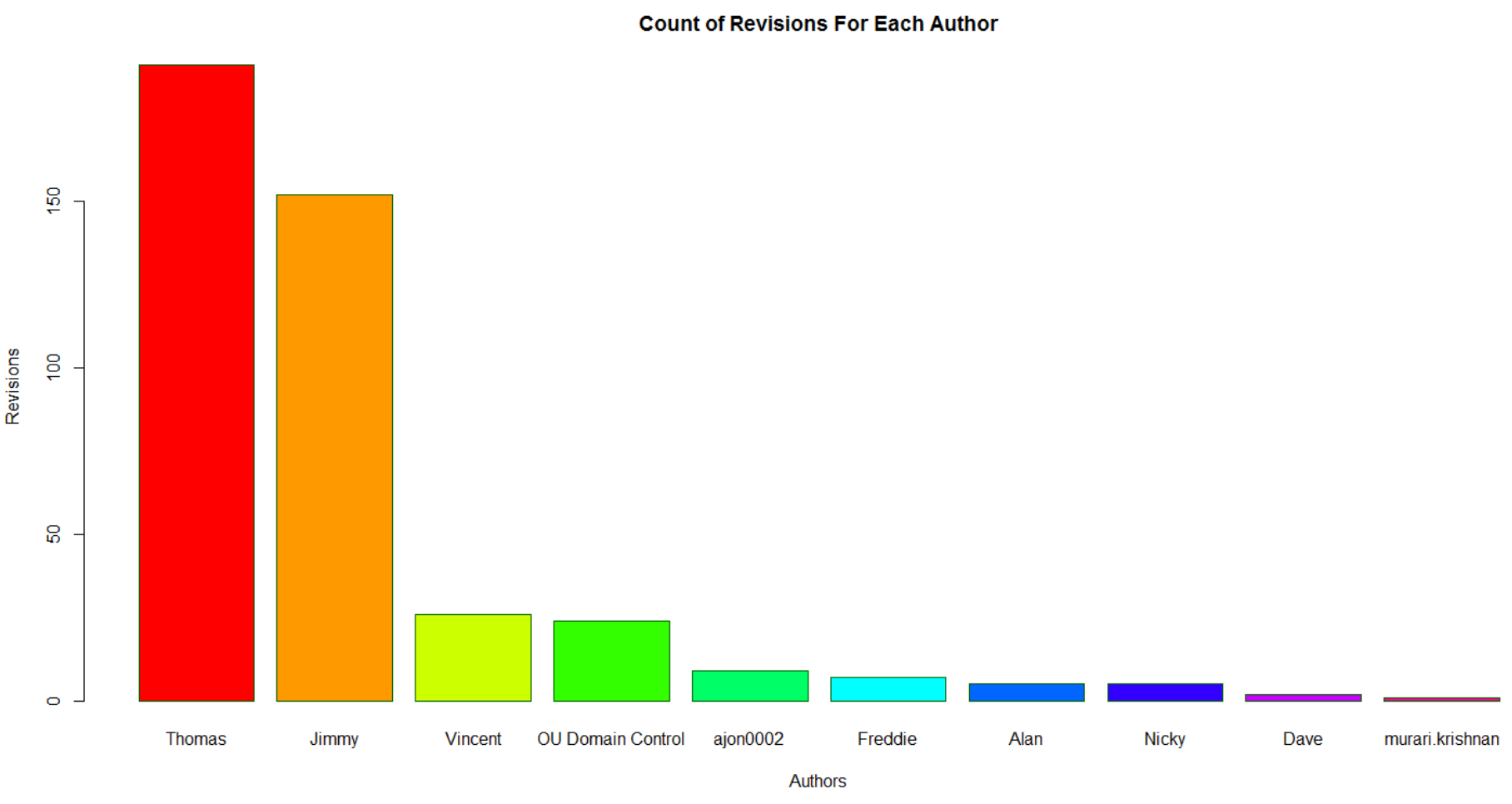
# 3.1 Analysis - By Authors

The table analysis below indicates that the number of revisions by author is not equally distributed across all authors.

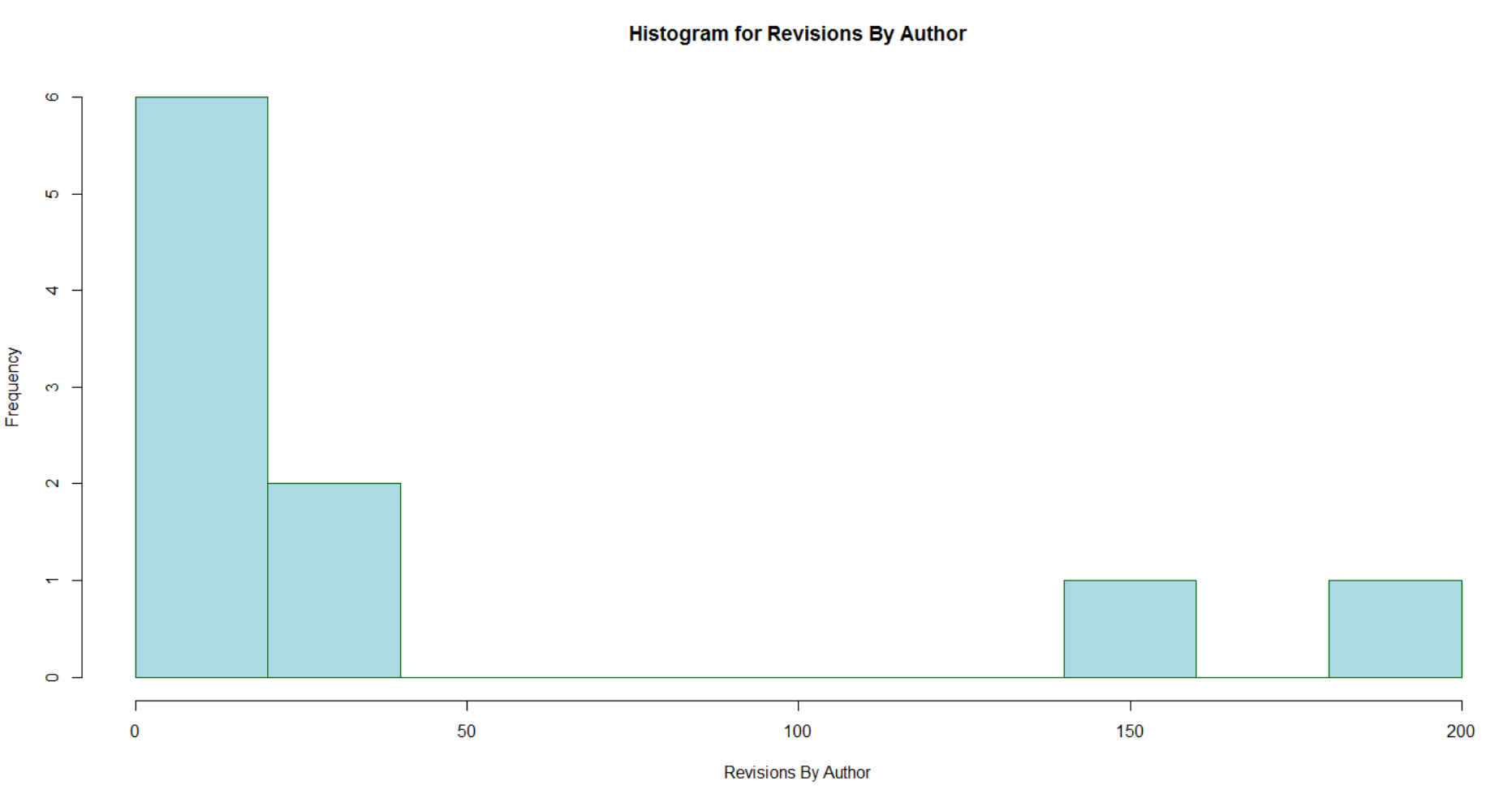
The maximum number is Thomas who has authored 191 revisions. This is a little ahead of Jimmy with 152, and significantly ahead of Vincent in third, who has 26 revisions to his name.

|  |  |
| --- | --- |
| **Author** | **Number Revisions** |
| Thomas | 191 |
| Jimmy | 152 |
| Vincent | 26 |
| /OU=Domain Control Validated/CN=svn.company.net | 24 |
| ajon0002 | 9 |
| Freddie | 7 |
| Alan | 5 |
| Nicky | 5 |
| Dave | 2 |
| murari.krishnan | 1 |

This significant variance is clearly seen in the bar chart below which shows that after Thomas, Jimmy, and Vincent the other authors are responsible for a disproportionately lower number of revisions.



The histogram below generated in R, reinforces this point, as it indicates that the clear majority of authors (8) are responsible for authoring less than 50 revisions. Thomas and Jimmy are the exception.



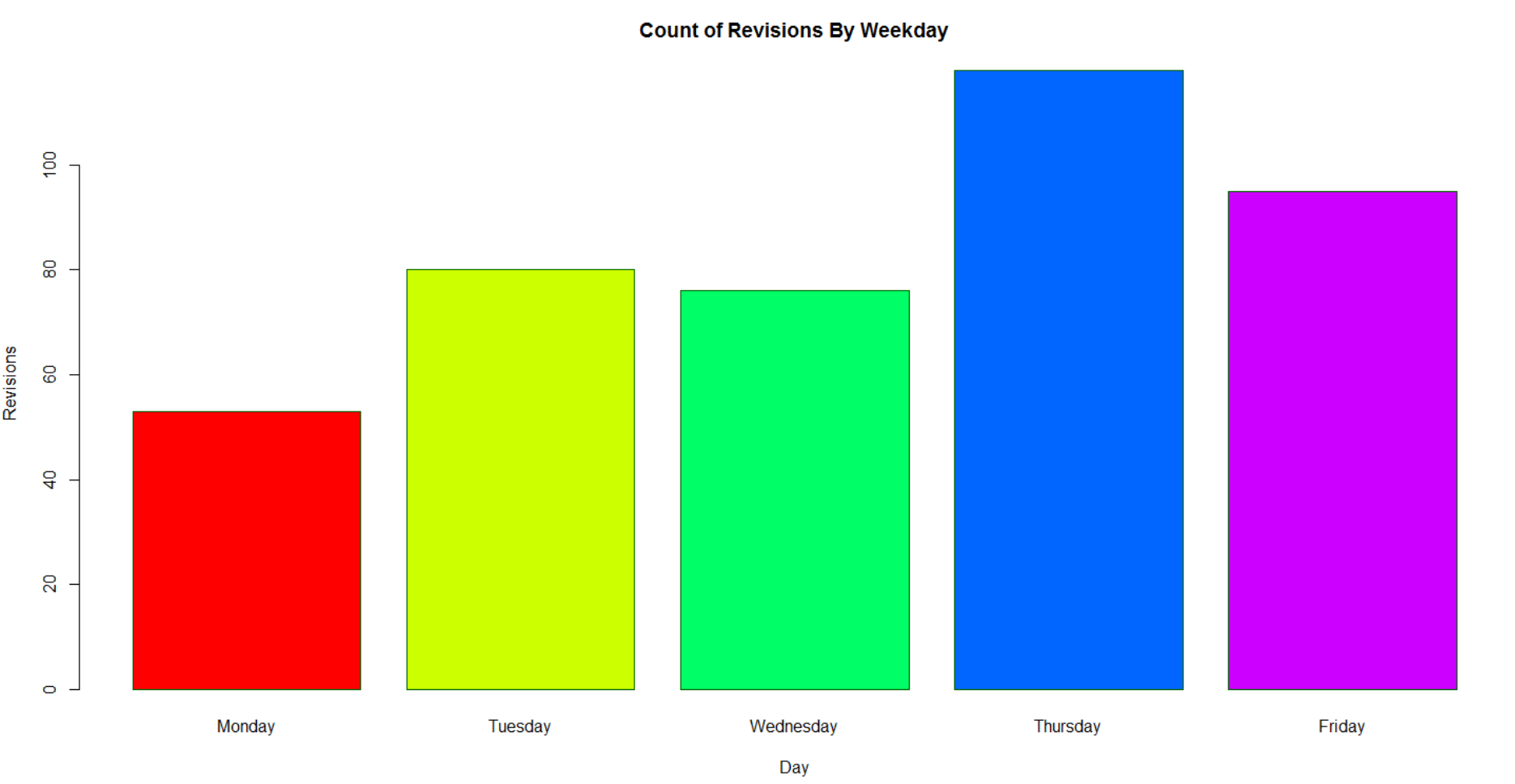
# 3.2 Analysis – By Day

The table analysis below indicates that for the time (Jul – Nov 2015 inclusive) the number of revisions is not equally distributed across all days in the week.

The day with the maximum number of revisions is Thursday with 118. The day with the minimum number of revisions is Monday with 53 revisions. The weekday average is 84.5 revisions.

|  |  |
| --- | --- |
| **Day** | **Number Revisions** |
| Mon | 53 |
| Tue | 80 |
| Wed | 76 |
| Thu | 118 |
| Fri | 95 |

The spread can be seen more clearly from the bar plot below, that was generated in R.



# 3.3 Analysis – By File Change Action

The table analysis below indicates that when all 422 revisions are considered, there was a total number 3011 files actioned as part of the revision process. These revisions are classified as follows:

Action Type - A

Action Type - D

Action Type - M

Action Type - R

The maximum number on any one revision is 297 files, and the average across all 422 revisions is 7.13 files per revision.

The breakout of revisions by author and by action type is seen in the table below. Thomas has made actions on the higher number of files (1359) followed by Jimmy (1158). This is probably to be expected seeing as we learned from 3.1 above, that Thomas and Jimmy are responsible for most revisions. This is also illustrated on the bar chart.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Author** | **Total Files** | **Action A** | **Action D** | **Action M** | **Action R** |
| Thomas | 1359 | 87 | 663 | 609 | 0 |
| Jimmy | 1158 | 690 | 66 | 401 | 1 |
| Vincent | 337 | 260 | 32 | 45 | 0 |
| Dave | 77 | 10 | 0 | 66 | 1 |
| Alan | 30 | 9 | 6 | 15 | 0 |
| /OU=Domain Control Validated/CN=svn.company.net | 24 | 0 | 0 | 24 | 0 |
| Freddie | 9 | 0 | 0 | 9 | 0 |
| ajon0002 | 9 | 0 | 0 | 9 | 0 |
| Nicky | 7 | 0 | 0 | 7 | 0 |
| murari.krishnan | 1 | 0 | 0 | 1 | 0 |
|  | **3011** | **1056** | **767** | **1186** | **2** |

The pie chart below signifies that the significant number of changes are ‘Action Type – M’ with 39%, followed by ‘Action Type A’ with 35%, and ‘Action Type D’ with 26%.

It might be expected that the proportion of change types would be seen equally across each author. However, the bar chart below which looks at the top 3 authors shows that the proportion of changes is not the same. Thomas has a greater number of ‘Action D’ file changes (663), compared to Jimmy (66). However, Jimmy has a greater number of ‘Action A’ file changes (690), compared to Thomas (87).

The different types of file changes indicated by Type A, Type D, Type M, may suggest a different level of effort which will needed for each type. This may also need to be considered as well as number of revisions completed, as a measure of author performance.