CA4 – Dataset Analysis

To generate the csv file I had to add some extra columns, as I couldn’t manage to clean the data to fit 6 columns only. Adding extra columns I have managed to get all the 422 rows when reading the csv.

My 3 observations on the csv file are as follows:

import pandas as pd

import csv

**1) The majority of the changes (over 80%) are affecting 1 line only**

df = pd.read\_csv("/Users/FabioDiGiovine/Desktop/Darren/CA4/changes.csv", error\_bad\_lines=False)

print(df.describe())

**2) Authors Thomas and Jimmy have high volumes of changes compared to the other users**

groupByAuthors = df.groupby(['author']).sum()

print('The sum of total line changed, grouped by the Author is: ' + str(groupByAuthors))

**3) November is the month with higher number of changes**

df['date'] = pd.to\_datetime(df['date'])

df.groupby(df['date'].dt.strftime('%B'))['number\_of\_lines'].sum().sort\_values()