Continuous Assessment 4 – Programing for Big Data

Ciaran O'Driscoll - 10357873

21/11/2017

Data has been placed in a dataframe ‘df’ and exported to a file ‘dataframe2.csv’

Using the data in this file I have tried to find some interesting details in the data.

pivot1 = pd.pivot\_table(df,index='user',values=['added','deleted','modified'],aggfunc=np.sum,margins=True)

pivot1

Out[99]:

added deleted modified

user

/OU=Domain Control Validated/CN=svn.company.net 0.0 0.0 24.0

Alan 9.0 6.0 15.0

Dave 10.0 0.0 66.0

Freddie 0.0 0.0 9.0

Jimmy 690.0 66.0 401.0

Nicky 0.0 0.0 7.0

Thomas 87.0 663.0 609.0

Vincent 260.0 32.0 45.0

ajon0002 0.0 0.0 9.0

murari.krishnan 0.0 0.0 1.0

All 1056.0 767.0 1186.0

Doing a pivot on the data we can see the total number of ‘added’, ‘deleted’ and ‘modified’ items were 1056, 767 and 1186 respectively and we can see the person that made the most ‘adds’ was Jimmy with 690, the person who made the most ‘deletes’ was Thomas with 663 and the person who made the most ‘modifications’ was also Thomas with 609.

We can group the timestamps by month and see the total values per month of each type of commit object :

df['timestamp']=pd.to\_datetime(df['timestamp'], format ='%Y-%m-%d %H:%M:%S’).dt.to\_period('M')

pivot1 = pd.pivot\_table(df,index='timestamp',values=['added','deleted','modified'],aggfunc=np.sum)

pivot1

Out[52]:

added deleted modified

timestamp

2015-07 53 463 371

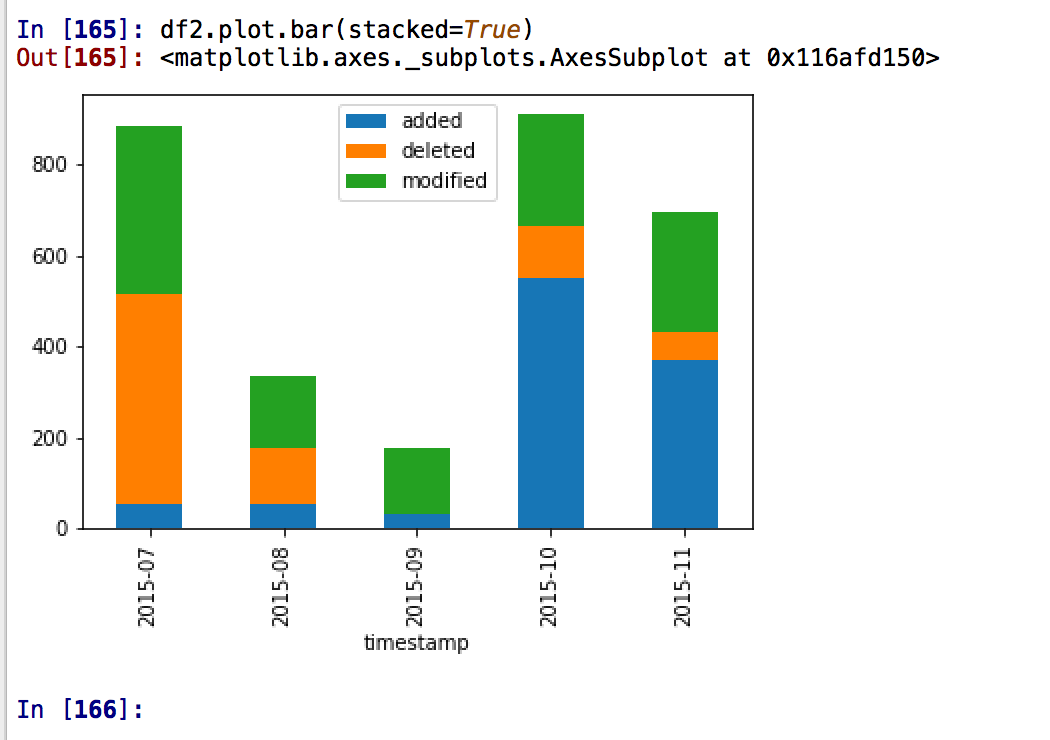
2015-08 53 125 158

2015-09 31 2 144

2015-10 550 116 246

2015-11 369 61 267

We can see that October was the busiest month for ‘additions’, July was the busiest month for ‘deleteions’ and July was also teh busiest month for ‘modifications.

We can view this as a stacked bar chart :

I then decided to creat pivots on each of ‘added’, ‘modified’ and ‘deleted’ per user showing the percentages by user.

pd.pivot\_table(df,index='user',values=['added'],aggfunc=np.sum,margins=True).div(sum(df.added)).mul(100)

Out[101]:

added

user

/OU=Domain Control Validated/CN=svn.company.net 0.000000

Alan 0.852273

Dave 0.946970

Freddie 0.000000

Jimmy 65.340909

Nicky 0.000000

Thomas 8.238636

Vincent 24.621212

ajon0002 0.000000

murari.krishnan 0.000000

All 100.000000

This shows 65.3% of the total adds were done by Jimmy

pd.pivot\_table(df,index='user',values=['modified'],aggfunc=np.sum,margins=True).div(sum(df.modified)).mul(100)

Out[105]:

modified

user

/OU=Domain Control Validated/CN=svn.company.net 2.023609

Alan 1.264755

Dave 5.564924

Freddie 0.758853

Jimmy 33.811130

Nicky 0.590219

Thomas 51.349073

Vincent 3.794266

ajon0002 0.758853

murari.krishnan 0.084317

All 100.000000

This shows 33.8% of the total ‘modifications’ were done by Jimmy

pd.pivot\_table(df,index='user',values=['deleted'],aggfunc=np.sum,margins=True).div(sum(df.deleted)).mul(100)

Out[107]:

deleted

user

/OU=Domain Control Validated/CN=svn.company.net 0.000000

Alan 0.782269

Dave 0.000000

Freddie 0.000000

Jimmy 8.604954

Nicky 0.000000

Thomas 86.440678

Vincent 4.172099

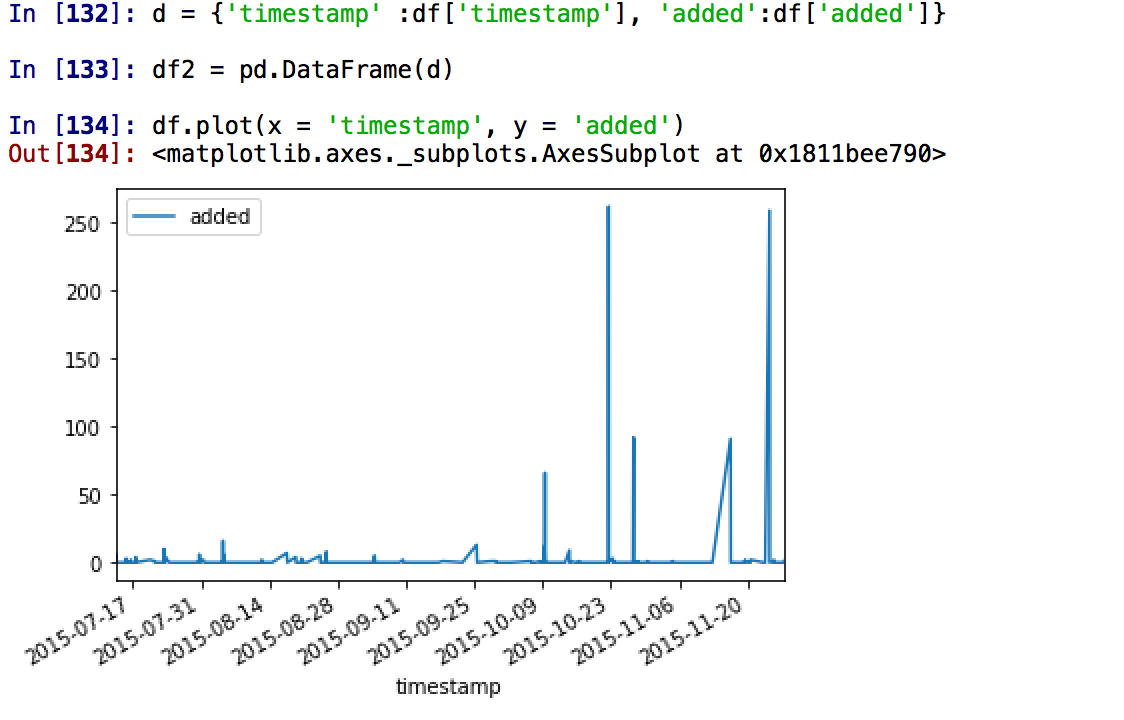
ajon0002 0.000000

murari.krishnan 0.000000

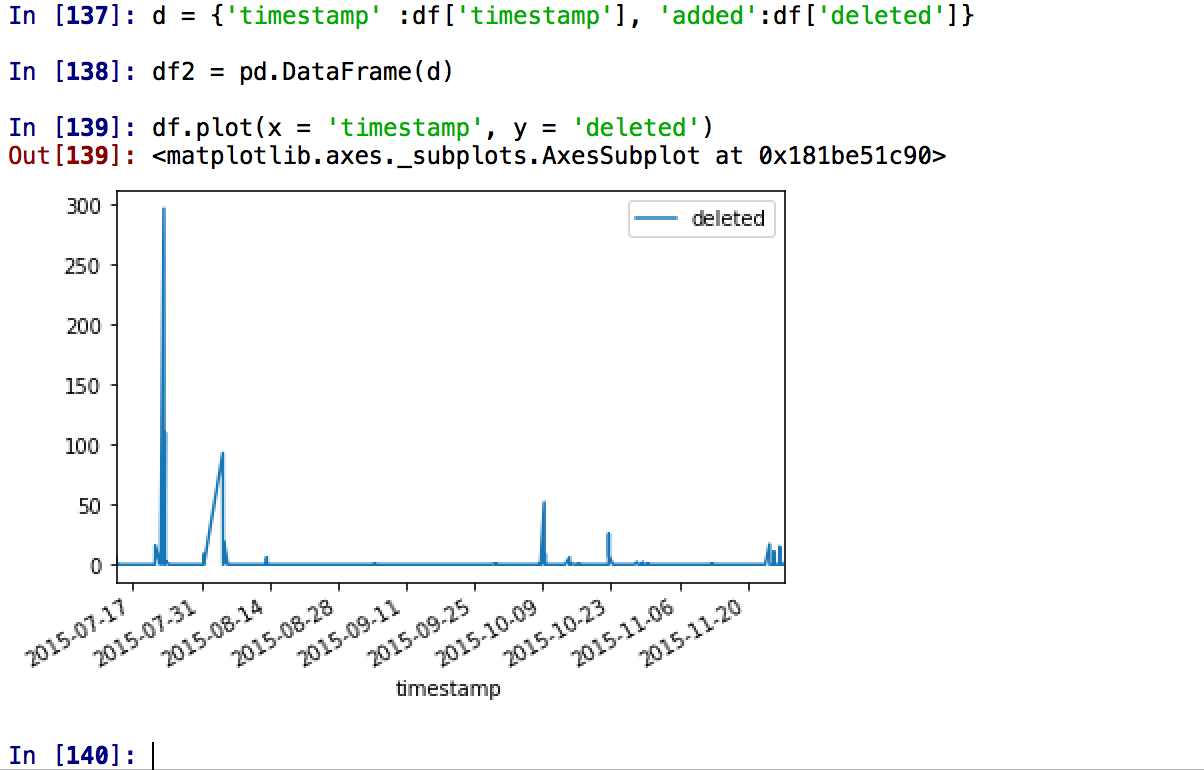
All 100.000000

This shows 86.4% of the ‘deletions’ were done by Thomas

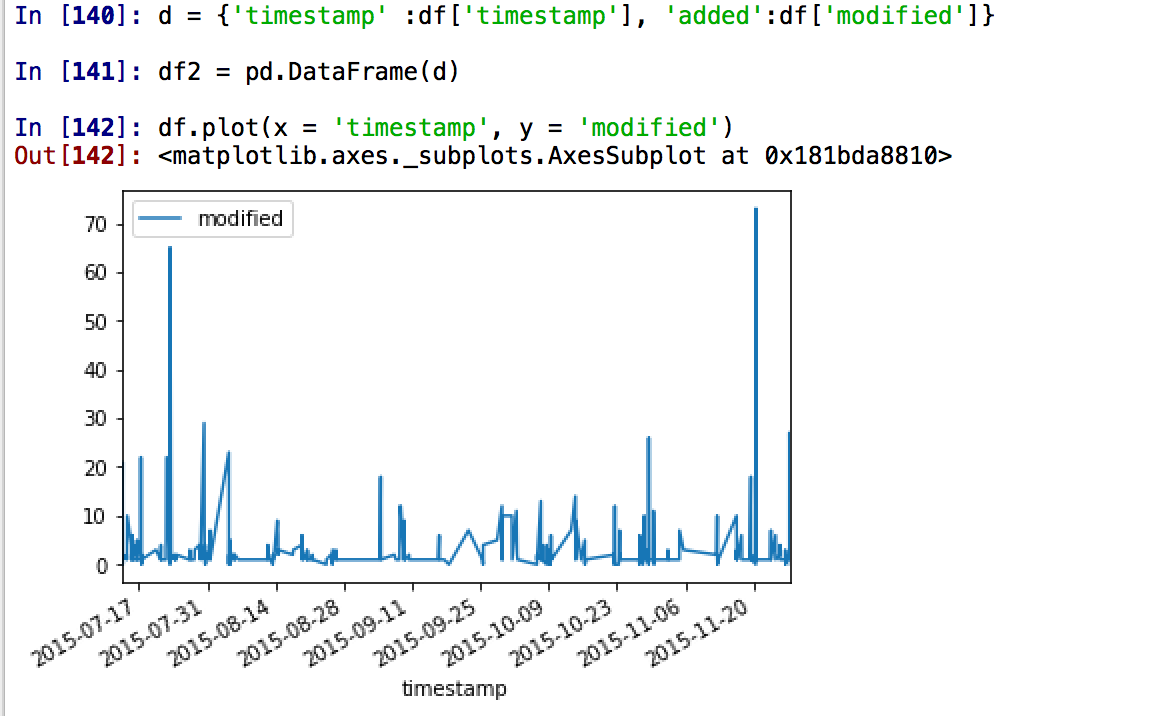
I then decided to see if I could see the variation in the activities with time.



We can see most of the ‘additions’ were done later in the year



Most of the deletions appear to have been done ealier in the year.



The modifications appear to be more evenly spread over the timeframe, with peaks at either end.

21/11/2017