check_data

August 11, 2022

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[2]:
```
df = pd.read_csv('number_calibrate.csv')
df2 = pd.read_csv('clean_calibrate.csv')
```

[3]:
```
df2 = df2.set_index('date')
```

[4]:
```
df = df.set_index('date')
```

[5]:
```
df
```

[5]:

date	Neil deGrasse Tyson	Cristiano Ronaldo	Earl K. Sneed	\
2017-10-01	1.352941	14.613730	0.0	
2017-10-08	1.058824	14.613730	0.0	
2017-10-15	1.058824	12.787014	0.0	
2017-10-22	1.117647	20.093879	0.0	
2017-10-29	1.294118	14.613730	0.0	
…	…	…	…	
2019-09-29	0.588235	10.960297	0.0	
2019-10-06	1.294118	10.960297	0.0	
2019-10-13	0.882353	12.787014	0.0	
2019-10-20	0.764706	10.960297	0.0	
2019-10-27	0.647059	10.960297	0.0	

date	Donald Trump	George H.W. Bush	Justin Caldbeck	Borris Miles	\
2017-10-01	49.321338	0.027027	0.00448	0.00000	
2017-10-08	56.628203	0.027027	0.00784	0.02800	
2017-10-15	43.841190	0.027027	0.01904	0.00448	
2017-10-22	42.014473	0.189189	0.02800	0.00336	
2017-10-29	47.494622	0.054054	0.00336	0.00000	
…	…	…	…	…	
2019-09-29	40.187757	0.027027	0.00560	0.00000	
2019-10-06	40.187757	0.000000	0.00000	0.00000	
2019-10-13	34.707608	0.027027	0.00000	0.00112	

```
2019-10-20     31.054176          0.027027          0.00224       0.00112
2019-10-27     38.361041          0.000000          0.00000       0.00000

            Dan Johnson  Ed Murray  Bobby Scott  …  Cody Wilson  \
date                                              …
2017-10-01     0.130435    0.03696         0.05  …     0.053763
2017-10-08     0.130435    0.04480         0.05  …     0.043011
2017-10-15     0.086957    0.02464         0.07  …     0.032258
2017-10-22     0.086957    0.03024         0.05  …     0.032258
2017-10-29     0.086957    0.02912         0.06  …     0.032258
…                 …          …          …   …          …
2019-09-29     0.130435    0.01456         0.05  …     0.043011
2019-10-06     0.130435    0.03024         0.06  …     0.021505
2019-10-13     0.130435    0.01456         0.05  …     0.032258
2019-10-20     0.130435    0.02128         0.05  …     0.032258
2019-10-27     0.086957    0.01680         0.05  …     0.032258

            Adam Berkowitz  Charles Schwertner  Albert J. Alvarez  \
date
2017-10-01            0.00                 0.0           0.000000
2017-10-08            0.00                 0.0           0.000000
2017-10-15            0.00                 0.0           0.000000
2017-10-22            0.00                 0.0           0.000000
2017-10-29            0.00                 0.0           0.000000
…                      …                   …                 …
2019-09-29            0.00                 0.0           0.000000
2019-10-06            0.00                 0.0           0.000000
2019-10-13            0.01                 0.0           0.001243
2019-10-20            0.00                 0.0           0.000000
2019-10-27            0.00                 0.0           0.000000

            Amit Singhal  Robert Scoble  Richard DeVaul  Roy Price  \
date
2017-10-01          0.01           0.01             0.0       0.09
2017-10-08          0.01           0.01             0.0       0.77
2017-10-15          0.01           0.18             0.0       0.70
2017-10-22          0.01           0.29             0.0       0.20
2017-10-29          0.00           0.05             0.0       0.16
…                    …              …                …         …
2019-09-29          0.01           0.01             0.0       0.06
2019-10-06          0.01           0.00             0.0       0.85
2019-10-13          0.01           0.00             0.0       0.12
2019-10-20          0.01           0.00             0.0       0.08
2019-10-27          0.01           0.00             0.0       0.08

            Lockhart Steele  Harvey Weinstein
date
```

```
2017-10-01             0.00         16.440446
2017-10-08             0.00        147.964015
2017-10-15             0.12         67.588501
2017-10-22             0.06         20.093879
2017-10-29             0.04         14.613730
…                       …              …
2019-09-29             0.00          1.826716
2019-10-06             0.00          1.826716
2019-10-13             0.00          1.826716
2019-10-20             0.00          3.653432
2019-10-27             0.00          1.826716

[109 rows x 331 columns]
```

[6]:
```
nameList = []
for name in df.columns:
    #print(name)
    #print(df[name].max(), df[name].min())
    if df[name].max() - df[name].min() >= 20:
        nameList.append(name)
```

[7]:
```
nameList2 = []
for name in df2.columns:
    #print(name)
    #print(df[name].max(), df[name].min())
    if df2[name].max() - df2[name].min() >= 20:
        nameList2.append(name)
```

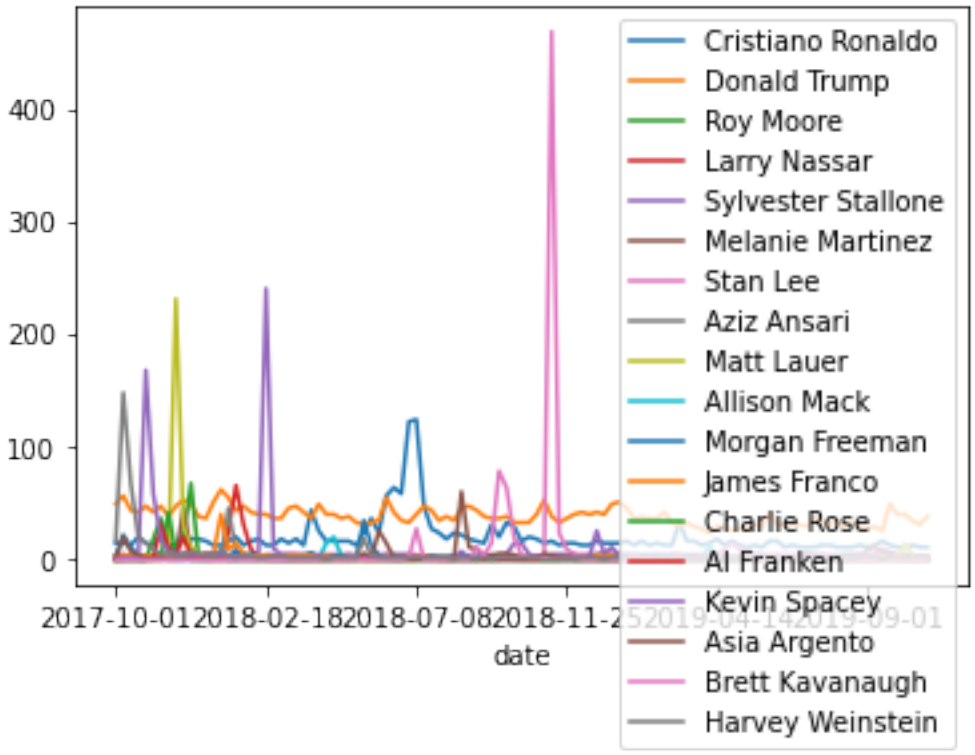
[8]:
```
df2[nameList2]
```

[8]:
```
            Al Franken  Cristiano Ronaldo  Aziz Ansari  Kevin Spacey  \
date
2017-10-01         0.0          14.613730          0.0      1.826716
2017-10-08         0.0          14.613730          0.0      1.826716
2017-10-15         0.0          12.787014          0.0      1.826716
2017-10-22         0.0          20.093879          0.0      1.826716
2017-10-29         0.0          14.613730          0.0    168.057894
…                  …                 …           …           …
2019-09-29         0.0          10.960297          0.0      1.826716
2019-10-06         0.0          10.960297          0.0      1.826716
2019-10-13         0.0          12.787014          0.0      1.826716
2019-10-20         0.0          10.960297          0.0      1.826716
2019-10-27         0.0          10.960297          0.0      1.826716

            Stan Lee  Sylvester Stallone  Charlie Rose  Morgan Freeman  \
date
2017-10-01  0.000000            2.403574           0.0        3.653432
```

```
2017-10-08  0.000000            4.807148          0.0       21.920595
2017-10-15  0.000000            2.403574          0.0       10.960297
2017-10-22  0.000000            4.807148          0.0        3.653432
2017-10-29  0.000000            2.403574          0.0        3.653432
…                 …                   …           …             …
2019-09-29  4.683888            4.807148          0.0        1.826716
2019-10-06  0.000000            4.807148          0.0        1.826716
2019-10-13  0.000000            2.403574          0.0        1.826716
2019-10-20  0.000000            2.403574          0.0        1.826716
2019-10-27  0.000000            2.403574          0.0        1.826716

            Allison Mack  James Franco  Harvey Weinstein  Roy Moore  \
date
2017-10-01           0.0      3.653432         16.440446        0.0
2017-10-08           0.0      3.653432        147.964015        0.0
2017-10-15           0.0      3.653432         67.588501        0.0
2017-10-22           0.0      3.653432         20.093879        0.0
2017-10-29           0.0      3.653432         14.613730        0.0
…                    …             …                 …         …
2019-09-29           0.0      3.653432          1.826716        0.0
2019-10-06           0.0      1.826716          1.826716        0.0
2019-10-13           0.0      1.826716          1.826716        0.0
2019-10-20           0.0      1.826716          3.653432        0.0
2019-10-27           0.0      1.826716          1.826716        0.0

            Matt Lauer
date
2017-10-01    0.000000
2017-10-08    0.000000
2017-10-15    0.000000
2017-10-22    0.000000
2017-10-29    0.000000
…                  …
2019-09-29    0.000000
2019-10-06   13.873794
2019-10-13    4.624598
2019-10-20    0.000000
2019-10-27    0.000000

[109 rows x 13 columns]
```

[9]:
```
df[nameList].plot()
```

[9]:
```
<AxesSubplot:xlabel='date'>
```

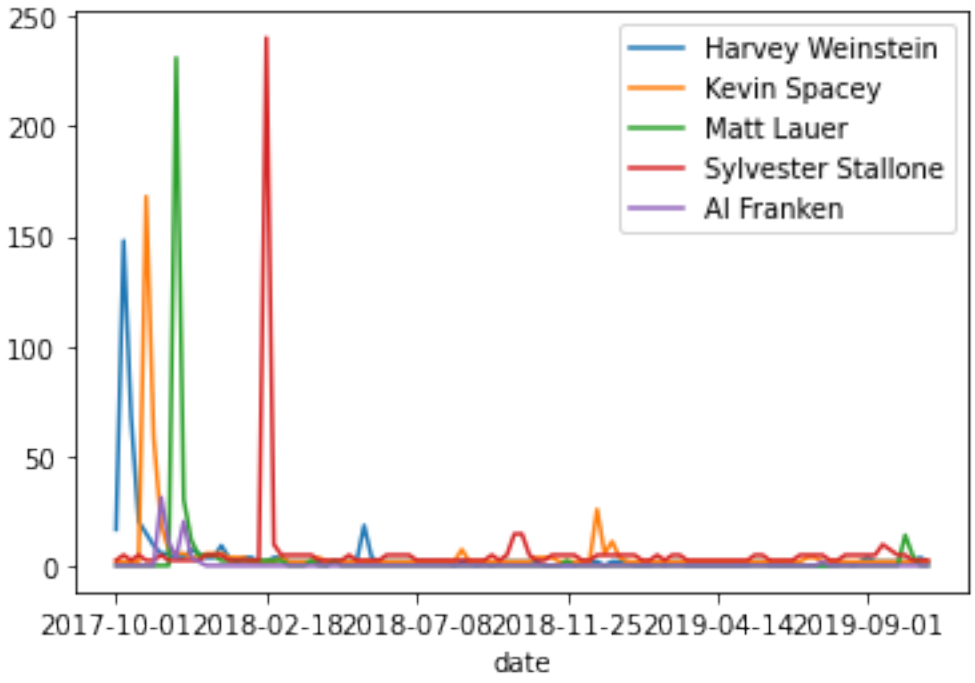


[41]:
```
period = (df.index > '2017-10-01') & (df.index < '2018-03-01')
period
```

[41]:
```
array([False,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False])
```

[10]:
```
df[['Harvey Weinstein', 'Kevin Spacey', 'Matt Lauer', 'Sylvester Stallone', 'Al␣
↪Franken']].plot()
```

[10]:
```
<AxesSubplot:xlabel='date'>
```



[37]:
```
clean_origin[clean_origin['name'].isin(['Harvey Weinstein', 'Kevin Spacey',␣
↪'Matt Lauer', 'Sylvester Stallone', 'Al Franken'])]
```

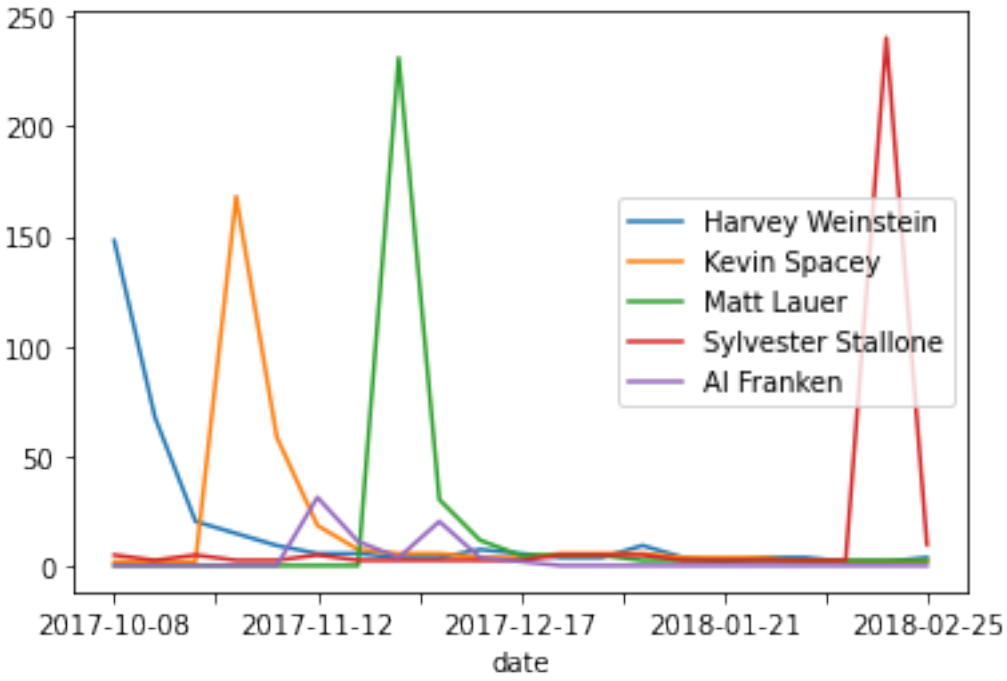
[37]:

	id	name	time	gender	is_work_influenced	confirmation_status	work_type	data_source	description	fb_followers
113	260	Sylvester Stallone	11/16/2017	1	0	0.0	Arts & Entertainment	Frankie Shaw	A woman has reported that he and another man s…	10,591,115
123	44	Matt Lauer	11/27/2017	1	1	1.0	Media	NY Times	Fired from NBC after being accused of inapprop…	FALSE
277	36	Al Franken	11/16/2017	1	1	1.0	Politics	NY Times	Resigned after accusations of groping and impr…	771,189
310	10	Kevin Spacey	10/29/2017	1	1	1.0	Arts & Entertainment	NY Times	Dropped from his Netflix show, "House of Cards…	3,718,803
333	0	Harvey Weinstein	10/5/2017	1	1	0.0	Arts & Entertainment	NY Times	Accused by dozens of women of sexual misconduc…	FALSE

```
                                    fb_pagelink  is_certified  \
113  https://www.facebook.com/SylvesterStallone/             1
123                                        FALSE             0
277      https://www.facebook.com/TeamAlFranken/             1
310        https://www.facebook.com/KevinSpacey/             1
333                                        FALSE             1

          twitter_name                     twitter_pagelink     tweets  \
113  Sylvester Stallone   https://twitter.com/TheSlyStallone  2,353
123          Matt Lauer     https://twitter.com/mattlauer012      0
277          Al Franken        https://twitter.com/alfranken  3,757
310        Kevin Spacey      https://twitter.com/KevinSpacey  2,089
333     Harvey Weinstein  https://twitter.com/HarveyWeinstein      0

      twitter_location twitter_since  \
113        Los Angeles      2010 8
123      New York, USA     2019 11
277          Minnesota      2007 7
310  London - NYC - LA      2009 3
333                NaN      2012 6

                                        twitter_intro twitter_follows  \
113  Actor, writer, director, artist. Instagram: of…              33
123                                             News              13
277                Al Franken, Democrat from Minnesota            4.2
310  Former shoe salesman now making a go at film a…              31
333                                                0               0

    twitter_followers
113             293.7
123                 5
277              96.5
310             420.5
333               1.4
```

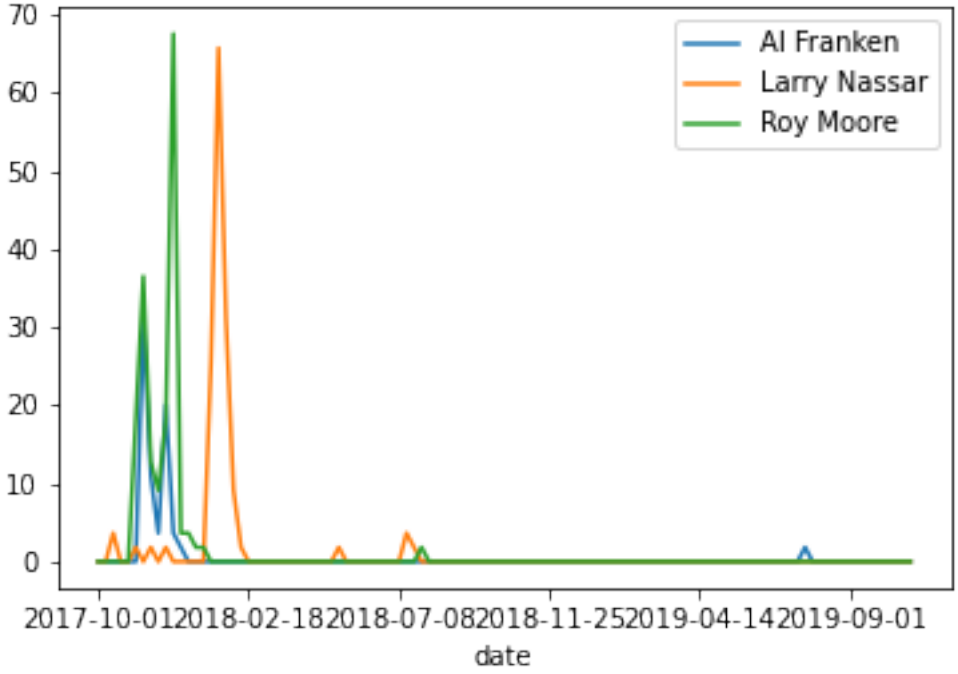
[42]:
```
df[['Harvey Weinstein', 'Kevin Spacey', 'Matt Lauer', 'Sylvester Stallone', 'Al␣
↪Franken']].loc[period].plot()
```

[42]:
```
<AxesSubplot:xlabel='date'>
```



[11]:
```
df[['Al Franken', 'Larry Nassar', 'Roy Moore']].plot()
```

[11]: <AxesSubplot:xlabel='date'>



[28]:
```
clean_origin = pd.read_csv('GoogleTrends/MeToo_20210302.csv')
clean_origin[clean_origin['name'].isin(['Al Franken', 'Larry Nassar', 'Roy␣
↪Moore'])]
#df[['Al Franken', 'Larry Nassar', 'Roy Moore']].plot()
```

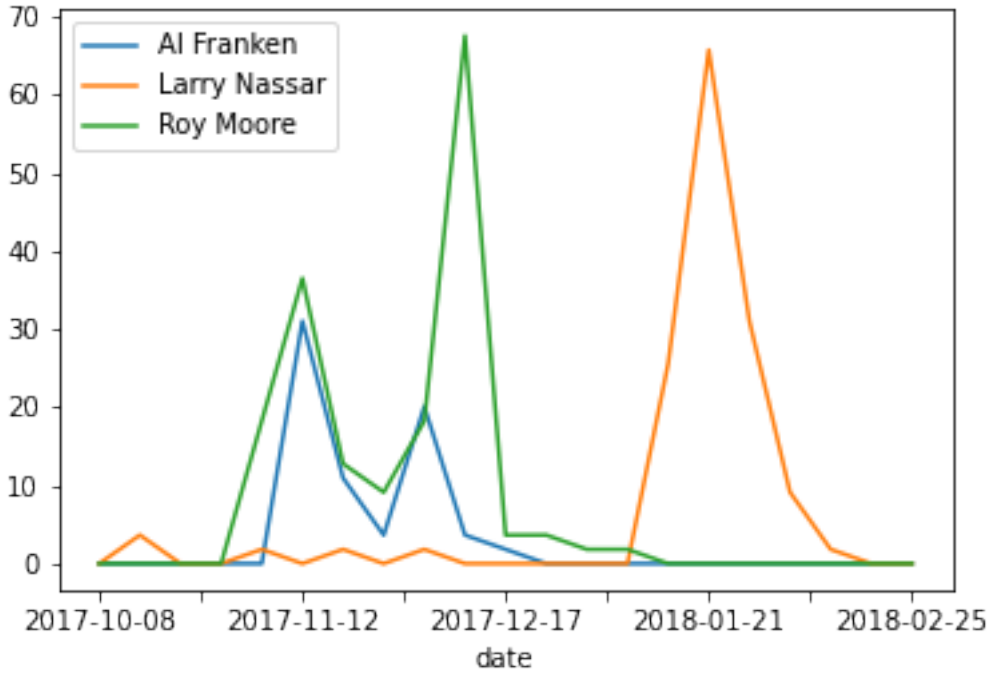
[28]:

	id	name	time	gender	is_work_influenced	confirmation_status	work_type	data_source	description	fb_followers	fb_pagelink	is_certified	twitter_name	twitter_pagelink	tweets	twitter_location	twitter_since	twitter_intro	twitter_follows	twitter_followers
81	420	Roy Moore	11/9/2017	1	0	0.0	Politics	Frankie Shaw	Multiple women have said he sexually abused or…	FALSE	FALSE	0	Roy Moore	https://twitter.com/roymoore03	4	NaN	2013 7	mothafuckaaa	48	32
93	455	Larry Nassar	11/10/2017	1	1	1.0	Other	Frankie Shaw	More than 100 women and girls have reported th…	FALSE	FALSE	0	Abigail Pesta	https://twitter.com/AbigailPesta	6,820	NaN	2012 1	"THE GIRLS is probably the most thorough accou…	1,120	2,844
277	36	Al Franken	11/16/2017	1	1	1.0	Politics	NY Times	Resigned after accusations of groping and impr…	771,189	https://www.facebook.com/TeamAlFranken/	1	Al Franken	https://twitter.com/alfranken	3,757	Minnesota	2007 7	Al Franken, Democrat from Minnesota	4.2	96.5

[43]:
```
df[['Al Franken', 'Larry Nassar', 'Roy Moore']].loc[period].plot()
"""
Nassar was sentenced to 60 years in federal prison on December 7, 2017,
after pleading guilty to child pornography and tampering with evidence charges␣
↪on July 11, 2017.
```

```
On January 24, 2018,
Nassar was sentenced to an additional 40 to 175 years in Michigan State prison␣
 ↪after pleading guilty in Ingham County to seven counts of sexual assault of␣
 ↪minors
"""
```

[43]:
```
'\nNassar was sentenced to 60 years in federal prison on December 7,
2017,\nafter pleading guilty to child pornography and tampering with evidence
charges on July 11, 2017. \nOn January 24, 2018, \nNassar was sentenced to an
additional 40 to 175 years in Michigan State prison after pleading guilty in
Ingham County to seven counts of sexual assault of minors\n'
```

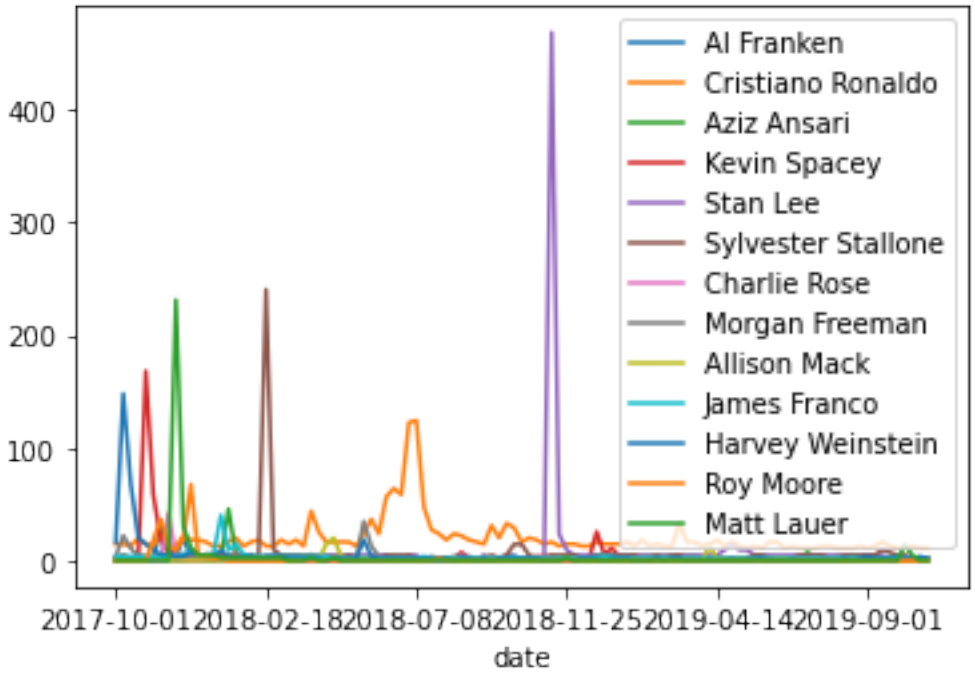


[36]:
```
df['Larry Nassar'].idxmax()
```

[36]:
```
'2018-01-21'
```

[12]:
```
df2[nameList2].plot()
```

[12]:
```
<AxesSubplot:xlabel='date'>
```



0.1 Regression

- How to select τ?
- Use the maximun likelihood to eslimate?
- Estimate α or use the same α for regression 2
- MLE for regression 2

[ ]: