Lab 2 - Jackson Nahom

September 7, 2021

[1]:

```
import pandas as pd
import numpy as np
```

[2]:

```
# load txt file
names = pd.read_csv('data/kddcup.names.txt', header=None, delimiter=':
 ↪',skiprows=1)

# make column 0 into a list
name_list = names[0].tolist()

# add the last column with type
name_list.append('type')

print(name_list)
```

```
['duration', 'protocol_type', 'service', 'flag', 'src_bytes', 'dst_bytes',
'land', 'wrong_fragment', 'urgent', 'hot', 'num_failed_logins', 'logged_in',
'num_compromised', 'root_shell', 'su_attempted', 'num_root',
'num_file_creations', 'num_shells', 'num_access_files', 'num_outbound_cmds',
'is_host_login', 'is_guest_login', 'count', 'srv_count', 'serror_rate',
'srv_serror_rate', 'rerror_rate', 'srv_rerror_rate', 'same_srv_rate',
'diff_srv_rate', 'srv_diff_host_rate', 'dst_host_count', 'dst_host_srv_count',
'dst_host_same_srv_rate', 'dst_host_diff_srv_rate',
'dst_host_same_src_port_rate', 'dst_host_srv_diff_host_rate',
'dst_host_serror_rate', 'dst_host_srv_serror_rate', 'dst_host_rerror_rate',
'dst_host_srv_rerror_rate', 'type']
```

[3]:

```
netattacks = pd.read_csv('data/kddcup.data_10_percent.gz', names=name_list,␣
 ↪header=None, index_col=None)
```

[4]:

```
netattacks.head()
```

[4]:

```
   duration protocol_type service flag  src_bytes  dst_bytes  land  \
0         0           tcp    http   SF        181       5450     0
1         0           tcp    http   SF        239        486     0
2         0           tcp    http   SF        235       1337     0
3         0           tcp    http   SF        219       1337     0
4         0           tcp    http   SF        217       2032     0
```

```
   wrong_fragment  urgent  hot  …  dst_host_srv_count  \
0               0       0    0  …                   9
1               0       0    0  …                  19
2               0       0    0  …                  29
3               0       0    0  …                  39
4               0       0    0  …                  49

   dst_host_same_srv_rate  dst_host_diff_srv_rate  \
0                     1.0                     0.0
1                     1.0                     0.0
2                     1.0                     0.0
3                     1.0                     0.0
4                     1.0                     0.0

   dst_host_same_src_port_rate  dst_host_srv_diff_host_rate  \
0                         0.11                          0.0
1                         0.05                          0.0
2                         0.03                          0.0
3                         0.03                          0.0
4                         0.02                          0.0

   dst_host_serror_rate  dst_host_srv_serror_rate  dst_host_rerror_rate  \
0                   0.0                       0.0                   0.0
1                   0.0                       0.0                   0.0
2                   0.0                       0.0                   0.0
3                   0.0                       0.0                   0.0
4                   0.0                       0.0                   0.0

   dst_host_srv_rerror_rate     type
0                       0.0  normal.
1                       0.0  normal.
2                       0.0  normal.
3                       0.0  normal.
4                       0.0  normal.

[5 rows x 42 columns]
```

```
[5]: netattacks.describe(include='all')
```

```
[5]:              duration protocol_type service    flag     src_bytes  \
     count   494021.000000        494021  494021  494021  4.940210e+05
     unique            NaN             3      66      11           NaN
     top               NaN          icmp   ecr_i      SF           NaN
     freq              NaN        283602  281400  378440           NaN
     mean        47.979302           NaN     NaN     NaN  3.025610e+03
     std        707.746472           NaN     NaN     NaN  9.882181e+05
```

```
min          0.000000           NaN     NaN     NaN  0.000000e+00
25%          0.000000           NaN     NaN     NaN  4.500000e+01
50%          0.000000           NaN     NaN     NaN  5.200000e+02
75%          0.000000           NaN     NaN     NaN  1.032000e+03
max      58329.000000           NaN     NaN     NaN  6.933756e+08

           dst_bytes           land  wrong_fragment         urgent  \
count   4.940210e+05  494021.000000   494021.000000  494021.000000
unique           NaN            NaN             NaN            NaN
top              NaN            NaN             NaN            NaN
freq             NaN            NaN             NaN            NaN
mean    8.685324e+02       0.000045        0.006433       0.000014
std     3.304000e+04       0.006673        0.134805       0.005510
min     0.000000e+00       0.000000        0.000000       0.000000
25%     0.000000e+00       0.000000        0.000000       0.000000
50%     0.000000e+00       0.000000        0.000000       0.000000
75%     0.000000e+00       0.000000        0.000000       0.000000
max     5.155468e+06       1.000000        3.000000       3.000000

                  hot  …  dst_host_srv_count  dst_host_same_srv_rate  \
count   494021.000000  …       494021.000000           494021.000000
unique            NaN  …                 NaN                     NaN
top               NaN  …                 NaN                     NaN
freq              NaN  …                 NaN                     NaN
mean         0.034519  …          188.665670                0.753780
std          0.782103  …          106.040437                0.410781
min          0.000000  …            0.000000                0.000000
25%          0.000000  …           46.000000                0.410000
50%          0.000000  …          255.000000                1.000000
75%          0.000000  …          255.000000                1.000000
max         30.000000  …          255.000000                1.000000

        dst_host_diff_srv_rate  dst_host_same_src_port_rate  \
count            494021.000000                494021.000000
unique                     NaN                          NaN
top                        NaN                          NaN
freq                       NaN                          NaN
mean                  0.030906                     0.601935
std                   0.109259                     0.481309
min                   0.000000                     0.000000
25%                   0.000000                     0.000000
50%                   0.000000                     1.000000
75%                   0.040000                     1.000000
max                   1.000000                     1.000000

        dst_host_srv_diff_host_rate  dst_host_serror_rate  \
count                 494021.000000         494021.000000
```

```
unique                          NaN                   NaN
top                             NaN                   NaN
freq                            NaN                   NaN
mean                       0.006684              0.176754
std                        0.042133              0.380593
min                        0.000000              0.000000
25%                        0.000000              0.000000
50%                        0.000000              0.000000
75%                        0.000000              0.000000
max                        1.000000              1.000000

        dst_host_srv_serror_rate  dst_host_rerror_rate  \
count              494021.000000         494021.000000
unique                       NaN                   NaN
top                          NaN                   NaN
freq                         NaN                   NaN
mean                    0.176443              0.058118
std                     0.380919              0.230590
min                     0.000000              0.000000
25%                     0.000000              0.000000
50%                     0.000000              0.000000
75%                     0.000000              0.000000
max                     1.000000              1.000000

        dst_host_srv_rerror_rate    type
count              494021.000000  494021
unique                       NaN      23
top                          NaN  smurf.
freq                         NaN  280790
mean                    0.057412     NaN
std                     0.230140     NaN
min                     0.000000     NaN
25%                     0.000000     NaN
50%                     0.000000     NaN
75%                     0.000000     NaN
max                     1.000000     NaN

[11 rows x 42 columns]
```

[6]:

```python
# store stats in a dataframe
df_stats = netattacks.describe(include='all')
# save dataframe to file
df_stats.to_csv('output/netattack_summary.csv')
```

[7]:

```python
# The first two are good if you want other stats besides count
# e.g. mean or standard deviation
type_counts = netattacks.groupby('type').count()
```

```python
type_means = netattacks.groupby('type').mean()

# get a multi-index with multiple stats
type_counts = netattacks.groupby('type').agg(['count', 'mean'])
print(type_counts)
# cleanest for just counts
type_counts = netattacks['type'].value_counts()
print(type_counts)
type_counts.head()
```

```
                 duration             src_bytes                dst_bytes  \
                    count        mean     count           mean     count
type
back.                2203    0.128915      2203   54156.355878      2203
buffer_overflow.       30   91.700000        30    1400.433333        30
ftp_write.              8   32.375000         8     220.750000         8
guess_passwd.          53    2.716981        53     125.339623        53
imap.                  12    6.000000        12     347.583333        12
ipsweep.             1247    0.034483      1247      10.083400      1247
land.                  21    0.000000        21       0.000000        21
loadmodule.             9   36.222222         9     151.888889         9
multihop.               7  184.000000         7     435.142857         7
neptune.           107201    0.000000    107201       0.000000    107201
nmap.                 231    0.000000       231      24.116883       231
normal.             97278  216.657322     97278    1157.047524     97278
perl.                   3   41.333333         3     265.666667         3
phf.                    4    4.500000         4      51.000000         4
pod.                  264    0.000000       264    1462.651515       264
portsweep.           1040 1915.299038      1040  666707.436538      1040
rootkit.               10  100.800000        10     294.700000        10
satan.               1589    0.040277      1589       1.337319      1589
smurf.             280790    0.000000    280790     935.772300    280790
spy.                    2  318.000000         2     174.500000         2
teardrop.             979    0.000000       979      28.000000       979
warezclient.         1020  615.257843      1020  300219.562745      1020
warezmaster.           20   15.050000        20      49.300000        20

                                  land          wrong_fragment           ...  \
                          mean   count     mean          count     mean  ...
type                                                                     ...
back.             8.232650e+03    2203  0.00000           2203 0.000000  ...
buffer_overflow.  6.339833e+03      30  0.00000             30 0.000000  ...
ftp_write.        5.382250e+03       8  0.00000              8 0.000000  ...
guess_passwd.     2.161887e+02      53  0.00000             53 0.000000  ...
imap.             5.494867e+04      12  0.00000             12 0.000000  ...
ipsweep.          2.718524e-01    1247  0.00000           1247 0.000000  ...
land.             0.000000e+00      21  1.00000             21 0.000000  ...
loadmodule.       3.009889e+03       9  0.00000              9 0.000000  ...
```

```
multihop.         2.130163e+05       7  0.00000              7  0.000000  …
neptune.          0.000000e+00  107201  0.00000         107201  0.000000  …
nmap.             0.000000e+00     231  0.00000            231  0.000000  …
normal.           3.384651e+03   97278  0.00001          97278  0.000000  …
perl.             2.444000e+03       3  0.00000              3  0.000000  …
phf.              8.127000e+03       4  0.00000              4  0.000000  …
pod.              0.000000e+00     264  0.00000            264  0.981061  …
portsweep.        0.000000e+00    1040  0.00000           1040  0.000000  …
rootkit.          4.276600e+03      10  0.00000             10  0.000000  …
satan.            9.477659e-01    1589  0.00000           1589  0.000000  …
smurf.            0.000000e+00  280790  0.00000         280790  0.000000  …
spy.              1.193500e+03       2  0.00000              2  0.000000  …
teardrop.         5.720123e-02     979  0.00000            979  2.981614  …
warezclient.      7.193176e+02    1020  0.00000           1020  0.000000  …
warezmaster.      3.922088e+06      20  0.00000             20  0.000000  …

                 dst_host_srv_diff_host_rate           dst_host_serror_rate  \
                                       count      mean                count
type
back.                                   2203  0.000000                 2203
buffer_overflow.                          30  0.075000                   30
ftp_write.                                 8  0.117500                    8
guess_passwd.                             53  0.018868                   53
imap.                                     12  0.000000                   12
ipsweep.                                1247  0.602719                 1247
land.                                     21  0.545238                   21
loadmodule.                                9  0.207778                    9
multihop.                                  7  0.000000                    7
neptune.                              107201  0.000031               107201
nmap.                                    231  0.111558                  231
normal.                                97278  0.024123                97278
perl.                                      3  0.000000                    3
phf.                                       4  0.000000                    4
pod.                                     264  0.208977                  264
portsweep.                              1040  0.000058                 1040
rootkit.                                  10  0.010000                   10
satan.                                  1589  0.000195                 1589
smurf.                                280790  0.000000               280790
spy.                                       2  0.000000                    2
teardrop.                                979  0.000000                  979
warezclient.                            1020  0.099382                 1020
warezmaster.                              20  0.000000                   20

                           dst_host_srv_serror_rate            \
                      mean                    count      mean
type
back.             0.002138                     2203  0.002138
buffer_overflow.  0.000000                       30  0.000000
```

```
ftp_write.        0.000000                        8 0.000000
guess_passwd.     0.101509                       53 0.101509
imap.             0.572500                       12 0.572500
ipsweep.          0.000000                     1247 0.000000
land.             0.893810                       21 0.646667
loadmodule.       0.000000                        9 0.000000
multihop.         0.000000                        7 0.000000
neptune.          0.809110                   107201 0.809025
nmap.             0.434675                      231 0.445887
normal.           0.002121                    97278 0.001068
perl.             0.000000                        3 0.000000
phf.              0.000000                        4 0.000000
pod.              0.066705                      264 0.000000
portsweep.        0.011663                     1040 0.023250
rootkit.          0.000000                       10 0.025000
satan.            0.110252                     1589 0.108263
smurf.            0.000011                   280790 0.000000
spy.              0.220000                        2 0.315000
teardrop.         0.020603                      979 0.000000
warezclient.      0.010961                     1020 0.003500
warezmaster.      0.020000                       20 0.000000

                 dst_host_rerror_rate           dst_host_srv_rerror_rate  \
                                count      mean                    count
type
back.                            2203  0.050390                     2203
buffer_overflow.                   30  0.021333                       30
ftp_write.                          8  0.000000                        8
guess_passwd.                      53  0.879245                       53
imap.                              12  0.002500                       12
ipsweep.                         1247  0.069118                     1247
land.                              21  0.006190                       21
loadmodule.                         9  0.000000                        9
multihop.                           7  0.011429                        7
neptune.                       107201  0.190831                   107201
nmap.                             231  0.000000                      231
normal.                         97278  0.057717                    97278
perl.                               3  0.230000                        3
phf.                                4  0.000000                        4
pod.                              264  0.000758                      264
portsweep.                       1040  0.890173                     1040
rootkit.                           10  0.073000                       10
satan.                           1589  0.780302                     1589
smurf.                         280790  0.000028                   280790
spy.                                2  0.000000                        2
teardrop.                         979  0.220061                      979
warezclient.                     1020  0.003961                     1020
warezmaster.                       20  0.006000                       20
```

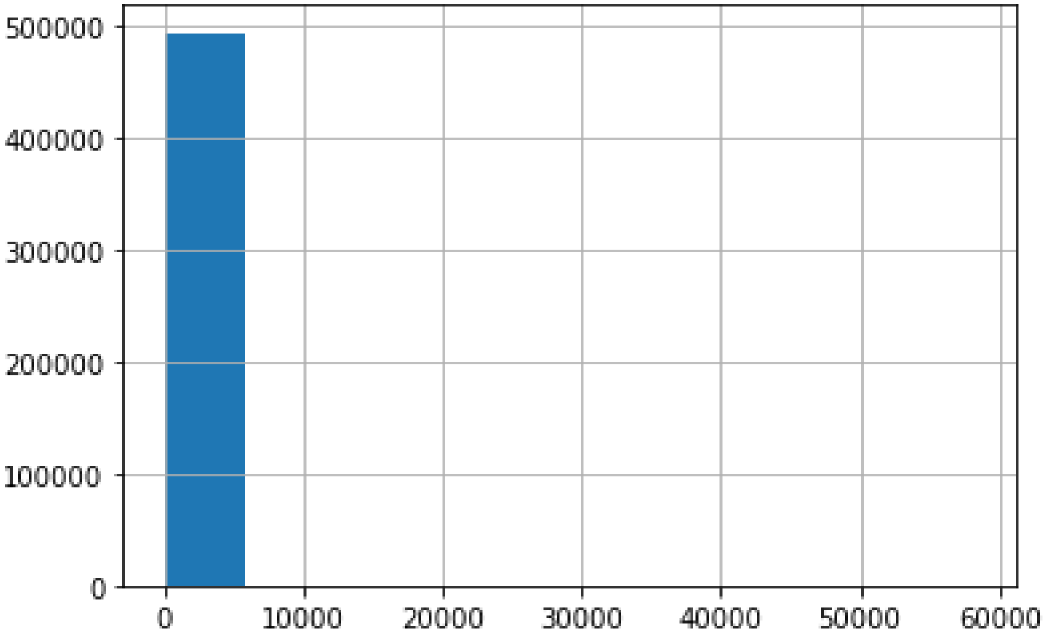
```
                      mean
type
back.             0.050390
buffer_overflow.  0.021333
ftp_write.        0.000000
guess_passwd.     0.879245
imap.             0.000000
ipsweep.          0.066215
land.             0.000000
loadmodule.       0.011111
multihop.         0.000000
neptune.          0.190828
nmap.             0.000000
normal.           0.055830
perl.             0.000000
phf.              0.000000
pod.              0.000000
portsweep.        0.965260
rootkit.          0.025000
satan.            0.773442
smurf.            0.000000
spy.              0.000000
teardrop.         0.000000
warezclient.      0.000520
warezmaster.      0.000000

[23 rows x 76 columns]
smurf.              280790
neptune.            107201
normal.              97278
back.                 2203
satan.                1589
ipsweep.              1247
portsweep.            1040
warezclient.          1020
teardrop.              979
pod.                   264
nmap.                  231
guess_passwd.           53
buffer_overflow.        30
land.                   21
warezmaster.            20
imap.                   12
rootkit.                10
loadmodule.              9
ftp_write.               8
```

```
multihop.                 7
phf.                      4
perl.                     3
spy.                      2
Name: type, dtype: int64
```

```
[7]: smurf.      280790
     neptune.    107201
     normal.      97278
     back.         2203
     satan.        1589
     Name: type, dtype: int64
```

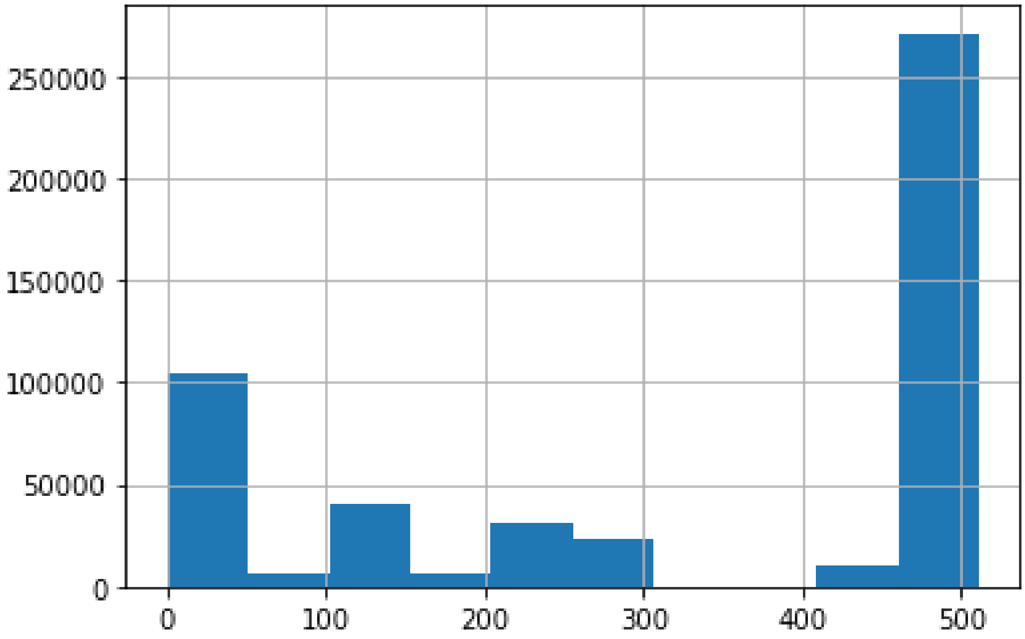
```
[8]: netattacks['duration'].hist()
```

```
[8]: <AxesSubplot:>
```



```
[9]: netattacks['count'].hist()
```

```
[9]: <AxesSubplot:>
```



[10]:
```
netattacks['duration'].corr(netattacks['count'])
```

[10]:
```
-0.10515338141725143
```

[11]:
```
# https://www.dataquest.io/blog/
 ↪tutorial-add-column-pandas-dataframe-based-on-if-else-condition/
netattacks['label'] = np.where(netattacks['type'] == 'normal.', 'good', 'bad')
netattacks['label'].value_counts()
```

[11]:
```
bad     396743
good     97278
Name: label, dtype: int64
```

[ ]:
```
netattacks['duration'].corr(netattacks['protocol_type'])
#Error: unsupported operand type(s) for /: 'str' and 'int'
```

[ ]:
```
netattacks['protocol_type'].corr(netattacks['duration'])
#Error
#you can not corrilate continous and categorical
#You can't run correlations with strings, to do so you need to use dummy␣
 ↪variables
```

1 Task 1:

- You can not run correlations between continous and categorical variables

- If you try to run the pandas .corr you will recieve and error: #Error: unsupported operand type(s) for /: ‘str’ and ‘int’.
- Regression is a mathamatical equation so we can not insert categorical data that would be strings.
- In order to use categorical variables we would need to covert the categories to dummy variables. For example for protocol_type convert TCP to be 1 and UDP to be 2. If we use protocol_type_dummy then we could run correlations between continous and categorical variables.

2 Task 2

```
[13]: netattacks.corr().to_csv('corr.csv')
      linear_reg = netattacks.corr()
      netattacks.corr()
```

```
[13]:                              duration  src_bytes  dst_bytes      land  \
duration                   1.000000   0.004258   0.005440 -0.000452
src_bytes                  0.004258   1.000000  -0.000002 -0.000020
dst_bytes                  0.005440  -0.000002   1.000000 -0.000175
land                      -0.000452  -0.000020  -0.000175  1.000000
wrong_fragment            -0.003235  -0.000139  -0.001254 -0.000318
urgent                     0.003786  -0.000005   0.016288 -0.000017
hot                        0.013213   0.004483   0.004365 -0.000295
num_failed_logins          0.005239  -0.000027   0.049330 -0.000065
logged_in                 -0.017265   0.001701   0.047814 -0.002784
num_compromised            0.058095   0.000119   0.023298 -0.000038
root_shell                 0.021340  -0.000022   0.031680 -0.000070
su_attempted               0.055853  -0.000010   0.075656 -0.000031
num_root                   0.056766  -0.000010   0.020746 -0.000038
num_file_creations         0.074562   0.000013   0.004958 -0.000075
num_shells                -0.000169   0.000005   0.000144 -0.000066
num_access_files           0.025661  -0.000052   0.008746 -0.000184
num_outbound_cmds               NaN        NaN        NaN       NaN
is_host_login                   NaN        NaN        NaN       NaN
is_guest_login             0.023424  -0.000082   0.001289 -0.000249
count                     -0.105153  -0.003098  -0.040373 -0.010260
srv_count                 -0.080250  -0.002501  -0.030544 -0.007886
serror_rate               -0.031416   0.001558  -0.011908  0.013898
srv_serror_rate           -0.031378   0.001114  -0.011930  0.014422
rerror_rate                0.012053   0.000591  -0.006166 -0.000777
srv_rerror_rate            0.012106   0.001379  -0.005808 -0.001659
same_srv_rate              0.021771  -0.001860   0.014002  0.002286
diff_srv_rate              0.051800   0.006207  -0.005702  0.002282
srv_diff_host_rate        -0.011790  -0.000015   0.008135  0.036985
dst_host_count             0.010074  -0.001743  -0.048869 -0.023671
dst_host_srv_count        -0.117515  -0.003212  -0.005850 -0.011587
dst_host_same_srv_rate    -0.118458  -0.002052   0.007058  0.001984
```

```
dst_host_diff_srv_rate       0.406233   0.000578  -0.005314 -0.000333
dst_host_same_src_port_rate  0.042642  -0.000724  -0.020143  0.003799
dst_host_srv_diff_host_rate -0.006983   0.001186   0.008707  0.083320
dst_host_serror_rate        -0.030400  -0.000718  -0.011334  0.012658
dst_host_srv_serror_rate    -0.030612   0.001122  -0.011235  0.007795
dst_host_rerror_rate         0.006739  -0.000393  -0.005000 -0.001511
dst_host_srv_rerror_rate     0.010465   0.001328  -0.005471 -0.001665

                             wrong_fragment    urgent       hot  \
duration                          -0.003235  0.003786  0.013213
src_bytes                         -0.000139 -0.000005  0.004483
dst_bytes                         -0.001254  0.016288  0.004365
land                              -0.000318 -0.000017 -0.000295
wrong_fragment                     1.000000 -0.000123 -0.002106
urgent                            -0.000123  1.000000  0.000356
hot                               -0.002106  0.000356  1.000000
num_failed_logins                 -0.000467  0.141996  0.008740
logged_in                         -0.019908  0.006164  0.105305
num_compromised                   -0.000271  0.014285  0.007348
root_shell                        -0.000504  0.034790  0.024065
su_attempted                      -0.000223 -0.000012 -0.000206
num_root                          -0.000269  0.009476  0.000998
num_file_creations                -0.000536  0.015211  0.025247
num_shells                        -0.000473 -0.000026  0.006373
num_access_files                  -0.001319  0.020068  0.001902
num_outbound_cmds                       NaN       NaN       NaN
is_host_login                           NaN       NaN       NaN
is_guest_login                    -0.001778 -0.000096  0.843572
count                             -0.061934 -0.003997 -0.068451
srv_count                         -0.047789 -0.003047 -0.052164
serror_rate                       -0.013969 -0.001193 -0.020264
srv_serror_rate                   -0.022119 -0.001192 -0.020217
rerror_rate                       -0.011529 -0.000638 -0.008305
srv_rerror_rate                   -0.011865 -0.000639 -0.005821
same_srv_rate                      0.017416  0.001381  0.022697
diff_srv_rate                     -0.007077 -0.000656 -0.002686
srv_diff_host_rate                 0.000153 -0.000524  0.001973
dst_host_count                    -0.005191 -0.007139 -0.026366
dst_host_srv_count                -0.058624 -0.004540 -0.038730
dst_host_same_srv_rate            -0.054903 -0.003279 -0.029117
dst_host_diff_srv_rate             0.071857  0.010536  0.001319
dst_host_same_src_port_rate       -0.031803 -0.002002 -0.052923
dst_host_srv_diff_host_rate        0.012092 -0.000408 -0.004467
dst_host_serror_rate              -0.019091 -0.001194 -0.019491
dst_host_srv_serror_rate          -0.022104 -0.001191 -0.020201
dst_host_rerror_rate               0.029774 -0.000648 -0.006541
dst_host_srv_rerror_rate          -0.011904 -0.000641 -0.007749
```

```
                            num_failed_logins  logged_in  num_compromised  \
duration                             0.005239  -0.017265         0.058095   
src_bytes                           -0.000027   0.001701         0.000119   
dst_bytes                            0.049330   0.047814         0.023298   
land                                -0.000065  -0.002784        -0.000038   
wrong_fragment                      -0.000467  -0.019908        -0.000271   
urgent                               0.141996   0.006164         0.014285   
hot                                  0.008740   0.105305         0.007348   
num_failed_logins                    1.000000  -0.001145         0.006907   
logged_in                           -0.001145   1.000000         0.013612   
num_compromised                      0.006907   0.013612         1.000000   
root_shell                           0.036983   0.025293         0.255557   
su_attempted                         0.117117   0.011207         0.701400   
num_root                             0.003250   0.013519         0.993828   
num_file_creations                   0.003948   0.026923         0.010934   
num_shells                          -0.000097   0.023776         0.009341   
num_access_files                     0.003305   0.066233         0.412238   
num_outbound_cmds                         NaN        NaN              NaN   
is_host_login                             NaN        NaN              NaN   
is_guest_login                      -0.000365   0.089318        -0.000212   
count                               -0.015184  -0.634643        -0.008792   
srv_count                           -0.011578  -0.478122        -0.006704   
serror_rate                         -0.003169  -0.191698        -0.002597   
srv_serror_rate                     -0.003850  -0.191113        -0.002618   
rerror_rate                          0.025167  -0.099137        -0.001049   
srv_rerror_rate                      0.025098  -0.094372        -0.000478   
same_srv_rate                        0.004581   0.219685         0.003012   
diff_srv_rate                        0.003850  -0.072692        -0.001338   
srv_diff_host_rate                  -0.001992   0.330673         0.000770   
dst_host_count                      -0.025444  -0.621029        -0.008361   
dst_host_srv_count                  -0.015413   0.119315        -0.004797   
dst_host_same_srv_rate               0.000507   0.161070        -0.002584   
dst_host_diff_srv_rate               0.001017  -0.061151         0.000359   
dst_host_same_src_port_rate         -0.009565  -0.461558        -0.006715   
dst_host_srv_diff_host_rate          0.016001   0.140493         0.000621   
dst_host_serror_rate                -0.001945  -0.190955        -0.001978   
dst_host_srv_serror_rate            -0.002453  -0.191704        -0.001631   
dst_host_rerror_rate                 0.024753  -0.090868        -0.000843   
dst_host_srv_rerror_rate             0.023584  -0.087885        -0.000873   

                            …  dst_host_count  dst_host_srv_count  \
duration                    …        0.010074           -0.117515   
src_bytes                   …       -0.001743           -0.003212   
dst_bytes                   …       -0.048869           -0.005850   
land                        …       -0.023671           -0.011587   
wrong_fragment              …       -0.005191           -0.058624   
```

```
urgent                       …       -0.007139           -0.004540
hot                          …       -0.026366           -0.038730
num_failed_logins            …       -0.025444           -0.015413
logged_in                    …       -0.621029            0.119315
num_compromised              …       -0.008361           -0.004797
root_shell                   …       -0.024384           -0.010026
su_attempted                 …       -0.006570           -0.006531
num_root                     …       -0.011611           -0.007985
num_file_creations           …       -0.019126           -0.013871
num_shells                   …       -0.017111           -0.012207
num_access_files             …       -0.021152           -0.000801
num_outbound_cmds            …             NaN                 NaN
is_host_login                …             NaN                 NaN
is_guest_login               …       -0.033787           -0.050016
count                        …        0.532632            0.514581
srv_count                    …        0.401536            0.718452
serror_rate                  …        0.156638           -0.775273
srv_serror_rate              …        0.156220           -0.774138
rerror_rate                  …       -0.088709           -0.329412
srv_rerror_rate              …       -0.087167           -0.327931
same_srv_rate                …       -0.181051            0.898955
diff_srv_rate                …        0.056804           -0.415498
srv_diff_host_rate           …       -0.382389            0.005096
dst_host_count               …        1.000000           -0.027236
dst_host_srv_count           …       -0.027236            1.000000
dst_host_same_srv_rate       …       -0.126892            0.973685
dst_host_diff_srv_rate       …        0.025170           -0.462367
dst_host_same_src_port_rate  …        0.290747            0.677580
dst_host_srv_diff_host_rate  …       -0.491162           -0.017112
dst_host_serror_rate         …        0.156728           -0.776070
dst_host_srv_serror_rate     …        0.157368           -0.774748
dst_host_rerror_rate         …       -0.092460           -0.332683
dst_host_srv_rerror_rate     …       -0.087791           -0.331625

                             dst_host_same_srv_rate  dst_host_diff_srv_rate  \
duration                                  -0.118458                0.406233
src_bytes                                 -0.002052                0.000578
dst_bytes                                  0.007058               -0.005314
land                                       0.001984               -0.000333
wrong_fragment                            -0.054903                0.071857
urgent                                    -0.003279                0.010536
hot                                       -0.029117                0.001319
num_failed_logins                          0.000507                0.001017
logged_in                                  0.161070               -0.061151
num_compromised                           -0.002584                0.000359
root_shell                                 0.000935               -0.000684
su_attempted                              -0.005582                0.000793
```

```
num_root                                  -0.006731                0.002863
num_file_creations                       -0.008808                0.006513
num_shells                                -0.007729                0.000641
num_access_files                            0.004395                0.002102
num_outbound_cmds                              NaN                     NaN
is_host_login                                  NaN                     NaN
is_guest_login                            -0.043189                0.009624
count                                       0.468775               -0.262107
srv_count                                   0.687993               -0.328593
serror_rate                               -0.799902                0.160595
srv_serror_rate                           -0.798736                0.158045
rerror_rate                               -0.319822                0.210199
srv_rerror_rate                           -0.318238                0.209200
same_srv_rate                               0.927808               -0.266172
diff_srv_rate                             -0.425760                0.524572
srv_diff_host_rate                          0.047734               -0.005613
dst_host_count                            -0.126892                0.025170
dst_host_srv_count                          0.973685               -0.462367
dst_host_same_srv_rate                      1.000000               -0.469706
dst_host_diff_srv_rate                    -0.469706                1.000000
dst_host_same_src_port_rate                 0.671338               -0.157512
dst_host_srv_diff_host_rate                 0.058055                0.006338
dst_host_serror_rate                      -0.800723                0.159359
dst_host_srv_serror_rate                  -0.799424                0.158195
dst_host_rerror_rate                      -0.323162                0.221739
dst_host_srv_rerror_rate                  -0.321608                0.211372

                             dst_host_same_src_port_rate  \
duration                                        0.042642
src_bytes                                      -0.000724
dst_bytes                                      -0.020143
land                                            0.003799
wrong_fragment                                 -0.031803
urgent                                         -0.002002
hot                                            -0.052923
num_failed_logins                              -0.009565
logged_in                                      -0.461558
num_compromised                                -0.006715
root_shell                                     -0.006273
su_attempted                                   -0.005745
num_root                                       -0.005216
num_file_creations                             -0.009924
num_shells                                     -0.003448
num_access_files                               -0.032291
num_outbound_cmds                                    NaN
is_host_login                                        NaN
is_guest_login                                 -0.045139
```

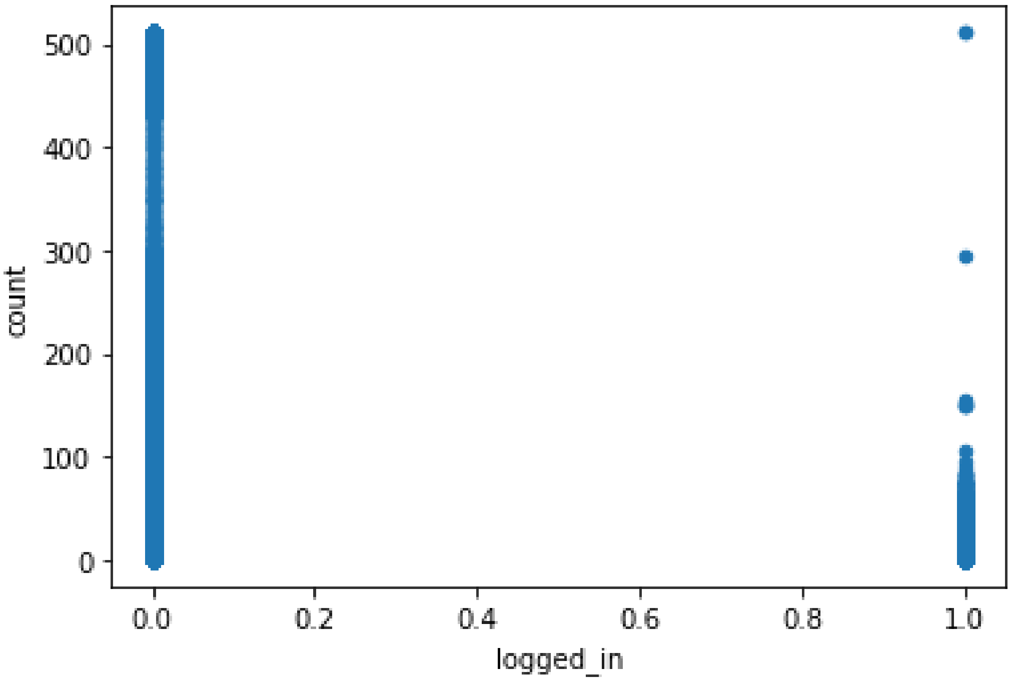
```
count                                           0.860579
srv_count                                       0.944926
serror_rate                                    -0.578416
srv_serror_rate                                -0.577825
rerror_rate                                    -0.265332
srv_rerror_rate                                -0.266598
same_srv_rate                                   0.660744
diff_srv_rate                                  -0.266495
srv_diff_host_rate                             -0.187881
dst_host_count                                  0.290747
dst_host_srv_count                              0.677580
dst_host_same_srv_rate                          0.671338
dst_host_diff_srv_rate                         -0.157512
dst_host_same_src_port_rate                     1.000000
dst_host_srv_diff_host_rate                    -0.064396
dst_host_serror_rate                           -0.578857
dst_host_srv_serror_rate                       -0.577618
dst_host_rerror_rate                           -0.268894
dst_host_srv_rerror_rate                       -0.268798

                             dst_host_srv_diff_host_rate  \
duration                                       -0.006983
src_bytes                                       0.001186
dst_bytes                                       0.008707
land                                            0.083320
wrong_fragment                                  0.012092
urgent                                         -0.000408
hot                                            -0.004467
num_failed_logins                               0.016001
logged_in                                       0.140493
num_compromised                                 0.000621
root_shell                                      0.016950
su_attempted                                    0.001848
num_root                                        0.002311
num_file_creations                              0.009938
num_shells                                      0.004835
num_access_files                                0.000740
num_outbound_cmds                                    NaN
is_host_login                                        NaN
is_guest_login                                 -0.005627
count                                          -0.244573
srv_count                                      -0.183215
serror_rate                                    -0.071137
srv_serror_rate                                -0.070892
rerror_rate                                     0.137343
srv_rerror_rate                                 0.135028
same_srv_rate                                   0.083454
```

```
diff_srv_rate                                  -0.028008
srv_diff_host_rate                              0.259985
dst_host_count                                 -0.491162
dst_host_srv_count                             -0.017112
dst_host_same_srv_rate                          0.058055
dst_host_diff_srv_rate                          0.006338
dst_host_same_src_port_rate                    -0.064396
dst_host_srv_diff_host_rate                     1.000000
dst_host_serror_rate                           -0.071132
dst_host_srv_serror_rate                       -0.071657
dst_host_rerror_rate                            0.139267
dst_host_srv_rerror_rate                        0.140224

                             dst_host_serror_rate  dst_host_srv_serror_rate  \
duration                                -0.030400                 -0.030612
src_bytes                               -0.000718                  0.001122
dst_bytes                               -0.011334                 -0.011235
land                                     0.012658                  0.007795
wrong_fragment                          -0.019091                 -0.022104
urgent                                  -0.001194                 -0.001191
hot                                     -0.019491                 -0.020201
num_failed_logins                       -0.001945                 -0.002453
logged_in                               -0.190955                 -0.191704
num_compromised                         -0.001978                 -0.001631
root_shell                              -0.004553                 -0.004208
su_attempted                            -0.000438                  0.000003
num_root                                -0.002078                 -0.001682
num_file_creations                      -0.002763                 -0.003685
num_shells                              -0.004433                 -0.004402
num_access_files                        -0.012330                 -0.012065
num_outbound_cmds                             NaN                       NaN
is_host_login                                 NaN                       NaN
is_guest_login                          -0.016196                 -0.017197
count                                   -0.311191                 -0.309709
srv_count                               -0.531342                 -0.529981
serror_rate                              0.998673                  0.997849
srv_serror_rate                          0.997835                  0.999304
rerror_rate                             -0.111550                 -0.111126
srv_rerror_rate                         -0.111856                 -0.114843
same_srv_rate                           -0.857835                 -0.857544
diff_srv_rate                            0.253278                  0.253452
srv_diff_host_rate                      -0.091336                 -0.092686
dst_host_count                           0.156728                  0.157368
dst_host_srv_count                      -0.776070                 -0.774748
dst_host_same_srv_rate                  -0.800723                 -0.799424
dst_host_diff_srv_rate                   0.159359                  0.158195
dst_host_same_src_port_rate             -0.578857                 -0.577618
```

```
dst_host_srv_diff_host_rate             -0.071132                -0.071657
dst_host_serror_rate                     1.000000                 0.998156
dst_host_srv_serror_rate                 0.998156                 1.000000
dst_host_rerror_rate                    -0.113403                -0.113184
dst_host_srv_rerror_rate                -0.112256                -0.115320

                             dst_host_rerror_rate  dst_host_srv_rerror_rate
duration                                 0.006739                 0.010465
src_bytes                               -0.000393                 0.001328
dst_bytes                               -0.005000                -0.005471
land                                    -0.001511                -0.001665
wrong_fragment                           0.029774                -0.011904
urgent                                  -0.000648                -0.000641
hot                                     -0.006541                -0.007749
num_failed_logins                        0.024753                 0.023584
logged_in                               -0.090868                -0.087885
num_compromised                         -0.000843                -0.000873
root_shell                               0.000003                -0.001799
su_attempted                             0.000646                -0.000873
num_root                                -0.001003                -0.001383
num_file_creations                      -0.001513                -0.001478
num_shells                              -0.001918                -0.002315
num_access_files                        -0.006611                -0.006780
num_outbound_cmds                             NaN                      NaN
is_host_login                                 NaN                      NaN
is_guest_login                          -0.007556                -0.008693
count                                   -0.211758                -0.208051
srv_count                               -0.291104                -0.288450
serror_rate                             -0.112393                -0.112288
srv_serror_rate                         -0.113219                -0.115346
rerror_rate                              0.986995                 0.985200
srv_rerror_rate                          0.982166                 0.986571
same_srv_rate                           -0.330291                -0.331992
diff_srv_rate                            0.227657                 0.233983
srv_diff_host_rate                       0.024295                 0.020311
dst_host_count                          -0.092460                -0.087791
dst_host_srv_count                      -0.332683                -0.331625
dst_host_same_srv_rate                  -0.323162                -0.321608
dst_host_diff_srv_rate                   0.221739                 0.211372
dst_host_same_src_port_rate             -0.268894                -0.268798
dst_host_srv_diff_host_rate              0.139267                 0.140224
dst_host_serror_rate                    -0.113403                -0.112256
dst_host_srv_serror_rate                -0.113184                -0.115320
dst_host_rerror_rate                     1.000000                 0.984804
dst_host_srv_rerror_rate                 0.984804                 1.000000

[38 rows x 38 columns]
```

[14]:
```
netattacks.plot.scatter('logged_in', 'count')
```

[14]:
```
<AxesSubplot:xlabel='logged_in', ylabel='count'>
```

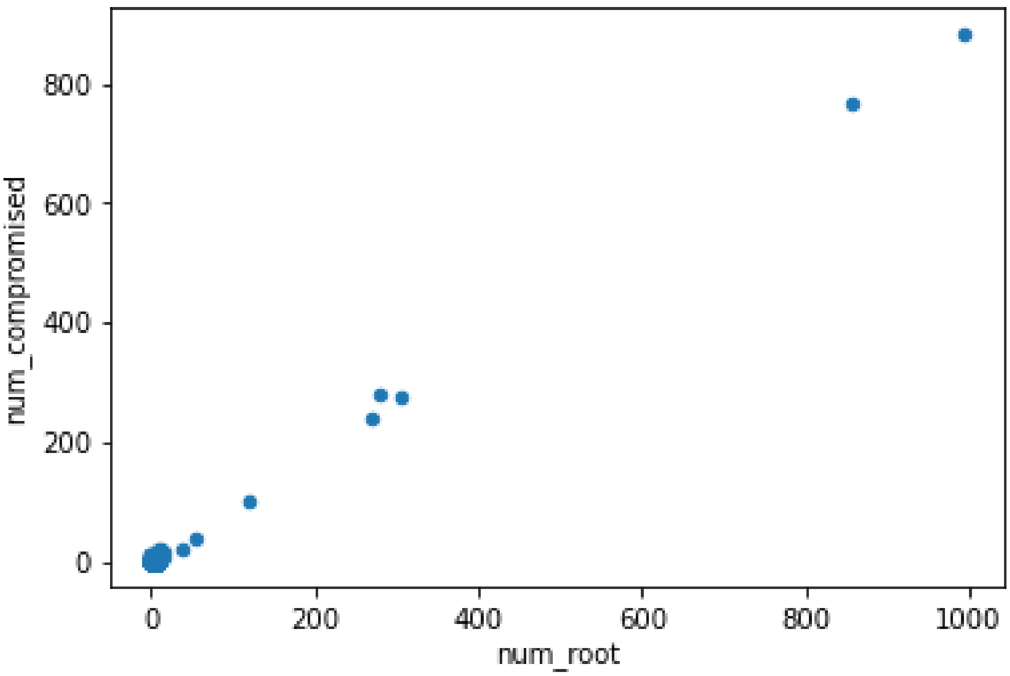


[15]:
```
cols = linear_reg.columns.tolist()
frames = []
for col in cols:
    frame = linear_reg.query(f'({col} > .5 or {col} < -.5) and {col} != 1')
    frames.append(frame)
    #print(frame)
df_interest = pd.concat(frames)
df_interest.drop_duplicates(inplace=True)
df_interest.to_csv('Check_corr.csv')
```

[16]:
```
netattacks.corr(method ='pearson').to_csv('corr.csv')
```

[34]:
```
netattacks.plot.scatter('num_root', 'num_compromised')
netattacks['num_root'].corr(netattacks['num_compromised'])
```

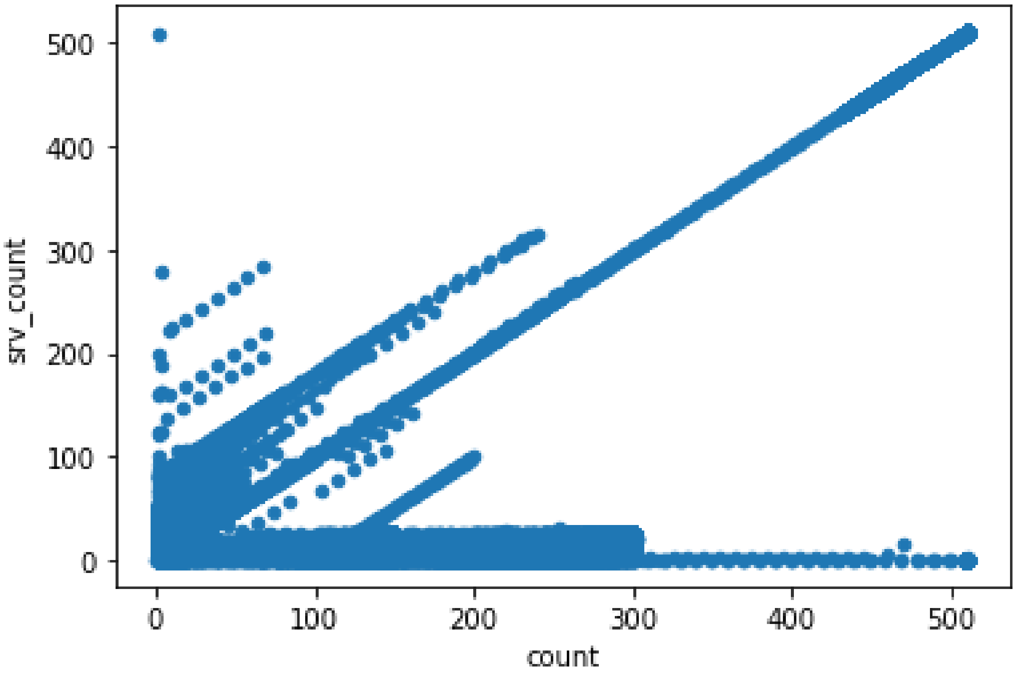
[34]:
```
0.9938277978738366
```



This correlation show us that it is almost 1 to 1 for the more root accesses there are it is likely a connection is comprimised.

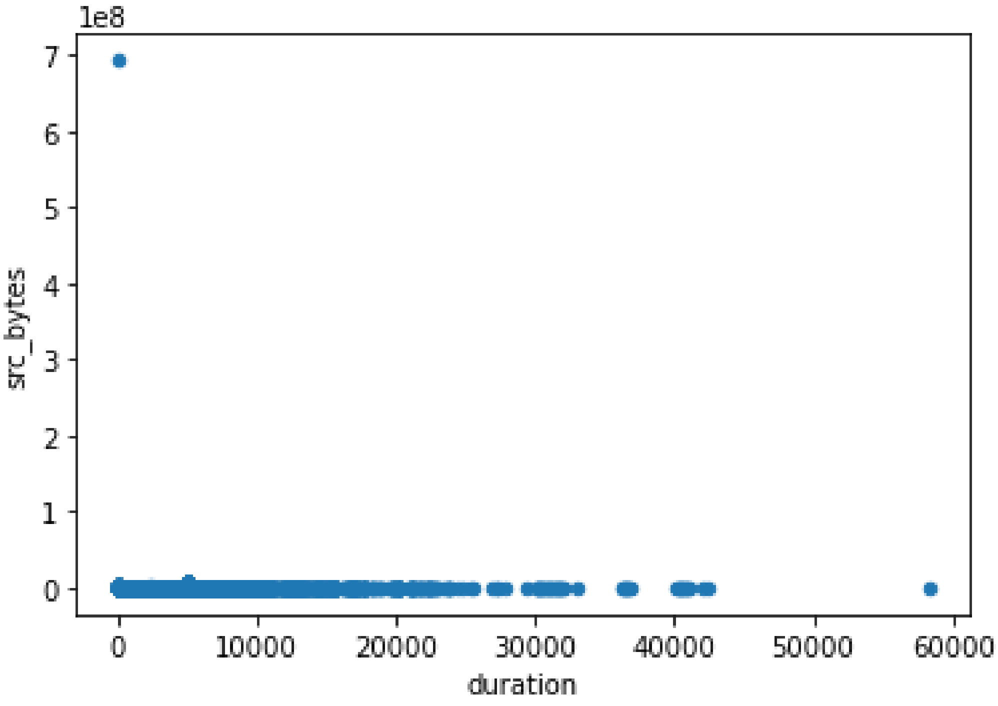
[33]:
```
netattacks.plot.scatter('count', 'srv_count')
netattacks['count'].corr(netattacks['srv_count'])
```

[33]:
```
0.9436670688882656
```



```
[32]: netattacks.plot.scatter('duration', 'src_bytes')
      netattacks['duration'].corr(netattacks['src_bytes'])
```

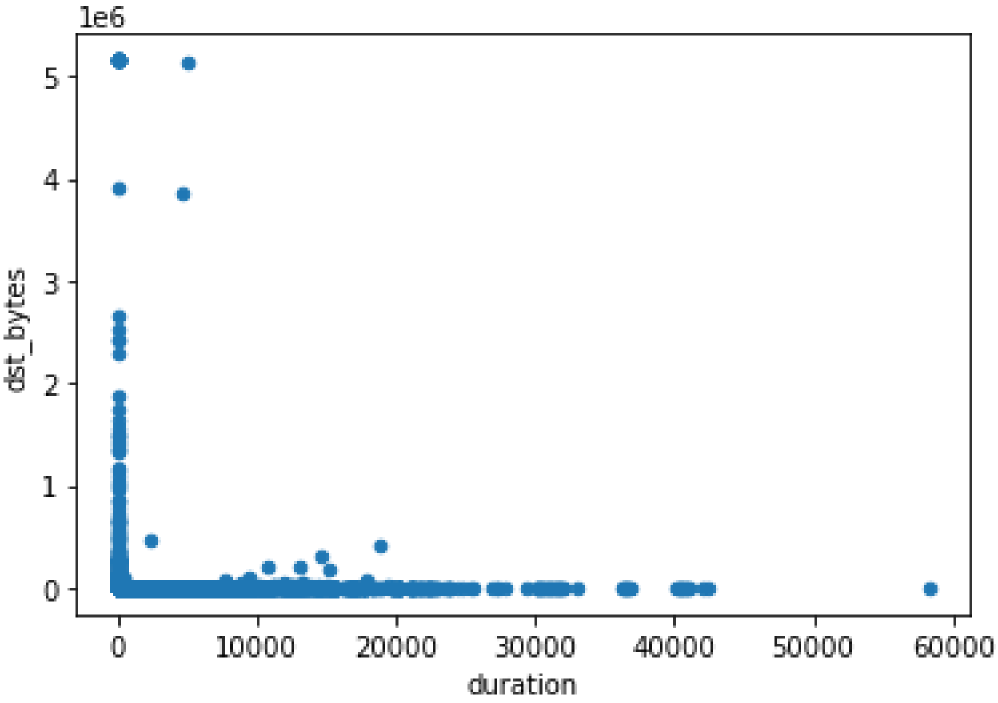
```
[32]: 0.0042582302695068665
```



I had a thought that the longer a connection was the more bytes from the source to destination. The correlation proves otherwise showing that there is not any relationtion between the two variables.

[36]:
```
netattacks.plot.scatter('duration', 'dst_bytes')
netattacks['duration'].corr(netattacks['dst_bytes'])
```

[36]:
```
0.00543953447823358
```



I had a similar thought for there might be a correlation with duration and bytes destination to source. Although it slightly better than the pervious correlation however, it is still not significant.

3 Task 3

[17]:
```
netattacks['protocol_type'].value_counts()
```

[17]:
```
icmp    283602
tcp     190065
udp      20354
Name: protocol_type, dtype: int64
```

[19]:
```
netattacks['protocol_type'].describe()
```

```
[19]: count     494021
      unique         3
      top         icmp
      freq      283602
      Name: protocol_type, dtype: object
```

There are more TCP attacks than UDP.