In [1]:
```
## Import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [2]:
```
## Import CSV with pandas
data = pd.read_csv("C:\\Users\\Jehanne\\Desktop\\RTU\\Telecommunication Software\\R
```

In [6]:
```
## Let's analyze the data roughly

print("\n\n\n###### Shape : ")
print(data.shape)

print("\n\n\n###### Columns : ")
print(data.columns)

print("\n\n\n###### Their type : ")
pd.set_option('display.max_rows', None) #by default it will only show a handful so
print(data.dtypes)
pd.set_option('display.max_rows',10)

print("\n\n\n###### First values : ")
print(data.head())

print("\n\n\n###### Statistics : ")
print(data.describe())
```

```
###### Shape :
(4234, 65)



###### Columns :
Index(['id_flow', 'nw_src', 'tp_src', 'nw_dst', 'tp_dst', 'nw_proto',
       'forward_pc', 'forward_bc', 'forward_pl', 'forward_piat', 'forward_pps',
       'forward_bps', 'forward_pl_mean', 'forward_piat_mean',
       'forward_pps_mean', 'forward_bps_mean', 'forward_pl_var',
       'forward_piat_var', 'forward_pps_var', 'forward_bps_var',
       'forward_pl_q1', 'forward_pl_q3', 'forward_piat_q1', 'forward_piat_q3',
       'forward_pl_max', 'forward_pl_min', 'forward_piat_max',
       'forward_piat_min', 'forward_pps_max', 'forward_pps_min',
       'forward_bps_max', 'forward_bps_min', 'forward_duration',
       'forward_size_packets', 'forward_size_bytes', 'reverse_pc',
       'reverse_bc', 'reverse_pl', 'reverse_piat', 'reverse_pps',
       'reverse_bps', 'reverse_pl_mean', 'reverse_piat_mean',
       'reverse_pps_mean', 'reverse_bps_mean', 'reverse_pl_var',
       'reverse_piat_var', 'reverse_pps_var', 'reverse_bps_var',
       'reverse_pl_q1', 'reverse_pl_q3', 'reverse_piat_q1', 'reverse_piat_q3',
       'reverse_pl_max', 'reverse_pl_min', 'reverse_piat_max',
       'reverse_piat_min', 'reverse_pps_max', 'reverse_pps_min',
       'reverse_bps_max', 'reverse_bps_min', 'reverse_duration',
       'reverse_size_packets', 'reverse_size_bytes', 'category'],
      dtype='object')



###### Their type :
id_flow                 object
nw_src                  object
tp_src                   int64
nw_dst                  object
tp_dst                   int64
nw_proto                 int64
forward_pc               int64
forward_bc               int64
forward_pl             float64
forward_piat           float64
forward_pps            float64
forward_bps            float64
forward_pl_mean        float64
forward_piat_mean      float64
forward_pps_mean       float64
forward_bps_mean       float64
forward_pl_var         float64
forward_piat_var       float64
forward_pps_var        float64
forward_bps_var         object
forward_pl_q1          float64
forward_pl_q3          float64
forward_piat_q1        float64
forward_piat_q3        float64
```

```
forward_pl_max          float64
forward_pl_min          float64
forward_piat_max        float64
forward_piat_min        float64
forward_pps_max         float64
forward_pps_min         float64
forward_bps_max         float64
forward_bps_min         float64
forward_duration          int64
forward_size_packets      int64
forward_size_bytes        int64
reverse_pc                int64
reverse_bc              float64
reverse_pl              float64
reverse_piat            float64
reverse_pps             float64
reverse_bps             float64
reverse_pl_mean         float64
reverse_piat_mean       float64
reverse_pps_mean        float64
reverse_bps_mean        float64
reverse_pl_var          float64
reverse_piat_var        float64
reverse_pps_var         float64
reverse_bps_var         float64
reverse_pl_q1           float64
reverse_pl_q3           float64
reverse_piat_q1         float64
reverse_piat_q3         float64
reverse_pl_max          float64
reverse_pl_min          float64
reverse_piat_max        float64
reverse_piat_min        float64
reverse_pps_max         float64
reverse_pps_min         float64
reverse_bps_max         float64
reverse_bps_min         float64
reverse_duration          int64
reverse_size_packets      int64
reverse_size_bytes        int64
category                 object
dtype: object



###### First values :
                           id_flow         nw_src  tp_src         nw_dst  \
0  b2bb77a570fcfa9325eb9e51b6116d2a  172.16.25.104   41402  34.107.221.82
1  f07977b0d1d6645c4fe1e9efea080ff3  172.16.25.104   41406  34.107.221.82
2  e4026ba9b6c1957516e92bdd0d04878f  172.16.25.104   38232    52.84.77.43
3  e2d747932e41500b1463fe8ae4299ecb  172.16.25.104   38234    52.84.77.43
4  56325703391225ad65e013e7a2b02fac  172.16.25.104   60166    52.32.34.32

   tp_dst  nw_proto  forward_pc  forward_bc  forward_pl  forward_piat  ...  \
0      80         6           5         300       60.00           6.0  ...
1      80         6           5         300       60.00           6.0  ...
```

```
2     443         6           3         198       66.00          10.0  ...
3     443         6           3         198       66.00          10.0  ...
4     443         6           4         265       66.25           7.5  ...

   reverse_piat_max  reverse_piat_min  reverse_pps_max  reverse_pps_min  \
0         10.333333              6.00         0.166667         0.096774
1         10.000000              6.20         0.161290         0.100000
2         10.333333             10.00         0.100000         0.096774
3         10.333333             10.00         0.100000         0.096774
4          7.750000              7.75         0.129032         0.129032

   reverse_bps_max  reverse_bps_min  reverse_duration  reverse_size_packets  \
0        15.133333         5.806452               121                    15
1        15.133333         6.000000               121                    15
2         6.000000         5.806452                91                     9
3         6.000000         5.806452                91                     9
4         8.548387         8.548387                31                     4

   reverse_size_bytes category
0                1114      WWW
1                1114      WWW
2                 540      WWW
3                 540      WWW
4                 265      WWW

[5 rows x 65 columns]



###### Statistics :
             tp_src        tp_dst     nw_proto     forward_pc    forward_bc  \
count   4234.000000   4234.000000  4234.000000    4234.000000  4.234000e+03
mean   39994.956542   8540.046528     6.660132    3835.848370  7.356521e+06
std    17331.881734  17575.486397     3.815368   18375.794566  3.585172e+07
min        0.000000      0.000000     1.000000       0.000000  0.000000e+00
25%    35248.500000     80.000000     6.000000       2.000000  1.200000e+02
50%    44009.000000    443.000000     6.000000       3.000000  1.980000e+02
75%    52130.250000    443.000000     6.000000       6.000000  3.850000e+02
max    65534.000000  60949.000000    17.000000  181104.000000  3.558093e+08

          forward_pl  forward_piat   forward_pps   forward_bps  \
count    4234.000000   4234.000000  4.234000e+03  4.234000e+03
mean      316.336560     15.261581  4.788105e+02  2.576202e+06
std      3732.045349    182.065520  2.021312e+04  1.200390e+08
min         0.000000      0.000000  0.000000e+00  0.000000e+00
25%        60.000000      0.048051  6.451613e-02  4.000000e+00
50%        66.000000      3.500000  1.666667e-01  1.260000e+01
75%        79.811688      7.500000  5.161290e-01  4.600000e+01
max    154375.000000   4125.000000  1.303625e+06  7.422774e+09

       forward_pl_mean  ...  reverse_pl_min  reverse_piat_max  \
count      4234.000000  ...     4234.000000       4234.000000
mean       1582.814224  ...       54.418871         23.652912
std        9644.341190  ...      269.495303        229.416470
min           0.000000  ...        0.000000          0.000000
25%          43.000000  ...        0.000938          0.000433
```

```
50%          61.250000  ...       15.500000          7.500000
75%          98.000000  ...       60.000000         15.500000
max      162975.000000  ...     5573.208202       4125.000000

       reverse_piat_min  reverse_pps_max  reverse_pps_min  reverse_bps_max  \
count      4.234000e+03     4.234000e+03     4.234000e+03     4.234000e+03
mean       5.189081e+02     1.263424e+03     6.683260e+04     1.270755e+05
std        2.792340e+04     4.689801e+04     2.674774e+06     4.139731e+06
min        0.000000e+00     0.000000e+00     0.000000e+00     0.000000e+00
25%        3.225807e-02     3.030303e-02     3.125000e-02     1.935484e+00
50%        6.559140e-01     9.677419e-02     1.000000e-01     6.026316e+00
75%        8.500000e+00     2.903226e-01     2.325000e+01     4.167742e+01
max        1.816375e+06     2.316875e+06     1.556534e+08     1.707531e+08

       reverse_bps_min  reverse_duration  reverse_size_packets  \
count     4.234000e+03       4234.000000          4.234000e+03
mean      6.747949e+04       3224.000000          2.750047e+05
std       3.034402e+06      20429.627234          1.519335e+06
min       0.000000e+00          0.000000          0.000000e+00
25%       2.242424e+00          5.000000          2.000000e+00
50%       9.258333e+00         30.000000          1.800000e+01
75%       4.900000e+01         60.000000          6.860000e+02
max       1.488506e+08     232137.000000          1.717689e+07

       reverse_size_bytes
count        4.234000e+03
mean         2.592156e+05
std          2.875554e+06
min          0.000000e+00
25%          0.000000e+00
50%          0.000000e+00
75%          2.460000e+02
max          1.214242e+08

[8 rows x 60 columns]
```

In [16]:

```
## Now that we have a better view of the data, we can manipulate it.
## The data is already nicely formatted so not many operations are necessary.

print(data["category"].unique())


data_category = data.groupby("category")[["tp_dst", "forward_size_packets", "revers
data_id_src = data.groupby("nw_src")[["tp_dst", "forward_size_packets", "reverse_si
data_id_dst = data.groupby("nw_dst")[["tp_dst", "forward_size_packets", "reverse_si


print("\n\n\n###### per category : ")
print(data_category.head())

print("\n\n\n###### per source : ")
print(data_id_src.head())

print("\n\n\n###### per destination : ")
print(data_id_dst.head())
```

```
print("\n\n\n###### stats per category : ")
print(data_category.describe())

print("\n\n\n###### stats per source : ")
print(data_id_src.describe())

print("\n\n\n###### stats per destination : ")
print(data_id_dst.describe())
```

```
['WWW' 'DNS' 'VOIP' 'ICMP' 'FTP' 'P2P']
```

```
###### per category :
  category    tp_dst  forward_size_packets  reverse_size_packets
0      DNS     75713             296924721               9044166
1      FTP    208803           16731539682             382059857
2     ICMP    625608            6164534717             176377710
3      P2P  31405590             263289124               6465986
4     VOIP   1527226            3399064747              82085318
```

```
###### per source :
        nw_src  tp_dst  forward_size_packets  reverse_size_packets
0  1.01136E+11   38705                     0                     3
1  1.02129E+11   50275                     0                     1
2  1.02134E+11   60309                     0                     2
3  1.02165E+11   37945                     0                     1
4  1.02222E+11  145385                     0                     5
```

```
###### per destination :
        nw_dst  tp_dst  forward_size_packets  reverse_size_packets
0  1.04198E+11     886                    11                     8
1  1.04237E+11     443                   145                   180
2  1.07155E+11     160                     4                     0
3  1.14141E+11       0                  5586                     0
4  1.30225E+11       0                  6083                  6084
```

```
###### stats per category :
             tp_dst  forward_size_packets  reverse_size_packets
count  6.000000e+00          6.000000e+00          6.000000e+00
mean   6.026426e+06          7.074529e+09          1.940617e+08
std    1.246226e+07          7.381980e+09          2.080323e+08
min    7.571300e+04          2.632891e+08          6.465986e+06
25%    3.130042e+05          1.072460e+09          2.730445e+07
50%    1.076417e+06          4.781800e+09          1.292315e+08
75%    2.118519e+06          1.323500e+10          3.306393e+08
max    3.140559e+07          1.673154e+10          5.083369e+08
```

```
###### stats per source :
              tp_dst  forward_size_packets  reverse_size_packets
count     483.000000          4.830000e+02          4.830000e+02
mean    74862.436853          8.788235e+07          2.410704e+06
std     83313.921056          1.393963e+09          3.186155e+07
min     32783.000000          0.000000e+00          0.000000e+00
25%     43980.000000          0.000000e+00          1.000000e+00
50%     53723.000000          0.000000e+00          2.000000e+00
75%     85348.000000          4.000000e+00          4.000000e+00
```

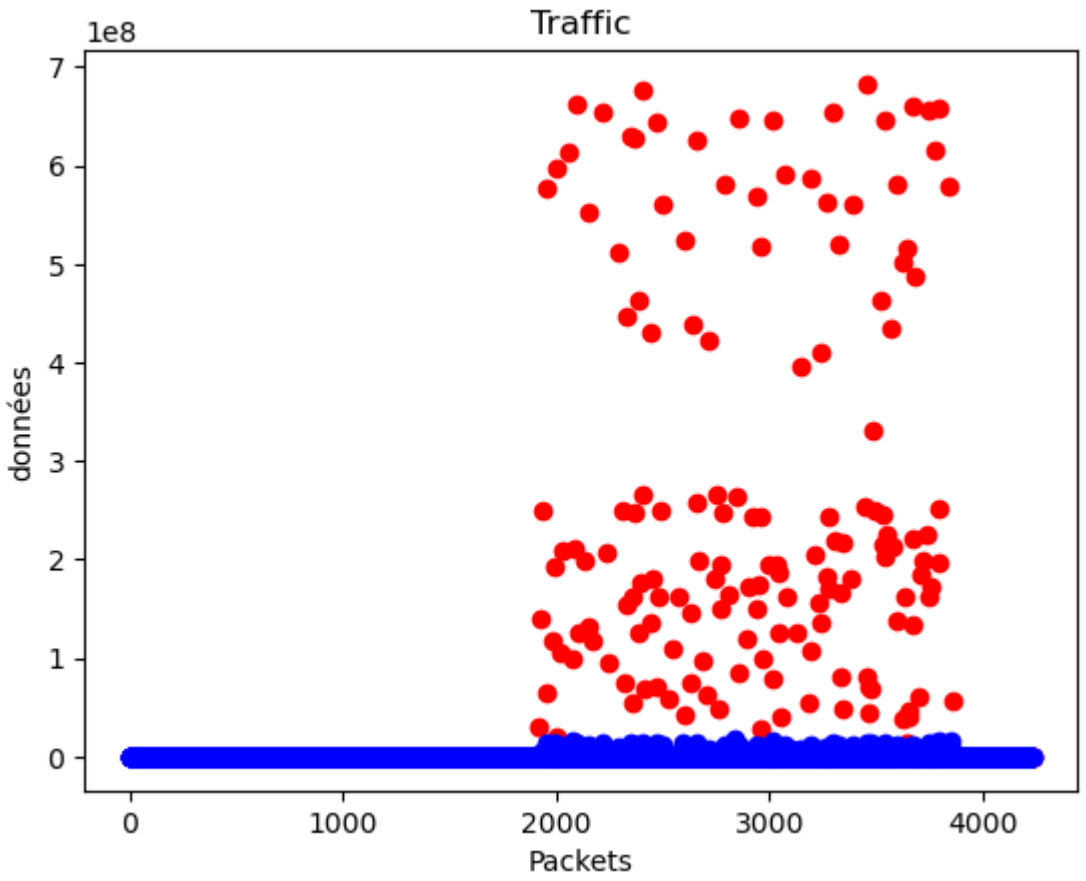
```
max    940337.000000          2.574205e+10          5.463225e+08


###### stats per destination :
            tp_dst  forward_size_packets  reverse_size_packets
count  8.800000e+02          8.800000e+02          8.800000e+02
mean   4.108927e+04          4.823543e+07          1.323148e+06
std    1.101733e+06          1.427903e+09          3.013184e+07
min    0.000000e+00          0.000000e+00          0.000000e+00
25%    4.430000e+02          2.300000e+01          1.400000e+01
50%    4.430000e+02          3.845000e+02          2.725000e+02
75%    8.860000e+02          4.208000e+03          5.134250e+03
max    3.265328e+07          4.235854e+10          8.923850e+08
```

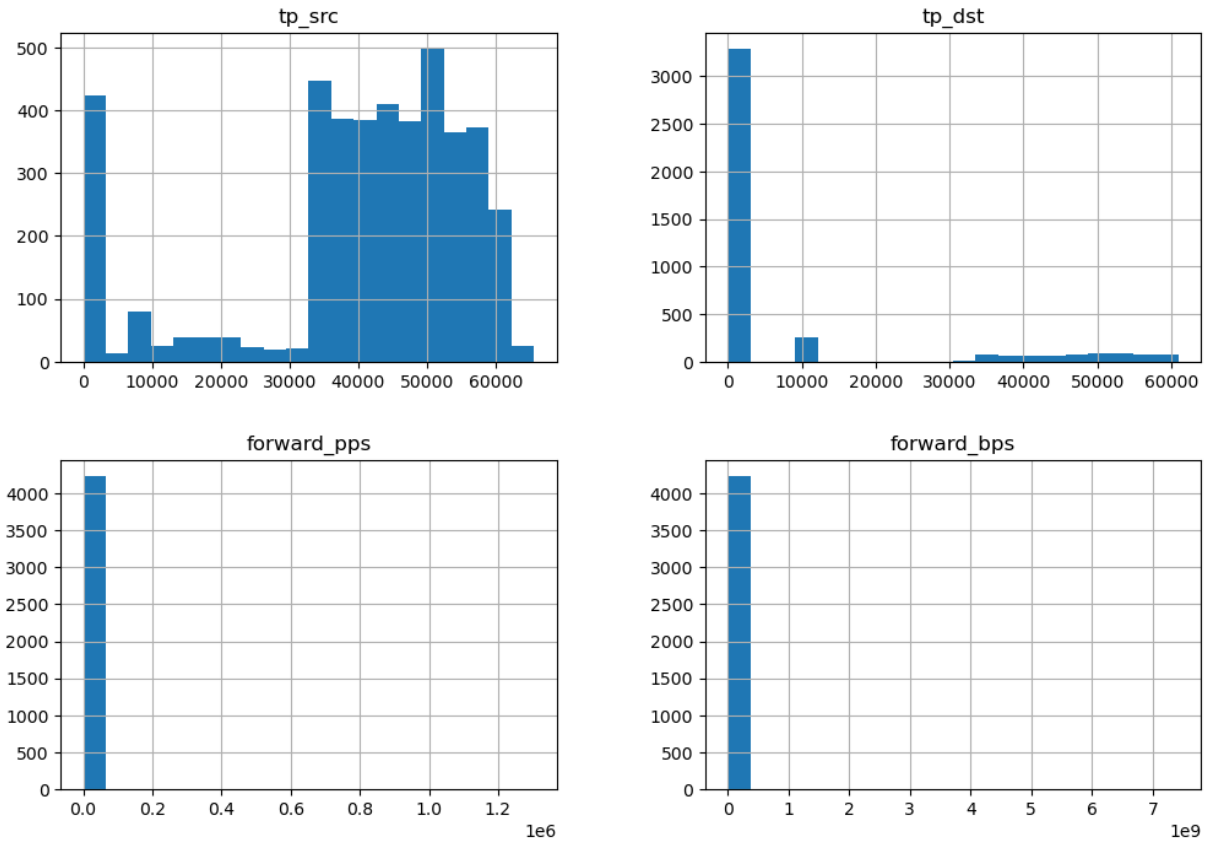
In [27]:
```
#Some matplotlib graphs:

plt.scatter(np.arange(0, len(data)), data["forward_size_packets"], color = "r", lab
plt.scatter(np.arange(0, len(data)), data["reverse_size_packets"], color = "b", lab
plt.title("Traffic")
plt.xlabel("Packets")
plt.ylabel("données")
```

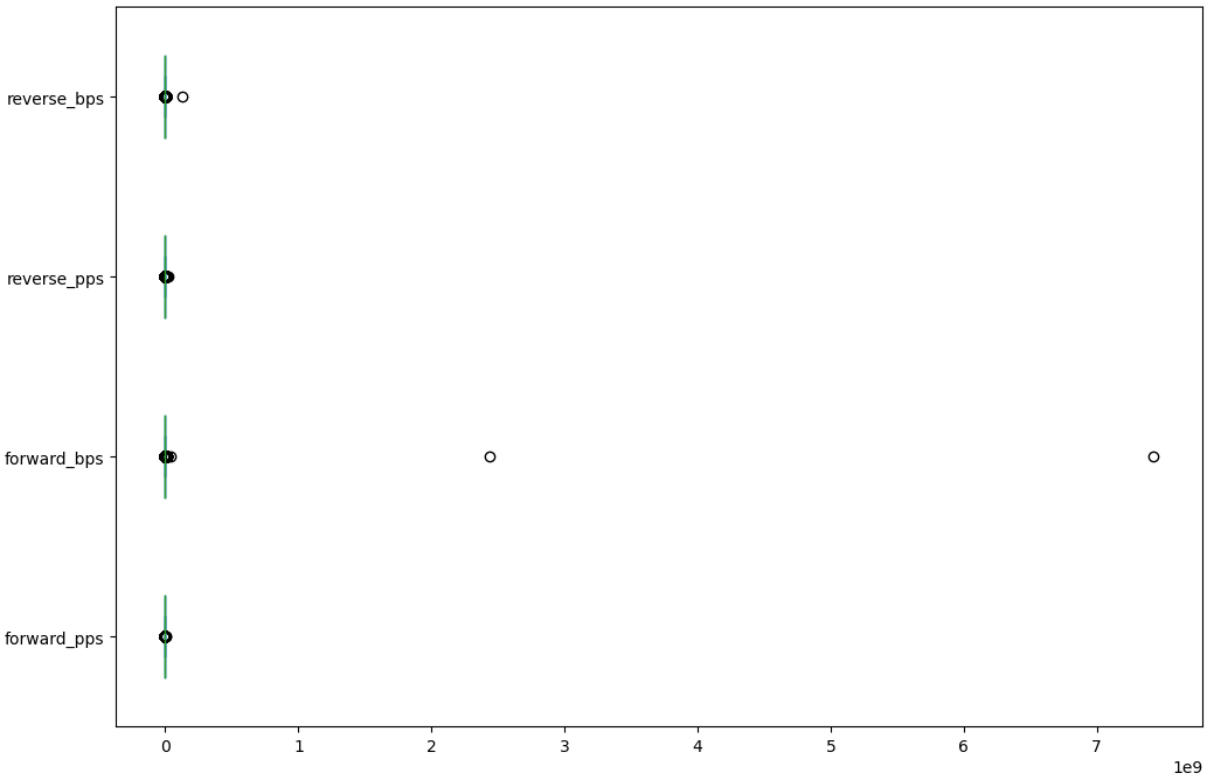
Out[27]:
```
Text(0, 0.5, 'données')
```



In [19]:
```
data.hist(column=['tp_src', 'tp_dst', 'forward_pps', 'forward_bps'], bins=20, figsi
plt.show()
```

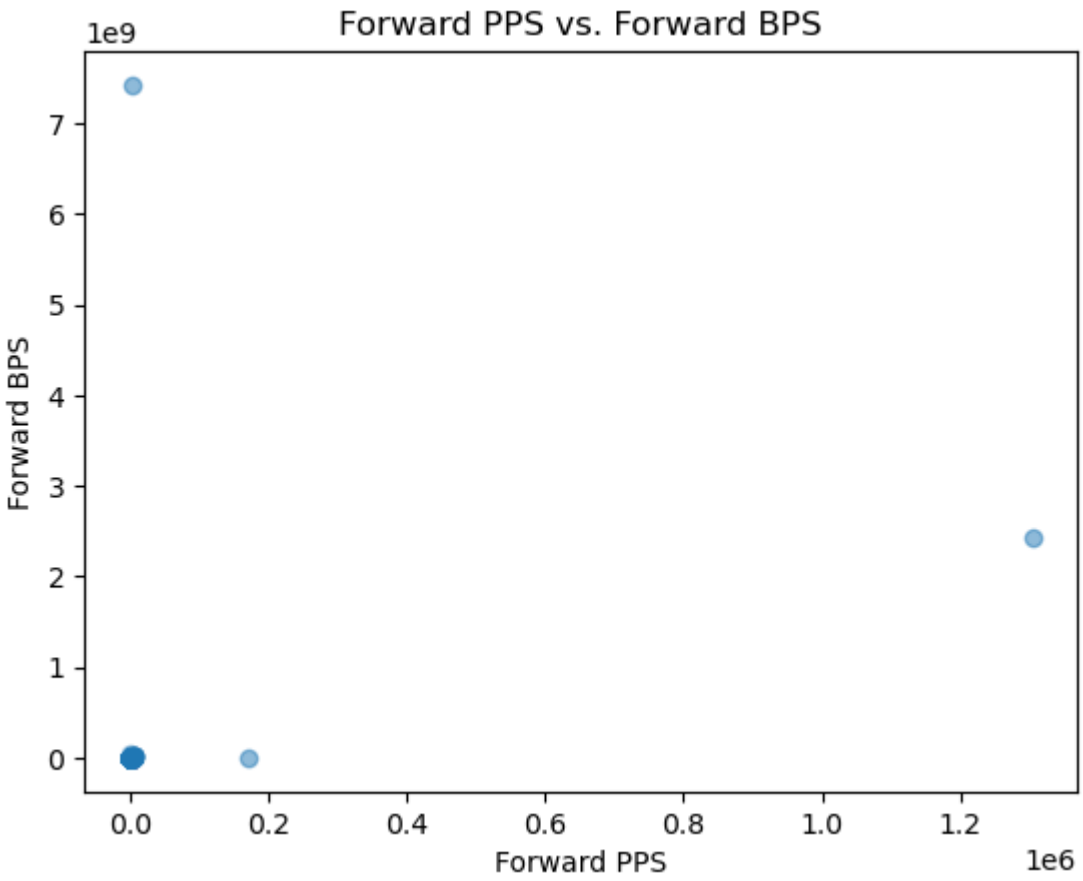


In [20]:
```
data[['forward_pps', 'forward_bps', 'reverse_pps', 'reverse_bps']].plot(kind='box',
plt.show()
```

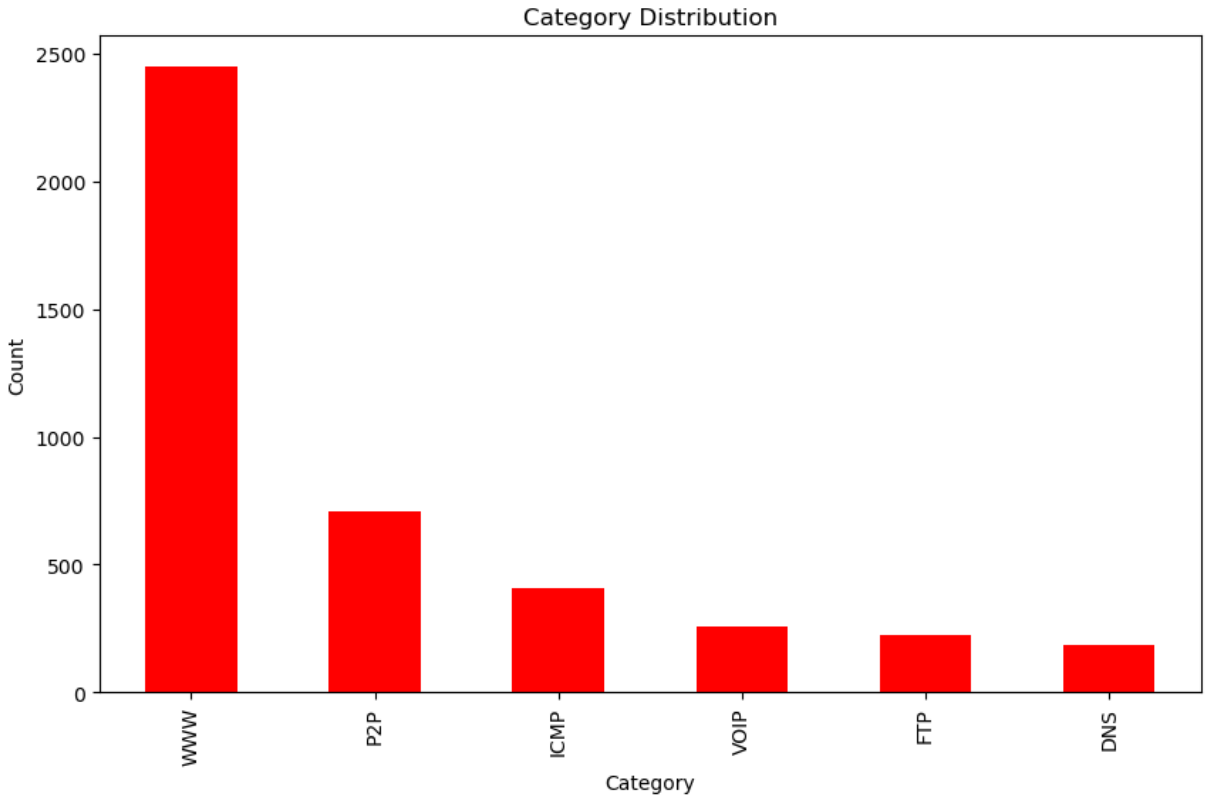


In [21]:
```
plt.scatter(data['forward_pps'], data['forward_bps'], alpha=0.5)
plt.title('Forward PPS vs. Forward BPS')
plt.xlabel('Forward PPS')
```

```
plt.ylabel('Forward BPS')
plt.show()
```



In [26]:
```
category_counts = data['category'].value_counts()
category_counts.plot(kind='bar', figsize=(10, 6), color='red')
plt.title('Category Distribution')
plt.xlabel('Category')
plt.ylabel('Count')
plt.show()
```


Category Distribution
2500
2000
1500
1000
500
0
Count
WWW
P2P
ICMP
VOIP
FTP
DNS
Category

In [ ]: