ENGS 108 Assignment 1

Cameron Wolfe 9/15/23

```
import pandas as pd
data = pd.read_csv('Flows.csv')
```

Questions 1 and 2

```
# Use shape to get number of records and number of fields
shape = data.shape
rows = shape[0]
columns = shape[1]

print("".join(['Number of records: ', str(rows)]))
print("".join(['\nNumber of fields: ', str(columns), '\n']))
```

```
Number of records: 73126

Number of fields: 50
```

Questions 3 and 4

```
# Extract source and destination IP addresses from dataframe
src_ip_addresses = data['src_ip']
dst_ip_addresses = data['dst_ip']

# Create dictionaries to store number of times an IP address has occured
src_dict = {}
dst_dict = {}

# Go through each row
for i in range(rows):
    # Grab IP in that row
    src_ip = src_ip_addresses[i]
    dst_ip = dst_ip_addresses[i]

    # If IP is in dictionary, increment count by 1, otherwise this is the first time it has been seen, so set the count to 1
    if src_ip in src_dict:
        src_dict[src_ip] += 1
    else:
        src_dict[src_ip] = 1

    # Same as above but with destination IPs as opposed to source IPs
    if dst_ip in dst_dict:
        dst_dict[dst_ip] += 1
    else:
        dst_dict[dst_ip] = 1

# Sort the keys in the dictionary in reverse order by the number of times it appears
sorted_src_keys = sorted(src_dict, key=lambda x: src_dict[x], reverse=True)
sorted_dst_keys = sorted(dst_dict, key=lambda x: dst_dict[x], reverse=True)

# Print first 10 IPs for each
print('Source IPs with most records')
for key in sorted_src_keys[:10]:
    print(key, ': ', src_dict[key])
print('\nDestination IPs with most records')
for key in sorted_dst_keys[:10]:
    print(key, ': ', dst_dict[key])
```

```
Source IPs with most records
192.168.121.62 :  7242
192.168.122.7 :  4070
192.168.121.67 :  3383
192.168.121.24 :  3040
192.168.122.46 :  2897
192.168.122.52 :  2240
192.168.122.51 :  2093
192.168.122.4 :  2087
192.168.121.27 :  1985
192.168.122.34 :  1877

Destination IPs with most records
172.16.255.200 :  23198
172.16.141.250 :  5813
172.16.255.183 :  5807
10.200.7.6 :  2159
10.200.7.5 :  1435
10.200.7.8 :  1413
216.58.192.46 :  866
```

```
10.200.7.7 :  853
172.217.2.67 :  689
172.217.8.78 :  669
```

Questions 5 and 6

```
# Extract flow duration and size from dataframe
flow_durations = data['flowDuration']
flow_sizes = data['octetTotalCount']

# Find max flow duration as well as IPs associated with that flow
max_duration = flow_durations.max()
max_duration_index = flow_durations.idxmax()
max_duration_src = src_ip_addresses[max_duration_index]
max_duration_dst = dst_ip_addresses[max_duration_index]

# Find max flow size as well as IPs associated with that flow
max_size = flow_sizes.max()
max_size_index = flow_sizes.idxmax()
max_size_src = src_ip_addresses[max_size_index]
max_size_dst = dst_ip_addresses[max_size_index]

print("".join(['\nLongest duration flow:\n', max_duration_src, ' to ', max_duration_dst, ': ', str(max_duration), ' seconds\n']))
print("".join(['Largest flow:\n', max_size_src, ' to ', max_size_dst, ': ', str(max_size), ' bytes']))
```

```
Longest duration flow:
192.168.121.62 to 192.168.190.13: 1799.50165295601 seconds

Largest flow:
192.168.122.53 to 10.200.7.4: 288154316 bytes
```

✓ 0s completed at 10:09 AM ● ×