Relatório - Análise Inicial dos Dados

Introdução e Motivação

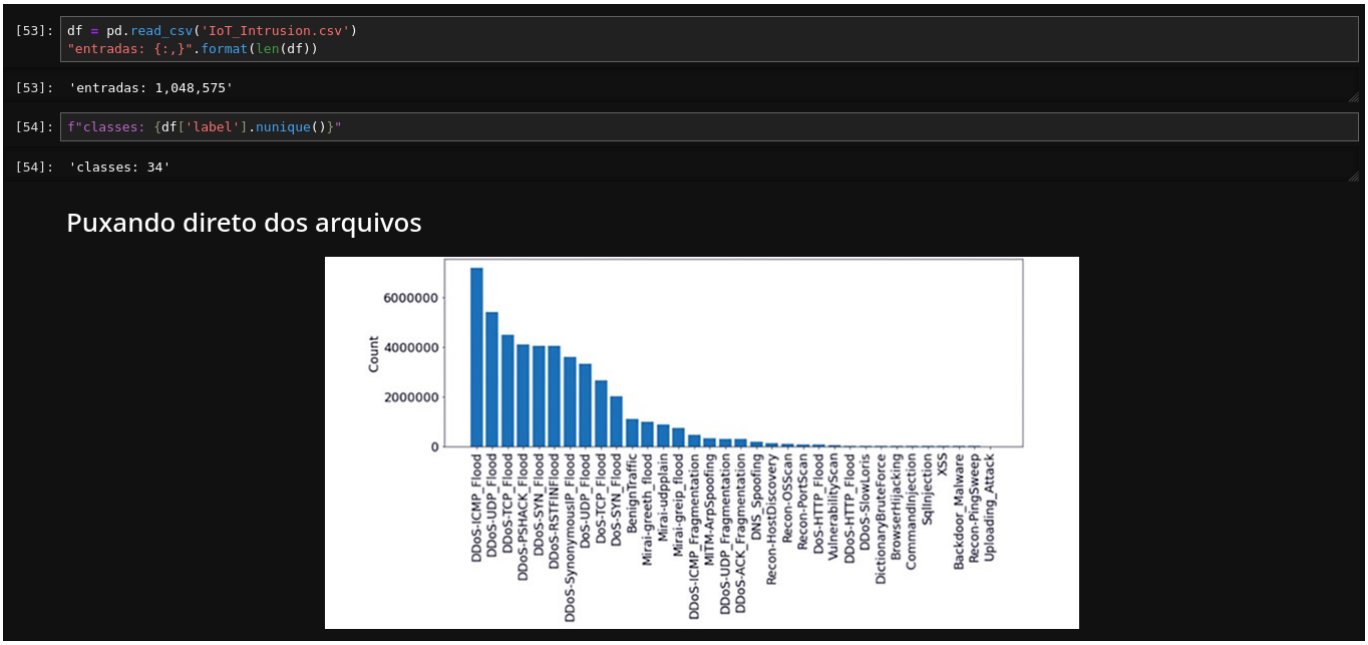
Uma grande dificuldade que eu tinha com o dataset que irei usar vinha do fato de eu *nunca ter conseguido baixá-lo integralmente*, logo minha motivação inicial para essa análise/exploração foi buscar obtê-lo por conta própria, ou mesmo de algum lugar confiável (como o Kaggle, por exemplo).

Dataset Online

Fazendo uma busca por "CICIOT2023" no Kaggle, consegui encontrar poucos resultados, surpreendentemente. Entretanto, consegui *1* arquivo, que no início parecia promissor: um `csv` único, contendo (supostamente) todos os dados - e ainda com rótulos!

- Link do Kaggle: https://www.kaggle.com/datasets/subhajournal/iotintrusion

Infelizmente, ao carregar os dados logo percebi que haviam poucos fluxos, o que não era condizente com a própria página do CIC:



Sendo assim, fui atrás de reproduzir integralmente o que os autores originais haviam disponibilizado.

Dataset Original

No passado eu já tentei usar de *scripts* para tentar baixar o *directory listing* do CIC - o que não deu certo. Porém, dessa vez, pesquisei por um outro método e felizmente encontrei um programa que foi chamado de "*swiss army knife of downloading*".

- Link do CIC: http://205.174.165.80/IOTDataset/CIC_IOT_Dataset2023/Dataset/
- Comando usado: `lftp -c 'mirror --parallel=100 --use-pget-n=10 --exclude-glob *.pcap --exclude PCAP/* <link-d> ;exit'`

Importante notar que eu explicitamente excluí os arquivos `pcap` do download, já que, se fosse incluí-los, o tamanho explodiria e nem mesmo caberia no meu HD.

```
λ rclone size --http-url
http://205.174.165.80/IOTDataset/CIC_IOT_Dataset2023/Dataset/ :http:
Total objects: 635
Total size: 579.928 GiB (622693003083 Byte)
```

Deixando eles de fora, consegui iniciar meus primeiros esforços de *entender a disposição de todos esses tráfegos*.

```
jp in ~/Documents/IC λ ee CICIoT2023
drwxr-xr-x    - jp 20 ago 08:23 CSV
drwxr-xr-x    - jp  6 fev 07:23 example
drwxr-xr-x    - jp 20 ago 08:31 'Supplementary Materials'
.rw-r--r-- 3,8M jp 20 ago 11:11 README.pdf
jp in ~/Documents/IC λ du --human-readable --total --summarize CICIoT2023/
8,4G CICIoT2023/
8,4G total
```

Também há a vantagem de que as classes (ataques e tráfego normal) já estão separados devidamente por pastas:

```
jp in ~/Documents/IC λ tree CICIoT2023/CSV | tail -n1
35 directories, 297 files
```

Contagem das classes

Depois de (re)lembrar como lidar com o Python e o Jupyter no geral, obtive a especificação de quantos fluxos (linhas nos `csv`'s) haviam para cada tipo de ataque, e o resultado me surpreendeu um pouco:

```
[60]: class_sizes_series = pd.Series(class_sizes, index=class_names)
      class_sizes_series_sorted = class_sizes_series.sort_values()
      class_sizes_series_sorted_formatted = class_sizes_series_sorted.apply(lambda x: f"{x:,}")

[60]: DDoS-ACK_Fragmentation             0
      Uploading_Attack               1,252
      Recon-PingSweep                2,262
      Backdoor_Malware               3,218
      XSS                            3,846
      SqlInjection                   5,245
      CommandInjection               5,409
      BrowserHijacking               5,859
      DictionaryBruteForce          13,064
      DDoS-SlowLoris                23,426
      DDoS-HTTP_Flood               28,790
      DoS-HTTP_Flood                71,861
      Recon-PortScan                82,284
      Recon-OSScan                  98,259
      Recon-HostDiscovery          134,378
      DNS_Spoofing                 178,898
      DDoS-UDP_Fragmentation       286,925
      MITM-ArpSpoofing             307,560
      VulnerabilityScan            373,351
      DDoS-ICMP_Fragmentation      452,490
      Mirai-greip_flood            751,646
      Mirai-udpplain               890,574
      Mirai-greeth_flood           991,834
      Benign_Final               1,098,191
      DoS-SYN_Flood              2,028,836
      DoS-TCP_Flood              2,671,430
      DoS-UDP_Flood              3,318,634
      DDoS-SynonymousIP_Flood    3,598,133
      DDoS-RSTFINFLOOD           4,045,279
      DDoS-SYN_Flood             4,059,179
      DDoS-PSHACK_FLOOD          4,094,772
      DDoS-TCP_Flood             4,497,649
      DDoS-UDP_Flood             5,412,231
      DDoS-ICMP_Flood            7,200,501
      dtype: object
```

Zero entradas!? Fui investigar diretamente no link do CIC.

Zero entradas?

Na realidade, para o DDoS-ACK_Fragmentation, não foram disponibilizados dados tabulares, somente os fluxos brutos:

Index of /IOTDataset/CIC_IOT_Dataset2023/Dataset/CSV/DDoS-ACK_Fragmentation

Name	Last modified	Size	Description
Parent Directory		-	
DDoS-ACK_Fragmentation.pcap	2023-03-31 12:06	1.9G	
DDoS-ACK_Fragmentation1.pcap	2023-03-31 12:07	1.9G	
DDoS-ACK_Fragmentation2.pcap	2023-03-31 12:07	1.9G	
DDoS-ACK_Fragmentation3.pcap	2023-03-31 12:07	1.9G	
DDoS-ACK_Fragmentation4.pcap	2023-03-31 12:08	1.9G	
DDoS-ACK_Fragmentation5.pcap	2023-03-31 12:09	1.9G	
DDoS-ACK_Fragmentation6.pcap	2023-03-31 12:09	1.9G	
DDoS-ACK_Fragmentation7.pcap	2023-03-31 12:10	1.9G	
DDoS-ACK_Fragmentation8.pcap	2023-03-31 12:10	1.9G	
DDoS-ACK_Fragmentation9.pcap	2023-03-31 12:11	1.9G	
DDoS-ACK_Fragmentation10.pcap	2023-03-31 12:11	1.9G	
DDoS-ACK_Fragmentation11.pcap	2023-03-31 12:12	1.9G	
DDoS-ACK_Fragmentation12.pcap	2023-03-31 12:12	816M	

Apache/2.4.41 (Ubuntu) Server at 205.174.165.80 Port 80

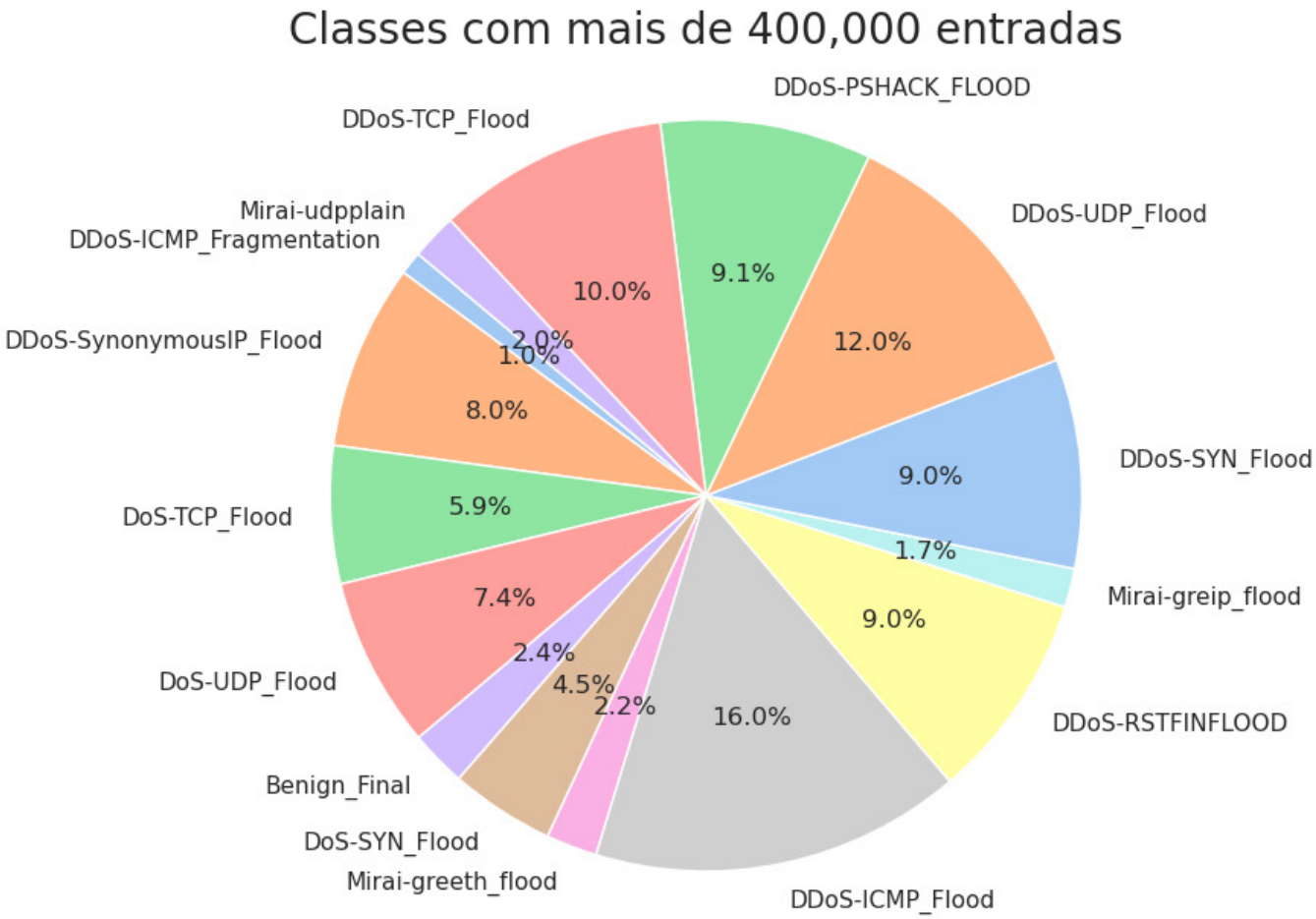
```
[61]: "tamanho total: {:.1f}GB".format(1.9 * 12 + 0.816)

[61]: 'tamanho total: 23.6GB'
```

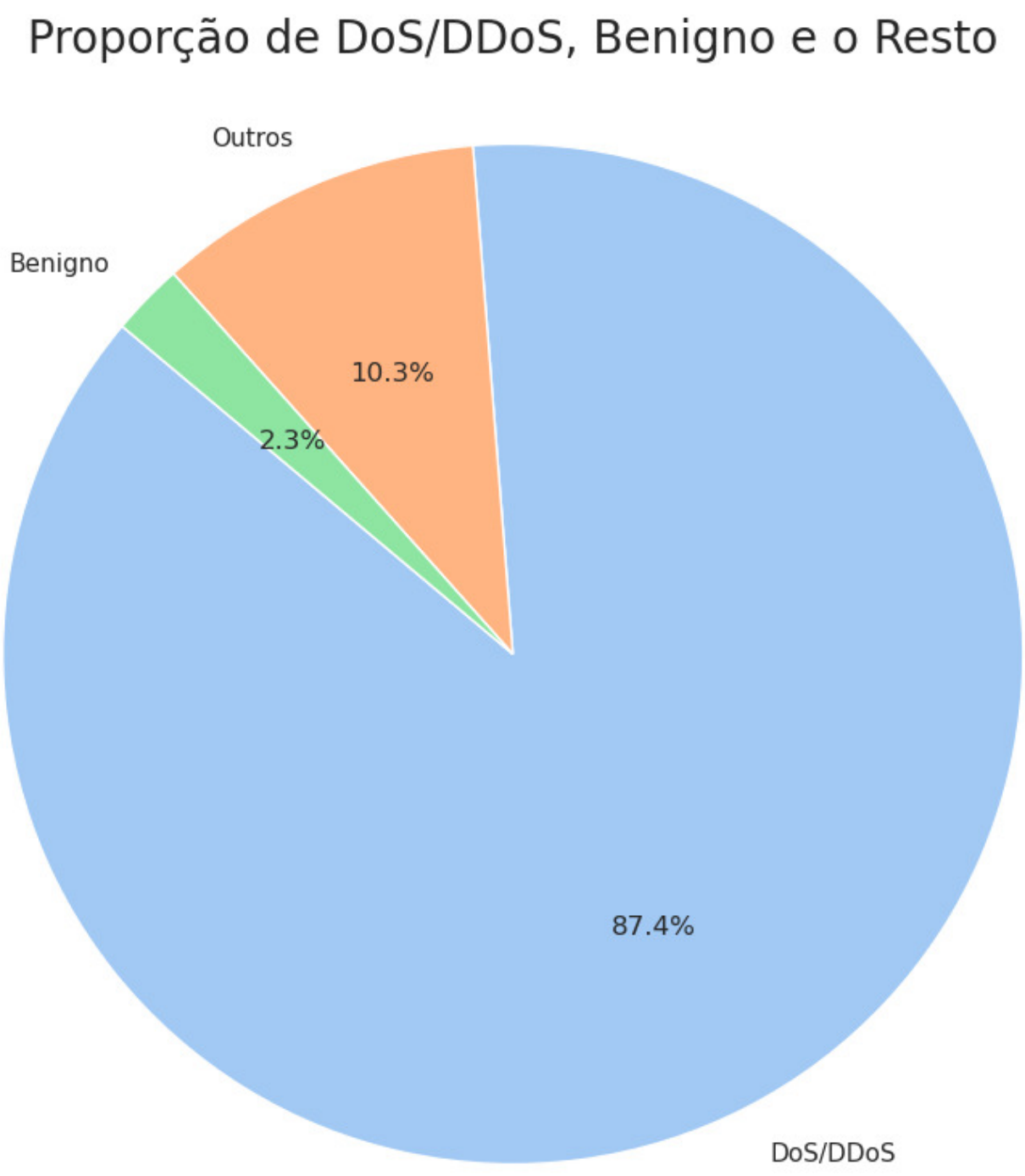
Como não irei utilizar dos `pcap`'s por agora, não busquei meios de baixar esses 23.6GB.

Distribuição das classes

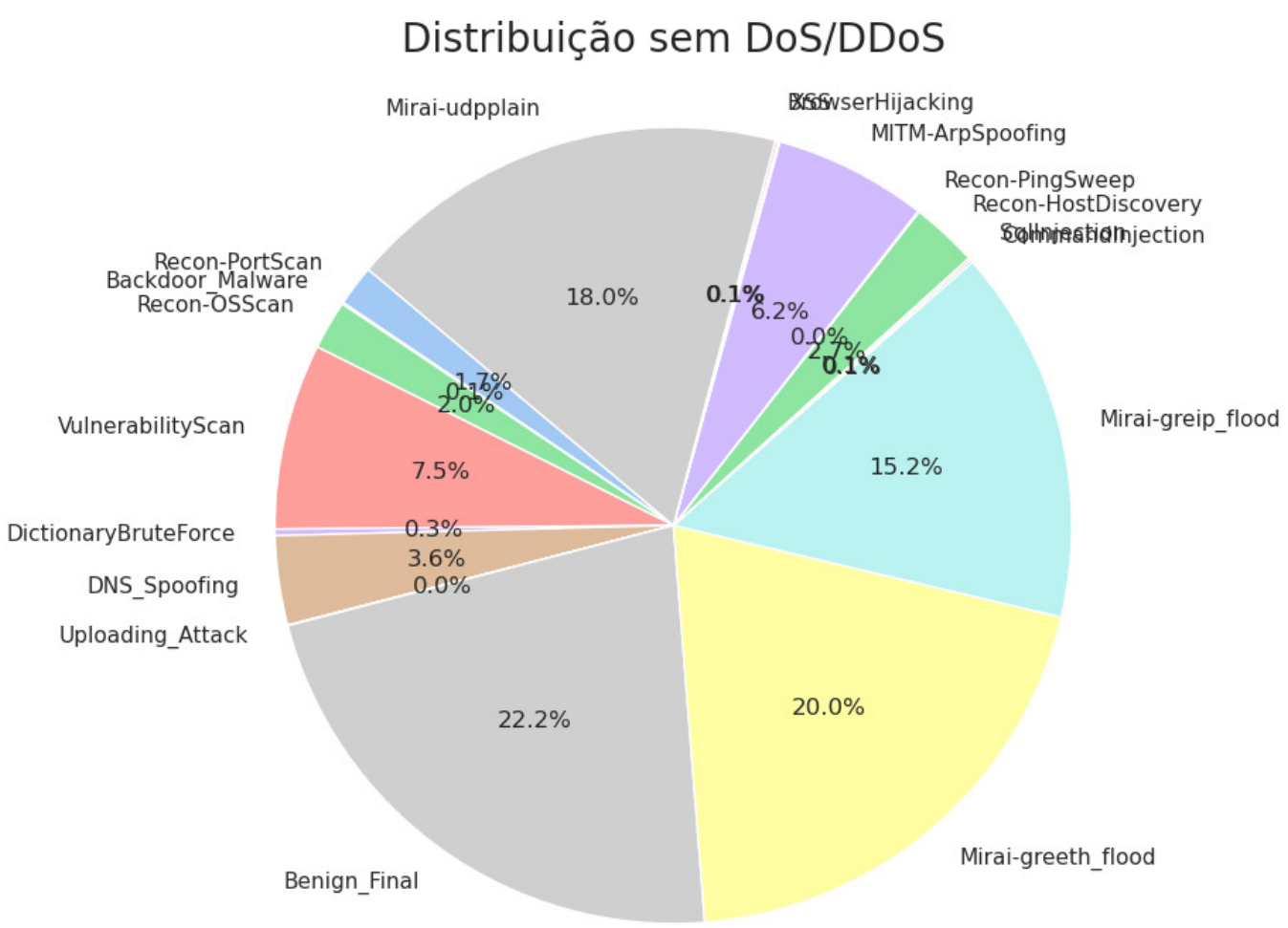
Finalmente, os resultados: li *cada um* dos diretórios contendo os dados e plotei alguns gráficos de pizza para entender visualmente a distribuição que encontrei acima:



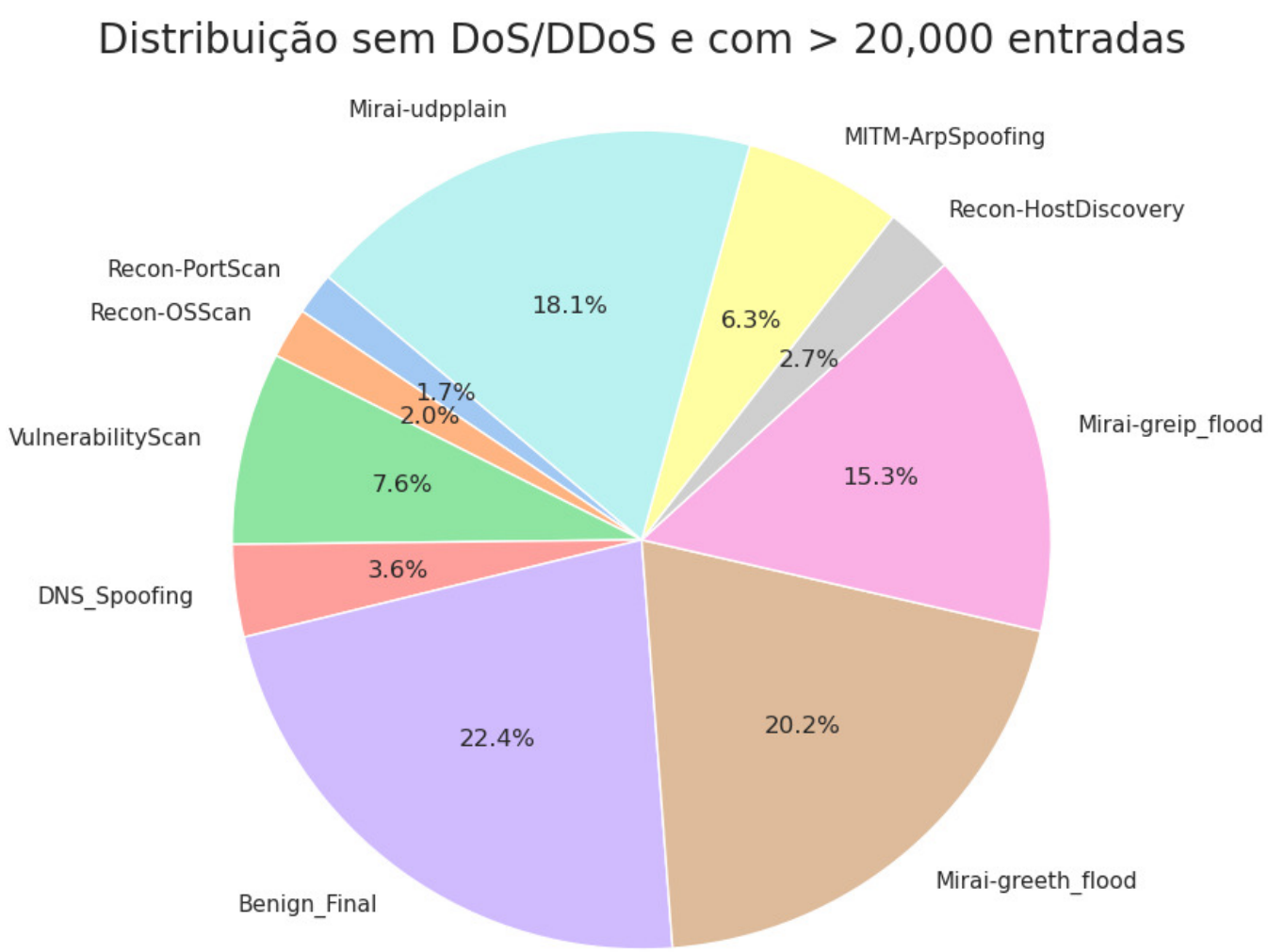
Olhando cuidadosamente, notei que a maioria das classes com maior proporção eram de negação de serviço, assim tive a ideia de ver qual espaço elas tomavam para o resto:



Tirando elas de cena, temos um *dataset* bem mais diverso:



Ainda assim, o gráfico fica muito amarrotado, então fazendo um filtro de quantidade de fluxos deixa ele um pouco mais palatável:



Um fato que me atentei, após essas análises, é que mesmo com a quantidade pequena de 20,000 entradas, a quantidade de classes já caiu de 34 (33 ataques + 1 benigno) para 10 (9 ataques + 1 benigno). Ainda, os representantes mais significativos de ataque todos eram do tipo "Mirai".