In [1]:
```
pip install kagglehub
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: kagglehub in c:\users\tipqc\appdata\roaming\python\py
thon312\site-packages (0.3.11)
Requirement already satisfied: packaging in c:\programdata\anaconda3\lib\site-packag
es (from kagglehub) (23.2)
Requirement already satisfied: pyyaml in c:\programdata\anaconda3\lib\site-packages
(from kagglehub) (6.0.1)
Requirement already satisfied: requests in c:\programdata\anaconda3\lib\site-package
s (from kagglehub) (2.32.2)
Requirement already satisfied: tqdm in c:\programdata\anaconda3\lib\site-packages (f
rom kagglehub) (4.66.4)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\programdata\anaconda3
\lib\site-packages (from requests->kagglehub) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\programdata\anaconda3\lib\site-pac
kages (from requests->kagglehub) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\programdata\anaconda3\lib\si
te-packages (from requests->kagglehub) (2.2.2)
Requirement already satisfied: certifi>=2017.4.17 in c:\programdata\anaconda3\lib\si
te-packages (from requests->kagglehub) (2024.6.2)
Requirement already satisfied: colorama in c:\programdata\anaconda3\lib\site-package
s (from tqdm->kagglehub) (0.4.6)
Note: you may need to restart the kernel to use updated packages.
```

Extract the provided dataset using FLAT FILE.

In [2]:
```
import kagglehub
```

In [3]:
```
# You get extra points for loading it through Kaggle API.
# Download latest version
path = kagglehub.dataset_download("supplejade/rt-iot2022real-time-internet-of-thing

print("Path to dataset files:", path)
```

```
Path to dataset files: C:\Users\tipqc\.cache\kagglehub\datasets\supplejade\rt-iot202
2real-time-internet-of-things\versions\3
```

In [4]:
```
import pandas as pd

rt_iot = pd.read_csv(path + '/RT_IOT2022.csv')
```

In [6]:
```
rt_iot.head()
```

Out[6]:

	no	id.orig_p	id.resp_p	proto	service	flow_duration	fwd_pkts_tot	bwd_pkts_tot	fwd
0	0	38667	1883	tcp	mqtt	32.011598	9	5	
1	1	51143	1883	tcp	mqtt	31.883584	9	5	
2	2	44761	1883	tcp	mqtt	32.124053	9	5	
3	3	60893	1883	tcp	mqtt	31.961063	9	5	
4	4	51087	1883	tcp	mqtt	31.902362	9	5	

5 rows × 85 columns

Transform the dataset

In [12]:
```
# I noticed the columns are many, so I think I can melt the data
# for it to be more readable
rt_iot.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 123117 entries, 0 to 123116
Data columns (total 85 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   no                        123117 non-null  int64
 1   id.orig_p                 123117 non-null  int64
 2   id.resp_p                 123117 non-null  int64
 3   proto                     123117 non-null  object
 4   service                   123117 non-null  object
 5   flow_duration             123117 non-null  float64
 6   fwd_pkts_tot              123117 non-null  int64
 7   bwd_pkts_tot              123117 non-null  int64
 8   fwd_data_pkts_tot         123117 non-null  int64
 9   bwd_data_pkts_tot         123117 non-null  int64
 10  fwd_pkts_per_sec          123117 non-null  float64
 11  bwd_pkts_per_sec          123117 non-null  float64
 12  flow_pkts_per_sec         123117 non-null  float64
 13  down_up_ratio             123117 non-null  float64
 14  fwd_header_size_tot       123117 non-null  int64
 15  fwd_header_size_min       123117 non-null  int64
 16  fwd_header_size_max       123117 non-null  int64
 17  bwd_header_size_tot       123117 non-null  int64
 18  bwd_header_size_min       123117 non-null  int64
 19  bwd_header_size_max       123117 non-null  int64
 20  flow_FIN_flag_count       123117 non-null  int64
 21  flow_SYN_flag_count       123117 non-null  int64
 22  flow_RST_flag_count       123117 non-null  int64
 23  fwd_PSH_flag_count        123117 non-null  int64
 24  bwd_PSH_flag_count        123117 non-null  int64
 25  flow_ACK_flag_count       123117 non-null  int64
 26  fwd_URG_flag_count        123117 non-null  int64
 27  bwd_URG_flag_count        123117 non-null  int64
 28  flow_CWR_flag_count       123117 non-null  int64
 29  flow_ECE_flag_count       123117 non-null  int64
 30  fwd_pkts_payload.min      123117 non-null  float64
 31  fwd_pkts_payload.max      123117 non-null  float64
 32  fwd_pkts_payload.tot      123117 non-null  float64
 33  fwd_pkts_payload.avg      123117 non-null  float64
 34  fwd_pkts_payload.std      123117 non-null  float64
 35  bwd_pkts_payload.min      123117 non-null  float64
 36  bwd_pkts_payload.max      123117 non-null  float64
 37  bwd_pkts_payload.tot      123117 non-null  float64
 38  bwd_pkts_payload.avg      123117 non-null  float64
 39  bwd_pkts_payload.std      123117 non-null  float64
 40  flow_pkts_payload.min     123117 non-null  float64
 41  flow_pkts_payload.max     123117 non-null  float64
 42  flow_pkts_payload.tot     123117 non-null  float64
 43  flow_pkts_payload.avg     123117 non-null  float64
 44  flow_pkts_payload.std     123117 non-null  float64
 45  fwd_iat.min               123117 non-null  float64
 46  fwd_iat.max               123117 non-null  float64
 47  fwd_iat.tot               123117 non-null  float64
 48  fwd_iat.avg               123117 non-null  float64
 49  fwd_iat.std               123117 non-null  float64
 50  bwd_iat.min               123117 non-null  float64
```

```
 51  bwd_iat.max               123117 non-null  float64
 52  bwd_iat.tot               123117 non-null  float64
 53  bwd_iat.avg               123117 non-null  float64
 54  bwd_iat.std               123117 non-null  float64
 55  flow_iat.min              123117 non-null  float64
 56  flow_iat.max              123117 non-null  float64
 57  flow_iat.tot              123117 non-null  float64
 58  flow_iat.avg              123117 non-null  float64
 59  flow_iat.std              123117 non-null  float64
 60  payload_bytes_per_second  123117 non-null  float64
 61  fwd_subflow_pkts          123117 non-null  float64
 62  bwd_subflow_pkts          123117 non-null  float64
 63  fwd_subflow_bytes         123117 non-null  float64
 64  bwd_subflow_bytes         123117 non-null  float64
 65  fwd_bulk_bytes            123117 non-null  float64
 66  bwd_bulk_bytes            123117 non-null  float64
 67  fwd_bulk_packets          123117 non-null  float64
 68  bwd_bulk_packets          123117 non-null  float64
 69  fwd_bulk_rate             123117 non-null  float64
 70  bwd_bulk_rate             123117 non-null  float64
 71  active.min                123117 non-null  float64
 72  active.max                123117 non-null  float64
 73  active.tot                123117 non-null  float64
 74  active.avg                123117 non-null  float64
 75  active.std                123117 non-null  float64
 76  idle.min                  123117 non-null  float64
 77  idle.max                  123117 non-null  float64
 78  idle.tot                  123117 non-null  float64
 79  idle.avg                  123117 non-null  float64
 80  idle.std                  123117 non-null  float64
 81  fwd_init_window_size      123117 non-null  int64
 82  bwd_init_window_size      123117 non-null  int64
 83  fwd_last_window_size      123117 non-null  int64
 84  Attack_type               123117 non-null  object
dtypes: float64(56), int64(26), object(3)
memory usage: 79.8+ MB
```

In [24]:

```
# Set 'no' column as index
test = rt_iot.set_index('no')
test
```

Out[24]:

no	id.orig_p	id.resp_p	proto	service	flow_duration	fwd_pkts_tot	bwd_pkts_tot	fwd_
0	38667	1883	tcp	mqtt	32.011598	9	5	
1	51143	1883	tcp	mqtt	31.883584	9	5	
2	44761	1883	tcp	mqtt	32.124053	9	5	
3	60893	1883	tcp	mqtt	31.961063	9	5	
4	51087	1883	tcp	mqtt	31.902362	9	5	
...	...	...	...	...	...	...	...	
2005	59247	63331	tcp	-	0.000006	1	1	
2006	59247	64623	tcp	-	0.000007	1	1	
2007	59247	64680	tcp	-	0.000006	1	1	
2008	59247	65000	tcp	-	0.000006	1	1	
2009	59247	65129	tcp	-	0.000006	1	1	

123117 rows × 84 columns

In [ ]:
```
# New column that indicates if MAX, MIN, AVG, STD, TOTAL
```

In [ ]:
```
# s
```

Load

In [ ]:
```
# Convert to csv
rt_iot.to_csv('clean_RT_IoT.csv')
```