For this quiz, you must build an end-to-end data analysis notebook following the ETL pipeline.

1.On a separate document, create a table with 3 columns. Column 1: Phase, Column 2: Activity, and Column 3: Code and Output. -Phase will only have 3 major rows: Extract, Transform, Load -Activity will list down all the activities you performed per phase. -Code and output will correspond to the activity.

In [52]:
```
import pandas as pd

RT = pd.read_csv('RT_IOT2022.csv')
RT.head()
```

Out[52]:

	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

In [9]:
```
RT.size
```

Out[9]:
```
10464945
```

In [10]:
```
RT.columns
```

Out[10]:
```
Index(['no', 'id.orig_p', 'id.resp_p', 'proto', 'service', 'flow_duration',
       'fwd_pkts_tot', 'bwd_pkts_tot', 'fwd_data_pkts_tot',
       'bwd_data_pkts_tot', 'fwd_pkts_per_sec', 'bwd_pkts_per_sec',
       'flow_pkts_per_sec', 'down_up_ratio', 'fwd_header_size_tot',
       'fwd_header_size_min', 'fwd_header_size_max', 'bwd_header_size_tot',
       'bwd_header_size_min', 'bwd_header_size_max', 'flow_FIN_flag_count',
       'flow_SYN_flag_count', 'flow_RST_flag_count', 'fwd_PSH_flag_count',
       'bwd_PSH_flag_count', 'flow_ACK_flag_count', 'fwd_URG_flag_count',
       'bwd_URG_flag_count', 'flow_CWR_flag_count', 'flow_ECE_flag_count',
       'fwd_pkts_payload.min', 'fwd_pkts_payload.max', 'fwd_pkts_payload.tot',
       'fwd_pkts_payload.avg', 'fwd_pkts_payload.std', 'bwd_pkts_payload.min',
       'bwd_pkts_payload.max', 'bwd_pkts_payload.tot', 'bwd_pkts_payload.avg',
       'bwd_pkts_payload.std', 'flow_pkts_payload.min',
       'flow_pkts_payload.max', 'flow_pkts_payload.tot',
       'flow_pkts_payload.avg', 'flow_pkts_payload.std', 'fwd_iat.min',
       'fwd_iat.max', 'fwd_iat.tot', 'fwd_iat.avg', 'fwd_iat.std',
       'bwd_iat.min', 'bwd_iat.max', 'bwd_iat.tot', 'bwd_iat.avg',
       'bwd_iat.std', 'flow_iat.min', 'flow_iat.max', 'flow_iat.tot',
       'flow_iat.avg', 'flow_iat.std', 'payload_bytes_per_second',
       'fwd_subflow_pkts', 'bwd_subflow_pkts', 'fwd_subflow_bytes',
       'bwd_subflow_bytes', 'fwd_bulk_bytes', 'bwd_bulk_bytes',
       'fwd_bulk_packets', 'bwd_bulk_packets', 'fwd_bulk_rate',
       'bwd_bulk_rate', 'active.min', 'active.max', 'active.tot', 'active.avg',
       'active.std', 'idle.min', 'idle.max', 'idle.tot', 'idle.avg',
       'idle.std', 'fwd_init_window_size', 'bwd_init_window_size',
       'fwd_last_window_size', 'Attack_type'],
      dtype='object')
```

In [18]:
```
RT1 = RT['Attack_type']
RT1
```

Out[18]:
```
0                MQTT_Publish
1                MQTT_Publish
2                MQTT_Publish
3                MQTT_Publish
4                MQTT_Publish
                     ...
123112    NMAP_XMAS_TREE_SCAN
123113    NMAP_XMAS_TREE_SCAN
123114    NMAP_XMAS_TREE_SCAN
123115    NMAP_XMAS_TREE_SCAN
123116    NMAP_XMAS_TREE_SCAN
Name: Attack_type, Length: 123117, dtype: object
```

2.Extract the provided dataset using FLAT FILE. You get extra points for loading it through Kaggle API.

import kagglehub

Download latest version

path = kagglehub.dataset_download("supplejade/rt-iot2022real-time-internet-of-things")

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

In [46]:
```
import kagglehub as kaggle

# Download latest version
RT = kagglehub.dataset_download("supplejade/rt-iot2022real-time-internet-of-things"

print("RT_IOT2022.csv", RT)
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
Cell In[46], line 1
----> 1 import kagglehub as kaggle
      3 # Download latest version
      4 RT = kagglehub.dataset_download("supplejade/rt-iot2022real-time-internet-of-
things")

ModuleNotFoundError: No module named 'kagglehub'
```

In [ ]:

3.

Transform the dataset. List down all activities included in the transformation.

In [ ]:

In [ ]:

4. Show the transformed dataset.

In [ ]:

In [ ]:

5. Load the dataset and perform statistical analysis and visualization. List down all activities included in the "load" phase.

In [57]:
```
import matplotlib as plt

plt.hist(RT['Attack_type'])
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
Cell In[57], line 3
      1 import matplotlib as plt
----> 3 plt.hist(RT['Attack_type'])

File C:\ProgramData\anaconda3\Lib\site-packages\matplotlib\_api\__init__.py:217, in
caching_module_getattr.<locals>.__getattr__(name)
    215 if name in props:
    216     return props[name].__get__(instance)
--> 217 raise AttributeError(
    218     f"module {cls.__module__!r} has no attribute {name!r}")

AttributeError: module 'matplotlib' has no attribute 'hist'
```

In [ ]:

Provide a summary of all activities performed and your insights derived from the dataset.

To sum this all up, I am still lacking knowledge in comprehension and in pandas. I have to practice on this lessons after and try to be better and learn from my mistakes since I didnt get to answer a lot of parts, just the extracting.

Also I apologize for opening other tabs, I had no intentions of cheating and have not searched for an answer, only tried to find the latest download file for kagglehub.