step2

June 1, 2024

[ ]:

```
!pip install -r requirements.txt
```

[1]:

```
from pandas import set_option

def append_suffix(name: str, suffix: str) -> str:
    return f"{name.removesuffix('.parquet')}{suffix}.parquet"

set_option('display.max_columns', None)
```

0.1 Remove unnecessary data

- remove non-TCP flows (https://en.wikipedia.org/wiki/List_of_IP_protocol_numbers)
- remove rows containing infinity/NaN values in numeric columns
- remove columns that contain low amount of unique values
- remove columns with high correlation

[11]:

```
from pandas import read_parquet
from collections import Counter
from numpy import isfinite


files = ['Wednesday-14-02-2018.parquet', 'Thursday-15-02-2018.parquet', 'Friday-16-02-2018.parquet']

for file in files:
    print(f"Reading file {file}")
    df = read_parquet(file)
    df.reset_index(drop=True, inplace=True)
    # Drop old index column, as it was not in original dataset.
    # It was added due to ORC file format requirements
    df.drop(columns=['index'], inplace=True)
    print(df.columns)

    print(f"Starting dataframe shape: {df.shape}")
    print(f"Counted rows by protocol: {Counter(df['Protocol'])}")
    df = df[df['Protocol'] == 6]
    print("Removing non-TCP rows...")
    df.drop(columns=['Protocol'], inplace=True)
```

```python
cols_to_remove = []

UNIQUE_VALUES_THRESHOLD = 10

for column in df:
    unique_vals = len(set(df[column]))
    print(f"column: '{column}', unique values: {unique_vals}")
    if unique_vals < UNIQUE_VALUES_THRESHOLD and column != 'Label':
        cols_to_remove.append(column)
        print(f"Removing column {column}")
        print(f"Column values: {Counter(df[column])}")

df.drop(columns=cols_to_remove, inplace=True)

print("Correlation matrix")
df_features = df.drop(columns=['Label'], inplace=False)

# removing fows with Inf/Nan values
finite_indexes = isfinite(df_features).all(1)
df_features = df_features[finite_indexes]
df = df[finite_indexes]

correlation_matrix = df_features.corr()

threshold = 0.95

high_correlation_rows = correlation_matrix[abs(correlation_matrix) > threshold].stack().reset_index()
high_correlation_rows.columns = ['Column 1', 'Column 2', 'Correlation']
high_correlation_rows = high_correlation_rows[high_correlation_rows['Column 1'] != high_correlation_rows['Column 2']]\
    .drop_duplicates(subset='Column 1')\
    .reset_index()

save_columns = []
drop_columns = []
for index, row in high_correlation_rows.iterrows():
    col1 = row['Column 1']
    col2 = row['Column 2']
    if col2 not in save_columns:
        drop_columns.append(col2)
        save_columns.append(col1)
        print(f"Removing column {col2} as its corellation to {col1} is {row['Correlation']}")

df.drop(columns=drop_columns, inplace=True)
```

```
    df.reset_index(drop=True, inplace=True)
    print(f"""

###################################
Final data state for file {file}
Shape (rows, columns): {df.shape}
Label counts: {Counter(df['Label'])}
###################################

    """)
    df.to_parquet(append_suffix(file, '_pruned'))
```

```
Reading file Wednesday-14-02-2018.parquet
Index(['Dst Port', 'Protocol', 'Timestamp', 'Flow Duration', 'Tot Fwd Pkts',
       'Tot Bwd Pkts', 'TotLen Fwd Pkts', 'TotLen Bwd Pkts', 'Fwd Pkt Len Max',
       'Fwd Pkt Len Min', 'Fwd Pkt Len Mean', 'Fwd Pkt Len Std',
       'Bwd Pkt Len Max', 'Bwd Pkt Len Min', 'Bwd Pkt Len Mean',
       'Bwd Pkt Len Std', 'Flow Byts/s', 'Flow Pkts/s', 'Flow IAT Mean',
       'Flow IAT Std', 'Flow IAT Max', 'Flow IAT Min', 'Fwd IAT Tot',
       'Fwd IAT Mean', 'Fwd IAT Std', 'Fwd IAT Max', 'Fwd IAT Min',
       'Bwd IAT Tot', 'Bwd IAT Mean', 'Bwd IAT Std', 'Bwd IAT Max',
       'Bwd IAT Min', 'Fwd PSH Flags', 'Bwd PSH Flags', 'Fwd URG Flags',
       'Bwd URG Flags', 'Fwd Header Len', 'Bwd Header Len', 'Fwd Pkts/s',
       'Bwd Pkts/s', 'Pkt Len Min', 'Pkt Len Max', 'Pkt Len Mean',
       'Pkt Len Std', 'Pkt Len Var', 'FIN Flag Cnt', 'SYN Flag Cnt',
       'RST Flag Cnt', 'PSH Flag Cnt', 'ACK Flag Cnt', 'URG Flag Cnt',
       'CWE Flag Count', 'ECE Flag Cnt', 'Down/Up Ratio', 'Pkt Size Avg',
       'Fwd Seg Size Avg', 'Bwd Seg Size Avg', 'Fwd Byts/b Avg',
       'Fwd Pkts/b Avg', 'Fwd Blk Rate Avg', 'Bwd Byts/b Avg',
       'Bwd Pkts/b Avg', 'Bwd Blk Rate Avg', 'Subflow Fwd Pkts',
       'Subflow Fwd Byts', 'Subflow Bwd Pkts', 'Subflow Bwd Byts',
       'Init Fwd Win Byts', 'Init Bwd Win Byts', 'Fwd Act Data Pkts',
       'Fwd Seg Size Min', 'Active Mean', 'Active Std', 'Active Max',
       'Active Min', 'Idle Mean', 'Idle Std', 'Idle Max', 'Idle Min', 'Label'],
      dtype='object')
Starting dataframe shape: (1048575, 80)
Counted rows by protocol: Counter({6: 829309, 17: 207384, 0: 11882})
Removing non-TCP rows…
column: 'Dst Port', unique values: 18538
column: 'Timestamp', unique values: 31941
column: 'Flow Duration', unique values: 335283
column: 'Tot Fwd Pkts', unique values: 749
column: 'Tot Bwd Pkts', unique values: 988
column: 'TotLen Fwd Pkts', unique values: 5778
column: 'TotLen Bwd Pkts', unique values: 15437
column: 'Fwd Pkt Len Max', unique values: 1354
column: 'Fwd Pkt Len Min', unique values: 48
column: 'Fwd Pkt Len Mean', unique values: 17872
```

```
column: 'Fwd Pkt Len Std', unique values: 27408
column: 'Bwd Pkt Len Max', unique values: 837
column: 'Bwd Pkt Len Min', unique values: 46
column: 'Bwd Pkt Len Mean', unique values: 24383
column: 'Bwd Pkt Len Std', unique values: 28595
column: 'Flow Byts/s', unique values: 325382
column: 'Flow Pkts/s', unique values: 361819
column: 'Flow IAT Mean', unique values: 360868
column: 'Flow IAT Std', unique values: 373314
column: 'Flow IAT Max', unique values: 256214
column: 'Flow IAT Min', unique values: 37431
column: 'Fwd IAT Tot', unique values: 314511
column: 'Fwd IAT Mean', unique values: 334938
column: 'Fwd IAT Std', unique values: 335659
column: 'Fwd IAT Max', unique values: 273699
column: 'Fwd IAT Min', unique values: 53706
column: 'Bwd IAT Tot', unique values: 278910
column: 'Bwd IAT Mean', unique values: 293878
column: 'Bwd IAT Std', unique values: 329738
column: 'Bwd IAT Max', unique values: 210086
column: 'Bwd IAT Min', unique values: 98141
column: 'Fwd PSH Flags', unique values: 2
Removing column Fwd PSH Flags
Column values: Counter({0: 801423, 1: 27886})
column: 'Bwd PSH Flags', unique values: 1
Removing column Bwd PSH Flags
Column values: Counter({0: 829309})
column: 'Fwd URG Flags', unique values: 1
Removing column Fwd URG Flags
Column values: Counter({0: 829309})
column: 'Bwd URG Flags', unique values: 1
Removing column Bwd URG Flags
Column values: Counter({0: 829309})
column: 'Fwd Header Len', unique values: 1234
column: 'Bwd Header Len', unique values: 1841
column: 'Fwd Pkts/s', unique values: 355982
column: 'Bwd Pkts/s', unique values: 319597
column: 'Pkt Len Min', unique values: 12
column: 'Pkt Len Max', unique values: 995
column: 'Pkt Len Mean', unique values: 35540
column: 'Pkt Len Std', unique values: 41147
column: 'Pkt Len Var', unique values: 41300
column: 'FIN Flag Cnt', unique values: 2
Removing column FIN Flag Cnt
Column values: Counter({0: 825563, 1: 3746})
column: 'SYN Flag Cnt', unique values: 2
Removing column SYN Flag Cnt
Column values: Counter({0: 801423, 1: 27886})
```

```
column: 'RST Flag Cnt', unique values: 2
Removing column RST Flag Cnt
Column values: Counter({0: 787781, 1: 41528})
column: 'PSH Flag Cnt', unique values: 2
Removing column PSH Flag Cnt
Column values: Counter({1: 550954, 0: 278355})
column: 'ACK Flag Cnt', unique values: 2
Removing column ACK Flag Cnt
Column values: Counter({0: 553041, 1: 276268})
column: 'URG Flag Cnt', unique values: 2
Removing column URG Flag Cnt
Column values: Counter({0: 702278, 1: 127031})
column: 'CWE Flag Count', unique values: 1
Removing column CWE Flag Count
Column values: Counter({0: 829309})
column: 'ECE Flag Cnt', unique values: 2
Removing column ECE Flag Cnt
Column values: Counter({0: 787782, 1: 41527})
column: 'Down/Up Ratio', unique values: 33
column: 'Pkt Size Avg', unique values: 35392
column: 'Fwd Seg Size Avg', unique values: 17872
column: 'Bwd Seg Size Avg', unique values: 24382
column: 'Fwd Byts/b Avg', unique values: 1
Removing column Fwd Byts/b Avg
Column values: Counter({0: 829309})
column: 'Fwd Pkts/b Avg', unique values: 1
Removing column Fwd Pkts/b Avg
Column values: Counter({0: 829309})
column: 'Fwd Blk Rate Avg', unique values: 1
Removing column Fwd Blk Rate Avg
Column values: Counter({0: 829309})
column: 'Bwd Byts/b Avg', unique values: 1
Removing column Bwd Byts/b Avg
Column values: Counter({0: 829309})
column: 'Bwd Pkts/b Avg', unique values: 1
Removing column Bwd Pkts/b Avg
Column values: Counter({0: 829309})
column: 'Bwd Blk Rate Avg', unique values: 1
Removing column Bwd Blk Rate Avg
Column values: Counter({0: 829309})
column: 'Subflow Fwd Pkts', unique values: 749
column: 'Subflow Fwd Byts', unique values: 5778
column: 'Subflow Bwd Pkts', unique values: 988
column: 'Subflow Bwd Byts', unique values: 15437
column: 'Init Fwd Win Byts', unique values: 3307
column: 'Init Bwd Win Byts', unique values: 3547
column: 'Fwd Act Data Pkts', unique values: 112
column: 'Fwd Seg Size Min', unique values: 8
```

```
Removing column Fwd Seg Size Min
Column values: Counter({20: 444401, 40: 193877, 32: 189921, 28: 972, 24: 111,
36: 25, 48: 1, 44: 1})
column: 'Active Mean', unique values: 71914
column: 'Active Std', unique values: 54865
column: 'Active Max', unique values: 69811
column: 'Active Min', unique values: 37375
column: 'Idle Mean', unique values: 73109
column: 'Idle Std', unique values: 56056
column: 'Idle Max', unique values: 52950
column: 'Idle Min', unique values: 62860
column: 'Label', unique values: 3
Correlation matrix
Removing column Fwd IAT Tot as its corellation to Flow Duration is
0.9919617405800067
Removing column Fwd Header Len as its corellation to Tot Fwd Pkts is
0.98723295015188
Removing column TotLen Bwd Pkts as its corellation to Tot Bwd Pkts is
0.9969485891335339
Removing column Subflow Fwd Byts as its corellation to TotLen Fwd Pkts is 1.0
Removing column Fwd Pkt Len Std as its corellation to Fwd Pkt Len Max is
0.9509477502935201
Removing column Fwd Pkt Len Std as its corellation to Fwd Pkt Len Mean is
0.953993181292946
Removing column Bwd Pkt Len Std as its corellation to Bwd Pkt Len Max is
0.968353191745189
Removing column Pkt Len Mean as its corellation to Bwd Pkt Len Mean is
0.9728679795759958
Removing column Fwd Pkts/s as its corellation to Flow Pkts/s is
0.9623216215358824
Removing column Flow IAT Min as its corellation to Flow IAT Mean is
0.9837769243116777
Removing column Fwd IAT Max as its corellation to Flow IAT Max is
0.9740383720535271
Removing column Bwd Pkt Len Std as its corellation to Pkt Len Std is
0.9705190745106906
Removing column Active Min as its corellation to Active Mean is
0.973153202109275
Removing column Idle Max as its corellation to Idle Mean is 0.9932573483413581


###################################
Final data state for file Wednesday-14-02-2018.parquet
Shape (rows, columns): (825485, 48)
Label counts: Counter({'Benign': 444542, 'FTP-BruteForce': 193354, 'SSH-
Bruteforce': 187589})
###################################
```

```
Reading file Thursday-15-02-2018.parquet
Index(['Dst Port', 'Protocol', 'Timestamp', 'Flow Duration', 'Tot Fwd Pkts',
       'Tot Bwd Pkts', 'TotLen Fwd Pkts', 'TotLen Bwd Pkts', 'Fwd Pkt Len Max',
       'Fwd Pkt Len Min', 'Fwd Pkt Len Mean', 'Fwd Pkt Len Std',
       'Bwd Pkt Len Max', 'Bwd Pkt Len Min', 'Bwd Pkt Len Mean',
       'Bwd Pkt Len Std', 'Flow Byts/s', 'Flow Pkts/s', 'Flow IAT Mean',
       'Flow IAT Std', 'Flow IAT Max', 'Flow IAT Min', 'Fwd IAT Tot',
       'Fwd IAT Mean', 'Fwd IAT Std', 'Fwd IAT Max', 'Fwd IAT Min',
       'Bwd IAT Tot', 'Bwd IAT Mean', 'Bwd IAT Std', 'Bwd IAT Max',
       'Bwd IAT Min', 'Fwd PSH Flags', 'Bwd PSH Flags', 'Fwd URG Flags',
       'Bwd URG Flags', 'Fwd Header Len', 'Bwd Header Len', 'Fwd Pkts/s',
       'Bwd Pkts/s', 'Pkt Len Min', 'Pkt Len Max', 'Pkt Len Mean',
       'Pkt Len Std', 'Pkt Len Var', 'FIN Flag Cnt', 'SYN Flag Cnt',
       'RST Flag Cnt', 'PSH Flag Cnt', 'ACK Flag Cnt', 'URG Flag Cnt',
       'CWE Flag Count', 'ECE Flag Cnt', 'Down/Up Ratio', 'Pkt Size Avg',
       'Fwd Seg Size Avg', 'Bwd Seg Size Avg', 'Fwd Byts/b Avg',
       'Fwd Pkts/b Avg', 'Fwd Blk Rate Avg', 'Bwd Byts/b Avg',
       'Bwd Pkts/b Avg', 'Bwd Blk Rate Avg', 'Subflow Fwd Pkts',
       'Subflow Fwd Byts', 'Subflow Bwd Pkts', 'Subflow Bwd Byts',
       'Init Fwd Win Byts', 'Init Bwd Win Byts', 'Fwd Act Data Pkts',
       'Fwd Seg Size Min', 'Active Mean', 'Active Std', 'Active Max',
       'Active Min', 'Idle Mean', 'Idle Std', 'Idle Max', 'Idle Min', 'Label'],
      dtype='object')
Starting dataframe shape: (1048575, 80)
Counted rows by protocol: Counter({6: 684486, 17: 345524, 0: 18565})
Removing non-TCP rows…
column: 'Dst Port', unique values: 21703
column: 'Timestamp', unique values: 33774
column: 'Flow Duration', unique values: 408713
column: 'Tot Fwd Pkts', unique values: 693
column: 'Tot Bwd Pkts', unique values: 1189
column: 'TotLen Fwd Pkts', unique values: 7669
column: 'TotLen Bwd Pkts', unique values: 22522
column: 'Fwd Pkt Len Max', unique values: 1427
column: 'Fwd Pkt Len Min', unique values: 72
column: 'Fwd Pkt Len Mean', unique values: 25677
column: 'Fwd Pkt Len Std', unique values: 46872
column: 'Bwd Pkt Len Max', unique values: 1094
column: 'Bwd Pkt Len Min', unique values: 58
column: 'Bwd Pkt Len Mean', unique values: 37950
column: 'Bwd Pkt Len Std', unique values: 46508
column: 'Flow Byts/s', unique values: 345089
column: 'Flow Pkts/s', unique values: 418105
column: 'Flow IAT Mean', unique values: 416434
column: 'Flow IAT Std', unique values: 399049
column: 'Flow IAT Max', unique values: 326876
column: 'Flow IAT Min', unique values: 72583
```

```
column: 'Fwd IAT Tot', unique values: 381761
column: 'Fwd IAT Mean', unique values: 389112
column: 'Fwd IAT Std', unique values: 343771
column: 'Fwd IAT Max', unique values: 330737
column: 'Fwd IAT Min', unique values: 96481
column: 'Bwd IAT Tot', unique values: 306090
column: 'Bwd IAT Mean', unique values: 309552
column: 'Bwd IAT Std', unique values: 323811
column: 'Bwd IAT Max', unique values: 240929
column: 'Bwd IAT Min', unique values: 105367
column: 'Fwd PSH Flags', unique values: 2
Removing column Fwd PSH Flags
Column values: Counter({0: 630759, 1: 53727})
column: 'Bwd PSH Flags', unique values: 1
Removing column Bwd PSH Flags
Column values: Counter({0: 684486})
column: 'Fwd URG Flags', unique values: 1
Removing column Fwd URG Flags
Column values: Counter({0: 684486})
column: 'Bwd URG Flags', unique values: 1
Removing column Bwd URG Flags
Column values: Counter({0: 684486})
column: 'Fwd Header Len', unique values: 1361
column: 'Bwd Header Len', unique values: 2327
column: 'Fwd Pkts/s', unique values: 415266
column: 'Bwd Pkts/s', unique values: 351340
column: 'Pkt Len Min', unique values: 19
column: 'Pkt Len Max', unique values: 1289
column: 'Pkt Len Mean', unique values: 56598
column: 'Pkt Len Std', unique values: 69955
column: 'Pkt Len Var', unique values: 70185
column: 'FIN Flag Cnt', unique values: 2
Removing column FIN Flag Cnt
Column values: Counter({0: 678334, 1: 6152})
column: 'SYN Flag Cnt', unique values: 2
Removing column SYN Flag Cnt
Column values: Counter({0: 630759, 1: 53727})
column: 'RST Flag Cnt', unique values: 2
Removing column RST Flag Cnt
Column values: Counter({0: 620537, 1: 63949})
column: 'PSH Flag Cnt', unique values: 2
Removing column PSH Flag Cnt
Column values: Counter({1: 381134, 0: 303352})
column: 'ACK Flag Cnt', unique values: 2
Removing column ACK Flag Cnt
Column values: Counter({0: 384985, 1: 299501})
column: 'URG Flag Cnt', unique values: 2
Removing column URG Flag Cnt
```

```
Column values: Counter({0: 634352, 1: 50134})
column: 'CWE Flag Count', unique values: 1
Removing column CWE Flag Count
Column values: Counter({0: 684486})
column: 'ECE Flag Cnt', unique values: 2
Removing column ECE Flag Cnt
Column values: Counter({0: 620539, 1: 63947})
column: 'Down/Up Ratio', unique values: 48
column: 'Pkt Size Avg', unique values: 56270
column: 'Fwd Seg Size Avg', unique values: 25677
column: 'Bwd Seg Size Avg', unique values: 37949
column: 'Fwd Byts/b Avg', unique values: 1
Removing column Fwd Byts/b Avg
Column values: Counter({0: 684486})
column: 'Fwd Pkts/b Avg', unique values: 1
Removing column Fwd Pkts/b Avg
Column values: Counter({0: 684486})
column: 'Fwd Blk Rate Avg', unique values: 1
Removing column Fwd Blk Rate Avg
Column values: Counter({0: 684486})
column: 'Bwd Byts/b Avg', unique values: 1
Removing column Bwd Byts/b Avg
Column values: Counter({0: 684486})
column: 'Bwd Pkts/b Avg', unique values: 1
Removing column Bwd Pkts/b Avg
Column values: Counter({0: 684486})
column: 'Bwd Blk Rate Avg', unique values: 1
Removing column Bwd Blk Rate Avg
Column values: Counter({0: 684486})
column: 'Subflow Fwd Pkts', unique values: 693
column: 'Subflow Fwd Byts', unique values: 7669
column: 'Subflow Bwd Pkts', unique values: 1189
column: 'Subflow Bwd Byts', unique values: 22522
column: 'Init Fwd Win Byts', unique values: 4492
column: 'Init Bwd Win Byts', unique values: 5123
column: 'Fwd Act Data Pkts', unique values: 138
column: 'Fwd Seg Size Min', unique values: 9
Removing column Fwd Seg Size Min
Column values: Counter({20: 615130, 32: 52376, 40: 14798, 28: 1976, 24: 133, 44:
42, 36: 28, 48: 2, 56: 1})
column: 'Active Mean', unique values: 121088
column: 'Active Std', unique values: 92294
column: 'Active Max', unique values: 115950
column: 'Active Min', unique values: 66057
column: 'Idle Mean', unique values: 142114
column: 'Idle Std', unique values: 94375
column: 'Idle Max', unique values: 114719
column: 'Idle Min', unique values: 124089
```

```
column: 'Label', unique values: 3
Correlation matrix
Removing column Fwd IAT Tot as its corellation to Flow Duration is
0.993781361910096
Removing column Fwd Header Len as its corellation to Tot Fwd Pkts is
0.9759575735557326
Removing column TotLen Bwd Pkts as its corellation to Tot Bwd Pkts is
0.9925035947290912
Removing column Subflow Fwd Byts as its corellation to TotLen Fwd Pkts is 1.0
Removing column Fwd Seg Size Avg as its corellation to Fwd Pkt Len Mean is 1.0
Removing column Bwd Pkt Len Std as its corellation to Bwd Pkt Len Max is
0.972276179239887
Removing column Pkt Len Mean as its corellation to Bwd Pkt Len Mean is
0.9630250792364126
Removing column Fwd Pkts/s as its corellation to Flow Pkts/s is
0.9854381209708611
Removing column Flow IAT Min as its corellation to Flow IAT Mean is
0.9523807718643995
Removing column Fwd IAT Max as its corellation to Flow IAT Max is
0.9816228230855114
Removing column Fwd IAT Min as its corellation to Fwd IAT Mean is
0.9762199553311824
Removing column Bwd Pkt Len Std as its corellation to Pkt Len Std is
0.9532881414224581
Removing column Active Min as its corellation to Active Mean is
0.9661673367048954
Removing column Idle Max as its corellation to Idle Mean is 0.9829703833537147


###################################
Final data state for file Thursday-15-02-2018.parquet
Shape (rows, columns): (676461, 47)
Label counts: Counter({'Benign': 623963, 'DoS attacks-GoldenEye': 41508, 'DoS
attacks-Slowloris': 10990})
###################################



Reading file Friday-16-02-2018.parquet
Index(['Dst Port', 'Protocol', 'Timestamp', 'Flow Duration', 'Tot Fwd Pkts',
       'Tot Bwd Pkts', 'TotLen Fwd Pkts', 'TotLen Bwd Pkts', 'Fwd Pkt Len Max',
       'Fwd Pkt Len Min', 'Fwd Pkt Len Mean', 'Fwd Pkt Len Std',
       'Bwd Pkt Len Max', 'Bwd Pkt Len Min', 'Bwd Pkt Len Mean',
       'Bwd Pkt Len Std', 'Flow Byts/s', 'Flow Pkts/s', 'Flow IAT Mean',
       'Flow IAT Std', 'Flow IAT Max', 'Flow IAT Min', 'Fwd IAT Tot',
       'Fwd IAT Mean', 'Fwd IAT Std', 'Fwd IAT Max', 'Fwd IAT Min',
       'Bwd IAT Tot', 'Bwd IAT Mean', 'Bwd IAT Std', 'Bwd IAT Max',
       'Bwd IAT Min', 'Fwd PSH Flags', 'Bwd PSH Flags', 'Fwd URG Flags',
       'Bwd URG Flags', 'Fwd Header Len', 'Bwd Header Len', 'Fwd Pkts/s',
       'Bwd Pkts/s', 'Pkt Len Min', 'Pkt Len Max', 'Pkt Len Mean',
       'Pkt Len Std', 'Pkt Len Var', 'FIN Flag Cnt', 'SYN Flag Cnt',
       'RST Flag Cnt', 'PSH Flag Cnt', 'ACK Flag Cnt', 'URG Flag Cnt',
       'CWE Flag Count', 'ECE Flag Cnt', 'Down/Up Ratio', 'Pkt Size Avg',
       'Fwd Seg Size Avg', 'Bwd Seg Size Avg', 'Fwd Byts/b Avg',
       'Fwd Pkts/b Avg', 'Fwd Blk Rate Avg', 'Bwd Byts/b Avg',
       'Bwd Pkts/b Avg', 'Bwd Blk Rate Avg', 'Subflow Fwd Pkts',
       'Subflow Fwd Byts', 'Subflow Bwd Pkts', 'Subflow Bwd Byts',
       'Init Fwd Win Byts', 'Init Bwd Win Byts', 'Fwd Act Data Pkts',
       'Fwd Seg Size Min', 'Active Mean', 'Active Std', 'Active Max',
       'Active Min', 'Idle Mean', 'Idle Std', 'Idle Max', 'Idle Min', 'Label'],
      dtype='object')
Starting dataframe shape: (1048574, 80)
Counted rows by protocol: Counter({6: 1048397, 0: 162, 17: 15})
Removing non-TCP rows…
column: 'Dst Port', unique values: 14134
column: 'Timestamp', unique values: 3071
column: 'Flow Duration', unique values: 453244
column: 'Tot Fwd Pkts', unique values: 24
column: 'Tot Bwd Pkts', unique values: 21
column: 'TotLen Fwd Pkts', unique values: 460
column: 'TotLen Bwd Pkts', unique values: 696
column: 'Fwd Pkt Len Max', unique values: 164
column: 'Fwd Pkt Len Min', unique values: 2
Removing column Fwd Pkt Len Min
Column values: Counter({0: 1048390, 64: 7})
column: 'Fwd Pkt Len Mean', unique values: 908
column: 'Fwd Pkt Len Std', unique values: 948
column: 'Bwd Pkt Len Max', unique values: 157
column: 'Bwd Pkt Len Min', unique values: 2
Removing column Bwd Pkt Len Min
Column values: Counter({0: 1048395, 272: 2})
column: 'Bwd Pkt Len Mean', unique values: 1442
column: 'Bwd Pkt Len Std', unique values: 1221
column: 'Flow Byts/s', unique values: 466676
column: 'Flow Pkts/s', unique values: 489676
column: 'Flow IAT Mean', unique values: 481579
column: 'Flow IAT Std', unique values: 495168
column: 'Flow IAT Max', unique values: 407842
column: 'Flow IAT Min', unique values: 58968
column: 'Fwd IAT Tot', unique values: 444139
column: 'Fwd IAT Mean', unique values: 453574
column: 'Fwd IAT Std', unique values: 483847
column: 'Fwd IAT Max', unique values: 413422
column: 'Fwd IAT Min', unique values: 60164
column: 'Bwd IAT Tot', unique values: 201721
column: 'Bwd IAT Mean', unique values: 216406
column: 'Bwd IAT Std', unique values: 237285
```

```
column: 'Bwd IAT Max', unique values: 185325
column: 'Bwd IAT Min', unique values: 39097
column: 'Fwd PSH Flags', unique values: 2
Removing column Fwd PSH Flags
Column values: Counter({0: 1048345, 1: 52})
column: 'Bwd PSH Flags', unique values: 1
Removing column Bwd PSH Flags
Column values: Counter({0: 1048397})
column: 'Fwd URG Flags', unique values: 1
Removing column Fwd URG Flags
Column values: Counter({0: 1048397})
column: 'Bwd URG Flags', unique values: 1
Removing column Bwd URG Flags
Column values: Counter({0: 1048397})
column: 'Fwd Header Len', unique values: 52
column: 'Bwd Header Len', unique values: 53
column: 'Fwd Pkts/s', unique values: 470610
column: 'Bwd Pkts/s', unique values: 406241
column: 'Pkt Len Min', unique values: 1
Removing column Pkt Len Min
Column values: Counter({0: 1048397})
column: 'Pkt Len Max', unique values: 18
column: 'Pkt Len Mean', unique values: 2003
column: 'Pkt Len Std', unique values: 2093
column: 'Pkt Len Var', unique values: 2093
column: 'FIN Flag Cnt', unique values: 2
Removing column FIN Flag Cnt
Column values: Counter({0: 1047130, 1: 1267})
column: 'SYN Flag Cnt', unique values: 2
Removing column SYN Flag Cnt
Column values: Counter({0: 1048345, 1: 52})
column: 'RST Flag Cnt', unique values: 2
Removing column RST Flag Cnt
Column values: Counter({0: 1048394, 1: 3})
column: 'PSH Flag Cnt', unique values: 2
Removing column PSH Flag Cnt
Column values: Counter({0: 894248, 1: 154149})
column: 'ACK Flag Cnt', unique values: 2
Removing column ACK Flag Cnt
Column values: Counter({1: 892981, 0: 155416})
column: 'URG Flag Cnt', unique values: 2
Removing column URG Flag Cnt
Column values: Counter({0: 1021995, 1: 26402})
column: 'CWE Flag Count', unique values: 1
Removing column CWE Flag Count
Column values: Counter({0: 1048397})
column: 'ECE Flag Cnt', unique values: 2
Removing column ECE Flag Cnt
```

```
Column values: Counter({0: 1048394, 1: 3})
column: 'Down/Up Ratio', unique values: 5
Removing column Down/Up Ratio
Column values: Counter({0: 847962, 1: 197878, 2: 1727, 3: 828, 4: 2})
column: 'Pkt Size Avg', unique values: 1997
column: 'Fwd Seg Size Avg', unique values: 908
column: 'Bwd Seg Size Avg', unique values: 1442
column: 'Fwd Byts/b Avg', unique values: 1
Removing column Fwd Byts/b Avg
Column values: Counter({0: 1048397})
column: 'Fwd Pkts/b Avg', unique values: 1
Removing column Fwd Pkts/b Avg
Column values: Counter({0: 1048397})
column: 'Fwd Blk Rate Avg', unique values: 1
Removing column Fwd Blk Rate Avg
Column values: Counter({0: 1048397})
column: 'Bwd Byts/b Avg', unique values: 1
Removing column Bwd Byts/b Avg
Column values: Counter({0: 1048397})
column: 'Bwd Pkts/b Avg', unique values: 1
Removing column Bwd Pkts/b Avg
Column values: Counter({0: 1048397})
column: 'Bwd Blk Rate Avg', unique values: 1
Removing column Bwd Blk Rate Avg
Column values: Counter({0: 1048397})
column: 'Subflow Fwd Pkts', unique values: 24
column: 'Subflow Fwd Byts', unique values: 460
column: 'Subflow Bwd Pkts', unique values: 21
column: 'Subflow Bwd Byts', unique values: 696
column: 'Init Fwd Win Byts', unique values: 23
column: 'Init Bwd Win Byts', unique values: 22
column: 'Fwd Act Data Pkts', unique values: 15
column: 'Fwd Seg Size Min', unique values: 3
Removing column Fwd Seg Size Min
Column values: Counter({32: 906232, 40: 139890, 20: 2275})
column: 'Active Mean', unique values: 5683
column: 'Active Std', unique values: 16
column: 'Active Max', unique values: 5684
column: 'Active Min', unique values: 5683
column: 'Idle Mean', unique values: 45987
column: 'Idle Std', unique values: 18
column: 'Idle Max', unique values: 45987
column: 'Idle Min', unique values: 45987
column: 'Label', unique values: 3
Correlation matrix
Removing column TotLen Fwd Pkts as its corellation to Dst Port is
0.9569207374359241
Removing column Flow IAT Mean as its corellation to Flow Duration is
```

```
0.99504178052204
Removing column Fwd Header Len as its corellation to Tot Fwd Pkts is
0.9954504263648369
Removing column Bwd Header Len as its corellation to Tot Bwd Pkts is
0.99428761766386
Removing column Subflow Bwd Byts as its corellation to TotLen Bwd Pkts is 1.0
Removing column TotLen Fwd Pkts as its corellation to Fwd Pkt Len Mean is
0.9944728515992636
Removing column Bwd Pkt Len Std as its corellation to Bwd Pkt Len Max is
0.9916275126984898
Removing column Bwd Pkt Len Std as its corellation to Bwd Pkt Len Mean is
0.9668734436677687
Removing column Fwd Pkts/s as its corellation to Flow Pkts/s is
0.9998774752166524
Removing column Bwd IAT Min as its corellation to Fwd IAT Min is
0.9942416949890744
Removing column Bwd IAT Mean as its corellation to Bwd IAT Tot is
0.9880254240737173
Removing column TotLen Fwd Pkts as its corellation to Pkt Len Max is
0.9796825083935715
Removing column TotLen Fwd Pkts as its corellation to Pkt Len Mean is
0.9726760776888527
Removing column TotLen Fwd Pkts as its corellation to Pkt Len Std is
0.9780823953181207
Removing column TotLen Fwd Pkts as its corellation to Pkt Len Var is
0.9726482530867464
Removing column TotLen Fwd Pkts as its corellation to Pkt Size Avg is
0.9718163281340162
Removing column TotLen Fwd Pkts as its corellation to Fwd Seg Size Avg is
0.9944728515992636
Removing column Active Max as its corellation to Active Mean is
0.9575588512890134


###################################
Final data state for file Friday-16-02-2018.parquet
Shape (rows, columns): (1048397, 46)
Label counts: Counter({'DoS attacks-Hulk': 461912, 'Benign': 446595, 'DoS
attacks-SlowHTTPTest': 139890})
###################################
```

0.2 Normalization

Additional step with results saved in separate files, to compare if normalization improves efficiency

[12]:

```python
from sklearn.preprocessing import MinMaxScaler
from pandas import DataFrame, concat
from numpy import isfinite

files = ['Wednesday-14-02-2018_pruned.parquet', 'Thursday-15-02-2018_pruned.
 ↪parquet', 'Friday-16-02-2018_pruned.parquet']

for file in files:
    print(f"Reading file {file}")
    df = read_parquet(file)
    df_features = df.drop(columns=['Label'], inplace=False)

    scaler = MinMaxScaler()

    normalized_features = DataFrame(scaler.fit_transform(df_features),␣
 ↪columns=df_features.columns)

    # Combine the normalized numeric data with the non-numeric data
    df = concat([normalized_features, df['Label']], axis=1, ignore_index=False)
    print("Example row after normalization:")
    # df = df.reset_index(drop=True)
    print(df.head(1))
    print(df.shape)
    print("dtypes:")
    print(df.dtypes)
    df.to_parquet(append_suffix(file, '_normalized'))
```

```
Reading file Wednesday-14-02-2018_pruned.parquet
Example row after normalization:
   Dst Port  Timestamp  Flow Duration  Tot Fwd Pkts  Tot Bwd Pkts  \
0  0.000153   0.639205       0.053783      0.002738      0.001087

   TotLen Fwd Pkts  Fwd Pkt Len Max  Fwd Pkt Len Min  Fwd Pkt Len Mean  \
0         0.000144         0.011546              0.0          0.007364

   Bwd Pkt Len Max  Bwd Pkt Len Min  Bwd Pkt Len Mean  Flow Byts/s  \
0         0.668493              0.0          0.155766     0.000001

    Flow Pkts/s  Flow IAT Mean  Flow IAT Std  Flow IAT Max  Fwd IAT Mean  \
0  9.641875e-07       0.002264      0.002939      0.005633      0.003881

   Fwd IAT Std  Fwd IAT Min  Bwd IAT Tot  Bwd IAT Mean  Bwd IAT Std  \
0     0.001462     0.001934     0.046983      0.006496     0.005395

   Bwd IAT Max  Bwd IAT Min  Bwd Header Len    Bwd Pkts/s  Pkt Len Min  \
0     0.009763     0.000006        0.001783  7.747174e-07          0.0

   Pkt Len Max  Pkt Len Std  Pkt Len Var  Down/Up Ratio  Pkt Size Avg  \
0     0.015146     0.031534     0.000994            0.0      0.055699

   Fwd Seg Size Avg  Bwd Seg Size Avg  Subflow Fwd Pkts  Subflow Bwd Pkts  \
0          0.007364          0.155766          0.002738          0.001087

   Subflow Bwd Byts  Init Fwd Win Byts  Init Bwd Win Byts  Fwd Act Data Pkts  \
0           0.00017                1.0           0.003571            0.00582

   Active Mean  Active Std  Active Max  Idle Mean  Idle Std  Idle Min   Label
0          0.0         0.0         0.0        0.0       0.0       0.0  Benign
(825485, 48)
dtypes:
Dst Port                    float64
Timestamp                   float64
Flow Duration               float64
Tot Fwd Pkts                float64
Tot Bwd Pkts                float64
TotLen Fwd Pkts             float64
Fwd Pkt Len Max             float64
Fwd Pkt Len Min             float64
Fwd Pkt Len Mean            float64
Bwd Pkt Len Max             float64
Bwd Pkt Len Min             float64
Bwd Pkt Len Mean            float64
Flow Byts/s                 float64
Flow Pkts/s                 float64
Flow IAT Mean               float64
Flow IAT Std                float64
Flow IAT Max                float64
Fwd IAT Mean                float64
Fwd IAT Std                 float64
Fwd IAT Min                 float64
Bwd IAT Tot                 float64
Bwd IAT Mean                float64
Bwd IAT Std                 float64
Bwd IAT Max                 float64
Bwd IAT Min                 float64
Bwd Header Len              float64
Bwd Pkts/s                  float64
Pkt Len Min                 float64
Pkt Len Max                 float64
Pkt Len Std                 float64
Pkt Len Var                 float64
Down/Up Ratio               float64
Pkt Size Avg                float64
Fwd Seg Size Avg            float64
Bwd Seg Size Avg            float64
Subflow Fwd Pkts            float64
Subflow Bwd Pkts            float64
Subflow Bwd Byts            float64
Init Fwd Win Byts           float64
Init Bwd Win Byts           float64
Fwd Act Data Pkts           float64
Active Mean                 float64
Active Std                  float64
Active Max                  float64
Idle Mean                   float64
Idle Std                    float64
Idle Min                    float64
Label                string[python]
dtype: object
Reading file Thursday-15-02-2018_pruned.parquet
Example row after normalization:
   Dst Port  Timestamp  Flow Duration  Tot Fwd Pkts  Tot Bwd Pkts  \
0  0.000259   0.623741        0.31139      0.001441      0.000626   

   TotLen Fwd Pkts  Fwd Pkt Len Max  Fwd Pkt Len Min  Fwd Pkt Len Mean  \
0         0.000248         0.011049              0.0          0.009369   

   Fwd Pkt Len Std  Bwd Pkt Len Max  Bwd Pkt Len Min  Bwd Pkt Len Mean  \
0          0.01385         0.360414              0.0          0.171075   

    Flow Byts/s   Flow Pkts/s  Flow IAT Mean  Flow IAT Std  Flow IAT Max  \
0  1.063669e-07  1.697423e-07       0.012582      0.046197      0.130669   

   Fwd IAT Mean  Fwd IAT Std  Bwd IAT Tot  Bwd IAT Mean  Bwd IAT Std  \
0       0.02406     0.060543      0.31139      0.028435     0.065928   

   Bwd IAT Max  Bwd IAT Min  Bwd Header Len    Bwd Pkts/s  Pkt Len Min  \
0     0.133203     0.000007        0.001084  1.605705e-07          0.0   

   Pkt Len Max  Pkt Len Std  Pkt Len Var  Down/Up Ratio  Pkt Size Avg  \
0     0.015146     0.030258     0.000916            0.0       0.05964   

   Bwd Seg Size Avg  Subflow Fwd Pkts  Subflow Bwd Pkts  Subflow Bwd Byts  \
0          0.171075          0.001441          0.000626          0.000107   

   Init Fwd Win Byts  Init Bwd Win Byts  Fwd Act Data Pkts  Active Mean  \
0           0.445563           0.003555            0.00545     0.009147   

   Active Std  Active Max  Idle Mean  Idle Std  Idle Min   Label  
0    0.010785    0.014297   0.095644  0.056435  0.074969  Benign  
(676461, 47)
dtypes:
Dst Port                    float64
Timestamp                   float64
```

```
Flow Duration               float64
Tot Fwd Pkts                float64
Tot Bwd Pkts                float64
TotLen Fwd Pkts             float64
Fwd Pkt Len Max             float64
Fwd Pkt Len Min             float64
Fwd Pkt Len Mean            float64
Fwd Pkt Len Std             float64
Bwd Pkt Len Max             float64
Bwd Pkt Len Min             float64
Bwd Pkt Len Mean            float64
Flow Byts/s                 float64
Flow Pkts/s                 float64
Flow IAT Mean               float64
Flow IAT Std                float64
Flow IAT Max                float64
Fwd IAT Mean                float64
Fwd IAT Std                 float64
Bwd IAT Tot                 float64
Bwd IAT Mean                float64
Bwd IAT Std                 float64
Bwd IAT Max                 float64
Bwd IAT Min                 float64
Bwd Header Len              float64
Bwd Pkts/s                  float64
Pkt Len Min                 float64
Pkt Len Max                 float64
Pkt Len Std                 float64
Pkt Len Var                 float64
Down/Up Ratio               float64
Pkt Size Avg                float64
Bwd Seg Size Avg            float64
Subflow Fwd Pkts            float64
Subflow Bwd Pkts            float64
Subflow Bwd Byts            float64
Init Fwd Win Byts           float64
Init Bwd Win Byts           float64
Fwd Act Data Pkts           float64
Active Mean                 float64
Active Std                  float64
Active Max                  float64
Idle Mean                   float64
Idle Std                    float64
Idle Min                    float64
Label                string[python]
dtype: object
Reading file Friday-16-02-2018_pruned.parquet
Example row after normalization:
   Dst Port  Timestamp  Flow Duration  Tot Fwd Pkts  Tot Bwd Pkts  \
0  0.583564   0.621204       0.227711      0.096491      0.049494

   TotLen Bwd Pkts  Fwd Pkt Len Max  Fwd Pkt Len Mean  Fwd Pkt Len Std  \
0         0.000858         0.082418          0.150578          0.09096

   Bwd Pkt Len Max  Bwd Pkt Len Mean  Flow Byts/s  Flow Pkts/s  Flow IAT Std  \
0         0.007366          0.015617     0.000105     0.000002      0.021745

   Flow IAT Max  Flow IAT Min  Fwd IAT Tot  Fwd IAT Mean  Fwd IAT Std  \
0      0.099214      0.000077     0.226725      0.010301     0.049825

   Fwd IAT Max   Fwd IAT Min  Bwd IAT Tot  Bwd IAT Std  Bwd IAT Max  \
0     0.100161  7.228916e-07     0.226725      0.03646     0.099214

   Bwd Pkts/s  Pkt Len Max  Pkt Len Mean  Pkt Len Std  Pkt Len Var  \
0    0.000002     0.027624      0.039343     0.041823     0.001749

   Pkt Size Avg  Fwd Seg Size Avg  Bwd Seg Size Avg  Subflow Fwd Pkts  \
0      0.039893          0.150578          0.015617          0.096491

   Subflow Fwd Byts  Subflow Bwd Pkts  Init Fwd Win Byts  Init Bwd Win Byts  \
0          0.146035          0.049494           0.004469           0.009522

   Fwd Act Data Pkts  Active Mean  Active Std  Active Min  Idle Mean  \
0           0.100457     0.298335         0.0    0.362174   0.099214

   Idle Std  Idle Max  Idle Min   Label
0       0.0  0.099214  0.099214  Benign
(1048397, 46)
dtypes:
Dst Port                    float64
Timestamp                   float64
Flow Duration               float64
Tot Fwd Pkts                float64
Tot Bwd Pkts                float64
TotLen Bwd Pkts             float64
Fwd Pkt Len Max             float64
Fwd Pkt Len Mean            float64
Fwd Pkt Len Std             float64
Bwd Pkt Len Max             float64
Bwd Pkt Len Mean            float64
Flow Byts/s                 float64
Flow Pkts/s                 float64
Flow IAT Std                float64
Flow IAT Max                float64
Flow IAT Min                float64
Fwd IAT Tot                 float64
Fwd IAT Mean                float64
Fwd IAT Std                 float64
Fwd IAT Max                 float64
Fwd IAT Min                 float64
Bwd IAT Tot                 float64
Bwd IAT Std                 float64
Bwd IAT Max                 float64
Bwd Pkts/s                  float64
Pkt Len Max                 float64
Pkt Len Mean                float64
Pkt Len Std                 float64
Pkt Len Var                 float64
Pkt Size Avg                float64
Fwd Seg Size Avg            float64
Bwd Seg Size Avg            float64
Subflow Fwd Pkts            float64
Subflow Fwd Byts            float64
Subflow Bwd Pkts            float64
Init Fwd Win Byts           float64
Init Bwd Win Byts           float64
Fwd Act Data Pkts           float64
Active Mean                 float64
Active Std                  float64
Active Min                  float64
Idle Mean                   float64
Idle Std                    float64
Idle Max                    float64
Idle Min                    float64
Label                string[python]
dtype: object
```

```
[ ]:
```