**Exploratory Data Analysis**

1. We are Using the data from Canadian Institute for Cybersecurity (CIC) project funded by Canadian Internet Registration Authority (CIRA) data set <https://www.unb.ca/cic/datasets/dohbrw-2020.html>
2. It has 3 different data set Non-DoH, malicious DoH, Bengine DoH so we merged those data. The Non-DoH and Bengine data is generated by using the simulated web browsers and the Malicious DoH is generated by using the combination of tools used to create DoH tunneling.
3. Importing data
   1. Loaded the data using the python library pandas and display the data set information using the shape(), info(), describe() functions.

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

* 1. Checking for null values duplicate values and removing if there are any.

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

* 1. Visualization

sns.violinplot(data=df, x="FlowBytesReceived", y="Label", scale="count")

sns.violinplot(data=df, x="PacketLengthMean", y="Label", scale="count")

sns.violinplot(data=df, x="PacketLengthVariance", y="Label", scale="count")

sns.boxplot (data=df, x="FlowBytesReceived", y="Label", scale="count")

sns.boxplot(data=df, x="PacketLengthMean", y="Label", scale="count")

sns.boxplot(data=df, x="PacketLengthVariance", y="Label", scale="count")

Chart, box and whisker chart

Description automatically generated Chart, box and whisker chart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated with medium confidence

Chart

Description automatically generatedChart

Description automatically generated

sns.histplot(data=df, x='Duration', hue='Label')

Chart, histogram

Description automatically generated

sns.histplot(data=df, x='ResponseTimeTimeSkewFromMedian', hue='Label')

Chart, histogram

Description automatically generated

sns.kdeplot(data=df, x="PacketLengthVariance", hue="Label")

Chart

Description automatically generated

* 1. Co-relation

Co-relation between the malicious-DoH and features

A picture containing text

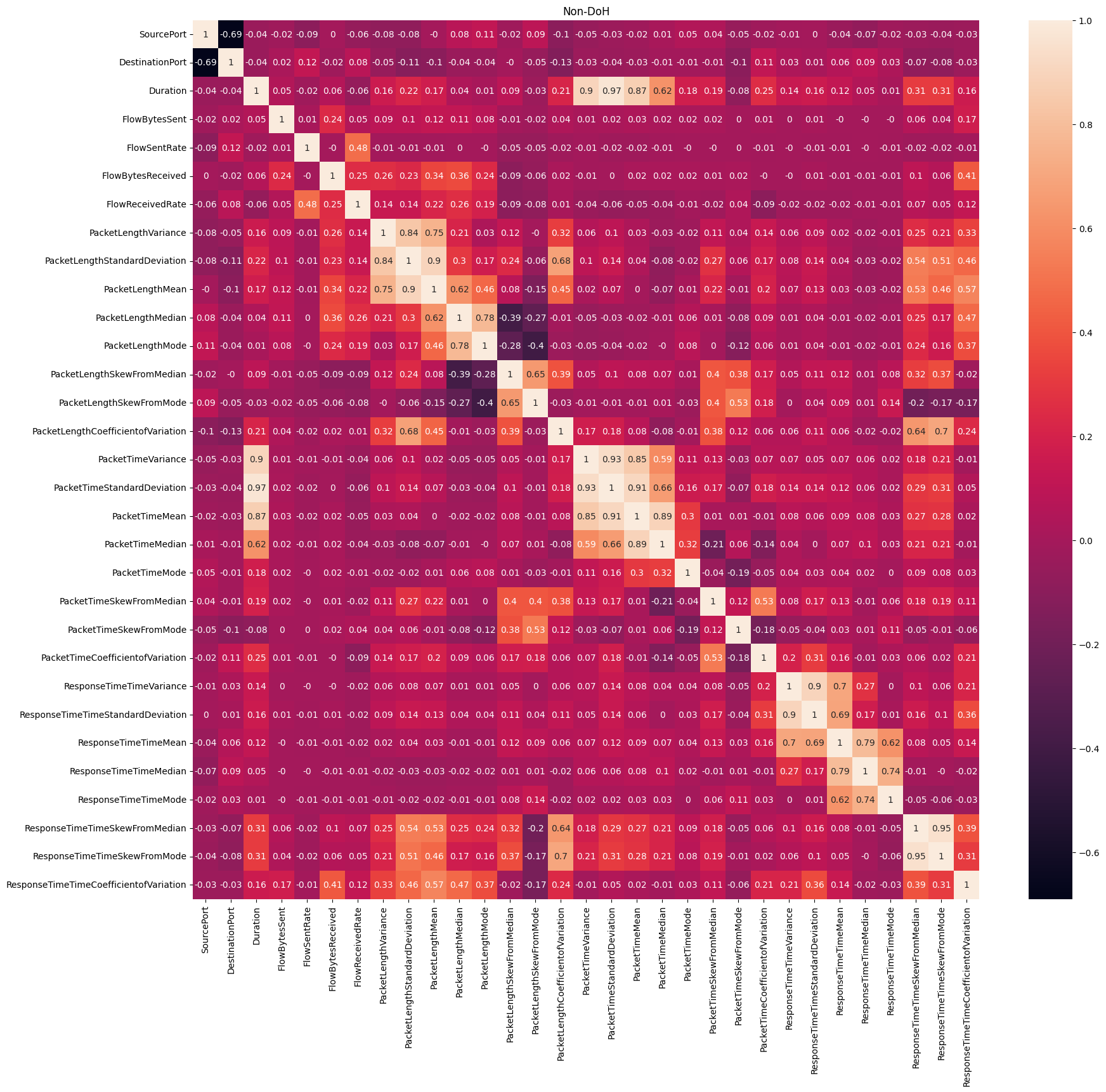
Description automatically generated

Corelation between Bengine DoH

A picture containing text

Description automatically generated

Corelation for Non-DoH.



* 1. Outliers

Chart, scatter chart

Description automatically generated

* 1. Data training

