Chapter 4

**Data preprocessing, analysis, visualization and feature engineering**

* 1. **Original shape of the dataset: -** (9167581, 59)

# Checking and removing duplicate records: -

* + - 310 duplicate records were fetched, which were removed. Thus, the new shape of the dataset is (9167271, 59).
    - 0.0042% of duplicate records for Label=Benign were removed.
    - 0.0016% of duplicate records for Label=DDOS-NTP were removed.
    - 0.00016% of duplicate records for ClassLabel=DDOS were removed.
    - As the result, it was observed that very small proportion of records were removed from the above category of records in the dataset. Thus, the overall distribution of records with respect to Label and ClassLabel have remained the same.

# Summarized view of distribution of data plotted on log scale for each independent feature: -

* + - Matplotlib library was used to plot the distribution of all features with chart type: histogram. But, due to large difference in scale, patterns were not observed.
    - Thus, again the histograms were plotted using log scale which helped to find pattern of distribution for each feature in the dataset.

Figure 4.3.1 Histogram of all independent features plotted on normal scale.

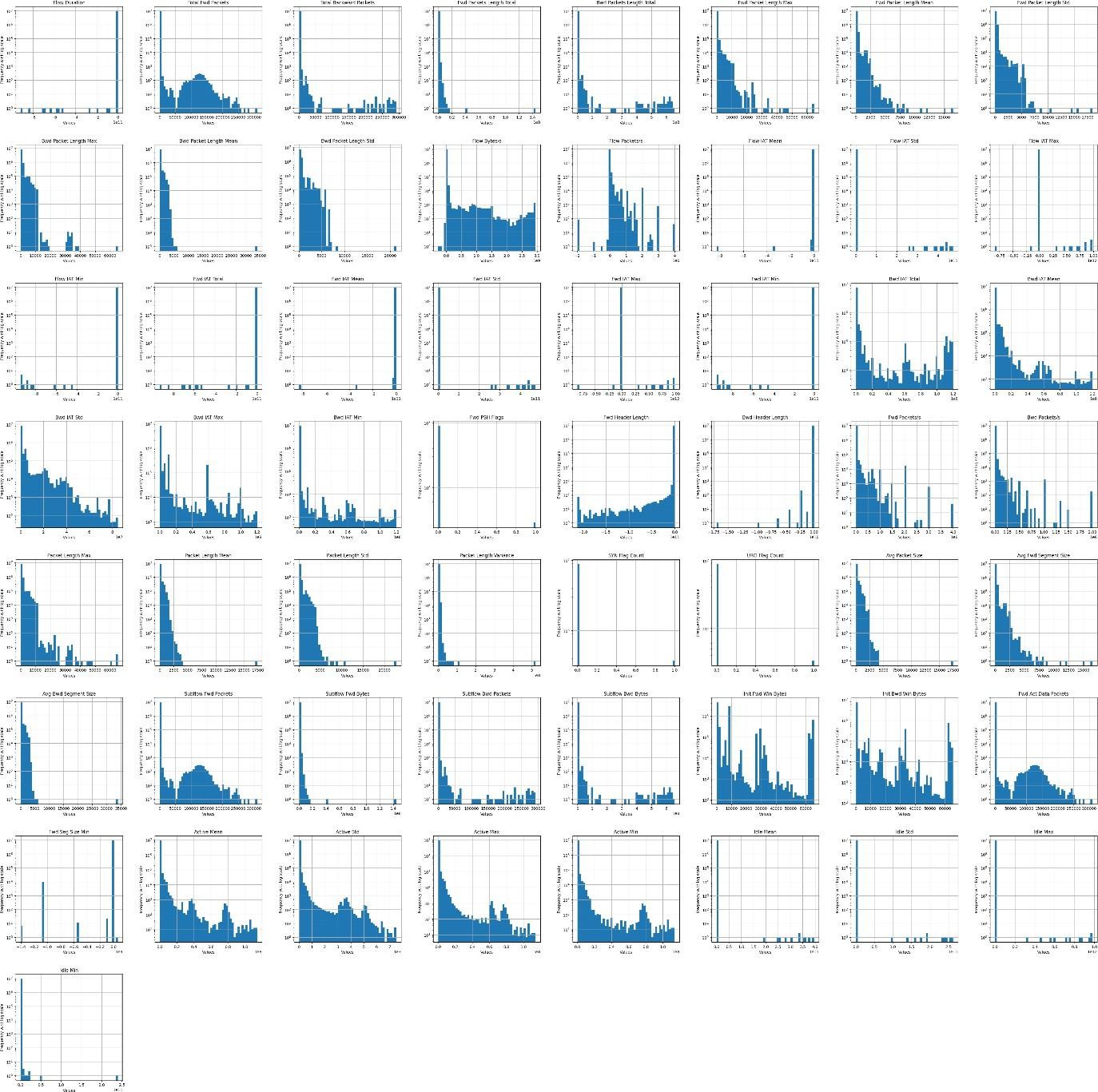


Figure 4.3.2 Histogram of all independent features plotted on log scale.

# Distribution of datapoints plotted on log scale for each independent feature: -

Flow Duration: The duration of the flow

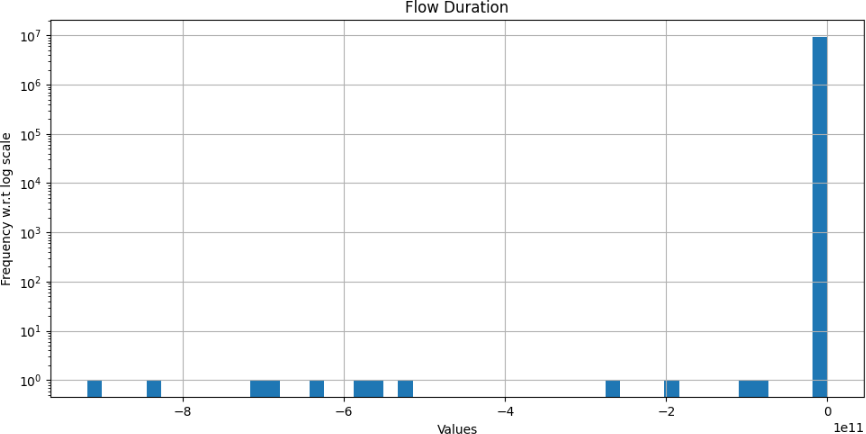


Figure 4.4.1 Histogram of Flow Duration plotted on log scale

* + - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.
    - Peak was observed at extreme right, Flow Duration=0.
    - There are some scattered bins of count=1.

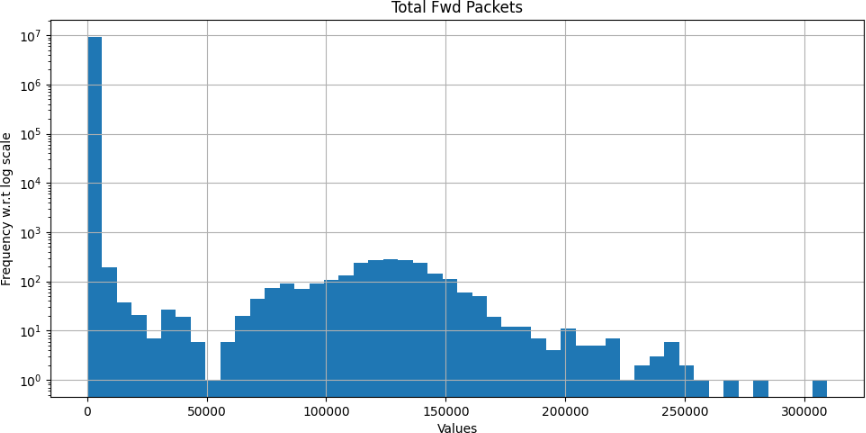
Total Fwd Packets: Total number of forward packets

Figure 4.4.2 Histogram of Total Fwd Packets plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from the left, after which we saw sharp decline.
    - There is another small peak around Total Fwd Packets=125000, but it is in plateau shape. Thus, we see many values around 125000.
    - The first bin (Peak) is in the range around 0 to 6250.
    - After the second bin, there is consistent decline.
    - Since there are two peaks at significant distance apart, we can also call the graph bi-modal.
    - We observed value for Total Fwd Packets>300000. This may indicate outlier in the data.

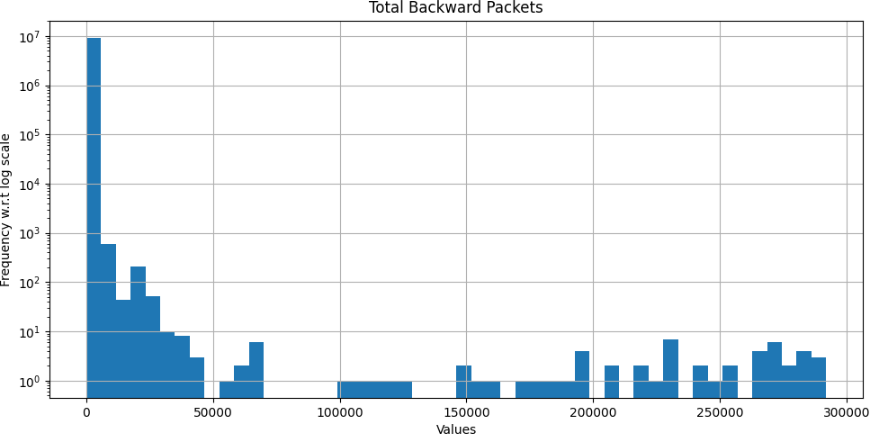
Total Backward Packets: Total number of backward packets

Figure 4.4.3 Histogram of Total Backward Packets plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from left, Total Backward Packets: 0 to 6250.
    - After the peak, there is significant decline in results.
    - Some records were observed at regular intervals but with very less frequency.

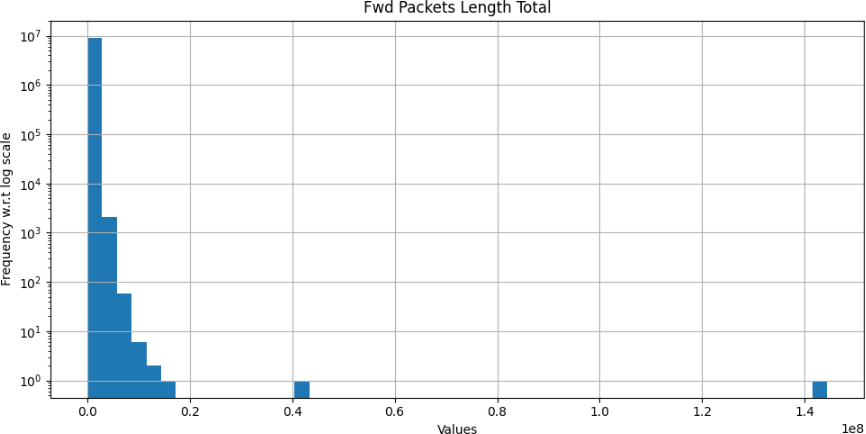
Fwd Packets Length Total: Total length of forward packets

Figure 4.4.4 Histogram of Fwd Packets Length Total plotted on log scale

* + - Peak was observed on first bin from left.
    - Most values are stacked on the left side of X-axis and they continuously decline as we move towards right hand side of X-axis.
    - There are a couple of observations at a distance on right hand side after long gap. They may indicate outliers in the data.

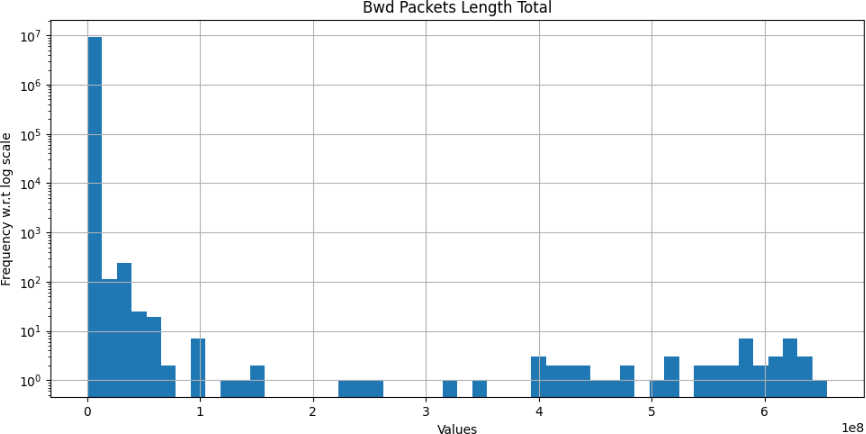
Bwd Packets Length Total: Total length of backward packets

Figure 4.4.5 Histogram of Bwd Packets Length Total plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from left.
    - After the peak, there is significant decline in results.
    - There are some observations spread out on X-axis, but all have frequency less than 10.

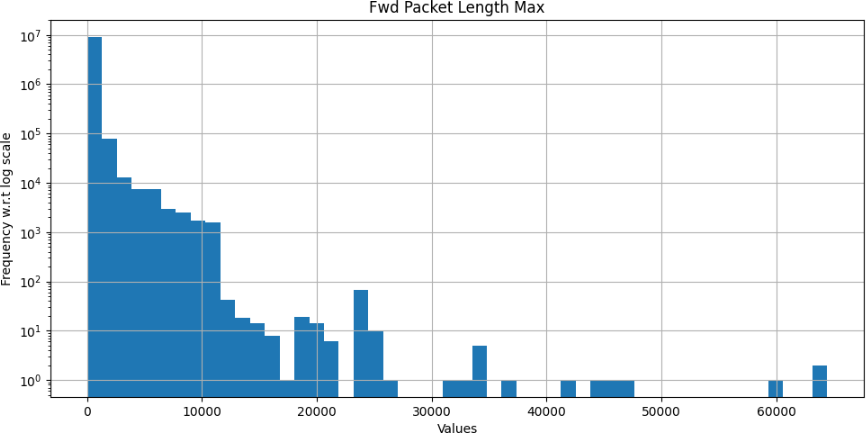
Fwd Packet Length Max: Maximum length of forward packets

Figure 4.4.6 Histogram of Fwd Packet Length Max plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from left.
    - The first bin (Peak) is in the range around 0 to 1250.
    - Most number of observations lie between Fwd Packet Length Max>0 and Fwd Packet Length Max<10000.
    - A small peak was observed around Fwd Packet Length Max>20000 and Fwd Packet Length Max<300000. However, the frequency is relatively very less compared to the peak observed in first bin.
    - There are some observations around Fwd Packet Length Max=60000. This may indicate outliers in the data.

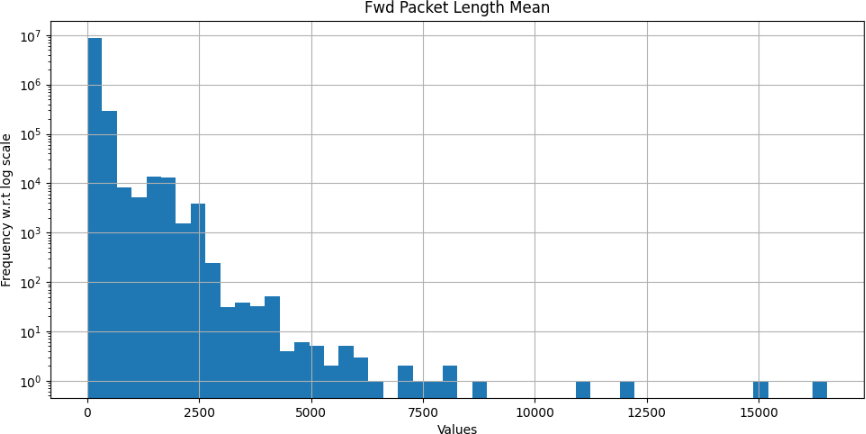
Fwd Packet Length Mean: Mean length of forward packets

Figure 4.4.7 Histogram of Fwd Packet Length Mean plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from left.
    - Most number of observations lie between Fwd Packet Length Mean>=0 and Fwd Packet Length Mean<=2500.
    - There are some small number of observations around Fwd Packet Length Mean=15000 and above. This may indicate outliers in the data.

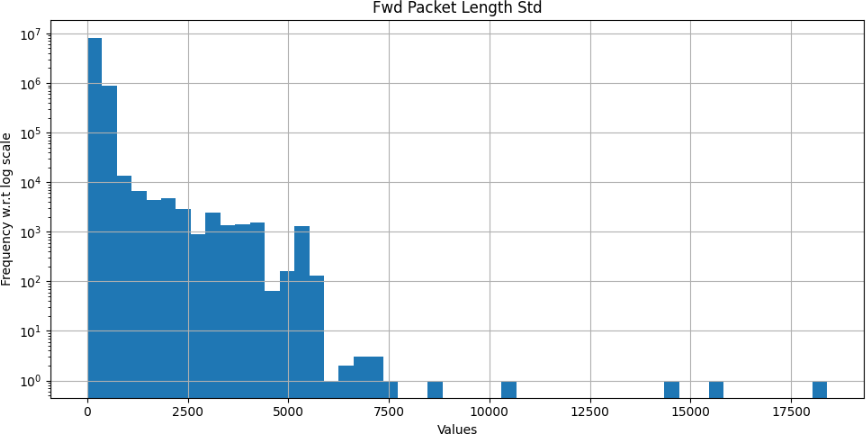
Fwd Packet Length Std: Standard deviation length of forward packets

Figure 4.4.8 Histogram of Fwd Packet Length Std plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from left.
    - Most number of observations lie between Fwd Packet Length Std>=0 and Fwd Packet Length Std<=5000.
    - There are some very small number of observations at Fwd Packet Length Std>7500. This may indicate outliers in the data.

Bwd Packet Length Max: Maximum length of backward packets

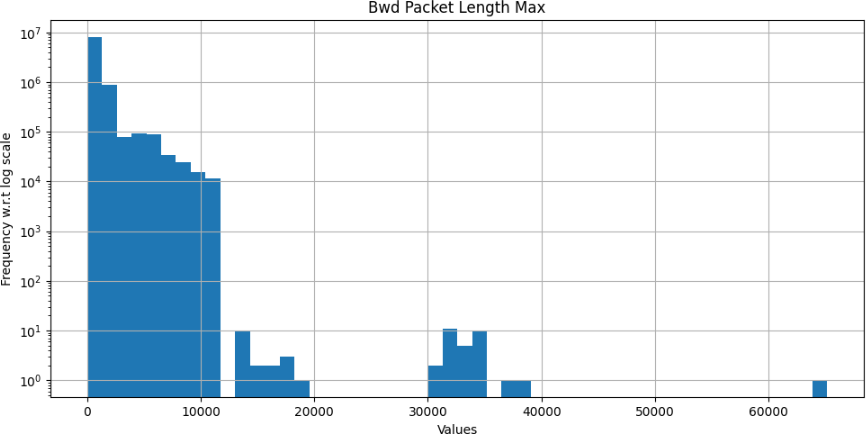


Figure 4.4.9 Histogram of Bwd Packet Length Max plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from left. Peak lies around Bwd Packet Length Max>=0 and Bwd Packet Length Max<=1250.
    - Most number of observations lie between Bwd Packet Length Max>=0 and Bwd Packet Length Max<=10000.
    - There are few observations in the range: - Bwd Packet Length Max>=11250 and Bwd Packet Length Max<=20000, Bwd Packet Length Max>=30000 and Bwd Packet Length Max<=35000.
    - There is an observation at Bwd Packet Length Max>60000. This may indicate outliers in the data.

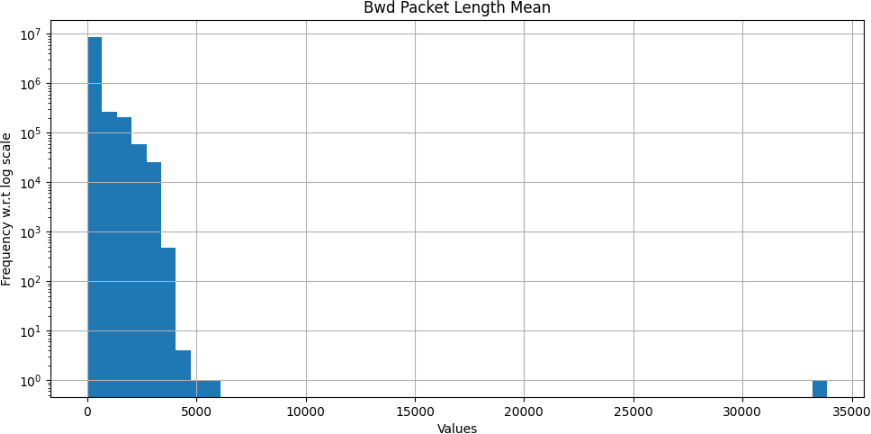
Bwd Packet Length Mean: Mean length of backward packet

Figure 4.4.10 Histogram of Bwd Packet Length Mean plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from left. Peak lies around Bwd Packet Length Mean>=0 and Bwd Packet Length Mean<=666.67.
    - After the peak, there is significant decline in results.
    - Between Bwd Packet Length Mean=0 and Bwd Packet Length Mean=5000, we observed J- shaped graph.
    - There is an observation at Bwd Packet Length Mean=35000. This may indicate outliers in the data.

Bwd Packet Length Std: Standard deviation length of backward packets

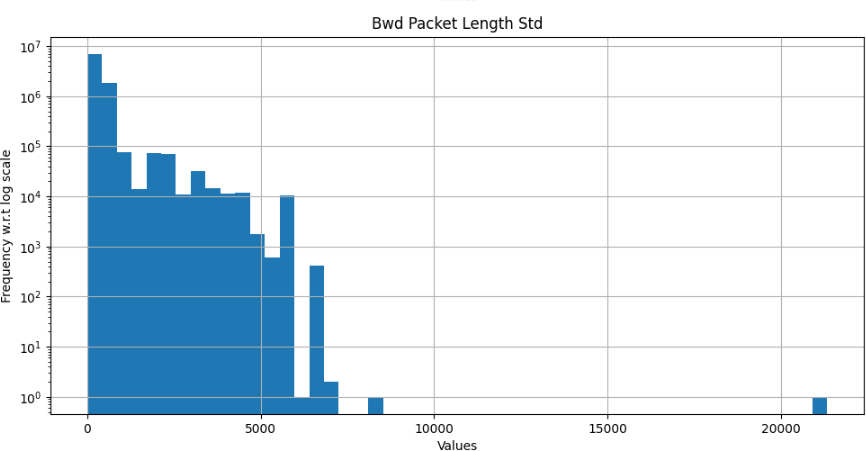


Figure 4.4.11 Histogram of Bwd Packet Length Std plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed on first bin from left. Peak lies around Bwd Packet Length Std>=0 and Bwd Packet Length Std<=416.67.
    - After the peak, there is significant decline in results.
    - There is plateau region observed around Bwd Packet Length Std>=2083 and Bwd Packet Length Std<=2500.
    - There is another plateau region observed (smaller than the above) around Bwd Packet Length Std>=3750 and Bwd Packet Length Std<=4166.
    - There is an observation at Bwd Packet Length Std>20000. This may indicate outliers in the data.

Flow Bytes/s: Flow bytes per second

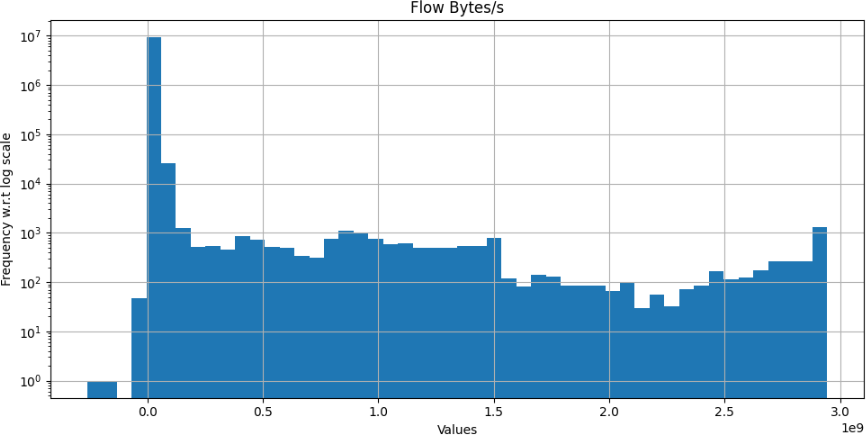


Figure 4.4.12 Histogram of Flow Bytes/s plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Flow Bytes/s=0.
    - After the peak, there is consistent decline in results.
    - Towards right hand side of the graph, there is increase in number of observations compared to other bins prior to it excluding the peak.
    - Between the two extremes of the graph there were some plateau regions.
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.

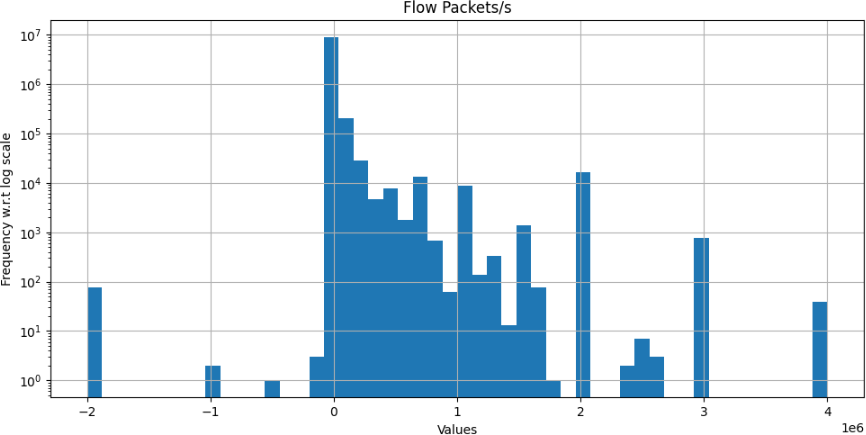
Flow Packets/s: Flow packets per second

Figure 4.4.13 Histogram of Flow Packets/s plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Flow Packets/s=0
    - After the peak, there is consistent decline in results.
    - At Flow Packets/s=2 and Flow Packets/s=3, there relatively small peaks.
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.

Flow IAT Mean: Mean time between flows

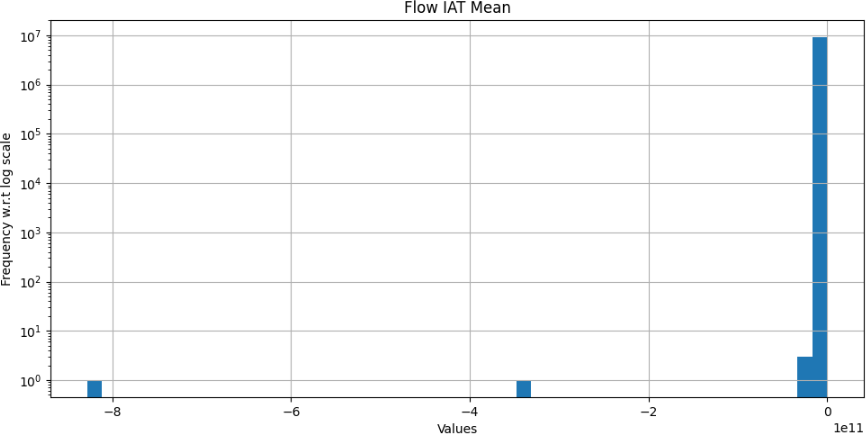


Figure 4.4.14 Histogram of Flow IAT Mean plotted on log scale

* + - Peak was observed at Flow IAT Mean=0.
    - Most values are concentrated in bin represented by the peak.
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.

Flow IAT Std: Standard deviation of time between flows

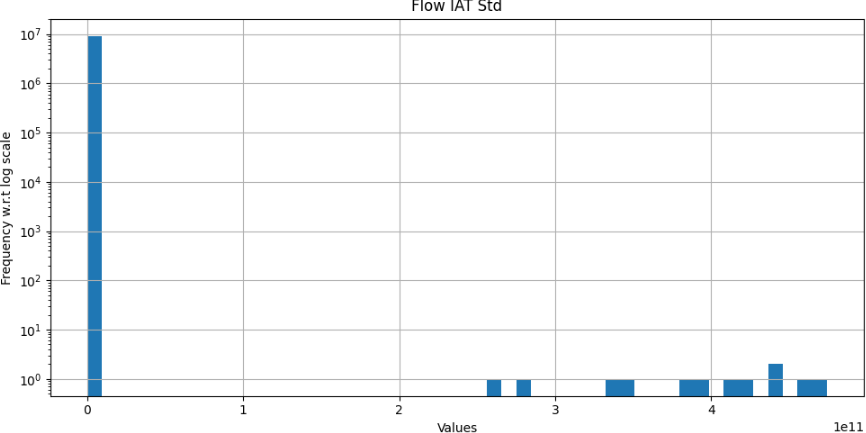


Figure 4.4.15 Histogram of Flow IAT Std plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Flow IAT Std=0.
    - Most values are concentrated in bin represented by the peak.
    - There are a few observations in the range: - Flow IAT Std>=2 and Flow IAT Std<=3, Flow IAT Std>=3 and Flow IAT Std<=4 and Flow IAT Std>4.

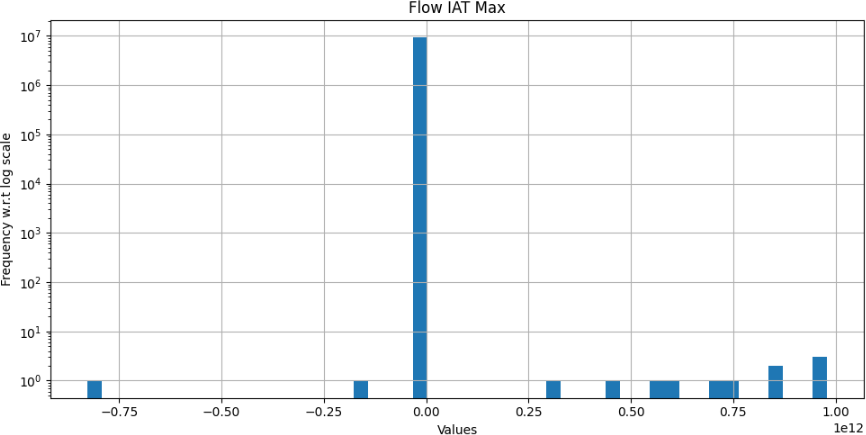
Flow IAT Max: Maximum time between flows

Figure 4.4.16 Histogram of Flow IAT Max plotted on log scale

* + - Peak was observed around Flow IAT Max=0.
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.
    - On X-axis values lie in the range -1.0 to +1.0

Flow IAT Min: Minimum time between flows

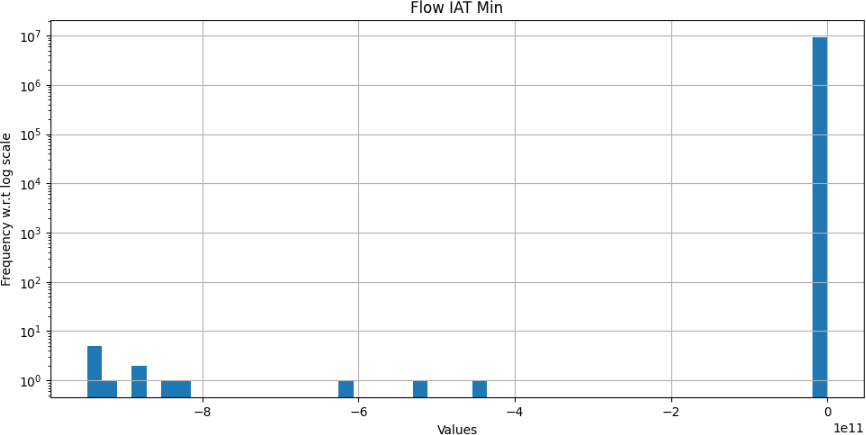


Figure 4.4.17 Histogram of Flow IAT Min plotted on log scale

* + - Peak was observed around Flow IAT Mean=0.
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.

Fwd IAT Total: Total time between forward packets

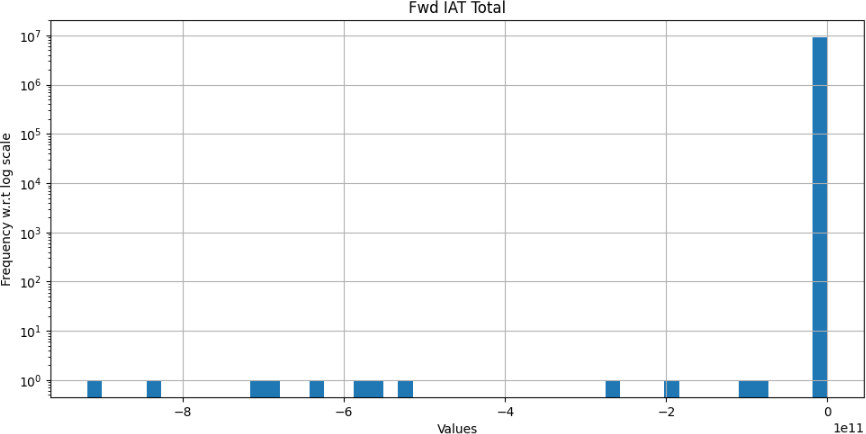


Figure 4.4.18 Histogram of Fwd IAT Total plotted on log scale

* + - Peak was observed around Fwd IAT Total=0.
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.

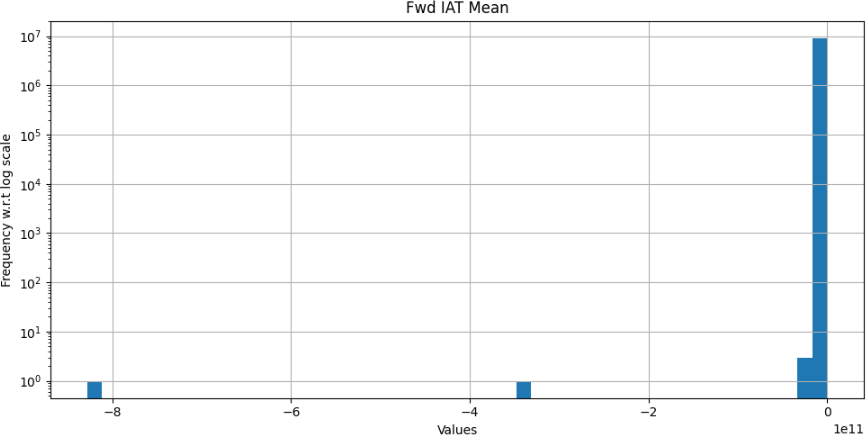
Fwd IAT Mean: Mean time between forward packets

Figure 4.4.19 Histogram of Fwd IAT Mean plotted on log scale

* + - Peak was observed around Fwd IAT Mean=0.
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.

Fwd IAT Std: Standard deviation of time between forward packets

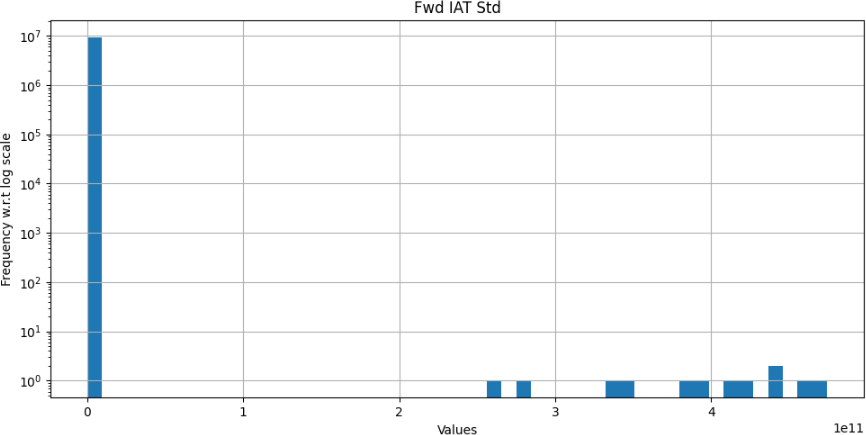


Figure 4.4.20 Histogram of Fwd IAT Std plotted on log scale

* + - Peak was observed around Fwd IAT Std=0.
    - There are small number of observations in the range: Fwd IAT Std>=2 and Fwd IAT Std<=3, Fwd IAT Std>=3 and Fwd IAT Std<=4, Fwd IAT Std>4.

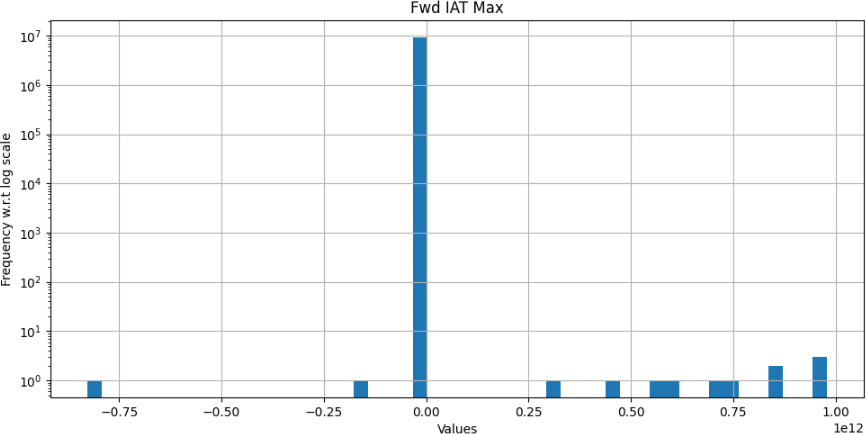
Fwd IAT Max: Maximum time between forward packets

Figure 4.4.21 Histogram of Fwd IAT Max plotted on log scale

* + - Peak was observed around Fwd IAT Max=0.
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.
    - There are scattered but very small number of observations between Fwd IAT Max=0.0 and Fwd IAT Max=1.0

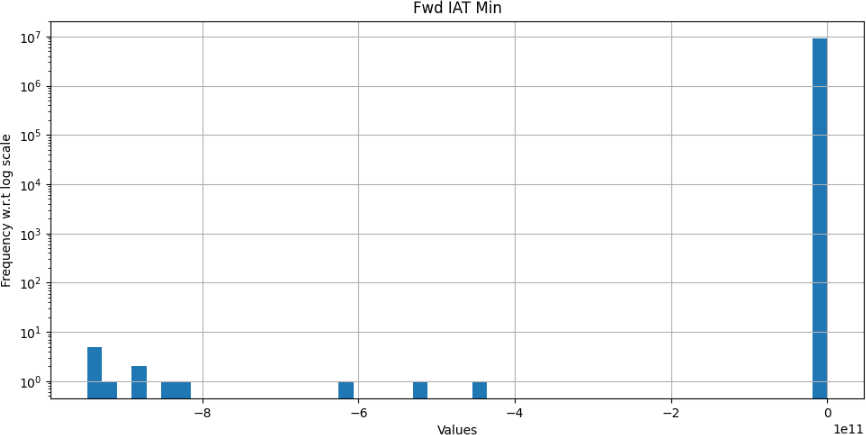
Fwd IAT Min: Minimum time between forward packets

Figure 4.4.22 Histogram of Fwd IAT Min plotted on log scale

* + - Peak was observed around Fwd IAT Min=0
    - We observed negative values on X-axis, thus, we need to check the actual values under the column to determine if data is accurate or invalid.

Bwd IAT Total: Total time between backward packets



Figure 4.4.23 Histogram of Bwd IAT Total plotted on log scale

* + - Peak was observed around Bwd IAT Total=0.
    - After the peak, there is consistent decline in results.
    - There are relatively smaller peaks at Bwd IAT Total=0.6 and Bwd IAT Total=1.125
    - There was a plateau region observed between Bwd IAT Total>=0.625 and Bwd IAT Total<=0.675

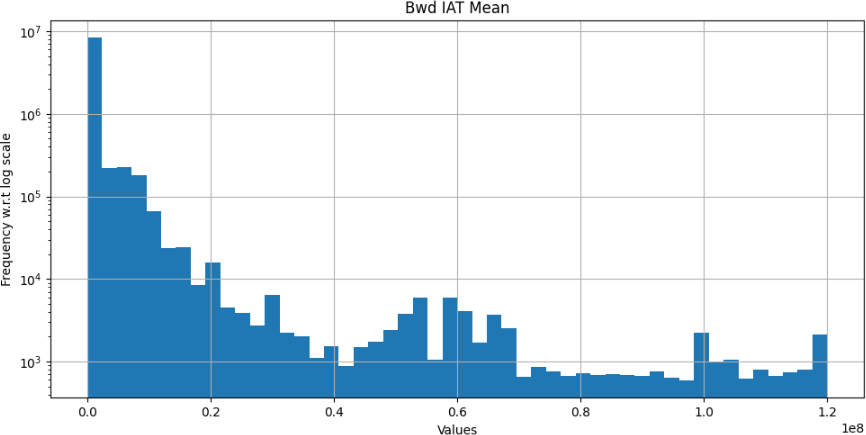
Bwd IAT Mean: Mean time between backward packets

Figure 4.4.24 Histogram of Bwd IAT Mean plotted on log scale

* + - Peak was observed around Bwd IAT Mean=0.
    - After the peak, there is consistent decline in results.
    - Most observations are stacked on left side of the graph, near the peak.
    - On X-axis values lie in the range 0.0 to +1.2

Bwd IAT Std: Standard deviation of time between packets

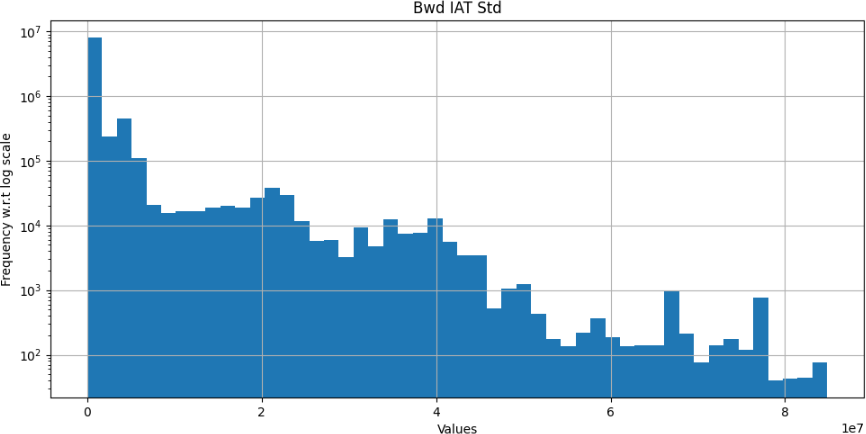


Figure 4.4.25 Histogram of Bwd IAT Std plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Bwd IAT Std=0
    - After the peak, there is consistent decline in results.
    - There was plateau region observed between Bwd IAT Std>=1.169 and Bwd IAT Std=2.
    - As the value of Bwd IAT Std increases, the size of bins decreases. In between there are a few exceptions where size of bin is greater than their neighbors.

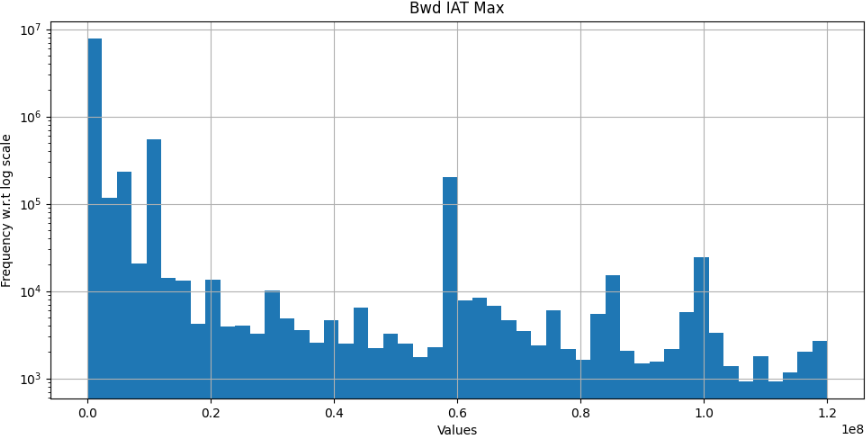
Bwd IAT Max: Maximum time between packets

Figure 4.4.26 Histogram of Bwd IAT Max plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Bwd IAT Max=0.0
    - After the peak, there is consistent decline in results.
    - There are relatively smaller peaks at Bwd IAT Max=0.125 and Bwd IAT Max=0.575
    - Since there are multiple peaks at significant distance apart, we can also call the graph multi-modal.
    - The bins prior and after all three peaks are very small.

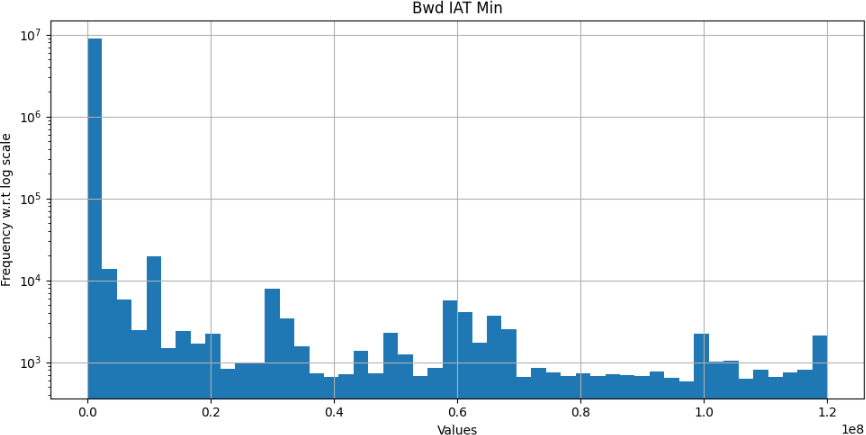
Bwd IAT Min: Minimum time between packets

Figure 4.4.27 Histogram of Bwd IAT Min plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Bwd IAT Min=0
    - After the peak, there is significant decline in results.
    - On X-axis values lie in the range 0.0 to 1.2

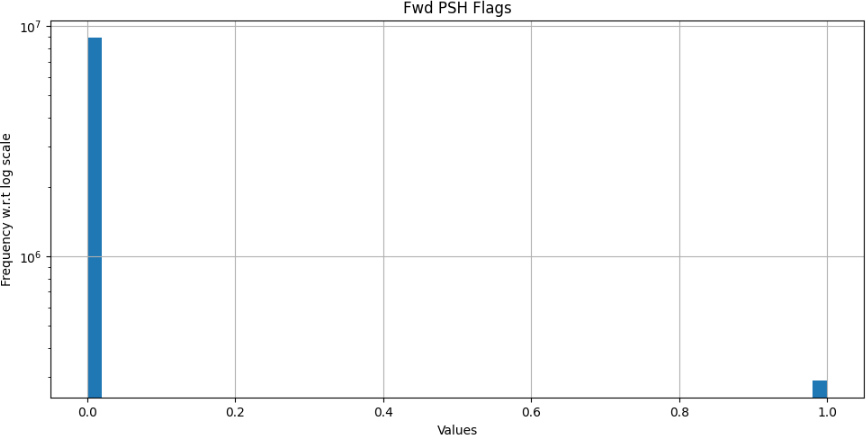
Fwd PSH Flags: Forward packets with PUSH flags

Figure 4.4.28 Histogram of Fwd PSH Flags plotted on log scale

* + - Most of the values are concentrated in the first bin at Fwd PSH Flags=0.0
    - There were few observations at Fwd PSH Flags=1.0. This may indicate outlier in the data.
    - There were no results between Fwd PSH Flags=0.0 and Fwd PSH Flags=1.0

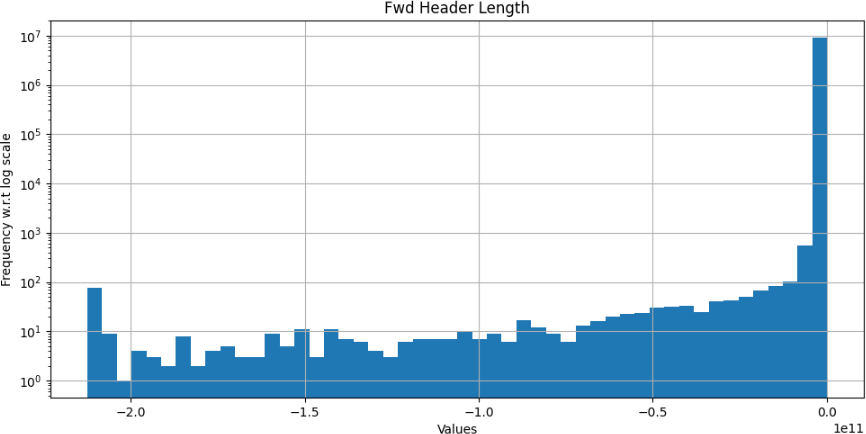
Fwd Header Length: Length of header in forward packets

Figure 4.4.29 Histogram of Fwd Header Length plotted on log scale

Figure 4.4.29 Histogram of Fwd Header Length plotted on log scale

* + - The distribution is skewed towards left: Negatively skewed.
    - Peak was observed around Fwd Header Length=0.0
    - There were no results for Fwd Header Length>0.0
    - There are relatively smaller size bins of left hand side of the peak.
    - On X-axis values lie in the range -2.0 to 0.0

Bwd Header Length: Length of header in backward packets

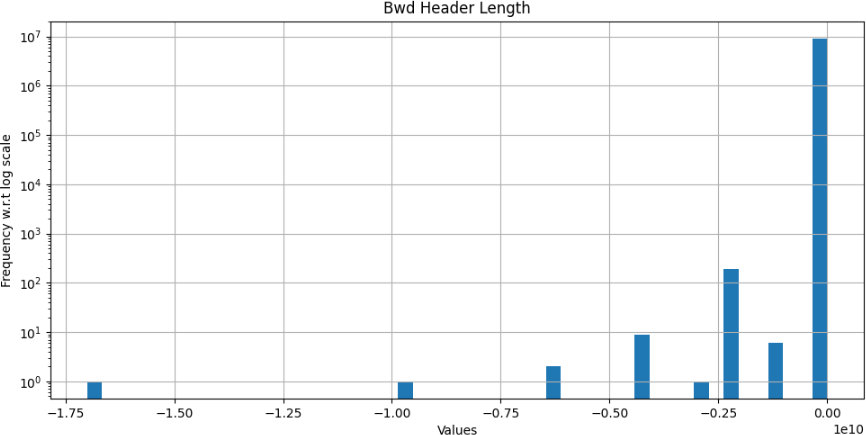


Figure 4.4.30 Histogram of Bwd Header Length plotted on log scale

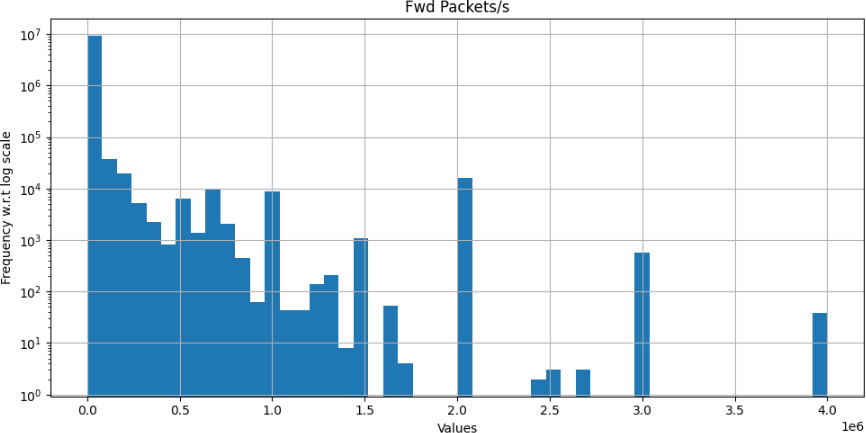
* + - Peak was observed around Bwd Header Length=0.0
    - Most values are concentrated at the peak.
    - There were few observations at Bwd Header Length=-1.75, -1, -0.6, -0.30
    - There were no results for Bwd Header Length>0.0
    - On X-axis values lie in the range -1.75 to 0.0 Fwd Packets/s: Forward packets per second

Figure 4.4.31 Histogram of Fwd Packets/s plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Fwd Packets/s=0
    - From Fwd Packets/s=0.0 to 1.5, the values are stacked to the right hand side of peak.
    - There are relatively smaller peaks at Fwd Packets/s= 2.0, 3.0, 4.0
    - There is a wide gap (no results) between Fwd Packets/s=3.0 and Fwd Packets/s=4.0
    - Most values are concentrated between Fwd Packets/s=0.0 and Fwd Packets/s=1.5. Between this range the graph also resembles to J-shaped graph.

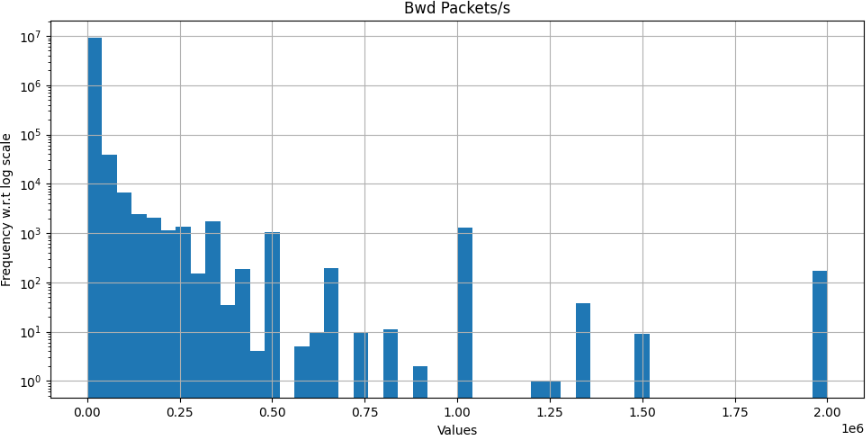
Bwd Packets/s: Backward packets per second

Figure 4.4.32 Histogram of Bwd Packets/s plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Bwd Packets/s=0.0
    - Most values are concentrated between Bwd Packets.s=0.0 and Bwd Packets/s=0.5. Between this range the graph also resembles to J-shaped graph.
    - There are relatively smaller peaks at Bwd Packets/s=0.5, 1.0 and 2.0
    - After Bwd Packets/s>=1.0, the bins are scattered and gaps were observed at irregular intervals on the x-axis.

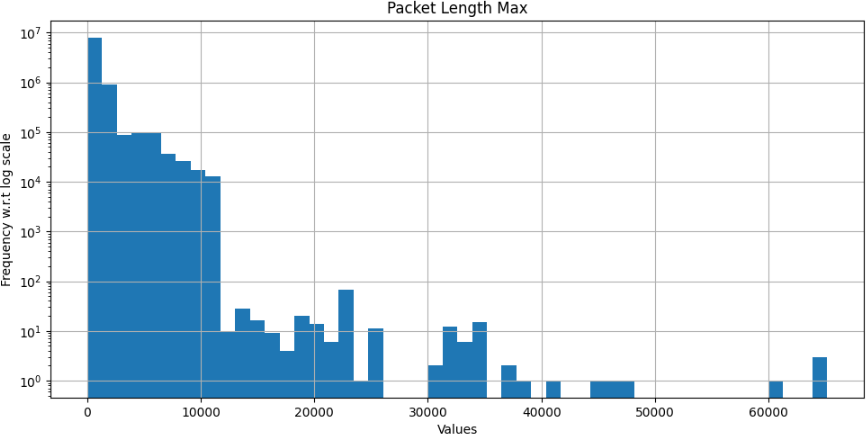
Packet Length Max: Maximum length of packets

Figure 4.4.33 Histogram of Packet Length Max plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Packet Length Max=0
    - After the peak, there is significant decline in results between Packet Length Max>=0 and Packet Length Max<=10000.
    - Between Packet Length Max=0 and Packet Length Max=10000, the graph also resemble to J-shaped graph.
    - Between Packet Length Max=10000 to 26000, the results are significantly lower than Packet Length Max=0 to 10000.
    - There were no results observed between Packet Length Max=26000 to 30000, 50000 to 60000.
    - There are some results observed between Packet Length Max=30000 to 50000.
    - There are small number of results observed for Packet Length Max>60000. This may indicate outlier in the data.

Packet Length Mean: Mean length of packets

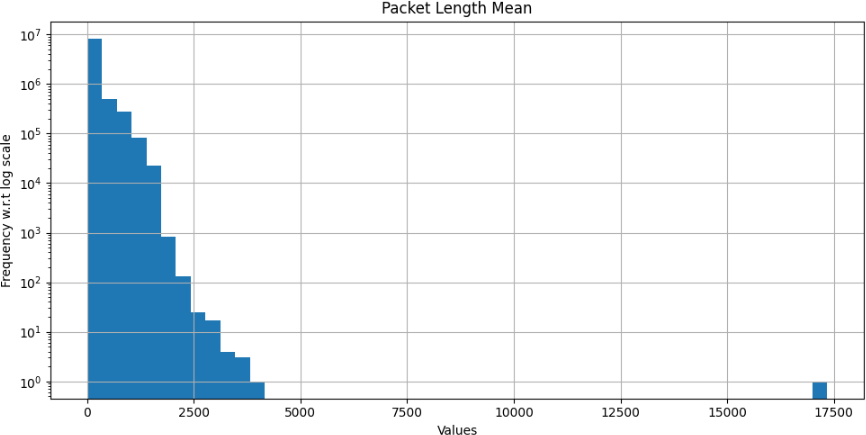


Figure 4.4.34 Histogram of Packet Length Mean plotted on log scale

* + - On X-axis, values lie in the range 0 to 17500.
    - The distribution is a J-shaped graph.
    - Peak was observed around Packet Length Mean=0.
    - All other bins are stacked against the peak on its right hand side.
    - There is a constant decline of results as we move towards right side of the graph.
    - The results are concentrated between Packet Length Mean>=0 and Packet Length Mean<5000.
    - There is a small observation at Packet Length Mean=17500. This may indicate an outlier in the data.

Packet Length Std: Standard deviation length of packets

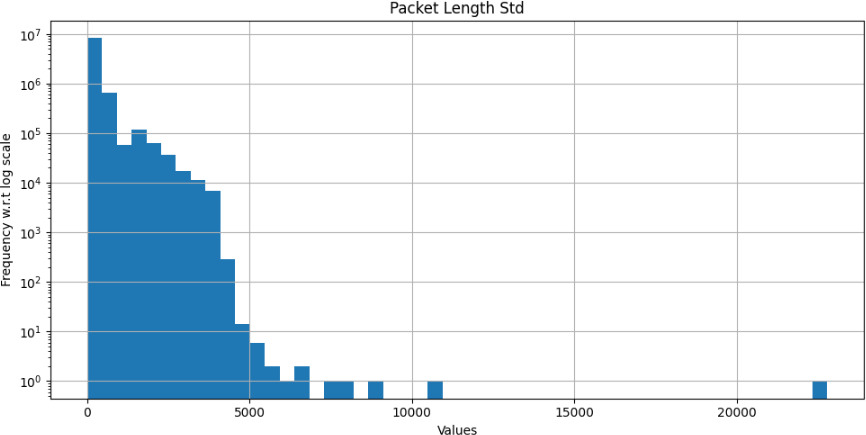


Figure 4.4.35 Histogram of Packet Length Std plotted on log scale

* + - The distribution is a J-shaped graph.
    - Peak was observed around Packet Length Std=0
    - All other bins are stacked against the peak on its right hand side.
    - Most values are concentrated between Packet Length Std>=0 and Packet Length Std<=5000.
    - There is an observation at Packet Length Std>20000. This may indicate an outlier in the data.
    - On X-axis, values lie in the range 0 to 20000.

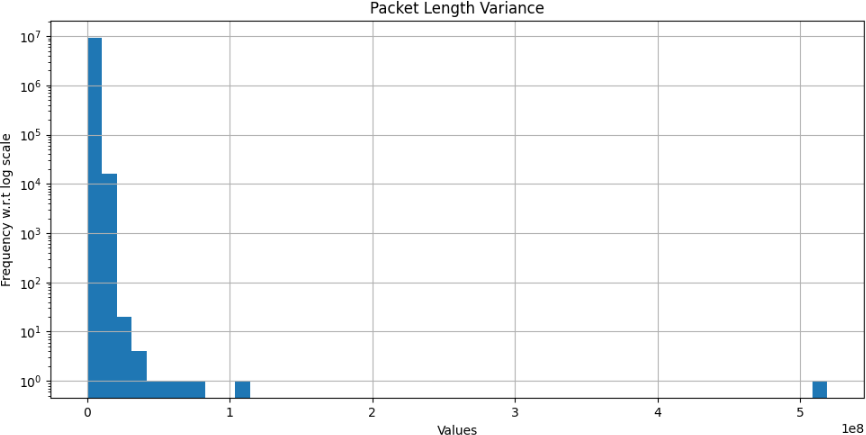
Packet Length Variance: Variance of length of packets

Figure 4.4.36 Histogram of Packet Length Variance plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed around Packet Length Variance=0
    - Most values are concentrated between Packet Length Variance>=0 and Packet Length Variance<=1.
    - There is an observation after long gap at Packet Length Variance>5. This may indicate an outlier in the data.

SYN Flag Count: Number of SYN flags

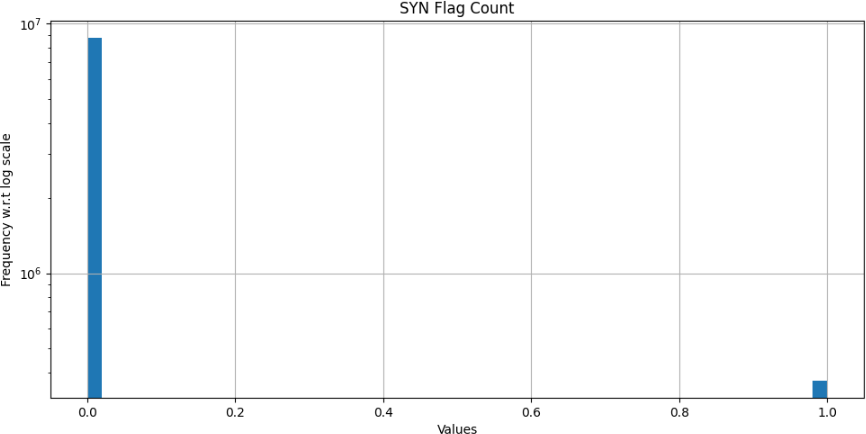


Figure 4.4.37 Histogram of SYN Flag Count plotted on log scale

* + - Peak was observed at SYN Flag Count=0.
    - Most values are concentrated at the peak.
    - There are a few observations at SYN Flag Count=1.0. This may indicate outlier in the data.

URG Flag Count: Number of URG flags

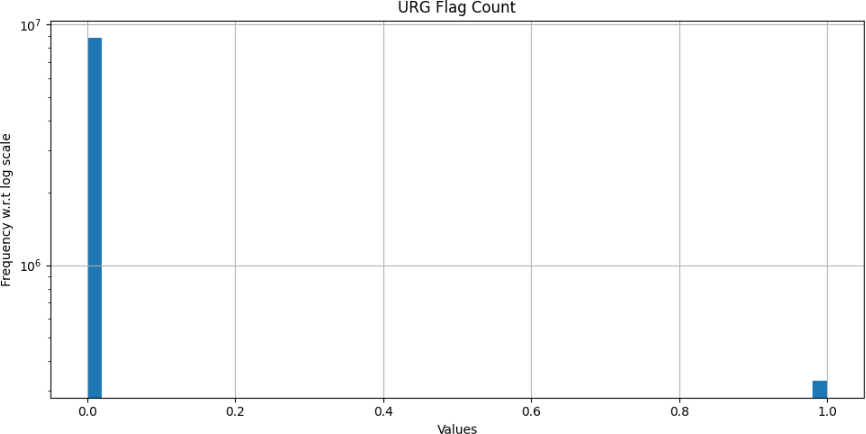


Figure 4.4.38 Histogram of URG Flag Count plotted on log scale

* + - Peak was observed at URG Flag Count=0.
    - Most values are concentrated at the peak.
    - There are a few observations at URG Flag Count=1.0. This may indicate outlier in the data.

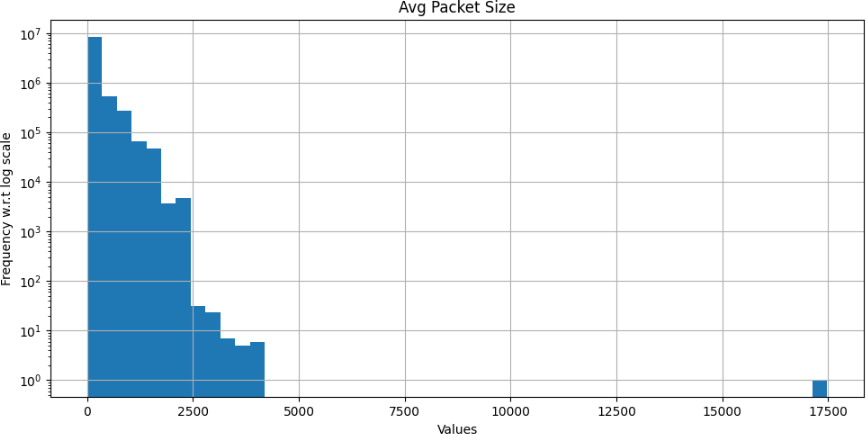
Avg Packet Size: Average packet size

Figure 4.4.39 Histogram of Avg Packet Size plotted on log scale

* + - The distribution is J-shaped graph.
    - Most of the values are stacked at left end and then it continuously declines as we move towards right hand side of the x-axis.
    - Peak was observed at Avg Packet Size=0.
    - Most values are concentrated between Avg Packet Size>=0 and Avg Packet Size<5000.
    - There were some values after a long gap between Avg Packet Size>5000 and Avg Packet Size

<=17500. This may indicate outlier in the data.

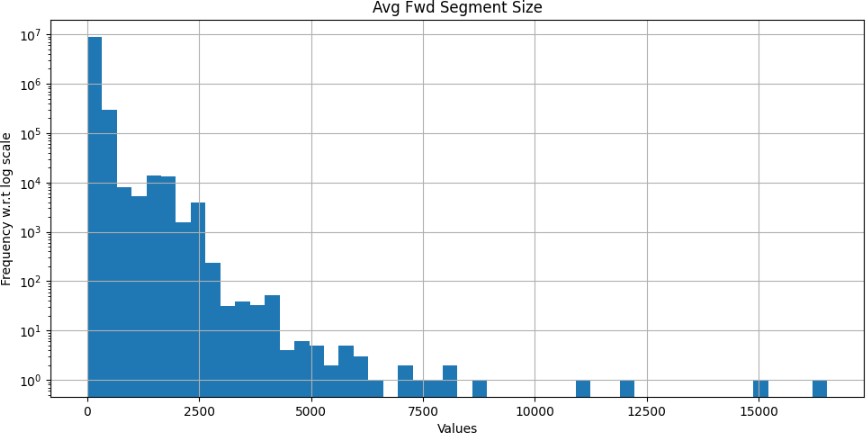
Avg Fwd Segment Size: Average forward segment size

Figure 4.4.40 Histogram of Avg Fwd Segment Size plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed at Avg Fwd Segment Size=0.
    - After the peak, there is consistent decline in results.
    - Most values are concentrated between Avg Fwd Segment Size>=0 and Avg Fwd Segment Size<=5000.
    - There were some values around Avg Fwd Segment Size=7500.
    - There were couple of values observed in range Avg Fwd Segment Size>10000 and Avg Fwd Segment Size<12500, Avg Fwd Segemnt Size>=15000. This may indicate outlier in the data.

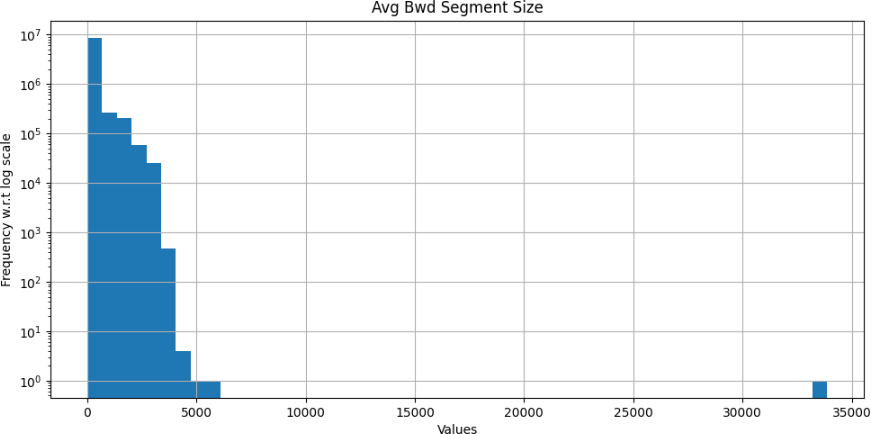
Avg Bwd Segment Size: Average backward segment size

Figure 4.4.41 Histogram of Avg Bwd Segment Size plotted on log scale

* + - The distribution is J-shaped graph.
    - Most of the values are stacked at left end and then it continuously declines as we move towards right hand side of the x-axis.
    - Peak was observed at Avg Bwd Segment Size=0.
    - Most values are concentrated between Avg Bwd Segment Size>=0 and Avg Bwd Segment Size<=5000.
    - There is a long gap observed after Avg Bwd Segment Size>5000.
    - On extreme right end side of the graph, between Avg Bwd Segment Size>=30000 and Avg Bwd Segment Size<=35000, few values were observed. This may indicate outlier in the data.

Subflow Fwd Packets: Subflow forward packets

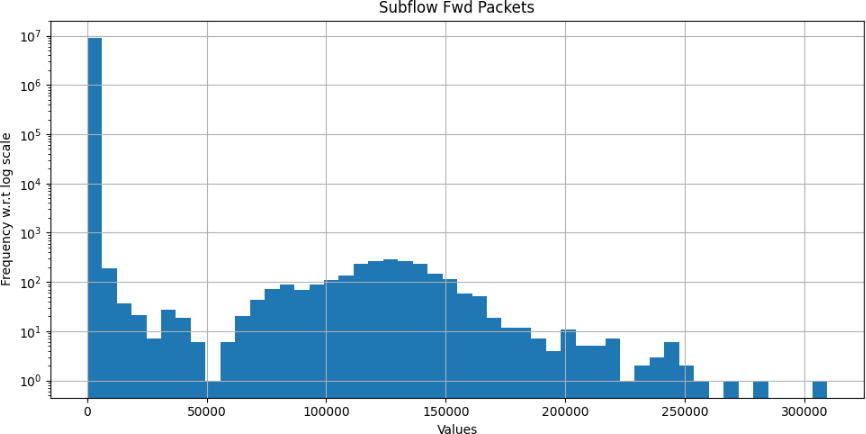


Figure 4.4.42 Histogram of Subflow Fwd Packets plotted on log scale

* + - Peak was observed at Subflow Fwd Packets=0.
    - After the peak, there is significant decline in results up to Subflow Fwd Packets=50000.
    - There is a plateau region observed between Subflow Fwd Packets>=100000 and Subflow Fwd Packets<=150000.
    - There were decline in the number of results observed after Subflow Fwd Packets>=150000.
    - There are many values between Subflow Fwd Packets>=50000 and Subflow Fwd Packets<=150000.
    - There is a value after Subflow Fwd Packets>300000. This may indicate outlier in the data.

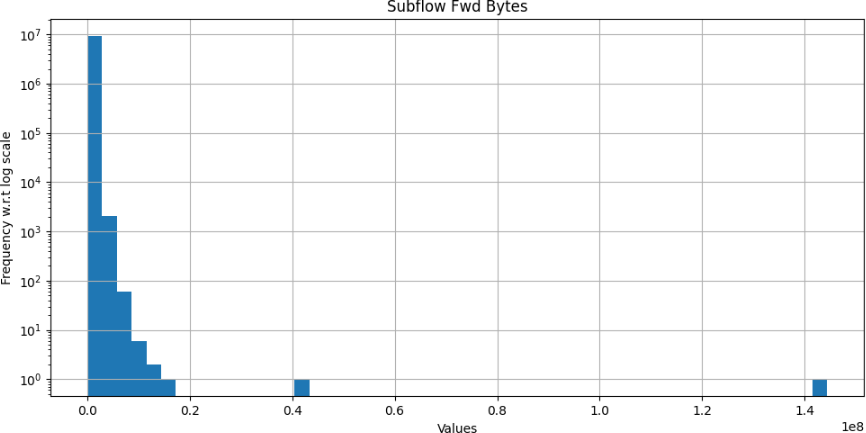
Subflow Fwd Bytes: Subflow forward bytes

Figure 4.4.43 Histogram of Subflow Fwd Bytes plotted on log scale

* + - The distribution is J-shaped graph.
    - Most of the values are stacked at left end and then it continuously declines as we move towards right hand side of the x-axis.
    - Most values are concentrated between Subflow Fwd Bytes>=0 and Subflow Fwd Bytes<0.2
    - There were couple of values observed around Subflow Fwd Bytes=0.4 and Subflow Fwd Bytes>1.4. This may indicate outlier in the data.

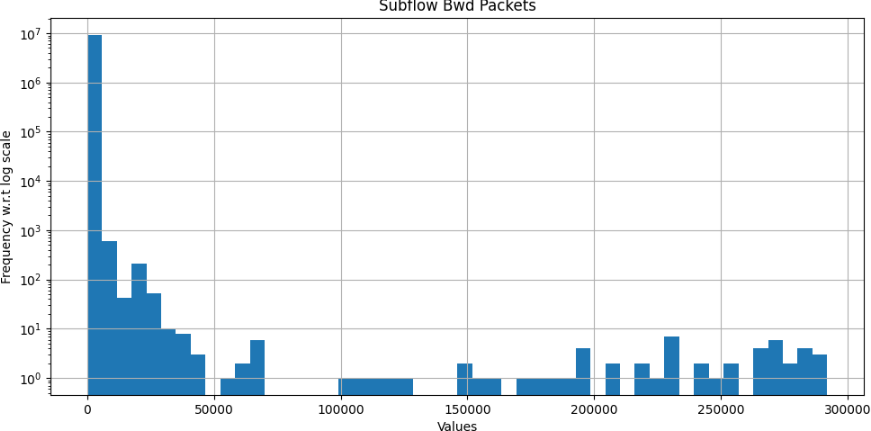
Subflow Bwd Packets: Subflow backward packets

Figure 4.4.44 Histogram of Subflow Bwd Packets plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed at Subflow Bwd Packets=0.
    - After the peak, there is consistent decline in results.
    - Most values are concentrated between Subflow Bwd Packets>=0 and Subflow Bwd Packets<=50000.
    - After Subflow Bwd Packets>50000, there are many small plateau regions at irregular gaps up to Subflow Bwd Packets<300000.

Subflow Bwd Bytes: Subflow backward bytes

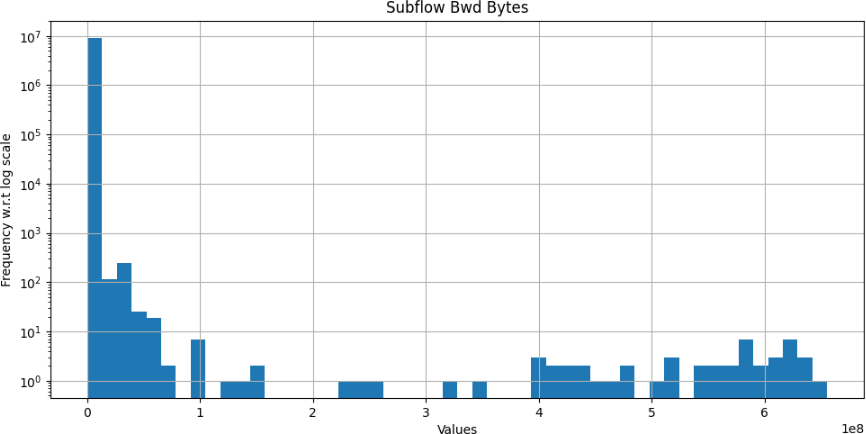


Figure 4.4.45 Histogram of Subflow Bwd Bytes plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed at Subflow Bwd Bytes=0.
    - Between Subflow Bwd Bytes>=0 and Subflow Bwd Bytes<=1, the graph appeared similar to J-shaped graph.
    - Most values are concentrated between Subflow Bwd Bytes>=0 and Subflow Bwd Bytes<=1.
    - There are some plateau regions on right hand side of the peak at irregular gaps.
    - On the X-axis values lie in the range 0 to 7.

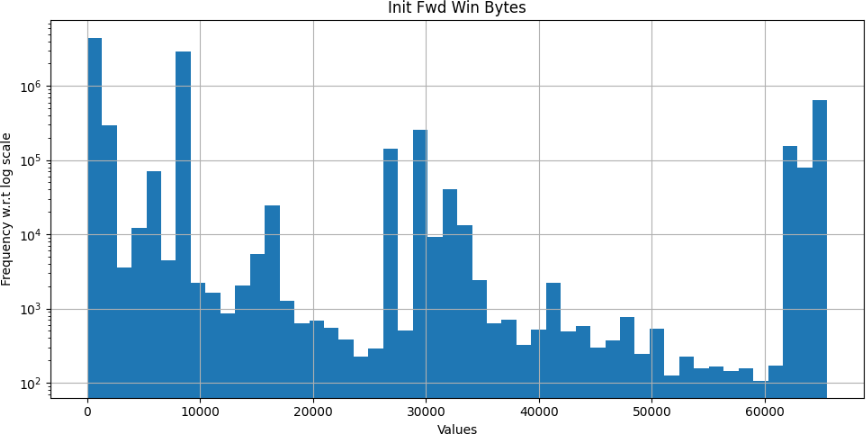
Init Fwd Win Bytes: Initial forward window size

Figure 4.4.46 Histogram of Init Fwd Win Bytes plotted on log scale

* + - There are two large peaks at Init Fwd Win Bytes=0 and Init Fwd Win Bytes=10000.
    - There are smaller peaks at Init Fwd Win Bytes=30000 and Init Fwd Win Bytes>60000.
    - Between the peaks, the frequency of bins is relatively very less.
    - There are no gaps in the results observed on X-axis of the graph.
    - Since the graph has multiple peaks, we can also call it multi-modal.
    - From broad overview, as we move from left to right hand side of the graph, the results decrease. But, due to tall peaks observed in between, we cannot conclude consistent decline of results.

Init Bwd Win Bytes: Initial backward window size

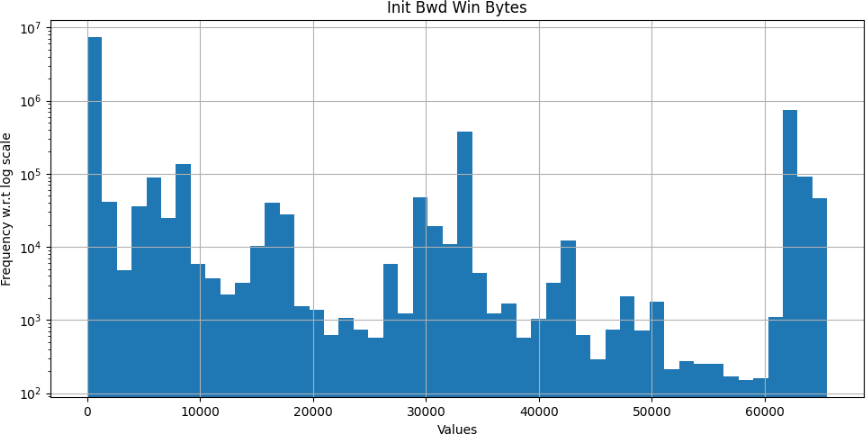


Figure 4.4.47 Histogram of Init Bwd Win Bytes plotted on log scale

* + - There are three main peaks from overall observation of the graph: Init Bwd Win Bytes=0, 30000, 60000.
    - The tallest peak was observed at Init Bwd Win Bytes=0, the second tallest was at Init Bwd Win Bytes=60000 and the smallest peak among the three was observed at Init Bwd Win Bytes=30000.
    - Between the peaks, the frequency of bins is relatively very less.
    - There are no gaps in the results observed on X-axis of the graph.
    - Since the graph has multiple peaks, we can also call it multi-modal.

Fwd Act Data Packets: Forward packets with actual data

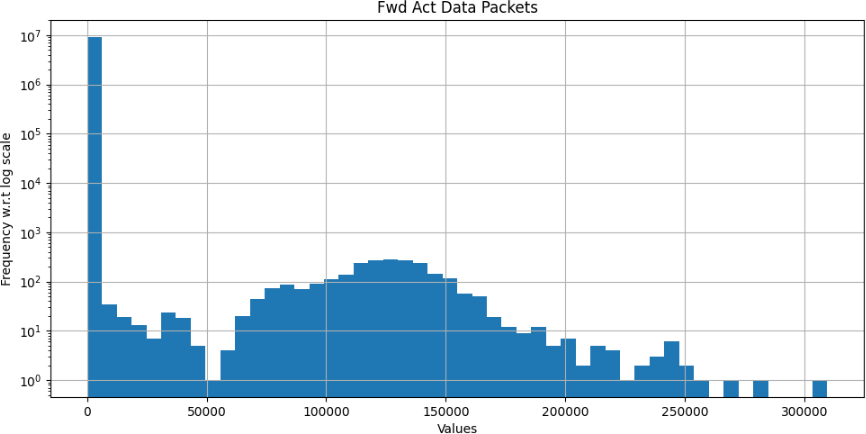


Figure 4.4.48 Histogram of Fwd Act Data Packets plotted on log scale

* + - Peak was observed at Fwd Act Data Packets=0.
    - After the peak, there is significant decline in results up to Fwd Act Data Packets=50000.
    - There is a plateau region observed between Fwd Act Data Packets>=100000 and Fwd Act Data Packets<=150000.
    - There are some values observed after Fwd Act Data Packets>300000. This may indicate outlier in the data.

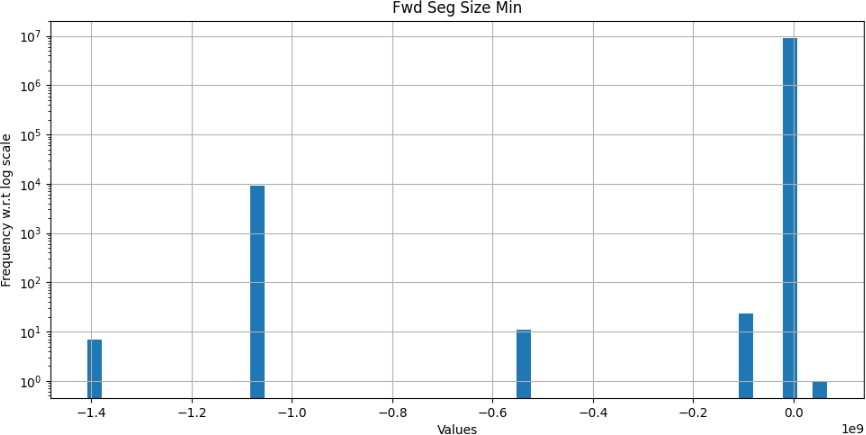
Fwd Seg Size Min: Minimum segment size in forward packets

Figure 4.4.49 Histogram of Fwd Seg Size Min plotted on log scale

* + - Peak was observed at Fwd Seg Size Min=0.0
    - On the X-axis value lie in the range -1.4 to 0.0. Thus, the values on X-axis are all negative, we need to check the actual values under the column to determine if data is accurate or invalid.
    - There are some values observed at Fwd Seg Size Min=-1.4, Fwd Seg Size Min>-1.2 and Fwd Seg Size Min<-1.0, Fwd Seg Size Min>-0.6 and Fwd Seg Size Min<-0.4

Active Mean: Mean active time

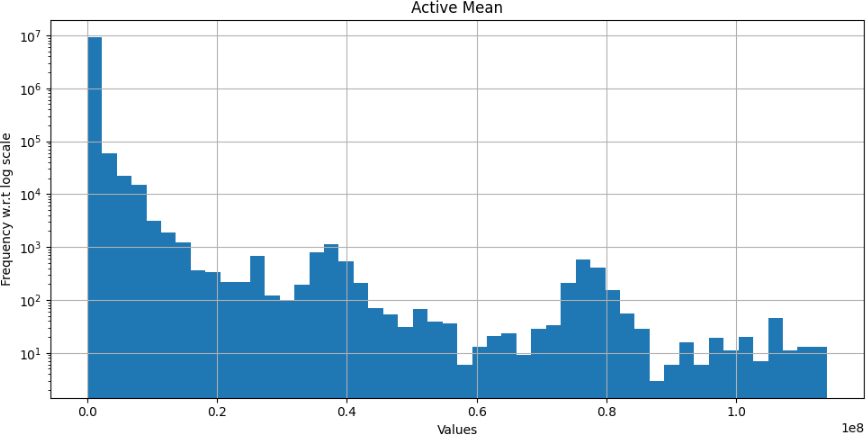


Figure 4.4.50 Histogram of Active Mean plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed at Active Mean=0.
    - After the peak, there is significant decline in results.
    - There are two plateau regions observed at Active Mean=0.4 and Active Mean=0.6
    - There are no gaps in the results observed on X-axis of the graph.

Active Std: Standard deviation of active time

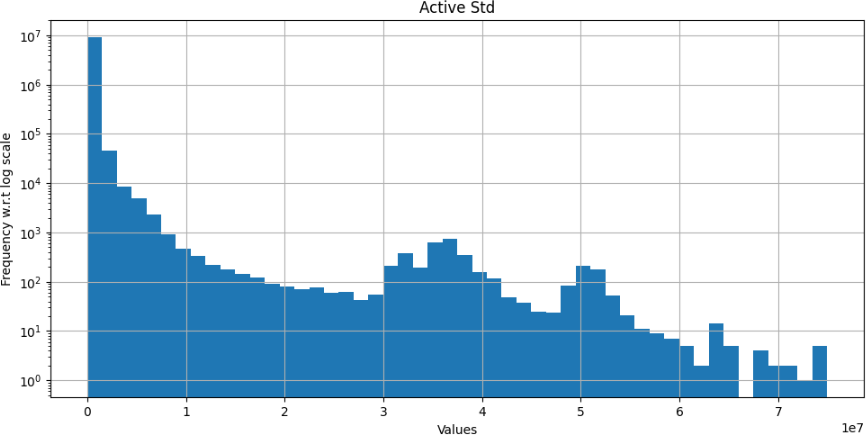


Figure 4.4.51 Histogram of Active Std plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed at Active Std=0.
    - After the peak, there is consistent decline in results up to Active Std=3.
    - There are two plateau regions observed between Active Std>=3 and Active Std<=4.
    - There is decline in the results between Active Std>=4 and Active Std<=5.
    - There is second plateau in the graph observed near Active Std=5.
    - There is decline in the results after Active Std>5.

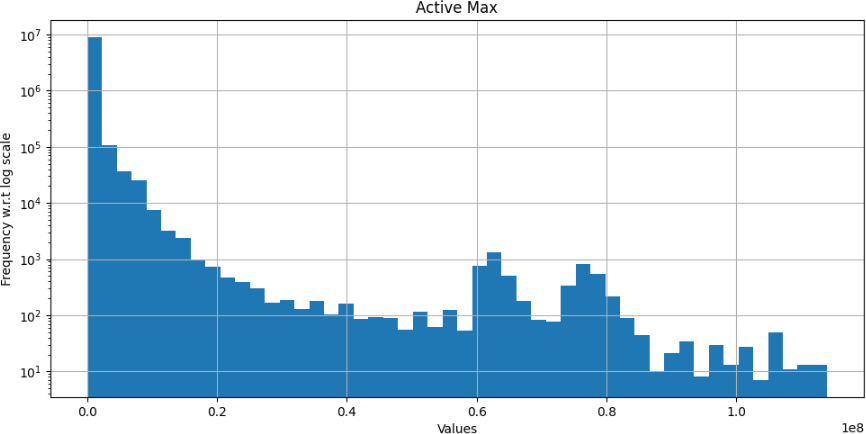
Active Max: Maximum active time

Figure 4.4.52 Histogram of Active Max plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed at Active Max=0.
    - After the peak, there is consistent decline in results up to Active Max=0.6
    - Around Active Max=0.6, there is a relatively smaller peak compared to main peak, and a plateau

region of 2 bins around it.

* + - Similarly, around Active Max=0.8, there is a relatively smaller peak compared to main peak, and a plateau region of 2 bins around it.
    - On X-axis values lie in the range 0 to 1.2
    - There are no gaps in the results observed on X-axis of the graph.

Active Min: Minimum active time

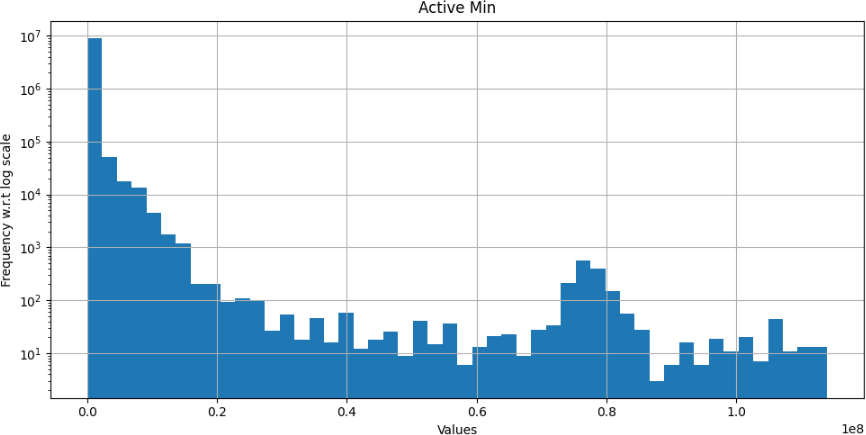


Figure 4.4.53 Histogram of Active Min plotted on log scale

* + - The distribution is skewed towards right: Positively skewed.
    - Peak was observed at Active Min=0.
    - There is relatively smaller peak at Active Min=0.8 and a plateau region around it.
    - On X-axis value lie in the range 0 to 1.2
    - There are no gaps in the results observed on X-axis of the graph.

Idle Mean: Mean idle time

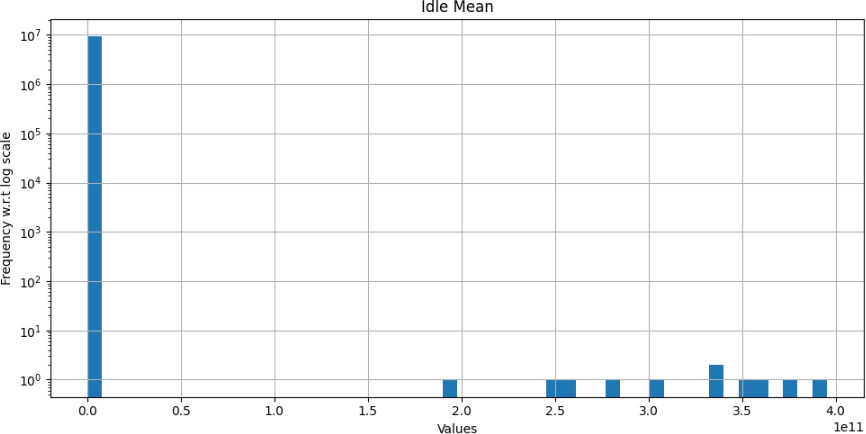


Figure 4.4.54 Histogram of Idle Mean plotted on log scale

* + - Peak was observed at Idle Mean=0.
    - Most values are concentrated in bin represented by the peak.
    - There are some values observed at Idle Mean=2.0, 3.0, 3.5, 4.0
    - There are large gaps observed on X-axis of the graph after the peak.

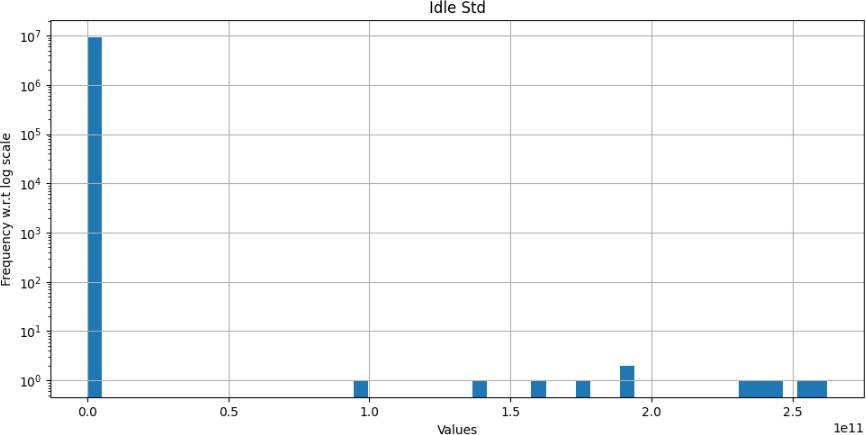
Idle Std: Standard deviation of idle time

Figure 4.4.55 Histogram of Idle Std plotted on log scale

* + - Peak was observed at Idle Std=0.0
    - Most values are concentrated in bin represented by the peak.
    - There are some values observed at Idle Std=1.0, 1.5, 2.0 and 2.5
    - There are large gaps observed on X-axis of the graph after the peak.

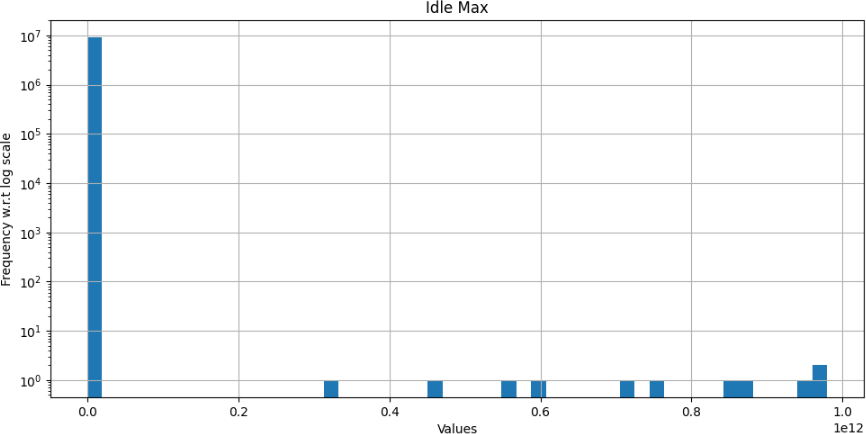
Idle Max: Maximum idle time

Figure 4.4.56 Histogram of Idle Max plotted on log scale

* + - Peak was observed at Idle Max=0.0
    - Most values are concentrated in bin represented by the peak.
    - There are some values observed at Idle Max=0.4, 0.6, 0.8, 1.0.

47

* + - There are large gaps observed on X-axis of the graph after the peak.

Idle Min: Minimum idle time

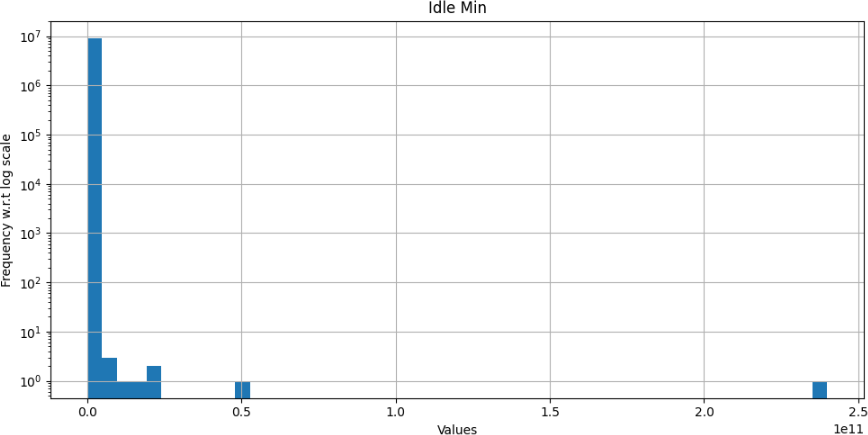


Figure 4.4.57 Histogram of Idle Min plotted on log scale

* + - Peak was observed at Idle Min=0.0
    - Most values are concentrated in bin represented by the peak.
    - There is a value observed after at Idle Min=2.5, which is after a large gap on X-axis. This may indicate outlier in the data.

# Distribution of target features using Bar chart: -

* + - Bar chart for all category of records under target feature: Label and ClassLabel were plotted.
    - The bar charts strongly indicated the imbalanced nature of the dataset in the direction of Benign records.
    - The dataset has 78% records classified as Benign and 22% records classified as Malicious.
    - Thus, it can be classified under Long-Tailed distribution, because lesser category of records (Benign events) has highest frequency, and more category of records (Malicious events) have lower frequency in the dataset.
    - As the result, we will need to address these issues while training the model to avoid bias for classifying an unknown event as Benign and also reasonably distinguish a Malicious event from Benign event.



Figure 4.5.1 Bar chart of ClassLabel plotted on log scale



Figure 4.5.2 Bar chart of Label plotted on log scale

# Analysing and handling negative values: -

Among the features where the negative values were observed, the proportion of negative values were computed.

List of features where negative values were observed are: -

1. Flow Duration = 0.001047%
2. Flow Bytes/s = 0.000578%
3. Flow Packets/s = 0.001047%
4. Flow IAT Mean = 0.001047%
5. Flow IAT Max = 0.000927%
6. Flow IAT Min = 0.030718%
7. Fwd IAT Total = 0.000153%
8. Fwd IAT Mean = 0.000153%
9. Fwd IAT Max = 0.000033%
10. Fwd IAT Min = 0.000349%
11. Fwd Header Length = 0.555312%
12. Bwd Header Length = 0.002803%
13. Init Fwd Win Bytes = 29.002950%
14. Init Bwd Win Bytes = 41.084015%
15. Fwd Seg Size Min = 0.808768%

Among the above 15 features, 2 features have relatively higher percentage of negative values: -

* + Init Fwd Win Bytes: 29%
  + Init Bwd Win Bytes: 41%

For remaining 13 features, the negative values were imputed with respective median value. This led to increase in concentration of values among those 13 features in their mid-range (at median), but percentage of negative values per feature is less than 1%, thus, the impact was very small.

If the rows having negative values for ‘Init Fwd Win Bytes’ and ‘Init Bwd Win Bytes’, we will lose massive volume of information for all features.

If the two columns were dropped, then the valid datapoints from those two columns will also be lost which may later play important role.

Init Fwd Win Bytes: - Among the negative values, 88% records are Benign and 12% records are Malicious.

Init Bwd Win Bytes: - Among the negative values, 78% records are Benign and 22% records are Malicious.

Thus, the negative values for the two features do not give any different characteristic of events when compared with characteristics of the complete dataset.

As the result, it indicates data quality issues.

We can perform prediction of data points by taking negative values as the unknown data and positive values as training and test data to build a regression model. But, due to time constraints, this approach was not adopted.

As the result, imputation with respective median values were performed against negative values. This led to large hump of values at median, thus the concentration of values around mid-range has increased.

# Defining a new feature in the dataset: -

A new feature was defined and added in the dataset: isMalicious, this will enable later to perform binary classification and differentiate between Malicious and Benign events.

Definition of isMalicious: -

* + - If ClassLabel=Benign -> isMalicious=0
    - If ClassLabel!=Benign -> isMalicious=1

Thus, number of records with isMalicious=0 : 7185881 and isMaliious=1 : 1981390

# Dropping an existing feature: -

‘Label’ was dropped from the dataset because it gives further sub-type of the attack, which will not be in scope of the project.

As the result, at this stage, the two target features in the dataset are: -

* + - isMalicious: For binary classification
    - ClassLabel: For multi-class classification

# Handling large size of the dataset: -

* + - The dataset was too large to carryout analysis and make updates as required, leading to

over utilization of system’s memory and notebook getting stalled.

* + - Thus, a sample of the dataset was taken by taking 20% of records as the sample size. Number of records with isMalicious=0 : 1437467 and isMalicious=1 : 395987
    - It was observed that the sampled dataset has similar imbalanced nature as the original dataset. And all the categories under ClassLabel are observed in the same as original dataset with similar proportions.

# Analyzing and handling outliers: -

The count and percentage of outliers in each independent feature of sampled dataset were computed.

Definition of outlier: - Given a datapoint x, if Equation (1) or Equation (2) is satisfied, then x is an outlier.

Table 4.10.1 Count and percentage of outliers for each feature

|  |  |  |
| --- | --- | --- |
| **Field name** | **Number of outliers** | **Percentage of outliers** |
| Flow Duration | 362139 | 19.75 |
| Total Fwd Packets | 167485 | 9.13 |
| Total Backward Packets | 176328 | 9.62 |
| Fwd Packets Length Total | 71763 | 3.91 |
| Bwd Packets Length Total | 265389 | 14.47 |
| Fwd Packet Length Max | 24476 | 1.33 |
| Fwd Packet Length Mean | 74245 | 4.05 |
| Fwd Packet Length Std | 21018 | 1.15 |
| Bwd Packet Length Max | 69888 | 3.81 |
| Bwd Packet Length Mean | 140674 | 7.67 |
| Bwd Packet Length Std | 56299 | 3.07 |
| Flow Bytes/s | 377550 | 20.59 |
| Flow Packets/s | 380170 | 20.74 |
| Flow IAT Mean | 346826 | 18.92 |
| Flow IAT Std | 284585 | 15.52 |
| Flow IAT Max | 255816 | 13.95 |

|  |  |  |
| --- | --- | --- |
| Flow IAT Min | 404158 | 22.04 |
| Fwd IAT Total | 352629 | 19.23 |
| Fwd IAT Mean | 355920 | 19.41 |
| Fwd IAT Std | 395445 | 21.57 |

Among the 57 independent features: -

1. 12 features have outliers whose percentage of difference between Malicious and Benign events is greater than or equal to 10%.
2. Remaining 45 features have nearly equal percentage of outliers labelled as Malicious and Benign.

Out of the 12 features, 4 features have relatively higher percentage of outliers.

1. Init Fwd Win Bytes: 39.24%
2. Init Bwd Win Bytes: 37.32%
3. Fwd Seg Size Min: 37.02%
4. Bwd IAT Mean: 14.04%

Among the above 4 features, there were a greater number of records classified as Benign than Malicious. Thus, the features with relatively higher number of outliers do not indicate any anomaly or provide differentiation to detect Malicious events.

Remaining 8 features have lesser than or equal to 6% of records as outliers: -

1. Fwd Packets Length Total
2. Bwd Packet Length Max
3. Bwd Packet Length Std
4. Fwd Header Length
5. Packet Length Max
6. Packet Length Std
7. Avg Fwd Segment Size
8. Subflow Fwd Bytes

All above 8 features have a greater number of outliers classified as Malicious than Benign. Thus, the features with relatively lesser number of outliers help to provide small differentiation of Malicious events over Benign events.

To handle outliers, two approaches were applied and tested: -

* 1. Winsorization
  2. Robust Scaling

The tests were performed on 4 features: -

1. Init Fwd Win Bytes
2. Init Bwd Win Bytes
3. Fwd Seg Size Min
4. Bwd IAT Mean

Winsorization: -

Winsorization was performed by replacing each feature’s lower range outliers with the 5th percentile value and higher range outliers with the 95th percentile value.

Histograms for each feature prior and post winsorization were plotted to observe the change in pattern of distribution.

Along with histogram, statistical computations were also done to compare the results for each feature and analyse the impact of the process.

Init Fwd Win Bytes: -

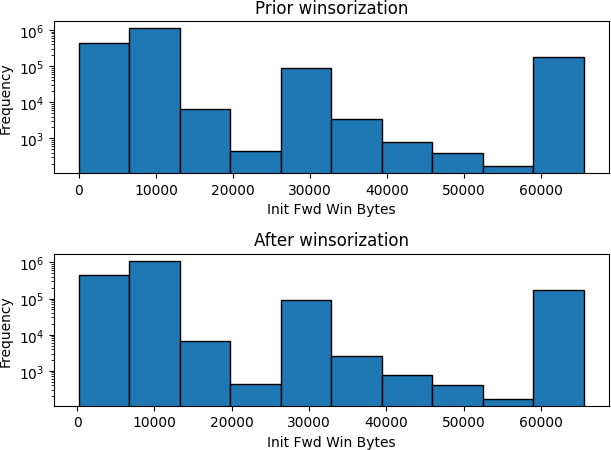


Figure 4.10.1 Histogram tocompareimpact of winsorizationon Init Fwd Win Bytes

* The distribution of the feature pre and post winsorization is similar.
* Median value has remained constant=8192.0
* Standard deviation value reduced from 17920 to 17917.

Init Bwd Win Bytes: -

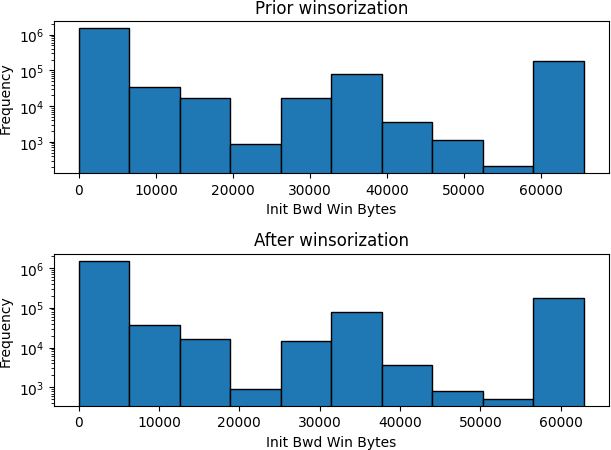


Figure 4.10.2 Histogram to compareimpactofwinsorizationon Init BwdWin Bytes

* The distribution of the feature pre and post winsorization is similar.
* Median value remained constant=235.0
* Standard deviation value reduced from 19414 to 19358.

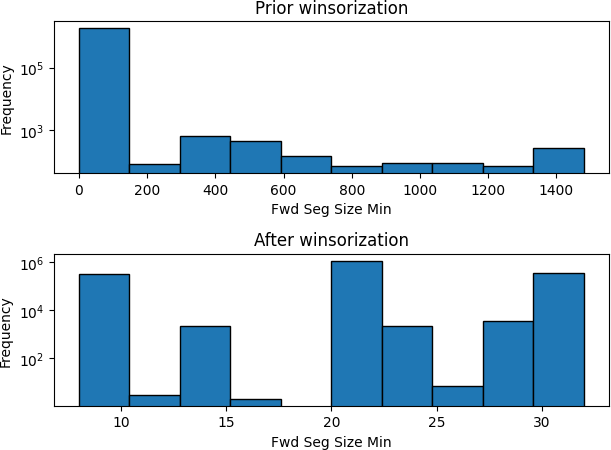
Fwd Seg Size Min: -

Figure 4.10.3Histogram to compareimpactofwinsorizationon FwdSeg Size Min

* Based on visual comparison of the two histograms, the distribution of data has changed drastically after winsorization.
* Median value remained constant=20.0
* Standard deviation value reduced from 25.97 to 7.26
* Maximum value prior winsorization was 1480 and after winsorization was 32. Number of values in the sampled dataset prior winsorization between 32 and 1480 = 17305. Thus,

many values were impacted due to the process.

Bwd IAT Mean: -

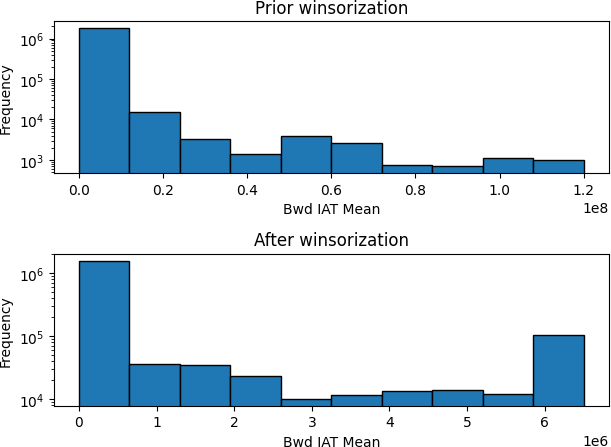


Figure 4.10.4 Histogram tocompareimpact of winsorization on BwdIAT Mean

* The distribution of data changed after performing winsorization on the feature.
* The main peak on the first bin (left hand side) has remained unchanged.
* There is a new peak observed towards the right hand side of the histogram plotted after winsorization. This may have occurred due to the outlier values that have got replaced by 95th percentile and thus, the frequency of last bin increased.
* Median value remained constant=647.
* Standard deviation value reduced from 6192044.5 to 1657361.2
* Maximum value in the sampled dataset prior winsorization was 120000000.0 and maximum value in the sampled dataset post winsorization was 6501929. Number of values in the sampled dataset prior winsorization between 6501929 and 120000000.0 = 91673. Thus, many values were impacted due to the process.

Robust Scaling: -

Robust Scaling was performed as the second option to handle outliers. If the given data point is x, then its value after Robust Scaling = x-Median/IQR.

Robust Scaling uses Median and IQR value to transform the data. Median and IQR value are mostly resistant to outliers. Thus, Robust Scaling is also resilient to outliers in data.

Init Fwd Win Bytes: -

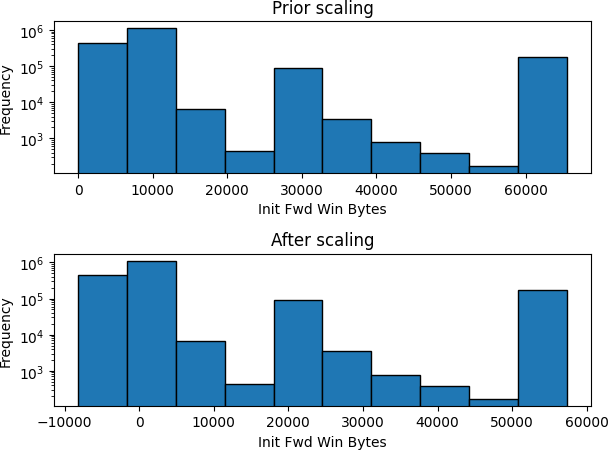


Figure4.10.5HistogramtocompareimpactofRobustScaling onInitFwdWinBytes

Init Bwd Win Bytes: -

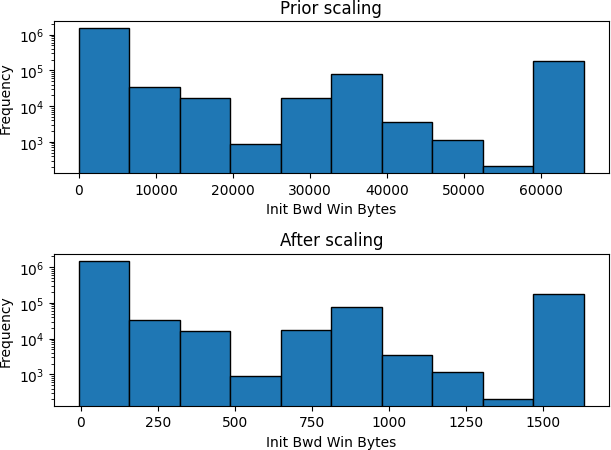


Figure4.10.6HistogramtocompareimpactofRobustScaling onInitBwdWinBytes

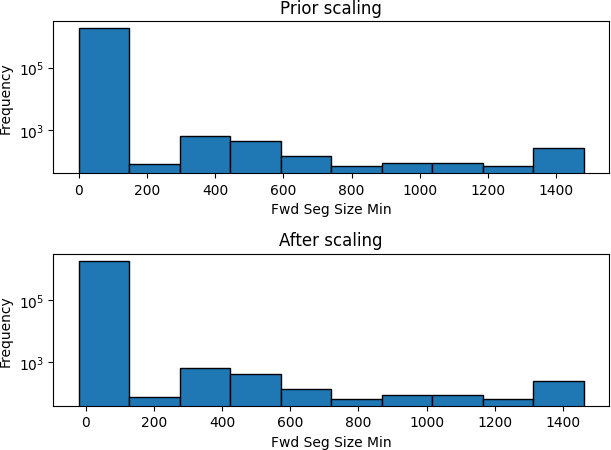
Fwd Seg Size Min: -

Figure 4.10.7 Histogram to compareimpactof Robust Scaling on Fwd Seg Size Min

Bwd IAT Mean: -

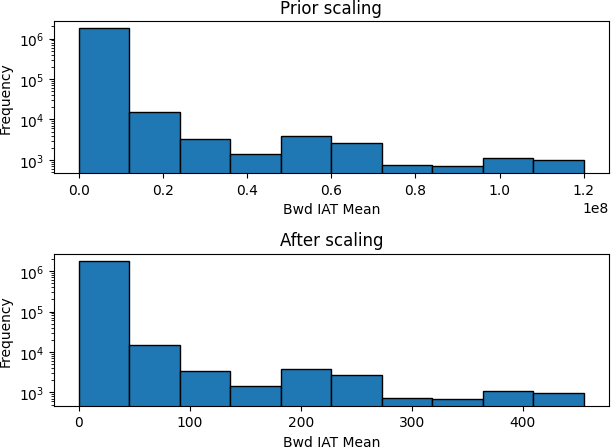


Figure 4.10.8 Histogram to compare impact of Robust Scaling on Bwd IAT Mean

After performing Robust Scaling on the four features, it was observed that although the distribution of data remained similar, the values on X-axis changed and it also led to transformation of existing data into negative values.

Summary of the two tests performed for handling outliers: -

* Winsorization impacts large number of values which are outliers and brings them closer to the normal range of data. As the result, the distribution pattern of the features changed by different magnitudes.
* Winsorization helped to reduce the influence of outliers by handling the extreme values in each of the four features.
* Robust Scaling kept the distribution pattern same pre and post operation. However, it led to negative values. This may be due to right skewed nature of the dataset. Since we subtract a

datapoint with median, in right skewed dataset, many data points are on the left hand side, that is closer to zero. Thus, subtraction of median from datapoints closer to zero led to generation of negative values.

Approach adopted for handling of outliers: -

* In order to prevent generation of negative values in the dataset, Winsorization was opted for handling outliers among the four features which have relatively large number of outliers.
* Since the four features have a greater number of outliers, they also have higher likelihood of having noisy data. As the result, Winsorization will help to reduce the impact of noise among the four features.
* For the remaining features, outliers were handled by performing imputation with median values. Reason for this approach: -
  + Since the number of outliers among these features were very less, imputing them with respective median value will help to approximate the entries having outliers.
  + Most of the features are skewed, thus, the imputation of outliers was performed with respective median values.

Thus, all outliers among the independent features were handled by combining the approach of Winsorization and Imputation with Median, and mitigated loss of data by preventing deletion of records having outliers.

After handling outliers on the original dataset, again the sample size of 20% was selected as sampled dataset.

# Distribution of data plotted on log scale for each independent feature after handling negative values and outliers: -

Based on the sampled dataset, histograms on log scale were plotted for all independent features, to compare how the distribution of each feature changed prior and after handling of negative values and outliers in the dataset.

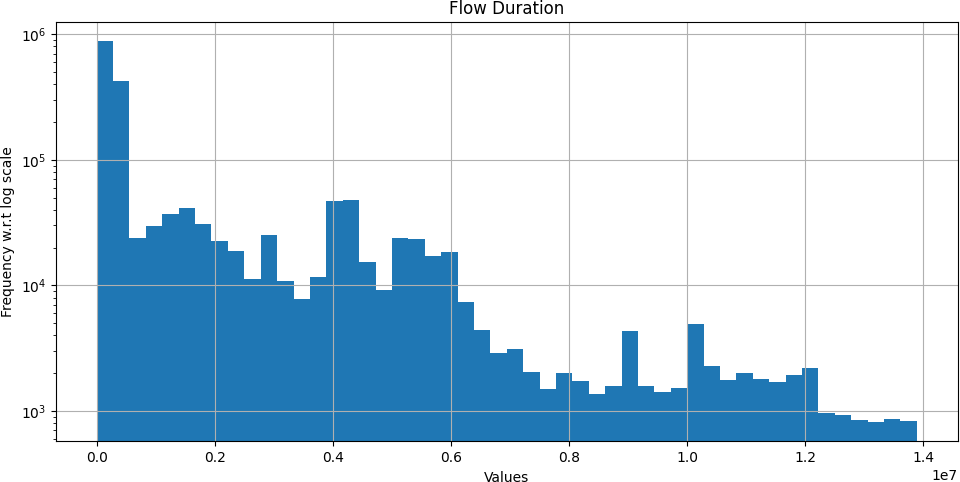


Figure 4.11.1 Histogram of Flow Duration plotted on log scale after handling negative values and outliers

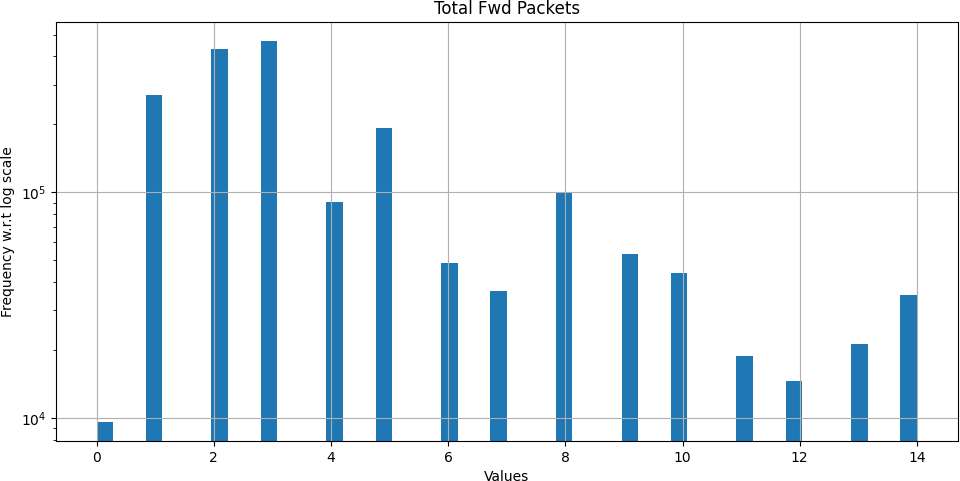


Figure 4.11.2 Histogram of Total Fwd Packets plotted on log scale after handling negative values and outliers

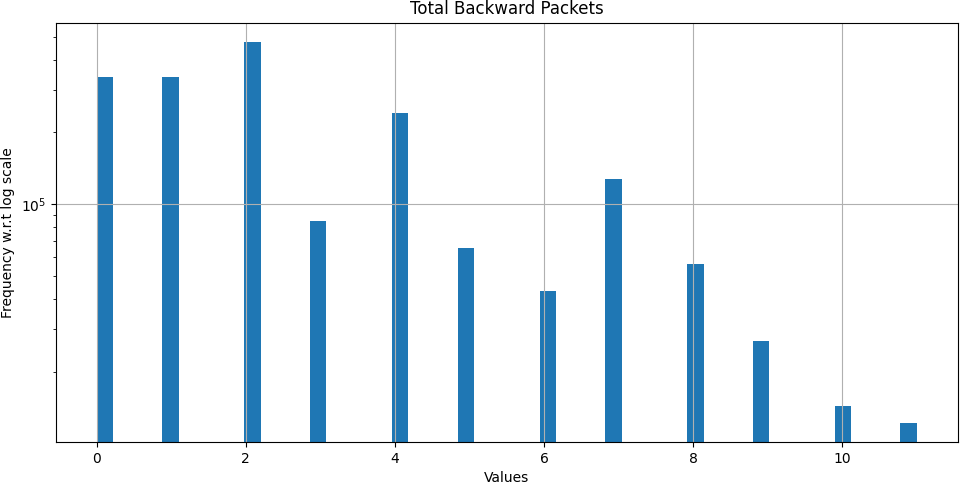


Figure 4.11.3 Histogram of Total Backward Packets plotted on log scale after handling negative values and outliers

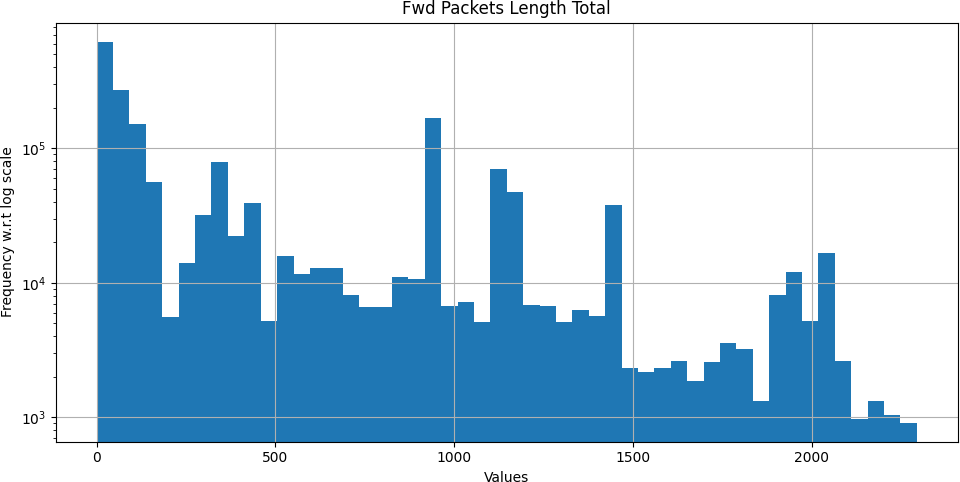


Figure 4.11.4 Histogram of Fwd Packets Length Total plotted on log scale after handling negative values and outliers

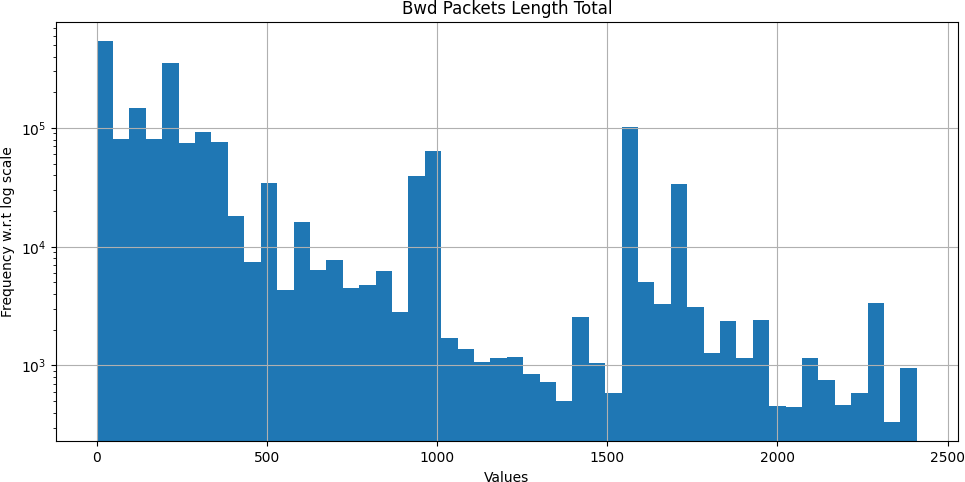


Figure 4.11.5 Histogram of Bwd Packets Length Total plotted on log scale after handling negative values and outliers

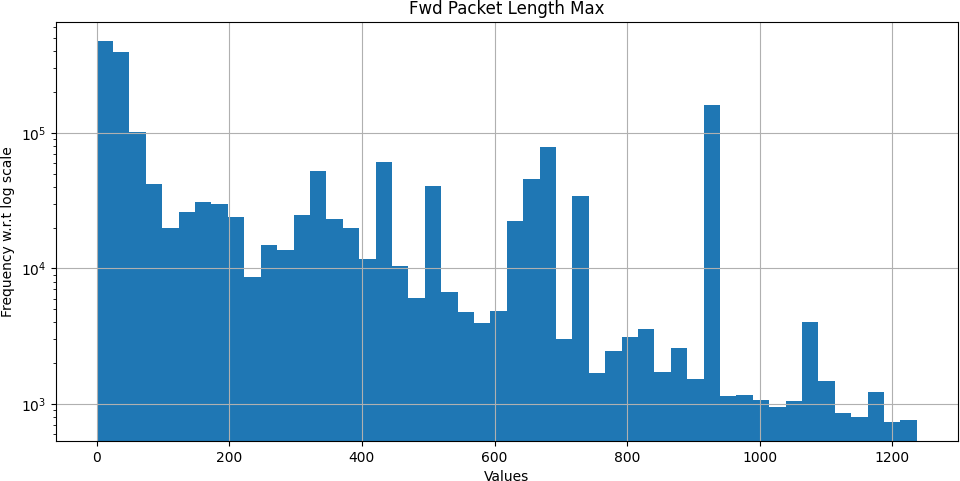


Figure 4.11.6 Histogram of Fwd Packet Length Max plotted on log scale after handling negative values and outliers

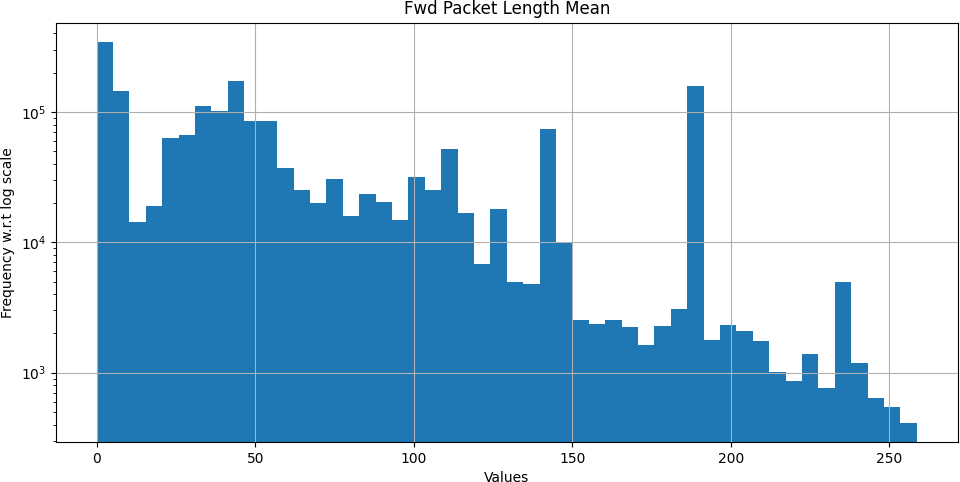


Figure 4.11.7 Histogram of Fwd Packet Length Mean plotted on log scale after handling negative values and outliers

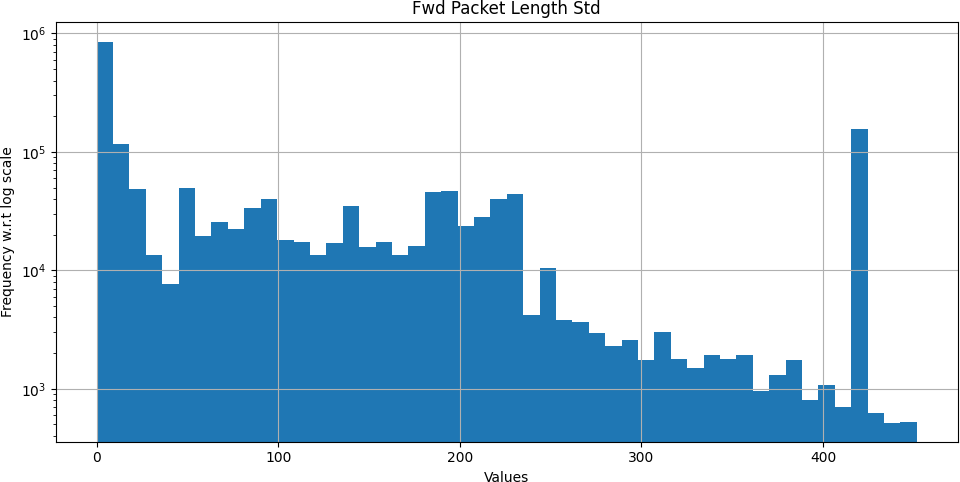


Figure 4.11.8 Histogram of Fwd Packet Length Std plotted on log scale after handling negative values and outliers

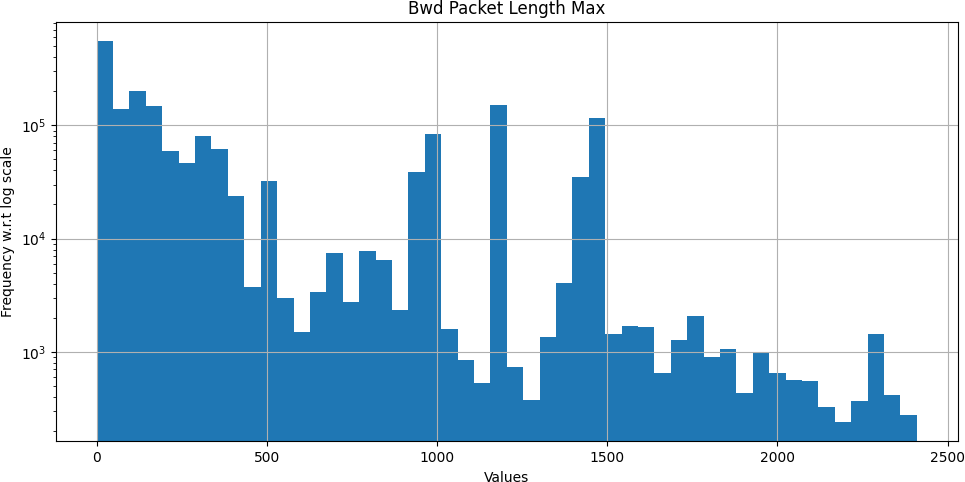


Figure 4.11.9 Histogram of Bwd Packet Length Max plotted on log scale after handling negative values and outliers

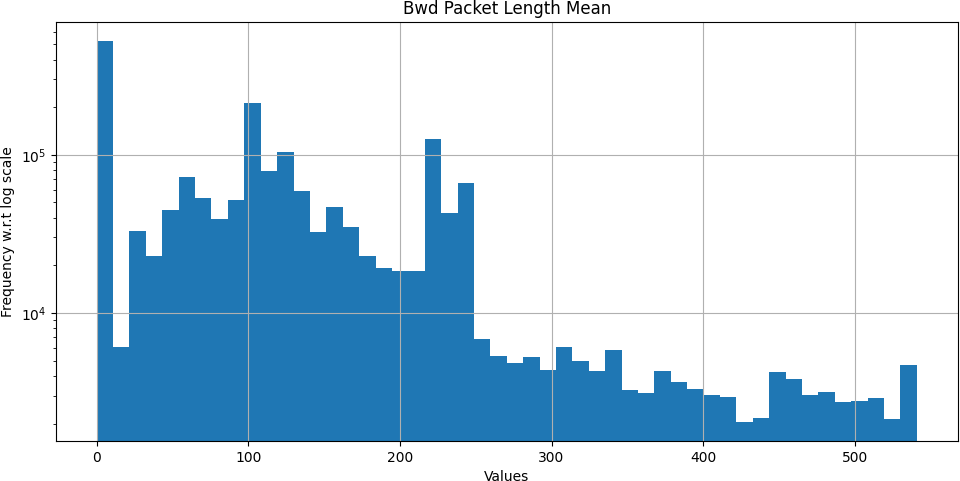


Figure 4.11.10 Histogram of Bwd Packet Length Mean plotted on log scale after handling negative values and outliers

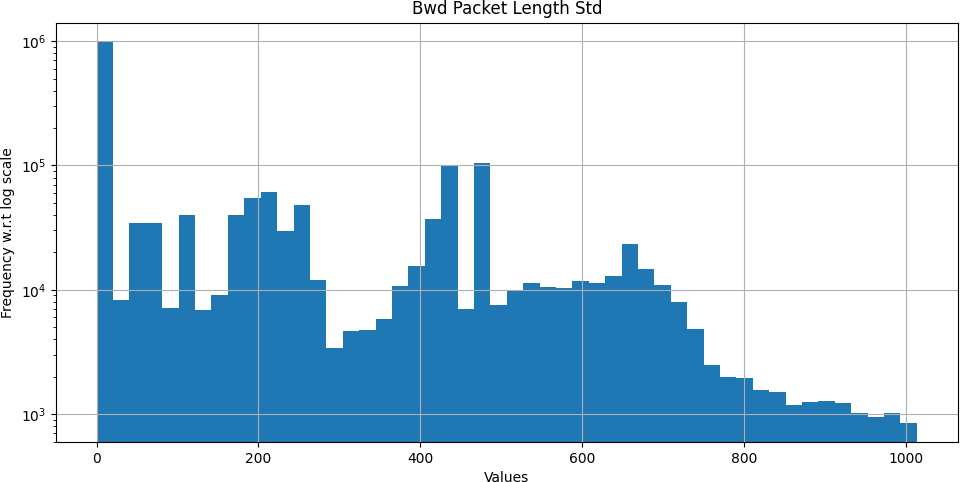


Figure 4.11.11 Histogram of Bwd Packet Length Std plotted on log scale after handling negative values and outliers

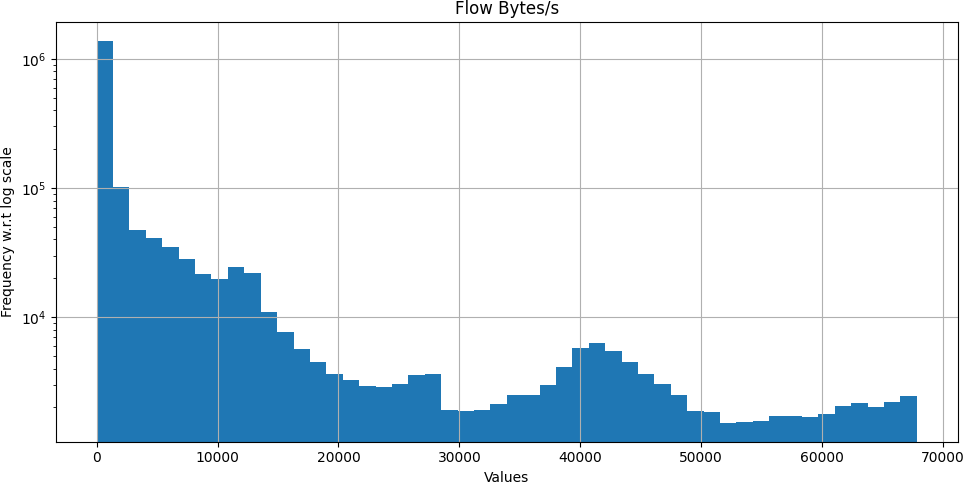


Figure 4.11.12 Histogram of Flow Bytes/s plotted on log scale after handling negative values and outliers

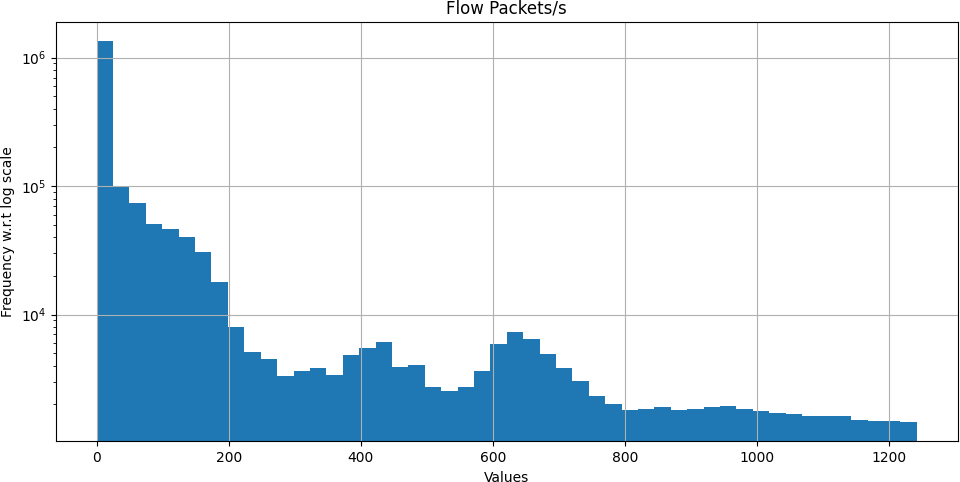


Figure 4.11.13 Histogram of Flow Packets/s plotted on log scale after handling negative values and outliers

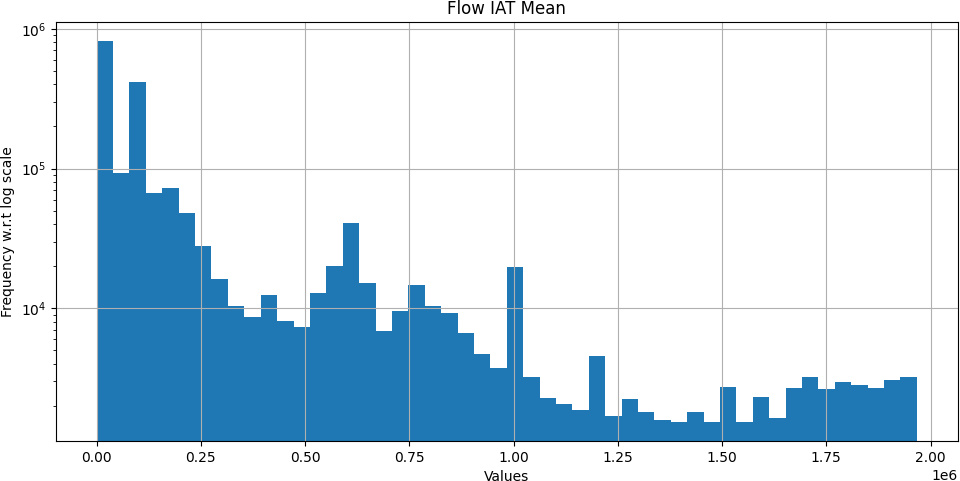


Figure 4.11.14 Histogram of Flow IAT Mean plotted on log scale after handling negative values and outliers

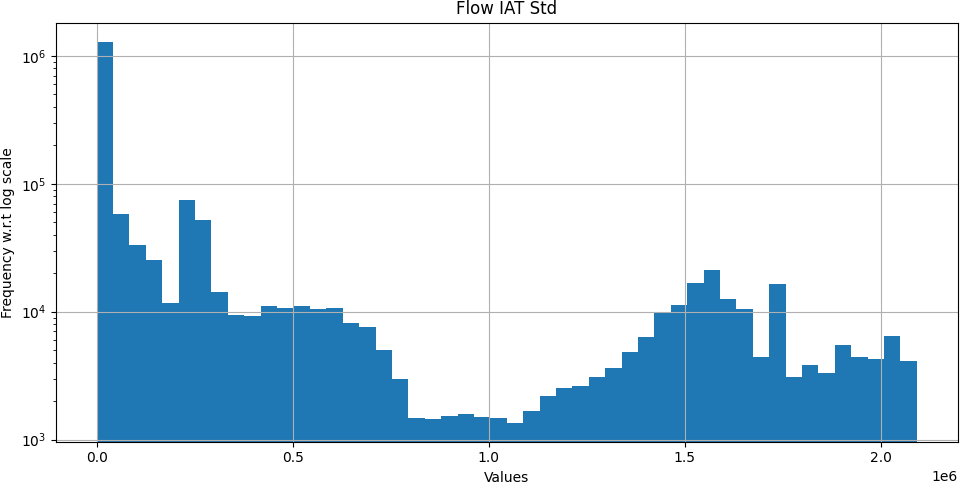


Figure 4.11.15 Histogram of Flow IAT Std plotted on log scale after handling negative values and outliers

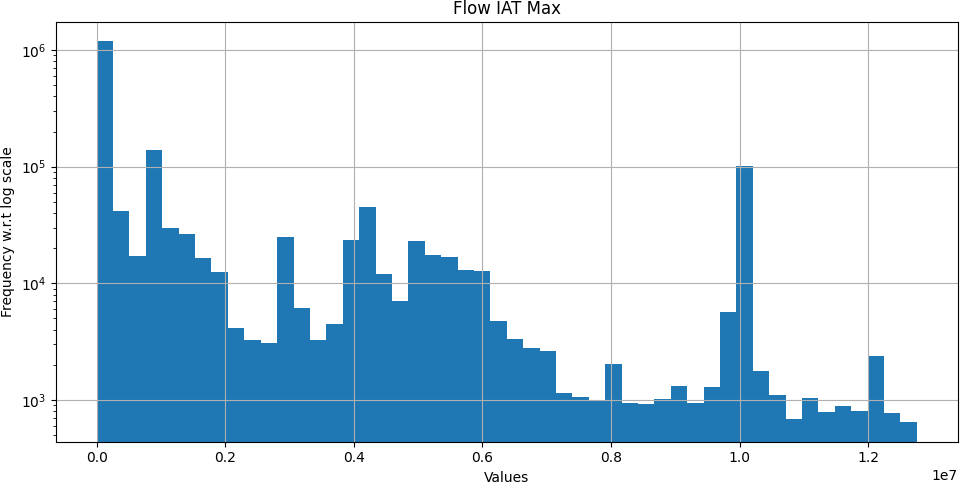


Figure 4.11.16 Histogram of Flow IAT Max plotted on log scale after handling negative values and outliers

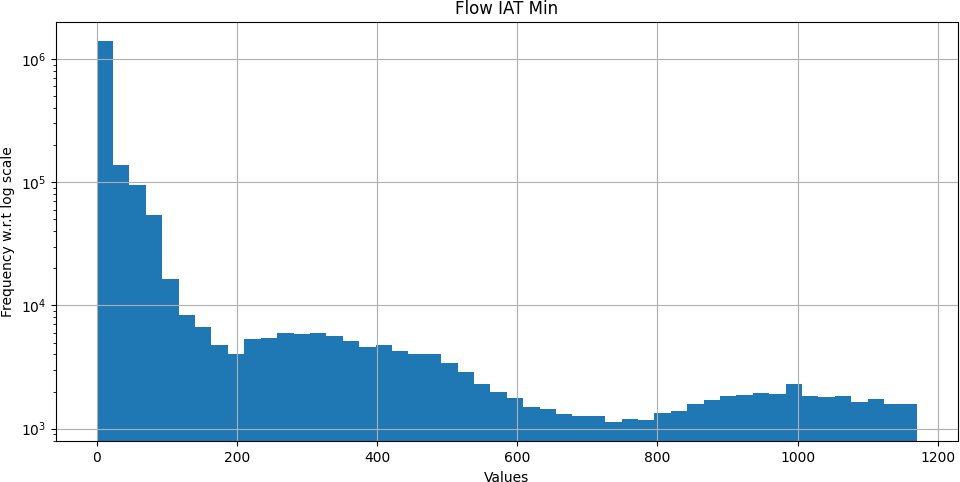


Figure 4.11.17 Histogram of Flow IAT Min plotted on log scale after handling negative values and outliers

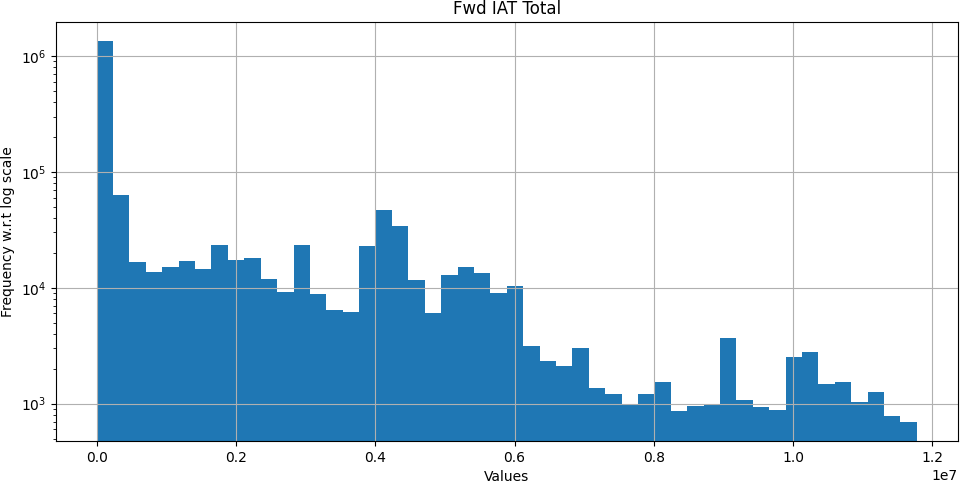


Figure 4.11.18 Histogram of Fwd IAT Total plotted on log scale after handling negative values and outliers

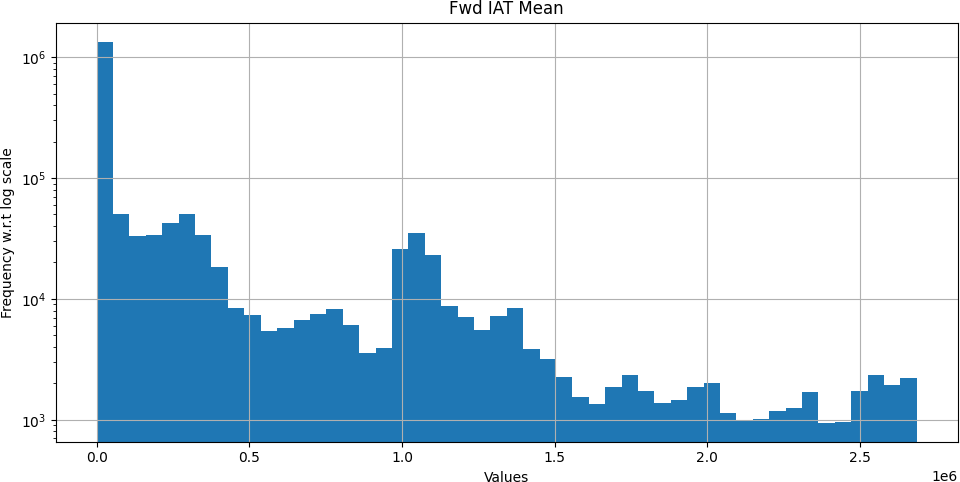


Figure 4.11.19 Histogram of Fwd IAT Mean plotted on log scale after handling negative values and outliers

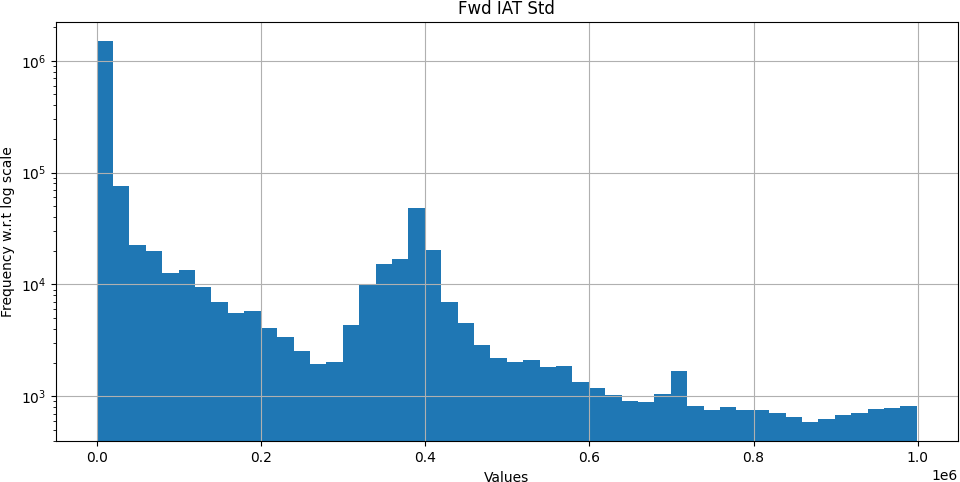


Figure 4.11.20 Histogram of Fwd IAT Std plotted on log scale after handling negative values and outliers

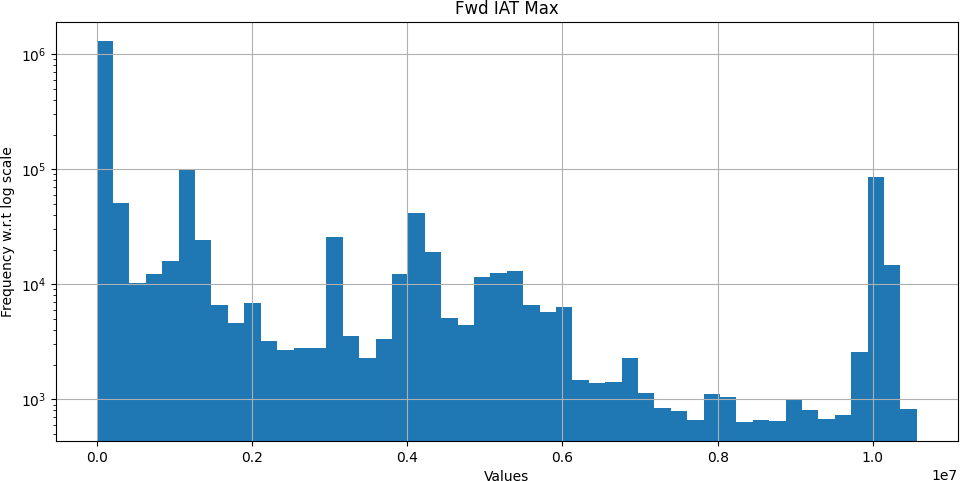


Figure 4.11.21 Histogram of Fwd IAT Max plotted on log scale after handling negative values and outliers

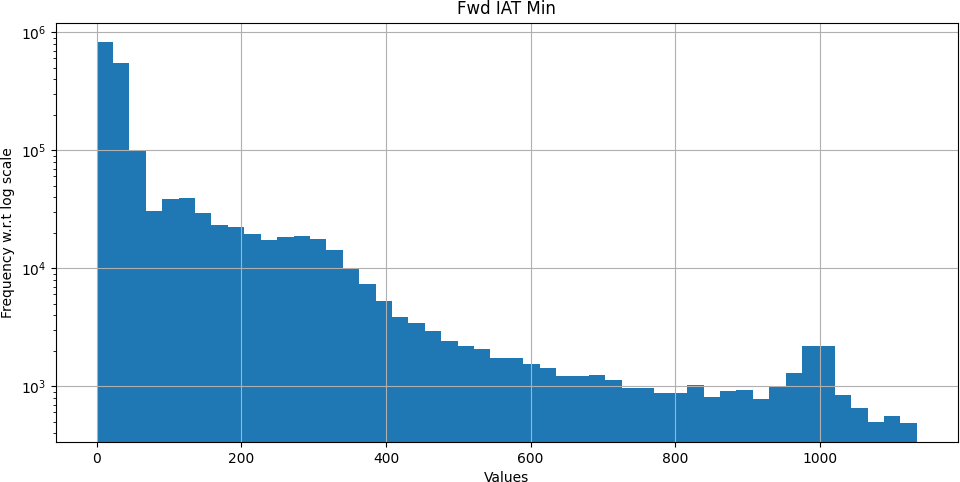


Figure 4.11.22 Histogram of Fwd IAT Min plotted on log scale after handling negative values and outliers

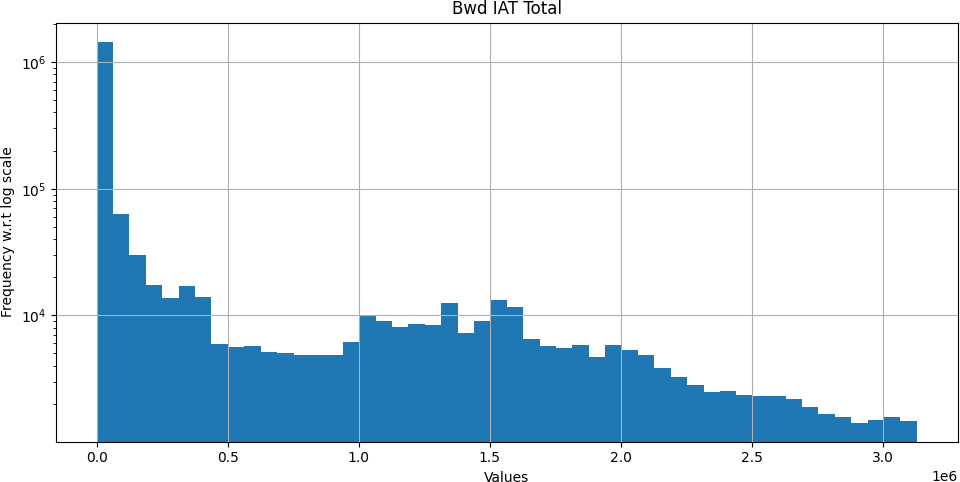


Figure 4.11.23 Histogram of Bwd IAT Total plotted on log scale after handling negative values and outliers

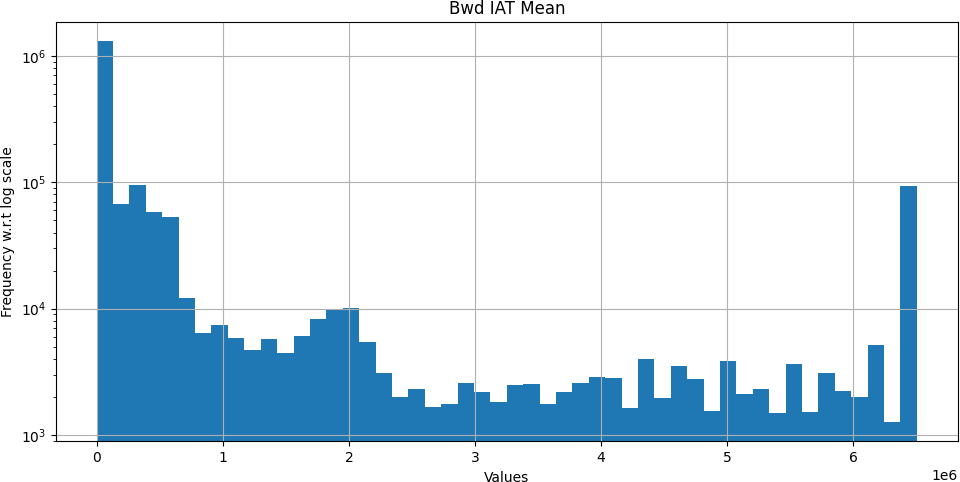


Figure 4.11.24 Histogram of Bwd IAT Mean plotted on log scale after handling negative values and outliers

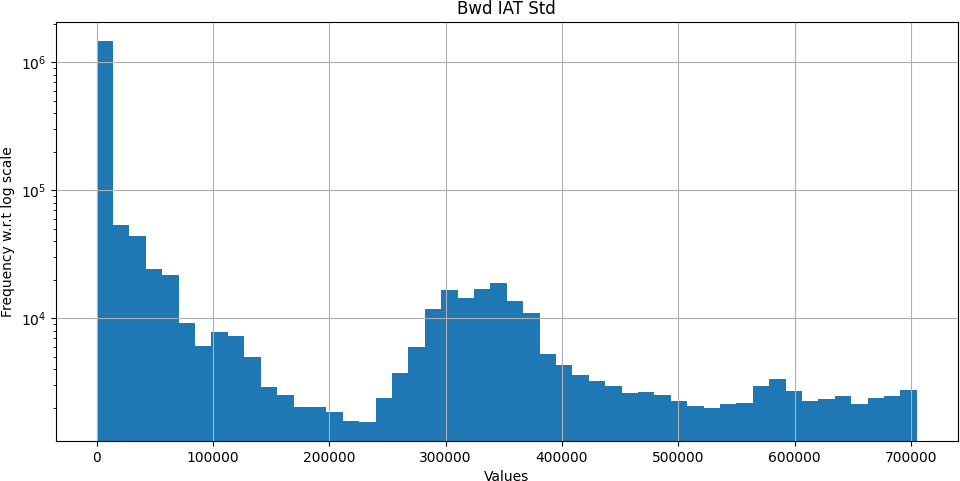


Figure 4.11.25 Histogram of Bwd IAT Std plotted on log scale after handling negative values and outliers

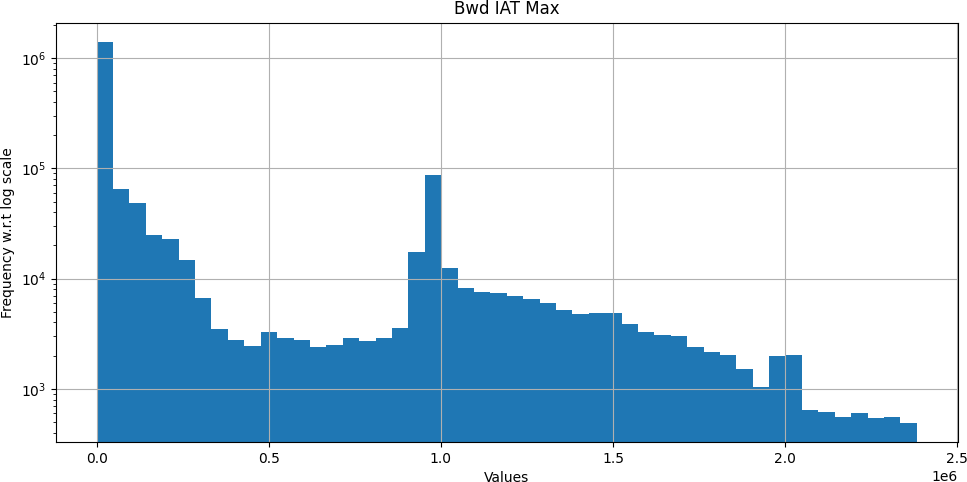


Figure 4.11.26 Histogram of Bwd IAT Max plotted on log scale after handling negative values and outliers

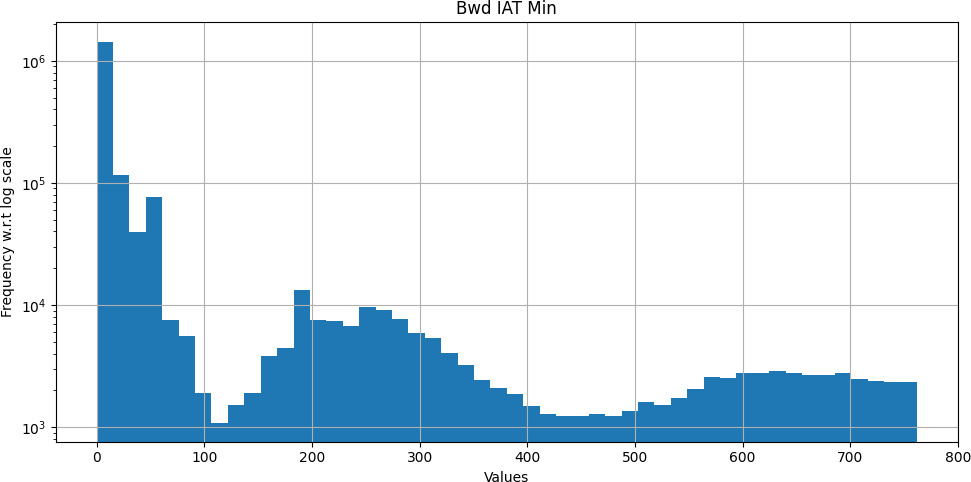


Figure 4.11.27 Histogram of Bwd IAT Min plotted on log scale after handling negative values and outliers

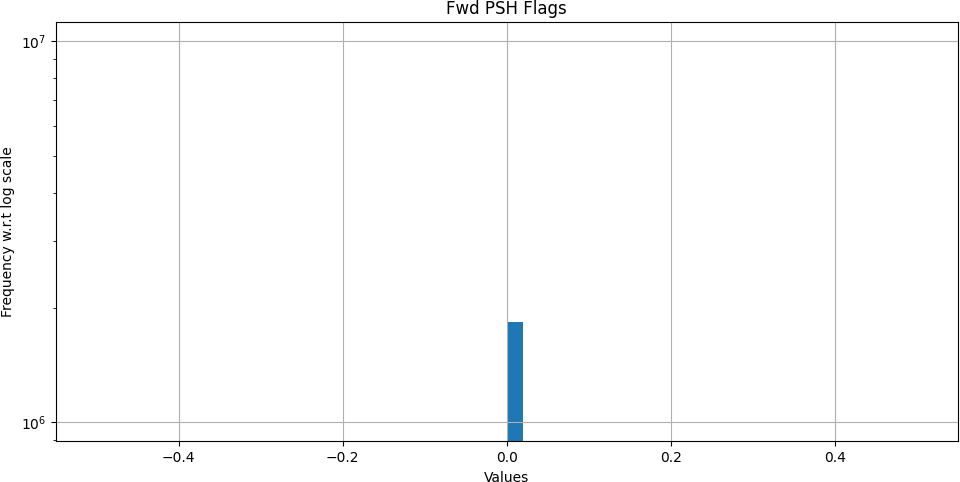


Figure 4.11.28 Histogram of Fwd PSH Flags plotted on log scale after handling negative values and outliers

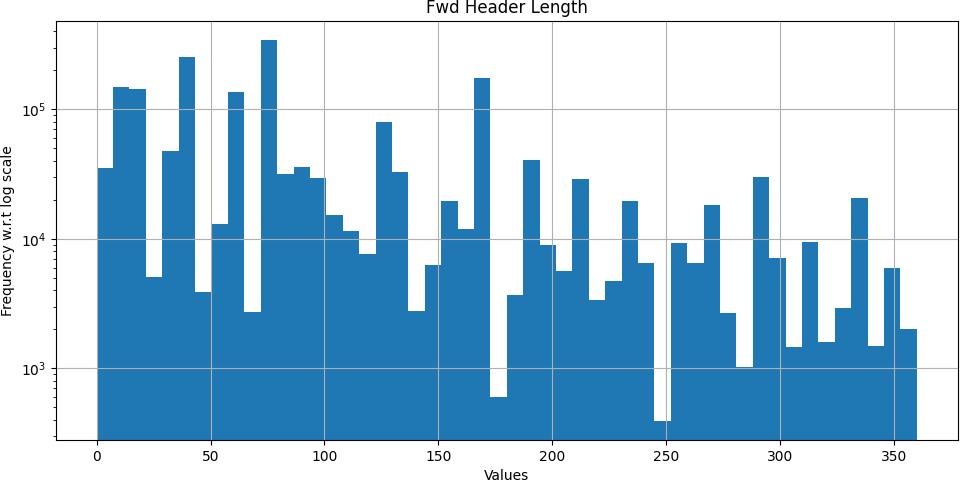


Figure 4.11.29 Histogram of Fwd Header Length plotted on log scale after handling negative values and outliers

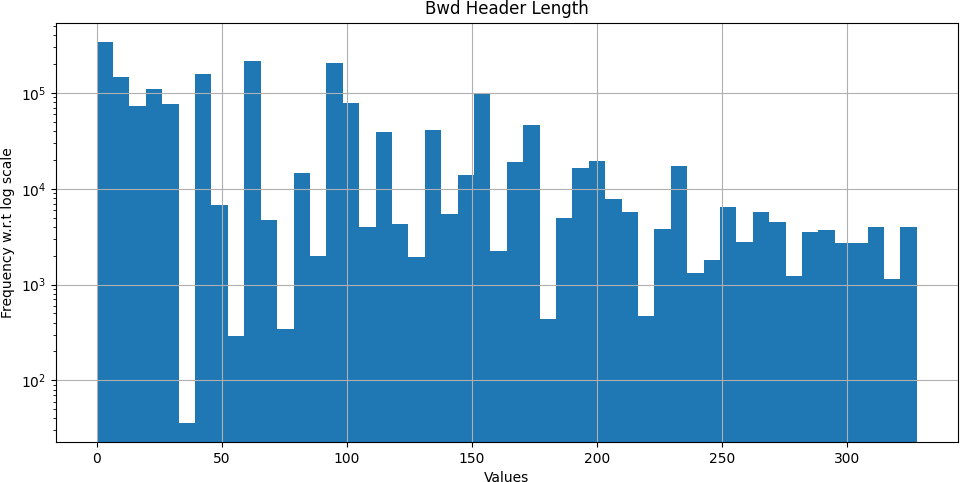


Figure 4.11.30 Histogram of Bwd Header Length plotted on log scale after handling negative values and outliers

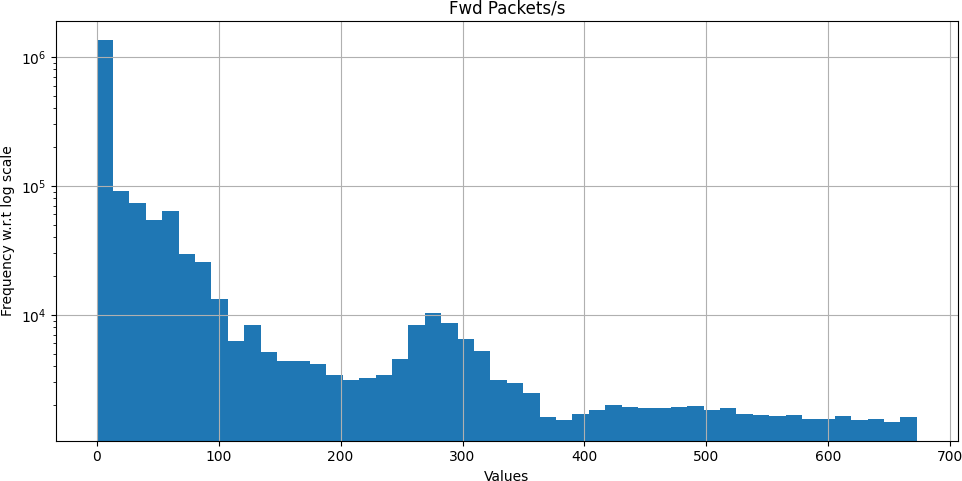


Figure 4.11.31 Histogram of Fwd Packets/s plotted on log scale after handling negative values and outliers

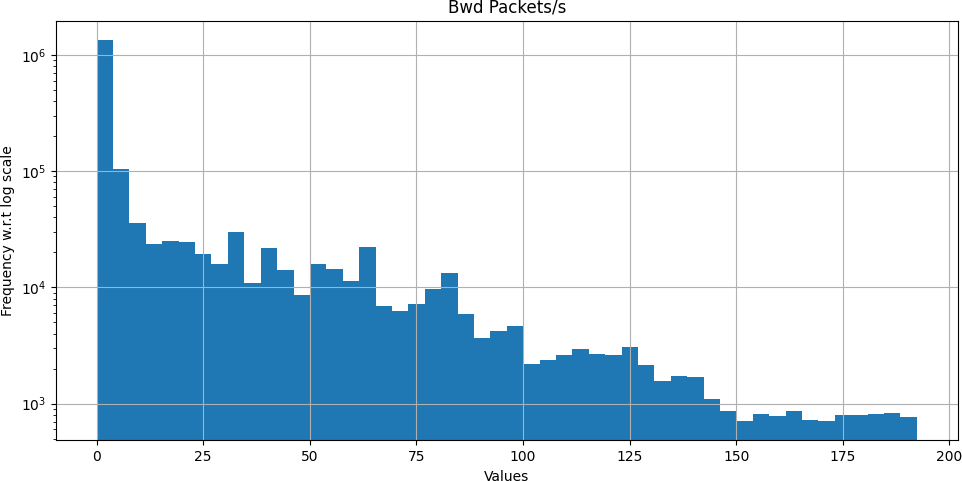


Figure 4.11.32 Histogram of Bwd Packets/s plotted on log scale after handling negative values and outliers

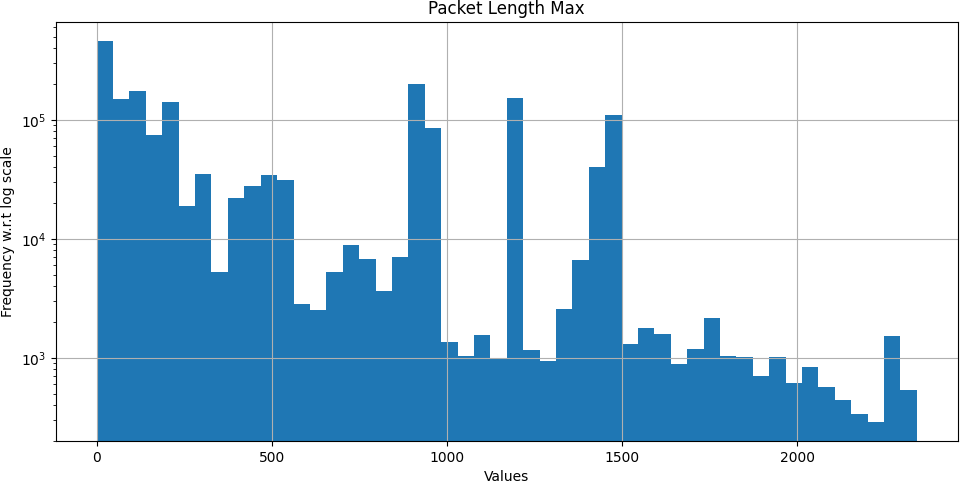


Figure 4.11.33 Histogram of Packet Length Max plotted on log scale after handling negative values and outliers

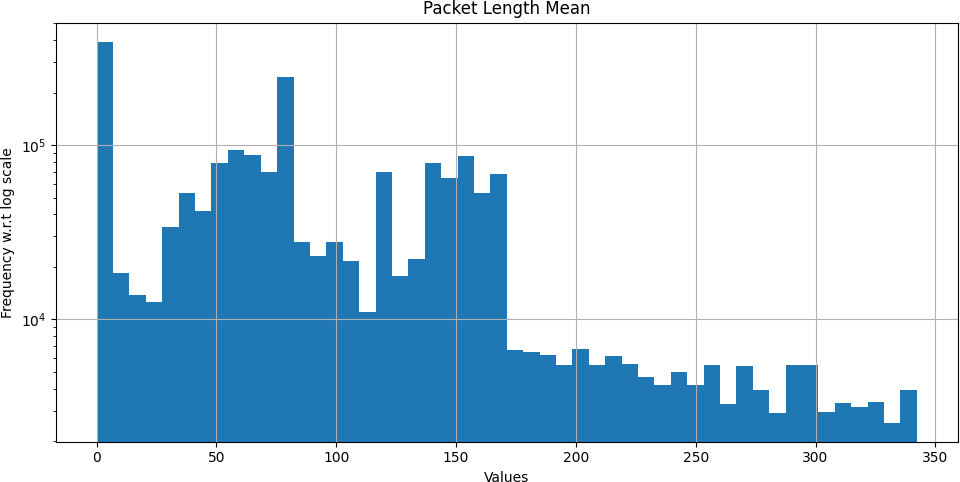


Figure 4.11.34 Histogram of Packet Length Mean plotted on log scale after handling negative values and outliers

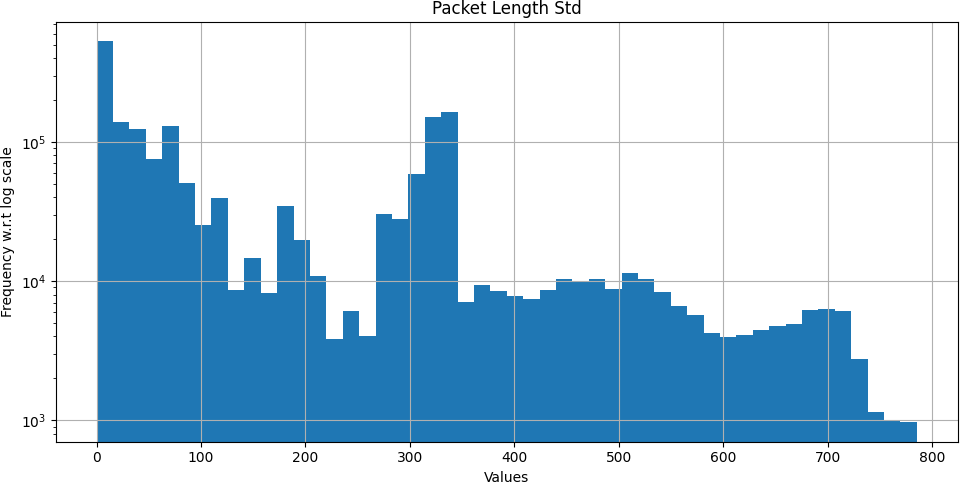


Figure 4.11.35 Histogram of Packet Length Std plotted on log scale after handling negative values and outliers

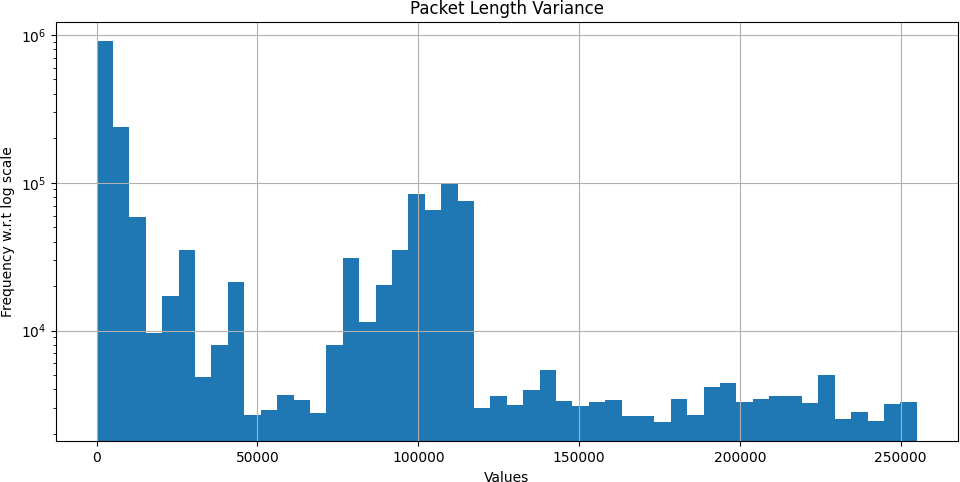


Figure 4.11.36 Histogram of Packet Length Variance plotted on log scale after handling negative values and outliers

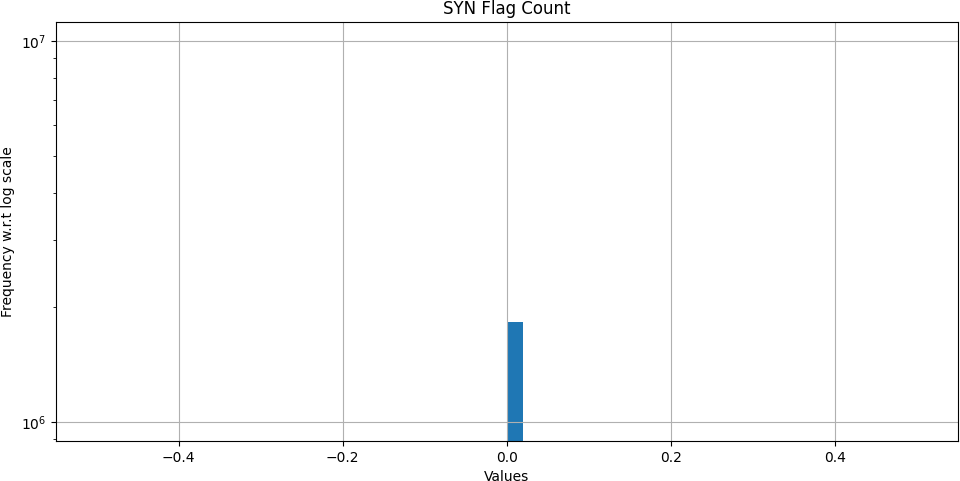


Figure 4.11.37 Histogram of SYN Flag Count plotted on log scale after handling negative values and outliers

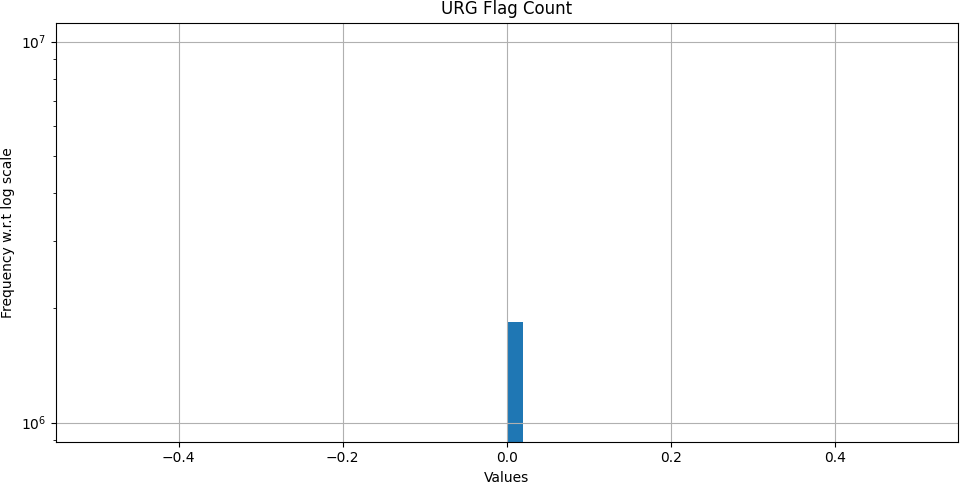


Figure 4.11.38 Histogram of URG Flag Count plotted on log scale after handling negative values and outliers

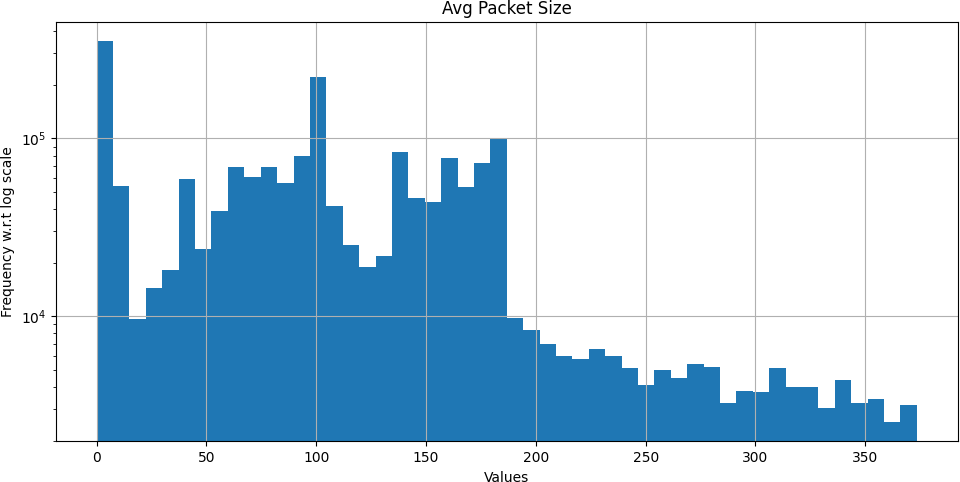


Figure 4.11.39 Histogram of Avg Packet Size plotted on log scale after handling negative values and outliers

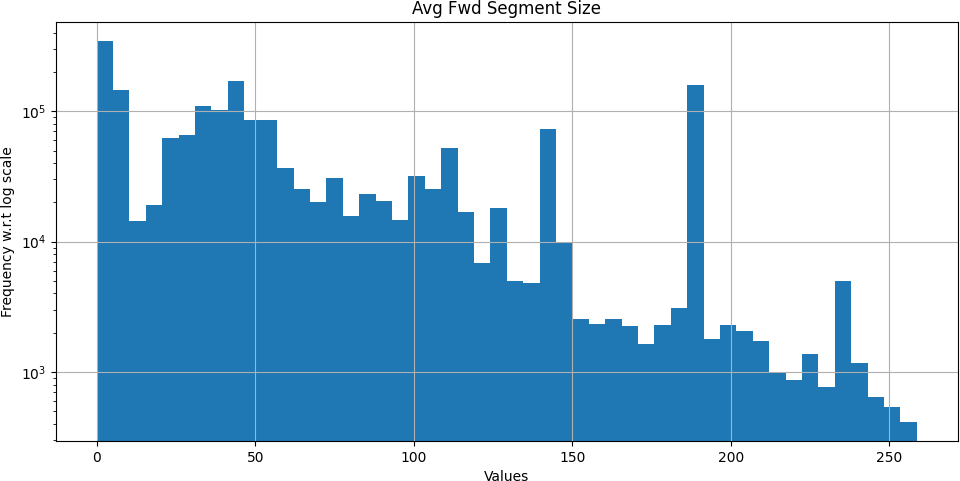


Figure 4.11.40 Histogram of Avg Fwd Segment Size plotted on log scale after handling negative values and outliers

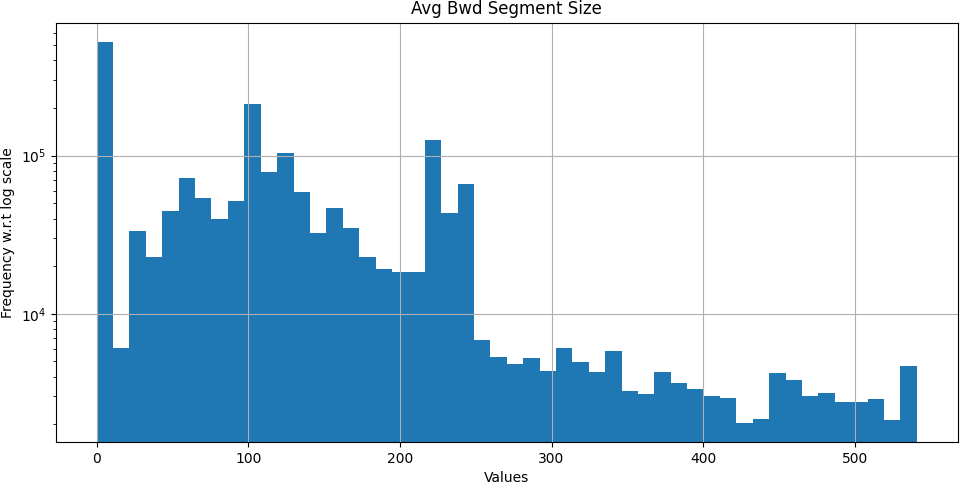


Figure 4.11.41 Histogram of Avg Bwd Segment Size plotted on log scale after handling negative values and outliers

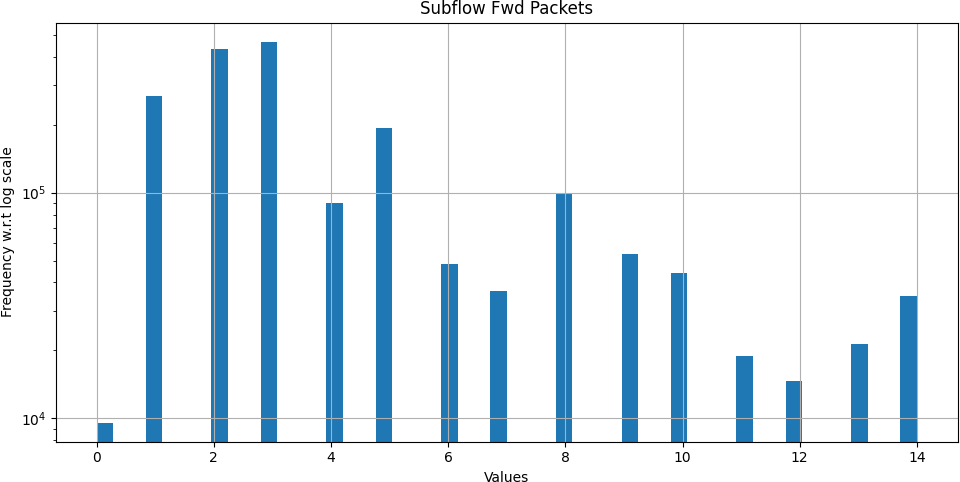


Figure 4.11.42 Histogram of Subflow Fwd Packets plotted on log scale after handling negative values and outliers

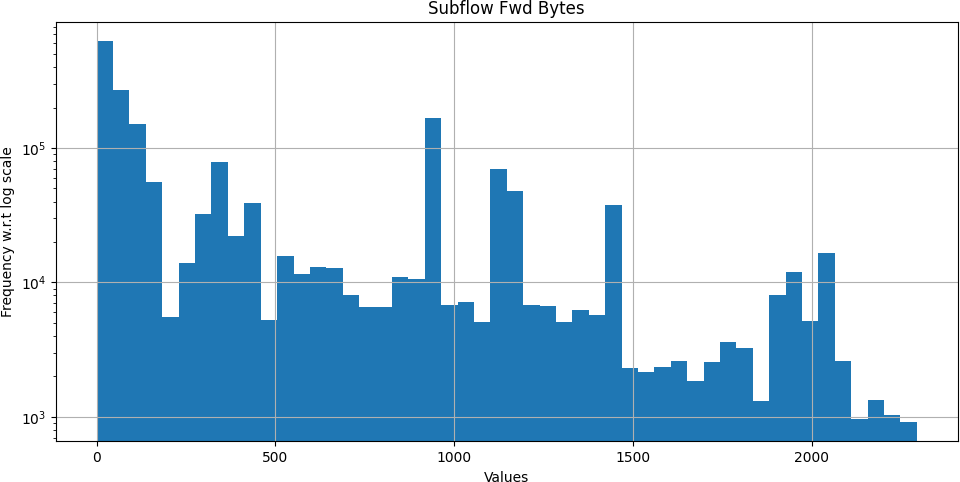


Figure 4.11.43 Histogram of Subflow Fwd Bytes plotted on log scale after handling negative values and outliers

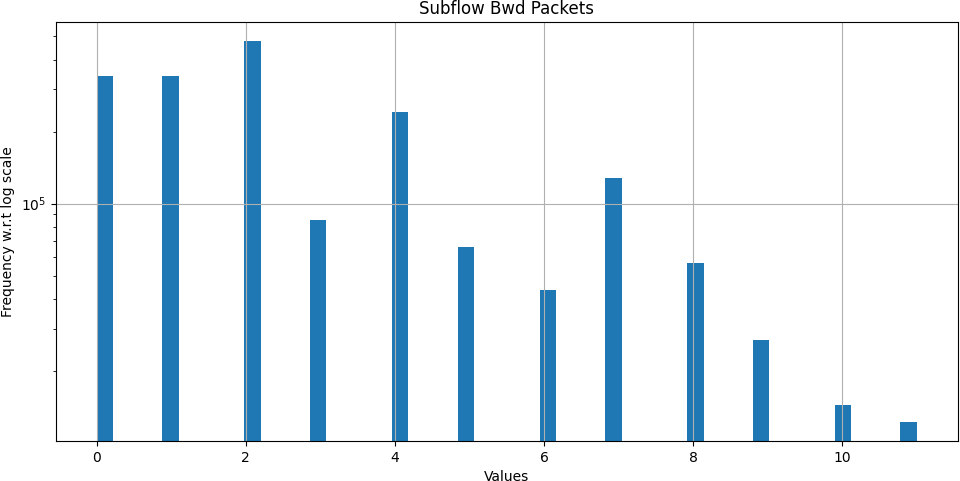


Figure 4.11.44 Histogram of Subflow Bwd Packets plotted on log scale after handling negative values and outliers

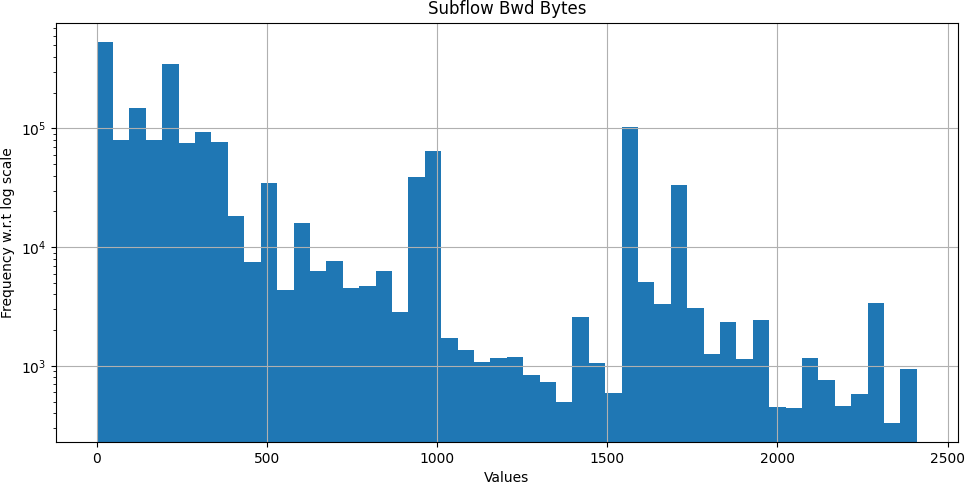


Figure 4.11.45 Histogram of Subflow Bwd Bytes plotted on log scale after handling negative values and outliers

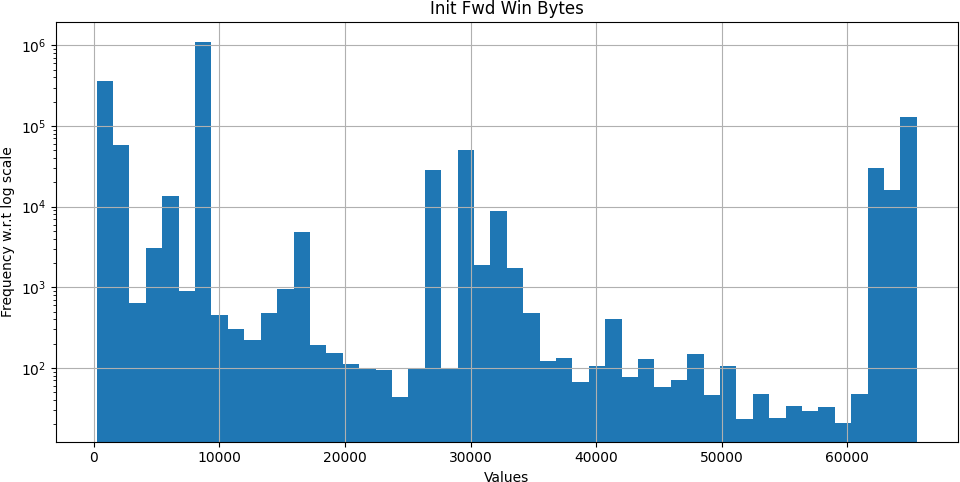


Figure 4.11.46 Histogram of Init Fwd Win Bytes plotted on log scale after handling negative values and outliers

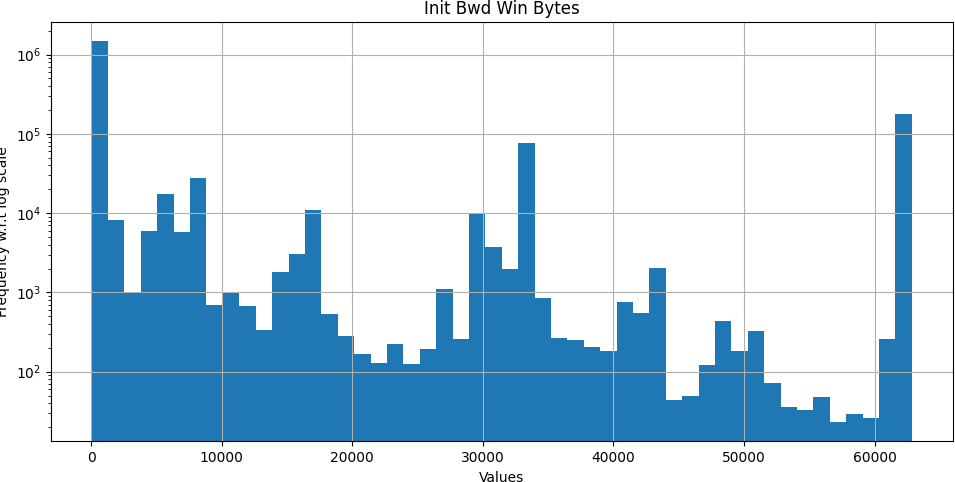


Figure 4.11.47 Histogram of Init Bwd Win Bytes plotted on log scale after handling negative values and outliers

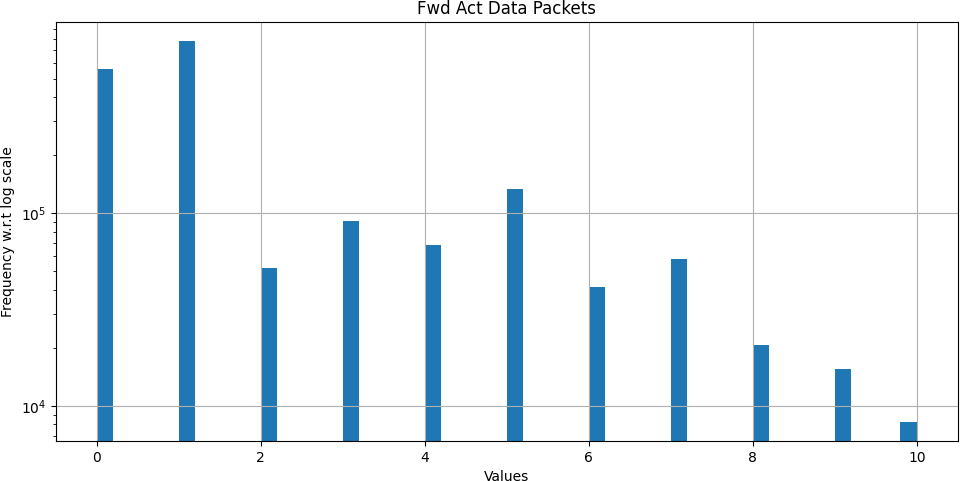


Figure 4.11.48 Histogram of Fwd Act Data Packets plotted on log scale after handling negative values and outliers

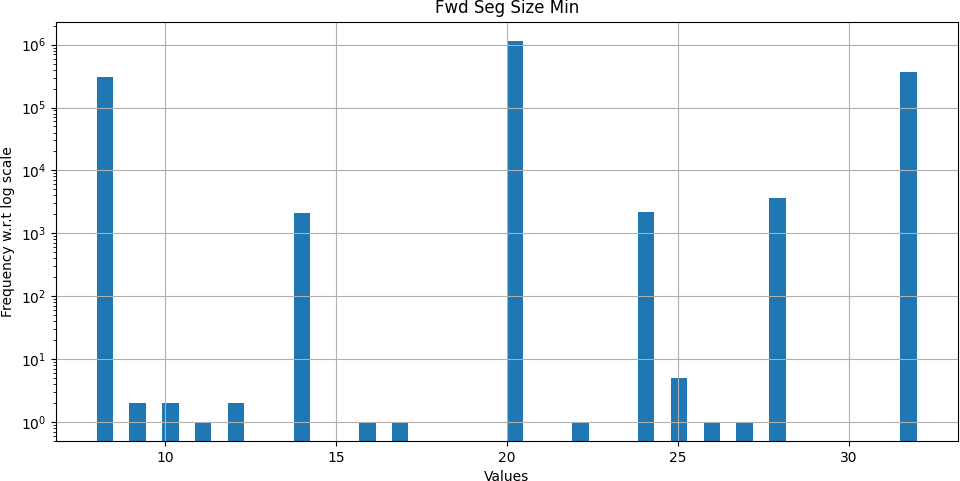


Figure 4.11.49 Histogram of Fwd Seg Size Min plotted on log scale after handling negative values and outliers



Figure 4.11.50 Histogram of Active Mean plotted on log scale after handling negative values and outliers

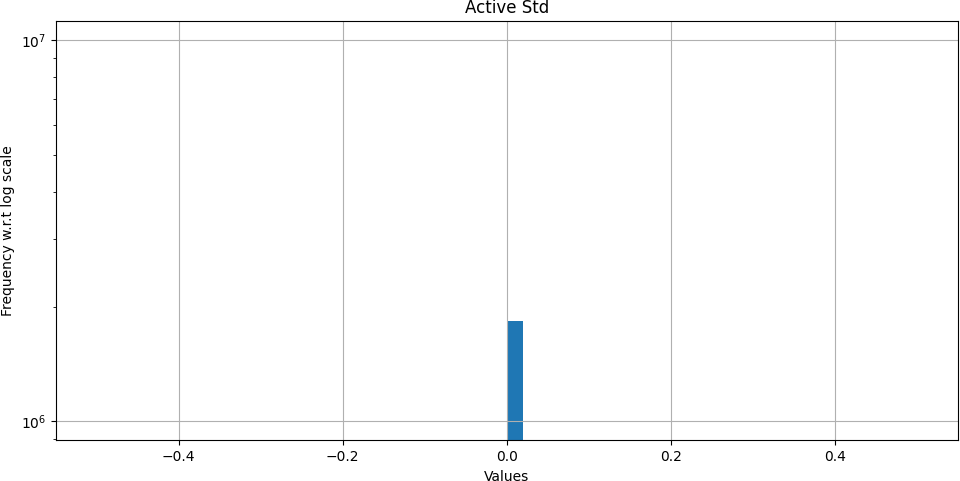


Figure 4.11.51 Histogram of Active Std plotted on log scale after handling negative values and outliers

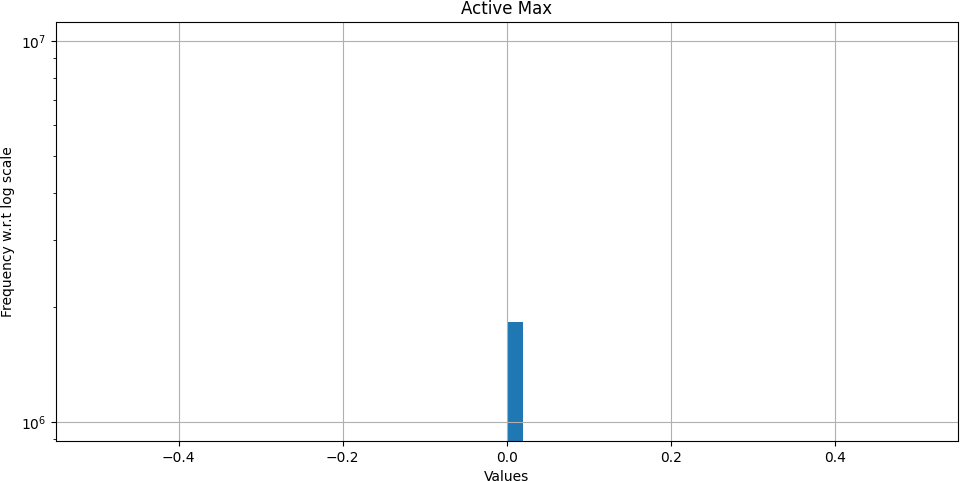


Figure 4.11.52 Histogram of Active Max plotted on log scale after handling negative values and outliers

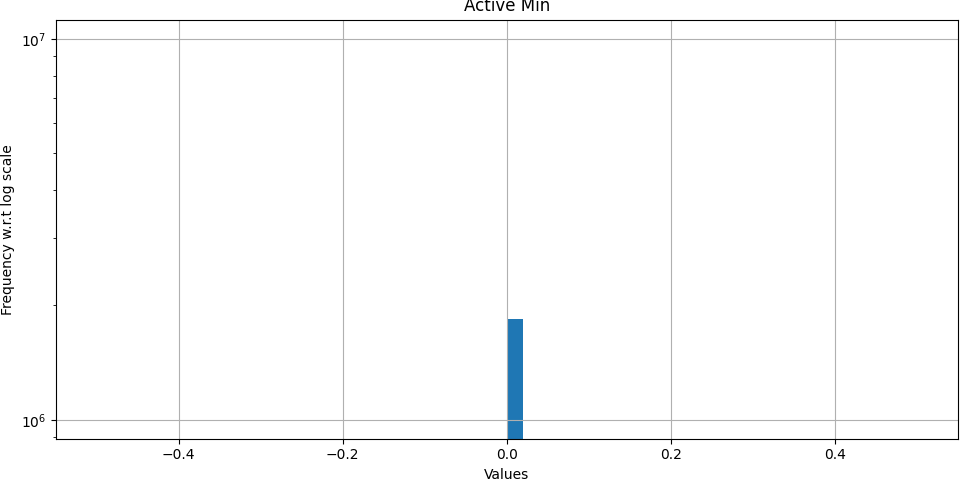


Figure 4.11.53 Histogram of Active Min plotted on log scale after handling negative values and outliers

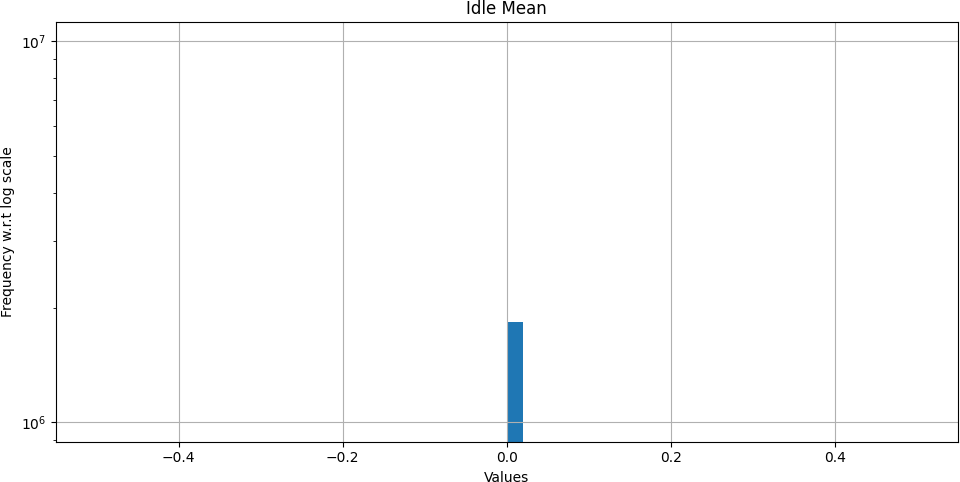


Figure 4.11.54 Histogram of Idle Mean plotted on log scale after handling negative values and outliers

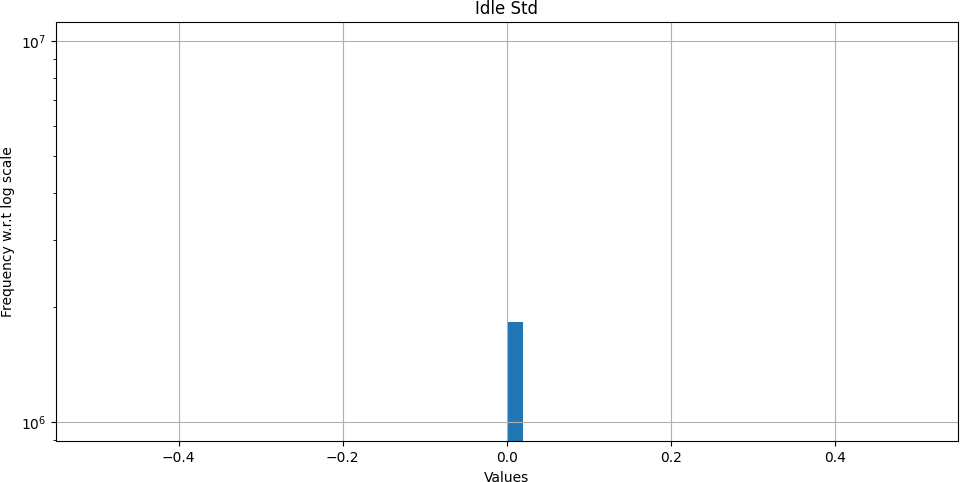


Figure 4.11.55 Histogram of Idle Std plotted on log scale after handling negative values and outliers

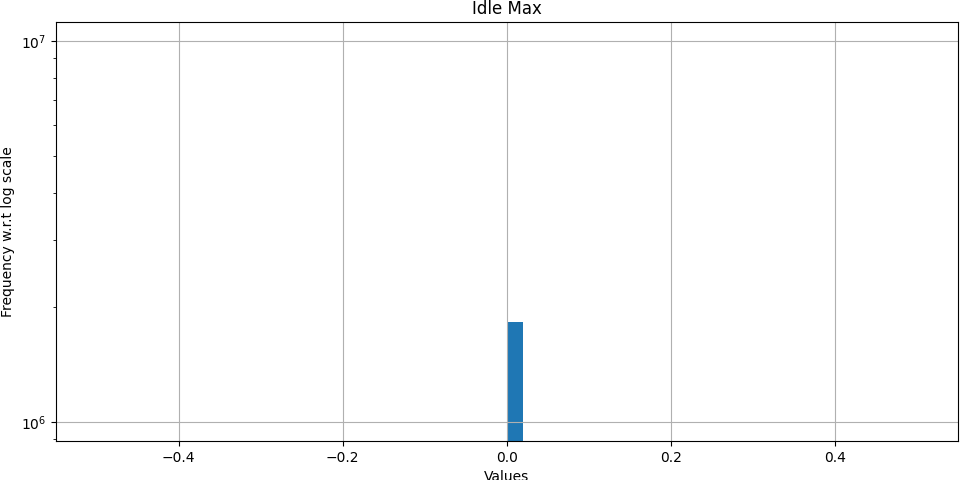


Figure 4.11.56 Histogram of Idle Max plotted on log scale after handling negative values and outliers

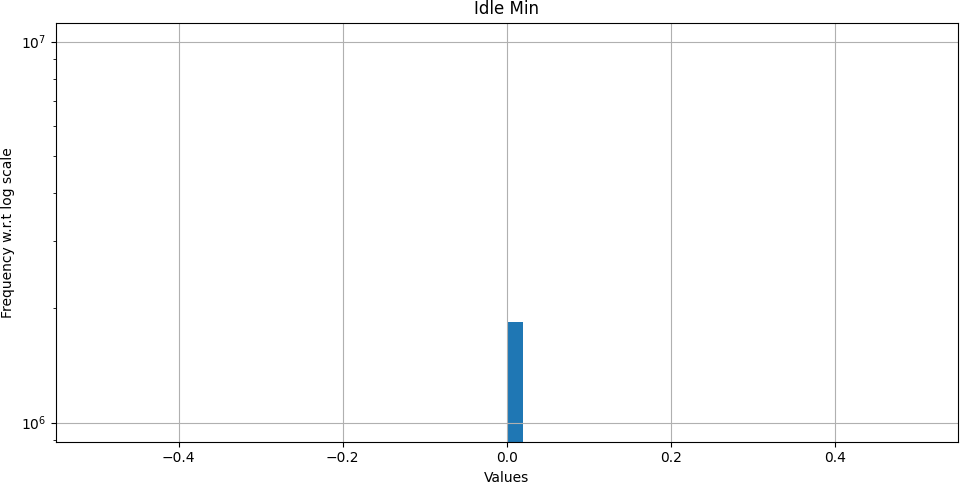


Figure 4.11.57 Histogram of Idle Min plotted on log scale after handling negative values and outliers

Among the new histograms, there were some features with datapoints only in single bin. Thus, they may be the features with single value. Such features will not help to train the classifier model because irrespective of category, those features will remain unchanged.

Thus, the list of features having a single value was fetched: -

1. Fwd PSH Flags
2. SYN Flag Count
3. URG Flag Count
4. Active Mean
5. Active Std
6. Active Max
7. Active Min
8. Idle Mean
9. Idle Std
10. Idle Max
11. Idle Min

The above features were dropped from the dataset because they will not contribute towards training the model.

New shape of the main dataset: (9167271, 48). New shape of sampled dataset: (1833454, 48).

# Pyramid chart with respect to isMalicious: -

Pyramid chart was plotted for each independent feature with respect to the target binary feature: isMalicious.

The continuous data in each feature was transformed into discrete categorical bins and then the charts were plotted to fetch new information from the dataset.

If the number of bins were too less, the graph will be too smooth and thus, no relationship with different ranges of data can be determined.

If the number of bins were too many, we will get a line for almost every datapoint.

Thus, it was essential to determine the optimal number of bins to plot these charts for each independent feature.

Following are commonly known methods to determine the number of bins: -

1. Sturge’s rule
2. Doane’s rule
3. Rice rule
4. Square root rule
5. Scott’s rule
6. Freedman-Diaconis rule
7. Knuth’s rule
8. Scargle’s Bayesian blocks

Bayesian algorithms such as Knuth’s rule and Scargle’s Bayesian blocks are useful when the data

points are skewed, heavy-tailed and have multi-modal distribution.

However, plotting Pyramid charts based on the Bayesian algorithm was not feasible due to

limitations of the system’s configurations.

Thus, to compute the number of bins for each feature, Fredman-Diaconis rule was used.

Freedman-Diaconis rule was selected because: -

* 1. It helps to compute bin width based on each feature’s IQR. Thus, it helps to reduce the impact of skewness in data.
  2. It does not assume the feature to be normally distributed.
  3. Since it uses IQR, the rule also helps to handle values at extreme end and compute optimal the number of

bins.

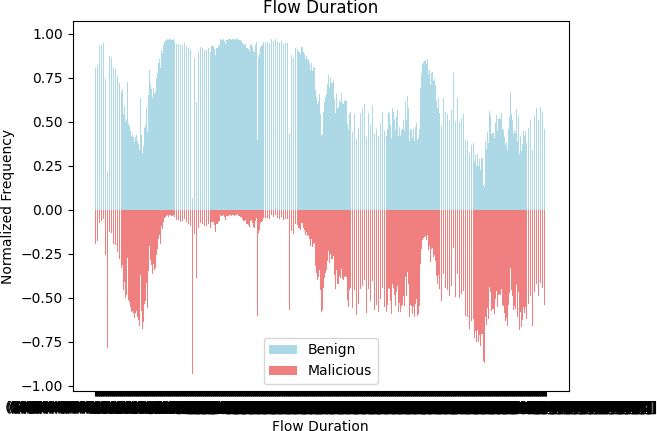


Figure 4.12.1 Pyramid chart of Flow Duration w.r.t isMalicious

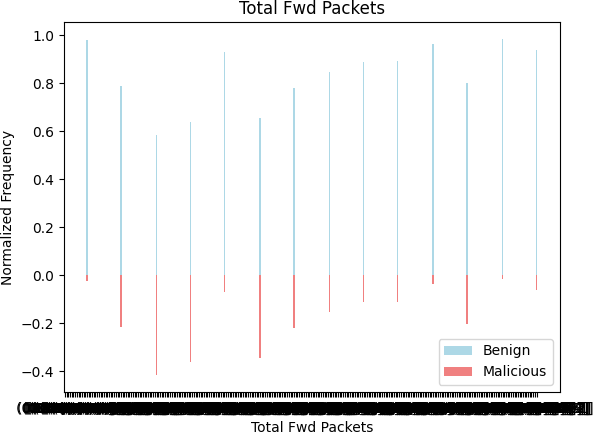


Figure 4.12.2 Pyramid chart of Total Fwd Packets w.r.t isMalicious

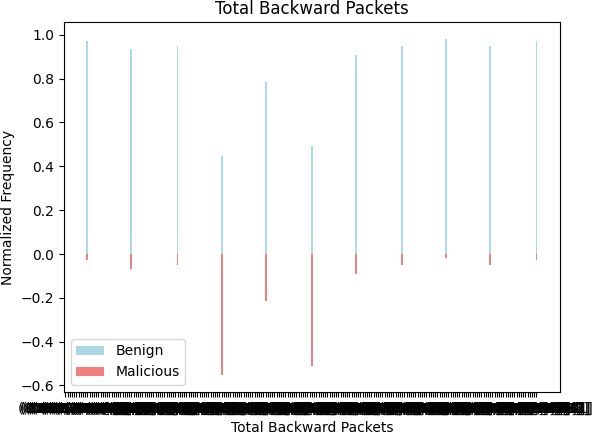


Figure 4.12.3 Pyramid chart of Total Backward Packets w.r.t isMalicious

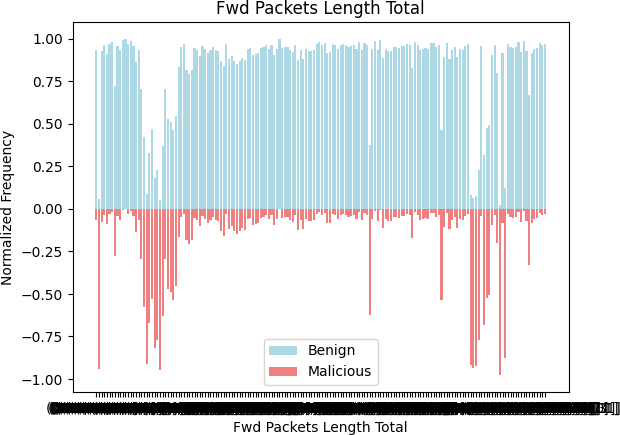


Figure 4.12.4 Pyramid chart of Fwd Packets Length Total w.r.t isMalicious

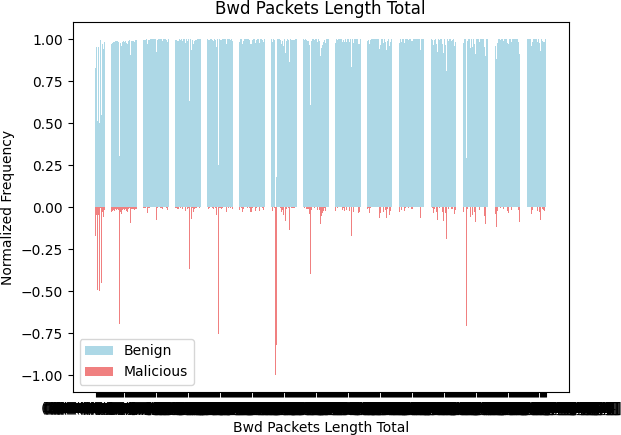


Figure 4.12.5 Pyramid chart of Bwd Packets Length Total w.r.t isMalicious

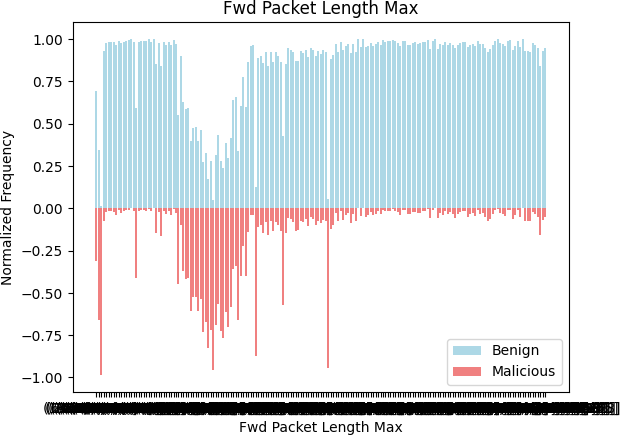


Figure 4.12.6 Pyramid chart of Fwd Packet Length Max w.r.t isMalicious

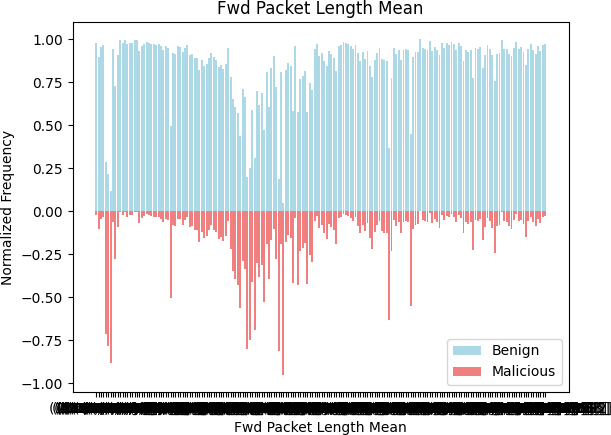


Figure 4.12.7 Pyramid chart of Fwd Packet Length Mean w.r.t isMalicious

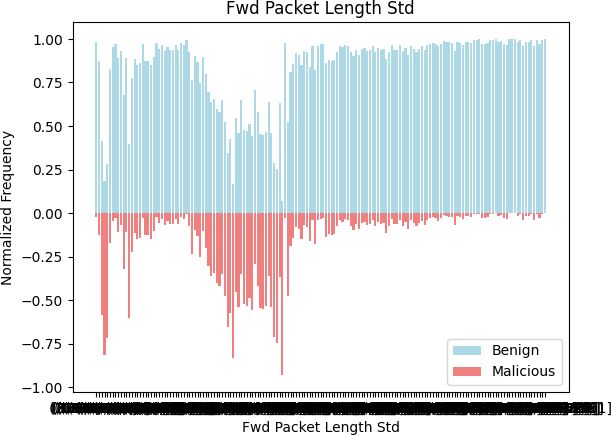


Figure 4.12.8 Pyramid chart of Fwd Packet Length Std w.r.t isMalicious

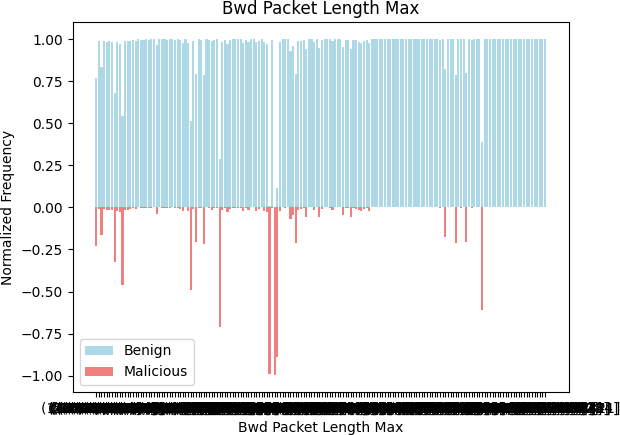


Figure 4.12.9 Pyramid chart of Bwd Packet Length Max w.r.t isMalicious

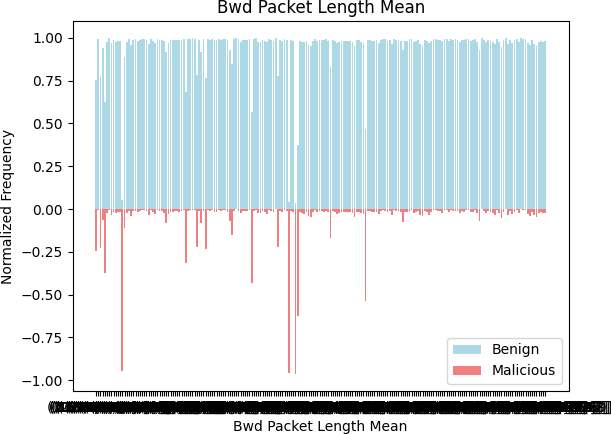


Figure 4.12.10 Pyramid chart of Bwd Packet Length Mean w.r.t isMalicious



Figure 4.12.11 Pyramid chart of Bwd Packet Length Std w.r.t isMalicious

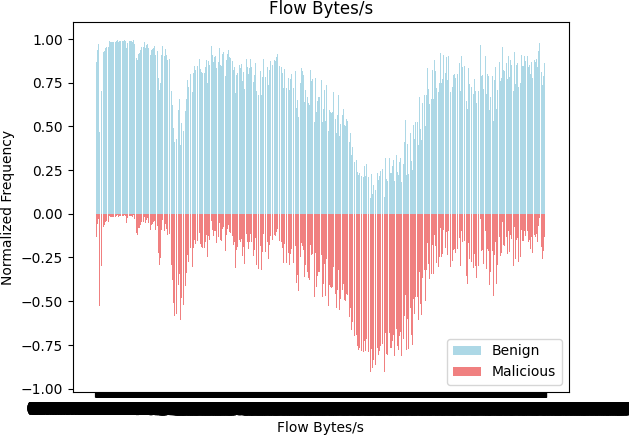


Figure 4.12.12 Pyramid chart of Flow Bytes/s w.r.t isMalicious

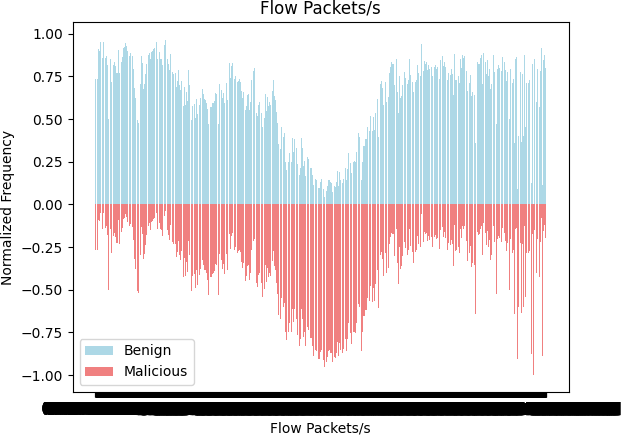


Figure 4.12.13 Pyramid chart of Flow Packets/s w.r.t isMalicious

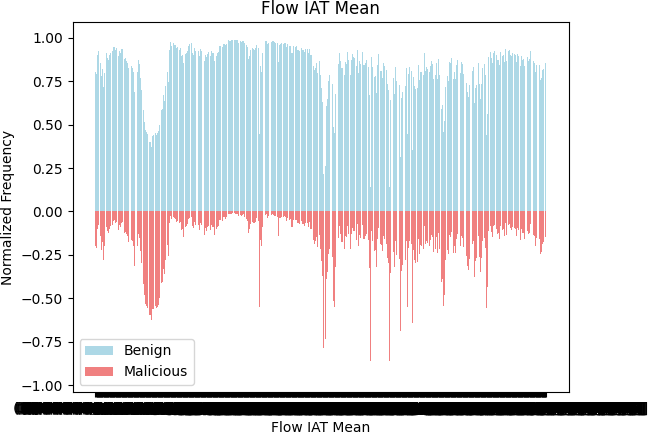


Figure 4.12.14 Pyramid chart of Flow IAT Mean w.r.t isMalicious

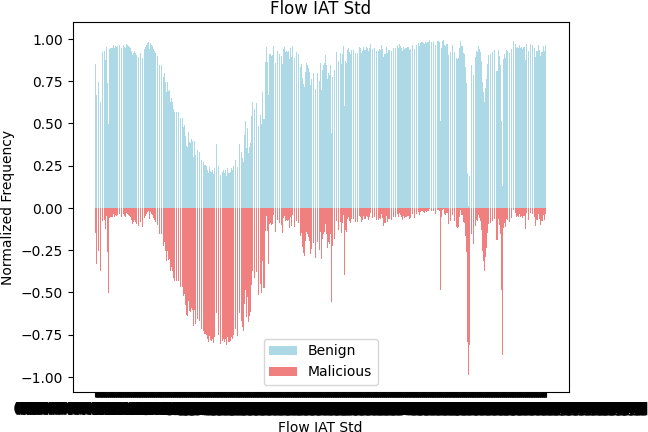


Figure 4.12.15 Pyramid chart of Flow IAT Std w.r.t isMalicious

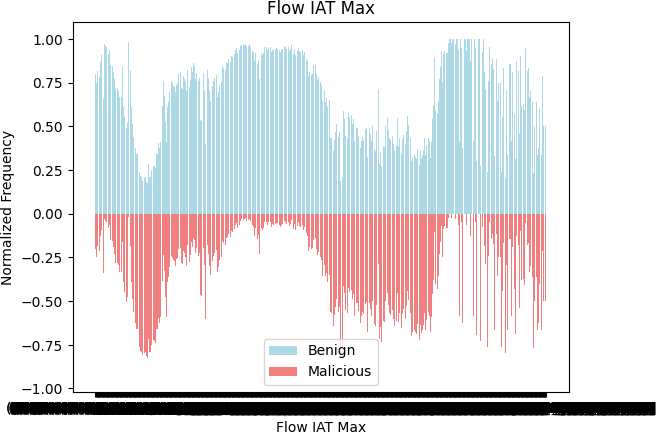


Figure 4.12.16 Pyramid chart of Flow IAT Max w.r.t isMalicious

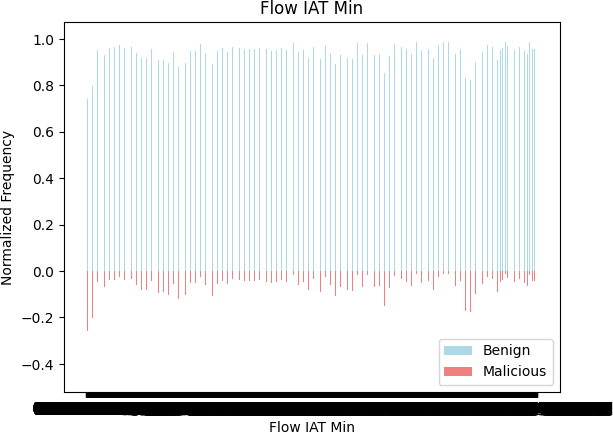


Figure 4.12.17 Pyramid chart of Flow IAT Min w.r.t isMalicious

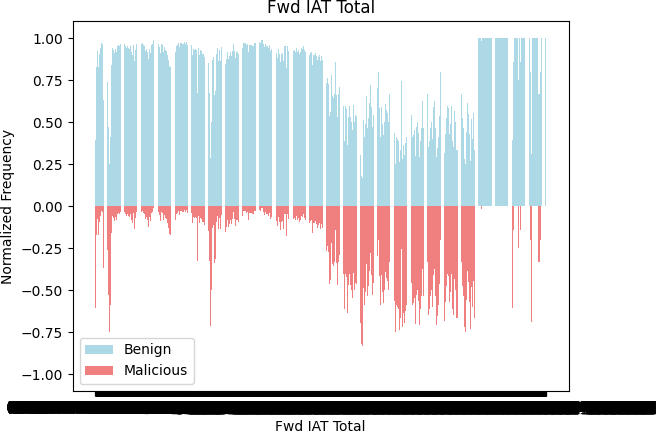


Figure 4.12.18 Pyramid chart of Fwd IAT Total w.r.t isMalicious

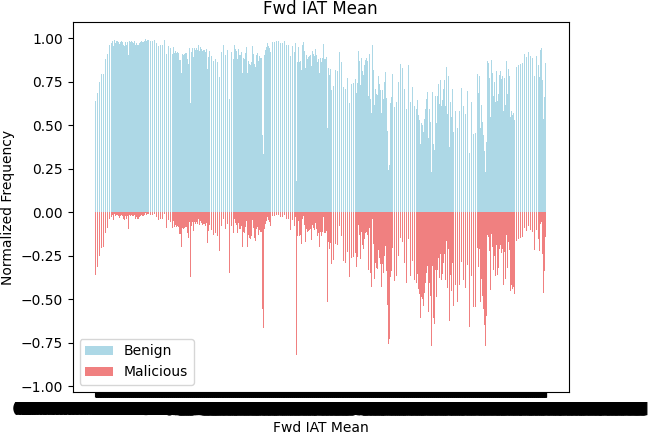


Figure 4.12.19 Pyramid chart of Fwd IAT Mean w.r.t isMalicious

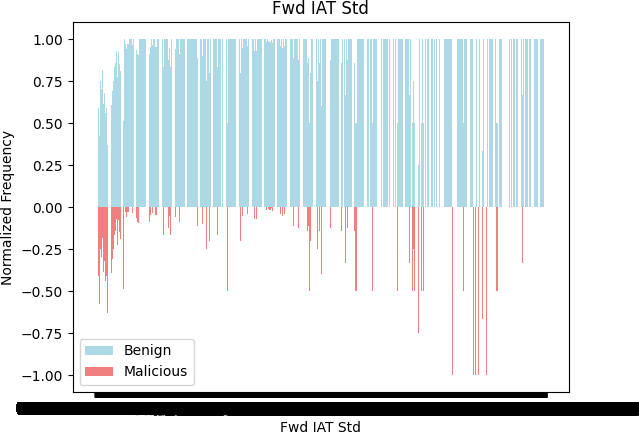


Figure 4.12.20 Pyramid chart of Fwd IAT Std w.r.t isMalicious

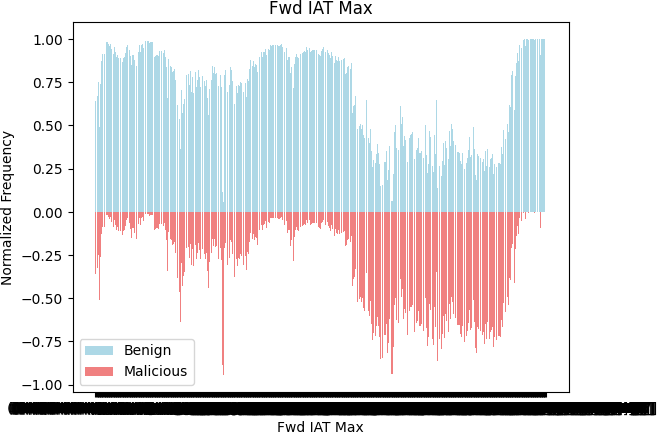


Figure 4.12.21 Pyramid chart of Fwd IAT Max w.r.t isMalicious

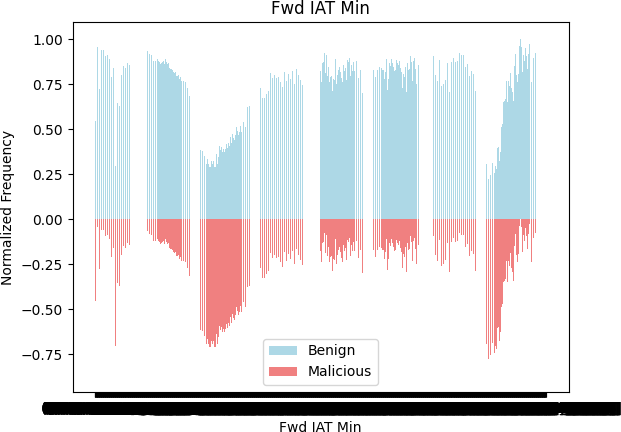


Figure 4.12.22 Pyramid chart of Fwd IAT Min w.r.t isMalicious

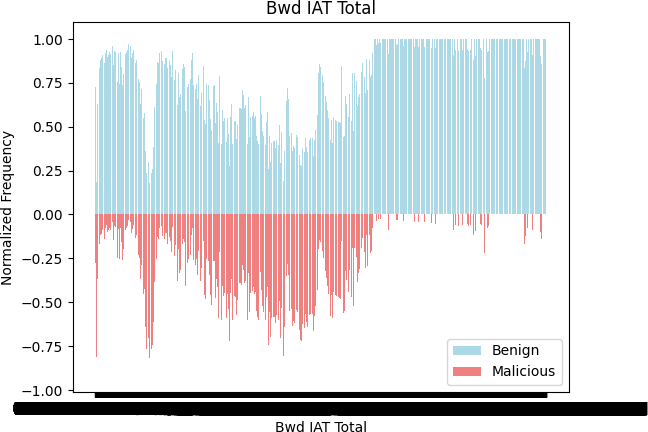


Figure 4.12.23 Pyramid chart of Bwd IAT Total w.r.t isMalicious

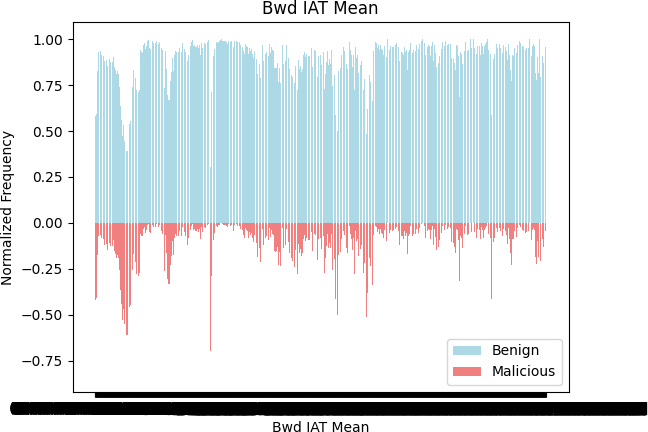


Figure 4.12.24 Pyramid chart of Bwd IAT Mean w.r.t isMalicious

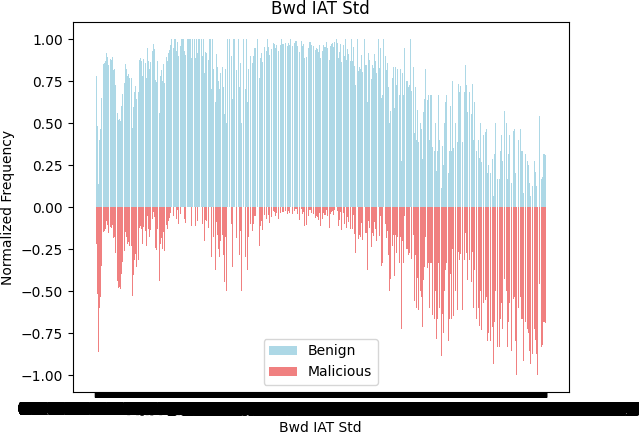


Figure 4.12.25 Pyramid chart of Bwd IAT Std w.r.t isMalicious



Figure 4.12.26 Pyramid chart of Bwd IAT Max w.r.t isMalicious

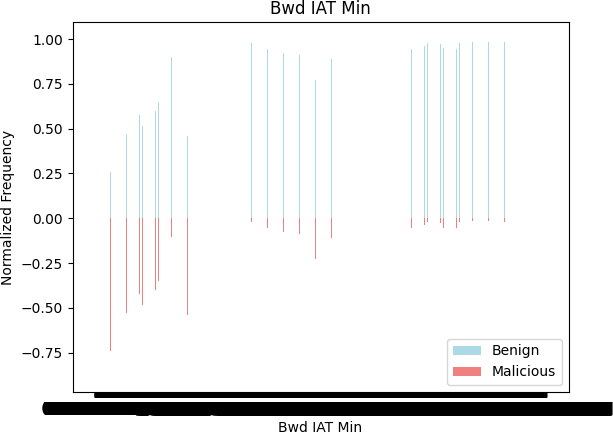


Figure 4.12.27 Pyramid chart of Bwd IAT Min w.r.t isMalicious

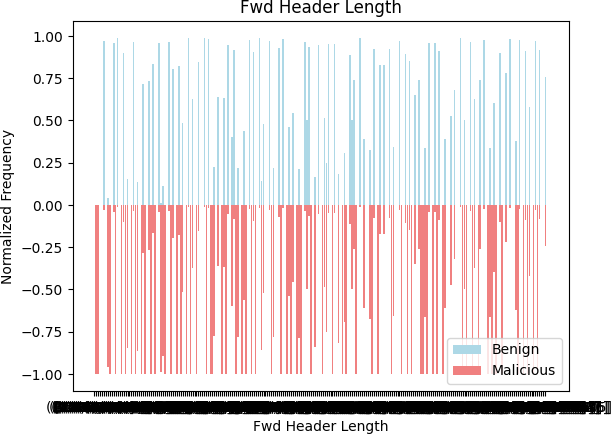


Figure 4.12.28 Pyramid chart of Fwd Header Lenngth w.r.t isMalicious

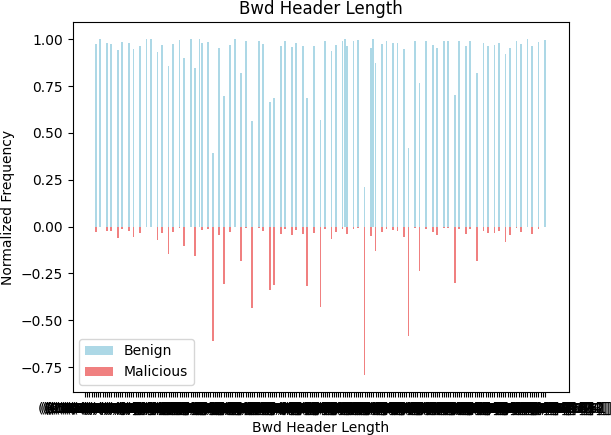


Figure 4.12.29 Pyramid chart of Bwd Header Lenngth w.r.t isMalicious

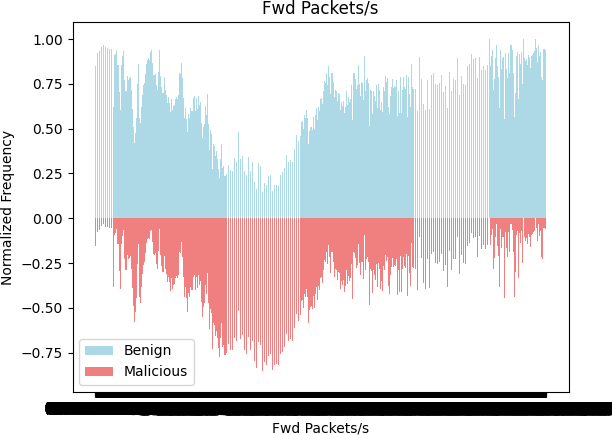


Figure 4.12.30 Pyramid chart of Fwd Packets/s w.r.t isMalicious

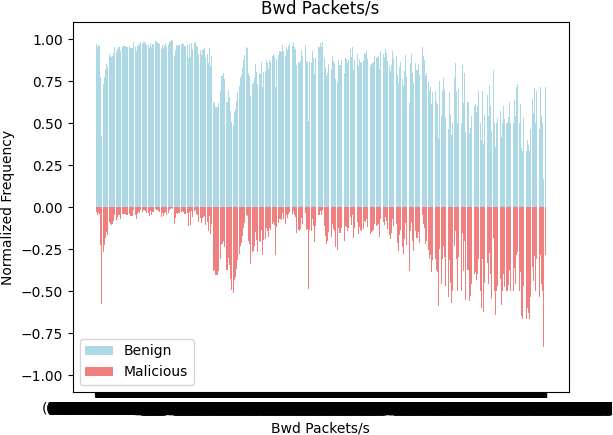


Figure 4.12.31 Pyramid chart of Bwd Packets/s w.r.t isMalicious

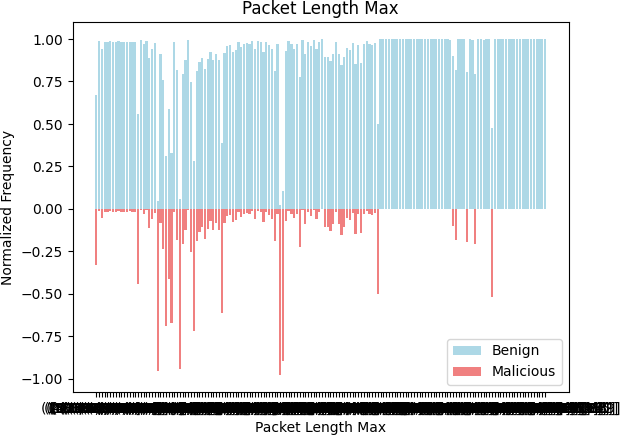


Figure 4.12.32 Pyramid chart of Packet Length Max w.r.t isMalicious

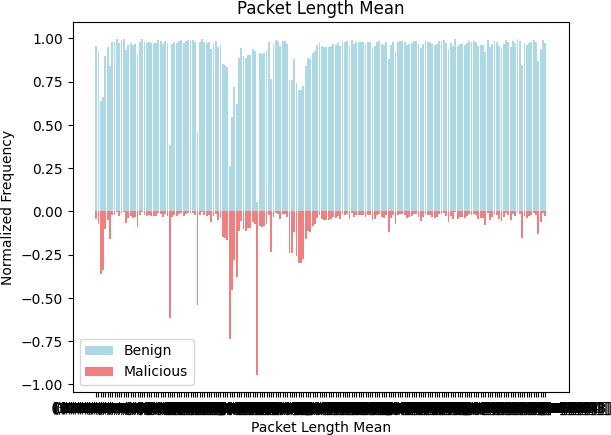


Figure 4.12.33 Pyramid chart of Packet Length Mean w.r.t isMalicious

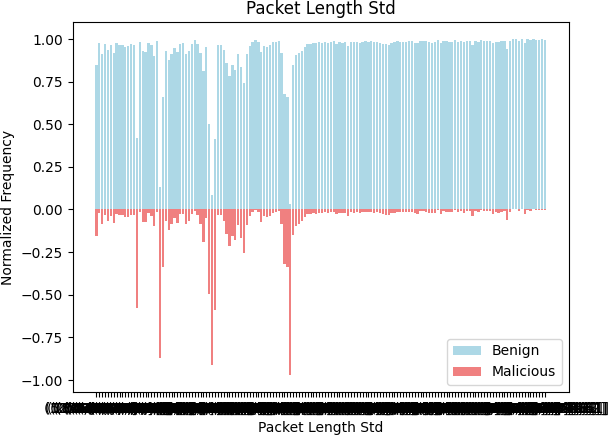


Figure 4.12.34 Pyramid chart of Packet Length Std w.r.t isMalicious

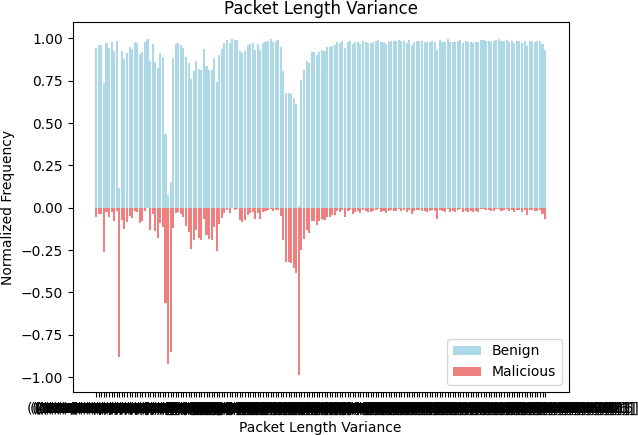


Figure 4.12.35 Pyramid chart of Packet Length Variance w.r.t isMalicious

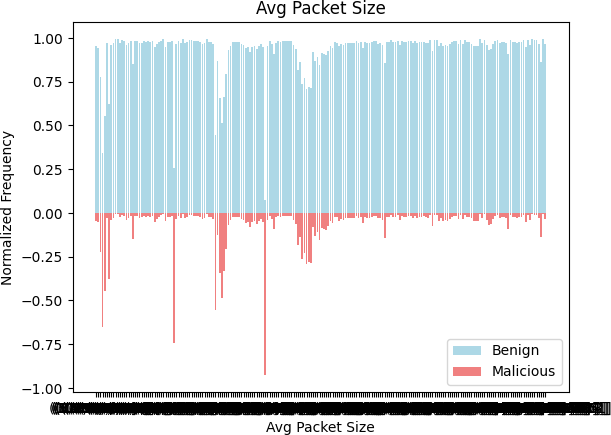


Figure 4.12.36 Pyramid chart of Avg Packet Size w.r.t isMalicious

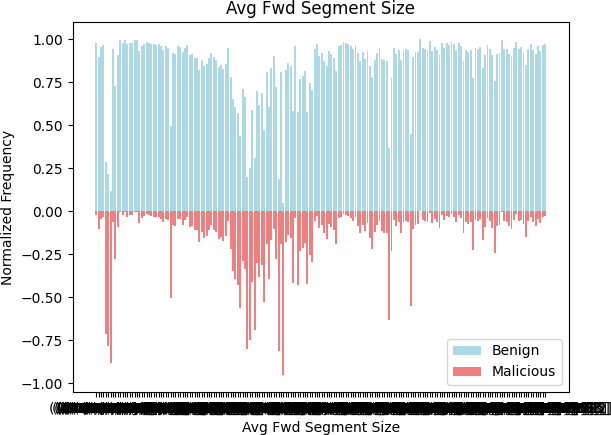


Figure 4.12.37 Pyramid chart of Avg Fwd Segment Size w.r.t isMalicious

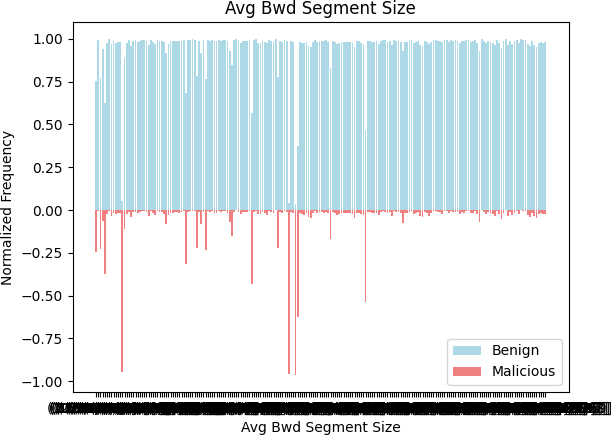


Figure 4.12.38 Pyramid chart of Avg Bwd Segment Size w.r.t isMalicious

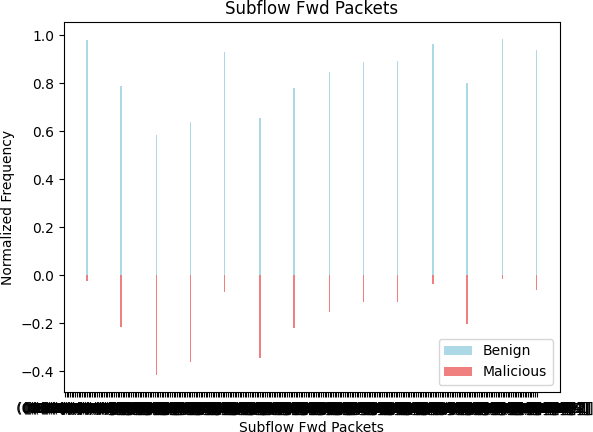


Figure 4.12.39 Pyramid chart of Subflow Fwd Packets w.r.t isMalicious

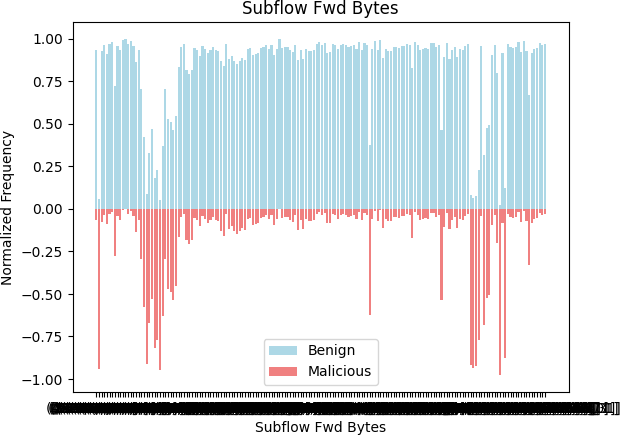


Figure 4.12.40 Pyramid chart of Subflow Fwd Bytes w.r.t isMalicious

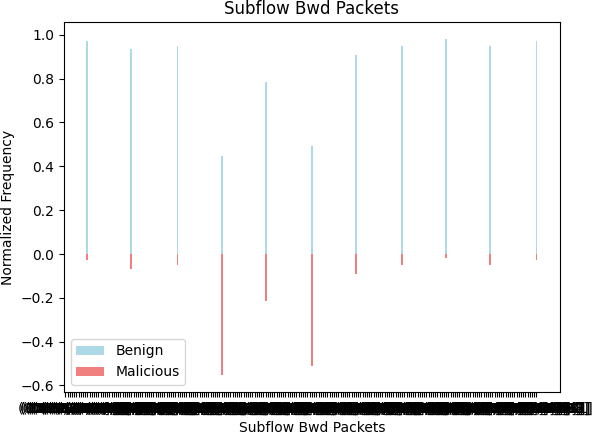


Figure 4.12.41 Pyramid chart of Subflow Bwd Packets w.r.t isMalicious

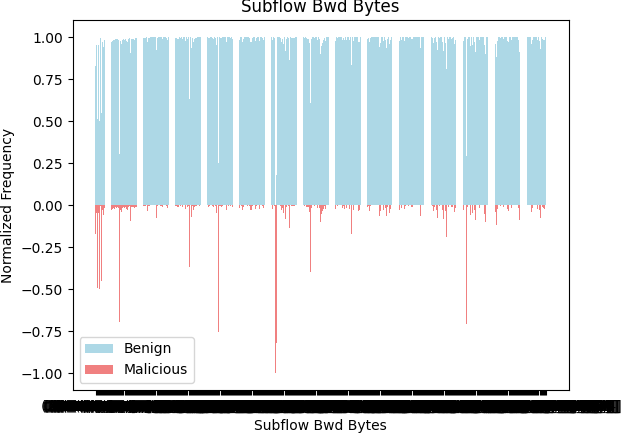


Figure 4.12.42 Pyramid chart of Subflow Bwd Bytes w.r.t isMalicious

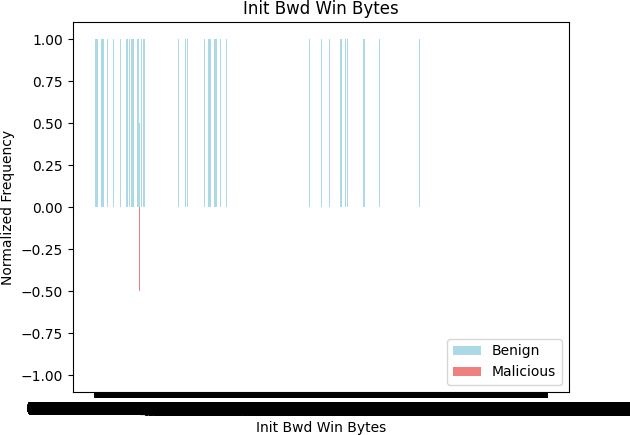


Figure 4.12.43 Pyramid chart of Init Bwd Win Bytes w.r.t isMalicious

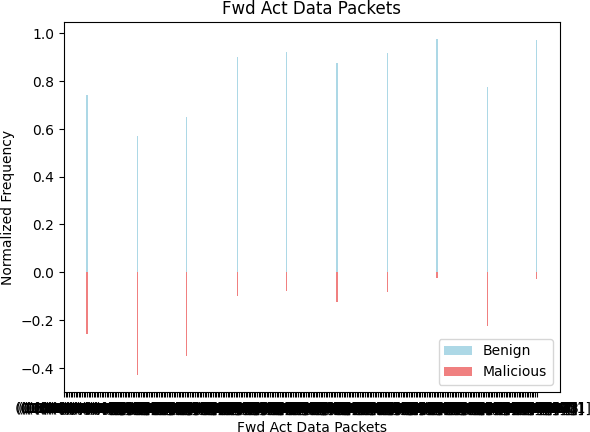


Figure 4.12.44 Pyramid chart of Fwd Act Data Packets w.r.t isMalicious

Following features have almost equal number of Malicious and Benign records in most of the bins: -

1. Flow Duration
2. Flow IAT Max
3. Fwd Header Length

Following features have some bins where number of Malicious records are relatively more than the

number of Benign records and thus, change in patterns were observed over a set of bins: -

1. Flow Bytes/s
2. Flow Packets/s
3. Flow IAT Std
4. Fwd IAT Max
5. Bwd IAT Std
6. Bwd IAT Max
7. Fwd Packets/s
8. Bwd Packets/s

‘Init Bwd Win Bytes’ was a rare feature which had only 1 bin with Malicious records and rest all bins had Benign records.

Remaining all features have relatively very high number of Benign records compared to Malicious records in most of the bins.

While carrying out the above interpretation small variations and changes were not recorded as decisions based on minor changes may result in incorrect analysis. Only the patterns which were thick and broadly visible were recorded from Pyramid charts plotted with respect to target binary feature: isMalicious.

# Label encoding: -

Label encoding on target feature: ClassLabel was done and results were stored in a new feature: attack\_id. Thus, after label encoding: -

Table 4.13.1: Encoded values of ClassLabel

|  |  |
| --- | --- |
| **ClassLabel** | **attack\_id** |
| Benign | 0 |
| Botnet | 1 |
| Bruteforce | 2 |
| DDoS | 3 |
| DoS | 4 |
| Infiltration | 5 |
| Portscan | 6 |
| Webattack | 7 |

# Correlation matrix: -

Using the sampled dataset, correlation matrix could not get plotted due to limitations of system’s

configurations, which led to insufficient memory error.

As the result, 20% of records from sampled dataset (4% of the original dataset) were taken and used to plot the heat map for correlation matrix, with target feature as attack\_id.

Shape of the new sub-sampled dataset on which correlation matrix was computed: (366690, 47).

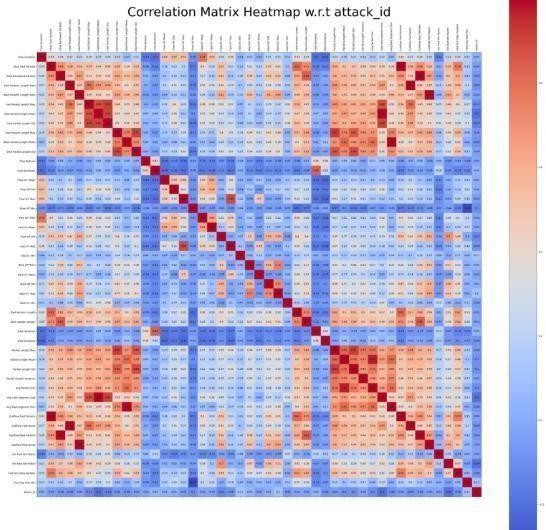


Figure 4.14.1 Correlation matrix based on 4% of the original dataset.

Observations from the above heat map: -

1. Many independent features have red and dark red squares, indicating strong relation among them.
2. Some of the examples are: -
   * Fwd Packet Length Max – Fwd Packet Length Mean = 0.88
   * Fwd Packet Length Max – Fwd Packet Length Std = 0.91
   * Bwd Packet Length Std – Packet Length Max = 0.91
   * Bwd Packet Length Std – Packet Length Mean = 0.71
3. All independent features have weak relation with attack\_id.

However, due to sub-sampled dataset used for plotting the correlation matrix, it was difficult to determine whether to use the results observed in correlation matrix on the main dataset or on the sampled dataset. As the result, the results of the above heat map were not used.

# Renaming of feature names: -

The columns were renamed for ease of use my replacing space with underscore.

# Analysis based on descriptive statistics: -

Three lists were created: -

1. columns\_equal\_min\_and\_Q1 = Features with equal minimum and Q1 value.
2. columns\_equal\_Q1\_and\_Q3 = Features with equal Q1 and Q3 value.
3. columns\_equal\_Q3\_and\_max = Features with equal Q3 and maximum value.

Following features were captured in columns\_equal\_min\_and\_Q1: -

1. Bwd\_Packet\_Length\_Total
2. Fwd\_Packet\_Length\_Std
3. Bwd\_Packet\_Length\_Max
4. Bwd\_Packet\_Length\_Mean
5. Bwd\_Packet\_Length\_Std
6. Flow\_IAT\_Std
7. Fwd\_IAT\_Std
8. Bwd\_IAT\_Total
9. Bwd\_IAT\_Mean
10. Bwd\_IAT\_Std
11. Bwd\_IAT\_Max
12. Bwd\_IAT\_Min
13. Avg\_Bwd\_Segment\_Size
14. Subflow\_Bwd\_Bytes
15. Fwd\_Act\_Data\_Packets

Thus, for the above list of features it was inferred that a large number of records are clustered in lower range.

These features may have many zero values or many constant values in lower range of data points. For all of the 15 features, minimum value and Q1 value equal to 0.0

Thus, 25% of the values are zero in all of the 15 features. And thus, the features may also be categorized as Zero-inflated features due to their high percentage of zero values.

Since the features are having data concentrated in lower range, they are positively skewed. The results were grouped into two categories: Non-zero, Zero.

1. Non-zero: Data points having value not equal to 0.
2. Zero: Data points having value equal to 0.

Based on the above two categories, the frequency of data points with respect to the target binary feature: isMalicious was plotted for each feature.

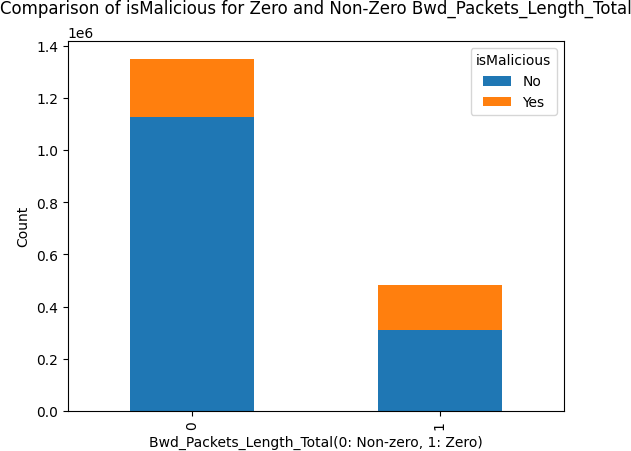


Figure 4.16.1 Stacked bar chart for Bwd Packets Length Total plotted for values which are zero and non-zero w.r.t isMalicious

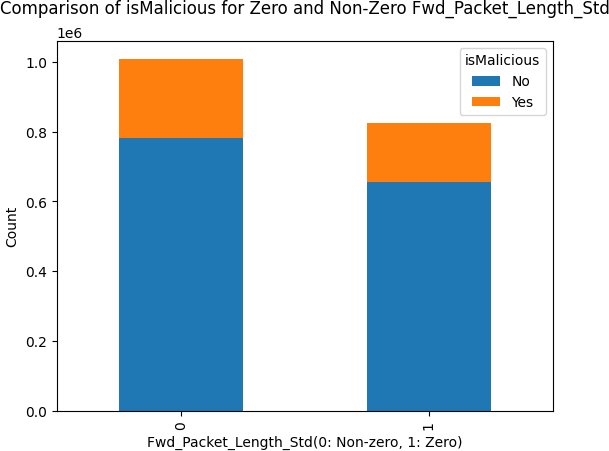


Figure 4.16.2 Stacked bar chart for Fwd Packet Length Std plotted for values which are zero and non-zero w.r.t isMalicious

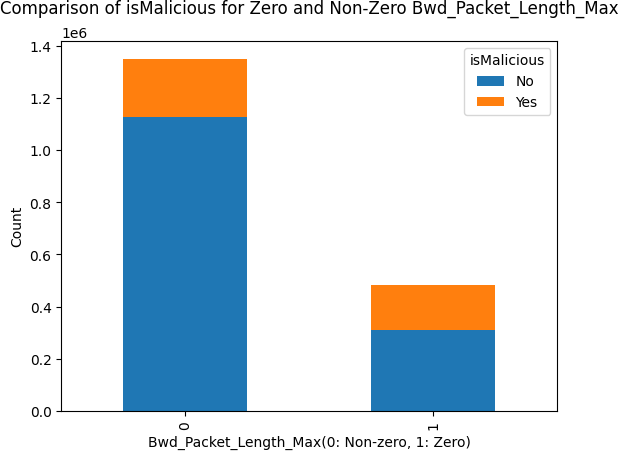


Figure 4.16.3 Stacked bar chart for Bwd Packet Length Max plotted for values which are zero and non-zero w.r.t isMalicious

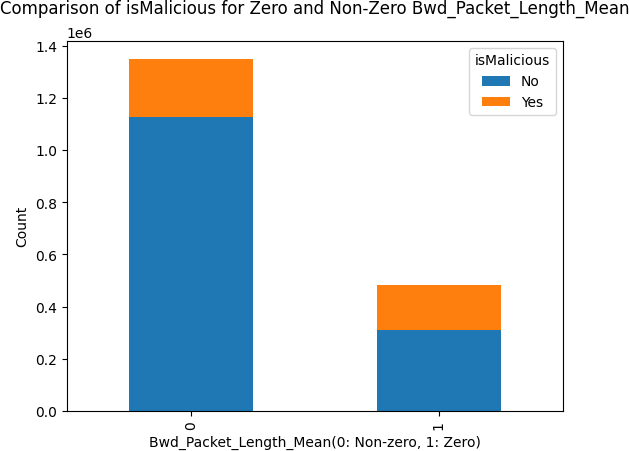


Figure 4.16.4 Stacked bar chart for Bwd Packet Length Mean plotted for values which are zero and non-zero w.r.t isMalicious

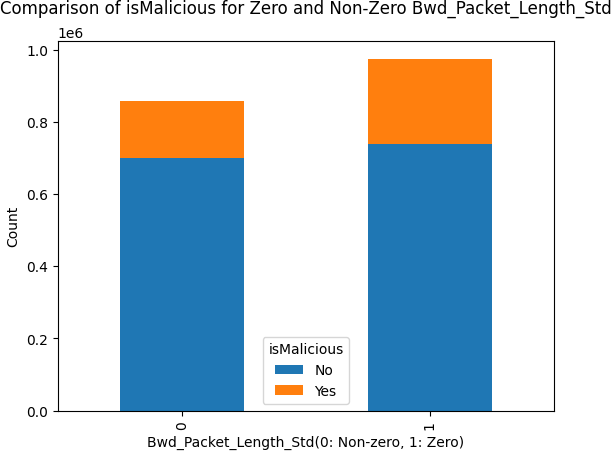


Figure 4.16.5 Stacked bar chart for Bwd Packet Length Std plotted for values which are zero and non-zero w.r.t isMalicious



Figure 4.16.6 Stacked bar chart for Flow IAT Std plotted for values which are zero and non-zero w.r.t isMalicious

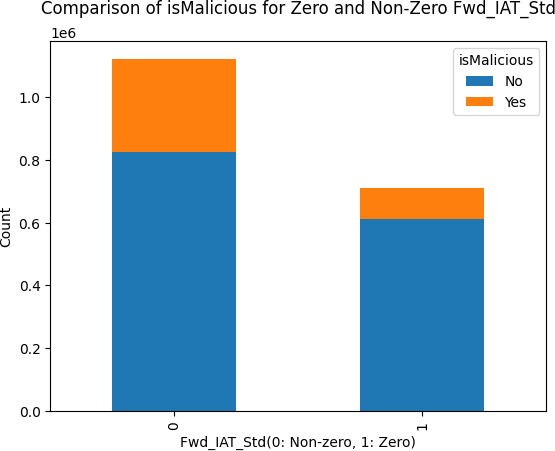


Figure 4.16.7 Stacked bar chart for Fwd IAT Std plotted for values which are zero and non-zero w.r.t isMalicious

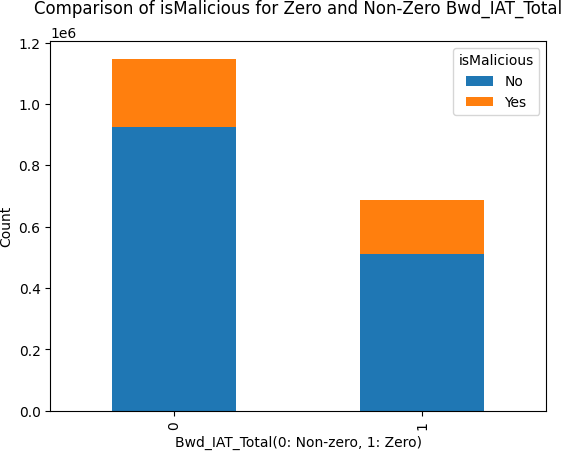


Figure 4.16.8 Stacked bar chart for Bwd IAT Total plotted for values which are zero and non-zero w.r.t isMalicious

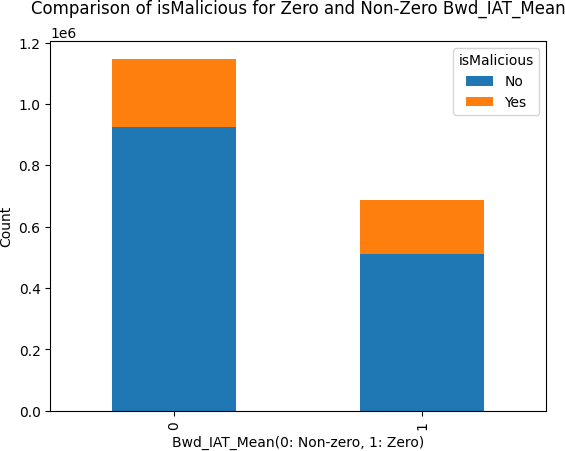


Figure 4.16.9 Stacked bar chart for Bwd IAT Mean plotted for values which are zero and non-zero w.r.t isMalicious

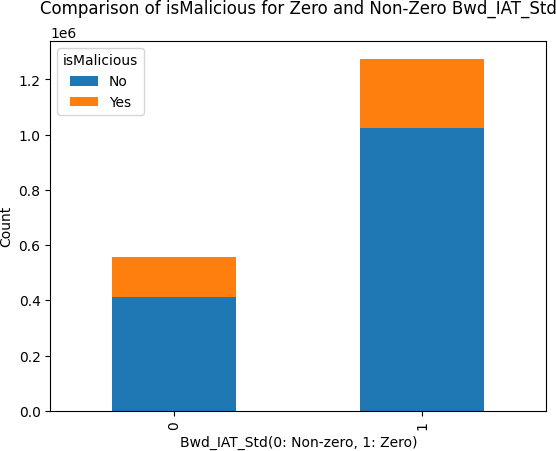


Figure 4.16.10 Stacked bar chart for Bwd IAT Std plotted for values which are zero and non-zero w.r.t isMalicious

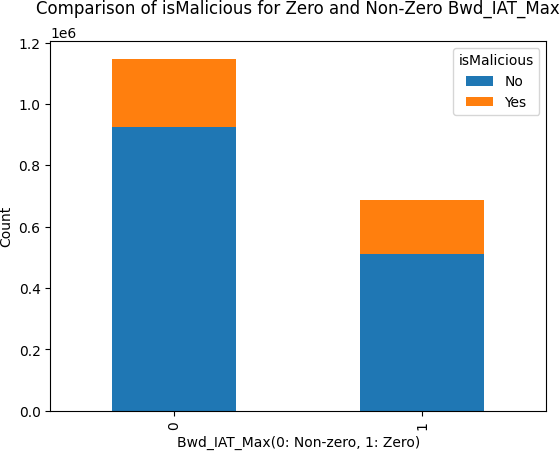


Figure 4.16.11 Stacked bar chart for Bwd IAT Max plotted for values which are zero and non-zero w.r.t isMalicious

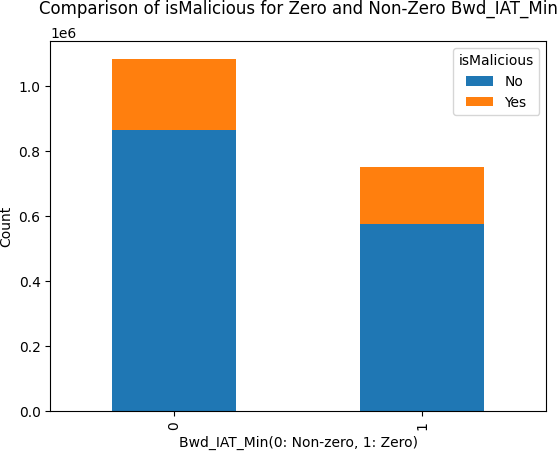


Figure 4.16.12 Stacked bar chart for Bwd IAT Min plotted for values which are zero and non-zero w.r.t isMalicious

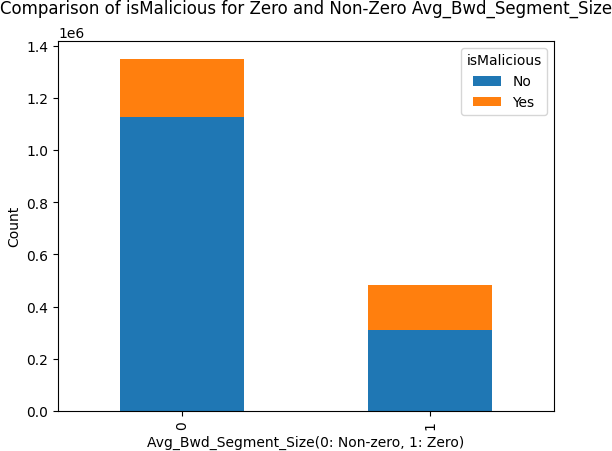


Figure 4.16.13 Stacked bar chart for Avg Bwd Segment Size plotted for values which are zero and non-zero w.r.t isMalicious

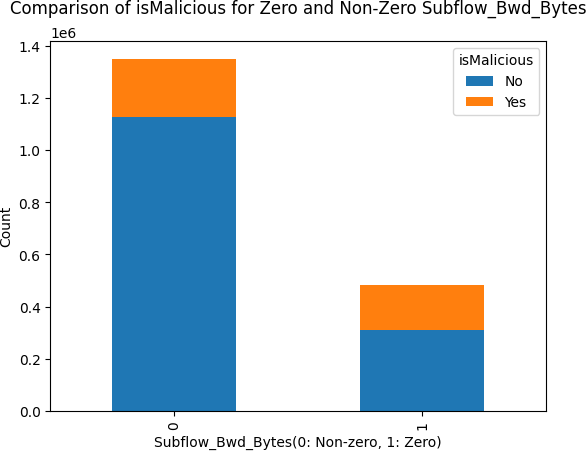


Figure 4.16.14 Stacked bar chart for Subflow Bwd Bytes plotted for values which are zero and non-zero w.r.t isMalicious

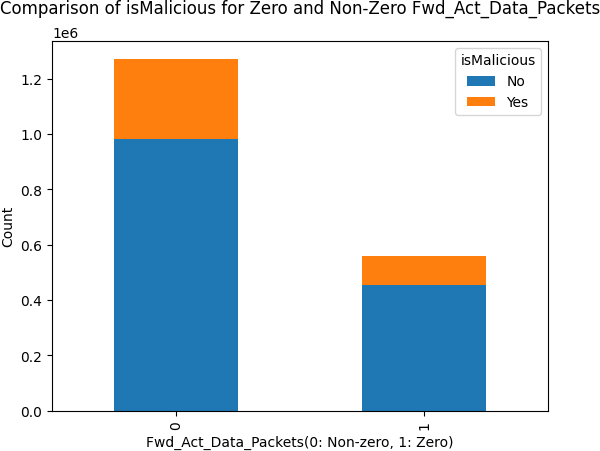


Figure 4.16.15 Stacked bar chart for Fwd Act Data Packets plotted for values which are zero and non-zero w.r.t isMalicious

There was no differentiation found among the 15 features to get more information to identify malicious events based on zero and non-zero values.

Following features were captured in columns\_equal\_Q1\_and\_Q3: -

1. Init\_Fwd\_Win\_Bytes
2. Fwd\_Seg\_Size\_Min

Thus, for the above list of features it was inferred that 50% of the datapoints are clustered at a single value.

Init\_Fwd\_Win\_Bytes: Q1=Q2=Q3=8192.0 Fwd\_Seg\_Size\_Min: Q1=Q2=Q3=20.0

These features may have very low variability and many constant values.

The results were grouped into two categories: Not mid-range, Mid-range. Not mid-range: Data points having value not equal to median (Q2).

Mid-range: Data points having value equal to median (Q2).

Based on the above two categories, the frequency of data points with respect to the target binary feature: isMalicious was plotted for each feature.

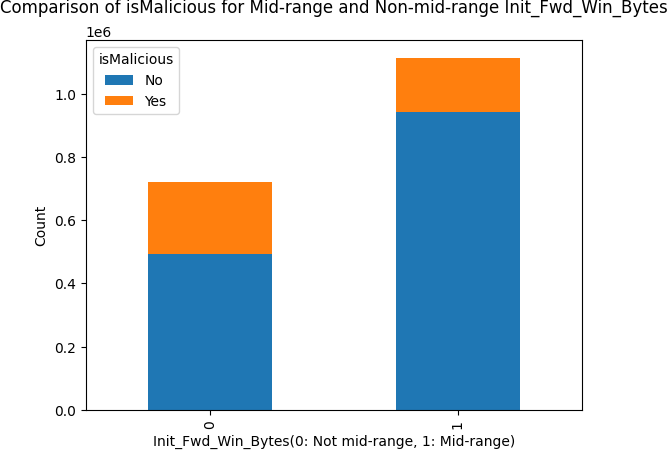


Figure 4.16.16 Stacked bar chart for Init Fwd Win Bytes plotted for values which are mid-range and not mid- range w.r.t isMalicious

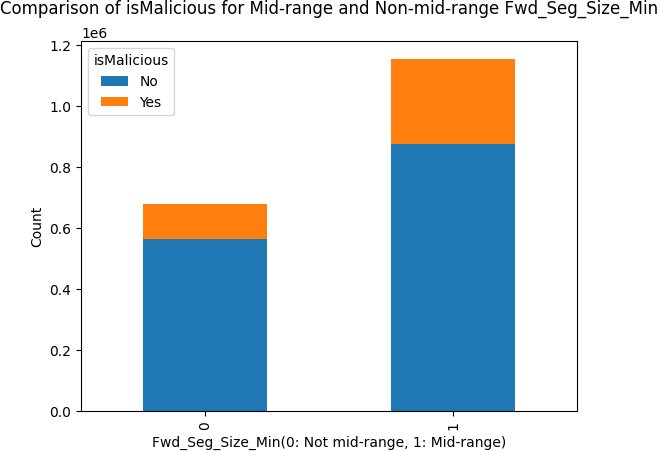


Figure 4.16.17 Stacked bar chart Fwd Seg Size Min plotted for values which are mid-range and not mid- range w.r.t isMalicious

There was no differentiation found among the 2 features to get more information to identify malicious events based on mid-range and not mid-range values.

No features were captured in columns\_equal\_Q3\_and\_max.

Thus, from the above it was inferred that all features in upper range have high variability and are

spread out. As the result, there are no negatively skewed features in the dataset.

# Fetching most category of records: -

Number of records for each category of event in ClassLabel were fetched in the sampled dataset:

* + - Benign : 1437467
    - DDoS : 246982
    - DoS : 79186
    - Botnet : 29348
    - Bruteforce : 20546
    - Infiltration : 18870
    - Webattack : 625
    - Portscan : 430

It was checked if all records of each category are unique or duplicate.

* + - 1437467 records for Benign were duplicate.
    - 246982 records for DDoS were duplicate.
    - 79186 records for DoS are duplicate.
    - 18870 records for Infiltration are duplicate.
    - 625 records for Webattack are duplicate.
    - 29348 records for Botnet are unique.
    - 20546 records for Bruteforce are unique.
    - 430 records for Portscan are unique.

A subset of data was selected: -

* + - 1. Top two categories having duplicate records: Benign, DDoS.
      2. Top two categories having unique records: Botnet, Bruteforce. Reason: -

1. Given the size of sampled dataset, it becomes extremely difficult to perform further processing and tasks such as feature selection and training the model.
2. Webattack and Portscan have too less number of records compared to other categories for training the model. Thus, it becomes extremely difficult to have a common model that can be trained to identify events with vast difference in frequency.

New shape of the sampled dataset: (1734343, 49).

Due to limitations of system’s configurations and memory, the sampled dataset will be used further for training the model.

# Label encoding on newly sampled dataset: -

Since the label encoding was previously performed on ClassLabel and stored the results in attack\_id, after dropping the rows, the encoded values will have gap.

Thus, the attack\_id column was dropped, label encoding was again performed on ClassLabel, and new results were stored in attack\_id.

Table 4.18.1: Encoded values of ClassLabel after dropping the rows

|  |  |
| --- | --- |
| **ClassLabel** | **attack\_id** |
| Benign | 0 |
| Botnet | 1 |
| Bruteforce | 2 |
| DDoS | 3 |

The column ClassLabel was dropped, since its equivalent numerical feature is available in the form of attack\_id.

Thus, at this stage the two target features in the dataset are: -

* + - isMalicious: For binary classification
    - attack\_id: For multi-class classification

# Dropping the feature: ‘Init Bwd Win Bytes’: -

Based on the results of Pyramid chart, we observed ‘Init Bwd Win Bytes’ will not enable to train the classifier model for differentiating malicious events from benign events.

Thus, we dropped the feature from our dataset. New shape of the sampled dataset: (1734343, 47).

# Writing pre-processed data in a new file: -

The dataset was written in a new file: processed\_dataaset.parquet

This will allow to perform further activities on a new notebook and prevent the overhead of loading the complete dataset, and running all the tasks performed for pre-processing, analysis and feature engineering.

# Handling imbalanced nature of dataset: -

Two types of classification models need to be built: -

1. binary\_cic\_df
2. multiclass\_cic\_df

For binary\_cic\_df: -

Table 4.21.1: Distribution of records in sampled dataset based on isMalicious

|  |  |
| --- | --- |
| **isMalicious** | **Number of records** |
| 0 | 1437467 |
| 1 | 296876 |

For multiclass\_cic\_df: -

Table 4.21.2: Distribution of records in sampled dataset based on attack\_id

|  |  |
| --- | --- |
| **attack\_id** | **Number of records** |
| 0 | 1437467 |
| 1 | 29348 |
| 2 | 20546 |
| 3 | 246982 |

Since dataset is imbalanced, building classifier models using it will result in: -

1. Biased predictions
2. Low sensitivity
3. Poor generalization

Approaches to handle imbalanced nature of dataset: -

1. Oversampling of minority class
   1. Here we create duplicate of records having minority class and make the count same as majority class.
2. Undersampling of majority class
   1. Here we reduce the records of majority class and make the count same as minority class.
3. Cost sensitive learning
   1. The algorithm is forced to correctly identify minority class by adding penalty for incorrect classification.
4. Anomaly detection approach
   1. The minority class is treated as an anomaly and majority class is treated as baseline.
   2. Algorithms that can help with this approach: -
      1. Isolation forest
      2. One-class SVM