**Data pre-processing, analysis and feature engineering**

Following activities were carried out on the dataset: -

1. Shape of the dataset: (9167581, 59).
2. None of the features have NA values.
3. 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.

1. 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.
2. Thus, again the histograms were plotted using log scale which helped to find pattern of distribution for each feature in the dataset.

Observations and interpretations from above Histograms with Logarithmic scale: -

1. Flow Duration: The duration of the flow
   * 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
2. Total Fwd Packets: Total number of forward packets
   * 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.
3. Total Backward Packets: Total number of backward packets
   * 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.
4. Fwd Packets Length Total: Total length of forward packets
   * 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.
5. Bwd Packets Length Total: Total length of backward packets
   * 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.
6. Fwd Packet Length Max: Maximum length of forward packets
   * 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.
7. Fwd Packet Length Mean: Mean length of forward packets
   * 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.
8. Fwd Packet Length Std: Standard deviation length of forward packets
   * 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.
9. Bwd Packet Length Max: Maximum length of backward packets
   * 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.
10. Bwd Packet Length Mean: Mean length of backward packets
    * 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.
11. Bwd Packet Length Std: Standard deviation length of backward packets
    * 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.
12. Flow Bytes/s: Flow bytes per second
    * 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.
13. Flow Packets/s: Flow packets per second
    * 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.
14. Flow IAT Mean: Mean time between flows
    * Peak was observed at Flow IAT Mean=0.
    * Most values are concenterated 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.
15. Flow IAT Std: Standard deviation of time between flows
    * The distribution is skewed towards right: Positively skewed.
    * Peak was observed around Flow IAT Std=0.
    * Most values are concenterated 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.
16. Flow IAT Max: Maximum time between flows
    * 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
17. Flow IAT Min: Minimum time between flows
    * Peak was observed around Flow 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.
18. Fwd IAT Total: Total time between forward packets
    * 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.
19. Fwd IAT Mean: Mean time between forward packets
    * 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.
20. Fwd IAT Std: Standard deviation of time between forward packets
    * 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.
21. Fwd IAT Max: Maximum time between forward packets
    * 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
22. Fwd IAT Min: Minimum time between forward packets
    * Peak was observed around Fed 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.
23. Bwd IAT Total: Total time between backward packets
    * 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
24. Bwd IAT Mean: Mean time between backward packets
    * 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
25. Bwd IAT Std: Standard deviation of time between packets
    * 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.
26. Bwd IAT Max: Maximum time between packets
    * 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.
27. Bwd IAT Min: Minimum time between packets
    * 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
28. Fwd PSH Flags: Forward packets with PUSH flags
    * Most of the values are concenterated 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
29. Fwd Header Length: Length of header in forward packets
    * 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
30. Bwd Header Length: Length of header in backward packets
    * Peak was observed around Bwd Header Length=0.0
    * Most values are concenterated 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
31. Fwd Packets/s: Forward packets per second
    * 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 concenterated between Fwd Packets/s=0.0 and Fwd Packets/s=1.5. Between this range the graph also resembles to J-shaped graph.
32. Bwd Packets/s: Backward packets per second
    * The distribution is skewed towards right: Positively skewed.
    * Peak was observed around Bwd Packets/s=0.0
    * Most values are concenterated 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.
33. Packet Length Max: Maximum length of packets
    * The distribution is skewed towards right: Positively skewed.
    * Peak was observed around Packet Length Max=0
    * After the peak, there is signifcant 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 resembes 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 resuls observed for Packet Length Max>60000. This may indicate outlier in the data.
34. Packet Length Mean: Mean length of packets
    * 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 concenterated 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.
35. Packet Length Std: Standard deviation length of packets
    * 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 concenterated 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.
36. Packet Length Variance: Variance of length of packets
    * The distribution is skewed towards right: Positively skewed.
    * Peak was observed around Packet Length Variance=0
    * Most values are concenterated 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.
37. SYN Flag Count: Number of SYN flags
    * Peak was observed at SYN Flag Count=0.
    * Most values are concenterated at the peak.
    * There are a few observations at SYN Flag Count=1.0. This may indicate outlier in the data.
38. URG Flag Count: Number of URG flags
    * Peak was observed at URG Flag Count=0.
    * Most values are concenterated at the peak.
    * There are a few observations at URG Flag Count=1.0. This may indicate outlier in the data.
39. Avg Packet Size: Average packet size
    * 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 concenterated between Avg Packet Size>=0 and Avg Packet Size<5000.
    * There were some values afer a long gap between Avg Packet Size>5000 and Avg Packet Size <=17500. This may indicate outlier in the data.
40. Avg Fwd Segment Size: Average forward segment size
    * 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 concenterated between Avg Fwd Segment Size>=0 and Avg Fwd Segment Size<=5000.
    * There were some values around Avg Fwd Sgement 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.
41. Avg Bwd Segment Size: Average backward segment size
    * 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 concenterated 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.
42. Subflow Fwd Packets: Subflow forward packets
    * 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.
43. Subflow Fwd Bytes: Subflow forward bytes
    * The distibution 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 concenterated 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.
44. Subflow Bwd Packets: Subflow backward packets
    * 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 concenterated 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.
45. Subflow Bwd Bytes: Subflow backward bytes
    * 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 concenterated 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.
46. Init Fwd Win Bytes: Initial forward window size
    * 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.
47. Init Bwd Win Bytes: Initial backward window size
    * 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.
48. Fwd Act Data Packets: Forward packets with actual data
    * 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.
49. Fwd Seg Size Min: Minimum segment size in forward packets
    * 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
50. Active Mean: Mean active time
    * 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
51. Active Std: Standard deviation of active time
    * 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.
52. Active Max: Maximum active time
    * 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 plateay 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.
53. Active Min: Minimum active time
    * The distibution 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.
54. Idle Mean: Mean idle time
    * Peak was observed at Idle Mean=0.
    * Most values are concenterated 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.
55. Idle Std: Standard deviation of idle time
    * Peak was observed at Idle Std=0.0
    * Most values are concenterated 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.
56. Idle Max: Maximum idle time
    * Peak was observed at Idle Max=0.0
    * Most values are concenterated in bin represented by the peak.
    * There are some values observed at Idle Max=0.4, 0.6, 0.8, 1.0.
    * There are large gaps observed on X-axis of the graph after the peak.
57. Idle Min: Minimum idle time
    * 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.
58. 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.
59. Statistical summary for each feature was also computed to understand how the features are spread and their scale. From statistical summary, negative values for many features were observed. As per CIC dataset documentation, the negative values cannot exist.
60. 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%
61. Among the features where the negative values were observed, the proportion of negative values were computed.  
    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 constraint of time, 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 have increased.

1. A new feature was defined and created: isMalicious. This will work as target feature for binary classification. If ClassLabel=Benign, isMalicious=0 and if ClassLabel!=Benign, isMalicious=1.  
   Number of records for isMalicious=0 are 7185881 (78%).  
   Number of records for isMalicious=1 are 1981390 (22%).
2. Feature: Label was dropped from the dataset because it provides information about sub-type of malicious events which will not be in the scope of the work.
3. As the result, two target features at this stage are: -  
   isMalicious : Binary classification  
   ClassLabel : Multi-class classification
4. Since the original dataset was too large, carrying out analysis and making updates led to over utilization of system’s memory and Jupyter notebook stalls.  
   Thus, a sampled of the original dataset was created by taking 20% records.
5. In the sampled dataset: -  
   Number of records for isMalicious=0 are 1437467 (78%).  
   Number of records for isMalicious=1 are 395987 (22%).  
     
   All the category of attacks in ClassLabel have similar proportion as the original dataset.
6. The number of outliers and percentage of outliers for each feature were computed in the sample dataset.   
   Definition of outlier: - If a datapoint x is smaller than Q1 – 1.5\*IQR   
    OR  
    x is greater than Q3+1.5\*IQR  
   where IQR = Q3-Q1  
    Q1 = 25th percentile (First quartile)  
    Q3 = 75th percentile (Third quartile)
7. The outliers were grouped in two categories based on isMalicious.
8. Similarly, outliers were grouped into multiple categories based on ClassLabel.
9. Observations from the data collected for outliers per independent feature in the sampled dataset: -
   1. 12 features have outliers whose percentage of difference between Malicious and Benign events was greater than or equal to 10%
      1. Out of 12 features, 4 features have relatively higher outlier percentage: -
         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%
      2. Among the above 4 features, we have more number of records labelled as Benign than Malicious. Thus, the features with relatively higher number of outliers do not indicate any anomaly or provide any differentiation to detect Malicious events.
      3. Out of 12 features, remaining 8 features have less than or equal to 7% of outliers: -
         1. Fwd Packets Length Total : 1.33%
         2. Bwd Packet Length Max : 3.81%
         3. Bwd Packet Length Std : 3.07%
         4. Fwd Header Length : 6.97%
         5. Packet Length Max : 4.22%
         6. Packet Length Std : 3.73%
         7. Avg Fwd Segment Size : 4.05%
         8. Subflow Fwd Bytes : 3.91%
      4. All above 8 features have more 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.
   2. 45 features have nearly equal percentage of outliers labelled as Benign and Malicious.
10. To handle outliers, two methods were tested on the four features with higher percentage: -
    1. Winsorization
    2. Robust Scaling

The test will be performed on the below four features: -

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

1. 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.