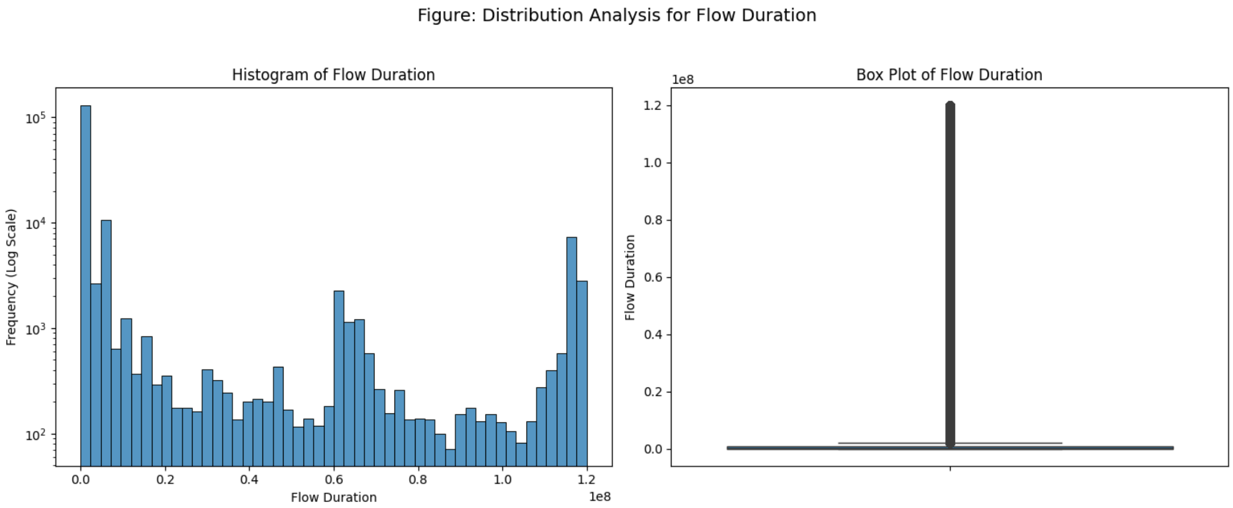
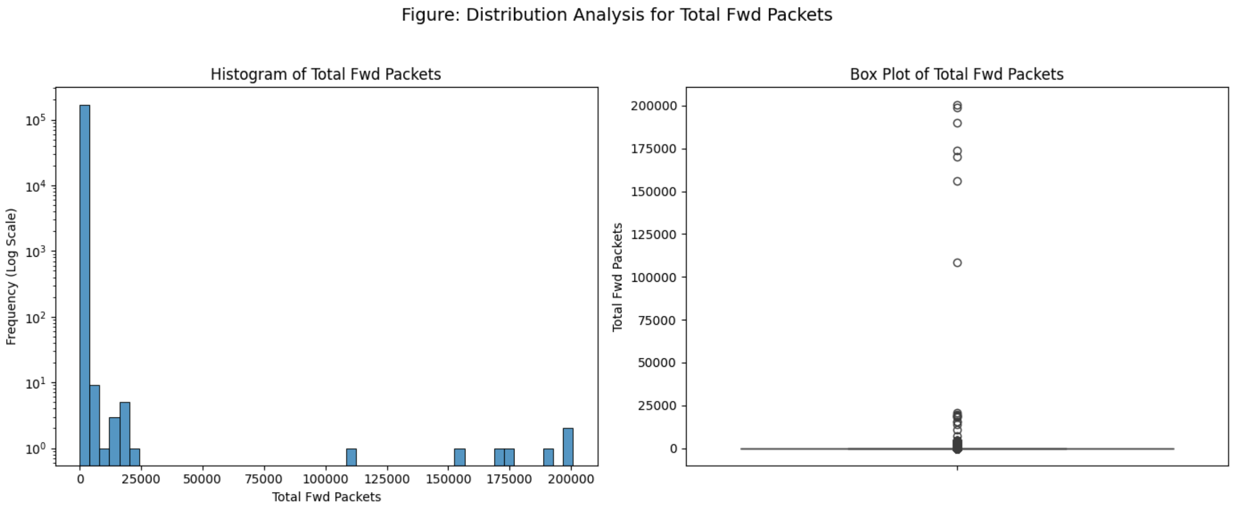
Figure 1: Distribution of Flow Duration



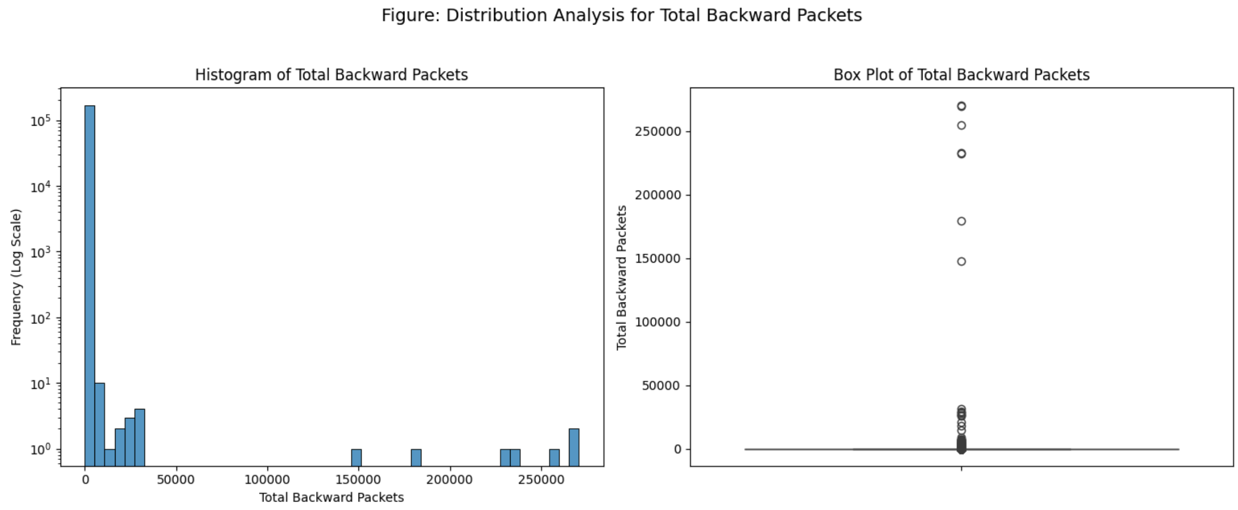
This figure shows the total duration of network flows. The log-scale histogram (left) indicates most flows are very short, while the box plot (right) highlights numerous outliers representing exceptionally long-duration flows, confirming extreme right-skewness.

Figure 2: Distribution of Total Forward Packets



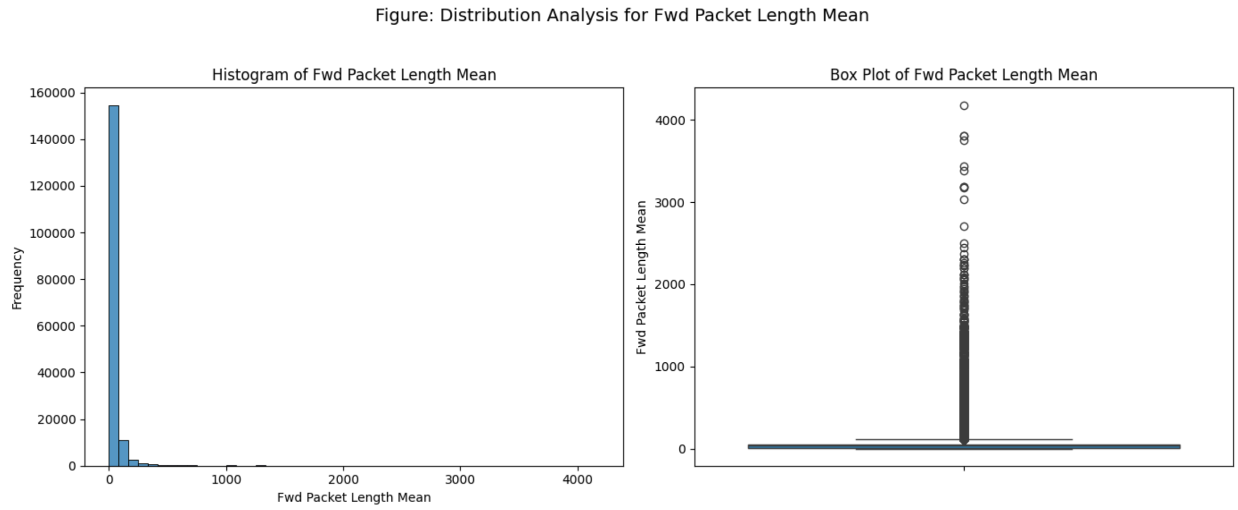
The log-scale histogram (left) and box plot (right) both show that most network flows contain very few forward packets (1-5). The data is highly right-skewed, with numerous outliers indicating flows with significantly higher packet counts.

Figure 3: Distribution of Total Backward Packets



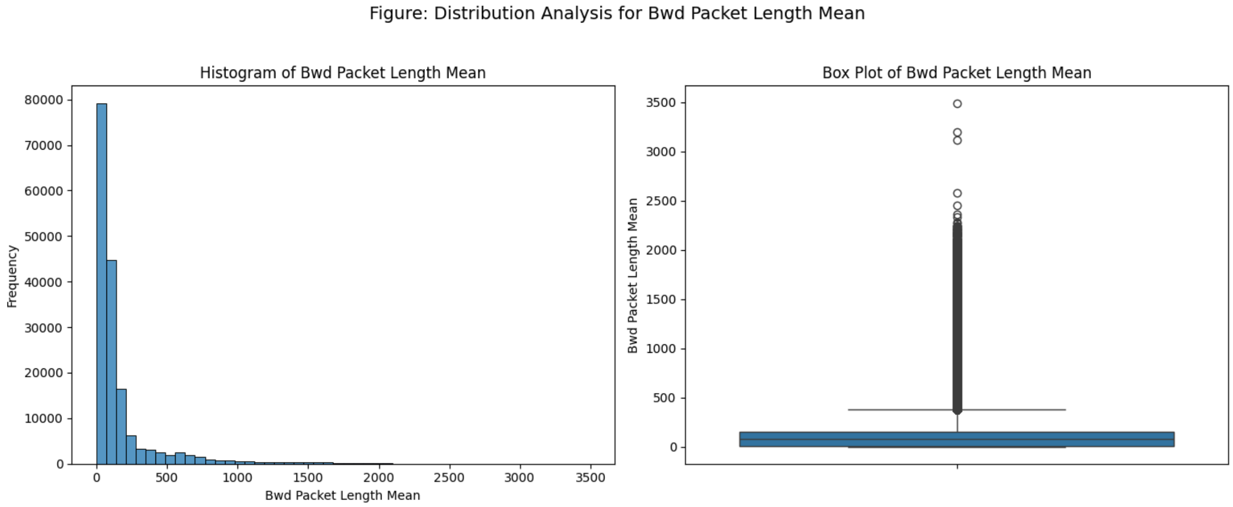
Similar to forward packets, the log-scale histogram (left) and box plot (right) indicate that most flows have very few backward packets (often 0-5). The distribution is extremely right-skewed, with many outliers representing flows containing a high volume of packets returning from the destination.

Figure 4: Distribution of Mean Forward Packet Length



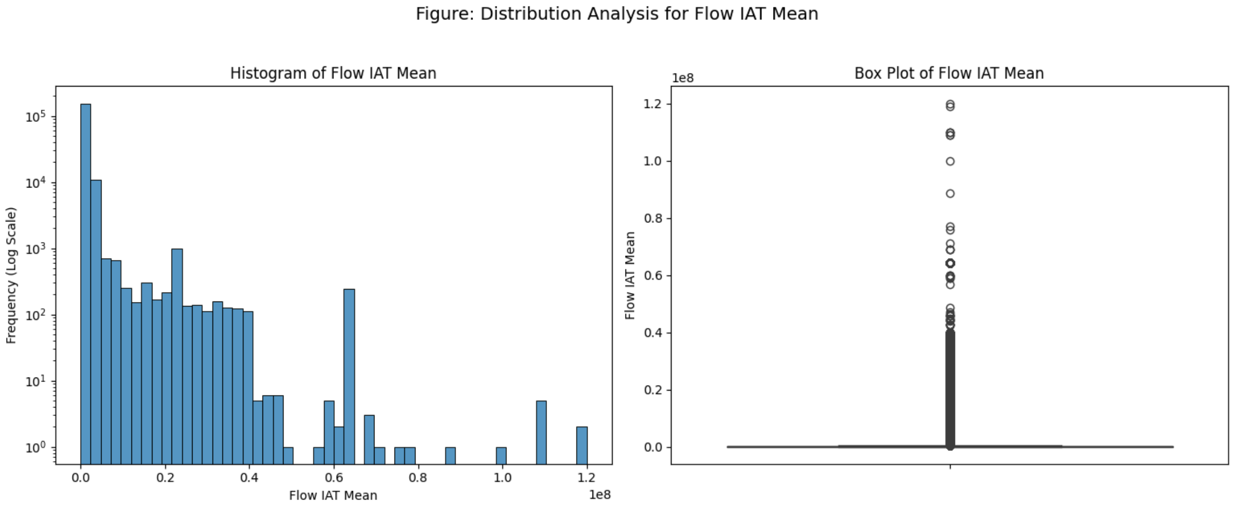
This shows the average size of packets sent from source to destination. The histogram (left) indicates several common average sizes, suggesting different types of traffic or protocols. The box plot (right) shows the central tendency and identifies flows with unusually high average forward packet sizes as outliers.

Figure 5: Distribution of Mean Backward Packet Length



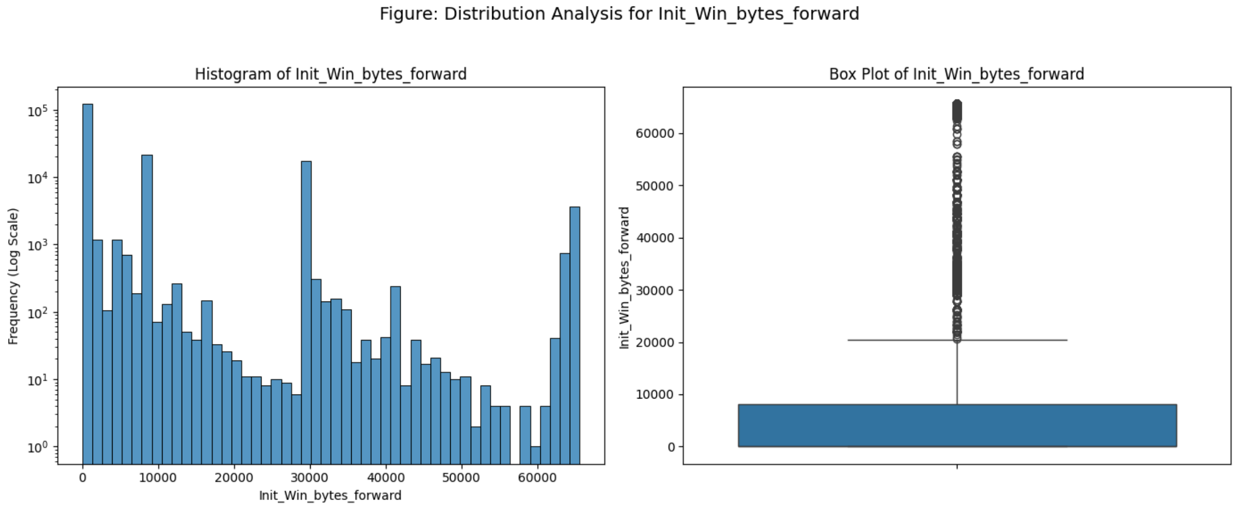
This displays the average size of packets sent from the destination back to the source. The histogram (left) shows a large peak at zero (likely flows with no backward packets) and other peaks representing common backward packet sizes. The box plot (right) indicates the central tendency is relatively low, but identifies flows with high average backward packet sizes as outliers.

Figure 6: Distribution of Mean Flow Inter-Arrival Time (IAT)



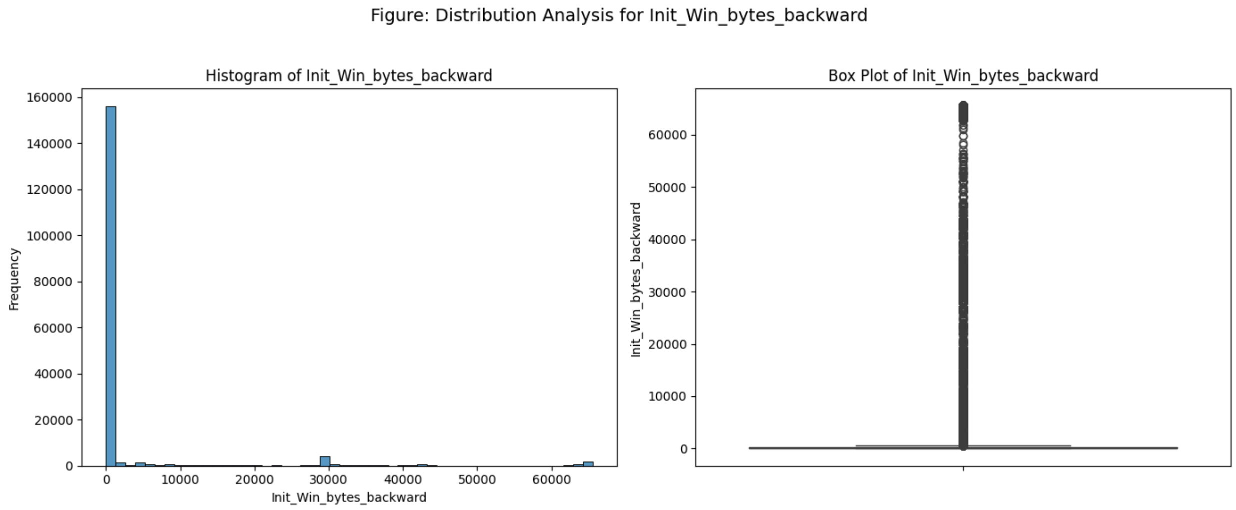
This shows the average time between packets within a flow. The log-scale histogram (left) and box plot (right) reveal that most flows have very short average times between packets. The distribution is extremely right-skewed, with numerous outliers representing flows with significantly longer average pauses between packets.

Figure 7: Distribution of Initial Forward Window Bytes



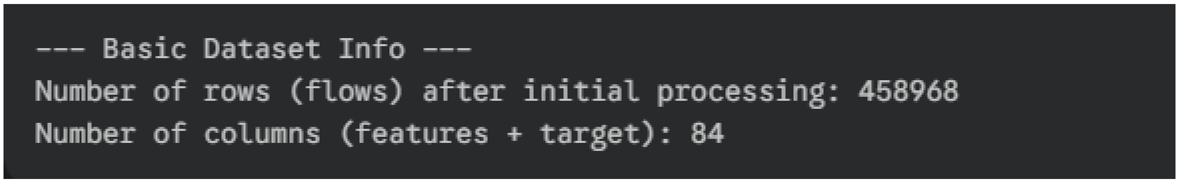
This shows the initial TCP window size advertised by the source. The histogram (left) indicates common values, potentially including -1 (for non-TCP or uncaptured flows) and standard window sizes like 8192 or 65535. The box plot (right) reflects these common values and shows the spread of other observed window sizes.

Figure 8: Distribution of Initial Backward Window Bytes



This displays the initial TCP window size advertised by the destination. The histogram (left) shows a large number of flows with a value of -1 (likely non-TCP or uncaptured) and another peak around a common window size (e.g., 256 or similar). The box plot (right) reflects these frequent values and the overall range observed.

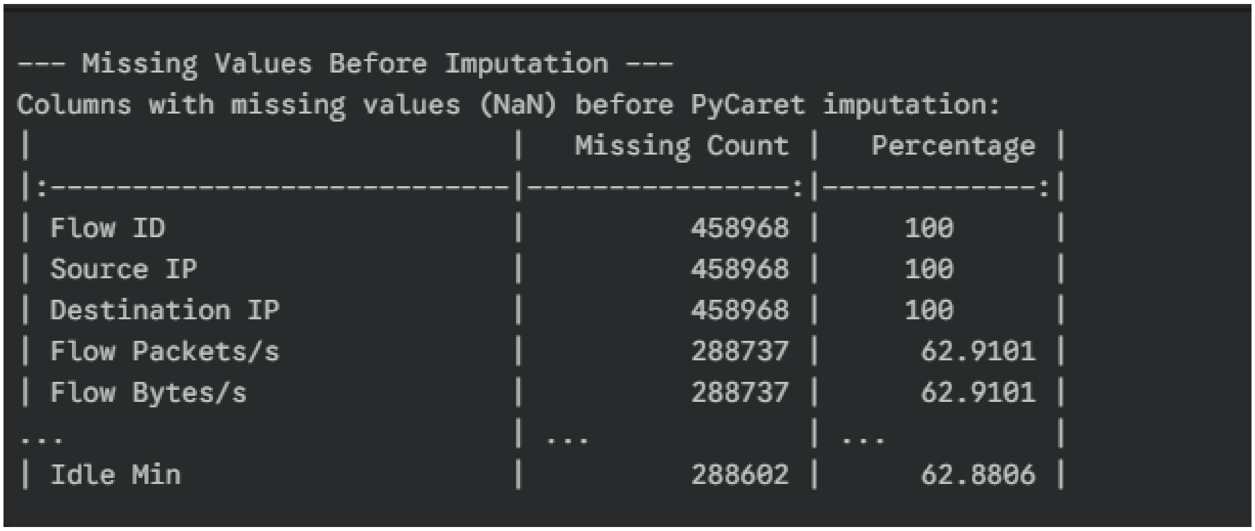
Table 1: Basic Dataset Dimensions



```
--- Basic Dataset Info ---
Number of rows (flows) after initial processing: 458968
Number of columns (features + target): 84
```

This table shows the initial dimensions of the `Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX.csv` dataset after loading, indicating 458,968 network flows (rows) and 84 columns including features and the original label before preprocessing.

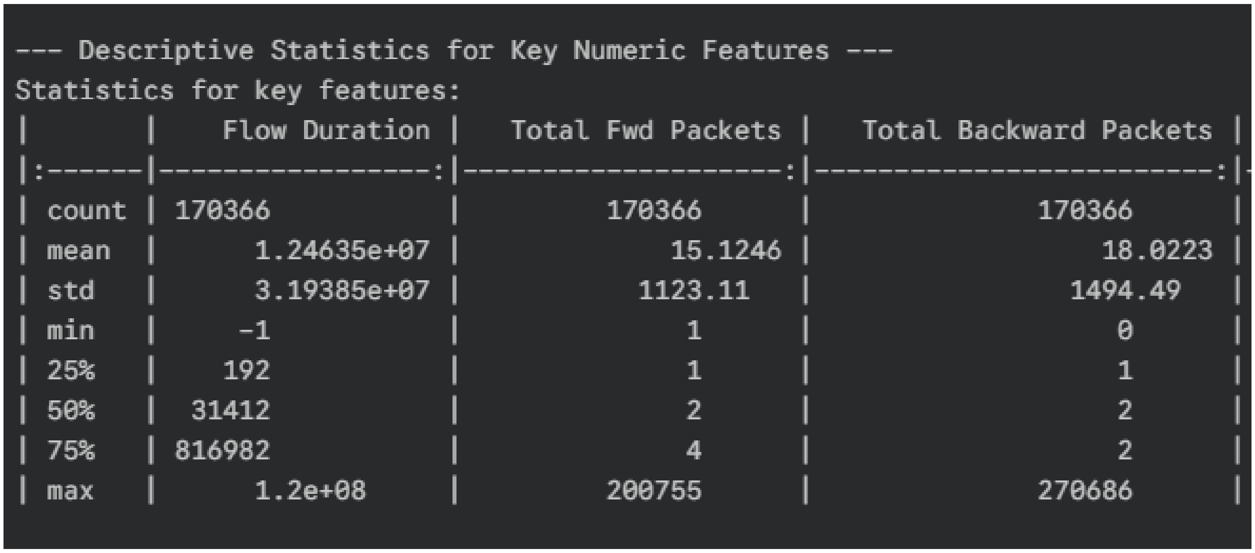
Table 2: Missing Value Summary Before Imputation



```
--- Missing Values Before Imputation ---
Columns with missing values (NaN) before PyCaret imputation:
|                              |   Missing Count |   Percentage |
|:-----------------------------|----------------:|-------------:|
| Flow ID                      |          458968 |     100      |
| Source IP                    |          458968 |     100      |
| Destination IP               |          458968 |     100      |
| Flow Packets/s               |          288737 |      62.9101 |
| Flow Bytes/s                 |          288737 |      62.9101 |
...                            | ...             | ...          |
| Idle Min                     |          288602 |      62.8806 |
```

This table lists columns containing missing (`NaN`) or infinite values after initial numeric conversion. Columns like `Flow ID`, `Source IP`, and `Destination IP` show 100% missing values, likely due to conversion errors from non-numeric identifiers originally present. Rate-based features (`Flow Packets/s`, `Flow Bytes/s`) and many others show significant missing data (~63%), often caused by division-by-zero errors (e.g., zero flow duration) or failed conversions. These missing values require imputation, which PyCaret handles during the setup phase.

Table 3: Descriptive Statistics for Key Numeric Features



```
--- Descriptive Statistics for Key Numeric Features ---
Statistics for key features:
|       |    Flow Duration |   Total Fwd Packets |   Total Backward Packets |
|:------|-----------------:|--------------------:|-------------------------:|-
| count | 170366           |         170366      |              170366      |
| mean  |      1.24635e+07 |             15.1246 |                  18.0223 |
| std   |      3.19385e+07 |           1123.11   |                1494.49   |
| min   |     -1           |              1      |                   0      |
| 25%   |    192           |              1      |                   1      |
| 50%   |  31412           |              2      |                   2      |
| 75%   | 816982           |              4      |                   2      |
| max   |      1.2e+08     |         200755      |              270686      |
```

This table summarizes key numerical network features after initial loading and conversion but before imputation. The vast differences between the mean/median (50%) and maximum values for `Flow Duration`, `Total Fwd Packets`, `Total Backward Packets`, and `Flow IAT Mean` confirm extreme right-skewness and the presence of significant outliers. The negative minimum values observed in several features (e.g., `Flow Duration`, `Flow IAT Mean`, `Init_Win_bytes_forward`, `Init_Win_bytes_backward`) might indicate data artifacts or specific flags from the data generation tool (CICFlowMeter) and were included prior to imputation.