**Data Dictionary**

This dataset has 1000 rows and 9 columns (8 features + 1 target). All data is synthetic but designed to mimic real-world variability, including intentional outliers (~5% of rows inflated by 5-10x) for your demo on model robustness.

| **Column Name** | **Data Type** | **Description** | **Possible Values/Range** |
| --- | --- | --- | --- |
| duration\_sec | float | Duration of the network flow in seconds. Longer durations typically involve more data transfer. | 1 to 20 seconds |
| packet\_count | integer | Number of packets exchanged in the flow. More packets often mean higher data volume. | 50 to 150 packets |
| latency\_ms | float | Latency (delay) in milliseconds. Higher latency might correlate with retries and more bytes sent. | 10 to 100 ms |
| protocol | string | Network protocol used. TCP is reliable (more overhead/bytes), UDP is faster (fewer bytes), ICMP is diagnostic (low bytes). | 'TCP', 'UDP', 'ICMP' |
| device\_type | string | Type of device initiating the flow. Servers handle high-load traffic (more bytes), IoT devices are lightweight (fewer bytes). | 'IoT', 'Server', 'Router', 'Workstation' |
| location | string | Location of the device. Data centers have centralized, high-volume traffic; remote sites vary. | 'Data Center', 'Remote' |
| event\_type | string | Associated security event (if any). Malware events might spike bytes (e.g., data exfiltration); no event means normal traffic. | 'No Event', 'Login Fail', 'Malware Detect', 'Firewall Block' |
| severity | string | Severity of any security event. High severity often correlates with anomalous high-byte flows (e.g., attacks). | 'None', 'Low', 'Medium', 'High' |
| bytes\_sent | float | **Target variable (y)**: Total bytes sent in the flow (outgoing traffic volume). This is what we predict. | ~1000 to ~80,000+ (with outliers up to 500,000+) |