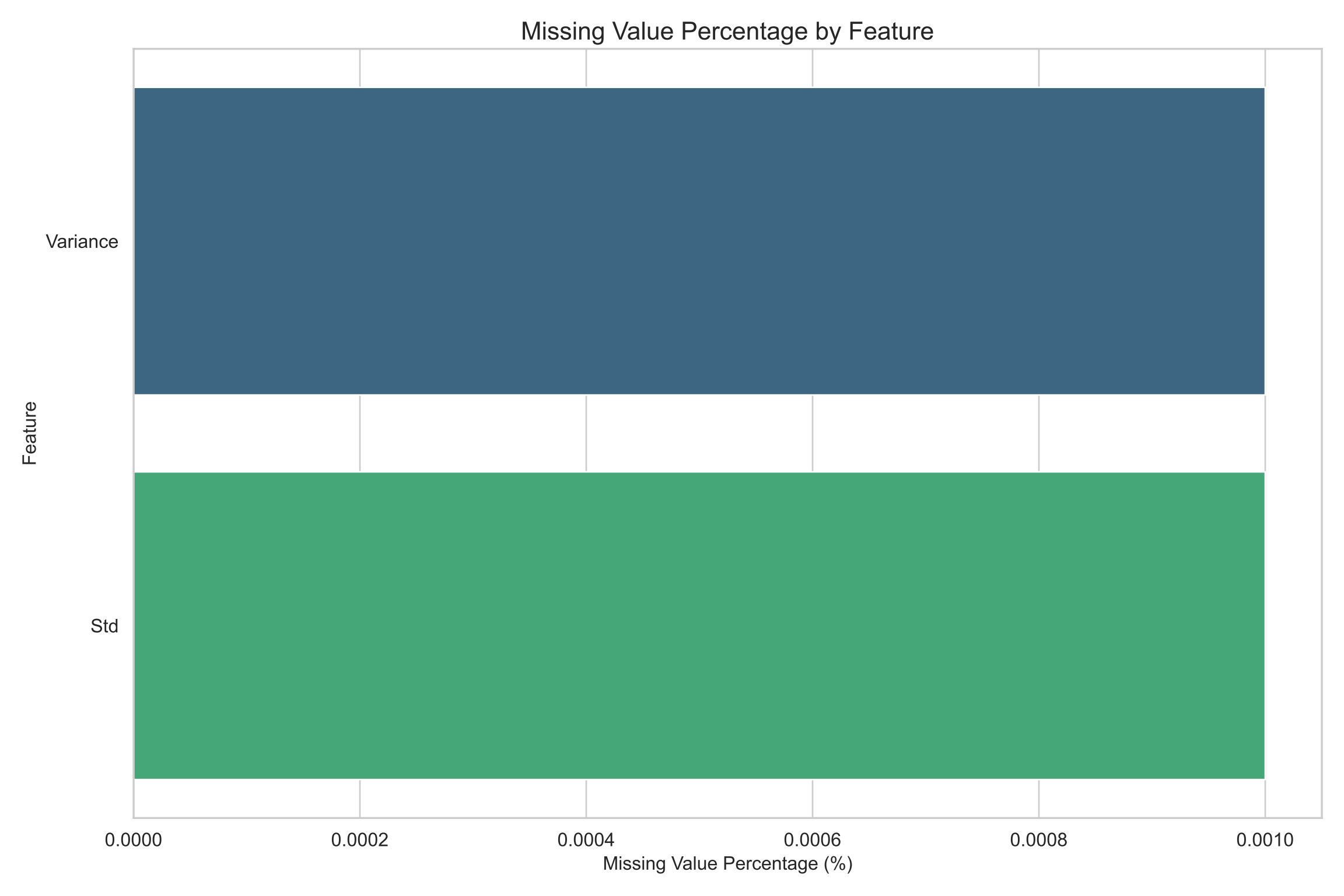
IoT Security Threat Detection for SMEs:

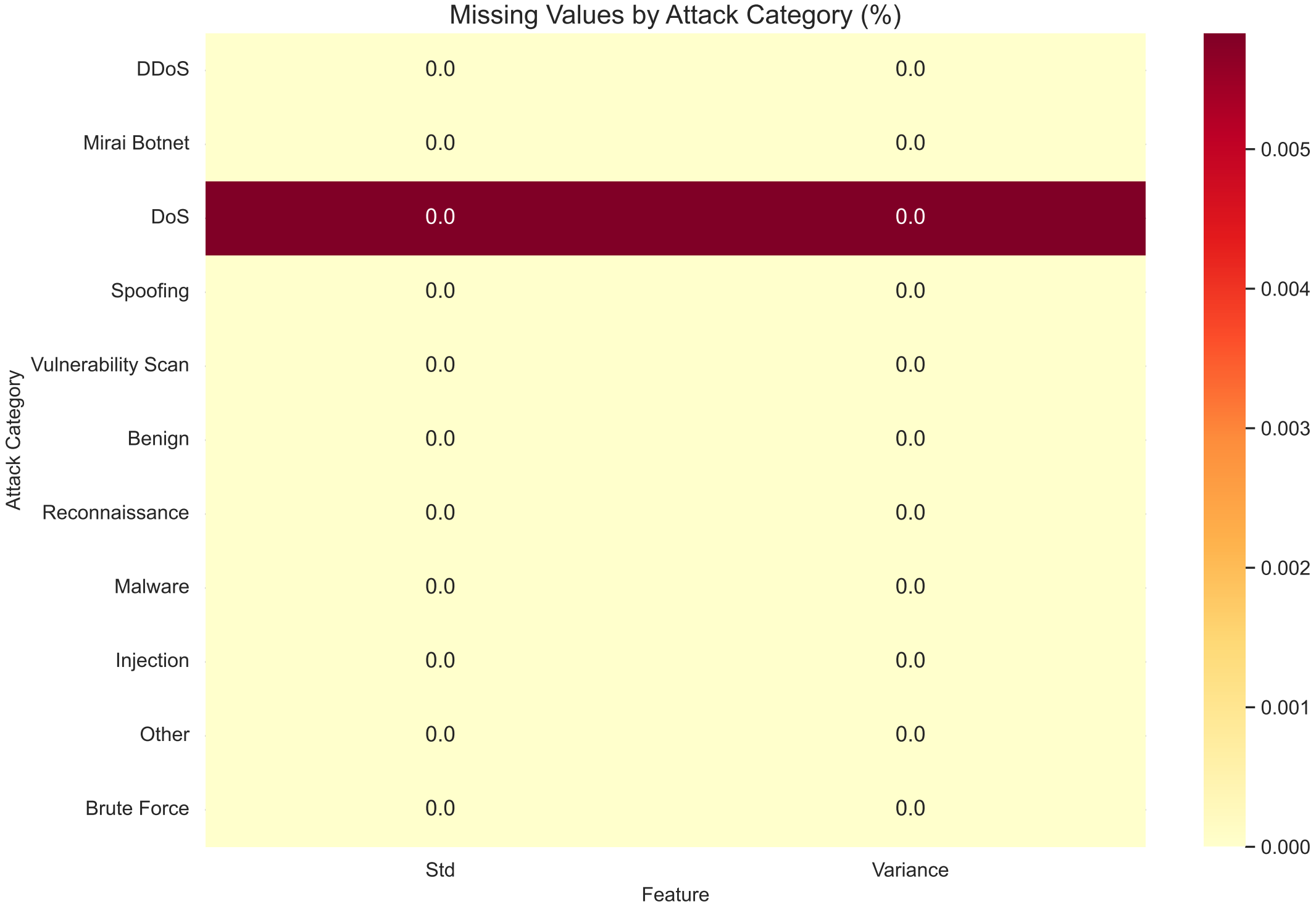
A Machine Learning Approach Using CIC-IoT Dataset

STAGE 2, STEP 1: QUALITY ASSESSMENT

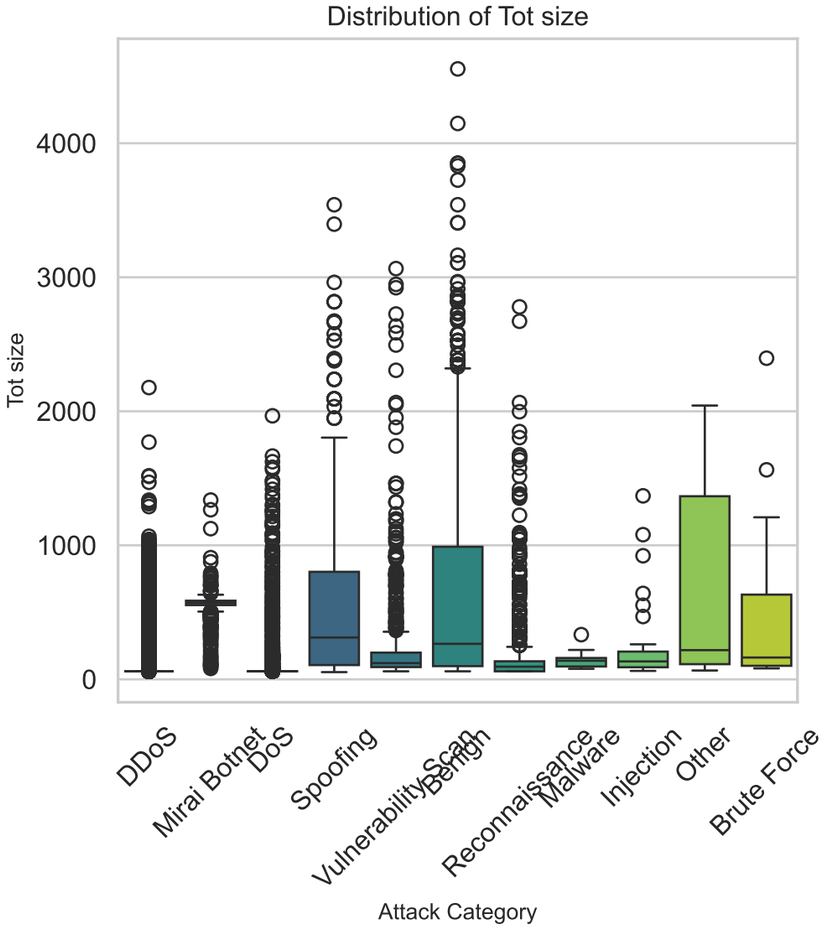
This report provides a comprehensive quality assessment of the CIC-IoT dataset, focusing on missing values, data consistency, feature completeness, timestamp validation, and class balance issues to ensure robust machine learning model development.

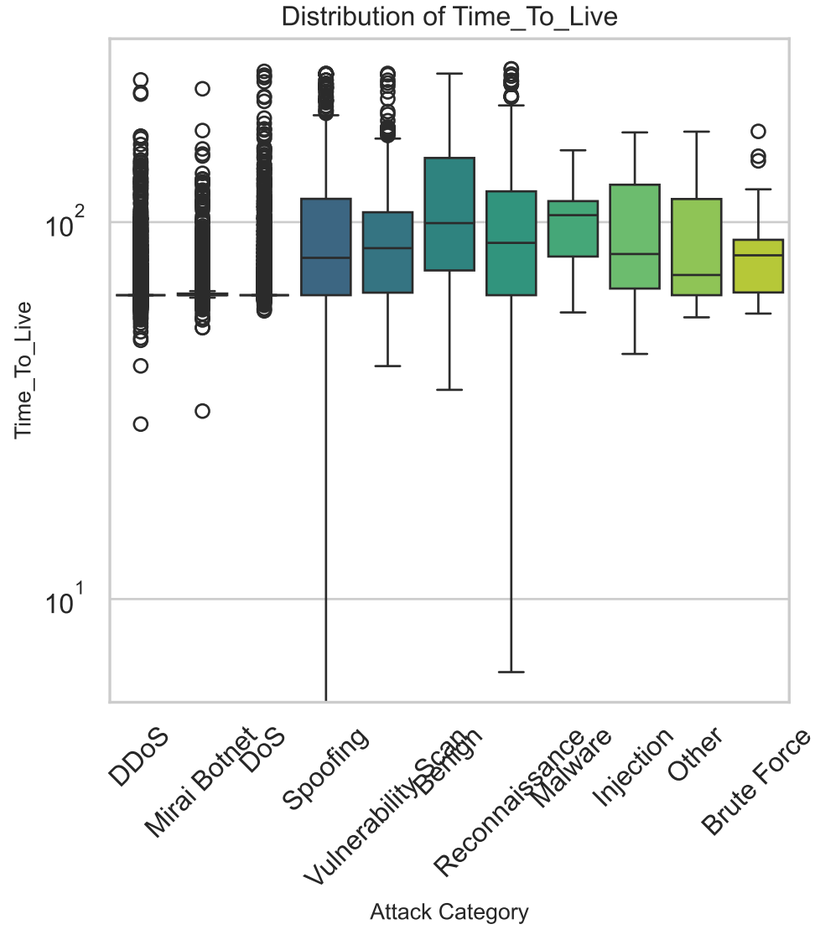

Missing Value Percentage by Feature
Variance
Std
Feature
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
Missing Value Percentage (%)

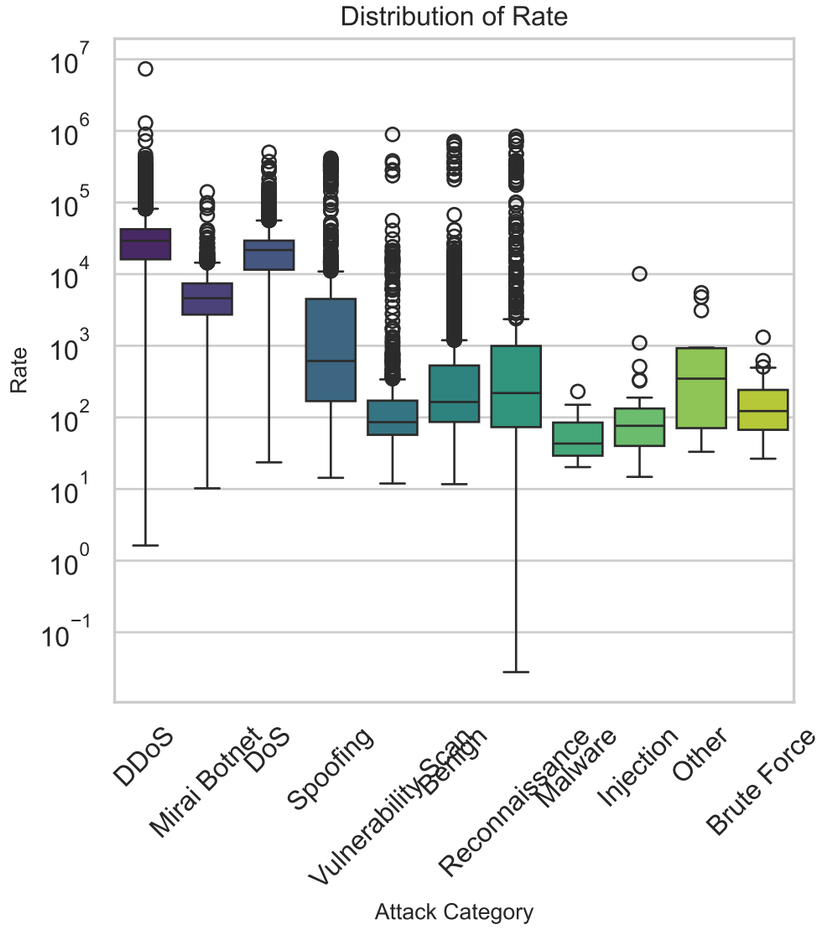
This visualization shows the percentage of missing values for each feature in the CIC-IoT dataset. Features are sorted by their missing value percentage in descending order, allowing us to quickly identify which features have data quality issues. Missing values can significantly impact machine learning model performance, especially for security-critical applications like IoT threat detection. Features with high missing value percentages may need imputation strategies or could potentially be dropped if they aren't critical. For SMEs with limited data science expertise, understanding these data quality issues is essential for building reliable threat detection systems. Features with close to 100% missing values might be protocol-specific and naturally absent in certain traffic types rather than representing data collection errors.

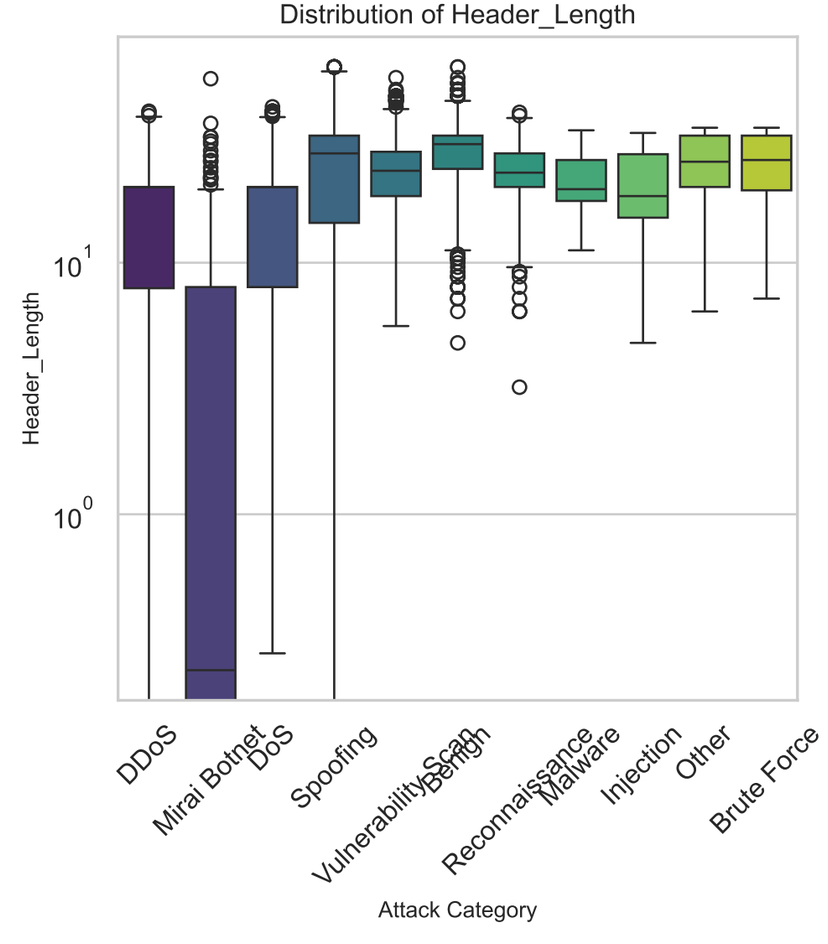

Missing Values by Attack Category (%)
DDoS
Mirai Botnet
DoS
Spoofing
Vulnerability Scan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
0.0
Attack Category
Std
Variance
Feature
0.005
0.004
0.003
0.002
0.001
0.000

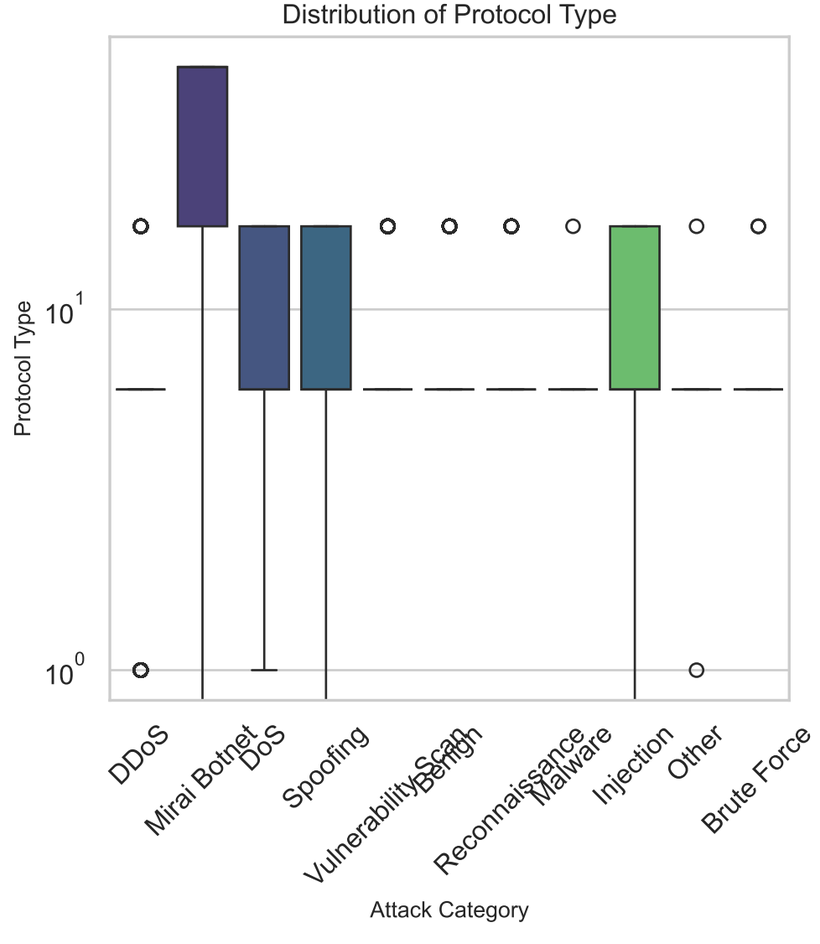
This heatmap reveals the percentage of missing values for key features across different attack categories.
The color intensity represents the missing value percentage, with darker colors indicating higher percentages.
This visualization is particularly valuable for understanding if missing data patterns are correlated with specific attack types, which could introduce bias in threat detection models. For example, if certain attack
categories consistently have missing values for specific features, it could indicate either a characteristic of that attack (such as protocol-specific fields being legitimately absent) or data collection issues during attack simulation. For SMEs implementing security solutions, this analysis helps understand potential blind
spots in detection capabilities related to specific attack types and features.

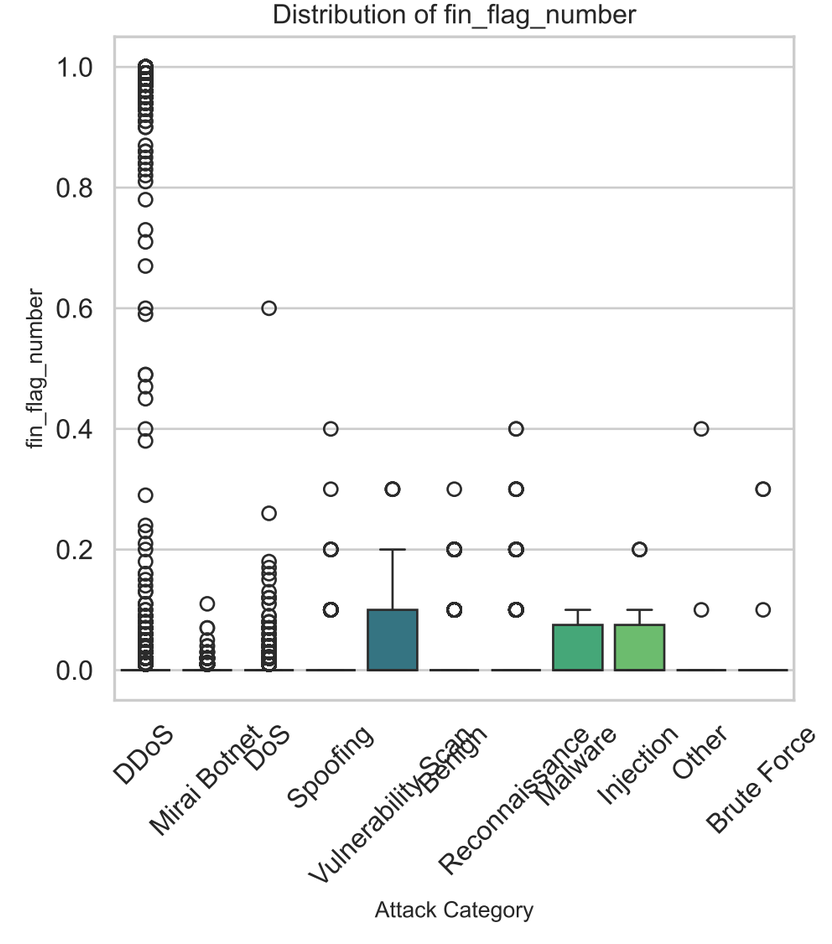

Distribution of Tot size
Tot size
0
1000
2000
3000
4000
DDoS
Mirai Botnet
DoS
Spoofing
VulnerabilityScan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
Attack Category


Distribution of Time_To_Live
Time_To_Live
10^2
10^1
DDoS
Mirai Botnet
DoS
Spoofing
VulnerabilityScan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
Attack Category

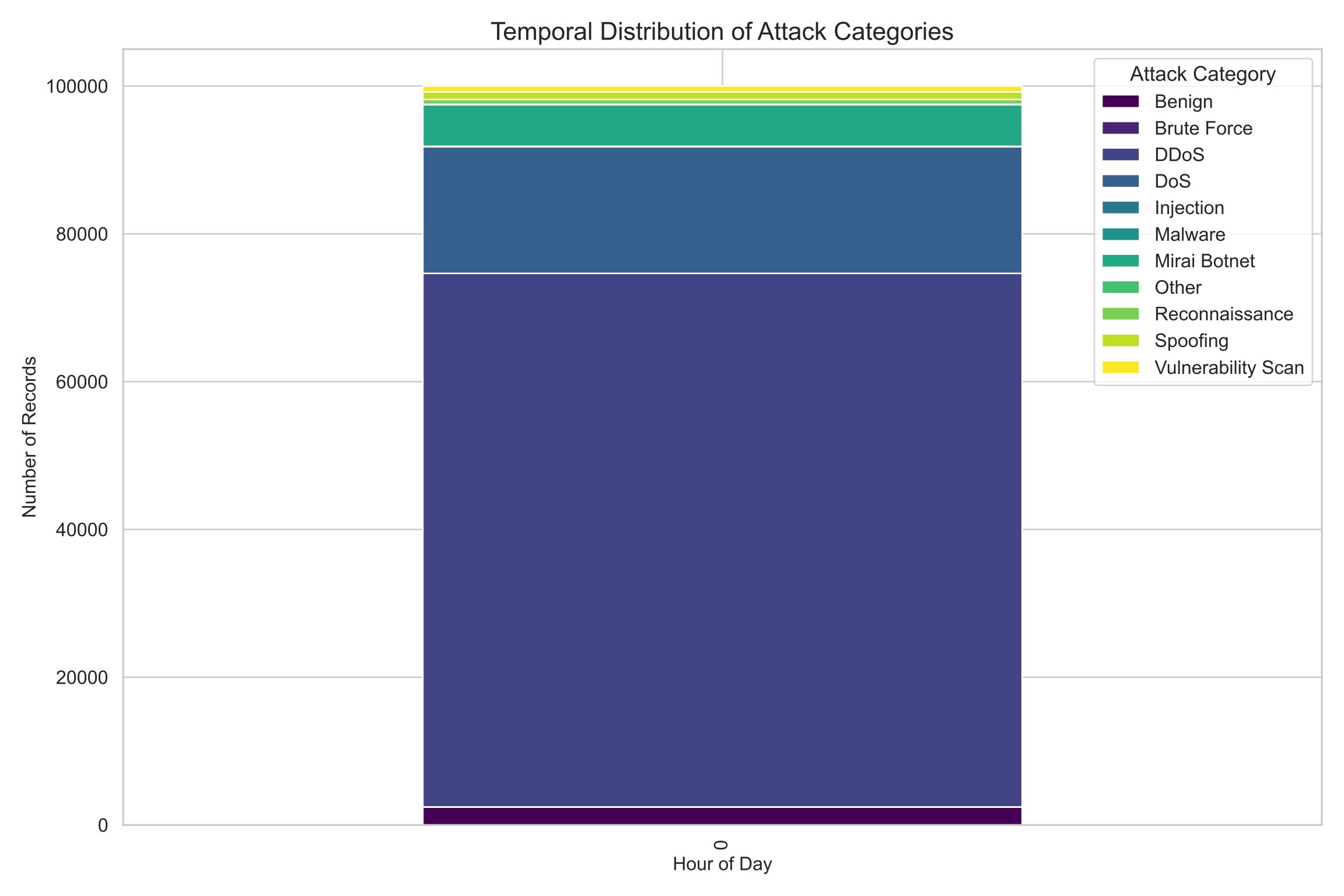

Distribution of Rate
Rate
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
DDoS
Mirai Botnet
DoS
Spoofing
VulnerabilityScan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
Attack Category


Distribution of Header_Length
Header_Length
10^1
10^0
DDoS
Mirai Botnet
DoS
Spoofing
VulnerabilityScan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
Attack Category

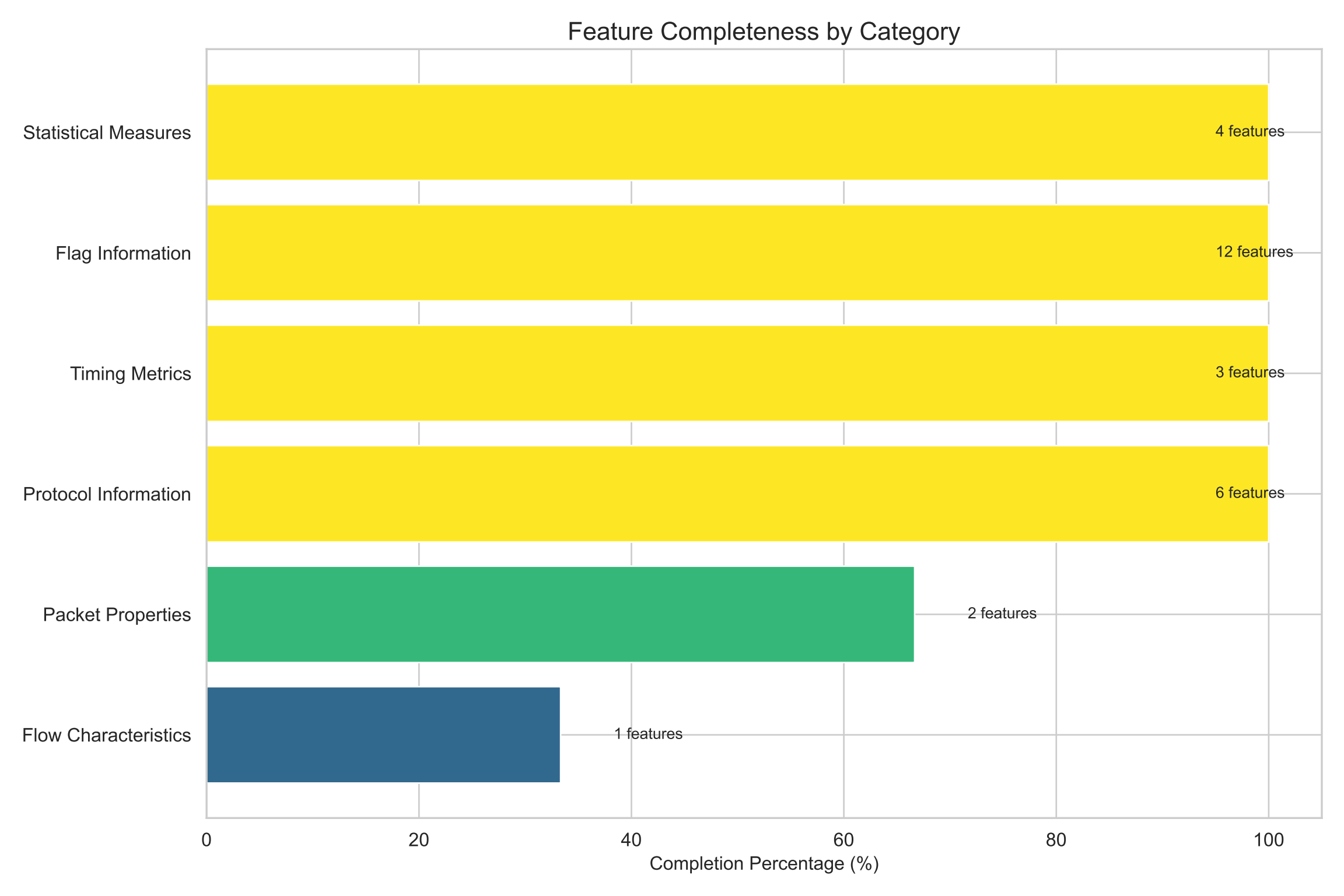

Distribution of Protocol Type
Protocol Type
10^1
10^0
DDoS
Mirai Botnet
DoS
Spoofing
VulnerabilityScan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
Attack Category


Distribution of fin_flag_number
fin_flag_number
1.0
0.8
0.6
0.4
0.2
0.0
DDoS
Mirai Botnet
DoS
Spoofing
VulnerabilityScan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
Attack Category

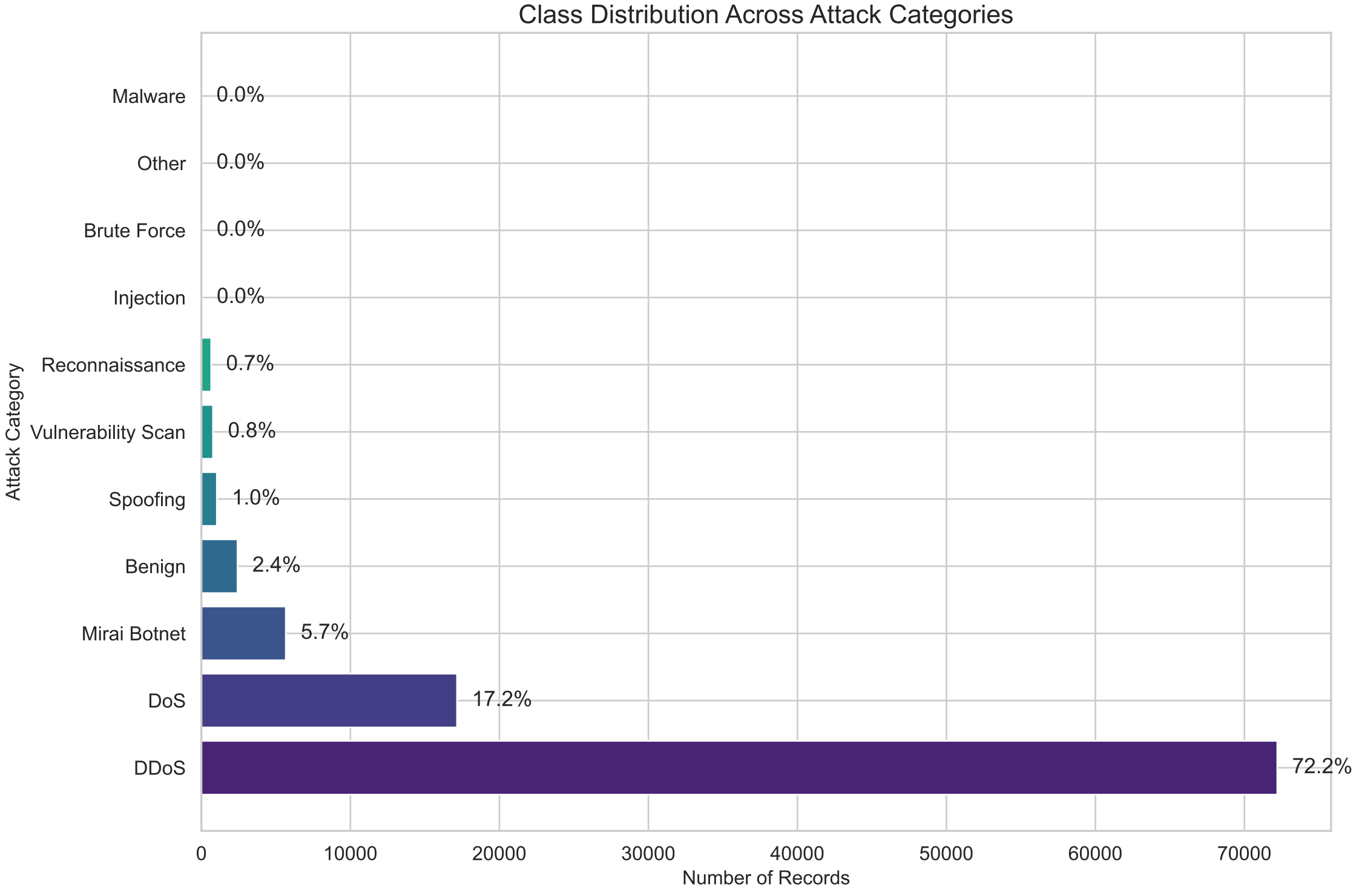
These box plots illustrate the distribution of key numerical features across different attack categories, providing insights into data consistency and potential outliers. The box represents the interquartile range (IQR) with the median shown as a horizontal line, while whiskers extend to 1.5 times the IQR. Points beyond
the whiskers are potential outliers. Significant variations in feature distributions between attack categories are expected and reflect the distinctive traffic patterns of different attack types. However, extreme outliers may indicate data quality issues or particularly severe attack instances. For SMEs building threat detection systems, understanding these distributions helps establish appropriate thresholds and detection rules, while
also identifying potential data quality issues that might affect model performance.


Temporal Distribution of Attack Categories
Number of Records
100000
80000
60000
40000
20000
0
0
Hour of Day
Attack Category
Benign
Brute Force
DDoS
DoS
Injection
Malware
Mirai Botnet
Other
Reconnaissance
Spoofing
Vulnerability Scan

This visualization shows the temporal distribution of different attack categories across hours of the day,
helping validate the timestamp accuracy in the dataset. The stacked bars represent the number of records
for each attack category during each hour. Natural variations in this distribution can indicate realistic traffic patterns, while abrupt changes or uniform distributions might suggest simulated data or timestamp
inaccuracies. For SMEs developing IoT security solutions, understanding the temporal patterns of attacks
is crucial for establishing baseline traffic models and detecting anomalies. This analysis also helps validate whether the dataset represents realistic attack scenarios, including time-of-day patterns that might be exploited by attackers targeting businesses during specific operational hours.


Feature Completeness by Category
Statistical Measures
Flag Information
Timing Metrics
Protocol Information
Packet Properties
Flow Characteristics
4 features
12 features
3 features
6 features
2 features
1 features
0
20
40
60
80
100
Completion Percentage (%)

This visualization assesses the completeness of the CIC-IoT dataset by categorizing features into key areas relevant for IoT security threat detection. Each bar represents a feature category, with the length indicating the completion percentage based on having adequate representative features in that category. The number of features found for each category is displayed next to the bars. A complete dataset for IoT security analysis should have good coverage across all these categories to effectively detect various attack types. For SMEs implementing security monitoring, this analysis helps identify potential blind spots in the dataset that might affect detection capabilities. It also informs feature engineering efforts, highlighting categories where derived features might need to be created to complement existing data.


Class Distribution Across Attack Categories
Malware
Other
Brute Force
Injection
Reconnaissance
Vulnerability Scan
Spoofing
Benign
Mirai Botnet
DoS
DDoS
0.0%
0.0%
0.0%
0.0%
0.7%
0.8%
1.0%
2.4%
5.7%
17.2%
72.2%
Attack Category
0
10000
20000
30000
40000
50000
60000
70000
Number of Records

This horizontal bar chart illustrates the class distribution across attack categories in the CIC-IoT dataset, with the count of records for each category and the percentage of the total dataset shown next to each bar.

The significant imbalance between classes is immediately apparent, with some attack categories having orders of magnitude more samples than others. This imbalance is a critical consideration for machine learning model development, as models tend to bias toward majority classes without proper mitigation strategies. For SMEs implementing security solutions, understanding this imbalance helps in selecting appropriate techniques such as class weighting, oversampling minority classes, or using anomaly detection approaches. It also highlights which attack types might have less reliable detection due to limited training examples.

Class Balance Issues and Mitigation Strategies

Majority Class Dominance

Description: DDoS and DoS attacks dominate the dataset

Impact: Models may bias toward detecting only the most common attacks

Mitigation: Class weighting, stratified sampling, or cost-sensitive learning

Rare Attack Types

Description: Some attack categories have very few samples

Impact: Poor detection capabilities for uncommon but potentially critical attacks

Mitigation: Oversampling techniques like SMOTE or targeted data augmentation

Benign-to-Attack Ratio

Description: Imbalance between benign and attack traffic

Impact: May not reflect realistic traffic proportions in SME environments

Mitigation: Adjust class weights to reflect expected real-world distributions

Category Granularity

Description: Some categories contain diverse attack subcategories

Impact: Potential masking of important subcategory patterns

Mitigation: Hierarchical classification or separate models for attack subtypes

This summary outlines the key class balance issues identified in the CIC-IoT dataset and provides mitigation strategies relevant for SME environments. Class imbalance is a significant challenge in security datasets, reflecting the real-world rarity of certain attack types compared to common ones. However, without proper handling, this imbalance can lead to security blind spots where less common but potentially severe attacks go undetected. For SMEs with limited cybersecurity resources, addressing these class balance issues is essential for creating effective IoT security monitoring systems. The suggested mitigation strategies aim to improve detection capabilities across all attack types while optimizing for the resource constraints typical in SME environments.