Unauthorized Network Access Analysis

Rhoda Diana Ndibalekera, Benard Wakyaya
Makerere University, College of Computing and Information Sciences
Uganda

ABSTRACT

Unauthorized access remains a critical cybersecurity challenge that can result in data breaches, network disruptions, and financial loss. This project utilizes the CIC-IDS 2018 dataset to detect and visualize unauthorized access patterns through Exploratory Data Analysis (EDA) and visualization techniques. We highlight key trends such as protocol usage, attack type distributions, flow durations, and packet characteristics. Our analysis provides valuable insights to guide intrusion detection strategies and network security improvements.

KEYWORDS

Unauthorized access, Network traffic, CIC-IDS 2018, Data visualization, Cybersecurity, Intrusion detection

1 INTRODUCTION AND MOTIVATION

Unauthorized access attempts in network environments are often precursors to serious cybersecurity incidents. Understanding these attempts through analysis of network traffic data can reveal patterns that assist in improving monitoring and response mechanisms. The CIC-IDS 2018 dataset provides a reliable benchmark for studying real-world attack scenarios. This project focuses on identifying attack trends, visualizing traffic characteristics, and informing defensive strategies using data analytics.

2 PROBLEM DEFINITION

The objective of our analysis is to detect and understand patterns in network traffic indicative of unauthorized access. We frame this as an unsupervised exploratory problem using labeled data for insight generation. While we do not build predictive models, our study lays a foundation for future supervised learning.

3 RELATED WORK

Previous studies using the CIC-IDS datasets have employed machine learning and visualization for intrusion detection. For example, Ring et al. (2019) compared ML classifiers on CICIDS2017, emphasizing the dataset's real-world value. Other works such as citesharafaldin2018toward explore time-based and flow-based features for detecting DDoS and infiltration attacks. Our work complements these by focusing on visualization-based insights using Tableau and Python.

Our analysis complements this body of work by focusing on descriptive analytics and visualization of network flows, providing an intuitive understanding of the patterns underlying various attack types.

4 DATASET OVERVIEW

The CIC-IDS 2018 dataset simulates real-world network traffic with both benign and malicious behaviors. It includes labeled attacks such as DDoS, PortScan, Bot, Infiltration, and others. The dataset contains 1,252,846 records and 79 features.

5 METHODOLOGY

Our workflow includes:

(1) **Dataset Acquisition**: CIC-IDS 2018 obtained from Kaggle, with over 1.2 million labeled network flow records.
(2) **Data Cleaning and Wrangling**: Used Google Colab and Pandas to preprocess data. Confirmed no missing values.
(3) **Feature Selection**: Selected relevant features including flow duration, protocol, destination port, packet size, and flags.
(4) **Visualization**: Used Google Colab to create bar charts, pie charts, box plots, Sankey diagrams, and heatmaps.

Tools: Google Colab (Python), Pandas, Matplotlib.

6 DESCRIPTIVE STATISTICS

- Total Records: 1,252,846
- Total Features: 79
- Label Distribution
  - Benign Traffic: ~77.5%
  - Malicious/Other Traffic: ~22.5%

This shows that most traffic is benign, but there's a notable proportion of potentially harmful traffic that warrants further analysis

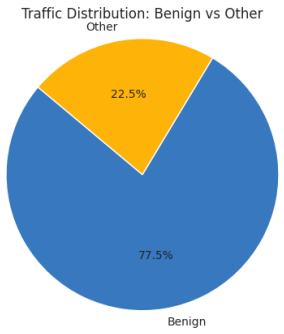

Figure 1: Traffic Distribution

7 DISTRIBUTION OF ATTACK TYPES

Here, we dive into the types of attacks observed to understand threat diversity. In other words, which threats are most prevalent? this guides where to focus mitigation efforts. From the visual above, it is observed that the Distributed Denial of Service(DDoS) attack traffic is the most prevalent among the recorded malicious attacks. Distributed Denial of Service (DDoS) attacks are the most prevalent among malicious traffic. These are often easier and cheaper to execute compared to other attack types.

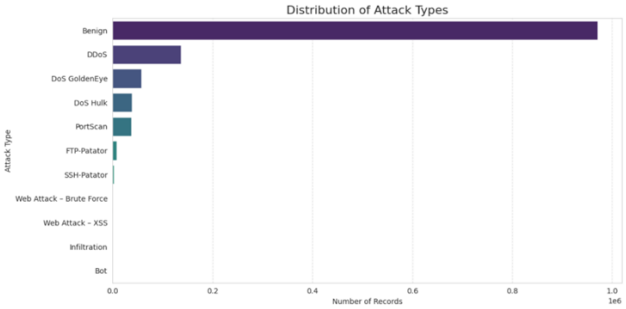

Figure 2: Most Common Destination Ports by Protocol

8 PORTS AND PROTOCOLS USAGE

We visualized this relationship using a bar chart showing top ports per protocol, clearly illustrating dominant services (e.g., HTTP/S, DNS).

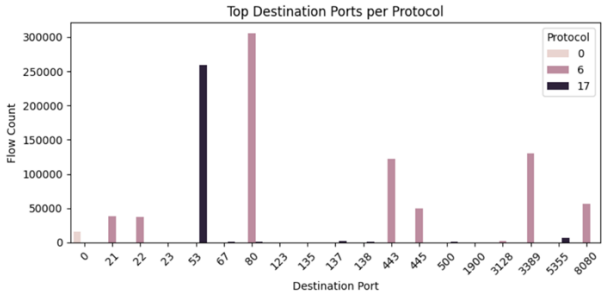

Figure 3: Most Common Destination Ports by Protocol

Most Common Destination Ports by Protocol:

- **TCP:** 80, 443, 22, 21, 8080
- **UDP:** 53, 123, 161
- **ICMP:** N/A (uses type/code)

9 FLOW DURATION DISTRIBUTION

In the Figure 4, we explore the duration of attack flows to assess severity and stealth. These visualizations help identify persistent threats like potential Denial of Service(DoS) attacks or other stealth behavior. Visualizations include:

- Log-transformed Flow Duration Distribution
- Flow Duration Categories

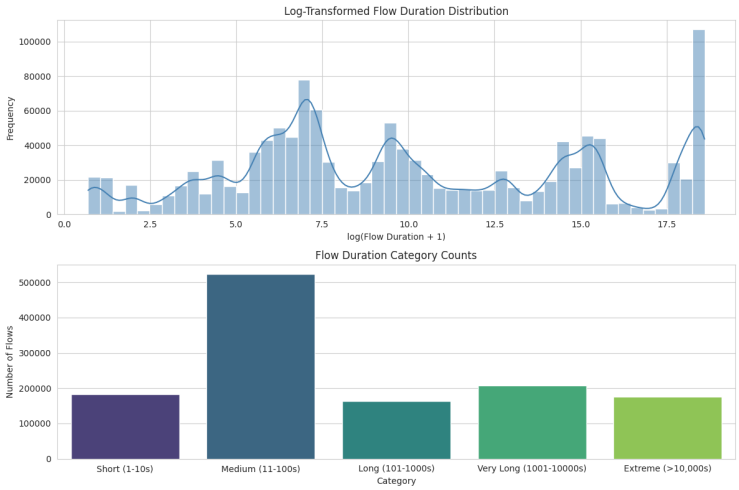

Figure 4: Flow Duration Distribution

Based on the results in figure 4, Most traffic was tiny. Almost every flow lasts only a short time, carries only a few packets, and moves very little data. These could be everyday, harmless browsing and emailing activities.

Few flows were significantly huge. The few spikes could indicate traffic from activities like big file dumps, denial-of-service blasts, or sneaky long-running intrusions. These are rare occurrences, but important to spot.

10 FLOW DURATIONS ACROSS ATTACK TYPES

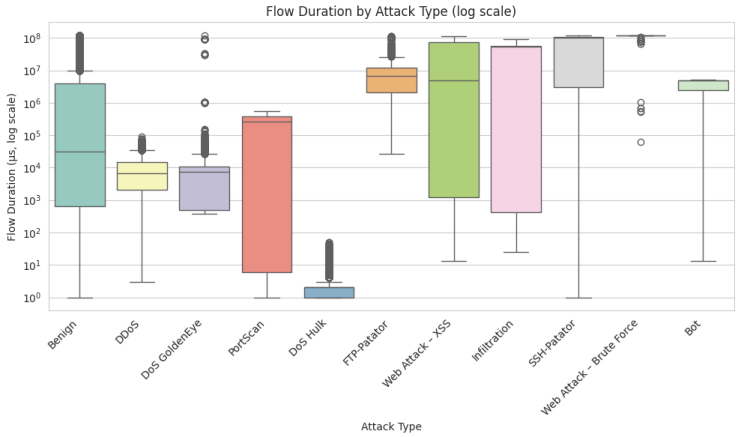

Figure 5: Flow Durations Across Attack Types

The objective of this visualization, was to understand the progression of different types of attacks in order to inform efforts to improve the security posture of the network.

By plotting the flow duration for each attach type, we can see that DDoS and DDoS attacks take a relatively short duration for each attack, compared to the XSS, infiltration and portscan.

XSS attacks were observed to take longer than DoS attacks because they often rely on indirect methods and require the victim to interact with the attack. This is the case with DoS and DDoS which directly overwhelm the system.

11 CORRELATION ANALYSIS

In order to analyze the correlations with in our dataset, we visualized all numerical fields using a heat map of Pearson correlation coefficients.

Correlation heatmaps of numerical fields revealed as in figure 6:

- `Tot Fwd Pkts` strongly correlates with `TotLen Fwd Pkts` - 0.7
- `Flow IAT Mean` strongly correlates with `Flow IAT Max` - 0.88
- `Flow Bytes/s` strongly correlates with `TotLen Fwd Pkts + TotLen Bwd Pkts`

In many network flows, a high maximum delay likely reflects consistent or occasional latency spikes, which in turn raise the overall average IAT. This kind of correlation is common in time-based metrics where outliers (e.g. long idle times) heavily influence averages.

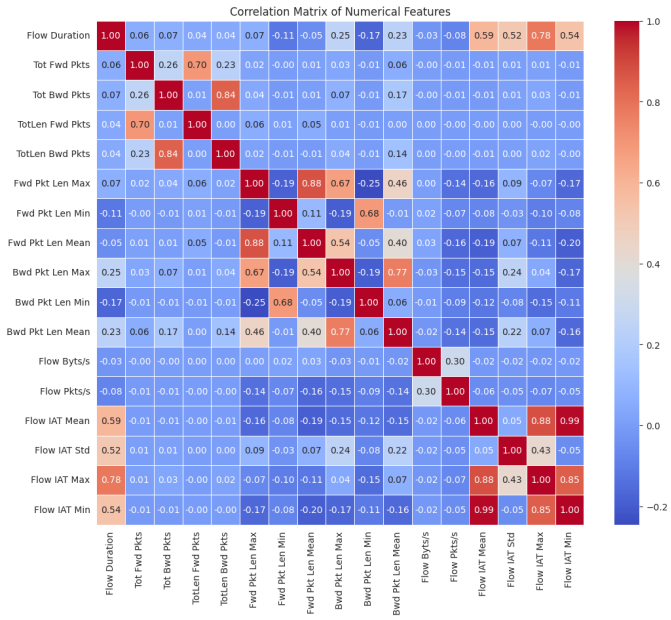

Figure 6: Correlation Analysis using Heatmap

Figure 6 still showcased a weak correlation between protocol flags and flow durations or byte counts. This means that these flags do not have a strong linear relationship with how long a flow lasts or how much data it transmits.

12 NETWORK TRAFFIC DISTRIBUTION ACROSS DESTINATION PORTS

Figure 7 shows that port 80, which is the standard port for HTTP traffic, is the most targeted port in the sampled network with over 30,000 connections. We attribute this kind of statistics, to the fact that port 80 is susceptible to various vulnerabilities due to its widespread use and the nature of HTTP requests.

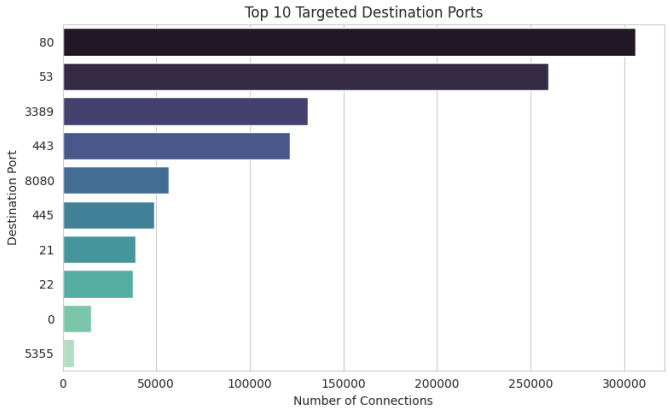

Figure 7: Network traffic distribution across destination ports

These vulnerabilities can allow attackers to compromise systems by exploiting flaws in web servers and applications. some of these vulnerabilities include disclosure of sensitive information, buffer overflows and path traversal.

13 ATTACK TYPES DISTRIBUTION ACROSS PORTS

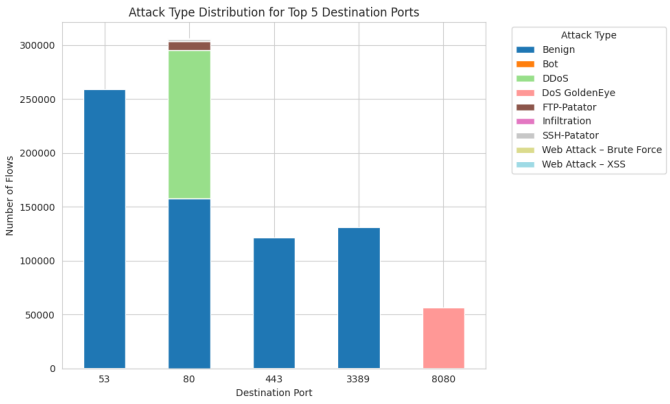

Figure 8: Attack types distribution across ports

This analysis Connects attack type to services that sit behind the individual ports. By looking at this visualization, we understand how attacks target specific services and applications.

The goal here is to enable more accurate threat detection and prevention. During configuration of firewalls, detail is key. A more detailed firewall/IPS policy is more effective than generic policies.

Other specific benefits of this analysis enabled:

- Targeted Attack Identification
- Vulnerability Assessment

In figure 8, we notice that ports 80, and 8080 face more malicious traffic compared to the likes of 443. It is important to note that port 443 is used for Secure Socket Layer(SSL) connections which might be the reason why we do not have a lot of malicious traffic hitting it.

14 PACKET SIZE ANALYSIS FOR EACH ATTACK TYPE

This relationship helps distinguish attack styles. Packet size analysis can help detect and differentiate between legitimate traffic and malicious activities, such as DDoS attacks or intrusion attempts.

By analyzing the packet size distribution and comparing it to known attack patterns, security professionals can identify anomalies and take appropriate action to protect their network.

From the visualization in figure 10, we note that Infiltration attack traffic packets are significantly larger, followed by bot and XSS attacks. This is due to various factors, including using fragmented packets, large data volumes. Infiltration attacks may also exploit vulnerabilities in protocols like UDP and ICMP to amplify the size of the attack traffic.

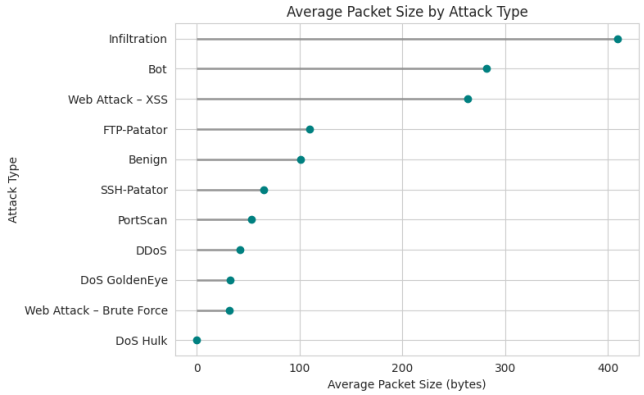

Figure 9: Packet Size analysis for each attack type

15 FLAGS USAGE ANALYSIS

With this visualization, we study the frequency of different flags used in packets. This can help identify attack characteristics based on the flags set in the packets. This information further informs the configuration of Intrusion Prevention Systems and Firewalls.

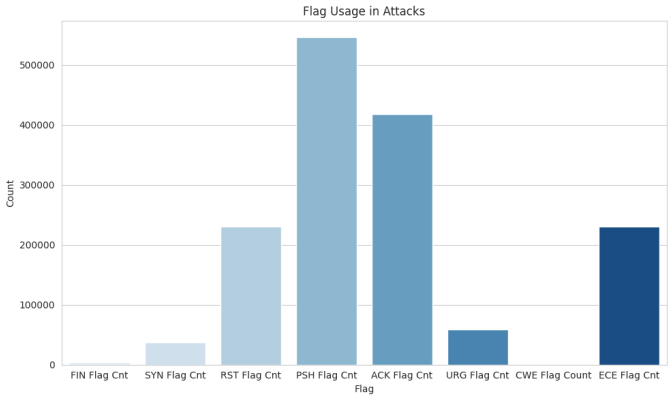

Figure 10: Flags Usage Analysis

We observe that the Push(PSH) Flag is the most used flag with the attack traffic. This flag is the most preferred in the attack traffic because of the following reasons:

- This flag Forces immediate processing.
- It is Used in payload delivery to deliver specific code or instructions quickly.
- The Push flag also Bypasses Some Defenses.
- It also Simulates Normal Behavior. Attackers mimic this behavior to make their traffic look normal and evade basic signature-based detection.

16 PACKET LENGTH DISTRIBUTION

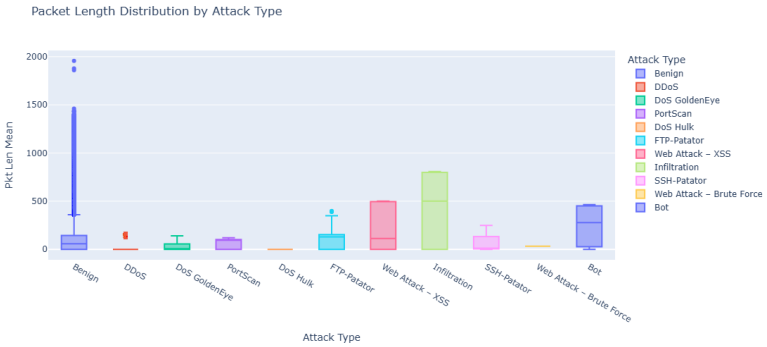

Figure 11: Packet Length Distribution

This visual helps us to distinguish between potential flooding (large packets) or probing (tiny packets). We note that Infiltration attacks have significantly longer packets across the network. This also explains why they take a significantly longer duration to traverse the network.

This analysis helped in:

i. Detecting Specific Attack Signatures. Many attacks have consistent packet length patterns.

ii. Differentiating Attack Types. Malware, scanning, brute-force, and data exfiltration all tend to show distinct packet size distributions. A port scan might consist of many very short packets, while Data exfiltration might show large payload packets.

iii. Behavioral Profiling. Normal applications (e.g., browsing, streaming) have predictable packet length ranges. Attack traffic often deviates from these norms.

iv. Evasion and Obfuscation Detection. Sophisticated attackers try to blend into normal traffic, but Slight but consistent differences in packet lengths can betray hidden patterns used in obfuscation or tunneling.

17 ATTACK TYPE DISTRIBUTION ACROSS THE TOP 10 DESTINATION PORTS

We observe that traffic related to Benign attacks majorly spans all ports except ports 22 and 21. We also notice that DDoS attack utilizes port 80 mostly.

This could be attributed to the fact that port 80 is easier prey. It is less secured compared to the likes of port 443. Port 80 is also a default port. This makes it easier for attackers to guess it.

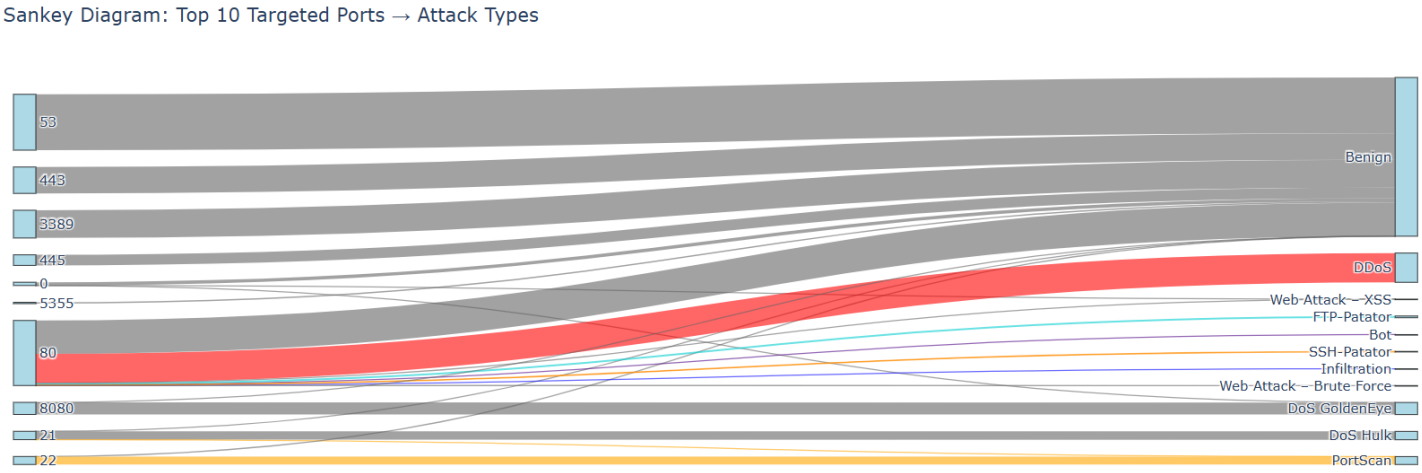

Figure 12: Attack Type Distribution across the top 10 destination ports.

18 TRAFFIC VOLUME BY ATTACK TYPE

Here, we observed that Brute force attack packets are the Forward Direction. This is because the attacker is the one initiating and aggressively sending repeated requests, while the victim responds minimally or not at all. Whereas Benign attack traffic move both ways(Forward and Backward).

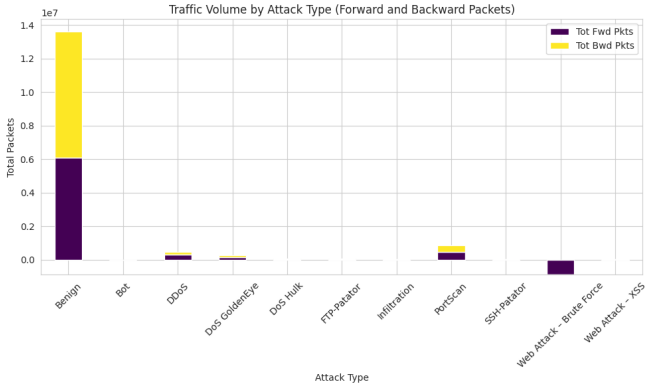

Figure 13: Traffic Volume by attack type

19 EVALUATION AND RESULTS

Key insights include:

- **Attack Types**: DDoS attacks are most prevalent. Benign traffic accounts for 77.5% of flows.
- **Port Usage**: Port 80 (HTTP) is most frequently targeted; secure ports like 443 face less malicious traffic.
- **Flow Duration**: Most attack flows are short-lived; XSS and infiltration last longer.
- **Correlation**: Strong correlation observed between flow IAT features and packet length.
- **Traffic Volume**: Benign traffic has balanced flow; attack traffic shows asymmetric patterns (e.g., brute-force floods).
- **Flags**: PSH flag heavily used in attacks, simulating urgency and mimicking normal behavior.

20 CONCLUSIONS AND FUTURE WORK

Our EDA of the CIC-IDS 2018 dataset uncovered several patterns of unauthorized access. Visualizations revealed differences in attack behavior across ports, packet size, and flow durations. These findings can assist in tuning IDS/IPS systems. In future, we aim to build classification models using these insights and explore streaming analytics for real-time threat detection.

REFERENCES

[1] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *ICISSP*.

[2] Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & Security*.

[3] SolarMainframe. (2022). CICIDS2018 - Intrusion Detection Dataset (CSV Format). Retrieved from https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv

[4] Moustafa, N., & Slay, J. (2015). UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set). In *2015 Military Communications and Information Systems Conference (MilCIS)*, IEEE.

[5] Sommer, R., & Paxson, V. (2010). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *IEEE Symposium on Security and Privacy*, pp. 305–316.

[6] Viegas, E., Santin, A., & Oliveira, C. (2017). A Visual Analytics Approach for Exploring Traffic Datasets in Computer Networks. In *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1036–1041.

APPENDIX

Guiding Questions

#	Question	Visualization / Relevance
1	What percentage of overall traffic is benign vs. malicious?	Pie Chart – Assess dataset bias and threat ratio
2	Top destination ports per protocol?	Bar Charts – Identify dominant targeted services
3	Distribution of attack types?	Bar Chart – Focus on prevalent threats
4	Flow duration distribution?	Box Plot – Identify anomalies and session characteristics
5	Flow durations by attack type?	Box Plot – Reveal behavioral traits
6	Correlation among numeric features?	Correlation Matrix – Aid modeling and pattern detection
7	Most targeted destination ports?	Bar Chart – Highlight high-risk services
8	Attack types vs. ports?	Stacked Bar – Map behaviors to services
9	Port-to-attack type relationship?	Sankey Diagram – Visualize multi-service attacks
10	Protocol usage across attacks?	Treemap – Detect abnormal usage
11	Avg. packet size per attack type?	Lollipop Chart – Reveal attack nature
12	Frequency of flags used?	Bar Chart – Indicate attack techniques
13	Traffic volume by attack type?	Stacked Bar – Measure impact
14	Packet length distribution per attack?	Box/Violin Plot – Identify unique signatures

Figure 14: Guiding Questions