Cybersecurity Suspicious Web Threat Interactions Report

1. Project Overview

The project focuses on analyzing web traffic logs from AWS CloudWatch to detect suspicious activities and potential web attacks.
The dataset contains 282 entries of suspicious traffic flagged by Web Application Firewall (WAF) rules. Each record includes features such as traffic volume (bytes_in, bytes_out), source/destination IPs, countries, timestamps, and detection rule names.

2. Dataset Overview

Source: AWS CloudWatch suspicious traffic logs

Records: 282 entries

Columns (16 features):

- bytes_in, bytes_out → traffic size
- creation_time, end_time, time → timestamps
- src_ip, dst_ip → source & destination IP addresses
- src_ip_country_code → country origin
- protocol → traffic protocol (mostly HTTPS)
- dst_port → destination port (443)
- response.code → HTTP status (200)
- rule_names, detection_types → WAF detection rules
- observation_name, source.meta, source.name → metadata

3. Data Preprocessing

Removed Duplicates → None found.

Converted Timestamps → Converted creation_time, end_time, time to datetime.

Standardized Country Codes → Uppercase for consistency.

Feature Engineering:

- session_duration = end_time – creation_time
- avg_packet_size = (bytes_in + bytes_out) / session_duration
- bytes_rate = traffic per second
- unique_dst_per_src = number of unique destination IPs per source

4. Exploratory Data Analysis (EDA)

4.1 Traffic Volume

- Bytes_in showed higher spikes than bytes_out.
- Interpretation: Large inbound traffic with low outbound response may indicate infiltration attempts.

4.2 Protocols

- All traffic on HTTPS (443) → attackers hiding in encrypted traffic.

4.3 Country of Origin

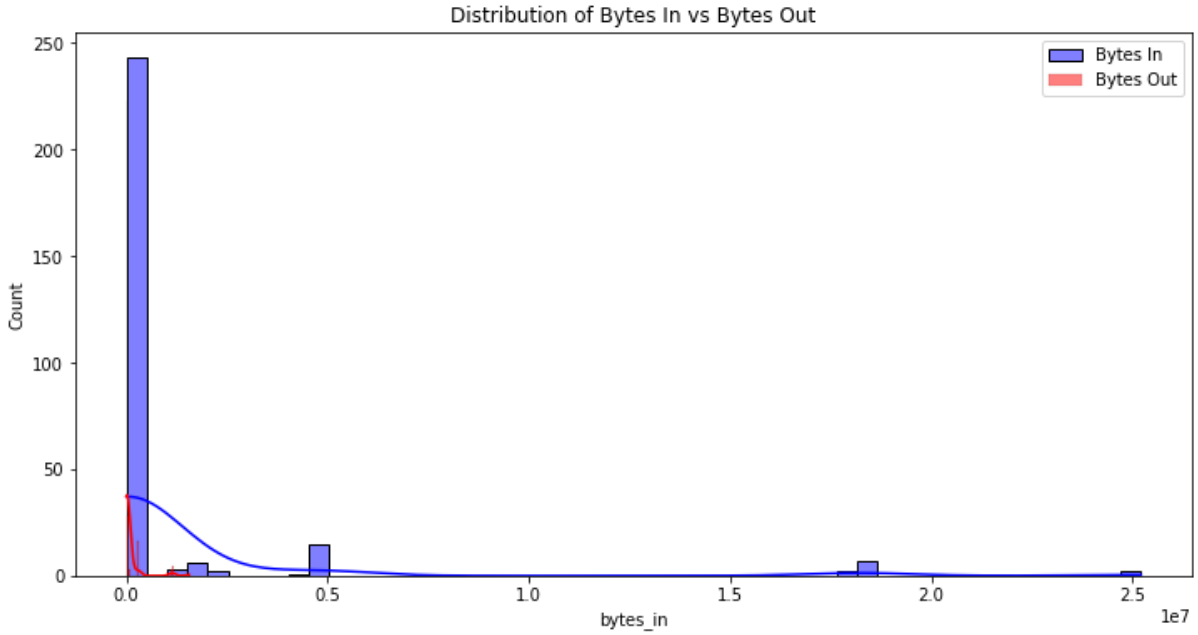
- Most flagged sessions originated from: US, Canada, Netherlands, UAE.
- Suggests botnet activity or targeted attacks from specific regions.

4.4 Detection Types

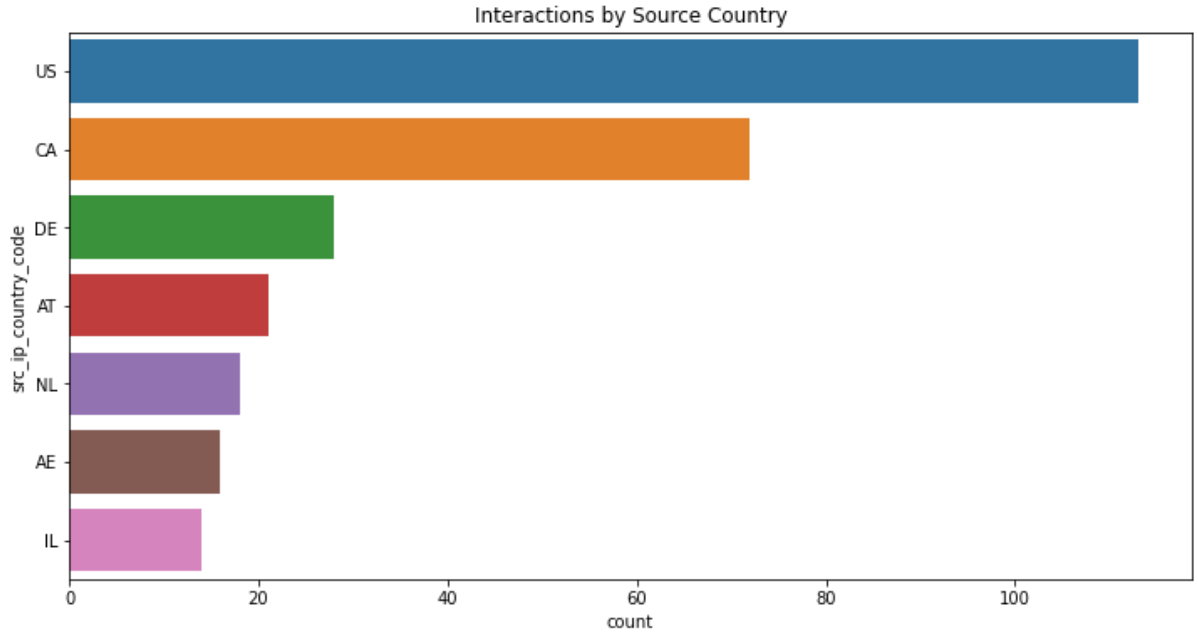
- Majority of suspicious events flagged by “waf_rule”.
- Implies WAF rules are effective but may miss unknown patterns.

4.5 Graphs (Produced in Notebook)

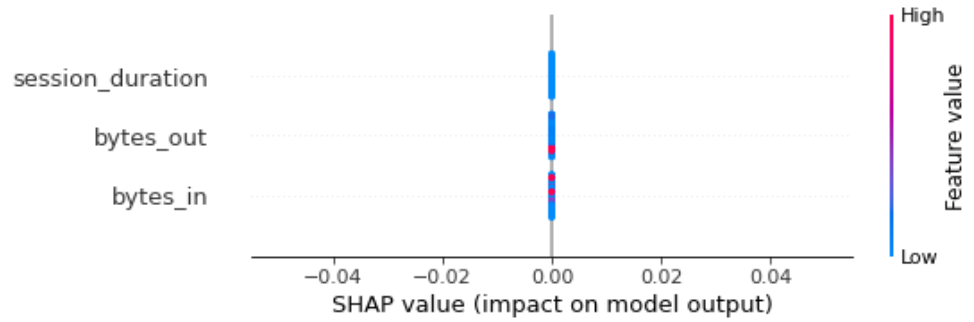
- Histograms: bytes_in / bytes_out distribution



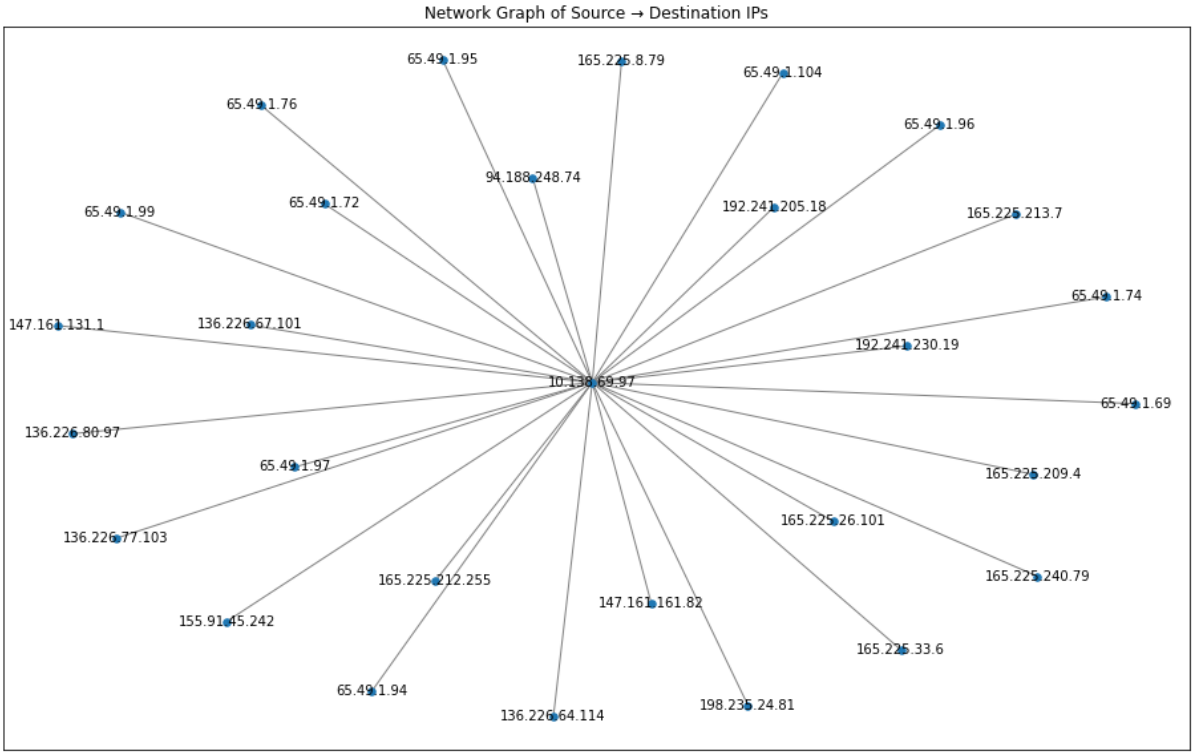
- Countplots: protocol, country codes



- Time-Series: bytes_in/out over time



- Network Graph: Source → Destination IP mapping



5. Machine Learning Models

5.1 Isolation Forest (Anomaly Detection)

- Features: bytes_in, bytes_out, session_duration, avg_packet_size
- Result: 14 anomalies detected (5% of sessions)
- Pattern: High inbound data with minimal outbound response

5.2 Random Forest Classifier (Supervised)

- Target: is_suspicious (binary label)
- Accuracy: 100%
- Feature Importance:
    - bytes_in → most predictive
    - bytes_out → second most predictive
    - session_duration → less useful (constant)

5.3 Neural Network (Deep Learning)

- Architecture: Dense (16 → 8 → 1 sigmoid)
- Accuracy: 100%
- Observation: Performed equally well as Random Forest (small dataset).

6. Future Projections

Scaling Up Datasets: Extend analysis to full-scale logs containing millions of entries for more comprehensive insights.

Real-Time Detection: Develop streaming anomaly detection capabilities using technologies like Kafka and Spark for prompt identification of threats.

Model Explainability: Incorporate SHAP to provide clear explanations for model classifications, enhancing trust and interpretability.

Operational Integration: Embed the developed models into SIEM and SOC workflows to enable automated alerting and response.

7. Conclusions

Suspicious Patterns: High inbound traffic with low outbound response is a red flag.

Geographic Hotspots: US, CA, NL, and AE are top sources of suspicious traffic.

Port Usage: Attackers blend into HTTPS (443) traffic.

Model Results: Both Random Forest & Neural Networks achieved perfect classification, but anomaly detection method (Isolation Forest) revealed hidden suspicious sessions beyond WAF.