Cybersecurity: Suspicious Web Threat Interactions

Project under Unified Mentor Data Science Internship

1. Project Overview

This project focuses on analyzing web traffic data to identify suspicious or potentially malicious activities. The goal was to understand how cyber threats behave through network traffic patterns and pinpoint sources of unusual activity.

The dataset used — **CloudWatch_Traffic_Web_Attack.csv** — contains details like source IP addresses, country codes, and data transfer volumes. By studying these attributes, we can detect abnormal traffic that might signal web attacks.

2. Data Preparation and Cleaning

Steps performed:

- Imported libraries (pandas, numpy, matplotlib, seaborn, plotly).
- Loaded dataset and checked dimensions with df.shape and df.info().
- Dropped columns with only one unique value to reduce redundancy.
- Converted time-related columns to datetime format.
- Converted categorical variables (src_ip, src_ip_country_code) to efficient data types.

Insight:
Data cleaning ensures more efficient analysis. Removing low-information columns reduces noise and improves focus on meaningful variables.

3. Exploratory Data Analysis (EDA)

a. Understanding the Data

A statistical summary was generated to get a sense of data ranges and variation.

b. Traffic Volume Analysis

Distributions of bytes_in and bytes_out were visualized using log-scaled histograms.

- Most traffic records showed very small byte sizes, while a few had extremely large values.
- These spikes could represent abnormal or suspicious activities like data breaches or DDoS attacks.

c. Countrywise Suspicious Activity

Using Plotly visualizations, the data was explored based on countries. Some countries had a higher share of suspicious traffic than others.

Insight: This analysis helps identify which regions or IPs generate the most unusual activity, making it easier to strengthen network security or block risky sources.

4. IP and Country Behavior

Further exploration focused on which specific IP addresses or countries were responsible for the highest amount of suspicious traffic or large data exchanges.

Insight: Identifying these “hotspot” IPs or regions allows cybersecurity teams to take preventive actions and apply targeted security measures.

5. Time-based Trends

Conversion of timestamps suggests analysis of:

- Attack frequency over time (hourly/daily trends).
- Peak periods for malicious activity.

Insight:
Helps in **predictive monitoring** — knowing when attacks are more likely can optimize defense resource allocation.

6. Key Takeaways

- The data shows strong **skewness**, where a small number of cases transfer very large volumes of data — a common sign of cyber threats.
- **Geographic analysis** highlights countries that need more monitoring or stricter access controls.
- **Time-based insights** help organizations stay alert during peak risk periods.
- Proper **data cleaning and feature preparation** improved the accuracy and efficiency of the overall analysis.