Cybersecurity Attacks Analysis – Final Report

Executive summary

The main finding is that response does not follow risk. Across 40,000 security events, a High-severity event was as likely to be ignored as a Low-severity one: 4,460 of 13,382 High-severity events (33.3%) were ignored, and 66.1% were not blocked.

This report covers 40,000 recorded security events from 1 January 2020 to 11 October 2023. The work ran from data cleaning and exploratory analysis in Python to a seven-page Power BI dashboard, with every headline figure cross-checked against an independent Python calculation.

Attack types, severity levels, protocols and response actions each split almost exactly into thirds. Statistical tests found no relationship between attack type and severity, between severity and response, or between anomaly score and severity. That evenness, together with every IP address appearing exactly once, strongly suggests the dataset is synthetic. The findings therefore demonstrate the analytical method rather than describe a real network.

The main recommendation is to make response risk-based: block High-severity and IoC-flagged events automatically, and review any that are ignored.

Objectives and dataset

The project set out to show what was attacked, how, when and how each event was handled, in a form security analysts and managers can explore themselves.

Specific questions:

1. Which attack types, severities, protocols and traffic types dominate?
2. Are some attack types riskier than others?
3. Do anomaly scores separate high-risk events from low-risk ones?
4. Is the response (Blocked, Logged, Ignored) matched to risk?
5. Are there time-of-day, weekday, yearly, device or geographic patterns?

The source is `cybersecurity_attacks_cleaned.csv`: 40,000 events (one row each) and 33 columns, cleaned from the original `cybersecurity_attacks.csv`.

Group	Columns	Notes
Time	timestamp, event_date, year, month, month_name, day_of_week, hour, is_weekend, time_period	1 Jan 2020 00:43 to 11 Oct 2023 19:34; 1,380 days, every day present. 2023 is a partial year.
Network	source/destination IP, source/destination port, protocol, packet_length, packet_type, traffic_type, network_segment, proxy_information	3 protocols (ICMP, UDP, TCP), 3 traffic types (DNS, HTTP, FTP); packet length 64–1,500 bytes; ports 1,024–65,535
Threat	attack_type, attack_signature, severity_level, anomaly_scores, malware_indicators	3 attack types, 2 signatures, 3 severities (no Critical level); anomaly score 0–100
Detection and response	alerts_warnings, ids_ips_alerts, firewall_logs, log_source, action_taken	Actions: Blocked, Logged, Ignored
User and device	user_information, device_information, geo_location_data	User names; browser user-agent strings; "City, State" locations in 28 Indian states
Payload	payload_data	Free text (placeholder Latin), not analysed

Five indicator columns are about half empty: malware_indicators 50.0%, alerts_warnings 50.17%, proxy_information 49.63%, firewall_logs 49.9% and ids_ips_alerts 50.12%. These are stored as empty strings rather than nulls, and an empty value means only that the indicator was not recorded.

Process

The work went through six stages, each producing an artefact the next stage relied on.

1. **Data cleaning (Python, `Cybersecurity_Attacks.ipynb`)**: produced `cybersecurity_attacks_cleaned.csv` from the original file. The cleaning steps are recorded in that notebook. The cleaned file adds derived time columns (year, month, month_name, day_of_week, hour, event_date, is_weekend, time_period).

2. **Exploratory data analysis (`cybersecurity_attacks-eda.ipynb`)**: univariate, categorical, numerical, bivariate and multivariate analysis, plus a Pearson correlation of the numeric columns.
3. **Visualisation and presentation (`Visualisation and Presentation.ipynb`)**: about 15 charts covering attack type, severity, protocol, response, anomaly-score and packet-length distributions, attack type by severity, protocol and year, events by hour, weekday and year, and a correlation heatmap. The notebook ends with a findings table of eight observations, each with evidence, implication and limitation.
4. **Statistical testing (Python, SciPy)**: chi-square tests of independence on attack type × severity, action × severity and action × attack type, and a Kruskal–Wallis test of anomaly score across severity levels. These test whether the differences seen in the charts are larger than chance.
5. **Power BI build**: Power Query transformations, a star-schema model, 52 DAX measures and a seven-page report with a dark theme. It is saved as a Power BI Project (PBIP: TMDL model and PBIR report) so every change is a readable text file.
6. **Validation**: every KPI was recalculated independently with pandas (`validate_kpis.py`, `expected_kpis.json`) and compared with the DAX results. The dashboard was then opened, refreshed and checked in Power BI Desktop.

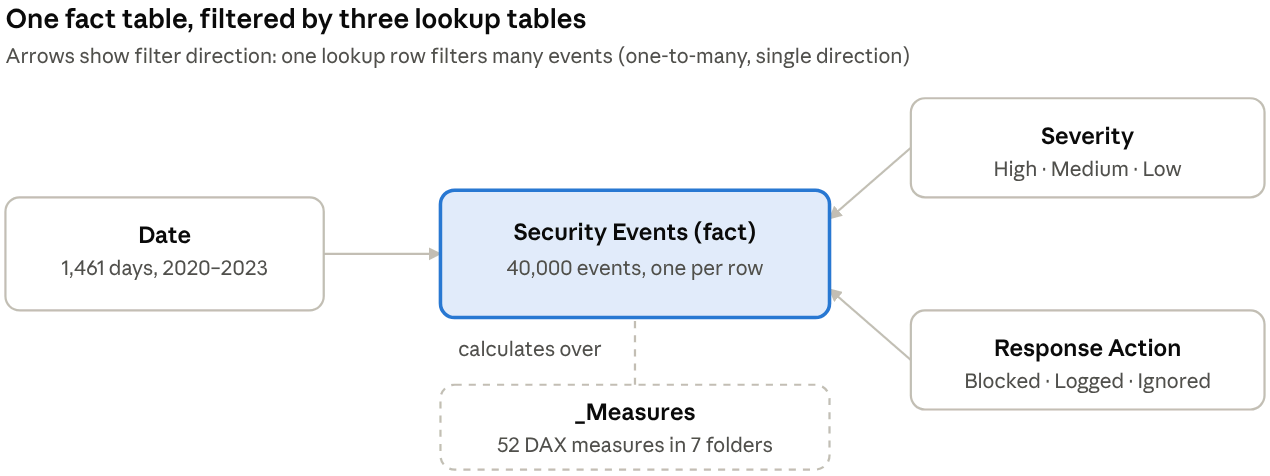
Power Query transformations

The original CSV is never modified; all changes happen in the query.

- **Loaded with quoted-field handling**: the payload text contains line breaks inside quotes, and this keeps each event on one row (40,000 rows in, 40,000 out).
- **Dates typed with the en-GB culture**: the file uses day/month/year, so this setting makes it load correctly on any PC's regional settings.
- **Empty strings in the five indicator columns converted to true blanks**.
- **Columns renamed to readable labels and an Event ID added** for row-level investigation.
- **Readable flag columns added**: Malware Flag, Alert Status, IDS/IPS Status, Firewall Log Status and Proxy Usage.
- **Location split into City and State**, and Device OS and Browser Family parsed from the user-agent string.
- **Bands for distribution charts**: 10-point anomaly bands, 200-byte packet bands and IANA port ranges, each with a sort key.

Power BI dashboard

The dashboard is a seven-page Power BI report built on a star schema. Every KPI is a DAX measure, so it updates with every slicer and cross-filter.



Power BI data model · 4 tables + measure table, 3 relationships

The three lookup tables give charts a logical sort order (High → Low, Blocked → Ignored) and consistent colours. The Date table also enables time intelligence (month-over-month and year-over-year). Auto date/time is switched off, and there are no many-to-many or bidirectional relationships.

Page	What it answers
Executive Overview	Six KPI cards, monthly attack trend, severity mix, top attack categories, response mix, key findings
Attack Analysis	Attack type share, frequency by weekday and by year → month (drill-down), type × severity, signatures, protocol mix, IPs ranked by anomaly score
Network Traffic	Traffic type, protocol, packet-size bands, packet type, source and destination port ranges, monthly traffic trend, port-range matrix
Threat & Severity	Severity distribution, anomaly bands, anomaly score by severity, IoC flags, indicators by attack type, device OS, states

Page	What it answers
Response & Operations	Blocked, Logged and Ignored counts, response by severity and by attack type, response trend, device × response, operational-gap table
Investigation	Row-level event log with Source IP and User search; drill-through target
Key Insights	Findings, recommendations and limitations

The interactivity works the same on every page:

- **Slicers**: date range, attack type, severity, protocol, traffic type and action taken, synced across pages.
- **Reset filters**: a button that clears every slicer.
- **Cross-filtering**: clicking a chart filters the rest of the page.
- **Drill-down**: the year → month chart.
- **Tooltips**: supporting measures appear on hover.
- **Styling**: a dark security-operations theme with fixed colours for severity (red, orange, teal) and response (green, orange, red).

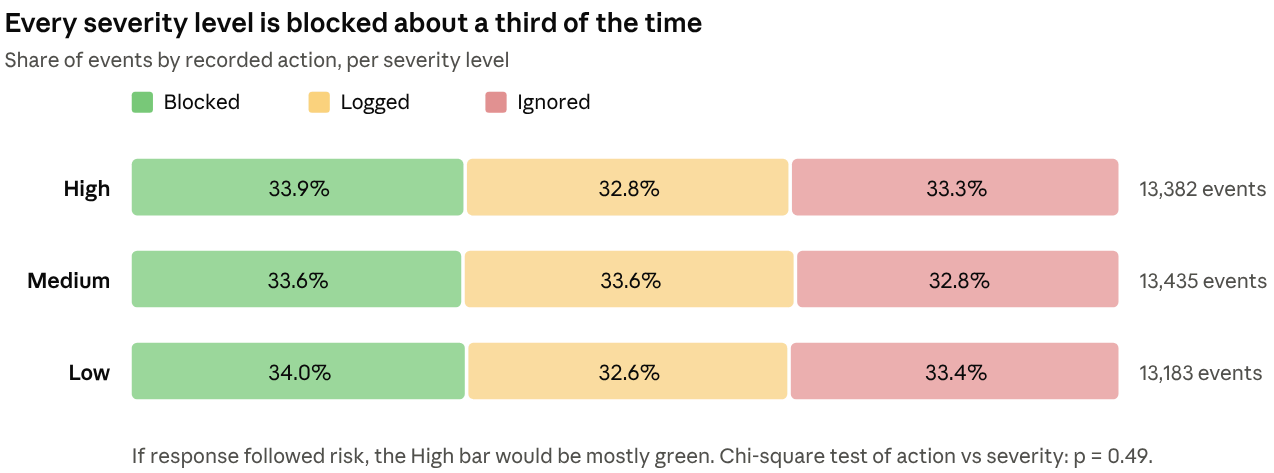
Measures are grouped into seven folders: Core KPIs, Severity, Threat Indicators, Response, Network, Time Intelligence and Formatting. All are documented in `DAX_Measures.pdf`.

Two requested measures were not built because the data cannot support them:

- **Critical-severity events**: no Critical severity level exists.
- **A true attack detection rate**: every row is already a labelled attack, and there are no benign or missed events to compare against. *IDS/IPS Detection Coverage %* (49.9% of attacks raised an IDS/IPS alert) is the closest valid alternative.

Findings

The most important finding is that response is not matched to risk. Every other dimension in the data is close to an even split, and none of the tested relationships is statistically significant.



cybersecurity_attacks_cleaned.csv · 40,000 events, Jan 2020 – Oct 2023

A High-severity event was blocked 33.9% of the time, a Low-severity event 34.0%. In total, 4,460 High-severity events (33.3%) and 6,576 IoC-flagged events (32.9%) were ignored, and 8,852 High-severity events (66.1%) had no blocking action recorded. Logging alone is not evidence of mitigation.

Area	Finding	Evidence
Volume and time	Volume is steady, with no growth trend	29.0 events per day on average (range 15–49); full months 788–950 events; 2020: 10,573, 2021: 10,538, 2022: 10,750, 2023: 8,139 (partial, to 11 Oct)
Time of day and week	No meaningful peak	Busiest hour 13:00 (1,737) vs quietest 12:00 (1,588); Tuesday 5,813 vs Wednesday 5,599; weekday 29.0 vs weekend 29.0 events per day
Attack types	Evenly split	DDoS 13,428 (33.6%), Malware 13,307 (33.3%), Intrusion 13,265 (33.2%); the mix is the same every year
Severity	Evenly split; no Critical level	Medium 33.6%, High 33.5%, Low 33.0%
Attack type vs severity	No attack type is riskier	Chi-square p = 0.77

Area	Finding	Evidence
Anomaly scores	Spread evenly from 0 to 100 and unrelated to severity or attack type	Mean 50.11, median 50.35; Kruskal–Wallis p = 0.32 across severities; DDoS 50.24, Malware 50.13, Intrusion 49.98
Response vs attack type	Response does not depend on attack type	Chi-square p = 0.48
Network	Protocols, traffic types and segments balanced	ICMP 33.6%, UDP 33.2%, TCP 33.2%; DNS, HTTP, FTP ≈ 33% each; packet length mean 781 bytes, spread evenly across 64–1,500
Numeric relationships	No correlation	All Pearson correlations among ports, packet length and anomaly score have $\|r\| \leq 0.005$
Indicators	Each flag present on about half of events	IoC detected 20,000 (50.0%); alert triggered 19,933; IDS/IPS alert 19,950 (49.9%); firewall log 20,039
IP addresses	No repeat sources or targets	All 40,000 source and 40,000 destination IPs are unique
Devices and geography	Small differences only	Windows 44.9% of devices, Linux 17.9%, macOS 14.5%, iOS 14.4%, Android 8.2%; top state Manipur with 1,498 events (3.7%) of 28 states

The uniformity is itself a finding. Real security logs are rarely this even: attacks cluster by time, source and target. This pattern, together with unique IPs and placeholder payload text, indicates a synthetic or simulated dataset.

Recommendations

The first two recommendations address the operational gap. The other four improve the data so that future analysis can say more.

#	Recommendation	Why (evidence)	Measure to track on the dashboard
1	Block High-severity and IoC-flagged events automatically	66.1% of High-severity events were not blocked; 6,576 IoC events were ignored	High Severity Not Blocked %, Malware Ignored Events
2	Send every ignored High-severity or IoC event to a review queue with an owner	4,460 High-severity events were ignored with no recorded follow-up	High Severity Ignored Events
3	Validate or recalibrate the anomaly model before using it for triage	Scores are uniform from 0 to 100 and unrelated to severity (p = 0.32)	Avg Anomaly Score by severity
4	Record outcomes: true or false positive, time to contain, resolved yes/no	"Logged" cannot currently be told apart from "resolved"	New outcome measures
5	Add asset criticality and stable host and account keys	Every IP is unique, so repeat attackers and exposed assets cannot be ranked	Unique Source IPs vs events
6	Log benign traffic alongside attacks	Without non-attack events a true detection rate cannot be calculated	A true detection rate

For the portfolio project itself:

- Fix the notebook so the findings table is saved (see Limitations).
- Parse dates with an explicit day-first format.
- Swap pie charts for bar charts where categories are near-equal, since 33% vs 34% slices are hard to compare.

Limitations, validation and quality checks

The headline numbers are verified. What is unverified is whether the data reflects a real environment.

Limitations

- **Likely synthetic data**: near-uniform distributions, unique IPs and placeholder payloads mean the findings demonstrate the method, not real-world attack behaviour.
- **Labels are not ground truth**: attack type, severity and action are as recorded, with no confirmation that an event was malicious or that a block succeeded.
- **Correlation is not causation**: the tests show associations (or their absence), not causes.
- **Partial final year**: 2023 ends on 11 October, so its yearly total and October's monthly total are not comparable with earlier periods.
- **No Critical severity and no benign events**: some standard KPIs (Critical events, true detection rate) cannot be calculated.

Validation

Check	Result
Rows loaded in Power BI vs CSV	40,000 vs 40,000
KPIs recalculated independently in pandas (`validate_kpis.py`)	Expected values stored in `expected_kpis.json`. The dashboard cards match: 40K events, 13K High severity, average anomaly score 50.11, 20K IoC, 33.8% block rate.
Power BI project opened, refreshed and rendered	Done in Power BI Desktop: all seven pages load, and the severity and response colours apply
Fixes after review	KPI cards had duplicated titles that clipped values; slicers had no room to draw. Both were fixed and re-checked.
Not yet checked	Exact DAX-vs-pandas comparison of every measure (run `KPI_Validation.dax` in DAX query view), slicer sync across all pages, drill-through setup on the Investigation page

Notebook quality notes

- **The findings CSV was saved empty.** In `Visualisation and Presentation.ipynb`, the cell that defines the eight findings ran (execution 23) before the cell that creates an empty findings table (execution 24). The empty table then overwrote them, and the save cell (execution 25) wrote only the header row. Running the cells top to bottom, or deleting the empty-table cell, fixes it.
- **Date parsing warning**: `read_csv` warns that day-first dates are being parsed without `dayfirst=True`. The parse is correct here, but an explicit `format='%d/%m/%Y %H:%M'` is safer.
- **Minor chart issues**: the day-of-week chart cell is duplicated. Two pie charts set x/y-axis labels, which pies ignore. Events by year is shown as a pie even though 2023 is partial, which makes 2023 look like a real decline.

Deliverables

All of these sit in `Cyber Security Attacks` Folder:

Deliverable	Location	Contents
Cleaned dataset	`cybersecurity_attacks_cleaned.csv`	40,000 events × 33 columns
Analysis notebooks	`Cybersecurity_Attacks.ipynb`, `cybersecurity_attacks-eda.ipynb`, `Visualisation and Presentation.ipynb`	Cleaning, EDA and presentation charts
Power BI project	`Cybersecurity_PowerBI_Project\PowerBI_Project\PBIP_Source\`	The `.pbip` file plus its `.Report` and `.SemanticModel` folders
Power Query	`Cybersecurity_PowerBI_Project\Power_Query\Transformations.m`	All five queries
Documentation	`Cybersecurity_PowerBI_Project\Documentation\`	Project overview, DAX measures, data model, user guide (PDF + Markdown)
Validation	`Cybersecurity_PowerBI_Project\Validation\`	`KPI_Validation.dax`, `validate_kpis.py`, `expected_kpis.json`

Appendix: statistical tests

Test	Variables	p-value	Reading
Chi-square	Attack type × severity	0.77	No association
Chi-square	Action taken × severity	0.49	No association: response is not risk-based
Chi-square	Action taken × attack type	0.48	No association
Kruskal–Wallis	Anomaly score across severity levels	0.32	No difference
Pearson correlation	Source port, destination port, packet length, anomaly score	$\lvert r\rvert \leqslant 0.005$ for all pairs	No linear relationship

A p-value above 0.05 means the observed differences are within what chance alone would produce.