Cybersecurity Dataset: Data Analysis Report

Data Analysis Project

Abstract

This project analyzes a cybersecurity event log dataset containing **20,000** events with timestamps, source/destination IPs, attack type and severity, response actions, data exfiltration flags, user agents, and unstructured threat-intelligence text. The goal is to describe event volume patterns over time, identify the most common and most risky attack categories, and summarize response outcomes using Python (EDA/statistics), SQL (schema + queries), and Power BI (interactive dashboards).

Contents

1 Introduction

This project analyzes a synthetic cybersecurity event log dataset using an end-to-end analytics workflow: data ingestion, cleaning and feature engineering, exploratory analysis, metric design, dashboard development, and interpretation of operational insights. The dataset contains timestamped security events with attributes describing *attack type*, *attack severity*, *source/destination*, *response action*, *threat intelligence text*, and a binary indicator for *data exfiltration*. The primary goal is to answer practical SOC-style questions such as:

- Which attack types dominate overall volume, and which contribute disproportionately to exfiltration?
- How does severity relate to exfiltration rate, and where is risk concentrated?
- When do events occur (hour/day/month), and are there temporal patterns useful for staffing and monitoring?
- How do response actions (blocked/contained/eradicated/recovered) relate to exfiltration outcomes?

The final deliverables include three Power BI dashboards for operational storytelling and two data model diagrams (schema and ERD) to document the dataset structure.

Security operations teams often need to quickly understand *what* is happening (attack types and severity), *when* it is happening (time patterns), and *how well* defenses are responding (response actions and residual risk such as data exfiltration). This report presents an end-to-end analysis workflow:

- **Python** for exploratory data analysis, feature engineering, and statistical summaries.
- **SQL** for structured querying and reproducible metric validation.
- **Power BI** for dashboarding and presentation-ready visualizations.

2 Dataset Overview

Each row represents a security event. Key fields include:

- **Timestamp**: event time used to derive Year, Month, Day of Week, and Hour.
- **Source IP / Destination IP**: network endpoints associated with the event.
- **Attack Type**: Malware, Phishing, DDoS, Ransomware, Insider Threat.
- **Attack Severity**: Low, Medium, High, Critical.
- **Response Action**: Blocked, Contained, Eradicated, Recovered.
- **Data Exfiltrated**: binary flag indicating exfiltration occurred.
- **User Agent**: client identifier string (high-cardinality).
- **Threat Intelligence**: unstructured text field with contextual notes.

2.1 Key Summary Metrics

Table 1 provides the headline metrics computed from the dataset.

Metric	Value
Total events	20,000
Exfiltration events	1,919
Overall exfiltration rate	9.59%
Critical severity events	5,025
High severity events	5,053
Blocked actions	5,020
Unique source IPs	*(computed in Python/Power BI)*
Unique destination IPs	*(computed in Python/Power BI)*
Unique user agents	16,327
Most frequent attack type	Malware (4,081 events)

Table 1: Headline dataset metrics.

3 Project Questions

The analysis focuses on the following questions:

1. **Volume patterns:** How do events vary by month, hour of day, and day of week?
2. **Threat composition:** What are the most common attack types and severity levels?
3. **Risk / impact:** Which categories have the highest data-exfiltration rate?
4. **Response outcomes:** How often are attacks blocked/contained/eradicated/recovered, and which outcomes correlate with higher exfiltration?
5. **Entity behavior:** What user agents dominate the dataset, and do they show distinct activity patterns?

4 Data Preparation

4.1 Cleaning and Type Casting

- Parsed `Timestamp` into a datetime type.
- Verified categorical columns (`Attack Type`, `Attack Severity`, `Response Action`) have consistent labels.
- Ensured `Data Exfiltrated` is treated as a binary numeric field for aggregation.

4.2 Feature Engineering

Derived fields were created for time-based analysis:

- **Year**, **Month**, **Month Name**
- **Day of Week**

- **Hour of Day**

These fields enable flexible slicing in Power BI and simpler grouping in SQL.

5 Python Analysis

Python was used to compute distributions, rates, and cross-tabulations.

5.1 Attack Type and Severity Distributions

- **Most frequent attack type:** Malware (4,081 events).
- Severity distribution is approximately balanced across the four categories, which supports comparisons of relative risk.

5.2 Exfiltration Rate by Category

Exfiltration rate is defined as:

$$ExfilRate = \frac{\#(DataExfiltrated = 1)}{\#(TotalEvents)}.$$

Key observations:

- **By attack type:** Malware shows the highest exfiltration rate (**10.88%**), followed by DDoS (**9.44%**).
- **By severity:** Low severity has the highest observed exfiltration rate (**9.93%**), while Critical is slightly lower (**9.33%**). (This pattern may reflect dataset generation/labeling rather than real-world behavior.)
- **By response action:** Events labeled *Blocked* still show the highest exfiltration rate (**10.14%**), suggesting either delayed blocking or that "blocked" indicates the attempted action rather than a fully prevented outcome.

5.3 Temporal Patterns

Time-based breakdowns were used to identify workload patterns:

- **Monthly trend:** event counts remain in a relatively stable band with moderate fluctuations, enabling comparison of periods without extreme seasonality.
- **Hour of day:** events are spread throughout the day, with mild peaks around mid-day and late afternoon.
- **Day of week:** weekday volume is slightly higher than weekend volume, consistent with enterprise activity patterns.

5.4 Threat Intelligence Text (Lightweight NLP)

Because `Threat Intelligence` is free text, a lightweight token-frequency scan was used to surface commonly repeated terms. This helps produce a quick "keyword cloud" style summary (shown on the dashboard) and can be extended to topic modeling or classification in future work.

6 SQL Component

SQL supports reproducible queries that validate metrics used in the dashboards (counts, rates, group-bys). Two diagrams were prepared to document the model.

6.1 Schema Diagram

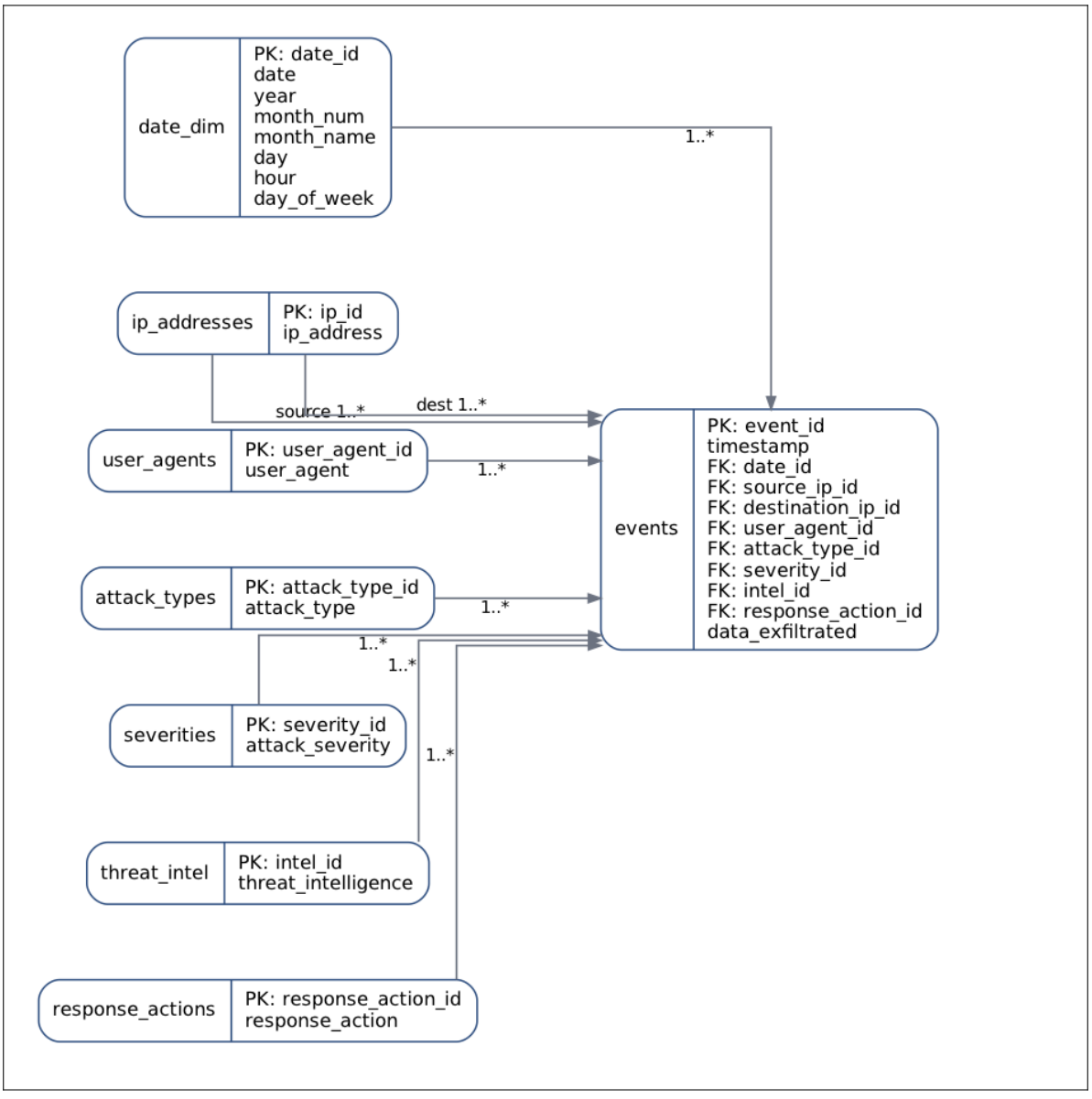

Figure 1: Schema-style diagram for the cybersecurity event dataset.

6.2 ER Diagram

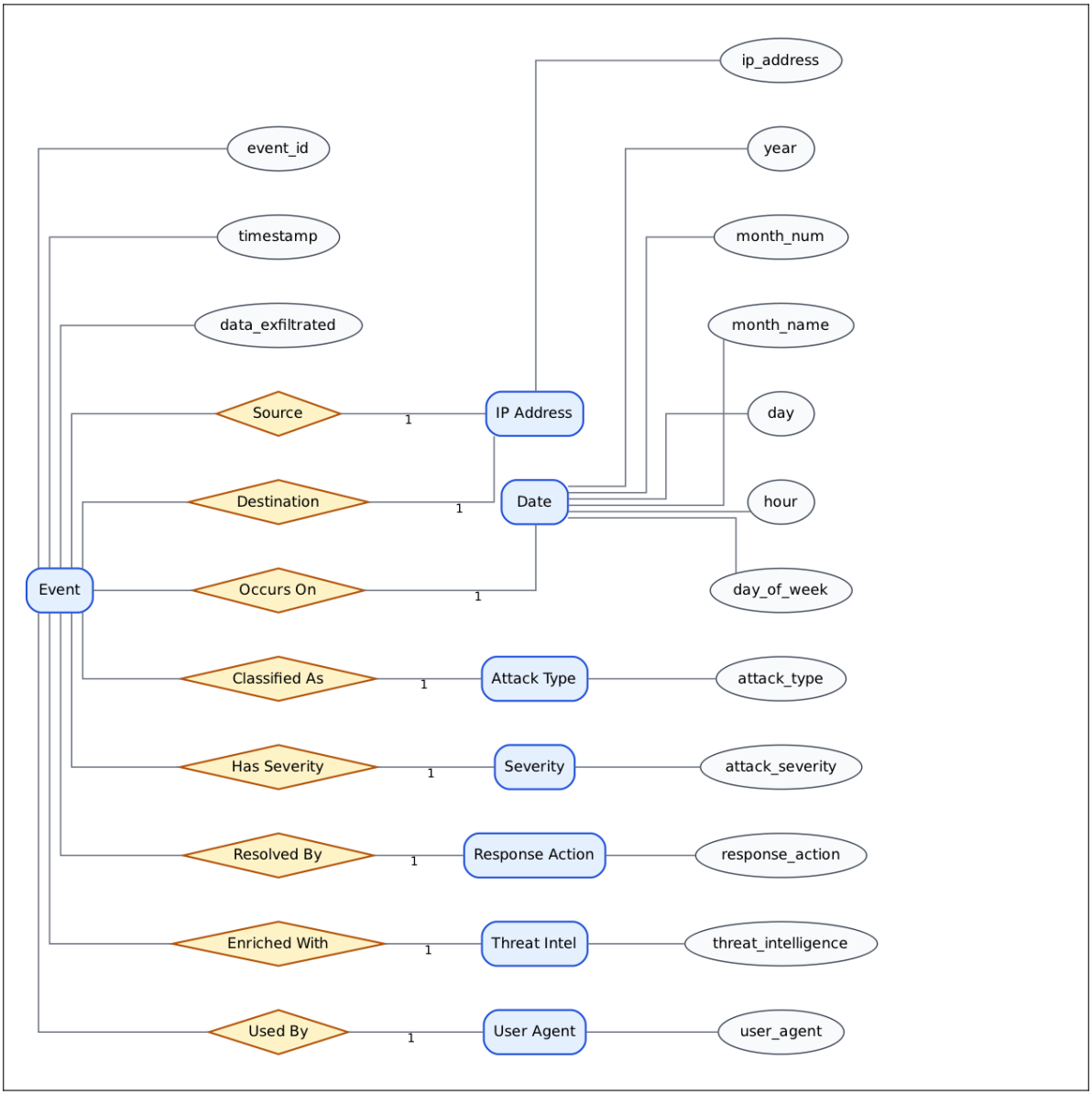

Figure 2: ER-style diagram representing cybersecurity event attributes and logical groupings.

6.3 Example SQL Queries

Below are representative SQL queries (syntax may vary slightly by database).

```
-- Total events
SELECT COUNT(*) AS total_events
FROM cybersecurity_events;

-- Exfil events and rate
SELECT
```

```
  SUM(CASE WHEN data_exfiltrated = 1 THEN 1 ELSE 0 END) AS exfil_events,
  AVG(CASE WHEN data_exfiltrated = 1 THEN 1.0 ELSE 0.0 END) AS exfil_rate
FROM cybersecurity_events;

-- Events by attack type
SELECT attack_type, COUNT(*) AS events
FROM cybersecurity_events
GROUP BY attack_type
ORDER BY events DESC;

-- Exfil rate by attack type
SELECT attack_type,
       AVG(CASE WHEN data_exfiltrated = 1 THEN 1.0 ELSE 0.0 END) AS exfil_rate
FROM cybersecurity_events
GROUP BY attack_type
ORDER BY exfil_rate DESC;

-- Monthly volume (example for PostgreSQL)
SELECT DATE_TRUNC('month', timestamp) AS month,
       COUNT(*) AS events
FROM cybersecurity_events
GROUP BY 1
ORDER BY 1;
```

7 Power BI Dashboards

Three dashboards were built to match the style of the healthcare project and to provide a clear narrative flow: (1) overall threat overview, (2) entities and behavior, and (3) response and intelligence.

7.1 Dashboard 1: Threat Overview

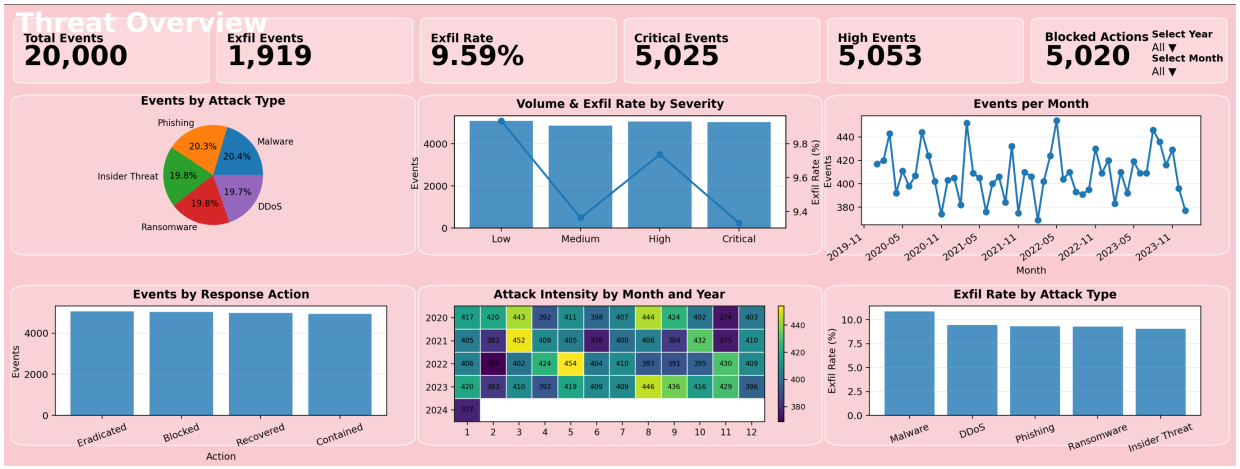

Figure 3: Threat Overview dashboard.

What it shows:

- High-level KPIs (total events, exfil events/rate, critical/high event volume, blocked actions).
- Attack type composition and exfiltration rate by type and severity.
- Month-by-month volume trend and an intensity heatmap (year × month).

7.2 Dashboard 2: Entity & Behavior Overview

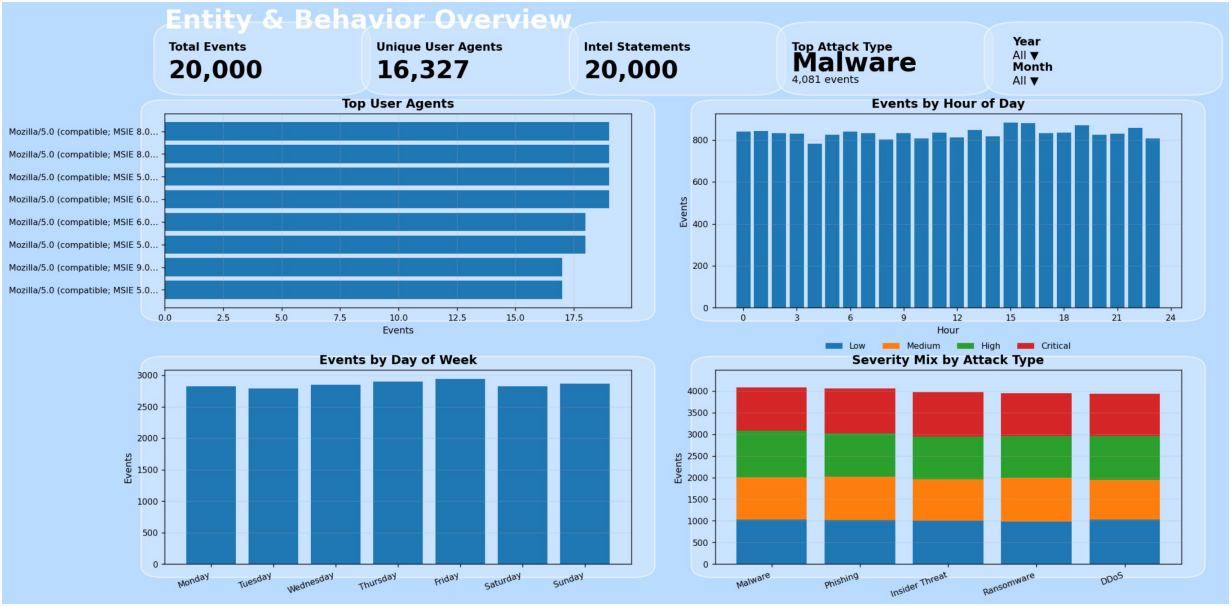

Figure 4: Entity & Behavior dashboard.

What it shows:

- High-cardinality **User Agent** field summarized with top-N bars.
- Time-of-day and day-of-week activity patterns.
- Severity mix by attack type (stacked bar).

7.3 Dashboard 3: Response & Intelligence

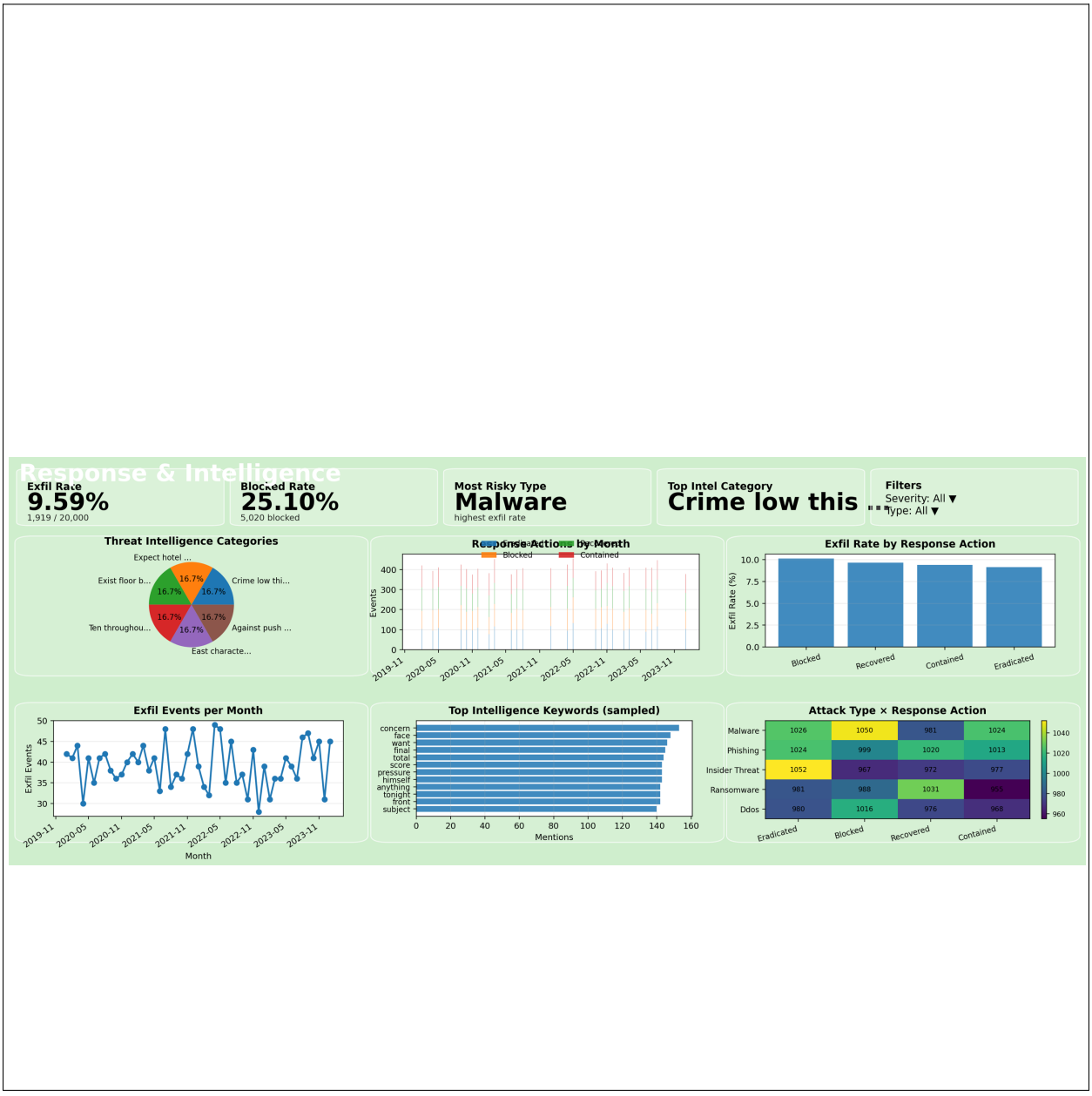

Figure 5: Response & Intelligence dashboard.

What it shows:

- Response action trends over time and exfil rate by response outcome.
- A keyword-based summary of **Threat Intelligence** notes.
- Attack type × response action heatmap for quick cross-comparison.

8 Key Findings

- **Malware dominates volume and risk:** Malware is the most frequent attack type (4,081 events) and has the highest exfiltration rate (10.88%).
- **Exfiltration exists across all categories:** Overall exfiltration rate is 9.59%, with relatively close rates across types and severities.
- **Response actions need careful interpretation:** "Blocked" events still show a non-trivial exfiltration rate, suggesting labels represent the response action taken rather than guaranteed prevention.
- **Activity is steady over time:** Monthly volumes fluctuate moderately without extreme spikes, making year-over-year comparisons feasible.

9 Limitations and Future Work

- **Label semantics:** Severity and response labels may not reflect real-world causal ordering (e.g., exfiltration despite blocking).
- **Text field is unstructured:** Threat-intelligence notes can be strengthened with NLP methods (topic modeling, clustering, supervised classification).
- **Entity enrichment:** IP reputation, geo lookups, and ASN mapping could add stronger behavioral insights.
- **Anomaly detection:** Time-series anomaly detection could highlight unusual bursts by type or by response action.

10 Dashboards and Findings

10.1 Dashboard 1: Threat Overview

The *Threat Overview* dashboard provides a high-level summary of the environment: overall event volume, exfiltration rate, and counts of high-severity activity. It combines KPI cards with distributions by attack type and severity, along with time-series plots and a heatmap.

What it shows and why it matters:

- **Events by Attack Type (Pie)**: Quickly highlights which attack categories dominate the dataset. A high share for a category can indicate a common attack vector or noisy telemetry.
- **Volume & Exfil Rate by Severity (Combo Chart)**: Separates *how often* each severity occurs from *how risky* it is. Even if a severity has lower volume, a higher exfil rate makes it operationally important.
- **Events per Month (Line)**: Reveals trends and bursts that might correspond to campaigns, changes in detection rules, or seasonal effects.
- **Attack Intensity Heatmap (Year $\times$ Month)**: Compresses multi-year data into an at-a-glance grid to identify recurring peaks.
- **Exfil Rate by Attack Type (Bar)**: A risk-focused view that can guide prioritization (e.g., hardening email controls for Phishing if exfil rate is high).

10.2 Dashboard 2: Entity & Behavior Overview

This dashboard shifts the lens from "what happened" to "who/what is involved" and "how activity behaves."

Key analytical angles:

- **Top User Agents**: Identifies frequent client signatures. Extremely repetitive patterns can indicate automated tooling or scripted activity.
- **Events by Hour of Day**: Helps determine whether activity clusters during business hours or off-hours (often a useful SOC staffing signal).
- **Events by Day of Week**: Captures weekly rhythms; for example, some attack traffic can peak mid-week.
- **Severity Mix by Attack Type (Stacked Bar)**: Shows whether certain attack types tend to appear at higher severities and whether severity labeling is consistent.

10.3 Dashboard 3: Response & Intelligence

The third dashboard connects outcomes and contextual intelligence.

Operational insights supported:

- **Threat Intelligence Categories (Pie)**: Summarizes the distribution of intelligence labels/keywords to understand narrative context.
- **Response Actions by Month**: Helps detect changes in containment posture or response automation over time.
- **Exfil Rate by Response Action**: A validation lens for response effectiveness. Ideally, "Blocked" should align with the lowest exfil rates.
- **Attack Type × Response Action (Heatmap)**: Highlights which responses are most common for each attack type, and can indicate playbook alignment (e.g., ransomware being more frequently contained/eradicated).
- **Exfil Events per Month**: Provides a direct measure of loss events over time.
- **Top Intelligence Keywords**: A lightweight text-mining view to surface repeated terms and themes from the Threat Intelligence field.

11 Interpretation and Discussion

Across the dashboards, the analysis emphasizes separating *frequency* from *impact*. High-volume categories demand efficiency and automation, while high-exfil-rate categories demand prioritization and prevention controls. Time-based views support staffing decisions and anomaly detection; entity/behavior views help explain whether activity resembles user behavior or automation.

12 Limitations

- The dataset is synthetic and may not reflect the true joint distributions found in real SOC telemetry.
- Exfiltration is represented as a binary label, which does not capture magnitude (bytes exfiltrated) or business impact.
- The Threat Intelligence field is short and not a full NLP corpus; keyword charts are indicative rather than definitive.

13 Future Work

- Add magnitude features (bytes, file counts) and build severity-weighted risk scores.
- Create an "entity enrichment" table (geo/IP reputation, ASN, known-bad lists) and join to events.
- Apply clustering or anomaly detection (e.g., unusual source IPs or off-hour surges) and surface results in a dedicated dashboard.
- Expand text analysis with TF-IDF topics or embedding-based clustering for richer intelligence summaries.

14 Conclusion

This project demonstrates a complete analytics workflow for cybersecurity event data. Python provides flexible EDA and metric validation, SQL supports reproducible query logic, and Power BI delivers clear, presentation-ready dashboards. The resulting visuals communicate both threat composition and response outcomes, helping stakeholders quickly understand risk and prioritize investigation.