Data Immersion-Achievement 6 Project Documentation

Ryan Wick – 05/13/2025

Data Source

Data Source Summary

- **Data Set Title:**
  - "Global_Cybersecurity_Threats_2015-2024"
- **Source:**
  - Kaggle
- **Brief Summary:**
  - This dataset provides comprehensive information on global cybersecurity incidents over a 10-year period, from 2015 to 2024. It contains data on various threat types, attack methods, impacted industries, geographic distribution, severity levels, and estimated financial losses. This dataset is particularly useful for threat intelligence analysis, trend forecasting, and the development of machine learning models to bolster cybersecurity strategies.
- **Variable Breakdown:**
  - Country: Country where the attack occurred
  - Year: Year of the incident
  - Attack Type: Type of cybersecurity threat (e.g., Malware, DDoS)
  - Target Industry: Targeted industry (e.g., Finance, Healthcare)
  - Financial Loss (in Million $): Estimated financial loss in millions
  - Number of Affected Users: Count of users that were impacted by the attack
  - Attack Source: The classification of individual or group responsible for the attack
  - Security Vulnerability Type: Type of exploit utilized by the bad actor(s) as their primary tool to carry out the attack
  - Defense Mechanism Used: Description of the countermeasures taken
  - Incident Resolution Time (in Hours): Time taken to respond and mitigate the threat
- **Notes:**
  - The file is provided in CSV format and is well-structured for data analysis.

Why this I choose this data set

- I selected this dataset because it aligns with my professional background in IT and my growing focus on data analytics. Cybersecurity is an increasingly critical area across all sectors, and this dataset offers an opportunity to explore patterns in digital threats, quantify their impacts, and identify which sectors and regions are most vulnerable.
  - **Specifically, this dataset allows for:**
    - Insightful trend analysis over a multi-year timeline
    - Quantification of both data loss and financial consequences

        - Exploration of response effectiveness and mitigation strategies
        - A rich mix of categorical and numerical data for analysis and visualization
    - **The dataset meets the requirements set out in the project brief:**
        - it is open-source, ethically sourced, recent, and sufficiently complex to support meaningful insights

---

Data Profile

Highlights:

- Financial losses range from **$0.5M to $100M**, with a median around **$50M**.
- Number of affected users ranges from **a few hundred to around a million**, highlighting wide attack impact variability.
- Incident resolution times vary from **1 hour to 72 hours**, suggesting inconsistent preparedness or resource allocation.

Common Dataset Profile Information

- **Variables and Data Types:**
    - Country - Qualitative / Nominal / Time-variant
    - Year - Quantitative / Discrete / Time-variant
    - Attack Type - Qualitative / Nominal / Time-variant
    - Target Industry - Qualitative / Nominal / Time-variant
    - Financial Loss (in Million $) - Quantitative / Continuous / Time-variant
    - Number of Affected Users - Quantitative / Discrete / Time-variant
    - Attack Source - Qualitative / Nominal / Time-variant
    - Security Vulnerability Type - Qualitative / Nominal / Time-variant
    - Defense Mechanism Used - Qualitative / Nominal / Time-variant
    - Incident Resolution Time (in Hours) - Quantitative / Continuous / Time-variant
- **Data Integrity Issues:**
    - No Missing or Duplicates located
- **Changed/Fixed Records:**
    - Column: Country
        - ***Then***: UK
        - ***Now:*** United Kingdom
        - ***Why:*** One of only two countries to be abbreviated. So to give a cleaner look and uniformity
    - Column: Country
        - ***Then:*** USA
        - ***Now:*** United States
        - ***Why:*** One of only two countries to be abbreviated. So to give a cleaner look and uniformity
    - Column: Financial Loss (in Million $)
        - ***Then:*** No $
        - ***Now:*** $0.00

        - ***Why:*** To give a cleaner output for any row to easily show it represents a dollar amount
    - Column: Number of Affected Users
        - ***Then:*** No (,) Separator
        - ***Now:*** 000,000
        - ***Why:*** Gives a cleaner look for output to tell when numbers are in the thousand and millions
- **Summary:**
    - 10 variables, 3000 records

Descriptive Statistics:

- **Quantitative Variables Analysis Results**

	Year	Financial Loss (in Million $)	Number of Affected Users	Incident Resolution Time (in Hours)
count	3,000.00	3,000.00	3,000.00	3,000.00
mean	2,019.57	50.49	504,684.14	36.48
std	2.86	28.79	289,944.08	20.57
min	2,015.00	0.50	424.00	1.00
25%	2,017.00	25.76	255,805.25	19.00
50%	2,020.00	50.8	504,513.00	37.00
75%	2,022.00	75.63	758,088.50	55.00
max	2,024.00	99.99	999,635.00	72.00

- **Qualitative Variables Analysis Results**
    - Top three most attacked countries: **United Kingdom, Brazil, India**
    - Most common attack types: **DDoS, Phishing, SQL Injection**
    - Heavily targeted industries: **IT, Banking, Healthcare**
    - Common attack sources: **National-State, Unknown, Insider**
    - Top security vulnerability types: **Zero-day, Social Engineering, Unpatched Software**
    - Most used defense mechanisms: **Antivirus, VPN, Encryption**

Limitations and Ethical Considerations

1. **Data Completeness & Accuracy**
    - Missing values may exist in key fields like financial loss or resolution time.
    - Underreporting is likely due to reputational risk or limited forensic capabilities in some regions.
2. **Bias in Representation**
    - Data may be skewed toward countries or organizations with better incident reporting systems.
    - Smaller organizations or countries may be underrepresented.
3. **Privacy & Ethics**
    - Even anonymized data about affected users could potentially be sensitive.
    - Use of this data must ensure no reverse identification of victims or vulnerabilities.
4. **Data Origin**

- While the dataset is public and educational, its original collection methods (news scraping, reports, etc.) may have subjective interpretations of severity or cause.

Questions to Explore

Top 3 Questions:

- How effective are different defense mechanisms in reducing resolution time or financial damage?
- How many millions are lost per hour of incident resolution and has this increased or decreased since 2015?
- Are there specific vulnerability types that are commonly exploited in high-impact attacks and which result in the highest losses?

Possible Reserve Questions:

- How have cyberattacks evolved from 2015 to 2024 in terms of volume and severity?
- Which countries report the most severe or frequent cybersecurity incidents?
- Which attack types are associated with the most significant financial impact or user data loss?
- Are certain attack types becoming more or less common over time?
- Which industries face the highest number of attacks and greatest financial losses?