NTRODUCTION

The rapid proliferation of cyber threats in the digital age has made it critical to develop innovative approaches to cybersecurity analysis. In response to this growing need, our project focuses on creating a robust and enriched dataset that consist of two primary data sources:

1. The Cyber Security Attacks dataset from Kaggle, which provides valuable information about cyber-attacks, attackers, and their victims.
2. The IP Geolocation API, offering geolocation metadata, including country, city, longitude, and latitude, based on IP addresses from Kaggle datset.

Our Methodology:

Unlike a simple merging of datasets, our approach actively leverages geolocation API to provide deeper insights. For each IP address from Kaggle dataset, we extracted:

- Geospatial data: Country, city, longitude, and latitude for both attackers and victims.
- Behavioral characteristics: Indicators of bots, anonymity tools like Tor, and other attack-related metadata.

Impact and Applications:

The enriched dataset has been made publicly available on Kaggle, ensuring open access for researchers, practitioners, and enthusiasts. Its applications could include:

- Exploratory data analysis to uncover trends in cyber-attacks.
- Machine learning model development for predictive analytics.
- Advanced cybersecurity strategies.

RESEARCH QUESTIONS

Cyber Threat Visualization and Communication

How can geolocation maps improve policymakers' and professionals' understanding of cyber threats?

Hotspot of Cyber Threats

Which countries are most frequently targeted or serve as origins for cyber-attacks?

Temporal Patterns in Cyber Attacks

Do cyber-attacks follow seasonal or periodic trends, and how can these inform predictions for risk periods?

Protocol Vulnerabilities in Cybersecurity

Which communication protocols are most frequently exploited, and how do these vary across attack types?

Severity and Type of Cyber Attacks

What factors influence the severity of cyber-attacks, and how do these differ by attack type?

Geopolitical Interactions in Cybercrime

What are the most common attacker-victim country pairs?

Behavioral Patterns of Attackers

What are the most common traits of attackers, such as using TOR, bots, or proxies...

Device and Operating System Vulnerabilities

Are attacker behaviors like bot usage more prevalent on specific devices or operating systems?

DATA SOURCES DESCRIPTION

In this project, two distinct datasets were utilized to generate a comprehensive and enriched dataset aimed at advancing the analysis of cybersecurity attacks. The datasets were sourced from Kaggle and an IP Geolocation API, each offering unique and complementary insights.

1 . KAGGLE

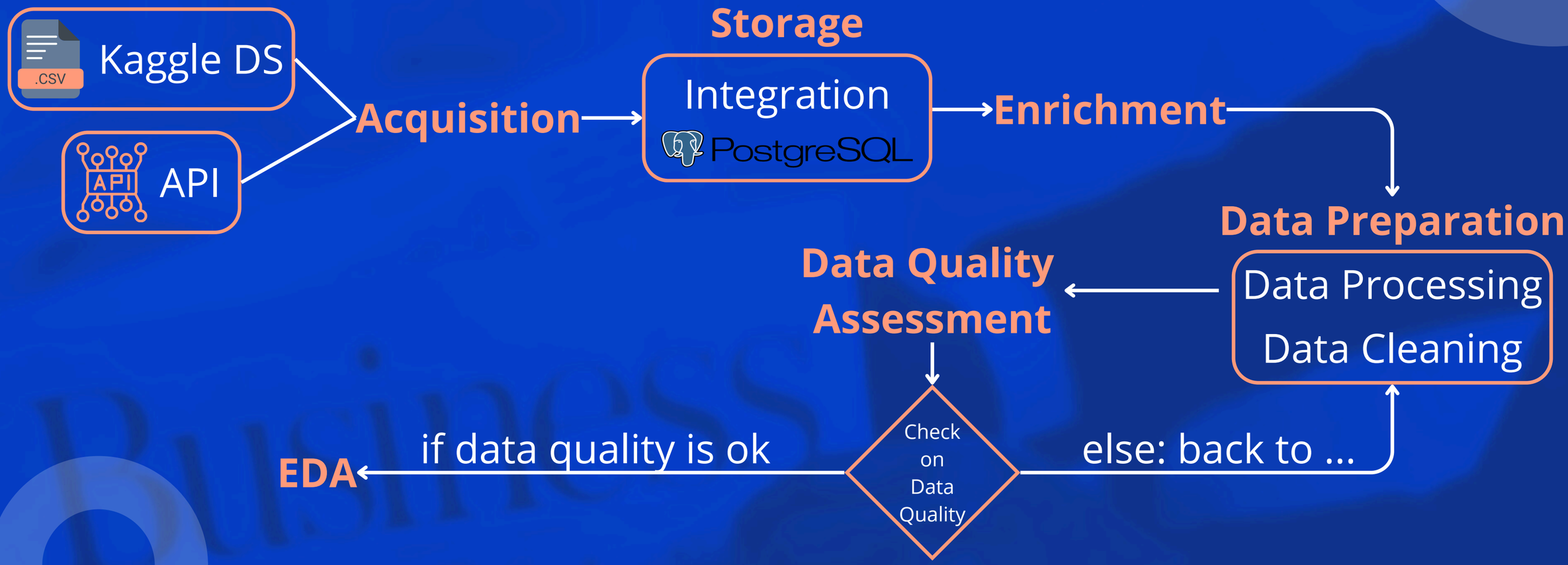
The Kaggle dataset, titled Cyber Security Attacks Cleaned, is a refined version of the original Cyber Security Attacks dataset, curated to address missing values and enhance data quality. This dataset is designed to provide a realistic portrayal of cybersecurity incidents, including attributes such as attack signatures, types, and travel histories.

2 . IP GEOLOCATION API

The second data source utilized in this project is derived from the IP Geolocation API, which provides real-time and precise geolocation and security information for any IPv4 or IPv6 address.
The API allow users to access geolocation information, including country, city, latitude, longitude, and additional security attributes, such as the use of anonymity tools like TOR or bots.

The Kaggle dataset served as the foundational source, offering structured information on attackers and victims, including their IP addresses. To enrich this data, the IP Geolocation API was employed to fetch additional geolocation details for both attackers and victims. Attributes such as country, city, latitude, longitude, and security-related features (e.g., TOR usage) were appended to the original dataset.

OUR DATA SCIENCE PIPELINE



DATA ACQUISITION

- In certain scenarions, data can be simply downloaded (e.g., the Kaggle dataset, our **first data source**)
- In other cases, such as with our second data source, a direct download is not possible. Specifically, **the second data source requires a sequence of steps to collect data from scratch** using the API ipgeolocation.

Why Collect More Data?

The initial dataset from Kaggle is insufficient for comprehensive analysis. Lacks critical information such as:

- IP location of hackers and targets.
- Cybersecurity insights tied to IP addresses.

Key Questions:

- Can we extract geolocation and cybersecurity info related to IPs involved in hacker attacks?
- Can this data be integrated with our existing dataset for a richer analysis?

Solution: Leveraging Additional Data Sources

By using IPGeolocation API, we can gather:

- Geolocation data (e.g., City, Country Name).
- Security-related information (e.g., Is Tor, Is Proxy , Threat Score).

DATA ACQUISITION

How we use the API to collect more data ?

Kaggle Dataset...

Source IP Address	Destination IP Address	---	other info about the attack

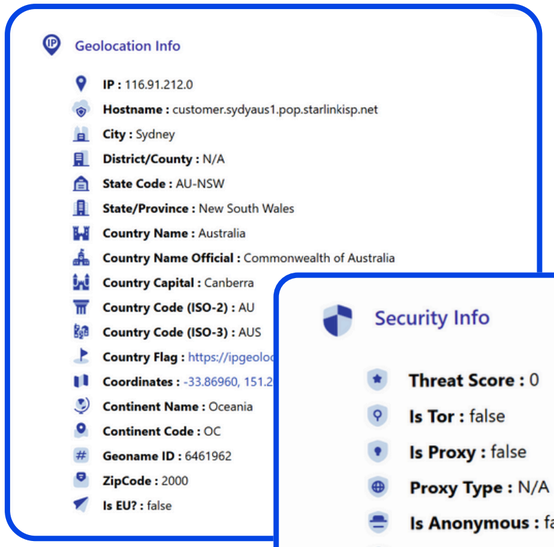
- The info contained is not enough...
- We want/need more info...

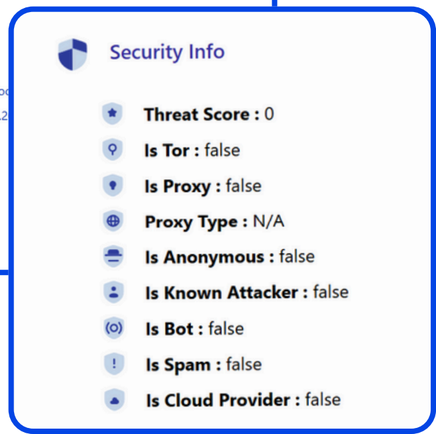
So we use the IPGeolocation API

INPUT:
an IP address (e.g., 116.91.212.0)

→ IPGeolocation →

OUTPUT:
geolocation (City, Continent Name, ...) and Security (Threat Score, Is Tor, ...) info about the IP address given as input





- The process on the right is repeated (more or less in the same way) **for each IP address in our dataset** (so for both the Hackers and Targets IP address).
- **How ?** By using the following function: *"JSON_frame_generation"*
  - This function takes a **list of IP addresses as input** and **generates (saves)** a JSON file containing a substantial amount of critical information about these IPs.
- The **output** of this entire process will be **two json files** (one for the info generatend and collected about the Hackers IP address and one for the info about the Targets IP address).
- Finally, we have builded a specially function to convert this two json files as csv (the ideal format for the final task, EDA).

DATA INTEGRATION

At the conclusion of the data acquisition phase, we need to manage **three datasets**, all in CSV format and for which, due their structure, each of them can be considered as a **table**:

- **Attacks:** Kaggle Dataset
- **Hackers:** Info about Hackers collected using the ipgeolocation API
- **Targets:** similar to the previous one but for the Targets

Why is a relational database ideal ?

1. **Structure and integrity**
   - A **fixed schema** ensures data consistency.
   - Primary and foreign keys enforce referential integrity, preventing duplicates and invalid data.
2. **Query efficiency**
   - SQL enables **complex queries** for aggregation and filtering, which are essential for exploratory data analysis (EDA).
3. **Processing efficiency**
   - Relational databases are designed to handle **large datasets** efficiently.
4. **Scalability and management**
   - Can **easily manage 40,000 observations** per table without significant performance issues.

Why is PostrgreSQL and ideal choiche ?

1. **Performance and scalability**
   - Optimized for managing and quering large datasets.
2. **Support for complex operations**
3. **Relationship management**
   - Fully suppports foreign keys, making it suitable for complex relational data models.

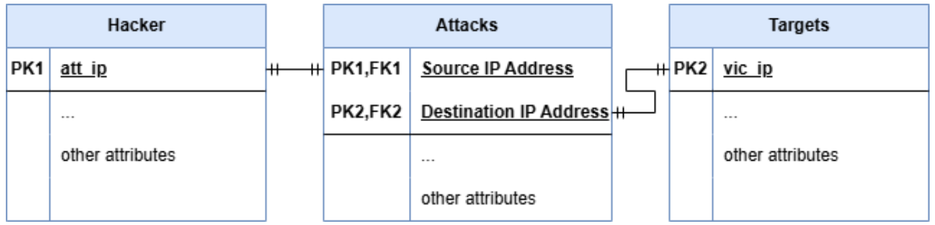
DATA INTEGRATION

How did we integrate our three tables ?

To ensure clarity regarding the **design of our database,** we emphasize that it has been **tailored specifically to the nature of the data we collected**. Here are the key considerations that guided our design decisions:

1. **Static Nature of the Data:**
   - Data are static and will not be updated in the future.
   - Primary goal of this project: answer the research question defined at the beginning, relying on EDA of the current dataset.
2. **Relationships Between Entities Based on the structure and characteristics of our data:**
   - One hacker is related to one and only one attack, which in turn is related to one and only one target.
   - Conversely, one target is related to one and only one attack, which is subsequently related to one and only one hacker.
3. **Implications on Database Design**:
   - This strict one-to-one relationship between hackers, attacks, and targets simplifies the design of our database, as each entity corresponds uniquely to another.
   - Foreign key relationships remain critical to enforce referential integrity and maintain the linkage between the entities.

The design of our database at glance...



DATA ENRICHMENT

In many projects, data integration is a precursor to data enrichment.

Starting Point:

- Data is **itegrated** into a **single system**

Enrich them by using additional useful attributes:

- Related to the **geolocation and security info**;
- And taken from the other two tables (Hackers and Targets).

Output:

- A **more insightful dataset** for analytics (but also for machine learning or decision-making).

How did we do it ?

- By **quering our database**
- We built a **specially query** named *“extractData4EDA”*:
  - **Designed to extract and enrich data from three tables:**
    - **Attacks** (the dataset to be enriched)
    - **Hackers and Targets** (the tables providing enrichment data)

Preparation for cleaning and Analysis:

- The resulting dataset serves as a **single, consolidated source** of truth for subsequent data cleaning, trasformation and exploratory tasks.

Next Step: Data Preparation →

DATA PREPARATION

Some of the Steps Taken:

- Check on missing values;
- Check on data formats;
- Check on duplicate data;
- Kept vs. removed columns;
- ...

Why Prepare the Data Now?
Reasons for Post-Acquisition Data Preparation

- During data acquisition, we used a **paid API** with a **limit of 150,000 requests**. Exceeding this limit would require additional payment.
- **The acquisition process required 80,000 requests** to generate the necessary project data:
- 40,000 for building the *Hackers* table.
- 40,000 for building the *Targets* table.
- **Adopted strategy:**

1. We collected all extractable data in one request to **fully utilize the service**.
2. Additional variables, while not currently necessary, might be useful for future projects.
3. After collection (through acquisition and integration), we extracted the data of interest (through enrichment) and performed preparation and cleaning only on that portion.

- **Advantages:** Cost optimization; Future flexibility; Error reduction.
- **Disadvantages:** Increased complexity (process large amounts of data); Longer processing times (additional steps to filter, clean and prepare required data); Overload risk.

DATA QUALITY

Processing and cleaning are essential but not always sufficient to guarantee the reliability and usability of the dataset.

- We need examining **various dimensions of data quality**, such as completeness, consistency, accuracy, timeliness, and relevance, to ensure that the dataset is both **robust and fit for purpose**.
- Data quality assessment is not a one-time activity but an **iterative process**:
  - If the assessment reveals additional issues, **further rounds of processing and cleaning may be necessary** to address these shortcomings.
  - By systematically evaluating and enhancing the quality of our data:
    - We strengthen the foundation for **meaningful insights**;
    - **Minimize** the risk of **erroneous conclusions**;
    - Ensures that the **subsequent analysis** is both **reliable and impactful**.
- **Objective:** incrementally improve the dataset until it **achieves a level of quality** that can be deemed **acceptable for proceeding with the EDA**.

Data Quality Assessmet - Step 1:
Checks on overall completeness and consistency for numeric variables.

Check on	Constraints	Results
Overall Completeness	Number of missing values	0 missing
Latitude	Must fall whitin [-90, 90]	[-45, 72]
Longitude	Must fall whitin [-180, 180]	[-158, 179]
IP Address'Threat Score	Must fall within [0, 100]	[0, 90]
Packet Lenght	Must fall within [20, 1500] bytes	[64, 1500]

Results:

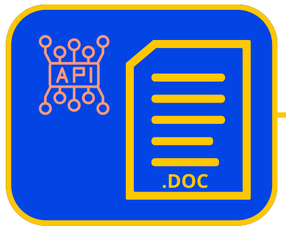
- **100% (perfect) of completeness (0 missing) in the current data;**
- **All the numeric variable are consistent with respect to their consraints.**

Step 2, 3 and 4 in the next page →

Data Quality Assessmet - Step 2: Checks for nominal variables.

- **Approach:** Reviewed all categorical or binary variables to **identify any anomalies or inconsistencies in their attributes**.
- **Results:** No anomalies or suspected issues were detected in the attributes.
- **Conclusion:** The nominal variables in the dataset meet the expected standards of **quality and consistency**.

Data Quality Assessmet - Step 3: Checks for the validity of IPs.



Looking at how our API works...
... If information about an IP address has been successfully collected, it means that the IP meets all constraints (e.g., it is a valid IP)...

Results:

- 0 missing values implies that **all IP addresses are valid**.
- **Perfect accuracy** for the IP addresses.

Data Quality Assessmet - Step 4: Checks on geospatial information.

We validated the data by **cross-checking** it with **two official resources** (datasets) containing lists of **officially recognized cities, countries, and continents**.

Potential Issues: Since we are using the FREE VERSION of the official datasets, some valid geospatial information might not appear in the dataset and, therefore, cannot be validated, even if it is correct.

Results:

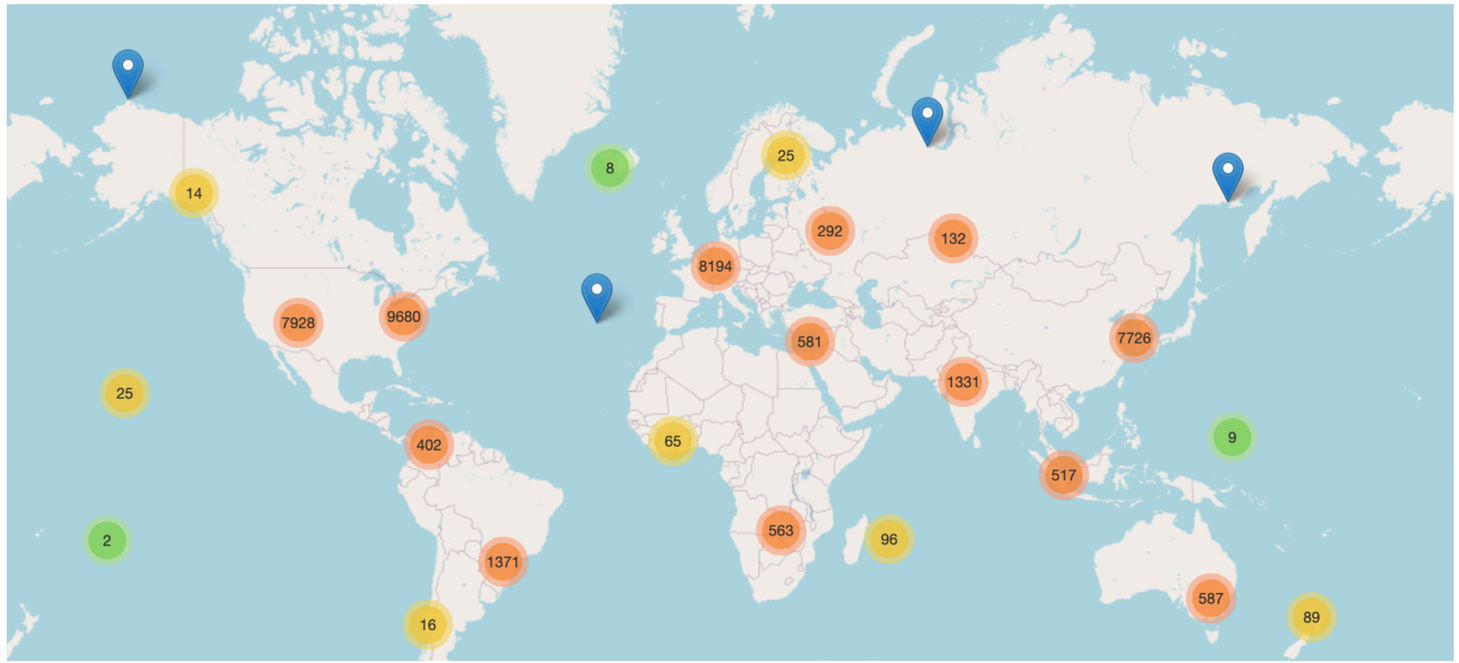
- 5.6% of cities not validated;
- 0.4% of countries not validated;
- Perfect results for the rest (country codes 2/ISO2 and continent validity)

Final Considerations:

- Our dataset presents **excellent data quality** in terms of completeness, validity, and consistency.
- The amount of remaining **dirty data is minimal** and will not impact the subsequent analysis (EDA).
- Our dataset for the EDA will yield real, **unbiased results** and uncover **valuable insights without the risk of incorrect analysis**.

EDA

GEOLOCATION DATA OF HACKERS

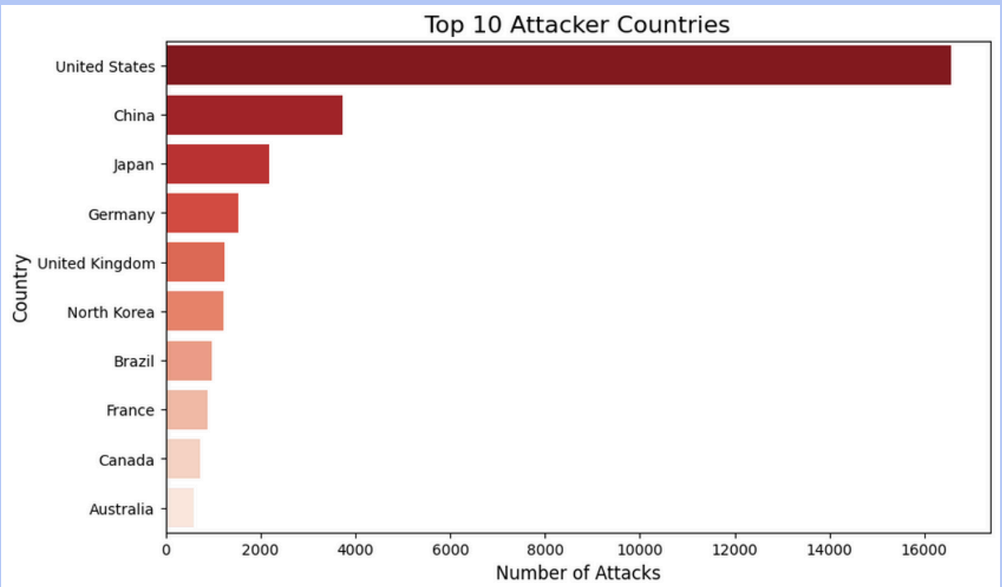


Clusters dynamically adjust as users zoom in or out, providing an intuitive way to explore areas of interest.
Each marker within the cluster contains a popup that provides detailed information about that specific attacker.

The interactive map is seamlessly integrated into a Jupyter Notebook environment, offering a dynamic and engaging interface for users to explore the geolocation data. The map allows users to zoom in on specific regions, investigate individual markers, and gain insights into the spatial distribution of cyber threats. This interactive visualization approach not only enhances the interpretability of complex data but also aids in identifying geographic hotspots and potential regions of heightened cybersecurity risks.

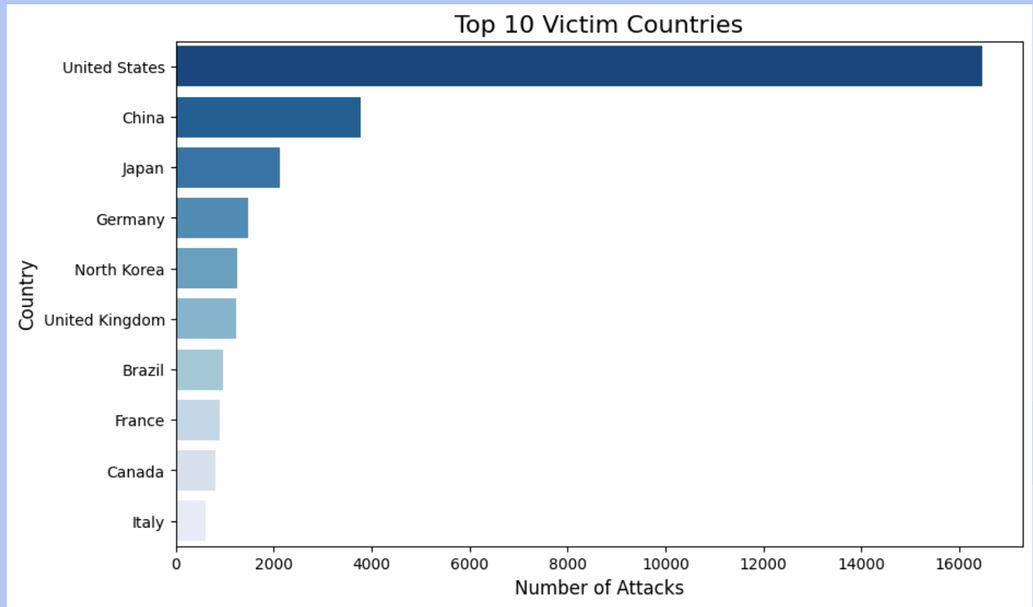
EDA

TOP HACKER AND TARGET COUNTRIES

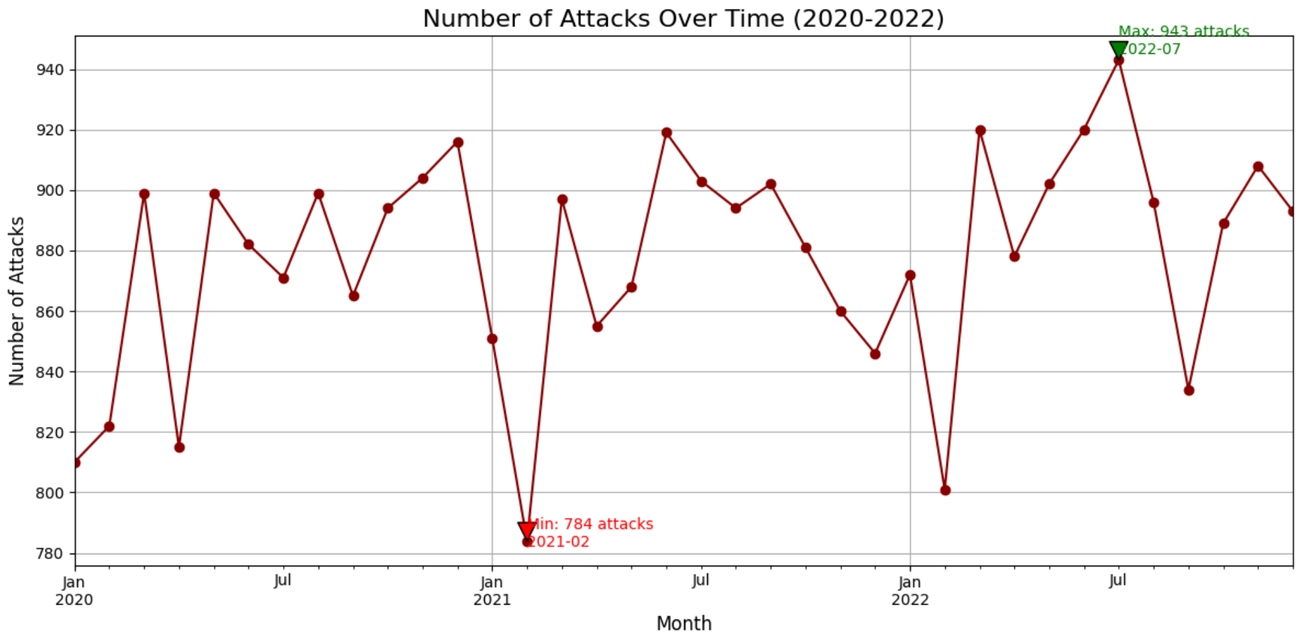


Our findings reveal that the United States consistently ranks highest among the top 10 countries, both as a leading source of attackers and as a primary victim of cyberattacks.

This dual role reflects the complexity of cybersecurity dynamics in regions with advanced digital infrastructures. The United States prominence as both an attacker and victim highlights the critical need for enhanced cybersecurity measures, not only to safeguard its digital ecosystems but also to address vulnerabilities that make it a target for external threats. These patterns provide valuable intelligence for policymakers, researchers, and cybersecurity professionals, enabling them to craft more effective strategies to combat cybercrime.



TEMPORAL ANALYSIS OF ATTACKS

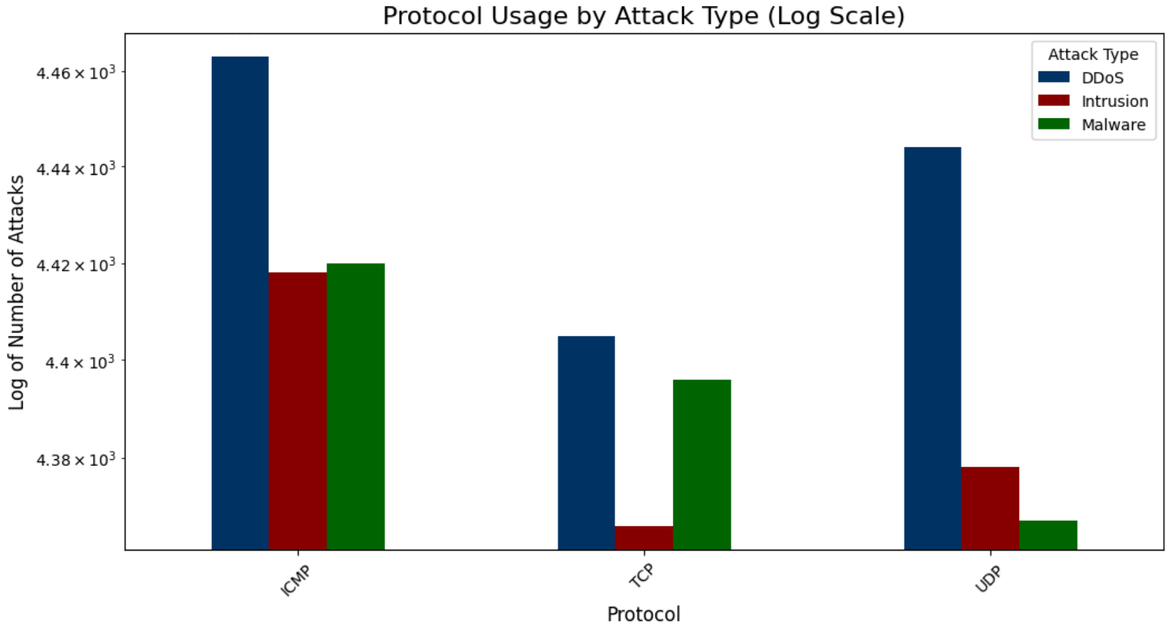


This analysis investigates potential seasonal or periodic patterns in cyberattacks, examining fluctuations in attack frequencies over time.

the graph highlights notable trends in attack activity over time. The data reveals that the highest number of attacks, reaching a peak of 943 incidents, occurred in July 2022. This surge may indicate specific factors driving increased malicious activity during that month, such as targeted campaigns or exploitation of seasonal vulnerabilities. On the other hand, the lowest number of attacks, recorded at 748 incidents, was observed in February 2021. This marked decline might reflect a temporary lull in attacker activity or the effectiveness of mitigation measures during that period.

COMMON PROTOCOLS USED BY ATTACKERS

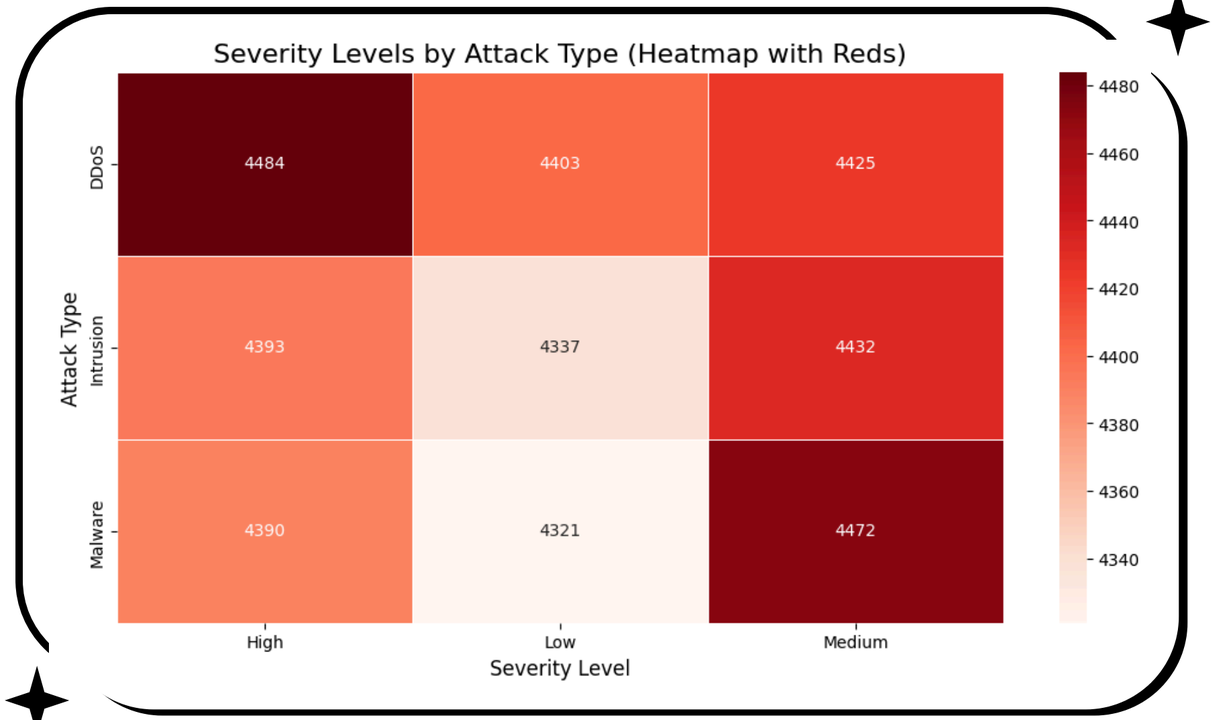
DDoS (Distributed Denial of Service): Overwhelm a system with traffic.
Intrusion: Unauthorized access to exploit systems or steal data.
Malware: Software designed to harm, spy, or hold systems hostage.



ICMP Attacks: Exploit diagnostic protocol for floods or tunneling.
TCP Attacks: Abuse connection-oriented protocol via SYN floods or hijacking.
UDP Attacks: Exploit the connectionless nature for amplification or flooding.

As it shows on the graph DoS attacks overwhelmingly dominate all protocol usage, indicating that this attack type is the most prevalent and widely distributed across different network protocols. The chart highlights the extensive reliance of DDoS attacks on ICMP, UDP, and TCP, with ICMP being the primary protocol of choice. This dominance suggests that attackers often exploit the characteristics of ICMP and other protocols to achieve their goal of overwhelming a target's network resources.

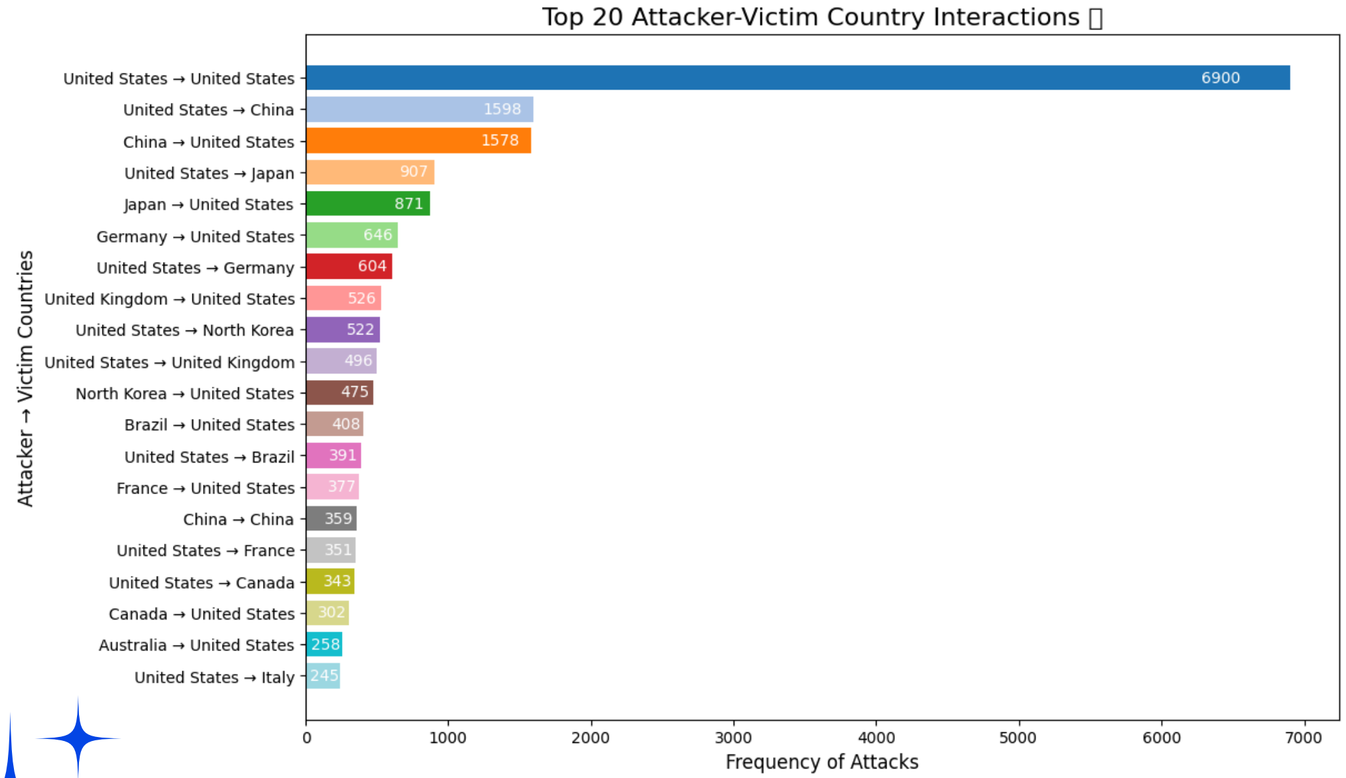
SEVERITY LEVEL ACROSS ATTACK TYPES



HEATMAP

This heatmap reveals the distribution of attack severity levels across different attack types, providing critical insights into their impact. DDoS attacks stand out as the most severe, with the "High" severity level exhibiting the highest count among all categories. This makes DDoS attacks the most critical type to address in terms of cybersecurity measures. Intrusion attacks show a more balanced distribution across severity levels, with the "High" severity category slightly leading, indicating a moderate but varied threat level. On the other hand, malware attacks predominantly fall into the "Medium" severity category, suggesting they are less critical but still capable of causing significant disruption.

TOP 20 ATTACKER-VICTIM COUNTRY INTERACTIONS



The graph displays the frequency of cyberattacks between different countries.

US Dominance: The United States is heavily involved in both launching and receiving cyberattacks. It appears in 12 of the top 20 interactions, both as an attacker and a victim. This suggests the US is a major player in the global cyber threat landscape.

- **US-China Rivalry:** The interaction with the highest frequency is between the United States and China, with both countries attacking and being attacked by each other. This highlights a significant cyber rivalry between these two nations.
- **Other Prominent Players:** Other countries frequently involved in cyberattacks include Japan, Germany, the United Kingdom, Brazil, and Canada. These countries often appear as both attackers and victims.
- **Attacker-Victim Asymmetry:** In many cases, the frequency of attacks is not reciprocal. For instance, the US attacks China more often than China attacks the US. This suggests that cyberattacks are not always symmetrical and can be strategically targeted.

ANALYZING HACKER CHARACTERISTICS

Category	YES	NO
Hacker used TOR	2	39655
Hacker used spam	221	39436
Hacker used bot	11	39646
Hacker used proxy	1036	38621
The IP belongs to a cloud provider	7630	32027
Hacker is anonymous	1038	38619
Hacker is known attacker	1021	38636

The data reveals that most hackers do not use TOR, making TOR usage rare. Spam appears as a more frequent method (**221**) compared to bots or anonymity, while proxies (**1,036**) are a preferred choice for masking identities. Notably, **7,630** attacks originate from cloud provider IPs, highlighting their exploitation in cyberattacks. The number of anonymous hackers (**1,038**) slightly exceeds known attackers (**1,021**), indicating a balance between anonymity and traceability. Bot usage is minimal (**11**), making it the least common method in the dataset.

- **Proxy Usage** and **Cloud Provider** Exploitation stand out as key methods, suggesting these are areas requiring more stringent monitoring and preventive measures.
- **Known Attacker** IPs indicate that previously identified malicious IPs are still active, underscoring the need for continuous blacklisting efforts.
- Despite popular assumptions, the low **usage of TOR** and bots suggests that these methods might not be as prevalent as **proxies** or **cloud IPs** for hiding hacker activities.

INVESTIGATING ATTACKER TRAITS ACROSS DEVICE/OS TYPES

Variable	Chi2	p-value
att_is_bot	23.148151	0.000748
att_is_tor	7.041048	0.317068
att_is_proxy	7.165424	0.305820
att_is_anonymous	6.894278	0.330735
att_is_known_attacker	5.609734	0.468294
att_is_spam	9.797785	0.133430
att_is_cloud_provider	3.431856	0.753013

This analysis is crucial for uncovering whether certain attacker traits are more prevalent on specific devices or operating systems. By identifying these associations, cybersecurity teams can develop targeted defenses and mitigation strategies. For instance, if a strong relationship is found between bots and certain OS types, additional security measures can be implemented to protect these systems. The highest Chi-square value, 23.148 for ‘att_is_bot‘, indicates a strong association between the presence of bots and the type of device/OS used during attacks. This finding suggests that certain devices or OS types are more likely to be targeted by or associated with bot-related attacks. For the trait ‘att_is_bot‘, the p value of 0.000748 indicates a statistically significant relationship with the type of device/OS. This means bots are more commonly associated with specific device/OS types, potentially revealing usage or targeting patterns.

CONCLUSION

In this project, geographical and temporal analyses highlighted regional hotspots and seasonal spikes in cyber activity, providing guidance for targeted and proactive defense strategies. Examining attack protocols revealed communication channels requiring stronger defenses, while severity analysis pinpointed high-severity attack scenarios that demand intensified mitigation efforts.

Our investigation into attacker traits, such as TOR, proxy, and bot usage, uncovered patterns to guide security measures. A chi-square test revealed a strong association between bot activity and specific devices/OS types, emphasizing the need for tailored defenses. Additionally, analyzing attacker-victim interactions exposed critical geopolitical patterns, informing cybersecurity policies at national and international levels.

This analysis deepens our understanding of cybersecurity trends and offers actionable insights for strengthening defenses.

Future Improvements: Enhancements could include integrating more diverse datasets (e.g., timestamps, attribution data), applying machine learning for anomaly detection, attack prediction, and clustering, and enabling real-time analysis with automated alerts. Collaborating with cybersecurity entities could improve data access, threat intelligence, and research impact. Optimizing data preparation, expanding geospatial validation sources, and standardizing the data pipeline for flexibility and reproducibility would further refine the analysis.