LLP113 Advanced programming and Data visualisation

Sohag Noman - F434565

Global Cybersecurity Threats:

This coursework aims to explore the cybersecurity landscape over the last 10 years. The dataset provides extensive data on cyberattacks, mainly designed for threat impact analysis and machine learning model development.

The dataset was provided by Atharva Soundankar, author of the dataset on Kaggle.

The analysis is structured as follows:

1. Aims & Objectives

This report investigates incidents between 2015 and 2024.

The core objective is to explore some useful insights that may be key to understanding patterns in the ongoing battle against cyber adversaries.

The analysis explores:

- Where and which types of cyberattacks had the highest prevalence and inflicted the most damage?
- How the cyberattacks evolved over the past decade?
- Which sectors have been targeted by cybercriminals?
- What is the financial impact of these cyber incidents?

The tools used for this project are the following:

Tools	*Description*
Python3	Programming language to implement the project.
Pandas	Data Analysis tool, a Python library.
Numpy	Array manipulation tool, a Python library.
Matplotlib	Python library for data visualization
Plotly	Python library for data visualization
Jupyter	ipykernel is Jupyter kernel built on top of IPython (That provides .ipynb formato)
VScode	IDE used for coding
Obsidian	Tool to write
Sk-learn	ML library to train models

2. Dataset Description

The dataset provides the following:

Column	Type	Description
Country	Categorical	Country where attack happened
Year	Numeric	Year of the attack
Attack Type	Categorical	Type of cyber attack
Target Industry	Categorical	Industry targeted
Financial Loss (in Million $)	Numeric	Target variable (continuous) — financial loss in million dollars
Number of Affected Users	Numeric	Number of users affected
Attack Source	Categorical	Source of the attack (Hacker group, insider, etc.)
Security Vulnerability Type	Categorical	Type of vulnerability exploited
Defense Mechanism Used	Categorical	Defense used to counter the attack
Incident Resolution Time (in Hours)	Numeric	Time taken to resolve incident

3. Exploratory Data Analysis

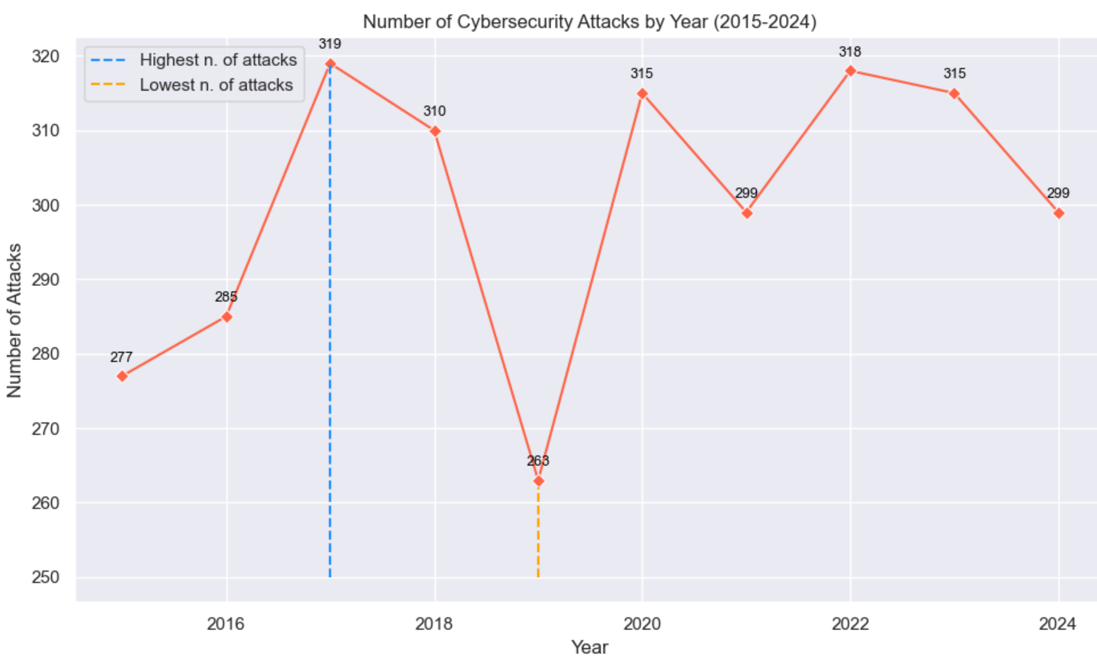
```python
# Visualize the dataset
cyberAttacks.head(10)
```

✓ 0.0s Python

	Country	Year	Attack Type	Target Industry	Financial Loss (in Million $)	Number of Affected Users	Attack Source	Security Vulnerability Type	Defense Mechanism Used	Incident Resolution Time (in Hours)
0	China	2019	Phishing	Education	80.530	773169	Hacker Group	Unpatched Software	VPN	63
1	China	2019	Ransomware	Retail	62.190	295961	Hacker Group	Unpatched Software	Firewall	71
2	India	2017	Man-in-the-Middle	IT	38.650	605895	Hacker Group	Weak Passwords	VPN	20
3	UK	2024	Ransomware	Telecommunications	41.440	659320	Nation-state	Social Engineering	AI-based Detection	7
4	Germany	2018	Man-in-the-Middle	IT	74.410	810682	Insider	Social Engineering	VPN	68
5	Germany	2017	Man-in-the-Middle	Retail	98.240	285201	Unknown	Social Engineering	Antivirus	25
6	Germany	2016	DDoS	Telecommunications	33.260	431262	Insider	Unpatched Software	VPN	34
7	France	2018	SQL Injection	Government	59.230	909991	Unknown	Social Engineering	Antivirus	66
8	India	2016	Man-in-the-Middle	Banking	16.880	698249	Unknown	Social Engineering	VPN	47
9	UK	2023	DDoS	Healthcare	69.140	685927	Hacker Group	Unpatched Software	Firewall	58

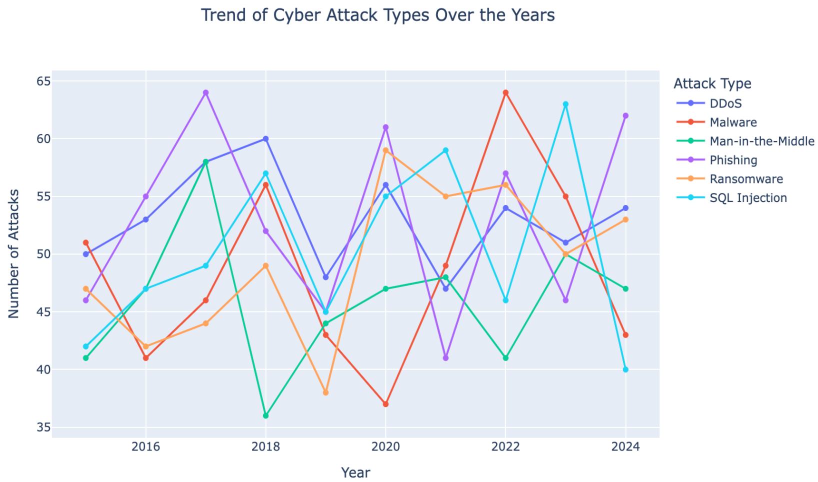
How has the number of cyberattacks varied over the years from 2015 to 2024?

By analyzing the yearly data, we aim to uncover whether there are any significant increases or decreases in cyberattack activity over the given period.

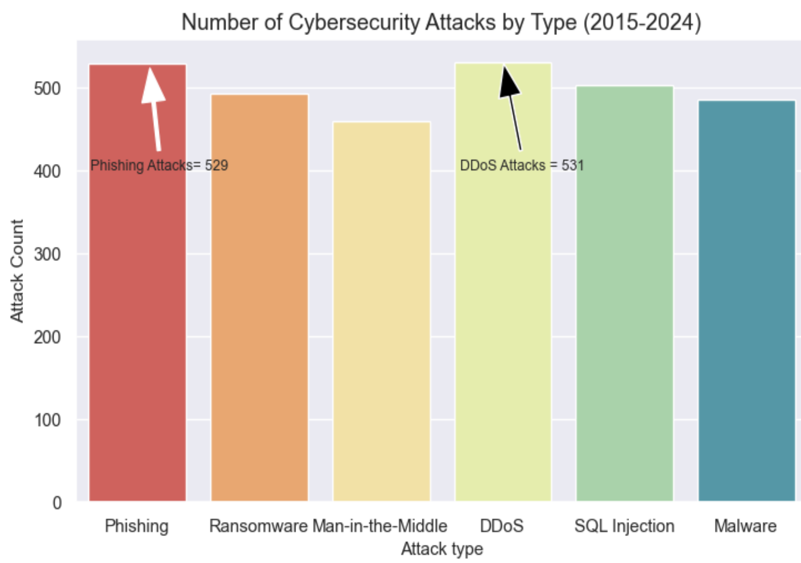


Insights:

- Stable Increase in Attacks: The overall increase of 7.94% in attacks over the past 9 years indicates a slight upward trend in cyber threats globally.
- Peak in 2017: The year 2017 stands out as the peak year, with the highest number of attacks (319).
- Low in 2019: The 2019 dip, with only 263 attacks, could indicate a temporary reduction in cyberattack activity.
- Consistency Post-2020: This may reflect a steady phase where cybersecurity measures are managing the threat levels but not preventing a significant increase in incidents.



The most common types of cyberattacks over the period from 2015 to 2024, quantifying their occurrence.

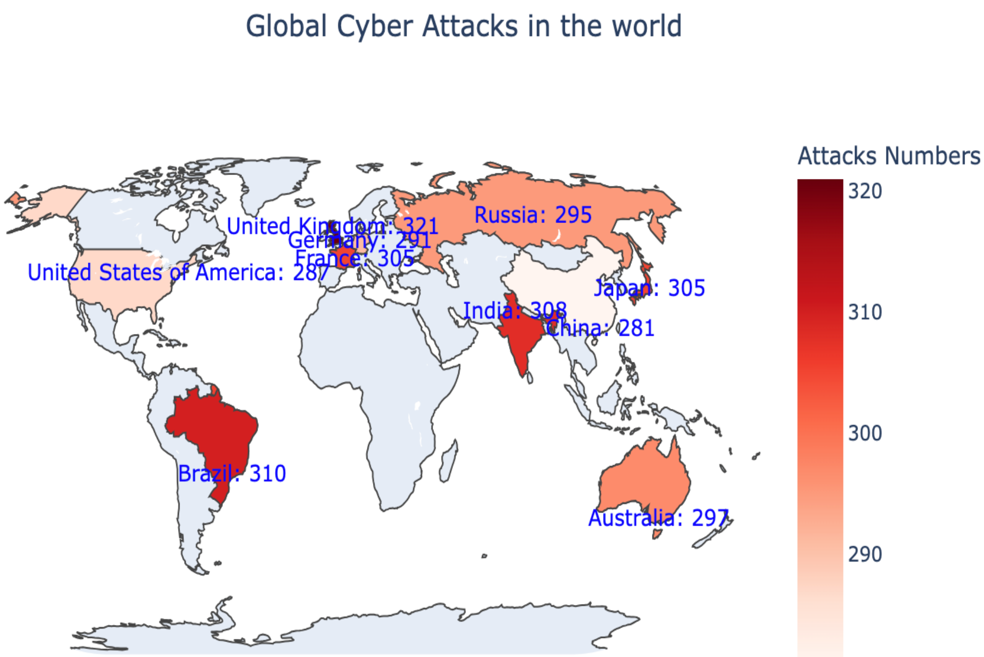


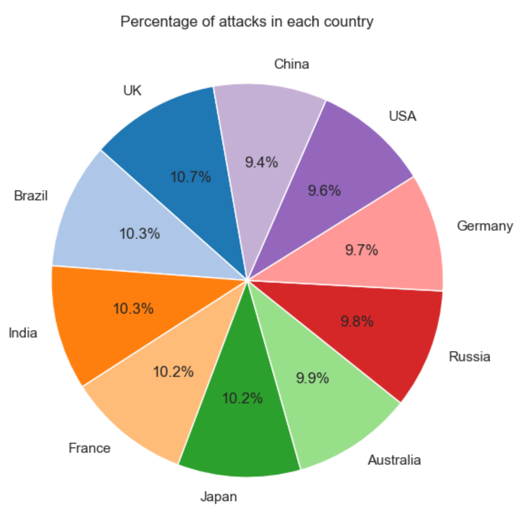
Insights:

- DDoS and Phishing most common methods.
- DDoS attacks often target organisations' networks, while Phishing exploits human vulnerabilities.
- The close distribution between attack types suggests a balance between long-standing attack strategies (e.g., Phishing, DDoS) and evolving tactics (e.g., Ransomware, SQL Injection).

Cyberattack Counts by Country:

Which countries are most frequently targeted by cyberattacks, and how are attacks distributed globally?



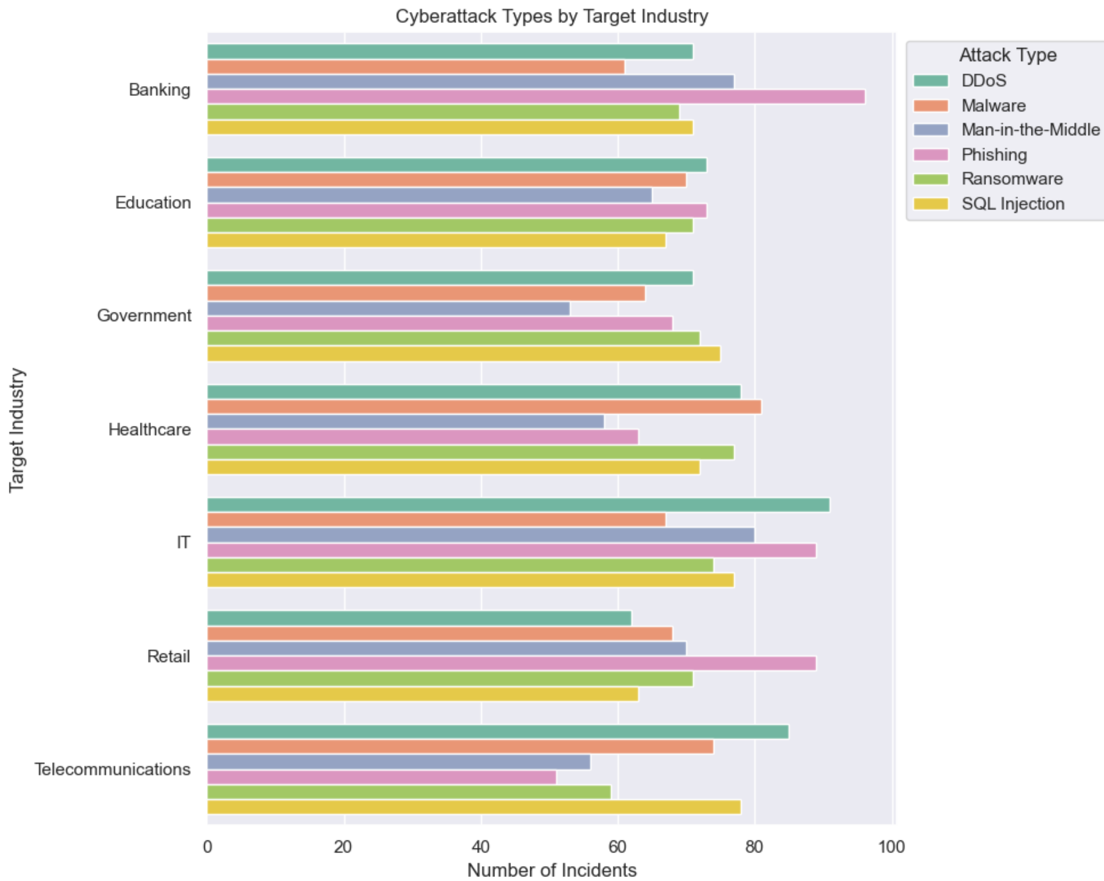


Countries such as Russia and China, which are frequently mentioned in cybersecurity reports for state-sponsored cyberattacks, also appear prominently in the data, reflecting ongoing geopolitical tensions and cyber espionage activities.

Strategic Implications: - Emerging markets like India and Brazil are becoming increasingly targeted as their digital and economic footprints grow.

Top Targeted Industries by Cyberattacks

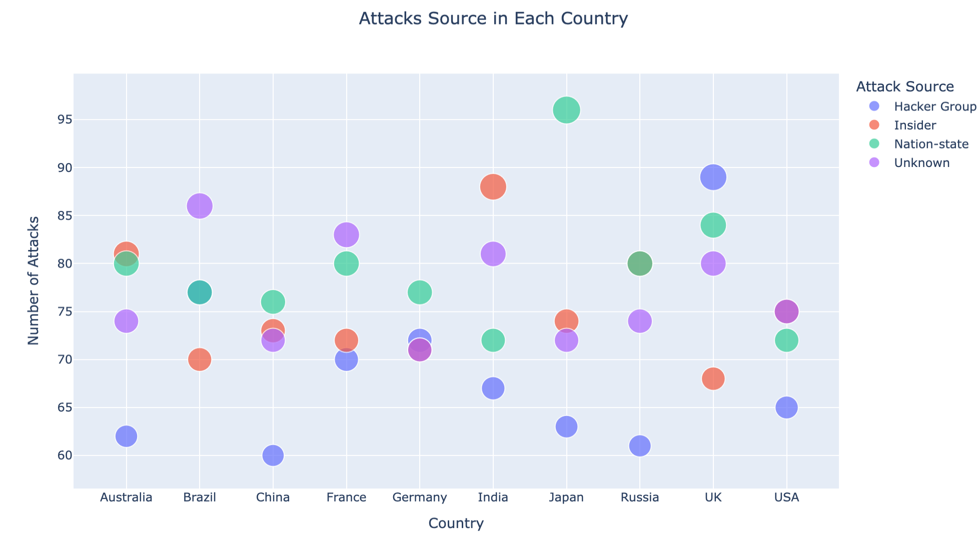
The visual shows which industries have been the most common targets of cyberattacks, by type.

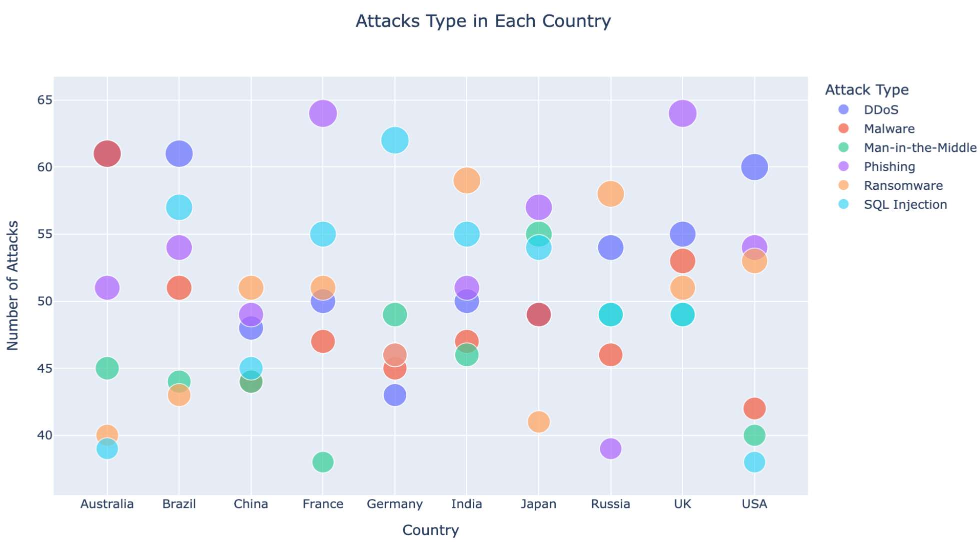


Insights:

- The IT and Banking industries are at the highest risk, which is expected due to the large amounts of sensitive data they manage, making them prime targets for attackers.
- The Retail and Telecommunications industries are also notably affected due to their reliance on consumer data
- Phishing is a Dominant Threat for Customer-Facing Industries
- DDoS Attacks Target Availability: DDoS attacks are the most frequent in Education, IT, and Telecommunications.
- SQL Injection Remains a Significant Threat in the Government Sector.
- Malware is a Key Concern for healthcare-sensitive patient data.

Attack Source in each country and the financial impacts:

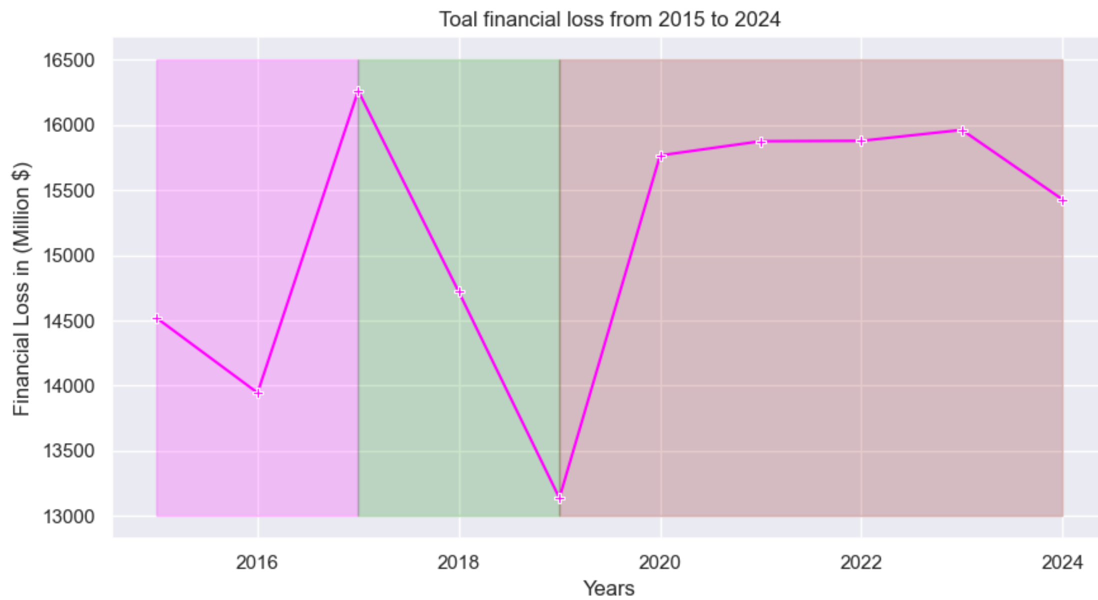




Insights:

- The US shows the highest frequency of Phishing attacks, highlighting its position as a prime target for credential theft and social engineering.
- China exhibits a spike in DDoS attacks, possibly linked to infrastructure disruption attempts or nation-state activity.
- India has a notable increase in Malware-based attacks, suggesting vulnerability in endpoint protection. Some attack sources show significant spikes in specific countries:
- Ransomware incidents surge in Germany, likely tied to attacks on industrial infrastructure or high-value targets.

Trend of financial losses over the past 10 years

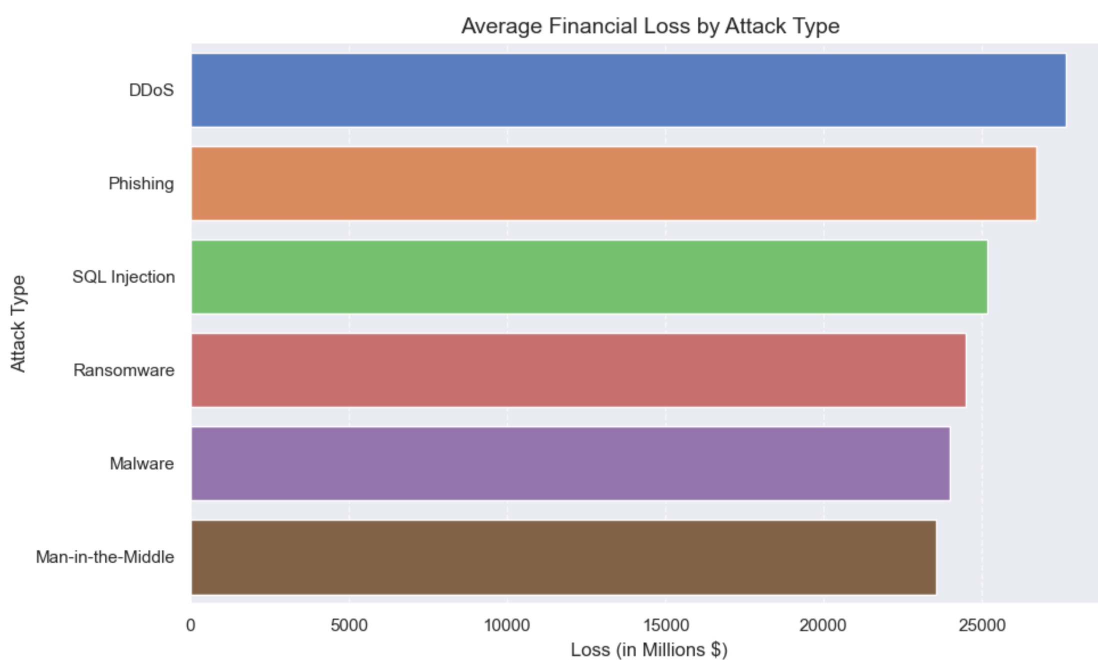


####

Insights:

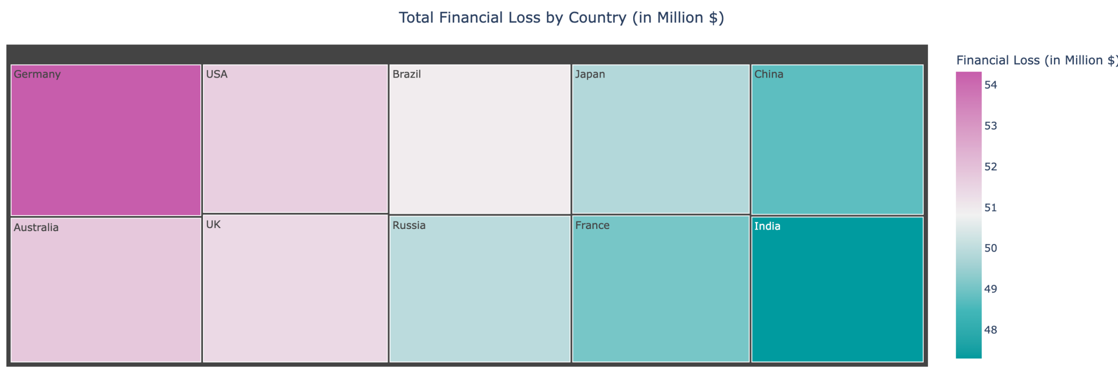
- 2017 has the highest financial loss in the last 10 years
- A drastic drop after 2 years, from 2017 to 2019
- After 2020, the losses follow a stable trend

The barplot presents the total financial loss attributed to each type of cyberattack:



Insights:

- DDoS is perceived as primarily disruptive, and the cost of mitigation over the years appears substantial.
- Phishing targets human vulnerabilities, underscoring the significant financial risks associated with social engineering.
- SQL Injection leads to data breaches, data corruption, and service disruption, all of which carry significant financial implications for affected organisations.



4. Feature Engineer

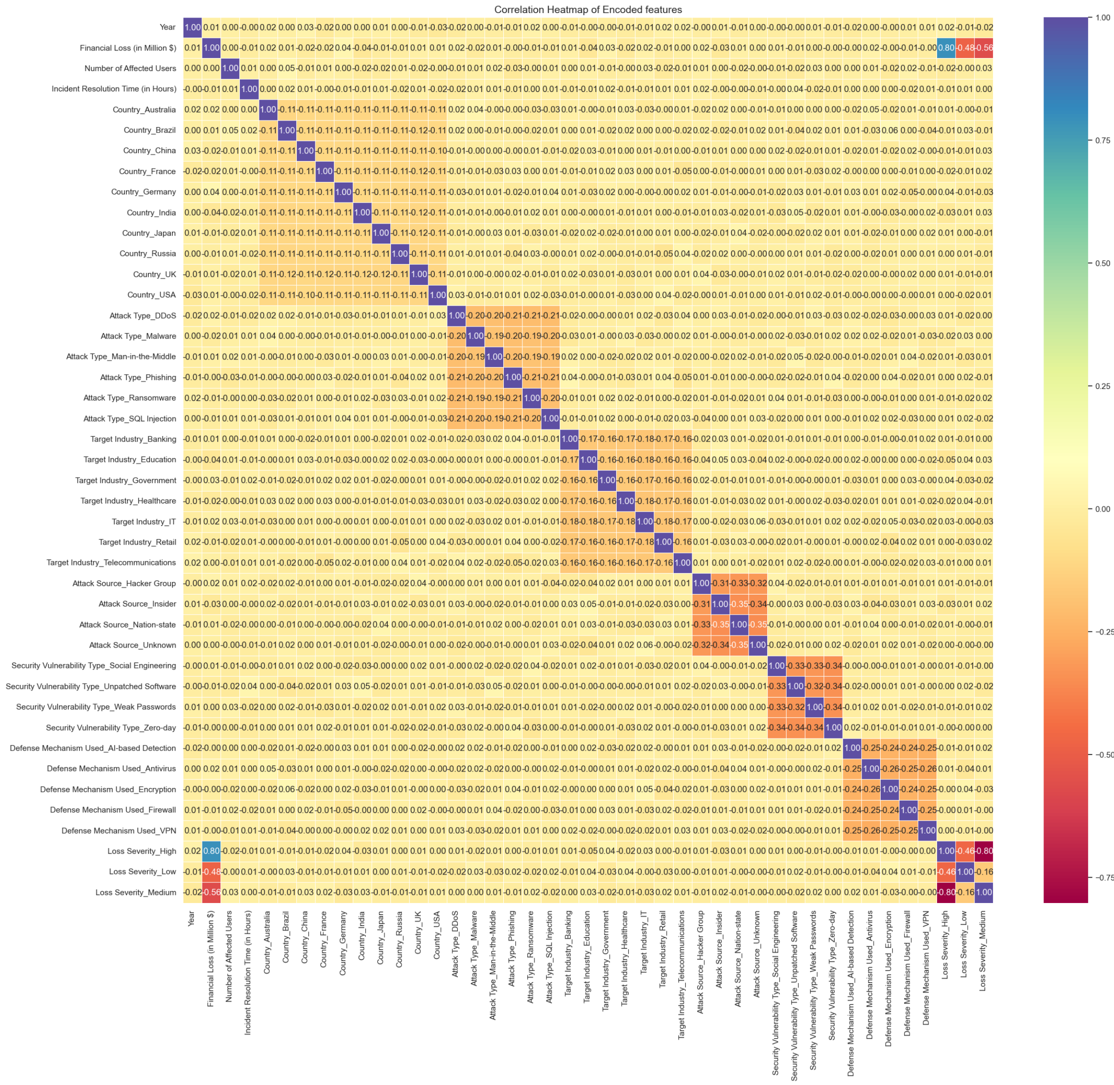
- The dataset has some columns that are categorical rather than numerical.
- The categorical values don’t show any kind of ‘order’, so AutoEncoder() is not the best option.
- Since we need to create our ML algorithms, OHE is the best option to avoid confusion.
- Created a new column that classifies the loss severity based on financial value

```python
# Classifying financial loss by severity
def loss_level(value):
    if value < 10:
        return "Low"
    elif value <= 30:
        return "Medium"
    else:
        return "High"

# Apply the function to create a new column
cyberAttacks["Loss Severity"] = cyberAttacks["Financial Loss (in Million $)"].apply(loss_level)

# Preview the changes
cyberAttacks.head(5)
```

Correlation Matrix with Hot Encoded Values:

In the Heatmap, only a few feature pairs go beyond 0.2, and just a handful cross 0.3, which are still considered weak correlations.

- Low Multicollinearity = Good for models like (Random Forest, XGBoost)
- Non-linear relationship
- We will apply these algorithms as ML models.

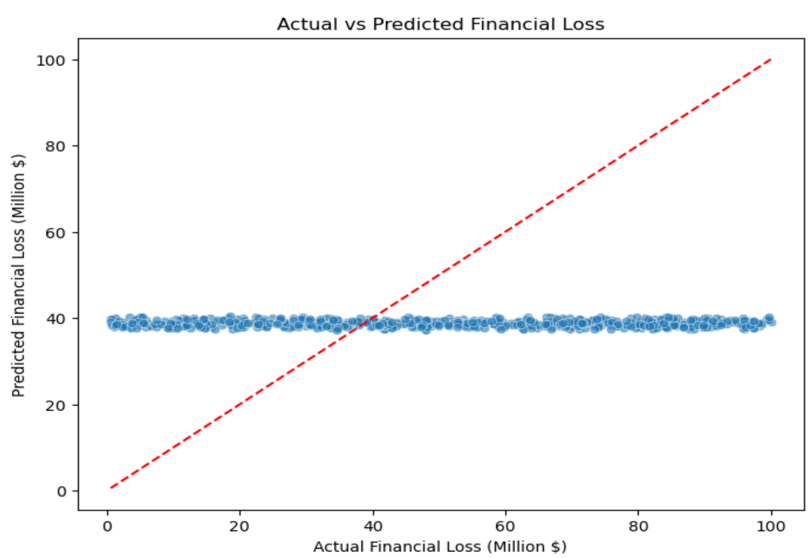

Correlation Heatmap of Encoded features

5. Machine Learning

ML Pipeline:

- Target Variable
  - Financial Loss (in Million $) is an excellent regression target.
- Numeric Features
  - Year
  - Number of Affected Users
  - Incident Resolution Time (in Hours)
- Categorical Features
  - Country
  - Attack Type
  - Target Industry
  - Attack Source
  - Security Vulnerability Type
  - Defense Mechanism Used

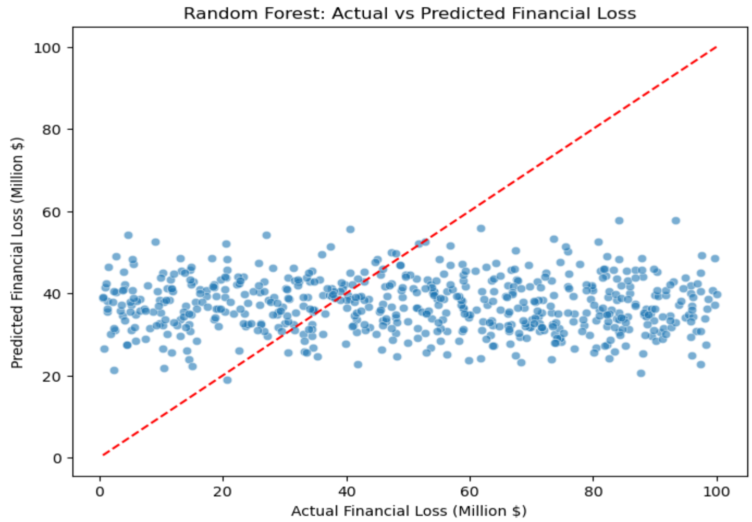
Simple Linear Regression




```
Mean Squared Error (original scale): 952.80
R^2 Score (original scale): -0.18
```

- **R^2 is near negative, which means linear regression doesn't explain the variance well.**

Trying Random Forest




```
Random Forest MSE: 1050.92
Random Forest R² Score: -0.30
```

The random forest model is performing poorly. It’s worse than just predicting the mean of the target every time.

Trying Classification:

- The financial loss feature performed well in both linear and Random Forest regression.
- Trying these features Attack Type”, “target industry”, “Security Vulnerability Type”, “Defence Mechanism used” for classification.
- The target is the Loss Type, which is an engineered feature that shows the size of financial losses.

```
Classification Report:
              precision    recall  f1-score   support

        High       0.33      0.34      0.34       203
         Low       0.33      0.34      0.33       194
      Medium       0.32      0.31      0.31       203

    accuracy                           0.33       600
   macro avg       0.33      0.33      0.33       600
weighted avg       0.33      0.33      0.33       600
```

The classification is also performing poorly, meaning the model has no predictive power with the selected features.

- The four chosen categorical features may not be strongly correlated with the size of financial loss.

Trying multi-class classification

- Balancing both numerical and categorical features, such as:
    - Number of Affected Users
    - Incident Resolution Time (in Hours)

```
# Define features and target
features = [
    "Target Industry",
    "Security Vulnerability Type",
    "Defense Mechanism Used",
    "Number of Affected Users",
    "Incident Resolution Time (in Hours)"
]
X = cyberAttacks3[features]
y = cyberAttacks3["Attack Type"]
```

The target variable is 'Attack Type'

```
Multi-Classification Report:
                   precision    recall  f1-score   support

             DDoS       0.17      0.17      0.17       111
          Malware       0.15      0.15      0.15        97
Man-in-the-Middle       0.18      0.15      0.16        99
         Phishing       0.13      0.15      0.14       103
       Ransomware       0.12      0.15      0.13        79
    SQL Injection       0.14      0.11      0.12       111

         accuracy                           0.15       600
        macro avg       0.15      0.15      0.15       600
     weighted avg       0.15      0.15      0.15       600
```

Accuracy of 15% is very low, even with 6 attack classes.

- **No model is finding meaningful predictive patterns in the data.**

7. Conclusion:

After running Linear regression, Random forest and classification, in all the cases, the results are really low.

Maybe the dataset has no correlation between the variables. Cyberattack impact and types can be influenced by many unobserved external factors (e.g., company size, response team skill, attacker motivation).