Cyber Attack Trends and Risk Analysis

Project Introduction

Cybersecurity incidents have become a critical global concern, impacting industries such as education, retail, IT, and telecommunications with severe financial losses, large-scale user data breaches, and prolonged recovery times. These incidents vary across multiple dimensions, including attack type, target industry, financial loss, number of affected users, attack source, exploited vulnerabilities, defense mechanisms, and resolution time. Analyzing these factors through Python, Pandas, NumPy, probability, and statistical methods can uncover meaningful patterns, highlight risk factors, and test significant relationships between variables. Such insights not only help understand the evolving nature of cyber threats but also provide data-driven guidance for strengthening defense strategies and minimizing organizational impact.

```
import numpy as np
import pandas as pd
```

```
df=pd.read_csv('/content/Global_Cybersecurity_Threats_2015-2024.csv')
```

```
df.head()
```

	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	80.53	773169	Hacker Group	Unpatched Software	VPN	63
1	China	2019	Ransomware	Retail	62.19	295961	Hacker Group	Unpatched Software	Firewall	71
2	India	2017	Man-in-the-Middle	IT	38.65	605895	Hacker Group	Weak Passwords	VPN	20

Next steps: Generate code with df | New interactive sheet

this command gave me the first 5 rows of dataset

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Country                             3000 non-null   object
 1   Year                                3000 non-null   int64
 2   Attack Type                         3000 non-null   object
 3   Target Industry                     3000 non-null   object
 4   Financial Loss (in Million $)       3000 non-null   float64
 5   Number of Affected Users            3000 non-null   int64
 6   Attack Source                       3000 non-null   object
 7   Security Vulnerability Type         3000 non-null   object
 8   Defense Mechanism Used              3000 non-null   object
 9   Incident Resolution Time (in Hours)  3000 non-null   int64
dtypes: float64(1), int64(3), object(6)
memory usage: 234.5+ KB
```

So, we have to convert columns like Attack_Type , Target_Industry, Attack Source, Security_vulnerability_Type, Defnce_mechanism_used into categories

```
cat_col=['Attack Type','Target Industry', 'Attack Source', 'Security Vulnerability Type', 'Defense Mechanism U

df[cat_col]=df[cat_col].astype('category')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column                              Non-Null Count  Dtype
```

```
---  ------                                ---------------  -----
 0   Country                               3000 non-null   object
 1   Year                                  3000 non-null   int64
 2   Attack Type                           3000 non-null   category
 3   Target Industry                       3000 non-null   category
 4   Financial Loss (in Million $)         3000 non-null   float64
 5   Number of Affected Users              3000 non-null   int64
 6   Attack Source                         3000 non-null   category
 7   Security Vulnerability Type           3000 non-null   category
 8   Defense Mechanism Used                3000 non-null   category
 9   Incident Resolution Time (in Hours)   3000 non-null   int64
dtypes: category(5), float64(1), int64(3), object(1)
memory usage: 133.1+ KB
```

So,we have converted all the required elements into categories

```
df.describe()
```

	Year	Financial Loss (in Million $)	Number of Affected Users	Incident Resolution Time (in Hours)
count	3000.000000	3000.000000	3000.000000	3000.000000
mean	2019.570333	50.492970	504684.136333	36.476000
std	2.857932	28.791415	289944.084972	20.570768
min	2015.000000	0.500000	424.000000	1.000000
25%	2017.000000	25.757500	255805.250000	19.000000
50%	2020.000000	50.795000	504513.000000	37.000000
75%	2022.000000	75.630000	758088.500000	55.000000
max	2024.000000	99.990000	999635.000000	72.000000

These are some statistical operations performed on numerical elements

```
df.isnull().sum()
```

	0
Country	0
Year	0
Attack Type	0
Target Industry	0
Financial Loss (in Million $)	0
Number of Affected Users	0
Attack Source	0
Security Vulnerability Type	0
Defense Mechanism Used	0
Incident Resolution Time (in Hours)	0

dtype: int64

There is no null value in our data

```
df.shape
```

```
(3000, 10)
```

our data contains 3000 rows and 10 columns

Descriptive & Exploratory Analysis (Pandas, NumPy)

1] The average and median financial loss across all countries

```
avg_loss=df['Financial Loss (in Million $)'].mean()
print(avg_loss)
```

```
50.49297
```

```
median_loss=df['Financial Loss (in Million $)'].quantile(0.50)
print(median_loss)
```

```
50.795
```

So, the average and median financial loss are almost same i.e. 50.49297 and 50.795 respectively

---

2] The attack type which caused the highest average loss

```
df.groupby('Attack Type')['Financial Loss (in Million $)'].mean()
```

```
/tmp/ipython-input-3644180547.py:1: FutureWarning: The default of observed=False is deprecated and will be changed to True in a
  df.groupby('Attack Type')['Financial Loss (in Million $)'].mean()
```

Attack Type	Financial Loss (in Million $)
DDoS	52.035631
Malware	49.418454
Man-in-the-Middle	51.309085
Phishing	50.459905
Ransomware	49.653793
SQL Injection	50.013042

dtype: float64

The variation is small (49–52M), meaning all attack types are financially damaging — but DDoS and MITM stand out as the most severe per incident.

3] The industry which is mostly targeted

```
df['Target Industry'].value_counts()
```

Target Industry	count
IT	478
Banking	445
Healthcare	429
Retail	423
Education	419
Government	403
Telecommunications	403

dtype: int64

1. IT sector is the most targeted with 478 incidents.
2. Banking (445) and Healthcare (429) are also highly targeted.
3. Retail (423) and Education (419)
4. Government (403) and Telecommunications (403) show equal incident counts.

4] The country which has the highest number of affected users overall

```
df.groupby('Country')['Number of Affected Users'].sum()
```

Country	Number of Affected Users
Australia	150011830
Brazil	168806980
China	139580938
France	156229142
Germany	147675358
India	149178659
Japan	148711814
Russia	152191835
UK	157464983
USA	144200870

dtype: int64

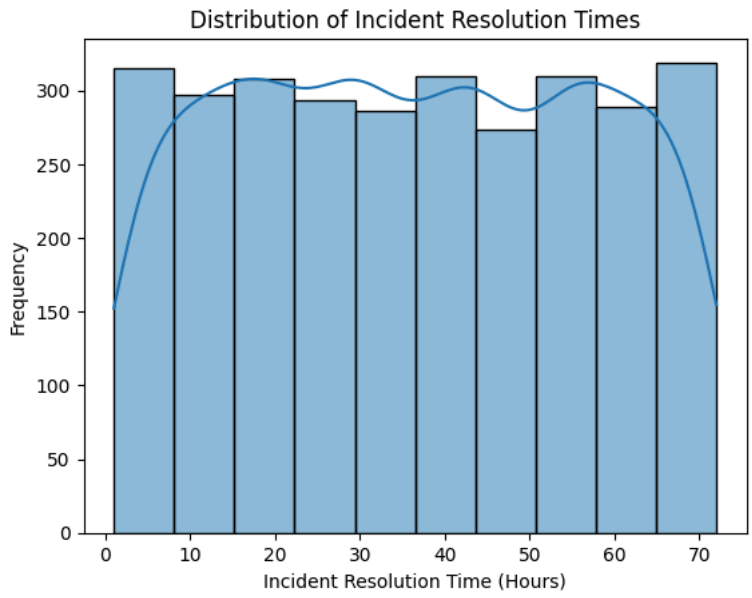
1. Brazil leads with the highest number of affected users (168.8M).
2. UK (157.5M), France (156.2M), and Russia (152.2M) also show very high numbers of compromised users.
3. India (149.2M), Japan (148.7M), Germany (147.7M), and Australia (150M) are all in a similar range.
4. USA (144.2M) surprisingly reports fewer affected users than Brazil, UK, or France

---

5] The distribution of the incedent resolution times

```
import matplotlib.pyplot as plt
import seaborn as sns

sns.histplot(df["Incident Resolution Time (in Hours)"], bins=10, kde=True)
plt.xlabel("Incident Resolution Time (Hours)")
plt.ylabel("Frequency")
plt.title("Distribution of Incident Resolution Times")
plt.show()
```



Comparative Analysis

1] To compare financial loss between industries

```
df.groupby('Target Industry')['Financial Loss (in Million $)'].sum()
```

```
/tmp/ipython-input-4246033073.py:1: FutureWarning: The default of observed=False is deprecated and will be changed to True in a
  df.groupby('Target Industry')['Financial Loss (in Million $)'].sum()
```

Target Industry	Financial Loss (in Million $)
Banking	22772.39
Education	20071.43
Government	21205.33
Healthcare	21041.29
IT	24809.83
Retail	21119.55
Telecommunications	20459.09

dtype: float64

IT has the highest loss ($24.8B) : most targeted industry.

Banking is second ($22.7B) : highly lucrative for attackers.

Government, Retail, Healthcare all face similar losses (~$21B).

Education has the lowest loss ($20B) but still significant.

Overall: Critical sectors (IT, Banking, Government) bear the biggest financial impact.

2] Comparison of Attack sources on the basis of Incident resolution time

```
# Average resolution time by attack source
df.groupby("Attack Source")["Incident Resolution Time (in Hours)"].mean()
```

```
/tmp/ipython-input-994679777.py:2: FutureWarning: The default of observed=False is deprecated and will be changed to True in a
  df.groupby("Attack Source")["Incident Resolution Time (in Hours)"].mean()
```

Attack Source	Incident Resolution Time (in Hours)
Hacker Group	37.212828
Insider	36.351064
Nation-state	36.329975
Unknown	36.091146

dtype: float64

Nation-state attacks have the fastest resolution (~36.33 hrs).

Unknown sources are close behind (~36.09 hrs)

Hacker Group incidents take slightly longer (~37.21 hrs).

Insider attacks are also relatively slower (~36.35 hrs), often harder to detect internally.

3] We have to check if there is any relation between attack source and vulnerability type

H0:There is no relation

H1:There is relation

```
from scipy.stats import chi2_contingency

crosstab = pd.crosstab(df["Attack Source"], df["Security Vulnerability Type"])

chi2, p, dof, expected = chi2_contingency(crosstab)

print("Chi-square Statistic:", chi2)
print("p-value:", p)
```

```
print("Degrees of Freedom:", dof)
```

```
Chi-square Statistic: 8.219881629569208
p-value: 0.5121484864912613
Degrees of Freedom: 9
```

We have applied chi-square test here because both are categorical.

Here,

our p-value is >0.05 significance level , so we fail to reject the H0

there is no significance relation between Attack Source and Vulnerability type

---

Probability and Stats

1] The probability that a randomly selected incident involves

1. Ransomware
2. Education sector
3. A vulnerability due to weak passwords

So, this is the joint probability example

joint probability = (No. of rows satisfying all 3)/Total rows

```
total= len(df)

joint_prob=len(df[(df["Attack Type"]=="Ransomware") &
                  (df["Target Industry"]=="Education") &
                  (df["Security Vulnerability Type"]=="Weak Passwords")])/total
round(joint_prob,4)
```

```
0.0053
```

so the probability is 0.0053

2] Compute the conditional probability:

P(Phishing | Education industry)

so,the conditional probability is = len(phishing and education industry)/len(education industry)

```
edu_count = len(df[df["Target Industry"]=="Education"])

phishing_and_edu = len(df[(df["Attack Type"]=="Phishing") &
                          (df["Target Industry"]=="Education")])


p_phishing_given_edu = phishing_and_edu / edu_count

p_phishing_given_edu
```

```
0.17422434367541767
```

So,the probability of Attack Type is Phishing given that Target Industry is Education is 0.1742

3] Compute the conditional probability :

P(Ransomware| Nation-state attack) =

len (Ransomware and Nation-state attack)/len(Nation-state attack)

```
nation_state_count = len(df[df["Attack Source"]=="Nation-state"])

ransomware_and_ns = len(df[(df["Attack Type"]=="Ransomware") &
                           (df["Attack Source"]=="Nation-state")])
```

```
p_ransomware_given_ns = ransomware_and_ns / nation_state_count

p_ransomware_given_ns
```

```
0.15491183879093198
```

So,conditional probability is 0.1549

Hypothesis Testing

1] Whether there is significant difference in financial losses across industries

Null Hypothesis (H_0): Mean financial losses are the same across all industries.

Alternative Hypothesis (H_1): At least one industry has a different mean financial loss.

```
from scipy.stats import f_oneway
f_stat, p_val = f_oneway(*df.groupby("Target Industry")['Financial Loss (in Million $)'].apply(list))

print("F-statistic:", f_stat)
print("p-value:", p_val)
```

```
F-statistic: 1.378370725938995
p-value: 0.21937002783005477
/tmp/ipython-input-1653631015.py:2: FutureWarning: The default of observed=False is deprecated and will be changed to True in a
  f_stat, p_val = f_oneway(*df.groupby("Target Industry")['Financial Loss (in Million $)'].apply(list))
```

So,here p-value is greater than 0.05 significance level, so we fail to reject H0

and can conclude that Mean financial losses are the same across all industries.

2] Do different attack types result in significantly different numbers of affected users?

Null Hypothesis (H_0): The mean number of affected users is the same across all attack types.

Alternative Hypothesis (H_1): At least one attack type affects a significantly different number of users.

```
f_stat, p_val = f_oneway(*df.groupby("Attack Type")["Number of Affected Users"].apply(list))

print("F-statistic:", f_stat)
print("p-value:", p_val)
```

```
F-statistic: 0.7743625189209407
p-value: 0.5680979959487062
/tmp/ipython-input-567711301.py:1: FutureWarning: The default of observed=False is deprecated and will be changed to True in a
  f_stat, p_val = f_oneway(*df.groupby("Attack Type")["Number of Affected Users"].apply(list))
```

The ANOVA test shows no significant difference in the average number of affected users across different attack types (p = 0.5681). This suggests that the scale of impact in terms of affected users is not strongly dependent on the type of attack.

3] Are attack sources independent of security vulnerability?

Null Hypothesis (H_0): Attack source and vulnerability type are independent (no relationship).

Alternative Hypothesis (H_1): Attack source and vulnerability type are dependent (there is a relationship).

```
#Attack Source  Security Vulnerability Type
from scipy.stats import chi2_contingency

contingency=pd.crosstab(df['Attack Source'],df['Security Vulnerability Type'])
contingency

stats,pvalue,dof,x=chi2_contingency(contingency)

print('stats:',stats)
print('pvalue:',pvalue)
print('dof:',dof)
```

```
stats: 8.219881629569208
pvalue: 0.5121484864912613
dof: 9
```

So here pvalue>0.05 significance level so we fail to reject the null hypothesis,

Attack source and vulnerability type are independent (no relationship).

Correlation

1] correlation between financial loss and number of affected users

```
corr = df["Financial Loss (in Million $)"].corr(df["Number of Affected Users"])
print("Correlation:", corr)
```

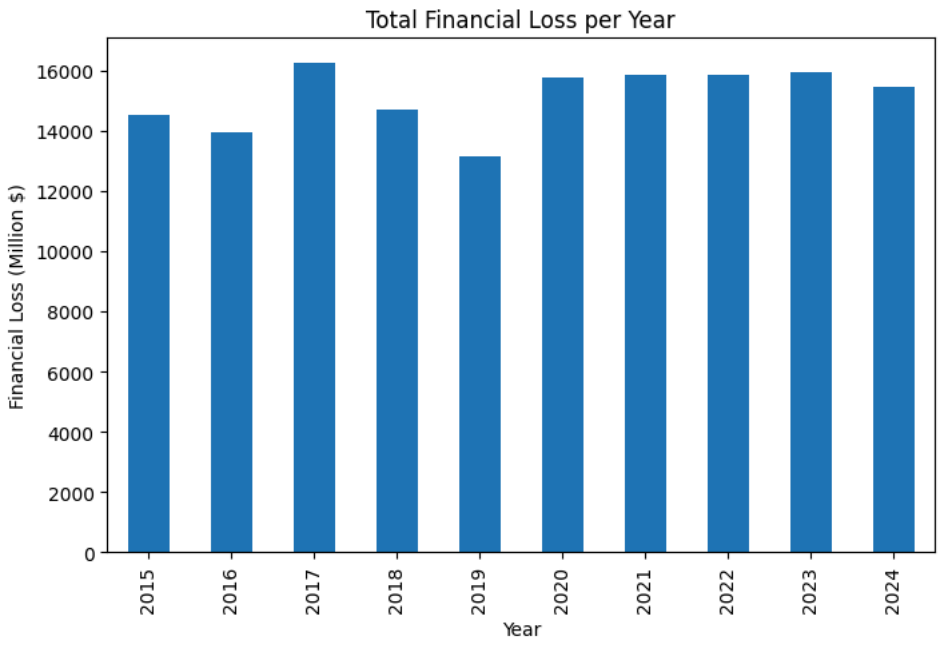
```
Correlation: 0.0017867355720258393
```

Value very close to 0 :means there is almost no linear relationship between Financial Loss (in Million $) and Number of Affected Users.

Visualizations

1] Bar plot: Total financial loss per year

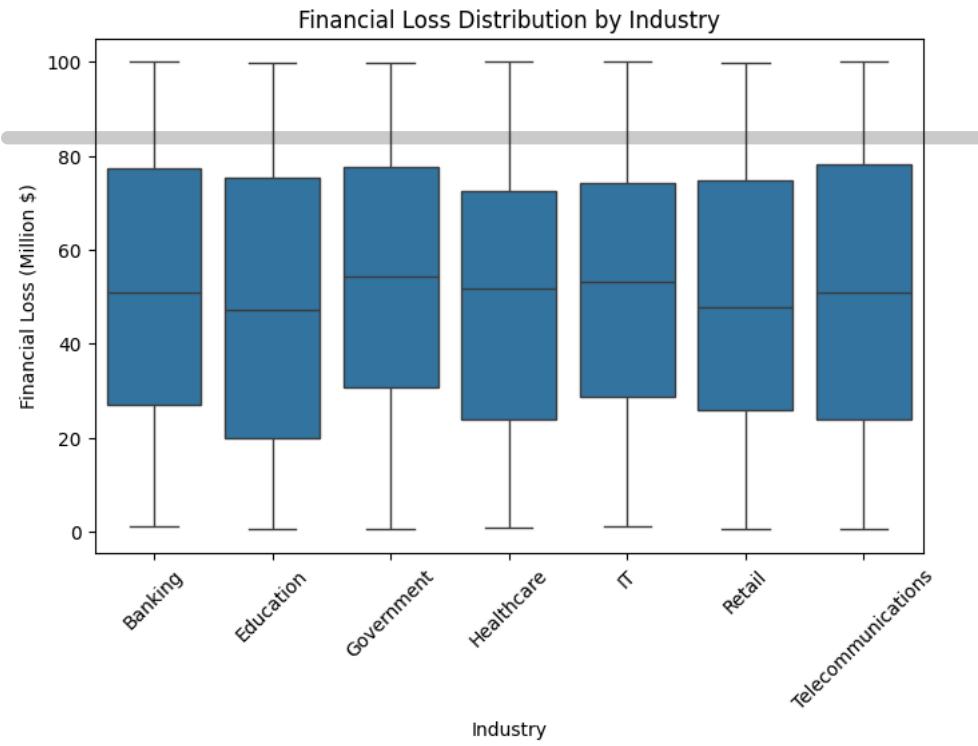
```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8,5))
yearly_loss = df.groupby("Year")["Financial Loss (in Million $)"].sum()
yearly_loss.plot(kind="bar")
plt.title("Total Financial Loss per Year")
plt.ylabel("Financial Loss (Million $)")
plt.xlabel("Year")
plt.show()
```



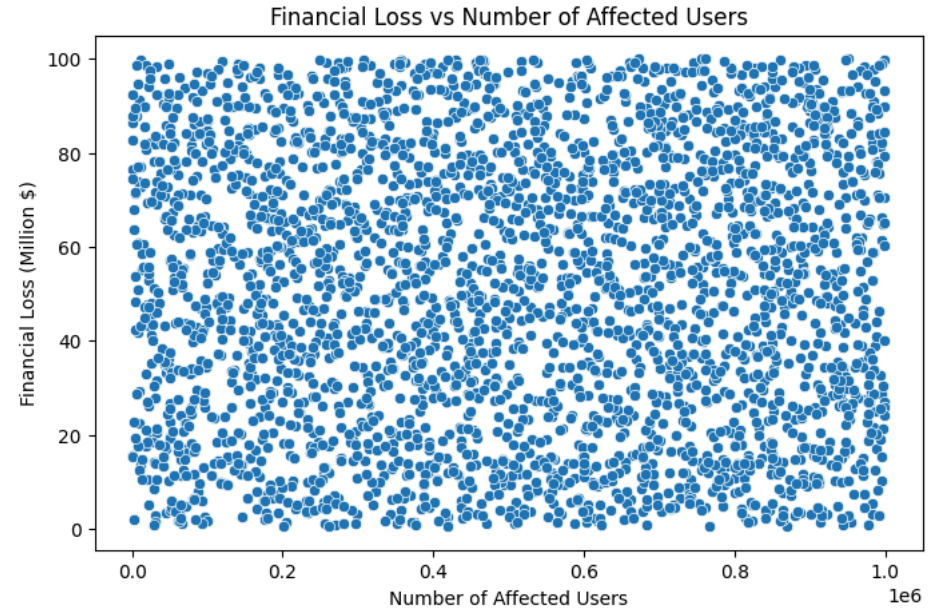
2] Box plot: Distribution of financial loss across industries

```
plt.figure(figsize=(8,5))
sns.boxplot(x="Target Industry", y="Financial Loss (in Million $)", data=df)
plt.title("Financial Loss Distribution by Industry")
plt.ylabel("Financial Loss (Million $)")
plt.xlabel("Industry")
```

```
plt.xticks(rotation=45)
```




```
plt.figure(figsize=(8,5))
sns.scatterplot(x="Number of Affected Users", y="Financial Loss (in Million $)", data=df)
plt.title("Financial Loss vs Number of Affected Users")
plt.xlabel("Number of Affected Users")
plt.ylabel("Financial Loss (Million $)")
plt.show()
```



Key Insights

1] Consistent Financial Impact Across Threats and Sectors: The analysis reveals that financial losses are consistently high regardless of the specific attack type or the targeted industry. Hypothesis testing showed no statistically significant difference in mean financial losses across industries or in the number of affected users across different attack types. This suggests that the financial risk from cyber attacks is a pervasive and uniform threat.

2] High-Value Industries are Prime Targets: The Information Technology (IT) and Banking sectors are the most frequently targeted industries, consequently suffering the highest cumulative financial losses, amounting to approximately 24.8 & 22.7 billion, respectively.

3] Attack Source and Vulnerability are Independent: A Chi-square test indicated that there is no statistically significant relationship between the source of an attack (e.g., Hacker Group, Nation-state) and the type of security vulnerability exploited. This implies that any attacker group is likely to exploit any available vulnerability opportunistically.

4] Financial Loss is Decoupled from the Number of Affected Users: The analysis found almost no linear correlation (correlation coefficient of 0.0017) between the financial loss incurred in an incident and the number of users whose data was compromised. This highlights that direct financial theft and large-scale data breaches are separate risk dimensions.

5] Attack Complexity Does Not Drastically Alter Resolution Time: The average incident resolution time shows minimal variation across different attack sources, with all sources (Hacker Group, Insider, Nation-state, Unknown) having resolution times between 36 and 37.5 hours

Actionable Recommendations

1] Adopt a Universal Security Posture: Given that financial damages are severe and consistent across all industries and attack types, organizations in every sector (including Education, Retail, and Healthcare) should implement robust, foundational cybersecurity frameworks. Over-investing in defenses for one specific attack type (e.g., DDoS) while neglecting others is not a data-supported strategy.

2] Enhance Defenses for High-Risk Sectors: Organizations in IT and Banking must recognize their status as primary targets and invest in advanced, layered security solutions. This should include proactive threat intelligence, continuous security monitoring, and specialized defenses to protect high-value assets.

3] Implement Comprehensive Vulnerability Management: Since attack sources do not specialize in specific vulnerabilities, organizations must develop a holistic vulnerability management program. This strategy should prioritize patching known software

vulnerabilities, enforcing strong password policies, and securing configurations, as any weakness can be a potential entry point for any attacker.

4] Develop Dual-Pronged Incident Response Plans: The lack of correlation between financial loss and the number of affected users necessitates separate risk management strategies. Organizations should create distinct response playbooks for:

Financial Containment: Focused on business continuity, isolating financial systems, and mitigating direct monetary loss.

Data Breach Management: Focused on user notification, public relations, regulatory compliance, and providing support to affected individuals.

5] Focus on Foundational Defense Mechanisms: The analysis shows that common vulnerabilities like unpatched software and weak passwords are consistently exploited. Therefore, prioritizing the deployment and consistent maintenance of fundamental defense mechanisms like Firewalls and VPNs is a critical and effective strategy for risk reduction.

```
Start coding or generate with AI.
```