Importing necessary libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading dataset

In [2]:
```
data_breach_df = pd.read_csv("/Users/srinivaspallapu/Documents/INFO 57
data_breach_df
```

Out[2]:

	Name of Covered Entity	State	Covered Entity Type	Individuals Affected	Breach Submission Date	Type of Breach	Location of Breached Information	B As
0	Wichita Urology Group	KS	Healthcare Provider	5314	12/07/2023	Hacking/IT Incident	Network Server	
1	Pan-American Life Insurance Group, Inc.	LA	Business Associate	94807	12/04/2023	Hacking/IT Incident	Network Server	
2	Leggett & Platt Incorporated Employee Benefit ...	MO	Health Plan	1200	12/04/2023	Hacking/IT Incident	Network Server	
3	Pan-American Life Insurance Group, Inc.	LA	Health Plan	105387	12/04/2023	Hacking/IT Incident	Network Server	
4	EMS Management and Consultants Inc.	NC	Business Associate	2654	12/01/2023	Unauthorized Access/Disclosure	Paper/Films	
...	...	...	...	...	...	...	...	
864	Daniel J. Edelman Holdings, Inc.	IL	Health Plan	184500	12/22/2021	Hacking/IT Incident	Network Server	
865	Rhode Island Public Transit Authority	RI	Health Plan	5015	12/21/2021	Hacking/IT Incident	Network Server	
866	Chaddock	IL	Healthcare Provider	777	12/21/2021	Theft	Paper/Films	
867	Northwest Broward Orthopaedics Associates	FL	Healthcare Provider	500	12/17/2021	Hacking/IT Incident	Desktop Computer, Network Server	
868	Youth Consultation Service	NJ	Healthcare Provider	2756	07/19/2021	Hacking/IT Incident	Network Server	

869 rows × 9 columns

In [3]:
```
# Display the first few rows of the DataFrame
print(data_breach_df.head(5))
```

```
                                Name of Covered Entity State  \
0                               Wichita Urology Group     KS
1           Pan-American Life Insurance Group, Inc.    LA
2  Leggett & Platt Incorporated Employee Benefit ...    MO
3           Pan-American Life Insurance Group, Inc.    LA
4               EMS Management and Consultants Inc.    NC

   Covered Entity Type  Individuals Affected Breach Submission Date
\
0  Healthcare Provider                  5314             12/07/2023
1   Business Associate                 94807             12/04/2023
2          Health Plan                  1200             12/04/2023
3          Health Plan                105387             12/04/2023
4   Business Associate                  2654             12/01/2023

                    Type of Breach Location of Breached Information  \
0             Hacking/IT Incident                   Network Server
1             Hacking/IT Incident                   Network Server
2             Hacking/IT Incident                   Network Server
3             Hacking/IT Incident                   Network Server
4  Unauthorized Access/Disclosure                      Paper/Films

  Business Associate Present  Web Description
0                        Yes              NaN
1                        Yes              NaN
2                        Yes              NaN
3                         No              NaN
4                        Yes              NaN
```

Info about data types and missing values

In [4]:
```
print("\nInfo about data types and missing values:")
print(data_breach_df.info())
```

```
Info about data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 869 entries, 0 to 868
Data columns (total 9 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   Name of Covered Entity            869 non-null    object 
 1   State                             865 non-null    object 
 2   Covered Entity Type               868 non-null    object 
 3   Individuals Affected              869 non-null    int64  
 4   Breach Submission Date            869 non-null    object 
 5   Type of Breach                    869 non-null    object 
 6   Location of Breached Information  869 non-null    object 
 7   Business Associate Present        869 non-null    object 
 8   Web Description                   0 non-null      float64
dtypes: float64(1), int64(1), object(7)
memory usage: 61.2+ KB
None
```

Data types of the columns

In [5]:
```
data_breach_df.dtypes
```

Out[5]:
```
Name of Covered Entity              object
State                               object
Covered Entity Type                 object
Individuals Affected                 int64
Breach Submission Date              object
Type of Breach                      object
Location of Breached Information    object
Business Associate Present          object
Web Description                    float64
dtype: object
```

Number of missing values

In [6]:
```
print(data_breach_df.isnull().sum())
```

```
Name of Covered Entity                0
State                                 4
Covered Entity Type                   1
Individuals Affected                  0
Breach Submission Date                0
Type of Breach                        0
Location of Breached Information      0
Business Associate Present            0
Web Description                     869
dtype: int64
```

Removing "Web Description" since the whole columns is empty

In [7]:
```
data_breach_df = data_breach_df.drop(columns=['Web Description'])
data_breach_df
```

Out[7]:

	Name of Covered Entity	State	Covered Entity Type	Individuals Affected	Breach Submission Date	Type of Breach	Location of Breached Information	B As
0	Wichita Urology Group	KS	Healthcare Provider	5314	12/07/2023	Hacking/IT Incident	Network Server	
1	Pan-American Life Insurance Group, Inc.	LA	Business Associate	94807	12/04/2023	Hacking/IT Incident	Network Server	
2	Leggett & Platt Incorporated Employee Benefit ...	MO	Health Plan	1200	12/04/2023	Hacking/IT Incident	Network Server	
3	Pan-American Life Insurance Group, Inc.	LA	Health Plan	105387	12/04/2023	Hacking/IT Incident	Network Server	
4	EMS Management and Consultants Inc.	NC	Business Associate	2654	12/01/2023	Unauthorized Access/Disclosure	Paper/Films	
...	...	...	...	...	...	...	...	
864	Daniel J. Edelman Holdings, Inc.	IL	Health Plan	184500	12/22/2021	Hacking/IT Incident	Network Server	
865	Rhode Island Public Transit Authority	RI	Health Plan	5015	12/21/2021	Hacking/IT Incident	Network Server	
866	Chaddock	IL	Healthcare Provider	777	12/21/2021	Theft	Paper/Films	
867	Northwest Broward Orthopaedics Associates	FL	Healthcare Provider	500	12/17/2021	Hacking/IT Incident	Desktop Computer, Network Server	
868	Youth Consultation Service	NJ	Healthcare Provider	2756	07/19/2021	Hacking/IT Incident	Network Server	

869 rows × 8 columns

In [8]:
```
data_breach_df['State'].fillna('Unknown', inplace=True)
data_breach_df['Covered Entity Type'].fillna('Unknown', inplace=True)
```

In [9]:
```
print(data_breach_df.isnull().sum())
```

```
Name of Covered Entity              0
State                               0
Covered Entity Type                 0
Individuals Affected                0
Breach Submission Date              0
Type of Breach                      0
Location of Breached Information    0
Business Associate Present          0
dtype: int64
```

In [10]:
```
data_breach_df.describe()
```

Out[10]:

	Individuals Affected
count	8.690000e+02
mean	1.697074e+05
std	7.851357e+05
min	5.000000e+02
25%	1.352000e+03
50%	5.973000e+03
75%	4.307100e+04
max	1.127000e+07

In [11]:
```
data_breach_df['State'].value_counts()
```

Out[11]:

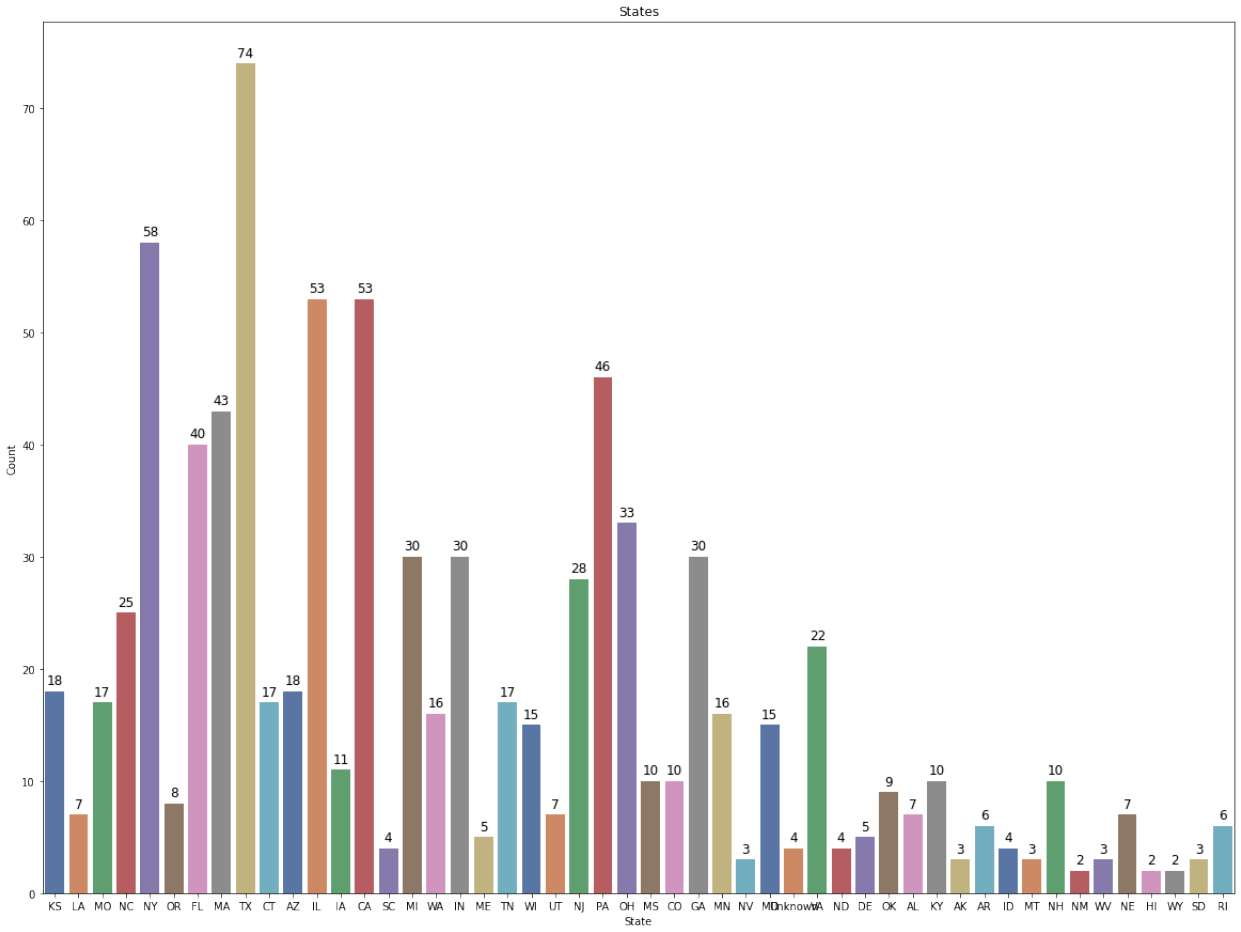
```
TX    74
NY    58
CA    53
IL    53
PA    46
MA    43
FL    40
OH    33
MI    30
IN    30
GA    30
NJ    28
NC    25
VA    22
KS    18
AZ    18
```

```
TN         17
CT         17
MO         17
WA         16
MN         16
WI         15
MD         15
IA         11
KY         10
NH         10
MS         10
CO         10
OK          9
OR          8
NE          7
UT          7
LA          7
AL          7
AR          6
RI          6
DE          5
ME          5
SC          4
ND          4
ID          4
Unknown     4
MT          3
WV          3
NV          3
SD          3
AK          3
NM          2
HI          2
WY          2
Name: State, dtype: int64
```

In [12]:
```
plt.figure(figsize=(20, 15))
plot = sns.countplot(x='State', data=data_breach_df, palette='deep')

for p in plot.patches:
    plot.annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2.
                  ha='center', va='center', xytext=(0, 10), textcoords

plt.title('States')
plt.ylabel('Count')
plt.show()
```



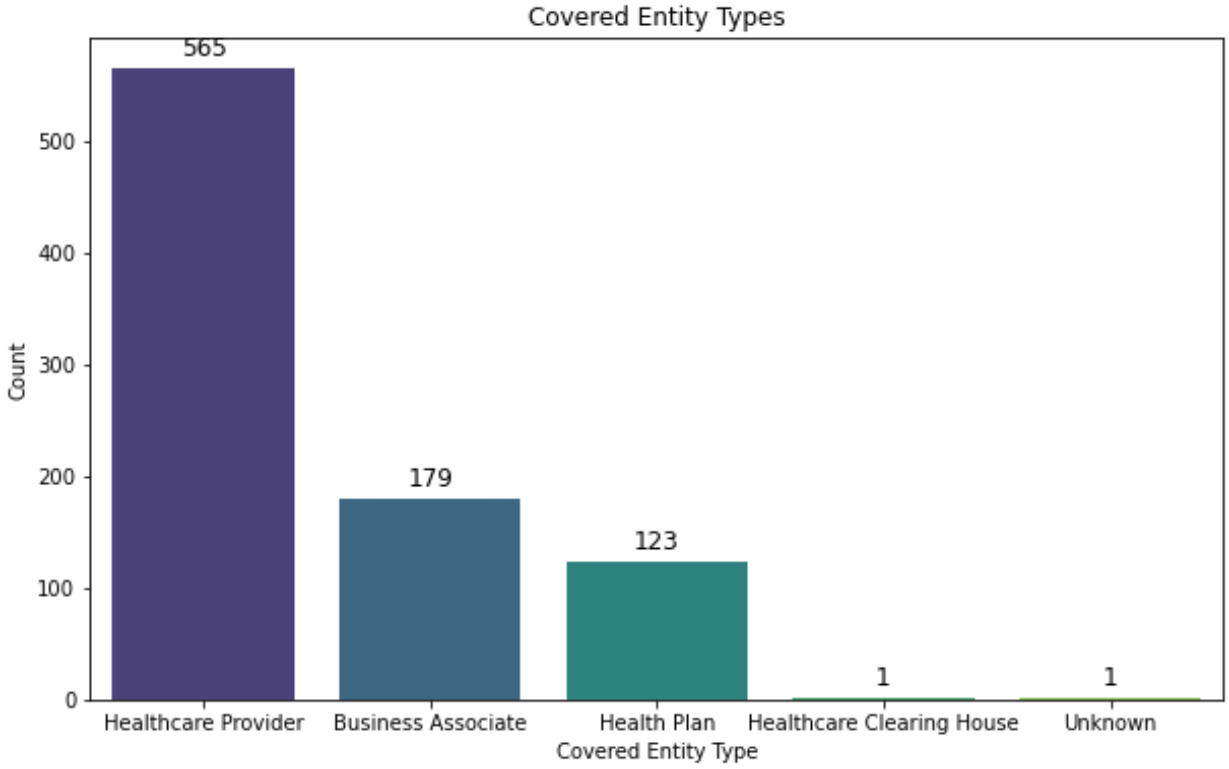
In [13]:
```
count_CoveredEntityType=data_breach_df['Covered Entity Type'].value_co
count_CoveredEntityType
```

Out[13]:
```
Healthcare Provider          565
Business Associate           179
Health Plan                  123
Healthcare Clearing House      1
Unknown                        1
Name: Covered Entity Type, dtype: int64
```

In [14]:
```
plt.figure(figsize=(10, 6))
plot = sns.countplot(x='Covered Entity Type', data=data_breach_df, pal

for p in plot.patches:
    plot.annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2.
                  ha='center', va='center', xytext=(0, 10), textcoords

plt.title('Covered Entity Types')
plt.xlabel('Covered Entity Type')
plt.ylabel('Count')
plt.show()
```



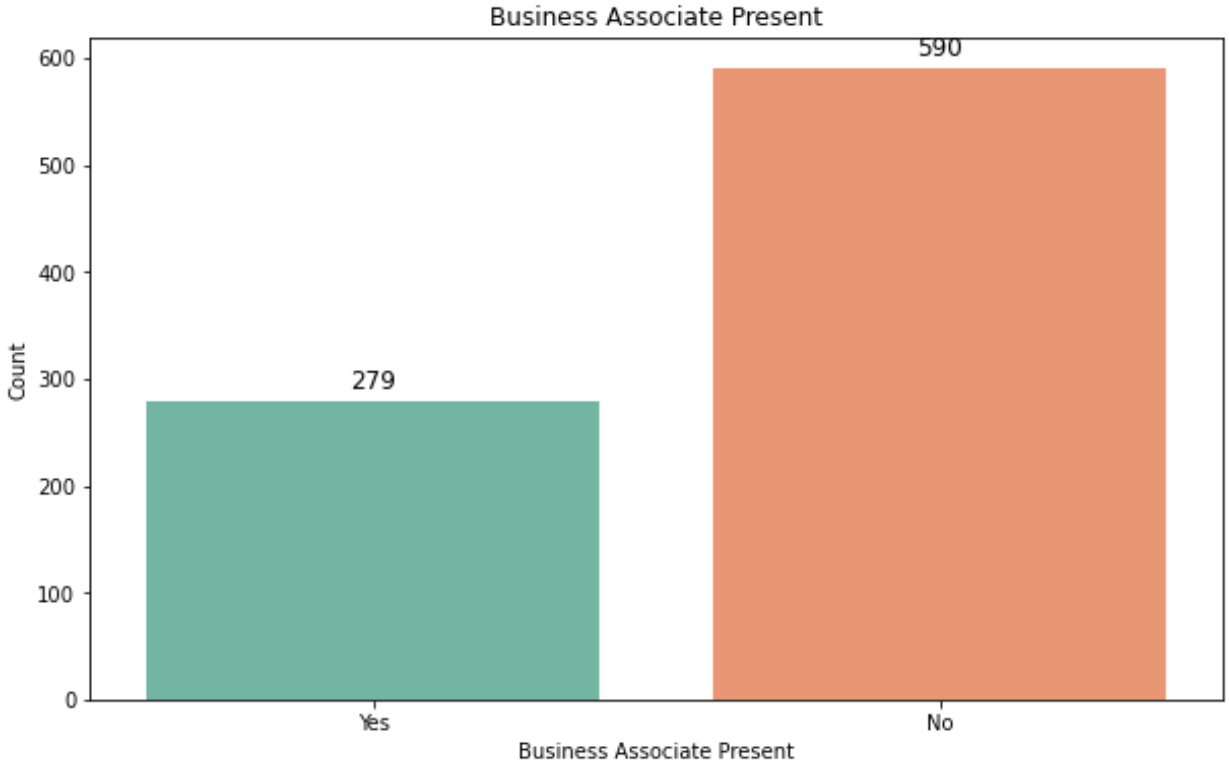
In [15]:
```
data_breach_df['Business Associate Present'].value_counts()
```

Out[15]:
```
No     590
Yes    279
Name: Business Associate Present, dtype: int64
```

In [16]:
```
plt.figure(figsize=(10, 6))
plot = sns.countplot(x='Business Associate Present', data=data_breach_

for p in plot.patches:
    plot.annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2.
                  ha='center', va='center', xytext=(0, 10), textcoords

plt.title('Business Associate Present')
plt.xlabel('Business Associate Present')
plt.ylabel('Count')
plt.show()
```



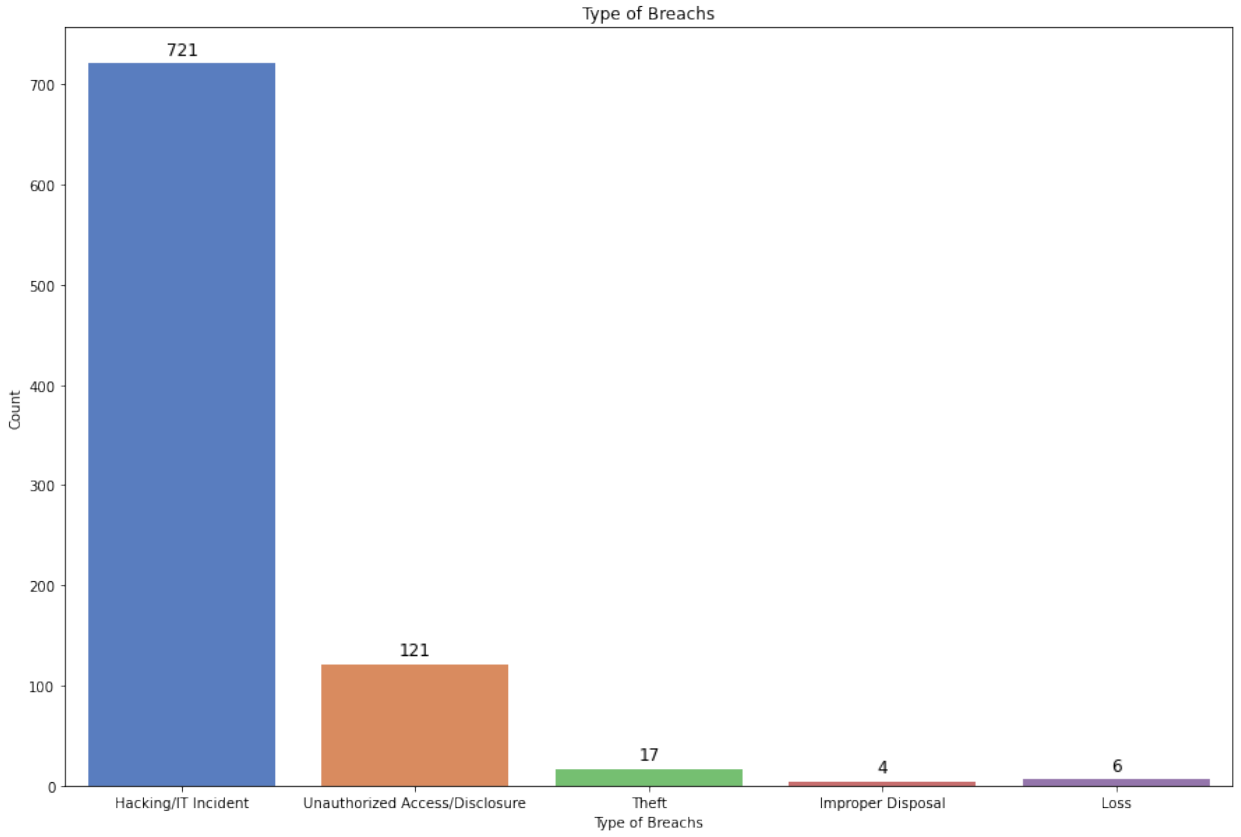
In [17]:
```
data_breach_df['Type of Breach'].value_counts()
```

Out[17]:
```
Hacking/IT Incident               721
Unauthorized Access/Disclosure    121
Theft                              17
Loss                                6
Improper Disposal                   4
Name: Type of Breach, dtype: int64
```

In [20]:
```
plt.figure(figsize=(15, 10))
plot = sns.countplot(x='Type of Breach', data=data_breach_df, palette=

for p in plot.patches:
    plot.annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2.
                  ha='center', va='center', xytext=(0, 10), textcoords

plt.title('Type of Breachs')
plt.xlabel('Type of Breachs')
plt.ylabel('Count')
plt.show()
```



In [19]:
```
data_breach_df['Location of Breached Information'].value_counts()
```

Out[19]:

Network Server	580
Email	169
Paper/Films	35
Electronic Medical Record	21
Other	13
Other Portable Electronic Device	9
Network Server, Other	8
Laptop	8
Electronic Medical Record, Network Server	6
Desktop Computer, Network Server	5
Desktop Computer	4
Desktop Computer, Email	1
Email, Other	1
Desktop Computer, Electronic Medical Record, Email, Laptop	1
Other, Paper/Films	1
Electronic Medical Record, Other	1
Desktop Computer, Laptop, Network Server	1
Electronic Medical Record, Paper/Films	1
Electronic Medical Record, Laptop	1
Desktop Computer, Electronic Medical Record	1
Email, Network Server	1
Electronic Medical Record, Email, Paper/Films	1

Name: Location of Breached Information, dtype: int64

In [ ]: