Correlation

Correlation measures the statistical relationship between two variables. It quantifies how much one variable changes when the other one does. The correlation coefficient (r) ranges from -1 to 1, where:

- +1: Perfect positive correlation
- 0: No correlation
- -1: Perfect negative correlation
- Key points about Correlation:

Types: Pearson (linear), Spearman (rank-based), Kendall (rank correlation).

Usage: Identifying relationships between variables in data analysis and feature selection in machine learning.

Practical Example in Cybersecurity

Problem: Analyzing the correlation between different features of network traffic to identify potential indicators of attacks.

Data: We'll simulate network traffic data.

Python Code

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

# Simulated network traffic data
data = {
   'duration': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100],
   'protocol_type': [1, 1, 2, 2, 1, 2, 2, 1, 1, 2],
   'bytes': [1000, 1500, 1600, 1100, 1800, 1200, 1300, 1400, 1700, 1900],
   'attack': [0, 0, 1, 0, 1, 0, 1, 1, 0, 1]
}

# Converting to DataFrame
df = pd.DataFrame(data)

# Calculating correlation matrix
correlation_matrix = df.corr()

# Plotting correlation matrix
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
```

```
plt.title('Correlation Matrix of Network Traffic Features')
plt.show()
```