Task 3. Understanding Correlation

Correlation refers to a statistical measure that describes the extent to which two variables change together. It indicates both the strength and direction of the relationship between two variables. The correlation coefficient, typically denoted by 'r', ranges from -1 to 1:

- If 'r' is close to 1, it indicates a strong positive correlation, meaning that as one variable increases, the other variable also tends to increase.
- If 'r' is close to -1, it indicates a strong negative correlation, meaning that as one variable increases, the other variable tends to decrease.
- If 'r' is close to 0, it indicates a weak or no correlation between the variables.

Practical Example in Cybersecurity:

In cybersecurity, correlation analysis can be incredibly valuable for identifying patterns and relationships within large datasets, which can help detect anomalies or potential security threats. Let's consider a scenario where we have data on network traffic and cybersecurity incidents. We want to analyze the correlation between certain network behaviors and the occurrence of security incidents.

Example Data:

We have collected data on network traffic and cybersecurity incidents over a period of time. Our dataset includes the following variables:

1. **Network Traffic:** Total bytes transmitted, number of packets sent, average packet size, etc.
2. **Cybersecurity Incidents:** Number of malware detections, phishing attempts, unauthorized access attempts, etc.

Python Code:

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

# Read data from the CSV file
data = pd.read_csv('cybersecurity_data.csv')

# Calculating correlation coefficients
correlation_matrix = data.corr()

# Extracting correlation coefficients between network traffic and
cybersecurity incidents
correlation_network_traffic = correlation_matrix.loc[['TotalBytes',
'NumPackets', 'AvgPacketSize'],
```

```
                                                    ['MalwareDetections',
'PhishingAttempts', 'UnauthorizedAccess']]

print("Correlation between Network Traffic and Cybersecurity Incidents:")
print(correlation_network_traffic)

# Calculate correlation matrix
correlation_matrix = data.corr()

# Extract correlation coefficients between network traffic and cybersecurity
incidents
correlation_network_traffic = correlation_matrix.loc[['TotalBytes',
'NumPackets', 'AvgPacketSize'],
                                                    ['MalwareDetections',
'PhishingAttempts', 'UnauthorizedAccess']]

# Plot heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_network_traffic, annot=True, cmap='coolwarm',
fmt=".2f", linewidths=0.5)
plt.title('Correlation between Network Traffic and Cybersecurity Incidents')
plt.xlabel('Cybersecurity Incidents')
plt.ylabel('Network Traffic Variables')
plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.show()
```

Below is the example of the data used:

TotalBytes	NumPackets	AvgPacketSize	MalwareDetections	PhishingAttempts	UnauthorizedAccess
3732	88	309	8	1	2
4264	69	468	5	3	1
5859	92	391	0	3	2
8891	165	476	8	4	0
5373	127	447	5	4	1
6874	80	268	4	2	2
7744	74	984	7	1	0
4468	175	904	4	1	2
1705	52	276	1	0	1
3599	53	637	3	0	2
3222	144	423	3	2	1
8768	157	971	9	4	0
3897	63	608	2	3	0
1537	162	903	5	3	1
7216	90	214	2	2	0
7921	122	898	3	1	2
7036	69	129	5	0	0
3163	145	853	7	4	1
6072	122	389	2	1	2