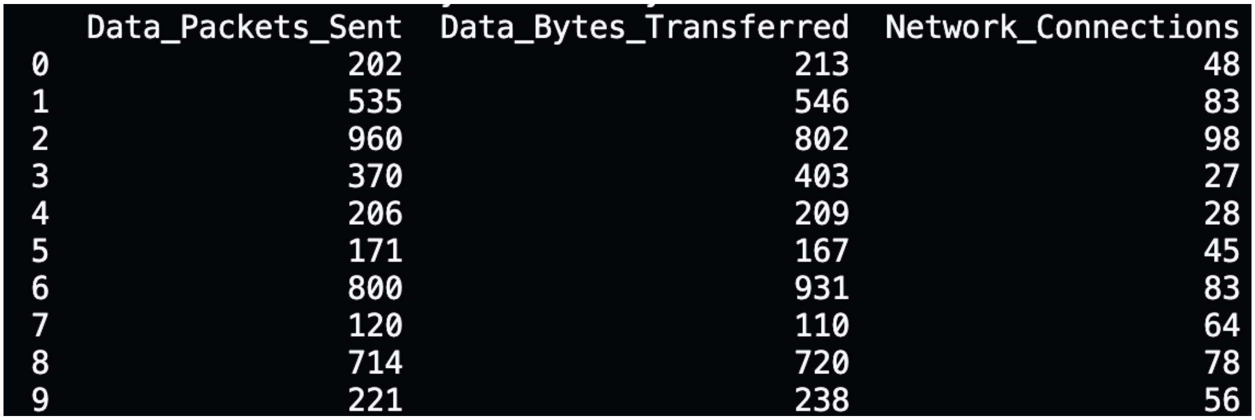
In machine learning, we use correlation to have a statistical measurement between two features. Correlation can be between -1 to 1, and in general we have three types of correlation between two variables, strong, negative, and no correlation. Strong correlation means that when one variable increases, the other also tends to increase. As an example, with more applications and services introduced in the company, the attack surface would increase. Negative correlation is the opposite of strong correlation – when one variable increases, the other decreases. For example, with more investment in cybersecurity tools and services, we would have less data breaches. Lastly, we would have no correlation meaning there is no pattern between two variables. For example, this could be something such as frequency of data breaches and sector of organizations, since most of the sectors face similar vulnerabilities and threats.

In the practical example, I decided to show three variables, data packets sent, data bytes transferred, and network connections. First two have high correlation, while the third one doesn't. I created an arbitrary dataset and created two plots to show the correlation.

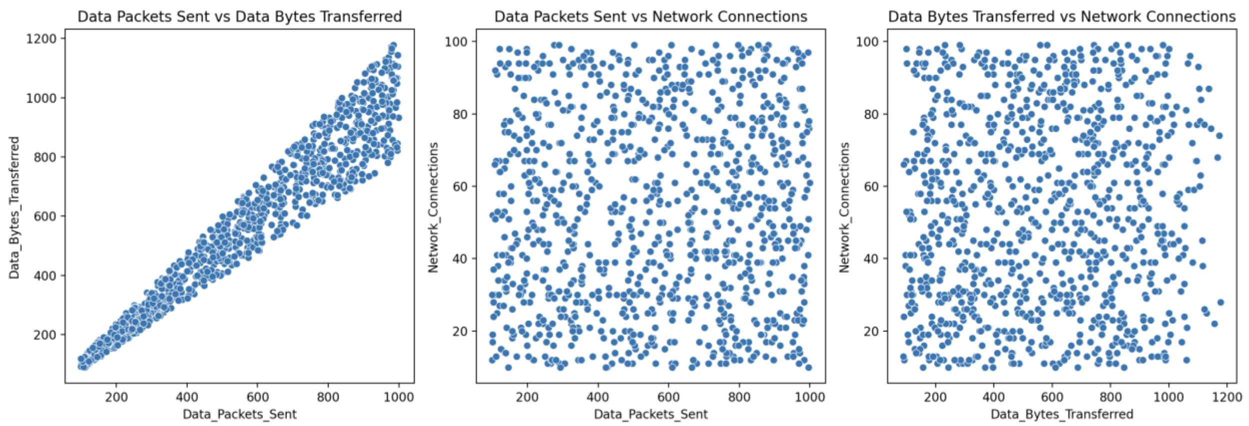
First 10 rows:

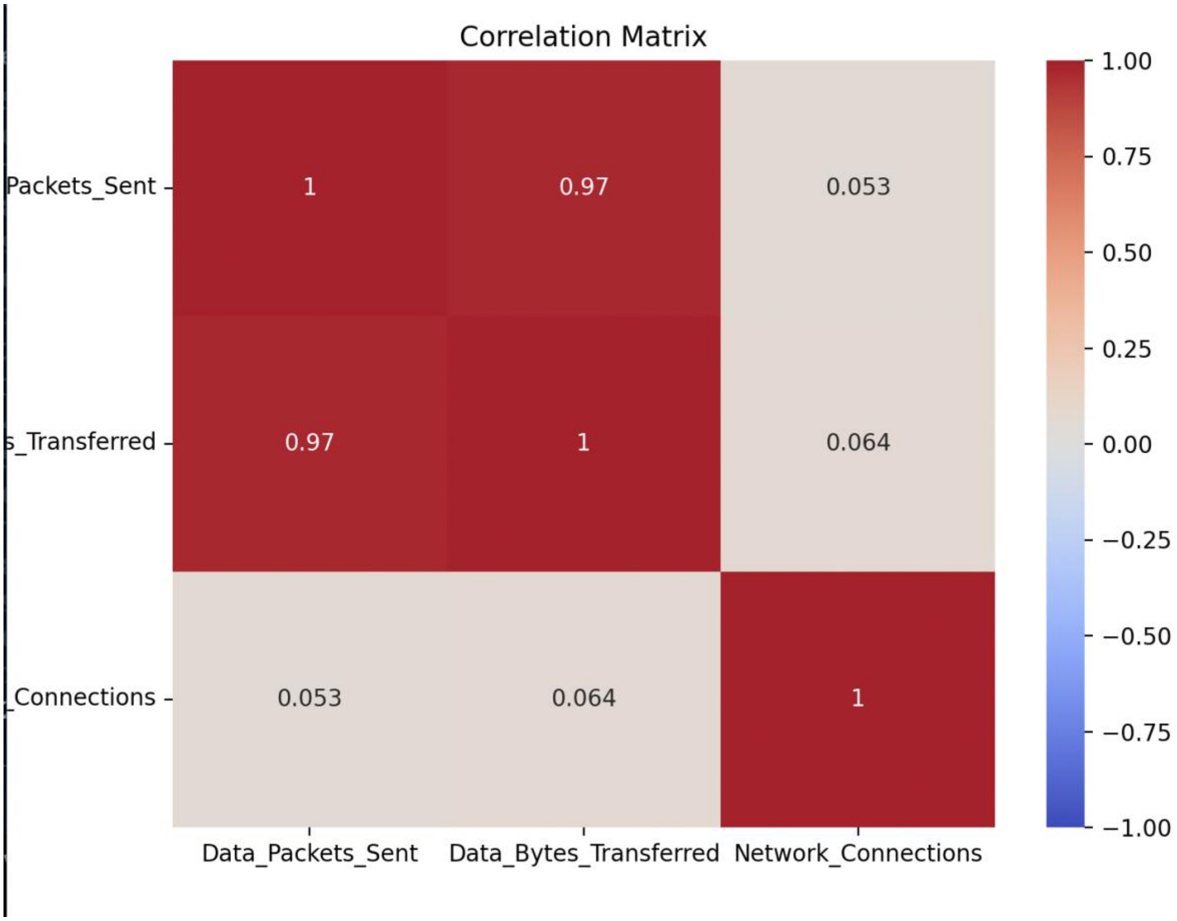


```
   Data_Packets_Sent  Data_Bytes_Transferred  Network_Connections
0                202                     213                   48
1                535                     546                   83
2                960                     802                   98
3                370                     403                   27
4                206                     209                   28
5                171                     167                   45
6                800                     931                   83
7                120                     110                   64
8                714                     720                   78
9                221                     238                   56
```

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

# Generating cybersecurity-themed dataset
np.random.seed(42)
n_samples = 1000

# Generating first two features with high correlation
data_packets_sent = np.random.randint(100, 1000, size=n_samples)
data_bytes_transferred = data_packets_sent * np.random.uniform(0.8, 1.2,
size=n_samples)

# Generating the third feature with low correlation
```

```
network_connections = np.random.randint(10, 100, size=n_samples)

# Combine features into a DataFrame
cyber_data = pd.DataFrame({'Data_Packets_Sent': data_packets_sent,
'Data_Bytes_Transferred': data_bytes_transferred, 'Network_Connections':
network_connections})

# Plotting the features
plt.figure(figsize=(15, 5))
plt.subplot(1, 3, 1)
sns.scatterplot(x='Data_Packets_Sent', y='Data_Bytes_Transferred', data=cyber_data)
plt.title('Data Packets Sent vs Data Bytes Transferred')
plt.subplot(1, 3, 2)
sns.scatterplot(x='Data_Packets_Sent', y='Network_Connections', data=cyber_data)
plt.title('Data Packets Sent vs Network Connections')
plt.subplot(1, 3, 3)
sns.scatterplot(x='Data_Bytes_Transferred', y='Network_Connections', data=cyber_data)
plt.title('Data Bytes Transferred vs Network Connections')
plt.tight_layout()
plt.show()

# Calculating correlation matrix
correlation_matrix = cyber_data.corr()

# Plotting correlation matrix
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Correlation Matrix')
plt.show()
```




Correlation Matrix
Packets_Sent
s_Transferred
_Connections
1
0.97
0.053
0.97
1
0.064
0.053
0.064
1
Data_Packets_Sent
Data_Bytes_Transferred
Network_Connections
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00