# LAPORAN TUGAS BESAR

# MATA KULIAH WI2002

# LITERASI DATA DAN INTELEGENSI ARTIFISIAL

# TAHUN 2025


**Disusun oleh Kelompok 14 – Kelas 32**

1. **Benedict Darrel Setiawan 13524057**
2. **Marcel Luther Sitorus 13524063**
3. **Muhammad Akmal 13524099**
4. **Daniel Putra Rywandi S 13524143**


**MATA KULIAH WAJIB KURIKULUM**

**WAKIL REKTOR BIDANG AKADEMIK DAN KEMAHASISWAAN**

**INSTITUT TEKNOLOGI BANDUNG**

**BANDUNG**

**2025**

# DAFTAR ISI

Daftar Tabel ..... 3
Daftar Bagan ..... 4
Bab 1 – Pendahuluan ..... 5
A. Latar Belakang ..... 5
B. Pertanyaan Penelitian ..... 5
C. Data ..... 5
Dataset 1 – Lalu-Lintas Normal ..... 5
Dataset 2 – Lalu-Lintas Serangan ..... 6
D. Atribut Data ..... 6
Bab 2 – Statistik Deskriptif dan Visualisasi Data ..... 7
A. Statistika Deskriptif ..... 7
1. Data Set 1 – Lalu-Lintas Normal ..... 7
2. Data Set 2 – Lalu-Lintas Serangan ..... 8
B. Visualisasi Data ..... 9
1. Diagram Batang ..... 9
2. Perubahan Terhadap Waktu ..... 10
3. Hierarki dan Hubungan Keseluruhan-Bagian ..... 11
*4. Plotting Relationships* ..... 14
Bab 3 – Pengolahan Data ..... 19
Bab 4 – Kesimpulan ..... 28

# DAFTAR TABEL

Tabel 1 – Atribut Relevan Data Set ..... 6
Tabel 2 – Statistik Deskriptif Data Set 2..... 7
Tabel 3 – Statistik Deskriptif Data Set 2..... 8

# DAFTAR BAGAN

Bagan 1 – Perbandingan Chart Rata-rata SYN dan ACK Flag Count pada Hari Senin dan Rabu ..... 9

Bagan 2 – Grouped Bar Chart Rata-rata Fwd dan Bwd Packet Length Mean pada Hari Senin dan Rabu..... 11

Bagan 3 – Perbandingan Flow Byte Terhadap Waktu ..... 11

Bagan 4 – Sunburst Chart Penyebab Gangguan Aktivitas Lalu Lintas pada Tanggal 5 Juli 2017..... 11

Bagan 5 – Sunburst Chart Aktivitas Lalu Lintas di Selang Waktu 8:00 - 16:00 pada Tanggal 5 Juli 2017..... 11

Bagan 6 – Scatter Plot *Total Forward Packets* dan *Total Backward Packets* pada dataset hari Rabu, 5 Juli 2017 ..... 11

Bagan 7 – *Zoom Out* Scatter Plot *Total Forward Packets* dan *Total Backward Packets* pada dataset hari Rabu, 5 Juli 2017..... 16

Bagan 8 – Scatter Plot Forward IAT Mean dan Backward IAT Mean pada dataset hari Rabu, 5 Juli 2017..... 17

Bagan 9 – Scatter Plot *Idle Mean* dan *Active Mean* pada dataset hari Rabu, 5 Juli 2017 ..... 18

Bagan 10 – Korelasi Data Set 1 ..... 19

Bagan 11 – Korelasi Data Set 2 ..... 20

# BAB 1 – PENDAHULUAN

## A. Latar Belakang

Di era digital saat ini, pertumbuhan layanan berbasis internet telah menjadikan server web sebagai infrastruktur penting bagi organisasi dan individu. Namun, ketergantungan yang meluas pada teknologi web ini juga membuat sistem rentan terhadap berbagai ancaman keamanan siber, termasuk serangan *Distributed Denial of Service* (DDoS), upaya *Brute-Force Login*, dan aktivitas jahat lainnya. Deteksi dan analisis yang efektif terhadap perilaku anomali dalam lalu lintas jaringan sangat penting untuk menjaga integritas sistem dan mencegah kerusakan.

Bidang analisis lalu lintas jaringan memanfaatkan data tingkat aliran dalam jumlah besar, yang merekam atribut seperti jumlah paket, durasi koneksi, kecepatan transfer data, dan penggunaan protokol. Dengan memeriksa atribut kuantitatif dan kategoris ini, menjadi mungkin untuk mengungkap pola tersembunyi yang membedakan perilaku normal dari serangan potensial.

Proyek ini bertujuan untuk menyelidiki hubungan antara fitur jaringan berbasis aliran dan keberadaan serangan siber. Dengan menganalisis lalu lintas yang ditangkap selama skenario jinak dan serangan, kami berupaya mengidentifikasi atribut mana yang berfungsi sebagai indikator kuat anomali. Secara khusus, fokusnya adalah pada pemodelan bagaimana karakteristik lalu lintas yang berbeda berkorelasi dengan perilaku jahat, menggunakan teknik pembelajaran mesin statistik dan sederhana.

## B. Pertanyaan Penelitian

Berdasarkan latar belakang tersebut, kami merumuskan pertanyaan penelitian sebagai berikut:

1. Bagaimana metrik jaringan tingkat aliran (*flow-level network metrics*) seperti *packet counts*, ukuran byte, dan durasi aliran berkorelasi dengan keberadaan serangan siber dalam lalu lintas server web?
2. Dapatkah kita memprediksi anomali atau perilaku tidak teratur dalam lalu lintas jaringan berdasarkan fitur statistik berbasis aliran (*flow-based statistics*)?
3. Apakah terdapat pola tertentu pada fitur lalu lintas jaringan (misalnya durasi aliran, jumlah paket, ukuran rata-rata paket) yang dapat digunakan untuk mendeteksi keberadaan serangan DDoS?

## C. Data

Untuk menjawab pertanyaan penelitian yang dirumuskan, proyek ini menggunakan *dataset* CIC-IDS2017 yang disediakan oleh Canadian Institute for Cybersecurity (CIC). Dataset CIC-IDS2017 berisi serangan umum yang jinak dan terkini, yang menyerupai data dunia nyata yang sebenarnya, sering disebut *analyzing packet captures* (PCAP). Dataset ini juga mencakup hasil analisis lalu lintas jaringan menggunakan CIC FlowMeter dengan aliran berlabel berdasarkan cap waktu, IP sumber dan tujuan, port sumber dan tujuan, protokol dan serangan.

Dataset CIC-IDS2017 dipilih karena representasinya yang komprehensif terhadap perilaku lalu lintas jaringan modern, yang menggabungkan aktivitas jinak dan berbagai skenario serangan siber dalam kondisi yang terkontrol dan realistis. Kumpulan data ini menyediakan metrik berbasis aliran yang terperinci, pelabelan yang ekstensif, dan komponen deret waktu, yang semuanya selaras erat dengan tujuan analitis proyek ini.

Meskipun CIC-IDS2017 secara resmi merupakan satu dataset terpadu, dataset ini terbagi menjadi koleksi harian yang menyimulasikan kondisi operasional. Pada hari-hari tertentu, dataset secara eksklusif terdiri dari lalu lintas normal yang bebas serangan, sementara pada hari-hari lainnya mencakup campuran aktivitas jinak dan berbahaya. Untuk memenuhi persyaratan penggunaan dua set data, kami memperlakukan lalu lintas normal yang direkam pada hari Senin, 3 Juli 2017, sebagai set data pertama (yang mewakili operasi server web dasar), dan lalu lintas serangan campuran yang direkam pada hari Rabu, 5 Juli 2017, sebagai set data kedua yang berbeda. Pemisahan ini dibenarkan secara struktural dan analitis, karena memungkinkan analisis komparatif antara perilaku jaringan yang umum dan lalu lintas yang terkena serangan.
**URL Sumber Dataset:** https://www.unb.ca/cic/datasets/ids-2017.html

**Dataset 1 – Lalu-Lintas Normal**

Spesifikasi data set pertama kami sebagai berikut:

- Sumber: Canadian Institute for Cybersecurity (CIC) – CIC-IDS2017 Dataset
- Nama File: Monday-WorkingHours.pcap_ISCX.csv
- Format: *Comma-Separated Values* (CSV)
- Ukuran: 262 MB
- Deskripsi: Simulasi rekaman aktivitas lalu-lintas jaringan normal tanpa aktivitas serangan.

**Dataset 2 – Lalu-Lintas Serangan**

Spesifikasi data set kedua kami sebagai berikut:

- Sumber: Canadian Institute for Cybersecurity (CIC) – CIC-IDS2017 Dataset
- Nama File: Wednesday-WorkingHours.pcap_ISCX.csv
- Format: *Comma-Separated Values* (CSV)
- Ukuran: 278 MB
- Deskripsi: Menggambarkan skenario serangan nyata yang dipadukan dengan perilaku normal.

### D. Atribut Data

Kedua dataset menggunakan format *Comma-Separated Values* (CSV) yang memiliki *header* atribut yang sama. Karena besarnya ukuran data dan banyaknya label dari dataset induk yang mencapai 85 atribut, kami memutuskan untuk memilih 18 atribut yang relevan dalam penelitian kami untuk dimasukkan dianalisis. Berikut ini adalah label-label yang relevan yang akan digunakan dalam penelitian ini.

**Tabel 1 – Atribut Relevan Data Set**

| No. | Nama Label | Deskripsi | Tipe | Alasan Pemilihan |
|---|---|---|---|---|
| 1. | Source IP | Alamat IP pengirim | Kategorikal (Nominal) | Untuk mengidentifikasi wilayah yang berpotensi menjadi penyerang. |
| 2. | Source Port | Nomor Port sumber | Numerik (Diskret) | Mirip, tetapi untuk lokal. |
| 3. | Protocol | Jenis protocol jaringan (TCP/UDP/ICMP) | Kategorikal (Nominal) | Indikator jenis lalu lintas tingkat tinggi |
| 4. | Timestamp | Waktu mulai koneksi | Kategorikal (Ordinal – Waktu) | Syarat atribut data pada tugas. |
| 5. | Flow Duration | Durasi total koneksi | Numerik (Kontinu) | Ukuran lalu lintas fundamental (kandidat regresi primer) |
| 6. | Total Fwd Packets | Jumlah paket terkirim maju (*forward)* | Numerik (Diskret) | Intensitas aliran pada arah pengiriman |
| 7. | Total Backward Packets | Jumlah paket terkirim mundur (*backward)* | Numerik (Diskret) | Intensitas aliran pada arah penerimaan |
| 8. | Flow Bytes/s | Kecepatan alir (*bytes/sec*) | Numerik (Kontinu) | Kecepatan volume lalu lintas (dapat melonjak selama serangan) |
| 9. | Flow Packets/s | Kecepatan alir (*packets/sec*) | Numerik (Kontinu) | Perilaku laju paket (dapat mengidentifikasi DDoS) |
| 10. | Fwd Packet Length Mean | Ukuran rerata (*forward packets*) | Numerik (Kontinu) | Berguna untuk mengkarakterisasi profil lalu lintas |
| 11. | Bwd Packet Length Mean | Ukuran rerata (*backward packets*) | Numerik (Kontinu) | Mirip, tetapi untuk respons |
| 12. | SYN Flag Count | Jumlah SYN (*synchronizes sequence numbers*) *flags* | Numerik (Diskret) | Penting untuk mengidentifikasi *SYN flood attack*. |
| 13. | ACK Flag Count | Jumlah ACK (*acknowledgment*) *flags* | Numerik (Diskret) | Indikator pembentukan sesi |
| 14. | Down/Up Ratio | Rasio unduhan terhadap unggahan | Numerik (Kontinu) | Pola perilaku: eksfiltrasi, aliran yang banyak diunduh |
| 15. | Average Packet Size | Ukuran rerata besar paket | Numerik (Kontinu) | Efisiensi ukuran atau petunjuk anomali |
| 16. | Active Mean | Durasi aktif rata-rata selama aliran | Numerik (Kontinu) | Berapa lama koneksi tetap aktif bertukar data |
| 17. | Idle Mean | Durasi idle rata-rata selama aliran | Numerik (Kontinu) | Idle yang lama dapat menunjukkan aktivitas yang mencurigakan/laju rendah |
| 18. | Label | Klasifikasi serangan/jinak | Kategorikal (Nominal) | Variabel target untuk analisis dan visualisasi |

# BAB 2 – STATISTIK DESKRIPTIF DAN VISUALISASI DATA

## A. Statistika Deskriptif

### 1. Data Set 1 – Lalu-Lintas Normal

**Tabel 2 – Statistik Deskriptif Data Set 1**

| No | Nama Label | Rata-rata | Standar Deviasi | P10 | P25 | Median | P75 | P90 | Max | Min |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | Flow Duration | 10389271.16 | 28751927.25 | 39 | 176 | 31303 | 355744.75 | 44204485.90 | 119999987 | -1 |
| 2. | Total Fwd Packets | 10.39 | 892.41 | 1 | 2 | 2 | 4 | 13 | 219759 | 1 |
| 3. | Total Backward Packets | 11.51 | 1173.32 | 0 | 1 | 2 | 3 | 11 | 291922 | 0 |
| 4. | Flow Bytes/s | 16132666.79 | 24627660.87 | 0.0 | 169.87 | 5556.26 | 236794.17 | 17699115.5 | 2071000000.0 | -12000000.0 |
| 5. | Flow Packets/s | 66238.02 | 235212.31 | 0.57 | 19.22 | 114.48 | 21164.02 | 58823.53 | 3000000.0 | -2000000.0 |
| 6. | Fwd Packet Length Mean | 50.78 | 91.99 | 0.0 | 6.0 | 38.0 | 53.0 | 93.86 | 4638.92 | 0.0 |
| 7. | Bwd Packet Length Mean | 164.9 | 277.06 | 0.0 | 0.0 | 81.0 | 164.0 | 485.89 | 2976.32 | 0.0 |
| 8. | SYN Flag Count | 0.06 | 0.24 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 9. | ACK Flag Count | 0.31 | 0.46 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 |
| 10. | Down/Up Ratio | 0.64 | 0.55 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 108.0 | 0.0 |
| 11. | Average Packet Size | 116.61 | 166.17 | 0 | 9 | 75.75 | 126.5 | 289.78 | 3684 | 0 |
| 12. | Active Mean | 68434.82 | 587232.21 | 0 | 0 | 0 | 0 | 63208.38 | 101659665 | 0 |
| 13. | Idle Mean | 34639177.94 | 12970567.22 | 0 | 0 | 0 | 0 | 9958582.53 | 119999735 | 0 |

Pada dataset yang dianalisis, ditemukan beberapa anomali data seperti nilai negatif pada kolom *Flow Duration*, *Flow Bytes/s*, dan *Flow Packets/s*, yang secara logis tidak masuk akal karena durasi dan kecepatan seharusnya bernilai positif. Hal ini menunjukkan perlunya proses pembersihan data (*data cleaning*) sebelum dilakukan analisis lebih lanjut.

Distribusi protokol dalam dataset didominasi oleh TCP (protokol 6) dan UDP (protokol 17), sesuai dengan karakteristik umum lalu lintas jaringan. Sebagian besar flow bersifat pendek, dengan rata-rata jumlah paket per flow berkisar 10–11, namun terdapat standar deviasi yang tinggi, mengindikasikan keberadaan beberapa flow dengan jumlah paket yang sangat besar (outlier). Analisis terhadap flag TCP menunjukkan bahwa sebagian besar flow tidak memulai koneksi baru (rata-rata SYN flag rendah), namun banyak yang berada dalam fase komunikasi (ACK flag relatif lebih sering muncul), yang bisa dimanfaatkan untuk mendeteksi pola serangan seperti SYN flood atau koneksi abnormal.

Dari segi ukuran paket, rata-rata panjang paket arah maju (forward) dan balik (backward) tergolong kecil, namun terdapat nilai maksimum yang sangat tinggi, menunjukkan bahwa sebagian kecil flow membawa data dalam jumlah besar. Waktu aktif dan idle juga menunjukkan distribusi yang ekstrem, dengan sebagian besar flow memiliki durasi aktif sangat singkat namun waktu idle yang sangat panjang, yang bisa menjadi indikator serangan

berbasis botnet atau aktivitas mencurigakan lain. Selain itu, nilai *Down/Up Ratio* sebagian besar berada di angka 1, menandakan lalu lintas dua arah yang seimbang, namun nilai maksimum yang sangat tinggi menunjukkan adanya flow yang hanya mendownload atau mengupload secara ekstrem, yang bisa mengindikasikan serangan seperti DoS atau pencurian data (*data exfiltration*).

### 2. Data Set 2 – Lalu-Lintas Serangan

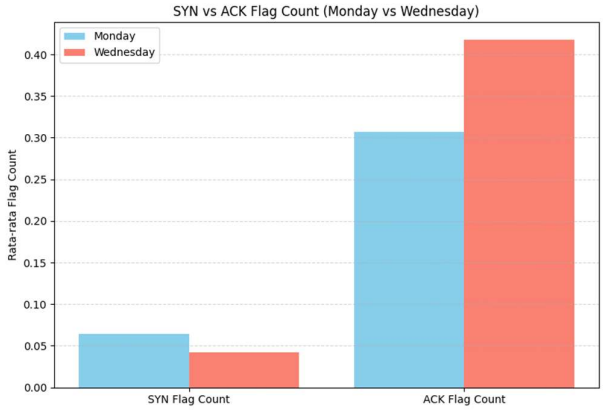
**Tabel 33 – Statistik Deskriptif Data Set 2**

| No | Nama Label | Rata-rata | Standar Deviasi | P10 | P25 | Median | P75 | P90 | Max | Min |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | Flow Duration | 28001680.75 | 42766771.33 | 5 | 201 | 61437 | 830240373 | 99417961.8 | 119999998 | -1 |
| 2. | Total Fwd Packets | 9.56 | 747.20 | 1 | 2 | 2 | 7 | 10 | 203943 | 1 |
| 3. | Total Backward Packets | 10.213 | 984.20 | 0 | 1 | 2 | 6 | 8 | 272353 | 0 |
| 4. | Flow Bytes/s | 1729533.08 | 29615636.04 | 0.0 | 102.79 | 515.51 | 18702.26 | 1277777.78 | 2070000000.0 | -12000000.0 |
| 5. | Flow Packets/s | 99631.51 | 323148.85 | 0.13 | 0.28 | 63.00 | 18181.82 | 500000.0 | 3000000.0 | -2000000.0 |
| 6. | Fwd Packet Length Mean | 60.64 | 157.74 | 0.0 | 6.0 | 41.0 | 56.8 | 99.022 | 4640.76 | 0.0 |
| 7. | Bwd Packet Length Mean | 552.75 | 797.75 | 0.0 | 0.0 | 102.0 | 929.0 | 1932.5 | 4370.69 | 0.0 |
| 8. | SYN Flag Count | 0.04 | 0.20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| 9. | ACK Flag Count | 0.42 | 0.49 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 |
| 10. | Down/Up Ratio | 0.56 | 0.57 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 43.0 | 0.0 |
| 11. | Average Packet Size | 305.66 | 398.05 | 0 | 9 | 91.0 | 696.07 | 926.23 | 2612 | 0 |
| 12. | Active Mean | 92244.78 | 700704.89 | 0 | 0 | 0 | 991 | 29476.83 | 100000000.0 | 0 |
| 13. | Idle Mean | 22111218.77 | 38124153.51 | 0 | 0.0 | 0 | 15900000 | 98300000 | 120000000 | 0 |

## B. Visualisasi Data

### 1. Diagram Batang

**Bagan 1 – Perbandingan Chart Rata-rata SYN dan ACK Flag Count pada Hari Senin dan Rabu**



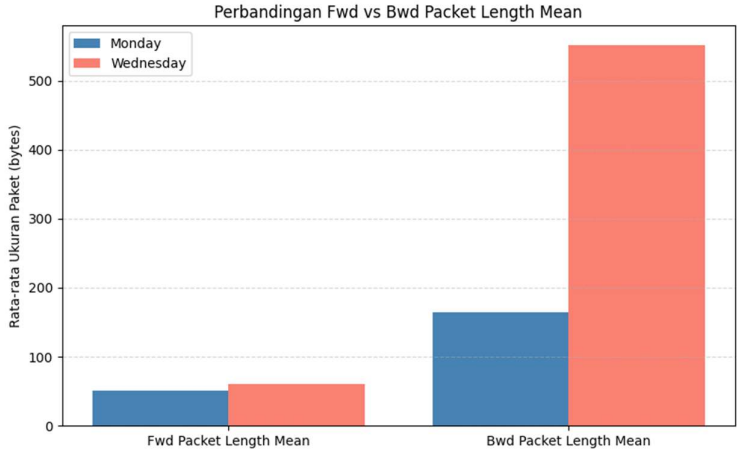

Grafik ini menunjukkan perbedaan jumlah rata-rata flag SYN dan ACK antara Hari Senin dan Rabu. Terlihat bahwa pada Hari Rabu, terjadi peningkatan signifikan pada ACK Flag Count dibandingkan Hari Senin, yang bisa mengindikasikan lebih banyak koneksi yang terkonfirmasi. Sementara SYN Flag Count tetap rendah dan stabil.

Berikut kode python yang digunakan untuk melakukan visualisasi data

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

df_sun = pd.read_csv(r"Monday-WorkingHours.pcap_ISCX.csv")
df_wed = pd.read_csv(r"Wednesday-workingHours.pcap_ISCX.csv")

df_sun.columns = df_sun.columns.str.strip()
df_wed.columns = df_wed.columns.str.strip()


labels_v1 = ['SYN Flag Count', 'ACK Flag Count']

mon_v1 = [
    df_sun['SYN Flag Count'].mean(),
    df_sun['ACK Flag Count'].mean()
]

wed_v1 = [
    df_wed['SYN Flag Count'].mean(),
    df_wed['ACK Flag Count'].mean()
]

x1 = range(len(labels_v1))
bar_width = 0.4

plt.figure(figsize=(8, 5))
plt.bar([i - bar_width/2 for i in x1], mon_v1, width=bar_width, label='Monday', color='skyblue')
plt.bar([i + bar_width/2 for i in x1], wed_v1, width=bar_width, label='Wednesday', color='salmon')
plt.xticks(x1, labels_v1)
plt.ylabel('Rata-rata Flag Count')
plt.title('SYN vs ACK Flag Count (Monday vs Wednesday)')
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()

labels_v2 = ['Fwd Packet Length Mean', 'Bwd Packet Length Mean']

sun_v2 = [
    df_sun['Fwd Packet Length Mean'].mean(),
```

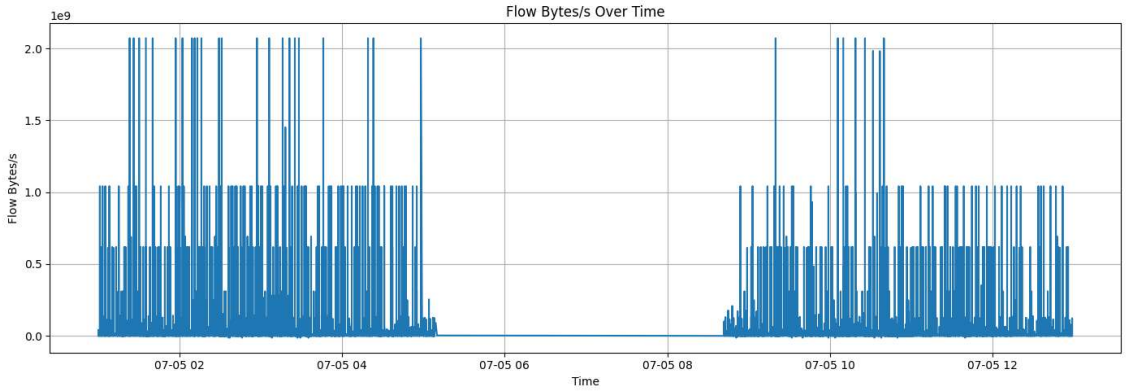
```
    df_sun['Bwd Packet Length Mean'].mean()
]

wed_v2 = [
    df_wed['Fwd Packet Length Mean'].mean(),
    df_wed['Bwd Packet Length Mean'].mean()
]

x2 = range(len(labels_v2))

plt.figure(figsize=(8, 5))
plt.bar([i - bar_width/2 for i in x2], sun_v2, width=bar_width, label='Sunday', color='skyblue')
plt.bar([i + bar_width/2 for i in x2], wed_v2, width=bar_width, label='Wednesday', color='salmon')
plt.xticks(x2, labels_v2)
plt.ylabel('Rata-rata Ukuran Paket (bytes)')
plt.title('Fwd vs Bwd Packet Length Mean (Sunday vs Wednesday)')
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()
```

**Bagan 2 – Grouped Bar Chart Rata-rata Fwd dan Bwd Packet Length Mean pada Hari Senin dan Rabu**




Terlihat bahwa pada Hari Rabu terjadi lonjakan signifikan pada rata-rata panjang paket backward dibandingkan Monday. Hal ini dapat menunjukkan pola transfer data besar dari server ke klien atau potensi serangan DDoS.

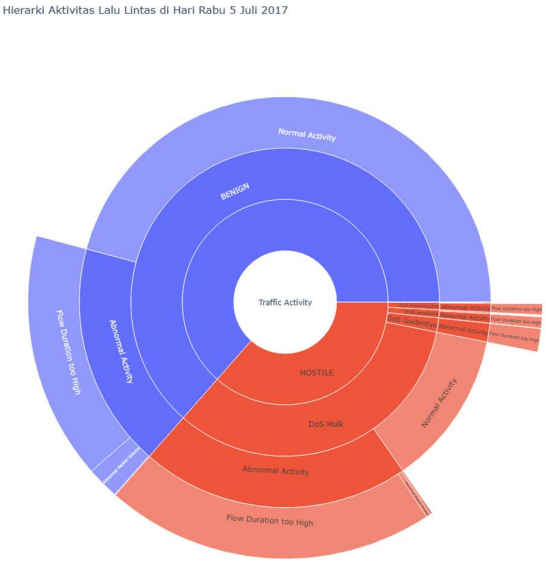
**Bagan 2 – Perbandingan Flow Byte Terhadap Waktu**




Sumbu horizontal (x) merepresentasikan waktu dengan format Bulan-Tanggal dan Waktu, sedangkan sumbu vertikal (y) menunjukkan jumlah *Flow Bytes/s* dalam satuan byte per detik. Dari grafik ini, terlihat adanya fluktuasi yang signifikan pada jumlah *Flow Bytes/s* sepanjang waktu. Terdapat beberapa puncak aktivitas yang tinggi, yang kemungkinan menunjukkan adanya pola lalu lintas jaringan yang tidak biasa atau serangan siber. Di sisi lain, terdapat juga periode dengan aktivitas rendah atau mendekati nol, yang mungkin mencerminkan waktu idle atau minimnya aktivitas jaringan.

## 2. Hierarki dan Hubungan Keseluruhan-Bagian

**Bagan 4 – Sunburst Chart Penyebab Gangguan Aktivitas Lalu Lintas pada Tanggal 5 Juli 2017**




Dari Gambar 1.5 dapat terlihat bahwa penyebab utama dari gangguan lalu lintas yang muncul pada tanggal 5 Juli 2017 adalah serangan dari DoS Hulk yang aktivitas abnormalnya ditandai dengan Flow Duration yang terlalu tinggi. Grafik tersebut juga memperlihatkan bahwa secara umum sebagian besar aktivitas abnormal disebabkan oleh Flow Duration yang terlalu tinggi.

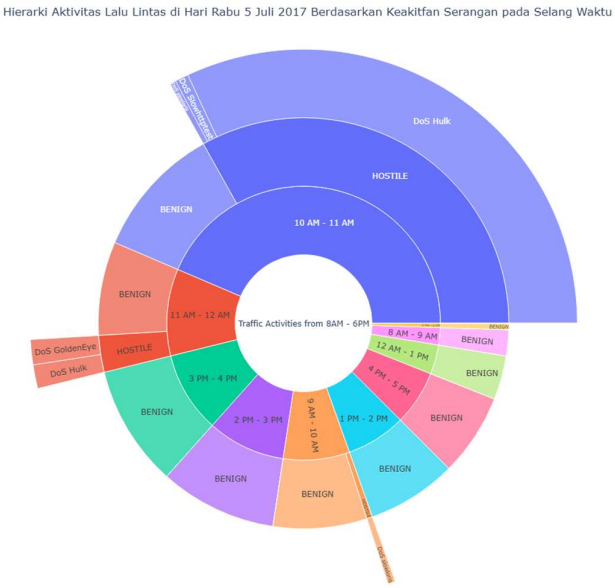
Untuk menentukan apakah aktivitas normal atau abnormal, digunakan batas atas pencilan dari data di hari Senin sebagai basis data pada saat aktivitas normal. Berikut kode python yang digunakan untuk melakukan visualisasi data di Gambar 1.5

```
import matplotlib.pyplot as plt
import numpy as np
import pandas
```

```
import plotly.express as px

data = pandas.read_csv(r"Wednesday.csv")

FD = data[' Flow Duration'].tolist()
TFwdP = data[' Total Fwd Packets'].tolist()
TBwdP = data[' Total Backward Packets'].tolist()
L = data[' Label'].tolist()

LReasoning = []
LActivity = []
LCategorized = []
LParent = []
LAmount = []
LWhole = []



for x in L :
    found = False
    for y in range(len(LCategorized)) :
        if x == LCategorized[y] :
            found = True

    if found == False :
        LCategorized.append(x)
        LCategorized.append(x)
        LCategorized.append(x)
        LCategorized.append(x)
        LActivity.append("Normal Activity")
        LActivity.append("Abnormal Activity")
        LActivity.append("Abnormal Activity")
        LActivity.append("Abnormal Activity")
        LReasoning.append(None)
        LReasoning.append("Flow Duration too High")
        LReasoning.append("No Flow Duration")
        LReasoning.append("Abnormal Packet Volume")
        LAmount.append(0)
        LAmount.append(0)
        LAmount.append(0)
        LAmount.append(0)

for i in range(len(FD)) :
    y = L[i]
    if ((TFwdP[i] <= 7) and (TBwdP[i] <= 6) and (FD[i] <= 889097) and (FD[i] > 0)) :
        x = "Normal Activity"
        z = None
    else :
        x = "Abnormal Activity"
        if (FD[i] <= 0) :
            z = "No Flow Duration"
        elif (FD[i] > 889097):
            z = "Flow Duration too High"
        else :
            z = "Abnormal Packet Volume"

    for j in range(len(LCategorized)) :
        if (LCategorized[j] == y) and (LActivity[j] == x) and (LReasoning[j] == z):
            LAmount[j] += 1


for x in LCategorized :
    if (x == "BENIGN") :
        LParent.append("")
    else :
        LParent.append("HOSTILE")

for x in LParent :
    LWhole.append("Traffic Activity")


sun_data = dict(E=np.array(LReasoning), C=np.array(LActivity), A = np.array(LCategorized), B =
np.array(LParent), D = np.array(LWhole), value = np.array(LAmount))
```

```
fig = px.sunburst(
    sun_data,
    path = ['D', 'B', 'A', 'C', 'E'],
    values = 'value',
    height= 1000, width=1000,
    title="Hierarki Aktivitas Lalu Lintas di Hari Rabu 5 Juli 2017 Berdasarkan Penyebab Gangguan Lalu Lintas")

fig.show()
```

**Bagan 5 – Sunburst Chart Aktivitas Lalu Lintas di Selang Waktu 8:00 - 16:00 pada Tanggal 5 Juli 2017**




Dari Gambar 1.6 dapat terlihat bahwa selang waktu yang memiliki serangan yang terdeteksi paling banyak proporsional dengan banyak data yang diambil adalah selang waktu jam 10 pagi hingga jam 11 pagi. Jenis DoS yang mendominasi selang waktu tersebut merupakan DoS Hulk. Dari grafik ini dapat diketahui selang waktu serangan yang dilakukan oleh setiap jenis DoS.

Berikut kode python yang digunakan untuk melakukan visualisasi data di Gambar 1.6

```
import matplotlib.pyplot as plt import numpy as np import pandas
import plotly.express as px
data = pandas.read_csv(r"Wednesday.csv")

Timestamp = data[' Timestamp'].tolist()
L = data[' Label'].tolist()

def getHour(time) :
    time = time[9:]
    hour = ""
    for x in time :
        if (x == ':') :
            ihour = int(hour)
            if ihour < 8 :
                hour += " PM"
            else :
                hour += " AM"
            iihour = ihour + 1
            if iihour > 12 :
                iihour -= 12
            hour += " - " + str(iihour)
            if iihour < 8 :
                hour += " PM"
```

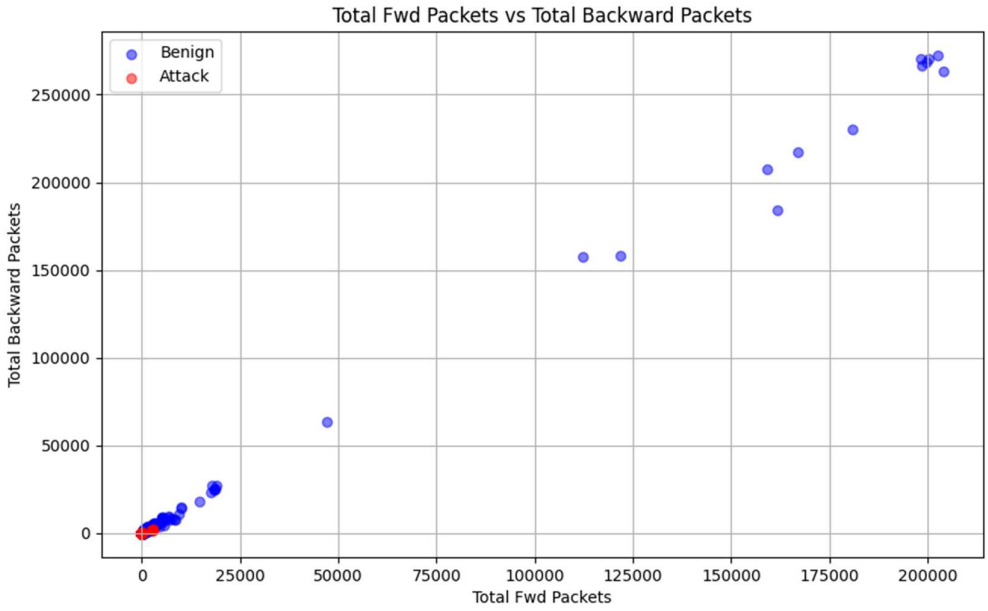
```
            else :
                hour += " AM"
            return hour

        hour += x

LReasoning = []
LActivity = []
LCategorized = []
LTimed = []
LParent = []
LAmount = []
LWhole = []
for x in Timestamp :
    found = False
    for y in range(len(LTimed)) :
        if getHour(x) == LTimed[y] :
            found = True
    if found == False :
        LTimed.append(getHour(x))



LTimed_ = []
for x in L :
    found = False
    for y in range(len(LCategorized)) :
        if x == LCategorized[y] :
            found = True
    if found == False :
        for w in LTimed :
            LTimed_.append(w)
            LCategorized.append(x)
            LAmount.append(0)

LCategorized_ = []
for i in range(len(L)) :
    y = L[i]
    x = getHour(Timestamp[i])
    for j in range(len(LCategorized)) :
        if (y == LCategorized[j]) and (x == LTimed_[j]) :
            LAmount[j] += 1

for x in LCategorized :
    if (x == "BENIGN") :
        LParent.append("BENIGN")
    else :
        LParent.append("HOSTILE")

for x in LCategorized :
    if (x == "BENIGN") :
        LCategorized_.append(None)
    else :
        LCategorized_.append(x)

for x in LParent :
    LWhole.append("Traffic Activities from 8AM - 6PM")

sun_data = dict(A = np.array(LCategorized_),B = np.array(LParent), C=np.array(LTimed_), D =
np.array(LWhole), value = np.array(LAmount))

fig = px.sunburst( sun_data, path = ['D', 'C', 'B', 'A'], values = 'value', height= 1000,
width=1000, title="Hierarki Aktivitas Lalu Lintas di Hari Rabu 5 Juli 2017 Berdasarkan Keakitfan
Serangan pada Selang Waktu")
fig.show()
```

### *3. Plotting Relationships*

Dalam proses visualisasi data khususnya plotting relationship, masing-masing baris data dikelompokkan berdasarkan atribut ‘Label’ pada dataset. Kelompok pertama merupakan data dengan atribut Label bernilai “BENIGN” yang merujuk pada jaringan normal dan ditandai dengan plot berwarna biru (Benign) dan “tidak BENIGN” yang merujuk pada serangan dalam jaringan dan ditandai dengan plot berwarna merah (Attack). Data

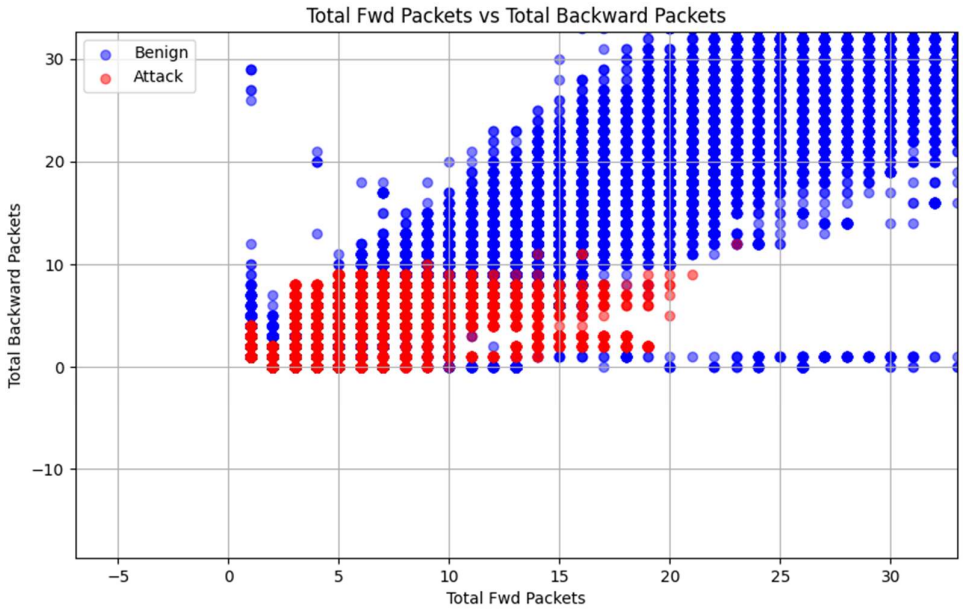
yang digunakan secara umum berasal dari file *csv* pada hari Rabu karena memiliki atribut label yang mengindikasi serangan jaringan dan penggunaan jaringan normal sekaligus. Proses visualisasi dilakukan dengan bantuan Python. Berikut merupakan hubungan-hubungan yang cukup mencolok dari berbagai atribut yang telah dikombinasikan.

**Bagan 6 – Scatter Plot *Total Forward Packets* dan *Total Backward Packets* pada dataset hari Rabu, 5 Juli 2017**




Plotting relationship yang pertama merujuk pada visualisasi hubungan *Total Forward Packets* dengan *Total Backward Packets*. *Total Forward Packets* merujuk pada seberapa banyak paket yang dikirim oleh *client* kepada *server*, sementara *Total Backward Packets* merujuk pada total paket jaringan yang dikirimkan oleh *server* kepada klien yang umumnya berupa *acknowledge* atau konfirmasi atas permintaan yang diajukan oleh *client*. Berdasarkan data tersebut didapatkan informasi bahwa, pada umumnya serangan jaringan dilaksanakan degan proporsi *Total Forward Packets* dan *Total Backward Packets* yang cenderung rendah dibandingkan dengan penggunaan jaringan normal pada umumnya. Pernyataan ini dapat dilihat lebih jelas melalui perbesaran berikut.

**Bagan 7 – *Zoom Out* Scatter Plot *Total Forward Packets* dan *Total Backward Packets* pada dataset hari Rabu, 5 Juli 2017**

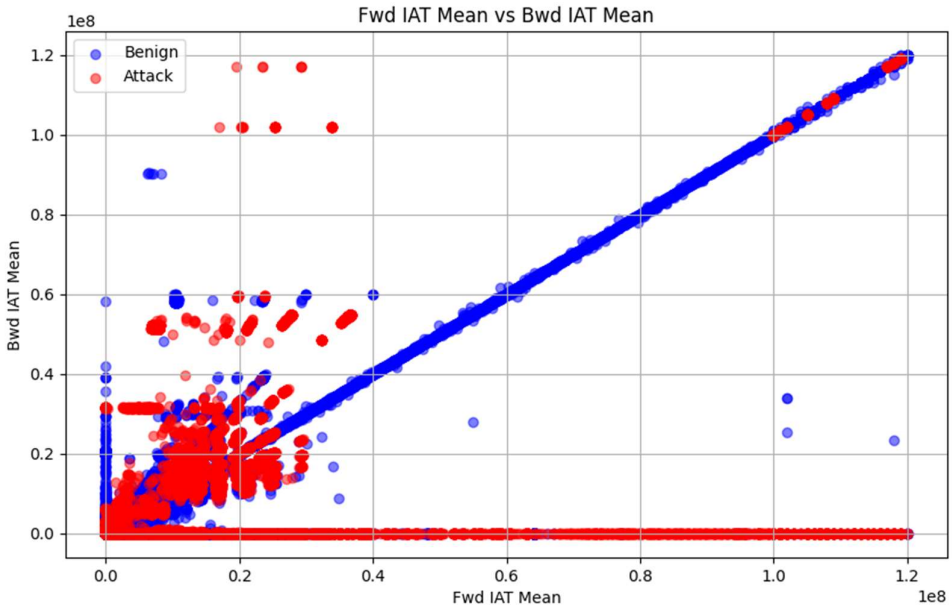



Dari data tersebut kita dapat melihat bahwa, pada umumnya serangan jaringan terjadi dengan proporsi *Total Forward Packets* yang berkisar pada 0-20 *packets* dan *Total Backward Packets* yang berkisar antara 0-10 *packets*. Walaupun demikian secara umum hubungan antara *Total Forward Packets* dengan *Total Backward Packets* memiliki korelasi positif.

Berikut kode python yang digunakan untuk visualisasi *Total Forward Packets* dan *Total Backward Packets.*

```
import pandas as pd
from matplotlib import pyplot as plt

wednesday_file = pd.read_csv('./TrafficLabelling/Wednesday-workingHours.pcap_ISCX.csv',
engine='python')

benign = wednesday_file[wednesday_file[' Label'] == 'BENIGN']
attack = wednesday_file[wednesday_file[' Label'] != 'BENIGN']

plt.figure(figsize=(10, 6))
plt.scatter(benign['Bwd Packet Length Max'], benign[' Fwd Packet Length Max'],
label='Benign', alpha=0.5, color='blue')
plt.scatter(attack['Bwd Packet Length Max'], attack[' Fwd Packet Length Max'],
label='Attack', alpha=0.5, color='red')
plt.xlabel('Bwd Packet Length Max')
plt.ylabel('Fwd Packet Length Max')
plt.title('Bwd Packet Length Max vs Fwd Packet Length Max')
plt.legend()
plt.grid(True)
plt.show()
```

**Bagan 8 – Scatter Plot *Forward IAT Mean* dan *Backward IAT Mean* pada dataset hari Rabu, 5 Juli 2017**



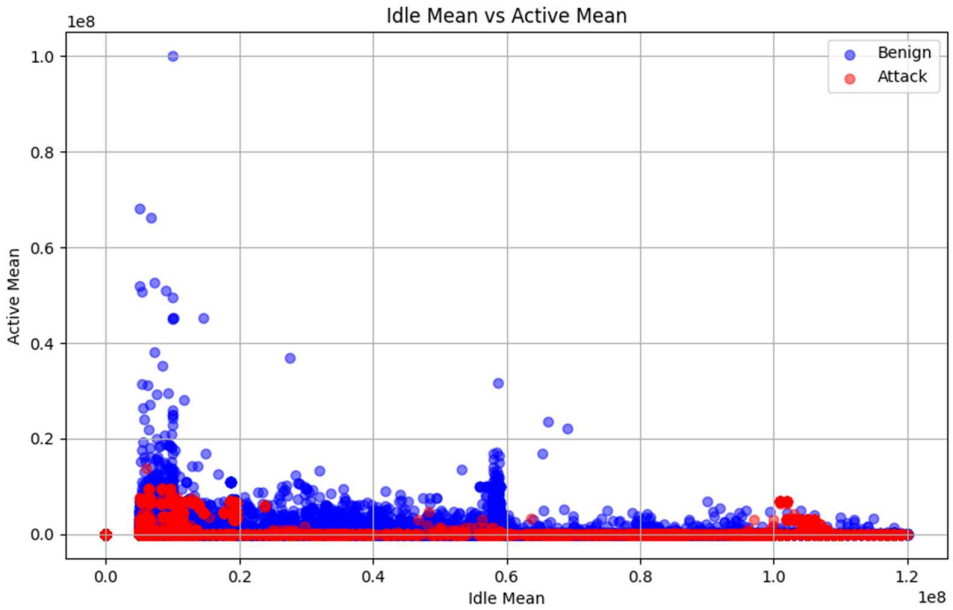

Plotting relationship yang kedua merujuk pada visualisasi antara *Forward IAT Mean* dengan *Backward IAT Mean. Forward IAT Mean* adalah waktu rata-rata (dalam *microsecond*) yang dibutuhkan oleh masing-masing *forward packets* dalam satu *flow* jaringan yang sama. Sementara *Backward IAT Mean* adalah waktu rata-rata yang dibutuhkan oleh masing-masing *backward packets* dalam satu *flow* jaringan.

Berdasarkan data tersebut terdapat hubungan yang unik antara *Forward IAT Mean* dan *Backward IAT Mean* pada label *Benign* dan *Attack.* Label *Benign* memiliki kecenderungan korelasi positif antara *Forward IAT Mean* dan *Backward IAT Mean* walaupun pada rentang awal distribusinya yang kurang merata. Namun pada label *Attack* memiliki waktu respon server (*Backward IAT Mean*) yang cenderung sangat rendah. Hal ini mengindikasi perilaku otomatis seperti *bot, bruteforce, DDoS attack,* dan lain sebagainya. Sehingga dapat disimpulkan bahwa apabila korelasi antara *Forward IAT Mean* dan *Backward IAT Mean* bernilai 0, besar kemungkinan terjadi serangan jaringan.

Berikut kode python yang digunakan pada visualisasi *Forward IAT Mean* dan *Backward IAT Mean.*

```
import pandas as pd
from matplotlib import pyplot as plt

wednesday_file    =    pd.read_csv('./TrafficLabelling/Wednesday-workingHours.pcap_ISCX.csv',
engine='python')

benign = wednesday_file[wednesday_file[' Label'] == 'BENIGN']
attack = wednesday_file[wednesday_file[' Label'] != 'BENIGN']

plt.figure(figsize=(10, 6))
plt.scatter(benign[' Fwd IAT Mean'], benign[' Bwd IAT Mean'], label='Benign', alpha=0.5,
color='blue')
plt.scatter(attack[' Fwd IAT Mean'], attack[' Bwd IAT Mean'], label='Attack', alpha=0.5,
color='red')
plt.xlabel('Fwd IAT Mean')
plt.ylabel('Bwd IAT Mean')
plt.title('Fwd IAT Mean vs Bwd IAT Mean')
plt.legend()
plt.grid(True)
plt.show()
```

**Bagan 9 – Scatter Plot *Idle Mean* dan *Active Mean* pada dataset hari Rabu, 5 Juli 2017**



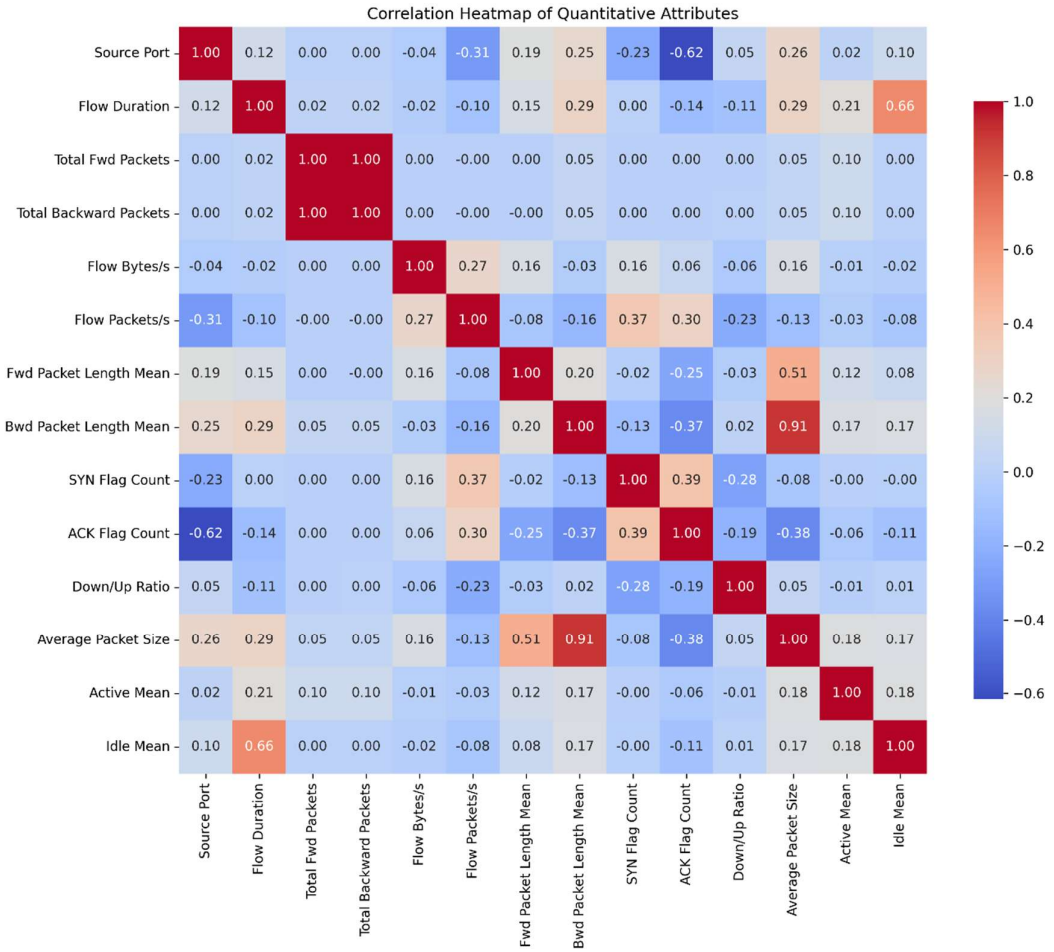

Plotting relationship yang ketiga memvisualisasikan hubungan antara *Idle Mean* dengan *Active Mean. Idle Mean* merujuk pada rata-rata lamanya aliran data terhenti/menganggur dari satu *flow* jaringan, sedangkan *Active Mean* adalah rata-rata lamanya waktu aktif dalam satu *flow* jaringan. Informasi yang bisa didapatkan dari visualisasi tersebut adalah secara umum, serangan jaringan terjadi pada *Active Mean* yang mendekati nol. Dengan kata lain semakin dekat *Active Mean* suatu jaringan dengan nol, semakin besar kemungkinan terjadi serangan jaringan.

Berikut kode python yang digunakan dalam visualisasi korelasi antara *Idle Mean* dengan *Active Mean.*

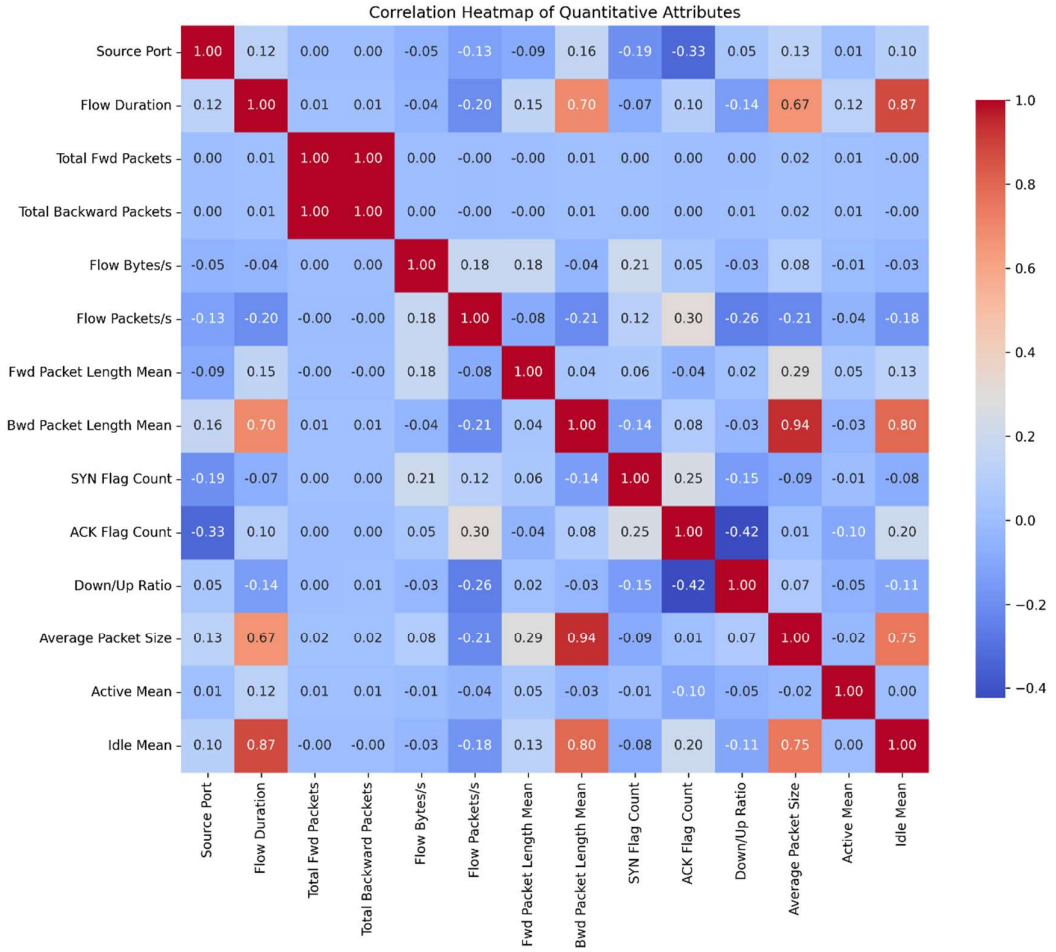
```
import pandas as pd
from matplotlib import pyplot as plt

wednesday_file     =     pd.read_csv('./TrafficLabelling/Wednesday-workingHours.pcap_ISCX.csv',
engine='python')

benign = wednesday_file[wednesday_file[' Label'] == 'BENIGN']
attack = wednesday_file[wednesday_file[' Label'] != 'BENIGN']

plt.figure(figsize=(10, 6))
plt.scatter(benign['Idle Mean'], benign['Active Mean'], label='Benign', alpha=0.5, color='blue')
plt.scatter(attack['Idle Mean'], attack['Active Mean'], label='Attack', alpha=0.5, color='red')
plt.xlabel('Idle Mean')
plt.ylabel('Active Mean')
plt.title('Idle Mean vs Active Mean')
plt.legend()
plt.grid(True)
plt.show()
```

# BAB 3 – PENGOLAHAN DATA

## A. Korelasi Data




**Bagan 10 – Korelasi Data Set 1**




**Bagan 3 11 – Korelasi Data Set 2**

Peta korelasi atribut kuantitatif pada dataset di atas menunjukkan hubungan linier antar variabel numerik yang dihitung menggunakan koefisien korelasi Pearson. Secara umum, korelasi dengan nilai mendekati +1 menunjukkan hubungan positif yang kuat, sementara nilai mendekati -1 menunjukkan hubungan negatif yang kuat. Dalam heatmap ini, ditemukan beberapa hubungan yang sangat signifikan. Misalnya, atribut Bwd Packet Length Mean memiliki korelasi sangat tinggi dengan Average Packet Size (0.94) dan Idle Mean (0.80), menandakan bahwa saat panjang rata-rata paket backward meningkat, ukuran rata-rata paket dan waktu idle juga cenderung meningkat. Selain itu, Flow Duration juga menunjukkan korelasi kuat dengan Idle Mean (0.87), Bwd Packet Length Mean (0.70), dan Average Packet Size (0.67), yang mengindikasikan bahwa aliran data yang berlangsung lebih lama berkaitan dengan ukuran dan waktu idle yang lebih besar.

Sebaliknya, terdapat juga korelasi negatif yang cukup kuat, seperti antara ACK Flag Count dan Down/Up Ratio (-0.42), serta antara Source Port dan ACK Flag Count (-0.33). Hal ini menunjukkan bahwa peningkatan jumlah flag ACK cenderung terjadi ketika rasio aliran data menurun atau ketika port tertentu digunakan. Di sisi lain, beberapa atribut seperti Total Fwd Packets, Total Backward Packets, dan Flow Bytes/s tampak memiliki korelasi yang rendah dengan sebagian besar atribut lain, yang menunjukkan bahwa mereka mungkin menyimpan informasi unik dan tidak redundant.

Berdasarkan hasil ini, analisis lanjutan dapat mempertimbangkan teknik reduksi dimensi atau pemilihan fitur (feature selection), terutama pada atribut yang memiliki korelasi sangat tinggi satu sama lain agar menghindari redundansi data. Sementara itu, atribut yang tidak terlalu berkorelasi bisa menjadi kandidat yang baik untuk fitur-fitur baru atau input pada model prediksi. Selain itu, pola korelasi seperti hubungan antara durasi aliran, jumlah flag ACK, dan rasio arah lalu lintas data dapat dimanfaatkan dalam deteksi anomali atau identifikasi serangan siber yang mungkin terjadi dalam jaringan.

**B. Data Cleansing**

| Atribut yang Dihapus | Syarat | Alasan |
|---|---|---|
| Flow Duration | Nilai $\leq 0$ | Flow duration tidak mungkin dibawah atau sama dengan 0 karena merupakan interval waktu |
| Total Fwd Packets | Nilai $\leq 0$ | Total Fwd Packets tidak mungkin negatif karena merupakan jumlah |
| Total Backward Packets | Nilai $\leq 0$ | Total Backward Packets tidak mungkin negatif karena merupakan jumlah |
| Average Packet Size | Nilai $\leq 0$ | Ukuran Packet size tidak mungkin 0 karena dapat menyebabkan data error |
| Active Mean | Nilai $\leq 0$ | Waktu aktif harus positif, nilai $\leq 0$ menunjukkan data tidak valid |
| Idle Mean | Nilai $\leq 0$ | Waktu idle tidak boleh nol atau negatif |

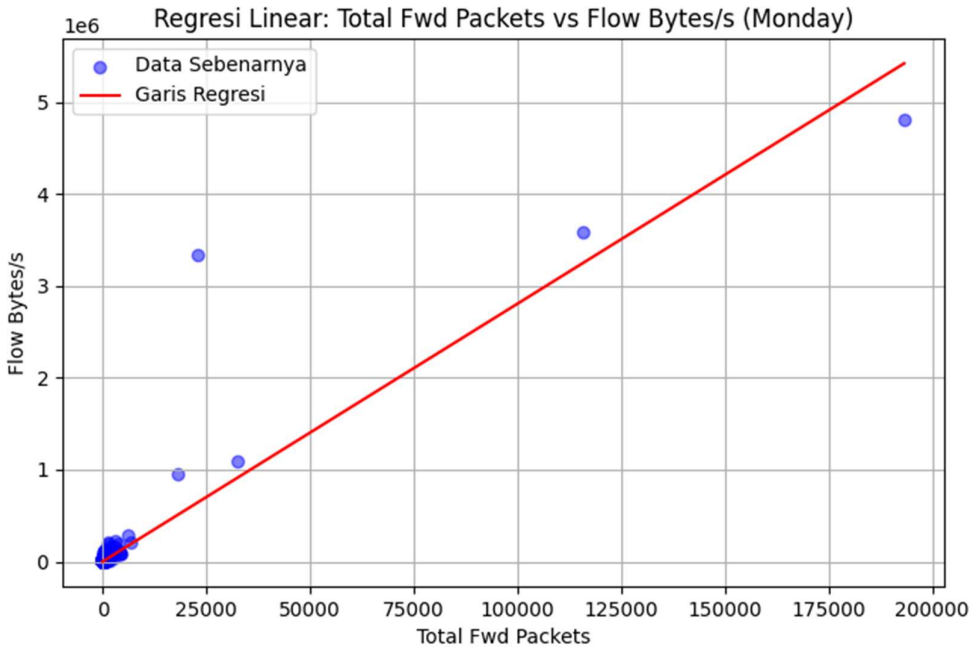
```
import pandas as pd
import numpy as np

input_csv = '../data/Monday-filtered.csv'
output_csv = '../data/Monday-cleaned.csv'

df = pd.read_csv(input_csv, low_memory=False)

numeric_cols = [
    'Flow Duration',
    'Total Fwd Packets',
    'Total Backward Packets',
    'Flow Bytes/s',
    'Flow Packets/s',
    'Fwd Packet Length Mean',
    'Bwd Packet Length Mean',
    'Average Packet Size',
    'Active Mean',
    'Idle Mean',
    'SYN Flag Count',
    'ACK Flag Count',
    'Down/Up Ratio',
    'Source Port'  # Port numbers >= 0
]

for col in numeric_cols:
    df[col] = pd.to_numeric(df[col], errors='coerce')

df = df.dropna(subset=numeric_cols)

for col in numeric_cols:
    df = df[df[col] >= 0]

df['Label'] = df['Label'].str.strip().str.upper()

df = df[df['Label'] != '']

df.to_csv(output_csv, index=False)
```

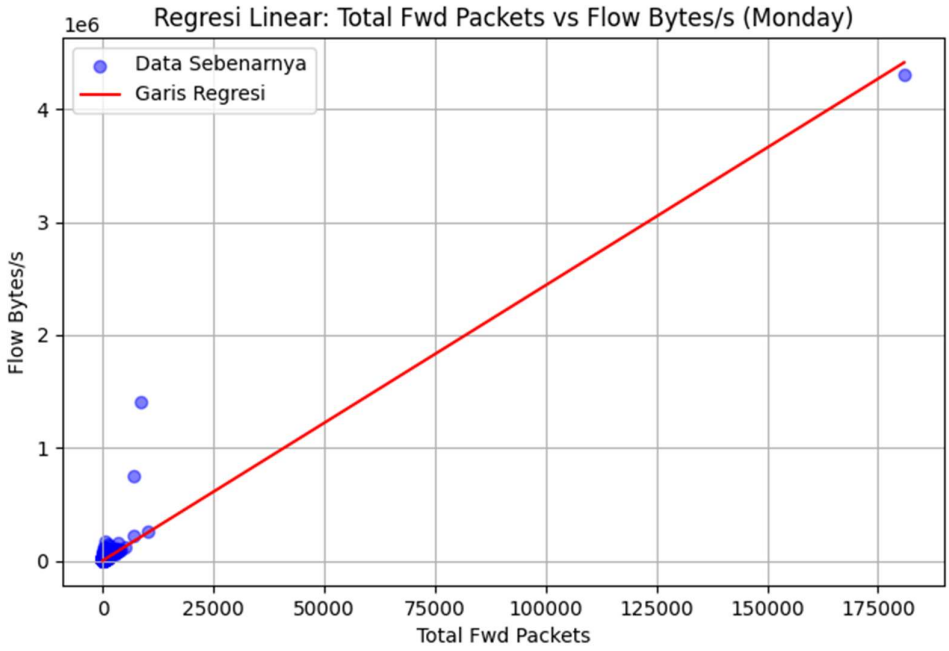
```
print(f"[INFO] Cleaned dataset saved to '{output_csv}'")
```

## C. TRANSFORMASI DATA

Tranformasi data dilakukan pada dataset Monday-WorkingHours.pcap_ISCX.csv khususnya pada atribut " Down/Up Ratio". Tipe data yang digunakan pada dataset tersebut adalah float64 sementara pada dataset Wednesday-WorkingHours.pcap_ISCX.csv digunakan tipe data int64 pada atribut yang sama. Walaupun menggunakan tipe data float64, dataset Monday-WorkingHours.pcap_ISCX.csv tetap menyimpan hasil berupa bilangan dengan desimal nol. Oleh karena itu dilakukan perubahan tipe data dari float64 menjadi int64 untuk mendukung konsistensi tipe data pada kedua dataset.

## D. Analisis Data Sederhana




Regresi Linear: Total Fwd Packets vs Flow Bytes/s (Monday)
1e6
Data Sebenarnya
Garis Regresi
Flow Bytes/s
Total Fwd Packets
0
1
2
3
4
0
25000
50000
75000
100000
125000
150000
175000

```
IMPORT PANDAS AS PD

IMPORT NUMPY AS NP

FROM SKLEARN.LINEAR_MODEL IMPORT LINEARREGRESSION

IMPORT MATPLOTLIB.PYPLOT AS PLT

FILE_MONDAY = R"WEDNESDAY-WORKINGHOURS.PCAP_ISCX.CSV"

DF = PD.READ_CSV(FILE_MONDAY)

DF.COLUMNS = DF.COLUMNS.STR.STRIP()

X = DF[['TOTAL FWD PACKETS', 'FLOW DURATION']]

Y = DF['FLOW BYTES/S']

MODEL = LINEARREGRESSION()

MODEL.FIT(X, Y)

PLT.FIGURE(FIGSIZE=(8,5))

PLT.SCATTER(X['TOTAL FWD PACKETS'], Y, COLOR='BLUE', ALPHA=0.5, LABEL='DATA
SEBENARNYA')



# PREDIKSI UNTUK MEMBUAT GARIS REGRESI

X_LINE = NP.LINSPACE(X['TOTAL FWD PACKETS'].MIN(), X['TOTAL FWD PACKETS'].MAX(), 100)

X_PRED = PD.DATAFRAME({'TOTAL FWD PACKETS': X_LINE, 'FLOW DURATION': NP.MEAN(X['FLOW
DURATION'])})

Y_PRED = MODEL.PREDICT(X_PRED)



PLT.PLOT(X_LINE, Y_PRED, COLOR='RED', LABEL='GARIS REGRESI')

PLT.XLABEL('TOTAL FWD PACKETS')

PLT.YLABEL('FLOW BYTES/S')

PLT.TITLE('REGRESI LINEAR: TOTAL FWD PACKETS VS FLOW BYTES/S (MONDAY)')

PLT.LEGEND()

PLT.GRID(TRUE)
```

```
PLT.SHOW()
```

# BAB 4 – KESIMPULAN

Berdasarkan analisis data sederhana yang telah dilakukan menggunakan model regresi linear, dapat disimpulkan bahwa:

1. **Hubungan Linear yang Kuat**
   Terdapat hubungan linier yang cukup kuat antara variabel Total Fwd Packets dan Flow Duration dengan Flow Bytes/s. Hal ini menunjukkan bahwa jumlah paket forward dan durasi koneksi berpengaruh terhadap kecepatan alir data dalam jaringan.

2. **Model Regresi Linear Sederhana Efektif**
   Model regresi linear sederhana yang dibangun mampu memprediksi Flow Bytes/s dengan cukup baik, meskipun menggunakan variabel independen terbatas. Model ini memberikan pemahaman awal tentang pola lalu lintas jaringan.

3. **Potensi Deteksi Anomali**
   Dengan memanfaatkan model regresi linear, dapat dikembangkan metode sederhana untuk mendeteksi anomali jaringan. Perbedaan signifikan antara prediksi dan realisasi Flow Bytes/s bisa menjadi indikasi awal adanya potensi serangan siber.

4. **Kebutuhan Pengembangan Lanjutan**
   Analisis sederhana ini perlu dikembangkan lebih lanjut dengan menambahkan variabel independen lain yang relevan, misalnya SYN Flag Count, ACK Flag Count, dan Down/Up Ratio, untuk meningkatkan akurasi prediksi dan mendeteksi pola serangan yang lebih kompleks.