DataCraft Trio

Debnath Kundu (MT22026) , Pijush Bhuyan (MT22049), Snehal Buldeo (MT22074)

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

Data Description

- Dataset URL - UCI Repository [detection_of_IoT_botnet_attacks_N_BaIoT]

- This public dataset contains real-time network traffic data recorded from nine commercial IOT devices that were infected with common botnet malware - **BASHLITE** and **MIRAI** to carry out ten types of network-based attacks.

- **Types of IOT devices:** Thermostat, Baby Monitor, Webcam, Doorbells, and Security Cameras.

Types of Attacks

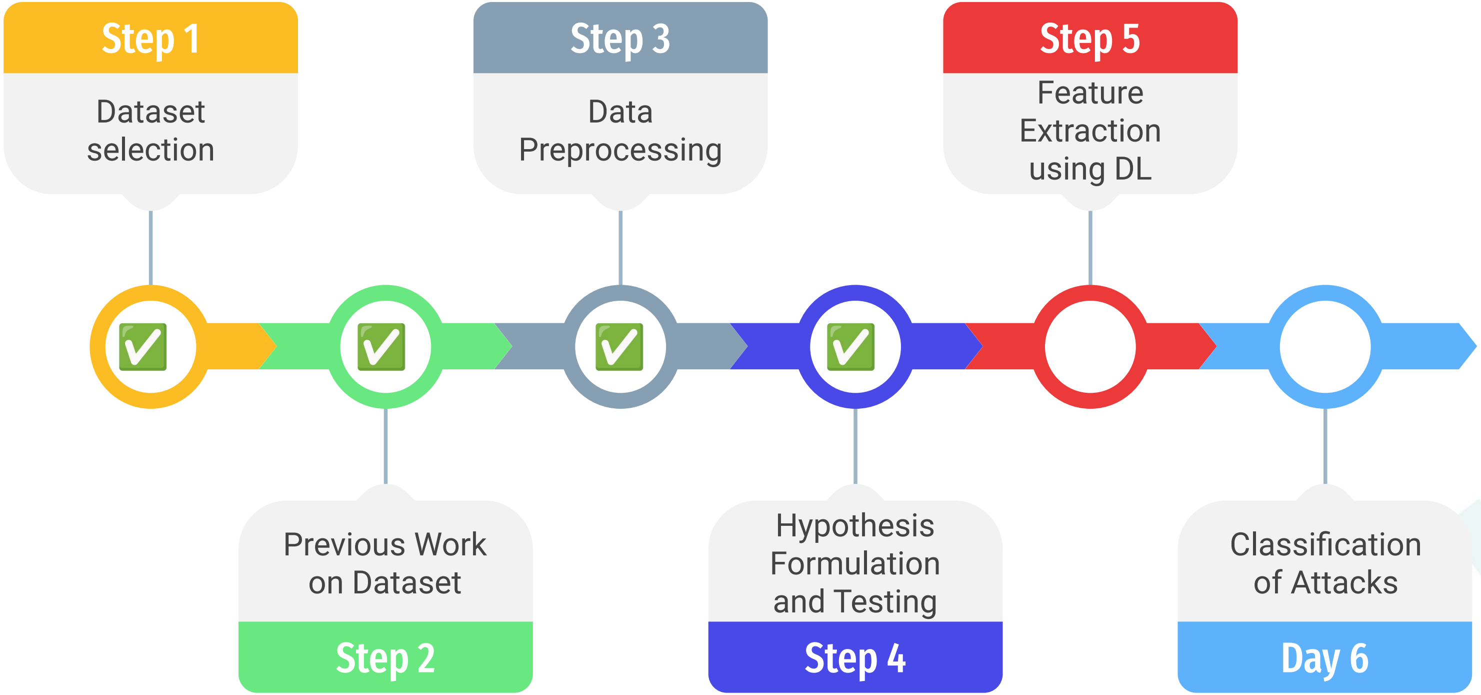
1. **BASHLITE MALWARE-MIRAI MALWARE -**

- Scan - Scanning the network for vulnerable devices
- Junk - Sending spam data packets
- UDP - Flooding the network with UDP packets
- TCP - Flooding the network with TCP packets
- Combo - Sending spam data and opening a connection to a specified IP address and

2. **MIRAI MALWARE -**

- Scan - Automatic scanning for vulnerable devices
- Ack - Flooding the network with Ack packets
- Syn - Flooding the network with Syn packets
- UDP - Flooding the network with UDP packets
- UDP Plain - UDP flooding with fewer options, optimized for higher PPS

Plan of Action and Progress So Far



Step 1
Dataset selection
✅

Step 2
Previous Work on Dataset
✅

Step 3
Data Preprocessing
✅

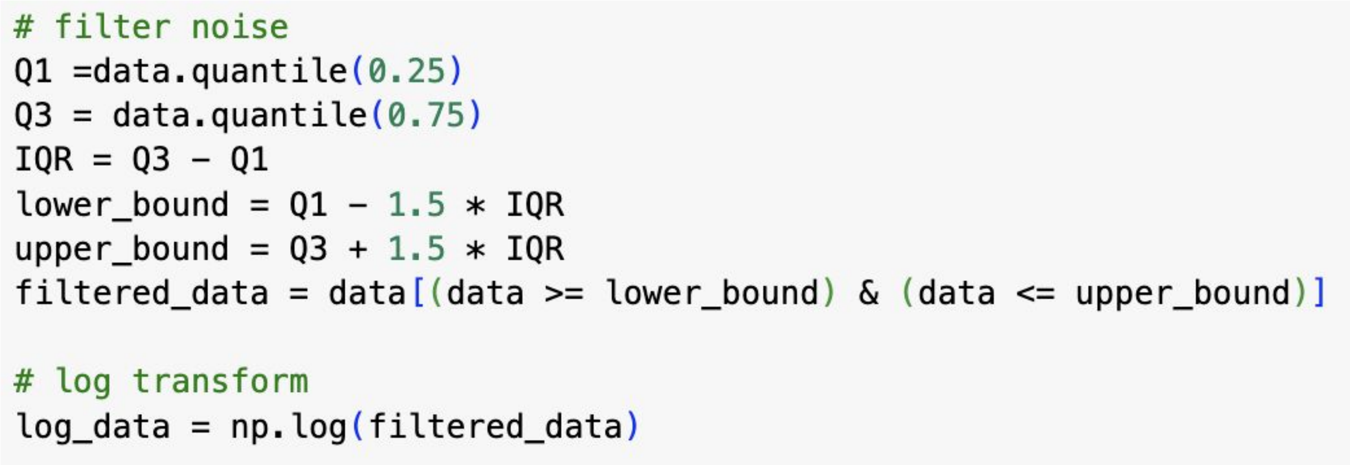
Step 4
Hypothesis Formulation and Testing
✅

Step 5
Feature Extraction using DL

Day 6
Classification of Attacks

Data Preprocessing

- Each one sample had outliers at the extremities
- Removed such extreme values using **InterQuartile Range (IQR)**

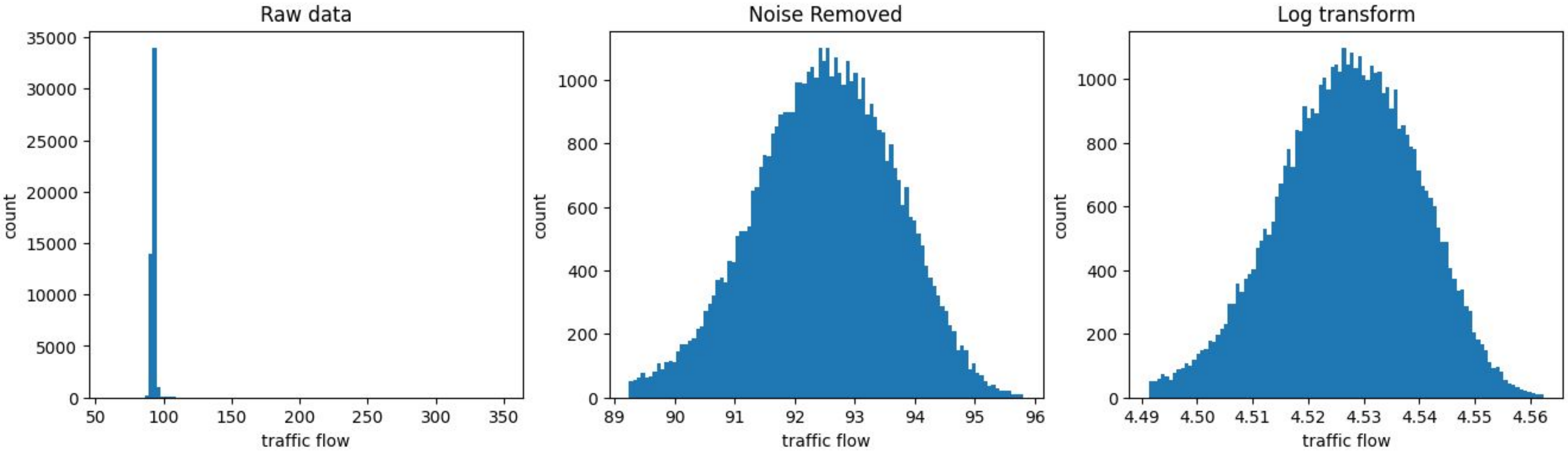


```
# filter noise
Q1 =data.quantile(0.25)
Q3 = data.quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
filtered_data = data[(data >= lower_bound) & (data <= upper_bound)]

# log transform
log_data = np.log(filtered_data)
```

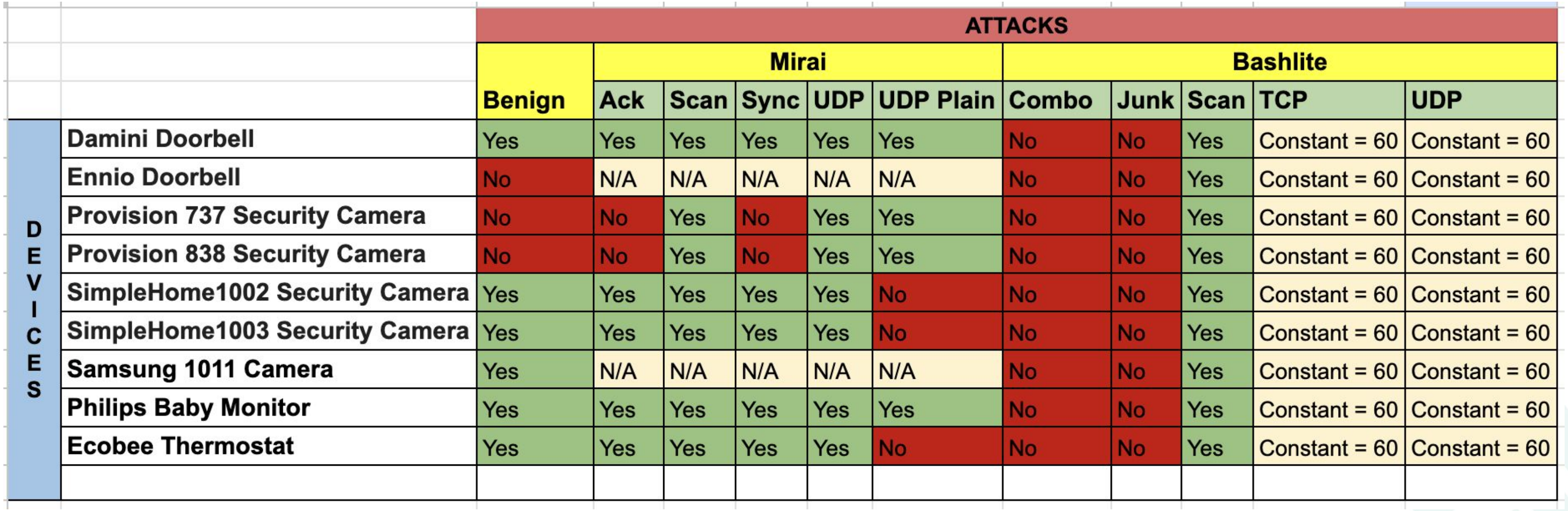
Data Preprocessing

- Sample distribution after noise removal



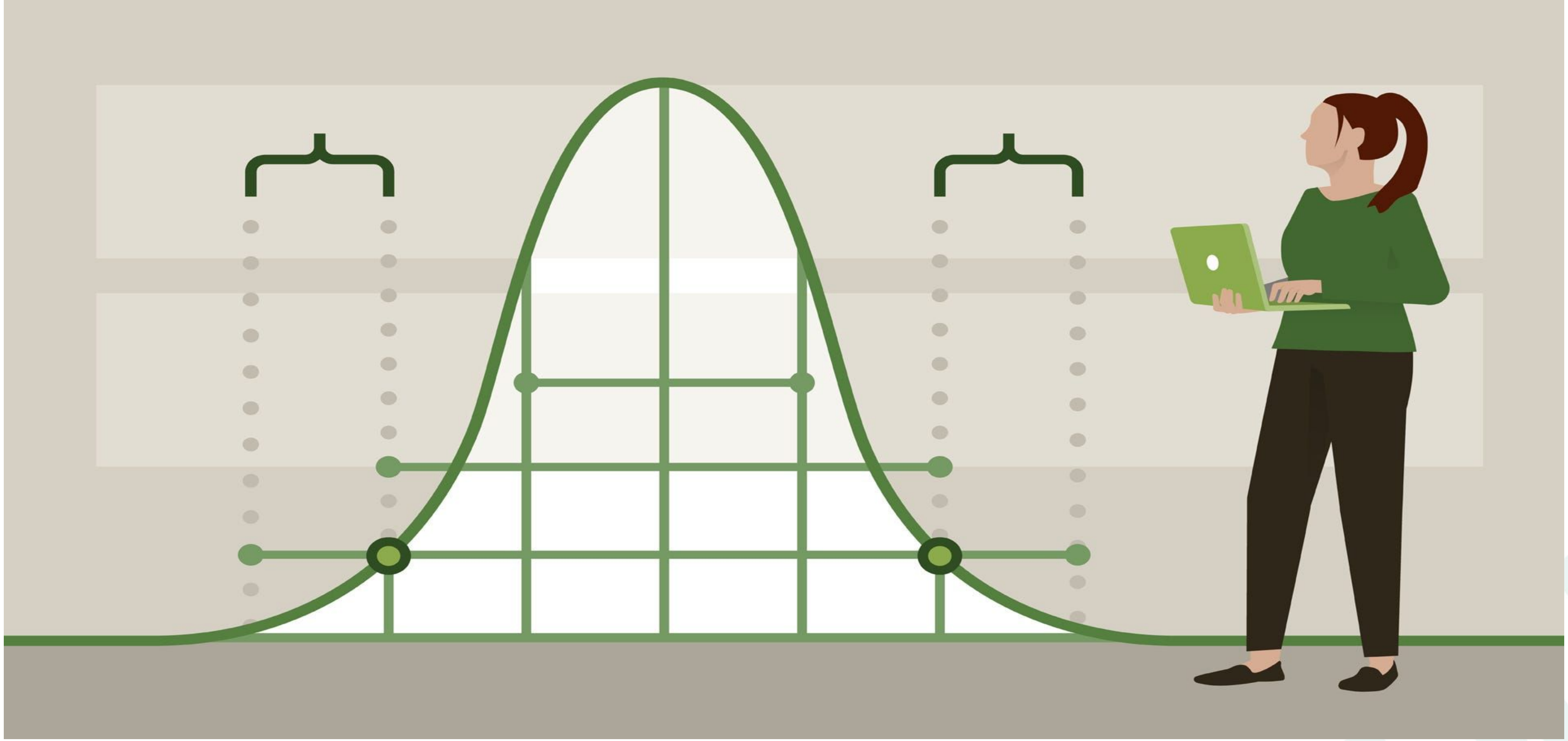
Attacks Distribution

- Mirai Attacks **not** present in **Emnio Doorbell & Samsung 1011 Camera.**

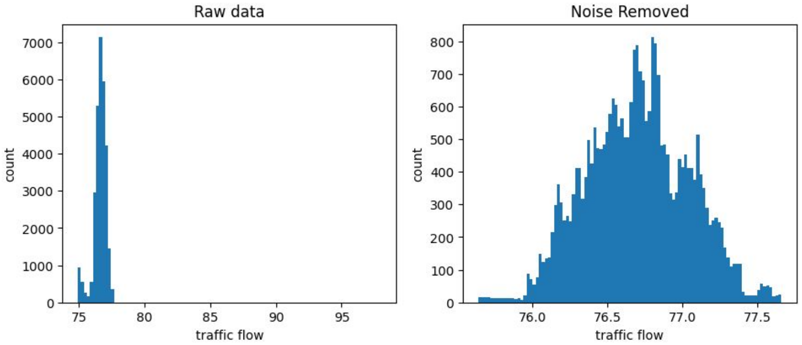


		ATTACKS										
			Mirai					Bashlite				
		Benign	Ack	Scan	Sync	UDP	UDP Plain	Combo	Junk	Scan	TCP	UDP
DEVICES	Damini Doorbell	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Constant = 60	Constant = 60
	Ennio Doorbell	No	N/A	N/A	N/A	N/A	N/A	No	No	Yes	Constant = 60	Constant = 60
	Provision 737 Security Camera	No	No	Yes	No	Yes	Yes	No	No	Yes	Constant = 60	Constant = 60
	Provision 838 Security Camera	No	No	Yes	No	Yes	Yes	No	No	Yes	Constant = 60	Constant = 60
	SimpleHome1002 Security Camera	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Constant = 60	Constant = 60
	SimpleHome1003 Security Camera	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Constant = 60	Constant = 60
	Samsung 1011 Camera	Yes	N/A	N/A	N/A	N/A	N/A	No	No	Yes	Constant = 60	Constant = 60
	Philips Baby Monitor	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Constant = 60	Constant = 60
	Ecobee Thermostat	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Constant = 60	Constant = 60

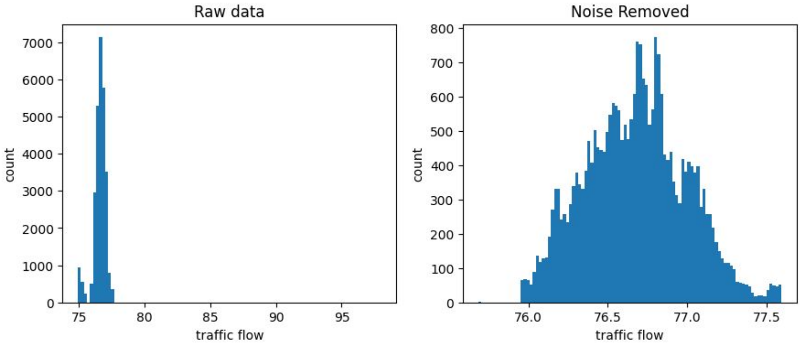
Hypothesis Testing and Validation



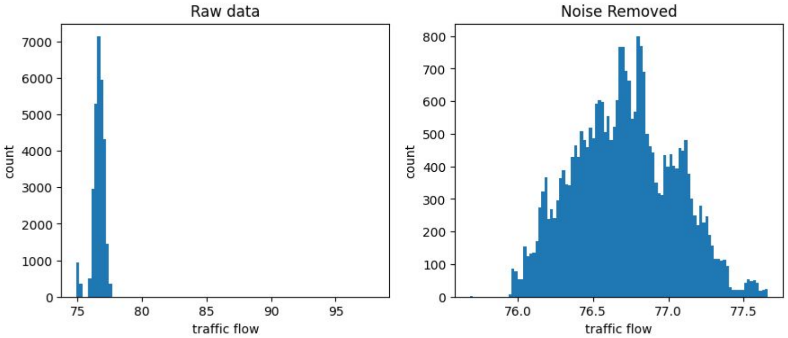
Data Distribution :: Bashlite Scan Attacks



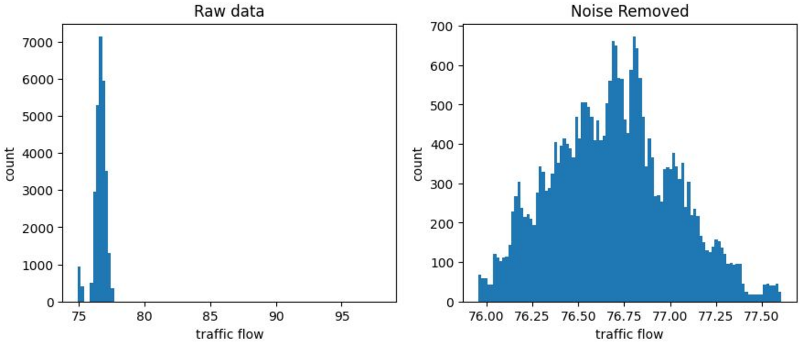
Danmani Doorbell



Ennio Doorbell

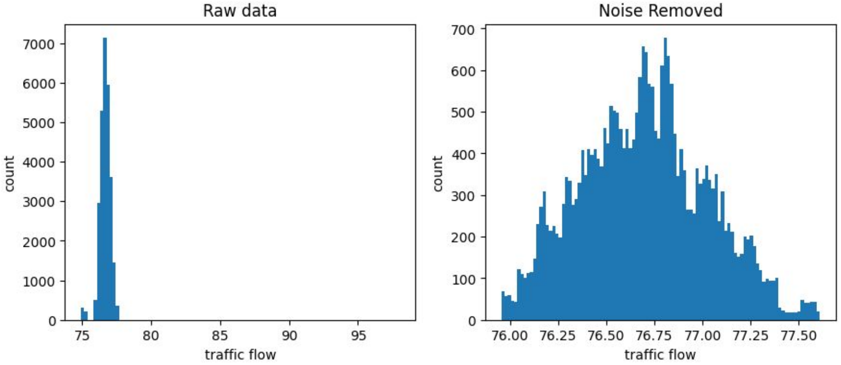


Provision PT 737E Security Camera

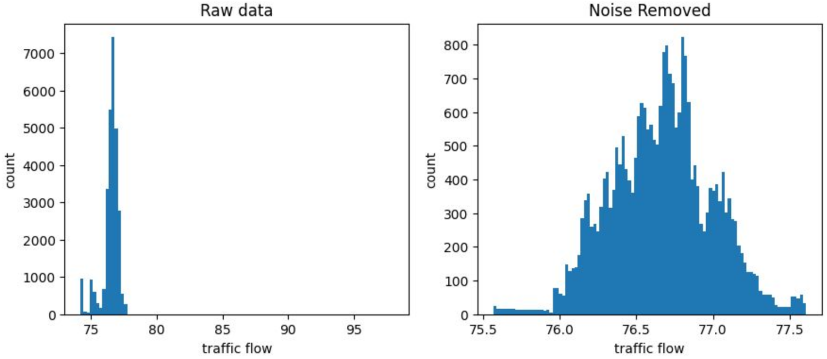


Provision PT 838 Security Camera

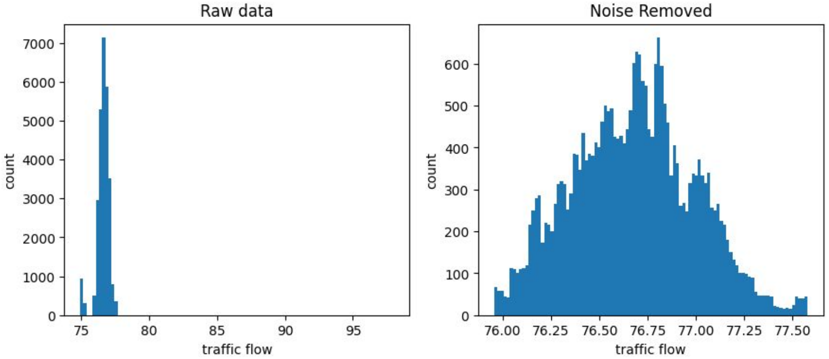
Data Distribution :: Bashlite Scan Attacks



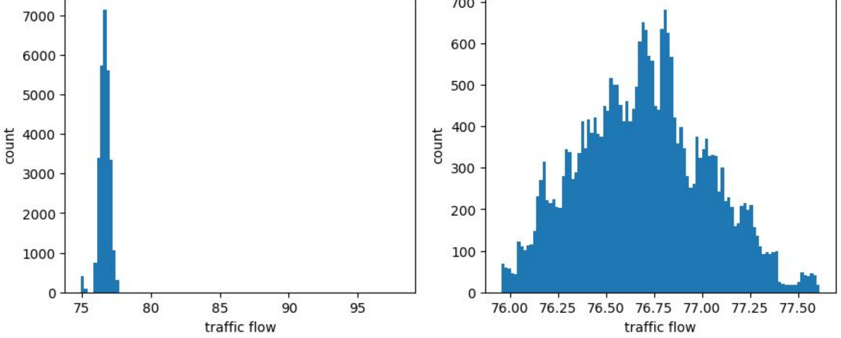
SimpleHome XCS7 1003 WHT Security Camera



SimpleHome XCS7 1003 WHT Security Camera

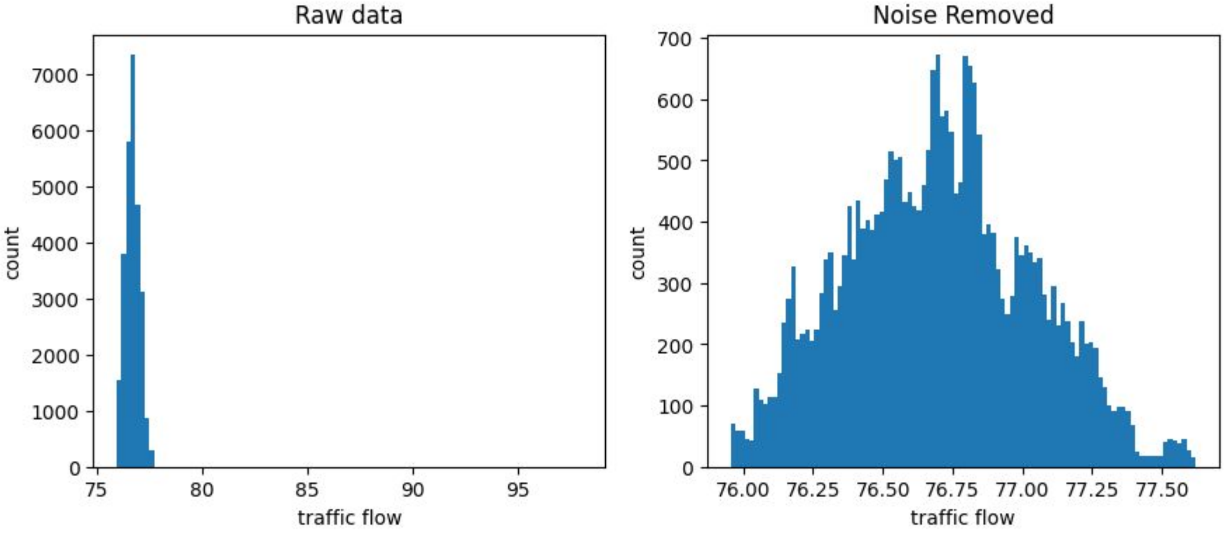


Samsung SNH 1011 N Webcam



Philips B120N10 Baby Monitor

Data Distribution :: Bashlite Scan Attacks



Ecobee Thermostat

Hypothesis Formulation :: 1 (Bashlite)

- **H_0 :** bashlite scan attacks have similar packet flow across all devices

  i.e. $\mu_1 = \mu_2 = \dots \mu_9$

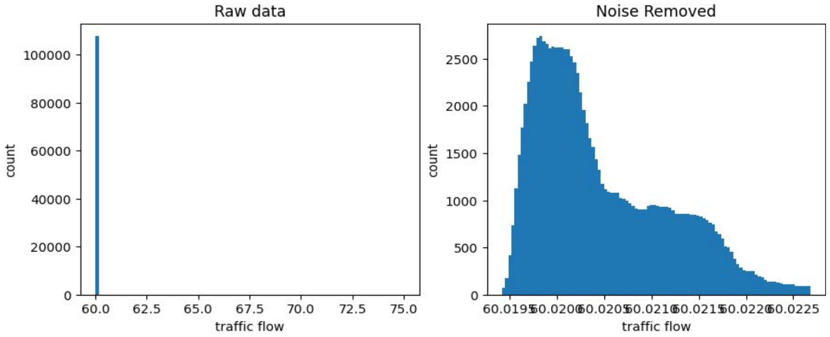
- **H_a :** the packet flow during a bashlite scan attack differs in at least one device
- Test Used : **ANOVA** (Analysis of Variance)
- **Sample Size :** 50 samples with random sampling
- **Assumptions :** normally distributed, equal variances and independent samples

Hypothesis Formulation - 1 (Contd)

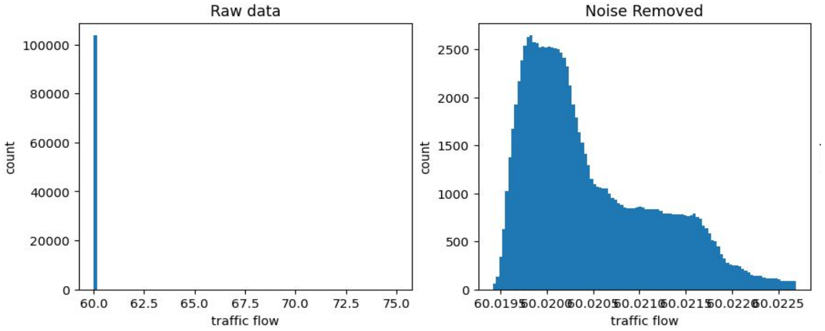
- P-value obtained = **$9e^{-4}$ << 0.05**
- Since p-value is less than level of significance, we reject the null hypothesis.
- Note - If we remove device no 6, P-value obtained = **0.863 >> 0.05**
- In such a case we **do not** reject the null hypothesis.
- Validation - (comparing the population mean packet flows)

Device	1	2	3	4	5	7	8	9
Mean(μ)	76.611559	76.589637	76.638644	76.620499	76.670631	76.611498	76.672633	76.704413
Population size	29849	28120	29297	28397	27825	27698	27859	27494

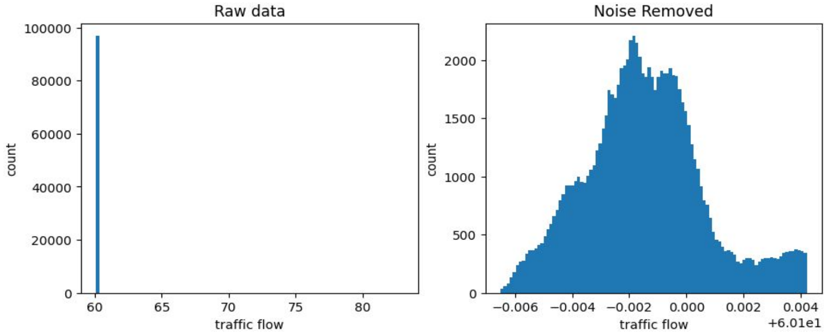
Data Distribution :: Mirai Scan Attacks



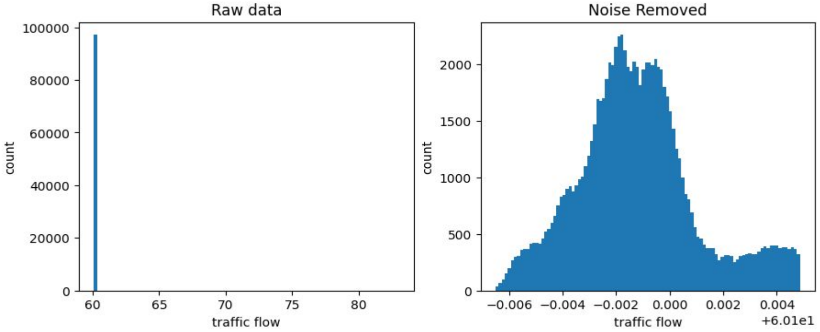
Danmini Doorbell



Philips B120N10 Baby Monitor

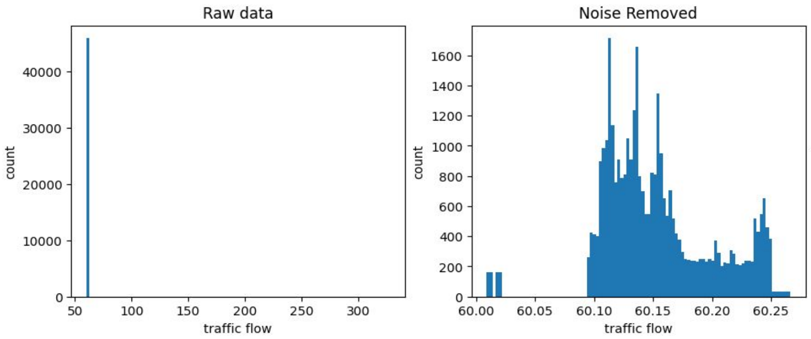


Provision PT 737E Security Camera

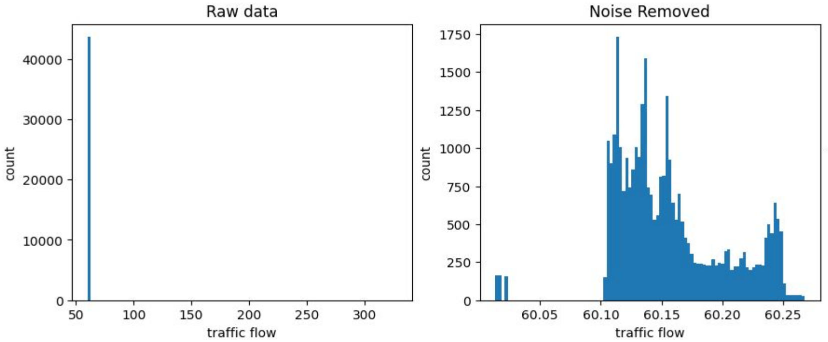


Provision PT 838 Security Camera

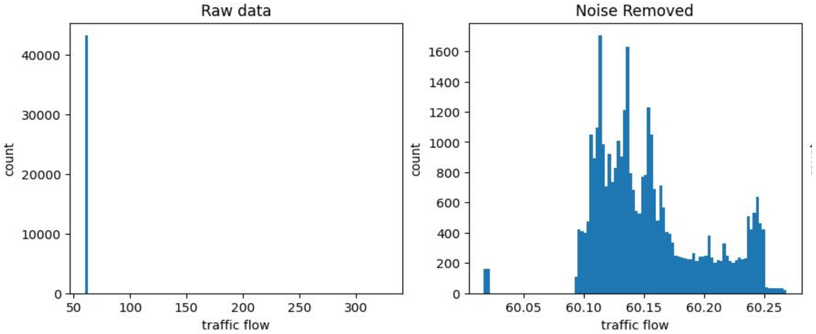
Data Distribution :: Mirai Scan Attacks



SimpleHome XCS7 1003 WHT Security Camera



SimpleHome XCS7 1003 WHT Security Camera



Ecobee Thermostat

Hypothesis Formulation - 2 (Mirai)

- **H_0 :** mirai scan attacks have similar packet flow for **Danmani Doorbell** and **Philips Baby Monitor**

  i.e. $\mu_1 = \mu_2$

- **H_a :** Mirai scan attacks don’t have similar packet flow for **Danmani Doorbell** and **Philips Baby Monitor**

- **Test Used :** T-test
- **Sample Size :** 50 samples with random sampling

- **Assumptions :** normally distributed, equal variances and independent samples

Hypothesis Formulation - 2 (Contd)

- P-value obtained = **0.1955 >> 0.05**

- Since P-value is more than level of significance, we **don’t** reject the null hypothesis.

- Validation - (comparing the population mean packet flows)

Device	1	6
Mean(μ)	60.020962	60.020907
Population Size	107685	103621

Hypothesis Formulation - 3 (Mirai)

- H_0 : mirai scan attacks have similar packet flow for both the **Provision camera models**

  i.e. $\mu_1 = \mu_2$

- H_a : mirai scan attacks don’t have similar packet flow for both the **Provision camera models**
- **Test Used :** T-test
- **Sample Size :** 50 samples with random sampling
- **Assumptions :** normally distributed, equal variances and independent samples

Hypothesis Formulation - 3 (Contd)

- P-value obtained = **0.1552 >> 0.05**
- Since p-value is more than level of significance, we **<u>don’t</u>** reject the null hypothesis.
- Validation - (comparing the population mean packet flows)

Device	2	3
Mean(μ)	60.099604	60.100516
Population Size	96781	97096

Hypothesis Formulation - 4 (Mirai)

- **H_0 :** mirai scan attacks have similar packet flow across the **SimpleHome cameras** and **Ecobee thermostat devices**

  i.e. $\mu_1 = \mu_2 = \mu_3$

- **H_a :** the packet flow during a bashlite scan attack differs in at least one device
- **Test Used :** ANOVA (Analysis of Variance
- **Sample Size :** 50 samples with random sampling
- **Assumptions :** normally distributed, equal variances and independent samples

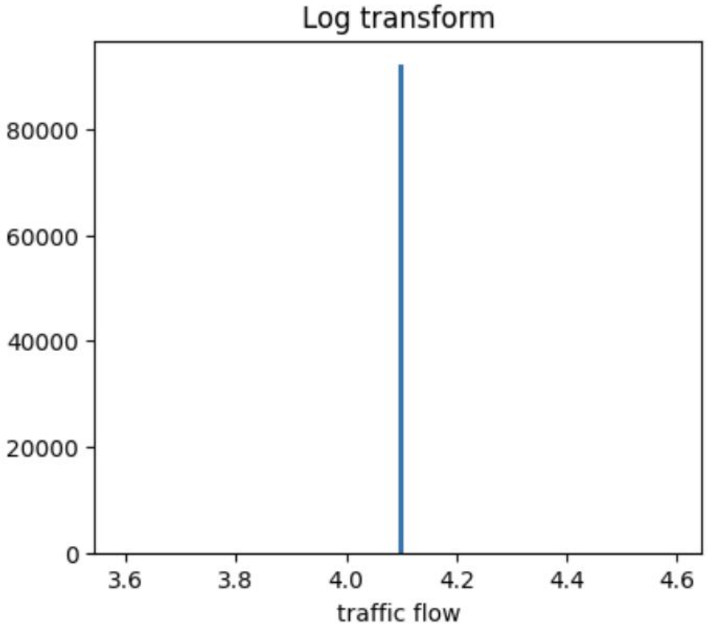
Hypothesis Formulation - 4 (Contd)

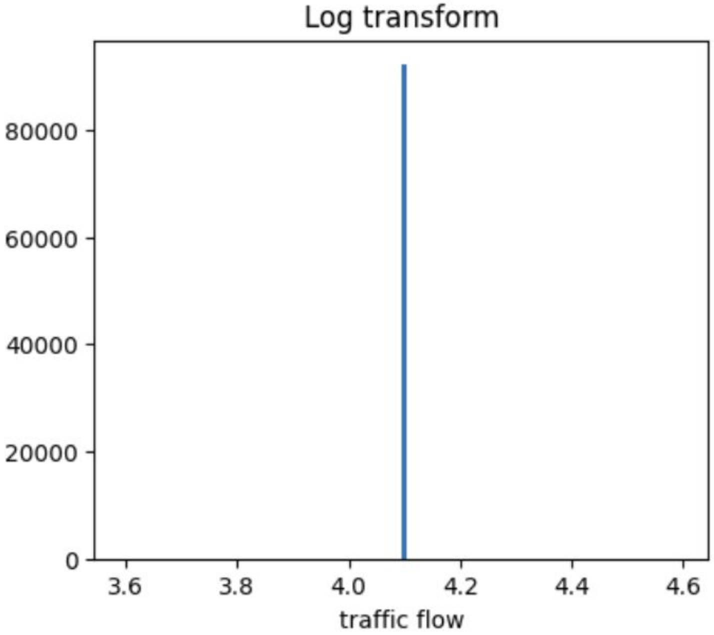
- P-value obtained = **0.3458 >> 0.05**
- Since P-value is more than level of significance, we don’t reject the null hypothesis.
- Validation - (comparing the population mean packet flows)

Device	4	5	7
Mean(μ)	60.151278	60.155650	60.161105
Population Size	45930	43674	43192

Other findings, so far

- Bashlite **TCP** and **UDP** attacks have a constant traffic flow value and follows no distribution across all devices.





Thank You