Santa Clara
Leavey School of Business

ISBA-2401 - Data Analytics with Python (Winter 2025)

Instructor & Project guide: Yu-Wei Lin, Email: ylin12@scu.edu

Date : 03/11/2025

Project Team (Team 4): **H**imanshu Jangra
Elizabeth Nguyen
Rohit Agarwal
Oliver Gonsalves

About the Dataset

Data analytics with Python - Cybersecurity Dataset

What is cybersecurity?

Cybersecurity is the protection of computer software, systems and networks from threats that can lead to unauthorized information disclosure, theft or damage to hardware, software, or data, as well as from the disruption or misdirection of the services they provide

Why do we need it?

Attributed to its definition, cyber security is important to protect devices, data, hardware and softwares (indirectly protect the owners of these devices)

what is team **HERO** doing in the field of cybersecurity?

We are building a persona of a Security Operations Manager/Leader (referred as **SecOps**) division in Acme India Pvt. Ltd.

- As a SecOps manager/leader, **what are hotspots in client-base and how many attacks were made and what devices were impacted?**
- As a SecOps manager/leader, **can we predict the type of protocol being used in the traffic with anomalous behavior and correlate that with the most prevalent attack?**
- As a SecOps manager/leader, **how can we bolster the security in our systems?**

Santa Clara
Leavey School of Business

Data Cleaning Techniques:

String type transformation

- **Geo-location Data**:
  - Breakdown City vs State into different columns
  - Original column Geo-Location Data series with values in the form **<city_name>, <state_name>**.
  - The cleaning process involves splitting the column into two **new columns** called ***location_city*** and **location_state** individually.
- **Device information**:
  - Blurred the values of the columns and transformed it to categorical column based on string operations to match the requirements of our finding.
  - New values are **Laptop/Desktop** and **Mobile/Tablet**
  - Created new column called 'device_types': Laptop/Desktop and Mobile/Tablet.
  - Dropped the original column 'Device Information'
- **Removed 8 columns** from original dataset as they were not used in the findings:
  - Log Source, Proxy Information, Firewall Logs, Payload Data, Malware Indicators, Packet Length, Source Port and Destination Port
- **Column Name processing**:
  - Change column names to all lower case, and spaces, '-' and '/' replaced with '_'
  - Change timestamp to date only values

Santa Clara Leavey School of Business

Data Cleaning Techniques:

Categorical column

- **Traffic type**: **HTTP, FTP, DNS**
  - Created categorical columns using get_dummies: traffic_type_dns, traffic_type_http, traffic_type_ftp
- **Protocol:** **TCP, UDP, ICMP**
  - Created categorical columns using get_dummies: protocol_icmp, protocol_udp, protocol_tcp
- **Severity:** **Low, Medium, High**
  - Created categorical columns using get_dummies: severity_level_low, severity_level_medium, severity_level_high
- **Segments:** **Segment A, Segment B, Segment C**
  - Created categorical columns using get_dummies and renamed column names: network_segement_a, network_segement_b, network_segement_c
- **Attack Type:** **DDoS, Malware, Intrusion**
  - Created categorical columns using get_dummies: attack_type_ddos, attack_type_malware, attack_type_intrusion.
- **Action Taken:** **Logged, Blocked, Ignored**
  - Created categorical columns using get_dummies: action_taken_logged, action_taken_blocked, action_taken_ignored.
- **Attack Signatures: Pattern A, Pattern B**
  - Created categorical columns using get_dummies: attack_signature_a, attack_signature_b

Santa Clara
Leavey School of Business

Data Cleaning Techniques:

Binary column

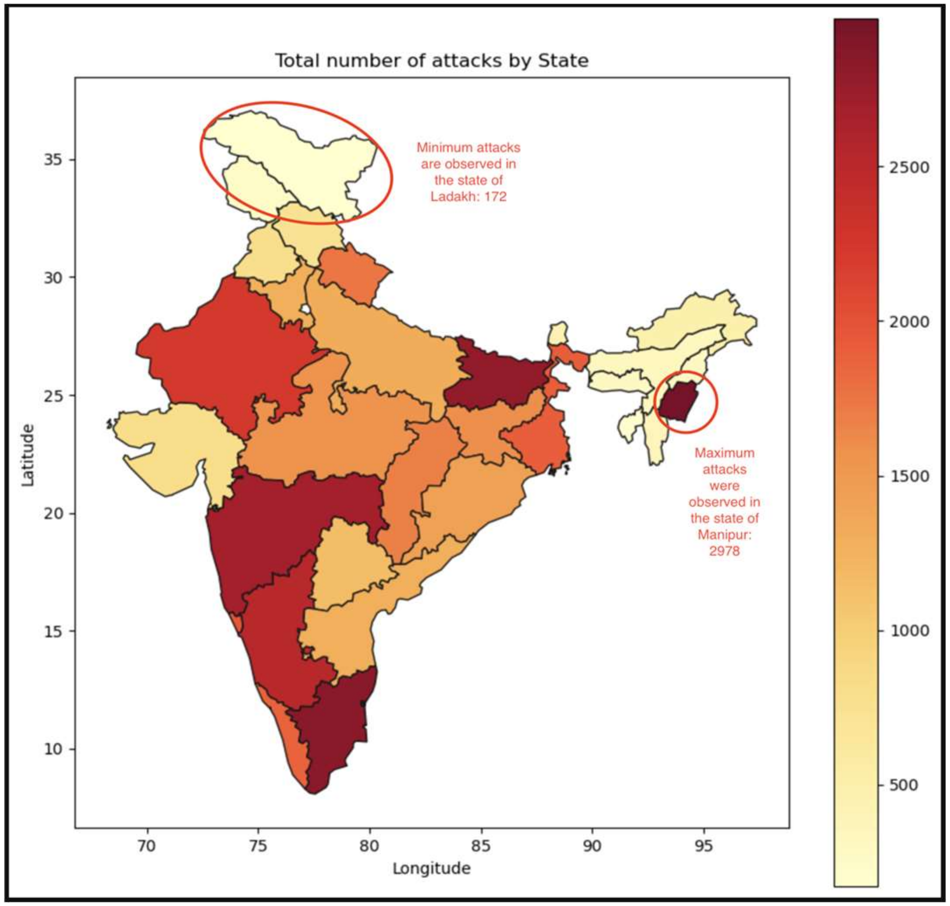
- **Packet Type - is Data? Or is Control?**
  - New column: is_data_packet; dropped original column - 'Packet Type'
  - Values: 1=Data and 0=Control
  - If the packet type is Data? Then 1. Otherwise, 0.
- **IDS/IPS alerts - alerted or not?**
  - Same column: ids_ips_alerts
  - Values: NaN = 0 and Alert Data = 1
  - Represents if there was an alert for IDS/IPS type of attack, it's 1. Otherwise, it's 0.
- **Alert Warnings: alert triggered or not?**
  - New column: alert_triggered; dropped original column - 'Alert Warnings'
  - Values: NaN = 0, Alert Triggered = 1
  - Represents if there was an alert triggered, it's 1. Otherwise, it's 0.

Cleaned dataset: ~7,375 KB.

dataset **size reduced by 57.7%** of ~17.9MB

Finding 1: Interesting finding

As a SecOps manager/leader, **what are hotspots in client-base** and how many attacks were made and what devices were impacted?


Total number of attacks by State
Minimum attacks are observed in the state of Ladakh: 172
Maximum attacks were observed in the state of Manipur: 2978
Latitude
35
30
25
20
15
10
Longitude
70
75
80
85
90
95
2500
2000
1500
1000
500

- Among the 28 states in the dataset, we find that all the states are attacked between the range **172 (Ladakh) - 2978 (Manipur)**

- Let's drill-down Manipur and find out what kind of attacks are prevalent.

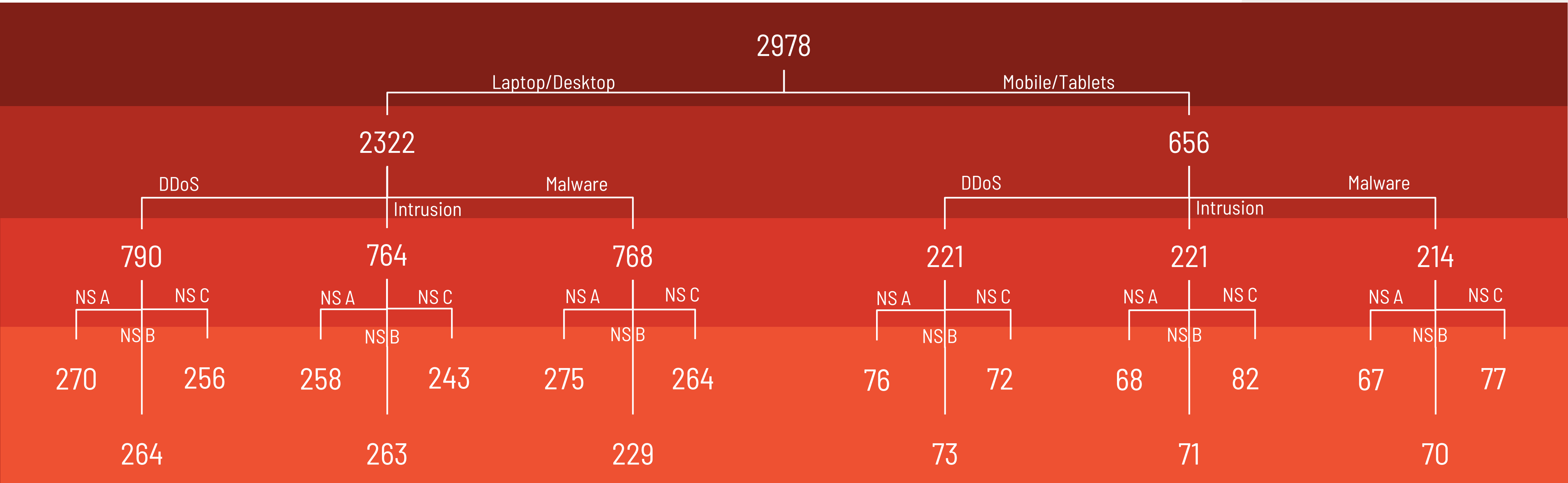
Top 5 states:

- ***Manipur***
- ***Tamil Nadu***
- ***Bihar***
- ***Maharashtra***
- ***Karnataka***

Santa Clara
Leavey School of Business

Finding 1: Interesting finding

As a SecOps manager/leader, what are hotspots in client-base and **how many attacks were made and what devices were impacted?**



- 2978
  - Laptop/Desktop: 2322
    - DDoS: 790
      - NS A: 270
      - NS B: 264
      - NS C: 256
    - Intrusion: 764
      - NS A: 258
      - NS B: 263
      - NS C: 243
    - Malware: 768
      - NS A: 275
      - NS B: 229
      - NS C: 264
  - Mobile/Tablets: 656
    - DDoS: 221
      - NS A: 76
      - NS B: 73
      - NS C: 72
    - Intrusion: 221
      - NS A: 68
      - NS B: 71
      - NS C: 82
    - Malware: 214
      - NS A: 67
      - NS B: 70
      - NS C: 77

NS A = Network Segment for Public sector and Defense

NS B = Network Segment for Commercial Enterprise

NS C = Network Segment for Commercial Individual

In the state of manipur, it is concluded that:

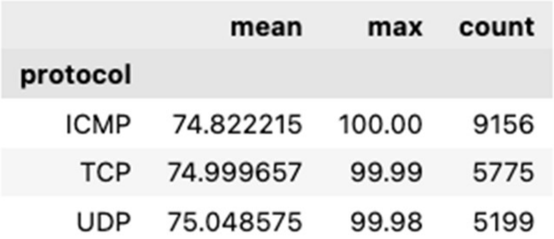
- Public Sector customer face a majority of DDoS attacks to their laptops/desktops
- This is an indicator that we have to provide more protection to customer's laptop/desktop devices for all customer segments.

Santa Clara
Leavey School of Business

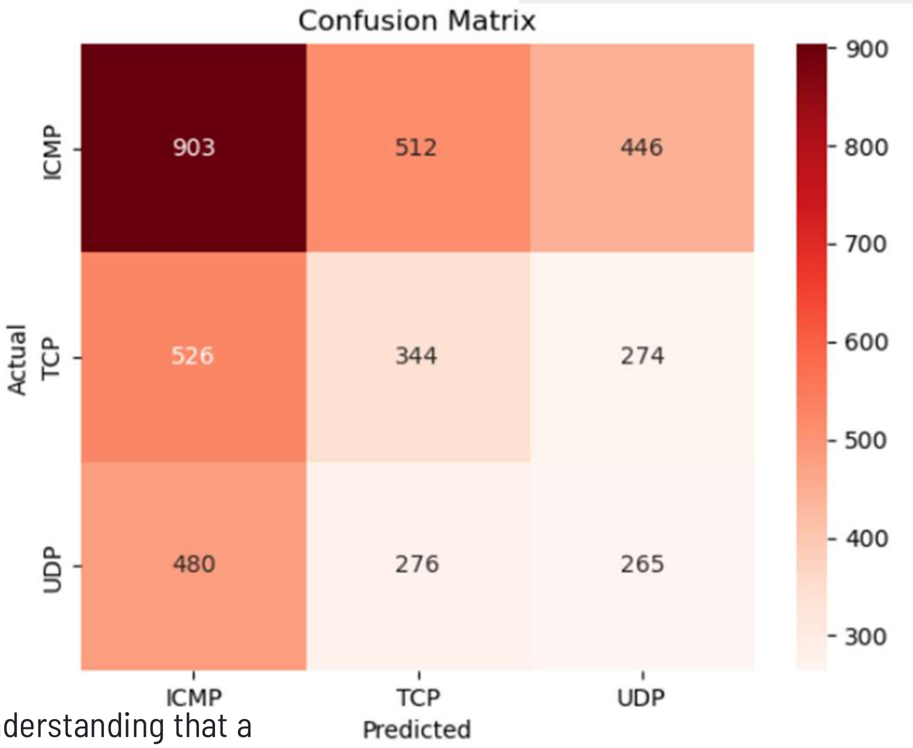
Finding 2: Non-trivial finding

As a SecOps manager/leader, **can we predict the type of protocol being used in the traffic with anomalous behavior and correlate that with the most prevalent attack?**

- Using RandomForestClassifier, team **HERO** is training the dataset containing anomaly score (measure of anomalous behavior) and predicting the protocol which was used with an accuracy of 37.56%
- Using this prediction and the insight in **Finding 1**, a correlation can be derived:
    - ICMP protocol was most prevalent in the anomalous traffic which were captured during the DDoS attacks



	mean	max	count
protocol			
ICMP	74.822215	100.00	9156
TCP	74.999657	99.99	5775
UDP	75.048575	99.98	5199



Confusion Matrix

Actual \ Predicted	ICMP	TCP	UDP
ICMP	903	512	446
TCP	526	344	274
UDP	480	276	265

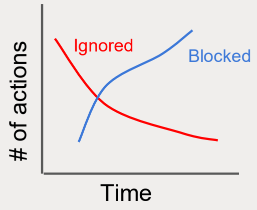
Managerial Insight: From the outcomes of findings 1 and 2, we are arrive at an understanding that a large number of DDoS attacks were made using ICMP protocol. The learning for SecOps team is to tighten up the security rules to rate traffic (especially with ICMP protocol) to the Acme's datacenter and also monitor suspicious traffic patterns (for example: large traffic spikes during off seasons).

Finding 3: Somewhat unexpected finding

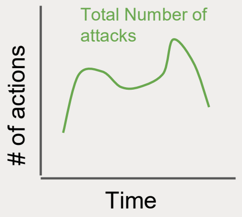
As a SecOps manager/leader, **how can we bolster the security in our systems?**

What we expected to find?

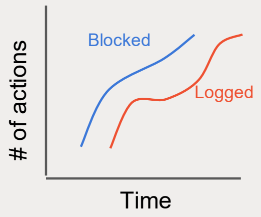
- **A steady decline in ignore actions, steady increase in blocked actions over a period of time**
  - SecOps engineers improve security measures and processes to help prevent attacks *For example: SecOps engineer may decide to limit the traffic flowing to an application by introducing an firewall to filter traffic to the application.*



- **Unstable number of attacks which are marginally above/below the over a period of time**
  - the Number of attacks are generally not a flatline, number of attack fluctuate based on events. *For example: After a product launch, a malicious users have a new product area to break into OR After a publication of a security advisory or a security patch, a malicious user might feel challenged to or may want to verify if the security vulnerability have been eliminated or not.*

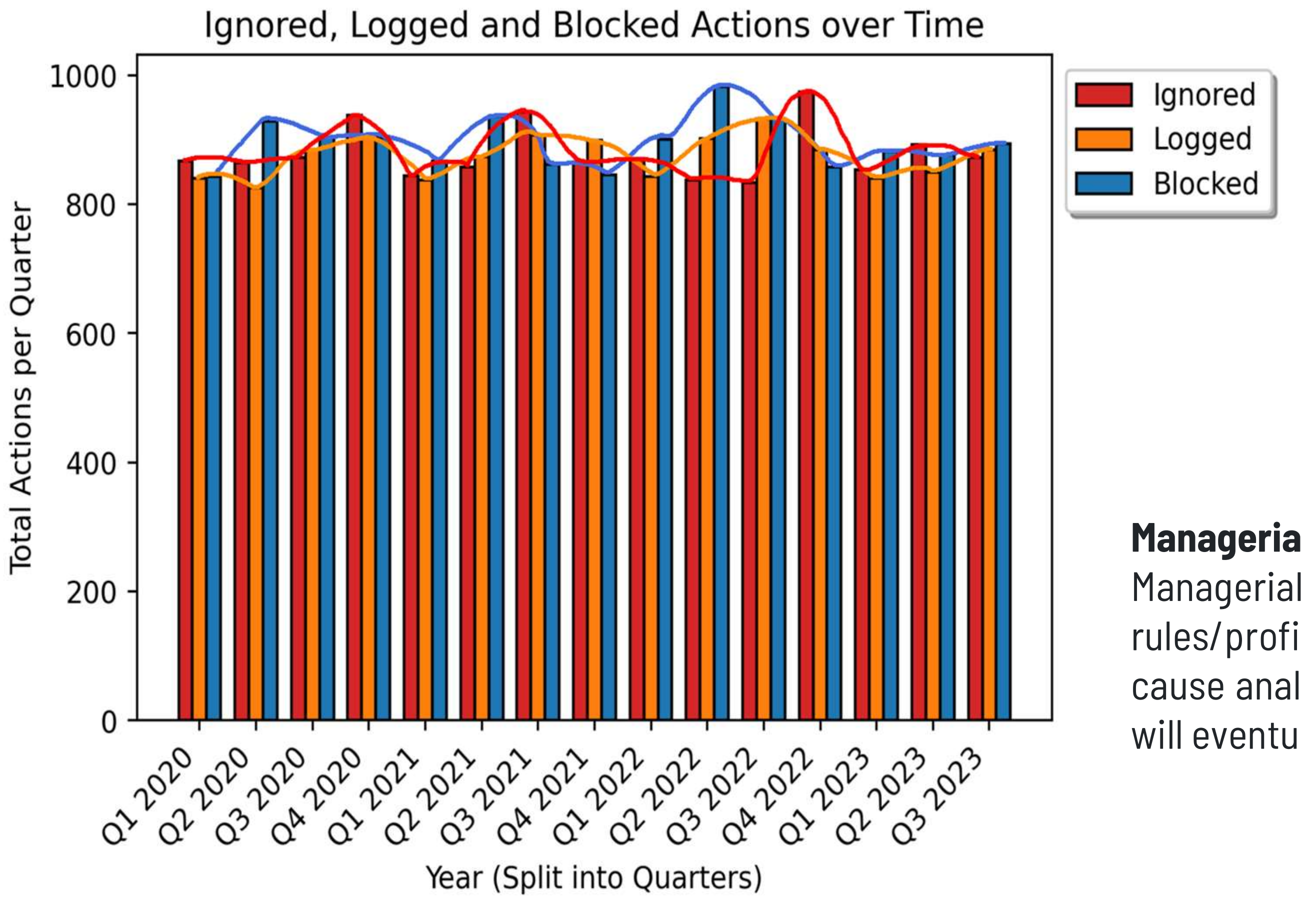


- **An increase in logged action would result in increase to the blocked action as time passes**
  - In the SecOps team, security engineers learn from logged events and improve their preventive measures for the suspicious activities. *This is attributed to ever evolving security posture of the organization.*



Finding 3: Somewhat unexpected finding

As a SecOps manager/leader, **how can we bolster the security in our systems?**



What we found instead?

- The logged actions follow a consistent peaks and troughs corresponding to second half and first half of the year.
- We observe from the dataset that an increase in ignored actions, lead to a decrease in the logged and blocked actions in same and next quarter.
- The action taken Q1 are consistently low than the rest of the quarters.

Managerial insight:

Managerial Insight : Security processes and new security rules/profiles can be implemented by performing root-cause analysis on the Ignored and Logged actions which will eventually lead to more Blocked actions in future.

Santa Clara Leavey School of Business

Backup slides

Sources:

- Icons & logo:
  https://www.rawpixel.com/image/14588430/abstract-particle-technology-background-security-person-human
  https://www.pngall.com/cleaning-logo-png/download/119899/
  https://media0.giphy.com/media/v1.Y2lkPTc5MGI3NjExdXR1eGtjYXJ5azZwYzY1eXh1bjVjdDZmMGFxY3B0NGxhaWx1dWY0NCZlcD12MV9pbnRlcm5hbF9naWZfYnlfaWQmY3Q9cw/KHEqwcxQ0zKfje8rVJ/giphy.gif

- Definition of cybersecurity: https://en.wikipedia.org/wiki/Computer_security

- Reference for DDoS attack: https://www.cyber.gc.ca/en/guidance/defending-against-distributed-denial-service-ddos-attacks-itsm80110
- ChatGPT

THANK YOU!!!!

Santa Clara
Leavey School of Business