ANALYSIS AND PREPROCESSING OF NETWORK TRAFFIC FOR CYBERSECURITY

Group: G-08, Section: C

Submitted To
Dr. Ashraf Uddin
Assistant Professor, CS, AIUB

AI Usage Declaration

We declare that this project and its accompanying report/code have been primarily prepared by our group.

We acknowledge that the use of Artificial Intelligence (AI) tools such as ChatGPT, GitHub Copilot, Grammarly, or similar systems was permitted only to assist in learning, idea generation, code debugging, or language improvement.

We further declare that:

1. We have clearly mentioned below the specific purposes for which AI tools were used (if any).
2. The core design, implementation, analysis, and conclusions are our own original work.
3. We collectively take full academic responsibility for the content of this submission.

AI Usage Details:

☐ No AI tools were used.

☐ AI tools were used for the following purposes (please specify clearly):

We use ChatGPT, Gemini and Perplexity to assist in learning, idea generation, code debugging and language improvement

	Name	Student ID	Signature with Date
1.	KABID, KAZI AL	21-45365-2	
2.	SHIHAB, MD FAHIM AL	22-46945-1	
3.	TALHA, ABDULLAH AL MARJUC	22-47294-1	
4.	PATWARY, HEDAYET ULLAH	22-47904-2	

Table of Contents

Data Understanding 4

Data Exploration & Visualization (EDA) 5

Univariate Analysis: 5

- Histograms & Boxplots 5
- Bar & Pie Charts 5

Bivariate Analysis: 6

- Correlation Heatmap 6
- Boxplots (Cat vs Num) 6

Data Preprocessing 7

- Handling Missing Values 7
- Handling Outliers 7
- Data Encoding 8
- Normalization 8
- Feature Selection 8

Conclusion 8

Reference 8

List of Figures

Figure 1: Sample rows of dataset.. 4

Figure 2: Histograms & Boxplots. 5

Figure 3: Bar & Pie Charts.. 5

Figure 4: Correlation Heatmap. 6

Figure 5: Boxplots (Cat vs Num)... 6

Figure 6: Handling Missing Values... 7

Figure 7: Handling Outliers. 7

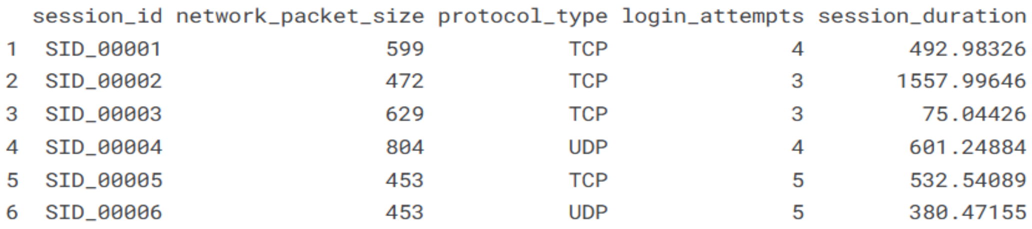
Analysis and Preprocessing of Network Traffic Data for Cybersecurity

In the face of the blistering development of digital infrastructure, network security has gained the top priority on the agenda of all organisations across the globe. The given project is aimed at analysing and preprocessing the data on the cybersecurity network traffic in order to locate possible threats and anomalies. The aim of the primary task was to convert raw and noisy network logs into a clean and structured format that would be passed through machine learning algorithms. The methodology we used was the R programming language and to perform the full Exploratory Data Analysis (EDA), including the missing values as statistically imputed values, and the outliers using the Interquartile Range (IQR) tool. Besides, nominal variables were coded, and the numerical characteristics were normalised through Z-score scaling. The most important result of this project would be to have the refined data in which prominent features, such as the packet size and duration of a session, are streamlined to ensure that the input to be fed into later intrusion-detecting models is of a high standard.

Data Understanding

At this step, we loaded the dataset with 9, 537 records and 10 features. The first few rows show in fig. 1. The network traffic modelled on simulated data reflects the real world.

- Dataset Shape: 9537 rows x 10 columns.
- Primary Numerical Attributes: packet size, logins attempts, duration of session, failed logins, ip reputation score.
- Key Categorical Features protocol type (TCP, UDP, etc.), browser type, used encryption, attack detected (Target variable), unusual time access.
- Observation: Preliminary look of the data showed the absence of the missing values (NA) and outliers in such columns as login attempts, which needed some cleaning.



```
  session_id network_packet_size protocol_type login_attempts session_duration
1  SID_00001                 599           TCP              4        492.98326
2  SID_00002                 472           TCP              3       1557.99646
3  SID_00003                 629           TCP              3         75.04426
4  SID_00004                 804           UDP              4        601.24884
5  SID_00005                 453           TCP              5        532.54089
6  SID_00006                 453           UDP              5        380.47155
```

Figure 1: Sample rows of dataset

Data Exploration & Visualization (EDA)

To understand the underlying patterns, we performed both univariate and bivariate analyses using various visualization techniques (saved in the /visualizations folder):

Univariate Analysis:

- **Histograms & Boxplots:** Next, we visualized the distribution of packet_size and session_duration. Boxplots clearly pointed out outliers in login_attempts, which could indicate possible brute-force attacks.

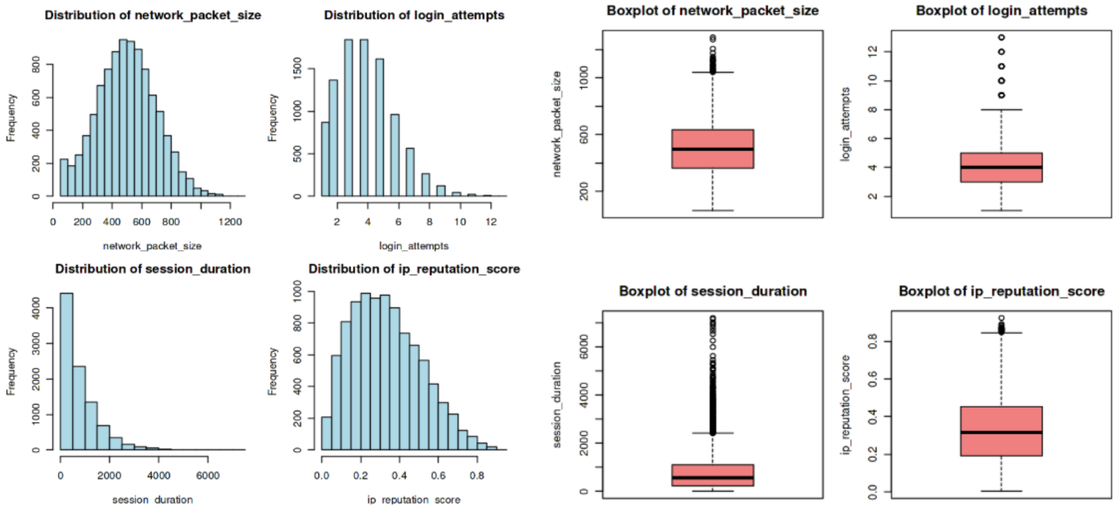


Figure 2: Histograms & Boxplots.

- **Bar & Pie Charts:** These showed the frequency of protocols used. An example would be that Chrome was one of the leading browser types, and TCP was a very frequent protocol.

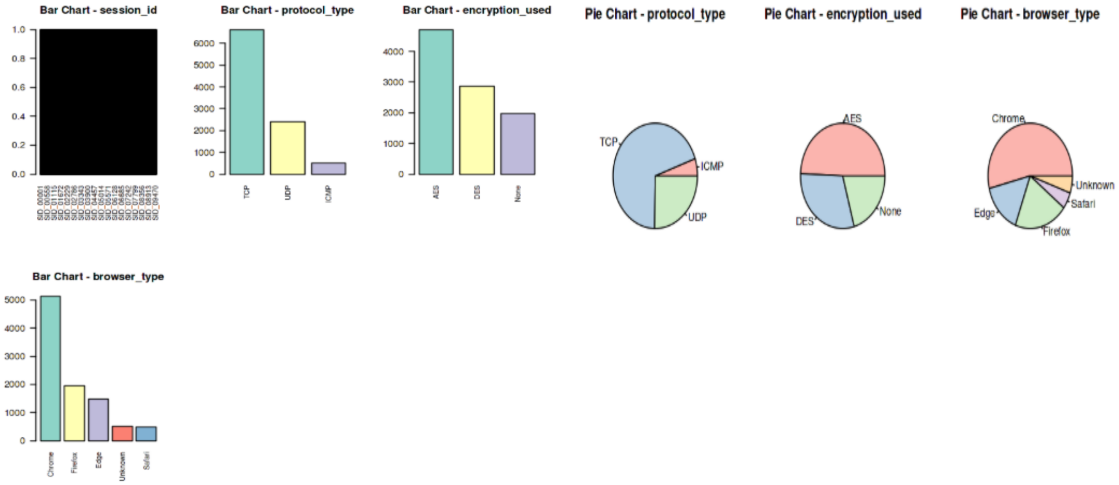


Figure 3: Bar & Pie Charts.

Bivariate Analysis:

- **Correlation Heatmap:** We developed a heatmap to observe the existence of relationships between numerical variables. This helped in identifying multicollinearity among network features.

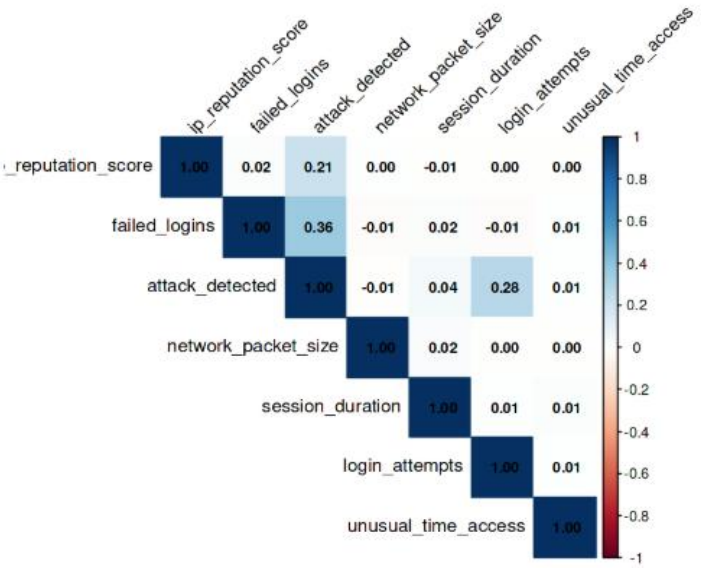


Figure 4: Correlation Heatmap.

- **Boxplots (Cat vs Num):** We studied the variation in packet_size across protocol_types to understand network traffic behavior.

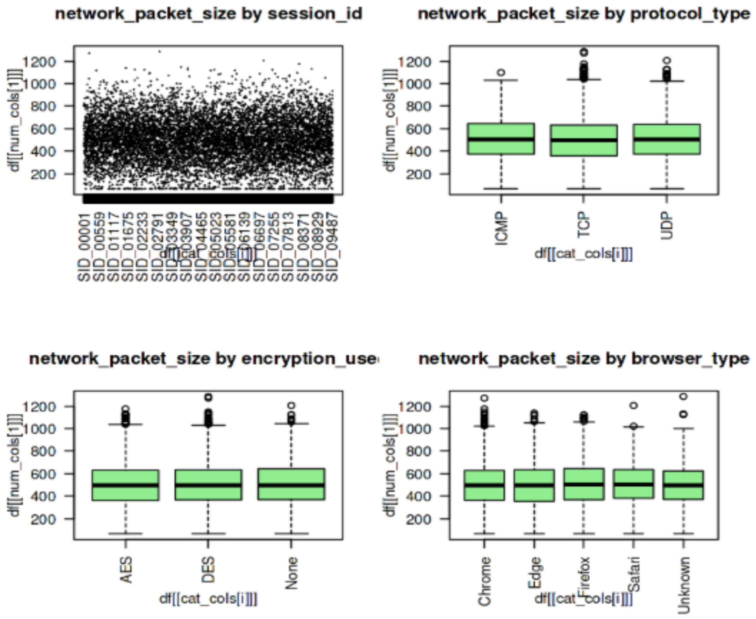
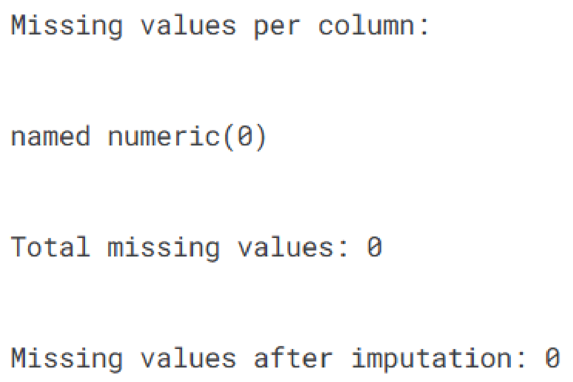


Figure 5: Boxplots (Cat vs Num).

Data Preprocessing

Data preprocessing was the most critical aspect of this pipeline. The following steps were implemented in the R script:

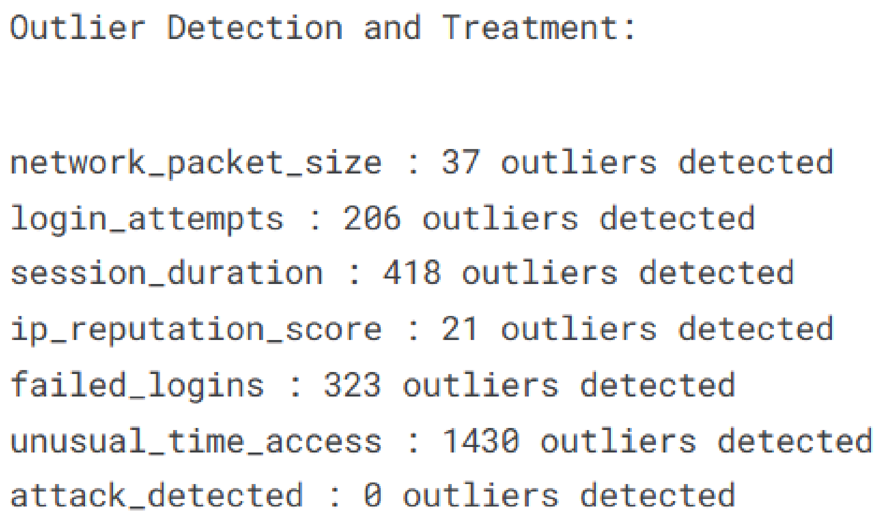
- **Handling Missing Values:** We identified missing values in columns such as packet_size and session_duration.
    - Strategy: Numerical nulls replaced with the Median and categorical nulls replaced with the Mode to maintain data integrity.



```
Missing values per column:

named numeric(0)

Total missing values: 0

Missing values after imputation: 0
```

Figure 6: Handling Missing Values.

- **Handling Outliers:** We identified outliers in numerical columns by the IQR method. Capping was used to set upper and lower bounds on extreme values without losing information.



```
Outlier Detection and Treatment:


network_packet_size : 37 outliers detected
login_attempts : 206 outliers detected
session_duration : 418 outliers detected
ip_reputation_score : 21 outliers detected
failed_logins : 323 outliers detected
unusual_time_access : 1430 outliers detected
attack_detected : 0 outliers detected
```

Figure 7: Handling Outliers.

- **Data Encoding:** Categorical variables such as protocol_type and browser_type were encoded in numeric format using One-Hot Encoding; though the dimensionality of the dataset increased, it became machine-readable.
- **Normalization:** Since some features such as packet_size, e.g., 1500 bytes and login_attempts e.g., 3, are of different magnitudes, we implemented Z-score Normalization to scale the numerical features to a standard scale.
- **Feature Selection:** We removed features with near-zero variance and highly correlated features (cutoff > 0.9) to reduce the noise and improve model efficiency.

Conclusion

This project successfully demonstrated the end-to-end data science pipeline applied to cybersecurity network data. We took a raw, disparate dataset and transformed it into a high-quality analytics asset by fastidiously cleaning the data, handling outliers, and normalizing feature scales. Exploratory analysis showed striking patterns in protocol usage and login behaviors that are indicative of potential attacks. Among the notable accomplishments was that feature selection greatly reduced the dimensions of the dataset while maintaining information variance. However, one of the limitations of this project was the utilization of synthetic/simulated data, which may not capture the full complexity of zero-day attacks. This pipeline could be extended in future work to take live network streams as input and apply further dimensionality reduction methods such as PCA before model training.

Reference

1. Dataset available: **https://drive.google.com/file/d/14PbXDWMuQeq4kNdT-pJny21HhQbWIdCS/view**