This presentation describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the presentation do not necessarily represent the holistic view of the of the entire data analyzed.

# Project Discussion: ANALYSIS OF MALWARE DETECTION


PRESENTED BY

***BABATUNDE ABDULLAHI OLATUNJI***

***ESTHER ABIODUN OSIKOYA***

***AUGUSTINE OLOCHE NOAH***

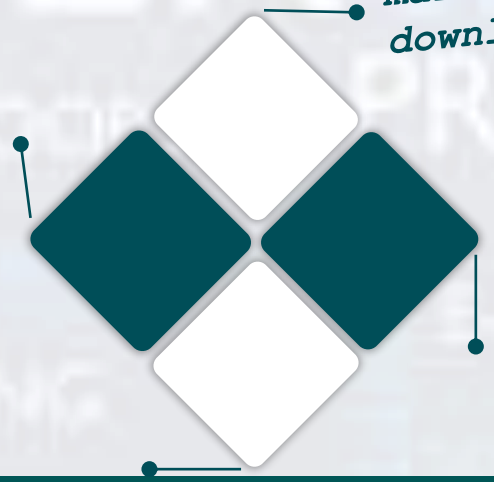
# Table of Content

# Highlight on Malware

***Malware is a broad term used to refer to malicious software designed to damage, disrupt, or gain unauthorized access to computer systems, networks, or individual devices. The term "malware" is a combination of "malicious" and "software," highlighting its harmful intent. Malware is typically created by cybercriminals with malicious intent, including stealing sensitive information, disrupting computer systems, extorting money through ransomware, or using infected devices to carry out further cyberattacks.***

**Malware** is a broad category that includes various types of harmful software, such as **viruses, worms, Trojans, ransomware, spyware, adware,** and more.

**Malware** can be distributed through various means, including **email attachments, infected websites, malicious links,** and **compromised software downloads.**



**Malware and cyber attacks are occurring at an increasing rate**

- Heartbleed
- Ransomware
- Attacks on power systems…

**A common thread is that something has to be executed on the host system**

- Signature based malware detection susceptible to obfuscation attacks
- Add null operators changes hashes
- Cannot detect novel variants of executables
- Brittle

## *Current ML approaches :Static & Dynamic*

- *In a dynamic approach, data is processed and analyzed in real-time or on-the-fly as it becomes available. This means that the model adapts and updates itself based on new incoming data, allowing it to continuously learn and improve over time. Dynamic approaches are well-suited for scenarios where the data distribution changes frequently, or where there is a need to respond quickly to evolving patterns and trends*

- *In a static approach, data is analyzed and processed in batches, and the model is trained on a fixed dataset, which is not continuously updated with new data. This means that the model remains unchanged until it is explicitly retrained with an updated dataset. Static approaches are suitable for scenarios where the data distribution is relatively stable over time, and frequent retraining is not necessary.*

# Prevalence of Malware and Cyber Attacks

## New Linux malware mines cryptocurrency and steals your password

### Amazon hit with major data breach days before Black Friday

**Customers' names and email addresses posted on website, tech giant confirms**


### Nasty New Linux Crypto Malware Compromise Root, Launches DDoS Attacks

With the value of Bitcoin once again experiencing a big drop this past week, you may begin to think that malware developers would begin shifting focus elsewhere. Unfortunately, that's far from being the

### Ukraine detects new Pterodo backdoor malware, warns of Russian cyberattack

*Revived Gamaredon threat group just part of wave of new attacks tied to Russia's FSB.*


### Cybersecurity Firm Detects Cryptojacking Malware on Make-A-Wish Foundation Website

### *Emotet malware runs on a dual infrastructure to avoid downtime and takedowns*

**Researchers spot unique design in the server infrastructure propping up the Emotet malware.**

### *Google removes 13 malware apps from its Play Store*

***The 13 malware apps have been downloaded over 5,60,000 times from the Play Store***

# Current Antivirus Techniques

***To protect against malware, users and organizations should adopt good cybersecurity practices, such as using up-to-date antivirus software, keeping operating systems and applications patched, being cautious with email attachments and links, and avoiding downloading software from untrusted sources. Regular backups of important data can also help mitigate the impact of ransomware attacks.***

## Signature Matching

- Easiest to defeat
- Trivially modify with non-used
- Code

## Heuristics-based Detection

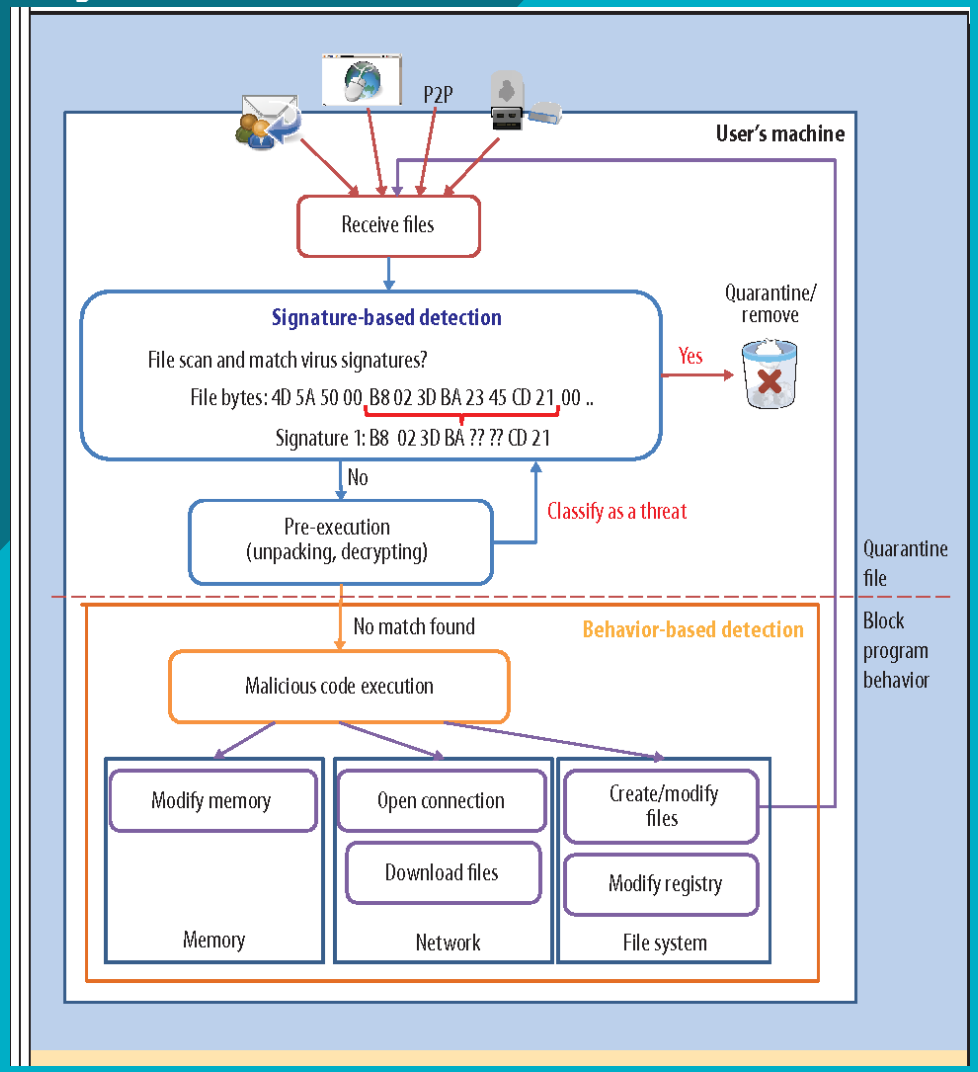
- Look for generic characteristics
  - Specific, rare, operations
  - Specific registry modifications

## Behavioral Detection

- Run the executable in a sandbox
- Observe behavior

## Each is vulnerable to minor changes in the code

- Not able to adapt to new changes




*J. Hoe, H. Kim and O. Sukwong, "Commercial Antivirus Software Effectiveness: An Empirical Study," in Computer, vol. 44, no. , pp. 63-70, 2010. doi:10.1109/MC.2010.187*

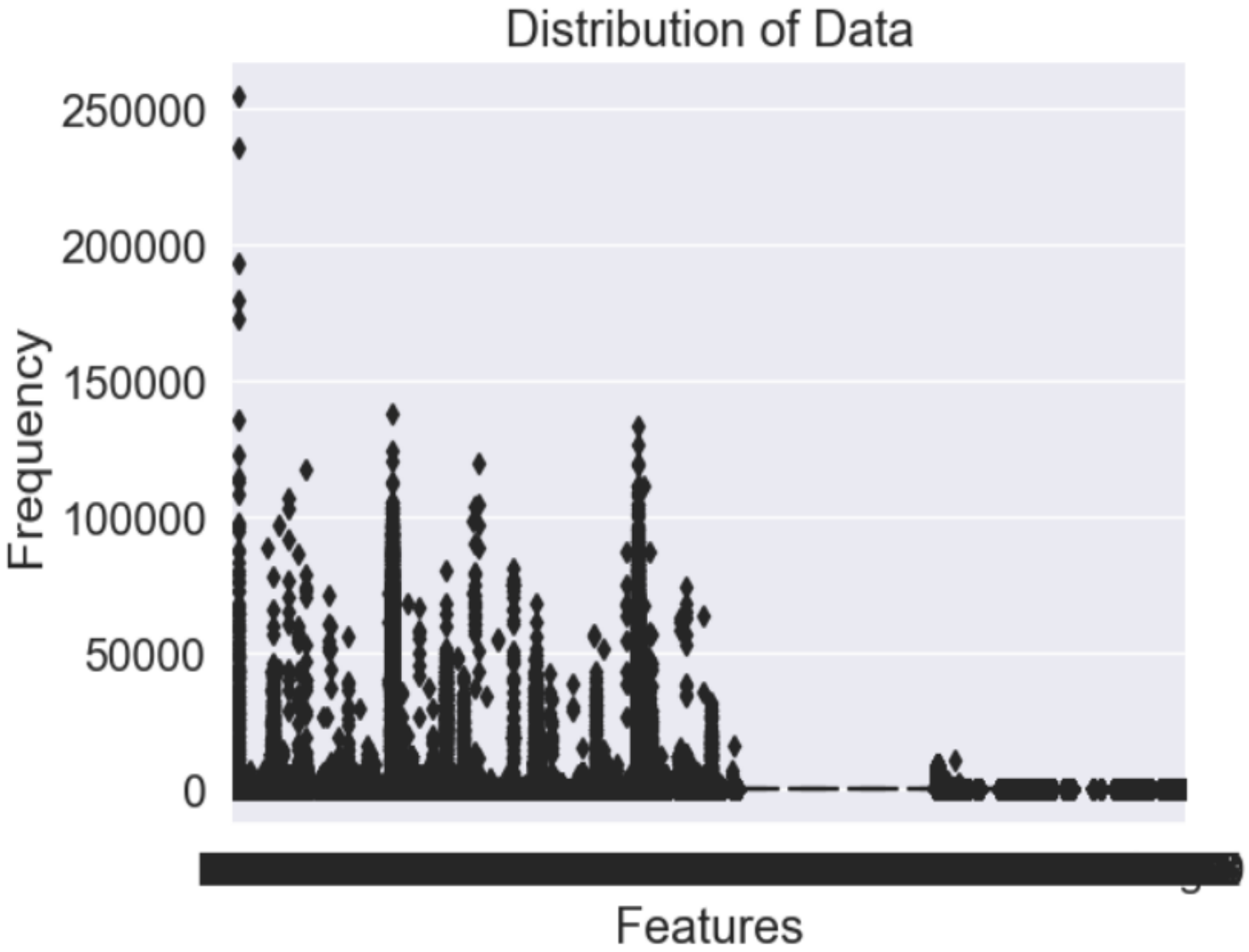
# Thwarting Anti-Virus

- Code packing and encryption
- Code mutation
- Polymorphic code
- Stealth techniques
  - *Process injection*
  - *Process hiding*
- Turning off anti-virus
- Adding noop

Solution: Use machine learning to monitor system calls

- ML can generalize away from static signatures
- System calls are the base level for interacting with the operating system


# Dataset Description




## Description

The initial data set provided for malware detection was in the form of a sparse text file, using the LIBSVM format, to enhance its usability, the data underwent a transformation and was consolidated into a more user-friendly CSV file.

Key details about the dataset:

Observation: 107,850
Number of Features: 480

# Dataset Cleaning

1

```
In [9]: df1
```

Out[9]:

| | FT1 | FT2 | FT3 | FT4 | FT5 | FT6 | FT7 | FT8 | FT9 | FT10 | ... | FT471 | FT472 | FT473 | FT474 | FT475 | FT476 | FT477 | FT478 | FT479 | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.074074 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.684211 |
| **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.814815 |
| **3** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.814815 |
| **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.421053 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **107851** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.438596 |
| **107852** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.666667 |
| **107853** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.719298 |
| **107854** | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.403509 |
| **107855** | 1 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.403509 |

107856 rows × 480 columns

2

| | FT1 | FT3 | FT4 | FT5 | FT6 | FT7 | FT9 | FT10 | FT12 | FT13 | ... | FT463 | FT468 | FT469 | FT470 | FT471 | FT473 | FT476 | FT478 | FT479 | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.074074 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.684211 |
| **2** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.814815 |
| **3** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0.814815 |
| **4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.421053 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **107851** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.438596 |
| **107852** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.666667 |
| **107853** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.719298 |
| **107854** | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.403509 |
| **107855** | 1 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.403509 |

107856 rows × 266 columns

Upon analyzing Table 1, we found that the raw dataset consists of 480 features, out of which 214 features contain only zeros.

As a result of eliminating these entirely zero features, we obtained the modified dataset represented in Table 2.

# Dataset Cleaning

```
selected_df['target'] = np.where(selected_df['target'] > 0.5, 1, 0)

# 0 - Benign 1 - Malware
# Print the updated DataFrame
selected_df
```

| | FT152 | FT402 | FT473 | FT394 | FT210 | FT81 | FT96 | FT392 | FT233 | FT356 | target |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 45 | 0 | 0 | 0 | 1 | 33 | 0 | 0 | 2 | 96 | 0 |
| **1** | 9 | 0 | 0 | 0 | 4 | 5085 | 0 | 0 | 3 | 54 | 1 |
| **2** | 1691 | 0 | 0 | 0 | 5 | 0 | 0 | 1 | 5 | 11 | 1 |
| **3** | 3 | 0 | 1 | 0 | 1 | 55 | 0 | 1 | 2 | 63 | 1 |
| **4** | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **107851** | 1 | 0 | 0 | 0 | 3 | 2 | 0 | 1 | 2 | 56 | 0 |
| **107852** | 12 | 0 | 0 | 0 | 12 | 24 | 0 | 1 | 3 | 54 | 1 |
| **107853** | 0 | 0 | 0 | 0 | 6 | 2 | 0 | 1 | 6 | 5 | 1 |
| **107854** | 218 | 0 | 0 | 0 | 110 | 4 | 0 | 1 | 2 | 60 | 0 |
| **107855** | 217 | 0 | 0 | 0 | 108 | 5 | 0 | 1 | 2 | 78 | 0 |

107856 rows × 11 columns

## Modifying the dataset to select the best 10 features

Given that correlation does not imply causation, we proceeded by implementing an additional regressor to identify the top 10 features based on their significance in predicting the target variable. Moreover, we converted the target variable into binary values, where "0" indicates Benign (No malware) and "1" indicates Malware.

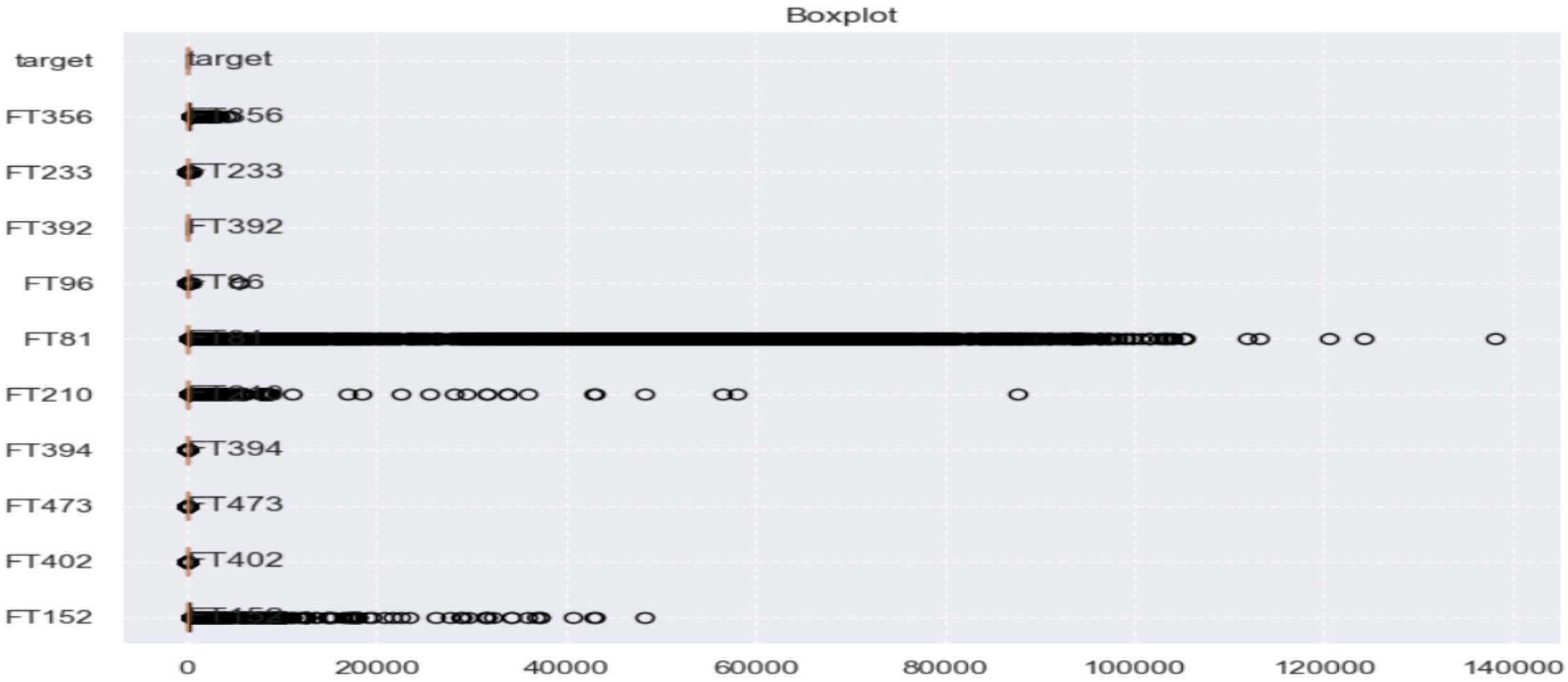
Consequently, the refined dataset we are currently working with contains the following key information:

Total Data Points: 107,850
Number of Features: 11

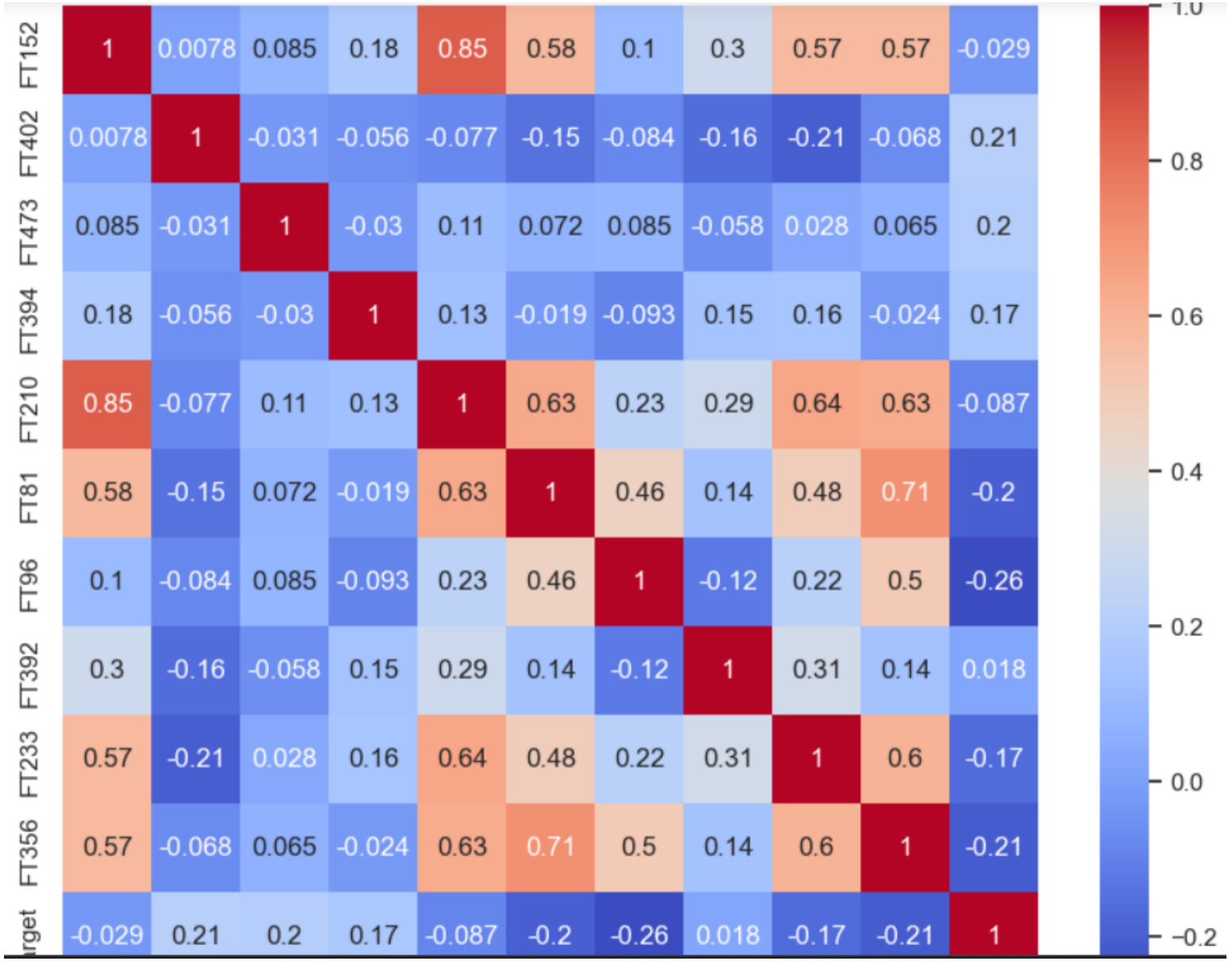
# Boxplot

**A well spread dataset**

- All options are used
- Well distributed



# Correlation matrix




Here the correlation matrix conveniently summarized the ten features with greater contribution and effect on the target variable.

# Imbalanced Data

The imbalance data helps to know the method to correct biasness of the data


```
Class Counts:
1    78013
0    29843
Name: target, dtype: int64
```

# Supervised Learning

## Confusion Matrix

| | Metric | Score |
|---|---|---|
| 0 | Train accuracy | 0.852025 |
| 1 | Test accuracy | 0.835955 |
| 2 | F1 score | 0.889144 |
| 3 | Precision | 0.871489 |
| 4 | Recall | 0.907529 |
| 5 | Balanced accuracy | 0.777436 |

**XGBOOST**

| | Metric | Score |
|---|---|---|
| 0 | Train accuracy | 0.918833 |
| 1 | Test accuracy | 0.835059 |
| 2 | F1 score | 0.888040 |
| 3 | Precision | 0.871854 |
| 4 | Recall | 0.904839 |
| 5 | Balanced accuracy | 0.778912 |

**RANDOM FOREST**

| | Metric | Score |
|---|---|---|
| 0 | Train accuracy | 0.717122 |
| 1 | Test accuracy | 0.717804 |
| 2 | F1 score | 0.831609 |
| 3 | Precision | 0.732426 |
| 4 | Recall | 0.961862 |
| 5 | Balanced accuracy | 0.519008 |

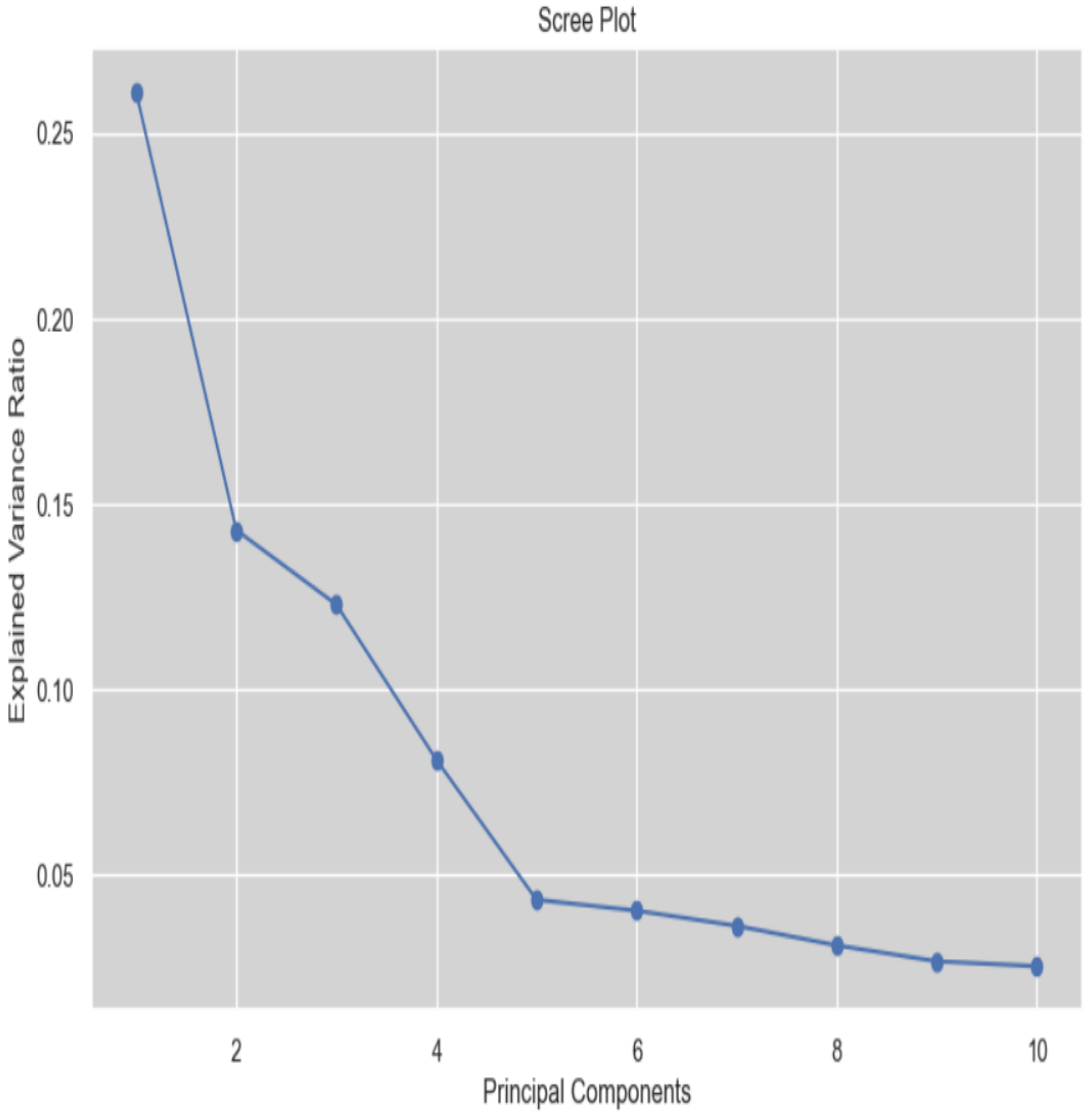
**LOGISTIC REGRESSION**

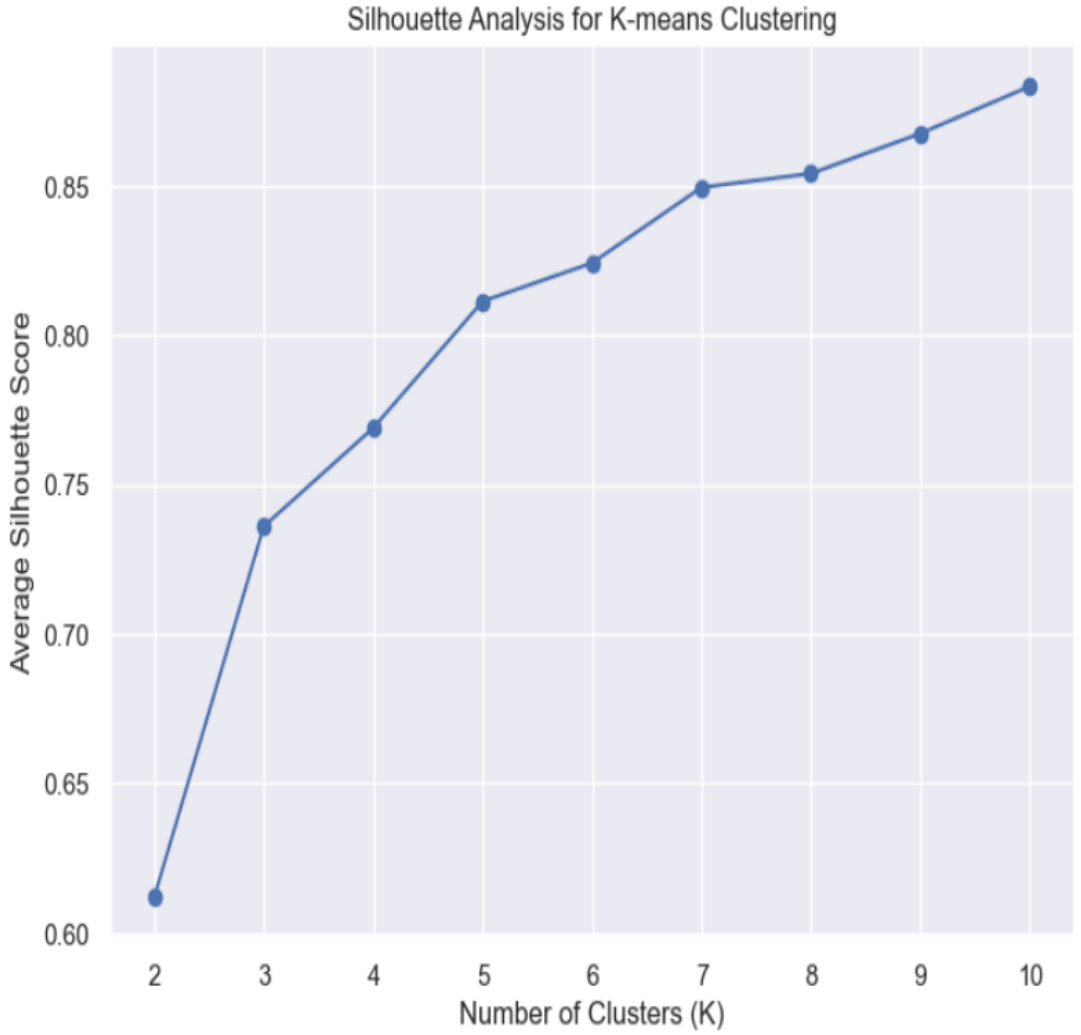
| | Metric | Score |
|---|---|---|
| 0 | Train accuracy | 0.769017 |
| 1 | Test accuracy | 0.768644 |
| 2 | F1 score | 0.850793 |
| 3 | Precision | 0.797660 |
| 4 | Recall | 0.911510 |
| 5 | Balanced accuracy | 0.653026 |

**ADABOOST CLASSIFIER**

# Unsupervised Learning

## Scree plot and Silhouette Analysis for K-means clustering



# Conclusion

- In this work, we carried out Supervised and Unsupervised learning to predict the presence of Malware and uncover insight respectively

# Thank you