# Executive Summary

This report presents an analysis of a dataset consisting of over 3.6 million entries related to iOS devices. The dataset includes information on iOS versions, device models, process IDs (PIDs), timestamps, device identifiers, and process names. The primary goal of this analysis is to identify any potential threats or anomalies within the dataset, particularly in the context of threat hunting within the mobile phone industry. The analysis focuses on understanding the structure and distribution of the data to uncover any unusual patterns or outliers.

# Data Overview

The dataset consists of 9 columns:

- iOSVersion: The version of the iOS operating system running on the device.
- iOSModel: The model of the iOS device (e.g., iPhone 12, iPhone SE).
- pid: The process ID, a unique identifier for each process running on the device.
- readableTimestamp: A human-readable timestamp indicating when the data was recorded.
- timestamp: The same timestamp in integer format, likely representing Unix time.
- id: A unique identifier for each entry in the dataset.
- device: The specific device identifier.
- scan: Information about the scanning process, potentially including details about the scan type or result.
- procName: The name of the process running on the device.

# Analytical Approach

The analysis was conducted in several stages:

## Data Cleaning and Preprocessing

- Check for Missing Values: Identified that there are 13,074 missing PID values, which were dropped from the dataset. Noted the need to investigate why these PID values were missing in the future.
- Convert Timestamp Fields to Datetime Objects: The dataset has two timestamp fields: readableTimestamp (currently in object format) and timestamp (in integer format). To focus on the most relevant data and improve analysis efficiency, any entries with timestamps before 2024 were removed from the dataset.
- Ensure PID is an Integer: The pid field should be an integer rather than a float. Converted the pid field to an integer data type to ensure that process IDs are correctly represented. PID is treated as both continuous and categorical variable for analysis.

- Assume No Duplicate Values: Assumed that there are no duplicate values in the dataset, as the timestamps are at the seconds level. Note that for future work, the timestamps should be extracted at the millisecond level, and an analysis should be performed to check for any duplicate entries.
- Check for Privacy Issues: Determined that there does not seem to be any data that requires special handling for privacy issues.

# Exploratory Data Analysis (EDA)

After data cleaning, a dataset comprising 2,565,753 rows of data collected from 25 unique devices over 910 unique scans. The data reflects processes running on these devices across various iOS versions.

- Total Rows of Data: 2,565,753
- Number of Devices: 25
- Total Scans Conducted: 910
- Distinct iOS Versions: 14
- Unique Process IDs (pid): 23,071
- Unique Process Names: 1,519

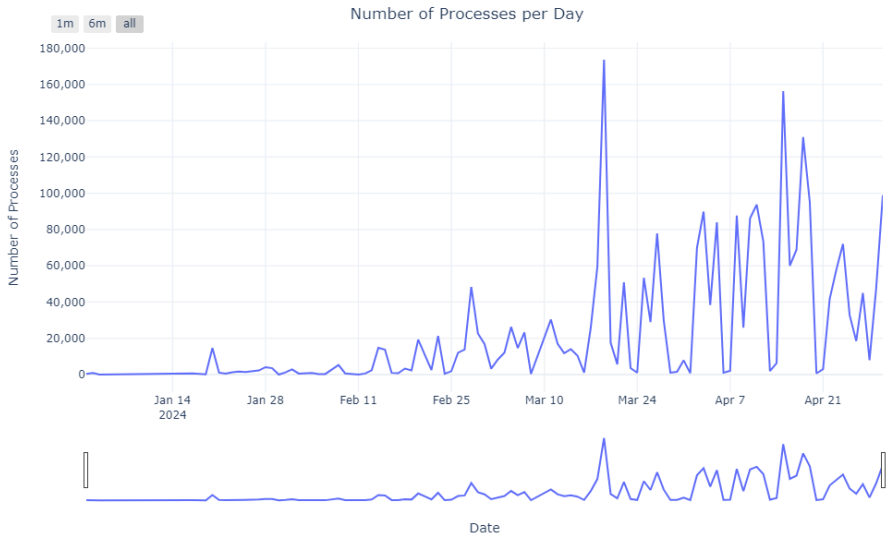
Each scan provides a snapshot of the processes active on a device at a specific time. The dataset spans multiple iOS versions, with iOS 17.4.1 being the most prevalent, representing nearly 50% of the data. This suggests that many devices were using this version during the data collection period, making it a critical focus for analysis.

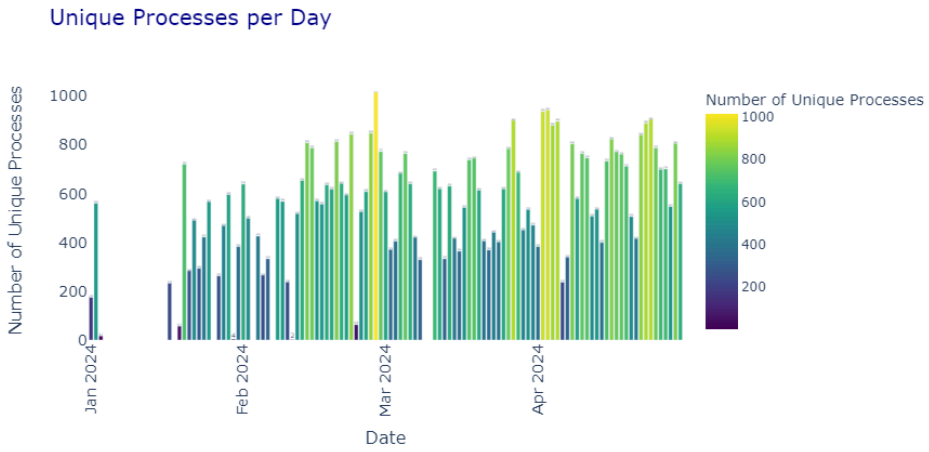
## Time Series Analysis

The number of processes executed per day and the number of unique processes per day from January 2024 to April 2024.

**Interactive Plots**

- number_processes.html,
- unique_processes.html



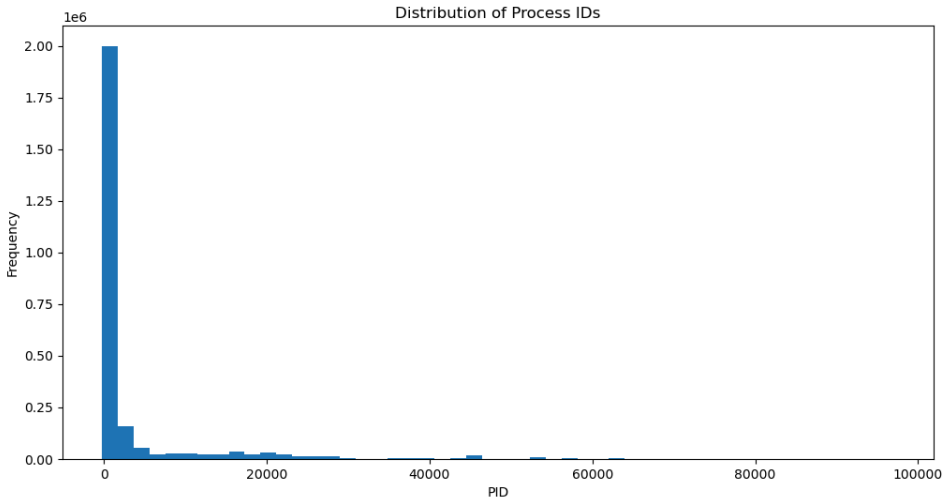

The main observations were:

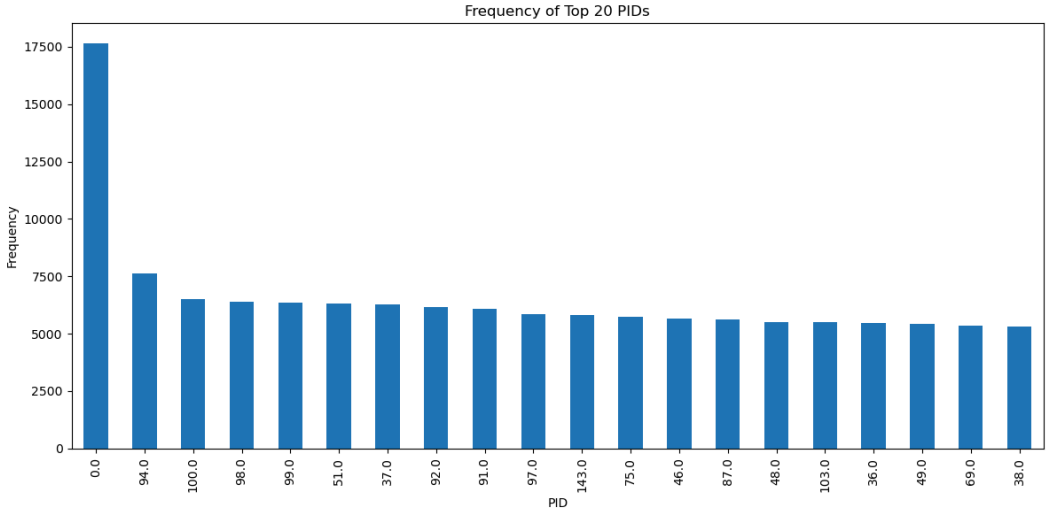
- Early January to late February: Relatively stable process counts, generally below 20,000 per day.
- Late February to March: Gradual increase in process counts, with significant variability and dramatic spikes up to over 160,000 processes per day.
- Late March to April: Continued fluctuating activity, with lower but still irregular process count spikes.
- Unique process counts started low in January, then increased steadily through February and March, reaching peaks of nearly 1,000 unique processes per day in mid-March.
- The fluctuations and spikes in both total and unique process counts could indicate anomalous behaviour, system changes, or potential security threats that warrant further investigation and monitoring.

# Process Behaviour Analysis

## Process Frequency Distribution

- The dataset includes 1,519 unique processes, with a significant skew towards high-frequency processes.
- The top 10 most common processes account for the majority of occurrences, and these are likely integral to the normal operation of the iOS system.
- The dataset also includes 67 rare processes that appear only once, which could represent less common applications or potential indicators of unauthorised software.




## Scans and Process Counts

- The average number of processes per scan is around 2,820, with a large variance, indicating some scans capture significantly more processes than others.

- The maximum number of processes in a single scan is 422,302, which is an extreme outlier and warrants further investigation as it could be indicative of an abnormal or compromised state.

### Multiple Process Lists in Scans

- 203 out of 910 scans contain multiple process lists, which could be indicative of concurrent or repeated process captures, suggesting irregular behaviour or a misconfiguration in the data collection process.

### Common Process Prefixes and Frequent, Persistent Processes

- The top process name prefixes, such as com, MTLCompilerService, and extensionkitservice, are typical of mobile systems, indicating the data contains standard system processes.
- Certain processes, like CommCenter, ContextService, and AppleCredentialManagerDaemon, are present in over 95% of scans, suggesting they are essential system processes.

### Similar Process Pairs

The dataset includes examples of similar process pairs, such as:

- ('dprivacyd', 'adprivacyd')
- ('com.apple.SafariServices.Conten', 'com.apple.SafariServices.ContentBlockerLoader')
- ('com.apple.SafariServices.Conten', 'com.apple.SafariServices.Content')

These similar process pairs may warrant further investigation to understand their purpose and relationship within the system.

### PID Reset Detection and Analysis (Future Work)

- Over 1.19 million PID resets were detected in the dataset, which are normal system behaviors but could obscure anomalies if not properly accounted for.
- Processes like MTLCompilerService, kernel, and CommCenter are often observed around these resets, indicating their involvement in or resilience to system resets.

### Anomaly Detection (Future work )

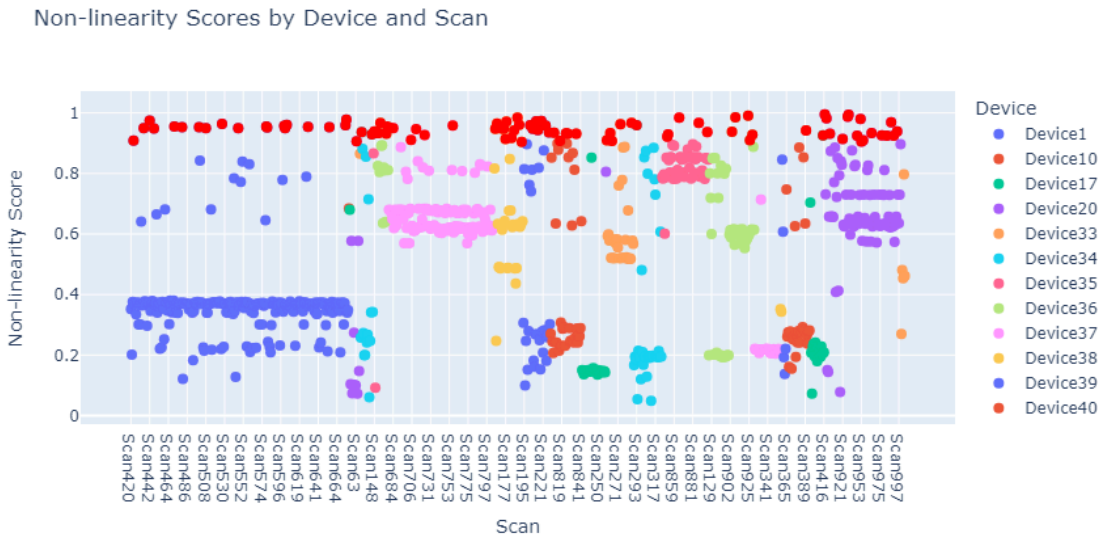
- Anomaly detection methods did not identify any clear, high-confidence anomalies across multiple techniques, suggesting the dataset largely contains standard, non-malicious system activities.
- However, some processes flagged for potential anomalies, such as filecoordinationd, duetexpertd, and IOMFB_bics_daemon, warrant further investigation and monitoring.

## Non-Linearity Scores Analysis

- The purpose is to provide insights into non-linearity scores measured across various devices and scans.

**Interactive Plot:**

- pid_nonlinearity.html



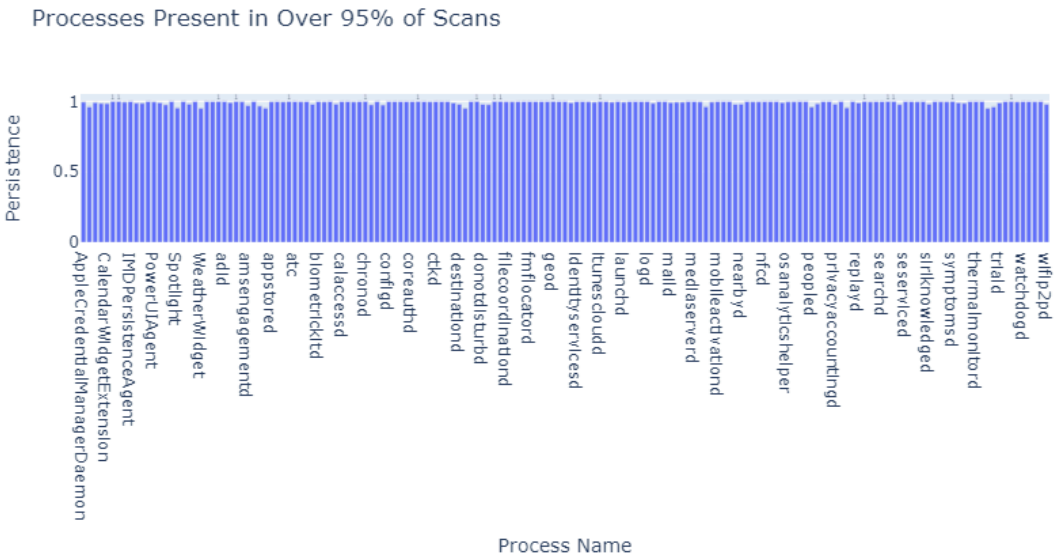

- The plot shows non-linearity scores for multiple devices across different scans, with each dot representing a device's score in a particular scan.
- Key findings:
  - Certain devices exhibit consistently low nonlinearity scores, suggesting stable performance.
  - Some devices demonstrate significant fluctuations, with intermittent high nonlinearity scores, warranting further investigation.
  - Devices like Device38 and Device35 have several high nonlinearity scores, which could indicate potential vulnerabilities or unexpected performance behaviours.

## Process Persistence Analysis

- The objective is to understand the behaviour and persistence of processes identified in a dataset.

**Interactive plot:**

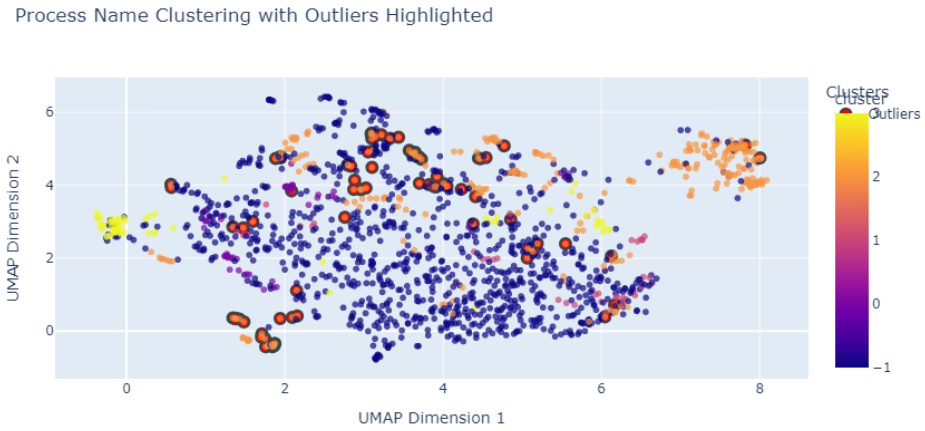
- Persistent_processes.html




- The plot shows the persistence ratio of various processes found in over 95% of the scans.
- Key findings:
    - The majority of processes have a persistence ratio close to 1, indicating they are frequently detected across scans and are likely integral to device operations.
    - There are no immediate outliers or anomalies that stand out as rare or potentially suspicious processes.
    - The process names suggest they represent legitimate applications and system processes, given their high persistence rates.
- No significant concerns or anomalies were identified in the process persistence analysis.

## Clustering of Process Names

**Interactive plot:** procName_outliers.html




- The analysis employed UMAP (Uniform Manifold Approximation and Projection) to visualise the high-dimensional process name data in a 2D plot.
- The data points are grouped into distinct clusters, suggesting there are several categories of process names that share similar characteristics or behaviours.
- Each cluster is represented by a specific colour, indicating that the processes within a cluster may have analogous functions or behave in a similar manner.

2. Outliers:
   - Certain data points are highlighted as outliers, using a different colour scheme.
   - The presence of outliers could point to unique process behaviours or potentially malicious activities that warrant further investigation.
3. Dimensionality Reduction:
   - The UMAP technique has effectively captured the underlying structure of the data in a simplified 2D format.
   - The distribution of the clusters in the plot reveals that they are separable, suggesting there is meaningful information within the process names that can be leveraged for further analysis.

# Conclusion and Recommendations

The reports recommend analysing the specific processes involved during peak activity periods. This should include correlating the spikes in activity with known events. Enhanced monitoring should be implemented to quickly detect any abnormal increases in processes, as this could be an early indicator of system or security issues.
The overall analysis suggests that the dataset predominantly contains common, expected system processes. However, there are a few potential areas of interest that require additional investigation. This is to determine if they represent genuine anomalies or simply normal variations in system behaviour.
The key findings from these reports highlight the need for further investigation into devices with high nonlinearity scores. The process persistence analysis suggests a generally stable and expected behaviour across the scanned devices.

The recommendation is to conduct more in-depth analysis of scans with high nonlinearity scores. This should include cross-referencing with known threat patterns to assess potential risks.
The clustering analysis provides insights into the various categories of process names within the dataset. The identification of well-defined clusters and outliers presents an opportunity for further investigation. This is to ascertain the nature and intent of the outliers. This analysis can aid in developing strategies to enhance security measures in the mobile phone industry and improve threat detection capabilities.
It is recommended to conduct a deeper analysis of the highlighted outliers. This is to understand their potential impact or relation to any unusual or harmful behaviour in mobile applications.

# Future Work

- Investigate Missing PID Values: The missing 13,074 PID values need to be investigated further to understand why they were missing in the first place. This information could provide valuable insights into the data collection or processing pipeline, and help improve the completeness of the dataset going forward.
- Investigate Timestamps from 1970: The presence of timestamps from 1970 is unusual, as this is often an indicator of issues with parsing or handling of timestamp data. It will be important to investigate the source of these 1970 timestamps and ensure they are handled correctly, as they may not represent valid or relevant data.
- Extract Timestamps at Millisecond Level: For future work, the timestamps should be extracted at the millisecond level, rather than just the second level. This higher level of precision will enable a more thorough analysis of the data, including the ability to check for any duplicate entries that may have been missed due to the coarser second-level timestamps.
- PID Reset Detection and Analysis: Perform an analysis to detect any instances of PID resets, where the process IDs may have been reused or recycled over time. This could provide insights into system behaviour and potentially identify any anomalies or security-related issues.
- Anomaly Detection: Implement advanced anomaly detection techniques, such as Isolation Forest (IForest) or Local Outlier Factor (LOF), to identify any unusual or suspicious patterns in the data that may warrant further investigation.
- Improve Visualisations: Enhance the visualisations of the data to better communicate the insights and findings. This could include creating more informative graphs, charts, or dashboards to help stakeholders understand the key trends and anomalies.
- Cross-reference with Domain Knowledge: Leverage any available domain knowledge or external data sources to provide context and better interpret the findings from the analysis. This could involve cross-referencing the data with known security threats, system architectures, or industry best practices.

By addressing these additional points, the data preprocessing and analysis can be further strengthened, leading to more comprehensive and actionable insights from the dataset.