
PROJECT REPORT

# Security Data Analysis - Detecting Anomalous Login Activity

**Project in Data Analysis : Report**

**Submitted To :**
**Dr. Mayank Dave**
**Professor**

**Submitted By:**
**Paramveer Singh**
**Candidate ID : 6422**

# ABSTRACT

Anomalous login detection is a critical security activity that can help to identify and prevent unauthorized access to systems and data. By analyzing login data for patterns that deviate from normal behavior, security analysts can identify potential threats and take action to mitigate them.

There are a number of different methods that can be used to detect anomalous logins. Some common approaches include:

- Rule-based detection: This involves defining rules that specify what constitutes anomalous login behavior. For example, a rule might specify that a login attempt from a new IP address should be considered anomalous if it occurs during off-peak hours.
- Machine learning: This involves using machine learning algorithms to identify patterns in login data that are indicative of anomalous behavior. Machine learning algorithms can be trained on historical data to learn what constitutes normal login behavior, and then used to identify login attempts that deviate from this norm.
- User behavior analytics: This involves analyzing login data to identify changes in user behavior that may be indicative of unauthorized access. For example, if a user suddenly starts logging in from a different IP address or at a different time of day, this could be a sign of a compromised account.

The use of anomalous login detection can help to improve the security of systems and data by identifying and preventing unauthorized access. By analyzing login data for patterns that deviate from normal behavior, security analysts can identify potential threats and take action to mitigate them.

# TABLE OF CONTENTS

# INTRODUCTION

In today's digital landscape, where cybersecurity threats are becoming increasingly sophisticated, organizations face the constant challenge of protecting their sensitive data and networks. One critical aspect of ensuring data security is detecting anomalous login activity. Anomalous logins refer to unauthorized or unusual login attempts that deviate from the normal patterns observed in a system. Analyzing security data to detect such anomalies has become a crucial practice for organizations aiming to identify potential breaches, prevent data loss, and safeguard their networks.

Security data analysis involves collecting and examining vast amounts of data generated by various systems, such as firewalls, intrusion detection systems, and authentication logs. By leveraging advanced analytical techniques and machine learning algorithms, organizations can sift through this data and identify patterns indicative of normal login behavior. This baseline understanding of regular login activity serves as a reference point for detecting any aberrations or deviations that might suggest unauthorized access attempts.

Detecting anomalous login activity typically involves analyzing several key parameters, including login frequency, time of day, geographic location, device type, and user behavior. By establishing thresholds and rules based on historical data, organizations can create models that identify deviations from the norm. For example, if a user typically logs in from a specific geographic location but suddenly attempts a login from a different continent, it could indicate a potential security breach.

Implementing robust security data analysis techniques not only helps in detecting external threats but also assists in identifying insider threats. Insider threats refer to instances where individuals within an organization with authorized access to sensitive information intentionally or unintentionally misuse or expose it. By monitoring and analyzing login behavior, organizations can identify unusual patterns exhibited by internal users, such as unusual login times, excessive failed login attempts, or accessing unauthorized resources.

The benefits of detecting anomalous login activity are manifold. By proactively identifying potential security breaches, organizations can respond swiftly and implement appropriate countermeasures to prevent unauthorized access. Furthermore, security data analysis assists in the forensic investigation of security incidents by providing a detailed log of login activities, aiding in incident response and post-incident analysis.

In conclusion, security data analysis plays a critical role in detecting anomalous login activity and safeguarding organizational networks and sensitive data. By leveraging advanced analytical techniques, organizations can identify deviations from normal login behavior, enabling them to respond swiftly to potential security breaches. With the constant evolution of cyber threats, implementing robust security data analysis practices is essential for organizations to stay one step ahead in the ongoing battle against unauthorized access and data breaches.

# METHODOLOGY

- **Data Preprocessing :** Data preprocessing is an essential step in data analysis and machine learning. It involves transforming raw data into a clean, structured, and organized format that can be effectively used for further analysis or modeling.

- **Exploratory Data Analysis (EDA):** Exploratory Data Analysis (EDA) is an essential step in the data analysis process that focuses on gaining insights and understanding the data before applying more advanced analysis techniques. EDA helps in uncovering patterns, identifying anomalies, and forming initial hypotheses.

- **Anomaly Detection:** Anomaly detection is the process of identifying patterns or instances within a dataset that deviate significantly from the expected or normal behavior. Anomalies, also known as outliers, can be caused by various factors such as errors, rare events, malicious activities, or system failures. Anomaly detection techniques aim to flag these anomalies for further investigation or action.

- **IP Address Geolocation Analysis:** IP address geolocation analysis is the process of determining the geographical location of an IP address based on its numeric value. It involves mapping the IP address to a specific location, typically represented by latitude and longitude coordinates, or by other geographical information such as city, region, or country.

# FINDINGS

Findings on the topic "Security Data Analysis - Detecting Anomalous Login Activity" can vary depending on the specific research or analysis conducted. However, here are some common findings and observations in this area:

1. Unusual Login Times: Analyzing login data can reveal patterns of unusual login times, such as logins occurring outside of regular business hours or during times when the user is not typically active. These findings can indicate potential unauthorized access attempts.

2. Geographic Anomalies: Monitoring the geographic locations of login attempts can help detect anomalous activity. For example, if a user typically logs in from a specific country or region but suddenly attempts a login from a different and unrelated location, it could be a sign of unauthorized access.

3. Device and IP Address Anomalies: Analyzing login data can also help identify anomalies related to device types and IP addresses. Unusual devices or IP addresses used for logins, especially those not associated with the user's typical devices or locations, can indicate unauthorized access attempts.

4. Failed Login Attempts: Monitoring the frequency of failed login attempts can provide insights into potential brute-force attacks or unauthorized access attempts. A sudden increase in failed login attempts for a specific user or from a specific IP address may indicate a security threat.

5. User Behavior Analysis: Analyzing login activity can help establish baseline behavior for each user. Deviations from this behavior, such as accessing unauthorized resources or performing unusual actions, can be indications of suspicious or malicious activity.

6. Insider Threat Detection: Security data analysis can be instrumental in identifying insider threats. By analyzing login data and comparing it against established user profiles and access privileges, organizations can identify unusual patterns of access or suspicious activities by employees or authorized users.

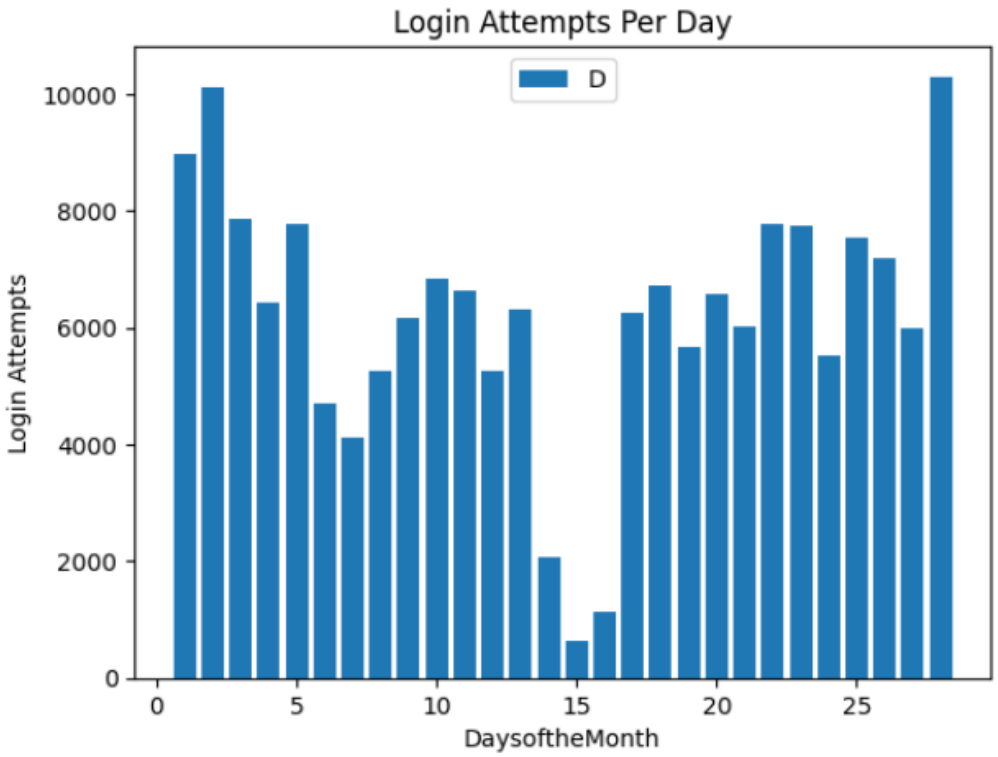
7. Real-time Alerting: Implementing real-time monitoring and analysis of login activity allows for immediate detection and response to anomalous login events. Automated alerting systems can notify security teams when potential security breaches or suspicious login activities are detected.

8. Continuous Improvement: Security data analysis should be an ongoing process that involves regularly updating models and thresholds based on new data and emerging threats. Continuous monitoring and analysis help organizations adapt and stay ahead of evolving security risks.

It's important to note that the specific findings and observations may vary depending on the dataset, analysis techniques, and the organization's unique security requirements and context. Conducting a comprehensive analysis of security data is crucial to tailor the findings to the specific needs of an organization and enhance its overall cybersecurity posture.

## Data Analysis :-

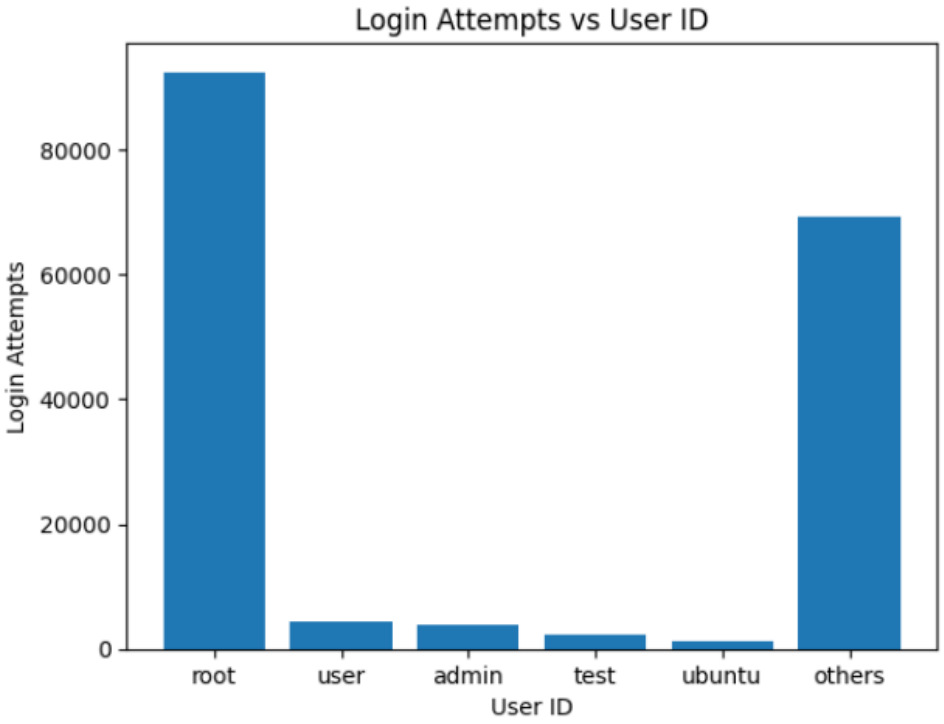
1. Graphical Analysis of Login Attempts per Day :




Graphical analysis of login attempts per day provides a visual representation of the frequency of login attempts over a period of time. This analysis helps in understanding login patterns, identifying anomalies, and detecting potential security threats.

Graphical analysis of login attempts per day serves as a valuable tool for visualizing and understanding login activity. It should be complemented with other security measures, such as anomaly detection algorithms, user authentication mechanisms, and real-time monitoring, to ensure comprehensive security for the system or network

2. Graphical Analysis of Login Attempts per User :

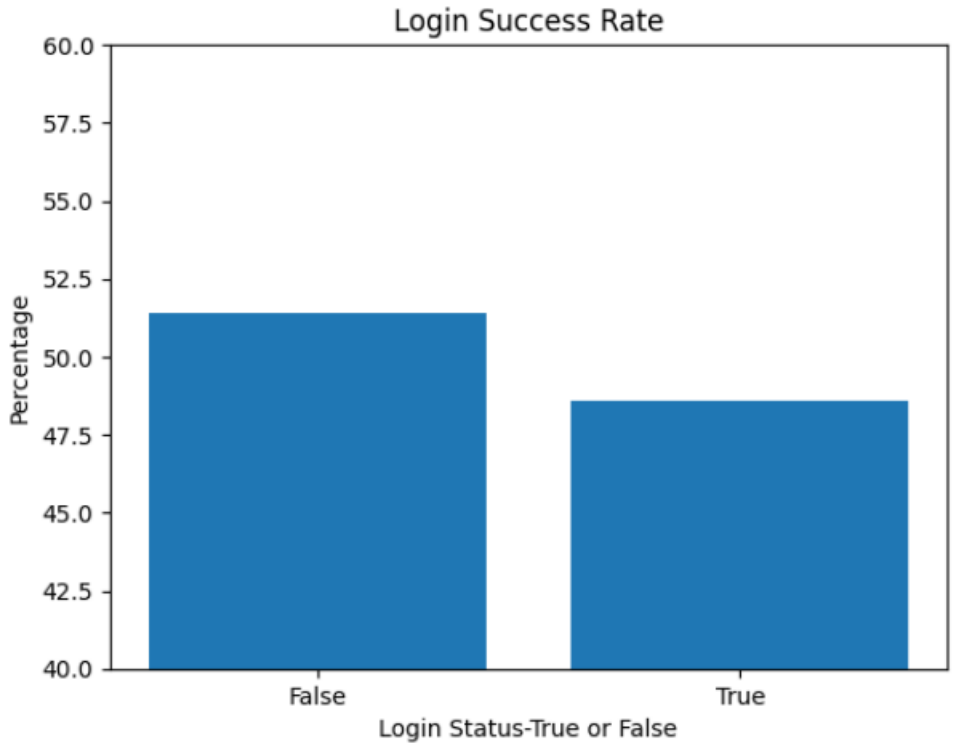



Graphical analysis of login attempts per user provides insights into individual user behavior, helping to detect anomalies, identify potential security threats, and understand patterns of login activity.

Graphical analysis of login attempts per user helps in understanding individual user behavior and detecting potential security issues at a granular level. It can be used to identify compromised accounts, detect insider threats, or investigate suspicious login patterns.

However, it is important to consider other security measures, such as multi-factor authentication, user access controls, and real-time monitoring, in conjunction with graphical analysis to ensure comprehensive security.

3. Graphical Analysis of Login Success Rate :




Graphical analysis of login success rate involves visualizing the percentage of successful login attempts over a given period. This analysis helps in understanding the effectiveness of login processes, detecting authentication issues, and identifying potential security concerns.

Graphical analysis of login success rate provides a clear and intuitive representation of the effectiveness of the login process. It helps in identifying areas of improvement, optimizing authentication mechanisms, and detecting potential security vulnerabilities. It should be complemented with other security measures, such as monitoring failed login attempts, implementing strong password policies, and conducting regular security assessments, to ensure a robust security posture.

# Anomaly Detection :

Anomaly detection is a technique used to identify patterns or instances that deviate significantly from the expected or normal behavior within a dataset. It plays a crucial role in various domains, including cybersecurity, fraud detection, network monitoring, and system health monitoring. Here are some key points to understand about anomaly detection:

1. Definition of Anomalies: Anomalies, also known as outliers, are data points or patterns that do not conform to the expected or typical behavior within a dataset. They can be caused by various factors, including errors, rare events, malicious activities, system failures, or novel patterns.
2. Unsupervised and Supervised Approaches: Anomaly detection techniques can be categorized into unsupervised and supervised methods. Unsupervised approaches do not require labeled data and focus on identifying patterns that are different from the majority. Supervised approaches rely on labeled data, where anomalies are explicitly identified and used to train a model.
3. Statistical Methods: Statistical methods involve analyzing data distribution and using statistical measures such as mean, standard deviation, or probability distribution models to detect anomalies. Data points that fall outside certain statistical thresholds or exhibit significant deviations are flagged as anomalies.
4. Machine Learning Techniques: Machine learning techniques, such as clustering, classification, and dimensionality reduction algorithms, can be applied to anomaly detection. These methods learn patterns and behaviors from training data and identify instances that do not fit into the learned model as anomalies.
5. Time Series Analysis: Anomaly detection in time series data involves identifying deviations from expected temporal patterns. Techniques like autoregressive integrated moving average (ARIMA), exponential smoothing, or change point detection can be employed to detect anomalies in time-dependent data.

6. Ensemble Methods: Ensemble methods combine multiple anomaly detection algorithms to improve accuracy and robustness. They leverage the strengths of different techniques and aggregate their results to provide more reliable anomaly detection outcomes.
7. Domain-specific Approaches: Anomaly detection methods can be tailored to specific domains or applications. For example, in network security, intrusion detection systems analyze network traffic patterns to identify suspicious activities. In fraud detection, anomaly detection algorithms can flag transactions that deviate from typical spending patterns.
8. Evaluation and Validation: Anomaly detection algorithms should be carefully evaluated and validated to ensure their effectiveness. This typically involves using labeled datasets with known anomalies and assessing the algorithm's performance in terms of detection rate, false positive rate, precision, and recall.
9. Continuous Learning and Adaptation: Anomaly detection systems should be designed to adapt and learn from new data, as normal behavior patterns may evolve over time. Continuous monitoring and retraining of models are essential to maintain accuracy and effectively detect emerging anomalies.
10. Human-in-the-Loop: While automated anomaly detection systems are valuable, human involvement is crucial for interpreting and investigating flagged anomalies. Human experts can provide domain knowledge, contextual understanding, and make final judgments about whether an identified anomaly is indeed a threat or a benign outlier.

Anomaly detection is a powerful technique for detecting unusual and potentially malicious activities or patterns within datasets. By identifying anomalies, organizations can proactively respond to security threats, prevent fraud, maintain system health, and improve overall operational efficiency.

**Anomaly Detection Methods :**

There are several methods commonly used for anomaly detection, each with its own strengths and suitable applications. Here are some of the commonly used methods for anomaly detection:

1. Statistical Methods:
    - Z-Score: This method calculates the standard deviation of the data and identifies data points that fall outside a specified threshold.
    - Modified Z-Score: Similar to the Z-score method but uses the median and median absolute deviation for better robustness against outliers.
    - Gaussian Distribution: Assumes that the data follows a Gaussian (normal) distribution and flags data points that deviate significantly from the expected distribution.
    - Box Plot: Uses quartiles and interquartile range to identify outliers based on data distribution.

2. Machine Learning Techniques:
    - Clustering: Unsupervised clustering algorithms like k-means or DBSCAN can identify data points that do not belong to any cluster, treating them as anomalies.
    - Classification: Supervised learning algorithms like decision trees, random forests, or support vector machines can be trained on labeled data to classify instances as normal or anomalous.
    - One-Class SVM: Trains on normal data and constructs a boundary to identify deviations as anomalies.
    - Autoencoders: Neural network models that learn to reconstruct normal data and flag instances with high reconstruction error as anomalies.

3. Time Series Analysis:
    - Moving Average: Compares data points to a rolling average over time and flags deviations outside a specified threshold.
    - Exponential Smoothing: Uses weighted averages to identify unexpected changes in the time series.
    - Seasonal Decomposition: Separates time series data into trend, seasonal, and residual components, enabling the identification of anomalies in each component.

4. Density-Based Methods:
    - Local Outlier Factor (LOF): Measures the density of data points relative to their neighbors and identifies outliers with lower density.
    - Isolation Forest: Constructs isolation trees to isolate anomalies that can be identified with fewer splits compared to normal instances.

5. Spectral Methods:
    - Principal Component Analysis (PCA): Projects high-dimensional data into a lower-dimensional space while preserving the most important information. Anomalies can be identified by their significant deviation from the expected projection.
    - Singular Value Decomposition (SVD): Decomposes the data matrix into singular values and vectors, highlighting anomalies that exhibit unexpected singular values.

6. Deep Learning Methods:
    - Recurrent Neural Networks (RNN): Designed for sequential data, RNNs can capture temporal dependencies and identify anomalies based on deviations from expected patterns.

- Generative Adversarial Networks (GAN): Employed to generate realistic data samples, GANs can flag instances that do not conform to the learned data distribution as anomalies.

It's important to note that the choice of the most suitable anomaly detection method depends on factors such as the data characteristics, available labeled data, computational resources, and the specific problem domain. In practice, a combination of methods or an ensemble approach is often used to improve detection accuracy and robustness. Additionally, the effectiveness of anomaly detection methods relies on proper preprocessing, feature engineering, and regular evaluation against labeled datasets or expert feedback to fine-tune the detection thresholds.

Out of these methods, we have applied the z-score method to compute the anomalies from the data frame and also used it to obtain the anomality free data frame. The snippet of the same is as follows:

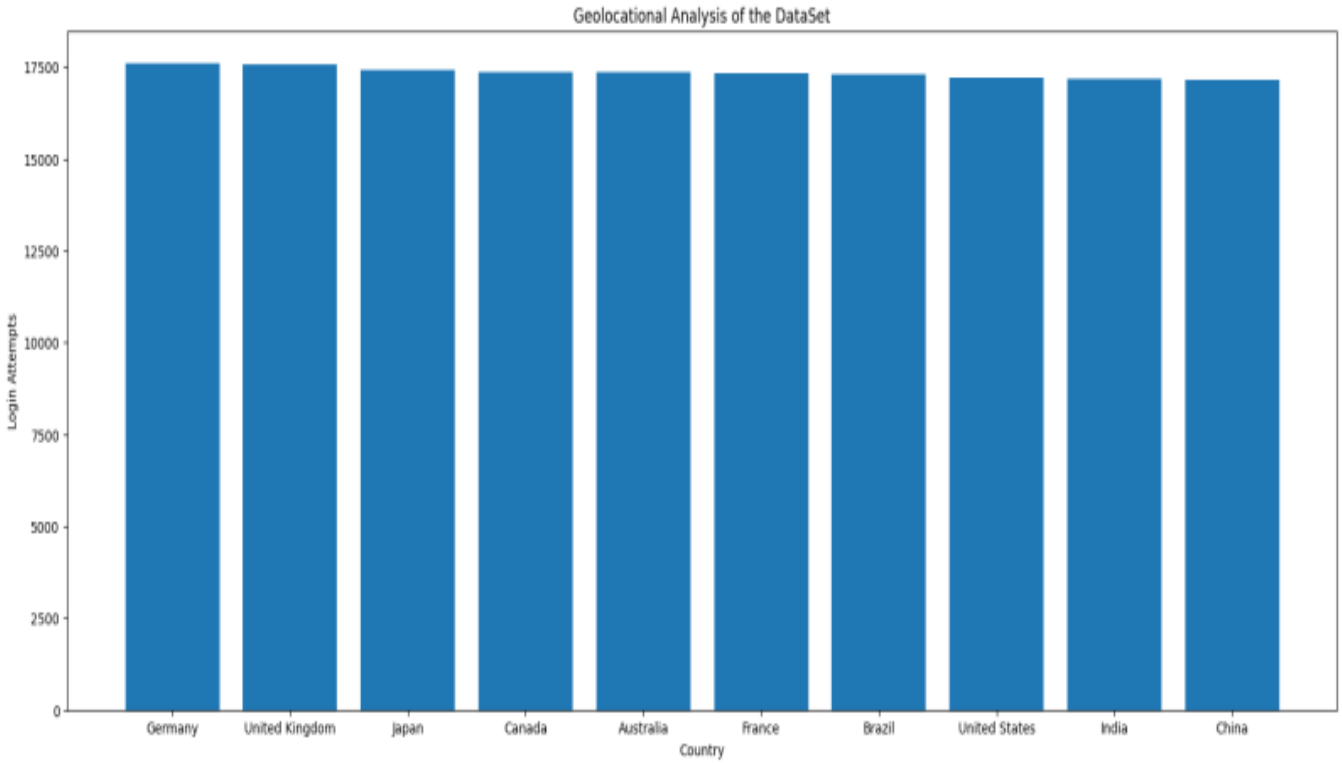
```
In [69]: data['Timestamp_Num'] = (data['Time'] - data['Time'].min()).dt.total_seconds()

         # Calculate mean and standard deviation for both successful and failed logins
         mean_successful = data[data['Login Status'] == True]['Timestamp_Num'].mean()
         std_successful = data[data['Login Status'] == True]['Timestamp_Num'].std()

         mean_failed = data[data['Login Status'] == False]['Timestamp_Num'].mean()
         std_failed = data[data['Login Status'] == True]['Timestamp_Num'].std()

         # Calculate z-scores for successful and failed login attempts
         data['Z-Score'] = np.where(data['Login Status'] == True,
                                    (data['Timestamp_Num'] - mean_successful) / s
                                    (data['Timestamp_Num'] - mean_failed) / std_f

         # Set a threshold for anomaly detection
         anomalous_logins = data[(data['Z-Score'] > 1.5) | (data['Z-Score'] < -1.5)]

         print("Anomalous Login Attempts:")
         print(anomalous_logins)
```

By, using the method we were able to obtain the anonymous data from the data frame.

## Geolocational Analysis :

Geolocation analysis is a process that involves analyzing and interpreting data based on geographical information. It leverages location data, such as latitude and longitude coordinates or addresses, to gain insights, understand spatial patterns, and make informed decisions.

Geolocation analysis offers valuable insights by integrating spatial data and analysis techniques. By understanding the spatial relationships and patterns in data, organizations can make informed decisions, optimize operations, and gain a deeper understanding of various phenomena within a geographic context.



## Results:

- Results in an anomaly detection project can vary depending on the specific dataset, anomaly detection methods used, and the objectives of the project.
- But, as in this project we successfully analyzed the dataset by using Machine Learning with the help of Python Modules.
- We analyzed the whole dataset by loading it to a data frame , the data is then used to figure out the login success rate as well as helped us to identify the anomalies in the dataframe.
- The anomalies then is detected and removed by using the statistical methods like Z-Score to figure out the anomalies in the given data and helped to obtain a anomaly free dataframe.
- We also analyzed geolocation patterns in the given data frame which gave us a idea about the login activities made from different locations.

## RECOMMENDATIONS:

Here are some recommendations for security data analysis to effectively detect anomalous login activity:

1. Establish Baseline Behavior: Build a baseline of normal login behavior by analyzing historical data. Understand typical patterns such as login times, locations, devices, and user behavior for different user roles. This baseline will serve as a reference to identify anomalies.
2. Implement Machine Learning Algorithms: Utilize machine learning algorithms to analyze login data and detect anomalies. Algorithms such as clustering, classification, or anomaly detection models can be trained on historical data to identify deviations from normal behavior.
3. Multi-factor Authentication (MFA): Implement multi-factor authentication as an additional layer of security. MFA helps mitigate the risk of unauthorized access even if login credentials are compromised. Analyze MFA logs and flag any abnormalities in the authentication process.

4. Real-time Monitoring: Implement real-time monitoring of login activity to detect anomalies as they occur. Set up alerts or notifications to promptly notify security teams of suspicious login behavior, enabling them to take immediate action.
5. User Behavior Analytics (UBA): Utilize user behavior analytics to analyze login patterns and detect anomalies specific to individual users. UBA models can learn and understand normal user behavior, making it easier to identify deviations and potential insider threats.
6. Contextual Analysis: Incorporate contextual information such as IP reputation, device reputation, and geographic data into the analysis. Anomalies that involve unusual IP addresses, unknown devices, or login attempts from unexpected locations can be flagged for further investigation.
7. Collaborative Threat Intelligence: Leverage shared threat intelligence platforms or collaborate with other organizations to stay informed about known threats and attack vectors. This shared knowledge can help identify anomalous login activity associated with known malicious actors or patterns.
8. Regular Data Review: Continuously review and update the anomaly detection models based on new data and emerging threats. Regularly analyze and refine the thresholds, rules, and algorithms to improve detection accuracy and stay ahead of evolving threats.
9. Incident Response and Forensics: Establish incident response procedures and protocols to effectively handle detected anomalies. This includes preserving relevant log data, conducting forensic analysis, and taking appropriate actions to mitigate the impact of security incidents.
10. Employee Education and Awareness: Educate employees about best practices for login security, including password hygiene, recognizing phishing attempts, and reporting suspicious activity. Well-informed employees can serve as an additional line of defense against anomalous login attempts.

Remember, security data analysis for detecting anomalous login activity should be a comprehensive and ongoing process.