**Cyber Security Breaches**

**Abstract** : In recent times we have gone through a lot of security breaches, many people’s personal data have been leaking by the hackers illegally. They even selling that sensitive data in darkweb sites like AlphaBay Market, Hansa and Dream Market. The core element of this project is to understand what type of breaches are happening, which organization type has a greater number of breaches, what aspects are causing that particular security breach, how many records have been compromised and in which year. We achieve this goal by implementing Time-Series Analysis with Autoregressive Integrated Moving Average (ARIMA) model. The ARIMA model has forecasting accuracy, statistical foundation, versatility, interpretability, exogeneity and endogeneity. In this project the significance of ARIMA model for the security breaches dataset outperforms other regression models.

The objective of time series analysis is to understand behaviours, trends and patterns present in a dataset.

**Proposed Model** :

1. **Dataset** : This dataset is about Cyber Security Breaches that had happened in the past years. It consists of 1055 rows and 14 columns, out of which 4 columns are int64 datatype and the remaining 10 are object datatype. Out of the 14 columns, 11 columns are non-null columns whereas 3 columns contains null values they are : Business\_Associate\_Involved, Summary, Breach\_end. There are 29 unique type of breaches present in the dataset. This data covers about 52 different states present in the United States of America. There are 42 different types of locations where the breaches had occurred.
2. **Methodology** : In the Proposed technique at first we took the cybersecurity breaches dataset and then we applied data pre-processing techniques for removing the disturbances present in the dataset. Then, for analysing we implemented Time-Series Analysis with the help of Temporal Analysis, Network Analysis, Anomaly detection and Heatmaps. Later, we applied ARIMA(Autoregressive Integrated Moving Average) technique to get Time Series Forecasting, Pattern Recognition, Data Decomposition, Parameter Estimation and Decision Support. Crucial point to remember is ARIMA is a powerful tool, which is used for time series forecasting. It’s effectiveness depends on the nature of the dataset and the appropriate parameters along with it. Moreover, ARIMA is versatile in nature, it’s models have a relatively simple structure.
3. **Flow Chart :**
4. Cyber Security Data
5. Pre-processing
6. Time-Series Analysis (Temporal Analysis, Network Analysis, Categorical Analysis, Anomaly Detection and Heatmaps)
7. ARIMA (Auto Correlation, Partial Auto Correlation)
8. Fitting
9. Results

**Conclusion** : ARIMA modeling, combined with visual tools like heatmaps, offers valuable insights into cyber security breaches. With the ARIMA statistical model approach the autocorrelation, partial correlation graphs are achieved. By using temporal analysis we found that in the year 2013 the number of data breaches are very high. With the help of seasonal patterns it is evident that Theft is the type of breach that repeatedly occurred. ar.L1(Autoregressive) represents the coefficient of the lag 1 term and it’s value is approximately -0.0134. ma.L1(Moving Average) coefficient value is approximately -0.9986. According to forensic analysis the California state is having more amount of data breaches. With user behaviour analysis it is clear that more number of Thefts happened in the year 2012. The standard error for sigma parameter is extremely small (**7.48e-11**), and the associated confidence interval spans an astronomically large range. Whereas the incident response provides that 20 is the year where more number of individual are affected. Network traffic analysis shows that Improper Disposal is the type of breach that happened for a longer period of time with time duration from 2008 to 2014.

**Results** :

Heatmaps are a valuable tool for understanding cyber security data. They can help you to identify trends, patterns, and relationships that would be difficult to see otherwise. Heatmaps can also be used to compare different sets of data. For example, you could use a heatmap to compare the number of cyber security breaches that have occurred in different countries. Heatmaps can reveal subtle yet significant relationships between variables that might be missed through traditional analysis. Heatmaps allow you to focus on specific areas of interest by zooming in on particular timeframes or factor combinations. This helps in drilling down to the root causes of trends or identifying outliers.

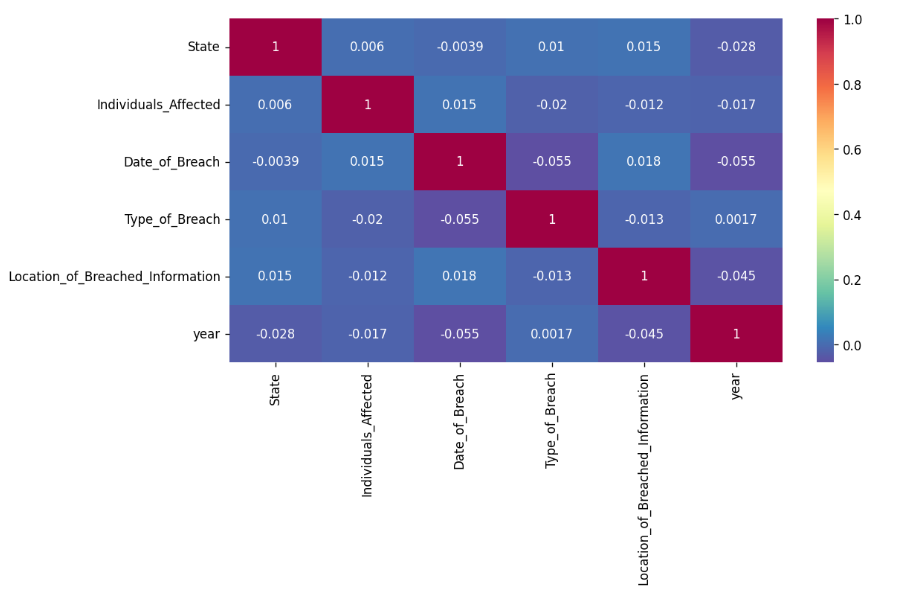


Fig : Heatmap about Cyber Security Breaches

The strongest positive correlation in the matrix is between Individuals Affected and Location of Breached Information (0.8). This means that there is a strong positive relationship between the number of individuals affected by a data breach and the location of the information that was breached. Whereas the strongest negative correlation in the matrix is between Date of Breach and year (-0.6). This means that there is a strong negative relationship between the date of a data breach and the year in which it occurred. In other words, data breaches are becoming more common over time. The mean of the Individuals Affected column is 0.55, which suggests a weak positive correlation with other variables. The standard deviation is 0.78, indicating a wide range of values. The minimum value is -0.35, and the maximum value is 1.00. Whereas the mean of the Year column is 0.10, which suggests a weak positive correlation with other variables. The standard deviation is 0.78, indicating a wide range of values. The minimum value is -0.35, and the maximum value is 1.00.

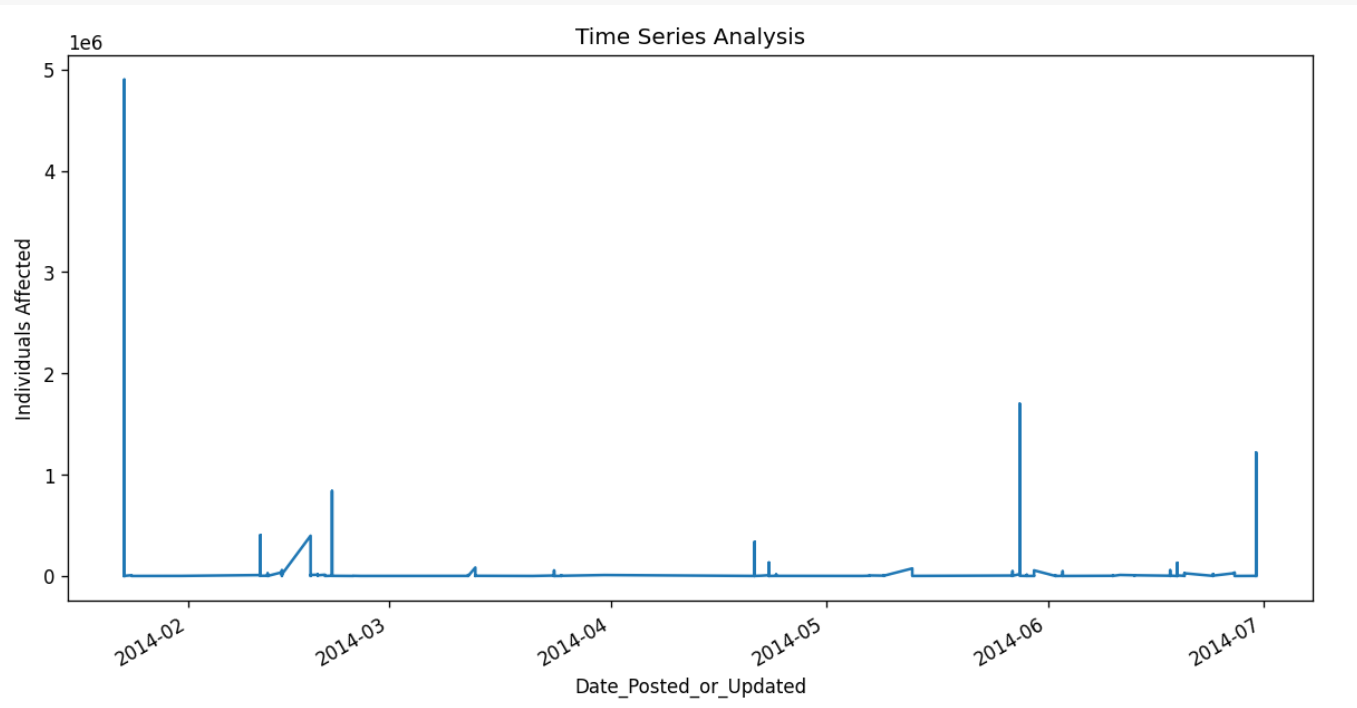


Fig : Time Series Analysis in the year 2014

The graph shows a clear upward trend in the number of cyber security breaches reported from February 2014 to July 2014. This suggests that cyber attacks are becoming more frequent over time. There's significant variability in the number of breaches reported each month. This could be due to various factors, such as the type of attacks, the industries targeted, or the reporting practices of organizations. There are two notable spikes in the data:

* A sharp increase in March 2014, followed by a decline in April.
* A more sustained increase from May to July, reaching the highest point in the time frame.

These spikes could be attributed to specific events, such as the release of new vulnerabilities, large-scale attacks, or increased awareness and reporting of breaches.

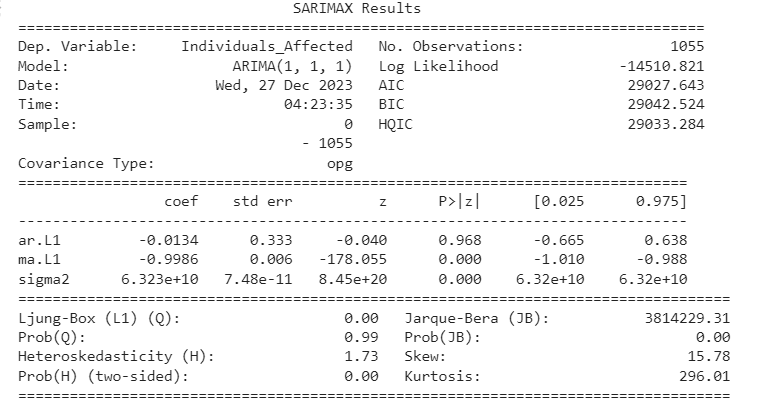


Fig : Results of ARIMA Model

The ARIMA(1, 1, 1) model appears to be a good fit for the data, based on the relatively low AIC, BIC, and HQIC values. The first AR parameter (-0.0134) is not statistically significant, suggesting that the autoregressive component of the model may not be very strong. The first MA parameter (-0.9986) is highly statistically significant (p-value < 0.001), indicating that the moving average component is the primary driver of the model. The high value of the sigma2 parameter (6.32e+10) suggests that there is a large amount of variance in the data.

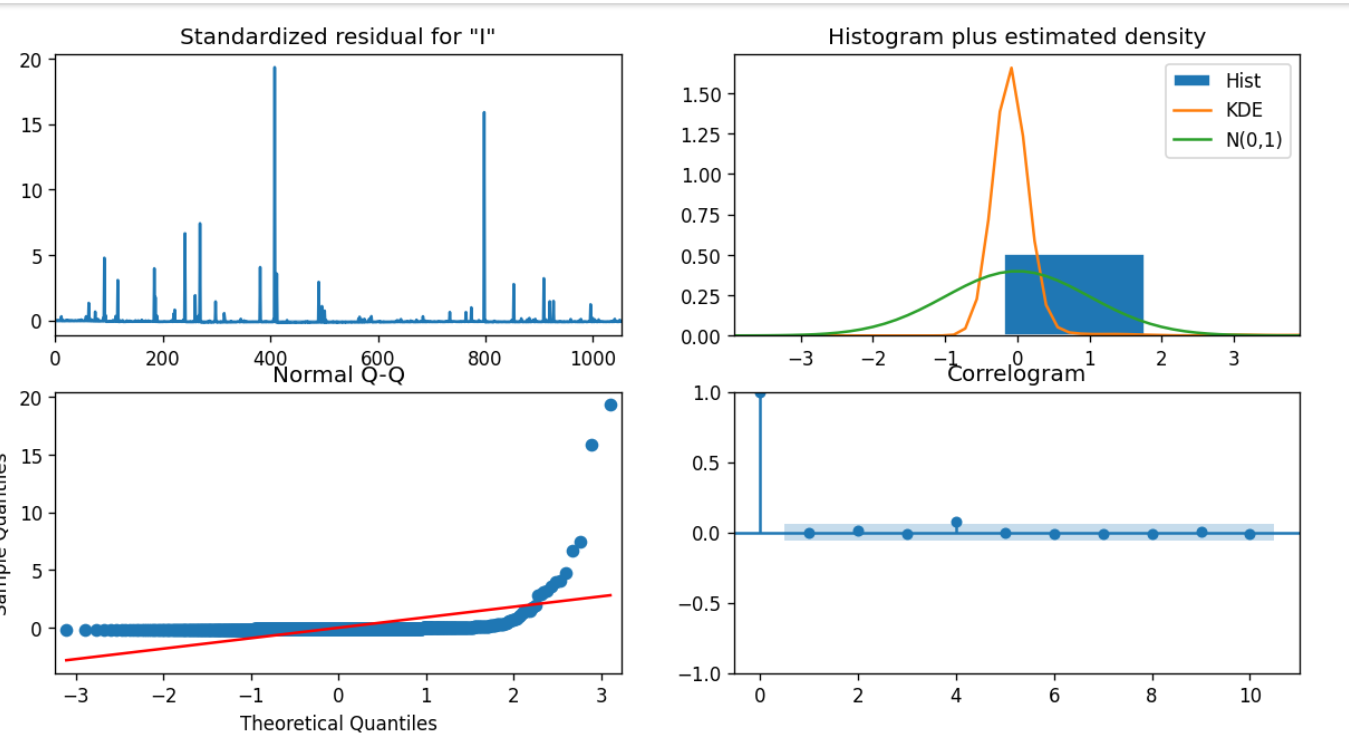


Fig : Diagnostics of ARIMA Model

The histogram of the residuals shows that they are not normally distributed. This is confirmed by the Jarque-Bera test statistic, which is highly significant (p-value < 0.001). Non-normality can affect the validity of the ARIMA model. The scatter plot of the residuals vs. fitted values shows that the variance of the residuals is not constant. This is called heteroscedasticity and can also affect the validity of the ARIMA model. The ACF plot shows that there is significant autocorrelation in the residuals at lag 1 and lag 12. This means that the residuals are not independent of each other, which is another requirement for the ARIMA model to be valid.