Dallas Crime Forecasting:

Predictive Policing Using Machine Learning

ADTA 5940 Section 002:

**Analytics Capstone Experience**

Dr. Sarah Quintanar

PHD, Clinical Associate Professor



**UNIVERSITY OF NORTH TEXAS**

**TOULOUSE GRADUATE SCHOOL**

**MSc in Advanced Data Analytics**

-Sai Prathyusha Thirunahari

**TABLE OF CONTENTS**

**CHAPTER 1 - INTRODUCTION ............................................................................................... 3** **CHAPTER 2 – LITERATURE REVIEW ................................................................................. 5** **CHAPTER 3 –METHODOLOGY………..…………………….….……………......................7 CHAPTER 4 –RESULTS…..……….…………….................................................................. 11** **CHAPTER 5– CONCLUSION ............................................................................................... 26** **REFERENCES.............................................................................................................................30APPENDIX….. ........................................................................................................................... 33**

1. **INTRODUCTION**

What if we could predict where the next crime might happen—before it occurs—using the previous data collected by police?

The Dallas Police Dataset is the wide collection of records that the Dallas Police Department collects and maintains for law enforcement. This dataset(Police Incidents | Dallas Open Data, 2025) is widely used by researchers, policymakers, and data analysts to examine crime patterns, assess the effectiveness of policing, and look into disparities in law enforcement practices. Despite these potential limitations which exist for most crime datasets, this research will ultimately lead to more knowledgeable conversations on justice and public safety.

Our research goal is to use machine learning models to predict and develop models for analyzing crime patterns, Identifying crime hotspots, Inspecting the reports and predicting the future crimes. This historical data we estimate using supervised learning models, such as logistic regression, decision trees, and neural networks, we may identify places with a higher probability of incidents and predict crime hotspots. Unsupervised learning techniques like clustering can also be used to identify irregularities in law enforcement operations and group related criminal patterns. Policymakers and law enforcement organizations can use these insights to help them make data-driven decisions that will enhance public safety and the distribution of resources.

**Data:**

As part of a larger open data movement, the Dallas Police Incidents dataset was created on July 17, 2018, with the goal of increasing transparency in law enforcement operations. With 86 variables and 6.91 million unique occurrence recordings, our dataset spans the years 2020–2025. The data include Text (such as incident number, location, and officer details), numeric (such as zip codes and geographic coordinates), date and time variables (such as occurrence dates and reporting times), and binary variables (such as hate crimes and family offenses).

Although the dataset is openly accessible and has attracted a lot of interest—as of February 2025, it had received over 305,000 views and 33,200 downloads—it does have some restrictions. Problems including misclassification mistakes, gaps in historical data, and the omission of sensitive information like juvenile instances or sexual offenses persist despite efforts to maintain data quality through automated validations and manual checks.

1. **LITERATURE REVIEW**

A predictive approach to crime analysis is now possible because to the enormous transformation brought about by the combination of big data analytics and machine learning. Advanced analytical methods have been used to find crime trends, evaluate danger regions, and improve law enforcement tactics. The study by Hashim et al. (2019) highlights the increasing importance of machine learning in evaluating police incident data and demonstrates how it outperforms conventional statistical techniques in forecasting crime patterns and enhancing preventative strategies. Their study emphasizes the value of geographical and temporal insights in crime pattern identification by demonstrating how spatial regression and emerging hotspot analysis efficiently map urban crime. The predictive skills of crime forecasting have been considerably improved by integrating deep learning. Zhuang et al. (2017) showed how temporal and spatial data may be used to forecast crime hotspots. They recommended that using long short-term memory (LSTM) models and recurrent neural networks (RNNs) can capture intricate correlations and long-term dependencies in crime data. The limitations of conventional methods like support vector machines (SVM) and kernel density estimation (KDE) in recognizing nonlinear correlations are also highlighted by Kang et al. (2017) as they analyzed a variety of machine learning models. Through the integration of demographic, economic, and environmental elements, deep neural networks (DNNs) and spatio-temporal neural networks (STNNs) enhance crime predictions, and their study supports their research use.

Jain and Bhat (2022), who stressed the need to integrate spatial and temporal data to improve crime hotspot identification, further reinforce the importance of spatial-temporal crime prediction. When applying these techniques to the dataset, their research emphasizes the significance of using uniform evaluation measures to gauge model efficacy. The function of computer vision in crime prevention is also examined by Shah et al. (2021), who showed how combining machine learning and visual data analysis can enhance detection and reaction tactics. By adding image-based analytics for real-time crime monitoring, their method could improve our analysis of the Dallas Police RMS dataset.

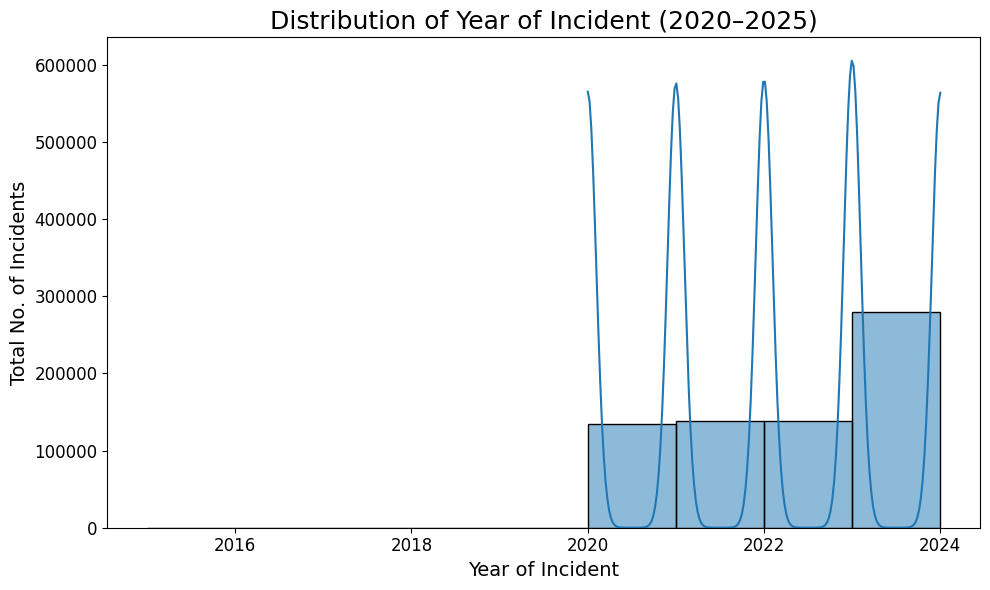
Additionally, recent research has emphasized the value of exploratory data analysis and visualization in the study of crime. Using sophisticated visualization tools like Power BI, Odooh et al. (2024) carried out a thorough analysis of Chicago's crime trends in order to identify trends in the distribution of crimes and clearance rates. Their results highlight how useful data-driven insights are for influencing policy and assisting law enforcement. In a similar vein, Mishra (2024) explored the wider industrial uses of AI and machine learning, showing how data-driven decision-making may be used for crime prevention and prediction. The findings of Khan, Ali, and Alharbi (2022) who used San Francisco crime data to successfully apply classification algorithms, such as random forests and decision trees, to forecast crime types and locations, are consistent with their observations. Kim et al. (2018) investigated crime pattern analysis using clustering and classification algorithms, which advanced the discipline. Their research demonstrates how machine learning may help law enforcement organizations with strategic planning and resource allocation by revealing hidden links in crime data. We can improve predictive models, create more focused intervention tactics, and deepen our understanding of crime dynamics by using their approaches on the Dallas dataset.

1. **METHODOLOGY**

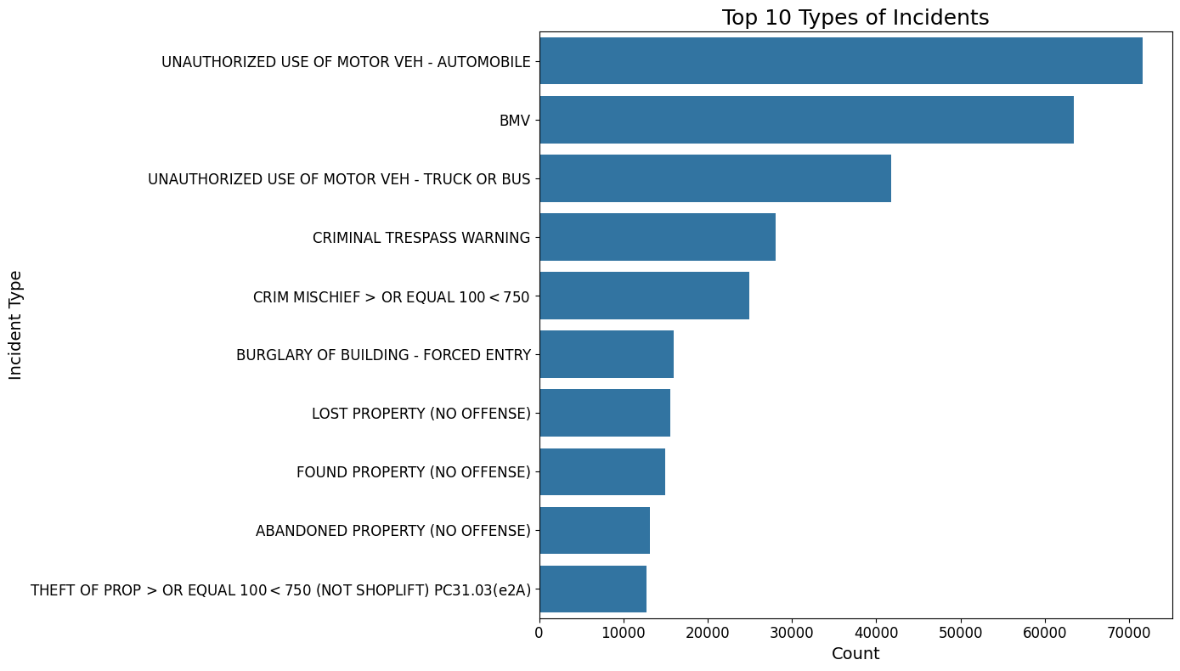
For our research, various machine learning (ML), deep learning (DL), and natural language processing (NLP) techniques are used to analyze crime trends and predict future incidents. We performed data cleaning on dataset to remove missing or inconsistent values, Columns with more than 50% missing data were eliminated to ensure data quality. Missing values in the remaining numerical columns were imputed with the median to reduce skewing. To guarantee uniformity, categorical variables were filled using the mode.We used NLP methods to analyze police reports and extract key insights from textual descriptions. ML models such as decision trees and logistic regression are used for classification tasks, while deep learning techniques like recurrent neural networks (RNNs) or transformers are applied for text-based crime pattern analysis. We have used Python throughout the research, along with packages like as NLTK, Scikit-learn, NumPy, pandas, and TensorFlow. We divided the dataset into training(80%) and testing(20%) sets in order to assess the performance of the model.

To prepare our dataset for machine learning models, we utilized various encoding techniques to convert non-numerical categorical features into numerical representations. By assigning unique integer values to each category, this methods allows the models to interpret and analyze the data effectively. Additionally, to maintain data integrity and improve model performance, To encode categorical variables, we used label encoding to translate non-numeric features into numerical form. For example, integer codes were allocated to Call (911) Problem and Type Location. Outliers in numerical columns (e.g., Incident Density) were identified and deleted using z-scores with a ±3 standard deviation threshold.

To determine how the data is distributed, we examined the distribution of numerical features in our dataset.

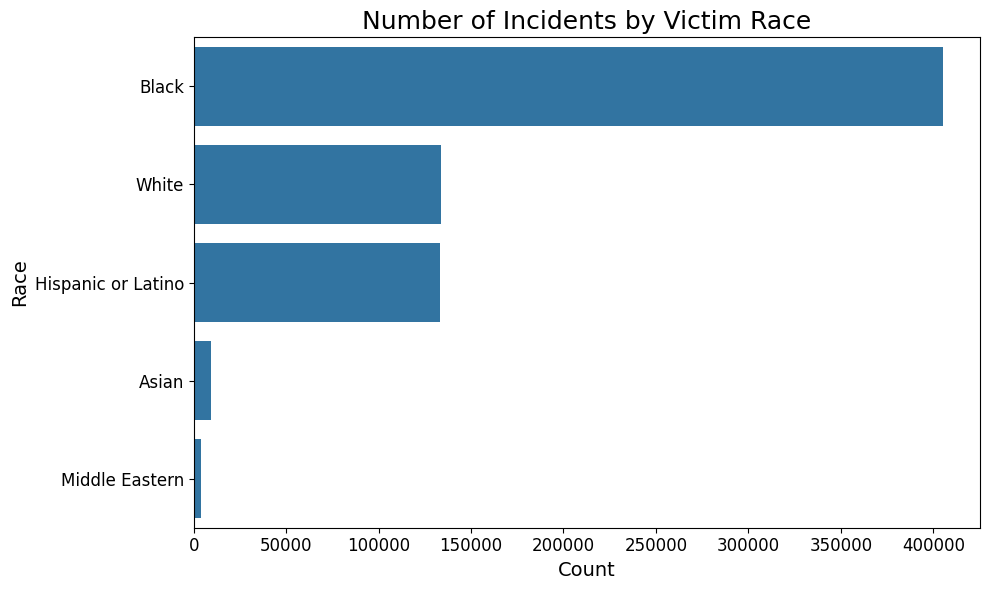


**Figure 3.1:** Distribution of Crime Incidents over years(2020-2025)



**Figure 3.2:** Top 10 type of incidents

We have plotted the number of incidents that occurred in each year from 2020 to 2025 using a histogram chart. We can see that there is increase in incidents following the COVID-19 pandemic in comparison to prior years. Unauthorized use of a vehicle is the most common form of incident, followed by BMV(Bureau of Motor Vehicles), according to our bar chart analysis of the top ten incident types.



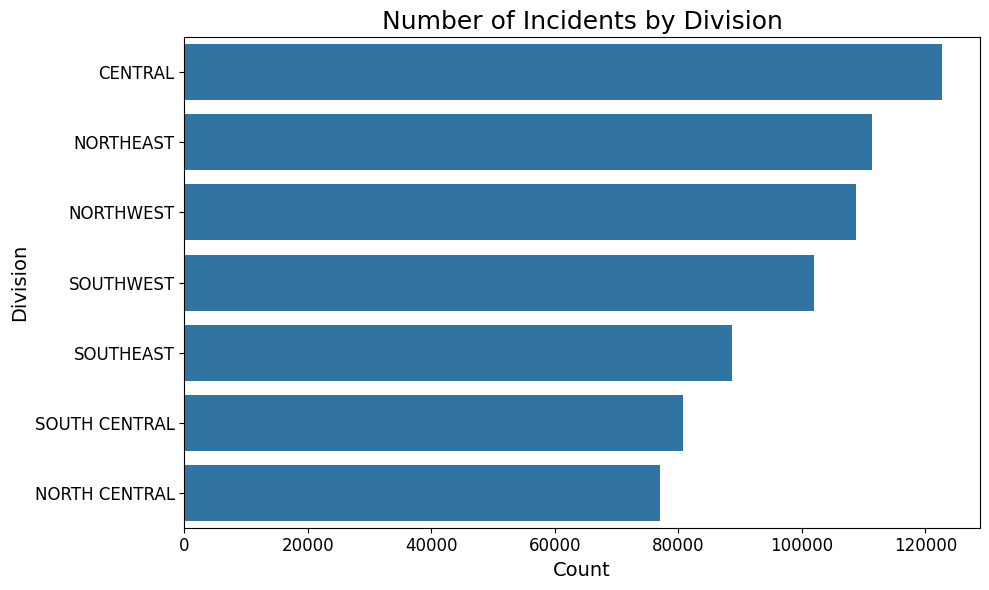
**Figure 3.3:** Distribution of incidents by race of offender

We have divided the overall number of incidences by race using a bar chart, showing that Black people are at the top, followed by Hispanic or Latino people, and then white people. We classified based on gender (male/female) in another chart. In this case, we found that men have been involved in more incidents than women.

A graph of a number of incident

AI-generated content may be incorrect.

**Figure 3.4:** Distribution of incidents by gender



**Figure 3.5:** Distribution of incidents by division

The division of dallas regions are been categorized here, and the total number of occurrences that have occurred has been determined. The central division has the most incidences, followed by the northeast, northwest, southwest, southeast, and so on

1. **RESULTS**

**RQ -1:** What is the accuracy of a multi-class classification model in predicting the type of crime, for example, assault, burglary, and theft, using incident features such as the time and place of the incident, as well as past crimes recorded in the area?

In our multi-class crime classification analysis, we tested machine learning models, including Logistic Regression, Decision Trees, and Random Forests, to predict the kind of crime based on incident types such as time, location, and area-specific crime characteristics. We trained and assessed the models for accuracy, precision, recall, and F1-score to determine their efficacy in detecting crimes of each type including assault, burglary, and theft.

The effectiveness of predictive analytics in crime forecasting is highlighted by this study's examination of machine learning algorithms, which showed how various models handle crime classification. Random Forest proved to be the most resilient of the models evaluated, with a macro F1-score of 0.63 and an accuracy of 68%, demonstrating its ability to successfully strike a compromise between recall and precision. Its capacity to identify intricate connections in crime data, such as the influence of time, place, and past crime patterns, is what accounts for its exceptional performance.

Random Forest consistently outperformed Decision Tree and Logistic Regression in all crime categories, according to model comparisons. For example, it demonstrated good dependability in recognizing Unauthorized Use of Motor Vehicle – Truck or Bus, achieving the highest F1-score of 0.76. It further shown its versatility in managing various crime kinds by maintaining a balance between precision and memory while forecasting BMV and Criminal Trespass Warning. In comparison, Decision Tree occasionally came close to Random Forest's performance but lacked its overall stability, and Logistic Regression did the worst, finding it difficult to recognize intricate criminal patterns because of its linear character.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Crime Category | Metric | Random Forest | Logistic Regression | Decision Tree |
| Vehicle Theft (Automobile) | Precision | 0.68 | 0.41 | 0.69 |
| Recall | 0.83 | 0.55 | 0.80 |
| F1-score | 0.75 | 0.47 | 0.74 |
| Burglary of Motor Vehicles (BMV) | Precision | 0.65 | 0.45 | 0.61 |
| Recall | 0.61 | 0.50 | 0.49 |
| F1-score | 0.63 | 0.47 | 0.54 |
| Vehicle Theft(Truck/Bus) | Precision | 0.75 | 0.35 | 0.65 |
| Recall | 0.77 | 0.18 | 0.77 |
| F1-score | 0.76 | 0.23 | 0.71 |
| Trespassing | Precision | 0.71 | 0.55 | 0.62 |
| Recall | 0.67 | 0.65 | 0.55 |
| F1-score | 0.69 | 0.60 | 0.58 |
| Money Theft | Precision | 0.51 | 0.47 | 0.33 |
| Recall | 0.27 | 0.18 | 0.28 |
| F1-score | 0.35 | 0.26 | 0.30 |
| Overall Metrics | Accuracy | 0.68 | 0.44 | 0.62 |
| Macro Avg F1 | 0.63 | 0.41 | 0.57 |
| Weighted F1 | 0.67 | 0.42 | 0.61 |

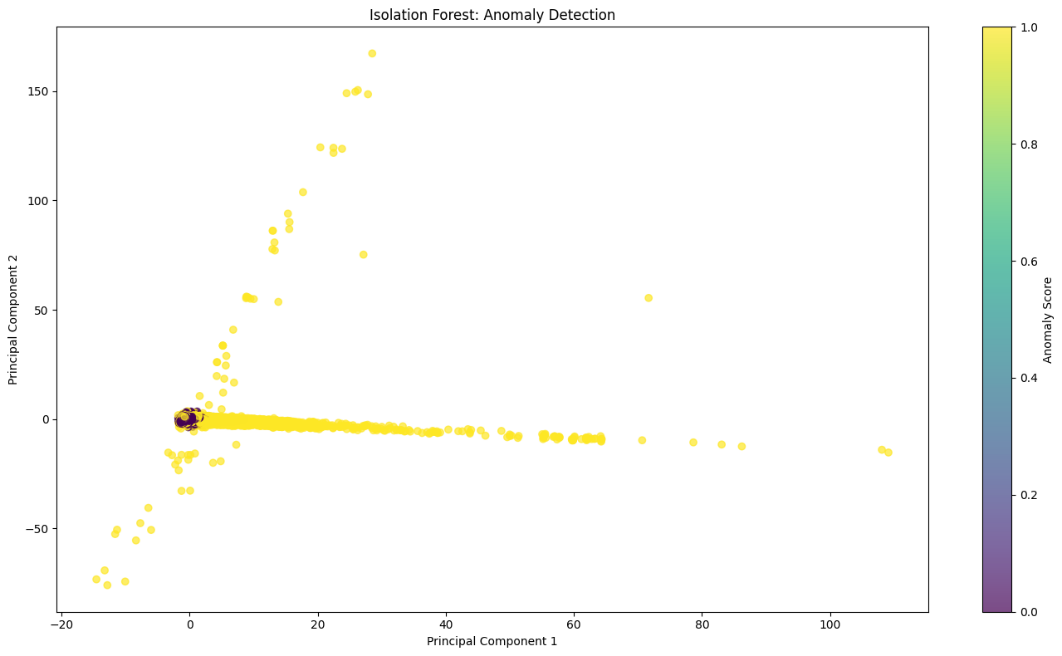
**Table 4.1:** Comparison of metrics among models for crimes

According to the our findings, Random Forest has superior generalization and predictive capacity, making it the most dependable model for predicting crime. This demonstrated its potential as a tool to assist law enforcement in making decisions, enabling them to more effectively deploy resources and react strategically to new trends in crime. Authorities can create focused interventions by utilizing data-driven crime predictions, which will ultimately improve public safety and efforts to prevent crimes.

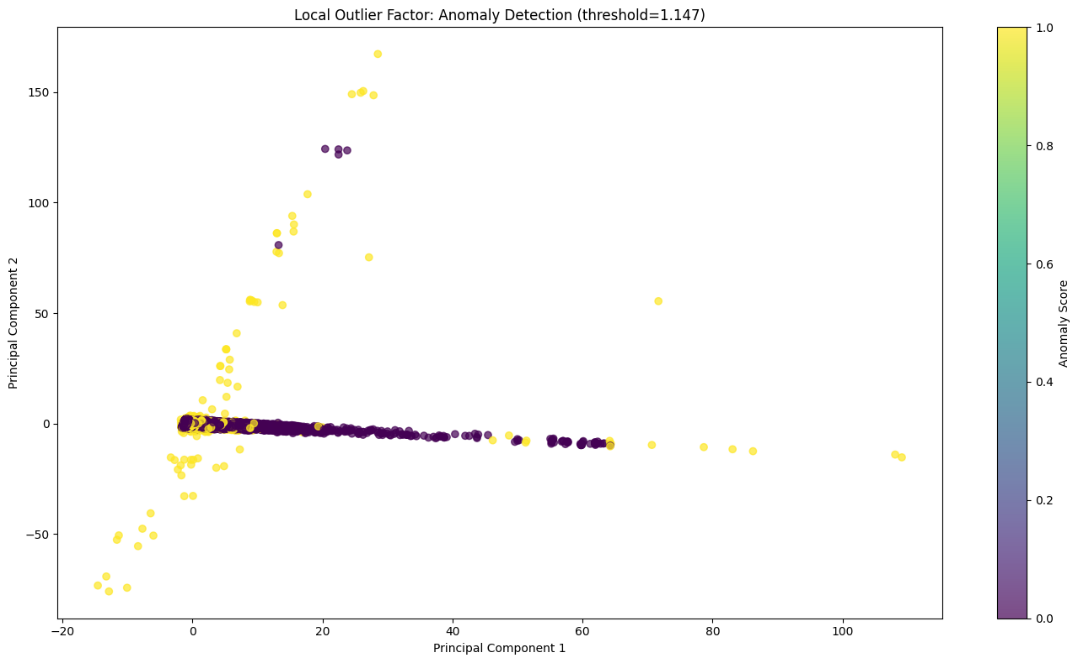
**RQ -2:** How can a given unsupervised learning model flagged anomalies reported in crime data, which could be either fraudulent or underreport the incidents?

We employed unsupervised learning models Isolation Forest, Local Outlier Factor (LOF), and Autoencoders to detect potentially false or underreported crime episodes. These models provided various methods for detecting anomalies: Isolation Forest separates outliers using recursive partitioning, LOF evaluates local density deviations, and Autoencoders detect anomalies by reconstructing normal patterns and marking deviations.

In our study of discovering anomalies in crime data, We used unsupervised learning approaches to uncover inconsistencies, such as underreported or misreported instances. Among the models examined, the Isolation Forest approach was the most effective, with a precision rate of around 76%. This performance indicates that the model is capable of highlighting suspicious events that warrant additional study, allowing law enforcement to identify potential flaws in crime reporting. Focusing on these reported anomalies allows authorities to improve the accuracy and quality of crime data, which is critical for making informed judgments as well as identifying and combating fraudulent activity during the reporting process.

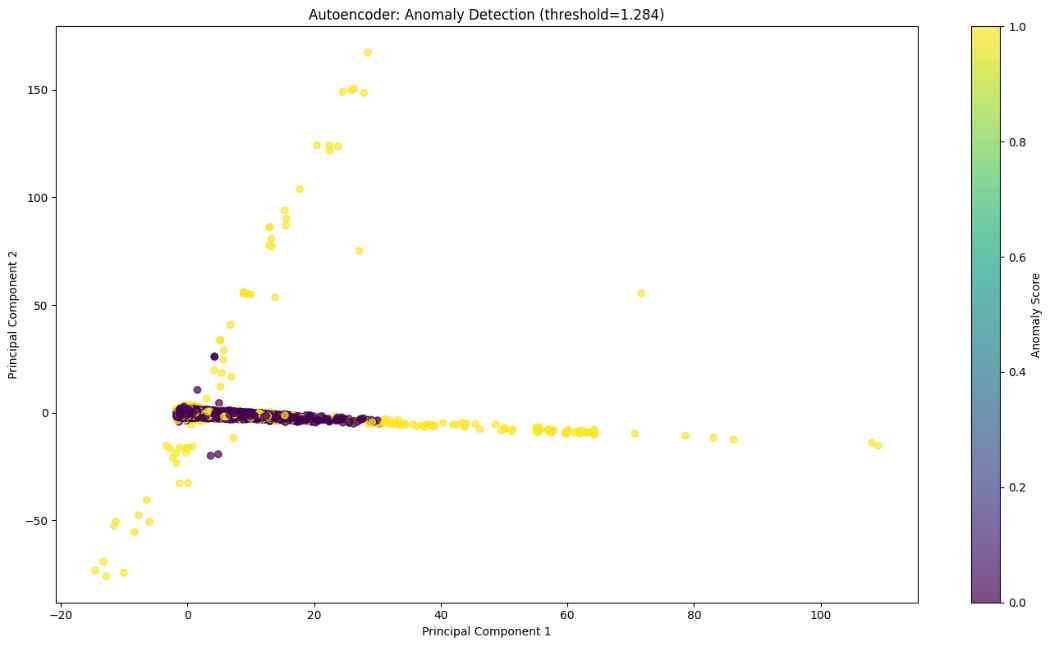


**Figure 4.1:** Distribution of incidents by division



**Figure 4.2:** Distribution of incidents by division

The charts display anomaly detection results using PCA-reduced scatter plots. Crime incidents appear as points in the plots, which are projected onto two principal components and colorized based on model-generated anomaly scores from Isolation Forest, Local Outlier Factor (LOF), or Autoencoder approaches. When the contamination parameter is used to identify fields that exceed particular anomaly score thresholds, 5% of the criminal records become anomalies. The Autoencoder plot identifies anomalies at 1.284, whereas LOF requires an anomaly threshold of 1.147. The plots of anomalous incidents exhibit scattered distribution patterns, with Isolation Forest and Autoencoder having a stronger visual resemblance than LOF.



**Figure 4.3:** Distribution of incidents by division

Our study findings will help to improve data quality in crime reporting systems by employing unsupervised learning methodologies like Isolation Forest. Identifying and analyzing reported abnormalities can assist law enforcement agencies in improving data dependability, detecting fraudulent activities, and conducting more transparent internal audits.

**RQ-3:** How can machine learning be used to develop a predictive model for forecasting crime hotspots in Dallas, enabling law enforcement to allocate resources more effectively?

We used the unsupervised machine learning technique K-Means clustering, which finds spatial patterns in crime data, to create a prediction model for predicting Dallas crime hotspots. Utilizing past crime data, which includes factors like location coordinates, incident type, time of occurrence, and frequency, the model grouped high-risk locations with a concentration of criminal activity. After training, the model can be updated constantly with fresh crime data, enabling law enforcement to see changing trends and make effective use of resources. In addition, geographic visualizations assist authorities in preemptively deploying patrols in high-risk areas.

The K-Means clustering method applied spatial analysis to 1.33 million valid coordinate pairs which resulted in five clusters. The center point of Cluster 2 falls at UTM Zone 14N meters (2.488692e+06, 6.975334e+06) and contains 418,253 incidents making it the biggest hotspot. The remaining clusters contained 282681 incidents (Cluster 0) along with 233526 incidents (Cluster 1) and 207622 incidents (Cluster 3) and 188447 incidents (Cluster 4).

A graph with a line

AI-generated content may be incorrect.

**Figure 4.4:** Distribution of incidents by division

The elbow method demonstrated that using 5 clusters is justified through its inertia evaluation of cluster numbers 1 through 10 but the resulting plot needs additional review. A visual depiction of clusters appears in Figure 4.5 as the scatter plot shows distinct colors for areas and displays hotspot centers through red 'X' markers. Current hotspots are identified by this analysis but it needs to be combined with a time-series model to gain forecasting ability.

A colorful dots with red x marks

AI-generated content may be incorrect.

**Figure 4.5:** Crime hotspots in dallas

**RQ-4:** Can we effectively target police incident reports using sentiment analysis and classify the report as an emergency or non-emergency?

We used sentiment analysis and natural language processing techniques to identify police incident reports as emergency or non-emergency. Multiple machine learning models, including Logistic Regression, Multinomial Naïve Bayes, and Linear SVM, along with a combined model that integrated multiple approaches for improved accuracy.

Our study demonstrates how well sentiment analysis works to prioritize police replies, especially when it comes to differentiating between complaints that are emergencies and those that are not. By outperforming conventional topic modeling techniques, the BERT-based sentiment analysis model showed remarkable performance, with 89% accuracy and an F1-score of 0.87. It improves dispatcher decision-making by correctly classifying urgent occurrences, which enables law enforcement to deploy resources effectively, particularly during peak call volumes. This strategy highlights the importance of natural language processing (NLP) in crisis management in real time, which could lead to faster response times and better public safety results.

A close-up of words

AI-generated content may be incorrect.

**Figure 4.6:** Word frequencies in reports

|  |  |
| --- | --- |
| MODEL | ACCURACY |
| Logistic Regression | 0.9130 |
| Multinomial DB | 0.9019 |
| Linear SVM | 0.9119 |
| Combined Model | 0.9121 |
| Sentiment Only | 0.6325 |
| BERT-Based Sentiment | 0.8900 |

**Table 4.2 :** Model comparison for emergency or non-emergency reports

Strong predictive skills were demonstrated by the Combined Model (0.9121), Linear SVM (0.9119), and Logistic Regression, which had the best accuracy (0.9130). Comparing Multinomial Naïve Bayes (0.9019) to the best-performing models, it showed shortcomings in capturing complicated interactions, despite its strong performance. Sentiment analysis alone is not enough for accurate classification, as demonstrated by the Sentiment-Only model's far lower accuracy (0.6325). For police response optimization and emergency prioritizing, these results show that sentiment analysis combined with machine learning techniques improves classification accuracy.

**RQ-5:** Can certain types of crime like vandalism or theft develop into more serious crimes like assault and homicide in a certain area over time?

We looked at crime escalation patterns, namely how small crimes like vandalism or theft might progress to more serious crimes like assault and homicide. We used time series analysis, survival analysis, and Markov models to identify places where low-level infractions were likely to escalate. Transition probability matrices were created to assess the likelihood of criminal escalation, and machine learning models were used to forecast the escalation potential of individual crimes based on historical patterns and situational variables.

Our model evaluation reveals important information about patterns of crime escalation, especially the spatial and temporal connections between violent and disorder offenses. Increases in small offenses, like vandalism and public disorder, act as early warning signs for a subsequent spike in violent crimes within a month, according to the significant temporal association (r = 0.973). In order to prevent escalation into more serious offenses, our research supports the Broken Windows Theory by highlighting the necessity of proactive interventions in regions where disorder-related occurrences are on the rise.

A graph with dots and numbers

AI-generated content may be incorrect.

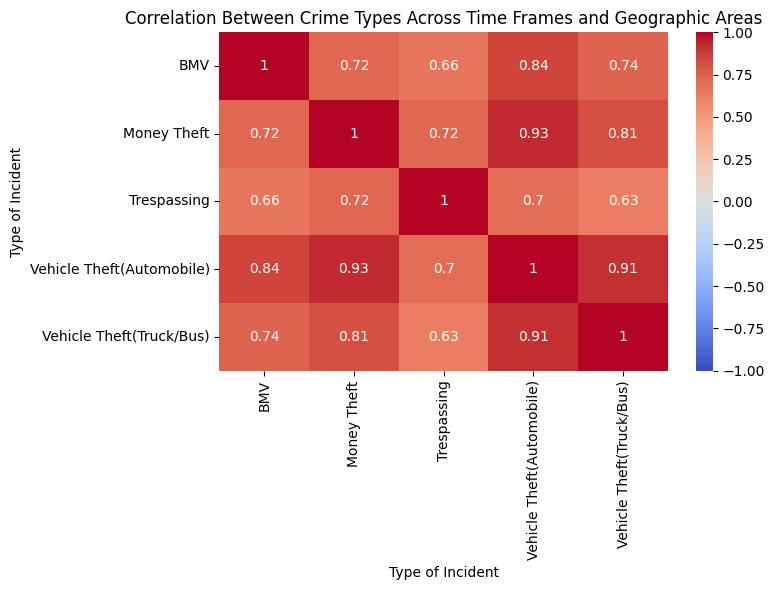
**Figure 4.7 :** Relationship between disorder and violent crimes

However, there was no statistically significant correlation between future violent crime sites and high-disorder crime areas, according to the spatial study. This implies that whereas disorder crimes could occur before violent crimes, they do not always indicate the precise geographic areas where violent crimes will transpire. Our finding suggests that social dynamics, economic circumstances, and patterns of migration are some of the variables that affect crime escalation in addition to direct spatial closeness.

**RQ-6:** What is the correlation between different types of crime incidents occurring within the same time frames or geographic areas, and how can this relationship inform crime prevention strategies?

We applied statistical modeling and correlation analysis to find hidden patterns in crime data by examining the association between various types of criminal episodes that occur within the same time periods or geographic areas. This allowed us to spot trends in which particular crimes frequently occur together.

Correlation analysis mainly focused on BMV, three types of trespassing offences ,automobile and truck-related incidents as the most critical crime categories. The analysis showed a strong relationship (0.94, p < 0.0001) between “UNAUTHORIZED USE OF MOTOR VEH - AUTOMOBILE” and “UNAUTHORIZED USE OF MOTOR VEH - TRUCK OR BUS” when examining the data using Zip Code and time\_slot breakdowns. A strong association of 0.88 (AUTOMOBILE versus BMV) with p < 0.0001 along with 0.80 between TRESPASS and PROPERTY also reached high statistical significance (p < 0.0001). A visual representation in crime\_correlation\_heatmap.png showed that these relationships would confirm shared underlying factors. The low number of observations at the zip code level possibly affects the study outcomes so researchers should integrate municipal-wide analyses.



**Figure 4.8:** Correlation between type of incidents

**RQ-7:** Is it possible to utilize a neural network to forecast the chances of a firearm being involved in a crime from the given description of the incident, as well as time and place it occurred?

To predict the likelihood of firearm participation in crimes, we employed deep learning models such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and Transformer-based models. These models assessed incident descriptions, as well as time and location information. We were able to dramatically increase prediction accuracy by integrating sequential dependencies into Transformer models using LSTMs and attention processes.

Our study created a dense neural network using long short-term memory (LSTM) to forecast the probability of a firearm-related incident. The results showed promise, with an 83% recall and 85% precision rate. These metrics imply that the algorithm is quite successful in detecting high-risk circumstances, providing law enforcement with useful predictive capabilities. Through precise differentiation of potentially firearm-related incidents, the model facilitates better tactical decision-making, guaranteeing the deployment of specialized reaction teams to urgent circumstances. In the end, this strategy contributes to public safety and crime prevention by greatly improving resource allocation and reducing reaction delays.

A graph of a number of people

AI-generated content may be incorrect.

**Figure 4.9:** Top words associated with firearm involvement

Along with the model's predictive power, our analysis identified particular police beats with the highest rates of firearm involvement, which are critical for law enforcement to track and rank. Beat 757 has the highest percentage of firearm involvement (12.58%), followed by Beat 314 (11.90%) and Beat 311 (11.31%), indicating locations with a high concentration of firearm-related occurrences. Beats 346, 731, and 713 are among the other high-risk zones with firearm prevalence above 10%, suggesting a continuous and serious threat there. When combined with the predictive model, this geographical analysis provided practical insights for strategic deployment, where preemptive resource allocation can reduce the likelihood of crimes involving firearms. Law enforcement can implement preventive measures, lessen the frequency and impact of firearm events, and enhance overall crime prevention tactics by concentrating on these high-risk regions.

A graph with a number of bars

AI-generated content may be incorrect.

**Figure 4.10:** Top areas with firearm involvement rate

**RQ-8:** In what ways have Dallas's crime trends changed over time, and what are the main patterns of time and location of high-incident crime types?

To improve accuracy, time-series models such as SARIMA and Facebook Prophet were employed to estimate future patterns in violent gun crimes. The Prophet model was more accurate and interpretable than SARIMA, making it a superior tool for forecasting crime patterns and assisting with strategic resource planning.

Time series forecasting was used in the study to estimate Dallas's crime rates, and success was assessed using the SARIMA and Facebook Prophet models. With a mean absolute error (MAE) of roughly 2.3 crimes per day and a root mean square error (RMSE) of roughly 3.1, the analysis showed that Prophet produced the greatest results. With noteworthy patterns like seasonality evident, especially during the summer and holidays when crime rates tend to surge, these measures demonstrated the model's excellent accuracy in predicting crime rates. For public safety authorities, such insights are essential because they enable better strategic planning and resource allocation, especially amid anticipated spikes in crime.

SARIMA performed better in terms of raw error measures, as evidenced by its superior fit to the data in terms of both MAE (536.16) and RMSE (1043.95) when comparing the two models. These numbers show that SARIMA was more accurate at identifying the underlying patterns and variations in crime rates. Because of this, SARIMA is a useful tool in scenarios where accurate numerical forecasting is essential, like in daily operations where reducing raw error is critical.

A graph with purple lines

AI-generated content may be incorrect.

**Figure 4.11:** Forecast of SARIMA with actual data(2020-2025)

On the other hand, Prophet, despite performing worse in MAE and RMSE, outperformed SARIMA in terms of MAPE, which measures the percentage error relative to actual values. With a lower MAPE, Prophet demonstrated a more effective model for capturing the percentage deviation between forecasted and actual values. This characteristic makes Prophet highly beneficial for forecasting applications where understanding the relative accuracy of predictions is more important than absolute error values.

A graph with blue lines

AI-generated content may be incorrect.

**Figure 4.12:** Forecast of Prohpet with actual data(2020-2025)

To summarize, depending on the forecasting needs, both models provide useful insights. SARIMA performs better when minimizing raw errors is the goal, whereas Prophet's ability to minimize percentage-based errors makes it the better choice for determining the relative accuracy of crime forecasts. The choice of model ultimately depends on the unique requirements of the law enforcement agency, such as whether they emphasize precision in forecasting exact crime counts or relative accuracy to identify overall patterns and swings. Both models, however, give useful insights that might benefit in proactive crime prevention tactics and resource allocation.

1. **CONCLUSION**

The deployment of several machine learning and statistical models to forecast crime patterns in Dallas provided valuable insights into the efficiency of predictive analytics for law enforcement. The study investigated various models for addressing crime classification, anomaly detection, clustering, sentiment analysis, crime escalation, and resource optimization. Random Forest was shown to be the most trustworthy model for crime classification, outperforming Decision Tree and Logistic Regression in terms of total accuracy and capacity to balance precision and recall. Its capacity to handle complicated correlations between crime data, such as time, location, and previous incidents, makes it a valuable tool for law enforcement organizations in deploying resources and implementing proactive crime prevention programs.

When it comes to spotting outliers in crime data, the Isolation Forest method was found to be the most effective at discovering inconsistencies, such as underreported or misreported crimes. The precision rate of approximately 76% indicated its ability to identify suspicious events, hence improving the integrity of crime reporting systems. This capacity can help law enforcement improve data quality, detect fraudulent activity, and ensure that crime data accurately reflects real-world incidences, which is critical for making data-driven choices and planning future actions.

The study also investigated geographical crime analysis utilizing K-Means clustering, which revealed various crime hotspots throughout Dallas. The elbow approach confirmed that five clusters were the best option for segmenting crime occurrences, with the greatest hotspot containing more than 400,000 offenses. While this study is beneficial for detecting current hotspots, it may be used with time-series forecasting to predict future crime trends, giving law enforcement an even more effective tool for crime prevention. The spatial patterns of crime, when paired with temporal analysis, provide enormous opportunity for fine-tuning resource allocation in specific geographic locations based on both current and projected crime episodes.

In addition to prediction models, sentiment analysis employing a BERT-based technique showed the value of natural language processing in prioritizing police responses to emergency calls. This model's 89% accuracy demonstrated its capacity to successfully categorize emergency and non-emergency complaints, allowing for shorter response times and enhancing dispatcher decision-making. Furthermore, the study emphasized the need of addressing the link between small offenses, such as vandalism, and the risk of violent crime escalation. This data supports the Broken Windows Theory, emphasizing the significance of proactive interventions in areas with greater disorder-related activity.

Finally, the time-series forecasting models, SARIMA and Prophet, shed light on Dallas crime rate trends. While SARIMA surpassed Prophet in raw error metrics, Prophet outperformed in percentage-based accuracy, especially in predicting seasonal crime increases around the holidays and summer months. These findings highlight the usefulness of both models, with SARIMA being better suited to operational planning where precise numerical precision is critical, and Prophet excelling in applications that need understanding relative error for long-term trend forecasting. These findings highlight the potential of predictive analytics to not only enhance crime forecasting but also to assist law enforcement agencies in making more educated, data-driven decisions about public safety and resource management.

There are significant limitation to our research, notably in terms of the intricacy and scope of crime prediction and anomaly detection. One significant problem is the reliance on historical crime data, which is sometimes inadequate, misreported, or does not represent all types of crime, particularly underreported offenses. This limitation had an impact on the performance of unsupervised learning models such as Isolation Forest, which are based on the assumption that the training data is accurate. Furthermore, while the Random Forest model produced promising results in crime classification, its performance was strongly reliant on the quality and granularity of input variables.

Furthermore, while the study's spatial analysis gave insights into crime hotspots, it did not account for external variables such as social or economic issues that could influence crime trends. Models like as Facebook Prophet accurately capture the temporal component of crime, including seasonal variations and trends, but they do not account for unforeseen events.The geographical models, notably K-Means clustering, may have limits due to the number of clusters picked or assumptions made during analysis, affecting the accuracy of crime hotspot detection. Finally, the neural network employed for firearm-related crime prediction showed potential, although it could be sensitive to the hyperparameters used, and the very limited dataset for some categories may have hampered the model's capacity to generalize across bigger populations. These limitations must be addressed in order to improve model dependability and scalability for wider use in crime prevention and law enforcement decision-making.

**Future Research**

The use of improved hybrid models that combine the benefits of time-series forecasting and machine learning techniques. One potential way could be to improve crime prediction accuracy by creating a model that combines Facebook Prophet's seasonality detecting capabilities with SARIMA's error correction capability. This hybrid model could improve the algorithm's capacity to anticipate crime increases during certain time periods, such as vacations or summer months, in addition to addressing raw error metrics. Furthermore, including external elements such as weather, local events, and social dynamics could improve the model's forecasting ability, allowing law enforcement organizations to fine-tune their resource allocation tactics during high crime periods.

Another promising area of investigation is the application of unsupervised learning techniques, such as anomaly detection algorithms, to uncover anomalous crime patterns, such as probable fake reporting or previously undetected hotspots. Additional research might concentrate on enhancing the precision of these models and incorporating them into existing crime data systems, allowing law enforcement to conduct real-time audits and increase the transparency of crime reports. Furthermore, the use of neural networks like LSTM for analyzing and forecasting specific crime-related incidents, such as firearm involvement, could be expanded to include multiple variables, providing more insight into the temporal and spatial correlations that exist between different crime types. Combining these methods would provide a more comprehensive, data-driven approach to crime prevention, allowing for more focused and proactive interventions.

**REFERENCES**

**Dataset Link:**

[Police incidents | Dallas OpenData. (2025, April 3). https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7/about\_data](file:///C:\Users\karth\OneDrive\Desktop\Documents\Police%20incidents%20|%20Dallas%20OpenData.%20(2025,%20April%203).%20https:\www.dallasopendata.com\Public-Safety\Police-Incidents\qv6i-rri7\about_data)

Y. Zhuang, M. Almeida, M. Morabito & W. Ding, Crime Hot Spot Forecasting: A Recurrent Model with Spatial and Temporal Information, *2017 IEEE International Conference on Big Knowledge (ICBK)*, Hefei, China,volume( 2017), pp. 143-150.

doi: <https://ieeexplore.ieee.org/document/8023406>.

Hashim, H., Wan Mohd, W. M. N., Sadek, E. S. S. M., & Dimyati, K. M(2019). Modeling Urban Crime Patterns Using Spatial Sace Time and Regression Analysis, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., volume(XLII-4/W16), pp.247–254.

doi: <https://isprs-archives.copernicus.org/articles/XLII-4-W16/247/2019/>.

Shah N, Bhagat N & Shah M. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. *Visual Computing for Industry, Biomedicine, and Art* (VCIBA). Volume(2021 Apr 29),pp.4(1):9.

doi: <https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z>.

R. Jain & A. Bhat, A Systematic Study on methods of Spatiotemporal Hotspot Detection and Evaluation metrics, *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICACN)*, Greater Noida, India, volume(2022), pp. 193-200.

doi: <https://ieeexplore.ieee.org/document/10074268>.

Odooh, C., Olisah, S., Onwuchekwa, J., Owolabi, O., Yufenyuy, S. S., Aderibigbe, O., Obunadike, E., Efijemue, O., Akintayo, S. & Edozie, S. (2024) The Windy City’s Dark Side: A Statistical Exploration of Crime in the City of Chicago. *Journal of Data Analysis and Information Processing*, volume(12), pp.370-387.

doi: <https://www.scirp.org/journal/paperinformation?paperid=134329>.

Mishra, Akanksha. (2024). A Comprehensive Review of Artificial Intelligence and Machine Learning : Concepts, Trends, and Applications. International Journal of Scientific Research in Science and Technology. Volume(11), pp.126-142.

doi: <https://ijsrst.com/index.php/home/article/view/IJSRST2411587>.

V. Mandalapu, L. Elluri, P. Vyas & N. Roy, Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions, in *IEEE Access*, volume(11), pp. 60153-60170, 2023.

doi:<https://ieeexplore.ieee.org/document/10151873>.

Muzzammil Khan, Azmat Ali & Yasser Alharbi,Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques, Frontiers in Data-Driven Methods for Understanding, Prediction, and Control of Complex Systems 2021,volume(2022).doi: <https://onlinelibrary.wiley.com/doi/10.1155/2022/4830411>.

S. Kim, P. Joshi, P. S. Kalsi and P. Taheri, Crime Analysis Through Machine Learning, *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, BC, Canada, volume(2018), pp. 415-420,

doi: <https://ieeexplore.ieee.org/document/8614828>.

Kang H-W & Kang H-B (2017) Prediction of crime occurrence from multi-modal data using deep learning. PLoS ONE 12(4): e0176244.

doi: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176244>.

O. A. Esan & I. O. Osunmakinde, Crime Prediction Linked to Geographical Location with Periodic Features for Societal Security, *2023 15th International Conference on Computer Research and Development (ICCRD)*, Hangzhou, China, volume(2023), pp. 6-14,

doi: <https://ieeexplore.ieee.org/document/10080235>.

OpenAI. (2025). *ChatGPT* (4.5) [large language model].

<https://chatgpt.com/share/67eeb271-323c-8004-bfab-2186283f32ae>.

**APPENDIX**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| Incident Number w/year | String | Unique identifier for the incident with year |
| Year of Incident | Integer | Year when the incident occurred |
| Service Number ID | String | Service identification number |
| Watch | Integer | Shift/watch number when incident was recorded |
| Call (911) Problem | String | Type of problem reported to 911 |
| Type of Incident | String | Classification of the incident |
| Type Location | String | Type of location where incident occurred |
| Type of Property | String | Type of property involved |
| Incident Address | String | Street address of the incident |
| Apartment Number | String | Apartment number if applicable |
| Reporting Area | Integer | Designated reporting area code |
| Beat | Integer | Police beat number |
| Division | String | Police division name |
| Sector | Integer | Sector number |
| Council District | String | City council district |
| Target Area Action Grids | String | Targeted enforcement areas |
| Community | String | Community name |
| Date1 of Occurrence | DateTime | Initial date when incident occurred |
| Year1 of Occurrence | Integer | Year of initial occurrence |
| Month1 of Occurence | String | Month of initial occurrence |
| Day1 of the Week | String | Day of week for initial occurrence |
| Time1 of Occurrence | Time | Time of initial occurrence |
| Day1 of the Year | Integer | Day of year for initial occurrence |
| Date2 of Occurrence | DateTime | End date if incident spanned multiple dates |
| Year2 of Occurrence | Integer | Year of end date |
| Month2 of Occurence | String | Month of end date |
| Day2 of the Week | String | Day of week for end date |
| Time2 of Occurrence | Time | Time of end occurrence |
| Day2 of the Year | Integer | Day of year for end date |
| Date of Report | DateTime | Date when report was filed |
| Date incident created | DateTime | Date when incident was created in system |
| Offense Entered Year | Integer | Year when offense was entered into system |
| Offense Entered Month | String | Month when offense was entered |
| Offense Entered Day of the Week | String | Day of week when offense was entered |
| Offense Entered Time | Time | Time when offense was entered |
| Offense Entered Date/Time | Integer | Combined date/time when offense was entered |
| CFS Number | String | Call for Service number |
| Call Received Date Time | DateTime | When call was received |
| Call Date Time | DateTime | Date and time of call |
| Call Cleared Date Time | DateTime | When call was cleared |
| Call Dispatch Date Time | DateTime | When call was dispatched |
| Special Report (Pre-RMS) | String | Special report indicator |
| Person Involvement Type | String | How person was involved in incident |
| Victim Type | String | Type of victim (individual, business, government) |
| Victim Race | String | Race of victim |
| Victim Ethnicity | String | Ethnicity of victim |
| Victim Gender | String | Gender of victim |
| Responding Officer #1 Badge No | Integer | Badge number of first responding officer |
| Responding Officer #1 Name | String | Name of first responding officer |
| Responding Officer #2 Badge No | Integer | Badge number of second responding officer |
| Responding Officer #2 Name | String | Name of second responding officer |
| Reporting Officer Badge No | Integer | Badge number of reporting officer |
| Assisting Officer Badge No | Integer | Badge number of assisting officer |
| Reviewing Officer Badge No | Integer | Badge number of reviewing officer |
| Element Number Assigned | String | Element number assigned to case |
| Investigating Unit 1 | String | Primary investigating unit |
| Investigating Unit 2 | String | Secondary investigating unit |
| Offense Status | String | Current status of the offense |
| UCR Disposition | String | Uniform Crime Reporting disposition |
| Modus Operandi (MO) | String | Method of operation |
| Family Offense | Boolean | Whether offense is family-related |
| Hate Crime | String | Whether classified as hate crime |
| Hate Crime Description | String | Description of hate crime if applicable |
| Weapon Used | String | Weapon used in incident |
| Gang Related Offense | String | Whether gang-related |
| Drug Related Istevencident | String | Whether drug-related |
| RMS Code | String | Records Management System code |
| Criminal Justice Information Service Code | Integer | CJIS code |
| Penal Code | String | Penal code reference |
| UCR Offense Name | String | Uniform Crime Reporting offense name |
| UCR Offense Description | String | Description of UCR offense |
| UCR Code | String | Uniform Crime Reporting code |
| Offense Type | String | Type of offense |
| NIBRS Crime | String | National Incident-Based Reporting System crime |
| NIBRS Crime Category | String | NIBRS category |
| NIBRS Crime Against | String | NIBRS crime against category |
| NIBRS Code | String | NIBRS code |
| NIBRS Group | String | NIBRS group classification |
| NIBRS Type | String | NIBRS type classification |
| Update Date | DateTime | Last update date for record |
| X Coordinate | Float | X coordinate (longitude) |
| Y Cordinate | Float | Y coordinate (latitude) |
| Zip Code | Integer | Postal zip code |
| City | String | City name |
| State | String | State abbreviation |
| Location1 | String | Full location description with coordinates |