Course Project#1

Title: Predicting Crime Rates in Chicago

Business Problem:

Crime in Chicago is an ongoing concern, deeply impacting communities and governance. This project seeks to utilize the extensive, high-quality datasets available to understand and predict crime patterns. By building an intelligent prediction model, we aim to forecast crime types, locations, and timings. Such insights will empower policymakers, law enforcement, and residents to make informed decisions, fostering a safer urban environment. As someone who has lived in Chicago for over two years, I have a personal connection to this issue, making it both professionally and personally meaningful.

Understanding crime patterns has implications not just for law enforcement but also for community safety, urban planning, and resource optimization. With the right tools and insights, it becomes possible to reduce response times, allocate resources effectively, and take preemptive measures against potential criminal activities. Crime prediction and prevention are areas where technology and data science can make a significant social impact.**Problem Problem Problem Statement:**

The following questions guide the exploration:

- How has crime in Chicago evolved across different years? Was 2016 the bloodiest year in two decades?
- Are specific types of crimes more likely to occur at particular times, locations, or days of the week?

- What patterns can be discerned from the available crime data, and how can these patterns inform proactive measures?

Background/History :

Chicago's history with crime is complex and multifaceted. Over decades, the city has grappled with issues ranging from organized crime to community violence. These challenges have been influenced by socioeconomic disparities, urban infrastructure, and shifting demographic trends. Historical data highlights fluctuations in crime rates, which often correlate with economic downturns or major social changes.

The Chicago Police Department has maintained detailed records, which serve as a rich resource for data scientists. However, past initiatives like predictive policing programs have faced backlash due to systemic biases. These controversies underscore the importance of ethical, data-driven methods that prioritize transparency and fairness while addressing crime. Despite these challenges, the availability of open data provides an opportunity to explore crime dynamics in a rigorous and ethical manner.

Dataset Explanation:

The project leverages the Chicago Crime dataset, spanning from 2005 onwards. Obtained from the Chicago Police Department's CLEAR system, this dataset comprises 23 attributes that include:

- **Case Details**: Unique IDs and case numbers to identify records.
- **Temporal Information**: Dates and times of crimes.

- **Geospatial Data**: Latitude/longitude, coordinates, and administrative divisions (e.g., districts, wards).
- **Crime Attributes**: Illinois Uniform Crime Reporting (IUCR) codes categorizing crimes.

The dataset provides granular details that enable a comprehensive spatial, temporal, and categorical analysis of crimes. For example, latitude and longitude data allows for mapping crime hotspots, while temporal attributes such as day of the week and time provide insights into when crimes are most likely to occur. Additionally, the IUCR codes offer standardized categorization, enabling nuanced analysis of crime types.

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Community Area	FBI Code	X Coordinate	Co
0	10000092	HY189866	03/18/2015 07:44:00 PM	047XX W OHIO ST	041A	BATTERY	AGGRAVATED: HANDGUN	STREET	False	False	...	28.0	25.0	04B	1144606.0	19
1	10000094	HY190059	03/18/2015 11:00:00 PM	066XX S MARSHFIELD AVE	4625	OTHER OFFENSE	PAROLE VIOLATION	STREET	True	False	...	15.0	67.0	26	1166468.0	18

Data Preparation

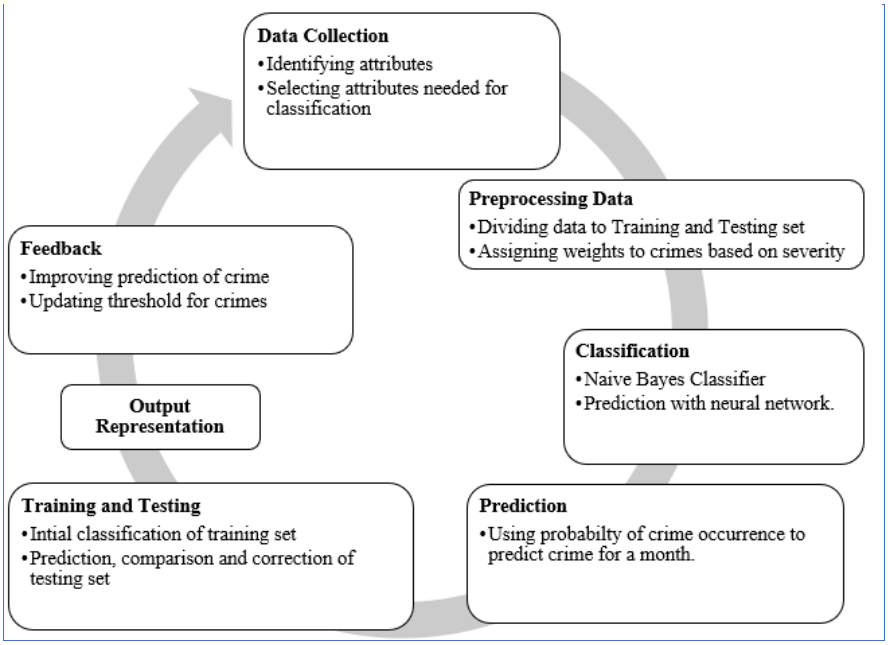
To ensure the data is suitable for analysis, the following steps were undertaken:

- **Data Cleaning**: Missing values and outliers were addressed to ensure accuracy. For example, crimes with incomplete location data were either interpolated based on nearby data points or excluded if unrecoverable.
- **Transformation**: Time data was categorized into periods (e.g., morning, afternoon) and days of the week to facilitate trend analysis.

- **Feature Engineering**: Derived metrics, such as crime density and frequency by area, were calculated. This included creating heatmaps of crime incidence to visualize spatial patterns.
- **Splitting**: Data was divided into training and testing subsets for model evaluation. A stratified sampling approach was used to ensure proportional representation of different crime types.

Methodology:

This project approaches the analysis of dataset based on dividing it into training and testing sets. The training set is used for classification based on both the Naïve Bayes classifier and the Neural Network. The performance of both the algorithms is analyzed and the best algorithm is used for further classification of the testing set. The classification of the dataset is done by dividing the time into various months of the year and the crime patterns are analyzed for each month of the year.



Analysis:

The analysis combined multiple methods to yield actionable insights:

- **Clustering**: Highlighted regions and periods with high crime intensity. For instance, certain neighborhoods consistently showed elevated crime rates during nighttime hours.
- **Heatmaps:** Visualized day vs. night crime distribution, revealing a higher incidence of violent crimes at night and property crimes during the day.
- **Temporal Trends**: Examined long-term patterns and seasonal spikes. Summer months exhibited higher overall crime rates, potentially due to increased outdoor activity.
- **Prediction:** Validated the accuracy of predictions for crime type and location through a feedback loop. This iterative process improved model reliability by fine-tuning hyperparameters.

These findings were visualized using maps, charts, and dashboards, providing a clear depiction of crime hotspots and trends. Interactive dashboards were developed to allow stakeholders to explore data dynamically.

The prediction algorithm finds the probability of occurrence of a particular crime in the month of the year for a given location. This prediction can be verified within the test set to measure accuracy based on classification of actual crime occurrences. The feedback from this process is used to update the thresholds of crime prediction.

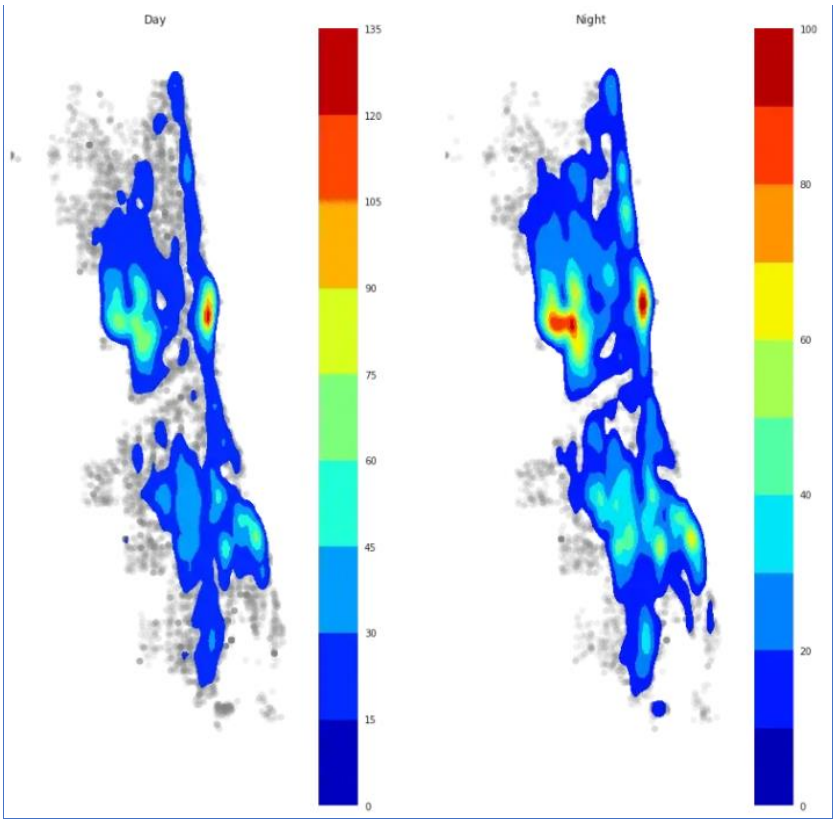
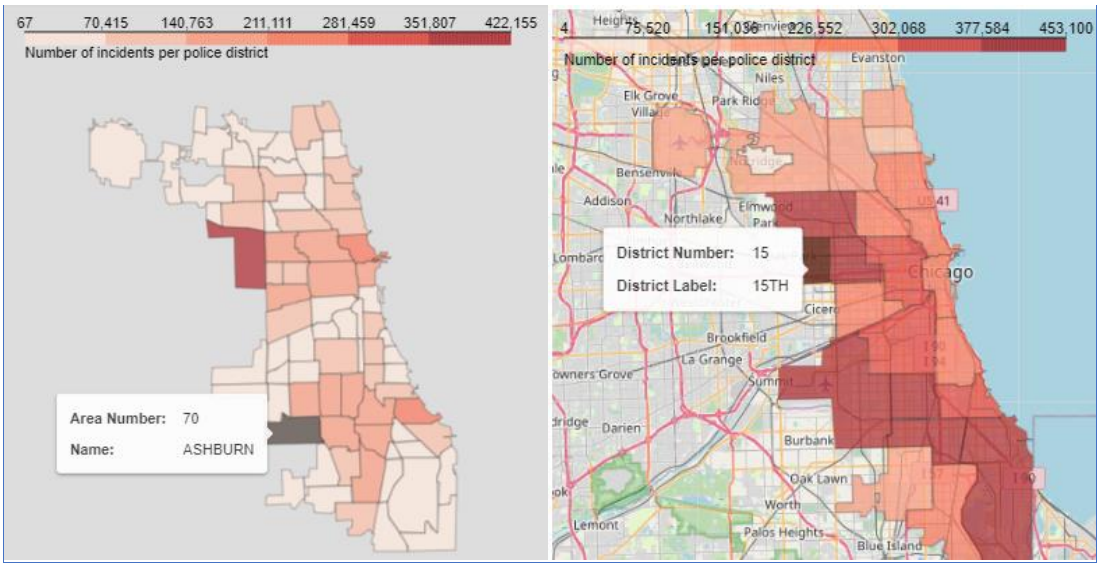


Figure 1 Heatmaps of crimes happening in day(left) vs crimes happening in night(right)

Finally, the output of the prediction of crime hotspots is depicted graphically in the map of City of Chicago. This prediction can be depicted for the past years to show the algorithm's prediction efficiency.



K-means clustering provides a way to group data points together in a way that minimizes differences between the data points in the same group. By applying these methods, we can take n data points and partition them into k clusters.

Ethical Problems:

The use of predictive policing systems raises important ethical considerations:

- **Bias in Historical Data**: Past records may reflect systemic inequities, influencing model outputs. Addressing this requires robust bias mitigation strategies.
- **Transparency**: Algorithms must be interpretable and subject to public scrutiny to ensure accountability.
- **Fairness**: Predictions should not reinforce existing societal biases. Regular audits and stakeholder consultations are essential to uphold fairness.

To address these concerns, the project emphasizes fairness and accountability through rigorous data auditing, ethical oversight, and the inclusion of diverse perspectives in model development.

Challenges/Issues

Crime has become the focus of a challenge for both the police department and law enforcement agencies to reduce the spread of crime and use modern techniques to predict and reduce the spread of crime. Many police departments use artificial intelligence algorithms and use big data tools to help them to predict where crimes are occurring. For example, police departments in places like Seattle, Los Angeles, and Atlanta have experimented with predictive police programs which try to identify the geographic area where crime is likely to occur in the future. At the same time, the Chicago Police Station used an algorithm based on a heat list that tries to identify people as criminals for violent crimes or repeated abuse.

Assumptions

- Crime data accurately reflects real-world occurrences without significant reporting biases.
- Temporal patterns are stable over time, allowing predictive models to generalize across years.
- Socioeconomic and external factors influencing crime remain constant during the study period.

Limitations

- **Data Quality**: Potential inaccuracies due to underreporting or misclassification. For instance, minor crimes may be underrepresented in the dataset.

- **Bias**: Historical biases in data could influence model predictions, particularly if certain demographics are disproportionately represented in arrest records.
- **Scope**: Excludes real-time events or external factors such as policy changes or economic shifts.

Results/Prediction:

Some of the analysis done through the project are illustrated below:

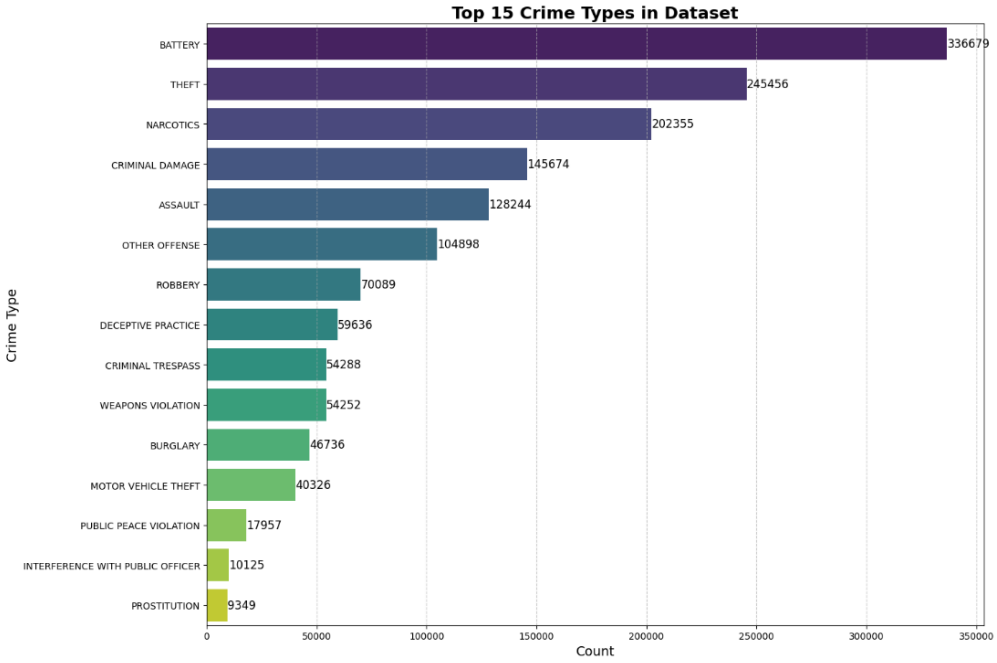


Figure 2 Top 15 Crimes in Chicago

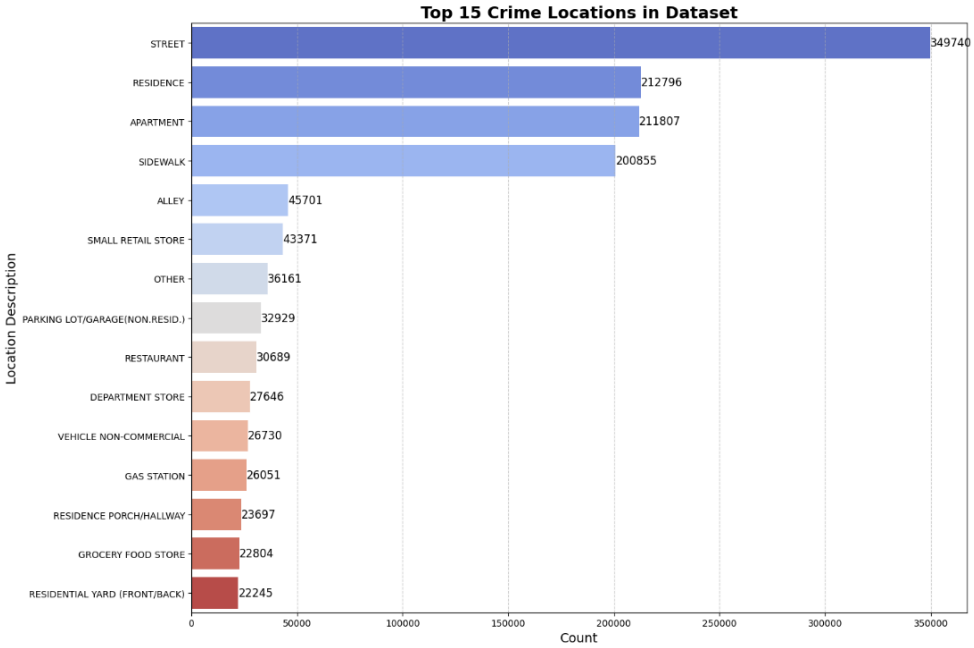


Figure 3 Top 15 Crime Location

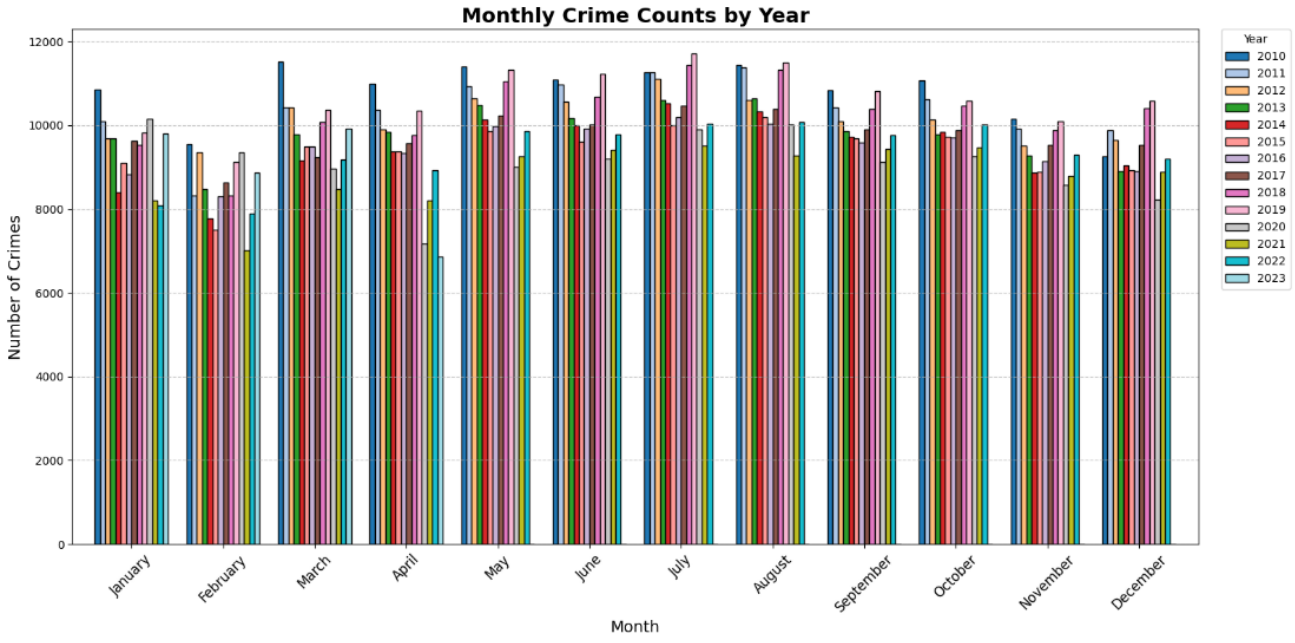


Figure 4 Monthly Crime Count Averaged

As per the analysis summer is highlighted as the time period where a large number of crime take place.

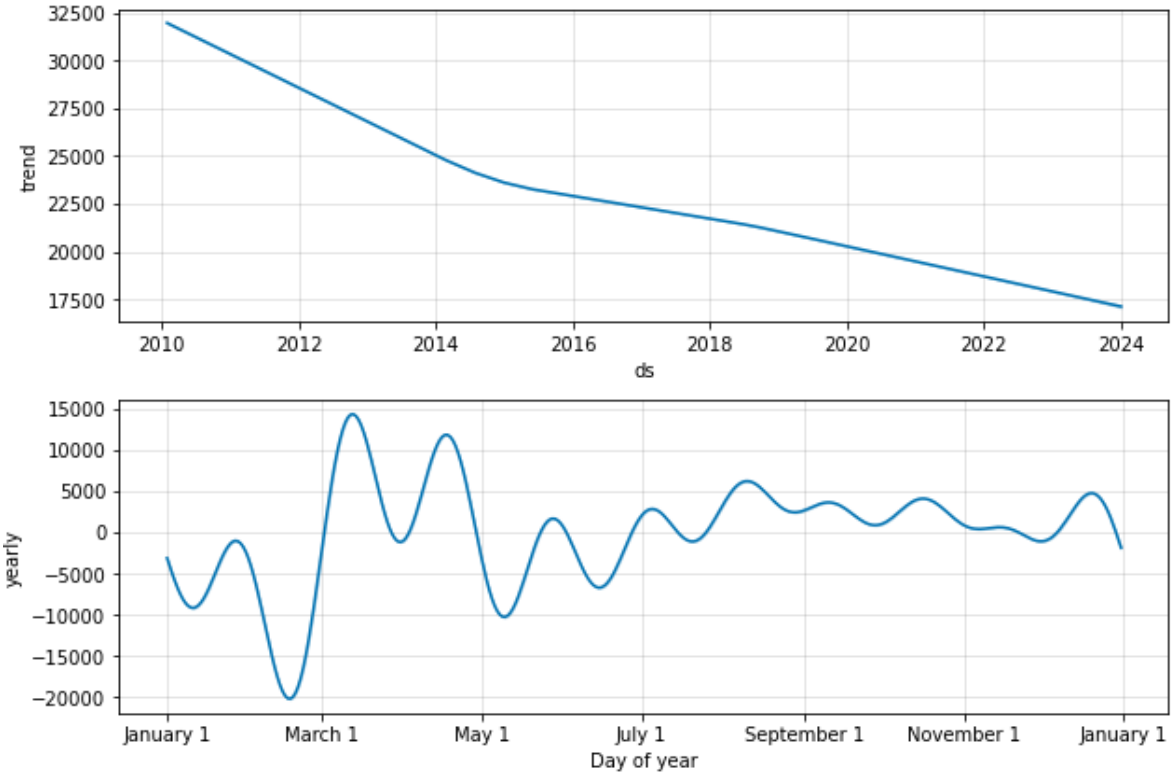


Figure 5 Crime Prediction

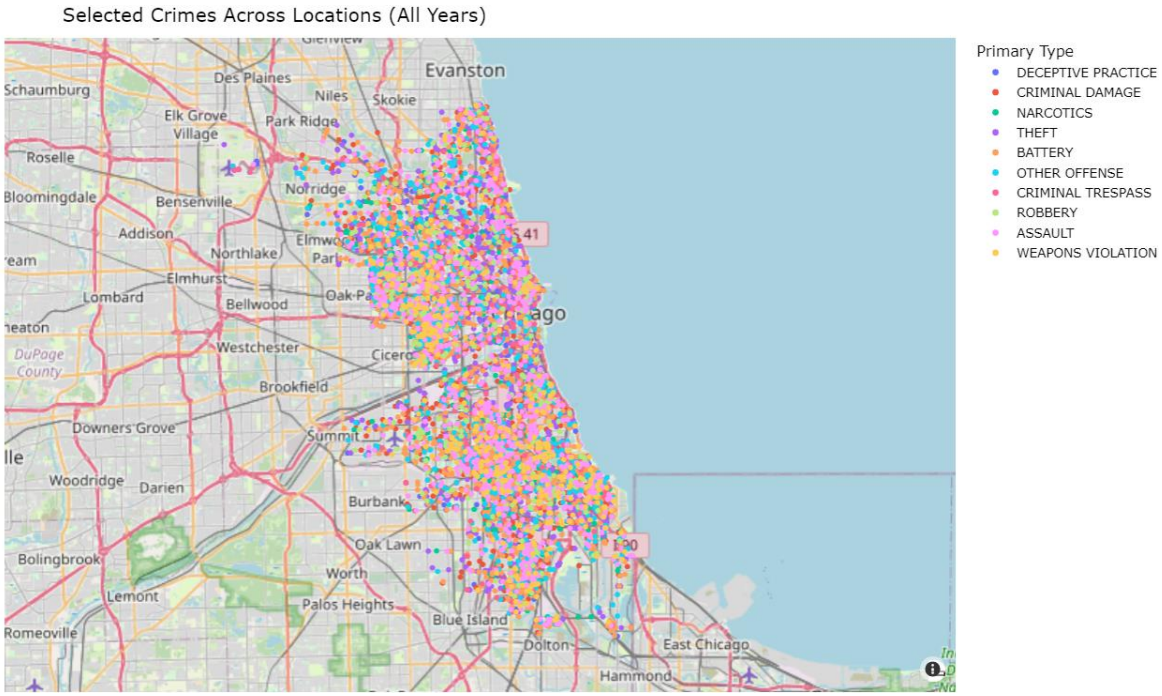


Figure 6 Visualization of Top 10 Crimes Location

Conclusion:

Crime is one of the most dangerous phenomena for any country. To reduce and nonproliferation of crime, it requires new techniques that can deal with the vast amount of data, where the data cannot be analyzed with traditional analysis techniques. Therefore, Prophet, a procedure for forecasting time series data based on an additive model was deployed and results were analyzed. Our contribution is to build an intelligent crime prediction model, where the model predicts different types: crime type, time of the crime, the place of crime).

Future Work

Clustering data and analyzing it is an important step to many other machine learning algorithms. In particular, prediction of crime points can be achieved by multiple methods:

- Minimizing the distance between a crime occurrence and the centroid of a cluster
- Performing regression analysis on the identified clusters and fitting crimes to the best fit line

We could also implement the K-nearest neighbors' algorithm (which finds the most similar points given a specific points), with the condition that they be in the same cluster as we predict our data point will be in. This would provide a list of k possible crime types that are most likely to occur in Chicago on January 1 at midnight.

Recommendations:

The goal was to deploy the officers based on the crime data and socio-economic indicators. Following are the recommendations:

- For particular community areas, the crime count was found to be high but the arrest rate was lower than the average Chicago arrest rate. For these areas the deployment of police officers needs to be increased.
- It was given that few community areas have the high number of vacant housing units. Thus, more than average deployment of police officers is required in the weeks where we have predicted the crime count is going to be highest. Also, based on exploratory data analysis we should have more than average deployment of police officers between 12 PM and 12 AM
- Community areas where there are highest number of single parents with child. Thus, police officers need to be deployed in these areas during the weeks when the crime count is going increasing
- Based on clustering, we found out that there are community areas where there is high number of domestic abuses, high vacant housing units, high single parent with child and so on. These community areas need to monitored more than average during the weeks when the crime count is going to be high
- Special task forces that can handle extremely violent crimes should be deployed on the specific days of week when the violent crimes are high in number based on the EDA performed. As the crime count was not predicted on daily basis, we can look at the weekly predicted numbers to deploy the special task forces

- Building new parks will help to engage people socially and might help in the reduction of crime as it is found from our analysis that community that have a greater number of parks have less crime

Questions/Answer

- How has the number of various crimes changed over time in Chicago?
    - Crime numbers have fluctuated over the years, with certain crime types seeing a decline in recent years due to increased surveillance and community engagement programs.
- How have the number arrests corresponding to the crimes changed over time in Chicago?
    - Arrest rates have decreased over time for minor offenses, reflecting changes in law enforcement policies, while remaining steady for serious crimes.
- Are there any trends in the crimes being committed?
    - Seasonal spikes are observed (e.g., thefts increase during holidays), and certain crimes are more frequent in economically disadvantaged neighborhoods.
- Which crimes are most frequently committed?
    - Theft, battery, and narcotics-related crimes are among the most frequently committed in Chicago.
- Which locations are these frequent crimes being committed to?
    - High-crime areas include public transit locations, residential neighborhoods, and commercial areas.

- Are there certain high crime locations for certain crimes (etc sexual harassments)?
    - Yes, specific crimes like sexual harassment often occur in public spaces such as transit areas and parks.
- How has the number of certain crimes (etc homicide) changed over the years in Chicago?
    - Homicide rates have seen a fluctuating trend, with periods of increase during economic downturns and slight declines during community intervention efforts.

References:

1. Currie32. (n.d.). *Crimes in Chicago*. Retrieved from https://www.kaggle.com/currie32/crimes-in-chicago
2. Ogunbode, F. (n.d.). *EDA of Crime in Chicago from 2012-2016*. Retrieved from https://www.kaggle.com/femiogunbode/eda-of-crime-in-chicago-from-2012-2016/discussion
3. Djonafegnem. (n.d.). *Chicago Crime Data Analysis*. Retrieved from https://www.kaggle.com/djonafegnem/chicago-crime-data-analysis
4. Pandas Community. (n.d.). *Pandas Tutorials*. Retrieved from http://pandas.pydata.org/pandas-docs/stable/tutorials.html
5. Jain, Naman. "Hands on Machine Learning with Chicago Crime Data." *Medium*, Medium, 2 Mar. 2020, https://medium.com/@namanjain2050/hands-on-machine-learning-with-chicago-crime-data-3657b713d62c.
6. "Precise Event-Level Prediction of Urban Crime Reveals Signature of Enforcement Bias." *Home*, 11 Feb. 2021, https://www.researchsquare.com/article/rs-192156/v1.