FEBRUARY 27, 2024

FINAL REPORT

X

CHICAGO: EMPOWERING A SAFER COMMUNITY THROUGH DATA-DRIVEN CRIME INSIGHTS

Presented by Team 51

Table of Contents

FINAL REPORT

Problem Statement and Overview

Chicago is the third largest city in the United States. It is a hub for finance, commerce, telecommunications, transportation, and is home to some of the most competitive schools in the country. By the end of 2023, **violent crime in Chicago was at an 11-year high** due to an increase in robbery and assault. Chicago followed a national trend with declines in homicide by 13%, but still has **the 6th highest homicide rate among top cities**, which is more than four times higher than New York City. **Automobile theft was at a 20 year high**, while burglaries were at historic lows.

For 2024, Chicago's mayor has increased the police budget along with raises for police officers in the hopes of reducing crime with more funds and resources. **One of Chicago's main goals for 2024 is to focus on crime prevention** through place-based and people-based approaches, which includes a range of intervention, socioeconomic, education, community involvement, and policing approaches. The **police chief supports data-driven policing** as a part of the strategy[1].

The City of Chicago has retained Team 51 to discover and analyze data-backed drivers of crime and recommend ways to reduce it.

Project Goals

There is a significant opportunity for Chicago to embrace data-driven insights as part of its holistic plan to reduce crime with the help of Team 51's years of experience in data science and consultation. To support Chicago in achieving its crime prevention objectives, Team 51 has developed and delivered crime prediction models, information access tools, and data-driven recommendations to help reduce crime. Team 51 organized the work into four substantial goals that address all aspects of the project.

[1] Ponce, Anthony. 2023. "Chicago police superintendent lays out 2024 goals for department". Fox News Chicago. (fox32chicago.com).Dec 11. https://www.fox32chicago.com/news/chicago-police-superintendent-lays-out-2024-goals-for-department

Understand Crime	Identify the top factors that predict crime and identify crime hotspots by Chicago community area.
Predict Crime	Develop a model that predicts crime as a function of community area and other factors by drawing from several datasets and machine learning techniques.
Segmentation	Investigate datasets for deeper insights into the combination of factors that contribute to differences in crime.
Information Access	Develop a dashboard, mobile app and chatbot that facilitates access to crime information for Chicago officials, police, and residents. **Since Team 51's last report, a website has also been added** to the suite of tools that provides crime information.

Data

DESCRIPTION OF DATA

The focus of our data is on Chicago crime and factors that may correlate with crime. Data compiled to support the analysis were obtained from six publicly available sources. The City of Chicago's publicly available data portal provided most of the data[2]. The initial data pull was more than 1.1 GB in size, represented over 14 million records (rows of data), and over 200 million data observations. After extensive data cleaning, a high-quality master dataset consisting of 250 thousand records, with each record representing a unique Chicago crime, and 61 columns, with each column representing a variable, was combined into a single comma-separated-value (.csv) file of 118 MB in size.

Table 1 summarizes the data feature category and the variables associated with that feature that are being included in this project. The "Overview and Transformation of Data" section describes details regarding extraction transformation, and loading (ETL) .

[2] See Appendix B for specific Data Sources.

Table 1. Feature and Variable Description of Master Dataset

Feature Category	Variable Description
Crime Type (Target Variable)	• Categorized Crimes: Violent, Non-Violent, and Property
Location	• Latitude and Longitude • Location Description (e.g. Street, Apartment) • Beat • District • Ward • Chicago Community Areas • Community Area Population Density • Chicago Region
Weather	• Average Temperature (Degrees Fahrenheit) • Average High Temperature (Degrees Fahrenheit) • Average Low Temperature (Degrees Fahrenheit) • Precipitation • Snow
Time	• Date • Month • Year • Day of Week • Day Type
Proximity to Crime	• Closest Police Station Address • Distance of Crime to Closest Police Station
Community Complaint	• Environmental Complaints (e.g. water quality, sewer concerns, lead tests, debris removal)
Sex Offenders	• Total Offenders in Each Community Area

DATA OVERVIEW

To assure that Team 51 was working with high quality data, meticulous attention was given to extracting, loading, cleaning and transforming the data with Python.

Data Extraction, Loading and Access Through Google Cloud Storage

The first phase was to extract the data from its sources, load it, and then transform it into a usable format for analysis. That majority of the data came from government sources, particularly the City of Chicago's data portal. Team 51 pulled together the Crimes dataset, the 311 Complaints dataset, the daily weather data from Chicago's Airport, the Environmental complaints dataset from the Department of Public Health, the Sex Offenders registry, and the location of every police station into a single combined dataset. This enabled faster analysis through Tableau and Python by allowing our analysts to skip the data gathering step by using a monolithic dataset then filtering down.

This monolithic dataset was uploaded to **Google Cloud Storage** enabling low-cost, long-term storage designed for quick retrieval and archival. The data was also added to **BigQuery** to enable analysis through Structured Query Language (SQL) which increases the accessibility of the data and allows for deeper exploration of each data source outside of the monolith.

TRANSFORMATION OF DATA

Data Transformation Overview

As the original datasets were merged, data needed to be reviewed and addressed for a number of potential issues. Removing duplicates and highly correlated variables, addressing missing data, reviewing outliers and irrelevant data were some of the steps taken to derive a clean master dataset that supported our modeling work. In some cases, additional variables were derived to aid in the analysis.

Removing Duplicates and Highly Correlated Variables

From conducting our Exploratory Data Analysis (EDA), we discovered that there were a few duplicate variables. We excluded these from our master dataset. These variables include 'name' for community regional name, 'location.human_address', 'location.latitude', 'location.longitude', 'month_num' for the numerical number for the month in the year, and 'updated_on' for the date and time when the crime was logged into the Chicago crime dataset. These were excluded to avoid overpredicting our models with multiple columns of the same value, to reduce model complexity, to improve model performance, and interpretability.

In addition, we explored both keeping and excluding other highly correlated variables so that an informed decision could be made on how to best treat these variables. We ran several experiments for comparison. We elected to remove highly correlated variables, and some redundant variables. During the modeling phase, we observed that incorporating these variables adversely impacted model overfitting, making it challenging to discern the individual predictive power of each variable. Additionally, advanced techniques such as dimensionality reduction methods (e.g., principal component analysis) can serve as alternatives for handling correlated variables. We have implemented principal component analysis (PCA) as part of the unsupervised modeling methods.

HANDLING MISSING DATA

When reviewing missing values in the master dataset, we discovered that there was a small percentage of rows, with "total_offenders_in_ca", or total number of sex offenders based on each community area, having the highest percentage of 1.58% missing values. Upon further investigation, it was discovered that community areas 35 and 36, which are also referred to as Douglas and Oakland[3], respectively, in the South Side do not have any registered sex offenders in these areas. With this information, we filled these missing values with zeros. In addition, for the location demographics, that include location_description and longitude, we filled the null values with -99 to indicate these values as outliers. Less than 0.5% of location information had null values. The remaining missing variables are in relation to environmental complaints such as water quality concerns, sewer concerns, debris removal and others. This data had less than 0.25% missing values and it was assumed that if a value is missing, it is an indication that there was no complaint and the missing values were filled with zeros.

Table 2. Summary of Missing Data Handling

Feature	Data used to fill Missing Value	Missing Value Rationale
Total Sex Offenders in Community Area	0	No sex offenders registered in the area
Location Description	-99	Deliberate choice to flag the location as an outlier
Longitude	-99	Deliberate choice to flag the location as an outlier
Environmental Complaints	0	Assumed no complaint was filed

[3] Refer to Appendix D for Map of Community Regions and Community Areas

Handling Outlier Data

We did not discover any major outliers that were deemed unfit to include in our analysis and modeling. However, we will continue to assess this as the model is used in production.

Additional Feature Engineering

- For additional feature engineering, Team 51 categorized the **31** different types of crime found in the 2023 Chicago Crime dataset into three categories for **nonviolent, violent**, and **property crimes**, as shown in Appendix C. This new variable is **the target, predicted variable** for the supervised and unsupervised methods explored.

- Team 51 cross-referenced the “community regions” with each “community area” in the master dataset, as shown in Appendix D. This provided clearer insight on how crime rates differentiate between each region and area. The community region and community area distinctions also align with how Chicago officials, police, and residents describe locational geography.

- Team 51 included day of week, daytype, month, and year based on the dates provided in the original data sources. These were essential for developing and interpreting the results for the Exploratory Data Analysis (EDA). This will also provide insight for the models developed, as crime rate may differentiate based on calendar and time features.

- There were 105 categorized community 311 complaints for 2023 which include Animal, Business, Drugs/Alcohol, Environmental, Neighborhood, and Vehicle. We excluded community 311 complaints that were unrelated to the environment. Each of the categorized complaints may, or may not hold great value in predicting crime; however, Team 51 is taking a measured approach to incorporating additional data. Therefore, initial models will include environmental complaints and additional 311 complaint categories may be added in future project phases.

- Finally, Team 51 is primarily focusing on year 2023 data in this project because our initial analysis showed that crime rates were heavily impacted by COVID-19 for 2020-2022 (Figure 1). The 2023 information provides a robust amount of data in the master dataset. That said, we took the additional step to further explore our analysis using data from different time periods between 2020 and 2023, which confirmed that year 2023 data provides quality results.

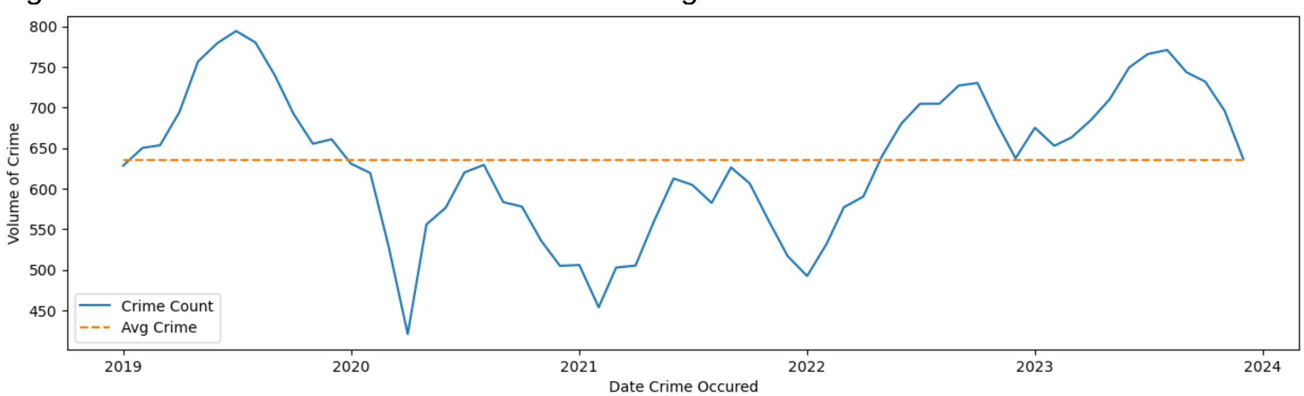
Project Process, Controls and Cloud Platform

Team 51 integrated robust project and documentation controls into our work. This assures that project data, code, artifacts, documentation, and versioning are well organized, secure, and available. Such governance is integral to delivering a robust project with dependable and durable results. The data and modeling work is following the well-known Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, an industry proven method for effectively guiding data intensive projects. Additionally, the project will be hosted on Google Cloud Platform (GCP), which greatly enhances the scalability, reliability, security, and accessibility of this work. Features of GCP should benefit the City of Chicago as their sophistication of data-driven policing and crime prevention grows.

Exploratory Data Analysis

Analysis of the cleaned master dataset was performed using Python and Tableau. Based on Chicago crime trends since 2019, there is a noticeable decrease in crime during the COVID-19 pandemic from 2020 to mid-2022, with the lowest volume of crime in early 2020 (Figure 1). Because the pandemic had a considerable effect on crime rates, Team 51 will primarily utilize crime data from 2023 in the analysis.

Figure 1. Number of Crimes Committed in Chicago from 2019-2023.



WHAT DOES CRIME IN CHICAGO LOOK LIKE?

One of Team 51's project goals is to understand what crime looks like in Chicago. Analysis shows that property and violent crimes have consistently higher crime rates in comparison to nonviolent crimes[4] (Figure 2). Further, crime of all types is the highest in the South and Far Southeast side of the City[5] (Figure 3).

Figure 2. Crime Rates Per 1000 People by Crime Type for 2023

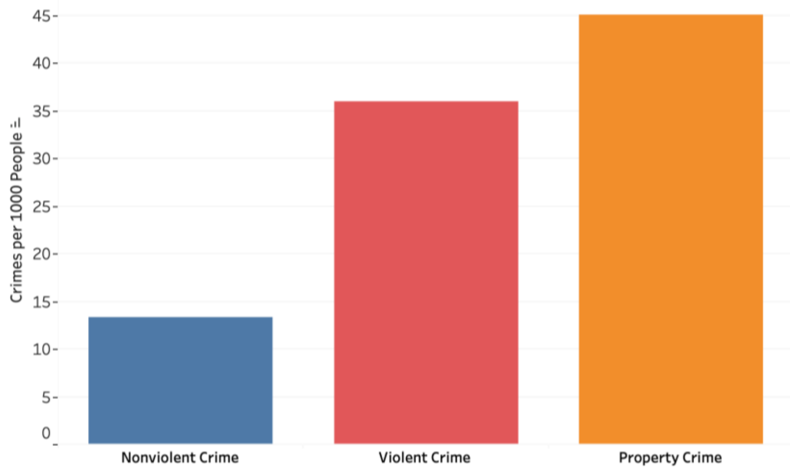
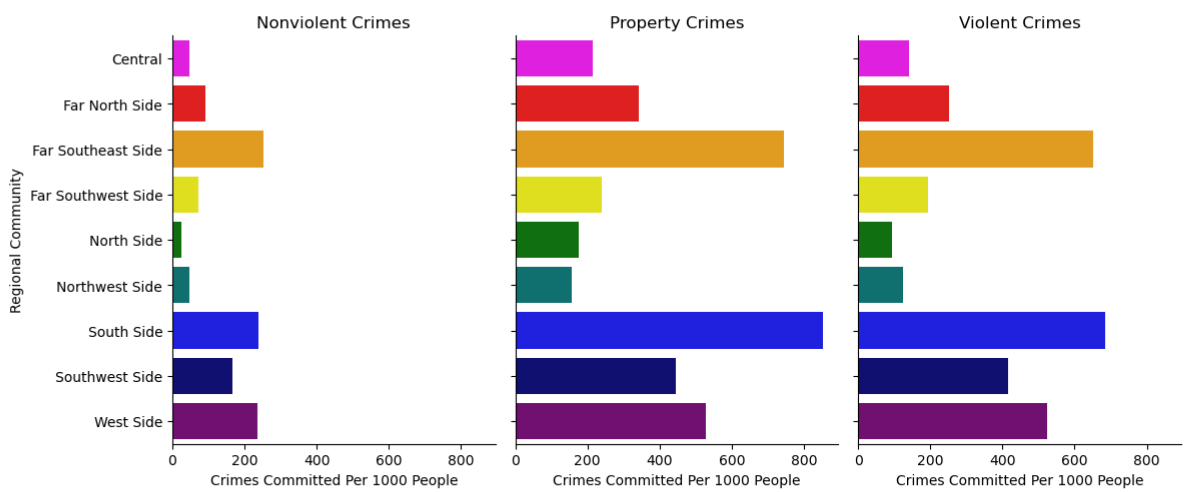


Figure 3. Crime Rates Per 1000 People by Crime Type and Community Region for 2023



---

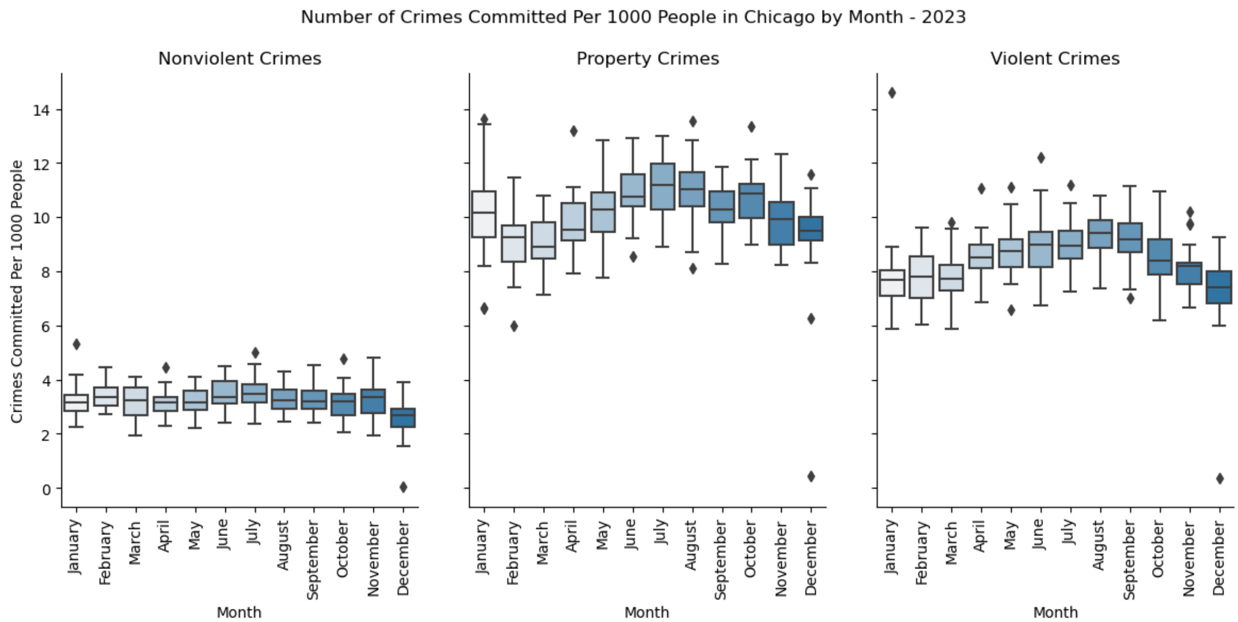
4 See Appendix B for list of crimes categorized by crime type

5 See Appendix C for map of community regions and community areas.

WHEN DOES CRIME MOST FREQUENTLY OCCUR?

There is more crime in the warmer months from May to September for nearly all crime types (Figure 4).

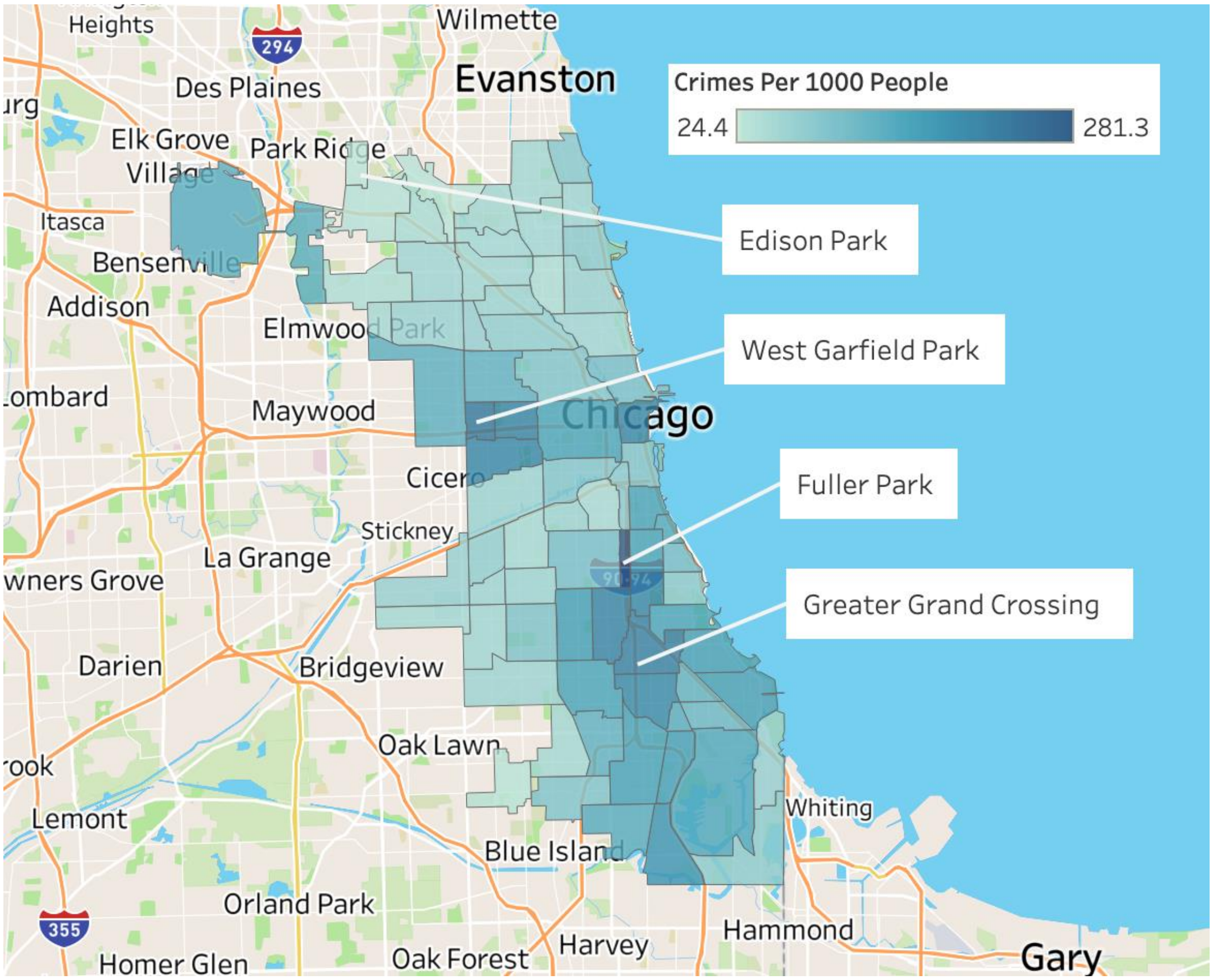
Figure 4. Chicago Monthly Crime Rate Compared to Average Temperature for 2023



WHERE ARE THE CRIME HOTSPOTS?

There is a wide range of crime rates between Chicago's 77 Community Areas. The lowest crime rate is 24 crimes per 1000 people in the Edison Park community, whereas the highest crime rate is more than 10 times greater at 281 crimes per 1000 people in Fuller Park. The top three crime hotspot areas[6] are Fuller Park, West Garfield Park, and Greater Grand Crossing communities (Figure 5).

Figure 5. Crime Per 1000 People by Chicago Community Area in 2023

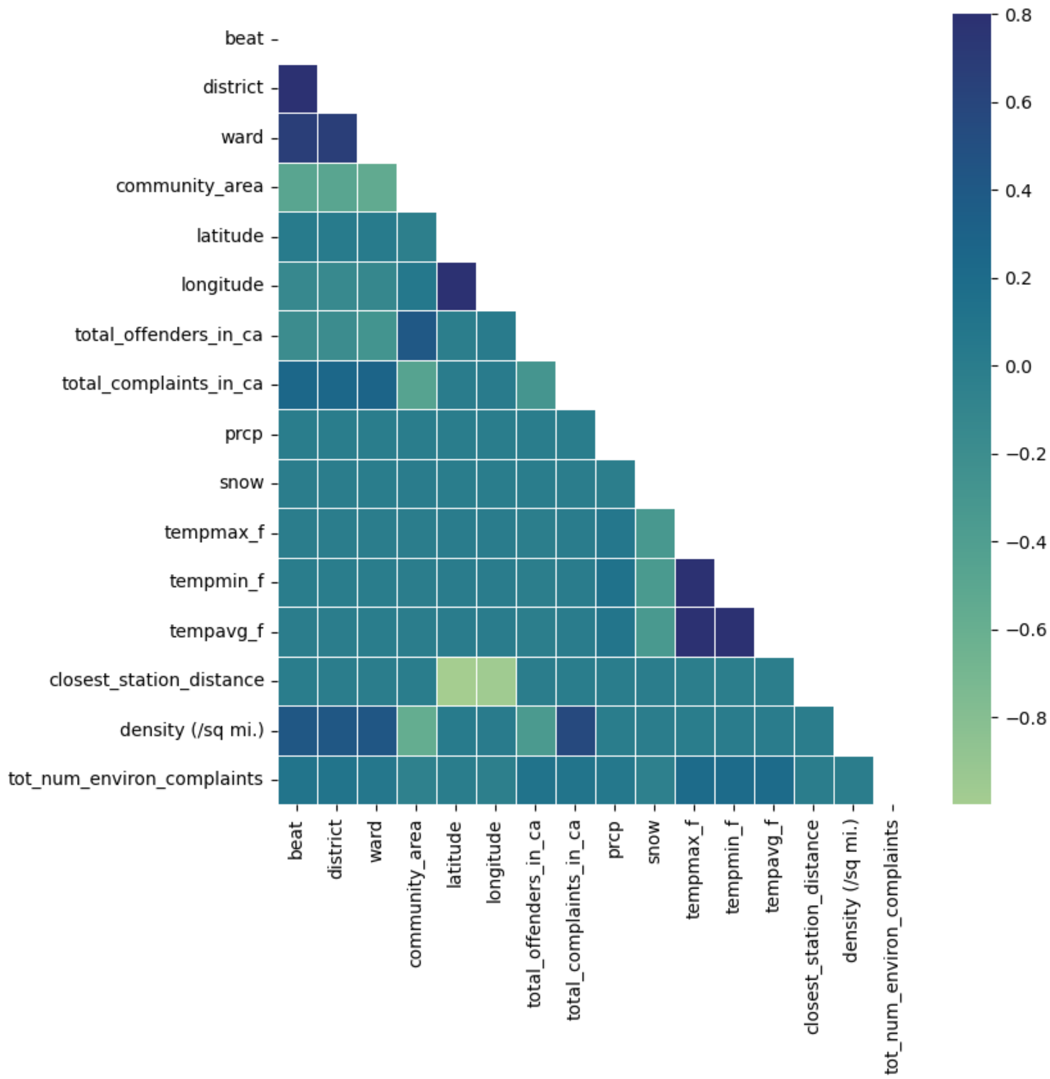


[6] See Appendix C for map of community regions and areas.

HOW ARE THE FEATURES (VARIABLES) CORRELATED?

A correlation matrix shows closer relationships based mostly on location features, weather, and sex offender variables (Figure 6).

Figure 6. Correlation Matrix



Modeling

The primary modeling approaches consist of both supervised and unsupervised methods in support of our goals to develop a model that predicts crime and identifies the combination of factors that contribute to differences in crime through segmentation. The programming language Python was the primary tool used for modeling.

Supervised methods are used to train predictive models, and are the types of methods Team 51 is using to predict crime types and crime counts. They classify outcomes based on input data by learning the relationship between inputs and corresponding outputs from labeled training examples. Team 51's supervised models include linear and logistic regression, decision tree, random forest, and gradient boosting.

Unsupervised methods are intended to identify relationships, patterns and clustered groupings in data that does not have a predefined target variable. Unsupervised methods are being used for segmentation to identify crime profile clusters that share similar characteristics. Team 51's unsupervised methods include K-means, Clustering and Principal Component Analysis (PCA) methods.

These models can be a powerful tool. Great care is being taken to use data and modeling in a transparent and responsible way that enables community engagement and understanding throughout the process. That is why Team 51's work uses publicly available data, transparent "glass-box" analytic methods, and data visualization tools that make information accessible to everyone

SUPERVISED METHODS

The intent of this section is to discuss results from supervised models that were primarily used to predict crime type (property, violent and non-violent) as well as trends in the number of crimes (crime count trends).

CRIME TYPE PREDICTION

Our models are based on predicting the categorized type of crime for nonviolent, violent, and property crimes. These crimes are predicted based on a variety of different factors, which include the demographics of crimes for 2023, police station proximity, weather, sex offenders by community area, environmental complaints, and by time of year. Crime type information will be useful for the Chicago police, officials, and residents to proactively identify interventions and resources that may need to be deployed preemptively to address crime in the different community areas. Team 51 is committed to using models that are easier to explain to our project stakeholders, so the recommended crime prediction model balances model accuracy with explainability. For comparison purposes, Team 51 ran several models to determine the best performing algorithm.

To prepare our data for the machine learning algorithms we investigate, which include decision trees, random forest, gradient boosting, and regression models, a split of 80% training data and 20% testing data was used. A training dataset is vital for these models to gain confidence in accurately predicting the type of crime with new, unfamiliar data such as the test data. The model's performances were based on their accuracy scores and Root Mean Square Error (RMSE). The accuracy score provides how accurate the model was in predicting the type of crime. The RMSE is used to measure the distance of the predicted values in the models to the observed values from the test data. It is better to have a smaller RMSE value, as this would indicate that the predicted values are fairly close to the observed values. Based on our findings in Tables 3 and 4, Gradient Boosting has the highest accuracy for the test data of 59.7% with the second lowest RMSE of 0.782. Although Linear Regression has the second lowest accuracy, it does have the best RMSE of 0.725. Figure 7 illustrates the accuracy score and RMSE for each supervised learning method based on the test data.

Table 3: Performance Summary for Supervised Models Based on Train Data

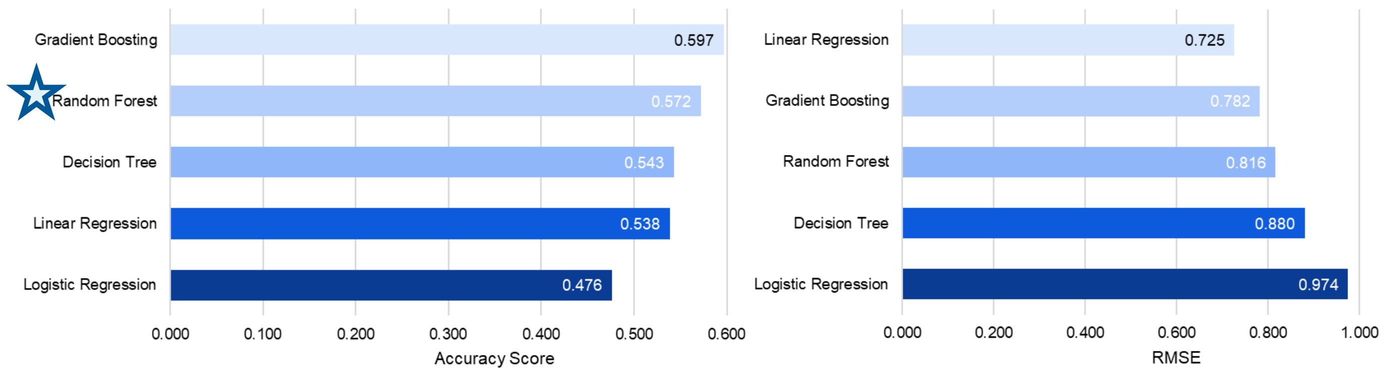
Model	Accuracy	RMSE
Decision Tree	0.543	0.878
Random Forest	0.995	0.082
Gradient Boosting	0.598	0.780
Linear Regression	0.536	0.728
Logistic Regression	0.478	0.973

Table 4: Performance Summary for Supervised Models Based on Test Data[7]

Model	Accuracy	RMSE
Decision Tree	0.543	0.880
Random Forest	0.572	0.816
Gradient Boosting	0.597	0.782
Linear Regression	0.538	0.725
Logistic Regression	0.476	0.974

[7] “Crime type” prediction accuracies for related, but less sophisticated studies: Philadelphia 27%; Chicago 37%; San Francisco 43 & 64%, Vancouver 44%.See Reference section for sources.

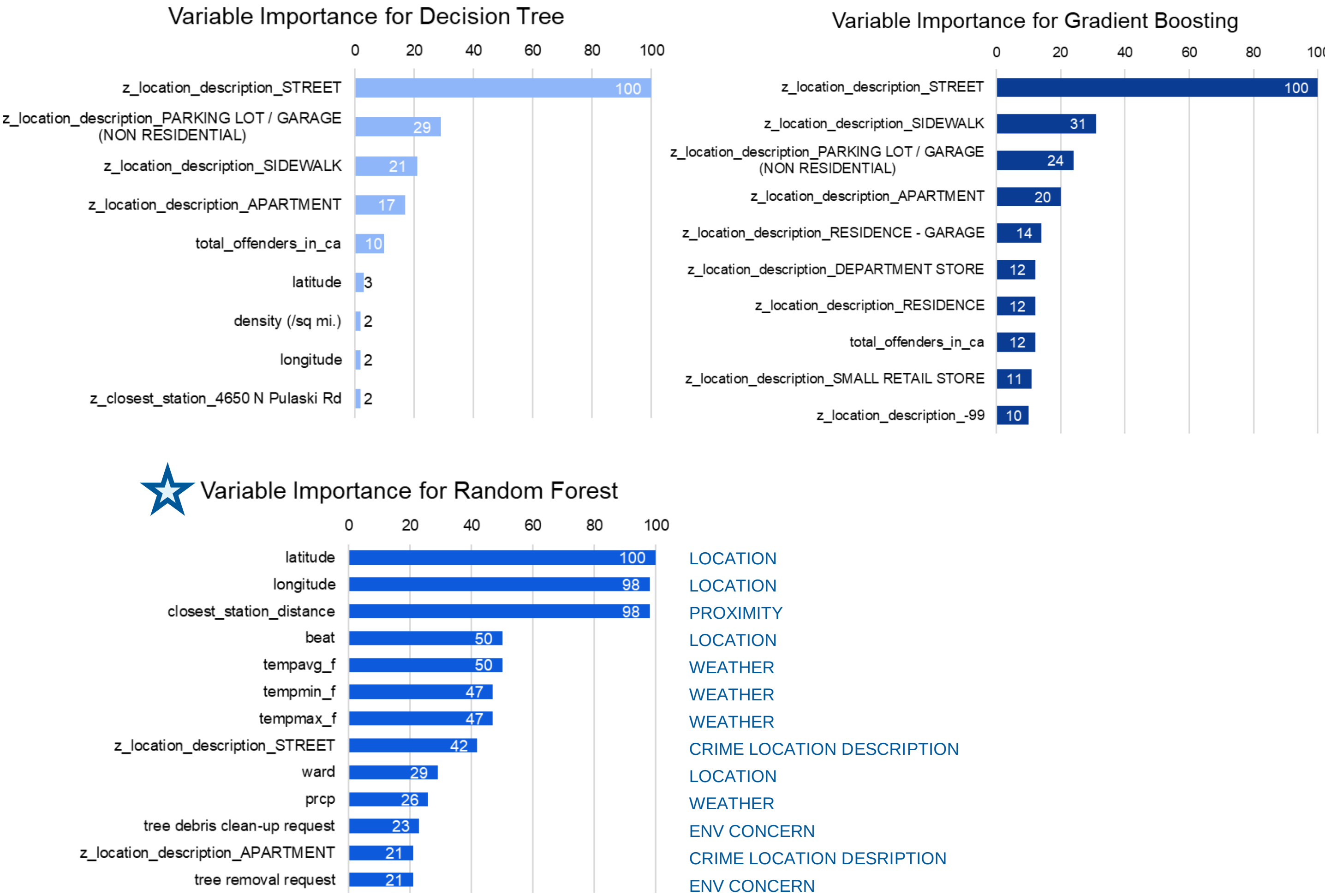
Figure 7: Performance for Supervised Models Based on Accuracy Score (left) and RMSE (right) for Test Data



The results from Tables 3 and 4 are based on excluding duplicate and highly correlated variables as described in the "Transformation of Data" section. Each of the highly correlated variables are direct indicators of the type of crime committed, which means these variables were preemptively predicting the type of crime before it had occurred. This resulted in the accuracy rates to be high and the models to overfit. After excluding the highly correlated variables, the interpretability of the model improved. We discovered that the location description and total number of sex offenders in each community area had the highest variable of importance for the Decision Tree and Gradient Boosting models while latitude, longitude, and distance to the closest police station were the highest for Random Forest, as shown in Figure 8. It was interesting to discover that although gradient boosting has the highest accuracy score, it has a similar ranking of variance importance for the Decision Trees, as they both contain the same top three variables.

It is important to understand the factors that predict crime as they provide a data-backed perspective on what is driving crime in Chicago. By targeting resources and interventions that ameliorate these factors, the crime rate should improve.

Figure 8. Feature (variable) Importance of Predicting Crime for Each Model



Even though the Gradient Boosting model has the highest accuracy of all the models, the Random Forest model is just slightly less accurate than Gradient Boosting. Random Forest models are more intuitive to understand in that the Random Forest model essentially builds multiple decision trees and combines their predictions to improve model accuracy. Additionally, the feature importance is more useful for the Random Forest Model in that it contains elements of location, proximity to police station, weather and environmental complaints as predictive variables. For these reasons, **the Random Forest Model is our selected model**.

The Random Forest model was further explored with a variety of different time periods between 2020 and 2023. Team 51 discovered that there is a slight decrease in accuracy and increase in RMSE with each extended time period, where the period 2020 to 2023 had the lowest accuracy of 55.4% and the year 2023 had the highest accuracy of 57.2% (Table 5). This change in accuracy can be the result of the effects of COVID-19, which experienced a significant decrease in crime rate from 2020-2022. The highest variables of importance were the same as well, where latitude, longitude, and distance to the closest police station were the highest variables of importance for each time period.

Table 5. Performance Summary for Random Forest Model Based on Different Time Periods with Test Data[8]

Time Period	Accuracy	RMSE
2020 to 2023	0.554	0.824
2021 to 2023	0.562	0.817
2022 to 2023	0.569	0.819
2023	0.572	0.816

Crime Count Prediction and Forecasting

The intent of this section is to present two methods for predicting the trend in crime counts. (1) The Crime Count Prediction method uses a Random Forest model and most of the same Crime Type Prediction variables to determine the trend in crime counts by community area. Predicted crime count trends from 2020 through 2023 are compared to actuals and show good correlation. (2) The Crime Count Forecasting method uses an exponential smoothing model in Tableau Desktop to predict crime count trends for 2024 for each community area and Chicago overall. These predictions are based on a simplified dataset that includes the number of crimes committed for each community area by date from 2021 to 2023. Interestingly, this model predicts a lower overall crime trend in 2024. Details of each method are presented in the next section.

Crime Count Prediction

With the Random Forest model as our selected model, Team 51 further explored this method to predict crime counts in each community area. The intent of this prediction is to indicate how crime is trending as opposed to providing definitive crime numbers for

[8] "Crime type" prediction accuracies for related, but less sophisticated studies: Philadelphia 27%; Chicago 37%; San Francisco 43 & 64%, Vancouver 44%.See Reference section for sources.

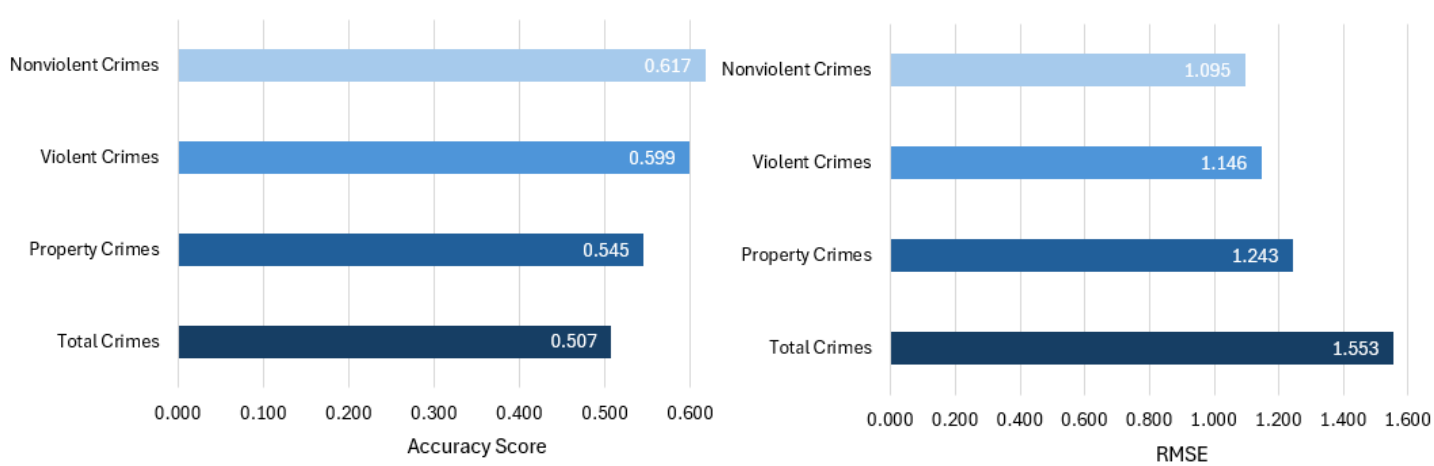
the future. These predictions are based on factors similar to the crime type analysis, with the exception of using a longer time period from 2020 to 2023 and the exclusion of police station proximity. Results from this analysis will be useful for Chicago police and officials as they will be able to proactively plan resources in community areas expected to have higher crime counts based on the time of year. Team 51 has provided trends for total combined crime counts as well as crime count trends for nonviolent, violent, and property crimes in each community area.

The data used for the Random Forest model is prepared in a similar way to the categorized crime type prediction models, where a split of 80% training data and 20% test data was used. Based on our findings, the predicted crime counts for nonviolent crimes has the highest accuracy of 61.7% and lowest RMSE of 1.095. Interestingly, the predicted crime count for total combined crimes has the lowest accuracy of 50.7% and the highest RMSE of 1.553 (Table 6). Figure 9 illustrates the accuracy score and RMSE for the total and categorized crimes based on the test data.

Table 6. Crime Count Trend Prediction Performance Summary for Random Forest Model Based on Categorized Crime Type with Test Data

Time Period	Accuracy	RMSE
Total Crimes	0.507	1.553
Property Crimes	0.545	1.243
Violent Crimes	0.599	1.146
Nonviolent Crimes	0.617	1.095

Figure 9. Crime Count Trend Prediction Performance of Random Forest Model Based on Accuracy Score (left) and RMSE (right) for Categorized Crime Type with Test Data



Upon further understanding the variables used to determine the accuracy scores and RMSE in the Random Forest model for crime rate predictions, we discovered that the total number of sex offenders in each community area, population density, and community area are the highest variables of importance for both the total combined crimes and each categorized crime type (Figure 10). It was interesting to discover that both total crimes and nonviolent crimes had the same rankings in the same order, despite having the largest differences in their accuracies. It should be noted that nonviolent crimes have the lowest count for the total number of crimes committed. This can be an indication of an easier prediction for the number of crimes to occur by date in each community area.

Figure 10. Feature (variable) Importance of Predicting Trends in the Number of Crimes for Total and Categorized Crimes

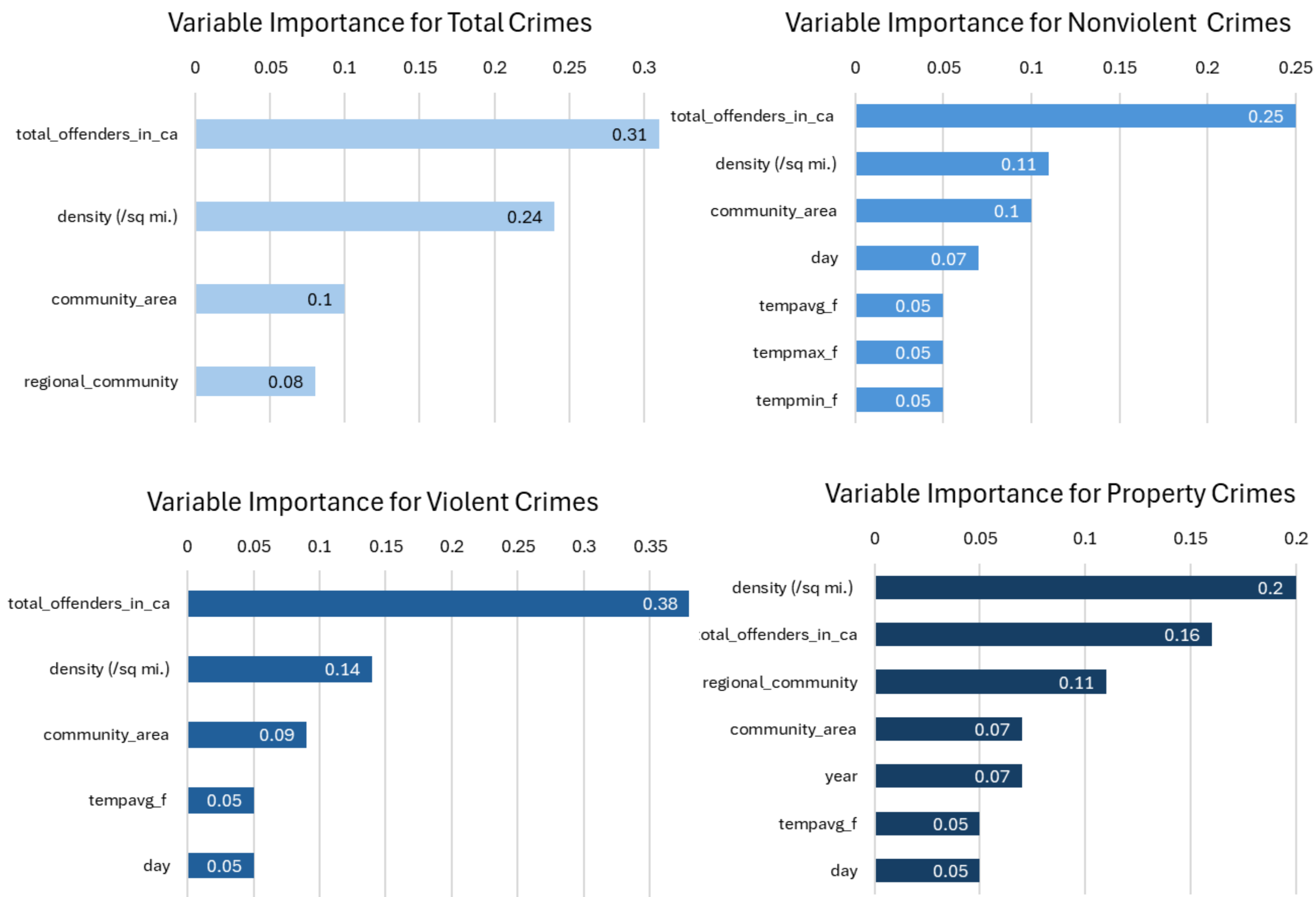
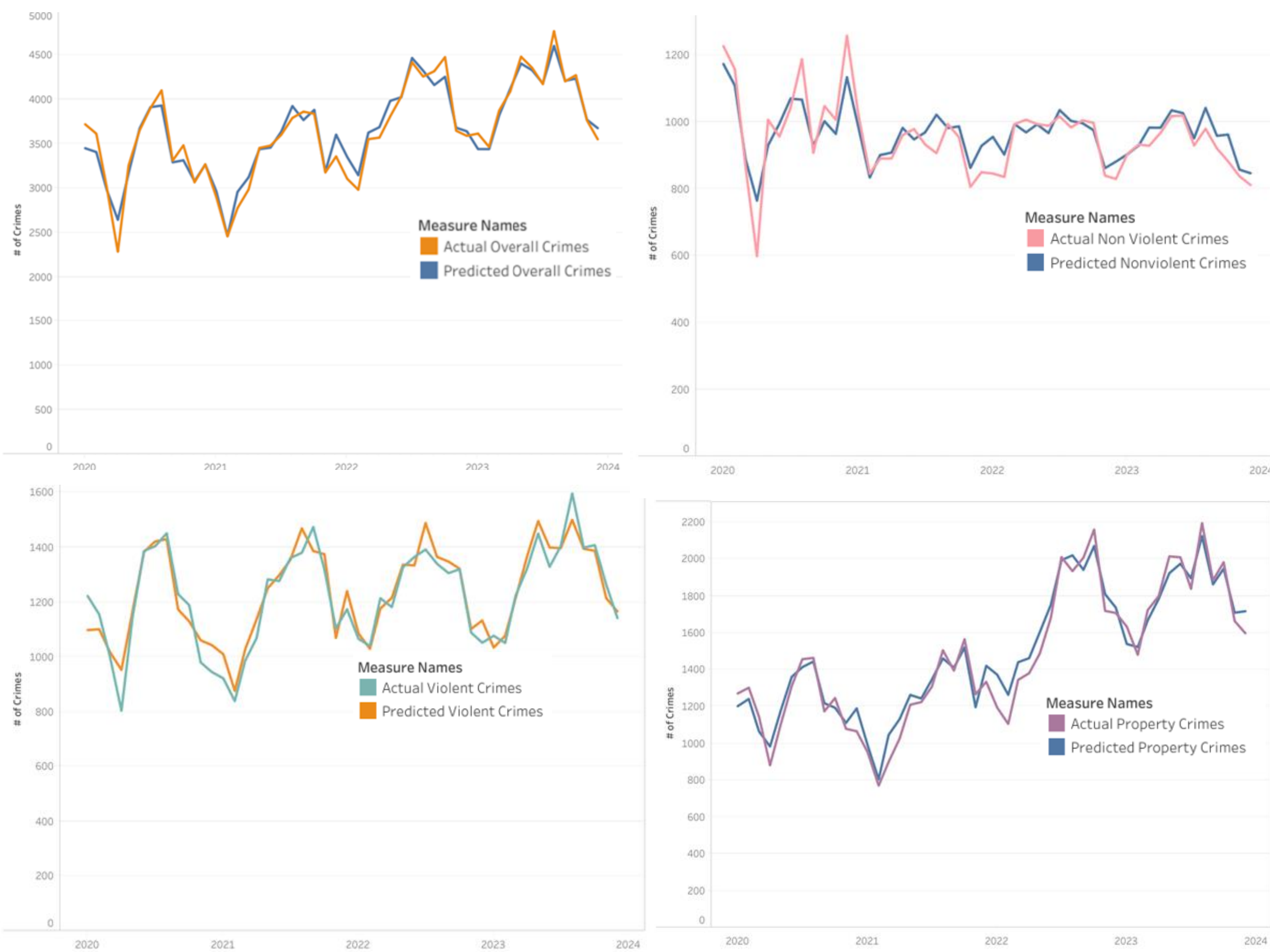


Figure 11 illustrates the actual versus predicted number of crimes for total crimes as well as each categorized crime type for violent, nonviolent, and property. It is important to note that these results are based on the test data that uses 20% of the original dataset from the train-test split. When reviewing each categorized crime type, there is a larger difference between actual and predicted number of crimes in the year 2020, especially for the month of April when the COVID-19 pandemic began to peak. Despite having the highest accuracy score, the actual versus predicted number of crimes for nonviolent crimes has more variation during the years 2020 and 2021, while in more recent years the trends are more consistent. This provides useful insight in how our model is able to perform better with more recent dates.

Figure 11. Crime Trend Predictions Based on Test Data (20%) For Overall (top left), Non-Violent (top right), Violent (bottom left) and Property (bottom-right) Crimes



CRIME COUNT FORECASTING

Team 51 further extrapolated our crime trend prediction models by using Tableau forecasting modeling to predict the crime count trends for the year 2024 for each community area and Chicago overall. Exponential smoothing models in Tableau Desktop enable the forecasting of quantitative time-series data. The models can effectively capture the seasonal patterns or changing patterns in the data to extend them into the future. This method also assigns greater weight to recent observations compared to older ones.

These predictions are based on a more simplified dataset, where we used the number of crimes committed for each community area by date from 2021 to 2023. This type of dataset was used for both the total number of crimes as well as each categorized crime type. Although this model does not take into consideration other crime demographics, this can still be a useful tool for the Chicago police, as it provides initial insight on how officials can be more prepared for the potential future changes in crime rate.

Based on the results in Figure 12 for the overall prediction, when comparing an aggregated total for all community areas, there is a decrease in the number of crimes committed for the year 2024. The highest peak for this year is in July, which is significantly lower compared to the highest peaks in 2022 and 2023 by about 2,000 crimes. It is also interesting to note that the margin of error is quite wide, as it predicts the peak number of crimes to exceed 24,000 with a significant drop in December 2024 of less than 12,000. This is an indication that this model should be further explored with other metrics, in addition to community area and date.

Figure 12. Total Crimes Per Month Prediction for 2024 (Using Tableau Forecasting)

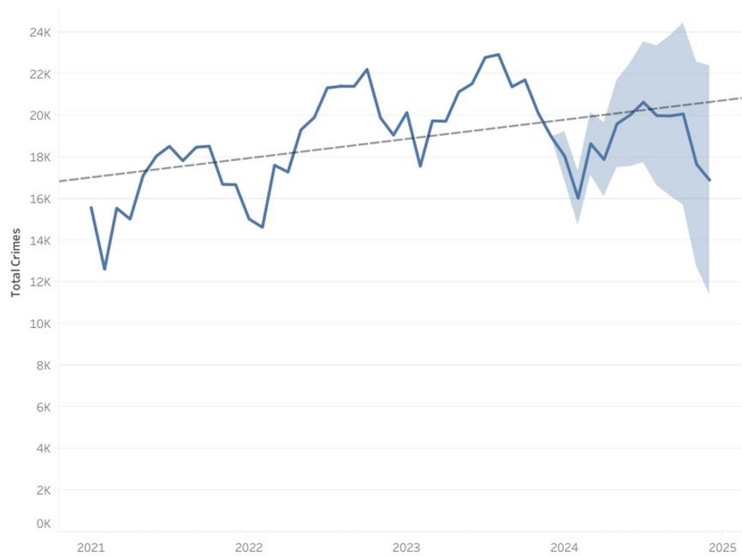
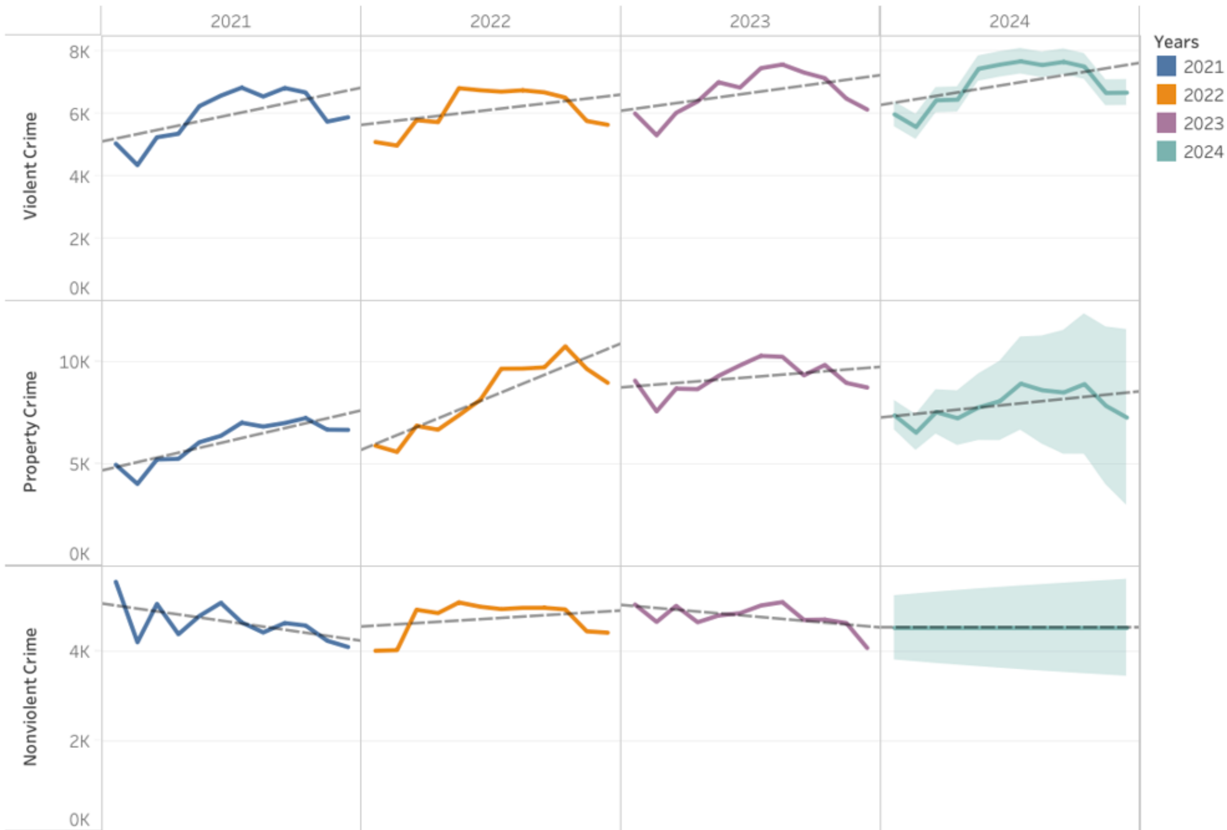


Figure 13 further explored these predictions for each categorized crime and discovered that violent crimes had the most consistent trend compared to the previous years with the smallest margin of error throughout the year 2024. However, we discovered that the predicted number of nonviolent crimes was the same throughout the year with a relatively wide margin of error. This may be an indication that more data needs to be used to provide more accurate predictions for future nonviolent crime rates.

Figure 13. Crimes Per Year Prediction for 2024 for Each Categorized Crime Type (Using Tableau Forecasting)

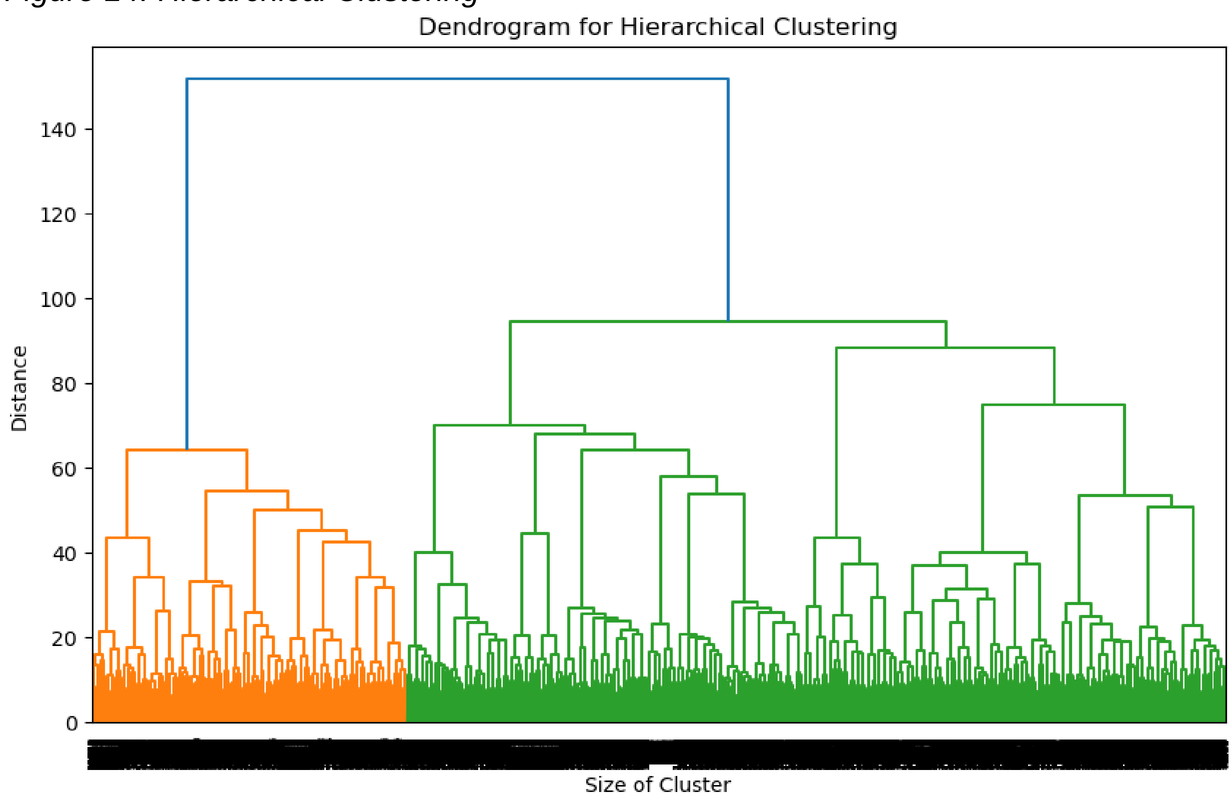


UNSUPERVISED METHODS

The intent of this section is to describe three unsupervised methods utilized in our **segmentation** models. The goal of these models is to understand the combination of factors that help identify crime profile clusters that share similar characteristics. Armed with this information Chicago officials, police, and residents can direct resources and develop multi-faceted programs that will address a combination of characteristics that are associated with crime. For segmentation, Team 51 examined K-means, clustering, and PCA models. Additional data preparation for these models consisted of scaling the dataset using MinMaxScaler. This type data transformation linearly scales the data between zero and one. This is a necessary step for unsupervised learning methods in order to have certainty that the distance calculations for the K-means and clustering models are not affected by any biases from how the variables are scaled. This ensures that the results will be accurate and reliable as well. As we have explored these unsupervised learning methods, we discovered that due to compute power limitations, we are only able to explore one month's worth of data. For the purposes of providing insightful results, we have focused on exploring the month of January in 2023. Both the K-means and PCA models provided insight as to characteristics that may be indicative of crime that were less obvious in the supervised models (environmental complaints). Details on each method and results follow.

Hierarchical Clustering: A dendrogram was created using agglomerative hierarchical clustering, employing a bottom-up approach (Figure 14). This entails initiating the process with each case considered as its own cluster, resulting in a total of N clusters. Subsequently, by employing a similarity measure such as Euclidean distance, the two closest clusters are grouped together, leading to an N-1 cluster solution. This procedure is then iteratively repeated until all observations form a single cluster. It is important to note that all cluster solutions are nested within the dendrogram.

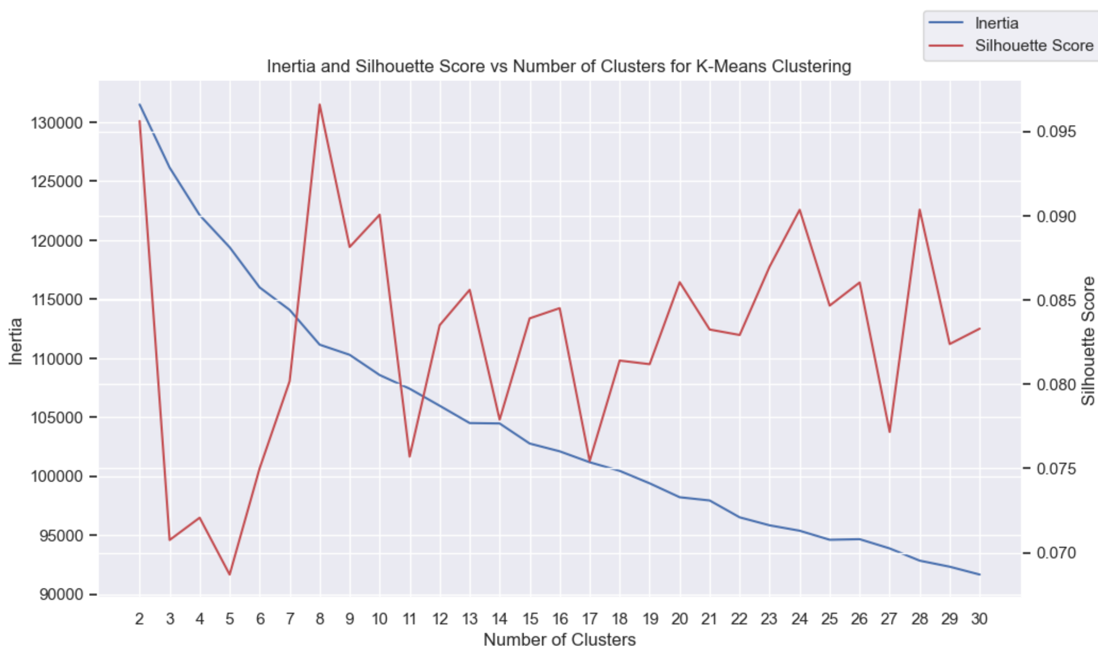
Figure 14. Hierarchical Clustering



Recalling that larger distances between two links indicate greater differences in terms of features, it becomes evident that the green and orange clusters are dissimilar. The goal here is to visually represent the structure of hierarchical clusterings, and as observed, the dendrogram effectively captures these relationships. Based on the outcome of the clustering, we identified **two** distinct clusters. However, we wanted to further explore the number of clusters we can use, as two clusters would not provide enough information to make meaningful distinction between the different clusters.

K-Means: Figure 15 represents the inertia and silhouette scores based on the number of clusters used in our normalized dataset. Inertia is used to measure how well a dataset is clustered when using K-means. This is calculated as the sum of squared distances between each data point and the centroid of its assigned cluster. The optimal number of clusters is based on the “elbow” point of the line for each inertia score. The silhouette score is used to measure how similar a data point is to its own cluster in comparison to other clusters. The higher a silhouette score for a cluster, the more distinct each cluster will be from one another. Based on the results from Figure 8, the optimal number of clusters to use is 8 clusters, as this has the highest silhouette score and has a distinct “elbow” in the line for the inertia scores.

Figure 15. Inertia and Silhouette Score vs Number of Clusters



Our findings indicate that an increase in the number of clusters generally results in decreased inertia, as smaller clusters lead to diminished within-cluster distances.

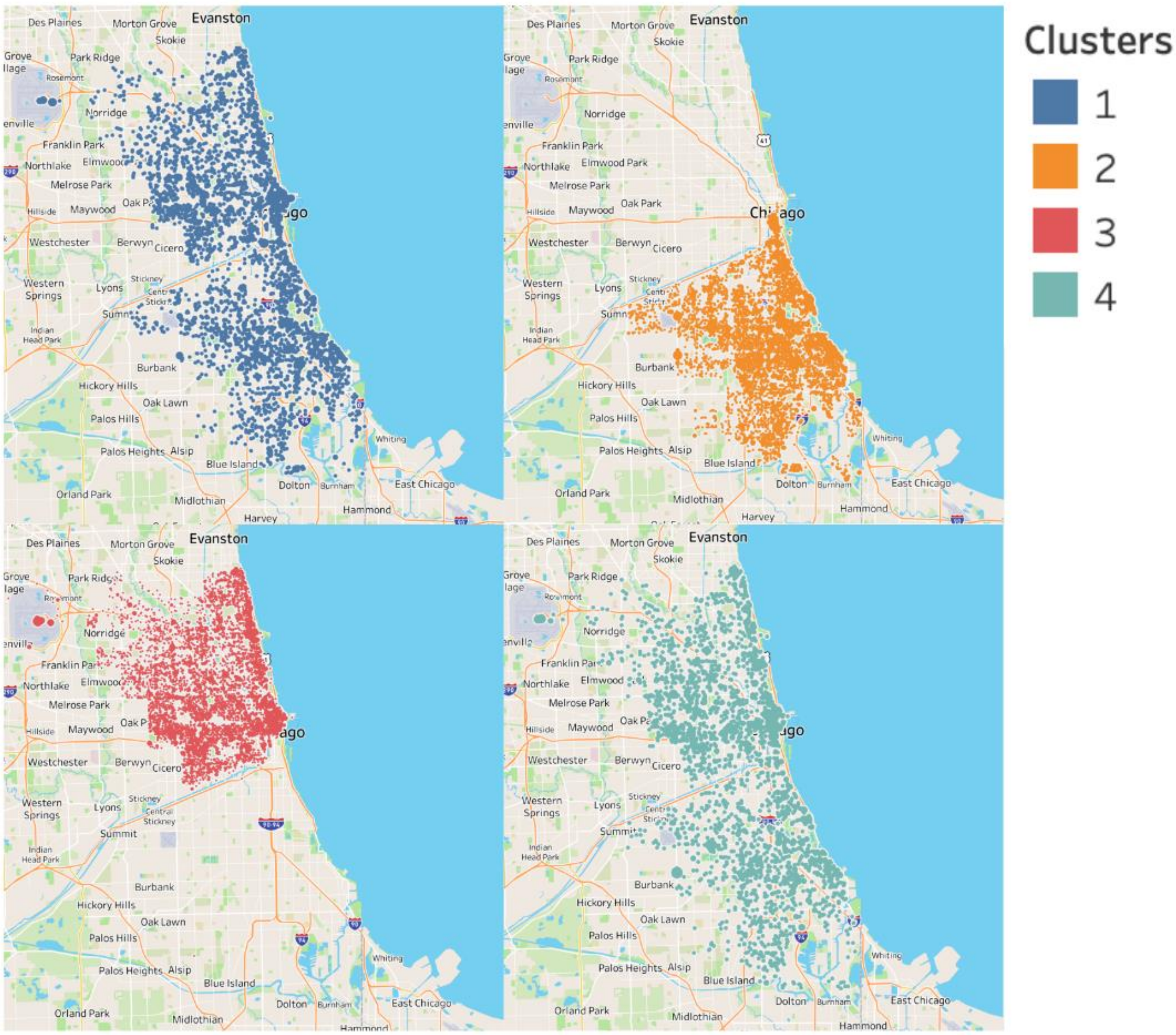
Silhouette scores nearing 1 indicate well-separated clusters, while scores close to 0 suggest overlapping clusters. While this point often serves as a good estimate for the optimal number of clusters, we will adjust this number by accruing domain knowledge to enhance readability and scalability of the model. While 8 clusters are ideal, we started with 4 clusters to more easily gather insights and keep it to a manageable amount for research (Table 7).

Table 7: Crime Profile Clusters and Characteristics

Cluster	Characteristics	Cluster Size
Cluster 1	• Crime occurred primarily on **Sundays** • Located across city but slightly heavier on the East side	3,575
Cluster 2	• Located on the **South Side** • Crime occurred primarily on **weekdays**	7,016
Cluster 3	• Located on the **North Side** • Crime occurred primarily on **weekdays** • Highest level of requests for **lead tests** • Highest level of **snow related complaints**	7,625
Cluster 4	• Crime occurred primarily on **Saturdays** • Located across city but slightly heavier on the East side	2,844

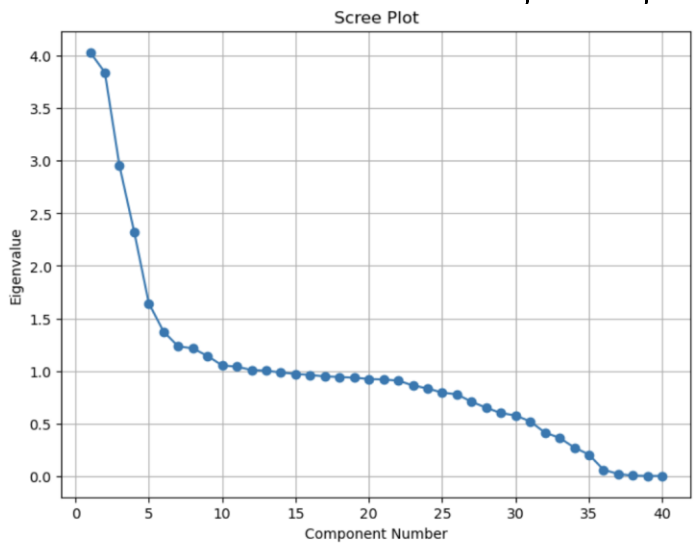
Armed with the crime profile clusters, we placed them on a map of Chicago (Figure 16). What we see is cluster 2 is located on the South side and Cluster 3 on the North side. These are the most distinctly grouped clusters by location. Cluster 1 and 4 are across the city but do tend to be more heavily clustered on the east side.

Figure 16. Map of Chicago Based on Crime Profile Clusters



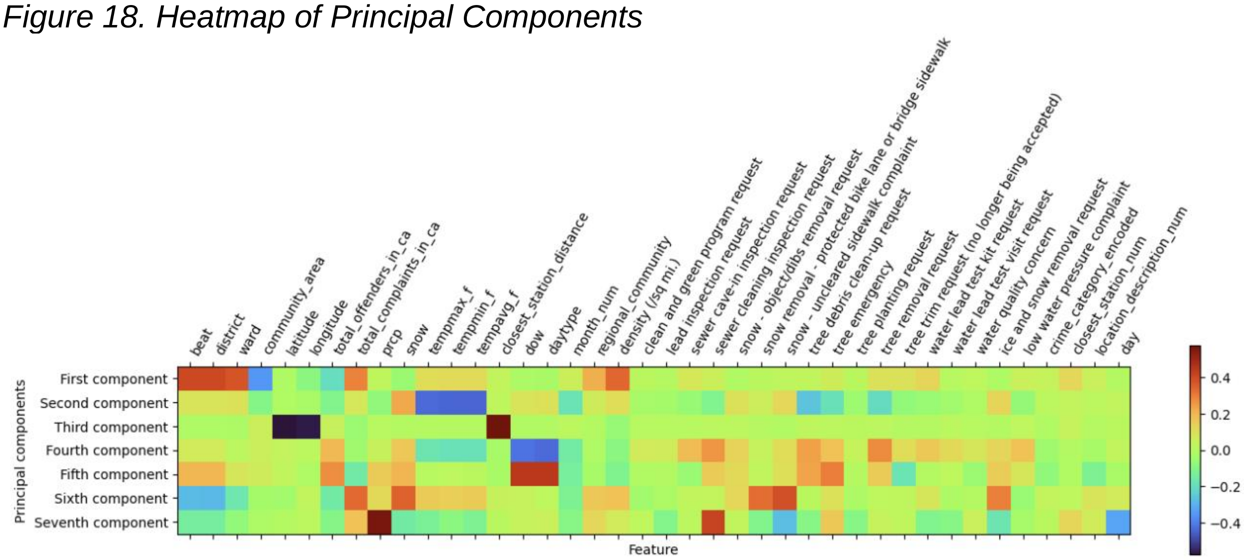
PCA: Principal Component Analysis was employed to identify underlying patterns and trends in the data by emphasizing the principal components that contribute the most to the variance in the dataset. For implementation purposes, a scree plot was utilized to ascertain the number of principal components (PCs) to retain and to discern the point at which adding more components did not significantly contribute to explaining additional variance (Figure 17).

Figure 17. Scree Plot to Determine the Number of Principal Components (PCs)



We used the “elbow method” to determine the optimal number of principal components to retain, which in our case was seven. Additionally, before the elbow, the eigenvalues drop steeply, and after the elbow, the decline is less pronounced. Moreover, we also utilized a heatmap (Figure 18) to get a visual summary of how each variable contributes to the overall structure captured by the principal components. As depicted in the figure, we can see positive loadings are represented by shades of red, and negative loadings are represented by shades of blue. For example, it is clear that there is a positive correlation between the first component and the variables beat, district, and ward. This essentially means that these variables contribute most to the principal component. Additionally, it is indicated that there is a strong negative correlation between the first component and the variable "community area."

Figure 18. Heatmap of Principal Components



We retained the PC scores to understand the projection of a data point onto the principal components (Table 8). The scores provide a way to represent the original data in a reduced-dimensional space, capturing the most important patterns in the data. Each observation's score for a particular principal component can be associated with the value of a variable. A positive correlation is that as the variable's value increases, the corresponding observation's score on the principal component also tends to increase. On the other hand, a negative correlation occurs if the variable's value increases, the corresponding observation's score on the principal component tends to decrease. As we can see from Table 8, there is quite a bit of variety in the associated variables and one interesting observation is that PC4 and PC5 are identical. This indicates that the day of the week, as well as whether it's the weekend or a weekday, significantly influences whether a crime is committed.

Table 8. Variable Association with Each Component and Their Variable Score

Principal Component	Associated Variable with score
PC1 (Community Location)	• Beat: 0.4015 • District: 0.4016 • Ward: 0.3748 • Community Area: -0.3525 • Density (sq/mi): 0.3376
PC2 (Temperature)	• Tempmax_f: -0.4362 • Tempmin_f: -0.4418 • Tempavg_f: -0.4447
PC3 (Geographic Location)	• Latitude: -0.5761 • Longitude: -0.5657 • Closest station distance: 0.5753
PC4 (Day/Day type)	• Day of week: -0.4219 • Day type: -0.4351
PC5 (Day/Day type)	• Day of week: 0.4464 • Day type: 0.4473
PC6 (Environmental Complaint)	• Total Complaints in Chicago: 0.3326 • Snow: 0.3488

	• Snow removal - protected bike lane or bridge sidewalk: 0.3244 • Snow-uncleared sidewalk complaint: 0.3820
PC7 (Precip and Sewer)	• Precipitation: 0.5623 • Sewer Cleaning inspection request: 0.4140 • Day: -0.3262

The utilization of PCA throughout our investigative process, as illustrated through the scree plot, elbow method, heatmap analysis, and examination of PC scores, has been useful in distilling intricate patterns and relationships within the crime data. Specifically, through PCA, variables related to community location, weather, geographic location, day, and environmental complaints are associated with crime occurrences. By reducing dimensionality and emphasizing key components, PCA not only enhanced the interpretability of the data but also provided valuable insights into the variables contributing most significantly to crime occurrences.

Information Access

DASHBOARDS

As an integral part of Team 51's goal to make information from this project visible and accessible, several informative dashboards have been developed and delivered. The "Chicago Crime Story" [9] dashboard provides visibility to crime insights at a glance using 2023 data. The dashboard provides both summary information as well as drill-down detail information for more granular community area analysis by community, crime type, and month. Using the dashboard is intuitive and it requires no special skills. Figure 19 shows a dashboard visualization of a map of Chicago broken out by community with additional stats based on population. Figure 20 shows a line graph visualization of the three crime types and the number of crimes by month. Additional visualizations in "Chicago Crime Story" include heatmaps of crime, bar graphs with trend lines of crime counts, sexual offenders by the community, and a table of Crimes per 1000 people shown in Appendix E Figure E1.

Two additional dashboards were also created called "Chicago Crime - Predictions" [10] and "Chicago Crime - Tableau Forecasting"[11]. The predictions dashboard as shown in Appendix E, Figure E2 shows the performance of our Random Forest model by crime type and community level with line graphs. The dashboard is used to monitor the performance of that model. The Tableau Forecasting dashboard shows the models' predictions for crime in 2024 by community and crime type as shown in Appendix E, Figure E3. The forecasting dashboard will help show which way crime is trending for 2024 in Chicago overall and in each community.

[9] https://public.tableau.com/app/profile/matthew.riegsecker/viz/shared/9469FJ5MB
[10] https://public.tableau.com/app/profile/matthew.riegsecker/viz/ChicagoCrime-Predictions/CrimePredictions
[11] https://public.tableau.com/app/profile/matthew.riegsecker/viz/ChicagoCrime-TableauForecasting/CrimePredictions-TableauForcast

Figure 19. 2023 Chicago Crime per 1000 People – Dashboard View

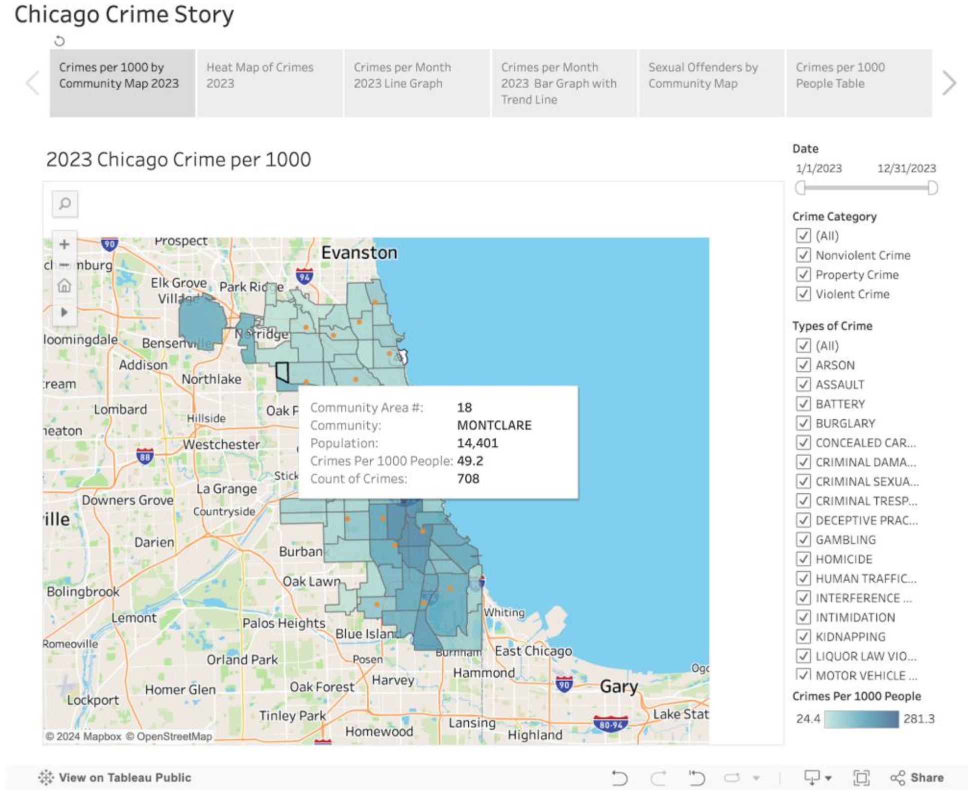
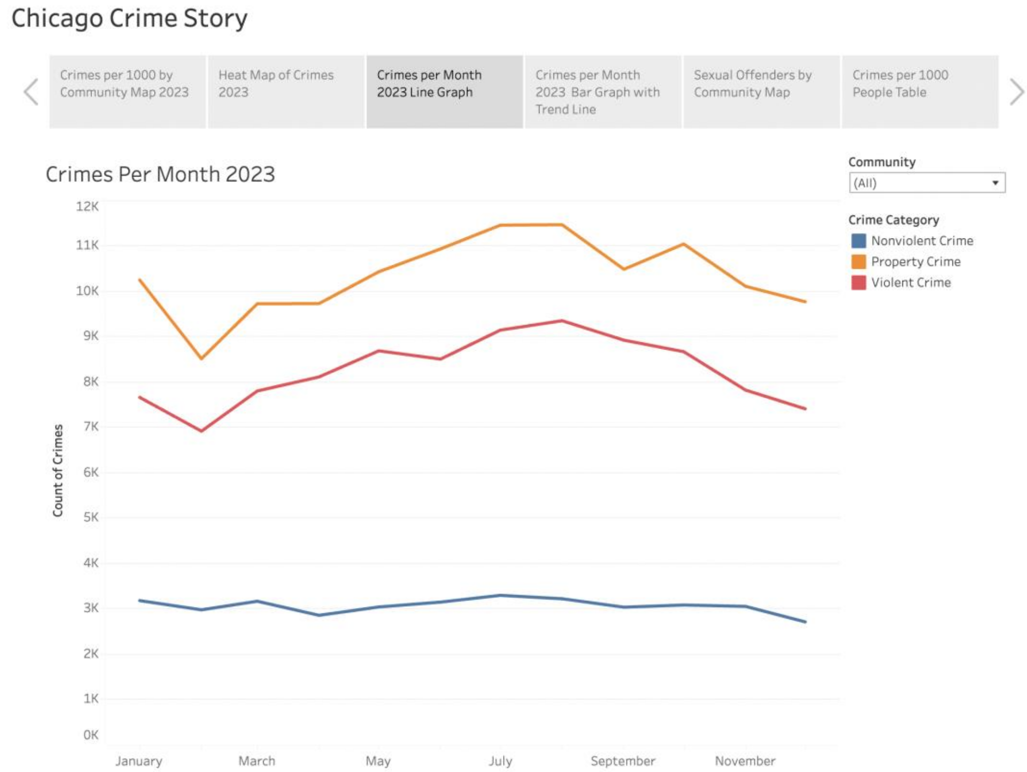


Figure 20. Crimes per Month 2023 – Dashboard View



MOBILE APP

The Mobile Application provides Chicago crime information that is available to Chicago officials, police, and residents anywhere and anytime. The mobile app has three main sections which are Crime by Community, Chicago Overall Stats, and a Police Map. The app provides historical crime data for 2023 at the community and city levels and uses the Tableau Forecasting models to predict overall crime trends at the city level. (Figure 21 and Figure 22). Due to a lack of crime prediction accuracy at the community level, it was not added to the community section at this time. Future updates can add this if given additional time to produce more accurate models to predict crime.   Refer to Appendix F for screenshots of the police map part of the application where you can get directions and phone numbers to the local police stations in the area.

Figure 21. Mobile App Crime by Community

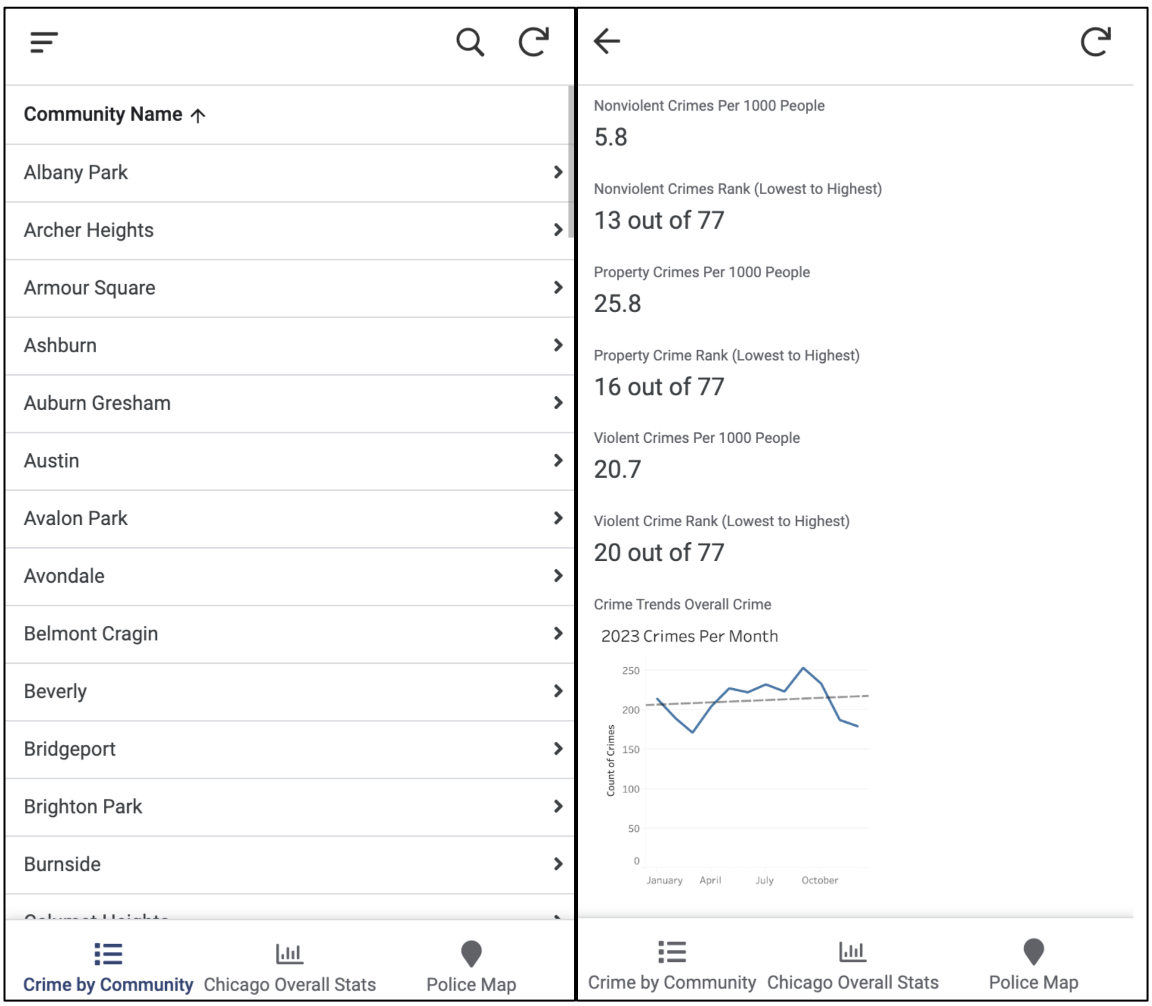
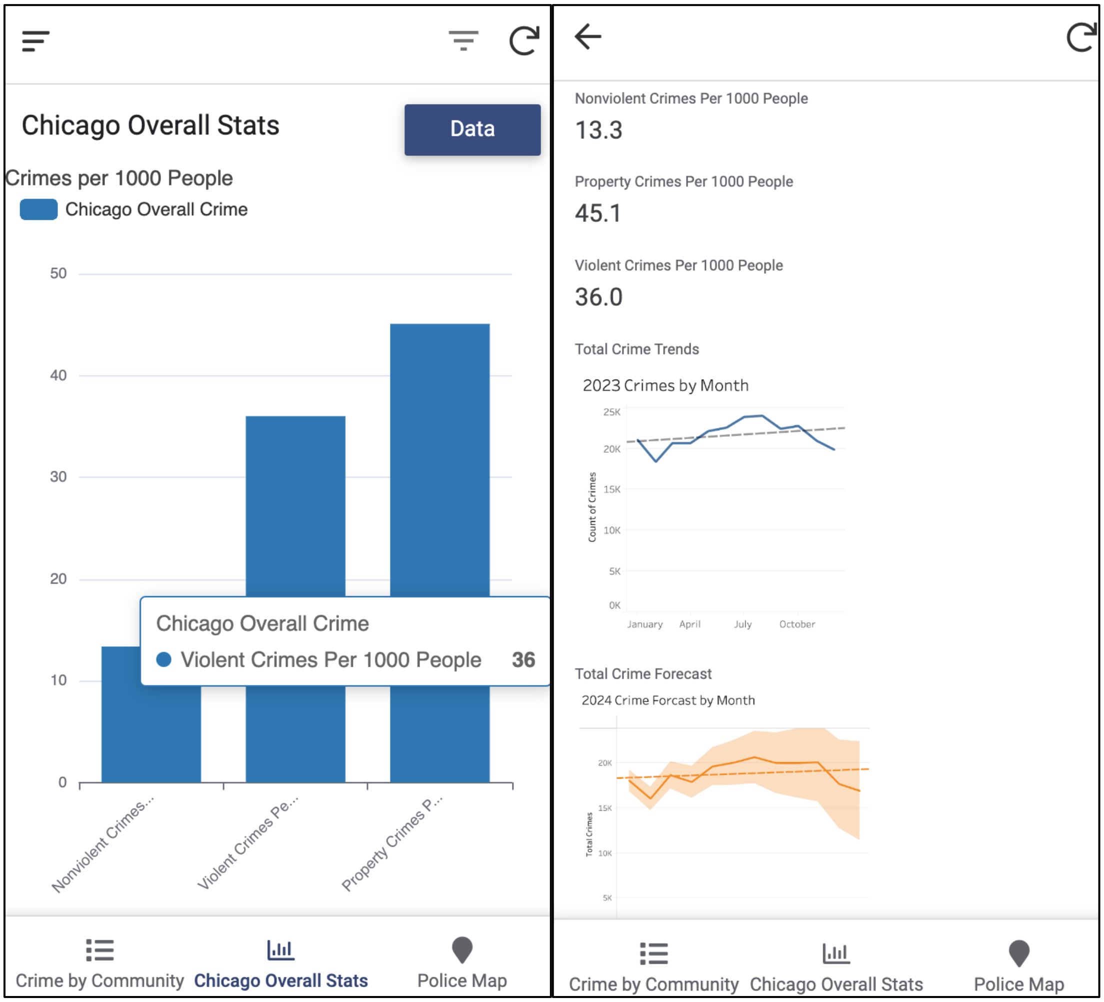


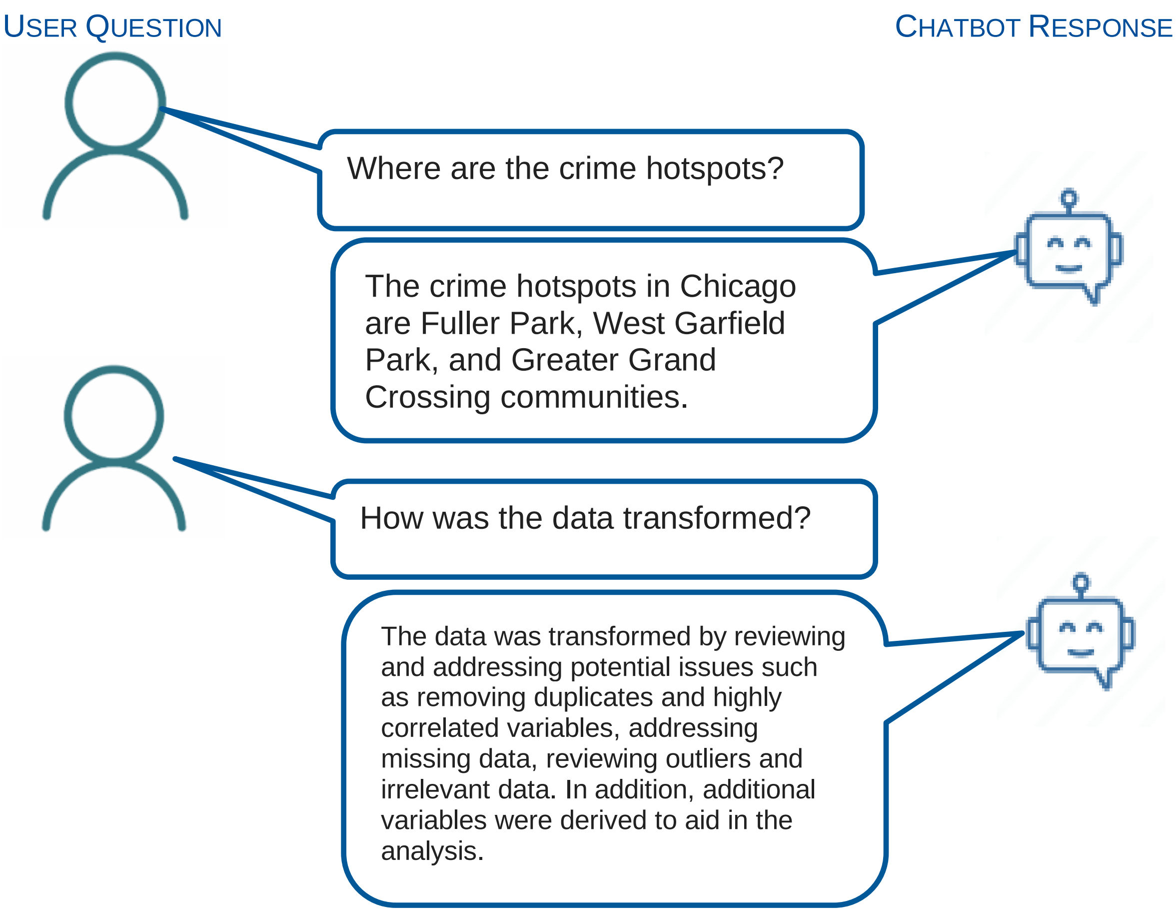
Figure 22. Mobile App Chicago Overall Stats



CHATBOT

A custom Chicago crime chatbot has been developed and delivered which provides quick access to all of the information from this project. The chatbot was developed using Python and the power of ChatGPT through an interface with Open AI's ChatGPT 3.5; it was trained and customized using text from project materials. Twelve test questions were given to the chatbot, and it performed very well (Figure 23). Additional actual chatbot Q&A can be found in Appendix G.

Figure 23. Example Chatbot Q&A for the Chicago Crime Project



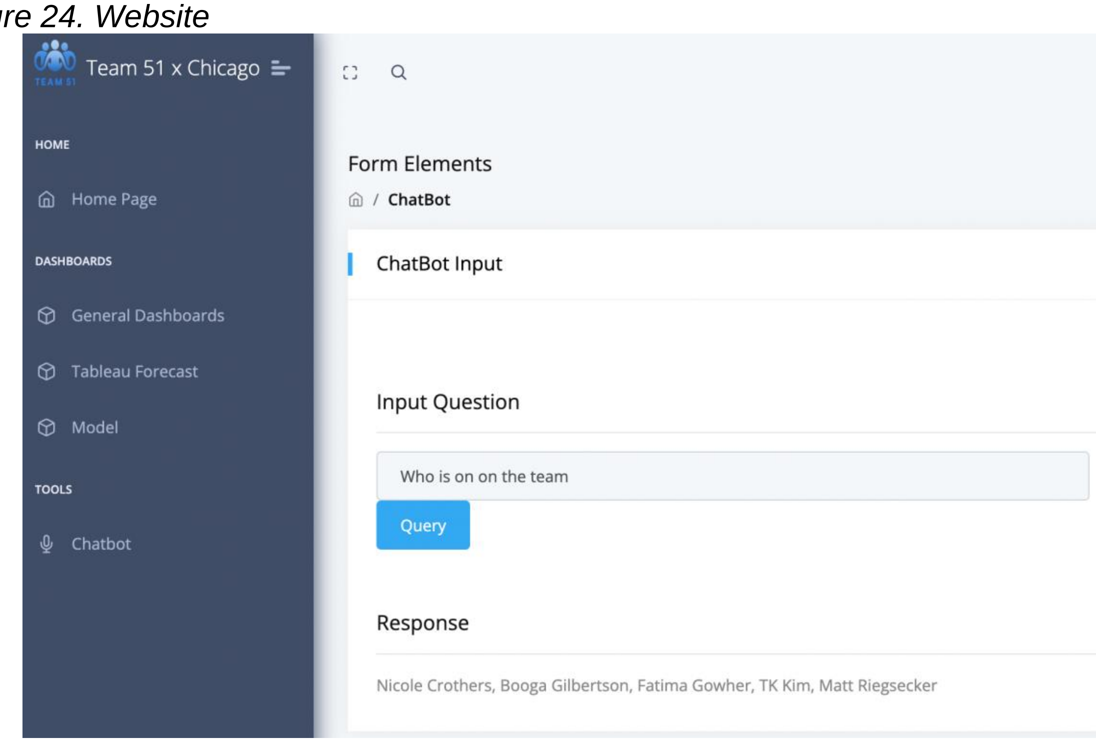
WEBSITE[12]

A website (Figure 24) was built to pull together the dashboards and the chatbot into one easy-to-access location where the data can be monitored and analyzed. The website has three main sections. The homepage has high-level stats and team information and is where app usage and engagement metrics could be added in the future. Some placeholder information was added to the homepage to show potential uses. The dashboard section houses the “Chicago Crime Story”, “Chicago Crime - Tableau Forecasting”, and “Chicago Crime - Predictions” dashboards. And there is a section where the Chatbot is hosted as shown in Figure 24.

This website is also easily editable and expandable. Load balancing is done through Google’s Google Front Ends which provides hands-free scalability. Web pages are rendered through Flask and templates with examples on how to construct various visualizations or widgets are included in the code base. Docker provides containerization enabling rapid deployment of additional websites or fast transfers to other cloud service providers.

The website is also updated through GitHub + Google Cloud Build as a continuous integration/continuous development workflow. As soon as changes are pushed to the repository, Google will automatically rebuild the application to incorporate the latest updates. Google’s process also means 100% uptime through updates as the load balancer will only send traffic to the new page once the build is fully complete.

Figure 24. Website



12 https://team51-v7s24t2e3q-uc.a.run.app/

Conclusions

To support Chicago in achieving its crime prevention objectives, Team 51 has delivered on the four robust goals identified at the beginning of this project. Team 51's data-backed analysis has demonstrated an understanding of the characteristics of crime in Chicago. The top factors that predict crime have been identified and models that predict crime type and crime rate trends have been developed. Segmentation analysis identified a combination of factors that are associated with crime, and four information access tools have been developed that put crime data in the hands of Chicago officials, police, and residents. Throughout Team 51's work, we have adhered to our commitment to use publicly available data and transparent "glass box" analytic methods.

Understand Crime	• Crime activity is dominated by property and violent crime. • The top three crime hotspots are the community areas of Fuller Park, West Garfield Park, and Greater Grand Crossing. • Higher crime rates correspond with higher temperatures in the summer months. • Top factors that predict the type of crime are latitude, longitude, distance to the closest police station, temperature, crime location description (such as "street" or "apartment"), followed by a combination of location, weather, and environmental factors. • Population density and the number of sex offenders in a community area are the top two predictive factors for crime rates. • Environmental complaints regarding lead, snow removal, tree removal, and sewer cleaning are also associated with crime occurrences.
Predict Crime	Predictive modeling using supervised methods has shown that a Gradient Boosting model is the most accurate overall for predicting crime; however, the Random Forest model is nearly as accurate, is easier to explain, and has more useful predictive variables. Therefore, **Random Forest is the selected model** for predicting both crime type and crime rate. Additionally, a simplified Tableau exponential smoothing model was also used to provide indicative 2024 forecasts of crime count trends by crime type.

Segmentation	Segmentation modeling using unsupervised methods has shown that the K-means model identified four crime profile clusters which highlight a combination of factors that influence crime. Two of the clusters are clearly marked by location in the South and North. One of the more interesting insights was that cluster three had high environmental complaints around lead and snow removal requests. Other groupings appear to be around days of the week such as crimes that were committed on Saturday and Sunday. The PCA model identified seven components associated with crime and generally relate to community location, temperature, geographic location, day of week, and environmental complaints around snow removal and sewer cleaning. It is noteworthy that unsupervised modeling picked up an association between environmental complaints and crime occurrences that supervised modeling did not emphasize. This might indicate a link between potentially underserved/neglected parts of the city and crime.
Information Access	Through development of dashboards, a mobile application, a chatbot, and a website, historical and near real-time crime information is available for the entire city of Chicago by community area. Additionally, crime prediction trends by type of crime (violent, non-violent, and property) for each community area are also available. The chatbot provides efficient access to all of the information contained in the Chicago crime project. A website provides easy access to the dashboards and chatbot where the data can be monitored and analyzed.

Market Expansion: Leveraging the Chicago Crime Project in New Markets

Team 51 has demonstrated its commitment to supporting the City of Chicago's success in reducing crime through use of our innovative, state-of-the-art "glass-box" machine learning and data science methods, along with our dashboard, mobile app, chatbot and website. A value-added benefit from Team 51's Chicago Crime project is that it is transferable and scalable to other metropolitan areas such as New York City, Los Angeles and other major cities. This makes us an ideal partner in empowering safer communities in other large cities.

Team 51's models have been developed and configured in anticipation of utilizing other metropolitan area data sets with minimal customization. This would be an immediately marketable offering for other cities with a favorable profit margin. Expanding our state-of-the-art, data-driven solutions that "Empower Safer Communities Through Data-Driven Crime Insights" presents an exciting opportunity to reduce crime in other areas. Team 51 has already assessed publicly available data for Los Angeles and New York City. **We recommend that we engage Los Angeles and New York City officials to present and market our solution** that will equip police, businesses, and residents with the tools they need to proactively address and mitigate crime. At the same time, given that Team 51 has already invested in the analytic, data-science, dashboard, mobile app, chatbot, and website infrastructure to support the Chicago crime project, **we expect new projects will deliver a substantial profit for our company while still being of incredible value for future clients.**

x TEAM 51

NEW YORK CITY: EMPOWERING A SAFER COMMUNITY THROUGH DATA-DRIVEN CRIME INSIGHTS

x TEAM 51

LOS ANGELES: EMPOWERING A SAFER COMMUNITY THROUGH DATA-DRIVEN CRIME INSIGHTS

Recommendations

Predicting crime is complex. Team 51 has explored and extensively analyzed numerous data-backed insights to crime in Chicago using machine-learning and other state-of-the-art data analysis techniques. Our analysis reveals that a combination of factors drive crime; therefore, multi-faceted efforts that leverage the most important crime drivers will have the greatest impact in helping Chicago meet crime reduction objectives. As a result of insights gained from the Chicago crime project, Team 51 developed three broad areas of recommendations: (1) for the City of Chicago, (2) for future project phases and model enhancements, and (3) for market expansion.

1. Recommendations for the City of Chicago:

a. **Evaluate and align crime response resources in the crime hotspot community areas** to address staffing, shift design and police support with crime activity and volume. Geographic location (latitude and longitude) is the top feature that is the most predictive of the type of crime, while population density and sex offender numbers are the most predictive for crime counts.

b. **Evaluate and promote crime awareness and prevention strategies targeted to of the top factors associated with crime:**

- **Geographic location** - crime hotspot community areas
- **Weather** – warmer summer months
- **Crime location** – streets and apartments

All of these factors are higher predictors of crime. Supporting awareness, police and community resources, and other prevention strategies that target these factors will have a positive impact on crime. Additionally, greater awareness by citizens alerts them to be mindful of these factors with the goal of preventing crime from happening in the first place.

c. **Deploy and promote information access tools** for Chicago officials, police, and residents. These tools provide data availability and promote data and information transparency.

- A convenient **website** offers easy access to the "Chicago Crime story" **dashboards** and **chatbo**t. The **dashboards** provide access and easy to understand data visibility to crime information by community area. Increasing awareness of crime information will help keep Chicago attuned to crime trends and changes. The **chatbot** provides easy access and reference to information contained in the Chicago Crime project.

- The **mobile app** puts historical crime data for 2023 at the community and city levels in the hands of Chicago officials, police, and residents. It also uses the Tableau Forecasting models to predict overall crime trends at the city level.

d. **Review mitigation of 311 environmental complaints**, especially related to requests for lead tests, snow removal, tree clean-up/removal, and sewer cleaning to assure that potential crime factors are also considered and addressed along with mitigation of the 311 complaints.

2. RECOMMENDATIONS FOR FUTURE PROJECT PHASES AND MODEL ENHANCEMENTS:

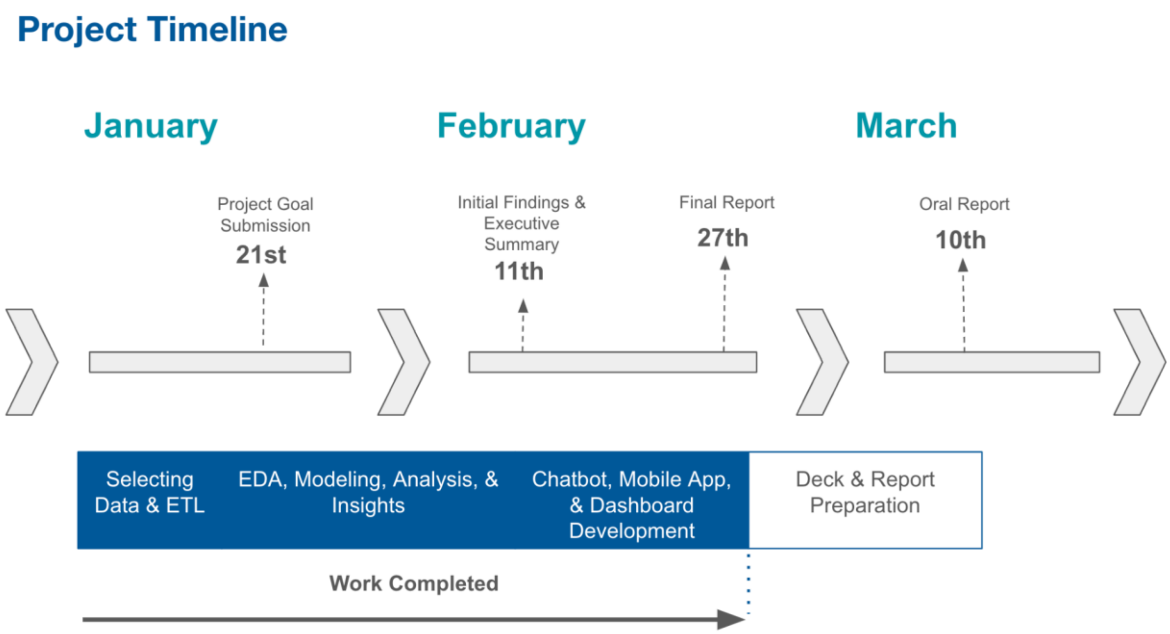
As more data is gathered along with resultant insights about Chicago crime mitigation strategies, the models and information access tools should be maintained and augmented with the newest machine-learning enhancements when appropriate.

3. RECOMMENDATIONS FOR FUTURE MARKET EXPANSION:

Team 51's innovations and investment from the Chicago project are scalable to other markets with minimal additional investment. Similar offerings should be marketed by our company to other large metropolitan areas in order to increase revenue from the intellectual property that has already been developed while offering future clients innovative solutions at a great value. **We recommend that we engage Los Angeles and New York City officials to present and market our solution** that will equip police, businesses, and residents with the tools they need to proactively address and mitigate crime

Project Status

Team 51’s project is **on track** for the final Oral Presentation deliverable as planned.



Completed	Next Steps
• Final Report, Executive Summary, and all precursor activities	• Prepare and present Oral Report

REFERENCES

Dabrowski, Ted and John Klingner. 2024. “Chicago led nation in homicides for 12th year in a row in 2023, murder rate still 5 times higher than NYC’s”. Wirepoints. Jan 3. https://wirepoints.org/chicago-led-nation-in-homicides-for-12th-year-in-a-row-in-2023-murder-rate-still-5-times-higher-than-nycs-wirepoints/

Ponce, Anthony. 2023. “Chicago police superintendent lays out 2024 goals for department”. Fox News Chicago. Dec 11. (fox32chicago.com)https://www.fox32chicago.com/news/chicago-police-superintendent-lays-out-2024-goals-for-department

Ramos, Elliot. 2023. “Here's what's happening with crime in Chicago in 2023 - CBS Chicago”. (cbsnews.com). Dec 29. https://www.cbsnews.com/chicago/news/heres-whats-happening-with-crime-in-chicago-in-2023/

Rivera, Mark. 2023. “Chicago crime: Mayor Brandon Johnson unveils 'People's Plan for Community Safety' to address city's crime problem”. ABC7 Chicago. Dec 14. https://abc7chicago.com/chicago-crime-mayor-brandon-johnson-peoples-plan-for-community-safety-police-department/14183500/.

Wolff, Kevin and Christopher Thomas. 2023. “How Volatile Temperatures Shape Violent Crime”. CUNY Graduate Center. Sep 21. https://www.gc.cuny.edu/news/how-volatile-temperatures-shape-violent-crime.

Crime Type Prediction Studies

Alsubayhin, Abdulrahman and Muhammad Sher Ramzan. 2024,” Crime Prediction Model using Three Classification Techniques: Random Forest, Logistic Regression, and LightGBM”. (thesai.org) *International Journal of Advanced Computer Science and Applications,* Vol. 15, No. 1, 2024. https://thesai.org/Downloads/Volume15No1/Paper_23-Crime_Prediction_Model_using_Three_Classification_Techniques.pdf
Uses a kaggle crime dataset with 9 features for San Francisco to predict crime type. Random Forest accuracy of 43%.

Khan, Muzammil, Azmat Ali, Yasser Alharbi. 2022. "Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques", Complexity, vol. 2022. https://doi.org/10.1155/2022/4830411
Uses a kaggle crime dataset with 9 features for San Francisco to predict crime type for the top 10 out of 39 crimes. Random Forest accuracy of 64%.

Kim, Suhong , Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. 2020. “Crime-Analysis-Through-Machine-Learning”. (researchgate.net) https://www.researchgate.net/profile/Pooya-

Taheri/publication/330475412_Crime_Analysis_Through_Machine_Learning/links/5c467bba92851c22a386ffbd/Crime-Analysis-Through-Machine-Learning.pdf
Uses an open portal dataset from Vancouver BC with 11 features to predict crime type. Boosted Tree Accuracy of 44%.

Malhotra, Vaibhav. 2020. “Chicago-crime-analysis: BigData analytics on Chicago crime dataset”. https://github.com/Vaibhav3M/Chicago-crime-analysis.
Uses the Kaggle Chicago crime dataset to predict crime type according to location. Random Forest Accuracy of 37%.

Tabedzki, Christian, AmruteshThirumalaiswamy, and Paul van Vliet. 2018. “Yo home to Bel-Air: predicting crime on the streets of Philadelphia”. https://www.seas.upenn.edu/~tabedzki/machine-learning-report-final.pdf.
Uses an open portal dataset from Philadelphia with 14 features along with local weather data to predict crime type. Random Forest Accuracy of 27%.

APPENDIX

Appendix A: About Team 51

Nicole Crothers
Data Scientist, Data Exploration, Predictive Models

Booga Gilbertson
Data Scientist, NLP, Communications

Fatima Gower
Data Scientist, Data Exploration, Visualization

TK Kim
Data Scientist, Data Engineer, Cloud Platforms

Matt Riegsecker
Data Scientist, Project Manager, Data Analysis

PARTNERING WITH YOU TO EMPOWER A SAFER COMMUNITY THROUGH DATA-DRIVEN CRIME INSIGHTS.

Appendix B: Data Sources

Affordable Housing Locations https://data.cityofchicago.org/Community-Economic-Development/Affordable-Housing-Units-by-Community-Area/yvj4-y3fb
584 records (not used in the current modeling)

Chicago Fire Station Locations
https://data.cityofchicago.org/Public-Safety/Fire-Stations/28km-gtjn/about_data
93 records (not used in the current modeling)

Chicago Police Station Locations
https://data.cityofchicago.org/Public-Safety/Police-Stations/z8bn-74gv/about_data
23 records

Community Boundaries
https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

Community Populations
cmap.illinois.gov/documents/10180/126764/_Combined_AllCCAs.pdf/
Community areas in Chicago - Wikipedia

Crimes Data - https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data
7.97 million records

Definition and Description of Crime Types in Chicago -
https://gis.chicagopolice.org/pages/crime_details

Environmental Data - https://data.cityofchicago.org/Environment-Sustainable-Development/CDPH-Environmental-Complaints/fypr-ksnz/data_preview
57,598 records

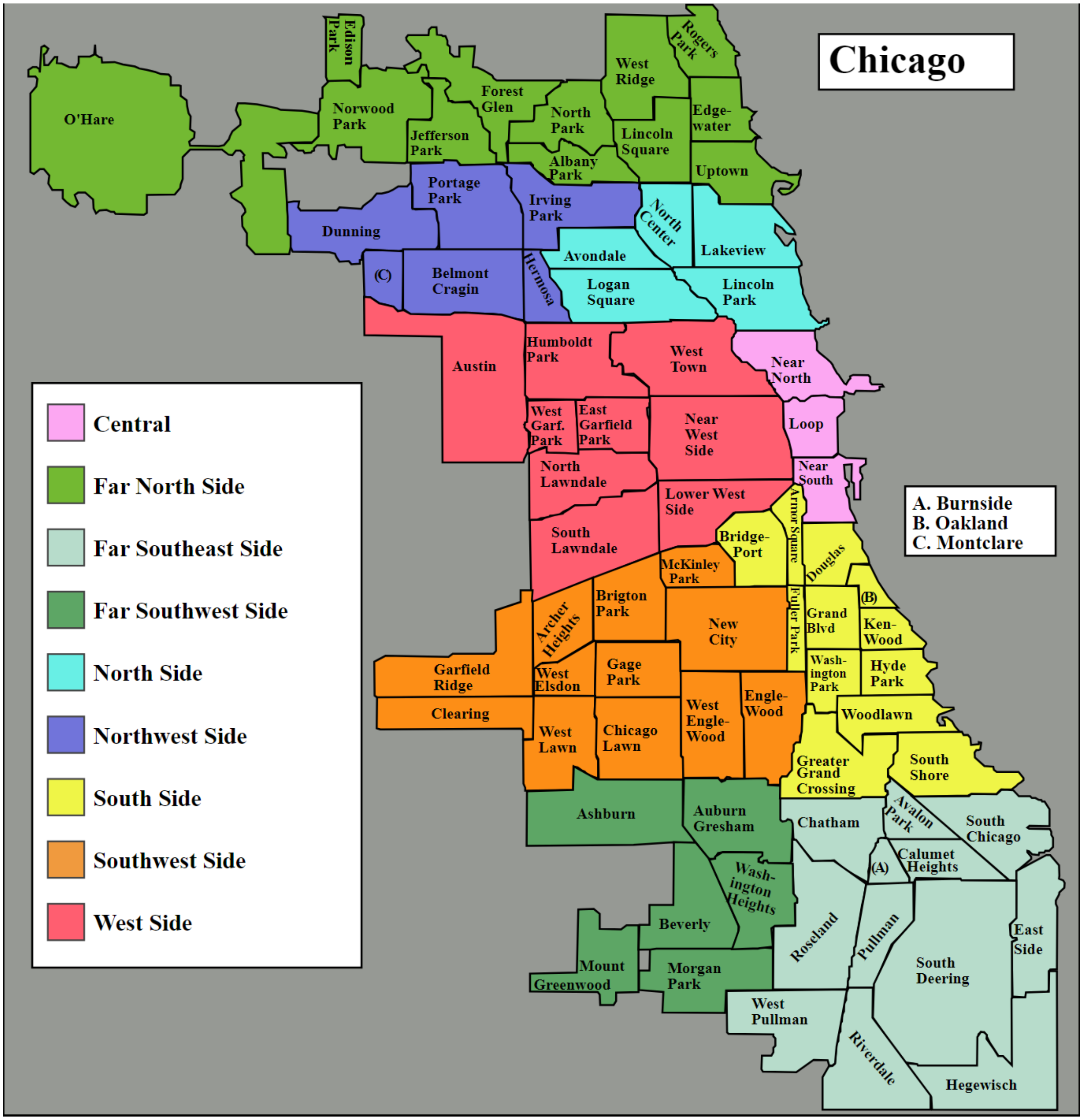
311 Complaints in Chicago
https://data.cityofchicago.org/Service-Requests/311-Service-Requests-Graffiti-Removal-No-Duplicate/8tus-apua/data_preview
5,874,796 records

Sex offenders data - https://data.cityofchicago.org/Public-Safety/Sex-Offenders/vc9r-bqvy/data_preview
336 records

Appendix C: Crimes Categorized by Crime Type

Violent Crimes	Property Crimes	Nonviolent Crimes
Criminal Sexual Assault	Theft	Other Offense
Offense Involving Children	Burglary	Criminal Trespass
Deceptive Practice	Criminal Damage	Weapons Violation
Battery	Motor Vehicle Theft	Stalking
Assault	Arson	Obscenity
Sex Offense		Public Peace Violation
Robbery		Liquor Law Violation
Homicide		Narcotics
Kidnapping		Concealed Carry License Violation
Human Trafficking		Interference with Public Officer
Ritualism		Intimidation
		Prostitution
		Gambling
		Non-Criminal
		Other Narcotic Violation
		Public Indecency

Appendix D: Map of Community Regions and Community Areas



By Peter Fitzgerald - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=3902994

Appendix E: Dashboard Screenshot Examples

Figure E1. 2023 Crime Per 1000 People Table

Community	crime_category	Date January	February	March	April	May	June	July	August	Septemb..	October	November	December
	Property Crime	1.44	1.39	1.44	1.32	1.67	1.37	1.92	1.81	1.58	1.67	2.06	1.51
EAST GARFIELD PARK	Nonviolent Crime	3.50	2.95	2.80	2.70	2.95	3.50	3.00	2.90	3.45	2.85	3.00	2.35
	Violent Crime	5.70	5.90	6.90	6.65	8.25	8.60	8.00	8.40	9.80	8.30	7.50	6.25
	Property Crime	6.30	5.45	6.80	6.75	7.80	8.00	9.00	6.25	6.95	7.25	6.15	6.25
EAST SIDE	Nonviolent Crime	0.69	0.92	0.97	0.46	0.69	1.15	1.98	1.66	1.33	1.20	0.92	0.97
	Violent Crime	1.84	1.66	2.26	2.12	1.98	2.16	2.49	2.30	2.49	1.75	1.80	2.16
	Property Crime	2.12	1.84	1.66	1.43	2.76	2.30	2.76	1.61	3.31	5.16	2.21	1.98
EDGEWATER	Nonviolent Crime	0.43	0.52	0.55	0.39	0.52	0.50	0.53	0.55	0.57	0.52	0.59	0.52
	Violent Crime	1.44	1.44	1.53	1.69	1.65	1.49	2.03	2.01	1.94	1.92	1.46	1.35
	Property Crime	2.75	2.27	2.66	2.24	2.70	2.50	2.97	3.27	3.09	2.58	2.74	2.66
EDISON PARK	Nonviolent Crime	0.52	0.26	0.52	0.35	0.61	0.95	0.35	0.52	0.35	0.17	0.52	0.09
	Violent Crime	0.78	0.78	0.87	0.95	0.95	0.69	0.69	0.69	0.43	0.87	1.65	0.95
	Property Crime	0.61	0.26	0.52	1.39	0.87	0.35	0.78	0.26	1.04	0.87	1.13	0.78
ENGLEWOOD	Nonviolent Crime	2.83	2.71	3.24	2.71	3.04	3.16	3.57	2.87	3.12	2.54	3.57	2.46
	Violent Crime	6.36	6.89	8.13	7.96	8.86	8.17	7.51	8.04	7.88	7.39	6.73	6.81
	Property Crime	7.80	5.50	6.73	6.73	8.13	7.43	7.63	6.11	6.77	6.57	5.09	5.75
FOREST GLEN	Nonviolent Crime	0.26	0.31	0.20	0.20	0.36	0.36	0.56	0.15	0.36	0.26	0.41	0.41
	Violent Crime	0.82	0.61	0.66	0.82	1.12	0.61	0.71	0.71	0.97	0.61	0.46	0.41
	Property Crime	1.68	1.43	1.02	0.87	0.92	1.07	1.02	1.22	1.17	1.33	1.38	1.48
FULLER PARK	Nonviolent Crime	4.29	4.29	3.12	4.67	3.90	3.51	5.45	1.95	5.06	3.12	5.06	3.51
	Property Crime	10.91	6.62	4.67	10.52	7.79	10.13	7.79	12.08	10.13	9.35	9.74	10.91
	Violent Crime	9.35	10.52	10.13	13.25	9.35	11.30	8.18	14.80	9.74	8.18	8.57	9.35
GAGE PARK	Nonviolent Crime	0.71	0.76	0.38	0.96	0.73	0.78	0.73	0.94	0.89	1.04	0.68	0.73
	Property Crime	1.90	2.15	1.80	1.77	1.54	2.23	2.18	1.92	2.30	2.68	2.30	2.58
	Violent Crime	1.57	1.47	1.64	1.69	2.30	2.02	2.23	2.88	2.07	2.18	2.43	2.28

Figure E2. Random Forest Model Crime Prediction Dashboard

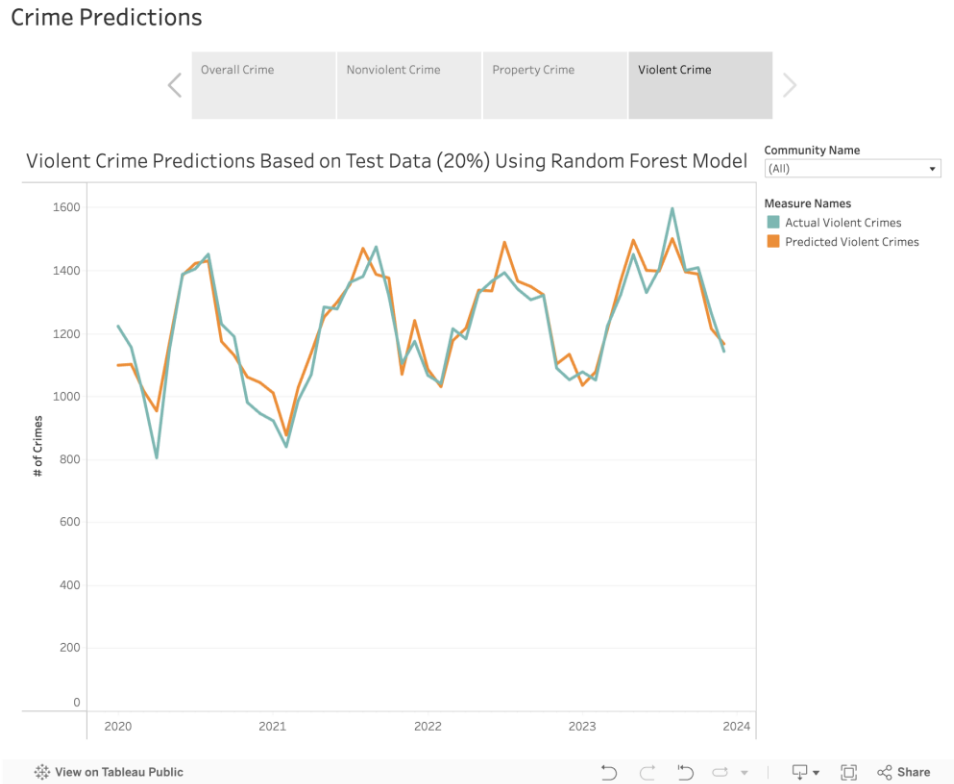
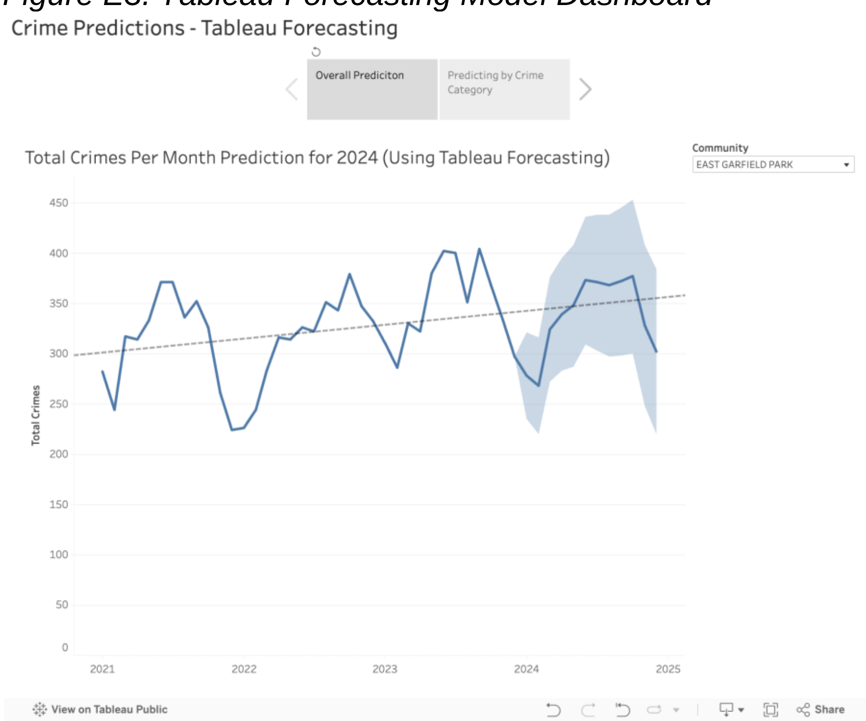
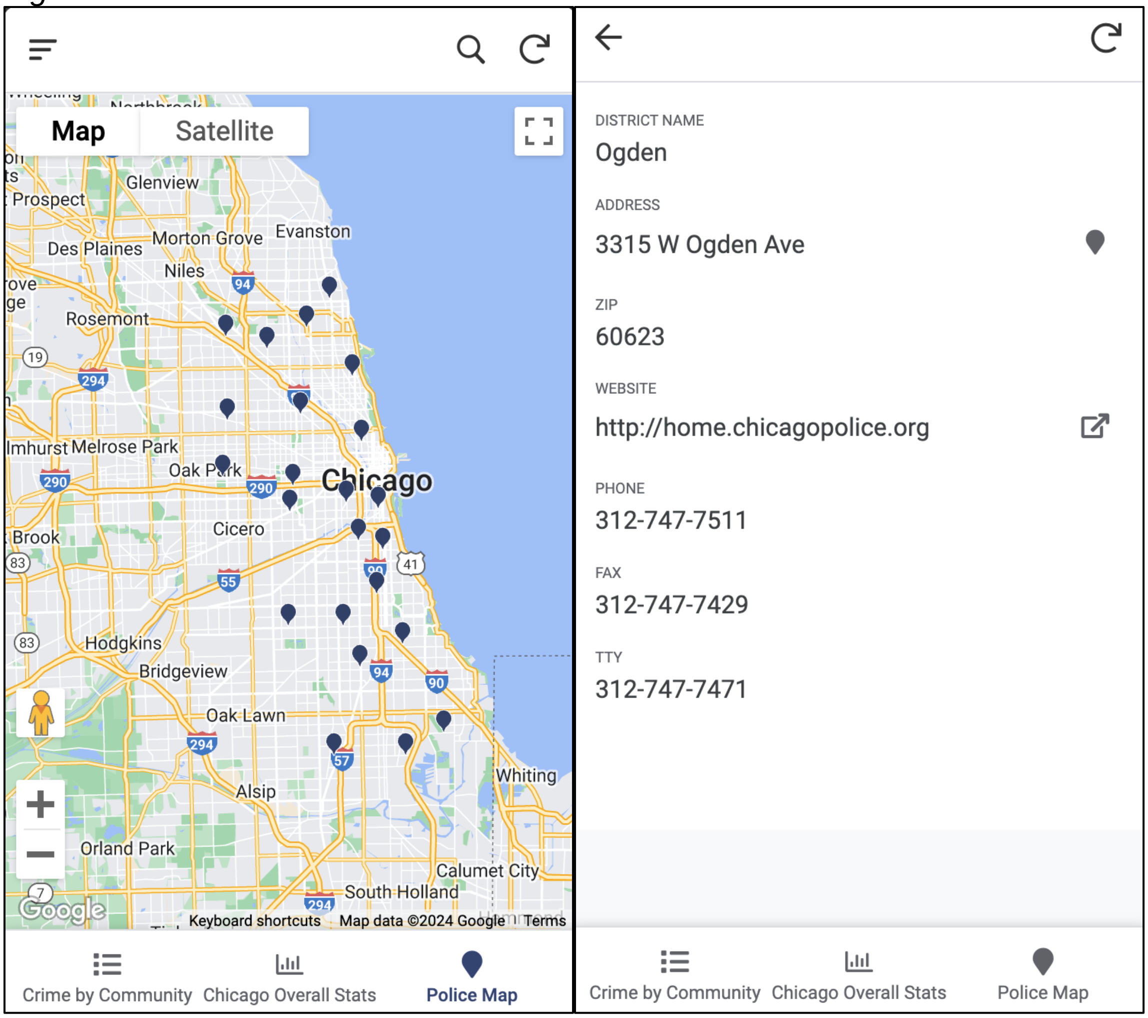


Figure E3. Tableau Forecasting Model Dashboard

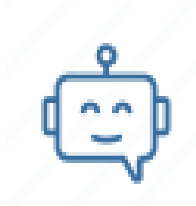


Appendix F: Mobile App Screenshot Examples

Figure F1. Mobile



Appendix G: Actual Chatbot Question and Answer Examples



Question	Chatbot Answer
what are the project deliverables	RESPONSE: The project deliverables include analytics, recommendations, a dashboard, a mobile application, and a chatbot.
what are team 51 project goals	RESPONSE: The team 51 project goals are to understand crime by identifying the top factors that predict crime and identify crime hotspots by Chicago neighborhood, predict crime by developing a model that predicts crime as a function of neighborhood and other factors, segment data sets for deeper insights into the combination of factors that contribute to differences in crime, and develop a dashboard, mobile app, and chatbot that facilitates access to crime information for Chicago officials, police, and residents.
what are the modeling methods	RESPONSE: The modeling methods used in this context consist of both supervised and unsupervised methods.
what supervised methods are used	RESPONSE: The supervised methods used by Team 51 to train predictive models include linear and logistic regression, decision tree, random forest, and gradient boosting.
what unsupervised methods are used	RESPONSE: K-means and Clustering methods are the unsupervised methods used.
what is the recommended model	RESPONSE: The recommended model is the Random Forest model.
why is random forest the recommended model	RESPONSE: Random Forest is the recommended model because it is nearly as accurate as the Gradient Boosting model for predicting crime, but it is easier to explain and has more useful predictive variables. Additionally, the top factors that predict crime for Random Forest, such as latitude, longitude, and distance to the closest police station, are considered to be highly important.
where are the crime hotspots	RESPONSE: The crime hotspots in Chicago are Fuller Park, West Garfield Park, and Greater Grand Crossing communities.

how was the data transformed	RESPONSE: The data was transformed by reviewing and addressing potential issues such as removing duplicates and highly correlated variables, addressing missing data, reviewing outliers and irrelevant data. In addition, additional variables were derived to aid in the analysis.
what is the description of the data	RESPONSE: The data consists of six publicly available datasets primarily obtained from government sources. These datasets include the Crimes dataset, the 311 Complaints dataset, the daily weather data from Chicago's Airport, the Environmental complaints dataset from the Department of Public Health, the Sex Offenders registry, and the location of every police station. The data has been extensively cleaned and transformed into a high-quality master dataset consisting of 250 thousand individual crime records and 61 variables. This master dataset is in a comma-separated-value (.csv) format and represents the calendar year 2023 crime.
who is on the team	RESPONSE: Nicole Crothers, Booga Gilbertson, Fatima Gowher, TK Kim, and Matt Riegsecker are on the team.

Figure G1. Screenshot of chatbot code with question and answer

```
**CHICAGO CRIME CHATBOT: Please ask a question. For example, ask
where are the crime hotspots
To end the session please hit the 'enter' key.**

 where are the crime hotspots
CHATBOT RESPONSE:  The crime hotspots in Chicago are Fuller Park, West Garfield Park, and Greater
Grand Crossing communities.
```

```
**CHICAGO CRIME CHATBOT: Please ask a question. For example, ask
where are the crime hotspots
To end the session please hit the 'enter' key.**

 What is the description of the data
CHATBOT RESPONSE:  The data consists of six publicly available datasets primarily obtained from go
vernment sources. These datasets were extracted, loaded, and transformed into a usable format for
analysis. The master dataset, developed in a comma-separated-value (.csv) format, consists of 250
thousand individual crime records and 61 variables. It represents the calendar year 2023 crime.
```

```
**CHICAGO CRIME CHATBOT: Please ask a question. For example, ask
where are the crime hotspots
To end the session please hit the 'enter' key.**

 how was the data transformed
CHATBOT RESPONSE:  The data was transformed by reviewing and addressing potential issues such as r
emoving duplicates and highly correlated variables, addressing missing data, reviewing outliers an
d irrelevant data. Additionally, in some cases, additional variables were derived to aid in the an
alysis.
```

Appendix H: Links to Project Resources and Artifacts

Links to Github Repository

Chatbot	https://github.com/tdotkim/498_team51_code/tree/main/chatbot Possible test questions: (1) where are the crime hotspots, (2) how was the data transformed, (3) who is on the team, (4) what are the modeling methods, (5) what supervised methods are used
ETL	https://github.com/tdotkim/498_team51_code/tree/main/etl
EDA	https://github.com/tdotkim/498_team51_code/tree/main/eda
Modeling	https://github.com/tdotkim/498_team51_code/tree/main/modeling
Website	https://team51-v7s24t2e3q-uc.a.run.app/