Capstone Project: Exploring Crime Patterns and Predictions in Philadelphia Through Machine Learning

For the Requirement of *HarvardX: PH125.9x Data Science: Capstone Course*

Guler Arsal

November 15, 2024

Abstract

This project aimed to develop a predictive model for crime types in Philadelphia, leveraging a robust dataset with over two million crime records from 2006 to 2017. Using both descriptive analytics and machine learning, this study investigated patterns in crime distribution and developed a Random Forest model to classify crime types based on key factors like time, location, and contextual variables. The analysis reveals the model's ability to distinguish between classes, albeit with varying accuracy across categories, highlighting areas for model refinement. Key findings emphasize the model's strengths in capturing frequent crime types and identifying temporal and spatial patterns, yet underscore the challenges in accurately classifying all categories.

Contents

List of Tables

List of Figures

1 Introduction

Understanding patterns in criminal activity is essential for developing strategies that enhance public safety and optimize resource allocation (Chainey & Ratcliffe, 2013). This project leverages the Philadelphia Crime Dataset to construct a machine learning model that predicts the type of crime based on factors such as location, time, and additional contextual variables. Using predictive modeling to anticipate crime types can enable law enforcement agencies to identify high-risk areas, anticipate crime trends, and make data-informed decisions, all of which can contribute to preventing incidents and improving urban safety measures (Perry et al., 2013).

Machine learning techniques have become increasingly valuable in crime analysis, as they provide robust methods to detect complex, often hidden patterns within large datasets (Ahmed et al., 2021). Predictive models not only classify crime types but also support proactive policing efforts, guiding the allocation of resources where they are most needed and maximizing the efficacy of public safety initiatives (Kang & Kang, 2017).

The Philadelphia Crime Dataset, sourced from Kaggle (a prominent platform that curates real-world datasets), is an extensive collection of over two million crime reports spanning from 2006 to 2017. Each record captures detailed attributes of a crime event, including date, time, geographic coordinates, and specific crime category. Analyzing these features provides a foundation for building predictive models that can offer valuable insights into the temporal and spatial dynamics of crime in Philadelphia.

This project has three primary objectives:

1. **Exploration:** To investigate crime patterns in Philadelphia through descriptive analytics, uncovering trends by location and time.
2. **Prediction**: To develop a machine learning model that predicts crime types based on factors like time and location.
3. **Evaluation**: To assess the model's performance, identifying key features that contribute most to accurate crime classification.

2 Method

2.1 Philadelphia Crime Dataset Overview

The Philadelphia Crime Dataset is a publicly accessible resource that provides a detailed record of crime incidents in Philadelphia. Each incident is recorded with distinct attributes, such as date, time, location (latitude and longitude), and type of crime. The dataset's structure and level of detail make it suitable for a variety of analytical applications.

For this project, the dataset will be used to train and test machine learning models that predict crime types based on spatiotemporal factors. In Table 1 below, a comprehensive description of the dataset's variables is provided.

Table 1
Summary of Variables in Philadelphia Crime Dataset

Variable Name	**Description**	**Category**
Dc_Dist	District Code of the crime occurred	Spatial
Psa	Police Service Area code, which further divides a district for local policing	Spatial
Dispatch_Date_Time	The date and time when the crime dispatch occurred	Temporal
Dispatch_Date	The date of the crime dispatch	Temporal
Dispatch_Time	The time of the crime dispatch (hours and minutes)	Temporal
Hour	Hour of the day	Temporal
Dc_Key	A unique identifier for each crime incident	Crime Type
Location_Block	The block location where the crime occurred	Spatial
UCR_General	Uniform Crime Reporting General category code, used to classify crimes	Crime Type
Text_General_Code	A textual general code for the type of crime	Crime Type
Police_Districts	The police district where the crime occurred	Spatial
Month	The month of the dispatch	Temporal
Lon	Longitude of the crime location	Spatial
Lat	Latitude of the crime location	Spatial

2.2 Missing values

The completeness of the Philadelphia Crime Dataset was analyzed, specifically examining the presence of missing data across its 14 variables. Among these, 10 variables display full data integrity, containing no missing entries. However, four variables exhibit some degree of missing data, as detailed in Table 2. Notably, the proportion of missing data in these variables remains minimal, with the highest missing percentage being slightly above 0.89%.

Table 2
Number and Percentage of Missing Values in Columns

Variable Name	Missing Values	Percentage (%)
UCR_General	663	0.03
Text_General_Code	663	0.03
Police_Districts	19930	0.89
Lon	17349	0.78
Lat	17349	0.78

2.3 Software and Tools

All analyses were conducted using R version 4.4.1, a comprehensive software environment developed by the R Foundation for Statistical Computing, renowned for its robust data analysis and visualization capabilities (R Core Team, 2024). The report was compiled using R Markdown within RStudio, an integrated development environment specifically designed for R programming (RStudio Team, 2024). Both R and RStudio are open-source applications freely available to the public.

3 Exploratory Data Analysis

3.1 Crime Type

Figure 1 presents the distribution of crime types in Philadelphia using horizontal bar plots, arranged with the most frequent crimes at the top and the least frequent at the bottom. It offers a comprehensive overview of 33 distinct categories of criminal activity, highlighting the relative prevalence of each type.

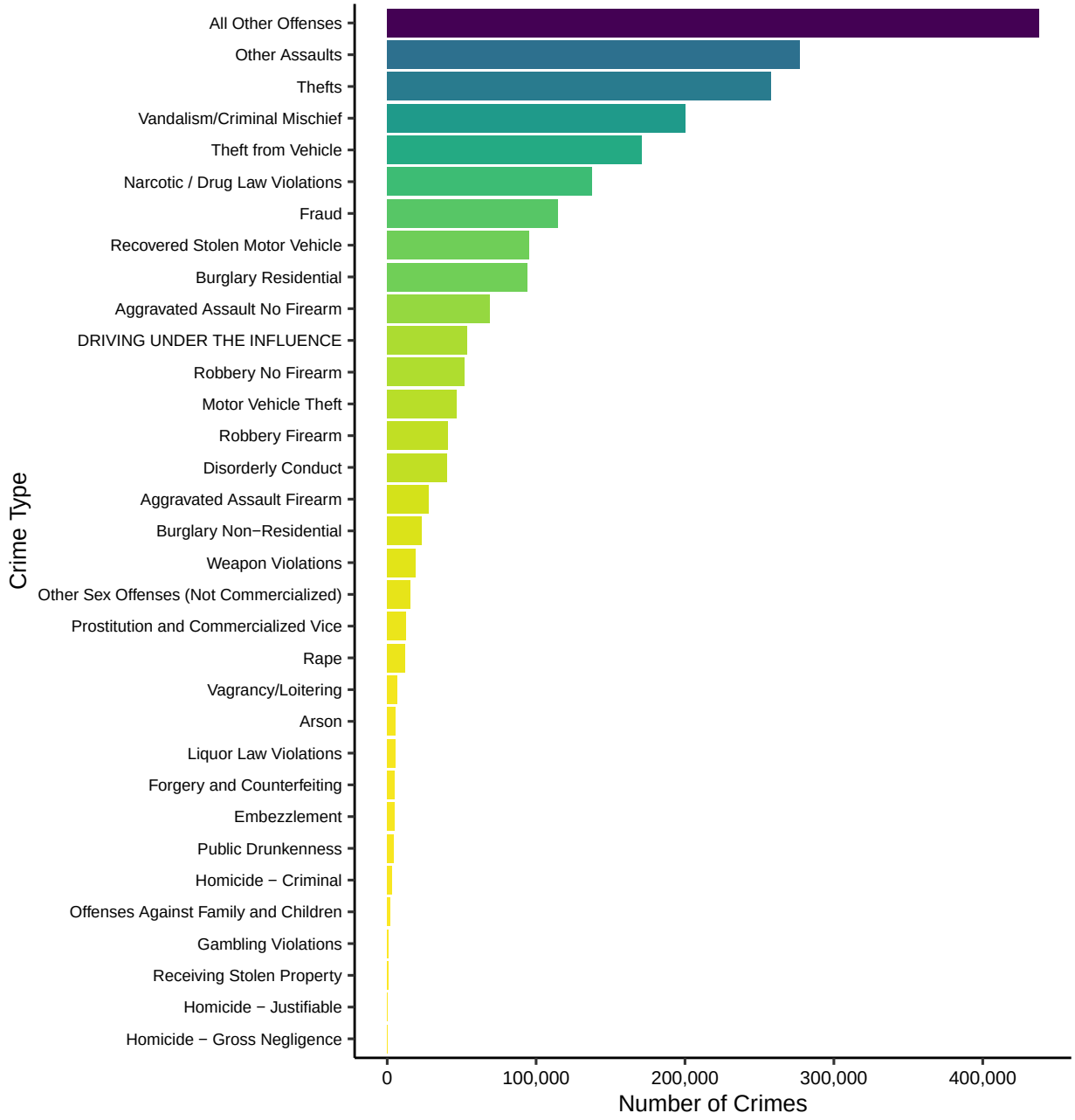

Figure 1: Number of Crimes by Crime Type

At the top of the distribution, "All Other Offenses" emerges as the most frequent crime, followed closely by "Other Assaults" and "Thefts." These categories account for a significant portion of Philadelphia's overall crime rates, a pattern consistent with findings from other urban studies, where assaults and theft-related offenses dominate the crime landscape (Malleson & Andresen, 2016).

Mid-range crimes, including "Vandalism/Criminal Mischief," "Theft from Vehicle," and "Narcotic/Drug Law Violations," reflect common urban challenges. These offenses are often linked to socioeconomic factors, such as poverty and inequality, that contribute to property damage, vehicle theft, and drug-related activities (Wikström & Treiber, 2016).

At the lower end of the spectrum, rarer but more severe crimes like "Homicide - Criminal" and "Homicide - Gross Negligence" are observed. These crimes occur less frequently but have significant social and legal consequences. Due to their severity, these violent crimes tend to draw more attention, reflecting national patterns where offenses such as homicide, though rare, are central to public and policy discussions (Daly & Wilson, 2017).

3.1.1 Thematic Organization of Crime Types

The crime categories can be thematically organized into five distinct groups to facilitate a clearer understanding of their characteristics and implications.

1. Other Offenses: This category includes offenses that do not fit neatly into other classifications, comprising:
    - All Other Offenses
    - Other Assaults
    - Other Sex Offenses (Not Commercialized)
2. Property Crimes: This group encompasses a wide range of offenses aimed at property, reflecting the impact of crime on individuals and communities through financial loss and property damage. It includes:
    - Thefts
    - Vandalism/Criminal Mischief
    - Theft from Vehicle
    - Fraud
    - Recovered Stolen Motor Vehicle
    - Burglary Residential
    - Motor Vehicle Theft
    - Burglary Non-Residential
    - Arson
    - Forgery and Counterfeiting
    - Embezzlement
    - Receiving Stolen Property

3. Drug and Alcohol-Related Crimes: This category addresses offenses associated with substance abuse, underscoring societal concerns related to addiction and public safety. It includes:

    - Narcotic / Drug Law Violations
    - Driving Under the Influence
    - Liquor Law Violations
    - Public Drunkenness

4. Crimes Against Public Order and Safety: This group encompasses offenses that disrupt community harmony and public safety, including:

    - Disorderly Conduct
    - Weapon Violations
    - Prostitution and Commercialized Vice
    - Vagrancy/Loitering
    - Gambling Violations

5. Violent Crimes: This category includes serious offenses that threaten personal safety and involve physical harm. It consists of:

    - Aggravated Assault No Firearm
    - Robbery No Firearm
    - Robbery Firearm
    - Aggravated Assault Firearm
    - Rape
    - Homicide - Criminal
    - Offenses Against Family and Children
    - Homicide - Justifiable
    - Homicide - Gross Negligence

3.2 Temporal Crime Analysis

The dataset has been enriched with several temporal features derived from the *Dispatch_Date_Time* variable. The month has been extracted as a label (e.g., "Jan," "Feb"), and both the year and numeric month values have been added as separate columns. Additionally, the day of the month and the day of the week (with Monday as 1 and Sunday as 7) have been extracted. These new features enable a more detailed analysis of crime trends across different time periods.

3.2.1 Crime Date

Figure 2 presents a line plot depicting the daily distribution of reported crimes. The black dashed line represents a LOESS smoothing curve, highlighting the overall trend in crime rates over time. From 2006 to 2017, the trend indicates a gradual decline in the number of reported crimes.

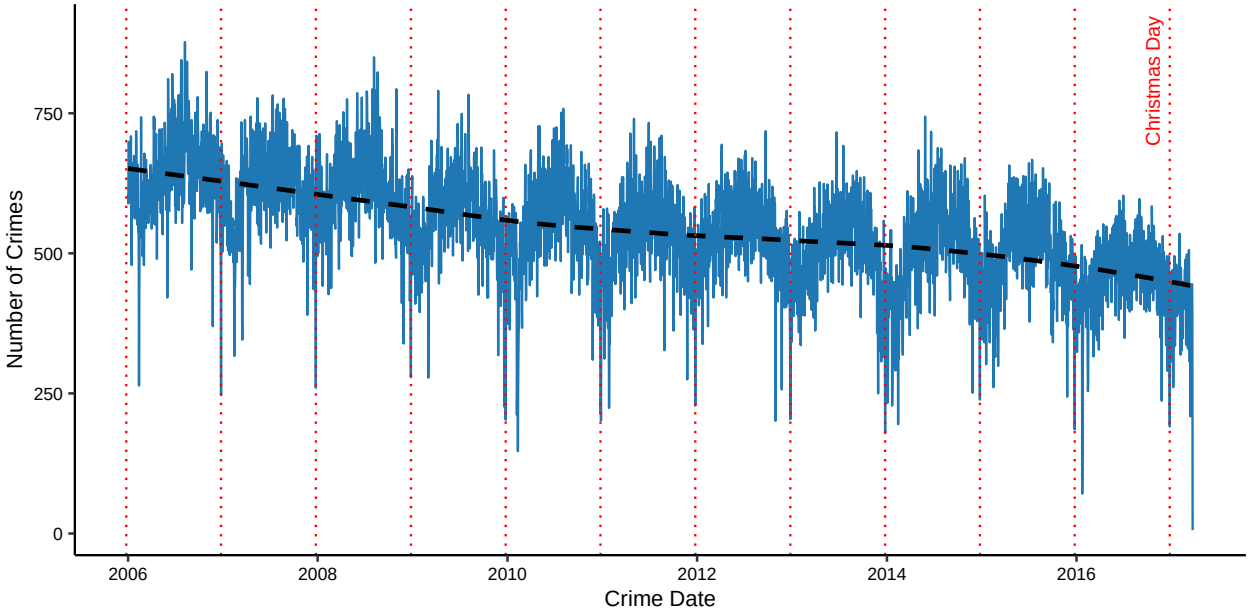

Figure 2: Number of Crimes by Crime Date

Seasonal fluctuations are evident, with recurring peaks and dips in crime rates throughout the year. Notably, there is a significant reduction in crime during the Christmas period, starting in mid-December and extending through the end of the month. This seasonal decline is marked by vertical red dashed lines indicating Christmas Day each year and may reflect behavioral, environmental, or social factors, such as increased community engagement or heightened security, that lead to a temporary reduction in criminal activity during the holiday season.

Additionally, the figure reveals a recurring pattern of crime peaking in the middle of each year, typically during the summer months. This suggests a seasonal trend, where crime rates tend to rise during warmer weather, potentially driven by increased outdoor activity and social interactions.

3.2.2 Crime Month

Figure 3 presents a line plot depicting the monthly distribution of reported crimes. This figure highlights a gradual overall decline in crime rates from 2006 to 2017 and further emphasizes the seasonal fluctuations shown in Figure 1. Crime rates consistently peak during the middle of the year, particularly in the summer months, suggesting that warmer weather and increased outdoor activities may contribute to the rise in crime. Conversely, crime occurrences tend to be lower in the winter, especially around the Christmas period, as highlighted in Figure 1.

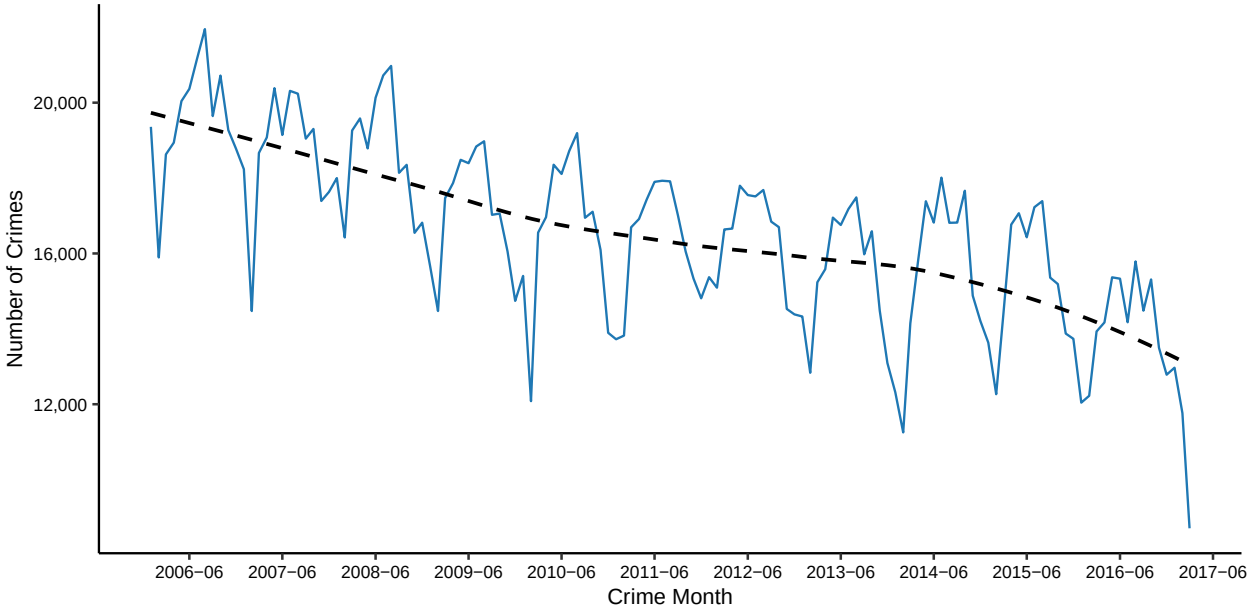

Figure 3: Number of Crimes by Crime Month

To gain a deeper understanding of crime trends, the analysis focused on monthly data for individual offense categories, allowing for the identification of detailed patterns that may be obscured in the aggregate data. By disaggregating the data, trends within specific crime types were examined, offering insights into the fluctuations and variations that occur over time. This approach helps pinpoint areas where crime prevention efforts may have been effective or where unusual patterns could warrant further investigation.

Figures 4 through 8 present the monthly trends for various crime categories. These figures highlight the fluctuations in crime rates, shedding light on how specific offenses evolve over time. By focusing on these individual crime types, we can better understand their seasonal variations and potentially identify the impact of external factors, such as law enforcement strategies or socio-economic conditions, on crime trends.

Figure 4 illustrates the monthly trends for the **Other Offenses** category, which includes crime types such as "All Other Offenses," "Other Assaults," and "Other Sex Offenses (Not Commercialized)." These categories generally show a decline in crime rates, consistent with the broader downward trends observed in the overall data. Similar to the seasonal fluctuations seen in the aggregated data, "All Other Offenses" and "Other Assaults" categories also peak during the warmer months and decrease in the winter.

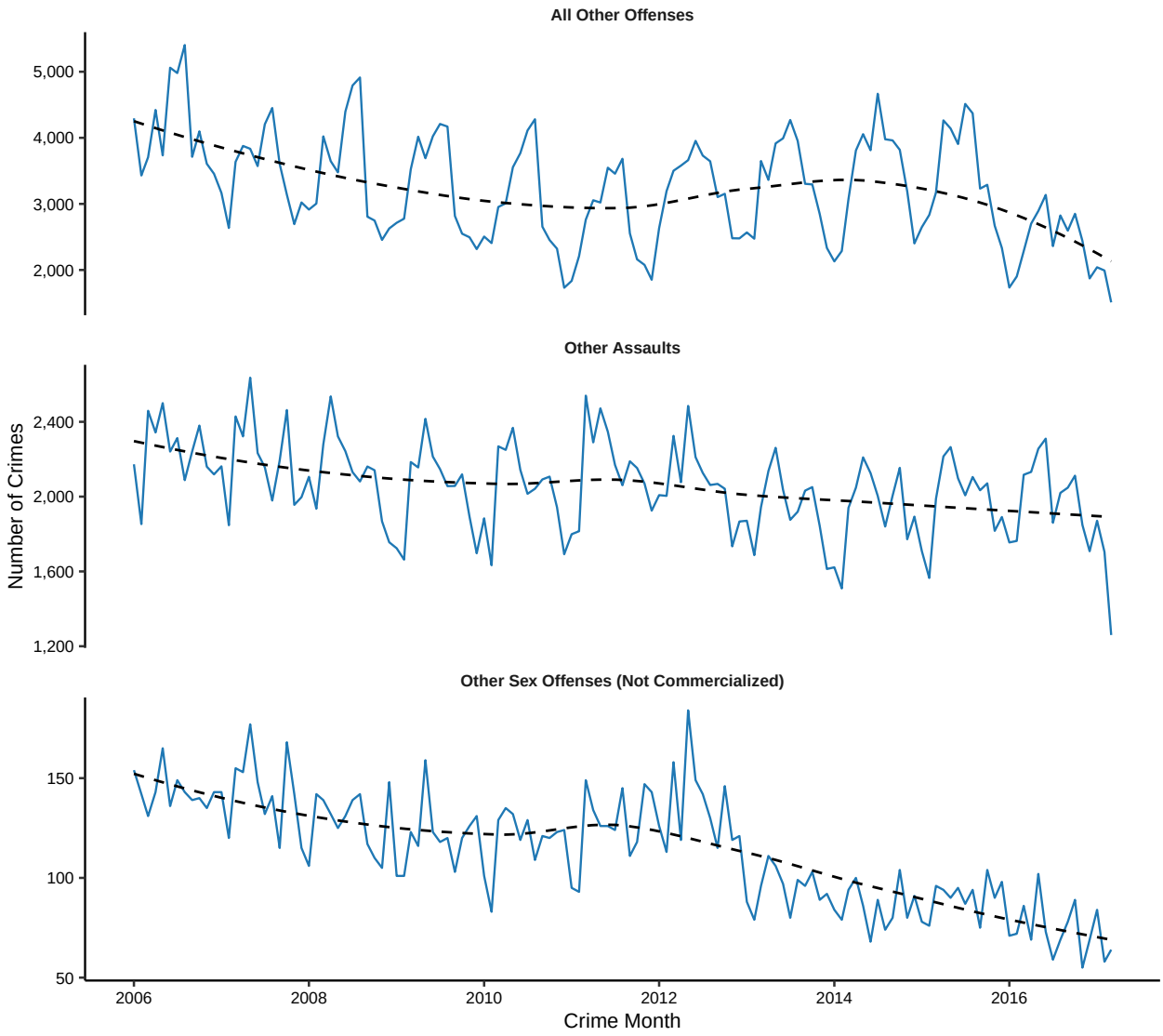

Figure 4: Number of Crimes by Crime Month for Other Crimes

Figure 5 presents a series of line plots illustrating the monthly trends for **Property** crime categories. Most of these categories, such as "Vandalism/Criminal Mischief" and "Burglary Non-Residential" show a clear downward trend in crime rates over time, reflecting a broader decline in property-related offenses.

In contrast, the "Fraud" category deviates from this general trend. The "Fraud" category exhibit relatively stable rates until around 2014, followed by a sharp increase in 2015. This sudden spike suggests that external factors, such as changes in economic conditions, policy shifts, or perhaps new opportunities for fraud (e.g., technological developments), may have played a role in the rise of these offenses. The anomaly in 2015 stands out as an area requiring further investigation to understand the underlying causes and to inform targeted crime prevention strategies for fraud.

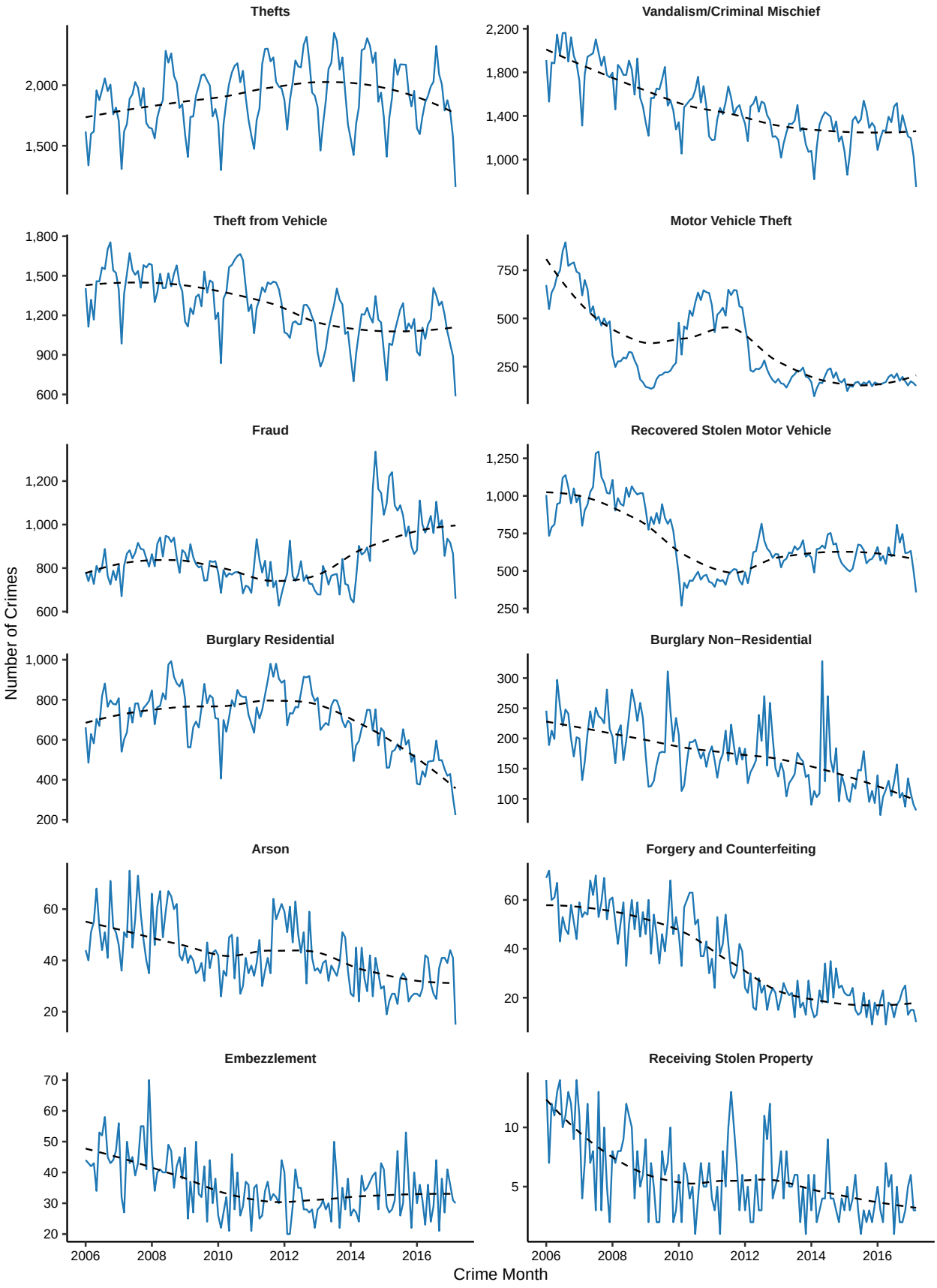

Figure 5: Number of Crimes by Crime Month for Property Crimes

Figure 6 displays a series of line plots depicting the monthly trends for **Drug and Alcohol-Related** crimes. Categories such as "Narcotic / Drug Law Violations," "Driving Under the Influence," and "Liquor Law Violations" exhibit relatively similar patterns, with parallel fluctuations over time. In contrast, "Public Drunkenness" shows a slightly different trend, although the variation is minimal. It's important to note, however, that the number of occurrences for "Public Drunkenness" is lower compared to the other categories, which may account for the less pronounced trend.

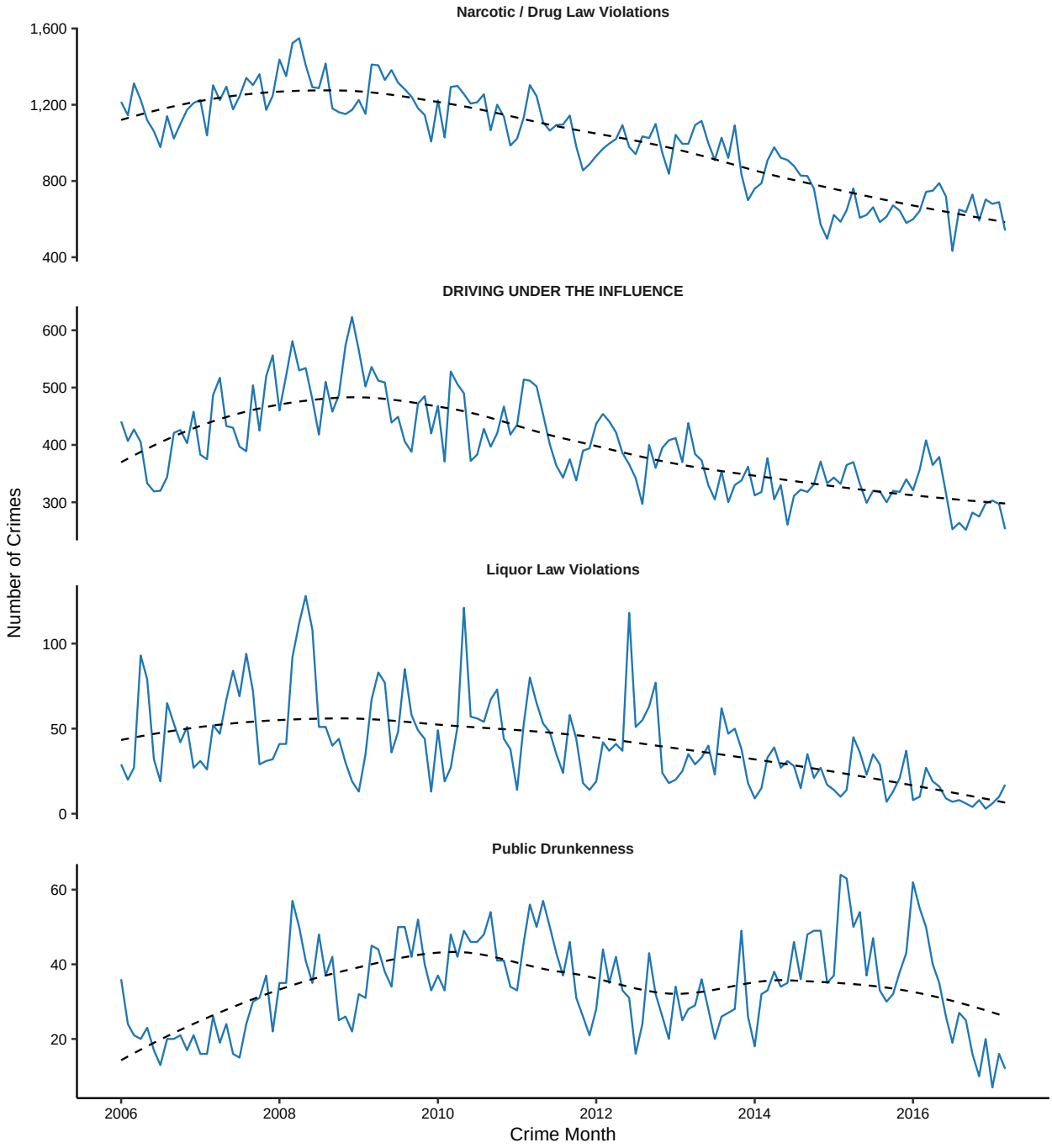

Figure 6: Number of Crimes by Crime Month for Drug and Alcohol-Related Crimes

Figure 7 displays the line plots for **Crimes Against Public Order and Safety**. "Disorderly Conduct" shows clear downward trends over time. Similarly, "Weapon Violations" also exhibit a general decline; however, there is a noticeable uptick in rates following 2013, suggesting a potential resurgence in these types of offenses. "Gambling Violations", though generally low in frequency, show a gradual decline, while "Prostitution and Commercialized Vice" exhibit significant fluctuations with occasional sharp spikes. Despite this variability, there is a slight downward trend in prostitution-related crimes. In contrast, "Vagrancy/Loitering" stands out due to its sharp spike around 2015.

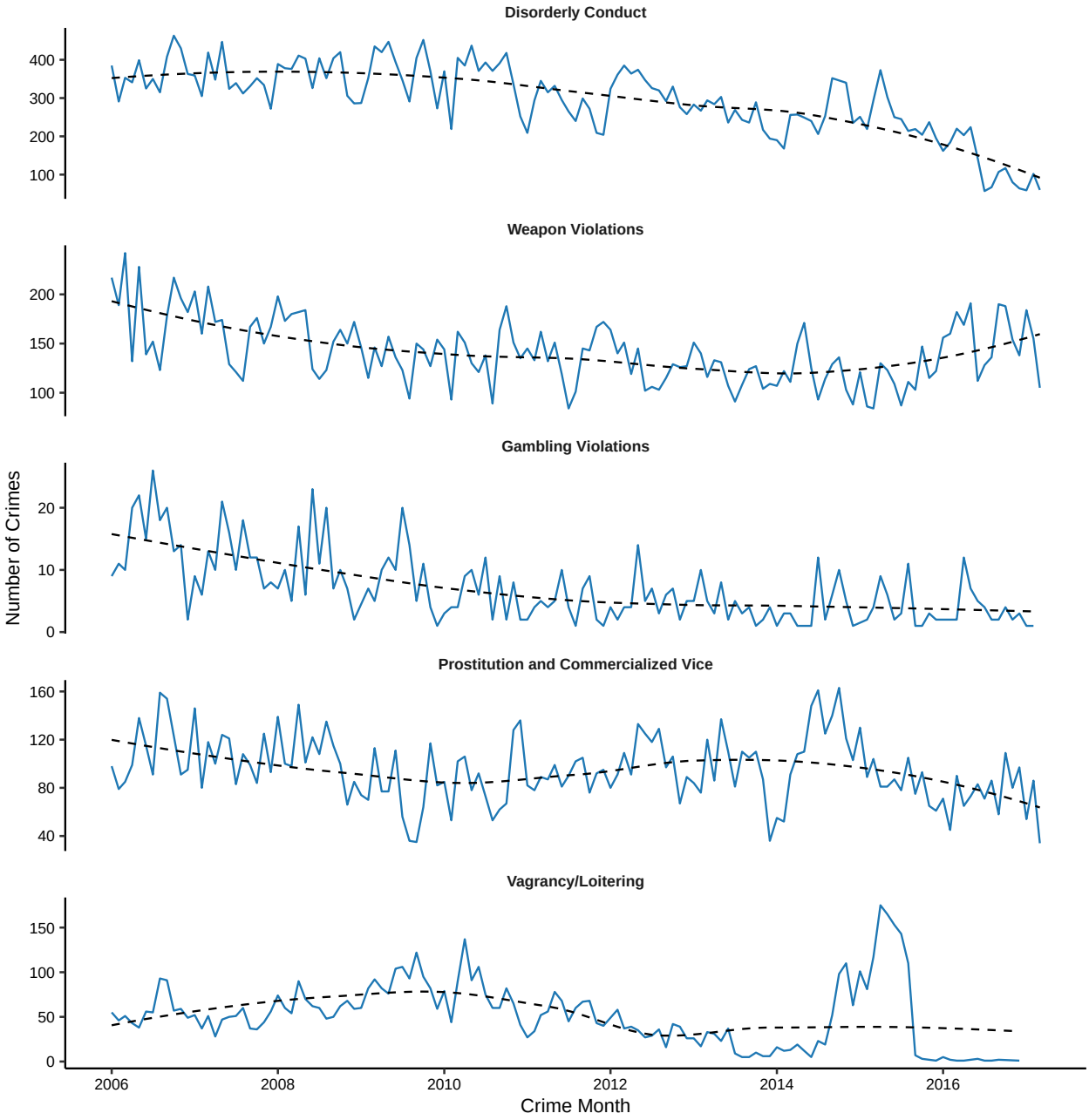

Figure 7: Number of Crimes by Crime Month for Crimes Against Public Order and Safety

Figure 8 displays a series of line plots depicting the monthly trends for **Violent** crimes. Both "Aggravated Assault" categories (firearm and non-firearm) and both "Robbery" categories (with and without firearms) exhibit a clear downward trend, with seasonal spikes typically occurring in the summer months.

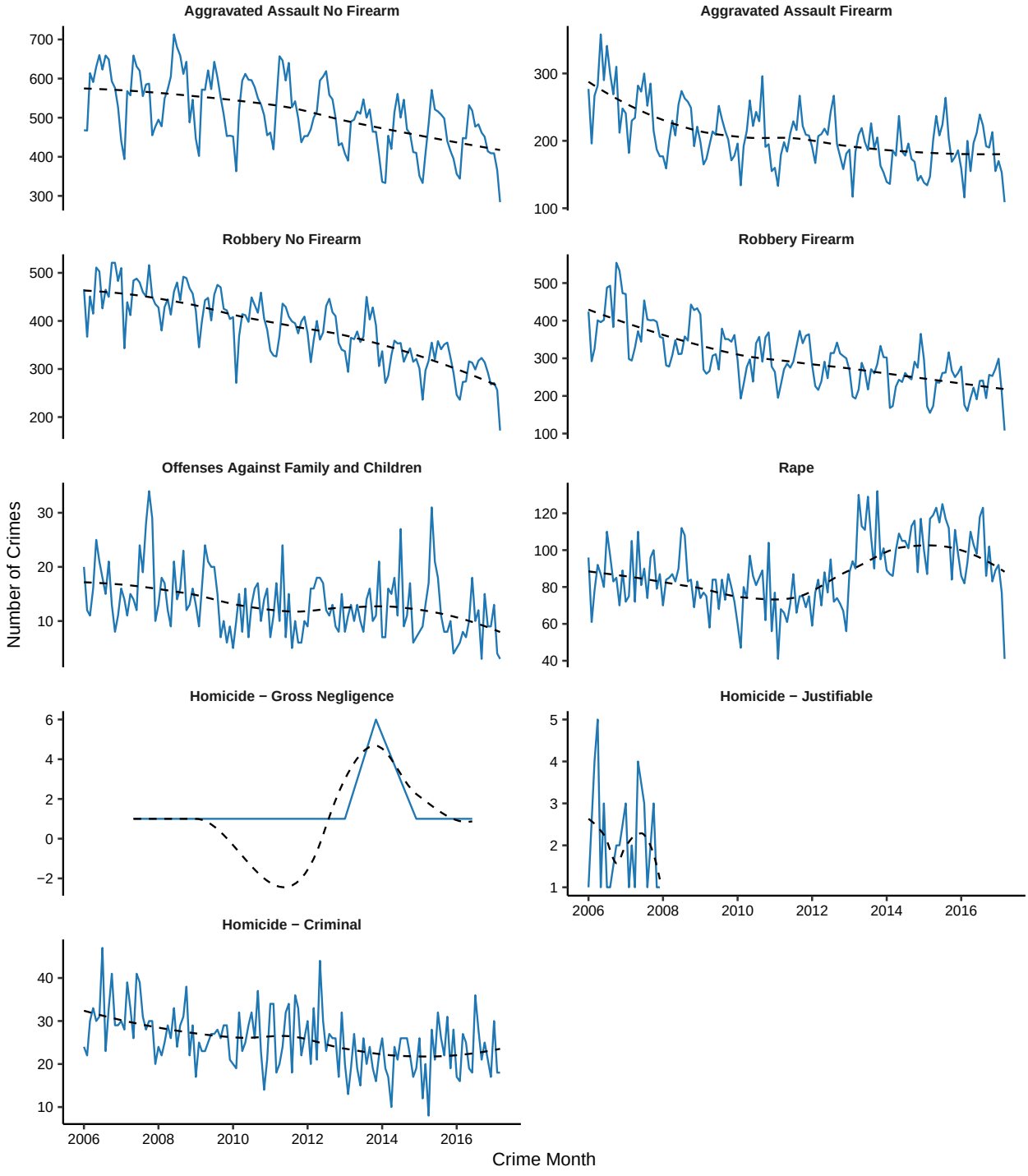

Figure 8: Number of Crimes by Crime Month for Violent Crimes

"Offenses Against Family and Children" exhibit fluctuations but no long-term trend, while "Rape" shows a more distinct downward trend, particularly in the later years.

"Homicide - Gross Negligence", although rare, shows a clear spike around 2014-2015, potentially indicating an isolated event or a change in reporting during this time period. "Homicide - Justifiable" remains low throughout the period, with no discernible trend."Homicide - Criminal" remains relatively stable, without a clear upward or downward trend, suggesting that the factors influencing this crime type may differ from those affecting robbery or aggravated assault.

Overall, the downward trends in many of these categories reflect broader crime reduction patterns, though certain anomalies, like the spike in gross negligence homicides, warrant further investigation.

3.2.3 Monthly Aggregates

Figure 9 presents a bar plot that aggregates crime data for each month, without distinguishing between specific years. The x-axis represents the months (January to December), while the y-axis reflects the number of recorded incidents. The bar plot confirms the trend observed previously, with crime rates rising during the summer and declining in the winter.

February shows a notably lower frequency of crimes, which might be related to the fact that it has fewer days compared to other months, though this would require further investigation to confirm.

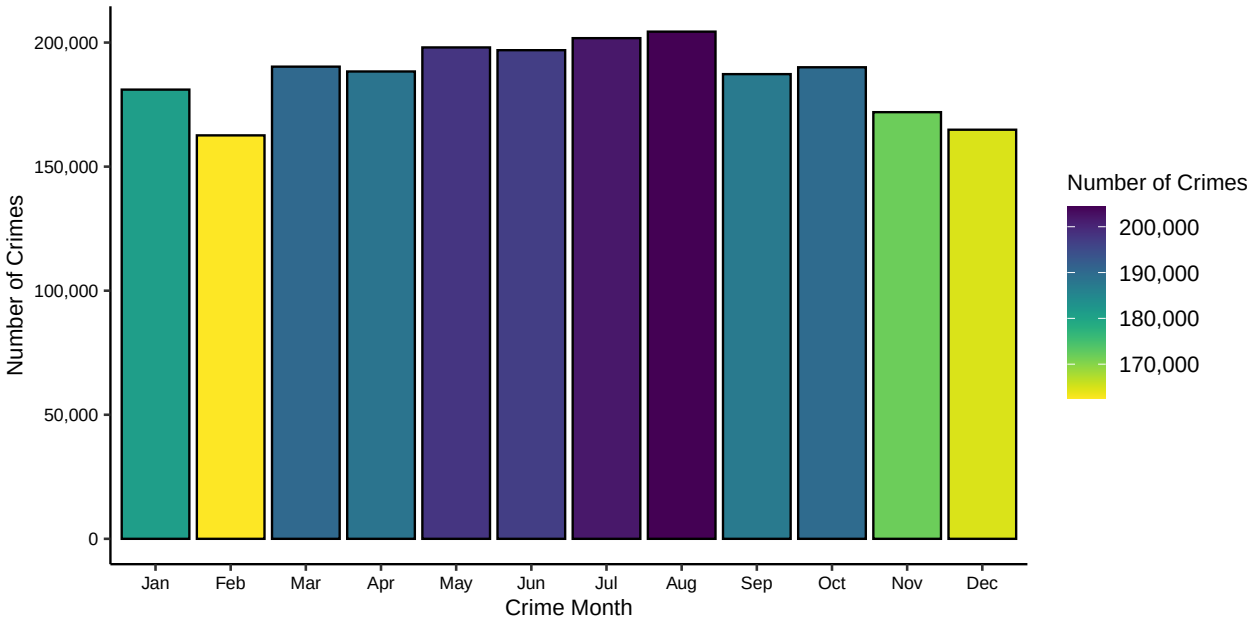

Figure 9: Number of Crime by Monthly Aggregation

3.2.4 Crime Hour

Figure 10 shows a histogram of crime frequency by hour of the day, highlighting clear variations in crime activity. Crime peaks occur between 10 AM and 1 PM, with another rise in late afternoon. These periods likely coincide with increased public activity, such as work hours and commuting, suggesting that crimes may be more opportunistic when more people are out and about.

In contrast, crime frequency drops significantly during the early morning hours, particularly from midnight to 6 AM. This period of low activity aligns with reduced public presence in urban areas, which may lower the likelihood of both crimes and their reporting during these quieter hours. The overall pattern reflects the rhythms of urban life, where higher crime rates occur during periods of greater social interaction and economic activity, while quieter hours see fewer opportunities for crime to take place or be reported.

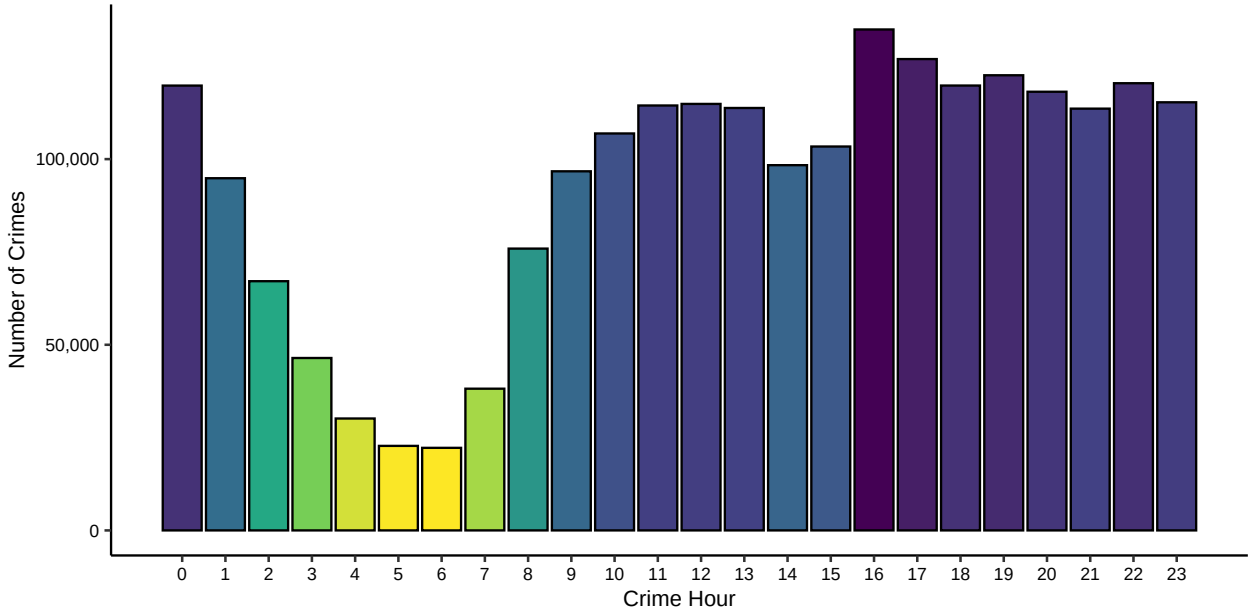

Figure 10: Number of Crimes by Crime Hour

To deepen the analysis, Figure 11 introduces a line graph that incorporates year-wise data, illustrating the hourly distribution of crimes from 2006 to 2017. As with the previous figure, crime activity tends to be at its lowest during the early morning hours (midnight to 6 AM), with a steep decline from midnight. After 6 AM, there is a steady increase in crime frequency, peaking sharply between 3 PM and 6 PM.Post 6 PM, crime numbers gradually decrease, but with a noticeable rise between 10 PM and midnight, indicating late-night activity.

The general pattern of crime activity remains consistent across years, though 2017 (yellow line) stands out with notably lower crime counts. However, this lower crime counts is largely influenced by incomplete data, as the dataset for 2017 only covers the months from January to March. The absence of data from the remaining months distorts the overall crime trend for 2017. Therefore, any conclusions drawn from this particular year must be interpreted with caution, as it does not offer a comprehensive picture of the entire year's crime patterns.

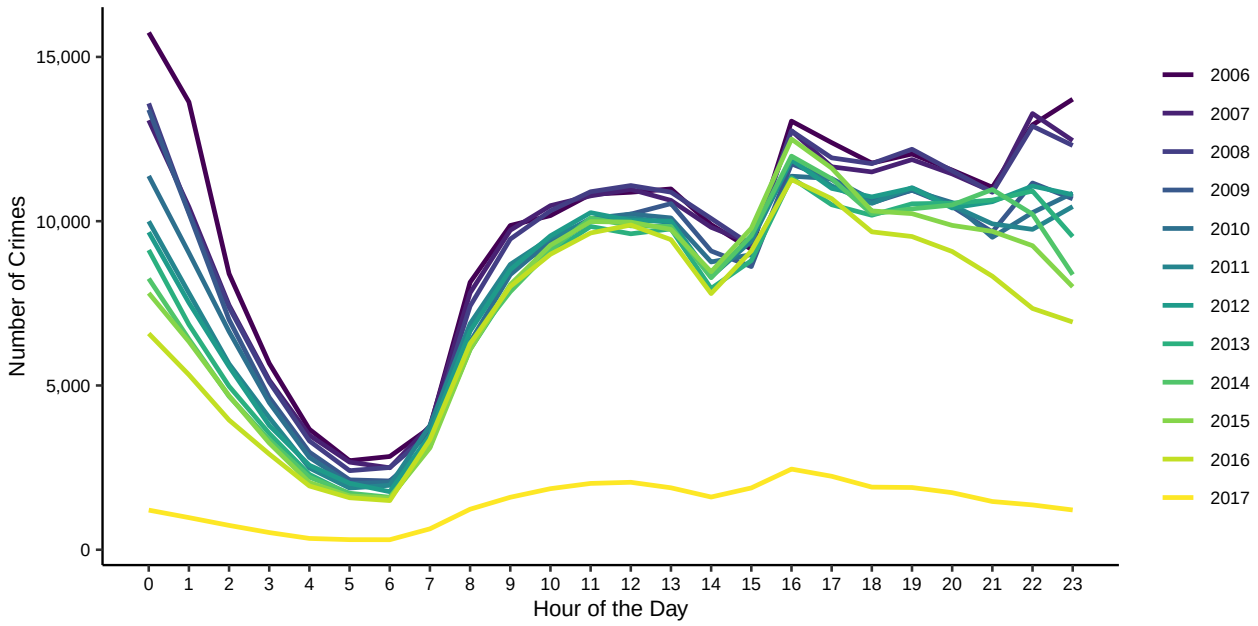

Figure 11: Number of Crimes by Crime Hour for Each Year

Figure 12 illustrates the hourly distribution of crime types using a color scale from yellow to purple, highlighting crime frequency at different times of the day.

The results reveal clear temporal trends across Drug and Alcohol-Related crimes. "Driving Under the Influence" and "Public Drunkenness" show a sharp increase in frequency during the late evening and early morning hours (from around 10 PM to 3 AM), coinciding with common social activities and nightlife. This pattern is expected, as these crimes are often linked to alcohol consumption and nighttime behavior. "Narcotic / Drug Law Violations" and "Liquor Law Violations" exhibit a broader distribution throughout the day, peaking in the late afternoon and evening.

Most of the Violent crimes, such as "Aggravated Assault" (both firearm and non-firearm) and "Robbery" (both with and without firearms) categories peak between the evening and midnight hours, suggesting a higher likelihood of confrontations during the evening.

"Disorderly Conduct" follows a similar trend, with higher frequencies in the evening and nighttime, reflecting social tensions that may escalate during this period. "Homicide - Criminal" is most frequent in the late evening hours as well.

The Property crimes of "Thefts", "Theft from Vehicle", "Motor Vehicle Theft", "Forgery and Counterfeiting", "Fraud" and "Embezzlement" have a very similar distribution with one peak at around 4pm. Some other property crimes, such as "Burglary Residential" displays a dual peak, one in the early morning hours and another around 4pm.

In summary, Figure 12 highlights clear temporal trends in crime patterns, which can inform more effective resource allocation and crime prevention strategies based on time of day.

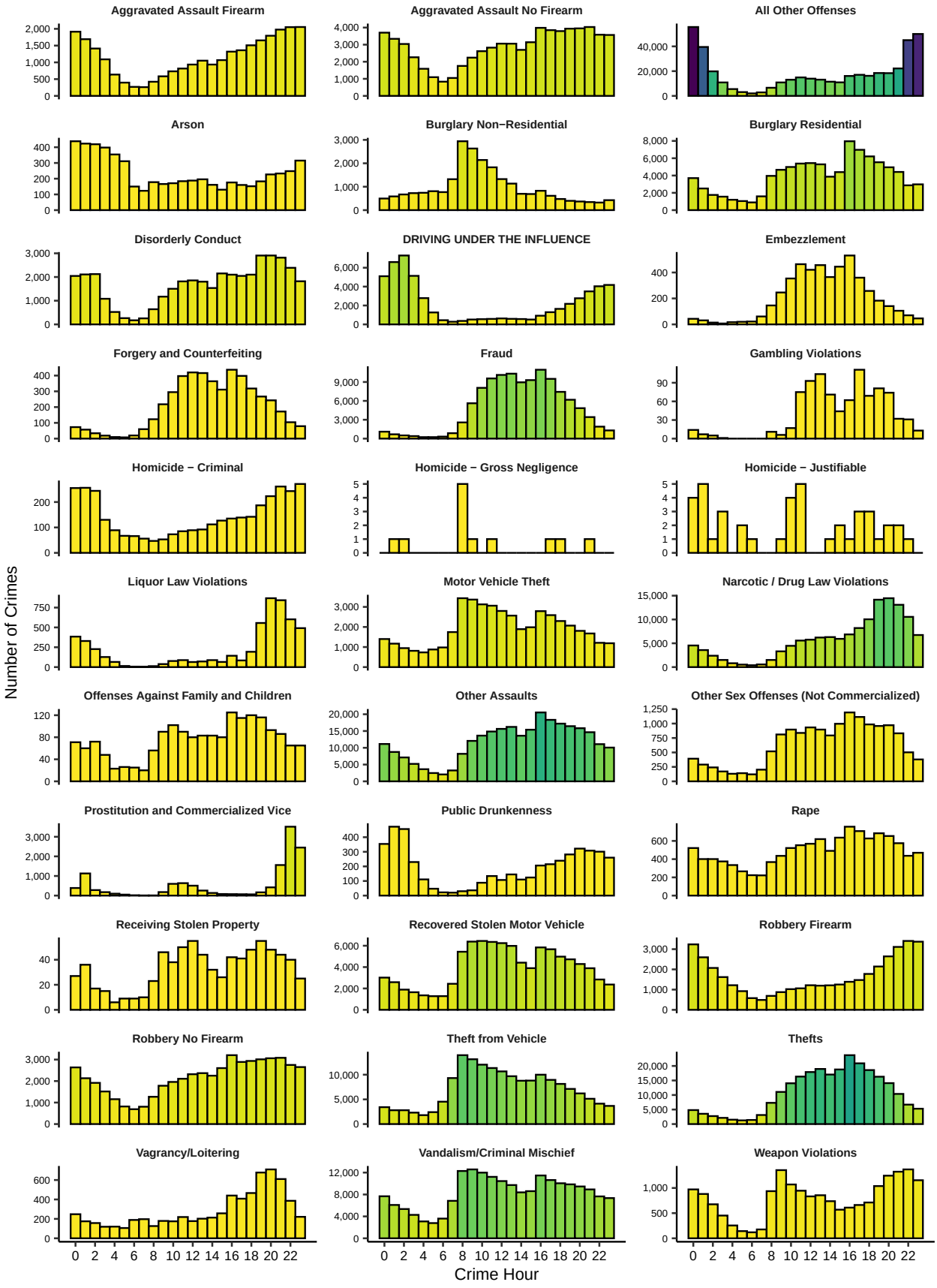

Figure 12: Number of Crimes by Crime Hour for Each Crime Type

3.2.5 Crime Weekday

Figure 13 illustrates the number of crimes that occurred on each day of the week. The data reveals that crime counts remain relatively high and consistent from Monday through Friday, with Tuesday and Wednesday showing the highest crime counts, each exceeding 330,000 incidents. Crime numbers start to decline slightly toward the weekend, with Sunday showing the lowest count. The lower crime activity on weekends may reflect differences in daily activities or law enforcement practices.

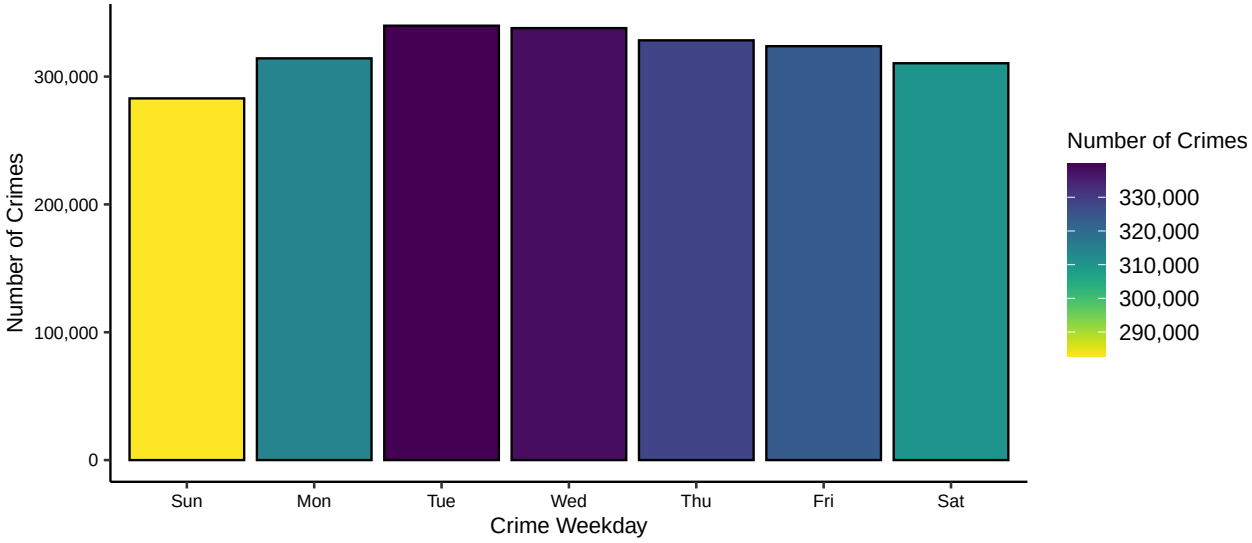

Figure 13: Number of Crime by Weekday

3.3 Spatial Crime Analysis

To incorporate police district boundaries for mapping, a GeoJSON file from OpenDataPhilly.org was downloaded and loaded into R as a spatial object (see code snippet). During this process, discrepancies were identified between district codes in the dataset and those in the boundary file. Specifically, the dataset contained four additional district codes (4, 6, 23, and 92), which reflect earlier police district divisions before departmental restructuring. These districts were merged into others as follows:

- The old 4th District was merged into the 3rd District.
- The old 6th District was merged into the 9th District.
- The old 23rd District was merged into the 22nd District.
- The old 92nd District was merged into the 16th District.

To align with the current district boundaries, the Dc_Dist variable was updated to reflect these changes accurately, ensuring consistency between the dataset and boundary file for accurate spatial analysis.

Figure 14 illustrates a bar chart representing the distribution of crimes across different district codes. The results show that District 15 has the highest number of recorded crimes, exceeding 150,000 incidents, followed closely by District 09. District 24, 22, and 25 have significant crime rates as well. These districts show a substantial crime burden, indicating that they may encompass larger or more densely populated areas, or areas with higher crime incidence rates.

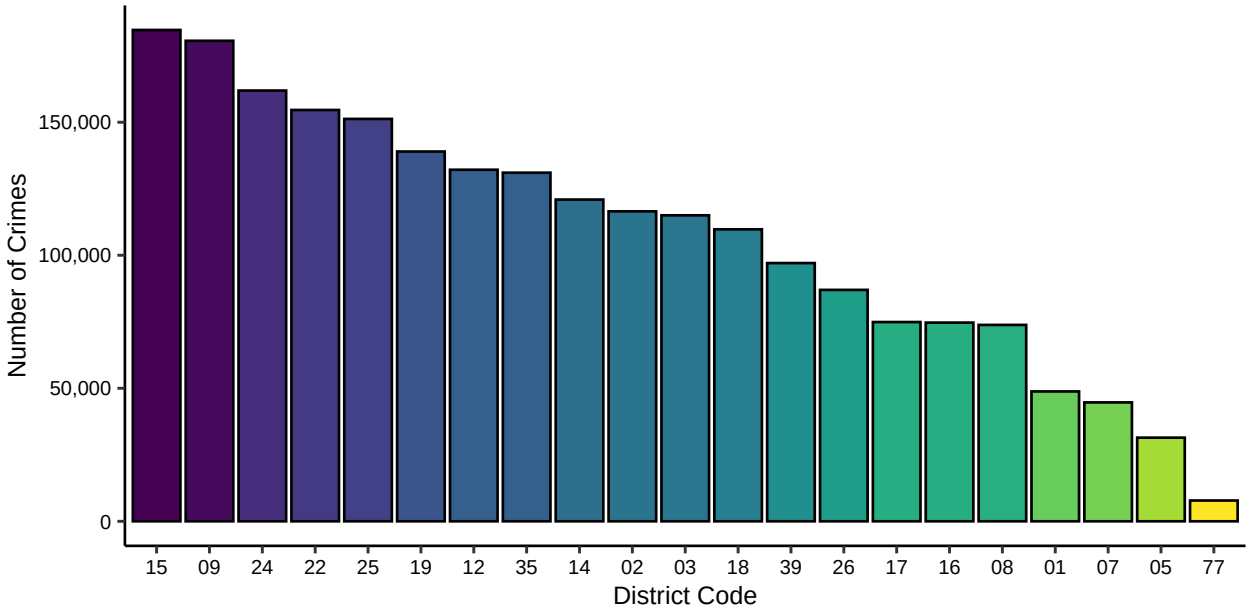

Figure 14: Number of Crimes by District Code

In contrast, the districts toward the right side of Figure 14, such as District 77 and District 05, report the fewest number of crimes, each with less than 50,000 incidents. The drop-off in crime frequency between the left and right ends of the chart is visually stark, indicating that certain districts experience far fewer criminal activities compared to others.

Figure 15 shows a map of Philadelphia with district boundaries outlined in black and orange-shaded density contours representing crime distribution. Darker orange areas indicate higher crime density, while lighter areas show lower crime concentrations. The visualization, created using kernel density estimation, is based on a 10% random sample of the dataset to manage computational complexity and highlight crime hotspots effectively.

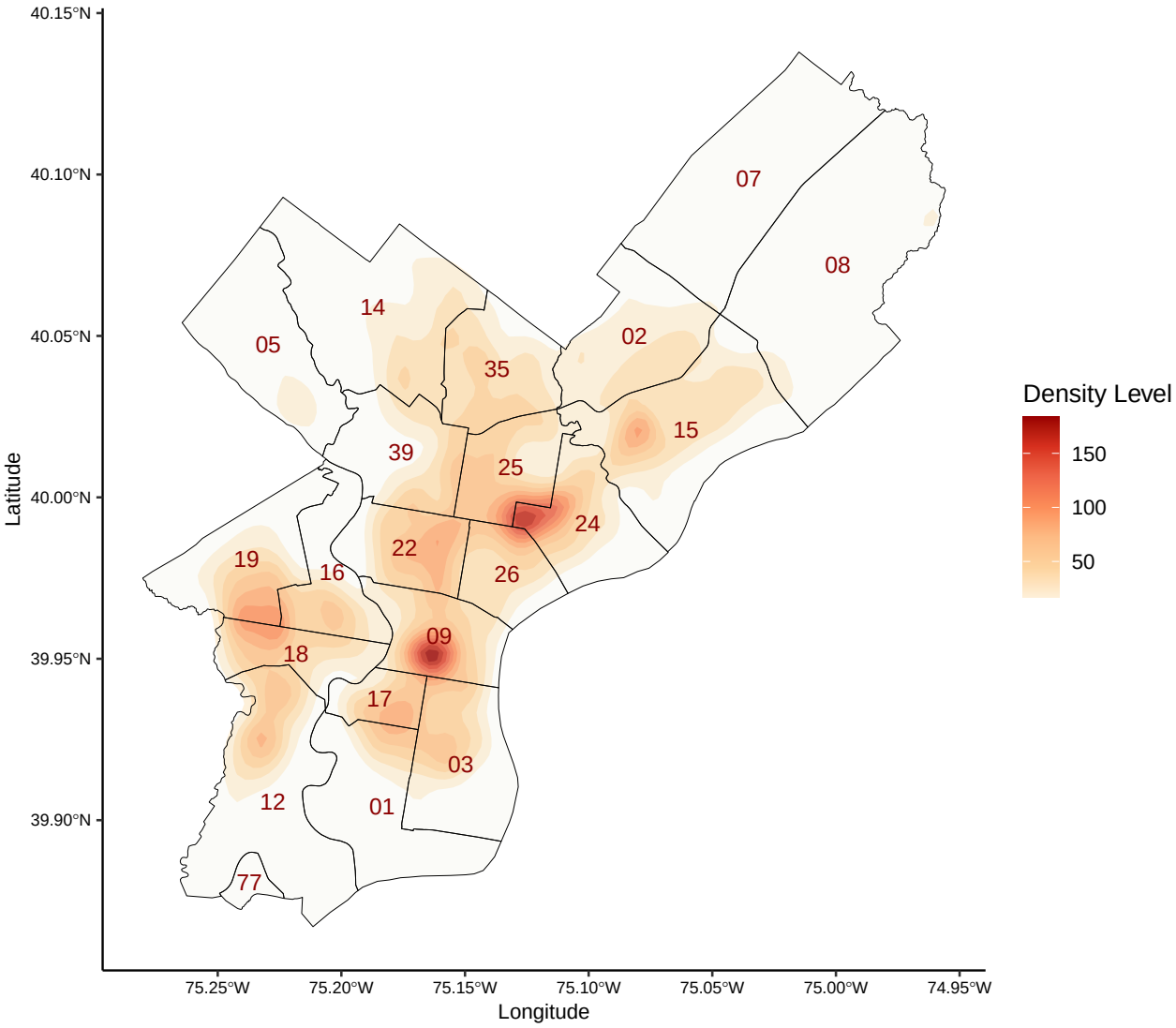

Figure 15: Spatial Distribution of Crime Density Across Districts

Figure 16 presents the distribution of crime categories across districts. Certain districts are showing heightened criminal activity across multiple crime types.

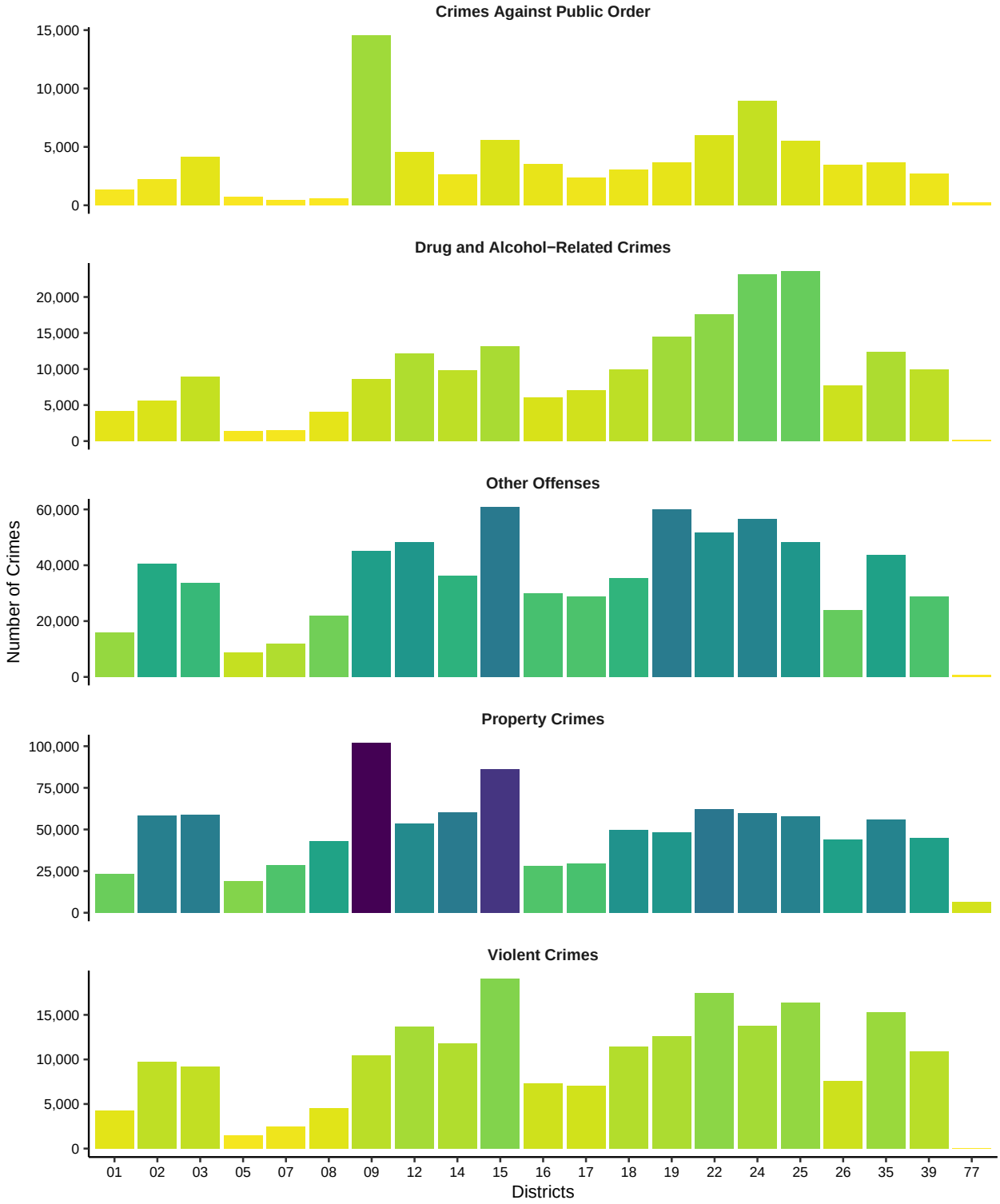

Figure 16: Number of Crimes by Crime Type and Districs

Crimes Against Public Order: District 9 shows a notably high number of public order crimes compared to other districts, with a sharp peak at around 15,000 incidents. Other districts have relatively lower occurrences, with values ranging from around 2,000 to 10,000, showing a more uniform spread.

Drug and Alcohol-Related Crimes: There is a significant spike in Districts 24 and 25, which both report over 20,000 incidents. Other districts, such as 19 and 22, also have elevated numbers at around 15,000 incidents.

Other Offenses: These crimes appear more evenly distributed, with Districts 15, 19, 22, and 24 reporting higher crime counts, reaching over 50,000 incidents.

Property Crimes: District 09 stands out with over 100,000 property crimes, followed by District 15 with a significant count of around 85,000.

Violent Crimes: District 15 again stands out, with a peak of around 19,000 violent crimes. Other districts, like 22 and 25, have elevated numbers, reaching over 16,000 incidents.

Figure 17 presents a heatmap illustrating the distribution of crime frequency across districts and Police Service Areas (PSAs). The x-axis displays the districts. The y-axis represents the PSAs, labeled alphabetically from A to Z, alongside numbers for PSAs 1 through 4.

The color gradient, ranging from yellow (representing lower frequencies) to deep purple (indicating higher crime frequencies), clearly highlights disparities in crime rates between different districts and PSAs. Notably, the areas corresponding to districts 9, 15, 19, 22, 24 and 25 show concentrated areas of darker shades, particularly in lower PSAs, suggesting significantly higher crime rates in these districts.

Many of the higher-numbered PSAs exhibit lighter colors, indicating relatively fewer recorded incidents. Additionally, several gaps or light patches appear in specific PSA and district combinations, potentially reflecting regions with no reported data.

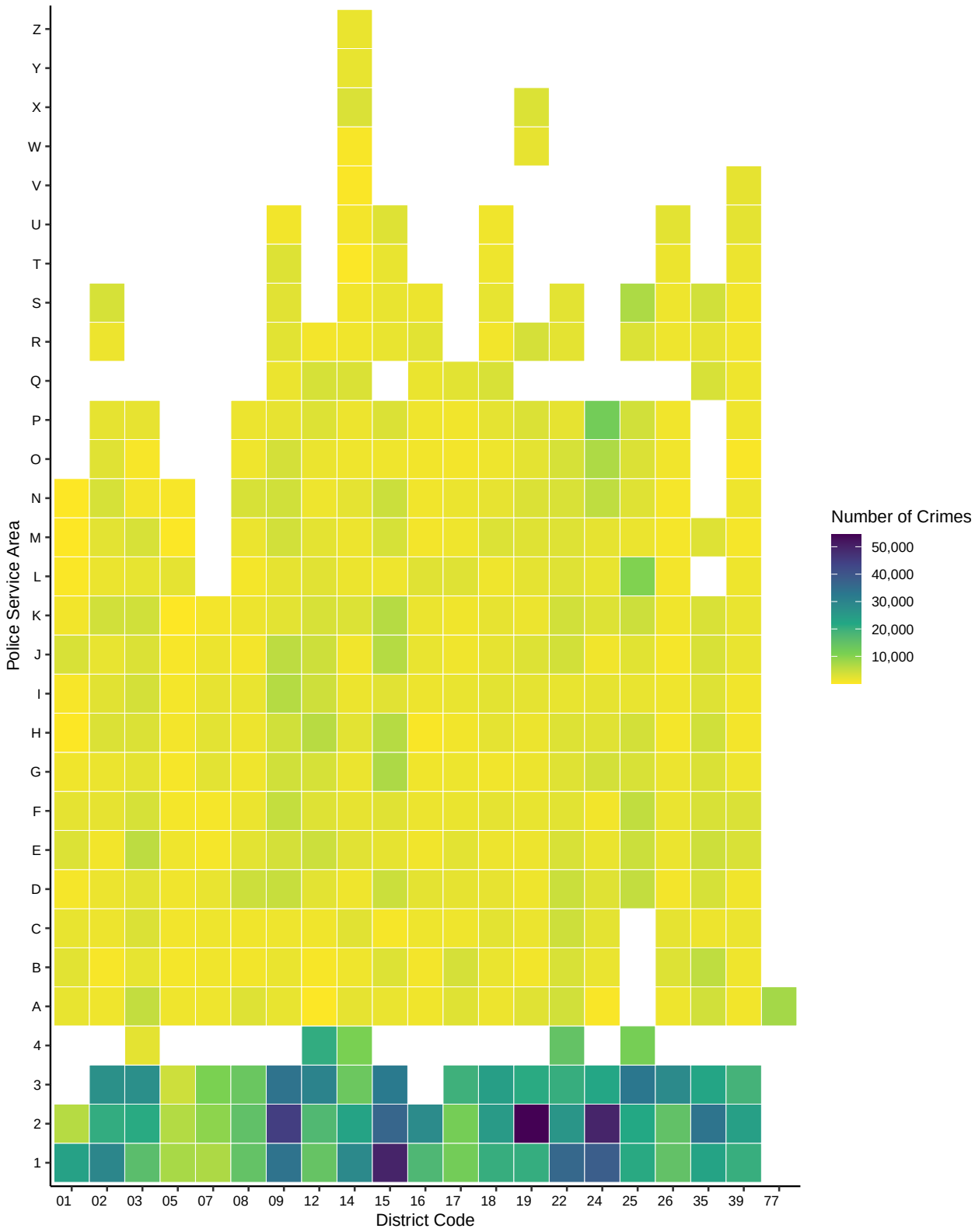

Figure 17: Heatmap of Number of Crimes by Districts and Police Service Area

4 Machine Learning Models

The objective is to predict crime type based on historical patterns related to time and location, focusing specifically on the 10 most frequent crime categories. By limiting the analysis to the most common types, this approach addresses data imbalance, reducing the influence of rare crime types and enhancing model simplicity, predictive performance, and stability.

1. Dependent Variable:
    - The target variable consists of the 10 most frequently occurring crime types in the dataset.
2. Independent Variables:
    - Year: Crime occurrence year, ranging from 2006 to 2017.
    - Month: Month of occurrence (1 to 12).
    - Day: Day of the month on which the crime occurred (1 to 31).
    - Hour: Hour of the day when the crime was reported (0 to 23).
    - Weekday: Day of the week, from Monday to Sunday.
    - Longitude (Lon): Longitude coordinate of the crime location.
    - Latitude (Lat): Latitude coordinate of the crime location.
    - District: The district where the crime occurred.

4.1 Data Splitting Strategy

To develop predictive models, the dataset was split into training and testing sets, with an 80/20 split. The training set, comprising 80% of the data, was further divided using an additional 80/20 split, resulting in 64% of the original data for training, 16% for validation, and 20% for testing.

This approach offered several advantages. The validation set enabled effective hyperparameter tuning, optimizing model performance while minimizing overfitting. Testing various configurations on the validation set allowed for selecting the best parameter values. Additionally, when comparing models (e.g., Decision Tree, Random Forest), the validation set provided an unbiased basis for performance assessment, facilitating a fair comparison in selecting the most accurate model for crime type prediction. By reserving the test set solely for final evaluation, the model's ability to generalize was thoroughly tested on unseen data, ensuring reliable and realistic insights into its real-world performance.

4.2 Handling Class Imbalance: SMOTE Application

Exploratory data analysis revealed a notable class imbalance in the dataset, with certain crime types occurring far more frequently than others. This imbalance poses a risk of model bias, as the model could become predisposed to the more common categories, resulting in higher accuracy for these classes while potentially underperforming on the less frequent crime types. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic examples for the minority classes, creating a more balanced dataset.

In the subsequent analysis, models will be trained on both the original (imbalanced) dataset and the SMOTE-transformed dataset. This approach aims to assess the impact of class balancing on model performance across crime types.

4.3 Model Training and Validation

4.3.1 Model 1a: Decision Tree Model (Original Data)

The Decision Tree model, trained on the original imbalanced dataset, achieved the following performance metrics:

- Accuracy: The model achieved an accuracy of 26.04%, meaning it correctly predicted the class for approximately 26% of cases. Although this is slightly better than random guessing (which would yield around 10% accuracy in a 10-class classification problem), the model's performance is still quite limited.
- Kappa: The Kappa statistic was 5.80%, suggesting that the model's predictions exhibit only minimal agreement with the true labels beyond what would be expected by chance.
- Confidence Interval for Accuracy: The 95% confidence interval for accuracy ranged from 25.88% to 26.20%. This narrow range indicates stable, but low, model performance.
- Accuracy Null: The null accuracy represents the accuracy achieved by always predicting the most common class. Here, the null accuracy is around 23.65%. The model's accuracy (26.04%) is only slightly better than if it were predicting solely the majority class.
- Accuracy p Value: The accuracy p-value was $1.152422 \times 10^{-200}$, indicating that the model's accuracy is statistically significant compared to random guessing.

Overall, these results indicate that the model trained on the original dataset may struggle to accurately classify minority classes, highlighting the need for further refinement.

4.3.2 Model 1b: Decision Tree Model with SMOTE-Enhanced Data

The Decision Tree model, trained on the SMOTE-transformed data, achieved the following performance metrics:

- Accuracy: The model achieved an accuracy of 21.39%, meaning it correctly predicted the class for approximately 21% of cases.
- Kappa: The Kappa statistic was 11.59%, suggesting that the model's predictions exhibit only minimal agreement with the true labels beyond what would be expected by chance.
- Confidence Interval for Accuracy: The 95% confidence interval for accuracy ranged from 21.24% to 21.54%.
- Accuracy Null: The null accuracy is around 23.65%, which is actually higher than the model's actual accuracy of 21.39%.

- Accuracy p Value: The p-value for accuracy is 1, indicating that the model's performance is not significantly better than random guessing.

After training and evaluating models on both the original (imbalanced) and SMOTE-transformed datasets, results indicated that models trained on the original, imbalanced dataset performed better, achieving higher accuracy compared to those trained on the SMOTE-transformed data.

4.3.3 Model 1c: Tuned Decision Tree Model (Original Data)

To improve Model 1a's performance, hyperparameter tuning was conducted by adjusting the complexity parameter (cp), which controls the depth of the decision tree to help prevent overfitting. An initial range of cp values from 0.001 to 0.1, incrementing by 0.005, indicated that smaller cp values produced better results. Based on this finding, the tuning search was refined to focus on a range from 0.0005 to 0.005 with increments of 0.0005, ultimately identifying 0.0005 as the optimal cp value.

Model 1c (Tuned Decision Tree Model), using cp = 0.0005, was then evaluated through cross-validation, resulting in the following performance metrics:

- Accuracy: The model achieved an accuracy of approximately 30.10%, indicating a modest improvement over the initial results.
- Kappa: The Kappa statistic was %14.88, showing minimal but increased agreement between the model predictions and true labels beyond random chance.
- Confidence Interval for Accuracy: The 95% confidence interval for accuracy ranged from 29.94% to 30.27%.
- Accuracy Null: The null accuracy was 23.65%, slightly lower than the model's actual accuracy, confirming that the tuned model provides a marginal advantage over always predicting the most common class.
- Accuracy p Value: With an exceptionally low p-value, the model's performance is statistically significant, indicating better-than-random results.

These results indicate that, while tuning has modestly improved the Decision Tree model's performance, it still struggles with predictive accuracy and remains only marginally better than a majority-class prediction strategy.

4.3.4 Model 2a: Random Forest Model (Original Data)

In Model 2a, the Random Forest algorithm was employed, leveraging its strength as an ensemble method capable of handling complex, non-linear relationships and capturing intricate interactions among features. By averaging the results of multiple decision trees, Random Forest reduces the risk of overfitting, a common issue with individual decision trees. This approach resulted in a more robust and reliable model, with the following performance metrics:

- Accuracy: The model achieved an accuracy of 36.91%
- Kappa: The Kappa statistic was 25.01%, indicating a modest level of agreement beyond chance.
- Confidence Interval for Accuracy: The 95% confidence interval for accuracy ranged from 36.74% to 37.09%, reflecting a stable model performance.
- Accuracy Null: The null accuracy, which represents the accuracy of always predicting the most frequent class, was 23.65%.
- Accuracy p Value: The p-value for accuracy was 0, indicating that the model's performance is significantly better than random guessing.

4.3.5 Model 2b: Random Forest Model with SMOTE-Enhanced Data

In Model 2b, which utilized the Random Forest algorithm with SMOTE-enhanced data, the following performance metrics were observed:

- Accuracy: The model achieved an accuracy of 35.49%, indicating that it correctly predicted the crime types approximately 35% of the time.
- Kappa: The Kappa statistic was 24.78%, suggesting a modest level of agreement between the predicted and actual crime categories beyond what would be expected by chance.
- Confidence Interval for Accuracy: The 95% confidence interval for accuracy ranged from 35.32% to 35.66%, reflecting a stable model performance.
- Accuracy Null: The null accuracy, which represents the accuracy of always predicting the most frequent class, was 23.65%.
- Accuracy p Value: The p-value for accuracy was 0, indicating that the model's performance is significantly better than random guessing.

4.4 Model Comparison and Final Selection

Table 3 provides a comparison of performance metrics across all developed models. After assessing the performance of the five models, Model 2a, the Random Forest model trained on the original data, was selected as the final model. While SMOTE was applied in Model 2b to address class imbalance, it did not deliver a substantial improvement over the original model. Consequently, Model 2a was prioritized for its robust, reliable results without data augmentation, maintaining both data integrity and simplicity in model interpretation.

Table 3
Comparison of Model Performance Metrics

Model	Description	Accuracy (%)	Kappa (%)	Accuracy Lower (%)	Accuracy Upper (%)	Accuracy Null (%)	Accuracy P-Value
Model 1a	Decision Tree	26.04	5.80	25.88	26.20	23.65	0
Model 1b	Decision Tree, SMOTE	21.39	11.59	21.24	21.54	23.65	1
Model 1c	Decision Tree, Tuned	30.10	14.88	29.94	30.27	23.65	0
Model 2a	Random Forest	36.91	25.01	36.74	37.09	23.65	0
Model 2b	Random Forest, SMOTE	35.49	24.78	35.32	35.66	23.65	0

4.4.1 Variable Importance

Random Forest provides feature importance scores, highlighting the variables that most affect crime type prediction. Figure 18 shows these values based on Model 2a. The MeanDecreaseGini scores indicate which features are most important, with higher values reflecting greater influence on the model's predictions. The analysis reveals that geographic location (latitude and longitude) plays the largest role in predicting crime types.

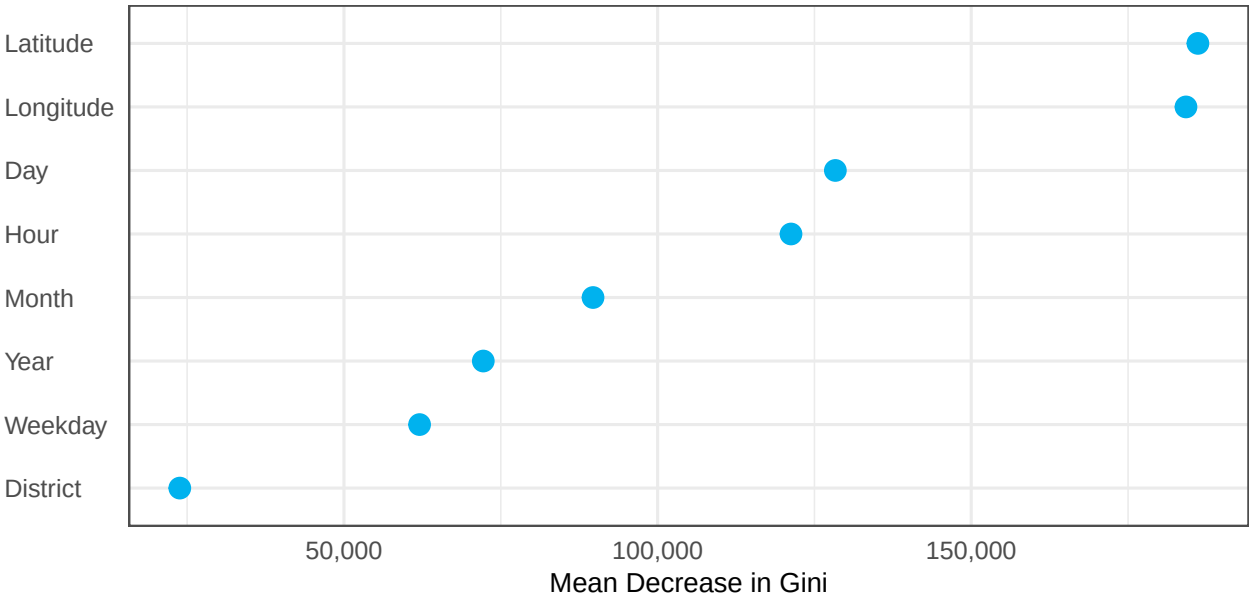

Figure 18: Feature Importance for Model 2a

4.5 Final Model Performance and Generalization

When evaluated on the test set, the model demonstrated an accuracy of **36.90%**, reflecting its ability to correctly predict approximately 37% of crime instances. The Kappa statistic was **24.99%**, indicating a moderate level of agreement between the predicted and actual crime categories, beyond what would be expected by chance. The confidence intervals for accuracy range from **36.74% to 37.05%**, providing further insight into the model's stability and consistency across different test data subsets.

The class-wise metrics for Model 2a are summarized in Table 4, which provides important performance indicators for each crime type. These metrics, including sensitivity, specificity, and F1-score, offer a detailed view of how well the model performs across different crime categories. Areas with lower sensitivity or specificity could indicate challenges in classifying specific crime types, which might be due to insufficient data.

Table 4
Class-Wise Metrics for Model 2a

Class	Sensitivity (%)	Specificity (%)	Precision (%)	F1 (%)	Balanced Accuracy (%)	Prevalence (%)
All Other Offenses	71.06	76.63	48.50	57.65	73.84	23.65
Other Assaults	35.38	81.80	25.49	29.63	58.59	14.96
Thefts	47.22	89.77	42.56	44.77	68.49	13.83
Vandalism/Criminal Mischief	15.48	93.22	21.69	18.07	54.35	10.83
Theft from Vehicle	27.66	93.58	30.39	28.96	60.62	9.21
Narcotic / Drug Law Violations	33.52	95.75	38.71	35.93	64.63	7.42
Fraud	12.09	97.46	23.83	16.04	54.78	6.17
Recovered Stolen Motor Vehicle	5.71	98.77	19.97	8.88	52.24	5.12
Burglary Residential	5.72	98.79	20.31	8.93	52.26	5.10
Aggravated Assault No Firearm	5.74	99.52	31.74	9.72	52.63	3.72

The Sensitivity of the model, which measures the proportion of true positives, varies significantly across crime types, with "All Other Offenses" having the highest sensitivity at 71.06%, indicating that the model is better at identifying these crimes compared to others. In contrast, "Recovered Stolen Motor Vehicle" has the lowest sensitivity at 5.71%, suggesting that the model struggles to detect this crime type.

Specificity, which measures the proportion of true negatives, is relatively high across all classes, with "Aggravated Assault No Firearm" showing the highest specificity at 99.52%, indicating that the model is good at correctly identifying non-occurrences of this crime.

Precision, which reflects the proportion of positive predictions that are correct, ranges from a low of 20.31% for "Burglary Residential" to 48.50% for "All Other Offenses".

The F1 score is a performance metric that combines precision and sensitivity (also known as recall) into a single value, providing a balanced measure of a model's ability to correctly identify positive instances. "All Other Offenses" has the highest F1 score of 57.65%. On the other hand, "Recovered Stolen Motor Vehicle" has an F1 score of 8.88%, indicating poor performance.

Balanced Accuracy, which accounts for both sensitivity and specificity, ranges from 52.24% for "Recovered Stolen Motor Vehicle" to 73.84% for "All Other Offenses", showing that the model performs better in balancing true positive and true negative rates for the more prevalent crime types.

Finally, Prevalence, which shows the proportion of each class in the data, is highest for "All Other Offenses" (23.65%) and decreases for less frequent classes such as "Aggravated Assault No Firearm" (3.72%), which is consistent with the imbalance observed in the dataset.

In conclusion, the model performs better at identifying more frequent crimes, such as "All Other Offenses" and "Thefts", while struggling with less frequent crimes.

4.5.1 Confusion Matrix

The confusion matrix visualized in Figure 19 illustrates the performance of Model 2a across different crime categories. Each cell in the matrix represents the frequency of predicted versus actual crime types, with the intensity of the red color indicating the prediction count. Diagonal cells, where predicted and actual labels match, indicate correctly classified instances. As can be seen in this figure, misclassification remains prevalent in certain lower-frequency classes, indicating areas where the model may be struggling, likely due to limited distinctiveness among categories.

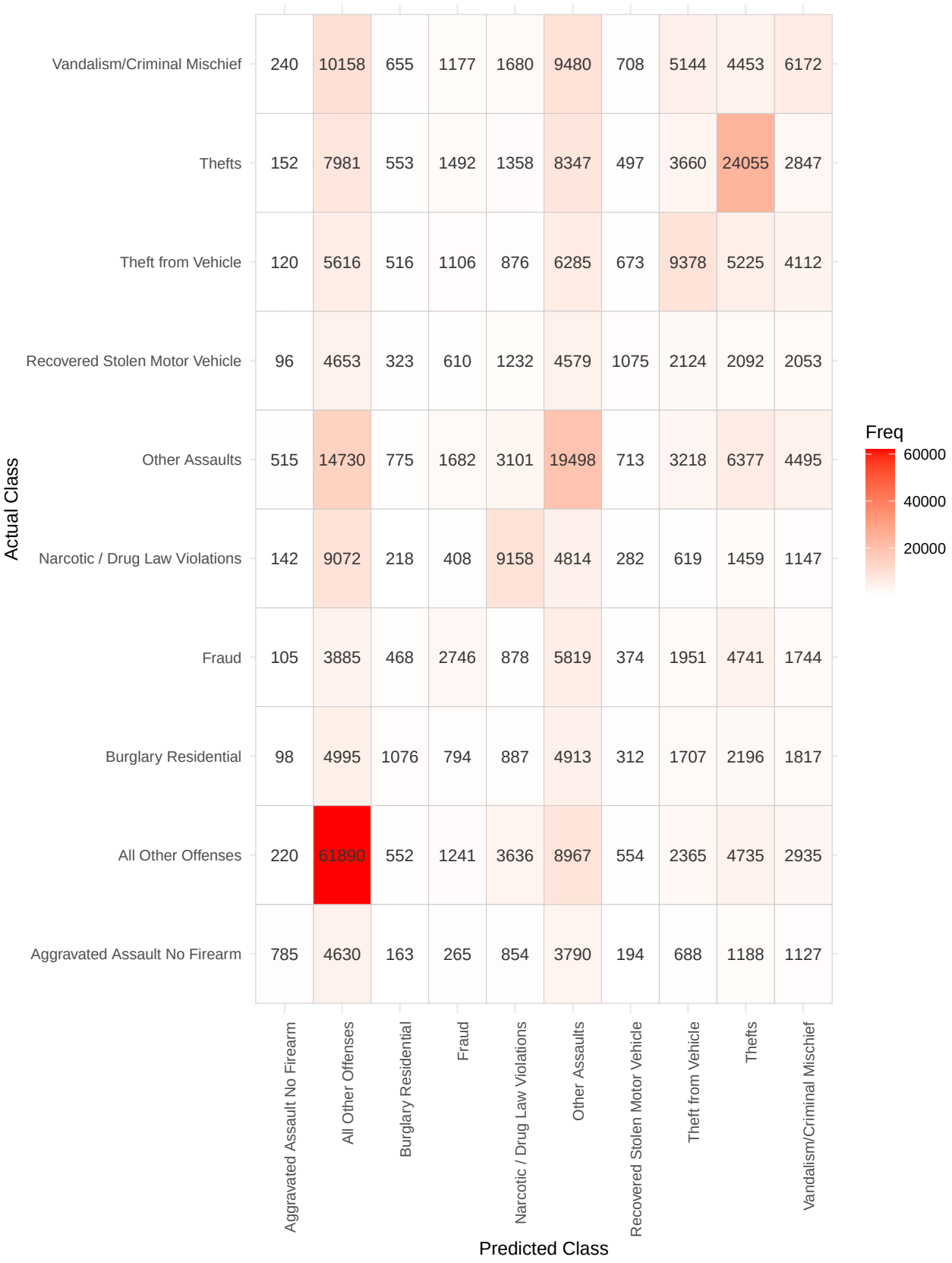

Figure 19: Confusion Matrix for Model 2a

4.5.2 Receiver Operating Characteristic (ROC) Curve

Figure 20 presents the Receiver Operating Characteristic (ROC) curve for the Random Forest model (Model 2a) on the test set, showcasing its performance in a multiclass classification scenario using the One-vs-Rest approach. Here, each class is evaluated individually by treating it as the positive class while considering all other classes as negative. The plot illustrates the model's ability to distinguish each class, with varying levels of Area Under the Curve (AUC) for each ROC curve. Overall, the model exhibits limited success in predicting certain classes effectively, indicating room for improvement in capturing distinct features across classes.

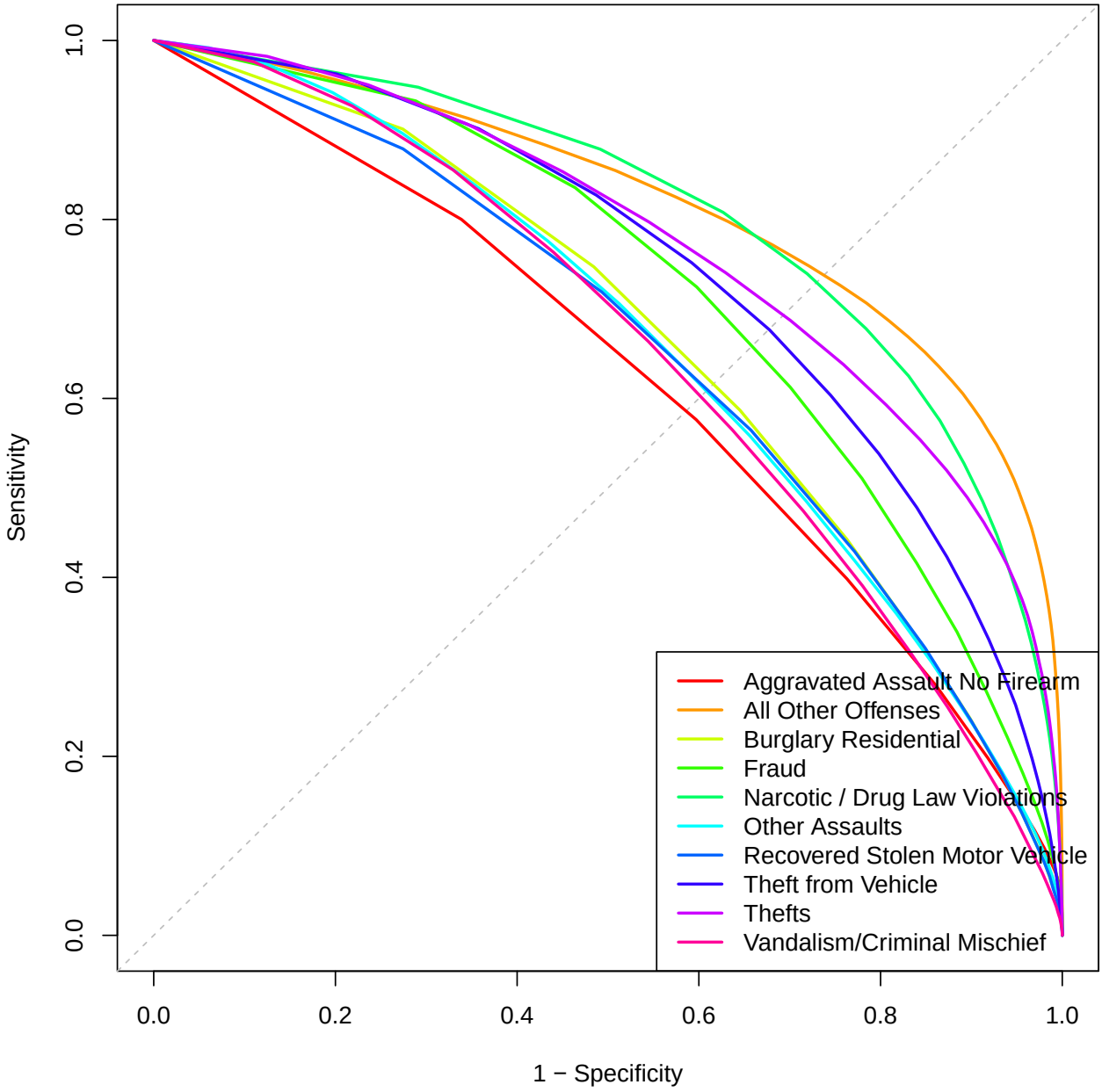

Figure 20: One-vs-Rest ROC Curves for Each Class

5 Discussion

This project aimed to explore, predict, and evaluate crime patterns in Philadelphia using machine learning and descriptive analytics. The key objectives were to explore trends in crime data, predict crime types, and evaluate the performance of the predictive model. Each of these objectives provided valuable insights into crime patterns, the effectiveness of machine learning in predicting crime types, and the factors that most influence the accuracy of crime classification.

The first objective of the project was to investigate crime patterns in Philadelphia through descriptive analytics. By analyzing crime data across various dimensions—such as location, time, and crime type—we were able to identify several notable trends. For instance, certain areas of the city exhibited higher crime frequencies, which aligns with prior knowledge of urban crime distribution. Temporal patterns, such as increased crime rates during specific months or times of day, were also evident. These insights can inform city planners, law enforcement, and community leaders about areas and times that require increased attention and resources. However, the exploratory analysis also highlighted some limitations, including potential biases in the dataset and the need for more granular geographic data to better capture crime hotspots.

The second objective was to develop a machine learning model to predict crime types based on factors such as time and location. Using features such as the time of day, day of the week, and geographic location, the model aimed to classify the type of crime. After training and tuning multiple models, Random Forest emerged as a strong performer in terms of predictive accuracy, providing valuable insights into the relationships between features and crime types. Notably, latitude and longitude were significant predictors, with certain crime types being more prevalent in specific neighborhoods.

The third objective was to evaluate the performance of the predictive model and identify the key features contributing to accurate crime classification. The model's performance was assessed using metrics such as accuracy, precision, recall, and F1 score, as well as more advanced techniques like ROC curves. The results indicated that the Random Forest model achieved a moderate level of accuracy, particularly for the more frequent crime types. However, its performance varied across different crime categories, with some classes being more difficult to predict than others.

5.1 Limitations

A key limitation of this project was the computational constraints of the hardware used for model training and analysis. Due to limited processing power, the dataset was sampled when visualizing the spatial distribution of crime density across districts, and certain computationally demanding processes, such as extensive cross-validation for parameter tuning, were restricted. These constraints may have limited the full potential of the Random Forest model, which typically benefits from fine-tuning over large parameter grids to optimize its predictive capabilities. Future iterations could use more powerful hardware to conduct a comprehensive analysis, potentially enhancing model performance.

6 Conclusion

This project offered a comprehensive exploration of crime patterns in Philadelphia, showcasing the practical application of machine learning for predicting crime types. By leveraging the Random Forest model—chosen for its robustness and ability to capture complex feature interactions—I achieved reasonable accuracy across various crime types, particularly for more frequent categories. Looking ahead, refining the model through hyperparameter tuning, incorporating additional data, and experimenting with advanced techniques could further improve accuracy and support proactive crime prevention strategies.

Beyond refining predictive modeling techniques, this project provided an invaluable opportunity to address the challenges of working with large-scale data, developing models, and evaluating their real-world performance. This experience deepened my understanding of Random Forests, highlighting their strengths and limitations in tackling multiclass classification tasks.

The hands-on nature of this work also allowed me to apply and build upon the skills I developed through the Professional Certificate in Data Science. From data manipulation to predictive modeling and performance evaluation, I've grown significantly more confident in applying these techniques to practical problems. This process has been both challenging and fulfilling, reinforcing my passion for data science and its ability to uncover actionable insights.

Generative artificial intelligence tools, including ChatGPT (OpenAI, 2024), played a key role throughout the project. These tools enhanced efficiency by assisting with report drafting, clarifying concepts, and even troubleshooting occasional coding issues. By streamlining workflows and improving the clarity of analysis and presentation, they contributed meaningfully to the project's goal of delivering effective and interpretable crime data analysis.

7 References

1. Ahmed, M., Danti, A., Mohammed, M. A., Zeebaree, S. R., & Abdulkareem, K. H. (2021). *The role of machine learning algorithms for diagnosing diseases.* Journal of Applied Sciences, 13(5), 426–442. https://doi.org/10.3923/jas.2021.426.442

2. Chainey, S., & Ratcliffe, J. (2013). *GIS and crime mapping.* Wiley.

3. Daly, M., & Wilson, M. (2017). *Homicide: Foundations of human behavior.* Taylor & Francis.https://doi.org/10.4324/9780203789872

4. Kang, H., & Kang, J. (2017). *Prediction of crime occurrence from multi-modal data using deep learning.* PLoS ONE, 12(4), e0176244. https://doi.org/10.1371/journal.pone.0176244

5. Malleson, N., & Andresen, M. A. (2016). An exploratory study of crime: Examining lived experiences of crime through socioeconomic, demographic, and physical characteristics. *Urban Science, 4*(5), 1-20. https://doi.org/10.3390/urbansci4010005

6. OpenAI. (2024). ChatGPT [Large language model]. https://chatgpt.com

7. Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations.* RAND Corporation.

8. R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.1). R Foundation for Statistical Computing. https://www.r-project.org/

9. RStudio Team. (2024). *RStudio: Integrated Development for R* (Version 2024.09.0). RStudio, PBC. https://posit.co/download/rstudio-desktop/

10. Wikström, P. O. H., & Treiber, K. (2016). Social disadvantage and crime: A criminological puzzle. *American Behavioral Scientist, 60*(10), 1232-1259. https://doi.org/10.1177/0002764216657304