**Crime Analysis and Prediction in Washington, D.C.**

Alice Pascalev, Sabrina Pierre, and Abhimanyu Barun

Department of Data Science, The George Washington University

DATS 6501: Data Science Capstone

Dr. Abdi Awl

December 8, 2024

**Table of Contents**

[**Glossary of Terms 3**](#_xfidh26yzdyj)

[**Introduction 4**](#_egmk9k4pr9fp)

[Introduction and Background 4](#_u7gpqaqm23c6)

[Problem Statement 4](#_ou8tyorne8u8)

[Problem Elaboration 4](#_m3h99h1sczez)

[**Literature Review 5**](#_p1k605j02p9d)

[Relevant Research 5](#_rxfvtfic0dlb)

[**Methodology 7**](#_1skuf25ktjvb)

[Dataset and Source 7](#_1fga8g1a7q1o)

[Data Collection 8](#_r1fvnzy3k5nj)

[Data Cleaning and Preprocessing 8](#_pbkbbjbfrp2q)

[Data Modeling & Visualizations 9](#_rq28fyykl554)

[EDA and Statistics 9](#_vo2v87ymejeq)

[GIS 12](#_x0vrqpvr85ua)

[**Results & Analysis 12**](#_t29jcrdndihu)

[Machine Learning Modeling 12](#_85ft1ojrzy6z)

[Regression 12](#_tpea9j6vaijh)

[Classification 21](#_wagwev22gx5z)

[GIS Analysis 24](#_rofweio4mmil)

[Natural Language Processing 24](#_wstsb7krkp4h)

[**Conclusion 27**](#_26av02v0re15)

[Project Limitation 27](#_btrni9aw9h2m)

[Future Research 28](#_1c4zj07hjw61)

[**References 29**](#_rmgd09p4cyqg)

[**Appendix 30**](#_dvtoewmfh3be)

# Glossary of Terms

# 

# Introduction

## Introduction and Background

Over the past 15 years, Washington, DC has experienced significant changes in crime trends that are influenced by factors such as gentrification, public policy shifts, and community interventions. Crime is not just a series of isolated events; it’s a reflection of the broader sociological landscape of a city. Understanding these trends is critical for policymakers, community organizations, and citizens alike, as they seek to create safer neighborhoods and address underlying issues.

## Problem Statement

This project aims to analyze crime trends in Washington, DC over a 15-year period using historical data, online public discourse sentiment, and geographical visualizations. The ultimate goal is to identify geographic hotspots, understand public sentiment, and predict future trends, to help law enforcement agencies understand crime dynamics and improve current strategies.

## Problem Elaboration

Crime is a persistent issue that affects the safety and well-being of communities across major cities, including Washington, D.C. Law enforcement agencies strive to prevent and reduce crime, but their efforts are often hindered by limited resources and the unpredictable nature of criminal activity. To effectively allocate resources and reduce crime, it’s essential to predict future crime trends accurately, identify geographic hotspots, and understand public sentiment surrounding crime-related issues. However, this task is complex due to the dynamic interplay of various factors such as location, time, and public perceptions.

By leveraging machine learning and statistical analysis, the project will identify patterns and anomalies in crime data, and by doing so, assist stakeholders in developing more effective, targeted interventions to reduce crime and improve public safety.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Literature Review

## Relevant Research

Crime has been increasingly influenced by a variety of environmental, social, and technological factors. Browning et al. (2017) suggest that neighborhoods with stronger ecological networks tend to experience lower levels of both violent and property crime. Their study poses that stronger community ties and neighborhood cohesion may have a mitigating effect on crime, highlighting the importance of environmental and social factors in understanding crime patterns. In Washington D.C., crime trends have shown significant fluctuations over the years, with a sharp increase in 2023, marking the largest crime spike compared to other cities, followed by a decrease in 2024 (Open Data DC, 2024). This dramatic increase in crime during 2023 underscores the dynamic nature of crime trends and the need for timely, data-driven analyses to understand the underlying causes of such shifts.

Several factors have been suggested as contributors to crime trends in Washington D.C. over the years. The Washington Examiner (2023) points to a combination of socio-political elements, including gentrification, a shrinking police force, and the impact of the D.C. crime lab’s loss of accreditation from 2021 to 2023, which may have contributed to delayed investigations and unresolved cases. Additionally, pandemic-related backlogs in case processing and changes in law enforcement policy have been identified as factors influencing crime patterns during the COVID-19 period. These elements may have further complicated efforts to maintain public safety and predict crime trends accurately.

LightGBM, a gradient boosting framework, is particularly well-suited for our crime prediction models due to its ability to handle large, high-dimensional datasets efficiently while capturing complex interactions. Its unique capability to process categorical variables, such as crime type and ward, directly without requiring extensive preprocessing like one-hot encoding, is especially advantageous. This feature allows LightGBM to better uncover underlying patterns in our dataset, which includes features such as time, geographic location, and crime types. Moreover, LightGBM excels at modeling intricate, non-linear relationships between variables—an essential capability for crime prediction, where factors like time, location, and crime type interact in complex ways. Research by Ke et al. (2017) has demonstrated LightGBM’s superior performance over other machine learning models, including XGBoost, in terms of training speed, memory efficiency, and predictive accuracy. These advantages make LightGBM an optimal choice for analyzing crime trends in large datasets, where both computational efficiency and high predictive performance are critical.

In recent years, advances in technology have introduced new tools for analyzing crime data. One such tool is Natural Language Processing (NLP), which has proven to be particularly useful in extracting insights from unstructured data, such as social media posts. Skeem et al. (2021) highlight the application of NLP techniques, including sentiment analysis and topic modeling, to analyze public sentiment and crime trends through social media platforms. This innovative approach allows for a more granular understanding of public perception and can aid in predicting crime hotspots based on social media conversations. By combining these new technological tools with traditional crime data, researchers are beginning to develop more accurate models for forecasting crime trends and understanding the factors that drive crime in urban areas.

# Methodology

## Dataset and Source

Open Data DC is Washington, D.C.'s open data portal, providing access to datasets that cover a wide range of topics related to the city. This data is publicly available for research, analysis, and transparency initiatives.

Users can explore datasets across multiple categories, including public safety, health, education, transportation, and more. The platform allows for data browsing, downloading in various formats, and integration into applications via APIs. This accessibility enables stakeholders to analyze information, develop data-driven solutions, and gain insights into the city's operations and services.

We utilized the Crime Incidents data sets, which provide police report data for 2008 (“Crime Incidents in 2008”) through September 2024 (“Crime Incidents in 2024”).

The attributes are as follows:

CCN Report ID

REPORT\_DATE

SHIFT

METHOD

OFFENSE

BLOCK

XBLOCK

YBLOCK

WARD

ANC

DISTRICT

PSA

NEIGHBORHOOD\_CLUSTER

BLOCK\_GROUP

CENSUS\_TRACT

VOTING\_PRECINCT

LATITUDE

LONGITUDE

BID

START\_DATE

END\_DATE

## 

## Data Collection

Data collected from OpenData DC was all data from January 1, 2008 (“Crime Incidents in 2008”) through the end of September 2024 (“Crime Incidents in 2024”), with over 554,000 rows of data.

These datasets were downloaded in both CSV and shapefile formats. Census tracts shapefiles for the years 2010 and 2020 were also pulled from OpenData DC; specifically the TRACT and P0010001 fields were used to join according to attribute with crime incidents datasets in ArcGIS Online.

## Data Cleaning and Preprocessing

Initially, individual datasets for each year were loaded as separate Pandas DataFrames. To ensure consistency before merging, column names across datasets were compared. After validating the structure, all datasets were concatenated into a single comprehensive DataFrame (df\_combine), consolidating the data for further analysis.

To prepare the data for geospatial analysis, a subset of location-related columns, including BLOCK, XBLOCK, YBLOCK, WARD, and others, was extracted into a new DataFrame (df\_location\_geog). Missing values in this subset were addressed by dropping rows containing NaN. The cleaned subset was then used to create mapping dictionaries that link (XBLOCK, YBLOCK) coordinates to specific geographical attributes (BLOCK, WARD, and ANC). These mappings were employed to impute missing values in the main DataFrame, ensuring a more complete and reliable dataset

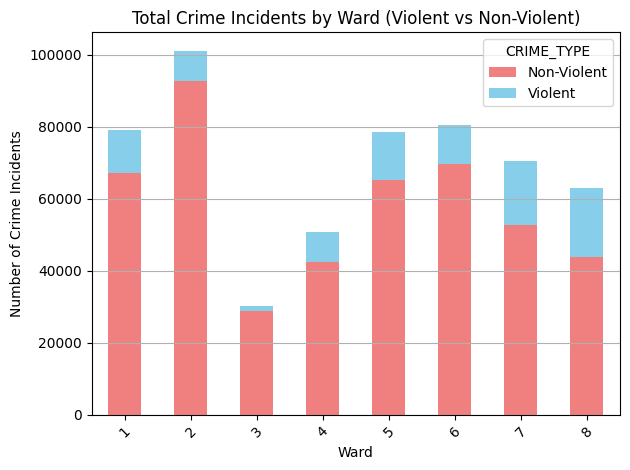
Additional steps included inspecting the combined dataset for structure and integrity using .info() and grouping data by geographical columns to identify patterns or validate the imputations. This workflow not only standardized the dataset but also leveraged existing relationships in the data to minimize the impact of missing values.

The resulting dataset is well-prepared for geospatial and temporal analysis, with consistent columns and minimal missing values. Further validation steps, such as checking for anomalies in imputed values and optimizing performance, could enhance the workflow’s robustness.

## Data Modeling & Visualizations

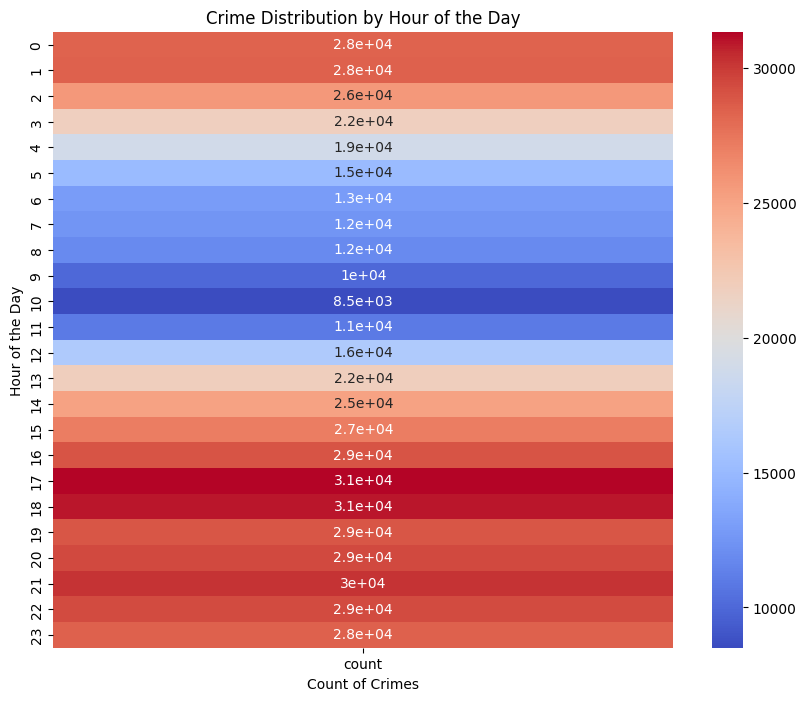
### EDA and Statistics

**Figure 1**



*Note.* This figure shows the makeup of violent versus non-violent crime according to ward, from 2008 to 2023.

**Figure 2**



*Note.* This temporal heatmap



Over the course of the 15-year period analyzed, theft accounted for 76% of all reported crimes, making it the most prevalent crime type. In terms of geographic distribution, Wards 2, 6, and 1 had the highest levels of overall crime, in that order. Conversely, Wards 7 and 8 recorded the highest rates of violent crimes (figure x). Looking at trends over time (figure xxx), the years 2014, 2015, and 2016 saw the highest crime volumes. However, from 2008 to 2023, overall crime decreased by 0.34%, indicating a slight downward trend. A more significant decrease of 10.78% was observed from 2014 to 2023, with 2014 being the peak year for crime. In contrast, the period from 2020 to 2023, which overlaps with the onset of the COVID-19 pandemic, experienced a notable 22.38% increase in crime, suggesting a possible impact of the pandemic on crime rates. Looking further into types of crime, the homicide rate experienced a significant increase in the 15 year period at nearly 54%, with Wards 7 & 8 seeing a little over 58% of homicides in Washington, D.C. Regarding trends across time, summer and fall were the months that saw the most crime throughout the years with the hours 5pm, 6pm, and 9pm seeing the most crime (figure xx).

### GIS

The ArcGIS Story Map titled DC Crime 2008-2023 (Pierre et al., 2024) presents a comprehensive spatial visualization of crime shifts in Washington, D.C. Heat Maps are used to indicate shifts in crime hotspots across the district and highlight constant areas of high crime density. Choropleth maps show the frequency of crime occurring within tracts according to a set number of inhabitants to accurately point out where inhabitants are experiencing the most crime taking into account the distribution of the population. Dot density maps highlight shifts in the frequency of certain types of crime across the city as well as in specific wards; these maps can help to identify hotspots for certain types of crimes.

### 

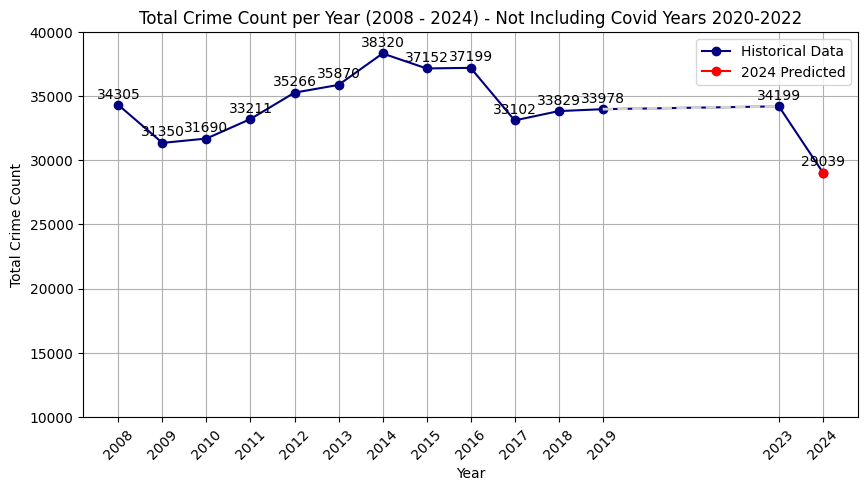
# Results & Analysis

## Machine Learning Modeling

### *Regression*

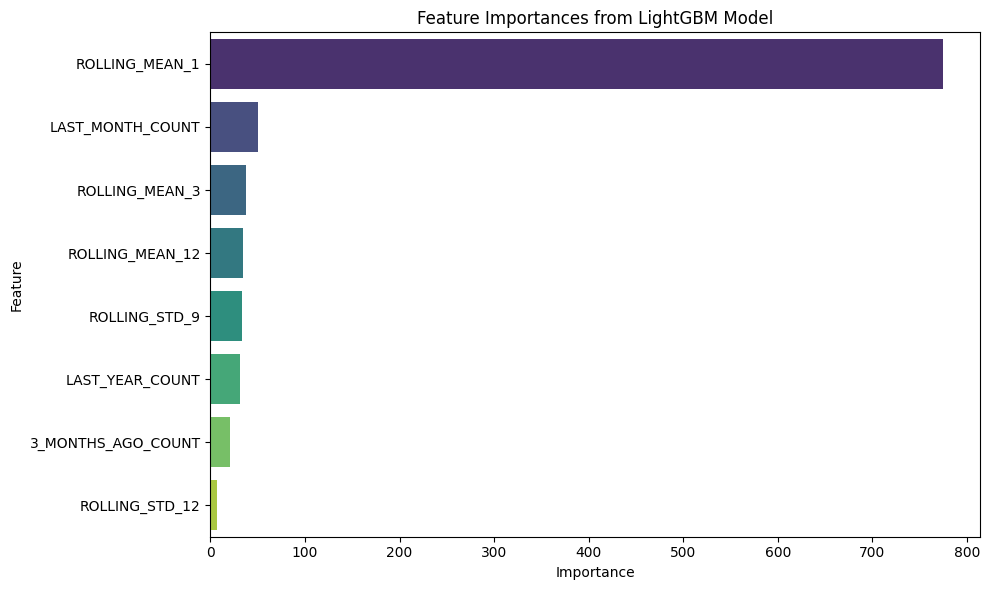
Three regression-based machine learning models were employed to predict crime-related outcomes through the end of 2024. Model 1 predicts total crime counts for the remaining months of 2024 (October, November, and December). Model 2 focuses on predicting the distribution of crimes across different wards in the city, again extending through the end of 2024. Model 3 is designed to forecast the distribution of crime types (e.g., theft, violent crime, etc.) for the same time period.

All three models utilize the LightGBM regression model due to its numerous advantages. Studies, such as Ke et al. (2017), have shown that LightGBM outperforms other models like XGBoost in terms of training speed, memory usage, and predictive accuracy, which is particularly advantageous for predicting crime trends in large datasets.



*Model 1: Predicting Crime Counts.*

Model 1 forecasts crime counts for October, November, and December 2024, with projected values of 2,542, 2,591, and 2,652, respectively. Adding these projections to the actual crime counts recorded through September 30, 2024, the model predicts a sharp decline in overall crime, positioning 2024 as potentially the year with the lowest total crime count in the 15-year period covered by our dataset.

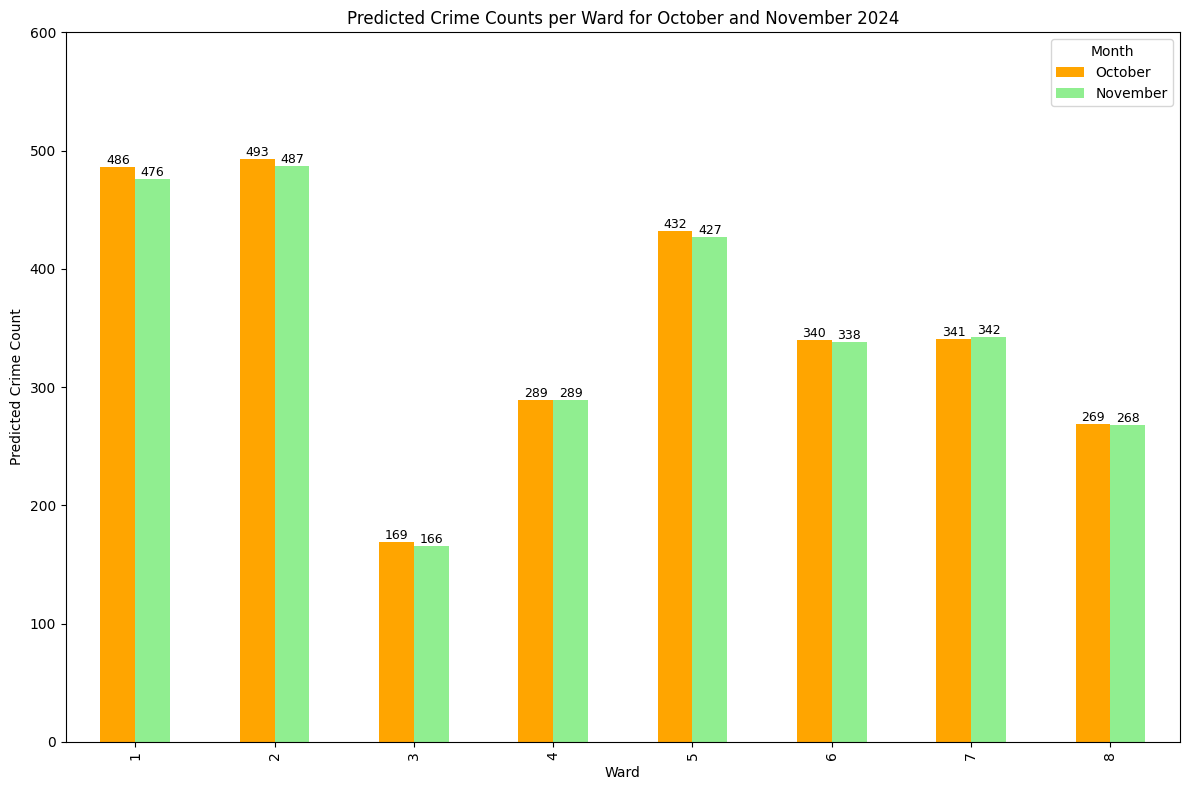


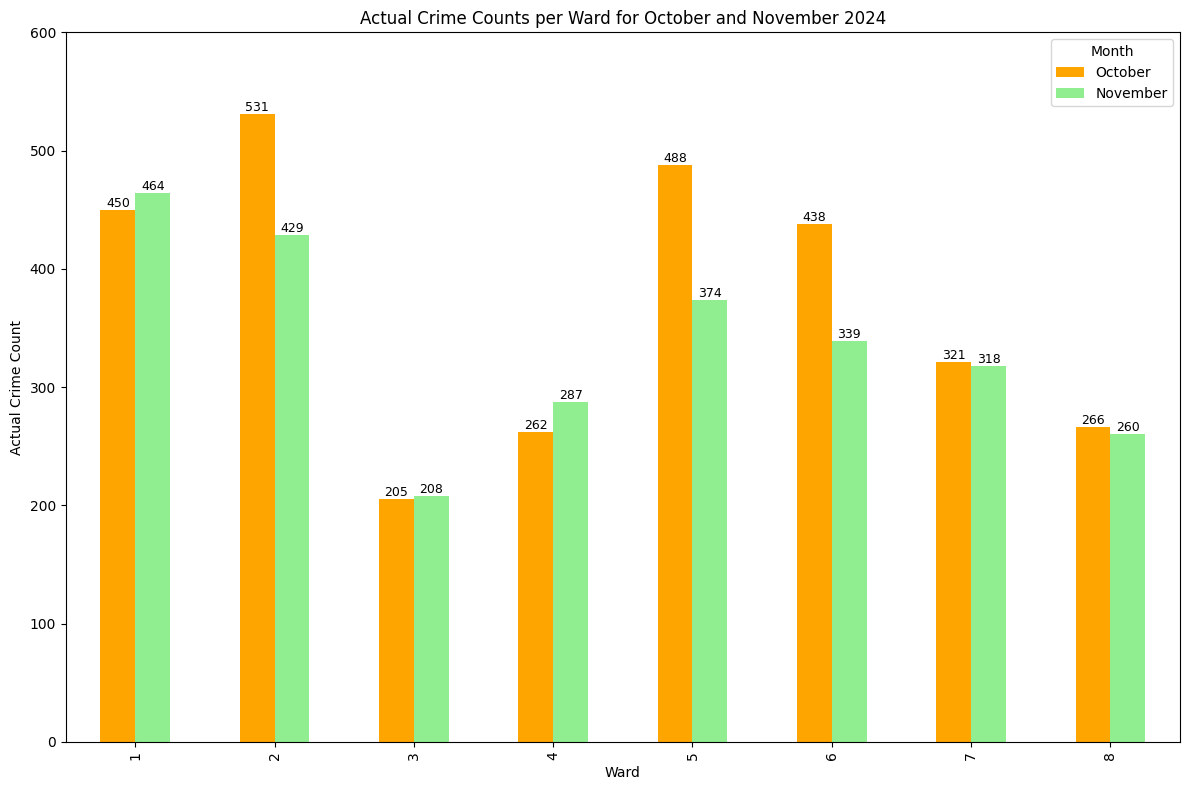
The model incorporates lag features (last month's count, three months ago, last year's count) to capture short-term spikes and seasonal trends, as well as rolling averages and rolling standard deviations to smooth out noise while capturing variability and long-term patterns in crime data. The rolling statistics used include rolling\_mean\_1 (past month’s average), rolling\_mean\_3, and rolling\_mean\_12, as well as rolling standard deviations (rolling\_std\_9 - STD for the past 9 month period, rolling\_std\_12 - STD for the past 12 month period) to track fluctuations in crime rates.

For hyperparameter selection, we prioritized efficiency while maintaining the complexity needed to handle the rolling and lag features. The selected hyperparameters aim to balance computational cost with the model's ability to capture intricate crime patterns. For hyperparameter tuning, the selected parameters included N-Estimators, with values of 100, 200, and 300, to control the number of boosting rounds and balance simplicity with the ability to capture complex patterns. The learning rate was set to 0.01, 0.1, and 0.2, where the lower rate ensured stable learning and higher rates sped up training. The max depth ranged from 3 to 7, with a depth of 3 promoting generalization, 5 balancing complexity, and 7 capturing intricate feature interactions. Additionally, the number of leaves was set to 31, 50, and 100, allowing finer-grained partitions while managing the risk of overfitting. To ensure reliability and computational efficiency, GridSearchCV with 5-fold cross-validation was used for robust hyperparameter tuning, leading to a highly reliable and efficient model.

Evaluation metrics indicate strong model performance, with low average errors relative to the target values, which are in the tens of thousands. The model achieves a Mean Squared Error (MSE) of approximately 15,000, a Root Mean Squared Error (RMSE) of around 123, and an R² of 0.91, explaining 91% of the variance in crime counts. The Mean Absolute Error (MAE) is about 61, and the Mean Absolute Percentage Error (MAPE) is 2.6%, reflecting the model's high accuracy and predictive power.

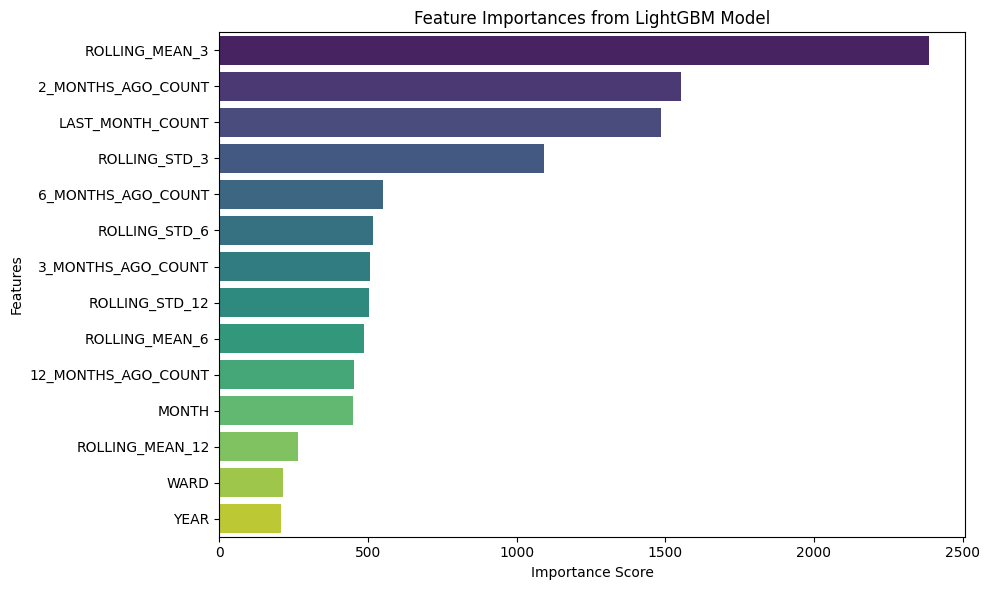
Our second model is a LightGBM regression model designed to predict crime distribution across wards. It aligns closely with our exploratory data analysis (EDA) findings, which highlighted Ward 2 as having the highest crime counts. However, the model provides new insights, predicting Ward 1 as the second-highest and Ward 5 as third-highest in crime counts, whereas EDA positioned Ward 6 and Ward 1 in these respective places. This discrepancy highlights the model's ability to uncover patterns not immediately evident in the raw data.





*Model 2: Predicting Crime Distribution across wards.*

The model's predictions for October and November align closely with the actual ward-level crime distributions, as visualized in the accompanying plot.

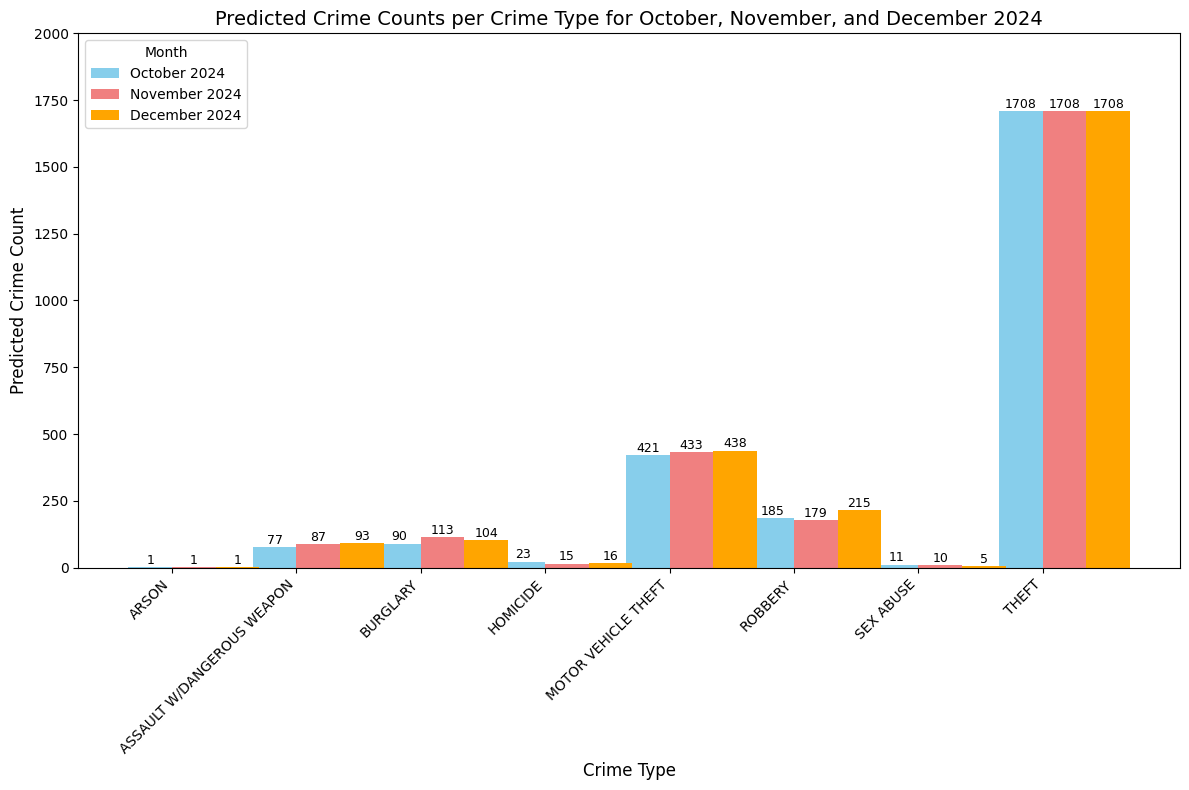


The model incorporates a robust feature set, including lag and rolling features, along with temporal and spatial indicators such as month, ward, and year. To balance computational efficiency and complexity, we carefully tuned hyperparameters using GridSearchCV with 5-fold cross-validation, ensuring reliable performance without excessive computational cost.

Key hyperparameter choices for the model were carefully selected to optimize performance while managing complexity. The default number of leaves (31) was used as a baseline, with values of 50 and 70 enabling the model to better capture non-linear relationships while mitigating the risk of overfitting. For the number of estimators, 500 and 1000 were tested, with 1000 ultimately chosen to effectively capture nuanced trends from lag and rolling features. Learning rates of 0.01 and 0.03 provided stability when using a large number of trees, while a rate of 0.05 expedited training but introduced a slightly higher risk of overshooting. Max depth values of 5, 10, and 15 were evaluated, with a depth of 10 offering an optimal balance between interpretability and complexity. Cross-validation with 5 folds was employed as a best practice to ensure reliable validation without imposing excessive computational demands.

The model demonstrates strong predictive accuracy, with errors that are relatively small compared to target values in the hundreds. Key evaluation metrics include an RMSE of 17.93, indicating an average deviation of approximately 18 units; an MSE of 321.71, reflecting low variability in squared errors; and an MAE of 13.21, showing an average error of about 13. Additionally, the model achieved an R² value of 0.9769, explaining 97.69% of the variance in the target variable, and a MAPE of 3.88%, indicating that predictions are off by less than 4% on average.

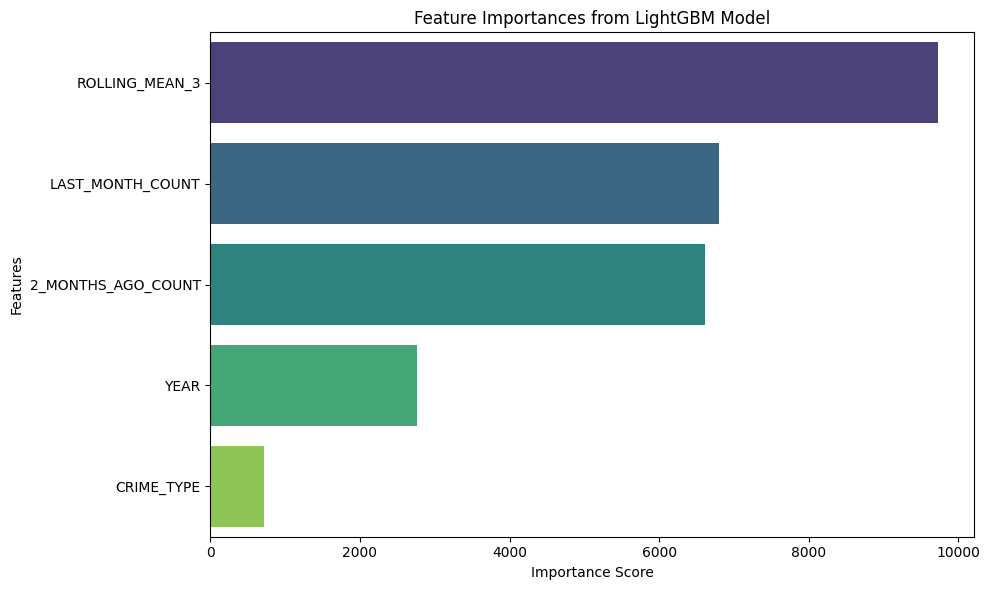
Our final model is a LightGBM regression model predicting monthly crime counts by category.





*Model 3: Predicting Crime Types.*

Consistent with EDA findings, theft remains the most prevalent crime, followed by motor vehicle theft and robbery. The model effectively replicates real-world patterns, with its predicted distributions closely matching actual data.



This model incorporates a smaller feature set, including two lag features (last\_month\_count and 2\_months\_ago\_count), one rolling feature (rolling\_mean\_3), year, and crime category. While it performs slightly weaker due to rare crimes such as arson (1–2 counts) and sex abuse (low teens), these higher errors are acceptable as these categories do not significantly influence broader trends.

Key hyperparameter selections were tailored to balance complexity, computational efficiency, and overfitting risks. The number of leaves was set to the default of 31 as a baseline, with 50 and 70 providing additional flexibility to capture non-linear relationships while managing overfitting. For N-Estimators, values ranged from 200, which was useful for quick evaluations, to 500, which balanced complexity, and up to 1000, which captured nuanced trends introduced by lag and rolling features. The learning rate ranged from 0.01 to 0.05, where lower rates ensured stability in training with many trees, and higher rates accelerated training while slightly increasing the risk of overshooting optimal solutions. Max depth values of 5, 10, and 15 were tested, with 5 promoting generalization, 10 balancing interpretability and complexity, and 15 enabling the model to learn from complex interactions, particularly involving lag and rolling features. Hyperparameter tuning was conducted using GridSearchCV with 5-fold cross-validation, a widely accepted standard that provided reliable validation results without imposing excessive computational demands.

Evaluation metrics indicate that the model achieved strong predictive performance. It attained an R² of 0.9927, explaining nearly 99.3% of the variance in the data, an RMSE of 53.85, reflecting moderate deviations for predictions in the hundreds, and an MAE of 21.17, representing a small average error. The Mean Absolute Percentage Error (MAPE) was 10.72%, signifying an acceptable level of relative error. Despite slightly higher errors in rare categories, these results confirm the model’s ability to reliably predict overall trends and patterns in the crime data.

### Classification

The dataset earlier underwent a rigorous preprocessing phase to address missing values through imputation or removal. The next steps of data preprocessing for classification models involve several essential steps to prepare the dataset for model training. The target variable, "OFFENSE", was separated from the predictors, with columns like "METHOD" dropped as they were not relevant for prediction. Categorical features such as "SHIFT", "ANC", "VOTING\_PRECINCT", and "BLOCK" were label-encoded to transform them into numerical representations, while numerical features like "YEAR", "MONTH", "DAY", "HOUR", "LATITUDE", and "LONGITUDE" were retained in their original form for analysis.

To address the significant class imbalance between theft (majority class) and violent crimes (minority class), undersampling techniques were applied to the majority class to ensure the models could learn effectively from the minority class. The data was then split into training and testing subsets using an 80-20 split to facilitate robust evaluation. These steps ensured the dataset was clean, balanced, and formatted for compatibility with machine learning models, creating a strong foundation for reliable model training and testing.

Several machine learning models were evaluated using multiple machine learning models for crime prediction, focusing on their ability to classify theft (majority class) and violent crimes (minority class). Precision and recall were prioritized as evaluation metrics to assess how well each model balanced false positives and false negatives, especially given the class imbalance in the dataset.

Across the evaluated models, performance for theft was consistently strong, with precision values ranging from 0.96 to 0.97, indicating the models' ability to minimize false positives for the majority class. Recall for theft varied between 0.61 and 0.72, demonstrating that most models captured a substantial portion of true theft instances. For violent crimes, however, precision was significantly lower, ranging between 0.13 and 0.17, reflecting difficulties in accurately identifying true positives without overpredicting. Despite this, recall for violent crimes was relatively robust, ranging from 0.63 to 0.76, highlighting the models' strength in identifying a large proportion of true instances for this minority class.

Among the ensemble methods, Random Forest, Gradient Boosting (XGBoost), LightGBM, and CatBoost exhibited strong performance. For theft, these models achieved precision close to 0.97 and recall values between 0.61 and 0.72, demonstrating consistent accuracy in predicting the majority class. For violent crimes, recall reached up to 0.76 (as seen in LightGBM), indicating effective sensitivity to minority class instances. Random Forest and CatBoost provided balanced performance across both classes, while LightGBM emphasized recall for violent crimes, albeit at the expense of precision.

Baseline models such as Logistic Regression and Naive Bayes performed moderately well, achieving precision of 0.96 for theft but struggling with precision for violent crimes (0.14) and maintaining recall for violent crimes between 0.63 and 0.67. K-Nearest Neighbors (KNN) demonstrated similar patterns, offering reasonable recall for violent crimes (0.69) but low precision (0.14) for this class. These models, while interpretable and computationally efficient, were less capable of managing the inherent imbalance between classes compared to ensemble methods.

With recall - we want to ensure that offenses are detected at a higher rate, while some over-classifications may occur but with the intention to mitigate risks associated with undetected offenses. It is better that there are false positives than that one undetected offense.

In conclusion, ensemble models, particularly Random Forest, Gradient Boosting, and CatBoost, emerged as the most reliable for crime prediction, effectively balancing recall for violent crimes with high precision for theft.

The feature importance analysis highlighted key variables influencing crime predictions, with location, and time-related variables consistently ranking highest across models such as Random Forest and XGBoost. Pandemic-related factors also emerged as significant, reflecting their impact on public behavior and crime, particularly for violent crimes. These findings underscore the central role of geographic and temporal features in understanding crime trends.

A clear pattern showed a steep decline in importance after the top features, suggesting that only a subset of variables significantly contributes to model accuracy. Demographic and auxiliary features had lower importance, indicating potential for exclusion or refinement to improve efficiency. Models like XGBoost also revealed sensitivity to temporal patterns, offering deeper insights into crime dynamics.

## GIS Analysis

Heatmaps (Pierre et al., 2024) depicting the state of crime across our chosen 15 year period highlight times when crime was dense in all wards across the district except Ward 3. These visualizations showed that the years 2014 and 2015 saw the most density in the most areas, and the year 2020 saw less dense crime across less areas. Crime incidents were highly congregated in wards 2, 6 and 1 in comparison with other wards who had partially dense areas throughout the years.

Choropleth maps showed more crime is being experienced in 2023 per 1,000 inhabitants in more tracts/areas in D.C. compared to earlier years (e.g. 2008, 2014). Dot density maps showing violent crime by type displayed an overwhelming prevalence of Robbery across the city following major roads in wards 1, 2, 6, 7 and 8.

Looking at dot density maps of violent crime by type for these wards, the Shaw area is seeing more instances of violent crime as opposed to other neighborhoods within ward 1, in ward 8 close to all violent crimes occurred south of I-295, and wards 7 and 8 are experiencing more homicides in 2023 versus 2008. These maps also showed more homicides occurring in Wards 7 and 8 as confirmed by heatmaps showing the number and size of hotspots for this crime type in these wards increasing. Heatmaps generated for all theft crimes showed more hotspots in ward 6 appearing recently in 2023 as opposed to 2008.

## Natural Language Processing

### Data Preprocessing

The analysis involves processing Reddit post data sourced from The-Eye.eu, an open data repository, to perform sentiment analysis and derive insights from user-generated content. The dataset contains fields such as author, created\_utc, selftext, and title, which are cleaned and prepared through a structured preprocessing pipeline. The workflow ensures a consistent, feature-rich dataset, enabling a robust foundation for sentiment analysis.

The data preparation begins with cleaning and standardizing the created\_utc column, which is a UNIX timestamp. This column is converted to a numeric format, with invalid values coerced to NaN and subsequently dropped. The cleaned timestamps are then converted to a human-readable datetime format and decomposed into year, month, day, hour, and minute components. These granular features allow for detailed time-series analysis and the identification of trends over specific periods.

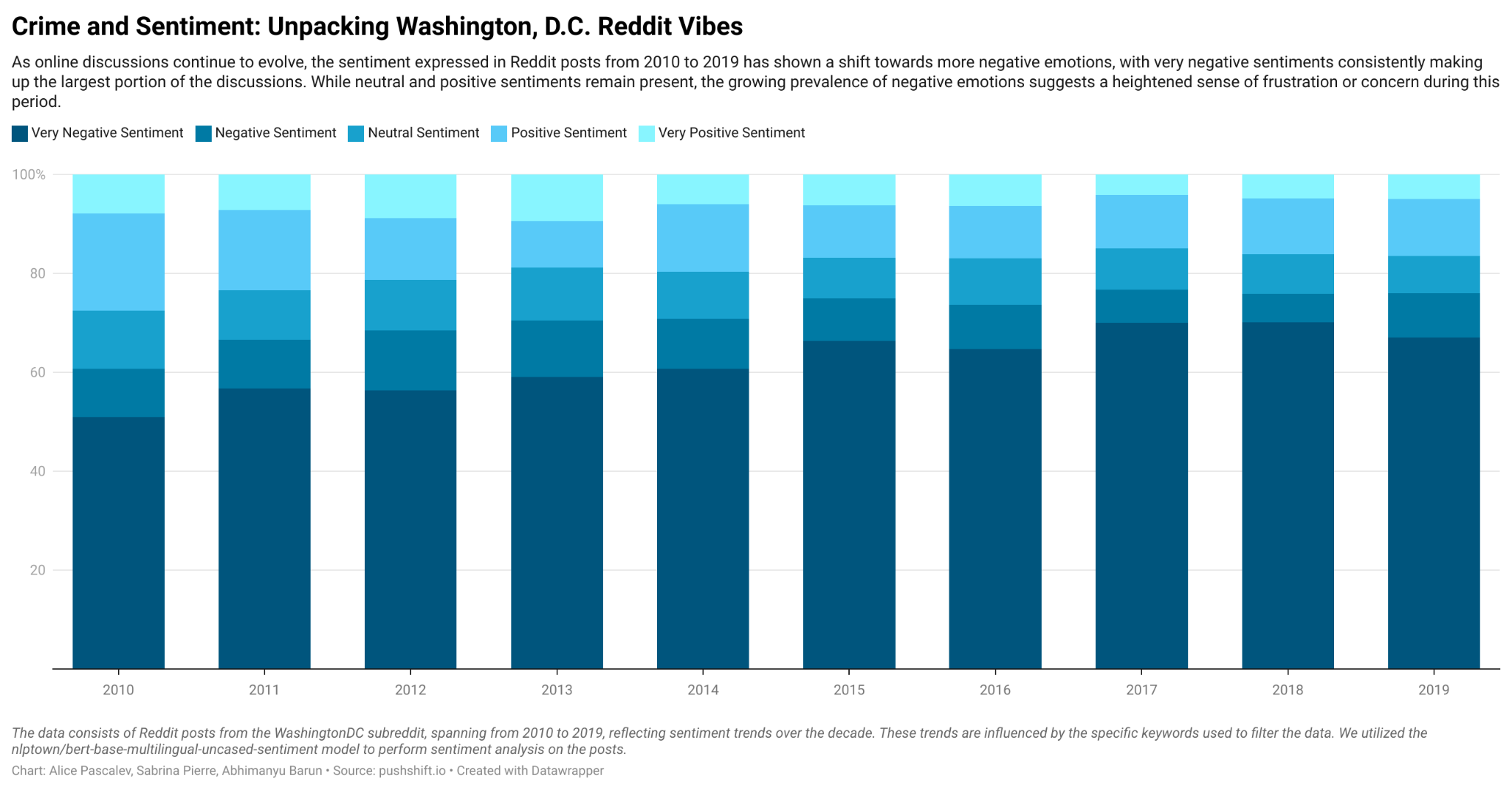
The dataset is further refined by filtering out years with insufficient post counts, such as 2008 and 2009. This ensures the dataset is robust and minimizes noise for downstream tasks. Post counts are grouped by year to assess overall activity trends, providing context for further analysis. Additionally, the selftext and title columns are combined into a single textual field, creating a unified input for sentiment classification.

### Sentiment Analysis

For sentiment analysis, the pre-trained model nlptown/bert-base-multilingual-uncased-sentiment is used. This model, known for its ability to classify sentiments across multiple languages, assigns sentiment scores ranging from very negative to very positive. Each post in the dataset is passed through the model, and the predicted sentiment is stored as a new column. This step enables sentiment trends to be analyzed across time, providing insights into user emotions and reactions.

The sentiment labels are aggregated and visualized to identify patterns and trends in user sentiments over different years. This analysis offers valuable insights into the emotional landscape of Reddit posts and how it evolves in response to specific events or topics.

In the Reddit discussions from the r/WashingtonDC subreddit between 2010 and 2019, we found a notable prevalence of negative emotions over the decade. Negative and Very Negative sentiments consistently dominate, suggesting a critical or concerned tone in the discussions. While Positive and Very Positive sentiments remain present, their relative proportions appear stable, indicating a smaller share of optimistic conversations. Neutral sentiments remain stagnant yet overshadowed by more polarizing sentiments.



# Conclusion

This research project sought out to examine crime patterns in Washington D.C. from 2008 to 2023 as well as predict future trends with the objective of gaining context and an understanding of how crime evolves. Machine learning techniques and statistical analyses were used to uncover sudden shifts and anomalies with the potential to inform leaders who aim to improve public safety.

During this period, violent crime experienced an increase with a slight decrease in overall crime. Despite the huge spike in crimes of all types from 2022 to 2023, both actual and predicted crime counts for 2024 show a decrease in overall crime for that year.

Overall, crime has undergone many spatial shifts in Washington D.C. The overarching trend is an increase in the density of violent crime towards the end of this period, with specific areas of interest emerging according to crime type. Regarding the most common crime, theft, wards 1, 2, 6, and 5 should be of focus for leaders, and for violent crimes such as homicide, wards 7 and 8 are of interest. The increase in people and spaces experiencing crimes within a D.C. tract is telling of the urgent need for interventions and policies. This increased frequency of crime is mirrored by our NLP findings which found growing negative sentiments regarding the state of crime in D.C. The gradual increase in negative sentiments could point towards growing frustrations or concerns among users, potentially tied to recurring topics like crime or urban challenges. These trends emphasize the need to explore specific issues driving such sentiments to better understand community concerns. These findings may also help reveal what crimes are most important to inhabitants as well as study public reception of policies.

## Project Limitation

There are several limitations to consider in this analysis. The lack of yearly census data may affect the accuracy of data joins used to create map visualizations. The shorter timespan of the dataset, covering only 15 years, could impact the accuracy and reliability of the predictive models. The inclusion of incomplete 2024 data may further skew results, leading to potential misinterpretations. Furthermore, imbalanced data has resulted in low recall and precision for certain models, highlighting a need for caution in interpreting results for less frequent crimes. Lastly, the absence of recent data in the analysis may fail to capture current public sentiment or shifts in crime-related discussions, such as those on Reddit, limiting the relevancy of findings.

## Future Research

Future research should aim to finalize predictions for the rest of 2024 and to further evaluate the effectiveness of chosen models. Additionally, other factors may be examined as drivers of crime such as gentrification, socioeconomic status, and changes in crime policies. Further research can also explore more subreddit pages based in the city to depict a more comprehensive outlook on the state of crime in D.C. by its inhabitants.

**Final Statement**

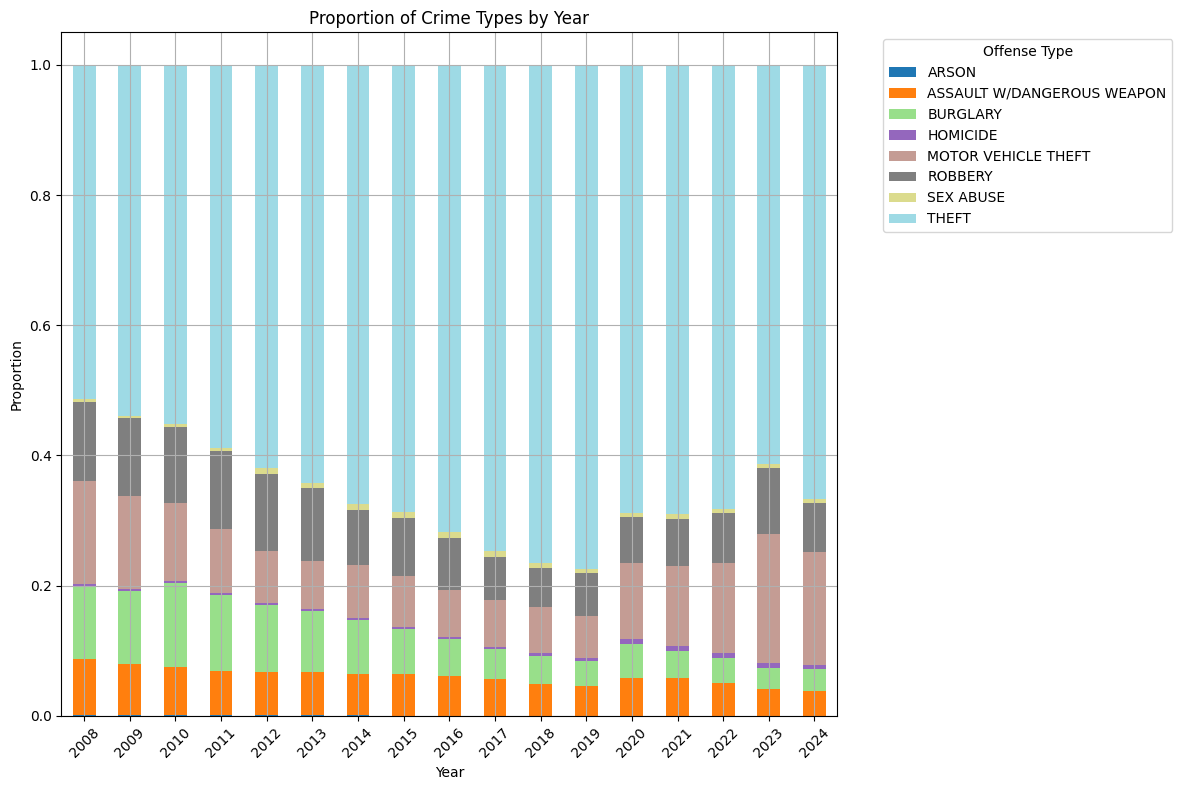
Understanding the state of crime in Washington D.C. is crucial to the well-being of its inhabitants and allows for the development of relevant and effective programs and interventions to mitigate crime, and particularly violent offenses, in areas where it is more needed.

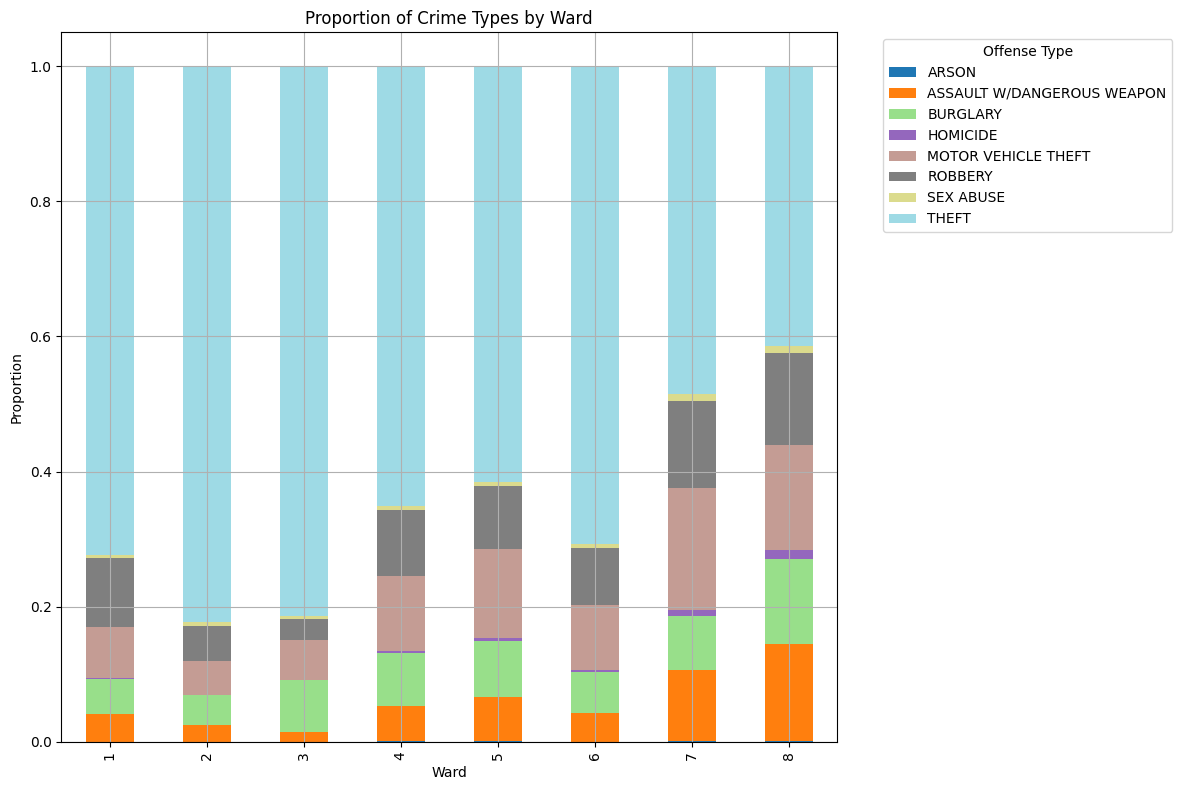
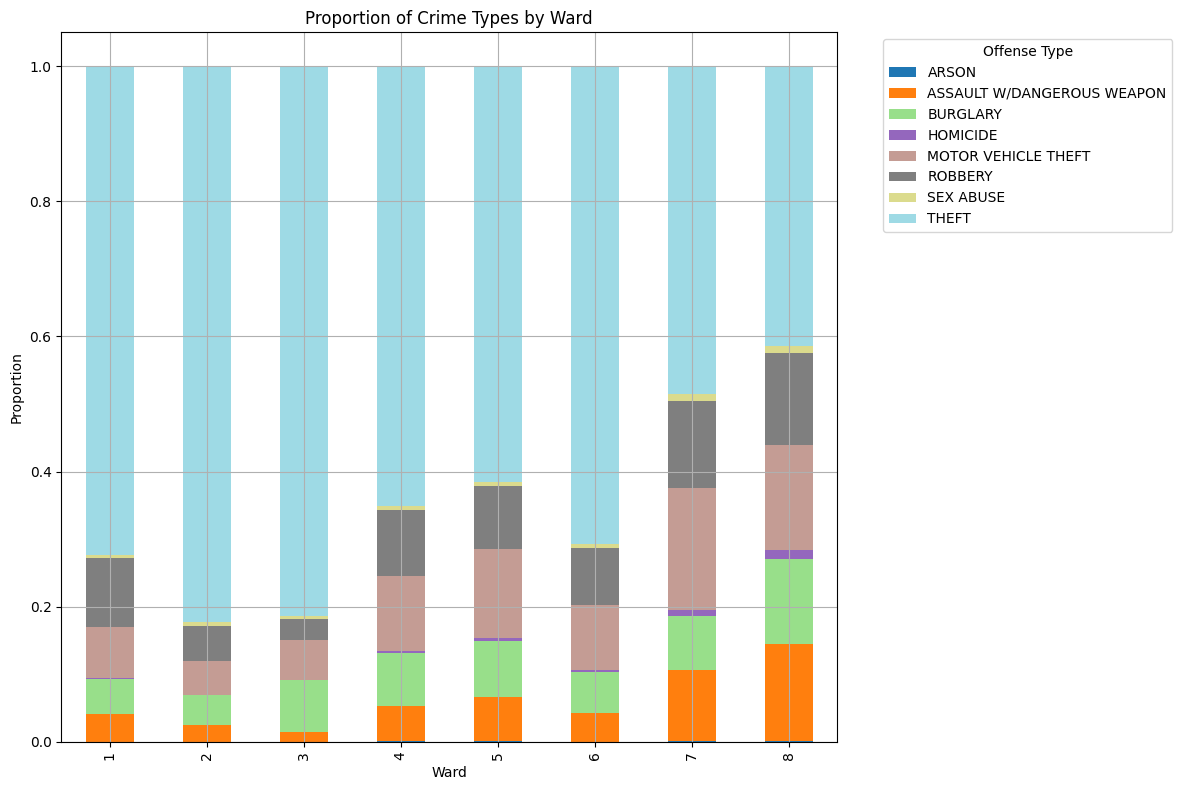
# References

* Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., & Ma, W. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the 31st International Conference on Neural Information Processing Systems, 3146-3154. https://arxiv.org/abs/1712.09913
* Pierre, S., Barun, A., & Pascalev, A. (2024, December). DC Crime 2008-2023: The evolution of crime in D.C. through map visualizations. https://storymaps.arcgis.com/stories/74f124ddc6e4440f948dfb931f5030f8
* Browning, C., et al. (2017). Ecological Networks and Urban Crime: The Structure of Shared Routine Activity Locations and Neighborhood-Level Informal Control Capacity. Criminology. https://pmc.ncbi.nlm.nih.gov/articles/PMC5815399/
* WUSA9. (2024, June 4). DC violent crime stats. WUSA9. https://www.wusa9.com/article/news/crime/dc-violent-crime-stats/65-dd6634de-9b66-4209-86c8-e765950f6c5f
* Villalovas, E. (2024, January 13). What’s behind the rise in violent crime in Washington, D.C.? Washington Examiner. https://www.washingtonexaminer.com/news/2789854/whats-behind-rise-violent-crime-washington-dc/
* Open Data DC. (n.d.). DC crime data. Open Data DC. https://opendata.dc.gov/
* Skeem, J. L., & Kenneson, M. A. (2021). Sentiment analysis of social media as a tool for crime prediction: A systematic review. Journal of Criminal Justice, 73, 101809.
* Wang, X., et al. (2022). Time series analysis of crime-related social media data: Predicting crime trends from user-generated content. Journal of Crime Science, 11(3), 215-227.
* CNN. (2024, February 5). Washington, DC crime: Carjackings on the rise. CNN. https://www.cnn.com/2024/02/05/us/washington-dc-crime-carjackings/index.html
* Lal, S., et al. (2020). Analysis and Classification of Crime Tweets. Procedia Computer Science, 167, 1911-1919. https://www.sciencedirect.com/science/article/pii/S1877050920306761#section-cited-by
* Predicting Violent Crime in Urban Cities. Springer. https://link.springer.com/article/10.5210/bsi.v14i2.334
* Kedia, P. (2016). Crime Mapping and Analysis Using GIS. International Institute of Information Technology.
* Washington Post Editorial Board. (2024, March 17). Can Anything Stop the DC Crime Wave? The Washington Post. https://www.washingtonpost.com/opinions/2024/03/17/dc-crime-blame-bowser-council-prosecutors/
* Muzammil, K., Azmat, A., & Yasser, A. (2022). Predicting and Preventing Crime: A Crime Prediction Model Using San Francisco Crime Data by Classification Techniques. Wiley. https://onlinelibrary.wiley.com/doi/full/10.1155/2022/4830411

# Appendix

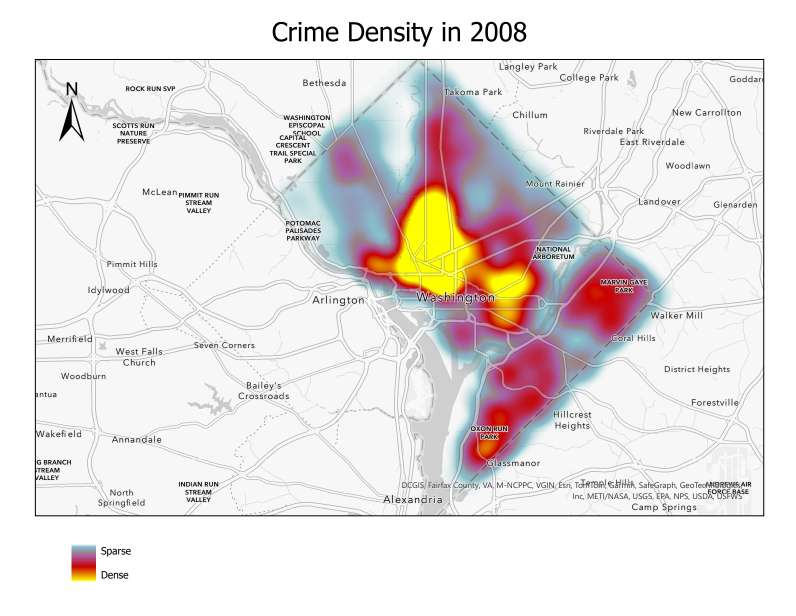
EDA and Statistics

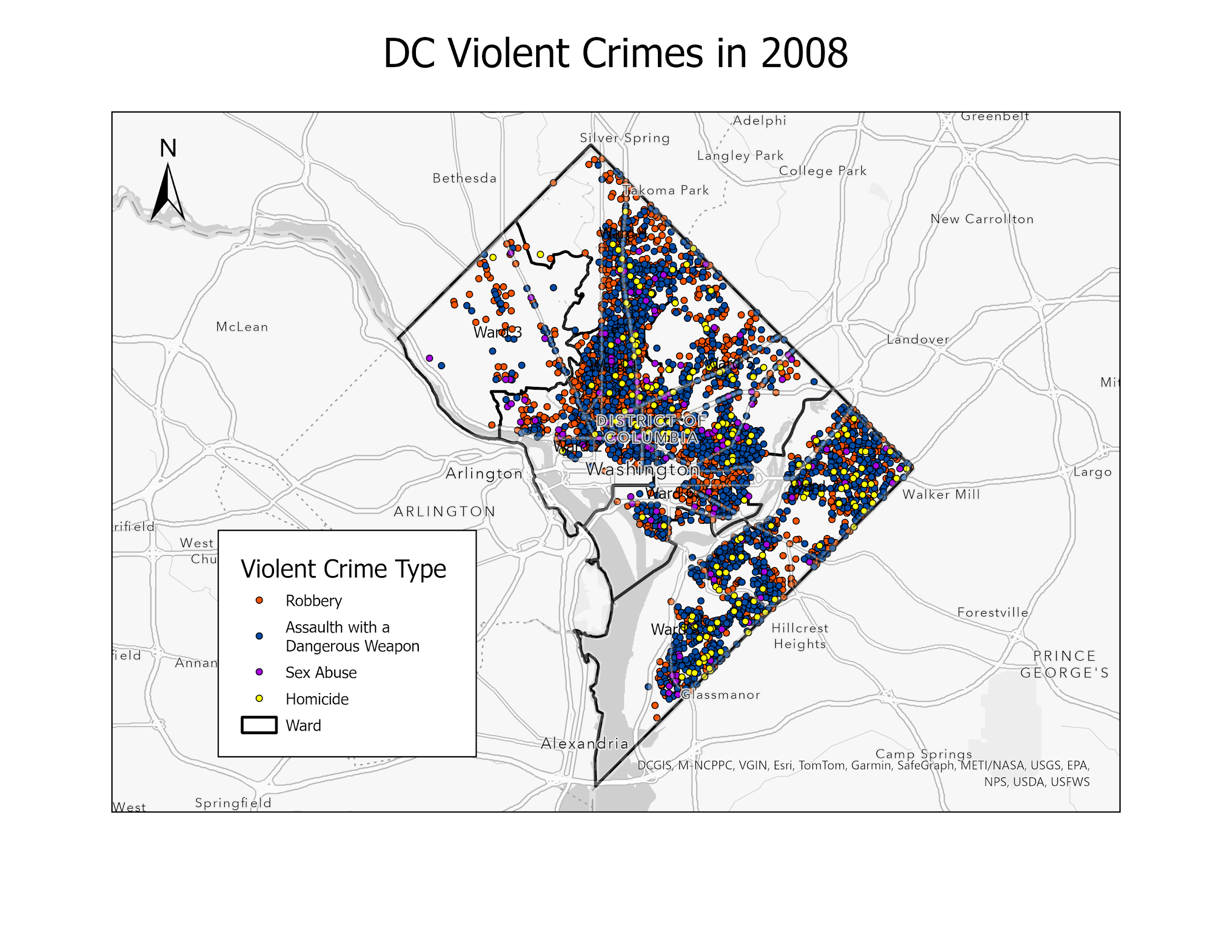


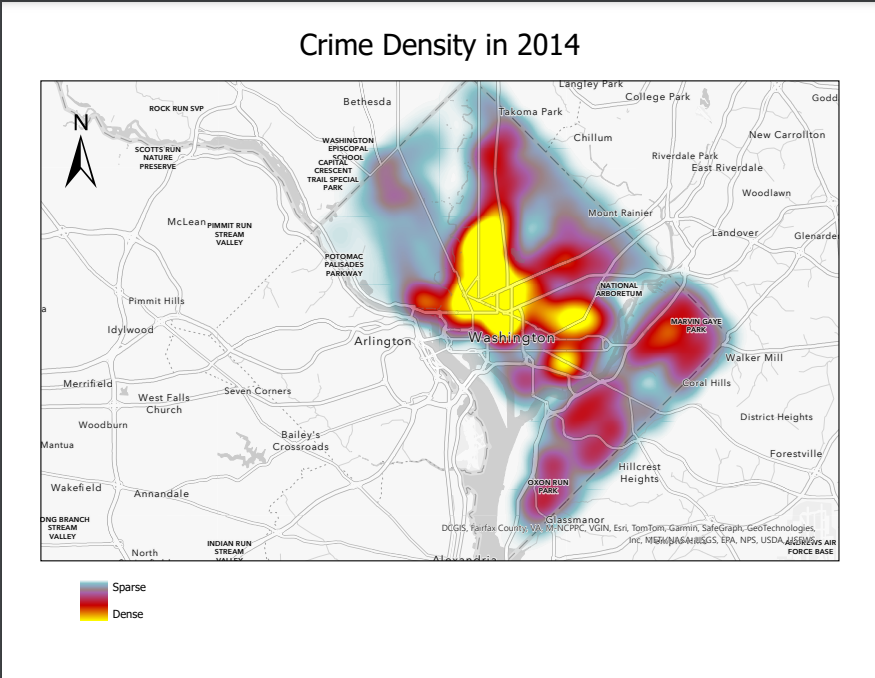


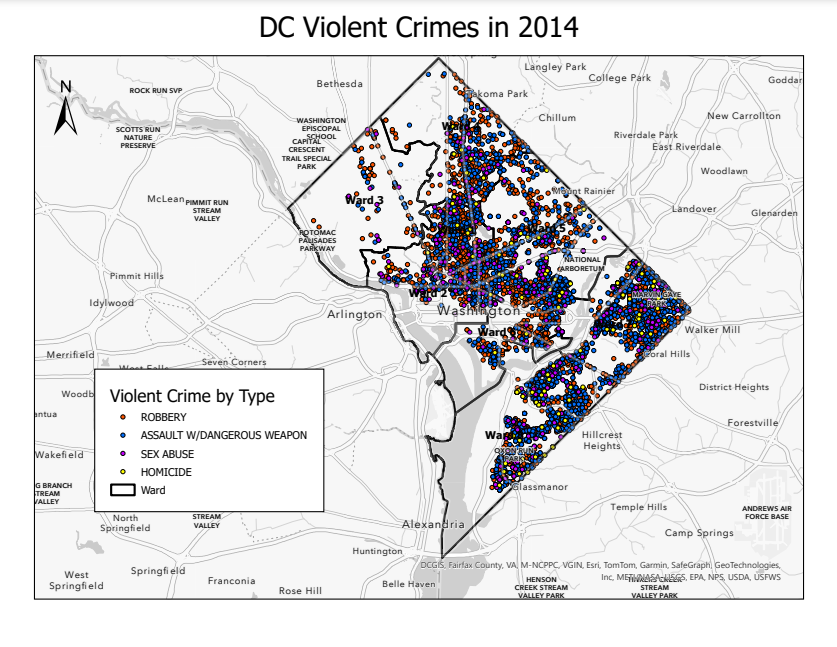
‘

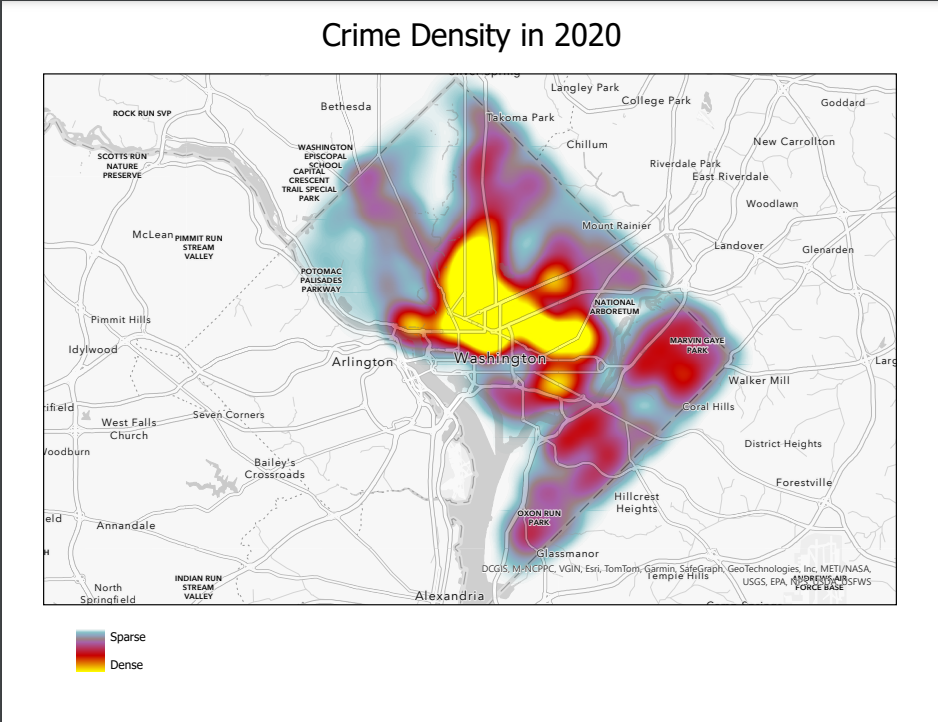
ArcGIS Pro Maps

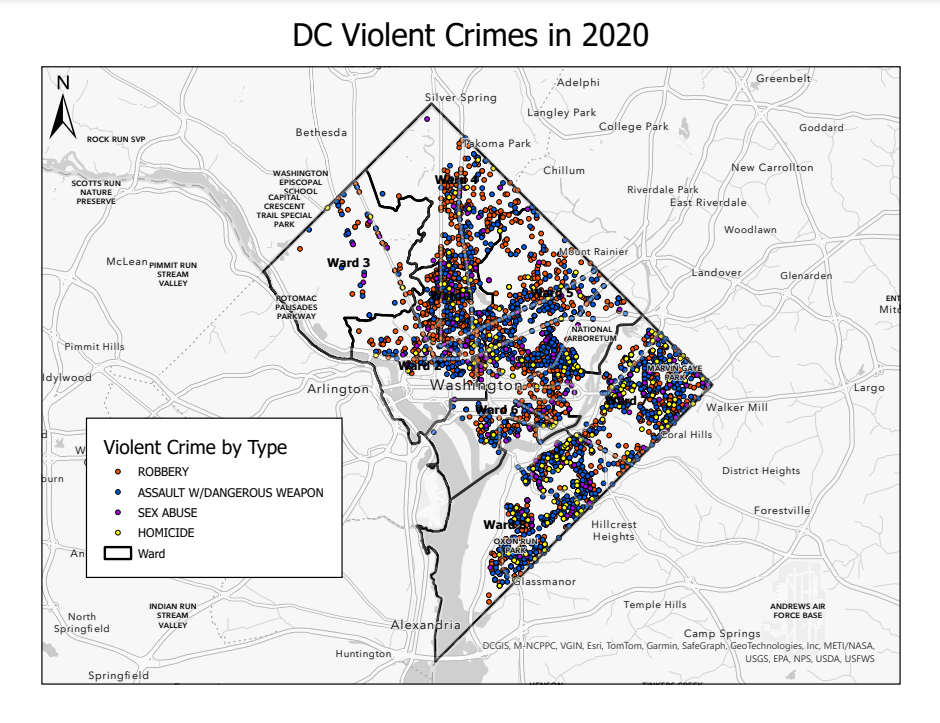










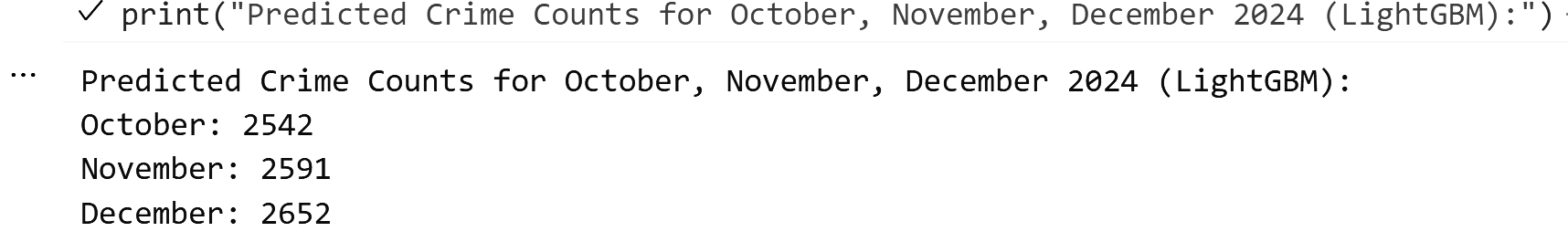


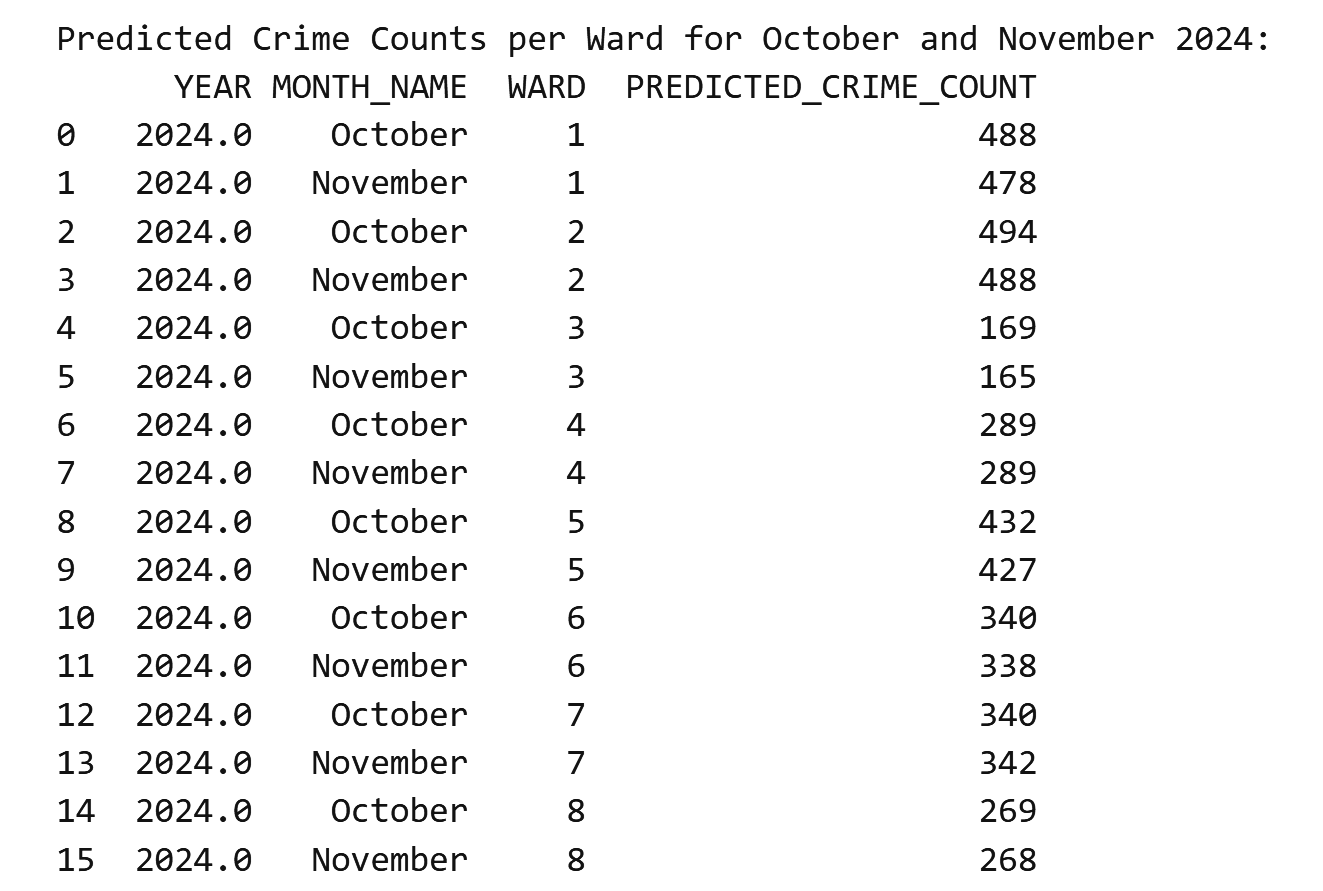
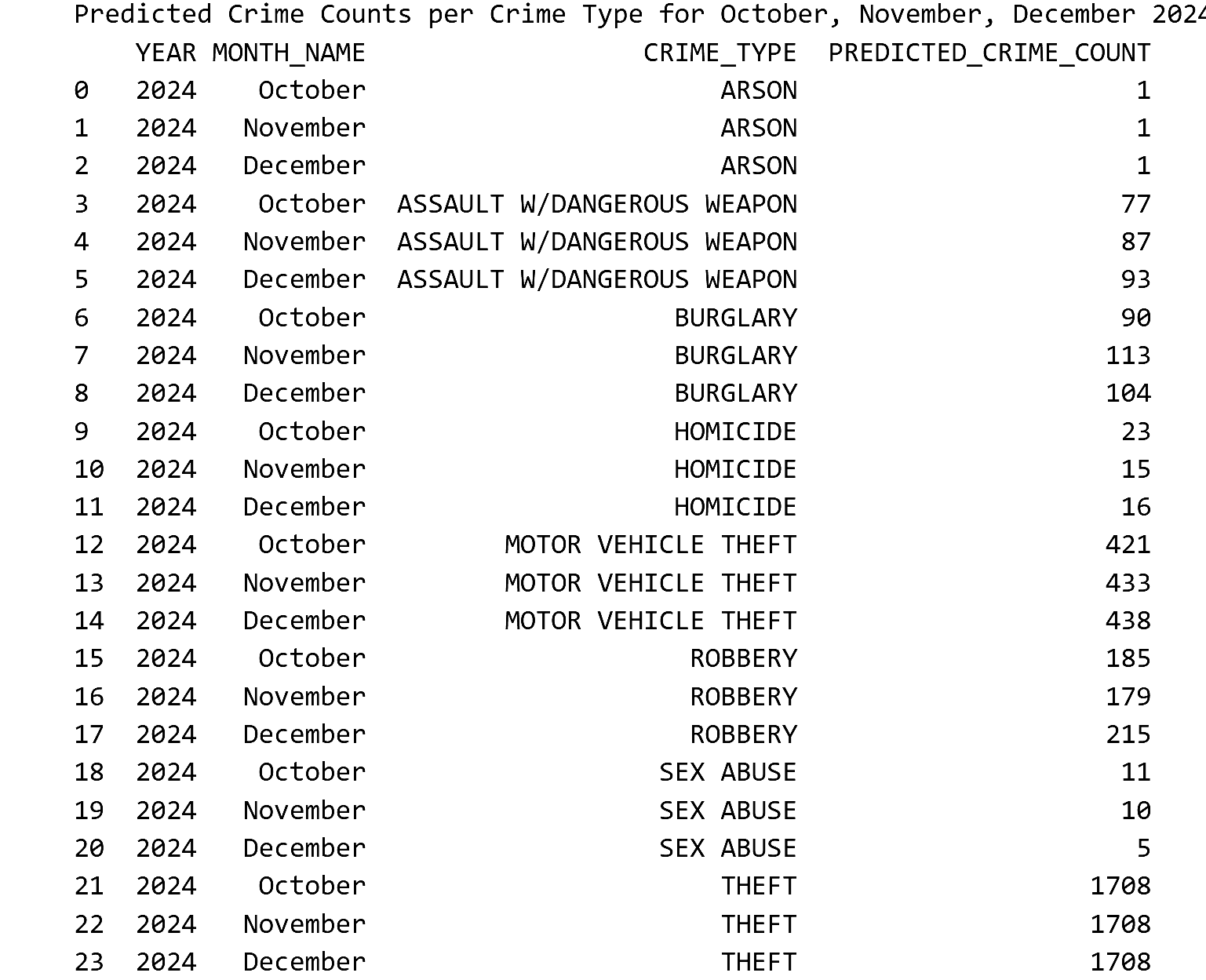
Regression Models - About Rolling Statistics

Rolling features provide valuable insights into temporal trends and fluctuations by calculating statistical measures over sliding windows of past data. In this model, rolling features include rolling means and rolling standard deviations computed over 1-, 3-, 6-, 9-, and 12-month windows. These features enhance the model's ability to understand both short-term and long-term patterns in crime counts.

The rolling mean represents the average crime count within a specified window. For example, the 1-month rolling mean reflects the most recent crime count, while the 9-month rolling mean smooths data over a longer period, revealing broader trends and reducing the impact of isolated fluctuations. The rolling standard deviation measures the variability of crime counts within each window, indicating periods of stability or volatility. High standard deviation values highlight significant fluctuations, while low values suggest steadier trends.

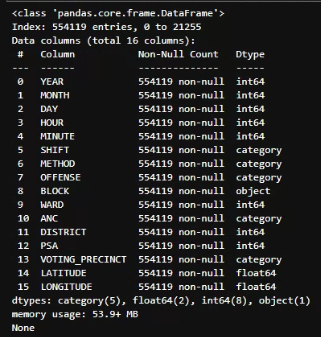
By incorporating multiple window sizes, the model gains a comprehensive view of temporal dynamics. The 1-month window captures very recent changes, the 3- and 6-month windows provide insights into seasonal trends, and the 9- and 12-month windows reflect broader year-over-year patterns. This multi-scale approach enables the model to balance short-term responsiveness with long-term context, improving its ability to predict future crime counts.

LightGBM Regression Model - Crime Counts Predictions

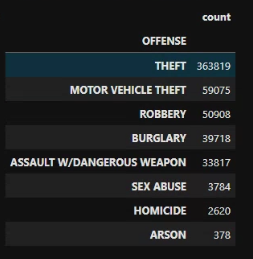
LightGBM Regression Model - Ward Distribution Predictions:LightGBM Regression Model - Crime Type Predictions

Dataset Info

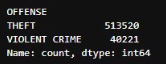
Dataset after Cleaning and Preprocessing



Offense Count Original

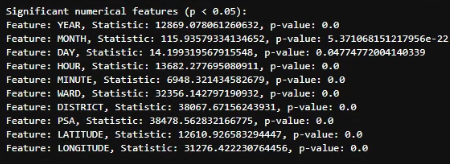


Offense Count after merging Categories

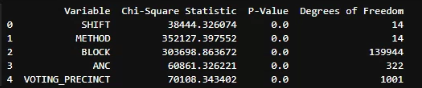


Significance Tests

Offense vs Numerical Columns - Kruskal Wallis Test

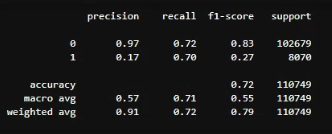


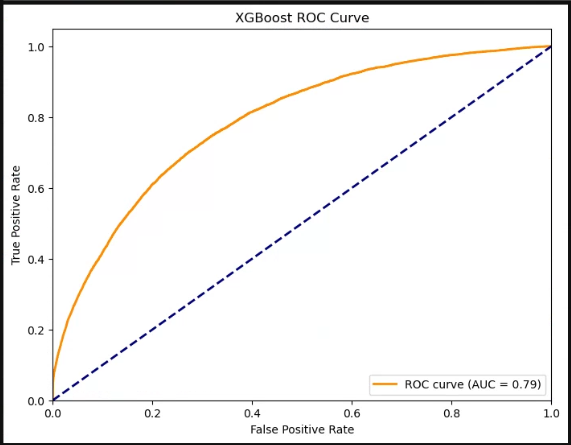
Offense vs Categorical Columns - Chi Square Test

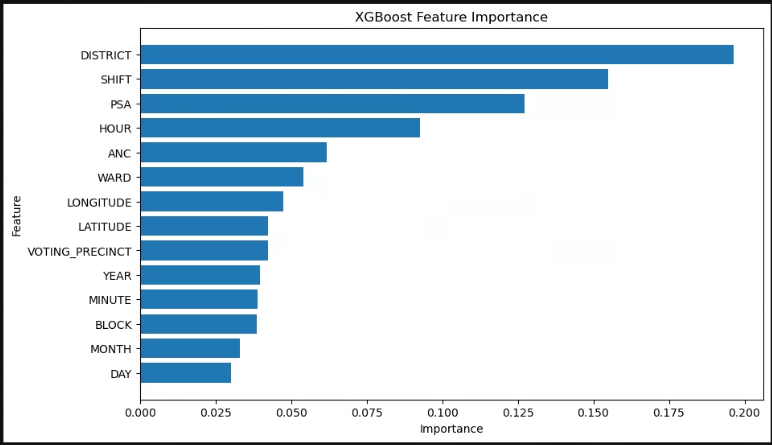


Classification Metrics

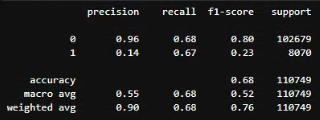
XGBoost

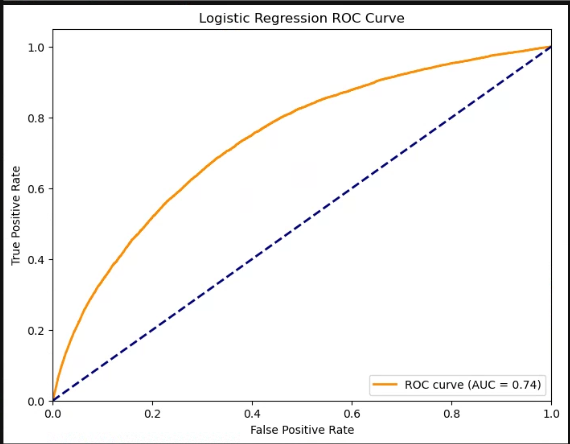




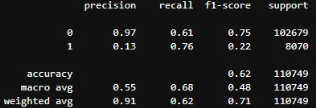


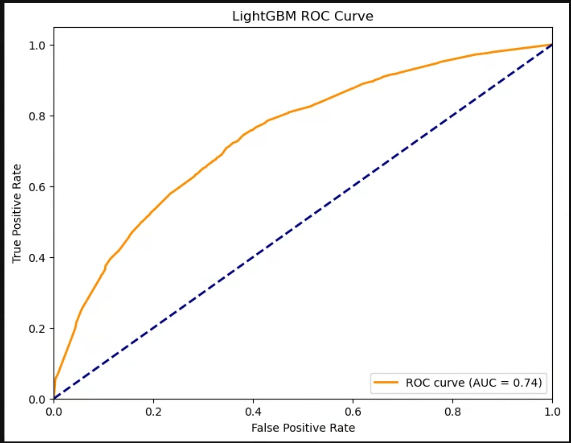
Logistic Regression





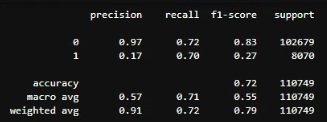
Light GBM

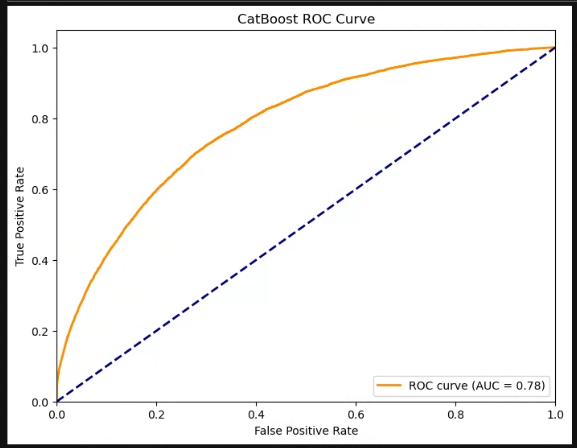


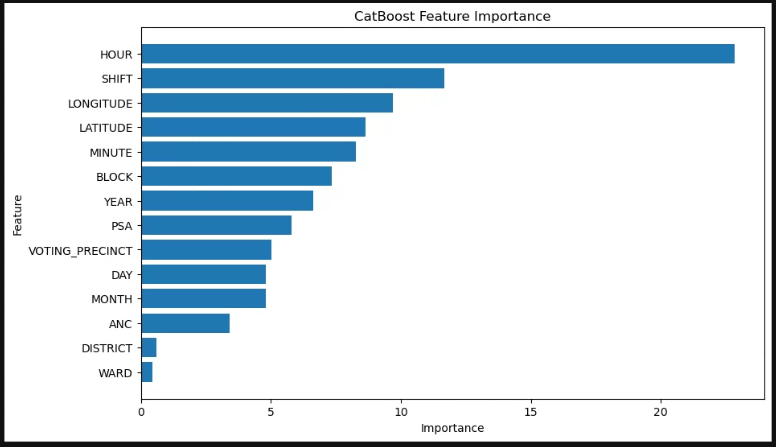




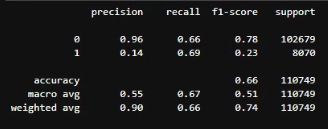
CatBoost Classifier

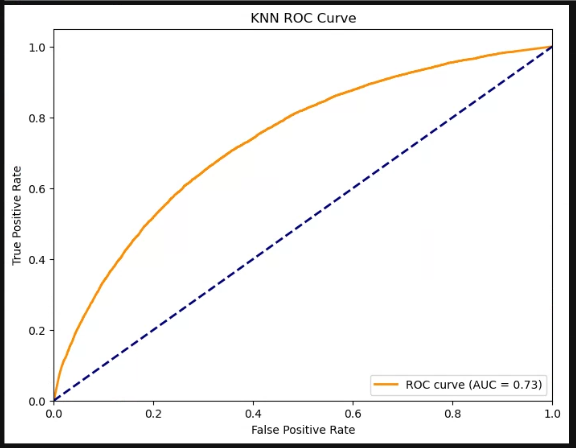




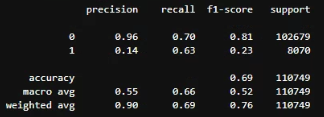


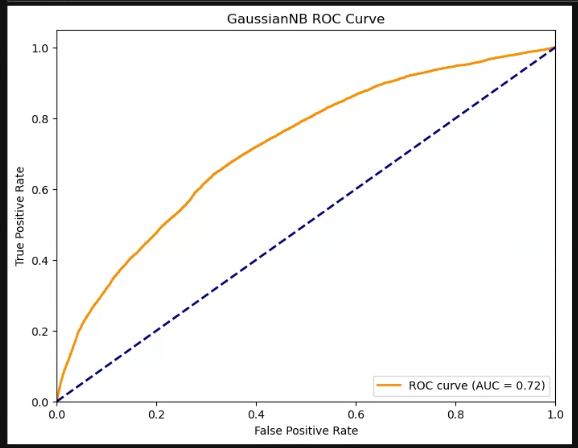
KNN



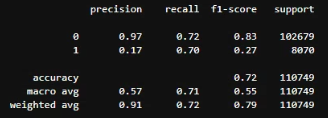


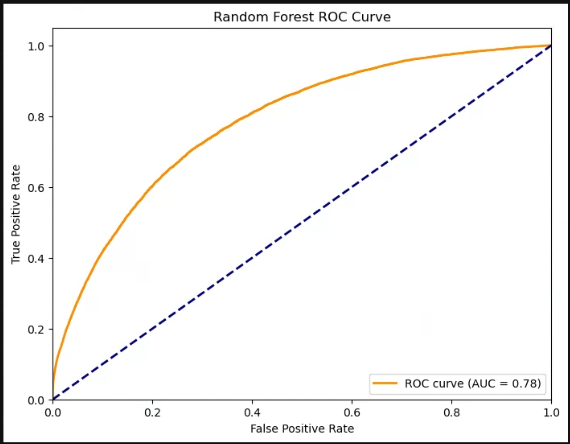
Gaussian Naive Bayes

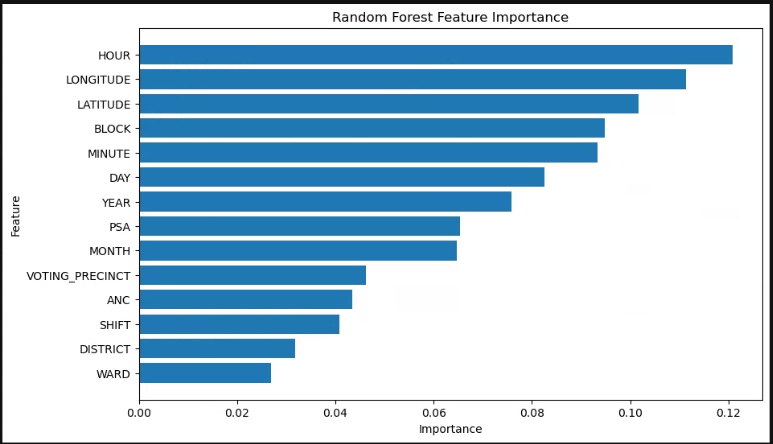




Random Forest



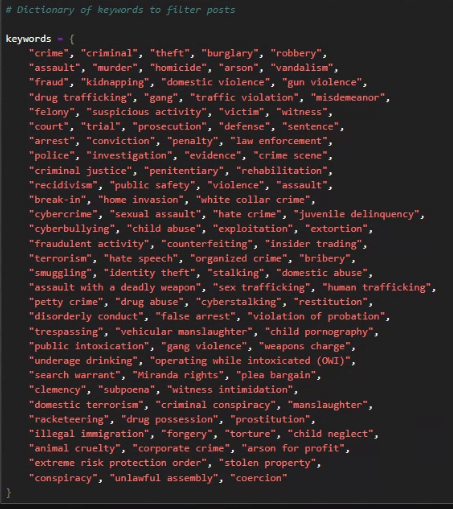




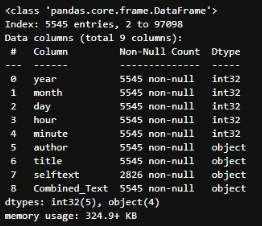
NLP Data Prep



Dictionary of Keywords used for Filtering



After filtering -



Post Count by year

