Crime Through the Years: Unveiling Trends and Patterns (2017-Present)

DATA602 - Principle of Data Science (Fall 2024)

December 15, 2024

https://github.com/drduron/DATA602_Group_Project

Berkay Yenilmez
Dimitri Rogelio-Mason Duron
Manuela O Deigh
Neha Dutt
Siva Durga Sai Prasad Buthada

Contributions

Dimitri Duron:
Searched and suggested the dataset to use. Assisted in generating an optimal predictive Random Forest model to classify incident types at a high score. Helped in all write ups throughout the entire project.

Neha Dutt:
Worked on the 'Number of Incidents in a PGPD Reporting Area' EDA analysis and created the Decision Tree, Neural Network, and Random Forest models to classify incident types, and created corresponding visualizations for. Helped in writing and cleaning the Github tutorial and final report, as well as formatting the final report. Contributed to all sections of the project.

Manuela Deigh:
Worked on 'Relationship between Clearance_Code_Inc_Type and PGPD_Reporting_Area' which is part of the EDA analysis. Created visual/ graphs such as the PCA, edited the k-means code, created maps and scatterplot in the ml section. Created and edited our Github tutorial and helped write our final report. (Sections worked on: Data Exploration and Summary Statistics, ML Algorithm Design/Development, Final Tutorial Report Creation)

Berkay Yenilmez:
Developed the SARIMA model for forecasting crime trends and the KNN classifier for incident type prediction. Created visualizations to analyze crime data, including temporal analyses of incident cases over time, a monthly crime rate chart, geographic maps displaying crime types and their locations, and interactive maps allowing users to explore localized incidents within a specific radius. Contributed to all parts of the project.

Siva Buthada:
Contributed to data curation and preprocessing by standardizing date formats, renaming columns, merging datasets, grouping related values in the clearance_code_inc_type column, removing duplicates, and handling missing values effectively. Reset the index post-cleaning and categorized incident types for better analysis. Conducted exploratory data analysis by visualizing incident types by clearance codes and documented data science ethics in the project report and tutorial.

1 Introduction

Understanding the presence of crime, including its frequency, spread, and diversity, is of critical importance in determining the safety of a community. Associated data can help drive decisions on whether or not to move into a community, whether a community requires additional resources to reduce crime, identifying crime hotspots, and how to overall improve the safety of a community. Furthermore, to make such conclusions, understanding how the number and types of crime incidences have evolved overtime is imperative. An overarching question here is if it is possible to determine the prominence of crime in a community based on the characteristics within that community, and how it has evolved over time? Can we identify the type of crime given the location where it occurred? Which regions have higher or lower crime rates, and do they coincide with specific crimes? Insights from this analysis would help all members of the community - ranging from homeowners to law enforcement officials to policy makers.

2 Data Curation

To answer these questions, this analysis focuses on crime incidents in Prince George's County, Maryland from February 2017 to September 2024. Two datasets from the county's data portal were selected and merged - "Crime Incidents July 2023 to Present" and the "Crime Incidents February 2017 to 5th July 2023". These datasets include information on the date, clearance code, case identification, coordinates, street address, and Prince George's Police Department (PGPD) sector, area, and beat of crime incidents. The crime incidents include traffic accidents, assaults, burglaries, homicides, robberies, sex offenses, stolen vehicles, thefts and vandalisms for which a report is written, thus it's important to note, that these datasets do not include every occurrence of an incident or call for service by PGPD.

To merge the two datasets, the date columns were converted to datetime using the Pandas library, and columns were renamed to have the same letter case across both datasets. As only two rows were missing PGPD beat information and one row was missing PGPD sector information, those values were dropped. Street number was missing for 42,090 records and replaced with 'Unknown', but since street address was populated for every record, street number can be extrapolated if required. Additionally, the "Crime Incidents July 2023 to Present" is regularly updated, but at the time of data processing, the data was only up until the first 8 days of October 2024, therefore, the dataset was capped until the end of September 2024 to avoid skewing the monthly trend analysis. Finally, as there were numerous clearance codes, the clearance codes were re-assigned to one of "ACCIDENT","ASSAULT", "AUTO, STOLEN", "B & E", "HOMICIDE", "ROBBERY", "SEX OFFENSE", "THEFT", or "VANDALISM" as incident type. The final dataset was organized into a Pandas dataframe with 184,583 records.

3 Exploratory Data Analysis

After cleaning and preprocessing we move on to our Exploratory Data Analysis (EDA). Due to our dataset being mostly categorical variables, we discussed with Dr. Alam and, as per her recommendation, performed only two kinds of statistical tests - Chi Square and ANOVA - rather than the required three different statistical tests for the assignment.

3.1 Distribution of Incident Types

We conducted a Chi-Square test of independence to evaluate whether there is a significant relationship between the frequency of incident types and the year in which the incidents occurred. Our null hypothesis for this test is that there is no significant relationship between the incident type and the year of occurrence. In contrast, the alternative hypothesis is that there is a significant relationship between the incident type and the year of occurrence. To visualize the results of the Chi-Square test, we created a heatmap (Figure 1). The analysis revealed that the top five incident types showed a statistically significant difference in frequency across eight years. The Chi-Square statistic was 16,077.283, with a

p-value of approximately 0.0 at an alpha level of 0.05. These results led us to reject the null hypothesis, concluding that the frequency of incident types is significantly dependent on the year of occurrence.

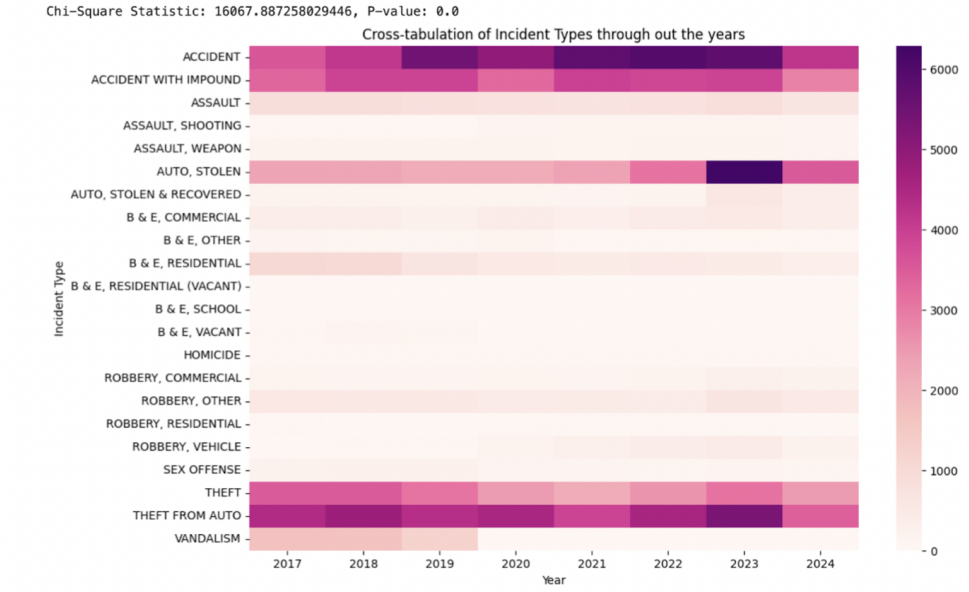


Figure 1. A heatmap showing the results from the Chi square test for the distribution of incident types.

3.2 Relationship between Clearance_Code_Inc_Type and PGPD Reporting Area

We conducted another Chi-Square test to examine whether there is a significant relationship between incident types and PGPD_Reporting_Area at an alpha level of 0.05. Our null hypothesis was that there is no significant association between the top ten incident types and the PGPD reporting areas where they occur, and our alternative hypothesis was that there is a significant association between the two variables.To visualize the results, we created a heatmap (Figure 2). A Chi-Square statistic of 41,434 and a p-value of 0.0, allowed us to reject the null hypothesis. This indicates a significant relationship between incident types and reporting areas at the specified alpha level.

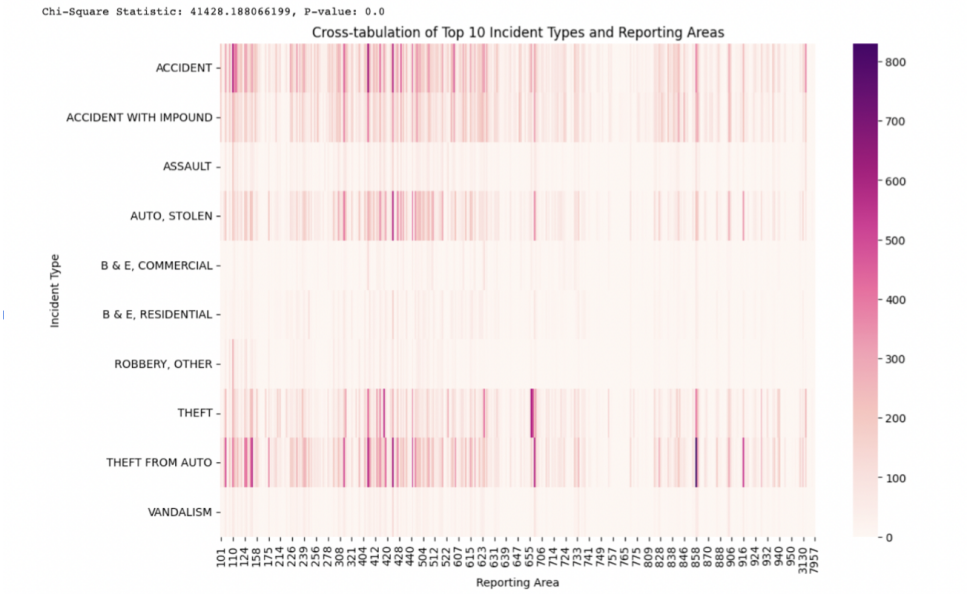


Figure 2. A heatmap of our Chi square test that looks at incident types based on PGPD reporting area.

3.3 Number of Incidents in a PGPD Reporting Area

Our final test we conducted was ANOVA (Analysis of Variance) to determine whether there is a statistically significant relationship between the year and the frequency of incidents at an alpha level of 0.05. Our null hypothesis is that there is no statistically significant difference in the number of incidents across the years, while the alternative hypothesis is that there is a significant difference. The results of the ANOVA show an F-value of 1.56, indicating high variance when categorizing incidents by year. However, the p-value of 0.15 is much larger than the alpha level of 0.05, suggesting low confidence in rejecting the null hypothesis. Based on these results, we conclude that the year is not a reliable predictor of incident frequency within the reporting areas. At an alpha level of 0.05, we fail to reject the null hypothesis.

4 Additional Analysis & Visualizations

4.1 Number of Incident Cases Over Time

Appendix A shows the number of incident cases over time from February 2017 to September 2024. A clear seasonal pattern is visible, with fluctuations in the number of incidents throughout the years. A significant drop can be observed around March 2020, coinciding with the onset of the COVID-19 pandemic and resulting lockdowns, which likely reduced activities such as driving and public gatherings. Such outliers, caused by unique events, may influence the analysis if the data is not normalized or accounted for during modeling. The overall trend indicates a gradual increase in reported incidents in recent years, emphasizing the importance of analyzing seasonal and long-term patterns in crime data.

4.2 Total Crime Rate at Each Month

Appendix B presents the total crime rate for each month, aggregated across all years. The bar chart reveals that the summer months—May, June, July, and August—consistently exhibit higher crime counts compared to other months. This trend may be attributed to increased outdoor activities or seasonal factors, which create more opportunities for crimes to occur. The analysis highlights the importance of accounting for seasonal variations when studying crime patterns and planning preventative measures.

4.3 PGPD Reporting Area Code Finder

Appendix C visualizes the geographic centers of PGPD reporting areas based on the latitude and longitude of incidents. Each pin represents the average coordinates of reported incidents in a specific area, with a popup displaying the corresponding PGPD area number. This interactive map provides an overview of the spatial distribution of crime incidents, allowing users to identify and explore high-activity regions. Such visualizations are invaluable for law enforcement and urban planning, enabling targeted interventions and resource allocation to areas with higher crime rates.

4.4 Crime Incident Map by PGPD Reporting Area

Figure 3 displays crime incidents within a specified PGPD reporting area on an interactive map. Each incident type is represented with a unique color-coded marker, making it easy to distinguish between categories like accidents, thefts, assaults, and more. The map is centered around the selected area's incidents, allowing users to observe the geographic distribution of crimes and identify potential clusters or hotspots. By exploring this map, users can gain insights into the types and frequencies of crimes in specific neighborhoods, providing valuable information for community safety and resource allocation.

Figure 3. Interactive map with colored incident type labels based on latitude and longitude.

4.5 Interactive Visualization of Localized Crime Incidents

Appendix D dynamically displays incidents occurring within a 0.2-mile radius of a user-specified location on the map. Users can interactively click on any point to set a central location, and incidents near that point are shown using color-coded markers based on their types, such as accidents, thefts, or assaults. This localized mapping approach provides an intuitive way to explore crime density and diversity in specific neighborhoods, enhancing the understanding of nearby incidents for residents, policymakers, or researchers. The inclusion of color-coded categories further aids in distinguishing the nature of incidents quickly and effectively.

5 Machine Learning Analysis & Results

5.1 SARIMA Analysis

The SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model was applied to analyze and forecast the frequency of crime incidents over time. The dataset exhibited clear seasonal trends and periodicity, making SARIMA an appropriate choice for modeling and prediction. The model was trained on historical crime data, with the last 12 months reserved for testing to evaluate its performance. The analysis showed that the SARIMA model effectively captured the seasonal trends in the data, with a MSE of 96908.49, a RMSE of 311.30, and a Standardized RMSE of 0.1383 relative to the mean of the test data. These metrics suggest that the model performed well in predicting the frequency of crimes, as its errors were relatively small compared to the scale of the test data.

The performance of the model is visualized in Figure 4, which shows the training data, test data, and forecasted values along with confidence intervals. The SARIMA model successfully identified repeating seasonal patterns in the crime data, allowing for accurate short-term forecasting. Additionally, the decomposition of the time series into trend, seasonal, and residual components is shown in Appendix E, further illustrating the patterns in the data. The trend line highlights a gradual increase in the frequency of crime incidents over the years, peaking in 2023. The seasonal component captures recurring yearly fluctuations, while the residual component represents unexplained variance after accounting for the trend and seasonality. While the SARIMA model demonstrated strong predictive capabilities, some discrepancies between the forecasted values and observed data, particularly within the confidence intervals in Figure 4, suggest potential areas for improvement. Overall, the SARIMA model provides valuable insights into crime trends over time and serves as a useful tool for resource allocation, policy-making, and understanding the seasonal nature of crime in Prince George's County.

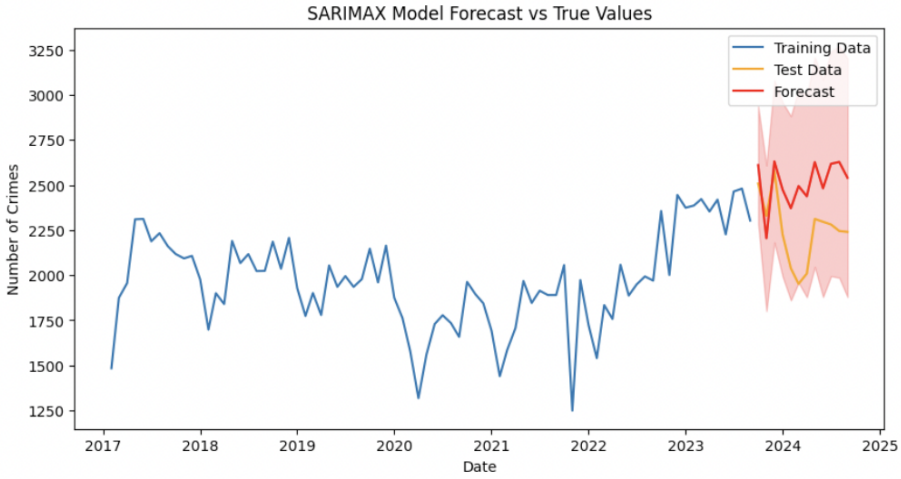


Figure 4. Time Series Plot from the SARIMAX analysis showing number of crimes per year.

5.2 KNN Classifier

Appendix F shows the process of finding the best k-value for the KNN classifier that predicts incident type. Testing k values ranging from 1 to 20 using cross-validation, the graph visualizes the calculated accuracy for each k. The best k value is determined to be 18, as it maximizes accuracy. The graph indicates that increasing k generally improves the classifier's performance, although it eventually plateaus, confirming k=18 is optimal. This selection balances classifier performance by smoothing the influence of outliers and ensuring enough neighbors are considered for more consistent predictions.

The classification report in Appendix G provides a summary of the KNN classifier's performance with K=18. The overall accuracy of the model is 0.51. The class with the largest number of samples, "ACCIDENT," shows higher performance compared to other classes, with a precision of 0.62, recall of 0.76, and an F1-score of 0.68. Classes like "SEX OFFENSE" and "HOMICIDE," which are minority classes, have a precision, recall, and F1-score of 0.00, indicating that the classifier struggled to predict their incident types accurately due to the imbalanced nature of the dataset. These results highlight the limitations of the KNN classifier in achieving satisfactory recall and precision for underrepresented incident types. Despite acceptable performance for frequent classes like "ACCIDENT" and "THEFT," the classifier's inability to predict minority classes impacts its overall effectiveness.

5.3 Decision Tree & Random Forest Analysis

In order to answer the question of whether a crime's incident type (clearance code) can be determined based on spatial and temporal data, a decision tree classification model was developed. The features used in this model were PGPD sector, area, and beat, which are all encoded as they are categorical values. The model itself was trained on 80% of the data, after removing outliers, for which the hyperparameters were tuned using a five-fold Grid Search Cross Validation algorithm with 'roc_auc_ovr' scoring as this is a multi-class classification scenario. Since there is an imbalance representation of each incident type in the dataset, ROC_AUC was used for tuning the model rather than accuracy. The parameter grid included a max_depth range of (3,6) and max_samples_split and max_samples_leaf ranges of (10,20) to minimize overfitting and achieve a simpler tree that would be ideal for real-world application. However, this only resulted in a best accuracy, precision, recall, and F1-score of 0.4139, ROC_AUC of 0.5722, and log loss of 1.52 (Appendix H). The poor model performance is most likely due to the imbalanced nature of the dataset. Different hyperparameters were also tested to increase complexity of the tree, but they resulted in similar or worse performance.

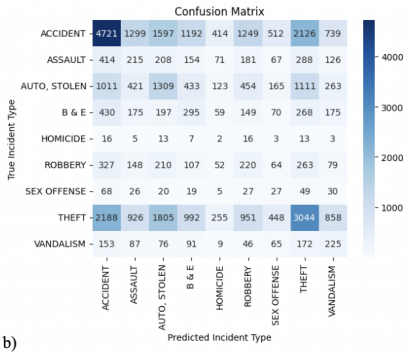
To improve model performance, a Random Forest Classifier model was created using the same train-test

split and encoded features. Additional encoded temporal features were added through feature engineering including month, year, day of the week, and whether or not the incident occurred on a weekend. Another difference here was the application of SMOTE sampling, which oversamples the minority incident type classes. Due to computational constraints, a three-fold Grid Search Cross Validation was used with n_estimators in [100, 200], max_features in ['auto', 'sqrt'], max_depth in [10, 20], min_samples_split in [2, 5], and min_samples_leaf in [1, 2]. This did improve model performance, albeit not much, with ROC_AUC of 0.6238. The classification report and confusion matrix can be seen in Figure 5. Though the ROC_AUC is moderate, as seen in Figure 5A, there are still a high number of misclassifications resulting in low precision, recall, and F1-scores across the classes.

Best Model Classification Report:

	precision	recall	f1-score	support
0	0.51	0.34	0.41	13849
1	0.07	0.12	0.09	1724
2	0.24	0.25	0.24	5290
3	0.09	0.16	0.12	1818
4	0.00	0.03	0.00	78
5	0.07	0.15	0.09	1470
6	0.02	0.10	0.03	271
7	0.42	0.27	0.32	11467
8	0.09	0.24	0.13	924
accuracy			0.27	36891
macro avg	0.17	0.18	0.16	36891
weighted avg	0.37	0.27	0.31	36891

a)



b)

Figure 5. Results of Random Forest Classifier with ROC_AUC of 0.6238. Classification report (a) shows weak metrics across the classes, and confusion matrix (b) shows how many misclassifications there were.

Afterwards, we pivoted to feature engineering where we transitioned the labels of incident types to become "violent" and made it a binary label so that either the incident was 0 for nonviolent or 1 for violent. We then incorporated another column by including an external dataset with the coordinates of nearby police stations. By calculating the distance from the incident location to the police station, we then converted this to miles and used this distance as another feature for the Random Forest. We then applied the coordinate columns and the distance from the nearest police station as features to train the model on the labels. We then attained an accuracy of over 0.9 but an F1 score of less than 0.1 so we tried random over sampling to compensate for the disparity of labels and reran the training to then get an F1 score of over 0.8, exceeding all previous iterations of the incident labeling.

5.4 Neural Network Analysis

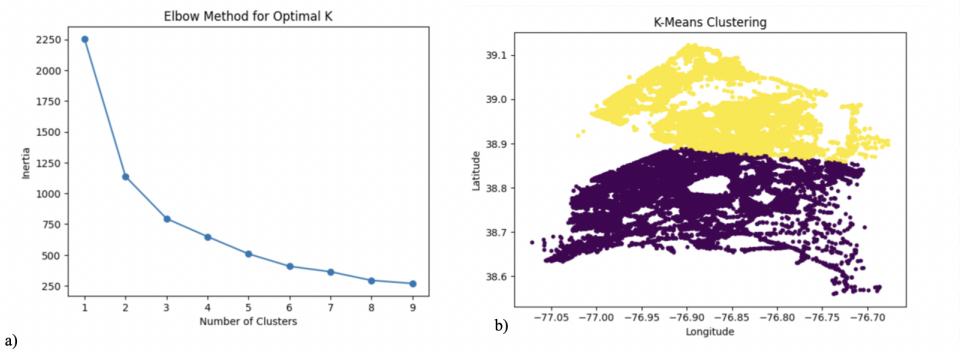
A Neural Network was also created to further build upon the Random Forest analysis by introducing complexity. Various optimizers like Adam and RMSProp, activation functions like ReLu and Softmax, train-test splits, and epochs were used to train the model. Ultimately, three hidden layers with Leaky ReLu activation (alpha=0.1) and dropouts of 0.2-0.3. Each of these were Keras dense layers of sizes 256, 128, and 64, respectively. The final output layer had Softmax activation for multiclass classification for the eight incident types. The final model used the RMSProp optimizer with 'sparse_categorical_crossentropy' loss and accuracy as the metric for 50 epochs that utilized early stopping to prevent overfitting. It also implemented SMOTE Sampling to account for class imbalance. The results indicated that this network performed similarly to the Decision Tree and preliminary Random Forest regression with accuracy of 0.1753, ROC_AUC of 0.6153, and test loss of 1.9596. Other combinations of the above mentioned parameters resulted in similar or worse performance. Even with the selected model, the results are comparable to the previous incident classification models, indicating that further network tuning, feature engineering, or additional data may be required.

5.5 PCA K-Means Analysis

We utilized Principal Component Analysis (PCA) and K-Means clustering to analyze and determine crime severity in different areas. Our analysis began with K-Means clustering to identify crime hotspots.

To determine the optimal number of clusters, we applied the Elbow Method (Figure 6A), which evaluates the inertia to identify the best value for K. The results indicated that K=2 was the optimal number of clusters for our analysis. Using K=2, we implemented K-Means clustering (Figure 6B), and followed it with PCA to better understand the data structure and reduce dimensionality. The dataset for PCA included K-Means cluster, latitude, longitude, crime count, crime diversity, and severity score. PCA allowed us to simplify the visualization of complex spatial data.

The clustering results revealed that Cluster 1 had a crime count of 92,406, while Cluster 0 had 92,045. Crime diversity, representing the number of unique crime types (eg. theft, assault), was consistent between the two clusters at 22, indicating that both regions experience a similar variety of crimes. To evaluate the quality of clustering, we calculated the Silhouette Score, which ranges from -1 to 1. Our Silhouette Score was 0.234 (Figure 6C), indicating moderately defined clusters with some overlap, likely due to the similarity in crime diversity across clusters. When analyzing severity, Cluster 1 exhibited a standardized severity score of 1.0, representing higher crime activity and severity. In contrast, Cluster 0 had a score of -1.0, indicating lower crime activity (Figure 6C). These standardized scores (ranging from -1 to 1) allow for easy comparison, with positive values reflecting higher severity and negative values reflecting lower severity.



c)

kmeans_cluster	latitude	longitude	crime_count	crime_diversity	severity_score
1	38.965131	-76.896339	92406	22	1.0
0	38.811666	-76.922154	92045	22	-1.0

Figure 6. Results of elbow method, K-means, and PCA metrics. Elbow method for optimal K (a) shows the best value for K. Scatter plot (b) reveals the two crime cluster we have determine. Summary table (c) shows the summary of the PCA model.

To visualize the results, we created a scatter plot showing crime severity by cluster. Cluster 1 is shown in red, indicating higher severity and increased crime activity, while Cluster 0 is shown in blue, reflecting lower severity and reduced crime activity (Figure 7a). Additionally, we created a geographical map to highlight areas with varying crime severity. The map identified New Carrollton as the area with the highest crime severity, whereas Camp Springs exhibited the lowest (Figure 7B). This visual can help us analyze areas with higher crime which can help guide law enforcement in focusing on those areas. While this analysis offers valuable insights, there can be room for improvement.

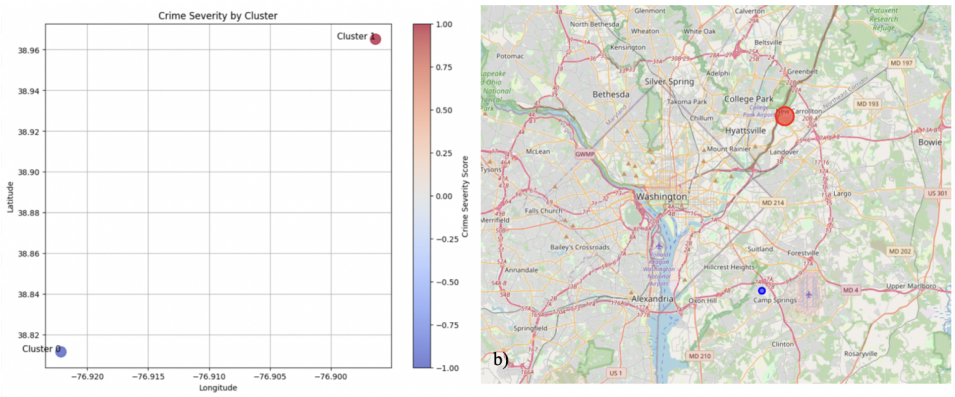


Figure 7. Scatterplot (a) visual of the two cluster from the PCA model. Map (b) visual of a geographical crime severity map.

6 Insights & Conclusions

A crime dataset from Prince George's County provides a critical foundation for leveraging data science to improve public safety, inform policy making, and support community well-being. Analyzing crime trends and patterns allows stakeholders, including governments, law enforcement, and urban planners, to allocate resources effectively and implement targeted interventions. Predictive analytics, such as SARIMAX models, have proven to be effective in forecasting crime frequency over time by identifying repeating seasonal patterns. These insights enable stakeholders to anticipate and mitigate potential crime hotspots proactively. Additionally, geolocation data plays a pivotal role in identifying areas with high incident frequency through methods like K-means clustering, enabling urban planners and community organizations to prioritize safety measures in these hotpots.

Delving deeper into the dataset reveals the complexities of categorizing incidents by type and addressing imbalances in the data. Early modeling efforts using algorithms such as K-nearest neighbors and decision trees faced challenges such as over-fitting and low predictive performance, with F1 scores failing to meet expectations. By applying feature engineering-such as relabeling incident types into binary categories of 'violent' and 'non-violent crimes' and leveraging techniques such as SMOTE and imbalanced sampling, the Random Forest model emerged as the most effective approach, achieving a significantly improved F1 score of 0.8. Although neural networks were also explored, they failed to outperform the optimized Random Forest model, underscoring the importance of careful data preparation and model selection in achieving accurate and actionable predictions.

This iterative analysis highlights the power of combining statistical techniques, machine learning models, and exploratory data analysis (EDA) to extract meaningful insights from crime data. For example, PCA was used to enhance clustering results, assigning severity scores to crime hotspots for a more nuanced understanding of local trends. These findings emphasize the value of integrating multiple data sources, such as geolocation and incident details, to inform policy and safety strategies. By identifying key factors like violent crime prevalence, proximity to law enforcement, and crime density, this approach equips decision-makers with the tools needed to develop targeted interventions, improve urban planning, and foster safer communities. The comprehensive workflow-spanning EDA, visualization, model experimentation, and refinement-demonstrates how data science can unlock the full potential of this crime dataset to benefit public safety and well-being.

7 Data Science Ethics

This project analyzed crime data from Prince George's County, Maryland (February 2017–present) with a focus on fairness, transparency, and reliability. The dataset, derived from real-time police reports, minimized selection bias but acknowledged under-reporting of certain crimes due to societal stigma or

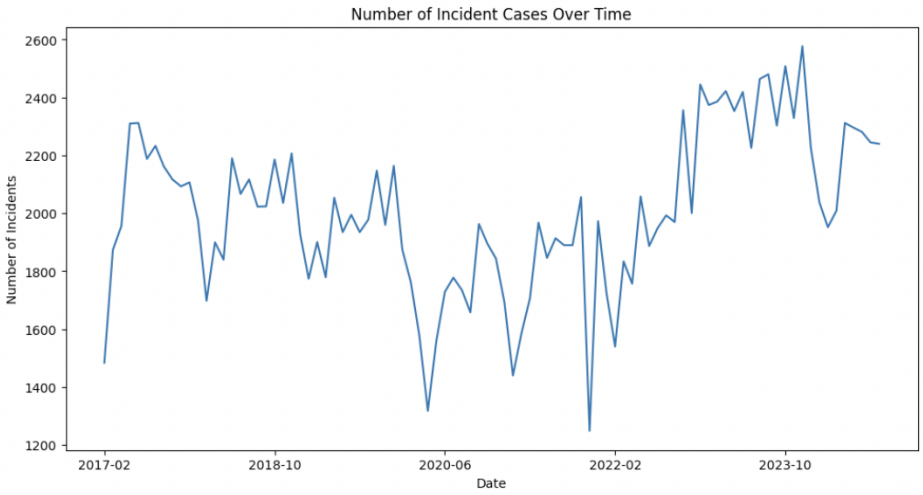
mistrust in law enforcement. The analysis emphasized the trends within the reported data, recognizing its limitations.

Sensitive demographic and victim information was excluded from the dataset to prevent discrimination and ensure privacy. Data imbalances were mitigated through pre-processing steps, such as removing duplicates, handling missing values, and categorizing incident types to simplify analysis while maintaining data integrity. Predictive models like SARIMAX and Neural Networks faced challenges due to imbalanced data, with plans to enhance the dataset and explore advanced techniques.

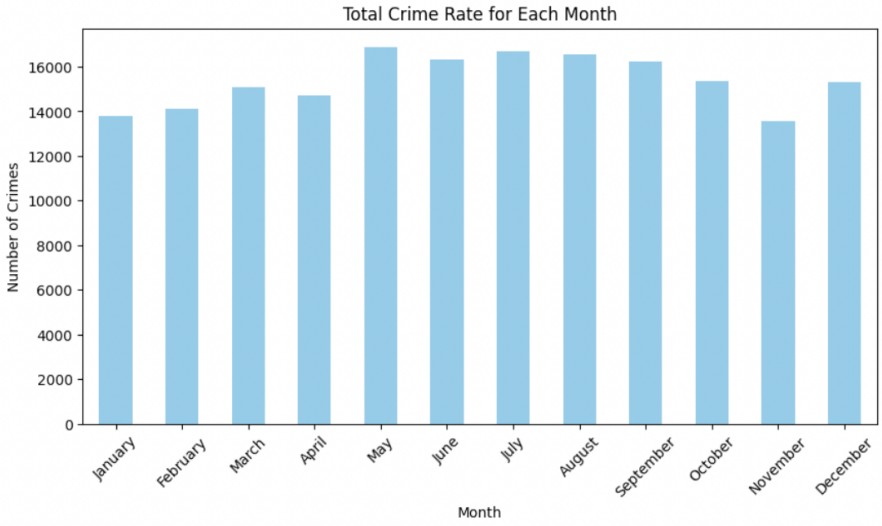
Transparency was maintained through detailed documentation of workflows, enabling independent verification. By focusing on spatial and temporal crime trends, the analysis avoided stereotypes and aimed to inform equitable public safety measures while maintaining privacy. These efforts ensured actionable, reliable findings aligned with ethical standards, fostering community trust and enabling informed policymaking.

8 Appendix

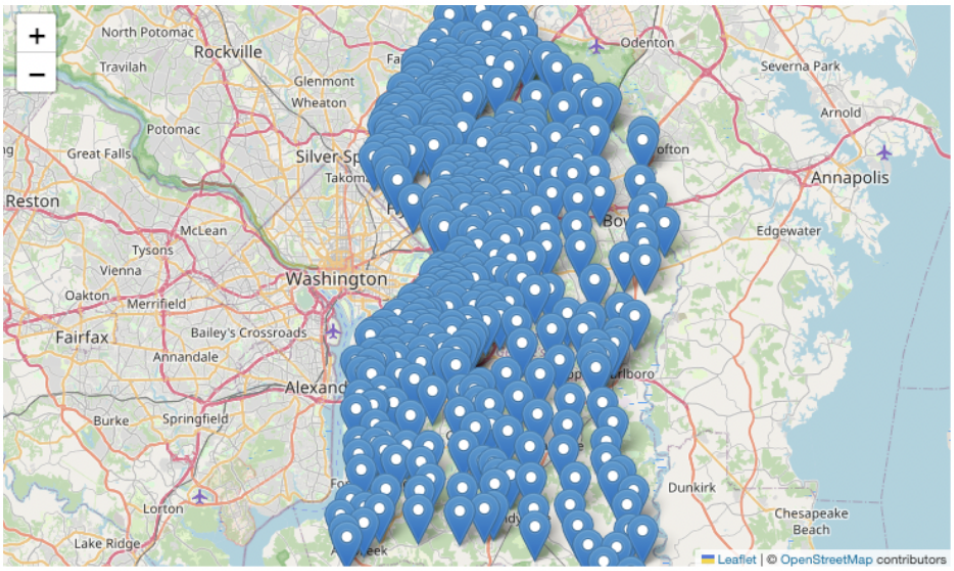
8.1 Appendix A



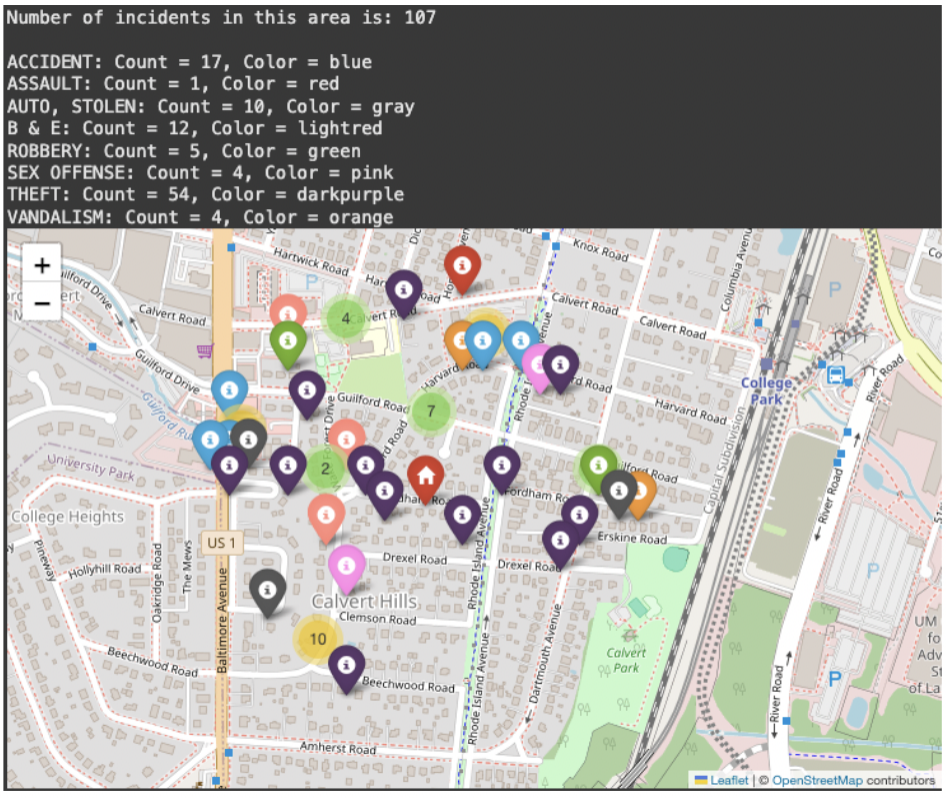
8.2 Appendix B



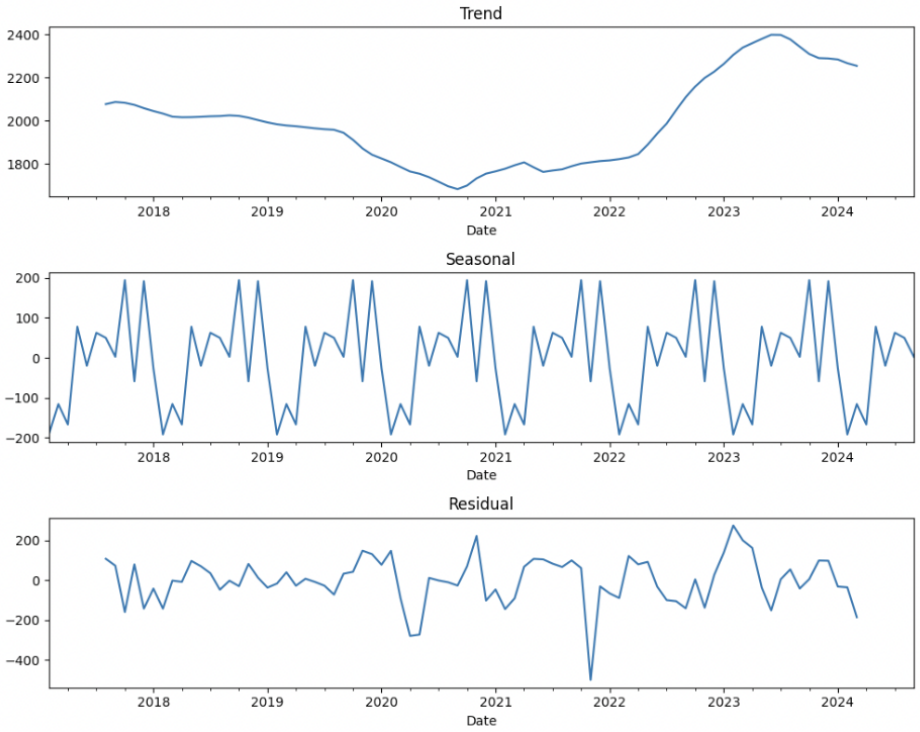
8.3 Appendix C



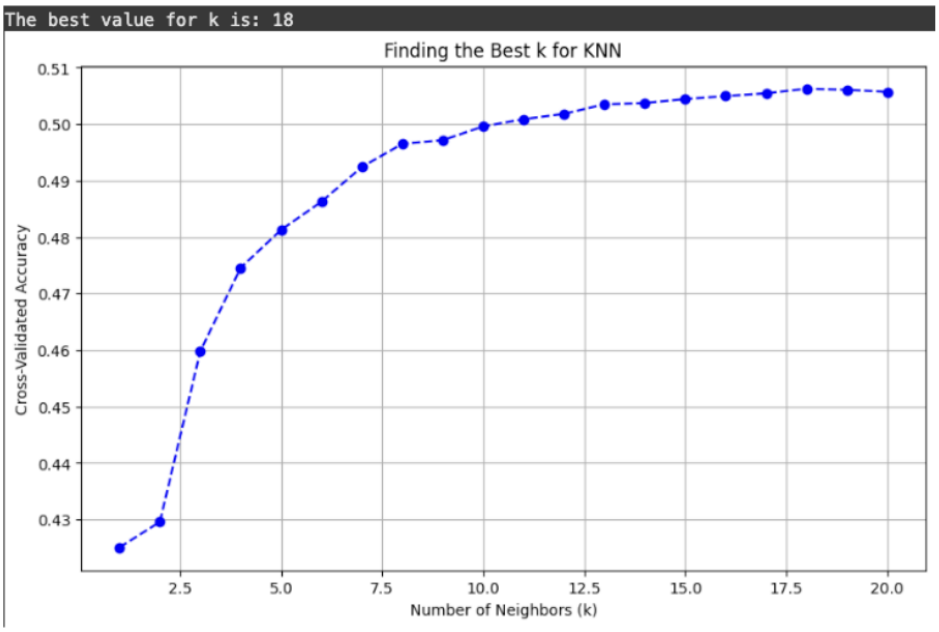
8.4 Appendix D



8.5 Appendix E



8.6 Appendix F



8.7 Appendix G

```
X_train shape: (147562, 2)
X_test shape: (36891, 2)
y_train shape: (147562,)
y_test shape: (36891,)
Accuracy of KNN Classifier with k=18: 0.51

Classification Report:
              precision    recall  f1-score   support

    ACCIDENT       0.62      0.76      0.68     13718
     ASSAULT       0.27      0.04      0.08      1725
AUTO, STOLEN       0.29      0.19      0.23      5274
       B & E       0.27      0.05      0.08      1824
    HOMICIDE       0.00      0.00      0.00        68
     ROBBERY       0.23      0.04      0.06      1472
 SEX OFFENSE       0.00      0.00      0.00       264
       THEFT       0.47      0.63      0.54     11622
   VANDALISM       0.08      0.00      0.00       924

    accuracy                           0.51     36891
   macro avg       0.25      0.19      0.19     36891
weighted avg       0.46      0.51      0.47     36891
```

8.8 Appendix H