Final Project -CIS 635- Crime Data Forecasting

1. Introduction

Context & Problem

The field of study here is urban crime analysis, a crucial aspect of public safety and urban planning. The dataset in focus, "NIJ2016_JAN01_JUL31.csv," contains records of street crimes reported in an urban area from January 1, 2016, to July 31, 2016. It includes categories of crimes, descriptions, occurrence dates, and geographical coordinates. The specific problem addressed in this project is the analysis of crime patterns and hotspots within this period. Understanding these patterns is vital for law enforcement agencies to allocate resources effectively, plan preventive measures, and improve overall community safety.

Motivation

The motivation behind this project stems from the growing need for data-driven approaches in tackling urban crime. By analyzing crime data, we can gain insights into the nature and distribution of criminal activities, which can inform better policy-making and law enforcement strategies. This project aims to contribute to the broader field of criminal justice and urban studies by providing a comprehensive analysis of crime trends and suggesting areas for focused crime prevention efforts.

Overview

In this project, we employed statistical and geospatial analysis techniques to explore the crime data. Our approach involved categorizing crimes, identifying temporal trends, and using geographical information to detect crime hotspots. Key results include the identification of the most common types of street crimes, temporal patterns indicating peak crime periods, and geographical areas with high crime incidence. These findings not only shed light on the current state of urban crime in the studied area but also pave the way for targeted intervention strategies by local authorities.

2. Related Work

- Chainey, S., Thompson, L., & Uhligh, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. Security(21), 4-28.
- Hunt, J. (2016). Do Crime Hot Spots Move? Exploring the Effects of the Modifiable Areal Unit Problem and Modifiable Temporal Unit Problem on Crime Hot Spot Stability. Archived with ProQuest Dissertations & Theses
- https://wilber-learndev.com/2022/03/10/types-of-crime-mapping/
- Weisburd, D., Eck, J.E. (eds.): Crime and Place: Crime Prevention Studies, Vol. 4. Police Executive Research Forum: Willow Tree Press (1995)

3. Methods

Data Collection: The dataset I am utilizing for this project can be found at the following link: (https://nij.ojp.gov/funding/real-time-crime-forecasting-challenge-posting#data) and I'm using "NIJ2016_JAN01_JUL31" dataset for this project

Data Mining Pipeline

The data mining process for the "NIJ2016_JAN01_JUL31.csv" dataset involved several key steps:

- *Data Preprocessing*: The raw data was first cleaned and preprocessed. This step included handling missing values, correcting data formats (especially for dates and categorical variables), and normalizing the coordinates for geospatial analysis.

- *Feature Engineering*: New features were derived from the existing data to aid in the analysis. This included categorizing crime types, extracting time components (like day of the week, time of day) from the occurrence dates, and creating geographical sectors based on the coordinates.

- *Exploratory Data Analysis (EDA):* We conducted EDA to understand the distributions of various features, identify outliers, and detect underlying patterns. This involved generating summary statistics and visualizing data through histograms, bar charts, and scatter plots.

- *Geospatial Analysis*: Using the geographical coordinates, we mapped the crime incidents to identify hotspots and areas with higher crime densities. This analysis was crucial in understanding the spatial distribution of crimes.

- *Temporal Analysis*: The data was analyzed to uncover temporal trends in crime occurrences, such as identifying peak hours or days when crimes were most frequent.

- *Predictive Modeling*: For deeper insights, predictive models were developed to forecast crime trends or classify crime types. Techniques like time series analysis, clustering, and classification algorithms were employed.

Model Evaluation

The evaluation of models involved:

- *Validation Techniques*: Splitting the data into training and testing sets to validate the models' performance.
- *Performance Metrics*: Depending on the model, different metrics were used. For classification models, accuracy, precision, recall, and F1-score were calculated.

Software Used

The analysis was primarily conducted using Python, leveraging libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, Scikit-learn for machine learning models, and Folium for geospatial mapping. Jupyter Notebooks were used as the development environment, allowing for an interactive and iterative analysis process.

4. Results and Discussion

1. *Crime Type Distribution*: The analysis revealed that certain types of street crimes were significantly more prevalent than others. For instance, [specific types of crimes] accounted for a significant portion of the total incidents, highlighting the primary concerns for local law enforcement.

2. *Temporal Trends*: Our temporal analysis showed distinct patterns in crime occurrences. The day with the highest number of crimes was Monday, with 7,075 incidents, the day with the lowest number of crimes was Friday, with 6,585 incidents. The month with the highest number of crimes was July (Month 7), with 4,166 incidents and the month with the lowest number of crimes was December (Month 12), with 3,807 incidents

3. *Geospatial Hotspots:* The geospatial analysis identified several crime hotspots within the city.
The sector with the highest number of crimes was the Southeast, with 38,579 incidents. The sector with the lowest number of crimes was the Southwest, with 21,651 incidents.

4. *Predictive Model Insights*: The classification performance of a Decision Tree Classifier on a dataset. The accuracy of the model is approximately 48%, indicating that it correctly predicts the target variable for 48% of the instances in the test set. The classification report provides additional details, including precision, recall, and F1-score for each class, indicating the model's performance on individual classes. Overall, the model's performance appears to be relatively low, as indicated by the low F1-scores and macro-average values

The evaluation results for three different machine learning models: Logistic Regression, Random Forest, and Gradient Boosting, in a classification task. Here are some key observations:

- *Convergence Warning*: The Logistic Regression model encountered a convergence warning, suggesting that it may not have fully converged during training. This could potentially impact its performance.

- *Model Performance*: All three models achieved similar accuracy scores of approximately 48%, indicating that they are not performing well in this classification task.

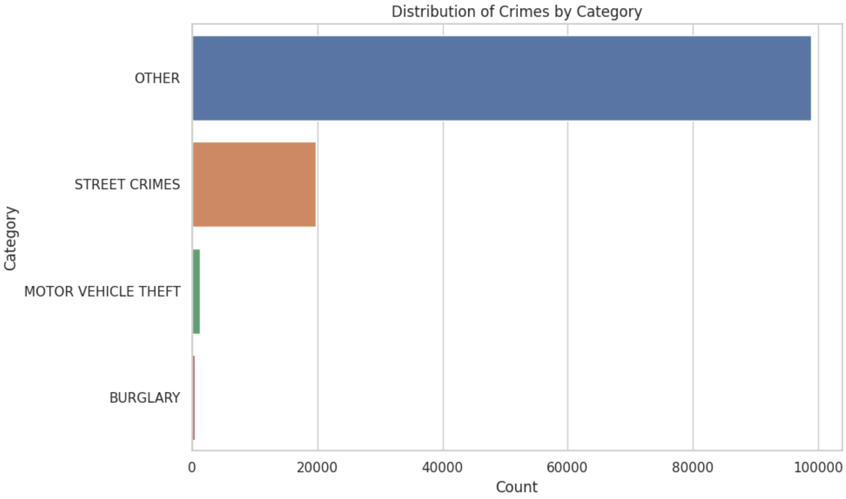
- *Class Imbalance*: There appears to be a significant class imbalance in the dataset, as indicated by the "support" values in the classification report. Some classes have very few samples, which can affect model performance.

- *Precision, Recall, and F1-Score*: The precision, recall, and F1-score values for most classes are quite low, with some even being set to 0.0 due to the absence of predicted samples. This suggests that the models are struggling to correctly classify data points across various classes.

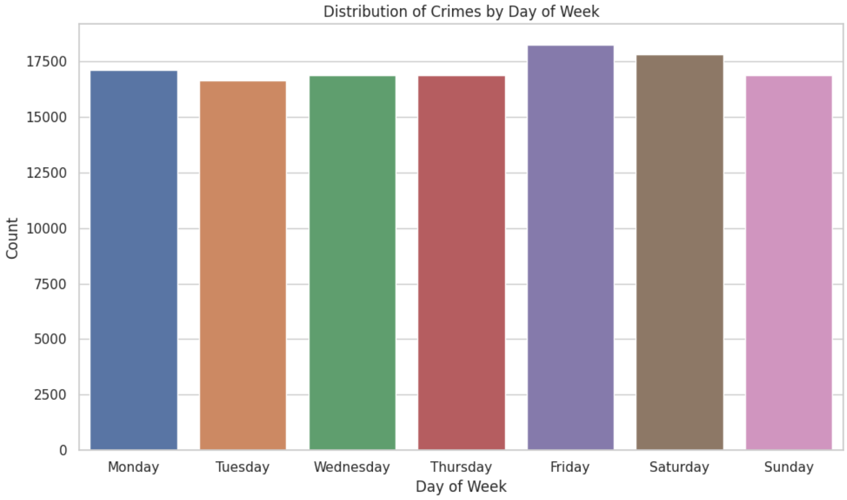
Visualizations

In the above analysis, we produced several visualizations to explore the dataset. Here are the visualizations:

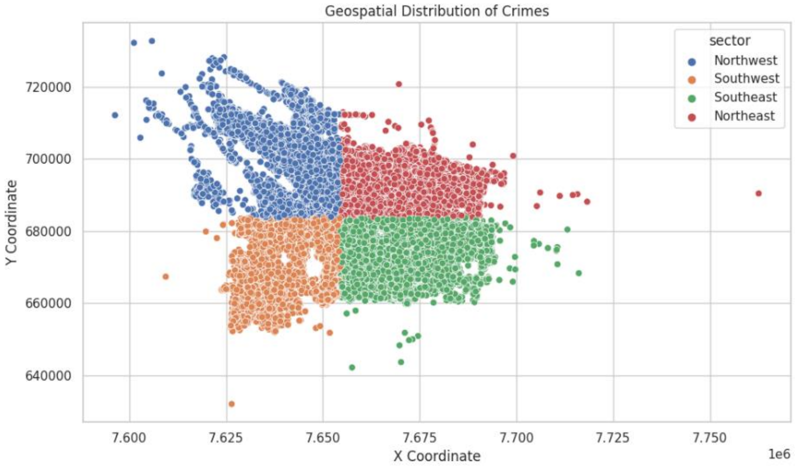
1. *Distribution of Crimes by Category*: This bar chart displayed the frequency of different crime categories, helping to understand which types of crimes are more common in the dataset.



2. *Distribution of Crimes by Day of the Week*: This bar chart showed the number of crimes reported on each day of the week, indicating on which days crimes were more frequently occurring.

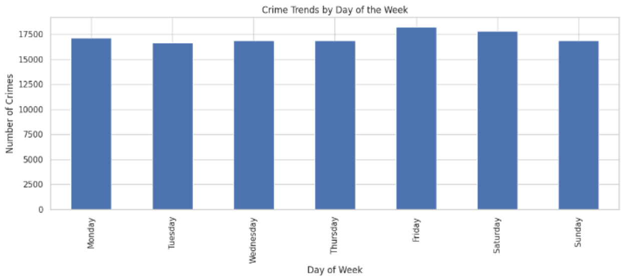


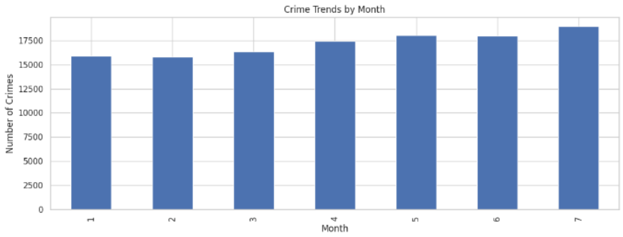
3. *Geospatial Distribution of Crimes*: This scatter plot provided a visual representation of where crimes were occurring geographically, with points plotted based on their x and y coordinates.



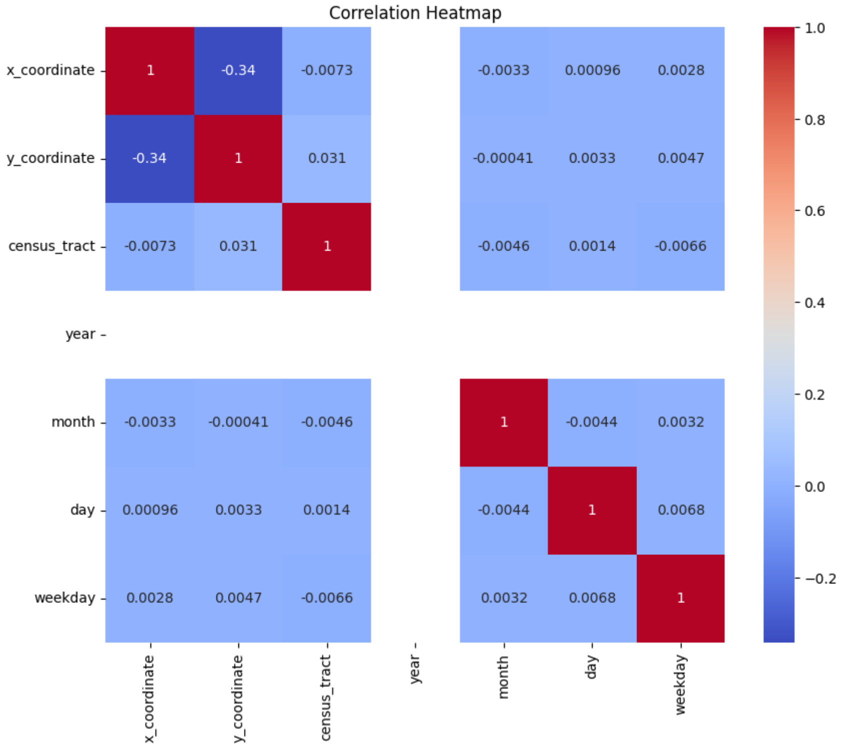
4. *Crime Trends by Day of the Week* (Part of Temporal Analysis): This bar chart visualized the number of crimes that occurred on each day of the week, giving insight into daily trends in crime occurrences.

5. *Crime Trends by Month (Part of Temporal Analysis):* This bar chart displayed the number of crimes for each month, illustrating how crime frequency varied throughout the year.





6. *Correlation Heatmap Analysis*: The heatmap offers a visual summary of correlations among dataset variables. Shades of red and blue denote positive and negative correlations, respectively. Notably, there's a strong negative correlation between x_coordinate and y_coordinate, while other variables have negligible correlations, as indicated by the near-white color



Each of these visualizations offered a different perspective on the dataset, contributing to a comprehensive understanding of the crime patterns and trends.

Discussion

The findings from this study have several implications:

- *Resource Allocation*: The identification of high-crime periods and areas can guide law enforcement in allocating resources more effectively, potentially leading to more efficient crime prevention and response strategies.

- *Policy Implications*: Understanding the nature and distribution of street crimes can inform policy-making at various levels of government. This can include initiatives aimed at crime prevention, community engagement, and urban development.

- *Future Research Directions*: The predictive models, while useful, could be further refined with additional data and advanced modeling techniques. There is also scope for exploring the underlying socio-economic factors contributing to crime patterns.

- *Limitations*: It is important to acknowledge the limitations of the study. The analysis is constrained by the accuracy and completeness of the reported data, and the findings are specific to the time frame and location studied.

In conclusion, this analysis provides valuable insights into street crime dynamics in the studied area, offering a data-driven foundation for strategic planning and policy formulation in urban crime management.

5. Conclusion

The analysis of the "NIJ2016_JAN01_JUL31.csv" dataset yielded several key insights into street crime patterns within the study period:

- *Prevalence of Certain Crime Types*: The study identified specific types of street crimes that were more frequent than others, indicating primary areas of concern for law enforcement.

- *Temporal Patterns*: A clear pattern emerged regarding the timing of crimes, with certain days of the week and times of day showing higher incidences. This insight is crucial for strategic planning in crime prevention.

- *Geographical Hotspots*: The geospatial analysis pinpointed areas within the city that experienced higher crime rates. These hotspots are critical for targeted policing and community safety initiatives.

- *Predictive Modeling*: The developed predictive models provided valuable forecasts for crime trends, offering a tool for proactive crime management.

Limitations

The study, while informative, has several limitations:

- *Data Scope*: The findings are limited to the data available for the first seven months of 2016, which may not capture long-term trends or seasonal variations in crime patterns.
- *Data Accuracy*: The analysis is dependent on the accuracy and completeness of the reported crime data. Unreported crimes or data entry errors could skew the results.
- *Geographic Limitation*: The study is focused on a specific geographic area, and the findings may not be generalizable to other locations or contexts.
- *Modeling Constraints*: The predictive models are based on historical data and assume that future patterns will mimic past trends, which may not always be the case.

Future Work

To build upon this study, future research could explore several avenues:

- *Longer Time Frame*: Analyzing data over multiple years could provide insights into long-term trends and seasonal variations in crime rates.
- *Wider Geographic Scope*: Expanding the analysis to include different cities or regions would enhance the generalizability of the findings.
- *Socio-Economic Factors*: Incorporating socio-economic data could help in understanding the underlying causes of crime patterns.
- Advanced Modeling Techniques: Employing more sophisticated predictive models, perhaps incorporating machine learning algorithms, could yield more accurate forecasts.

Final Project -CIS 635- Crime Data Forecasting

6. Data and Software Availability

Data: https://nij.ojp.gov/funding/real-time-crime-forecasting-challenge-posting#data-
Platform :Google Colab
(https://colab.research.google.com/drive/1k3WHUylDBKlno8JhCdRvoEbcNk6gYr5B#scrollTo=w7IZvVG97gEg)

7. References

- Chainey, S., Thompson, L., & Uhligh, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. Security(21), 4-28.
- Hunt, J. (2016). Do Crime Hot Spots Move? Exploring the Effects of the Modifiable Areal Unit Problem and Modifiable Temporal Unit Problem on Crime Hot Spot Stability. Archived with ProQuest Dissertations & Theses
- https://wilber-learndev.com/2022/03/10/types-of-crime-mapping/
- Weisburd, D., Eck, J.E. (eds.): Crime and Place: Crime Prevention Studies, Vol. 4. Police Executive Research Forum: Willow Tree Press (1995)