Analysis of Criminal Details and Predict the Severity of Crime in New York

Group 10: Xuchao Gao, Yiqun Liu, Yuxuan Liu, Zhuoqing Xu, Shuhua Yu, Yi Li
Vinay Mahadeo
Apan 5205 Project report
April 29th, 2024

Introduction & Scenario Description

Over recent years, urban crime has been increasingly perceived as a critical issue dramatically affecting the safety and quality of life within metropolitan environments. New York City is characterized by diverse population features and dense boroughs and has shown varying patterns of criminal activity. Such fluctuations have raised considerable concerns among residents and authorities. It is therefore particularly important to understand and predict the spatial and temporal distribution of crimes, not only in relation to public safety issues but also in relation to urban planning and law enforcement strategies. This research will, therefore, dissect the types of crimes happening in the different boroughs and further identify the periods when these crimes happen most often based on New York Police Department alarm data. Although the basic data analysis gave the first overview, it did not supply the necessary groundwork to enforce measures that could have effectively reduced the crime in New York. Hence, this study aims to bridge this gap by advancing from descriptive to predictive analytics. With an accurate predictive model built, the research would be able to empower the civilian and the law enforcement, with the power of foresight. This anticipatory form of policing and urban safety, driven through community awareness, is set to give a major fillip to hopes that cities will become better places to live in. More broadly, it is our goal that through this research, we contribute to greater efforts to ensure public safety and peace of mind for all citizens of New York City by building a stronger knowledge base and understanding of how crime is most effectively and proactively deterred.

Research Problem

How to analyze ten years of criminal details and predict the severity of crime in New York based on multiple factors(time, area, etc.), and optimize the allocation of police resources based on the data prediction results.

The problem of social security has become more and more serious in the United States, and more and more people living in the United States have been seriously threatened with their lives. Everyone, whether students or people in the society, may face life threats. The research focuses on predicting the types of crimes and severity that occur based on police records, based on street, time of day, jurisdiction, etc., through modeling or machine learning. This research will help more people avoid entering areas of high risk of crime in advance, thus reducing the probability of crime in the population and reducing the crime rate in the city. We hope to know the analysis of criminal details and predict the severity of crime in New York for the last ten years based on multiple factors(time, area, etc.), and optimize the allocation of police resources based on the data prediction results. According to the breakdown, we will first analyze the most common crime months, the most common crime time of day, and the top 10 most common crime types during the year 2017-2023, and then use a predicting model to predict the severity of the crime

level. Based on the analysis and prediction, we will advise the New York Police Department on how to optimize the allocation of police resources and effective crime prevention strategy.

Literature Review

Crime prevention has to be approached in a different way as urban environment complexities are growing. Technological advancement in data analytics and machine learning is now being applied by researchers to urban data in predictive analytics to predict crime events. This literature review covers two important studies in this field, their methods, results, and implications for predicting urban crime.

In its International Journal of Geoinformatics, a recent article delved into how point-of-interest (POI) data can be used to predict urban crime risk. For the case of predicting crime risk based on the attributes of points of interest in urban areas, the researchers applied a plethora of algorithms (logistic regression, SVM, decision trees, random forests) (Cichosz, 2020). It undertook a 10-fold cross-validation to establish reliable performance estimates. Results have indicated that POI attributes hold the potential for better prediction performances of the Random Forest algorithm across different urban areas and types of crime. The approach puts in emphasis the urban data to improve predictive policing and public safety strategies.

Another useful recent systematic review was published in the Journal of Crime Science, which adds important insights to the broader spatial crime prediction research with existing empirical studies on POI data. This review aims at finding the answers to some of the key questions like: What are the types of predictive information that space is playing an important role in? What are the common prediction methods that are in practice? What are the model validation strategies?(Kounadi et al., 2020). A review based on a range of studies using methods such as decision trees, neural networks, and support vector machines, underscores the growing importance of hybrid models to enhance predictive accuracy. The review summarizes the latest technological and methodological advances in the field and points to the effectiveness of spatial approaches to crime prediction; it will provide new illumination for researchers in this sphere. Overall, this comprehensive review points to a growing consensus of the utility of data-driven methods in crime prediction and towards important future research directions and potential pitfalls. In general, the studies have highlighted the emergent trend that is crime prediction research and its important value for informing urban planning, enforcement strategies, and community safety initiatives.

Database Overview

Data Collection

We collect the data from the NYPD official website which contains the crime data from 2009 to 2024. This large database enables us to get an overall view of how crime happened in New York during this year. Although there may be some null value, compared with the existing data we have, it can be a noise for the whole data collection. Therefore, we choose this dataset as our data collection basement.

Variable:

1. **CMPLNT_FR_DT** - Exact date of occurrence for the reported event (Date & Time)
2. **CMPLNT_FR_TM** - Exact time of occurrence for the reported event (Plain Text)
3. **ADDR_PCT_CD** - The precinct in which the incident occurred (Number)
4. **RPT_DT** - Date event was reported to police (Date & Time)
5. **KY_CD** - Three-digit offense classification code (Number)
6. **OFNS_DESC** - Description of offense corresponding with key code (Plain Text)
7. **PD_CD** - Three-digit internal classification code (more granular than Key Code) (Number)
8. **PD_DESC** - Description of internal classification corresponding with PD code (Plain Text)
9. **CRM_ATPT_CPTD_CD** - Indicator of whether crime was successfully completed or attempted (Plain Text)
10. **LAW_CAT_CD** - Level of offense: felony, misdemeanor, violation (Plain Text)
11. **BORO_NM** - The name of the borough in which the incident occurred (Plain Text)
12. **PREM_TYP_DESC** - Specific description of premises; grocery store, residence, street, etc. (Plain Text)
13. **JURISDICTION_CODE** - Jurisdiction responsible for incident (Number)
14. **Latitude** - Midblock Latitude coordinate for Global Coordinate System (Number)
15. **Longitude** - Midblock Longitude coordinate for Global Coordinate System (Number)
16. **X_COORD_CD** - X-coordinate for New York State Plane Coordinate System (Number)
17. **Y_COORD_CD** - Y-coordinate for New York State Plane Coordinate System (Number)
18. **SUSP_AGE_GROUP** - Suspect's Age Group (Plain Text)
19. **SUSP_RACE** - Suspect's Race Description (Plain Text)
20. **VIC_SEX** - Victim's Sex Description (Plain Text)

Research Plan

Sample Selection:

Ignoring uncontrollable factors such as unreported crimes and false reports, we assume that this dataset represents all crimes in the entire New York area during a specified time period and scope from 2009 to 2024 (specific circumstances include crime type, specific location, specific time, etc.). Given that this dataset was collected over a period of 15 years and the size of the data, we believe that factors such as false crime reports can be ignored and that this dataset can be used as a population representative of all crime incidents in the New York area.

Considering that the sample will be used for the subsequent prediction model development and statistical analysis, the sample selection will be a subset of all the data in the population (that is, the data set) involved in the subsequent crime prediction and comparison (the sample may be included, for example, day-time of crime. crime type, place type, and specific location, etc.). Considering that the huge data set may contain false positives, duplicates, omissions, and other error information, the final sample selection will involve data processing methods such as cleaning.

Operational Procedures

The process of obtaining a complete statistical element involves data science processing of the original data set (population) to obtain professional and representative statistical results. The first step will be to review the data set and filter the categories of dependent variables, independent variables, data subsets, etc. that will be used in the subsequent analysis and construction of the prediction model. Secondly, the unprocessed original data set contains error information (such as duplicate data, missing data, etc.), so it is necessary to use basic data set cleaning and filtering methods to eliminate useless data, reduce the error of the original data set, and obtain samples. Then, we will segment the sample (i.e. the processed data set) again. The sample will be segmented into training, validation, and test sets for model development and evaluation. This ensures that our subsequent predictions and comparisons will be trained, tested, and validated on a separate subset of data. At this stage, we can basically complete the preliminary acquisition and processing of experimental data for data analysis.

Next, we will divide into two groups and use two technical means to analyze the problems mentioned in the research problem. First, we will use the neural network model to build a highly accurate prediction model for predicting the severity of criminal activities, so as to provide reference for police actions and police force allocation. At the same Time, we will also use the Time series technology to provide detailed analysis of criminal activities and police situation in New York City in the past ten years to assist the prediction results and provide more references for NYPD's actions and police resource allocation. Next, we'll detail how to apply these two techniques to achieve the results we want.

Analytical Procedure and Interpretation

Part 1: **Neural Network Prediction**

- In the first three versions, we mainly used multinomial logistic regression to process the data, but achieved less accuracy. However, after research, in our final version of the code, we used the neural network in the Keras library to make predictions for more complex nonlinear patterns.

Initial model (V1-V3) :

- 1st Edition: Uses date-time conversions, factor encoding of categorical variables, and simple multi-categorical logistic regression. However, the initial accuracy is very low, which indicates that the complexity or features of the model are not enough.

```
data$CMPLNT_FR_DT <- as.Date(data$CMPLNT_FR_DT, "%m/%d/%Y")
data$CMPLNT_FR_TM <- as.POSIXct(data$CMPLNT_FR_TM, format="%H:%M:%S")
categorical_columns <- c("OFNS_DESC", "PD_DESC", "BORO_NM")
data[categorical_columns] <- lapply(data[categorical_columns], as.factor)
```

- 2nd Edition: Binary variables for the month and time of day (e.g., morning, afternoon, evening, evening) were added to capture time patterns more clearly.

```
data <- mutate(data,
               Morning = ifelse(Hour >= 6 & Hour < 12, 1, 0),
               Afternoon = ifelse(Hour >= 12 & Hour < 18, 1, 0),
               Evening = ifelse(Hour >= 18 & Hour < 24, 1, 0),
               Night = ifelse(Hour >= 0 & Hour < 6, 1, 0))
```

- 3rd Edition: Improved data set, including data from 2016 onwards, and continued the enhanced features of the second edition.

Result:
Accuracy: 0.489285455336427. This result indicates that our model has only a low accuracy rate and there is still a lot of room for improvement.

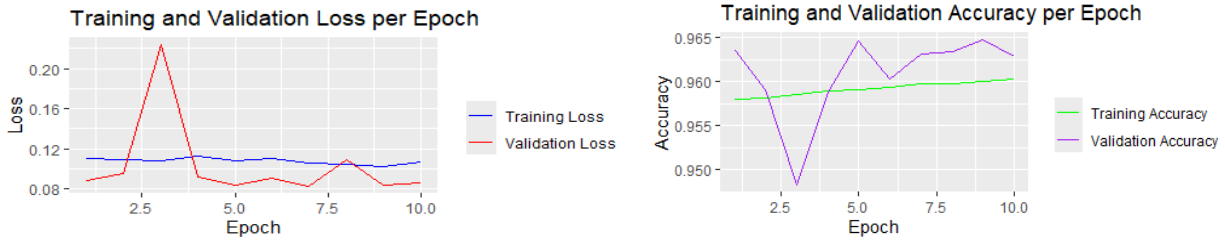
Final model:

- A neural network approach based on the Keras library enables deeper learning and more complex data representation. The model includes a dense input layer and a softmax output layer, optimized by an Adam optimizer and a classification cross entropy loss function.

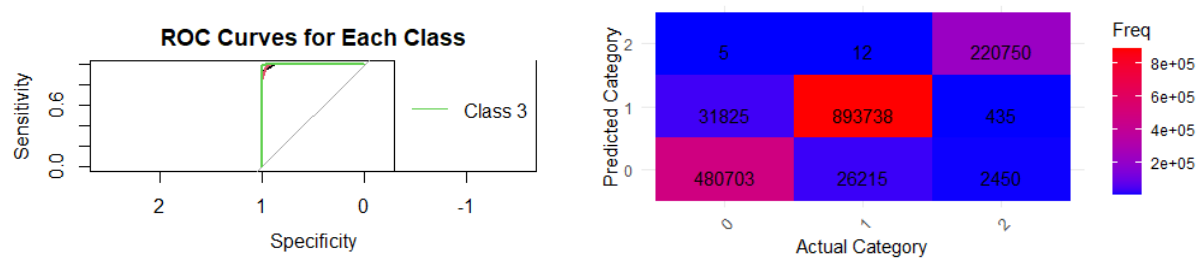
```
model <- keras_model_sequential()
model$add(layer_dense(units = 64, input_shape = c(ncol(train_features)), activation = 'relu'))
model$add(layer_dense(units = 3, activation = 'softmax'))  # Adjusted to match the number of classes
model$compile(loss = 'categorical_crossentropy', optimizer = 'adam', metrics = list('accuracy'))
```

- Improved data preprocessing methods, removing less correlated features and treating missing values more rigorously.

- The model demonstrated high accuracy, showing that it effectively learned the data patterns associated with crime prediction.



- These screenshots show the losses during model training. As the number of Epochs increases, the losses decrease, which is usually a good sign. After the initial Epoch, accuracy hovered around 96%. This shows that the model performs well.



- From the two graphs above, the ROC curve and confusion matrix, we can see that our model has very high accuracy, indicating many correct predictions.

Part 2：**Time Series Analysis**

- Independent analysis of crime occurrence trends based on time series. This method uses historical data to identify patterns and make predictions independent of the neural network's predictions.

Reasons for code changes and key insights

- From Basic to Complex Models: The main reason for moving from logistic regression to neural networks is that the initial model failed to capture complex patterns in the data. Neural networks provide greater flexibility and the ability to learn nonlinear relationships.

- Feature Engineering: Adding binary indicators for months and time periods directly addresses the timing of crime occurrence, which is critical for accurate forecasting.

- Improved data quality: Enhanced data cleaning and preprocessing, such as removing rows with too many missing values and irrelevant features, to help the model focus on important predictors.

Key modifications to better understand the problem:

After switching to neural networks, the accuracy of predictions improved significantly, from less than 50% to more than 96%.

Given the ever-changing nature of urban crime, focusing the analysis on recent data (post-2016) may have improved the relevance and accuracy of the analysis.

Research results:

Our model is able to predict the probability of different crime types in New York City with high accuracy, which allows law enforcement to allocate resources more efficiently. The insights provided by this study could help the NYPD make more informed decisions about resource allocation and preventive strategies. By predicting crime probabilities, our program helps strengthen community safety by allowing residents and law enforcement officials to identify and prevent potentially high-risk crime areas in advance.

Impact and recommendations:

- Continue to refine the model to include more detailed data such as specific events or demographic changes that may affect crime patterns.

- Explore integrating real-time data sources in order to implement mobility State crime prediction and immediate law enforcement response.

- Extend the analysis to other metropolitan areas to adapt the model to different urban environments and crime dynamics.

How accurate are the results on the scale of the problem?

- The second model got it right 96% of the time, which is very impressive, showing its powerful predictability. Such a high accuracy rate means that it can greatly help the police plan ahead and take measures. But in a high-stakes field like crime prediction, even a small improvement can make a huge difference. In addition, it is important to ensure that the model makes equally accurate predictions for all types of crime.

What more is needed to better understand the problem?

- We should use more data to test whether our model works properly, such as other seasons and regions.

- In-depth study of the process of model prediction to better discover possible bias problems and avoid overfitting.

- It is necessary to combine the actual situation of the real world to judge whether the prediction of the data is reliable. And the continuous improvement to ensure the accuracy of the model.

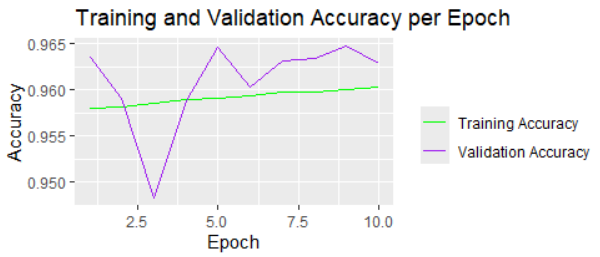
How well can analytical procedures answer research questions?

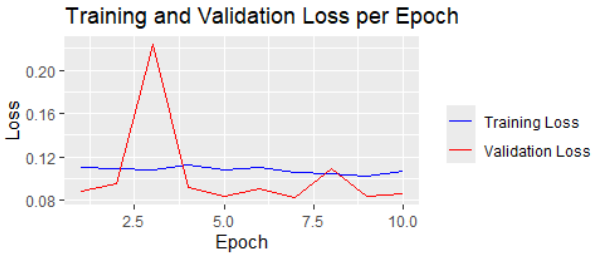
In the above analytical procedure, we have used time series and neural network models to obtain descriptive analysis results and models for predicting crime severity. The prediction model built by the neural network model (Part 1) can predict the severity of the upcoming crime with an accuracy of 96%. It's an accurate and credible answer to the Research Problem "how to predict the severity of crime in New York based on multiple factors(time, area, etc.) ". The analysis results obtained through the Time Series (Part 2) can be directly used to answer the question "how to analyze the most common crime months, the most common crime time of day, and the top 10 most common crime types during the year 2017-2023 ", and a statistically significant distribution pattern was obtained. Next, we will combine the analysis and prediction results to give answers and strategic suggestions for optimizing police resource allocation and crime prevention proposed in the research problem.

Strategic proposal from Analysis Interpretation

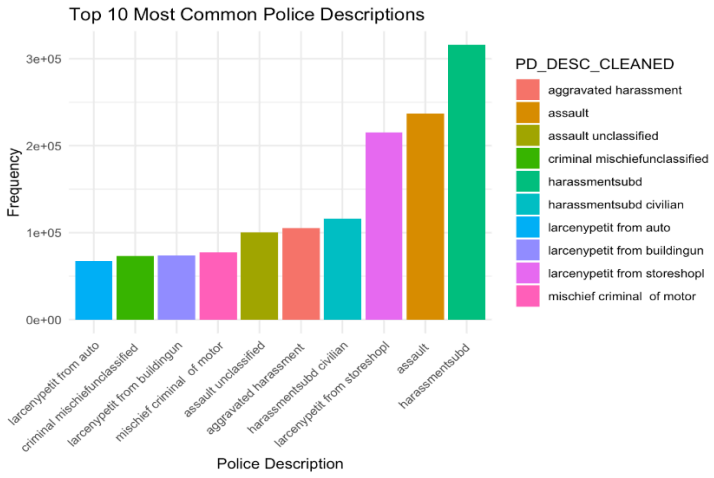
We are divided into two parts, one is the analysis of crime details based on the time series method, and the other is the prediction of crime severity based on prediction. As for the Prediction result, we suggest that NYPD may consider adjusting the allocation of police resources according to the prediction result of crime severity as appropriate. However, because the prediction model needs to be perfected, and crime situations are often random and unpredictable events, it is suggested to make a reasonable judgment based on the 10-year police situation analysis given by us, rather than relying entirely on the prediction.

Specific Suggestion and Choices for decision-makers
Decision Based on Prediction



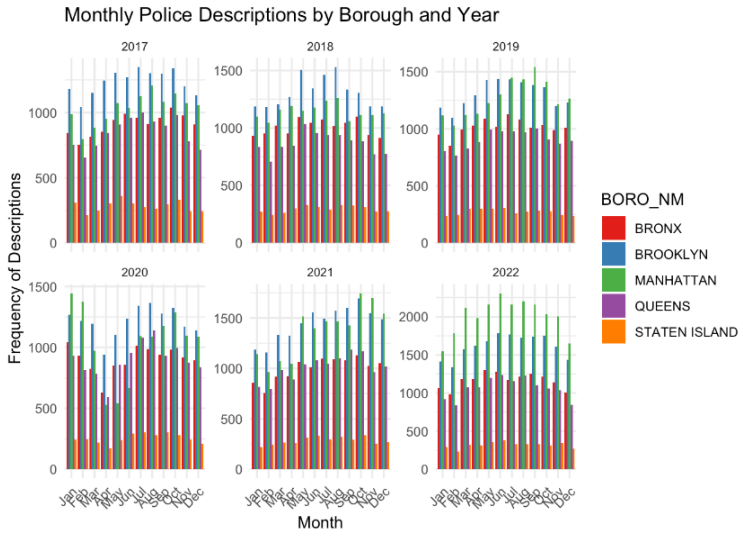


Test loss: 0.08603013 Test accuracy: 0.9632022



After testing, the accuracy of the NN prediction model built by us reaches 96%, so we believe that the police can refer to the predicted crime severity when they act, deploy or pre-arrange the distribution of police resources. If the severity of the crime is predicted to be high, then it is recommended to increase the number of police officers in a specific area or time period, patrol locations, or arrange for sufficient backup of police officers to be available at any time. At the same time, the allocation of police materials can also be adjusted according to the predicted severity of the crime, if the severity of the crime is low, there is no need to equip with advanced assault weapons or protective gear. If the severity of the crime is predicted to be high, it is recommended to equip the police with good protective materials to ensure the efficiency and safety of the operation.

Decision Based on Prediction
Crime Type



According to the above analysis, we can learn that the top three major crime types with the highest frequency are harassment, assault and larceny petit from stores or shop, which are less harmful to the police and the public. We recommend that

the NYPD equip most officers on duty with regular police protection for these three types of crime that are significantly more frequent than other types of crime, and for less frequent criminal cases or shootings, or for special operations, officers involved in operations with high-grade protection or firearms.

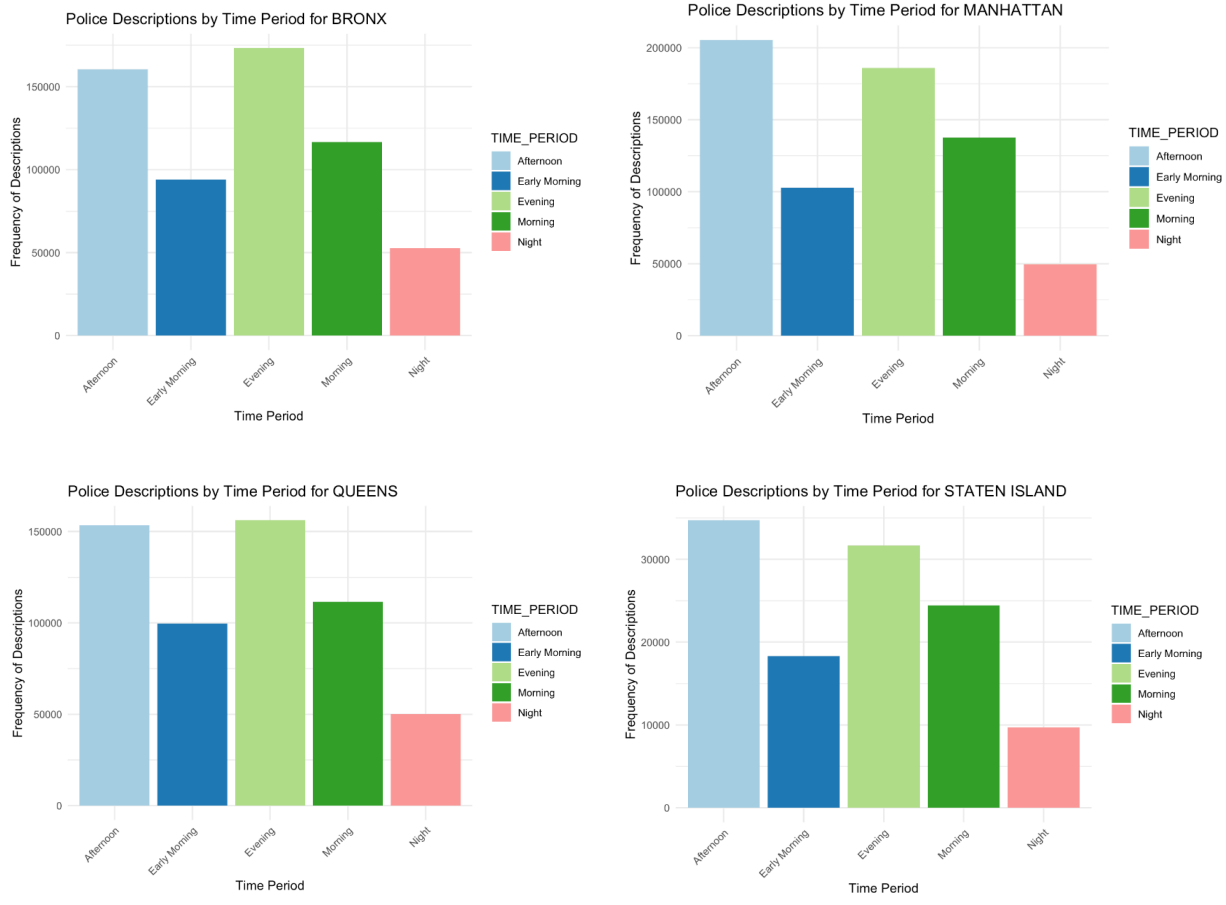
Monthly Police by Borough and Year

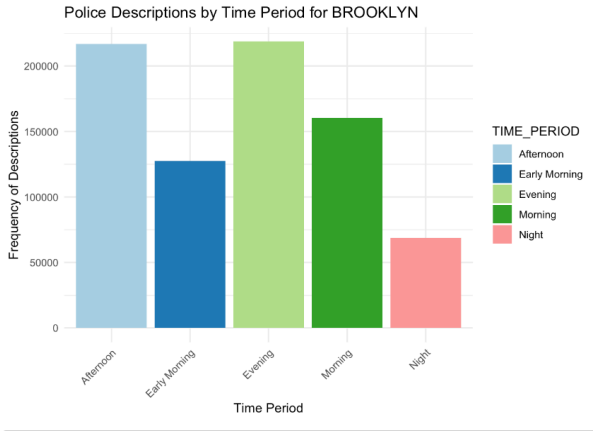
The most frequent crime areas are Manhattan and Brooklyn, and the frequency of crime incidents in these two areas is significantly higher than in other areas. So the idea is to focus on these two areas and divert most police resources. Crime control or prevention by increasing police patrol points, increasing manpower, rotating duty, and being ready to deploy police from other districts to the area. At the same time, because of the high frequency of crime in these two areas, it is suggested that the budget and police resources should be appropriately tilted towards these two areas.

In the absence of significant high crime months throughout the year, it is suggested that NYPD does not need to make special police arrangements according to the month, but only needs to be fully prepared during the holidays and focus on the high crime periods in each single day.

Time Period for each Borough

In all areas, the frequency of crime is significantly higher in the afternoon and evening than in the other two time periods, and it is recommended that more officers be on patrol or on duty during these two times and be ready to respond at any time. It can be reduced in other time periods.





Conclusion - Improvements

Based on the above independent strategic analysis, we can conclude a comprehensive action plan: **Combined with the crime severity forecast, focus on the areas of Manhattan and Brooklyn throughout the year, increase the number of police officers, patrol stations, or better police supplies in these areas during the two high crime time period(afternoon and evening).** Because New York City is very busy, the population is large, and the gap between rich and poor is very large, so crime is frequent. Especially in crowded, high-traffic areas such as Manhattan or Brooklyn, it is extremely difficult to carry out duty. These objective environmental factors cannot be changed, so it is suggested that NYPD can only start from optimizing internal police efficiency and optimizing the distribution of police resources to provide better protection for New York citizens. Especially in areas like downtown Manhattan, the NYPD should also take into account the increased difficulty of enforcement due to unexpected and unpredictable objective factors such as traffic jams.Since it is difficult to directly quantify how to improve the allocation of police resources, we propose to make gradual adjustments according to the direction given by us, gradually increase manpower or resources by 0.2 times in high-incidence areas and high-incidence periods, and make subsequent adjustments at any time according to prediction and realistic feedback. At the same time, because improvements are difficult to quantify in the short term by direct observation, it is recommended to continue to monitor the details of the New York City police for subsequent analytical adjustments.

* The complete code will be uploaded as an attachment.

Reference

Cichosz, P. (2020). Urban Crime Risk Prediction using point of interest data.

ISPRS International Journal of Geo-information, *9*(7), 459.
https://doi.org/10.3390/ijgi9070459

Kounadi, O., Ristea, A., De Araujo, A., & Leitner, M. (2020). A systematic review on spatial crime forecasting. *Crime Science*, *9*(1).
https://doi.org/10.1186/s40163-020-00116-7