Safeguarding NYC: Analysing Crime Patterns using Big Data

Arsalan Anwar
New York University
New York, USA
ax2134@nyu.edu

Somya Gupta
New York University
New York, USA
sg7885@nyu.edu

Raunak Shukla
New York University
New York, USA
rs8668@nyu.edu

ABSTRACT

The study "Safeguarding NYC: Analyzing Crime Patterns using Big Data" offers a thorough examination aimed at carrying out a detailed and comprehensive analysis of crime data obtained from several sources in New York City. The data sources include shooting data, crime data, frisk and search data obtained from NYC Open Data. The principal goal is to identify trends, levels of intensity, and geographic distribution of criminal activity, with a tactical intent to support various stakeholders, such as residents, travellers, NYPD and students, in their understanding and proactive reduction of criminal activity in the city. Through the use of sophisticated big data analytics tools such as Airbyte, Snowflake, Spark and Power BI, we propose to setup automated data pipelines which output insightful information that helps community organizations and law enforcement agencies alike make decisions that will improve public safety and foster a more secure urban environment.

KEYWORDS

Crime Analysis, NYC Crimes, NYC Safety, Big Data, Airbyte, Snowflake, Spark, Data Mining

ACM Reference Format:
Arsalan Anwar, Somya Gupta, and Raunak Shukla. 2023. Safeguarding NYC: Analysing Crime Patterns using Big Data. In *Proceedings of NYU Tandon (Big Data).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

New York City, a metropolis renowned for its vibrant culture and diversity, faces a persistent challenge: urban crime. Understanding and mitigating crime patterns is crucial for ensuring the safety and well-being of its residents. Traditional approaches to crime analysis, often relying on isolated datasets and limited methodologies, often fail to provide a comprehensive picture of the city's crime landscape. This gap necessitates a more holistic and data-driven approach.

In this proposal, we perform thorough crime analysis in New York City, leveraging the power of big data to gain deeper insights into crime patterns, intensities, and distributions. By consolidating and analyzing multiple datasets such as Use of Force (Subjects and Incidents) dataset, Shooting Incidents dataset, Stop, Question and Frisk dataset and Complaint Map dataset, we aim to identify spatiotemporal patterns, pinpoint crime hotspots, and uncover hidden relationships between crime incidents and sociodemographic factors.

We aim to setup robust, scalable and flexible data pipelines that ingest the huge amounts of data from various sources, transform it and then visualize comprehensive set of analytical patterns, including temporal, event-based, density-based, and people-focused (age, gender, origin, race) analysis. This multifaceted approach enables us to delve into the intricacies of crime patterns, revealing insights that would otherwise remain hidden.

Additionally, by addressing various social issues such as elder abuse, community-police relations, safety for newcomers, and student safety, our holistic strategy is more likely to yield effective results in reducing crime and improving overall safety. Our findings will empower stakeholders, including law enforcement agencies, community organizations, and policymakers, to make informed decisions that address the root causes of crime and foster a safer and more inclusive city.

2 PROBLEM STATEMENT

To provide an in-depth and integrated analysis of crime data from multiple sources in New York City. This study aims to identify patterns, intensities, and distributions of crimes, with the ultimate goal of aiding various stakeholders in understanding and mitigating criminal activities in NYC.

3 RELATED WORK

C. R. Herrmann [5] employs a geographical information system framework to investigate spatial and temporal variations in violent crime at the street level. By focusing on small-scale clusters and their temporal trends, the analysis delves into hotspot geographies generated by density maps, utilizing longitudinal crime data to discern the underlying causes of these hot spots. The advantages of this approach include its efficacy in directing police patrols efficiently, particularly through targeted efforts in specific small-scale areas. However, potential challenges include the need for robust data management tools, limited discussion on ethical considerations, and the risk of over-policing. While emphasizing the importance of small-scale research, the paper acknowledges the necessity for future studies to incorporate additional factors for a more comprehensive understanding of spatiotemporal relationships at the street level.

H. Corman and H. N. Mocan [3] propose an econometric approach to analyze the crime-supply equation, utilizing high-frequency time-series data from the New York City Police Department spanning nearly three decades (1970-1996). The model incorporates variables such as police force size, crime arrests, poverty (as a legal-market opportunities proxy), and drug use, providing a comprehensive understanding of crime dynamics. Advantages include

Big Data, Nov 2023, New York, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

the novel exploration of drug consumption's impact on criminal activity. However, the study faces challenges related to the simultaneous determination of crime, police, arrests, and drug use, impacting identification. Additionally, reliance on specific proxies and assumptions of linear relationships between variables introduce potential measurement errors and oversimplification of the complex dynamics of criminal behavior.

Chhaya Chauhan and Smriti Sehgal [2] outline a methodology of integrating advanced algorithms and tools, including the Advanced ID3 Algorithm, Hidden Link Algorithm, Apriori Algorithm, Naive Bayes Algorithm, Forensic ToolKit 4.0, and the Criminal Investigation Analysis Tool (CIA). This approach prioritizes attributes, detects covert connections in co-offender networks, identifies trends, and incorporates behavioral analysis for insights into unknown offenders. The advantages of this methodology lie in its automation for efficient crime analysis, predictive capabilities aiding in crime prevention, high accuracy rates, and remote investigation tools. However, limitations include potential accuracy issues in tools like CIA, dependence on the quality of crime data, and challenges in less technologically advanced areas. Ethical considerations related to predictive policing and data mining are acknowledged as crucial for responsible and unbiased crime analysis practices.

Rachel Boba [1] offers a comprehensive exploration of crime analysis, crime mapping, and problem-solving utilizing the SARA approach. The study employs both qualitative and quantitative methods to analyze crime and law enforcement data, providing a valuable resource for apprehending criminals, reducing crime, and evaluating organizational procedures. The incorporation of crime mapping, including manual pin mapping and computer mapping, aids in pinpointing high-crime areas, supporting targeted interventions. The SARA problem-solving approach is emphasized for its systematic methodology in community policing. Spatial analysis techniques, such as single symbol mapping and graduated mapping, are discussed to enhance practitioners' analytical capabilities. Despite its strengths, the guide lacks detailed instructions for practical implementation, is dependent on the quality of available data, and does not explicitly address ethical considerations associated with crime mapping, potentially leaving room for misuse or privacy concerns.

David F. Greenberg [4] tackles methodological challenges in studying the crime drop in New York City, employing various spatial units and mathematical models, including time series analysis, precinct-level panel analysis, and cross-city panel analysis. The author utilizes static and dynamic fixed effects panel models, recognizing the importance of choosing an appropriate spatial unit due to crime rate variations across neighborhoods. While the strengths lie in the meticulous consideration of methodological issues, the paper acknowledges potential data inaccuracies and justifies the use of NYPD crime counts. Advantages include insights into localized patterns, the depth provided by panel models, and transparency in addressing data limitations. However, limitations include the impact of data inaccuracies, potential oversimplification of crime decline uniformity across neighborhoods, and a focus on felonies omitting a comprehensive view of the overall crime landscape.

Jerry H. Ratcliffe and Michael J. McCullagh [7] introduce a novel methodology for crime pattern analysis, focusing on temporal exploration using aoristic crime data. Aoristic analysis, compared to averaging and rigid searches, proves effective in handling temporal uncertainty by considering all possible time intervals for each incident. This approach reveals detailed temporal patterns, as evidenced by the identification of a weekly Monday peak in motor vehicle crime. Despite its advantages in uncovering intricate temporal trends, aoristic analysis introduces complexity due to potentially larger numbers of incidents, requiring careful consideration in subsequent statistical operations. While offering a comprehensive understanding of temporal patterns, it may not provide straightforward causal inferences, emphasizing its role as a tool for pattern identification rather than elucidating underlying causes. The methodology's effectiveness depends on factors such as database size and the accuracy of temporal fields.

The tool, NeighborhoodScout [6], employs a methodology that involves collecting and analyzing diverse data from public sources to offer users comprehensive insights into the safety, livability, and real estate conditions of U.S. neighborhoods. By utilizing proprietary algorithms, the platform processes crime reports, school quality assessments, demographic details, and real estate data. This allows users to make informed decisions about where to live based on factors important to them. While providing valuable information for free, in-depth reports require a subscription. The free version offers a generic overview of crime without detailed specifics, such as crime types or temporal analyses, potentially limiting its scope for certain users.

4 METHODS, ARCHITECTURE AND DESIGN

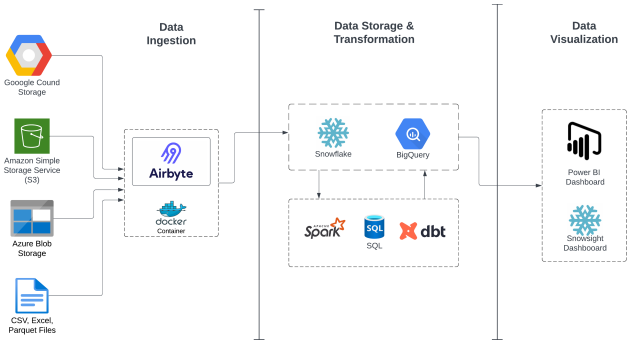

Figure 1: Proposed Data Pipeline

The overall architecture of the data pipeline described above is as follows:

- **Data Ingestion:** The data ingestion phase is responsible for bringing data from various sources into the data pipeline. In the image provided, Airbyte is used to ingest data from various sources, including Google Cloud Storage, Amazon S3, and Azure Blob Storage. Airbyte is a cloud-based data integration platform that can connect to over 400 different data sources. It provides a unified view of all data sources, making it easy to ingest data into the data pipeline.
- **Data Storage and Transformation:** The data storage and transformation phase is responsible for storing and transforming the data ingested in the previous phase. In the image provided, Snowflake is used to store the data. Snowflake is

a cloud-based data warehouse that is designed to handle large amounts of data. It is scalable and can support a variety of different data types. Spark is used to transform the raw data stored in Snowflake, process it and store it back on Snowflake.

- **Data Visualization:** The data visualization phase is responsible for visualizing the data transformed in the previous phase. Snowsight dashboard would be used to visualize the data which is based on query results on the Snowflake database.

5 RESULTS

Our comprehensive analysis of New York City crime data reveals several key patterns and distributions that are essential for stakeholders to understand the dynamics of criminal activities within the city.

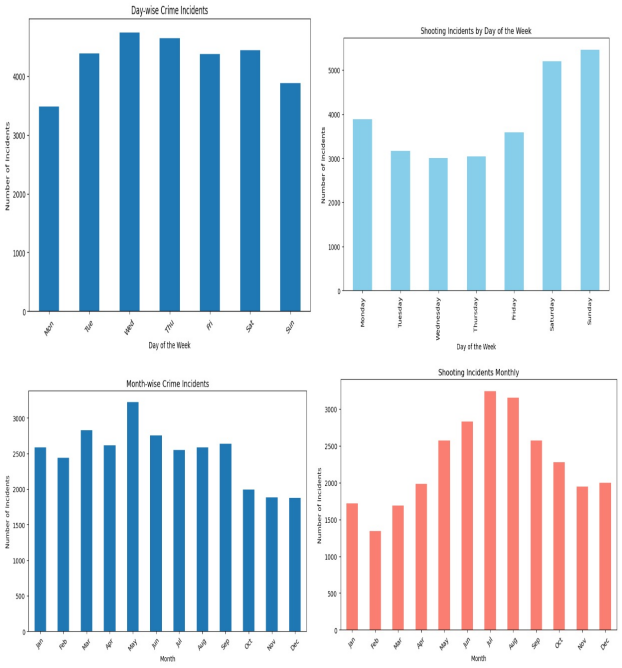

Figure 2: Temporal Crime Analysis of NYC

Fig 2. shows that crime incidents fluctuate throughout the week, with the highest occurrences midweek and a noticeable drop towards the weekend. This trend could influence how community policing is scheduled. Similarly, our month-wise analysis indicates a peak in criminal activities during the summer months, with a significant drop as the year ends. Such insights can assist in seasonal adjustments to policing strategies.

Fig 3. depicts the suspect age distribution that shows a concentration in the 20-40 age bracket, indicating a demographic that may require targeted intervention programs. In contrast, victim age distribution highlights that individuals between 25-44 years are the most affected group, which can inform preventive measures for this cohort.

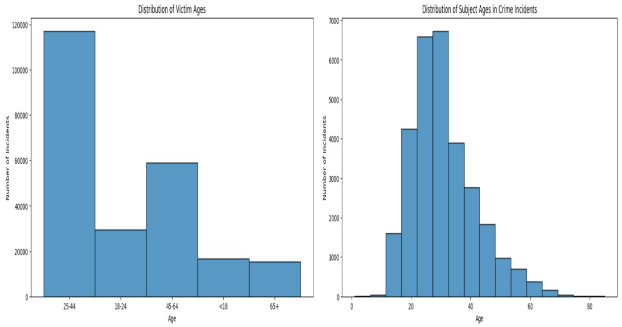

Figure 3: Victim and Suspect Age Distribution vs. No. of Crimes

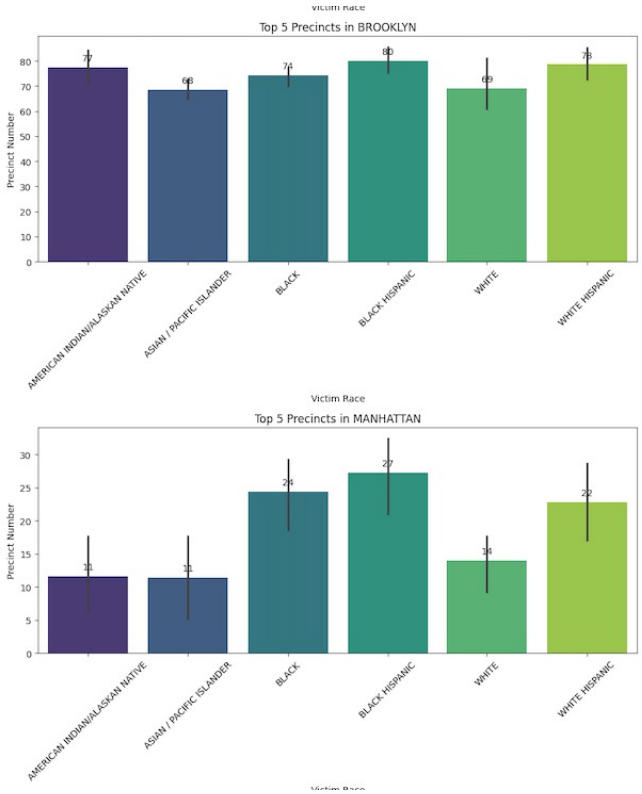

Figure 4: Geospatial Distribution of Crimes in NYC

As seen in Fig 4., investigating the most dangerous precincts across the boroughs reveals that crime is not uniformly distributed throughout the city. This information is crucial for New Yorkers, especially the elderly, new immigrants, and students, to understand and avoid dangerous areas and for law enforcement to allocate resources effectively.

Fig 5. indicates disparities that may need to be addressed to improve community-police relations. The data on crime incidents by race in various precincts provides a granular view of how crime

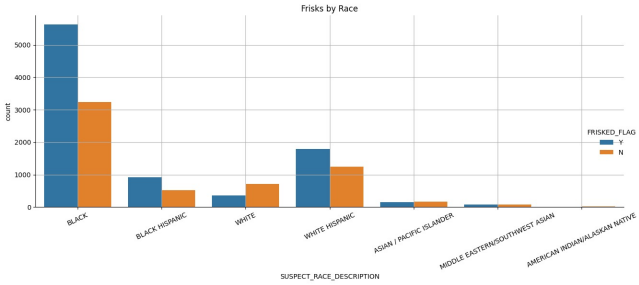

Figure 5: Analysis of Frisks Distribution by Race

affects different communities, which can help in creating informed, inclusive, and equitable policing policies.

REFERENCES

[1] Rachel Boba. 2001. Introductory guide to crime analysis and mapping. https://www.ojp.gov.

[2] Chhaya Chauhan and Smriti Sehgal. 2017. A review: crime analysis using data mining techniques and algorithms. *Proceedings of the 2017 International Conference on Computing, Communication and Automation (ICCCA)*, 21–25. DOI: 10.1109/CCAA.2017.8229823.

[3] H. Corman and H. N. Mocan. 2000. A time-series analysis of crime, deterrence, and drug abuse in new york city. *American Economic Review*, 90, 3, 584–604. DOI: 10.1257/aer.90.3.584.

[4] David F. Greenberg. 2014. Studying new york city's crime decline: methodological issues. *Justice Quarterly*, 31, 1, 154–188. DOI: 10.1080/07418825.2012.752026.

[5] C. R. Herrmann. 2013. Street-level spatiotemporal crime analysis: examples from bronx county, ny (2006–2010). *Crime Modeling and Mapping Using Geospatial Technologies*, 73–104. DOI: 10.1007/978-94-007-4997-9_4.

[6] NeighborhoodScout. 2000. New york crime. https://www.neighborhoodscout.com.

[7] Jerry H. Ratcliffe and Michael J. McCullagh. 1998. Aoristic crime analysis. *International Journal of Geographical Information Science*, 12, 7, 751–764. DOI: 10.1080/136588198241644.