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

"Safeguarding NYC: Analyzing Crime Patterns using Big Data" is a pivotal study that delves into the complex landscape of urban crime in New York City, utilizing extensive datasets from NYC Open Data, including shooting incidents, use of force, and stop, question, and frisk records. This project aims to discern crime trends, intensity levels, and geographic distributions to aid various stakeholders—from local residents and immigrants to the NYPD—in understanding and mitigating criminal activities. We establish automated data pipelines to extract, process, and visualize crime data employing various big data tools/technologies such as Spark, Snowflake and Airbyte. The insights generated and presented through the study are designed to empower community organizations and law enforcement agencies with actionable intelligence, ultimately enhancing public safety and fostering a more secure urban environment.

KEYWORDS

Crime Analysis, NYC Crimes, NYC Safety, Big Data, Snowflake, Spark, Airbyte, ELT, Data Mining, Medallion Architecture, Data Transformation, Data Pipelines

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, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

New York City, a vibrant mosaic of cultures, confronts the intricate challenge of urban crime. Traditional crime analysis methods, constrained by limited data scopes, fail to fully capture the city's complex crime patterns. Our study, 'Safeguarding NYC: Analyzing Crime Patterns using Big Data,' addresses this shortfall by leveraging expansive big data techniques. By analyzing diverse datasets, including Use of Force, Shooting Incidents, and Stop, Question and Frisk records, we explore spatio-temporal crime trends, identify critical hotspots, and examine the relationship between crime and socio-demographic factors.

By integrating and scrutinizing diverse datasets, including the Use of Force: Subjects and Incidents, Shooting Incidents, and Stop, Question and Frisk, we aim to uncover spatiotemporal crime patterns, pinpoint critical hotspots, and discern the interplay between crime and socio-demographic elements.

Our methodology involves establishing comprehensive data pipelines, capable of processing large data volumes and enabling detailed analysis across temporal, event, density, and demographic dimensions. This extensive big data investigation aims to provide insights beyond traditional analyses, deepening our understanding of crime trends.

Through this comprehensive big data analysis, our study seeks to reveal insights that transcend traditional analyses, shedding light on underlying factors and offering a deeper understanding of crime trends. This endeavor is not just an academic exercise; it aims to inform and empower a wide range of stakeholders, from law enforcement to community groups and policymakers. Ultimately, our goal is to contribute to more effective crime reduction strategies and improve urban safety, ensuring New York City remains a secure home for its diverse population.

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 the novel exploration of drug consumption's impact on criminal

Big Data, Dec 2023, New York, USA

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

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

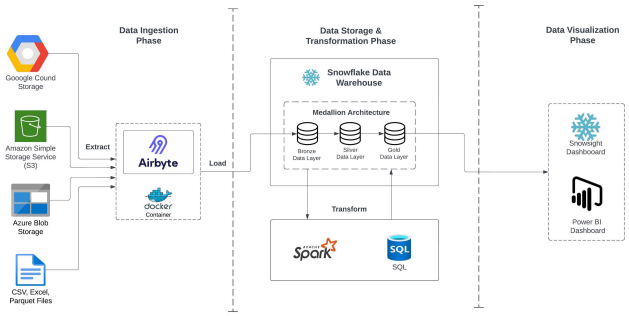

Figure 1: Proposed Data Pipeline

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

- **Data Ingestion:** The initial phase of our data architecture involves the meticulous ingestion of data. Leveraging Airbyte as shown in Fig. 2, a cloud-based integration platform, we seamlessly imported datasets from Google Cloud into Snowflake. This phase dealt with four primary datasets: Shooting Incidents, Use of Force (Subjects and Incidents), and Stop, Question and Frisk, collectively encompassing over 218,000 records. The efficient handling of data ingestion lays the groundwork for subsequent analysis, ensuring data integrity and availability.
- **Data Storage:** The data storage phase is responsible for storing the data ingested in the previous phase. In the image

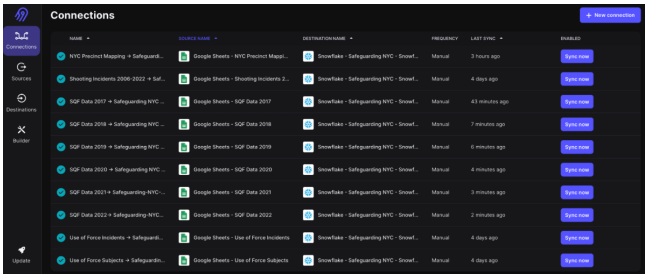

Figure 2: Data Ingestion using Airbyte into Snowflake

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.

A key component of our architecture is the Medallion Architecture, characterized by its three-tiered structure:

(1) **Bronze Layer (Raw Data):** This foundational layer houses raw, unprocessed data. It serves as a repository for data in its original form, ensuring traceability, flexibility and authenticity.
(2) **Silver Layer (Processed Data):** At this intermediate stage, the data undergoes significant processing. This includes cleansing of duplicates, standardization of formats, and correction of errors. Additionally, data is enriched with contextual dimensions like geospatial and temporal references, enabling more sophisticated analyses.
(3) **Gold Layer (Business Data):** Representing the culmination of the data processing and transformation pipeline, this layer holds the most refined form of data. Here, the focus is on deriving business-centric metrics and insights, such as crime rates and emerging patterns. This data is primed for strategic applications and decision-making processes.

- **Data Transformation:** The transformation phase was both extensive and meticulous, involving several key steps. Spark is used to transform the raw data stored in Snowflake, process it and store it back on Snowflake in the respective layers. Some of the transformations performed on the data are as follows:

(1) **Merging and Dropping:** To create a cohesive dataset, we merged data from various sources. Simultaneously, we eliminated superfluous columns to streamline the dataset for analysis. We orchestrated a comprehensive merger of datasets, integrating 'Use of Force: Subjects' with 'Use of Force: Incidents', and synchronizing the 'Neighborhoods' dataset with our primary datasets. This strategic consolidation enhances the robustness and granularity of our analysis, providing a multifaceted view of the data.
(2) **Formatting for Consistency:** We ensured that diverse data points were uniformly formatted, adhering to a standardized business format. This step was crucial for maintaining consistency across datasets. In the data harmonization process, specific racial categorizations like 'AMER INDIAN' and 'ASIAN / PAC. ISL' were standardized to 'AMERICAN INDIAN' and 'ASIAN/PACIFIC ISLANDER' respectively. This meticulous reclassification ensures uniformity across the various datasets, thereby enhancing the integrity and comparability of the analytical results.
(3) **Mapping and Categorization:** This involved addressing null values and appropriately categorizing data based on boroughs, neighborhoods, race, and gender. For enhanced clarity and business relevance, borough codes such as 'PBMN' for Manhattan and 'PBBN' for Brooklyn were meticulously mapped to their full borough names. This strategic data normalization fosters a more intuitive understanding and facilitates a smoother business analysis process.
(4) **Outlier Removal:** We meticulously identified and removed outliers, particularly in parameters like dates, weights, heights, and ages, to enhance data accuracy. Additionally, anomalous high-age values such as 944 and 1020 were identified and subsequently removed to maintain data quality and realism in the dataset, ensuring only feasible and representative age data were retained for analysis.

- **Data Visualization:** The final phase involves translating the processed data into actionable insights through visualization. Using the Snowsight dashboard, we created intuitive and informative visual representations of our findings, directly querying the Snowflake database. These visualizations not only illustrate the crime trends and patterns but also make the data accessible and understandable to various stakeholders, including community organizations, law enforcement agencies, and policymakers.

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.

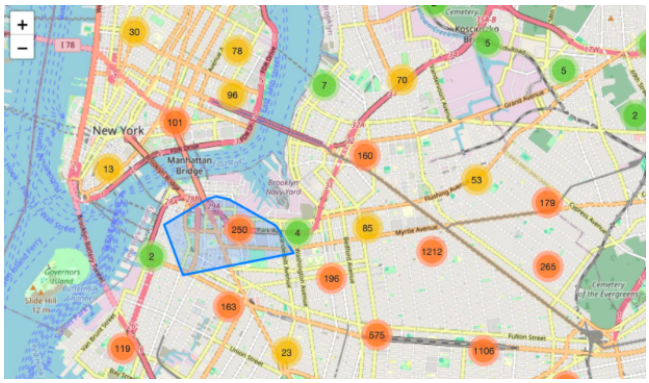

Figure 3: Geospatial Analysis of Shooting Incidents across New York City

In Fig. 3, the map depicts the geospatial distribution of crimes, mainly the shooting incidents across New York City, with specific focus areas highlighted, indicating the prevalence of such events. This visual analysis aids the NYPD in optimizing resource allocation, as

it clearly identifies areas with higher frequencies of incidents, suggesting locations that may benefit from increased police presence or community-based interventions to enhance public safety.

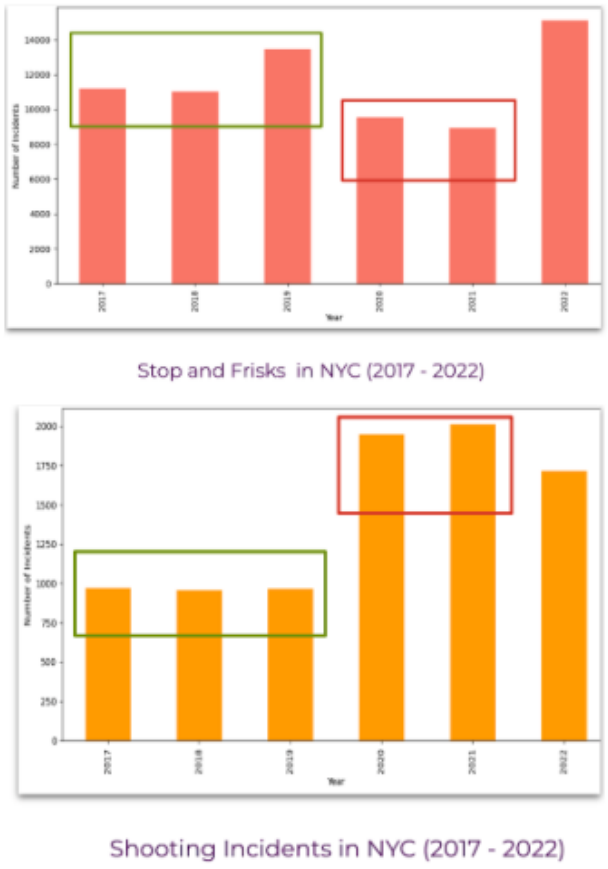

Figure 4: Impact of Search and Frisk Practices on Shooting Incidents (2017-2022)

Fig. 4 depicts the impact of NYPD's Inspection practices on shooting incidents in the last 6 years. We observed a clear trend: as frisks increase, shootings tend to decrease and as frisks decrease, shootings tend to increase. This inverse relationship highlights the deterrent effect of active policing measures on gun-related crimes. Therefore, it's important for policymakers to consider these findings in strategizing law enforcement activities to reduce future gun violence in the city.

Neighborhood analysis identifying areas with a high incidents of shooting cases is shown in Fig. 5. In the graph, we can see the neighborhoods with the most shooting incidents coming from Brooklyn with East NY being the top one which also saw a rise of 97

Fig. 6 presents a compelling visualization of weapon incident hotspots across New York City neighborhoods. Fordham leads, indicative of a significant need for targeted policing and community

Neighborhood	Shooting Increases (2017-22)
East Harlem	190%
Brownsville	116%
Central Harlem	100%
East New York	97%

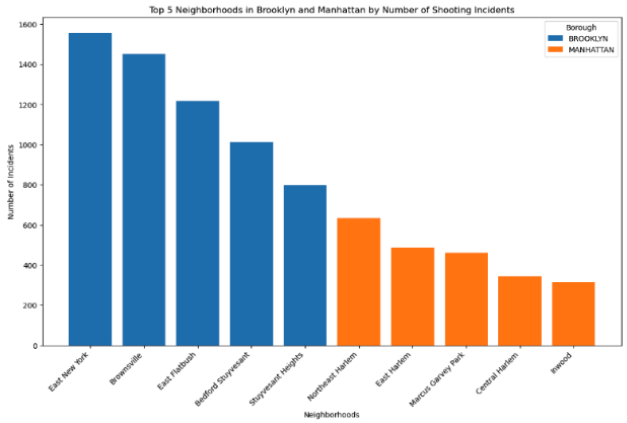

Figure 5: Neighborhood analysis highlighting areas with high incidents of Shooting Cases

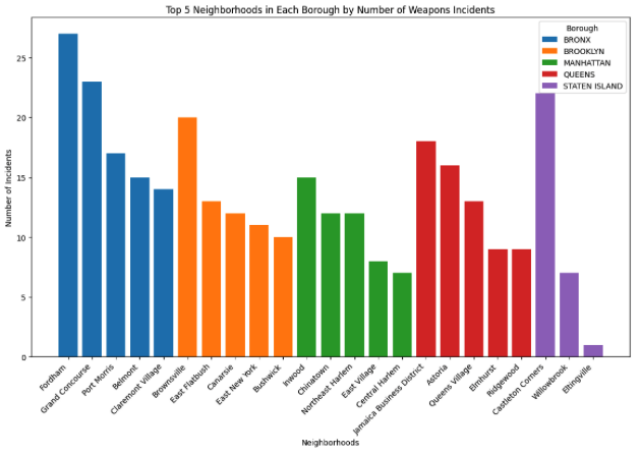

Figure 6: Analysis of Weapon Incident Hotspots across New York City

engagement strategies. Contrasting this, Staten Island neighborhoods like Willowbrook and Eltingville report notably fewer incidents, reflecting variances in social dynamics and urban infrastructure. This geographic data is critical for directing police resources efficiently and fostering crime prevention measures. Moreover, it

underscores the importance of community-specific strategies, integrating local insights with public safety initiatives to ensure tailored and respectful solutions.

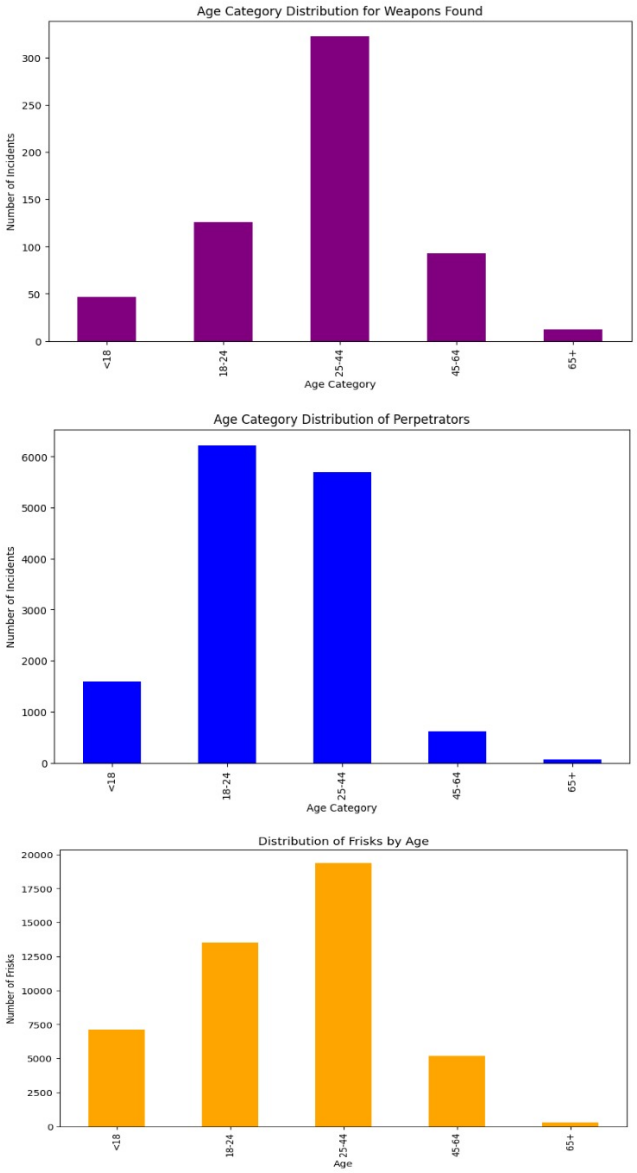

Figure 7: Analysis of Age Distribution by Crimes and Inspections

In Fig. 7, a significant correlation is observed between the age distribution of shooting perpetrators, distribution of weapons found and NYPD's frisk practices. The number of weapons found is significant within the 18-24 and 25-44 age bracket, which coincides with the peak age range for shooting perpetrators. This alignment likely influences the NYPD's strategy, leading to a higher number of frisks conducted on individuals within these age categories, as evidenced by the prominent peak in the "Distribution of Frisks by Age" graph. The data suggests a targeted approach in stop-and-frisk practices, correlating with the age groups most implicated in shooting incidents and weapon retrievals.

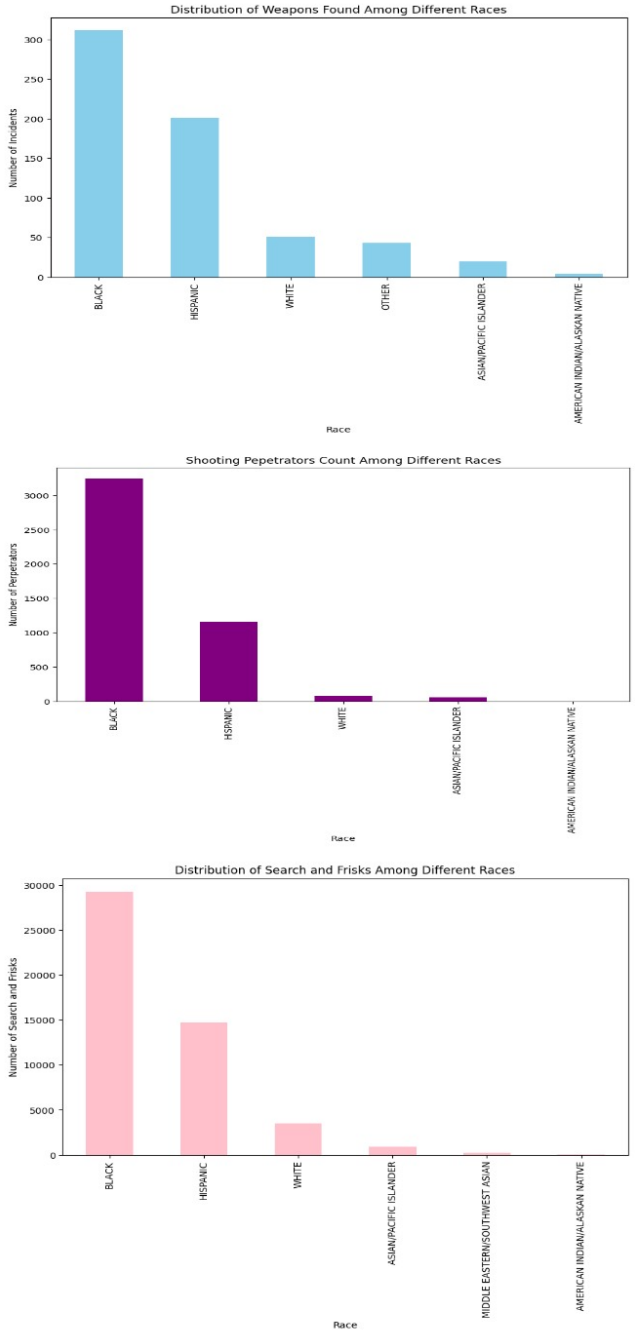

Figure 8: Analysis of Race Distribution by Crimes and Inspections

Fig. 8 sheds light on the distribution of crime across different racial groups in New York City. A notable aspect of our findings

is the correlation between the race of shooting perpetrators and the discovery of weapons, and how this impacts NYPD's search and frisk practices among different racial groups. It has been observed that people of African and Hispanic descent were stopped, searched and frisked more frequently than whites and people of other descents.

6 CONCLUSION

Our study, 'Safeguarding NYC: Analyzing Crime Patterns using Big Data,' has effectively navigated through complex crime data to unearth significant insights. We established sophisticated data pipelines using GCP, Airbyte, and Snowflake, enabling efficient data management. Our analysis revealed crucial crime trends and hotspots, evaluating NYPD's current strategies. These insights not only provide valuable information for stakeholders, including residents, law enforcement, and immigrants, but also lay a foundation for future urban safety research. This study underscores the importance of big data in understanding and enhancing urban security, offering a model for similar analyses in other metropolitan areas.

7 FUTURE WORK

In future expansions of our New York City crime and safety analysis, we plan to integrate machine learning models, especially clustering algorithms, to refine our understanding of crime locations. This approach, aimed at identifying intricate patterns in spatial distributions, builds upon the limitations of our current data-centric methods. Additionally, we propose to incorporate time series models, capitalizing on their potential to forecast crime trends. By addressing the current gap in predictive analytics, these models will offer valuable foresight for law enforcement and policymakers, enhancing proactive measures against crime. The integration of these advanced machine learning and predictive techniques will elevate our crime analysis framework, moving from mere observation to proactive prevention.

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.