AIT 664-009: Represent, Process & Visualize Applied Information Technology

PROJECT MILESTONE - I

Prof. Ebrima N Ceesay, Ph.D., CISSP

ANALYZING NEW YORK CITY POLICE DEPARTMENT (NYPD) ARREST DATA FOR CRIME INSIGHTS AND POLICING STRATEGIES

I. INTRODUCTION:

New York City, one of the world's busiest cities, is home to a diverse range of neighborhoods, cultures, and dynamics. As such, policing this city requires evidence-based strategies that are both effective and equitable. The NYPD Arrest Data (Year to Date) dataset provides a wealth of information, including the type of crime, location, time of enforcement, and suspect demographics. The New York City Police Department (NYPD) Arrest Data (Year to Date) dataset, available on data.gov, is a comprehensive repository of information regarding arrests made within the city. This dataset offers a unique opportunity to investigate the particulars of law enforcement, crime, and community dynamics in New York City's diverse and dynamic geography. The NYPD's goal is to safeguard public safety, uphold the law, and build positive relationships with the community, thus it is vital to scrutinize and interpret the massive amounts of data generated by these operations.

This project aims to analyze the NYPD Arrest Data using data science and analytics to answer crucial questions regarding crime patterns, law enforcement practices, and community impact. We recognize the significance of this endeavor considering its potential to not only inform policing policies, but also to improve transparency, accountability, and public trust in law enforcement institutions. [1] Through an in-depth analysis of the NYPD arrest data, this project seeks to shed light on the current state of crime and the effectiveness of law enforcement strategies. By leveraging this data, we aim to uncover patterns and trends that can inform policing strategies and contribute to public safety. [2]

II. DESCRIPTION OF THE PROBLEM:

Policing a city as diverse and dynamic as New York needs constant adaptation. Crime is a significant problem in New York City, and understanding crime patterns and trends is crucial for developing successful crime prevention and reduction tactics. The New York Police Department (NYPD) keeps a detailed record of all arrests made throughout the year. While this dataset is rich in information, it is also complex and large, making it difficult to extract useful insights manually. Although the NYPD Arrest Data (Year to Date) dataset contains a large amount of arrest information, it is not used to guide crime prevention measures, analyze community policing activities, or analyze the dynamics of crime in New York City. The aim of this project is to analyze and predict crime patterns using machine learning techniques based on this dataset. In the broader context, this project aligns with the growing trend in law enforcement towards data-driven decision-making. It aims to bridge the gap between the massive amounts of arrest data generated by the NYPD and the practical use of this data to enhance public safety and the efficiency of law enforcement operations. By identifying crime trends, hotspots, and crime types, governments and police departments can allocate resources more effectively and deploy their resources more strategically. By addressing these challenges and objectives, this project aims to provide valuable insights and recommendations for improving policing strategies, enhancing public safety, and building stronger community relations in the dynamic and diverse environment of New York City.

III. IMPORTANCE OF THE PROBLEM:

Crime is a major problem in New York City, and it is important to develop effective crime prevention and reduction strategies. Effective crime analysis can significantly enhance public safety and resource allocation in law enforcement. Understanding and analyzing the NYPD Arrest Data has a direct impact on public safety because it serves as the foundation for building successful law enforcement techniques that aid in the maintenance of security in a metropolis as dynamic and diverse as New York. Law enforcement organizations can strategically deploy resources for maximum impact by identifying potential crime hotspots and periods of increased criminal

activity. Furthermore, understanding the factors influencing crime can inform policy decisions and community engagement efforts. By examining the impact of community policing activities and resolving demographic disparities in arrests, this study can also strengthen community bonds and promote fairness and equity within the criminal justice system. Furthermore, using this dataset to feed predictive policing models, proactive law enforcement can be enabled, optimizing resource allocation for more efficient crime prevention and response.

IV. MOTIVATION:

The motivation behind selecting the NYPD Arrest Data (Year to Date) dataset for this project is driven by several key factors:

- **Enhancing Public Safety:** In cities like New York City, crime is a major concern. By analyzing this dataset, law enforcement might better understand crime trends and take proactive measures to deter and prevent criminal activity, which will eventually improve public safety.
- **Community Policing Enhancement:** The dataset enables an evaluation of community policing initiatives. Understanding the influence of community engagement on crime rates is crucial for establishing confidence between law enforcement and the community.
- **Equity and Fairness:** The dataset allows us to look into demographic differences in arrests. It is essential to identify and eliminate potential biases in law enforcement practices in order to ensure equity and fairness in the criminal justice system.
- **Resource Allocation Optimization:** This data can be used to create predictive policing models. These models can help with the optimal allocation of law enforcement resources, resulting in more effective crime prevention and response techniques.
- **Historical Crime Trends:** Comparing current arrest trends to historical data offers insight into how crime dynamics have changed over time. This historical framework is essential for understanding and adapting to changing patterns of crime.
- **Data-Driven Decision-Making:** The idea is in line with a broader trend in law enforcement to use data-driven techniques. The insights gained from this analysis can contribute to more informed decision-making and evidence-based policing strategies.

In summary, the NYPD Arrest Data (Year to Date) dataset was chosen to improve public safety, improve community relations, promote fairness within law enforcement, optimize resource allocation, and provide a data-driven approach to addressing crime dynamics in New York City.

V. LITERATURE REVIEW:

Several literature searches were conducted for the purpose of this study. In this literature review, we can explore relevant research, including studies on predictive policing, law enforcement strategies, and geospatial analysis, to provide valuable context and insights for our analysis.

The research “Minority Report” a Reality? The NYPD’s Big Data Approach to Predicting Crime by Clemens discusses how the NYPD has developed a data-driven approach to fight, and even predict, crime. Clemens' study investigates the utilization of big data and predictive policing by the New York City Police Department (NYPD) as a means to address and potentially predict criminal activity. The study highlights the incorporation of vast amounts of data, including crime locations, times, and types, into sophisticated computer models and algorithms. These models enable the prediction of areas with a higher likelihood of criminal activity. The study also emphasizes the use of “Big Data”-driven predictive policing by analyzing regularly recorded crime data (location, time, and crime), using sophisticated computer models and algorithms, to predict places of expected criminal activity. This approach is consistent with the core principles of our project since it prioritizes the implementation

of data-driven strategies. This study offers valuable insights on the utilization of crime data for the purpose of predicting areas with high criminal activity. [3]

The research article "The Effects of Local Police Surges on Crime and Arrests in New York City" published in PLOS ONE tested the effects of Operation Impact on reported crimes and arrests from 2004 to 2012 using a difference-in-differences approach. According to the findings of the study, Operation Impact was significantly associated with reductions in total reported crimes, assaults, burglaries, drug violations, misdemeanor offences, felony property crimes, robberies, and felony violent crimes. This study holds significant relevance with our project as it investigates the consequences of particular police techniques, providing valuable insights into the potential outcomes of data-driven policing initiatives. [4]

The study by Chainey et al. highlighted hotspot mapping as a useful method for predicting spatial crime patterns. These insights are critical when we analyze the NYPD Arrest Data, emphasizing data-driven ways to improving New York City law enforcement strategies. Chainey stresses the importance of geographical analysis, particularly the value of hotspot mapping, as a complement to the temporal dimension. Their study demonstrates how spatial analysis methods can reveal spatial trends in crime, an important component of crime analysis. Geospatial analysis reveals crime hotspots—specific geographic areas with higher concentrations of criminal activity—when applied to datasets like the NYPD Arrest Data. This knowledge enables law enforcement organizations to concentrate their efforts on specific regions, using manpower, resources, and preventive measures to effectively target and deter criminal activity. It also helps to comprehend crime displacement by demonstrating whether criminal activity moves from one site to another because of law enforcement efforts, allowing for the development of adaptive measures. These studies provide unique and significant insights into the study of NYPD arrest data, which may be relevant to our project. They offer insights on predictive policing, the influence of various law enforcement techniques, and the significance of geospatial data. We can obtain a better knowledge of the background for our analysis, identify areas of interest, and evaluate the findings of our investigation by reviewing these papers. They collectively emphasize the role of data-driven methods in improving law enforcement strategies and enhancing public safety. [5]

Each of these papers offers a distinctive viewpoint on the analysis of NYPD arrest data, and they may be able to provide us useful information for this project. These help us in understanding the larger context of our analysis, identifying prospective areas of interest, and evaluating the results of our investigation.

VI. RESEARCH QUESTIONS:

The research questions that can be answered by exploring the dataset include:

1. Are certain crimes more common during specific seasons or times of the day?
2. Where are the areas with the most arrests for different crimes in NYC?
3. Are there differences in arrest rates for different racial or ethnic groups?
4. Which age group has committed more crimes?
5. Which area has more sex crimes?

VII. DATA SOURCE:

The primary data source for this project is the "NYPD Arrest Data (Year to Date)" dataset, which is publicly available on the NYC Open Data portal and on the DATA.gov website. This dataset provides comprehensive information about every arrest made by the NYPD during the current year, including details about the type of crime, the location and time of enforcement, and suspect demographics. The data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning. This dataset will serve as the foundation for our analysis and insight generation. [1]

VIII. DESCRIPTION OF THE DATASET:

The given dataset is made up of arrest records, each of which is uniquely recognized by an **ARREST_KEY**. It also includes a variety of attributes pertaining to each arrest. The date of the arrest is indicated by the **ARREST_DATE**, while the offense description and penal code are specified by the **PD_DESC** and **PD_CD**, respectively. **KY_CD** and **OFNS_DESC** provide additional classification and explanation of the transgression. Legal codes and the associated legal category of the offense (such as felony) are provided by **LAW_CODE** and **LAW_CAT_CD**. **ARREST_PRECINCT** (precinct) and **ARREST_BORO** (borough) are two specifics regarding the arrest location. The jurisdiction that is involved in the arrest is indicated by **JURISDICTION_CODE**. The offender's **PERP_SEX**, **PERP_RACE**, and **AGE_GROUP** demographic data are included. The terms **Latitude**, **Longitude**, **X_COORD_CD**, and **Y_COORD_CD** are used to denote geographic coordinates.

Furthermore, for every arrest record, a New Georeferenced Column displays the georeferenced point that combines latitude and longitude. This dataset is crucial for helping with crime analysis and law enforcement activities by illuminating trends and demographics related to arrests.

COLUMN	DESCRIPTION
ARREST_KEY	A unique identifier for each arrest record.
ARREST_DATE	The date when the arrest occurred.
PD_CD:	The penal code associated with the offense.
PD_DESC	Description of the offense corresponding to the penal code.
KY_CD:	The internal classification code for the offense.
OFNS_DESC	Description of the offense category.
LAW_CODE:	The legal code associated with the offense.
LAW_CAT_CD	The legal category of the offense (e.g., felony).
ARREST_BORO	The borough where the arrest occurred.
ARREST_PRECINCT	The precinct where the arrest occurred.
JURISDICTION_CODE	The jurisdiction code related to the arrest.
AGE_GROUP:	The age group of the perpetrator.
PERP_SEX:	The gender of the perpetrator.
PERP_RACE:	The race of the perpetrator.
X_COORD_CD:	The X-coordinate (geospatial coordinate) associated with the arrest location.
Y_COORD_CD:	The Y-coordinate (geospatial coordinate) associated with the arrest location.
LATITUDE	Latitude of the arrest location
LONGITUDE	Longitude of the arrest location.
NEW GEOREFERENCED COLUMN	A georeferenced point that shows where the arrest happened by combining longitude and latitude.

IX. PROPOSED APPROACH:

1. **Data Preparation, Cleaning and Preprocessing:** Data preparation and cleaning are essential to ensure the quality and integrity of the dataset. We will address missing values, outliers, and inconsistencies to create a reliable dataset for analysis.
2. **Exploratory Data Analysis (EDA) and Statistical Analysis:** To find patterns, trends, and linkages in the data, EDA and statistical analysis will be used. To acquire preliminary insights, we will use descriptive statistics and visualize the data.
3. **Clustering and Identifying Influencing Factors:** To discover groups or trends in the data, we will use clustering algorithms. This step seeks to identify connections or influencing elements that lead to crime trends.
4. **Data Visualization:** Extensive data visualization will be used to illustrate and make accessible the findings. Visualization will be critical in identifying high- and low-crime regions, as well as the SDOH characteristics connected with them**.** Visualizations will be used to analyze and pick the most and least observed locations for further investigation.
5. **Communicating results:** Finally, the data analysis insights and results are clearly and concisely presented.

We propose to use a combination of exploratory data analysis and statistical analysis techniques. Explorative data analysis helps in understanding the data and identifying potential trends and patterns. Following that, we will use statistical analysis to validate these patterns and gain deeper insights. We plan to experiment with various methods and then choose the one that provides the most useful insights.

X. PROPOSED METHOD FOR EVALUATION:

The insights derived from the analysis will be evaluated based on their relevance to the problem statement and their potential impact on policing strategies. We will also use a portion of the dataset as a hold-out test set to evaluate the robustness of our insights on unseen data. The visualizations and outputs will be achieved using various techniques, analytics methods, and tools.

XI. PRELIMINARY ANALYSIS:

The NYPD Arrest data analysis can make us understand the crime trends and take proactive measures to determine and prevent criminal activity, which will eventually improve public safety. The following methodology has been employed for the analysis of the NYPD Arrest Data:

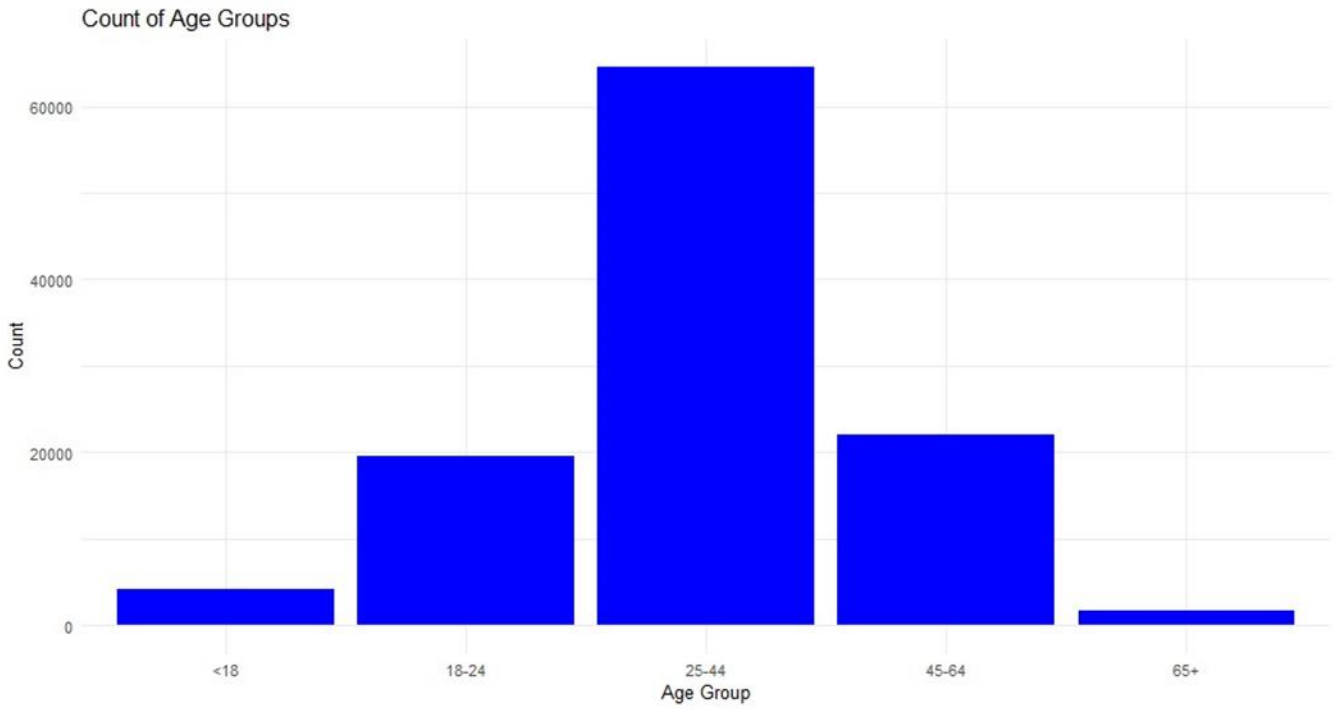
1) Data Collection and Cleaning:

- The NYPD Arrest Data (Year-to-Date) dataset was successfully collected and subsequently put through a thorough examination process, that involved a detailed analysis of the entire dataset. During the course of this analysis, several instances of missing and null values were identified within the dataset.
- Python was used to carry out the data cleaning process to keep the data's dependability and integrity. The impacted rows were eliminated where there were only a small number of missing values, and they did not significantly affect the analysis.

2) Exploratory Data Analysis and Visualization:

i. Analyses of Crime by Age Group:

We created a thorough and insightful visualization by utilizing R's ggplot2 function. The plot successfully depicted the predominant age groups involved in a majority of criminal activities, providing a clear visual representation of the age groups associated with higher crime rates.

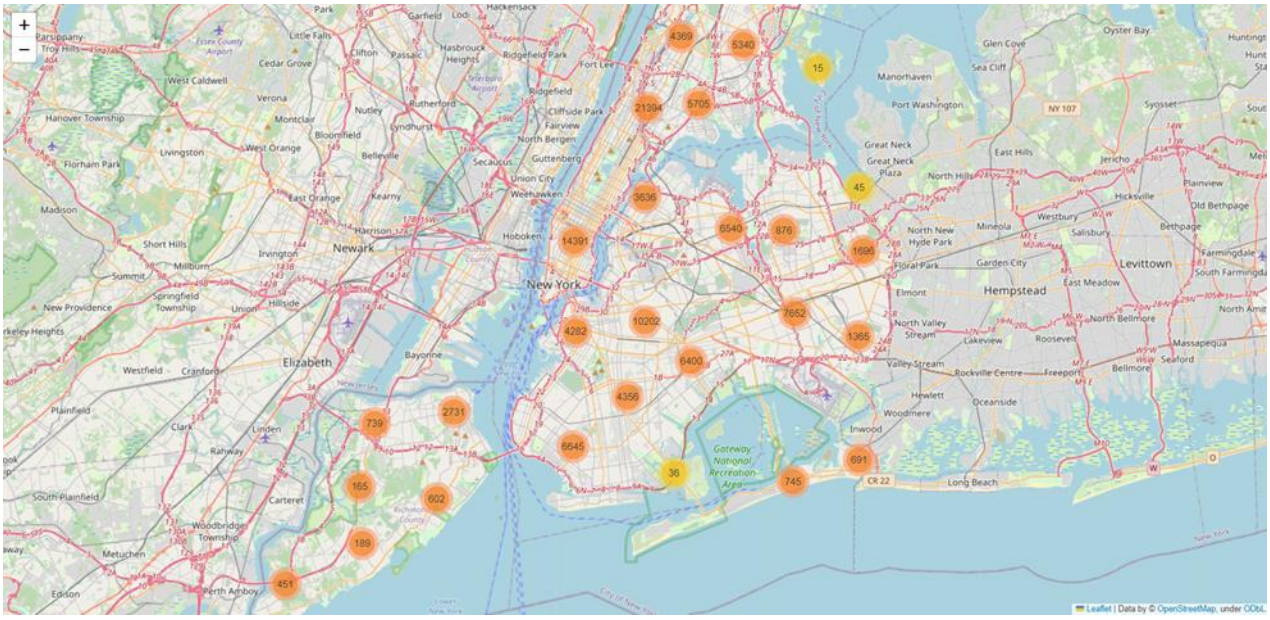


Insights:

- A much higher crime incidence among a particular age group, such as teenagers or young adults, may indicate the need for youth-focused intervention programs, mentorship programs, and educational support.
- If there are any differences in crime rates between age groups, they should be made public to provide insight on the social, economic, and/or cultural variables that influence criminal conduct in different populations.
- In order to reduce recidivism rates across various age groups, it is essential to highlight the implementation of age-specific rehabilitation and reintegration programs.
- Lack of access to school and employment possibilities may be associated with high crime rates in particular age groups. Taking care of these problems can aid in lowering crime.

ii. Geospatial Visualization:

- As a noteworthy aspect of our research, we used the longitude and latitude coordinates found in the dataset to visualize the geographic information. We developed visualizations to identify regions on a map with higher crime rates using geospatial frameworks in Python. To create a map display of the dataset's longitude and latitude information, we used the Python module Folium. With the help of this visualization tool, we were able to pinpoint geographical areas with higher crime rates, giving our project's goals a valuable perspective.
- Our project's deeper research and more insights into the mechanisms behind these patterns will be built on this basic analysis and the mapping of crime distribution.
- Below is an example of the map view to get a detailed understanding about the map and the details of the crimes. Please look into the pdf document attached in the assignment for navigating through the map and better view.



Insights:

- The map identifies regions with larger concentrations of criminal activity, which can help law enforcement better use their resources to combat crime in those areas.
- Local governments and politicians can better customize crime prevention initiatives to match the problems faced in each area by keeping track of how crimes are distributed around the boroughs.
- There may be a higher requirement for security and enforcement at and around transportation hubs like train stations and bus terminals because crime tends to cluster there.

XII. PROJECT TIME PLAN:

The project is expected to be completed in 10 to 12 weeks, with the following timeline:

PROJECT TIMELINE	*TASK DESCRIPTION*
WEEK 1-2	Project Proposal and Data Collection
WEEK 3-4	Data Cleaning, Pre-processing and Exploratory data analysis
WEEK 5-6	Descriptive analysis, Insight Generation and Data visualization
WEEK 7-8	Interpretation of Statistical Analysis and Findings
WEEK 9-10	Evaluation, Refinement of Insights and Finalization
WEEK 11-12	Report Writing and Project presentation

XIII. References

[1] ., "NYPD Arrest Data (Year to Date)," 14 July 2023. [Online]. Available: https://catalog.data.gov/dataset/nypd-arrest-data-year-to-date. [Accessed 20 September 2023].

[2] L. E. P. V. a. N. R. V. Mandalapu, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10151873. [Accessed 20 September 2023].

[3] Clemens, ""“Minority Report” a Reality? The NYPD’s Big Data Approach to Predicting Crime," [Online]. Available: https://d3.harvard.edu/platform-rctom/submission/minority-report-a-reality-the-nypds-big-data-approach-to-predicting-crime/. [Accessed 20 September 2023].

[4] J. F. a. A. G. J. MacDonald, "The Effects of Local Police Surges on Crime and Arrests in New York City," [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157223. [Accessed 20 September 2023].

[5] L. T. a. S. U. S. Chainey, "‘The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime’, Secur J, vol. 21, no. 1, pp. 4–28,," Feb 2008. [Online]. Available: https://doi.org/10.1057/palgrave.sj.8350066. [Accessed 20 September 2023].