A picture containing text, sign, alcohol

Description automatically generated

ALY 6110 – Data Management & Big Data

Module 6 Final Project  
**Final Project Report**

Devansh Hukmani

Yuning Chen

Instructor – Daya Rudhramoorthi

28 October 2023

1. **Introduction**

In the ever-evolving landscape of law enforcement, the analysis of arrest data plays a pivotal role in understanding the dynamics of policing within a major metropolis like New York City. This report delves into the extensive NYPD Arrest Data for the current year, a dataset comprising over 113,000 rows and 19 columns. It represents a comprehensive breakdown of every arrest conducted by the New York City Police Department (NYPD) during this year.

This data, manually extracted and periodically reviewed by the Office of Management Analysis and Planning, provides invaluable insights into the nature of law enforcement activities in the city. Each record within the dataset encapsulates essential information, encompassing the type of crime, the geographical location of enforcement, and the precise timing of each arrest. Furthermore, it offers a unique perspective on the demographics of the individuals involved in these arrests.

The purpose of this report is to leverage this rich dataset to explore and analyze the patterns and trends within NYPD's arrest activities. By examining various aspects such as crime types, locations, and suspect demographics, we aim to uncover significant insights that can contribute to a better understanding of policing in New York City. This data is not only of interest to law enforcement agencies but also serves as a valuable resource for policymakers, researchers, and the public seeking transparency and accountability in the realm of law enforcement.

* Data source: <https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc>

1. **Business Question**

**How do the distribution of arrests across New York City boroughs correlate with the types of crimes committed in different age groups, and can we identify patterns and variations that can inform regulated suggestions for the NYPD?**

In this report, we will explore the interplay between the geographic distribution of arrests across New York City's boroughs and the types of crimes committed within distinct age groups. By analyzing the NYPD Arrest Data for the current year, we aim to uncover correlations and patterns that shed light on whether certain boroughs exhibit unique crime profiles associated with specific age cohorts. Our ultimate objective is to provide evidence-based, regulated suggestions to the NYPD, considering the multifaceted relationship between crime types, age demographics, and borough-specific dynamics.

1. **Variables in the Dataset**

* **ARREST\_KEY**: Randomly generated persistent IDs for each arrest.
* **ARREST\_DATE**: Exact date of arrest for the reported event.
* **PD\_CD**: Three-digit internal classification code (more granular than Key Code).
* **PD\_DESC**: Description of internal classification corresponding with PD code (more granular than Offense Description).
* **KY\_CD**: Three-digit internal classification code (more general category than PD code).
* **OFNS\_DESC**: Description of internal classification corresponding with KY code (more general category than PD description).
* **LAW\_CODE**: Law code charges corresponding to the NYS Penal Law, VTL, and other various local laws.
* **LAW\_CAT\_CD**: Level of offense: felony, misdemeanor, violation.
* **ARREST\_BORO**: Borough of arrest (B for Bronx, S for Staten Island, K for Brooklyn, M for Manhattan, Q for Queens).
* **ARREST\_PRECINCT**: Precinct where the arrest occurred.
* **JURISDICTION\_CODE**: Jurisdiction responsible for arrest (0 for Patrol, 1 for Transit, 2 for Housing representing NYPD; 3 and higher for non-NYPD jurisdictions).
* **AGE\_GROUP**: Perpetrator’s age within a category.
* **PERP\_SEX**: Perpetrator’s sex description.
* **PERP\_RACE**: Perpetrator’s race description.
* **X\_COORD\_CD**: Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, measured in feet (FIPS 3104).
* **Y\_COORD\_CD**: Midblock Y-coordinate using the same coordinate system as X\_COORD\_CD.
* **Latitude**: Latitude coordinate for Global Coordinate System, WGS 1984, in decimal degrees (EPSG 4326).
* **Longitude**: Longitude coordinate using the same coordinate system as Latitude.
* **New Georeferenced Column**: A new column likely combining geographical information, forming a point for geospatial analysis.

1. **Tableau Analysis**

Creating visualizations to understand the distribution of arrests across New York City boroughs in correlation with types of crimes committed in different age groups is a valuable analysis that can potentially inform regulated suggestions for the NYPD.

The graph Arrest Count by Race illustrates the distribution of arrests by race in New York City, revealing notable disparities in law enforcement interactions across different racial categories. The highest number of arrests is among Black individuals, totaling around 54,000, followed by White Hispanic individuals with approximately 28,500 arrests. White individuals account for around 11,000 arrests, while Black Hispanic individuals are close behind with roughly 10,500 arrests. Asian/Pacific Islander individuals have the fewest arrests, with approximately 6,000. While this visualization highlights these differences, it doesn't provide the underlying reasons, which may include various socioeconomic and demographic factors. Nevertheless, it underscores the importance of examining and addressing potential racial disparities within the criminal justice system to ensure fairness and equity.

The graph Arrest by Age Group for Every Borough is depicting arrests by age group for every New York City borough. The age group with the highest number of arrests in every borough is individuals aged 25-44, highlighting a demographic segment frequently engaged in law enforcement interactions. Following closely behind are those aged 45-64, reflecting a substantial portion of arrests. The 18-24 age group, representing young adults, also shows a notable arrest rate, potentially due to factors like urban living and socioeconomic conditions. Conversely, individuals under 18, often benefiting from juvenile justice approaches, have fewer arrests. The 65+ age group consistently exhibits the lowest arrest count, underscoring minimal involvement in criminal activities among the elderly. This trend emphasizes the importance of understanding age-related arrest patterns to inform law enforcement strategies, social policies, and community initiatives aimed at reducing crime and enhancing public safety, particularly among the age groups most frequently implicated in arrests.

A screenshot of a data report

Description automatically generated

Dashboard 1

The visualization of Offenses Caused by Age Groups presents a clear pattern in New York City's crime data. Assault-related offenses stand out as the most frequently committed crimes, reflecting a significant concern for interpersonal violence. What's particularly striking is the consistent trend of individuals aged 25-44 having the highest offense count, especially in assault-related incidents. This demographic appears to be at the center of a significant portion of criminal activity.

The graph depicting the Number of Arrests Every Month, marked by recurring fluctuations, signifies the dynamic nature of law enforcement activities in New York City. These fluctuations can be attributed to various factors, including seasonal variations in criminal behavior, adjustments in policing strategies, the influence of special events and holidays, socioeconomic conditions, and potential data-related anomalies. This ever-changing pattern of arrests emphasizes the need for a responsive and adaptive approach to law enforcement and crime prevention. Understanding the driving forces behind these fluctuations is essential for law enforcement agencies and policymakers, enabling them to allocate resources effectively, time community engagement initiatives, and develop strategies that align with the evolving landscape of criminal activity and public safety concerns in the city.

A screenshot of a computer screen

Description automatically generated

Dashboard 2

The graph illustrating Offense by Area with longitude on the y-axis and latitude on the x-axis offers a geospatial perspective of arrest data in New York City. Each point on the graph corresponds to a specific location within the city, with information about the borough, and the type of crime committed. This geospatial representation enables a clear understanding of where different types of crimes occur, pinpointing crime hotspots and trends in specific areas across the city. Such visualizations are essential tools for law enforcement and policymakers as they provide actionable insights for resource allocation, targeted policing strategies, and crime prevention efforts tailored to the needs of specific geographic areas within New York City.

Looking at the second bar chart of regional criminal records, we can see that Brooklyn and the Bronx have consistently high criminal arrest records.

The Top 5 Arrested Offenses graph is a descriptive statistical analysis of the types of arrested offenses reveals the types of crimes that are most likely to win conflicts in New York City. Summarizing above three analysis of crime concentration areas, crime concentration age groups and top arrested offenses, we can draw the following insights and recommend that NYPD direct the development of targeted interventions and crime prevention strategies to reduce violence and enhance public safety, particularly in areas with high rates of assault-related crime.

1. **Python Analysis**

Python packages play pivotal roles in data analysis and visualization. NumPy is fundamental for numerical operations and efficient array manipulation, while Matplotlib offers a versatile platform for creating various types of visualizations. Seaborn builds upon Matplotlib to provide aesthetically pleasing statistical graphics, simplifying the process of data exploration. Pandas is indispensable for data manipulation, offering data structures and functions for data cleaning, selection, and transformation. Additionally, the %matplotlib inline command, when used in Jupyter Notebook, ensures that Matplotlib plots are displayed directly within the notebook, facilitating interactive data analysis. These packages collectively empower data analysts and scientists to efficiently process, explore, and represent data for insights and decision-making.

Before proceeding with Data Pre-processing, we will make comprehensive use of these charts and graphs to gain a deeper understanding of the dataset and provide valuable insights for further analysis and modeling. Furthermore, through visual analysis, we can better convey the meaning and discoveries of the data to our classmates. Visualizations, powered by these Python packages, serve as a bridge between raw data and meaningful insights. They allow us to uncover patterns, trends, and anomalies, making the data more accessible and understandable. This visual exploration sets the stage for robust data preprocessing and informs the subsequent steps in our analysis, ultimately enhancing the quality and depth of our findings.

* **The relationship between borough and age of suspects**

By creating the graph of Arrest Distribution by Borough and age Group, let’s look at concentration trends within the data concerning age and region of criminal suspects. Highest crimes happened in Kings (18071 records) followed by Bronx (15264 records), and criminal suspects age are concentrated in 25-to 44.

A screenshot of a computer

Description automatically generated

* **Crime distribution by borough and age group**

By creating the graph of Crime Distribution by Borough and Age Group. We can specifically derive the highest and lowest age groups for crime in each Borough to assist the NYPD in providing the appropriate prevention techniques for individual Boroughs. And most of the crimes committed in Kings are Arson, in Manhattan are Dangerous Drugs

A screenshot of a computer

Description automatically generated

* **The relationship between Crimes and sex of suspects**

By analyzing the relationship between arrests by law category and the gender of the offender, we can see that the crime rate of men is about 10 percent higher than that of women.

A screenshot of a graph

Description automatically generated

* **Trends in arrests over time**

We want to understand the trends in crime rates in New York arrest rates in 2023. 2023 is the year that New York City is fully open after COVID-19 and people don't need to wear masks and observe social distances, but due to the resurgence of the epidemic in the month of February we can see that New York's crime rate is at a record low. After February, people's lives returned to normal again and their crime rates began to rise rapidly. Secondly, we find that when the last day of each month is the day with the highest crime rate for that month, so we guess a reasonable assumption is that the NYPD sets the date of the crime on the last day of the month when the exact date of the crime is uncertain.

A graph with a line

Description automatically generated

* **Correlation between variables**

A heatmap is a powerful visualization tool that represents data in a matrix format, where each cell's color indicates the strength and direction of the correlation between two variables. This visual representation allows you to quickly identify patterns and relationships in the data.

A screenshot of a computer

Description automatically generated

1. **Data Cleaning**

The first step in data preprocessing was to clean a dataset containing the records. We start by identifying the missing values in the dataset. Then the missing data that do not meet the criteria are removed manually by using mode as a criterion.

A screenshot of a computer

Description automatically generated

The mode represents the most frequently occurring value in a dataset and, when used to fill missing data, it ensures that the overall distribution and consistency of the data are preserved. This approach maintains data transparency, simplicity, and applicability to different scales while avoiding significant alterations to the data's characteristics.

A screenshot of a computer code

Description automatically generated

After the removal, check the missing value and white-space cells again until they are all zero.

A screenshot of a computer code

Description automatically generated

Next, I suspected there are missing values in it, represented by the question mark symbol '?'. After the columns have been checked, it then shows the names of those columns one by one. We can see that there were none found.

A screenshot of a computer code

Description automatically generated

1. **Logistic Regression Model**

It begins with the import of necessary packages, including scikit-learn's modules for model selection, preprocessing, and evaluation. The dataset is prepared by selecting the target variable 'LAW\_CAT\_CD' and relevant features, 'ARREST\_BORO' and 'AGE\_GROUP.' Categorical features are encoded using LabelEncoder, ensuring that the model can handle them. The dataset is then split into training and testing sets, with 80% used for training and 20% for testing. Feature standardization is applied to ensure that all variables are on a consistent scale. A logistic regression model is chosen for classification, and it is trained on the training data. Predictions are made on the test set, and the model's performance is evaluated.

A screenshot of a computer program

Description automatically generated

The results of the logistic regression model's performance evaluation provide a detailed overview of its ability to classify law categories based on the selected features 'ARREST\_BORO' and 'AGE\_GROUP.' The accuracy of approximately 57.78% indicates that the model correctly predicts the law category for a considerable portion of the test dataset. However, a closer examination of the confusion matrix reveals variations in performance across different law categories. For instance, the model excels in identifying misdemeanor cases (class 'M') with a high recall of 0.99, suggesting it captures most of these cases accurately. In contrast, the model struggles to identify infractions (class 'I') and violations (class 'V') with no true positives in the former and limited accuracy in the latter. Precision values vary, with the model showing the highest precision for felonies (class 'F'). These results offer valuable insights for understanding the model's strengths and areas requiring improvement, making it clear that further refinements are needed to enhance its performance, particularly in recognizing infractions and violations.

A screenshot of a computer

Description automatically generated

1. **Random Forest Classifier Model**

This model is employed to predict the 'LAW\_CAT\_CD' (law category) based on the selected features 'ARREST\_BORO' and 'AGE\_GROUP' from the dataset. The dataset is divided into training and testing sets, with 80% used for training and 20% for testing. The Random Forest Classifier is configured with 100 decision trees (n\_estimators=100) and a fixed random state for reproducibility. The model is then trained on the training data, allowing the decision trees to collectively make predictions on the test set. After making predictions, the model's performance is assessed using various metrics, including accuracy, which measures the proportion of correctly predicted law categories. The confusion matrix provides a detailed breakdown of true positives, true negatives, false positives, and false negatives for each law category, offering insights into the model's classification accuracy. The classification report further details precision, recall, F1-score, and support for each category, providing a comprehensive evaluation of the model's performance in law category classification.

A screenshot of a computer program

Description automatically generated

The accuracy of approximately 58.37% indicates that the model correctly predicts the law category for a substantial portion of the test dataset. Delving deeper into the confusion matrix, we observe variations in the model's performance across different law categories. The model excels in identifying misdemeanor cases (class 'M') with a high recall of 0.98, signifying its capability to accurately capture most of these cases. Conversely, the model struggles to identify infractions (class 'I') and violations (class 'V) with no true positives in either category. The precision values vary, with the model showing the highest precision for felonies (class 'F'). These results provide valuable insights for understanding the model's strengths and areas requiring improvement. Like the previous logistic regression model, it is evident that further refinements are needed to enhance its performance, particularly in recognizing infractions and violations.

A screenshot of a computer

Description automatically generated

1. **Model Comparison**

The Logistic Regression and Random Forest Classifier models were evaluated for their ability to classify law categories based on the selected features 'ARREST\_BORO' and 'AGE\_GROUP.'

In terms of accuracy, the Random Forest Classifier model outperforms the Logistic Regression model with an accuracy of approximately 58.37%, while the Logistic Regression model achieved an accuracy of approximately 57.78%. This means that the Random Forest model correctly predicted the law category for a larger portion of the test dataset.

Both models display variations in performance across different law categories. The Logistic Regression model excels in identifying misdemeanor cases (class 'M') with a high recall of 0.99, indicating that it accurately captures most of these cases. However, it struggles to identify infractions (class 'I') and violations (class 'V), with no true positives in the former and limited accuracy in the latter. On the other hand, the Random Forest Classifier also struggles to identify infractions and violations, with no true positives in either category.

For precision, the Logistic Regression model shows the highest precision for felonies (class 'F'), while the Random Forest Classifier exhibits a good balance in precision across law categories.

In summary, while both models have strengths and weaknesses, the Random Forest Classifier model is slightly better in terms of overall accuracy. However, both models need further refinements, particularly in recognizing infractions and violations.

**10. Results and Conclusion**

The analysis of NYPD Arrest Data for the current year offers profound insights into the dynamics of law enforcement in New York City. It underscores the significance of age demographics, with the age group of 25 to 44 consistently emerging as the most engaged in law enforcement interactions. This knowledge can guide targeted crime prevention strategies and community initiatives, acknowledging the specific needs of different age groups. Gender disparities in arrest rates point to the importance of tailoring interventions based on gender-specific factors. The temporal trends in arrests unveil the impact of the COVID-19 pandemic and its subsequent recovery on law enforcement activities, underlining the importance of adaptability in policing. Geospatial analysis identifies crime hotspots, informing resource allocation and focused policing efforts. Moreover, the prevalence of assault-related offenses among individuals aged 25 to 44 highlights the need for strategies to address interpersonal violence and enhance public safety. While predictive models exhibit promise, further refinements are required to enhance their accuracy in classifying different law categories, especially infractions and violations. Overall, these findings are invaluable for policymakers, law enforcement agencies, and researchers, offering actionable insights for addressing crime trends and improving public safety in New York City.

**11. References**

*NYPD Arrest Data (Year to Date) | NYC Open Data*. (2023, July 14). <https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc>

Li, S. (2019, February 27). *Building A Logistic Regression in Python, Step by Step*. Medium. <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

Shafi, A. (2023, February 24). *Random Forest Classification with Scikit-Learn*. <https://www.datacamp.com/tutorial/random-forests-classifier-python>