Analyzing NYPD Arrest Data: Trends, Patterns, and Predictive Modeling for Public Safety

Data Processing Framework Final Project Report

By:

Hadi Knaiber

Rim Zeaiter

Nour Azakir

Lara Baltaji

American University of Beirut

April 2023

Professor: Mireille Makary, Ph.D.

Contents

Abstract

The New York City Police Department (NYPD) arrests dataset contains a wealth of information on arrests made in the city. Throughout this study, the dataset was analyzed to uncover trends and patterns to understand the nature of violations, distribution of the arrests, and the demographics of the arrested individuals. With the help of Python, Apache Spark, Microsoft SQL Server, and Tableau, we were able to preprocess the dataset for analysis and visualize the data to spot trends and key findings such as the demographics of the arrests, distribution of the arrests among boroughs, the most frequent offenses and their level. Our findings show that some boroughs have a higher rate of arrests compared to others. Moreover, certain crimes, such as assaults and thefts, are more common than others. Also, we observed that the majority of the arrested perpetrators are more likely to be black adult males. One of the key highlights in our study is the development of a predictive model to identify unclassified arrests in terms of the level of offense that took place. By identifying patterns in criminal activity and defining the characteristics of the offenders, we can develop strategies and approaches to reduce crime and improve public safety. This can include sending extra police units to areas that witness high criminal activity, developing prevention programs for vulnerable groups, and offering educational and career opportunities to address the root causes of crime. Additionally, our analysis emphasizes the need for transparency in police activity and the necessity to ensure that all individuals are treated fairly within the criminal justice system.

1 Introduction

In recent years, there has been an increasing concern about the activities of law enforcement agencies, including the police. Due to this, the need for transparency in police enforcement activities has become more important. As a response, many law enforcement organizations have made their data available to the public. One such organization is the New York Police Department (NYPD), which provides information on every arrest made in New York City (NYC).

The NYPD's primary objective is to protect and serve the people of New York City. To achieve this, they keep a detailed record of every arrest made within their jurisdiction. This record includes information such as the date and time of the arrest, the type of offense, the relevant laws, and the demographics of the perpetrator.

This project focuses on analyzing the NYPD Arrest dataset, which contains information about the arrests made by the NYPD in 2022. The dataset includes various factors, such as the type of crime committed, the age and race of the perpetrator, and the location and time of the arrest. To handle this large dataset more effectively and efficiently, we have divided it into four tables: detailed_offense_table, law_table, offense_table, and arrest_table. By doing so, we can analyze the data more effectively and run queries on SQL and Apache Spark.

2 Literature Review

New York City has been a center for crime and law enforcement, providing researchers with ample data on arrests. Advances in technology and digital records have made this data more accessible. However, this also raises concerns about the fairness, accountability, and transparency of policing practices.

Arrest Insights about crime trends, policing techniques, and the criminal justice system can be gained from arrest statistics. A focus for arrest data studies in recent years has been New York City because of its sizable population, diversified communities, and high crime rates. There are several studies that were published related to our topic.

A study conducted by Braga in 2022 examined the implementation and impact of a pilot body-worn camera program on the New York City Police Department (NYPD). The study involved 3,889 NYPD officers and 40 police precincts, and found that equipping officers with body-worn cameras resulted in a 21% decrease in public complaints against treatment officers. However, the study also revealed that stop reports involving minorities were more likely to be evaluated as non-compliant with constitutional reasons for stops, frisks, and searches. Nevertheless, the findings suggest that body-worn cameras may increase officer compliance with documentation requirements and reduce illegal policing by facilitating the identification of problematic police-citizen interactions.

A study by Levine (2017) examined the Domain Awareness System (DAS) used by the NYPD. The system includes various technologies, such as pattern recognition, machine learning, and data visualization, and is designed to help law enforcement officers make better decisions by providing specialized data and analytics. Since 2008, the NYPD has been using DAS to maximize its use and has sold the technology to other organizations, generating revenue for the city. The study found that DAS has saved the NYPD an estimated $50 million annually and has been useful in reducing terrorism and increasing law enforcement efficiency.

Regarding police-related predictive models, a study was conducted (Catlett,2019) that emphasized on cities that are undergoing tremendous economic and social transformation as a result of urbanization, creating a variety of management and service delivery issues. Ensuring public safety in cities with high crime rates is getting more difficult. In order to deal with this complexity, new technologies are giving police agencies access to increasing volumes of data on crimes that can be examined to spot patterns and trends. These technologies may increase the effective use of police resources in a specific area and ultimately promote crime prevention. This article proposes a predictive method for automatically locating high-risk crime locations in urban areas and accurately predicting crime trends in each area. The method makes use of spatial analysis and auto-regressive models. The algorithm's output is a spatio-temporal crime forecasting model made up of a number of crime-prone areas with corresponding crime predictors, each of which represents a model for predicting the number of crimes that are most likely to occur in the area to which it is connected. Two real-world datasets from the cities of

Chicago and New York City were used to evaluate the technique, which demonstrated that it can predict crime with reasonable accuracy over rolling time horizons in both space and time.

These studies are relevant to our research because they emphasize how crucial it is to comprehend the fundamental causes of arrest risk. Additionally, they highlight the significance of taking into account both individual-level variables, such as the cause of the arrest, and societal and environmental factors, when analyzing arrest data. We may build more effective policies and initiatives to lessen crime and advance social justice by taking these broader contextual elements into account. By doing so, we can acquire a more nuanced understanding of the patterns and trends in arrest statistics.

Our investigation on the arrest statistics in New York City relies heavily on the literature review. It enables us to review previous studies and research on related subjects, which can provide us with knowledge of the patterns, trends, and influences affecting arrests in the city. We can find gaps in the existing literature and choose the course of our research by examining and synthesizing this data. The literature review also aids in the development of our study's theoretical framework, which directs the analysis and interpretation of our results.

3 Experimental Design and Evaluation

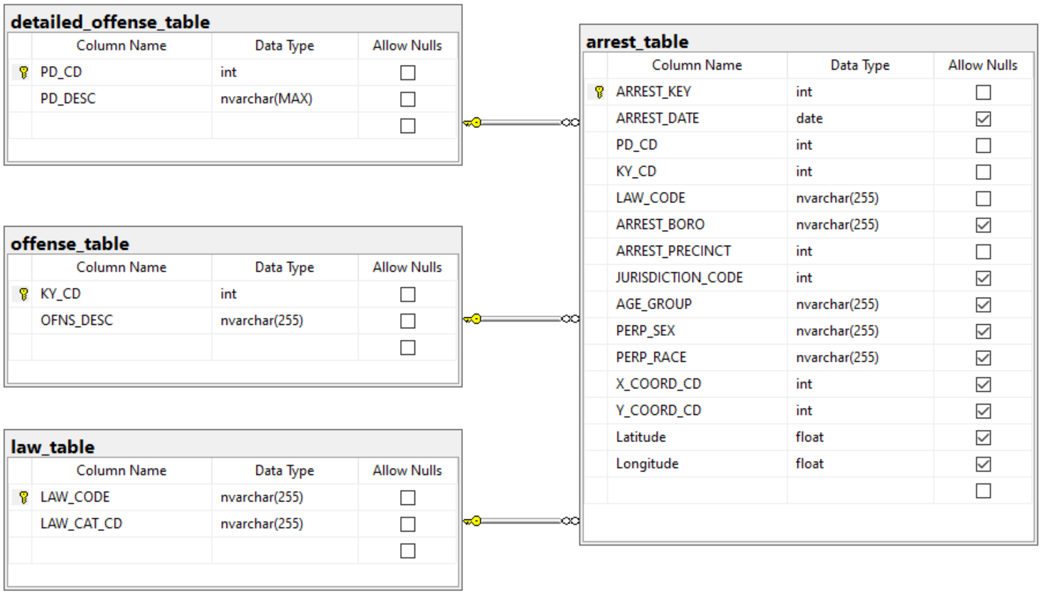
The methods utilized to preprocess, analyze, and visualize the arrest data are thoroughly discussed in the experimental design and evaluation section in order to produce significant insights and create a predictive model.

The process included several steps where we first divided the dataset into tables and then cleaned and transformed the data for analysis using Python. We verified the data types, content, shape, and null values. We removed the duplicates and the null values after locating them.

The dataset was divided into 4 tables in order to make them applicable for relational databases using SQL. The first table is called the “Offense Table” and it consists of the three-digit internal classification code for the general offense, labeled “KY_CD”, and the description of the offense, “OFNS_DESC”. Using this table, we can analyze the frequency and types of offenses committed in the area covered by the NYPD. The second table is called the “Law Table” and it consists of the LAW_CAT_CD, which indicates the level of offense—felony, misdemeanor, or violation, infraction—as well as the law code. Using this table, we can analyze the severity and the trends of the offenses committed that are more likely to be considered. The third table is called the “Detailed Offense Table” and it consists of the three-digit internal classification number for the detailed offense, “PD_CD”, and the description of the detailed offense, “PD_DESC”. This table provides a more detailed analysis of the crimes committed and can be used to determine how common particular offenses are in the area covered by the NYPD dataset. The “Arrest Table” is the fourth table and it also includes details about each arrest made by the NYPD, such as the unique key for the arrest, the arrest date , the keys of both the offense and the detailed offense, the law code, the arrest borough (neighborhood), the arrest precinct, the jurisdiction code, the perpetrator’s age group, sex and race, as well as the x and y coordinates based on the New York State Plane Coordinate System and latitude and longitude of the arrest site. The most comprehensive information on the arrests is provided in this table, and this can be used to examine trends in the arrests made by the NYPD, such as which precincts or areas have greater arrest rates, which age groups, or which races, are more likely to be arrested.

We examined the “PD_CD” compatibility between the offense table and the arrest table in order to guarantee dataset compatibility for relational databases using a function that returns values that are incompatible. The values were compatible since we applied the function (if statement) that returned no value. The function once more returned no values when we tested the KY_CD's compatibility between the arrest table and the offense table. The arrest table and the law table were both tested for consistency concerning the law code and compatibility of values was found. The four tables were then saved as CSV files.

Afterwards, we created the tables on Microsoft SQL server. The primary keys were the KY_CD, PD_CD, ARREST_KEY, LAW_CODE for the offense table, detailed offense table, arrest table and the law table respectively. The foreign keys were the PD_CD, the KY_CD and the LAW_CODE in the arrest table. We then connected the tables using primary key-foreign key relations on SQL, then adjusted the data types in the tables to be compatible with its respective feature in another table. Following that, we inserted the data into the tables and applied queries to gain insights. The queries are important in order to learn more about the arrests made by the NYPD, including the arrests committed, the characteristics of the criminals, the places and times of the arrests. Below is a screenshot of the database’s diagram created on Microsoft SQL Server.



detailed_offense_table

	Column Name	Data Type	Allow Nulls
🔑	PD_CD	int	☐
	PD_DESC	nvarchar(MAX)	☐
			☐

offense_table

	Column Name	Data Type	Allow Nulls
🔑	KY_CD	int	☐
	OFNS_DESC	nvarchar(255)	☐
			☐

law_table

	Column Name	Data Type	Allow Nulls
🔑	LAW_CODE	nvarchar(255)	☐
	LAW_CAT_CD	nvarchar(255)	☐
			☐

arrest_table

	Column Name	Data Type	Allow Nulls
🔑	ARREST_KEY	int	☐
	ARREST_DATE	date	☑
	PD_CD	int	☐
	KY_CD	int	☐
	LAW_CODE	nvarchar(255)	☐
	ARREST_BORO	nvarchar(255)	☑
	ARREST_PRECINCT	int	☐
	JURISDICTION_CODE	int	☑
	AGE_GROUP	nvarchar(255)	☑
	PERP_SEX	nvarchar(255)	☑
	PERP_RACE	nvarchar(255)	☑
	X_COORD_CD	int	☑
	Y_COORD_CD	int	☑
	Latitude	float	☑
	Longitude	float	☑
			☐

Figure 1: NYPD Database Diagram

Spark and Tableau were used to visualize the data and create visual dashboards. We used PySpark to carry out deeper data analysis, which included combining tables, querying data, and building visualizations in order to obtain insights into the patterns and trends of arrests in New York City. This is relevant since we were able to extract useful information from the dataset. The visualizations helped present the data in a more digestible and informative format and this allowed us to draw out meaningful conclusions. Below are the links for the interactive tableau dashboards that we created.

- NYPD Arrest Count Over Time: https://public.tableau.com/views/NYPDArrestCountOverTime/Dashboard1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link
- NYPD Arrest Locations: https://public.tableau.com/views/NYPDArrestLocation/Dashboard2?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link
- NYPD Demographics and Jurisdiction Information: https://public.tableau.com/views/NYPDDemographicsandJurisdiction/Dashboard3?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link
- NYPD Offense Description:

https://public.tableau.com/views/NYPDOffenseDescription/Dashboard4?:language=en-US&:display_count=n&:origin=viz_share_link

In order to address the issue of unclassified arrests in terms of level of offense, we developed a predictive model that can precisely predict the level of offense in cases involving unclassified arrests. We started by extracting some important date features from our dataset such as the arrest month, day of the month and day of the week. Then we dropped the features that we believed are insignificant in terms of predictive power including the original date feature, the geolocation-related features (longitude, latitude, X coordinate, Y coordinate), in addition to unique key features (LAW_CODE, PD_CD, KY_CD, ARREST_KEY) and the detailed description of the offense PD_CD (to avoid redundancy of information). Next, we removed all the observations that included unknown levels of offense. Our next step is to split the resulting data in training and testing subsets with 80% for training and 20% for validation and we specified our target variable which is the level of offense (LAW_CAT_CD). In the feature engineering step, we used StringIndexer which is a feature transformer in PySpark that maps a string column of labels to an ML column of label indices which changes all categorical variables into numerical types. We then scaled the numerical features using MinMaxScaler. Finally, we defined our model, Random Forest Classifier. All the above steps were staged as steps in a pipeline. The pipeline was fitted on the training data and transformed on the test data.

This model is noteworthy because unclassified arrests might make it difficult to identify the patterns and trends in NYC arrest data. We can better comprehend the criminal justice system in New York City if we correctly categorize these offenses. The predictive model can also improve the efficiency of the NYPD's work by automatically classifying offenses, by doing this, much time and resources that would otherwise be used to manually classify infractions can be saved. Additionally, the predictive model might help in knowing the sentencing duration of the perpetuator. By estimating the level of offense, we are facilitating the estimation of the sentence that someone would receive based on their demographic data and the specifics of their arrest. We can also eliminate bias in sentencing and detect unjust decisions in the criminal justice system. This could make it easier to spot trends in the types of crimes committed and provide information for measures to lower crime rates in particular places.

4 Results

After conducting our analysis, we were able to draw out meaningful information and useful insights from the dataset. The results provide a thorough look at the patterns and trends of the arrest data in New York City. This can aid in understanding the city's nature of crime and work on initiatives to improve public safety.

We utilized SQL queries to gain significant findings and insights. First, we aimed to view the total number of arrests made by each jurisdiction, in this case, the NYPD. This gave us information about the competence with which the NYPD makes arrests. The results showed that the highest number of arrests was during police patrol (arrests made on regular police duties) with 171,972 arrests followed by 6,704 arrests in transit (arrests made by transit bureau, responsible for subway system and other transit facilities such as buses, train stations) and 6,473 in housing jurisdiction (arrests made by policing for housing authority properties). The number of patrols is the highest since NYPD sends regular patrols to investigate areas in NYC and this is where trouble is mostly witnessed by the police.

We looked at the precincts per borough to understand the distribution of arrests among them. The results showed the borough with the highest number of precincts is Brooklyn. This is because Brooklyn is the most popular borough in the city and has a high crime rate which requires a greater police presence to maintain public safety.

In addition, we looked at the borough with the most arrests and the most crimes committed in the top borough. The results were also Brooklyn and the most crimes committed were assaults with 8,035 arrests related to assaults.

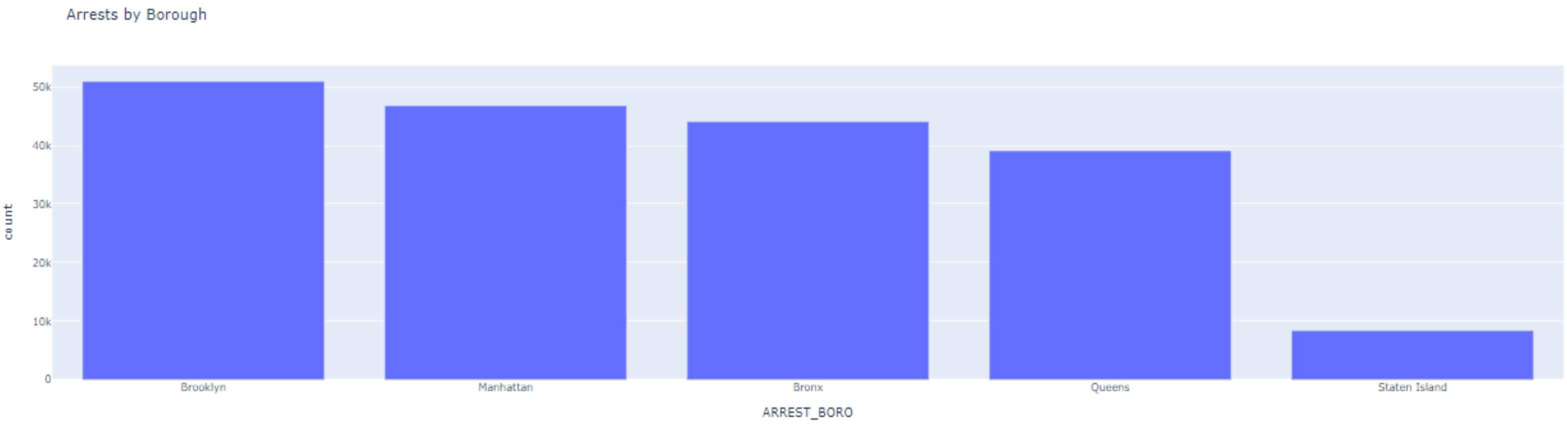


Figure 2: Total Number of Arrests by Borough

The reason for this could be due to several factors, for example, socioeconomic factors such as inequality and poverty, plus, the demographics of Brooklyn include a high number of young people and a variety of color.

The most frequent offense was then determined by examining the description of the offense which was Assault 3 and Related Offenses as seen in the figure below.

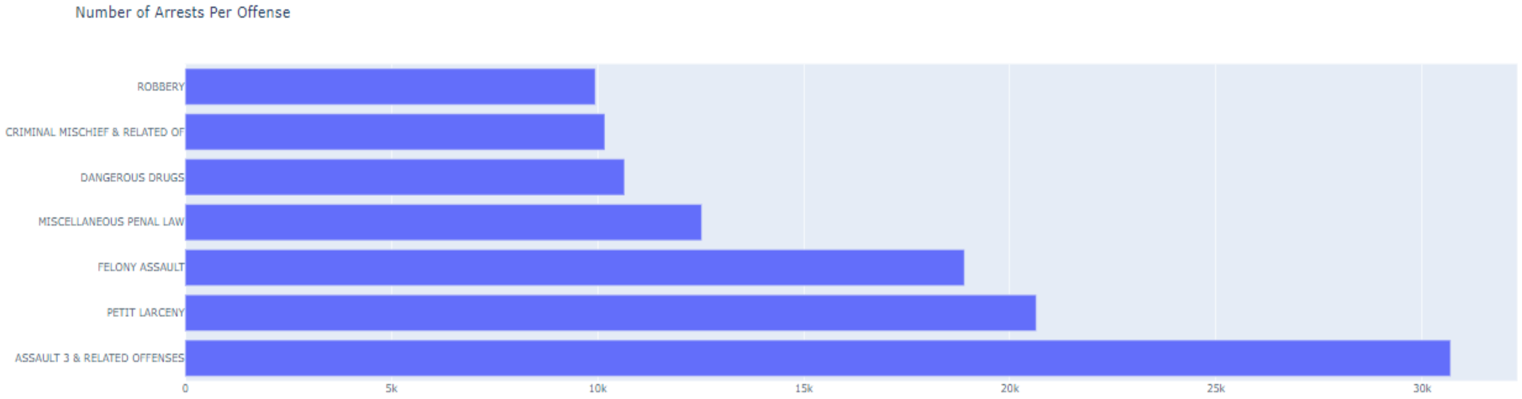


Figure 3: Number of Arrests per Offense

Similar to that, we looked at the detailed offense with the top offense that happens most frequently and the results also pointed to assault-related arrests with 30,687 arrests linked to assaults.

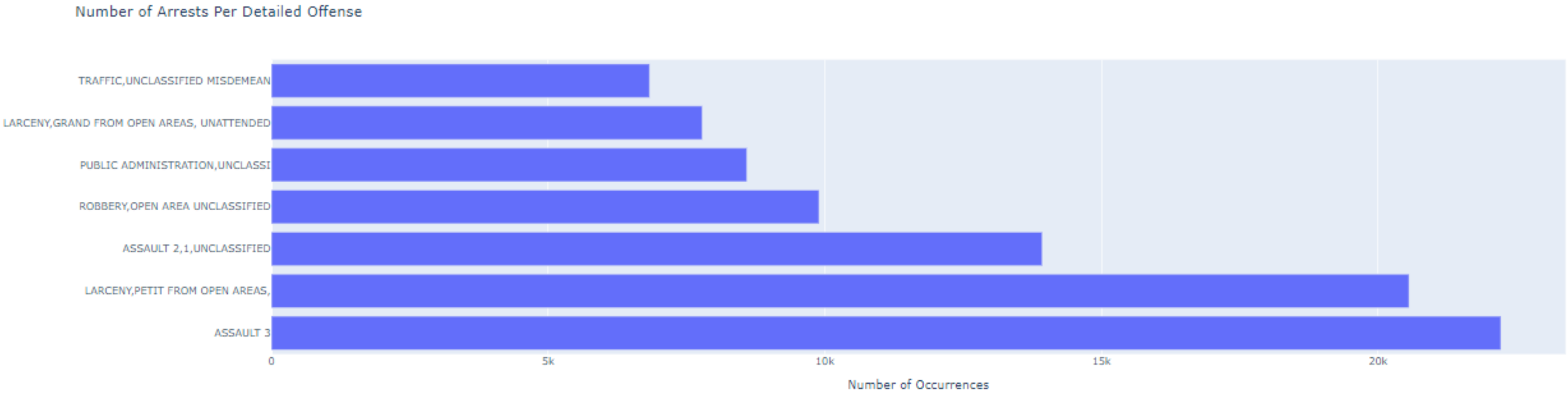
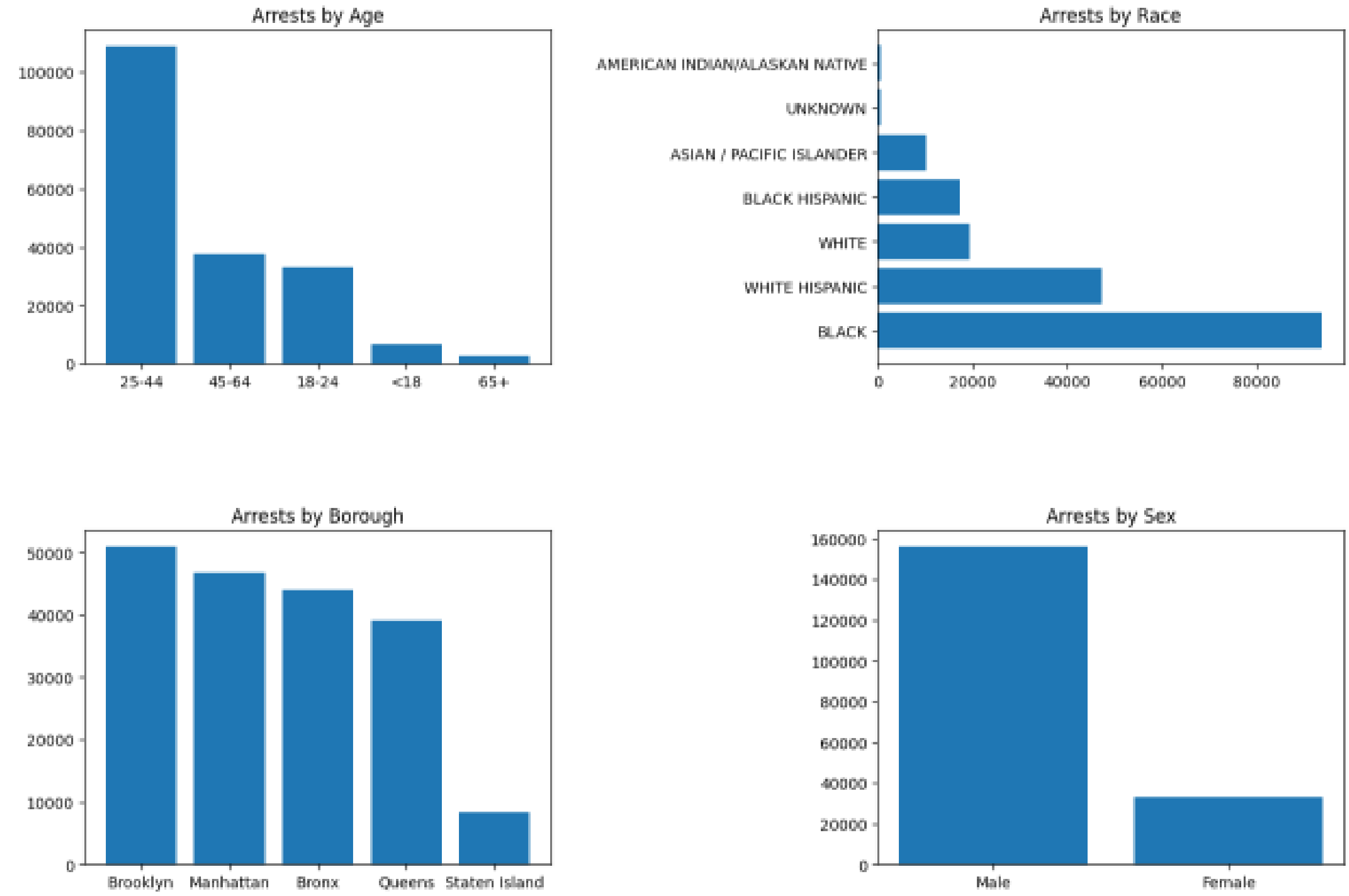


Figure 4: Number of Arrests per Detailed Offense

Additionally, we looked at the number of arrests according to the racial, age, and gender of the perpetrators. This enabled us to spot any trends or prejudices in the arrests that were made. Based on the results of these queries and visuals, we found that black race witnessed the highest number of arrests with 93,851; followed by white Hispanics with 47,241 arrests. The reason for these numbers might be because of racism and bias in the criminal justice system which is a modern topic nowadays in the US. Also it can be due to factors related to poverty, lack of access to education or healthcare facilities among certain demographics such as black and Hispanic groups. The age groups with the highest arrest rates were ages from 25-44 with 108,968 arrests and as the age increases, the number of arrests decreases. Based on gender, the arrested people were mostly males with 156,112 arrests compared to 33,092 to females. Usually, young adults and adult males of this age group have the highest levels of aggression and risk taking behavior and this could explain the reason behind these numbers. The following figure shows the different demographical information.


Arrests by Age

Arrests by Race

Arrests by Borough

Arrests by Sex

Figure 5: Arrests by Age, Race, Borough and Gender

We studied the number of arrests per unique location using longitude and latitude in each borough to understand the distribution of arrests among distinct locales. The most dangerous locations found in each of the boroughs in our dataset are were:

- In Manhattan: Harlem, Midtown, Chelsea, East Village, Lower East Side
- In Bronx: Mott Haven, Port Morris, Hunts Point
- In Queens: Corona, Jackson Heights, Elmhurst
- In Brooklyn: Bedford-Stuyvesant, Crown Heights, East New York, Brownsville.

According to the Property Club in NY, the most dangerous neighborhood in New York is Hunts Point. We examined the number of arrests taking place there based on latitude and longitude and found that they hold around 20 arrests. The following figures show geolocation related figures.

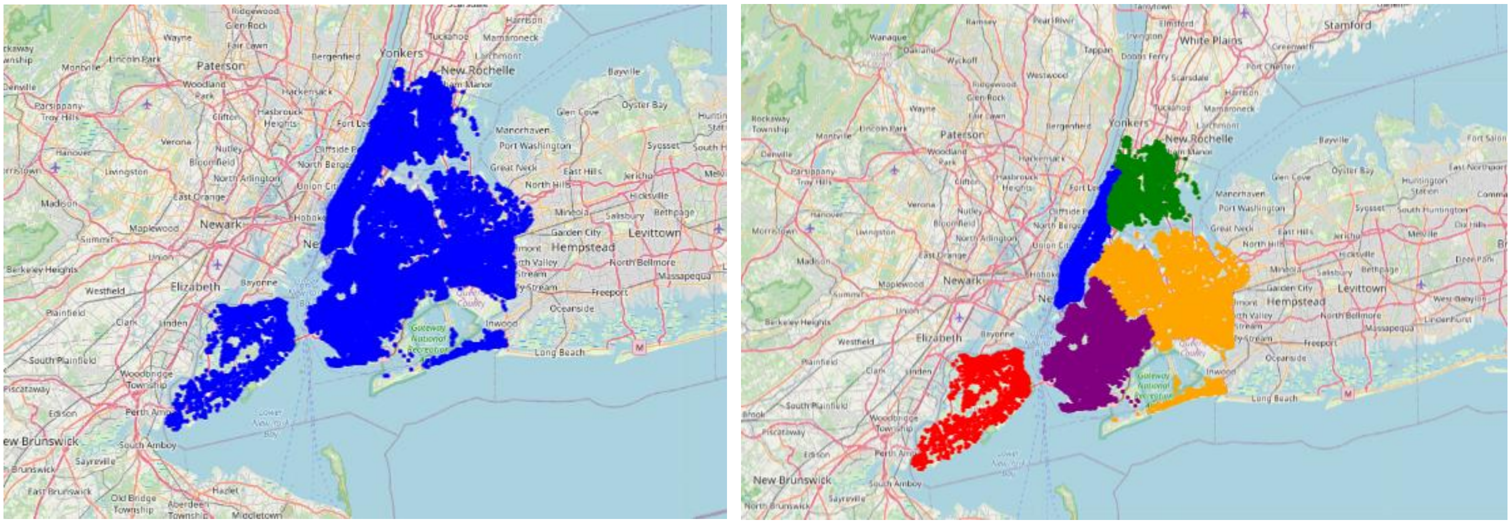


Figure 6: Maps showing the locations of arrests

In order to find any seasonal patterns or variations in criminal activity, we finally examined the number of arrests made per day and per month. According to the results obtained, October showed the highest number of arrests with 17,087 arrests followed by March and November with 16,901 and 16,607 respectively. The lowest arrest per month was in January with 13,148 arrests. The reason for this might be due to several reasons such as holidays and weather conditions. It is possible that public places and increased levels of alcohol consumption might lead to extra trouble during holidays in NYC (Columbus Day and Halloween in October, Thanksgiving and Veterans day in November) with January lacking holidays besides New year’s eve. Also, colder weather might lead to an increase in indoor criminal activity. In Brooklyn however, the highest number of arrests was in March with 4,687. According to the New York Times, several crimes took place, especially in the subway station where shootings took place and at least 23 people were injured. The following figures show time-related visuals.

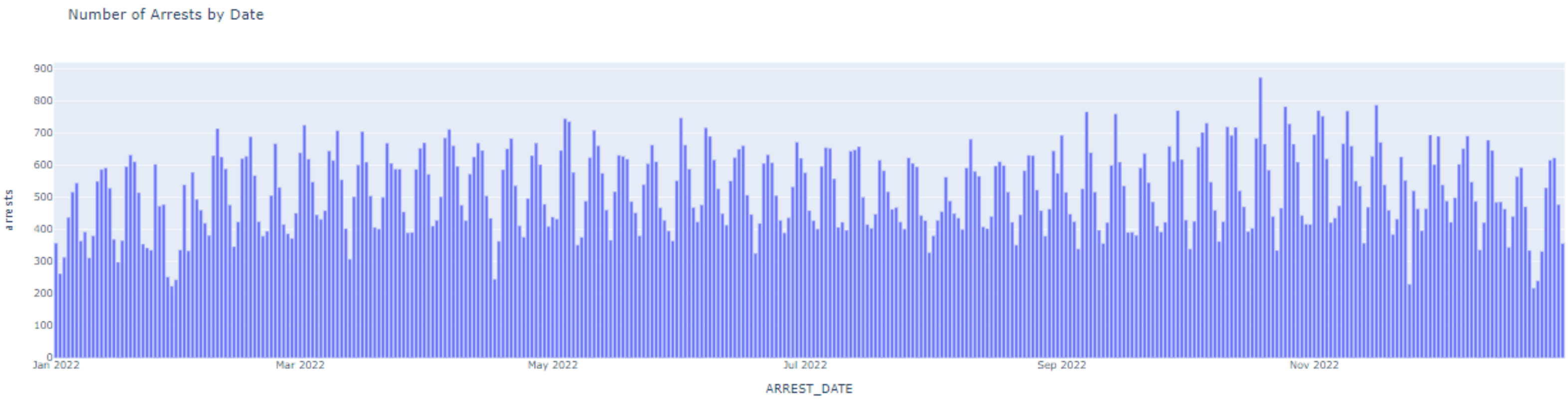


Figure 7: Total Number of Arrests by Date (Clear Seasonality)

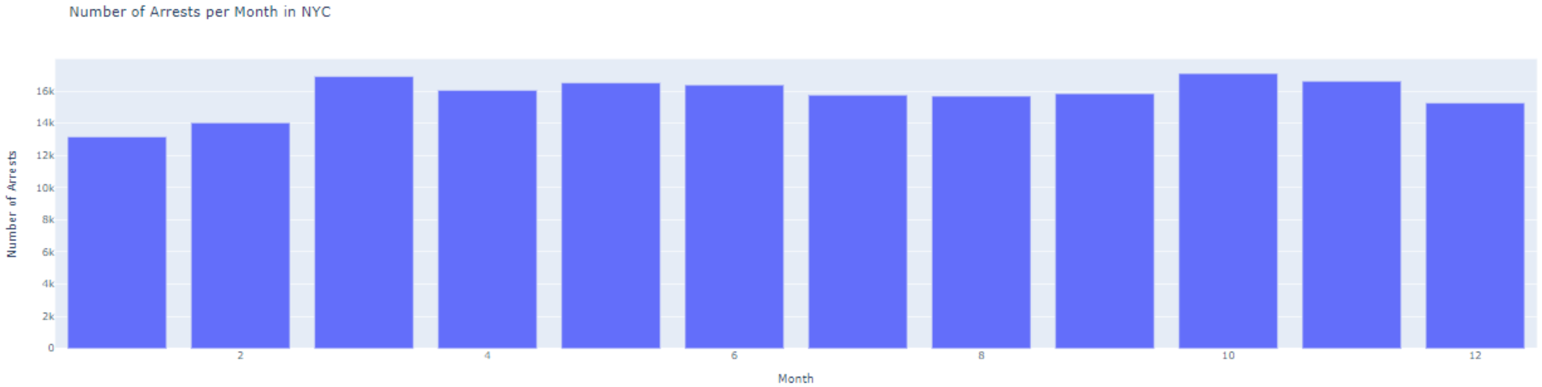


Figure 8: Total Number of Arrests by Month (October is the highest)

We then studied the number of arrests per Level of Offense and Jurisdiction Group. We noticed that the dominant level of offense was misdemeanor with around 54% of arrests, followed by felony with around 50% of arrests. We also noticed that almost 90% of all arrests were captured by the Patrol group while only 2.1% were captured by non-NYPD groups. The following figures show the percentages of arrests per Jurisdiction Group and per level of offense.

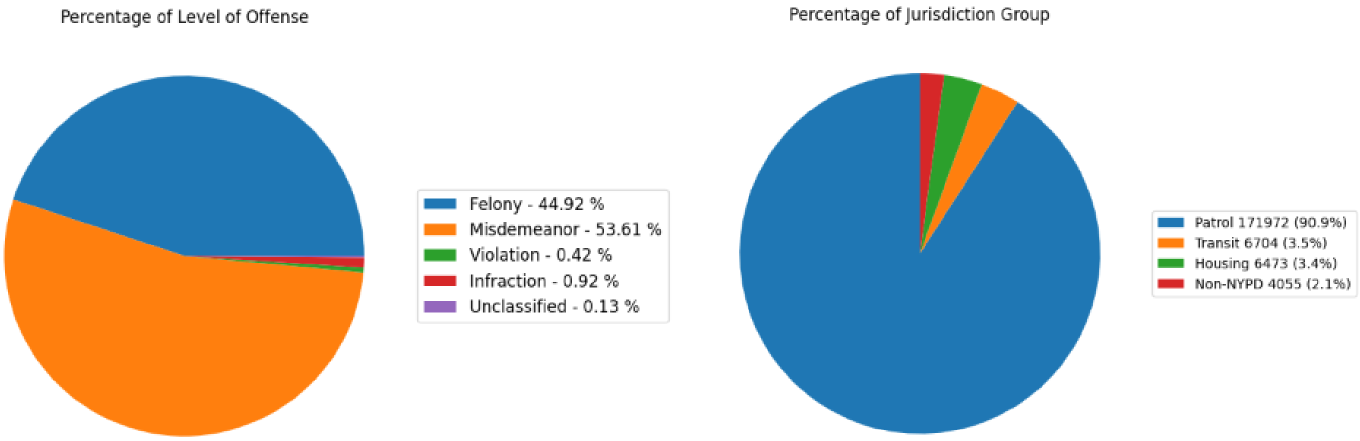


Figure 9: Percentage of Arrests per Level of Offense and per Jurisdiction Group

We also looked at all the arrests precinct and all the Brooklyn arrest precincts. Precinct 14 has the highest number of arrests in New York as a whole while Precinct 75 has the highest number of arrests in Brooklyn with around 5,706. According to CBS news, “The 75th is the largest precinct in Brooklyn North. It's a diverse neighborhood where community policing efforts are part of a shared responsibility”.

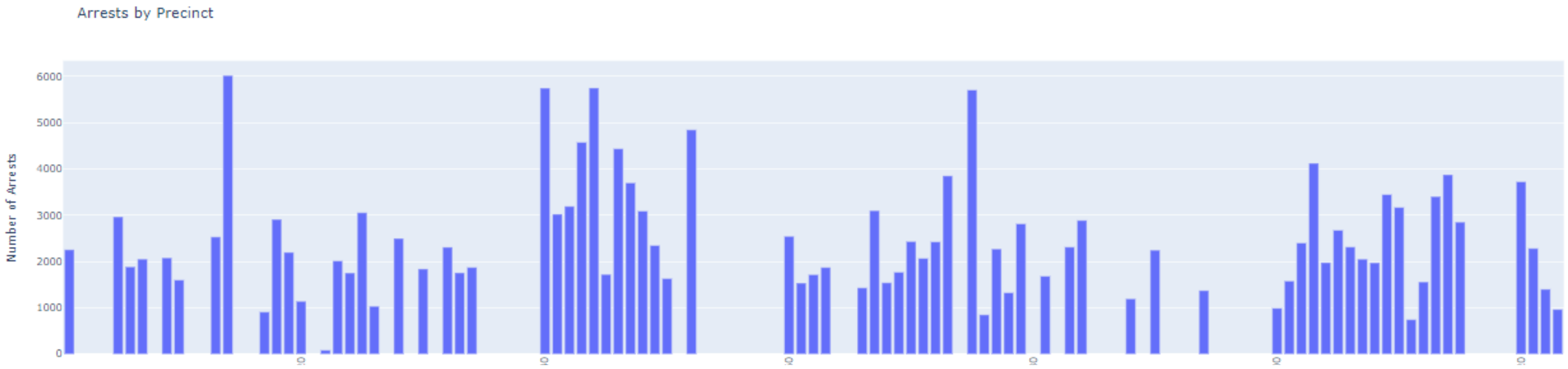


Figure 10: Arrests by Precinct (14 has the highest number of arrests)

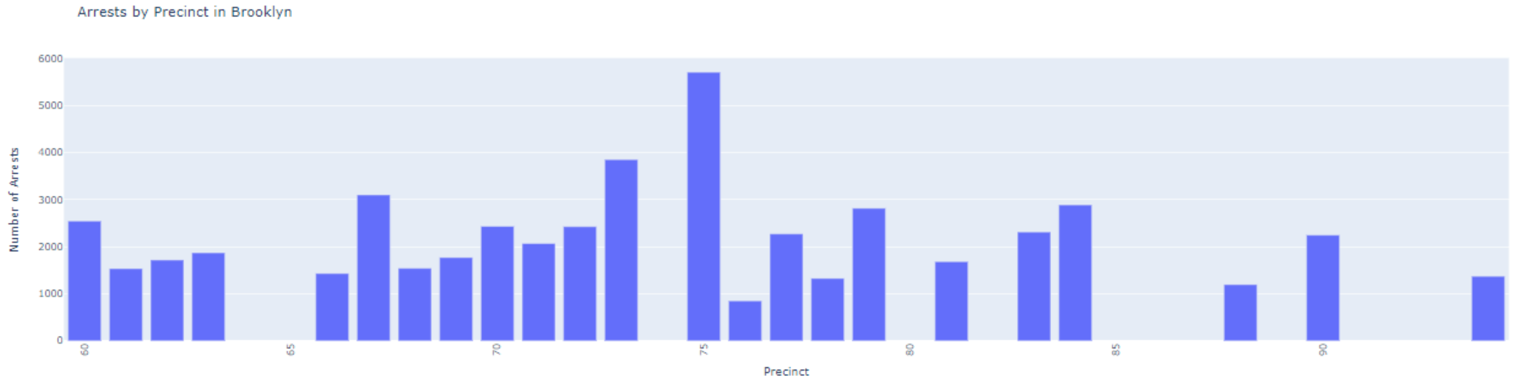


Figure 11: Arrests by Precinct in Brooklyn (75 has the highest number of arrests)

After conducting the necessary analysis, we plotted a number of visualizations as they provide an easy way to examine and convey complex information. These queries and visualizations played a crucial role in facilitating our understanding of the data and in deriving insights.

Finally, we applied the model which was described in the previous section. Our predictive model was a success from our first trial, with an accuracy of of 90%, a weighted recall of 90$, and a weighted precision of 90%.

5 Conclusion

In conclusion, analyzing the NYPD arrest dataset has revealed important information about the criminal justice system and the nature of crime in New York City. Our study was able to spot patterns and trends in the data through our analysis, including the distribution of arrests between boroughs, the most frequent offenses with their description, and the demographics of people who were arrested. These discoveries can guide the development of programs and policies that enhance public security, lower crime rates, and advance equity in the criminal justice system.

One of the key findings in our study was the significant racial difference in arrest numbers. More black people were arrested compared to any other race. These discrepancies cause serious concerns regarding the role of prejudice and discrimination in the criminal justice system overall and in law enforcement. Even though there are probably numerous factors, such as socioeconomic factors and historical inequalities that contribute to these disparities it is obvious that addressing these problems must be a top priority in efforts to reform the criminal justice system. Regardless of a person's color or ethnicity, policymakers and law enforcement organizations must cooperate to guarantee that policing procedures are equitable and that all individuals are treated fairly, regardless of their race or ethnicity.

In addition, more arrests are happening in Brooklyn and Manhattan in comparison to other locations in Brooklyn. Therefore, extra security measures should be taken in those areas the number of crimes are expected to be higher than other places. We also noticed that arrests are happening in areas with huge gatherings (just like the protest shooting which happened in March), which suggests that extra caution should be taken during such occasions.

Based on our findings, we recommend the NYPD and other authorities to look more closely at the factors behind the high arrest rates in particular demographics, location and times to put up initiatives to deal with these problems. This includes increasing police patrols in certain locations and times and offering police officers advanced training programs as well as funding community and educational initiatives to give less fortunate people greater opportunities.

6 References

Catlett, C., Cesario, E., Talia, D., & Vinci, A. (2019). Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments. *Pervasive and Mobile Computing,* 62-74.

CBS News. (2018, March 24). *March for Our Lives: Gun Control Protests NYC*. Retrieved from CBS News: https://www.cbsnews.com/newyork/live-updates/march-for-our-lives-gun-control-protests-nyc/

CBS News. (2021, January 11). *Work being done to change reputation of Brooklyn's 75th precinct, the inspiration for new CBS show "East New York"*. Retrieved from CBS New York: https://www.cbsnews.com/newyork/news/work-being-done-to-change-reputation-of-brooklyns-75th-precinct-the-inspiration-for-new-cbs-show-east-new-york/

Levine, E. S., Tisch, J., Tasso, A., & Joy, M. (2017). The New York City Police Department’s Domain Awareness System. *Interfaces,* 70-84.

McCabe, J. E., Braga, A., & Macdonald, J. M. (2022). Body-worn Cameras, Lawful Police Stops, and NYPD Officer Compliance: A Cluster Randomized Controlled Trial. *Criminology 60(1),* 124-158.

The New York Times. (2022, April 12). *Brooklyn Subway Shooting*. Retrieved from The New York Times: https://www.nytimes.com/live/2022/04/12/nyregion/brooklyn-subway-shooting