**ALY6110 – Data Management and Big Data**

**Final Project Report**

A logo of a university

Description automatically generated

**Shubham Nilesh Palande**

**Group 12**

**October 15, 2023**

**Introduction:**

This report is to present the arrests in New York City and its neighborhood by the New York Police Department (NYPD) from January 2023 to June 2023. The dataset provides thorough records of every arrest, including vital details about the type of offense, the exact place, and the time of enforcement. Furthermore, the dataset offers insightful information on the characteristics such as age, gender, and race of the implicated suspects. This publicly available dataset serves as a valuable foundation for in-depth exploration of police enforcement operations within the city.

The dimensions of the dataset are 112572 rows and 19 columns. This dataset includes comprehensive statistics on arrests that have occurred in New York City. Every entry contains the precise date and time of the arrest occurrence and is uniquely recognized by a persistent Arrest Key that is produced at random. The dataset includes textual descriptions (PD\_DESC and OFNS\_DESC) for a more thorough understanding in addition to internal classification codes (PD\_CD and KY\_CD) that offer both general and detailed categorizations of the violation. To emphasize the legal background of the arrests, law-related facts including the Law Code, Law Category, and the Jurisdiction Code are also presented. The precinct in which the arrest occurred, the borough of the arrest, the demographics of the perpetrator (age, sex, and race), and the exact location coordinates in several coordinate systems (X\_COORD\_CD, Y\_COORD\_CD, Latitude, Longitude) are additional attributes. The tabular format of the attributes is displayed below:

|  |  |
| --- | --- |
| Variable Names | Variable Information |
| ARREST\_KEY | Unique ID for each arrest |
| ARREST\_DATE | Date of the arrest event |
| PD\_CD | Internal classification code for the offense |
| PD\_DESC | Description of the offense corresponding to PD code |
| KY\_CD | General classification code for the offense |
| OFNS\_DESC | General description of the offense |
| LAW\_CODE | Charges based on NYS Penal Law, VTL, and local laws |
| LAW\_CAT\_CD | Level of offense (felony, misdemeanor, violation) |
| ARREST\_BORO | Borough where the arrest occurred |
| ARREST\_PRECINCT | Precinct where the arrest took place |
| JURISDICTION\_CODE | Jurisdiction responsible for the arrest |
| AGE\_GROUP | Age category of the perpetrator |
| PERP\_SEX | Sex of the perpetrator |
| PERP\_RACE | Race of the perpetrator |
| X\_COORD\_CD | X-coordinate in New York State Plane Coordinate System |
| Y\_COORD\_CD | Y-coordinate in New York State Plane Coordinate System |
| Latitude | Latitude coordinate in WGS 1984 |
| Longitude | Longitude coordinate in WGS 1984 |

The tools used for the analysis of this NYPD arrest dataset are Tableau for visualization, Python as well as Apache Spark for programming.

Business Questions:

1. What are the most common kinds of crimes in New York City?
2. Which neighborhoods are most likely to have crime?
3. What are the characteristics of people who are more likely to commit crimes?
4. What are the environmental factors that are linked to crime?

**Analysis:**

**A screenshot of a computer screen

Description automatically generated**

Figure 1: Bubble Chart representing different crimes by different gender

The study of the NYPD Arrest Data - Year to Date dataset, which was shown using a Tableau dashboard, from January 2023 to June 2023 resulted in significant findings regarding the distribution of arrests by gender in New York City. The above bubble graph represents the overall types of crime in NYC by Gender. With a count of 90,596 arrests, males accounted for the biggest proportion of the arrests, as shown by the gender-specific pie chart in figure 2. Female arrests accounted for 18,975 cases, whilst arrests of people of other genders represented 3,000 cases. This breakdown highlights the disproportionate number of arrests made of men, suggesting that there may have been differences in the participation of different gender groups in criminal activity during the first half of the year.

A pie chart of arrest

Description automatically generated

Figure 2: Pie Chart of Arrests by Gender

The data shows a sharp rise in the number of arrests in May 2023, with a considerable increase in criminal activity for both males and females as compared to the previous months. Interestingly, the data shows that males were arrested at the highest rate during this time, with 15,703 arrests in May 2023. On the other hand, there were 3,501 arrests made of women in May 2023, and 845 arrests of people of other genders in the same month. This apparent increase in arrests, especially in May, raises the idea that certain incidents, societal norms, or environmental factors are responsible for the noted rise in criminal activity among certain gender categories.

A screenshot of a graph

Description automatically generated

Figure 3: Area Plot of Monthly Arrests by Gender

According to the data analysis, the age group of 25-44 accounted for the highest percentage of arrests, indicating a concentration of law enforcement activities involving members of this demographic. Moreover, concerning racial demographics, the data indicates that the Black racial group accounted for the greatest proportion of arrests, with a total of almost 53,000 arrests made within the designated period. The White Hispanic population was next in line, accounting for almost 29,000 arrests, suggesting a significant presence in the total number of arrest records. Later next in line were White, Black Hispanic and Asian / Pacific Islander demographic proportion.

A screenshot of a graph

Description automatically generated

Figure 4: Stacked Bar Plot of Arrests by Age Group and Race

A summary of the number of arrests broken down by Borough of Arrest. Brooklyn has had the greatest number of arrests during the time, with 31,104, as the statistics clearly shows. The Bronx came in second with 26,757 arrests, followed by Manhattan with 25,903 and Queens with 23,922 arrests. At 4,886 arrests throughout the designated period, Staten Island had the fewest number of arrests. This distribution draws attention to the differences in law enforcement activity between the boroughs and raises the possibility of regional differences in crime rates and enforcement tactics.

A screenshot of a computer

Description automatically generated

Figure 5: Tile Plot of Total Arrests by Borough

The data reveals an important difference in the number of weekly arrests across different age groups, with the 25-44 age group showing more than thrice as many arrests as the 18-24 and 45-64 age groups. On the other hand, the data shows that those 65 years of age and older, as well as those under the age of 18, regularly record the fewest weekly arrests, usually fewer than 200 cases per week. The age group of 25-44 accounts for a huge number of arrests, on average of 2,600 per week, which highlights the need for focused interventions and all-encompassing programs that address the unique issues and situations that this group faces. It can be observed that the beginning of spring gradually increased the crime as all the age groups from 25-44, 45-64, and 18-24 hit the highest arrest days as 2734 in the week of June 18th, 953 in the week of April 30th and 886 in the week of June 4th arrests respectively.

A screenshot of a graph

Description automatically generated

Figure 6: Line Plot of Weekly Arrests by Age Group

The city's most common crime categories, which include theft, assault, and drug-related events, are centered in important districts like Manhattan, Brooklyn, and the Bronx. This causes a noticeable spike in criminal activity in May 2023, suggesting a possible temporal trend. The high arrest rate among the 25–44 age range emphasizes the vulnerabilities within this demographic, and the disproportionate arrest rates among the Black and White Hispanic communities emphasize the critical need to address the socioeconomic inequities that these communities face. Furthermore, the disproportionate number of men involved in arrest incidents highlights the significance of comprehending gender-specific elements that contribute to criminal activity, and the differences in arrest rates among boroughs highlight the impact of environmental factors, like unemployment and poverty, on differing crime rates in various areas.

A screenshot of a computer

Description automatically generated

Figure 7: Dashboard related to demographic characteristics

The dashboard's stacked bar chart analysis shows that, with 16,000 arrests, Assault Level 3 and related offenses account for the most arrests. These are followed by hazardous narcotics, felony assault, and petit larceny. This shows that to address these kinds of criminal activity, focused intervention techniques are desperately needed. Additionally, the use of a second stacked bar chart that shows the total number of arrests by month in each of the boroughs as Bronx is represented in Blue, Manhattan in Green, Staten Island in Purple, Brooklyn in Orange, and Queens in Red. This graph offers insightful information about the type of criminal activities in various areas of New York from January 2023 to June 2023. It can also be observed that as the beginning of Spring, the crime rates started to increase gradually.

A screenshot of a graph

Description automatically generated

Figure 8: Dashboard related to the type of crime and location

The analysis was further proceeded using Spark to read the file and get all the insights of Criminal Arrests in NYC. The figure below displays the Top 20 rows of the dataset with respect to their attribute names.

A screenshot of a computer

Description automatically generated

Figure 9: Top 20 Rows of NYPD Arrest Dataset

This code prints the PySpark DataFrame schema, displaying the data types and structure of the columns in the NYPD dataset for report analysis. It can be also observed that most the attributes are in string format while the rest of them are integers and doubles and do not have any Boolean type. This can be observed in the figure 10 below, as well as figure 11 displays the NYPD dataset's columns summary statistics are computed by this code, which then presents the results as a Pandas DataFrame.

A screenshot of a computer code

Description automatically generated

Figure 10: Schema of NYPD Arrest Dataset

A computer screen shot of a computer

Description automatically generated

Figure 11: Summary Statistics

To clean the data, this code first looks for missing values in each column of the NYPD dataset, shows the columns that have missing values, and then removes any rows that have any missing values. It drops rows with missing values from these columns and specifically addresses null values in the following columns: "PD\_CD" (three-digit internal classification code), "KY\_CD" (three-digit internal classification code, a more general category than PD code), and "LAW\_CAT\_CD" (Level of offense: felony, misdemeanor, violation).

A close up of a code

Description automatically generated

Figure 12: Dataset with Null values

A close-up of a computer code

Description automatically generated

Figure 13: Dataset without Null values

Later, this code looks for duplicate rows in the cleaned NYPD dataset, counts the number of duplicate rows it finds, and prints the result "No duplicate rows found" if none are found. As it can be observed that are no duplicate rows in the dataset.

A screenshot of a computer program

Description automatically generated

Figure 14: Duplicates not found

For solving and predicting the previous business question, Data Modeling was conducted on this dataset. The provided code snippet shows how to create a binary classification model in PySpark to determine the gender of people involved in arrests based on a variety of features. The dataset is first loaded and preprocessed, with all features combined into a single vector and string columns converted into numerical values using StringIndexer. These models were trained with 80%, while tested with 20%. This task's model is a Logistic Regression classifier with specific parameter sets. The model is assessed using the accuracy metric on the test data after being trained on the training set. With an accuracy of 80.83%, the model does a respectable job of predicting, given the provided features, the gender of those involved in arrests. Later, Random Forest Classifier was used to predict the gender, where the accuracy was around 98%.

A screenshot of a computer code

Description automatically generated

Figure 15: Training and Testing Data

A screen shot of a computer code

Description automatically generated

Figure 16: Logistic Regression Model predicting Gender of arrests

A screen shot of a computer code

Description automatically generated

Figure 17: Random Forest Model predicting Gender of arrests

This sample of code illustrates how to create a Multinomial Logistic Regression model in PySpark to forecast the location (borough) of arrests according to specific features. The model is used to generate predictions on the test dataset after being trained on the training dataset. The accuracy score that is obtained can be expressed as a percentage, which shows how well the model predicts the borough of arrest using the provided features. This model's prediction accuracy for the Arrest Location, also known as the Arrest Borough, was 79.37%. Brooklyn had the most arrests, as shown by the earlier graphs.

A computer screen shot of a computer code

Description automatically generated

Figure 18: Logistic Regressor predicting Arrest Borough

Answering the Business Questions:

1. Hazardous drugs, felony assault, petit larceny, and assault level 3 and related offenses are the most frequent types of crimes in New York City, according to an analysis of the NYPD Arrest Data from January 2023 to June 2023. These results emphasize the necessity of targeted intervention strategies to deal with these kinds of illegal activities.
2. In terms of the likelihood of crime in neighborhoods, Brooklyn had the most arrests (31,104), followed by the Bronx (26,757), Manhattan (25,903), and Queens (23,922). With 4,886 arrests, Staten Island had the fewest. This distribution indicates differences in law enforcement activity across boroughs, as well as possible differences in crime rates and enforcement strategies across neighborhoods.
3. According to the analysis, the number of arrests increased significantly in May 2023 compared to previous months, indicating a peak in criminal activity during that time. This increase in arrests was notable for both men and women. The data also shows that the 25-44 age group had the highest percentage of arrests. The rise in crime rates as spring arrived indicates a temporal trend in criminal activity. Furthermore, the Black racial group had the highest proportion of arrests, followed by White Hispanic, White, Black Hispanic, and Asian/Pacific Islander demographics. These findings reveal the characteristics of people who are more likely to commit crimes.
4. The presence of specific crime categories, such as theft, assault, and drug-related events, which are concentrated in important districts such as Manhattan, Brooklyn, and the Bronx, appears to be one of the environmental factors linked to crime. In May 2023, there was a noticeable increase in criminal activity in these areas. After peak winters, people start getting out of their houses and visit places, which can have an impact on theft and other related crimes.

**Conclusion:**

Finally, the NYPD Arrest Data Analysis from January 2023 to June 2023 provides valuable insights into the dynamics of law enforcement and criminal activity in New York City. The most common types of crimes in the city are hazardous drugs, felony assault, petit larceny, assault level 3 and related offenses, emphasizing the need for targeted interventions to address these issues. Brooklyn had the highest number of arrests, indicating differences in law enforcement practices and potential differences in crime rates across boroughs. The temporal trend of increased arrests in May 2023 highlights the role of environmental factors, as the arrival of spring may have resulted in higher crime rates. Furthermore, the data sheds light on the demographic characteristics of those who are more likely to commit crimes, with the 25-44 age group and specific racial demographics prominently featured. The significance of comprehending and tackling the various elements that contribute to criminal behavior is underscored by these findings, which offer a thorough basis for well-informed decision-making and tactical law enforcement initiatives in New York City.

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>

*PySpark Overview — PySpark 3.5.0 documentation*. (n.d.). <https://spark.apache.org/docs/latest/api/python/index.html>

*MLlib (DataFrame-based) — PySpark 3.5.0 documentation*. (n.d.). <https://spark.apache.org/docs/latest/api/python/reference/pyspark.ml.html>