REPORT

Names: Dias Amangeldy**,** Ilya Rogov, Beybarys Rysbek

Group: CS-2117

Discipline: Big Data in Law Enforcement 2

Case: Endterm Project.

Link to dataset: https://data.world/city-of-ny/uip8-fykc/workspace/file?filename=nypd-arrest-data-year-to-date-1.csv.

Link to GitHub: https://github.com/deadA1R/Big-Data-2-Endterm

Analysis plan:

1. Introduction: Presentation of the dataset, its meaning and application in the context of law enforcement agencies.
2. Data preparation: Using tools to clean, transform and load data, ensuring that it is ready for analysis.
3. Data Analysis: In-depth analysis using statistical methods and visualization to identify trends and patterns.
4. Conclusion: Summarizing the key findings, discussing the significance of the analysis and suggestions for further research.

The purpose of the analysis:

The purpose of this project is a comprehensive analysis of law enforcement data to identify key trends, patterns and anomalies in committed offenses. This includes an analysis of the types of offenses, the time and place of their commission, as well as the demographic characteristics of offenders. The main task is to identify potential patterns that can help law enforcement agencies in planning measures to prevent crime and improve public safety.

Introduction

This report focuses on a comprehensive dataset detailing arrests made by the NYPD. It's critical for understanding patterns in law enforcement activities, including the types of crimes leading to arrests, demographic information about suspects, and the locations and times of these incidents.

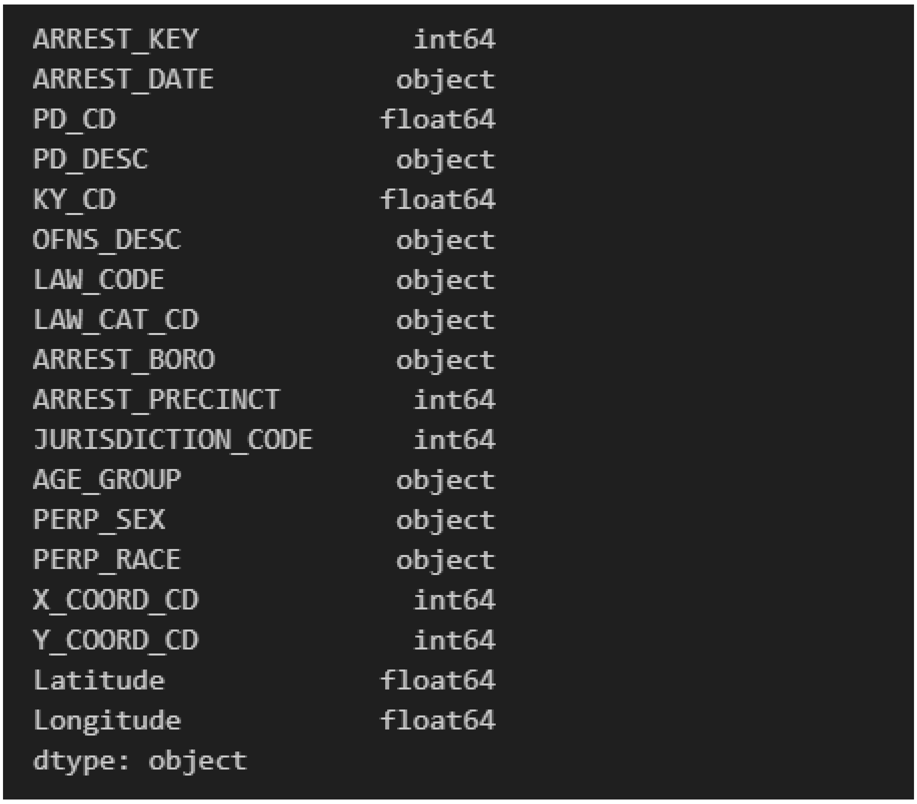
Data Preparation

The dataset underwent an extensive ETL process using Pandas for cleaning and transformation. This included removing duplicates, handling missing values, and ensuring correct data types for analysis. SQL was utilized for structured queries to refine the data further for meaningful insights.

Analysis. The analysis was conducted through several lenses:

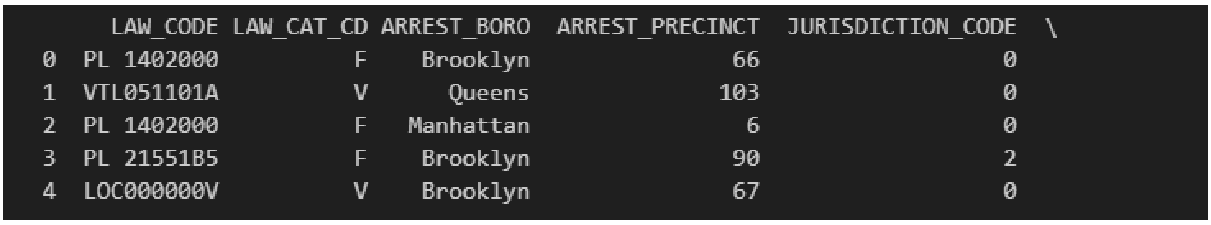
- Pandas and NumPy were used for preliminary explorations, revealing trends in crime types and demographic distributions.

We conducted an analysis of arrest data provided by the New York Police Department. The original dataset was loaded from the 'nypd-arrest-data-year-to-date-1.csv' file using the pandas library. After loading, we examined the data types of each column (screenshot 1) and replaced values in the 'ARREST_BORO' column using a dictionary to match abbreviations with full borough names (screenshot 2).



```
ARREST_KEY               int64
ARREST_DATE             object
PD_CD                  float64
PD_DESC                 object
KY_CD                  float64
OFNS_DESC               object
LAW_CODE                object
LAW_CAT_CD              object
ARREST_BORO             object
ARREST_PRECINCT          int64
JURISDICTION_CODE        int64
AGE_GROUP               object
PERP_SEX                object
PERP_RACE               object
X_COORD_CD               int64
Y_COORD_CD               int64
Latitude               float64
Longitude              float64
dtype: object
```

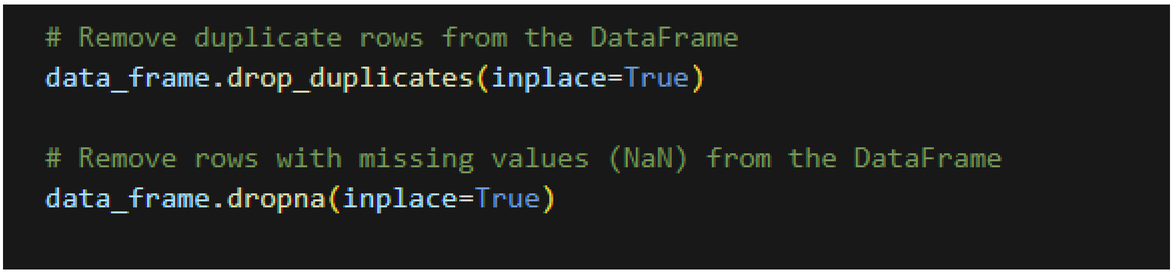
Screenshot 1



```
     LAW_CODE LAW_CAT_CD ARREST_BORO  ARREST_PRECINCT  JURISDICTION_CODE  \
0  PL 1402000          F    Brooklyn               66                  0
1  VTL051101A          V      Queens              103                  0
2  PL 1402000          F   Manhattan                6                  0
3  PL 21551B5          F    Brooklyn               90                  2
4  LOC000000V          V    Brooklyn               67                  0
```

Screenshot 2

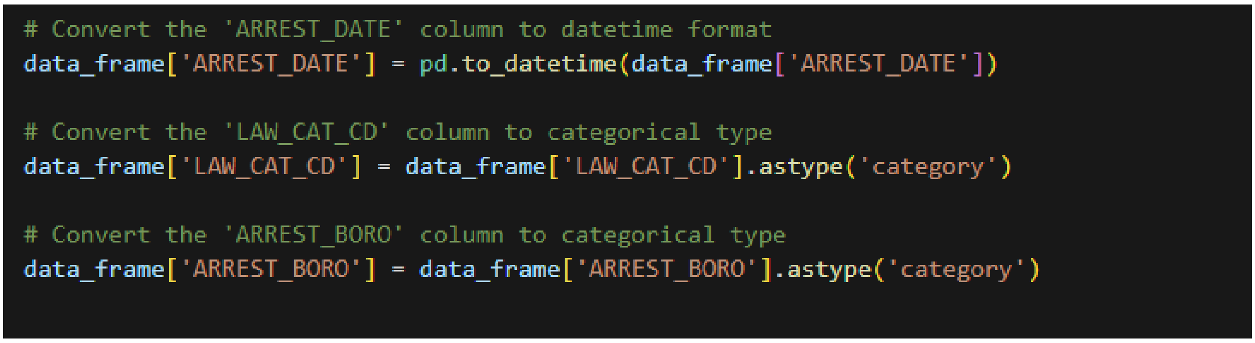
For data processing, we removed duplicate rows and rows with missing values. Then, we converted the 'ARREST_DATE' column to datetime format and the 'LAW_CAT_CD' and 'ARREST_BORO' columns to categorical data types (screenshot 3).



```
# Remove duplicate rows from the DataFrame
data_frame.drop_duplicates(inplace=True)

# Remove rows with missing values (NaN) from the DataFrame
data_frame.dropna(inplace=True)
```

Screenshot 3



```
# Convert the 'ARREST_DATE' column to datetime format
data_frame['ARREST_DATE'] = pd.to_datetime(data_frame['ARREST_DATE'])

# Convert the 'LAW_CAT_CD' column to categorical type
data_frame['LAW_CAT_CD'] = data_frame['LAW_CAT_CD'].astype('category')

# Convert the 'ARREST_BORO' column to categorical type
data_frame['ARREST_BORO'] = data_frame['ARREST_BORO'].astype('category')
```

Screenshot 3

Next, we conducted various data analyses, including counting unique types of crimes, distribution of age groups among arrestees, number of arrests over different time periods (by month), and number of arrests in each borough of the city. Here are the summarized results:

- Types of Crimes (OFNS_DESC):
  - DANGEROUS DRUGS: 19,710 occurrences
  - ASSAULT 3 & RELATED OFFENSES: 17,379 occurrences
  - PETIT LARCENY: 11,606 occurrences
  - VEHICLE AND TRAFFIC LAWS: 11,018 occurrences
  - FELONY ASSAULT: 7,580 occurrences
- Age Groups (AGE_GROUP):
  - 25-44: 67,567
  - 18-24: 28,997
  - 45-64: 25,612
  - <18: 6,944
  - 65+: 1,471
- Arrests Over Time (ARREST_DATE):
  - January: 23,797
  - February: 21,651
  - March: 22,382
  - April: 21,241
  - May: 21,970
  - June: 19,550
- Arrests by Borough (ARREST_BORO):
  - Brooklyn: 33,769
  - Manhattan: 32,386
  - Bronx: 28,651
  - Queens: 25,959
  - Staten Island: 5,341

These analyses provided valuable insights into the distribution of crimes, age demographics of arrestees, temporal patterns of arrests, and geographical distribution of arrests across different boroughs of the city.

Additionally, using the NumPy library, we calculated the average age of arrestees, the most common type of crime, total number of arrests, median age of arrestees, total count of unique crime descriptions, percentage of arrests in each borough relative to the total number of arrests, and average number of arrests per month.

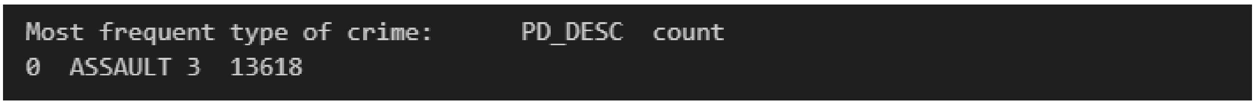
- Total Number of Arrests Recorded: 130591
- Total Number of Unique Crime Descriptions: 62
- Percentage of Arrests by Borough:
    - Brooklyn 28.153548
    - Manhattan 25.939000
    - Bronx 21.939490
    - Queens 19.878093
    - Staten Island 4.089868
- Average Arrests per Month: 21765.166666666668

All analysis results were saved in CSV files for further use and processing. This analysis provides an overview of actions taken and results obtained during the analysis of New York Police Department arrest data, which can be valuable for decision-making and further research.

- SQL queries identified specific patterns, such as the most frequent crime types and areas with high arrest rates.

New York Law Enforcement Analysis Report using SQL

1. Total number of arrests.
   The dataset includes a total of 130,591 arrests registered in the New York Law Enforcement database.
2. The most common type of crimes leading to arrests.
   The most common type of crime leading to arrests is "Assault 3", with a total of 13,618 cases. Attack 3 represents a significant portion of the arrests in the dataset, indicating its prevalence as a reason for law enforcement intervention.

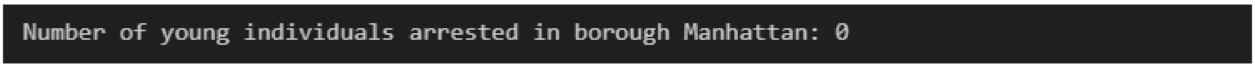


```
Most frequent type of crime:      PD_DESC  count
0  ASSAULT 3  13618
```

3. Borough with the Highest Number of Arrests in 2021
   The analysis to identify the borough with the highest number of arrests in 2021 did not yield any results. This might be due to a lack of data within the specified date range (January 1, 2021, to December 31, 2021), or it could indicate that no arrests were recorded within that period. Further investigation or clarification is necessary to determine the cause of this absence of data.

4. Arrests of Young Individuals in Borough Manhattan

5. There were no arrests of young individuals (age group labeled as "Y") recorded in Borough M according to the provided dataset.



```
Number of young individuals arrested in borough Manhattan: 0
```

Conclusion

The analysis provides insight into trends and patterns in the New York law enforcement dataset. It highlights the prevalence of assault-related crimes as a significant cause of arrests and the lack of arrests of youth in the Manhattan area.

- Apache Spark enabled the processing of large volumes of data, focusing on aggregating information by crime type and examining arrest trends over time and across different boroughs.

The provided code snippet utilizes PySpark to analyze NYPD arrest data. Here's a comprehensive analysis of the code:

The script begins by configuring a SparkConf and SparkContext, setting the application name as "ArrestAnalysis," and reading the data from a CSV file located at "/content/nypd-arrest-data-year-to-date-1.csv" into an RDD (Resilient Distributed Dataset). The header line is skipped to ensure that only data lines are processed.

The *parseLine* function is defined to parse each line of the RDD, extracting relevant fields such as age group, arrest borough, and offense type. This function ensures that the data is properly structured and filters out any invalid records.

Parsed data is transformed into key-value pairs, where the key represents the offense type, and the value is a tuple containing the arrest borough, age group, and a count of 1 indicating one arrest. This transformation prepares the data for aggregation and analysis.

The script aggregates the parsed data by offense type using the *reduceByKey* function, which sums up the counts of arrests for each offense type. This step provides insight into the overall frequency of different types of offenses.

To identify the top 5 offenses with the highest number of arrests, the script utilizes the *takeOrdered* function, which retrieves the top offenses based on the number of arrests.

In terms of analysis, the code primarily focuses on determining the top 5 offenses by the number of arrests made. However, it does not delve into further analysis beyond identifying these top offenses.

To enhance the analysis, the script could be extended to calculate the distribution of arrests by age group and borough for each offense type. Additionally, incorporating error handling and data validation would improve the robustness of the analysis.

The top 5 offenses by the number of arrests are as follows:

- 'UNCLASSIFIED': 22,789 arrests
- 'ASSAULT 3 & RELATED OFFENSES': 15,289 arrests

- '235': 13,494 arrests
- '348': 10,683 arrests
- '117': 5,736 arrests

These results indicate that offenses categorized as 'UNCLASSIFIED' and 'ASSAULT 3 & RELATED OFFENSES' are the most frequent causes of arrest, followed by offenses labeled '235', '348', and '117'.

In summary, while the provided code snippet offers valuable insights into the top offenses based on arrest frequency, there is potential for further exploration and refinement to gain deeper insights into the NYPD arrest data.

Conclusion

The analysis unearthed significant insights into the nature and distribution of arrests in NYC, highlighting prevalent crime types and areas with higher arrest frequencies. These findings are vital for law enforcement strategies, allowing for targeted interventions and resource allocation to address specific crime patterns effectively. This report underscores the importance of data-driven approaches in enhancing public safety and law enforcement efficacy.