Analysis of NYPD Arrests Data (2019 - 2024)

1. Introduction

This report provides an analysis of the NYPD arrests data from 2019-2024. It is complemented by the code and comments provided in NYCC_Data_Challenge.ipynb. This report discusses the trends in arrest rates, most common types of arrests and a comparison of crime levels in specific precincts. Additionally, it includes visualizations of geographical hotspots as well as temporal patterns of arrests in NYC. It also proposes predictive models to forecast crime in order to assist NYPD in future resource allocation. This analysis aims to support the Council's Public Safety Committee in understanding trends in police enforcement in the city.

2. Data Validation and Cleaning

The data was sourced from NYPD Arrests Data (Historic) and NYPD Arrest Data (Year to Date). The following steps were taken to validate and clean the data:

i. Initial Data Inspection

The datasets were inspected to understand their structure, including data types, formats and potential inconsistencies.

ii. Data Cleaning

- **Datetime Consistency:** The "ARREST_DATE" columns in both the historic and year-to-date datasets were converted to a consistent datetime format.
- **Data Filtering:** The historic data was filtered to include only data from 2019 to 2023. The year-to-date dataset was checked to consist of arrests only from 2024.
- **Column Renaming:** The "Lon_Lat" column in historic data and the "New Georeferenced Column" in the year-to-date data represent the same information. To maintain consistency and prevent data loss during merging, the "New Georeferenced Column" was renamed to "Lon_Lat".
- **Combining Datasets:** The historic and year-to-date datasets were then merged into a single dataframe for a unified analysis of arrests from 2019-2024.

iii. Handling Missing Values

- **Crime Descriptions and Codes:** It was observed that "PD_CD", "PD_DESC", "KY_CD" and "OFNS_DESC" columns contained missing values, typically occurring together. It might have been possible to infer one column value from another as these categories are more general or granular versions of each other. However, since they were usually missing simultaneously and constituted a very small proportion of the dataset (less than 0.1%), these rows were dropped to maintain data quality.
- **Offense Category:** The column "LAW_CAT_CD" categorizes the level of offense and had approximately 0.7% null values. To retain maximum amount of usable data, these missing values were replaced with the category "9" ("Unknown or Not Applicable").

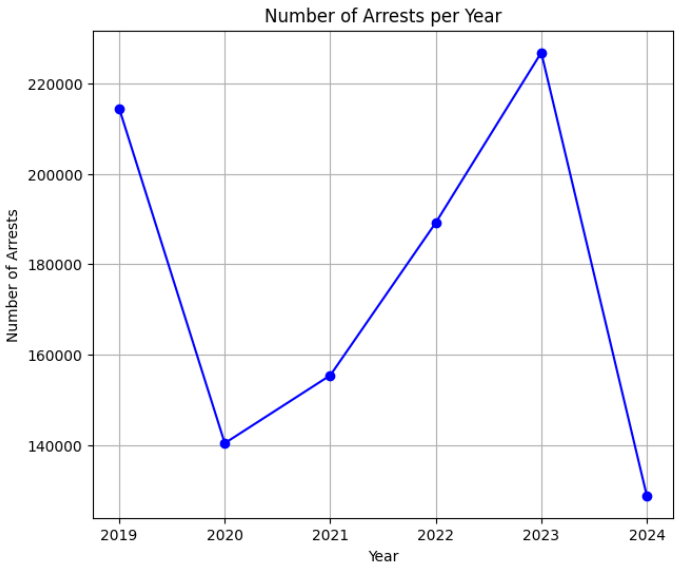
3. Arrest Rate Trend (2019-2024)

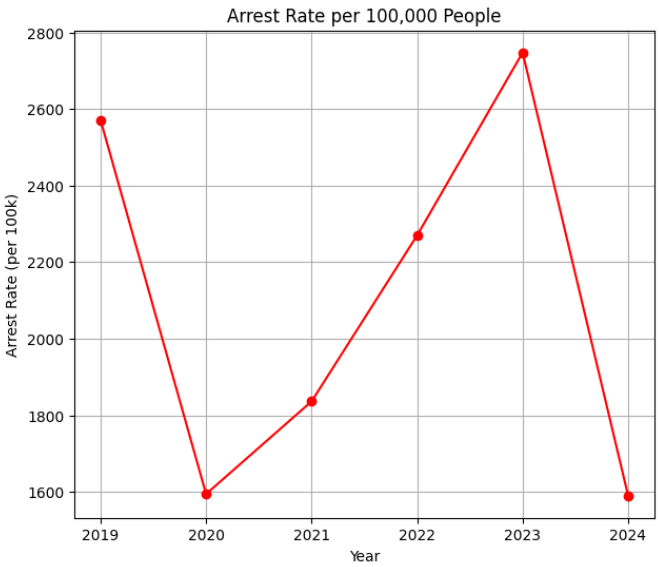
I obtained population estimates for New York City from 2019 to 2024 from different sources. The official estimates for 2020-2023 were released by the NYC Department of City Planning and can be found in their report. Estimates for 2024 and 2019 were sourced from the World Population Review website.

In order to visualize the trend in arrest rate, I used both the raw number of arrests each year as well as the arrest rate per 100,000 people calculated as:

$$Arrest\ Rate = \left(\frac{Number\ of\ Arrests}{Population}\right) \times 100{,}000$$

i. Visual Trend Analysis

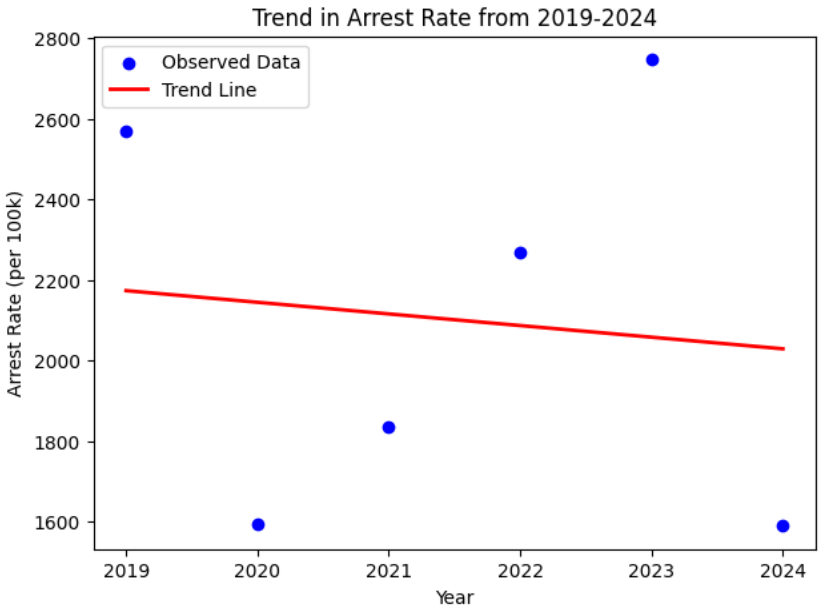




Based on visual inspection, the plots exhibit similar trends in arrest rate over the years. The following observations can be made:

- The arrest rate exhibited a significant decline in 2020, likely due to the COVID-19 pandemic and quarantine restrictions. After 2020, there was a notable increase in arrest numbers, reaching a peak in 2023.
- The plot shows a substantial decrease in arrests in 2024 so far, compared to the previous year. This may be because we're only halfway through the year. Additionally, it could also reflect a likely shift in enforcement patterns or external factors influencing arrest rates.

ii. Linear Regression Analysis

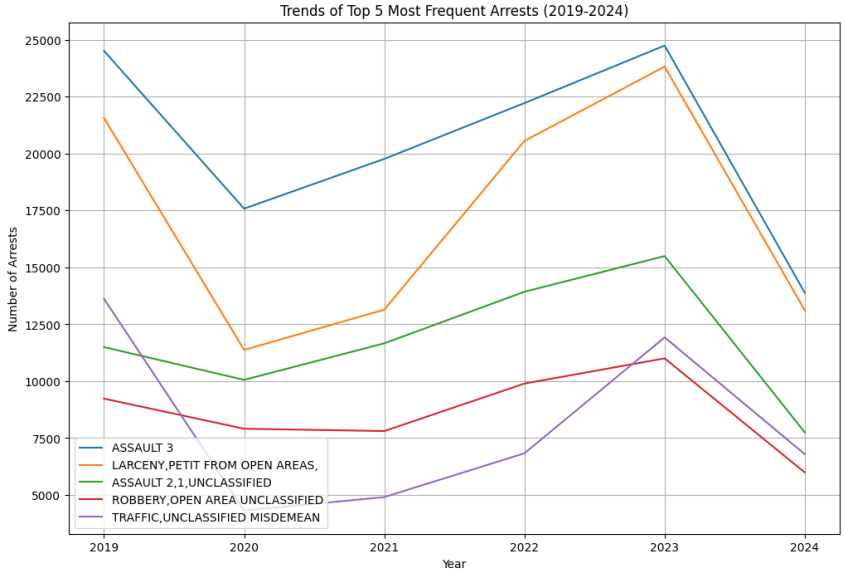


I applied a linear regression model to the data to quantify the trend in arrests rate. The regression analysis yielded a slope of -28.8679, indicating that, on average, the arrest rate decreased by approximately 28.87 arrests per 100,000 people for each additional year. This negative trend suggests an overall decline in arrest rates over the analyzed period.

4. Top 5 Arrests

The top 5 most frequent arrests analyzed from “PS_DESC” column from 2019-2024 were found to be:

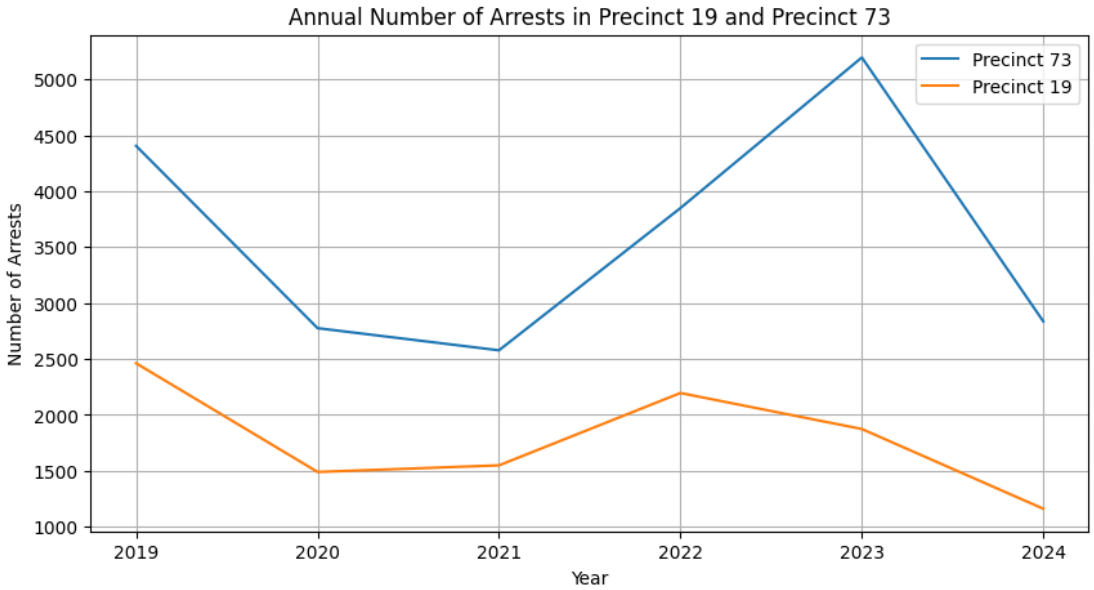
1. Assault in the third degree
2. Petit larceny from open areas
3. Assault in the second and first degree (unclassified)
4. Robbery in open area (unclassified)
5. Traffic misdemeanor (unclassified)



Across all categories, there's a common pattern of fluctuation. The most frequent arrests typically declined in 2020, rebounded and peaked in 2023, then dropped again in 2024.

5. Crime Comparison

By visualizing the arrests of Precinct 73 and 19 from 2019 to 2024, we see that Precinct 73 saw a significant spike in arrests in 2023, whereas Precinct 19 experienced a slight decrease. Over the period from 2019 to 2024, Precinct 73 consistently reported higher arrest numbers compared to Precinct 19.



i. T-test for Crime Comparison Between Precinct 19 and 73

To further compare the crime reported between these two precincts, I used the t statistical test to evaluate if there's a significant difference in the arrest counts between Precinct 73 and 19.

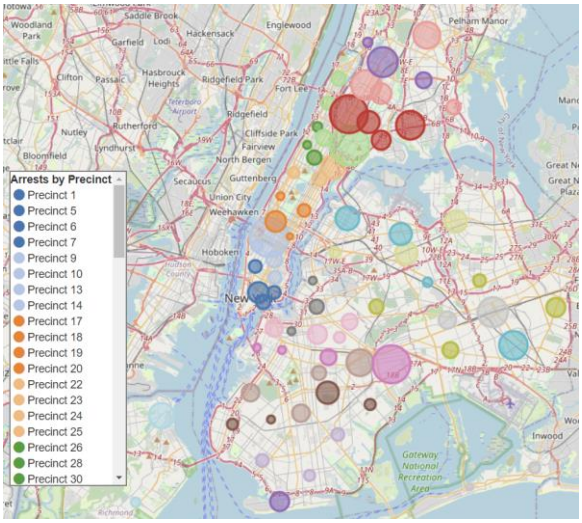
To perform the t-test, we establish the following hypotheses:

- **The Null Hypothesis H_0:** There is no significant difference in the mean number of arrests between Precinct 19 and Precinct 73.
- **The Alternative Hypothesis H_A:** There is a significant difference in the mean number of arrests between Precinct 19 and Precinct 73.

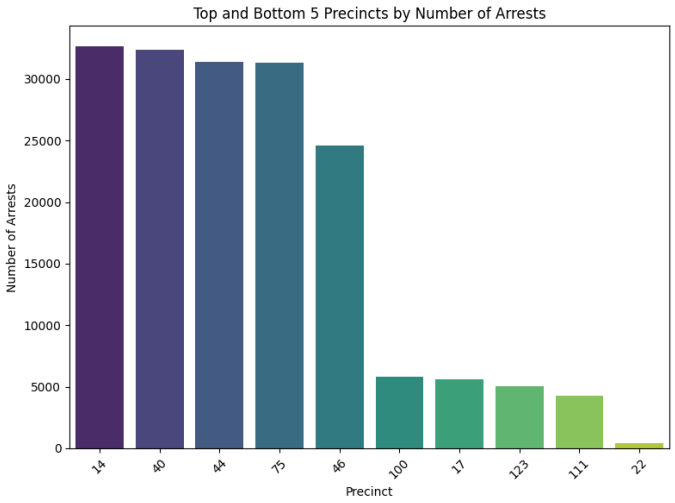
The t-test results: t-statistic = -3.8383 and p-value = 0.0033 indicate a significant difference between the number of arrests in Precinct 19 and Precinct 73. Where Precinct 73 has a significantly higher average number of arrests compared to Precinct 19. The p-value of 0.0033 confirms that this difference is statistically significant.

6. Future Resource Allocation

i. Overview of Arrests by Precinct

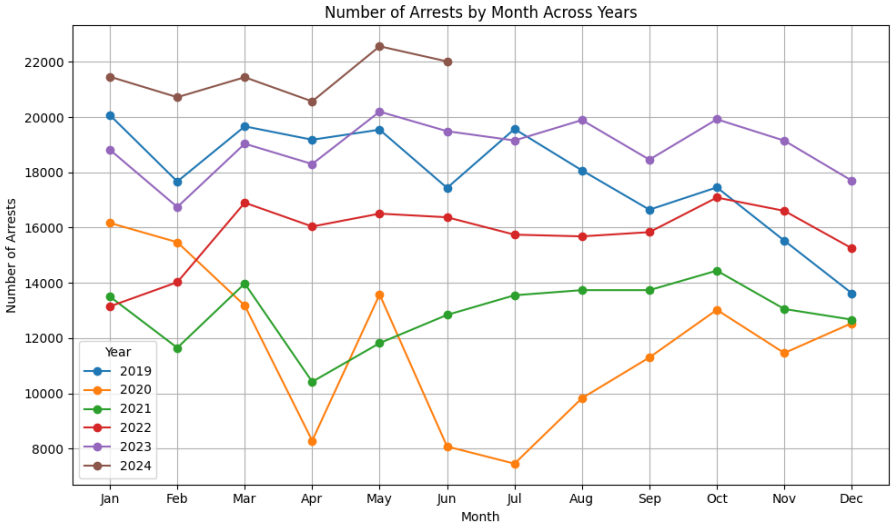


To gain insights into the spatial distribution of arrests, I created an interactive HTML file, 'precinct_arrests.html', which is available in the GitHub repository. This visualization maps the concentration of arrests in each precinct, with marker sizes proportional to the number of arrests.



- **Top Precincts**: Precincts 14, 40, 44, 75 and 46 show the highest numbers of arrests, ranging from approximately 24,000 to 32,000 arrests. These precincts may require additional resources such as specialized units, increased patrol presence and support staff to manage the higher volume of arrests compared to others.
- **Bottom Precincts:** Precincts 100, 17, 123, 111 and 22 exhibit much lower arrest frequencies, with counts ranging from about 400 to 5,800. For these units, the NYPD can consider optimizing resource use based on specific needs and crime trends.

ii. Temporal Analysis of Arrests



- **Seasonal Trends:** Arrests generally peak during spring to summer (April to June) and dip during fall to winter (October to February). This suggests a higher need for resources during spring/summer when arrest volumes are consistently higher.
- **2024 Arrests:** Despite the overall annual arrest numbers for 2024 showing a decline, it's important to note that this is due to the year being only halfway through. Monthly arrest data for 2024 shows each month's figures surpassing those of previous years.

iii. Predictive Modeling

- **Model Choice**: To better plan future resource allocation by predicting arrests count, a Random Forrest Regressor can be used. This model can handle non-linear relationships and interactions between features. It also provides feature importance, which can help understand which factors most influence crime rates.
- **Independent Variables**:
    1. **Historical Crime Data:** Past crime counts (lagged features), types of crimes and locations.
    2. **Temporal Features**: Time of the year, day of the week and time of day.
    3. **Demographic Data:** Population density, age distribution and socioeconomic indicators of precincts.
    4. **Geospatial Features:** Proximity to high-risk areas (e.g., bars, clubs) and neighborhood safety ratings
    5. **Weather Data:** Conditions such as temperature, rainfall and extreme weather events.
- **Dependent Variable:** Crime count (number of reported crimes within a specific time frame and precinct)
- **Model Evaluation:**
    1. **Train-Test Split:** Split the data into training and testing sets to assess model performance.
    2. **Cross-Validation:** Use k-fold cross-validation to ensure the model generalizes well across different subsets of the data.
    3. **Performance Metrics:**
        - **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions
        - **Mean Squared Error (MSE):** Assesses the average squared difference between predicted and actual values.

    - **R-Squared (R^2):** Indicates the proportion of variance explained by the model.
- **Potential Challenges:**
    1. **Data Quality:** Missing or inaccurate data can impact model performance. Implement imputation strategies and data validation.
    2. **Feature Selection:** Choosing relevant features is crucial to avoid overfitting and ensure interpretability.
    3. **Temporal Changes:** Crime patterns can shift over time, so regularly updating the model with recent data is essential.
    4. **Privacy and Ethics:** Ensure that demographic and geospatial data usage complies with privacy laws and ethical guidelines.

This model can aid the NYPD in efficiently allocating resources by identifying high-risk areas and predicting crime trends.