NYPD ARREST DATA

TRENDS & INSIGHTS FOR CRIME PREVENTION

CONTRIBUTORS

POOJA SHAH
ADITYA GHUGE
TANVI SALIAN
ROHAN RAMAKRISHNA

Data link: https://catalog.data.gov/dataset/nypd-arrest-data-year-to-date

GOALS & OBJECTIVES

The primary goal of analyzing the NYPD Arrest Data and Crime Prevention Insights dataset is to enhance public safety through data-driven strategies. Objectives include identifying crime patterns and trends across time and locations, uncovering demographic factors contributing to arrest disparities, and pinpointing high-crime areas for targeted interventions. This analysis aims to support law enforcement in resource allocation, improve community policing efforts, and foster equitable policy-making to reduce crime effectively and fairly.

Research Questions:

1. How does the frequency of various offenses differ across different age groups?
2. Are there notable seasonal or temporal patterns in arrests for specific age groups?
3. How do the top 10 offenses vary across different boroughs in terms of arrest frequency?
4. How does the distribution of arrests by race vary across different boroughs?

DATASET USED & ITS CHARACTERISTICS

01 Contains detailed records of arrests made by the NYPD, including information such as offense type, arrest date, and location.

02 Includes demographic data (e.g., age, gender, ethnicity) of individuals arrested, enabling analysis of trends and disparities.

03 Features location-based data (boroughs and precincts), allowing for crime mapping and hotspot identification.

04 Provides time-stamped data, facilitating analysis of crime patterns across different days, months, and years for predictive modeling.

```
dataset = pd.read_csv("library/ist652/fall2024/Scripting_data.csv")
print(dataset.head())
```

```
   ARREST_KEY ARREST_DATE  PD_CD                    PD_DESC  KY_CD  \
0   281240883  01/28/2024  105.0         STRANGULATION 1ST  106.0
1   282884120  02/27/2024  263.0                ARSON 2,3,4  114.0
2   283137868  03/03/2024  109.0  ASSAULT 2,1,UNCLASSIFIED  106.0
3   287001362  05/16/2024  109.0  ASSAULT 2,1,UNCLASSIFIED  106.0
4   287829614  06/02/2024  105.0         STRANGULATION 1ST  106.0

        OFNS_DESC    LAW_CODE LAW_CAT_CD ARREST_BORO  ARREST_PRECINCT  \
0  FELONY ASSAULT  PL 1211200          F           Q              105
1           ARSON  PL 1501001          F           Q              107
2  FELONY ASSAULT  PL 1200502          F           B               48
3  FELONY ASSAULT  PL 1200512          F           S              121
4  FELONY ASSAULT  PL 1211200          F           Q              100

   JURISDICTION_CODE AGE_GROUP PERP_SEX PERP_RACE  X_COORD_CD  Y_COORD_CD  \
0                  0     25-44        M     WHITE     1057545      207911
1                 71     45-64        M     WHITE     1037489      206343
2                  0     25-44        M     BLACK     1013900      250835
3                  0     25-44        M     WHITE      938928      168468
4                  0     25-44        M     BLACK     1039777      155013

    Latitude  Longitude      New Georeferenced Column
0  40.737043 -73.735514  POINT (-73.735514 40.737043)
1  40.732881 -73.807899  POINT (-73.807899 40.732881)
2  40.855109 -73.892818  POINT (-73.892818 40.855109)
3  40.628967 -74.163275  POINT (-74.163275 40.628967)
4  40.591980 -73.800066   POINT (-73.800066 40.59198)
```

DATA CLEANING & TRANSFORMATION

01 We started by standardizing the dataset to ensure consistency. This included renaming columns to lowercase and replacing ambiguous or coded values with descriptive labels, like mapping borough codes ('M', 'K') to their full names ('Manhattan', 'Brooklyn')

02 Key columns with missing values, such as law_cat_cd and ky_cd, were addressed. We replaced them with meaningful placeholders like 'Unknown' for law_cat_cd and -1 for numerical identifiers to avoid losing valuable data.

03 We ensured the arrest_date column was in a standardized datetime format. Invalid dates were removed, and the remaining data was formatted to MM/DD/YYYY for consistency.

04 We mapped singular values to full descriptions for better interpretability. For instance, law_cat_cd values like 'F', 'M', and 'V' were replaced with 'Felony', 'Misdemeanor', and 'Violation'. This step was crucial for generating insights.

DATA CLEANING

```
Unique values in arrest_boro after replacement:
['Brooklyn' 'Queens' 'Staten Island' 'Bronx' 'Manhattan']
Unique values in perp_sex after replacement:
['Male' 'Female']
Unique values in law_cat_cd after replacement:
['Felony' 'Misdemeanor' 'Violation' 'Unknown' 'Infraction']
```

```
Remaining missing values:
 pd_cd          0
ky_cd          0
law_cat_cd     0
dtype: int64
Rows filled with -1 in pd_cd: 4
Rows filled with -1 in ky_cd: 15
Rows filled with 'Unknown' in law_cat_cd: 770
```

```
Missing Values:
 ARREST_KEY                   0
ARREST_DATE                  0
PD_CD                        4
PD_DESC                      0
KY_CD                       15
OFNS_DESC                    0
LAW_CODE                     0
LAW_CAT_CD                 770
ARREST_BORO                  0
ARREST_PRECINCT              0
JURISDICTION_CODE            0
AGE_GROUP                    0
PERP_SEX                     0
PERP_RACE                    0
X_COORD_CD                   0
Y_COORD_CD                   0
Latitude                     0
Longitude                    0
New Georeferenced Column     0
dtype: int64
```

01

We started by standardizing the dataset to ensure consistency. This included renaming columns to lowercase and replacing ambiguous or coded values with descriptive labels, like mapping borough codes ('M', 'K') to their full names ('Manhattan', 'Brooklyn')

02

Key columns with missing values, such as law_cat_cd and ky_cd, were addressed. We replaced them with meaningful placeholders like 'Unknown' for law_cat_cd and -1 for numerical identifiers to avoid losing valuable data.

DATA CLEANING

```
  arrest_date
0  01/03/2024
1  01/03/2024
2  01/04/2024
3  01/15/2024
4  01/07/2024
Date range: 01/01/2024 06/30/2024
```

ARREST_DATE
01/03/24
01/03/24
01/04/24
01/15/2024
01/07/24
01/08/24
01/14/2024
01/23/2024
01/18/2024
02/05/24
02/07/24

```
Unique values in arrest_boro after replacement:
['Brooklyn' 'Queens' 'Staten Island' 'Bronx' 'Manhattan']
Unique values in perp_sex after replacement:
['Male' 'Female']
Unique values in law_cat_cd after replacement:
['Felony' 'Misdemeanor' 'Violation' 'Unknown' 'Infraction']
```

03

We ensured the arrest_date column was in a standardized datetime format. Invalid dates were removed, and the remaining data was formatted to MM/DD/YYYY for consistency.

04

We mapped singular values to full descriptions for better interpretability. For instance, law_cat_cd values like 'F', 'M', and 'V' were replaced with 'Felony', 'Misdemeanor', and 'Violation'. This step was crucial for generating insights.

DATA TRANSFORMATION

```
Date-based features added:

  arrest_date  arrest_year  arrest_month day_of_week  is_weekend
0  01/03/2024         2024             1   Wednesday       False
1  01/03/2024         2024             1   Wednesday       False
2  01/04/2024         2024             1    Thursday       False
3  01/15/2024         2024             1      Monday       False
4  01/07/2024         2024             1      Sunday        True
```

```
Severity Index values:

      law_cat_cd  severity_index
0         Felony             3.0
1         Felony             3.0
2         Felony             3.0
3    Misdemeanor             2.0
4         Felony             3.0

Unique severity index values:
 [ 3.   2.   1. nan]
```

01 From the arrest_date, we derived features such as the day of the week, month, and whether the arrest occurred on a weekend or a weekday. This enabled us to analyze temporal patterns effectively.

02 To quantify crime severity, we introduced a numeric severity index, assigning values like 3 for felonies, 2 for misdemeanors, and 1 for violations. This allowed us to assess trends based on severity.

DATA TRANSFORMATION

```
    latitude  longitude location_zone
0  40.674496 -73.930571     South NYC
1  40.602740 -73.750081     South NYC
2  40.698323 -73.917495     South NYC
3  40.623238 -74.149217     South NYC
4  40.734681 -73.810626     South NYC

Borough-Severity combination:

     arrest_boro   law_cat_cd              boro_severity
0       Brooklyn       Felony            Brooklyn_Felony
1         Queens       Felony              Queens_Felony
2       Brooklyn       Felony            Brooklyn_Felony
3  Staten Island  Misdemeanor  Staten Island_Misdemeanor
```

03

We categorized locations into 'North NYC' and 'South NYC' zones based on latitude, and combined borough and severity into a new feature, boro_severity, for deeper insights.

```
Date-based features added:

  arrest_date  arrest_year  arrest_month day_of_week  is_weekend
0  01/03/2024         2024             1   Wednesday       False
1  01/03/2024         2024             1   Wednesday       False
2  01/04/2024         2024             1    Thursday       False
3  01/15/2024         2024             1      Monday       False
4  01/07/2024         2024             1      Sunday        True
```

04

We grouped data by categories like age_group, law_cat_cd, and is_weekend to prepare for visualizations and understand trends across different dimensions.

EXPLORATORY DATA ANALYSIS

01

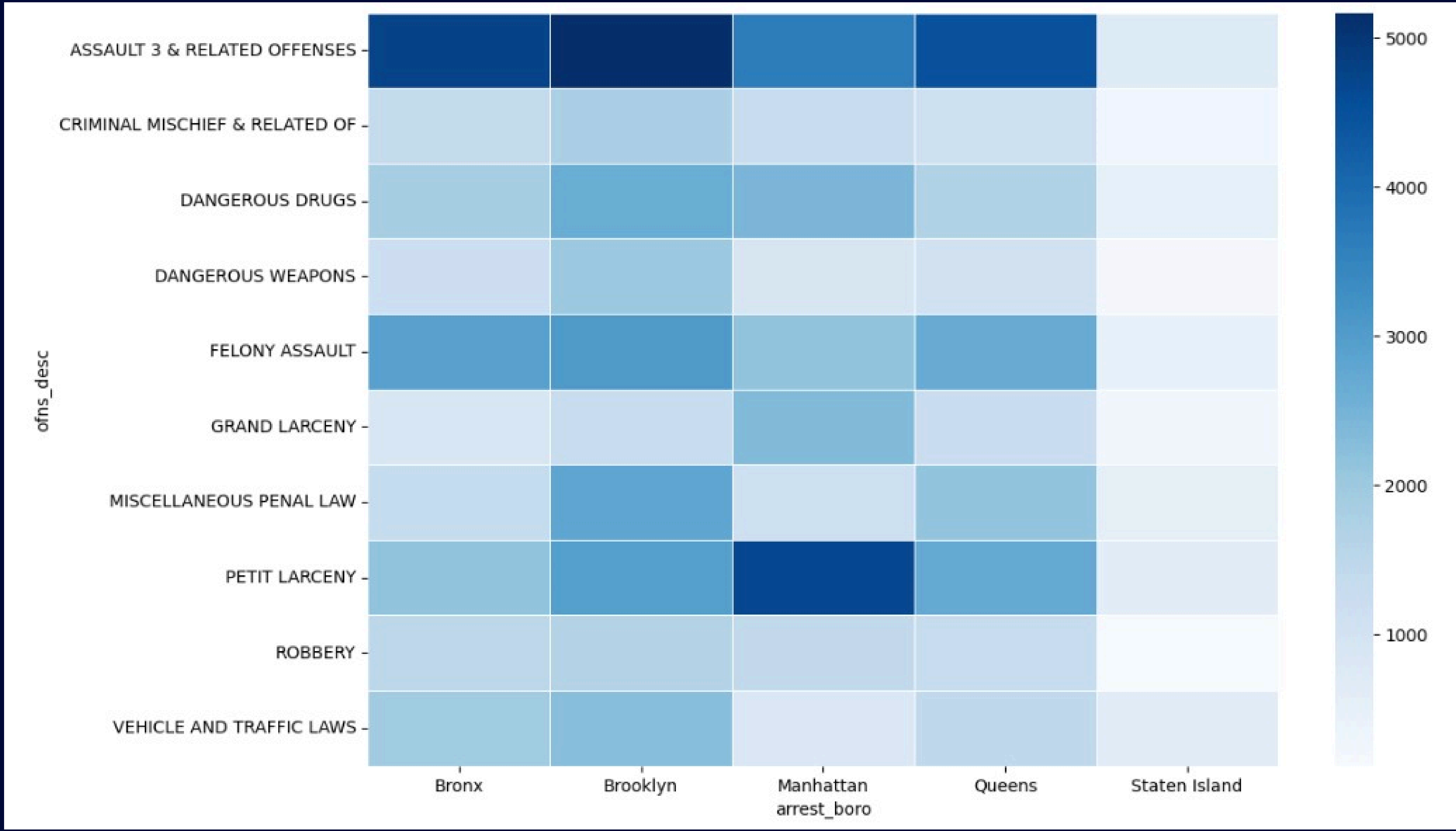
From the arrest_date, we derived features such as the day of the week, month, and whether the arrest occurred on a weekend or a weekday. This enabled us to analyze temporal patterns effectively.

02

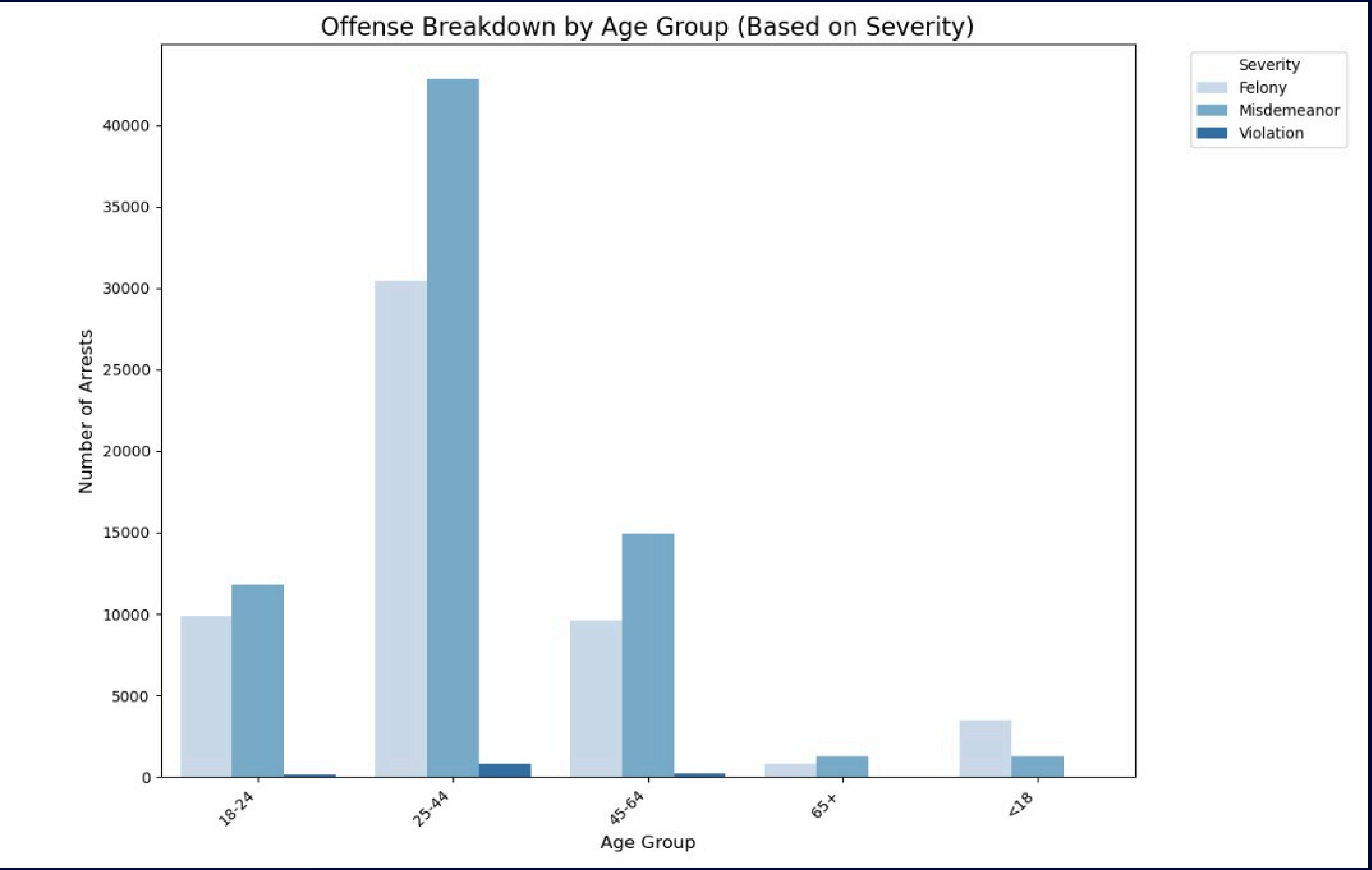
To quantify crime severity, we introduced a numeric severity index, assigning values like 3 for felonies, 2 for misdemeanors, and 1 for violations. This allowed us to assess trends based on severity.

03

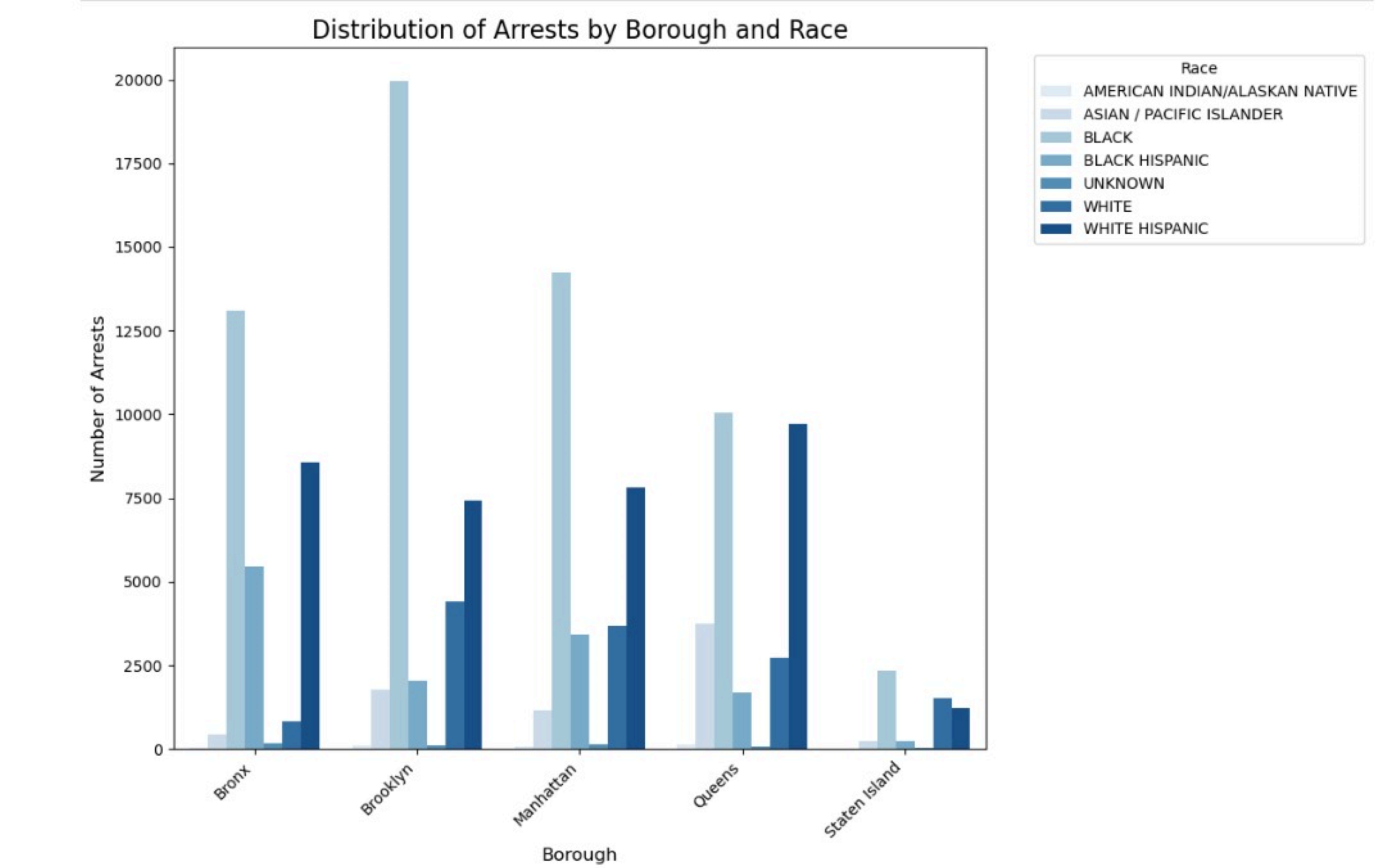
We categorized locations into 'North NYC' and 'South NYC' zones based on latitude, and combined borough and severity into a new feature, boro_severity, for deeper insights.



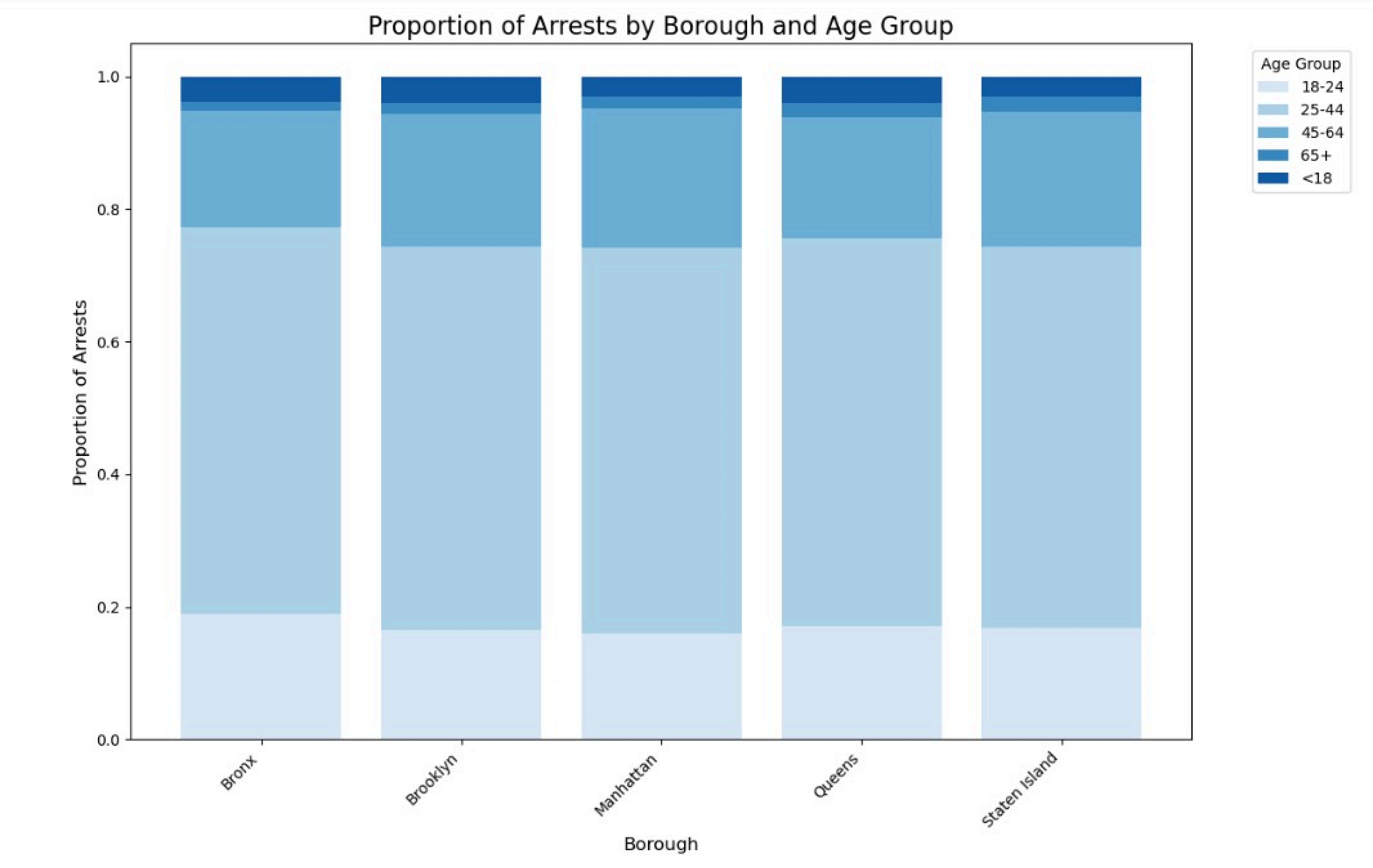
Borough-Offense Patterns: The heatmap highlights the frequency of top 10 offenses across boroughs, showcasing distinct crime hotspots for specific offenses.



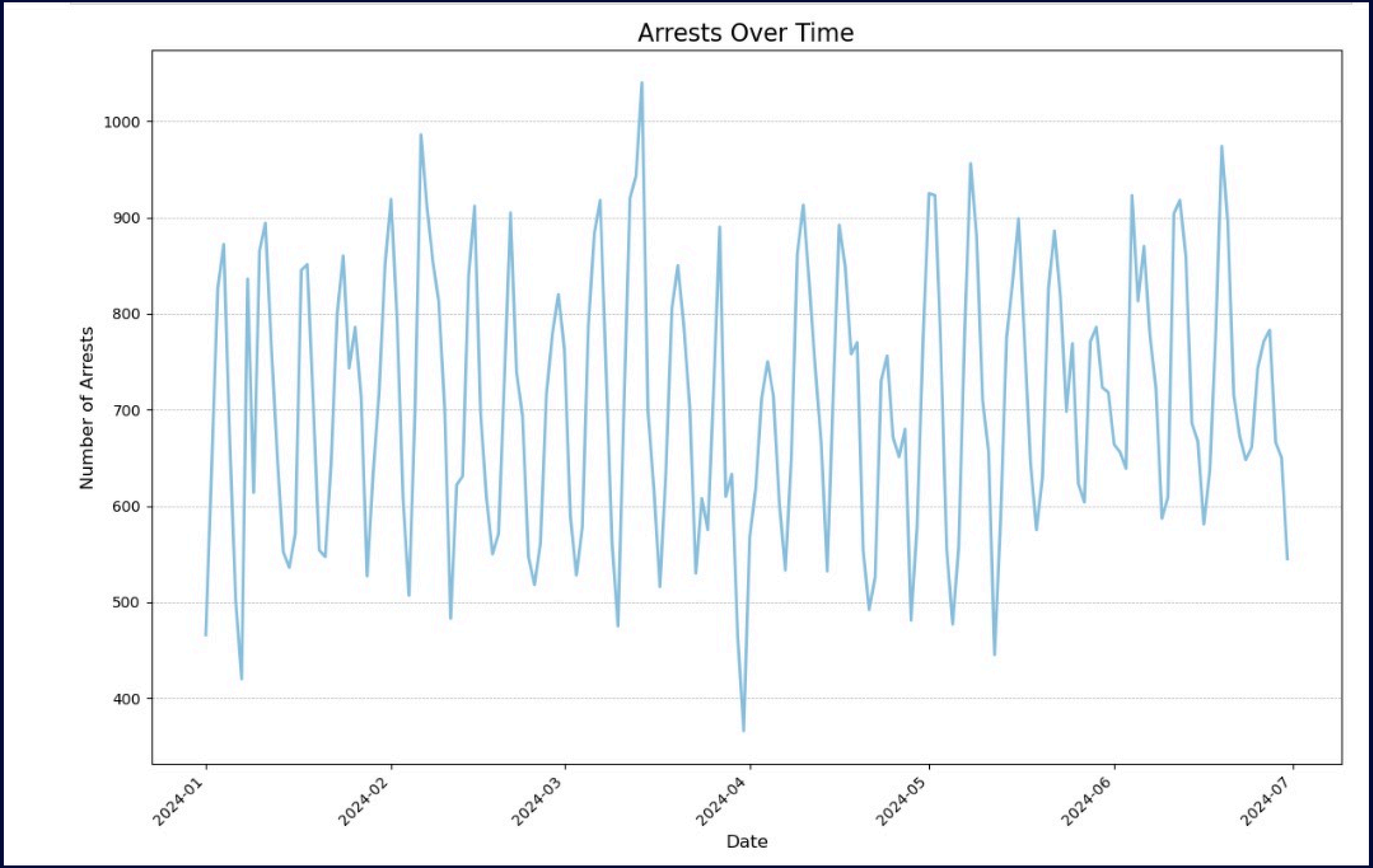
Offense-Age Correlation: The analysis highlights specific offenses prevalent in different age groups, aiding targeted interventions.



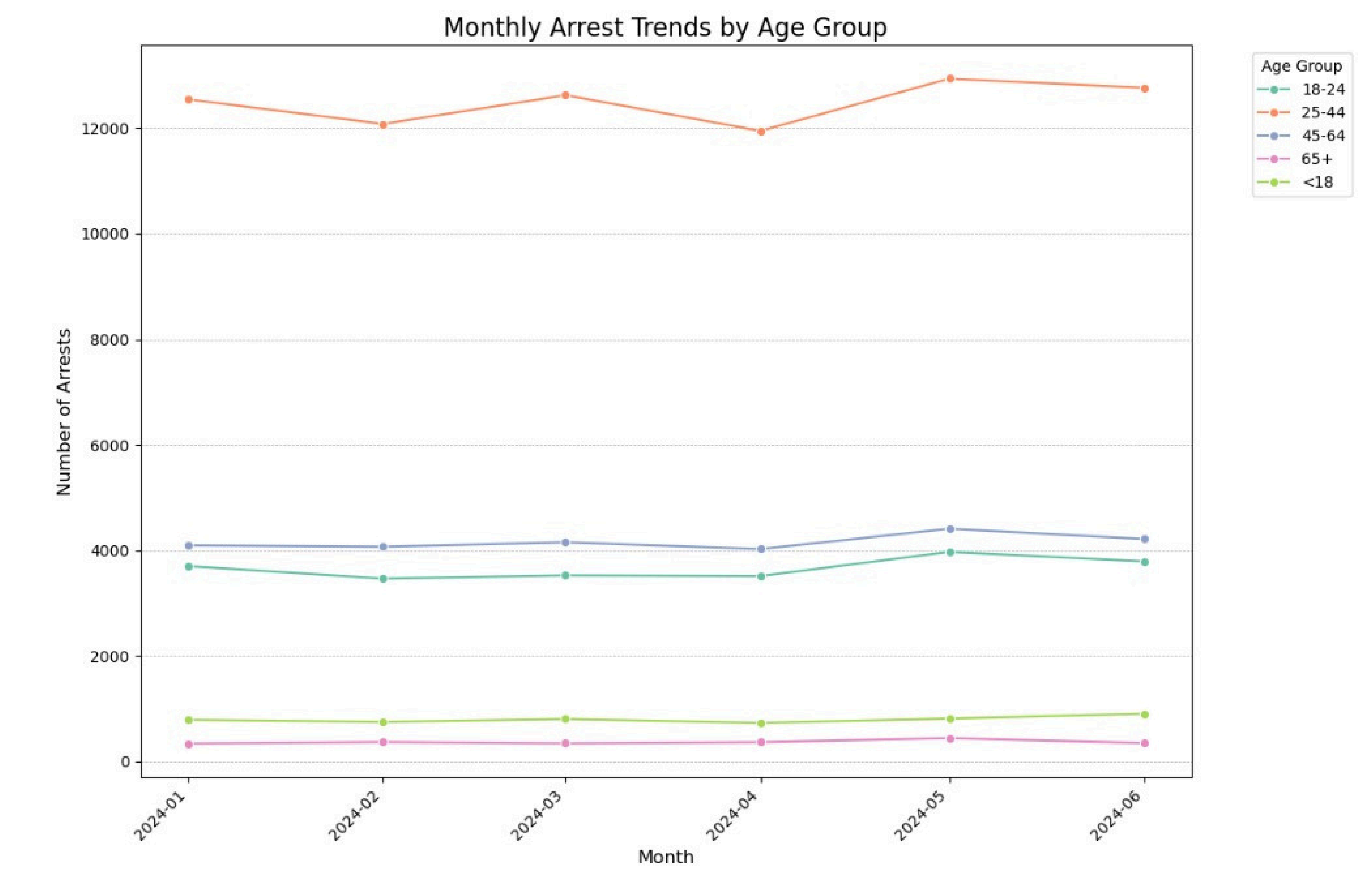
Borough-Race Distribution: The Borough-Race Distribution graph shows how arrests are distributed across different boroughs in New York City by race.



Borough-Specific Insights: Differences in age group proportions suggest varying arrest patterns, potentially influenced by borough-specific population demographics or crime trends.



Trend Analysis: The line plot visualizes fluctuations in arrest counts over time, helping identify periods of increased or decreased arrests.



Temporal Trends: Monthly data reveals seasonal fluctuations or consistent patterns in arrest rates over time.

CONCLUSION

- The analysis of NYPD Arrest Data highlights trends and patterns that can help predict potential criminal activities. This enables proactive measures rather than reactive responses.

- Identifying high-crime areas and understanding demographic factors allows for optimized deployment of resources, such as increased patrolling or community policing, in vulnerable regions.

- The insights gained can guide the implementation of fair and equitable policies to address arrest disparities, ensuring a balanced approach to law enforcement.

- Leveraging the dataset for strategic interventions fosters safer neighborhoods, builds community trust, and ultimately creates a stronger collaboration between law enforcement and the public.


THANK YOU