DAV PROJECT

Pratik Guram RUID 196004669

I explored a variety of datasets related to the demographics, job applications, crime statistics, motor vehicle crashes, in NYC and finalized the Arrests made my NYC police department dataset, as I found it to be informative and the type of data I would love to work with. I found the dataset to be intriguing, and certainly made me curious about the conclusion I would arrive at after working through this dataset. Following is the link to the dataset: <https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc>

The following paper consists of two sections, Exploratory Data Analysis in Part 1 and the Project explanation in Part 2.

Part 1: Exploratory Data Analysis

My dataset has 65,535 rows and 18 columns, where each row is an arrest made by NYPD in NYC.

The following documentation is available for the dataset,

1. Information is accurate as of the date it was queried from the system of record but should be considered a close approximation of current records, due to arrest revisions and updates.

2. Data is available as of the date that technological enhancements to information systems allowed for data capture. Null values appearing frequently in certain fields may be attributed to changes on official department forms where data was previously not collected. Null values may also appear in instances where information was not available or unknown at the time of the report and should be considered as either “Unknown/Not Available/Not Reported.”

3. Arrests which involve multiple charges are classified according to the top charge.

4. Arrests occurring near an intersection are represented by the X coordinate and Y coordinates of the intersection. Arrests occurring anywhere other than at an intersection are geo-located to the middle of the nearest street segment where appropriate.

5. Any attempt to match the approximate location of the incident to an exact address or link to other datasets is not recommended.

6. Many other arrests that were not able to be geo-coded (for example, due to an invalid address) have been located as occurring at the police station house within the precinct of occurrence.

7. Arrests occurring in open areas such as parks or beaches may be geo-coded as occurring on streets or intersections bordering the area.

8. Some arrests do not have a specific law codes due to the voluminous amount of laws. The NYPD database contains all NYS Penal Laws but only includes commonly enforced laws from the VTL, Admin Code, etc. As a result, arrests that have law codes similar to “LOC00000UM” represent arrests associated with violations from Transit Rules and Regulations and other such categories.

9. Arrests occurring on a moving train on transit systems are geo-coded as occurring at the train’s next stop (street intersection).

10. All Arrests occurring within the jurisdiction of the Department of Correction have been geo-coded as occurring on Riker’s Island.

11. X and Y Coordinates are in NAD 1983 State Plane New York Long Island Zone Feet (EPSG 2263).

12. Latitude and Longitude Coordinates are provided in Global Coordinate System WGS 1984 decimal degrees (EPSG 4326).

13. The data represents criminal offenses according to New York State Penal Law definitions, not FBI Uniform Crime Report definitions, and are therefore not comparable to UCR reported crime.

15. Only valid arrests are included in this release. Arrests that were voided when further investigation reveals person did not commit offense or it is determined no offense has been committed are excluded from the data set.

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| ARREST\_KEY | Randomly generated persistent ID for each arrest |
| ARREST\_DATE | Exact date of arrest for the reported event |
| PD\_CD | Three-digit internal classification code (more granular than Key Code) |
| PD\_DESC | Description of internal classification corresponding with PD code (more granular than Offense Description) |
| KY\_CD | Three-digit internal classification code (more general category than PD code) |
| OFNS\_DESC | Description of internal classification corresponding with KY code (more general category than PD description) |
| LAW\_CODE | Law code charges corresponding to the NYS Penal Law, VTL and other various local laws |
| LAW\_CAT\_CD | Level of offense: felony, misdemeanor, violation |
| ARREST\_BORO | Borough of arrest. B(Bronx), S (Staten Island), K(Brooklyn), M(Manhattan), Q(Queens) |
| ARREST\_PRECINCT | Precinct where the arrest occurred |
| JURISDICTION\_CODE | Jurisdiction responsible for arrest. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non-NYPD jurisdictions |
| AGE\_GROUP | Perpetrator’s age within a category |
| PERP\_SEX | Perpetrator’s sex description |
| PERP\_RACE | Perpetrator’s race description |
| X\_COORD\_CD | Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) |
| Y\_COORD\_CD | Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) |
| Latitude | Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |
| Longitude | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |

The variables I am interested in are ARREST\_KEY, ARREST\_DATE, OFNS\_DESC, ARREST\_PRECINCT, ARREST\_BORO, AGE\_GROUP, PERP\_SEX, PERP\_RACE, and the geospatial locations.

I have attached some initial visualizations that are necessary to point out certain trends and outright visible inclinations towards certain variables in the data. These visualizations show the correlations between variables and are simple to decipher.

A screenshot of text

Description automatically generated

This implies the number of arrests comprising of male population far outweigh the female population.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The age group between 25-44 has the highest number of perpetrators in NYC, almost thrice the next age group.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The Brooklyn borough(K) is a highly volatile one with the maximum arrests followed closely by Manhattan borough(B).

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

As it can be clearly seen, the Black population has seen the maximum number of arrests across NYC, and the bar graph, although simple, works exceptionally well in visualizing the vast difference in arrests between the races.

A close up of a map

Description automatically generated

Based on the dataset, a vast number of arrests were made in September and October, with the maximum number of arrests in October comprising of the Black population. American Indian/ Alaskan Native population arrests have been extremely low, albeit this can be attributed to their low population overall. This points to another factor that should be taken into consideration, the ratio of people arrested of any race and the total population of it.

A screenshot of a cell phone

Description automatically generated

The above visualization lets us work with the description of the crimes for which the arrests were made, and by assimilating these details with the other variables, various conclusions can be arrived at. Right now, we can clearly conclude, Assault and related offences were the prime reason for arrests, which begets the question, how safe exactly is NYC? Although this question would need a lot more variables to compute an answer to, it certainly is a start.

The dataset I have selected does not require data cleaning, as it has been curated already.

I intend to combine another dataset that allows me to correlate the arrests with the economic conditions of the area of arrests, to arrive at a better conclusion. This combination of datasets depends on the availability of the same.

Part 2: Project

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The above visualization points to Precinct 41 being the hotspot of arrests in NYC, and that is the Precinct we will be focusing on further.

Precinct 41 is located in The Bronx, notoriously infamous for being unsafe, and rightly so, as per the statistics.

The neighborhood predominantly consists of [Latin Americans](https://en.wikipedia.org/wiki/Latin_Americans) (Longwood has one of the highest concentrations of [Puerto Ricans](https://en.wikipedia.org/wiki/Puerto_Rican_American) in New York City) but includes a small yet diverse mix of races, ethnic groups, religious affiliations, sexual orientations, and national origins. Like most neighborhoods in New York City, the vast majority of households are renter occupied. Almost half of the population lives below the federal poverty line and receives public assistance (AFDC, Home Relief, Supplemental Security Income, and Medicaid). This community was once part of New York's 16th congressional district, the poorest such district in the United States. After [redistricting](https://en.wikipedia.org/wiki/Redistricting), it is now part of [New York's 15th congressional district](https://en.wikipedia.org/wiki/New_York%27s_15th_congressional_district) and no longer the poorest district in the United States. There is significant income diversity on a block by block basis.

The 41st Precinct ranked 67th safest out of 69 patrol areas for per-capita crime in 2010.

In my visualizations, I have tried to answer the questions, WHO, WHAT,WHERE and WHEN, as simply as possible. A lot of conclusions could be arrived at by working on the dataset, and I did decipher a lot of relevant information, but to put into perspective, decided to enlighten Precinct 41 only.

The visualizations are relevant to and can be referred by a variety of potential users. These include potential buyers looking to rent/ buy a property in a certain neighborhood covered by Precinct 41. Safety is the most important aspect of buying/ renting a property in New York, or for that matter, anywhere in the world, and analysis like these educate potential customers.

These can also be used by Precinct 41 itself, to understand the demographics of their arrests and increase patrolling and surveillance in hotspots as located by the analysis. Keeping a tab on the kind of offences taking place right under their nose is a crucial step in eliminating the offences itself. Understanding where certain offences take place on a larger scale, and knowing the severity of the offence, helps in having a management plan ready to mitigate them.

A close up of a map

Description automatically generated

In the above visualization, we encounter our first outlining face, that although the number of arrests were highest for New York City in the month on October as seen in the EDA, Precinct 41 suffered more in September, and that is irrespective of the race. It can be clearly deciphered from the visualization that all arrests irrespective of race maxed out in September. So, to know WHO exactly contributed to these arrests, I have bifurcated the races, and found out that the maximum number of arrests in Precinct 41 throughout the months available in the dataset were the Black race. This concludes the fact that a sizeable number of Black population indulges in illegal activities in Precinct 41 followed by Black Hispanics, and White Hispanics. Another conclusion from the visualization is that the number of arrests of Black population did considerably reduce after September, but still were the maximum each month.

A close up of a map

Description automatically generated

This next visualization is crucial solely based on the fact that it allows us to pinpoint the location of the arrests. To know the areas which are affected the most by certain offences, in this case, the top three offences by arrest numbers, allows us to differentiate between the safer localities and the ones that need extra surveillance and patrolling. Assault 3 and Related offences, Dangerous Drugs and Petit Larceny are the three offences that took place the most in Precinct 41. I have added Offences to the Pages filter, to allow for smoother viewing of the locations of these arrests and reduce clutter. The large the circle, the greater number of arrests at that location, and that certainly says a lot about the neighborhood. To know exactly where certain offences take place allows the police to increase Narcotics patrolling in certain areas and Riot police in the relevant ones. And this answers the question WHAT and WHERE.

Moving on, we’ll follow up with WHEN do certain offences increase in numbers and if certain actions were taken to curb them, their effectiveness in doing so.

A close up of a map

Description automatically generated

In the aforementioned visualization, a 42%drop in Dangerous drugs, and Assault 3 and related offences indicated certain measures taken by the Precinct were effective in curbing them.

Also seen is a 30% increase in Petit Larceny arrests. Nonetheless, September certainly lives on to be the month with the maximum number of arrests for the three top offences by number of arrests. Dangerous drugs and Petit Larceny seem to follow each other inversely, whereas Assault 3 and related offences is a volatile domain. By answering a few of the most important questions about Precinct 41, and with the history that comes with it, Precinct 41 does have a lot to work on. And continued analysis like this would allow it to be a safer place, for good.