Capstone Technical Write-up

Rachalle Ann Olila

[Rachalle.olila@gmail.com](mailto:Rachalle.olila@gmail.com)

General Assembly Data Analytics Bootcamp

DAB 0304

## Introduction

Problem Description/Scenario:

In the dynamic world of New York City real estate, ensuring safety is a high priority. As a data analyst for our real estate company, I've been tasked with exploring NYPD arrest data to help people find the safest neighborhoods to live in NYC.

Objective:

The primary objective is to utilize NYPD arrest data to evaluate and identify the safest boroughs/neighborhoods in New York City. By analyzing arrest frequency and types across different areas, I aim to provide insights that can guide decisions on where to live, work, and invest.

#### Data Collection, Data Preprocessing, & EDA

Primary Source:

The Historic NYPD dataset was originally ranging from the years 2006 to 2023 is taken from the Data.Gov website which was provided by NYC OpenData. “This dataset is a breakdown of every arrest effected in NYC by the NYPD going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents an arrest effected in NYC by the NYPD and includes information about the type of crime, the location and time of enforcement. In addition, information related to suspect demographics is also included” (*NYPD Arrests Data (Historic),*2023).

[Link: NYPD Arrest Data](https://catalog.data.gov/dataset/nypd-arrests-data-historic)

[Footnotes and Dictionary](https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/about_data)

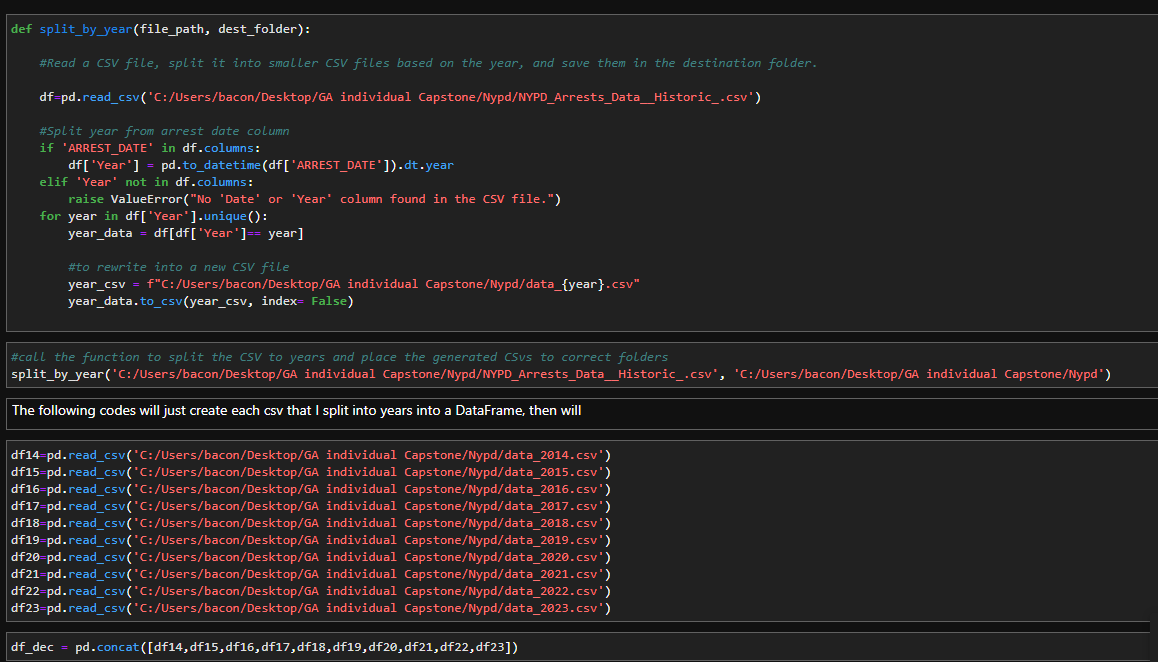
[Geospatial Precinct](https://catalog.data.gov/dataset/police-precincts)

The original dataset contained 5,725,522 rows and 19 columns.

Below is the Data dictionary:

|  |  |
| --- | --- |
| Column name | Column 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) |
| Lon\_Lat | Column with both longitude and latitude in a single line. |

**Data Cleaning**

I imported a .CSV file into Jupyter Labs using pandas. Deciding to focus exclusively on the previous decade (2014-2023), I split the original CSV into individual years to collect the specific data I needed. Afterwards, I concatenated the desired years into my main CSV file. 

\*still was not sure what years I wanted to keep\*

I eliminated some redundant and unnecessary columns which are:

1. PD\_CD
2. PD\_DESC
3. KY\_CD
4. PERP\_RACE
5. JURISDICTION\_CODE
6. Lon\_Lat
7. LAW\_CD

The columns "PD\_CD," "PD\_DESC," and "KY\_CD" are internal codes utilized by the NYPD for their documentation purposes. I do not require "PERP\_RACE" demographic data to assess any safety-related decisions. "JURISDICTION\_CODE" and "Lon\_Lat" columns were removed because redundant information is available in the dataset. The column “LAW\_CD” contains law code charges corresponding to NYS penal law codes.

The new dataset has 2,502,242 rows and 13 columns (additional “YEAR” column).

Null Handling and formatting:

I encountered some null values and errors in the "OFNS\_DESC" and "LAW\_CAT\_CD" columns. Following documentation suggestions, I renamed the nulls as "unknown" or "uncategorized." Additionally, I discovered some unknown categories in the "LAW\_CAT\_CD" column, such as "i" or "9." Despite external research, I could not find any explanations for these categories, so I categorized them as "uncategorized" to maintain data accuracy. I converted the new dataset from uppercase to lowercase, just for housekeeping purposes.

EDA/Visualization/Tableau:

I imported the concatted CSV to Tableau as my main data source. I'll be conducting spatial analysis and crime profiling focused specifically on the boroughs.

The task I needed to complete for the initial EDA.

1. How has the overall arrest count changed over the years?
   1. Create a line graph to show changes over the years. Make note that 2019 to 2021 was “COVID” era. Also include custom shape to accompany graph.
2. What is the distribution of arrests by crime type in New York City, and which crime types are the most common offenses in each borough/neighborhood according to NYPD arrest data?
   1. Which borough has the highest arrest rate?
      1. Which crime?
      2. Type of offenses?
3. How do arrest rates vary across different boroughs?
   1. Are there any high-crime areas or patterns of spatial clustering?
4. How do arrest patterns differ across different time periods, boroughs, or neighborhoods? Are there any notable differences or similarities in arrest trends between various groups or locations? Create a map to see any high crime patterns. Explore any disparities in arrest rates across NYC. Using "Latitude" & "Longitude" to create a map to see high crime patterns.
5. Classify crimes based on their severity (e.g., misdemeanor vs. felony) and analyze the distribution of arrests by crime severity level.
6. How many precincts are there?
   1. Overall arrest in the precinct

I created a bunch of custom icons and images in Adobe Photoshop to accompany the dashboard.

Secondary Source:

For further context, I sought to ascertain the recent median rent by borough. To obtain this information, I sourced data from the StreetEasy website. The StreetEasy Price Indices offer monthly indices tracking changes in resale prices of condos, co-op, townhouses, and single-family homes. Currently, these indices are available for Manhattan, Brooklyn, and Queens. The median rent represents the exact middle asking price among all listed homes available during the specified month, quarter, or year (starting in 2010 to 2023).

[Link: Street Easy](https://streeteasy.com/blog/data-dashboard/[object%20Object]?agg=Total&metric=Inventory&type=Sales&bedrooms=Any%20Bedrooms&property=Any%20Property%20Type&minDate=2010-01-01&maxDate=2024-04-01&area=Flatiron,Brooklyn%20Heights)

Data Cleaning

The original dataset comprised 198 rows and 175 columns. I cleaned the data in Excel, retaining only the median prices from 2014 to 2023. Using formulas, I calculated the average asking price for each year and created new columns indicating the average price for each year by borough.

Null Handling

I encountered numerous null values in the dataset, which I replaced with 'not available' to maintain accuracy. This was necessary as certain years or neighborhoods did not contain price points.

Although this data is not directly connected to the NYPD dataset, I believe it can serve as valuable supporting evidence for my recommendations.

Evaluation

1. The year 2014 recorded the highest arrests at 387,000, followed by a decline during the years 2019-2020, coinciding with the COVID-19 epidemic. Subsequently, from 2021 to the present, the arrest rate began to rise once again.
2. Despite the Bronx being rumored to have the highest crime rates, Brooklyn had the highest reported number of arrests at 27% of arrest in NYC. However, Brooklyn and Manhattan are closely tied in terms of reported arrest numbers. Staten Island has the lowest overall arrest reported.
3. From the heat map I generated, it appears that Manhattan has a relatively even distribution of arrests across its neighborhoods. However, there is an exception in Central Park, where arrests are notably absent. This observation aligns with expectations, as it's unlikely for law enforcement to make arrests in the middle of a field or pond within the park.
4. The distribution of the top 5 offenses remains consistent across all five boroughs: dangerous drugs, assault in the third degree, petit larceny, felony assault, and vehicle traffic laws. Misdemeanors emerge as the predominant crime type across all boroughs.
5. The latest average asking price in 2023 is the highest in Manhattan at $4,538 and the lowest in Staten Island at $2,124.

Recommendation:

Based on the analysis of NYPD arrest data and real estate trends, it's evident that Staten Island has the lowest number of arrests and the lowest rent prices in NYC. However, individuals considering relocating to Staten Island should be aware of the transportation limitations. Staten Island primarily relies on designated buses and the free ferry to access the city, making car ownership or careful planning essential for commuting. Additionally, tolls for bridge crossings can be significant, especially during peak hours. Therefore, prospective residents should consider transportation costs and their commuting needs before deciding to move to Staten Island.