Project: New York City's Crime Data in 2024

Developers:

- T1: Olivia LaCroix
- T2: Swathi Danturi

File: PROPOSAL.md

Date: 11/23/2024

Authorship Information

- Data Source, Pre-processing, Cleaning and Labeling, Questions 3 and 4: `Swathi`
- Data Documentation, Data Privacy, Questions 1 and 2: `Olivia`
- Data Contents: `Olivia` and `Swathi`

Data Source

- Name: `NYPD Arrest Data (Year to Date)`
- Link:
  - `NYPD Arrest Data (Year to Date)`: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc/about_data
  - `Catalog of the dataset`: https://catalog.data.gov/dataset/nypd-arrest-data-year-to-date

Data Documentation

- DatasetName: NYPD Arrest Data (Year to Date)
- DatasetVersion and Number: Dataset Changelog version 15
- Dataset owner: NYC OpenData
- Who can access this dataset? Public as of 11/1/2018
- How can the data be accessed? Online via NYC OpenData, CSV download

Dataset Contents:

- What does each item/record represent?
  - Arrest in New York City by the New York Police Department
- How many items in the dataset?
  - 195K rows/records
  - 19 columns/attributes
- What data is available about each item?
  - There are 19 columns and the columns are as follows:
    - `ARREST_KEY` - randomly generated ID number for each arrest
    - `ARREST_DATE` - exact date of arrest for the reported event
    - `PD_CD` - three digit 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 than PD code)
    - `OFNS_DESC` - description of internal classification corresponding to KY code (more general than PD description)
    - `LAW_CODE` - law code charges corresponding to penal law, and local laws
    - `LAW_CAT_CD` - level of offense: felony(F), misdemeanor(M) or violation(V)
    - `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)
    - `New Georeferenced Column` - randomly generated geocoded column based on Latitude and Longitude fields
- what is the Timeframe of the dataset?
    - Jan 2024 - Sept 2024
- What are the intended purposes for this dataset?
    - To explore the nature of police enforcement activity in New York City, New York.
- How/Who collected the data?
    - The data was collected and provided by the New York Police Department (NYPD).
    - Available on NYD OpenData site and also data.gov
- What demographic groups are identified in the group?
    - Gender, Race, Age are the demographic groups included in the dataset.
- How was the data validated/verified?
    - This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning.

Pre-processing, Cleaning and Labeling

- What pre-processing, cleaning, and/or labeling was done on this dataset?
    - Pre-processing was done to take out the following columns from the original dataset:
        - X_COORD_CD
        - Y_COORD_CD
        - Latitude
        - Longitude
        - New Georeferenced Column

Privacy

- What are potential data confidentiality issues a user of this dataset needs to be aware of? How might a dataset user protect data confidentiality?
    - This data is public and there is no issue with privacy for this dataset.
- Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
    - No, there is no way to identify a person from the data provided.

Questions from each team member

1. Which day of the year had the highest number of arrests and what was the most common race of the perpetrators on that day? `(T1)`
    - Columns used
        - `ARREST_DATE`, `PERP_RACE`
    - Approach
        - Need to filter data by `ARREST_DATE` and count the number of arrests per day in the year.
        - Then filter for arrests that occurred on that day and count which race `PERP_RACE` offended the most on that specific day.
    - How are the fields related?
        - The data relates to each other because a certain number of arrests are occurring every year, and to be able to target enforcement in a certain timeframe of the year, you want to identify the highest crime days. From there, you are able to target further by addressing the question of which race is most likely to commit the crimes during those times.
    - Additional data that might be useful
        - Other data that would provide additional insights for this question would be if there was a field for `TIME`. Time would help narrow down the output even more by allowing law enforcement to identify not only the highest day of the year in arrests but also the most popular time to commit a crime.
2. Which borough has the highest number of arrests for felony offenses and therefore what felony offense was most committed? `(T1)`
    - Columns used
        - `ARREST_BORO`, `LAW_CAT_CD`, `PD_DESC`
    - Approach
        - Need to filter data for `felony` arrests under `LAW_CAT_CD`.
        - Then count the number of arrests in each borough from the `ARREST_BORO` column.
        - Then find the highest number of a specific type of felony offense that is the highest.
    - How are the fields related?
        - This data relates to each other because we want to find out the amount of felony offenses that are taking place in the year of 2024 and out of those felony arrests, we want to know which offense was most common to be able to find patterns in offenses that are occurring.
    - Additional data that might be useful
        - Other data that would provide additional insights for this question would be if the suspect had a `PRIOR ARREST RECORD`. This information gathered would allow police to identify if there are systematic issues with bail reform and recurring crimes with the same suspect.

3. Which age group committed the crime of 'Intoxicated and Impaired Driving' the most and what is the ratio of male to female in the crime? `(T2)`
    - Columns used
        - `PD_DESC`, `AGE_GROUP`, `PERP_SEX`
    - Approach
        - Need to filter the data for `Intoxicated and Impaired Driving` under the `PD_DESC` column.
        - Then count the number of occurrences for each `AGE_GROUP`.
        - Find the count for each `PERP_SEX` (Gender) in the above occurences.
        - Calculate the ratio by dividing both the counts obtained for each `PERP_SEX` from the previous step.
    - How are the fields related?
        - These fields are related because they help us understand the crime rate of `Intoxicated and Impaired Driving` committed by different `AGE_GROUP` and how it varies between males and females, `PERP_SEX`.
    - Additional data that might be useful
        - A new field such as `Age` instead of `Age Group` would allow to determine the crime rate according to a particular age instead of a range.

4. For each borough, what is the average time between arrests for the violence crime type according to the age group?`(T2)`
    - Columns used
        - `ARREST_BORO`, `LAW_CAT_CD`, `AGE_GROUP`, `ARREST_DATE`
    - Approach
        - For each `ARREST_BORO`:
            - For each `AGE_GROUP` where `LAW_CAT_CODE` is `V` get the `ARREST_DATE`s, two at a time.
            - Then calculate the difference between those two dates.
        - Calculate the average time by diving the sum of those differences by the total count of number of time differences.
    - How are the fields related?
        - The fields are connected because they help us see where crime happens, how serious the crime is, the age group of the person arrested, and when the crime happened. This shows the patterns of crime, who commits them, and when and where they happen.
    - Additional data that might be useful
        - Recording `Time` of crime and adding it to the dataset might be helpful to analyze the patterns in the crime if any.