Analysis of Gun Violence in NYC

Rushabh Kenia
rmk9680@nyu.edu
New York Univeristy

Shreyas Umare
ssu4614@nyu.edu
New York University

Yufan Ren
yr975@nyu.edu
New York Univeristy

ABSTRACT

Monitoring the gun violence activities in NYC is the main problem that will be assessed in the project. We will draw out conclusions taking the impact of the pandemic into account as well as the historic data. Also, we are going to take into considerations various other factors such as socio-economic impact an area has on such incidents, jurisdiction, time of occurrence of such events, perpetrator background, etc. to list a few. Another important factor we will be examining is the impact of occurrence of such events before and after the Covid-19 pandemic to answer a few questions like if there is a decrease in the count of such incidents or did it increase? are there any other factors contributing to the growth or decline in gun violence?

KEYWORDS

data sets, Gun violence, NYC, Covid19, median incomes, zip codes

1 INTRODUCTION AND APPROACH

We have acquired data sets containing data related to the gun violence scene in NYC aand also some other data sets that have information about the locations in NYC and it's socio-economic status. We will use all of this data to analyze and get some conclusions/answers for some questions, only a few are listed below:

- What is the most common time when the shooting incidents occurs?
- Which is the safest area in NYC in terms gun violence?
- Are there any factors that led to increase/decrease of such occurences?
- Which jurisdiction has the most occurrences?
- How has the pandemic impacted the whole scenario?
- Is there any correlation between them? etc.

Approach: We are going to follow the standard procedure for making this analysis:

- Data Acquisition
- Data Integration
- Data Cleaning and Wrangling
- Data Analysis
- Information Visualization

Over the course of this project each of the above mentioned concepts and their use in our project will be explained in detail. Let's start with of a few of those below:

2 DATA ACQUISITION

Data acquisition is the process of gathering data. We have kept in mind the four V's of Big Data during this whole process: Volume, Variety, Velocity and Value. All the data sets used in this projects were acquired from different sources:
1. NYPD Shooting Incident Data (Historic) -
https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8
2. NYPD Shooting Incident Data (Year To Date) -
https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Year-To-Date-/5ucz-vwe8
3. Median Incomes -
https://data.cccnewyork.org/data/table/66/median-incomes#66/107/62/a/a

2.1 List and Description of Data sets

- `Data set 1: NYPD Shooting Incident Data (Historic)`: List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This is a breakdown of every shooting incident that occurred in NYC 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 a shooting incident in NYC and includes information about the event, the location and the time of occurrence.
- `Data set 2: NYPD Shooting Incident Data (Year To Date)`: List of every shooting incident that occurred in NYC during the current calendar year. This is a breakdown of every shooting incident that occurred in NYC during the current 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.
- `Data set 3: New York City household income data`: To explore the relationship between gun violence and socioeconomic conditions, geographically tabulated income data is needed. New York City's regional income data from Citizen's Committee for Children (CCC) is composed of sections based on boroughs, neighborhoods and zip codes.

2.2 Challenges involved in acquiring data

- **Purpose and Approach :**
Data set 1 : According to the requirement of the topic of the project, we had to acquire this data from the NYC open data portal. We were able to get the required columns in this data set. The data set contained the columns which were suitable for answering the questions which our title concludes.
Data set 2 : This data is of the current year, which was

needed for the analysis of the analysis between pre and post covid scenario. This data set was also acquired from the NYC open data portal.

- **Challenges faced :**
  Data set 1 and 2 : Initially our plan was to make this analysis on a wider scope of data, but we were not able to find the required data set. We were easily able to find the NYC data set, so we went with a limited scope. We found many data sets for Gun Violence, but those were only available from paid sources, so we acquired this data set which was readily available.
  Data set 3 : After examination of the data, it was determined that dividing the city by zip codes was the best option. Boroughs are too general and lacking granularity since there are good and bad neighborhoods in all boroughs. Neighborhoods on the other hand can have poorly defined borders. In some descriptions of New York City, there are around 300 named neighborhoods in NYC. In the data set only around 50 neighborhoods were listed. Regions described by zip codes are very actively used in everyday life and there is no ambiguity in its definition.

3 TECHNOLOGIES USED

- Jupyter Notebook
- PyCharm
- OpenRefine
- Geocoding engine
- Geoservice platform

4 DATA INTEGRATION

Data integration is the practice of consolidating data from disparate sources into a single data set with the ultimate goal of making analysis of data across the spectrum of subjects and structure types, and to meet the information needs of all applications and business processes.

- `Zip codes and reverse geocoding`: To derive the relationship between gun violence and area income, data sets with gun violence incidents and regional income must be joined together. The incidents data set uses geographical coordinates to describe location instead of zip codes. To obtain zip codes for locations involved in these incidents, reverse geocoding was used. We explored two different ways to obtain this data, both involving web services. Two approaches used:
  - The first method was submitting bulk jobs to an online geocoding engine. A file with index numbers, latitude, longitude and state as required fields (2020_NYC_Shooting_Coordinates.csv) was prepared using getLatLon.py and submited to Texas AM University's geoservice platform at **https://geoservices.tamu.edu/Services/ReverseGeocoding**
    Result file was then downloaded from the platform (2020_result.csv).
  - The file with historical gun violence was much larger than the 2020 data. It contains more than 21,000 entries. To process this amount of data in bulk, a fee could not be avoided. The free method we employed was the Geopy library. 2006-19_NYC_Shooting_Coordinates.csv was originally prepared for a bulk process using getLatLon.py. It was instead fed into get_zipcode.py to produce a few fragmented files (reason explained in the challenges section). These files were joined together using the "cat" command into historic_result.csv.
  - This file consisting of the zip codes was then joined with the data sets consisting of gun violence data.

4.1 Challenges involved in integrating data

The purpose of utilizing the Geopy library was to avoid paying a fee. The web services backing the Geopy library (Nominatim) do not like users submitting bulk requests. The service tried to disconnect users with large amounts of requests. Although there was a built in function inside the library limiting the rate of requests, our program was still getting interruption and the process must be restated at the recorded where interruption occurred. Additionally, the unreliable nature of home internet connection also caused disruption in data processing. After disconnecting from the services a few times, we modified our code to use random sleeping time between each request. The goal was to trick the web service into believing that we were not processing data in bulk. We also implemented our code to automatically retry after server response timed out. After several modifications, the algorithm was able to run without interruption. It took a total of 10 hours to obtain all the zip codes by using reverse geocoding

5 DATA CLEANING

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

- `Zip codes and reverse geocoding`: Both 2020 and historic data with zip codes were processed using Openrefine. The process included deleting unused data and fixing bad data from reverse geocoding. Around 20 bad return data were verified with Google Maps and changed manually in Openrefine. The steps are described in openrefine_zipcode.json. 2020_result.csv was then joined with 2020 shooting incidents data to produce 2020_incident_with_zipcode.csv. historic_result.csv was joined together with historic data to produce historic_incident_with_zipcode.csv. Join_zipcode.py was used for the joining operations.
- **Data sets 1 and 2 :**
  After successfully, acquiring the data set, we found that the it had irrelevant and unnecessary data, which was not required for the analysis. This process is crucial and emphasized because wrong data can drive an analysis to wrong decisions, conclusions, and poor analysis, especially if the huge quantities of big data are into the picture.
  In this particular case, we removed :
  'LOCATION_DESC', 'PRECINCT', 'STATISTICAL_MURDER_FLAG', 'PERP_AGE_GROUP', 'VIC_SEX', 'VIC_RACE', 'X_COORD_CD',

'Y_COORD_CD', 'Latitude', 'Longitude', 'Lon_Lat', 'New Georeferenced Column' as these columns, which were totally irrelevant to our analysis.
Then we cleaned all the rows which contained NaN values for which we used HeatMap to detect the NaN value, which is in fact a very good way to check for NaN and inconsistent values.
Next, we removed all the duplicate entries and made the data consistent.
Dropped all the entries in which the column PERP_SEX contains: UNKNOWN. This made all the entries clean and consistent.

- **Data sets 3 :**
  The aforementioned database was already cleaned and integrated by CCC. No obvious error was found in the data.

5.1 Challenges involved in cleaning data

Although the data obtained from reverse geocoding was largely good, there were still a few inconsistencies. For example, some zip codes were returned "100". The reason for such error was not apparent. After manually checking the lat-lon points using Google Map, it was determined that all the "100" zip codes were in fact "10023". It was fortunate that there were only a few such occurrences. It would not be possible to check them hand by hand if the numbers were larger. This mistake was not found in the bulk job submitted to the Texas AM geoservices. It is possible that error did not occur because the bulk job was much smaller than the batch processed with our own algorithm. Another possibility is that commercial geoservices are more reliable than the Geopy library. A possible experiment for the future would be to filter out the records with incorrect zip code format and resubmit them to a service where small batches are free.
Some basic challenges faced while cleaning:
A lot of rows had UNKNOWN written in its column, this we had to overcome by detecting and removing that entry. We also faced a challenge as we encountered NaN entries, we detected it by using a heatmap and then removed all the NaN values. We also had duplicate entries in the dataset, by using drop_duplicates we removed it and overcame the issue.

6 STEPS TO REPRODUCE :

(Please refer to the github link available at the end of this paper for code and datasets:)
1. To derive the relationship between gun violence and area income, datasets with gun violence incidents and regional income must be joined together.
2. The incidents dataset uses geographical coordinates to describe location instead of zip codes.
3. To obtain zip codes for locations involved in these incidents, reverse geocoding was used.
4. We explored two different ways to obtain this data, both involving web services. The first method was submitting bulk jobs to an online geocoding engine.
4a. A file with index numbers, latitude, longitude and state as required fields (2020_NYC_Shooting_Coordinates.csv) was prepared using getLatLon.py and submited to Texas AM University's geoservice platform at
https://geoservices.tamu.edu/Services/ReverseGeocoding.
Result file was then downloaded from the platform (2020_result.csv) available at the following path: Github->datasets->Integrated and Intermediate Data
4b. The file with historical gun violence was much larger than the 2020 data. It contains more than 21,000 entries. To process this amount of data in bulk, a fee could not be avoided. The free method we employed was the Geopy library.
2006-19_NYC_Shooting_Coordinates.csv was originally prepared for a bulk process using getLatLon.py. It was instead fed into get_zipcode.py to produce a few fragmented files (reason explained in the challenges section). These files were joined together using the "cat" command into historic_result.csv available at the following path : Github->datasets->Integrated and Intermediate Data
5. Both 2020 and historic data with zip codes were processed using Openrefine. The process included deleting unused data and fixing bad data from reverse geocoding. Around 20 bad return data were verified with Google Maps and changed manually in Openrefine. The steps are described in openrefine_zipcode.json available at the following path: Github->code
6. 2020_result.csv was then joined with 2020 shooting incidents data to produce 2020_incident_with_zipcode.csv available at the following path : Github->datasets->Integrated and Intermediate Data
7. historic_result.csv was joined together with historic data to produce historic_incident_with_zipcode.csv available at the following path : Github->datasets->Integrated and Intermediate Data. Join_zipcode.py was used for the joining operations and is available at Github->code.
8. After acquiring the data set, we are cleaning the NYPD_Shooting_Incident_Data__Historic_.csv which is present in Github->Datasets->Raw Data dataset by using dataCleaning_Historic.ipynb this file which is present in the GitHub->code folder.
9. Next, we are cleaning the historic_incident_with_zipcode2020.csv which is present in Github->Datasets->Raw Data dataset by using dataCleaning_Historic_2020.ipynb this file which is present in the GitHub->code folder.
10. After the integration of historic_result.csv and NYPD_Shooting_Incident_Data__Historic_.csv, we were able to obtain historic_incident_with_zipcode.csv which is present in Github->Datasets->Raw Data. We cleaned the dataset using dataCleaning_HistoricWithZip.ipynb which is present in the GitHub-> code folder.
11. The same process is followed for the historic_incident_with_zipcode2020.csv file, which is present in Github->Datasets->Raw Data, we used the dataCleaning_HistoricWithZip_2020.ipynb file to clean the dataset which is present in the GitHub->code folder.

NOTE: Please update the .read path and .write path.

7 GITHUB REPOSITORY :

All the data sets, python files(codes) and Jupyter notebooks used to integrate, transform, clean and wrangle the data can be found at the link below:
https://github.com/rushabhkenia/Big-Data-Project