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? We propose to study the occurrence and trends of such activities over the period of 2006 to 2020 in New York City but on a more detailed level to understand what is the most common time or location of such activities and which people are involved keeping into mind a variety of factors such as socio-economic impact, income groups, age groups, gender, race etc.

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. some other data sets that have information about the locations in NYC and it's socio-economic status, labor force participation data and arrest data by NYPD. 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 occur?
- Which is the safest area in NYC in terms gun violence?
- Are there any factors that led to increase/decrease of such occurrences?
- 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
4. Historical NYC labor force participation data -
https://dol.ny.gov/labor-statistics-new-york-city-region#:~:text=The%20city's%20seasonally%20adjusted%20unemployment,8.5%20percent%20in%20March%202021.
5. NYC geojson for drawing maps -
https://data.beta.nyc/dataset/nyc-zip-code-tabulation-areas

2.1 List and Description of Data sets

We have listed down a description of few of the important datasets required for our project:

- `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.
  Data set 4 and 5: Challenge we faced was the correctness and inconsistency in geological data. Our original idea was to use neighborhood names as local units on our maps. However, we soon found that both the definition of 'neighborhood' and its boundary are subject to interpretations. The definition can change according to different sources or over time. Additionally, many of the datasets we found are already cleaned and tabulated. Datas that are aggregated according to neighborhoods are often incomplete. According to some sources, there are well over 200 distinct neighborhoods in New York City. Data sets that divide NYC into neighborhoods often have much less categories. It was impossible to verify whether there is an unified version of such division. By the end we decided to use postal code areas to represent subareas. It was much more consistent compared to neighborhood divisions. Additionally, there are different versions of geojson files of New York city on the web with different numbers of postal codes. We used the version we believed to most accurately reflect the number of postal areas in NYC.

3 TECHNOLOGIES USED

- Jupyter Notebook
- Google Colab
- 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 A&M 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 inconsistent but necessary data, which would be required for the analysis. This process is crucial and emphasized because wrong or inconsistent data can drive an analysis to wrong decisions, conclusions, and poor analysis, especially if the huge quantities of big data are into the picture.
  Thus, first we checked for any duplicate entries in the data set. We removed all the duplicate entries and made the data consistent.
  Next, we checked for the missing values in the data set. To our surprise a lot of data was missing but cannot be ignored as in such cases sometimes the perpretators are not found or victim cannot be identified and hence, After finding missing values in column PERP_RACE, we normalized the values by using random function. This helped us make the data more consistent and accurate.

  Similarly, we normalized the values for the below columns too: PERP_AGE_GROUP,PERP_SEX,LOCATION_DESC, JURISDICTION_CODE.

- **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, initially we thought this we had to overcome by detecting and removing that entries but after a detailed inspection of data we found out that these records are of importance and can be used as is it as for such a project where we are trying to analyse the shooting activities there is a huge scope of people and their records not being reported and hence the unknown values. We also faced a challenge as we encountered NaN entries, we detected it by using a heatmap and then normalized 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_YTD.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_Historic.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_YTD.ipynb file to clean the dataset which is present in the GitHub->code folder. NOTE: Please update the .read .write path.

7 DATA ANALYSIS

Data analysis is a process of inspecting, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Once the datasets are cleaned, it can then be analyzed. Analysts may apply a variety of techniques, referred to as exploratory data analysis, to begin understanding the messages contained within the obtained data. The process of data exploration may result in additional data cleaning or additional requests for data; thus, the initialization of the iterative phases may be required. Descriptive statistics, such as, the average or median, can be generated to aid in understanding the data. Data visualization is also a technique used, in which the analyst is able to examine the data in a graphical format in order to obtain additional insights, regarding the messages within the data. The main strategy for analysis and visualization was to compare the count of such occurrences with respect to various locations in New York City area over the years keeping into account various factors such as age, sex, race etc. of an individual involved. We did our analysis using Jupyter and Google Colab notebooks by creating various kinds of visualizations. For this we used Numpy, Pandas, Geopandas, Matplotlib, Plotly and Seaborn.
We started our data analysis by using various visualizations such as bar graph, maps , pie-chart, line-chart, to give a better understanding of the data which we used. We compared the analysis of historic data and the Covid period data.

7.1 Location based Analysis

New York City over the years

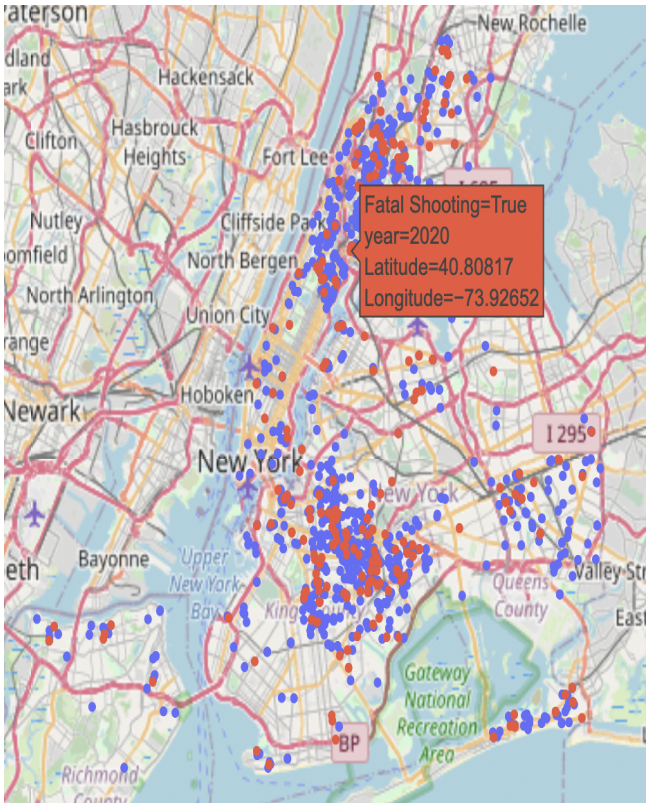

Figure 1: Fatal Shooting Incidents subjected on Interactive Map for NYC

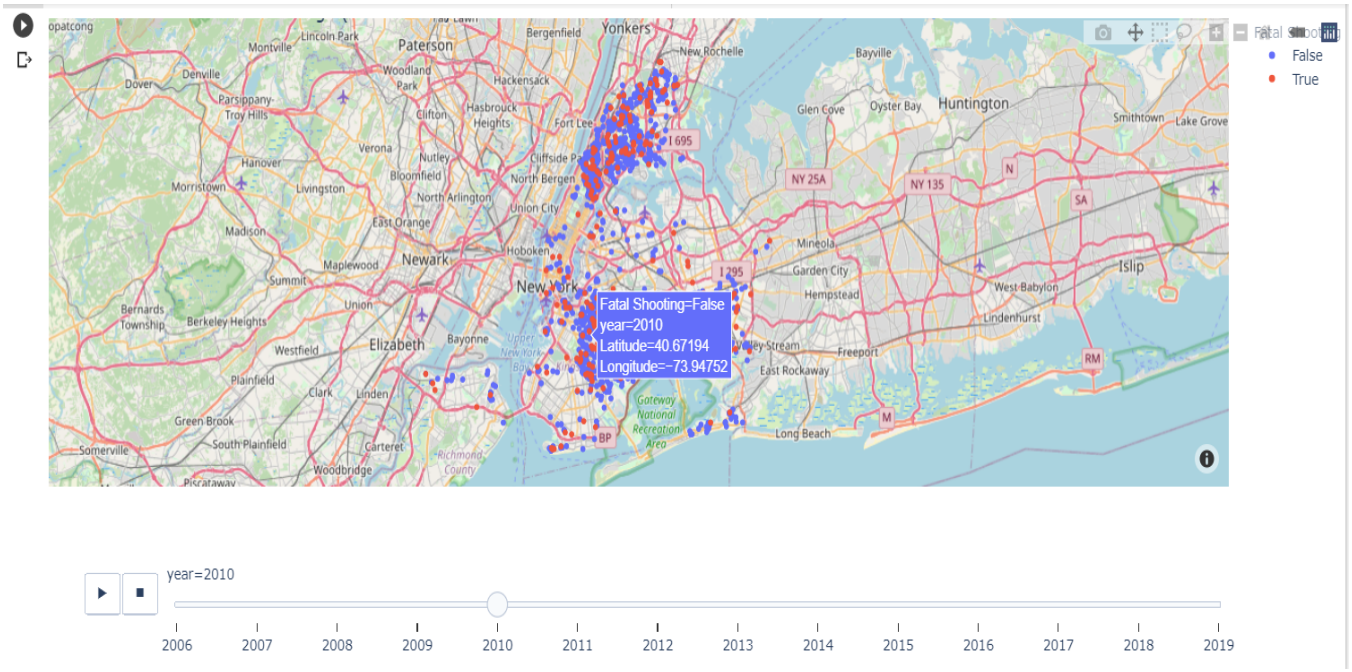

Figure 2: Total Shooting Incidents subjected on Interactive Map for NYC

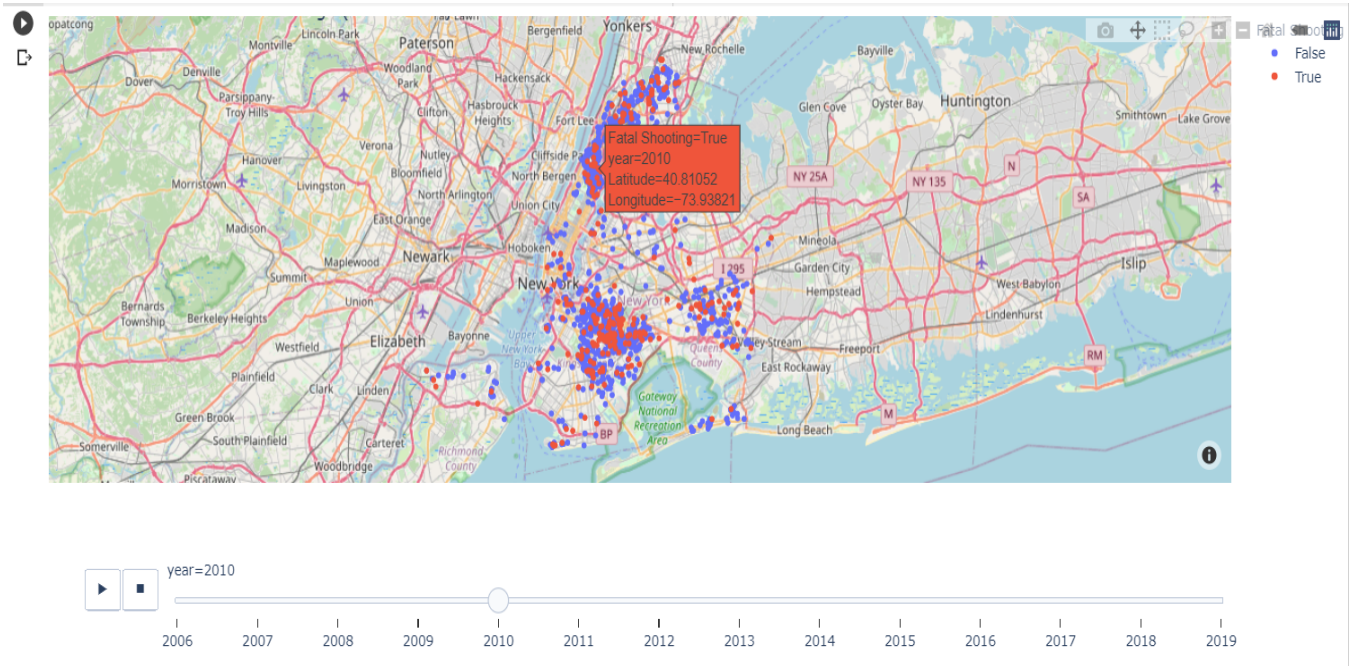

Figure 3: Fatal Shooting Incidents subjected on Interactive Map for NYC

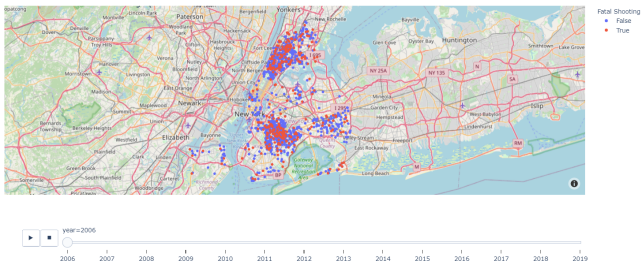

Figure 4: Shooting Incidents subjected on Interactive Map for 2006

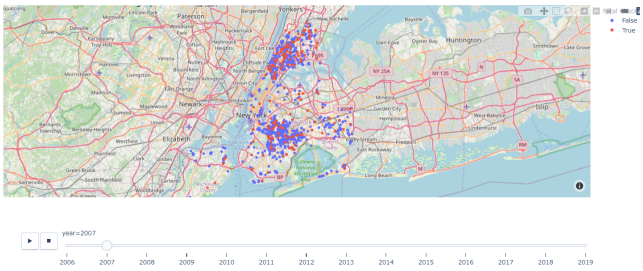

Figure 5: Shooting Incidents subjected on Interactive Map for 2007

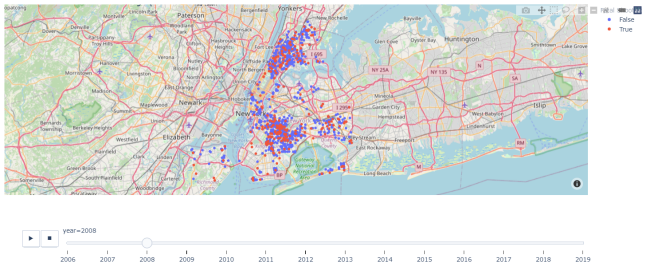

Figure 6: Shooting Incidents subjected on Interactive Map for 2008

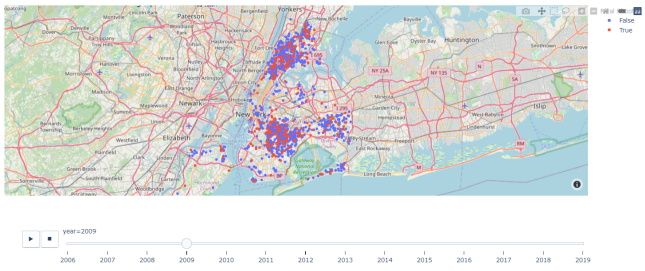

Figure 7: Shooting Incidents subjected on Interactive Map for 2009

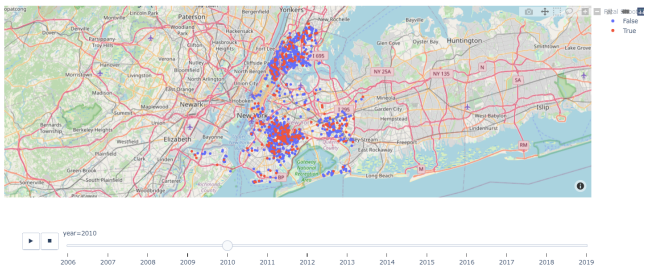

Figure 8: Shooting Incidents subjected on Interactive Map for 2010

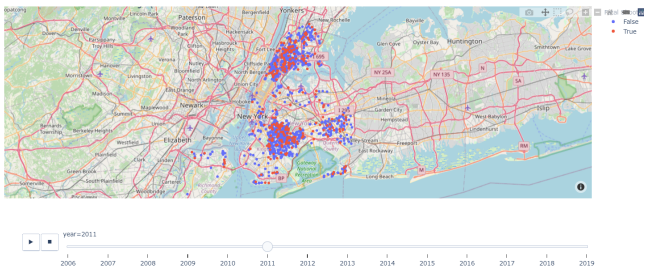

Figure 9: Shooting Incidents subjected on Interactive Map for 2011

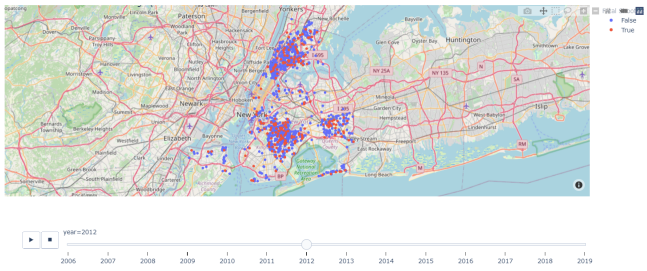

Figure 10: Shooting Incidents subjected on Interactive Map for 2012

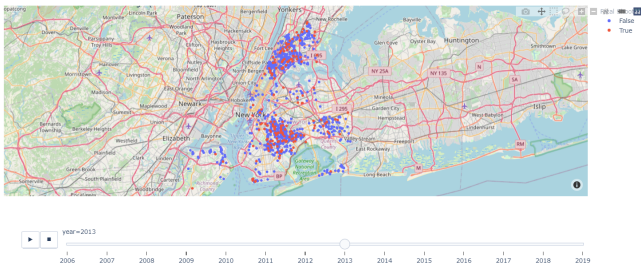

Figure 11: Shooting Incidents subjected on Interactive Map for 2013

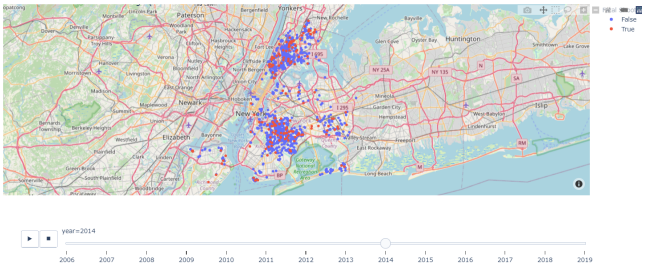

Figure 12: Shooting Incidents subjected on Interactive Map for 2014

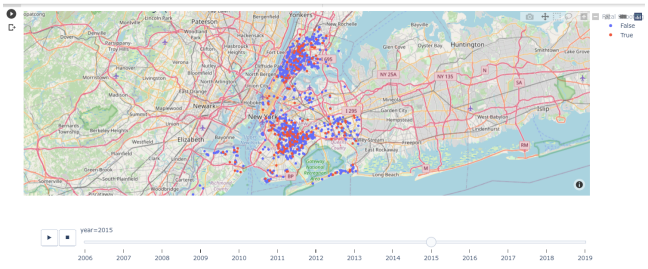

Figure 13: Shooting Incidents subjected on Interactive Map for 2015

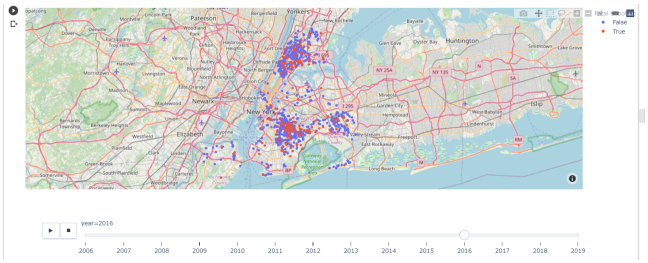

Figure 14: Shooting Incidents subjected on Interactive Map for 2016

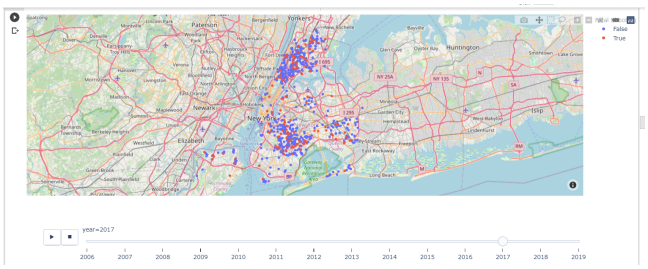

Figure 15: Shooting Incidents subjected on Interactive Map for 2017

The above seen visualizations from Fig 1-17 display the total shooting activities that took place for the period of 2006 to 2020. We can see that over all the years these shooting activities follow a similar kind of pattern and although the count of occurrences of

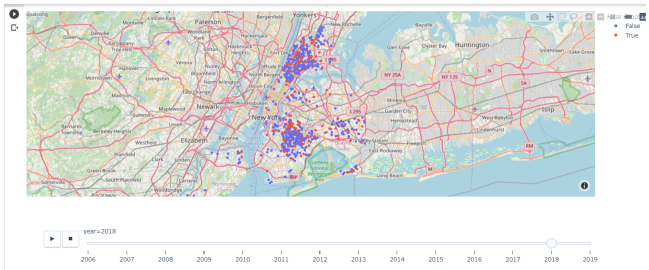

Figure 16: Shooting Incidents subjected on Interactive Map for 2018

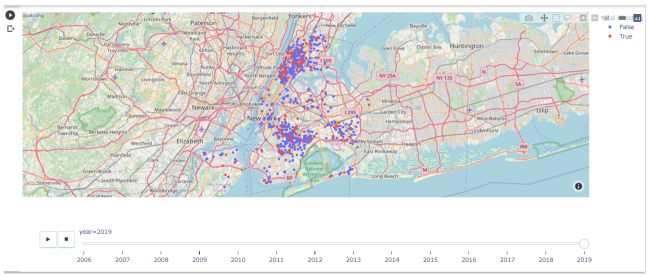

Figure 17: Shooting Incidents subjected on Interactive Map for 2019

such activities is different the maps still display a similarity. Another important thing found was that such activities are concentrated in some specific areas of the city. If seen carefully we can make an analysis is that some specific neighbourhoods are contributing to most of the occurrences. One interesting thing to observe is that even during 2020 i.e. while the pandemic was going on and lockdown was imposed, we can still see a similar pattern that was seen during all the previous years although the movement was restricted because of pandemic.

The blue dots in the map represent the non-fatal shooting incidents where as the red dots indicate all the fatal shooting incidents which were characterized based on the the Statistical_Murder_Flag. Thus, the total such incidents can be summarized by all the dots in the map.

To get a proper analysis we did a detailed study and a detailed location based analysis on the shooting activities in Boroughs, Precincts, Jurisdiction codes and housing arrangements.

Figure 18: Boroughwise Shooting Incidents from 2006-2019

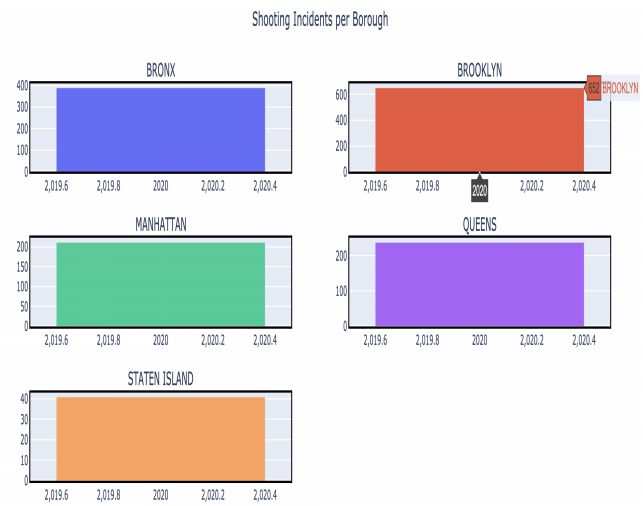

Figure 20: Boroughwise Shooting Incidents for 2020

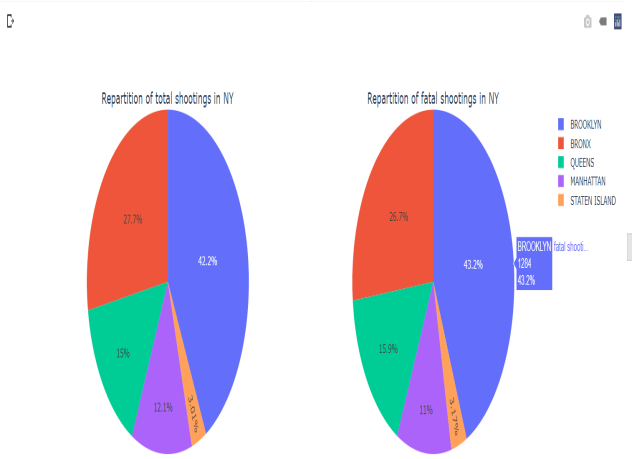

Figure 19: Boroughwise ratio of Shooting Incidents from 2006-2019

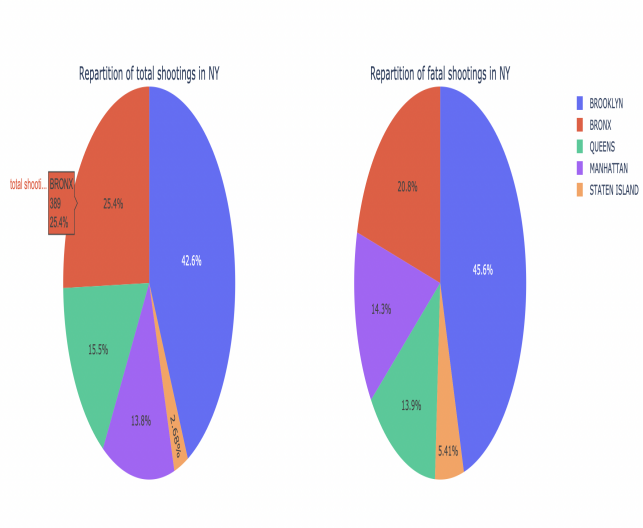

Figure 21: Boroughwise ratio of Shooting Incidents for 2020

Based on the analysis done borough wise, as shown in Fig 18 - 21, most of the shooting incidents are concentrated in two of the boroughs namely Brooklyn and Bronx. Statistically, the total count of these boroughs is almost 70 percent of the total count. The interactive map tells us that even within the boroughs, such activities are concentrated in certain areas which show a significant number of shootings than the rest which proves that even within a borough that shows very high concentration of such incidents there are many areas that are very safe. The occurrence of such gun violence activities has seen a steady decline in Brooklyn and Bronx after 2012 but surged back in 2020.Reported incidents have been the lowest during 2017 -2019. Another important finding was that, the above statement does not hold true for Manhattan and Staten island boroughs as here the shooting incidents have happened at random locations.The number of shooting incidents have significantly reduced in each of the boroughs except for Manhattan and Staten Island. Although in 2020 it should be noted that the cases of fatal shootings was more in Manhattan than in Queens as opposed to the previous year trends.

List of Boroughs - Most Dangerous to Safest.

- Brooklyn
- Bronx
- Queens
- Manhattan
- Staten Island

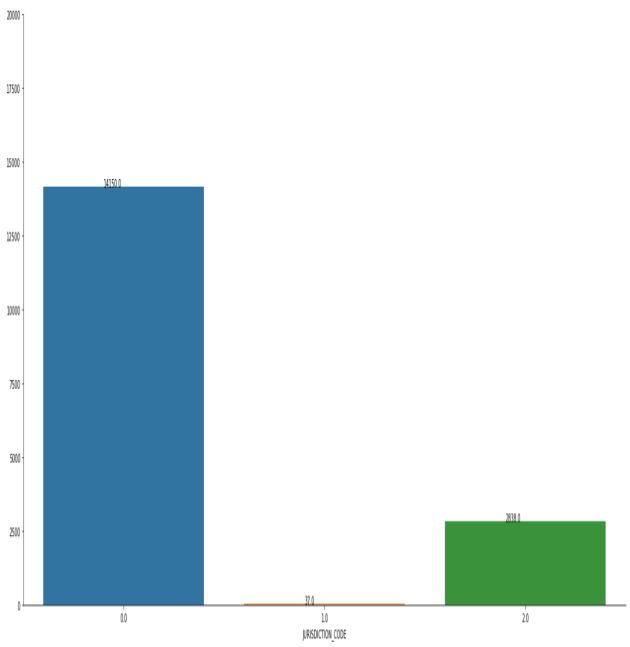

Figure 22: Jurisdiction - Shooting Incidents from 2006-2019

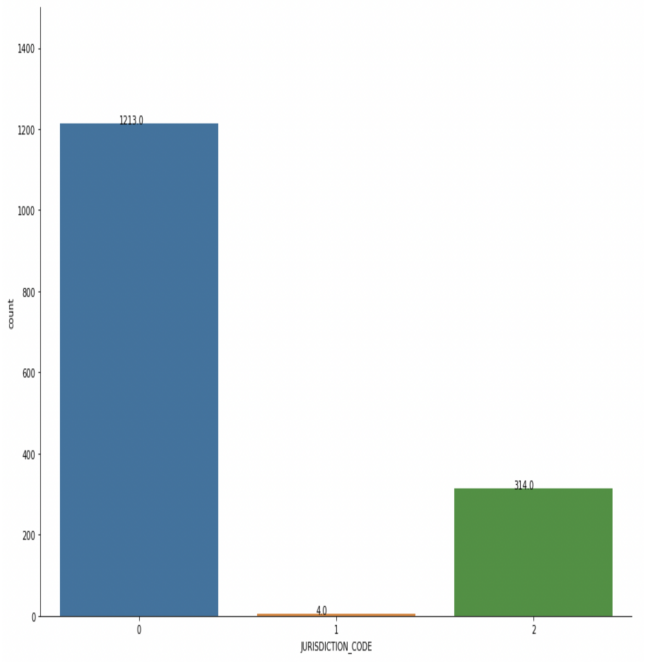

Figure 24: Jurisdiction - Shooting Incidents in 2020

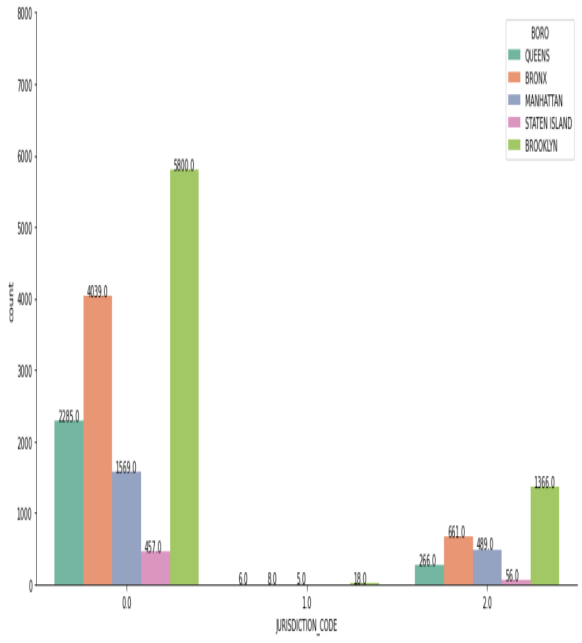

Figure 23: Precincts - Shooting Incidents from 2006-2019

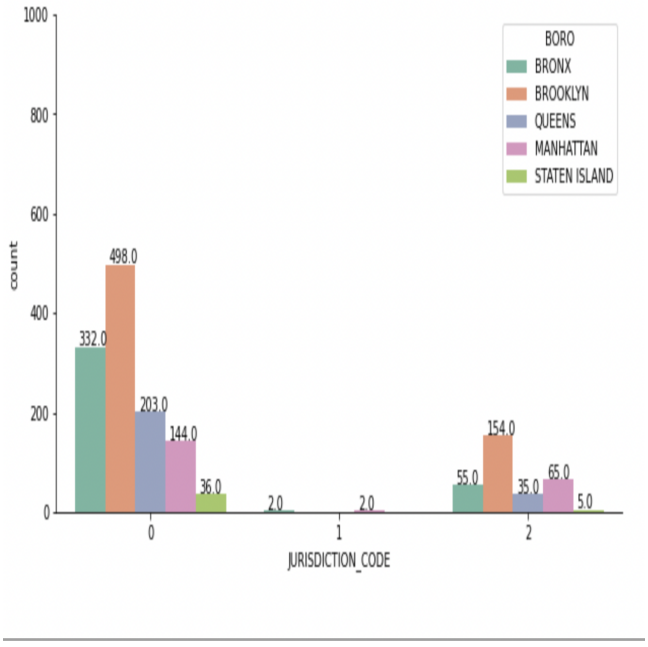

Figure 25: Jurisdiction - Shooting Incidents in 2020

Based on analysis done on Jurisdiction codes, precincts and housing arrangements as shown in Fig 23 -29, we found that out of all the shooting activities within a reported place – Multi-dwelling houses are more prone to shooting incidents and commercial buildings see low occurrences of such incidences .

But this is not true in the case of Fatal shooting incidents as they are unpredictable and all type of shootings have lowered in the apartments over the years.

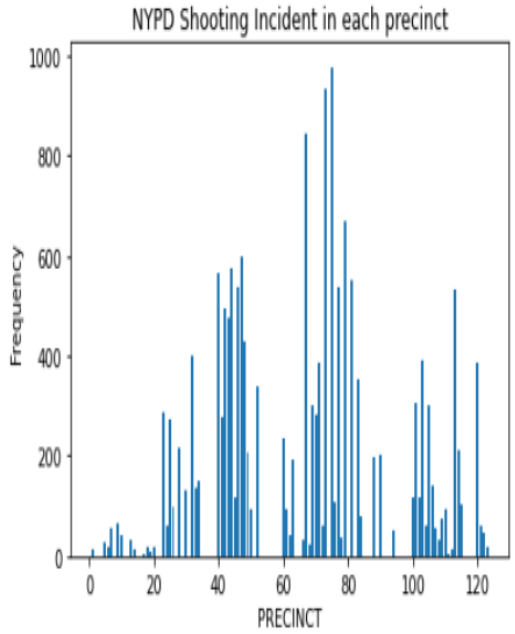

Figure 26: Precinct - Shooting Incidents from 2006-2019

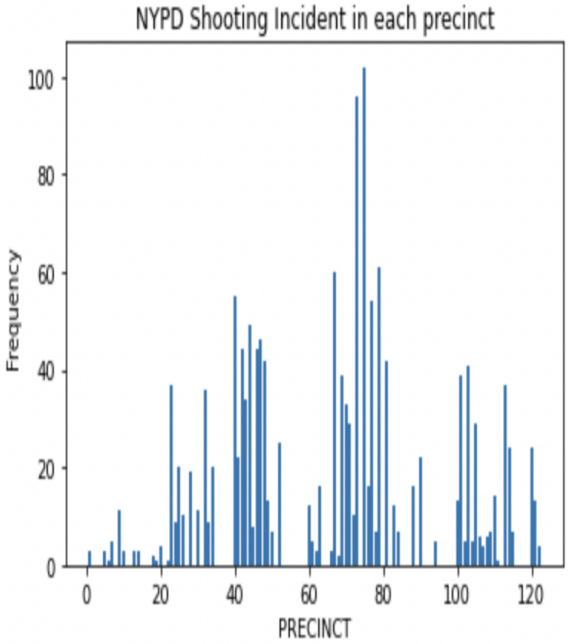

Figure 28: Precincts - Shooting Incidents in 2020

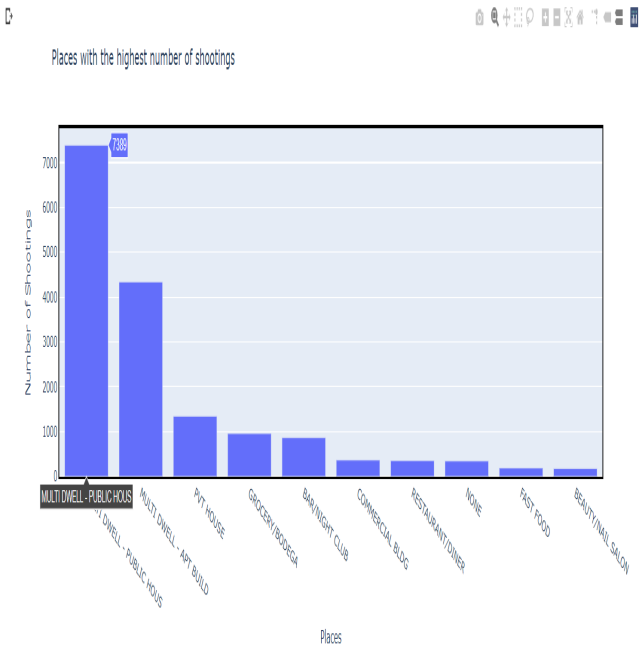

Figure 27: Household Highest- Shooting Incidents from 2006 -2019

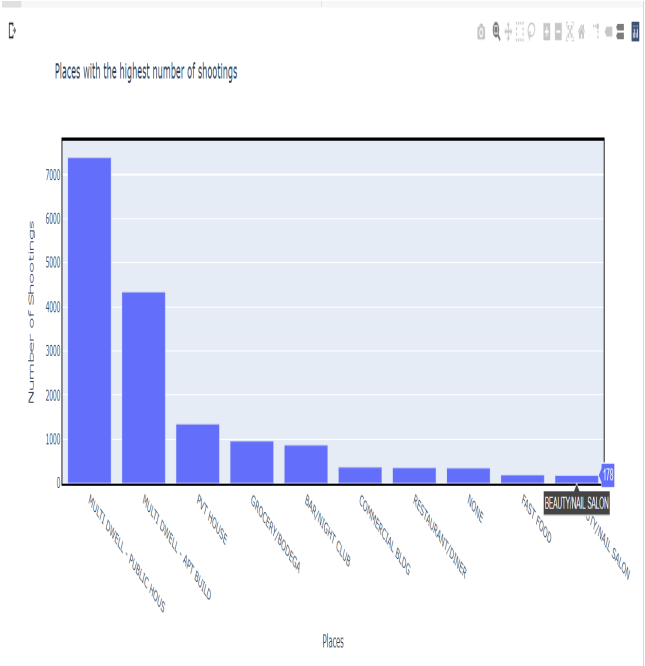

Figure 29: Household Lowest- Shooting Incidents from 2006 -2019

Also there has been no significant impact of the pandemic over shooting incidents occurring at certain places as the trend is similar over the years.

Highest number of cases are reported in precinct are 40-80 i.e. Brooklyn. Most incidents were recorded in jurisdiction 0 i.e. Patrol in Brooklyn followed by Bronx.

Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions. Thus we can see that housing has a lower occurrences of such cases than Patrol which is where most such activities can be seen followed Transit which has the lowest registered activities among the three.

To answer our question as to which was the safest area the above analysis were still insufficient, hence to get the exact scenario for location based shooting incidents, after looking at the data we started the analysis by finding the safest area in the city. We divided NYC into postal code areas and drew the map using geojson data, then we coded the number of fire-arms related crimes in all historical data to the area using color intensity.

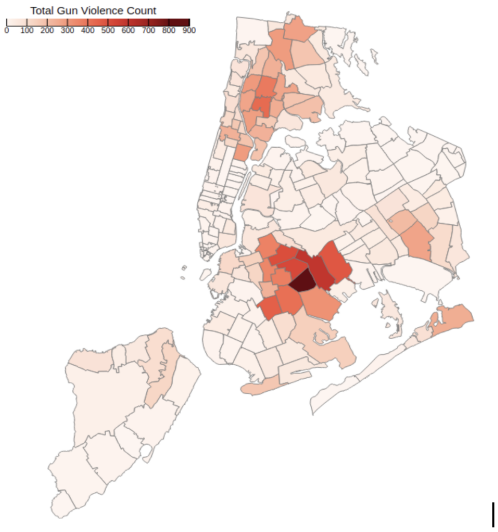

Figure 30: Gun Violence subjected on Map

Most of the firearms related crimes occur in northeastern Brooklyn boarding Queens (around East Flatbush) and Bronx.

Top 10 most dangerous postal code areas are:

- 11212
- 11207
- 11208
- 11233
- 11221
- 11226
- 10456
- 10457
- 11206
- 10466

This map displayed in Fig 30 showed us which is the safest area in the City. Safer areas including most of , Queens, Staten Island and eastern/southern parts of Brooklyn.

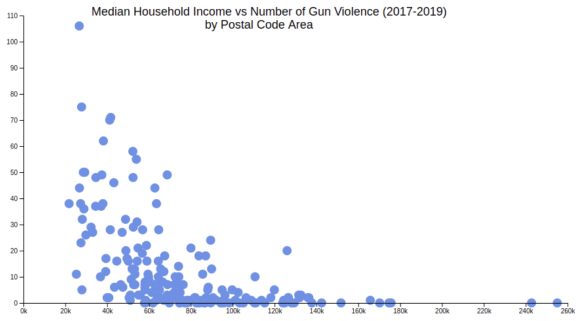

Figure 31: Median household income vs Gun Violence (2017-2019)

7.2 Location based Analysis w.r.t Income

Next, we looked into the income of people and tried to figure out is there any relation between the income of the people in that area and the number of shooting incidents. We found data of median household income divided by postal code areas for the years of 2017-2019. We calculated the average gun crimes occurred in 2017-2019 per area and plotted against median household income of the area.

We can see a clear inverse relationship between income and gun crime counts. High end neighborhoods rarely have any gun crimes occur. It may be related to the ability of families in these communities to procure security systems or hire management personnel. Such communities usually have a much lower population density, which lower the likelihood of crimes occurring.

The areas with low incomes are more varied however. All areas with high gun crime rates are exclusively low income neighborhoods. On the other hand, there are some low income neighborhoods that are relatively safe. There must be other factors at play. These factors may be related to immigrant composition, ethnicity or unemployment rate.

7.3 Location based Analysis on Arrest Data

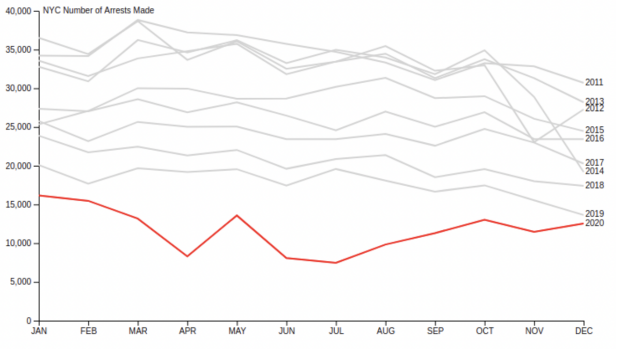

Figure 32: Year wise trend of number of arrests by NYPD

After analysing the income stats with the shooting incidents, we tried to figure out, what factors directly affect the sudden increase in fire-arms related crimes in NYC in 2020?

In May 2020, the death of George Floyd caused rage and a series of protests across the country. Many demanded punishment of the perpetrating officers, a reform of the police department and even defund the police. Some believe the police force became more passive amid the criticism and their activities slowed down. We compared data of arrests made by the NYC department for the past 10 year and found that there are indeed usual patterns in police activity.

Number of arrests per month for 2020 is plotted together with data from 2011-2019. Year 2020's data is highlighted in red.(Figure 3)

In the past 10 year the overall trend has been a slow decrease of arrests made. In any given year the number of arrests per month gradually decreased. In 2020 a different trend emerged. The number of arrests took a sudden dip in April, which may be correlated to the covid shutdown. With everyone staying home, it is sensible that crime greatly decreased. However in May the arrests picked up again, which may be related to the number of crimes going back up. In late May the George Floyed incident happened. Although the lockdown had not ended officially, people were less afraid of the pandemic after a period of isolation. A need for people to go outdoors might have caused crimes to pick back up. At the same time police arrests made another dip in June, which can only be explained by the police department's own decision to reduce activity. The arrest numbers then recovered very slowly for the rest of the year.

7.4 Location based Analysis w.r.t. economy and labor participation

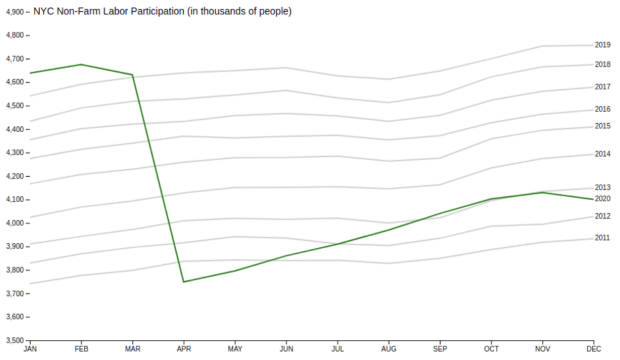

Figure 33: NYC Non-Farm Labor participation

The second factor may be related to the economy. From the previous section we established the relationship between economic condition and likelihood of gun crimes. It is natural to deduce that the economy has a big part to play in our analysis. Because of the covid lockdown, almost all industries were negatively impacted. Up until today, economic stimuli are still being handed out to mitigate the impact. The total number of non-farm labor participation in

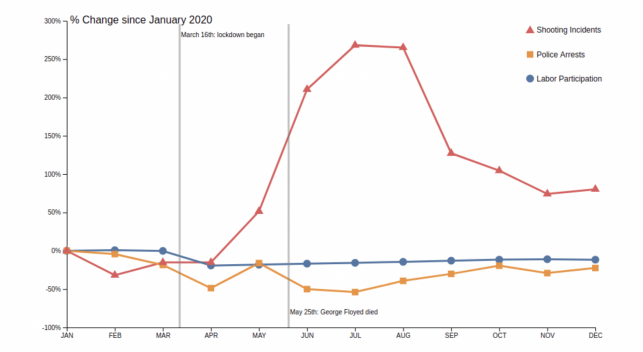

Figure 34: Change since January 2020

NYC per month is plotted together with the data from 2011-2019. Year 2020's data is highlighted in green.(Figure 34)

In the nine years before 2020 the number of labor participation is very consistent. It smoothly increases throughout the years. In April of 2020 it took a nosedive from 4.6 million to 3.7 million people. Events of this magnitude have not been seen in decades. With these many people out of job and others' pay or hours cut, it is possible that many become desperate and turn to crime.

Now we put these factors together with the number of gun crimes in 2020.

Number of gun crimes, labor participation and number of police arrests are plotted on the same time scale.(Figure 5) The y-axis scale is normalized to percent changes from January 2020. The covid outbreak and the ensuing lockdown likely caused a sharp downturn in the economy. Although not the most obvious element on the chart, the down tick of labor participants from March to April represents almost 1 million people losing their jobs. The George Floyed incident in May caused the police department in many places in the country to reduce their activity. Police arrest counts fell to half compared to beginning of the year. This chain of events contributed to the sharp increase of gun crime to almost 400% during the summer compared to beginning of 2020. Later in the year, with the city opening up in June, the economy started churning again and the police department ramping up their effort in law enforcement, the number of shooting crimes subsided somewhat, to around 180% compared to beginning of 2020.

7.5 Trends in Shooting incidents over time

Once the location based analysis was completed and we knew which were the safest areas in New York City now it was necessary to find out whether the above analysis still holds true over time or there are different patterns in such activities over the years.

We shifted our focus to analyse how the shooting incidents occurred over the years, months, weeks, hours to get a better idea of the trends and to understand the relationship between time vs location and significance of how these incidents are concentrated over a specific location at a specific point of time. For this even the data acquired during the pandemic was used to get a better idea.

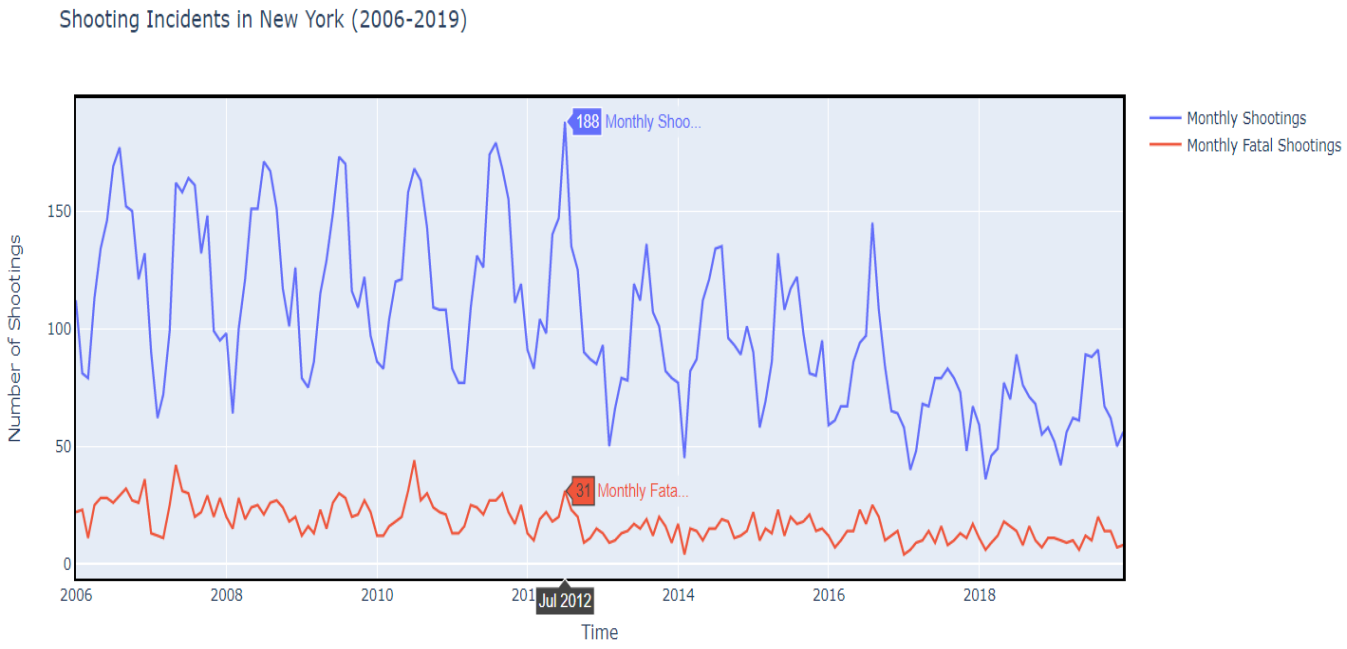

Figure 35: Shooting incidents from 2006- 2019

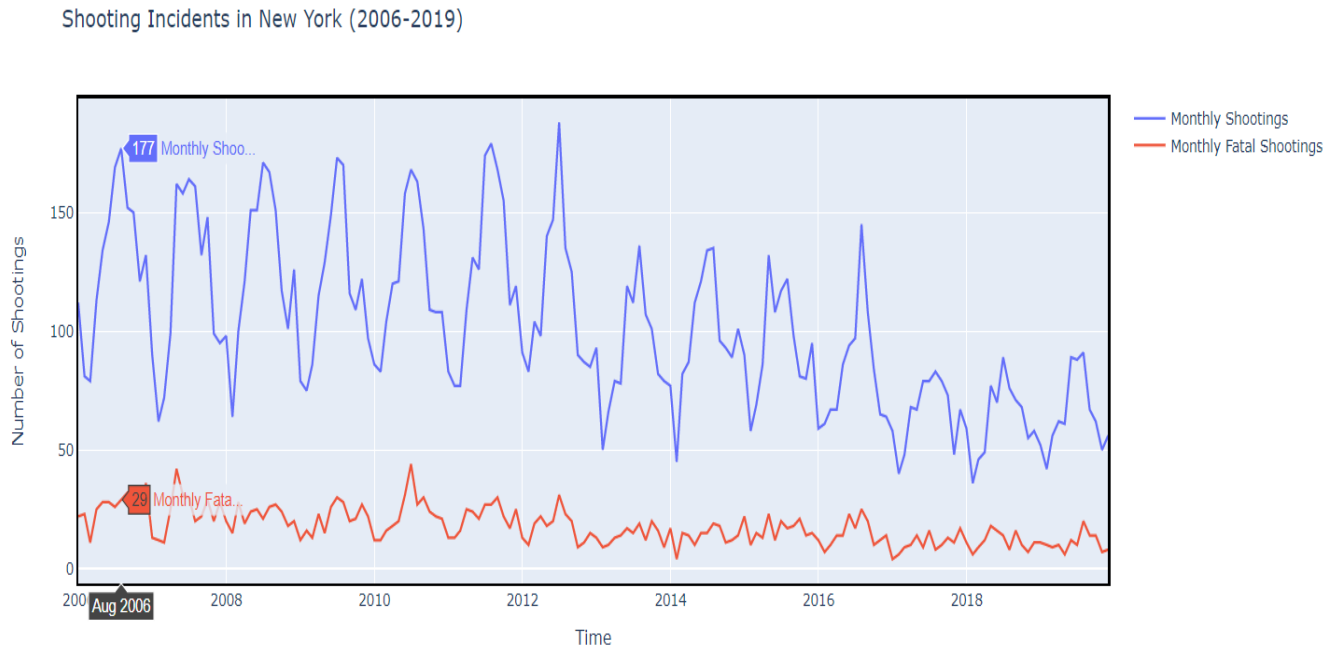

Figure 36: Shooting incidents from 2006- 2019

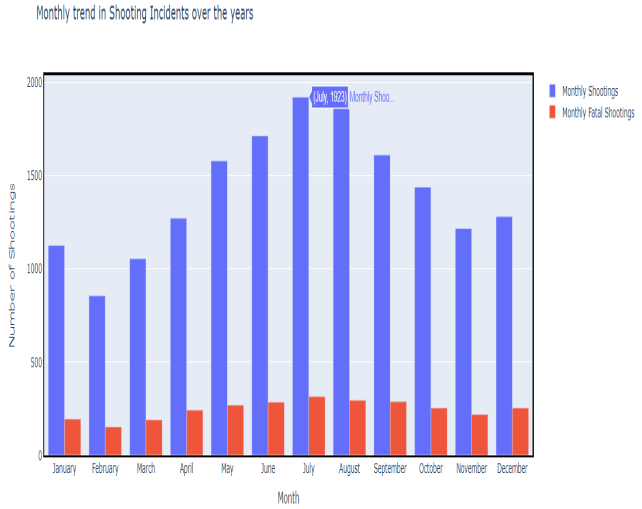

Figure 37: Shooting incidents from 2006- 2019 - Monthly Trends

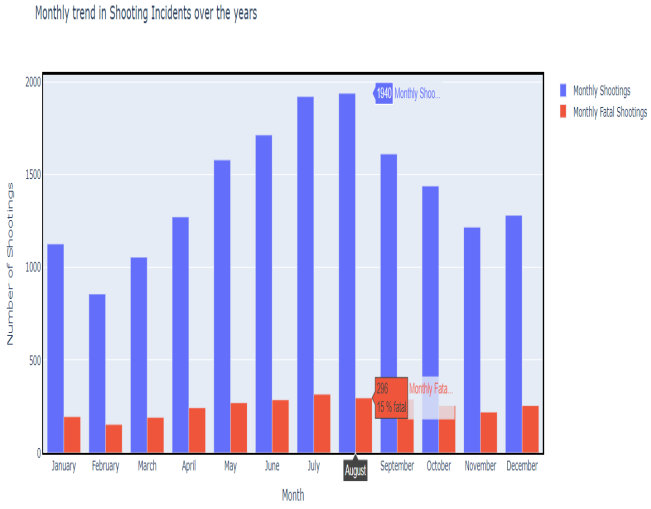

Figure 38: Shooting incidents from 2006- 2019 - Monthly Trends

The graph as shown in figure 35 and 36 denotes the trends in occurrences of such activities over the years on a monthly basis. As analyzed above the it can be confirmed that the latter years i.e. after 2012 have been more peaceful that the previous years.
Also, the number of fatal shooting incidents were approximately 20% of the total number of shooting incidents that occurred in every month for the year 2006 to 2019.
There are seasonal shooting incidents that peak during the summer, especially the months of July and August. These peaks are less pronounced after 2013. A steady decline in such incidents can be seen in the months of January to March.
Thus, it can be said that the Winters are less prone to such incidents occurring than the summer.
Comparing the total vs fatal shooting activities, the number of fatal shooting incidents always stays between 5-50 per month every year with an exception of July 2010 as there were 61 such cases registered for that period but no specific event happened during that time which means that the occurrence of so many incidents had no specific reason but followed the trend of peaks seen in July over the years. Number of total shooting incidents is more from 2006-2012 and started declining from 2012 -2016 and saw further decline from 2016 - 2019.
Although a lockdown was imposed during 2020 the number of shooting incidents still peaked during the month of July and we saw a similar trend as seen in the previous years.
Thus, it can be concluded that the months of July and August are more dangerous to be travelling to New York City and the months from January to March are the safest.

As deduced from the graphs in Fig 37 -39 there is a peak in shooting activities during the summers i.e. the months from May to August. The fatal shooting incidents follow a similar trend with it's count being more stable from one month to another. Similar trend was observed during 2020 even though a lockdown was imposed due to the Pandemic and hence we can conclude that the pandemic had no significant effect on the number of shooting activities.

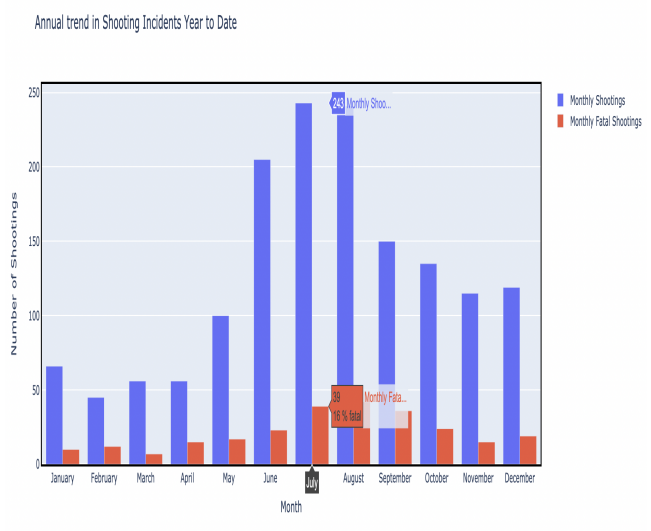

Figure 39: Shooting incidents in 2020 - Monthly Trends

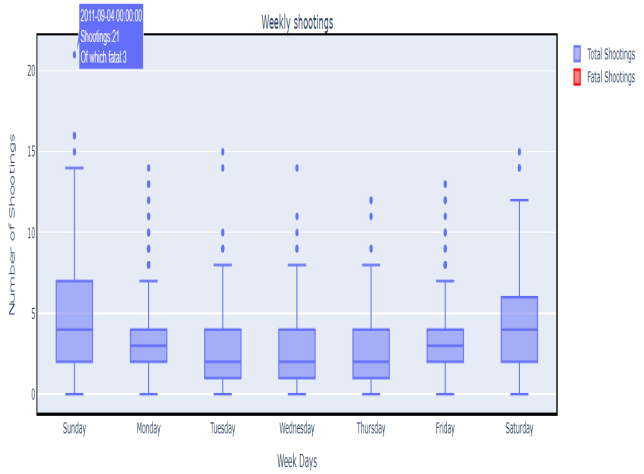

Figure 40: Shooting incidents from 2006- 2019 - Weekly Trends

To dig deeper we analysed the pattern to check if there were any weekly trends and an important thing observed is that Shooting activities are more common on the weekends whereas the most peaceful day as per our observation is Thursday.
During the 14 years that we have the data for i.e. 742 weeks only 18 Saturdays were peaceful. Based on detailed analysis it can be said that only 2.5% of the Saturdays were peaceful.
The day with most number of shootings is 4th September 2011 with 31 such incidents as shown in Fig 40

Data available for 2020 was insufficient to make any such analysis but a similar kind of trend was observed as shown in Fig 41

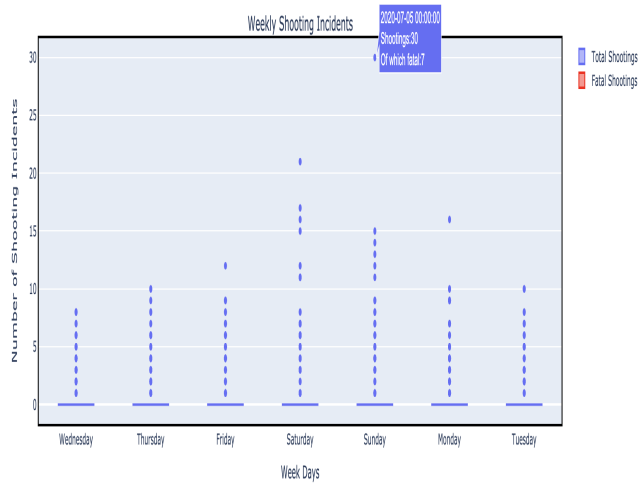

Figure 41: Shooting incidents from 2020 - Weekly Trends

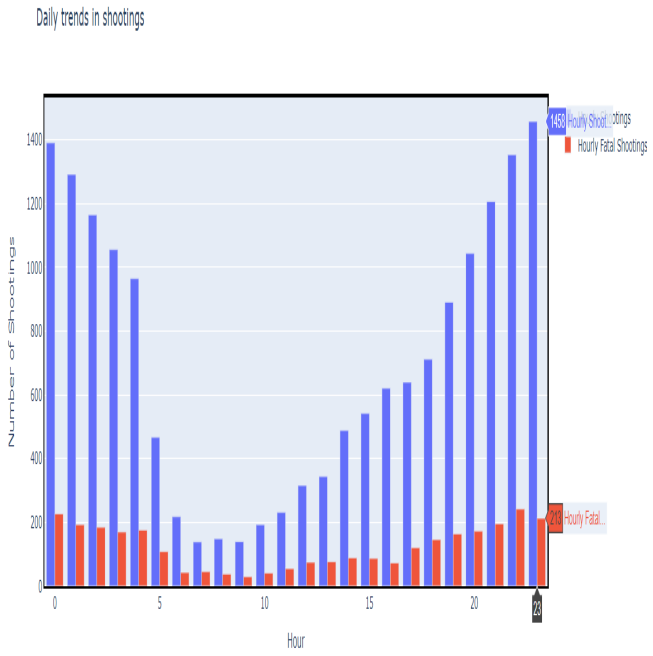

Figure 42: Shooting incidents from 2006- 2019 - Daily Trends

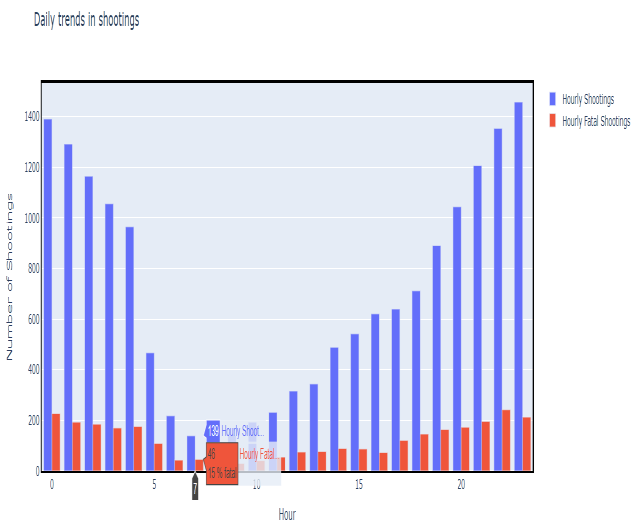

Figure 43: Shooting incidents from 2006- 2019 - Daily Trends

Now. since we knew what days of the weeks saw more shooting activities we analysed the hours at which such activities were most common. As seen in Fig 43 and 44, most number of shooting incidents(fatal or non-fatal) occur at the night especially showing peaks during midnight i.e. from 11 pm to 2 am. This pattern has been constant throughout the years but it can be seen diminishing as the count of shooting activities can be seen reducing from the year 2013. The most peaceful hours are between 7 am to 9 am.

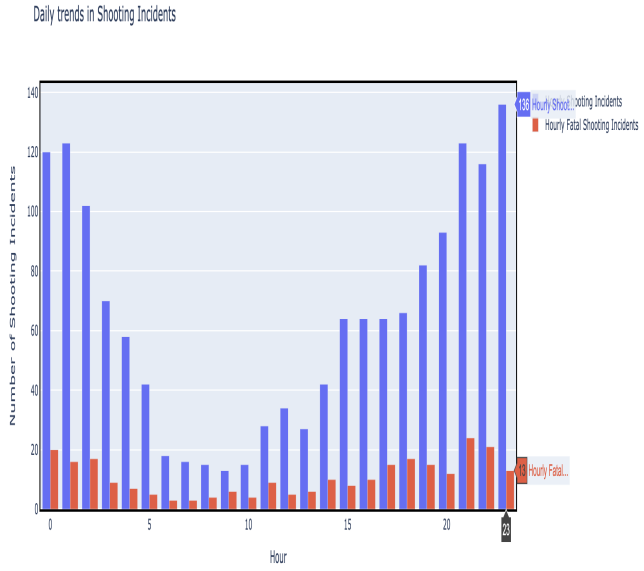

Figure 44: Shooting incidents in 2020 - Daily Trends

A similar curve was seen even during the pandemic as the most active hours were around midnight and the least active hours were early morning as shown in Fig 45-48.
The total incidents saw a decline after 2 am until 10 am but again started increasing after that for the year 2020.
But the graph for 2006 - 2019 has a more steady curve that 2020 which has a few more peaks even during the peaceful hours.

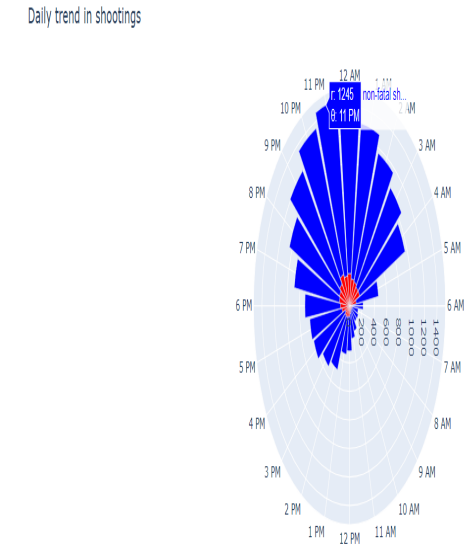

Figure 45: Shooting incidents from 2016 - 2019 - Daily Trends

Thus to summarize, we found out that over the years 2006 - 2012 have seen the most shooting activities which were followed by the years 2013 - 2016 and then 2017 to 2019 as there was a steady decline after 2012 but again during the year 2020 a significant increase can be seen.
Then, during all these years a similar trend was seen i.e. the months of July and August were the most violent and the months from January to March were the most peaceful.
Weekends have seen the most amount of such activities and Thursdays have been the most peaceful.
And on an hourly basis midnight is the most dangerous time of the day whereas early morning 7 -9 am is the least common time for such incidents.

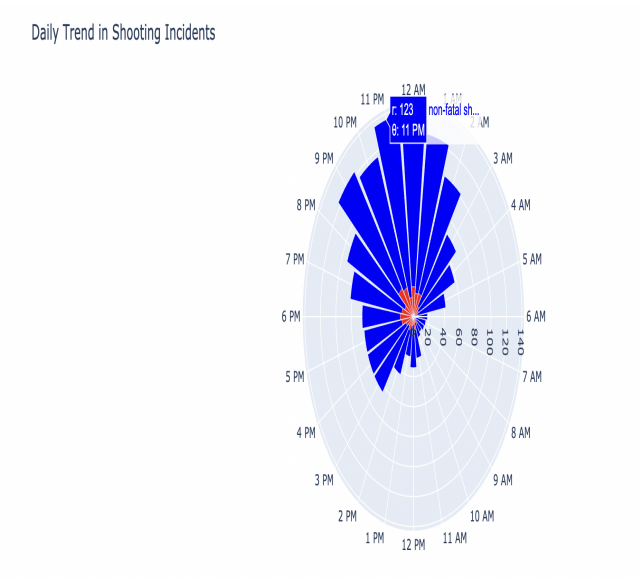

Figure 46: Shooting incidents in 2020 - Daily Trends

Please find below a complete hourly analysis of shooting incidents over the years:

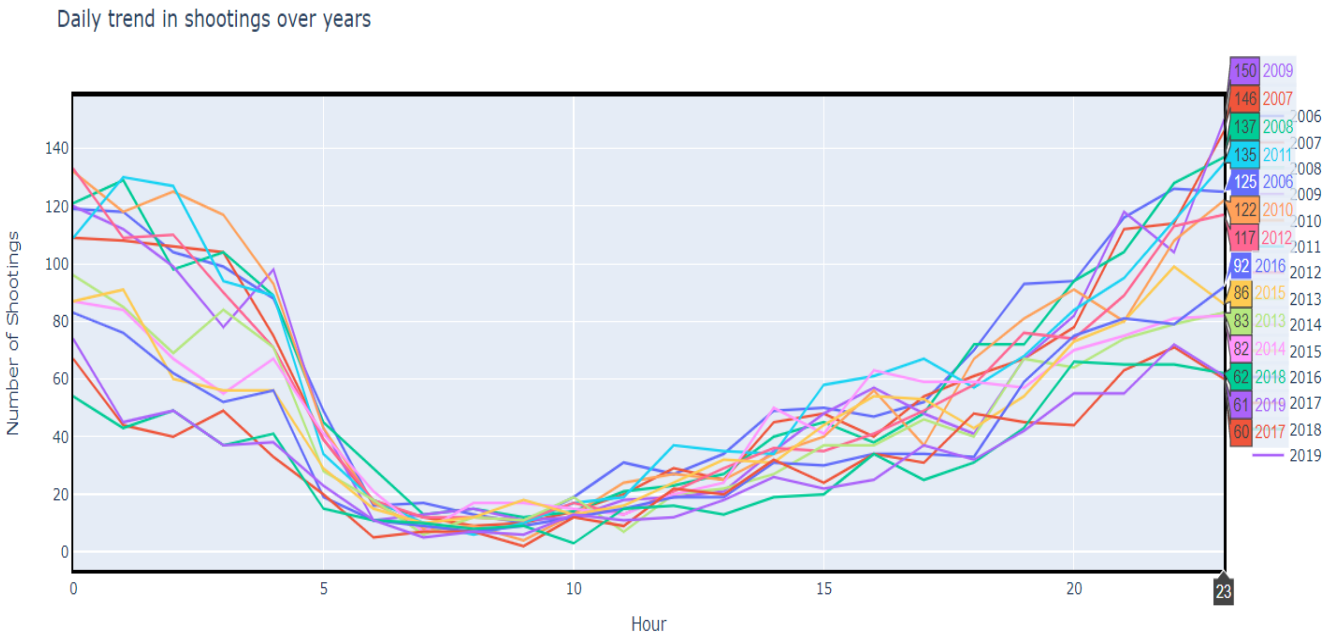

Figure 47: Shooting incidents from 2006 -2019 - Daily Trends

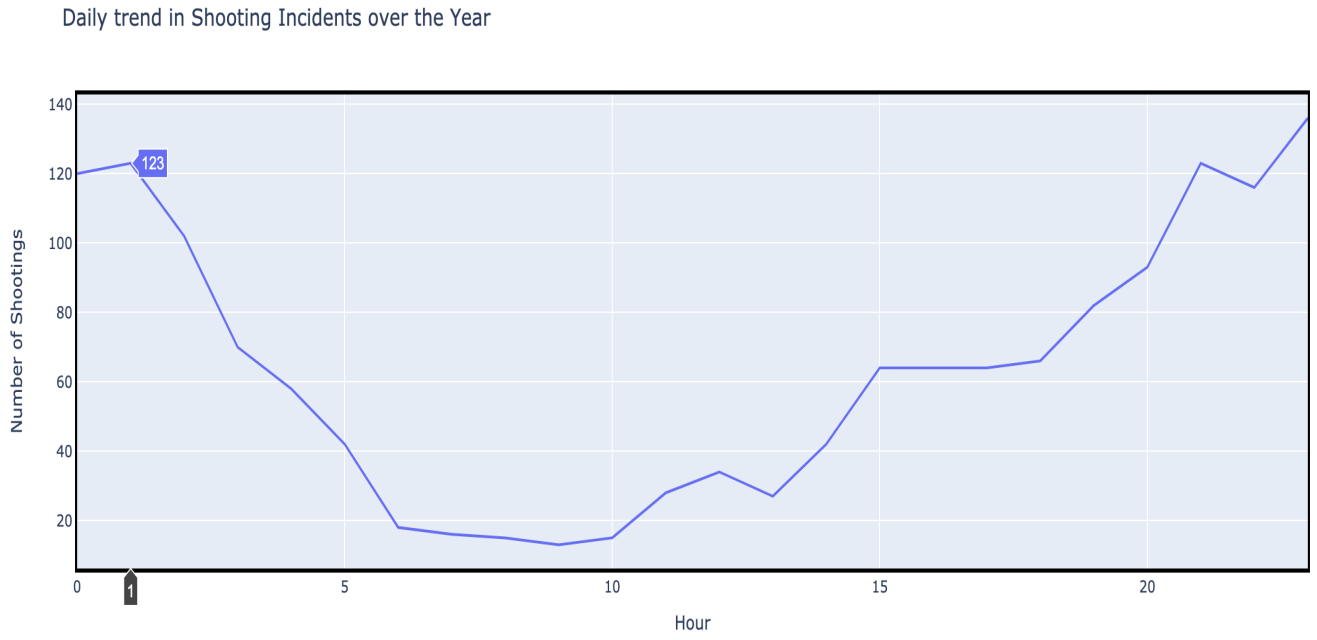

Figure 48: Shooting incidents in 2020 - Daily Trends

7.6 Analysis based on Individuals - Perpetrators and Victims

Since we have a better idea of the time and location of such activities over the years we tries to analyse how an individual behaviour trends over the years and contributes to such incidents.

PERP_SEX	F	M	U	All
PERP_AGE_GROUP				
<18	26	1337	69	1432
18-24	97	5422	285	5804
25-44	98	4574	247	4919
45-64	14	431	20	465
65+	1	51	1	53
UNKNOWN	29	2843	1480	4352
All	265	14658	2102	17025

Figure 49: Shooting incidents in previous years - Perpetrator Sex Trends

VIC_SEX	F	M	U	All
VIC_AGE_GROUP				
<18	210.0	1541.0	NaN	1751
18-24	373.0	6274.0	2.0	6649
25-44	476.0	6944.0	2.0	7422
45-64	164.0	897.0	NaN	1061
65+	34.0	73.0	NaN	107
UNKNOWN	4.0	26.0	5.0	35
All	1261.0	15755.0	9.0	17025

Figure 50: Shooting incidents in previous years - Victim Sex Trends

As seen from Fig 49 -52, Men have been more violent than women in general, a lot of data is missing, most probably that is because most shooters were not reported. Young men dominate the table and we observe that men under 45years old have committed approximately, 60% of total crimes. We see that the blacks and Hispanic blacks have committed more than 70% of crime. This can be linked to poverty and social distress.

Black males are the most common victims followed by White Hispanic males. Maximum number of victims belong to age grp of 18-24 followed by 25-44 and less than 18 respectively.
American Indian/Alaskan Natives are least common victims

PERP_SEX	F	M	U	All
PERP_AGE_GROUP				
<18	5.0	97.0	NaN	102
18-24	21.0	494.0	1.0	516
25-44	33.0	750.0	2.0	785
45-64	2.0	118.0	NaN	120
65+	NaN	6.0	NaN	6
UNKNOWN	NaN	1.0	1.0	2
All	61.0	1466.0	4.0	1531

Figure 51: Shooting incidents in 2020 - Perpetrator Sex Trends

VIC_SEX	F	M	U	All
VIC_AGE_GROUP				
<18	6.0	87.0	2.0	95
18-24	45.0	404.0	1.0	450
25-44	54.0	790.0	3.0	847
45-64	24.0	100.0	NaN	124
65+	4.0	5.0	NaN	9
UNKNOWN	NaN	5.0	1.0	6
All	133.0	1391.0	7.0	1531

Figure 52: Shooting incidents in 2020 - Victim Sex Trends

PERP_AGE_GROUP	PERP_RACE AMERICAN INDIAN/ALASKAN NATIVE	ASIAN / PACIFIC ISLANDER	BLACK	BLACK HISPANIC	UNKNOWN	WHITE	WHITE HISPANIC	All
<18	NaN	6.0	1023.0	93.0	102.0	14.0	194.0	1432
18-24	2.0	22.0	4143.0	453.0	386.0	76.0	722.0	5804
25-44	1.0	49.0	3435.0	328.0	337.0	129.0	640.0	4919
45-64	NaN	6.0	287.0	34.0	36.0	40.0	62.0	465
65+	NaN	NaN	28.0	3.0	1.0	14.0	7.0	53
UNKNOWN	NaN	15.0	2073.0	188.0	1724.0	40.0	312.0	4352
All	3.0	98.0	10989.0	1099.0	2586.0	313.0	1937.0	17025

Figure 53: Shooting incidents in previous years - Perpetrator Race Trends

VIC_AGE_GROUP	VIC_RACE AMERICAN INDIAN/ALASKAN NATIVE	ASIAN / PACIFIC ISLANDER	BLACK	BLACK HISPANIC	UNKNOWN	WHITE	WHITE HISPANIC	All
<18	2.0	8.0	1331.0	160.0	4.0	18.0	228.0	1751
18-24	4.0	55.0	4984.0	616.0	19.0	82.0	889.0	6649
25-44	2.0	103.0	5421.0	663.0	30.0	213.0	990.0	7422
45-64	NaN	41.0	669.0	103.0	9.0	109.0	130.0	1061
65+	NaN	2.0	53.0	12.0	NaN	24.0	16.0	107
UNKNOWN	NaN	1.0	10.0	3.0	11.0	6.0	4.0	35
All	8.0	210.0	12468.0	1557.0	73.0	452.0	2257.0	17025

Figure 54: Shooting incidents in previous years - Victim race Trends

PERP_AGE_GROUP	PERP_RACE ASIAN / PACIFIC ISLANDER	BLACK	BLACK HISPANIC	UNKNOWN	WHITE	WHITE HISPANIC	All
<18	1.0	79.0	4.0	5.0	NaN	13.0	102
18-24	10.0	370.0	33.0	28.0	5.0	70.0	516
25-44	7.0	549.0	61.0	30.0	17.0	121.0	785
45-64	1.0	77.0	17.0	NaN	6.0	19.0	120
65+	1.0	1.0	1.0	NaN	1.0	2.0	6
UNKNOWN	NaN	1.0	NaN	1.0	NaN	NaN	2
All	20.0	1077.0	116.0	64.0	29.0	225.0	1531

Figure 55: Shooting incidents in 2020 - Perpetrator Race Trends

VIC_AGE_GROUP	VIC_RACE ASIAN / PACIFIC ISLANDER	BLACK	BLACK HISPANIC	UNKNOWN	WHITE	WHITE HISPANIC	All
<18	1.0	73.0	10.0	NaN	NaN	11.0	95
18-24	7.0	321.0	35.0	3.0	12.0	72.0	450
25-44	16.0	635.0	65.0	1.0	13.0	117.0	847
45-64	2.0	83.0	9.0	NaN	6.0	24.0	124
65+	2.0	4.0	NaN	NaN	NaN	3.0	9
UNKNOWN	1.0	1.0	NaN	2.0	1.0	1.0	6
All	29.0	1117.0	119.0	6.0	32.0	228.0	1531

Figure 56: Shooting incidents in 2020 - Victim Race Trends

A key observation is that the race, age and sex of the perpetrators is follows in almost the same category and hence presumably the reason behind these shootings can be presence of gangs and evidence that only a certain areas have more concentration of such cases makes it more evident.

One more thing observed was that there are a lot of perpetrators and victims that were not identified or reported and hence a lot of NAN and unknown values can be seen in the visualizations. These values were kept as it is as this is not inconsistent or incorrect data and can be justified.

If observed carefully these values do not hinder our analysis much and hence were not removed during cleaning and kept for analysing as they provide significant insights.

7.7 Challenges involved in analyzing data

The most prominent challenge was finding the appropriate data that is reflective of the questions asked. Sometimes the data that directly answers our questions was not available, and we had to find other data which may indirectly answer our question. For example, when we decided to explore how COVID damaged household income in NYC, there was no income data available. We thought maybe the number of food stamp benefit claimants would increase coupling the economic downturn, but it turned out there was no correlation between SNAP program participants and economic health. We finally settled on labor participation stats from the department of labor. Usually it is not apparent what data is useful for our analysis and what data is not useful before processing and visualizing it. It would be useful to have some kind of automatic engine that can automatically search and match data trends based on specified parameters.

8 STEPS TO REPRODUCE THE ANALYSIS:

(Please refer to the github link available at the end of this paper for code and datasets:)

1. For reproducing the analysis please navigate to GitHub->code folder and make sure the correct datasets are loaded as mentioned in each notebook. Please run the notebooks on Google Colab and Jupyter Notebook as required respectively:
Analysis_Historic.ipynb , Analysis_YearToDate.ipynb,
Shooting_Map.ipynb, Shooting_vs_Income.ipynb ,
Arrest_Data_toDate.ipynb,Labor_Stats_10Years.ipynb,
Conclusion_Shooting_Increase.ipynb

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

9 GITHUB REPOSITORY :

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

10 CONCLUSION

After deep analysis of the data we could find many interesting observations which are listed below:

- We can see a clear inverse relationship between income and gun crime counts. High end neighborhoods rarely have any gun crimes occur. It may be related to the ability of families in these communities to procure security systems or hire management personnel. Such communities usually have a much lower population density, which lower the likelihood of crimes occurring.
- Safest areas include most of Manhattan, Queens, Staten Island and eastern/southern parts of Brooklyn.
- Most of the shooting incidents are concentrated in two of the boroughs namely Brooklyn and Bronx. The total count of these boroughs is almost 70% of the total count. The interactive map tells us that even within the boroughs, such activities are concentrated in certain areas which show a significant number of shootings than the rest which proves that even within a borough that shows very high concentration of such incidents there are many areas that are very safe.
- Out of all the shooting activities within a reported place i.e. Multip-dwellig houses are more prone to shootings and commercial buildings see low occurrences of such incidences.
- Over the years 2006 - 2012 we have seen the most shooting activities which were followed by the years 2013 - 2016 and then 2017 to 2019 as there was a steady decline after 2012 but again during the year 2020 a significant increase can be seen.
- Then, during all these years a similar trend was seen i.e. the months of July and August were the most violent and the months from January to March were the most peaceful.
- Weekends have seen the most amount of such activities and Thursdays have been the most peaceful.
- And on an hourly basis midnight is the most dangerous time of the day whereas early morning 7 -9 am is the least common time for such incidents.
- Men have been more violent than women in general, a lot of data is missing, most probably that is because most shooters weren't reported. Young men dominate the table and we observe that men under 45years old have committed approximately, 60% of total crimes.
- Black males are the most common perpetrators as well as the most common victims followed by White Hispanic males.

Application of the above findings: So these major finding can be used by many applications, agencies and associations to help there clients and customers. In 2017 visitors to New York City spent more than USD 44.2 billion while staying here, generating an economic impact totaling more than USD 70 billion. The economic importance of tourism to New York City's economy is substantial, so these findings can be used by tourists to make there itinerary. All the mobile applications such as Citizen, LexisNexis Community Crime Map can use the findings in the report. The police department can also get a major analysis for gun-voilence and can help reduce the shooting incidents by taking required actions at the most common locations as analyzed earlier at the most common times.