NYPD Shooting Incident Data Report

Student

2024-08-20

This report provides a comprehensive analysis of the NYPD crime data, focusing on the spatial and temporal distribution of incidents across New York City. The data set, sourced from the NYPD, encompasses a range of variables including incident types, dates, locations, and demographic information about victims and perpetrators. The primary objective of this analysis is to explore the trends in crime rates over time and identify patterns based on geographic locations. This report will leverage statistical and visualization tools to provide insights into crime patterns, aiming to assist policymakers, law enforcement agencies, and community stakeholders in making informed decisions to enhance public safety. Key aspects of the analysis include Temporal Analysis and Spatial Analysis. We will examine crime trends across different months and years to identify any significant increases or decreases in incident rates. We will map the distribution of incidents to detect hotspots and areas with higher crime rates.

We will use the libraries dplyr, lubridate, ggplot2, leaflet, sf and caret for this project.We will start by reading in the data from the main csv file "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD. This file contains the list of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

```
## Get current data
nypd_data <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
head(nypd_data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME      BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    244608249 05/05/2022   00:10:00 MANHATTAN            INSIDE       14
## 2    247542571 07/04/2022   22:20:00     BRONX           OUTSIDE       48
## 3     84967535 05/27/2012   19:35:00    QUEENS                        103
## 4    202853370 09/24/2019   21:00:00     BRONX                         42
## 5     27078636 02/25/2007   21:00:00  BROOKLYN                         83
## 6    230311078 07/01/2021   23:07:00 MANHATTAN                         23
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC             LOCATION_DESC
## 1                 0         COMMERCIAL               VIDEO STORE
## 2                 0             STREET                    (null)
## 3                 0
## 4                 0
## 5                 0
## 6                 2                    MULTI DWELL - PUBLIC HOUS
##   STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## 1                    true          25-44        M     BLACK         25-44
## 2                    true         (null)   (null)    (null)         18-24
## 3                   false                                           18-24
## 4                   false          25-44        M   UNKNOWN         25-44
## 5                   false          25-44        M     BLACK         25-44
## 6                   false                                           25-44
##   VIC_SEX VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
## 1       M    BLACK     986050   214231.0 40.75469 -73.99350
## 2       M    BLACK    1016802   250581.0 40.85440 -73.88233
## 3       M    BLACK    1048632   198262.0 40.71063 -73.76777
## 4       M    BLACK    1014493   242565.0 40.83242 -73.89071
## 5       M    BLACK    1009149   190104.7 40.68844 -73.91022
## 6       M    BLACK     999061   229912.0 40.79773 -73.94651
##                                        Lon_Lat
## 1                   POINT (-73.9935 40.754692)
## 2                  POINT (-73.88233 40.854402)
## 3  POINT (-73.76777349199995 40.71063412500007)
## 4 POINT (-73.89071440599997 40.832416753000075)
## 5  POINT (-73.91021857399994 40.68844345900004)
## 6  POINT (-73.94650786199998 40.79772716600007)
```

We do not need some of the columns for our analysis so we will remove them.

```
nypd_data <- nypd_data %>%
  select(-c(LOC_OF_OCCUR_DESC, PRECINCT, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_DESC, STATISTICAL_MURDER
_FLAG,PERP_AGE_GROUP,PERP_SEX,PERP_RACE,X_COORD_CD,Y_COORD_CD,Lon_Lat))
```

Now, let us check the datatypes of the columns.

```
str(nypd_data)
```

```
## 'data.frame':    28562 obs. of  9 variables:
##  $ INCIDENT_KEY : int  244608249 247542571 84967535 202853370 27078636 230311078 229224142 231246224 228559720
238210279 ...
##  $ OCCUR_DATE   : chr  "05/05/2022" "07/04/2022" "05/27/2012" "09/24/2019" ...
##  $ OCCUR_TIME   : chr  "00:10:00" "22:20:00" "19:35:00" "21:00:00" ...
##  $ BORO         : chr  "MANHATTAN" "BRONX" "QUEENS" "BRONX" ...
##  $ VIC_AGE_GROUP: chr  "25-44" "18-24" "18-24" "25-44" ...
##  $ VIC_SEX      : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE     : chr  "BLACK" "BLACK" "BLACK" "BLACK" ...
##  $ Latitude     : num  40.8 40.9 40.7 40.8 40.7 ...
##  $ Longitude    : num  -74 -73.9 -73.8 -73.9 -73.9 ...
```

OCCUR_DATE is not a Date type so we will make it a Date type. OCCUR_TIME is not time type so we will convert it to a Time type. We will use library(lubridate) for this.

```
nypd_data <- nypd_data %>%
     mutate(OCCUR_DATE = mdy(OCCUR_DATE),
            OCCUR_TIME = hms(OCCUR_TIME))
```

Lets look at the summary to see if we have changed the datatype correctly.

```
summary(nypd_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE           OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Min.   :0S
##  1st Qu.: 65439914   1st Qu.:2009-09-04   1st Qu.:3H 30M 0S
##  Median : 92711254   Median :2013-09-20   Median :15H 15M 0S
##  Mean   :127405824   Mean   :2014-06-07   Mean   :12H 44M 16.7131153281007S
##  3rd Qu.:203131993   3rd Qu.:2019-09-29   3rd Qu.:20H 45M 0S
##  Max.   :279758069   Max.   :2023-12-29   Max.   :23H 59M 0S
##
##      BORO           VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Latitude       Longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.92
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##  NA's   :59      NA's   :59
```

We will now add a new column 'Year' to extract the year from OCCUR_DATE.

```
## Extract Year from OCCUR_DATE
nypd_data <- nypd_data %>%
  mutate(Year = year(OCCUR_DATE))
head(nypd_data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME      BORO VIC_AGE_GROUP VIC_SEX VIC_RACE
## 1    244608249 2022-05-05     10M 0S MANHATTAN         25-44       M    BLACK
## 2    247542571 2022-07-04 22H 20M 0S     BRONX         18-24       M    BLACK
## 3     84967535 2012-05-27 19H 35M 0S    QUEENS         18-24       M    BLACK
## 4    202853370 2019-09-24  21H 0M 0S     BRONX         25-44       M    BLACK
## 5     27078636 2007-02-25  21H 0M 0S  BROOKLYN         25-44       M    BLACK
## 6    230311078 2021-07-01  23H 7M 0S MANHATTAN         25-44       M    BLACK
##   Latitude Longitude Year
## 1 40.75469 -73.99350 2022
## 2 40.85440 -73.88233 2022
## 3 40.71063 -73.76777 2012
## 4 40.83242 -73.89071 2019
## 5 40.68844 -73.91022 2007
## 6 40.79773 -73.94651 2021
```

Analysis of the number of incidents by month

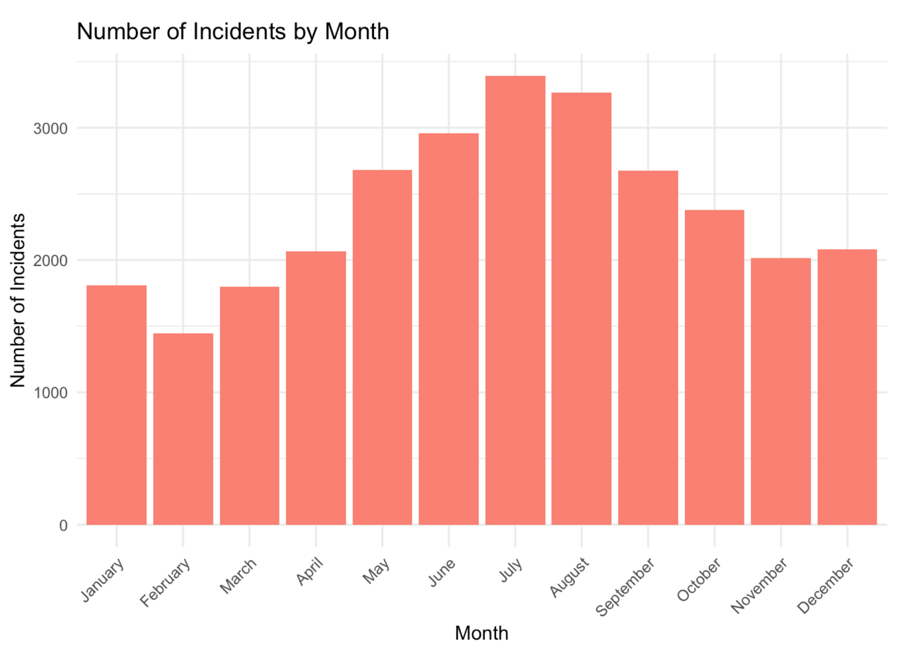
Let us analyze how the number of incidents vary by month.

```
nypd_data_bymonth <- nypd_data %>%
## Extract month name and order it
     mutate(
         Month = format(OCCUR_DATE, "%B"),
         Month = factor(Month, levels = month.name)
    ) %>%
## Group by Month and count incidents
    group_by(Month) %>%
     summarize(Incident_Count = n(), .groups = 'drop')
head(nypd_data_bymonth)
```

```
## # A tibble: 6 × 2
##   Month     Incident_Count
##   <fct>              <int>
## 1 January             1809
## 2 February            1444
## 3 March               1797
## 4 April               2068
## 5 May                 2682
## 6 June                2959
```

Now, let us visualize this.

```
ggplot(nypd_data_bymonth, aes(x = Month, y = Incident_Count)) +
    geom_bar(stat = "identity", fill = "salmon") +
    labs(title = "Number of Incidents by Month", x = "Month", y = "Number of Incidents") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We see that most number of incidents occur in summer season. From the graph we can see that the number of incidents increased from February to July and then decrease from August to November.

Interactive map to see the incidents in different boroughs

Now, let us visualize the incidents in different boroughs on the map. We need to convert our data to an sf object for this but before that, we need to filter out rows with missing values in the Longitude or Latitude columns.

```
# Filtering out rows with missing coordinates
nypd_filtered <- nypd_data %>%
  filter(!is.na(Longitude) & !is.na(Latitude))
```

```
# Converting to sf object
nypd_sf <- st_as_sf(nypd_filtered, coords = c("Longitude", "Latitude"), crs = 4326)
```

```
# Ensuring that the coordinates are numeric
nypd_sf <- nypd_sf %>%
  mutate(Latitude = as.numeric(st_coordinates(.)[, "Y"]),
         Longitude = as.numeric(st_coordinates(.)[, "X"]))
```

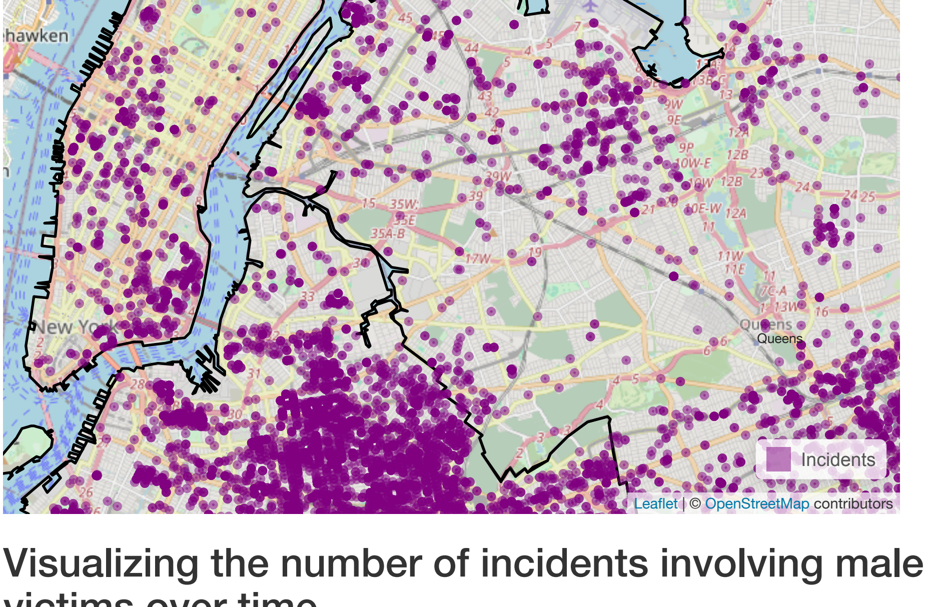
We will use GeoJSON file from "https://data.cityofnewyork.us/api/geospatial/tqmj-j8zm?method=export&format=GeoJSON" to add the boundaries to separate the boroughs.

```
# Loading the GeoJSON file
boroughs_sf <- st_read("https://data.cityofnewyork.us/api/geospatial/tqmj-j8zm?method=export&format=GeoJSON")
```

```
## Reading layer `OGRGeoJSON' from data source
##   `https://data.cityofnewyork.us/api/geospatial/tqmj-j8zm?method=export&format=GeoJSON'
##   using driver `GeoJSON'
## Simple feature collection with 5 features and 4 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -74.25559 ymin: 40.49613 xmax: -73.70001 ymax: 40.91553
## Geodetic CRS:  WGS 84
```

We will use leaflet library to create the interactive map showing the incidents.

```
# Creating an interactive map
leaflet() %>%
  addProviderTiles(providers$OpenStreetMap) %>%
  addPolygons(data = boroughs_sf,
              fillColor = "lightgrey",
              color = "black",
              weight = 2,
              opacity = 1,
              fillOpacity = 0.3) %>%
  addCircles(data = nypd_sf,
             radius = 2,
             color = "purple",
             opacity = 0.5,
             fillOpacity = 0.5) %>%
  addLegend(position = "bottomright",
            colors = "purple",
            labels = "Incidents") %>%
  addLabelOnlyMarkers(
    data = boroughs_sf,
    ~ st_coordinates(st_centroid(geometry))[,1],
    ~ st_coordinates(st_centroid(geometry))[,2],
    label = ~ boro_name,
    labelOptions = labelOptions(noHide = TRUE, textOnly = TRUE, direction = 'auto', offset = c(0, -10))
  ) %>%
  setView(lng = mean(nypd_sf$Longitude, na.rm = TRUE),
          lat = mean(nypd_sf$Latitude, na.rm = TRUE),
          zoom = 12)
```



Visualizing the number of incidents involving male and female victims over time

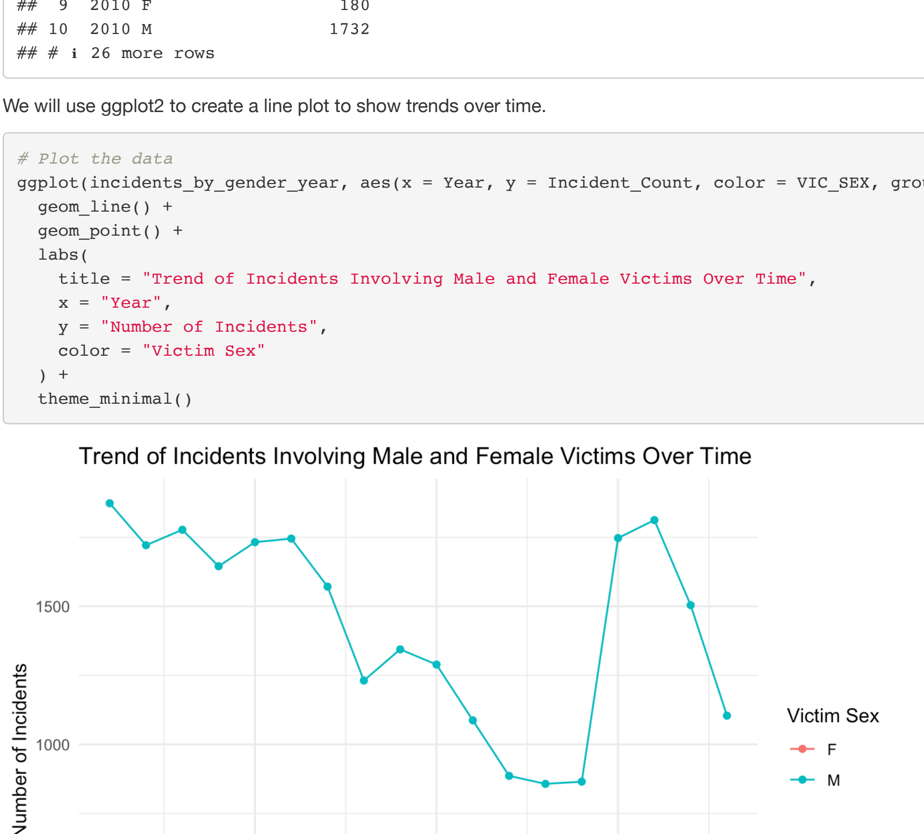
We would like to visualize how the number of incidents involving male and female victims has changed over time.

```
# Count incidents by year and gender
incidents_by_gender_year <- nypd_data %>%
  filter(VIC_SEX %in% c("M", "F")) %>%
  group_by(Year, VIC_SEX) %>%
  summarize(Incident_Count = n(), .groups = 'drop')
incidents_by_gender_year
```

```
## # A tibble: 36 × 3
##     Year VIC_SEX Incident_Count
##    <dbl> <chr>            <int>
##  1  2006 F                  182
##  2  2006 M                 1873
##  3  2007 F                  162
##  4  2007 M                 1721
##  5  2008 F                  182
##  6  2008 M                 1777
##  7  2009 F                  181
##  8  2009 M                 1645
##  9  2010 F                  180
## 10  2010 M                 1732
## # ℹ 26 more rows
```

We will use ggplot2 to create a line plot to show trends over time.

```
# Plot the data
ggplot(incidents_by_gender_year, aes(x = Year, y = Incident_Count, color = VIC_SEX, group = VIC_SEX)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Trend of Incidents Involving Male and Female Victims Over Time",
    x = "Year",
    y = "Number of Incidents",
    color = "Victim Sex"
  ) +
  theme_minimal()
```



From the graph we see that the number of incidents were decreasing from 2006 to 2019 but then suddenly increased from 2019 to 2021 and then they drastically decreased from 2021 to 2023 for males. Number of incidents for females are considerably lower compared to males.

Predictive Modeling to predict the number of incidents in 2024

We will create a model to predict the number of incidents where the victim would be female or male in 2024. We will use the train function from the caret package for this.

```
# Group the data and get the incident count
demo_data <- nypd_data %>%
  group_by(VIC_SEX, Year) %>%
  mutate(Incident_Count = n()) %>%
  ungroup()
```

We will use Generalized Linear Model (GLM) with a Poisson distribution for this as we need to get the incident count.

```
# Train the Generalized Linear Model (GLM)
model_demographic <- train(
  Incident_Count ~ VIC_SEX + Year,
  data = demo_data,
  method = "glm",
  family = "poisson"
)
```

```
# Prepare new data for prediction
future_demographics <- data.frame(
  VIC_SEX = c("M", "F"),
  Year = c(2024, 2024)
)
```

```
# Predict incident counts based on new demographics
demographic_predictions <- predict(model_demographic, newdata = future_demographics)

# View the predictions
print(demographic_predictions)
```

```
##         1         2
## 1258.9808  136.7189
```

From the above prediction we see that the number of predicted incidents in 2024 is 1258.9808 for males and 136.7189 for females.

Bias

When analyzing and reporting NYPD Shooting Incident Data (Historic), several potential biases and limitations can affect the accuracy and interpretation of the results. Understanding these biases is crucial for ensuring that the insights derived from the analysis are reliable and actionable. Some of these biases are:

Reporting Bias: Certain types of crimes may be underreported, especially sensitive incidents such as domestic violence and sexual assault. This underreporting can skew the data and lead to inaccurate conclusions about the prevalence and distribution of crimes. On the other hand, some areas may have higher reporting rates due to increased community vigilance or more proactive policing, which might not necessarily reflect a higher actual crime rate.

Temporal Bias: Focusing on data from a limited time frame without considering seasonal patterns can lead to misleading conclusions. Changes in crime reporting practices, law enforcement policies, or socio-economic conditions over time can impact crime rates and trends. Failing to account for these changes may result in incorrect interpretations.

Demographic Bias: The analysis of crime data by demographic factors such as age, sex, and race can be biased if certain groups are overrepresented or underrepresented in the data. For example, from the "Trend of Incidents Involving Male and Female Victims Over Time" above we see that the from 2006 to 2023, the number of incidents with female victims is considerably lower than the number of incidents with male victims. But it is possible that females are underrepresented in this data.

Socio-Economic Bias: Changes in socio-economic conditions, such as unemployment rates or housing instability, can influence crime rates. For example, increase in the number of incidents from 2019 to 2021. If these factors are not included in the analysis, the results may not fully capture the underlying drivers of crime.

Conclusion

The analysis of the NYPD crime data has provided valuable insights into crime trends and patterns across New York City. The analysis revealed variations in crime rates over different months and years, highlighting periods of increased incidents (summer months and from 2019 to 2021) or decreased incidents. These trends can help in understanding seasonal or year-specific fluctuations in crime rates. The mapping of incident locations identified specific hotspots and areas with higher crime rates. This is crucial for targeted law enforcement interventions and resource allocation to improve public safety in high-crime areas. By leveraging these insights, stakeholders can develop more effective strategies for crime prevention, resource management, and community engagement. Future analyses could build upon these findings by incorporating additional variables, such as economic factors or changes in law enforcement practices, to gain a deeper understanding of crime dynamics in New York City.