NYPD Shooting Incident Data: Cleaning and Summary

Rasha Ahmed

2025-06-24

reading in the data from this link below https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

```
# Load the data directly from the NYC Open Data URL (updated quarterly)
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

# Read the CSV into R
shootings_raw <- read_csv(url)

# Preview the structure of the raw data
glimpse(shootings_raw)
```

```
## Rows: 29,744
## Columns: 21
## $ INCIDENT_KEY            <dbl> 231974218, 177934247, 255028563, 25384540, 726~
## $ OCCUR_DATE              <chr> "08/09/2021", "04/07/2018", "12/02/2022", "11/~
## $ OCCUR_TIME              <time> 01:06:00, 19:48:00, 22:57:00, 01:50:00, 01:58~
## $ BORO                    <chr> "BRONX", "BROOKLYN", "BRONX", "BROOKLYN", "BRO~
## $ LOC_OF_OCCUR_DESC       <chr> NA, NA, "OUTSIDE", NA, NA, NA, NA, NA, NA, NA,~
## $ PRECINCT                <dbl> 40, 79, 47, 66, 46, 42, 71, 69, 75, 69, 40, 42~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 2, 0, 0, 2, 0~
## $ LOC_CLASSFCTN_DESC      <chr> NA, NA, "STREET", NA, NA, NA, NA, NA, NA, NA, ~
## $ LOCATION_DESC           <chr> NA, NA, "GROCERY/BODEGA", "PVT HOUSE", "MULTI ~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, TRUE, F~
## $ PERP_AGE_GROUP          <chr> NA, "25-44", "(null)", "UNKNOWN", "25-44", "18~
## $ PERP_SEX                <chr> NA, "M", "(null)", "U", "M", "M", NA, NA, "M",~
## $ PERP_RACE               <chr> NA, "WHITE HISPANIC", "(null)", "UNKNOWN", "BL~
## $ VIC_AGE_GROUP           <chr> "18-24", "25-44", "25-44", "18-24", "<18", "18~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "F", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 1006343.0, 1000082.9, 1020691.0, 985107.3, 100~
## $ Y_COORD_CD              <dbl> 234270.0, 189064.7, 257125.0, 173349.8, 247502~
## $ Latitude                <dbl> 40.80967, 40.68561, 40.87235, 40.64249, 40.845~
## $ Longitude               <dbl> -73.92019, -73.94291, -73.86823, -73.99691, -7~
## $ Lon_Lat                 <chr> "POINT (-73.92019278899994 40.80967347200004)"~
```

```
# Step 1: Clean column names (lowercase and remove spaces)
shootings_clean <- shootings_raw %>%
  rename_with(~ str_replace_all(., "\\s+", "_")) %>%
  rename_with(~ str_to_lower(.))
```

```r
# Step 2: Convert appropriate columns to correct types
# Dates (use correct format as in the actual data)
# Times (hms() parses HH:MM:SS format)
# Factors for categorical fields

shootings_clean <- shootings_clean %>%
  mutate(
    occur_date = mdy(occur_date),                       # Convert date
    occur_time = parse_time(as.character(occur_time)), # Convert time safely
    boro = as.factor(boro),
    location_desc = as.factor(location_desc),
    perp_sex = as.factor(perp_sex),
    perp_race = as.factor(perp_race),
    vic_sex = as.factor(vic_sex),
    vic_race = as.factor(vic_race)
  )

# Step 3: Drop columns not needed for analysis
# Adjust the columns based on your analysis goals

shootings_clean <- shootings_clean %>%
  select(
    occur_date, occur_time, boro, precinct, location_desc,
    perp_sex, perp_race, vic_sex, vic_race,
    latitude, longitude
  )
```

```r
# Summarize number of missing values per column
missing_summary <- shootings_clean %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "missing_count") %>%
  arrange(desc(missing_count))

# Display missing data summary
missing_summary
```

```
## # A tibble: 11 x 2
##    variable      missing_count
##    <chr>                 <int>
##  1 location_desc         14977
##  2 perp_sex               9310
##  3 perp_race              9310
##  4 latitude                 97
##  5 longitude                97
##  6 occur_date                0
##  7 occur_time                0
##  8 boro                      0
##  9 precinct                  0
## 10 vic_sex                   0
## 11 vic_race                  0
```

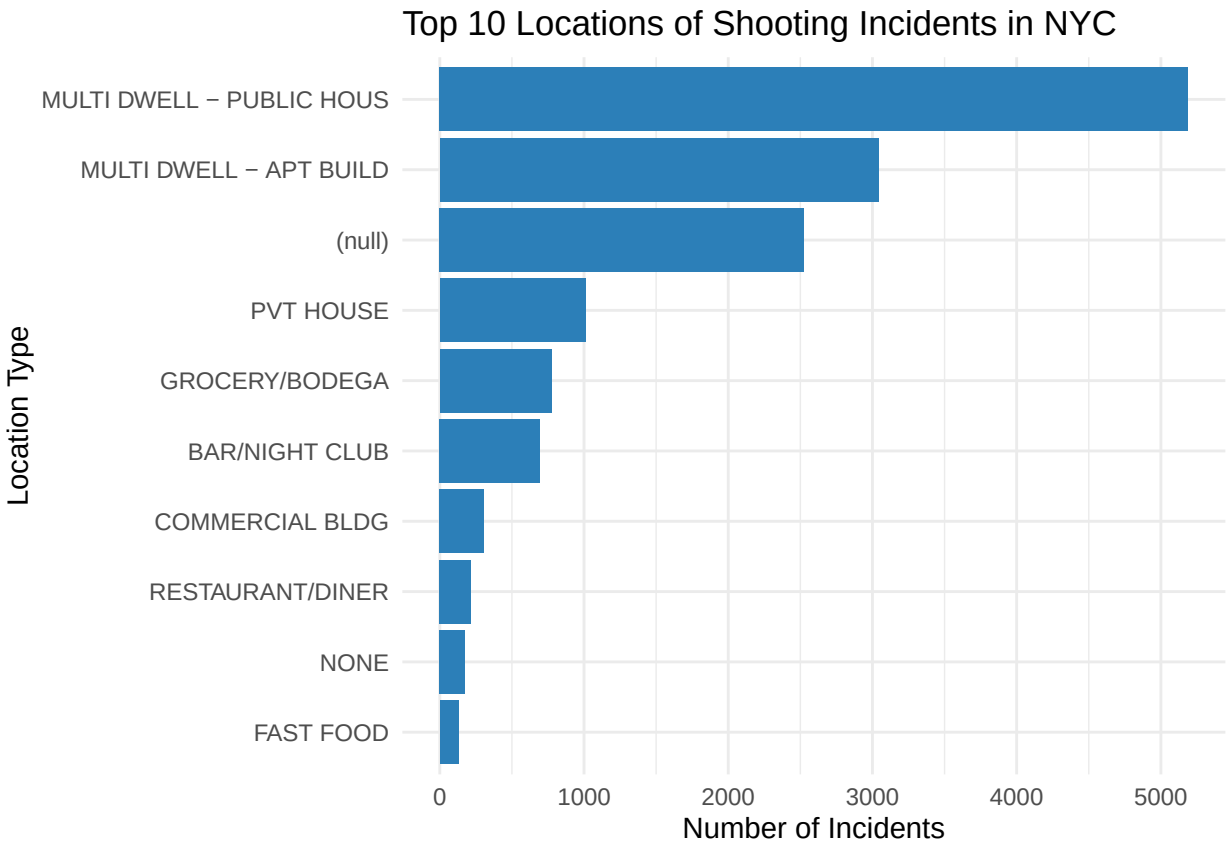
```r
# Plan:
# - Drop rows where key analysis variables are missing: boro, occur_date
# - Retain rows with missing perp info (often not available for open cases)

shootings_final <- shootings_clean %>%
  drop_na(occur_date, boro)

# Recheck summary to confirm data integrity
summary(shootings_final)
```

```
##    occur_date           occur_time                          boro
##  Min.   :2006-01-01   Min.   :00:00:00.000000   BRONX        : 8834
##  1st Qu.:2009-10-29   1st Qu.:03:30:45.000000   BROOKLYN     :11685
##  Median :2014-03-25   Median :15:15:00.000000   MANHATTAN    : 3977
##  Mean   :2014-10-31   Mean   :12:46:10.874798   QUEENS       : 4426
##  3rd Qu.:2020-06-29   3rd Qu.:20:44:00.000000   STATEN ISLAND:  822
##  Max.   :2024-12-31   Max.   :23:59:00.000000
##
##     precinct                        location_desc     perp_sex
##  Min.   :  1.00   MULTI DWELL - PUBLIC HOUS: 5188   (null): 1628
##  1st Qu.: 44.00   MULTI DWELL - APT BUILD  : 3042   F     :  461
##  Median : 67.00   (null)                   : 2526   M     :16845
##  Mean   : 65.23   PVT HOUSE                : 1010   U     : 1500
##  3rd Qu.: 81.00   GROCERY/BODEGA           :  775   NA's  : 9310
##  Max.   :123.00   (Other)                  : 2226
##                   NA's                     :14977
##           perp_race     vic_sex                              vic_race
##  BLACK         :12323   F: 2891   AMERICAN INDIAN/ALASKAN NATIVE:   13
##  WHITE HISPANIC: 2667   M:26841   ASIAN / PACIFIC ISLANDER      :  478
##  UNKNOWN       : 1838   U:   12   BLACK                         :20999
##  (null)        : 1628             BLACK HISPANIC                : 2930
##  BLACK HISPANIC: 1487             UNKNOWN                       :   72
##  (Other)       :  491             WHITE                         :  741
##  NA's          : 9310             WHITE HISPANIC                : 4511
##     latitude       longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.91
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.83   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##  NA's   :97      NA's   :97
```

This breakdown shows racial disparities in who is affected by shootings across different boroughs. Certain groups appear to be disproportionately affected depending on location. This raises questions about underlying systemic or community-level factors contributing to violence.

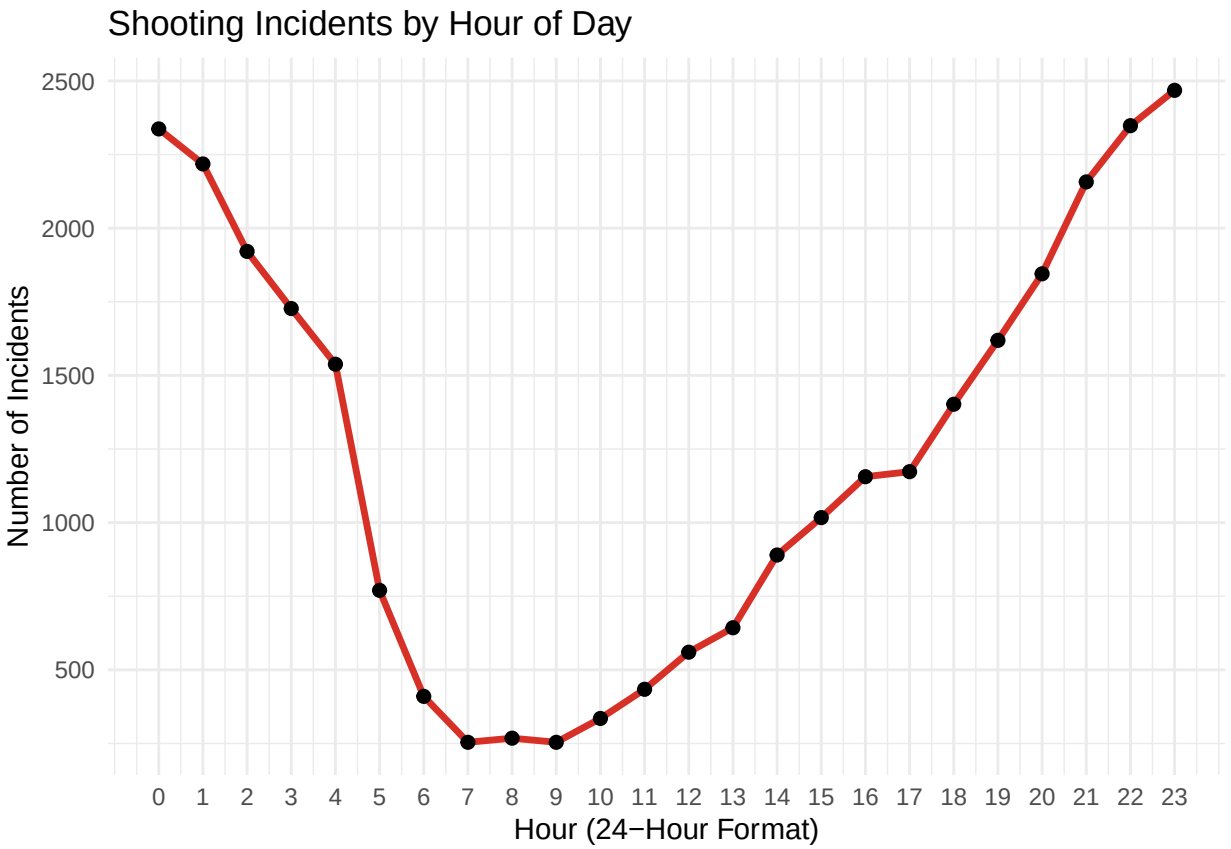
```r
# Count the number of shootings by location type
shootings_final %>%
  filter(!is.na(location_desc)) %>%
  count(location_desc, sort = TRUE) %>%
  slice_max(n, n = 10) %>%  # Show top 10 most common locations
  ggplot(aes(x = reorder(location_desc, n), y = n)) +
```

```
  geom_col(fill = "#2c7fb8") +
  coord_flip() +
  labs(title = "Top 10 Locations of Shooting Incidents in NYC",
       x = "Location Type",
       y = "Number of Incidents") +
  theme_minimal()
```



This chart reveals where shootings most often take place. These trends can reflect patterns in public safety, housing, or neighborhood infrastructure and may help city agencies prioritize patrols or interventions.

```
# Extract hour from time of day
shootings_final <- shootings_final %>%
  mutate(hour = hour(occur_time))  # lubridate::hour()

# Plot number of incidents by hour (0-23)
shootings_final %>%
  count(hour) %>%
  ggplot(aes(x = hour, y = n)) +
  geom_line(color = "#d73027", size = 1.2) +
  geom_point(color = "black", size = 2) +
  scale_x_continuous(breaks = 0:23) +
  labs(title = "Shooting Incidents by Hour of Day",
       x = "Hour (24-Hour Format)",
       y = "Number of Incidents") +
  theme_minimal()
```



There is often see a sharp increase in the evening and night (between 8 PM and 2 AM). Less incidents occur in the early morning (3–7 AM), when less people are active.

Conclusion: This analysis of the NYPD Shooting Incident Data highlighted key trends, such as the peak of incidents occurring during late evening and night, with most shootings taking place in public spaces like streets. The findings emphasize the need for targeted interventions in high-risk areas and times.

However, potential biases exist in the data, such as underreporting, geographical discrepancies, and incomplete demographic information. As the analyst, I recognize my personal biases shaped by societal narratives and have mitigated them by focusing on data-driven findings and maintaining objectivity throughout the process.

Further research could expand this analysis to include socio-economic factors and explore the impact of police presence, offering deeper insights into the factors contributing to urban crime.