NYPD Shooting Analysis

JBrien Barcoma

2025-01-23

Introduction

Overview of Report

This report expounds the data analysis, visualization and insights derived from the dataset "NYPD Shooting Incident Data (Historic)" by the New York Police Department.

Objectives of Analysis

Rather than simply exploring the dataset, we aim to purposely have the motivation to identify opportunities to reduce shooting incidents, in addition to obtaining any other insights that may assist or improve government operations in the topic of concern.

Dataset Overview

Data Source

The data was downloaded from the US Government's Open Data site (https://catalog.data.gov/dataset) on 1 January 2025. Incidents from January 2006 to December 2023

Dataset Description

```
df <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv")
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

The dateset consists of 21 columns and 28562 rows.

Below are the 21 column names and their respective data types.

```r
summary(df)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms         Class :character
##  Median : 92711254   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE            X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

```r
generate_summary <- function(df) {
  # Create a summary data frame
  summary_df <- data.frame(
    Column_Name = names(df),                                  # Column name
    Data_Type = I(sapply(df, class)),                        # Include all classes as a list
    Sample_Value = sapply(df, function(col) {
      if (length(na.omit(col)) > 0) na.omit(col)[1] else NA
    }), # First non-NA value
    NA_Count = sapply(df, function(col) sum(is.na(col))), # Number of missing values
    Unique_Values = sapply(df, function(col) length(unique(na.omit(col)))), # Unique non-NA values
    Min_Value = sapply(df, function(col) {
      if (is.numeric(col)) min(col, na.rm = TRUE) else NA
    }), # Min value for numeric columns
    Max_Value = sapply(df, function(col) {
      if (is.numeric(col)) max(col, na.rm = TRUE) else NA
    }), # Max value for numeric columns
    Mean_Value = sapply(df, function(col) {
      if (is.numeric(col)) mean(col, na.rm = TRUE) else NA
    }), # Mean value for numeric columns
    Mode_Value = sapply(df, function(col) {
      if (is.character(col) || is.factor(col)) {
        # Return the most frequent value for categorical columns
        tab <- table(col)
        names(tab)[which.max(tab)]
      } else {
        NA
      }
    }) # Mode for categorical columns
  )

  return(summary_df)
}

generate_summary(df)
```

```
##                                    Column_Name    Data_Type
## INCIDENT_KEY                       INCIDENT_KEY      numeric
## OCCUR_DATE                           OCCUR_DATE    character
## OCCUR_TIME                           OCCUR_TIME hms, dif....
## BORO                                       BORO    character
## LOC_OF_OCCUR_DESC             LOC_OF_OCCUR_DESC    character
## PRECINCT                               PRECINCT      numeric
## JURISDICTION_CODE             JURISDICTION_CODE      numeric
## LOC_CLASSFCTN_DESC           LOC_CLASSFCTN_DESC    character
## LOCATION_DESC                     LOCATION_DESC    character
## STATISTICAL_MURDER_FLAG STATISTICAL_MURDER_FLAG      logical
## PERP_AGE_GROUP                   PERP_AGE_GROUP    character
## PERP_SEX                               PERP_SEX    character
## PERP_RACE                             PERP_RACE    character
## VIC_AGE_GROUP                     VIC_AGE_GROUP    character
## VIC_SEX                                 VIC_SEX    character
## VIC_RACE                               VIC_RACE    character
## X_COORD_CD                           X_COORD_CD      numeric
```

```
## Y_COORD_CD                          Y_COORD_CD      numeric
## Latitude                              Latitude      numeric
## Longitude                            Longitude      numeric
## Lon_Lat                                Lon_Lat    character
##                                                  Sample_Value NA_Count
## INCIDENT_KEY                                        231974218        0
## OCCUR_DATE                                         08/09/2021        0
## OCCUR_TIME                                               3960        0
## BORO                                                    BRONX        0
## LOC_OF_OCCUR_DESC                                     OUTSIDE    25596
## PRECINCT                                                   40        0
## JURISDICTION_CODE                                           0        2
## LOC_CLASSFCTN_DESC                                     STREET    25596
## LOCATION_DESC                                   GROCERY/BODEGA    14977
## STATISTICAL_MURDER_FLAG                                 FALSE        0
## PERP_AGE_GROUP                                          25-44     9344
## PERP_SEX                                                    M     9310
## PERP_RACE                                       WHITE HISPANIC     9310
## VIC_AGE_GROUP                                           18-24        0
## VIC_SEX                                                     M        0
## VIC_RACE                                                BLACK        0
## X_COORD_CD                                            1006343        0
## Y_COORD_CD                                             234270        0
## Latitude                                          40.809673472       59
## Longitude                                      -73.9201927889999       59
## Lon_Lat                POINT (-73.92019278899994 40.80967347200004)       59
##                        Unique_Values     Min_Value     Max_Value    Mean_Value
## INCIDENT_KEY                   22394 9953245.00000  2.797581e+08  1.274058e+08
## OCCUR_DATE                      6095            NA            NA            NA
## OCCUR_TIME                      1423            NA            NA            NA
## BORO                               5            NA            NA            NA
## LOC_OF_OCCUR_DESC                  2            NA            NA            NA
## PRECINCT                          77       1.00000  1.230000e+02  6.549601e+01
## JURISDICTION_CODE                  3       0.00000  2.000000e+00  3.218838e-01
## LOC_CLASSFCTN_DESC                10            NA            NA            NA
## LOCATION_DESC                     40            NA            NA            NA
## STATISTICAL_MURDER_FLAG            2            NA            NA            NA
## PERP_AGE_GROUP                    11            NA            NA            NA
## PERP_SEX                           4            NA            NA            NA
## PERP_RACE                          8            NA            NA            NA
## VIC_AGE_GROUP                      7            NA            NA            NA
## VIC_SEX                            3            NA            NA            NA
## VIC_RACE                           7            NA            NA            NA
## X_COORD_CD                     12706  914928.06250  1.066815e+06  1.009424e+06
## Y_COORD_CD                     12918  125756.71875  2.711277e+05  2.083801e+05
## Latitude                       13385      40.51159  4.091082e+01  4.073857e+01
## Longitude                      13373     -74.24930 -7.370205e+01 -7.390910e+01
## Lon_Lat                        13403            NA            NA            NA
##                                                    Mode_Value
## INCIDENT_KEY                                             <NA>
## OCCUR_DATE                                         07/05/2020
## OCCUR_TIME                                               <NA>
## BORO                                                 BROOKLYN
## LOC_OF_OCCUR_DESC                                     OUTSIDE
```

```
## PRECINCT                                                    <NA>
## JURISDICTION_CODE                                           <NA>
## LOC_CLASSFCTN_DESC                                        STREET
## LOCATION_DESC                                MULTI DWELL - PUBLIC HOUS
## STATISTICAL_MURDER_FLAG                                     <NA>
## PERP_AGE_GROUP                                             18-24
## PERP_SEX                                                       M
## PERP_RACE                                                  BLACK
## VIC_AGE_GROUP                                              25-44
## VIC_SEX                                                        M
## VIC_RACE                                                   BLACK
## X_COORD_CD                                                  <NA>
## Y_COORD_CD                                                  <NA>
## Latitude                                                    <NA>
## Longitude                                                   <NA>
## Lon_Lat                 POINT (-73.88151014499994 40.67141260500006)
```

We will remove INCIDENT_KEY, X_COORD_CD, Y_COORD_CD, and Lon_Lat. We will also rename the columns appropriately.

```
df <- df %>%
  select(-c(
    INCIDENT_KEY,
    LOC_OF_OCCUR_DESC,
    PRECINCT,
    JURISDICTION_CODE,
    LOC_CLASSFCTN_DESC,
    STATISTICAL_MURDER_FLAG,
    X_COORD_CD,
    Y_COORD_CD,
    Lon_Lat))
```

Key Features and Variables

Data Preprocessing

Missing Values and Handling

Let us look at the percentage of missing values for each column, this will inform us on our interpretation of results in our later analysis.

```
#
```

Data Cleaning & Transformation

Dates

We will mutate the given date values and create new columns for Year, Month and Day of the Week and Hour.

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

```r
df <- df %>%
  mutate(date_column = dmy(OCCUR_DATE),
         month = month(OCCUR_DATE),
         year = year(OCCUR_DATE),
         day_of_week = wday(OCCUR_DATE, label = TRUE, abbr = FALSE),
         hour = hour(OCCUR_TIME))
```

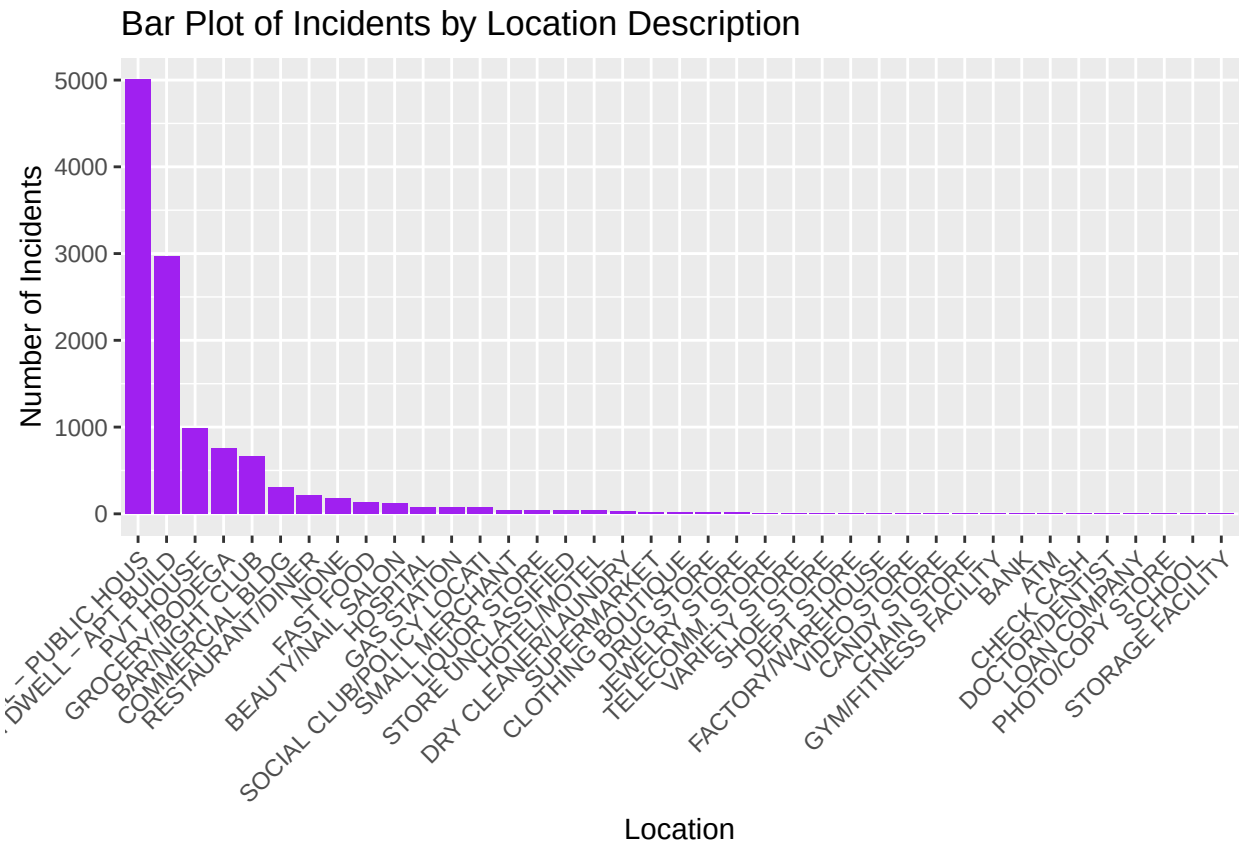
```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `date_column = dmy(OCCUR_DATE)`.
## Caused by warning:
## ! All formats failed to parse. No formats found.
```

Exploratory Data Analysis

Incidents by burough

This graph shows the count of incidents by burough x <- Burough y <- # of Incidents

```r
ggplot(df %>% filter(!is.na(LOCATION_DESC) & !is.null(LOCATION_DESC) & !LOCATION_DESC %in% c("(null)")),
       aes(x = fct_infreq(LOCATION_DESC))) +
      geom_bar(fill = "purple") +
      theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
      labs(title = "Bar Plot of Incidents by Location Description",
           x = "Location",
           y = "Number of Incidents")
```



Incidents by Year by Burough

Let us look at the unique entries. We notice NA, null and NONE entries. We will remove them, but first we shall find out what percentage of incidents have NA, null and NONE entries.

```
#list unique entries
unique(df$LOCATION_DESC)
```

```
##  [1] NA                          "GROCERY/BODEGA"
##  [3] "PVT HOUSE"                 "MULTI DWELL - APT BUILD"
##  [5] "MULTI DWELL - PUBLIC HOUS" "(null)"
##  [7] "BAR/NIGHT CLUB"            "COMMERCIAL BLDG"
##  [9] "FAST FOOD"                 "HOSPITAL"
## [11] "BEAUTY/NAIL SALON"         "LIQUOR STORE"
## [13] "CHAIN STORE"               "RESTAURANT/DINER"
## [15] "SMALL MERCHANT"            "GAS STATION"
## [17] "JEWELRY STORE"             "GYM/FITNESS FACILITY"
## [19] "STORE UNCLASSIFIED"        "SOCIAL CLUB/POLICY LOCATI"
## [21] "DRY CLEANER/LAUNDRY"       "NONE"
## [23] "VIDEO STORE"               "SUPERMARKET"
## [25] "VARIETY STORE"             "FACTORY/WAREHOUSE"
## [27] "CLOTHING BOUTIQUE"         "SHOE STORE"
## [29] "HOTEL/MOTEL"               "CANDY STORE"
## [31] "DEPT STORE"                "BANK"
## [33] "TELECOMM. STORE"           "DRUG STORE"
```

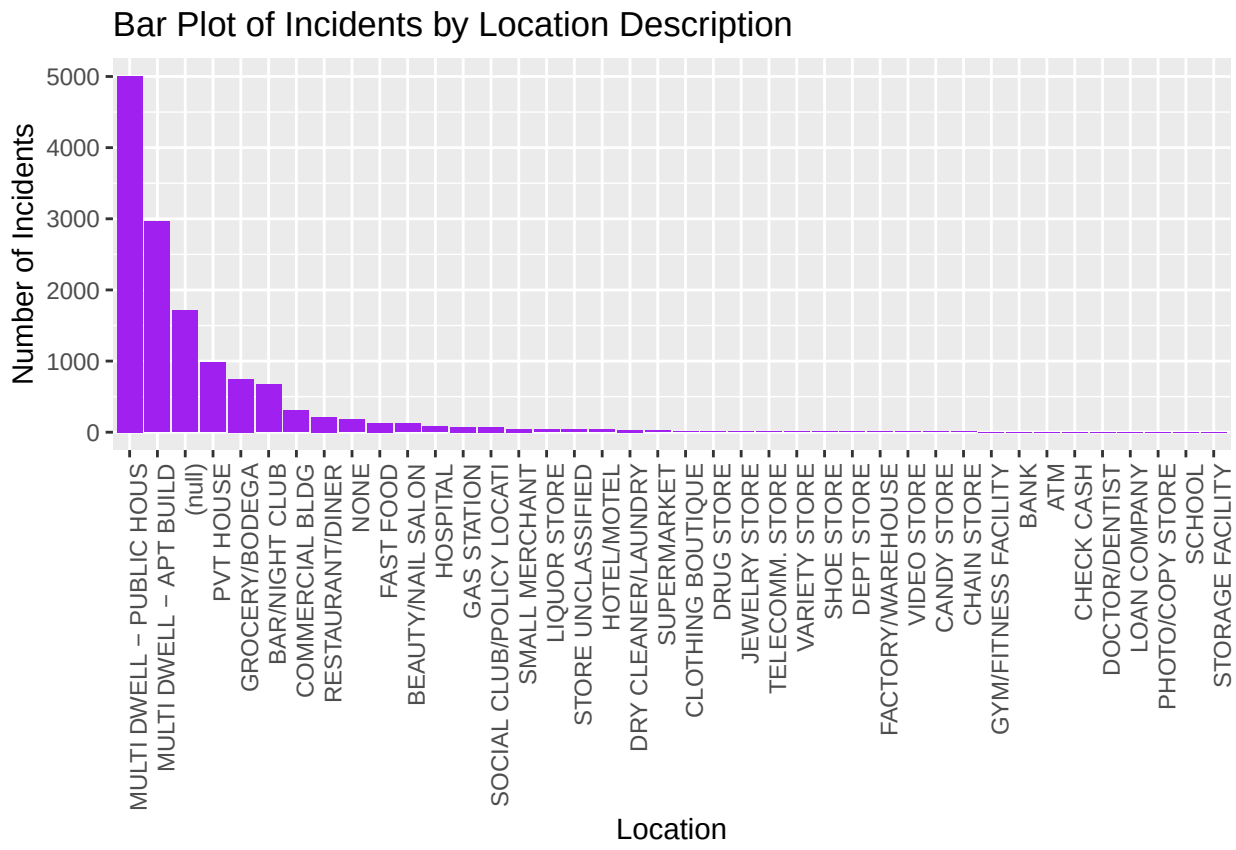
```
## [35] "LOAN COMPANY"              "CHECK CASH"
## [37] "SCHOOL"                    "STORAGE FACILITY"
## [39] "PHOTO/COPY STORE"          "ATM"
## [41] "DOCTOR/DENTIST"
```

```
#convert null and none to NA
#create new column "isNA?"
#stacked bar chart of non-NA/NULL vs NA/NULL
```

```
# Create new Month & Year Column
df <- df %>%
  mutate(date_column = dmy(date_column),
         month = month(date_column),
         year = year(date_column),
         day_of_week = wday(date_column, label = TRUE, abbr = FALSE))  # Extract full weekday name
```

Below is the chart for the number of incidents by location, for all years.

```
ggplot(df %>% filter(!is.na(LOCATION_DESC) & !is.null(LOCATION_DESC)), aes(x = fct_infreq(LOCATION_DESC)
      geom_bar(fill = "purple") +
      theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
      labs(title = "Bar Plot of Incidents by Location Description",
           x = "Location",
           y = "Number of Incidents")
```



Let us look at the top 10 locations and break them down by year.

Incidents by Month by Year by Burough

Data Visualization & Insights

Gender Ratio by Location

Map / Heat Map

Plot 3

Conclusion

References

Session Info

```
sessionInfo()
```

```
## R version 4.4.2 (2024-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 26100)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: Asia/Singapore
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.4 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.5     tidyr_1.3.1     tibble_3.2.1
##  [9] ggplot2_3.5.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.5.0.1       gtable_0.3.6      crayon_1.5.3      compiler_4.4.2
##  [5] tidyselect_1.2.1  parallel_4.4.2    scales_1.3.0      yaml_2.3.10
##  [9] fastmap_1.2.0     R6_2.5.1          labeling_0.4.3    generics_0.1.3
## [13] curl_6.1.0        knitr_1.49        munsell_0.5.1     pillar_1.10.1
## [17] tzdb_0.4.0        rlang_1.1.5       stringi_1.8.4     xfun_0.50
## [21] bit64_4.6.0-1     timechange_0.3.0  cli_3.6.3         withr_3.0.2
## [25] magrittr_2.0.3    digest_0.6.37     grid_4.4.2        vroom_1.6.5
## [29] rstudioapi_0.17.1 hms_1.1.3         lifecycle_1.0.4   vctrs_0.6.5
```

```
## [33] evaluate_1.0.3    glue_1.8.0        farver_2.1.2      colorspace_2.1-1
## [37] rmarkdown_2.29    tools_4.4.2       pkgconfig_2.0.3   htmltools_0.5.8.1
```

Resources