NYC Shooting Data Analysis

K. Katzmann

2024-07-31

Introduction

This analysis explores the NYPD Shooting Incident Data (Historical) dataset from Data.gov. The dataset contains information on shooting incidents reported to the NYPD from 2006 to the end of the previous calendar year. It includes the date, time and location of each reported incident, as well as the age, sex and race of the victim and perpetrator.

This analysis aims to identify temporal and spatial trends in shooting incidents in NYC and to understand the availability of perpetrator information in these incidents. Can trends in the data help the NYPD better allocate resources to prevent or solve future incidents?

```
# load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)

# set theme for ggplot
theme_set(theme_minimal())
```

Importing Data

We'll begin the analysis by importing the data from Data.gov and inspecting the first few rows to understand the structure of the dataset. After looking at the data, we noticed that missing values are coded in multiple ways. We'll address this by including them in the `na` parameter of `read_csv`.

```
# import data as a csv from Data.gov
csv_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

# after inspecting the data, we see that there are some missing values that are not coded as NA
nypd_shooting <- read_csv(csv_url, na = c("", "(null)", "UNKNOWN"))
```

```
## 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.
```

```
# take a glimpse at the data
nypd_shooting %>%
  glimpse()
```

```
## Rows: 28,562
## Columns: 21
## $ INCIDENT_KEY            <dbl> 244608249, 247542571, 84967535, 202853370, 270~
## $ OCCUR_DATE              <chr> "05/05/2022", "07/04/2022", "05/27/2012", "09/~
## $ OCCUR_TIME              <time> 00:10:00, 22:20:00, 19:35:00, 21:00:00, 21:00~
## $ BORO                    <chr> "MANHATTAN", "BRONX", "QUEENS", "BRONX", "BROO~
## $ LOC_OF_OCCUR_DESC       <chr> "INSIDE", "OUTSIDE", NA, NA, NA, NA, NA, NA, N~
## $ PRECINCT                <dbl> 14, 48, 103, 42, 83, 23, 113, 77, 48, 49, 73, ~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ LOC_CLASSFCTN_DESC      <chr> "COMMERCIAL", "STREET", NA, NA, NA, NA, NA, NA~
## $ LOCATION_DESC           <chr> "VIDEO STORE", NA, NA, NA, NA, "MULTI DWELL - ~
## $ STATISTICAL_MURDER_FLAG <lgl> TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, TRUE, ~
## $ PERP_AGE_GROUP          <chr> "25-44", NA, NA, "25-44", "25-44", NA, NA, NA,~
## $ PERP_SEX                <chr> "M", NA, NA, "M", "M", NA, NA, NA, NA, "M", NA~
## $ PERP_RACE               <chr> "BLACK", NA, NA, NA, "BLACK", NA, NA, NA, NA, ~
## $ VIC_AGE_GROUP           <chr> "25-44", "18-24", "18-24", "25-44", "25-44", "~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 986050, 1016802, 1048632, 1014493, 1009149, 99~
## $ Y_COORD_CD              <dbl> 214231.0, 250581.0, 198262.0, 242565.0, 190104~
## $ Latitude                <dbl> 40.75469, 40.85440, 40.71063, 40.83242, 40.688~
## $ Longitude               <dbl> -73.99350, -73.88233, -73.76777, -73.89071, -7~
## $ Lon_Lat                 <chr> "POINT (-73.9935 40.754692)", "POINT (-73.8823~
```

Tidying and Transforming Data

We'll tidy and transform the data by removing unneeded columns, converting data types and creating new columns to be used in our analysis.

```
# remove columns we don't need
nypd_shooting <- nypd_shooting %>%
  select(-c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))

# convert date columns to date format
nypd_shooting <- nypd_shooting %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME))
```

```r
# convert numeric columns to factors
nypd_shooting <- nypd_shooting %>%
  mutate(PRECINCT = as_factor(PRECINCT),
         JURISDICTION_CODE = as_factor(JURISDICTION_CODE))

# create new columns for year, month, day of week, and hour
nypd_shooting <- nypd_shooting %>%
  mutate(OCCUR_YEAR = year(OCCUR_DATE),
         OCCUR_MONTH = month(OCCUR_DATE, label = TRUE),
         OCCUR_WDAY = wday(OCCUR_DATE, label = TRUE),
         OCCUR_HOUR = as_factor(hour(OCCUR_TIME)),
         OCCUR_TIME_OF_DAY = case_when(hour(OCCUR_TIME) >= 5 & hour(OCCUR_TIME) < 12 ~ "Morning",
                                       hour(OCCUR_TIME) >= 12 & hour(OCCUR_TIME) < 17 ~ "Afternoon",
                                       hour(OCCUR_TIME) >= 17 & hour(OCCUR_TIME) < 21 ~ "Evening",
                                       TRUE ~ "Night"))

# create column for perpetrator information
nypd_shooting <- nypd_shooting %>%
  mutate(PERP_INFO = !(is.na(PERP_AGE_GROUP) & is.na(PERP_SEX) & is.na(PERP_RACE)))

summary(nypd_shooting)
```

```
##   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           LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE
##  Length:28562       Length:28562       75     : 1628   0   :23923
##  Class :character   Class :character   73     : 1500   1   :   81
##  Mode  :character   Mode  :character   67     : 1259   2   : 4556
##                                        44     : 1076   NA's:    2
##                                        79     : 1045
##                                        47     : 1006
##                                        (Other):21048
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:28562       Length:28562       Mode :logical
##  Class :character   Class :character   FALSE:23036
##  Mode  :character   Mode  :character   TRUE :5526
##
##
##
##
##  PERP_AGE_GROUP       PERP_SEX           PERP_RACE         VIC_AGE_GROUP
##  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_SEX            VIC_RACE           OCCUR_YEAR    OCCUR_MONTH
##  Length:28562       Length:28562       Min.   :2006   Jul    : 3390
##  Class :character   Class :character   1st Qu.:2009   Aug    : 3264
##  Mode  :character   Mode  :character   Median :2013   Jun    : 2959
##                                        Mean   :2014   May    : 2682
##                                        3rd Qu.:2019   Sep    : 2677
##                                        Max.   :2023   Oct    : 2378
##                                                       (Other):11212
##  OCCUR_WDAY   OCCUR_HOUR    OCCUR_TIME_OF_DAY  PERP_INFO
##  Sun:5669   23     : 2397   Length:28562       Mode :logical
##  Mon:4062   0      : 2267   Class :character   FALSE:10451
##  Tue:3331   22     : 2264   Mode  :character   TRUE :18111
##  Wed:3145   1      : 2142
##  Thu:3169   21     : 2074
##  Fri:3759   2      : 1878
##  Sat:5427   (Other):15540
```

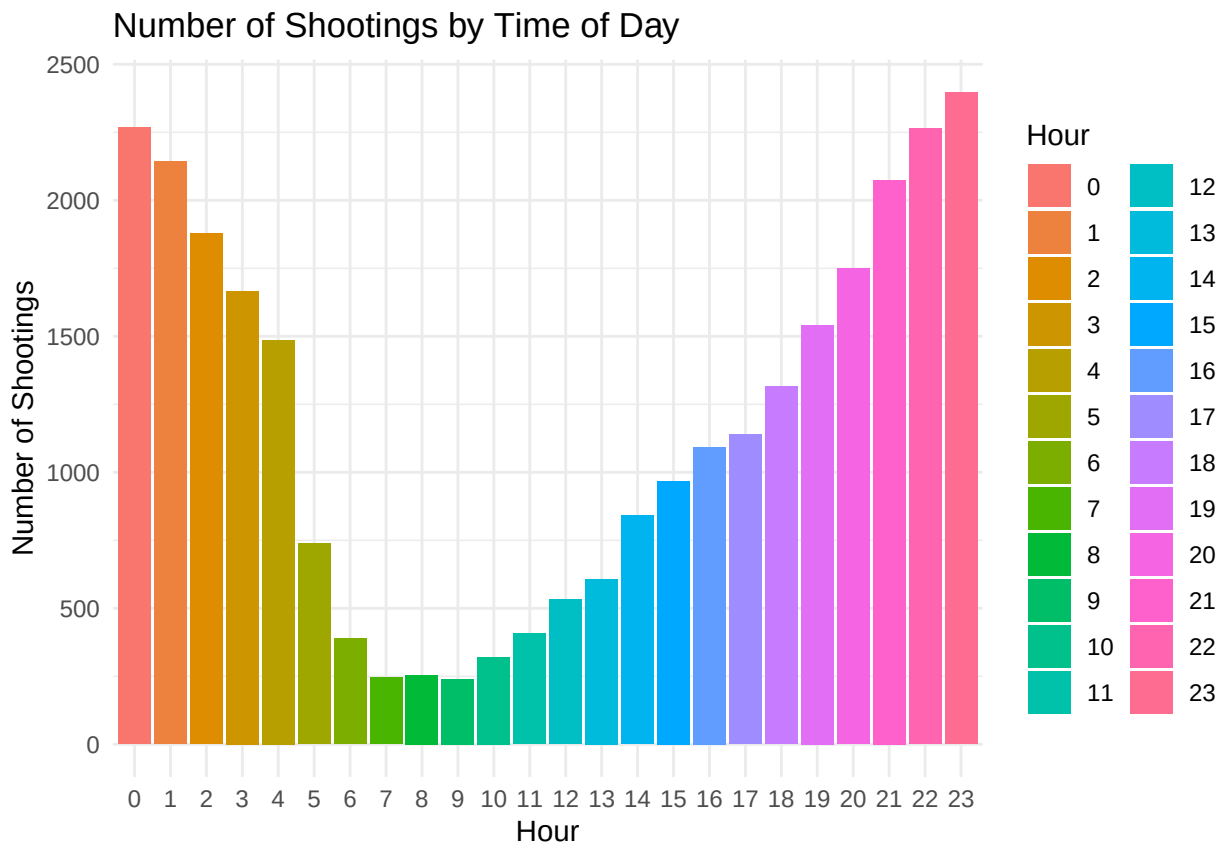
Exploratory Data Analysis

Temporal Trends

We'll begin the exploratory analysis by checking for a few temporal trends in the data.

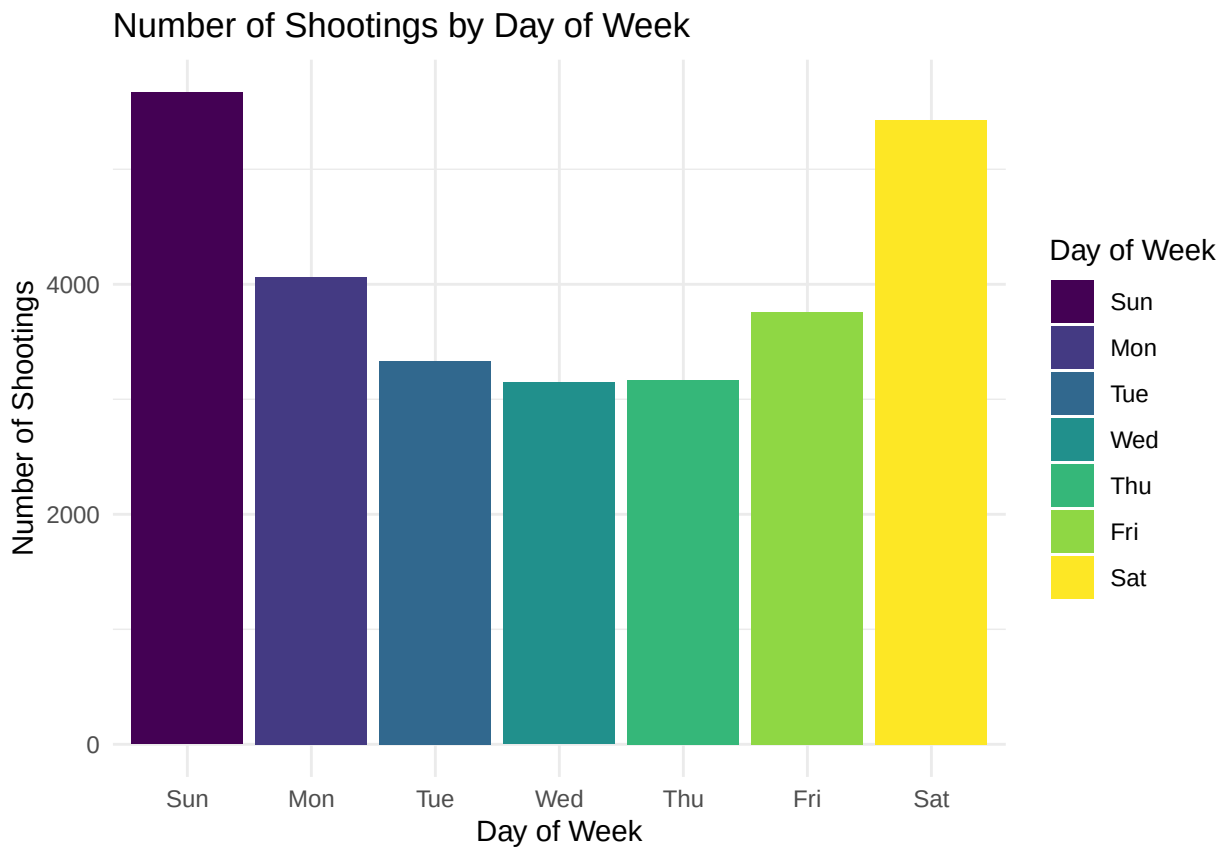
First, let's see if the number of shootings varies by time of day.

```
# plot number of shootings by time of day
nypd_shooting %>%
  group_by(OCCUR_HOUR) %>%
  summarise(count_incidents = n()) %>%
  ggplot(aes(x = OCCUR_HOUR, y = count_incidents, fill = OCCUR_HOUR)) +
  geom_col() +
  labs(title = "Number of Shootings by Time of Day",
       x = "Hour",
       y = "Number of Shootings",
       fill = "Hour")
```



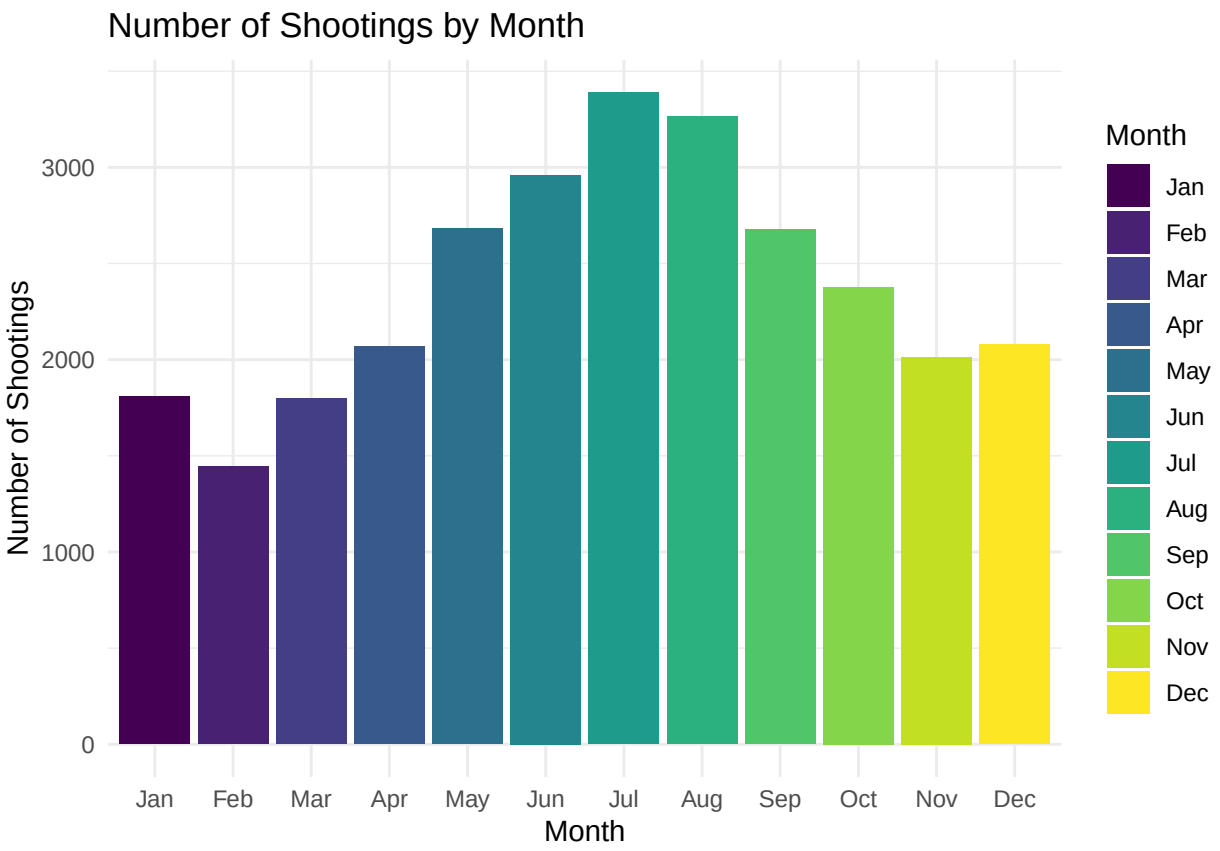
Next, let's see if the number of shootings varies by day of the week.

```
# plot number of shootings by day of week
nypd_shooting %>%
  group_by(OCCUR_WDAY) %>%
  summarise(count_incidents = n()) %>%
  ggplot(aes(x = OCCUR_WDAY, y = count_incidents, fill = OCCUR_WDAY)) +
  geom_col() +
  labs(title = "Number of Shootings by Day of Week",
       x = "Day of Week",
       y = "Number of Shootings",
       fill = "Day of Week")
```



Finally, let's see if the number of shootings varies by month.

```
# plot number of shootings by month
nypd_shooting %>%
  group_by(OCCUR_MONTH) %>%
  summarise(count_incidents = n()) %>%
  ggplot(aes(x = OCCUR_MONTH, y = count_incidents, fill = OCCUR_MONTH)) +
  geom_col() +
  labs(title = "Number of Shootings by Month",
       x = "Month",
       y = "Number of Shootings",
       fill = "Month")
```

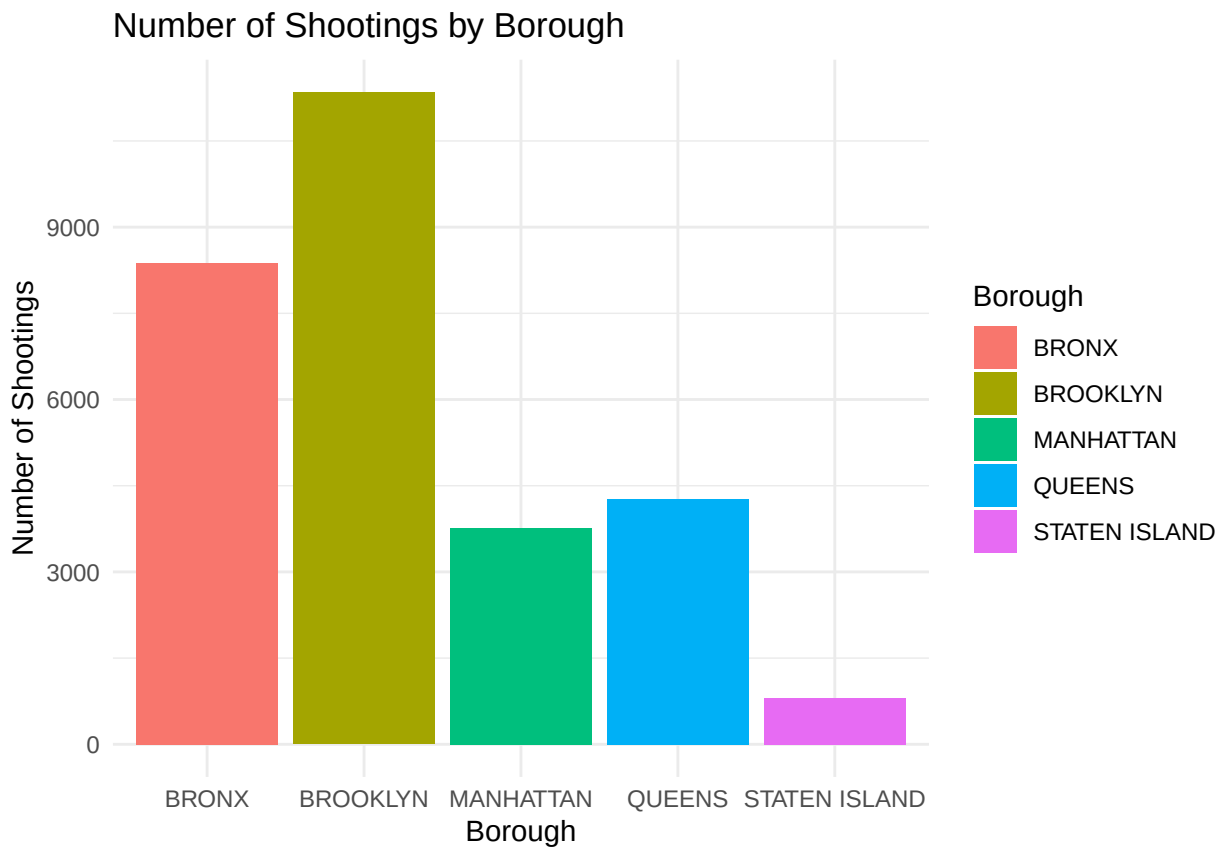


Spatial Trends

Now that we've looked at some temporal trends, let's see if there are any patterns in the data related to the location of the shootings.

First, let's see if the number of shootings varies by borough.

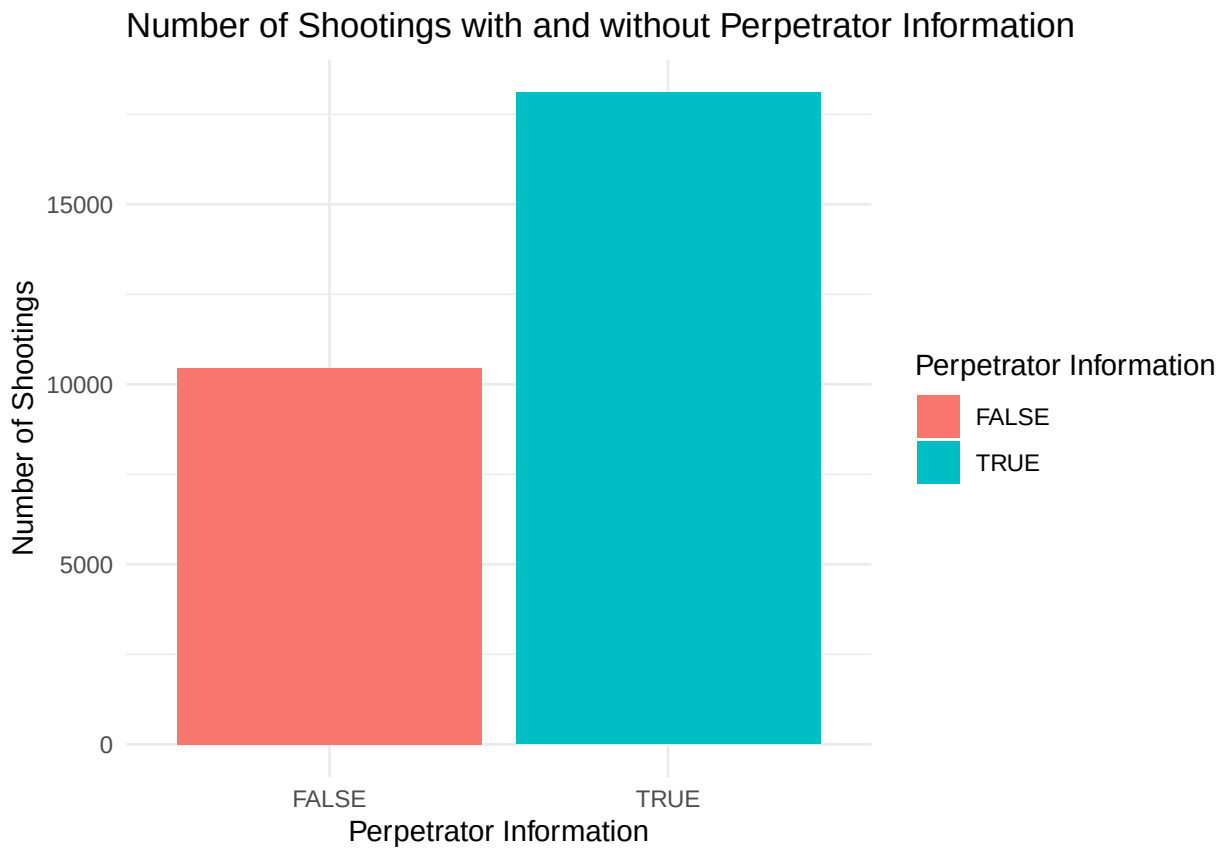
```
# plot number of shootings by borough
nypd_shooting %>%
  group_by(BORO) %>%
  summarise(count_incidents = n()) %>%
  ggplot(aes(x = BORO, y = count_incidents, fill = BORO)) +
  geom_col() +
  labs(title = "Number of Shootings by Borough",
       x = "Borough",
       y = "Number of Shootings",
       fill = "Borough")
```



Perpetrator Information

Now, let's see how many shootings have information about the perpetrator.

```
# plot number of shootings with and without perpetrator information
nypd_shooting %>%
  group_by(PERP_INFO) %>%
  summarise(count_incidents = n()) %>%
  ggplot(aes(x = PERP_INFO, y = count_incidents, fill = PERP_INFO)) +
  geom_col() +
  labs(title = "Number of Shootings with and without Perpetrator Information",
       x = "Perpetrator Information",
       y = "Number of Shootings",
       fill = "Perpetrator Information")
```



Modeling

Now that we've explored the data a bit, let's build a linear model to better understand the relationship between some of these variables.

Let's see if we can predict the availability of perpetrator information based on the time of the day and the borough where the shooting occurred.

```
model <- lm(PERP_INFO ~ OCCUR_TIME_OF_DAY + BORO, data = nypd_shooting)

summary(model)
```

```
##
## Call:
## lm(formula = PERP_INFO ~ OCCUR_TIME_OF_DAY + BORO, data = nypd_shooting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8804 -0.6019  0.3163  0.3751  0.4569
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.748503   0.008747  85.573  < 2e-16 ***
## OCCUR_TIME_OF_DAYEvening   -0.064805   0.009794  -6.617 3.73e-11 ***
## OCCUR_TIME_OF_DAYMorning   -0.040250   0.011997  -3.355 0.000795 ***
```

```
## OCCUR_TIME_OF_DAYNight   -0.123650   0.008392 -14.734  < 2e-16 ***
## BOROBROOKLYN             -0.081794   0.006876 -11.896  < 2e-16 ***
## BOROMANHATTAN             0.032239   0.009363   3.443 0.000575 ***
## BOROQUEENS               -0.021179   0.008973  -2.360 0.018266 *
## BOROSTATEN ISLAND         0.131874   0.017582   7.500 6.55e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 28554 degrees of freedom
## Multiple R-squared:  0.01961,    Adjusted R-squared:  0.01937
## F-statistic:  81.6 on 7 and 28554 DF,  p-value: < 2.2e-16
```

Bias

Before discussing any insights and drawing conclusions, it is important to recognize the bias that exists in this analysis.

One source of bias is reporting bias. It is likely that not all shootings have been reported, and some might be under-reported in certain areas or at certain times.

Another source of bias is in the analysis itself. This analysis focuses primarily on the temporal patterns of reported shooting incidents and the availability of perpetrator information. Although the availability of perpetrator information might seem important for identifying the perpetrator, it's uncertain just how important it is, and too much weight might be placed on it here.

This analysis also avoids any exploration of demographic information, which could be useful in understanding the underlying causes of these incidents.

Conclusion

There are a few potentially valuable insights we've gained by analyzing the NYPD Shooting Incident Data (Historical) dataset.

By digging into the temporal component of these incidents, we noticed that shootings occur more often at night, on the weekends, and during Summer months. The police department should be extra vigilant during these times!

We also learned that Brooklyn has the highest frequency of shooting incidents reported. However, that might be explained by Brooklyn having a larger population. We should adjust for the population of each borough in a follow-up analysis.

We also noticed that not all incidents had perpetrator information available. Since information on the perpetrator could be very useful in identifying and arresting the perpetrator, we decided to explore this variable a bit further. We modeled the relationship between the availability of perpetrator information and the time of day and borough where the shooting occurred. We found that shooting incidents at night are less likely to have perpetrator information available (estimate = -0.12, p-value < 0.001). We also found that shooting incidents in Brooklyn are less likely to have perpetrator information available (estimate = -0.08, p-value < 0.001). However, this model only explains a small portion of the variance in perpetrator information avialable on shooting incidents (R-squared = 0.02).

The insights gained in this analysis could be very useful to the NYPD in allocating resources in order to prevent or solve future shooting incidents.

Session Info

```
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin23.4.0
## Running under: macOS Sonoma 14.6
##
## Matrix products: default
## BLAS:   /opt/homebrew/Cellar/openblas/0.3.27/lib/libopenblasp-r0.3.27.dylib
## LAPACK: /opt/homebrew/Cellar/r/4.4.0_1/lib/R/lib/libRlapack.dylib;  LAPACK version 3.12.0
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 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] utf8_1.2.4        generics_0.1.3    stringi_1.8.4     hms_1.1.3
##  [5] digest_0.6.35     magrittr_2.0.3    evaluate_0.23     grid_4.4.0
##  [9] timechange_0.3.0  fastmap_1.1.1     fansi_1.0.6       viridisLite_0.4.2
## [13] scales_1.3.0      cli_3.6.2         rlang_1.1.3       crayon_1.5.2
## [17] bit64_4.0.5       munsell_0.5.1     withr_3.0.0       yaml_2.3.8
## [21] tools_4.4.0       parallel_4.4.0    tzdb_0.4.0        colorspace_2.1-0
## [25] curl_5.2.1        vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4
## [29] bit_4.0.5         vroom_1.6.5       pkgconfig_2.0.3   pillar_1.9.0
## [33] gtable_0.3.5      glue_1.7.0        xfun_0.43         tidyselect_1.2.1
## [37] highr_0.10        rstudioapi_0.16.0 knitr_1.46        farver_2.1.2
## [41] htmltools_0.5.8.1 rmarkdown_2.26    labeling_0.4.3    compiler_4.4.0
```