NYPD Shooting Data

Kendall

2025-11-12

NYPD Shooting Data

#Analysis of shooting trends per time of year and area. The data file contains information of every shooting in NY since 2006, including location, date, perpetrator age, victim age, etc.

#Read in and clean the file data

```
#read in the data file
url_in <- "https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic"

file_name <- c("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")

data <- read.csv(file_name)

#Since this analysis will focus on location and time of the shooting, we will clean up data by removing columns not related to location or
data = data[, -c(1,5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21)]

#convert OCCUR_DATE column to date format
data$OCCUR_DATE = as.Date(data$OCCUR_DATE, format = "%m/%d/%Y")
#convert OCCUR_TIME column to time format
library(hms)
```

```
## Warning: package 'hms' was built under R version 4.5.2
```

```
##
## Attaching package: 'hms'
## The following object is masked from 'package:lubridate':
##
##     hms
```

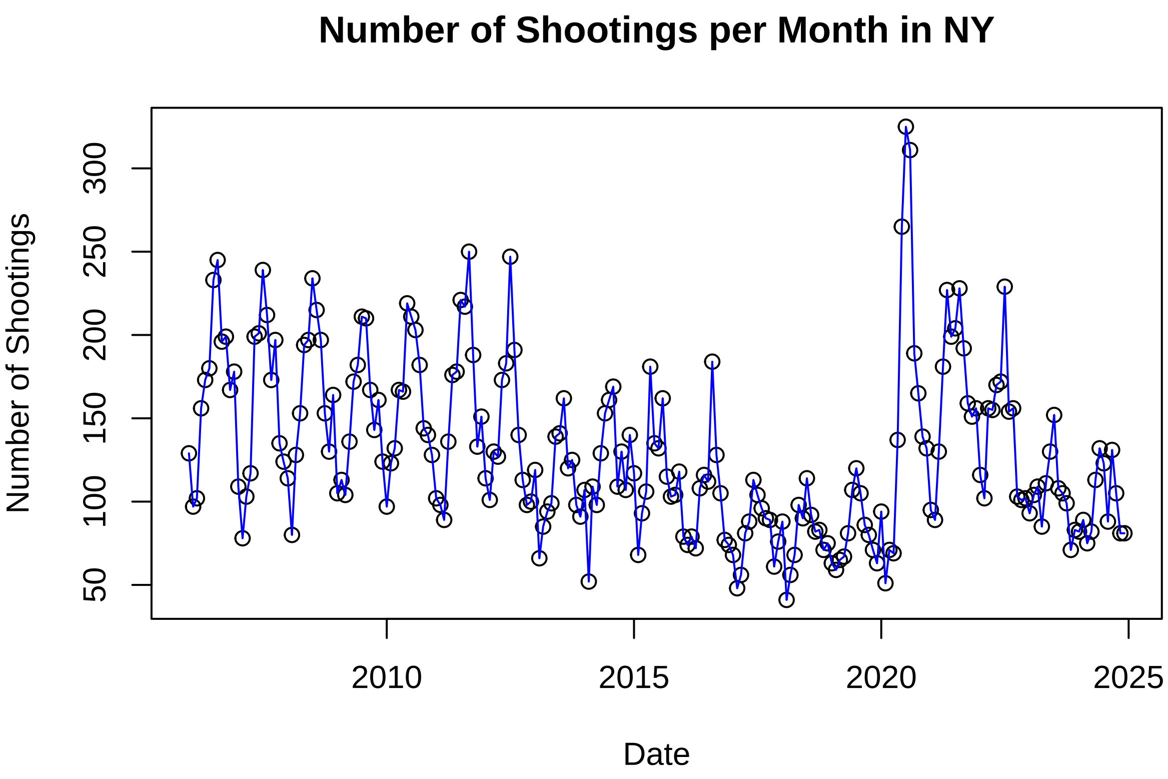
```r
data$OCCUR_TIME = as_hms(data$OCCUR_TIME)
#print summary of data
summary(data)
```

```
##    OCCUR_DATE           OCCUR_TIME                 BORO          
##  Min.   :2006-01-01   Min.   :00:00:00.000000   Length:29744      
##  1st Qu.:2009-10-29   1st Qu.:03:30:45.000000   Class :character  
##  Median :2014-03-25   Median :15:15:00.000000   Mode  :character  
##  Mean   :2014-10-31   Mean   :12:46:10.874798                     
##  3rd Qu.:2020-06-29   3rd Qu.:20:44:00.000000                     
##  Max.   :2024-12-31   Max.   :23:59:00.000000                     
##     PRECINCT     
##  Min.   :  1.00  
##  1st Qu.: 44.00  
##  Median : 67.00  
##  Mean   : 65.23  
##  3rd Qu.: 81.00  
##  Max.   :123.00
```

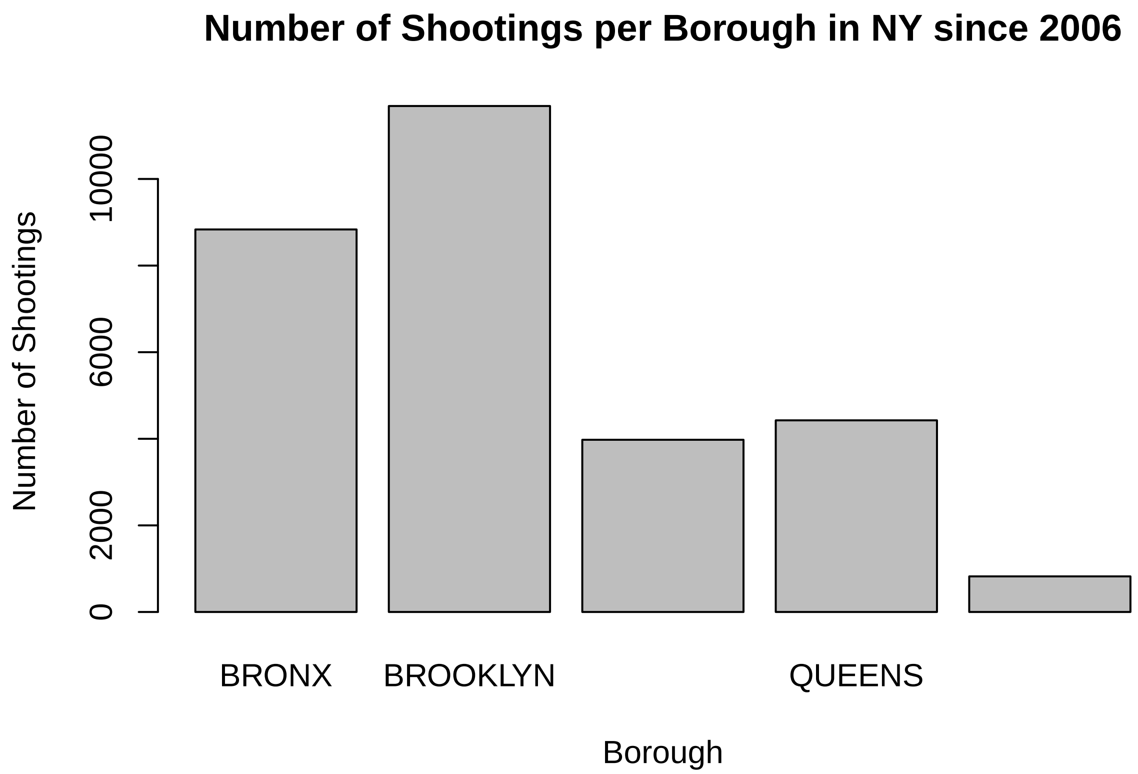
Data Visualization

```r
#plot the number of shootings per month in NY
library(lubridate)
#summarize number of shootings by month and year
sumdate = table(floor_date(data$OCCUR_DATE, "month"))
#change table to data frame
sumdatedf = as.data.frame.table(sumdate)
#add column names
colnames(sumdatedf) = c("Date", "Occurrences")
#truncate date off of the date column as we only will look at month and year
library(lubridate)
sumdatedf$Date = as.Date(sumdatedf$Date, format = "%Y-%m-%d")
```

```
#plot the data
plot(sumdatedf$Date, sumdatedf$Occurrences, xlab = "Date", ylab = "Number of Shootings", main = "Number of Shootings per Month in NY")
lines(sumdatedf$Date, sumdatedf$Occurrences, col = "blue")
```



```
#plot the number of shootings per borough of NY
sumboro = table(data$BORO)
barplot(sumboro, xlab = "Borough", ylab = "Number of Shootings", main = "Number of Shootings per Borough in NY since 2006")
```



```
#when exporting to PDF, not all of the y-labels will print
```

Data Analysis

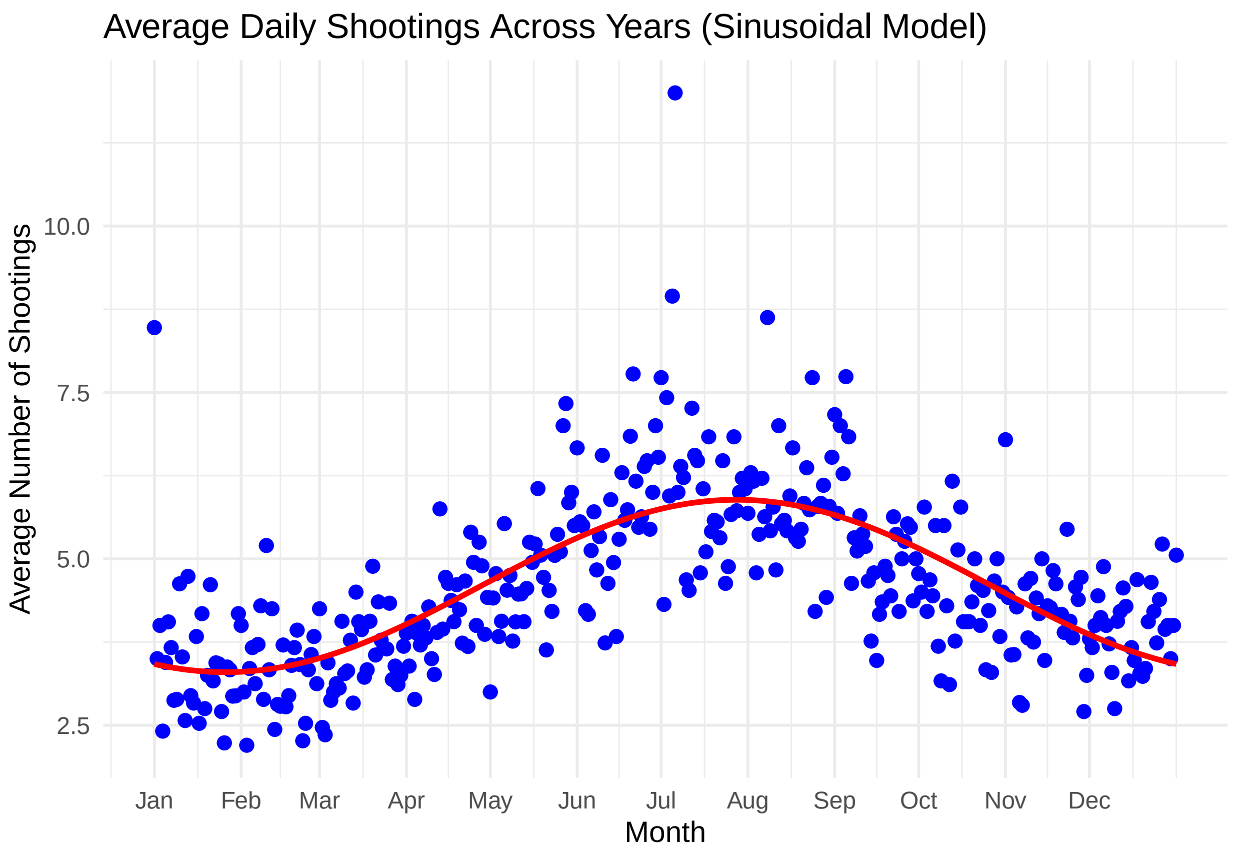
```
#Count shootings per time
# Count shootings per day
shootings_by_day <- data %>%
  group_by(OCCUR_DATE) %>%
  summarise(num_shootings = n())
```

```r
# Extract month and day to create a "day-of-year" ignoring the year
shootings_by_day <- shootings_by_day %>%
  mutate(
    month = month(OCCUR_DATE),
    day = day(OCCUR_DATE),
    day_of_year = yday(as.Date(paste("2000", month, day, sep = "-")))  # arbitrary non-leap year
  )

# Summarize across all years for each day-of-year
shootings_by_doy <- shootings_by_day %>%
  group_by(day_of_year) %>%
  summarise(num_shootings = mean(num_shootings))  # average over years

# Fit sinusoidal model
model_sin <- lm(num_shootings ~ sin(2 * pi * day_of_year / 365) +
                                cos(2 * pi * day_of_year / 365),
                data = shootings_by_doy)

summary(model_sin)
```

```
##
## Call:
## lm(formula = num_shootings ~ sin(2 * pi * day_of_year/365) +
##     cos(2 * pi * day_of_year/365), data = shootings_by_doy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9613 -0.5489 -0.0585  0.4131  6.2033
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.59264    0.04713  97.456  < 2e-16 ***
## sin(2 * pi * day_of_year/365) -0.56833    0.06674  -8.516 4.44e-16 ***
## cos(2 * pi * day_of_year/365) -1.16357    0.06655 -17.483  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9015 on 363 degrees of freedom
## Multiple R-squared:  0.5103, Adjusted R-squared:  0.5076
## F-statistic: 189.1 on 2 and 363 DF,  p-value: < 2.2e-16
```

```
# Add predictions
shootings_by_doy <- shootings_by_doy %>%
  mutate(predicted = predict(model_sin))

# Plot actual vs predicted, using day-of-year as x-axis
ggplot(shootings_by_doy, aes(x = day_of_year)) +
  geom_point(aes(y = num_shootings), color = "blue", size = 2) +
  geom_line(aes(y = predicted), color = "red", size = 1) +
  scale_x_continuous(
    breaks = c(1, 32, 60, 91, 121, 152, 182, 213, 244, 274, 305, 335),
    labels = month.abb
  ) +
  labs(
    title = "Average Daily Shootings Across Years (Sinusoidal Model)",
    x = "Month",
    y = "Average Number of Shootings"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Conclusion

In this analysis I looked at three things: whether there were any trends in number of shootings depending on the time of year, the number of shootings per borough, and analyzed the number of shootings per season and tried modeling this analysis.

The first visualization “Number of Shootings per Month in NY” shows how the seasons can affect the number of shootings. My personal bias was that there would be an increase in shootings in the summer months due to hotter weather. There does look like there is a spike in shootings every summer, however more data/analysis would be needed before I can draw a conclusion if it is significant. For example, it would be helpful to have temperature data for each month.

The second visualization, “Number of Shootings per Borough in NY since 2006” shows that the Brooklyn has the largest number of shootings compared to other boroughs. A conclusion could be drawn from this graph that the Brooklyn is the most unsafe borough to live in, but this graph does not take into account population density or shootings per capita, so this graph is biased and can lead to inaccurate conclusions.

The analysis I chose was to summarize the number of shootings per month/date combination. To do this I summarized the data by day, not taking into account the year and summed the number of shootings that occurred per that date each year. My theory was that shootings would increase in hotter months, so I modeled this using a sinusoidal model. I then mapped the data from the model with the actual data. The R-squared value for this model was 0.5103, so there is additional variables at play, however, overall, it does seem that the data follows some sort of sinusoidal trend.

```r
sessionInfo()
```

```
## R version 4.5.1 (2025-06-13 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 26100)
## 
## Matrix products: default
##   LAPACK version 3.12.1
## 
## 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: America/Denver
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] hms_1.1.4       lubridate_1.9.4 forcats_1.0.1   stringr_1.6.0  
##  [5] dplyr_1.1.4     purrr_1.2.0     readr_2.1.6     tidyr_1.3.1    
##  [9] tibble_3.3.0    ggplot2_4.0.1   tidyverse_2.0.0
## 
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.6       crayon_1.5.3       compiler_4.5.1     tidyselect_1.2.1  
##  [5] scales_1.4.0       yaml_2.3.10        fastmap_1.2.0      R6_2.6.1          
##  [9] labeling_0.4.3     generics_0.1.4     knitr_1.50         pillar_1.11.1     
## [13] RColorBrewer_1.1-3 tzdb_0.5.0         rlang_1.1.6        stringi_1.8.7
## [17] xfun_0.54          S7_0.2.1           timechange_0.3.0   cli_3.6.5
## [21] withr_3.0.2        magrittr_2.0.4     digest_0.6.37      grid_4.5.1
## [25] rstudioapi_0.17.1  lifecycle_1.0.4    vctrs_0.6.5        evaluate_1.0.5
## [29] glue_1.8.0         farver_2.1.2       rmarkdown_2.30     tools_4.5.1
## [33] pkgconfig_2.0.3    htmltools_0.5.8.1
```