NYPD Shooting Incidents

Adam Krull

2024-04-28

About the data

The data is every recorded shooting incident in New York City since 2006. It includes information about each shooting incident, including the date, time, and borough. My analysis will only focus on these bits of information, so they will be the only features displayed for the duration of the report. I decided to display the first few rows of the raw data so you can see what it looks like.

I had difficulties acquiring the data directly from the source into this RMarkdown document, because the connection kept timing out after 60 seconds. In order to successfully run this file on your own computer, you will need to visit this link and download the data as a csv file named "nypd_data.csv". Make sure your current working directory in RStudio is pointed to the folder containing this csv file.

```
data <- read.csv("nypd_data.csv")
head(data, n = c(5, 4))
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME      BORO
## 1    244608249 05/05/2022   00:10:00 MANHATTAN
## 2    247542571 07/04/2022   22:20:00     BRONX
## 3     84967535 05/27/2012   19:35:00    QUEENS
## 4    202853370 09/24/2019   21:00:00     BRONX
## 5     27078636 02/25/2007   21:00:00  BROOKLYN
```

Data summary

I will summarize the data to check for appropriate data types and irregularities. If I encounter strange or missing values, I will investigate them and determine how to handle them.

```
summary(data[1:4])
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:28562       Length:28562       Length:28562
##  1st Qu.: 65439914   Class :character   Class :character   Class :character
##  Median : 92711254   Mode  :character   Mode  :character   Mode  :character
##  Mean   :127405824
##  3rd Qu.:203131993
##  Max.   :279758069
```

Cleaning up the data

This dataset has a lot to offer for detailed analysis. The scope of this project asks for a couple visualizations and accompanying analysis. I will focus on the following attributes for my analysis: date, time of day, and borough. I will create a new dataframe that only contains the relevant columns. I noticed that OCCUR_DATE is currently a character field: I will recast it as a datetime object. I also noticed that OCCUR_TIME is currently a character field. I have plans to extract the first two characters as an "hour of the day" field and create three categories from the resulting information: morning, afternoon, and night.

```
subset <- data %>% select(OCCUR_DATE, OCCUR_TIME, BORO)
subset <- subset %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
subset$OCCUR_MONTH <- format(subset$OCCUR_DATE, "%m")
subset$OCCUR_HOUR <- substr(subset$OCCUR_TIME, start=1, stop=2)
subset <- transform(subset, OCCUR_HOUR = as.numeric(OCCUR_HOUR))
subset <- subset %>% mutate(TIME_OF_DAY = case_when(
  (OCCUR_HOUR > 4) & (OCCUR_HOUR < 12) ~ 'Morning',
  (OCCUR_HOUR > 11) & (OCCUR_HOUR < 20) ~ 'Afternoon',
  TRUE ~ 'Night'
  )
)
subset <- subset %>% select(OCCUR_MONTH, TIME_OF_DAY, BORO)
head(subset, n = 5)
```

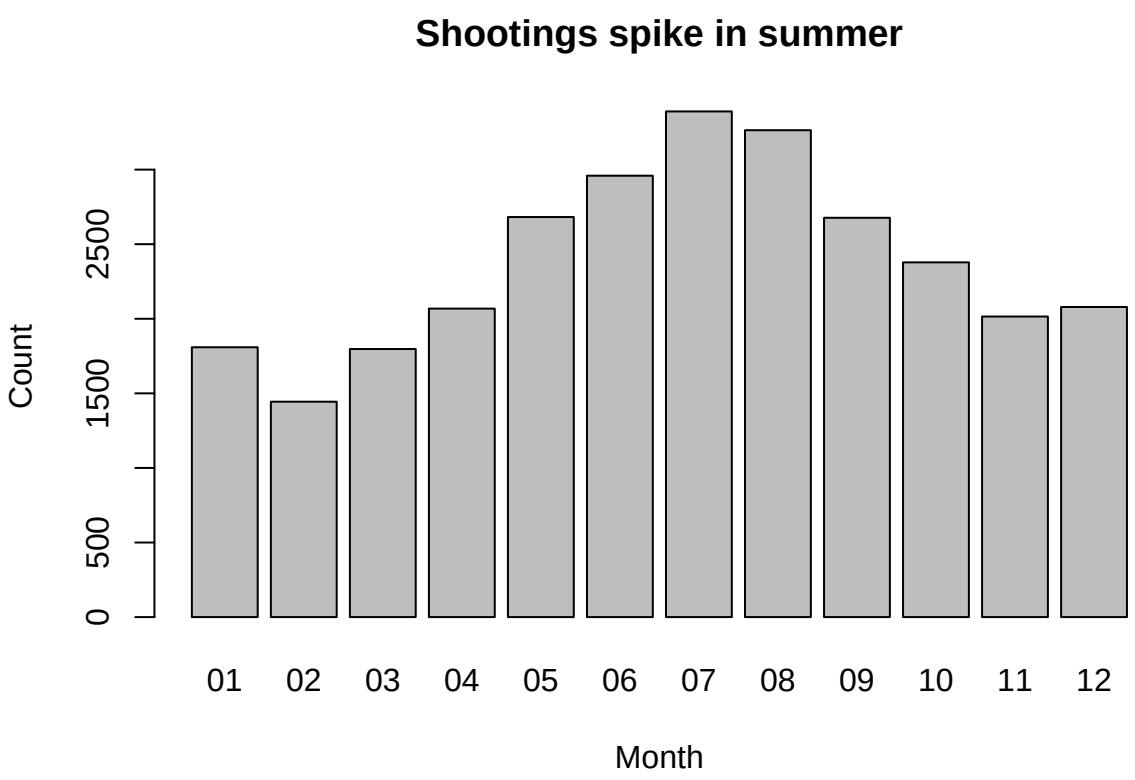
```
##   OCCUR_MONTH TIME_OF_DAY      BORO
## 1          05       Night MANHATTAN
## 2          07       Night     BRONX
## 3          05   Afternoon    QUEENS
## 4          09       Night     BRONX
## 5          02       Night  BROOKLYN
```

Analyzing my data

Now that the data is ready to go, I have a few questions I will attempt to answer with the aid of some visualizations.

Month Does the month of the year affect the number of shootings?

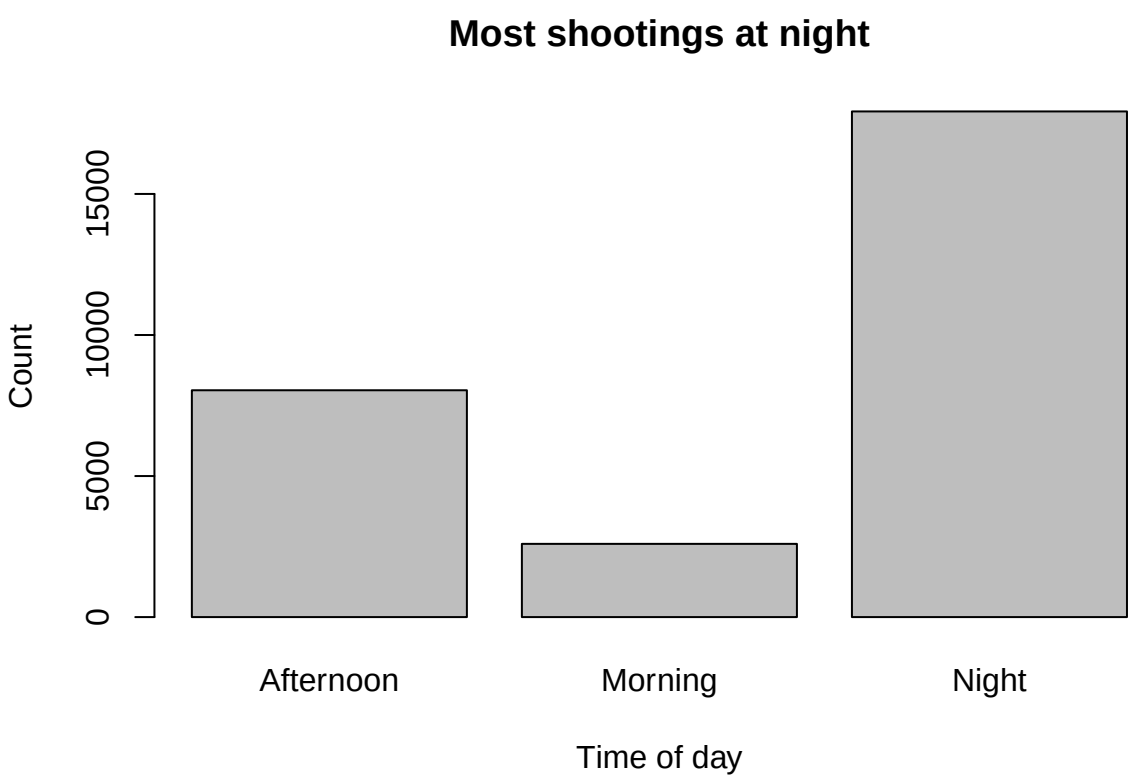
```
barplot(table(subset$OCCUR_MONTH), main="Shootings spike in summer", xlab="Month", ylab="Count")
```



Yes, the time of year appears to affect the number of shootings. There are more recorded shooting incidents in the summer months than there are in winter.

Time of day Is there a time of day when most shootings occur? For reference, morning is from the hours 5-11am, afternoon is from 12-7pm, and night is from 8pm-4am.

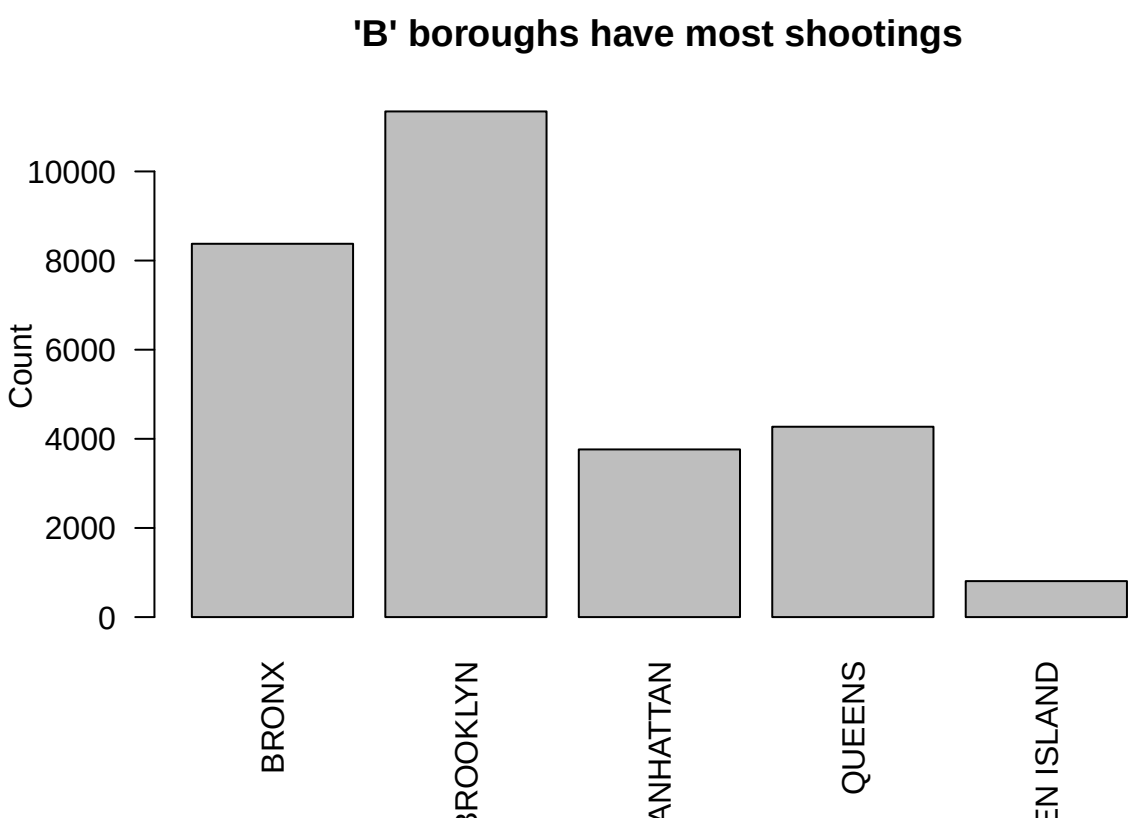
```
barplot(table(subset$TIME_OF_DAY), main="Most shootings at night", xlab="Time of day", ylab="Count")
```



Over half of all shootings occur in the night category.

Borough Are some boroughs of NYC more violent than others?

```
barplot(table(subset$BORO), main="'B' boroughs have most shootings", ylab="Count", las=2)
```



Yes, the boroughs that begin with the letter B (Bronx and Brooklyn) have many more shooting incidents than the other boroughs.

Conclusion

The dataset was acquired from data.gov, with information on every recorded shooting by the NYPD since 2006. This analysis focused on the following pieces of information: the month, the time, and the borough. The dataset was pared down to these features, and the features were encoded in a way that was conducive to analysis. After visualizing the data, it was clear that all three factors may be related to the likelihood of a shooting. We saw that most shootings happen in the summer, at night, and in the Bronx and Brooklyn boroughs.

It's possible that other factors cause these shooting incidents to occur: the factors investigated today don't tell the full story. The only possible source of bias in my analysis that I can identify are the arbitrary cut-offs assigned for morning, afternoon, and night. These cut-offs were determined according to my interpretation, and other bins may result in a different analysis. The session info is posted below, to allow for reproducibility of this work.

```
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 22631)
##
## 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: America/Chicago
## 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 magrittr_2.0.3
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.5      highr_0.10        compiler_4.4.0    tidyselect_1.2.1
##  [5] scales_1.3.0      yaml_2.3.8        fastmap_1.1.1     R6_2.5.1
##  [9] generics_0.1.3    knitr_1.46        munsell_0.5.1     pillar_1.9.0
## [13] tzdb_0.4.0        rlang_1.1.3       utf8_1.2.4        stringi_1.8.3
## [17] xfun_0.43         timechange_0.3.0  cli_3.6.2         withr_3.0.0
## [21] digest_0.6.35     grid_4.4.0        rstudioapi_0.16.0 hms_1.1.3
## [25] lifecycle_1.0.4   vctrs_0.6.5       evaluate_0.23     glue_1.7.0
## [29] fansi_1.0.6       colorspace_2.1-0  rmarkdown_2.26    tools_4.4.0
## [33] pkgconfig_2.0.3   htmltools_0.5.8.1
```