NYPD COVID Shooting

OB

2024-10-08

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

1. Importing Data

First we install one of R's most important libraries (tidyverse)

```
library(tidyverse)
```

Next, we obtain the url that we will be downloading our NYPD shooting data from and extract it into a df using read_csv from the tidyverse library

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_shoot <- read_csv(url_in)
```

We have now imported and stored our data in a variable.

2. Tidying and Transforming Data

Let's call on our data to see what it looks like to have a general understanding of variables we would need to change or affect in our tidying and transformation process.

```
nypd_shoot
```

```
## # A tibble: 28,562 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
##  1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
##  2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
##  3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
##  4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
##  5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
##  6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
##  7    229224142 06/07/2021 19:55      QUEENS    <NA>                   113
##  8    231246224 07/22/2021 01:47      BROOKLYN  <NA>                    77
##  9    228559720 05/22/2021 18:39      BRONX     <NA>                    48
```

```
## 10    238210279 12/22/2021 23:17      BRONX     <NA>                    49
## # i 28,552 more rows
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

I have decided to remove all the location based columns. Analysis will be done by borough/precinct. Then, load the lubridate library and change the occur_date column to type date.

```
nypd_shoot <- nypd_shoot %>% select(-c(Latitude, Longitude, X_COORD_CD, Y_COORD_CD, Lon_Lat))
library(lubridate)
nypd_shoot$OCCUR_DATE <- mdy(nypd_shoot$OCCUR_DATE)
summary(nypd_shoot)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:28562      Length:28562
##  1st Qu.: 65439914   1st Qu.:2009-09-04   Class1:hms        Class :character
##  Median : 92711254   Median :2013-09-20   Class2:difftime   Mode  :character
##  Mean   :127405824   Mean   :2014-06-07   Mode  :numeric
##  3rd Qu.:203131993   3rd Qu.:2019-09-29
##  Max.   :279758069   Max.   :2023-12-29
##
##  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
##  Length:28562
##  Class :character
##  Mode  :character
##
```

```
##
##
##
```

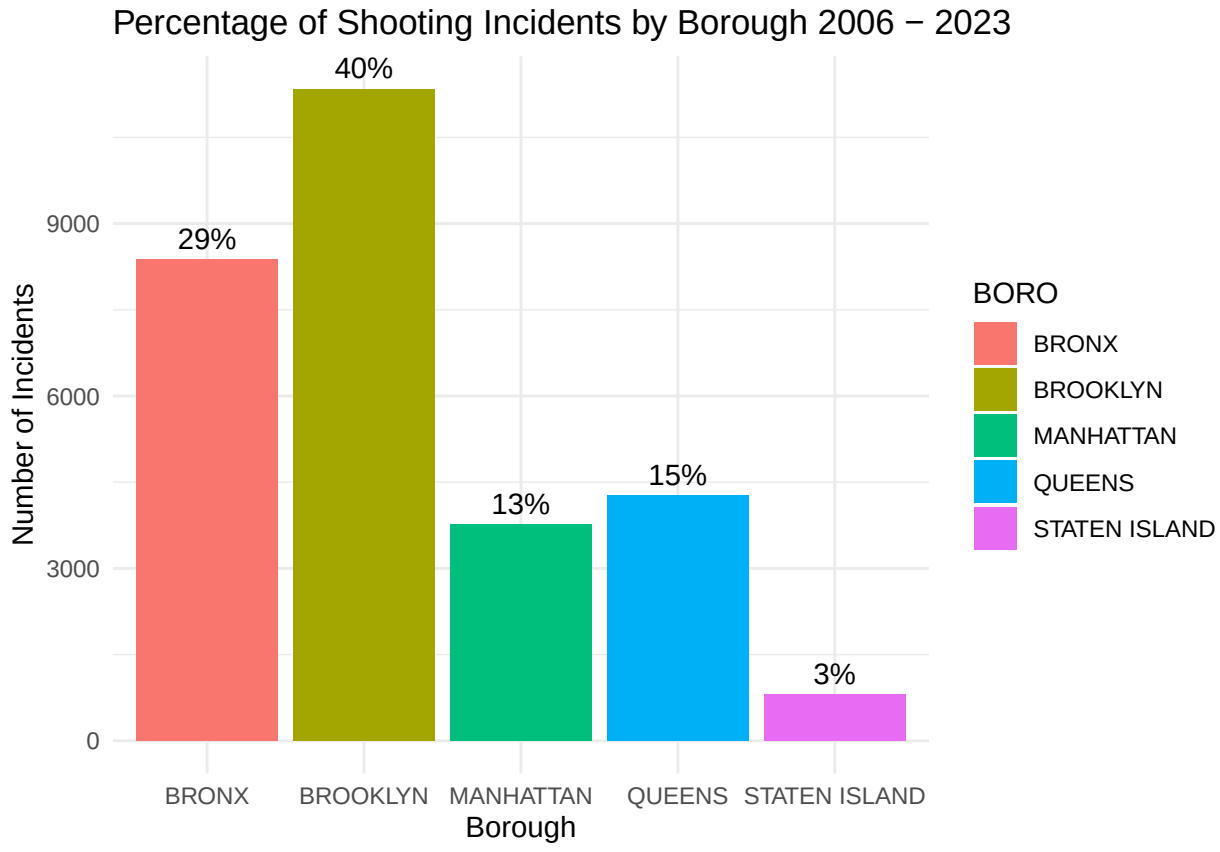
3. Visualization and Model(s)

The first visualization I am making is what percentage of shootings take place by Borough.

```
 borough_counts <- nypd_shoot %>%
     group_by(BORO) %>%
     summarize(count = n())

borough_counts <- borough_counts %>%
      mutate(percentage = count / sum(count) * 100)

ggplot(borough_counts, aes(x = BORO, y = count, fill = BORO)) +
      geom_bar(stat = "identity") +
      geom_text(aes(label = paste0(round(percentage, 0), "%")), vjust = -0.5) +
      labs(title = "Percentage of Shooting Incidents by Borough 2006 - 2023",
           x = "Borough",
           y = "Number of Incidents") +
      theme_minimal()
```



The next visualization I make is one that shows what time these shootings occur.

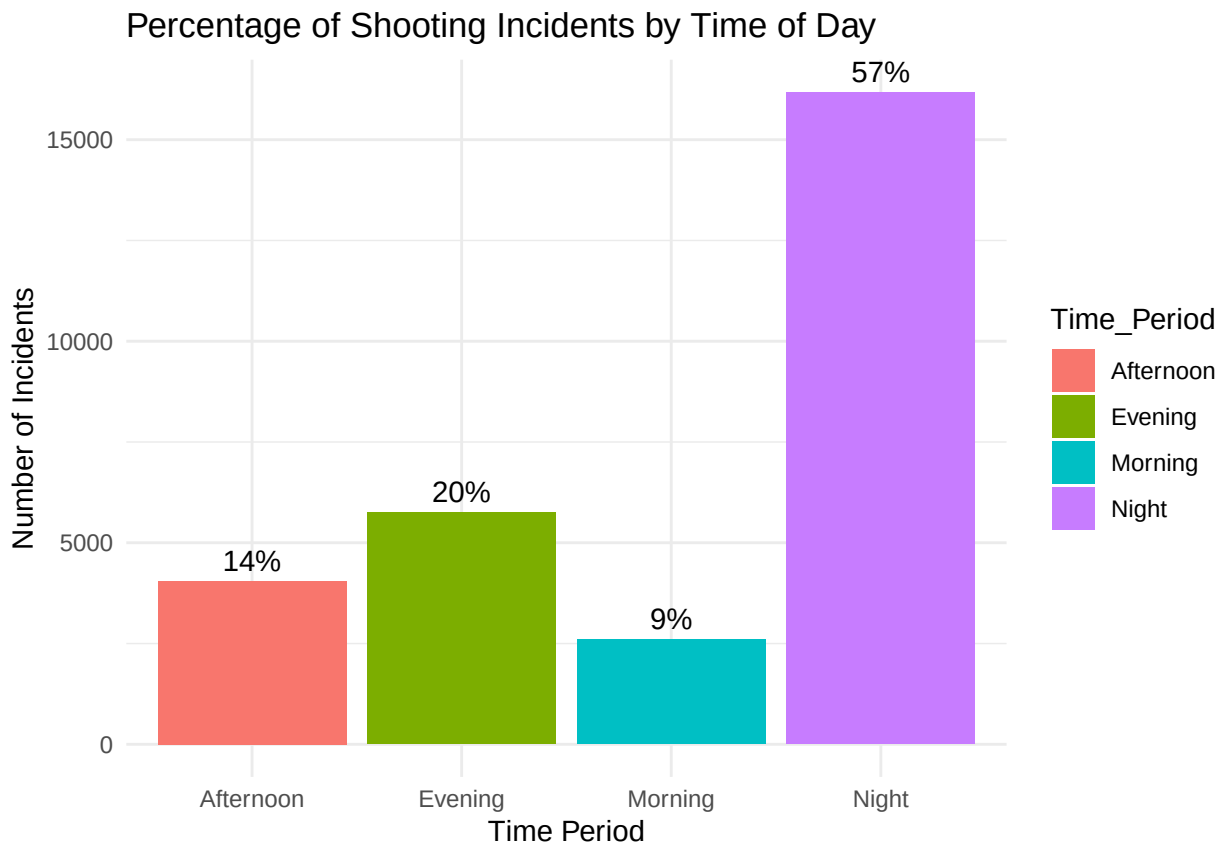
```r
classify_time <- function(time) {
      hour <- as.numeric(format(time, "%H"))  # Extract hour as numeric

      if (hour >= 5 & hour < 12) {
          return("Morning")
      } else if (hour >= 12 & hour < 17) {
          return("Afternoon")
      } else if (hour >= 17 & hour < 21) {
          return("Evening")
      } else {
          return("Night")
      }
}

nypd_shoot <- nypd_shoot %>%
     mutate(OCCUR_TIME = as.POSIXct(OCCUR_TIME, format = "%H:%M:%S", tz = "UTC"))
nypd_shoot <- nypd_shoot %>%
     mutate(Time_Period = sapply(OCCUR_TIME, classify_time))

time_counts <- nypd_shoot %>%
           group_by(Time_Period) %>%
           summarize(count = n())
time_counts <- time_counts %>%
           mutate(percentage = count / sum(count) * 100)

ggplot(time_counts, aes(x = Time_Period, y = count, fill = Time_Period)) +
     geom_bar(stat = "identity") +
     geom_text(aes(label = paste0(round(percentage, 0), "%")), vjust = -0.5) +
     labs(
         title = "Percentage of Shooting Incidents by Time of Day",
         x = "Time Period",
         y = "Number of Incidents"
     ) +
     theme_minimal()
```



Using a linear regression model to showcase relationships between boroughs and the times of night that shootings occur.

```
incident_counts <- nypd_shoot %>%
      group_by(OCCUR_DATE, BORO, Time_Period) %>%
      summarize(incident_count = n())
```

```
lm_model <- lm(incident_count ~ BORO + OCCUR_DATE + Time_Period, data = incident_counts)

summary(lm_model)
```

```
##
## Call:
## lm(formula = incident_count ~ BORO + OCCUR_DATE + Time_Period,
##     data = incident_counts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9808 -0.6194 -0.4142  0.1801 17.1564
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.871e+00  8.172e-02  22.900  < 2e-16 ***
## BOROBROOKLYN         1.816e-02  2.277e-02   0.797 0.425257
## BOROMANHATTAN       -2.390e-01  2.955e-02  -8.088 6.47e-16 ***
## BOROQUEENS          -2.367e-01  2.837e-02  -8.346  < 2e-16 ***
```

```
## BOROSTATEN ISLAND  -4.166e-01  5.159e-02  -8.076 7.10e-16 ***
## OCCUR_DATE         -2.714e-05  4.662e-06  -5.821 5.94e-09 ***
## Time_PeriodEvening  9.873e-02  2.957e-02   3.339 0.000844 ***
## Time_PeriodMorning -2.798e-02  3.547e-02  -0.789 0.430187
## Time_PeriodNight    4.481e-01  2.582e-02  17.353  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.219 on 17926 degrees of freedom
## Multiple R-squared:  0.03927,    Adjusted R-squared:  0.03884
## F-statistic: 91.59 on 8 and 17926 DF,  p-value: < 2.2e-16
```

Three Key Insights that can be drawn from the data:

1. Shootings are statistically more likely to occur at night
2. Shootings are decreasing over time, a possible decline in gun violence
3. Manhattan, Queens, and Staten Island are less likely to havve shooting incidents

4. Ethics/Bias and Conclusion

As data scientists, it is crucial to prioritize ethical considerations in our analyses to ensure that the insights we produce are not harmful or misleading. For instance, while exploring the relationship between shootings and ethnicity was a potential avenue of inquiry, we decided not to pursue it to avoid reinforcing harmful biases. Instead, we focused on insights that can drive constructive action. Our analysis revealed several key findings: shootings are statistically more likely to occur at night, there is a potential decline in gun violence over time, and Manhattan, Queens, and Staten Island show lower likelihoods of shooting incidents compared to other areas. These insights offer valuable information for policy-makers and law enforcement, promoting data-driven strategies to reduce gun violence without perpetuating harmful narratives.