NYPD_Shooting

Chitra G

2024-05-28

```
knitr::opts_chunk$set(echo = TRUE)
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(readr)
library(lubridate)
library(ggplot2)

# Define the URL for the data
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

# Read the data from the URL
shoot_in <- read_csv(url_in)
```

```
## 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.
```

```
# Display the specification of all columns
spec(shoot_in)
```

```
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   LOC_OF_OCCUR_DESC = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOC_CLASSFCTN_DESC = col_character(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_double(),
##   Y_COORD_CD = col_double(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

```r
# Renaming columns for better readability
shoot_in <- shoot_in %>%
  rename(
    Incident_Key = `INCIDENT_KEY`,
    Occurrence_Date = `OCCUR_DATE`,
    Borough = `BORO`,
    Precinct = `PRECINCT`,
    Jurisdiction_Code = `JURISDICTION_CODE`,
    Victim_Age_Group = `VIC_AGE_GROUP`,
    Victim_Sex = `VIC_SEX`,
    Victim_Race = `VIC_RACE`
  )

# Convert date columns to Date type
shoot_in <- shoot_in %>%
  mutate(Occurrence_Date = as.Date(Occurrence_Date, format = "%m/%d/%Y"))

# Select all columns except Latitude and Longitude
shoot_in <- shoot_in %>%
  select(-Latitude, -Longitude)

# Remove rows with any missing values
shoot_in <- shoot_in %>%
  drop_na()

# Display the first few rows of the tidied dataset
head(shoot_in)
```

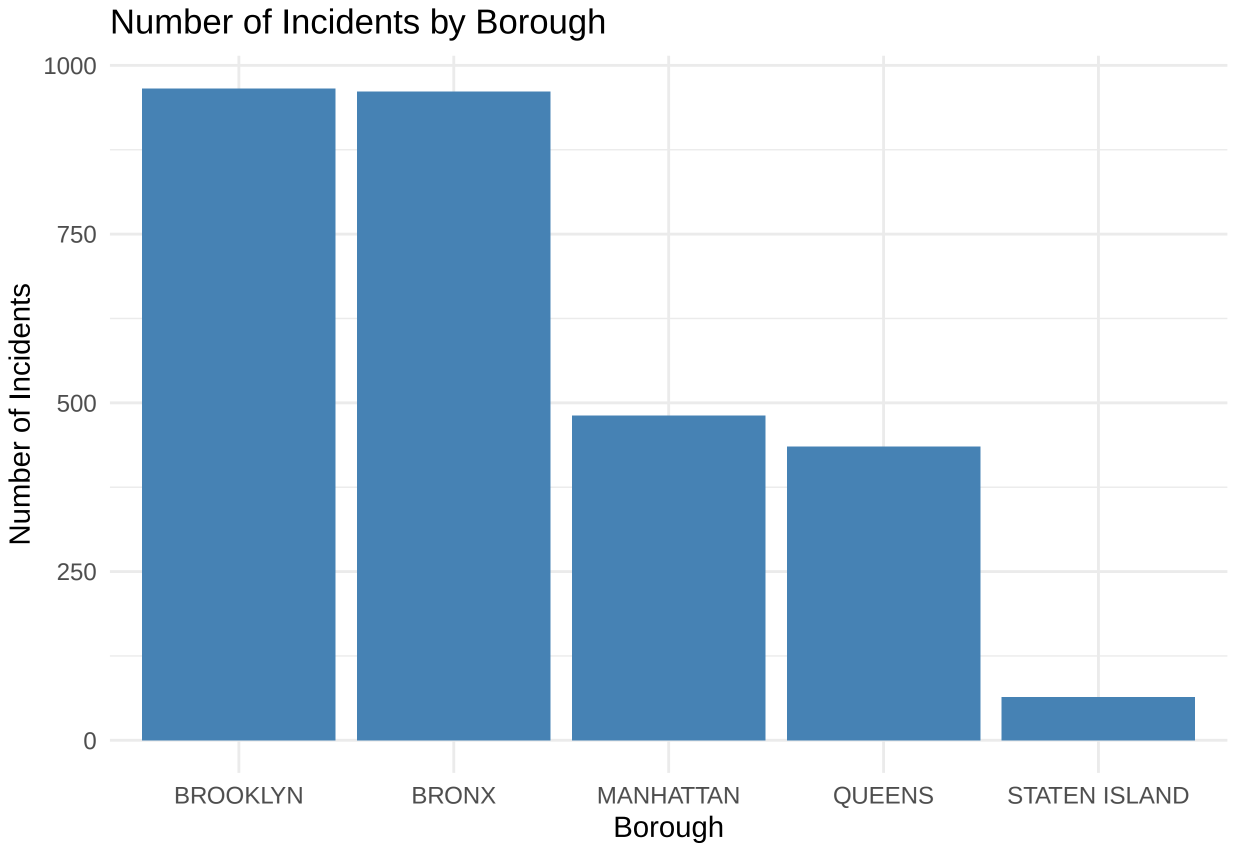
```
## # A tibble: 6 x 19
##   Incident_Key Occurrence_Date OCCUR_TIME Borough   LOC_OF_OCCUR_DESC Precinct
##          <dbl> <date>          <time>     <chr>     <chr>                <dbl>
## 1    244608249 2022-05-05      00:10      MANHATTAN INSIDE                  14
## 2    247542571 2022-07-04      22:20      BRONX     OUTSIDE                  48
## 3    254911480 2022-11-30      21:15      BRONX     OUTSIDE                  46
## 4    249623757 2022-08-15      18:21      QUEENS    OUTSIDE                 101
## 5    243433246 2022-04-10      17:00      BRONX     OUTSIDE                  49
## 6    253757468 2022-11-07      11:35      BROOKLYN  OUTSIDE                  75
## # i 13 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>, Victim_Age_Group <chr>, Victim_Sex <chr>,
## #   Victim_Race <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Lon_Lat <chr>
```

```r
#Data Analysis and Visualization
#Visualization 1: Incident Counts by Borough
##We start by visualizing the number of incidents by borough.
## Count the number of incidents by borough
incidents_by_borough <- shoot_in %>%
  count(Borough, sort = TRUE)

## Display the count of incidents by borough
incidents_by_borough
```

```
## # A tibble: 5 x 2
##   Borough           n
##   <chr>         <int>
## 1 BROOKLYN        966
## 2 BRONX           961
## 3 MANHATTAN       481
## 4 QUEENS          435
## 5 STATEN ISLAND    64
```

```r
# Visualization: Incidents by Borough
ggplot(incidents_by_borough, aes(x = reorder(Borough, -n), y = n)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Number of Incidents by Borough",
       x = "Borough",
       y = "Number of Incidents") +
  theme_minimal()
```



Looking at this visualization, we can see that:

- With 966 incidents, Brooklyn has the highest number of incidents among all boroughs.
- Following closely, the Bronx reports 961 incidents.
- Both Brooklyn and the Bronx experience a substantial number of shooting incidents, making them areas with higher levels of gun violence.

On the other hand:

- Staten Island reports the lowest number of shooting incidents among all boroughs, with only 64 incidents reported.
- This suggests that Staten Island has relatively lower levels of gun violence compared to other boroughs in New York City.

Questions arise:

- Why are shooting incidents more common in Brooklyn and the Bronx?
- Why is Staten Island more peaceful?

```r
# Visualization 2: Trend Over Time
## Next, we visualize the trend of incidents over time.
## Aggregate incidents by occurrence date
incidents_by_date <- shoot_in %>%
  group_by(Occurrence_Date) %>%
  summarize(Incident_Count = n())
```

```r
# Check the structure of the incidents_by_date dataframe
str(incidents_by_date)
```

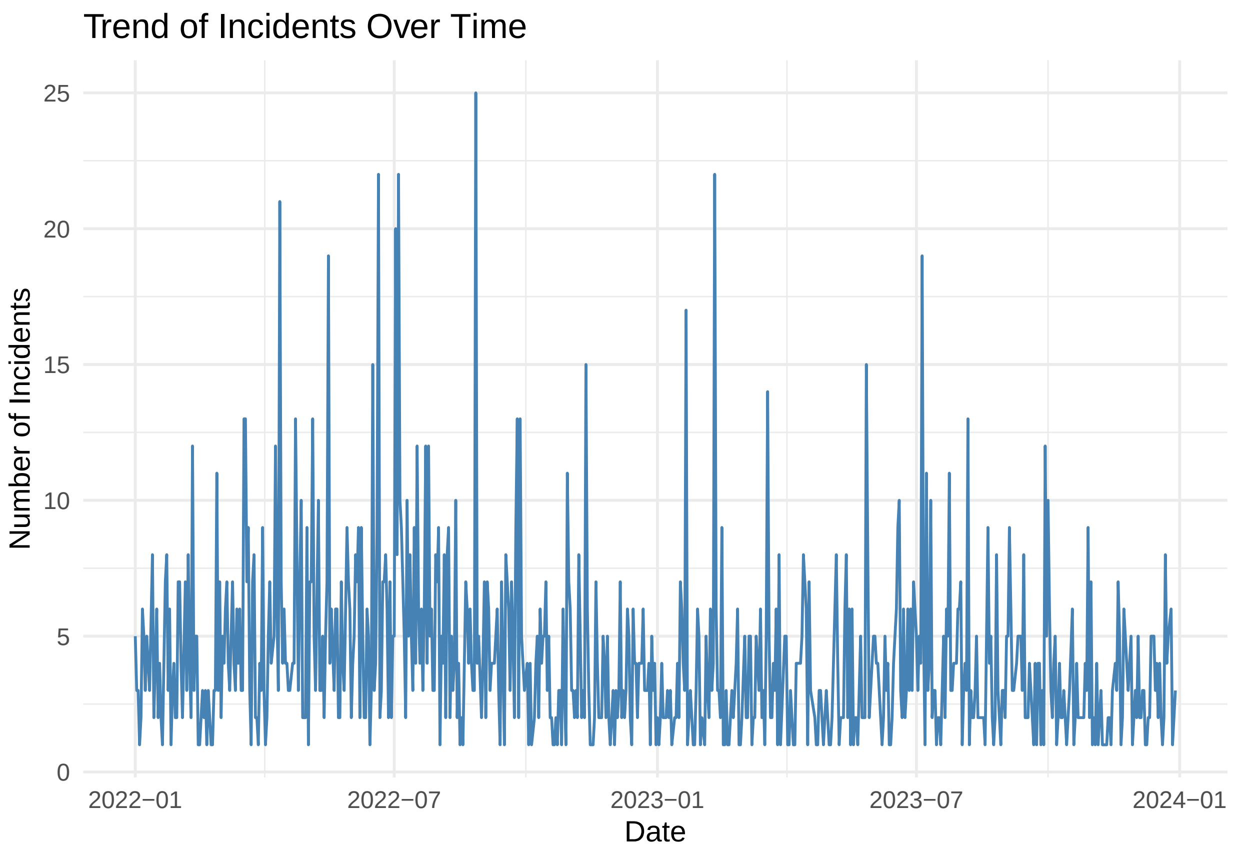
```
## tibble [684 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Occurrence_Date: Date[1:684], format: "2022-01-01" "2022-01-02" ...
##  $ Incident_Count : int [1:684] 5 3 3 1 2 6 5 3 5 4 ...
```

```r
# Add a numeric date column for the model
incidents_by_date <- incidents_by_date %>%
  mutate(Date_Num = as.numeric(Occurrence_Date))

# Check the structure to ensure Date_Num is created
print(head(incidents_by_date))
```

```
## # A tibble: 6 x 3
##   Occurrence_Date Incident_Count Date_Num
##   <date>                   <int>    <dbl>
## 1 2022-01-01                   5    18993
## 2 2022-01-02                   3    18994
## 3 2022-01-03                   3    18995
## 4 2022-01-04                   1    18996
## 5 2022-01-05                   2    18997
## 6 2022-01-06                   6    18998
```

```r
# Visualization: Incidents Over Time
ggplot(incidents_by_date, aes(x = Occurrence_Date, y = Incident_Count)) +
  geom_line(color = "steelblue") +
  labs(title = "Trend of Incidents Over Time",
       x = "Date",
       y = "Number of Incidents") +
  theme_minimal()
```



Second visualization shows more spikes in 2022. *Question to ponder why 2022 saw so much shooting incidents. Can we find the reason?*

```
#Simple Linear Regression Model
#We fit a simple linear regression model to predict #the number of incidents based on the date.
# Add a numeric date column for the model
shoot_in <- shoot_in %>%
  mutate(Date_Num = as.numeric(Occurrence_Date))

# Linear regression model
model <- lm(Incident_Count ~ Date_Num, data = incidents_by_date)

# Model summary
summary(model)
```

```
##
## Call:
## lm(formula = Incident_Count ~ Date_Num, data = incidents_by_date)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3465 -2.1794 -0.7333  1.3340 20.3771
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 63.8419024 11.3829306   5.609 2.97e-08 ***
## Date_Num    -0.0030794  0.0005882  -5.236 2.19e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.227 on 682 degrees of freedom
## Multiple R-squared:  0.03864,    Adjusted R-squared:  0.03723
## F-statistic: 27.41 on 1 and 682 DF,  p-value: 2.194e-07
```

This summary is about a model that tried to predict how many incidents happen based on date

- **Coefficients**
  - *The first number 63.84 is starting point*
  - *The second number -0.00308 shows that as the date increases, the number of shooting incidents decreases little bit*
- **Statistical Significance:**
  - *Both numbers are so unlikely to be zero that we can be confident they're real* -*This means that the date does seem to have an effect on the number of incidents.* -**Model Performance:** -*The model isn't perfect. It's like it can't explain much of what's going on. It only gets about 3.86% of the way there.* -*This means that knowing the date doesn't tell us much about how many incidents there will be.* -*But overall, the model does seem to be doing something useful, according to the tests it went through.*
- **Bias Identification:Data bias in NYPD shooting incident data means that the information we have might not give us the full picture. Here's how that might have happened:**:
  - *Some Incidents Aren't Reported: Sometimes, not all shooting incidents get reported to the police. This could happen for various reasons, like people being afraid to come forward or not trusting the police.*
  - *Certain Areas Get More Attention: Police might focus more on certain neighborhoods, which could make it seem like more incidents happen there, even if they happen elsewhere too*
  - *How Data is Collected Matters: Sometimes, the way police collect data on shootings can lead to bias. For example, if they only count certain types of incidents or if their methods change over time, it can affect the numbers we see.*
  - *Who Reports Matters: People might be more or less likely to report a shooting based on their trust in the police or fear of getting in trouble.*
  - *Data Definitions Can Change: If the rules for what counts as a shooting change, it can make comparisons over time tricky.*

All these factors can make it challenging to understand the true extent and nature of shooting incidents in New York City. Recognizing and addressing these biases is important for making better decisions and policies to keep communities safe.