Analysis of New York City Shooting Incidents from 2006 to Present

John Creath

2025-12-04

Source of Analysis

Data is supplied by the New York City Police Department and served up at data.gov.

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

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```
library(dplyr)
```

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

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

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

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

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

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

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

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

## Warning: package 'lubridate' was built under R version 4.5.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.1     v stringr   1.6.0
## v lubridate 1.9.4     v tibble    3.3.0
## v purrr     1.2.0     v tidyr     1.3.1
## v readr     2.1.6

## -- 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)
library(stringr)

nypd_url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?"

nypd_file_names <-
     c("rows.csv")
nypd_urls <- str_c(nypd_url_in, nypd_file_names)

nypd_shootings <- read_csv(nypd_urls[1])
```

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): 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.
```

```
nypd_shootings <- nypd_shootings %>%

  # --- Date & time cleaning ---
  mutate(
    OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"),
    OCCUR_YEAR = year(OCCUR_DATE),
    OCCUR_HOUR = hour(hms(OCCUR_TIME))
  ) %>%

  # --- Borough cleaning ---
  mutate(
    BORO = str_to_title(BORO)
  ) %>%
```

```r
  # --- Sex cleaning ---
  mutate(
    PERP_SEX = case_when(
      PERP_SEX == "M" ~ "Male",
      PERP_SEX == "F" ~ "Female",
      TRUE ~ "Unknown"
    ),
    VIC_SEX = case_when(
      VIC_SEX == "M" ~ "Male",
      VIC_SEX == "F" ~ "Female",
      TRUE ~ "Unknown"
    )
  ) %>%

  # --- Race cleaning ---
  mutate(
    PERP_RACE = case_when(
      PERP_RACE %in% c("UNKNOWN", "(null)", NA) ~ "Unknown",
      TRUE ~ str_to_title(str_to_lower(PERP_RACE))
    ),
    VIC_RACE = case_when(
      VIC_RACE %in% c("UNKNOWN", "(null)", NA) ~ "Unknown",
      TRUE ~ str_to_title(str_to_lower(VIC_RACE))
    )
  ) %>%

  # --- Age cleaning ---
  mutate(
    numeric_age_perp = suppressWarnings(as.numeric(PERP_AGE_GROUP)),
    PERP_AGE_GROUP = case_when(
      PERP_AGE_GROUP %in% c("(null)", NA) ~ "UNKNOWN",
      !is.na(numeric_age_perp) & (numeric_age_perp < 0 | numeric_age_perp > 120) ~ "UNKNOWN",
      PERP_AGE_GROUP == "224" ~ "18-24",
      TRUE ~ PERP_AGE_GROUP
    ),

    numeric_age_vic = suppressWarnings(as.numeric(VIC_AGE_GROUP)),
    VIC_AGE_GROUP = case_when(
      VIC_AGE_GROUP %in% c("1022", "(null)", NA) ~ "UNKNOWN",
      !is.na(numeric_age_vic) & (numeric_age_vic < 0 | numeric_age_vic > 120) ~ "UNKNOWN",
      VIC_AGE_GROUP == "224" ~ "18-24",
      TRUE ~ VIC_AGE_GROUP
    )
  ) %>%

  # Drop helper numeric columns
  select(-numeric_age_perp, -numeric_age_vic)
```

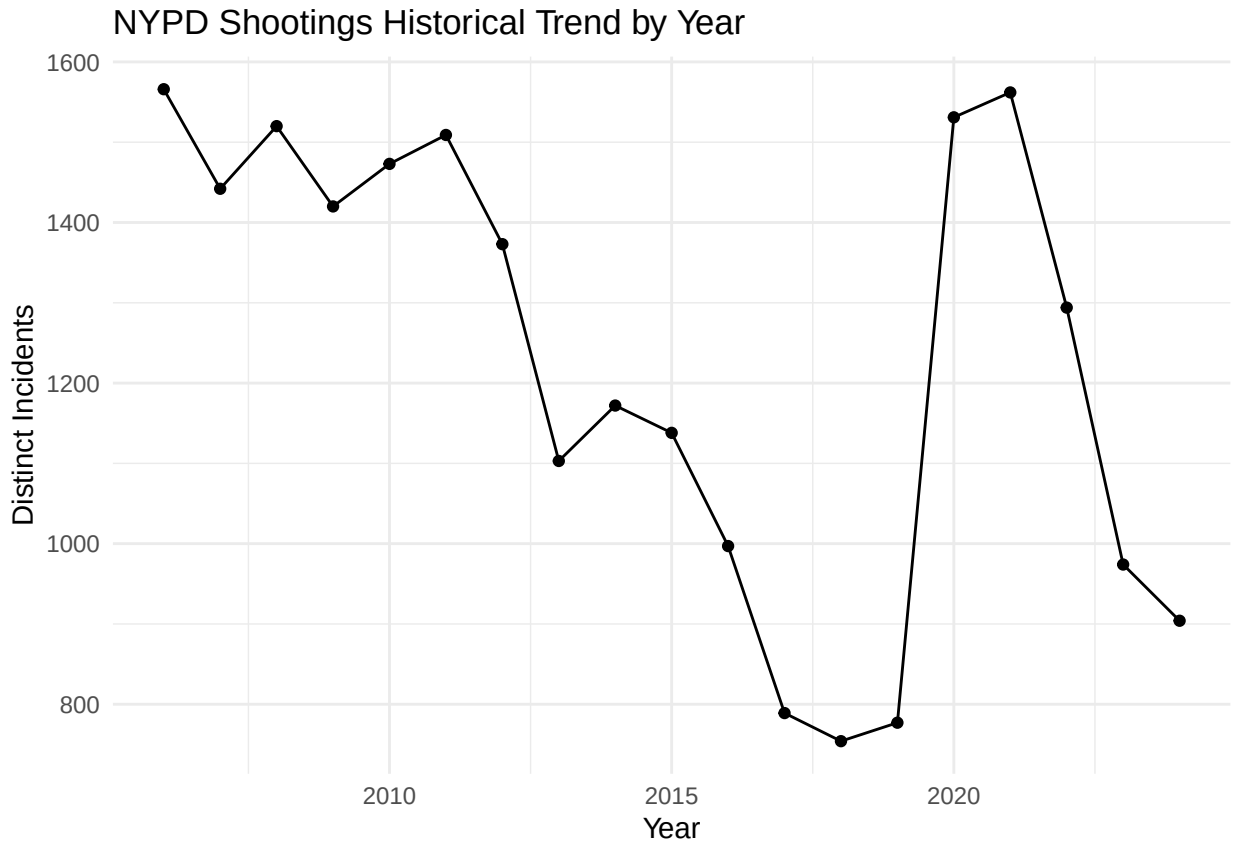
Findings

This analysis seeks to understand shootings in New York City, New York.

History

Since 2006, overall we can see that shootings in New York City have been generally decreasing. Much progress was made especially between the years of 2006 and 2019 to curb this form of violence seeing a 50% reduction in shootings. However, sadly with the arrival of the global COVID-19 pandemic we see an alarming resurgence that in essence unraveled the previous 10 years worth of work.

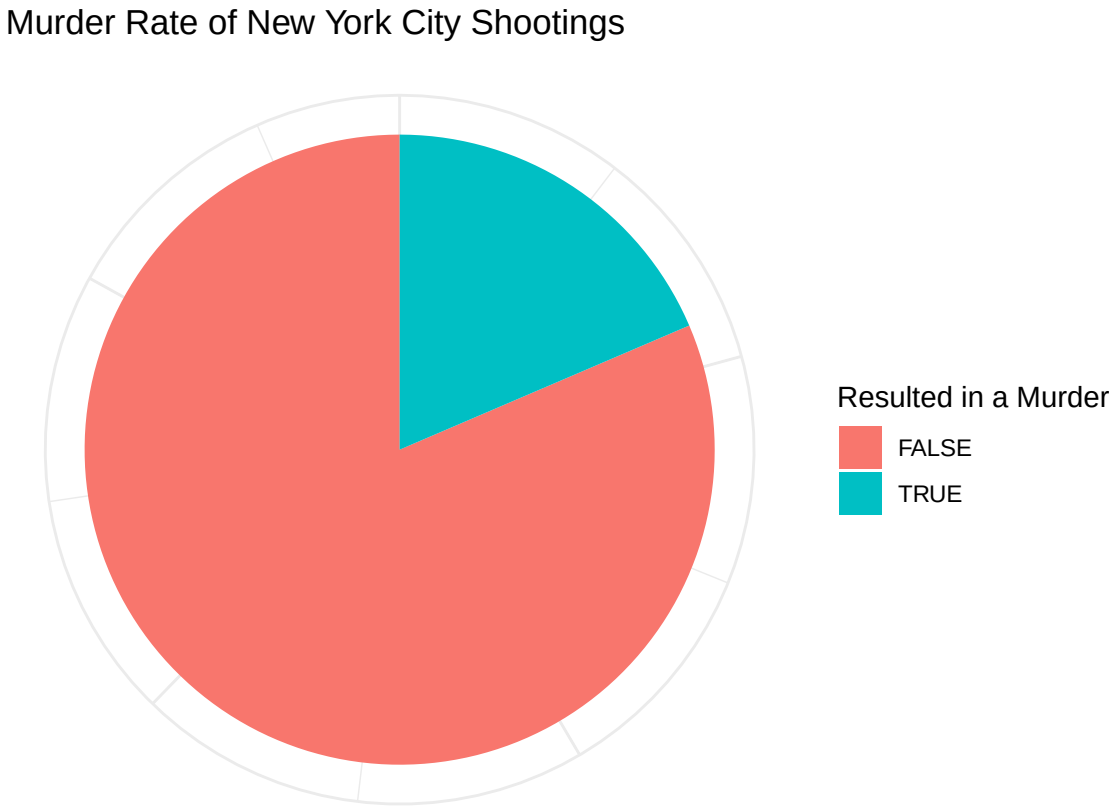
```
# Group by year and count distinct INCIDENT_KEY
shootings_by_year <- nypd_shootings %>%
  group_by(OCCUR_YEAR) %>%
  summarise(distinct_incidents = n_distinct(INCIDENT_KEY))

# Create a ggplot for visualization
ggplot(shootings_by_year, aes(x = OCCUR_YEAR, y = distinct_incidents)) +
  geom_line() +
  geom_point() +
  labs(title = "NYPD Shootings Historical Trend by Year",
       x = "Year",
       y = "Distinct Incidents") +
  theme_minimal()
```



While the resurgence of gun violence in New York City during the pandemic is disheartening, it is encouraging to see that since the height of the resurgence in 2021 we can see signs that perhaps NYC is heading back in the right direction. A 15% reduction in shootings can be observed between 2021 and 2022, a dramatic year-over-year decline that had not been seen in nearly 10 years.

As we wrap the background and history of shootings in NYC it could be helpful to understand the associated murder rate of these incidents. Since 2006 we see a murder rate of nearly 20%.

```
murder_flag_counts <- nypd_shootings %>%
  group_by(STATISTICAL_MURDER_FLAG) %>%
  summarize(distinct_incidents = n_distinct(INCIDENT_KEY))

ggplot(murder_flag_counts, aes(x = "", y = distinct_incidents, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Murder Rate of New York City Shootings",
       fill = "Resulted in a Murder") +
  theme_minimal() +
  theme(axis.text = element_blank(),
        axis.title = element_blank(),
        legend.position = "right")
```

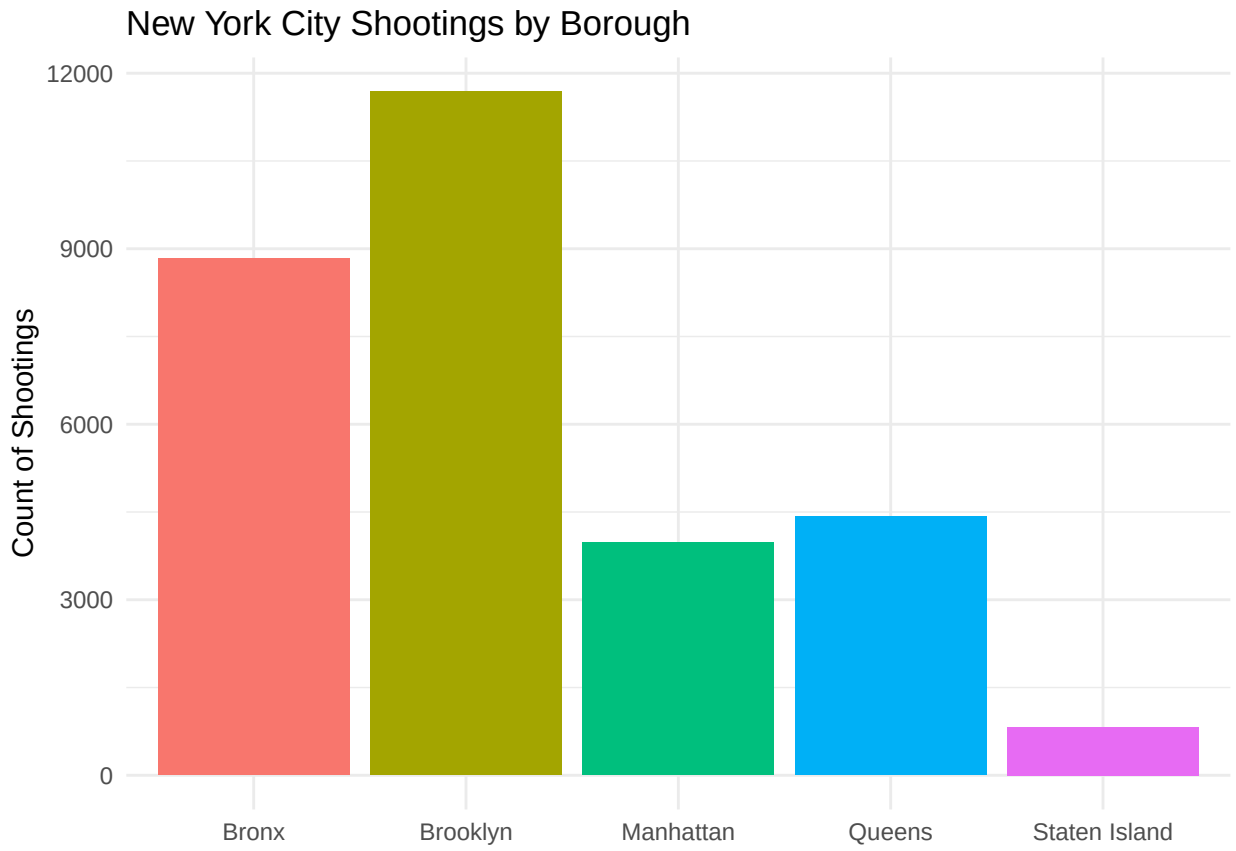


Plotting the murder rate by year does not yield any meaningful results. The rate fluctuates between 17-22% with no observed correlation to other dimensions.

Distribution

As the city allocates resources to help reduce the threat of gun violence it will be helpful to understand *where* the problem is most acute. The distribution of shootings by borough immediately draws our attention to the Bronx and Brooklyn, and for good reason.

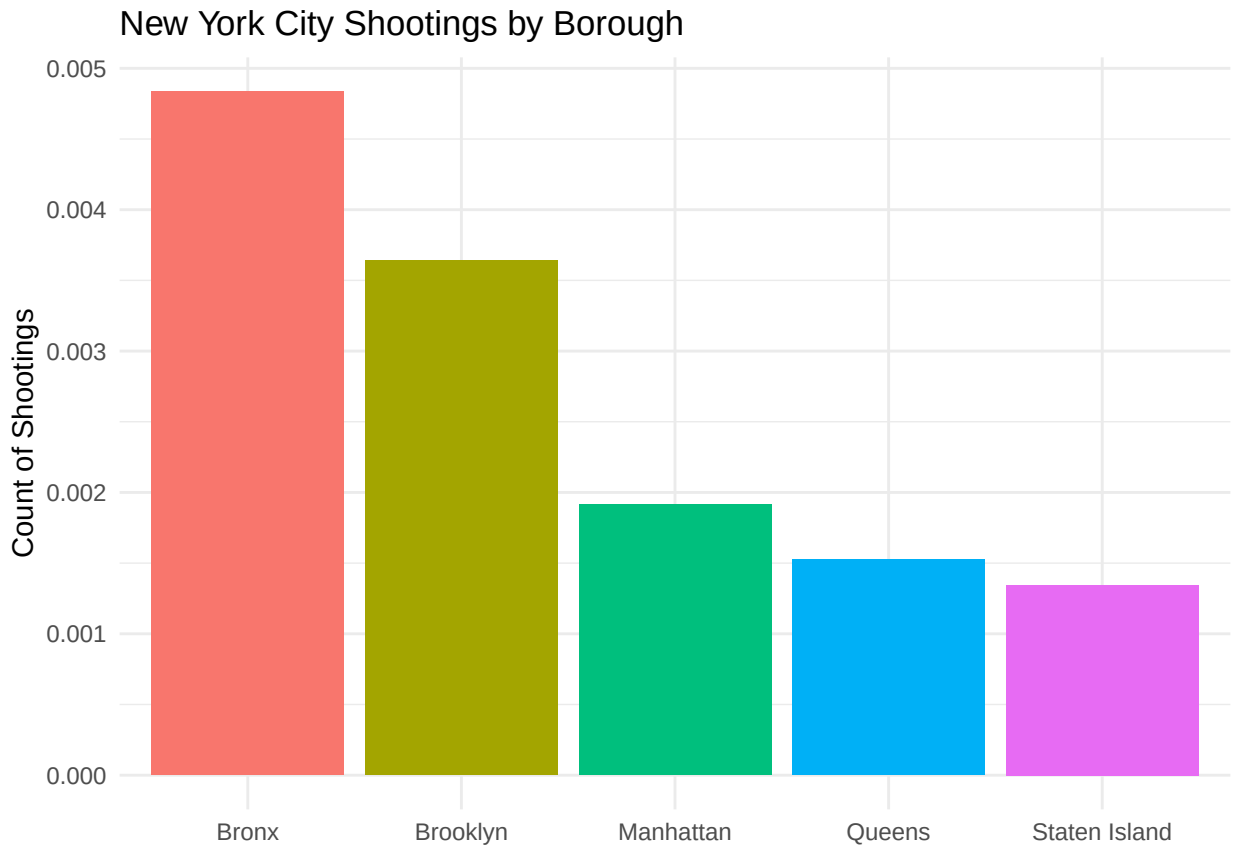
```r
nypd_shootings %>%
  ggplot(aes(x = BORO, y = after_stat(count), fill = BORO)) +
  geom_bar(stat = "count") +
  labs(
    title = "New York City Shootings by Borough",
    x = NULL,                  # remove x-axis title
    y = "Count of Shootings"
  ) +
  theme_minimal() +
  theme(legend.position = "none")    # remove legend
```



But how significant are these numbers considering the populations each borough represents? To answer this question, we take 2023 population estimates. Then we take the number of shootings per borough divided by the population to get shootings per capita, which will help contextualize these numbers.

```r
# Source: https://www.citypopulation.de/en/usa/newyorkcity/
ny_population_by_borough <- data.frame(
  borough = c("Bronx", "Brooklyn", "Manhattan", "Queens", "Staten Island"),
  population = c(1379946, 2590516, 1596273, 2278029, 491133))

nypd_shootings_agg <- nypd_shootings %>%
  group_by(BORO) %>%
  summarize(distinct_incidents = n_distinct(INCIDENT_KEY))

nypd_shootings_agg <- left_join(nypd_shootings_agg, ny_population_by_borough, by = c("BORO" = "borough")
```

```r
nypd_shootings_agg <- nypd_shootings_agg %>%
  mutate(shootings_per_capita = distinct_incidents / population)

nypd_shootings_agg %>%
  ggplot(aes(x = BORO, y = shootings_per_capita, fill = BORO)) +
  geom_bar(stat = "sum") +
  labs(
    title = "New York City Shootings by Borough",
    x = NULL,                  # remove x-axis title
    y = "Count of Shootings"
  ) +
  theme_minimal() +
  theme(legend.position = "none")    # remove legend
```

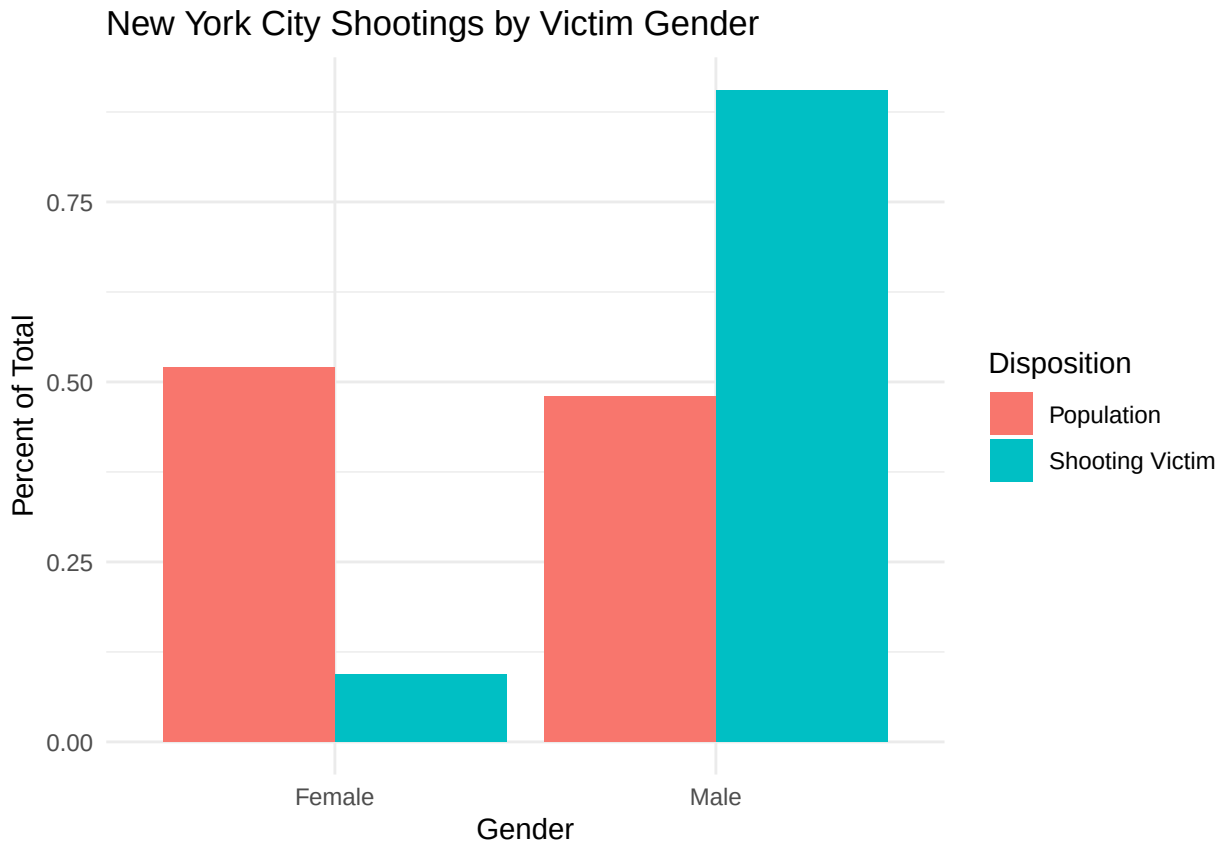


This confirms that while gun violence is especially prominent in the Bronx and Brooklyn, when correlated with population, we see that the problem is most acute in the Bronx.

When it comes to who gun violence affects, we can see from the graphs below that the victims and perpetrators of shootings in New York City are disproportionately male by a wide margin.

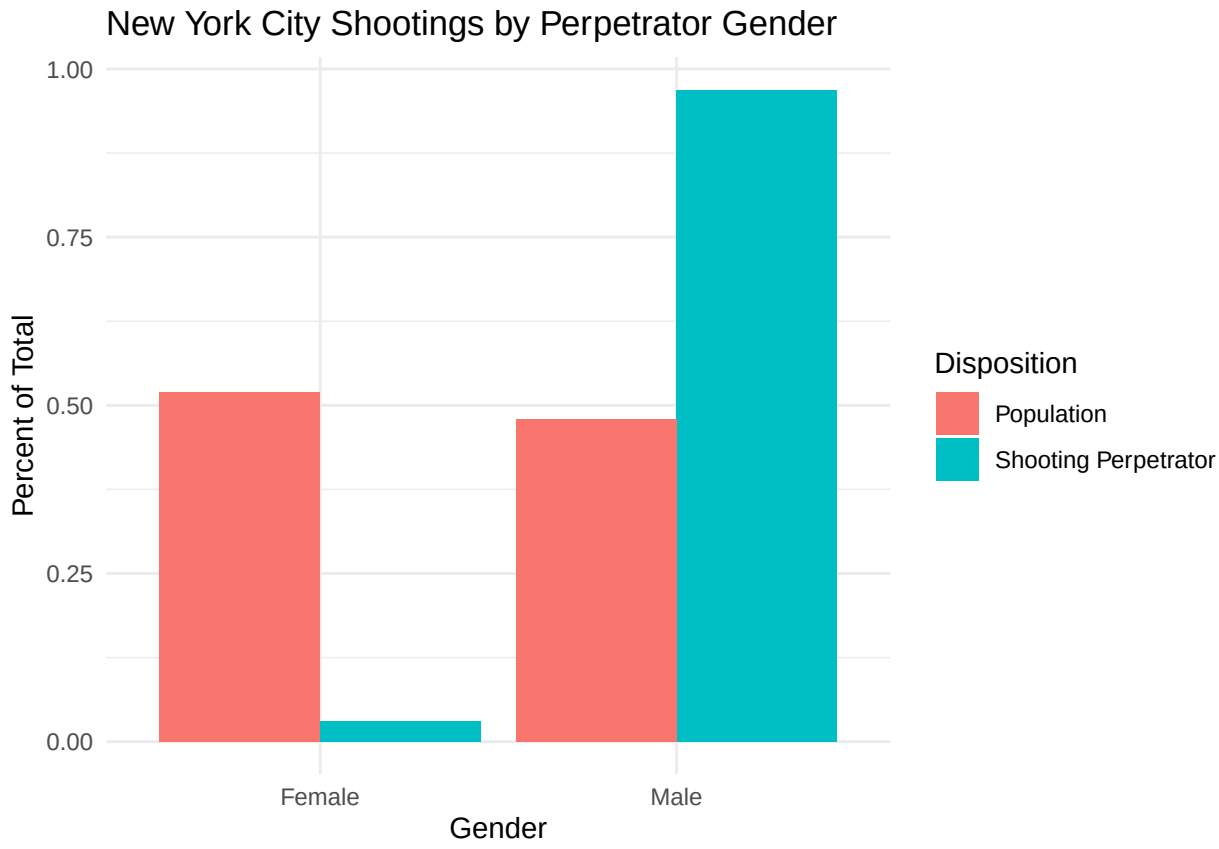
```r
# Source: https://www.neilsberg.com/insights/new-york-ny-population-by-gender/
ny_gender_distribution <- data.frame(
  gender = c("Male", "Female"),
  # population = c(4192240, 4543807),
  percent_of_total = c(.4799, .5201),
  disposition = c('Population'))
```

```r
nypd_shootings_by_victim_gender <- nypd_shootings %>%
  group_by(VIC_SEX) %>%
  rename(gender = VIC_SEX) %>%
  filter(gender != "Unknown") %>%
  summarize(distinct_incidents = n_distinct(INCIDENT_KEY)) %>%
  mutate(percent_of_total = distinct_incidents / sum(distinct_incidents)) %>%
  select(-distinct_incidents) %>%
  mutate(disposition = "Shooting Victim")

combined_victim_gender_analysis <- bind_rows(
  nypd_shootings_by_victim_gender,
  ny_gender_distribution
)

ggplot(combined_victim_gender_analysis, aes(x = gender, y = percent_of_total, fill = disposition)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "New York City Shootings by Victim Gender",
       x = "Gender",
       y = "Percent of Total",
       fill = "Disposition") +
  theme_minimal()
```




```r
nypd_shootings_by_perpetrator_gender <- nypd_shootings %>%
  group_by(PERP_SEX) %>%
  rename(gender = PERP_SEX) %>%
```

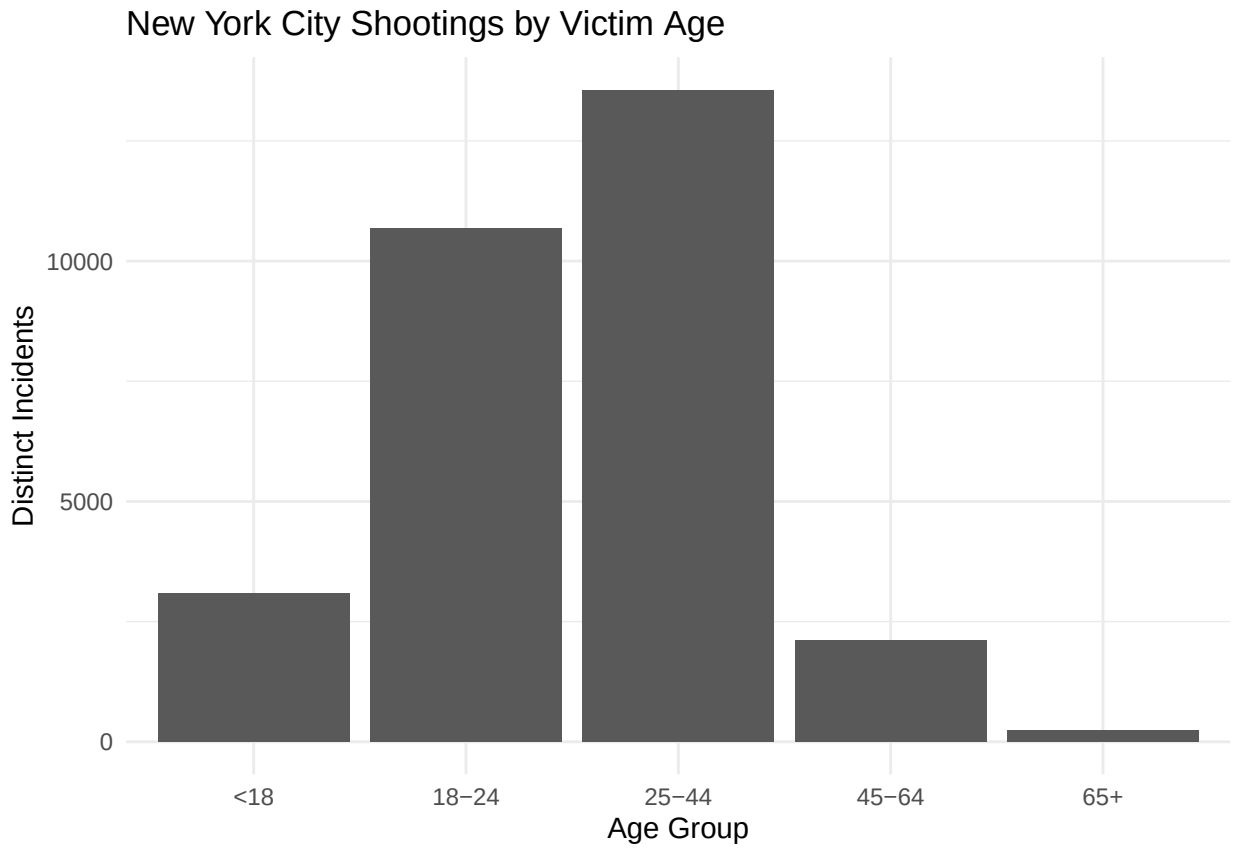
```
  filter(gender != "Unknown") %>%
  summarize(distinct_incidents = n_distinct(INCIDENT_KEY)) %>%
  mutate(percent_of_total = distinct_incidents / sum(distinct_incidents)) %>%
  select(-distinct_incidents) %>%
  mutate(disposition = "Shooting Perpetrator")

combined_perpetrator_gender_analysis <- bind_rows(
  nypd_shootings_by_perpetrator_gender,
  ny_gender_distribution
)

ggplot(combined_perpetrator_gender_analysis, aes(x = gender, y = percent_of_total, fill = disposition)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "New York City Shootings by Perpetrator Gender",
       x = "Gender",
       y = "Percent of Total",
       fill = "Disposition") +
  theme_minimal()
```



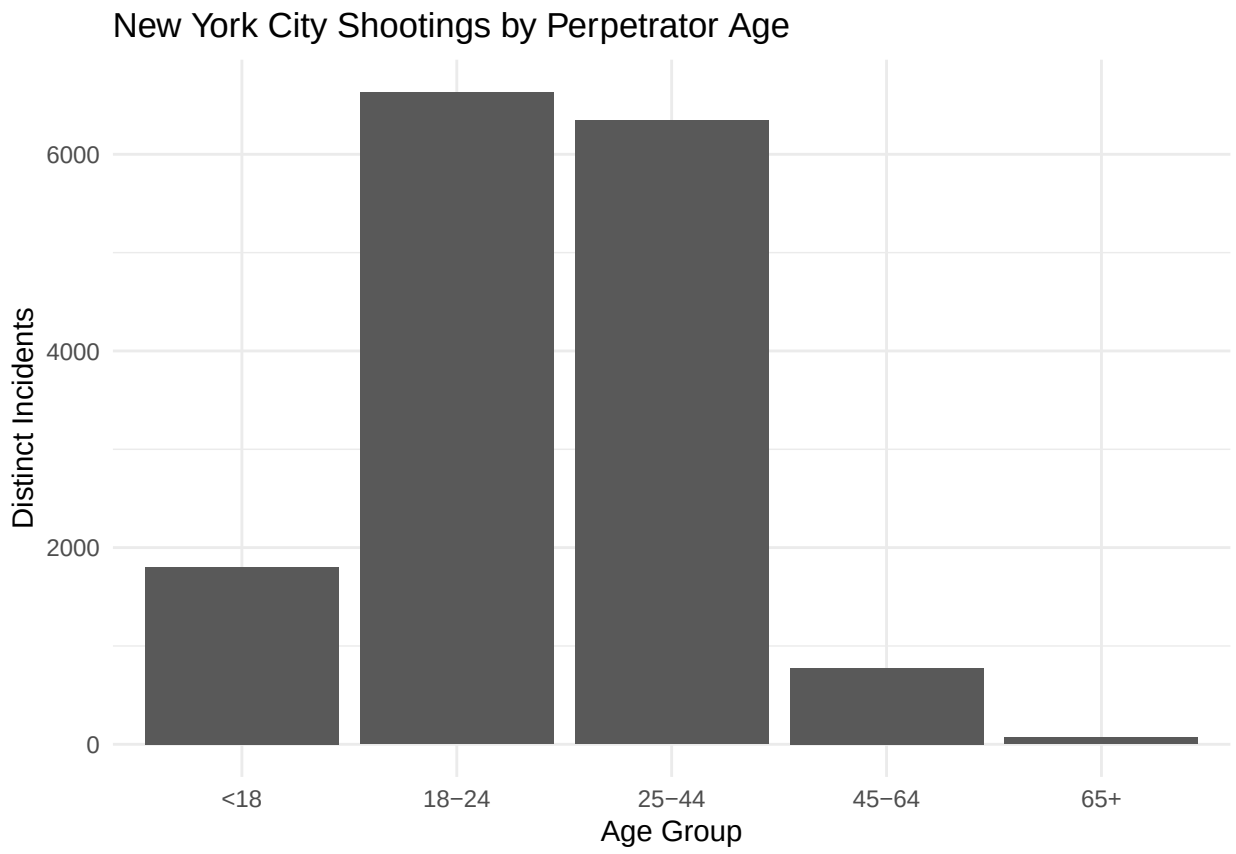
With respect to age of the victims and perpetrators where the age is known, we see the following:

```
# Exclude "UNKNOWN" values
filtered_vic_age <- nypd_shootings %>%
  filter(VIC_AGE_GROUP != "UNKNOWN")

# Create a bar plot
```

```r
ggplot(filtered_vic_age, aes(x = VIC_AGE_GROUP)) +
  geom_bar() +
  labs(title = "New York City Shootings by Victim Age",
       x = "Age Group",
       y = "Distinct Incidents") +
  theme_minimal()
```

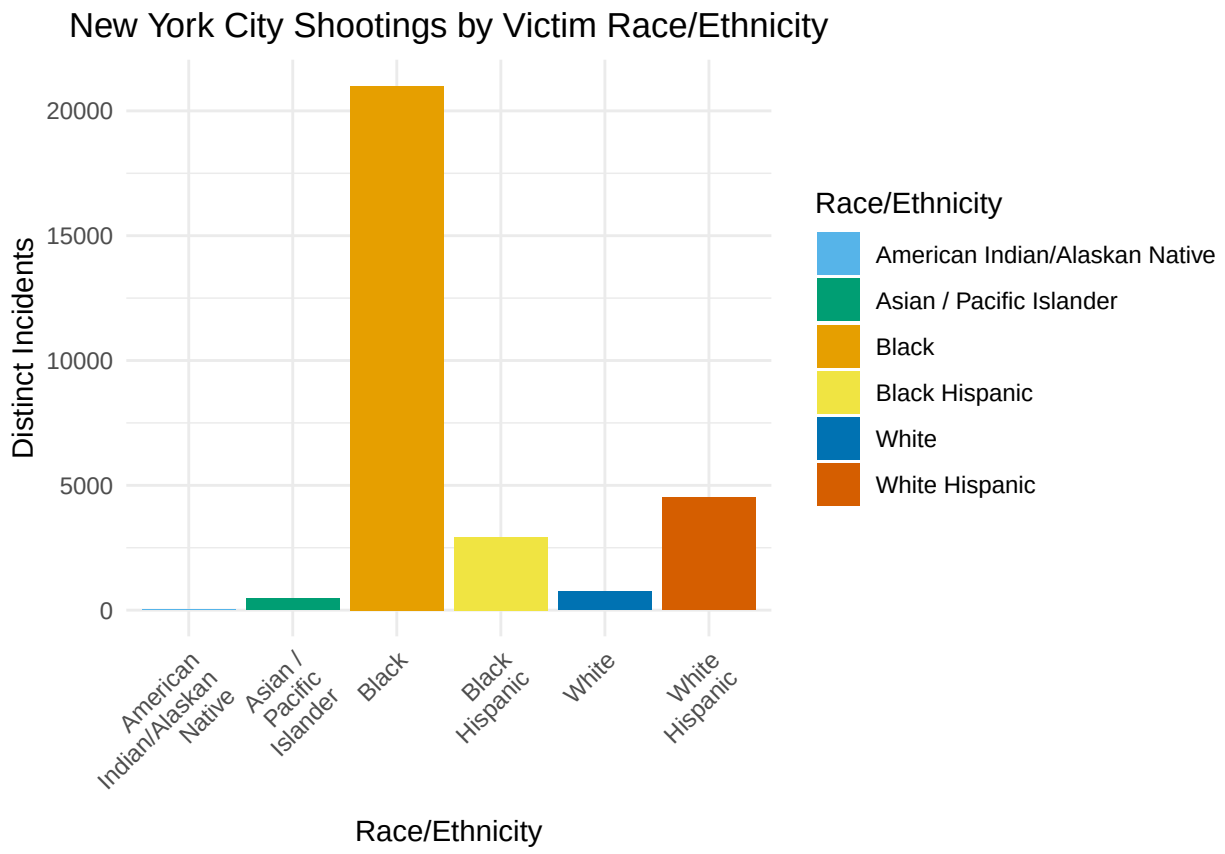



```r
# Exclude "UNKNOWN" values
filtered_perp_age <- nypd_shootings %>%
  filter(PERP_AGE_GROUP != "UNKNOWN")

# Create a bar plot with the filtered data
ggplot(filtered_perp_age, aes(x = PERP_AGE_GROUP)) +
  geom_bar() +
  labs(title = "New York City Shootings by Perpetrator Age",
       x = "Age Group",
       y = "Distinct Incidents") +
  theme_minimal()
```



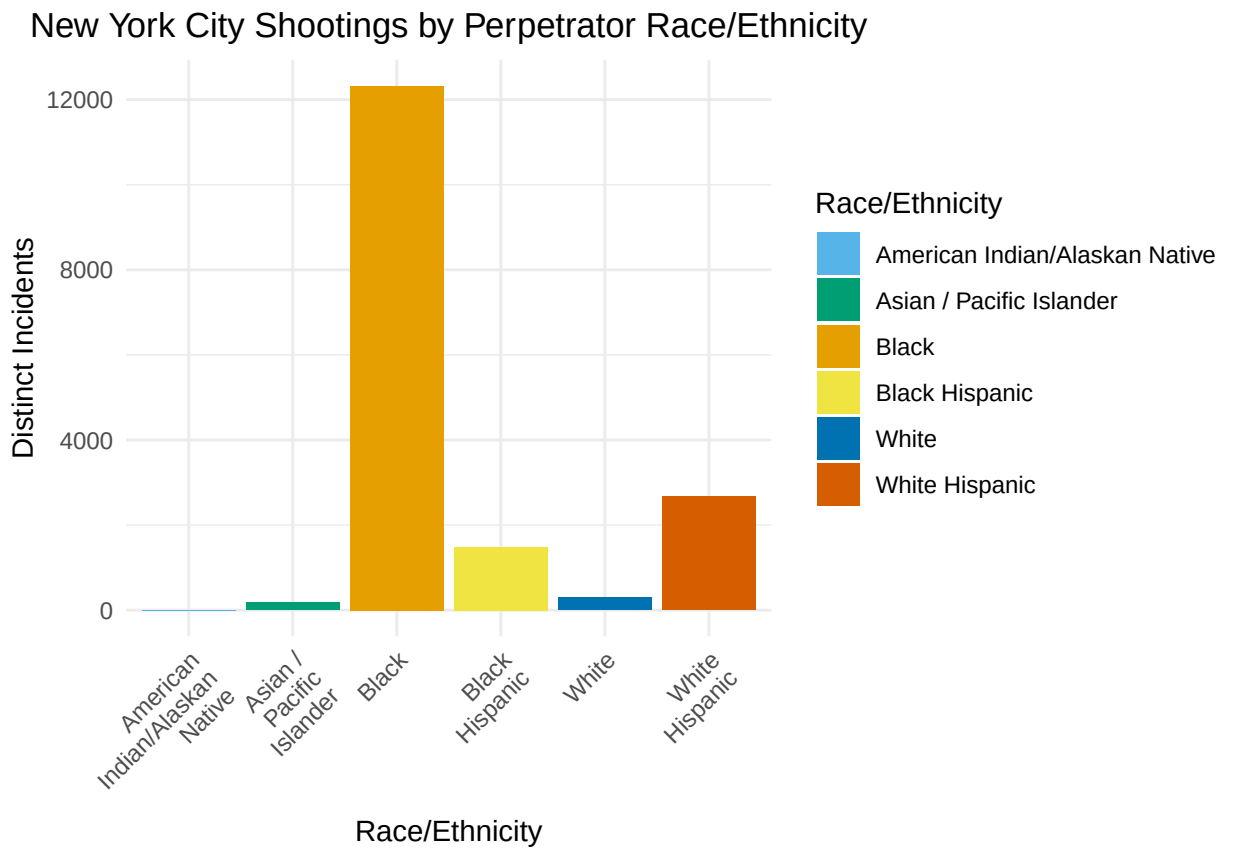
With respect to race/ethnicity of the victims and perpetrators where the race/ethnicity is known, we see the following:

```r
# Exclude "Unknown" values
filtered_vic_race <- nypd_shootings %>%
  filter(VIC_RACE != "Unknown")

# Create a bar plot
ggplot(filtered_vic_race, aes(x = VIC_RACE, fill = VIC_RACE)) +
  geom_bar() +
  scale_fill_manual(values = c(
    "#56B4E9", "#009E73", "#E69F00",
    "#F0E442", "#0072B2", "#D55E00", "#CC79A7"
  )) +
  labs(
    title = "New York City Shootings by Victim Race/Ethnicity",
    x = "Race/Ethnicity",
    y = "Distinct Incidents",
    fill = "Race/Ethnicity"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),  # angled labels
    plot.title = element_text(hjust = 0.5)
  ) +
  scale_x_discrete(labels = function(x) stringr::str_wrap(x, width = 12))
```



With respect to age of the perpetrator where the race/ethnicity is known, we see:

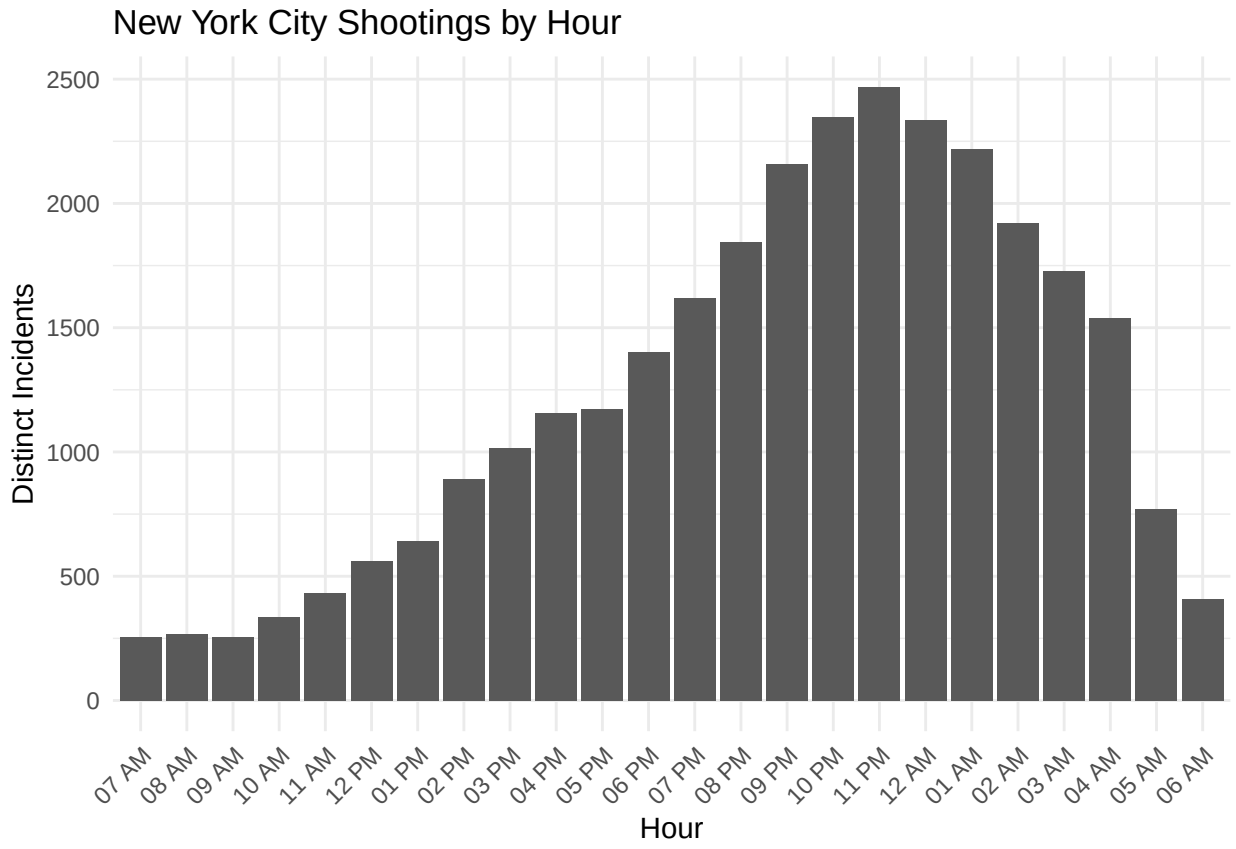
```
# Exclude "Unknown" values
filtered_perp_race <- nypd_shootings %>%
  filter(PERP_RACE != "Unknown")

# Create a bar plot
ggplot(filtered_perp_race, aes(x = PERP_RACE, fill = PERP_RACE)) +
  geom_bar() +
  scale_fill_manual(values = c(
    "#56B4E9", "#009E73", "#E69F00",
    "#F0E442", "#0072B2", "#D55E00", "#CC79A7"
  )) +
  labs(
    title = "New York City Shootings by Perpetrator Race/Ethnicity",
    x = "Race/Ethnicity",
    y = "Distinct Incidents",
    fill = "Race/Ethnicity"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),  # angled labels
    plot.title = element_text(hjust = 0.5)
  ) +
  scale_x_discrete(labels = function(x) stringr::str_wrap(x, width = 12))
```



As NYPD considers how best to allocate resources from a scheduling perspective, understanding time-of-day when shootings are occurring is important.

```
hour_counts <- nypd_shootings %>%
  count(OCCUR_HOUR)

min_hour <- hour_counts$OCCUR_HOUR[which.min(hour_counts$n)]

rotated_hours <- c(min_hour:23, 0:(min_hour-1))

nypd_shootings <- nypd_shootings %>%
  mutate(
    OCCUR_HOUR = factor(
      OCCUR_HOUR,
      levels = rotated_hours,
      labels = format(strptime(rotated_hours, format = "%H"), "%I %p")
    )
  )

ggplot(nypd_shootings, aes(x = OCCUR_HOUR)) +
  geom_bar() +
  labs(
    title = "New York City Shootings by Hour",
    x = "Hour",
    y = "Distinct Incidents"
  ) +
```

```
theme_minimal() +
theme(
  axis.text.x = element_text(angle = 45, hjust = 1)
)
```



Perhaps unsurprisingly we can see that 80% of shootings occur between the hours of 4PM and 5AM, and 63% between 9PM and 5AM. While inconvenient, these are the hours where having an alert police force is optimal for patrolling and responding to incidents.

Conclusion

There are many reasons that gun violence occurs; it is a complex issue. This research does not even begin to scratch the surface of the many facets and correlations of this type of crime. But, hopefully it gives my audience at least a little more information about the topic than they had before reading it.

Data is one of the primary keys to unlocking solid decisions-making. So, some key takeaways for how the NYPD can engage their community around the issue of gun violence:

1. Focus engagement especially with the male 18-44 population.
2. Consider disproportionate scheduling of patrols for the hours between 9PM and 5AM.
3. Direct more attention to The Bronx as the shootings per capita is higher in that borough than the rest.
4. Know that the resurgence of shootings in 2020 was more likely to be correlated with the global pandemic and less likely the result of something the NYPD did wrong. Shootings are likely to continue to decrease over the coming years to pre-pandemic levels.

Biases

- While on the surface the data strongly suggests that males 18-44 are statistically more likely to be involved in a shooting in New York City, either as a perpetrator or victim, it is possible that further study *could* suggest that by having NYPD engage a different segment of the population (for example: parents, partners, siblings of the male 18-44 population) it could yield equal or greater results in reducing gun violence with the male 18-44 population. It is my personal bias that police won't go wrong when they engage any at risk segment of the population, but that isn't to say there aren't other, effective strategies.
- One technique employed in this analysis that I have found effective in *removing* biases within the data I'm analyzing is converting a measurement from one of raw volume and contextualizing it with some other control variable. This can be seen in the shootings per capita by borough metric where we divide the raw volume of shootings by the population which acts as a control to help us interpret the true severity of the volume.