NYPD

2025-08-31

1-Load Library and read in Data

```
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   3.5.2     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(lubridate)
library(ggplot2)
```

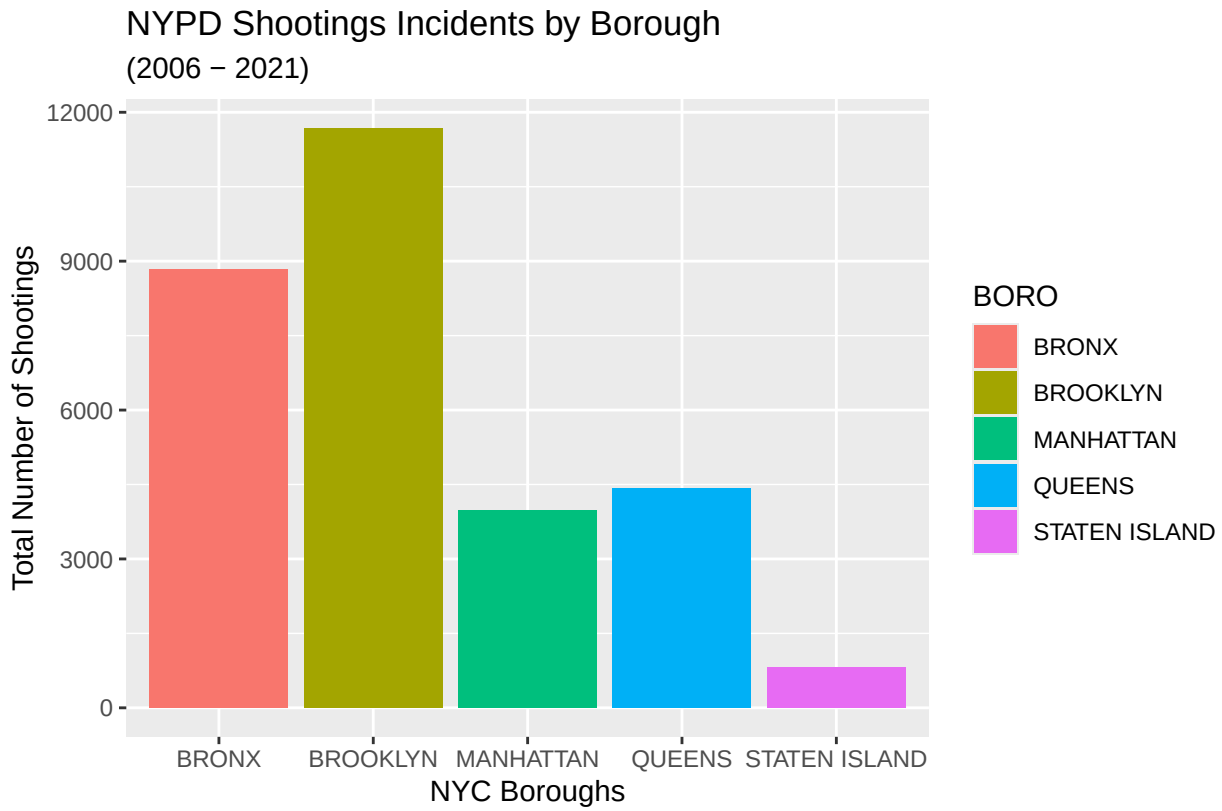
2-Read NYPD Data

```
url_NYPD <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

NYPD <- read.csv(url_NYPD)
```

3- Data Retrieval and Cleaning

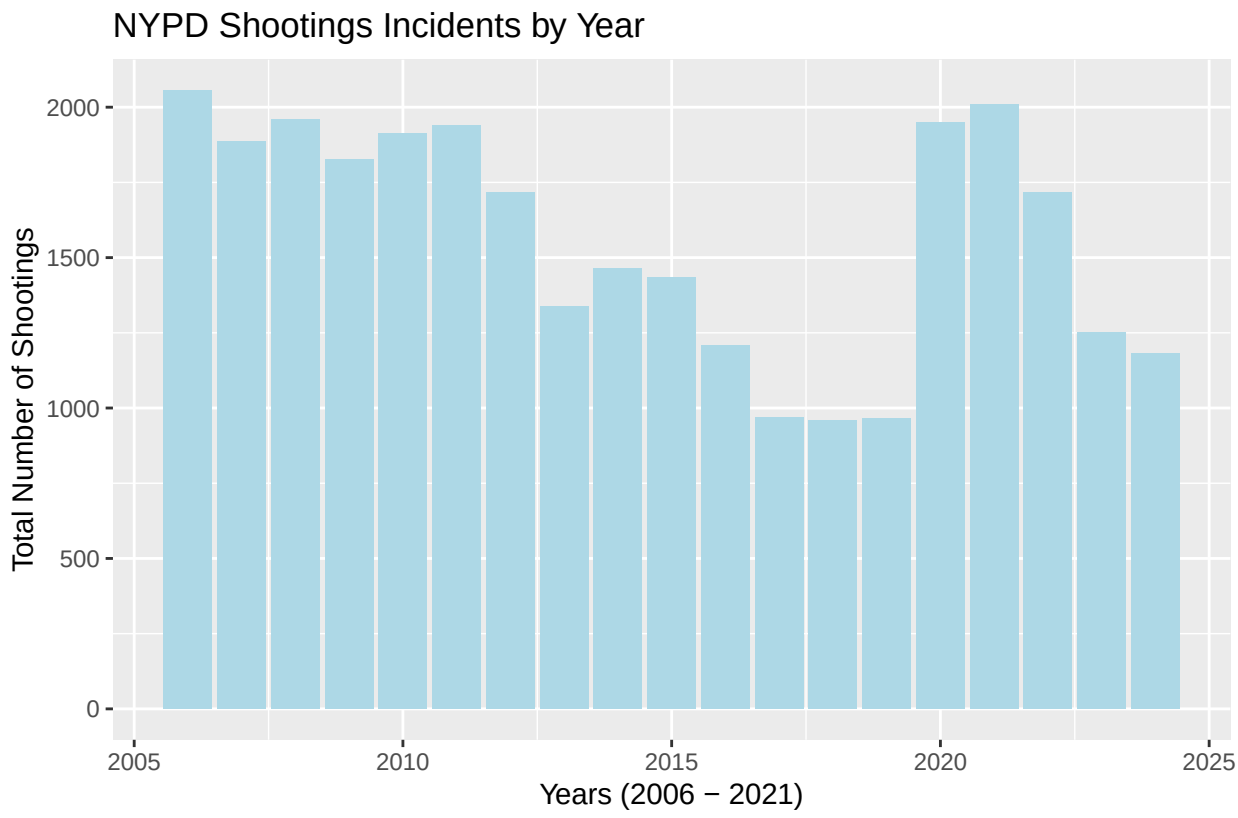
```
NYPD_clean <- NYPD %>%
  select(c("OCCUR_DATE","OCCUR_TIME","BORO","PRECINCT",
           "STATISTICAL_MURDER_FLAG","VIC_AGE_GROUP","VIC_SEX","VIC_RACE")) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         OCCUR_TIME = hms(OCCUR_TIME),
         STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG),
         Shootings = 1,
         Year = year(OCCUR_DATE))

NYPD_clean %>%
```

```
  ggplot(aes(x = BORO, fill = BORO)) +
  geom_bar() +
  labs(title = "NYPD Shootings Incidents by Borough",
       subtitle = "(2006 - 2021)",
       x = "NYC Boroughs",
       y = "Total Number of Shootings",
       caption = "(Figure - 1)")
```



(Figure − 1)

```
NYPD_clean %>%
  ggplot(aes(x = Year)) +
  geom_bar(fill = "lightblue",show.legend = FALSE) +
  labs(title = "NYPD Shootings Incidents by Year",
       x = "Years (2006 - 2021)",
       y = "Total Number of Shootings",
       caption = "(Figure - 2)")
```



(Figure − 2)

NYPD shooting incidents covering calendar years 2006 through 2021

(Figure - 1) Shootings by Borough

Visualization: Bar chart showing total shootings by borough. Insights: -Brooklyn has the highest number of shootings. -Staten Island has the lowest number.

(Figure - 2) Shootings by Year

Visualization: Bar chart displaying total shootings per year. Insights: -Shootings peaked around 2011. -A gradual decline in recent years.

Shootings Trends Over Years

```
NYPD_year <- NYPD_clean %>%
  group_by(Year,Shootings) %>%
  summarize(Shootings = sum(Shootings),
            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(Year,Shootings,STATISTICAL_MURDER_FLAG) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```
NYPD_year %>% slice_max(Shootings, n = 4)
```

```
## # A tibble: 4 x 3
##    Year Shootings STATISTICAL_MURDER_FLAG
```

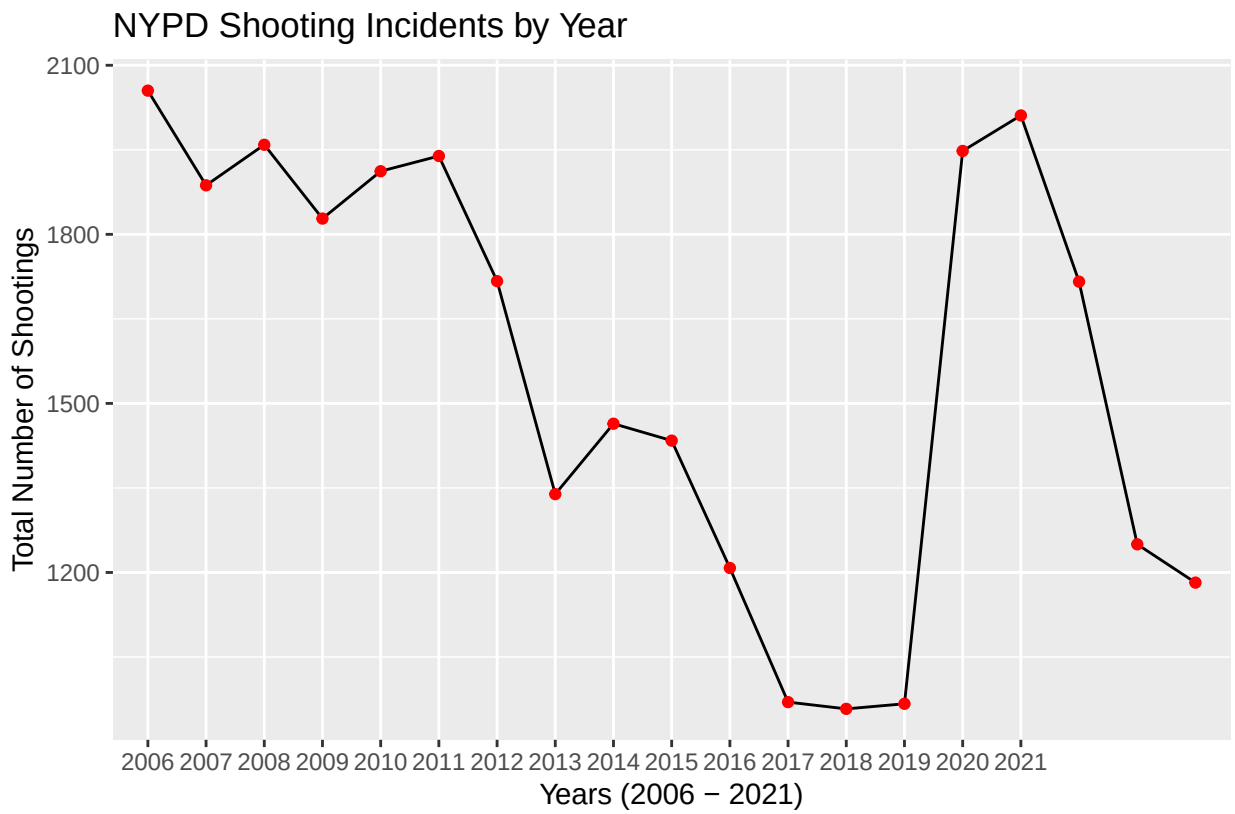
```
##   <dbl>     <dbl>                   <int>
## 1  2006      2055                     445
## 2  2021      2011                     428
## 3  2008      1959                     362
## 4  2020      1948                     366
```

```
NYPD_year %>% slice_min(Shootings, n = 4)
```

```
## # A tibble: 4 x 3
##    Year Shootings STATISTICAL_MURDER_FLAG
##   <dbl>     <dbl>                   <int>
## 1  2018       958                     204
## 2  2019       967                     184
## 3  2017       970                     174
## 4  2024      1182                     239
```

```
NYPD_year %>%
  ggplot(aes(x = Year, y = Shootings)) +
  geom_line() +
  geom_point(color="red") +
  scale_x_discrete(limits = c(2006:2021)) +
  labs(
    title = "NYPD Shooting Incidents by Year",
    x = "Years (2006 - 2021)",
    y = "Total Number of Shootings",
    caption = "(Figure - 3)")
```

```
## Warning in scale_x_discrete(limits = c(2006:2021)): Continuous limits supplied to discrete scale.
## i Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```



(Figure − 3)

(Figure - 3) Shootings Trends Over Years

Visualization: Line chart representing shootings trend over 2006-2021. Insights: -Overall decreasing trend in shootings. -Anomaly in 2021 (possible data incompleteness). Borough-wise Shootings Over Time

```
NYPD_boro <- NYPD_clean %>%
  group_by(BORO, OCCUR_DATE,Shootings) %>%
  summarize(Shootings = sum(Shootings),
            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(BORO,OCCUR_DATE,Shootings,STATISTICAL_MURDER_FLAG) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'BORO', 'OCCUR_DATE'. You can override
## using the `.groups` argument.
```

```
NYPD_boro_year <- NYPD_clean %>%
  mutate(Year = year(OCCUR_DATE)) %>%
  group_by(BORO, Year, Shootings) %>%
  summarize(Shootings = sum(Shootings),
            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(BORO, Year,Shootings,STATISTICAL_MURDER_FLAG) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'BORO', 'Year'. You can override using the
## `.groups` argument.
```

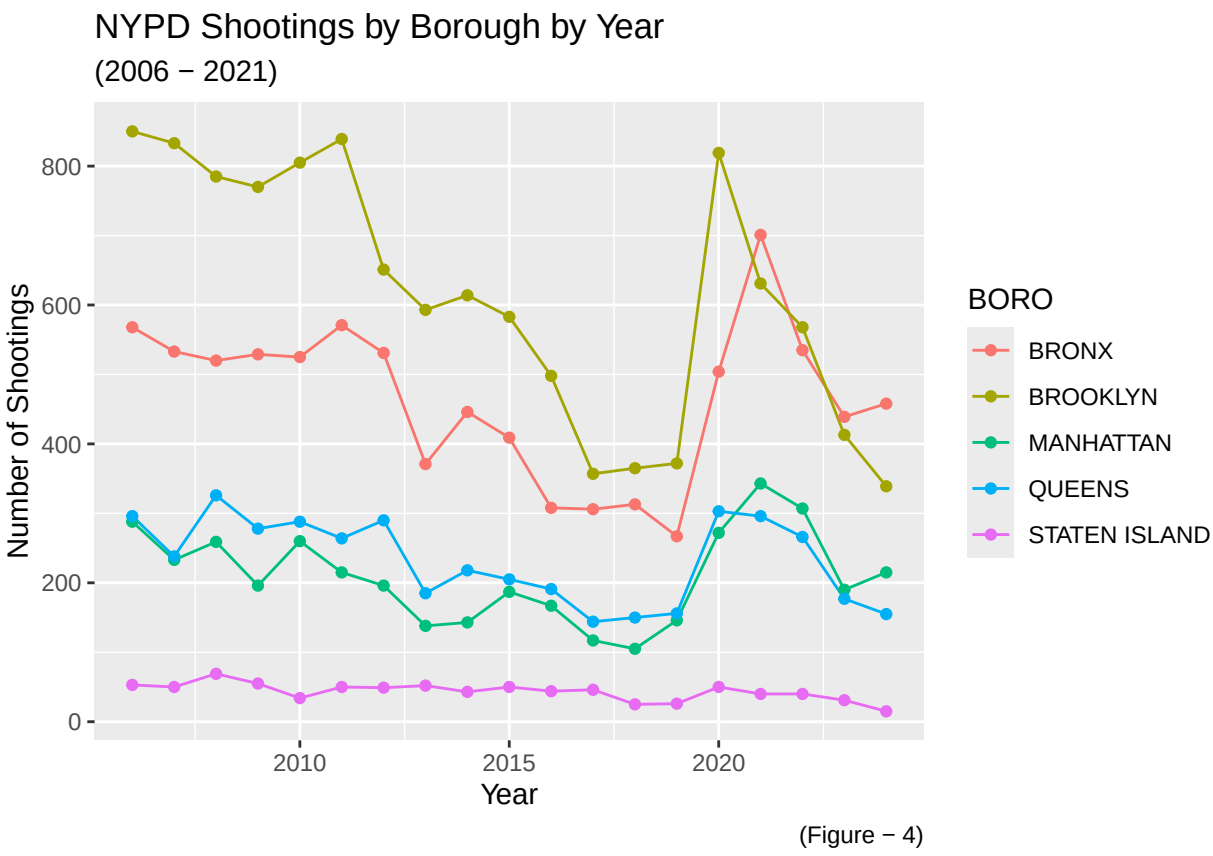
```
NYPD_boro_total <- NYPD_boro_year %>%
  group_by(BORO) %>%
  summarize(Shootings = sum(Shootings))
(7402 + 10365)/ sum(NYPD_boro_total$Shootings)
```

```
## [1] 0.5973306
```

```
736/ sum(NYPD_boro_total$Shootings)
```

```
## [1] 0.02474449
```

```
NYPD_boro_year %>%
  ggplot(aes(x = Year, y = Shootings,color = BORO)) +
  geom_line() +
  geom_point() +
  labs(title = "NYPD Shootings by Borough by Year",
       subtitle = "(2006 - 2021)",
       x = "Year",
       y = "Number of Shootings",
       caption = "(Figure - 4)")
```



(Figure - 4)

Visualization: Line chart depicting borough-wise shootings by year. Insights: -Brooklyn consistently has the highest shootings. -Other boroughs show varying trends. NYPD Shootings Per Day

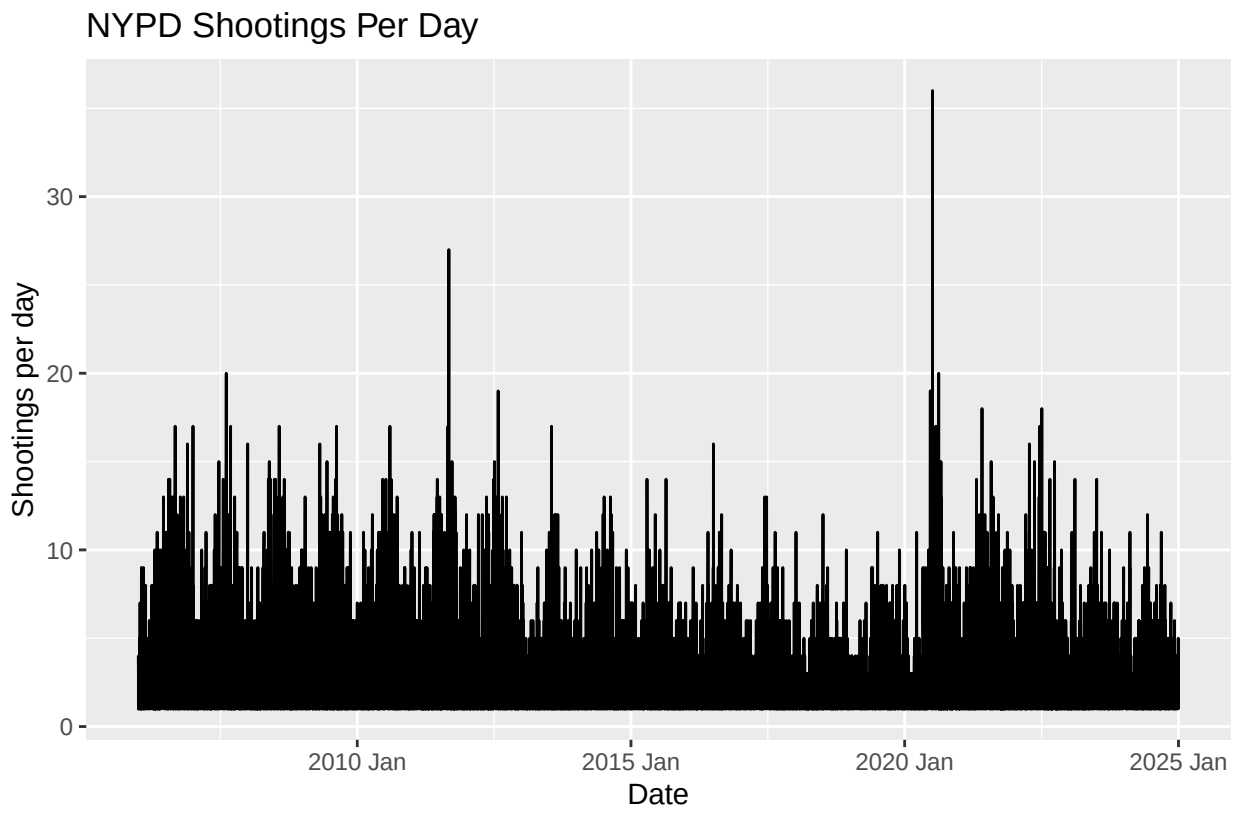
```
NYPD_date <- NYPD_clean %>%
  group_by(OCCUR_DATE,Shootings,STATISTICAL_MURDER_FLAG) %>%
  summarize(Shootings = sum(Shootings),

            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(OCCUR_DATE,Shootings,STATISTICAL_MURDER_FLAG) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'OCCUR_DATE', 'Shootings'. You can override
## using the `.groups` argument.
```

```
NYPD_date %>% slice_max(Shootings, n=2)
```

```
## # A tibble: 2 x 3
##   OCCUR_DATE Shootings STATISTICAL_MURDER_FLAG
##   <date>         <dbl>                   <int>
## 1 2020-07-05        36                       0
## 2 2011-09-04        27                       0
```

```
NYPD_date %>%
  ggplot(aes(x = OCCUR_DATE, y = Shootings)) +
  geom_line() +
  scale_x_date(date_labels = "%Y %b") +
  labs(title = "NYPD Shootings Per Day",
       subtile = "(2006 - 2021)",
       x = "Date",
       y = "Shootings per day",
       caption = "(Figure - 5)")
```



(Figure – 5)

(Figure - 5)

Visualization: Line chart illustrating the daily count of shootings from 2006 to 2021. Insights: -Peaks and valleys in shooting incidents over the years. -Highest shooting days around July 4th, 2023.

Shootings on July 5, 2023

```
NYPD_time_year <- NYPD_clean %>%
  mutate(Time_year = format(as.Date(OCCUR_DATE), "%m/%d")) %>%
  mutate(Time_year = as.Date(Time_year,"%m/%d")) %>%
  group_by(Time_year,Shootings) %>%
  summarize(Shootings = sum(Shootings),
            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(Time_year,Shootings,STATISTICAL_MURDER_FLAG) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Time_year'. You can override using the
## `.groups` argument.
```

```
NYPD_time_year %>% slice_max(Shootings, n = 2)
```

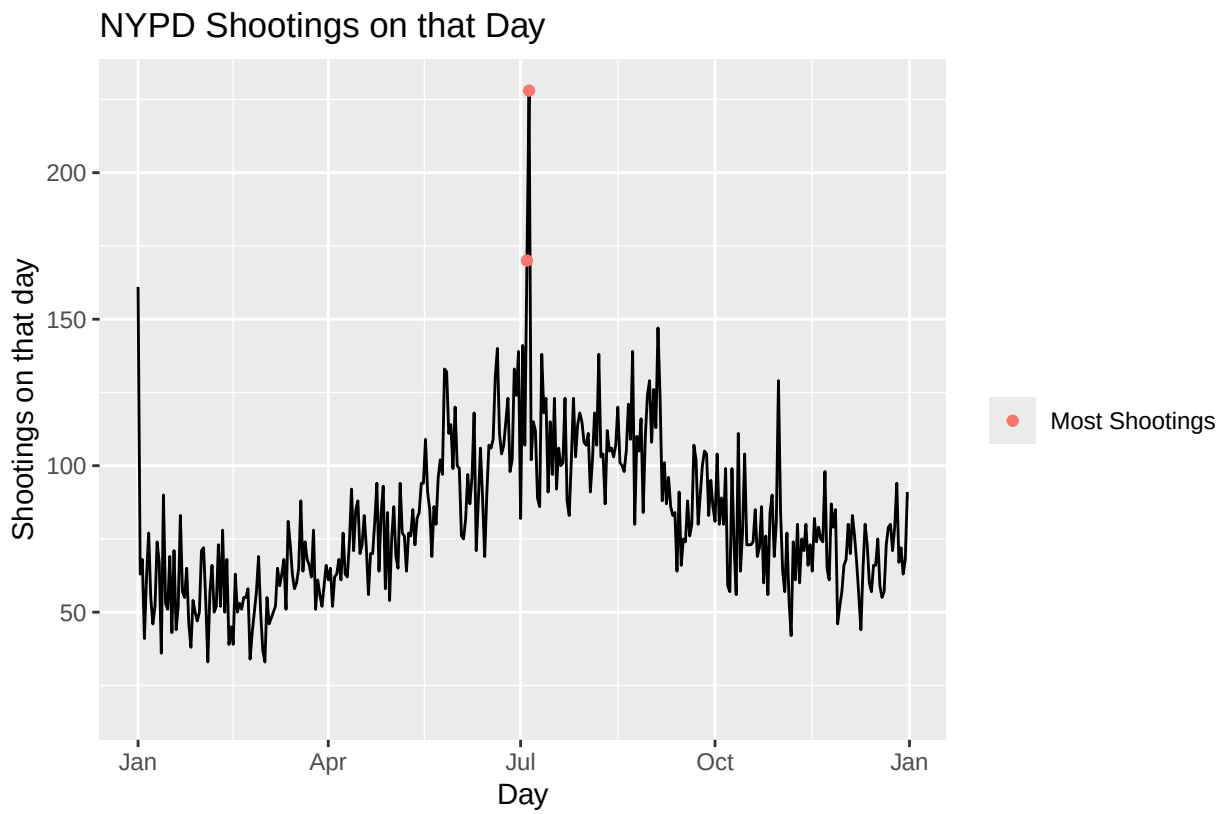
```
## # A tibble: 2 x 3
##   Time_year  Shootings STATISTICAL_MURDER_FLAG
##   <date>         <dbl>                  <int>
## 1 2025-07-05       228                     39
## 2 2025-07-04       170                     28
```

```
NYPD_July_5 <- NYPD_clean %>%
  mutate(Time_year = format(as.Date(OCCUR_DATE), "%m/%d"),
         Hour = hour(OCCUR_TIME)) %>%
  mutate(Time_year = as.Date(Time_year,"%m/%d")) %>%
  filter(Time_year == "2022-07-05") %>%
  group_by(Hour,Shootings) %>%
  summarize(Shootings = sum(Shootings))
```

```
## `summarise()` has grouped output by 'Hour'. You can override using the
## `.groups` argument.
```

```
NYPD_time_year %>%
  ggplot(aes(x = Time_year, y = Shootings)) +
  geom_line() +
  geom_point(data = NYPD_time_year %>% slice_max(Shootings, n = 2),
             aes(color="Most Shootings")) +
  scale_x_date(date_labels = "%b") +
  labs(title = "NYPD Shootings on that Day",
       subtile = "(2006 - 2021)",
       colour = "",
       x = "Day",
       y = "Shootings on that day",
       caption = "(Figure - 6)")
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_line()`).
```



(Figure - 6)

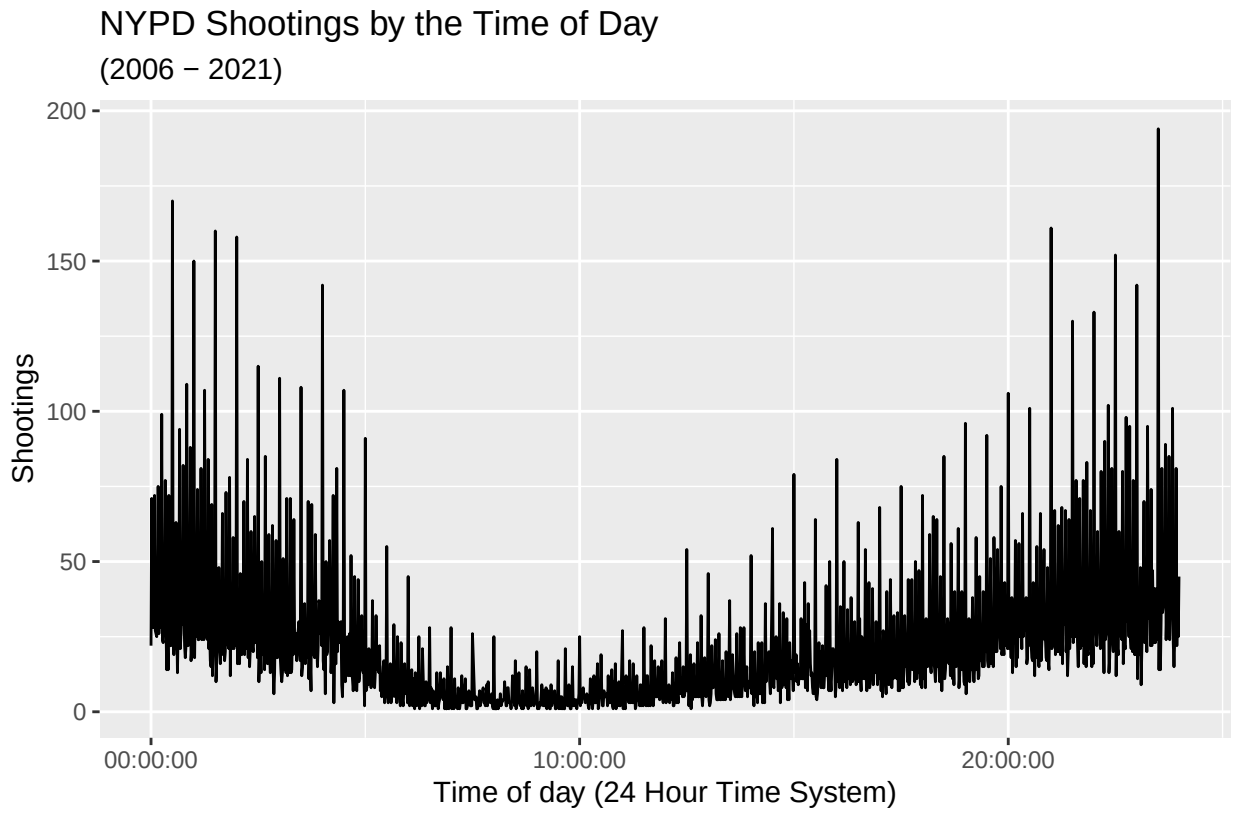
Visualization: Line chart showing shootings on July 5, 2023, with a focus on the top two shooting days. Insights: -July 5, 2023, had the highest number of shootings. -The second-highest shootings occurred on July 4, 2023.

NYPD Shootings by the Time of Day

```
NYPD_time_day <- NYPD_clean %>%
  group_by(OCCUR_TIME,Shootings) %>%
  summarize(Shootings = sum(Shootings),
            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(OCCUR_TIME,Shootings,STATISTICAL_MURDER_FLAG)
```

```
## `summarise()` has grouped output by 'OCCUR_TIME'. You can override using the
## `.groups` argument.
```

```
NYPD_time_day %>%
  ggplot(aes(x = OCCUR_TIME, y = Shootings)) +
  geom_line() +
  scale_x_time() +
  labs(title = "NYPD Shootings by the Time of Day",
       subtitle = "(2006 - 2021)",
       x = "Time of day (24 Hour Time System)",
       y = "Shootings",
       caption = "(Figure - 7)")
```



(Figure − 7)

```
NYPD_time_hour <- NYPD_clean %>%
  mutate(Hour = hour(OCCUR_TIME)) %>%
  group_by(Hour,Shootings) %>%
  summarize(Shootings = sum(Shootings),
            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG)) %>%
  mutate(Hour2 = Hour^2) %>%
  select(Hour,Shootings,STATISTICAL_MURDER_FLAG, Hour2)
```

```
## `summarise()` has grouped output by 'Hour'. You can override using the
## `.groups` argument.
```

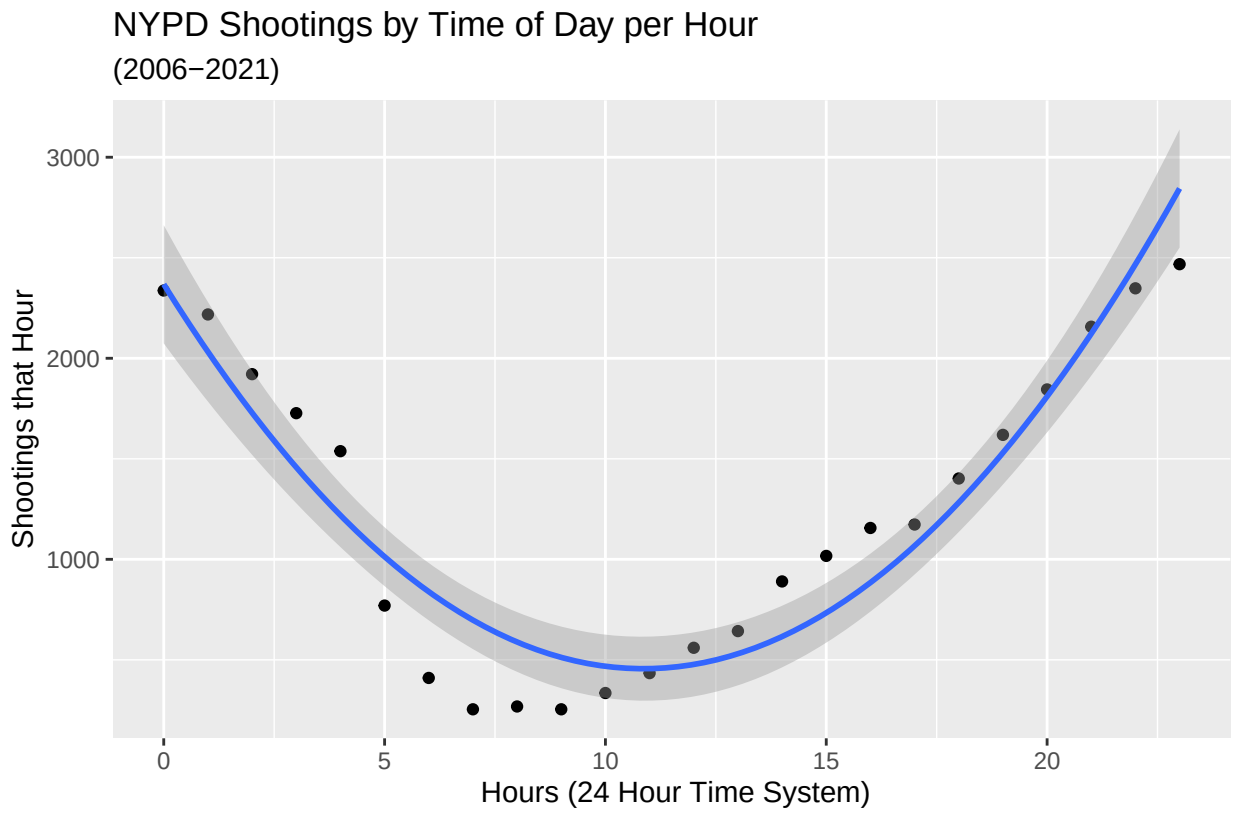
```
NYPD_time_hour_model <- lm(data = NYPD_time_hour, Shootings ~ Hour + Hour2)
summary(NYPD_time_hour_model)
```

```
##
## Call:
## lm(formula = Shootings ~ Hour + Hour2, data = NYPD_time_hour)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -443.63 -160.45   59.06  187.37  318.96
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2368.156    141.318   16.76 1.25e-13 ***
```

```
## Hour        -352.119     28.461  -12.37 4.14e-11 ***
## Hour2         16.210      1.195   13.56 7.37e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 250.3 on 21 degrees of freedom
## Multiple R-squared:  0.9013, Adjusted R-squared:  0.8919
## F-statistic: 95.93 on 2 and 21 DF,  p-value: 2.744e-11
```

```
NYPD_time_hour %>%
  ggplot(aes(x = Hour, y = Shootings)) +
  geom_point() +
  stat_smooth(method = "lm", formula = y ~ x + I(x^2), size = 1) +
  labs(title = "NYPD Shootings by Time of Day per Hour",
       subtitle = "(2006-2021)",
       x = " Hours (24 Hour Time System)",
       y = "Shootings that Hour",
       caption = "(Figure - 8)")
```

```
## 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.
```



(Figure − 8)

(Figure - 7)

Visualization: Line chart presenting the distribution of shootings throughout the 24-hour time system. Insights: -Shootings tend to be higher during the early morning hours.

(Figure - 8) Time of Day Analysis with Polynomial Regression

Model: Polynomial regression model applied to analyze shootings per hour.

Model Insights: -The model suggests a quadratic relationship between the hour of the day and shootings. -Residuals indicate a good fit to the data. -Shooting incidents are concentrated during specific hours.

Conclusion:

Summary: The analysis provides a comprehensive understanding of NYPD shootings incidents, considering temporal and spatial dimensions. Key Findings: Peaks in shootings around specific days (e.g., July 4, 2023). Concentration of shootings during certain hours of the day. A decreasing trend in overall shootings from 2006 to 2021.

Recommendations: Investigate anomalies in 2021 and explore potential contributing factors. Conduct further analysis to understand the temporal patterns in more detail. Consider additional data sources to enrich the analysis.