NYPD Shooting Incident

Lin

2024-03-19

```
library(stringr)
library(readr)
library(dplyr)
library(ggplot2)
library(rpart)
library(leaflet)
library(htmlwidgets)
```

Import Data

Import NYPD Shooting Data from https://data.cityofnewyork.us

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

Clean data

```
NYPD_cases <- ORG_NYPD_cases %>%
 select(-c(INCIDENT_KEY, X_COORD_CD, Y_COORD_CD, Lon_Lat)) %>%
 mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y"))
```

Date with MURDER_FLAG

Through analyzing the relationship between date and the occurrence of MURDER_FLAG cases, it was found that the proportion of cases with MURDER_FLAG per month typically ranged between 16% and 23%. However, upon further examination, it was noted that in many months, the proportion of MURDER_FLAG cases exceeded 23%, indicating a concerning trend.

```
# Select case for MURDER_FLAG = TRUE
murder_cases <- NYPD_cases %>%
  mutate(date = format(OCCUR_DATE, "%Y/%m")) %>%
  group_by(date) %>%
  summarize(
    flag_cases = sum(STATISTICAL_MURDER_FLAG == TRUE),
    total_cases = n(),
    case_rate=flag_cases/total_cases)

summary(murder_cases)
```

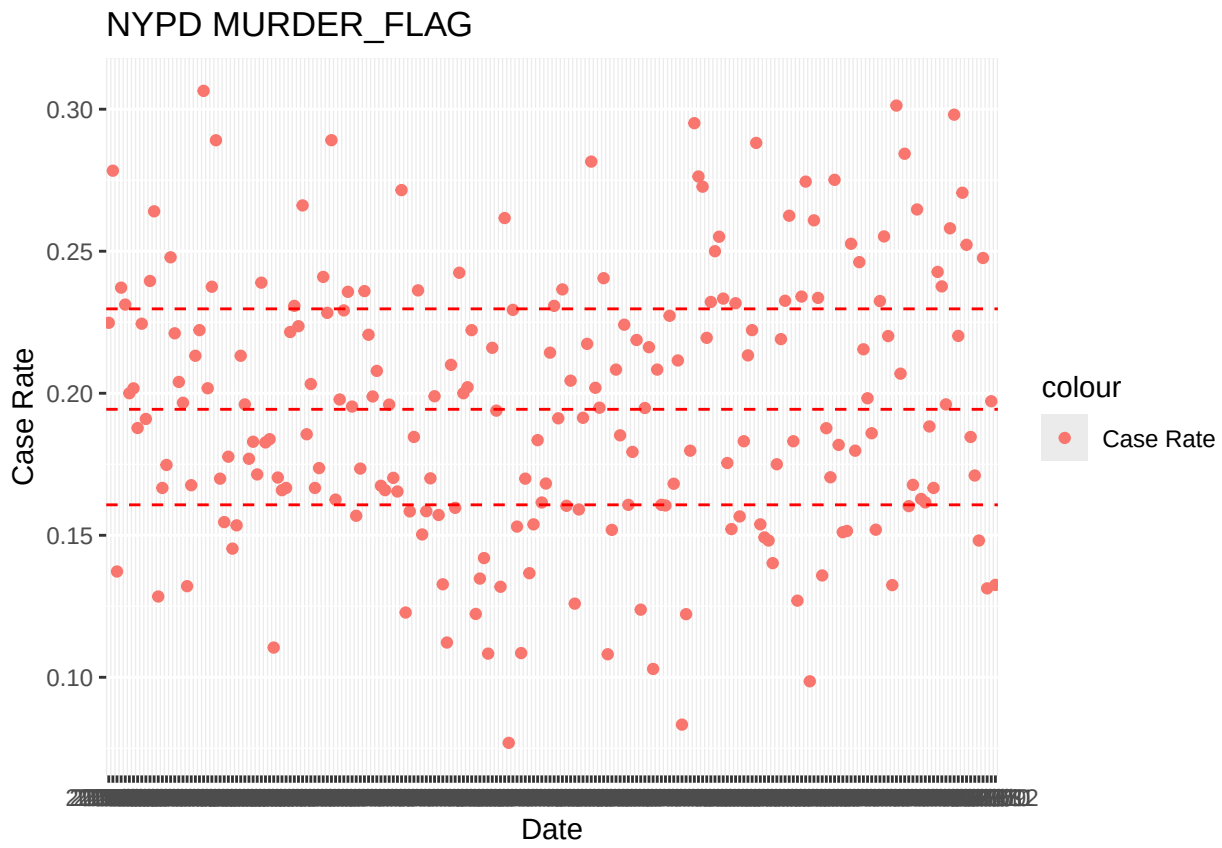
```
##      date              flag_cases     total_cases       case_rate
##  Length:216         Min.   : 4.00   Min.   : 41.00   Min.   :0.07692
##  Class :character   1st Qu.:17.00   1st Qu.: 95.75   1st Qu.:0.16071
##  Mode  :character   Median :24.00   Median :120.00   Median :0.19434
##                     Mean   :25.58   Mean   :132.23   Mean   :0.19495
##                     3rd Qu.:33.00   3rd Qu.:167.00   3rd Qu.:0.22971
##                     Max.   :61.00   Max.   :325.00   Max.   :0.30645
```

```r
# calculate quantile
vline_positions <- quantile(murder_cases$case_rate, probs = c(0.25, 0.5, 0.75))

vline_positions
```

```
##       25%       50%       75%
## 0.1607143 0.1943414 0.2297107
```

```r
# ggplot with quantile line
murder_cases %>%
  ggplot(aes(x = date, y = case_rate)) +
  geom_point(aes(color = "Case Rate")) +
  labs(title = "NYPD MURDER_FLAG", x = "Date", y = "Case Rate") +
  geom_hline(yintercept = vline_positions, linetype = "dashed", color = "red")
```

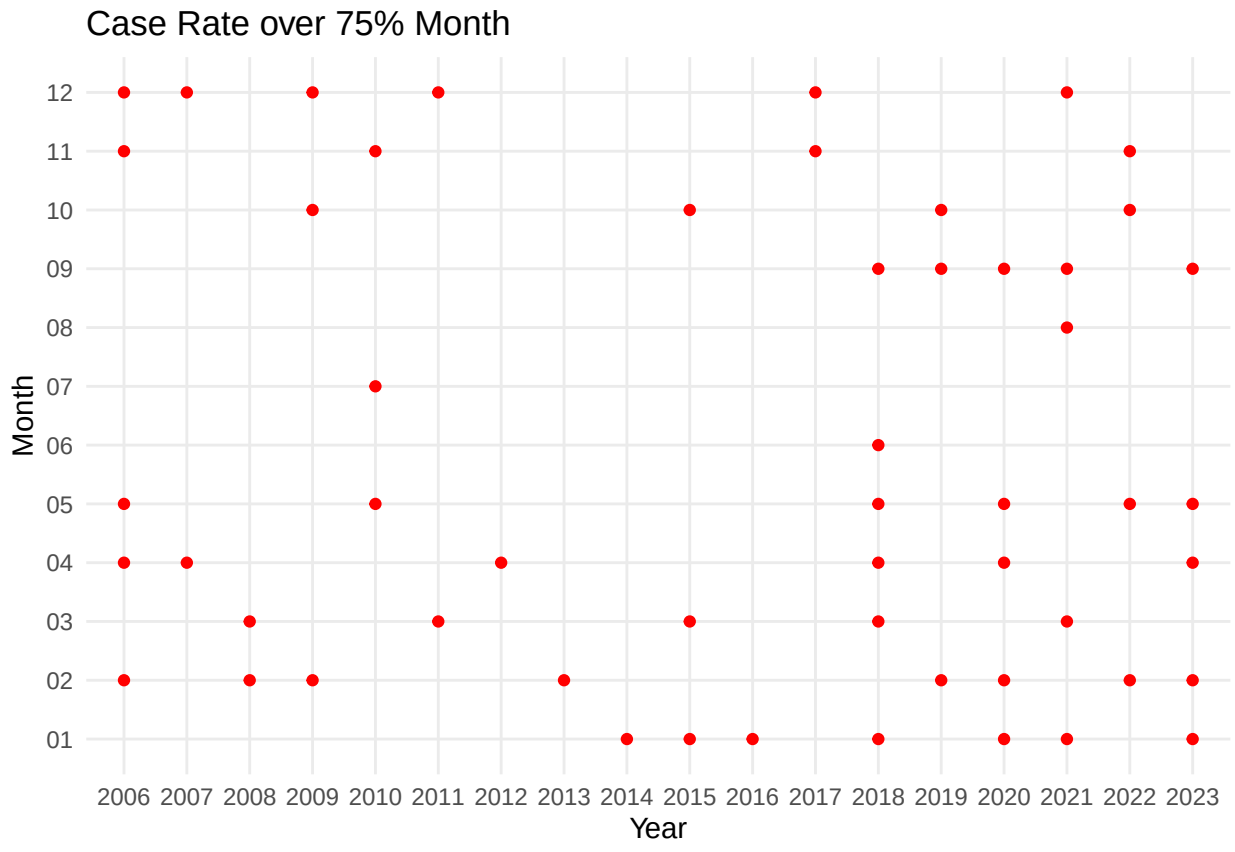


Case Rate over 75%

By plotting a scatter plot, it was observed that since 2018, there has been a significant increase in the proportion of MURDER_FLAG cases, consistently exceeding the average for three-month periods. This upward trend in the proportion of MURDER_FLAG cases suggests a concerning rise in violent crimes over time.

```
# select case over 75%
murder_cases_over <- murder_cases %>%
  mutate(year_month = strsplit(date, "/"),
         year = sapply(year_month, `[`, 1),
         month = sapply(year_month, `[`, 2)) %>%
  filter(case_rate > vline_positions[3]) %>%
  select(year, month)

ggplot(murder_cases_over, aes(x = year, y = month)) +
  geom_point(color = "red") +
  labs(title = "Case Rate over 75% Month", x = "Year", y = "Month") +
  scale_color_brewer(palette = "Set1") +
  theme_minimal()
```




```
# calculate how many months in a year
cases_count_per_year <- murder_cases_over %>%
  group_by(year) %>%
  summarize(count = (n()))
```

```r
summary(cases_count_per_year)
```

```
##      year               count
##  Length:18          Min.   :1.00
##  Class :character   1st Qu.:2.00
##  Mode  :character   Median :3.00
##                     Mean   :3.00
##                     3rd Qu.:4.75
##                     Max.   :6.00
```

```r
# Filter the years with case-month greater than the average value.
cases_count_per_year %>%
  filter(count>3)
```

```
## # A tibble: 6 x 2
##   year  count
##   <chr> <int>
## 1 2006      5
## 2 2018      6
## 3 2020      5
## 4 2021      5
## 5 2022      4
## 6 2023      5
```

Coordinate with MURDER_FLAG

Analyzing the geographical distribution of MURDER_FLAG cases over three-year intervals revealed that the majority of these cases were concentrated around the "Yankee Stadium" area and the "Brooklyn" district. While there was a decreasing trend in the density of cases in the "Brooklyn" district over time, the density remained high around the "Yankee Stadium" area. Further analysis is warranted to determine if the cases around the "Yankee Stadium" area are related to sporting events.

```r
# Flag with coordinate
Location_cases <- NYPD_cases %>%
  mutate(year = format(OCCUR_DATE, "%Y"),month = format(OCCUR_DATE, "%m")) %>%
  rename(Lat = Latitude, Lon = Longitude) %>%
  select(year,month,STATISTICAL_MURDER_FLAG,Lat,Lon) %>%
  group_by(year, month,Lat,Lon) %>%
  summarize(
    flag_cases = sum(STATISTICAL_MURDER_FLAG == TRUE),
    total_cases = n(),
    case_rate=flag_cases/total_cases) %>%
  filter(flag_cases>0)

# Create NY map
start_year <- "2006"
end_year <- "2022"

for (i in seq(start_year, end_year, by = 3)) {
  ny_map <- leaflet() %>%
    setView(lng = -73.90882, lat = 40.73346, zoom = 10)
```

```
  ny_map <- ny_map %>% addTiles()

  color <- colorFactor(palette = "Dark2", domain = Location_cases$flag_cases)

  ny_map <- ny_map %>% addCircleMarkers(
  data = Location_cases %>%
    filter(year >= i & year <= i+2),
  lng = ~Lon,
  lat = ~Lat,
  radius = 1,
  color = ~color(flag_cases),
  label = ~paste("Value:", flag_cases)
  )

  ny_map <- ny_map %>%
    addControl(
      html = paste("<b>Year:</b>", i, "-", i + 2),
      position = "topright"
    )

  print(ny_map)
  saveWidget(ny_map, paste0("ny_map_", i, ".html"))
}
```

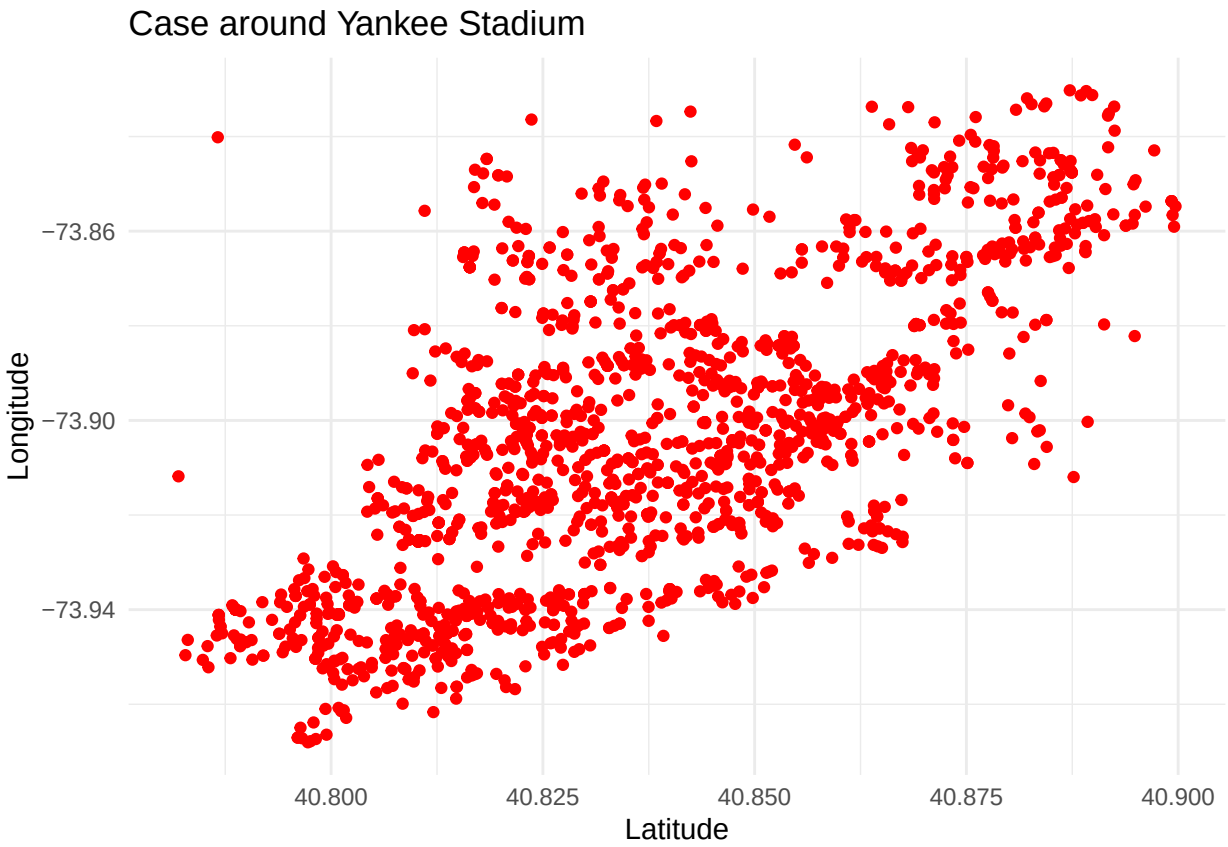
```
## Warning in validateCoords(lng, lat, funcName): Data contains 4 rows with either
## missing or invalid lat/lon values and will be ignored
```

Case/Pred around "Yankee Stadium" and "Brooklyn"

Given the high density of MURDER_FLAG cases around the "Yankee Stadium" area and the "Brooklyn" district, additional analysis was conducted to understand the variation in cases and develop predictive models. The regression line for the "Yankee Stadium" area remained relatively flat, indicating a persistent high rate of cases that may require a reevaluation of law enforcement strategies. Conversely, the regression line for the "Brooklyn" district showed a downward trend, indicating a continual decrease in MURDER_FLAG cases and an improvement in public safety.

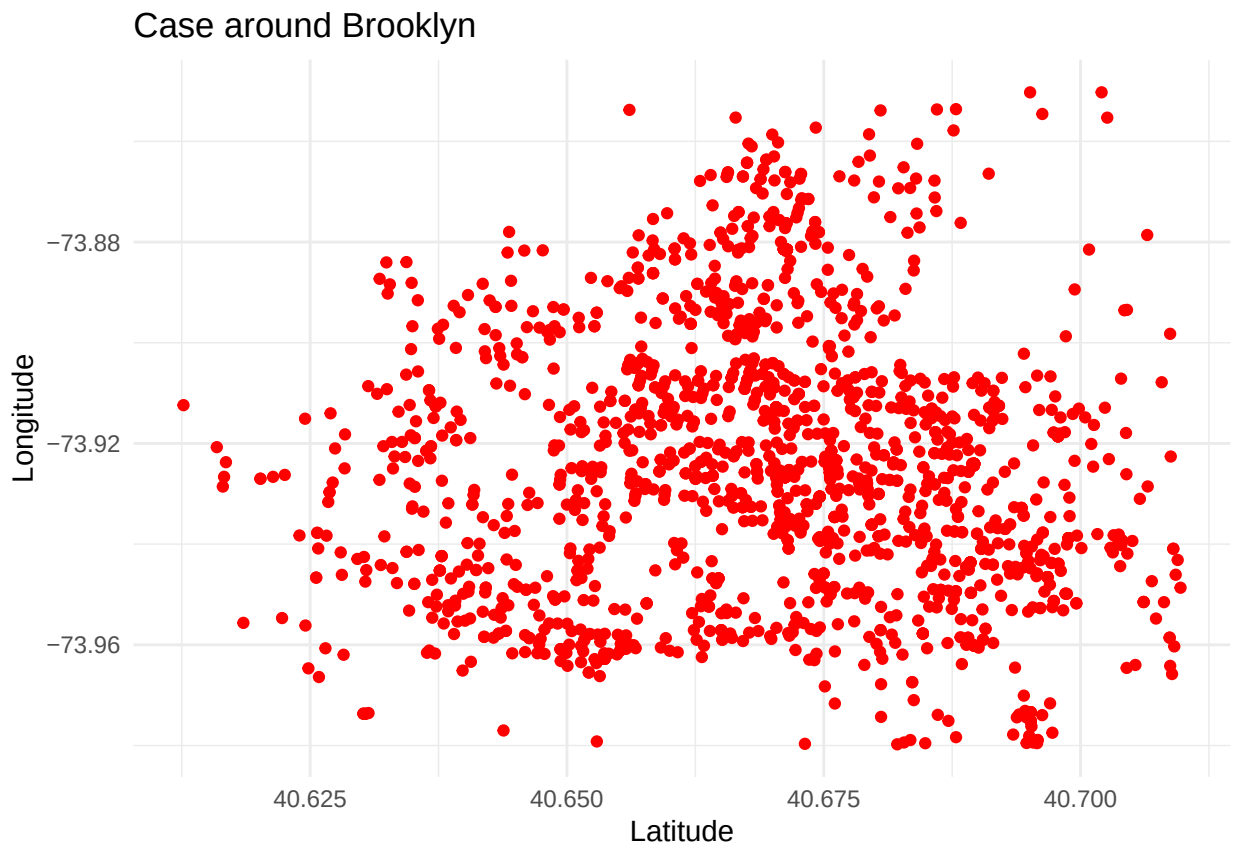
```
# MOD for Yankee_case and Brooklyn_case
Yankee_case <- Location_cases %>%
  filter(Lat >= 40.779 & Lat <= 40.9
         & Lon >= -73.97 & Lon <= -73.83
         )

ggplot(Yankee_case, aes(x = Lat, y = Lon)) +
  geom_point(color = "red") +
  labs(title = "Case around Yankee Stadium", x = "Latitude", y = "Longitude") +
  scale_color_brewer(palette = "Set1") +
  theme_minimal()
```




```
Brooklyn_case <- Location_cases %>%
  filter(Lat >= 40.61 & Lat <= 40.71
         & Lon >= -73.98 & Lon <= -73.85
         )

ggplot(Brooklyn_case, aes(x = Lat, y = Lon)) +
  geom_point(color = "red") +
  labs(title = "Case around Brooklyn", x = "Latitude", y = "Longitude") +
  scale_color_brewer(palette = "Set1") +
  theme_minimal()
```

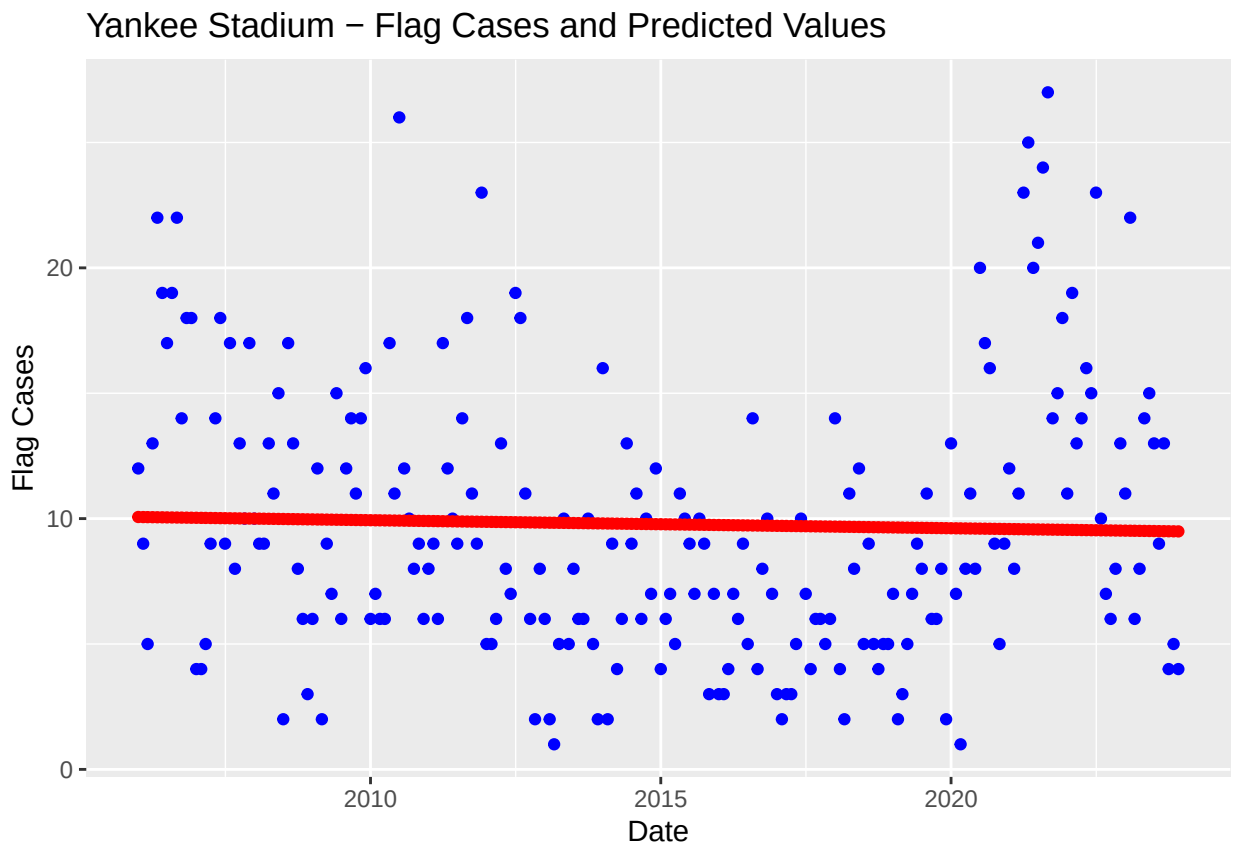



```
# Yankee case pred
Yankee_case <- Yankee_case %>%
  mutate(date = as.Date(paste(year, month, "01", sep = "-"))) %>%
  group_by(date) %>%
  summarise(total_flag_cases = sum(flag_cases))

mod <- lm(total_flag_cases ~ date, data = Yankee_case)

Yankee_case_pred <- Yankee_case %>% mutate(pred = predict(mod))

Yankee_case_pred %>% ggplot() +
  geom_point(aes(x = date, y = total_flag_cases), color = "blue")+
  geom_point(aes(x = date, y = pred), color = "red")+
  labs(title = "Yankee Stadium - Flag Cases and Predicted Values",
       x = "Date",
       y = "Flag Cases")
```

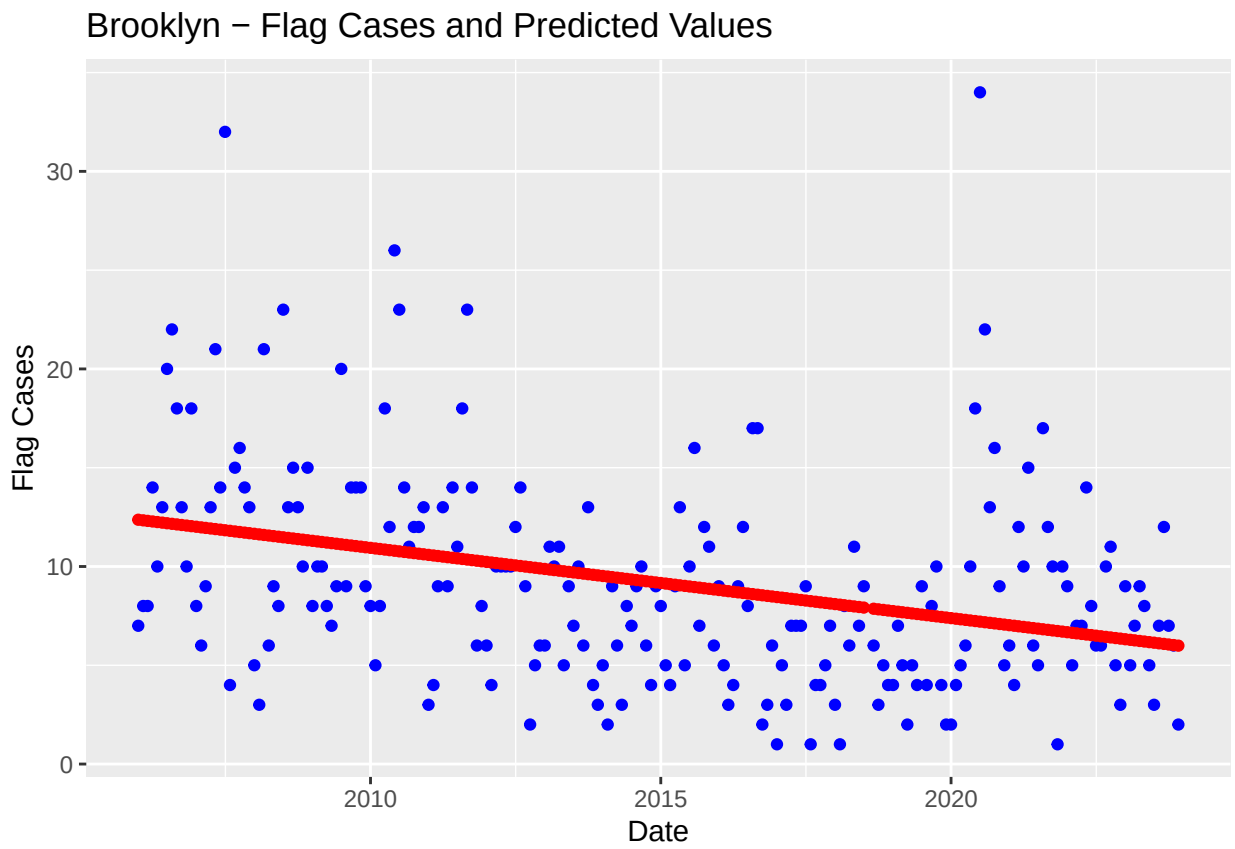



```
# Brooklyn case pred
Brooklyn_case <- Brooklyn_case %>%
  mutate(date = as.Date(paste(year, month, "01", sep = "-"))) %>%
  group_by(date) %>%
  summarise(total_flag_cases = sum(flag_cases))

mod <- lm(total_flag_cases ~ date, data = Brooklyn_case)

Brooklyn_case_pred <- Brooklyn_case %>% mutate(pred = predict(mod))

Brooklyn_case_pred %>% ggplot() +
  geom_point(aes(x = date, y = total_flag_cases), color = "blue")+
  geom_point(aes(x = date, y = pred), color = "red")+
  labs(title = "Brooklyn - Flag Cases and Predicted Values",
       x = "Date",
       y = "Flag Cases")
```



Conclusion and Potential biases

The analysis highlights the concerning trend of increasing MURDER_FLAG cases since 2018 and the persistent high rate of cases around the "Yankee Stadium" area. Efforts to address the underlying factors contributing to violent crimes in these areas, such as potential socio-economic issues or inadequate law enforcement resources, may be necessary to mitigate the risk of further escalation in violent crimes.

Potential biases in the analysis may arise from limitations in the dataset, such as incomplete or inaccurate data on crime incidents or socio-economic factors. Additionally, the analysis may be influenced by the researcher's subjective interpretation of the data and prior assumptions about the factors influencing crime rates. Further research incorporating a broader range of variables and perspectives may provide a more comprehensive understanding of the factors contributing to MURDER_FLAG cases and inform more effective strategies for crime prevention and intervention.