NYPDShooting

```
library(tidyverse)
library(lubridate)
library(zoo)
```

```
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv"
NYPDShooting <- read_csv(url_in)
shooting <- NYPDShooting %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  select(-c(JURISDICTION_CODE, LOCATION_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))
summary(shooting)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Min.   :2006-01-01   Length:27312      Length:27312
##  1st Qu.: 63860880   1st Qu.:2009-07-18   Class1:hms        Class :character
##  Median : 90372218   Median :2013-04-29   Class2:difftime   Mode  :character
##  Mean   :120860536   Mean   :2014-01-06   Mode  :numeric
##  3rd Qu.:188810230   3rd Qu.:2018-10-15
##  Max.   :261190187   Max.   :2022-12-31
##  LOC_OF_OCCUR_DESC     PRECINCT      LOC_CLASSFCTN_DESC STATISTICAL_MURDER_FLAG
##  Length:27312       Min.   :  1.00   Length:27312       Mode :logical
##  Class :character   1st Qu.: 44.00   Class :character   FALSE:22046
##  Mode  :character   Median : 68.00   Mode  :character   TRUE :5266
##                     Mean   : 65.64
##                     3rd Qu.: 81.00
##                     Max.   :123.00
##  PERP_AGE_GROUP       PERP_SEX           PERP_RACE         VIC_AGE_GROUP
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    VIC_SEX            VIC_RACE
##  Length:27312       Length:27312
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```
shooting$PERP_AGE_GROUP <- shooting$PERP_AGE_GROUP %>% replace_na("UNKNOWN")
shooting$PERP_SEX <- shooting$PERP_SEX %>% replace_na("U")
shooting$PERP_RACE <- shooting$PERP_RACE %>% replace_na("UNKNOWN")
```

During the exploration, I encountered some missing values in the dataset. Missing data can distort the analysis and lead to inaccurate conclusions. To handle this, I replaced missing values with "UNKNOWN" or "U". This approach maintains the integrity of the analysis while acknowledging the incomplete information.

```
shooting_by_month_year <- shooting %>%
  group_by(PRECINCT, BORO, OCCUR_DATE) %>%
  summarize(INCIDENTS = n(), MONTH_YEAR = as.yearmon(paste(month(OCCUR_DATE, label = TRUE), year(OCCUR_
  select(MONTH_YEAR, PRECINCT, BORO, OCCUR_DATE, INCIDENTS) %>%
  ungroup()

shooting_by_precinct <- shooting_by_month_year %>%
  group_by(PRECINCT, BORO, MONTH_YEAR) %>%
  summarize(INCIDENTS = n()) %>%
  select(MONTH_YEAR, PRECINCT, BORO, INCIDENTS) %>%
  ungroup()

shooting_by_boro <- shooting_by_precinct %>%
  group_by(BORO, MONTH_YEAR) %>%
  summarize(INCIDENTS = sum(INCIDENTS)) %>%
  select(BORO, MONTH_YEAR, INCIDENTS) %>%
  ungroup()
```

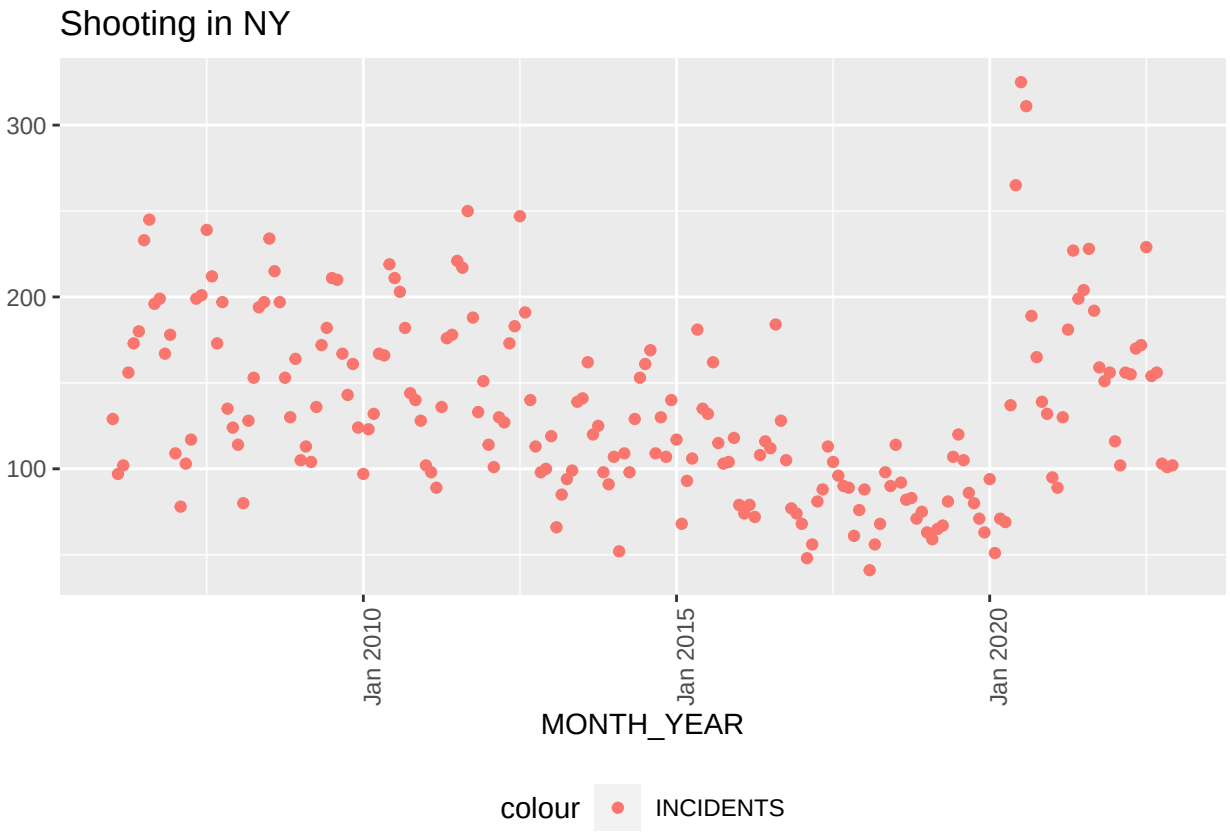
In this phase, I undertook data processing by grouping the information based on Month-Year, Precinct, Borough, and Year. This grouping will be beneficial for the later analysis.

```
shooting_year_total <- shooting_by_boro %>%
  group_by(MONTH_YEAR) %>%
  summarize(INCIDENTS = sum(INCIDENTS)) %>%
  select(MONTH_YEAR, INCIDENTS) %>%
  ungroup()

summary(shooting_year_total)
```

```
##    MONTH_YEAR     INCIDENTS
##  Min.   :2006   Min.   : 41.00
##  1st Qu.:2010   1st Qu.: 96.75
##  Median :2014   Median :124.50
##  Mean   :2014   Mean   :133.88
##  3rd Qu.:2019   3rd Qu.:169.25
##  Max.   :2023   Max.   :325.00
```

```
shooting_year_total %>%
  filter(INCIDENTS > 0) %>%
  ggplot(aes(x = MONTH_YEAR, y = INCIDENTS)) +
  geom_point(aes(color = "INCIDENTS")) +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Shooting in NY", y = NULL)
```

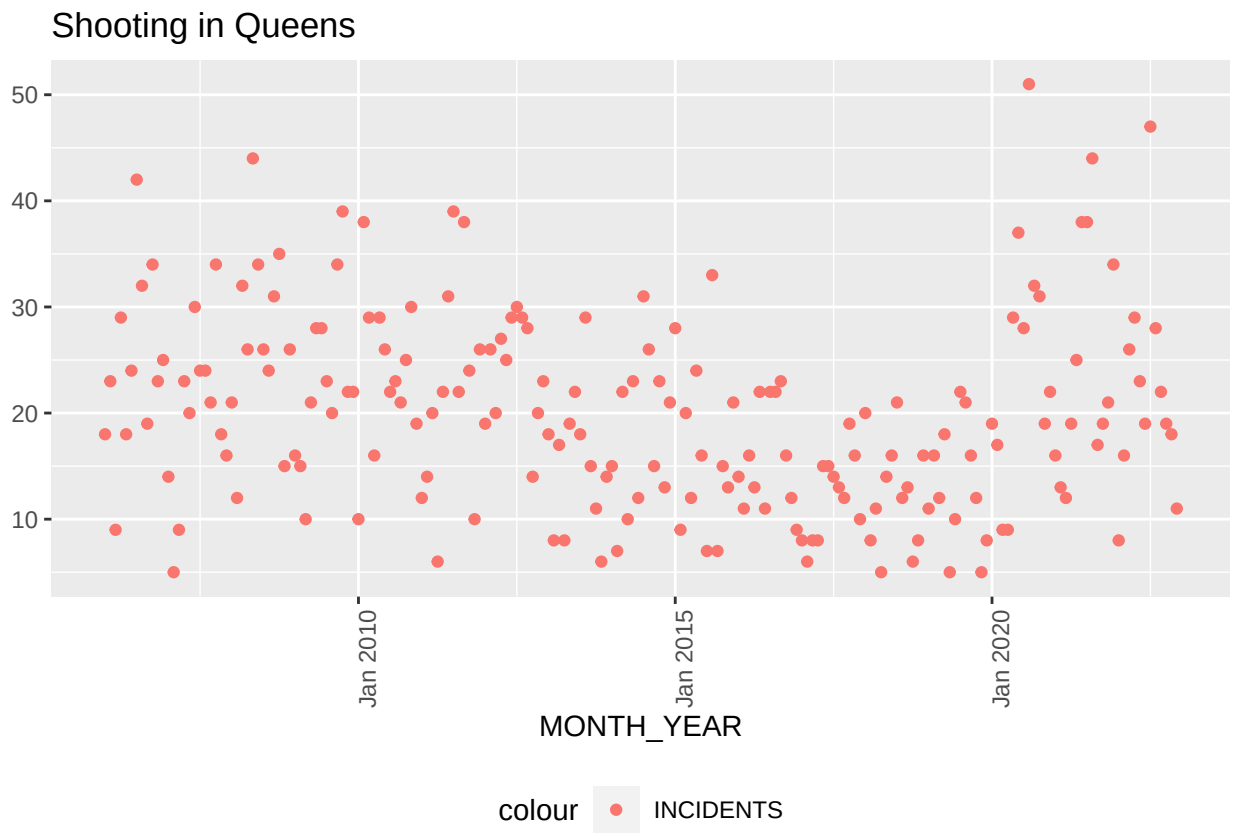


This plot illustrates shooting incidents across the entirety of New York. It demonstrates an initial gradual decline trend until around January 2020, followed by a sharp and significant upward trend thereafter.

```
shooting_queens_year_total <- shooting_by_boro %>%
  filter(BORO == "QUEENS") %>%
  group_by(MONTH_YEAR) %>%
  summarize(INCIDENTS = sum(INCIDENTS)) %>%
  select(MONTH_YEAR, INCIDENTS) %>%
  ungroup()
summary(shooting_queens_year_total)
```

```
##    MONTH_YEAR     INCIDENTS
##  Min.   :2006   Min.   : 5.00
##  1st Qu.:2010   1st Qu.:13.00
##  Median :2014   Median :19.00
##  Mean   :2014   Mean   :20.07
##  3rd Qu.:2019   3rd Qu.:26.00
##  Max.   :2023   Max.   :51.00
```

```
shooting_queens_year_total %>%
  filter(INCIDENTS > 0) %>%
  ggplot(aes(x = MONTH_YEAR, y = INCIDENTS)) +
  geom_point(aes(color = "INCIDENTS")) +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Shooting in Queens", y = NULL)
```

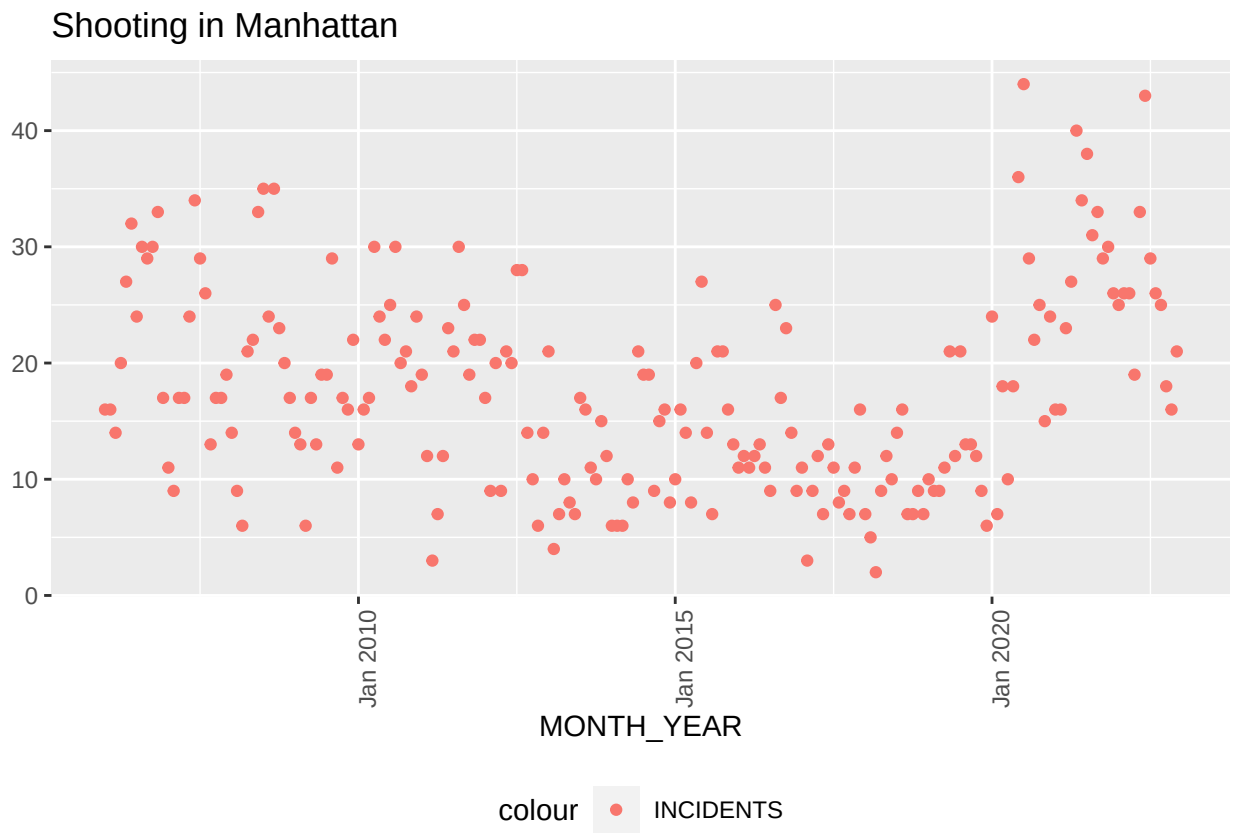


This plot contains shooting incidents specifically within the Queens borough. It shows a comparable trend to that of the entire New York, with a decrease preceding January 2020, followed by a subsequent increase.

```
shooting_manhattan_year_total <- shooting_by_boro %>%
  filter(BORO == "MANHATTAN") %>%
  group_by(MONTH_YEAR) %>%
  summarize(INCIDENTS = sum(INCIDENTS)) %>%
  select(MONTH_YEAR, INCIDENTS) %>%
  ungroup()
summary(shooting_manhattan_year_total)
```

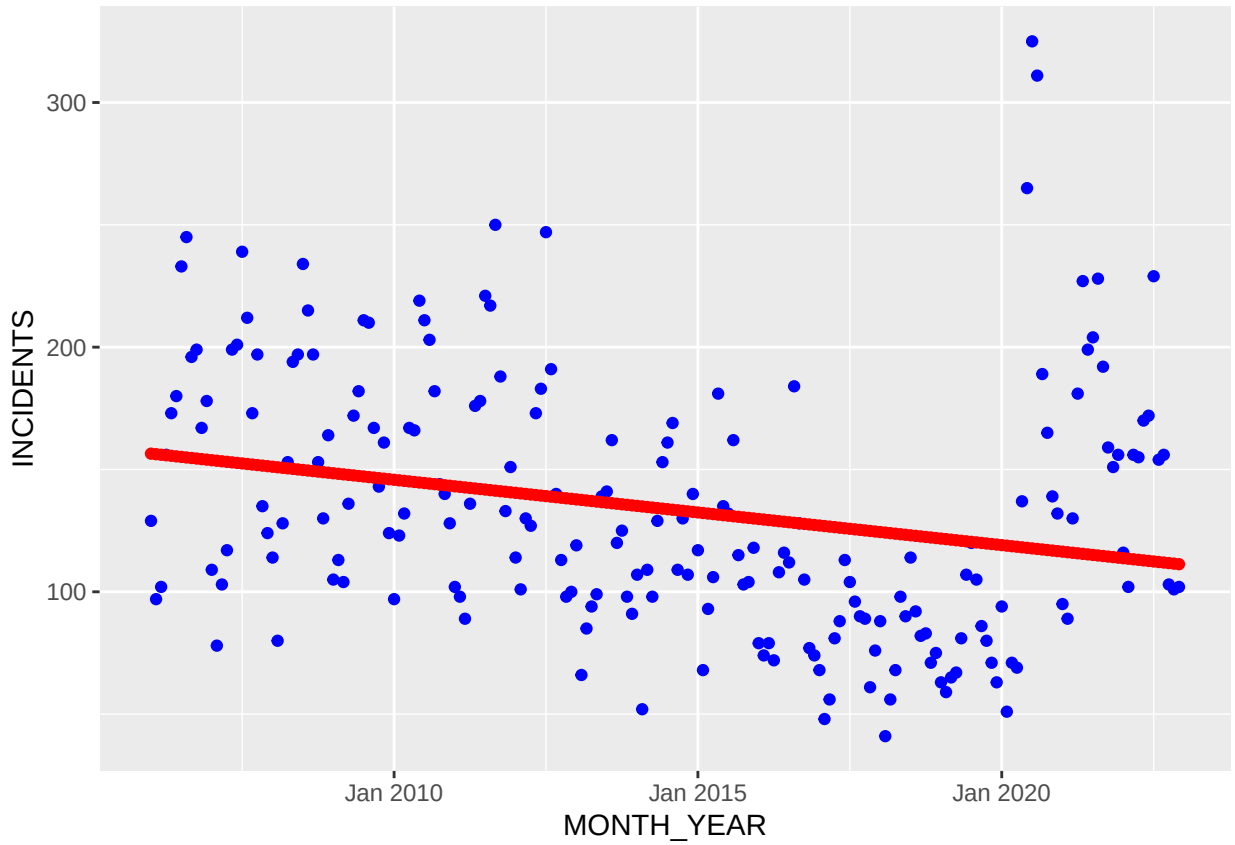
```
##    MONTH_YEAR     INCIDENTS
##  Min.   :2006   Min.   : 2.00
##  1st Qu.:2010   1st Qu.:11.00
##  Median :2014   Median :16.00
##  Mean   :2014   Mean   :17.51
##  3rd Qu.:2019   3rd Qu.:23.00
##  Max.   :2023   Max.   :44.00
```

```
shooting_manhattan_year_total %>%
  filter(INCIDENTS > 0) %>%
  ggplot(aes(x = MONTH_YEAR, y = INCIDENTS)) +
  geom_point(aes(color = "INCIDENTS")) +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Shooting in Manhattan", y = NULL)
```



This plot shows shooting incidents within the Manhattan borough. Its trend closely mirrors the one of the Queens borough. However, the accompanying summary reveals a difference in incident count, indicating there are fewer occurrences in comparison to Queens.

```
mod <- lm(INCIDENTS ~ MONTH_YEAR, data = shooting_year_total)
x_grid <- seq(as.yearmon("JAN 2006"), as.yearmon("DEC 2023"))
new_df <- tibble(MONTH_YEAR = x_grid)
shooting_year_total_pred <- shooting_year_total %>% mutate(pred = predict(mod))
shooting_year_total_pred %>% ggplot() +
  geom_point(aes(x = MONTH_YEAR, y = INCIDENTS), color = "blue") +
  geom_point(aes(x = MONTH_YEAR, y = pred), color = "red")
```



When analyzing the data as a whole or by borough, a noticeable spike emerges around January 2020. However, predictions from a linear model indicate a declining trend due to the drops before this point. The presence of the pandemic might have influenced shooting incident numbers, introducing potential bias to the data.