NYPD Shooting: Time Analysis

2024-02-05

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## 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)
library(dplyr)
```

NYPD Shooting Incident Data

I will use the data collected by the New York Police Department. This data contains the shooting incidents that arose in the boroughs in New York City. Here I import the NYPD data:

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

Here I read the data via read.csv

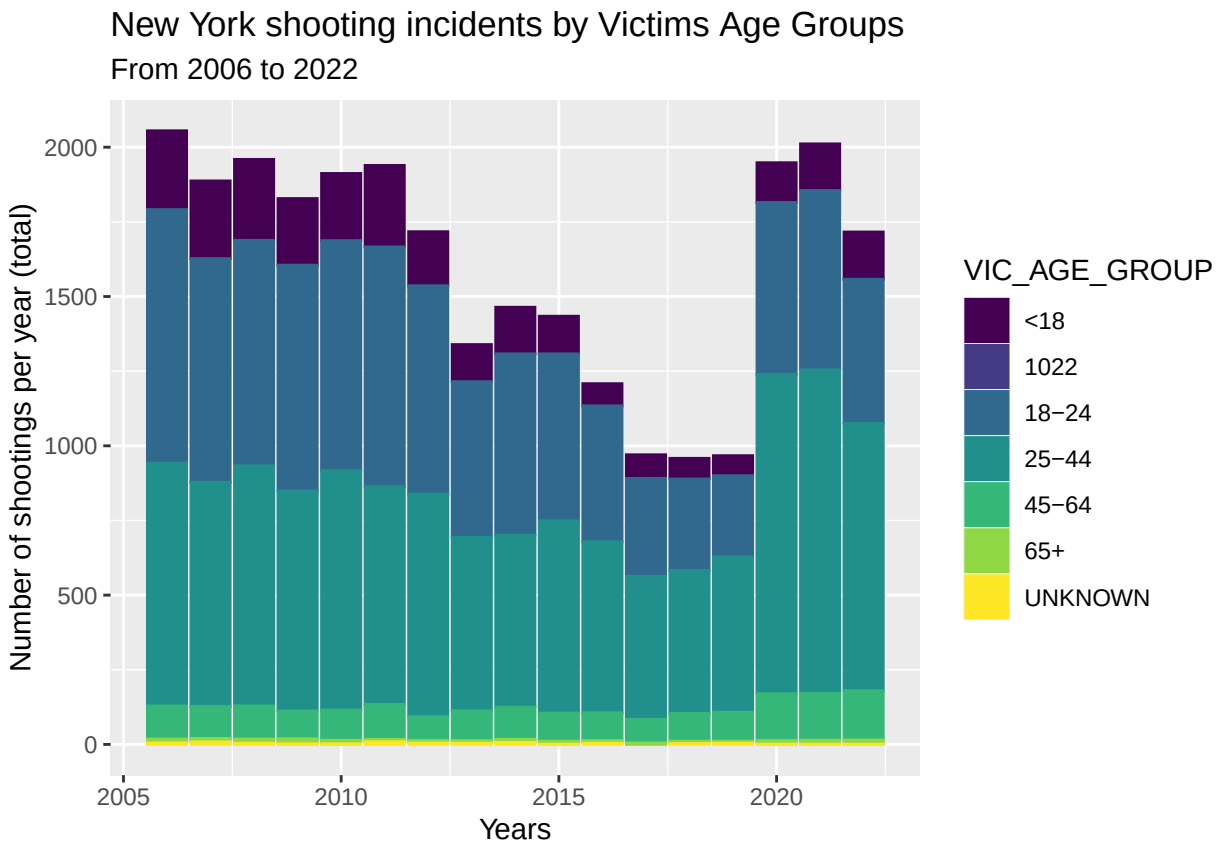
```
data_NYPD <- read.csv(url_import_NYPD)
```

Here I do the tidying of the data and select the columns that I need for the project.

```
tid_data <- data_NYPD %>%
  select(c("OCCUR_DATE", "OCCUR_TIME", "BORO", "STATISTICAL_MURDER_FLAG", "VIC_AGE_GROUP", "VIC_SEX", "
  mutate(OCCUR_DATE = mdy(OCCUR_DATE),
         Time = hms(OCCUR_TIME),
         Year = year(OCCUR_DATE),
         STATISTICAL_MURDER_FLAG = as.logical(STATISTICAL_MURDER_FLAG),
         Shootings = 1,
         Month = month(OCCUR_DATE))
```

I plot here the first view of the data

```
tid_data %>%
  ggplot(aes(x = Year, color = VIC_AGE_GROUP, fill = VIC_AGE_GROUP)) +
  geom_bar() +
  scale_color_viridis_d() +
  scale_fill_viridis_d() +
  labs(title = "New York shooting incidents by Victims Age Groups",
       subtitle = "From 2006 to 2022",
       x = "Years",
       y = "Number of shootings per year (total)")
```

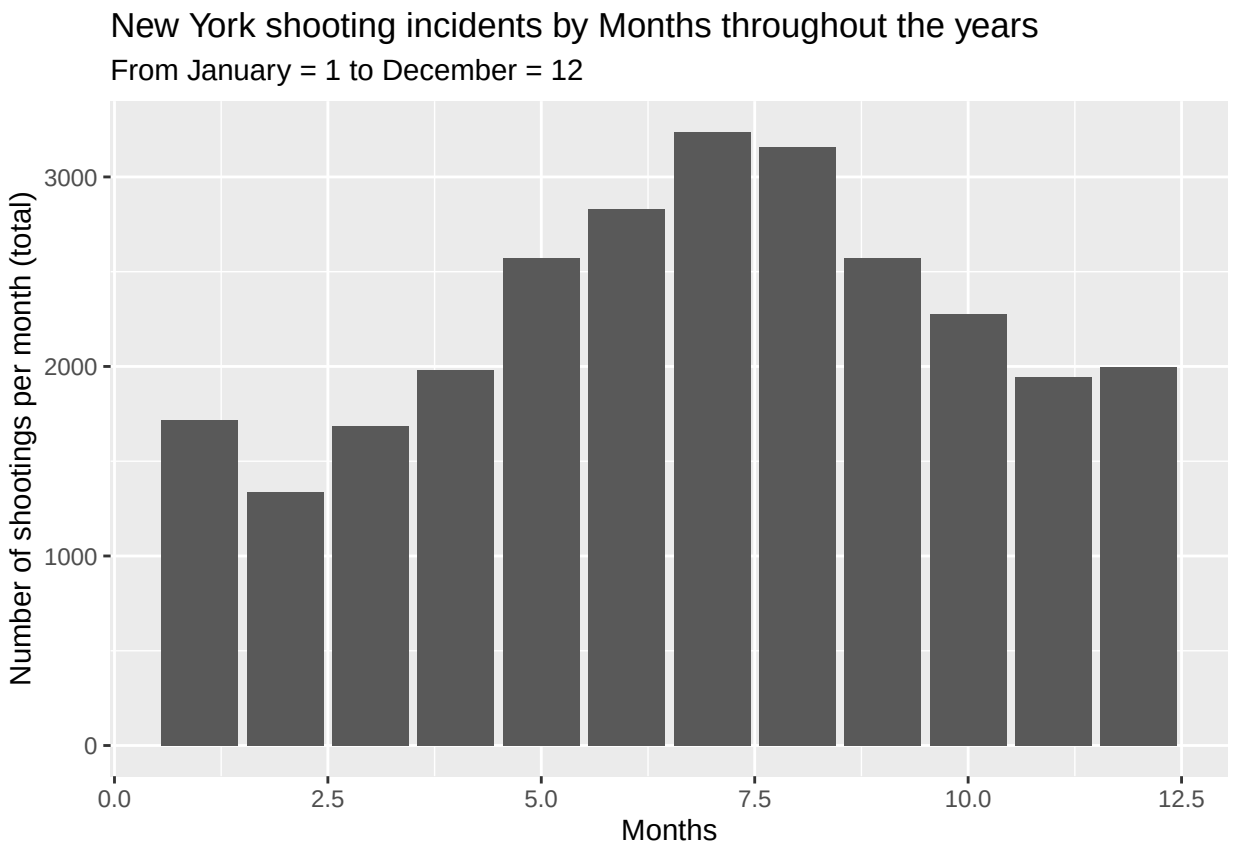


I plot the second view of the data.

```
tid_data %>%
  ggplot(aes(x = Month, color = Year, fill = Year)) +
  geom_bar() +
  scale_color_viridis_d() +
  scale_fill_viridis_d() +
  labs(title = "New York shooting incidents by Months throughout the years",
       subtitle = "From January = 1 to December = 12",
       x = "Months",
       y = "Number of shootings per month (total)")
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour, fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
```

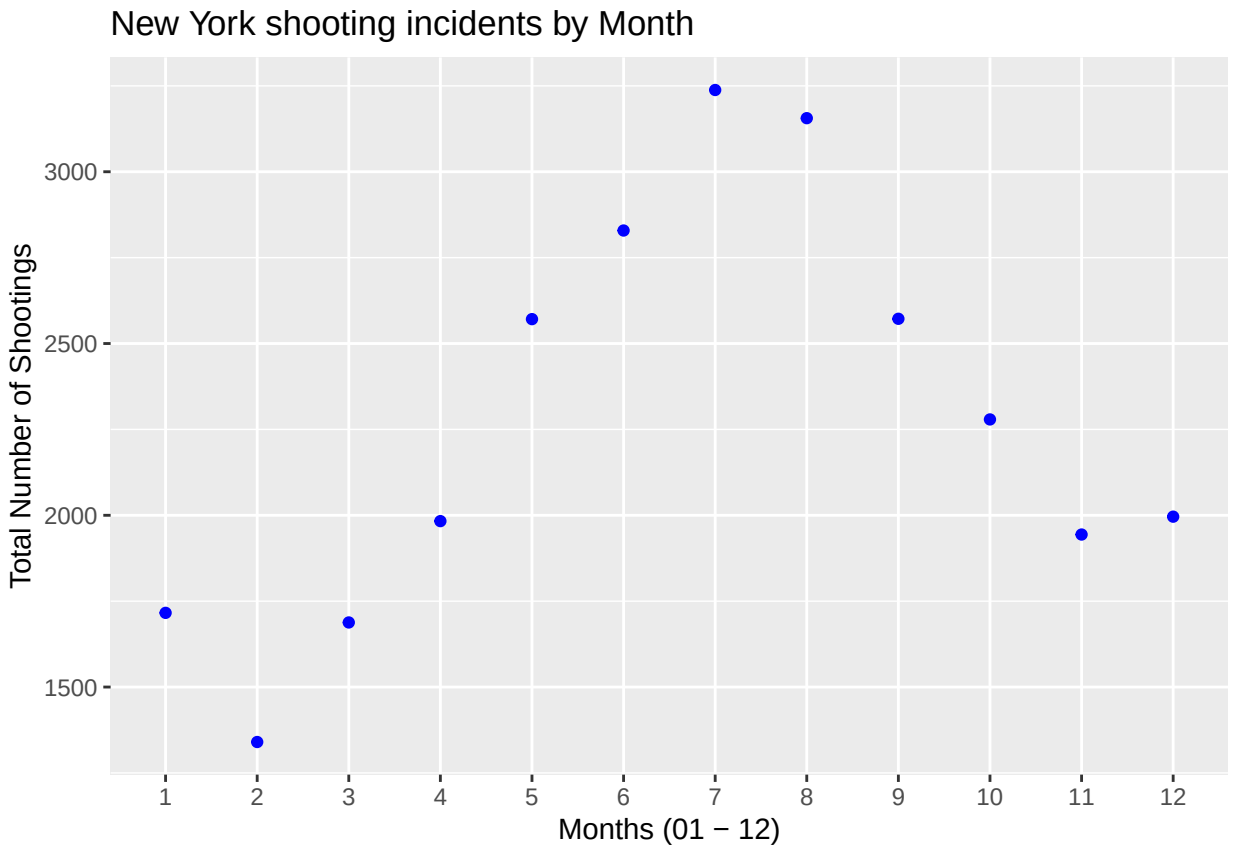
```
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



And now I plot the tendencies of the shooting incidents by month.

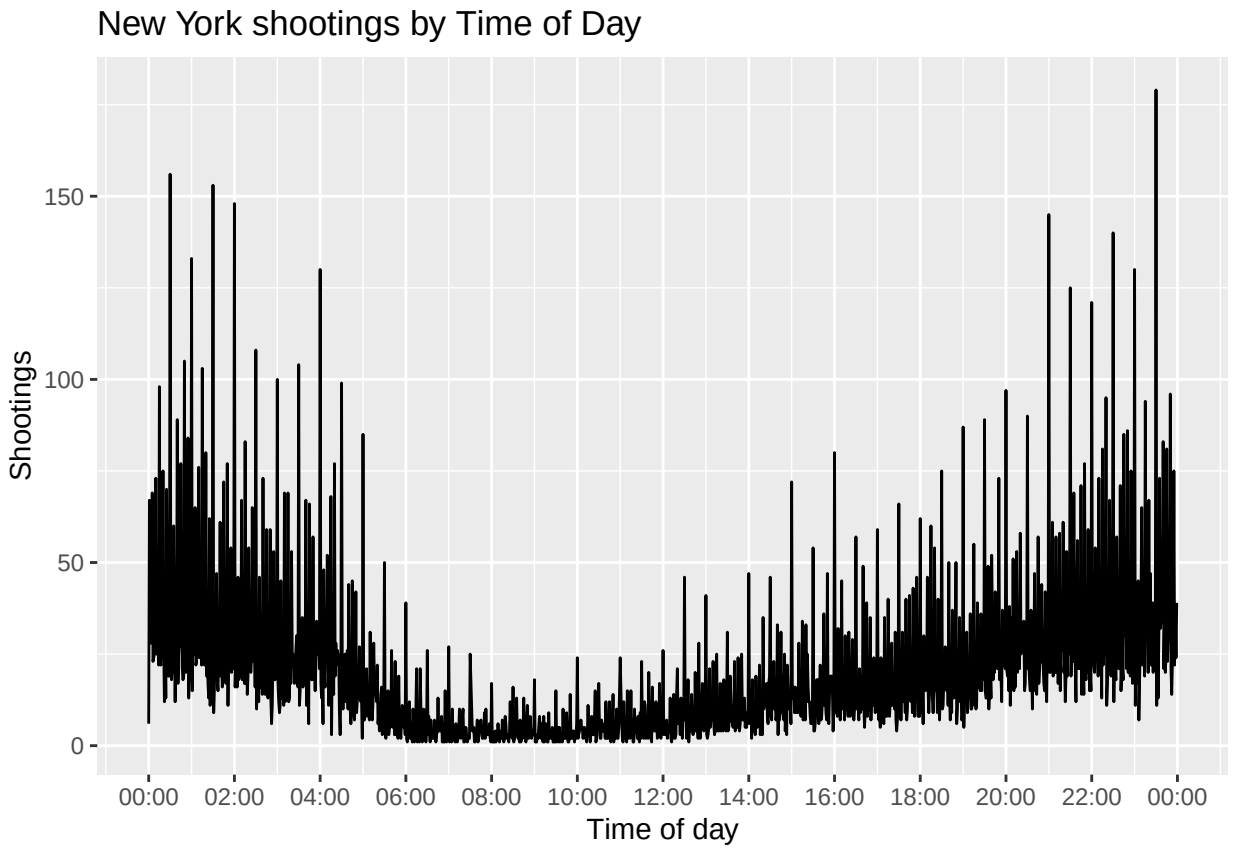
```
tid_month <- tid_data %>%
group_by(Month) %>%
summarize(Shootings = sum(Shootings))

tid_month %>%
ggplot(aes(x = as.factor(Month), y = Shootings)) +
geom_line() +
geom_point(color = "blue") +
scale_x_discrete(labels = as.character(01:12)) +
labs(
title = "New York shooting incidents by Month",
x = "Months (01 - 12)",
y = "Total Number of Shootings")
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```



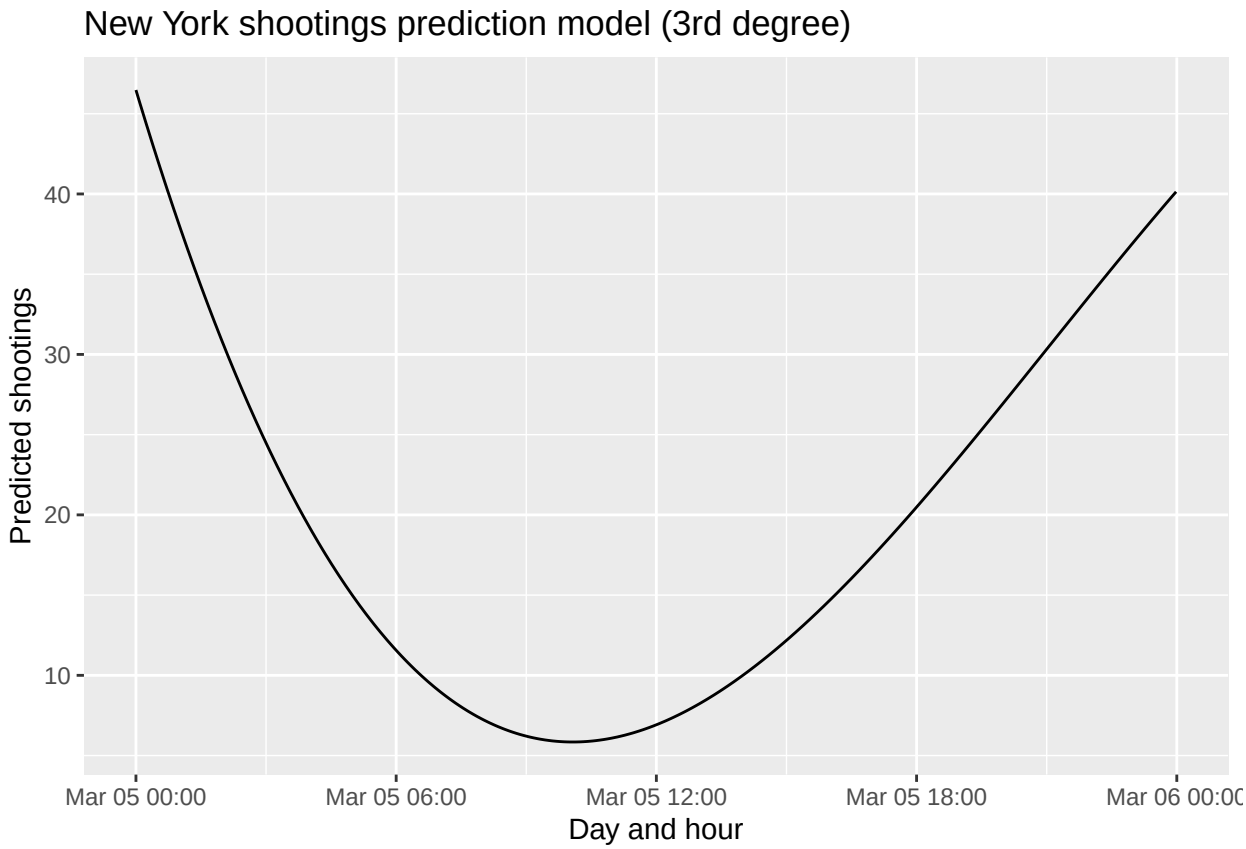
I plot the time-day shootings.

```
tid_time_day <- tid_data %>%
  group_by(OCCUR_TIME,Shootings) %>%
  summarize(Shootings = sum(Shootings),
            STATISTICAL_MURDER_FLAG = sum(STATISTICAL_MURDER_FLAG),
            .groups = 'drop') %>%
  select(OCCUR_TIME,Shootings,STATISTICAL_MURDER_FLAG)

tid_time_day$OCCUR_TIME <- as.POSIXct(tid_time_day$OCCUR_TIME, format = "%H:%M:%S")

tid_time_day %>%
  ggplot(aes(x = OCCUR_TIME, y = Shootings)) +
  geom_line() +
  scale_x_datetime(date_labels = "%H:%M", date_breaks = "2 hours") +
  labs(title = "New York shootings by Time of Day",
       x = "Time of day",
       y = "Shootings")
```



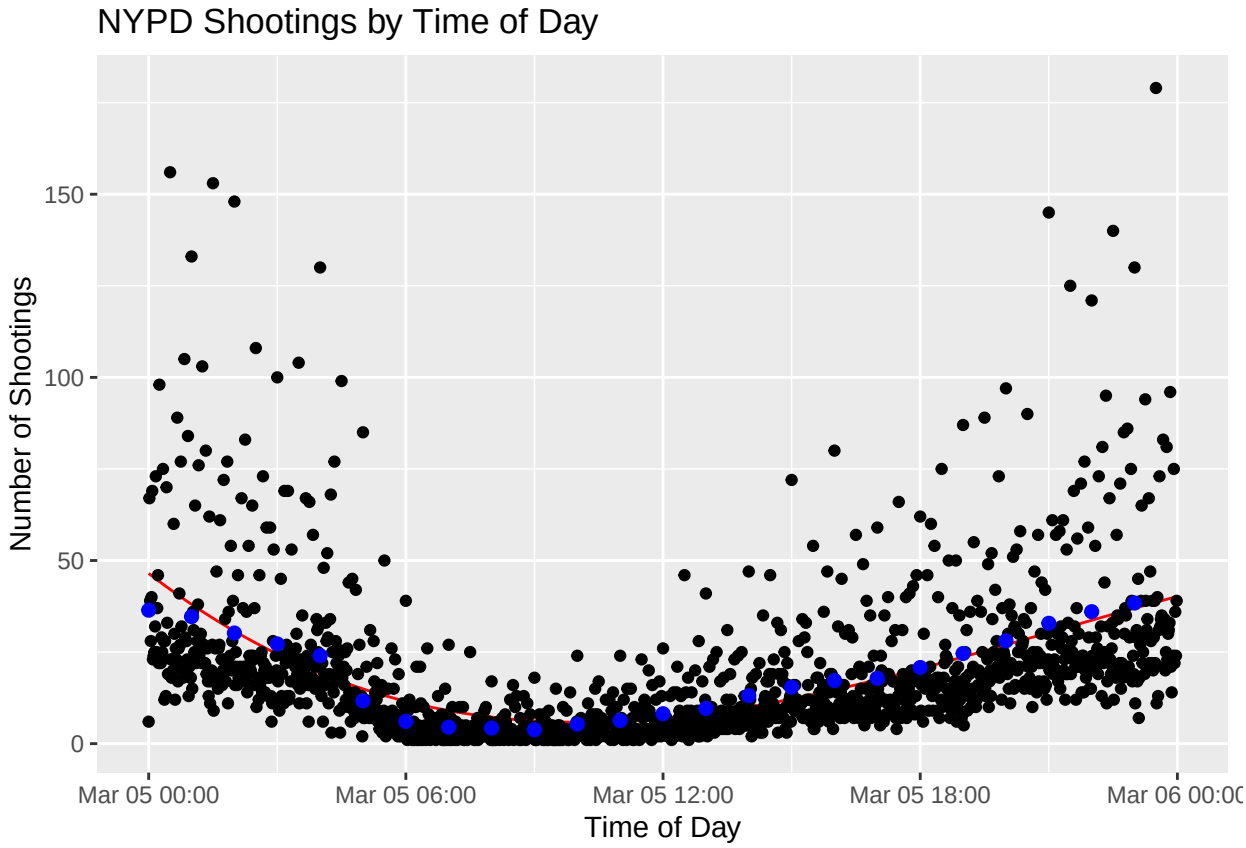
The interesting part is that there's a tendency, so I create here a model based on the data provided by the NYPD. (I thought it fitted 3rd degree)

```
shootings_data <- tid_time_day
shootings_model <- lm(Shootings ~ poly(OCCUR_TIME, 3), data = shootings_data)
predicted_shootings <- shootings_data %>%
  mutate(
    predicted = predict(shootings_model)
  )
ggplot(predicted_shootings, aes(x = OCCUR_TIME, y = predicted)) +
  geom_line() +
  labs(title = "New York shootings prediction model (3rd degree)",
       x = "Day and hour",
       y = "Predicted shootings")
```



Finally I create a data model with the prediction model I created before.

```
# Fit polynomial model
shootings_model <- lm(Shootings ~ poly(OCCUR_TIME, 3), data = shootings_data)

# Generate predictions
predicted_shootings <- shootings_data %>%
  mutate(
    predicted = predict(shootings_model)
  )

# Calculate historical hourly averages
hourly_averages <- shootings_data %>%
  mutate(hour = format(OCCUR_TIME, "%H")) %>%
  group_by(hour) %>%
  summarize(avg = mean(Shootings))

# Plot data
ggplot() +
  geom_line(data = predicted_shootings, aes(x = OCCUR_TIME, y = predicted), color = "red") +
  geom_point(data = shootings_data, aes(x = OCCUR_TIME, y = Shootings)) +
  geom_point(data = hourly_averages, aes(x = as.POSIXct(hour, format = "%H"), y = avg), color = "blue",
  labs(title = "NYPD Shootings by Time of Day",
       x = "Time of Day",
       y = "Number of Shootings")
```



Conclusions

The following conclusions are based on my model and my analysis.

The first thing I must note is that the most affected victim age group is from 25-44. However there's information in the database which I couldn't correct (the victim age group of 1022), which is a bias. Now I tried to focus more on the dates rather than the victim age groups. Thus leading to a mathematical tendency of the month of July. Which led to my question: Did they occur because of the opening of gun sales in New York? It would be interesting to see the perpetrator's ages, and retrieve the policies on gun's sales.

Now, when analyzing the data in terms of time throughout the day, it's very impressive to see the mathematical tendency that between 19:00 and 05:00 occurred the most of the shootings. Thus leading to my question, is this a time-based problem? As far as I'm concerned, the tendencies of risk is bigger late in the night.

This analysis led me to other questions like, is it possible that it's riskier in being a young man rather than an old woman, or viceversa? Is it also an economics problem? Or rather an economical politics problem?