NYC Shooting Report

Data source

The dataset I analyzed contains every shooting incident that took place in NYC from 2006 to 2020. This data can be used by the public to explore the nature of shooting/criminal activity. Data are manually extracted and reviewed quarterly. Each record represents a shooting incident including information about the event, the location and time of occurrence. You can also find information related to suspect and victim demographics.

Preliminary steps

After loading all the libraries needed to carry on the analysis, the first step is to get the data of NYPD shooting incidents report from the provided url
https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

```
library(tidyverse)
library(magrittr)
library(lubridate)
library(dplyr)
```

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

```
shootings_data <- read_csv(url_in)
```

In order to tidy up the data, columns `Longitude`, `Latitude`, `Lon_Lat`, `X_Coord_CD`, `Y_Coord_CD`, `INCIDENT_KEY` have been deleted since they are not useful to the analysis I want to carry on.

```
shootings_data <- subset(shootings_data,
  select=-c(INCIDENT_KEY, Latitude, Longitude, Lon_Lat, X_COORD_CD, Y_COORD_CD))
```

Some format conversions took place as well. `OCCUR_DATE` has been converted from character to date, while `OCCUR_TIME` has been transformed from seconds-from-midnight to "hours:minutes:seconds" to improve readability.

```
shootings_data %<>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

```
shootings_data %<>% mutate(OCCUR_TIME = format(as.POSIXct('1900-1-1') +
  OCCUR_TIME, '%H:%M:%S'))
```

On the other hand, `JURISDICTION_CODE` and `PRECINCT_CODE` have been converted from number to character since they are categorical data.

```
shootings_data %<>% mutate(JURISDICTION_CODE = as.character(JURISDICTION_CODE))

shootings_data %<>% mutate(PRECINCT = as.character(PRECINCT))
```

```
summary(shootings_data)
```

```
##    OCCUR_DATE          OCCUR_TIME            BORO              PRECINCT
##  Min.   :2006-01-01   Length:23585       Length:23585       Length:23585
##  1st Qu.:2008-12-31   Class :character   Class :character   Class :character
##  Median :2012-02-27   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2012-10-05
##  3rd Qu.:2016-03-02
##  Max.   :2020-12-31
##  JURISDICTION_CODE  LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:23585       Length:23585       Mode :logical
##  Class :character   Class :character   FALSE:19085
##  Mode  :character   Mode  :character   TRUE :4500
##
##
##
##  PERP_AGE_GROUP       PERP_SEX           PERP_RACE         VIC_AGE_GROUP
##  Length:23585       Length:23585       Length:23585       Length:23585
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    VIC_SEX             VIC_RACE
##  Length:23585       Length:23585
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Analysis 1: Cumulative incidents over the whole time interval

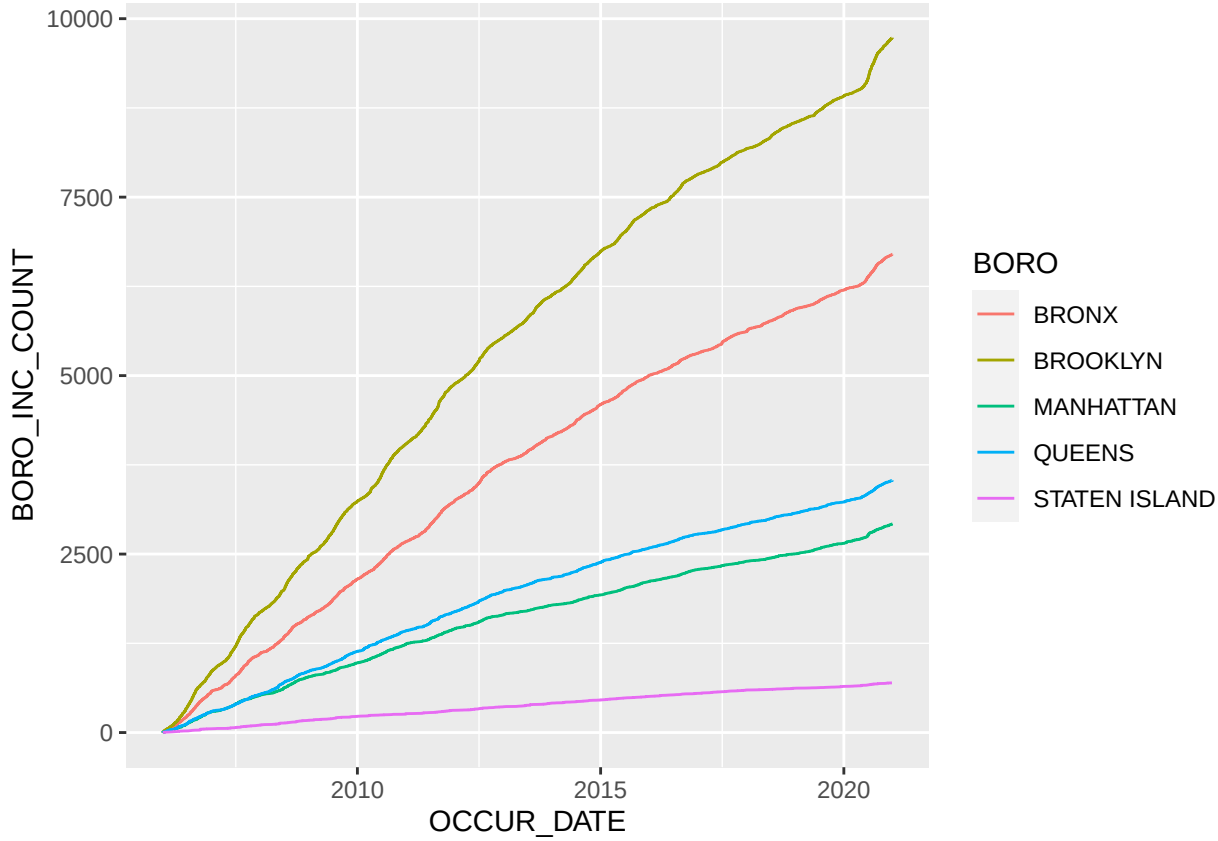
This analysis focuses on the trend of shooting incidents in NY boroughs over the whole timespan covered by the dataset. To do so, a new dataframe has been created ordering the original data by date and grouping them by boroughs.

```
shootings_by_boro <- shootings_data %>%
  mutate(across(OCCUR_DATE, ymd)) %>%
  arrange(OCCUR_DATE) %>%
  group_by(BORO)
```

In order to count the number of incidents occurred in each borough for each date, a new column called `BORO_INC_COUNT` has been set up.

```
shootings_by_boro %<>%
  group_by(BORO) %>%
  mutate(BORO_INC_COUNT=row_number())
```

```
shootings_by_boro %>%
  ggplot(mapping = aes(x = OCCUR_DATE, y = BORO_INC_COUNT, color = BORO)) +
  geom_line()
```



This graph shows the evolution of cumulative shooting incidents in each borough from year 2006 to 2020 included. The x-axis represents the timespan of analysis while the y-axis shows the number of shooting incidents. Each colored line describes the cumulative cases for a single borough.
It can be seen that the number of shooting incidents grew steadily throughout the years. The boroughs of Brooklyn and Bronx experienced the highest increment while Staten Island sees the slightest increase.
Finally, it is interesting to see how the trend increases in 2020 over all boroughs which could be a starting point for further analysis.

Analysis 2: Cumulative incidents in 24H by borough

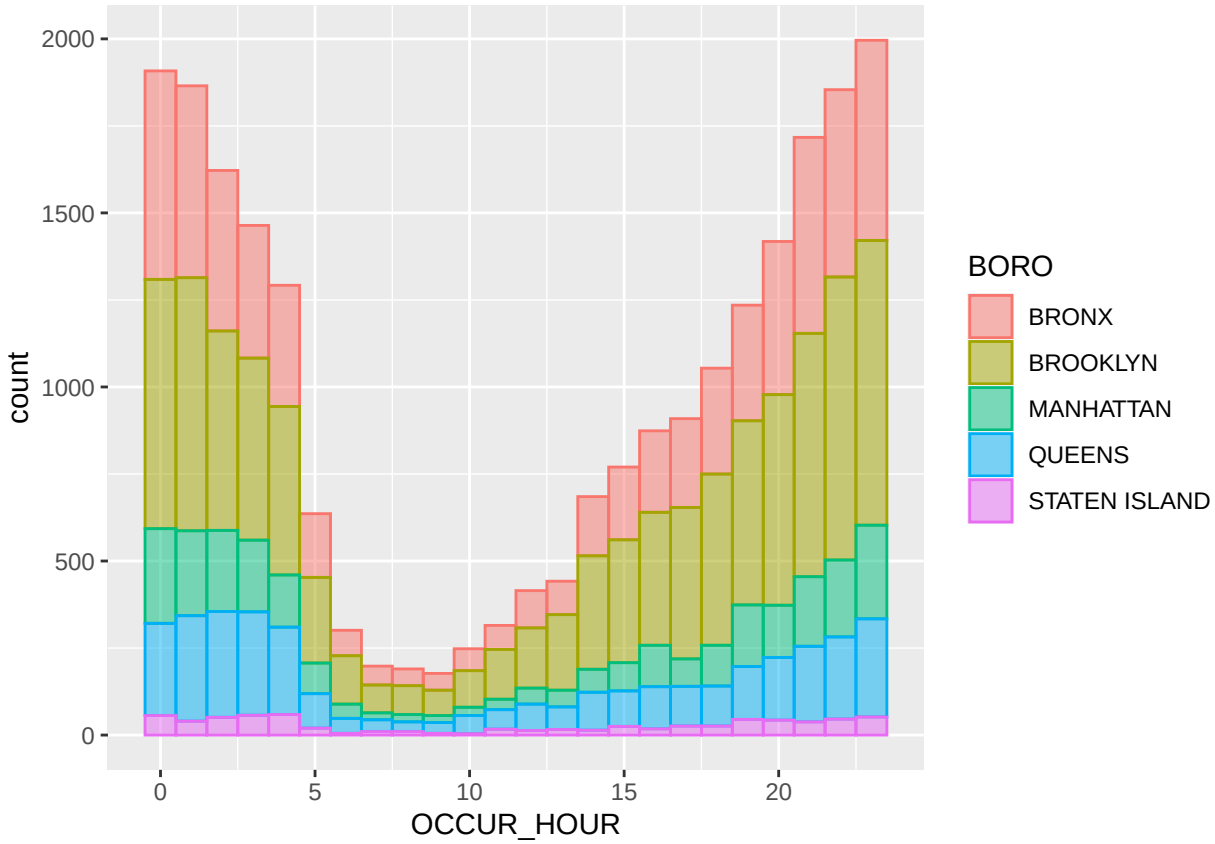
The purpose of this second analysis is to evaluate the cumulative incidents on a hourly basis within each borough, considering the full timespan covered by the dataset.
The same dataframe created in the previous analysis has been used, with the addition of two new columns. A first one called `OCCUR_HOUR` includes the hour part of the `OCCUR_TIME` variable, in order to have 24 hours in which each `OCCUR_TIME` can fall in.

```
shootings_by_boro %<>%
  mutate(OCCUR_HOUR=hour(hms(OCCUR_TIME)))

shootings_by_boro %<>%
  group_by(OCCUR_HOUR) %>%
  mutate(HOUR_INC_COUNT = row_number())
```

After grouping the data by `OCCUR_HOUR`, a second variable named `HOUR_INC_COUNT` act as a counter keeping track of the amount of shooting incidents for each hour of the day.

```
shootings_by_boro %>%
  ggplot(aes(x=OCCUR_HOUR)) +
  geom_histogram(aes(color=BORO, fill= BORO), bins = 24 , alpha = 0.5)
```



This histogram illustrates the amount of shooting incidents on a 24-hours basis. The total is composed by the amount of shooting incidents of each borough for that hour. On the x-axis the 24 hours of a day are represented, while on the y-axis are the counts of shooting incidents for each hour. Each colored bar describes a borough.
It can be seen that the amount of shooting incidents grows steadily after 10 am, reaching its peak at midnight. Conversely, from 6 am to 10 am the incidence appears to be the lowest.
This graph confirms the conclusions of the previous analysis, pointing out the boroughs with the higher shooting incidents rate.

Linear model: Murders in function of shooting incidents

In order to test a linear predictive model, I tried to correlate the case of murder with the number of shooting incidents on a daily basis, referring to the `STATISTICAL_MURDER_FLAG` variable of the original dataset.
To do so, a new dataframe has been created ordering data by `OCCUR_DATE`, along with two new variables.

```
aux_shooting <- shootings_data %>%
  mutate(across(OCCUR_DATE, ymd)) %>%
  arrange(OCCUR_DATE)
```

The first variable keeps track of the shooting incident per date, while the second one counts the number of murder cases per day.

```
aux_shooting %<>%
  group_by(OCCUR_DATE) %>%
  mutate(DATE_INC_COUNT=row_number())
```

```
aux_shooting %<>%
  group_by(OCCUR_DATE, STATISTICAL_MURDER_FLAG) %>%
  mutate(DATE_DEATH_COUNT=ifelse(STATISTICAL_MURDER_FLAG==TRUE, row_number(), 0))
```

Finally, the data have been aggregated to have one row for each day. Therefore, a new dataframe called `shootings_by_date` is created out of the previous one collapsing the same days on a single row.

```
shootings_by_date <- aux_shooting %>%
  group_by(OCCUR_DATE) %>%
  summarize(INC_PER_DATE = max(DATE_INC_COUNT), DEATHS_PER_DATE = max(DATE_DEATH_COUNT))
```

Then the linear model is set up on murders per day versus shooting incidents per day (i.e. using variables `DEATHS_PER_DATE` and `INC_PER_DATE`).

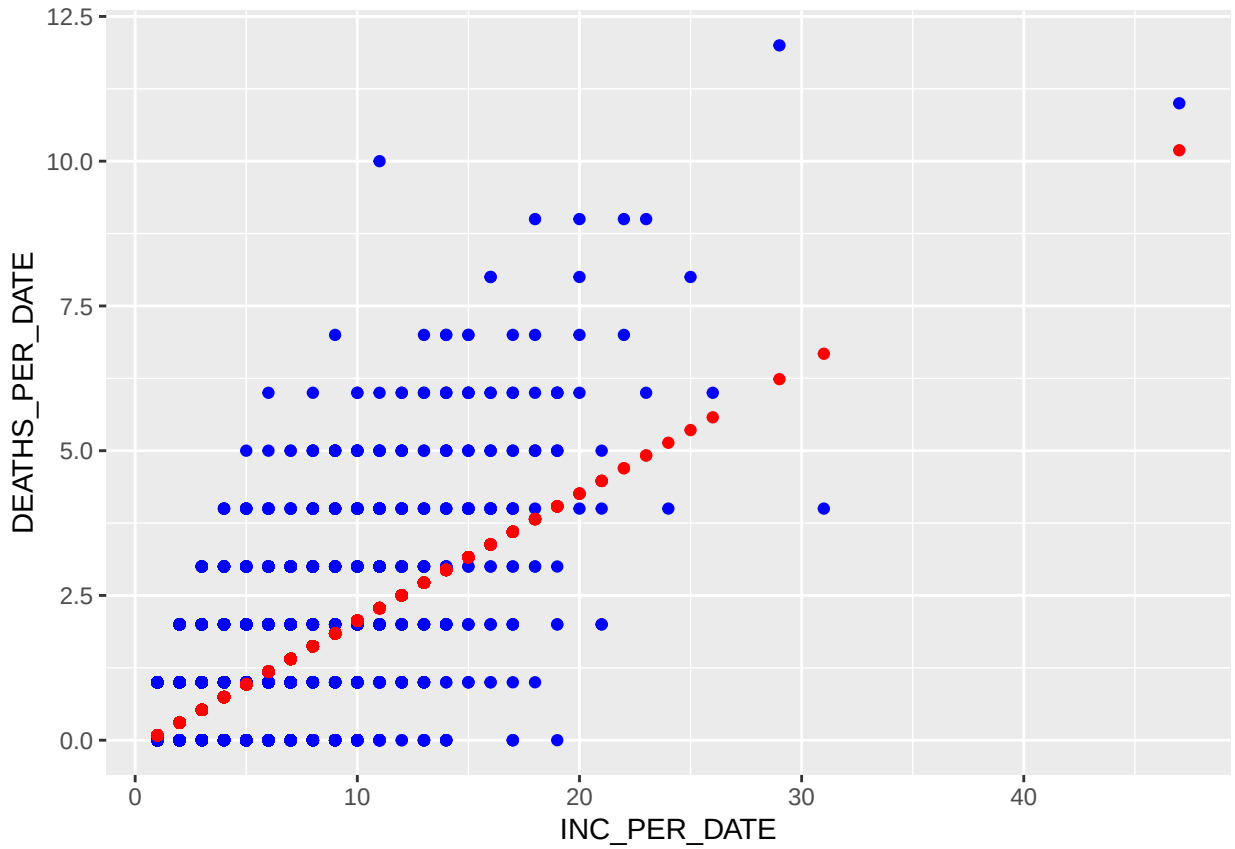
```
deathmodel <- lm(DEATHS_PER_DATE ~ INC_PER_DATE, data = shootings_by_date);
summary(deathmodel)
```

```
##
## Call:
## lm(formula = DEATHS_PER_DATE ~ INC_PER_DATE, data = shootings_by_date)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0387 -0.5243 -0.0850  0.4757  7.7185
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.134615   0.022655  -5.942    3e-09 ***
## INC_PER_DATE  0.219646   0.003859  56.924   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9774 on 5052 degrees of freedom
## Multiple R-squared:  0.3908, Adjusted R-squared:  0.3906
## F-statistic:  3240 on 1 and 5052 DF,  p-value: < 2.2e-16
```

A prediction column named `pred` has been created to include the predicted murder cases for each day.

```
shootings_by_date_w_pred <- shootings_by_date %>%
  mutate(pred=predict(deathmodel))
```

```
shootings_by_date_w_pred %>% ggplot() +
  geom_point(aes(x=INC_PER_DATE, y=DEATHS_PER_DATE), color = "blue") +
  geom_point(aes(x=INC_PER_DATE, y=pred), color = "red")
```



This plot shows how the model presents to some extent a linearity. The blue dots represent the actual behaviour of incidents and murder counts for each date, while the red dots picture the same data according to a linear predictive model. The presence of more murder cases for a single incident number (especially for low incident numbers), comes as no surprise, since there are of course more days with the same incident count, and each of them can have a different number of murders. It is worth to notice some outliers, like a day with 29 cases related to 11 murders.

Identification of bias

A possible source of bias may be related to social environment of each borough. One may think that shooting incidents are more frequent in areas with lower socio-economic status. To overcome this bias all boroughs have been equally and objectively considered throughout the analysis.
In order to avoid any possible bias related to age or race of both victims and perpetrators, this analysis excluded that kind of information.