NYPD Project

Laura McReady

2023-06-16

NYPD Shooting Data Project

Set up code chunk is used to load packages required.

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

This data includes every shooting incident in NYC from 2006 to the end of 2022. It includes information on both the suspect and victim as well as information about the event such as time and location.

Input Data

First, obtain data from the website.

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

Next, read the data and look at the summary.

```
NYPD_data <- read_csv(NYPD_data_url)
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(NYPD_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
```

```
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  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_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Tidy Data

Tidy the data by removing some columns and changing the format of other.

- Format of date column was changed
- Removed columns that were not going to be used
- Format of Precinct column changed from character to numeric

```
NYPD <- NYPD_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  select(-c(INCIDENT_KEY, LOC_OF_OCCUR_DESC, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, LOC_CLASSFCTN_DE
            Latitude, Longitude, Lon_Lat, LOCATION_DESC))
```

One area of interest was the time that shootings took place so new columns were added, one for the hour, one for the month, and one for the year of each event.

```
NYPD <- NYPD %>%
  mutate(OCCUR_HOUR = hour(OCCUR_TIME)) %>%
  mutate(OCCUR_MONTH = month(OCCUR_DATE)) %>%
  mutate(OCCUR_YEAR = year(OCCUR_DATE))
```

Created a data set to explore the differences between boroughs

```
NYPD_boro <- NYPD %>%
  group_by(BORO) %>%
  summarize(incidents = n()) %>%
  ungroup()
```

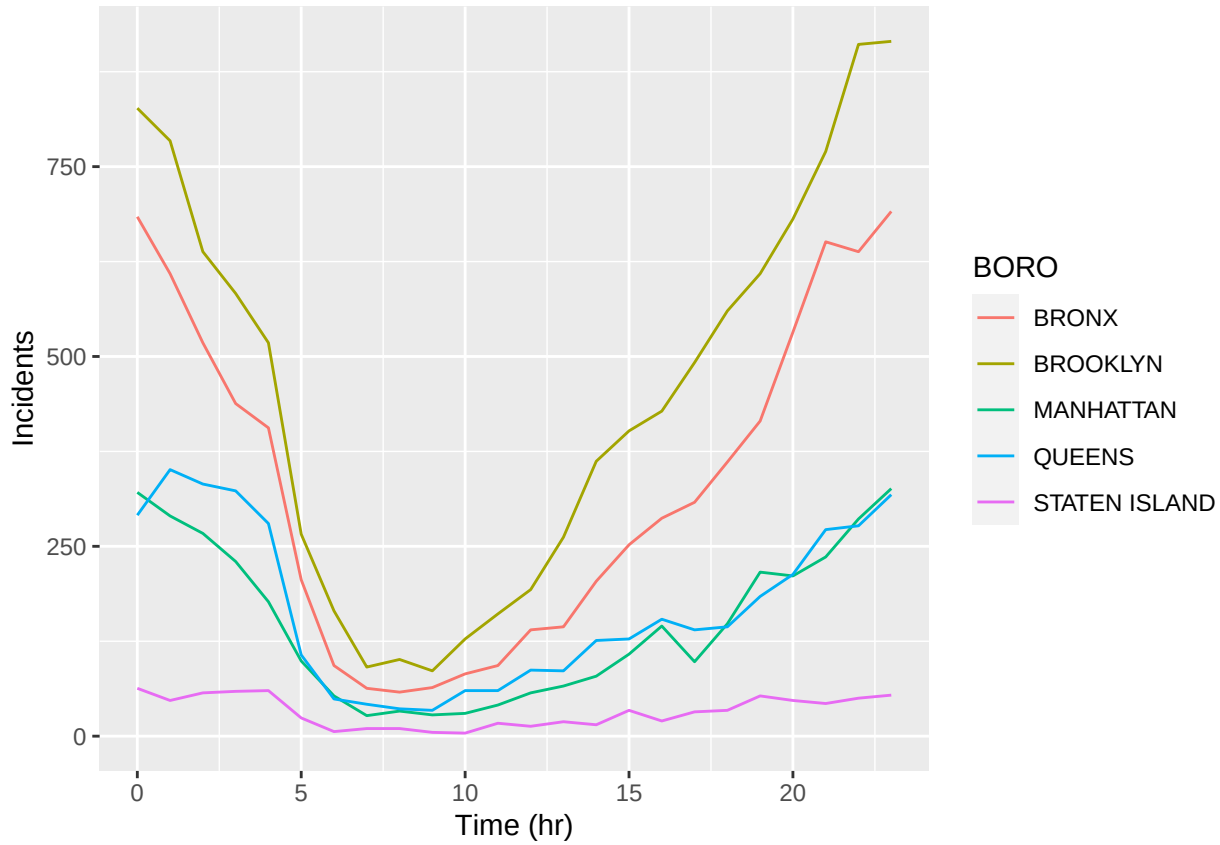
Created a data set to explore the differences by precinct

```
NYPD_precinct <- NYPD %>%
  group_by(PRECINCT) %>%
  summarize(incidents = n()) %>%
  ungroup()
```

Visualizations and Analysis

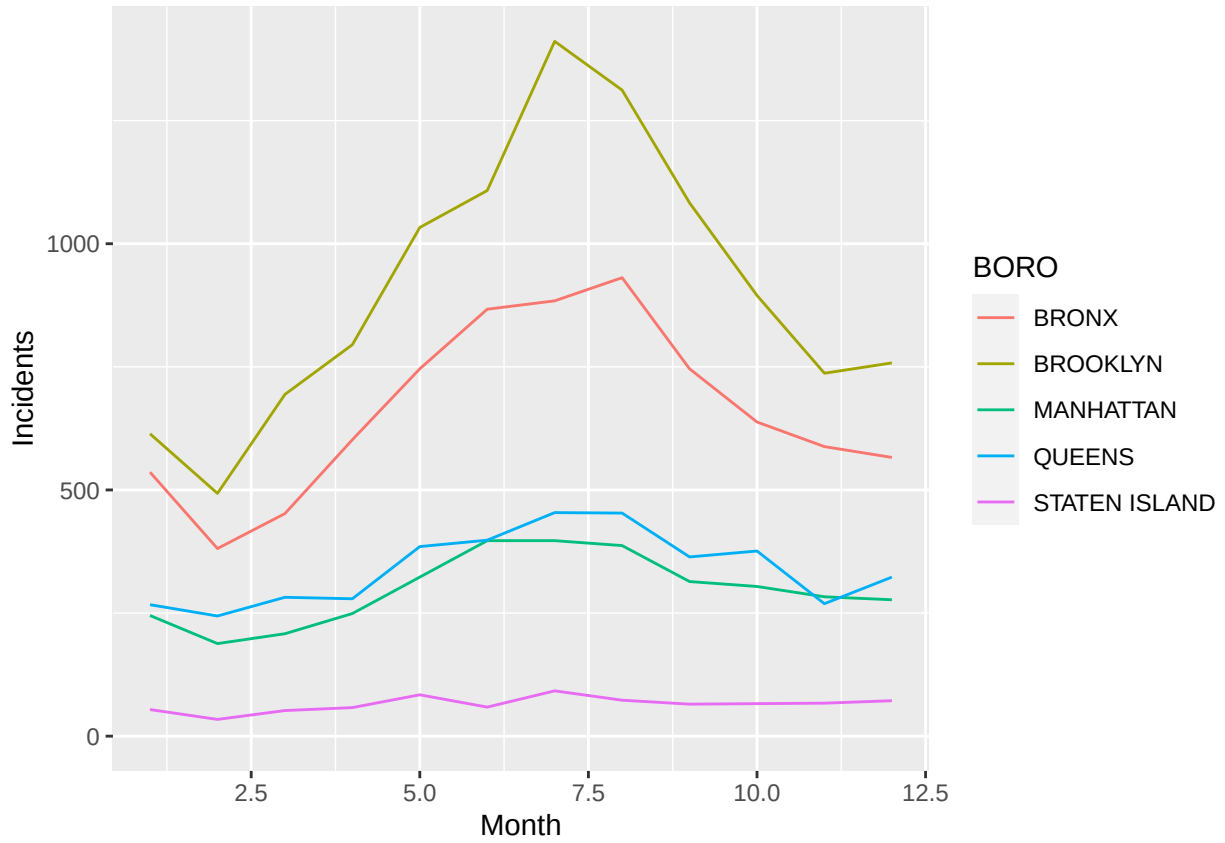
The first plot shows the number of shootings by time of day. Most take place in the overnight hours in all boroughs but Staten Island seemed to show less of a difference between time of day.

```
NYPD %>%
  ggplot(aes(x = OCCUR_HOUR, color = BORO)) +
  geom_freqpoly(binwidth = 1) +
  xlim(0, 23) +
  labs(x = "Time (hr)", y = "Incidents" )
```



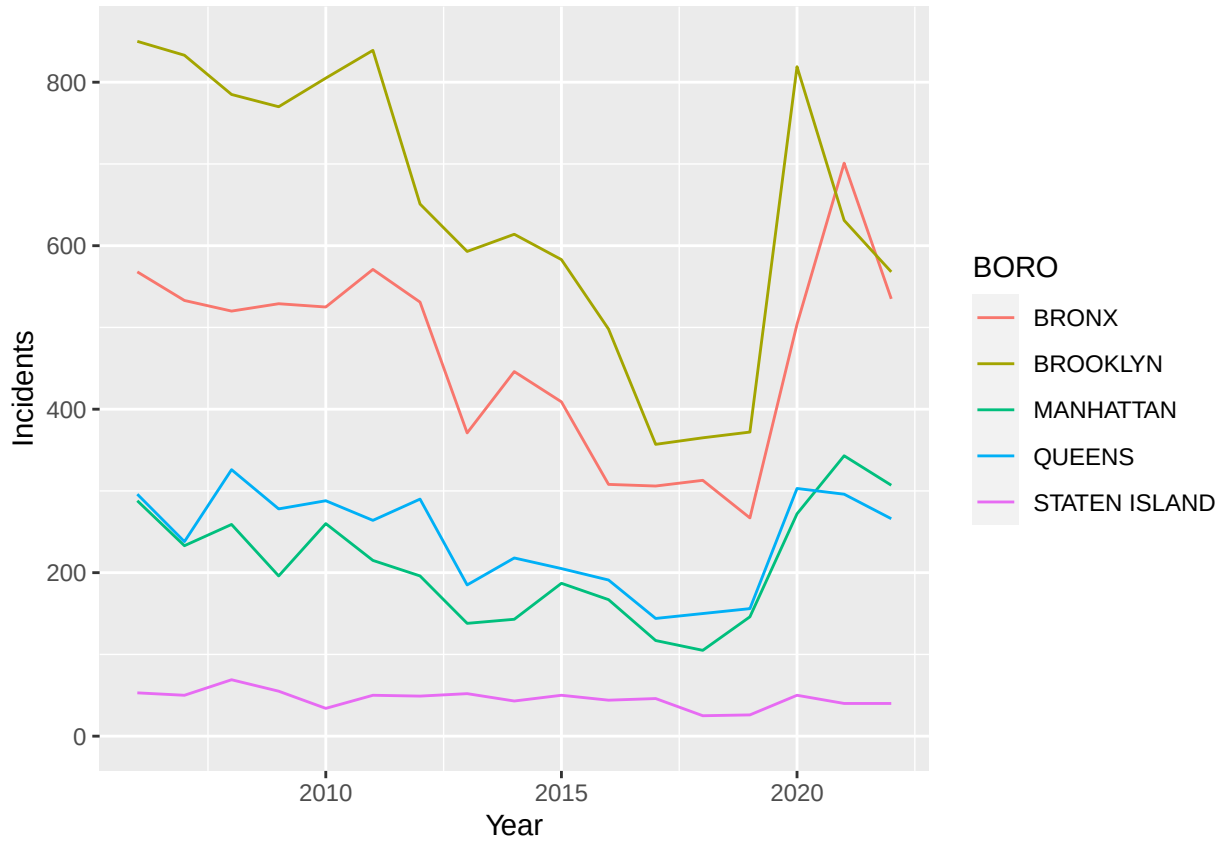
The next image shows the number of shootings throughout the year. The summer months had higher amounts, perhaps because people are spending more time outside and are more exposed to shootings.

```
NYPD %>%
  ggplot(aes(x = OCCUR_MONTH, color = BORO)) +
  geom_freqpoly(binwidth = 1) +
  xlim(1,12) +
  labs(x = "Month", y = "Incidents")
```



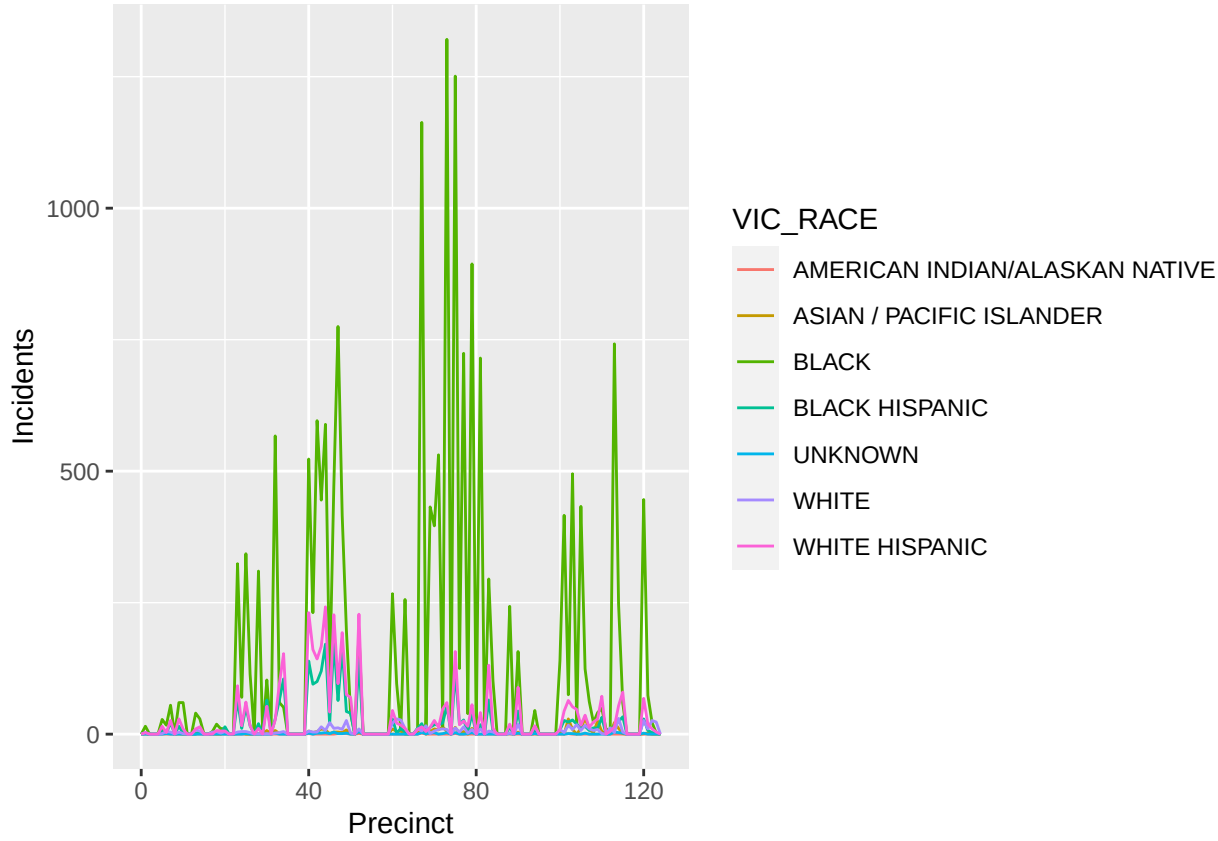
Next the number of incidents for the years 2006-2022 are shown. This shows that before 2020 the number of shootings had been declining. While shootings shot up for a couple years they seem to be on the decline again.

```
NYPD %>%
  ggplot(aes(x = OCCUR_YEAR, color = BORO)) +
  geom_freqpoly(binwidth = 1) +
  xlim(2006,2022) +
  labs(x = "Year", y = "Incidents")
```



The previous images, which were separated out by borough, all showed that Brooklyn had the highest number of shootins. This was unsurprising considering it has the largest population of the boroughs. I was interested to see whether specific neighborhoods had large amounts of shootings or if it was spread out evenly. The following image shows the number of shootings by precinct. While many precincts in Brooklyn have lots of shootings other areas

```
NYPD %>%
  ggplot(aes(x = PRECINCT, color = VIC_RACE)) +
  geom_freqpoly(binwidth = 1) +
  labs(x = "Precinct", y = "Incidents")
```



The following code chunk obtains the five precincts with the most incidents.

```
NYPD_precinct %>%
  slice_max(incidents, n = 5)
```

```
## # A tibble: 5 x 2
##   PRECINCT incidents
##      <dbl>     <int>
## 1       75      1557
## 2       73      1452
## 3       67      1216
## 4       44      1020
## 5       79      1012
```

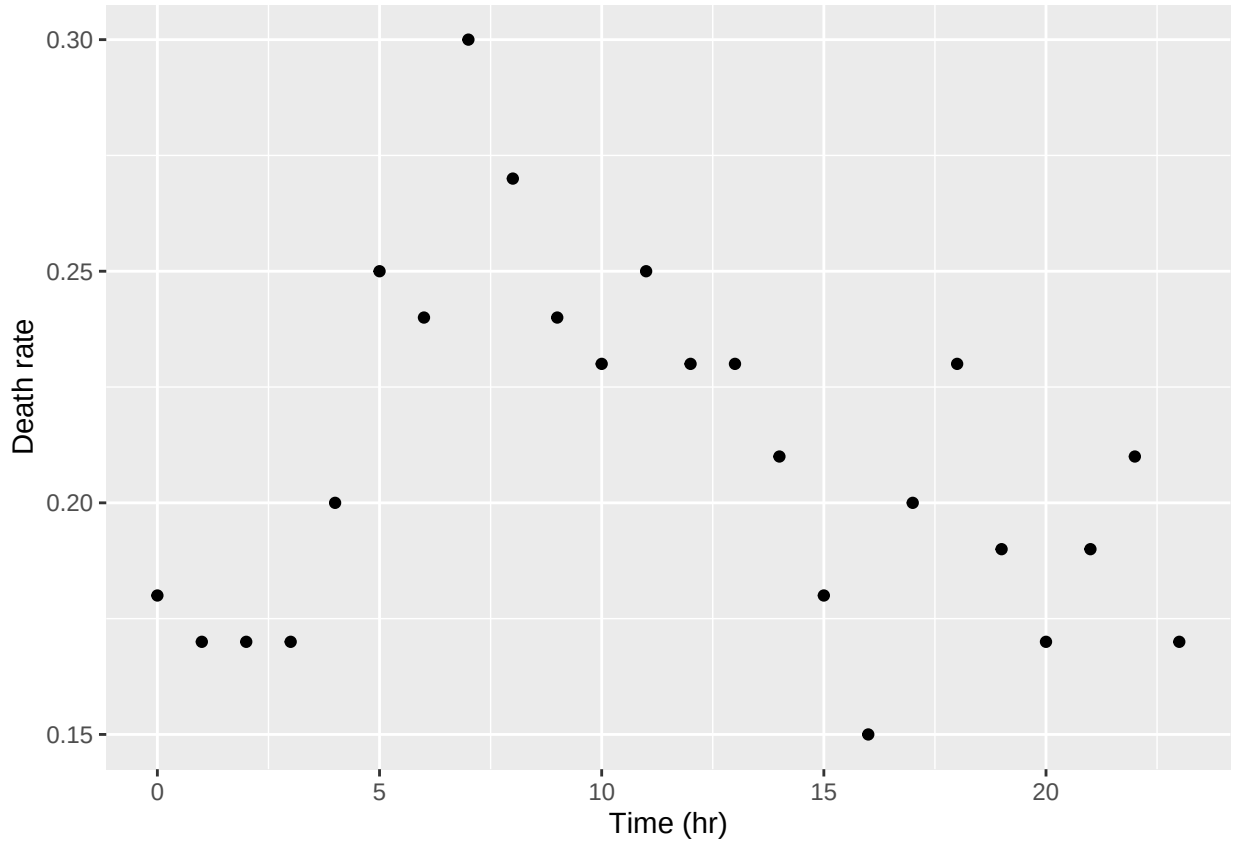
Model

Create a data frame that can be used to model the death rate by time. The data is grouped by the hour of each occurrence then the number of shootings and deaths are summed up. To determine the death rate the number of deaths was divided by the number of shootings.

```
NYPD_model <- NYPD %>%
  group_by(OCCUR_HOUR) %>%
  summarize(incidents = n(), deaths = sum(STATISTICAL_MURDER_FLAG)) %>%
  mutate(death_rate = round(deaths/incidents, digits = 2)) %>%
  ungroup()
```

The following model shows how the death rate varies with time.

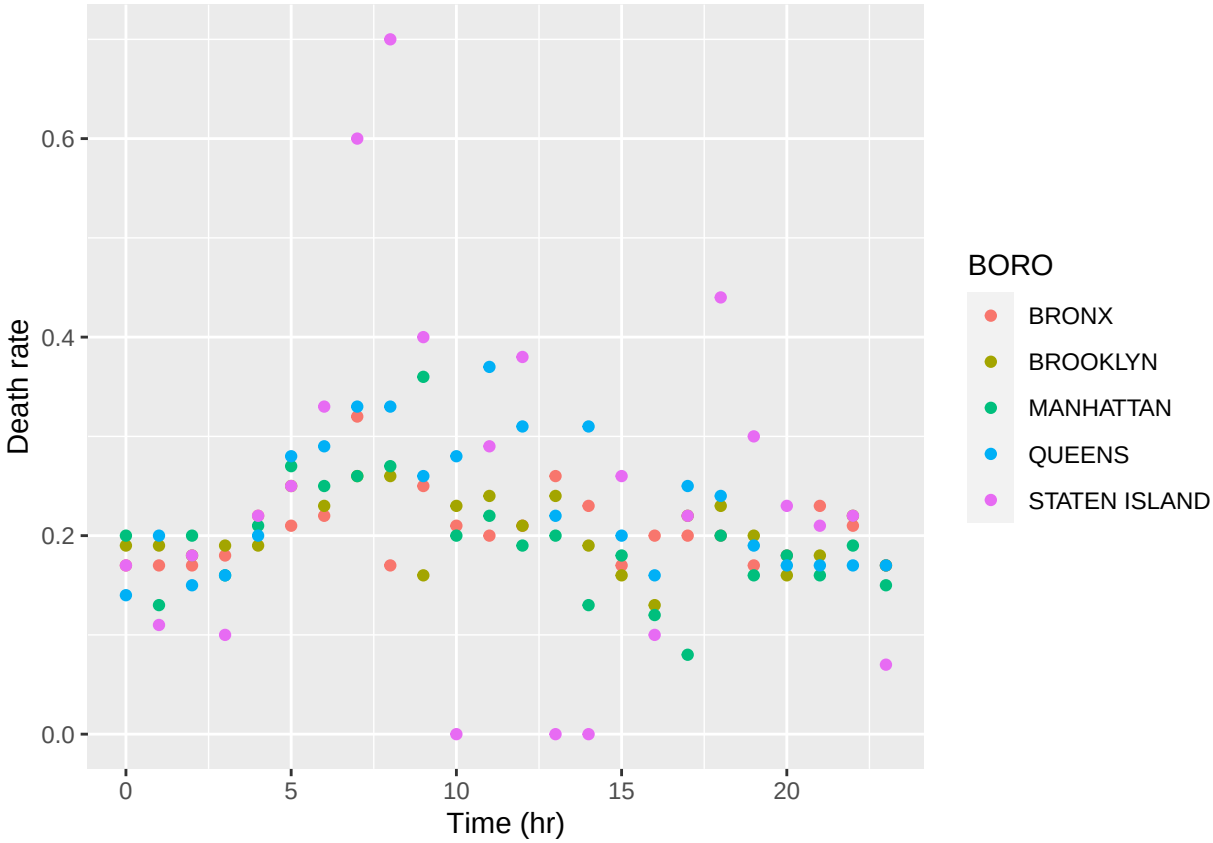
```
NYPD_model %>%
  ggplot(aes(x = OCCUR_HOUR, y = death_rate)) +
  geom_point() +
  labs(x = "Time (hr)", y = "Death rate")
```



I was surprised to see the highest death rate was early in the morning, at 7:00am. Another model was created to see how the death rates varied among the different boroughs.

```
NYPD_boro_model <- NYPD %>%
  group_by(OCCUR_HOUR, BORO) %>%
  summarize(incidents = n(), deaths = sum(STATISTICAL_MURDER_FLAG)) %>%
  mutate(death_rate = round(deaths/incidents, digits = 2)) %>%
  ungroup()
```

```
NYPD_boro_model %>%
  ggplot(aes(x = OCCUR_HOUR, y = death_rate, color = BORO)) +
  geom_point() +
  labs(x = "Time (hr)", y = "Death rate")
```



Sources of Bias

One area of bias for this project comes from the topics that I investigated. I visit NYC often so was interested in the timing and location of events to see if they overlapped with where I spend my time in NYC. Another source could be how the data was collected, especially if it is relying on witness testimonies.